Aalto University
School of Science
Master's Programme in Mathematics and Operations Research

Cosmo Jenytin

Should a machine be trusted?

A study on the effect of explainable artificial intelligence on trust in automated machine learning

Master's Thesis
Espoo, July 15, 2021

Supervisor: Professor Antti Punkka
Advisors: Jorge Torres MSc (Eng), MindsDB Inc.
Tuomas J. Lahtinen DSc (Tech), Aalto SCI

Aalto University
School of Science
Master's Programme in Mathematics and Operations Research

ABSTRACT OF MASTER'S THESIS

Author: Cosmo Jenytin

Title:
Should a machine be trusted? A study on the effect of explainable artificial intelligence on trust in automated machine learning

Date: July 15, 2021 **Pages:** vii + 132

Major: Systems and Operations Research **Code:** SCI3055

Supervisor: Professor Antti Punkka

Advisors: Jorge Torres MSc (Eng)
Tuomas J. Lahtinen DSc (Tech)

This thesis explores the effect of explainable artificial intelligence (XAI) on user trust in automated machine learning (AutoML). AutoML promises to democratize data science but can also decrease the transparency of machine learning (ML) models. To ensure appropriate reliance on AutoML systems, users must be able to calibrate their trust to match the actual reliability of the systems; transparency is essential for this. XAI has shown promising results, but its effect on user trust in AutoML, especially in information-heavy prediction contexts, such as time series forecasting, has received little attention.

To study the feasibility of utilizing XAI techniques to affect trust in AutoML we conduct a randomized controlled trial (RCT). We focus specifically on non-experts and multivariate time series predictions. We adapt and develop three XAI techniques for the experiment, based on a review of the literature on trust, AutoML, and XAI. Trust is measured both as attitudinal trust—based on a questionnaire—and behavioral trust, operationalized as a participant switching their prediction to align with that of the AutoML system.

We find that XAI has a statistically significant—but relatively small—effect on measured attitudinal trust, and that two XAI-enabled systems are trusted more and one less than the control. The effect on measured behavioral trust is the opposite: all treatment AutoML systems are trusted less, but this effect is not statistically significant. XAI has shown promise in increasing model transparency and user trust in previous research; nevertheless, our results indicate this effect is context-dependent. Earlier experiments have utilized somewhat simpler tasks than ours, and this is likely an important factor. Future work on XAI that provides transparency without overwhelming the users is encouraged, and limitations and additional avenues of future work are discussed.

Keywords: trust in automation, AutoML, XAI, RCT

Language: English

Aalto-universitetet
Högskolan för teknikvetenskaper
Master's Programme in Mathematics and Operations Research

SAMMANDRAG AV DIPLOMARBETET

Utfört av:	Cosmo Jenytin		
Arbetets namn:	Ska man lita på en maskin? En studie över effekten av förklarbar artificiell intelligens på förtroende för automatiserad maskininlärning		
Datum:	Den 15 juli 2021	**Sidantal:**	vii + 132
Huvudämne:	Systems and Operations Research	**Kod:**	SCI3055
Övervakare:	Professor Antti Punkka		
Handledare:	Diplomingenjör Jorge Torres Teknologie doktor Tuomas J. Lahtinen		

Detta diplomarbete utforskar effekten av förklarbar artificiell intelligens (XAI) på förtroende för automatiserad maskininlärning (AutoML). AutoML strävar efter att demokratisera data science, men kan också minska transparensen i maskininlärning (ML). För att säkerställa lämplig förlitan på AutoML-system, måste användare kunna kalibrera sitt förtroende så att det överensstämmer med systemens verkliga tillförlitlighet; transparens är väsentligt för detta. XAI har gett lovande resultat, men endast lite uppmärksamhet har ägnats åt dess effekt på användarnas förtroende för AutoML, speciellt i informationsdryga prediktionssammanhang, såsom tidsserieprediktion.

För att studera möjligheten att påverka förtroende för AutoML med XAI-tekniker utför vi en randomiserad kontrollerad studie (RCT). Vi fokuserar specifikt på icke-experter och multivariata tidsserieprognoser. Vi anpassar och utvecklar tre XAI-tekniker för experimentet, baserat på en litteraturöversikt över förtroende, AutoML och XAI. Förtroende mäts både som attityd – baserat på ett frågeformulär – och som beteende, definierat som att en deltagare ändrar sin prognos så att den stämmer överens med den prognos som AutoML-systemet gett.

Vi observerar att XAI har en statistiskt signifikant – men relativt liten – effekt på förtroende uppmätt som en attityd, och att förtroende är högre för två av AutoML-systemen med XAI och lägre för en, jämfört med kontrollgruppen. Effekten på förtroende uppmätt som beteende är motsatt: förtroende för alla behandlingsgruppers AutoML-system är lägre, men effekten är inte statistiskt signifikant. I tidigare forskning har XAI uppvisat lovande resultat med att öka transparensen hos modeller och användarnas förtroende för dem; trots detta tyder våra resultat på att denna effekt beror på sammanhanget. I tidigare experiment har något enklare uppgifter använts, vilket är sannolikt en viktig faktor. Framtida arbete kring XAI som ökar transparens utan att överväldiga användare uppmuntras, och begränsningar och andra möjligheter för framtida arbete diskuteras.

Nyckelord:	förtroende för automation, AutoML, XAI, RCT
Språk:	Engelska

Acknowledgments

This has been a long time coming.

I want to thank Jorge and MindsDB, for the idea for and your excellent guidance and support throughout this incredible interdisciplinary research project. Tuomas, thank you for encouraging me, for inspiring me to always aim higher, and for the endless brainstorming, discussions, and philosophizing around this thesis and other subjects. Antti, your insight and support was invaluable in shaping this project—which, admittedly, started as a rather vague idea—into the academically sound and—*absolutely impartially speaking*—extremely interesting thesis. Thank you Patrik and OpenOcean, for granting me this chance: this project turned out to be so much more than I could ever have hoped of a Master's thesis.

My friends and family: thank you for your patience, understanding, and support throughout. I am looking forward to spending time with you instead of spending all my waking hours studying. Thank you Alina, I could not have done this without you, or been able to gladly take the step out into the unknown future of life after graduation.

And the most special thanks, for making the last seven years so incredible, to all the wonderful people I have had the chance to meet and spend time with at Teknologföreningen. TF made me choose Aalto University, and I have not regretted it since. What TF turned out to be was so much more than I could have imagined: a place to grow as a human, and a community with whose help to find myself and my place in this world. As they say: "*Läroverket utbildar tjänstemannen, kamratskapet danar medborgaren*".

I wholeheartedly concur.

Otaniemi, July 15, 2021

Cosmo Jenytin

Contents

Glossary

Worker Amazon Mechanical Turk (MTurk) Worker; person who completes assignments (Human Intelligence Tasks) on MTurk

List of acronyms and abbreviations

ANCOVA	analysis of covariance
AutoML	automated machine learning
COVID-19	Coronavirus disease 2019
DL	deep learning
DTW	Dynamic Time Warping
GLMM	generalized linear mixed-effects model
HIT	Human Intelligence Task
ICU	intensive care unit
ML	machine learning
MTurk	Amazon Mechanical Turk
NN	neural network
PTT	propensity to trust
RCT	randomized controlled trial
SD	standard deviation
SHAP	SHapley Additive exPlanations
XAI	explainable artificial intelligence

Chapter 1

Introduction

Data-driven decision making is increasingly important in the world of today, with its abundance of data. Powerful data analysis tools are needed to uncover the insights hidden in this data. However, developing and applying advanced tools, such as machine learning (ML) models, and making data-driven decisions based on insights derived from them, require specialized skills, including data science skills, domain expertise, and stakeholder understanding. Clearly, one person seldom has all the required skills and knowledge. Automated machine learning (AutoML) aims to bridge this gap and "democratize" ML, by automating all or part of the data science workflow, to allow anyone to employ state-of-the-art ML models off-the-shelf (Elshawi and Sakr, 2020; He et al., 2021; Hutter et al., 2018; Waring et al., 2020). This allows those who maintain the data (such as developers and data engineers), domain experts with deep knowledge about the data itself, and decision makers to create and deploy ML models for their own uses, without extensive knowledge about the ML methods themselves.

To ensure appropriate reliance on new types of technology, including ML and AutoML systems, it is essential to enable users to calibrate their trust to a level corresponding to the actual reliability of technology (Drozdal et al., 2020; Hoff and Bashir, 2015; Lee and See, 2004; Ribeiro et al., 2016). This is especially important regarding AutoML systems, since humans are usually responsible for their predictions, even without insight into their inner workings. On the other hand, the general lack of transparency in ML models can impede this trust calibration, and AutoML can exacerbate this, since it abstracts away and automates parts of the data science process.

Unsurprisingly, explainable artificial intelligence (XAI) is quickly gaining popularity as a means to mitigate the opaqueness of models, and is actively

researched (Adadi and Berrada, 2018). XAI techniques can shed light on ML models and interpret their predictions in general or in specific instances, by utilizing, among others, visualizations, sensitivity analysis, variable importance values, decision rules extracted from the model, and example-based explanations. Previous research has indicated that employing XAI techniques, to explain the predictions of ML models, affects user trust in the system (Nourani et al., 2019; Yin et al., 2019). Similar results were observed for AutoML systems by Drozdal et al. (2020), in their case for students with previous ML experience. Nonetheless, research on the effect of XAI on trust in AutoML is scarce, and almost nonexistent for time series forecasting—a challenging, yet common, task for humans and AutoML alike.

This thesis studies the feasibility of utilizing XAI techniques to affect trust in AutoML, focusing specifically on non-expert users and a time series forecasting context. A randomized controlled trial (RCT) is conducted to study the effect of XAI techniques on trust in an AutoML system. Three different XAI techniques are developed and adapted for this experiment, based on a review and summary of XAI design recommendations.

This thesis is structured as follows. Chapter 2 starts with a brief overview of ML and AutoML, continuing with a review of trust in both interpersonal and human-automation contexts, followed by a summary of current XAI techniques and design frameworks. Chapter 3 describes the design and implementation of the RCT to study the effects of three XAI techniques on trust in an AutoML system based on MindsDB (MindsDB Inc, 2019); the XAI techniques used in the experiment are also described in detail. Chapter 4 presents the results of the experiment and offers a discussion on how explainability features affect trust in the underlying system, and discusses limitations and future work. Chapter 5 concludes with a discussion of the implications of the results for AutoML design and an outline of possible avenues of future research.

Chapter 2

Background: AutoML, trust, and XAI

2.1 ML: uncovering the insights hidden in data with the right tools and skills

Data-driven decision making is essential in the world of today, with an abundance of data. Within these bytes are obscured a massive number of insights that are hidden without the right tools at one's disposal.

To utilize this data and the insights it contains, powerful data analysis tools are required; tools such as ML and neural network (NN) are widely utilized for analyzing and predicting based on large data sets. As Jung (2020) summarized, ML tools are commonly employed for classification (labeling data into two or more classes), clustering (grouping data based on commonalities), prediction (predicting values of a target variable or variables based on explanatory variables), and dimensionality reduction (reducing high-dimensional data to lower dimensions).

An increasingly common type of data is time series: data is gathered continuously through sensors and other means, accruing in copious amounts in databases. Therefore, there is a growing supply of comprehensive time series data, which has enabled new applications of time series analysis, such as creating detailed forecasts of the electricity demand of individual households equipped with sensors (Coelho et al., 2017). In order to effectively analyze this data and make predictions of future developments, one needs specialized statistical tools that account for the temporal information and relations in the time series. ML models are finding use in time series forecasting, and ML

models (such as support vector machines) already "offer results similar to or better than those reached by state-of-the-art statistical methods" (such as seasonal autoregressive integrated moving average models) (Parmezan et al., 2019, p. 332).

Traditionally, developing and applying ML tools have required specialized skills; these requirements have led to a high demand for and shortage of data scientists to answer the demands of increasing amounts of data (Miller and Hughes, 2017). A typical data science workflow is shown in Figure 2.1; clearly, one person seldom has the data science skills, domain expertise, and stakeholder understanding required to execute all steps of this workflow and utilize the results in decision making. This fragmentation of skills and knowledge across specialists may introduce challenges in efficiently utilizing data and insights derived from ML (see, for example, Elshawi and Sakr, 2020; McAfee and Brynjolfsson, 2012; Schelter et al., 2018; Waring et al., 2020). For example, a CEO may make decisions based on the data and provide feedback on its relevance and validity (Model Validation and Model Improvement phases), and there may be data(base) engineers responsible for gathering and maintaining the data (Data Acquisition phase); facilitating the discussion between these persons and stakeholders with different roles and interests may be challenging or costly to arrange, especially in fast-paced business environments.

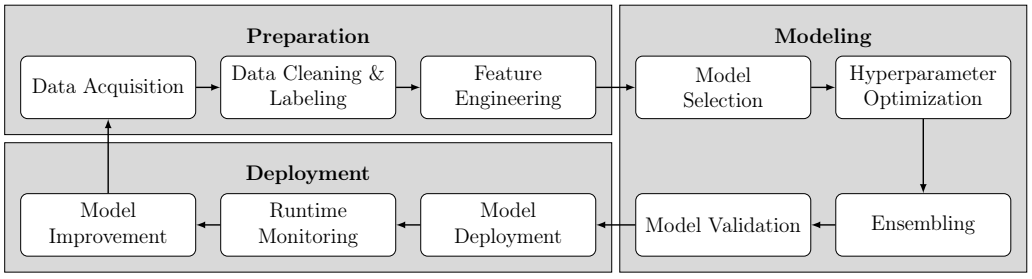


Figure 2.1: Typical data science workflow. Figure adapted from Drozdal et al. (2020).

2.2 AutoML: automating the data science workflow

AutoML is an emerging solution that aims to mitigate the heavy reliance of traditional ML development on human domain and task-specific expertise and the low supply of data scientists (Elshawi and Sakr, 2020; He et al., 2021;

Waring et al., 2020). In an AutoML system all or parts of the data science workflow are automated, helping people—including those without experience of data science or mathematical modeling—to utilize ML. With the help of AutoML, those who own and maintain the data (such as developers and data engineers), domain experts with deep knowledge about the data itself, and decision makers can create and deploy ML models for their own uses, without extensive knowledge about the ML methods themselves, and expert users can also benefit from less complicated parts of the development being automated.

AutoML software are quickly growing in numbers, spurred on by, among other things, competitions such as ChaLearn AutoML Challenge (Guyon et al., 2018). Some popular tools include:

- Auto-WEKA (`https://www.cs.ubc.ca/labs/beta/Projects/autoweka/`, Thornton et al., 2013)
- Auto-sklearn (`https://automl.github.io/auto-sklearn/`, Feurer et al., 2015)
- TPOT (`https://epistasislab.github.io/tpot/`, Fu et al., 2020)
- Amazon's AutoGluon (`https://auto.gluon.ai/stable/index.html`, Erickson et al., 2020)
- Google's Cloud AutoML (`https://cloud.google.com/automl`)
- Facebook's Prophet (for time series forecasting; `https://facebook.github.io/prophet/`, Taylor and Letham, 2018)

These and AutoML systems in general aim to provide advanced ML models off-the-shelf. Common methods employed by AutoML, outlined by Hutter et al. (2018), include model selection (selecting the best model for the task at hand), neural architecture search (selecting the best configuration of a neural network), hyperparameter optimization (optimizing the hyperparameters of the chosen model), and meta-learning (learning how to learn, that is, using experience from previous learning tasks to optimize learning in future tasks). These are all design decisions a data scientist would make when creating a ML model; automating these decisions may even make them more systematic and efficient, compared to the ad hoc nature of ML development today.

On the other hand, abstracting away and automating parts of the data science process introduces its own problems. For instance, ML models, and especially

deep learning (DL) models, already suffer from being opaque and uninterpretable or from inexplicable predictions (Adadi and Berrada, 2018). This perception of the black-box nature of the models can be further increased by not having insight into and control over every detail of the development and refinement of a model; this may decrease the *trust* users have in the model.

2.3 Trust

2.3.1 Interpersonal trust

Trust between humans is a familiar subject to everyone, and thus we start by examining trust as an interpersonal concept in order to define trust. Trust is essential for the smooth functioning of society. Each day, a person is faced with many kinds of potential threats and risks from other people, institutions, and nature. Trust helps an individual cope with this by reducing the perceived complexity of the world around them: "to trust is to live as if certain rationally possible [undesirable] futures will not occur. Thus, trust reduces complexity far more quickly, economically, and thoroughly than does prediction" (Lewis and Weigert, 1985); not trusting, on the other hand, leads to doubting everything and everyone around oneself (Giddens, 2013; Luhmann et al., 2018).

On the other hand, trust always incorporates a measure of risk: trusting someone or something means accepting the risk of being betrayed by that object of one's trust. This risk of negative consequences is central to trust and comes from the inability or unwillingness to control the trustee and their actions (McKnight and Chervany, 2001). Neoclassical economic theory sees trusting behavior as irrational for just this reason: a rational, profit-maximizing individual should not take the risk of trusting others; and yet, people generally and consistently exhibit trusting and other-regarding behavior (Evans and Krueger, 2009).

Interpersonal trust has been defined in many ways depending on context. It "has been defined as both a noun and a verb ... a personality trait ... a belief ... a social structure ... and a behavioral intention" (McKnight and Chervany, 2001, p. 28).

McKnight and Chervany (2001) identified four main characteristics of a *trustee* (the object of the trust of a *trustor*): *benevolence*, *competence*, *predictability*, and *integrity*. A *benevolent* trustee is willing to act in the best interest of a trustor. A *competent* trustee is able to act in the best interest

of a trustor. A *predictable* trustee acts in a consistent enough manner for a trustor to be able to predict their behavior on different situations. A trustee with *integrity* can be expected to make and keep good-will promises.

McKnight and Chervany (2001) then defined a typology of conceptual trust types, shown in Figure 2.2. *Disposition to trust* "means the extent to which one displays a consistent tendency to be willing to depend on general others across a broad spectrum of situations and persons" (McKnight and Chervany, 2001, p. 38). Thus, it refers to the general propensity of a person to be trusting of others, not whether they will trust a specific person. Disposition to trust is a stable personality characteristic, formed by previous experiences, and affects trust in a specific situation to varying degrees, depending on previous experience of that type of situation (Rotter, 1980).

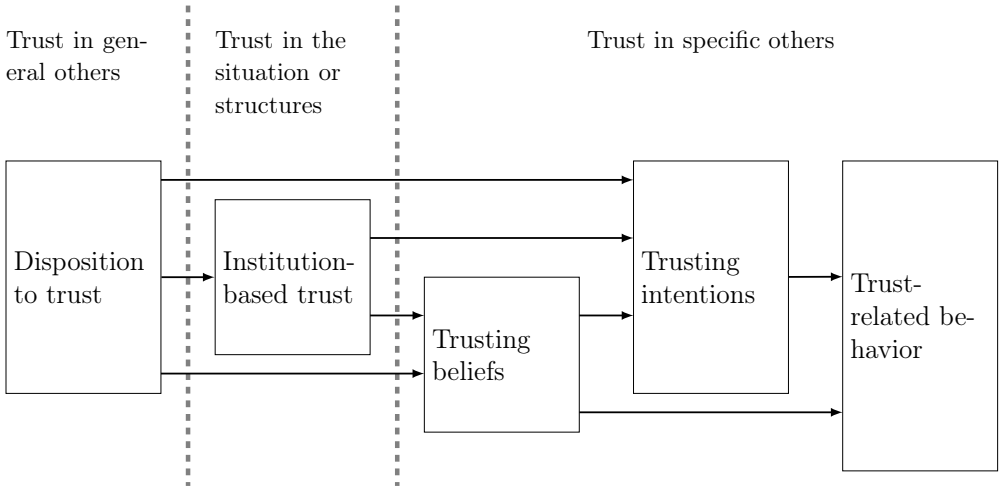


Figure 2.2: Trust constructs, adapted from McKnight and Chervany (2001). Arrows show causal links between the trust concepts.

Institution-based trust "means one believes, with feelings of relative security, that favorable conditions are in place that are conducive to situational success in a risky endeavor or aspect of one's life" (McKnight and Chervany, 2001, p. 37). Institution-based trust is not directed at a specific other, but instead the underlying structures of a situation or context itself, such as laws and law enforcement and government regulation.

Trusting beliefs "means the extent to which one believes, with feelings of relative security, that the other person has characteristics beneficial to one" (McKnight and Chervany, 2001, p. 36). Trusting beliefs are related to a specific other, and imply the trustor perceives that the trustee has one or more of the four main characteristics (benevolence, competence, predictability, integrity)

of a trustworthy person.

Trusting intentions "means one is willing to depend, or intends to depend, on the other party with a feeling of relative security" (McKnight and Chervany, 2001, p. 34). This willingness is directed at a specific person generally, that is, not in a specific situation.

Trust-related behavior means one "depends on another person with a feeling of relative security, even though negative consequences are possible", and thus gives "gives the trustee some measure of power over the trustor" (McKnight and Chervany, 2001, p. 34). Trusting behavior is directed at a specific person.

These trust types are of course related and affect each other, as shown by the arrows in Figure 2.2: trusting beliefs lead to trusting intentions, and they both lead to trusting behavior—indeed, "attitudes significantly predict future behavior" (Kraus, 1990); trust in a situation (institution-based trust) leads to also trusting the persons in the situation; disposition to trust describes what one believes about others generally, and should thus affect all other kinds of trust, at least in unfamiliar situations where there are no previous experiences to fall back on.

2.3.2 Trust in automation

Trust in automation is similar to interpersonal trust, and humans often apply social norms when interacting with machines (Hoff and Bashir, 2015; Madhavan and Wiegmann, 2007). Similar to interpersonal trust, Hoff and Bashir (2015) find that variations in trust in automation are explained by dispositional, situational, and learned trust factors. Situational trust refers to differences in context, such as characteristics of the automation (for example, its complexity and reliability) or the situation (such as situation novelty, task, user self-confidence in manual completion of task, user workload, freedom of the user to choose to rely on automation). Learned trust is based on pre-existing knowledge (such as knowledge of and experience with the specific automation system, or a similar one), in addition to dynamically learned trust (that is, trust based on perceived performance of the system during interaction). Just as interpersonal trust, trust in automation is also dynamic: dispositional trust (general propensity to trust machines) and pre-existing knowledge dominates trust at the beginning of interaction with a machine, but its effect gradually decreases as trust is calibrated by perceptions of the performance and reliability of the machine (Hoff and Bashir, 2015; Merritt and Ilgen, 2008).

However, there are some key differences between interpersonal and human-automation trust, especially due to the biases affecting trust formation and calibration in automation (Hoff and Bashir, 2015; Madhavan and Wiegmann, 2007). Differences are at least partly due to the different cognitive schemas humans have for automation and other humans (for example, human experts). Violations of expectations based on these schemas are especially likely to be noted and remembered, and thus have a disproportionate effect on, for instance, trust. Hence, this holds for automation errors, since humans in general have a schema of perfect automation and a positivity bias towards automation (Dzindolet et al., 2003); on the other hand, humans generally have a schema of imperfect human advisors, and are therefore more likely to be forgiving of mistakes. In addition, social norms that make people reluctant to claim they distrust someone do not apply to machines, exacerbating the labeling of machines as untrustworthy (Madhavan and Wiegmann, 2007).

Madhavan and Wiegmann (2007) also present differences between trust in humans and automation in three important factors affecting trust: the source of information (human advisor or automated aid), source credibility, and source reliability. Concerning the source of information, Madhavan and Wiegmann (2007) found that advice from an automated aid is perceived as more trustworthy, since automated aids are seen as more objective and rational. On the other hand, the fundamental attribution error often leads to people blaming computers for mistakes—which are also more likely to be noticed due to schema violations—and not giving credit for successes. In the context of decision support systems, the self-confidence of a human in their ability to perform a given task affects their trust in the automated aid: low confidence leads to higher trust in the aid, and vice versa.

Source credibility affects trust in both humans and automation, especially when a person is not capable of judging the trustworthiness of information based on its contents, but rather judges it based on superficial characteristics, such as presumed expertise of the source of information. In this regard, the credibility of automation is judged solely based on knowledge. Humans, on the other hand, are treated more comprehensively: in addition to their knowledge, their experience and effort are factored into this judgment.

Source reliability refers to the true reliability of the aid, either human or automated; reliability, and especially whether trust is calibrated to the true reliability, is an important determinant of reliance on automation. Users are sensitive to the true reliability (accuracy) of automation, and their trust in it is affected by both the stated and observed reliability (Yin et al., 2019); users are also sensitive to changing levels of reliability (Madhavan and Wiegmann,

2007). On the other hand, violations of expectations (expecting high reliability but observing low reliability) damages trust in automation more than trust in humans; moreover, automation errors in "easy" cases (as judged by the user) are penalized more than in "hard" cases, even though the automated aid may on average perform better than the user.

Furthermore, Deley and Dubois (2020) argue, based on the trust philosophy of Baier (1986), that humans cannot trust technology, only rely on it. Since trust must include the possibility of being betrayed by the trustee, technology can only be relied upon: it can disappoint us when it fails, but not betray us. Thus, Deley and Dubois (2020) argue that, of the four main characteristics of a trustee—benevolence, competence, predictability, and integrity—only competence and predictability apply to machines. However, the benevolence and integrity of the designers of the technology may be indirectly perceived as characteristics of the technology (Hoff and Bashir, 2015); for example, knowing that the developer of a software are not benevolent (for example, using the data gathered from their customers for malevolent purposes), may decrease the trust in the software. Nevertheless, this suggests efforts to calibrate trust in technology itself should focus on demonstrating competence and predictability, while showing the benevolence and integrity of the designers may also be beneficial to trust calibration.

2.3.3 Trust as a driver for technology adoption and appropriate use

To attain the benefits of new technology it must be adopted and utilized. New and improved technology has laid the foundation for and been an integral part of the rapid and continuing economic growth since the Industrial Revolution (Mokyr, 2005). On the other hand, new technology has not always been met with enthusiasm: in 1811 and 1812, a group of textile workers calling themselves Luddites protested the adoption of new machines replacing human workers by threatening machine owners and destroying machines (Jones, 2013).

Trust is central in the decision to adopt new types of technology, in human-technology interaction in general, and to appropriate utilization of and reliance on technology (Lee and See, 2004; Parasuraman and Riley, 1997). More recent examples, than the textile machines protested by the Luddites, of technology adoption where trust plays a significant role include mobile pay systems (Chandra et al., 2010), e-commerce (Van Slyke et al., 2004), and self-driving cars (Adnan et al., 2018).

The relationship of trust in and reliance on automation is also affected by situational factors (Hoff and Bashir, 2015). A high automation system complexity, high task novelty, high ability to compare automation performance to unaided performance of the user, and high degree of freedom of the user to choose whether to rely on the automation system all strengthen the relationship of trust and reliance.

To ensure appropriate reliance on automation and machines, the trust of the user must be calibrated to the actual reliability and trustworthiness of the system (Lee and Moray, 1994; Lee and See, 2004). Trust in automation that does not match its actual reliability can lead to misuse or disuse of automation. Misuse refers to inappropriate reliance, or "blind trust": for example, relying on automation in situations or tasks that it is not designed for. Disuse refers to inappropriately declining to utilize automation: for instance, failures resulting from refusing to use automation even in cases where its performance is far superior. With trust calibrated to an appropriate level, one is more likely to appropriately rely on automation.

Trust is likewise important for the adoption, deployment, and appropriate use of and reliance on ML and especially AutoML systems (Drozdal et al., 2020; Hoff and Bashir, 2015; Ribeiro et al., 2016; Yin et al., 2019). Appropriate use and reliance is especially important for AutoML systems, since they can be utilized to derive insights from data—possibly without human intervention—but users are still the ones to make the decision to trust the insights and recommendations and act upon them, while bearing the responsibility, regardless of outcome.

2.3.4 Measuring trust

Different ways of measuring trust have been employed when studying trust, most prominently in the form of questionnaires and directly through trusting behavior, or a combination of both (Hoff and Bashir, 2015). Both are related, as discussed above and shown in Figure 2.2: trusting attitudes lead to trusting behavior.

Questionnaires measure attitudinal trust, often with the underlying assumption that attitudinal trust will predict trusting behavior—which is the assumed outcome of trusting attitudes. Indeed, the meta-analysis of Kraus (1990, p. 4) showed that "attitudes significantly predict future behavior", with a combined $p << .000000000001$, and an average attitude-behavior correlation 0.38. Questions on attitudinal trust are nevertheless often fairly general, as for instance in the General Social Survey (`https://gss.norc.org/`),

which includes questions on attitudinal trust, such as, "Generally speaking, would you say that people can be trusted or that you can't be too careful in dealing with people?". Glaeser et al. (2000) found that general attitudinal survey questions such as these predict trustworthiness and trustworthy behavior in respondents, rather than trusting behavior; this may be due to the tendency of people to project their own beliefs and overestimate how similar other people are to them (false consensus effect, Ross et al., 1977). In the experiments of Glaeser et al. (2000), trusting behavior is instead predicted by past trusting behavior (such as lending money or personal possessions to friends), and some more precise attitudinal questions, related to the experiment task (for example, asking about one's general trust in strangers, when the task at hand is about trusting strangers).

Trust can also be measured more directly by measuring trusting behavior. Especially in the field of economics, trusting behavior is measured through experiments with different types of "games" (Evans and Krueger, 2009). In these games, the first player (trustor) can choose to trust the second player (trustee); this has the potential for greater returns than distrust, but also the risk of betrayal by the trustee (and thus lower returns). In this case, the feelings of trust of the trustor determine the level of trust demonstrated by their action. Similarly, trust, and specifically reliance, can also be operationalized by, for example, how often a subject chooses to align with or rely on a trustee (such as an automated decision aid) (Merritt, 2011; Yin et al., 2019).

Multiple scales have been specifically developed and validated for measuring attitudinal trust in different kinds of technology, automation, and computers (for example, Gulati et al., 2019; Jessup et al., 2019; Madsen and Gregor, 2000; Merritt, 2011; Merritt and Ilgen, 2008; Schaefer, 2016). Different scales have varying degrees of specificity, ranging from vague (for example, referring to technology in general) to context-specific (referring to the specific piece of technology at hand). Jessup et al. (2019) suggests that using a more specific and tailored scale better reflects the trust in the specific tool referred to; this is in agreement with the findings of Glaeser et al. (2000) for interpersonal trust.

Trust in automation has also been measured through trusting behavior, mainly as reliance on automation. For example, in addition to measuring attitudinal trust, Yin et al. (2019) operationalized trust in an automated decision aid as the reliance on it, measured by the "switch fraction", that is, the share of times the aid disagreed with the initial choice of the subject and the subject changed their choice to agree with the aid, after seeing its

advice. Merritt (2011) measured reliance similarly, and showed attitudinal trust predicted this reliance; likewise, Dzindolet et al. (2003) found trust to be important in automation reliance decisions, further supporting the use of reliance as an indirect measure of trust.

2.4 Calibrating trust in and reliance on AutoML through explainability

A lack of transparency in ML and AutoML can impede calibration of user trust in them, and thus increasing transparency is a key objective in AutoML design (Adadi and Berrada, 2018; Drozdal et al., 2020). Unsurprisingly, XAI and interpretable models are gaining popularity, aiming to both mitigate the general opacity of models and to calibrate the trust of users in the models, and are actively researched (Adadi and Berrada, 2018).

Calibration of trust in ML and AutoML through explainability assumes that trust formation requires information and transparency. This is a natural assumption, given the trust framework presented in this chapter: of the four characteristics of a trustworthy trustee, at least competence and predictability need to be present for trust in machines to form; in order for a user to gauge the competence and predictability of, for example, an AutoML system, the system must be able to demonstrate these characteristics. Without sufficient transparency this may not be possible: the user may be left asking more questions than the system is answering with its oracle-like predictions, without any explanations.

2.4.1 XAI: more transparency through explanations

Adadi and Berrada (2018) categorize XAI techniques based on three axes: intrinsic–post hoc, global–local (scope), model-specific–model-agnostic. The first axis, intrinsic–post hoc, is related to the complexity of the model to be explained and whether the model itself can be explained, or whether post hoc explanations must be utilized. On one end of the spectrum are intrinsically explainable models, such as decision tree based models: their predictions are based on rule-based decision trees, and these rules can simply be shown to the user to explain the prediction. However, Adadi and Berrada (2018) note that intrinsically explainable models like this come at a cost of accuracy: more accurate models are often more complex and black-box. These less transparent and intrinsically explainable models can be explained using post

hoc explanations. On the other end of the spectrum are post hoc explanations which are employed separately from the model, often requiring no insight into the model itself. A common approach is approximating the model with simpler and intrinsically explainable models, such as linear models; a simple such technique is LIME (Ribeiro et al., 2016). LIME is a special case of SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), another powerful and popular post hoc explanation technique.

The second axis, global–local scope, differentiates between XAI techniques that strive to explain the model and its predictions globally (in general), and techniques that explain instances locally (that is, specific predictions made by the model). Examples of global explanations include the intrinsically explainable rule and decision tree-based models; Adadi and Berrada (2018) also summarize other methods, among them a technique which builds a global interpretation tree based on local explanations. Nevertheless, they conclude that global explainability can be difficult to attain, due to the complexity of most models.

Thus, local explanations—explanations for single instances (predictions)—are more easily constructed, and more readily available. LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) are popular local explanation techniques. Other local explanation techniques include gradient-based techniques, especially common in image classification (known as saliency maps, which show regions that are influential in the classification of the picture), and variable importance explanations that compare the original prediction to predictions made by omitting a subset of the variables.

The third axis, model-specific–model-agnostic, separates explanation techniques that are tailored for specific models (such as for specific neural network architectures) and techniques that work with any (black-box) model. As Adadi and Berrada (2018) note, intrinsic explanations are by definition model-specific. Model-specific explanation techniques are limited in their application scope, and thus the majority of explanation techniques developed are model-agnostic.

Adadi and Berrada (2018) further divide model-agnostic explanations into four types. Firstly, visualization-based methods aim to explain the predictions of the model visually. For example, a simpler model can be fitted locally around the data, as LIME (Ribeiro et al., 2016) does; partial dependence plots, on the other hand, visualize the (conditional) effect a subset of variables (often one or two) has on the predictions of the model, allowing the user the analyze the general relationship of variables and the predictions (Friedman, 2001).

Secondly, knowledge extraction techniques strive to distill the knowledge and internal representations the model has learned into understandable form. For example, decision rules may be extracted from black-box models based on input and output.

Thirdly, influence methods show the impact of specific variables and model parameters. For example, sensitivity analysis can be used to study the impact of the input variables and model parameters, and feature importance methods show the impact of variables by changing the value of variables and recording the impact on the accuracy of the model—more influential variables will have a bigger (negative) effect on the accuracy.

The fourth and last type of model-agnostic explanations, example-based explanations, explain the model through specific examples of data. Prototypes that represent clusters of similar data can be constructed to show how the model behaves for distinct types of data (for example, how they are classified). Counterfactual explanations, on the other hand, explain through opposites: these methods show the minimum changes required to get an alternative prediction (such as another label in a classification task, or a value above or below a chosen threshold in a regression task). For example, Mothilal et al. (2020) propose an explanation technique that finds a set of diverse and feasible counterfactual examples to explain the prediction of any ML model.

On the other hand, designing and using ML for time series modeling is in itself challenging, and designing explanations for this context requires some tailoring to take into account the special characteristics of time series data. XAI research has been strongly focused on contexts other than time series, but previously developed methods have been adapted for time series use, and explanations methods specifically designed for time series are increasingly researched and developed.

Local and model-agnostic methods, such as SHAP and LIME, have been successfully adapted for time series classification and forecasting (Mujkanovic et al., 2020; Ozyegen et al., 2020); similar methods have also been developed specifically for time series classification (Guillemé et al., 2019). One of the challenges is the large number of "variables" due to the temporal dimension: variables vary through time, and each value can influence the prediction. This can affect the explanations: as Alvarez-Melis and Jaakkola (2018) showed, SHAP and LIME can be sensitive to small changes in the data. To mitigate this for time series explanations, Mujkanovic et al. (2020) suggested using a time slice mapping, that is, partitioning the time series into small slices (of more than one time step), and treating these as the variables whose influence

is calculated for the explanation.

Example-based explanations have also been applied for time series. Among others, counterfactual explanations have been proposed for time series classification (Ates et al., 2020; Delaney et al., 2020). These techniques work much like their non-time series counterparts, with just different distance metrics, whose choice naturally affects what constitutes as the smallest required perturbation in the data to get an alternative prediction (the counterfactual). Prototypes, another example-based explanation method previously employed in non-time series explanations, have also been proposed as an explanation technique for time series classification (Gee et al., 2019).

These and the XAI techniques discussed by Adadi and Berrada (2018) are designed for explaining ML models; as such, they apply to the models created by AutoML, but AutoML-specific explanations may also be needed. Since AutoML can automate the model creation and selection, explanations of these phases can increase the transparency of the AutoML system. With a transparency perspective as a focus for trust in AutoML, Drozdal et al. (2020) studied what kinds of information students with machine learning experience wished an AutoML system to present. They indeed found that users considered information about the data (both raw and preprocessed) and model creation and selection process important, in addition to information about the final model chosen by the AutoML system (provided by, for example, model performance metrics and the XAI methods outlined above).

2.4.2 Designing XAI that is meaningful for humans

AutoML can democratize ML, by enabling even those without extensive data science skills to employ powerful data analysis tools; on the other hand, this also means AutoML users are generally not data science experts, and thus it is especially important that the explanation techniques utilized by such tools should also be meaningful for non-experts. Human-meaningful explanations align with human logic and judgment and are understandable to humans (Nourani et al., 2019).

Indeed, Nourani et al. (2019) demonstrated that the human-meaningfulness of the explanations given by an automated system significantly affected how accurate the system was perceived to be: less human-meaningful explanations led to underestimation of system accuracy. Thus, the meaningfulness of explanations may affect the trust calibration of users, since system accuracy can be an important part of the information users utilize in calibrating their trust in the system. This suggests that, in order to calibrate trust in

automated systems (to match their actual accuracy and reliability), XAI development should focus on explanations meaningful to humans. Liao et al. (2020) also found that current XAI techniques are often not user-friendly and suggested a more user-centered approach for XAI design.

In this vein, Miller (2019) suggested utilizing the vast body of research on explanations that the social sciences have contributed, to design explanations that are "good" and meaningful to humans in general, not only researchers. They found that human-human explanations are

1. **contrastive**: people mostly seek explanations to counterfactual, often unexpected outcomes. That is, people do not ask questions like "Why did this event happen?", but instead ask "Why did this event happen instead of this other (expected) event?"

2. **selective**: explanations are not expected to describe the complete cause of an event, but instead people choose "one or two causes from a sometimes infinite number of causes to be *the* explanation" (Miller, 2019, p. 3).

3. **social**: information is transferred through interaction and conversation, and the explainer adjusts their explanation based on their beliefs of the knowledge and beliefs of the explainee. An explanation is most often given interactively, so the explainee can question the explainer to get an explanation that matches their knowledge; this also ensures that the explanations that are given are relevant.

4. **not based on probabilities**: basing an explanation on probabilities is less effective and satisfying than basing it on causes; indeed, "the most likely explanation is not always the *best* explanation for a person" (Miller, 2019, p. 3).

These findings suggest shifting the focus of XAI design towards explanations that are: *local*, since explanations are most often needed for single, unexpected instances; *contrastive*, that is, explanations that explain through, for example, counterfactual examples; *simple*, since people prefer selective, but sufficient, causes as explanations; *non-probabilistic*, that is, explanations should focus on direct causes, not likelihoods or statistical relationships; and *social and interactive*, since explanations are normally given as part of a conversation—allowing the user to prompt a system with questions would also ensure that superfluous explanations are not given.

Chapter 3

Experiment: the effect of explanations on trust in AutoML

We conducted a RCT to study the feasibility of affecting trust in AutoML by utilizing XAI techniques. The study compared a baseline (control) AutoML system to three different systems (treatments), each employing a different XAI technique developed and adapted for the experiment based on the findings in Chapter 2 and especially Section 2.4.2. The underlying models were identical in all four cases; the control system only provided predictions graphically, while the treatment systems supplemented the predictions with an explanation.

Figure 3.1 illustrates the experiment procedure for a participant in one of the four treatment groups. The general structure of the experiment consisted of the following six steps:

1. **Introduction and control variable measurement**: rough description of ML, AutoML, and the survey itself; giving informed consent for storing and analyzing anonymized survey data; agreeing to not go backwards in the survey; measuring control variables.

2. **Instructions**: detailed description of the data, task, and tools; in addition, for treatment groups, a description of the explanations of the AutoML system.

3. **Tasks 1–10**: based on historical data in four variables, making initial and final dichotomous predictions (yes/no)—before and after seeing the prediction (and, for treatment groups, an explanation of the prediction) of the AutoML system, respectively—to answer a question of the form "Will the adult intensive care unit (ICU) bed utilization exceed the

given threshold in the coming seven days?", in ten prediction tasks with different data.

4. **Trust measurement**: filling in a questionnaire to measure attitudinal trust.

5. **Amazon Mechanical Turk (MTurk)-specific questions**: filling in MTurk Worker ID for verifying completed surveys and awarding bonus based on correct predictions.

6. **Submission and debriefing**: description of the aims of the experiment; reviewing predictions compared to ground truth; group-specific completion code to submit in MTurk to get paid.

This chapter starts by describing the prediction context (data and prediction task), moving on to detail the treatment XAI techniques adapted and developed for the experiment. Then the experiment measurements—including control variables and trust metrics—are outlined, followed by a description of the participants and their recruitment. The chapter concludes with some remarks on the practical implementation of the experiment survey.

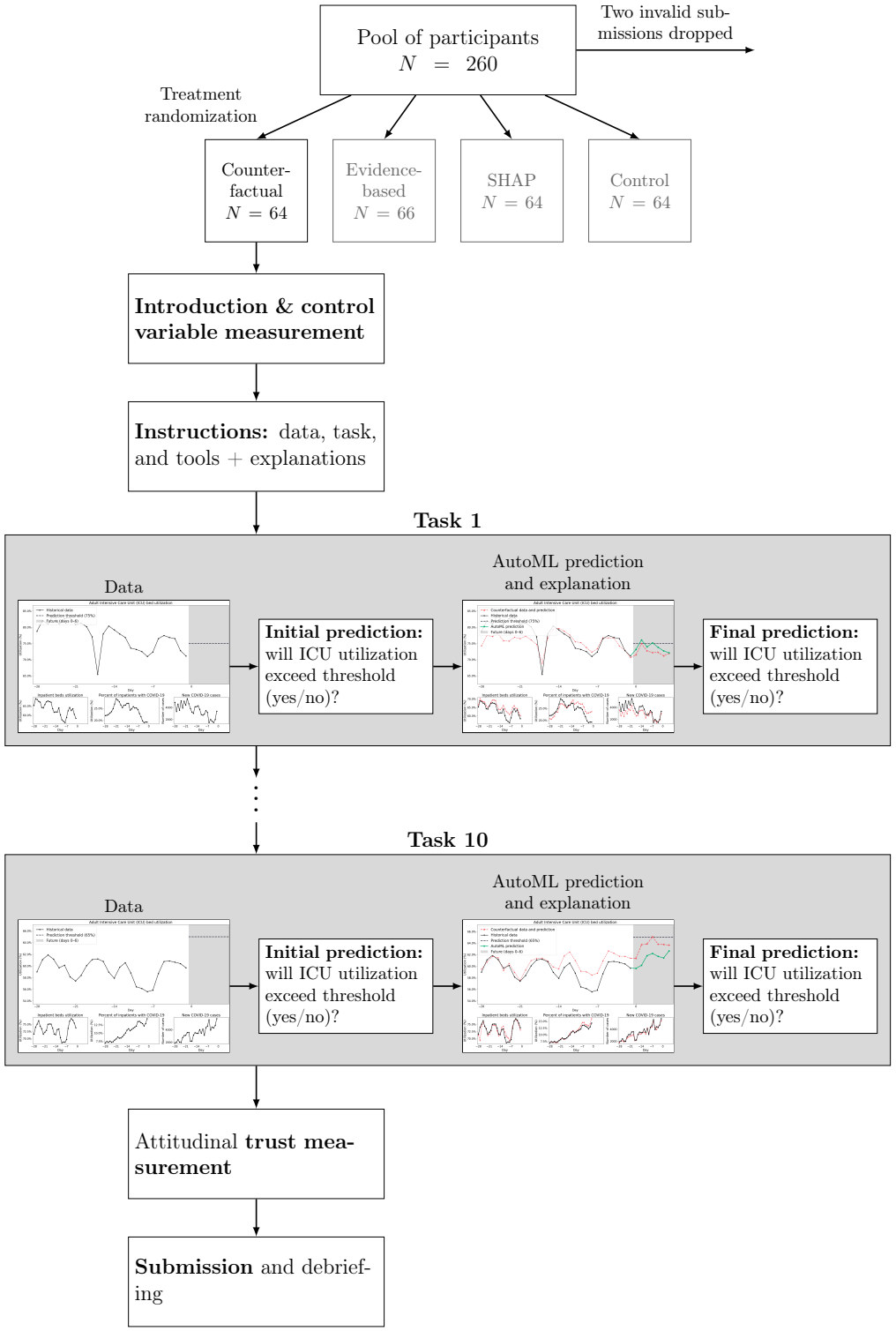


Figure 3.1: The RCT experiment procedure, illustrated for one treatment group.

3.1 Experiment task context: multivariate time series forecasting of ICU bed utilization

The AutoML, based on MindsDB (MindsDB Inc, 2019), was taught to predict seven days of adult ICU bed utilization in a state in the USA, based on daily state-level data of adult ICU bed utilization, inpatient bed utilization, percent of inpatients with (suspected or confirmed) Coronavirus disease 2019 (COVID-19), and new COVID-19 cases for the previous 28 days as explanatory variables. These variables were chosen from the following initial set:

- State (ID)
- Adult ICU bed utilization
- Inpatient bed utilization
- Percent of inpatients with (suspected or confirmed) COVID-19
- Percent of total inpatient beds occupied by patients with (suspected or confirmed) COVID-19
- Total number of hospitalized patients with confirmed COVID-19
- Total number of hospitalized patients with suspected or confirmed COVID-19
- New COVID-19 cases
- Probable new COVID-19 cases

We found that the two variables describing total numbers of hospitalized patients with COVID-19 (either "confirmed" or "suspected or confirmed") could vary significantly from day to day, due to varying degrees of reporting coverage and accuracy by hospitals; these were thus dropped before further analysis.

Data for all but the last seven days for each state was utilized as the training data for fitting the AutoML. The out-of-sample predictions for the last seven days were used when calculating accuracy metrics—for choosing the final subset of variables—and as the predictions utilized in the experiment.

The final subset of variables chosen for the experiment consisted of inpatient bed utilization, percent of inpatients with COVID-19, and new COVID-19

cases, in addition to the static state ID and the predicted variable adult ICU bed utilization. The final variables were chosen based on striking a balance between maximizing standard time series out-of-sample accuracy metrics (mean absolute error, mean absolute percentage error, root mean square error) and minimizing the number of variables to keep the task as simple as possible for both the participants and explanation methods. Different combinations of variables gave similar results on the accuracy metrics; therefore, the number of variables was the most important attribute, and resulted in the final set of variables that is described below.

Adult ICU bed utilization (ICU)—the predicted variable—is the percentage of ICU beds for adults in use in the hospitals of a state. An ICU bed is a "a unit in a hospital providing intensive care for critically ill or injured patients that is staffed by specially trained medical personnel and has equipment that allows for continuous monitoring and life support" (`https://www.merriam-webster.com/dictionary/intensive%20care%20unit`, accessed 5 May 2021).

Often, the ICU bed capacity is small compared to the whole capacity of the hospital. In addition, patients in need of ICU beds are in grave need of ICU care. Thus, a fully utilized ICU capacity is often not preferred, in case of a sudden increase in demand for ICU beds—such as due to the COVID-19-pandemic (Arabi et al., 2002).

Inpatient bed utilization (IB) is the percentage of inpatient beds that are in use in the hospitals of a state. Inpatient beds are hospital beds (of any kind) for patients that either require a bed or will have to stay one or more nights at the hospital. Inpatient beds include ICU beds, but ICU beds usually make up only a small portion of all inpatient beds.

Inpatient beds are often used to describe the capacity of a hospital, since no free inpatient beds means that a hospital cannot admit any new patients to inpatient care. If inpatient bed utilization is high, there may be a risk of ICU bed utilization also increasing soon: if the state of patients (whether with COVID-19 or not) worsens, they may be transferred to ICU beds.

Percent of inpatients with COVID-19 (ICov) is the percentage of inpatients (that is, patients occupying an inpatient bed) in the hospitals of a state that have suspected or confirmed COVID-19. Since COVID-19 patients can develop severe symptoms, there is a risk of them being transferred to ICU beds; about 30 % of hospitalized COVID-19 patients are transferred to the ICU (CDC, 2021b). Therefore, if the percentage of inpatients is high or increases, there can be a higher probability of patients being transferred to ICU beds. In addition, the median length of hospitalization among COVID-

19 survivors is 10–13 days, which means COVID-19 patients may occupy beds for a relatively long period.

New COVID-19 cases (NCov) is the number of new laboratory-confirmed COVID-19 cases in a state. Every COVID-19 case has some chance of resulting in a hospitalization. Every new case can therefore lead to a higher adult ICU bed utilization: the median time from the onset of COVID-19 to ICU admission is 9.5–12 days (CDC, 2021b).

State ID was only used to enable modeling differences between states. Since the state ID may include inherent information that only the participants may possess (such as knowing how a specific state has managed COVID-19 during the pandemic), it was not shown to the participants in order not to put the AutoML system at a disadvantage against the participants. Likewise, dates were not shown to participants, since time was utilized by the AutoML only implicitly, through the history in the data for each prediction task (and, for example, not as explicit dates).

We used data from HHS (2021) for data relating to hospital capacity and CDC (2021a) for data relating to COVID-19 cases. The data from HHS (2021) was downloaded on 22 December 2021 for the dates 7 September 2020–9 December 2020; data for the dates 26 July 2020–6 September 2020 was downloaded on 26 January 2021. The data from CDC (2021a) for the corresponding dates was downloaded on 18 January 2021. When merged based on the date, this combined dataset included 77 columns of daily data for 55 states and territories in the USA for the period 26 July 2020–9 December 2020.

Some preprocessing of the data was required. Missing values and clear outliers (hospital capacity utilization values greater than 120 %, as these were most likely due to reporting errors) in the variables were replaced with linearly interpolated values. In the case of the data from HHS (2021), two of the states (MP/Northern Mariana Islands and GU/Guam) were removed due to only having data for two dates each—not even enough to make the seven days of predictions. In addition, when there was conflicting data for the same date, the more recently uploaded one was used; the one exception was 2 August 2020, where two clearly different datasets were uploaded for that same date, while 3 August 2020 was completely missing from the database—in this case, the more recent of them was manually changed to have the date 3 August 2020, to get daily data without gaps.

The temporal structure of the data was retained through a transformation of the data, by concatenating historical and future data as columns, allowing the date column to be discarded. Figure 3.2 demonstrates the transformation

for the final subset of the variables; all variables (except the state ID, the sole static one) were transformed in the same way. After this transformation, each time step for each state was represented by one row with the state ID, 28 columns per variable for their history, and seven columns for the future of the predicted variable (adult ICU bed utilization).

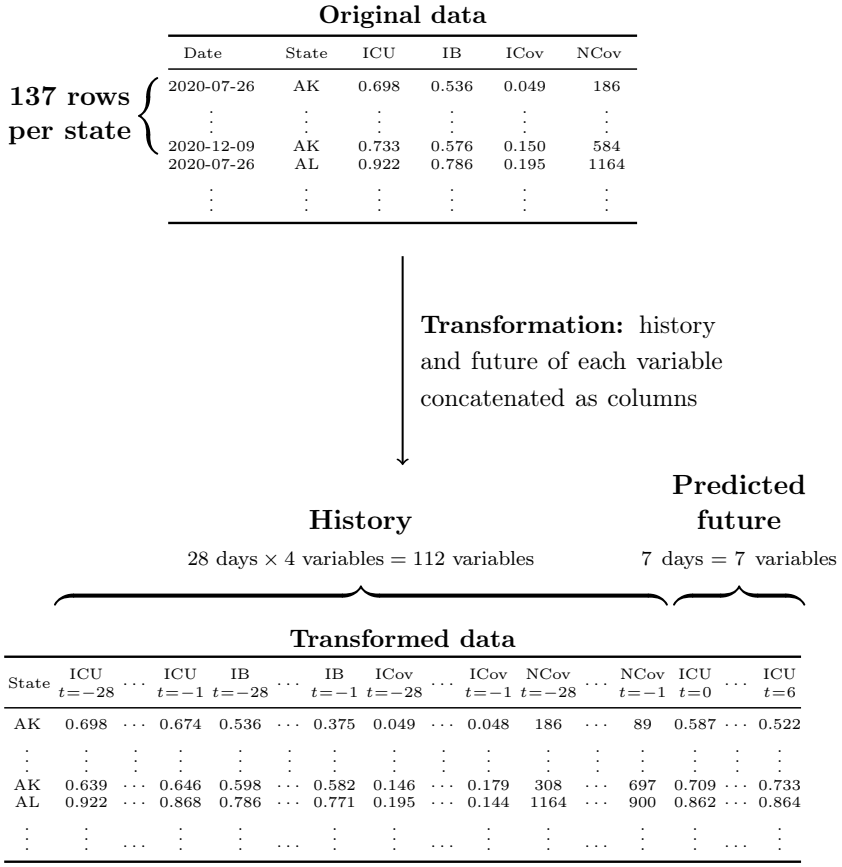


Figure 3.2: Illustration of how the data was transformed, by concatenating the history—28 days for each of the final four hospital and COVID-19 related variables chosen for the analysis—and the future—seven days of the target variable (adult ICU bed utilization)—as columns in the data.

3.1.1 Experiment tasks: dichotomization, two-part predictions, and choosing the final subset of tasks

Given the complicated context of multivariate time series predictions, the task of experiment participants was made as easy and concrete as possible

by making the predictions dichotomous: the participants answered questions of the form "Based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6)?". Predictions like these are relevant for real decision making during the COVID-19-pandemic: for example, in New York "regions must have at least 30 percent of their ICU beds available before a phased re-opening can begin" (that is, the ICU bed utilization must be below 70 %) (`https://forward.ny.gov/metrics-guide-reopening-new-york`, accessed 6 May 2021).

To allow us to measure directly the effect of the AutoML system and XAI technique on the predictions of the participants, the participants were asked to make two predictions per prediction task: an *initial prediction*, based only on the data and before seeing the prediction of the AutoML, followed by a *final prediction*, after seeing the prediction (and, for treatment groups, the explanation of the prediction) of the AutoML system. This way, if the initial prediction of participants did not match the prediction of the AutoML system, the participants could demonstrate trusting behavior by changing their mind and choosing their final prediction to match the one made by the AutoML system.

Figure 3.3 shows an example of the beginning of one of the ten prediction tasks in the experiment, asking for the initial prediction of the participant. The participants were also asked to indicate their confidence in each of their (initial and final) predictions, to include the possibility of us exploring changes in prediction confidence when analyzing the results.

The thresholds (75 % in the above example and in Figure 3.3) were tailored for each prediction task. The thresholds were chosen to both make the prediction tasks challenging enough (that is, both outcomes had to seem feasible) and to have multiple cases where the prediction of the AutoML was wrong (that is, where its dichotomized prediction differed from the dichotomized ground truth).

The thresholds and the final subset of ten tasks was chosen in three steps. First, candidate thresholds were chosen for each of the 53 predictions. Thresholds that seemed on a quick glance too obviously wrong or right were discarded (the threshold had to seem reachable); the final thresholds were chosen from these candidate thresholds so that they always seemed potentially reachable but the answer did not seem too obvious at a first glance. Secondly, a subset of 20 tasks were chosen based on the AutoML predictions: a mix of cases where the AutoML helped and where its prediction was incorrect. Thirdly, a final subset of ten tasks were chosen based on the explanations:

in six cases, all three explanation modes were deemed to work at least reasonably well (compared to the other ones, where one or more explanation technique was not that convincing), so these tasks were included. Four of the remaining 14 tasks were chosen randomly to get ten tasks. In the ten tasks chosen for the experiment, the AutoML predicted correctly in 60 % of the cases.

The order of these ten tasks was then randomized, except for the first and last task: the two tasks where all explanation methods seemed to be especially insightful were used as the first and last one. The one of these two tasks where the explanations methods seemed to fare better was chosen as the first task, and the other one as the last task. The order of the rest (in positions 2–9) was randomized.

This ordering of the ten tasks was kept constant for all participants, regardless of treatment group. The ordering was created to ensure that the participants would not get discouraged by poor explanations either as the first or last impression of the explanation methods, since this may create unnecessary bias against the explanations in lay persons or without more background information (of, for example, cases where the explanations may not provide good explanations).

We chose to only include ten tasks, to keep the workload of participants at a reasonable level, since the prediction tasks themselves could be challenging, and carefully reading the extensive instructions and background information was critical but could be time-consuming.

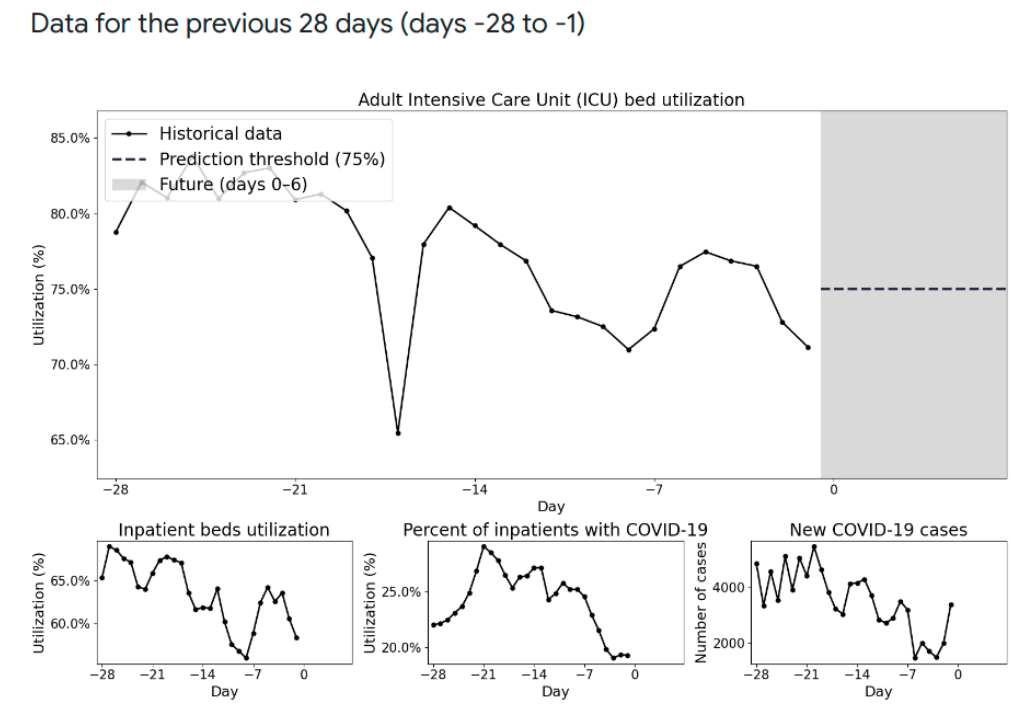


Figure 3.3: Example of initial prediction task, before seeing the prediction of the AutoML system, for the first task in the experiment.

3.2 Treatments: XAI techniques

Figure 3.4 shows an example prediction of the AutoML system used by the control group; that is, it gives no explanation for its prediction. Each of the treatment versions of the AutoML system employed a different XAI technique to explain its predictions. Therefore, in addition to seeing the prediction of the AutoML system (both visually in a figure with the data, and in writing), the participants in the treatment groups saw an explanation for the prediction of the system.

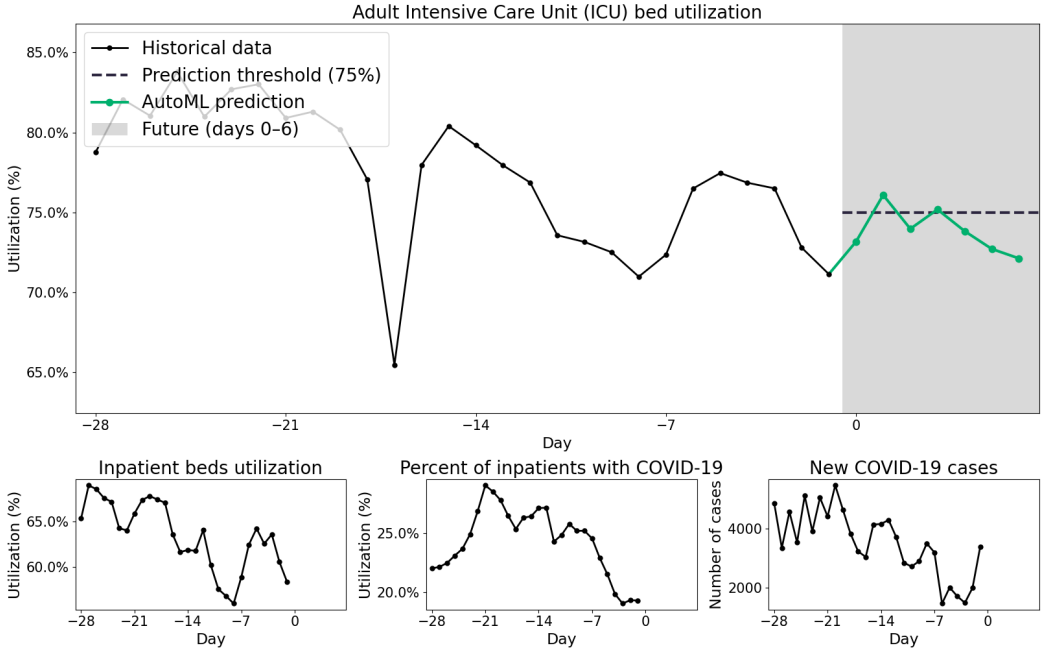


Figure 3.4: Example of AutoML prediction for control group (without explanation), for the first task in the experiment.

We chose and developed the employed XAI techniques with an audience without ML expertise in mind and based on the findings in Chapter 2 and especially Section 2.4.2, focusing on explanations that are *local*, *contrastive*, *simple* (selective and sufficient), and/or *non-probabilistic* (that is, they do not refer to probabilities or statistical relationships, but instead direct causes). To further support non-expert participants in correctly understanding the XAI techniques, treatment group participants were also provided with instructions describing the explanation technique employed by the AutoML system, including an example explanation, and general suggestions of how one might interpret both the given and general explanations and what they

may tell about the underlying model. The full explanation instructions can be found in Appendix A.

3.2.1 Evidence-based explanation

Evidence-based explanations—that is, in-sample predictions for the historical data shown and utilized for each experiment task—were used as a simple baseline XAI technique; Figure 3.5 shows an example of the evidence-based explanation. This explanation shows how the AutoML system would have predicted previously, for the earlier data: the system is made "blind" to the historical data of a task (the previous 28 days), and makes predictions for that time (in this case, in four increments of seven days).

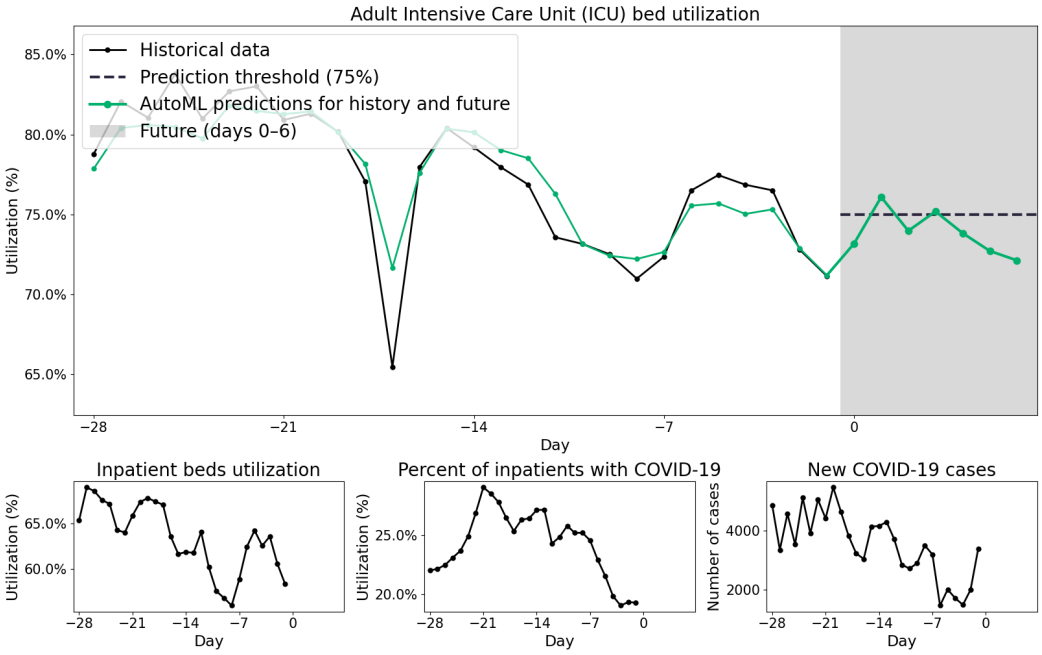


Figure 3.5: Example of evidence-based explanation, for the first task in the experiment.

This evidence of previous predictions allows one to judge how well the system predicts: if the predictions for the history do not make sense or seem incorrect, then the system may make future predictions that are as unreasonable. On the other hand, if the predictions for the history make sense and seem correct, the system may also make reasonable predictions for the future. Evidence-based explanations, or in-sample forecasts, are frequently provided by time series forecasting software and models (for example, Prophet:

`https://facebook.github.io/prophet/`); hence, it was included as a status quo alternative to more sophisticated XAI techniques.

This kind of explanation is local (focused on a specific prediction task and data), simple (small amount of extra information, focus on target variable), and non-probabilistic (actual predictions shown, not, for example, probabilities for different possible predictions).

3.2.2 SHAP

SHAP (Lundberg and Lee, 2017) is a popular XAI technique for interpreting the predictions of a (black-box) ML model. SHAP explains a prediction by assigning an importance value to each variable; in addition, these values add up to the predicted value explained.

The SHAP variable importance values are calculated relative to a chosen baseline value (for example, the mean of the variable). SHAP works by sampling random subsets of variables to "turn off"—that is, set to their baseline values—to see their effect on the prediction. The importance of a single variable is then the change in the prediction, compared to if the variable would have had the baseline value.

We interpreted every variable value as one feature for which to calculate the SHAP values. Mujkanovic et al. (2020) suggest using time slices instead of this kind of direct mapping of features: instead of interpreting every value in the time series as separate features, the time series are split into slices of a specified length. This is done partly to make the explanations more robust: Alvarez-Melis and Jaakkola (2018) have shown that explanation methods such as SHAP can be sensitive to small changes. On the other hand, using time slices lowers the resolution of the explanation. Since the data in each task of this experiment only consisted of 28 values per variable, we did not use time slices as features, but instead the individual variable values.

As suggested by Ozyegen et al. (2020), global variable means were used as the baseline values for replacement when calculating the SHAP values. The global means of each variable were calculated using all data except the last seven values (the unknown future) and for each state, to allow each explanation to be state-specific.

SHAP values were calculated separately for each of the seven predictions (days). Each of the seven predicted values of the AutoML is a distinct prediction, and thus seven SHAP values were calculated for each feature. To make the explanations simpler and more general, these SHAP values were

summed and divided by the largest absolute SHAP value to get overall relative SHAP feature importance scores. These relative SHAP values show the overall contribution of each feature to all seven predictions collectively and compared to the other features, and can be intuitively visualized; Figure 3.6 shows an example of the SHAP explanation.

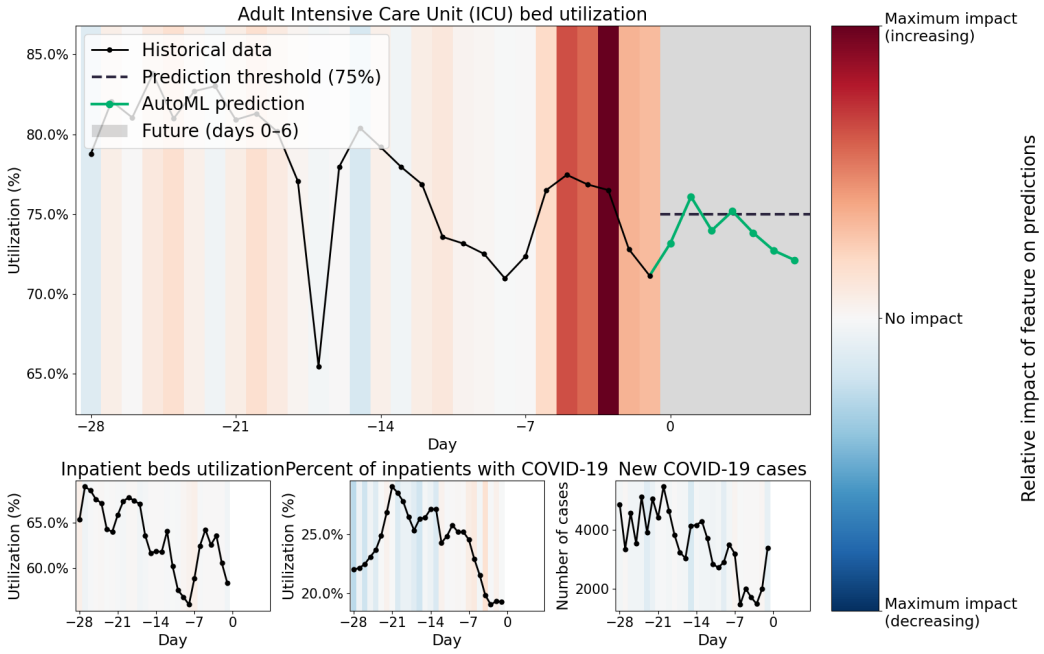


Figure 3.6: Example of SHAP explanation, for the first task in the experiment.

Explanations of this feature importance type show how important each feature of the data used for a specific prediction is: some values of features increase and some decrease the predicted value—SHAP shows both the magnitude and direction of these contributions. Thus, we get an explanation that is local (a specific prediction is explained), non-probabilistic, and potentially simple (depending on how many features are used for the explanation).

3.2.3 Counterfactual explanation

Counterfactual explanations explain through opposites: in order to explain why the system made a specific prediction, they show instead one example of how the data would need to change to make the AutoML system change its prediction; that is, they answer questions like "Why *this (unexpected) prediction* instead of *that (expected) prediction*?" and "What is the

smallest change to the data that is required for the system to make the opposite prediction (yes instead of no, no instead of yes)?"

Given these kinds of questions, counterfactual explanations are best suited to classification tasks; on the other hand, any continuous predicted variable can be discretized into one or more bins, representing classes, and counterfactual explanations can then be employed. In our case, we have a binary prediction: either at least one of the seven predicted values exceeds the given threshold, or none do. Thus, we can utilize counterfactual explanations to find the smallest necessary change to get the "opposite" prediction.

The Native-Guide method proposed by Delaney et al. (2020) was utilized in this experiment. Native-Guide finds a counterfactual—that is, a time series perturbed from the original data—that is as close as possible to the original data while still yielding the opposite prediction. It does this by first finding the closest neighbor in the (training) data (measured with some distance metric $d(T_i, T_j)$ for time series T_i and T_j) that would receive the opposite prediction, and then finding a linear combination of that and the original time series that is as close to the original time series as possible, while still yielding the opposite prediction. This enables finding counterfactuals that are as close as possible to the original time series, while still aiming to stay as close as possible to actual (and therefore more feasible) data instances.

Delaney et al. (2020) demonstrate the use of Native-Guide in univariate time series classification, but we generalized the concept to multivariate time series for this experiment by simply using a multivariate distance measure. We compared Independent Dynamic Time Warping (DTW) (Shokoohi-Yekta et al., 2017) and Euclidean distance measures, but given the time warping nature of DTW, Euclidean distance was chosen to favor ease of understanding for non-experts. Changes in all variables except the state ID were allowed, since changing the state of specific data would not be feasible (whereas changing the other variables could be).

Figure 3.7 shows an example of the counterfactual explanation. Counterfactual explanations fit all the criteria we set for the explanations in our experiment: they are local (focused on a specific prediction), contrastive (contrasts a prediction with an opposite one), simple (shows smallest necessary changes), and non-probabilistic.

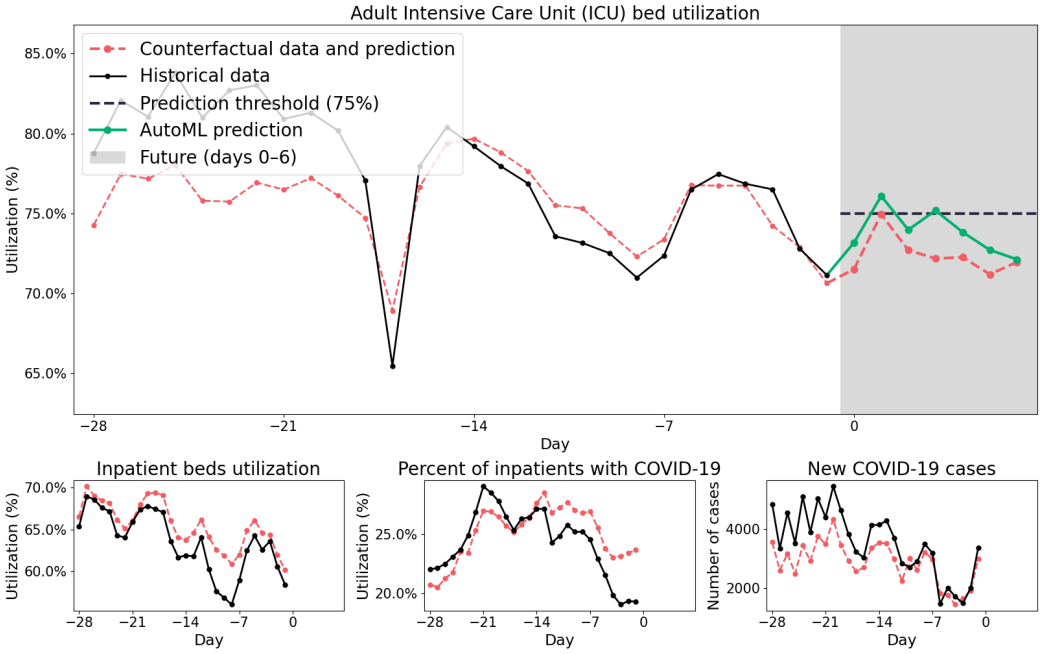


Figure 3.7: Example of counterfactual explanation, for the first task in the experiment.

3.3 Experiment measurements: trust metrics and control variables

Two different metrics were used to measure trust in the experiment: attitudinal trust and prediction switching. To measure **attitudinal trust** (trusting intentions and trusting beliefs), the 5-point Likert-style questionnaire of Merritt (2011) was employed. Merritt (2011) used and validated the scale in an experiment where participants rated screened airport luggage for weapons with an "Automatic Weapons Detection" (AWD) system: Cronbach's alphas for the scale ranged from $\alpha = 0.87$ to $\alpha = 0.92$. The scale was also utilized by Drozdal et al. (2020), and others have also used questionnaires to measure trust (for example, Dzindolet et al., 2003).

The questionnaire of Merritt (2011) was adapted to the context of this experiment by replacing "AWD" of the original scale with "AutoML system". The final scale used was:

1. I believe the AutoML system is a competent performer.
2. I trust the AutoML system.

3. I have confidence in the advice given by the AutoML system.
4. I can depend on the AutoML system.
5. I can rely on the AutoML system to behave in consistent ways.
6. I can rely on the AutoML system to do its best every time I take its advice.

The participants indicated their agreement with each statement on a 5-point scale:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

We also operationalized trust in the AutoML system as **prediction switching**, inspired by previous similar research (Merritt, 2011; Yin et al., 2019): that is, every time the initial prediction of the participant (made before seeing the prediction of the AutoML system) disagreed with the prediction of the AutoML and they switched their (final) prediction to agree with the AutoML, they are demonstrating trust (trusting behavior) towards the system. For each task where the initial prediction of a participant agrees with the prediction of the AutoML system, this prediction switching metric is undefined; thus, it is possible to get no measurement with this metric for some participants if they always happen to make the same initial prediction as the AutoML system.

Table 3.1 shows all the control variables and covariates that were measured in the experiment, together with their possible values. As suggested by Cameron and Stinson (2019), gender was measured with the open-ended question "I identify my gender as:" and recoded to the values

- 0 = Men
- 1 = Women
- 2 = Transgender and Non-Binary Individuals

using code adapted from Cameron and Stinson (2019).

Level of education, professional field, and job title all had an open-ended answer option. These were recoded to match the existing levels whenever possible. "Unemployed" and "No field" were added as values for professional field and "Unemployed" as a job title, to account for open-ended answers that fit best into these categories.

Overall confidence in one's ability to predict future developments based on data and previous experience with data analysis and/or predictions (yes/no) were included as control variables, since the confidence in one's own skills in a particular task can affect one's willingness to utilize automation (Lee and Moray, 1994; Madhavan and Wiegmann, 2007). Confidence in one's ability to predict was measured with a 11-point Likert-style item, ranging from "Not at all confident" (0) to "Very confident" (10).

Previous experience of using AutoML (yes/no) was included as a control variable since familiarity can affect the perceived expertise or credibility of a source of information (Madhavan and Wiegmann, 2007).

Propensity to trust affects the disposition of people to trust general others (McKnight and Chervany, 2001); this also holds for propensity to trust machines and actually trusting machines (Merritt, 2011). Therefore, we included propensity to trust (PTT) "automated agents" as a control variable, measured with the Likert-style scale suggested by Jessup et al. (2019), with the following six questions:

1. Generally, I trust automated agents.
2. Automated agents help me solve many problems.
3. I think it's a good idea to rely on automated agents for help.
4. I don't trust the information I get from automated agents.
5. Automated agents are reliable.
6. I rely on automated agents.

The above questions were prefaced with the following general description of an automated agent, adapted from Jessup et al. (2019): "An automated agent runs by computerized algorithms and interacts with humans. For example, a website predicting a medical diagnosis based on symptoms is an automated agent; likewise, an ATM is an automated agent."

Table 3.1: Control variables and covariates measured in the experiment, with their possible values indicated together with the compact variable names used in the analysis.

Control/covariate	Variable name	Values
Age	age	Positive integer
Gender	gender	Open-ended, recoded to four values
Level of education	education	Four values
Professional field	professional_field	18 values
Job title/position in job hierarchy	job_title	Seven values
Job experience	job_experience	Positive integer
Overall confidence in ability to predict based on data	general_confidence	11-point Likert-style item (range: 0–10)
Previous experience with data analysis and/or predictions	previous_data	Binary
Previous experience of using AutoML	previous_automl	Binary
Propensity to trust (Jessup et al., 2019)	ptt	5-point Likert-style scale, six questions (range: 1–5)

3.4 Participants: sample size, recruiting, and incentives

We calculated approximate statistical power before starting the experiment, to enable us to choose an appropriate sample size. The expected changes in trust were based on previous research. Merritt (2011) report changes of 0.35–0.42 times the standard deviation (SD) in the means of trust measurements, utilizing a within-subjects experiment design; on the other hand, Drozdal et al. (2020) report changes of over 1 SD in mean trust, measured on the same scale of Merritt (2011) and utilizing a within-subject design.

Based on this, we started with an assumption of a change of approximately 0.5 SD in the trust between the groups. To allow for all six pairwise comparisons between the groups, we used the p-adjusted formula of Chow et al. (2008, p. 71) to calculate the sample size. With type I error $\alpha = 0.05$ and statistical power 90 % (that is, type II error $\beta = 0.1$), we get a sample size $n_{i,j} \approx 123$ for any pairwise comparison of groups i and j; thus, each group k must have approximately $n_k \approx 61$, for a total of $N \approx 244$.

Initially, we sought a survey population of healthcare-related administration and management professionals, to have participants who were at least somewhat familiar with the experiment context and task. Recruiting through a LinkedIn (`https://www.linkedin.com/`) campaign proved inefficient, and therefore MTurk (`https://www.mturk.com/`) was instead used for recruiting participants.

To further explore whether our target population was feasible from a recruiting perspective, we conducted a pilot experiment on MTurk, with a base pay of $3 for completing the survey and a maximum of $3 bonus for correct predictions in the survey. This pilot showed that recruiting MTurk participants (Workers) matching the preferred profile (healthcare administration) was inefficient through MTurk too (no participants completed the survey within 5 days of publishing), whereas with less stringent requirements on field and job tasks five participants completed the survey in under four hours. Thus, the population was expanded to include any Workers that had the following qualifications:

- Masters qualification: Workers who have "consistently demonstrated a high degree of success in performing a wide range of [tasks] across a large number of [work requesters]" (`https://www.mturk.com/worker/help`)
- approval rate of previous assignments—that is, Human Intelligence

Tasks (HITs)—at least 95 %

- completed at least 50 MTurk HITs previously

The monetary incentive for participants was set at a base pay of $4 for completing the survey and a maximum of $3 bonus for correct predictions in the survey ($0.15 for each correct initial and final prediction). In addition, final adjustments to the instructions and visual layout of the experiment were made based on the feedback of the pilot participants. The pilot experiment data was discarded.

The participants were then recruited in batches and randomized to the treatment groups, dynamically adjusting the probability of being assigned to a group in order to make the group sizes as even as possible. In the end, a total of 260 Workers participated in the survey. Two of these were dropped from the final data, since they submitted the form multiple times with differing demographic information each time, suggesting foul play and making it impossible to determine their true answers. Two other participants also made multiple submissions, but with identical answers on demographic questions; therefore, their first submissions were deemed truthful, and were included. Thus, the final sample size was $N = 258$, distributed in the groups as shown in Table 3.2.

Table 3.2: Sample sizes in each treatment group and in total.

Treatment	**Size**
Control	64
SHAP	64
Evidence-based	66
Counterfactual	64
Total	**258**

The average score of the participants was approximately 12 correct predictions out of 20 (SD 2.4), yielding an average bonus of $1.8 per participant.

3.5 Experiment survey implementation

The survey was implemented in Google Forms (`https://www.google.com/forms/about/`). Each group had its own survey, identical except for explanations

(the explanations themselves for each AutoML prediction, in addition to mentions and descriptions of them); see Appendix A for the full forms.

In Google Forms, it is not possible to block anyone filling in a survey from going backwards in the survey; to avoid participants backing and changing their initial predictions after seeing the final prediction of the AutoML system (and thus invalidating the economic incentive and contaminating some of the data), we included warnings that doing so would result in an invalid submission, and forbade participants from going backwards in the survey at any point for any reason. To further discourage cheating by going backwards, participants were required to check a box saying they understand that they are not allowed to go backwards. A link to a PDF with all the instructions and background information presented in the survey was included to allow participants to view the instructions without going backwards in the survey.

Chapter 4

Results and discussion

4.1 Experiment results

The analysis of the experiment results is divided into two parts, one for each trust metric (attitudinal trust and prediction switching). For attitudinal trust, an analysis of covariance (ANCOVA) model was fit, and for prediction switching, logistic regression was used by fitting a generalized linear mixed-effects model (GLMM) with logit link function. In both cases, all control variables (see Table 3.1) were controlled for by including them as fixed effects. Effect sizes are reported as partial η^2 for the ANCOVA model; partial η^2 estimates the share of unexplained variance that can be explained by a predictor (after controlling for all other variables in the model) (Levine and Hullett, 2002). For the logistic regression, odds ratios of single model coefficients are discussed as a form of effect size. The analysis was performed using `R` (R Core Team, 2021).

We used parametric tests, even though some of the variables may not at first glance seem to fulfill the assumptions required for these tests. For example, Likert-style *items* (single Likert-style questions) are technically ordinal scales; on the other hand, 11 (and more) levels has been recommended as enough to allow treating it as an interval scale (Leung, 2011; Nunnally and Bernstein, 1994; Wu and Leung, 2017). Moreover, Likert-style *scales* (that is, sums or averages of multiple Likert-style items), exhibit linear properties, and are thus most often treated as interval scales (Streiner et al., 2015). Hence, the variables we measured with Likert items (overall confidence in one's ability to predict) and on Likert scales (PTT and attitudinal trust) were treated as continuous variables on interval scales, which enabled us to use parametric tests such as ANCOVA.

4.1.1 Descriptive statistics

Table 4.1 shows the means (with corresponding standard deviations) and adjusted means (with corresponding confidence intervals) for attitudinal trust in each treatment group, in addition to the total number of switches and non-switches in each group, and their sum—that is, the total number of potential switches (in other words, the number of times the initial prediction of a participant and the prediction of the AutoML system were in conflict). Prediction switches are also visualized in Figure 4.1. Table 4.2 shows descriptive statistics for the control variables.

Table 4.1: Attitudinal trust scale means, standard deviations, and adjusted means with corresponding 95 % confidence intervals, in addition to prediction switch counts for each treatment (with switches demonstrating agreement with the AutoML system and non-switches demonstrating disagreement).

	Attitudinal trust				Prediction switching		
Treatment	M	SD	M_{adj}	95 % CI	Switches	Non-switches	Total
Control	4.07	0.62	4.00	[3.87, 4.12]	144 (62 %)	88 (38 %)	232
SHAP	3.86	0.63	3.93	[3.80, 4.06]	123 (50 %)	121 (50 %)	244
Evidence-based	4.06	0.66	4.05	[3.93, 4.18]	128 (60 %)	87 (40 %)	215
Counterfactual	4.18	0.48	4.19	[4.06, 4.32]	130 (57 %)	98 (43 %)	228

4.1.2 The effect of XAI on attitudinal trust

There was a statistically significant difference in the means of attitudinal trust between the groups, controlling for all measured control variables (see Table 3.1), as determined by a one-way ANCOVA ($F(3, 222) = 2.899$, $p = 0.036$, partial $\eta_p^2 = 0.038$). Table 4.1 shows the adjusted means of attitudinal trust for each group: the system with counterfactual explanations was trusted most and evidence-based explanations second most, with control coming in third place and SHAP trusted the least.

Nevertheless, post hoc tests—simultaneous pairwise comparisons with Tukey's honestly significant difference (HSD)—indicated that only the means of SHAP and counterfactual had a statistically significant difference ($t(222) = -2.823$, $p = 0.027$), whereas the other differences were not statistically significant; Table 4.3 shows the full results of the post hoc test. The post hoc test was performed employing the `emmeans` package (Lenth, 2021).

Four control variables had a statistically significant effect on attitudinal trust:

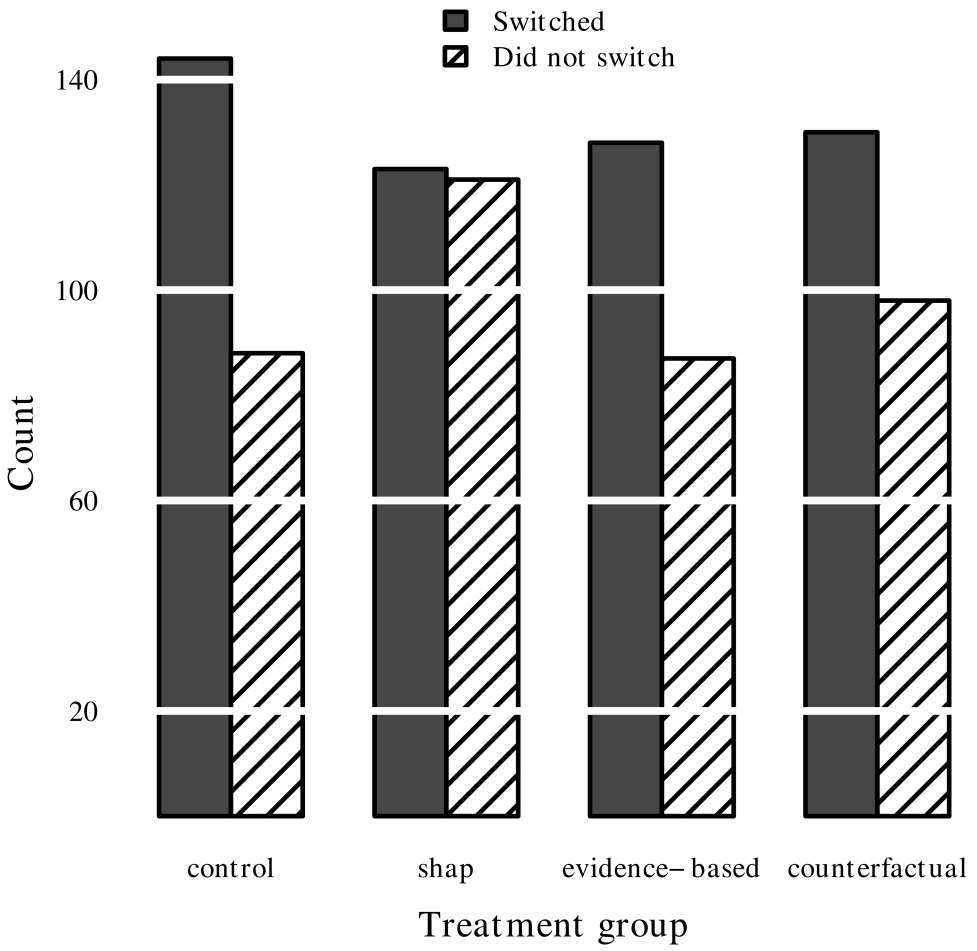


Figure 4.1: Number of switches and non-switches for each treatment group.

level of education ($F(3, 222) = 3.082$, $p = 0.028$, partial $\eta_p^2 = 0.040$), professional field ($F(17, 222) = 1.806$, $p = 0.028$, partial $\eta_p^2 = 0.125$), previous experience with data analysis and/or predictions ($F(1, 222) = 4.725$, $p = 0.031$, partial $\eta_p^2 = 0.021$), and PTT ($F(1, 222) = 82.258$, $p < 0.001$, partial $\eta_p^2 = 0.270$). Table 4.4 shows the full ANCOVA results, using Type III sums of square with orthogonal contrasts (control against one treatment at a time) defined; the full model coefficients are shown in Table 4.5. The ANCOVA model was fit with `R` (R Core Team, 2021), the significance tested using `Anova` in the `car` package (Fox and Weisberg, 2019), and the adjusted means calculated with the `effects` package (Fox and Weisberg, 2019).

4.1.3 The effect of XAI on prediction switching

The prediction switches are binary, and each participant could have 0–10 chances to switch (one switch/non-switch for each time their first prediction differed from the one made by the AutoML system). Thus, observations were not independent, and a simple logistic regression could not be employed.

Instead, a GLMM with a logit link function was fit, with participants entering the model as a random effect. The treatment group (independent variable) and all control variables (see Table 3.1) were included as fixed effects. All continuous variables were scaled and centered for a more stable regression. The GLMM was fit using the `lme4` package (Bates et al., 2015); the analysis of variance of the model, utilizing type III Wald χ^2 tests, was performed with the `car` package (Fox and Weisberg, 2019).

The analysis of variance based on the GLMM, controlling for all control variables, found no statistically significant effect of treatment group on prediction switching ($\chi^2(3) = 6.036$, $p = 0.110$). On the other hand, four covariates had a statistically significant effect on prediction switching: PTT ($\chi^2(1) = 19.237$, $p < 0.001$), overall confidence in ability to predict based on data ($\chi^2(1) = 21.067$, $p < 0.001$), previous experience with AutoML ($\chi^2(1) = 21.067$, $p < 0.001$), and professional field ($\chi^2(17) = 28.383$, $p = 0.041$). The full results of the analysis of variance on the GLMM are shown in Table 4.6.

Table 4.7 shows the full GLMM model coefficients, their standard errors and significance (using the z values and Wald tests), along with the corresponding odds ratios (e^{β}). The coefficients are given in standardized units for continuous variables, since the variables were scaled and centered before fitting the model. Thus, an odds ratio of x for a continuous variable means that every increase of one standard deviation in that variable makes it x times more likely that the person will switch their prediction. For binary variables, odds ratios simply mean that having the characteristic (for example, having previous experience with AutoML) makes a person x times more likely to switch their prediction.

Five control variable coefficients are significant: PTT (Wald $z = 4.386$, $p < 0.001$), confidence in one's ability to predict based on data (Wald $z = -4.590$, $p < 0.001$), previous experience with AutoML (Wald $z = 2.121$, $p = 0.034$), "Transportation, Distribution, and Logistics" (professional field; Wald $z = -2.073$, $p = 0.038$), and "Doctor's or equivalent" (level of education; Wald $z = 2.121$, $p = 0.034$).

The odds ratios also indicate the direction of the effect of XAI on behavioral trust: all treatment group odds ratios are less than 1 (all $\beta < 0$), meaning

participants using the treatment systems were less likely to switch their prediction than those using the control system (without explanations). However, since the analysis of variance on the model did not indicate a statistically significant effect of the treatment group, no strong conclusions should be drawn about these odds ratios: even though they may be statistically significant when tested separately, as in Table 4.7, this includes multiple comparisons, and thus inflates the type I error rate unless corrections are made to mitigate this.

Table 4.2: Descriptive statistics for control variables (counts and corresponding percentages) and covariates (means and standard deviations).

Variable	M	SD	N	%
age	40.54	9.84		
job_experience	16.67	9.83		
general_confidence	5.9	2.3		
ptt	3.58	0.73		
gender_recoded				
Man			145	56
Woman			111	43
Transgender/non-binary			2	1
education				
Upper secondary			78	30
Bachelor's or equivalent			146	57
Master's or equivalent			27	10
Doctoral or equivalent			7	3
professional_field				
Agriculture, Food and Natural Resources			3	1
Architecture and Construction			8	3
Arts, Audio/Video Technology and Communications			14	5
Business Management and Administration			21	8
Education and Training			15	6
Finance			16	6
Government and Public Administration			8	3
Health and Medicine			7	3
Hospitality and Tourism			8	3
Human Services			4	2
Information Technology			68	26
Law, Public Safety, Corrections and Security			4	2
Manufacturing			20	8
Marketing, Sales, and Service			25	10
Science, Technology, Engineering, and Mathematics			14	5
Transportation, Distribution, and Logistics			6	2
No field			9	3
Unemployed			8	3
job_title				
CEO, CxO			7	3
Manager			75	29
Individual Contributor			114	44
Entry-level			52	20
Unemployed			10	4
previous_data (yes)			101	39
previous_automl (yes)			42	16

Table 4.3: Tukey's HSD post hoc test results for attitudinal trust.

Comparison	Estimate	SE	df	t	p
Control − SHAP	0.070	0.094	222	0.749	0.877
Control − Evidence-based	−0.057	0.091	222	−0.625	0.924
Control − Counterfactual	−0.194	0.092	222	−2.103	0.155
SHAP − Evidence-based	−0.127	0.093	222	−1.367	0.521
SHAP − Counterfactual	−0.264	0.094	222	−2.823	0.026*
Evidence-based − Counterfactual	−0.137	0.093	222	−1.479	0.452

* $p < 0.05$

Table 4.4: ANCOVA results for attitudinal trust.

Variable	Sum Sq	df	F	Pr(>F)	Partial η^2
Intercept	5.016	1	20.341	< 0.001***	0.084
gender	0.245	2	0.496	0.609	0.004
education	2.280	3	3.082	0.028*	0.040
professional_field	7.573	17	1.806	0.029*	0.122
job_title	0.691	4	0.701	0.592	0.012
previous_data	1.165	1	4.725	0.031*	0.021
previous_automl	0.062	1	0.251	0.617	0.001
age	0.443	1	1.798	0.181	0.008
job_experience	0.462	1	1.875	0.172	0.008
general_confidence	0.087	1	0.354	0.553	0.002
ptt	20.285	1	82.258	< 0.001***	0.270
Treatment group	2.144	3	2.899	0.036*	0.038
Residuals	54.746	222			

* $p < 0.05$ *** $p < 0.001$

Table 4.5: Full ANCOVA model coefficients for attitudinal trust.

Variable/factor level	β
Intercept	2.101
gender (baseline: Man)	
Woman	0.003
Non-binary	0.363
education (baseline: Upper secondary)	
Bachelor's or equivalent	−0.160
Master's or equivalent	−0.128
Doctoral or equivalent	0.450
professional_field (baseline: Agriculture, Food and Natural Resources)	
Architecture and Construction	0.006
Arts, Audio/Video Technology and Communications	0.003
Business Management and Administration	−0.352
Education and Training	−0.137
Finance	−0.142
Government and Public Administration	0.408
Health and Medicine	−0.272
Hospitality and Tourism	0.101
Human Services	0.226
Information Technology	−0.035
Law, Public Safety, Corrections and Security	−0.201
Manufacturing	0.284
Marketing, Sales, and Service	−0.103
Science, Technology, Engineering, and Mathematics	0.002
Transportation, Distribution, and Logistics	−0.396
No field	−0.223
Unemployed	−0.375
job_title (baseline: CEO, CxO)	
Manager	0.315
Individual Contributor	0.338
Entry-level	0.302
Unemployed	0.377
previous_data	−0.179
previous_automl	0.052
age	0.009
job_experience	−0.009
general_confidence	0.011
ptt	0.436
Treatment group (baseline: control)	
SHAP	−0.115
Evidence-based	0.012
Counterfactual	0.149

Note. The baseline values for binary variables are false (No); baselines for categorical variables (factors) are indicated separately. Variables were not scaled, so the coefficients are in variable units.

Table 4.6: GLMM analysis of variance (using type III Wald χ^2 tests) results for prediction switching.

	χ^2	df	$\Pr(>\chi^2)$
Intercept	0.658	1	0.417
gender	2.073	2	0.355
education	5.263	3	0.153
professional_field	28.383	17	0.041^{*}
job_title	3.313	4	0.507
previous_data	0.294	1	0.588
previous_automl	11.871	1	$< 0.001^{***}$
age	0.196	1	0.658
job_experience	1.634	1	0.201
general_confidence	21.066	1	$< 0.001^{***}$
ptt	19.237	1	$< 0.001^{***}$
Treatment group	6.036	3	0.110

$^{*}\, p < 0.05 \quad ^{***}\, p < 0.001$

Table 4.7: GLMM results for prediction switching.

Variable/factor level	β	SE	z	Pr(>|z|)	Odds ratio
Intercept					
Fixed effect	0.966	1.191	0.811	0.417	
Random effect (subject)	(SD) 0.832				
gender (baseline: Man)					
Woman	0.180	0.245	0.735	0.463	1.197
Non-binary	1.256	1.004	1.251	0.211	3.512
education (baseline: Upper secondary)					
Bachelor's or equivalent	−0.005	0.262	−0.019	0.985	0.995
Master's or equivalent	−0.160	0.405	−0.394	0.693	0.852
Doctoral or equivalent	2.044	0.964	2.121	0.034*	7.718
professional_field (baseline: Agriculture, Food and Natural Resources)					
Architecture and Construction	0.014	1.198	0.012	0.990	1.014
Arts, Audio/Video Technology and Communications	0.575	1.074	0.535	0.592	1.777
Business Management and Administration	0.077	1.022	0.075	0.940	1.080
Education and Training	−0.249	1.077	−0.231	0.817	0.780
Finance	1.189	1.068	1.113	0.266	3.282
Government and Public Administration	0.072	1.121	0.064	0.949	1.074
Health and Medicine	−0.663	1.205	−0.550	0.582	0.515
Hospitality and Tourism	1.100	1.234	0.891	0.373	3.004
Human Services	−0.419	1.381	−0.303	0.762	0.658
Information Technology	−0.112	0.981	−0.114	0.909	0.894
Law, Public Safety, Corrections and Security	−1.310	1.496	−0.875	0.381	0.270
Manufacturing	−0.183	1.030	−0.178	0.859	0.833
Marketing, Sales, and Service	0.083	1.030	0.080	0.936	1.086
Science, Technology, Engineering, and Mathematics	0.896	1.069	0.839	0.402	2.450
Transportation, Distribution, and Logistics	−2.605	1.257	−2.073	0.038*	0.074
No field	−0.782	1.092	−0.716	0.474	0.458
Unemployed	−1.323	1.865	−0.710	0.478	0.266
job_title (baseline: CEO, CxO)					
Manager	−0.168	0.689	−0.244	0.808	0.846
Individual Contributor	−0.395	0.680	−0.581	0.561	0.673
Entry-level	0.015	0.714	0.021	0.984	1.015
Unemployed	1.351	1.647	0.821	0.412	3.863
previous_data	0.144	0.266	0.542	0.588	1.155
previous_automl	−1.129	0.328	−3.445	< 0.001***	0.323
age	0.095	0.214	0.442	0.658	1.099
job_experience	0.262	0.205	1.278	0.201	1.299
general_confidence	−0.662	0.144	−4.590	< 0.001***	0.516
ptt	0.502	0.115	4.386	< 0.001***	1.653
Treatment group (baseline: control)					
SHAP	−0.644	0.297	−2.169	0.030*	0.525
Evidence-based	−0.137	0.302	−0.455	0.649	0.872
Counterfactual	−0.483	0.299	−1.613	0.107	0.617
Log likelihood	−519.8				
AIC	1113.617				

Note. Subjects enter the model as random effects (intercepts). The baseline values for binary variables are false (No); baselines for categorical variables (factors) are indicated separately. Continuous variables were scaled and centered, and thus the coefficients β are in standardized units.
* $p < 0.05$ *** $p < 0.001$

4.2 Discussion

The experiment results indicate that the employed XAI technique in a time series forecasting context has a statistically significant effect on user attitudinal trust. Comparing the techniques pairwise showed that counterfactual explanations were trusted most (adjusted mean 4.18; see Table 4.1), but that only the difference between counterfactuals and SHAP was statistically significant (0.26 difference in adjusted means; see Table 4.3 for pairwise comparisons). This seems to support the findings of Miller (2019), that explanations should be contrastive and selective (simple): counterfactual explanations explain through opposites ("why not this other possible prediction") and show only a direct cause (not countless, possibly almost irrelevant, contributing factors), as compared to SHAP, which in our implementation shows the contribution of every single variable value to the final prediction. Evidence-based explanations—arguably the simplest XAI technique employed—were also trusted more than SHAP (though the difference was not statistically significant). This further supports the idea that simple and selective explanations are trusted more, especially in a context with a large amount of variables and data, such as time series forecasting, where the amount of information in the prediction task itself is great and explanations further add to it.

However, the effect size of XAI on attitudinal trust was relatively small: smaller than that of many control variables, and only a fraction of the effect size (partial $\eta^2 = 0.36$) reported by Drozdal et al. (2020). However, the prediction context in the study by Drozdal et al. (2020) is not related to time series and is one with less variables and data (loan approval, that is, binary classification); they were thus able to construct simple and convincing explanations, and subjects may also have had more knowledge of the task and therefore a better ability to gauge the performance and trustworthiness of the AutoML systems. Some of the difference may also be explained by differences in experiment design: their experiment was a within-subjects one, possibly reducing unexplained variance significantly and thus making the effect of XAI more prominent; in addition, their subjects consisted only of students with ML experience, and the participants had no "skin in the game" (nothing to lose by trusting or not trusting the AutoML), possibly biasing their trust measurements to a more positive direction.

When measured by prediction switching, the effect of XAI on trust was not statistically significant. In addition, based on the GLMM and ANCOVA model coefficients, the effect on prediction switching was opposite to that on attitudinal trust for evidence-based and counterfactual explanations: the

participants were less likely to switch their predictions to align with the ones of the AutoML system, even though they trusted the systems more (except for SHAP, which was trusted the least on both metrics). The odds ratios (see Table 4.7) indicate that with SHAP, participants were 0.525 times as likely (that is, 47.5 % less likely) to switch than with the control system; for counterfactual this distrust was weaker (odds ratio 0.617), and weakest for evidence-based explanations (odds ratio 0.872). On the other hand, the only statistically significant GLMM coefficient is the one for SHAP, for which the effect on trust was in the same direction on both metrics, that is, SHAP was trusted less by both metrics.

Such a discrepancy, between the self-reported attitudinal trust of participants and their actual trusting behavior, is in agreement with what, among others, Glaeser et al. (2000) found: attitudinal survey questions about trust predict trustworthy behavior, not necessarily trusting behavior. Even Kraus (1990, p. 5), whose meta-analysis showed that attitudes significantly predict behavior, found that "corresponding levels of specificity" of attitude and behavior measures, in addition to "attitudes formed by direct experience" and "held with certainty", increase the correspondence between attitudes and behavior. In our case, the attitudinal trust was generally not formed by experience or held with certainty (since the system was new to all participants, and only 16 % of participants had previous experience of using AutoML). Moreover, the attitudinal trust scale asked questions about the specific AutoML system but about its reliability and behavior generally, whereas prediction switches were related to specific instances. In addition, the small number of possible tasks, and therefore switches, may not have allowed the prediction switching to have a corresponding level of specificity (generality), and this lack of strong correspondence may further explain the discrepancy in the results.

Another contributing fact to these differences between attitudes and behavior may be the relatively low resolution and lack of robustness of the prediction switching metric. Given only ten tasks, at most ten switching opportunities (conflicts) were possible for each participant (eight conflicts was the maximum)—this does not allow for many opportunities for demonstrating (dis)trust by switching, and even one switch will result in a large difference in number of switches and the share of switches compared to non-switches. Moreover, the prediction task was in practice binary; this made (dis)agreement binary too, instead of allowing for degrees of (dis)agreement. This means that it was quite possible for participants to make a prediction they were not initially confident about, and seeing the AutoML system make the same prediction made them more confident in their prediction, and thus seldom switching to disagree; indeed, the confidence of participants in their

final predictions was for all groups on average higher than their confidence in their initial prediction. In addition, for all groups, prediction switches in cases where initial predictions *agreed* with the ones of the AutoML, participants switched in only a handful of cases (ranging from 3.1 % to 7.7 %, compared to the shares of non-switches for conflicting initial predictions, 38 %–50 %, in Table 4.1).

Interestingly, by both trust metrics, SHAP was trusted the least—even less than the (black-box) control system—whereas evidence-based explanations and counterfactuals were both trusted more than the control AutoML system (though these differences to the control group were not statistically significant). This may be due to the information-heavy nature of the SHAP explanations—the explanations were included as a separate figure, with a color for each value of each variable (indicating the importance of the variable values)—this may simply have been too much to take in, leading to too high a cognitive load and lack of trust in the system due to an explanation that was too complicated. This is also supported by the fact that the other treatments—both arguably simpler and with less information displayed and requiring less mathematical thinking and processing—were trusted more. This suggests, as Miller (2019) found, that one should strive to keep explanations simple and selective.

Another important perspective to consider when interpreting the results is that of trust calibration: it is possible that an XAI technique may even (justifiably) decrease trust if it sheds light on the shortcomings of the underlying model. Usually, the expert developer of a model is convinced of its reliability, for example, based on its performance with validation data, and wants to convince users of its reliability. In such a case, the developer may provide explanations to try to increase the trust of the users (or even themselves) in the model, until it matches the actual reliability of the model; users may even require some kinds of explanations to be able to trust the model. Explanations can also promote continuous calibration of trust, based on the future performance of the model. Especially with AutoML systems, where model development is automated and non-experts may be creating models, meaningful explanations can be crucial in understanding the model and its strengths and weaknesses, to ensure appropriate trust in it.

Indeed, given that the accuracy of the system in our experiment was only 60 % (in the ten prediction tasks in the experiment), a participant should rely (completely) on the system only if they are confident that the system has a better accuracy (that is, if the participant assumes their own accuracy is 50 % or less—equal to or worse than guessing). Interestingly, the scores of

those who switched in less than 50 % of cases were higher (mean 13.07 out of 20, SD 3.24) than for those who often (more than 50 % of the time) switched (mean 11.51, SD 1.74), indicating that "distrusting" participants (those who did not switch as often) may have simply (justifiably) decided not to rely on the AutoML system since their own accuracy was higher.

Examining the results from the trust calibration perspective, SHAP was trusted the least, but it may simply be that it was the explanation best suited to calibrating trust in the AutoML system: the SHAP explanation gave the most extensive information about the predictions of the AutoML system, thus allowing participants to gauge its performance in detail. This may have led participants to form strong views of the accuracy and validity of the system, which may not have been possible with the other, less detailed explanations or the control system with no explanations. Thus, this calibration view might explain the lack of a strong effect of XAI techniques on trust: the explanations may have provided valuable information to users, based on which they adjusted their (attitudinal and behavioral) trust to match the actual reliability of the system; that is, in the terminology of Lee and See (2004), they were (more) appropriately using the machine.

It may also be possible that users simply almost blindly trust any kind of advice by a machine; in such a case, the effect of XAI may not have much of an effect, or may even be detrimental, since users may instead be exposed to features of the AutoML system that seem to them like errors or inconsistencies. This view would align with the findings of Madhavan and Wiegmann (2007), that humans often have a schema of machines being perfect, and schema violations, such as even small errors (especially in easier tasks), would be likely to be spotted and remembered, hence negatively impacting trust.

Most of the control variables were statistically significant in at least one of the analysis models, in line with expectations based on previous research; in fact, only the most general control variables—not directly related to automation, data analysis, or AutoML (age, gender, job title, job experience)—did not have a statistically significant effect on either model. In addition, the statistically significant control variables all had relatively large effects. For example, education, professional field, and PTT all had larger effect sizes than treatment group for attitudinal trust (see Table 4.4). For example, the ANCOVA model coefficients (see Table 4.5) show that an increase of one unit in PTT increased trust by 0.436; interestingly, having a level of education of "Doctoral or equivalent" increased trust with 0.450 compared to having "Upper secondary" education, whereas "Bachelor's or eq." or "Master's or eq." education decreased trust.

The odds ratios for prediction switching (see Table 4.7) indicate similar results for professional field, previous AutoML experience, confidence in predictions, and PTT, which have large effects (that is, changes in likelihood of a switch, compared to the baseline of that variable), at least for some levels (for non-binary variables). For instance, a higher PTT increased the likelihood of a switch: increasing PTT by one standard deviation increased made the participant 1.653 times as likely to switch. Conversely, just as Lee and Moray (1994) and Madhavan and Wiegmann (2007) found, a higher confidence in one's own ability to predict (general_confidence) decreased the likelihood of a switch (that is, relying on a machine)—every increase of one standard deviation in confidence in one's own ability to predict (general_confidence) made a participant 0.516 times as likely to switch. Interestingly, previous experience of using AutoML made a participant 0.323 times as likely (that is, 67.7 % less likely) to switch. Bad previous experiences with using AutoML may contribute to a distrust of AutoML in general, especially since automation errors are often well-remembered (Madhavan and Wiegmann, 2007).

4.2.1 Limitations and future work

Our ability to draw general conclusions is restricted by many limiting factors. Firstly, our experiment utilized AutoML in a time series forecasting context. It is possible that this context in itself is too complicated for a general population to grasp fully, and therefore participants may have been unable to truly gauge the trustworthiness of the AutoML systems. Completing the survey online, without any oversight by or possibility of getting help from the experiment organizer, may have led to misunderstandings, further impacting the trust of participants in the AutoML systems.

Related to this, our experiment gave participants relatively little information about the underlying model, owing to the lack of information about previous and current performance of the system. Participants had no information on the previous performance of the system (for example, its prediction accuracy). On the other hand, the accuracy of a regression model is affected by the arbitrary choice of an accuracy metric, which may not have provided lay persons any sensible information; giving a prediction accuracy in terms of the kinds of binary predictions utilized in the experiment, transformed from continuous predictions based on prediction thresholds, would naturally have required choosing a large number of arbitrary thresholds for different cases, again making the accuracy information arbitrary and possibly misleading.

Moreover, participants did not know whether the predictions given by the system for the experiment tasks were correct, until after the experiment; thus, they had to form their own views of the accuracy of the system based on how convincing and reasonable its predictions and explanations seemed. Leaving out the ground truth was nonetheless a design choice we had to make, due to technical restrictions of the Google Forms platform: we could not block participants from going backwards and changing their responses, and hence some participants could have changed their predictions after seeing the ground truth and thus contaminated the data.

The information participants could get of the system was also restricted by the low number of prediction tasks. By increasing the number of predictions, participants would have been able to see the model in action in more cases, possibly giving them a more detailed picture of its performance in a diverse set of situations.

As discussed previously, given more tasks, the prediction switching measure would also have been more robust. The binary nature of the predictions also affected the sensitivity of the prediction switching metric: no degree of agreement was possible, since (dis)agreement was binary. Including a metric of trust such as prediction switching was nevertheless important, since it can provide valuable information: if users only say they trust an AutoML system, but do not demonstrate trusting behavior by actually relying on it, the measured attitudinal trust may be misleading.

Another limiting factor is our sample of subjects, which may not be representative enough to generalize to a general population anywhere. Our subjects were drawn from the pool of available MTurk Workers, whose demographic distribution is heavily biased towards US and, especially at specific times of day, Indian Workers (Difallah et al., 2018). Simple issues such as time zones may have affected our sampling: depending on when the MTurk batches were published and how quickly they were completed, the Workers who were able to accept the tasks may have been limited to specific time zones. We did not ask the participants about their nationality or country of residence, so we could not control for nationality or culture possibly influencing trust in AutoML; previous research has indicated differences between cultures in interpersonal and human-automation trust (Hoff and Bashir, 2015). On the other hand, our sample is clearly more diverse and representative than the all-student samples commonly utilized in research, as demonstrated by Table 4.2.

The incentives of participants and the time invested by participants in the experiment may also have contributed to the relatively small effect of XAI

on trust: it may be that someone with more skin in the game (for instance, a real decision maker) would have more thoroughly considered the data, tasks, and the predictions and explanations of the AutoML system, and in this case the effect of XAI might have been more pronounced. A Worker, in a hurry to complete the survey, may not have carefully considered and utilized the information provided by the XAI techniques, but may instead only have been further burdened by having to also understand the explanations of the AutoML system, thus reducing the effect of XAI. Adjusting the incentives to focus more on correct predictions (even penalizing incorrect ones) or requiring participants to use a certain amount of time to complete the survey may have contributed to participants having a more thorough understanding of the AutoML model and its explanations.

The lack of oversight by and connection with the experiment organizer, inherent in work done by Workers, might have also led to some not doing their best, or even cheating (for example, by going backwards, regardless of being explicitly forbidden to do so). Nevertheless, only a handful of Workers clearly cheated (by submitting multiple surveys), and we are inclined to trust in the honesty of the participants.

Future work can mitigate the limitations outlined above

- **by utilizing AutoML in a simpler context**, to reduce the potentially confounding effect of an unfamiliar and complicated context and more accurately gauge the effect of XAI on trust;
- **by increasing the number of tasks in the experiment**, to give more participants more first-hand information on the AutoML system, and to make prediction switching a more robust metric;
- **by providing more information on the performance and accuracy of the AutoML system**, both in terms of previous accuracy and the correctness of the predictions it gives for the experiment tasks;
- **by adjusting incentives and time spent by participants**, to encourage and ensure thorough understanding of the AutoML model and the information provided by the XAI technique; and
- **by conducting the study with an experiment organizer present**, to be able to interact with the participants, clear up misconceptions, and oversee the process to avoid cheating.

Chapter 5

Conclusion

This thesis studied the feasibility of using XAI to affect user trust in AutoML. We conducted an RCT comparing the effect of three XAI techniques on user trust in an AutoML system, and our results showed a statistically significant difference in attitudinal trust but no statistically significant difference in behavioral trust (switching predictions to align with the AutoML). Attitudinal trust was highest for the AutoML system employing counterfactual explanations, and SHAP was trusted the least; indeed, the only statistically significant pairwise comparison was the difference between attitudinal trust in the system giving counterfactual explanations and SHAP. Nevertheless, affecting trust proved challenging: the effect size of treatment group on attitudinal trust was relatively small compared to previous research and the effects of control variables. Measured by behavioral trust, participants were significantly less likely to trust all the treatment systems, with SHAP again trusted the least; nevertheless, this effect was not statistically significant.

The results of our experiment indicate that, although XAI has shown promise as a tool for increasing the transparency of and trust in (Auto)ML models, especially in somewhat simpler prediction tasks, these findings may not generalize to contexts with more variables and data, such as the multivariate time series forecasting context of our experiment. Such a context is a challenge for humans and models alike, and understanding it may be challenging for lay persons; the cognitive load of a AutoML user can be further exacerbated by (possibly complicated) explanations, negatively affecting trust in the system—this might explain the observed difference in trust between the relatively simple counterfactual explanations and SHAP, which provided users with extensive information. Future work could aim to gauge the effect of the amount of information, by examining the effects of XAI in simpler prediction contexts or with similar explanations, with only the amount of

information varied.

Nevertheless, our results show some differences in user trust, and given the promising results of previous research on employing XAI techniques to increase the transparency of ML and AutoML systems, future work could further develop these ideas in the AutoML context. Using our experiment as a starting point, one could form a view of how to increase or decrease trust, by developing new XAI techniques and observing how they affect user trust. The art of explanation is challenging, yet a crucial skill: for example, meaningful and effective explanations are important when trying to persuade someone to adopt a new method over another, especially in complex tasks with high stakes, however convinced one may be of the superiority of that method. Explanations that are too detailed and complicated or too vague or general could even be harmful and lower trust.

Thus, an important question for future research is how much transparency and information is optimal. Users should be provided enough and sufficient information on an AutoML system, to enable them to gain an initial understanding of it. Further information is required to allow users to continuously calibrate their trust in the system to a justifiable level that matches the actual reliability demonstrated by the system in future tasks; this can ensure appropriate and continued reliance on the system. Users should, however, not be overwhelmed by the amount of information provided, possibly lowering their trust unjustifiably.

After all, calibrating user trust should be one of our most important missions as designers and developers of AutoML systems and algorithms in general: in a world that is increasingly governed by data and models, people should be able to trust the decisions made based on the predictions of these models, and that means we must not strive for blind trust in algorithms—only calibrated trust and appropriate reliance.

Bibliography

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adnan, N., Md Nordin, S., bin Bahruddin, M. A., & Ali, M. (2018). How trust can drive forward the user acceptance to the technology? in-vehicle technology for autonomous vehicle. *Transportation Research Part A: Policy and Practice*, *118*, 819–836. https://doi.org/10.1016/j.tra.2018.10.019

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. In B. Kim, K. R. Varshney, & A. Weller (Eds.), *Proceedings of the 2018 icml workshop on human interpretability in machine learning (whi 2018)* (pp. 66–71).

Arabi, Y., Venkatesh, S., Haddad, S., Al-Shimemeri, A., & Al-Malik, S. (2002). A prospective study of prolonged stay in the intensive care unit: Predictors and impact on resource utilization. *International Journal for Quality in Health care*, *14*(5), 403–410.

Ates, E., Aksar, B., Leung, V. J., & Coskun, A. K. (2020). Counterfactual explanations for machine learning on multivariate time series data. arXiv preprint arXiv:2008.10781[cs.LG].

Baier, A. (1986). Trust and antitrust. *Ethics*, *96*(2), 231–260. https://doi.org/10.1086/292745

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Cameron, J. J., & Stinson, D. A. (2019). Gender (mis) measurement: Guidelines for respecting gender diversity in psychological research. *Social and Personality Psychology Compass*, *13*(11), e12506. https://doi.org/10.1111/spc3.12506

Chandra, S., Srivastava, S. C., & Theng, Y.-L. (2010). Evaluating the role of trust in consumer adoption of mobile payment systems: An empirical

analysis. *Communications of the Association for Information Systems*, *27*, 561–588.

Chow, S.-C., Shao, J., & Wang, H. (2008). *Sample size calculations in clinical research* (2nd ed.). Chapman & Hall/CRC.

Coelho, I. M., Coelho, V. N., da S. Luz, E. J., Ochi, L. S., Guimarães, F. G., & Rios, E. (2017). A GPU deep learning metaheuristic based model for time series forecasting. *Applied Energy*, *201*, 412–418. https://doi.org/10.1016/j.apenergy.2017.01.003

Delaney, E., Greene, D., & Keane, M. T. (2020). Instance-based counterfactual explanations for time series classification. arXiv preprint arXiv:2009.13211[cs.LG].

Deley, T., & Dubois, E. (2020). Assessing trust versus reliance for technology platforms by systematic literature review. *Social Media + Society*, *6*(2), 1–8. https://doi.org/10.1177/2056305120913883

Difallah, D., Filatova, E., & Ipeirotis, P. (2018). Demographics and dynamics of mechanical turk workers. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 135–143. https://doi.org/10.1145/3159652.3159661

Drozdal, J., Weisz, J., Wang, D., Dass, G., Yao, B., Zhao, C., Muller, M., Ju, L., & Su, H. (2020). Trust in automl: Exploring information needs for establishing trust in automated machine learning systems. *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 297–307.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International journal of human-computer studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Elshawi, R., & Sakr, S. (2020). Automated machine learning: Techniques and frameworks. In R.-D. Kutsche & E. Zimányi (Eds.), *Big data management and analytics* (pp. 40–69). Springer International Publishing.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). Autogluon-tabular: Robust and accurate automl for structured data. arXiv preprint arXiv:2003.06505[stat.ML].

Evans, A. M., & Krueger, J. I. (2009). The psychology (and economics) of trust. *Social and Personality Psychology Compass*, *3*(6), 1003–1017. https://doi.org/10.1111/j.1751-9004.2009.00232.x

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems 28*

(pp. 2962–2970). Curran Associates, Inc. http://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). SAGE Publications. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, *29*(5), 1189–1232. http://www.jstor.org/stable/2699986

Fu, W., Olson, R., Nathan, Jena, G., PGijsbers, Augspurger, T., Romano, J., Saha, P., Shah, S., Raschka, S., sohnam, DanKoretsky, kadarakos, Jaimecclin, bartdp1, Bradway, G., Ortiz, J., Smit, J. J., Menke, J.-H., ... Carnevale, R. (2020). *Epistasislab/tpot: V0.11.5* (Version v0.11.5). Zenodo. https://doi.org/10.5281/zenodo.3872281

Gee, A. H., Garcia-Olano, D., Ghosh, J., & Paydarfar, D. (2019). Explaining deep classification of time-series data with learned prototypes. arXiv preprint arXiv:2008.10781[cs.LG].

Giddens, A. (2013). *The consequences of modernity*. Stanford University Press. https://books.google.fi/books?id=SVmkJEwWGwAC

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring trust. *The Quarterly Journal of Economics*, *115*(3), 811–846. https://doi.org/10.1162/003355300554926

Guillemé, M., Masson, V., Rozé, L., & Termier, A. (2019). Agnostic local explanation for time series classification. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 432–439. https://doi.org/10.1109/ICTAI.2019.00067

Gulati, S., Sousa, S., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, *38*(10), 1004–1015. https://doi.org/10.1080/0144929X.2019.1656779

Guyon, I., Sun-Hosoya, L., Boullé, M., Escalante, H. J., Escalera, S., Liu, Z., Jajetic, D., Ray, B., Saeed, M., Sebag, M., Statnikov, A., Tu, W.-W., & Viegas, E. (2018). Analysis of the automl challenge series 2015-2018 (F. Hutter, L. Kotthoff, & J. Vanschoren, Eds.) [In press, available at http://automl.org/book.], 191–236.

He, X., Zhao, K., & Chu, X. (2021). Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, *212*, 106622. https://doi.org/10.1016/j.knosys.2020.106622

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hutter, F., Kotthoff, L., & Vanschoren, J. (Eds.). (2018). *Automated machine learning: Methods, systems, challenges* [In press, available at http://automl.org/book.]. Springer.

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In J. Y. Chen & G. Fragomeni (Eds.), *Virtual, augmented and mixed reality. applications and case studies.* (pp. 476–489). Springer International Publishing.

Jones, S. E. (2013). *Against technology: From the luddites to neo-luddism.* Routledge.

Jung, A. (2020). Machine learning: Basic principles. arXiv preprint arXiv:1805.05052[cs.LG].

Kraus, S. J. (1990). Attitudes and the prediction of behavior: A meta-analysis. *98th Annual Convention of the American Psychological Association, Boston, MA*.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153–184. https://doi.org/10.1006/ijhc.1994.1007

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lenth, R. V. (2021). *Emmeans: Estimated marginal means, aka least-squares means* [R package version 1.6.0]. https://CRAN.R-project.org/package=emmeans

Leung, S.-O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point likert scales. *Journal of Social Service Research*, *37*(4), 412–421. https://doi.org/10.1080/01488376.2011.580697

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, *28*(4), 612–625. https://doi.org/10.1111/j.1468-2958.2002.tb00828.x

Lewis, J. D., & Weigert, A. (1985). Trust as a social reality. *Social Forces*, *63*(4), 967–985. https://doi.org/10.1093/sf/63.4.967

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the ai: Informing design practices for explainable ai user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3313831.3376590

Luhmann, N., Davis, H., Raffan, J., Rooney, K., King, M., & Morgner, C. (2018). *Trust and power.* Wiley. https://books.google.fi/books?id=CKBRDwAAQBAJ

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. *Proceedings of 11th Australasian Conference on Information Systems*, *53*, 6–8.

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, *90*(10), 60–68.

McKnight, D. H., & Chervany, N. L. (2001). Trust and distrust definitions: One bite at a time. *Trust in cyber-societies* (pp. 27–54). Springer.

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors*, *53*(4), 356–370. https://doi.org/10.1177/0018720811411912

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, *50*(2), 194–210. https://doi.org/10.1518/001872008X288574

Miller, S., & Hughes, D. (2017). The quant crunch: How the demand for data science skills is disrupting the job market. *Burning Glass Technologies*. http://burning-glass.com/wp-content/uploads/The_Quant_Crunch.pdf

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

MindsDB Inc. (2019). *MindsDB*. https://github.com/mindsdb/mindsdb

Mokyr, J. (2005). Chapter 17 - long-term economic growth and the history of technology. In P. Aghion & S. N. Durlauf (Eds.). Elsevier. https://doi.org/10.1016/S1574-0684(05)01017-8

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617. https://doi.org/10.1145/3351095.3372850

Mujkanovic, F., Doskoč, V., Schirneck, M., Schäfer, P., & Friedrich, T. (2020). Timexplain – a framework for explaining the predictions of time series classifiers. arXiv preprint arXiv:2007.07606[cs.LG].

Nourani, M., Kabir, S., Mohseni, S., & Ragan, E. D. (2019). The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *7*(1), 97–105. https://ojs.aaai.org/index.php/HCOMP/article/view/5284

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Ozyegen, O., Ilic, I., & Cevik, M. (2020). Evaluation of local explanation methods for multivariate time series forecasting. arXiv preprint arXiv:2009.09092[cs.LG].

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Parmezan, A. R. S., Souza, V. M., & Batista, G. E. (2019). Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. *Information Sciences*, *484*, 302–337. https://doi.org/10.1016/j.ins.2019.01.076

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*(3), 279–301. https://doi.org/10.1016/0022-1031(77)90049-X

Rotter, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist*, *35*(1), 1–7.

Schaefer, K. E. (2016). Measuring trust in human robot interactions: Development of the "trust perception scale-hri". In R. Mittu, D. Sofge, A. Wagner, & W. F. Lawless (Eds.), *Robust intelligence and trust in autonomous systems* (pp. 191–218). Springer. https://doi.org/10.1007/978-1-4899-7668-0_10

Schelter, S., Biessmann, F., Januschowski, T., Salinas, D., Seufert, S., & Szarvas, G. (2018). On challenges in machine learning model manage-

ment. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, *41*, 5–15.

Shokoohi-Yekta, M., Hu, B., Jin, H., Wang, J., & Keogh, E. (2017). Generalizing DTW to the multi-dimensional case requires an adaptive approach. *Data Mining and Knowledge Discovery*, *31*(1), 1–31.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use* (5th ed.). Oxford University Press.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, *72*(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. *Proc. of KDD-2013*, 847–855.

United States Centers for Disease Control and Prevention. (2021a). *Covid-19 case surveillance public data access, summary, and limitations: United states covid-19 cases and deaths by state over time*. Retrieved January 18, 2021, from https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36/

United States Centers for Disease Control and Prevention. (2021b). *Management of patients with confirmed 2019-ncov*. Retrieved April 26, 2021, from https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html

United States Department of Health Human Services. (2021). *Covid-19 reported patient impact and hospital capacity by state timeseries* [Accessed 22 December 2020 and 26 January 2021 for the data with dates 7 September 2020–9 December 2020 and 26 July 2020–6 September 2020, respectively.]. https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/g62h-syeh

Van Slyke, C., Belanger, F., & Comunale, C. L. (2004). Factors influencing the adoption of web-based shopping: The impact of trust. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, *35*(2), 32–49. https://doi.org/10.1145/1007965.1007969

Waring, J., Lindvall, C., & Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine*, *104*, 101822. https://doi.org/10.1016/j.artmed.2020.101822

Wu, H., & Leung, S.-O. (2017). Can likert scales be treated as interval scales?—a simulation study. *Journal of Social Service Research*, *43*(4), 527–532. https://doi.org/10.1080/01488376.2017.1329775

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300509

Appendix A

Experiment surveys

The Google Forms surveys utilized in the experiment are included in the sections below as printed versions; the layout of these printed versions slightly differ from that of the online forms, and they display some answer formatting requirements, which are automatically enforced by the online forms and thus not shown by default. After completing the survey, the participants were shown a debriefing page, with the following text:

> Thank you for your time and for participating in the survey! The survey is now finished, and your answers have been submitted.
>
> The experiment's goal is to study whether and how explainability features (in this case, explaining the system's predictions in different ways)—or the lack of them—affect the users' trust in the system and its predictions. Trust is measured using the trust questionnaire you filled in after your predictions, and by the share of times you switched your predictions when your initial prediction differed from that made by the AutoML system.
>
> The experiment is done as part of the research for a Master's thesis, and the results will be published and available publicly as part of that thesis later this year.
>
> If you have any feedback or questions, you can contact the thesis author and experiment organizer Cosmo Jenytin (cosmo@mindsdb.com).
>
> To verify that you have completed the survey, fill in this code in your HIT on MTurk:
>
> mdb-experiment-group-X
>
> You can now view your results by clicking "View your score" below, and close this window.

The "X" in "mdb-experiment-group-X" was replaced with the number of the experiment group (1, 2.1, 2.2, or 2.3). From the feedback page, the participants could also click "View score" to review their answers and to see which were correct.

A.1 Survey, group 1: control

AutoML survey group 1

This study is done as part of a Master's Thesis, in collaboration with MindsDB (mindsdb.com), and the results and the thesis will be published later this year.

In this survey, you will use the predictions of an Automated Machine Learning (AutoML) system. AutoML systems automate one or more phases of the process of creating a Machine Learning (ML) model; this way, anyone can create a complex ML model without deep knowledge of ML itself. Machine Learning (ML) describes a type of computer program that learns in a way similar to human beings. Rather than employ statistical techniques or simple mathematics to analyze a large amount of data, the ML model learns by example and is able to digest a large amount of data, and be trained to give incrementally more and more accurate predictions if it is fed more and more data.

Your task in this survey will be to make predictions about adult Intensive Care Unit (ICU) bed utilization during the COVID-19 pandemic. Your goal is to estimate the percentage of adult ICU beds occupied, given a certain scenario. You will also be shown the predictions of an AutoML system to take into account.

Your goal should be to make as many correct predictions as possible.

The survey proceeds as follows.
1. Before the tasks begin, we will ask you to fill in some necessary information about yourself and give your informed consent about us gathering and storing your anonymized answers.
2. Then you will be shown a description of the data used and how you should make your predictions.
3. Then, you will make predictions in ten scenarios, with support from an AutoML system. After these predictions, you will fill in a questionnaire based on your experience with the AutoML system.

After this the study is done. You will see how your predictions compare to the actual adult ICU bed utilization. You will also have the opportunity to give feedback on the experiment and the AutoML systems used.

If you have any questions, please contact the survey organizer and thesis author Cosmo Jenytin (cosmo@mindsdb.com).

* Required

1. Informed consent: I consent to my answers being used and published anonymously. The background information asked here will be connected to your survey answers and anonymized. *

   Mark only one oval.

   ◯ Yes

Note: you are not allowed to go backwards in the survey for any reason. If you go back to a previous page, your submission will be invalid.
All the instructions can be found as a PDF here: https://drive.google.com/file/d/1nq-8rKE0FR_2vz-UNiMW1isANYcRXVSq/view?usp=sharing. Open this PDF in a separate window, so you have access to it throughout the experiment without going backwards in the form (which results in an invalid submission).

2. I understand that I am not allowed to go backwards (go back to previous pages) in the survey, and that doing so will result in my submission being invalid. *

   Mark only one oval.

   ◯ Yes

Background information

Please answer the following questions. Your answers will be used and stored anonymously.

3. Age *

   ____

4. I identify my gender as *

   ____

5. Education *

   Mark only one oval.

   ◯ Upper secondary
   ◯ Bachelor's or equivalent
   ◯ Master's or equivalent
   ◯ Doctoral or equivalent
   ◯ Other: __________

6. Professional field *

Mark only one oval.

- ◯ Agriculture, Food and Natural Resources
- ◯ Architecture and Construction
- ◯ Arts, Audio/Video Technology and Communications
- ◯ Business Management and Administration
- ◯ Education and Training
- ◯ Finance
- ◯ Government and Public Administration
- ◯ Health and Medicine
- ◯ Hospitality and Tourism
- ◯ Human Services
- ◯ Information Technology
- ◯ Law, Public Safety, Corrections and Security
- ◯ Manufacturing
- ◯ Marketing, Sales, and Service
- ◯ Science, Technology, Engineering, and Mathematics
- ◯ Transportation, Distribution, and Logistics
- ◯ Other: _____

7. Job title *

Mark only one oval.

- ◯ Chairperson, member of Board of Directors
- ◯ CEO, CxO
- ◯ Vice President
- ◯ Manager
- ◯ Individual Contributor
- ◯ Entry-level
- ◯ Other: _____

8. Job experience (years) *

_____

9. How confident are you, in general, in your ability to predict future developments based on data? *

NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

10. Have you previously worked with data analysis or predictions? *

This can, for example, mean working with simple data analysis tools (e.g., Excel) or more advanced tools, such as Machine Learning (ML) models.

Mark only one oval.

- ◯ Yes
- ◯ No

11. Have you previously used an Automated Machine Learning (AutoML) system? *

An AutoML system is a Machine Learning system that can automate one or more phases of a data science or Machine Learning model development process.

Mark only one oval.

- ◯ Yes
- ◯ No

12. Please answer the following questions on automated agents. *

An automated agent runs by computerized algorithms and interacts with humans. For example, a website predicting a medical diagnosis based on symptoms is an automated agent; likewise, an ATM is an automated agent. NOTE: there are five levels to choose from, please scroll if you cannot see all alternatives.

Mark only one oval per row.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Generally, I trust automated agents.	◯	◯	◯	◯	◯
Automated agents help me solve many problems.	◯	◯	◯	◯	◯
I think it's a good idea to rely on automated agents for help.	◯	◯	◯	◯	◯
I don't trust the information I get from automated agents.	◯	◯	◯	◯	◯
Automated agents are reliable.	◯	◯	◯	◯	◯
I rely on automated agents.	◯	◯	◯	◯	◯

Tasks, tools, and data

These instructions can be found as a PDF here: https://drive.google.com/file/d/1nq-8rKE0FR_2vz-UNiMW1isANYcRXVSg/view?usp=sharing. Open this PDF in a separate window, so you have access to it throughout the experiment without going backwards in the form (which results in an invalid submission).

Tasks and tools

Your task is to make predictions of adult Intensive Care Unit (ICU) bed utilization in a state in the USA. Predictions like these are relevant for real decision making during the COVID19-pandemic: for example, in New York "regions must have at least 30 percent of their ICU beds available before a phased re-opening can begin" (i.e., the ICU bed utilization must be below 70 %). (https://forward.ny.gov/metrics-guide-reopening-new-york)

For each prediction you will see data for the previous 28 days on ICU bed utilization and other variables to consider (days -28 to -1). Based on these, you are to make a prediction to answer a question of the form:

> *"Based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 70 % on any of the coming seven days (days 0–6)?"*

The threshold X % will vary and is indicated for each prediction task.

Therefore, your prediction is either:

- **Yes** (the value will exceed the threshold on at least one of the days); or
- **No** (the value will not exceed the threshold on any of the days).

This is your initial prediction. Then, an Automated Machine Learning (AutoML) system will show its prediction, based on the same data. Then you will make your final prediction, so you have a chance to adjust your prediction and take into account the prediction of the AutoML system. Remember that neither you nor the AutoML system knows the true future values you are predicting.

This is repeated ten times. Finally, you will fill in another short questionnaire based on your experiences with the AutoML system.

Your goal in this experiment should be to make as many correct initial and final predictions as possible. For each correct prediction, you will get a bonus of $0.15, yielding a maximum bonus of $3 (with a total of 10+10=20 correct predictions).

Data

The state-level data and variables used in this experiment are described below. Note that the figure scales of the variables can change between tasks. The description can also be found in the PDF (see the link at the top of this page) with the rest of the instructions.

Description of data variables

The state-level data and variables used in this experiment are described below. Note that the figure scales of the variables can change between tasks.

Predicted variable: Adult Intensive Care Unit bed utilization

- Percentage of Intensive Care Unit (ICU) beds for adults in use in the hospitals of this state.

  "Intensive care units cater to patients with severe or life-threatening illnesses and injuries, which require constant care, close supervision from life support equipment and medication in order to ensure normal bodily functions. They are staffed by highly trained physicians, nurses and respiratory therapists who specialize in caring for critically ill patients." [1]

  Often, the ICU bed capacity is small compared to the whole capacity of the hospital. In addition, patients in need of ICU beds are in grave need of ICU care (i.e., a bed). Thus, a fully utilized ICU capacity is often not preferred, in case of sudden increased demand on ICU beds (such as due to the COVID19-pandemic).

Other variables relevant for making predictions about the adult ICU bed utilization:

- **New COVID-19 cases**
  - Number of new laboratory-confirmed COVID-19 cases in this state.

    Every COVID-19 case has some chance of resulting in a hospitalization. Every new case can therefore lead to a higher adult ICU bed utilization: the median time from the onset of COVID-19 to ICU admission is 9.5–12 days. [2]
- **Inpatient beds utilization**
  - Percentage of inpatient beds that are being utilized in the hospitals in this state.

    Inpatient beds are hospital beds (of any kind) for patients that either require a bed or will have to stay one or more nights at the hospital. Inpatient beds include ICU beds, but ICU beds usually make up only a small portion of total inpatient beds. Inpatient beds are often used to describe the capacity of a hospital, since no free inpatient beds means that a hospital cannot admit any new patients to inpatient care.

    If inpatient bed utilization is high, there may be a risk of ICU bed utilization also increasing soon: if the state of patients' (whether with COVID-19 or not) worsens, they may be transferred to ICU beds.
- **Percent of inpatients with COVID-19**
  - Percentage of inpatients (i.e., patients occupying an inpatient bed) in the hospitals of this state that have suspected or confirmed COVID-19.

    Since COVID-19 patients can develop severe symptoms, there is a risk of them being transferred to ICU beds; about 30 % of hospitalized COVID-19 patients are transferred to the ICU [2]. Therefore, if the percentage of inpatients (whether in ICU beds or not) is high or increases, there can be a higher probability of patients being transferred to ICU beds.

    In addition, the median length of hospitalization among [COVID-19] survivors is 10–13 days [2], which means COVID-19 patients may occupy beds for a relatively long period.

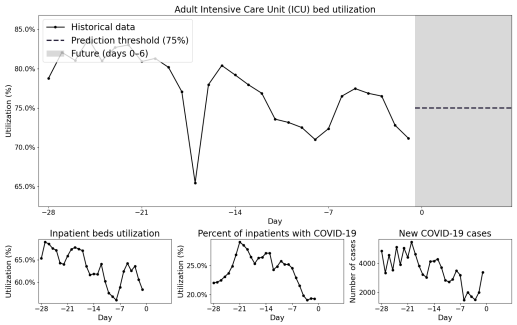
References:
[1] https://en.wikipedia.org/wiki/Intensive_care_unit
[2] https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html

Tasks 1–10

Now you will make ten predictions without and ten prediction with the help of an AutoML system, as described in the instructions. Remember, you are not allowed to go back at any point; doing so will result in your submission being invalid.

Task 1: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



13. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

14. How confident are you in your prediction? *

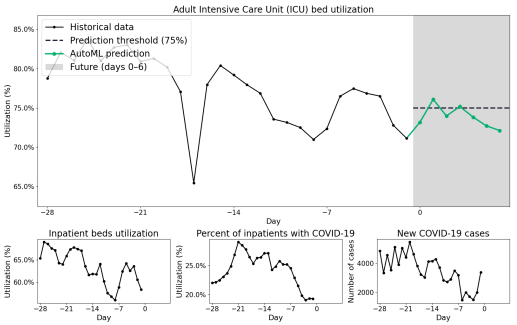
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 1: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



AutoML prediction: Yes

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL exceed the indicated threshold in the following seven days (days 0–6 in the figure).

15. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

16. How confident are you in your prediction? *

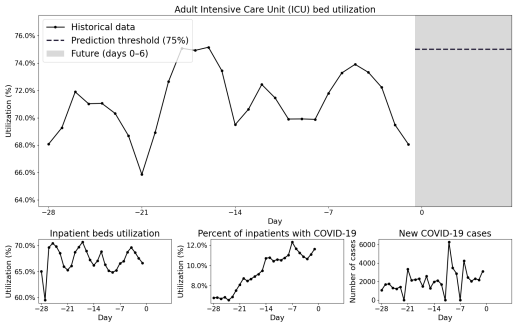
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 2: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



17. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

18. How confident are you in your prediction? *

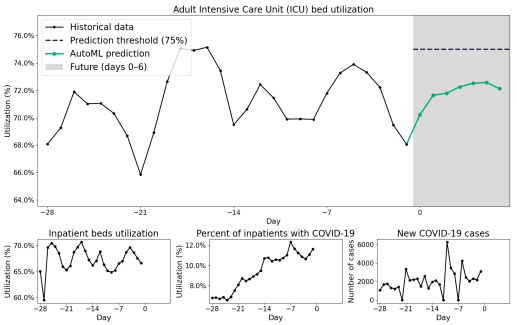
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 2: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

19. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

20. How confident are you in your prediction? *

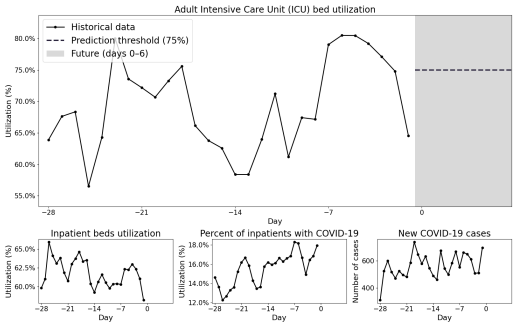
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 3: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



21. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

22. How confident are you in your prediction? *

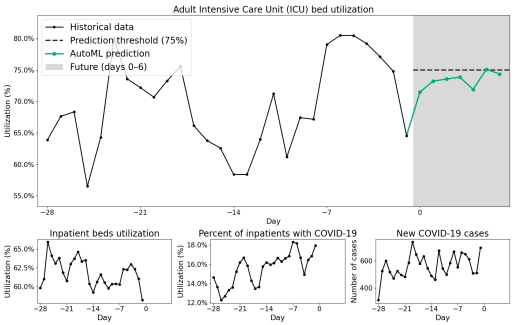
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 3: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



AutoML prediction: Yes

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL exceed the indicated threshold in the following seven days (days 0–6 in the figure).

23. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

24. How confident are you in your prediction? *

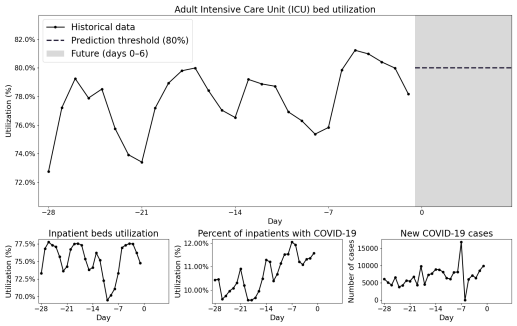
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 4: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



25. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

26. How confident are you in your prediction? *

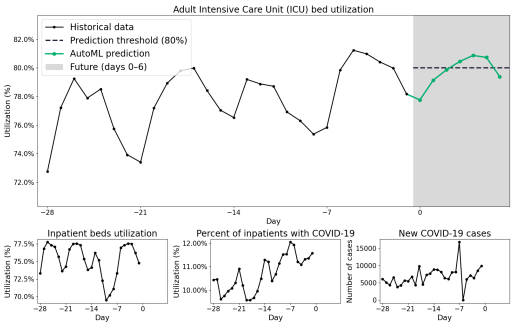
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 4: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



AutoML prediction: Yes

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL exceed the indicated threshold in the following seven days (days 0–6 in the figure).

27. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

28. How confident are you in your prediction? *

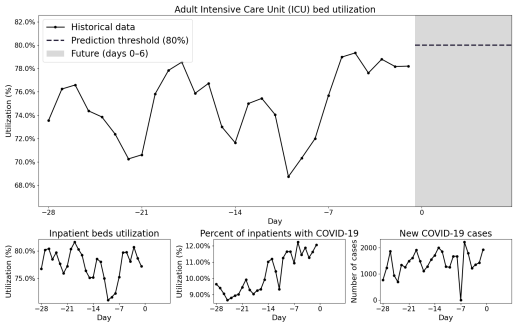
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 5: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



29. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

30. How confident are you in your prediction? *

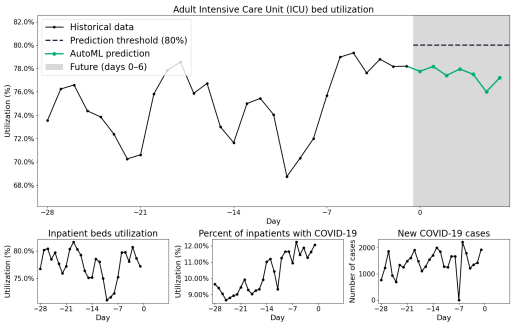
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 5: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

31. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

32. How confident are you in your prediction? *

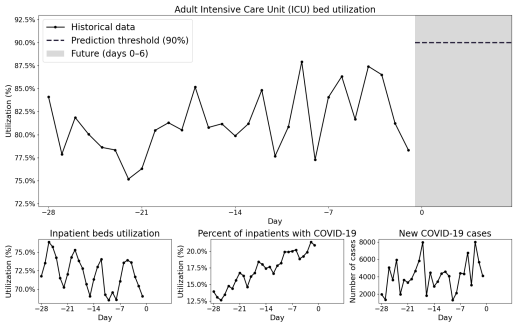
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 6: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



33. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 90 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

34. How confident are you in your prediction? *

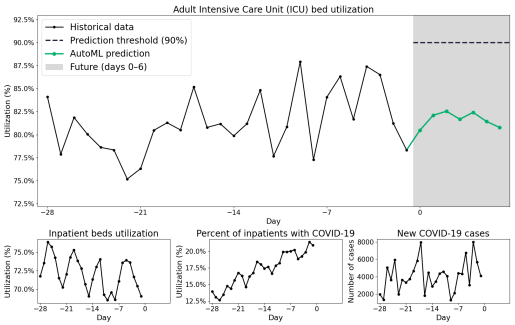
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 6: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

35. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 90 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

36. How confident are you in your prediction? *

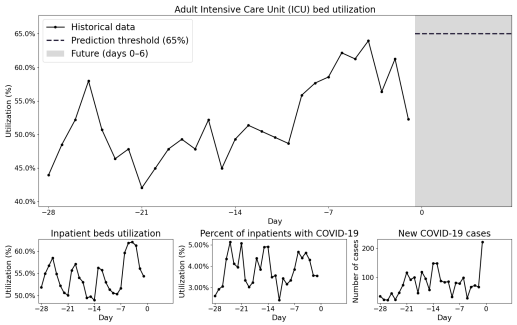
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 7: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



37. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

38. How confident are you in your prediction? *

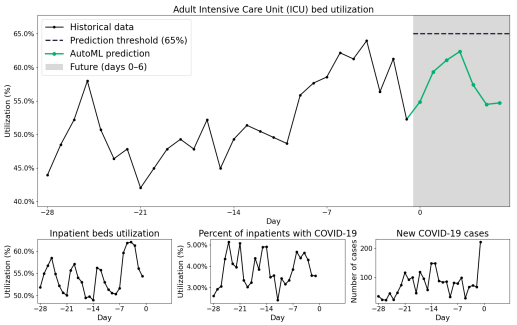
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 7: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

39. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

40. How confident are you in your prediction? *

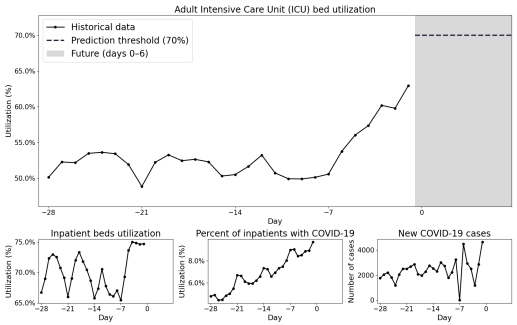
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 8: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



41. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 70 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

42. How confident are you in your prediction? *

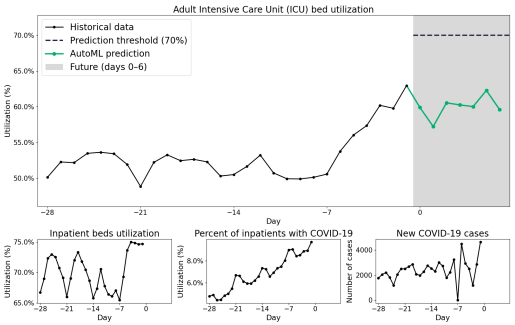
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 8: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

43. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 70 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

44. How confident are you in your prediction? *

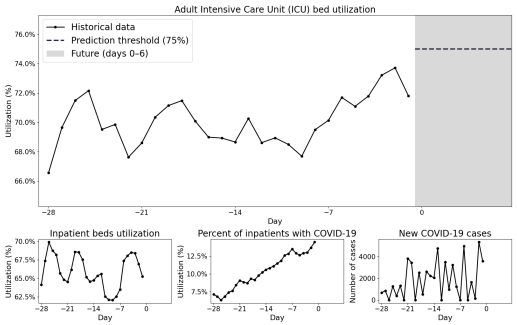
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 9: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



45. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

46. How confident are you in your prediction? *

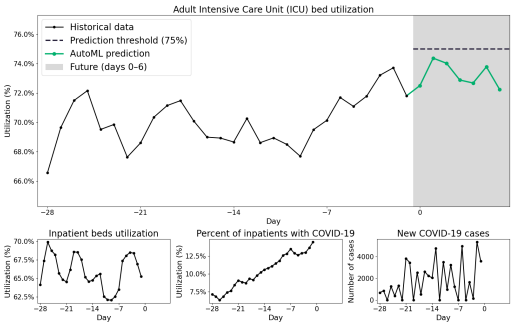
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 9: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

47. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

48. How confident are you in your prediction? *

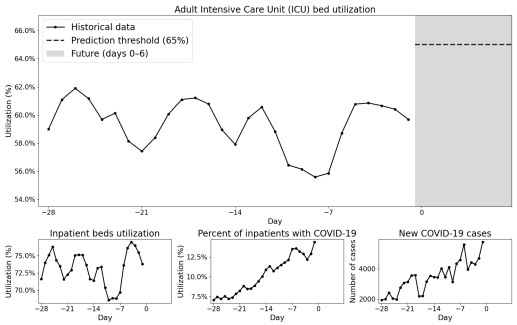
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 10: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



49. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

50. How confident are you in your prediction? *

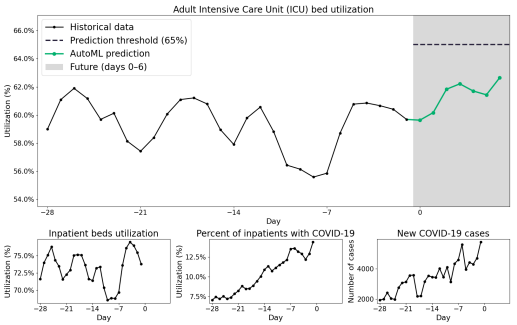
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 10: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

51. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

52. How confident are you in your prediction? *

NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Questionnaire on your experiences with the AutoML system

53. Please answer the following questions based on your experiences with the AutoML system that provided you with the predictions for the previous ten tasks. *

NOTE: there are five levels to choose from for each row, please scroll if you cannot see all alternatives.

Mark only one oval per row.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I believe the AutoML system is a competent performer.					
I trust the AutoML system.					
I have confidence in the advice given by the AutoML system.					
I can depend on the AutoML system.					
I can rely on the AutoML system to behave in consistent ways.					
I can rely on the AutoML system to do its best every time I take its advice.					

MTurk questions and information

The following information is needed for approving your task (and possible bonus) payment.

54. Please fill in your MTurk Worker ID *

This will be used to verify that you have completed the task before approving and paying. See e.g. https://blog.mturk.com/get-to-know-the-new-worker-site-4a69967d90c3 for instructions to find your WorkerID.

Feedback

55. Approximately how long did you take to complete the survey? *

Example: 4:03:32 (4 hours, 3 minutes, 32 seconds)

56. Do you have feedback on the AutoML system? *

E.g., what made them (un)trustworthy, what would make them more or less trustworthy, or what would you require of an AutoML system to trust it enough to use it as support for decision making?

57. Do you have any feedback on the experiment itself? *

E.g., was the structure of the experiment clear, were the instructions clear, would you have wanted more or less instructions, did the order of tasks make sense, ...

58. Do you have any other feedback? *

Next: submission

Next, you will submit your answers. In the survey completion message, you will be shown a debriefing of the survey. Do not close the window yet, instead read through the debriefing. At the end you will find the code you will need to fill in in your HIT on MTurk to verify that you have completed the survey and get your payment.

In addition, after submitting you can see your results, i.e. how many of your (10) initial and (10) final predictions were correct (which is the basis for any bonus to be paid to you).

A.2 Survey, group 2.1: SHAP

AutoML survey group 2.1

This study is done as part of a Master's Thesis, in collaboration with MindsDB (mindsdb.com), and the results and the thesis will be published later this year.

In this survey, you will use the predictions of an Automated Machine Learning (AutoML) system. AutoML systems automate one or more phases of the process of creating a Machine Learning (ML) model; this way, anyone can create a complex ML model without deep knowledge of ML itself. Machine Learning (ML) describes a type of computer program that learns in a way similar to human beings. Rather than employ statistical techniques or simple mathematics to analyze a large amount of data, the ML model learns by example and is able to digest a large amount of data, and be trained to give incrementally more and more accurate predictions if it is fed more and more data.

Your task in this survey will be to make predictions about adult Intensive Care Unit (ICU) bed utilization during the COVID-19 pandemic. Your goal is to estimate the percentage of adult ICU beds occupied, given a certain scenario. You will also be shown the predictions of an AutoML system to take into account, and the AutoML system will also explain its predictions.

Your goal should be to make as many correct predictions as possible.

The survey proceeds as follows.
1. Before the tasks begin, we will ask you to fill in some necessary information about yourself and give your informed consent about us gathering and storing your anonymized answers.
2. Then you will be shown a description of the data used and how you should make your predictions. You will also be shown instructions on how to interpret the AutoML system's explanations for its predictions.
3. Then, you will make predictions in ten scenarios, with support from an AutoML system. After these predictions, you will fill in a questionnaire based on your experience with the AutoML system.

After this the study is done. You will see how your predictions compare to the actual adult ICU bed utilization. You will also have the opportunity to give feedback on the experiment and the AutoML systems used.

If you have any questions, please contact the survey organizer and thesis author Cosmo Jenytin (cosmo@mindsdb.com).

* Required

1. Informed consent: I consent to my answers being used and published anonymously. The background information asked here will be connected to your survey answers and anonymized. *

   Mark only one oval.

   - ◯ Yes

Note: you are not allowed to go backwards in the survey for any reason. If you go back to a previous page, your submission will be invalid.

All the instructions can be found as a PDF here: https://drive.google.com/file/d/1h_mj3mZBT15C63PLb1864J2I5ekQ3R1r/view?usp=sharing. Open this PDF in a separate window, so you have access to it throughout the experiment without going backwards in the form (which results in an invalid submission).

2. I understand that I am not allowed to go backwards (go back to previous pages) in the survey, and that doing so will result in my submission being invalid. *

   Mark only one oval.

   - ◯ Yes

Background information

Please answer the following questions. Your answers will be used and stored anonymously.

3. Age *

   _____

4. I identify my gender as *

   _____

5. Education *

   Mark only one oval.

   - ◯ Upper secondary
   - ◯ Bachelor's or equivalent
   - ◯ Master's or equivalent
   - ◯ Doctoral or equivalent
   - ◯ Other: _____

6. Professional field *

Mark only one oval.

- Agriculture, Food and Natural Resources
- Architecture and Construction
- Arts, Audio/Video Technology and Communications
- Business Management and Administration
- Education and Training
- Finance
- Government and Public Administration
- Health and Medicine
- Hospitality and Tourism
- Human Services
- Information Technology
- Law, Public Safety, Corrections and Security
- Manufacturing
- Marketing, Sales, and Service
- Science, Technology, Engineering, and Mathematics
- Transportation, Distribution, and Logistics
- Other: _____

7. Job title *

Mark only one oval.

- Chairperson, member of Board of Directors
- CEO, CxO
- Vice President
- Manager
- Individual Contributor
- Entry-level
- Other: _____

8. Job experience (years) *

_____

9. How confident are you, in general, in your ability to predict future developments based on data? *

NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

10. Have you previously worked with data analysis or predictions? *

This can, for example, mean working with simple data analysis tools (e.g., Excel) or more advanced tools, such as Machine Learning (ML) models.

Mark only one oval.

- Yes
- No

11. Have you previously used an Automated Machine Learning (AutoML) system? *

An AutoML system is a Machine Learning system that can automate one or more phases of a data science or Machine Learning model development process.

Mark only one oval.

- Yes
- No

12. Please answer the following questions on automated agents. *

An automated agent runs by computerized algorithms and interacts with humans. For example, a website predicting a medical diagnosis based on symptoms is an automated agent; likewise, an ATM is an automated agent. NOTE: there are five levels to choose from, please scroll if you cannot see all alternatives.

Mark only one oval per row.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Generally, I trust automated agents.	◯	◯	◯	◯	◯
Automated agents help me solve many problems.	◯	◯	◯	◯	◯
I think it's a good idea to rely on automated agents for help.	◯	◯	◯	◯	◯
I don't trust the information I get from automated agents.	◯	◯	◯	◯	◯
Automated agents are reliable.	◯	◯	◯	◯	◯
I rely on automated agents.	◯	◯	◯	◯	◯

Tasks, tools, and data

These instructions can be found as a PDF here: https://drive.google.com/file/d/1h_mj3mZBT15C63PLb1864J2l5ekQ3R1r/view?usp=sharing. Open this PDF in a separate window, so you have access to it throughout the experiment without going backwards in the form (which results in an invalid submission).

Tasks and tools

Your task is to make predictions of adult Intensive Care Unit (ICU) bed utilization in a state in the USA. Predictions like these are relevant for real decision making during the COVID19-pandemic: for example, in New York "regions must have at least 30 percent of their ICU beds available before a phased re-opening can begin" (i.e., the ICU bed utilization must be below 70 %). (https://forward.ny.gov/metrics-guide-reopening-new-york)

For each prediction you will see data for the previous 28 days on ICU bed utilization and other variables to consider (days -28 to -1). Based on these, you are to make a prediction to answer a question of the form:

> *"Based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 70 % on any of the coming seven days (days 0–6)?"*

The threshold X % will vary and is indicated for each prediction task.

Therefore, your prediction is either:

- **Yes** (the value will exceed the threshold on at least one of the days); or
- **No** (the value will not exceed the threshold on any of the days).

This is your initial prediction. Then, an Automated Machine Learning (AutoML) system will show its prediction, based on the same data. Then you will make your final prediction, so you have a chance to adjust your prediction and take into account the prediction of the AutoML system. Remember that neither you nor the AutoML system knows the true future values you are predicting.

This is repeated ten times. Finally, you will fill in another short questionnaire based on your experiences with the AutoML system.

Your goal in this experiment should be to make as many correct initial and final predictions as possible. For each correct prediction, you will get a bonus of $0.15, yielding a maximum bonus of $3 (with a total of 10+10=20 correct predictions).

Data

The state-level data and variables used in this experiment are described below. Note that the figure scales of the variables can change between tasks. The description can also be found in the PDF (see the link at the top of this page) with the rest of the instructions.

Description of data variables

The state-level data and variables used in this experiment are described below. Note that the figure scales of the variables can change between tasks.

Predicted variable: Adult Intensive Care Unit bed utilization

- Percentage of Intensive Care Unit (ICU) beds for adults in use in the hospitals of this state.

  "Intensive care units cater to patients with severe or life-threatening illnesses and injuries, which require constant care, close supervision from life support equipment and medication in order to ensure normal bodily functions. They are staffed by highly trained physicians, nurses and respiratory therapists who specialize in caring for critically ill patients." [1]

  Often, the ICU bed capacity is small compared to the whole capacity of the hospital. In addition, patients in need of ICU beds are in grave need of ICU care (i.e., a bed). Thus, a fully utilized ICU capacity is often not preferred, in case of sudden increased demand on ICU beds (such as due to the COVID19-pandemic).

Other variables relevant for making predictions about the adult ICU bed utilization:

- **New COVID-19 cases**
  - Number of new laboratory-confirmed COVID-19 cases in this state.

    Every COVID-19 case has some chance of resulting in a hospitalization. Every new case can therefore lead to a higher adult ICU bed utilization: the median time from the onset of COVID-19 to ICU admission is 9.5–12 days. [2]
- **Inpatient beds utilization**
  - Percentage of inpatient beds that are being utilized in the hospitals in this state.

    Inpatient beds are hospital beds (of any kind) for patients that either require a bed or will have to stay one or more nights at the hospital. Inpatient beds include ICU beds, but ICU beds usually make up only a small portion of total inpatient beds. Inpatient beds are often used to describe the capacity of a hospital, since no free inpatient beds means that a hospital cannot admit any new patients to inpatient care.

    If inpatient bed utilization is high, there may be a risk of ICU bed utilization also increasing soon: if the state of patients' (whether with COVID-19 or not) worsens, they may be transferred to ICU beds.
- **Percent of inpatients with COVID-19**
  - Percentage of inpatients (i.e., patients occupying an inpatient bed) in the hospitals of this state that have suspected or confirmed COVID-19.

    Since COVID-19 patients can develop severe symptoms, there is a risk of them being transferred to ICU beds; about 30 % of hospitalized COVID-19 patients are transferred to the ICU [2]. Therefore, if the percentage of inpatients (whether in ICU beds or not) is high or increases, there can be a higher probability of patients being transferred to ICU beds.

    In addition, the median length of hospitalization among [COVID-19] survivors is 10–13 days [2], which means COVID-19 patients may occupy beds for a relatively long period.

References:

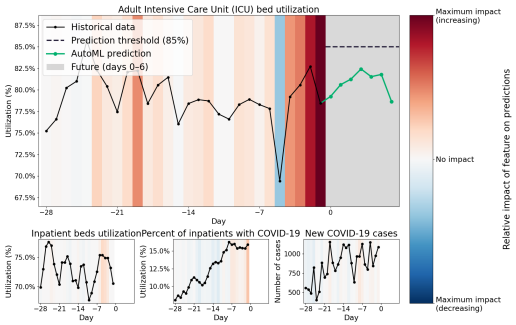
[1] https://en.wikipedia.org/wiki/Intensive_care_unit
[2] https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html

Tasks 1–10

Now you will make ten predictions without and ten prediction with the help of an AutoML system, as described in the instructions. Remember, you are not allowed to go back at any point; doing so will result in your submission being invalid.

The AutoML system will make predictions for you to consider; in addition, it explains its predictions. Below is an example of an explanation, and some remarks on how one can interpret it.

Example explanation of the AutoML system



Prediction explanations: data importances, i.e. "What parts of the data contributed most to this specific prediction?"

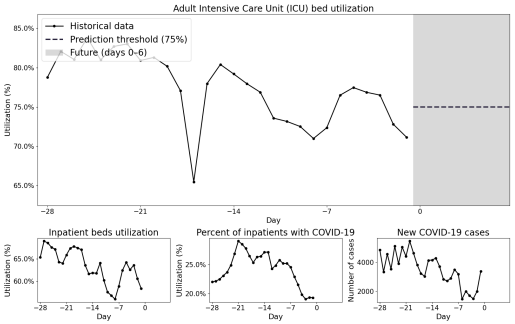
For this explanation, the AutoML system has used mathematical techniques to determine how much an individual data point contributes to the final prediction.

The explanations show visually how different points of data contributed to the predictions, relative to each other. A red background indicates the data point had a positive impact (i.e. increased the predicted utilizations); a blue background indicates the data had a negative impact (i.e. decreased the predicted utilizations); a white background indicates no significant impact on the predictions. A deeper color indicates a more significant impact, relative to the other data. The deepest color (red or blue) indicates that data had the biggest (positive or negative) impact of all data.

In the above example, one could, for example, look at the colors to see how reasonable the model seems to be. The large dip in adult ICU bed utilization on day -5 has a significant negative impact on the predictions (i.e., this decreases the predicted values), which seems to make sense given that a lower previous ICU bed utilization probably means the utilization in the near future will be somewhat lower than otherwise.

Task 1: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



13. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

14. How confident are you in your prediction? *

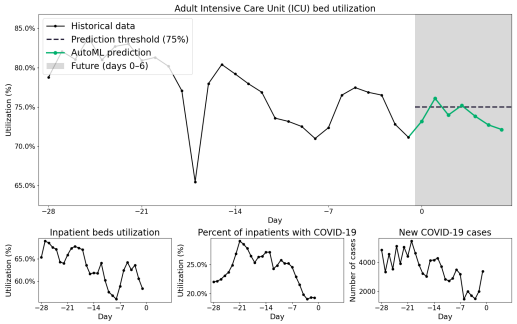
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

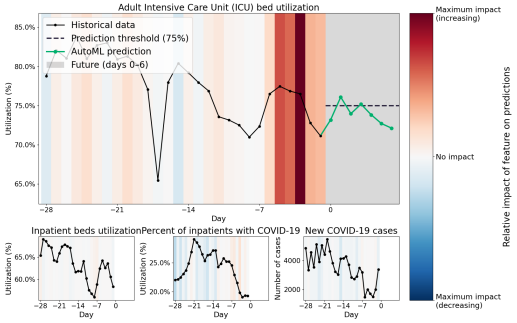
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 1: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: Yes

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL exceed the indicated threshold in the following seven days (days 0–6 in the figure).

15. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

16. How confident are you in your prediction? *

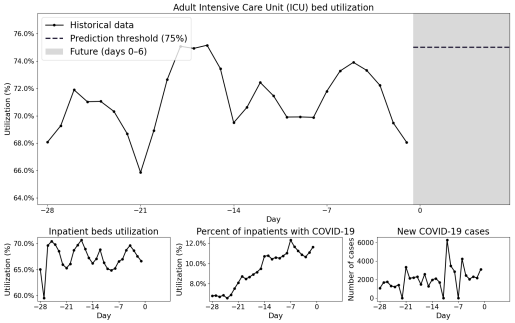
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 2: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



17. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

18. How confident are you in your prediction? *

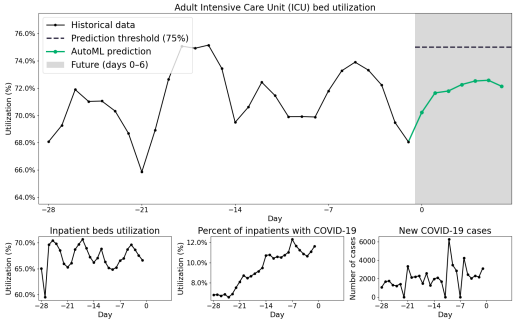
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

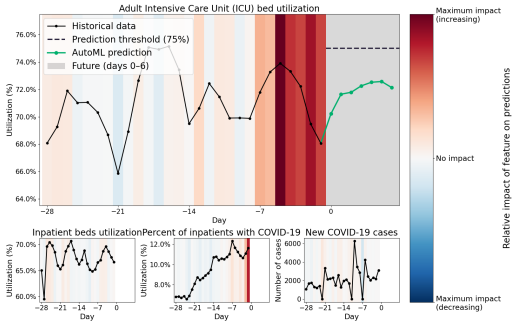
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 2: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

19. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

20. How confident are you in your prediction? *

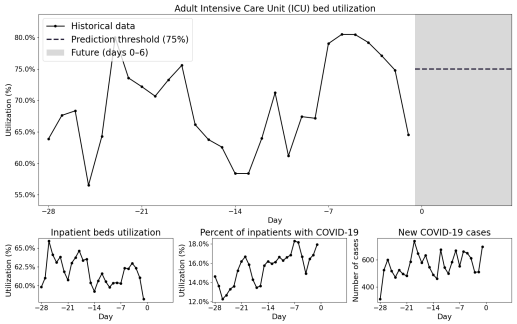
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 3: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



21. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

22. How confident are you in your prediction? *

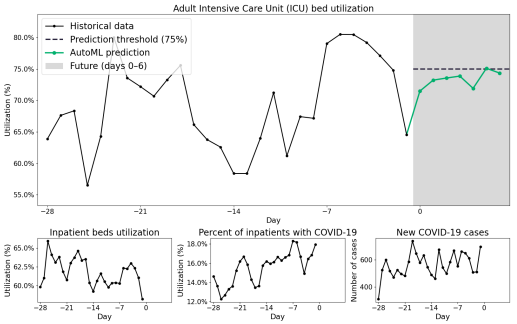
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

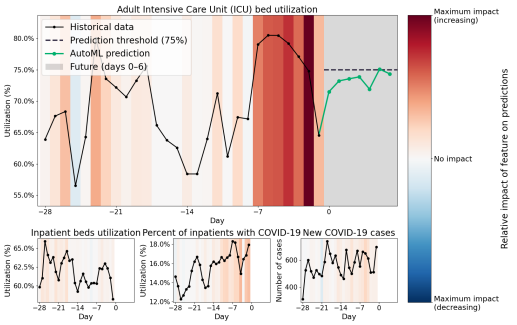
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident												Very confident

Task 3: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: Yes

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL exceed the indicated threshold in the following seven days (days 0–6 in the figure).

23. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

24. How confident are you in your prediction? *

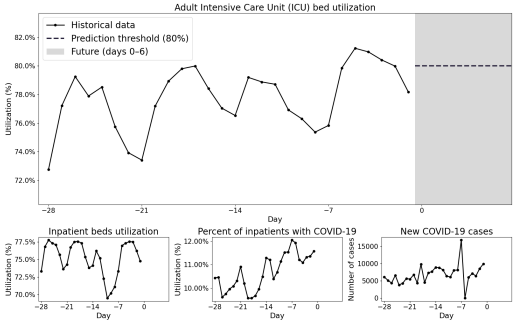
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 4: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



25. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

26. How confident are you in your prediction? *

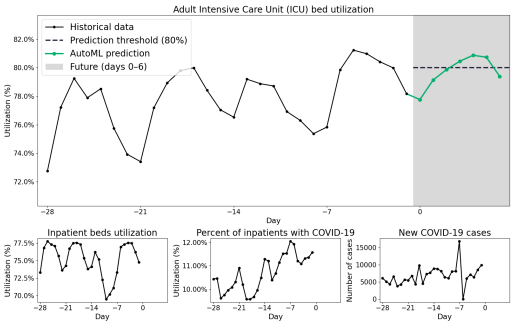
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

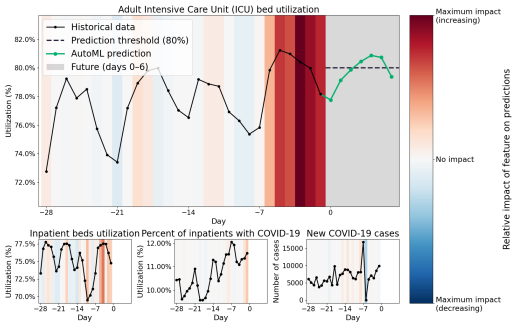
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 4: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: Yes

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL exceed the indicated threshold in the following seven days (days 0–6 in the figure).

27. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

28. How confident are you in your prediction? *

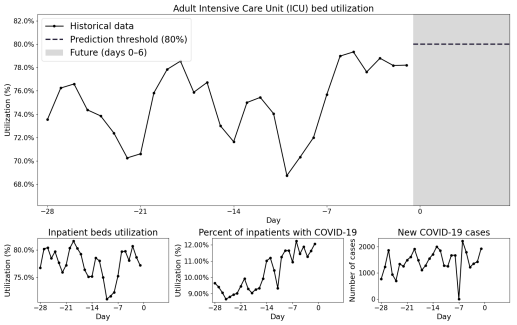
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 5: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



29. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

30. How confident are you in your prediction? *

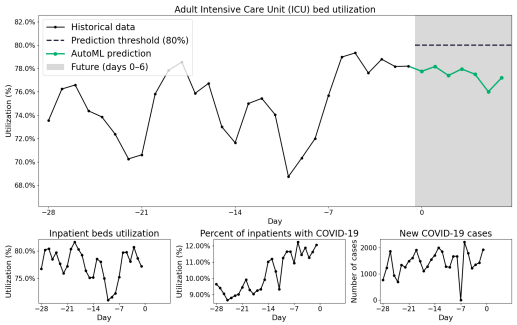
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

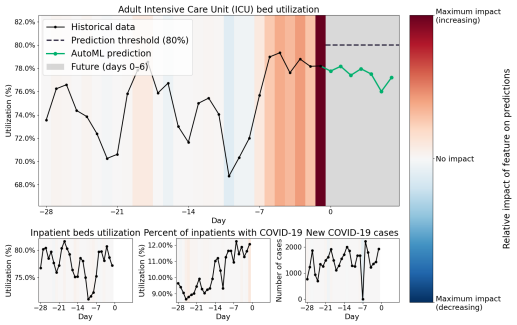
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 5: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

31. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

32. How confident are you in your prediction? *

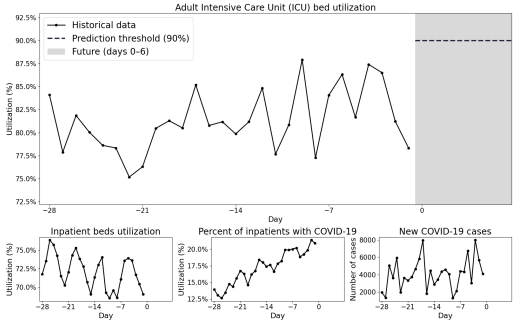
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 6: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



33. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 90 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

34. How confident are you in your prediction? *

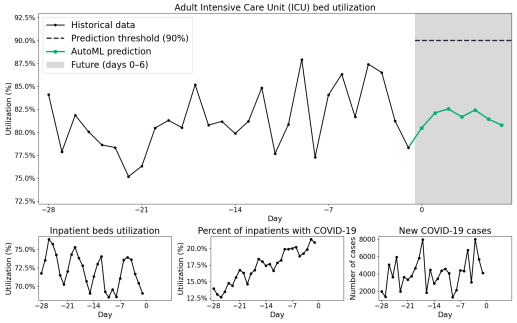
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

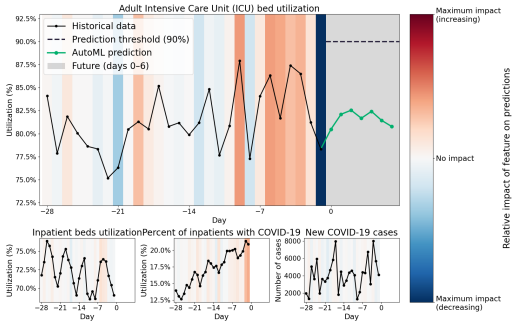
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 6: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

35. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 90 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

36. How confident are you in your prediction? *

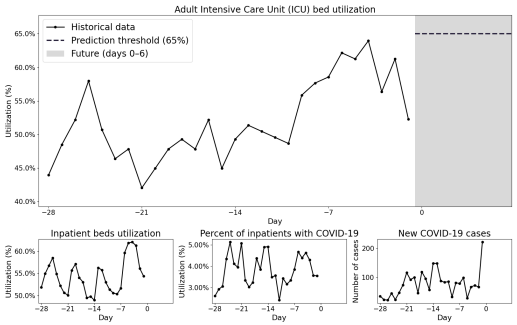
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 7: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



37. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

38. How confident are you in your prediction? *

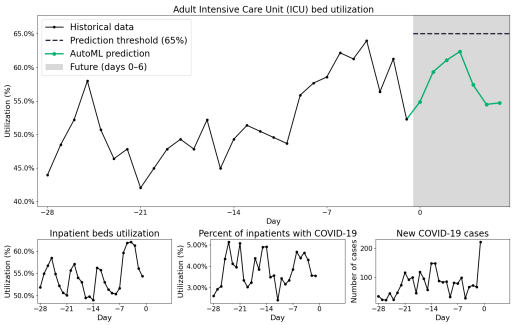
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

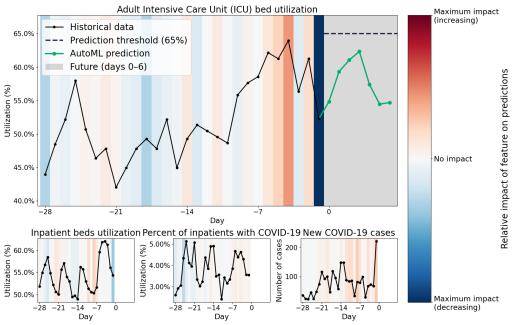
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	⬭	⬭	⬭	⬭	⬭	⬭	⬭	⬭	⬭	⬭	⬭	Very confident

Task 7: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

39. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

40. How confident are you in your prediction? *

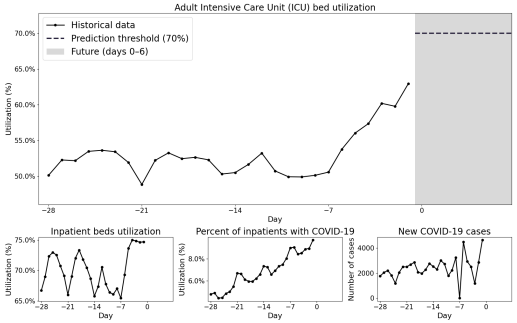
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 8: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



41. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 70 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

42. How confident are you in your prediction? *

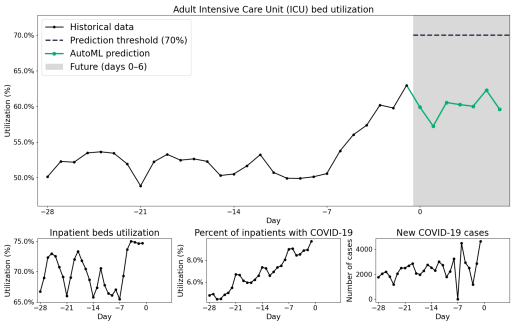
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

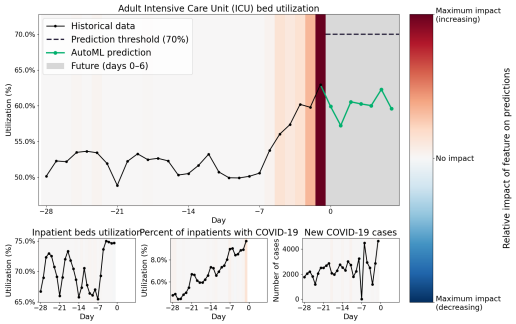
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 8: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

43. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 70 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

44. How confident are you in your prediction? *

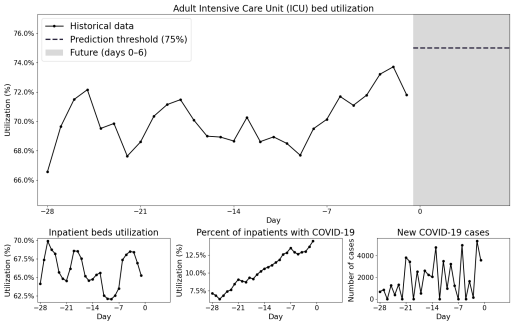
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 9: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



45. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

46. How confident are you in your prediction? *

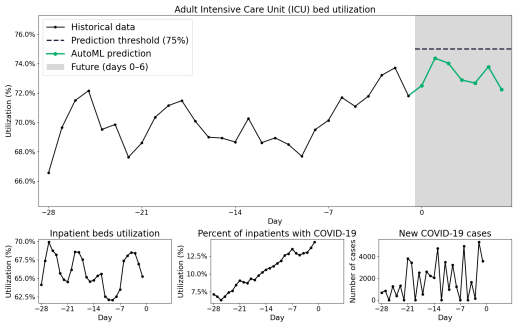
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

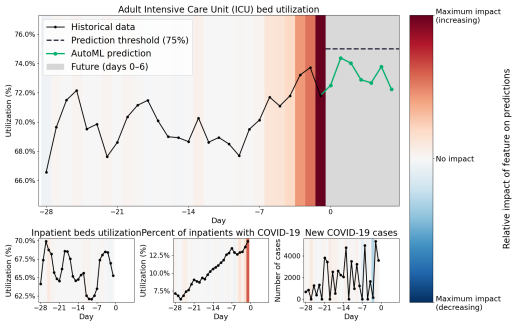
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 9: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

47. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

48. How confident are you in your prediction? *

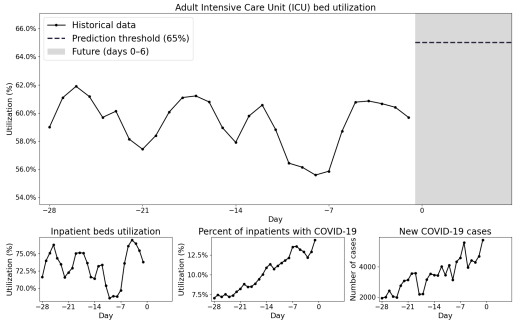
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 10: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



49. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

50. How confident are you in your prediction? *

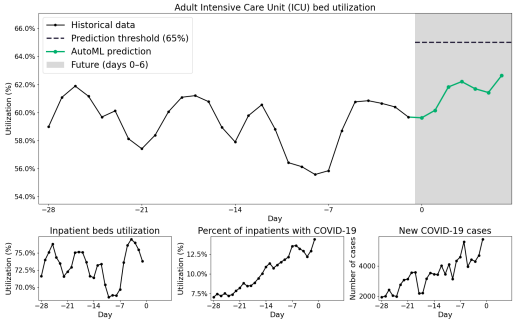
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

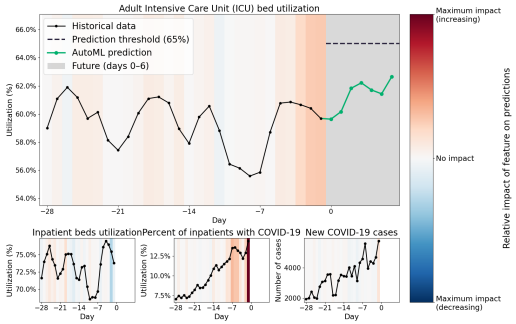
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 10: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

51. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

52. How confident are you in your prediction? *

NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Questionnaire on your experiences with the AutoML system

53. Please answer the following questions based on your experiences with the AutoML system that provided you with the predictions for the previous ten tasks. *

NOTE: there are five levels to choose from for each row, please scroll if you cannot see all alternatives.

Mark only one oval per row.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I believe the AutoML system is a competent performer.	◯	◯	◯	◯	◯
I trust the AutoML system.	◯	◯	◯	◯	◯
I have confidence in the advice given by the AutoML system.	◯	◯	◯	◯	◯
I can depend on the AutoML system.	◯	◯	◯	◯	◯
I can rely on the AutoML system to behave in consistent ways.	◯	◯	◯	◯	◯
I can rely on the AutoML system to do its best every time I take its advice.	◯	◯	◯	◯	◯

MTurk questions and information

The following information is needed for approving your task (and possible bonus) payment.

54. Please fill in your MTurk Worker ID *

This will be used to verify that you have completed the task before approving and paying. See e.g. https://blog.mturk.com/get-to-know-the-new-worker-site-4a69967d90c3 for instructions to find your WorkerID.

Feedback

55. Approximately how long did you take to complete the survey? *

Example: 4:03:32 (4 hours, 3 minutes, 32 seconds)

56. Do you have feedback on the AutoML system? *

E.g., what made them (un)trustworthy, what would make them more or less trustworthy, or what would you require of an AutoML system to trust it enough to use it as support for decision making?

57. Do you have any feedback on the experiment itself? *

E.g., was the structure of the experiment clear, were the instructions clear, would you have wanted more or less instructions, did the order of tasks make sense, ...

58. Do you have any other feedback? *

Next: submission

Next, you will submit your answers. In the survey completion message, you will be shown a debriefing of the survey. Do not close the window yet, instead read through the debriefing. At the end you will find the code you will need to fill in in your HIT on MTurk to verify that you have completed the survey and get your payment.

In addition, after submitting you can see your results, i.e. how many of your (10) initial and (10) final predictions were correct (which is the basis for any bonus to be paid to you).

A.3 Survey, group 2.2: evidence-based

AutoML survey group 2.2

This study is done as part of a Master's Thesis, in collaboration with MindsDB (mindsdb.com), and the results and the thesis will be published later this year.

In this survey, you will use the predictions of an Automated Machine Learning (AutoML) system. AutoML systems automate one or more phases of the process of creating a Machine Learning (ML) model; this way, anyone can create a complex ML model without deep knowledge of ML itself. Machine Learning (ML) describes a type of computer program that learns in a way similar to human beings. Rather than employ statistical techniques or simple mathematics to analyze a large amount of data, the ML model learns by example and is able to digest a large amount of data, and be trained to give incrementally more and more accurate predictions if it is fed more and more data.

Your task in this survey will be to make predictions about adult Intensive Care Unit (ICU) bed utilization during the COVID-19 pandemic. Your goal is to estimate the percentage of adult ICU beds occupied, given a certain scenario. You will also be shown the predictions of an AutoML system to take into account, and the AutoML system will also explain its predictions.

Your goal should be to make as many correct predictions as possible.

The survey proceeds as follows.
1. Before the tasks begin, we will ask you to fill in some necessary information about yourself and give your informed consent about us gathering and storing your anonymized answers.
2. Then you will be shown a description of the data used and how you should make your predictions. You will also be shown instructions on how to interpret the AutoML system's explanations for its predictions.
3. Then, you will make predictions in ten scenarios, with support from an AutoML system. After these predictions, you will fill in a questionnaire based on your experience with the AutoML system.

After this the study is done. You will see how your predictions compare to the actual adult ICU bed utilization. You will also have the opportunity to give feedback on the experiment and the AutoML systems used.

If you have any questions, please contact the survey organizer and thesis author Cosmo Jenytin (cosmo@mindsdb.com).

* Required

1. Informed consent: I consent to my answers being used and published anonymously. The background information asked here will be connected to your survey answers and anonymized. *

   Mark only one oval.

   ◯ Yes

Note: you are not allowed to go backwards in the survey for any reason. If you go back to a previous page, your submission will be invalid.

All the instructions can be found as a PDF here: https://drive.google.com/file/d/1E3J84KQm9OIzSpYaEuFmDwjEkwqtzxa2/view?usp=sharing. Open this PDF in a separate window, so you have access to it throughout the experiment without going backwards in the form (which results in an invalid submission).

2. I understand that I am not allowed to go backwards (go back to previous pages) in the survey, and that doing so will result in my submission being invalid. *

   Mark only one oval.

   ◯ Yes

Background information

Please answer the following questions. Your answers will be used and stored anonymously.

3. Age *

   ____

4. I identify my gender as *

   ____

5. Education *

   Mark only one oval.

   ◯ Upper secondary
   ◯ Bachelor's or equivalent
   ◯ Master's or equivalent
   ◯ Doctoral or equivalent
   ◯ Other: ______________________

6. Professional field *

Mark only one oval.

- ◯ Agriculture, Food and Natural Resources
- ◯ Architecture and Construction
- ◯ Arts, Audio/Video Technology and Communications
- ◯ Business Management and Administration
- ◯ Education and Training
- ◯ Finance
- ◯ Government and Public Administration
- ◯ Health and Medicine
- ◯ Hospitality and Tourism
- ◯ Human Services
- ◯ Information Technology
- ◯ Law, Public Safety, Corrections and Security
- ◯ Manufacturing
- ◯ Marketing, Sales, and Service
- ◯ Science, Technology, Engineering, and Mathematics
- ◯ Transportation, Distribution, and Logistics
- ◯ Other: _____

7. Job title *

Mark only one oval.

- ◯ Chairperson, member of Board of Directors
- ◯ CEO, CxO
- ◯ Vice President
- ◯ Manager
- ◯ Individual Contributor
- ◯ Entry-level
- ◯ Other: _____

8. Job experience (years) *

_____

9. How confident are you, in general, in your ability to predict future developments based on data? *

NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

10. Have you previously worked with data analysis or predictions? *

This can, for example, mean working with simple data analysis tools (e.g., Excel) or more advanced tools, such as Machine Learning (ML) models.

Mark only one oval.

- ◯ Yes
- ◯ No

11. Have you previously used an Automated Machine Learning (AutoML) system? *

An AutoML system is a Machine Learning system that can automate one or more phases of a data science or Machine Learning model development process.

Mark only one oval.

- ◯ Yes
- ◯ No

12. Please answer the following questions on automated agents. *

An automated agent runs by computerized algorithms and interacts with humans. For example, a website predicting a medical diagnosis based on symptoms is an automated agent; likewise, an ATM is an automated agent. NOTE: there are five levels to choose from, please scroll if you cannot see all alternatives.

Mark only one oval per row.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Generally, I trust automated agents.	◯	◯	◯	◯	◯
Automated agents help me solve many problems.	◯	◯	◯	◯	◯
I think it's a good idea to rely on automated agents for help.	◯	◯	◯	◯	◯
I don't trust the information I get from automated agents.	◯	◯	◯	◯	◯
Automated agents are reliable.	◯	◯	◯	◯	◯
I rely on automated agents.	◯	◯	◯	◯	◯

Tasks, tools, and data

These instructions can be found as a PDF here: https://drive.google.com/file/d/1E3J84KQm9OIzSpYaEuFmDwjEkwqtzxa2/view?usp=sharing. Open this PDF in a separate window, so you have access to it throughout the experiment without going backwards in the form (which results in an invalid submission).

Tasks and tools

Your task is to make predictions of adult Intensive Care Unit (ICU) bed utilization in a state in the USA. Predictions like these are relevant for real decision making during the COVID19-pandemic: for example, in New York "regions must have at least 30 percent of their ICU beds available before a phased re-opening can begin" (i.e., the ICU bed utilization must be below 70 %). (https://forward.ny.gov/metrics-guide-reopening-new-york)

For each prediction you will see data for the previous 28 days on ICU bed utilization and other variables to consider (days -28 to -1). Based on these, you are to make a prediction to answer a question of the form:

> *"Based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 70 % on any of the coming seven days (days 0–6)?"*

The threshold X % will vary and is indicated for each prediction task.

Therefore, your prediction is either:

- **Yes** (the value will exceed the threshold on at least one of the days); or
- **No** (the value will not exceed the threshold on any of the days).

This is your initial prediction. Then, an Automated Machine Learning (AutoML) system will show its prediction, based on the same data. Then you will make your final prediction, so you have a chance to adjust your prediction and take into account the prediction of the AutoML system. Remember that neither you nor the AutoML system knows the true future values you are predicting.

This is repeated ten times. Finally, you will fill in another short questionnaire based on your experiences with the AutoML system.

Your goal in this experiment should be to make as many correct initial and final predictions as possible. For each correct prediction, you will get a bonus of $0.15, yielding a maximum bonus of $3 (with a total of 10+10=20 correct predictions).

Data

The state-level data and variables used in this experiment are described below. Note that the figure scales of the variables can change between tasks. The description can also be found in the PDF (see the link at the top of this page) with the rest of the instructions.

Description of data variables

The state-level data and variables used in this experiment are described below. Note that the figure scales of the variables can change between tasks.

Predicted variable: Adult Intensive Care Unit bed utilization

- Percentage of Intensive Care Unit (ICU) beds for adults in use in the hospitals of this state.

  "Intensive care units cater to patients with severe or life-threatening illnesses and injuries, which require constant care, close supervision from life support equipment and medication in order to ensure normal bodily functions. They are staffed by highly trained physicians, nurses and respiratory therapists who specialize in caring for critically ill patients." [1]

  Often, the ICU bed capacity is small compared to the whole capacity of the hospital. In addition, patients in need of ICU beds are in grave need of ICU care (i.e., a bed). Thus, a fully utilized ICU capacity is often not preferred, in case of sudden increased demand on ICU beds (such as due to the COVID19-pandemic).

Other variables relevant for making predictions about the adult ICU bed utilization:

- **New COVID-19 cases**
  - Number of new laboratory-confirmed COVID-19 cases in this state.

    Every COVID-19 case has some chance of resulting in a hospitalization. Every new case can therefore lead to a higher adult ICU bed utilization: the median time from the onset of COVID-19 to ICU admission is 9.5–12 days. [2]
- **Inpatient beds utilization**
  - Percentage of inpatient beds that are being utilized in the hospitals in this state.

    Inpatient beds are hospital beds (of any kind) for patients that either require a bed or will have to stay one or more nights at the hospital. Inpatient beds include ICU beds, but ICU beds usually make up only a small portion of total inpatient beds. Inpatient beds are often used to describe the capacity of a hospital, since no free inpatient beds means that a hospital cannot admit any new patients to inpatient care.

    If inpatient bed utilization is high, there may be a risk of ICU bed utilization also increasing soon: if the state of patients' (whether with COVID-19 or not) worsens, they may be transferred to ICU beds.
- **Percent of inpatients with COVID-19**
  - Percentage of inpatients (i.e., patients occupying an inpatient bed) in the hospitals of this state that have suspected or confirmed COVID-19.

    Since COVID-19 patients can develop severe symptoms, there is a risk of them being transferred to ICU beds; about 30 % of hospitalized COVID-19 patients are transferred to the ICU [2]. Therefore, if the percentage of inpatients (whether in ICU beds or not) is high or increases, there can be a higher probability of patients being transferred to ICU beds.

    In addition, the median length of hospitalization among [COVID-19] survivors is 10–13 days [2], which means COVID-19 patients may occupy beds for a relatively long period.

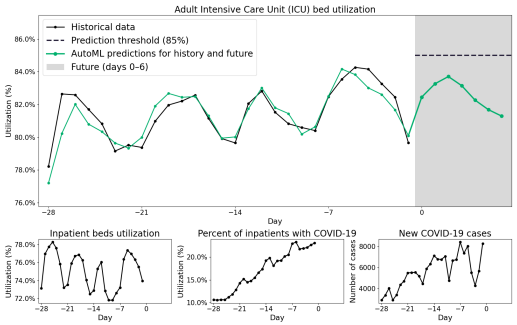
References:
[1] https://en.wikipedia.org/wiki/Intensive_care_unit
[2] https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html

Tasks 1–10

Now you will make ten predictions without and ten prediction with the help of an AutoML system, as described in the instructions. Remember, you are not allowed to go back at any point; doing so will result in your submission being invalid.

The AutoML system will make predictions for you to consider; in addition, it explains its predictions. Below is an example of an explanation, and some remarks on how one can interpret it.

Example explanation of the AutoML system



Prediction explanations: predictions for the history, i.e. "How did the system predict for earlier data?"

Only the future is truly unknown to the system. The system has been trained with data for a longer period than is shown here (going backward), so it knows the true values for the history you see (days -28 to -1); but the future (days 0–6 is not known to it. Therefore, only the predictions for days 0–6 are truly predictions of an unknown future; but since the system has access to older data, it can show how it's predictions look for the history.

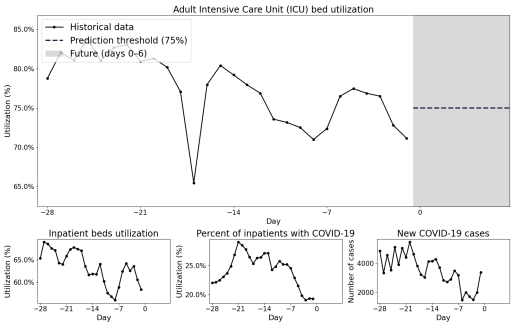
This explanation shows what the predictions look like for the previous 28 days. That is, we make the system "blind" to the values shown for days -28 to -1, and let the system make predictions for the days -28 to -1 as if they were the unknown future, using only even older data.

This way, we can judge how well the system predicts: if the predictions for the history do not make sense or seem incorrect, then the system may be making even worse predictions for the truly unknown future. On the other hand, if the predictions for the history make sense and seem correct, the system may also make reasonable predictions for the future.

In the above example, the system seems to have predicted the history very well when blinded to the true values (the green line is close to the black one, and follows its general shape). Therefore, the system seems to work well for previous data (it would have predicted well the past), so it may make reasonable predictions for the future (grey area) too.

Task 1: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



13. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

14. How confident are you in your prediction? *

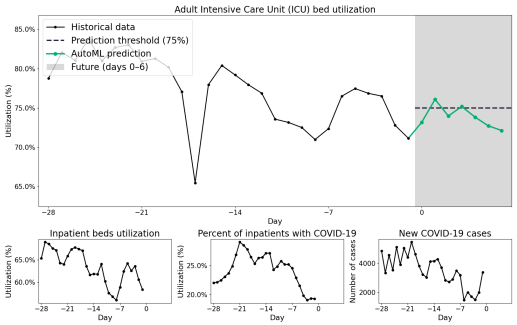
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

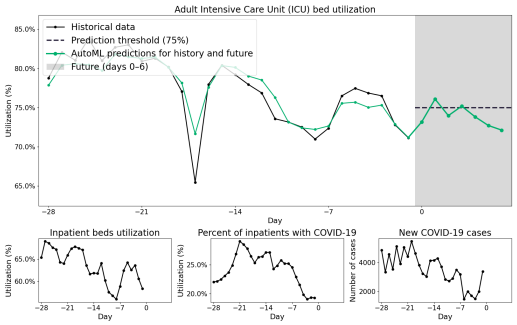
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 1: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: Yes

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL exceed the indicated threshold in the following seven days (days 0–6 in the figure).

15. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

16. How confident are you in your prediction? *

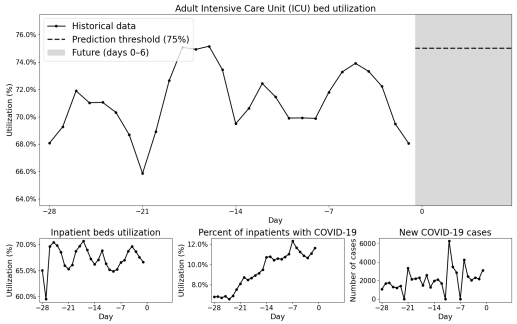
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 2: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



17. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

18. How confident are you in your prediction? *

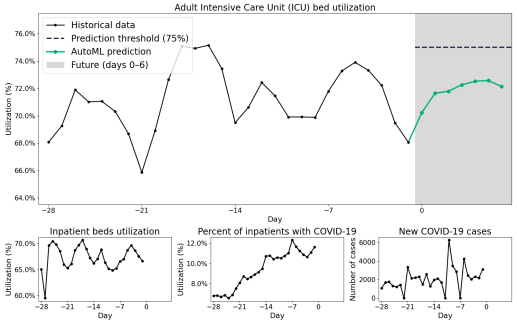
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

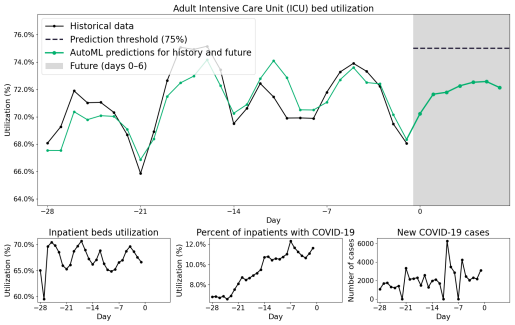
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 2: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

19. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

20. How confident are you in your prediction? *

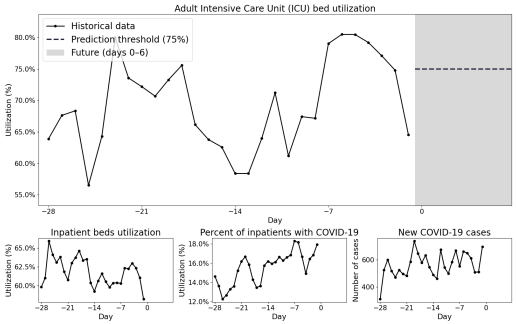
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 3: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



21. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

22. How confident are you in your prediction? *

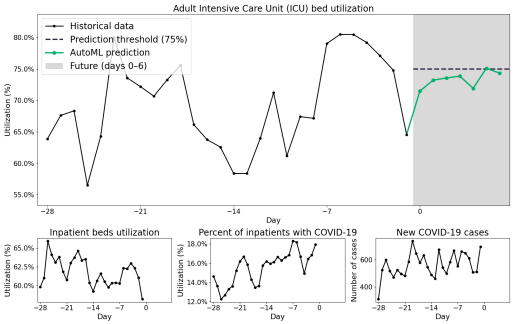
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

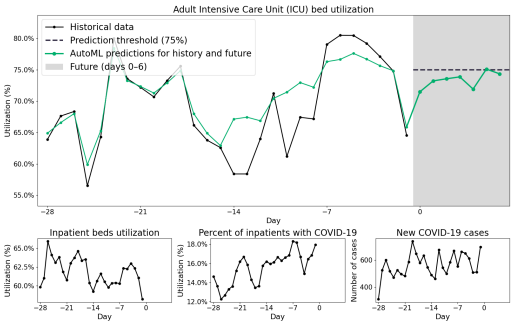
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 3: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: Yes

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL exceed the indicated threshold in the following seven days (days 0–6 in the figure).

23. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

24. How confident are you in your prediction? *

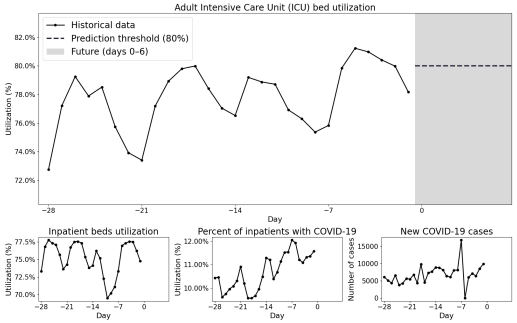
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 4: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



25. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

26. How confident are you in your prediction? *

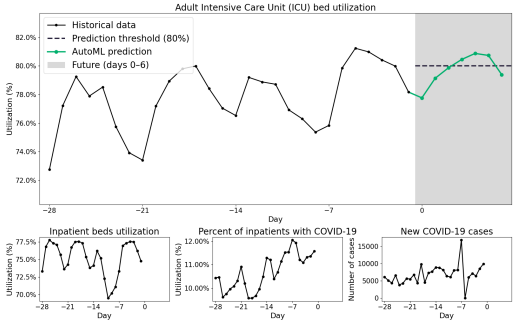
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

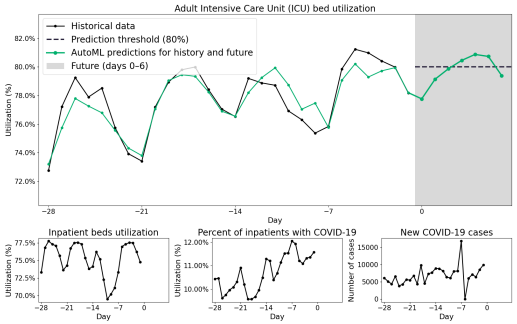
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 4: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: Yes

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL exceed the indicated threshold in the following seven days (days 0–6 in the figure).

27. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

28. How confident are you in your prediction? *

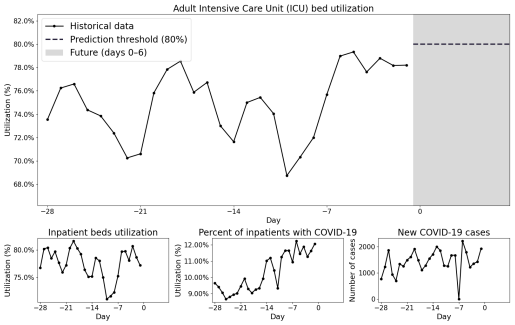
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 5: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



29. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

30. How confident are you in your prediction? *

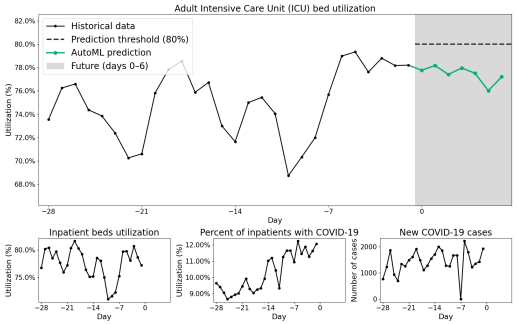
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

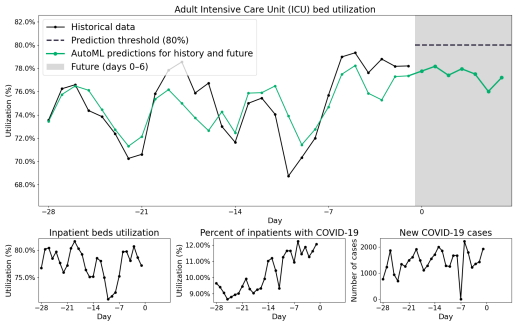
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 5: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

31. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

32. How confident are you in your prediction? *

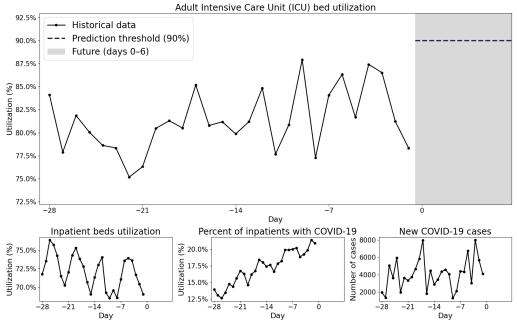
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 6: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



33. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 90 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

34. How confident are you in your prediction? *

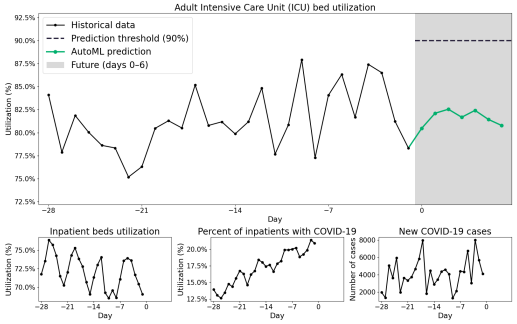
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

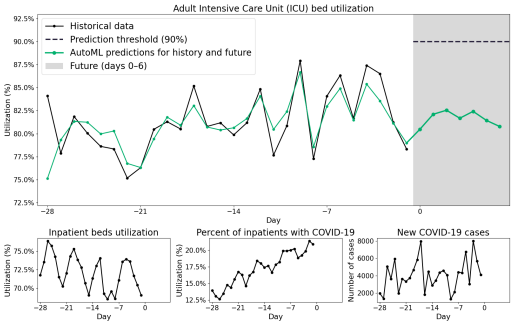
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 6: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

35. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 90 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

36. How confident are you in your prediction? *

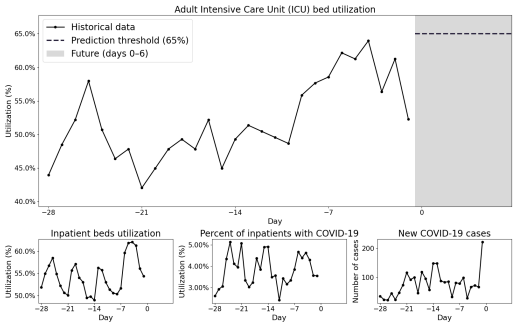
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 7: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



37. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

38. How confident are you in your prediction? *

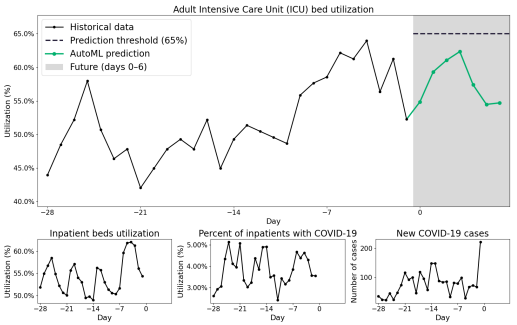
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

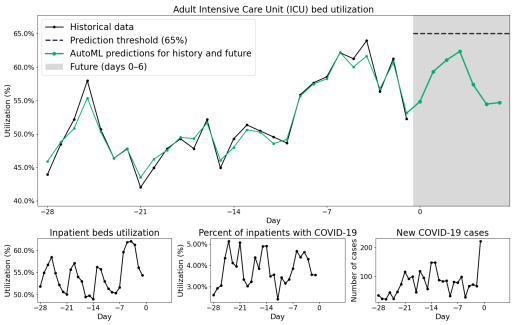
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident												Very confident

Task 7: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

39. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

40. How confident are you in your prediction? *

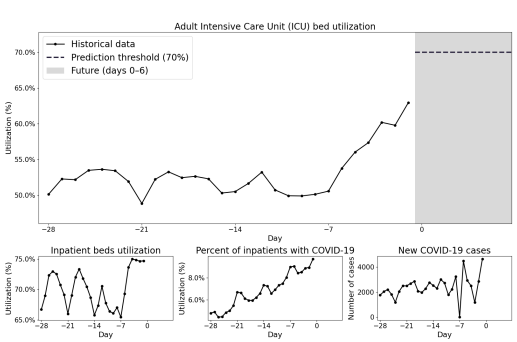
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 8: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



41. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 70 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

42. How confident are you in your prediction? *

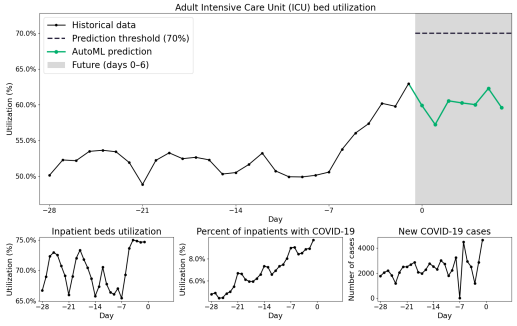
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

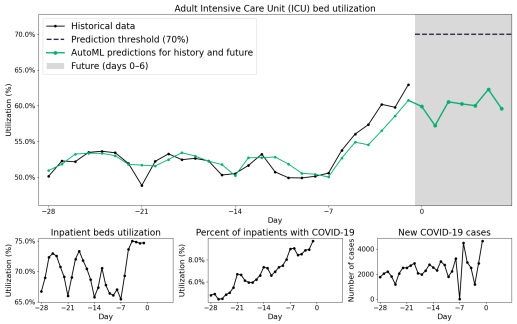
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 8: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

43. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 70 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

44. How confident are you in your prediction? *

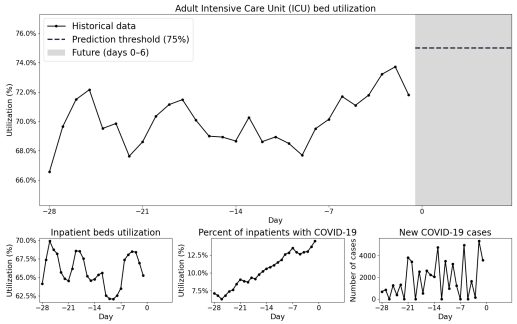
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 9: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



45. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

46. How confident are you in your prediction? *

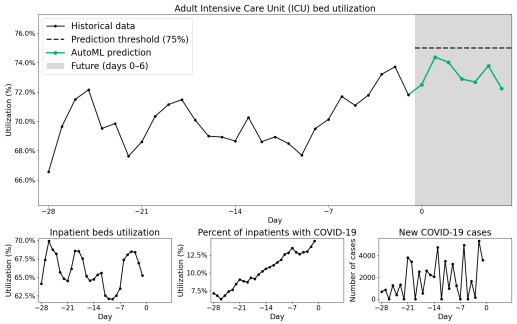
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

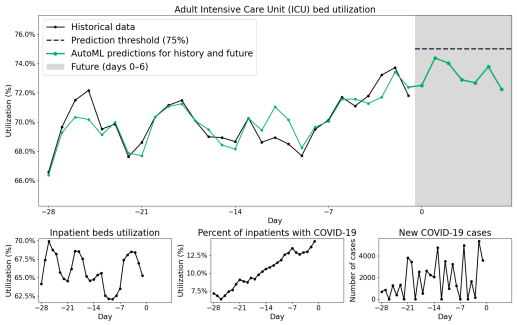
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 9: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

47. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

48. How confident are you in your prediction? *

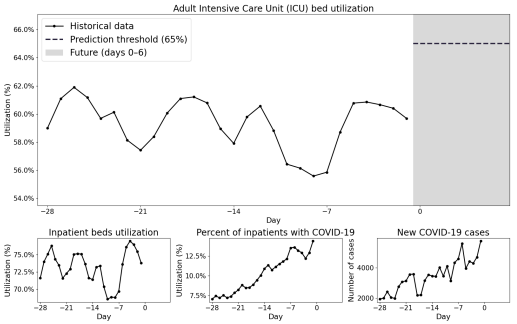
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 10: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



49. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

50. How confident are you in your prediction? *

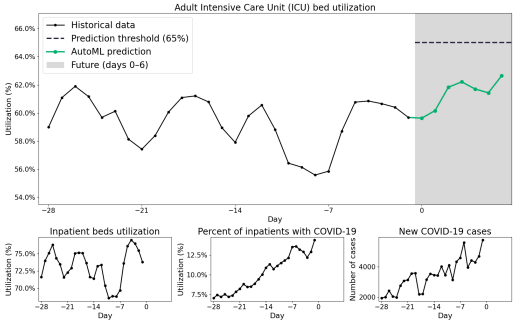
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

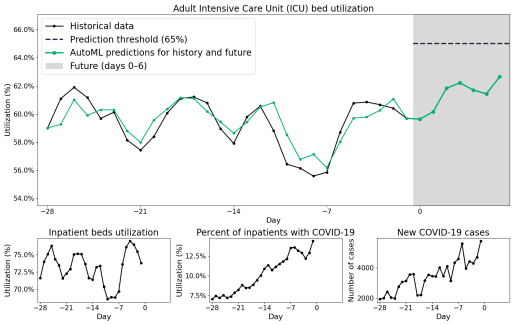
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 10: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

51. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

52. How confident are you in your prediction? *

NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Questionnaire on your experiences with the AutoML system

53. Please answer the following questions based on your experiences with the AutoML system that provided you with the predictions for the previous ten tasks. *

NOTE: there are five levels to choose from for each row, please scroll if you cannot see all alternatives.

Mark only one oval per row.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I believe the AutoML system is a competent performer.					
I trust the AutoML system.					
I have confidence in the advice given by the AutoML system.					
I can depend on the AutoML system.					
I can rely on the AutoML system to behave in consistent ways.					
I can rely on the AutoML system to do its best every time I take its advice.					

MTurk questions and information

The following information is needed for approving your task (and possible bonus) payment.

54. Please fill in your MTurk Worker ID *

This will be used to verify that you have completed the task before approving and paying. See e.g. https://blog.mturk.com/get-to-know-the-new-worker-site-4a69967d90c3 for instructions to find your WorkerID.

Feedback

55. Approximately how long did you take to complete the survey? *

Example: 4:03:32 (4 hours, 3 minutes, 32 seconds)

56. Do you have feedback on the AutoML system? *

E.g., what made them (un)trustworthy, what would make them more or less trustworthy, or what would you require of an AutoML system to trust it enough to use it as support for decision making?

57. Do you have any feedback on the experiment itself? *

E.g., was the structure of the experiment clear, were the instructions clear, would you have wanted more or less instructions, did the order of tasks make sense, ...

58. Do you have any other feedback? *

Next: submission

Next, you will submit your answers. In the survey completion message, you will be shown a debriefing of the survey. Do not close the window yet, instead read through the debriefing. At the end you will find the code you will need to fill in in your HIT on MTurk to verify that you have completed the survey and get your payment.

In addition, after submitting you can see your results, i.e. how many of your (10) initial and (10) final predictions were correct (which is the basis for any bonus to be paid to you).

A.4 Survey, group 2.3: counterfactual

AutoML survey group 2.3

This study is done as part of a Master's Thesis, in collaboration with MindsDB (mindsdb.com), and the results and the thesis will be published later this year.

In this survey, you will use the predictions of an Automated Machine Learning (AutoML) system. AutoML systems automate one or more phases of the process of creating a Machine Learning (ML) model; this way, anyone can create a complex ML model without deep knowledge of ML itself. Machine Learning (ML) describes a type of computer program that learns in a way similar to human beings. Rather than employ statistical techniques or simple mathematics to analyze a large amount of data, the ML model learns by example and is able to digest a large amount of data, and be trained to give incrementally more and more accurate predictions if it is fed more and more data.

Your task in this survey will be to make predictions about adult Intensive Care Unit (ICU) bed utilization during the COVID-19 pandemic. Your goal is to estimate the percentage of adult ICU beds occupied, given a certain scenario. You will also be shown the predictions of an AutoML system to take into account, and the AutoML system will also explain its predictions.

Your goal should be to make as many correct predictions as possible.

The survey proceeds as follows.
1. Before the tasks begin, we will ask you to fill in some necessary information about yourself and give your informed consent about us gathering and storing your anonymized answers.
2. Then you will be shown a description of the data used and how you should make your predictions. You will also be shown instructions on how to interpret the AutoML system's explanations for its predictions.
3. Then, you will make predictions in ten scenarios, with support from an AutoML system. After these predictions, you will fill in a questionnaire based on your experience with the AutoML system.

After this the study is done. You will see how your predictions compare to the actual adult ICU bed utilization. You will also have the opportunity to give feedback on the experiment and the AutoML systems used.

If you have any questions, please contact the survey organizer and thesis author Cosmo Jenytin (cosmo@mindsdb.com).

* Required

1. Informed consent: I consent to my answers being used and published anonymously. The background information asked here will be connected to your survey answers and anonymized. *

   Mark only one oval.

   ◯ Yes

Note: you are not allowed to go backwards in the survey for any reason. If you go back to a previous page, your submission will be invalid.

All the instructions can be found as a PDF here: https://drive.google.com/file/d/17_PDACdFOGPK1DhT6GxjdkptllxmDI-x/view?usp=sharing. Open this PDF in a separate window, so you have access to it throughout the experiment without going backwards in the form (which results in an invalid submission).

2. I understand that I am not allowed to go backwards (go back to previous pages) in the survey, and that doing so will result in my submission being invalid. *

   Mark only one oval.

   ◯ Yes

Background information

Please answer the following questions. Your answers will be used and stored anonymously.

3. Age *

   ____

4. I identify my gender as *

   ____

5. Education *

   Mark only one oval.

   ◯ Upper secondary
   ◯ Bachelor's or equivalent
   ◯ Master's or equivalent
   ◯ Doctoral or equivalent
   ◯ Other: __________

6. Professional field *

Mark only one oval.

- ◯ Agriculture, Food and Natural Resources
- ◯ Architecture and Construction
- ◯ Arts, Audio/Video Technology and Communications
- ◯ Business Management and Administration
- ◯ Education and Training
- ◯ Finance
- ◯ Government and Public Administration
- ◯ Health and Medicine
- ◯ Hospitality and Tourism
- ◯ Human Services
- ◯ Information Technology
- ◯ Law, Public Safety, Corrections and Security
- ◯ Manufacturing
- ◯ Marketing, Sales, and Service
- ◯ Science, Technology, Engineering, and Mathematics
- ◯ Transportation, Distribution, and Logistics
- ◯ Other: _____

7. Job title *

Mark only one oval.

- ◯ Chairperson, member of Board of Directors
- ◯ CEO, CxO
- ◯ Vice President
- ◯ Manager
- ◯ Individual Contributor
- ◯ Entry-level
- ◯ Other: _____

8. Job experience (years) *

_____

9. How confident are you, in general, in your ability to predict future developments based on data? *

NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

10. Have you previously worked with data analysis or predictions? *

This can, for example, mean working with simple data analysis tools (e.g., Excel) or more advanced tools, such as Machine Learning (ML) models.

Mark only one oval.

- ◯ Yes
- ◯ No

11. Have you previously used an Automated Machine Learning (AutoML) system? *

An AutoML system is a Machine Learning system that can automate one or more phases of a data science or Machine Learning model development process.

Mark only one oval.

- ◯ Yes
- ◯ No

12. Please answer the following questions on automated agents. *

An automated agent runs by computerized algorithms and interacts with humans. For example, a website predicting a medical diagnosis based on symptoms is an automated agent; likewise, an ATM is an automated agent. NOTE: there are five levels to choose from, please scroll if you cannot see all alternatives.

Mark only one oval per row.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Generally, I trust automated agents.	◯	◯	◯	◯	◯
Automated agents help me solve many problems.	◯	◯	◯	◯	◯
I think it's a good idea to rely on automated agents for help.	◯	◯	◯	◯	◯
I don't trust the information I get from automated agents.	◯	◯	◯	◯	◯
Automated agents are reliable.	◯	◯	◯	◯	◯
I rely on automated agents.	◯	◯	◯	◯	◯

Tasks, tools, and data

These instructions can be found as a PDF here: https://drive.google.com/file/d/17_PDACdFOGPK1DhT6GxjdkptIIxmDI-x/view?usp=sharing. Open this PDF in a separate window, so you have access to it throughout the experiment without going backwards in the form (which results in an invalid submission).

Tasks and tools

Your task is to make predictions of adult Intensive Care Unit (ICU) bed utilization in a state in the USA. Predictions like these are relevant for real decision making during the COVID19-pandemic: for example, in New York "regions must have at least 30 percent of their ICU beds available before a phased re-opening can begin" (i.e., the ICU bed utilization must be below 70 %). (https://forward.ny.gov/metrics-guide-reopening-new-york)

For each prediction you will see data for the previous 28 days on ICU bed utilization and other variables to consider (days -28 to -1). Based on these, you are to make a prediction to answer a question of the form:

> *"Based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 70 % on any of the coming seven days (days 0–6)?"*

The threshold X % will vary and is indicated for each prediction task.

Therefore, your prediction is either:

- **Yes** (the value will exceed the threshold on at least one of the days); or
- **No** (the value will not exceed the threshold on any of the days).

This is your initial prediction. Then, an Automated Machine Learning (AutoML) system will show its prediction, based on the same data. Then you will make your final prediction, so you have a chance to adjust your prediction and take into account the prediction of the AutoML system. Remember that neither you nor the AutoML system knows the true future values you are predicting.

This is repeated ten times. Finally, you will fill in another short questionnaire based on your experiences with the AutoML system.

Your goal in this experiment should be to make as many correct initial and final predictions as possible. For each correct prediction, you will get a bonus of $0.15, yielding a maximum bonus of $3 (with a total of 10+10=20 correct predictions).

Data

The state-level data and variables used in this experiment are described below. Note that the figure scales of the variables can change between tasks. The description can also be found in the PDF (see the link at the top of this page) with the rest of the instructions.

Description of data variables

The state-level data and variables used in this experiment are described below. Note that the figure scales of the variables can change between tasks.

Predicted variable: Adult Intensive Care Unit bed utilization

- Percentage of Intensive Care Unit (ICU) beds for adults in use in the hospitals of this state.

  "Intensive care units cater to patients with severe or life-threatening illnesses and injuries, which require constant care, close supervision from life support equipment and medication in order to ensure normal bodily functions. They are staffed by highly trained physicians, nurses and respiratory therapists who specialize in caring for critically ill patients." [1]

  Often, the ICU bed capacity is small compared to the whole capacity of the hospital. In addition, patients in need of ICU beds are in grave need of ICU care (i.e., a bed). Thus, a fully utilized ICU capacity is often not preferred, in case of sudden increased demand on ICU beds (such as due to the COVID19-pandemic).

Other variables relevant for making predictions about the adult ICU bed utilization:

- **New COVID-19 cases**
  - Number of new laboratory-confirmed COVID-19 cases in this state.

    Every COVID-19 case has some chance of resulting in a hospitalization. Every new case can therefore lead to a higher adult ICU bed utilization: the median time from the onset of COVID-19 to ICU admission is 9.5–12 days. [2]
- **Inpatient beds utilization**
  - Percentage of inpatient beds that are being utilized in the hospitals in this state.

    Inpatient beds are hospital beds (of any kind) for patients that either require a bed or will have to stay one or more nights at the hospital. Inpatient beds include ICU beds, but ICU beds usually make up only a small portion of total inpatient beds. Inpatient beds are often used to describe the capacity of a hospital, since no free inpatient beds means that a hospital cannot admit any new patients to inpatient care.

    If inpatient bed utilization is high, there may be a risk of ICU bed utilization also increasing soon: if the state of patients' (whether with COVID-19 or not) worsens, they may be transferred to ICU beds.
- **Percent of inpatients with COVID-19**
  - Percentage of inpatients (i.e., patients occupying an inpatient bed) in the hospitals of this state that have suspected or confirmed COVID-19.

    Since COVID-19 patients can develop severe symptoms, there is a risk of them being transferred to ICU beds; about 30 % of hospitalized COVID-19 patients are transferred to the ICU [2]. Therefore, if the percentage of inpatients (whether in ICU beds or not) is high or increases, there can be a higher probability of patients being transferred to ICU beds.

    In addition, the median length of hospitalization among [COVID-19] survivors is 10–13 days [2], which means COVID-19 patients may occupy beds for a relatively long period.

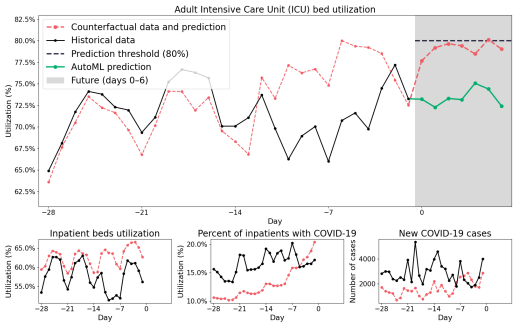
References:
[1] https://en.wikipedia.org/wiki/Intensive_care_unit
[2] https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-guidance-management-patients.html

Tasks 1–10

Now you will make ten predictions without and ten prediction with the help of an AutoML system, as described in the instructions. Remember, you are not allowed to go back at any point; doing so will result in your submission being invalid.

The AutoML system will make predictions for you to consider; in addition, it explains its predictions. Below is an example of an explanation, and some remarks on how one can interpret it.

Example explanation of the AutoML system



Prediction explanations: counterfactual explanations, i.e. "How would the data have to change to get the system to make the opposite prediction?"

Counterfactual explanations explain through opposites: in order to understand why the system made a specific prediction, it shows instead how the data would need to change to make it change its prediction; that is, it shows the answer to the question: "What is the smallest change to the data that is required for the system to make the opposite prediction (yes instead of no, no instead of yes)?"

This means that the system looks for data that is as close to the original one (i.e., the different variables' values are close), but that gets the opposite prediction (i.e., if the system made the prediction "Yes", it looks for data that gets the prediction "No", and vice versa).

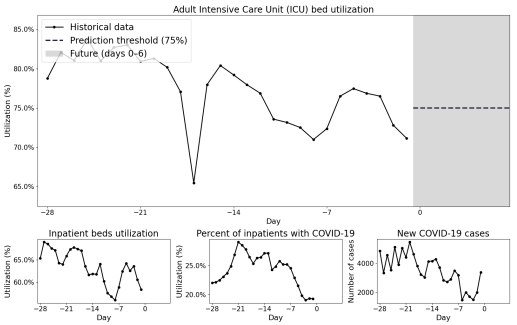
This kind of explanation can help us judge how well the system works in this specific case. If the required change in the data is large or requires some unfeasible changes, then the prediction is probably quite good. On the other hand, if even a slight change in the data gives an opposite prediction, we should be careful, since even a small change can result in the opposite prediction – and all models have some error.

We can also learn about how the model predicts in general. If it seems that changes in a specific variable or at a specific time is the only major change that makes the prediction significantly different, then that specific variable or time may have a (relatively) large influence on the predictions.

In the above example, significant changes are required to make the system give the opposite prediction ("Yes, the adult ICU bed utilization will exceed the threshold"). The changes required are also in multiple variables, and they seem to make sense: increasing the adult ICU bed utilization in the near past and having an increasing trend, in addition to the inpatient beds utilization being higher and the percent of inpatients with COVID-19 having a clear increasing trend and larger values in the last few days all seem to support a higher predicted adult ICU bed utilization in the coming seven days. Therefore, the system's prediction seems to make sense and the data would need to change significantly to get the opposite prediction, so the system's prediction is probably reasonable.

Task 1: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



13. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

14. How confident are you in your prediction? *

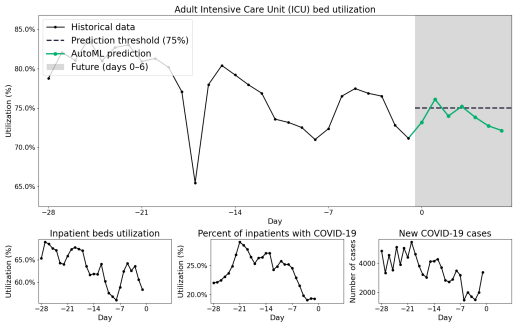
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

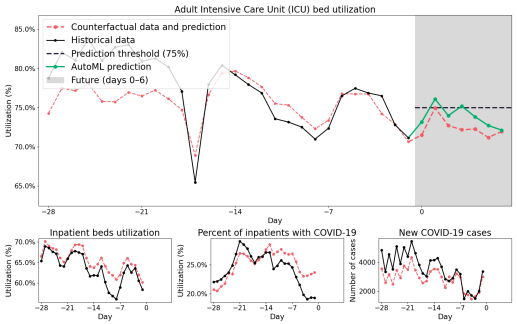
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 1: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: Yes

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL exceed the indicated threshold in the following seven days (days 0–6 in the figure).

15. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

16. How confident are you in your prediction? *

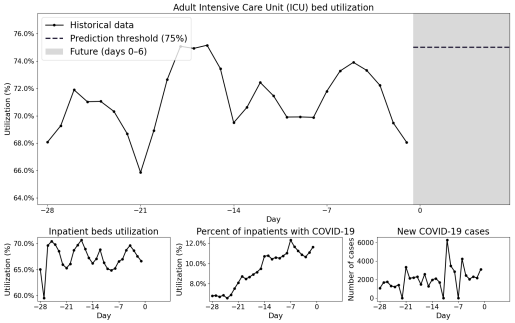
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 2: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



17. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

18. How confident are you in your prediction? *

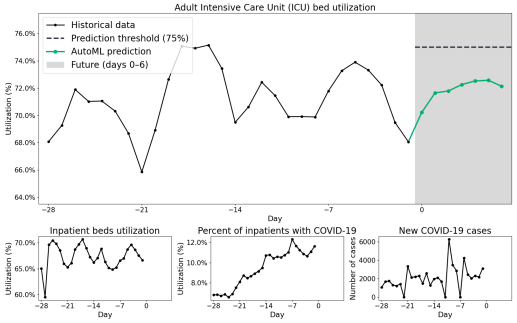
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

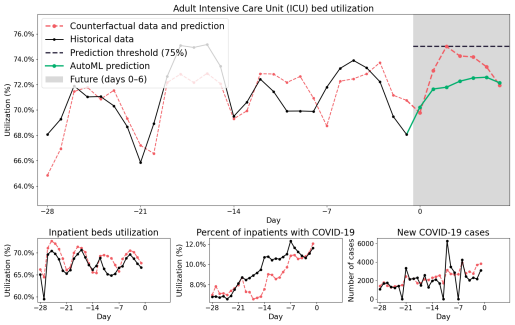
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 2: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

19. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

20. How confident are you in your prediction? *

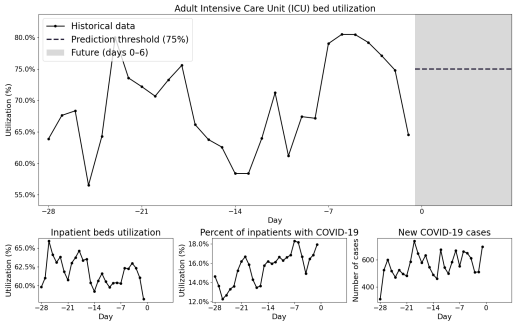
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 3: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



21. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

22. How confident are you in your prediction? *

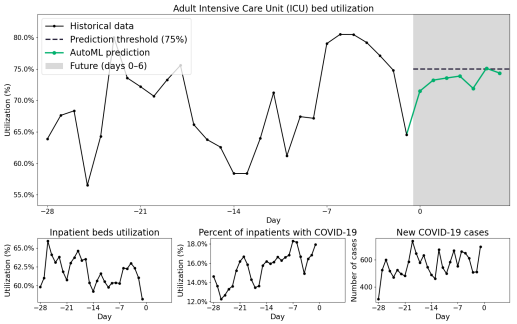
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

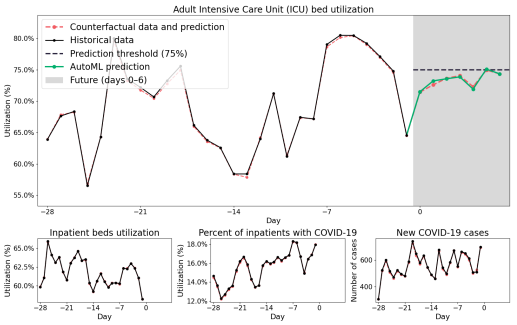
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident												Very confident

Task 3: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: Yes

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL exceed the indicated threshold in the following seven days (days 0–6 in the figure).

23. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

24. How confident are you in your prediction? *

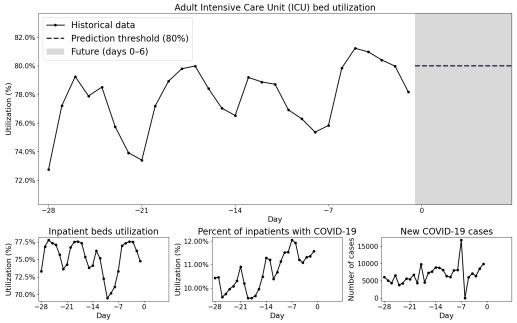
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 4: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



25. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

26. How confident are you in your prediction? *

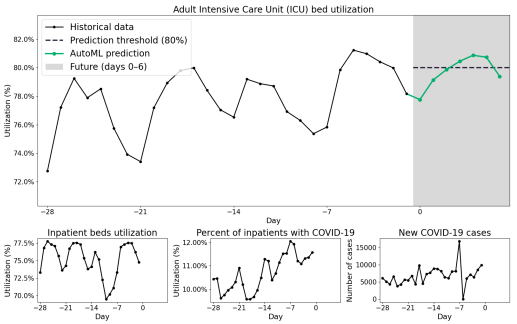
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

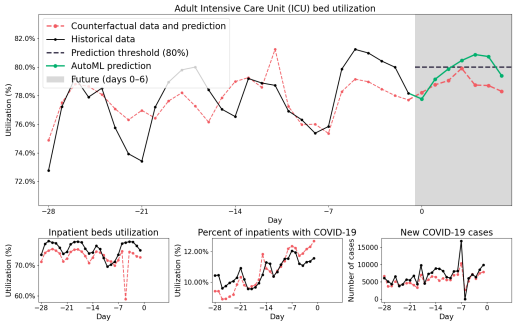
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 4: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: Yes

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL exceed the indicated threshold in the following seven days (days 0–6 in the figure).

27. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

28. How confident are you in your prediction? *

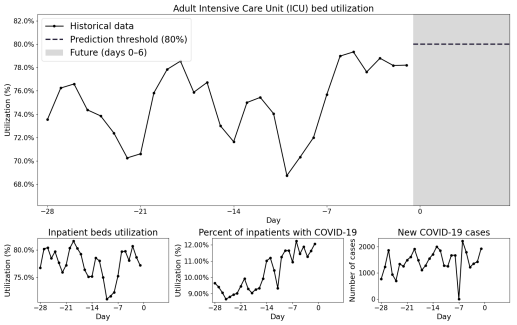
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 5: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



29. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

30. How confident are you in your prediction? *

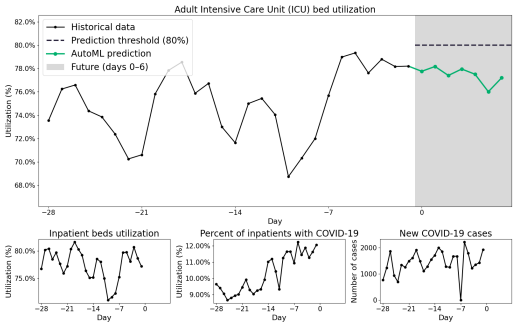
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

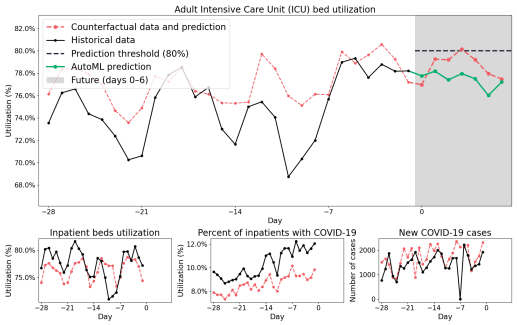
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 5: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

31. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 80 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

32. How confident are you in your prediction? *

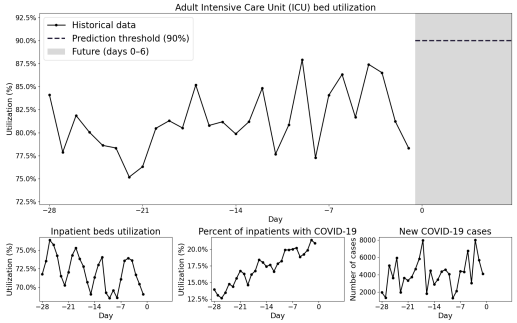
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 6: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



33. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 90 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

34. How confident are you in your prediction? *

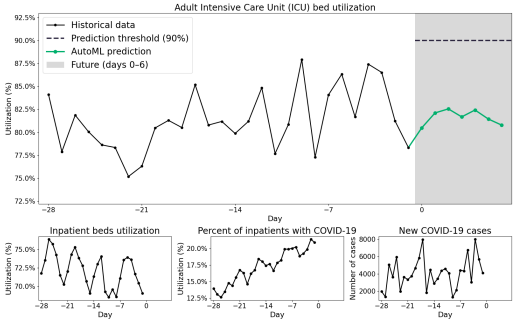
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

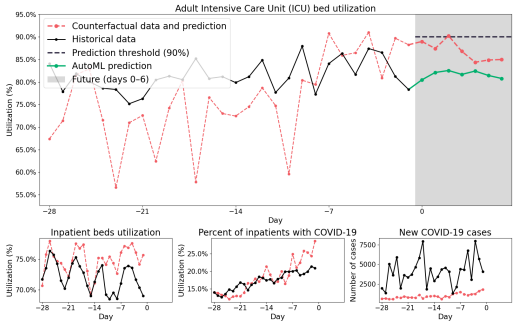
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 6: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

35. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 90 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

36. How confident are you in your prediction? *

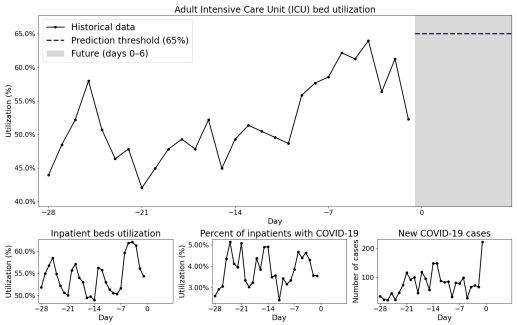
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 7: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



37. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

38. How confident are you in your prediction? *

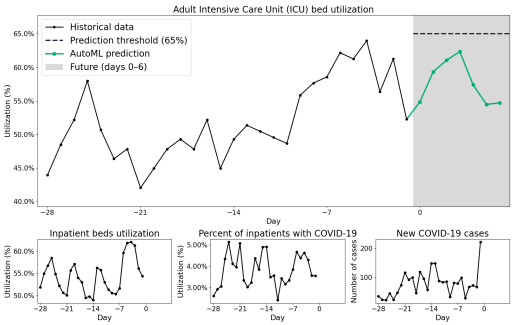
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

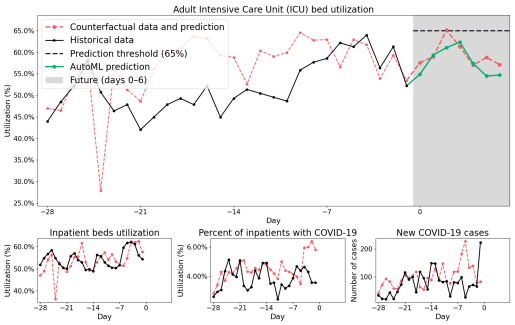
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 7: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

39. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

40. How confident are you in your prediction? *

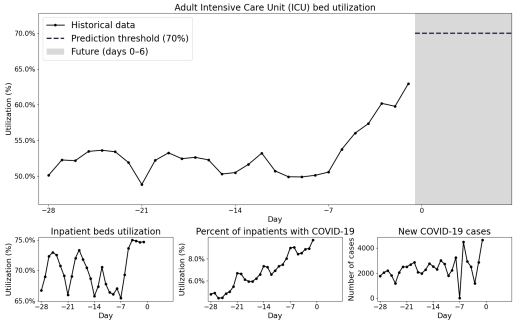
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 8: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



41. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 70 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- Yes
- No

42. How confident are you in your prediction? *

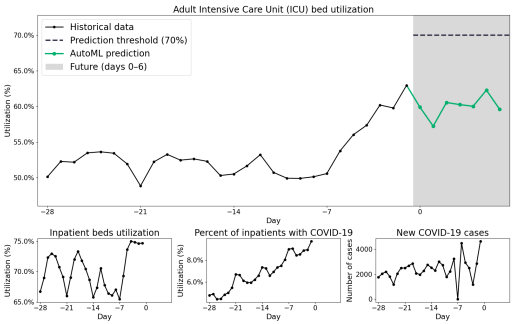
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

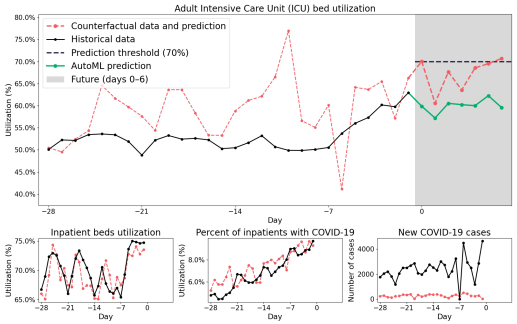
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 8: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

43. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 70 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

44. How confident are you in your prediction? *

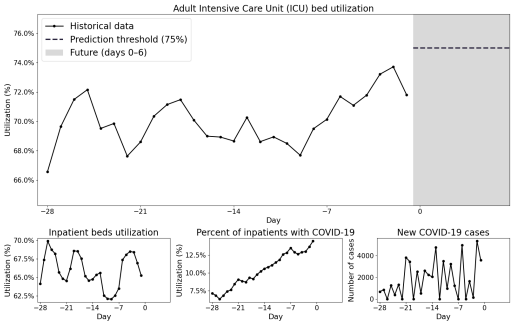
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 9: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



45. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

46. How confident are you in your prediction? *

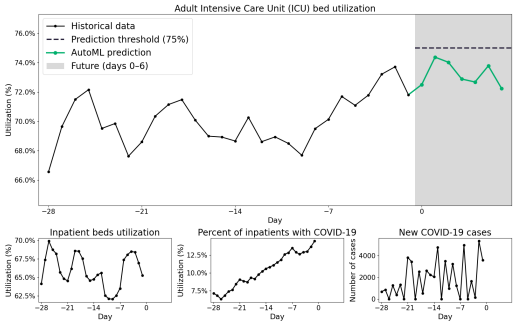
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

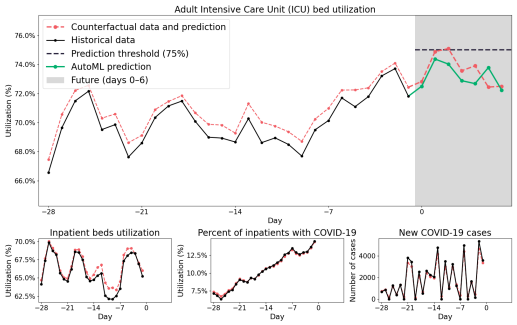
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 9: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

47. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 75 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

48. How confident are you in your prediction? *

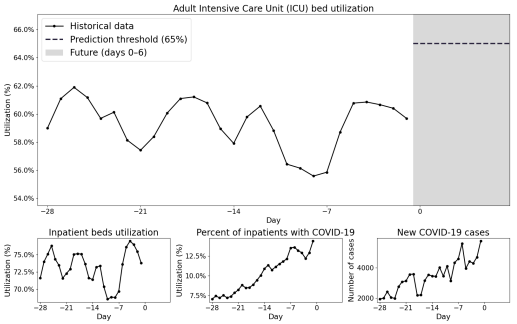
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 10: data and your initial prediction

Data for the previous 28 days (days -28 to -1)



49. Your initial prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

50. How confident are you in your prediction? *

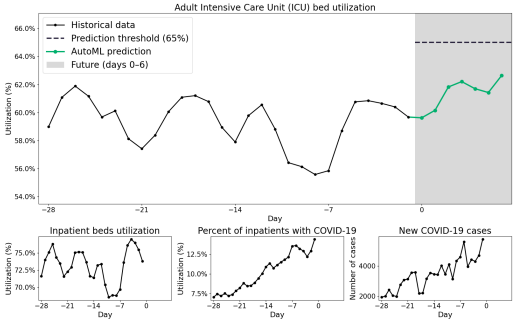
NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

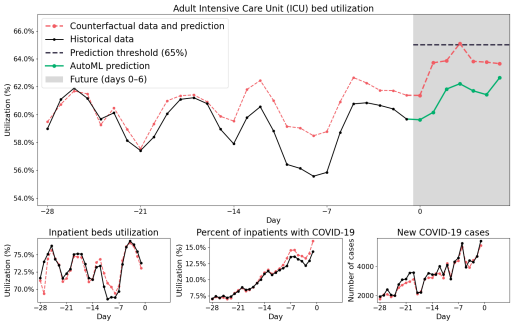
	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Task 10: AutoML prediction and your final prediction

Data for the previous 28 days (days -28 to -1) and the prediction of the AutoML system (days 0 to 6)



Data for the previous 28 days (days -28 to -1), the prediction of the AutoML system (days 0 to 6), and the AutoML system's explanation



AutoML prediction: No

Based on the prediction of the AutoML system, shown in green in the figure above, the adult ICU bed utilization WILL NOT exceed the indicated threshold in the following seven days (days 0–6 in the figure).

51. Considering your own views and the prediction of the AutoML system, what is your final prediction: based on the previous 28 days' data, will the adult ICU bed utilization (percent of adult ICU beds in use) exceed 65 % on any of the coming seven days (days 0–6 in the figure)? * 1 point

Mark only one oval.

- ◯ Yes
- ◯ No

52. How confident are you in your prediction? *

NOTE: there are 11 levels to choose from (0–10), please scroll if you cannot see all alternatives.

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all confident	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Very confident

Questionnaire on your experiences with the AutoML system

53. Please answer the following questions based on your experiences with the AutoML system that provided you with the predictions for the previous ten tasks. *

NOTE: there are five levels to choose from for each row, please scroll if you cannot see all alternatives.

Mark only one oval per row.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I believe the AutoML system is a competent performer.					
I trust the AutoML system.					
I have confidence in the advice given by the AutoML system.					
I can depend on the AutoML system.					
I can rely on the AutoML system to behave in consistent ways.					
I can rely on the AutoML system to do its best every time I take its advice.					

MTurk questions and information

The following information is needed for approving your task (and possible bonus) payment.

54. Please fill in your MTurk Worker ID *

This will be used to verify that you have completed the task before approving and paying. See e.g. https://blog.mturk.com/get-to-know-the-new-worker-site-4a69967d90c3 for instructions to find your WorkerID.

Feedback

55. Approximately how long did you take to complete the survey? *

Example: 4:03:32 (4 hours, 3 minutes, 32 seconds)

56. Do you have feedback on the AutoML system? *

E.g., what made them (un)trustworthy, what would make them more or less trustworthy, or what would you require of an AutoML system to trust it enough to use it as support for decision making?

57. Do you have any feedback on the experiment itself? *

E.g., was the structure of the experiment clear, were the instructions clear, would you have wanted more or less instructions, did the order of tasks make sense, ...

58. Do you have any other feedback? *

Next: submission

Next, you will submit your answers. In the survey completion message, you will be shown a debriefing of the survey. Do not close the window yet, instead read through the debriefing. At the end you will find the code you will need to fill in in your HIT on MTurk to verify that you have completed the survey and get your payment.

In addition, after submitting you can see your results, i.e. how many of your (10) initial and (10) final predictions were correct (which is the basis for any bonus to be paid to you).