Ryan Canfield

Dr. Salemi

DSE6211

May 19th, 2024

Analytic Plan

**Introduction**

Given the business needs of ABC Hotels and a data set containing a list of 35,000 bookings it is known whether the booking was canceled. I would address the business needs by proposing to build a predictive model that can accurately estimate the probability of cancellation for each booking. Since the data given contains previous bookings where the cancellation is known I will be using supervised classification techniques to build a machine learning model. By identifying bookings with high risks, the hotel will be able to proactively go after and target these bookings, with tailored advertisements and/or offers, to prevent them from being canceled.

**Problem Statement**

Using the given data on booking rooms, which include various features such as arrival date, meal plan, room type, price, as well as many others, the task is to predict the likelihood of a booking being canceled. The predicted likelihood will be a continuous value between 0 and 1, representing the probability of cancellation.

**Objective**

The objective is to use supervised classification techniques to build a machine learning predictive model that can accurately estimate the probability of cancellation for each booking. So that they can try to prevent customers from canceling by targeting them with additional advertisements and/or offers.

**Approach**

The label or target for this supervised classification problem is the booking status column. The column contains two values either canceled or not canceled. These two values will be changed to numerical values. 0 to signify if a guest has canceled their reservation and a 1 will be used to signify if they have not.

The data preprocessing steps that are included to answer the business need of ABC Hotels will include the following steps. First, I will check for missing values found in the dataset and if there are any they will be removed unless there are many of them. I am also going to check for any duplicates and remove those as well. This is to keep the data as pure as possible to get the most accurate answers. Afterwards, we are going to introduce label encoding into all the variables that are not already numerical. An example of this would be in the type\_of\_meal\_plan column, instead of having meal\_plan\_1 and meal\_plan\_2 you would have a 1 replace meal\_plan\_1 and a 2 replace meal\_plan\_2 etc. Since the hotel wants us to use all the variables, we do not have to worry about dimension reduction like PCA, but if some columns have sparse data or other ones do not explain much of the variance they might be taken out. The next step, before we split our data, I will normalize the dataset making sure all the values in the dataset will fall between 0 and 1. This is to help our models and avoid overfitting if an outlier is greater than most of the rest of the data. Finally, we will split the dataset into two sets, a training set, and a testing set. The training set will have 80% of the data and is used to create our models, and the testing set will contain 20% of the data and will be used to evaluate how well the model performs. After doing some light statistical analysis we can see that most of the data is numerical. The columns will have to be changed so it can be run in the models. It is important to note that the summary statistics do not provide much since there are a lot of columns that only have just a couple of numbers so the means and different quarters do not provide much. I have also created 12 histograms just to get an idea of the distribution of data. The graphs are found at the bottom of the code appendix.

All the 16 features will be included when first creating the models. After an initial creation further analysis will be used to judge if some variables will be removed. Eight models will be initially included and evaluated. There will be two logistic regression models, two random forest models, one extreme boosting model, one support vector machine and two neural networks. The reason why there will be two of the same models is because R has multiple functions that can create the same model with some slight differences. So, after building all of them and fine tuning their hyperparameters. I will analyze each of them and choose the best one to give the client.

For analytical outcomes, the model will be able to produce the expected probability of cancellation based on certain features. This will give us an indication of some of the most important variables to look for when a customer books a room. This way when a high-risk target appears the hotel can do more for them to make sure they don’t cancel.  The models will also give us other evaluation metrics such as accuracy, precision, recall, F1-score, and ROC, but these just show the model’s performance and do not aid in the hotel’s decision making. For informational outcomes it gives the hotels better decision making on which customer to focus more of its resources on. Again, by prioritizing high risk bookings, they can use more of their resources to provide more tailored offers and advertisements in hopes of keeping people from calling them. By increasing this they will become more efficient and make more money since less people will cancel their books and get a refund. Overall, the machine learning model will serve as a valuable tool for ABC Hotels because it will proactively manage booking cancellations and optimize their business.

Code Appendix

Ryan Canfield

2024-05-20

## R Markdown

library(glmnet) #For a Logistic Regression model.

library(randomForest) #For a basic Random model.

library(ranger) # For an alternative Random Forest model.

library(xgboost) # For a Extreme Gradient Boosting model.  
library(e1071) # For a Support Vector Machine model.  
library(neuralnet) #For a basic Neural Network model.  
library(keras) #For a TensorFlow Neural Network model.  
library(tensorflow) #For a TensorFlow framework.  
library(reticulate) #For a python interface.  
library(caret) # For model training and evaluation.

library(ROCR) # For ROC analysis.

library(tidyverse) #For data manipulation.

library(ggplot2) #For data visualizations.   
library(RColorBrewer) #For coloring visuals.

# Reading the CSV file into a data frame  
df <- read.csv("..\\Dataset\\project\_data.csv")

# Get summary statistics of the data frame  
summary(df)

## Booking\_ID no\_of\_adults no\_of\_children no\_of\_weekend\_nights  
## Length:36238 Min. :0.000 Min. : 0.0000 Min. :0.0000   
## Class :character 1st Qu.:2.000 1st Qu.: 0.0000 1st Qu.:0.0000   
## Mode :character Median :2.000 Median : 0.0000 Median :1.0000   
## Mean :1.845 Mean : 0.1052 Mean :0.8105   
## 3rd Qu.:2.000 3rd Qu.: 0.0000 3rd Qu.:2.0000   
## Max. :4.000 Max. :10.0000 Max. :7.0000   
## no\_of\_week\_nights type\_of\_meal\_plan required\_car\_parking\_space  
## Min. : 0.000 Length:36238 Min. :0.00000   
## 1st Qu.: 1.000 Class :character 1st Qu.:0.00000   
## Median : 2.000 Mode :character Median :0.00000   
## Mean : 2.204 Mean :0.03093   
## 3rd Qu.: 3.000 3rd Qu.:0.00000   
## Max. :17.000 Max. :1.00000   
## room\_type\_reserved lead\_time arrival\_date market\_segment\_type  
## Length:36238 Min. : 0.00 Length:36238 Length:36238   
## Class :character 1st Qu.: 17.00 Class :character Class :character   
## Mode :character Median : 57.00 Mode :character Mode :character   
## Mean : 85.28   
## 3rd Qu.:126.00   
## Max. :443.00   
## repeated\_guest no\_of\_previous\_cancellations  
## Min. :0.00000 Min. : 0.00000   
## 1st Qu.:0.00000 1st Qu.: 0.00000   
## Median :0.00000 Median : 0.00000   
## Mean :0.02555 Mean : 0.02335   
## 3rd Qu.:0.00000 3rd Qu.: 0.00000   
## Max. :1.00000 Max. :13.00000   
## no\_of\_previous\_bookings\_not\_canceled avg\_price\_per\_room no\_of\_special\_requests  
## Min. : 0.000 Min. : 0.00 Min. :0.00   
## 1st Qu.: 0.000 1st Qu.: 80.30 1st Qu.:0.00   
## Median : 0.000 Median : 99.45 Median :0.00   
## Mean : 0.153 Mean :103.44 Mean :0.62   
## 3rd Qu.: 0.000 3rd Qu.:120.00 3rd Qu.:1.00   
## Max. :58.000 Max. :540.00 Max. :5.00   
## booking\_status   
## Length:36238   
## Class :character   
## Mode :character

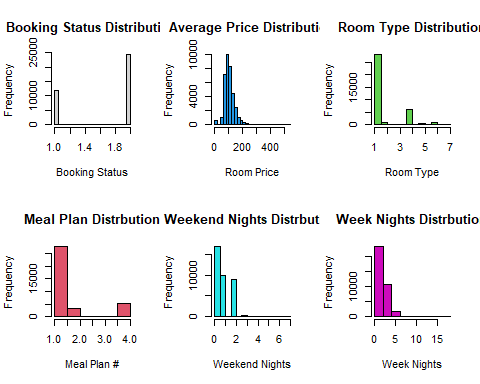
# Check for missing values in a data frame df  
missing\_values <- is.na(df)  
num\_missing <- sum(missing\_values)  
num\_missing

## [1] 0

library(forcats)  
  
# Loop through each column of the dataframe  
for (col in names(df)[-c(1, which(sapply(df, class) == "arrival\_date"))]) { # Exclude the first column, and Date columns  
 # Check if the column is non-numeric  
 if (!is.numeric(df[[col]])) {  
 # Perform label encoding  
 df[[col]] <- as.integer(factor(df[[col]]))  
   
 # Print a message indicating label encoding  
 cat("Label encoding applied to column:", col, "\n")  
 }  
}

## Label encoding applied to column: type\_of\_meal\_plan   
## Label encoding applied to column: room\_type\_reserved   
## Label encoding applied to column: arrival\_date   
## Label encoding applied to column: market\_segment\_type   
## Label encoding applied to column: booking\_status

# Looking at different histograms to get a better idea of the data  
par(mfrow = c(2,3))  
hist(df$booking\_status, xlab = "Booking Status", main = "Booking Status Distribution")  
hist(df$avg\_price\_per\_room, col = 4, breaks = 20, xlab = "Room Price", main = "Average Price Distribution")  
hist(df$room\_type\_reserved, col = 3, xlab = "Room Type", main = "Room Type Distribution")  
hist(df$type\_of\_meal\_plan, col = 2, breaks = 10, xlab = "Meal Plan #", main = "Meal Plan Distrbution")  
hist(df$no\_of\_weekend\_nights, col = 5, xlab = "Weekend Nights", main = "Weekend Nights Distrbution")  
hist(df$no\_of\_week\_nights , col = 6, breaks = 8, xlab = "Week Nights", main = "Week Nights Distrbution")



# Looking at different histograms to get a better idea of the data  
par(mfrow = c(2,3))  
hist(df$no\_of\_adults, col = 12, breaks = 4, xlab = "# of Adults", main = "Adult Distribution")  
hist(df$no\_of\_children, col = 7, breaks = 15, xlab = "# of Children ", main = "Children Distribution")  
hist(df$required\_car\_parking\_space, col = 8, breaks = 4, xlab = "No or Yes", main = "Do They Need a Parking Space?")  
hist(df$lead\_time, col = 9, breaks = 10, xlab = "Lead Time", main = "Distrbution of Lead Time")  
hist(df$repeated\_guest, col = 10, breaks = 4, xlab = "No or Yes", main = "Are They a Repeated Guest?")  
hist(df$no\_of\_previous\_cancellations, col = 11, breaks = 4, xlab = "No or Yes", main = "Have They Canceled Before?")

