Evaluating the sudden change in flight cancellations in Paris during November 2015

Gallego Andreu, C.; Gironés Sangüesa, R.; Grau Gil, J.; Lillo Collado, L.; Losa Brito; A.; Romero Alvarado, D.

Escola Tècnica Superior d'Informàtica - Universitat Politècnica de València

Abstract

The air travel industry is a really volatile market, where sudden events can lead to unexpected results in flight transactions. Companies of the sector must be ahead of the curve in predicting and understanding the impact of these events.

ForwardKeys, a data collection and analysis company, has provided us data about flight transactions in Paris during November 2014 and 2015. In this study, we use association rules to discover the main repeating patterns among the data, specially those who are related to cancellations, to better explain and outline the clients profile of the market.

Furthermore, to better understand what defines the profile of a client, a linear discriminant analysis and a decision tree are performed to determine importance to attributes of our dataset.

Keywords

Terrorism; Commercial Flights; Airline Tickets; Pattern Recognition; Passengers Profile

Background

The analysis of air travel transaction data is useful, not only for airlines but also for everything that surrounds them, such as the hotel or retail sector or other marketing or tourism companies. With the help of such data and good analysis, businesses will be able to react to special situations. One example of a company that focuses on this type of data is ForwardKeys, a specialist in travel intelligence. From this company, they are able to obtain information from data on airline ticket transactions. In these observations (transactions), reference is made to traveller profiles.

The company's purpose is to provide its customers (other companies, usually in the previously mentioned sectors) with knowledge about the flights they get by answering questions such as where, when, for how long and who travels. In addition, they try to study the behaviour and preferences of people who use aeroplanes as a means of transportation.

This company has provided us with some data that we will work with during this project. More specifically, the data deals with air transactions made in November 2014 and November 2015. The question we seek to answer, in a very general way, is what happened in 2015 for there to be irregularities between these two months. Thus, our first action was to look for what happened in Paris in November 2015, since the data provided focuses on this city. As expected, we found an event that considerably changed air traffic in Paris. On November 13, 2015, the terrorists attacks that shocked the city of Paris took place.

At first sight, we could see an increase in flight cancellations in November 2015 during the second half of the month. Given that the coincidence of the date of the terrorist attack with the first peak in cancellations is quite obvious, this abnormal behaviour has been attributed to the attacks. The impact of such events on the air transport sector is not a widely studied topic, although some studies show that most of the costs in these situations are cancelled and delayed flights (Janić, 2015).

The following is a project that aims to study the data in depth so that we can obtain detailed information on cancellations. Additionally, and thanks to the use of association rules, we may be able to extract common trends and patterns during both years.

Data

The data employed throughout this project was provided directly by Forwardkeys. There are two datasets, one for November 2014 and one for November 2015, both corresponding to flight transactions (which may be new flight bookings, some kind of alterations to existing ones or cancellations) and some details regarding the

transaction. Each dataset has around five million transactions and fifteen different attributes regarding characteristics of the flight or the booking.

In between these variables we can find 3 main groups: important dates related to the bookings, characteristics of the booking itself or the client that made it and variables related to characteristics of the flight.

Both datasets were given as .csv files, each one being around 620 MB in disk. However, this size can be improved via reconverting some attributes into more appropriate data types.

Data preparation

In order to perform data preparation and feature engineering we have used both Python and R libraries such as pandas, numpy and dplyr.

Despite having a pretty clean dataset with barely any missing values or outliers, some of the variables required further exploration. These attributes are *lengthofstay*, *leadtime* and *pax*.

After this study and a brief meeting with the company we discovered:

- Negative values in *lengthofstay* are 100% correlated with the categorical variable *losname* (type of stay or transfer). If the value is positive, the categorical value associated with it is always STAY. Otherwise, it corresponds to a specific type of category that represents that the flight did not get to the destination.

- Negative values in *leadtime* happen because the booking data is following the arrival date. This is due to a delay in the processing of said transaction, hence they are considered errors and were removed.

- Values in *pax* represent differences in passengers for a reservation: if a transaction has a value of -2 in pax, said booking had removed two passengers to its original number of passengers.

In addition to perform changes to existing attributes, two more features were created:

- *Weekday*: name of the day of the week. Can be useful to analyse weekly patterns.

- *DaysSinceNov:* since the data is constrained to a really specific time frame, we can extract the day of the month and use it as a numerical variable.

A precise description of the variables can be found in the material provided (Data and Code availability).

Regarding data size, some variables (such as the categorical and datetime ones) are not assigned correctly upon loading the dataset in pandas (they are assigned the generic “pandas.object” data type) that occupy more disk space and can slow calculations.

This can be quickly solved by converting them into their appropriate data types using pandas and numpy functions, reducing the size so that, even though as .csv files they still have the same size, upon loading them in a Python or R program, their size is reduced to around 350 MB.

Having performed these changes, we have two final datasets, Nov2014.csv and Nov2015.csv, both with around five million transactions each and seventeen different attributes.

Models employed

In order to achieve our main objectives (finding patterns in the data and better understanding our dataset to study cancellations), a variety of machine learning models can be used. In this case, three approaches will be utilised: an exploratory analysis, an unsupervised model and a supervised model.

Exploratory analysis

To create this graphic we have used Power BI. In order to compare both years, we have extracted data from graphics generated thanks to an option of Power BI and used python to integrate both datasets between years in one unique database. Moreover, we have generated a new metric that represents the number of transactions divided by the number of arrivals. This solves the problem of absolute values since we can see a bar higher than others just because arrivals on that day were higher, and we want to know the number of transactions independently of the arrivals of that day.

Finding patterns among the clients: association rules

One of the most useful things that an exploratory analysis can show is which combinations of conditions lead to more cancellations or what are the characteristics of people that cancel the most (given a terrorist attack has happened). This could be extrapolable to other situations

of fear and uncertainty. Moreover, the information is going to be extremely helpful to travel retailers, hotels, and tourism-related companies.

For this analysis we have used association rules, as they have the ability to find common relationships between sets of elements of every distinct airline ticket, discovering general patterns that could lead to personalised offers that mitigate the cost of cancellations.

Given a collection of transactions, each one containing binary variables called items, an association rule is an expression $X \Rightarrow Y$, where X and Y are a subset of said items. These patterns indicate a probabilistic statement that transactions containing the item subset X tend to also have the item subset Y. This technique can be really powerful to analyse market trends, as these rules can be easily translated as "customers that buy X tend to also purchase Y". When using this $X \Rightarrow Y$ expression, the X is called the antecedent of the rule and the Y is named the consequent. (Kotsiantis, S 2006)

Since association rules are an unsupervised model (there is no labelled data to have as ground truth), we need metrics that describe how good these rules are.

A good metric to evaluate the quality of these associations is the confidence of the rule, which describes the proportion of transactions that satisfies said rule given the antecedent:

$$C(X \Rightarrow Y) = \frac{P(X \cap Y)}{P(X)} = P(Y \mid X)$$

This metric could also be described in terms of the support of a rule (proportion of transactions where antecedent and consequent appear simultaneously):

$$S(X \Rightarrow Y) = P(X \cap Y) \rightarrow C(X \Rightarrow Y) = \frac{S(X \Rightarrow Y)}{S(X)}$$

Good rules have a confidence value close to 1. The lift is also a metric to take into consideration:

$$L(X \Rightarrow Y) = \frac{C(X \Rightarrow Y)}{P(Y)}$$

A lift value closer or equal to 1 indicates independence between antecedent and consequent, meaning the rule doesn't describe an interesting pattern, since there is no correlation between X and Y. The furthest this value is to 1, the more useful the rule is.

Given our dataset, we will be focusing on the following variables:

Variable	Description	Unique values
bookingsign	Type of transaction. Can be a new booking, some kind of modification or a cancellation	4
paxprofile	Estimated client profile by ForwardKeys' classification algorithm	4
losname	Type of stay or transfer from the flight.	8
distchannel	Cabin class code	4
cabinclass	Type of travel agency	4

Table 1. Variables used for association rules model. There are 32 unique categories in total

In this case, we will be looking at rules that only lead to FULL CANCELLATIONS in the variable *bookingsign*, which accounts for < 9% of the dataset. With this scope, we are aware that a classification algorithm could also be applied, since we are not taking advantage of one of the strongest benefits from association rules, which is that the consequent can't be different or even have multiple items.

Another condition is that a minimum of 5000 instances have to meet the criteria (~1% of the data if we take into account the condition before). This criteria discards many combinations whose support is so low they could not really be extrapolable. The other four variables straightly depend on the nature of the passenger and are easy to assess for the companies interested.

Before computing any rules, the data provided must be reconverted into a set of transactions with binary variables. This can be easily achieved by transforming said factors in dummy variables. In this case, the function *as(data, "transactions")* from base R automatically prepares the data.

For the model, an apriori algorithm from arules package in R is used in both 2014 and 2015 separately, which works well with our amount of data (5,000,000 rows x 32 columns), for a later comparison. Alternatively, a more suitable algorithm for big data, such as *fpgrowth*, could have been used if more variables would have been taken into account.

Regarding hyper parameters, the minimum requirement for a rule to appear is a confidence > 0.08 and count > 5000 (support > 0.1%) as previously mentioned.

Due to the fact that 9% of all bookingsigns are FULL_CANCELLATIONS, finding rules with a lower confidence value would not be significant (for our study, as we want rules with big impact on cancellations), in short, a lower threshold of confidence would show us rules with lift < 1, which we do not want.

Understanding the client profile: linear discriminant analysis for paxprofile

As previously mentioned, the client profile is a really interesting and important variable to explore and fully comprehend. According to ForwardKeys, this variable is "the purpose of the trip according to the company's classification algorithm"[1]. This classification decision is unknown to us, and thus, considered as a black box model. An interpretable model that can explain the decisions the classification algorithm makes when categorising was trained in order to help companies reach their target and fully comprehend passenger preferences. We opted for a linear discriminant analysis model to predict the variable paxprofile.

Linear discriminant analysis is a generalisation of Fisher's linear discriminant, a statistical method that tries to find a linear combination of predictor variables to describe the differences between the different classes of a categorical target variable. This method can be later used to reduce the dimensionality of the data or as a linear classifier (Izenman, A.J. 2013).

The main goal of linear discriminant analysis in two populations is to find a linear combination of features w such that maximises the ratio of inter-class variance to intra-class variance: this means that in an ideal LDA model, populations are far away from each other, but instances within a group are quite close together (on the dimensional space):

$$max \frac{(w'x_2 - w'x_1)^2}{w'S_w w}$$

Where the numerator indicates the inter-class variance -difference of the projected mean values for each population squared- and the denominator indicates the intra-class variance -projected intra-class covariance matrix..

When dealing with a multiclass variable, the first discriminant function maximises this ratio, and subsequent discriminant functions maximise the ratio of the residual variance (left variance upon applying the previous discriminant functions). All discriminant functions are independent from each other.

There are some assumptions or conditions that have to be met to properly use this technique: namely that the predictor variables must follow a normal multivariate distribution and must be homoscedastic (intra-group variance must be the same). Also, since it is a linear model, multicollinearity can prove detrimental. Therefore, linear discriminant analysis is used with numerical predictors. However, it has been suggested that categorical variables may not significantly reduce the performance of the model or that the model can be modified accordingly to suit both data types, specially if the dimensionality of the data is pretty high (Jiang et al., 2019), (Krzanowski, 1980).

Taking these limitations into consideration, our model uses eight variables, continuous and categorical that have been converted into dummy variables:

Variable	Description	Type (unique values)
leadtime	Difference in days between the date of booking and the date of the flight	numerical
cabinclass	Cabin class code	factor (4)
distchannel	Type of travel agency	factor (4)
pax	Difference of passengers from last modification of the reservation	numerical
numpss	Number of previous steps in the itirenary	numerical
numnss	Number of next steps in the itirenary	numerical
losname	Type of stay or transfer from the flight	factor (8)

Table 2. Attributes used for linear discriminant analysis. Using dummy variables for encoding categorical attributes, the final train dataset contains 1 700 000 observations and 19 variables

One aspect to take into consideration is the class distribution of *paxprofile:*

Figure 1. Relative frequency of different profiles of clients before and after undersampling. There is clearly an imbalance between LEISURE and the other profiles

[1] Please refer to the file 202012 Data Layout Actual Air Reservations.pdf in the materials provided (Data and Code availability) for the exact definition

The target variable is clearly unbalanced, with LEISURE being the most common profile client. Therefore, common metrics to evaluate the performance of a classification model such as the accuracy are not good indicators. However, there are other metrics more suitable for this unbalanced situation, such as the Cohen's Kappa coefficient and the balanced accuracy, which will be the ones we will use to evaluate the model's efficacy.

Preliminary models without any transformations to mitigate this imbalance problem proved to give poor results, so our final presented model uses undersampled data that equalises the two main classes (LEISURE and BUSINESS). Even with this treatment, there still exists an imbalance, since the other two profiles (VFR and GROUP) have fewer instances.

We also built a toy model with only numerical variables, but it performed worse in the test set, so we will be discussing the model that mixes categorical and numerical variables, using the data treated with undersampling.
Before training the model, and given the size of the data, the dataset can be split into 75%-25% training/test subsets. Also, since the data is plentiful and flight transactions from 2015 may have an atypical distribution in the target variable due to the attacks, the model will be built using data from 2014.

Further investigation could be oriented to perform a discriminant analysis that is not linear (quadratic discriminant analysis or discriminant analysis using a kernel), allowing for some restrictions such as homoscedasticity to be removed or relaxed.

Other interpretable model: recursive partitioning trees

Decision trees are another easy-to-interpret model. These models divide the data according to splits in the input features so that they maximise the decrement in the impurity of the resulting children nodes (Strobl, C, 2009). The purity is the homogeneity of the target variable in the node.

There are multiple ways of measuring this impurity, like the entropy or the estimate of positive correctness, but in this case we will be using the Gini index.
For a given node p, and a target variable with J classes, let p_i be the
proportion of examples in the node p with label i:

$$I_G(p) = 1 - \sum_{i=1}^{J} p_i^2$$

Once the model is built, one sample can be easily classified following its path in the tree. In this case, we will be using a binary tree, which only has two outcomes regarding the split, so that it behaves like a yes/no question. We will be using the rpart R package to build this model, which uses a complexity parameter to decide to prune (if a split does not decrease the impurity by a factor equal to this parameter, the split is discarded).

In order to compare this model with the LDA one, we will be using the same dataset (with undersampling to take into account class imbalance) and the same seed for the traint-test partition. Since the data is the same, and it is still unbalanced, the metrics used to evaluate the effectiveness of the model are the same as in LDA: the Kappa Coefficient and the weighted balanced accuracy.

However, and in contrast to DLA, decision trees do not need categorical variables to be one-hot encoded. Therefore, the minable view is the same as Table 2.

Results and discussion

Exploratory analysis

Firstly, we have represented in a graphic the number of transactions of each different value regarding the bookingsign variable. We can highlight full cancellations, since there is a big difference between the number of cancellations in both periods mentioned in Figure 2.1.

After knowing this, we created a graphic (Figure 2.2) that shows the evolution of cancellations and bookings over time. From the first moment, results were as expected. This means, since November 14th 2015, flight cancellations were duplicated, and this number was kept high. This contrast in cancellations is a good indicator that our first thoughts (there were differences between the number of cancellations in November 2014 and November 2015) are headed in the right direction.

Moreover, as we can see in Figure 2.3 there is a difference between the number of cancellations made in both periods. After the attack people tend to cancel indepently of the day of the week and the magnitude of cancellations is higher than before.

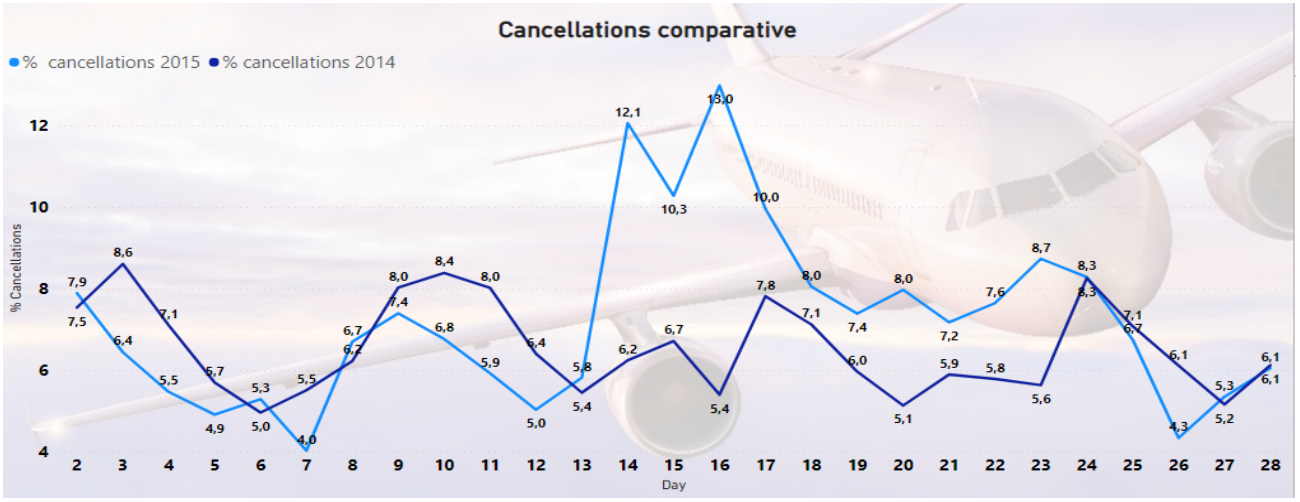


Figure 2.2. Relative % of cancellations per day for both November of 2015 and 2014. The spike and relative increase of cancellations in the following days in 2015 is due to the terrorist attack and reaches its peak three days after.

Association rules

Figure 3 and Figure 4 show the association rules obtained for 2014 and 2015, respectively.

Rules are sorted by lift as it is the most important metric for our analysis. A value > 1 means that the condition (lhs) helps the rhs happen, the higher, the better. Confidence is also important; for instance, 13% of the time a client had business paxprofile, stayed in Paris and used a cabin of class B, they decided to cancel (in 2015) which is higher than the average of instances of cancellations (< 9%)

On the other hand, it could also be interesting to study which passengers are the least likely to cancel. This could be done through the inspection of the worst conditions that decide to do a FULL CANCELLATION. The results can be found in Figure 5 and Figure 6.

In this case, a value of lift < 1 means that the condition is helping the rhs NOT happen, or at least happen less than if we did not have the condition. People that meet the lhs don't care about the attack as much as the people on the other 2 tables or they are just less likely to cancel a trip.

Some of the ideas that we can discuss that derive from these tables are:

- In 2015 people that went to Paris to STAY had the highest chance of cancelling as it appears in every rule, and because this condition is nowhere to be seen in 2014, we could argue that these people are the first to cancel for fear.

- Cabinclass B, “Business”, is another condition that people who cancel usually have. But because it appears on both years we dont know how impactful the terrorist attack was for them. Perhaps they just cancel more often.

- Paxprofile is an interesting one as both LEISURE (passengers that come for tourism) and BUSINESS are strong for cancellations on both years. But LEISURE is also strong for not cancelling in 2015.

- The distchannel = OTHER condition (travel agencies different from online, retail or corporate) appears in more rules in 2014 than in 2015, thus arguing that people that take these flights are not really fearful about the attack, as it's importance on cancellations during 2015 has at least not kept up with the other conditions.

- Confidence and count about the rules is higher in 2015, which is just another reason to think that something abnormal has happened.

- A LONG TRANSFER length of stay is a strong condition for not cancelling on both years, same goes for distchannel = OTHER and cabinclass T and E (Tourist/Economy).

- Because the conditions of 2014 and 2015 on the last tables are similar, it is hard to tell which passengers did not care about the attack, DWELLING TRANSFER on length of stay seems the most characteristic attribute of these customers. (Passengers that stay for a very short period of time)

Linear Discriminant Analysis

Figure 7 shows the confusion matrix and some performance metrics on the test set for the final model, that uses numerical and categorical variables and on which undersampling was performed to reduce class imbalance:

```
Confusion Matrix and Statistics

          Reference
Prediction BUSINESS  GROUP LEISURE    VFR
  BUSINESS   181824   1805   48068   1924
  GROUP         501   2545    2211    292
  LEISURE    135096  16331  267721  33013
  VFR            57    294       0   1861

Overall Statistics

               Accuracy : 0.6545
                 95% CI : (0.6534, 0.6557)
    No Information Rate : 0.4585
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3678

 Mcnemar's Test P-Value : < 2.2e-16
```

Figure 7. Confusion matrix in the set dataset of linear discriminant model. Since the target variable still presents imbalance, the accuracy is not relevant to evaluate the model. The Kappa Coefficient is pretty low

Even though undersampling was performed (Figure 1), the model still has a tendency to classify most samples as LEISURE. The Kappa Coefficient is not good either: sitting below the 0.4 mark, the model does not distinguish well enough between classes.

If we analyse the classifier class by class using ROC curves, we can see in Figures 12 and 13 that the model is indeed able to distinguish between some classes, but not all of them, so it is not a good approximation for the black box model we assume ForwardKeys is using.

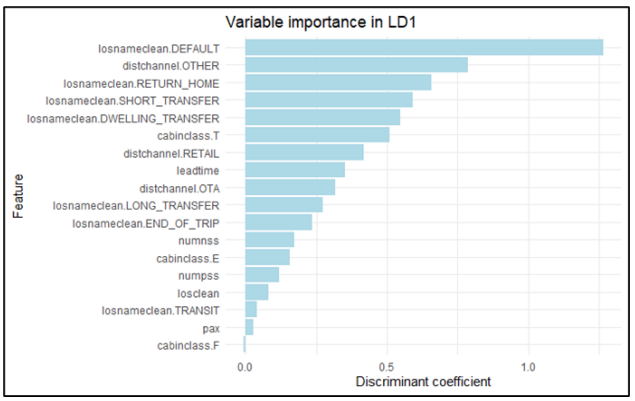


Figure 8. Feature importance in the first discriminatory dimensions (that explains 73% of the data variability). Different categories of the type of stay are very important

The overall area under the curve of the model is 0.7895, and the weighted balanced accuracy (the weights being the class frequency) is 0.683.

Regarding which variables are more important when building the discriminant dimensions, Table 3 and Figures 8 and 9 show that some variables such as the number of passengers of having a cabin class code F or E do not contribute that much to any dimension, whereas staying at the destination (losnameclean DEFAULT) has a great contribution in the two main dimensions.

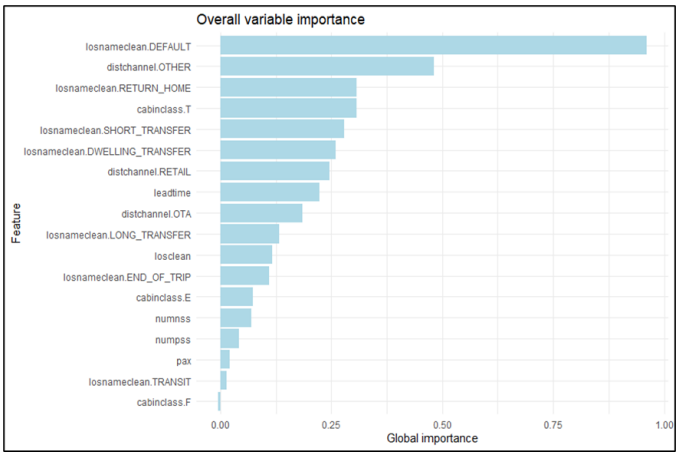


Figure 9. Overall feature importance weighted by the % of variability explained per dimension. The same tendencies as in the first discriminant dimension (Figure 8) can be seen in the global behaviour

If we take a look class-wise, Figure 10 shows similar results: in the groups VFR and GROUP, the length of stay is the most important factor, while in the most common groups (LEISURE and BUSINESS), this attribute is not so important.

A complete table of every variable importance per class can be found in Annex II (Table 4).

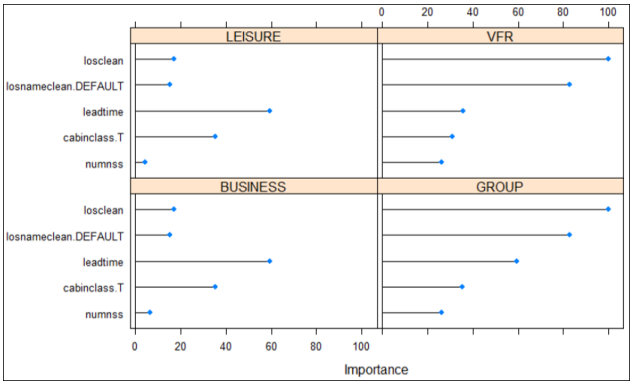


Figure 10. Top 5 most important variables for each class in paxprofile. As Table 4, losnameclean DEFAULT is a very decisive attribute to decide the profile of a passenger

Recursive partitioning tree

Figure 11 shows the confusion matrix and some performance metrics on the test set (which is the same as in the LDA model, and hence, with undersampling) for the final tree model:

```
Confusion Matrix and Statistics

          Reference
Prediction BUSINESS  GROUP LEISURE    VFR
  BUSINESS   152250    719   31563      0
  GROUP           0  14200       0      0
  LEISURE    163658   5816  286437      0
  VFR          1570    240       0  37090

Overall Statistics

               Accuracy : 0.7065
                 95% CI : (0.7054, 0.7076)
    No Information Rate : 0.4585
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4879

 Mcnemar's Test P-Value : NA
```

Figure 11. Confusion matrix in the set dataset of rtree model. Although slightly better than in the LDA model, the Kappa coefficient is still below 0.5. The influence of LEISURE is clearly shown in its prevalence.

The use of a recursive partitioning tree shows a small enhancement on the performance, improving the Kappa Coefficient by 0.12 and the weighted balanced accuracy by 0.05. Still, these values are pretty low, and the prevalence of the class LEISURE is still significantly greater than the prevalence of the other classes.

Class-wise, the classifier performs greatly in distinguishing the minor classes from the most common ones. However, it continues to struggle differentiating between the main two ones BUSINESS and LEISURE.

When comparing ROC curves of the two models, Figure 12 shows that, in fact, LDA performs better in the BUSINESS/LEISURE portion. The tree excels at every other pairing (Figure 13). The curves for all pairings can be found in Annex I, in Figure 14.

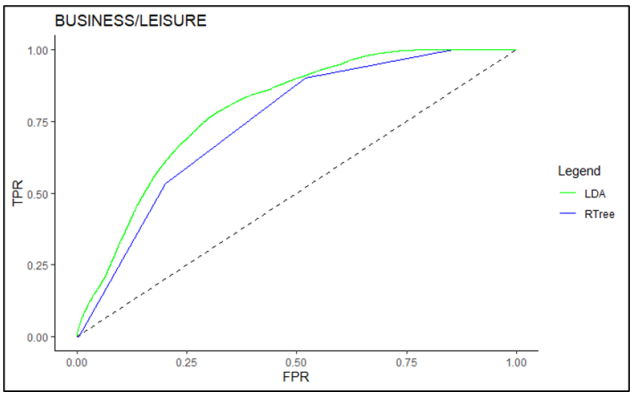


Figure 12. ROC curves comparison between BUSINESS vs LEISURE, The LDA model performs slightly better

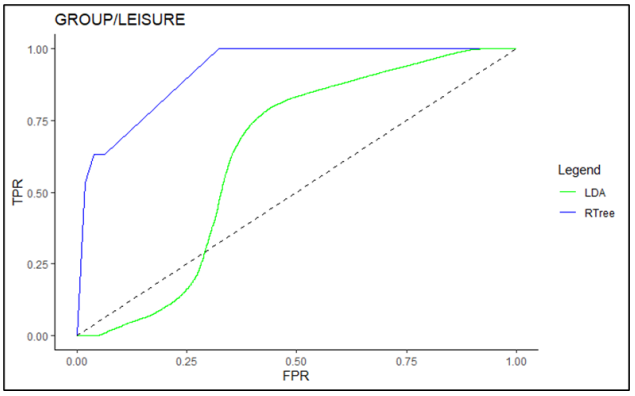


Figure 13. ROC curves comparison between GROUP vs LEISURE, The tree model is much better in distinguishing these classes

Regarding variable importance, the model does not use the number of previous stops and the day of the month. The most important features are *distchannel, cabinclass* and the type of stay (*losname)*, as shown in Table 5.

feature <chr>	importance <dbl>
disthannel	191859.12
losclean	168377.70
cabinclass	162917.55
losnameclean	153160.89
pax	146424.71
leadtime	129294.42
bookingday	5757.67
numnss	3612.47
numpss	0.00
days_sinceNov2014	0.00

Table 5. Feature importance in the rtree model. *distchannel* is clearly the most influential attribute

Finally, the tree is shown in Figure 15, in Annex I. Matching the results in Table 5, some splits clearly

differentiate classes: for instance, distchannel = CORPORATE helps distinguish clients with profile BUSINESS.

Other models and limitations

One of the models that we tried to implement was clustering, trying to find other ways to group the data than the already provided by the data itself (mainly by client profile and type of transaction). However, due to the immense number of observations (5 million), the model could not be created given the limited computational resources. The guidance we were supposed to follow started by creating a distance matrix, from which we would have obtained the number of necessary clusters. It should be noted that we managed to get a distance matrix, but it was considered useless because it only included 1% of the transactions, that's why we can not consider it as a representative sample. Given the time constraints we have to complete the project and having other models to focus on, we have decided not to carry out this analysis. Even being aware that it would be possible to find solutions.

Moreover, we tried to create a model to predict if the transactions are new bookings, cancellations or modifications. This is interesting from the economic point of view since knowing if a flight is going to be cancelled or not given some characteristics allows the company to take advantage and make decisions based on the prediction. We used a support vector machine with data related to 2014, however, with this amount of data and this model we spent three days executing the code in order to obtain a result. So, we will have to find solutions, possibly reducing the number of files of our dataset or trying another model. A distributed approach to computing the model or the access to a server could be other solutions.

Also, we had the same limitation with the second SVM model. The objective of this one was to train a model with data of 10 days after the attack since in this period people have a different behaviour. In this way, we will provide to the company a model that can be used if something similar occurs.

We would also like to comment on one model that couldn't be done due to different encounterments: PLS-DA. We tried to predict *paxprofile* through other variables via an interpretable model. The main problem was the huge amount of data we are working with. The model could only be done with a non-representative sample of the total. Trying with a bigger sample, would probably not give us a result, and if it did, the best case scenario would still give us a very long waiting time.
There is also a second PLS-DA model with *bookingsign* as target variable to understand which variables help us to know why flights are cancelled or a booking is made.

Given the results of the model we can conclude that with the data we have it is not possible to predict this variable accurately enough. Most predictions are classified as new bookings and the model has no predictive power. It would be necessary to gather other characteristics in order to predict better.

To finish with, we also thought about a time series analysis of the cancellations to extract some information about the event or to model the peak and the following drop to normal levels (%) of cancellations. However this idea was quickly removed and derived into the exploratory analysis with association rules, which ended up being more useful. The problems were mainly the lack of data after the attack (only 13 days), and the difficulty to extrapolate something that the company could find useful.

Conclusions

Association rules

After comparing the ruleset for each year, the team has decided that clients with a business profile and staying in the city are the strongest conditions for a cancellation in November of 2015 (which are not for November of 2014). Now we will visualise it on a plot (Figure 16) to see the evolution of relative cancellations throughout the month (% of transactions that are total cancellations that day). The cabin class code B is not taken into account because it was also a common condition in 2014.

The other lines are for reference and comparison of the same study without the conditions on both years, they are the same as Figure 3.

When it comes to the day of the attack, we can see a huge peak. About 35% of that day's transactions were cancellations, when usually it stays around 10% (red and blue lines for 2014 and 2015 with the condition overlap at the start of the month)

Also because we remove the cabinclass condition, there are more samples and the conclusions are more robust. People willing to STAY with a BUSINESS paxprofile are the most likely to cancel due to a surprising/fearful event.

The same can be done (Figure 17) over the passengers least likely to cancel. It is harder in this case due to the

similarities between 2014 and 2015, so our only condition will be *losname = DWELLING_TRANSFER,* meaning that people do not stay for very long. Short transfer is also an option.

Now the blue dotted line is below the other line for all of the month, and the peak is practically invisible in comparison to 2015, is does not even look as if a terrorist attack has happened, we could argue that these people were unfaced by the attack.

Notice how the dotted red line and red line are almost the same, this is because this condition is not important to determine regular cancellations (like cabinclass = B), but it is for 2015. These people behave the same despite a sudden terrible event happening.

The biggest flaw of this exploratory model are the conditions, which lower the amount of cases that meet the requirements, this can be seen on the association rules.

Around 800k people had a losname = dwelling transfer (only stay for a short period of time), so it's more useful for companies to study them than it is to study a very specific profile that only appears 30k times. This may appear as a lot of people, however it's only 1000 passengers each day, in comparison to 25000 passengers per day on the other group.

Something else to take into account is that we are assuming that all the changes in cancellations are due to the terrorist attack, but there are many different reasons for which someone could decide not to go to a city or to take a flight. However due to a terrorist attack being incredibly impactful on a city, this is not a big problem.

Linear discriminant analysis and partitioning trees

Despite not fully following all the assumptions and conditions to perform an ideal linear discriminant analysis, and possibly due to the amount of the data provided by Forwardkeys, our LDA model with 3 discriminant dimensions gives a weighted balanced accuracy of around 68% of the clients' profile from the test set, whereas a recursive partitioning tree improves this metric by 5%. Despite the improvements, both models fail to reach a good Kappa coefficient (both below 0.5), not providing a good model to understand ForwardKeys' black box model.

In both models the most important attributes to determine the client's profile are type of stay, the cabin class code and the type of travel agency.

Since the main objective of these models was trying to provide interpretable and contrastive explanations to a profile classification done by an unknown model, rather than precisely predicting said profile (therefore giving really good results in metrics such as the Kappa Coefficient and the weighted balanced accuracy score), we can conclude that we may be able to have a loose intuition of what attributes are more determinant to assign a client a specific profile. One thing we can be sure of is that more unprovided variables could help construct a more robust and complete model.

Deployment

A study based on association rules could be of extremely importance and useful to FORWARDKEYS. Our team has the idea of creating an application that the company could use to evaluate how passengers react to catastrophic events in a country. As for now, we have a working prototype. This is how it works:

1. Given 2 datasets like the ones we were provided (event year and year before/ same month without a catastrophic event), calculate the association rules that lead to a certain bookingsign (with corresponding hyperparameters).
2. Use a metric to compare both rulesets to find which rules are the most important for the event year. We have done this manually, "humanly", by looking at the results, but on the application an algorithm compares the rulesets of both years and gives them a "score" to find the best and worst ones. Find more about it in Annex III. Deployment.
3. Once we have the rules, visualise them on a plot and get relevant information like the one in Figure 17.
4. For the company it would be extraordinary to know how the passengers react to a catastrophic event as soon as possible. However, as we have approached it, this is an a posteriori study (we need the data to know how the clients have reacted to an event).
   But if we had all the possible data available up to this point in time, we could replicate the same study and discover the association rules for other kinds of fatal events (earthquakes, tornados, war attacks, even a change in the political power), creating a table similar to this one:

Passenger attributes	Event	Date	City	Passengers cancelling
losname = STAY, cabinclass = B	Terrorist attack	13/11/2015	Paris	200 000
paxprofile = LEISURE	Hurricane	23/08/2005	New Orleans	450 000
losname = SHORT TRANSFER	War attack	22/03/2022	Mariupol	125 000

Table 5. Possible scenario of all fatal events which we have available data of. This is an extreme oversimplification of what information it could contain

With the assumptions that previous reactions to an event lead to the same reactions as in the future, the company is ready to know what kind of clients will cancel, how big the impact would be, or for how long the drought in flights will be. Obviously a further analysis of the data created in Table 5 to study each kind of event and find common patterns is needed.

The most beneficial thing of this approach is that we could do exactly the same study but analysing those passengers that do not care about the event, to know which clients would fly to the destination notwithstanding what has happened. This is another information that could be of extreme importance to travel retailers, hotels, and hospitality of the country/city.

In this project, we will only work as far as step 3, but our will is to leave everything ready for a possible implementation of step 4, where the company could benefit the most.

The deployment has been built up to the 3rd step using RStudio, with shiny and flexdashboard, it is an interactive working dashboard where ForwardKeys can upload similar data to the one we were provided (data of event and data of other year, same month), so it is ready to be used.

On the Dashboard page we find what we call the "best attributes plot" which is the score that the algorithm gives to each rule. There is also the relative frequency of the bookingsign plot, which in reality is an extension of the exploratory analysis. It serves as a way to better understand the attack. Finally, there are a variety of attack metrics that will also make the impact of the attack very easy to understand. It's possible to choose between bookingsigns, the attributes of the passengers, the attack day and the temporal impact in days.

The team has deployed 2 models in the cloud (shinyapps). One for ForwardKeys, where they can upload their own data, and one with the data from Paris already loaded, so anyone can see and try it for themselves.
Data-ready application (sample with 500,000 transactions on each dataset, otherwise it won't load on the free version)
Original Deployment (In the same way, it needs small (~500k transactions) on each file at max)

Acknowledgments
We want to thank Sonia Tarazona Campos, for guiding us as the project moved forward and especially when we got stuck; José Alberto Conejero Casares y José Hernández Orallo, for teaching us the phases of a project and how to get things done properly; and, finally, Laura Rodríguez for answering our questions.

Code

The main software employed was Power BI, Python and R. In these last two, a variety of libraries were used, such as pandas, matplotlib, numpy, sklearn, caret, dplyr, dbplyr, pROC, tidyr, tidyverse, rpart, ggplot2, shiny, flexdashboard, arules and biotools.

Every dataset and code used in this project are provided in the following GitHub repository: https://github.com/Daniframe/PROY-III-FORWARDKEYS

References

(Izenman, A.J. 2013). *Linear Discriminant Analysis. In: Modern Multivariate Statistical Techniques. Springer Texts in Statistics. Springer, New York, NY.* https://doi.org/10.1007/978-0-387-78189-1_8

(Janić, 2015) Janić, M. (November 2015). Reprint of *Modelling the resilience, friability and cost of an air transport network affected by a large-scale disruptive event*. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0965856415002049?casa_token=RCvO-zusLsQAAAAA:d5hNvEfZdU_ul5Yt1UjMQHoVBzZuRNg8LLa6udup9Q65mXcMFJZSh4iA6_UWDSj5fK-FFLtm

(Jiang et.al 2019) Jiang et al, *Linear Discriminant Analysis with High-dimensional Mixed Variables*, https://arxiv.org/abs/2112.07145

(Kotsiantis, S., Kanellopoulos, D 2006) *Association Rules Mining: A Recent Overview,* https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.6295&rep=rep1&type=pdf

(Krzanowski, 1980), *Krzanowski W.J., Mixtures of continuous and categorical variables in Discriminant Analysis*, https://www.jstor.org/stable/2530217?origin=crossref

(Strobl, C., Malley, J., & Tutz, G., 2009). *An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests.* Psychological Methods, 14(4), 323–348, https://doi.org/10.1037/a0016973

Annex I. Figures used

A. Models employed

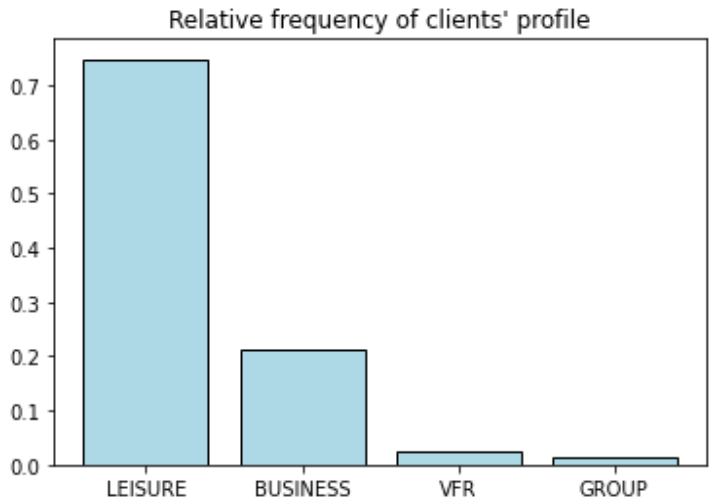


Figure 1. Relative frequency of different profiles of clients. There is clearly an imbalance between LEISURE and the other profiles

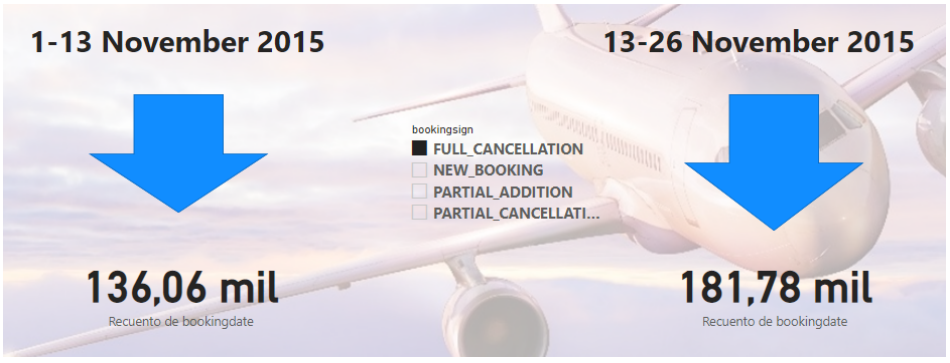


Figure 2.1. Shows how the attack augments the number of full cancellations.

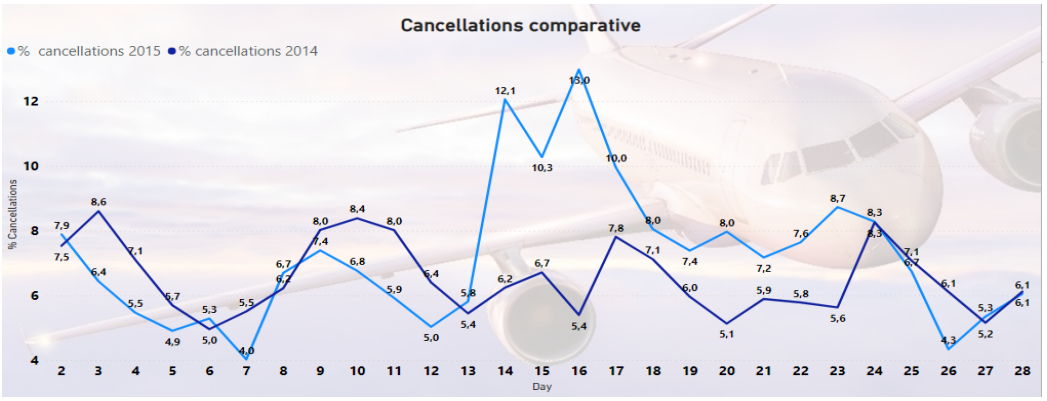


Figure 2.2. Relative % of cancellations per day for both November of 2015 and 2014. The spike and relative increase of cancellations in the following days in 2015 is due to the terrorist attack and reaches its peak three days after.

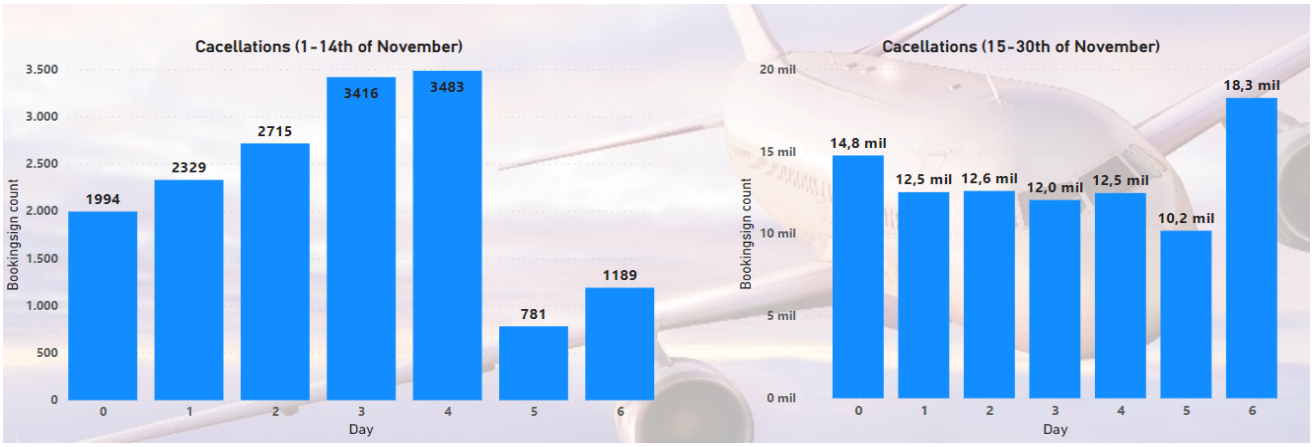


Figure 2.3. Shows the difference of cancellations made each day of the week between 1-14th of November and 15-30th of this month in 2015.

B. Association rules

	lhs <chr>		rhs <chr>	confidence <dbl>	lift <dbl>	count <int>
[1]	{paxprofile=BUSINESS,losname=SHORT_TRANSFER,cabinclass=B,distchannel=OTHER}	=>	{bookingsign=FULL_CANCELLATION}	0.1177077	1.737647	7084
[2]	{losname=SHORT_TRANSFER,cabinclass=B,distchannel=OTHER}	=>	{bookingsign=FULL_CANCELLATION}	0.1161998	1.715388	16918
[3]	{losname=DWELLING_TRANSFER,cabinclass=B,distchannel=OTHER}	=>	{bookingsign=FULL_CANCELLATION}	0.1156333	1.707025	9745
[4]	{paxprofile=LEISURE,losname=DWELLING_TRANSFER,cabinclass=B,distchannel=OTHER}	=>	{bookingsign=FULL_CANCELLATION}	0.1148947	1.696121	6522
[5]	{paxprofile=LEISURE,losname=SHORT_TRANSFER,cabinclass=B,distchannel=OTHER}	=>	{bookingsign=FULL_CANCELLATION}	0.1147562	1.694077	9749
[6]	{paxprofile=LEISURE,losname=DWELLING_TRANSFER,cabinclass=B}	=>	{bookingsign=FULL_CANCELLATION}	0.1137617	1.679395	7077

Figure 3. Top 6 rules which consequent is a full cancellation sorted by lift (2014). Notice how common cabinclass B and distchannel OTHER are. More info about the creation and other use cases of this table at the github repository models/Association_rules.Rmd

	lhs <chr>		rhs <chr>	confidence <dbl>	lift <dbl>	count <int>
[1]	{paxprofile=BUSINESS,losname=STAY,cabinclass=B}	=>	{bookingsign=FULL_CANCELLATION}	0.1336452	1.810246	7799
[2]	{losname=STAY,cabinclass=B}	=>	{bookingsign=FULL_CANCELLATION}	0.1259869	1.706513	19117
[3]	{losname=STAY,cabinclass=B,distchannel=OTHER}	=>	{bookingsign=FULL_CANCELLATION}	0.1247002	1.689085	15028
[4]	{losname=STAY,distchannel=RETAIL}	=>	{bookingsign=FULL_CANCELLATION}	0.1226451	1.661247	10084
[5]	{paxprofile=BUSINESS,losname=STAY}	=>	{bookingsign=FULL_CANCELLATION}	0.1216298	1.647495	23783
[6]	{paxprofile=LEISURE,losname=STAY,cabinclass=B}	=>	{bookingsign=FULL_CANCELLATION}	0.1211052	1.640389	10327

Figure 4. Top 6 rules which consequent is a full cancellation sorted by lift (2015). losname STAY is one of the strongest conditions, in addition, it is nowhere to be seen in 2014. This means these passengers are really sensitive to sudden events of fearful nature.

LHS <fctr>	RHS <fctr>	confidence <dbl>	lift <dbl>	count <int>
{losname=LONG_TRANSFER,cabinclass=T,distchannel=OTHER}	{bookingsign=FULL_CANCELLATION}	0.04533051	0.6691868	5050
{losname=LONG_TRANSFER,cabinclass=T}	{bookingsign=FULL_CANCELLATION}	0.04628090	0.6832169	5901
{paxprofile=BUSINESS,losname=RETURN_HOME,cabinclass=T,distchannel=OTHER}	{bookingsign=FULL_CANCELLATION}	0.04666632	0.6889066	5402
{paxprofile=BUSINESS,losname=RETURN_HOME,cabinclass=T}	{bookingsign=FULL_CANCELLATION}	0.04676642	0.6903843	5976
{losname=RETURN_HOME,cabinclass=E}	{bookingsign=FULL_CANCELLATION}	0.04942071	0.7295679	6360
{losname=RETURN_HOME,cabinclass=E,distchannel=OTHER}	{bookingsign=FULL_CANCELLATION}	0.04964505	0.7328798	6189

Figure 5. Top 6 rules which consequent is not a full cancellation (2014). Cabinclass T and losname = return home are the most common conditions. Notice how similar it is to 2015, the first rows are the same.

LHS	RHS	confidence	lift	count
{losname=LONG_TRANSFER,cabinclass=T,distchannel=OTHER}	{bookingsign=FULL_CANCELLATION}	0.04623495	0.6262599	5192
{losname=LONG_TRANSFER,cabinclass=T}	{bookingsign=FULL_CANCELLATION}	0.04757939	0.6444705	6113
{paxprofile=LEISURE,losname=DWELLING_TRANSFER,cabinclass=T,distchannel=OTHER}	{bookingsign=FULL_CANCELLATION}	0.05089682	0.6894056	27877
{paxprofile=LEISURE,losname=DWELLING_TRANSFER,cabinclass=T}	{bookingsign=FULL_CANCELLATION}	0.05121315	0.6936904	30724
{losname=DWELLING_TRANSFER,cabinclass=T,distchannel=OTHER}	{bookingsign=FULL_CANCELLATION}	0.05128297	0.6946360	33065
{paxprofile=LEISURE,losname=SHORT_TRANSFER,cabinclass=T}	{bookingsign=FULL_CANCELLATION}	0.05148291	0.6973444	42074

Figure 6. Top 6 rules which consequent is not a full cancellation (2015). Cabinclass T and a leisure paxprofile are the most common conditions. Notice how similar it is to 2014, and the first rows are the same.

C. Linear Discriminant Analysis and Recursive Partitioning Trees

```
Confusion Matrix and Statistics

          Reference
Prediction BUSINESS  GROUP LEISURE    VFR
  BUSINESS   181824   1805   48068   1924
  GROUP         501   2545    2211    292
  LEISURE    135096  16331  267721  33013
  VFR            57    294       0   1861

Overall Statistics

               Accuracy : 0.6545
                 95% CI : (0.6534, 0.6557)
    No Information Rate : 0.4585
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3678

 Mcnemar's Test P-Value : < 2.2e-16
```

Figure 7. Confusion matrix in the set dataset of linear discriminant model. Since the target variable still presents imbalance, the accuracy is not relevant to evaluate the model. The Kappa Coefficient is pretty low, and the majority of the samples are classified in the LEISURE class

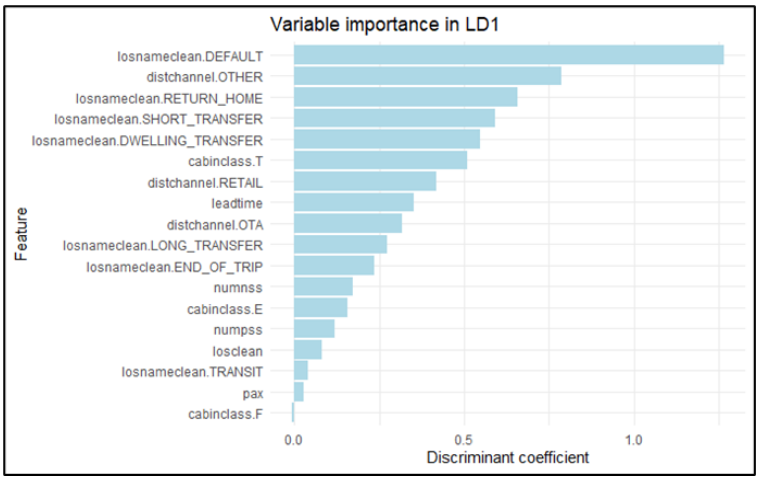


Figure 8. Feature importance in the first discriminatory dimensions (that explains 73% of the data variability). Different categories of the type of stay are very important

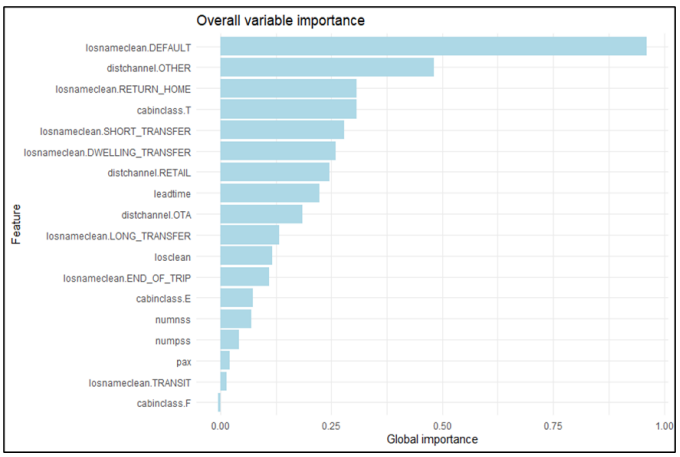


Figure 9. Overall feature importance weighted by the % of variability explained per dimension. The same tendencies as in the first discriminant dimension (Figure 8) can be seen in the global behaviour

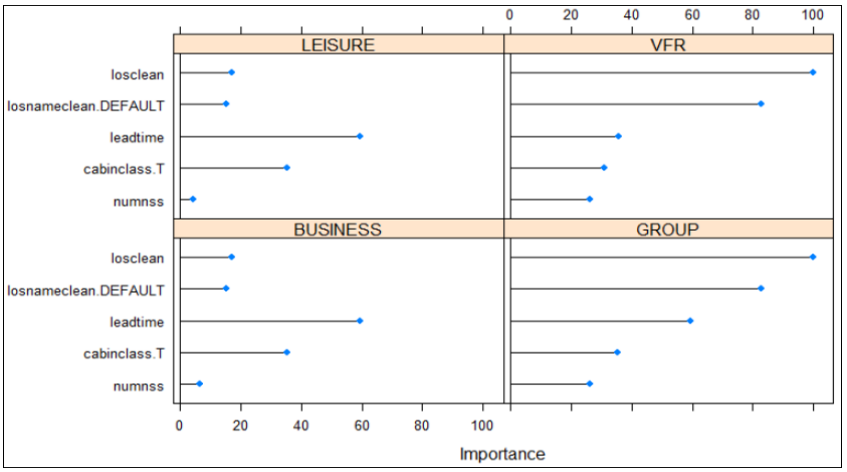


Figure 10. Top 5 most important variables for each class in paxprofile. As Table 4, losnameclean DEFAULT is a very decisive attribute to decide the profile of a passenger

```
Confusion Matrix and Statistics

          Reference
Prediction BUSINESS  GROUP LEISURE    VFR
  BUSINESS   152250    719   31563      0
  GROUP           0  14200       0      0
  LEISURE    163658   5816  286437      0
  VFR          1570    240       0  37090

Overall Statistics

               Accuracy : 0.7065
                 95% CI : (0.7054, 0.7076)
    No Information Rate : 0.4585
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4879

 Mcnemar's Test P-Value : NA
```

Figure 11. Confusion matrix in the set dataset of rtree model. Although slightly better than in the LDA model, the Kappa coefficient is still below 0.5. The influence of LEISURE is clearly shown in its prevalence.

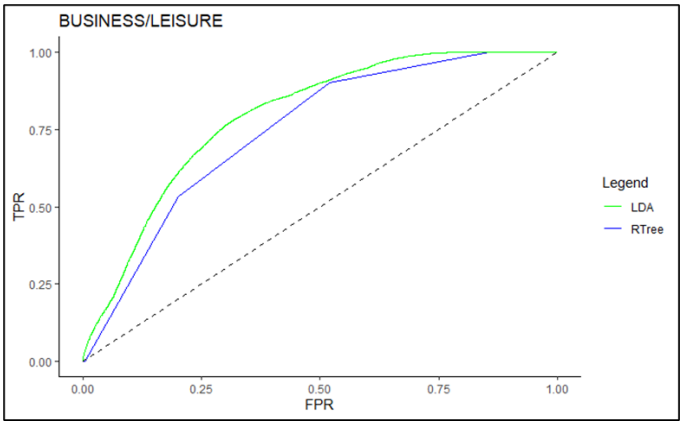


Figure 12. ROC curves comparison between BUSINESS vs LEISURE, The LDA model performs slightly better

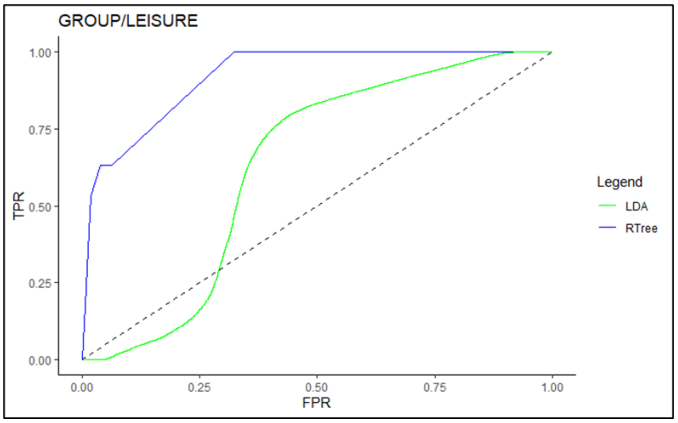


Figure 13. ROC curves comparison between GROUP vs LEISURE, The tree model is much better in distinguishing these classes

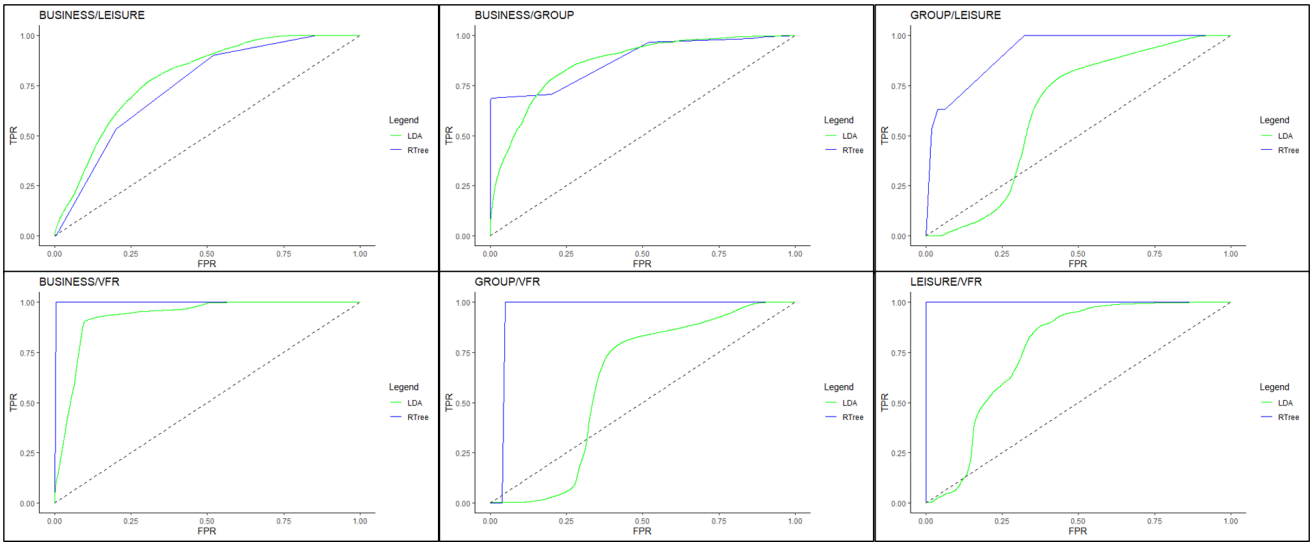


Figure 14. OvO ROC curves comparison between linear discriminant model and recursive partitioning tree model. The tree model excels at distinguishing all classes except for the pairing BUSINESS vs LEISURE, in which LDA has a slight advantage.

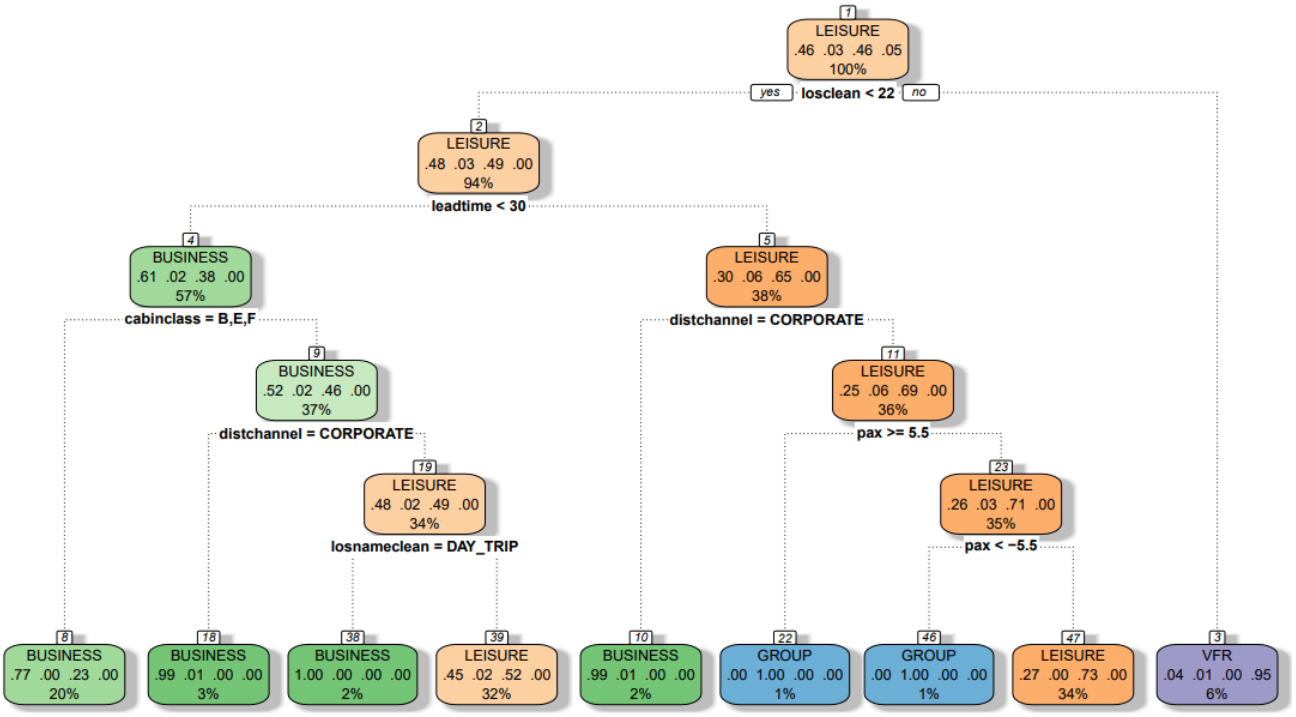


Figure 15. Final rpart decision tree model. It consists of 6 levels, and some variables clearly help distinguish between profiles: if a client has a corporate travel agency, it is very likely they have a business profile.

D. Conclusions

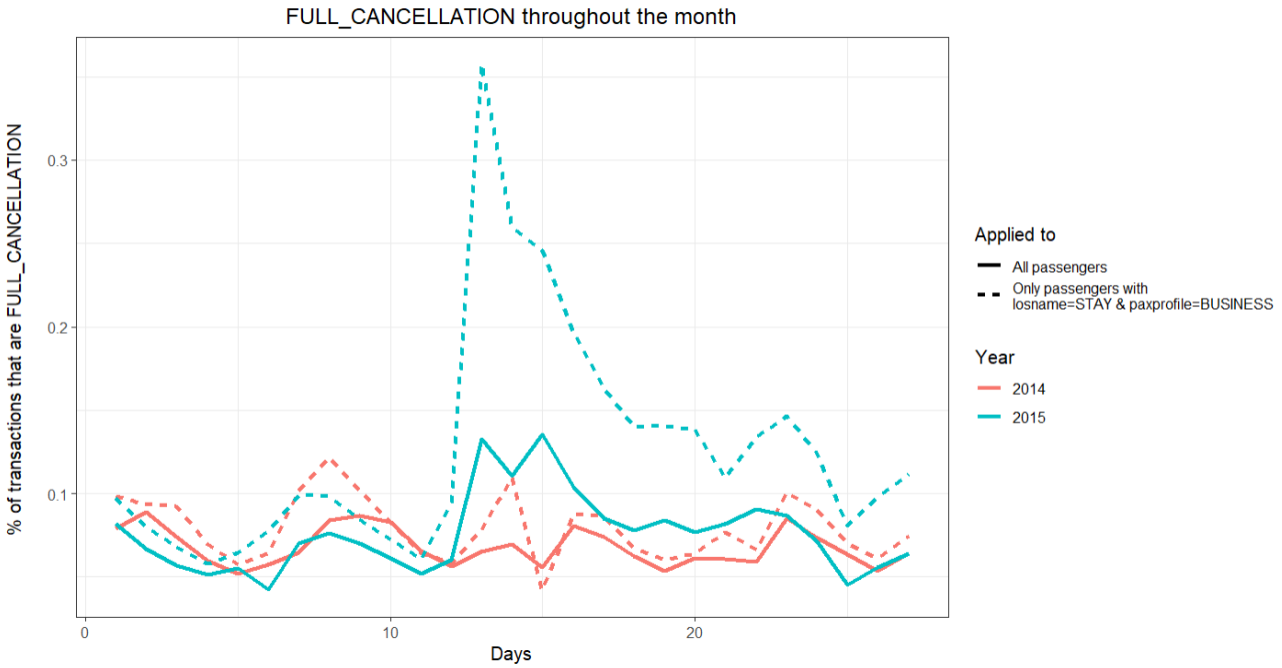


Figure 16. Relative % of cancellations per day for both 2014 and 2015 with and without the losname = STAY and paxprofile = BUSINESS.. Notice the strength at the peak and the following days, whereas 2014 is not altered as much. Built using RStudio and ggplot

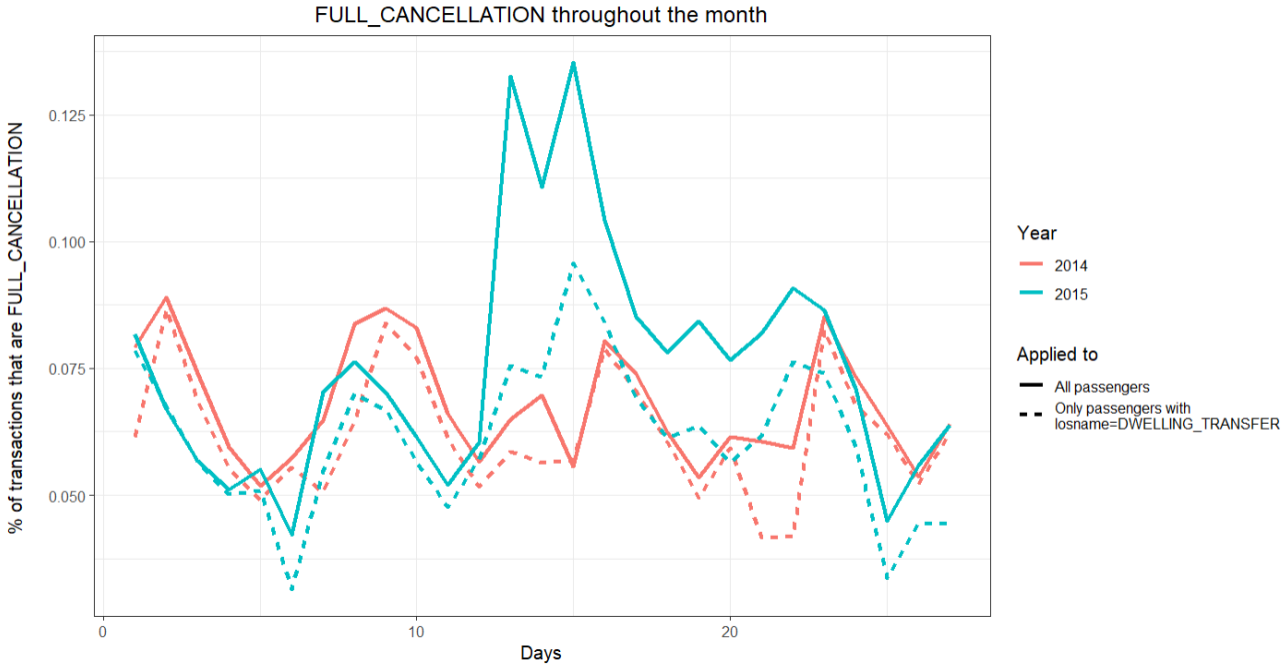


Figure 17. Relative % of cancellations per day for both 2014 and 2015 with and without the losname = dwelling transfer condition. Notice how the peak at the attack day disappears, and 2014 is not really altered. Built using RStudio and ggplot

Annex II. Tables employed

Variable	Description	Unique values
bookingsign	Type of transaction. Can be a new booking, some kind of modification or a cancellation	4
paxprofile	Estimated client profile by ForwardKeys' classification algorithm	4
losname	Type of stay or transfer from the flight.	8
distchannel	Cabin class code	4
cabinclass	Type of travel agency	4

Table 1. Variables used for association rules model. There are 32 unique categories in total

Variable	Description	Type (unique values)
leadtime	Difference in days between the date of booking and the date of the flight	numerical
cabinclass	Cabin class code	factor (4)
distchannel	Type of travel agency	factor (4)
pax	Difference of passengers from last modification of the reservation	numerical
numpss	Number of previous steps in the itirenary	numerical
numnss	Number of next steps in the itirenary	numerical
losname	Type of stay or transfer from the flight	factor (8)

Table 2. Attributes used for linear discriminant analysis. Using dummy variables for encoding categorical attributes, the final train dataset contains 1 700 000 observations and 19 variables

	LD1 (0.73) <dbl>	LD2 (0.24) <dbl>	LD3 (0.03) <dbl>
losnameclean.DEFAULT	1.266055679	0.227452031	-0.585901793
distchannel.OTHER	0.786033234	-0.326067847	-0.485155854
losnameclean.RETURN_HOME	0.656948503	-0.659236472	-0.460336330
losnameclean.SHORT_TRANSFER	0.591523664	-0.556323651	-0.630848660
losnameclean.DWELLING_TRANSFER	0.546293344	-0.511086239	-0.514290540
cabinclass.T	0.508983757	-0.265299336	-0.037834519
distchannel.RETAIL	0.417773157	-0.226367087	-0.161107112
leadtime	0.352185552	-0.230460325	0.710278370
distchannel.OTA	0.319154450	-0.157344523	-0.314537575
losnameclean.LONG_TRANSFER	0.272061842	-0.245150316	-0.226459937
losnameclean.END_OF_TRIP	0.235040236	-0.242759261	-0.121976868
numnss	0.172549567	-0.207852414	-0.185298075
cabinclass.E	0.156327323	-0.149512109	-0.147652467
numpss	0.119130432	-0.163003206	-0.209574721
losclean	0.082253010	0.236422640	-0.016956040
losnameclean.TRANSIT	0.040237613	-0.050049047	-0.091272770
pax	0.028720184	-0.058727558	0.501780574
cabinclass.F	-0.006217516	-0.002888266	-0.002455084

Table 3. Attribute importance per discriminant dimension, ranked by importance in the first dimension.

	BUSINESS	GROUP	LEISURE	VFR
leadtime	59.44585057	59.4458506	59.44585057	35.7187319
cabinclass.E	13.43853734	13.4385373	13.43853734	12.4837950
cabinclass.F	0.27849791	0.2784979	0.27849791	0.1773205
cabinclass.T	35.40179588	35.4017959	35.40179588	30.8013653
distchannel.OTA	0.93670904	0.5495719	0.53109989	0.9367090
distchannel.OTHER	13.22518856	13.7452355	10.87105150	13.7452355
distchannel.RETAIL	1.18351477	1.1835148	1.37011217	0.0000000
pax	16.27237206	16.2723721	16.27237206	3.8906009
numpss	12.17086659	15.7158969	12.17086659	15.7158969
numnss	6.28244383	26.1197036	4.19193297	26.1197036
losclean	17.12485128	100.0000000	17.12485128	100.0000000
losnameclean.DEFAULT	15.11519825	82.7478302	15.11519825	82.7478302
losnameclean.DWELLING_TRANSFER	2.12717481	14.7642084	2.70760848	14.7642084
losnameclean.END_OF_TRIP	2.51300201	5.4371658	1.39207241	5.4371658
losnameclean.LONG_TRANSFER	0.19068781	2.9306025	0.32424248	2.9306025
losnameclean.RETURN_HOME	3.28566248	23.2206677	3.28566248	23.2206677
losnameclean.SHORT_TRANSFER	6.09101493	24.4878599	6.75555491	24.4878599
losnameclean.TRANSIT	0.05746823	0.1363009	0.05746823	0.1363009

Table 4. Variable importance per client profile in linear discriminant analysis. The least common profiles have strong dependencies with losclean and losnameclean.DEFAULT, while the most common profiles are more related to leadtime.

feature <chr>	importance <dbl>
disthannel	191859.12
losclean	168377.70
cabinclass	162917.55
losnameclean	153160.89
pax	146424.71
leadtime	129294.42
bookingday	5757.67
numnss	3612.47
numpss	0.00
days_sinceNov2014	0.00

Table 5. Feature importance in the rtree model. *distchannel* is clearly the most influential attribute

Passenger attributes	Event	Date	City	Passengers cancelling
losname = STAY, cabinclass = B	Terrorist attack	13/11/2015	Paris	200 000
paxprofile = LEISURE	Hurricane	23/08/2005	New Orleans	450 000
losname = SHORT TRANSFER	War attack	22/03/2022	Mariupol	125 000

Table 6. Possible scenario of all fatal events which we have available data of. This is an extreme oversimplification of what information it could contain

Annex III. Deployment

Algorithm to compare rulesets, finds the most important rules for each year. In this case, we only compute a score for single condition rules ex. losname=STAY. However it takes into account all rules were that condition appears

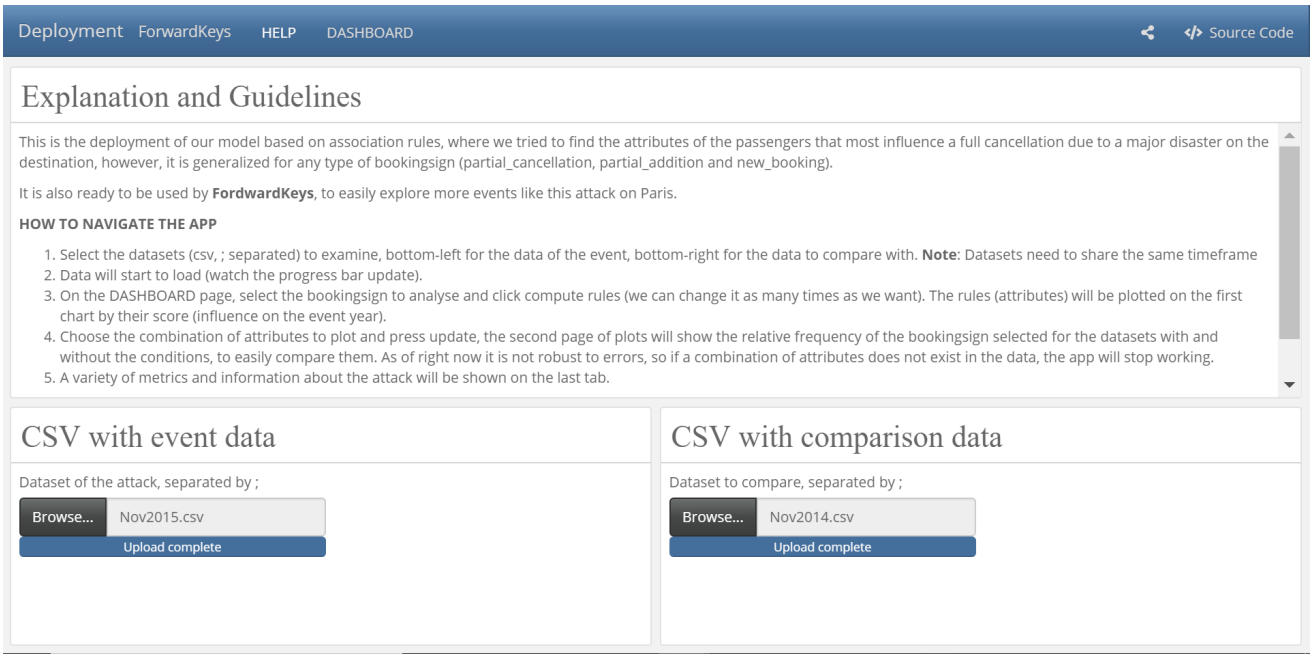
```
for (i in colnames(rules_attack_df)){
      sub = DATAFRAME(subset(rules_attack, subset = lhs %in% i))
      sub2 = DATAFRAME(subset(rules_compare, subset = lhs %in% i))
            dat = data.frame(condition = i, value = sum(log(sub$support) *
sub$confidence * sub$lift^2) / sum(log(sub2$support) * sub2$lift^2 *
sub2$confidence), quantity_attack = max(sub$count), quantity_compare =
max(sub2$count))
```

Basically for each condition (losname=STAY, paxprofile=BUSINESS, losname=RETURN_HOME…) , it subsets the rules for each ruleset to only those rules where the condition appears, it then calculates the sum of the product of log(support), lift^2 and condition for each rule on the ruleset

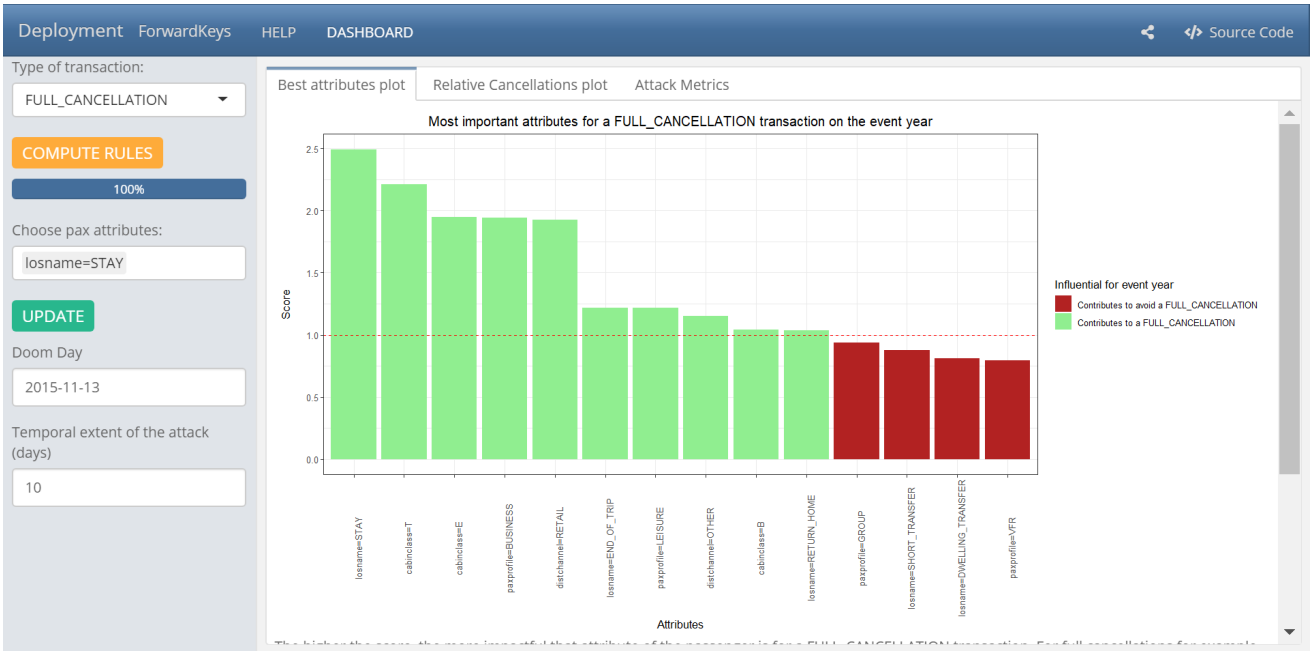
- Rules with higher lift, support and confidence will give a higher score
- Support is important but less than confidence or lift. That is why we apply a log transformation to give it less influence on the algorithm
- Lift is the most important metric, so it is squared to give it a higher influence.
- The dataset that contains more rules will generally have a higher score, as there will be more values on the sum()

Then the score for both rulesets is divided to normalise the data and easily compare values. Rules with a score way over 1 will be more significant for event data, meaning that passengers with those attributes will be prone to cancel when an event of this calibre happens. The same can be said for rules way under 1, passengers with those attributes will be significant.

The algorithm works fairly well on the data from Paris. Losname STAY and paxprofile BUSINESS are at the top, and losname DWELLING TRANSFER is at the bottom. Which is what the team had decided after studying the rulesets manually.



Deployment Page 1. Explanation of the App. The first step is to upload the corresponding datasets.



Deployment Page 2. The Dashboad. On the left, the inputs and options to show, on the right the plots and metrics, divided in tabs. It's a simple but working and generalized model to easily explore and mine information about an event.