Hotel Cancellation Analysis Report

University of Louisville

Data Analytics II

Presented By:
Rahela Jawadi

Abstract

This report presents a comprehensive analysis of hotel booking cancellations within the context of a case competition focused on hotel cancellations. The study was conducted as part of the Data Analytics II class and aimed to examine the factors and patterns influencing hotel cancellations. The primary objective was to understand the reasons behind cancellations and provide strategies to minimize their impact on the hospitality industry. Through the utilization of various data analytics techniques such as exploratory analysis, descriptive statistics, predictive modeling, and trend analysis, the report explored a dataset consisting of booking and cancellation records. The findings revealed key drivers of cancellations, including factors like booking lead time, customer profiles, seasonal variations, and hotel features. Furthermore, the report offers recommendations to hotel management on improving cancellation policies, optimizing revenue management, and enhancing the overall guest experience. In summary, the analysis provides valuable insights into hotel cancellations, empowering industry stakeholders to make informed decisions and implement effective strategies in response to this significant challenge.

Contents

Introduction

Hotel booking cancellations pose a significant challenge in the hospitality industry, impacting revenue and room capacity management. To address this problem, we utilized our expertise in software development, technology, and business operations. Leveraging a gradient boosting machine model (GBM) with Extreme Gradient Boosting Tree (xgbtree) algorithm, we aimed to predict cancellations and provide proactive solutions. Our analysis focused on understanding the reasons behind cancellations and identifying the most likely customer segment to cancel.

Through an iterative process, we optimized the performance of the GBM model by tuning hyperparameters. The model's performance was evaluated using the AUC (Area Under the Curve) metric, which measures the model's ability to discriminate between cancellations and non-cancellations. The GBM model, trained using the (xgbtree) algorithm, achieved an impressive AUC result of 0.85, indicating its strong predictive power.

By understanding the factors influencing cancellations and leveraging the predictive capabilities of the GBM model, hotels can implement proactive strategies. This includes flexible cancellation policies, personalized communication approaches, optimized pricing, and advanced analytics. These strategies aim to reduce cancellations, maximize revenue, improve room capacity management, and enhance customer satisfaction. Ultimately, our analysis provides valuable insights for hoteliers to tackle the challenge of hotel booking cancellations and optimize their operations in a competitive market.

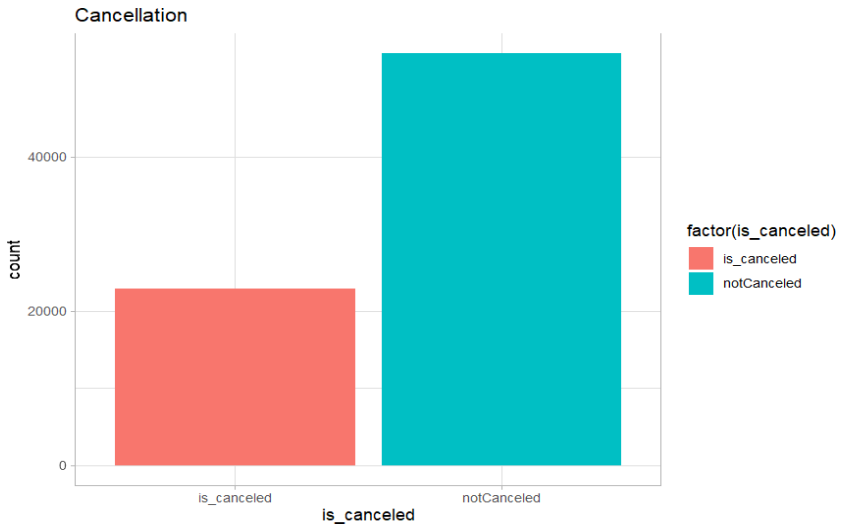
Exploratory Data Analysis

The training dataset provided for our analysis contained 66749 samples and 78 features, with a response variable (canceled and notCanceled). Features were divided into four categories of hotel type, the arrival date of year, month, and day, and guest preferences (room type, deposit type, and meal type). Among the variables, two require a definition: ADR and lead time.

ADR (Average Daily Rate): ADR is the average daily rate calculated by dividing the sum of all lodging transactions by the total number of staying nights in a hotel. It represents the average price guests pay per room for a specific period, indicating the hotel's revenue per night on average.

Lead time: Lead time refers to the number of days that elapsed between the entering date of a booking into the system and the guest's arrival date.

Each data row contained information about a guest's preference (type of food, meal, deposit type, and extra), arrival date, and other information related to guests. To further understand our features and gain insights for future feature engineering, we investigated some variables below:



The bar plot above displays the cancellation rate, which accounts for approximately 30% of the data. This indicates a substantial portion of the dataset being affected by cancellations. Consequently, it becomes crucial to identify the factors influencing these cancellations. Among a lot of the factors that we have, we will pick the lead time to see its effect on cancellation.

Also, since the data comes from two different hotels, we have analyzed the chart below to determine which hotel has the highest number of guests based on market segmentation.

The graph clearly indicates that the city hotel has a larger number of guests compared to the resort hotel across all market segments. Notably, the online travel agency segment stands out as having the highest number of bookings, surpassing all other market segments significantly.

Data Cleaning

For data cleaning, we began by addressing duplicate rows, which constituted a significant portion of the dataset. We removed all duplicated rows to ensure data integrity. Next, we encountered columns with missing values, which required applying various cleaning strategies. However, due to the large volume of data and the nature of these categorical columns, encoding them resulted in separate columns, potentially leading to high cardinality. To mitigate this issue, we made the decision to remove three columns: company, agent, and county. This was due to complications encountered while running the code.

Also, we decided to exclude the reservation status column. This decision was based on the high collinearity observed between the reservation status column and the ‘is_canceled’ response variable, as they contained redundant information. After removing the reservation status column, we also decided to remove the reservation status date column. Additionally, we addressed outliers in the dataset. In some cases, we chose to remove the outliers completely, while in other cases, we replaced them with either the median or mean values, depending on the specific feature. These steps were taken to ensure the quality and integrity of the data for further analysis.

Feature Engineering (One-Hot encoding)

One-hot encoding is a method used to convert categorical variables into a set of binary indicators, enabling the representation of categorical data in a numeric format that is compatible with modeling techniques. For instance, let's consider the "deposit_type" column, which has three categorical entries: ‘refundable’, ‘nonrefundable’, and ‘no deposit’. With one-hot encoding, these categories would be transformed into three separate columns, each serving as a binary indicator for a specific category as shown below.

deposit _type
refundable
nonrefundable
no deposit
refundable
no deposit

deposit_type.no deposit	deposit_type.nonrefundable	deposit_type.refundable
0	0	1
0	1	0
1	0	0
0	0	1
1	0	0

However, a drawback of one-hot encoding is the generation of a new feature for every unique value present in the categorical variable. This can lead to an increase in the dimensionality of the dataset, potentially causing challenges in computational efficiency and interpretation. For instance, the column country contains 175 categories and as a result of one-hot encoding, we will have 175 new columns. So, these are the issues that should be considered. Therefore, since tree-based models handle categorical variables with many levels without one-hot encoding, preventing excessive column generation and maintaining interpretability, we can ignore the one-hot encoding process in this case that our data has high cardinality.

Modeling

We ran four different models, namely random forest, lasso classification, subset selection, and xgboost. After evaluating their performance, we determined that the xgboost model exhibited a higher AUC accuracy compared to the other models. Therefore, we decided to proceed with the xgboost model as our final choice.

We further fine-tuned the xgboost model, which resulted in an optimal ROC curve (refer to Figure 1). When tested with the validation set, the model achieved an AUC score of 0.83. This indicates a good level of predictive accuracy and suggests that the xgboost model is well-suited for the given task.

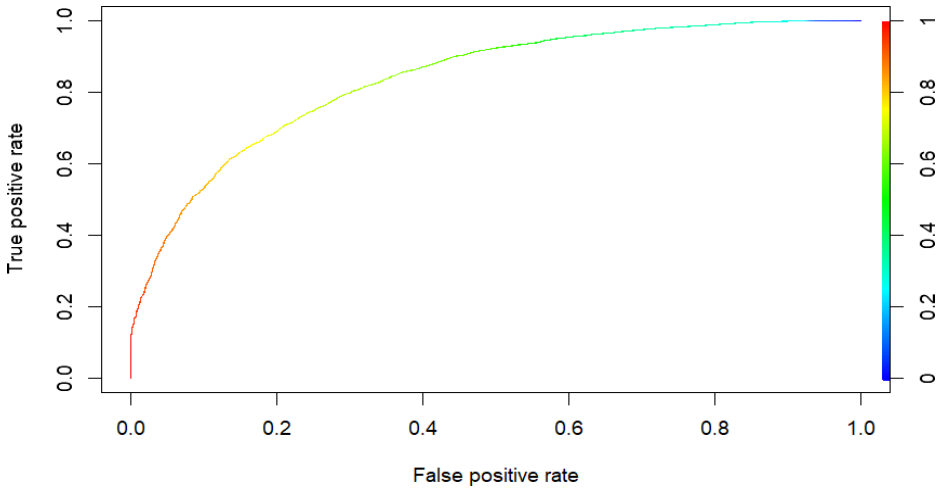


Figure 1. ROC curve for the tuned Gradient Boosting Tree model.

Feature Importance

To gain valuable insights from the model's output, it is crucial to identify the key features that significantly influence the target variable. To achieve this, we plotted the top 15 most important features as determined by the Gradient Boosting Tree model. This visualization provides a clear understanding of the influential factors that contribute the most to the model's predictions (refer to Figure 2).

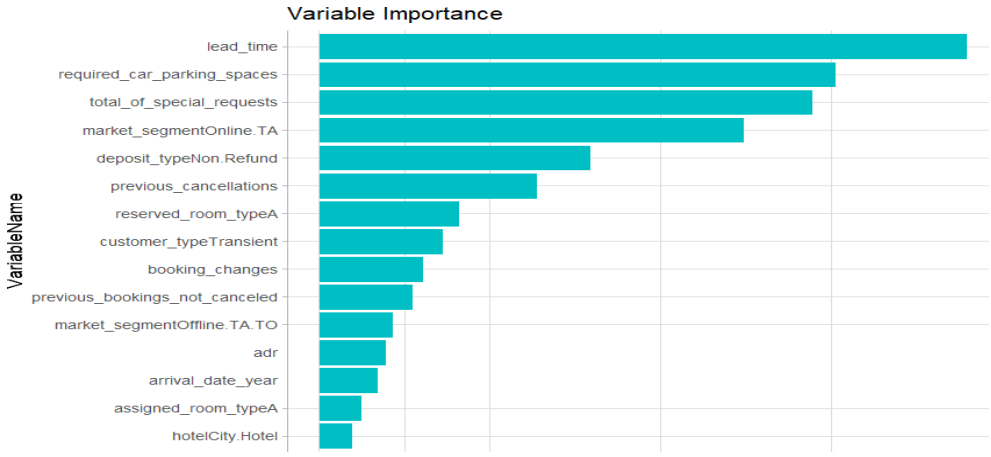
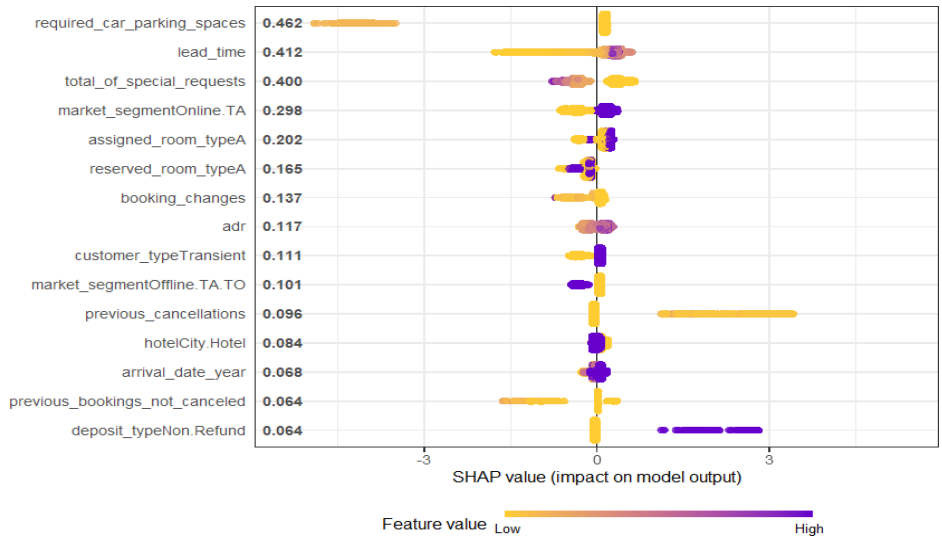


Figure 2. Variable Importance Gradient Boosting Tree model.

The SHAP plot below illustrates the impact of each feature on the final output or the response variable.



Features with higher SHAP values indicate a greater influence on the overall output of the target variable. By analyzing the SHAP plot, we can discern the relative importance of each feature in contributing to the model's predictions for the target variable. For instance, if we consider the number of required car parking spaces, we observe that an increase in the number of required car parking spaces has a negative impact on the model's output.

For better understanding and analysis, we have divided these important variables into three categories:

High Demand Time Frames:

- Arrival date by month
- Arrival date by year
- Adr

Guest's preferences and special requests:

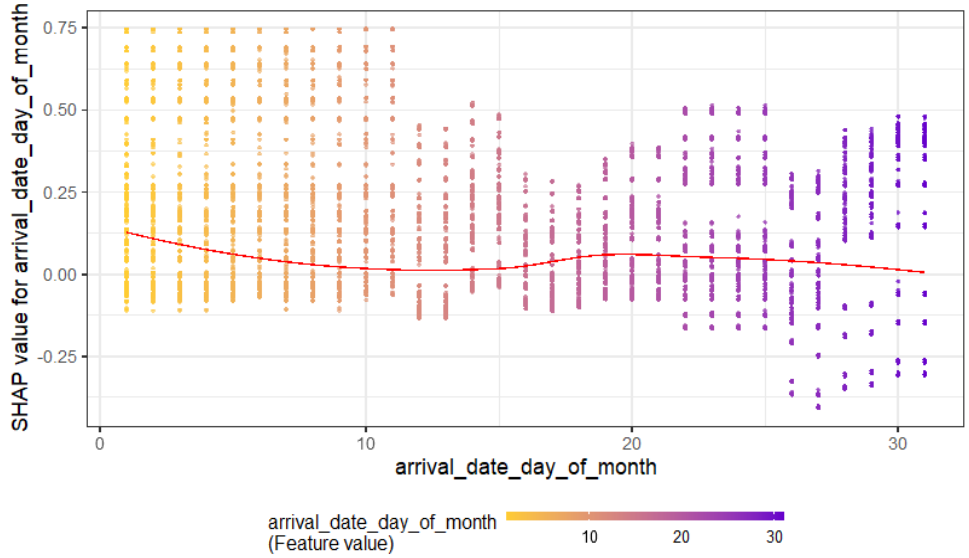
- Total number of requests
- Market segmentation
- Required car parking spaces
- Deposit Type non Refund

Lead Time

By categorizing these variables, we can gain deeper insights and effectively analyze their impact on the overall outcome.

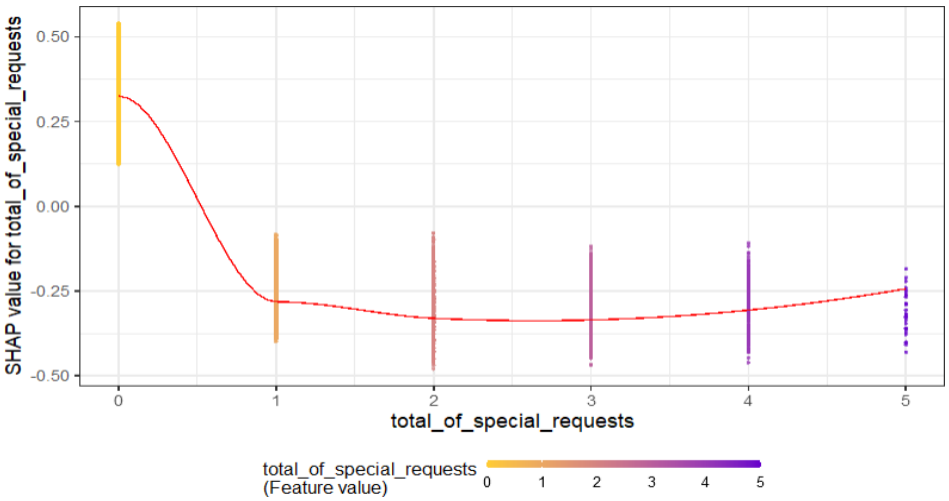
Individual Dependency Analysis

High Demand Time Frames



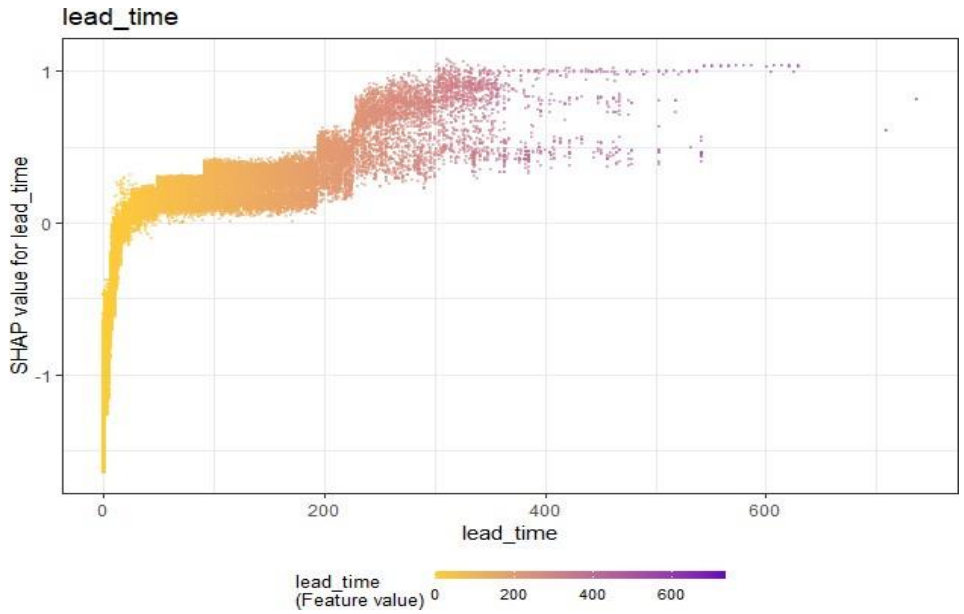
The dependence SHAP plot shows that arrivals during the first 15 days of the month have a positive impact on the predicted outcome, while arrivals during the last 15 days have a negative impact. Timing of arrivals within the month is influential, with early days being more favorable and later days less favorable for the desired outcome. This information helps understand and predict arrival patterns based on the day of the month.

Guest's preferences and special requests



In the dependence SHAP plot of the number of special requests, the negative side indicates that as the number of requests increases, it has a negative impact on the target variable. This suggests that a higher number of special requests may lead to less desirable or negative outcomes. Furthermore, the plot shows a decreasing trend as the number of special requests increases, implying that the impact of additional requests becomes less significant. This understanding can aid in optimizing processes and setting appropriate limits for special requests to achieve better outcomes.

Lead Time



A longer lead time has a positive impact on the outcome, while a shorter lead time has a negative impact. Booking in advance increases the likelihood of a positive experience. Last-minute bookings may lead to challenges. This information helps hospitality businesses optimize their booking strategies.

Recommendations

High Demand Time Frames

- Marketing and promotion: Focus on effective marketing and promotional strategies during high demand periods to attract more customers and increase bookings.
- Online Presence and Visibility: Enhance online presence through website optimization, social media engagement, and online advertising to increase visibility and reach a wider audience.

- Competitive Pricing: Analyze competitor pricing and adjust rates accordingly to remain competitive while maximizing revenue during high demand time frames.

Guest's Preferences and Special Requests

- Improve Website information: Enhance the website's content and information to provide comprehensive details about rooms, amenities, services, and special offerings, helping guests make informed decisions.
- Optimize operational efficiency: Streamline internal processes and operations to improve efficiency, reduce wait times, and enhance the overall guest experience.
- Simplify and streamline the process of handling special requests: Develop a streamlined system to handle special requests effectively, ensuring clear communication, prompt responses, and accurate fulfillment of guest preferences.

Lead Time

- Implement advance purchase discounts: Offer discounted rates or exclusive offers for guests who book in advance, incentivizing early reservations and improving cash flow.
- Monitor booking patterns and identify trends related to lead time: Analyze historical data to identify booking patterns and trends, helping to optimize pricing, staffing, and resource allocation based on lead time.

General Recommendation

- Consider offering incentives for guests who choose not to cancel: Provide rewards, discounts, or perks to guests who maintain their bookings, encouraging them to stay committed and reducing cancellations.
- Regularly analyze customer feedback and reviews: Pay close attention to customer feedback and reviews, gathering insights to identify areas for improvement and enhance the overall guest experience.
- Implement a loyalty program to reward repeat guests and encourage their continued patronage: Create a loyalty program that offers exclusive benefits, rewards, or discounts to frequent guests, fostering loyalty and incentivizing return visits.

Conclusion

By developing an XGBoost tree model with an AUC score of 0.83, we successfully identified the factors influencing booking cancellations. These features were categorized into three main areas: improving website information, monitoring booking patterns and lead time trends, and implementing competitive pricing strategies. Through solutions like enhancing website information, analyzing booking patterns, and adjusting pricing, BOOKNOW aims to enhance its hospitality services and minimize cancellations.