Taxi Cancellation

Indrani Mazumdar

Feb 6, 2018

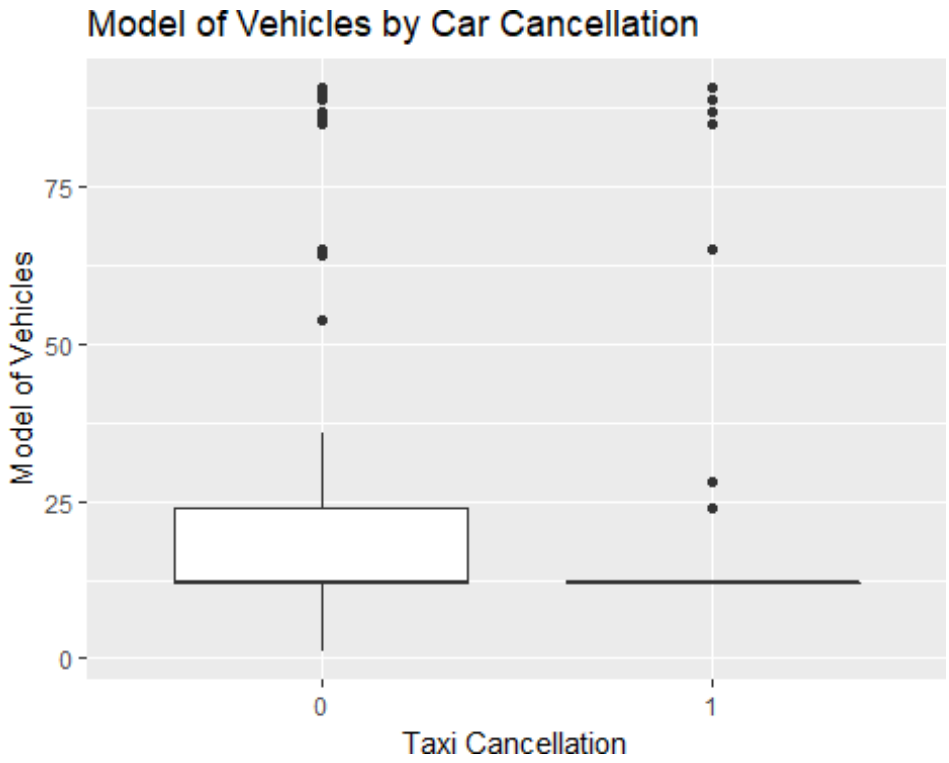
Part 1 - Preprocessing the data

```
library(tidyverse)
library(dplyr)
library(rpart)
library(rpart.plot)
library(lubridate)
library(geosphere)
library(ggplot2)
library(caret)

# read in the original data
taxi.orig.df <- read.csv("C:/Users/indrani/Desktop/My R
Work/BUS212/Data/Taxi_case.csv")
taxi.df <- read.csv("C:/Users/indrani/Desktop/My R
Work/BUS212/Data/Taxi_case.csv") # used for preprocessing

# summarize and describe the dataframe
str(taxi.df)

## 'data.frame':    9900 obs. of  19 variables:
##  $ row.              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ user_id           : int  17712 17037 761 868 21716 38966  22196 22200
22201 22202 ...
##  $ online_booking    : int  0 0 1 0 0 0 0 1 1 0 ...
##  $ mobile_site_booking: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ booking_created   : num  41275 41275 41276 41276 41276 ...
##  $ from_lat          : num  13 13 12.9 13 12.9 ...
##  $ from_long         : num  77.5 77.7 77.7 77.6 77.6 ...
##  $ to_lat            : Factor w/ 390 levels "12.77663","12.78091",..:
19 152 385 301 17 385 319 16 390 369 ...
##  $ to_long           : Factor w/ 387 levels "77.38693","77.38845",..:
293 358 360 192 118 360 69 386 387 266 ...
##  $ Car_Cancellation  : int  0 0 0 0 0 0 0 0 0 0 ...

summary(taxi.df)

##       row.          user_id      vehicle_model_id
package_id ##   Min.   :    1   Min.   :   16   Min.   : 1.00
NULL    :8166 ##   1st Qu.: 2506   1st Qu.:24450   1st Qu.:12.00
1       : 791
##  Median : 5002   Median :31510   Median :12.00   2       : 654
##  Mean   : 5002   Mean   :30666   Mean   :26.21   6       :
103 ##  3rd Qu.: 7500   3rd Qu.:39116   3rd Qu.:24.00   4       :
82 ##  Max.   :10000   Max.   :48729   Max.   :91.00   3       :
81
##                                                   (Other):  23
```

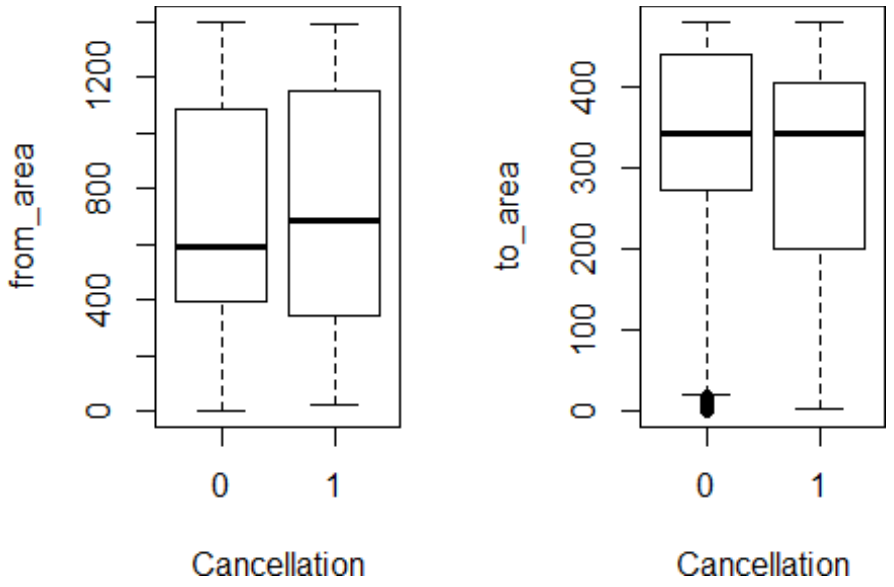
```
##  travel_type_id   from_area_id      to_area_id
    from_city_id ##  Min.   :1.000    Min.   :       2.0 NULL
    :2070  1        :27
##  1st Qu.:2.000   1st Qu.: 393.0   393    :1993   15  :3641
##  Median :2.000   Median : 590.0   585    : 521   NULL:6232
##   Mean   :2.141    Mean   : 709.4    1384    :
261 ##  3rd Qu.:2.000   3rd Qu.:1086.0   571    :
168 ##  Max.   :3.000   Max.   :1401.0   293    :
132 ##                  NA's                   :15
                    (Other):4755
##    to_city_id     from_date        to_date
    online_booking ## NULL :9564    Min.   :41275  NULL   :4134
    Min.   :0.0000
##  32     : 103   1st Qu.:41387   41406  :  13   1st Qu.:0.0000
##  55     :  35   Median :41463   41420  :  12   Median :0.0000
##  29     :  26   Mean   :41455   41427  :  12   Mean   :0.3536
##  146    :  17   3rd Qu.:41528   41447  :   9   3rd Qu.:1.0000
##  108    :  15   Max.   :41623   41434  :   8   Max.   :1.0000
##  (Other): 140                   (Other):5712
##  mobile_site_booking booking_created    from_lat       from_long
##  Min.   :0.00000     Min.   :41275   Min.   :12.78   Min.
    :77.39 ##           1st Qu.:0.00000 1st Qu.:41386   1st
Qu.:12.93  1st Qu.:77.59
##  Median :0.00000     Median :41462   Median :12.97   Median :77.64
##   Mean   :0.04232      Mean   :41453   Mean   :12.98   Mean   :
77.64 ##   3rd Qu.:0.00000      3rd Qu.:41526   3rd Qu.:13.01   3rd
Qu.:77.69 ##   Max.   :1.00000      Max.   :41603   Max.   :13.37
Max.   :77.79
##                                     NA's   :15      NA's   15
##       to_lat          to_long
Car_Cancellation ##   NULL     :2070   NULL    :2070
    Min.  :0.00000
##  13.19956 :1993   77.70688 :2046   1st Qu.:0.00000
##  12.97677 : 566   77.5727  : 566   Median :0.00000
##  13.02853 : 319   77.54625 : 319   Mean   :0.07394
##  12.95185 : 168   77.69642 : 168   3rd Qu.:0.00000
##  12.849482: 132   77.663187: 132   Max.   :1.00000
##  (Other)  :4652   (Other)  :4599
```

- First, we summarize the data. We have 9900 observations and 19 variables in the taxi database. Row number and user_id are identification variables which won't contribute to the decision of cancellation or not. We'll drop those two variables.

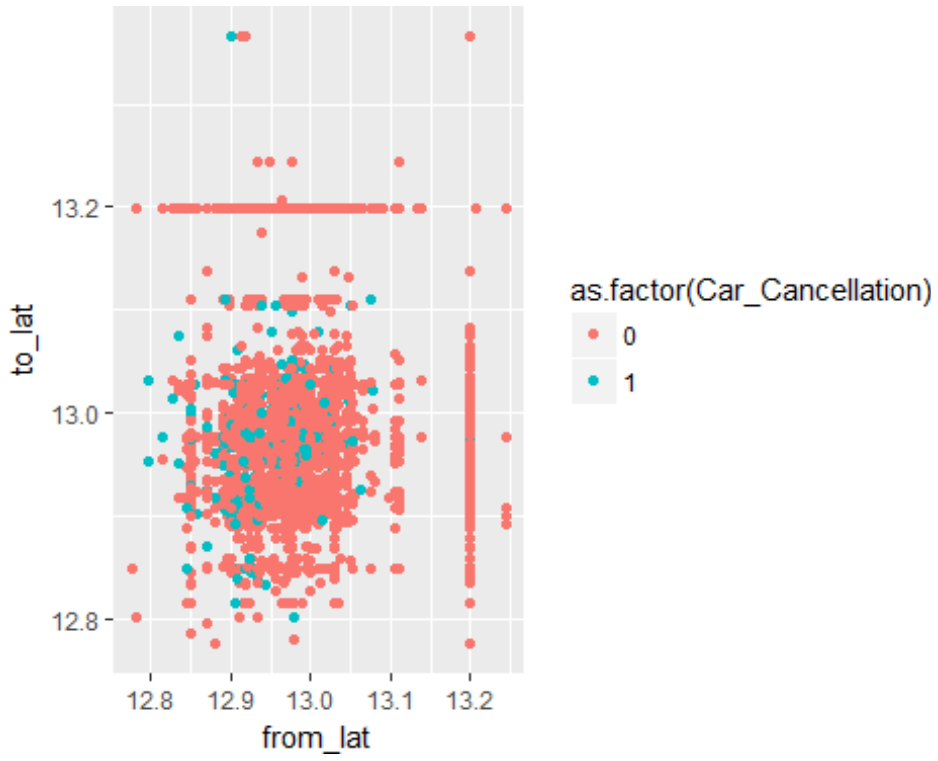
- From the summary of variables, we drop package_id, from_city_id, to_city_id and to_date which have more than half missing values. For factor variables, we will change the format of these variables into number in further analysis.

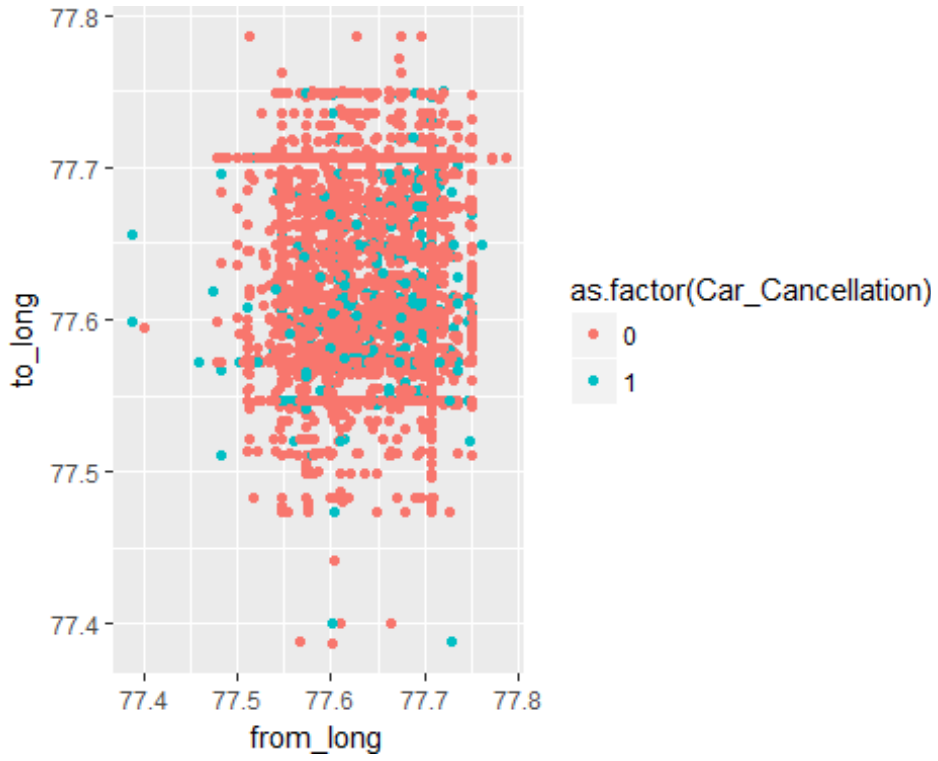


- After partioning the dataset, we investigate the vehicle_model_id variable.The boxplot clearly shows a variation of vehicle models by cancellation or not. The percentile range of vehicle model is roughly zero in the cancellation case, indicating that drivers with certain type of car are more likely to cancel the reservation. This variable should be included.



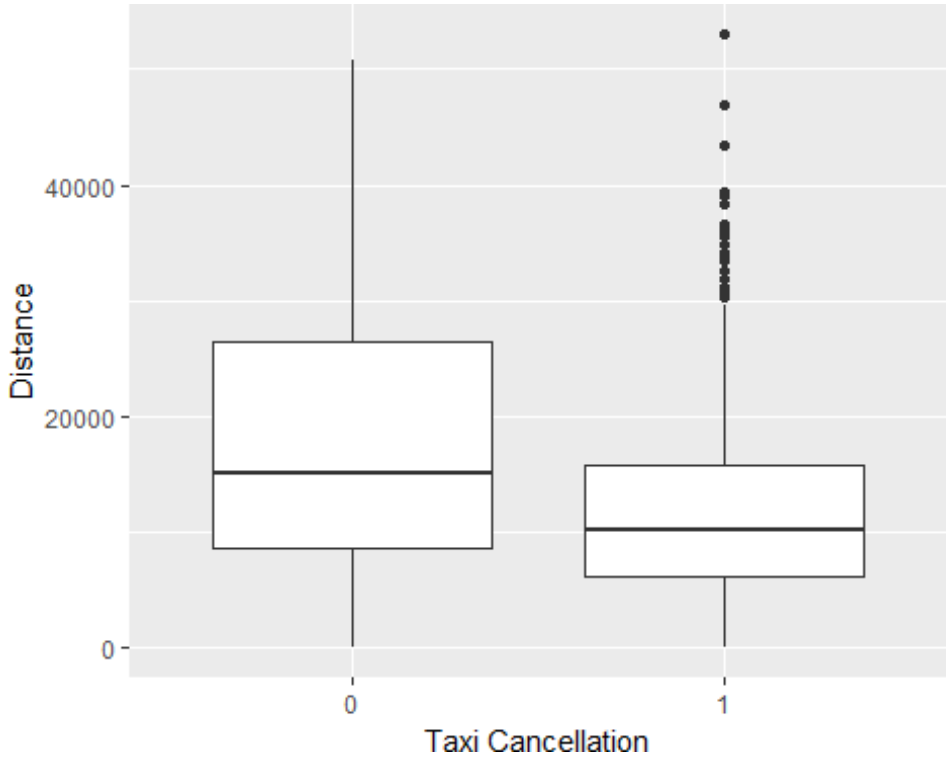
*The side-by-side boxplot above shows that the variation of both from_area_id and to_ared_id is not clear enough by cancellation. Therefore, we'll pick from_area_id which has extinct different mean value and which has less missing data.


13.2
13.0
12.8
to_lat
12.8
12.9
13.0
13.1
13.2
from_lat
as.factor(Car_Cancellation)
0
1



- We simply plot two scatter plot to investigate the pattern of longitude and latitude variables by cancellation.
- From two graphs above, we find that both cancelled and non-cancelled ride cluster together and it's hard to split a pure group. Therefore, we decide to calculate the distance between from place and to place, using the data more effeicienly.

Calculate the distance between the from_location and to_location



- From the boxplot above, we could clearly see the variation of distance by cancellation. There is just a little overlap of the percentile range of cancellation or not. Ride with more than 20000 is less likely to be cancelled by drivers than those short-distance ride. This information is really helpful for the classification tree.

Change the format of date variables

- Our group thinks that from_date and booking_created variable are both important. But the dataset showed those variables improperly, so, we decide to change the format of

the date variables and then extract month, day in a month, day in a week and hour variables from them. We decide to include all the variables the date and time of start and booking both matters.

- Our group also conducted simple grouping analysis which is not shown in this file. We find that travel_type_id is import because point-to-point ride tends to be cancelled by the drivers more frequently in the training data. Moreover, we pick up online_booking, instead of mobile_site_booking, in our tree model. These two variables are correlated to some degree and our analysis shows that online_booking matters more.

Part 2 - Input Variables explanation

Post the pre -processing and data cleaning the model uses the following columns listed below (with a few new ones extracted from the old columns)

- Vehicle_model_id - this column shares information on the type of vehicle the driver is driving and also depicts the information on whether it is a particular model which cancels more rides as compared to other models
- Travel_id_type - this column shares information on the type of trip requested and is chosen for our analysis to see if the
- Online_bkg - this column shares information on whether the booking has been made online or offline (is a binary variable, where 0 depicts no online booking and 1 depicts online booking)
- from_area_id - this column gives us the area code where the taxi is currently stationed and is there a relation between the area code and the cancellations of rides

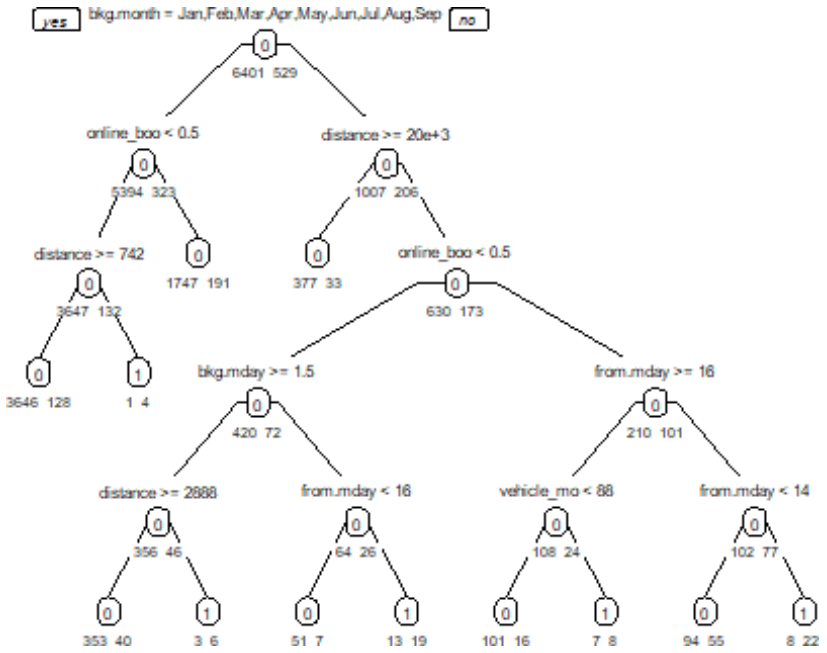
New Columns created - For the simple understanding of the tree we have extracted and simplified a few columns to give us a better picture of the scenario

- distance - the actual distance is a new column which is calculated using the latitude and longitude coordinates using the geoshpere package in R.
- from_month - this is the month of the booking of the taxi which is created to depict a trend in a particular month or is the cancellation distributed evenly across all months (this is extracted from the from_date column)
- from_mday - this is the day in a month when the trip has been requested for a taxi, this is usually evident for the point to point or the long distance trips
- from_wday - this is the day in a week when the trip has been requested, usually to check a trend between the weekdays and the weekends
- from_hour - this extracts the hour when the trip has been requested, this also is a good indicator to tell us whether it is the off-peak hours or the peak hours which observe more cancellations
- bkg_month - this column shares the information on the month when the booking has been requested
- bkg_mday - this shares information on the day in a month when the booking has been requested

- bkg_wday - this shares information on the day in a week when the booking has been requested
- bkg_hour - this column shares information on the hour when the booking has been made

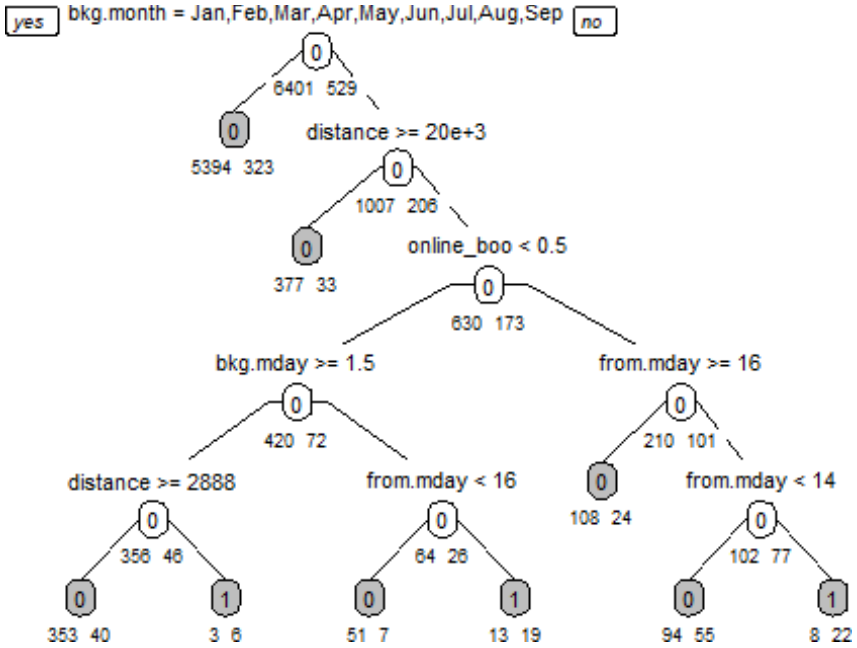
Note : the two columns extracted from_date and booking_created are important for us to understand that what type of requests, at which hour and month are getting cancelled.

Part 3&4 - Plot the tree and describe the decision rules




```
##
## Classification tree:
## rpart(formula = Car_Cancellation ~ ., data = train.df.mod, method =
"class",
##     cp = 1e-06, minsplit = 4, maxdepth = 5, xval =
5) ##
## Variables actually used in tree construction:
## [1] bkg.hour          bkg.mday          bkg.month         distance
## [5] from.mday         online_booking    vehicle_model_id
##
## Root node error: 529/6930 = 0.076335
##
## n= 6930
##
##           CP nsplit rel error  xerror     xstd
## 1 0.00529301      0   1.00000 1.00000 0.041786
## 2 0.00283554      8   0.95652 0.99055 0.041604
## 3 0.00189036     12   0.94518 0.99622 0.041713
## 4 0.00094518     15   0.93951 0.99622 0.041713
## 5 0.00000100     17   0.93762 1.00189 0.041822
```

- From the result of cross validation, we could see that the row 2, split 8 has the smallest xerror. We then prune the tree following the cross validation.



Decision rules for the tree

1. IF (bkg_month = Jan,Feb, Mar,Apr,May, Jun,Jul,Aug,Sep)
   - THEN CLASS = 0
2. IF (bkg_month <> Jan,Feb, Mar,Apr,May, Jun,Jul,Aug,Sep)
   - AND (distance < 20000) AND (online_booking > 0.5) AND (14<= from_mday <16) THEN CLASS = 1
3. IF (bkg_month <> Jan,Feb, Mar,Apr,May, Jun,Jul,Aug,Sep)

– AND (distance < 20000) AND (online_booking > 0.5) AND (from_mday >=16) THEN CLASS = 0

4. IF (bkg_month <> Jan,Feb, Mar,Apr,May, Jun,Jul,Aug,Sep)

– AND (distance < 20000) AND (online_booking > 0.5) AND (from_mday < 14) THEN CLASS = 0

5. IF (bkg_month <> Jan,Feb, Mar,Apr,May, Jun,Jul,Aug,Sep)
   - AND (distance >= 20000) THEN CLASS = 0
6. IF (bkg_month <> Jan,Feb, Mar,Apr,May, Jun,Jul,Aug,Sep)

– AND (distance < 20000) AND (online_bookinh < 0.5) AND (bkg_mday >= 1.5) AND

(distance >=2888) THEN CLASS = 0

7. IF (bkg_month <> Jan,Feb, Mar,Apr,May, Jun,Jul,Aug,Sep)

– AND (distance < 20000) AND (online_bookinh < 0.5) AND (bkg_mday >= 1.5) AND
(distance < 2888) THEN CLASS = 1

8. IF (bkg_month <> Jan,Feb, Mar,Apr,May, Jun,Jul,Aug,Sep)

– AND (distance < 20000) AND (online_bookinh < 0.5) AND (bkg_mday < 1.5) AND
(from_mday < 16) THEN CLASS = 0

9. IF (bkg_month <> Jan,Feb, Mar,Apr,May, Jun,Jul,Aug,Sep)

– AND (distance < 20000) AND (online_bookinh < 0.5) AND (bkg_mday < 1.5) AND
(from_mday >= 16) THEN CLASS = 1

Part 5 - Confusion

Matrix Treat the

valiation data

```
# change the format of validation data

valid.df$from_area_id <- as.numeric(valid.df$from_area_id)
valid.df$to_lat <- as.numeric(as.character(valid.df$to_lat))
valid.df$to_long <- as.numeric(as.character(valid.df$to_long))

# calcualte distance

for (i in 1:nrow(valid.df)) {
valid.df$distance[i] <- distm(c(valid.df$from_long[i], valid.df$from_lat[i]),
c(valid.df$to_long[i], valid.df$to_lat[i]))
}

# change format of date

valid.df$new.from.date <- as.POSIXct(as.Date(as.numeric(valid.df$from_date),
origin = "1899-12-31"))
valid.df$from.month <- month(valid.df$new.from.date, label =
TRUE) valid.df$from.mday <- mday(valid.df$new.from.date)
valid.df$from.wday  <- wday(valid.df$new.from.date, label = TRUE)
valid.df$from.hour <- hour(valid.df$new.from.date)

valid.df$new.bkg.create <-
as.POSIXct(as.Date(as.numeric(valid.df$booking_created), origin = "1899-12-
31"))
valid.df$bkg.month <- month(valid.df$new.bkg.create, label =
TRUE) valid.df$bkg.mday <- mday(valid.df$new.bkg.create)
valid.df$bkg.wday  <- wday(valid.df$new.bkg.create, label = TRUE)
valid.df$bkg.hour <- hour(valid.df$new.bkg.create)
```

Confusion Matrix

```
# for training data confusion matrix
class.tree.point.pred.train <- predict(pruned.ct,train.df,type = "class")
confusionMatrix(class.tree.point.pred.train, train.df$Car_Cancellation)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction        0
            1
##          0 6377  482
##          1   24   47
##
##                Accuracy : 0.927
##                  95% CI : (0.9206, 0.933)
##         No  Information  Rate  :
0.9237 ##      P-Value [Acc > NIR] :
0.1543 ##
##                   Kappa : 0.1412
##  Mcnemar's Test P-Value : <2e-
16 ##
##             Sensitivity : 0.99625
##             Specificity : 0.08885
##          Pos Pred Value : 0.92973
##          Neg Pred Value : 0.66197
##              Prevalence : 0.92367
##          Detection     Rate     :
0.92020 ##    Detection Prevalence :
0.98975 ##   Balanced    Accuracy    :
0.54255 ##
##        'Positive' Class :
0 ##
```

```
# for validation data confusion matrix
class.tree.point.pred.valid <- predict(pruned.ct,valid.df,type = "class")
confusionMatrix(class.tree.point.pred.valid, valid.df$Car_Cancellation)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction        0
            1
##          0 2750  186
##          1   17   17
##
##                Accuracy : 0.9316
##                  95% CI : (0.922, 0.9405)
##         No  Information  Rate  :
0.9316 ##      P-Value [Acc > NIR] :
0.5187 ##
##                   Kappa : 0.1263
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.99386
##             Specificity : 0.08374
```

```
##          Pos Pred Value : 0.93665
##          Neg Pred Value : 0.50000
##              Prevalence : 0.93165
##          Detection Rate : 0.92593
##    Detection Prevalence : 0.98855
##       Balanced Accuracy : 0.53880
##
##        'Positive' Class : 0
##
```

Comments on confusion matrix

For the first Confusion Matrix that evaluates the accuracy of of prediction in the training data, the accuracy rate is 92.7%. Our tree predicts that 71 orders will be cancelled over 6930 orders, while the training data shows that in fact 529 orders are cancelled. Regarding to sensitivity, 99.625% positives are identified almost correctly, so our model does a good job in predicting orders that won't be cancelled. The specificity tells us that only 8.885% negatives are correctly predicted, so our model needs improvements in its prediction of cancellation cases.

As for the second Confusion Matrix that evaluates the accuracy of prediction in validation data, the accuracy rate is 93.16%. Our model predicts that 34 orders will be cancelled over 2970 orders, while the validation data shows that actually 203 orders are cancelled.
Sensitivity is 99.386% for validation set, which is very similar to the one in training set, so our model fits the validation set quite well. And there won't be much overfitting issues in our model. However, our model has only 8.374% specificity in validation sets, indicating the model is not good at predicting negatives.

Thus, we can infer that the model is valid for the future test.

Part 6 - Predict new data

```r
# calcualte distance

for (i in 1:nrow(taxi.new.df)) {
  taxi.new.df$distance[i] <- distm(c(taxi.new.df$from_long[i],
taxi.new.df$from_lat[i]), c(taxi.new.df$to_long[i], taxi.new.df$to_lat[i]))
}

# change format of date

taxi.new.df$new.from.date <-
as.POSIXct(as.Date(as.numeric(taxi.new.df$from_date), origin = "1899-12-
31")) taxi.new.df$from.month <- month(taxi.new.df$new.from.date, label =
TRUE) taxi.new.df$from.mday <- mday(taxi.new.df$new.from.date)
taxi.new.df$from.wday  <- wday(taxi.new.df$new.from.date, label = TRUE)
taxi.new.df$from.hour <- hour(taxi.new.df$new.from.date)

taxi.new.df$new.bkg.create <-
as.POSIXct(as.Date(as.numeric(taxi.new.df$booking_created), origin = "1899-
12-31"))
taxi.new.df$bkg.month <- month(taxi.new.df$new.bkg.create, label = TRUE)
taxi.new.df$bkg.mday <- mday(taxi.new.df$new.bkg.create)
taxi.new.df$bkg.wday <- wday(taxi.new.df$new.bkg.create, label = TRUE)
taxi.new.df$bkg.hour <- hour(taxi.new.df$new.bkg.create)

# new data with pruned tree

class.tree.point.pred.new <- predict(pruned.ct,taxi.new.df,type =
"class") taxi.new.df$Car_Cancellation <- class.tree.point.pred.new
taxi.new.orig.df$Car_Cancellation <-  class.tree.point.pred.new

head(taxi.new.orig.df)

##   row. user_id vehicle_model_id package_id travel_type_id from_area_id
## 1 3469   32315               12       NULL              2          392
## 2 4344   34894               12          1              3          836
## 3 1391   26169               12       NULL              2          722
## 4 6843   40039               12       NULL              2         1281
## 5 5037   31995               12       NULL              2          776
## 6 9381   23473               65          4              3          625
##   to_area_id from_city_id to_city_id from_date     to_date online_booking
## 1       1371         NULL       NULL  41424.71        NULL              0
## 2       NULL         NULL       NULL  41451.38        NULL              0
## 3       1036         NULL       NULL  41339.78        NULL              0
## 4       1237           15       NULL  41528.69 41528.75007              0
## 5        393         NULL       NULL  41471.75 41471.80008              1
## 6       NULL           15       NULL  41602.01 41602.42708              1
##   mobile_site_booking booking_created from_lat from_long    to_lat
## 1                   0        41424.64 12.98628  77.73525 13.000418
## 2                   0        41451.33 12.90944  77.57295      NULL
## 3                   0        41339.76 12.91567  77.55465 12.925568
## 4                   0        41519.59 12.93463  77.61128  12.92645
## 5                   0        41471.54 13.01508  77.67796  13.19956
## 6                   0        41601.92 12.95431  77.65530      NULL
##     to_long Car_Cancellation
```

```
## 1 77.674835                0
## 2       NULL                0
## 3 77.580568                0
## 4  77.61206                0
## 5  77.70688                0
## 6       NULL                0

# export dataset

write.csv(taxi.new.orig.df, file = "Taxi_new_cancellation.csv")
```

- Our tree model predicts that in the new dataset, only the reservations with row number of #8776 will be cancelled.
- In the new sample, the predicted cancellation rate is 1%, which is far more below the average of 7% or 8%. Our model performs relatively weak to predict cancellation of reservation.

Part 7 - Future Improvement

Generally, we are satisfied with the model we trained. Still, there are many improvements can be done for better prediction.

First of all, we identified from_date (the exact time an order is made) to be an important indicator of cancellations. But, when we tried to extract year, month, date and specific time from this column, we find it very complicated to convert data into information we need. Thus, if the data is given separately as date and hour, that will be much more straight-forward for us to process.

Second, we determined to_area_id have major influence on car cancellation rate. Taxi drivers might cancel an order because they want to avoid certain destinations. Since to_area_id are numbers that has no numeric meaning but simply denote certain areas, they are difficult to emphasize how numbers contribute to cancellation rate. Better way to understand the correlation between destinations and cancellation rate is to draw a map showing frequencies of cancellation in different areas. Then we could explore the pattern in certain areas and group them together. If we have different group of areas, we can use the grouped data to train our model.

Third, as we processed the data before training model, we found out that the Model Vehicle Type has a significant impact on taxi cancellations. But since the data gives us only the number, there's fewer information we can extract from it. If more information about Model Vehicles, such as make, model and year, could be accessible, we will have a deeper understanding of how Model Vehicles shapes to Car Cancellation decisions. An intuitive

guess would be that taxi drivers might be reluctant to take long-distance customers if their cars are too old and worn-out.

Our group suggests that the variable for traffic zone (heavy zone - medium zone -low zone) will help us understand another factor for car cancellation as this can be an important factor for the taxi driver to cancel rides which are in heavy traffic zone.

With better understand of the data, the use of random forest or boosted trees can generate a more accurate data prediction as these models have more sampling.