Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 4: PREDICT CUSTOMER CHURN

Ana Carolina Ottavi, number: 20220541

Carolina Bezerra, number: 20220392

Duarte Girão, number: 20220670

João Pólvora, number: 20221037

Luca Loureiro, number: 20221750

Group Q: OptimaDataConsulting

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

May, 2023

INDEX

1. EXECUTIVE SUMMARY

In the hospitality industry, handling advance reservations is a crucial factor. The consumer has the contractual guarantee of hotel service at a predetermined price, as well as the option to cancel, when reservations are negotiated. When a guest cancels a reservation, however, the resulting empty rooms are crucial to the business because they incur expenses and miss out on business opportunities. From 2014 to 2018, the European cancellation rate increased from 33% to 40% based on the value of the reservation. Hotel chain C must address the high cancellation rate of hotel H2, so it will provide data for a business consultant to conduct an in-depth analysis of this issue.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. Background

There are many reasons and unforeseen circumstances that cause cancellations. One of them is the deal-seeking customers that made reservations but continue seeking for better options. This is caused by Online Travel Agencies (OTAs) that provide broad market exposure, conversely, force competition between hotels resulting in overbookings as practice to deal with the cancellations.

Overbooking creates issues as customer reallocation and negative experience damage the reputation of the hotel. Additionally, customers can decide not to rebook generating future revenue loss. Nonetheless, restrictive cancellation policies with non-refundable rates also bring problems, reducing booking since customers dislike severe policies and the hotels should offer more discounts to deal with the decrease in revenue.

2.2. Business Objectives

The main business objective is to create a predictive model that can forecast net demand for the hotel H2, considering the potential for cancellations. This will allow the hotel to manage their room availability more effectively, thereby reducing the losses associated with overbooking and increasing their overall revenue.

2.3. Business Success criteria

The objective is to accurately predict consumer preferences so as to reduce hotel H2 cancellations from 40% to 20%.

2.4. Situation assessment

The hotel H2 has an almost 42% cancellation rate. Therefore, in order to address this issue, A implemented an aggressive overbooking policy that resulted in increased costs. Regarding this, A adjusted to a less aggressive strategy that was ineffective, resulting in empty rooms even during peak demand periods. In order to analyze data from a city hotel H2 in Lisbon, the revenue manager director of hotel chain C engaged consultants to provide booking data for the period between July 1, 2015 and August 31, 2017.

The outcomes are provided in the form of a comprehensive report and a notebook detailing the modeling and success criteria. Finally, a presentation for the decision-makers must be available within 15 days.

2.5. Determine Data Mining goals

The goals of data mining in our project are: analyzing cancellations patterns, predictive modeling to forecast potential cancellations, optimize revenue by avoiding empty rooms, a less aggressive overbooking policy with cost reduction, and decision support by the data scientist team.

3. METHODOLOGY

3.1. Data understanding

The dataset comprises 70,330 rows and 31 columns. There are 16 metric features and 15 non metric features. The summary table provided initial insights into the variables and their unique traits.

Features	Takeaways
IsCanceled	There are fewer records of cancelled bookings (33102) than non-canceled (46228).
LeadTime	Bookings made on the same day (lead time of 0) have 3,109 instances. Not cancelled bookings have lower lead times.
ArrivalDateYear	Bookings span three years (2015-2017), with 2016 having the highest number (38,140) and 2015 the lowest (13,682).
ArrivalDateMonth	August has the highest bookings (8,983), while January has the lowest (3,736).
ArrivalDateDayOfMonth	The day with the highest number of bookings is the 17th, with 3,012 bookings. IsCanceled doesn’t seem to have an impact on the behavior of this variable.
StaysInWeekNights and StaysInWeekendNights	For ‘StaysInWeekendNights’, shorter stays are common. As for 'StaysInWeekNights', the most common weekday stay duration is 2 nights. There is a preference for weekday stays and shorter weekend getaways.
Adults, Children and Babies	Most bookings have two adults, few with 1 or 4 adults, and majority have no children or babies.
Meal	BB (Bed & Breakfast) is the most popular meal type, with many canceled and uncanceled appointments, while FB (Full Board) is the least. "IsCanceled" and "Meal" have no special relationship.
ReservedRoomType and AssignedRoomType	Room type 'A' is the most common and may be the default. Different assigned room types indicate upgrades.
DepositType	This variable suggests that the hotel primarily operates on a no-deposit or non-refundable deposit policy.

BookingChanges	Indicates that a considerable portion of guests modify their bookings, but the majority of bookings remain unchanged (69,062).
PreviousCancellations and PreviousBookingsNotCanceled	The majority of bookings (73,941) have no previous cancellations. As for PreviousBookingsNotCanceled, the majority of bookings (77,742) have no previous bookings that were not cancelled. This suggests a history of successful reservations.
DaysInWaitingList	The most common scenario is that bookings do not have any days on the waiting list (DaysInWaitingList = 0), with a count of 75,887 instances. Yet, this variable has no limit.
Agent and Company	The most frequent Agent ID in the dataset is “9" with a count of 31,955 bookings. The majority of bookings have the company recorded as 'NULL' (75,641). These 'NULL' values represent reservations that didn't come from any agents or companies; this is not a missing value but rather not applicable.
Country	With 30,960 bookings, Portugal (PRT) dominates the dataset. This dataset comprises clients from 166 countries. If a guest fails to check in, the country information may be incorrect.
MarketSegment and DistributionChannel	"Online TA" has the most bookings (38,748), followed by "Offline TA/TO" (16,747). "Complementary," "Aviation," and "Undefined" are rarely booked. "DistributionChannel" has 68,945 "TA/TO" bookings, showing a high reliance on travel agents. This channel cancels frequently. "Corporate," "Direct," and "GDS" book fewer than TA/TO. This column was missing "Undefined".
IsRepeatedGuest	The dataset contains a majority of bookings (77,298) from customers who are visiting for the first time, yet the hotel also has a loyal customer base.
TotalOfSpecialRequests	The majority of bookings have no special requests, but a significant number do. Most bookings without special requests end in cancellations.
ADR	The most frequent ADR value is 62.00 (3,593). Other common ADR values include 75.00, 90.00, 65.00, and 95.00. When the average daily rate is 62, there are more cancelled bookings than non-canceled ones. This variable contains a total of 5,405 unique values, indicating a diverse range of rates.
RequiredCarParkingSpaces	Most guests don't have a car or don't need parking facilities, and bookings with parking requirements have no cancellations.
ReservationStatus and ReservationStatusDate	Most ReservationStatus bookings (63,286) were completed. The most frequent ReservationStatusDate is "2015-10-21" (1,416) and there are 864 unique dates.
CustomerType	Majority of bookings are "Transient" (not part of any special arrangements), followed by "Transient-Party" (associated with a group). Groups make up a smaller percentage.

3.2. Data Exploration

The data exploration phase of CRISP-DM involves understanding the dataset and its characteristics, checking for missing values, outliers, and other issues that could affect the analysis, cleaning the data to remove or correct errors, visualizing and summarizing the data to gain insights, identifying relationships and correlations between variables, and selecting techniques for further analysis.

Cancelled bookings had a longer average lead time than non-canceled ones. This implies that longer lead times may increase cancellation risk. Subsequently, we compared the overall number of special requests for cancelled and non-canceled bookings. Guests with more special requests were less likely to cancel.

We also examined the lead time for cancelled and non-canceled bookings by month, finding that July, August, September, and October have longer lead periods. January, February, and March have lower lead times, implying shorter planning periods or greater cancellation flexibility. Cancelled bookings have longer lead times. We then examined market segment lead times for cancelled and uncanceled bookings. Aviation and Complementary have low lead times for both cancelled and uncanceled bookings. Corporate bookings, especially cancelled ones, have longer lead periods. Cancelled Direct and Online TA bookings have longer lead times. Cancelled bookings in the Groups segment have the highest lead times. Offline TA/TO bookings and cancelled bookings have substantial lead times. Cancelled appointments have slightly larger values for lead times in Online TA. Undefined bookings with missing values have lower lead times.

We examined the average lead time for cancelled and uncanceled bookings by client category. Contract clients book last-minute due to contractual obligations. Both cancelled and uncanceled group bookings had minimal lead times, implying closer-to-arrival bookings. Transient consumers prefer advanced preparation, with longer lead times for cancelled and uncanceled arrangements. Transient-party consumers have long lead times and low cancellation rates. Cross-checking lead time for various reservation statuses demonstrates that lead time is irrelevant for non-canceled reservations. Successful stays average 80.70 days.

In addition, we plotted the distribution of cancelled (***Figure 1***) and non-canceled bookings (***Figure 2***) by days on the List and Market Segment, and found that Groups cancel more after 50-100 Days in Waiting List and book more after 0-50 Days in Waiting List.

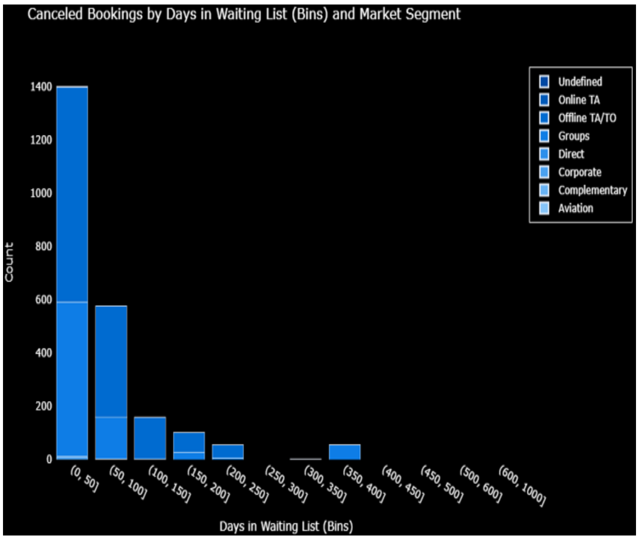


Figure 1 *"Canceled Day Waiting-Market Segment"*

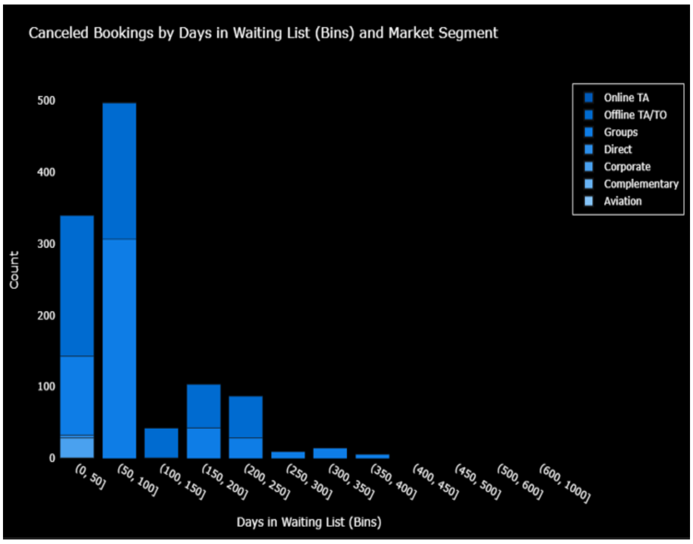


Figure 2 *"Not Canceled Day Waiting-Market"*

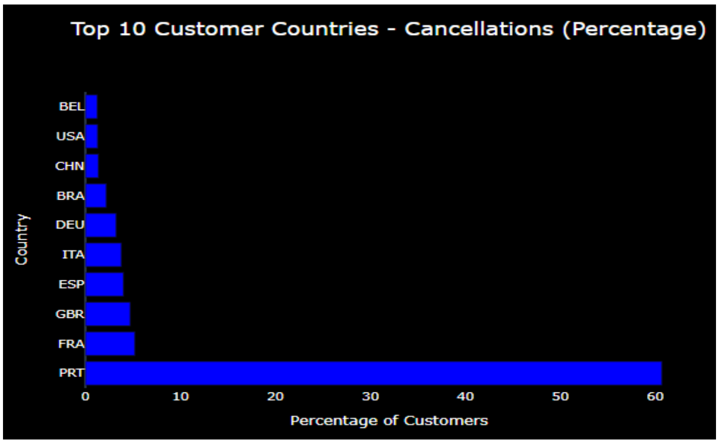


Figure 3 *"The top 10 countries with the majority of customers who cancel"*

In addition, we conducted an analysis to identify the countries with the highest cancellation rates in ***Figure 3***. Our findings showed that Portugal is the country with the highest cancellation rate, at 60%.

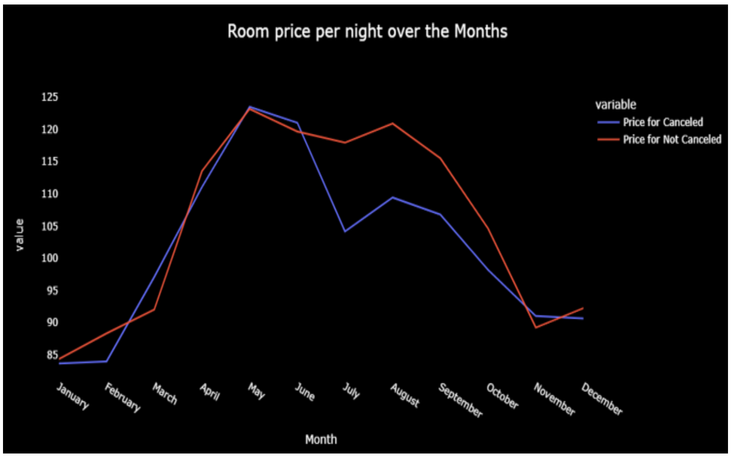


Figure 4 *"Room price per night over the Months".*

Figure 4 shows the annual average daily pricing for cancelled and non-canceled bookings. We found that the average costs for cancelled bookings range from 83.68 (January) to 123.47 (May) and for non-canceled bookings from 84.37 to 123.16. Both cancelled and non-canceled bookings have higher summer costs and lower winter prices.

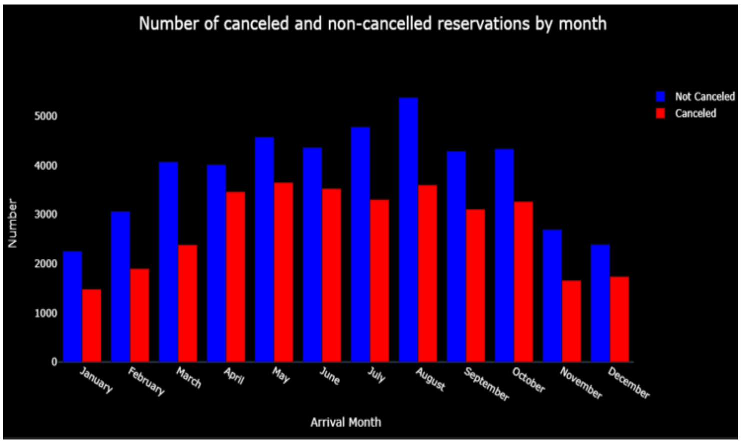


Figure 5 *"Months have the highest occupancy and cancellation rates"*

Additionally, we noticed that July saw the most cancellations, followed by August and May. The months with the lowest number of cancelled bookings are November and December. Overall, there is a higher count of not cancelled bookings compared to cancelled bookings for each month.

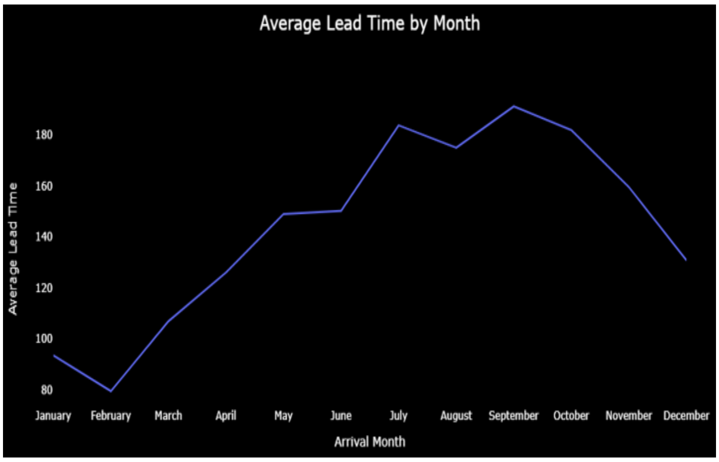


Figure 6 *"Average Lead Time by Month".*

February has the shortest cancellation lead time (Figure 6). LeadTime gradually increases from February to September, suggesting customers book their stays further in advance during this period. Bookings from October to December possibly indicate more last-minute bookings or shorter holiday planning periods.

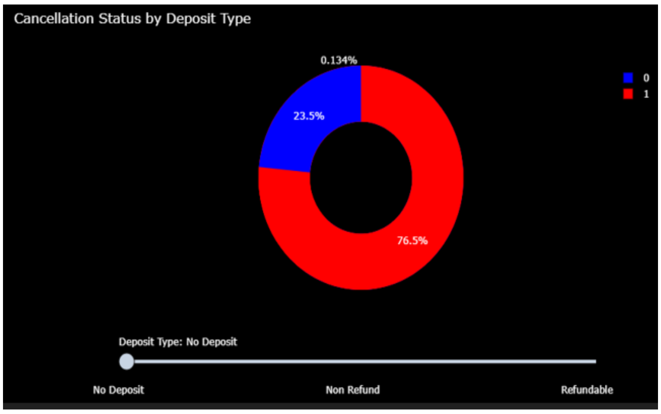


Figure 7 *"Cancelation Status by Deposit Type"*

Next, we looked at the cancellation status for different deposit types in Figure 7, and we could see that, for cancellations, only 40% are non-refundable deposit types. It means that this can have a major impact on the hotel's revenue, as it loses the reservation and, in 60% of cases, does not receive a deposit for it.

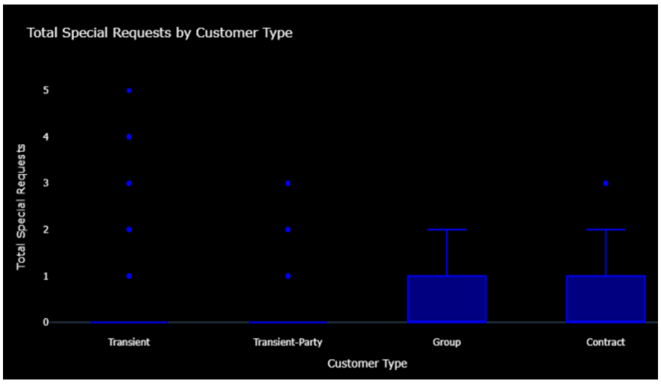


Figure 8 *"Total Special Requests by Customer Type"*

In order to better visualize the comparison of total number of special requests for different customer types for canceled bookings, we plotted box plot, ***Figure 8,*** where it's possible to see that Special Request 0 and 1 are specially for Contract and Group Customer Type and for the others is 0 for cancellations, and since we created box plot, we could observed some outliers.

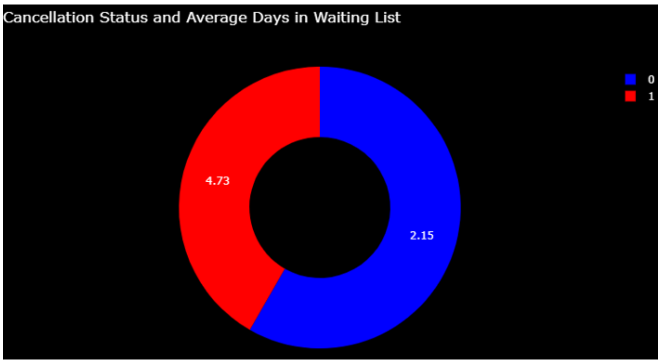


Figure 9 *"Cancellation Status and Average Days in Waiting List"*

Figure 9 assesses and compares the average days on the waiting list for cancelled and not cancelled bookings, and it was revealed that for both cancelled and not cancelled bookings, the concentration of days on the waiting list is between 0 and 250.

3.3. Data preparation

3.3.1 Cleaning the dataset

The goal of this step was to choose the necessary data to move on to the modeling stage. Following earlier insights acquired in the Data Understanding section, we began by defining metric and nonmetric features. Afterwards, cleaning the data removes unnecessary spaces from variables Agent, AssignedRoomType, Company, DepositType, Meal, and ReservedRoomType. We checked that it is impossible to have bookings with babies without adults and without adults, babies, and children at the same time; therefore, those observations were removed. Concerning missing values of 4 for children and 16 for country, we decided to replace the first one with 0 and remove all records in the variable country for a clearer analysis and to avoid creating new assumptions due to this reduced significance.

Regarding changing data types, we only change for integers the variables ADR and Children and convert to datetime ReservationStatusDate. The next step was removing outliers, in which we created a filter by outlier manual remove approach, checking histograms and boxplots (***Figure 10***), as well as descriptive statistics. We filtered ADR with greater or equal values than 20 and less or equal than 520; babies less or equal than 8. BookingChanges less/equal than 10, DaysInWaitingList less/equal than 250, LeadTime less/equal than 550, PreviousCancellations less/equal than 10, RequeriedCarParkingSpaces less/equal than 1, and TotalOfSpecialRequests less/equal than 3. The data after removing outliers is 97.03%.

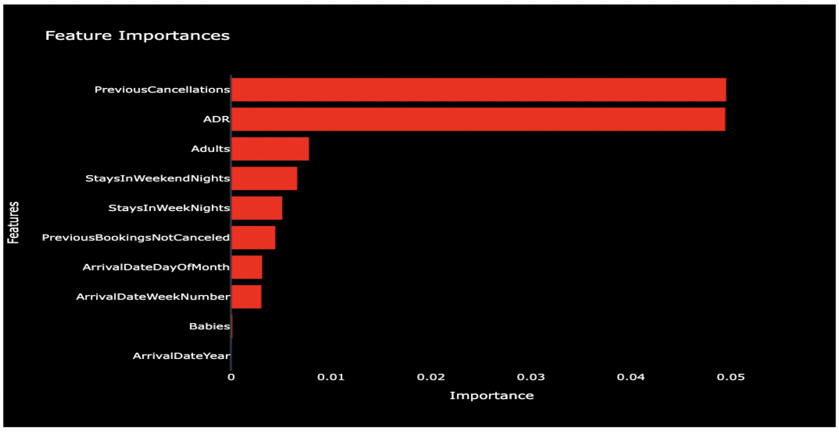
Relating cleaning non-metric features, we did an individual analysis with the following conclusions: removed observations for the variable Meal equal to 'FB', deleted one observation in DepositType equal to "Refundable', and found 2 entries in Country with 2 letters when the standard is 3, in this case China, and we replaced CN for CHN. We also create a new variable called Different Rooms, with 0 if the room didn't change and 1 if it did. In this case, we noticed that only 6643 cases had changed.

3.3.2 Feature Engineering

Firstly, we created a binary variable called ADR_NEW that represents bins of values for the price of the room (0–19, 20–50, 51–100, 101–150, 151-200). The second new variable is StaystotalNights_NEW, which is defined by the sum of StaysInWeekendNights and StaysInWeekNights binning into a total of nights (1, 2-3, 4-7, 8-30, 30+). For the new feature MarketSegment_NEW, firstly, we defined a function to check the cancellation ratio that groups the dataframe by the specified feature and counts the number of cancelled and non-cancelled instances. The cancellation is defined by dividing the count of cancelled instances by the sum of cancelled and non-cancelled instances. We used this function associated with the feature MarketSegment to categorise if the cancellation ratio is greater or less than 0.25; this removes unnecessary columns related to the cancellation ratio and the original MarketSegment column. We binned LeadTime into 4 bins (0-7, 8-30, 31-120, 121-365), resulting in the new variable LeadTime_NEW. The next is ArriveDateMonthSeason, grouped in categories as Peak Season, Winter Season, and Spring Season, and the last one is DaysInWaitingList_NEW, which was grouped in 3 bins (0, 1–30, 31–90). In conclusion, we dropped the features: MarketSegment, LeadTime, ArrivalDateMonth and DaysInWaitingList.

Continuing with the feature engineering, we defined some dummy variables, with values 0 and 1, as BookingChanges_NEW, RequiredCarParkingSpaces_NEW, TotalOfSpecialRequests_NEW, Company_NEW, Existing_Children, and Transient_Contract. Consequently, the features BookingChanfes, RequiredCarParkingSpaces, TotalOfSpecialRequests, Company, Children, and CustomerType were removed.

Figure 11 **-** *"Feature Importance Metric Features".*

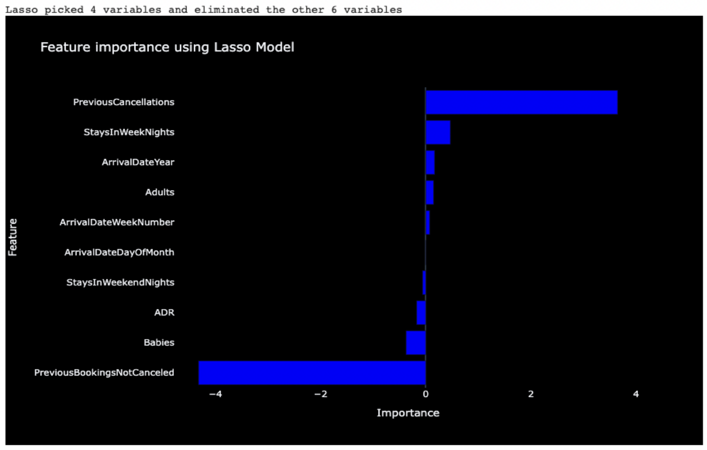


Metric feature Filter methods have no independent variable strongly connected with the objective. ArrivalDateYear is associated with ArrivalDateWeekNumber (-0.5) and PreviousCancellations (-0.3). ***Figure 11*** shows mutual information's characteristic relevance. TestIndependence, a Filter Methods function for non-metric features, uses chi-square to determine if an independent variable is a significant predictor. The function suggests keeping all non-metric attributes.

The next method, Wrapper Method, for the RFE we used Logistic Regression to select the features. Since we don't know a priori the number of features to select, we decided to create a loop to check it. The result indicates the most important are Babies, PreviousCancellations and PreviousBookingNotCancelled. We also apply RFE with Random Forest resulting in keep the features ArrivalDateYear, ArrivalDateWeekNumber, ArrivalDateDayOfMonth, Adults, PreviousCancellations, PreviousBookingNotCanceled, being Babies the unique variable to be discarded.

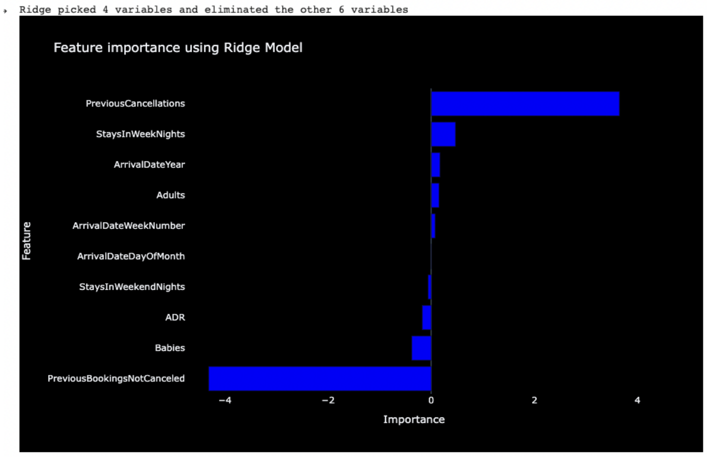
Lastly, the Embedded Method using LASSO in the first approach. This indicates in ***Figure 12*** that 4 metric features to be kept: StayInWeekNights, PreviousBookingsNotCanceled, PreviousCancellations, Babies and PreviousBookingCanceled. Further we used the Ridge Model as second approach, ***Figure 13,*** and it maintained the same variables kept. Lastly, the GradientBoosting indicates the feature importance as we can explore in ***Figure 14.***

Figure 12 **-** *"Feature Selection with LASSO"*



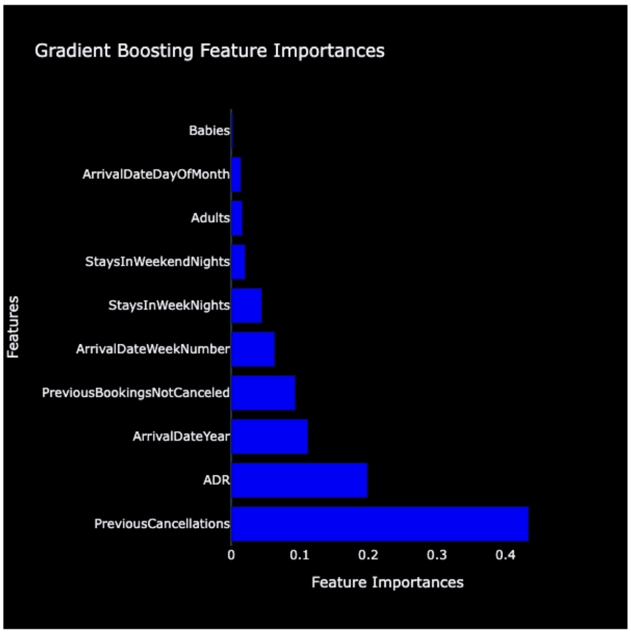
```
Lasso picked 4 variables and eliminated the other 6 variables
```

Figure 13 **-** *"Feature Selection with Ridge"*



```
Ridge picked 4 variables and eliminated the other 6 variables
```

Figure 14 - *"Feature Selection with Gradient Boosting."*



We kept Adults, PreviousCancellations, ADR, StaysInWeekNights, StaysInWeekendNights and PreviousBookinsNotCanceled for metric features in the appendix, Figure 15. Figure 16 in the appendix shows Company_NEW as the unique non-metric attribute. Afterwards, we encoded categorical features using fit_transform and scikit-learn OHE.

4. MODELING

4.1. Success Criteria

The two main metrics that we took into account to define our success criteria are precision and recall. First, precision tells us the accuracy of the positive predictions made by our model. In this sense, with this metric, we aim to evaluate, among the predicted cancellations, how many of them are actually real cancellations. For example, a precision of 80% tells us that among all our predicted cancellations, only 80% of them actually were cancellations. In this business context, high precision would mean that when the model forecasts a booking as canceled, it is very likely to be correct. A high level of precision is needed if the hotel wants to avoid unnecessary actions (such as contacting clients to avoid cancellations or making unnecessary offers) and decrease costs associated with false positive predictions.

Second, recall, which is usually used in binary classification tasks, measures the ability of a model to correctly identify positive instances. In other words, it tells us, among all the existing cancellations, how many of them were correctly predicted by our model. For example, a recall of 80% would mean that among all the real cancellations, we were able to predict 80% of them. In this business context, a high recall would suggest that the model is effective at detecting the majority of the cancellations. A high recall is important for the hotel to minimize missed cancellations and take appropriate actions to manage the remaining free rooms and optimize profit. By having a high recall, the hotel can proactively address canceled bookings, reassign resources, and potentially prevent revenue loss.

Having in mind our main goal, the strategy adopted for this project, to decrease the hotel cancellation ratio from 40% to 20%, we defined as our targets at least 70% for recall and at least 80% for precision. Our rationale tells us that we would need to be able to correctly identify and address at least 50% of the existing total cancellations. Once there are 40% of cancellations, we would need to obtain a total of 50% of total positives (according to the confusion matrix) in order to make sure that it would be possible to achieve our main goal.

The final goal could be achieved through a 50% recall and 100% precision, which is obviously unrealistic. As long as we increase our desired recall level, we decrease the necessity for higher precision. With 70% for recall and 80% for precision, we have a comfortable margin to be sure that we achieve our final goal. Taking into consideration our goal, although we could have opted to analyse exclusively the F1 score to assess our results, we preferred to make sure that our model results achieved a certain threshold regarding the recall and precision metrics in order to avoid big differences between them and to obtain more insights useful for recommendation suggestions. For example, with precision detail, we will be able to suggest a concrete strategy to avoid unnecessary costs over FP: in case the hotel uses some special commodities to try to avoid client cancellation, we must be careful regarding FP since we would be losing substantial amounts of money.

4.2. Generate Test Design

First, we applied a data partition where, in the variable 'X," we dropped the target, IsCanceled, and saved it in the variable "y." Then, the train_test_split function from Scikit-Learn breaks down X_train, X_val, y_train, and y_val into the validation set containing 30% of the data. For some models, the data must be scaled; therefore, we used the StandardScaler class from scikit-learn. We fit the method to the training data where the mean and standard deviation of each feature in X_train were calculated, and after the transform method, we scaled the transformation in X_train and X_val.

4.3. Build Model

4.3.1 Logistic Regression

Our first model was Logistic Regression Algorithm, which is a statistical model that uses a logistic function to model a binary dependent variable. In our case, the dependent variable is whether a booking is canceled. The model achieved an accuracy of 80.72%, a precision of 84.26%, a recall of 66.63%, and an F1 score of 74.41%, and by checking the confusion matrix for the model it's verified that the model correctly predicted 12155 non-cancellations and 6471 cancellations, while it incorrectly predicted 1209 cancellations and 3241 non-cancellations.

4.3.2 Random Forest

After we performed a Random Forest model, which is a learning method that constructs multiple decision trees during, and return class that is the mode of the classes of the individual trees. We obtained an accuracy of 84.38%, a precision of 83.60%, a recall of 78.22%, and an F1 score of 80.82%. By checking the confusion matrix for the model showed that the model correctly predicted 11874 non-cancellations and 7597 cancellations, while it incorrectly predicted 1490 cancellations and 2115 non-cancellations.

4.3.3 MLP

Then, we tried a Multilayer Perceptron (MLP) model. This is a type of feedforward artificial neural network that consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer. In this one, we reached an accuracy of 84.38%, a precision of 83.60%, a recall of 78.22%, and an F1 score of 80.82%. Looking at the confusion matrix for the model showed that the model correctly predicted 11874 non-cancellations and 7597 cancellations, while it incorrectly predicted 1490 cancellations and 2115 non-cancellations.

4.3.4 Decision Trees

Our fourth model was a Decision Tree model. This is a type of model used for both classification and regression and works by creating a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Here we achieved an accuracy of 76.51%, a precision of 99.56%, a recall of 44.39%, and an F1 score of 61.40% and the confusion matrix for the model showed that the model correctly predicted 13345 non-cancellations and 4311 cancellations, while it incorrectly predicted 19 cancellations and 5401 non-cancellations.

4.3.5 Gradient Boosting

Next, a Gradient Boosting model was performed, which is a derivation of decision tree models, being a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models. The Gradient Boosting model acquired an accuracy of 81.22%, a precision of 86.70%, a recall of 65.42%, and an F1 score of 74.57%. The confusion matrix for the model shows that the model correctly predicted 12389 non-cancellations and 6354 cancellations, while it incorrectly predicted 975 cancellations and 3358 non-cancellations.

4.3.6 AdaBoost

Additionally, we executed AdaBoost model (Adaptive Boosting), that is used as a classifier for machine learning solutions and also uses decision tree learning gathering information at each stage of the AdaBoost algorithm about the relative 'hardness' of each training sample is fed into the tree growing algorithm such that later trees tend to focus on harder to classify examples. We had an accuracy of 80.83%, a precision of 85.02%, a recall of 66.10%, and an F1 score of 74.38%. The confusion matrix for the model revealed that the model correctly predicted 12233 non-cancellations and 6420 cancellations, while it incorrectly predicted 1131 cancellations and 3292 non-cancellations.

4.3.7 Bagging

As well, we implemented our seventh model which was a Bagging model (Bootstrap Aggregating), a simple and very powerful ensemble method, which is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model. Analyzing the performance it indicated an accuracy of 83.82%, a precision of 83.18%, a recall of 77.16%, and an F1 score of 80.06%. The confusion matrix for the model displayed that the model correctly predicted 11849 non-cancellations and 7494 cancellations, while it incorrectly predicted 1515 cancellations and 2218 non-cancellations.

4.3.8 XGBoost

Our eighth model was an XGBoost model (Extreme Gradient Boosting), an implementation of gradient boosting with several additional features focused on performance and speed. Here it showed an accuracy of 83.71%, a precision of 85.76%, a recall of 73.50%, and an F1 score of 79.16%, and by checking the confusion matrix for the model it confirms that the model correctly predicted 12179 non-cancellations and 7138 cancellations, while it incorrectly predicted 1185 cancellations and 2574 non-cancellations.

4.3.9 SVM

The next algorithm was a Support Vector Machine (SVM) model, which is a type of machine learning algorithm that is used for classification and regression analysis. It achieved an accuracy of 81.71%, a precision of 87.45%, a recall of 66.02%, and an F1 score of 75.24% and the confusion matrix for the model showed that the model correctly predicted 12444 non-cancellations and 6412 cancellations, while it incorrectly predicted 920 cancellations and 3300 non-cancellations.

4.3.10 CatBoost

Additionally, we performed a CatBoost algorithm that uses gradient boosting on decision trees. It shows good results for handling categorical variables and for handling data leakage and overfitting. As results it returned an accuracy of 83.87%, a precision of 86.19%, a recall of 73.43%, and an F1 score of 79.30%, and the confusion matrix for the model displayed that the model correctly predicted 12221 non-cancellations and 7132 cancellations, while it incorrectly predicted 1143 cancellations and 2580 non-cancellations.

4.3.11 ExtraTreesClassifier

Then, we executed the ExtraTreesClassifier, which achieved an accuracy of 83.41%, a precision of 82.36%, a recall of 77.10%, and an F1 score of 79.30%. Having a look at the confusion matrix for the model showed that the model correctly predicted 11760 non-cancellations and 7488 cancellations, while it incorrectly predicted 1604 cancellations and 2224 non-cancellations.

4.3.12 PassiveAggressiveClassifier

After we tried this very interesting algorithm, PassiveAggressiveClassifier which reached an accuracy of 76.34%, a precision of 74.64%, a recall of 66.32%, and an F1 score of 79.30%. The confusion matrix for the model demonstrated that the model correctly predicted 11176 non-cancellations and 6441 cancellations, while it incorrectly predicted 2188 cancellations and 3271 non-cancellations.

4.3.13 Naive Bayes

We look at the Naive Bayes that acquired an accuracy of 61.45%, a precision of 52.48%, a recall of 88.80%, and an F1 score of 79.30%. Its confusion matrix indicated that the model correctly predicted 5556 non-cancellations and 8624 cancellations, while it incorrectly predicted 7808 cancellations and 1088 non-cancellations.

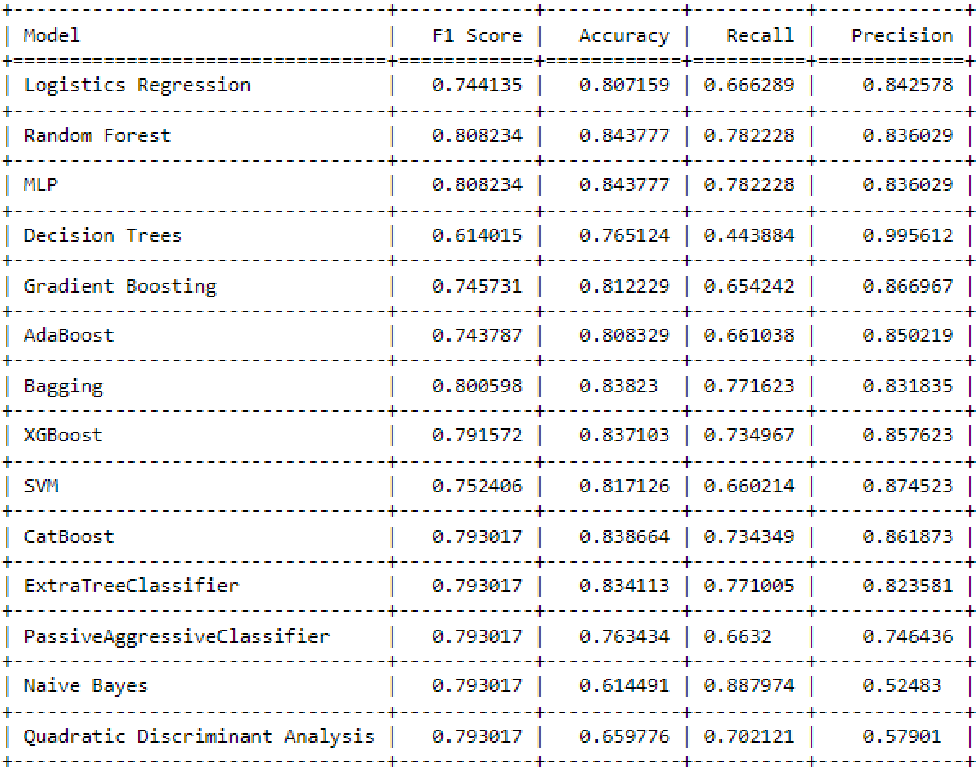
4.3.14 Quadratic Discriminant Analysis

The Quadratic Discriminant Analysis model performed an accuracy of 65.98%, a precision of 57.90%, a recall of 70.21%, and an F1 score of 79.30% and the confusion matrix for the model showed that the model correctly predicted 8406 non-cancellations and 6819 cancellations, while it incorrectly predicted 4958 cancellations and 2893 non-cancellations.

4.4 Assess model / Models Comptonization

As we conclude the model development and evaluation phase, it’s essential to compare the performance of each model based on key metrics such as F1 score, accuracy, recall, and precision. This comparative analysis, ***Figure 17,*** was instrumental in identifying the models that performed best. To make this information more accessible and business-friendly, we have consolidated these metrics into a comprehensive table and created bar plots for each metric.

Figure 17 *- "Comparative models performance."*



```
+---------------------------------+------------+------------+----------+-------------+
| Model                           |   F1 Score |   Accuracy |   Recall |   Precision |
+=================================+============+============+==========+=============+
| Logistics Regression            |   0.744135 |   0.807159 | 0.666289 |    0.842578 |
+---------------------------------+------------+------------+----------+-------------+
| Random Forest                   |   0.808234 |   0.843777 | 0.782228 |    0.836029 |
+---------------------------------+------------+------------+----------+-------------+
| MLP                             |   0.808234 |   0.843777 | 0.782228 |    0.836029 |
+---------------------------------+------------+------------+----------+-------------+
| Decision Trees                  |   0.614015 |   0.765124 | 0.443884 |    0.995612 |
+---------------------------------+------------+------------+----------+-------------+
| Gradient Boosting               |   0.745731 |   0.812229 | 0.654242 |    0.866967 |
+---------------------------------+------------+------------+----------+-------------+
| AdaBoost                        |   0.743787 |   0.808329 | 0.661038 |    0.850219 |
+---------------------------------+------------+------------+----------+-------------+
| Bagging                         |   0.800598 |   0.83823  | 0.771623 |    0.831835 |
+---------------------------------+------------+------------+----------+-------------+
| XGBoost                         |   0.791572 |   0.837103 | 0.734967 |    0.857623 |
+---------------------------------+------------+------------+----------+-------------+
| SVM                             |   0.752406 |   0.817126 | 0.660214 |    0.874523 |
+---------------------------------+------------+------------+----------+-------------+
| CatBoost                        |   0.793017 |   0.838664 | 0.734349 |    0.861873 |
+---------------------------------+------------+------------+----------+-------------+
| ExtraTreeClassifier             |   0.793017 |   0.834113 | 0.771005 |    0.823581 |
+---------------------------------+------------+------------+----------+-------------+
| PassiveAggressiveClassifier     |   0.793017 |   0.763434 | 0.6632   |    0.746436 |
+---------------------------------+------------+------------+----------+-------------+
| Naive Bayes                     |   0.793017 |   0.614491 | 0.887974 |    0.52483  |
+---------------------------------+------------+------------+----------+-------------+
| Quadratic Discriminant Analysis |   0.793017 |   0.659776 | 0.702121 |    0.57901  |
+---------------------------------+------------+------------+----------+-------------+
```

Figure 18 *- "F1 Score of Models"*

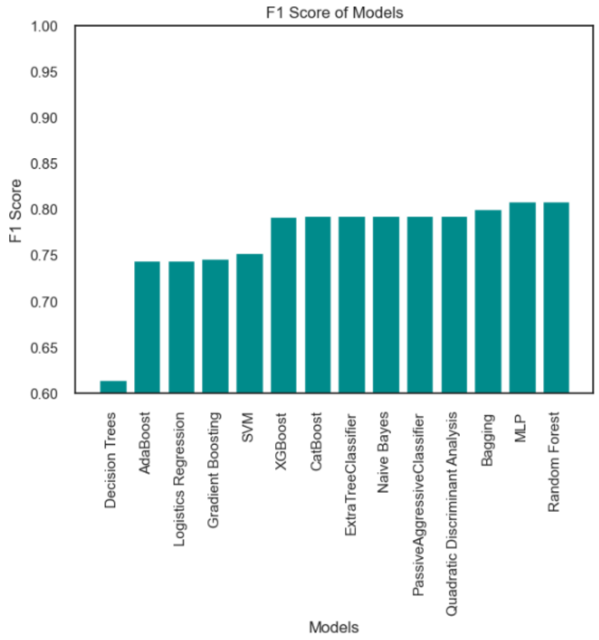


Figure 19 *- "Recall of Models"*

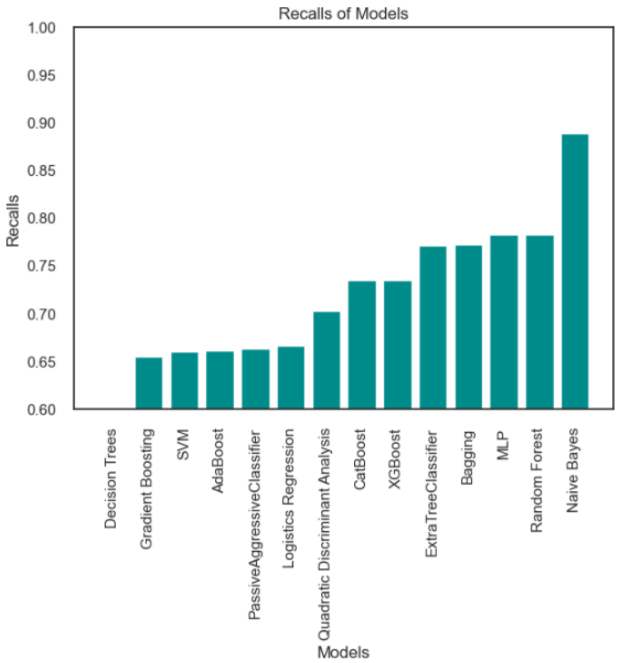


Figure 20 **-** *“Accuracies of Models”*

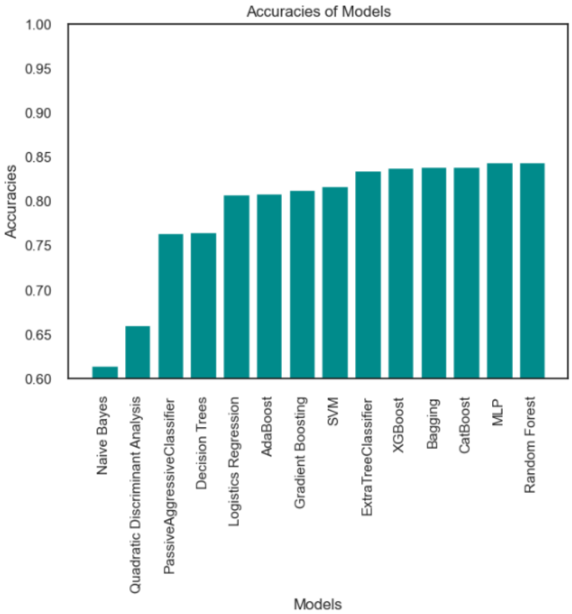
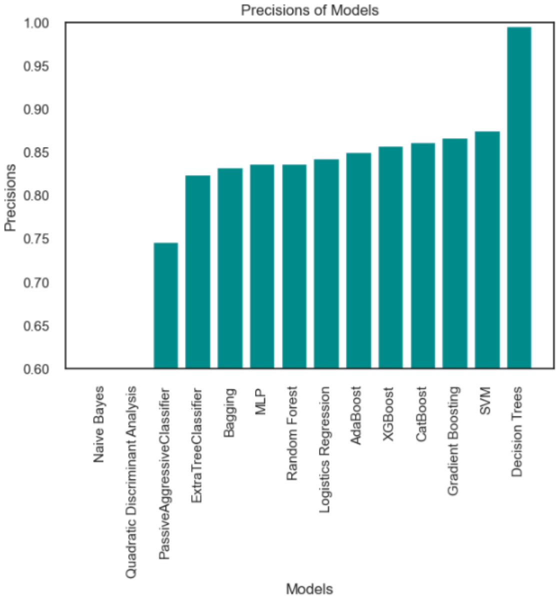
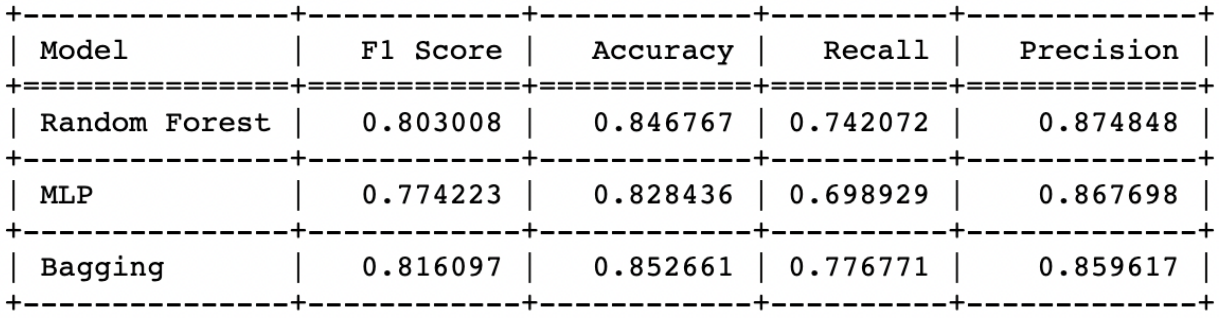


Figure 21 **-** *“Precision of Models”*



Analyzing our results, ***Figures 18, 19, 20, and 21,*** using the default parameters, we observe that several models achieved a recall higher than 70% and a precision higher than 80%, as we intended. At this point, we decided to select the three that have higher recall, above 75%, which are also the ones with the higher F1 Score (computed based both on recall and precision): **MLP, Random Forest, and Bagging**. So, we will apply a Hyperparameter Search over these three models and recheck which performed better (***Figure 22).***

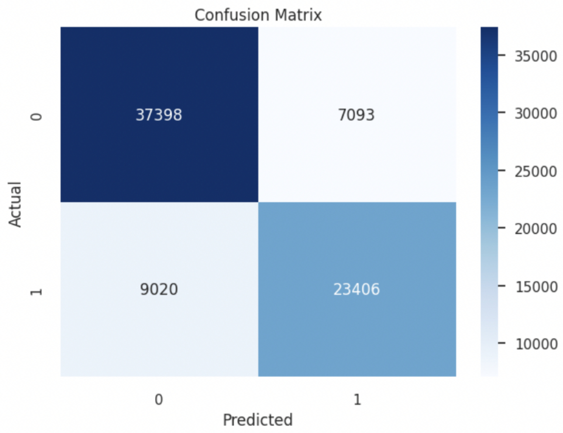
Figure 22 **-** *“Comparative 3 best models performance”*



```
+---------------+------------+------------+----------+-------------+
| Model         |   F1 Score |   Accuracy |   Recall |   Precision |
+===============+============+============+==========+=============+
| Random Forest |   0.803008 |   0.846767 | 0.742072 |    0.874848 |
+---------------+------------+------------+----------+-------------+
| MLP           |   0.774223 |   0.828436 | 0.698929 |    0.867698 |
+---------------+------------+------------+----------+-------------+
| Bagging       |   0.816097 |   0.852661 | 0.776771 |    0.859617 |
+---------------+------------+------------+----------+-------------+
```

5. DEPLOYMENT AND MAINTENANCE PLANS

Figure 23 **-** *“Confusion Matrix of Evaluate Results”*



After data preparation, data preprocessing and the entire modeling, including the comparison between all the models we have we can positively say that we obtained our machine learning success criteria:

- Required at least 70% on recall and at least 80% on precision;
- Achieved on our best(s) model(s) >64% on recall and >89% on precision;
- Although the recall was not achieved totally, we have confidence in our results, since a 70% value already included a comfortable margin on top of the minimum value required. Additionally, our precision lets us be very confident about our results, since it is far above our initial requirements.

In summary, we can positively conclude that we achieved the goal of reducing the cancellations ratio to 20%: Since, as clearly demonstrated above, we were able to obtain a final cancellation ratio of 23,26%, which represents a decrease of around almost 20% regarding the existing cancellation ratio at the date;

In summary, taking into account that our final dataset (after all the preprocessing and feature engineering) has a total of 76.917 Observations observing our confusion matrix, we were able to observe: total number of records of 7.093 False Positives, total number of records of 23.406 True Positives; total number of records of 9.020 False Negatives; total number of records of 37.398 True Negatives;

We got to some insights obtained from our model development:

- Regarding the technical point of view, we got to a point where we had to make a decision, due to the trade off between the overfitting and the cancellation ratio. To make sure that our final best model did not display overfitting, we had to accept a slightly higher cancellation ratio. At the end, we obtained a final model practically without overfitting and a cancellation ratio around 20%, as desired. <br>
- Regarding business point of view As we can see above, we obtained a final overbooking_loss value of 7.526.872 (7.5M Eur), which represents the total value (€), that the hotel would;

Finally, talking in consideration the final results, our final suggestions on how the model strategy should be deployed is the following:

- The False positives represent the clients which our model predicts with a big likelihood of cancellation and in reality, turn to not cancel, and checkIN. These cases represent 7.093, among all the 76.917 existing customers. In this cases, we believe that it will not bring significant additional costs to the company. These clients may cause some uncomfortable situations, due to overbooking situations and generate some conflicts between clients, who may reserve the same room, for example. However, as he have said, these situations are not expected to be very common, and are not very expensive to deal with. We believe that these situations could be dealt:
- Through e-mail or by telephone directly with the client, which means that the only significant setback would be the time consumed to take care of those situations. Monetary speaking, the cost would not be significant, in our opinion, even almost null.
- The false negatives represent the clients which our model does not predict as cancellations, but actual turn in cancellations. These cases represent the most complicated cases, where the hotels tend to lose more money, according to the opportunity. Additionally, and since this is the most sensitive point, there are some methodologies that the hotel could adopt in

order to address this problem of " how to avoid cancellations" and keep their profit margins attractive enough:

- First, develop more robust canceling policies, which encourage guests to provide early notice of cancellations. This includes providing incentives to the clients for early cancellations or even higher penalties for late cancellations. These policies must also be totally clear and help to reduce occurence of cancellations, without an aggressive posture near the client.
- Second, implement a more regular communication with the client, developing more dynamic communication channels with clients after their booking in order to remind them of their reservations and provide assistance if required. There can also be sent personalized emails or phone text messages a few days/weeks before their planned arrival to confirm their stay, clarify any doubt, and provide information about the cancellation process if needed.
- Thirdly, direct personalized upselling products/services opportunities or incentives to guests which we are predicting a higher risk of cancellation. This could include providing special offers, complimentary upgrades with existing partnerships that the hotel may have (with restaurants, cinemas, museums, etc), or exclusive services to incentive guests to keep their reservations enhancing their experience.
- Fourth, directly linked with the previous suggestion, we see a need to strengthen existing partnerships and create new ones, with a wide diversity of restaurants and cultural places. Additionally, collaborations with travel agents and online travel agencies (OTAs) to improve communication and coordination regarding cancellations.
- Fifth, carefully identify and monitor cancellations patterns through machine learning models, which can have two goals: first, target more accurate marketing campaigns, leveraging guest data related to cancellations. This allows to reach out to specific guest segments that are more prone to cancellations, proving tailored offers and addressing pain points; second, analyze trends and historical client data in order to identify areas of improvement.
- Lastly, there can be developed a more regular guest feedback collection, to identify common reasons for cancellations. This can allow the hotel to address those concerns more accurately in the future.
- The true positives represent the most significant customers that we most address, since are the actual cancellations that we were able to correctly predict. They represent a very significant percentage of our entire dataset, and taking this into consideration, we decided to look closely at the Random Forest (our best model) feature of importance map. This map, plotted above, told us the importance of each feature on our final target (Cancellation). Our most important features are:
- DepositType_No Deposit and DepositType_Non Refund, since they provide a level of assurance for the hotel that guests will honor their bookings, are quite relevant in predicting the cancellations. The deposit policy must be carefully reviewed and the hotel may increase the standards required for the deposit.
- TotalOfSpecialRequests_NEW_O and TotalOfSpecialRequests_NEW_1, although we do not see an obvious practical measure or insight that could be withdrawn from this.
- PreviousCancellations, since for obvious reasons, clients that have a higher number of previous cancellations tend to cancel more frequently. We suggest tracking and closely monitoring the behavior of these clients in order to be able to detect segments of customers who tend to have previous cancellations. This can be done with both Data

Mining and Machine Learning models and can provide good insights on what to address in a more original way this customers in order to avoid further cancellations.

- Different Room, which represents the difference between Assign and Reserved Rooms. Also for obvious reasons, the clients that have different (Assign and Reserved) rooms, represent customers that actually did CheckIN, and come to the hotel.

6. CONCLUSIONS

After conducting an in-depth analysis of the dataset from Hotel C, located in Lisbon, we have extracted significant valuable information. By leveraging the results of cross-validation on our best model, we have been able to optimize our operations and drive profitable improvements for the business by gaining an understanding of customer behavior. Our RandomForestClassifier model, with carefully selected parameters, achieved an accuracy of 82% on the training set and 81% on the validation set. This shows that the model is robust and not significantly overfitting, as the performance on the training and validation sets is similar. Importantly, the model has successfully reduced the cancellation rate from 42.16% to 23.26%. This is a significant improvement and achieves our objective.

In conclusion, these strategies can bring about a significant improvement in Hotel C's operations, reducing cancellations and overbookings and taking advantage of it’s data, using predictive modeling, and actively engaging with customers, the hotel can notably decrease cancellations, boost guest satisfaction, and in turn, increase revenue. This approach doesn't just position Hotel C as a mere accommodation provider, but as a place that understands and meets its customers' needs.

7. REFERENCES

PedroSancho. “BC2_Predicting_Hotel_Cancellations/GroupY_BC2_Hotel_Cancellations_RFC.ipynb at Main · PedroSancho/BC2_Predicting_Hotel_Cancellations.” GitHub, github.com/PedroSancho/BC2_Predicting_Hotel_Cancellations. Accessed 24 May 2023.

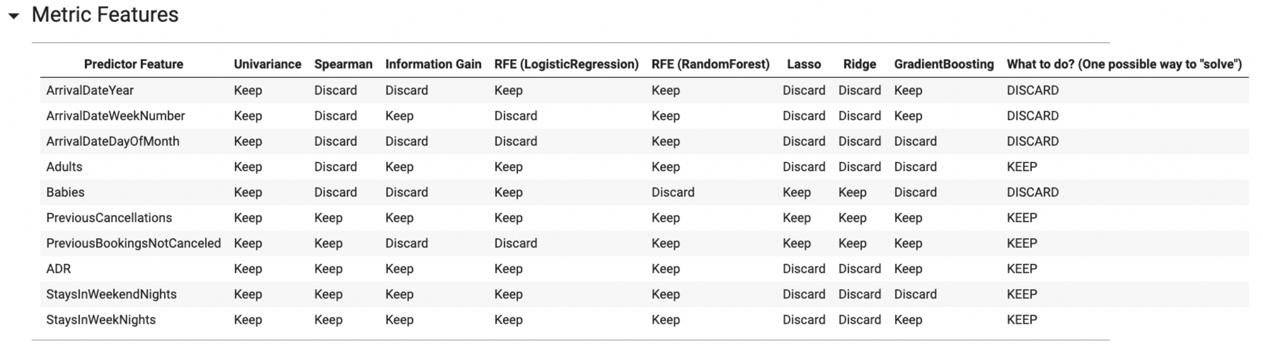
SaraxSilva. “Business-Cases-Projects-21-22/BC2_Predicting_Cancellations at Main · SaraxSilva/Business-Cases-Projects-21-22.” GitHub, github.com/SaraxSilva/Business-Cases-Projects-21-22. Accessed 24 May 2023.

Huilgol, Purva. “Precision and Recall | Essential Metrics for Data Analysis (Updated 2023).” Analytics Vidhya, 3 Sept. 2020, www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning.

OpenAI. (n.d.). ChatGPT. https://www.openai.com/

8. APPENDIX

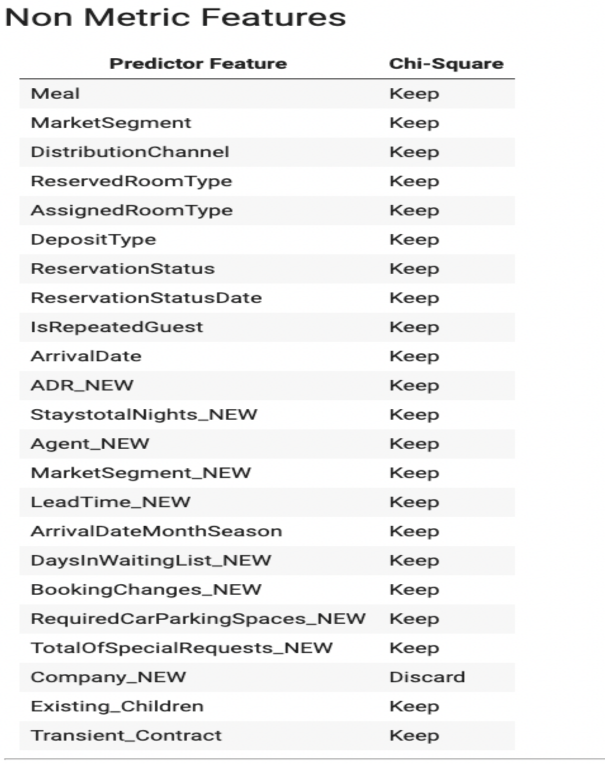
***Figure 15-** “Metric Features kept after Feature Selection ”*



Metric Features

Predictor Feature	Univariance	Spearman	Information Gain	RFE (LogisticRegression)	RFE (RandomForest)	Lasso	Ridge	GradientBoosting	What to do? (One possible way to "solve")
ArrivalDateYear	Keep	Discard	Discard	Keep	Keep	Discard	Discard	Keep	DISCARD
ArrivalDateWeekNumber	Keep	Discard	Keep	Discard	Keep	Discard	Discard	Keep	DISCARD
ArrivalDateDayOfMonth	Keep	Discard	Discard	Discard	Keep	Discard	Discard	Discard	DISCARD
Adults	Keep	Discard	Keep	Keep	Keep	Discard	Discard	Discard	KEEP
Babies	Keep	Discard	Discard	Keep	Discard	Keep	Keep	Discard	DISCARD
PreviousCancellations	Keep	Keep	Keep	Keep	Keep	Keep	Keep	Keep	KEEP
PreviousBookingsNotCanceled	Keep	Keep	Discard	Discard	Keep	Keep	Keep	Keep	KEEP
ADR	Keep	Keep	Keep	Keep	Keep	Discard	Discard	Keep	KEEP
StaysInWeekendNights	Keep	Keep	Keep	Keep	Keep	Discard	Discard	Discard	KEEP
StaysInWeekNights	Keep	Keep	Keep	Keep	Keep	Discard	Discard	Keep	KEEP

***Figure 16-** “ Non Metric Features kept after Feature Selection ”*



Non Metric Features

Predictor Feature	Chi-Square
Meal	Keep
MarketSegment	Keep
DistributionChannel	Keep
ReservedRoomType	Keep
AssignedRoomType	Keep
DepositType	Keep
ReservationStatus	Keep
ReservationStatusDate	Keep
IsRepeatedGuest	Keep
ArrivalDate	Keep
ADR_NEW	Keep
StaystotalNights_NEW	Keep
Agent_NEW	Keep
MarketSegment_NEW	Keep
LeadTime_NEW	Keep
ArrivalDateMonthSeason	Keep
DaysInWaitingList_NEW	Keep
BookingChanges_NEW	Keep
RequiredCarParkingSpaces_NEW	Keep
TotalOfSpecialRequests_NEW	Keep
Company_NEW	Discard
Existing_Children	Keep
Transient_Contract	Keep