NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 4: PREDICTING HOTEL BOOKING CANCELLATIONS

Ana Miguel Monteiro, number: 20221645

Ana Rita Viseu, number: 20220703

Miguel Cruz, number: 20221391

Rodrigo Brigham, number: 20221607

Sara Galguinho, number: 20220682

Group G

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

May, 2023

Index

1. EXECUTIVE SUMMARY

This project aims to develop a predictive model for hotel booking cancellations specifically tailored to H2, a city hotel located in Lisbon belonging to hotel chain C. By leveraging advanced data analysis techniques, our objective is to uncover meaningful patterns and trends in customer behavior that can help us predict and anticipate hotel booking cancellations. Understanding the factors that contribute to booking cancellations will enable us to gain valuable insights into customers' preferences, habits, and external factors that influence their decisions. With this knowledge, we can help hotel H2 to take proactive measures in reducing cancellation rates, implementing better pricing and overbooking polices, and improving operational efficiency.

The hospitality industry has witnessed a surge in booking cancellations due to various factors being very relevant the increase in competitive offerings and "deal-seeking" customers who make multiple bookings or continue to search for better deals. A main contributor is the rise of Online Travel Agencies (OTAs), which charge hotels commissions ranging from 15% to 30% and push for a free cancellation policy. In response to this problem, hotels use overbooking, which creates problems such as reallocation costs, social reputation damage, and loss of revenue. Restrictive cancellation policies, such as nonrefundable rates, can also lead to a decrease in revenue and the number of bookings. These cancellations have a direct impact on the hotel's revenue, making it essential for hotel H2 to develop effective strategies to mitigate cancellations and maximize revenue potential.

During this project, the CRISP-DM methodology was followed, starting by gaining a broader understanding of the business. To do this, the hotel H2's background was explored, the business objectives were clarified and the success criteria for the problem were defined, to afterwards measure the performance of the findings and predictions. The available resources were also assessed, as well as risks and costs/benefits that this project entailed. As a final step, the machine learning goals - objectives in technical terms - for this project were determined. Afterwards, the data was analyzed and prepared. This involved exploring, cleaning and organizing the data, as well as identifying and addressing any missing values or outliers, along with analyzing and creating new features. A thorough description of the cancellation patterns in the customer base was also performed using data visualization techniques. Then, many models were explored and, by applying LGBM Classifier, a F1-score of 0.773 and a recall of 0.695 were achieved on the test dataset.

Based on our evaluation, useful insights were found, such as the influence of non-refund deposits on the cancellation rate and how the number of special requests impacts the predictions. With these insights, recommendations to help counterbalance cancellations were developed, notably the implementation of a loyalty program and the preparation of customized offers to bookings predicted to cancel. More information would be needed to estimate the cost of these suggestions.

In summary, a robust predictive model for hotel booking cancellations at hotel H2 was developed, enabling them to proactively manage cancellations, optimize revenue, and deliver exceptional guest experiences. This strengthens hotel H2's competitive position in the market and, with the recommendations provided, ensures the reduction of the cancellation rate to 20%.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. BACKGROUND

Much like other hotel companies, chain C has been severely impacted by high cancellation rates. The chain's resort hotel (H1) has experienced a cancellation rate of nearly 28%, while the city hotel (H2) had an even higher cancellation rate of almost 42%. This project specifically focuses on addressing the challenges faced by hotel H2, located in Lisbon, Portugal. In response to the increasingly negative impact caused by cancellations, the Revenue Manager Director of hotel chain C initially implemented a strategy limiting the number of rooms sold with restrictive cancellation policies, while adopting a more aggressive overbooking policy, which proved unsuccessful and costly. Subsequently, the hotel director softened the overbooking policy, which also proved to be ineffective.

To overcome the challenge of cancellations, hotels must consider adopting more flexible and customer-centric policies. This includes implementing dynamic pricing strategies and offering personalized incentives based on customer behavior. It is also crucial for hotels to invest in building brand loyalty and direct booking channels to reduce the reliance on Online Travel Agencies (OTAs) and increase customer retention. By adopting a proactive and innovative approach to the issue of cancellations, hotel chain C can improve their business strategy and mitigate their financial losses.

Predicting hotel booking cancellations is a critical aspect of an effective business strategy, since it helps hotels to anticipate future net demand (demand minus cancellations). Accurate cancellation predictions allow for more informed decisions on everything from services planning to pricing and marketing. By anticipating cancellation patterns, hotels can also manage their inventory to meet the expected demand.

Thus, by working closely with hotel H2's team/stakeholders, our goal is to leverage the provided historical data and apply machine learning techniques to accurately predict hotel booking cancellations. Consequently, the hotel will be able to make precise predictions, allowing for more informed decisions regarding pricing, overbooking policies, and resource allocation. By effectively managing cancellations, Hotel H2 can optimize revenue generation and enhance their overall business performance.

2.2. BUSINESS OBJECTIVES

Our primary goal for this project is to accurately identify bookings with a high likelihood of cancellation of hotel H2. Through analyzing its customer base and implementing predictive models, we aim to assist the hotel and seek to address the following key business questions:

- Identifying the variations of bookings and cancellations over time;
- Gaining a better understanding of the customer base, like the clients' typical behaviours and preferences;
- Discovering interesting patterns that may indicate tendencies or seasonality;
- Determining the key factors that contribute to a higher likelihood of cancellations;
- Understanding how to manage booking cancellations and present recommendations aligned with the company's objectives.

Some of the expected benefits of this project include:

- Improving net demand prediction: by getting valuable insights into the customer behaviour and booking/cancellation patterns, the prediction models can be refined to provide more accurate estimates.
- Minimizing the impact of cancellations on the hotel's revenue and overall business performance by developing and implementing the appropriate strategies;

- Optimizing resource allocation: by accurately predicting booking cancellations, hotel H2 can ensure more efficient operations and cost savings;
- Improving customer retention rates and satisfaction: by anticipating cancellations, the hotel can offer proactive assistance to customers and a more personalized service, such as rebookings or discounts. Moreover, the hotel will not need to apply restrictive cancellation policies, which customers generally do not like;
- Minimizing overbooking situations: by predicting cancellations, hotel H2 will be able to define better overbooking policies, which can eliminate issues like reallocation costs, reputation damage and loss of immediate and future revenue;
- Enhancing revenue diversification: by anticipating cancellations, the hotel can explore alternative revenue streams, such as offering last-minute deals or partnering with travel agencies to fill vacant rooms resulting from cancellations;

2.3. BUSINESS SUCCESS CRITERIA

In order to evaluate the success of the project, it is necessary to define concrete criteria that enable us to evaluate its progress and outcomes. Therefore, we established the following business success criteria to be assessed by the relevant stakeholders:

- Reduce cancellations to a rate of 20%, which is the Revenue Manager Director's main goal;
- Increase total revenue by 4% compared to the previous year;
- Decrease customer acquisition costs by 3%;
- Increase customer retention rate by 1.5% (considering the hotel industry in general has low customer loyalty);
- Increase occupancy rate by 5%;
- Develop a scalable solution that can be applied to other hotels of chain C.

2.4. SITUATION ASSESSMENT

To develop this project, we used the file "Case4_H2.csv", a dataset with 31 features/columns and 79330 rows provided by hotel chain C. The data corresponds to bookings that were due to arrive to hotel H2 between July 1, 2015 and August 31, 2017.

To prepare the data for analysis, we resorted to software tools such as Python and some of its libraries: *Pandas*, *NumPy*, *Scikit-learn*, *ydata_profiling*, *SciPy*, *datetime*, *calendar*, *math* and *pycountry_convert* and *boruta*. For data visualization, which is essential to communicate insights and present results from our exploratory analysis, we also used libraries like *Seaborn* and *Matplotlib*. Finally, for the modeling phase of the project, we used libraries such as *xgboost*, *shap*, *catboost*, *lightgbm*, and *bayes_opt*.

Besides the given dataset, we added new sources of information as we believed it would be an important resource to get a better understanding of the data and to achieve greater results. Firstly, we added information regarding the holidays and relevant events that took place in Lisbon by creating two binary features: *IsHoliday* and *IsEvent*. Secondly, we imported one new dataset: the file "Lisbon 2015-07-01 to 2017-08-31.csv" is a dataset related to the atmospheric conditions in Lisbon, including daily information regarding the time period of the initial dataset. It consists of 33 columns/features. We removed some of the columns because they did not add useful information to the data that we already have or because they contained missing values. Then, the remaining columns were added to the original dataset, leaving us with five additional features: *Temp*, *Precip*, *Windspeed*, *Cloudcover* and *Visibility*.

Thus, overall, the original dataset ended up with seven additional columns. The information about all the features is present in Table 1.

Besides resource assessment, it is important to mention the multiple assumptions that we made about the provided dataset:

- The feature *ADR* (average daily rate) does not change if the assigned room differs from the booked room;
- The waiting list is for when the hotel is full and does not want to do more overbookings;
- The waiting list does not have a limit;
- A customer could have cancelled and no refund was made, meaning the customer had not made any payments before canceling;
- Regarding the types of rooms provided in the dataset, room type B is more expensive than room type A, C is more expensive than B and so on;
- If a customer has a previous booking, even if canceled, they should be considered a repeated guest (as confirmed by the stakeholders);
- There was a delay in the opening of the hotel that affected the bookings for which the booking date was 2014-10-17 and the values of reservation status until 2015-07-02. As advised by the stakeholders, these rows were removed;
- A no-show is considered a cancellation if the customer informs the hotel of the reason why.

During the realization of the project, some challenges may arise due to limitations in resources, such as computing power and time. It is also important to consider potential risks, as well as take measures to minimize their impact:

- Data quality issues: The data might contain missing values, incorrect values or inconsistencies that can affect our analysis. To minimize this problem, we resorted to data cleaning techniques, ensuring that the final data was clean, complete, and accurate.
- Unrepresentative samples: The dataset provided might not be representative of the entire population of the hotel's customers, since it only represents a limited period of time (from July 1, 2015 to August 31, 2017), which can limit the generalizability of the results. To help this problem, it is good to make sure the data is diverse and representative of the customer base.
- Problems with the implementation of some algorithm: In order to mitigate this risk, we made sure our data was carefully prepared for all algorithms to work without problems.
- Lack of domain knowledge: Without a good understanding of the hospitality industry and the specific context of this business, we could miss important patterns and make incorrect assumptions. To solve this issue, we conducted a thorough additional research to build some domain knowledge and gain a better understanding of the industry trends and factors that could impact the accuracy of our models.
- The insights and recommendations generated might not be accepted or adopted by the hotel's team: To ensure this does not happen, our findings should be communicated with clear and detailed explanations in a way that is easily understandable and actionable for the hotel's stakeholders.

Finally, this project may involve several direct and indirect costs, such as acquiring and processing the data, computing resources, hiring or training the hotel staff, and redefining and implementing business strategies. On the other hand, there are also potential benefits that the hotel can expect, like increasing its revenue, implementing better pricing and overbooking policies and improving its operational efficiency. Additionally, the insights gained can help the hotel to better understand its customers and tailor its services to them, consequently preventing future cancellations and possibly increasing customer retention.

2.5. DETERMINE MACHINE LEARNING GOALS

With the aim of evaluating our booking cancellations predictions, it is important to determine the data mining results we are expecting, meaning what are the technical objectives of this project.

Considering the already mentioned business objectives, our general goal, and the available information provided, we have defined the following data mining goals:

- Minimum Recall of 0.50: The Recall can often be called the true-positive rate (Rectenwald, 2019). Thus, by being able to correctly identify at least 50% of the cancellations, we will be able to prevent them, in theory. Therefore, we would reduce the cancellation rate by half (from 40% to 20%).
- Minimum F1-score of 0.70 on the test dataset: The F1-score is a commonly used metric for binary classification problems and considers both precision and recall, allowing us to evaluate the model's ability to correctly identify both the positive and negative cases. The F1-score measures the harmonic mean of precision and recall while also not taking into account true negatives, so it may be a more accurate measure of the model's performance (Rectenwald, 2019).
- Use multiple forms of performance evaluation: Although the main metric that we will use is the F1-score, our goal is to also use other forms of performance evaluation, such as precision and recall. There are multiple metrics for evaluating a classification model and each of them evaluates it in a different way (Bajaj, 2023), so it is important to use a range of metrics to assess the quality of our model.
- Avoid overfitting: The generalization of our models to new data is ultimately what allows us to use them to make predictions (IBM, n.d.), therefore, it is important that the models we train don't fit too closely to the training set, allowing them to perform accurately against unseen data.
- Test various machine learning models: According to The MathWorks (n.d.), "there is no cut-and-dried flowchart that can be used to determine which model you should use without grossly oversimplifying the considerations". Thus, we wish to take advantage of the several methods available and evaluate the performance of multiple classification models to identify the most effective ones for the case in question.
- Optimize the parameters of the selected models: Once we have selected the most effective models, our goal is to fine-tune their parameters for performance optimization. This will involve testing different combinations of parameters and evaluating the resulting performance metrics to find the optimal settings.
- Build a robust and scalable classification system: Outside the scope of this project, our goal is to build a prediction system that can handle large volumes of data and changing market conditions. Scalability is about handling huge amounts of data and performing a lot of computations in a cost-effective and time-saving way, and it includes benefits like productivity, portability, and cost reduction (Kansal, 2020).

3. METHODOLOGY

To present a solution for the proposed problem, we followed the CRISP-DM methodology, consisting of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The first phase (developed previously in section 2) involves defining the problem and project objectives. In the second phase, data is collected and analyzed to gain a better understanding of the available information. The third phase involves cleaning, transforming, and integrating the data to prepare it for modeling. In the modeling phase, various techniques can be used to build and test models. The fifth phase is evaluation, which involves assessing the quality and effectiveness of the respective models. Finally, in the deployment phase, the chosen model is put into action and monitored for performance. It is important to note that the sequence of these phases is not rigid, and we moved back and forth between steps when necessary.

3.1. DATA UNDERSTANDING

As mentioned before, we uploaded new information to get a better understanding of the business and the customers, which culminated in having seven additional features in our dataset. We also created a new feature called *ArrivalDate* to better analyze the data. After these additions, we began to perform an initial exploration of the available information and to try to identify data patterns, structural problems and draw early conclusions. We started by becoming familiar with the data and identifying the existing features, which can be seen in Table 1.

Afterwards, we used descriptive statistics and data visualization tools (like *pandas profiling*) to generate an overall report of the data and discover inconsistencies and potential insights, such as:

- The target variable *IsCanceled* has two unique values – 0 and 1 – as expected;
- The dataset is nearly balanced, with approximately 42% of cancellations;
- The variable *LeadTime* does not have values below 0, indicating that all bookings were made in advance, as expected;
- The variable *ArrivalDateYear* goes from 2015 to 2017, as referenced in the materials given to support this case;
- The variable *ArrivalDateWeekNumber* varies from 1 to 53, which seems reasonable considering 52×7=364. Therefore, a customer can arrive on the 53rd week corresponding to the 365th day of the year;
- *ArrivalDateDayOfMonth* varies from 1 to 31, as expected;
- There are bookings with no adults (the feature *Adults* has a value of 0);
- There are bookings with no people associated to them (the features *Adults*, *Children* and *Babies* are all 0), however, this was confirmed to be possible by the stakeholders, since it might be created for zero nights to take care of financial reversals;
- There are bookings with excessively high values for the feature *Babies*, which might be due to mini-group reservations that were not cancelled and were divided into "normal" reservations;
- The hotel has customers from 166 different countries;
- There are two bookings made from Antarctica;
- There are 1208 rows with an *ADR* of 0, indicating that the price was not defined, which could be due to an offer or a promotion;
- There are 75641 rows with null values in the variable *Company*, indicating that there was no company associated with those bookings;
- There are 8131 rows with null values in the variable *Agent*, indicating that there was no travel agent associated with those bookings;
- The variables *MarketSegment* and *DistributionChannel* have an Undefined category, which was not considered as missing values since this classification simply indicates that the booking does not fit into any of the other categories;
- The variable *AssignedRoomType* had one more category than *ReservedRoomType*, which was denoted by the letter “K”. Approximately 0.3% of the bookings were associated with this category;
- The possible categories of the variables *DepositType*, *CustomerType*, and *ReservationStatus* are coherent with the materials given to support this case;
- There are 25902 lines with the same values for every variable, but it was guaranteed that there were not any duplicate bookings;
- There are bookings where the check-out coincides with the arrival day, which is reasonable, considering that customers might arrive at the hotel and, for instance, because they do not like it or have some urgent situation, they check-out on the same day;

- There are customers that had stayed at the hotel before and were wrongly considered as a not repeated guest;
- There are customers that had previously cancelled a booking and were not considered as a repeated guest, which is incorrect, considering the stakeholders have indicated that if a customer has a previous booking, even if canceled, they are considered a repeated guest;
- There are bookings that were not cancelled, however, the guest stayed zero nights at the hotel, which is valid, considering that they can stay only for one day;
- The feature *MarketSegment* can be misleading considering that not every collaborator may classify the client in the same way. Nevertheless, it can bring important information to the model;
- There are bookings that have been cancelled but the status of the reservation appears as "No-Show". However, it has been confirmed by the stakeholders that these can be interpreted as cancellations;
- There is an unreasonable peak in the number of bookings made on 2014-10-17. The stakeholders have indicated that the hotel's opening was delayed, explaining this pattern.

FEATURE	DESCRIPTION
IsCanceled	Value indicating if the booking was canceled (1) or not (0)
LeadTime	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
ArrivalDateYear	Year of the arrival date
ArrivalDateMonth	Month of arrival date with 12 categories: "January" to "December"
ArrivalDateWeekNumber	Week number of the arrival date
ArrivalDateDayOfMonth	Day of the month of the arrival date
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
Adults	Number of adults
Children	Number of children
Babies	Number of babies
Meal	Type of meal booked, assuming one of four categories: Undefined/SC, BB, HB, FB
Country	Country of origin (represented in the ISO 3155-3:2013 format)
MarketSegment	Market segment designation
DistributionChannel	Booking distribution channel
IsRepeatedGuest	Value indicating if the booking name was from a repeated guest (1) or not (0)
PreviousCancellations	Number of previous bookings that were cancelled by the customer prior to the current booking
PreviousBookingsNotCanceled	Number of previous bookings not cancelled by the customer prior to the current booking
ReservedRoomType	Code of room type reserved
AssignedRoomType	Code for the type of room assigned to the booking
BookingChanges	Number of changes/amendments made to the booking from the moment the booking was
DepositType	Indication if the customer made a deposit to guarantee the booking, assuming one of three
Agent	ID of the travel agency that made the booking
Company	ID of the company/entity that made the booking or is responsible for paying the booking
DaysInWaitingList	Number of days the booking was in the waiting list before it was confirmed to the customer
CustomerType	Type of booking, assuming one of four categories: Contract, Group, Transient, Transient-party
ADR	Average Daily Rate
RequiredCarParkingSpaces	Number of car parking spaces required by the customer
TotalOfSpecialRequests	Number of special requests made by the customer
ReservationStatus	Reservation last status, assuming one of three categories: Canceled, Check-Out, No-Show
ReservationStatusDate	Date at which the last status was set
ArrivalDate	Date of the arrival (in "YYYY-MM-DD" format)
IsHoliday	Value indicating if the day of the arrival is a holiday (1) or not (0)
IsEvent	Value indicating if the day of the arrival corresponds to an event (1) or not (0)
Temp	Mean temperature (Fahrenheit) on the day of the arrival
Precip	The amount of precipitation (mm) that fell or was predicted to fall on the day of the arrival
Windspeed	Average wind speed (mph) on the day of the arrival
Cloudcover	The amount of sky that is covered by clouds expressed as a percentage
Visibility	Distance than could be seen in daylight on the day of the arrival

Table 1 – Features/ Columns in the provided dataset (including the added data)

To better explore the data, we also created a feature called *BookingDate*, which indicates the day the booking was made. Overall, these observations provide important insights into the available data and will be useful in developing a model to address the customer churn problem.

3.2. DATA PREPARATION

The following actions are taken to ensure that the data is as clean and accurate as possible, which is crucial for the subsequent analysis and modeling.

3.2.1. Inconsistencies/Incoherences

Firstly, we removed the bookings where that had no people associated to them (where the features *Adults*, *Children* and *Babies* are all 0), since these bookings would not help in training the model. Secondly, we removed the bookings where there are more than 9 babies because we considered these as outliers that would not add value to training the model as well. Then, we also

removed the two bookings made from Antarctica, considering they are outliers from the country perspective and are very few. Afterwards, we altered the variable *IsRepeatedGuest* so that every customer that has had a booking before, even if cancelled, is considered a repeated guest. Finally, considering the stakeholders' provided information on the hotel's delay, we removed the bookings that were made on 2014-10-17 and the bookings that had a *ReservationStatusDate* after 2015-07-02.

3.2.2. Feature Creation

As stated before, we initially added holidays, important events and weather conditions. Additionally, to achieve better predictions, we created new features (presented in Table 2) based on the combination of the original ones. These new variables are possibly more relevant for the problem at hand and allow a more effective analysis.

FEATURE	DESCRIPTION
Continent	Continent of origin
RoomChanged	Value indicating if the customer changed from the room they were assigned to (1) or not (0)
ReservationStatusDateYear	Year at which the last status was set
ReservationStatusDateMonth	Month at which the last status was set
ReservationStatusDateDayOfMonth	Day of month at which the last status was set
ReservationStatusDateWeekNumber	Week number at which the last status was set
ReservationStatusDateDayOfWeek	Day of the week at which the last status was set
ArrivalDateDayOfWeek	Indicates the day of the week when a customer arrives at the hotel
BookingDateYear	Year in which the reservation was made
BookingDateMonth	Month in which the reservation was made
BookingDateDayOfMonth	Day of the month in which the reservation was made
BookingDateWeekNumber	Week number in which the reservation was made
BookingDateDayOfWeek	Day of the week in which the reservation was made
TotalNights	Total number of nights a guest stayed or booked to stay at the hotel
NumberOfPeople	Total number of people
ChildrenAndBabies	Total number of non adults
NoCancelRepeatedGuest	Value indicating if the customer is a repeated guest and has never cancelled a booking (1) or not (0)
Season	Season of the year assuming one of four categories: Winter, Autumn, Summer and Spring
CancelationRatio	Cancellation rate of the customer

Table 2 – New Features Created

3.2.3. Company's KPIs and Customer Analysis

In order to measure the current overall performance of hotel H2, we calculated the company's standard KPIs, such as the average daily rate (106.67€), the average revenue per booking (323.71€) and the average revenue per guest (165.89€). We also determined that the average length of stay (ALOS) of the hotel is 2.99 nights, with the average weekend nights being 0.8 and the average weeknights 2.19. Besides, the average number of guests per booking is 1.95 and the customers make bookings with an average lead time of 103.15 days, with the guest repeat rate being 5.65%. Moreover, the cancellation rate is 39.85%, with the percentage of cancellations being 38.65% and the percentage of no-shows 1.2%, whereas the percentage of check-outs/check-ins is, naturally, 60.15%. Finally, the average number of special requests per booking (for example twin bed or high floor) is 0.57.

Upon analysis of the data visualizations related to the hotel's customers base, several conclusions can be drawn. Firstly, the average daily rate (*ADR*) shows a gradual increase over the three-year period, indicating a positive trend in revenue generation (Figure 8.1). Moreover, the *ADR* tends to rise from December to April, remains relatively steady from May to August, and then declines in the following months (Figure 8.2). Notably, May marks the peak *ADR*, while January exhibits the lowest rate. Additionally, the analysis reveals that December, August, and February experience the highest cancellation rates, while January and March have the lowest (Figure 8.3). If we analyse the number of cancellations over time, we can observe a trend of growth from November until May, upon then there is a decreasing trend over the months (Figure 8.4). In terms of lead time, June and July have the longest duration, while March, April, and May exhibit the shortest (Figure 8.5). The remaining months fall close to the average lead time.

Considering the nationalities of our customer base, we can conclude that most bookings originate from Portuguese customers, followed by German, French, British, and Spanish clients (Figure 8.6). Specifically, around 36% of bookings are attributed to Portuguese customers, while 52% come

from European clients excluding Portugal, and the remaining bookings stem from outside Europe. Interestingly, Portuguese clients demonstrate a significantly higher cancellation rate compared to their actual utilization of bookings, whereas the other nationalities exhibit relatively lower cancellation rates (Figure 8.7). If we take a deeper look into the top nationalities that cancel most bookings, we will find a similar structure to that of the top nationalities that did not cancel their room, with the exception of the presence of Italy in the top five, instead of Germany (Figure 8.8). Furthermore, German, Portuguese, and British clients tend to book their rooms in advance with higher lead times than the average, while French and Spanish clients tend to book with shorter lead times, below the average (Figure 8.9). Although the average *ADR* per nationality generally aligns with the overall average for the top five nations, Portuguese clients, who constitute a considerable portion of the bookings, display an *ADR* below the average (Figure 8.10).

The analysis of the hotel's booking data reveals significant insights into the distribution channels, market segments, and reservation statuses. Most bookings originate from Travel Agents/Tour Operators (TA/TO), with minimal Global Distribution Systems (GDS) bookings. Online TA dominates the market segment, followed by offline TA/TO, while complementary and aviation channels contribute minimally. Within market segments, online TA generates the highest number of checkouts but also experiences the highest cancellation rate, while Groups and offline TA/TO sources witness substantial cancellations, with Groups surpassing checkouts in cancellations (Figure 8.11). Notably, online TA also contributes significantly to No-Show bookings. If we separate our booking data into those that have been cancelled and that have not, we can once again visualize that a similar structure arises for both groups, with the exception of Groups having a bigger proportion on the market segment of cancelled bookings (Figures 8.12, 8.13, 8.14 and 8.15). Analysing reservation status by the day of the week reveals that Fridays observe the peak of cancellations, gradually increasing from Monday, while Fridays have the lowest number of checkouts and Sundays show the highest (Figure 8.16). No-show bookings exhibit minimal fluctuations throughout the week.

Notably, "Event days" show a peak in ADR on 7/11, and further examination of the average ADR per event highlights the Web Summit, Lisbon Triathlon Championships, and Lisbon Fashion Week as the events with the highest number of average bookings (Figure 8.17), while the Portuguese Soccer Cup Final, Web Summit, and NOS Alive Music Festival generate the highest average *ADR* (Figure 8.18). We also analysed the weather data, which did not reveal any particularly interesting insights.

By analysing the *BookingDate* variable – which tells us when the booking was originally done - we exhibit a peak in January and gradually decrease until June, with a local peak in July, followed by a relatively stable period until December (Figure 8.19). Similarly, the weekday analysis shows that Mondays, Wednesdays, and Fridays are the most popular days for performing the booking of a room, while weekends see relatively fewer bookings (Figure 8.20). It is also observed that a significant number of cancellations are made by clients scheduled to arrive in March and November (Figure 8.21). Furthermore, most cancellations are made by clients who booked their reservations between September and November (Figure 8.22). When examining the date of cancellation, a notable trend emerges: the cancellation pattern closely resembles the booking date pattern, indicating that a considerable portion of cancellations occurs close to the original booking date (Figure 8.22 and Figure 8.23). This suggests that many cancellations are made near the date in which the booking would supposedly occur.

3.2.4. Data Imputation and Feature Engineering

There were missing values in the features *Children* (4 values), *ADR* (1 value), *NumberOfPeople* (4 values) and *ChildrenAndBabies* (4 values). Considering the missing values correspond to a maximum of 0.01% of the dataset, the variable *Children* was filled with the mode (0) and the *ADR* was filled with the mean. The remaining two variables were filled in accordance to the features used to create them. After that, the data types of these variables (except *ADR*) were changed to integers.

Afterwards, we used *OneHotEncoder* to transform the appropriate features (the ones which are nominal and the concept of distance between its categories does not make sense) into dummy variables, since it is more useful for the training of the models. For the feature *Agent*, we only one hot encoded the categories that were above 5%, while all other categories to were encoded as "Other". For the feature *Company*, we noticed more than 95% of the bookings do not have a company associated, so we transformed it into a binary feature indicating if there is a company associated to the booking (1) or not (0). Finally, for the feature *ArrivalDateMonth*, we simply transformed the month names into their respective numbers.

3.3. OUTLIER DETECTION

Since we no longer have any missing values, we proceeded to split the data into train (80%) and test (20%). It is essential to identify outliers, since "when a model is built, these special points can skew the model training and result in less accurate predictions" (Fernández, Bella, Dorronsoro, 2022, p. 1).

When preparing data, detecting and processing outliers is an important step to consider. To ensure the data is treated in the most beneficial manner, a particular treatment must be adopted. We tried six different ways to detect outliers: Manual Filtering (we defined filters based on visualizations), IQR Method (we defined an upper and lower bound using the IQR multiplied by 15, which was chosen based on visualizations and the percentage of data that was kept), Manual & IQR (we combined the two previous methods), Z-score (we removed all values that are 4 standard deviations apart from the mean), DBSCAN (we excluded all points that are not considered core or border, using an epsilon value of 0.73, which was determined using the Elbow Method and considering both visualizations and the percentage of data kept) and Isolation Forest (we defined the contamination as 0.04 as we intended to exclude 4% of the data). The IQR removed too much data (we 53.66%) because there are binary variables that have the value 0 for both the 25th and 75th quantiles.

We decided to proceed with the manual filtering method as our final choice after considering visualizations, the percentage of data removed and testing we have done in the modelling stage. With this method, we kept 96.91% of the original data.

3.3.1. Feature Selection

With the aim of not having irrelevant or redundant data, as it could negatively impact the performance of our models, we performed a thorough feature selection of the data.

The variables Country and Continent cannot be used for the modelling phase because, for the guests that have not checked-in, these values can be incorrect. Thus, we removed these features. Besides, there were variables that were removed because they leak information, since a lot of their values are only known after the period of cancellation. The removed features were: *ReservationStatusDateWeekNumber*, *RoomChanged*, *AssignedRoomType*, *ReservationStatusDate*, *ReservationStatusDateYear*, *ReservationStatusDateMonth*, *ReservationStatusDateDayOfMonth*, *ReservationStatus* and *ReservationStatusDateDayOfWeek*. Besides, the dates in the format "YYYY-MM-DD" cannot be used for modelling, thus *ArrivalDate* and *BookingDate* were also removed.

Afterwards, we started our feature selection by checking the Spearman correlations between all metric features. Then, we checked the correlations between non-metric and metric features, as well as just the non-metric features through a *pandas profiling* (since this tool provides the Kendall's and Cramer's correlations). This was done in order to make sure that, in the end, we didn't leave any highly correlated features in the dataset. Besides, we also checked the Spearman correlation between the numeric features and the target, where we noticed there were several variables with a correlation lower than 1%.

Then, we used a feature selection method called Boruta, which is an algorithm that is statistically grounded and works extremely well even without any specific input by the user (Mazzanti, 2020). This algorithm dates back to 2010 and was born as a package for R. Using Boruta, we defined estimators of both Random Forest and XGBoost. Since we obtained better results with Random Forest, we performed sequential feature selection using this estimator without the parameter max_depth and with a tolerance of 0.00000001 (this value was defined considering the results of the models). Having a forward or backward direction in the sequential feature selection did affect the results. Consequently, we removed 13 features.

3.4. MODELING

For the modeling phase of our project, it was of the utmost importance that we defined a scalable and efficient way to train, test and tune the hyperparameters of each of our potential models.

Overall, we tested twelve different models: Logistic Regression, Gaussian Naive Bayes, Instance Based Learning (KNN Classifier), Neural Networks (MLP Classifier), XGBoost Classifier, AdaBoost Classifier, Decision Tree Classifier, Random Forest Classifier, Support Vector Classifier (SVC), Bagging Classifier (using Decision Trees), CatBoost Classifier and LGBM Classifier. For evaluating the models, we were confident that we would reach the minimum recall previously defined (50%), thus, the F1-score was used as our evaluation metric, since it is more robust, as mentioned in our machine learning goals (section 2.5).

First, we trained each of these models using k-fold cross validation (with k=5) with and without outliers to see which models gave better results. For this, we created a function that returned a dataframe with the scores so we could quickly analyse them. Except for the SVC, all models provided better results when trained with outliers. For the models that required scaling the data, we used the MinMax Scaler and, in case any of those models were selected as one of the best, we would then test different scaling methods (Standard and Robust Scaler). Afterwards, we performed grid search on each of the twelve models (also with 5-fold cross validation) with a basic and broad hyperparameter tuning, simply to have an idea of the performance of each model. However, for the Gaussian Naive Bayes, we did not perform grid search since the initial results were already low, we could only use numeric features and there were not many parameters to test. Based on the results, we chose three models for further testing: LGBM Classifer, XGBoost Classifier and Decision Tree Classifer. Multiple algorithms gave similar results to these models, mainly tree-based methods. Nevertheless, LGBM and XGBoost were the best in terms of validation score and did not suffer from overfitting. Moreover, we wanted to test the Decision Tree due to the simplicity of this model and its advantages in terms of cost and computational requirements.

Afterwards, we started with the LGBM Classifier and verified if it was beneficial to add new features, one by one (by seeing if there was a significant increase in the validation score without suffering from overfitting). We ended up adding five features, which resulted in the increase of the F1-score by one percentual point. Then, we performed a Bayesian hyperparameter tuning, which is an informed search method, so we could achieve the best combination of parameters in a more efficient way. To make sure the results converged to a model with no overfitting, we used the following performance evaluation metric: (Validation F1-score) – 2 × (Train F1-score – Validation F1-score). For the remaining two models, we also tested with and without the additional features (only in the Decision Tree the features decreased the scores) and we used a similar strategy of hyperparameter tuning.

3.5. EVALUATION

As our final model, the LGBM Classifier was chosen, with the following parameter combination: {random_state=1, learning_rate=0.009565450056953601, max_depth=10, min_child_weight=2, n_estimators=528, subsample=0.9922734588192919}. With this model, we achieved a training and

validation score of 0.778 and 0.771, respectively. On the test dataset, we attained a F1-score of 0.773, a recall of 0.695, an accuracy of 0.837 and a precision of 0.871. The confusion matrix can be seen below (Figure 3.1).

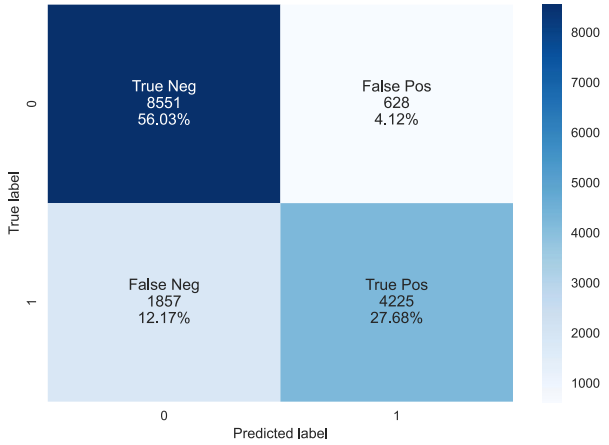

Figure 3.1 – Confusion Matrix

In Figure 3.2, we can see that the three features that have the most importance in the model are *DepositType_Non Refund* (binary variable indicating if the customer made a deposit in the value of the total stay cost), *TotalofSpecialRequests*, and *Agent_9* (one of the travel agencies that made the booking). Moreover, in Figure 3.3, we can analyse the patterns of some of the features. For example, a value of 1 in the feature *DepositType_Non Refund* is associated with more cancellations while the value 0 is associated with no cancellations (with a lower SHAP value). The feature *TotalofSpecialRequests* indicates that high values (more special requests) are associated with less cancellations and vice versa. We can also see that *Agent_9* having a value of 1 is associated with more cancellations and vice versa, meaning that customers who make bookings with this agency in particular tend to cancel more.

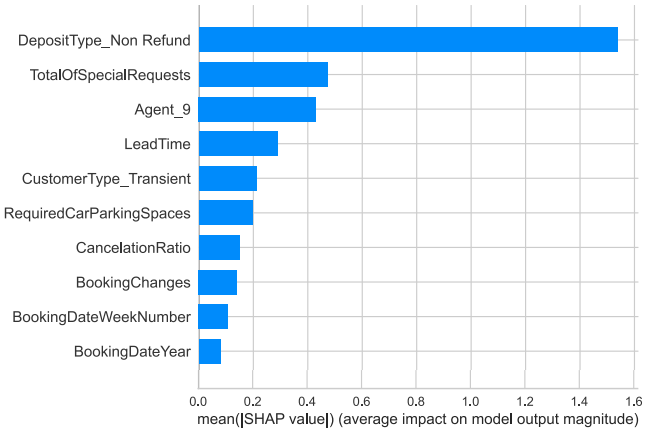

Figure 3.2 – SHAP Summary Plot 1

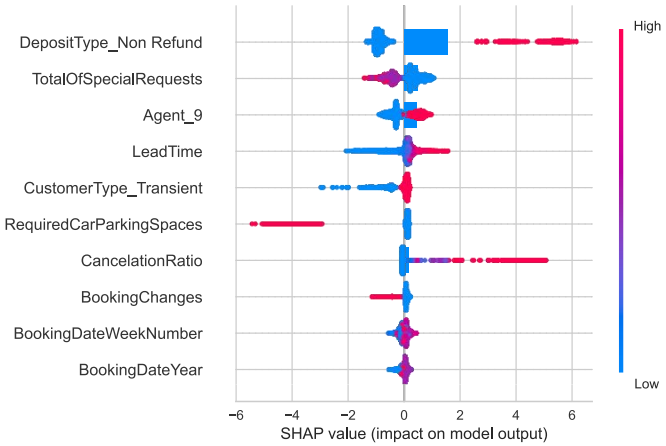

Figure 3.3 – SHAP Summary Plot 2

The defined machine learning goals (section 2.5) were achieved. Our model has a recall of almost 70% and a F1-score higher than 0.7. We were also able to avoid overfitting, utilize various forms of performance metrics and test multiple machine learning models, with optimized parameters. As for the scalability and robustness of our solution, "Light GBM can handle the large size of data and takes lower memory to run" (Banerjee, 2020), therefore, the goal of creating a model that can handle huge amounts of data and perform many computations in a cost-effective and time-saving way was achieved.

4. RESULTS EVALUATION

The strongest factor influencing the cancellation rate is the policy regarding the deposit (whether it is made on the value of the total stay cost). If the deposit is made, then the cancellation rate is much lower. If not, the cancellation rate is affected the other way around, although not as strongly. Even though it has negative consequences, as discussed, the data presents it as a fundamental part of reducing cancellations. The number of special requests is the second most important indicator (the higher the number the fewer the tendency to cancel and the same for the reverse), which could reflect that the customers' specific needs are being met. The third corresponds to reservations made through the agent called by the ID 9. If a booking is made through the Agent called ID 9 then the booking has a higher tendency of being canceled and the same the other way around. The fourth is the lead time and the data shows that the more in advance customers make a reservation, the greater the tendency to cancel (and the same for the reverse). This trend is compatible with industry metrics (D-EDGE, 2019). As the fifth, it was found that when the client booking is not part of a group or contract and is not associated to other transient booking it tends to be less cancelled, while the others cancel as much as the average bookings (could be due to different booking purposes or financial implications). The sixth indicator showed that the higher the number of requested places to park the car, the fewer the cancellations (still if the number is low then the prediction for the cancelation is not altered). This might happen because it reveals a greater need and effort to prepare the booking. Finally, the seventh revealed that the customers who have a higher cancellation ratio from previous bookings also tend to cancel more, as expected (again if the number is low then the booking is expected to cancel just like an average booking).

In consideration of these insights, the following recommendations were developed:

- Strengthen the development of packages where a deposit needs to be made in the value of the total stay cost, adding benefits such as complimentary breakfast or access to the hotel's fitness center or spa;
- Prepare customized promotions and offers to the clients, accounting for their unique needs and habits, to the bookings with a high likelihood of cancelling;
- Review the conditions and the agreements regarding the cancellation policy made with the Agency of ID 9;
- Reinforce the need of a deposit in the value of the total stay cost when it comes to bookings with a longer lead time;
- Promotional offers when a customer asks for a place to park his car and highlight the convenience, security, and cost-effectiveness of the parking options available;
- Implement a loyalty program that rewards customers based on their booking history and low cancellation ratios and advertise it to the bookings associated to customers with a high cancellation rate.

To better understand the costs associated to these recommendations, more information would have to be known regarding the current state of company and its revenue/cost structure.

5. DEPLOYMENT AND MAINTENANCE PLANS

Upon completion of the predictive model, the following deployment and maintenance plan has been defined, which is divided into eight different phases:

1. Firstly, the model and its results will be presented to the relevant stakeholders within the hotel, such as the executives, marketing, customer service, and financial teams, to ensure the needs of the company are being met. If not, we will go back and rework the model until the needs are met and the results are actionable. This phase will take about 1 week for completion.

2. Next, we will put the model into production, ensuring compatibility with the current technological infrastructure. The model must also undergo integration and system testing to ensure it is compatible with the other existing systems within the data ecosystem and the necessary changes are made. This process will take roughly 2 weeks to complete.
3. For an accelerated and smooth adoption, we will provide the necessary training and support to ensure the relevant employees can interpret the model's outputs effectively. We will present essential guidelines that enable the teams to make more informed decisions on customer service, marketing, and sales strategies. These will be regular since the workforce and the model change over time. These training sessions will be done every 3 months and will have the duration of 1 to 2 weeks, as long as there are interested participants.
4. Afterwards, new historical data will be incorporated to capture evolving patterns in cancellation behaviour, while reviewing and refining the model to optimize its performance and accuracy. This involves reviewing industry trends, customer feedback, economy's variability, and market dynamics to identify opportunities. It also entails exploring new data sources and variables to enhance prediction accuracy, collaborating with internal teams, data scientists, and industry experts to incorporate updated methodologies and assessing possible data/concept drift. As well as that, detailed reports will be generated highlighting the performance, efficiency, strengths, limitations, and impact of the predictive model on reducing cancellations during each update. Any potential issues that require further investigation will also be identified and reported. This will be reported to hotel H2's revenue managers, and other relevant stakeholders to provide valuable insights for decision-making and facilitate strategy adjustments. This phase will be done every three months.
5. Finally, the deployment and maintenance plan will be evaluated to ensure its cost-effectiveness. A thorough analysis of the expenses associated with acquiring and processing the necessary data, as well as updating the model, will be conducted. Simultaneously, the potential benefits of the model, such as reduced customer churn, improved customer satisfaction, and increased revenue, will be assessed. Based on this analysis, recommendations will be made to optimize the plan and ensure that the benefits outweigh the costs. Factors considered will also include the frequency of updates, data quantity and quality requirements, and the resources needed for maintenance and model updates. This should be analysed annually.

6. CONCLUSION

In conclusion, we were able to achieve most of our objectives and generate valuable insights to help hotel H2 understand its customer base and accurately predict their cancellations. These insights and predictions, coupled with the recommendations made to improve the business, will contribute to its growth and success in the long term.

Moreover, we have also outlined a comprehensive deployment and maintenance plan that aims to monitor the progress of the developed model, make new adjustments and updates and discover new insights periodically. We are confident that the insights and recommendations, along with the continual development of our model, will help differentiate the hotel from its competitors and reduce its cancellation rate to 20%.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

There are several ways to improve our model and the results we obtained:

- Incorporate data from additional years, which would help in the understanding of new patterns and tendencies over time;
- Explore additional data sources to enhance the model's predictive power;

- Create and/or improve automated data pipelines, for instance when it comes to pre-processing, which would help to speed-up and improve the current processes;
- Explore other machine learning models, such as Linear and Quadratic Discriminant Analysis;
- Explore and improve a broader set of the algorithms, both on the hyperparameters and the features used;
- Test alternative preprocessing and feature selection techniques;
- Consider the use of other libraries and softwares, such as Kedro and MLFlow, to create a more reproducible and modular data science code;
- Explore the feasibility of implementing a system for real-time model updates.

7. REFERENCES

Virtual Crossing Corporation (n.d.). *Weather Query Builder*. https://www.visualcrossing.com/weather/weather-data-services

Rectenwald, C. (2019). *Predicting Customer Churn*. https://medium.com/analytics-vidhya/predicting-customer-churn-58faa29bc836

Wu, S., Yau, W., Ong, T., Chong, S. (2021). *Integrated Churn Prediction and Customer Segmentation Framework for Telco Business*. https://ieeexplore.ieee.org/document/9406002

Bajaj, A. (2023). *Performance Metrics in Machine Learning [Complete Guide]*. Performance Metrics in Machine Learning [Complete Guide] - neptune.ai

IBM (n.d.). *What is overfitting?.* What is Overfitting? | IBM

The MathWorks, Inc. (n.d.). *Choosing the Best Machine Learning Classification Model and Avoiding Overfitting*. Choosing the Best Machine Learning Classification Model and Avoiding Overfitting - MATLAB & Simulink (mathworks.com)

Kansal, S. (2020). Machine Learning: Why Scaling Matters. Machine Learning: Why Scaling Matters (codementor.io)

Brown, T. (2021). *The 7 most important KPIs for hotel industry*. 7 most important hotel KPIs in the hospitality industry | Mews Blog

Fernández, Á., Bella, J., Dorronsoro, J. (2022). *Supervised outlier detection for classification and regression*. Supervised outlier detection for classification and regression - ScienceDirect

Mazzanti, S. (2020). *Boruta Explained Exactly How You Wished Someone Explained to You*. Boruta Explained Exactly How You Wished Someone Explained to You | by Samuele Mazzanti | Towards Data Science

Bannerjee, P. (2020). *LightGBM Classifier in Python*. LightGBM Classifier in Python | Kaggle

D-EDGE (2019). *How online hotel distribution is changing in Europe* https://www.d-edge.com/how-online-hotel-distribution-is-changing-in-europe/

8. APPENDIX

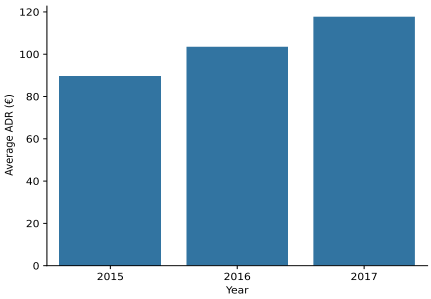

Figure 8.1 – Average ADR per Year

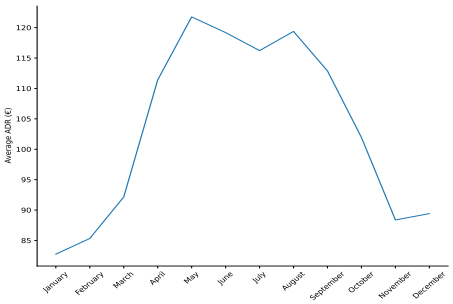

Figure 8.2 – Average ADR per Month

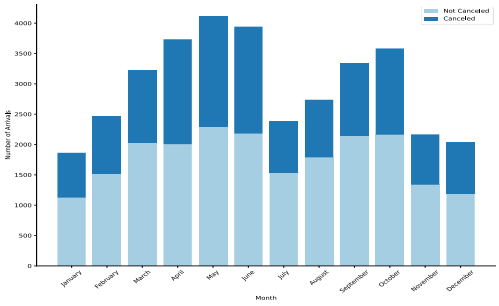

Figure 8.3 – Average Monthly Arrivals and Cancellations

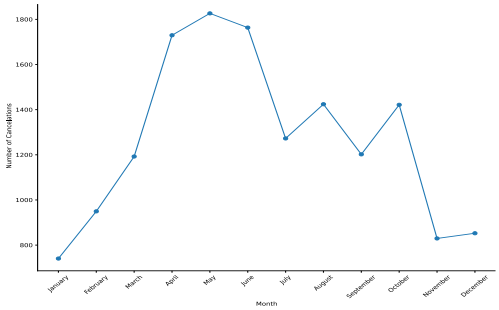

Figure 8.4 – Average Number of Cancellations per Month

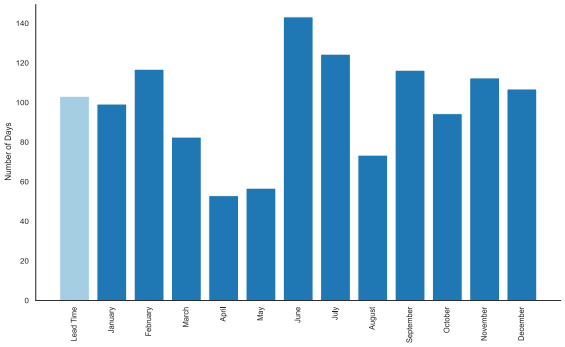

Figure 8.5 – Average Lead Time per Month

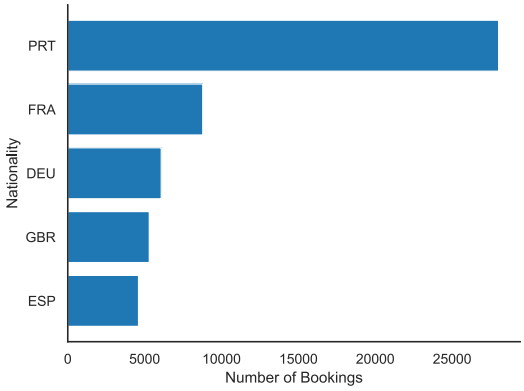

Figure 8.6 – Top 5 Nationalities

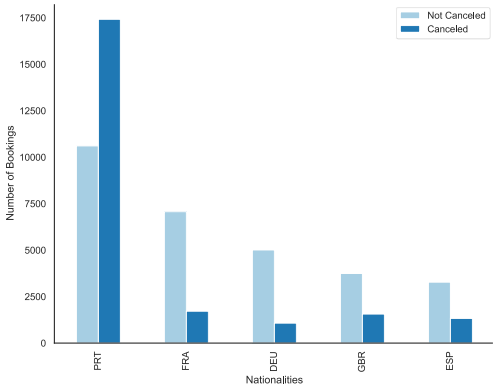

Figure 8.7 – Cancelled and Not Cancelled Bookings by Top 5 Nationalities

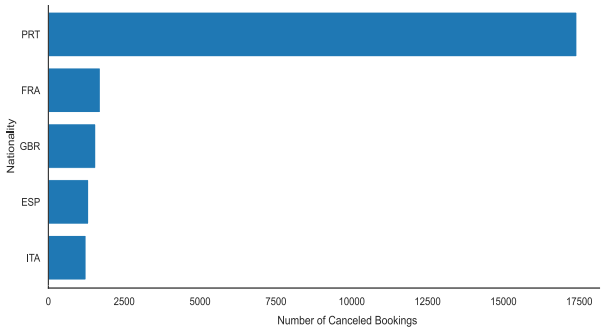

Figure 8.8 – Top 5 Nationalities: Cancelled Bookings

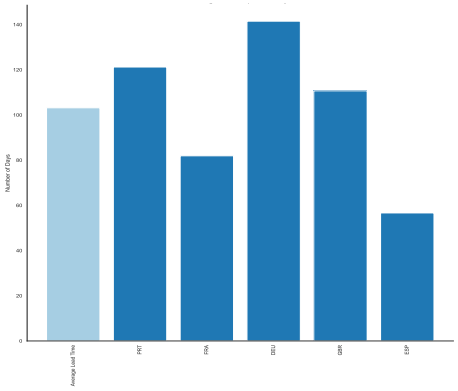

Figure 8.9 – Average Lead Time per Nationality

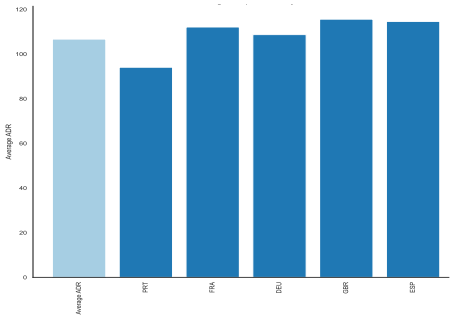

Figure 8.10 – Average ADR per Nationality

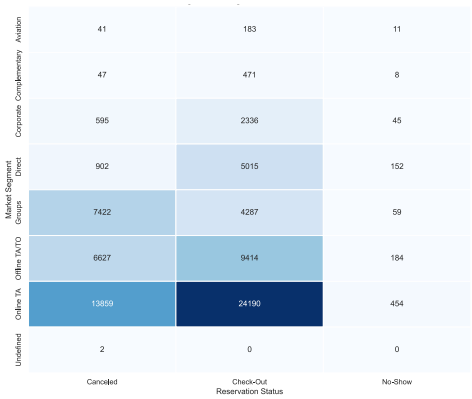

Figure 8.11 – Bookings: Market Segments to Reservation Status

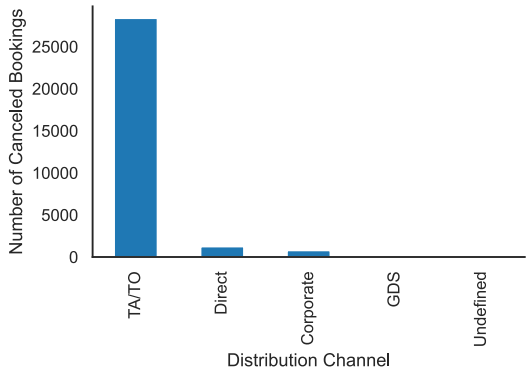

Figure 8.12 – Cancelled Bookings by Distribution Channel

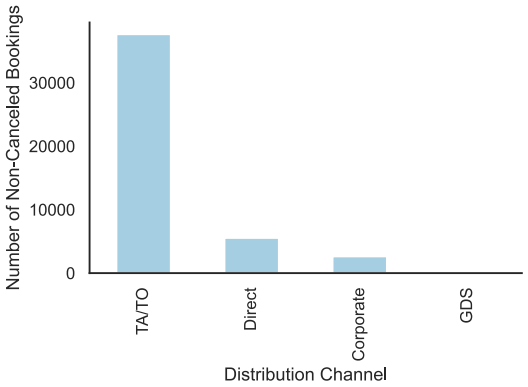

Figure 8.13 – Non-Cancelled Bookings by Distribution Channel

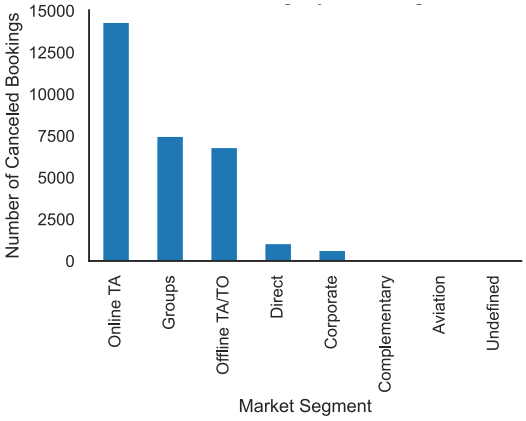

Figure 8.14 – Cancelled Bookings by Market Segments

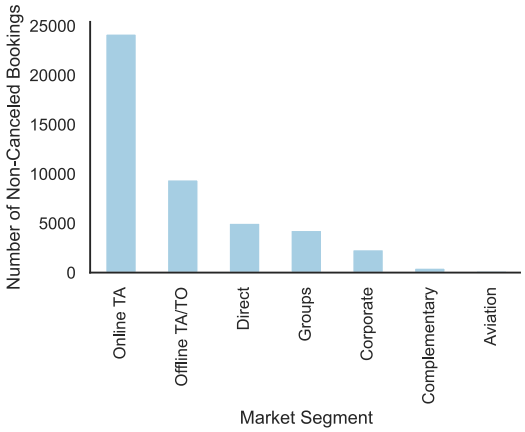

Figure 8.15 – Non-Cancelled Bookings by Market Segments

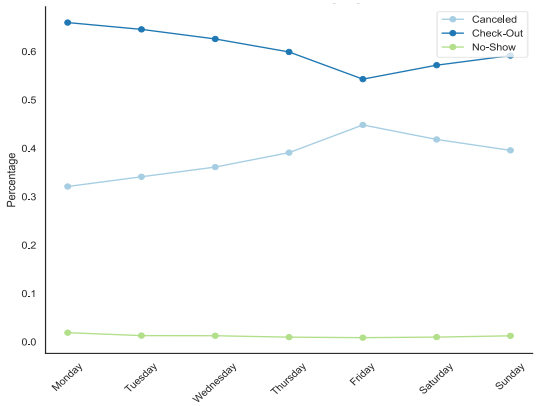

Figure 8.16 – Distribution of Reservation Status by Weekday

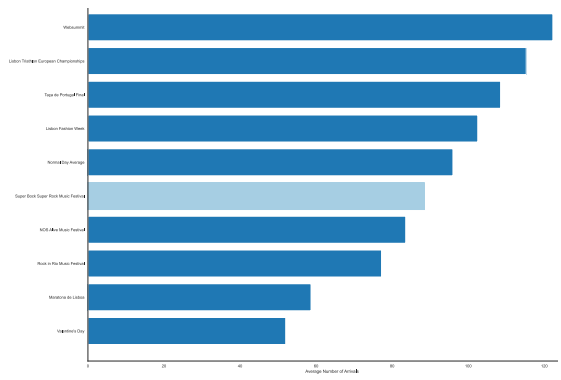

Figure 8.17 – Comparison of Average Arrivals on Normal Days and Events

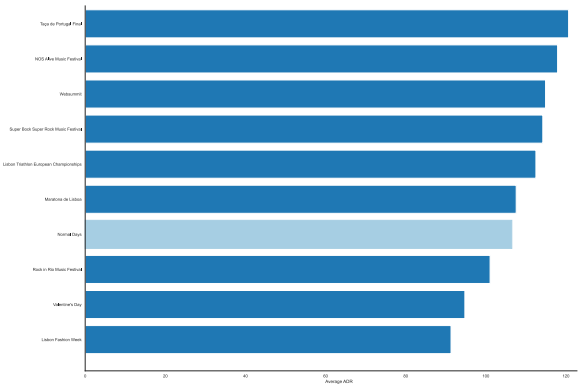

Figure 8.18 – Comparison of Average ADR on Normal Days and Events

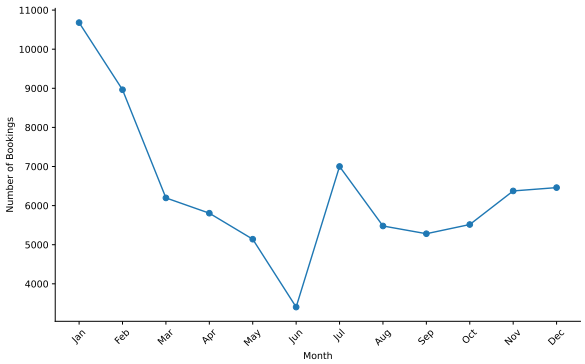

Figure 8.19 – Monthly Booking Count

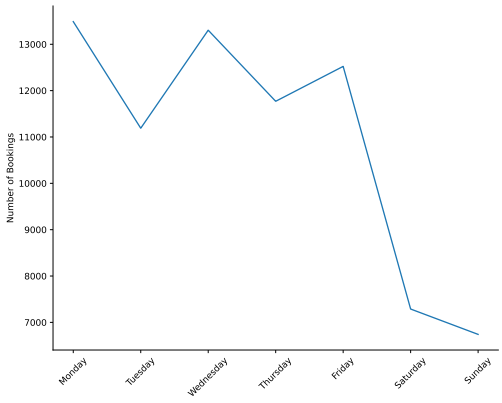

Figure 8.20 – Bookings by Weekday

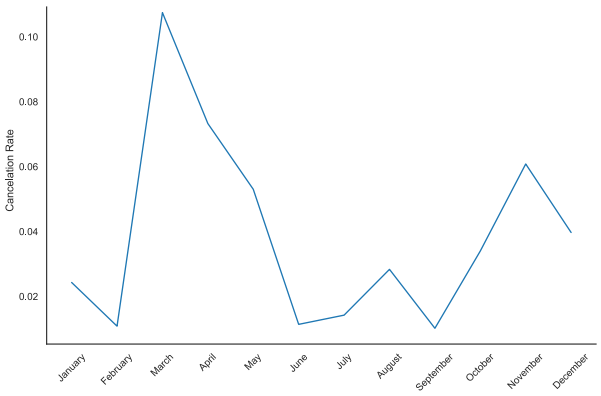

Figure 8.21 – Cancellation Rate by Month of Arrival Date

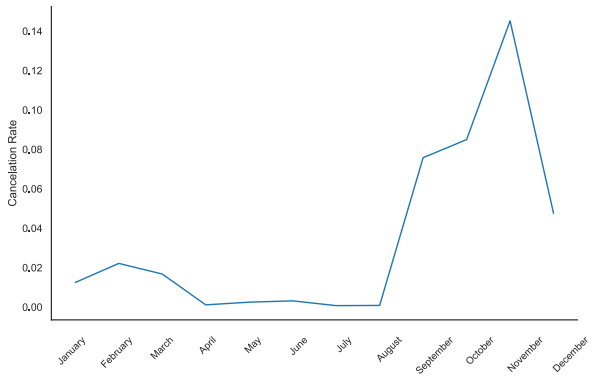

Figure 8.22 – Cancellation Rate by Month of Booking Date

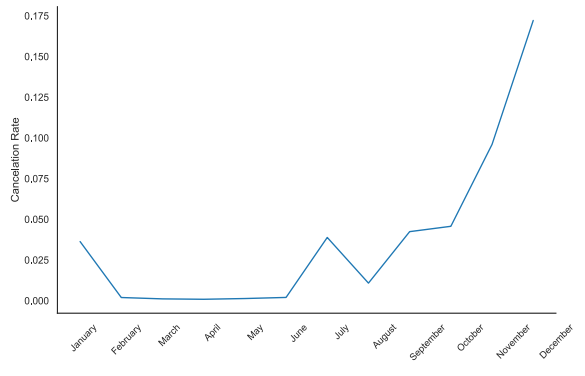

Figure 8.23 – Cancellation Rate by Month of Reservation Status