NOVA IMS
Information Management School

# BUSINESS CASE 2

MASTER'S DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

## PREDICTING CANCELATIONS

**Group PN:**

Mafalda Garcia, number: 20210763

Simão Pereira, number: 20210250

Tiago Santos, number: 20210548

Rui Ribeiro, number: 20211017

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

# INDEX

# 1. Introduction

From 2014 to 2018, the booking cancelation rate rose from 33% to 40%, cancelations are highly harmful to the hotel business as they put it at risk in terms of image and money. Cancelations happen from several reasons, ones more excusable than others, and that's why our goal with this project is to provide the Manager Director of cancelations of hotel chain C with the necessary tools to reduce these cancelations and improve the hotel business.

To carry out this project we followed a CRISP-DM methodology. Firstly, we start by reconnaissance of the business, analyzing its objectives, which materials we had access to and defining the data mining goals for this project. Next, we will perform a recognition of the data itself, seeing its composition in detail and we will start the necessary steps for the data preparation, making all the necessary changes so that it is possible to apply the respective prediction models in it.

# 2. Business Understanding

## 2.1 Business Objectives

Over the last few years Hotel Chain C has suffered a negative impact caused by high cancelation rates of around 28% at the Algarve resort (H1) and 42% at the city hotel in Lisbon (H2). The Manager Director of cancelation policies tried some techniques to deal with the situation such as a more aggressive overbooking policy and then a less aggressive one, however both approaches were unsuccessful. As a way of dealing with this situation, the objective is to create predictive models that allow the Manager Director of the Hotel chain C to create better pricing, more efficient overbooking policies and still be able to simply and automatically identify bookings with high likelihood of canceling through the behavioral patterns of its customers over the past years.

## 2.2 Assess Situation

The hotel provided us a dataset with 31 variables describing the 79,330 observations of H2 bookings due to arrive between the 1st of July of 2015 and the 31st of August 2017, including bookings arrived and bookings that were canceled.

This dataset includes information related not only to the dates information but also to previous cancelations made by that customer, the type of deposit mad, the number of week or weekend nights that the guest intends to stay at the hotel, among other information that will allow us to do an analysis of guest's behavioral patterns.

Finally, it is important to emphasize that the dataset provided, being this representative of an hotel, does not present any personal data that could jeopardize the privacy of its guests and therefore all the work that we will develop will be around behavior and non-personal characteristics.

## 2.3 Data Mining Goals

The business objectives already mentioned can also be described in a technical way , more precisely, as Data Mining goals.

Our goals focus on analyzing the given data which has information about previous bookings of the clients as well as some booking date related information. From these two previous points we want to find a pattern in groups of clients that have canceled their reservation and create a predictive machine learning model capable of identifying weather or not a future customer is likely to cancel his reservation. Using F1-score and accuracy we will be able to evaluate the performance of the returned models and therefore choose the one that best learns, operates, and also generates good results for the business problem in question.

From our final model we will make some business approaches and possible ways to implement it in future business operations.

# 3. Data Understanding

### 3.1 Collect Initial Data

The dataset was present in an CSV document given by the Hotel chain C and we loaded it into a Jupyter Notebook using the pandas library and there were no problems.

### 3.2 Description of data

The dataset has a total of 79330 records each representing a reservation and a total of 31 features. In these features we can find information about each booking like for example the number of adults and children, how many nights they stay, type of room and many more.

Also, the dataset has a lot of useful information, it is very complete in terms of information about hotel bookings.

### 3.3 Exploring data

Since the variable “IsCanceled” is our target in this work we started our data exploration by checking how many canceled bookings we had in order to see if dataset was unbalanced and we concluded that it's not because about 60% of the bookings was not canceled and around 40% of our data show us canceled bookings.

We also performed the info() method to acquire information about the dataset like the number of features; type of the variables; memory usage; null values, etc. and noticed that we had many different types of variables (object, float, datetime)

To better understand the data, we also analyzed statistical information about all the features, and it matched with the previous information we had about these attributes (e.g: mean , max and min values) and used visualization tools to identify possible outliers.

And finally, we visualized the distribution of the variables and their correlations with each other.

### 3.4 Verify data quality

The data had some null values in the variables “Children” and “Country and the data types of some variables were not totally correct.

We noticed that there were many duplicated records (about 26 000) and some incoherent values like bookings with zero adults, not making sense in this context.

These were the main issues with this dataset and we will explain later on how we dealt with these problems.

# 4.Data Preparation

### 4.1 Selection and Cleaning of Data

To be able to have the desired data quality for our predictive models we made some changes in the original data so we:

1. Started by changing the data type of the features “Children” and “ReservationStatusDate” to integer and Datatime respectively.
2. Then we changed the outputs of the variables “Agent” and “Company” to zeros and ones meaning that if the booking has a zero value on for example company means that the booking was paid by a particular person and if it has an one value means that booking was paid by a company. Given that most of the bookings had null values for these variables and many different values we thought it is easier to understand and for the model to encode manually these two features.
3. Deleted the rows that had null values in “Children” and “Country” features.

4. Removed the duplicates values and left only the original records among those duplicates because we thought that so many bookings with the same information might be bias for our predictive model. With this action the dataset ended up with 53410 records.
5. Removed the records that had the incoherent value of “adults = 0” because it does not make sense having reservation with no adults on it. (370 records eliminated).
6. Finally, by checking for outliers using boxplots we removed just a few records because it could be bias for our modelling process to remove more. So, we removed outliers from the variables “Babies” ( no more than 5 ), “ADR”(no more than 1000” and “RequestedCarParkingSpaces” (no more than 2).

Note: It is important to refer that we checked the correlation between variables but there was no big correlation between any specific features meaning that there is no pair of variables that give us similar information.

### 4.2 Constructing the data

Given this data we find it useful to do some feature engineering, so we replaced the name of the months in the variable “ArrivalDateMonth” by the respective number of the month (e.g: January = 1 and December = 12)

Then create a new feature called “Seasons” where: months 12 to 2 = Winter, 3 to 5 = Spring; 6 to 8 = Summer and 9 to 11 = Autumn.

We also add two more columns to this dataset by putting the date of arrival all together ( year-month – day) called “ArrivalDate” and using this new feature we created the column “CancelationTime” which is comes from “ArrivalDate” – “ReservationStatusDate” and allows to understand if the customers canceled their reservation near the day of the start of the booking or if they canceled a long time ago.

All these new elements added to the data had the goal to help us build a better predictive model but also for a better analysis afterwards.

### 4.3 Formatting the data

To finish the data preparation and since we had numerical and categorical variables we standardized/encode all the values so that the model can be more accurate. For the categorical variables we used the one hot encoding so that the model can interpret better these features.

However, for the numerical features we used the Robust Scaler method because deals well with the presence of outliers and since we did not remove many we thought this method might be helpful in the case we did miss some outlier.

And finally, after all these steps our data was ready to be used in our modelling methods.

### 4.4 Data Analysis

We used visual tools in order to better understand the impact caused by each variable on the target.

We concluded that some of the variables that mostly affect the decision to cancel the reservation: Lead Time, Deposit Type, Special Requests and Arrival Data/Seasons. These visuals can be found in the appendix for consultation.

# 5. Modelling

## 5.1 Selecting modelling technique

Considering the problem, we knew that we were facing a supervised machine learning problem, so our first step was to get the target column (“IsCanceled”) away from the rest of the data.

From the columns left we decided to drop: “ReservationStatus”, “Country”, “ReservationStatusDate”, “ArrivalDate”, “Meal”, “ArriveDateYear”, “ArrivalDateDayOfMonth” and “RequiredCarParkingSpaces” because they were not relevant for this particular problem or because they could interfere with our model.

### 5.2 Building the model

Before starting to test the models, we used the Train Test Split method in order to divide the dataset into 70% training and 30% testing.

With our data ready for the models, we researched and discussed which methods we should try. We ended up trying the logistic regression, naive bayes, decision and regression trees, random forest and neural networks.

In order to evaluate how good our models were we decided to use f1 scores and accuracy scores to compare them. The models with the higher f1 score and higher accuracy both for training and for validation were the neural networks using repeated k-fold and the random forest.

When doing the neural networks model, we used grid search to get the best parameters. In the case of random forest we tried many different parameters and chose the one with the higher score.

The parameters we found were:

- Neural Networks: “activation = "tanh", alpha = 0.001, hidden_layer_sizes = (10, 10, 10), learning_rate='adaptive', solver ='adam'”

- RandomForest: “n_estimators=125, criterion="gini", max_features=None, random_state=5, max_depth = 20”

And their final scores were:

- Neural Networks: Accuracy Score = 0.8 train, 0.79 test / F1 Score = 0.8 train, 0.8 test
- Random Forest: Accuracy Score = 0.91 train, 0.79 test / F1 Score = 0.9 train, 0.74 test

Even thought the best scores were with random forest, the difference between train and test made us believe that this model could be overfitting and because of this we concluded that the best model was the NeuralNetworks.

# 6. Evaluation

### 6.1 Evaluate Results

Using different Machine Learning algorithms, we got the best way to predict what are the instances in which a client is more likely to cancel their booking. We consider that we have built a very solid model, that can give great insight on a very important part of the business, through which the management can make thoughtful and confident key decisions.

### 6.2 Review Process

When it comes to the complete working process, we consider having done a plentiful job, and that are no specific issues that could undermine the results. We consider that our level of accuracy is very good and that with this model, the company can get all the information it needs, regarding their customers’ cancelation behavior.

### 6.3 The next steps

When it comes to the improvement of our work, we conclude that with more time we could use some ensemble classifiers or try to find another singular algorithm that would better learn from the data.

Although this could be an interesting endeavor, the budget impact it would have, and the fact that we consider the final algorithm we presented to be very suitable for the project the client wants it for, made us disregard this next possible step.

## 7. Deployment

This model makes it possible to predict in advance the percentage of reservations that may be canceled which enables the company to overbook the rooms with a certain level of confidence without damaging the brand and lose the trust of their customers.

As a maintenance measure, we suggest the model to be re-runed every 6 months in order to include the most recent information and, in turn, delete the information from the older reservations so that the database and the model are always updated.

We would also like to make some suggestions for the future such as:

- Require an *a priori* payment of a determined percentage of the reservation amount (without refund), in bookings that include 5 or more adults and/or 4 or more nights in the hotel;
- Based on the predictions verified in the model, contact the customers that correspond to the characteristics of guests who are more prone to cancel;
- Implement a symbolic cancelation fee for everyone.

## 8. Conclusion

Using our prective model, we believe that Michael's company will be able reduce their cancelation rate to 20%. This won't be achieved imediately, as the model isn't 100% accurate, but it's a very good tool to lower the unweighted overbooking and the loss of money due to "no-shows".

We concluded that the hotel should foccus their attention in some specific features such as getting a better performance in terms of Lead Time and delivering the guest's special requests.

Finally, we noticed that one of this conclusions, that reservations with Non-Refund deposit tend to be more canceled, doesn't add up to our initial thoughts before analysing the data(, but in this case we need to trust the data, as we don't have evidences that it is wrong).

# Appendix

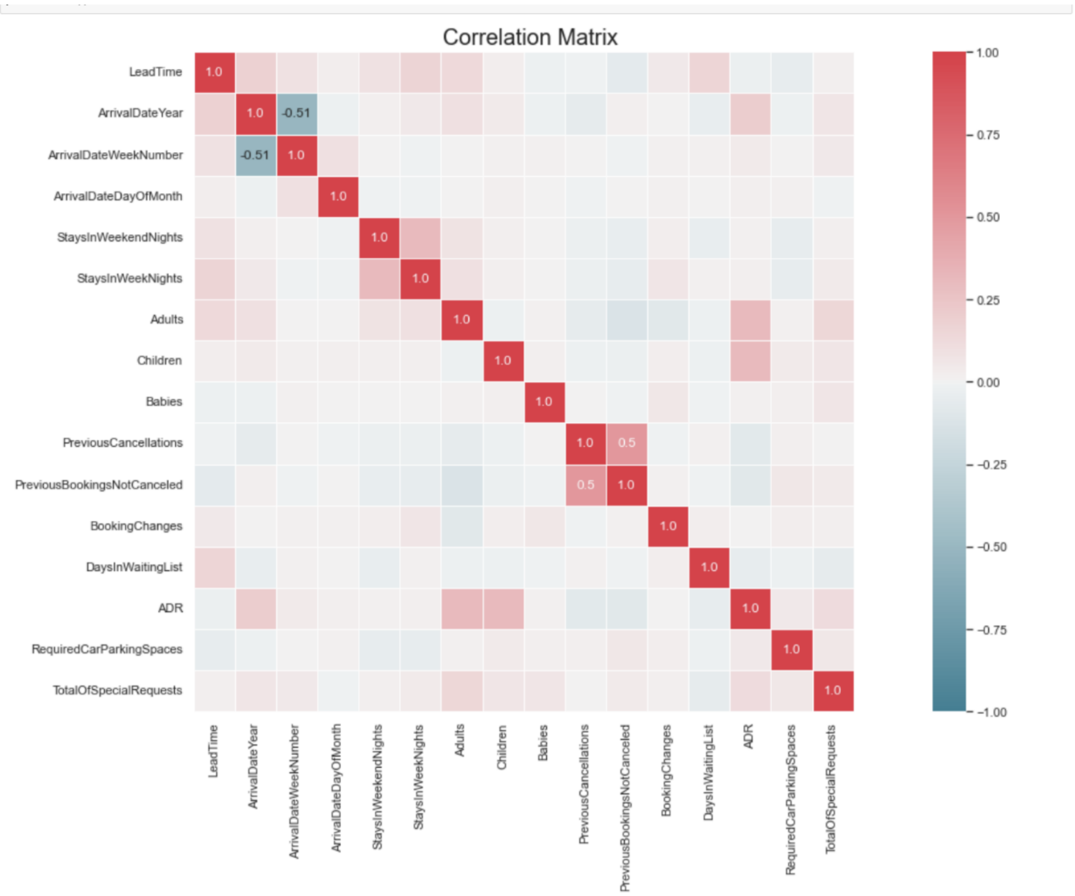

Fig. 1 – Correlation Matrix

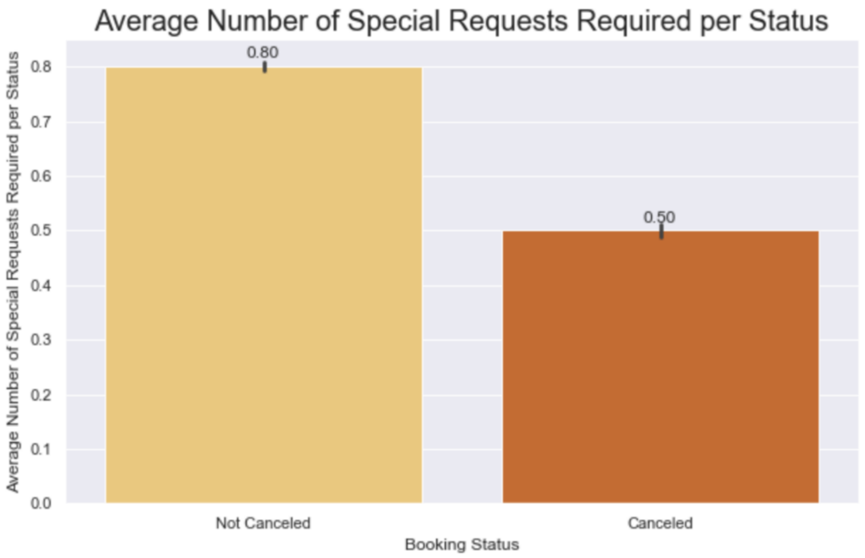

Fig. 2 – Average Number of Special Requests Required per status

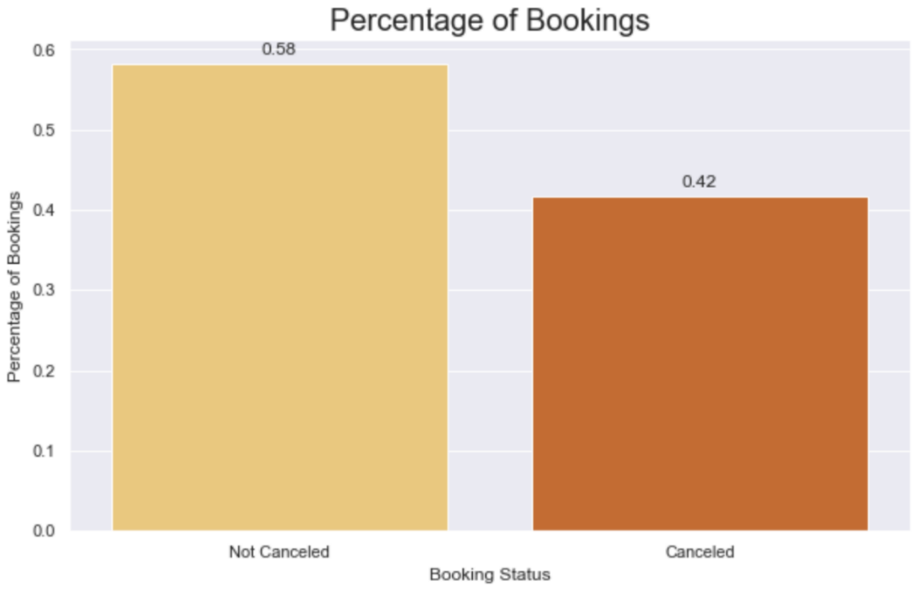

Fig. 3 – Percentage of Bookings

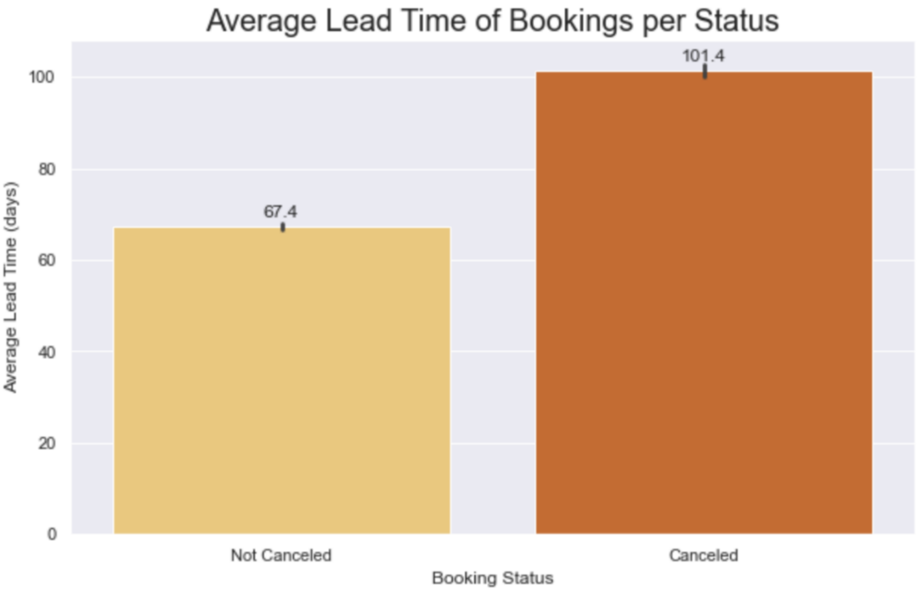

Fig. 4 – Average Lead Time of Bookings per status

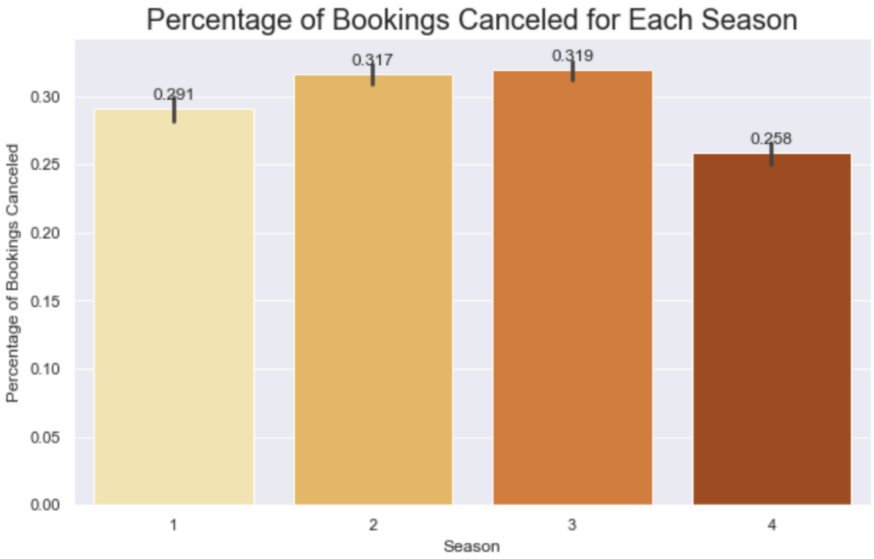

Fig. 5 – Percentage of Bookings Canceled for each season

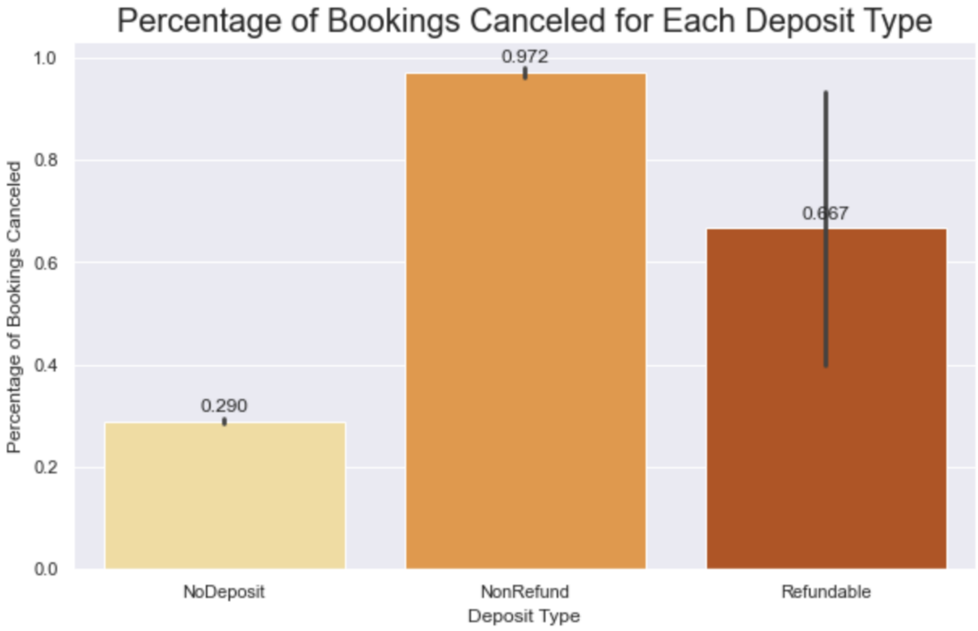

Fig. 6 – Percentage of Bookings Canceled for each deposit type

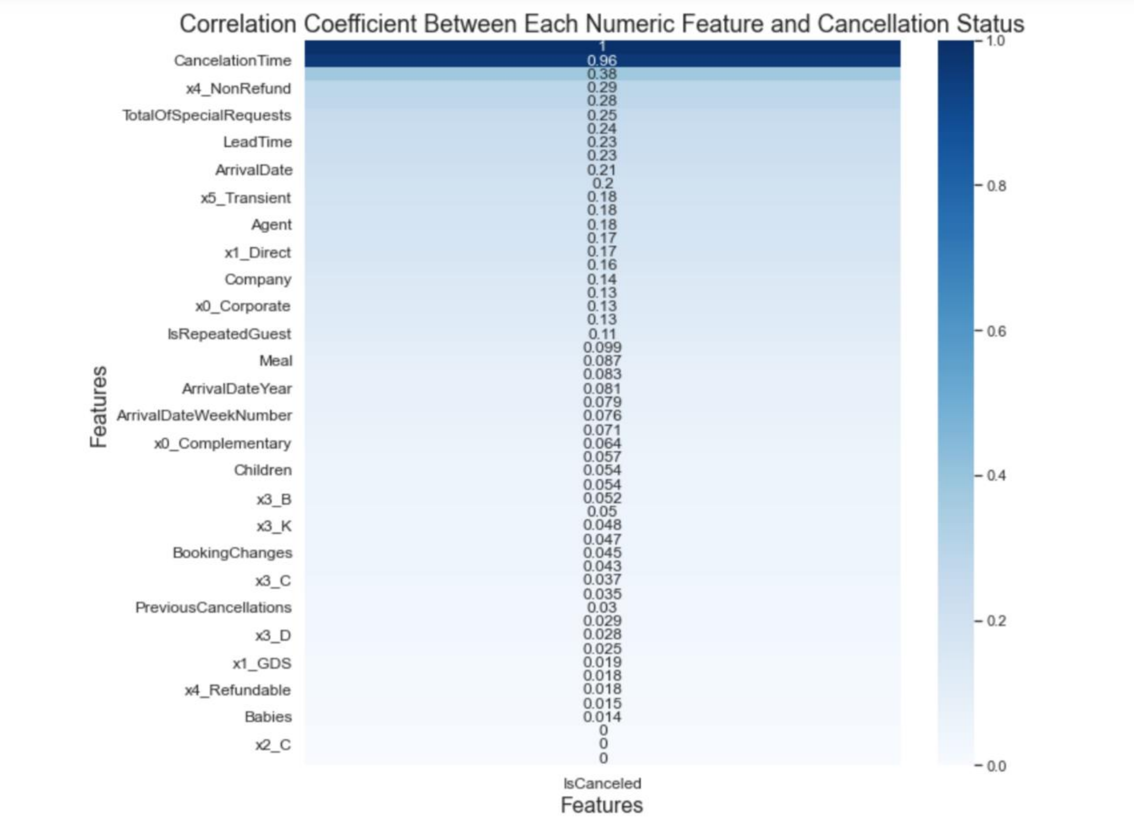

Fig. 7 – Correlation Coefficient between each Numeric Feature and Cancelation Status