Optimizing the prices for airline flight passes

Santos, Bruno F.; Gillis, Myrthe M.D.

DOI
10.1016/j.trpro.2018.12.192

Publication date
2019

Document Version
Final published version

Published in
Transportation Research Procedia

Citation (APA)

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.
Abstract

Flight pass is a new concept in which airline passengers pre-purchase a number of flights for a flat fee. This flat fee can be customized by the passenger and has an expiring date. Being an innovative concept, the industry is still lacking analytical support to define the prices for these passes. This research is aimed to fill this research gap. We propose a data-driven modeling framework that determines the value of each option per flight and that, consequently, estimates the recommended flight pass price. This is the first time in the literature that flight passes are discussed. The framework is divided in two models. First, a random forest regression is used to predict the ticket price of individual flights. Second, the flight pass prices are predicted using a Monte-Carlo simulation. The simulation is used to estimate the potential behavior of a passenger when using the flight pass. Since no reliable flight pass data was available, we make use of historical booking data from revenue management available from a major African airline to design, calibrate, and validate our models. With an average fit of 57 percent, the random forest regression algorithm can adequately predict the flight price, improving the current trial-and-error or linear regression approaches followed by most airlines. Moreover, the Monte-Carlo simulation is fast enough to support an online implementation of the proposed modeling framework to estimate flight pass prices.

Keywords: Airlines flight passes; random forest regression; revenue management; Monte-Carlo simulation.

1. Introduction

Airlines encounter challenges with respect to satisfying both the customer satisfaction and fidelity. One of the recent solutions offered by airlines are the flight passes. These are service-oriented products offered at a fixed fee price, just like a health insurance package or a public transportation pass. The concept of the flight pass is based on the fact that customers pre-purchase a number of flights for a flat price, in function of a set of options chosen by the customer. These options may include the number of flights, the travel period, how early the flights can be booked, which cabin is used, which routes are included, or the number of people that can make use of the flight pass. The price of the flight pass depends on the options chosen. By using this concept, the airline company wants to give flexibility to its customers pre-purchase a number of flights for a flat fee, in function of a set of options chosen by the customer. These options may include the number of flights, the travel period, how early the flights can be booked, which cabin is used, which routes are included, or the number of people that can make use of the flight pass. The price of the flight pass depends on the options chosen. By using this concept, the airline company wants to give flexibility to its customers.
passenger, increasing both customer satisfaction and fidelity, without compromising the capability to generate revenue in a highly competitive environment.

The flight pass is a rather new pricing structure. The practice started in 2015, but there are already several airlines making use of it, namely: Air Asia X, Air Canada, Air France, British Airways, Eurowings, Kenya Airways, KLM, Oman Air, and Vietnam Airlines. Currently, the flight pass pricing followed by these airlines is based on pricing controllers experience and market knowledge (i.e., following a trial-and-error approach) or in linear regression analysis of the influence of the options chosen by the customer on the price. These are simple approaches to solve the complex problem of determining the right price for flight passes. The flight passes involve many possible user configuration options, which do not necessarily have a linear and independent impact on the determination of a pass value that is, at the same time, attractive enough for the costumer and covers (most part of) the costs of the potential pass usage.

Therefore, the objective of this research is to develop a data-driven methodology to better determine flight passes prices as a function of the options selected by the customer. This is the first time that the concept of airline flight passes is discussed in the literature. Other two innovative aspects of this research include the use of a Random Forest Regression (RFR) algorithm (Breiman, 2001) to infer the importance of key features in the airline ticket price and the combination of this technique with Monte-Carlo simulation for pricing. The context of Kenya Airways (KQ), who started to sell the flight pass in 2017, is used as a reference case. Since the flight pass is a new concept, no useful data is available yet. However, historical booking data from revenue management was made available by the airline.

The remaining part of this paper is organized as follows: Section 2 describes the modeling framework proposed. Section 3 describes the Kenya Airways case study, while Section 4 reports results obtained. In this later section the performance of the RFR algorithm is compared with the performance of the traditional Multiple Linear Regression Analysis (MLRA) (Allison, 1999) used in practice to determine flight passes prices. Section 5 summarizes the main conclusions from this research and discusses future research opportunities.

2. Modeling Framework

The data-driven modeling framework proposed is based in two models. The first model comprises the computation of the estimation of the ticket price per individual flights, based on a set of features from the flight booking. The second model concerns the simulation of the passenger behavior when using the flight pass, determining a flight pass price that reflects the expected usage of the flight pass. Each one of the models is explained in the next subsections. The section is concluded with an explanation of the algorithms training and validation procedure.

2.1. Flight price

There are already several airline revenue management and pricing studies making use of data-driven techniques (e.g., Weatherford 2003; Martens 2011; Chen 2015; Ostaijen 2017). However, few studies consider the usage of RFR algorithms in airline applications. In this study we propose this algorithm following the conclusions from several studies that have proved the better performance of RFR algorithms over MLRA (e.g., Pow 2014; Borde 2017; Moosavi 2017).

Breiman introduced the general concept of a random forest in 2001 (Breiman, 2001). The RFR algorithm followed up the Classification & Regression Trees (CART) (Breiman, 1984) and the bootstrap aggregation (Breiman, 1996) algorithms developed by the same author. An RFR algorithm operates by constructing several parallel decision trees with different subsets of data (i.e., bootstrap aggregation). To create some variability between the different decision trees and to reduce the risk of overfitting, the subsets are drawn at random from the full training set. The different decision trees are combined and averaged when performing a regression or a classification. To illustrate the RFR algorithm, Figure 1 shows the decision tree that predicts the price of flight depending on only 2 input features. The first feature (‘Feature_1’) is a continuous variable, while the second feature (‘Feature_1’) is a categorical feature, with a value of 0 or 1.

A regression tree is built by recursively partitioning the sample (i.e., the “root node) into more and more homogeneous groups, so-called nodes, down to the “terminal nodes. The partitioning is called a split. Each split is based on the values of one feature and is selected according to a splitting criterion. In this work we consider the splitting
criterion to be the mean squared error (MSE), as defined in Equation 1 in which $y_{i<\delta}$ and $y_{i\geq\delta}$ are the mean value for the left and right branch of the split.

$$
\min_{s} \sum_{i:x_i<\delta} (y_i - y_{i,x_i<\delta}) + \sum_{i:x_i\geq\delta} (y_i - y_{i,x_i\geq\delta})
$$

The model takes into account all possible splits for each feature and it chooses the feature and the split that minimizes the MSE. The stopping condition is a pre-defined parameter, which can be either the minimum samples per split or the minimum samples per node.

There are two important advantages of using RFR, when comparing it with linear regression techniques. The first one is that RFR is not a probabilistic technique, but a binary split. This means that most of the well-known linear regression assumptions, like linear dependence between input features and dependent variable, no significant outliers exist, and no multicollinearity is observed, can be relaxed. The second advantage is that decision trees are non-parametric techniques. Trees are represented by a string of choices and decisions (Figure 1).

The RFR algorithm is defined by a set of hyper-parameters, which defined the structure of the trees included in the “forest” and of the “forest” itself. The values of these hyper-parameters highly influence the performance and accuracy of the RFR algorithm. We considered the following hyper-parameters for the study of the flight passes, with the respective sets of values considered:

- the number of trees in the forest [10, 50, 100, 200],
- the number of features to consider when looking for the best split [#all, sqrt(#all), log^2(#all)] (see Louppe 2014),
- the maximum depth of the tree [0, 10, 15, 20, 25],
- the minimum number of samples required to split an internal node [2, 4, 6, 8, 10],
- the minimum number of samples required to beat a leaf node [1, 2, 5, 10].

The values for those parameters were calibrated when the model was trained. This problem of optimizing the hyper-parameters is called tuning. We used a ‘grid search’ approach to tune the RFR algorithm. This is a traditional way of tuning such type of machine learning techniques, which exhaustively considers all parameter combinations pre-specified as sets and bounds. In this research, the coefficient of determination was used to evaluate the performance while tuning the algorithm.

![Decision tree example: predicting the ticket price value only using two features](image-url)
2.2. Flight passes price

To estimate the flight passes price, it was assumed that the offered price should be lower than the maximum sum of prices for the flights that the customer can book with the pass. It is easy to understand that a strategy in which the airline would estimate the flight pass price based on a worst-case scenario would be non-attractive at all for the customer. Therefore, it is considered that the airline is willing to take the risk of in N% of the times the flight pass price is lower than the sum of the prices of the realized customer bookings. This risk can be seen as the cost for the airline to guarantee the customer’s loyalty.

Since a customer can book different flights with one pass, we propose a simulation approach to estimate the distribution of the fares of the flights booked by a random customer. A Monte-Carlo simulation was used to simulate the booking behavior of a passenger (i.e., for instance, the duration of stay in a round trip, the day of the week for the flights, the number of days between booking and flying). The booking behavior of a random customer can be estimated by analyzing historical revenue data, ideally from flight passes usage or from airlines’ frequent flyer programs data. In addition, it was assumed that the costumer will make the best use of the flight pass. For instance, booking the flights close to the number of days between booking and flying limit. For each flight pass estimation, the Monte-Carlo simulation is run several times with a different set of random values. The outcome of the Monte-Carlo simulation is a probability distribution obtained from simulation of thousands of generated random customers. The suggested flight pass price is then estimated by extracting the price value that is not exceeded more than 100 – N% of the times.

2.3. Data sets and cross-validation

To properly train and validate the RFR algorithm, we divided the data provided by the airline into three data sets:

- *Training set*: used to train the algorithm.
- *Validation set*: used to tune the hyper-parameters.
- *Test set*: used to assess the performance of the model

Training and validation data sets were used to train and tune the algorithm, respectively, by minimizing the error between the individual flight prices estimated and the observed flight fares. The test data set was then used to evaluate the resulting model with observations not used during the calibration and tuning. A random sample of the data was used to obtain these three data sets. For the validation set, a cross-validation technique was used (see, e.g., Kohavi 1995). The algorithm is fitted on different training sets separately, and afterwards tested against the full test set. In this research a 10-fold cross validation is used, where the process is repeated 10 times.

3. Case Study - Kenya Airways

Kenya Airways (KQ) is the national carrier of Kenya and part of the Sky Team alliance. The airline operates a hub and spoke network with Jomo Kenyatta International Airport (NBO), located in Nairobi, as its main hub.

The flight pass is available since 2017 at KQ, and it includes every flight operated by the airline, with exception of the flights operated to Europe. The prices are determined with a route-independent linear model. This means that the same model is used in the case of the thrice per week intercontinental flights to Shanghai and in the case of the national route between Nairobi and Mombasa with more than 10 flights per day. In the same way, no distinction is made between days of the week in which the customer wants to fly and the price of the flight passes linearly increase with the increase number of days in advance in which the flights need to be booked.

3.1. Data available

Since the flight pass option is still recent, no reliable data was available when this research was initiated. Therefore, the data used for this work was obtained from the revenue management systems at KQ. It was assumed that the revenue management data obtained is representative of the willingness to pay of KQ customer under circumstances equivalent to the options to be considered for the flight passes. That is, for instance, that for a specific flight, the average difference
observed between case of booking 7 days in advance versus 14 days in advance is representative of the added value of changing the flight pass option from minimum booking time of 2 weeks to a minimum booking time of 1 week before the flight.

Some data specifics need to be mentioned:

- The data extracted is from August 2016 till July 2017 (i.e., 1 year of data).
- Only point to point passengers are considered, since data referring to connecting passengers did not have enough detail to be considered.
- Only flights considered in the current airline’s flight pass configuration are considered. These are 46 point to point routes departing or arriving in NBO.

A short example of the data obtained, after being filtered and processed, is shown in Figure 2. The one year data had to be split between training data (i.e., training and validation data sets) and testing data. A split of 70% and 30%, respectively, was considered for this study.

3.2. Input features

The input features were selected according to the data available and to the options being considered by the airline when defining the flight passes. This way, the following seven features were included in the RFR algorithm:

- **Point of sale (POS)**, divided in three categories - Kenya; other country of origin/destination; and other countries.
- **Length of stay (LOS)**, divided in three categories - less than 3 days; between 3 and 7 days; and more than 7 days.
- **Sunday stay (SUN)**, divided in two categories - included and not included in the stay.
- **Time-of-departure (TOD)** of the flight, divided in the twenty-four hours of the day and transformed into a cyclic variable in which hour 23 is as closed to hour 0 and as to hour 22.
- **Day-of-the-week (DOW)**, divided in the seven days of the week and transformed into a cyclic variable.
- **Month of the flight (Month)**, divided in the twelve months of the year and transformed into a cyclic variable.
- **Ticketing leading time (TLT)**, a continuous variable describing the number of days before departure of the booking.

<table>
<thead>
<tr>
<th>City</th>
<th>TOD</th>
<th>Cabin</th>
<th>Month</th>
<th>DOW</th>
<th>POS</th>
<th>Sunday stay</th>
<th>LOS</th>
<th>TLT</th>
<th>RevPerPas</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAN</td>
<td>21:40</td>
<td>Economy</td>
<td>8</td>
<td>Fri</td>
<td>Org/Dest</td>
<td>Yes</td>
<td>7</td>
<td>8</td>
<td>380</td>
</tr>
<tr>
<td>LVI</td>
<td>18:15</td>
<td>Economy</td>
<td>8</td>
<td>Wed</td>
<td>Other</td>
<td>Single Ticket</td>
<td>Single Ticket</td>
<td>36</td>
<td>470</td>
</tr>
<tr>
<td>JNB</td>
<td>21:15</td>
<td>Economy</td>
<td>9</td>
<td>Mon</td>
<td>Org/Dest</td>
<td>Yes</td>
<td>3</td>
<td>3</td>
<td>260</td>
</tr>
<tr>
<td>JNB</td>
<td>01:15</td>
<td>Economy</td>
<td>9</td>
<td>Fri</td>
<td>Org/Dest</td>
<td>Yes</td>
<td>3</td>
<td>3</td>
<td>260</td>
</tr>
<tr>
<td>CAN</td>
<td>23:59</td>
<td>Economy</td>
<td>10</td>
<td>Thu</td>
<td>Kenya</td>
<td>Yes</td>
<td>7</td>
<td>3</td>
<td>380</td>
</tr>
<tr>
<td>JNB</td>
<td>12:55</td>
<td>Business</td>
<td>3</td>
<td>Sat</td>
<td>Kenya</td>
<td>Single Ticket</td>
<td>Single Ticket</td>
<td>23</td>
<td>1680</td>
</tr>
<tr>
<td>CAN</td>
<td>22:10</td>
<td>Economy</td>
<td>8</td>
<td>Sat</td>
<td>Org/Dest</td>
<td>Yes</td>
<td>7</td>
<td>1</td>
<td>615</td>
</tr>
<tr>
<td>TNR</td>
<td>11:10</td>
<td>Economy</td>
<td>12</td>
<td>Sat</td>
<td>Other</td>
<td>Single Ticket</td>
<td>Single Ticket</td>
<td>15</td>
<td>630</td>
</tr>
<tr>
<td>CPT</td>
<td>14:35</td>
<td>Economy</td>
<td>5</td>
<td>Sat</td>
<td>Org/Dest</td>
<td>Yes</td>
<td>7</td>
<td>1</td>
<td>240</td>
</tr>
<tr>
<td>MPM</td>
<td>11:05</td>
<td>Economy</td>
<td>11</td>
<td>Mon</td>
<td>Kenya</td>
<td>Single Ticket</td>
<td>Single Ticket</td>
<td>6</td>
<td>530</td>
</tr>
<tr>
<td>CAN</td>
<td>21:40</td>
<td>Economy</td>
<td>8</td>
<td>Fri</td>
<td>Org/Dest</td>
<td>Yes</td>
<td>7</td>
<td>10</td>
<td>430</td>
</tr>
<tr>
<td>KGL</td>
<td>07:35</td>
<td>Economy</td>
<td>8</td>
<td>Fri</td>
<td>Kenya</td>
<td>Yes</td>
<td>7</td>
<td>3</td>
<td>440</td>
</tr>
<tr>
<td>CAN</td>
<td>22:10</td>
<td>Business</td>
<td>7</td>
<td>Tue</td>
<td>Org/Dest</td>
<td>Yes</td>
<td>7</td>
<td>48</td>
<td>1120</td>
</tr>
<tr>
<td>CAN</td>
<td>23:45</td>
<td>Business</td>
<td>8</td>
<td>Wed</td>
<td>Org/Dest</td>
<td>No</td>
<td>7</td>
<td>21</td>
<td>890</td>
</tr>
<tr>
<td>MPM</td>
<td>11:05</td>
<td>Economy</td>
<td>11</td>
<td>Wed</td>
<td>Other</td>
<td>Single Ticket</td>
<td>Single Ticket</td>
<td>3</td>
<td>630</td>
</tr>
<tr>
<td>JUB</td>
<td>10:20</td>
<td>Economy</td>
<td>8</td>
<td>Sat</td>
<td>Kenya</td>
<td>Single Ticket</td>
<td>Single Ticket</td>
<td>6</td>
<td>305</td>
</tr>
<tr>
<td>BGF</td>
<td>14:45</td>
<td>Economy</td>
<td>1</td>
<td>Sun</td>
<td>Kenya</td>
<td>Single Ticket</td>
<td>Single Ticket</td>
<td>2</td>
<td>690</td>
</tr>
</tbody>
</table>

Fig. 2: Extract of the data obtained from Kenya Airways, after being filtered and processed (RevPerPas: Revenue per Passenger)
when defining the flight passes. This way, the following seven features were included in the RFR algorithm:

3.2. Input features

respectively, was considered for this study. Had to be split between training data (i.e., training and validation data sets) and testing data. A split of 70% and 30%,

the flight. Observed between case of booking 7 days in advance versus 14 days in advance is representative of the added value of

The input features were selected according to the data available and to the options being considered by the airline

A short example of the data obtained, after being filtered and processed, is shown in Figure 2. The one year data

Ticketing leading time (TLT)

Month of the flight (Month)

Day-of-the-week (DOW)

Time-of-departure (TOD)

Sunday stay (SUN)

Length of stay (LOS)

Point of sale (POS)

- Only point to point passengers are considered, since data referring to connecting passengers did not have enough

- The data extracted is from August 2016 till July 2017 (i.e., 1 year of data).

booking.

, a continuous variable describing the number of days before departure of the

variable in which hour 23 is as closed to hour 0 and as to hour 22.

variable in which hour 23 is as closed to hour 0 and as to hour 22.

, divided in two categories - included and not included in the stay.

, divided in three categories - less than 3 days; between 3 and 7 days; and more than 7 days.

point routes departing or arriving in NBO.

detail to be considered.

Fig. 2: Extract of the data obtained from Kenya Airways, after being filtered and processed (RevPerPas: Revenue per Passenger)

3.3. Modeling approach and validation

KQ operates routes with different characteristics. It is expected that features have a different impact for domestic, continental, and intercontinental routes. Therefore, it was decided to develop separate models for each route and for each cabin class. This means that there are two models for each route, for business and economy cabins. When combining this with all the 46 routes considered, this results in 92 (46 routes x 2 cabins) models. Each one of these models have different decision trees and calibrated hyper-parameters.

4. Results

This section summarizes the results obtained for the case study from KQ, illustrating the results that can be obtained with the methodology proposed.

4.1. Flight price predictions

The RFR was compared with a multiple linear regression analysis (MLRA) algorithm. The former algorithm is used as a benchmark for the RFR, since it is the current state-of-the-art for flight passes price estimation. Both algorithms are compared based on two goodness of fit metrics, namely the coefficient of determination (also known as $R^2$) and the root mean square error (RMSE).

The random forest regression performs best for every of the 46 routes in terms of both goodness of fit - i.e., the coefficient of determination is higher and the RMSE value is lower. An aggregated summary of results, for all routes, are presented in Figure 3.

As one can see, RFR algorithm has a much better fit than the MLRA. The MLRA can predict 30.4% of prices correctly. However, the RFR algorithm significantly outperforms the other algorithm with a coefficient of determination of 0.56. Taking into account the human input in the determination of prices and the dependency on demand, one could say that an expected ‘good’ fit for this problem would have $R^2$ values between 0.4 and 0.8 (Taylor, 1990). The prediction results when using the RFR algorithm are for most routes within this range. Therefore, one can conclude that the RFR algorithm is a suitable prediction model for this problem.

The RMSE gives high weights for bigger errors, which is especially helpful to identify cases where particularly bad predictions take place. Again, the RFR clearly outperforms the MLRA. It can be inferred from these results that the data behaves non linear and that (multiple) linear regression is not the most suitable method to predict flight prices.

4.2. Features’ importance

Next to the prediction, the random forest regression shows the importance of the features in predicting the price. Figure 4 shows the average feature importance of all routes, where the addition of the different importances adds to
1. It can be read from this figure that, on average, the ticketing lead time explains 26.2% of the variability of the ticket prices, the month of the flight explains 23.9%, the day of week explains 11.4%, and so on. This way, it can be concluded that, for the aggregated analysis of the 46 routes from KQ, the most important features in predicting the ticket price are ticketing lead time and month of the flight.

4.3. Flight passes price estimations

The Monte-Carlo simulation was used to simulate the behavior of a passenger taking into account the distribution of the features in the revenue data provided by KQ. For each case tested, a total of 10,000 simulations were performed. These simulations can be computed in a matter of one or two seconds, supporting the idea that the proposed data-driven approach is suitable for online flight passes pricing.

The prediction of the price for two cases is shown in Figure 5. The y-axis represents the amount of times the prices occur. The x-axis represents the price of the flight. The figure presents a case in which a single route is considered and just a small set of the flight pass options are activated by the customer (Figure 5a); and a case in which multiple routes are included in the flight pass and several other options are activated (Figure 5b). The flight pass volatility is much higher for the second case. The red lines define, for both case, the flight passes prices that should be considered if the airline would like to take a 20% risk of having a flight pass price lower than the sum of the flight fares of the flights booked by the passenger.
5. Conclusions

A data-driven modeling framework was proposed in this work to estimate airline flight passes prices, based on the flight pass options selected by the customer. The modeling framework comprises a random forest regression algorithm which outperforms the multiple linear regression analysis currently used by airlines to compute the prices of these passes. Therefore, it can be concluded that airlines would benefit from using data-driven techniques, obtaining more accurate and route dependent estimations for the flight passes. Additionally, the algorithm also gives insights into the importance of the features in predicting the flight prices. This is a relevant information for the airline when defining (standard) single flights and flight passes pricing strategy. For the case study considered, the ticketing lead time was found to be the most relevant feature.

This was the first study done on the topic of airline flight passes. Future steps are necessary to consolidate this research. The most important one would be the validation of the results obtained with data-driven approaches. Since there is no flight pass booking data was available yet, no valid validation could be done. Furthermore, it would be interesting to consider training data from multiple years in order to better capture seasonality effect in the computation of the flight passes prices.

Acknowledgements

The authors thank the collaboration from Kenya Airways during this research. In particular, the authors are thankful to Dr. Marco van Vliet for the knowledge he shared and for his critical comments.

References