BUSINESS CASES WITH DATA SCIENCE

Mestrado em Estatística e Gestão de Informação
MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Group A

Ana Sofia Silva, number: 20200220

José Francisco Alves, number: 20200653

Miguel Nunes, number: 20200615

Mohammadali Gharghi, number: 20200997

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

INDEX

1. INTRODUCTION

In the Hotel Industry, the cancellation of hotel bookings through the online system is one of a problem for the hotel management. The cancellations represent a significant lost of profits, and for that reason, there was the need to create model predictions, to reduce the uncertainty about this demand.

Therefore, it is suggested two ways to overcome this unpredictability, but it entails a great risk for the business itself. The overbooking, which can generate in customer reallocation, generating costs and trust loss, and also in social reputation damage. The other way is the restrictive cancellation policy problems, which will rebound in a decrease in demand and in revenue.

At the moment, the expected outcome is to make suggestions of how a model could be deployed and what impact could it have on the hotel's business processes. As so, a predictive analysis using the CRISP-DM framework is conducted.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

Hotel chain C, a chain with resort and city hotels in Portugal, was severely impacted by cancellations, representing almost 28% in H1 and almost 42% in H2, as shown in the table below. The Revenue Manager Director, Michael, hired a consultant to build a predictive model in order to identify booking with high likelihood of being canceled and with the aim of reduce the cancellations in H2.

Table 1 - Table H1 & H2

Hotel	Metric	Not Canceled	Canceled	Total
H1	Bookings	28,938 (72.2%)	11,122 (27.8%)	40,060 (100%)
H1	Room Revenue	11,601,850€ (66.5%)	5,842,177€ (33.5%)	17,444,028€ (100%)
H2	Bookings	46,228 (58.3%)	33,102 (41.7%)	79,330 (100%)
H2	Room Revenue	14,394,410€ (56.9%)	10,885,060€ (43.1%)	25,279,470€ (100%)

2.2. BUSINESS OBJECTIVES

The objective of this business passes by adapting and implementing better strategies of pricing and overbooking policies, by identifying bookings with high likelihood of canceling and try to contact those booking's customers in order to make offers preventing cancellation. The goal of this project is to reduce the cancelation rate by 20%. With this drop of the cancelation rate, it will be possible to plan better every operation occurring inside the business.

2.3. BUSINESS SUCCESS CRITERIA

Given the fact that is crucial to develop a predictive model in order to classify booking's cancellation, was defined the criteria to achieve a prediction Accuracy above 0.8.

2.4. SITUATION ASSESSMENT

This project was made in an interactive computing environment called Jupyter Notebook using Python programming language. The databases were installed on a MacBookPro computer that ran Mac OS X and Windows 10. The dataset was provided by Github platform along complementary documents explaining the general context and the business situation. Data understanding, data preparation, modeling, evaluation and deployment phases were all conducted on Jupyter Notebbok.

2.5. DETERMINE DATA MINING GOALS

In order to reduce the cancelation rate to 20%, the principal Data Mining goal consists of creating a predictive model that define which bookings are more probable to be canceled based on the information of records of each attribute inside H2 dataset.

In order to reduce the cancellation rate to 20% and create an effective model, the dataset should be prepared in a good order by some actions:

- Removing duplicate records;
- Filling missing values;
- Removing all of outliers;
- Choosing right columns for prediction.

3. PREDICTIVE ANALYTICS PROCESS

3.1. DATA UNDERSTANDING

In the data understanding phase, we start with an initial dataset and continues with activities required to enable modelers to become familiarized with data, including finding patterns, tendencies, and anomalies.

The dataset has 79330 rows and 31 columns. The datatypes look correct althought it was detected that '*Company'* column had 75641 values as "NULL" with type string, therefore it was necessary to replace these values to "NaN" considering them as missing values. The '*Company'* variable wasn't deleted due to the fact that the "NaN" means that its not a company but a customer (person). This variable represents an ID of the company, thus it was created a binary column for '*Company'* represented by "0" (not a company) and "1" (company).

Besides '*Company'*, there was missing values in 3 features, '*Children'*, '*Country'* and '*Agent'*. '*Children' and* '*Country'* columns had ratio of missing values below 0.5%, therefore we proceed with deletion of that rows avoiding the fill of missing data. Only '*Agent*' feature was filled with mode due to fact that had, approximately, 10% of missing values and was represetend by an of the travel agency.

We also found out 3 variables with high cardinality, which are: '*Country*', '*Agent*' and '*ReservationStatusDate*'. There are also 18551 duplicated values that are coincidence (referent to different bookings even though they have the same values), thus our group decided to maintain this records in the dataset.

After the understanding of data, there was crucial to divide the variables into categorical and numerical features. To understand the distribution of each group of variables, for the categorical features, it was made bar plots with absolute frequencies, and for the numerical features, it was made histograms and boxplots.

- 42% of bookings were cancelled;
- Redundant variables: '*nr.employment'*, *euribor3m*, '*emp.var.rate';*
- Irrelevant variables: '*ArrivalDateDayOfMonth'* doesn't show any discriminancy on the target, all values have the same frequency and the variable itself doesn't seem to provide any value;
- *'LeadTime'*, '*StaysInWeekNights'*, '*PreviousBookingsNotCanceled'*, '*BookingChanges'* and '*DaysInWaitingList'* might have some outliers.

4.1.1 Data visualization

By using LASSO, RIDGE and the Decision Tree classifier (Figure 2), the most appropriate features for the machine learning model are built and this will give important insights for business strategy.

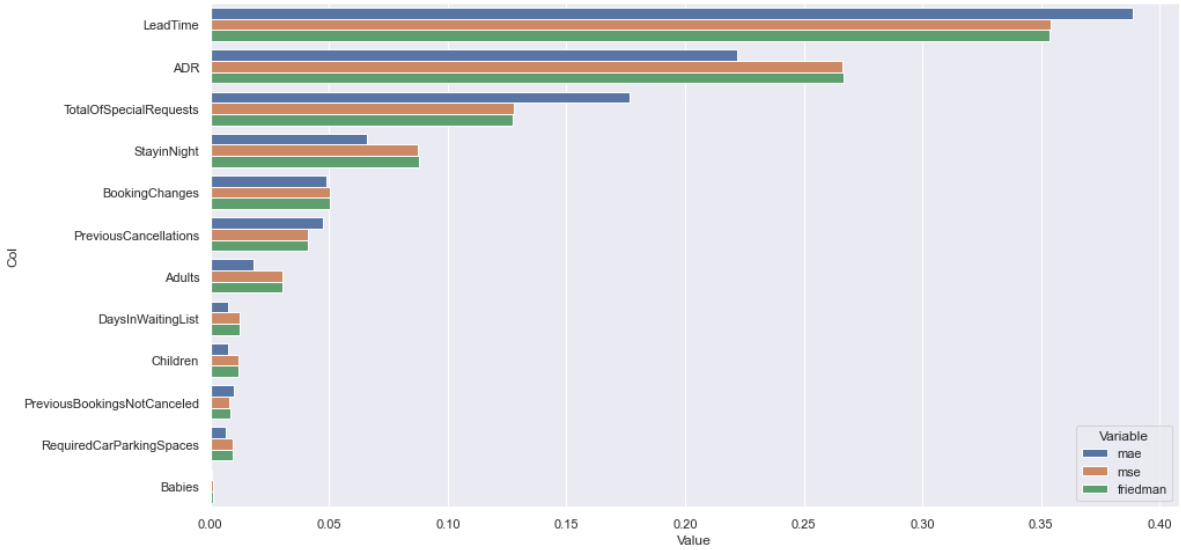

Figure 1 - Feature Importance with Decision Tree Classificator.

According to these methods, the variable *'LeadTime'* should have the most important role in the model's implementation - the number of days between the booking and the date when the customer arrives at the hotel is the most important variable for cancelling prediction.

Also, it can be seen in Figure 3, that with a short lead time is common to cancel the reservation, while bookings made more than one year before are very rare canceled.

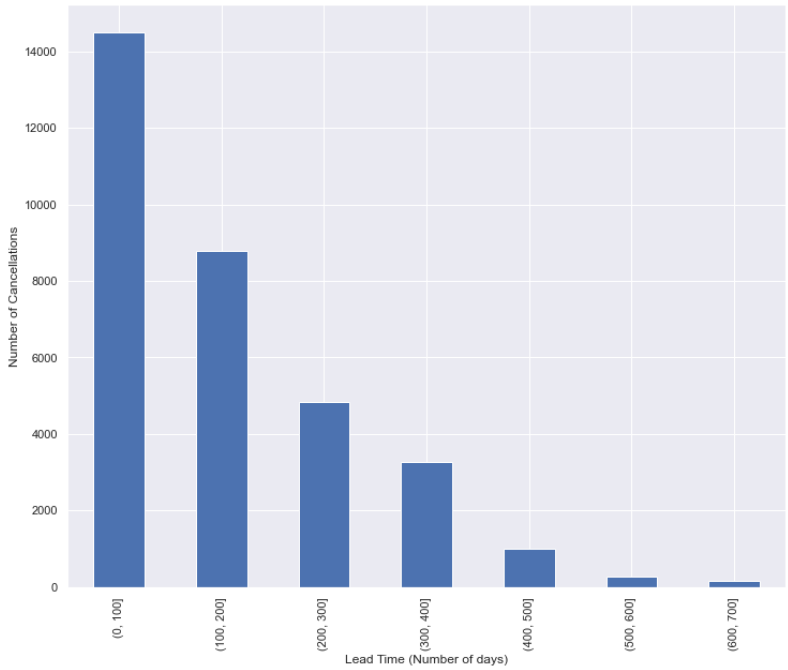

Figure 2 - Lead Time per Number of cancellations.

3.2. DATA PREPARATION

Inside the Feature Engineering topic, the variable *'IsRoomNotEqual'*, which returns a boolean answer, is the result of many of times that the room was booked corresponding to the real assigned room of the client. With this creation, the variables *'ReservedRoomType'* and *'AssignedRoomType'* were removed from the dataset. The variable *'StayinNight'* was also created and supported by the sum of variables *'StaysInWeekendNights'* and *'StaysInWeekNights'*, which were dropped right after. This new variable represents the number of nights that each client stayed at the Hotel. It was also necessary to use the one hot enconding for the variables '*Meal', 'MarketSegment', 'DistributionChannel'* and 'CustomerType' to convert the type of this variables into boolean in order to prepare these variables for the model analysis.

Relatively to the outlier analysis, it was decided to not remove a single outlier since the accuracy results were better with all the outliers.

There was also the need to proceed with the normalization of the metric features. With this normalization was possible to prepare the data for a correlation matrix analysis. From this correlation matrix analysis, it was possible to say that even knowing that the variable '*Agent'* has a high correlation to the most of variables present on data, it was decided to not remove this variable. This happens just because this variable represents the ID of the travel agency that made the booking, making it a possible important variable for an analysis in the future, because it defines a common characteristic between observations.

3.3. MODELING

In this phase, the selection and application of variations in machine learning models are carried out, so that the best model is obtained. Some techniques had specific requirements on the form of data. Therefore, going back to the data preparation phase was often necessary.

The data was divided into train and test set with a corresponding size of 80% and 20% of the entire dataset. This split will be based on random stratified (preserving target relative frequencies) sampling. We used a 10-fold cross-validation approach to evaluate each model 10 consecutive times and perform hyper-parameter tuning. Afterwards, we train the model on the entire train set with the best hyper-parameter configuration and use the test set to obtain a clean and unbiased estimate of the generalization capability of the model.

The model selection procedure will be based on a single measure: *accuracy* (number of correctly predicted data points out of all the data points). This measure was selected as one of the business goals in order to be possible to reduce the cancellations to a rate of 20%.

In order to provide different results, our group developed 3 models using different classification algorithms and then we selected the one with better accuracy. The models used were Logistic Regression, Decision Tree and Random Forest.

For this specific task we opted to use a Random Forest model where we applied the best parameters with GridSearchCV function. Random forest is a combination of decision trees that can be modeled for prediction and behavior analysis and present estimates for variable importance. Among all the available classification methods, random forests provide the highest accuracy.

4. RESULTS EVALUATION

To improve the final results, the method K-Fold was used for partitioning the dataset in order to ensure all of the observations would be included in all of the Train, Validation and Test partitions. The model was built based on Decision Tree, because there were many boolean and categorical values, which returned a good output in this technique. After that, it was used bagging ensemble to simulate a bigger dataset to have a more accurate outcome.

After understanding, preparing and modeling our problem, the results were:

Table 2- Models Assessment.

No.	**Modelling Result**	
	Model	**Accuracy**
1	Logistic Regression	0.7314
2	Decision Tree	0.7996
3	Random Forest	0.8341

The best model that gave the best accuracy, was the Random Forest, which also exceeds the business success criteria defined above. With this prediction, it is possible to predict cancellations rate at 83,41%.

5. Conclusions

Predicting cancellation of hotel room reservations is important to minimize the loss of income to the company. The use of predictive analysis in cases of cancellations of hotel rooms can be done using the CRISP-DM framework. By using the stages in CRISP-DM, the most appropriate features for the machine learning model are built. The best machine learning model is Random Forest, with an accuracy of 0,8341 and the time difference between bookings made and time arrival is the most influential feature to predict the cancellation rate of a hotel booking. It's possible to conclude that orders that have a short lead time have the higher number of cancellations

For the deployment and for a better management of expectations, the integration of these machine learning techniques will be crucial to minimize the loss of income, providing better predictions of future cancellations.