MS-E2177 Seminar on Case Studies in Operations Research 2017

# Predicting and managing group cancellations in air passenger traffic

Client: Finnair

# Interim report

April 11, 2018
Tuomas Koskinen (project manager)
Elias Axelsson
Olli Herrala

## Changes in objectives and scope

No major changes in the objectives or scope of the project have happened. The main goal of the project is the development of a satisfactory model for predicting group cancellations and this is well underway. The next step is to develop an overbooking management system based on the model. We are currently in the process of clarifying the methodological approaches which will be used to achieve this goal. The initial plan was to program a simulation and this is still the primary plan but other options need to be considered.

## Project status in relation to the initial project plan

The current status of the project is somewhat behind schedule compared to the original project plan. So far we have carried out a literature review of airline revenue and management and done some exploration and simplified modelling of the provided data. Tasks defined in the original project plan are listed below with steps taken to complete them.

1. **Literature review**

Relevant literature has been studied as needed. In addition to literature related to airline revenue management it has been necessary to familiarize ourselves with some literature concerning  predictive modeling. We still need to do some additional reading so this task is ongoing.

## 2. Exploring the data

This task is completed. We now have a good idea of what the data contains. Modalities of the categorical variables and their distributions have been examined. The information gained via exploratory analysis of the data was taken into account in modeling.

## 3. Data preprocessing

A lot of preprocessing was needed to fit any model on the data. Each row of the data is a single reservation/cancellation event identified by a reservation and flight number. The first problem was to define what exactly it is that we are trying to predict. The problem was simplified as much as possible. Most reservations are either cancelled completely or all initially reserved seats are used. However in some cases only some of the initially reserved seats are cancelled. In order to frame this problem as a classification problem (cancelled or not cancelled) each reservation is assigned either a “cancelled” or a “not cancelled” - label based on whether over 50% of the initially reserved seats were cancelled. This simplification causes some information to be lost but it is used as a first approach.

After this all the individual events needed to be appropriately aggregated into rows corresponding to individual reservations. Finally all the categorical variables needed to be converted into “one-hot” encoding. Preprocessing was done using pandas library available for Python. As we continue to explore different models additional processing of the data is necessary.

## 4. Model fitting

Different models were fitted to the simplified data. These include logistic regression and a random forest classifier. Models implemented in scikit-learn Python library were used. We are still trying to find better models.

## 5. Model testing and comparison

The data is sorted into chronological order and the last 20% is reserved for model validation. This is equal to 14 000 reservations. The data is sorted into chronological order by reservation date, the validation set consists of the last 20% of reservations. The goal is to simulate a real life scenario, where new reservations are predicted using data from previous ones. On the validation set the logistic regression model achieved a 78% accuracy while the random forest model performed better achieving an accuracy of 84%. Other models will be validated similarly.

# Changes to the initial project plan

Due to sickness and other projects consuming more time than expected, our group has fallen behind the planned schedule. The situation is not too serious as of yet, as our other projects will be finished in the near future, but we have to acknowledge the fact that our schedule has become quite tight. Finishing the literature review has been given three additional weeks to get a broader view of the topic. Model fitting, testing and comparison might still require substantial work, since we might change our approach. The simulation and overbooking model have also been postponed slightly because of the delay in model fitting.
A schedule for the tasks we have yet to complete is given below:

| Task \ Week | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|
| Literature review | 🟧 | 🟧 | 🟧 | | |
| Model fitting | 🟧 | 🟧 | 🟧 | | |
| Model testing and comparison | 🟧 | 🟧 | 🟧 | 🟧 | |
| Programming the simulation | | | 🟧 | 🟧 | |
| Formulate an overbooking model | | | 🟧 | 🟧 | 🟧 |
| Final report | | | | 🟧 | 🟧 |

# Updated risk management plan

Below you can find our updated risk management plan. The major change compared to the initial plan is that the risk of being late on schedule has increased due to the limited time remaining. Thus we are also more likely of having unsatisfying model in the end. On the other hand the probability of member dropout has decreased.

| **Risk** | **Probability** | **Effects** | **Impact** | **Mitigation measures** |
|---|---|---|---|---|
| Member inactivity or dropout | Low | Increase in the workload of other members | Low to high | Transparency in scheduling |
| Inability to stay on schedule | High | Workload grows large towards the end and implementation may | Low to high | Frequent meetings between team members |

| | | | | |
|---|---|---|---|---|
| | | remain incomplete | | |
| Weak communication with customer | Moderate | Implementation proves to be unsatisfying | Moderate | Frequent meetings with customer |
| Data proves to be too messy | Low | Inability to build a model | High | Active communication with customer |
| Inability to build a reasonable model | Moderate | Little or no value creation for customer | High | Studying the subject carefully and scoping with customer and course staff |
| Model doesn’t satisfy customer needs | Moderate | Low value creation for customer | Moderate | Model comparison and frequent meetings with customer |