CARS

Karthik Paranthaman

UNIVERSITY OF TEXAS

1. PROJECT OVERVIEW

The main aim of this project is to explore the data set provided by the client "cars-dataset" which contains 9 different attributes carries every employee's personal details to find business insights to whether or not a particular will use the car as the mode of transport over public transport or 2wheeler. First Chapter discusses the EDA analysis such as Univariate analysis and Bi-Variate analysis. 2nd Chapter illustrates the techniques to analyze the data. Chapter -3 will be discussing the various Predictive Algorithm such as K-Nearest Neighbor, Naïve-Bayes and Logistic Regression built and the model performance evaluation. Chapter -5 discusses the overall discussion while the Chapter-6 provides a conclusive findings.

2. ASSUMPTIONS

The information provided by the customer about the employee license status contains information of not having license and driving 2 wheelers. I have assumed that employee must be driving a 2wheeler like TVS Moped with lesser CC which does not require license and it's not a typo error created by the data entry operator. It was assumed that all information provided by the customers are accurate. Also, the data distribution is considered as normally distributed due to the limited number of obeservations.

3. EXPLORATORY DATA ANALYSIS

In this Chapter we will be discussing about the data exploration approaches to find business insights. The exploratory process consists of the following stages.

1. Environment Setup and Data Import
2. Missing Value Identification
3. Missing Value Treatment.
4. Variable Identification
5. Univariate Analysis
6. Bi-Variate Analysis

3.1 Environment Setup and Data Import

3.1.1 WORKING DIRECTORY

The working directory has been set. The dataset is called using the name "Customer_Details" in R-Studio (Refer to APPENDIX -1 – for R Code).

3.2 Missing Value Identification

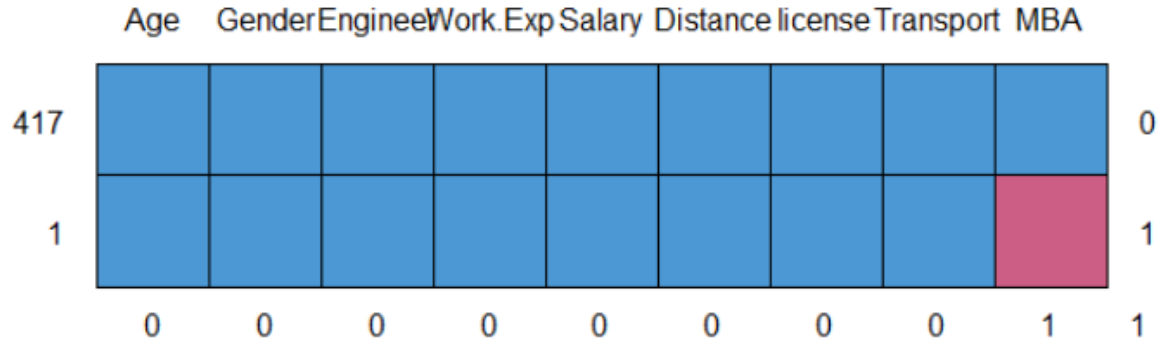
Based on the 5 line summary output produced by the "summary" R function, we can see that employee MBA attribute does contain some "Missing Values".

```
      Age          Gender       Engineer          MBA          Work.Exp
 Min.   :18.00   Female:121   Min.   :0.0000   Min.   :0.0000   Min.   : 0.000
 1st Qu.:25.00   Male  :297   1st Qu.:0.2500   1st Qu.:0.0000   1st Qu.: 3.000
 Median :27.00                Median :1.0000   Median :0.0000   Median : 5.000
 Mean   :27.33                Mean   :0.7488   Mean   :0.2614   Mean   : 5.873
 3rd Qu.:29.00                3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 8.000
 Max.   :43.00                Max.   :1.0000   Max.   :1.0000   Max.   :24.000
                                               NA's   :1
     Salary          Distance         license                  Transport
 Min.   : 6.500   Min.   : 3.20   Min.   :0.0000   2Wheeler        : 83
 1st Qu.: 9.625   1st Qu.: 8.60   1st Qu.:0.0000   Car             : 35
 Median :13.000   Median :10.90   Median :0.0000   Public Transport:300
 Mean   :15.418   Mean   :11.29   Mean   :0.2033
 3rd Qu.:14.900   3rd Qu.:13.57   3rd Qu.:0.0000
 Max.   :57.000   Max.   :23.40   Max.   :1.0000
```

This also can be confirmed by running the “anyNA” function for missing value identification and the output is shown below.

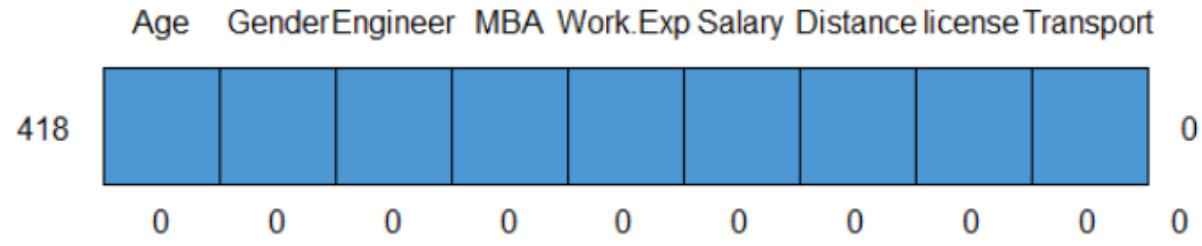
```
[1] TRUE
[1] 1
      Age    Gender  Engineer       MBA  Work.Exp    Salary  Distance
license Transport
        0         0         0         1         0         0         0
0         0
    Age Gender Engineer Work.Exp Salary Distance license Transport MBA
417   1      1        1        1      1        1       1         1   1 0
1     1      1        1        1      1        1       1         1   0 1
      0      0        0        0      0        0       0         0   1 1
```

Let's generate a graphical representation of the "Missing Values" in the given dataset.



The dataset contains 1 missing value in “MBA” attribute. The graphical representation below shows the each observation presence and absence of all 9 attributes in the dataset.

I used "mice" R library function algorithm to study the overall pattern of the dataset. Based on the pattern of the data distribution R had filled the missing value as "0". The graphical and sum calculation shown above on the treated new dataset confirms the elimination of NAs.



By studying the classes of all 9 attributes it can be said that "Engineer", "MBA" and "License" attributes needs to be converted into a factor. The overall dataset contains of 418 observations with 9 attributes. After the class transformation “str” function is used to study the class type of each variables.

```
'data.frame':   418 obs. of  9 variables:
 $ Age      : int  28 24 27 25 25 21 23 23 24 28 ...
 $ Gender   : Factor w/ 2 levels "Female","Male": 2 2 1 2 1 2 2 2 2 2 ...
 $ Engineer : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 2 1 2 2 ...
 $ MBA      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
 $ Work.Exp : int  5 6 9 1 3 3 3 0 4 6 ...
 $ Salary   : num  14.4 10.6 15.5 7.6 9.6 9.5 11.7 6.5 8.5 13.7 ...
 $ Distance : num  5.1 6.1 6.1 6.3 6.7 7.1 7.2 7.3 7.5 7.5 ...
 $ license  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
 $ Transport: Factor w/ 3 levels "2Wheeler","Car",..: 1 1 1 1 1 1 1 1 1 1 ...
```

The “5 Number summary” on the overall dataset was run and the output is given below.

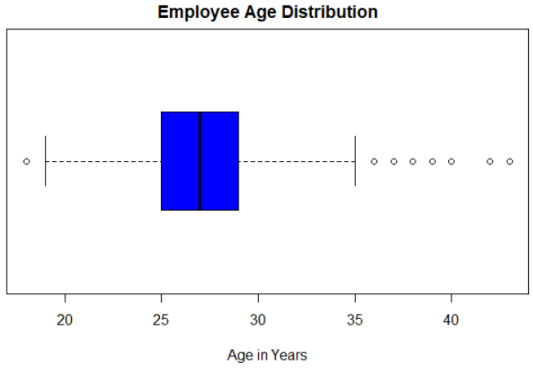
```
      Age           Gender    Engineer MBA        Work.Exp         Salary     
 Min.   :18.00   Female:121   0:105    0:309   Min.   : 0.000   Min.   : 6.500  
 1st Qu.:25.00   Male  :297   1:313    1:109   1st Qu.: 3.000   1st Qu.: 9.625  
 Median :27.00                                Median : 5.000   Median :13.000  
 Mean   :27.33                                Mean   : 5.873   Mean   :15.418  
 3rd Qu.:29.00                                3rd Qu.: 8.000   3rd Qu.:14.900  
 Max.   :43.00                                Max.   :24.000   Max.   :57.000  
    Distance      license            Transport  
 Min.   : 3.20   0:333   2Wheeler        : 83  
 1st Qu.: 8.60   1: 85   Car             : 35  
 Median :10.90           Public Transport:300  
 Mean   :11.29                                 
 3rd Qu.:13.57                                 
 Max.   :23.40                                 
```

As we can see from the figure above that all attributes are fixed with the appropriate class type for the analysis. Also, all “NA” entries have been treated and converted into a meaningful data for analysis.

3.3 Univariate Analysis

AGE:

As you can see the 5 number summary of the treated and prepared dataset that "Age" variable have a mean of 27.33 and median of 27 while the minimum value being 18 and maximum being 43. By theory if the mean is greater than the median, would make the distribution to be the Right Skewed. However, the difference between mean and median is not very noticeable. In the histogram plot shown below it can be said that "Age" variable observation is normally distributed and there could be some outliers at the upper value of the distribution. Let's analyze the outliers by doing an outlier examinations.



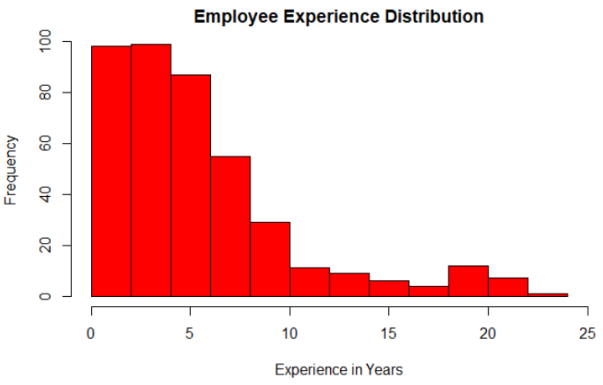
As observed from the "Age" histogram, most number of outliers were observed at the upper side of the age spectrum. This is confirmed by the boxplot diagram. The calculated 26 outliers are listed out below:

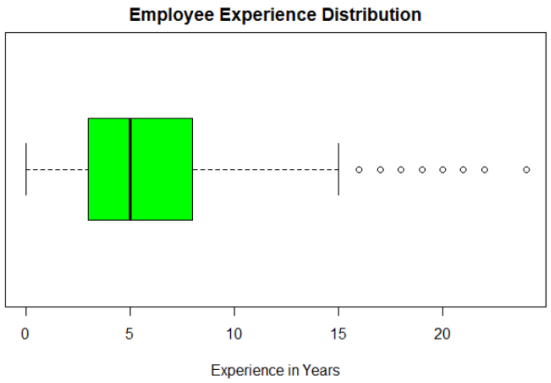
38 38 40 36 40 37 39 40 38 36 39 38 40 39 38 42 40 37 43 40 38 37 37 39 36 36

In addition to that there were two repeated outliers were observed at lower end of the distribution spectrum.i.e: 18, 18.

WORK EXPERIENCE:

The "Work.Exp" attribute have a mean of 5.873 and the median of 5 with a minimum of "No Experience" at all and maximum of 24 years. By theory and the graphical representation it can be said that work experience takes a Rightly Skewed distribution. Equal and tall bars were observed between the values of 0 to 5. Also, I suspect there could be outlier at the right side of the spectrum.



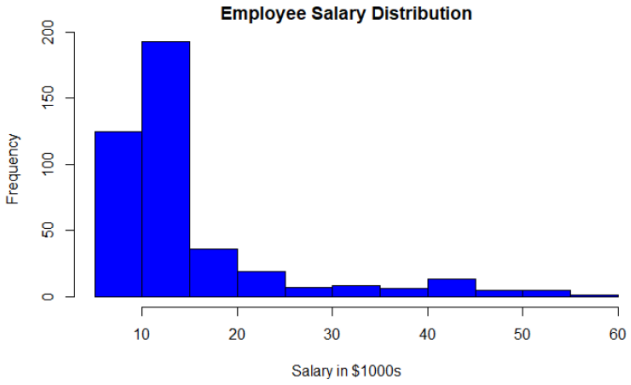


As suspected it can be seen from the outlier calculation that, there are no minimum outliers in the dataset. In contrast to that, 29 data points were identified as outliers located beyond the Max_IQR value for employee gained “Work Experience” attribute. The each observed outliers are listed below.

19 20 22 16 20 18 21 20 20 16 17 21 18 20 21 19 22 22 19 24 20 19 19 19 21 16 16 18 16.

SALARY:

The mean salary of the dataset is approximately 15.418k and median of 13k. The distribution contains employee’s salary in the range of 6.5k to 57k. The below histogram graph indicates that distribution is negatively skewed to the left. Therefore, without hesitation let us proceed to the outlier testing.



52 outlier points were identified at the right side of the "Salary" variable distribution spectrum.

The list of outliers are:

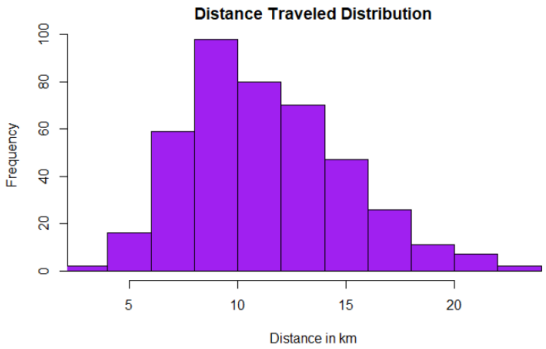
23.8 36.9 28.8 37.0 23.8 23.0 48.0 42.0 51.0 45.0 34.0 45.0 42.9 41.0 40.9 30.9 41.9 43.0 33.0

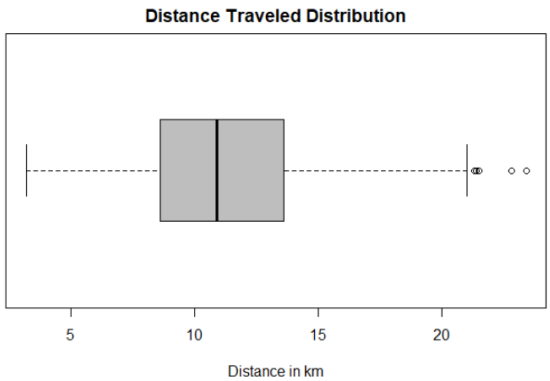
36.0 33.0 38.0 46.0 45.0 48.0 35.0 51.0 51.0 55.0 45.0 42.0 52.0 38.0 57.0 44.0 45.0 47.0 50.0

36.6 25.9 34.8 28.8 28.7 28.7 34.9 23.8 29.9 34.9 24.9 23.9 28.8 23.8.

DISTANCE:

The average distance travelled by each employee is about 11.9 km and the median distance is about 10.90km. The closest distance of the employee from the workplace is about 3.2km and furthest distance is about 23.40km. Both the theory and the histogram plot confirms the positively skewed nature in the employee traveled distance related information.





Let us begin the outlier analysis on the "Distance" attribute. The maximum number of outliers are observed in the right side of the spectrum which is 6 in total while none from the negative spectrum of the distribution.

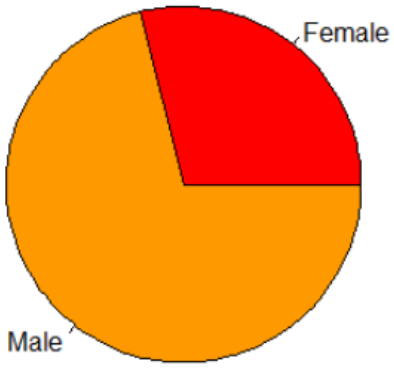
The observed outlier values are listed below:

21.3 21.4 21.5 21.5 22.8 23.4

GENDER:

Our dataset is been influenced by majority of Male employees which about 71% of the total dataset.

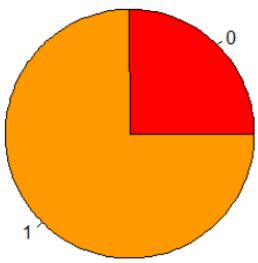
GENDER	NUMBER	%
Male	297	71%
Female	121	29%

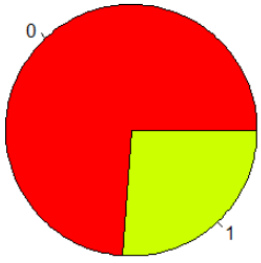


EDUCATION:

Total number of technically savvy people in the dataset is 313. That is around 75% of the total dataset.

EDUCATION	NUMBER	%
Engineer	313	74.8%
MBA	109	26.07%





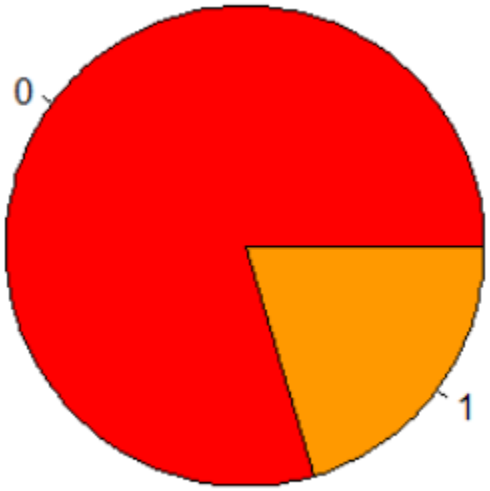
Pie-chart for Engineers *Pie-chart for MBA*

Number of employees from the Management division with a MBA qualification is 109. That is approximately a one quarter of the observations of the dataset.

LICENCE:

It was interesting to note that only 20% people carry license. Majority does not carry any driver's license. However, this attribute will influence from people choose cars as their mode of transport.

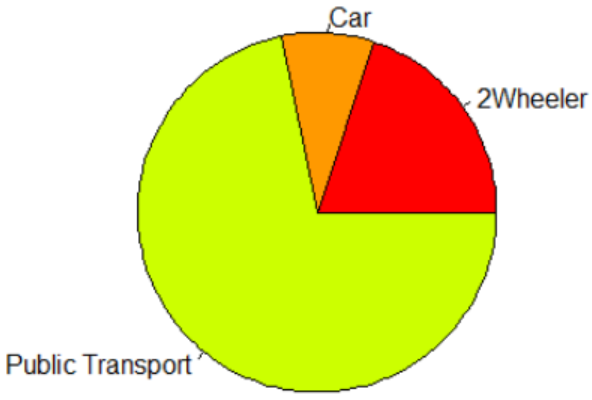
LICENSE	NUMBER	%
YES	85	20.33%
NO	333	79.67%



TRANSPORT:

Out of 418 observations 35 people either own or choose a car as a mode of transport and 83 of them own or choose 2wheeler as their mode of transport and the majority of people take the public transport.

LICENSE	NUMBER	%
Car	35	8.37%
2Wheeler	83	19.86%
Public	300	71.77%



3.4 Variable Transformation / Feature Creation

DUMMY VARIABLE – MODE OF TRANSPORT:

As we can see that the original dataset predicted variable mode of transport contains of 3 levels of categorical variables.

i.e: 2 Wheelers, Public Transport and Cars

Therefore, another columns have been created using "dummy variable" statistical tool to create minimize the number of levels in the predicted variable column. Therefore, a new column called "Transport.cars" have been created with a binary value 1 to represent all the employee who would take the car and "0" for other mode of transport.

Engineer <fctr>	MBA <fctr>	Work.Exp <int>	Salary <dbl>	Distance <dbl>	license <fctr>	Transport <fctr>	Transport.cars <fctr>
0	1	7	15.0	19.0	1	2Wheeler	0
1	0	4	13.0	19.1	1	2Wheeler	0
1	1	7	13.0	21.0	1	2Wheeler	0
1	0	19	48.0	14.1	1	Car	1
1	1	20	42.0	14.1	1	Car	1
1	0	22	51.0	14.1	1	Car	1
1	0	16	45.0	14.4	1	Car	1
1	0	12	34.0	14.4	1	Car	1
0	0	10	15.9	14.6	0	Car	1
1	1	10	15.8	14.6	1	Car	1

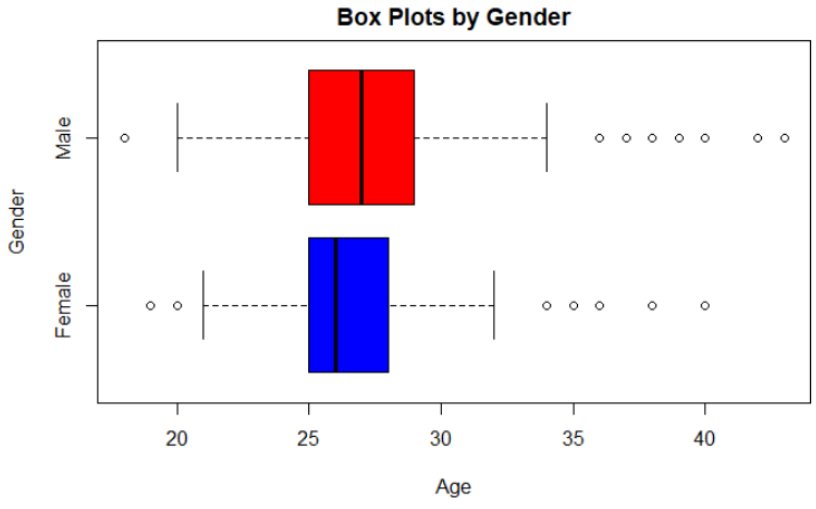
81-90 of 150 rows | 4-11 of 10 columns Previous 1 ... 7 8 9 10 11 ... 15 Next

DUMMY VARIABLE – MODE OF TRANSPORT:

3.5 BI-VARIATE ANALYSIS

AGE Vs GENDER

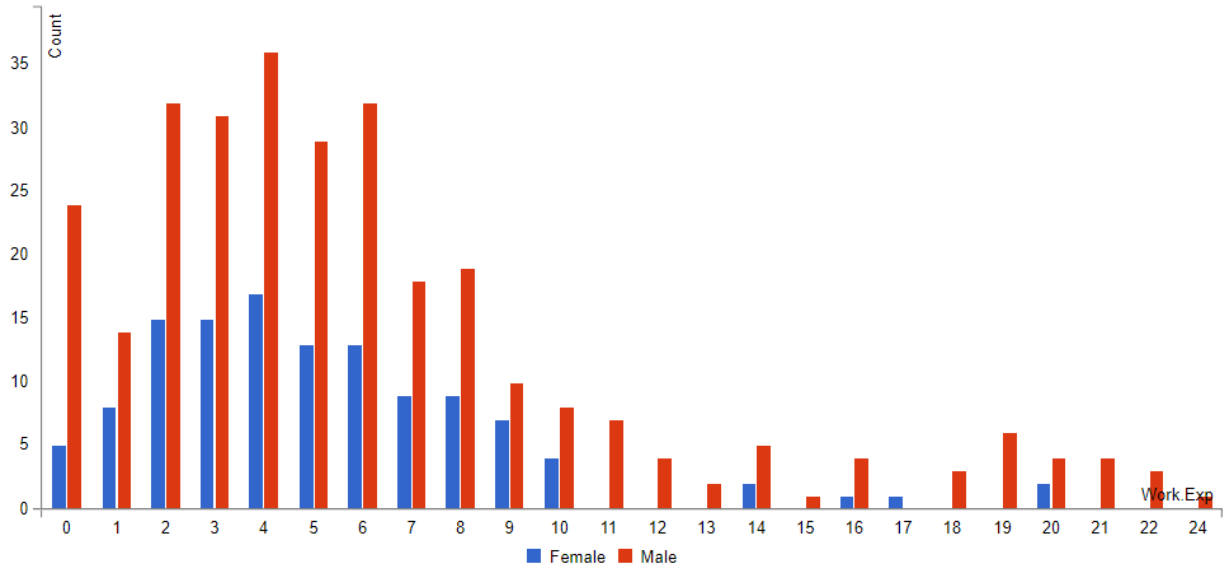
Let's investigate the type of people who are involved in the dataset.



The average mean age for male is about 28 and female being 27.

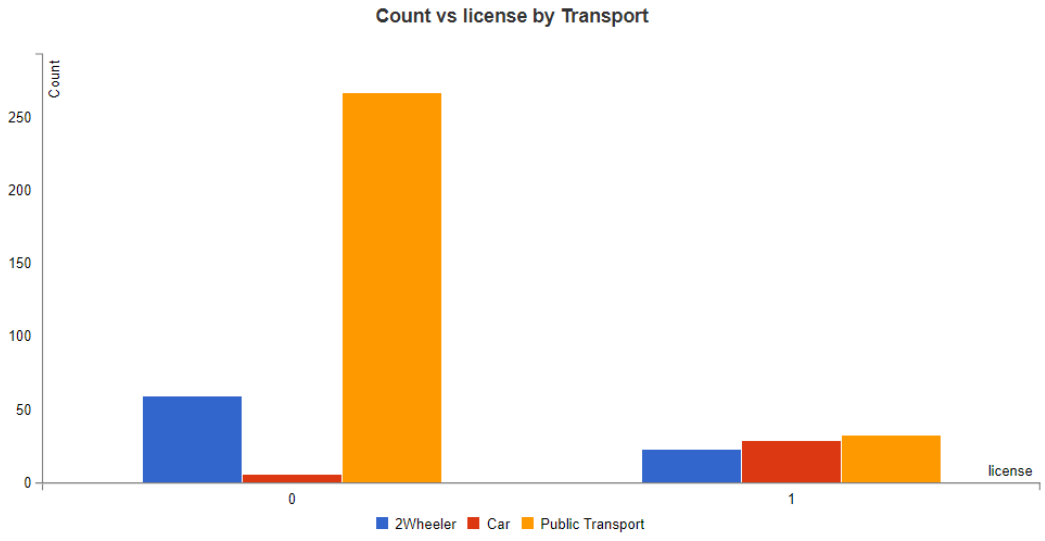
GENDER VS WORKEXPERIENCE

It can be said that at all experience level it's dominated by male. It's observable this dataset consists very large volume male employees at the early stage of the work experience level.



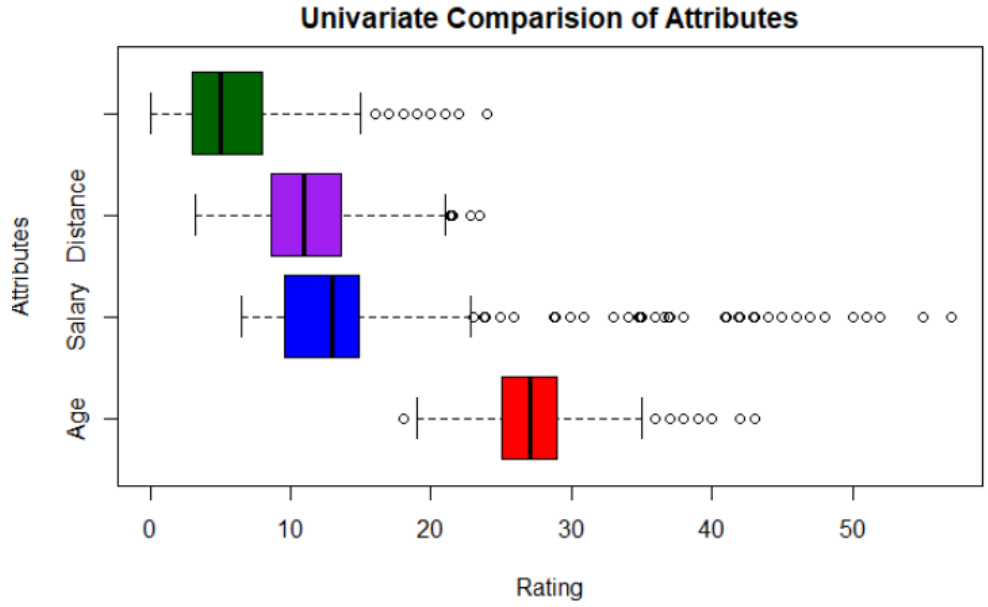
LICENSE Vs TRANSPORT AS CARS

Its interesting note that there are noticeable amount of people who take public transport have driving license while small proportion of the crowd does not hold a driver's license still take car as their preferred mode of transport.



3.6 MULTIVARIATE ANALYSIS

The box plot here gives the overall distribution of all the categorical variable in given dataset.



The plot below shows the multi-collinearity between each continuous variables.

We can see from the above correlation plot that Salary is very well correlated with both the Age and the Work experience to each other.

ATTRIBUTES	% of MULTI_COLINEARITY
Age Vs Salary	86%
Work.Exp Vs Salary	93.2%
Age Vs Work.Exp	93%
Distance Vs Salary	48%
Age Vs Distance	37.55%

4. CHALLENGING ASPECT OF THE PROJECT

The most challenging aspect of the project is identifying the correct independent variable due to their high multi-collinearity between each other. The size of the dataset is releatively smaller to make prediction especially with high correlated data.

Other than that all aspects of the projects are straight forward for me.

5. PREDICTIVE DATA ANALYSIS

5.1 LOGISTIC REGRESSION ANALYSIS

As we all know logistic regression predictive algorithm analysis used to predict a categorical outcome using a probabilities. Let's prepare the Train and Test dataset from the given dataset in the ratio of 70% to 30% using the set seed (1).

Let us create our first Logistic Regression model using "glm" function. By assuming that people who are financially sound will be very much interested in choosing a car as their mode of transport. So the first model is built using "Salary" as the independent variable.

DATASET	DIMENTIONS
Train	292 x 10
Test	126 x 10

Our first logistic regression model is built to predict the mode of transport against the Salary variable. Below is the code for the build Model-1.

```
model.lm <- glm(formula = Transport.cars~Salary,data=train,family="binomial")
summary(model.lm)
```

The output of the Model-1 is given below.

```
Call:
glm(formula = Transport.cars ~ Salary, family = "binomial", data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.75145  -0.10158  -0.07530  -0.04352   3.13702

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.53683    1.55526  -6.132 8.68e-10 ***
Salary       0.29264    0.05121   5.715 1.10e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 179.992  on 291  degrees of freedom
Residual deviance:  38.605  on 290  degrees of freedom
AIC: 42.605

Number of Fisher Scoring iterations: 8
```

The summary of model.lm output is shown above at which the P-Value for the variable Salary is lower than 5% which is significant in predicting the dependent variable mode of transport. Now let's evaluate the model performance by running the confusion matrix.

```
          Reference
Prediction   0   1
         0 263   2
         1   2  25
```

MODEL-1 PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL
Accuracy	98.63%
Sensitivity	99.25%
Specificity	92.59%
Concordance	98.12%
Discordance	1.19%

Concordance value indicates the selection of threshold level. However our output shows almost an equal value on both concordance and accuracy of the model.

Now, we proceed by including more variables into our model to see the possibility of building a model with better predicting power. Due to the high multi-collinearity between each variable, we use AIC (Akaike Information Criteria) function to choose the suitable independent variable for the logistic regression model prediction. The below summary is the AIC function output with the lowest value of 41.39. The function recommends "Salary "+"Distance" variables as best independent variable to predict the mode of transportation.

```
Step:  AIC=41.39
Transport.cars ~ Salary + Distance

           Df Deviance    AIC
<none>          26.389 41.389
+ Age       1   22.625 42.625
+ Work.Exp  1   23.807 43.807
+ Female    1   25.268 45.268
+ Male      1   25.268 45.268
+ license   1   25.713 45.713
+ MBA       1   26.312 46.312
+ Engineer  1   26.389 46.389
- Distance  1   38.605 48.605
- Salary    1   86.872 96.872
```

The model-2 is created with the "Salary"+"Distance" and the model would look like the screenshot shown below.

```
Call:
glm(formula = Transport.cars ~ Salary + Distance, family = "binomial",
    data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.20182  -0.06823  -0.03259  -0.01240   2.90919

Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -17.32448    4.08608  -4.240 0.0000224 ***
Salary        0.25176    0.05722   4.400 0.0000108 ***
Distance      0.62532    0.21515   2.906   0.00366 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 179.992  on 291  degrees of freedom
Residual deviance:  26.389  on 289  degrees of freedom
AIC: 32.389

Number of Fisher Scoring iterations: 9
```

As we can see both P-values of Salary and Distance look very significant. Now we can comfortably proceed with the prediction with this model. However, based on the nature of the given categorical variable, it's advisable to include a factor attribute of the dataset too.

The method used to select the suitable categorical variable into the model is by calculating the rate of particular class employees of selecting the mode of transport as cars. If the difference between two classes are significantly larger that categorical variable will be selected.

GENDER: Rate of Female Car owners is about 4.9% and the male car owners being 9.7%. The difference between these two classes is very low and not very significant. Let us try with the Engineer Class.

ENGINEER: Rate of Engineers Car owners is about 9.5% and the male car owners being 4.7%. The difference between these two classes is very low and not significant. Let us try with the MBA Class.

MBA: Rate of MBA degree holders who are Car owners is about 8.2% and the male car owners being 8.4%. The difference between these two classes is very low and not significant. Let us try with the License Class.

LICENSE: Rate of License holders who are Car owners is about 34% and the male car owners being 1.8%. The difference between these two classes is high and significant. Let us add License categorical variable into our model as an Independent Variable.

CREATING MODEL WITH A CATERGORICAL VARIABLE.

```
model.mlr.step.cat<- glm(formula =
Transport.cars~Salary+Distance+license,data=train.car,family="binomial")
```

The below shows the output logistic model prediction of top10 rows.

	real <fctr>	Salary.predict <dbl>	Salary.class <fctr>	mlr.step.cat.predict <dbl>	mlr.step.cat.class <chr>
1	0	0.0038457108	0	0.00131187190	0
2	0	0.0027902035	0	0.00009251173	0
3	0	0.0037352153	0	0.00009748642	0
4	0	0.0051463666	0	0.00284629607	0
5	0	0.0031356099	0	0.00043212200	0
6	0	0.0022091015	0	0.00011312108	0
7	0	0.0056159487	0	0.00134912910	0
8	1	0.9401391262	1	0.91532306876	1
9	0	0.2426704142	0	0.01962455671	0
10	0	0.0022091015	0	0.00172797294	0

Now let's evaluate the model performance by running the confusion matrix.

```
          Reference
Prediction   0   1
         0 264   1
         1   2  25
```

As you can see in this model misclassification have improved. Let's see the other aspect of the model. The specificity have improved by approximately 4%. The concordance have reduced by 6%.

MODEL-1 PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL
Accuracy	98.97%
Sensitivity	99.25%
Specificity	96.15%
Concordance	92.6%
Discordance	7.4%

Let's check the model performance with the optimal threshold value calculated by the "optimalCutoff" R studio function. Model prediction will be categorized with a different cut off value of 0.64.

The Confusion Matrix and the Concordance & Discordance Ratio are tabulated below.

```
          Reference
Prediction   0   1
         0 265   0
         1   2  25
```

Evidently the new cutoff value have given a better result. The accuracy and specificity have increased remarkably on the prediction conducted on the train dataset.

MODEL-1 PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL
Accuracy	99.32%
Sensitivity	99.25%
Specificity	100%
Concordance	92.6%
Discordance	7.4%

In the next section the predicting power of the model will be put against Test dataset.

PERFORMANCE MEASURE ON THE TEST DATASET – LOGISTIC REGRESSION MODEL

Same model is run and the Confusion matrix and Concordance & Discordance ratio is tabulated below.

```
          Reference
Prediction   0   1
         0 118   0
         1   2   6
```

MODEL-1 PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL
Accuracy	98.41%
Sensitivity	98.33%
Specificity	100%
Concordance	99.79%
Discordance	0.21%

DISCUSSION ON LOGISTIC REGRESSION MODEL PERFORMANCE MEASURE

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL COMPARISON BETWEEN TRAIN & TEST		
	TRAIN	TEST	% DIFFERENCE
Accuracy	99.32%	98.41%	0.91
Sensitivity	99.25%	98.33%	0.92
Specificity	100%	100%	0
Concordance	92.6%	99.79%	-7.19
Discordance	7.4%	0.21%	7.19

The variation between the Train and Test model performance values lies less than +/- 10% which is a good indication of a good model. Hence, we can say that the developed LOGISTIC REGRESSION model have good prediction making abilities.

5.2 K-NEAREST NEIGHBOURS ALGORITHM ANALYSIS

KNN is a non-parametric lazy learning algorithm used to analyze the datasets in which data points are separated into several classes to predict classification of a new predictive sample point.

The "class" R library is used to do the KNN analysis. The function used I "knn3". Refer to APPENDIX for the R-Studio Codes.

The figure below illustrates the top ten rows of the KNN prediction when K = 11 and Optimal Cutoff threshold value of 0.5.

	real <fctr>	nb.prob <dbl>	knn.prob.11 <dbl>	knn.class.11 <fctr>
1	0	4.391138e-07	0.0000000	0
2	0	5.005316e-11	0.0000000	0
3	0	1.966830e-08	0.0000000	0
4	0	6.020138e-05	0.0000000	0
5	0	2.599797e-11	0.0000000	0
6	0	1.656313e-11	0.0000000	0
7	0	4.947652e-08	0.0000000	0
8	1	1.000000e+00	1.0000000	1
9	0	5.239457e-03	0.1818182	0
10	0	1.164826e-08	0.0000000	0

Let's evaluate the above model with the confusion matrix and Concordance- Discordance Ratio.

CONFUSION MATRIX K=11:

The output of the confusion matrix is given below:

```
          Reference
Prediction   0   1
         0 264   1
         1   3  24
```

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL
Accuracy	98.63%
Sensitivity	98.88%
Specificity	96.00%
Concordance	99.65%
Discordance	0.35%

The accuracy of the model looks good. However, there are 4 mis-classification can be observed. Let's fine tune the model by increasing the k value to be 21.

The top 10 rows of the model with the K=21.

	real <fctr>	nb.prob <dbl>	knn.prob.11 <dbl>	knn.class.11 <fctr>	knn.prob.21 <dbl>	knn.class.21 <fctr>
1	0	4.391138e-07	0.0000000	0	0.0000000	0
2	0	5.005316e-11	0.0000000	0	0.0000000	0
3	0	1.966830e-08	0.0000000	0	0.0000000	0
4	0	6.020138e-05	0.0000000	0	0.0000000	0
5	0	2.599797e-11	0.0000000	0	0.0000000	0
6	0	1.656313e-11	0.0000000	0	0.0000000	0
7	0	4.947652e-08	0.0000000	0	0.0000000	0
8	1	1.000000e+00	1.0000000	1	0.9047619	1
9	0	5.239457e-03	0.1818182	0	0.2380952	0
10	0	1.164826e-08	0.0000000	0	0.0000000	0

CONFUSION MATRIX K=21:
The output of the confusion matrix is given below:

```
          Reference
Prediction   0   1
         0 263   2
         1   6  21
```

It can be observed that the total miss-classification increases to 8.

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL	
	K=11	K=21
Accuracy	98.63%	97.26%
Sensitivity	98.88%	97.77%
Specificity	96.00%	91.30%
Concordance	99.65%	99.45%
Discordance	0.35%	0.55%

It can be observed that by increasing the K value does not improve the accuracy, sensitivity and the specificity of the model.

The next option is to fine tune the KNN model prediction using the optimal cutoff threshold value by using the R "optimalCutoff" function.

Cutoff Threshold value for K=11 is 0.37

Cutoff Threshold Value for K =21 is 0.29

Let's check the model strength using the new cutoff threshold value for both the models (I:e: K=11,K=21)

Confusion Matrix, K=11, Threshold Value = 0.37

```
          Reference
Prediction   0   1
         0 264   1
         1   2  25
```

Confusion Matrix, K=21, Threshold Value = 0.29

```
          Reference
Prediction   0   1
         0 262   3
         1   2  25
```

Comparison of all model measures we have built so far,

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL			
	K=11 (0.5)	K=11 (0.37)	K=21 (0.5)	K=21 (0.29)
Accuracy	98.63%	98.97%	97.26%	98.29%
Sensitivity	98.88%	99.25%	97.77%	99.24%
Specificity	96.00%	96.15%	91.30%	89.29%
Concordance	99.65%	99.65%	99.45%	99.45%
Discordance	0.35%	0.35%	0.55%	0.55%

The table above shows the model strength with new optimal threshold values. Although the above output is somewhat satisfactory it can be observed that, after a trial and error method of running the

model with different K value, model gives the best performance at K=13 with the optimal threshold value of 0.5.

The CONFUSION MATRIX:

```
          Reference
Prediction   0   1
         0 264   1
         1   3  24
```

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL
Accuracy	98.63%
Sensitivity	98.88%
Specificity	96.00%
Concordance	99.65%
Discordance	0.35%

Let's try the train model to check the performance with the test dataset.

PERFORMANCE MEASURE ON THE TEST DATASET – KNN MODEL

The procedure is exactly the same to carry on the performance measure on the test dataset of the KNN model. Also, same type of packages have been used for this procedure.

THE CONFUSION MATRIX OUTPUT:

```
          Reference
Prediction   0   1
         0 118   0
         1   2   6
```

Actually the model performs well with predicting the mis-classification compared to the train data on test data. Misclassification is limited to 2 on test data.

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL
Accuracy	98.41%
Sensitivity	98.33%
Specificity	100%
Concordance	98.83%
Discordance	1.17%

DISCUSSION ON KNN MODEL PERFORMANCE MEASURE

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL COMPARISON BETWEEN TRAIN & TEST		
	TRAIN	TEST	% DIFFERENCE
Accuracy	98.63%	98.41%	0.22
Sensitivity	98.88%	98.33%	0.55
Specificity	96.00%	100%	4
Concordance	99.65%	98.83%	0.82
Discordance	0.35%	1.17%	-0.82

The variation between the Train and Test model performance values lies less than +/- 10% which is a good indication of a good model. Hence, we can say the developed KNN model have good prediction making abilities.

5.3 NAÏVE BAYES ALGORITHM ANALYSIS

Naïve Bayes is a predictive modeling algorithm at which works by computing a probability and stores in a table format. Every time a new prediction is expected for anew data point based on the stored event probabilities a prediction is done.

Naïve Bayes model shows below: All variables are included in prediction.

```
#Creating a new dataset with the results
nb.results.df <- data.frame(real=train.nb$Transport.cars)
# Naive Bayes Model Building
model_cars =caret::train(train.nb[,-9],as.factor(train.nb[,9]),'nb',
                         trcontrol=trainControl(method='cv',number=5))
```

The top 10 rows of the predicted probabilities and classes are tabulated below. The cutoff threshold value is 0.5

	real <fctr>	nb.prob <dbl>	nb.class <fctr>
1	0	4.391138e-07	0
2	0	5.005316e-11	0
3	0	1.966830e-08	0
4	0	6.020138e-05	0
5	0	2.599797e-11	0
6	0	1.656313e-11	0
7	0	4.947652e-08	0
8	1	1.000000e+00	1
9	0	5.239457e-03	0
10	0	1.164826e-08	0

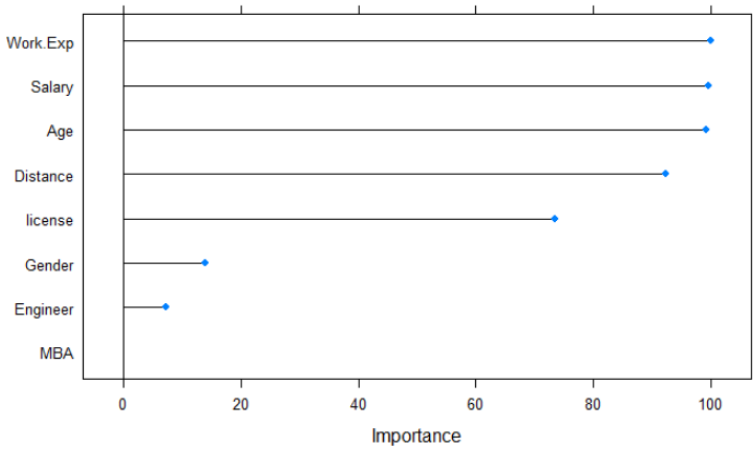
The Confusion Matrix and Concordance & Discordance ratio of the built model is summarized below.

```
          Reference
Prediction   0   1
         0 263   2
         1   2  25
```

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL
Accuracy	98.63%
Sensitivity	99.25%
Specificity	92.59%
Concordance	99.80%
Discordance	0.2%

The built model with a 50% threshold cutoff value have a good predicting power.

The plot below shows the Imprtance score against each variables. Based on the above plot it can be said that "Work.Exp","Salary,"Age" have almost 100% score and they are very significant. "Distance" attribute is also very significant due to second high score of 90%+. In my opinion I decided to add "license" attribute too due to the 77% of reasonably high score. The "Gender", "Enginner" and "MBA" attributes could assumed to be insignificant during Naive bayes analysis due to the low score.



Let us run the Naive Bayes using selected attributes:

```
# Naive Bayes Model Building
model_cars.t =caret::train(train.nb[,-c(2,3,4,9)],as.factor(train.nb[,9]),'nb',
                           trcontrol=trainControl(method='cv',number=10))
```

The top 10 rows of the predicted probability and the class. The selected cutoff threshold is 0.77.

	real <fctr>	nb.prob <dbl>	nb.class <fctr>	nb.prob.t <dbl>	nb.class.t <fctr>
1	0	4.391138e-07	0	1.097437e-06	0
2	0	5.005316e-11	0	3.682698e-11	0
3	0	1.966830e-08	0	1.447110e-08	0
4	0	6.020138e-05	0	6.996051e-05	0
5	0	2.599797e-11	0	2.000150e-11	0
6	0	1.656313e-11	0	2.620771e-11	0
7	0	4.947652e-08	0	3.640272e-08	0
8	1	1.000000e+00	1	1.000000e+00	1
9	0	5.239457e-03	0	8.265135e-03	0
10	0	1.164826e-08	0	8.570294e-09	0

The Confusion Matrix and Concordance & Discordance ratio to check the predicting strength of the new model.

```
          Reference
Prediction   0   1
         0 264   1
         1   2  25
```

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL	
	All Variables	Fined Tuned Variables
Accuracy	98.63%	98.97%
Sensitivity	99.25%	99.25%
Specificity	92.59%	96.15%
Concordance	99.80%	99.82%
Discordance	0.2%	0.18%

As you can see from the above table fine tuning have improved the predicting power of the model.

PERFORMANCE MEASURE ON THE TEST DATASET – NAÏVE BAYES MODEL

The procedure is exactly the same to carry on the performance measure on the test dataset of the NB model. Also, same type of packages have been used for this procedure.

THE CONFUSION MATRIX OUTPUT:

```
          Reference
Prediction   0   1
         0 117   1
         1   1   7
```

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL
Accuracy	98.41%
Sensitivity	99.15%
Specificity	87.50%
Concordance	99.68%
Discordance	0.32%

DISCUSSION ON KNN MODEL PERFORMANCE MEASURE

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL COMPARISON BETWEEN TRAIN & TEST		
	TRAIN	TEST	% DIFFERENCE
Accuracy	98.97%	98.41%	0.56
Sensitivity	99.25%	99.15%	0.1
Specificity	96.15%	87.50%	8.65
Concordance	99.82%	99.68%	0.14
Discordance	0.18%	0.32%	-0.14

The variation between the Train and Test model performance values lies less than +/- 10% which is a good indication of a good model. Hence, we can say the developed NB model have good prediction making abilities.

6. DATA ENSEMBLE METHODS

6.1 SMOTE

As we can see that data is pretty well balanced by having upper class as around 92% and 8% of lower class data. I believe data is pretty well balanced so there is no need of creating a synthetic data technique application like SMOTE.

6.2 BAGGING

Bagging technique is used to minimize the variance in the dataset. In the case of an outlier, bagging technique will focus on the entire dataset to create new rows which makes the outliers a less of an influential factor.

The Confusion Matrix of the built model is summarized below.

```
          Reference
Prediction   0   1
         0 264   1
         1   0  27
```

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL
Accuracy	99.66%
Sensitivity	100%
Specificity	96.43%

PERFORMANCE MEASURE ON THE TEST DATASET – BAGGING MODEL

The procedure is exactly the same to carry on the performance measure on the test dataset of the BAGGING model. Also, same type of packages have been used for this procedure.

THE CONFUSION MATRIX OUTPUT:

```
          Reference
Prediction   0   1
         0 117   1
         1   1   7
```

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL
Accuracy	98.41%
Sensitivity	99.15%
Specificity	87.50%

DISCUSSION ON BAGGING MODEL PERFORMANCE MEASURE

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL COMPARISON BETWEEN TRAIN & TEST		
	TRAIN	TEST	% DIFFERENCE
Accuracy	99.66%	98.41%	1.25
Sensitivity	100%	99.15%	0.85
Specificity	96.43%	87.50%	8.93

The variation between the Train and Test model performance values lies less than +/- 10% which is a good indication of a good model. Hence, we can say the developed BAGGING model have good prediction making abilities.

6.3 BOOSTING

This ensemble technique is used to create strong learner from a weak learner. This tool helps to minimize biasness of the model.

6.3.1 XGBOOST BOOSTING

Unlike other algorithm gradient boosting focus only on the accuracy of the model. If the error rate reduces the iteration will be continued. If not it will be stopped.

XG BOOST model:

```
xgb.fit <- xgboost(data=xg_features_train,
                   label=xg_label_train,
                   eta=0.001,
                   max_depth=3,
                   min_child_weight=3,
                   nrounds=10000,
                   nfold=5,
                   objective="binary:logistic",
                   verbose=0,
                   early_stopping_rounds=10)
```

The Confusion Matrix of the built model is summarized below.

```
          Reference
Prediction   0   1
         0 263   2
         1   2  25
```

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL
Accuracy	98.63%
Sensitivity	99.25%
Specificity	92.59%

PERFORMANCE MEASURE ON THE TEST DATASET – XGBOOST MODEL

The procedure is exactly the same to carry on the performance measure on the test dataset of the BAGGING model. Also, same type of packages have been used for this procedure.

THE CONFUSION MATRIX OUTPUT:

```
          Reference
Prediction   0   1
         0 114   4
         1   2   6
```

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL
Accuracy	95.24%
Sensitivity	98.28%
Specificity	60%

DISCUSSION ON XGBOOST MODEL PERFORMANCE MEASURE

MODEL PERFORMANCE TOOLS	GOODNESS OF THE TRAINING MODEL COMPARISON BETWEEN TRAIN & TEST		
	TRAIN	TEST	% DIFFERENCE
Accuracy	98.63%	95.24%	3.39%
Sensitivity	99.25%	98.28%	0.97%
Specificity	92.59%	60%	32.59%

The variation between the Train and Test model performance values lies less than +/- 10% Except the specificity of the model on test data is very low.

7. DISSCUSSION

Each of the Machine Learning algorithm have demonstrated different types of predicting power. The summary and a comparison is shown in the table below

PARAMETER	LOGISTIC REG	KNN	NB	BAGGING	XGBOOST
Accuracy	98.41%	98.41%	98.41%	98.41%	95.24%
Sensitivity	98.33%	98.33%	99.15%	99.15%	98.28%
Specificity	100%	100%	87.50%	87.50%	60%

According to the comparison above all model have shown a very good accuracy and sensitivity. In order to pick the best predicting model the last feature is used.ie: SPECIFICITY.

There are 9 different variables are used in this dataset at which they key most influencing attributes in predicting the mode of transport are Salary, Distance and License. All other variables can be eliminated due to the fact of multi-collinearity of the independent variables.

The XG boosting ensemble method focus only on accuracy except the misclassification. Therefore, the specificity of the model can be improved by fine tuning the XG boost model by converting the multi factor variables using dummy variables. Also, considering the given data size and the ease of implementation of other models over powers the XG Boost model.

Based on the overall findings it can be said that all models have shown a good level accuracy in predicting however, my personal choice is the logistic regression tool due to its nature of indicating the relationship of each co-efficient on the predicting variable. In addition to that logistic regression is very simple to train and test the model.

8. CONCLUSION

After the above findings it can be said that the distance travelled by the staff is a vital decision maker in choosing the mode of transport. It can be said approximately, staff who live at about 10km and more are very much interested in choosing the car. Most importantly, I see investing in a personal vehicle seems to be a very common thing among the people who get higher salaries. This is an important factor when it comes in purchasing a car by full cash or through bank loan. In conclusion we can say that most number of employees prefer car as their mode of transport.