MS-E2177 Seminar on Case Studies in Operations Research 2018

Predicting and managing group cancellations in air passenger traffic

Client: Finnair

Final report

May 31, 2018

Tuomas Koskinen (project manager)

Elias Axelsson

Olli Herrala

Introduction

Overbooking is arguably the most important component of airline revenue management. According to some estimates as many as 15% of all seats would go unsold without some form of overbooking due to the high rate of cancellations and no-shows observed in the business. Approximately 50% of all reservations end up being either cancelled or a no-show, because the customer does not cancel the reservation but does not arrive in time for departure. With these figures in mind it is easy to see why developing a sound overbooking strategy is integral for the success of an airline. As a result much work has been done to develop the mathematical theory pertaining to optimal overbooking.
Most of these models depend heavily on the assumption of single reservations being independent of each other.[1] Unfortunately, this assumption is violated by the presence of group reservations which are all too common in markets most important for Finnair's business.
Consequently, the objective of this project work is to investigate different methods for dealing with group reservations. Most importantly a predictive model is developed for forecasting group cancellations. The model is based on basic information that Finnair collects on all of its reservation such as destination and origin country, time of reservation and departure, travel agent information etc. In addition to relying on the assumption of reservation independence previous work on overbooking models is based on optimizing around known baseline cancellation and no-show rates within certain customer segments. If one can to accurately predict cancellation probabilities on a reservation-by-reservation basis it would be possible to further optimize overbooking based on the information supplied by the model.
The second objective is to develop a simulation model for investigating how such a predictive model would affect optimal revenue in a simplified overbooking scenario. The simulation is used to study how the accuracy of the model affects its usefulness. Based on the results conclusions can be made on whether the developed forecasting model could be helpful in practice.

Literature review

Revenue management

The practice of revenue management (RM) originates to the U.S airline industry in the 1970s. [1] It has been since continuously developed and presently it is a common and crucial practice among wide area of different industries. Throughout the history of airline RM, the main objective has been to maximize the revenue gained from a flight. This includes for example ticket pricing, fare class mix and overbooking.

Using different fare classes the airline attempts to attract customers with different desires. Usually there are at least two fare classes, low fare and high fare. The most simple problem is to decide how much low fare and high fare seats should be sold in a flight. It is also one of the first questions in the history of airline RM. The earliest known revenue management model is the *Littlewood's rule*, which provides a simplified solution for the problem. It assumes that low fare reservations arrive before high fare reservations.[5] Thus, the solution is the seat index x, when the sales of low fare tickets should be closed and the sales of high fare tickets should start. Our simulation is based on this assumption.

The overbooking is a major part of airline RM, because it increases the expected revenue significantly. This is because cancellations rates are high and also there is usually little or no deposit paid before cancellation. Figure 1 describes how the ultimate demand ("show demand") increases when overbooking is used. C is the capacity of the plane.

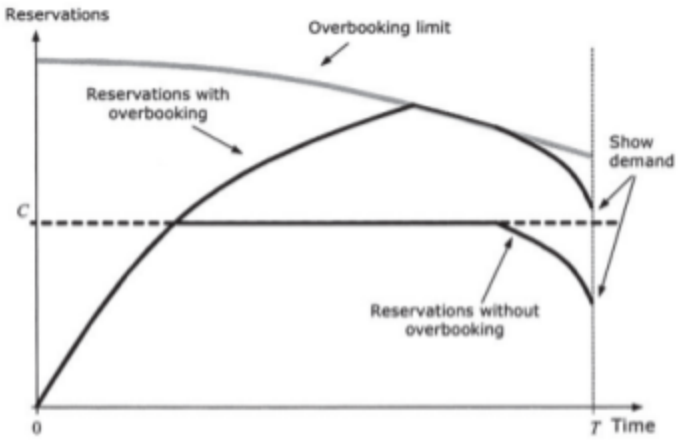

Figure 1: An illustration of overbooking limits and reservations over time. [1]

Of course overbooking leads sometimes to denied boardings, which have provoked even some legal issues in recent decades.[1] However, in the airline RM, the denied boardings are described with a penalty, which can be seen as a monetary value.

In order to use overbooking effectively the cancellations should be predicted as well as possible. Usually some kind of a mathematical forecasting model is used. However, most of them consider customers independent of each others, which does not hold in reality in the case of group reservations.

Forecasting models

In order to achieve the main objective of the project, that is predicting group cancellations, different statistical models and algorithms were studied. The ones which were used in this project are explained in what follows.

Logistic regression

Logistic regression is a simple linear model for binary classification. For each input the output is either 0 or 1. Logistic regression models the distribution of the dependent binary variable y as a function of the input vector $\mathbf{x}$ according to the equation

$$Pr(y = 1) = \sigma(\alpha + \beta^T \mathbf{x}),$$

where $\sigma(\cdot)$ is the sigmoid function

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

Parameters α, β are estimated using a training sample, for example with maximum likelihood estimation, and the labels of new observations can be predicted as follows:

$$y = \begin{cases} 1, \text{ if } \alpha + \beta^T \mathbf{x} > 0 \\ 0, \text{ else} \end{cases}$$

Logistic regression is thus a very simple model and its decision boundary is linear. As a result it may not be sufficient for modeling complex relationships between variables but it can often give satisfactory results with minimal effort. Its results are easy to interpret, simply put if $\beta_i > 0$ large values of $\mathbf{x}_i$ increase the probability of y being equal to 1 and similarly if $\beta_i < 0$ larger $\mathbf{x}_i$ increases the probability of $y = 0$. As an example, in the context of this project work label 0 is used to indicate that a flight was cancelled and label 1 indicates lack of cancellation, by looking at the estimated coefficients it can be seen how different variables affect probability of cancellation.

Random forest

While logistic regression is a simple and easily interpretable model, it will not often yield the best performance when the underlying relationships are not linear. A popular model often used to address such problems is the random forest model. The random forest is a so called ensemble model, which means that it is made up of multiple simple models. In the case a random forest, these simple models are decision trees. A number of decision trees are fitted on randomly chosen subsets of the data, and when making a prediction majority voting is used. If most of these decision trees output a 1 instead of a 0 the final prediction is 1 and vice versa. Ensemble methods help alleviate the common problem of overfitting. Individual decision trees often fall for overfitting the data, but random forests resolve this problem to some extent.

The details of decision tree fitting fall out of the scope of this brief review. There are multiple algorithms for building decision trees, each algorithm uses some metric which is to be minimized in the training set and splits the data step by step by seeking the variable which best minimizes the chosen metric.[3]

Survival analysis

Methods such as logistic regression or random forest can be used to make a simple binary prediction, in the context of this project work “0 - cancelled” or “1 - not cancelled”. However, such a prediction fails to capture the dynamic nature of a process such as group cancellation. As an example you might use a model such as a random forest to predict whether a reservation will be cancelled before departure at the time of reservation, but it is unclear how the probability of cancellation will evolve as a function of time to departure. Obviously, the lower the time to departure is the lower the predicted probability of cancellation should be and it should go zero as the last opportunity for cancellation approaches. The prediction should be able incorporate the passage of time to be more useful but it is not clear how one would implement this using the models discussed before. One way to approach this problem is survival analysis, which is a branch of statistics for analyzing the expected duration of time until an event happens, be it the death of an organism in a medical study or something entirely different depending on the context. The most central concept in survival analysis is the survival function

$$S(t) = Pr(T > t),$$

where T is the time until an event happens. That is, the survival function is the probability of survival until time t. The are various different methods for estimating the survival function with many of them making it possible to model the survival function as a function of some independent variables. One commonly used survival analysis method is the Kaplan-Meier estimator

$$\hat{S}(t) = \prod_{i:t_i<t} (1 - \frac{d_i}{n_i}),$$

where t_i is any time when at least one event happened, d_i is the number of events that happened at time t_i and n_i is the number of individuals who have have not had an event until time t_i.In the context of this project work the event being studied is cancellation, models such as this one can be used to model how the probability of cancellation evolves as a function of time and they are likely to provide more insight than a model with binary output.[4]

Data

It was clear from the beginning that the project would depend heavily on data. Since we were requested to make a classifier or predictor for whether a booking would be cancelled before flight, we knew we had to have some variable(s) determining whether there was a cancellation or not. In addition to that, explanatory variables were also required to make predictions with. In our problem, the response variable was decided to be a binary with a value 1 if at least 50% of the initial booking size was left on the day of departure and 0 otherwise. Based on our initial discussions with Finnair, the feature variables should include at least a variable about whether or not the deposit was paid, the group size and time from initial booking to the flight. In addition to these critical ones, we also used origin and destination information on different geographical levels, and which agent made the booking.

The data for the project was received from Finnair in the end of February. The data set consists of 151593 observations, each representing a change in a reserved flight. The number of unique booking identifiers is 33214, resulting in an average of 4,6 rows per booking. However, each booking consists of all the flights of the trip, meaning that a return journey has at least two rows in the data. The timestamps of the changes range from 2016/01/18 to 2017/05/29, and the flight dates from 2017/01/01 to 2018/02/01. In addition to these observations, we can find that 12157 booking numbers have paid deposits, resulting in 37% paid deposits. After processing the raw data to a form in which each row represents a single flight in a booking, the row count decreases to 84847 rows.

The processing of the data was done in Python, using the Pandas data analysis library. The main reason for choosing Pandas for the task was that one of the team members had prior experience with it, and it is quite straightforward and easy to learn. First the data was read into a Pandas dataframe from the received csv-file. Then, the initial grouping of the data was carried out, using the combination of booking identifier, flight number and flight date as the key. In addition to these, the origin and destination information, along with the agent information was included. For each such combination, two aggregated features were calculated in the initial grouping. The first one was the minimum of change dates, or the date of the initial booking. The second one was the

sum of all changes made to the number of passengers, resulting in the final number of passengers. For example, if the (chronologically) first row had 9 as the change number, and the rest would be -2 and 1, that would mean that 9 seats were originally reserved, of which 2 got cancelled before one returned. The final number of passengers in this case was therefore 9-2+1 = 8. After these simple aggregations, we added three more features by first creating a suitable dataframe, then doing an sql-style join to the processed data. First was grouping the data by the booking id and counting the number of distinct values of the deposit column. If the deposit was paid for that booking, there was a row with a value of ‘depo’ in addition to a number of null values, and otherwise only nulls, thus resulting in the binary variable mentioned earlier. This dataframe was then joined by the booking id, adding the binary deposit feature to the processed data. The next feature to add was the initial booking size, which was received by joining the original data to the grouped one by the three-feature key and the row timestamp. For the original data, the change date was joined with the initial booking date in the grouped data, and the change number from that row was used. The third and final feature was the size of the agent office, which we decided would be determined by the count of distinct booking identifiers associated with the agent. This count was extracted from the original data and then joined to the grouped data.

The first thing we can see from the data is the popularity of different airports. The most popular one is obviously Helsinki airport, as the data consists of Finnair flights. Counting the flights by origin or destination both give Helsinki-Vantaa 50% of the data, and only 0,1% of the flights do not have Helsinki as one of their endpoints. These exceptions are mostly flights between Arlanda and Bergen. Finnair also provided us with information about the “true” origin and destination in the data. These are more informative, as the previous statements about Helsinki-Vantaa were only good for proving the point about Helsinki having a good geographical location for Asian flights. The true origin and destination are not flight-specific, but booking-specific. As we can see in table 1, Helsinki still dominates in both categories, but is not even close to 50%. We can also observe that the origin cities are heavily focused in Asia, while the destination cities have no Asian cities. This is as expected, as the goal of the project was to gain a deeper understanding of Finnair’s Asian flights.

Origin code	Origin city	N	% of all bookings	Destination code	Destination city	N	% of all bookings
HEL	Helsinki	11483	35,1 %	HEL	Helsinki	7871	24,0 %
NRT	Tokyo	1718	5,2 %	RVN	Rovaniemi	1786	5,5 %
KIX	Osaka	1179	3,6 %	FCO	Rome	1334	4,1 %
NGO	Nagoya	1053	3,2 %	BCN	Barcelona	1227	3,7 %
HKG	Hong Kong	883	2,7 %	KTT	Kittilä	1065	3,3 %
ICN	Seoul	831	2,5 %	IVL	Ivalo	947	2,9 %
PVG	Shanghai	799	2,4 %	CDG	Paris	945	2,9 %
OUL	Oulu	692	2,1 %	MXP	Milan	868	2,7 %
ARN	Stockholm	663	2,0 %	PRG	Prague	815	2,5 %
PEK	Beijing	656	2,0 %	CPH	Copenhagen	810	2,5 %

Table 1: Distribution of reservations between destination and origin cities.

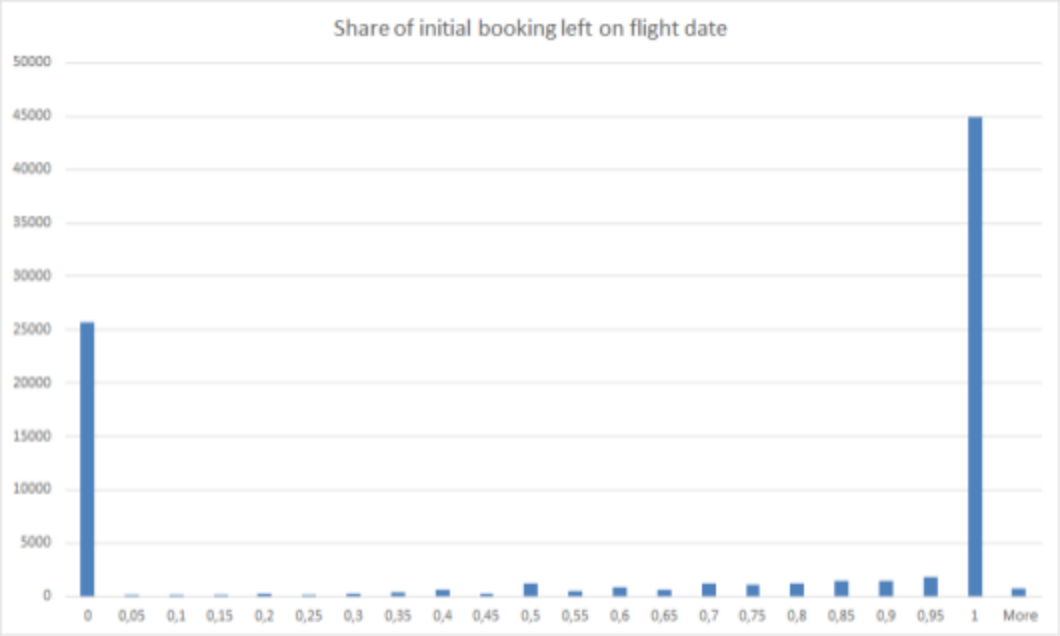

Figure 2: Number of reservations by proportion of non-cancelled seats.

As the topic of the project was cancellations, it seems reasonable to observe how many of the seats on booked flights get cancelled. For this purpose, a histogram was plotted with the share $\frac{booking\ size\ at\ time\ of\ flight}{initial\ booking\ size}$ on the x-axis, and the number of bookings on the y-axis. What we immediately see from this figure is that a very large share of the bookings get cancelled. The proportion of completely cancelled flight bookings is 30%, and the average size of the initial booking in those is 21 people, while the average for both cancelled and non-cancelled flights is 13 people. This illustrates the core of the research problem: big groups are more likely to cancel. This is further supported by figure 4 containing the number of observations and the survival rate of the initial booking as a function of initial booking size. Hardly half of the groups of over 10 people eventually materialize on the flight. In addition to this, we also observe that booking sizes of exactly 10, 20, 30 etc. are even more unlikely to materialize. This is most likely due to the fact that these bookings are made by traveling agencies with the plan of filling the booking before the actual flight. Based on our discussions with the client, it is pretty common for Asian travel agencies to book a large number of John Smiths and then gradually change the names as they get more passengers to the flight. However, these

often get cancelled, because there are no actual people making the reservation in the first place.

The same phenomenon can also be seen if we examine the time between the initial booking and the flight date. The longer this time is, the more likely it is that the booking

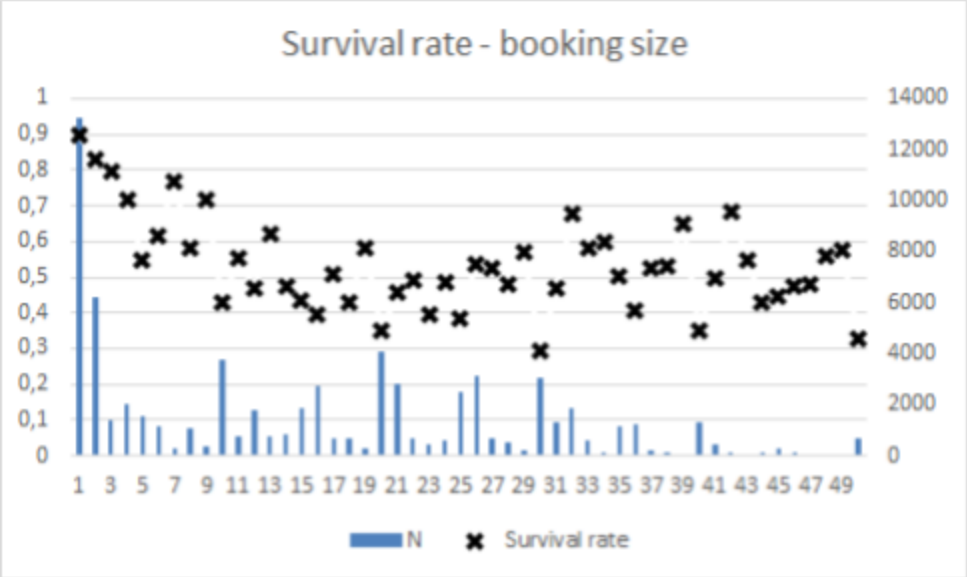

Figure 3: Proportion of non cancelled flights (left axis) and number of reservations (right axis) by size of reservation.

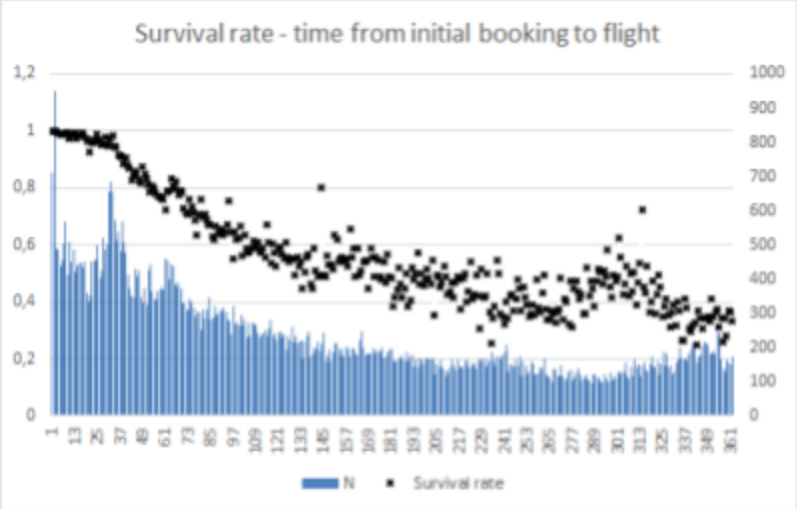

Figure 4: Proportion of non-cancelled reservations (left axis) and number of reservations (right axis) by days to departure.

will be cancelled. For flights booked less than one month prior to departure, the survival rate is over 0,9, but at roughly 4 months it reaches 0,5. If the flight is booked nearly a year in advance, the survival rate is below 0,4, which makes those bookings very risky for the airline company.

Because both the time from initial booking to flight date and initial group size correlate negatively with the survival rate, it is pertinent to examine the correlations between them. When plotting the average group size as a function of the described time period, we see that there is a significant positive correlation between these two variables. This might cause problems when building the model, as multicollinearity makes it harder to distinguish the effects of two variables from each other.

Figure 5: Average booking size as a function of time to departure.

Predicting cancellation

The main objective of this project assignment is to develop of a forecasting model to predict group cancellation based on the provided data. This would enable optimizing overbooking policy based on the predictions assuming a high enough accuracy is

achieved. To accomplish this it is first necessary to define what exactly is the quantity being forecasted and then to preprocess the data accordingly.

The data consists reservations identified by a reservation number (pnr id) along with the flight number. The original objective was to consider each reservation, often consisting of multiple individual flight, as one unit and try to predict whether it is going to be cancelled or not. However, defining cancellation in this context proved to be somewhat problematic. With most reservations there is no problem, if a given number of seats is cancelled on one of the flights the same number of seats is also cancelled on all the other flights that are a part of the same reservation. However, consider for example a reservation consisting of two connecting flights. It might be the case that the second flight is changed to another flight at a different time but the reservation for the first flight does not change. How would one define whether this reservation was cancelled or not in a meaningful, consistent manner? In addition to cases like this there were many cases in the data where for example 5 seats were cancelled on one of the connecting flights but not on other flights belonging to the same reservation. It is not clear why this happens or whether it's just a bug in the data but it further complicates defining the cancellation status of a reservation.

For these reasons instead of working with reservations as a whole a decision was made to simplify the problem and do the modelling on a flight by flight basis. This means that each reservation on a given flight is considered completely separately from the other flights in the same reservation. This simplifies the problem and makes it possible to easily frame it as either a statistical classification or regression problem but this assumption needs to be taken into account when interpreting and evaluating the results. After the forecasting problem had been properly defined the data needed to be preprocessed accordingly. For each flight the original group size and materialized group size were calculated from the data. In most cases the flight was either completely cancelled or not cancelled at all, but in some cases only a portion of the originally reserved seats were cancelled. In order to frame this as a binary classification problem, flights like these were labeled cancelled if more than half of the seats were cancelled and were otherwise labeled as uncancelled.

Next it was necessary to apply some feature engineering in order to extract relevant information from the data. The provided data was high dimensional and thus only features which were intuitively considered to be relevant were included. These are reservation month, flight month, days from reservation to flight, origin country, destination country, travel agent country, travel agent office identifier, office size, whether the deposit was paid or not and group size. The categorical variables needed to be further preprocessed because the contained a high number of rare modalities. In order to combat this all rare modalities, which had less than 2000 instances in the data, were all lumped into the same category “other”. For the office identifier value 0 is used instead. An example of the resulting data is given in the following image. After these preprocessing steps the data is in suitable form for different classification models to be fitted.

RESERVATION_MONTH	FLIGHT_MONTH	DAYS_TO_FLIGHT	ORIGIN_COUNTRY	DESTINATION_COUNTRY	AGENT_COUNTRY	OFFICE_IDENTIFIER	OFFICE_SIZE	DEPOSIT	GROUP_SIZE	CANCELLED
01	01	353.625	OTHER	FI	OTHER	0	1239	0	25	0
03	02	342.0868056	KR	ES	KR	0	1625	0	32	1

Figure 6: Two randomly chosen reservations after preprocessing.

Logistic regression

First a simple logistic regression model was fitted on the data. In contrast to more complex models the results of logistic regression are easily interpretable and thus may provide valuable insight into the problem. The data is first partitioned into a training set and a validation set. The data is sorted by reservation date and the first roughly 80% of the data is used for fitting the model and the following 20% is used for evaluating it is accuracy. The 80/20 partition is chosen because it is commonly used. Sorting by date is done to simulate a realistic setting, the model is fitted based on past observations and new ones are predicted. This also ensures that the validation and training set will not contain flights belonging to the same reservation. This would otherwise be a problem because flights belonging to the same reservation are clearly not independently distributed. This is also the reason why using cross-validation would be questionable in this case. In addition cross-validation is made impractical by limited computing power available. The training set consists of 55000 rows while the remaining 16478 rows are used for validation.

After fitting the model on the training set a prediction was made for the validation set. A predictive accuracy of 78% was reached in the training set, but a slightly lower accuracy of 75% is achieved on the validation set. To get a better image of the accuracy of the model the confusion matrix is calculated and it is presented in table 2.

True value / Predicted	Cancelled	Not cancelled	Total
Cancelled	1896	1342	3238
Not cancelled	2831	10409	13240
Total	4727	11751	16478

Table 2: Confusion matrix for validation of the logistic regression.

It can be seen that out of the 13240 non-cancelled flights in the validation set the model correctly predicts approximately 78,6% but it only manages to predict 58,6% of the cancelled flights. It can be noted that only roughly 20% of the flights in the validation set were cancelled in contrast to approximately 40% in the whole dataset. As a result a 80% accuracy could be reached by always predicting a flight not to be cancelled. Taking this into account the achieved accuracy is not great.
By looking at the signs of the regression coefficients it can be seen how different features affect the cancellation probability. Some of the most notable coefficients are given in table 3.

Feature	Deposit	Group size	Days to departure	Agent identifier = 8
Coefficient	2.51	-0.010	-0.0036	4.65
Feature	Agent identifier = 22	Agent country = KR	Agent country = FI	Origin country = JP
Coefficient	-1.78	2.50	-1.19	0.89

Table 3: Some of the regression coefficients.

Thus it can be seen that for example paying the deposit significantly decreases probability of cancellation but higher group size and time to departure seem to increase it. It also looks like agent number 8 rarely cancels but number 22 is somewhat likely to cancel. Korean agents are less likely to cancel than Finnish ones and flights originating from Japan also have a decreased probability of cancellation. These are just some handpicked examples. The model was fitted using an implementation available in scikit-learn library available for Python.[2]

Random forest

In similar fashion to the logistic regression a random forest model was fitted to the data. The logistic regression is a simple linear model that is easy to interpret but it is predictive power is not as high as some more complex models such as the random forest. A random forest consisting of 100 decision trees was fitted on the same data using the implementation provided in the sklearn library. An accuracy of approximately 15% was achieved on the validation set. The confusion matrix is given in table 4.
Notably, the random forest model predicts non-cancelled flights better than the logistic regression model at 92,3%. However, it is not better at predicting cancelled flights at only 55,9% accuracy. Using the random forest model the usefulness of different predictors is evaluated using a measure known as mean decrease in accuracy. This means that the values of one feature are randomly permuted and the model is fitted again and its accuracy on the validation set is calculated and compared with the original accuracy. Results are presented in table 5. Based on this analysis the deposit payment is the most

powerful predictor of cancellation followed by days to departure and group size. The model was fitted using an implementation available in scikit-learn library available for Python.[2]

True value / Predicted	Cancelled	Not cancelled	Total
Cancelled	1812	1426	3238
Not cancelled	1016	12224	13240
Total	4727	11751	16478

Table 4: Confusion matrix for validation of the random forest.

Permuted feature	Validation accuracy (%)	Change (%)
None	84.8	0
Agent office size	84.7	-0.1
Group size	83.6	-1.2
Days to departure	82.1	-2.7
Deposit	78.2	-6.6
Office Identifier	85.0	+0.2
Destination country	85.1	+0.3
Origin Country	85.0	+0.2
Agent country	84.7	-0.1
Reservation month	84.1	-0.7
Flight month	85.2	+0.4

Table 5: Mean decrease in accuracy for each feature.

Survival analysis

The problem with binary predictions such as those given by logistic regression or random forest models is that they do not properly reflect the dynamic nature of the reservation/cancellation process. That is, the prediction should change as a function of time to departure. It may have originally predicted a reservation to be cancelled but the probability of cancellation should approach zero as time to departure does so. There is no easy way to capture this using a binary classification scheme. Survival analysis is a natural method for modeling this type of process.

To fit survival analysis models on the data it needs to be processed into a different form. For example, when five seats are cancelled this is modeled by five individuals “dying”. Materialized reservations consist of individuals that “survived” and are thus censored. An attempt was made to fit a Cox proportional hazards model on the dataset. This model makes it possible to model the effect different covariates have on survival rates. However, after continuously running into problems with convergence and numerical stability of the implementation this plan had to be ditched. This is disappointing but a Kaplan–Meier estimator can still be fitted to visualize survival rates in the dataset. The resulting survival curve is presented in figure 7. The x-axis shows days since the reservation was made and the y-axis represents the probability that a reservation is not cancelled after a given number of days have passed. It can be seen that, for example, after 200 days roughly 50% of reservations were cancelled. The probability of surviving a full year is roughly 25%. This curve provides a accessible visualization of the survival rates observed in the dataset but it is not very useful for optimizing overbooking because it does not take into account information about the covariates. Because of this further analysis in this work is based on the simplified binary prediction scheme in order to keep the workload manageable.

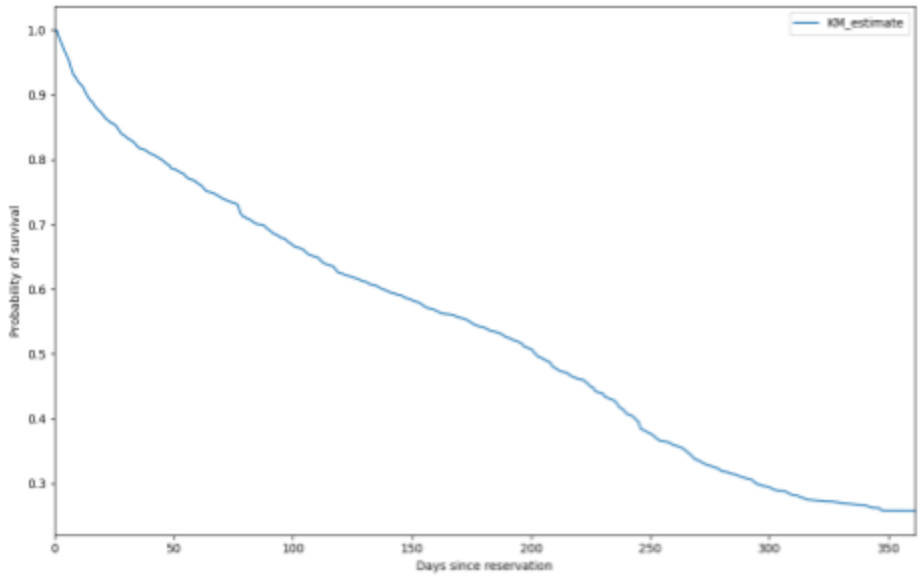

Figure 7: Survival curve given by the Kaplan-Meier estimator.

Simulation

The objective of the simulation is to find which accuracy a forecasting model should achieve in order to increase the expected revenue form a flight. The reference method is a fixed overbooking limit, which is also optimized using simulation. Put otherwise, the goal is to find which accuracy (P_{model}) is enough to create larger expected revenue than the optimal overbooking limit L.

The accuracy of a forecasting model is measured in percentage P_{model}, which describes how likely the model is to predict whether a reservation will be cancelled or not. This means that if the reservation will be cancelled, the model will state “cancelled” with probability P_{model} and “not cancelled” with probability 1 - P_{model}. It is assumed that the decision maker is strictly obedient to the model. The prediction will not change in time, which is of course a major simplification. But as it holds also for the reference method (fixed overbooking limit L), the results should be somewhat comparable.

The booking process of a flight with 100 seats and two fare classes will be simulated. To make it simple enough, multiple assumptions are made. First of all, the low fare reservations will arrive before high fare reservations. The sales of low fare tickets will be discontinued as the predicted seat index exceeds the limit x, which is also optimized in the simulation. Then, the high fare tickets will be in sale as long as the predicted seat index exceeds 100. Additionally, there is time limits for both sales, T_{low} and T_{high}. Once the sales are closed, no tickets will be sold even though there would be cancellations. In the reference simulation the sales are closed as the number of reserved seats reaches the overbooking limit.

The reservations will arrive as a Poisson process with rate parameters λ_{low} and λ_{high} such that $\lambda_{low} > \lambda_{high}$. The reservation size is also exponentially distributed with rate parameter λ_{size}, which is equal for both fare classes. A fixed cancellation rate of 30% is used in the simulation. The time of cancellation is randomly chosen from the remaining time interval using uniform distribution. If the model initially predicted the reservation not to cancel, the “prediction” will be of course changed at the time of cancellation.

The total revenue of the flight is calculated using ticket prices R_{low} and R_{high} as well as the penalty D for denied boardings. The denied boardings are assumed to be low fare seats.

The simulation consists of three nested loops. The outermost loop changes the model accuracy P_{model}. The second loop iterates over all possible values of x. The innermost loop simulates the booking process thousands of times to create reasonable result for expected revenue. The reference method simulation is highly similar, only the outermost loop changing the overbooking limit L.

The constants used in the simulation can be seen in the table below. Any of them will not decrease or increase during the booking process, which is once again a simplification.

seats	100
cancellation rate	30%
λ_{low}	0.1
λ_{high}	0.05
λ_{size}	0.1
T_{low}	100
T_{high}	50
ticket price R_{low}	50
ticket price R_{high}	100
penalty D	150

Table 6: Parameter values used in the simulation.

Final simulations used 100 000 iterations over each x. Before that, some approximate simulations were made to find the most interesting ranges for x, P_{model} and L. The ultimate values used are presented in table 7.

	min	max	interval
x	40	100	1
P_{model}	0.7	1	0.02
L	120	160	5

Table 7: Final values used in the simulation.

Results

The expected revenue over x for each P_{model} simulated can be seen in figure 8..
The curves are in descending order with respect to P_{model} meaning that the uppermost curve represents the accuracy of 100% and the lowermost stands for P_{model} = 0.7. It can be seen that with higher P_{model} the optimal x is also larger.

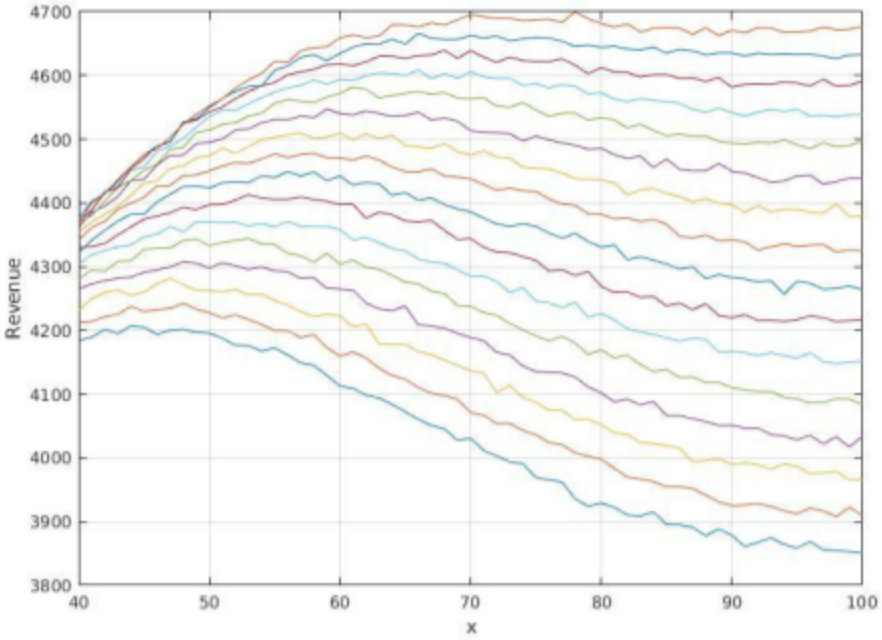

Figure 8: Expected revenue as a function of booking limit x. Each curve corresponds to a single forecast accuracy.

The corresponding plot for most optimal overbooking limits is given in figure 9. Plotting the maximum revenue of each overbooking limit generates the graph in figure 10. Based on this figure it can be seen that the maximum revenue using overbooking limit seems to be about 4360. A similar graph is plotted for model accuracy in figure 11. The revenue of 4360 is exceeded with the model accuracy of 80%. Therefore, based on this simulation, this is the accuracy when model becomes profitable compared to the fixed overbooking limit.

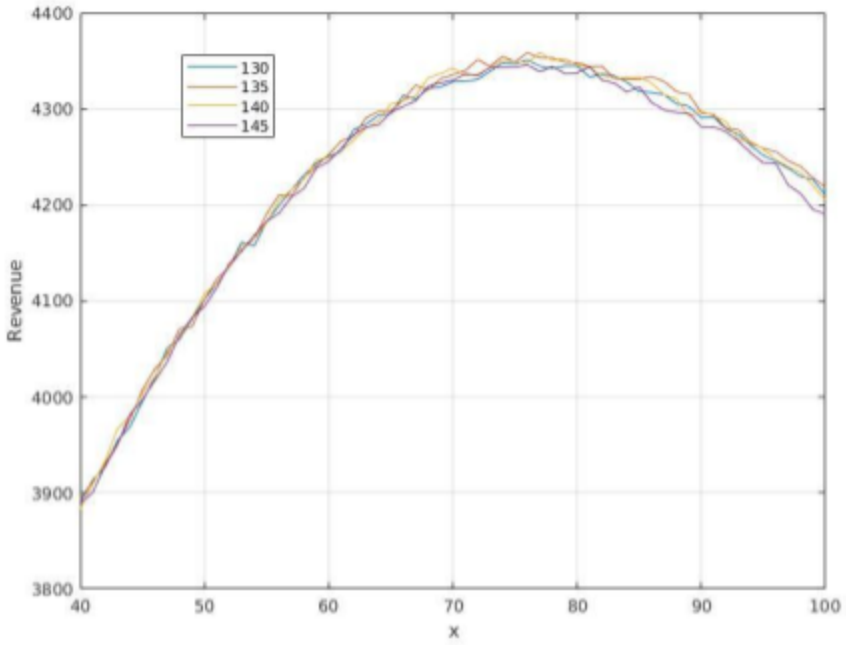

Figure 9: Expected revenue as a function of the booking limit x for each overbooking limit.

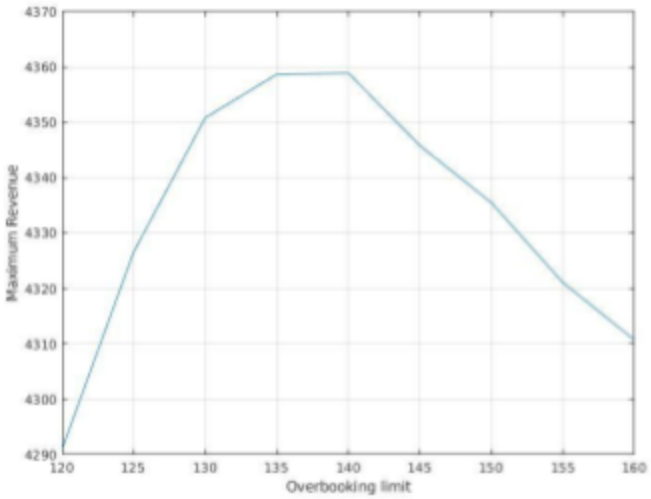

Figure 10: Maximal revenue as a function of the overbooking limit.

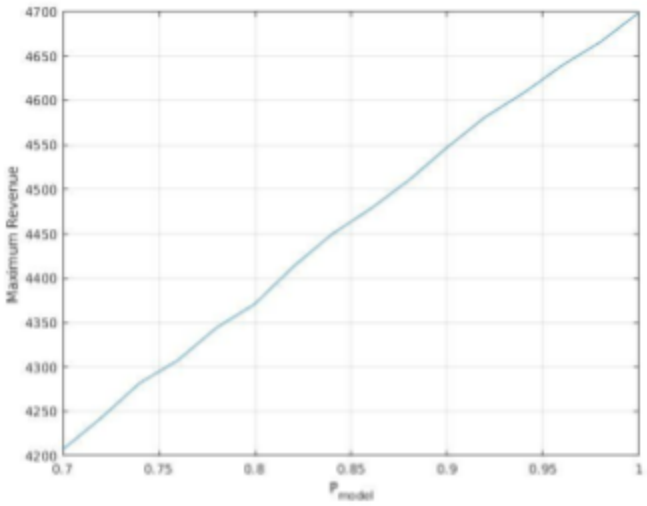

Figure 11: Maximal revenue as a function of forecast accuracy.

Discussion

The results of the simulation suggest that in order to be profitable compared to the traditional overbooking models, the prediction accuracy of the cancellation model needs to be considerably high. This is most likely explained by the high penalty of denied boardings. If the model falsely predicts a large group to be cancelled, the costs of denied boardings will quickly increase, while the traditional overbooking method does not involve a similar risk, as it only books seats up to a given capacity.

One of the main simplifications of the simulation is that none of the parameters are time-dependent. Some of the cancellations occur a significant time before the flight, and the seats thus cancelled could be sold again without the need to worry about the original group. This approach should lead to higher revenues due to less uncertainty. In this simplified model, we try to predict the behaviour of some groups over the period of one year, without updating the prediction as time goes by. If a different approach taking this into account would be used, we could have time left until flight as a feature in the prediction, and it would have an effect on the accuracy of the model, based on intuition. For a booking that has survived until a month to the flight, the cancellation probability

would be smaller than for a similar booking with six months until flight, as the latter would have 5 more months to cancel.

Another point about time-dependency is that the ticket prices in the simulation are constant for a fare class, while in reality the prices would change in time. As the flight date approaches, prices should not go up too much, as that would result in the risk of the tickets not being sold and the plane flying below the capacity. On the other hand, as the number of available seats goes down, prices should go up. This is implemented in our model by first selling the low fare class, then the higher class, but overall, the prices could definitely be simulated in more detail.

Another factor that should be taken into account when evaluating the simulation is that many of the parameters, such as the demands are constant over the simulation period. In reality, they would change in time, and this could have an effect on the results. It is however difficult to tell what the effect would be, and that will therefore not be speculated in this report.

Conclusion

Our models were able to predict cancellations decently, even though a similar overall prediction error percent of roughly 80% could have been reached by classifying all flights as not cancelled. The main problem, however, was predicting the cancellations. The models were able to classify cancelled flights with an accuracy of under 60%, while the simulation suggests that a 80% accuracy would be required. Even though the accuracy seems low, the performance in non-cancelled flights suggests that a similar performance could not be achieved simply by randomly classifying bookings with any probability of a flight classifying as cancelled. A possible next step would be to combine the two simulations, finding an optimal booking limit for different prediction accuracies. This could result in higher expected revenues for lower accuracies, but it has to also be noted that the uncertainty of such models would be relatively high. The model is not something Finnair should directly use in their business, but it provides a base idea for expanding further.

The obvious next step would be to allow the prediction to change after the initial booking, as more information about the booking becomes available. This would lead to the prediction accuracy rising as the flight approaches, thus allowing the airline to make

informed decisions on how many seats to keep open for sale. Another possible next step would be to sit down and think about what features to use now that we have a slightly better understanding on the topic. Feature selection is extremely important in prediction and classification problems, as missing features result in poor prediction ability, and unnecessary features might confuse the method and result in overfitting. In addition to these questions, multicollinearity should be examined further to prevent cases where the model tries to predict based on a feature dependent on the actual significant feature, resulting in high variance in the parameter estimates and possible overfitting.

References

[1] Talluri, K. T., & Van Ryzin, G. J. (2006). *The theory and practice of revenue management* (Vol. 68). New York: Springer Science & Business Media.

[2] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[3] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[4] Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival analysis part I: basic concepts and first analyses. *British Journal of Cancer*, *89*(2), 232.

[5] Littlewood, K. (2005). Special issue papers: Forecasting and control of passenger bookings. *Journal of Revenue and Pricing Management*, *4*(2), 111-123.

Self-assessment

In the original project topic description three objectives were listed:

- Build a predictive model for cancellation of group bookings

- Estimate denied boarding risk as a function of cancellation forecast error, group size and group instances

- Propose a simplified overbooking model minimizing the denied boarding risk while maximising flight revenue

While we did acquire significant results in terms of achieving these objectives we did not manage to accomplish all of them. The first objective, building a predictive model for cancellation of group bookings, was prioritized and satisfactory results were achieved. While the model built may not be useful in practice it may be viewed as a proof of concept which demonstrates that it is able to predict cancellations based on the provided features with a decent accuracy. The second objective was simplified somewhat as the project went on. The implemented simulation may not be as ambitious as originally hoped but it did yield some interesting results which indicate that the achieved accuracy might be sufficient. However this needs to be taken with a grain of salt as the simulation was a gross simplification of the true booking process. The third objective was not achieved. The reasons have to do with difficulty of the topic and problems with project coordination.

The project topic proved to be surprisingly complex. Firstly the literature related to airline revenue management and overbooking in specific is vast and despite investing a significant amount of time on studying these topics no thorough understanding of the subject was achieved. This proved to be a problem when programming the simulation as nobody felt like they truly understood the process being simulated. Thus multiple questionable assumptions and simplifications were made in order to make the task more approachable. The messiness of the provided data also cause problems and consumed

a lot of time. In addition to the usual “messiness” a few clear errors were also discovered.

The project execution left something to be desired. There were some problems with communicating with the client which caused some time to be lost. Distributing the workload effectively for three people proved to be more difficult than expected because the project objectives were sequential in nature. That is in order to start the simulation you need to understand the modeling part and you need to process the data before you can fit models on it. None of these tasks were easily parallelizable which in practice lead to one person working on the current problem while others waited for the results in order for them to be able to carry on with the next task. In hindsight some of these problems could have been avoided with more careful scheduling and planning.