BUSINESS CASES WITH DATA SCIENCE

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Business Case 2 – Predict Hotel Booking Cancellations

Group R

Andreia Bastos, number: 20210604

João Silva, number: 20211014

Pauline Richard, number: 20211019

Tiago Quaresma, number: 20210766

March, 2022

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

INDEX

Index of Figures

INTRODUCTION

In the hotel industry, as in many other travel-related industries, demand is managed through advanced bookings. Bookings (also known as reservations) are a forward contract between the hotel and the customer that gives the customer the right to use the service in the future at a settled price, but often with an option to cancel. This cancellation option puts the risk on hotels who must honour the bookings that they have on-the-books, but, at the same time, have to support the opportunity costs of having vacant rooms, when someone cancels, and there is no time to try to sell the room or sell it at a discounted price.

For this project, we are taking into consideration data about H2, a city hotel from hotel chain C. Our main goal is to develop a predictive model capable of forecasting the outcome of a booking, i.e., whether it's going to be cancelled or not. To achieve that we will be following the Crisp-DM method.

1. BUSINESS AND DATA UNDERSTANDING

1.1. Business Context and Objectives

Cancellations can occur for multiple reasons, such as vacations rescheduling, illness, or even when a better deal is found. "Deal-seeking" customers, tend to make multiple bookings for the same trip or make one booking, but continue to search for better deals. Hotel chain C was severely impacted by cancellations, representing almost 42% of the bookings in H2. To deal with this, two approaches were taken, restrictive cancellation policies and an overbooking policy. However, both methods can be prejudicial to the hotel and result in a loss of revenue and/or reputation, which we want to avoid. Therefore, since neither is the optimal way to proceed, they came up with the idea of using a booking cancellation prediction model. This model should allow the chain's hotel to forecast better net demand based on reservations on the books and know which booking to favor or not (minimizing the uncertainty). The end goal of this is to reach a cancellation rate of 20%. In order to do achieve that, we will go through three main steps: identify the features that have the best predicting power; build a predictive model and interpret its performance in the business context; lay out the deployment and application to the business.

1.2. Data Description and Assessing Data Quality

In order to do this, we have been handed a dataset composed of 79330 observations and 31 variables, including our target variable 'IsCanceled'. As 42% of the bookings are cancelations, we can describe our dataset as slightly imbalanced. To start off the data exploration, we checked for duplicates and found 25902 observations. After a brief discussion, we decided to eliminate them, which corresponds to 32.65% of the data. Even though the given variables cannot be used as a unique identifier of a specific booking, the predictive models would likely be biased.

When taking a deeper look at our data, we noticed there were only two variables with missing values. The missing values in 'Country' were deleted, since the quantity was very small, whereas in Children they were replaced with zero. However, there are two variables, 'Company' and 'Agent' that present 'NULL' values, which we concluded to be normal, as the correspondent bookings did not have agent or/and company. Furthermore, we also found observations with no adults. Even though this

could represent an event held at the hotel or other services the hotel might provide, we still decided to delete them since they are not relevant to our study.

1.3. Data Visualization

Additionally, through a series of plots, we also analyzed the relationships of variables we deemed important with our target, and also amongst themselves. We started by analyzing the cancelations per year but since we don't have data about all of 2015 and 2017, we ended up looking at it from a monthly perspective, where we can clearly see that cancelations increased significantly throughout the years, as we can see in Figure 1.

We also noticed that cancelations were more common amongst new guests, as expected. Regarding the countries, it is also worth mentioning that the higher number of successful reservations are booked by Portuguese customers, followed by the French. On the other hand, the majority of the cancelations are also done by Portuguese customers (7,4%). Finally, we can visualize on the bar chart in Figure 2, that most of the cancellations happen when it has been booked by the clients through a Travel Agent/Operator but also that the latter represents more than 83% of the bookings made (and only 11% for the Direct Channel).

2. DATA PREPARATION AND TRANSFORMATION

2.1. Removing Outliers

To detect outliers, we applied a Density-based clustering algorithm (DBSCAN) since it is robust to outliers. When performing it (with the parameters MinPts and epsilon defined as 4 and 32), the percentage of data kept after removing outliers was 99.4%.

2.2. Feature Engineering

We decided to create 2 new variables that will minimize the number of features and at the same time capture the same amount of information. We thought about generating the Total Number of Nights per clients and per stay, by adding up 'StaysInWeekendNights' and 'StaysInWeekNights'. Moreover, we also applied the same strategy to get the Total number of children (= under 18) per reservation, by adding up 'Children' and 'Babies'. We decided to NOT add the 'Adults' variable to this total because having a reservation with or without children seems to impact more a cancellation decision.

2.3. Feature Selection

Moreover, we ran different feature selection methods to decide which variables to keep and which ones to remove. The methods we applied helped us decrease the number of features we had and to keep only the most relevant ones to predict cancellation.

To determine the important features among **categorical variables**, we used the Chi-square method. After running the algorithm, the results we got were showing all variables with significant importance, so we deepened our analysis using the function cross tab of Pandas. We visualized the relevancy of our categorical variables with our target with a bar chart, as we can see on Figure 3.

Therefore, we were able to conclude that the categorical variable 'Meal' was not very relevant to our model. Regardless the value taken, the proportion of 'IsCanceled' was quite the same for all values, except for FB but it only represents 44 values of the dataset. Same conclusion for 'DistributionChannel' (also redundant information to 'Market Segment').

To determine the **useful numeric data** to our model, we followed two approaches. First, we checked for Redundancy, to see if some features are giving the same information to the model: a Spearman and Phix correlation heatmap (Figure 4.1 and 4.2). After analyzing it, we noticed that the 'ReservationStatus' variables were very correlated, because they were used to build other variables such as 'IsCanceled'. We also saw that Adults doesn't seem to have any correlation to our target. To complement this analysis, we used wrapper and embedded methods. First, we used recursive feature elimination (RFE), then Lasso Regression and decision tree importance were applied.

After considering all the methods applied, we removed the following variables: ADR, Adults, DaysInWaitingList, ArrivalDateMonth, ArrivalDateYear, ArrivalDateDayOfMonth, ReservationStatusYear, ReservationStatusMonth, ReservationStatusDay, Meal, DistributionChannel, Company, ReservationStatus, ArrivalDateWeekNumber.

3. IMPLEMENTING DIFFERENT MACHINE LEARNING MODELS

Since time variables are present within our data, our first approach was to consider the dataset time dependent and use the TimeSeriesSplit from sklearn to split the data. However, since the customer decision to cancel the reservation or not isn't directly affected by the date, we did the final decision to not take the time into account. Moreover, since we have an imbalance dataset, we decided to apply the Stratified K-Fold to split it, so both classes (0 and 1) are equally represented in both train and test. We applied 10 folds.

Finally, our target 'IsCanceled' can only result in a binary outcome (0: no and 1: yes), so we decided to try the following two-class classification algorithms: Random Forest, Catboost and XGBoost. For our best models we applied a Halving Grid Search to find the optimal hyperparameters.

4. EVALUATION OF THE RESULTS

Before arriving at this stage, several combinations (features, models and parameters) were tested and we even had to revisit previous phases of the CRISP-DM. The Table in Figure 5 reviews the global performance (average measured in each of the 10 folds) of our 3 most relevant algorithms. For the cross-validation results, since we are working on imbalanced data, Accuracy is not self-sufficient as a metric score, we also need to pay special attention to the Recall and Precision.

The importance of those results needs also to be looked at, using the confusion matrix (Figure 6). The focus on rather we should be more careful to the False Positives or Negatives depends on the main objective of the hotel, and we will come back to it when talking about the implication for the business. However, it is important to keep both numbers at a low level to avoid generating important extra costs or loss of profit. Therefore, considering those 2 evaluation criteria, we can conclude CatBoost is our best predictor.

5. DEPLOYMENT AND APPLICATION TO THE BUSINESS

5.1. Application to the Business

From what we observed, our model allows the hotel to predict with a good accuracy of 82% if a booking is going to be canceled or not and a precision-recall trade-off of 69% and 70% respectively.

According to the business' objectives, those metrics will not have the same impact and we can still make the decision of using another model or parameters to fit better the strategy chosen. For example, if the hotel wants to focus on bookings classified as "going to be canceled", then prioritizing a minimum of False positives, and therefore a good Precision %, is important. It will allow it to reduce drastically its spending in cash or services (no overbooking or relocation) and improve its reputation (less unsatisfied clients). On the other hand, if the strategy is to focus on the 'not going to be canceled' outcomes, Recall % and a minimal number of False Negatives have to be prioritized. It will allow the hotel to better forecast its net demand and to avoid loss of profit (no vacant rooms). Both choices have good and bad consequences and should always be chosen with the team managing risks.

However, as we saw in the past cancelation management, taking drastic decisions doesn't seem to be profitable for the hotel. That's why having a balanced Precision-Recall % is what we have been focused on and seems to be the most adequate for the business. It will result in a combination of better predictions of the actual demands and avoid selling above its capacity, and so a general gain in profit and reputation.

5.2. Deployment of the Model

Now that we built a trustworthy predictive model, it is essential to ensure a good implementation of it in the business processes. A major point to take into consideration, is the hotel distribution that had become more and more externalise and diverse. Indeed, we saw in the Data Understanding section that bookings via external distribution channels such as Travel Agents/Operators and Corporate, were representing 94% of the actual reservations (check Figure 2). Therefore, having good coordination between all parties, meaning the hotel management system (called PMS), the Central Reservation System (called CRS), and the External Agents, is of crucial importance for a successful deployment.

Therefore, taking to account the speed and the complexity of the booking options, we came up with the conclusion that the model should be implemented on the hotel Central Reservation System. It will allow an **accurate forecast of the demands** and it will also improve the general **management and planning of the channel capacity.** The CRS, by having already a direct link to the PMS and to the External Agents (via the Channel Manager), instant communication on the hotel capacity will be improved (see Figure 7). Additionally, the model should run every day since some predictors can vary and can be impacted by changes very fast ("BookingChanges", "Adults", "LeadTime"). Forecasting the demands on daily-time basis, will help the hotel to better prevent and react to any changes (booking cancellations) and adjust its inventory more accurately.

Finally, several **preventive actions** can be advised for the hotel to reduce its cancellation rate to 20%. The most intuitive approach is to react to the bookings with a high prediction of being cancelled. The hotel could target those clients and come up with special discounts on services (such as access to the spa, gym, rent a car, activities in the surroundings...etc) or special offers (such a free breakfast, airport pick up...etc) that could reduce the risk of the booking cancellation. This approach cannot be used on the 11% booking from Corporate that is insensitive to those offers (the bookings

are not handled by individuals). Those preventive actions will have some costs for the hotel but, on the other hand, it will reduce the need to overbook and even maybe increase the satisfaction of the client. So, overall, it will have long-term benefits on the reputation and profit of the hotel.

6. CONCLUSION

As expected, when following the CRISP-DM, an agile methodology, some steps were not sequential and required going back and forth, which was the case of the modeling and evaluation steps. Throughout the project we encountered a few challenges, such as the data being imbalanced (after removing the duplicates) and data quality. In general, we can conclude that by using data science techniques we were able to present a good solution and achieve the main goals such as successfully identifying the features with the greatest predicting power and creating an accurate predictive model to classify the outcome of a booking which can translate in more sales. For future research, additional variables could be of use such as weather information. Also, it could be interesting to add data from other hotels in the chain to see if the model's performance increases.

7. REFRENCES

- Talluri, K. T., & Ryzin, G. V. (2004). *The theory and practice of Revenue Management*. Kluwer Academic Publishers.
- Lewinson, E. (2021, August 11). *Phik (φk) - get familiar with the latest correlation coefficient*. Medium. Retrieved March 20, 2022, from https://towardsdatascience.com/phik-k-get-familiar-with-the-latest-correlation-coefficient-9ba0032b37e7
- Brain John Aboz. *CatBoost vs XGBoost and lighgbm: When to choose CatBoost?* neptune.ai. Retrieved March 20, 2022, from https://neptune.ai/blog/when-to-choose-catboost-over-xgboost-or-lightgbm

- Mishra, R. (2021, March 6). *Stratified-k-fold in machine learning*. Medium. Retrieved March 21, 2022, from https://rahulmishra-40030.medium.com/stratified-k-fold-in-machine-learning-1b767a0b1eea

- Gilde, K. (2021, January 17). Faster hyperparameter tuning with Scikit-Learn's new HALVINGGRIDSEARCHCV. Medium. Retrieved March 21, 2022, from https://towardsdatascience.com/faster-hyperparameter-tuning-with-scikit-learn-71aa76d06f12

8. APPENDIX

Figure 1: Bar plots showing the increase of cancellations throughout the Years

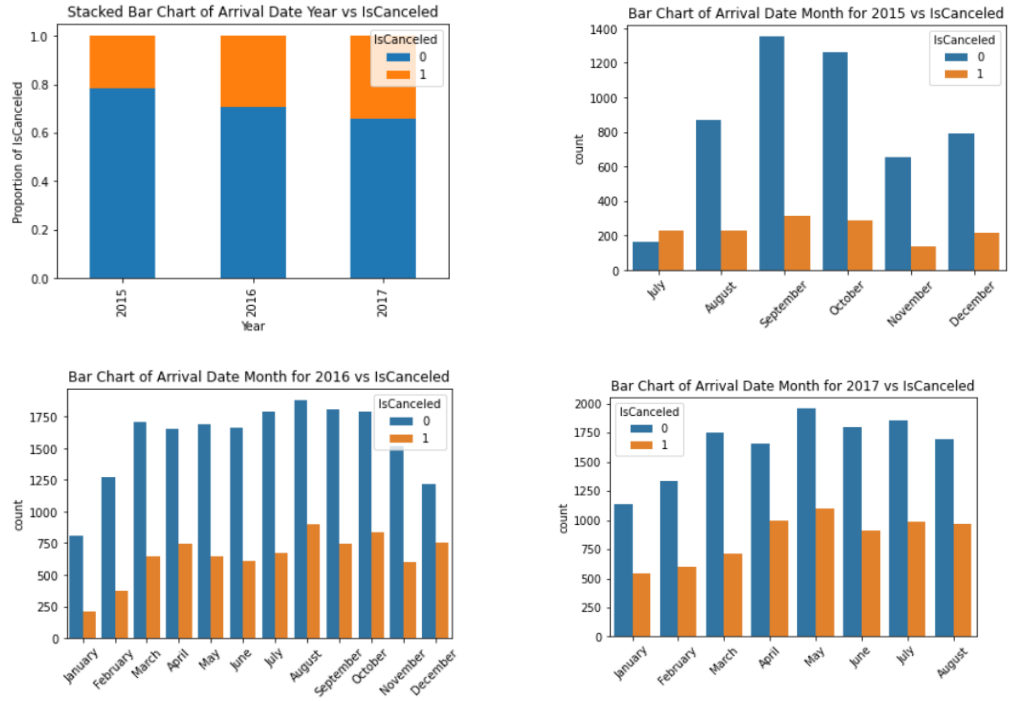


Figure 2: Distribution Channels %

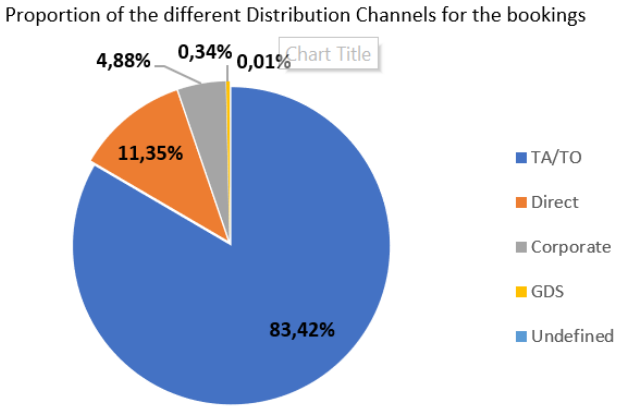


FIGURE 3: EXAMPLE OF BAR CHART WITH CROSS TAB FOR CATEGORICAL VARIABLES

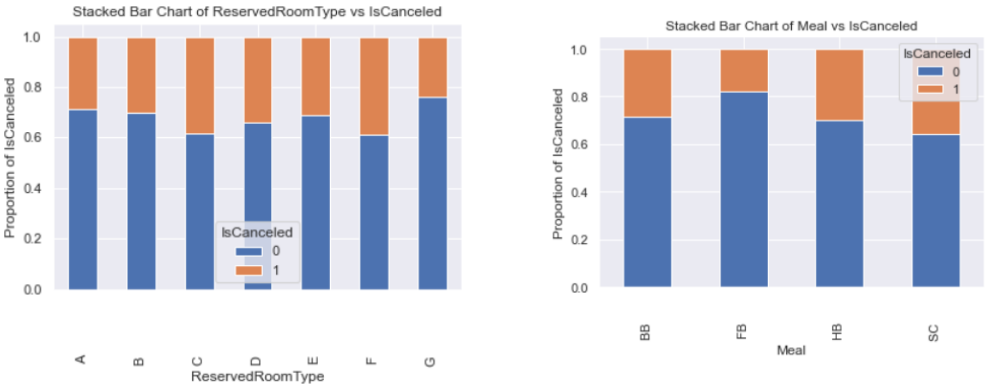


FIGURE 4.1: PHIK CORRELATION HEATMAP

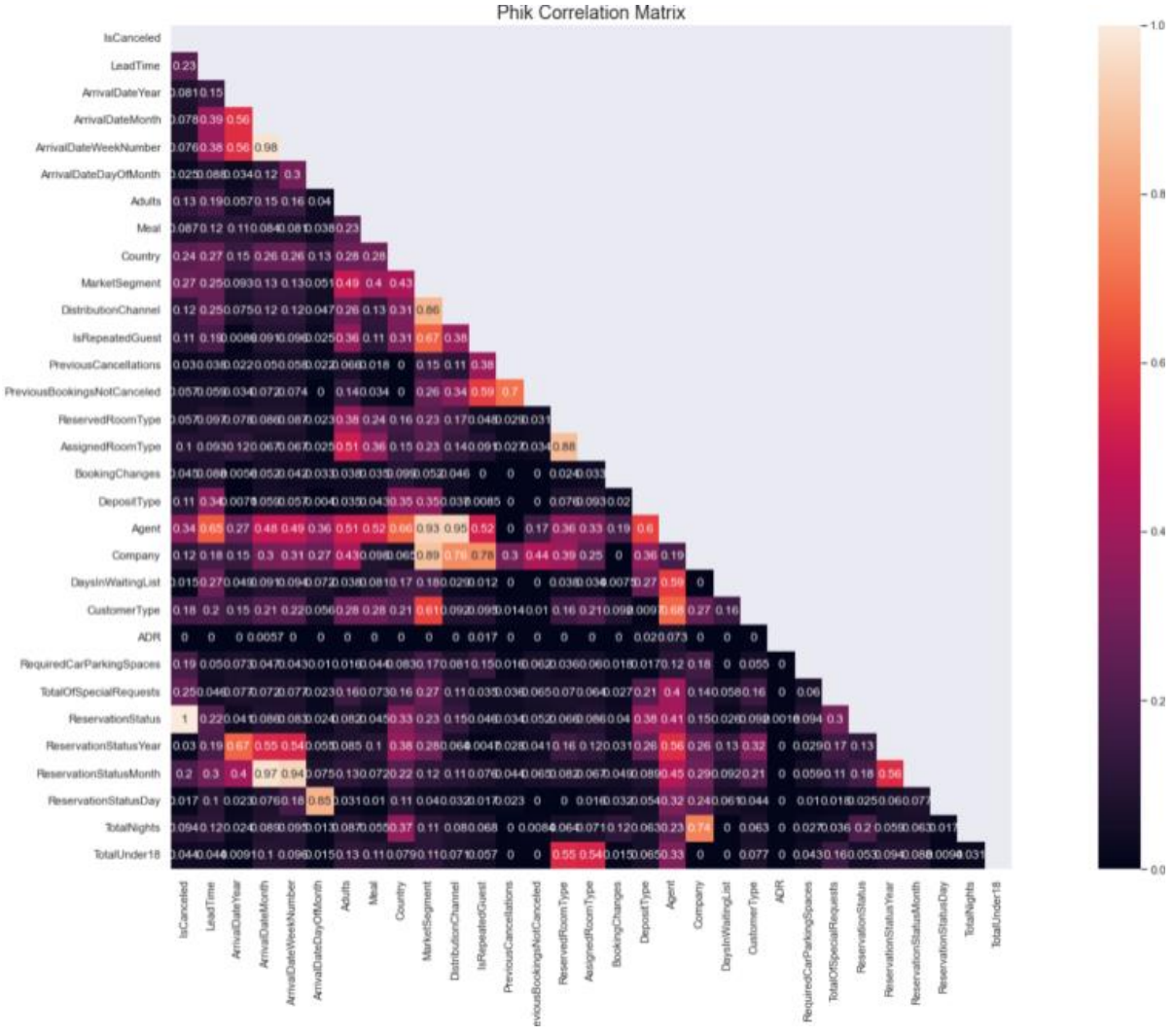


FIGURE 4.2: SPEARMAN CORRELATION HEATMAP

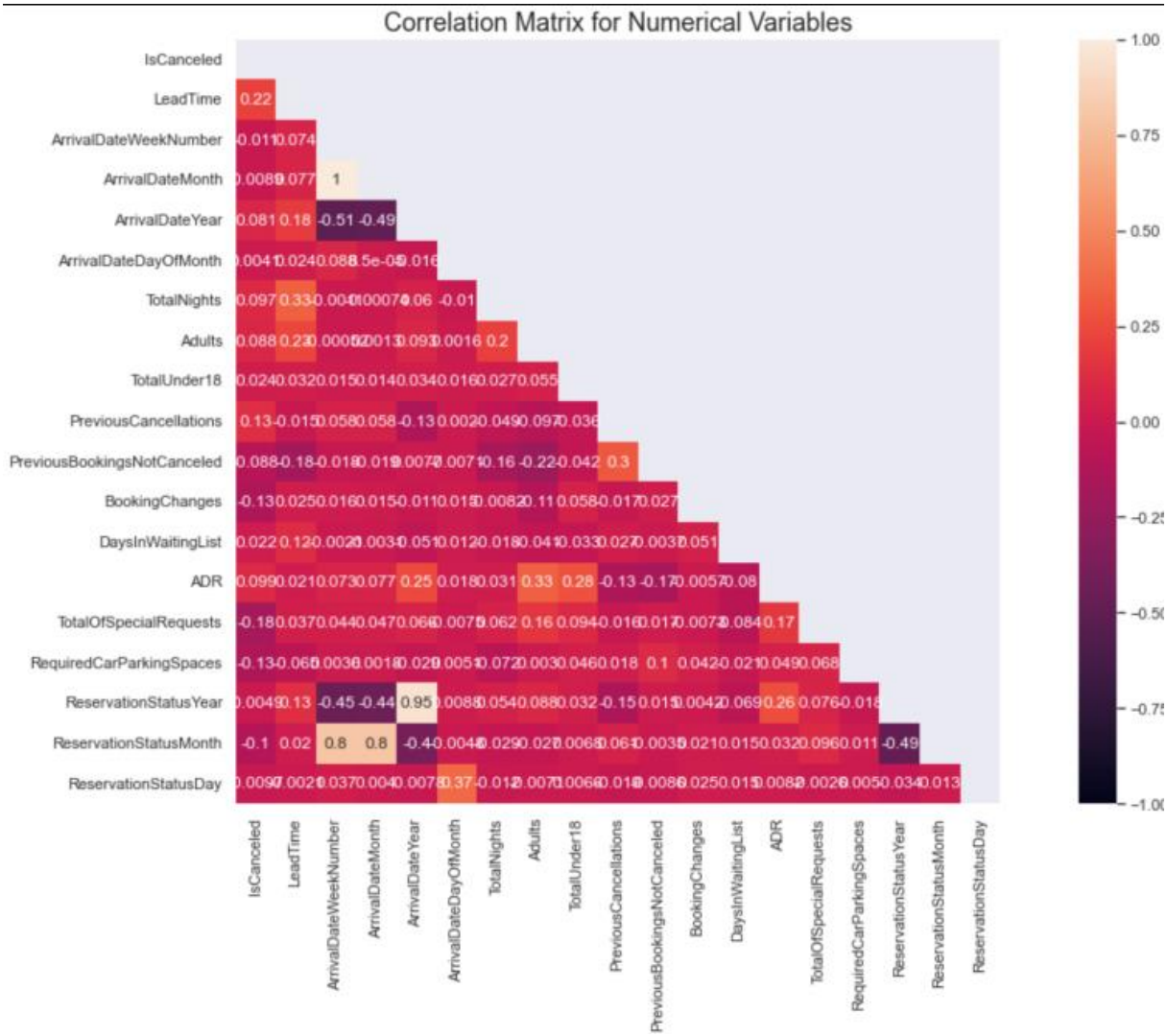
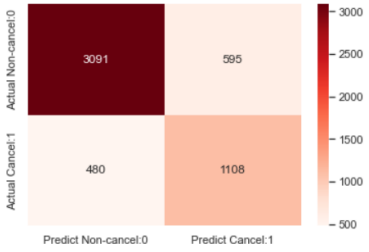


FIGURE 5: 10-FOLD CROSS-VALIDATION RESULTS ON VALIDATION

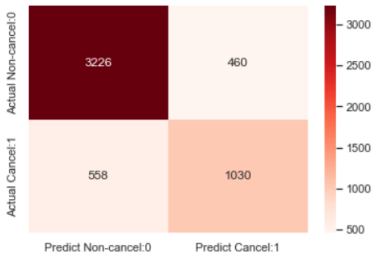
Algorithm	Measure	Accuracy	Precision	Recall	F1 Score
Catboost	Weighted Avg	0,81	0,69	0,7	0,87
XGBoost	Weighted Avg	0,80	0,65	0,70	0,67
Random Forest	Weighted Avg	0,81	0,69	0,65	0,67

FIGURE 6: COMPARISON OF THE CONFUSION MATRIX OF EACH PREDICTIVE MODEL

XGBOOST



RANDOM FOREST



CATBOOST

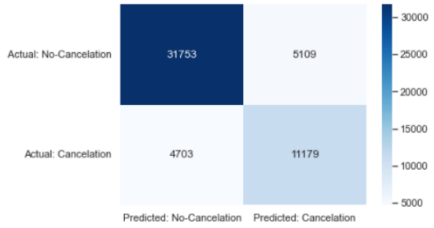


FIGURE 7: DEPLOYEMENT PLAN FOR THE HOTEL