Predicting Hotel Reservation Cancellations

Kevin Nguyen

Department of Statistics

University of Virginia

Data Source Location

- Sourced from two hotels in Portugal: one in the resort region of Algarve and the other in the city of Lisbon.

How Data Was Acquired

- Extracted from the hotels' Property Management System (PMS) databases using Structured Query Language (SQL), a programming language used to manage and manipulate data stored in databases.

Purpose of the Dataset

- Develop prediction models to classify a hotel booking's likelihood of being canceled.

Data Overview

- 36,275 observations with 17 variables.

Motivation

- Online hotel reservation channels have led to changes in customer behavior and an increase in cancellations, which can negatively impact hotel revenue.
- Hotels can formulate effective strategies to reduce cancellations, enhance their pricing and resource management policies, and tailor their services to suit the evolving preferences of customers by gaining insights into customer behavior and booking patterns.

Objectives

- Determine variables that are important in predicting whether a hotel guest will cancel their reservation or not.
- Evaluate various classification models that predict the likelihood of a guest canceling their hotel reservation and identify the best performing model.

Methodology

- Use logistic regression, k-nearest neighbors (KNN), and random forests to predict whether a guest will cancel their hotel reservation or not.

Cross-Validation

- All three classification models were evaluated using five-fold cross validation.
- Reduces risk of overfitting, especially when dealing with a high number of variables, and provides more reliable estimate of model's performance by using multiple evaluations.
- For KNN and random forests, a grid search was performed to find the optimal hyperparameters using cross-validation.

Majority Class Classification

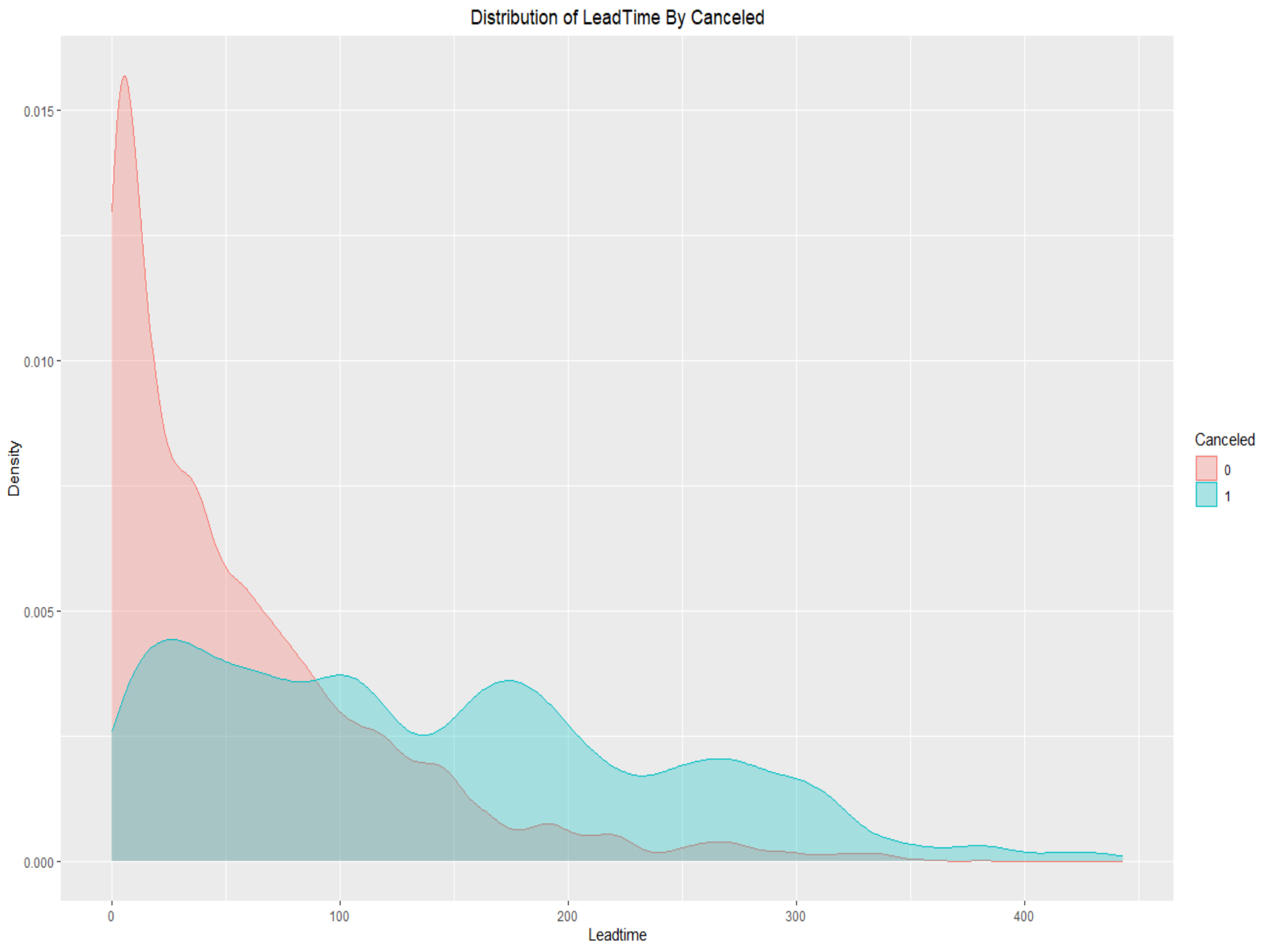
- Naïve method to serve as a baseline for comparison.
- In the training set, majority of the guests did not cancel their hotel reservations. As a result, the benchmark accuracy is 67.4%, which corresponds to the proportion of non-cancelled reservations in the test set. More specifically, 64.7% of the samples in the test set did not cancel their reservation.

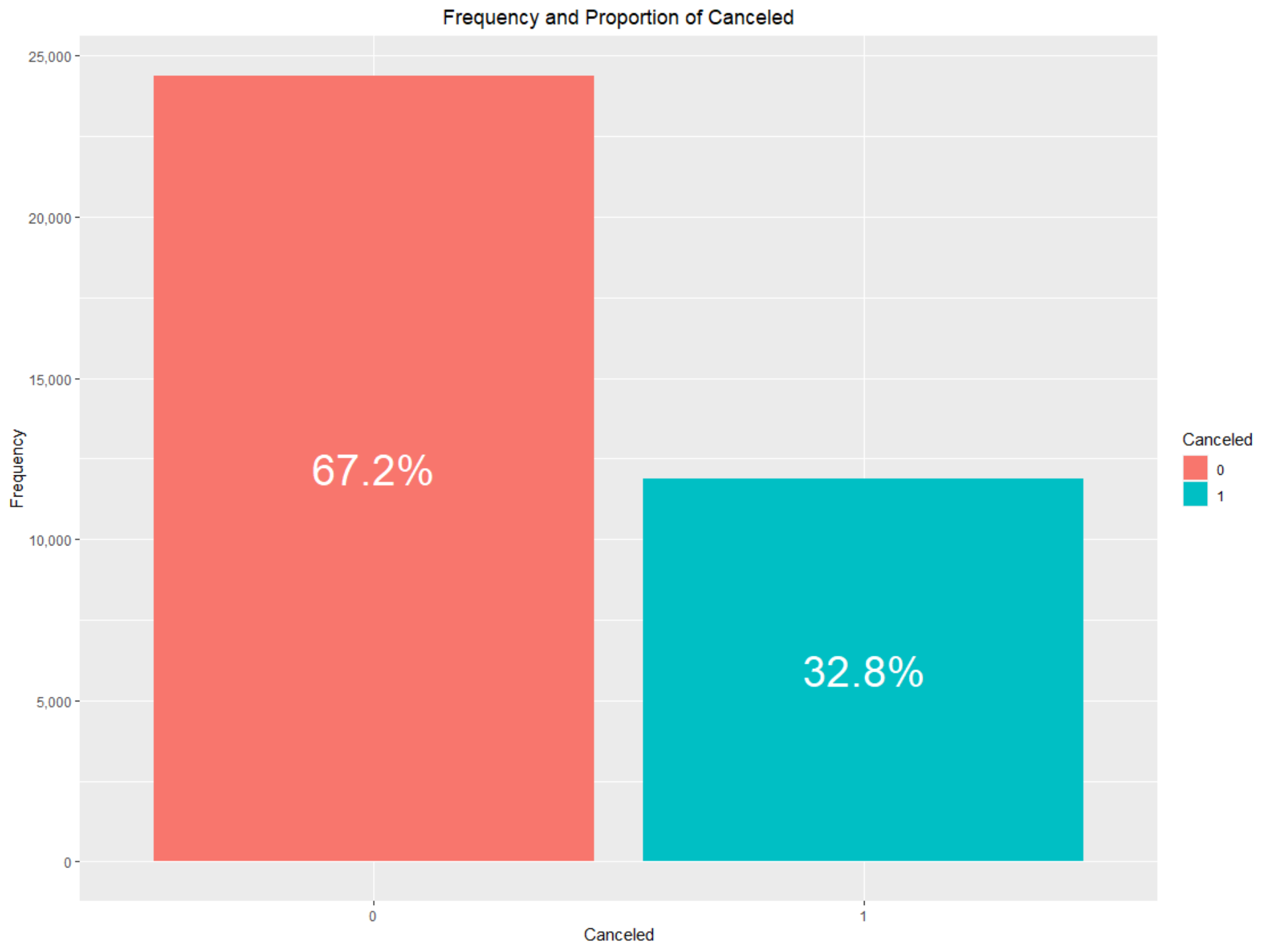
Model Fitting

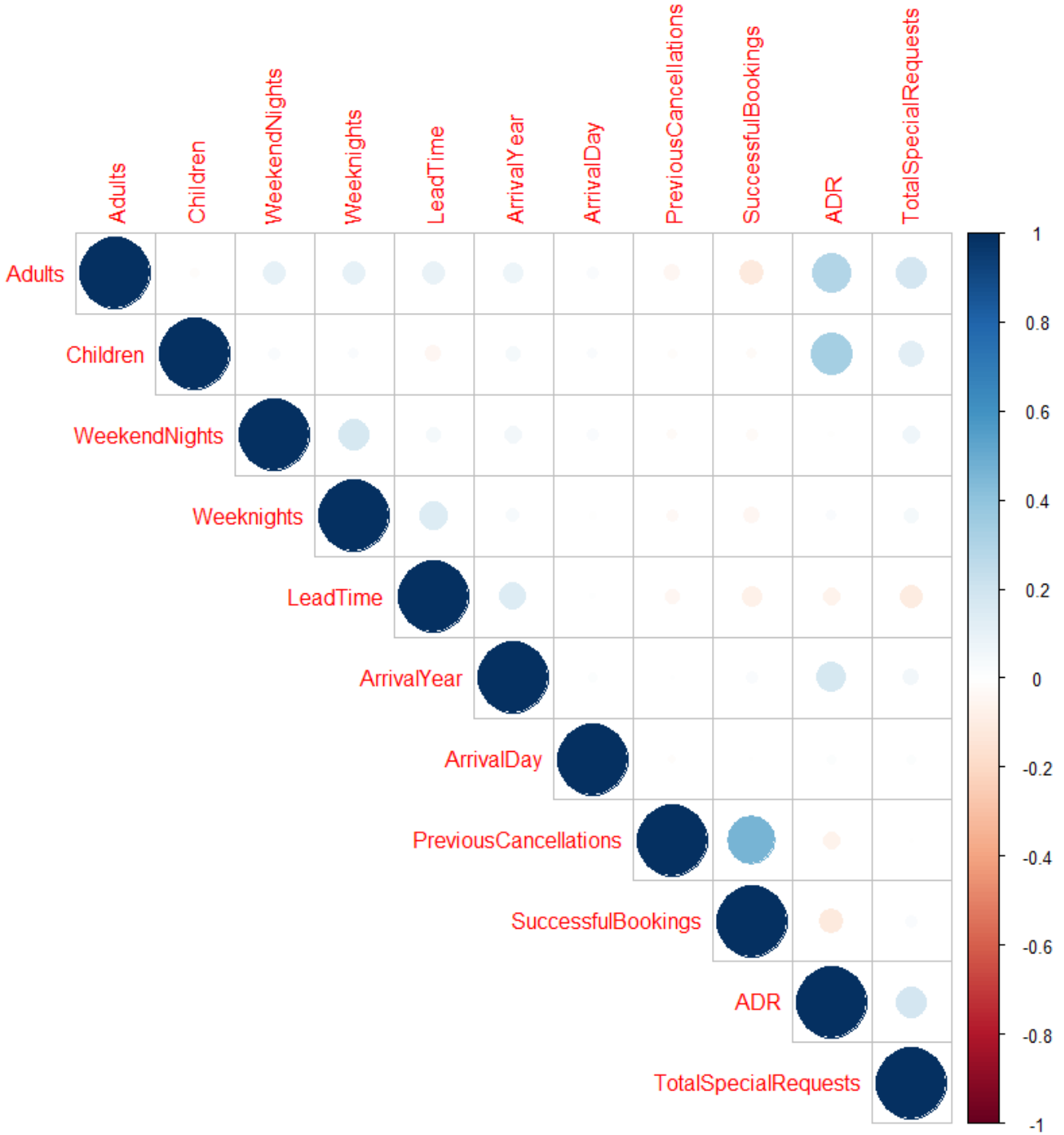
- For all three classification models, a full model was used.
- Variable selection was not performed as all variables were of interest.

Variable	Description
Adults	Number of adults
Children	Number of children
WeekendNights	Number of weekend nights the guest stayed or booked to stay at the hotel
Weeknights	Number of weeknights the guest stayed or booked to stay at the hotel
Meal	Type of meal booked
CarParking	Value indicating if the guest required a car parking space (1) or not (0)
ReservedRoomType	Type of room reserved by the guest
LeadTime	Number of days that elapsed between the booking date and the arrival date
ArrivalYear	Year of arrival date
ArrivalMonth	Month of arrival date
ArrivalDay	Day of arrival date
MarketSegment	Market segment designation
RepeatedGuest	Value indicating if the booking was from a repeated guest (1) or not (0)
PreviousCancellations	Number of previous bookings that were canceled by the guest prior to the current booking
SuccessfulBookings	Number of previous bookings not canceled by the guest prior to the current booking
ADR	Average Daily Rate (Sum of all lodging transactions / Total number of staying nights)
TotalSpecialRequests	Number of special requests made by the guest
Canceled	Value indicating if the booking was canceled (1) or not (0)

- ***All numerical variables are standardized.***

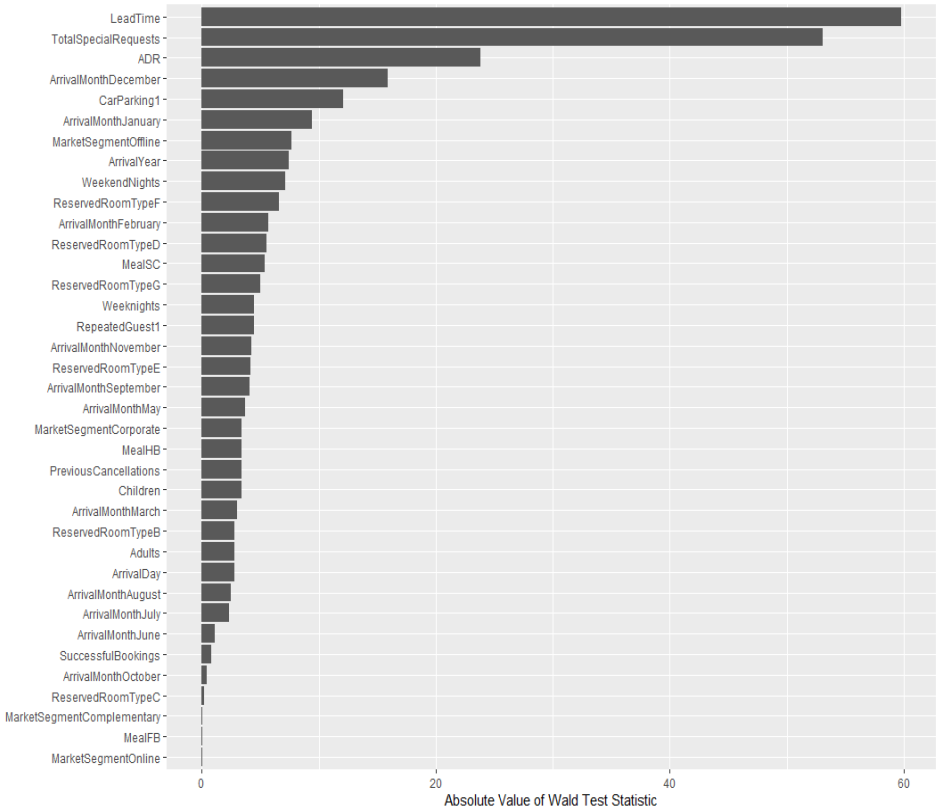

Distribution of LeadTime By Canceled
0.015
0.010
0.005
0.000
Density
0
100
200
300
400
Leadtime
Canceled
0
1

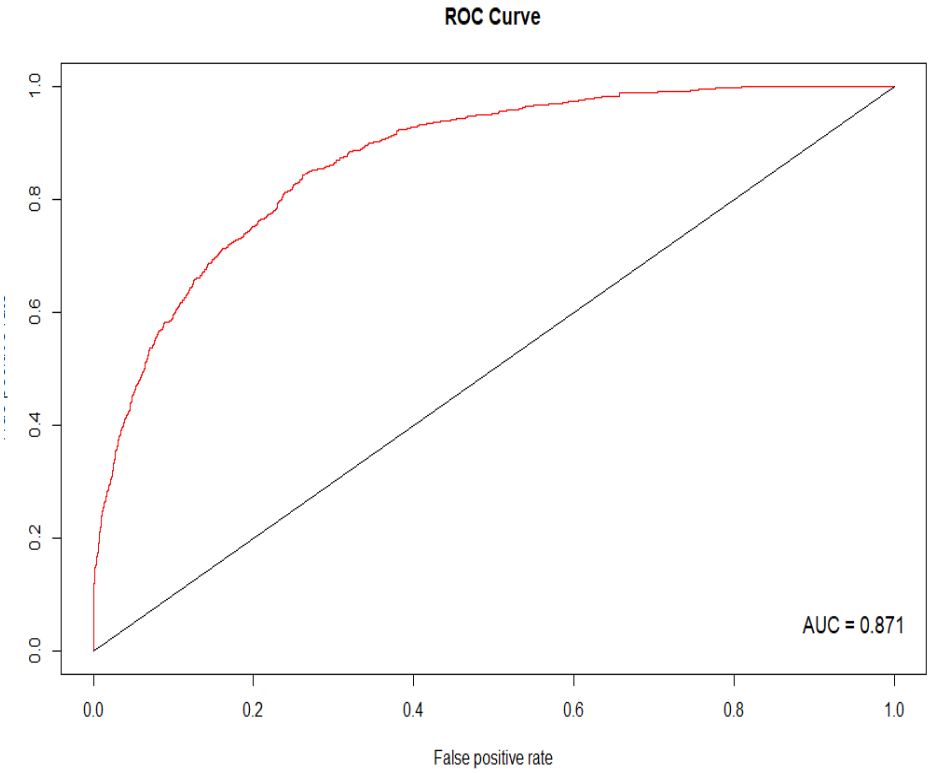

Frequency and Proportion of Canceled
25,000
20,000
15,000
10,000
5,000
0
Frequency
67.2%
32.8%
0
1
Canceled
Canceled
0
1


Adults
Children
WeekendNights
Weeknights
LeadTime
ArrivalYear
ArrivalDay
PreviousCancellations
SuccessfulBookings
ADR
TotalSpecialRequests
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

- Logistic regression is simple, interpretable, and is used widely for binary classification problems, making it a good starting point.

Accuracy Rate	False Positive Rate	False Negative Rate
80.2%	11.9%	36.2%

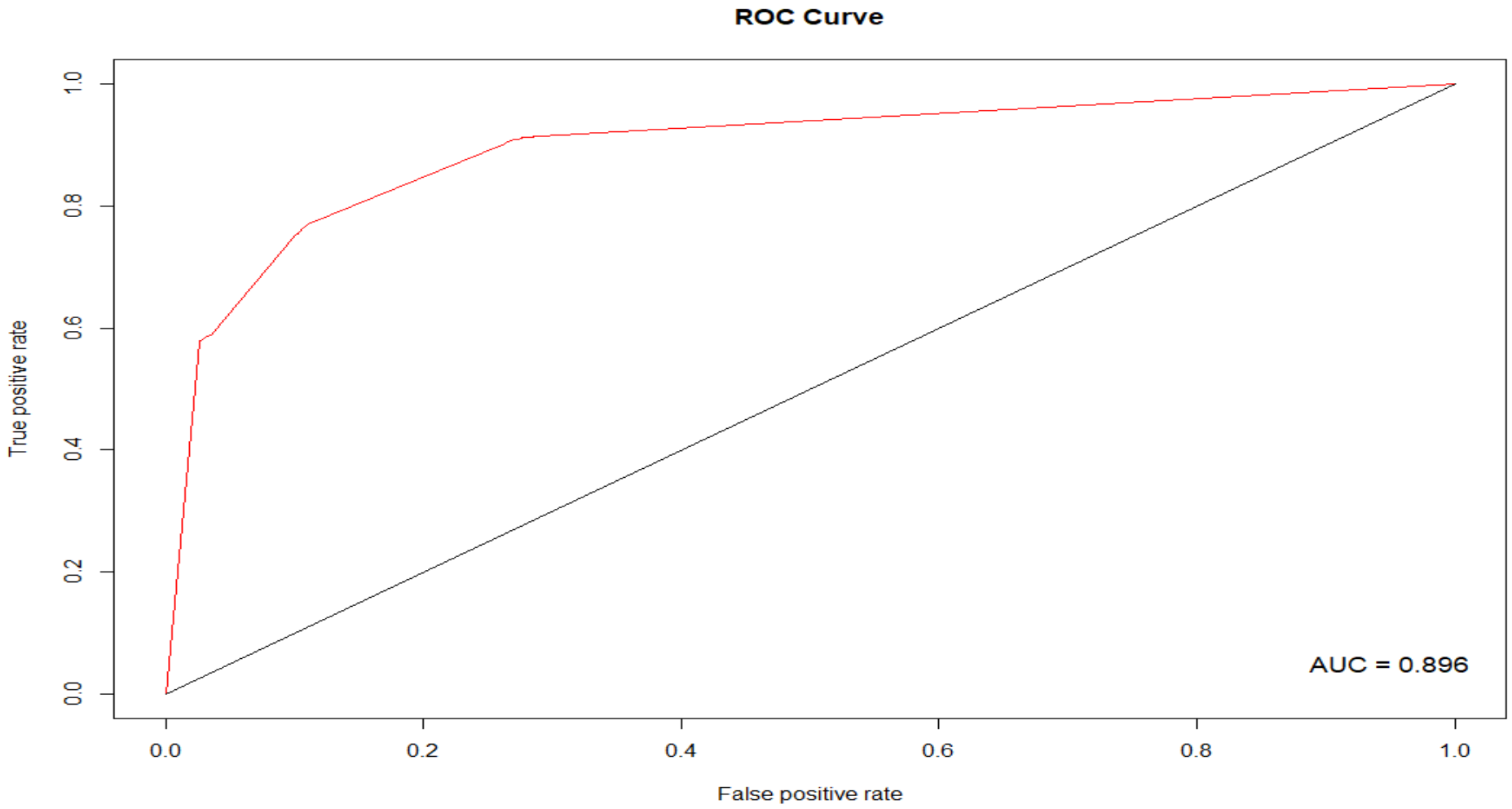




- Based on the absolute value of Wald test statistic, the top three most important variables are:
  1. Lead Time
  2. TotalSpecialRequests
  3. ADR

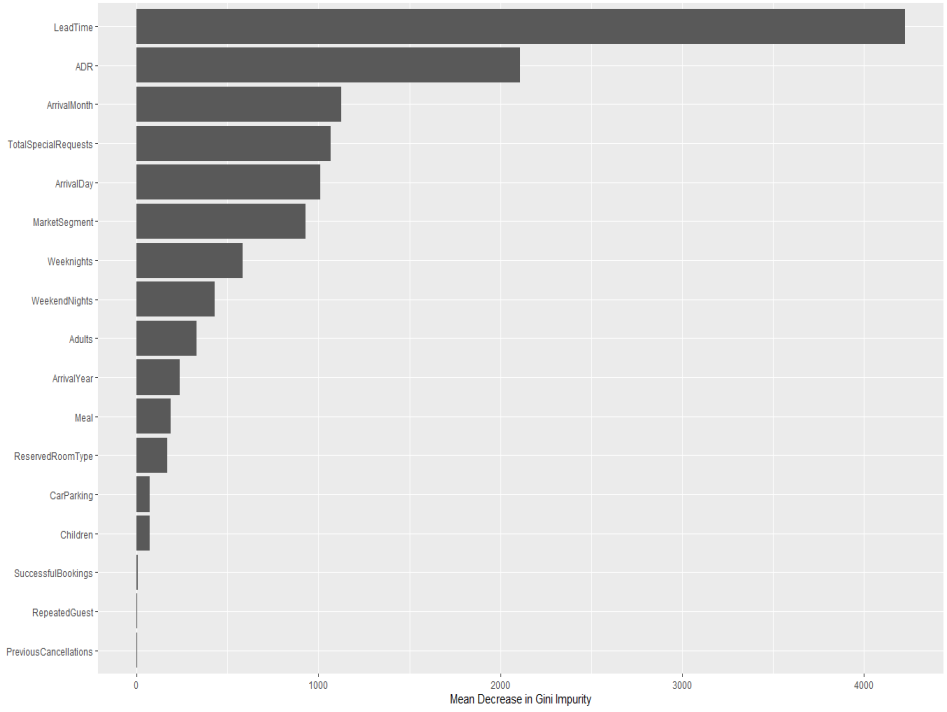
- KNN captures local patterns and relationships, make no assumptions of the distribution of the data, and is computationally efficient because it is a lazy learner.
- Based on a grid search of k values from 1 to 10, the optimal value of k is **k = 3**.

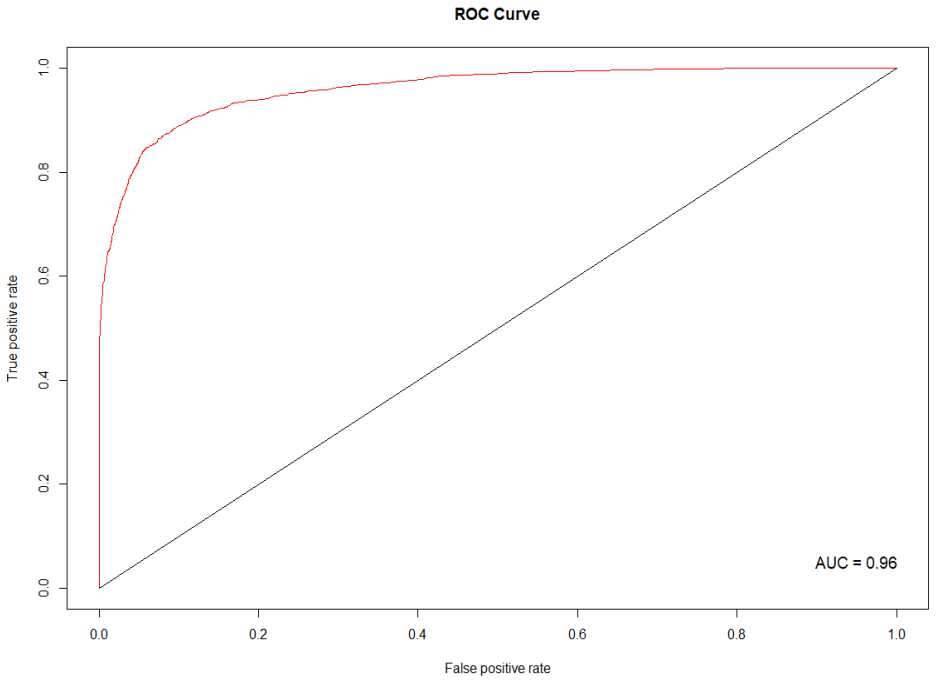
Accuracy Rate	False Positive Rate	False Negative Rate
85.2%	10.5%	23.6%



- Random forests can handle high-dimensional and noisy data well, prevents overfitting, and provides a built-in measure of variable importance.
- Based on a grid search of mtry values from 3 to 17 and ntree values from 100 to 1000, the optimal value for mtry and ntree is **mtry = 10** and **ntree = 1000**.

Accuracy Rate	False Positive Rate	False Negative Rate
91.0%	5.0%	17.4%





- Based on the mean decrease in Gini impurity, the top three most important variables are:
  1. Lead Time
  2. ADR
  3. Arrival Month

Summary

- Random forests is the best performing model.
- Based on the available data, LeadTime is the most significant variable in predicting whether a hotel guest will cancel their reservation or not.

Limitations

- Presence of imbalanced response variable can lead to biased model performance and inaccurate predictions.
- Dataset only includes information from hotels located in Portugal, which may not be sufficient to generalize findings and draw conclusions about hotels in other regions or countries.
- The captured data is limited to the period before the guests' arrival date, and thus, any subsequent changes or cancellations made during their stay are not reflected in the data.

Future Work

- Address class imbalance using undersampling or oversampling techniques.
- Explore other models like gradient boosting and support vector machine.
- Conduct variable selection to potentially reduce dimensionality of the dataset.