MDSAA

Master Degree Program in Data Science and Advanced Analytics

- Major in Business Analytics

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

Group C

Catarina Moreira, number: 20201034

Luisa Barral, number: 20201045

Madalena Valério, number: 20200657

Yu Song, number: 20200572

February, 2021

INDEX

1. INTRODUCTION

A Hotel Chain by the name of C has been severely affected by the cancellation of bookings in its partnership with OTA. For rising unpredictable cancellations, Michael, Revenue Manager Director had to balance between limiting the amount of rooms sold with restrictive cancellation policies and overbooking. As a result we are invited to develop predictive models to foresee the net demand for hotels based on a dataset with the bookings made in hotel H2, which were due to arrive between July 1, 2015, and August 31, 2017. Based on the estimates of these models, Michael hopes to implement better pricing and overbooking policies, and determines that reservations are likely to be cancelled. Make sure that these reservations allow the hotel to try to contact the customers of these reservations and offer discounts to avoid cancellation.

2. BUSINESS UNDERSTANDING

2.1. Background

Hotel Chain C is a chain with resort and city hotels in Portugal. Like many hotel chains it has been impacted by high cancellation rates, almost 42% in the case of its city hotel (H2), that we will be studying. To combat this problem it has employed several strategies, such as overbooking and restrictive cancellations policies, but these come with their own side effects and costs. For this reason they are looking to implement a predictive model to predict whether new bookings will be cancelled, manage their business accordingly, and increase their revenue.

2.2. Business Objectives

The hotel wants to reduce cancellation rates by predicting which customers are more likely to cancel. This way they can make more strategic choices about pricing and overbooking policies. In addition, they can contact the customers that are likely to cancel and offer them extra amenities in order to prevent the cancellation.

2.3. Business Success criteria

A useful outcome of this project would be to reduce the costs associated with high cancellation rates, by giving insights into what types of customers will cancel, so the hotel can preventively overbook those rooms and not lose revenue from inventory not sold, or make special offers to those that it expects to cancel. Ultimately, the goal of the Revenue Manager Director of the hotel chain, Michael, is to reduce cancellations to a rate of 20%.

2.4. Situation assessment

Our team was given a csv file that contains detailed booking information of Hotel H2 such as number of children and BookingChanges. Our main goal is to forecast actual demand taking into account the cancellation rate. The reservation status has three possible states: cancered, check-out (the customer has checked in but has left) and no-show (the customer booked but did not show up). Identifying these reservations allows the hotel to try to contact the customers of these reservations and provide discounts to reduce the cancellation rate to 20% (for example, dinner, parking, spa, discounts, or other benefits). In addition, we have some metadata about the dataset.

2.5. Determine data mining goals

We useed our hotel dataset and binary cancellation features as target variables to create models that can classify hotel reservations. We are aiming for the highest level of accuracy. So, we compare the baseline model to the logistic regression, a random forest, a decision tree, a gradient boost classifier and neural network. Despite the accuracy, data mining goals will be shown in the context of data understanding by identifying the Correlation Matrix of variables with relevant analysis. Those important variables are more understandable through the statistical data visualization.

3. PREDICTIVE ANALYTICS PROCESS

3.1. Data understanding

We started by reading the description of each variable from the dataset provided to us. After that, we loaded the data to a pandas Dataframe. Our dataset has 79330 rows and 31 columns. With some exploration, we identified that a couple of data types were incorrect (e.g. ‘Children’, ‘ReservationStatusDate’), we found that we have very few missing values in ‘Children’ and ‘Country’, there’s also features with values NULL (‘Agent’ and ‘Company’), there’s some high cardinality features (e.g. ‘Country’, ‘Company’) and finally, there are 31748 duplicate observations in the dataset.

As stated before approximately 42% of bookings resulted in a cancellation and 1% of these are No-Shows. Some of the variables that have a higher correlation with the target variable ‘IsCanceled’, according to the Phik Matrix (Figure 1), are ‘ReservationStatus’, which is obvious and will not be included in the model, ‘LeadTime’ and ‘TotalOfSpecialRequests’, which need to be studied further.

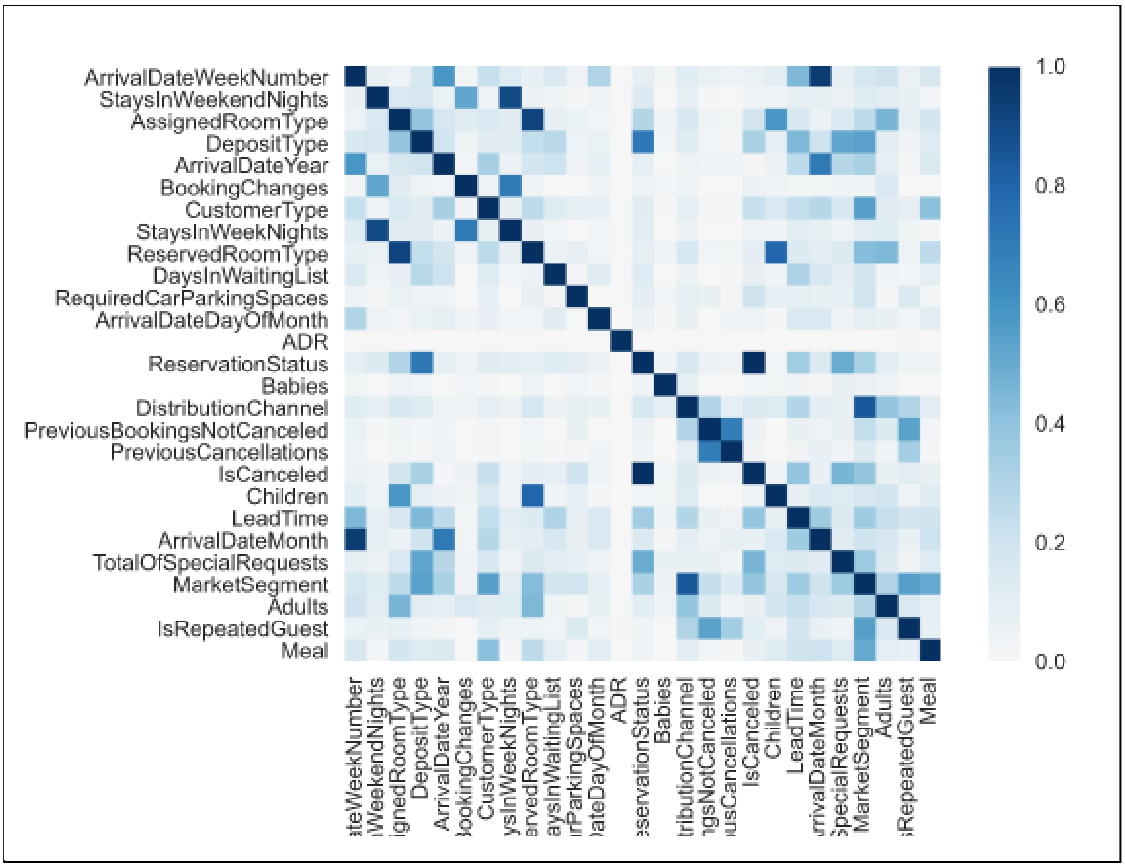


Figure 1 - Phik Correlation Matrix

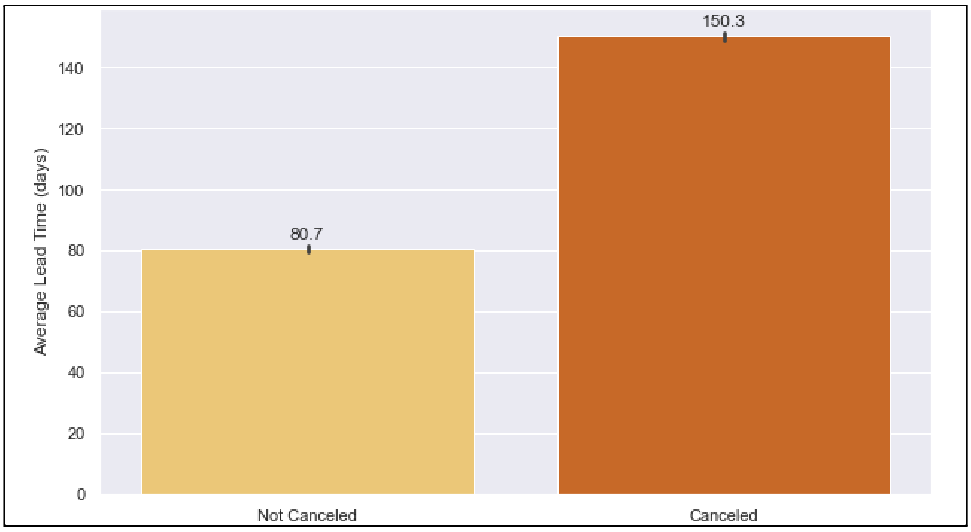


Figure 2 - Average Lead Time for each Category

As we can see in Figure 2, on average Canceled bookings have a longer Lead Time. This can be due to customers having more time to cancel and find a better deal in another hotel, or possibly having unexpected events occur that prevent them from traveling.

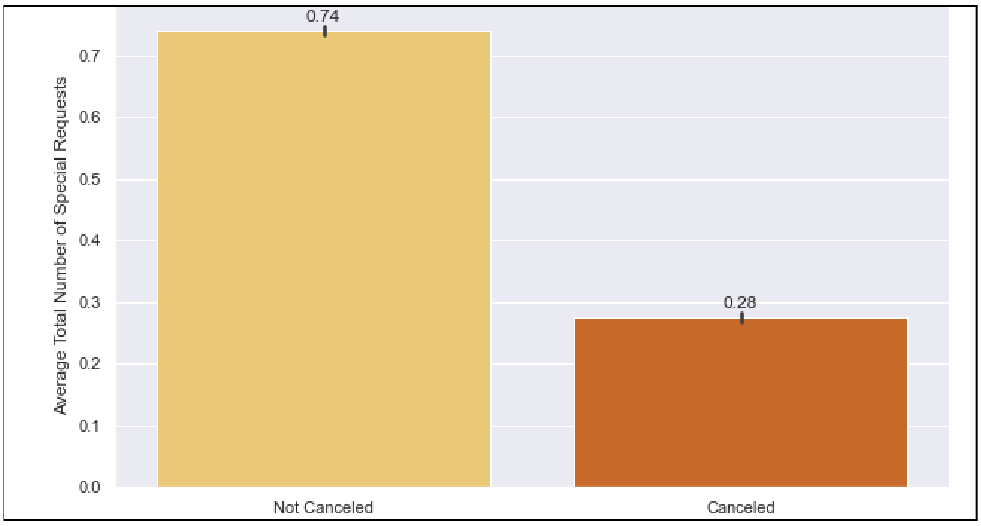


Figure 3 - Average Number of Requests for each Category

From Figure 3 we can conclude that bookings that are Canceled make on average fewer requests. This suggests that customers that make more requests are more likely to not cancel.

3.2. Data preparation

As stated before, we have many duplicate observations, but it is possible that different bookings have the same features. Since there are no IDs we can't say for sure if they are really duplicates, therefore we decided to keep them. In the future, the bookings should come with an ID so this situation can be avoided.

As for missing values, there were only 4 missing values in the variable 'Children' and 24 missing values in 'Country'. We decided to delete all the missing values in 'Children', because they were so few and for 'Country' we filled with *Unknown,* but we ended up not using this feature for our predictive model because its quality is dependent on check-in.

We found that we had a lot of outliers and using the IQR method, it showed that 72% of our data was outliers. However, since these observations are not errors and are still valid, we chose to keep them for our model.

Feature Engineering

For the variables 'Agent' and 'Company' we assumed that the NULL value meant that the booking was made without Agent or Company. In the case of 'Company' most of the values were NULL, and both 'Country' and 'Agent' had High Cardinality. Given that the specific ID of the Company or Agent was irrelevant for our analysis, we decided to convert them to binary variables, where 1 meant they were booked through a Company/Agent and 0 if they weren't.

We also created a new variable called 'StatusToArrivalDate', that represents the number of days a customer stays in the hotel, if the booking wasn't cancelled or the number of days before expected arrival the client cancels, if the booking was canceled. With this variable we plotted Figure 3.1, where we can see that on average people that don't Cancel stay 3 Nights at the Hotel and those that Cancel, do it on average 3 days before their Arrival Date.

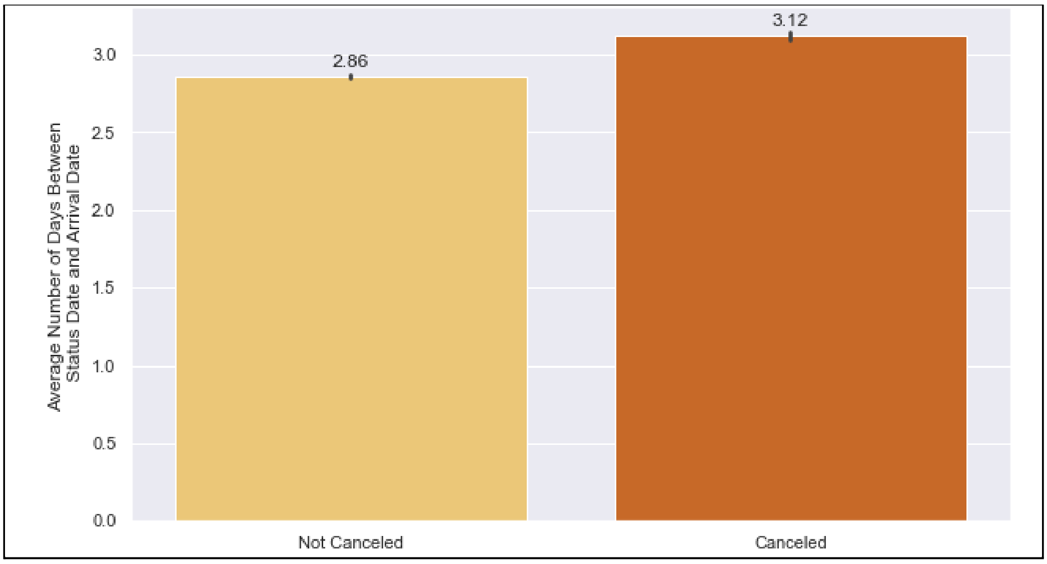


Figure 3.1 - Average Number of Days between Status and Arrival Date for each Category

3.3. Modeling

Before applying any model, we decided to do some data preparation before. Firstly, we applied the one hot encoder to convert the categorical variables into numerical categorical variables so that we can use it in the application of predictive models.

After that, we remove some variables from our X to do a better predictive model. At the beginning we removed 'IsCanceled' because our target variable. After this, we removed 'ReservationStatusDate' and 'ArrivalDateFull' because they are date_time. We also decided to remove the 'Country' feature because almost 50% of the bookings are made by Portuguese people.

With the date preparation done, we split matrices into random train and test subsets and we applied the models.

Our main goal was to build a model that predicted the accuracy of if the booking will or not be canceled and for that reason we took several approaches to find a solution to the problem.

For each model, we calculated the F1-Score, the Recall and the Accuracy. However, we decided to only compare the accuracy between the models.

The models that we applied were: Baseline Model that we compared with the Logistic Regression, Random Forest, Decision Tree, Gradient Boost Classifier and the Neural Network. We tried to use the Naive Bayes Classifier but we decided not to include it because its performance/accuracy was quite inferior compared to the other models.

Both in Decision Tree and the Gradient Boost models, we decided to check which variables were the most important. After that we removed (more or less) the 10 less important variables and we applied the model. In the Decision Tree, when removing these 10 variables, there was no significant change in the accuracy of this model. On other hand, when removing these 10 variables in the Gradient Boost Classifier, with the Grid Search the accuracy of the model decreased. However, using the Grid Search, we concluded that it improved the model.

To choose the best model we changed the different parameters and we concluded that the best model to predict the booking hotel cancelation were the Neural Networks.

Although the Gradient Boost presents a good accuracy, this model compared to the Neural Networks presents overfitting. So, to predict the booking cancelations we decided to use the Neural Networks.

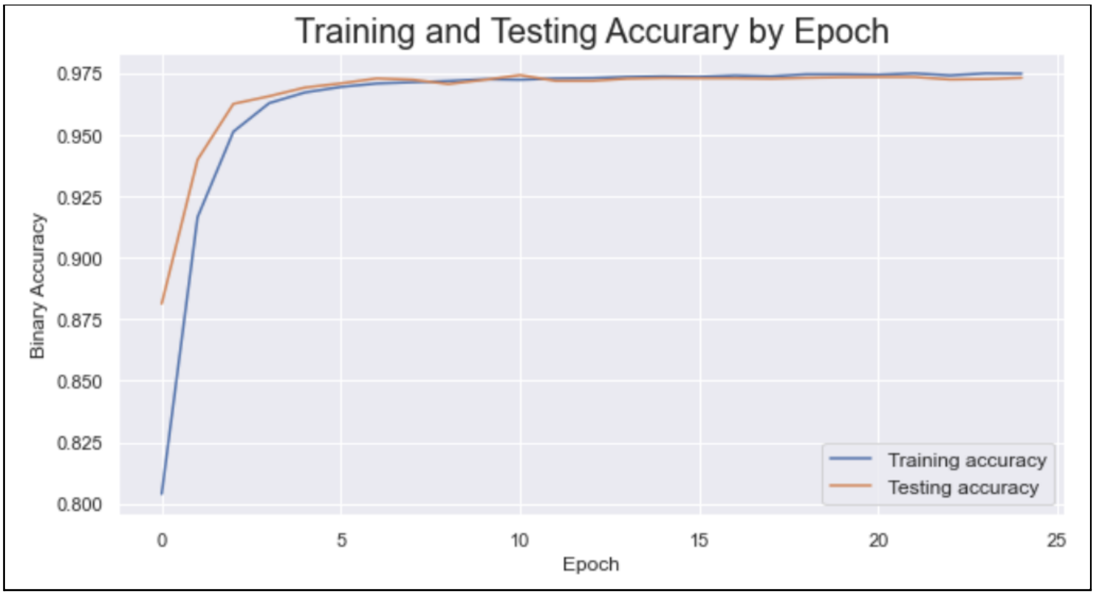
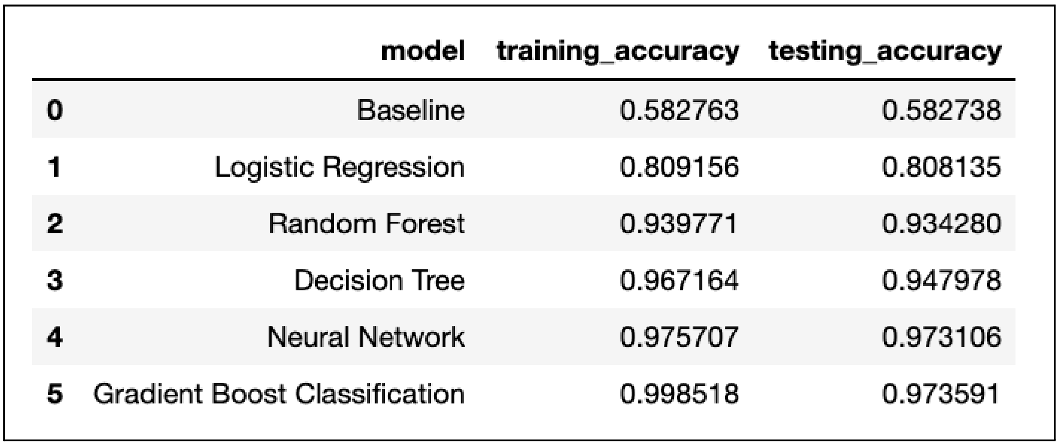


Figure 4 - Training and Testing Accuracy by Epoch

As we can see, the training and the test accuracy after the 5 epochs are very similar and after the 11 epochs this accuracy is 97.5%.

3.4. Evaluation

After applying the predictive models, we put the training and the testing accuracy of the different models in a table - ***Figure 5***. After that, we made a graphic - ***Figure 6*** - with the model's accuracy.

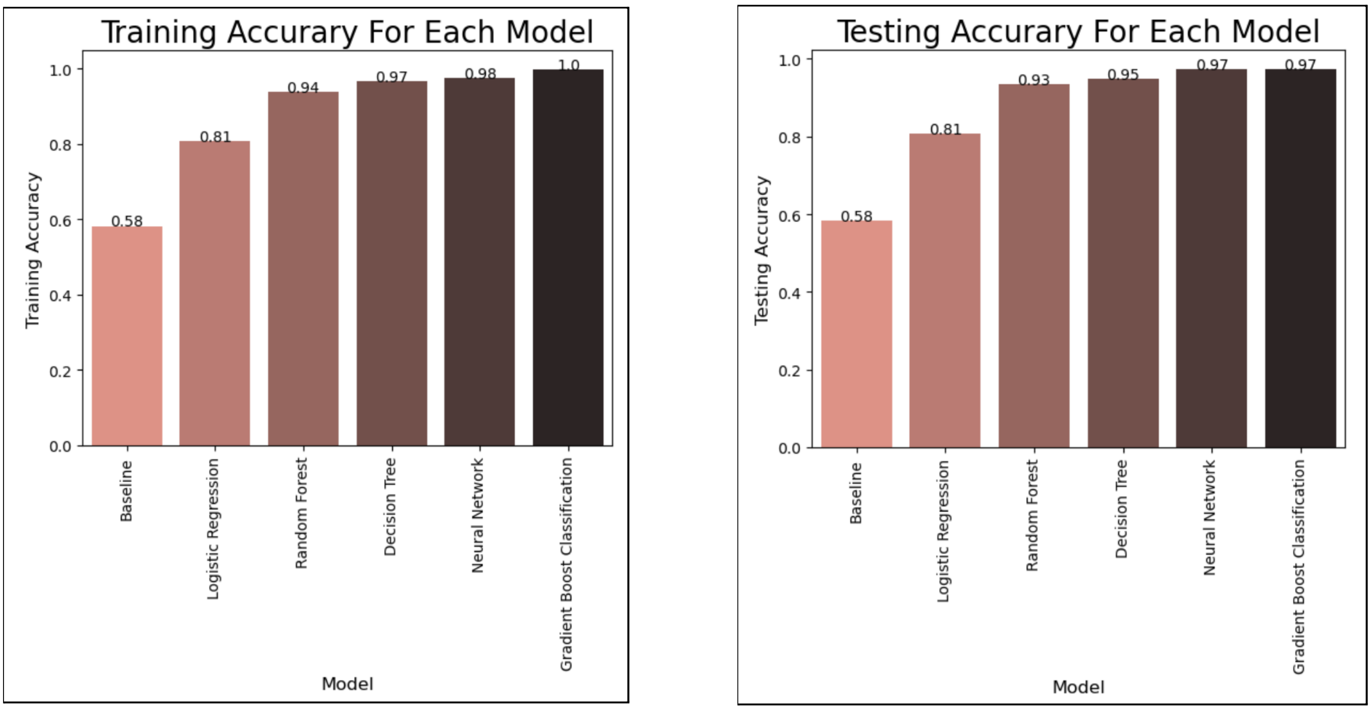


	model	training_accuracy	testing_accuracy
0	Baseline	0.582763	0.582738
1	Logistic Regression	0.809156	0.808135
2	Random Forest	0.939771	0.934280
3	Decision Tree	0.967164	0.947978
4	Neural Network	0.975707	0.973106
5	Gradient Boost Classification	0.998518	0.973591

***Figure 5* -** Table with the Different Models and the Training and the Testing Accuracy

Through ***Figure 5***, we can conclude that despite the Gradient Boost Classification presenting a higher training and testing accuracy, this model is slightly overfitted because the training accuracy is slightly higher than the testing accuracy.

On the other hand, we can notice that the Neural Network has higher training and higher testing accuracy. However, contrary to the Gradient Boos Classification it does not present overfitting.



Training Accurary For Each Model

Testing Accurary For Each Model

Figure 6 - Training and Testing Accuracy for Each Model

To conclude, through ***Figure 5*** and ***Figure 6***, we can conclude that the best model to predict the booking cancelations is the Neural Network since they present a higher accuracy and they are not overfitted.

After that, we made the predictions with the Neural Network.

4. RESULTS EVALUATION

After choosing the fittest predicting model which is Neural Network, we create and visualize the confusion matrix to see how our model works. We could classify 97% of the bookings to check if they are about to cancel or not, In which, we are correctly predicting 95% of the canceled bookings and 99% of the not canceled bookings. Looking into our confusion matrix, we can see that 271 bookings that our model predicted to be not canceled and that were actually canceled, which is 3.8% that we can accept to guarantee accuracy. Our model predicts 63 bookings that were canceled and they are not canceled, by communicating with these clients we could still keep the reservation.

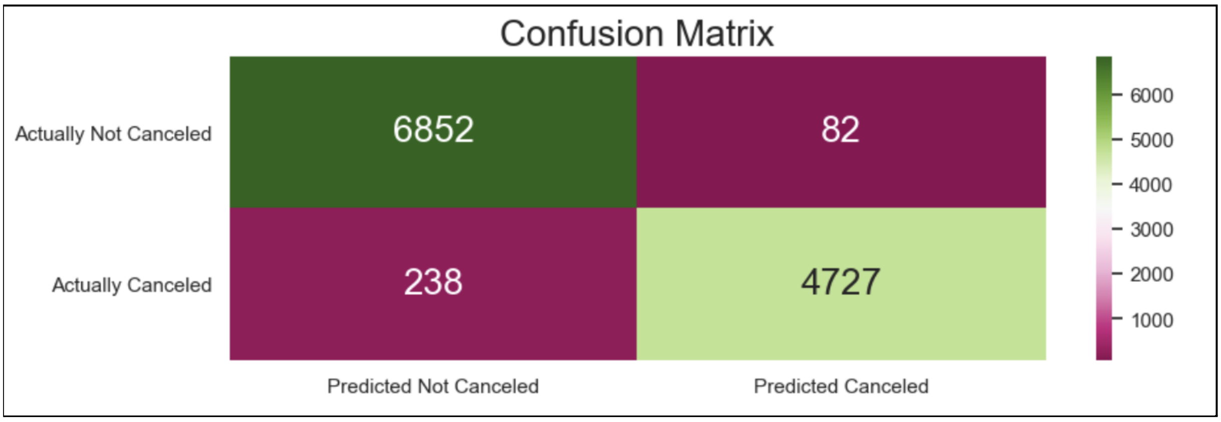


Figure 7 - Confusion Matrix

Hotel 2 is another experience of our group applying supervised learning algorithms. The dataset has 79,330 entities and 31 columns, we changed data types, deleted some columns and converted to meaningful ones. We used 'IsCanceled' as our target and plotted the importance of variables to the target.

The neural network structure won't give any insights on the structure of the function being approximated. So, we decide to use the Gradient Boos Classification as our interpretive model and to discover the most important variables in the prediction.

Using this mode, we search about the variable importance. In ***Figure 7***, we can see the 8 most important variables, the variables that have more weight in predicting if the booking is or not canceled.

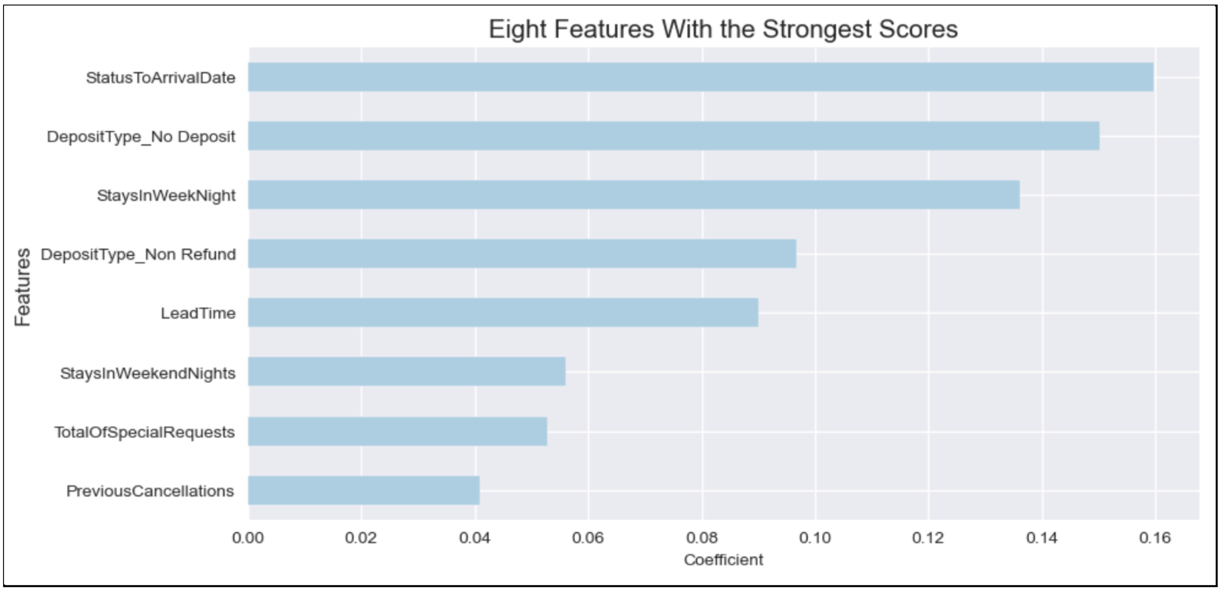


Figure 8 - Variable Importance

5. DEPLOYMENT AND MAINTENANCE PLANS

The group discovered a few problems in the data that can be solved in future data analysis, especially in collection and cleaning efforts. Firstly a variable with the full date of the arrival should be created, as well as a variable that calculates the number of days between the full date of the arrival and the date of the reservation status. This will either give us the number of days a customer stayed in that hotel or the number of days (before arrival) for a customer to cancel.

We believe this project should be deployed according to the MLOps principles, which means that we comply with automation and monitorization at all steps on the construction of a machine learning system. The group suggests the creation of an integration user as well as an API key for that same user. The API key is then used by the DevOps, which is an usual practice in developing and managing large software systems, for Maximo APM - Predictive Maintenance Insights to create a corresponding user account which enables Maximo to retrieve information and also generate predictions based on that information.

After deployment we should monitor the data, since it could change and therefore diverge from our original model, as well as our model performance, if it is still valid and accurate. For the monitorization of the data the group suggests the use of Maximo APM since it provides us with two main components to manage the “health” of our assets which is the MAHI - Maximo Asset Health Insight as well as PMI - Predictive Maintenance Insight. Maximo Asset Health Insights, gives us a higher level insight of the health status of our assets, which the most important thing to assess is basically if the asset is doing what it is supposed to. PMI - Predictive Maintenance Insights so that we can access predictive maintenance information from

6. CONCLUSIONS

As said before, the business objective is to reduce cancellation rates of the Hotel Chain C by predicting which customers are more likely to cancel. After a good data preprocessing the group deployed 6 different predictive models and ended up choosing the Neural Networks (NN) since it was the model with best accuracy and, unlike Gradient Boost, it didn’t overfit. After selecting NN as our predictive model for deployment we used the confusion matrix to calculate the net demand and therefore help the hotel, providing them with the number of reservations that they need to be ready for. As for reducing the cancellation rate, the group suggests a few different but pretty simple strategies that should be applied to the clients that might cancel their stay. We found out that most of the cancellations are done 3 days prior to the arrival date, as well as a lower number of special requests might indicate that this reservation might be canceled in the future. With that being said we suggest that the Hotel sends an email to the clients that will very likely cancel their stay, with an offer of a voucher to our restaurants, bars and SPA. Another possible marketing strategy, this time to avoid the negative impact of the Online Travel Agencies, the group suggests the creation of a membership system in order to cultivate our clients loyalty with more attractive privileges.

With the high accuracy score of our predictive model, alongside our prediction and possible marketing strategies, we hope that the Hotel Chain C will make the necessary decisions to reduce their number of cancellations

6.1. Considerations for model improvement

The group discovered that a deeper and better understanding of the situation may require more specific informations about the hotel, such as deposit policies, how the surrounding parking availability is or even how well-known the hotel is, if people enjoy staying there .The group suggests that working with the hotel’s stakeholders directly and trying to build specific models for the hotel, may be a good next step to take towards improvement.

REFERENCES

[1] Brownlee, J., 2021. *Your First Deep Learning Project in Python with Keras Step-By-Step*. Machine Learning Mastery.

[2] Scikit-learn.org. 2021. *sklearn.ensemble.GradientBoostingRegressor — scikit-learn 0.24.1 documentation*.

[3] Data School. 2021. *Simple guide to confusion matrix terminology*.

[4] Medium. 2021. *Understanding Gradient Boosting Machines*