**Machine Learning: Hotel Reservation Cancellation Prediction**

Exploring Predictive Models for Hotel Booking Cancellation Anticipation

Sai Mohan Anjesh

Aditya Dasari

# Abstract:

In today's world, hotel bookings have become an integral part of travel planning. This project utilizes advanced machine learning techniques to predict the likelihood of a hotel reservation being honored. By analyzing a vast dataset that includes diverse booking attributes and customer history, we have developed predictive models that can anticipate booking outcomes. These models have the ability to identify if customers are likely to cancel their reservations, allowing hotel managers to make informed decisions. Through training models such as Regularized Logistic Regression, Support Vector Machine, and Bagging and Boosting Models on over 36,000 cases, our team has generated two possible outcomes: a customer either cancels their reservation or honors it. Our models have proven valuable in providing the industry with insights to improve hotel bookings and customer satisfaction.

# Motivation:

A significant number of hotel reservations are called off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with. The news article on USA Today was the genesis of this project. We read that there had been bogus reservations and cancellations in recent times in the lodging industry and thought it would be better to solve this problem for the companies hence helping hotel management to properly manage their revenue and sales.

# Problem Definition and Goals:

The main objective is to address the challenge of managing reservation cancellations and no-shows in the hospitality industry. This approach aims to enhance revenue management strategies and operational efficiency by accurately predicting cancellations. This will empower hotels to optimize resource allocation, refine pricing strategies, and implement targeted interventions to minimize revenue loss and streamline operations effectively. The goal is to use binary classification methods to predict whether a hotel reservation will be cancelled. We are particularly interested in understanding the characteristics of people who are likely to cancel their booked rooms.

## Data Dictionary:

Booking\_ID: unique identifier of each booking

no\_of\_adults: Number of adults

no\_of\_children: Number of Children

no\_of\_weekend\_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

no\_of\_week\_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel

type\_of\_meal\_plan: Type of meal plan booked by the customer:

required\_car\_parking\_space: Does the customer require a car parking space? (0 - No, 1- Yes)

room\_type\_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels.

lead\_time: Number of days between the date of booking and the arrival date

arrival\_year: Year of arrival date

arrival\_month: Month of arrival date

arrival\_date: Date of the month

market\_segment\_type: Market segment designation.

repeated\_guest: Is the customer a repeated guest? (0 - No, 1- Yes)

no\_of\_previous\_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking

no\_of\_previous\_bookings\_not\_canceled: Number of previous bookings not canceled by the customer prior to the current booking

avg\_price\_per\_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)

no\_of\_special\_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)

booking\_status: Flag indicating if the booking was canceled or not.

Prior to training machine learning models, it is important to clean the data which involves discarding irrelevant features and preprocessing the data. The exact steps for data exploration and preprocessing are described in the section titled "Data Exploration and Preprocessing". After the dataset has been cleaned and prepared, different classifiers are trained and fine-tuned. Finally, the performance of the models is compared using accuracy and AUC as the performance measures.

# Related Work:

There are numerous papers in the field of hospitality demand and revenue management that talk about various methods, yet this topic is still a long-term challenge to solve. One such paper titled "Comparison and Analysis of Machine Learning Models to Predict Hotel Booking Cancellation" by Yiying Chen, Chuhan Ding, Hanjie Ye, and Yuchen Zhou provides three possible substitutes for the neural network, including logistic regression, k-Nearest Neighbor (k-NN), and CatBoost. The paper uses a dataset adapted from Kaggle, which includes booking data from two types of hotels (resort hotel and city hotel) in Portugal and corresponding customer information. Key variables were selected as the predictor to train and test the prediction models based on three machine learning algorithms. After preprocessing the raw data by standardizing, dealing with missing data, recoding some variables, and scaling, the researchers conducted the prediction and compared each model through three metrics (confusion matrix, accuracy score, and F1-score). [1]

Our project is based on a dataset from Kaggle and our modeling process was compared with another project on Kaggle called "Hotel Reservation Prediction ML|DL - 95% Accuracy". It's worth noting that in this project, the author tried out various machine learning models and eventually built a perfect model using neural networks with approximately 95% accuracy [2].

This project takes a different approach compared to [1] and [2]. Instead of using a single Machine Learning modelling technique, we attempted multiple approaches and identified the Random Forest Model (with a class weight) of around 98% AUC & Accuracy as the most suitable. Additionally, the data processing methods used in this project differ from the other two mentioned projects. We relied on AUC and Accuracy metrics to evaluate the performance of our models. In the following sections, we will delve into our techniques in greater detail.

# Data Exploration and Processing

When exploring data, the first step is to understand its structure. Our dataset contains 19 features and 36275 observations. The data spans from July 2017 to December 2018, covering one and a half years of bookings. All 19 features are significant, and the first column, Booking\_ID, is a unique identifier that we can discard. Unfortunately, the dataset doesn't include any demographic information about the customers. There are both numerical and categorical columns in the dataset. Numerical columns include "no\_of\_adults", "no\_of\_children", "no\_of\_weekend\_nights", "no\_of\_week\_nights", "required\_car\_parking\_space", "lead\_time", "repeated\_guest", "no\_of\_previous\_cancellations", "no\_of\_previous\_bookings\_not\_canceled", "avg\_price\_per\_room", "no\_of\_special\_requests", "arrival\_date", "arrival\_year", and "arrival\_month". The categorical columns include "type\_of\_meal\_plan", "room\_type\_reserved", "market\_segment\_type", and "booking\_status". Lastly, we assume that the prices in the dataset are in USD.

As we delved deeper into the dataset, we discovered instances where prices were set to $0. We investigated these cases by combining the arrival date, month, and year into a single date feature, and examining the pricing trends. This examination led us to identify a discrepancy in the data due to the entry of data for February 29th, 2018, which is not a valid date. This error could have occurred during data entry or due to corruption during data transfer or storage. To address this issue, we utilized a technique to correct the problem by analyzing the number of observations present on the day before February 29th (i.e., February 28th) and the day after February 29th (i.e., March 1st). The results of this analysis are presented below:

|  |  |
| --- | --- |
| Date | total |
| 2018-02-28 | 165 |
| 2018-03-01 | 61 |

Table 1 - Number of observations for the dates - Feb 28 & Mar 01

To ensure fairness during modelling, we updated the 37 records from February 29th to March 1st. Once the date issue was resolved, we analyzed the price trend and created a chart that displayed the minimum, average, and maximum prices for each day. Since there were multiple cases for a single day, we aggregated the data at three levels.

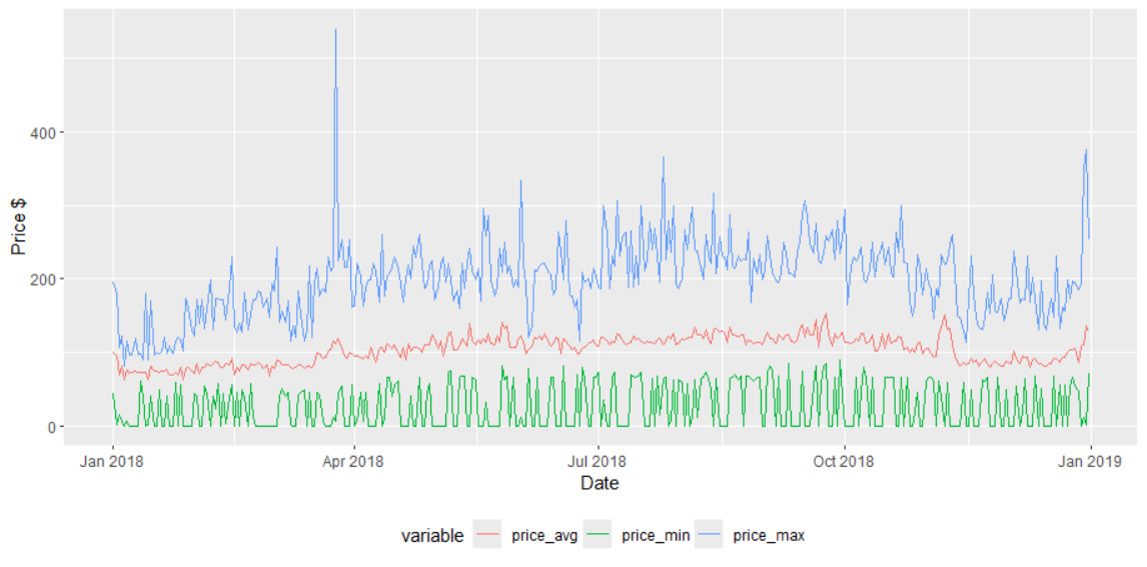


Figure 1 - Price Trend (2018)

It is clear from the graph that not all price levels are zero on a given date. This means that the instances where the price is zero (as indicated by the green line) are for specific booking options. We then looked into the market segmentation of these zero-price cases and found that they are either from the online or complimentary market segment. Based on this, we hypothesized that in the complimentary room or online offer option, one might be able to book with a zero price. To strengthen our hunch, we checked the number of cancellations and found that only six out of the 546 zero-price cases were cancelled. People who have free hotel rooms are less likely to cancel due to human nature.

In our data analysis, we have focused on the cancellation of booked rooms. We created a graph that plotted the total number of cancellations for each day in 2018. From the graph, we can understand that there were many cancellations during September, October, and November.

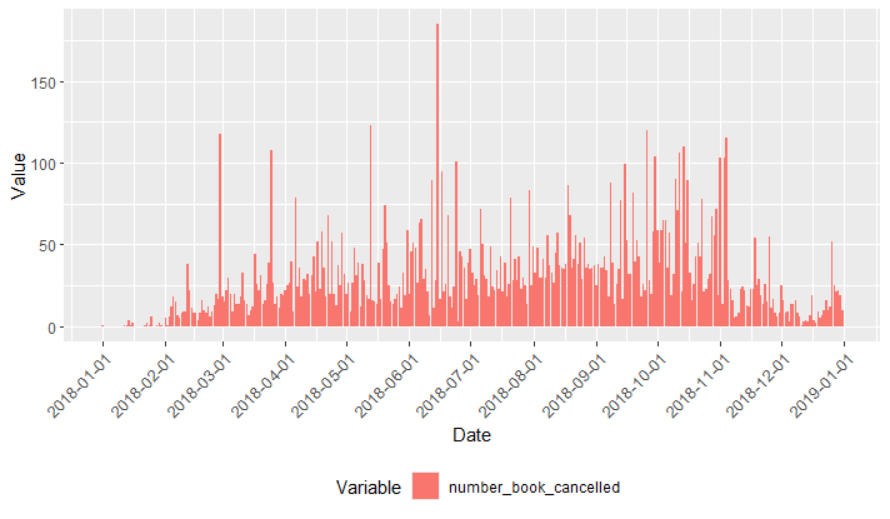


Figure 2 - Number of Cancellations (2018)

We focused on understanding the relationship between various variables and to perform this analysis, we had to remove the Booking\_ID column as it was a unique identifier. We simplified the outcome variable booking\_status to booking\_cancellation as our main interest was to predict whether a booking would be cancelled in the future. If the booking\_status variable has the value "Cancelled", it is redefined as "Yes" and "No" otherwise. Furthermore, we changed the data type of the features that were character to factor.

|  |  |
| --- | --- |
| Figure 3 | Figure 4 |
| Figure 5 | Figure 6 |
| Figure 7 | Figure 8 |
| Figure 9 | Figure 10 |
| Figure 11 | Figure 12 |
| Figure 13 | Figure 14 |
| Figure 15 | Figure 16 |
| Figure 17 | Figure 18 |
| Figure 19 | |
| Figure 20 | |

The mosaic plots in our analysis present a significant association between the booking cancellation variable and the other categorical variables. Furthermore, the Chi-Square test results and their significance level reinforce our statement. Some of the box plots indicate a marked association and difference in means between each of the cancellation categories (yes / no). Specifically, the average price, lead time, and number of special requests show noticeable associations. To test our hypothesis, we conducted two-sided tests, and the results further strengthen our hypothesis. Figure 20 clearly displays the test results and the quality of the features linked to our outcome variable.

We have identified that the features arrival\_date and arrival\_month have a periodic nature. To address this, we have transformed these features by mapping each value onto a point in a circle. The lowest value is placed next to the largest value in the circle. To create the "x" and "y" coordinates of each point, we have used sine and cosine transformations.

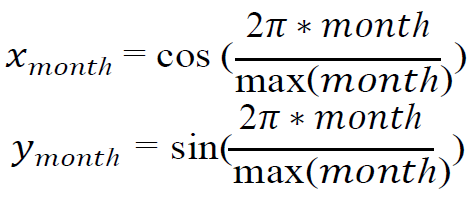


Figure 21

After applying feature transformation, the class balance of the dependent variable 'booking\_cancellation' was checked.

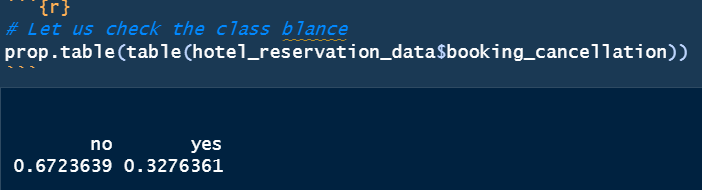


Figure 22

After reviewing the dataset, we identified an imbalance in the classes. Since our goal was to predict cancelled cases, we used class weights during the modeling process. We partitioned the dataset into a 90% training set and a 10% test set and calculated class weights based on the training data. To prepare for creating the neural network, we first created a validation dataset from the training data to fine-tune the network. Then, we one-hot encoded the categorical columns and normalized (standardized) the numerical columns for use in the binary classification neural network.

# Data analysis and experimental Results:

## Benchmark Model:

Before applying any modelling techniques, we conducted a basic heuristic calculation and established a benchmark model. We analyzed the feature lead time and concluded that based on the box plot, there is a significant statistical difference between the means of the lead\_time variable for each level of the categorical variable (i.e. booking\_cancellation).

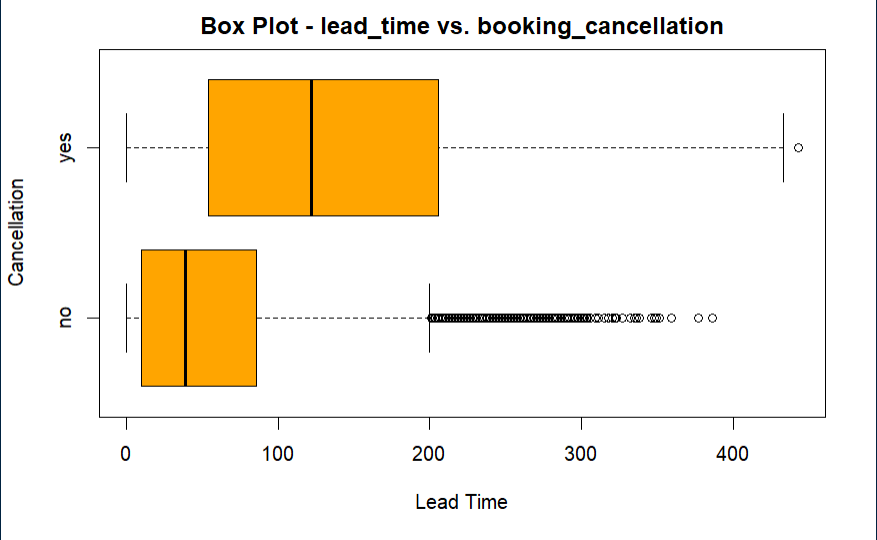


Figure 23 – Box Plot of the lead time across different categories of the booking\_cancellation variable of the train data.

The train data was used to determine the median lead time for cancelled cases. This median value was then used to classify cases in the test data. If the lead time in the test data was greater than or equal to the median, the case was classified as cancelled. If it was less than the median, the case was marked as not cancelled. The evaluation of the benchmark model is presented below.

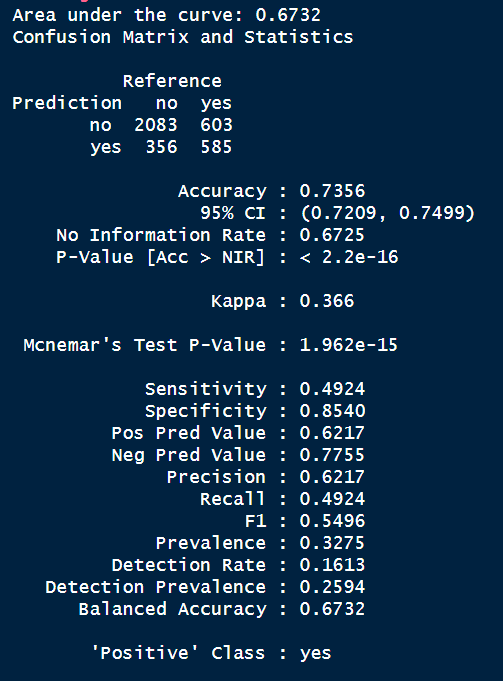


Figure 24

The next step involved constructing advanced models. The models employed were as follows:

* Regularized Logistic Regression
  + Lasso Regularization
* Random Forest
* Gradient Boosted
* Support Vector Machine
  + Linear
  + Radial
* Neural Network

Every model except the neural network was tuned using the caret package and 10-fold cross-validation. In most models, a custom tune grid was used.

## Sophisticated Models:

Below are the observations and results of the models.

### Regularized Logistic Regression – Lasso:

Performed hyperparameter tuning by adjusting lambda with 100 values between 10^-3 and 10^3.

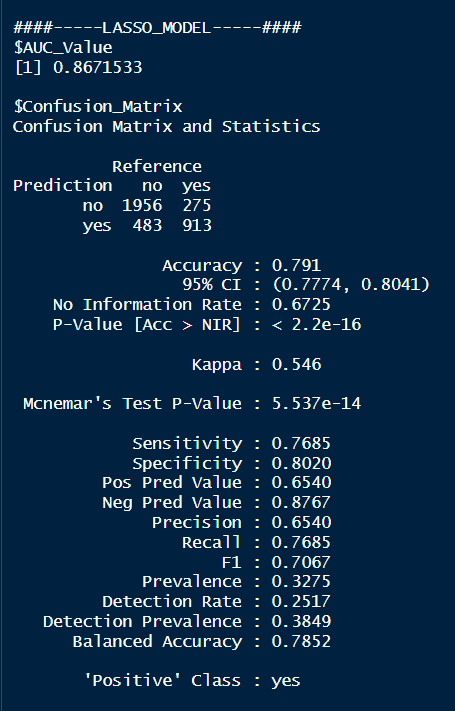


Figure 25

### Random Forest Model:

Performed manual hyperparameter tuning with ‘mtry’ picking any of these values (2, 4, 8).

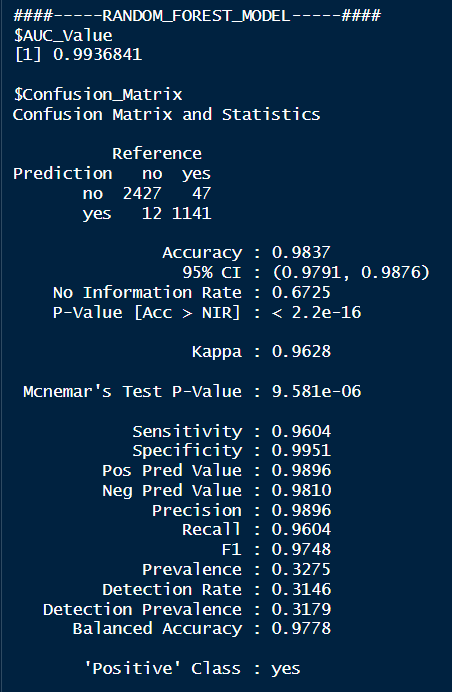


Figure 26

We also noted the important features.

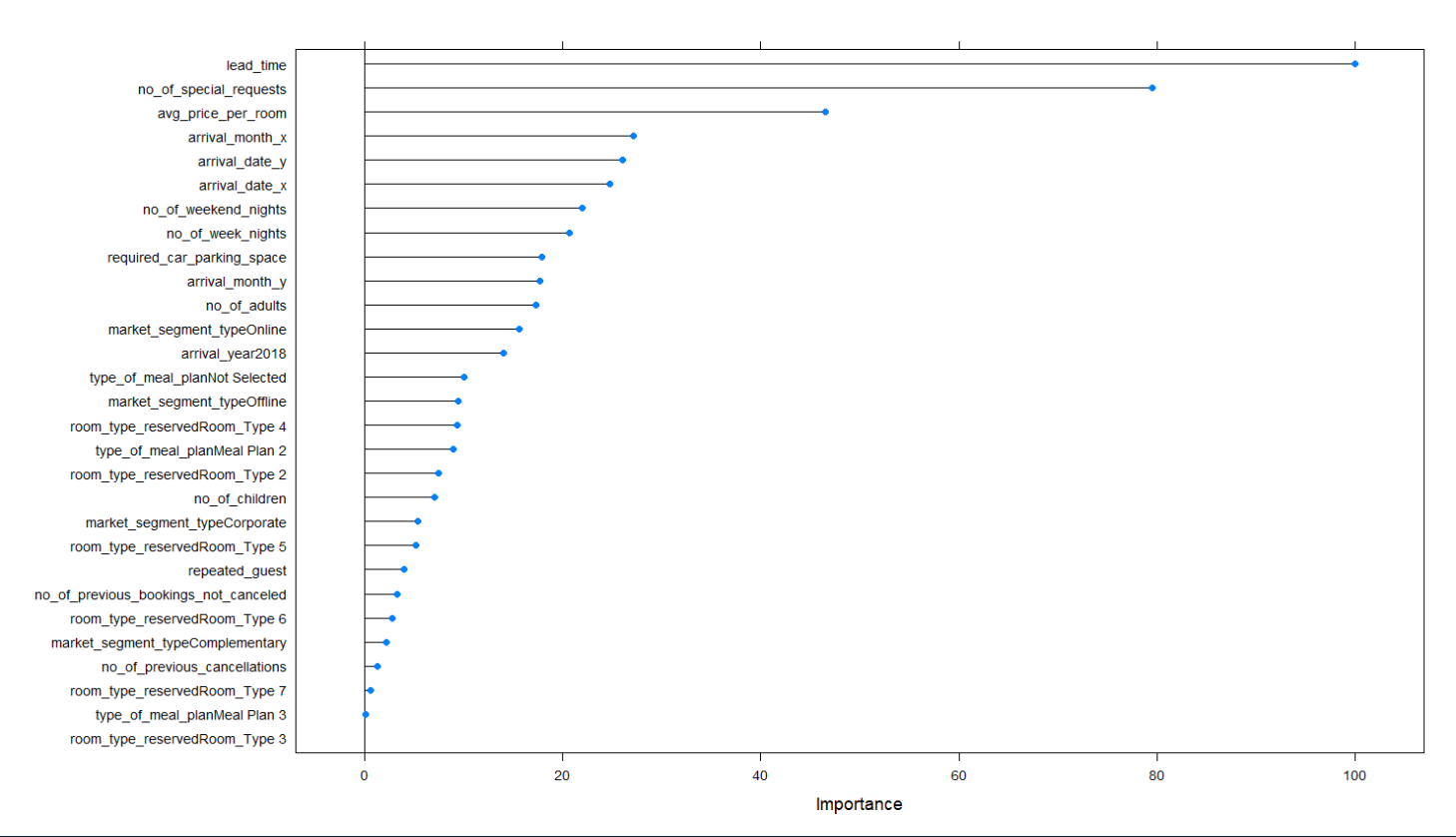


Figure 27

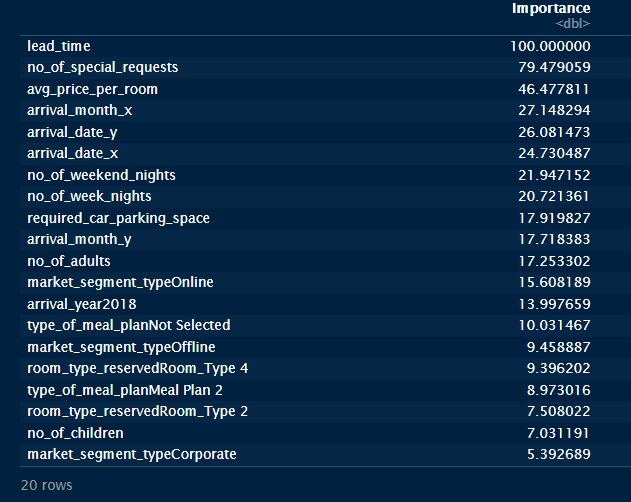


Figure 28

We have confirmed that our hypothesis regarding exploratory data analysis (EDA) was correct. During the EDA process, we observed a strong correlation between the lead time, average price, and number of special requests with our outcome variable. This suggests that these variables may be significant in predicting the outcome variable.

### Gradient Boost Model:

This model was automatically tuned.

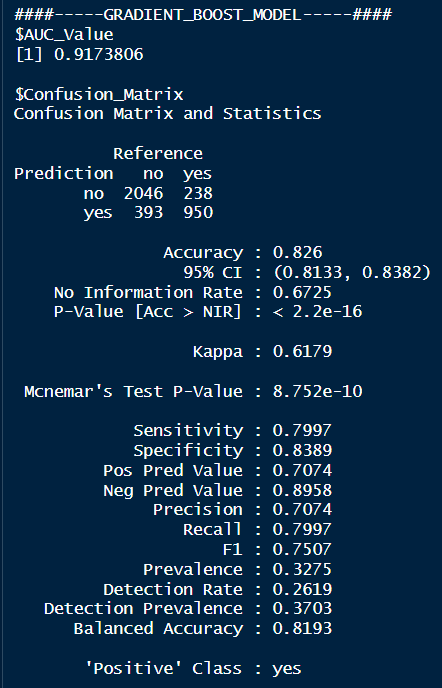


Figure 29

### Support Vectore Machine:

#### Linear Kernel:

Used a vector of 3 C values – (0.25, 0.5, 1) for tuning this model.

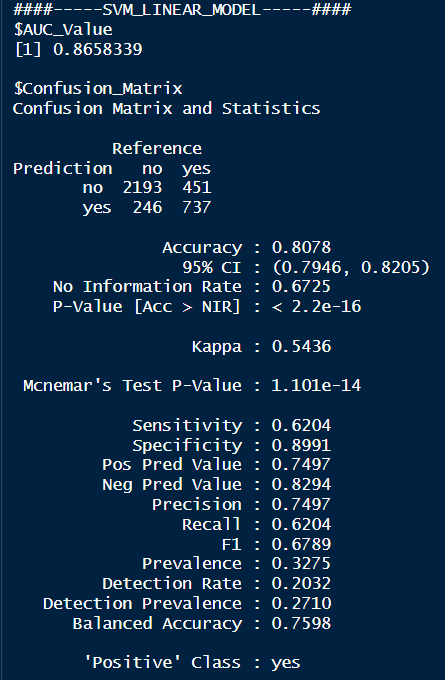


Figure 30

#### Radial Kernel:

This model was autotuned.

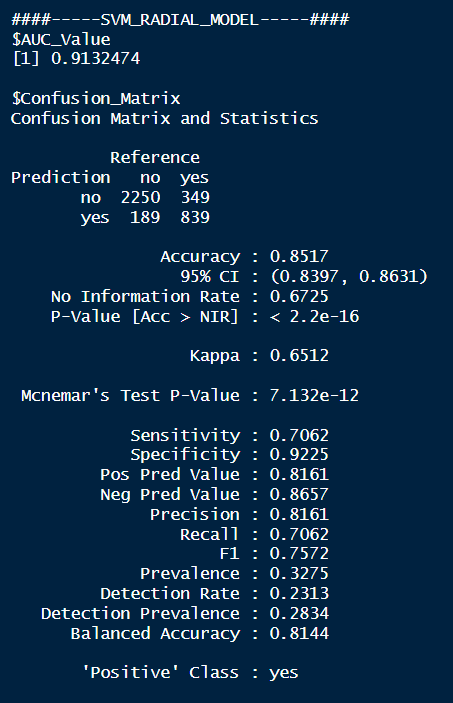


Figure 31

### Neural Networks:

Creating a dataset for a neural network model is different from creating a dataset for other types of models. Before fine-tuning the model, we first created a simple neural network without considering class weights. However, we noticed that this model mainly predicted the majority class. Below is the confusion matrix for this simple neural network model.

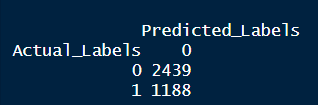


Figure 32

After tuning multiple models, I identified the best combination of parameters for the final model.

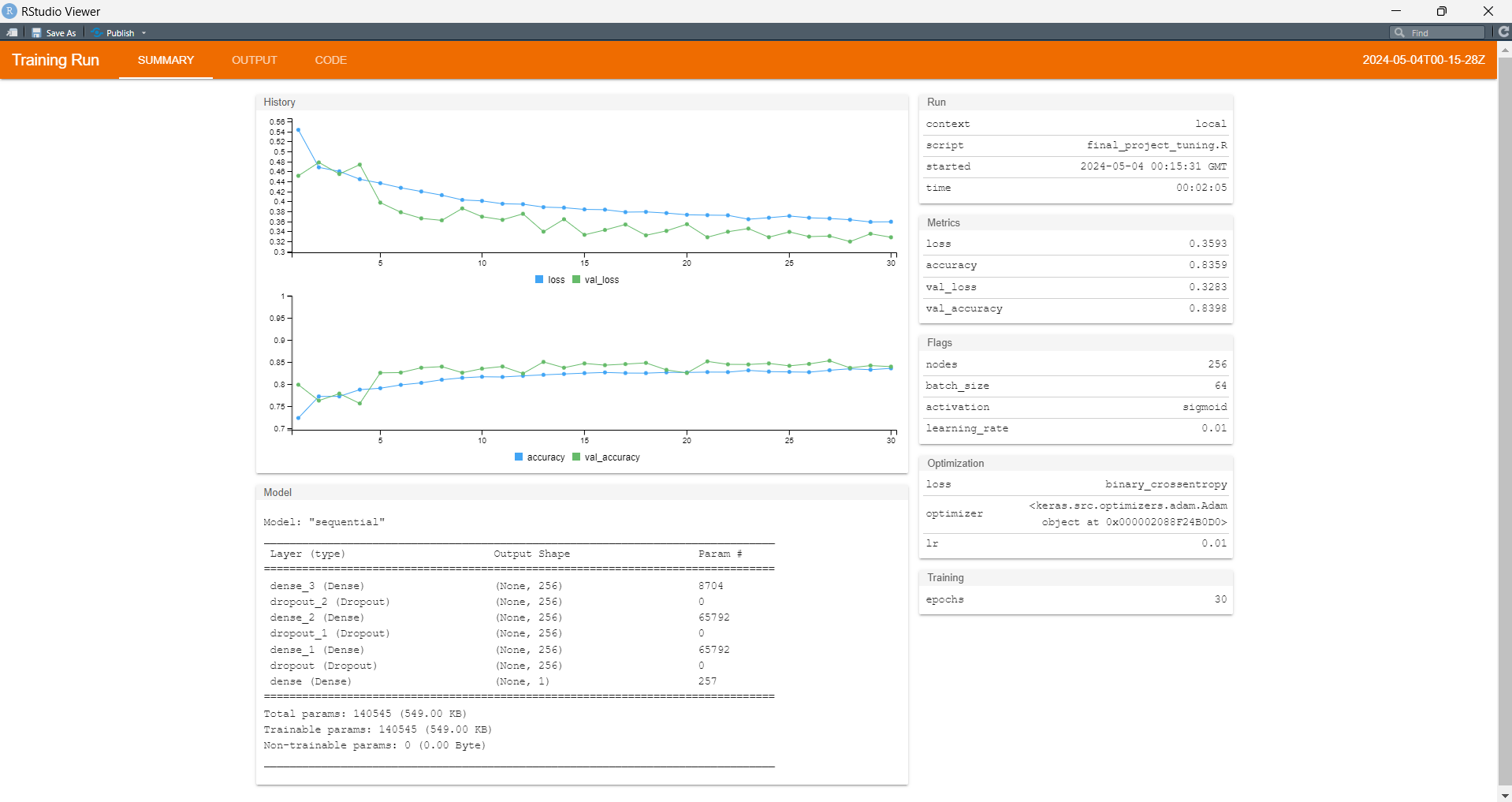


Figure 33

After configuring the parameters based on the above recommendation below are the results of the final model.

|  |  |
| --- | --- |
| Figure 34 | Figure 35 |

We evaluated all models based on AUC ROC and Accuracy and compiled the results..

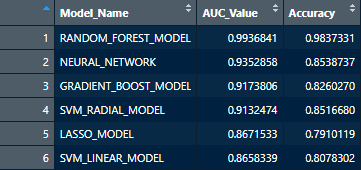


Figure 36

Although all the classifiers, namely Gradient Boost Mode, Neural Networks Model, and Random Forest Model, have shown spectacular performances, we have chosen the Random Forest Model due to its exceptional performance. However, in case of class imbalance, it is necessary to apply some form of class balancing technique during the modelling process, such as Upsampling/Downsampling or Class Weights based process. This ensures that the classifiers are generally successful and accurate.

# Conclusion:

The models used in this analysis have proven effective in predicting cancellations. Among all the models used, the Random Forest model produced the best results. Based on our analysis, we can conclude that the lead time, average price per room, and number of special requests are the major factors that impact room reservations.

We recommend that the organization who shared this dataset on Kaggle should check with customers 123 days after the reservation to confirm if they still want to keep their booking. This will help to reduce the number of cancellations and improve overall customer satisfaction.

# Bibiliogrpahy & References:

* Travelers’ rights: When reservations aren’t honored (usatoday.com) https://www.usatoday.com/story/travel/columnist/mcgee/2017/11/08/when-reservations-arent-honored/841685001/
* Predicting Hotel Booking Cancellations using Machine Learning Techniques" by João Lopes, Pedro Almeida, and Orlando Belo - Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model | IEEE Conference Publication | IEEE Xplore  
  https://ieeexplore.ieee.org/document/8260781/authors#authors
* Hotel Reservation Prediction ML|DL - 95% Accuracy (kaggle.com) https://www.kaggle.com/code/subhranilmondal12/hotel-reservation-prediction-ml-dl-95-accuracy#Cross-Validation-Method-To-Eliminate-Overfitting