NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Business Cases with Data Science

Case 4: Predict Hotel Booking Cancelations

Carolina Costa, number: 20220715

Martim Santos, number: 20220540

Pedro Pereira, number: 20220684

Rodrigo Silva, number: 20221360

Rúben Serpa, number: 20221284

Group C

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

Universidade Nova de Lisboa

May, 2023

INDEX

1. EXECUTIVE SUMMARY

Through the power of data, the Young Talent Consulting Group made an understanding of the existing booking records and identified unique cancelation patterns, building a model to help Hotel H2 forecast net demand and take their business to the next level. The outputs, including recommendations and timelines, are provided in detail at the end of the report.

In sum, the booking data shared by Hotel H2 comprised a great foundation for the success of this project although a detailed understanding of the provided data was needed to correctly form a model that match Hotel H2 client profile. External data was also captured in order to relate certain reservation fluctuations or patterns that the data could present. Proper visualizations such as bar and line charts were also a key factor in understanding bookings patterns composition and various factors influencing its distribution.

By conducting a detailed analysis with the support of tree-based models and other machine learning models such as Logistic Regression and MLP classifier, a net demand forecast was created in accordance with the findings and a list of business applications were meticulously thought out based on outcomes.

It is believed that by implementing the given recommendations and with appropriate model maintenance to keep the efficiency of the given predictions, Hotel H2 will be able to reduce its cancellation rate with proper customer experience focus as well as pricing and overbooking policies together with a strong and accurate model.

2. BUSINESS NEEDS AND REQUIRED OUTCOME

2.1. INDUSTRY OVERVIEW

On an industry level, hospitality can be defined as the act of making someone feel welcome, usually through entertainment and comfort (*What Is the Hospitality Industry? Your Complete Guide | Cvent Blog*, n.d.). Hotels operate under this concept, where their main services fall under lodging and complementary services. As one of the Portuguese most important markets, it is important to reinforce the rapid evolution and change of structure that this market has been having. Statistically speaking, in 2021, one year after the pandemic, a year with huge sales breakout, hotel overnight stays rose from 14.8 to 21.9 million overnight stays equivalent to a total increase of 47.8% (*Portal Do INE*, n.d.). Of the total overnight stays, 5.3 million (24.2%) were based in the Lisbon metropolitan area and of these, 47% were based on four-star hotels, like Hotel H2.

In recent years, digital advancements led to the creation of online travel agencies (OTAs) and started to completely reformulate hotel sell channels. These new businesses that focus on offer display throughout digital channels enable the customer to have a more user-friendly concise knowledge of the overall product offering, making more thoughtful choices based on a comparison perspective. This adds another whole level of complexity to the market because now not only do hotels have to beware of demand fluxes, but also have to take into account the complexity that converting a customer into the business has. Fierce competition, OTAs influence on online selling, and demand fluctuations are just examples of some forces playing in this market.

The success of a business within this market starts with understanding the impact of such forces mainly through cancellations. The 'new' customer, adapted to the sudden changing and evolving environment that is shaping society and the pool of information easily available to him, adopted a loyal-less, temporary buyer behavior. Adding this to the OTAs influence that ends up pressuring the market, the flexible cancellation reservations seem to be at trend. Cancellation costs include loss of potential income, meaning unsold rooms. It gets worse when they are cancelled close to the check-in since it does not give the hotel time to deal with the situation. A no-show is a cancellation with no notice (*Cancellations Shooting up: Implications, Costs and How to Reduce Them |*, n.d.). These are some of the worst cases since these happen at the day of the check-in it makes the hotel lose revenue for at least the first night. Cancellations also imply lower Revenue Per Available Room (RevPAR) due to the necessity of selling the room in a short time. Globally speaking, cancellation rates via online channels reached an average of 39.6% in Europe, in 2018. According to D-Edge Hospitality Solutions, from 2014, this value has been on a rise for every online distribution channel (Figure 1), including OTAs and wholesalers. D-Edge believes that guests have become accustomed to free cancellation policies that have been made popular (and encouraged) mainly by Booking.com and channels and apps such as Tingo or Service, designed to cancel and rebook hotel rooms at each rate drop ("Global Cancellation Rate of Hotel Reservations Reaches 40% on Average," 2019). One other key fact that is vital for understanding customer behavior is that bookings are 65% more likely to be cancelled when booked 60 or more days in advance.

CANCELLATION RATE BY RESERVATION VALUE

Percentage of on-the-books revenue cancelled before arrival in Europe

	2014	2015	2016	2017	2018	Change
Booking Group	43.4%	43.8%	48.2%	50.9%	49.8%	6.4
Expedia Group	20.0%	25.0%	25.8%	24.7%	26.1%	6.1
Hotelbeds Group	33.2%	37.8%	40.3%	38.3%	37.6%	4.4
HRS Group	58.5%	51.7%	55.2%	59.4%	66.0%	7.5
Other OTAs	13.7%	15.2%	27.0%	24.4%	24.3%	10.6
Other Wholesalers	31.2%	30.3%	34.6%	33.8%	32.8%	1.6
Website Direct	15.4%	17.7%	18.0%	18.4%	18.2%	2.8
AVERAGE	**32.5%**	**34.8%**	**39.6%**	**41.3%**	**39.6%**	**7.1**

Yearly average percentage of on-the-books revenue cancelled prior to guest arrival from a sample of 680 D-EDGE clients in Europe.

D-EDGE, Hospitality Solutions www.d-edge.com

Figure 1- Europe's Cancellation rates by online distribution channels

2.2. Business Objectives

The hotel industry is highly competitive, with various forces sustaining this including the appearance of Online Travel Agencies (OTAs). With this, hotels become even more sensitive to competition having to take measures to resolve this. Measures become difficult to determine because even though they contain one problem, they generally generate others, so leveraging measures is key for the business.

Based on this, A, the Revenue Manager Director of hotel chain C, hopes to prevent future cancelation rates from affecting their business performance by utilizing past reservations data. To do so, the company intends to gain data-driven insights from a net demand prediction to achieve the following business objectives:

- **Reduce uncertainty about demand**: Demand uncertainty strongly affects the overall performance of a hotel. By not being able to know beforehand an accurate number of guests

for a specific period of time, the business ends up having larger revenue fluctuations and more avoidable losses while also not being able to correctly allocate budgets.

- **Implement better pricing and overbooking policies**: By successfully achieving the first objective, the business then has the potential to set appropriate pricing and overbooking strategies according to demand, setting a right balance that allows the business to optimize revenue and occupancy rates.
- **Identify bookings with a high likelihood of canceling**: Nowadays customers have access to more information and diversified platforms. From a hotel perspective this means more power for the customer. For example, an informed customer can utilize free cancellation reservations to guarantee temporarily a room while they take time to search for other options or wait for sudden offers or discounts from competitors. It is then vital for the company to understand which reservations could be from these customers or generally speaking more likely to be cancelled.

2.3. Business Success Criteria

Business success is going to depend on the ability to **reduce cancellations from the current 42% to 20%**, in a data-driven and concise way. The Young Talent Consulting Group has defined the following objectives with the purpose of achieving success:

- Primarily, the focus is on generating a **prediction model** based on reservations on-the-books. By comprehending certain characteristics and behaviors of the data, customizing it accordingly, and utilizing appropriate machine learning algorithms and techniques, the success comes down to the ability of the model to **accurately forecast net demand**.
- In order to implement better pricing and overbooking policies, the hotel should use the information obtained from the model to restructure policies. For example, after the model was created, the hotel could apply the concept of **dynamic pricing**. Dynamic pricing is a strategy in which the price of a particular product tends to change as per the ongoing customers' demand and supply (Fuchs, 2022). By using the model to dynamically adjust rooms rates in real-time, the hotel can, besides optimizing revenue, prevent cancellation rates by offering appropriate promotions and price offers to customers.
- By having the ability to identify bookings that are likely to be cancelled, the business can better manage inventory and operational efforts towards lower occupancy rates or leverage overbooking policies to not overextend them. In a customer experience perspective, the hotel could benefit from these insights to **reach out to uncertain bookings** and make offers to try to prevent cancellations and closely monitor them.

2.4. Situation Assessment

For the project realization, various datasets were implemented including not only the one provided by the hotel but also five additional external datasets to help understand some possible patterns.

- Hotel dataset: This dataset contains information about bookings between July 1, 2015, and August 31, 2017. It contains 79.330 reservations and 31 variables.
- Weather dataset: It focuses on the minimum and maximum temperature registered in each day of analysis. It contains 5 variables throughout 793 days of analysis.

- Inflation dataset: This dataset contains inflation values of each year (2015, 2016 and 2017) for each country or region. It contains 5 variables and 255 different countries and regions.
- GDP per capita dataset: This dataset contains the GDP per capita values of each year (2015, 2016 and 2017) for each country or region. It contains 5 variables and 268 different countries and regions. Both GDP per capita and inflation datasets were retrieved by the World Bank.
- Events dataset: This dataset has 15 rows and 3 columns. Each row represents a major event happening in Lisbon with the interval of dates of that same event.
- Average Daily Rate (ADR) dataset: It contains information for each year of analysis with the ADR of hotels in Lisbon, by number of stars. The dataset has 3 rows and 5 columns. The data was retrieved by Statista.
- Holidays dataset: It contains all Portuguese holidays days between 2015 and 2017. The data was gathered from the Time and Data website (*Holidays and Observances in Portugal in 2017*).

2.5. Determine Data Mining Goals

This project is going to follow a sequence of phases, an industry known as CRISP-DM. To fulfil the proposed business objective of reducing cancellations to a 20% rate, 6 main technical checkpoints exist:

- **Prevent overfitting:** It is crucial in ensuring the ability of our models to generalize well to new and unseen data, enabling accurate predictions. It is imperative that the models we train do not excessively conform to the training set, as this would limit their effectiveness when faced with unfamiliar data.
- **Explore diverse machine learning models**: As there is no predefined step-by-step guide to determine the ideal model without oversimplifying the complexities involved. Hence, the aim is to leverage the multitude of available methods and assess the performance of numerous classification models.
- **Refine the parameters of the chosen models**: After identifying the most promising models, our objective is to optimize their performance by fine-tuning their parameters. This process entails experimenting with various parameter combinations and analyzing the resulting performance metrics to determine the most effective settings.
- **Achieve a minimum accuracy of 0.8**: Accuracy is defined as the proportion of reservations correctly identified as cancelled divided by all reservations, multiplied by 100. This is the simplest way of verifying the performance of a model.
- **Achieve a minimum Recall of 0.5**: It is crucial to correctly identify at least 50% of the cancellations. By doing so, it is possible theoretically to prevent a significant number of cancellations. Consequently, this would result in a reduction of the cancellation rate by half, from 40% to 20%.
- **Achieve a Minimum F1-score of 0.70**: The F1 score is also another metric that balances both recall and precision giving a more real-world accurate metric to work with. In sum, the F1 score is the harmonic mean of precision and recall. This means that a high F1 score represents both high recall and precision and a low F1 score the opposite being important to focus on the overall improvement of this metric over the other two alone.

3. METHODOLOGY

3.1. DATA UNDERSTANDING

In the project, several datasets were employed, contributing to a comprehensive and valuable analysis. Firstly, the client provided a dataset containing hotel demand data for a city hotel in Lisbon from 1st of July of 2015 to the 31st of August 2017. The integration of the external datasets with the hotel data was performed in the data preparation phase.

The main dataset contains **79,330** observations each representing a hotel booking and **31** variables describing the booking characteristics. The initial data acquired appears to provide already some essential information to satisfy the project requirements. Through data exploration, it was possible to extract preliminary insights and identify several key issues.

Firstly, **58.3%** are cancelled bookings and **97.4%** of the customers are new clients (Figure 2).

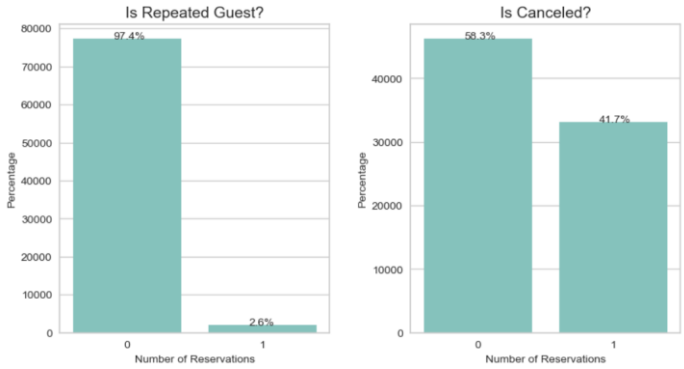

Figure 2- Class distribution of two variables

The majority of customers are from Portugal and from the Online TA Segment. However, Offline TA/TO and Groups are also common segments. Travel Agency 9 and Company 40 have a high number of reservations associated. Most reservations are made through a third-party entity. Furthermore, August is the peak month, Room Type A and D are the most assigned and reserved and in most reservations no deposit was made to guarantee the booking. A **typical customer** makes reservations 110 days in advance, for 1 or 2 nights in the hotel, with an Average Daily Rate (ADR) of 105€, and they frequently select the Bed Breakfast package. Special requirements and booking changes are not common. The bookings are composed, on average, of 2 adults, with the majority having no children or babies included. Second, some problems were identified in the data, namely **outliers** in the variables *LeadTime*, *ArrivalDateWeekNumber*, *Babies*, *Previous Cancelations*, *PreviousBookingsNotCanceled*, *DaysinWaitingList*, and *ADR*. Also, both *Children* and *Country* variables have missing values. There are 25,902 **repeated reservations**, however, different bookings can have the same characteristics, for instance, a group of employees going on a business trip. All these issues were further analysed and addressed in the data preparation stage.

Concerning the **business operations**, some conclusions were taken from the data. The high-demand seasons are Spring and Summer, while beginning and end of the year correspond to months with a smaller number of bookings. Regarding some relevant Hotel **KPIs**, the following information was gathered: Average Revenue per Booking: 298.55€; Average Revenue per Guest: 164.56€; Average

Length of Stay: 2.82 days; Average Lead Time: 98.45 days; Average Guests per Booking: 1.9 persons; Cancellation Rate: 41.53%; and Cancelation Lead Time: 84.79 days.

In terms of **customer demographics and room preferences**, it was possible to conclude that bookings with 1 and 2 adults usually prefer Room Type A, while 3 and 4 adults prefer Room Type D and G, respectively. Online TA and Direct Market Segments pay, on average, a higher ADR of 116.72€ and 114.89€, respectively. On the other hand, Groups and Corporate pay a lower ADR, possibly due to special offers. Additionally, it is relevant to understand how the behavior of different customer types differs in terms of **booking behavior**, namely, Group and Contract make reservations with less time in advance, and Contract customers ask for more requests on average. Additionally, more than 40% of contract and transient clients cancelled.

In what regards **cancelled and not cancelled reservations,** there are no differences between the ADR charged by the hotel, or in the length of stay, room type discrepancy and occurrence of events. Nevertheless, some important differences appear through the data analysis (Figure 3), namely, that 68% of groups clients and 43% of offline TA/TO Clients cancelled, Transient and Contract customers have similar number of cancelations and no cancelation, 99.9% of non-refundable deposit and 70% of refundable deposit reservations were cancelled, and Meal FB has more cancellations than no cancellations.

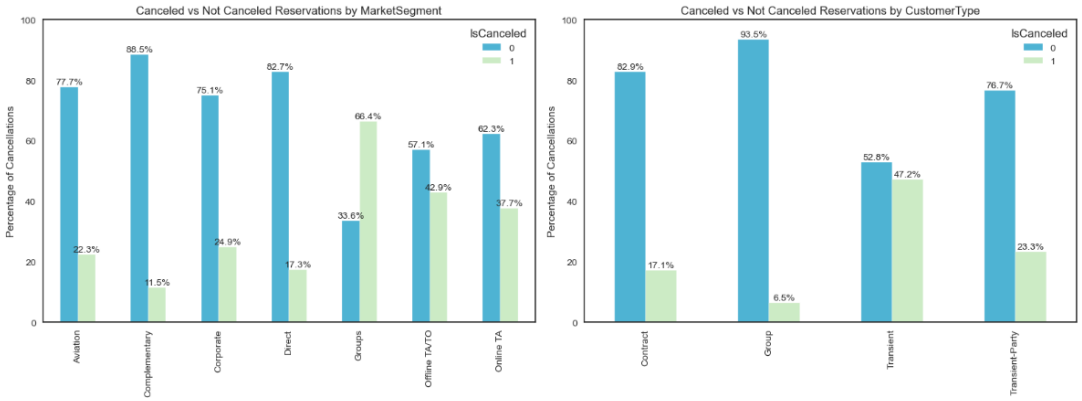

Figure 3- Cancelled vs Not Cancelled Reservations by Market Segment and Customer Type

3.2. DATA PREPARATION

3.2.1. Missing Values Treatment

In an initial search it was found two columns, '*Children*' and '*Country',* with 4 and 24, respectively, missing values. Since it only represents about **0.035%** of the data, it was decided to drop these rows instead of inputting values to not induce bias. The '*Company*' column was dropped as 95% of its data was missing, lacking relevant data for analysis. The 'NULL' categories for the 'Agent' column are not considered missing values but instead as reservations booked without an agent ID assigned, for example direct bookings. These cases the values are going to be substituted by 'Not Applicable'.

3.2.2. Duplicates Treatment

Regarding duplicates, there were **25,902** possible repeated reservation, which were considered valid, as are just normal reservations that were on the same day and have the same characteristics.

3.2.3. Inconsistencies

Data "inconsistency" causes problems, including a loss of information and results that are incorrect. Data consistency, on the other hand, promotes accuracy and the usability of available data and may be the difference between a business's success or its failure (Foote, 2022). Below are some of the possible inconsistencies checked that the dataset might have had.

1) No adults in the reservation (Adults = 0): Under this situation, three settings were analyzed. The first setting was reservations with at least one child and the second one with at least one baby (no children). Both settings make sense if parents have specifically booked the rooms for their children. The other setting represents customers that did not specify guest characteristics. There were **159** rows under this setting, which were dropped.

2) Unrealistic Lead Time: After creating the '*booking_date'* variable to understand the reservation date, it was verified that the first booking date (**17-10-2014**) contained 2509 bookings, an unusual number of bookings compared to the rest of the dates. Additionally, it was verified that 92.67% of these bookings were cancelled. This situation concerns an issue that arose during the opening of the hotel when reservations started on the opening day, instead of when the hotel was operational. This obliged the hotel to cancel almost all reservations and reallocate the other 7.33% to other places. Therefore, reservations booked on this date were removed to not influence the accuracy of the model.

3) Repeated Guests with PreviousBookingsNotCanceled/Previous Cancellations: A repeated guest is everyone that made a previous booking, even though that same booking was cancelled, or the person did not check-in. Firstly, there were **171** reservations where the guest had previous bookings (not cancelled) and the guest was categorized as 'not repeated'. Second, there were **2489** reservations where the guest had previous cancelled bookings and the guest was categorized as 'not repeated'. In these situations, the *IsRepeatedGuest* value will be changed and considered has previous guests.

4) Stay Nights: Focusing on stay nights, when '*StaysInWeekendNights'* is greater than 2 that means the customer stayed for at least the full week before or after the first weekend, meaning that '*StaysInWeekNights*' should be at least 5. All reservations respected this condition. On a second perspective, it was also verified bookings without stay nights, cancelled and non-cancelled, having **215** and **13** cases respectively. After confirming with the stakeholders, these bookings were kept as valid and considered day-use cases.

5) ArrivalDateWeekNumber: This column contains the week number of the reservation. After double-checking its values, it was found that **42,563** of its values were not accurate. Based on the correct week number intervals, the values were substituted.

6) Market Segment and Distribution Channel: There was an 'Undefined' category for both columns, four reservations had this category on Distribution Channel with two of them also having this category on market segment. Since these reservations had missing values in the column 'Children', these rows were removed in the previous phase. There were some doubts regarding the need for the two variables to have the same values for its common categories, like 'Direct', being this potentially an inconsistency. After talking to the stakeholders, it was concluded that they could happen to not be directly related so these cases were all considered valid.

7) Number of Babies: There were in total 2 cases with reservations with more than 8 babies, one with 9 and another with 10 babies. These were dropped from the dataset as it did not represent coherent values. Looking at the column value distribution, **99.52%** of the reservations did not have babies, with only 0.47% having one baby and approximately 0.01% having two babies. Due to very unbalanced distribution, the column was dropped.

8) Incorrect ISO Codes: Another possible inconsistency was around country ISO codes. After checking the ISO codes presented in the dataset there were two invalid ISO codes: TMP, CN. These codes were substituted by the correct format TLS for Timor-Leste and CHN for China.

9) Misspecified Countries: Two other ISO codes, ATA (Antarctica) and ATF (French Southern Territories) had a total of 2 and 1 bookings, respectively, that were removed due to not being relevant regions. For reservations that did not have previous checked-in stays and the reservation is cancelled, the country is not trustful and therefore the values will be replaced by 'Not Applicable' to not induce the analysis to misleading data.

3.2.4. Outliers Analysis

In the context of churn prediction, addressing outliers is crucial, therefore, various approaches were considered to handle outliers effectively. Prior to implementing any specific approach, a comprehensive analysis of the variable graphs was conducted to examine their distributions and identify potential outlier records, using Boxplots and Histograms. Different methods were applied to each metric feature were chosen for their ability to detect and handle outliers in diverse ways. Ultimately, the approach that demonstrated the best trade-off between retaining the majority of the records, thereby avoiding the loss of a significant percentage of values, and achieving a better normalized distribution was selected as the preferred method.

Using **Manual Detection**, the outliers are defined based on several criteria. Firstly, a reservation is considered an outlier if it involves less than one adult or more than three adults. Similarly, if a reservation includes more than one baby, it is also classified as an outlier. Additionally, if a reservation has experienced more than two booking changes or has more than two children, it is considered an outlier. Furthermore, we identified outliers based on the duration of being on the waiting list and lead time. A reservation is deemed an outlier if it has been on the waiting list for more than 105 days and has a lead time exceeding 338 days. Moreover, reservations with more than four bookings that were not canceled and more than two cancellations were also classified as outliers. In addition to these criteria, we considered certain assumptions as outliers. Reservations that requested more than one car parking space or had more than two special requests were regarded as outliers. Furthermore, clients who stayed for more than two nights during the week or more than three nights during the weekend were also considered atypical and treated as outliers. In addition to the aforementioned analysis, a separate examination of outliers within the ADR was conducted due to its variability across different room types. Consequently, specific thresholds were established for the ADR in conjunction with each respective room type. By applying this manual detection approach, we identified that 10,96% of the observations within the dataset met the criteria for outliers. Although the removal of these observations is justifiable for the continuation of the study, we decided to explore how the interquartile range (IQR) method could contribute to the identification of outliers, thereby minimizing the potential for human error

The **IQR method**, as described by Grant (2022), is a statistical approach used to measure the spread of data points within the middle 50% range around the mean. However, upon applying the IQR method to the dataset, it was observed that removing outliers using this technique would result in the elimination of more than approximately 51% of the observations, which is not sustainable since it would lead to a significant loss of data. Furthermore, the IQR method did not yield satisfactory results in comparison to the Manual approach employed earlier.

In order to enhance the outlier detection process, a combination of both **Manual and IQR methods** was attempted. This hybrid approach aimed to provide a more robust technique for outlier identification. Consequently, by applying this combined method, approximately 11% of the data was identified and removed as outliers. Although an 11% removal rate was acceptable, it was observed that the data did not exhibit a normal distribution as prominently as it did after applying the Manual Detection approach. These findings are further illustrated in subchapter 5.5 of the notebook.

In an attempt to further enhance the outlier detection process, **DBSCAN** was utilized and his resulted in the removal of approximately 16% of the data. Furthermore, it was observed that the distribution of the data after DBSCAN-based outlier removal was not as well-represented as that achieved through Manual Detection. This discrepancy suggests that DBSCAN may not be the most suitable method for outlier detection in the context of churn prediction in this particular dataset.

Among the various outlier detection methods explored in this study, **Isolation Forest** was considered for its effectiveness in handling skewed datasets and its efficiency in processing large datasets. The use of Isolation Forest aimed to identify and remove outliers while maintaining the integrity of the dataset. Furthermore, Isolation Forest was expected to have a minimal impact on the dataset, removing only a small portion of the data as it was confirmed that it removed less data, only about 7%. This removal rate was relatively low compared to other outlier detection methods previously explored. However, upon further analysis of the box plots and histograms, it became evident that the resulting data representation was not optimal.

Based on the comparison of the five outlier detection methods, it is evident that the Manual Detection approach yielded the most satisfactory results in terms of accurately identifying outliers while maintaining a reasonable proportion of the dataset. The predefined criteria and thresholds applied in Manual Detection allowed for a more precise identification of outliers that deviated from expected patterns. Still, the other methods were tested to make sure which method had the best validation result and it turned out that DBSCAN had the best impact.

3.2.5. Feature Engineering

According to the guidelines provided by Kelleher and Tierney (2018) for feature engineering, which involves combining, selecting, and transforming raw data to obtain new and more informative features, 26 new variables were extracted as shown by the following Table 1. In conformity with Lo Duca (2022), cyclicity can significantly enhance the performance of machine learning models when dealing with time-related features. In line with this principle, the Arrival Date features have been enriched with cyclicity to better capture and interpret the cyclical nature of temporal patterns in the dataset.

Variable	Description
Total Guests	The sum of the number of adults and any additional guests included in the reservation.
Total Nights	The sum of the number of nights the guest stayed at the hotel.
Total Revenue	Total price paid for accommodation.
Has Previous Stays	Represents whether a client has already stayed in the hotel.
Special Requests Ratio	The average number of special requests made by the customer per night of stay.

Room Type Discrepancy	Comparing the values of *AssignedRoomType* and *ReservedRoomType* to identify if there are any discrepancies between the room type reserved and the room type assigned.
Previous Cancellations Ratio	The ratio of previous bookings that were cancelled by the customer.
Agent Grouped	The percentage of bookings cancelled for each travel agency (*Agent*), grouped by the travel agency with the majority of reservations.
Assigned Room Type Grouped	The grouping of room types based on the number of reservations. The goal is to distinguish between rooms with a high number of reservations (room types 'A' and 'D') and rooms with fewer reservations.
Is No Deposit	Represents whether a customer made a deposit to guarantee the booking. It is a binary variable indicating if a deposit was not made (No Deposit) or made (Non Refund or Refundable).
Market Segment Grouped	The grouping of market segments based on customer behaviour and reservation frequency. It aims to categorize related segments and segments with few reservations into broader categories.
Distribution Channel Grouped	The grouping of distribution channels based on customer behaviour and reservation frequency. It aims to categorize different distribution channels into broader categories.
Booking Date	The date on which the booking was made, derived from the information of the arrival date and lead time.
Booking/Arrival Date: Weekday, month day, month, week number	To capture the cyclic nature of the weekdays, days of the months, months, week numbers of booking/arrival dates sine and cosine transformations have been applied.
Cancel Lead Time	The lead time between the reservation status date and the arrival date, specifically for cancelled bookings.
Total Revenue Normalized	Total price paid for accommodation normalized.
ADR Normalized	The normalized ADR for each booking in the dataset.
Customer Segment	Categorizes the customers based on the number of adults and children included in their reservation.

Table 1- Created variables description.

Following the appropriate data transformations, the modelling dataset has been extended to encompass a total of 80 variables.

3.2.6. Feature Selection

In the process of feature selection, firstly certain features were discarded that were deemed unfit for modelling, including variables like *'ReservationStatus'*, *'ReservationStatusDate'*, *'CancelLeadTime'*, *'RoomTypeDiscrepancy'*, *'AssignedRoomType'* that can contain future information not available at prediction time. The variables *'Inflation'*, *'Country'*, *'GDP per capita' were also removed, since the countries' information is not trustable before the arrival date.* The datetime features, ungrouped categories and non-normalized values (*ADR* and *Total Revenue*) were also removed from this segment.

Subsequently, the data was split into non-metric and metric features to adopt appropriate selection techniques for each. **Kendall Correlation Coefficient** was utilized as a filter method for metric features, which demonstrated small correlations between *'LeadTime', 'DailySpecialRequests', 'TotalOfSpecialRequests', 'RequiredCarParkingSpaces', 'BookingChanges', 'PreviousCancellations', 'TotalPreviousBookings', 'PreviousCancellationsRatio'* and the target variable 'IsCanceled'. To avoid redundancy, only *'PreviousCancellationsRatio'* was retained from the highly correlated trio of *'PreviousCancellations', 'TotalPreviousBookings', and 'PreviousCancellationsRatio'*. Consequently, the final metric features chosen for model construction are *'LeadTime', 'DailySpecialRequests', and 'PreviousCancellationsRatio', 'RequiredCarParkingSpaces', 'BookingChanges'*.

Regarding non-metric features, these were encoded, transforming the categorical variables into numerical values. The **Mutual Information (MI)** statistical method was first employed. This technique

gauges the dependency between the variables, providing insight into the shared information between the feature and the target variable.

The features '*DepositType_grouped'*, *'CustomerType_grouped'*, '*MarketSegment_grouped'*, *'DistributionChannel_grouped'* and '*BookingDate_weeknumber'* stood out with the highest mutual information scores, thus indicating a significant shared relationship with the target variable 'IsCanceled'. Conversely, certain features such as '*Meal'*, '*ArrivalDate_dayofmonth*' and '*IsEvent'* displayed a zero mutual information score, implying no shared information with the target variable.

Concerning Wrapper methods, Recursive Feature Elimination (RFE) method was also employed. It performs feature selection by gauging the model's performance based on a particular set of features. In this case, the Logistic Regression model was used as the estimator. The dataset was split into train and test. Moreover, the model was fitted using RFE with stipulated count of 10 features to be selected. RFE ranked the features according to their relevance. The selected features were accurate considering domain expertise, for instance, variables such as *'PreviousCancellationsRatio'*, *'TotalOfSpecialRequests'*, *'IsRepeatedGuest'*, and *'RequiredCarParkingSpaces'* can be expected to have significant impact on whether a booking gets cancelled or not.

The final features selected were: *'LeadTime'*, *'PreviousCancellationsRatio'*, *'TotalOfSpecialRequests'*, *'BookingDate_weeknumber'*, *'BookingDate_dayofmonth_sin'*, *'BookingDate_dayofmonth_cos'*, *'IsRepeatedGuest*', *'RequiredCarParkingSpaces*', *'CustomerType_grouped'*, *'DepositType_grouped'*, *'MarketSegment_grouped'*, *'DistributionChannel_grouped'*, *'BookingChanges'.*

3.3. MODELING

In this section of the report, the focus is on the modelling phase of the analysis, encompasses the preparation of the data, the encoding of categorical variables, and the application of various models.

3.3.1. Tree-based models

Tree-based models were employed due to their inherent feature selection capabilities and their ability to minimize the impact of non-important features. Consequently, explicit feature selection techniques were not applied, and variables deemed unsuitable for modelling were excluded. Additionally, since tree-based models do not accept categorical variables in their original text format, an encoding process was carried out. The *OneHotEncoder* was utilized to convert categorical variables into a binary format, which could be effectively utilized by machine learning algorithms to enhance prediction performance. Several models were implemented, specifically *DecisionTree, XGBoost, CatBoost, Random Forest, Gradient Boosting, Extra Trees Classifier,* and *LightGBM*. To ensure result reproducibility, all models were initialized with a consistent random state.

To assess the performance of the default models and establish a benchmark, the ***Stratified K-Fold cross-validation*** method was employed. This involved iterating over the default models and calculating the average training and validation scores for each model. These scores provided a baseline indication of the performance of each model before proceeding to the subsequent steps of hypertuning and performance evaluation. The Figure 4 contains the scores obtained for each model applied in this phase.

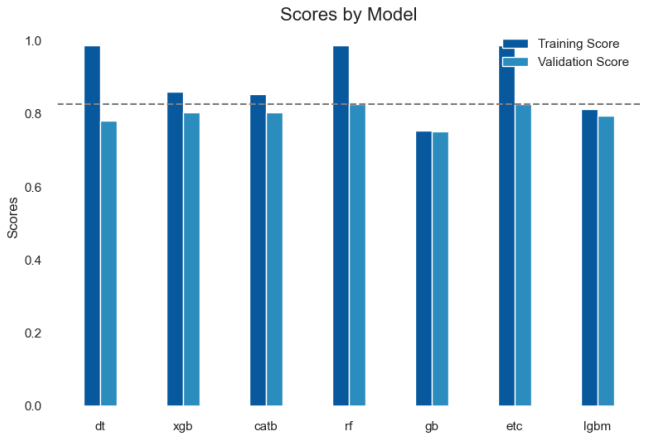

Figure 4- Tree based models' scores

3.3.2. Other models

This section extends the modelling analysis to incorporate alternative model types that are not based on tree-based algorithms. Unlike tree-based models, these models require attention to the selection of features, as employing all features does not yield satisfactory results. Therefore, only the features selected in the previous step were utilized. The initial step in preparing the data for these models involves encoding the categorical variables, following a similar approach as employed for tree-based models. After encoding the categorical variables using *OneHotEncoder*, it was performed feature scaling. Unlike tree-based models, other models can exhibit sensitivity to the scale of input features. Thus, to ensure equitable contribution of all features to the final decision function, scaling was applied to achieve a comparable range. The **MinMaxScaler** method was employed, which individually scales and translates each feature to fit within the specified range, typically between zero and one. Several models were employed in this analysis, namely Logistic Regression, K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), and Support Vector Machine (SVM).

Similar to the tree-based models, these models were initialized with a consistent random state to ensure reproducibility of results. In the case of KNN, the number of neighbors was set to 10. For MLP, the neural network architecture was defined with two neurons in the first hidden layer and one neuron in the second hidden layer. Subsequently, the average training and validation scores (Figure 5) were calculated for each model using the Stratified K-Fold cross-validation method.

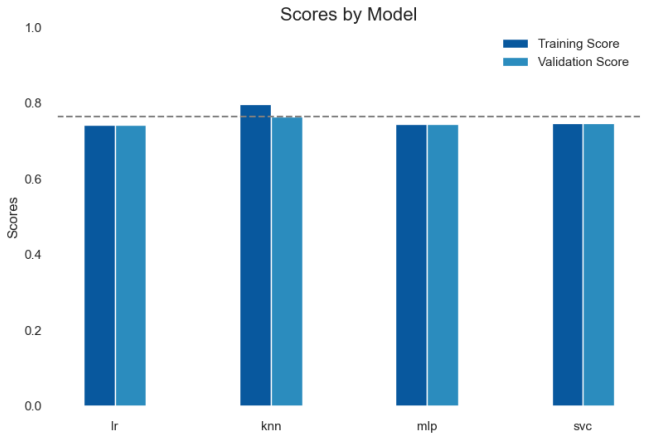

Figure 5- Other models' scores

3.4. FINAL CONSIDERATIONS ON THE TOP PERFORMING MODELS

After careful consideration, it has been determined that the two top performing models under default parameterized conditions are XGBoost and CatBoost, as both exhibit minimal overfitting and satisfactory validation scores. During the optimization phase, the hyperparameters of the XGBoost and CatBoost models were adjusted to enhance their performance. The XGBoost model that yielded the best results was configured with 30 estimators, a maximum depth of 5, and other optimized parameters. Likewise, the CatBoost model achieved optimal performance with 30 estimators, a maximum depth of 5, a learning rate of 0.4, and other optimized parameters. Upon comparing the training and validation scores, it was evident that the CatBoost model outperformed the XGBoost model by exhibiting equivalent validation performance while displaying reduced overfitting tendencies. Consequently, the CatBoost model was selected as the superior model. Figure 6 summarises the training and validation score of the two best models.

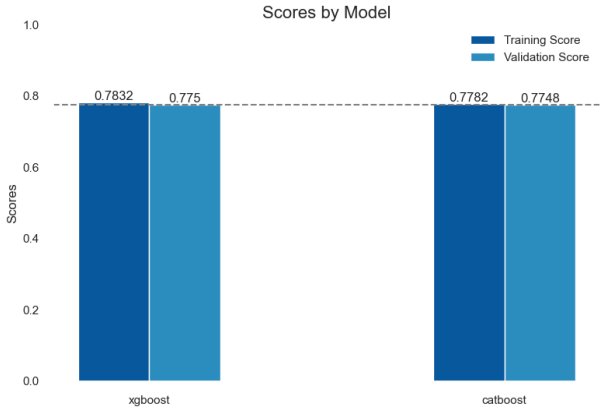

Figure 6- Optimized Models Scores

3.5. EVALUATION

In the context of comprehending the cancellation and churn behavior of hotel clients, the utilization of the confusion matrix (Figure 7) offers noteworthy insights into the predictive model's performance.

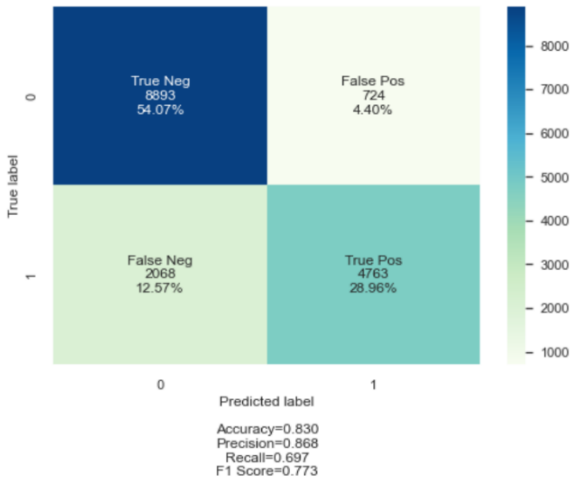

Figure 7- Confusion Matrix

These values enable the calculation of key metrics pertaining to the model's predictive and comprehension capabilities concerning client cancellations. The model attained an **accuracy** rate of approximately 83% of cancellations and non-cancellations. Looking at the **precision**, 86.8% of the predicted cancellations were correct. This metric is crucial for hotel management as it minimizes false positives, ensuring efficient allocation of resources to address genuine cancellation cases. On the other hand, the **recall** is **0.697**. In the context of hotel churn, a higher recall implies that the model can effectively capture a significant proportion of clients likely to cancel their bookings. The **F1** score is **0.773** (77.3%), providing a balanced assessment of both precision and recall. It combines these two metrics to evaluate the overall performance of the model in predicting cancellations, considering both false positives and false negatives. Regarding other methods to analyse the performance of the model, a ROC Curve and Precision-Recall curve (Figure 8) were calculated. Both presented and AUC of 0.9 which indicates a reliable and effective performance in identifying cancelled hotel reservations.

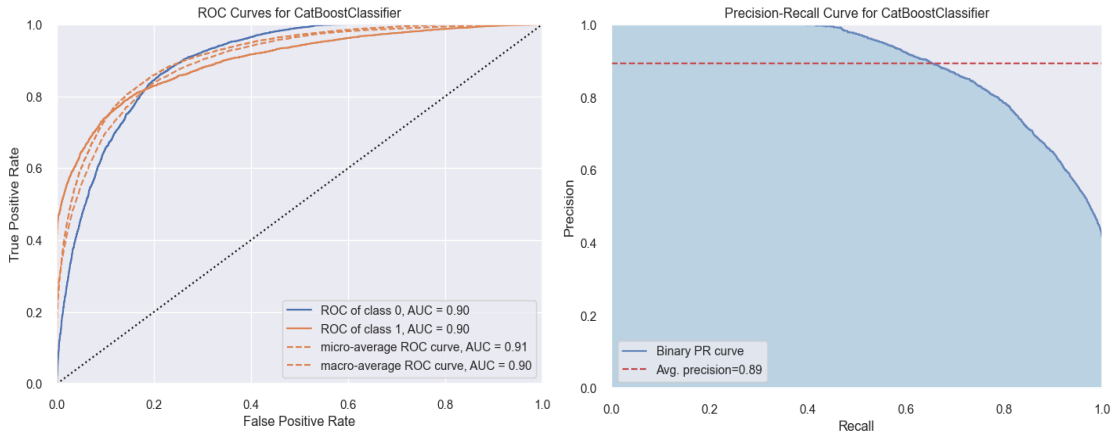

Figure 8- ROC and Precision-Recall Curve

4. RESULTS EVALUATION

The final goal of predicting customer churn using a Machine Learning model was to reduce the cancellations to **20%.** The usage of AI techniques helps the Hotel meet its Business Objectives of reducing demand uncertainty, optimizing pricing and overbooking strategies, and identifying high churn bookings. Firstly, since the model's recall is **70%**, it means that it can effectively identify a larger proportion of clients who are likely to cancel their bookings. Through the usage of the information of the clients who have a higher probability of cancelling their bookings, the Hotel can simply contact these clients to confirm their booking and express appreciation for their early commitment, or either to offer exclusive benefits or perks to incentivize them to retain their booking (including complementary upgrades, exclusive amenities, or discounts on-site services). Another alternative is to send personalized and timely pre-arrival emails, which may include some travel guides about Lisbon and the unique experiences that the hotel offers, to build some anticipation and to address customer concerns and strive to exceed their expectations.

Second, the model's results allow for the analysis of booking patterns of the identified high-churn bookings. Customers who make reservations with more time in advance with a non-refundable deposit type, without any special request or need for parkins space, and of type Transient are more likely to cancel. Notwithstanding the fact the deposit is non-refundable, offering some flexibility in modifying

their reservation by allowing data changes or providing credits for future stays can help avoiding cancelations. In addition to stay in touch as mentioned before, personalized upselling is a good strategy as well, meaning that the staff should identify opportunities to upsell or cross-sell additional services which are aligned with the customers' preferences and needs. Also, provide an exceptional pre-arrival service, by ensuring that all their enquiries and requests are promptly addressed and offer proactive assistance with any additional arrangements they may need. Furthermore, agents 9 and 1 have more cancelations, thus being important to strengthen communication with the Agency to stay updated on booking trends, and customer behaviour changes and to create exclusive incentives or packages that add value to their bookings and make cancellations less attractive. Correctly identifying customers who may cancel and avoid the cancellation can save the hotel 4763 bookings, which corresponds to **1,421,994€** in revenue (considering a 298.55€ ADR per booking).

Third, it is relevant to assess the cost of False Positives. In this case, the model predicts that **4.4%** of the customers will churn, but they do not. Consequently, mistakenly treating a loyal customer result in allocating resources to retain them unnecessarily or providing unnecessary discounts or offers that may impact profitability. Therefore, establishing a feedback loop with front-line employees is critical, by gathering insights about their observations and experiences with customers who were flagged as potential churners. Afterwards, a confirmation process should be implemented, which involves reaching out to customers who have been flagged as potential cancelations to verify their intention to cancel and rectify any mistake.

Finally, there are the False Negatives, which indicate some customers that were likely to churn but they were not identified by the model. In this case, it is important to understand whether there are customer segments consistently missed by the model and collect additional data to enhance the model's predictive power. Moreover, relying solely on the model for churn prediction is unsatisfactory, therefore, the hotel should implement proactive customer retention strategies to prevent churn:

1. Neutralizing OTAs' competitive advantage by ensuring that the hotel website is updated and user-friendly or booking engine is fast, easy, and intuitive. Also, offering discounts on direct bookings or value-added incentives like late-checkout or free breakfast. This is critical to increase retention rates.
2. Identify the right incentives, offer exclusive benefits which are accordance with your customer's characteristics and preferences. For example, for a businessperson room upgrade can be motivating, while for a business travellers may be less appealing.
3. Personalized customer service and experience: Millennials and specially Gen Z seek for authentic experiences. Therefore, the hotel can promote eco-friendly features or partner with cultural palces in the city. Integrating technology to simplify all the journey of the customer in the hotel can also be relevant.
4. Emphasis on room quality: according to a PWC's study in 2016 (*Customer Intelligence Series: What's Driving Customer Loyalty for Today's Hotel Brands?*, 2016), room quality is the most important aspect when selecting a hotel for both business and leisure guests.
5. Do not finish the customer journey when they leave the hotel. Send a personalized email to thank their visit, and to encourage them to post a review.
6. Enhance Online Presence and reputation: actively manage and answer to online reviews on OTAs platforms and take advantage of social medial platforms to engage with potential and current guests. The Hotel can share updates, promotes, showcase the hotel's unique features, or even share curiosities about Lisbon or news about events that are occurring in the city. Also, working with

influencers can really help in increasing hotel's reputation and brand awareness. Regarding overbooking strategies, the model can help identifying the hotel's acceptable level of actual guests overbooking. Nevertheless, placing a limit on the maximum number of overbookings allowed across their various booking channels is critical. Also, by implementing the model, a list of guests subject to potential overbooking versus those who have a guaranteed room can be created and it can be valuable to avoid check-in surprised.

5. DEPLOYMENT AND MAINTENANCE PLANS

5.1. DEPLOYMENT PLAN

To facilitate the deployment of the model, it is essential to undertake the **preparation of the infrastructure**. This involves establishing the required hardware, software, and network configurations necessary to support the model's deployment effectively. This setup should be completed **within two weeks** to ensure a seamless deployment process.

To ensure the model's optimal functioning, a **vigilant monitoring process** must be implemented. Continuous monitoring of the model's performance is imperative, wherein key metrics are tracked and assessed regularly. **Collaboration with relevant personnel** is essential to promptly address any issues or concerns that may arise and to ensure the smooth operation of the model. Within the first month of model deployment, the monitoring process should be established and continued throughout its operation. This collaboration should commence immediately after the deployment and continue on an ongoing basis. To enable the effective utilization of the churn predictions generated by the model, it is imperative to provide comprehensive **training and guidance to the employees** who will be utilizing these predictions. This entails equipping them with the necessary knowledge and understanding of how to interpret and utilize the churn predictions efficiently and effectively. Training sessions should be conducted within the first two weeks after the deployment, with follow-up sessions provided periodically to reinforce the understanding.

Overall, the successful deployment and maintenance of the model rely on the thorough preparation of the infrastructure, vigilant monitoring of performance, and comprehensive training provided to the relevant personnel. These steps collectively contribute to the smooth operation and optimal utilization of the churn prediction model. The entire process, from infrastructure setup to training completion, is expected to be accomplished within the first three months of the project initiation.

5.2. MAINTENANCE PLAN

It is crucial to regularly **update the model with new data** and retrain it when significant changes occur in the data or business environment. By doing so, the model can maintain its effectiveness and adapt to evolving conditions. The model should be updated and retrained at regular intervals, such as every month or quarter, depending on the rate of data and environmental changes. To enhance the model's performance, it is essential to **establish a feedback loop with front-line employees**. This will facilitate the gathering of insights and observations from employees who directly interact with customers, contributing to a better understanding of potential false positives or false negatives that the model may generate. The feedback loop should be established as an ongoing process, with regular meetings or surveys conducted every two weeks or monthly to collect and incorporate employee feedback.

Continuous measurement of key metrics such as accuracy, precision, recall, and F1 score is necessary to track the performance of the model. Additionally, it is important to monitor incoming data to **detect changes in the data distribution** or **underlying patterns** that could impact the model's performance. This monitoring process will help identify and address any drift in the model's behaviour. Key metric measurement and data monitoring should be conducted on a weekly or bi-weekly basis to ensure timely detection and resolution of any performance issues. By implementing these maintenance plan, the hotel can ensure that their models remain up-to-date, adaptable, and perform optimally in a changing business environment. The regular model updates and retraining, feedback loop establishment, and continuous monitoring should be incorporated as ongoing practices throughout the model's lifespan. The entire maintenance plan should be initiated and established within the first two months after the initial deployment of the model.

6. CONCLUSIONS

The analysis conducted by the Young Talent Consulting Group on the booking records of Hotel H2 has resulted in the development of a prediction model with the objective of reducing cancellations from 42% to 20% and providing insights for improved pricing and overbooking policies.

The data preparation phase encompassed addressing missing values, duplicates, and removing outliers using DBSCAN, following the CRISP-DM methodology. Key metrics were utilized to assess the model's performance. Through the utilization of selected features and the CatBoost model, the obtained results have been promising, achieving an accuracy of 83% and an F1 score of 77.3%, achieving the business objectives previously defined. Notably, the model exhibited a recall of 70%. While the model has demonstrated favourable performance, it is crucial to acknowledge the implications of false positives and false negatives. To enhance the predictive capabilities of the model and address missed customer segments, establishing a feedback loop with front-line employees and gathering additional data is recommended. Effectively preventing churn needs the implementation of proactive customer retention strategies in conjunction with the model's predictions. By implementing the provided recommendations and maintaining the model, H2 should encompass tasks such as infrastructure preparation, performance monitoring, staff training, regular updates, and drift monitoring.

In conclusion, the developed CatBoost model provides valuable insights into predicting customer churn, offering actionable strategies to reduce cancellations, optimize hotel operations, and enhance customer retention.

6.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

Although the current models have demonstrated effectiveness in sales forecasting, there exist opportunities for improvement.

One area of improvement is **feature engineering and selection**, wherein additional features should be explored to identify customer churn with predictive power. The set of features used in the model should be continuously evaluated and refined to enhance its performance. Another crucial aspect is **model training and updating**. Regularly retraining the model using new data is necessary to ensure its relevance and ability to capture any changes in customer behavior or market trends. Consideration should be given to implementing **automated or semi-automated processes** for model training and updating to streamline these tasks. Additionally, conducting a **cost-benefit analysis** is essential. It is important to assess the costs associated with false positives and false negatives in churn prediction. By understanding the potential impact on the business, adjustments can be made to optimize the

model's performance based on specific business objectives. Furthermore, **fostering collaboration** among data scientists, domain experts, and business stakeholders is vital. Encouraging collaboration and knowledge sharing allows for a comprehensive understanding of the insights and findings from the modelling process. This, in turn, facilitates informed decision-making and alignment of business strategies with the recommendations provided by the model.

In conclusion, by focusing on feature engineering and selection, model training and updating, cost-benefit analysis, and collaboration and knowledge sharing, improvements can be made to enhance the effectiveness of sales forecasting models and their ability to predict customer churn.

7. REFERENCES

Cancellations shooting up: implications, costs and how to reduce them |. (n.d.). https://www.mirai.com/blog/cancellations-shooting-up-implications-costs-and-how-to-educe-them/

Customer Intelligence Series: What's driving customer loyalty for today's hotel brands? (2016). PwC. https://www.pwc.com/id/en/pwc-publications/industries-publications/telecommunications-media-technology/customer-intelligence-series--what-s-driving-customer-loyalty-fo.html

Global Cancellation Rate of Hotel Reservations Reaches 40% on Average. (2019, April 24). *Hospitality Technology*. https://hospitalitytech.com/global-cancellation-rate-hotel-reservations-reaches-40-average

Grant, P. (2022, May 12). How to Find Outliers With IQR Using Python | Built In. Builtin.com. https://builtin.com/data-science/how-to-find-outliers-with-iqr

Holidays and Observances in Portugal in 2017. (n.d.). Time and Date. https://www.timeanddate.com/holidays/portugal/2017

Foote, K. D. (2022, June 3). Measuring Data Consistency - DATAVERSITY. DATAVERSITY. https://www.dataversity.net/measuring-data-consistency/

Fuchs, J. (2022, September 7). Dynamic Pricing: The Complete Guide. HubSpot. https://blog.hubspot.com/sales/dynamic-pricing

Kulkarni, U. (2017). Overcoming implementation challenges in a big data project. Journal of Emerging Technologies and Innovative Research, 4(4), 376-382.

Lo Duca, A. (2022, January 29). Make Your Machine Learning Model Work Better with DateTime Features. *Medium*. https://towardsdatascience.com/make-your-machine-learning-model-work-better-with-datetime-features-eb21de397fe8

Manning, M. (2021, November 24). *The Ten Best Customer Retention Strategies for Hotels - StayNTouch*. Stayntouch. https://www.stayntouch.com/blog/the-ten-best-customer-retention-strategies-for-hotels/

What Is the Hospitality Industry? Your Complete Guide | Cvent Blog. (n.d.). https://www.cvent.com/en/blog/hospitality/what-is-the-hospitality-industry