Bayesian model validation and selection metrics in retail time series forecasting

Leevi Rönty

School of Science

Bachelor's thesis
Espoo 26.9.2021

Supervisor

Prof. Fabricio Oliveira

Advisors

DSc (Tech.) Mikko Ervasti

DSc (Tech.) Paavo Niskala

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the bachelor's thesis

Author Leevi Rönty		
Title Bayesian model validation and selection metrics in retail time series forecasting		
Degree programme Engineering Physics and Mathematics		
Major Mathematics and Systems Sciences		**Code of major** SCI3029
Teacher in charge Prof. Fabricio Oliveira		
Advisors DSc (Tech.) Mikko Ervasti, DSc (Tech.) Paavo Niskala		
Date 26.9.2021	**Number of pages** 21+2	**Language** English

Abstract

Sales modelling is commonplace in modern retail business. The modelling process however is not reliable if model validation and selection is not done properly. Typically, cross validation (CV) is considered a robust validation method, but in practice it can not always be done. Lack of data can prevent splitting it to training and testing sets of sufficient sizes. The speed of cross validation can become a problem if the initial model fitting is slow.

In this thesis we studied known model validation and selection metrics. The chosen metrics were Akaike information criterion (AIC), deviance information criterion (DIC), Watanabe-Akaike information criterion (WAIC), mean absolute percentage error (MAPE), and 10-fold-CV. We modelled weekly retail sales with Bayesian models. The models were ranked in order of their relative goodness of fit using the selected metrics. The purpose of this thesis is to find the metrics most useful in a business setting.

The thesis showed that information criteria and 10-fold-CV, the metrics based on expected log pointwise predictive density, ranked the models similarly. MAPE is based on a different loss function and thus it sometimes did not agree with the ordering by other metrics. Even though the information criteria did not use out-of-sample data, they did end up with similar ordering of the models as computationally more complex metrics. Information criteria thus offer an attractive alternative for out-of-sample data based metrics.

Keywords Bayesian models, model validation, information criterion, Facebook Prophet

Aalto University
School of Science

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Tekniikan kandidaatintyön tiivistelmä

Tekijä Leevi Rönty

Työn nimi Bayesilaisten mallien validointi vähittäismyynnin aineistoissa

Koulutusohjelma Teknillinen fysiikka ja matematiikka

Pääaine Matematiikka ja systeemitieteet **Pääaineen koodi** SCI3029

Vastuuopettaja Prof. Fabricio Oliveira

Työn ohjaajat TkT Mikko Ervasti, TkT Paavo Niskala

Päivämäärä 26.9.2021 **Sivumäärä** 21+2 **Kieli** Englanti

Tiivistelmä

Myynnin mallinnus on arkipäivää moderneissa vähittäismyymäläketjuissa. Mallien valinta ja validointi ovat kuitenkin tärkeä osa mallinnusprosessia, sillä mallinnuksen tulokset eivät ole uskottavia ilman niitä. Ristiinvalidointia pidetään yleisesti hyvänä validointitapana, mutta käytännössä ristiinvalidointia (engl. cross-validation, CV) ei aina voida tehdä. Datan vähyys voi estää datan jakamisen testi- ja opetusjoukkoon. Mallin sovittaminen voi olla hidasta, jolloin ristiinvalidoinnin nopeus muodostuu käytännön ongelmaksi.

Tässä kandidaatintyössä tutkittiin tunnettuja mallin valinta- ja validointimetriikoita. Tutkitut metriikat olivat Akaiken-, devianssi- ja Watanabe-Akaike -informaatiokriteeri (AIC, DIC ja WAIC), keskimääräinen absoluuttinen virheprosentti (engl. mean absolute percentage error, MAPE) sekä 10-kertainen ristiinvalidointi (10-fold-CV). Työssä mallinnettiin päivittäistavaraketjun myyntiä Bayesilaisilla malleilla, joiden paremmuutta vertailtiin keskenään edellä mainituilla metriikoilla. Työn tarkoituksena on löytää kaupallisessa viitekehyksessä käyttökelpoisimmat metriikat.

Työ osoitti, että odotettuun logistiseen piste-ennusteen uskottavuuteen (engl. expected log pointwise predictive density) perustuvat menetelmät, eli informaatiokriteerit ja 10-fold-CV, yleisesti ottaen järjestävät mallit samaan paremmuusjärjestykseen. MAPE perustuu erilaiseen sakkofunktioon ja näin ollen se myös toisinaan päätyy erilaiseen paremmuusjärjestykseen kuin muut metriikat. Vaikka informaatiokriteerit eivät käytäkään opetusjoukon ulkopuolista dataa, päätyivät ne samoihin tuloksiin laskennallisesti raskaampien menetelmien kanssa. Informaatiokriteerit siis tarjoavat käytännöllisiä vaihtoehtoja testijoukkoon perustuville metriikoille.

Avainsanat Bayesilaiset mallit, mallin validointi, informaatiokriteeri, Facebook Prophet

Contents

1 Introduction

Modern businesses collect data and use it in analytics and predictions. Statistical modelling can be applied in multiple retail business areas, such as supply chain planning (appropriate amount of goods on the shelves), assortment (what to have on the shelves), pricing, media planning, promotions etc. Davenport et al. (2006); Chan and Perry (2017); Kök and Fisher (2007). Correctly forecasting demand for goods is important for any retail business, as shortages can cause loss of sales. As companies collect increasingly more data, the possibilities of forecastable subjects and possible features increase dramatically. However, one cannot just include more features in the model and expect it to perform well. Furthermore, not all models are suitable for all forecasting tasks. Virtually unlimited amount of models can be formulated, but most of them will perform poorly. To find the useful ones, we must be able to compare the goodness of a model. It is typically defined as expected predictive accuracy, accodring to Gelman et al. (2014). Being able to objectively compare models allows for systematic approaches of model selection to be adopted.

According to Landry et al. (1983), model validation can not be separated from the modelling process. Gelman et al. (2013) describe model validation as an integral part of a robust modelling framework. Bayesian models are a relatively novel model type, which has gained considerable popularity in recent years, as computational power of modern computers has kept rising. As a relatively new method, not too many validation metrics have been proposed with these kinds of models in mind.

Gelman et al. (2014) describe the current state of model validation for Bayesian models to be unsatisfying. Information criteria that try to predict out-of-sample fitness can have strong biases, and some may not take into consideration the nature of Bayesian models and distributions of parameters. Cross validation can be extremely computationally intensive. Additionally, most information criteria are based on a loss function proportional to the root mean square of errors, but that may not always be the most useful function to optimise for. In some businesses, the consequences of errors in forecasts can be described better with the mean absolute percentage error. All in all, there does not exist a universally optimal method for solving all Bayesian model validation problems.

This thesis attempts to study the usefulness of popular model selection metrics applied in business setting with retail data. How can the metrics be applied when the number of past observations is very limited? Does it significantly affect the performance or applicability of some metrics? Retail data can also be very detailed by having a separate time series for each product. Low-level modelling can lead to a very high count of fitted models or exceedingly complex ones. How computationally complex are the evaluation methods? This can significantly impact the usefulness of the metric. As the metrics are applied to Bayesian models, there is always some randomness involved as the sampling process is not deterministic. Will this play a significant role in the stability of the metrics or will the results be always consistent? By answering these questions, we try to conclude which metrics are useful in which circumstances in business applications.

In this thesis, we will use the dataset by Singh (2017) to model sales time series of

Walmart stores in the United States. We will use the Facebook's Prophet modelling framework as described by Taylor and Letham (2018). It consists of a flexible Bayesian model which can be customised to fit a wide range of possible time series modelling tasks. We will implement few popular information criteria metrics for the model and compare them to other methods of model validation.

Writing this thesis was part of a summer internship at Sellfote Solutions Ltd. The company uses Bayesian models in marketing mix modelling.

2 Background

Models can be viewed as simplifications of reality. This means that no model never really matches the true data generating processes, but if one can formulate a model that works sufficiently well, it might be possible to draw some conclusions from them. Most of the models we are concerned with link together some explanatory variables to the measured data. The parameters of the model are learned by fitting the data to the model. The fitted model can then be used to predict future observations, assuming that the explanatory variable values are known for the future data points. The sampled model parameters can likewise give insight on the behaviour of the physical world, assuming that said parameters have a sensible interpretation.

As described by, e.g., Bayarri and Berger (2004), Bayesian models differ from frequentist models in two major ways. Firstly, in Bayesian thinking, parameters are not expressed as point values, but as distributions. Secondly, the posterior distributions are affected also by the prior beliefs of the distribution. This means that the data is not the only source of information affecting the results, as the posterior beliefs are actually prior beliefs updated by Bayesian inference on the new data. These qualities allow for embedding uncertainty and business knowledge in the models. Highly informative priors can, in addition, make the models behave more robustly in case the data has some outliers.

Model selection methods can be categorised by the view they take on whether the considered models contain the actual data generating model. This data generating model would represent the actual real-world process by which some events result in the observed data. As proposed by Vehtari and Ojanen (2012), the M-closed view assumes that the data generating model is present in the considered models. M-open view instead attempts to model the data with minimal assumptions and is thus more applicable with real-world data.

Vehtari and Ojanen (2012) describe multiple approaches for model selection. The most often used methods are information criteria, hold-out predictive, and cross validation. None of the most popular methods make explicit assumptions of the "true" model and are thus rather simple to implement. The methods differ on how they reuse data when evaluating model performance. Information criteria are based on calculating the training utility of the model and adding a bias term that attempts to correct for the reuse of training data in model evaluation to prevent overfitting to the available data. Hold-out predictive methods separate the data to a training and testing set. The testing set is used to evaluate the utility function for the model that has been trained using the training data. The method is considered to be robust if enough training and testing data is available. Cross validation takes this hold-out approach a step further by splitting the dataset and training the model multiple times. This allows for all of the data to be used in testing once. If the amount of available past observations is limited this can help with robustness of evaluation.

3 Datasets and methods

3.1 Original data and feature engineering

The original retail dataset by Singh (2017) consists of department-level weekly sales data of 45 stores. The data spans a little over two years, from 20 February 2010 to 26 October 2012. Each store location is associated with features that may help with sales prediction. The features are temperature, unemployment, price of gas, and consumer price index CPI. Furthermore, the store size and type of store is known.

The data contains information about markdowns, but this feature is available only from 11 November 2011 onwards. This renders the feature useless, as the effect of the missing markdowns will not be attributed correctly, skewing the results. For example, a spike in sales could be due to markdown campaign or seasonality. If the markdown data would be complete for the whole dataset the effect could be separated from seasonality. However, as there is markdown data for less than two years, some spikes will not have data about possible markdowns. This leads to the uplift to be attributed to seasonality, which may not be correct. Leaving markdowns out of the features will effectively cause the uplift from markdowns to be attributed to noise if the markdown campaigns are not seasonal.

The original dataset contains sales data on a department level. There are a total of 81 departments, but not all are present in every store. We are still left with 3331 sales time series. Aggregated data was used to reduce computational complexity. The department-level sales are first aggregated to store-level sales by summing sales in each department in each store. This still leaves us with 45 time series, which would still require too much time to model each individually. To simplify this issue, the stores are further aggregated based on store type (A, B or C) as classified in the data. The aggregated sales are presented in Fig. 1. The features for these store groups are obtained by calculating a size-weighted average. Store size is given in the dataset. As the aggregation process of the features can render the features less meaningful, a small subset of stores is also selected from each store group. Comparing these stores to the store groups will reveal if the aggregation lessens the predictive performance of the models using the provided features. The aggregated features can be found in Fig. 2.

The sales data for store groups A and B demonstrate strong seasonality. Especially during Decembers, the turnover spikes due to strong holiday effect. For store group C the sales are more consistent throughout the year. Temperature is the only feature that clearly behaves seasonally. Originally, unemployment was expressed to equal the latest unemployment statistic. This made the feature behave rather discontinuously. To get around the discontinuity affecting the results, the unemployment for each

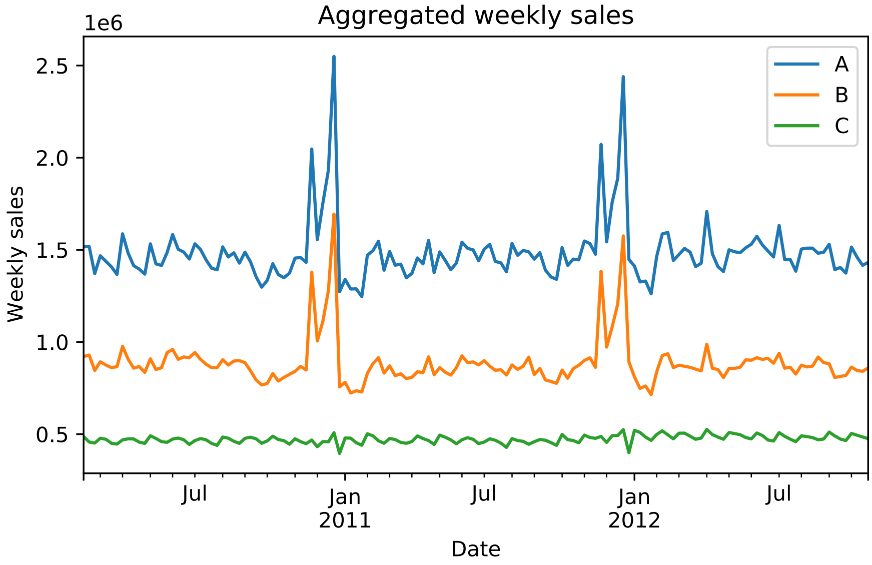


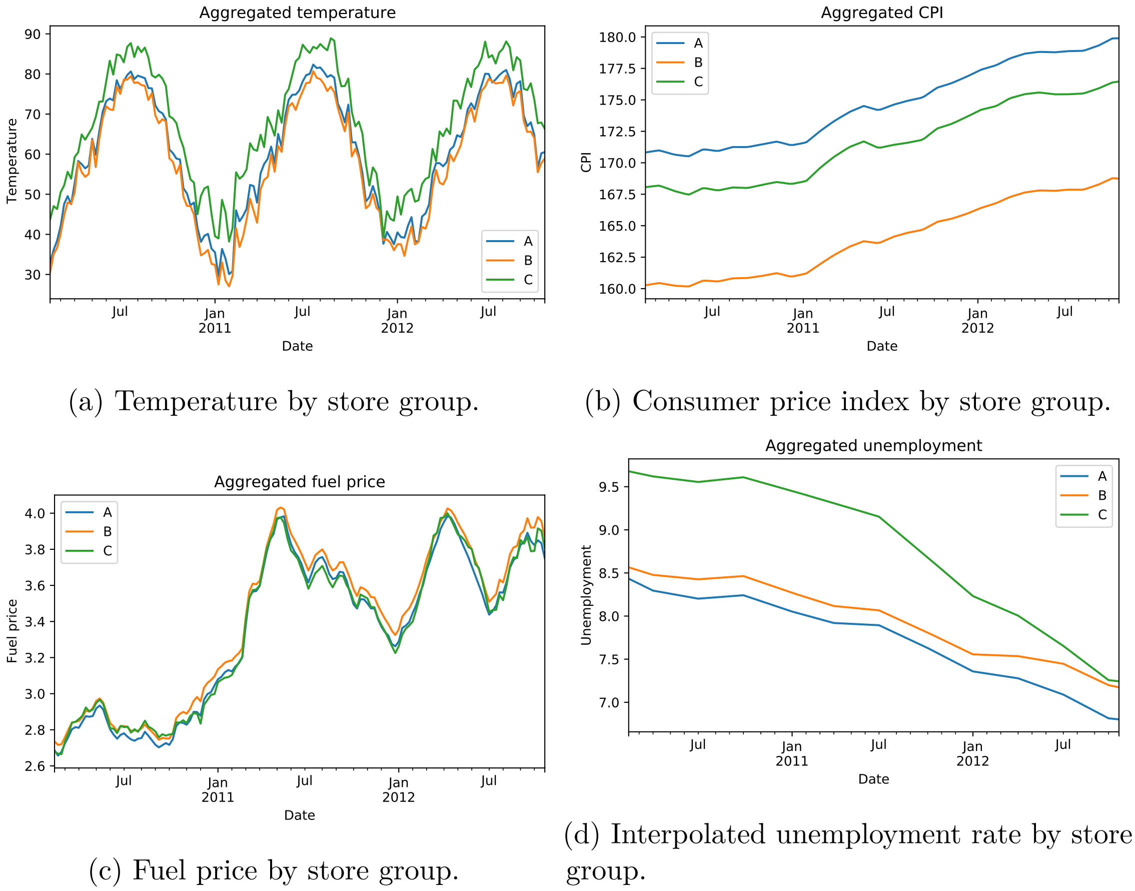
Figure 1: Aggregated weekly sales by store group. Groups A and B demonstrate holiday effects while group C remains relatively stable through each year.

week was calculated by interpolating unemployment from last and next changepoint. After the transformation, both unemployment and consumer price index (CPI) have a clear trend as seen in Fig. 2.

This thesis does not study how the features should be constructed to obtain best modelling accuracy. Feature engineering is thus left at minimum. We still acknowledge that some relatively simple transformation steps could improve modelling results. For example, temperature having strong yearly seasonality could be problematic. The model could attribute sales seasonality to temperature, which is not likely correct. Similarly, linear trends in sales can be misattributed to features with strong linear trends.

3.2 Selected models

The Prophet model by Taylor and Letham (2018) is a linear model that is used to predict time series. The model consists of seasonality components, a piece-wise linear trend, and possible explanatory variables such as holidays. An example of a component-wise breakdown of a fitted model is visualised in Fig. 3. The seasonalities are modelled using a finite Fourier series approximation. Trend is modelled linearly between several pre-determined changepoints. The Prophet model assumes a Gaussian prior for other parameters than trend changes. Trend changes have a double-exponential (or Laplacian) prior. This roughly equates to L1-regularisation as described by Tibshirani (2011). This is done to ensure a strong signal of trend change is needed for the corresponding parameter value to be non-zero. After the last change point, the model assumes a linear trend. The Prophet model treats time as a feature ranging from 0 to 1. Notably, the prediction does not directly depend on



(a) Temperature by store group.

(b) Consumer price index by store group.

(c) Fuel price by store group.

(d) Interpolated unemployment rate by store group.

Figure 2: Additional features used in Model M2.

previous observations but seasonality and trend. They are calculated also using the other observations, but that calculation does not require a constant time difference between observations. As such, the model can handle inconsistent time intervals between datapoints. This means that the training and testing set do not have to consist of sequential observations.

We chose to study five different models. The models are based on the Prophet framework. They differ from each other by the used settings and chosen features. The differences between used models are displayed in Table 1. The models are labelled from M1 to M5. The model M1 was chosen to act as a benchmark to be compared against the other models. This should demonstrate if changes to the base model yield any improved predictive performance. The M2 model adds on the benchmark model by including the additional features present in the dataset. These were temperature, CPI, unemployment rate and fuel price. The M4 model does not utilise the extra features. It attempts to demonstrate the behaviour of the metrics with overfitted models. The number of Fourier coefficients for seasonality are doubled from the benchmark model. M3 and M5 differ fundamentally from the other models as they are given noisy y-data, the sales time series, as a feature. As such M3 and M5 are not reasonably possible to implement in a real-world prediction

Table 1: Summary of the selected models. Extra features are fuel price, CPI, unemployment and temperature.

	M1	**M2**	**M3**	**M4**	**M5**
Yearly seasonality coefficients	10	10	10	20	None
Extra features	No	Yes	No	No	No
Trend changepoints	25	25	25	25	1
Response variable as regressor	No	No	Yes	No	Yes

scenario. This model formulation is done to study the behaviour of the metrics in situations where the models actually represent the real-world data generating process. In this case, the process is rather simple, as the best prediction should follow trivially from the features. The small noise is added to the feature to see how the models behave with non-perfect information. The hypothesis is that the model will try to fit the rest of the parameters to said noise. A model with smaller number of tuneable parameters should be more robust to overfitting to the noise. M3 is equal to the benchmark model extended by this noisy y feature. M5 is similar except it has less trend changepoints and disabled seasonality. M5 should be the superior model as it has the same information but it is more resistant to overfitting to noise present in the y feature. It should be capable of producing the same result as M3 but with less parameters.

3.3 Model selection metrics

The models were compared using five different metrics. The selected metrics were Akaike Information Criterion (AIC), Akaike (1974); Deviance Information criterion (DIC), Spiegelhalter et al. (2002); Watanabe-Akaike Information Criterion (WAIC), Watanabe and Opper (2010); Mean Absolute Percentage Error (MAPE), and 10-fold cross validation, Vehtari and Lampinen (2002). All the above information criteria are calculated using only in-sample fits. They aim to estimate the out-of-sample predictive performance of the model using the likelihood of past observations. Notably, only WAIC of the information criteria fully considers the Bayesian nature of the model, since it utilises the whole posterior distribution instead of some aggregated values.

3.3.1 Metric calculation

MAPE is calculated in the spirit of backtesting. First, a fifteen week test period is selected from the tail of the dataset. The model is then fitted on the preceding data. Prediction is calculated on the test period. This procedure is repeated three times, each time starting the test period seven weeks earlier. After the last predictions have been made, the absolute percentage errors are calculated. MAPE is the mean of those errors. 10-fold cross validation is calculated by assigning randomly 1/10 of the datapoints as test set and then calculating the likelihood of the prediction for the test set. This is performed ten times so that each datapoint belongs to the test set once. 10-fold CV is the average of the results. Another method typically used is to select the splits in continuous chunks. Assuming non-correlated residuals allows for selecting random splits. One of the 10-fold CV splits can be found in Fig. 4 and test set selection of MAPE is demonstrated in Fig. 5.

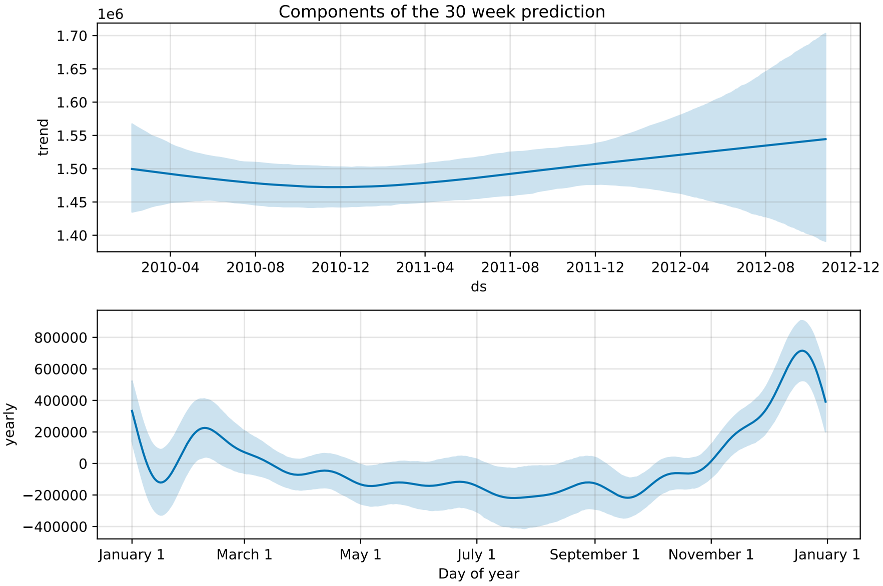


Figure 3: Trend and yearly seasonality components of the fitted model used in Fig. 5.

3.4 Bayesian model fitting

Bayesian models rely on inferring the posterior probability densities of the parameters from data and prior probabilities. For complex models, this cannot be done analytically but by using for instance Markov chain Monte Carlo sampling (MCMC). Stan (Carpenter et al. (2017)) is a Bayesian data modelling platform. In this thesis, Stan is used to mathematically model the Prophet model and fit it using MCMC sampling. To be more specific, the adaptation of MCMC used in Stan is the No-U-Turn sampler (NUTS), as described in Hoffman and Gelman (2014). This sampling process explores the possible parameter space of the model, searching for likely parameter values, and yields a large collection of point value combinations for the parameters. Usually, the sampling process utilises multiple chains to ensure proper convergence of the sampling process. Stan automatically tests model convergence by calculating multiple convergence test metrics, such as Bayesian factor of missing information, number of efficient samples per parameter, divergent transitions etc., as describe by Betancourt (2016).

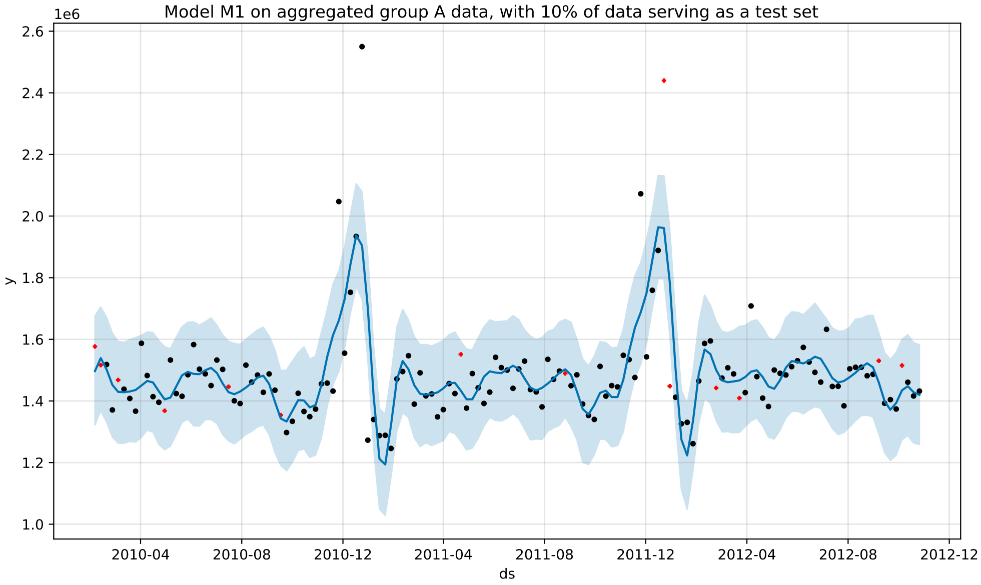


Figure 4: Demonstrating the 10-fold CV. Red dots represent the 10% of the datapoints that were selected for testing set. Model is fitted using training data represented by the black dots. This data splitting and evaluation is repeated ten times until all datapoints have belonged to the test set exactly once.

3.5 Workflow

This thesis focuses on the metrics used in model selection, not necessarily the model selection process itself. The workflow of metric evaluation was as follows: first, the data was analysed to check for which features to use or if some data processing had to be done. After this the data was aggregated to different datasets: three by store group level and nine store level datasets. The compared models were constructed from the available data. Each of the models were fitted for each of the available datasets. Each of the five metrics were then computed for all the model-data combinations. All the fitted models and metrics were saved using python's Pickle module (Van Rossum (2020)) to be analysed later. Ordering of the models by metrics results were then compared against each other within each dataset.

4 Results

The models were fitted using Stan for MCMC sampling with four chains and sampling parameters target metropolis acceptance rate $\delta = 0.9$ and maximum tree depth of 11. Each chain was sampled for a total of 1 000 times of which 500 were discarded as warmup samples. The other sampling parameters were left to their default values as suggested by Stan Development Team (2018). The obtained results for the store group level metrics are presented in Table 2. Results for store level data are presented in the Appendix A. The columns of the tables represent the evaluated metrics. Metrics other than MAPE are presented in deviance scale, i.e. as $-2lppd$. In deviance

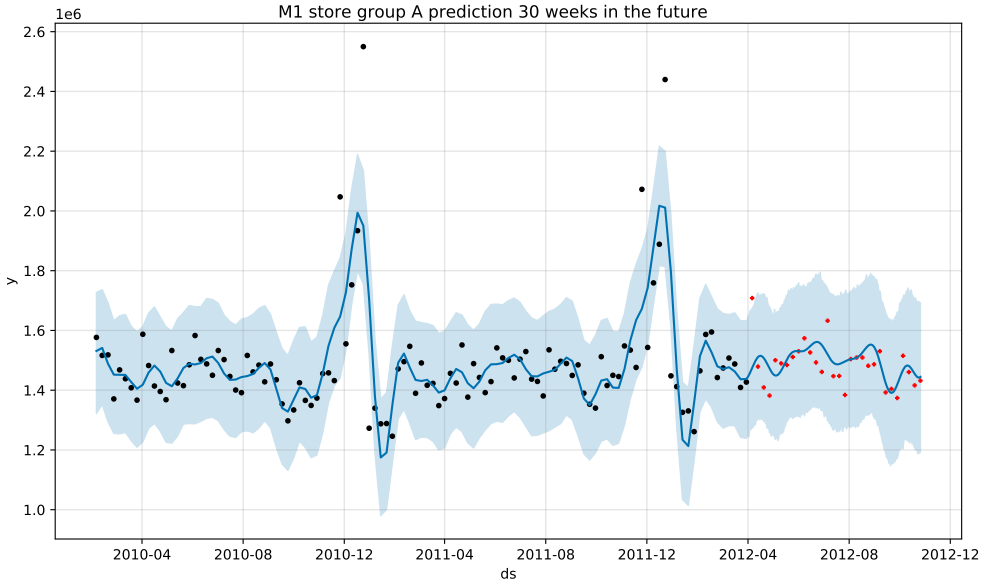


Figure 5: Hold-out prediction 30 weeks in the future. Training data in black, test data in red. This train / test split is used in MAPE metric.

scale, lower is better. The metric results can be used to rank the models by their performance. This is visualised in Fig. 6. From the figure, it can be seen that the results are somewhat consistent between the models and metrics.

The visual representation highlights the order different metrics rank the models. AIC and DIC ranked all the models similarly for all the datasets, but differences in deviance scores were not equal between the models. WAIC seemed to mostly agree with AIC and DIC, but in datasets B and C the order of models M1 and M4 was reversed.

Ranking by MAPE has many similarities with the information criterion rankings. However, Model M5 performed worse with all datasets, dropping behind M3 in A and C, but also behind M1 and M4 in B. In dataset B, MAPE ranked M1 better than M4. This agrees with the WAIC ranking, but it is the other way around with AIC and DIC. The other notable difference in MAPE scores is that in datasets A and B the Model M2 seems to perform much worse than the other models. However, this phenomenon is not present in C.

The 10-fold cross validation ranks the models M5 and M3 the best in all datasets, just like the information criteria do. However, the ranking of the rest varies. In dataset C, models M1, M2 and M4 are ranked similarly as by AIC and DIC. In both A and B, Model M4 takes the last place, while M2 and M1 get a score similar to each other. M2 was ranked better than M1 for dataset A and vice versa for B, but the margin was rather slim.

Overall, it seems that the metrics correctly recognise the "data generating models" from other models. Additionally, M5 mostly ranks better than M3, as it has less non-informative regressors. When it comes to the ranking of models M1, M2 and M4, the results are not as clear. In store groups A and B, M2 performs much worse than the other models when evaluating with MAPE. This likely indicates that the model overestimates the importance of the external regressors. As the regressors have trends, the prediction drifts off the actual sales. In the 10-fold CV this is not observed, as the datapoints in test set are picked randomly. The test points have similar-valued training points next to them, thus the trend does not have time to cause the predictions to drift off.

4.1 Metric evaluation times

Model fitting and metric evaluation times can be found in Table 3. Fitting and evaluation was performed using two cores on a 2017 MacBook Pro with an 2.3GHz Intel Core i5-7360U CPU and 8GB of RAM. For AIC and WAIC the times are averages of hundred evaluations, for DIC average of three evaluations, and rest are single runs. Fitting time depended heavily on the complexity of the model. The metrics can be categorised by their evaluation time. Both AIC and WAIC took under 10ms for all models. DIC evaluation took between six to nine seconds for all models, but this does not seem to depend on model complexity or initial fitting time. Lastly, there are both MAPE and 10-fold CV evaluation. Their evaluation times did indirectly depend on model complexity, as both metrics require the model to be refit with different splits of data. Their evaluation took between 2 to 4.5 and 7 to 11 times their respective model fitting time. The fastest category of AIC and WAIC are calculated by using only matrix operations on the sampled parameters. This makes their evaluation extremely rapid. DIC also relies only on the fitted parameters, but first maximum likelihood estimates are calculated using the prediction method of the model. This involves non-matrix operations that are considerably slower to execute.

One may wonder if it is fair to present MAPE evaluation times as including the fitting time. This choice is deliberate, as the underlying assumption is that the model needs to be fitted for the whole data for it to be useful. If using only out-of-sample MAPE for model selection, a model needs to be fitted at least once more that when using in-sample metrics. As the in-sample metrics are practically instantaneous to calculate when compared to fitting a model, it is justifiable to present information criteria based model selection as being faster.

4.2 Convergence of the metric values

The metrics were tested for stability and consistency by plotting the metric value as a function of used samples. First model M5 was fitted with 500 warmup iterations and 4000 samples per chain. The metrics were calculated by giving the evaluation function slices of the sampled parameters corresponding to the evaluated sample count. For AIC, WAIC and 10-fold CV metrics were evaluated every 10th sample, for DIC and MAPE every 100 samples. These resolutions were chosen to better manage the computation time, as both DIC and MAPE relied on predictions using sampled parameters. The results for metrics in deviance scale are presented in Fig. 7 and for MAPE in Fig. 8. Notably, for metric in deviance scale the level of noise after 500 samples is low, only under one unit of deviance. MAPE also behaves well with changes in metric value of under 0.01% after 500 samples. For all metrics the observed differences in model performances from Fig. 6 greatly exceed the level of noise in metric values.

When examining the results, it is important to keep in mind the different utility functions used in the metrics. Other metrics than MAPE are based on estimating the log-likelihood of out-of-sample predictions, while MAPE evaluates the mean absolute percentage error. Log-likelihood is proportional to MSE as the model consist of a normally distributed error term with a constant variance. In a perfect world, the utility function would represent the actual losses due to errors in predictions. However, it is difficult to estimate those accurately.

From a business application point of view, it can be concluded that the 10-fold CV is not a feasible model selection metric. In practice, the used models are very complex and may take hours to fit. This renders cross validation-based methods useless as the results are often needed rapidly. If the dataset has enough datapoints, hold-out based metrics like MAPE and others with possibly different utility functions may be used. In a scenario with limited data availability the information criteria may suffice.

Table 2: Numerical representation of the validation metric results applied by store group.

(a) Store group A

Model	**AIC**	**DIC**	**WAIC**	**MAPE**	**10-fold CV**
M1	-394.4	-438.8	-15.9	0.0336	-410.8
M2	-384.6	-431.5	136.5	0.0445	-416.4
M3	-654.5	-702.1	-669.3	0.0219	-699.2
M4	-457.4	-512.4	-41.6	0.0275	-328.2
M5	-703.9	-703.2	-707.8	0.0253	-702.6

(b) Store group B

Model	**AIC**	**DIC**	**WAIC**	**MAPE**	**10-fold CV**
M1	-374.3	-419.1	19.3	0.0279	-408.4
M2	-365.5	-412.2	315.9	0.0494	-393.8
M3	-647.8	-695.3	-663.2	0.0255	-694.6
M4	-446.9	-501.4	152.2	0.0286	-324.2
M5	-699.4	-698.3	-701.5	0.0297	-697.8

(c) Store group C

Model	**AIC**	**DIC**	**WAIC**	**MAPE**	**10-fold CV**
M1	-454.0	-501.2	-397.1	0.0339	-494.8
M2	-447.1	-496.6	-186.4	0.0349	-488.4
M3	-668.9	-715.6	-668.1	0.0136	-714.0
M4	-549.1	-605.1	-214.3	0.0253	-582.8
M5	-741.9	-740.6	-743.7	0.0140	-740.4

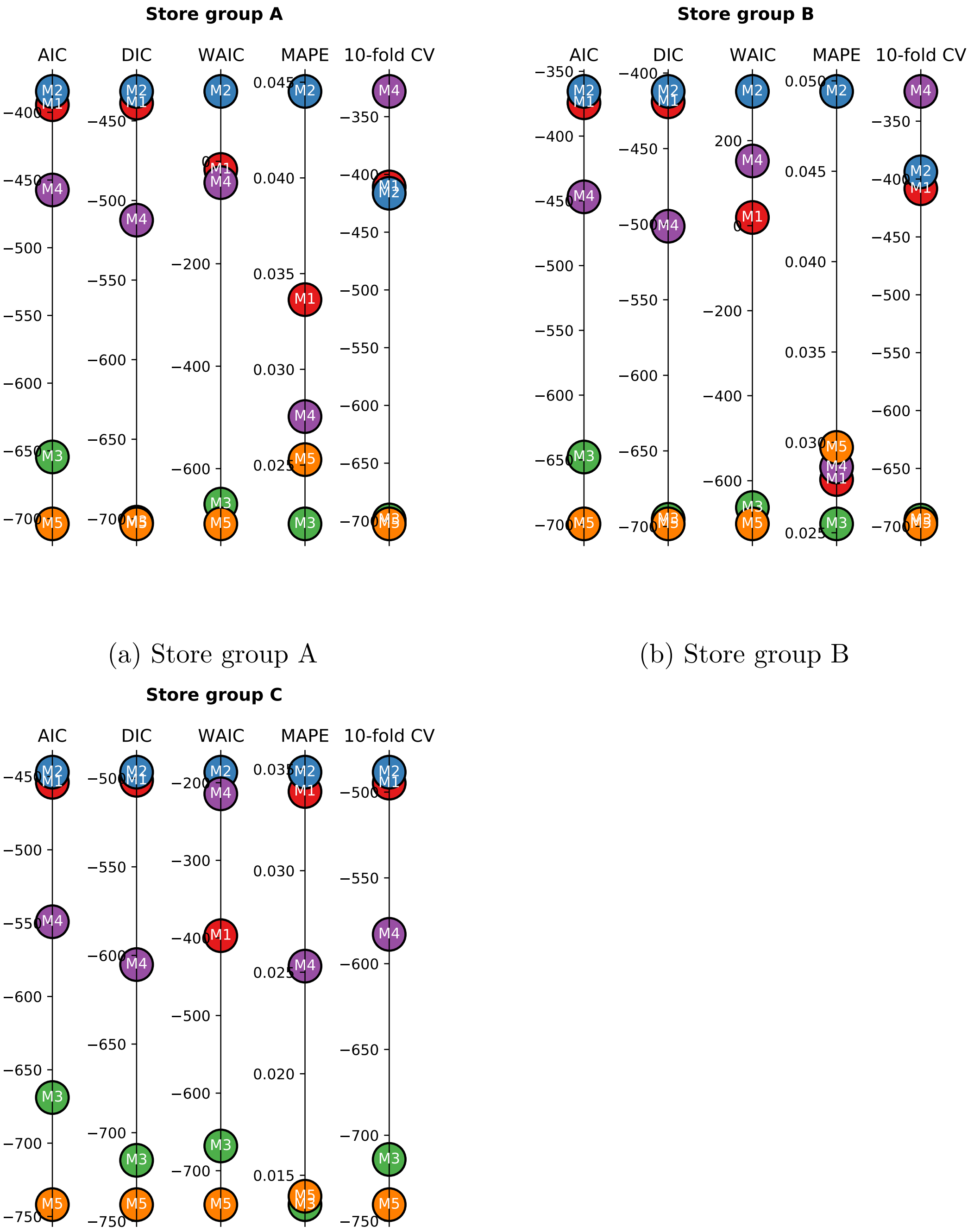
Table 3: Model fitting and metric evaluation times in seconds.

	Initial fitting time	AIC	DIC	WAIC	MAPE	10-fold-CV
M1	32.27	0.00207	8.31	0.00420	108.39	358.46
M2	59.47	0.00326	7.88	0.00412	163.33	491.66
M3	105.55	0.00740	8.06	0.00453	203.68	763.72
M4	65.30	0.00312	7.48	0.00426	151.36	490.07
M5	7.68	0.00376	6.69	0.00453	34.79	83.10

5 Summary

In this thesis, we studied the applicability of multiple model selection metrics using datasets and Bayesian models typical to business settings. The studied metrics demonstrated consistent results in ranking model performance, making them useful for model selection. However, cross validation may be too computationally intensive to apply in practice, but other studied methods can be used instead as their accuracy in ranking the models is adequate. Having only 140 weeks was not an issue for all metrics, as the in-sample-metrics were not suffering from the lack of data.

To study further the applicability of common model selection metrics one could experiment with a retail dataset with better data quality. The effect of marketing and discounts could greatly affect sales and thus make for an interesting modelling problem. Longer time series would be desirable, as it would help with separating seasonality from the effect of explanatory variables.



(a) Store group A

(b) Store group B

(c) Store group C

Figure 6: Visualisation of model validation metric results by store group. Arranged such that visually lower is better. Each axis is scaled so that extreme values are at the ends. For precise values please refer to Table 2.

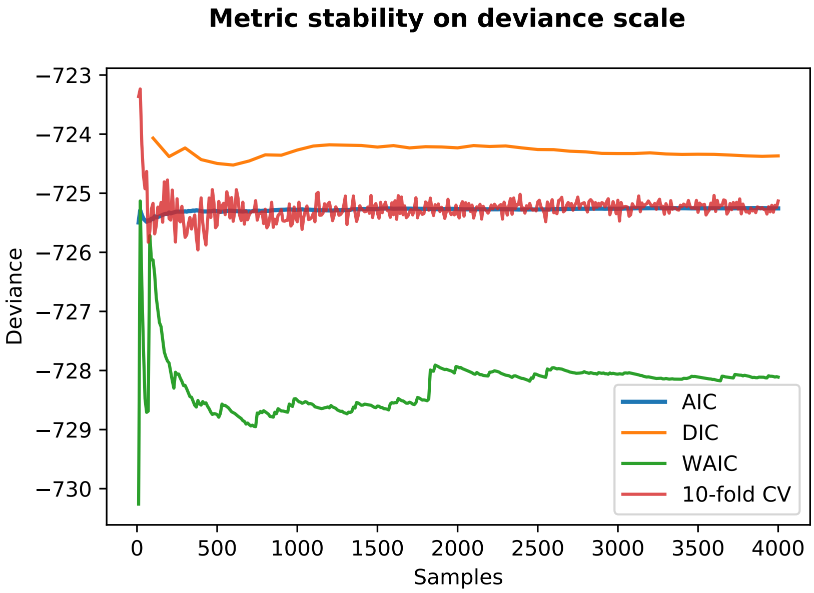


Figure 7: Metric values by sample count. Lower noise indicates more stable results. Metrics in deviance scale are present here, for MAPE see Fig. 8.

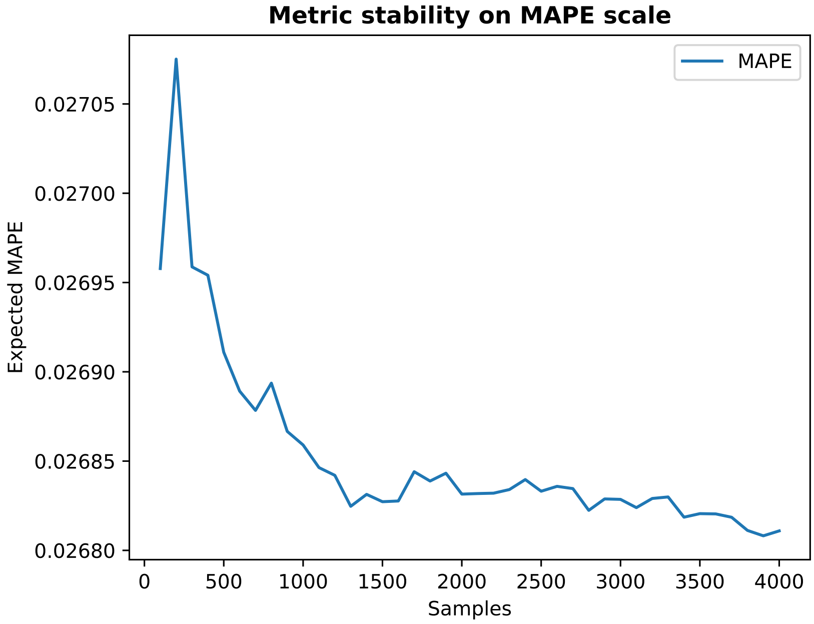


Figure 8: Expected MAPE by sample count. Lower noise indicates more stable results. Metrics in deviance scale are presented in Fig. 7.

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. doi: 10.1109/TAC.1974.1100705.

M Jésus Bayarri and James O Berger. The interplay of bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80, 2004.

Michael Betancourt. Diagnosing suboptimal cotangent disintegrations in hamiltonian monte carlo. *arXiv preprint arXiv:1604.00695*, 2016.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76 (1), 2017.

David Chan and Mike Perry. Challenges and opportunities in media mix modeling. 2017.

Thomas H Davenport et al. Competing on analytics. *Harvard business review*, 84 (1):98, 2006.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, Nov 2014. ISSN 1573-1375. doi: 10.1007/s11222-013-9416-2. URL https://doi.org/10.1007/s11222-013-9416-2.

Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014. URL http://jmlr.org/papers/v15/hoffman14a.html.

A Gürhan Kök and Marshall L Fisher. Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research*, 55(6):1001–1021, 2007.

Maurice Landry, Jean-Louis Malouin, and Muhittin Oral. Model validation in operations research. *European Journal of Operational Research*, 14(3):207–220, 1983. ISSN 0377-2217. doi: https://doi.org/10.1016/0377-2217(83)90257-6. URL https://www.sciencedirect.com/science/article/pii/0377221783902576. Methodology, Risk and Personnel.

Manjeet Singh. Retail dataset. Kaggle dataset, 2017. URL https://www.kaggle.com/manjeetsingh/retaildataset/.

David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002. doi: https://doi.org/10.1111/1467-9868.00353. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00353.

Stan Development Team. Hmc algorithm parameters, 2018. URL https://mc-stan.org/docs/2_27/reference-manual/hmc-algorithm-parameters.html. Version 2.27.0.

Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018. doi: 10.1080/00031305.2017.1380080. URL https://doi.org/10.1080/00031305.2017.1380080.

Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3): 273–282, 2011.

Guido Van Rossum. *The Python Library Reference, release 3.8.2*. Python Software Foundation, 2020.

Aki Vehtari and Jouko Lampinen. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10):2439–2468, 2002. doi: 10.1162/08997660260293292.

Aki Vehtari and Janne Ojanen. A survey of bayesian predictive methods for model assessment, selection and comparison. *Statist. Surv.*, 6:142–228, 2012. doi: 10.1214/12-SS102. URL https://doi.org/10.1214/12-SS102.

Sumio Watanabe and Manfred Opper. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12), 2010.

A Results for store-level data

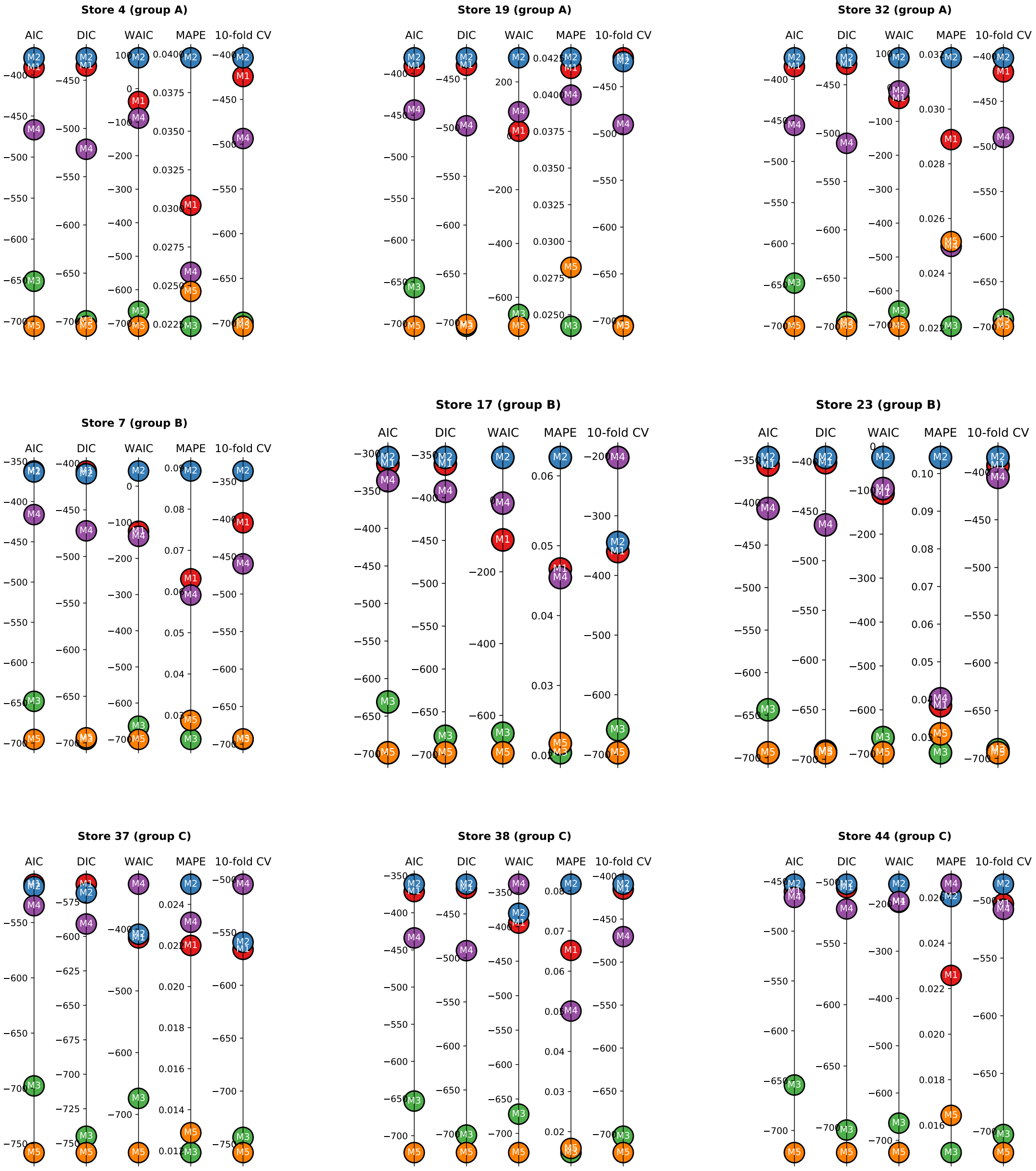


Figure A1: Visualisation of model validation metric results for single stores.

B Source code

This thesis and the used source code can be found online in Github. https://github.com/leevironty/BSC_thesis