Preferred Mode of Transport – Work

We need to predict whether or not an employee will use Car as a mode of transport. Also, which variables are a significant predictor behind this decision.

Problem Statement:

Employee will use Car as a mode of transport

Data in below format:

Variables	Description
Age	Age of employee
Gender	Gender of employee (Male or Female)
Engineer	Whether an Engineer 1- Yes 0 - No
MBA	Whether an MBA 1- Yes 0 - No
Work Exp	Number of work- exp
Salary	Salary in Numbers
Distance	Commute Distance
license	Whether Driving License 1 - Yes and 0 - No
Transport	Option for Transport - Public Transport / 2 - Wheeler / Car

Steps to be followed:

- Import Cars Input data files in R Studio
- Install required libraries for data analysis and building model
- Read and carry explanatory data analysis
- Prepare data (SMOTE)
- Study imported data using various graphs/plots
- Identify correlated variable
- Build predictive model with KNN model and Naïve Bayes Model
- Use Bagging / Boosting techniques

Exploratory Data Analysis –

Environment setup - install required libraries and built a function for doing EDA.

```
rm(list=ls())
library(xlsx)
setwd('C:/Users/ashwi/Downloads')
mydata=read.csv("Cars_edited.csv")
```

Running below commands in R to perform EDA which includes

```
str(mydata)
names(mydata)
summary(mydata)
head(mydata)
glimpse(mydata)
df_status(mydata)

profiling_num(mydata)
plot_num(mydata)
describe(mydata)
```

- Glimpse – This makes it possible to see every column in a data frame
- Df_status For each variable it returns: Quantity and percentage of zeros (q_zeros and p_zeros respectively). Same metrics for NA values (q_NA/p_na), and infinite values (q_inf/p_inf). Last two columns indicate data type and quantity of unique values. This function print and return the results.
- Freq - basis of estimating population sizes, their standard error, and symmetric as well as asymmetric confidence intervall.
- Profiling_num - Get a metric table with many indicators for all numerical variables, automatically skipping the non-numerical variables. Current metrics are: mean, std_dev: standard deviation, all the p_XX: percentile at XX number, skewness, kurtosis, iqr: inter quartile range, variation_coef: the ratio of sd/mean, range_98 is the limit for which the 98
- Plot_num - Retrieves one plot containing all the histograms for numerical variables. NA values will not be displayed
- Describe - Describes a vector or the columns in a matrix or data frame.
- Summary To produce result summaries of the results of various model fitting functions
- Names This function prints the header of each column.
- Head This function shows the first n number of rows. This helps to check if the data has been properly imported to R.

Observation noted after running above R commands for EDA

a) data. Frame: 444 obs. of 9 variables:

b) Variable name

"Age" "Gender" "Engineer" "MBA" "Work.Exp" "Salary" "Distance" "license" "Transport"

c) Summary observation

- Transport is a dependent variable and a factor variable
- Other factor variables are Gender / Engineer / MBA / License

```
> summary(mydata)
      Age           Gender       Engineer          MBA            Work.Exp         Salary        Distance         license                  Transport
 Min.   :18    Female:128   Min.   :0.00   Min.   :0.00   Min.   : 0.0   Min.   : 6   Min.   : 3.2   Min.   :0.00   2Wheeler        : 83
 1st Qu.:25    Male  :316   1st Qu.:1.00   1st Qu.:0.00   1st Qu.: 3.0   1st Qu.:10   1st Qu.: 8.8   1st Qu.:0.00   Car             : 61
 Median :27                 Median :1.00   Median :0.00   Median : 5.0   Median :14   Median :11.0   Median :0.00   Public Transport:300
 Mean   :28                 Mean   :0.75   Mean   :0.25   Mean   : 6.3   Mean   :16   Mean   :11.3   Mean   :0.23
 3rd Qu.:30                 3rd Qu.:1.00   3rd Qu.:1.00   3rd Qu.: 8.0   3rd Qu.:16   3rd Qu.:13.4   3rd Qu.:0.00
 Max.   :43                 Max.   :1.00   Max.   :1.00   Max.   :24.0   Max.   :57   Max.   :23.4   Max.   :1.00
                                           NA's   :1
```

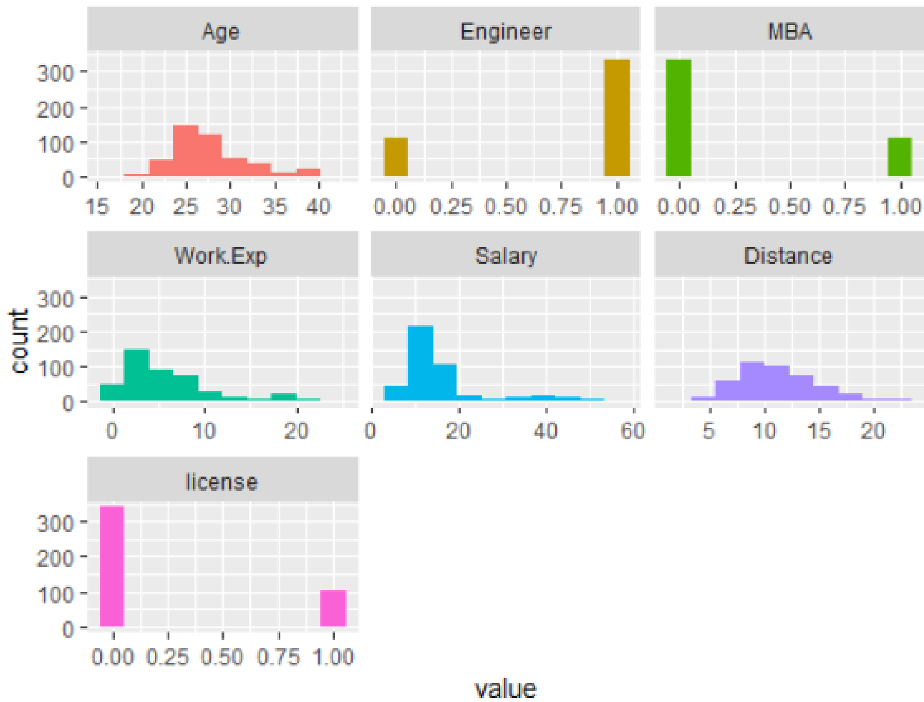
d) DF_STATUS observation:

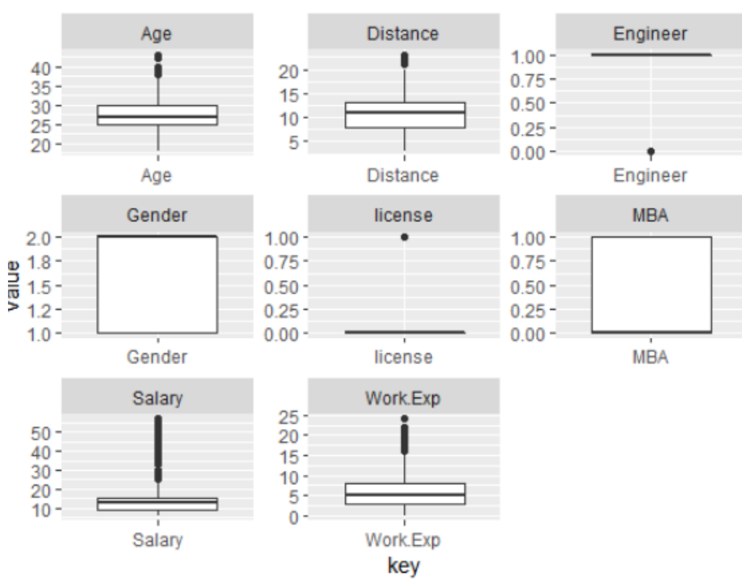
```
> df_status(mydata)
   variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
1       Age       0     0.0    0 0.00     0     0 integer     25
2    Gender       0     0.0    0 0.00     0     0  factor      2
3  Engineer     109    24.6    0 0.00     0     0 integer      2
4       MBA     331    74.5    1 0.23     0     0 integer      2
5  Work.Exp      29     6.5    0 0.00     0     0 integer     24
6    Salary       0     0.0    0 0.00     0     0 numeric    122
7  Distance       0     0.0    0 0.00     0     0 numeric    137
8   license     340    76.6    0 0.00     0     0 integer      2
9 Transport       0     0.0    0 0.00     0     0  factor      3
>
```

- **Numbers of zeroes** is indicated by q_zeros which will be higher for factor variable.
- Q_na column indicates **null values** on column which is on MBA column and number of occurrences are 1, which needs to be treated.

e) <u>Graph for the value distribution for each numeric variable:</u>

- Age is left skewed graph, where commuter is more in 25 – 30 age range
- Engineer are more compared to MBA
- Majority of people has experience in range from 2 – 8 years
- Majority for people do not have license

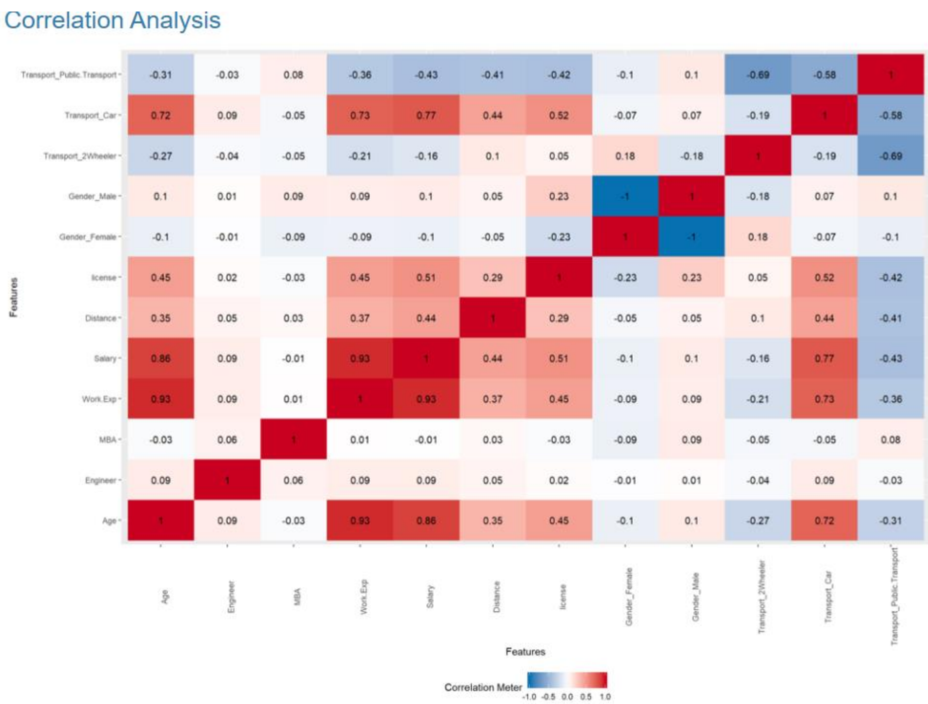




Observation from Co-relation graph:

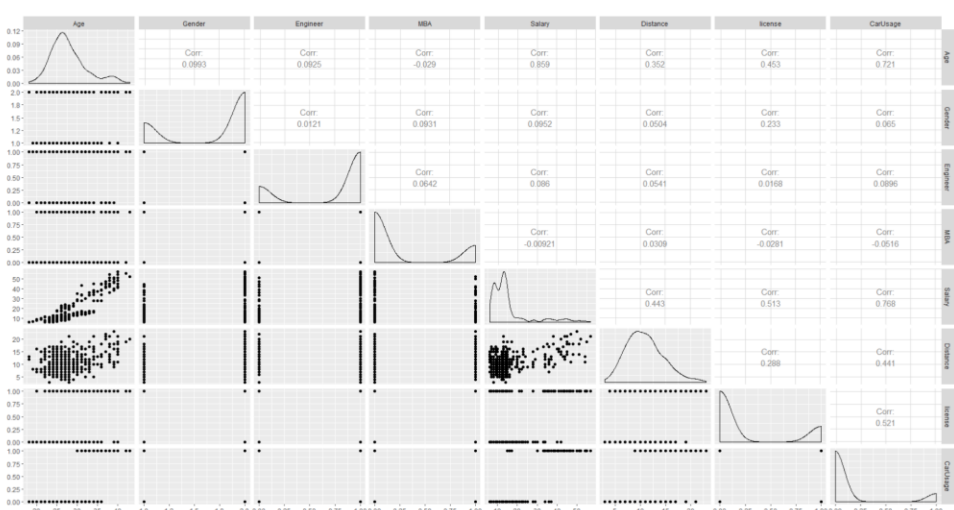
- ➢ Salary / Work Experience / Age are highly correlated with each other
- ➢ Age and Transport are also corelated

All the correlation would be considered when building model.



As the variables are correlated removed Work-experience which removed correlation, below are the vif values without work-experience variable:

```
> car::vif(m1)
   Age   Gender Engineer      MBA   Salary Distance  license
   1.8      1.2      1.1      1.3      1.6      1.2      1.3
>
```



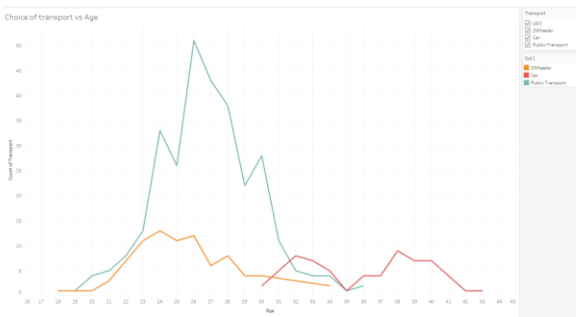
Additional Graph and its observations

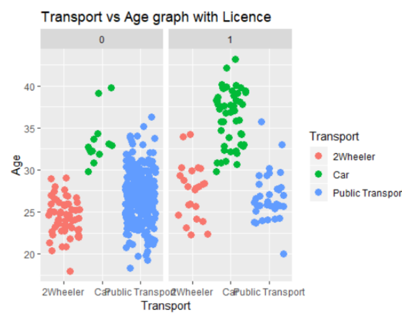
Age and mode of transport:

- ➢ Mode of transport car is preferred only after age 30
- ➢ Popular mode of transport is public and common till age 30

Age and mode of transport against license:

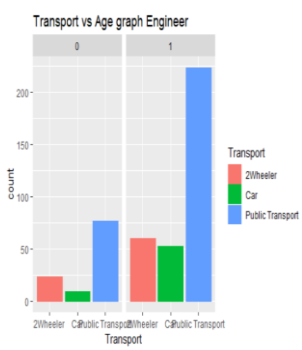
- ➢ Left side as 0 indicates as no license and right side as 1 indicates as license
- ➢ Car is popular after age – 30 and also driven by people with no license

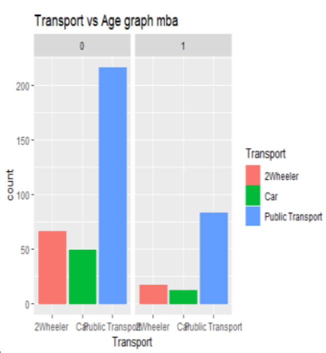


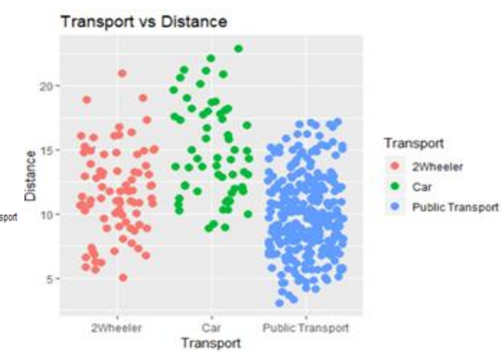


Age and mode of transport against Engineer:

- ➢ Engineer are driving more Cars compared to MBA degree employee
- ➢ For distance more than 18 people don't prefer public transport
- ➢ Preferred transport for longer distance is 2-wheeler or car







Data-Setup for building model

Train and Test data-set:

1) Split the data in train and test in proportion = 0.7

Observations after data split

	Train Dataset	Test Dataset
Number of rows	312	131
Number of columns	8	8
Percentage of Car users	0.1346	0.145

SMOTE Data-set:

The parameters perc.over and perc.under control the amount of over-sampling of the minority class and under-sampling of the majority classes, respectively

perc.over : means that 1 minority class will be added for every value of perc.over
perc.under : taking out of the majority class

Before SMOTE train dataset has 42 minority variables in total on 312 rows

In order to have equal distribution (126/126)
perc.over 200 : 2*42+42
perc.over 150 : 144 records has to be removed hence used 150

```
test$CarUsage<-as.factor(test$CarUsage)
train$CarUsage<-as.factor(train$CarUsage)
smote_train <- SMOTE(CarUsage ~ ., train , perc.over = 200, k = 5, perc.under = 150)
table(smote_train$CarUsage)
smote_features_train<-as.matrix(smote_train[,1:7])
smote_label_train<-as.matrix(smote_train$CarUsage)
```

Model Building

a) Logistic-regression:

Used STEP function to fist understand important variable and remove the other.

- Age / Distance / license / MBA are important variable in logistic model
- AIC score of model is 75.98 which will be used to relative quality of statistical model
- The model having least AIC Score would be the most preferred and optimized one

```
Call:
glm(formula = smote_train$CarUsage ~ . - Engineer - Salary -
    Gender, family = binomial(link = "logit"), data = smote_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0364  -0.0929  -0.0001   0.0220   1.9401

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -37.827      6.794   -5.57  2.6e-08 ***
Age            1.126      0.209    5.38  7.6e-08 ***
MBA           -1.408      0.834   -1.69    0.091 .
Distance       0.259      0.130    1.99    0.047 *
license        0.993      0.791    1.26    0.209
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 349.346  on 251  degrees of freedom
Residual deviance:  65.985  on 247  degrees of freedom
AIC: 75.98
```

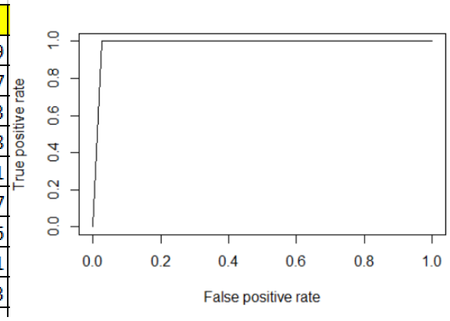
pR2(m1)

llh	llhNull	G2	McFadden	r2ML	r2CU
-32.99	-174.67	283.36	0.81	0.68	0.90

Model 1: smote_train$CarUsage ~ (Age + Gender + Engineer + MBA + Salary +
Distance + license) - Engineer - Salary - Gender
Model 2: smote_train$CarUsage ~ 1
#Df LogLik Df Chisq Pr(>Chisq)
1 5 -33
2 1 -175 -4 283 <2e-16 ***

Parameters for model / Graph :

	Logistic Regression
AUC	0.99
KS	0.97
Gini	0.83
Accuracy	0.98
Sensitivity	1
Specificity	0.97
Predicted Positive Rate	0.86
Predicted Negative Rate	1
Error	0.023



b) Navie Bayes:

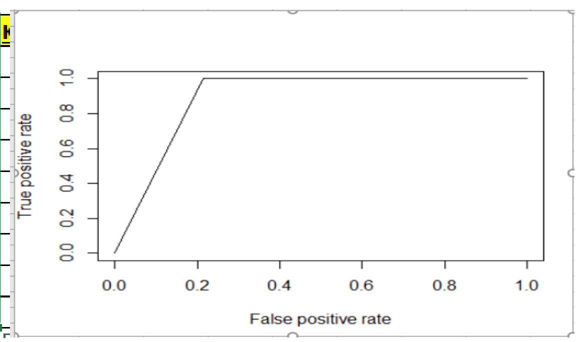
The e1071 package holds the naiveBayes function. It allows continuous and categorical features to be used in the naive bayes model. It is count-based classifier i.e. only thing it does is – count how often each variable's distinct values occur for each class.

Prior probabilities and conditional probabilities:

```
Conditional probabilities:
   Age
Y       0     1
  0 0.937 0.063
  1 0.063 0.937

   Gender
Y        1 1.10558998282067 1.27228792803362 1.27826669323258 1.31011119461618 1.43337579257786 1.81195830134675 1.96593745937571
  0 0.2937           0.0000           0.0000           0.0000           0.0000           0.0000           0.0000           0.0000
  1 0.1746           0.0079           0.0079           0.0079           0.0079           0.0079           0.0079           0.0079
   Gender
Y   1.99479480995797      2
  0           0.0000 0.7063
  1           0.0079 0.7619

   Engineer
Y        0 0.157906927168369 0.220026169205084 0.473365484271199 0.789027367718518      1
  0 0.1984            0.0000            0.0000            0.0000            0.0000 0.8016
  1 0.1429            0.0079            0.0079            0.0079            0.0079 0.8254

   MBA
Y        0 0.379208744736388 0.492872001836076 0.726590186823159 0.804266741033643 0.84550999943167 0.858901411760598      1
  0 0.6905            0.0000            0.0000            0.0000            0.0000           0.0000            0.0000 0.3095
  1 0.7778            0.0079            0.0079            0.0079            0.0079           0.0079            0.0079 0.1746

   Salary
Y       0     1
  0 0.881 0.119
  1 0.048 0.952

   Distance
Y      0    1
  0 0.61 0.39
  1 0.14 0.86

   license
Y        0 0.61393270897679 0.638174654217437 0.905037128599361      1
  0 0.8095           0.0000            0.0000            0.0000 0.1905
  1 0.1905           0.0079            0.0079            0.0079 0.7857
```

Model Performance and Graph:

	Navi Bayes
AUC	0.89
KS	0.79
Gini	0.67
Accuracy	0.82
Sensitivity	1
Specificity	0.79
Predicted Positive Rate	0.44
Predicted Negative Rate	1
Error	0.18



c) KNN – Model

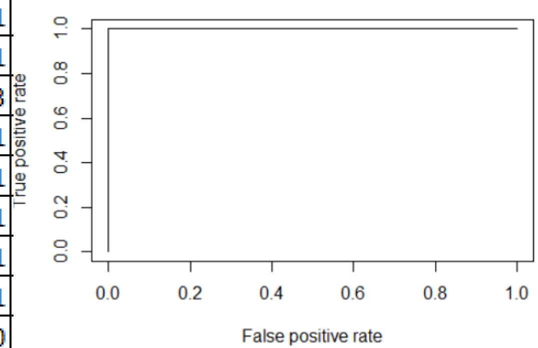
KNN is supervised classifier, which uses neighbor data points' information to predict outcome variable. Neighbors are identified using distance measures such as Euclidean distance.

Value used for model was K = 5

```
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 202, 201, 202, 202, 201
Resampling results across tuning parameters:

  k   Accuracy  Kappa
   5  0.88      0.75
   7  0.87      0.74
   9  0.84      0.68
  11  0.85      0.70
  13  0.85      0.69
  15  0.84      0.68
  17  0.84      0.68
  19  0.83      0.65
  21  0.82      0.64
  23  0.82      0.64

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
```

Model Performance and Graph:

	KNN
AUC	1
KS	1
Gini	0.83
Accuracy	1
Sensitivity	1
Specificity	1
Predicted Positive Rate	1
Predicted Negative Rate	1
Error	0

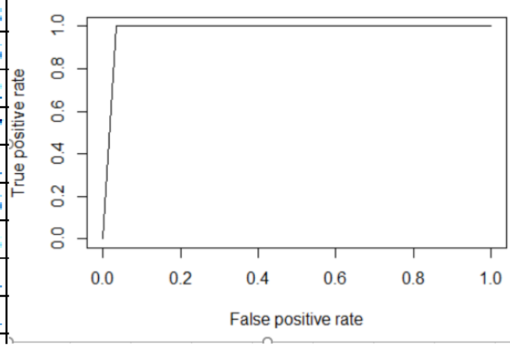


d) Bagging -Model

Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression

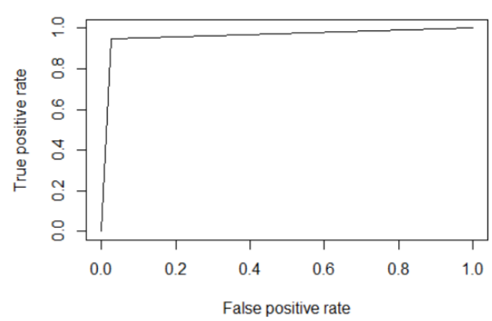
Model Performance and Graph:

	Bagging
AUC	0.98
KS	0.96
Gini	0.12
Accuracy	0.97
Sensitivity	1
Specificity	0.96
Predicted Positive Rate	0.83
Predicted Negative Rate	1
Error	0.031



e) Boosting -Model

	Boosting
AUC	0.96
KS	0.92
Gini	0.12
Accuracy	0.97
Sensitivity	0.95
Specificity	0.97
Predicted Positive Rate	0.86
Predicted Negative Rate	0.99
Error	0.031



Comparison for all models

- KNN model shows better predictive rate compared to all models
- Error rate is less in Logistic Regression model
- Navi Bayes is least preferred model compared to all model
- All models are showing good results in predicting Transport

	Logistic Regression	Navi Bayes	KNN	Bagging	Boosting
AUC	0.99	0.89	1	0.98	0.96
KS	0.97	0.79	1	0.96	0.92
Gini	0.83	0.67	0.83	0.12	0.12
Accuracy	0.98	0.82	1	0.97	0.97
Sensitivity	1	1	1	1	0.95
Specificity	0.97	0.79	1	0.96	0.97
Predicted Positive Rate	0.86	0.44	1	0.83	0.86
Predicted Negative Rate	1	1	1	1	0.99
Error	0.023	0.18	0	0.031	0.031

CONCLUSION

- Comparing build 5 models on Accuracy, we conclude KNN is competitively accurate in prediction data
- Accuracy was approx. 100 %, which indicates KNN is good for predictive model
- Variables below are factors due to which people discontinue existing service

  Age: odds of 1.13 and 76% of information can be predicted just with age parameter

  Distance: odds of .26 and 56% of information can be predicted just with distance parameter

  License: odds of .99 and 73 % of information can be predicted just with distance parameter

  MBA: Has a negative correlation odds of negative 1.41 and 20% of information can be predicted just with distance parameter