PREDICTING CAB CANCELLATION

PROBLEM STATEMENT

An upcoming problem in 2013 – a cab company in Bangalore – yourcabs.com – featured online booking system

Drivers using this platform – some wouldn't show up/abrupt cancellation

Cancellation occurs without prior notice – customers left waiting and unhappy

OBJECTIVES

Build a model to predict whether or not the driver has cancelled the client's call by using data obtained from the taxi company

YourCabs.com will be able to manage its vendors and drivers better by providing them with up-to-date information about cancellations and reduce the dissatisfaction incurred from drivers' no-show.

DATA SOURCE INFORMATION

- DATA COMPOSITION: 10,000 clients; 18 input variables; 1 target variable (Cab cancellation). Therefore, the dataset consists of 10,000 bookings and It has a total of 19 variables.

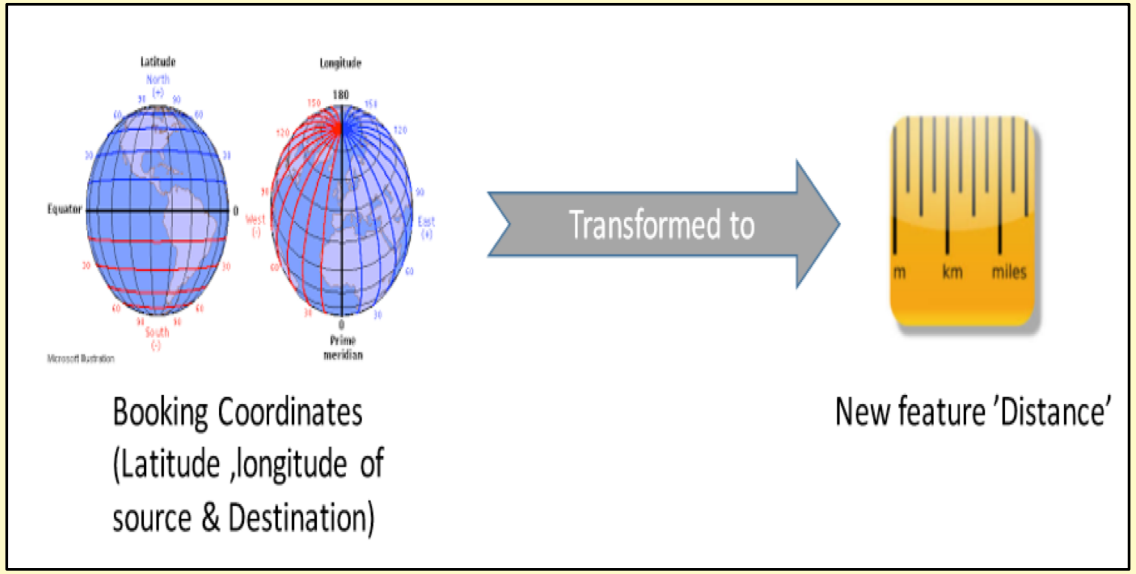
id - booking ID
user_id - the ID of the customer (based on mobile number)
vehicle_model_id - vehicle model type.
package_id - type of package (1=4hrs & 40kms, 2=8hrs & 80kms, 3=6hrs & 60kms, 4= 10hrs & 100kms, 5=5hrs & 50kms, 6=3hrs & 30kms, 7=12hrs & 120kms)
travel_type_id - type of travel (1=long distance, 2= point to point, 3= hourly rental).
from_area_id - unique identifier of area. Applicable only for point-to-point travel and packages
to_area_id - unique identifier of area. Applicable only for point-to-point travel
from_city_id - unique identifier of city
to_city_id - unique identifier of city (only for intercity)
from_date - time stamp of requested trip start
to_date - time stamp of trip end

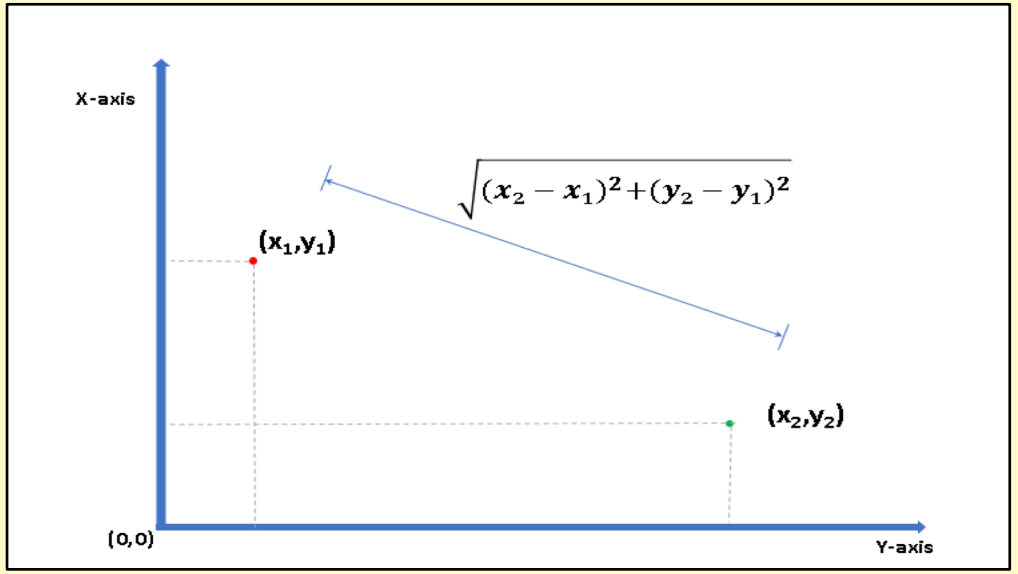
online_booking - if booking was done on desktop website
mobile_site_booking - if booking was done on mobile website
booking_created - time stamp of booking
from_lat - latitude of from area
from_long - longitude of from area
to_lat - latitude of to area
to_long - longitude of to area
Car_Cancellation (available only in training data) - whether the booking was cancelled (1) or not (0) due to unavailability of a car.

EXPLORATORY DATA ANALYSIS – DATA CLEANING

Step 1	identify number of missing values from each column and identify columns with a large number of na's/missing columns
Step 2	"package_id", "from_area_id", "to_area_id", "from_city_id", "to_city_id", "from_lat", "from_long", "to_lat", and "to_long“ found to have a large number of missing values
Step 3	identify inconsistent data types and convert them to a standard form; date entries Was found to be inconsistent and hence converted into standardized format “yyyy/mm/dd”
Step 4	the missing geocoordinates in the data set were filled using mean coordinates and concept of Manhattan distance
Step 5	implementation of feature engineering

FEATURE ENGINEERING : GPS DATA





Used the values of to_lat, to_long, from_lat and from_long to calculate effective distance between pick up and drop off points using the concept of Manhattan distance

FEATURE ENGINEERING : HANDLING THE DUMMY VARIABLES

- Vehicle_model_id
- travel_type_id
- Traditional booking

FEATURE ENGINEERING: VEHICLE_MODEL_ID

```
> taxi$vehicle_model_id[ taxi$vehicle_model_id %in% c(1,13,17,30,36,70,91) ] <- 100
> table(taxi$vehicle_model_id)

  12   24   28   65   85   87   89   90  100  101
7279  318  406  445  572  116  591   85   23  165
> taxi$vehicle_model_id[ taxi$vehicle_model_id %in% c(10,23,54,64,86,100) ] <- 101
> table(taxi$vehicle_model_id)

  12   24   28   65   85   87   89   90  101
7279  318  406  445  572  116  591   85  188
>
```

- The vehicles with frequency below 100 were categorized into one category -> 101 to reduce redundancy
- The vehicles with frequency above 100 are left as it is(most frequently used vehicles)

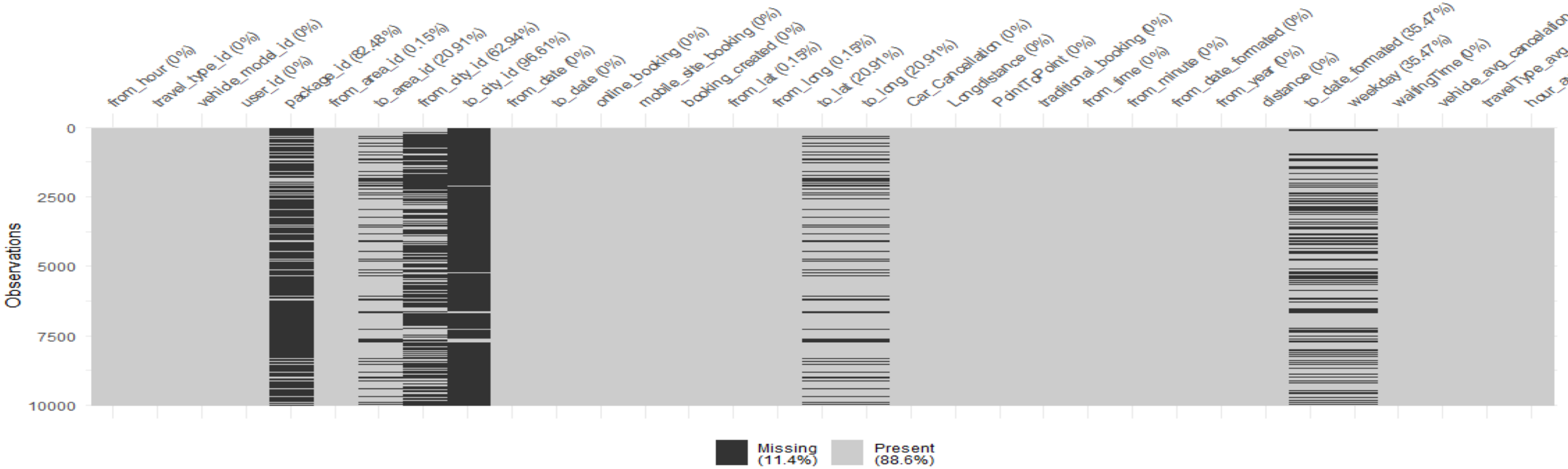
FEATURE ENGINEERING : TRAVEL_TYPE_ID

There are three categories under travel type id: 1=long distance, 2=point to point and 3=hourly rental

Hence, we converted the categorical variable travel_type_id into dummy variable: LongDistance and PointToPoint

taxi

booking_created	from_lat	from_long	to_lat	to_long	Car_Cancellation	Longdistance	PointToPoint
1/1/13 8:01	13.02853	77.54625	12.86980	77.65321	0	0	1
1/1/13 9:59	12.99987	77.67812	12.95343	77.70651	0	0	1
1/1/13 12:14	12.90899	77.68890	13.19956	77.70688	0	0	1
1/1/13 12:42	12.99789	77.61488	12.99474	77.60797	0	0	1
1/1/13 15:07	12.92645	77.61206	12.85883	77.58913	0	0	1
1/1/13 15:11	12.96298	77.71229	13.19956	77.70688	0	0	1
1/1/13 15:40	13.07746	77.60668	13.00446	77.56923	0	0	1
1/1/13 17:21	13.00042	77.67484	12.85773	77.78642	0	0	1
1/1/13 17:25	12.86980	77.65321	NA	NA	0	0	0
1/1/13 17:30	13.11084	77.60074	13.05845	77.64075	0	0	1
1/1/13 17:54	13.03726	77.58166	12.96937	77.64130	0	0	1
1/1/13 18:38	12.86980	77.65321	12.97677	77.57270	0	0	1
1/1/13 19:37	12.91281	77.60923	NA	NA	0	0	0
1/1/13 20:39	12.88963	77.60119	13.19956	77.70688	0	0	1
1/1/13 21:04	13.19956	77.70688	12.98628	77.73525	0	0	1
1/1/13 21:31	12.99067	77.65587	12.98712	77.56624	0	0	1
1/1/13 21:54	13.02853	77.54625	12.91280	77.58978	0	0	1
1/1/13 22:41	12.93331	77.56666	13.19956	77.70688	0	0	1
1/2/13 0:41	12.97896	77.67345	12.80257	77.70453	1	0	1
1/2/13 7:24	13.02239	77.59492	12.98275	77.61582	0	0	1



FEATURE ENGINEERING : HANDLING MISSING VALUES

- Deleted from_area_id, to_area_id, from_city_id, to_city_id,to_date and package_id due to the presence of too many missing values
- For the calculated distance column, we found out the mean value of distance and substituted the mean value in the missing rows
- Generated using vis_miss() function included in library(naniar)

FEATURE ENGINEERING : CALCULATION OF WAITING TIME

- Using the from_date, booking_created-> we extracted individual components of time(day, month, year, hour, minute, second) using the function strptime() and carried out respective date difference calculations using the function difftime() to calculate a more meaningful variable Waiting Time.

- NOTE: The functions strptime() and difftime() are included in the Library("stringr")

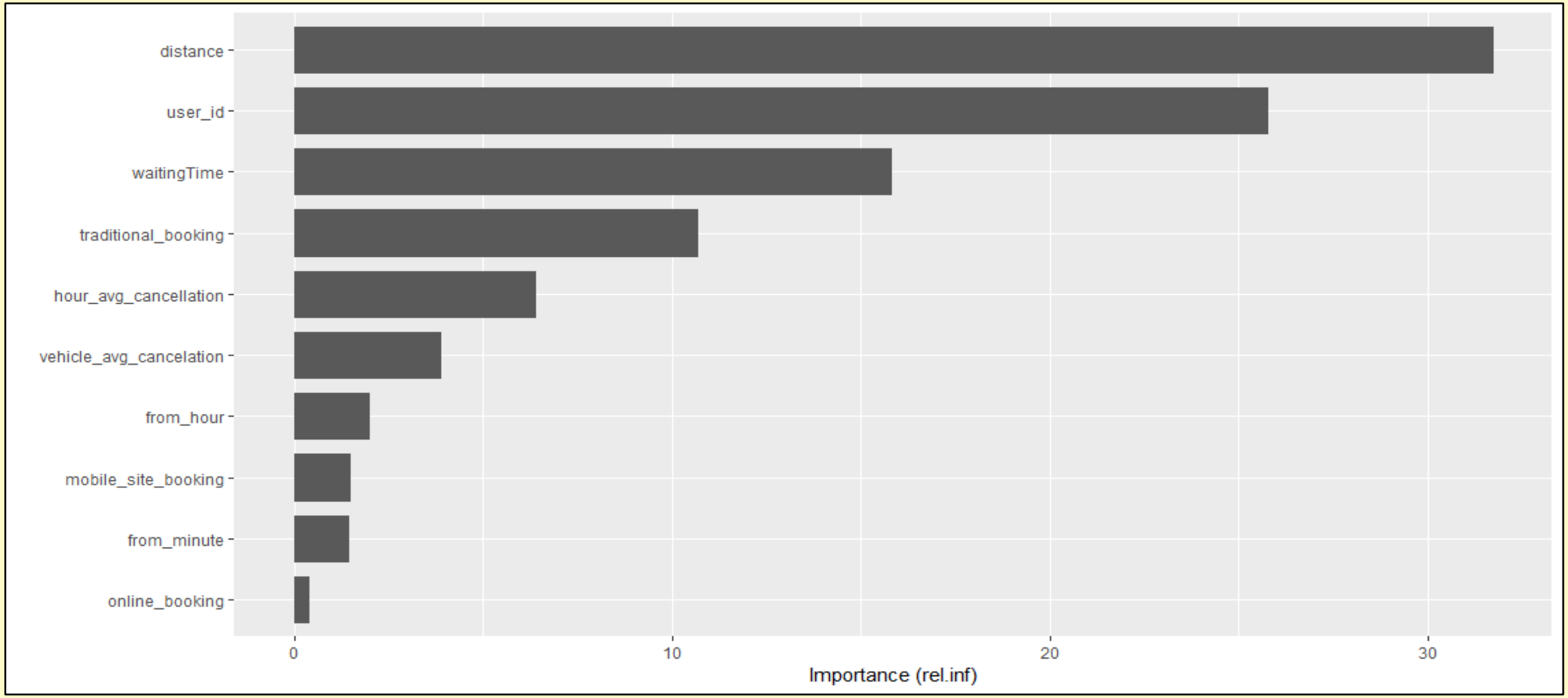
FEATURE ENGINEERING : AVERAGE CANCELLATION BY VEHICLE_MODEL_ID

- Using the different vehicle_model_id's, a table showing the average rate of cancellation for different vehicle_id's was generated.

A tibble: 9 x 5

vehicle_model_id	Count	vehicle_avg_cancelation	online	mobile_site_booking
<dbl>	<int>	<dbl>	<dbl>	<dbl>
1	12	7279	0.0846	0.352	0.0446
2	24	318	0.00314	0.355	0.0409
3	28	406	0.0542	0.350	0.0271
4	65	445	0.0404	0.306	0.0315
5	85	572	0.00524	0.329	0.0175
6	87	116	0.0517	0.422	0.0172
7	89	591	0.129	0.431	0.0677
8	90	85	0	0.388	0.0353
9	101	188	0.00532	0.309	0.0319

RESULT OF FEATURE ENGINEERING

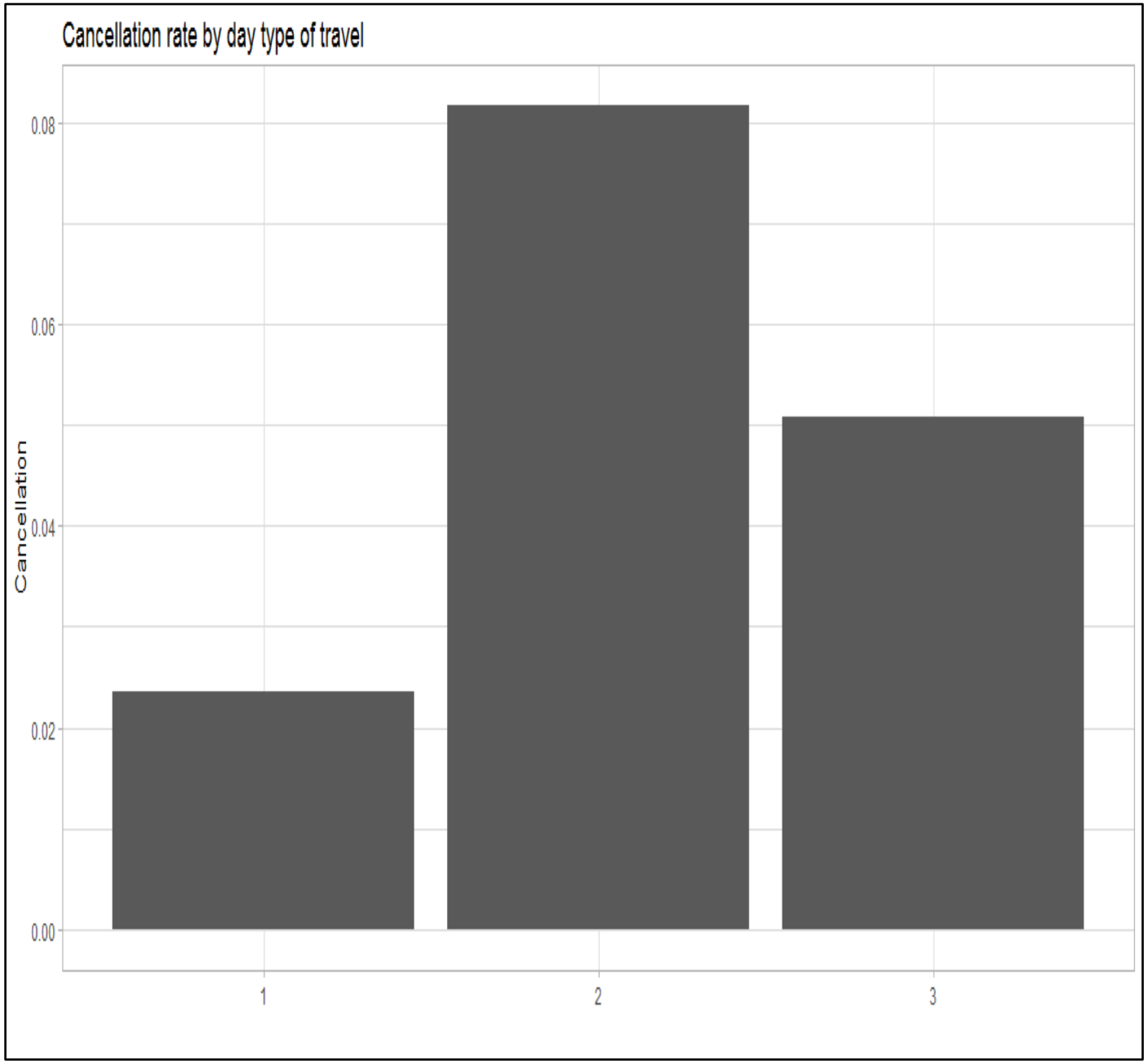


RESULT OF FEATURE ENGINEERING

ANALYSIS USING DIFFERENT PLOTS

(1) CANCELLATION RATE BY TYPE OF TRAVEL:
1- Long Distance
2- Point to Point
3- Hourly rental

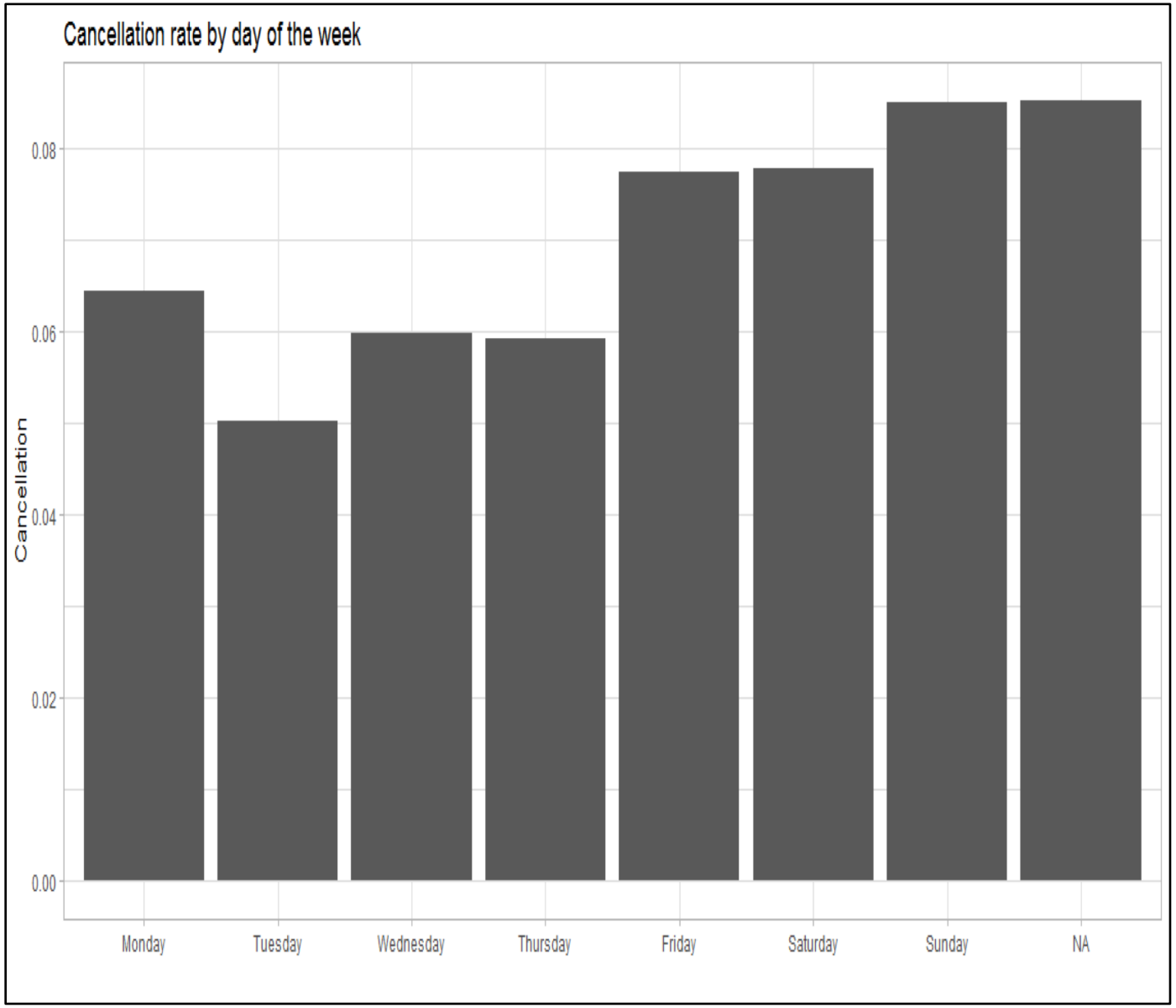
- INFERENCE: POINT TO POINT HAS HIGHEST RATE OF CANCELLATIONS



ANALYSIS USING DIFFERENT PLOTS

(2) CANCELLATION RATE BY DAY OF THE WEEK

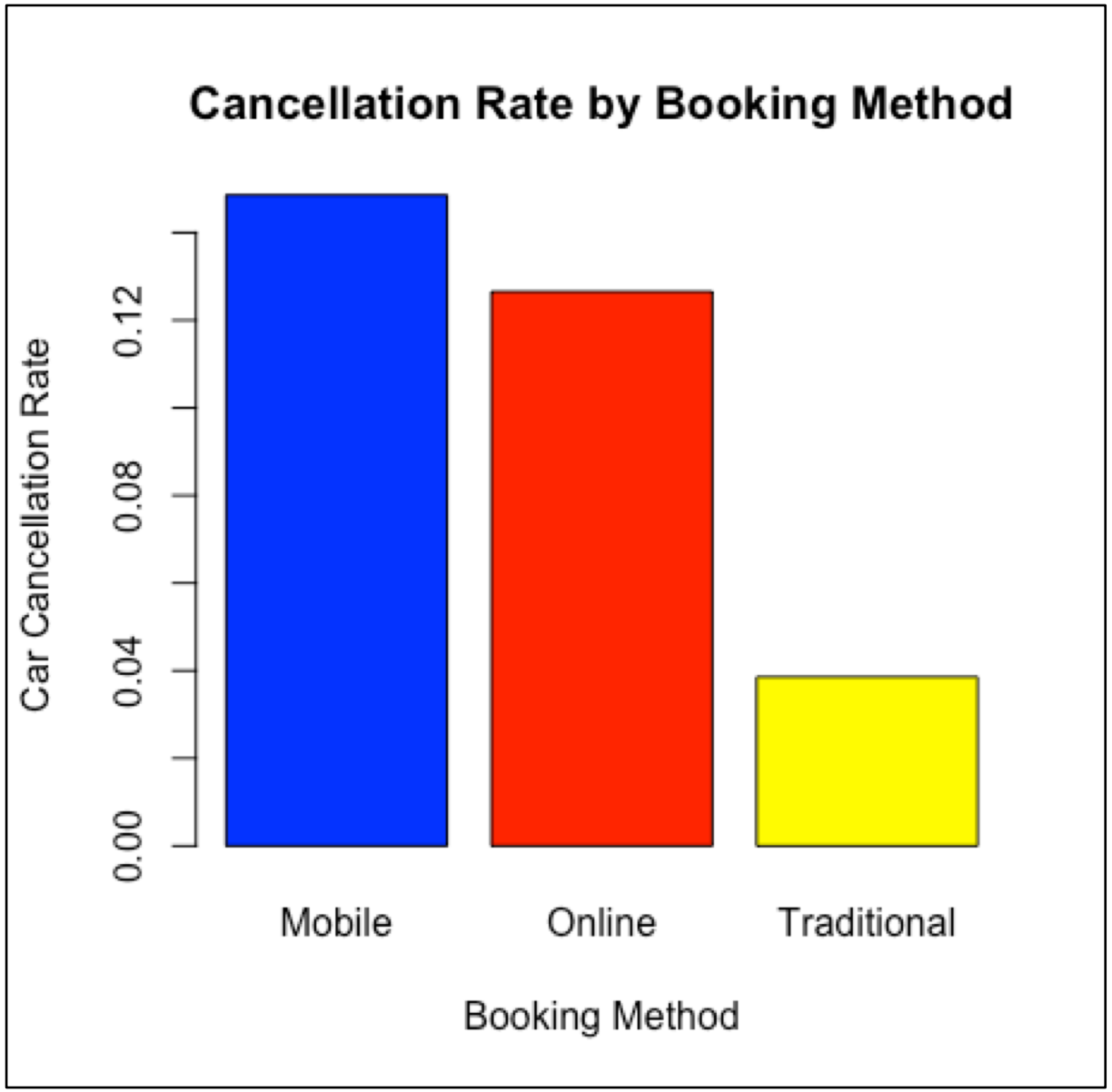
- INFERENCE: SUNDAY HAS HIGHEST RATE OF CANCELLATIONS

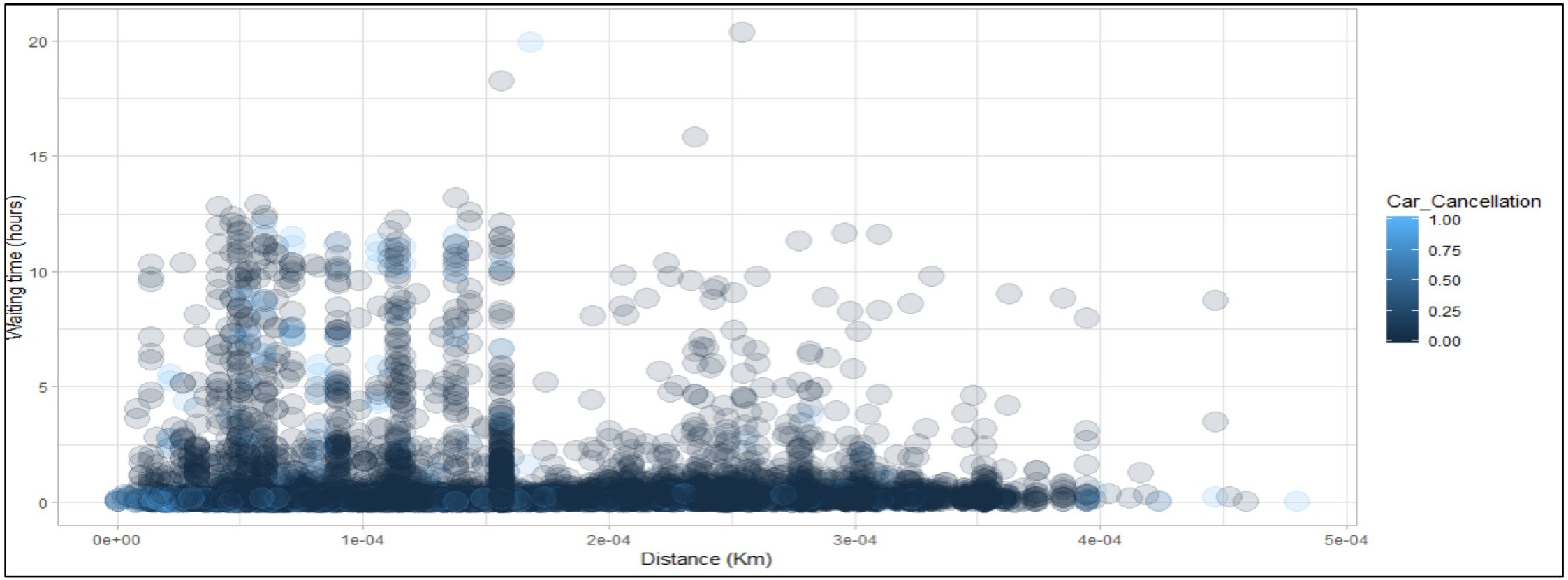


ANALYSIS USING DIFFERENT PLOTS

(3) CANCELLATION RATE BY BOOKING METHOD

- INFERENCE: MOBILE BOOKING HAS HIGHEST RATE OF CANCELLATION RATES

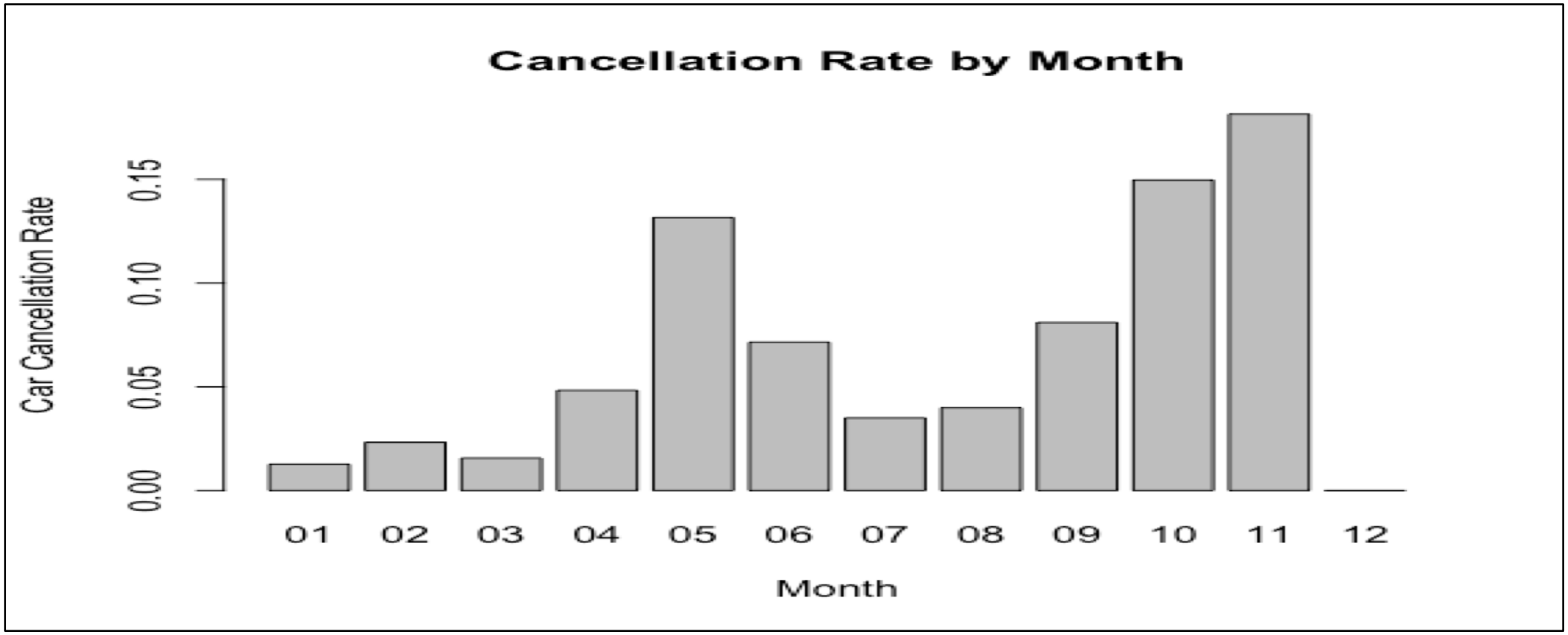




ANALYSIS USING DIFFERENT PLOTS

(4) CANCELLATION RATE BY WAITING TIME AND DISTANCE

- INFERENCE: RIDES WITH LONGER WAITING TIME AND FURTHER DISTANCE HAVE HIGHER RATES OF CANCELLATION



ANALYSIS USING DIFFERENT PLOTS

(5) CANCELLATION RATE BY MONTH

- INFERENCE: Month 11 has highest number of cancellation rates

DESCRIPTIVE STATISTICS

- From the table showing frequency of cancelled vs uncancelled bookings:

```
> table(taxi$Car_Cancellation)

   0    1
9257  743
```

- Very unbalanced data set

MODELLING TECHNIQUES USED

(1) Logistic Regression

(2) Decision Tree

(3) Random forest

DATA PARTITIONING USED :

70 % training set, 30% validation set – try to account for overfitting of data

LOGISTIC REGRESSION

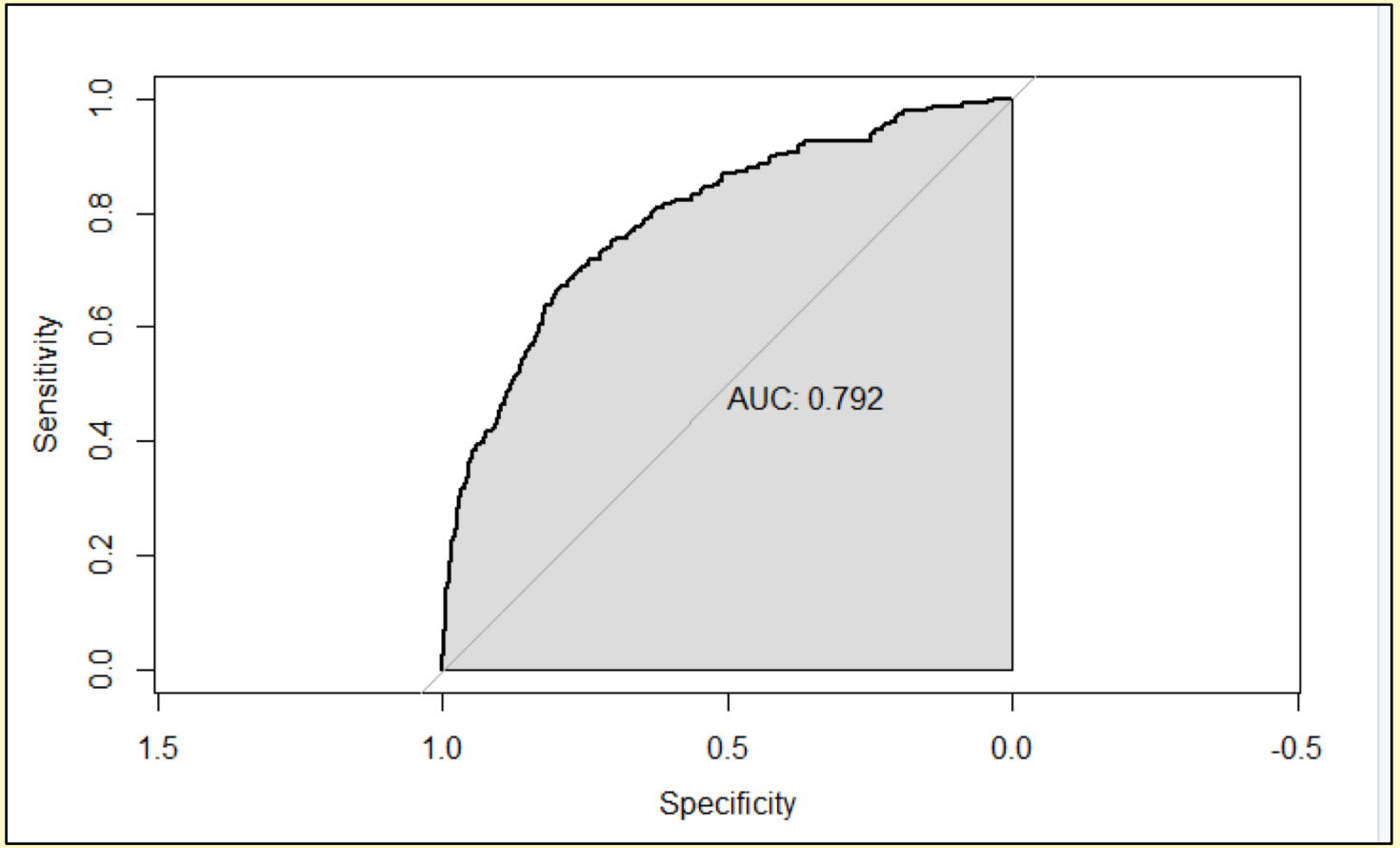
```
Call:
glm(formula = Car_Cancellation ~ ., family = "binomial", data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2922  -0.4129  -0.2790  -0.1798   3.1439

Coefficients: (2 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -4.399e+00  4.144e-01 -10.615  < 2e-16 ***
from_hour           1.573e-02  7.729e-03   2.035  0.04189 *
user_id             3.548e-05  5.102e-06   6.955 3.51e-12 ***
vehicle_model_id   -5.707e-03  1.897e-03  -3.008  0.00263 **
travel_type_id     -3.661e-02  1.467e-01  -0.250  0.80293
online_booking      1.360e+00  1.074e-01  12.664  < 2e-16 ***
mobile_site_booking 1.507e+00  1.924e-01   7.833 4.78e-15 ***
Longdistance       -1.319e+00  4.910e-01  -2.687  0.00722 **
PointToPoint               NA         NA      NA       NA
traditional_booking        NA         NA      NA       NA
from_minute        -1.364e-03  2.651e-03  -0.515  0.60676
distance           -6.768e+00  6.777e-01  -9.986  < 2e-16 ***
waitingTime         1.636e-05  7.241e-06   2.259  0.02386 *
hour_avg_cancellation 1.138e+01  1.223e+00   9.304  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3689.1  on 6999  degrees of freedom
Residual deviance: 3177.9  on 6988  degrees of freedom
AIC: 3201.9

Number of Fisher Scoring iterations: 6
```

summary(m1)

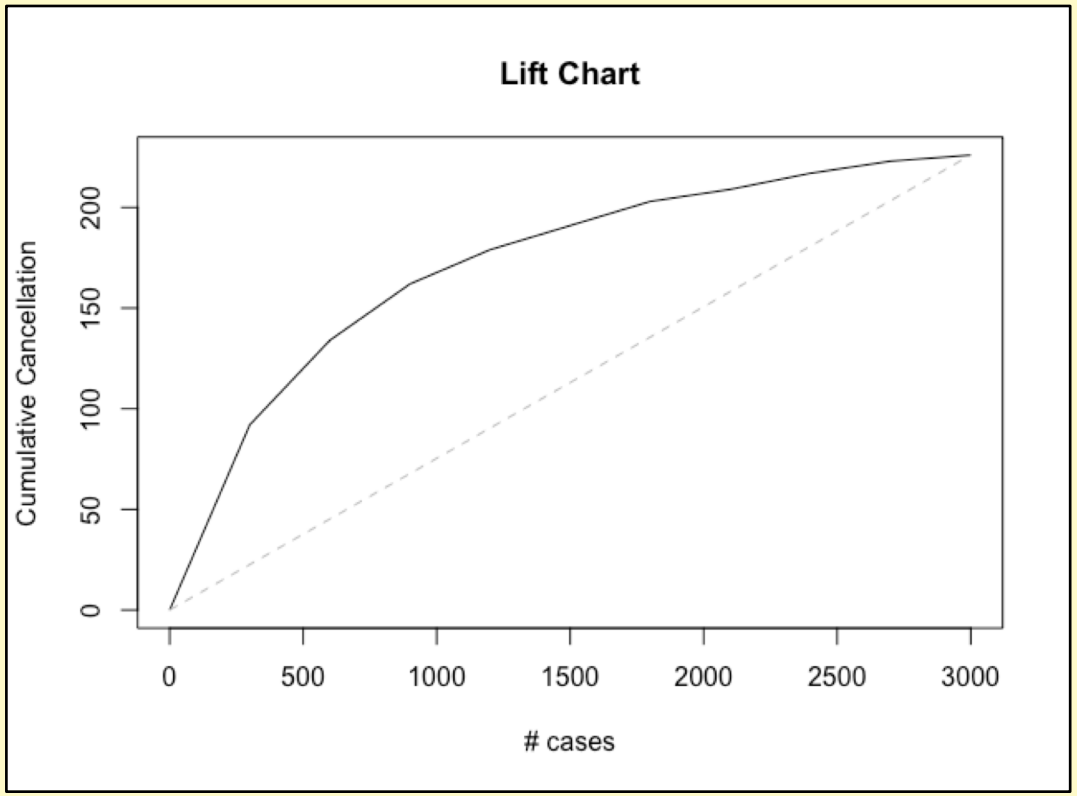
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2774  223
         1    0    3

               Accuracy : 0.9257
                 95% CI : (0.9157, 0.9348)
    No Information Rate : 0.9247
    P-Value [Acc > NIR] : 0.4351

                  Kappa : 0.0243

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.01327
            Specificity : 1.00000
         Pos Pred Value : 1.00000
         Neg Pred Value : 0.92559
             Prevalence : 0.07533
         Detection Rate : 0.00100
   Detection Prevalence : 0.00100
      Balanced Accuracy : 0.50664

       'Positive' Class : 1
```

Confusion matrix

LOGISTIC REGRESSION : ROC CURVE



LOGISTIC REGRESSION: LIFT CHART



Solid line - expected performance under predictive model

Dotted Line - expected response without using a predictive model

Lift = (Expected Response In A Specific Lot of Prospects Using Predictive Model) / (Expected Response In A Random Lot of Prospects Without Using Predictive Model)

LOGISTIC REGRESSION : INFERENCE

List of important predictors from summary(m1) table:

- user_id
- booking type -> mobile site booking or online booking
- distance
- average rate of cancellation/hour

- Accuracy of logistic regression model: 92.57 %
- The imbalance in the confusion matrix table is because of the imbalance in the data set
- Area under ROC curve : 0.792

DECISION TREE

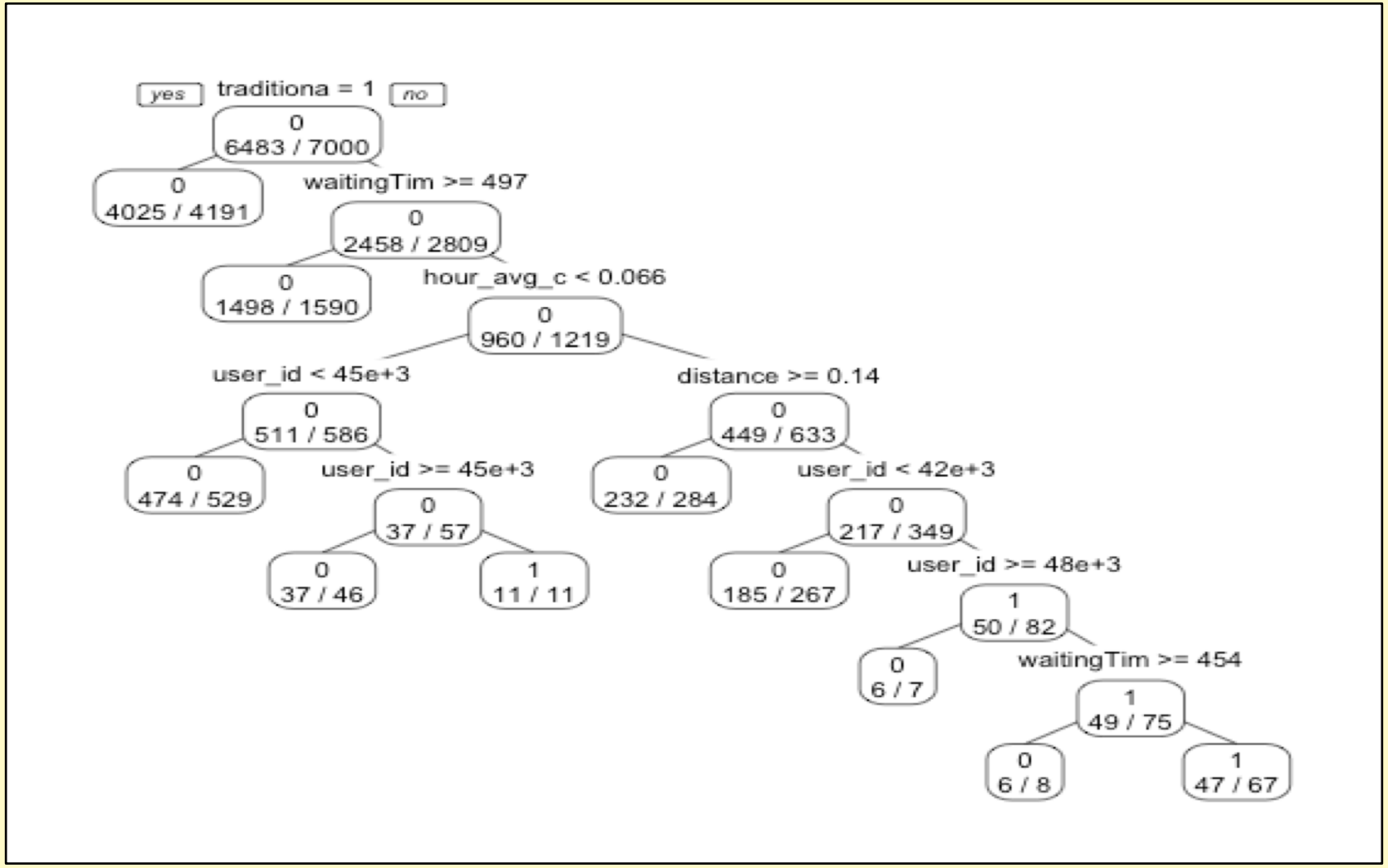
CP	nsplit	rel error	xerror	xstd
0.00696325	0	1.00000	1.00000	0.042325
0.00531915	9	0.92650	0.99613	0.042249
0.00483559	15	0.89362	0.99033	0.042136
0.00386847	19	0.87427	1.00193	0.042362
0.00338491	23	0.85880	1.01354	0.042587
0.00322373	27	0.84526	1.01354	0.042587
0.00221056	30	0.83559	1.02708	0.042848
0.00193424	39	0.81431	1.08124	0.043868
0.00096712	49	0.79304	1.11025	0.044400
0.00064475	51	0.79110	1.14507	0.045028
0.00048356	54	0.78917	1.14894	0.045097
0.00024178	58	0.78723	1.15667	0.045234
0.00001000	71	0.78337	1.16054	0.045303

min.xerror.cp = 0.00483559

Best Pruned Tree cp = 0.006963

	Reference	
Prediction	0	1
0	2765	208
1	9	18

DECISION TREE



DECISION TREE: INFERENCE

List of important predictors:

- distance
- hour_avg_cancellation
- traditional_booking
- user_id
- waitingTime

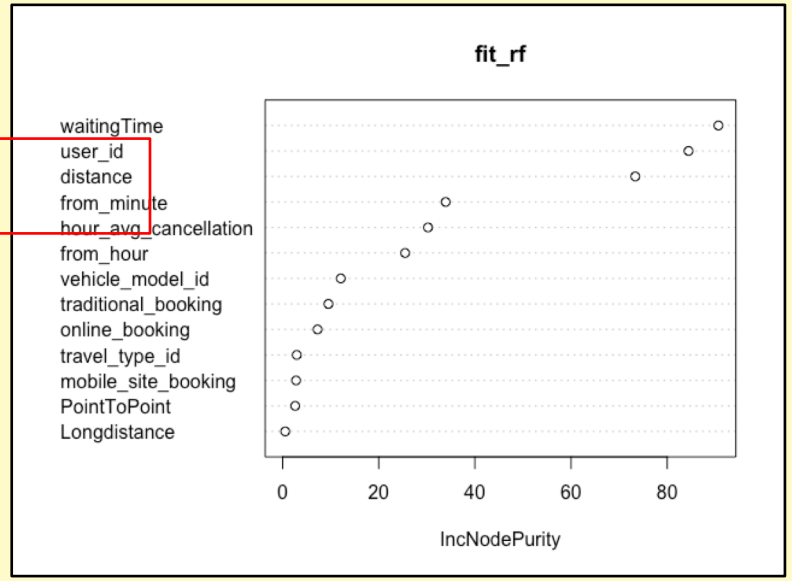
- **Accuracy of decision tree model: 92.77 %**
- **Area under ROC curve : 0.538**
- **The Decision Tree model has a higher accuracy and higher precision than the logistics model**

RANDOM FOREST

Random Forest model does not produce any tree diagram.

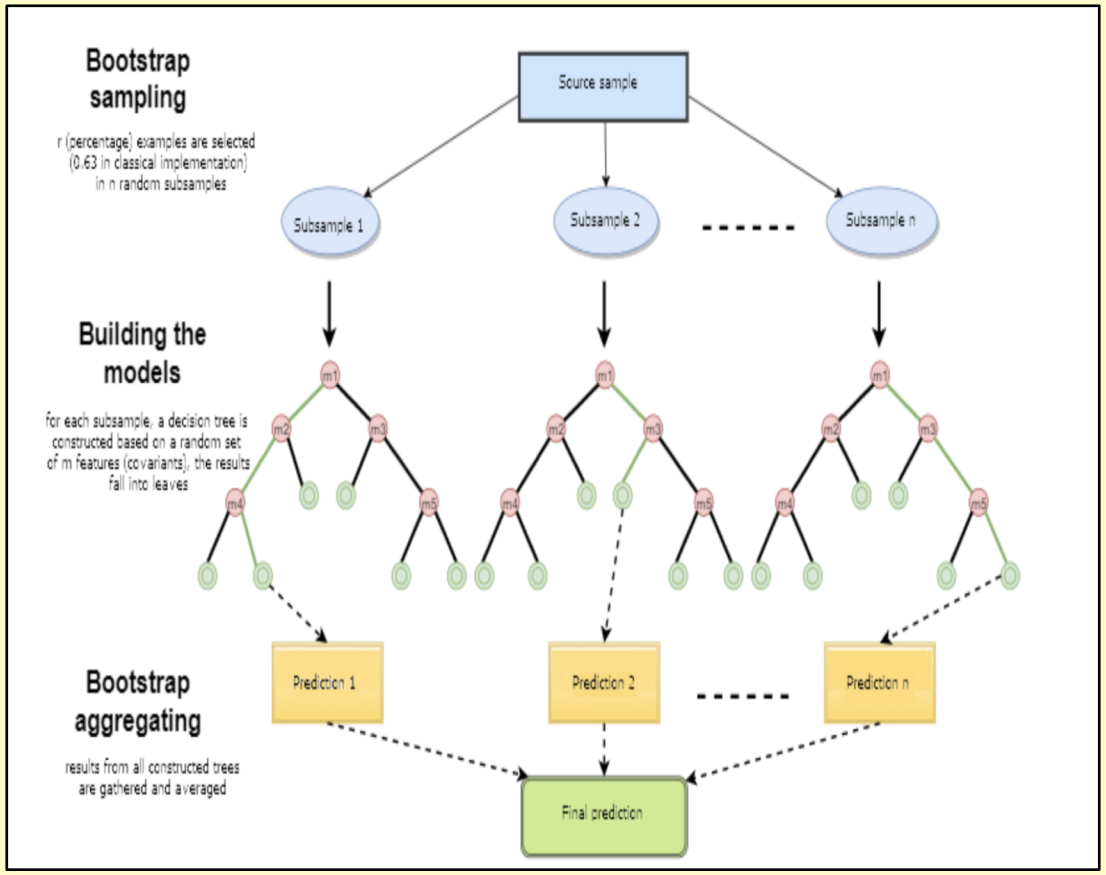
However, it shows “feature importance”.

The importance score for a particular predictor is computed by summing up the decrease in the Gini index for that predictor over all the trees in the forest.




```
> importance(fit_rf)
                    IncNodePurity
from_hour              25.4902430
user_id                84.4415728
vehicle_model_id       12.0979375
travel_type_id          2.9439209
online_booking          7.2694523
mobile_site_booking     2.8004039
```

RANDOM FOREST: INFERENCE



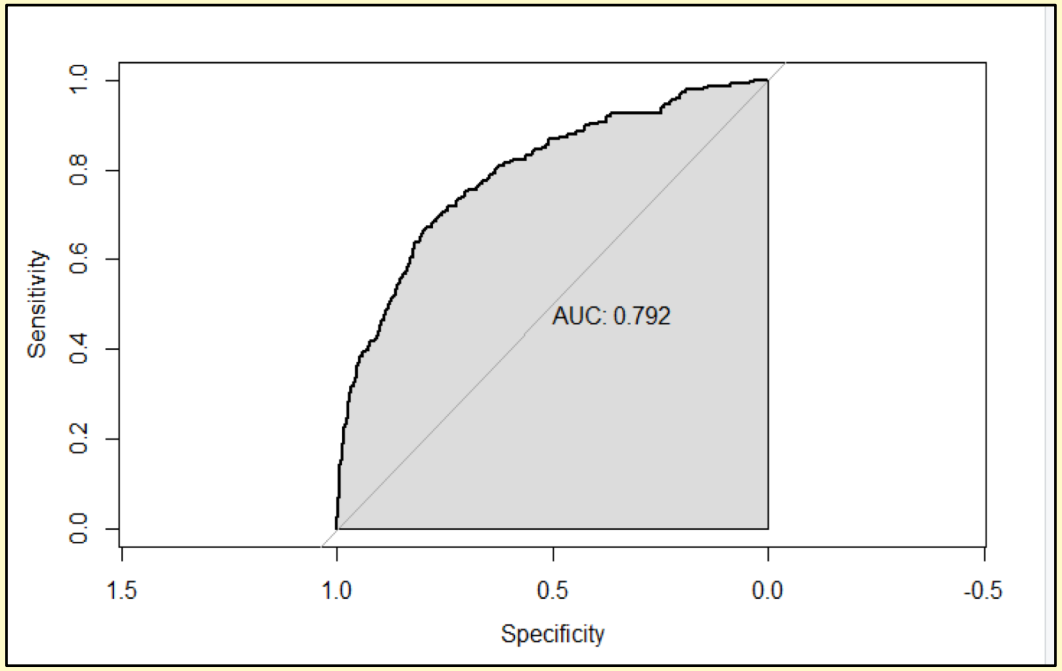
- **Accuracy of decision tree model: 93.77 %**
- **Area under ROC curve : 0.831**
- **The Random Forest model has a higher accuracy and higher precision than the other model**
- **It is very good predictive model, often the best choice for predictive modeling.**

COMPARISON OF ACCURACY BETWEEN THE MODELLING TECHNIQUES

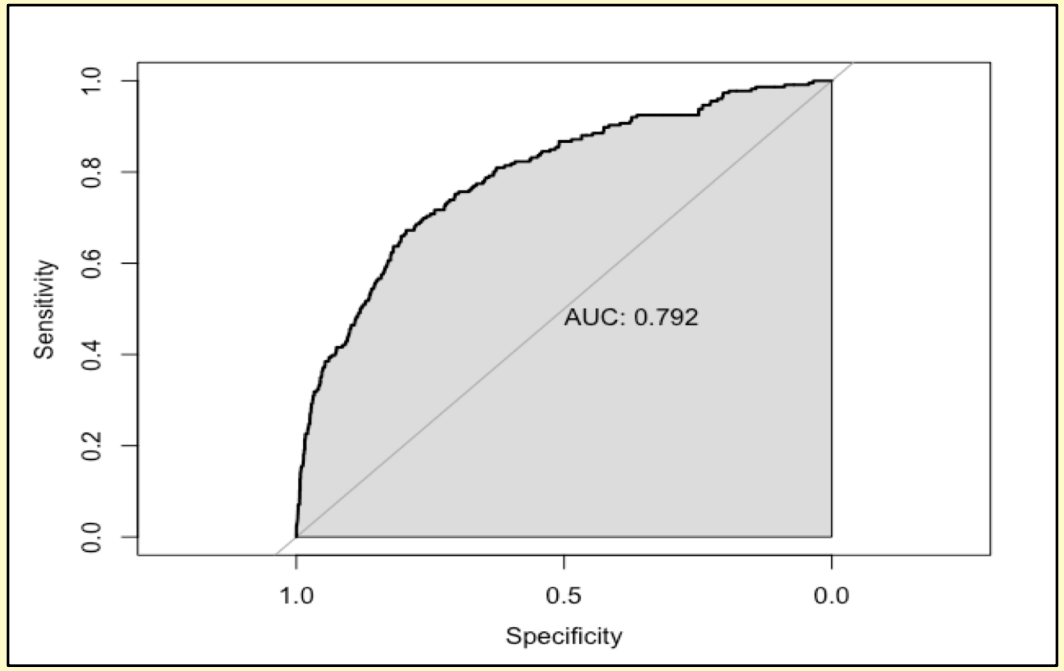
METHOD USED	ACCURACY
LOGISTIC REGRESSION	92.7
DECISION TREE	92.77
RANDOM FOREST	93.27

COMPARISON OF ROC CURVES

LOGISTIC REGRESSION

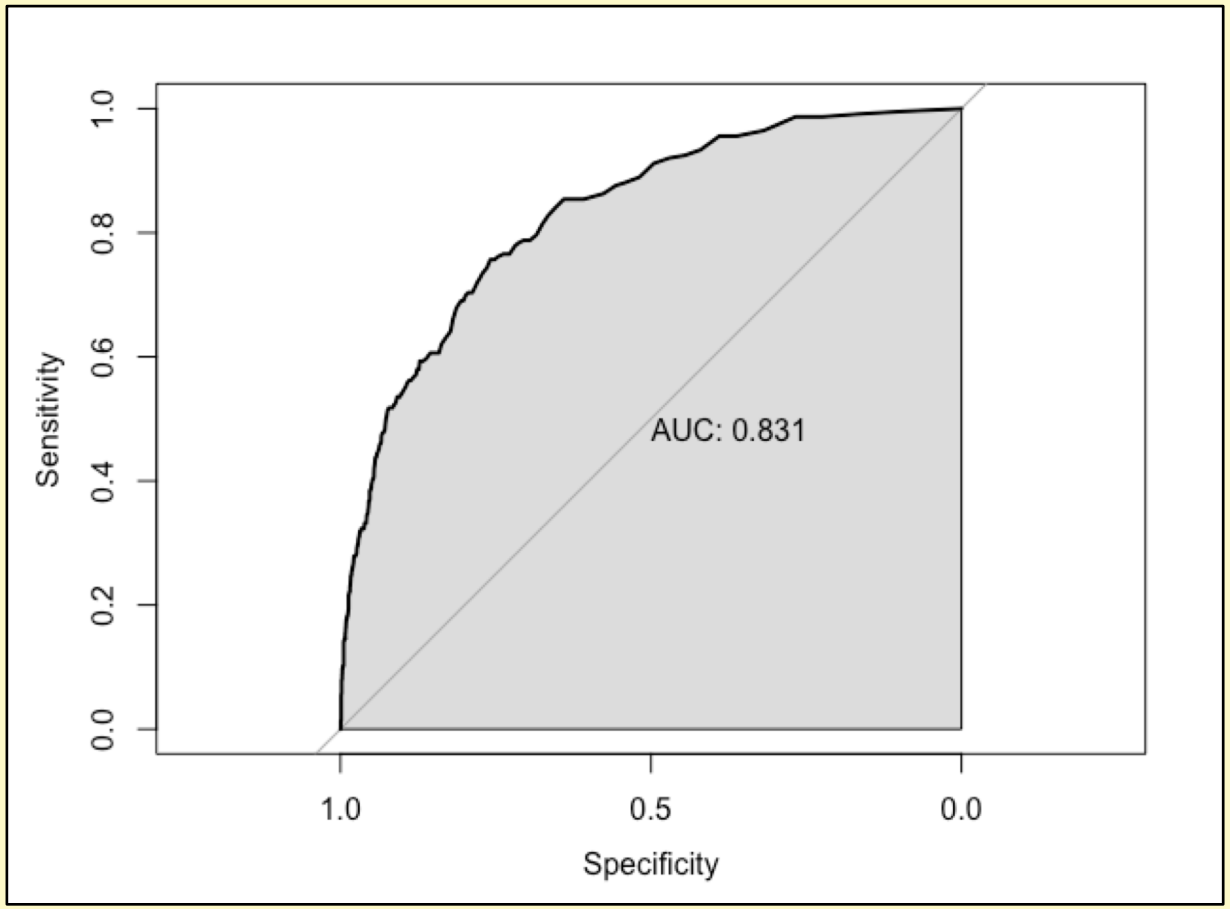


DECISION TREE



COMPARISON OF ROC CURVES

RANDOM FOREST



COMPARISON OF ROC CURVES : AREA UNDER CURVE

MODELLING TECHNIQUE	AREA UNDER CURVE
LOGISTIC REGRESSION	0.792
DECISION TREE	0.538
RANDOM FOREST	0.831

INFERENCE

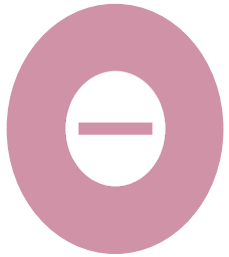
- From the results obtained, it seems that **random forest** has the highest accuracy(93.27) and highest value for ROC area under the curve(0.831) and hence, is the recommended model for usage.

IDEAS FOR IMPROVEMENT OF ANALYSIS

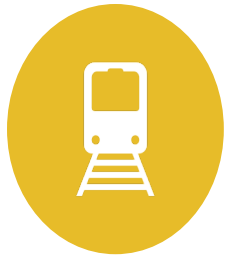
- Deeper data exploration and better feature engineering
- Increase number of observations and try to get a better balance of the data set
- Include factors in analysis such as geographic distance calculation, weather and climatological factors

RECOMMENDATIONS

Push Notifications : For booking that has a high chance of cancellations, send a push notification to the customer seeking a reconfirmation



Decline the booking : If the distance is too less and booking has a high probability for cancellation, don't accept the booking



Fleet Reduction : For those months of the year that have a high chance of cancellations consider reducing the fleet size

THANK YOU

Any Questions?