Master's Degree Program in

MDSAA

Data Science and Advanced Analytics

Business Cases with Data Science

Case 4: Improving Revenue in the C hotel chain: A Data-Driven Approach to Reduce Cancellations and Optimize Bookings

Jaime Duarte: 20220675
Shanjida Roman: 20221395
Yousef Ebrahimi: 20221382
Diogo Martins: 20221361
David Martins: 20221006

Group D – Data Vision Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

May, 2023

INDEX

EXECUTIVE SUMMARY

This report describes the process our team conducted to develop a solution that allows Hotel C to better address the problematic increase in booking cancellations. The analysis utilizes a dataset of bookings from the hotel chain in Portugal, covering the period between July 1, 2015, and August 31, 2017. The goals of our analysis are to develop predictive models to forecast net demand based on reservations on the books, enabling better pricing, overbooking policies, and identification of bookings with a high likelihood of cancellation.

Following the CRISP-DM methodology, the process consisted of four main steps. First, the dataset was explored and analyzed to gain insights into patterns and trends related to cancellations and overall customer characteristics. Second, the data was cleaned, preprocessed, and transformed to ensure data quality and model performance. Next, we used Lazy Classifier to identify the best performing models for the task. The selected models included XGB classifier, Random Forrest Classifier and Bagging Classifier. Out of those, the best-performing model during training was XGB classifier which was then selected and used to make the final prediction. Using XGB classifier, our final prediction obtained a precision score of 0.87, a recall score of 0.87 and an f1-score score of 0.87 on the test data, and a precision score of 0.89, a recall score of 0.89 and an f1-score score of 0.89 on the train data, suggesting low likelihood of overfitting.

The developed model should be implemented into the hotel chain's business processes and be integrated with existing systems to forecast net demand, enabling proactive measures to prevent cancellations. The performance of the model will be monitored to ensure continuous improvement and maximize its effectiveness in reducing cancellations and increasing revenue.

1. 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.

The cancellation option puts the risk on hotels who have to honor 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. In Europe, the cancellation rate by reservation value, from 2014 to 2018, rose from 33% to 40%. Concerned about the increasingly negative impact caused by cancellations, hotel chain C, a chain with resort and city hotels in Portugal, hired us to evaluate the possibility of developing predictive models to predict net demand for their hotels. The hotel provided us with datasets of city hotels (H2), of bookings that were due to arrive between July 1, 2015, and August 31, 2017.

Cancellations occur for understandable reasons such as business meeting changes, vacation rescheduling, illness, or adverse weather conditions. However, cancellations are also done by customers looking to find a better deal. “Deal-seeking” customers tend to make multiple bookings for the same trip or make one booking but continue to search for better deals (e.g., looking for hotels with better social reputations, better prices, or better locations). The number of “deal-seeking” customers has grown immensely with the appearance of Online Travel Agencies (OTAs).

Currently overbooking is deployed to deal with this problem, but it has many downsides, as do restrictive cancellation policies, so no current solution seems viable.

2. AIMS AND OBJECTIVES

2.1 BUSINESS PROBLEM AND OBJECTIVES

The C hotel chain operation is not much different from other chains, in that it is deeply impacted by cancelations, with cancellations representing almost 28% in H1 and almost 42% in H2. Although several solutions like changing overbooking strategies have been attempted, none have been able to change the tide in their favor, due to the current unpredictable nature of cancelations.

Thus, the hotel management has now a goal of predicting these cancelations, to be able to deploy their chosen strategies. By accurately predicting cancellations and implementing appropriate strategies, the chain aims to minimize the negative impact caused by cancellations, including financial costs, social reputation damage, and revenue loss. Management expects to implement better pricing and overbooking policies but also do identify bookings with a high likelihood of cancellation. Identifying those bookings could allow the hotels to try to contact those bookings' customers and make offers to try to prevent cancellation (e.g., dinner, car parking, spa treatments, discounts, or other perks). The ultimate goal is to reduce the cancellation rate to 20% for the hotel chain, which would aid immensely in maximizing profit.

2.2 DATA MINING OBJECTIVES

The first minor goal of our project is to explore the provided data, as well as import any necessary data. This will allow us to understand customer characteristics and behavior patterns. Based on this analysis we should focus on the primary objective, which is to develop predictive models that can predict cancelations. These models should be tested for reliability and the best one chosen to be integrated. The best model will be chosen based on its scores in the chosen performance metrics: Accuracy, Precision, Recall and f1-score. A minimum threshold score of 0.80 was set as a goal, to ensure that the model is not just the best from our selection, but an objectively reliable one, able to reach the business goals described above. The data mining process will generate actionable recommendations based on the analysis and insights gained from the models. These recommendations can be used by the Revenue Manager Director and hotel staff to make informed decisions and take proactive measures to mitigate the effects of cancellations and improve overall operational efficiency. We will also provide clear documentation, that will facilitate understanding, replication, and future enhancements of the predictive model, as well as make suggestions on how the model can be deployed.

3. DATA UNDERSTANDING

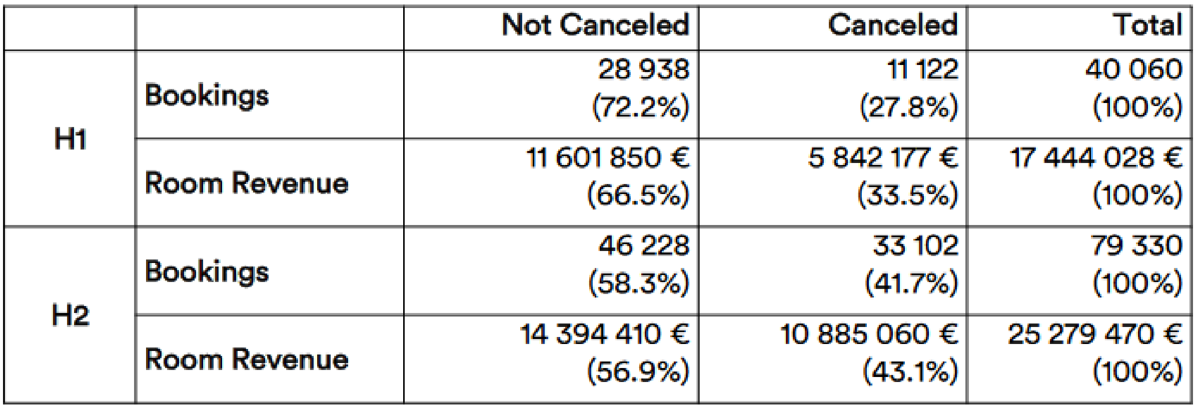
This phase began with data collection and proceeded with activities that enabled us to become acquainted with the data. The H2 dataset corresponds to a city hotel. It contains 79,330 observations and 31 variables. Each observation represents a hotel booking, including both bookings that were honored and bookings that were canceled. The data covers the period from July 1, 2015, to August 31, 2017. Below is a description of the features present in the dataset:

- ADR: Average Daily Rate
- Adults: Number of adults
- Agent: ID of the travel agency that made the booking
- ArrivalDateDayOfMonth: Day of the month of the arrival date
- ArrivalDateMonth: Month of arrival date with 12 categories: "January" to "December"
- ArrivalDateWeekNumber: Week number of the arrival date
- ArrivalDateYear: Year of the arrival date
- AssignedRoomType: Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation

reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons

- Babies: Number of babies
- BookingChanges: Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
- Children: Number of children
- Company: ID of the company/entity that made the booking or is responsible for paying the booking. ID is presented instead of designation for anonymity reasons
- Country: Country of origin. Categories are represented in the ISO 3155-3:2013 format
- CustomerType: Type of booking, assuming one of four categories:
- Contract - when the booking has an allotment or other type of contract associated to it;
- Group – when the booking is associated to a group;
- Transient – when the booking is not part of a group or contract, and is not associated to other transient booking;
- Transient-party – when the booking is transient, but is associated to at least other transient booking
- DaysInWaitingList: Number of days the booking was in the waiting list before it was confirmed to the customer
- DepositType: Indication if the customer made a deposit to guarantee the booking. This variable can assume three categories:
- No Deposit – no deposit was made;
- Non Refund – a deposit was made in the value of the total stay cost;
- Refundable – a deposit was made with a value under the total cost of the stay.
- DistributionChannel: Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
- IsCanceled: Value indicating if the booking was canceled (1) or not (0)
- IsRepeatedGuest: Value indicating if the booking name was from a repeated guest (1) or not (0)
- LeadTime: Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
- MarketSegment: Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
- Meal: Type of meal booked. Categories are presented in standard hospitality meal packages:
- Undefined/SC – no meal package;
- BB – Bed & Breakfast;
- HB – Half board (breakfast and one other meal – usually dinner);
- FB – Full board (breakfast, lunch and dinner)

- PreviousBookingsNotCanceled: Number of previous bookings not cancelled by the customer prior to the current booking
- PreviousCancellations: Number of previous bookings that were cancelled by the customer prior to the current booking
- RequiredCarParkingSpaces: Number of car parking spaces required by the customer
- ReservationStatus: Reservation last status, assuming one of three categories:
- Canceled – booking was canceled by the customer;
- Check-Out – customer has checked in but already departed;
- No-Show – customer did not check-in and did inform the hotel of the reason why
- ReservationStatusDate: Date at which the last status was set. This variable can be used in conjunction with the
- ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel
- ReservedRoomType: Code of room type reserved. Code is presented instead of designation for anonymity reasons
- StaysInWeekendNights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- StaysInWeekNights: Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
- TotalOfSpecialRequests: Number of special requests made by the customer (e.g. twin bed or high floor)

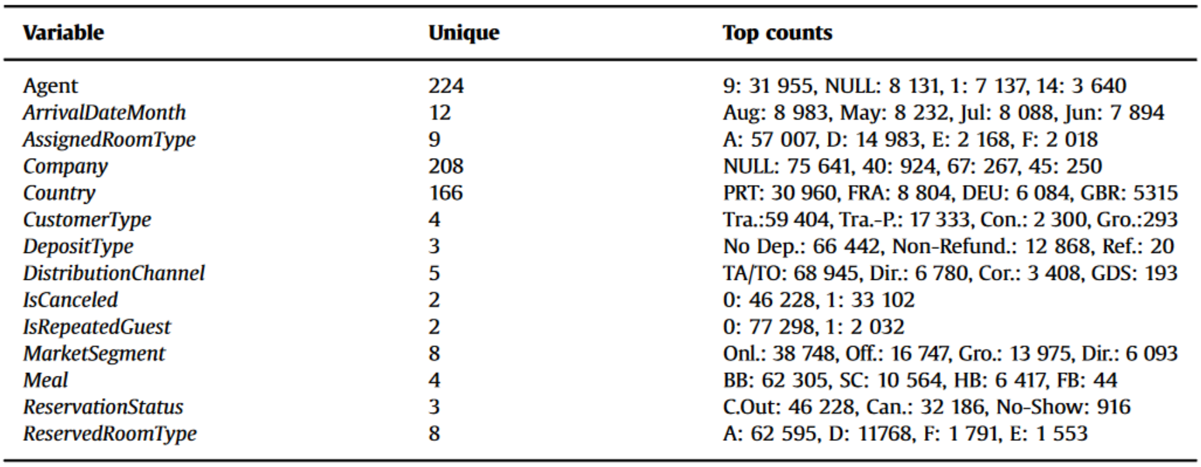


		Not Canceled	Canceled	Total
H1	Bookings	28 938 (72.2%)	11 122 (27.8%)	40 060 (100%)
	Room Revenue	11 601 850 € (66.5%)	5 842 177 € (33.5%)	17 444 028 € (100%)
H2	Bookings	46 228 (58.3%)	33 102 (41.7%)	79 330 (100%)
	Room Revenue	14 394 410 € (56.9%)	10 885 060 € (43.1%)	25 279 470 € (100%)

Table 1: Cancellations in H1 and H2

By checking the data in some columns, we get results for a correct understanding of the data, for example:

MarketSegment: With a count of 38,748, it demonstrates that the largest sector consisted of travelers who made their reservations through online travel agencies (online TAs). Visitors who used offline travel agencies or tour operators (offline TA/TO) make up the second largest sector with a total of 16,747. Also, 13,975 of the guests at the hotel were traveling in groups.



Variable	Unique	Top counts
Agent	224	9: 31 955, NULL: 8 131, 1: 7 137, 14: 3 640
ArrivalDateMonth	12	Aug: 8 983, May: 8 232, Jul: 8 088, Jun: 7 894
AssignedRoomType	9	A: 57 007, D: 14 983, E: 2 168, F: 2 018
Company	208	NULL: 75 641, 40: 924, 67: 267, 45: 250
Country	166	PRT: 30 960, FRA: 8 804, DEU: 6 084, GBR: 5315
CustomerType	4	Tra.:59 404, Tra.-P.: 17 333, Con.: 2 300, Gro.:293
DepositType	3	No Dep.: 66 442, Non-Refund.: 12 868, Ref.: 20
DistributionChannel	5	TA/TO: 68 945, Dir.: 6 780, Cor.: 3 408, GDS: 193
IsCanceled	2	0: 46 228, 1: 33 102
IsRepeatedGuest	2	0: 77 298, 1: 2 032
MarketSegment	8	Onl.: 38 748, Off.: 16 747, Gro.: 13 975, Dir.: 6 093
Meal	4	BB: 62 305, SC: 10 564, HB: 6 417, FB: 44
ReservationStatus	3	C.Out: 46 228, Can.: 32 186, No-Show: 916
ReservedRoomType	8	A: 62 595, D: 11768, F: 1 791, E: 1 553

Figure 2: H2 dataset summary statistics – Categorical variables

4. DATA PREPARATION AND EXPLORATION

4.1 DATA PREPARATION

4.1.1 Clean Data – Data was cleaned before visualizations. First, we replaced missing values with the mode (24 values for Country and 4 for Children). To help identify outliers, boxplots were used. This allowed for a manual process of handling outliers. Other methods like using the Inter-Quartile Range were performed but were not used to remove more data than desired. Using the Interquartile Range (IQR) method, approximately 66.48% of the data was retained after removing outliers based on the upper and lower limits calculated from the 20th and 80th percentiles. Additionally, through the manual removal of outliers based on specific criteria for each feature, approximately 99.39% of the data was retained. These outlier removal techniques help to ensure a cleaner dataset for further analysis.

4.1.2 Format data/Feature engineering – In this stage, we combined the “sub-region” and “region” columns from the country CSV file in ISO-3166 format with the already-existing dataset to include regional data.

Also, we merged the three distinct columns for the arrival date's year, month, and day into a single column of DateTime data called "ArrivalDate." In later calculations and visualizations, this makes it easier to manipulate and analyze the arrival dates. In addition, we extract additional temporal features from the 'ArrivalDate', including the day of the week (0-6, Monday-Sunday) and the quarter of the year (1-4).

Lastly, we created a new feature named 'TotalBookings' by summing the number of previous cancellations and the number of previous bookings that were not canceled. This combined metric provides a consolidated view of the overall booking history for each reservation.

In both metric and non-metric features, it seems that the algorithms both classify close to all variables as important for our target variable. In that sense, we are keeping all of them (except some that don't make sense for clustering, which will be addressed for that purpose), and then when we proceed with modeling, we will apply multiple feature selection methods to see which one gives us the best results.

To further examine customer booking behavior, we implemented the 'TotalBookings' feature. This feature provides a complete record of all reservations made by a customer by adding the number of prior cancellations with the number of prior reservations that were not canceled. In the case where hotels amend their policies and plans it is possible to take account of prior booking activity by consumers.

The 'Length of Stay' feature was then calculated by determining the difference in days between the 'ReservationStatusDate' and the 'ArrivalDate'. The duration of the customer's stays is quantized by this Metric, which may be used to detect any correlation between length of stay and likelihood of cancellation. The 'Length of Stay' was set at 0 to consider the actual duration of a stay in respect of reservations marked as canceled or without shows.

In addition, we calculated the 'time in system' parameter by subtracting the booking date from the status reservation date. To analyze the overall processing times and efficiency of hotel reservation management practices, this measure represents the number of days a booking is held in the hotel's system.

A 'RoomChange' column was introduced to indicate whether a room change occurred for each booking. This binary feature helps identify instances where customers requested or were assigned a different room type than what was initially reserved. Room changes can provide valuable insights into customer preferences and potential factors influencing cancellations.

Finally, the dataset was divided into two categories: metric and non-metric features. The 'metric_features' list included columns representing numerical or continuous variables, while the 'non_metric_features' list encompassed categorical or discrete variables. This categorization facilitated further analysis and modeling based on the nature of the features.

Overall, these transformations and calculations enriched the dataset by creating new features that provide valuable insights into booking patterns, customer behavior, and potential factors influencing cancellations. This enhanced dataset served as the foundation for subsequent analyses and predictive modeling to optimize revenue and minimize the impact of cancellations for the hotel chain.

5. DATA VISUALIZATION

Data visualization provided us with a better understanding of key patterns and trends. By employing various visualization techniques, we explore significant aspects such as guest nationalities, room changes, and the distribution of bookings across different periods.

A choropleth map was created to visualize the distribution of hotel guests by country. The complete, interactive map is available in the report.

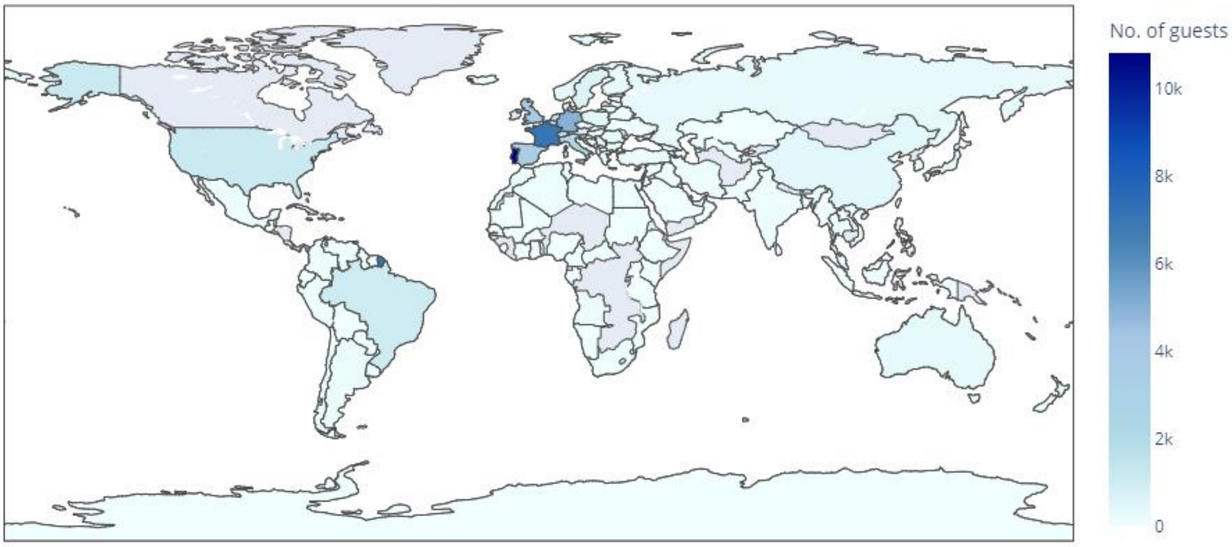


Figure 3: Choropleth map of hotel guests by country

Fig. 3 which is a bar chart, showcasing the top 10 countries based on the number of guests. This visualization aims to provide a clear and concise representation of the countries with the highest guest count in the dataset. The majority of the customers are coming from Portugal, followed by other European countries like France, Germany, the United Kingdom, and Spain.

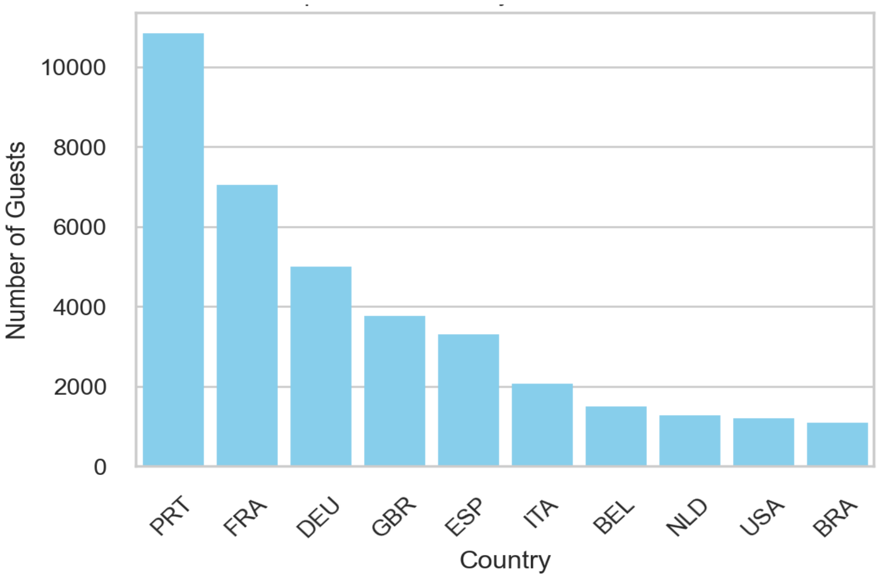


Figure 4: Top 10 countries by number of guests

The line chart below was used to analyze and illustrate the data regarding cancellation patterns. The monthly total of cancellations was displayed in the chart, allowing us to spot patterns and trends in client behavior.

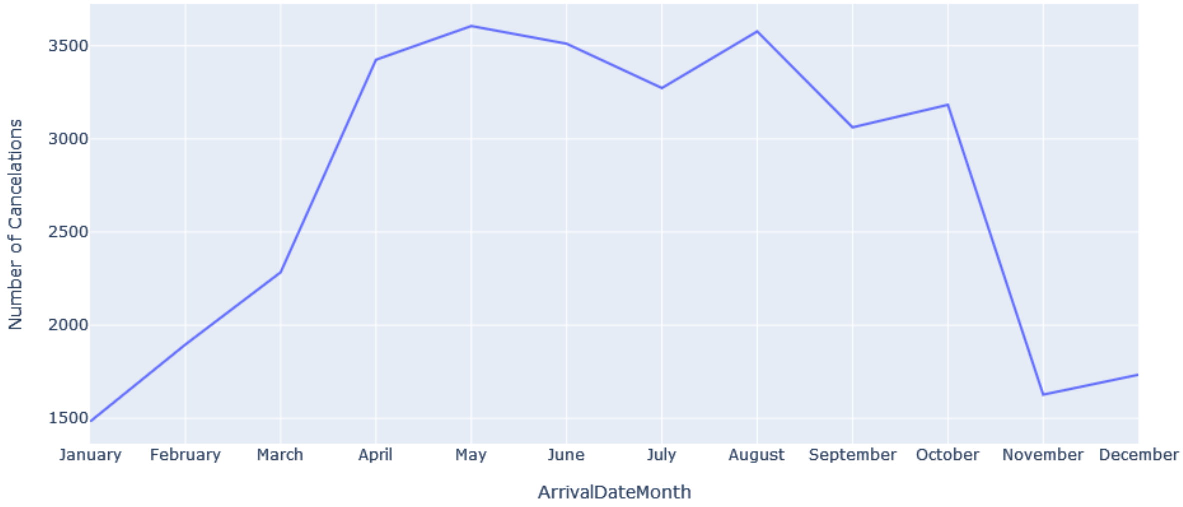


Figure 5: Total number of cancellations per month

It appears that longer lead times are associated with higher cancellation rates which imply that people who book further in advance are more likely to cancel their bookings. There could be various reasons for this behavior. For example, people who book early might change their plans or find better deals closer to their travel dates, leading to a higher likelihood of cancellation. Additionally, uncertainties or changes in personal circumstances may arise over time, prompting individuals to cancel their bookings. Solutions for this issue:

- Improve Customer Service
- Offer Flexible booking options - allow customers to modify or cancel their reservations without incurring heavy penalties. This can reduce the perceived risk of booking and make customers feel more comfortable making reservations in advance.
- Provide incentives for non-cancellation

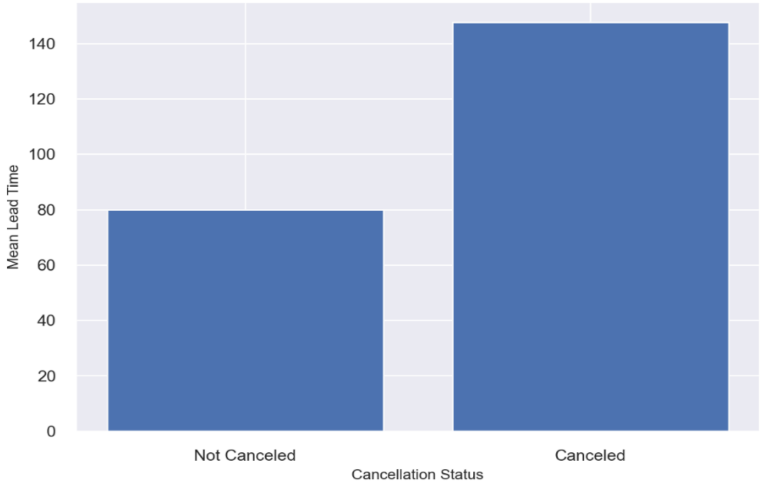


Figure 6: Mean Lead Time per cancelation status

In this analysis, we examined the distribution of reservation statuses across different months. By grouping the data and creating a table, we were able to observe the count of reservations for each month and their respective statuses. To visualize the results, a bar chart was used, allowing us to easily compare the distribution of reservation statuses across the months. The Fig. 6 shows a fairly even distribution, indicating a consistent pattern in reservation statuses throughout the year and it appears quite proportional.

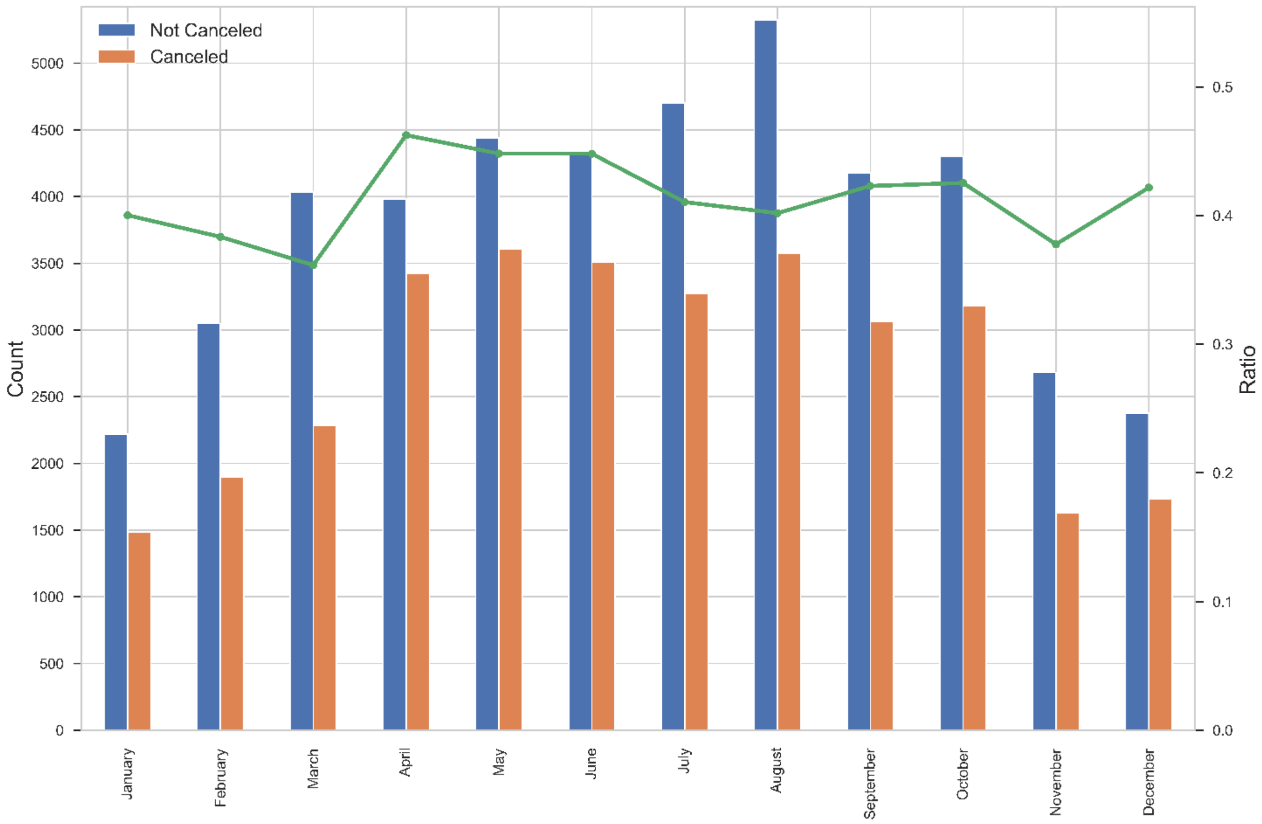


Figure 7: Distribution of Reservation Status by Month

The plot reveals the busiest months (May through August), but also the ratio between the number of people canceling and not. This information might be useful for the hotel, as they can choose to be more open to overbooking during the months with the highest ratio of cancelations, such as April, May and June.

We examine the price of different room types per night and per person. The boxplot visualization showcases the distribution of prices for each room type, allowing us to identify any variations and outliers. By analyzing this information, we can make informed decisions regarding pricing strategies and understand the value associated with each room type. The plot provides a clear overview of the price range, highlighting any potential trends or discrepancies.

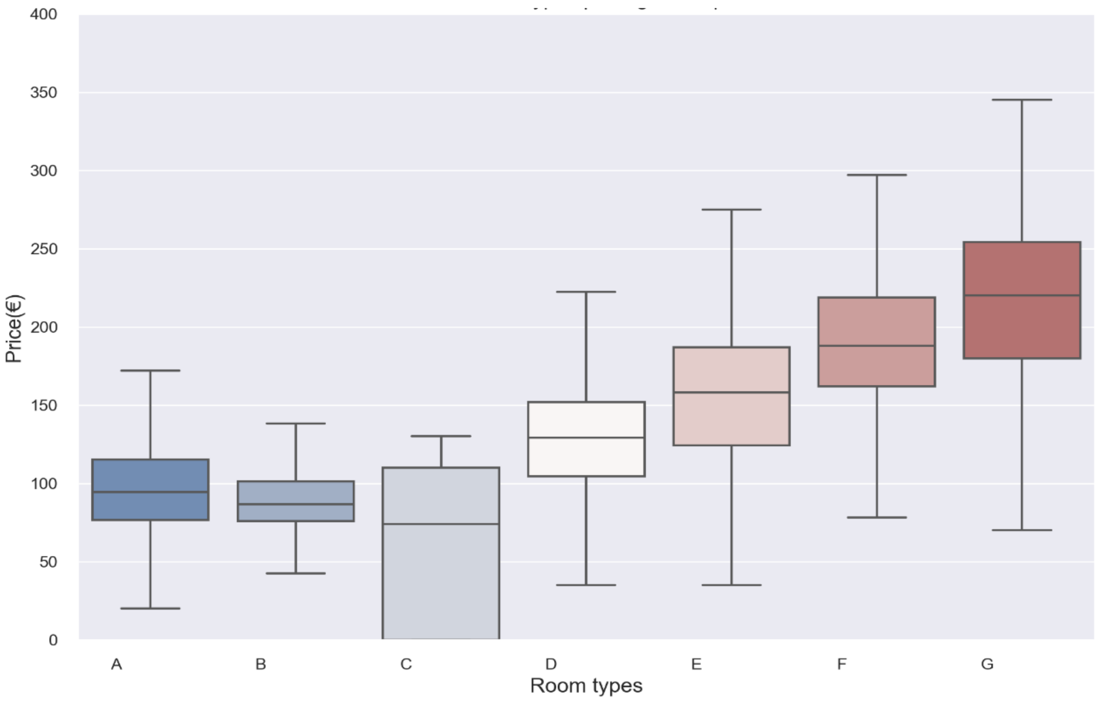


Figure 8: Price of room types per night and person

We filter the dataset to focus on the top 10 countries with the highest number of guests. By analyzing this subset of data, we calculate the average length of stay for each country. We create a histogram plot where each bar represents a country and its corresponding average length of stay. In Fig. 9 Germany and France are the top countries with long lengths of stay compared to other countries.

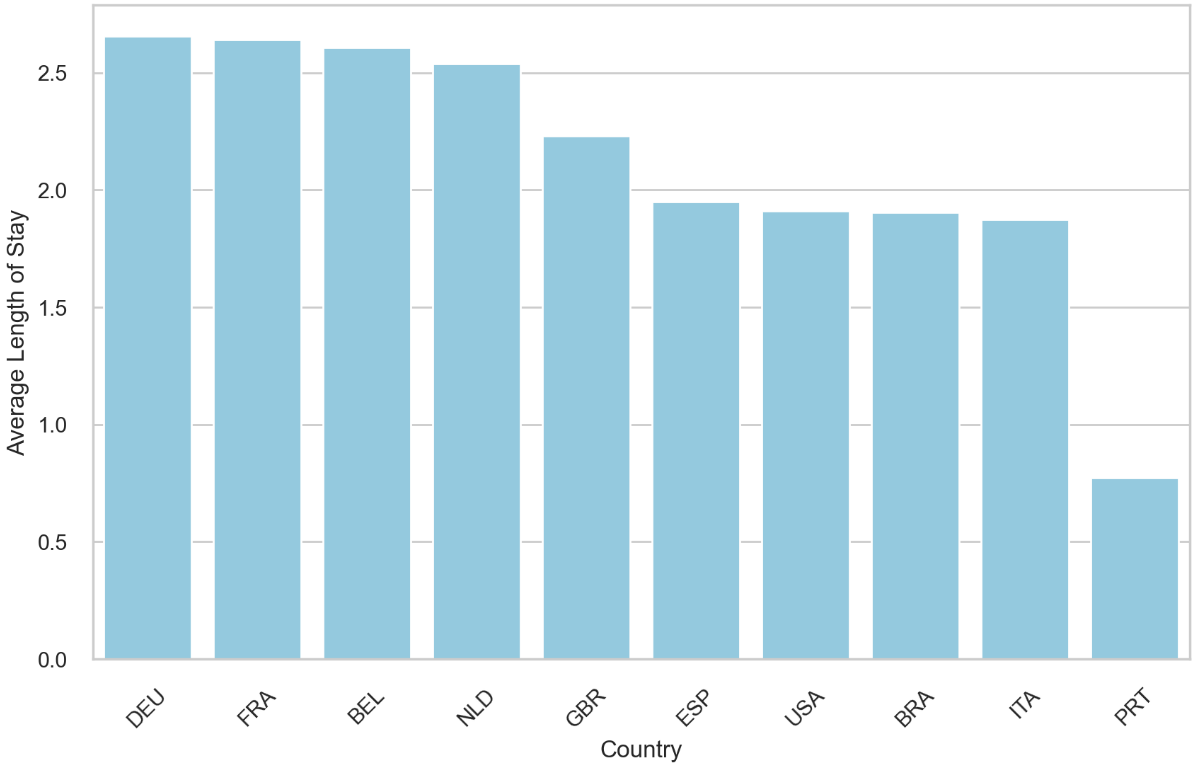


Figure 9: Average Length of Stay - Top 10 Countries

Also, we group the dataset by the month of arrival and calculate the average length of stay for each month. To visualize this information, we create a line chart where each point represents a month and its corresponding average length of stay. The chart shows us to observe any patterns or trends in the average length of stay over the months. We can see that during summer, July, and August, the guests stayed more. Also in March, there is a pick with a 1.9 average length of stay. On the other hand, in May and June, the length of stays reduces.

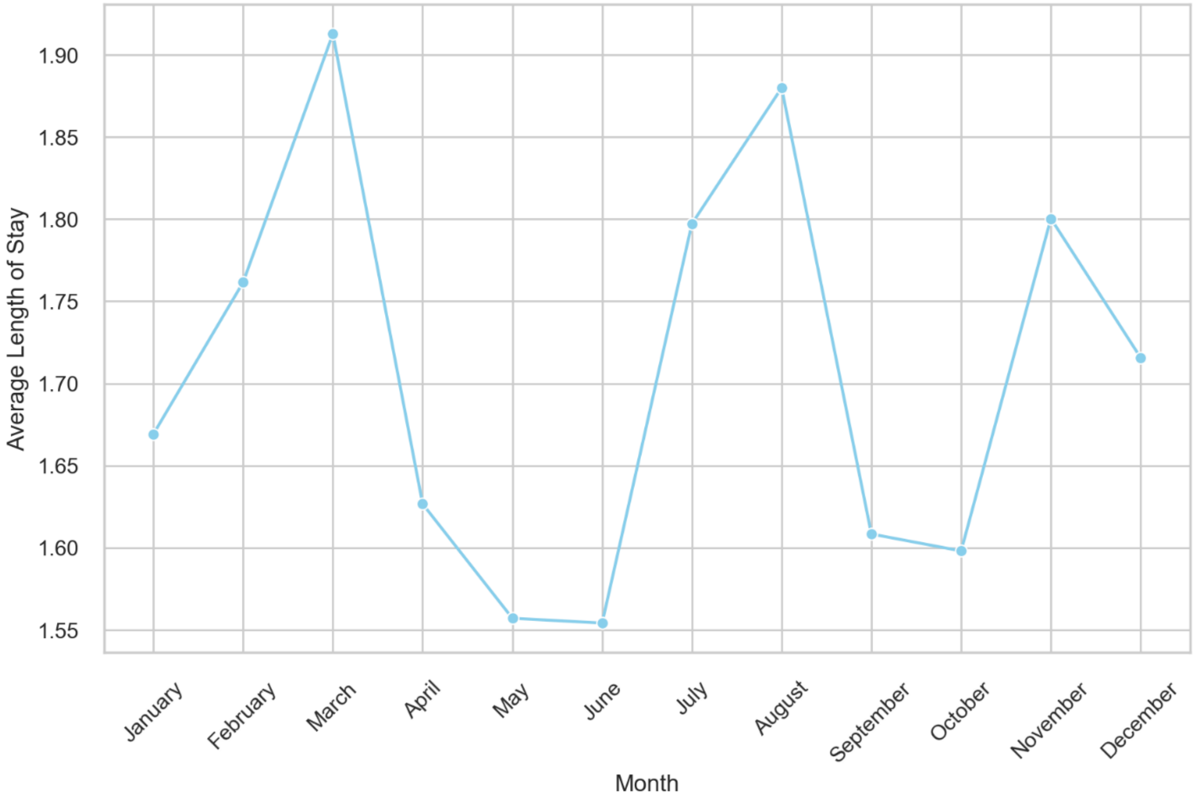


Figure 10: Average Length of Stay per Month

6. CLUSTERING

In this hotel chain, understanding customer behavior and segmenting guests based on their preferences and booking patterns is crucial for personalized marketing strategies and enhanced service delivery. One powerful technique for achieving this is step clustering, which allows hotels to identify distinct groups of guests with similar characteristics and behaviors. After normalization and encoding the data, we did clustering. Clustering, although not the main focus of the project, could aid in the prediction capabilities of our model, but also provide us and the Hotel with some good insights about their customers. By leveraging clustering, the hotel can also gain valuable insights into guest segmentation, enabling them to tailor their marketing campaigns, pricing strategies, and service offerings to cater to the unique needs and preferences of each guest segment.

6.1 PCA

In the visualization using PCA with 2 components, a scatter plot is created with the two principal components as axes. Each point represents an observation, and the color represents the "target" variable. It helps us understand how much of the total variance in the data is captured by each additional component. The factor loadings represent the correlation between the original features and the principal component. The top 16 features with the highest and lowest factor loadings for the 1st component are displayed.

```
                Top 16 Highest
----------------------------------------
   TotalBookings                    0.56
   PreviousBookingsNotCanceled      0.52
   PreviousCancellations            0.42
   LeadTime                         0.06
   DaysInWaitingList                0.05
   RequiredCarParkingSpaces         0.02
   Time in the System              -0.02
   Babies                          -0.03
   BookingChanges                  -0.05
   Children                        -0.12
   TotalOfSpecialRequests          -0.14
   StaysInWeekendNights            -0.15
   StaysInWeekNights               -0.17
   Adults                          -0.20
   Length of Stay                  -0.21
   ADR                             -0.25
```

Table 3: Features with highest and lowest factor loadings for the 1st component

We applied dimension reduction using PCA with 16 components on the normalized metric dataset. Then, we used the Elbow method and Silhouette method to determine the appropriate number of clusters (K) for K-means clustering. The Elbow method indicates K should be 6, but the Silhouette method points to higher values such as 17. However, since a large number of clusters are not very helpful for marketing purposes, we selected K as 6.

6.2 K-MEANS

In our hotel case, we employed the K-means clustering technique to gain insights into our guests' booking patterns and preferences. By analyzing data such as the number of stays, booking channel, lead time, and previous cancellations, we were able to identify distinct groups of guests with similar characteristics.

We began by employing methods to determine the optimal number of clusters such as the silhouette score and elbow method. This led us to select 7 clusters We visualized the clusters' sizes and distances in 2D using the Inter-cluster distance plot. We adopted this method, since it achieved the highest Calinski Harabasz Index score, with 13.457.

6.2.1 Cluster Solution

The clusters are quite heterogeneous in regard to customer average behavior, but also in size, with the vast majority of customers falling into cluster 0 and 6:

Cluster 0: 39,432 data points

Cluster 0 represents a group of guests who typically have shorter weekend and weekday stays, fewer adult guests per booking. They opt for more affordable room rates (3rd lowest "ADR") and have fewer special requests during their stay. It has a very even distribution between people who cancel bookings, and people who don't, as well as between people who are repeated guests and those that are not.

Cluster 1: 5,263 data points

Cluster 1 represents a group of guests who tend to cancel their bookings more frequently, book well in advance, stay for shorter durations, prefer more affordable room rates, make fewer special requests during their stay, have a higher proportion of repeated guests, have a history of previous cancellations and bookings, exhibit a higher frequency of room change requests, prefer booking during peak weeks or specific periods, and have a higher proportion of guests from Portugal. This group is also the one with more people being repeated guests.

Cluster 2: 4,649 data points

Cluster 2 sets itself apart by their number of children, which is on average much higher than any other cluster. This group appears to have no preference regarding staying on the weekend or week. They are also not repeated guests on average, which is unfortunate, since this group has by far the greatest average "ADR". They don't book with a lot of antecedence and have a middle-of-the-pack level of booking cancelation. This cluster, that also has the 2nd highest rate of foreigners, is very valuable, and attempts should be made to convert these customers into repeating guests.

Cluster 3: 1,870 data points

This group has the 2nd lowest average "Lead Time", meaning they don't usually book very much in advance. They are the group with the 2nd biggest rate of repeated guests, and the one with the lowest ratio of people canceling bookings. They are the most demanding group in terms of car parking spaces and their "ADR", while not close to the highest, is still one of the highest. This cluster is very valuable to do their aversion to canceling bookings

Cluster 4: 363 data points

Cluster 4 is characterized by the number of babies they usually bring, they are also the group making the most special requests and booking changes, likely to accommodate their infants. This group is the 2nd least likely to cancel a booking and also the group with the 2nd bigger average length of stay. Making sure the hotel complies with the requests of this group could lead to an improvement of their “IsRepeatedGuest” average, which is very middle-of-the-pack at the moment.

Cluster 5: 1,623 data points

This group books with great antecedence. Guests on this group usually aren’t coming back to the hotel. They prefer to stay during wee nights, usually towards the end of the month. They are the group with the 2nd highest rate of Portuguese people. They don’t have a lot of children or babies and make few special requests. Unfortunately, they are the 2nd most likely to cancel the booking and have the second lowest length of stay and ADR. Thus, we believe this group might be one of the most problematic in terms of profit for the hotel

Cluster 6: 25,079 data points

The second largest cluster, it has the highest ratio of foreigners. It has the highest average length of stay, as well as the second highest ADR. They show no real preference between staying on the weekends or during the week. They rarely bring children or babies with them. This group makes a lot of special requests, but not a lot of demands of a room change. They have a relatively low rate of cancelation, but also of being repeated guests unfortunately for the Hotel. This group should be regarded as very important, due to its size and characteristics. The Hotel should make efforts to retain and satisfy these customers as much as possible.

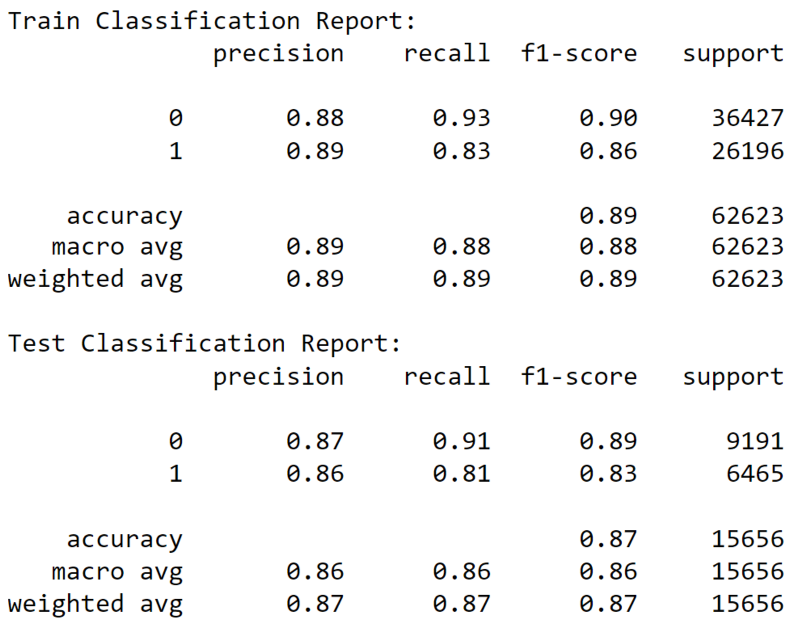
7. FEATURE SELECTION

Before beginning the model building stage, we performed the hold-out method to set aside 20% of the data to be used for the final testing. We then performed several feature selection techniques such as correlations, 2-Chi-Squared and Recursive Feature Extraction to reach our final selection of features for modeling. In total we were left with 68 features, which we then used for our modeling.

8. MODELING

The first step was using the Lazy Predict library’s Classifier, allowing us to access the adequacy of several different models before any hyper-tuning. The best performing models, XGBoost Classifier, Bagging Classifier and Random Forrest Classifier were then hyper-tunned. Using Research, we were able to exhaustively search for the best parameters for the model, by subjecting several parameter combinations to various folds of the train data. We evaluated the average performance during those folds, using measures such as accuracy, precision, recall, f1-score, and area under the ROC curve. XGBoost Classifier was the highest performing after hyperparameterization, and thus was chosen as the final model. It was then trained on the complete dataset once, before finally making the prediction with the hold-out method previously mentioned.

The final scores of our model in the selected performance metrics were the following:



```
Train Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.93      0.90     36427
           1       0.89      0.83      0.86     26196

    accuracy                           0.89     62623
   macro avg       0.89      0.88      0.88     62623
weighted avg       0.89      0.89      0.89     62623

Test Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.91      0.89      9191
           1       0.86      0.81      0.83      6465

    accuracy                           0.87     15656
   macro avg       0.86      0.86      0.86     15656
weighted avg       0.87      0.87      0.87     15656
```

Table 2: Final Test scores for the XGB Classifier model

As you can see the results achieved are promising. Our model achieved scores above 0.85 on the selected measures, with similar scores on both the training and test data, lowering the likelihood of overfitting having occurred. A recall (sometimes referred to as sensitivity) score over 0.80 tells us how many of the positive cases the classifier managed to predict correctly, over all the positive cases in the data. In terms of what this means to the hotel, out of 100 "cancelers", the model is able to correctly identify around 80 of them. With precision scores of 0.87, the Hotel could also be very certain that anyone identified as a potential "canceler" is very likely one.

7. DEPLOYMENT AND MAINTENANCE

7.1 DEPLOYMENT STAGE

The deployment stage of the model involves the process of making it accessible and operational for real-world usage. This stage focuses on integrating the model into the existing hotel booking system or a new dedicated system. The key activities in this stage include:

a) **System Integration**: Integrating the prediction model with the hotel booking system, ensuring seamless communication and data flow between the two components.

b) **API Development**: Creating an API that exposes the model's functionality, allowing other systems or applications to interact with the prediction model.

c) **Infrastructure Setup**: Setting up the necessary infrastructure, including servers, databases, and necessary resources to host and serve the model.

d) **User Interface**: Developing a user-friendly interface, such as a web application or dashboard, to facilitate user interaction with the prediction model and display relevant insights.

e) **Testing and Quality Assurance**: Conduct thorough testing to validate the deployment process, ensuring the model performs as expected and meets the required performance metrics.

Timeframe: The deployment stage typically takes around 2-4 weeks, depending on the complexity of the integration, infrastructure requirements, and testing efforts.

7.2 MAINTENANCE STAGE

The maintenance stage focuses on ensuring the long-term performance, reliability, and effectiveness of the Hotel Booking Cancellation Prediction Model. This stage involves ongoing activities to monitor, update, and enhance the model based on feedback, new data, and changing business requirements. The key activities in this stage include:

a) **Monitoring and Performance Evaluation**: Implementing monitoring mechanisms to track the model's performance, identify anomalies, and address issues promptly. Regular evaluation of model performance against predefined metrics is crucial to ensure its accuracy and reliability.

b) **Data Updates**: Incorporating new data into the model to maintain its relevance and adapt to evolving patterns or trends in hotel bookings. This may involve periodic data collection, preprocessing, and retraining of the model to improve its predictive capabilities.

c) **Model Updates and Enhancements**: Analyzing feedback and insights from users, stakeholders, and domain experts to identify areas for model improvement. This may involve refining features, incorporating new algorithms or techniques, and addressing any limitations or biases identified during model usage.

d) **Security and Privacy**: Ensuring the model and associated infrastructure adhere to security and privacy standards, protecting sensitive customer information, and preventing unauthorized access.

e) **Documentation and Communication**: Maintaining comprehensive documentation of the model, including its architecture, dependencies, and usage guidelines. Communicating model updates, enhancements, and performance reports to relevant stakeholders.

The maintenance stage is an ongoing process that extends throughout the model's lifecycle. Regular monitoring, updates, and enhancements can be expected to take place on a monthly or quarterly basis, depending on the availability of new data and the identified needs for improvement.

8. CONCLUSION

Using ML models to analyze hotel data has provided insightful information about market segmentation, visitor demographics, and booking for H2 hotel. Clustering techniques helped us identify distinct guest groups for targeted marketing. In the end, we were able to complete the task we set out to do, having developed a model capable of accurately predict customer behavior regarding booking cancelation. The models achieved scores on the selected measures above our initial targets on both the train and test data. We also provided a few insights using our data exploration that we believe will aid the Hotel management in reaching their business goals. We believe projects such as this one offer opportunities for revenue management, guest profiling, and enhancing customer satisfaction in the hotel industry.

9. REFERENCES

Antonio, N., de Almeida, A. and Nunes, L. (2019) 'Hotel booking demand datasets', Data in Brief, 22, pp. 41–49. doi:10.1016/j.dib.2018.11.126.

Hertzfeld, E. (2019) Study: Cancellation rate at 40% as otas push free to change policy, Hotel Management. Available at: https://www.hotelmanagement.net/tech/study-cancelation-rate-at-40-as-otas-push-free-change-policy (Accessed: 18 May 2023).

Talluri, K.T. and Ryzin, G.V. (2009) The theory and practice of Revenue Management. New York: Springer.

Agag, G. and El-Masry, A.A. (2016) 'Understanding the determinants of hotel booking intentions and moderating role of Habit', International Journal of Hospitality Management, 54, pp. 52–67. doi:10.1016/j.ijhm.2016.01.007.

Lukes (no date) Lukes/ISO-3166-countries-with-regional-codes: ISO 3166-1 country lists merged with their UN geoscheme regional codes in ready-to-use JSON, XML, CSV data sets, GitHub. Available at: https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes (Accessed: 19 May 2023).

Silhouette visualizer (no date) Silhouette Visualizer - Yellowbrick v1.5 documentation. Available at: https://www.scikit-yb.org/en/latest/api/cluster/silhouette.html (Accessed: 20 May 2023).

10. APPENDIX (LIST OF FIGURES AND TABLES)