MGT 6203 Final Report
GROUP PROJECT #78

Predicting Hotel Booking Cancellations to Minimize Cancellations and Optimize Hotel Revenue Management

GitHub

Team Members:
Jeffrey Sonola (msonola3)
Kristen Hart (khart44)
Xiaofan Jiao (xjiao34)
Kelsey Ritchey (kritchey3)

I. Introduction

Project Overview

The hotel industry is constantly navigating through the complexities of supply and demand and is significantly influenced by the unpredictability of reservation cancellations. These cancellations, driven by many personal and economic factors, posed challenges for hotels in forecasting revenue and managing inventory and pricing effectively. Despite access to detailed booking data, many hotels still struggle tsvo forecast cancellations accurately. To tackle these issues, this project utilized a dataset sourced from Kaggle (2019) of real-world hotel bookings to analyze factors that may influence consumer cancellation behaviors. We aim to delve into factors such as customer demographics, booking timings, and specific reservation details to identify the primary predictors of cancellations. The insights gained from this research will equip hoteliers with the tools to better predict cancellations, enabling them to enhance occupancy rates, tailor services to meet customer needs more effectively, and improve overall revenue management strategies.

Business Justification

From a business perspective, increasing revenue and operational efficiency in the hospitality industry depends on the ability to predict hotel booking cancellations accurately. From a financial standpoint, precise forecasts help hotels maximize room rates and overbooking strategies, which directly affects hotel revenues. An eHotelier article states that a hotel's revenue can be increased with careful revenue management, such as limiting overbooking. Hotels that used revenue management techniques performed well prior to the pandemic, particularly in the difficult 2020 and 2021 years. It did not take three or four years as many analysts had projected; instead, the recovery occurred more quickly. Numerous hotels reached 2019 levels again, and as early as 2021, they even surpassed RevPar (revenue per available room) by 15% to 40% (Terzulli, 2024). This is a considerable increase in a sector where profit margins are thin. Furthermore, anticipating cancellations makes it easier to allocate resources (i.e., housekeeping staff, room occupancies, and others) more effectively, which reduces costs and boosts customer satisfaction. For example, efficiently allocating housekeeping resources ensures timely room turnovers. Anticipating guest check-ins and check-outs can help the hotel allocate housekeeping staff accordingly.

Operationally, this predictive capacity reduces resource waste and ensures superior service delivery by streamlining workforce requirements and room inventory management. Marketing tactics also gain from this, as knowledge of cancellation trends enables customized offers and campaigns to occupy rooms that are expected to become vacant and allows for targeted campaigns to customers least likely to cancel. In conclusion, a strong model for forecasting cancellations of reservations not only solves a significant issue facing the hospitality industry, but it also offers a tactical instrument for improving overall company performance.

Supporting Research Questions

1. How do customer demographics and booking behaviors correlate with cancellation likelihood?
2. What impact do seasonal trends, room types, and advance booking periods have on cancellation rates?
3. How do different pricing strategies and payment conditions influence the probability of a booking being canceled?

4. Are there identifiable patterns or clusters in the data that signify higher risks of cancellations, and how can these be addressed through targeted strategies?

Initial Hypothesis

Based on preliminary data observations and industry insights, we hypothesized that specific variables such as customer demographics, time of booking, and chosen room types significantly influence the likelihood of cancellations. It is anticipated that customers from certain age groups or geographical locations may exhibit higher cancellation rates. Similarly, bookings made during certain seasons or well in advance are presumed to have a higher probability of being canceled. Additionally, premium pricing strategies and stringent payment conditions might deter cancellations by committing customers more firmly to their reservations. These hypotheses will guide the initial phase of the investigation and will lay the groundwork for developing a model to effectively predict future booking cancellations.

II. Methodology

Data Processing

We discovered that there is a substantial quantity of missing data in several columns during the data cleaning procedure. Specifically, the 'Company' and 'Agent' columns, which showed a high percentage of missing values (94.31% and 13.69%, respectively). We decided to remove these columns to maintain the integrity of the dataset going forward, as the excessive amount of missing data may introduce biases or inaccuracies in our subsequent analyses and these columns might not be essential for the intended analysis or predictive modeling.

The 'Children' column had a negligible number of missing values (less than 0.01%), and hence, these were replaced with the mean value of the column. This approach is standard for numerical variables, as it maintains the overall distribution of the data. For the 'Country' column, which is categorical and had 0.41% missing values, the missing entries were replaced with the most frequent value. This method helps in preserving the most common category within the dataset.

Exploratory Data Analysis

In our correlation analysis, we observed a variety of relationships between the different predictors related to hotel bookings. Most notably, certain predictors such as lead_time, total_of_special_requests, required_car_parking_spaces, booking_changes, and previous_cancellations demonstrate noteworthy correlations with the target variable is_canceled.

- Lead_time shows a positive correlation of approximately 0.293 with is_canceled, suggesting that the longer the interval between booking and arrival, the higher the probability of cancellation.
- A negative correlation of -0.235 between total_of_special_requests and is_canceled implies that bookings with more special requests tend to be honored rather than canceled.
- Required_car_parking_spaces have a notable negative correlation of -0.195 with is_canceled, suggesting that bookings made with car parking requirements are less likely to be canceled.
- The relationship between booking_changes and is_canceled shows a negative correlation of -0.144, which indicates that modifications to a booking could potentially reduce the likelihood of cancellation.

- Lastly, previous_cancellations are positively correlated with is_canceled at 0.110. This finding could highlight a pattern of behavior where guests who have canceled in the past may be more likely to cancel future bookings as well.

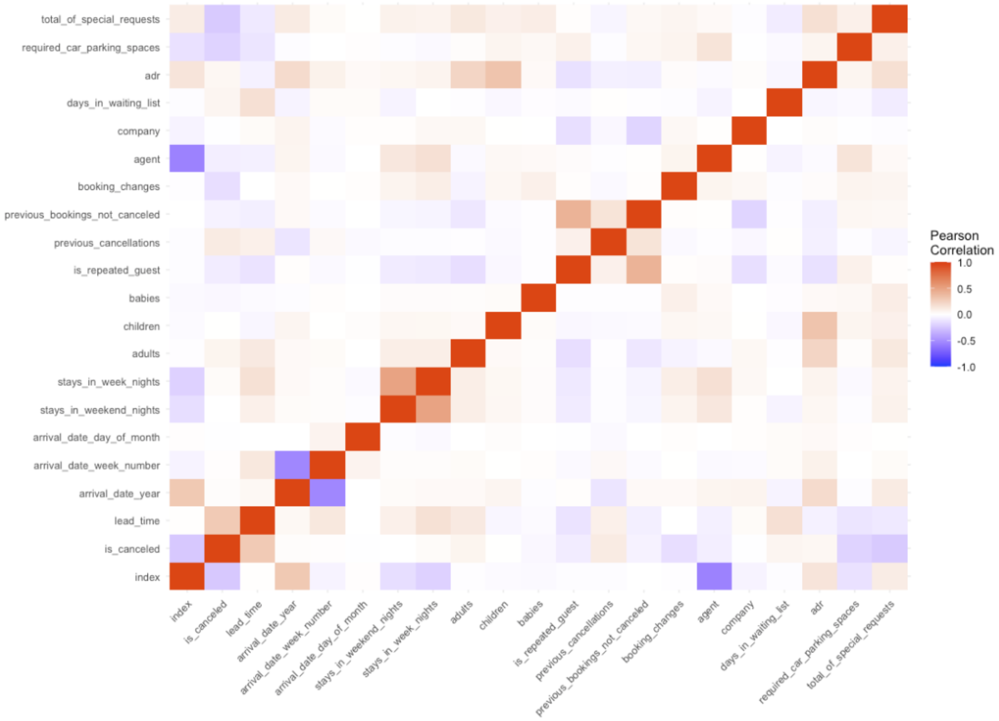

Figure 1: Correlation Analysis

Based on the insights gained from the correlation analysis, the selection of variables for our predictive model reflects significant correlations with `is_canceled` and covers key variables of booking dynamics. `Lead_time` and `previous_cancellations` are direct indicators of cancellation likelihood, with their positive correlations suggesting higher cancellation risks over longer planning periods and among guests with a history of cancellations. In contrast, `total_of_special_requests` and `required_car_parking_spaces` show negative correlations, indicating that bookings with specific commitments are less likely to be canceled. Including `booking_changes`, although less strongly correlated, aligns with the notion that flexibility might decrease cancellation. These chosen variables balance statistical significance with practical relevance, aiming for a model that captures the nuanced patterns of guest behavior in hotel bookings.

Statistical Models

From the previous variables selected, we conducted comprehensive predictive models to further our studies. We incorporated machine learning algorithms, such as Logistic Regression, Random Forest, and Support Vector Machines (SVM), to capture the non-linear relationships and complex interactions between predictors. These methods are well-suited to handle the high dimensionality of our data and are expected to enhance predictive performance. Also, we will conduct a Holt-Winters analysis to examine temporal trends and seasonality in booking cancellations. The combination of these advanced techniques will allow us to validate our initial findings and provide robust predictions.

Logistic Regression

We applied logistic regression to predict the likelihood of hotel booking cancellations. The model's coefficients indicate how each predictor variable affects the log odds of a cancellation. Notably, longer lead_time increases the probability of cancellation, suggesting that customers making reservations further in advance may face greater uncertainty in their plans. The variable stays_in_week_nights showed an initial negative impact on cancellations, but this relationship became positive when more variables were introduced, indicating a nuanced effect possibly influenced by the length of stay.

Interestingly, the deposit_type of 'Non Refund' was associated with a higher likelihood of cancellation, which could imply that guests are willing to risk losing a non-refundable deposit due to the discounted rates often associated with such bookings. is_repeated_guest had a negative coefficient, demonstrating that repeat guests are less likely to cancel their bookings, reflecting perhaps a loyalty effect or more certainty in their travel plans. For the market_segment, different booking sources influenced cancellation probabilities in varying ways, highlighting the role of the booking channel in predicting cancellations.

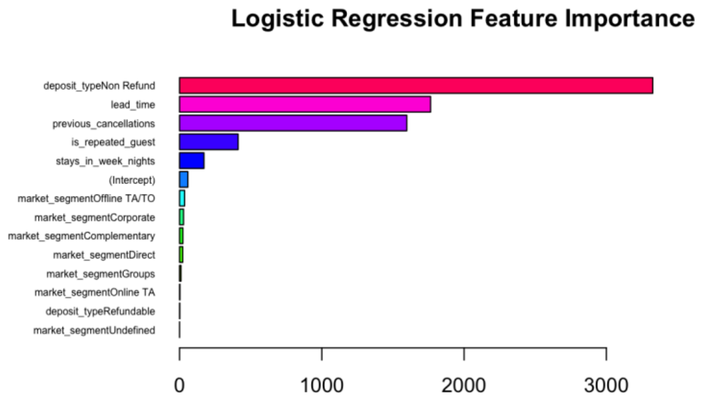

Figure 2: Logistic Regression Feature Importance

The confusion matrix showed that the model performed with an accuracy of 76.8%, a precision of 97.7%, and an F1 score of 84.1%. Though the model is effective at predicting when a booking will not be canceled, the number of false positives indicates that it is not very accurate at spotting cancellations. This might be an area where the model is further refined, perhaps by looking into different classification criteria or other variables that could increase the predicted accuracy.

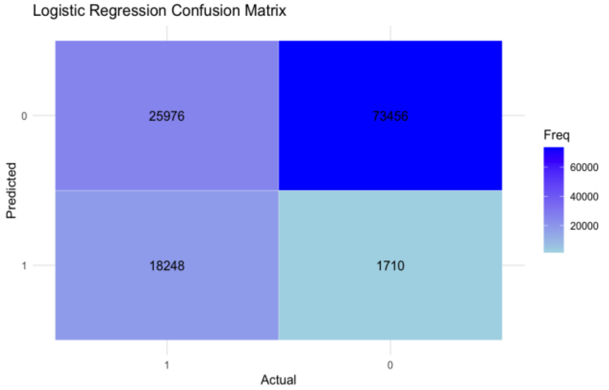

Figure 3: Logistic Regression Confusion Matrix.

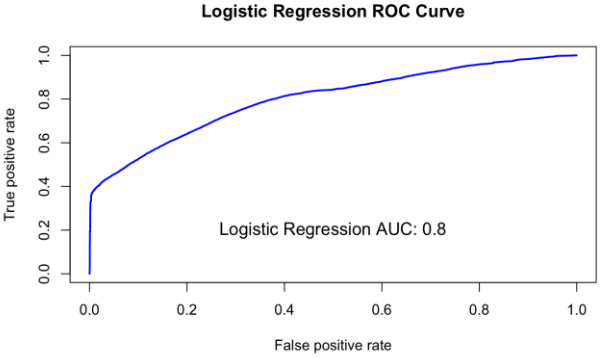

Figure 4: Logistic Regression ROC Curve

Random Forest

In parallel to logistic regression, our analysis explored the predictive capabilities of a Random Forest model, using an ensemble of 500 trees to enhance accuracy and robustness by combining numerous decision tree predictions to better discern the likelihood of hotel booking cancellations.

The feature importance plot reveals that the deposit_type variable stands out as the most significant predictor, overshadowing other variables like lead_time, previous_cancellations, and market_segment. This indicates the booking deposit's conditions play a crucial role in the cancellation decision. Notably, the duration of the booking (stays_in_week_nights) and whether the guest has booked previously (is_repeated_guest) have less influence on the model, but still contribute to its decision-making process.

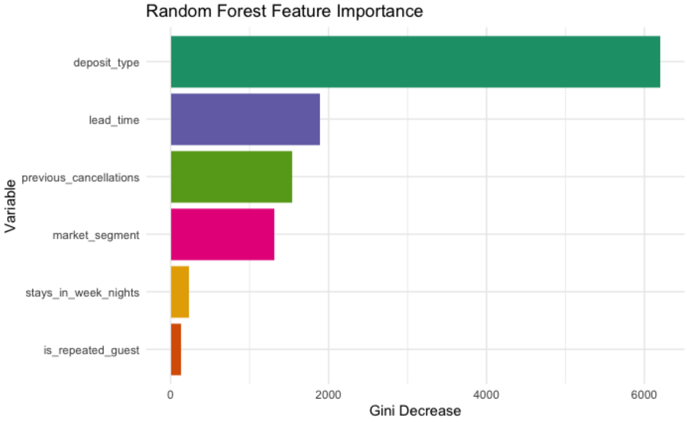

Figure 5: Random Forest Feature Importance

Contrary to typical expectations, the confusion matrix for the Random Forest model reveals its strong capability in correctly predicting cancellations (True Positives) with 5,442 bookings accurately identified as canceled, while demonstrating limited performance in identifying non-cancellations (True Negatives) with only 30 bookings correctly classified. The model shows an increasing number of False Positives, 3,405 cases, indicating a tendency to falsely predict a booking as canceled. Despite this, the model achieves a precision of 73.38% and an exceptionally high recall (sensitivity) of 99.80%, emphasizing its strength in identifying canceled bookings. The F1 score stands at 84.57%, reflecting a robust balance between precision and recall.

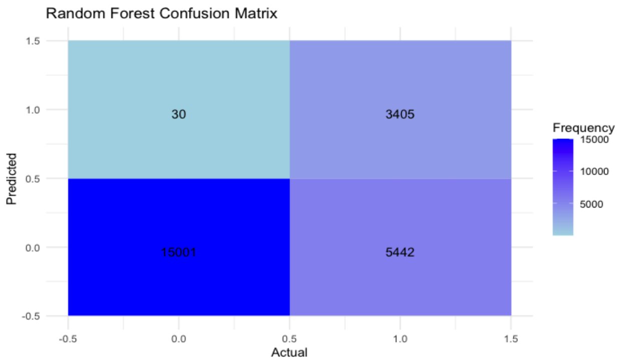

Figure 6: Random Forest Confusion Matrix

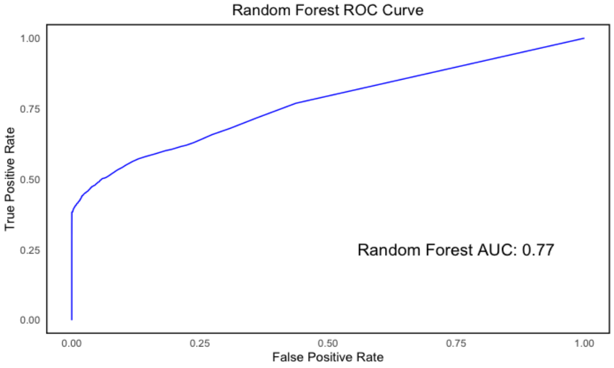

Figure 7: Random Forest ROC Curve

The AUC score of 0.77 further quantifies the model's performance. The AUC score reflects a good but not perfect classification ability, suggesting that while the model is adept at distinguishing between canceled and non-canceled bookings, there is still room for improvement. Strategies to refine the model could include rebalancing the dataset to better capture the minority class or incorporating additional predictive features. Fine-tuning model parameters and implementing advanced resampling techniques may also enhance the model's ability to identify true cancellations, ultimately leading to a more balanced predictive performance.

Support Vector Machines (SVM)

An exploration of SVM was conducted as a complementary analysis for our research, a model known for its effectiveness in binary classification problems. The model underwent a rigorous training and testing process, with a 70-30 split in the data to ensure robust validation. First, a learning curve was generated, serving as a visual presentation of the model's improvement trajectory as it enlarges its data during the training phase. The learning curve revealed important insights into its ability to generalize as it transitioned from a training environment to making predictions on unseen data. The initial high accuracy on the small training set suggested overfitting, which gradually stabilized as more data was introduced. Notably, the cross-validation score's upward trend with increased data volume indicated an enhancement in the model’s ability to generalize.

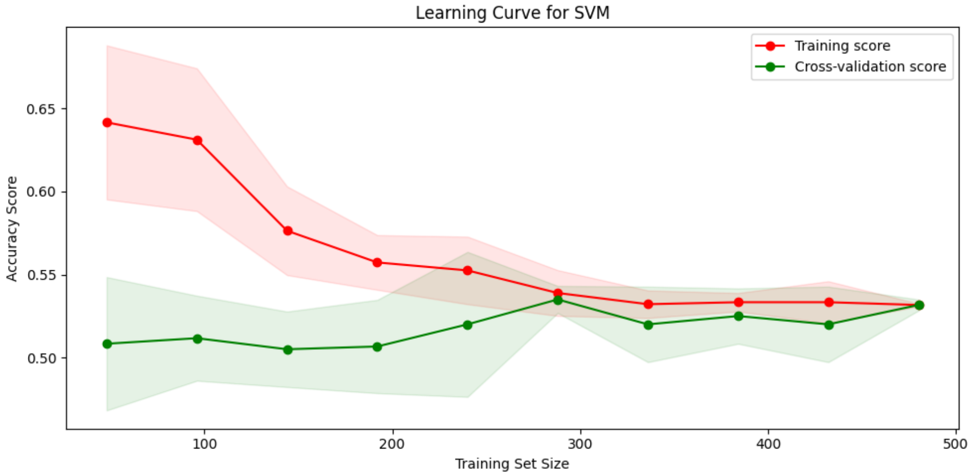

Figure 8: Learning Curve for SVM

Our SVM model demonstrated commendable predictive capabilities with an accuracy of 76.42%. Precision was particularly high at 97.92%, suggesting that when the model predicted cancellations, it was correct most of the time. Despite these strong points, the model's sensitivity, or recall, was somewhat moderate at 37.47%, indicating that it was less adept at identifying all true cancellations. The specificity, which is not quantified here, would still likely reflect a high success rate in identifying non-cancellations based on the nature of SVM's performance characteristics. The resulting F1 score of 54.2% reveals that there is room for improvement, especially in creating a more balanced relationship between precision and recall enhancing the model’s overall performance.

The ROC curve further reflected the model's fair discriminative ability with an AUC of 0.76, a somewhat strong ability of the SVM to differentiate between canceled and non-canceled bookings, marking a significant discriminative power. To enhance the SVM's performance, we would recommend strategies such as rebalancing the dataset or refining the features.

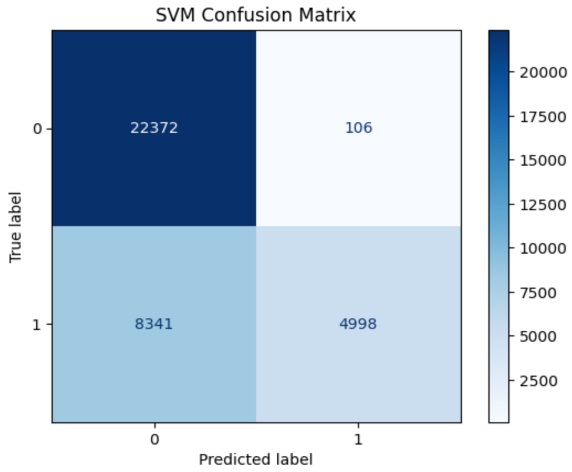

Figure 9: SVM Confusion Matrix.

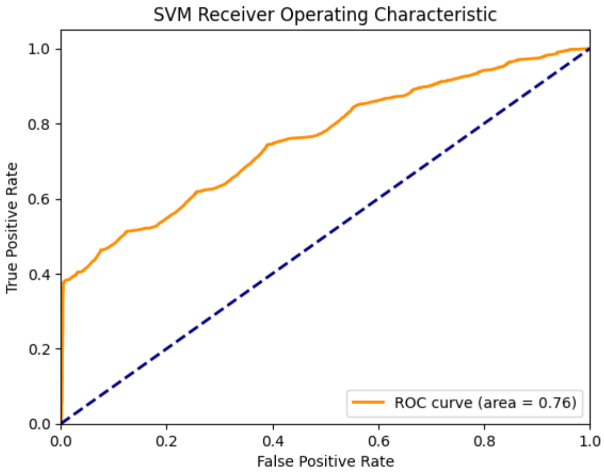

Figure 10: SVM ROC Curve

Holt-Winters Analysis

For further investigation, we employed the Holt-Winters seasonal method, examining data from 2015 to 2017. This approach decomposes the time series data into three primary components: level, trend, and seasonality. The method is particularly adept at handling data with a seasonal component, which is predominantly shaped by the patterns of travel that flow with the seasons. By applying exponential smoothing techniques, the Holt-Winters method meticulously estimates each of these components, offering a nuanced view of the underlying trends in cancellation rates.

The Holt-Winters model's application to our dataset revealed a distinct seasonal pattern in cancellation rates, peaking in the summer months and declining during the winter, aligning with conventional travel behaviors that surge due to holiday seasons. Additionally, the model detected a marginal but consistent upward trend in cancellations year after year, suggesting a potential area for closer scrutiny despite the limited temporal scope of our dataset. After extracting the seasonal and trend components, the remainder of the time series showed random fluctuations, indicating the absence of unaccounted systematic patterns. These insights confirm the travel industry's susceptibility to seasonal consumer behavior cycles and underscore the importance of tactical approaches like implementing non-refundable rates during peak seasons or adopting flexible pricing strategies to optimize occupancy rates, as the forecast suggests that the trend of higher summer cancellations is likely to persist.

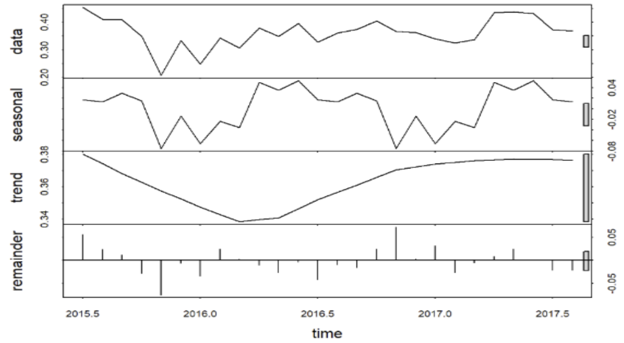

Figure 11: Holt-Winters Analysis

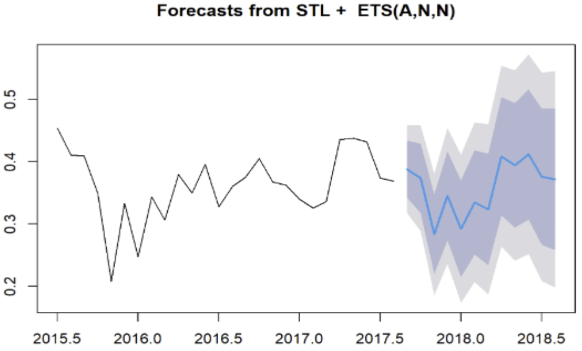

Figure 12: Forecasts

III. Conclusions

In conclusion, our research investigated various factors, including lead time, special requests, car parking requirements, booking changes, and previous cancellations, that influence hotel booking cancellations. To explore these factors, we utilized various predictive models, each offering unique perspectives and strengths. Logistic Regression was employed to establish a foundational understanding of how different variables impact the probability of cancellation. This model highlighted the critical role of advanced booking times and the presence of special requests in predicting cancellations. The Random Forest model allowed for an in-depth examination of guest behavior, uncovering significant effects of deposit type and market segment on the likelihood of cancellation. This model's ability to handle complex interactions between variables provided deeper insights into the dynamics of guest decisions. The SVM model was instrumental in classifying non-cancellations. However, it requires additional tuning to enhance its sensitivity towards actual cancellations, indicating a potential area for improvement in predictive performance. Additionally, the Holt-Winters method was applied to capture seasonal trends within the booking data, confirming the hotel industry's vulnerability to seasonal fluctuations in consumer behavior. This analysis was pivotal in understanding the temporal patterns that affect booking cancellations. Collectively, these models achieve high predictive accuracy, but challenges remain in reducing false positives and improving the sensitivity towards true cancellations. The nuanced understanding gleaned from this research lays a robust foundation for the implementation of more targeted and effective management strategies within the hotel industry, aiming to minimize cancellations and optimize guest satisfaction.

IV. Recommendations and Limitation

Based on our findings, we would recommend hotel managers to implement dynamic pricing models that adjust rates based on varying lead times and seasonal trends. This approach is standard and can discourage cancellations by offering non-refundable rates during peak periods, thereby mitigating risks while maximizing revenue. Hotels can also develop targeted marketing campaigns that focus on demographics and guest profiles with lower likelihoods of cancellation. Enhancing customer engagement through personalized offers and loyalty programs can significantly reduce cancellation rates. Future research studies could include economic factors or global events and fine tune these machine learning models to help refine the predictive models, enhancing their accuracy and relevance.

While the findings of this study offer valuable insights into hotel booking cancellations, there are several limitations to consider. Firstly, it is unclear where these hotels are located, but from the analysis, we see that most guests come from Portugal. These limitations could restrict how broadly the findings can be applied to other geographic areas with distinct cultural, economic, and tourism-related dynamics. Furthermore, the data was collected prior to the COVID-19 pandemic, so it does not consider the substantial changes in visitor behavior and hotel business changes due to the pandemic. Moreover, while the dataset includes a variety of booking details such as length of stay, number of guests, and so on, it may lack other potentially influential factors like reasons for travel (business vs leisure) or external economic conditions. These omissions could affect the comprehensiveness and applicability of the predictive models developed from this data.

V. Reference

Kaggle (2019). Hotel booking demand. Kaggle. Retrieved February 25, 2024, from https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand?resource=download

Terzulli, M. (2024, February 24). Transform your hotel's revenue in two years or less with this guide to revenue management. eHotelier. Retrieved February 25, 2024, from https://insights.ehotelier.com/insights/2024/02/24/transform-your-hotels-revenue-in-two-years-or-less-with-this-guide-to-revenue-management/