# 

# Machine Learning Group Assignment

**Submitted By:**

PGP-BABI, Marketing & Retail analytics Group -2

Arun Ukamanal

Lakshmi Sree Jakkaraju Meenakshi Sundaram

M.Ramachandran

Roy Cherian Cherukarayil

S.R.Vignesh

Contents

[Machine Learning Group Assignment 1](#_heading=h.gjdgxs)

[Problem Statement: 3](#_heading=h.1fob9te)

[Exploratory Data Analysis: 3](#_heading=h.3znysh7)

[Basic Data Summary: 3](#_heading=h.2et92p0)

[Univariate Analysis: 4](#_heading=h.tyjcwt)

[Bivariate Analysis: 6](#_heading=h.3dy6vkm)

[Missing value check: 9](#_heading=h.1t3h5sf)

[Outlier Analysis: 9](#_heading=h.4d34og8)

[Data preparation: 13](#_heading=h.2s8eyo1)

[KNN imputation: 13](#_heading=h.17dp8vu)

[Transport ratio: 13](#_heading=h.3rdcrjn)

[Car usage: 14](#_heading=h.26in1rg)

[Logistic Regression: 15](#_heading=h.lnxbz9)

[Model Interpretation: 16](#_heading=h.35nkun2)

[Logistic regression – Post ridge regression: 17](#_heading=h.1ksv4uv)

[Linear discriminant analysis: 19](#_heading=h.44sinio)

[Improve LDA Model by Regularization 20](#_heading=h.2jxsxqh)

[Prediction using CART 22](#_heading=h.z337ya)

[Boosting: 24](#_heading=h.3j2qqm3)

[Random forest: 25](#_heading=h.1y810tw)

[Bagging: 28](#_heading=h.4i7ojhp)

[References: 28](#_heading=h.2xcytpi)

## Problem Statement:

The attached data has records of 444 employees in a firm. The variables are described below

1. Age: Age of the employees
2. Gender: Gender of the employee
3. Engineer: Whether the employee is an Engineer
4. MBA: Whether the employee is a MBA
5. Work Exp (work experience in completed years)
6. Salary: Salary in Rs Lakhs
7. Distance (in Km): The distance between the employee’s residence and office
8. License: Whether the employee has a driving license
9. Transport: Main mode of transport taken by the employee
10. Data (first few records)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Gender | Engineer | MBA | Work Exp | Salary | Distance | license | Transport |
| 28 | Male | 0 | 0 | 4 | 14.3 | 3.2 | 0 | Public Transport |
| 23 | Female | 1 | 0 | 4 | 8.3 | 3.3 | 0 | Public Transport |
| 29 | Male | 1 | 0 | 7 | 13.4 | 4.1 | 0 | Public Transport |
| 28 | Female | 1 | 1 | 5 | 13.4 | 4.5 | 0 | Public Transport |
| 27 | Male | 1 | 0 | 4 | 13.4 | 4.6 | 0 | Public Transport |
| 26 | Male | 1 | 0 | 4 | 12.3 | 4.8 | 1 | Public Transport |
| 28 | Male | 1 | 0 | 5 | 14.4 | 5.1 | 0 | 2Wheeler |

Question.1: Build a model that best explains the employee’s decision to use cars as the main means of transport?

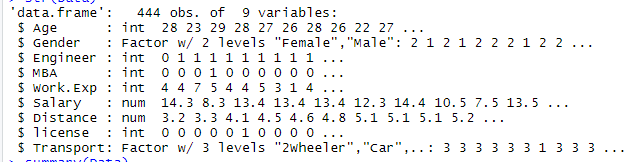
Question.2: What would your predictions regarding their choice of transport be for the following two employees?

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Age | Gender | Engineer | MBA | Work Exp | Salary | Distance | license |
| 25 | Male | 0 | 0 | 2 | 10 | 5 | 1 |
| 25 | Female | 1 | 0 | 2 | 10 | 5 | 0 |

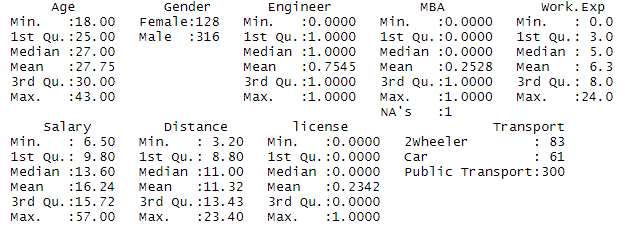
## Exploratory Data Analysis:

### Basic Data Summary:

1. Total no of rows: 444
2. Data Structure:



1. Data Summary:
   1. Age band: 18-43
   2. Gender: F:M=128:316



1. Engineers by Gender



1. MBA by Gender – There are no female MBA graduates



1. Work experience by Gender



1. Salary by Gender

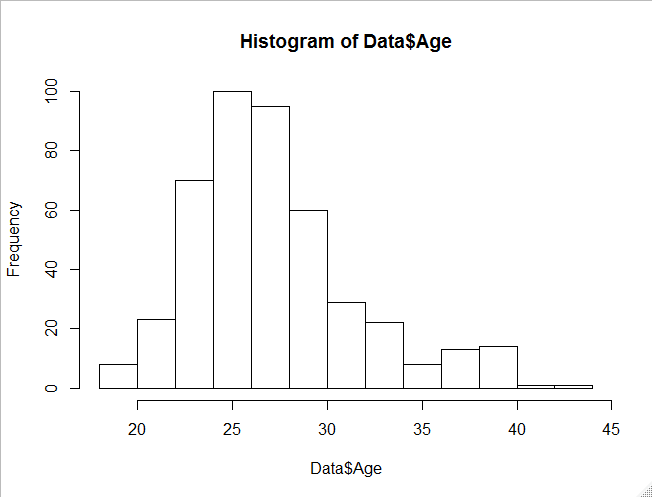
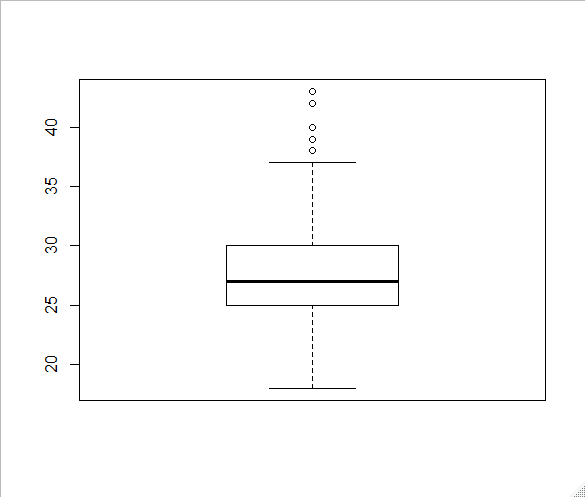


**Inferences:**

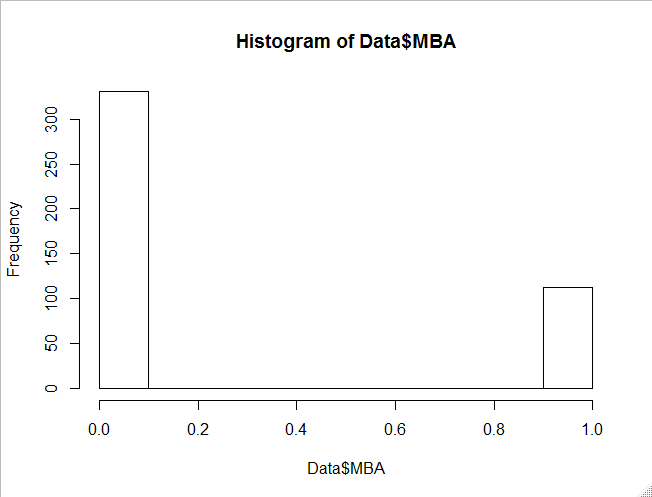
* More number of male employees (approximately 75%) when compared to female employees
* More male engineers (71%) than female engineers
* More number of engineers when compared to MBA’s
* No female MBA’s
* One data point for MBA is missing
* Salary is skewed towards male employees’ distribution
* Public transport is most common mode of transportation followed by 2-wheeler and car
* Marginally male employees experience is more than female employees

### Univariate Analysis:

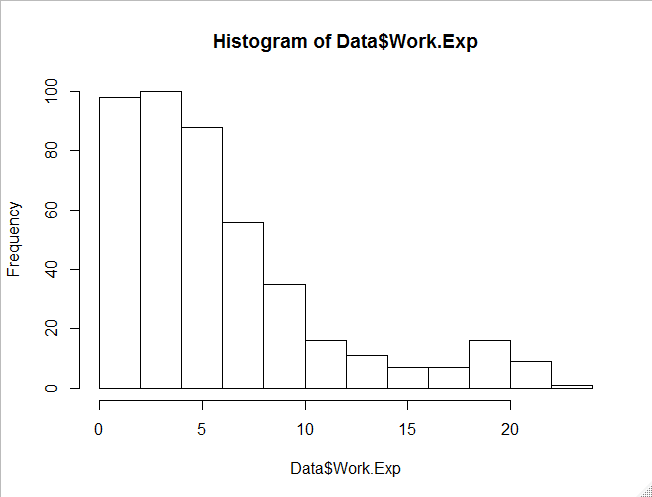
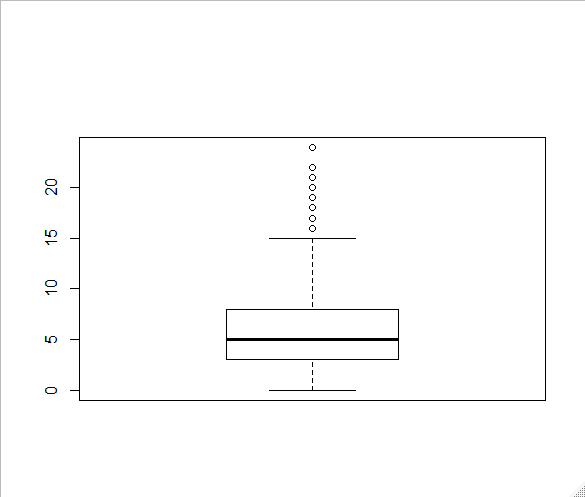
1. Age:



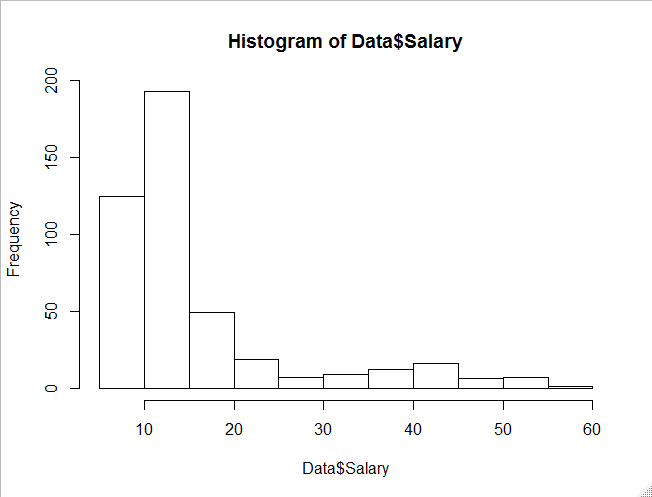
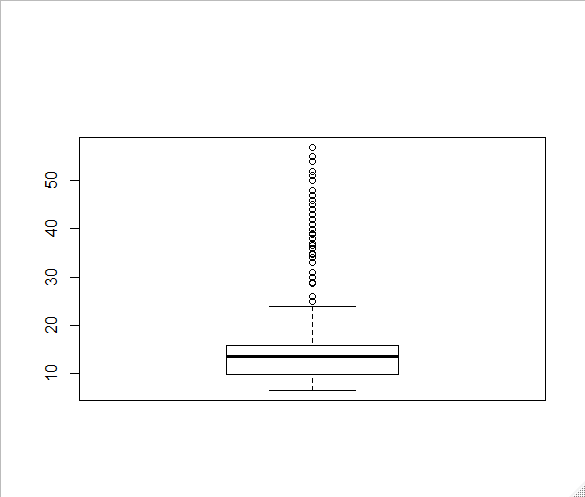
1. MBA:



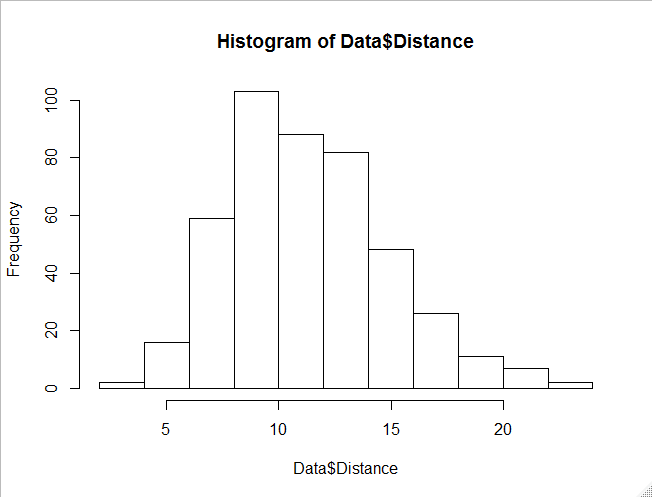
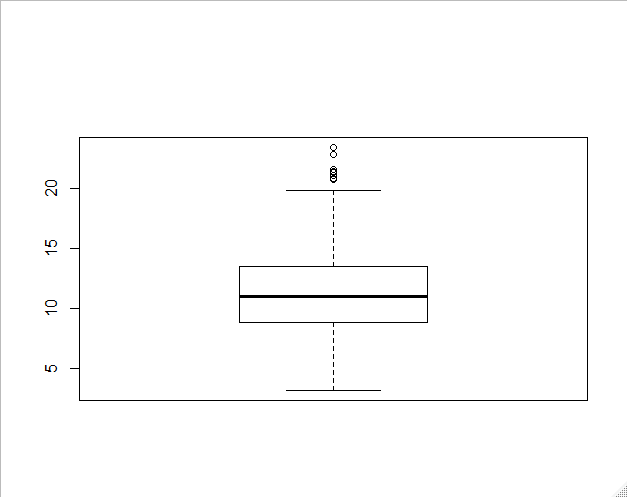
1. Work experience:



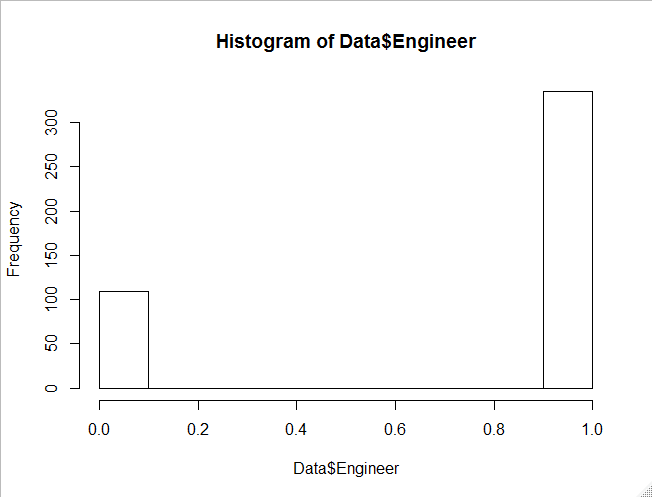
1. Salary



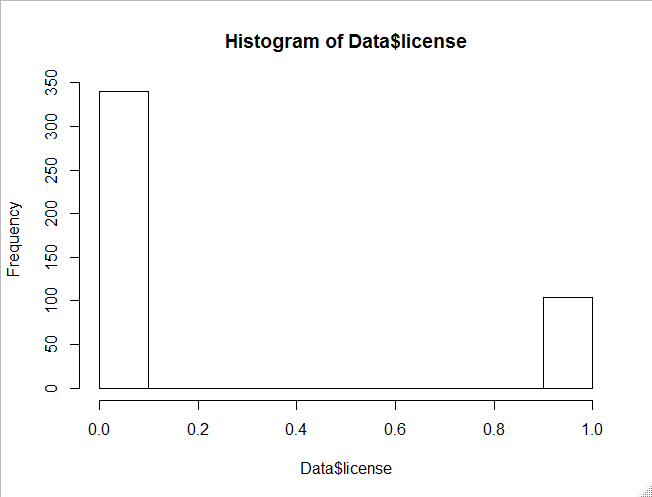
1. Distance



1. Engineer



1. License



**Inferences:**

Age:

* Right skewed distribution
* There are outliers in the Age which need to be analysed
* Majority of the sample is in the age group 22-30

Work experience:

* Right skewed distribution with more employees with work experience in the range 0-10 years.
* We could observe lot of outliers in the work experience

Salary:

* Right skewed distribution with more junior employees at lesser pay
* Salary also has lot of outliers which needs to be analysed
* Majority of the employee’s salary <=15 lakhs

More number of engineers than MBAs

Distance:

* Right skewed distribution with more employees commute from far distance
* There are very few outliers.
* Shapiro’s test of normality: p value<0.05, hence it is not normally distributed.
* Less number of employees (approximately 25%) have a driving licence.

Engineer:

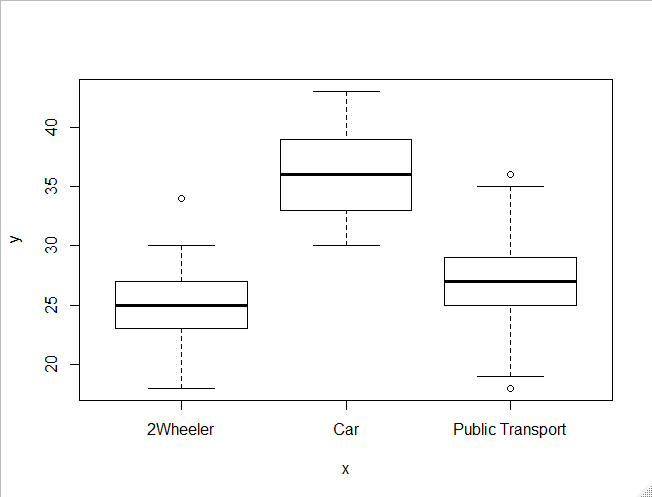
* Majority of employees are engineers (close to 300).

License:

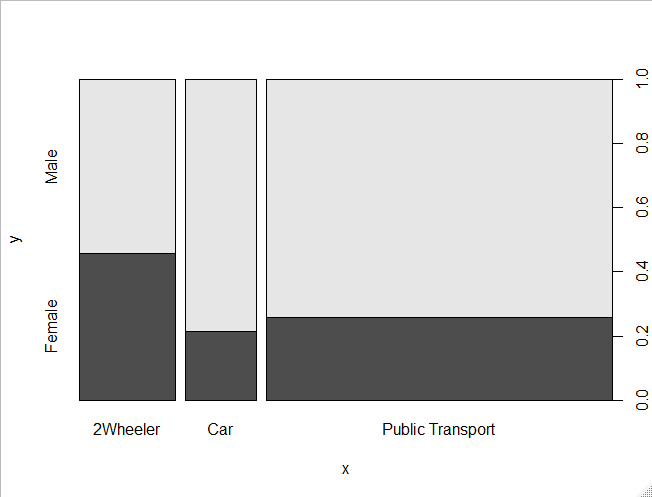
* There are more non license holders than the license holders. This will have a significant impact on the type of transport chosen.

### Bivariate Analysis:

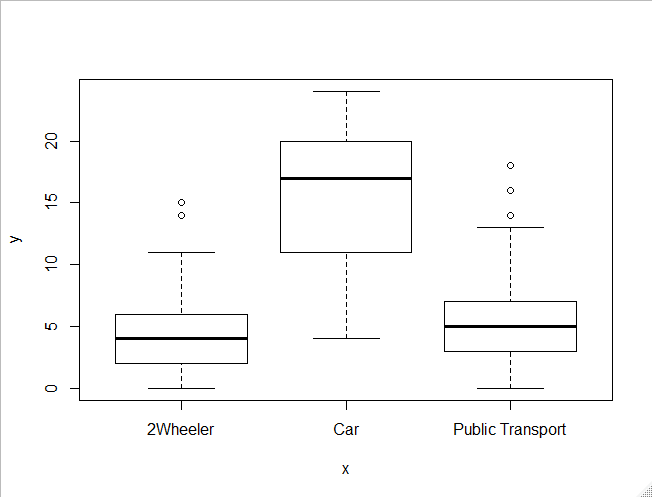
1. Transport by Age



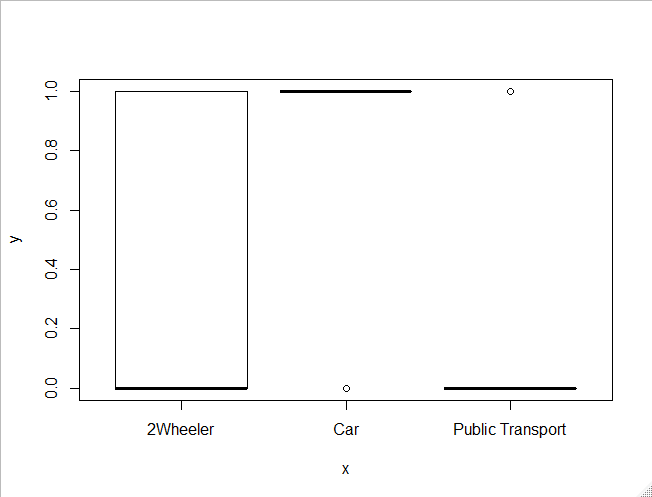
1. Transport by Gender



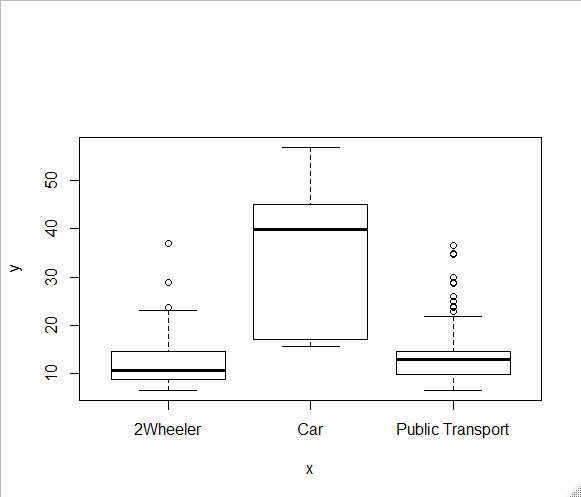
1. Transport by Work experience



1. Transport by License



1. Transport by Salary



**Inferences:**

Transport by age:

* Employees using cars as a mode of transport have a higher median age than non-car users
* So higher age seems to be a driving factor for selecting mode of transport
* Graph specifies that young aged people in the age group (23-28) are commuting by 2 wheeler
* Employees in the age group (25-30) prefer Public transport
* Employees above age 35 use car as mode of transport

Transport by gender:

* More number of male employees prefer cars as a mode of transport when compared with non-car & female employees. This may be due to gender-inequality in job & salary imparity
* In the graph, it appears that more than 70% of male employees and 60% of female employees use Public transport
* 30% of female employees and 10% of male employees use 2 wheeler
* Remaining use car as the mode of transport

Transport by work experience:

* Car users’ median work experience is much higher than non-car users.

So higher work experience seems to be a driving factor selecting mode of transport.

* With respect to work experience it is clear that people with less experience (<8 years) almost equally prefer public transport and 2 wheeler.

Transport by licence: Employees with license prefer to use their own transport and employees with no license prefer public transport

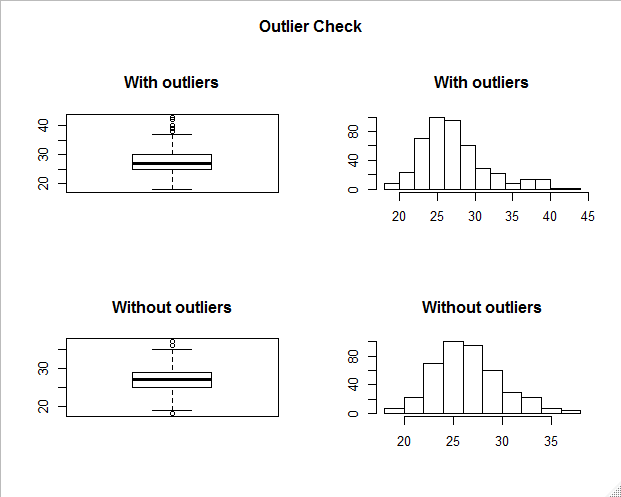
Transport by salary:

* Higher salary seems to be a driving factor for choosing a car as a preferred mode of transport, but there are few outliers for non-car owners whose salary is on the higher side.
* Employees with <10 Lakhs salary use 2 wheeler
* Employees with around 15 Lakhs salary use public transport
* Employees with salary around 30-40 Lakhs use car

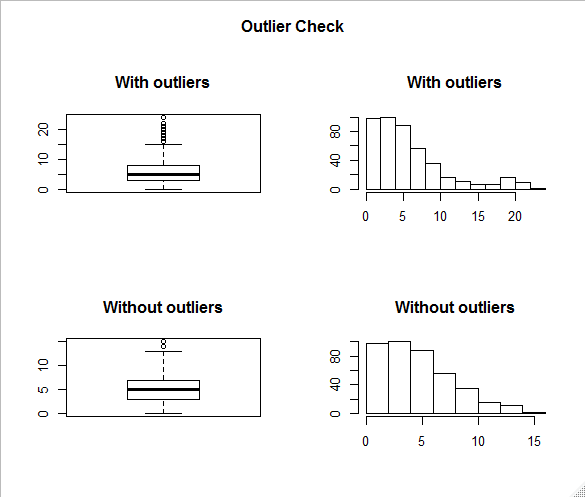
### Missing value check:

### Outlier Analysis:

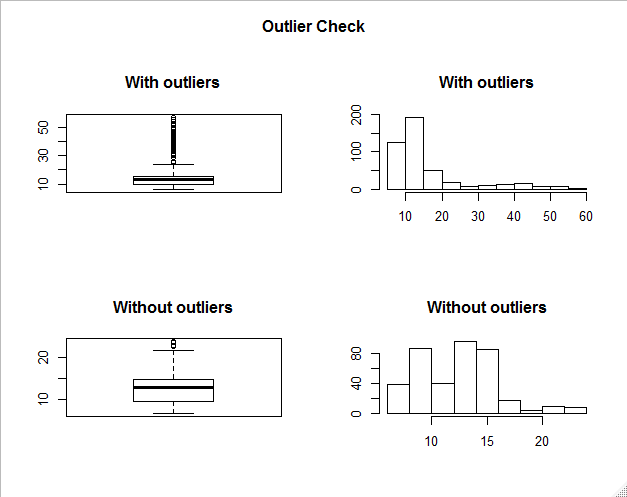
1. Age:
2. Outliers: Age > 37
3. Outlier values: 39 39 39 38 40 38 38 38 38 40 40 39 40 38 39 38 40 39 38 42 40 43 40 38 39
4. Outliers identified: 25
5. nProportion (%) of outliers: 6
6. nMean of the outliers: 39.2
7. nMean without removing outliers: 27.75
8. nMean if we remove outliers: 27.06
9. Action on outliers: Retained the outliers



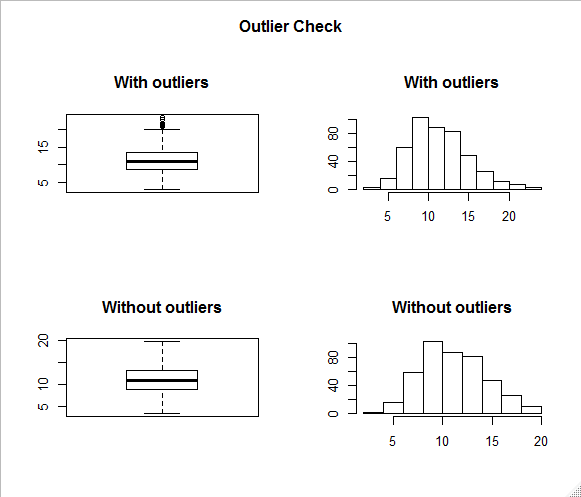
1. Work experience
2. Outliers: > 15
3. 19 16 21 17 16 18 19 18 21 16 19 19 18 19 20 22 16 20 18 21 20 20 16 17 21 18 20 21 19 22 22 19 24 20 19 19 19 21
4. Outliers identified: 38
5. nProportion (%) of outliers: 9.4
6. nMean of the outliers: 19.21
7. nMean without removing outliers: 6.3
8. nMean if we remove outliers: 5.09
9. Action on outliers: Retained the outliers



1. Salary
2. Outliers >24
3. 36.6 38.9 25.9 34.8 28.8 39.9 39.0 28.7 36.9 28.7 34.9 47.0 28.8 36.9 54.0 29.9 34.9 36.0 44.0 37.0 24.9 43.0 37.0 54.0 44.0 34.0 48.0 42.0 51.0 45.0 34.0 28.8 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
4. Outliers identified: 59
5. nProportion (%) of outliers: 15.3
6. nMean of the outliers: 40.41
7. nMean without removing outliers: 16.24
8. nMean if we remove outliers: 12.54
9. Action on outliers: Removed the outliers.



1. Distance
2. Outliers >20
3. 20.7 20.8 21.0 21.3 21.4 21.5 21.5 22.8 23.4
4. Outliers identified: 9
5. nProportion (%) of outliers: 2.1
6. nMean of the outliers: 21.6
7. nMean without removing outliers: 11.32
8. nMean if we remove outliers: 11.11
9. Action on outliers: Removed the outliers.

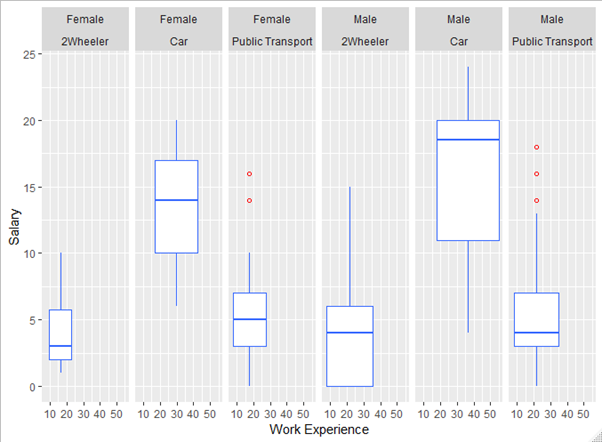


### **Multivariate Analysis:**

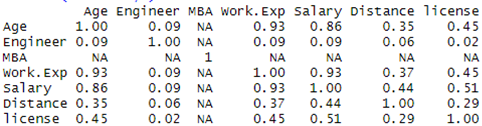
1. Below graph specifies that following parameters impact the

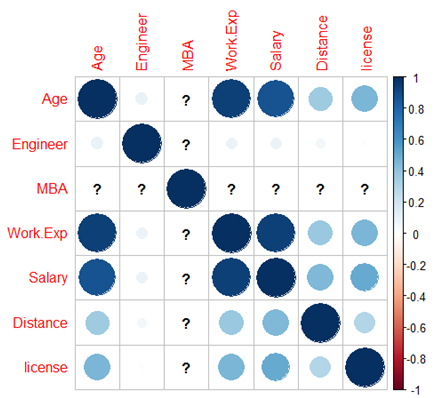
a. Salary

b. Work experience



### **Correlation Matrix:**





**Age vs Transport**

: As was the case with salary, we could see clear demarcation in

usage of transport. With lower age group 2-wheeler is preferable and with higher

work exp car is preferred.

**Distance vs Transport:**

As distance increase employee, would prefer car for comfort

and ease. There is a slight pattern that could be observed here. For greater distance

car is preferred for lower by 2-wheeler and then public transport.

**Gender vs Transport:**

We could see that around 40 % of females use private

transport and 10% use car compared to males where 15% prefers car and total of

30% uses private transport. Thus, even though percentage of car usage is high, but

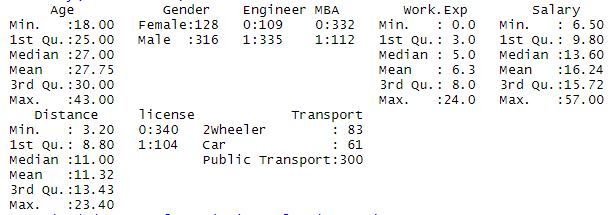
they are also high on public transport

### Data preparation:

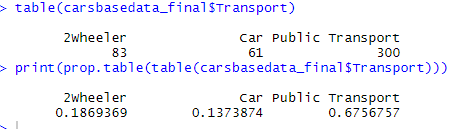
Following columns are converted to factors:

1. Engineer
2. MBA
3. License

### KNN imputation:



### **Data split:**

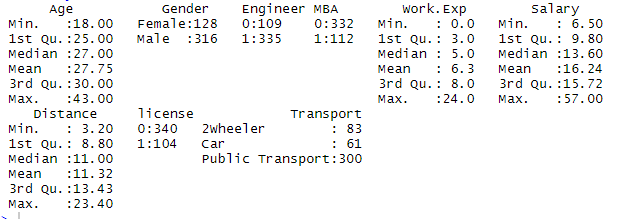


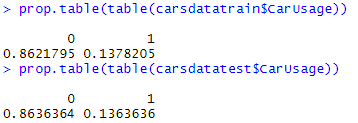
Car Vs other transport:

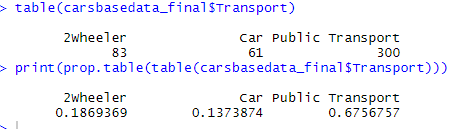


0.1373874

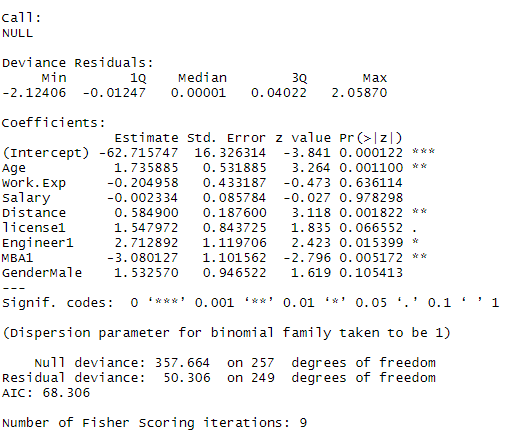
### Car usage:

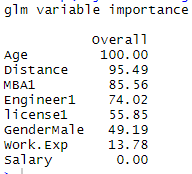


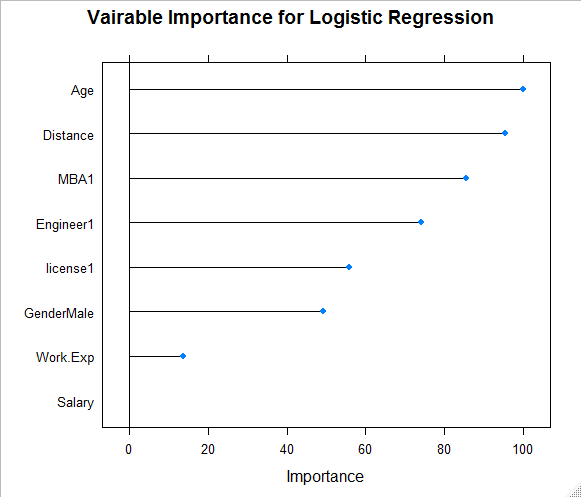


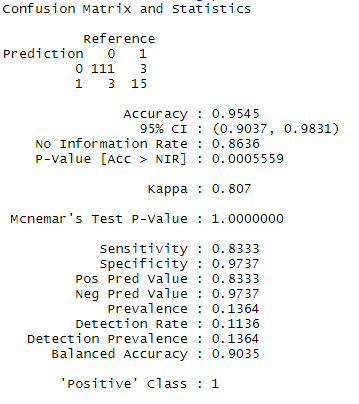


### Logistic Regression:









### Model Interpretation:

From the model it is seen that Age and Distance are more significant. When we look at the odds and probabilities table, we can see that Increase in age by 1 year implies that there is a 98% probability that the employee will use a car. Hence, if the employee has a license, then it implies a 99% probability that he/she will use a car. One lakh increase in salary increases the probability of car usage by 72% the null deviance of this model is 357.664 and the residual deviance is 17.959. This yields a McFadden R Square almost 0.95 yielding a very good fit. We get to see Accuracy and Kappa values are high and we shall do the prediction based on this model.

### **Prediction using Logit:**



### 

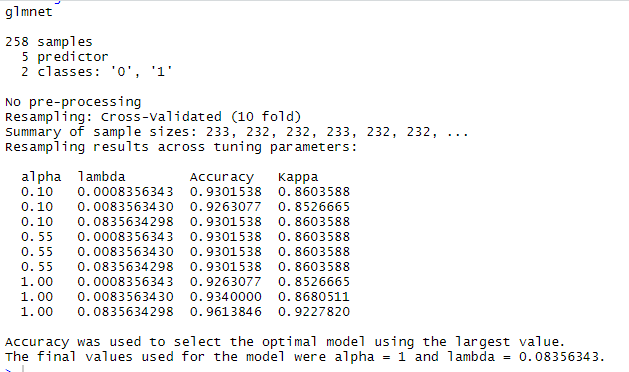
### **Logit Summary:**

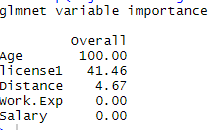
### Accuracy: 90.35%

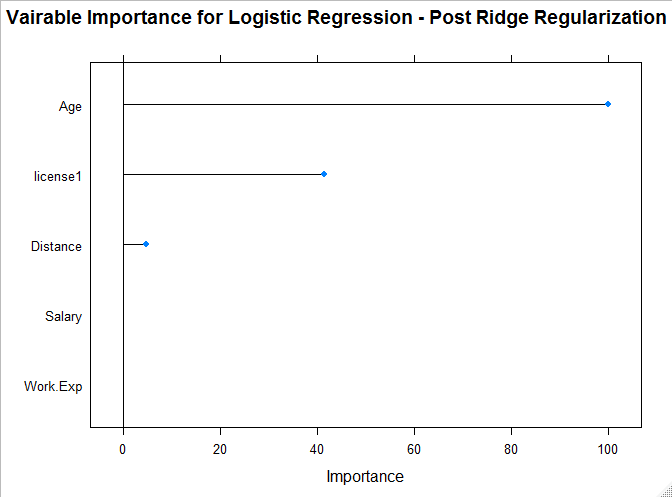
### For both the sample it predicts as 0 for car usage.

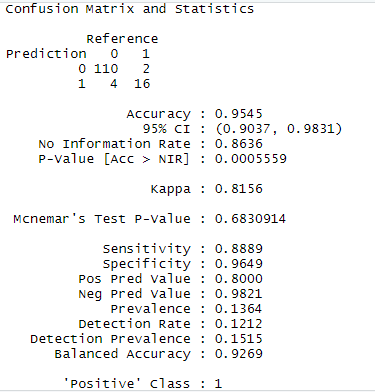
### 

### Logistic regression – Post ridge regression:

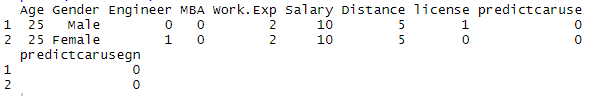








### **Logit – Post ridge Summary:**



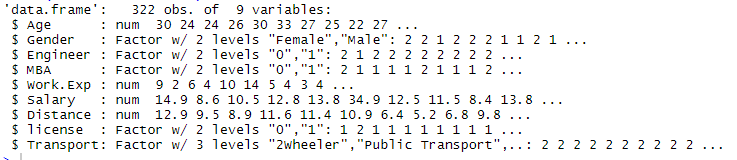
### 

### **Linear discriminant model:**



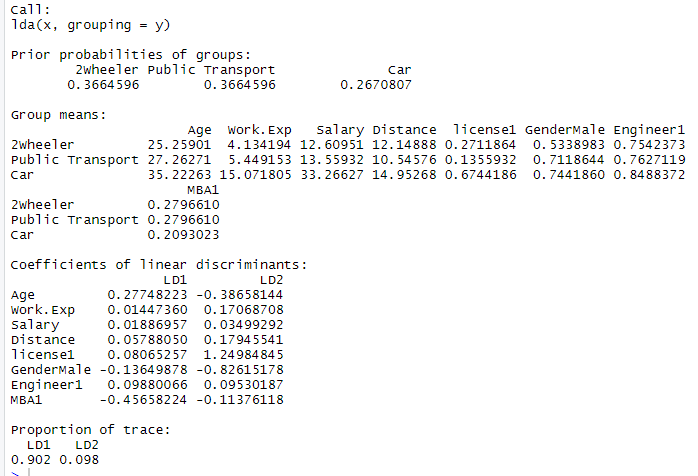


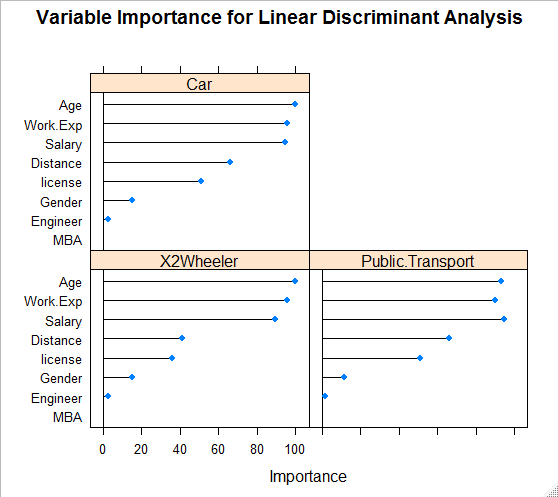




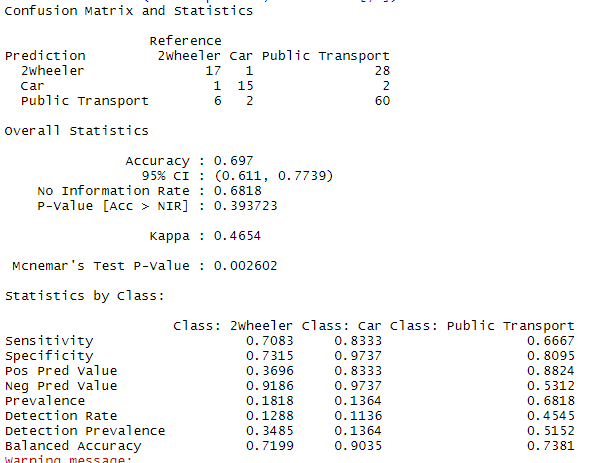


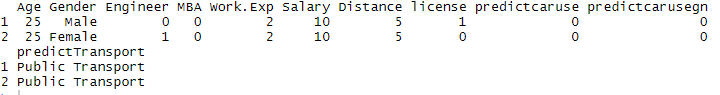
### Linear discriminant analysis:

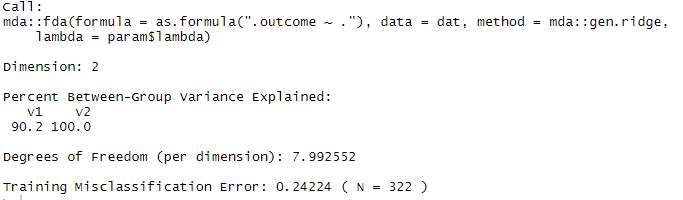




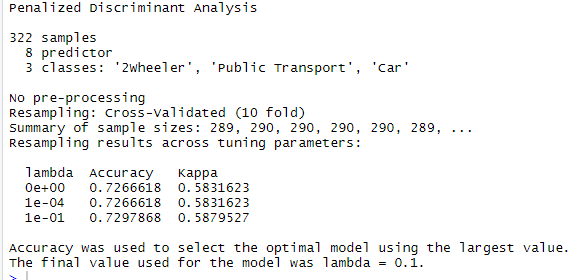
We can see that the Discriminant function achieves 90% separation. It means Age, Work Experience and Salary play key roles in the preference of transport, followed by Distance and License.

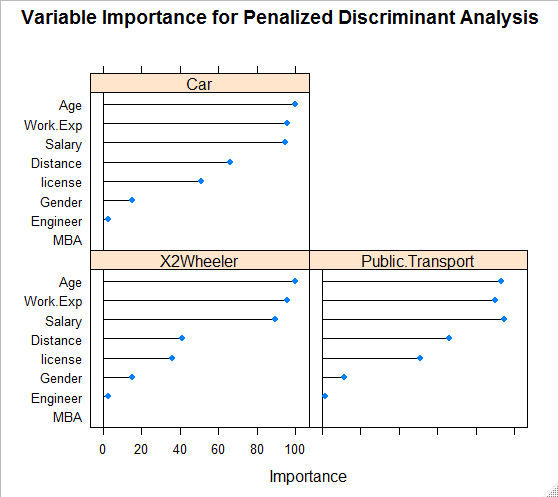


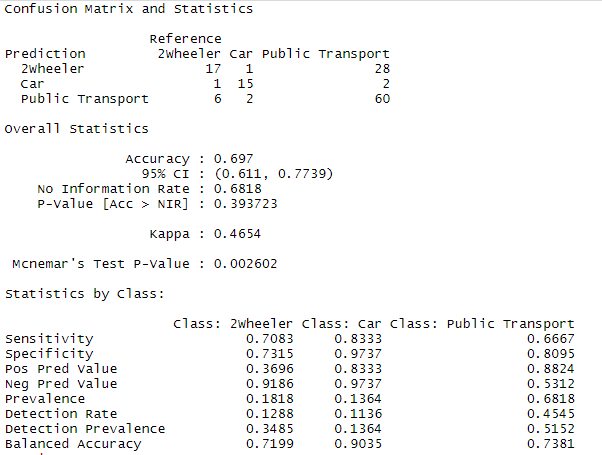




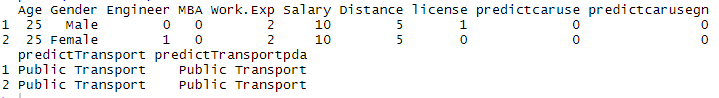
### Improve LDA Model by Regularization







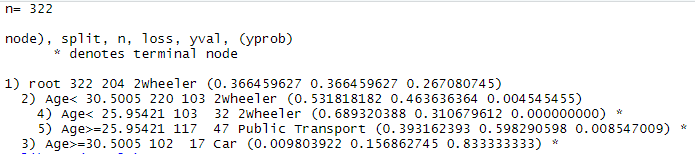
### **LDA summary:**

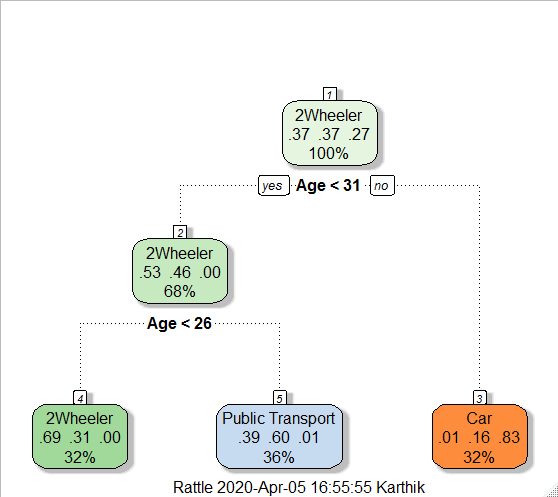


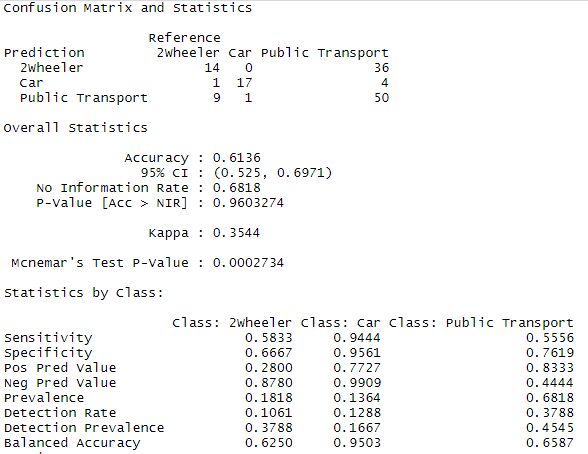
LDA Conclusion:

We can see as per the prediction by LDA, both employees would choose public transport as a means of transportation.

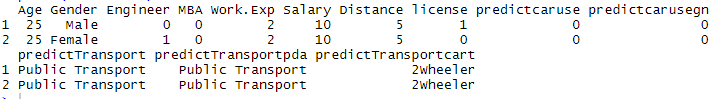
### Prediction using CART







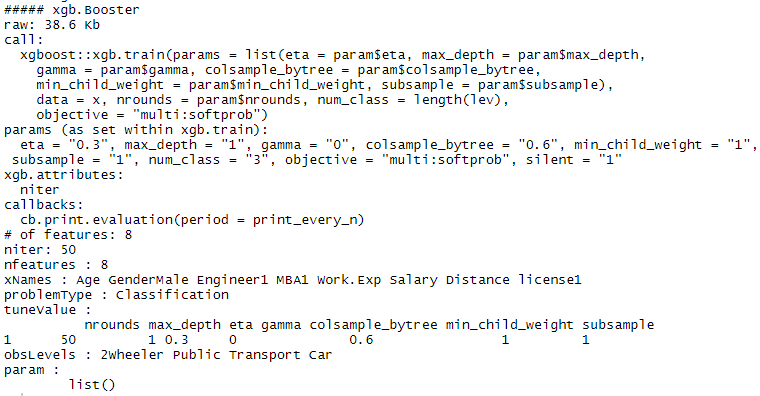
### **CART – Summary:**

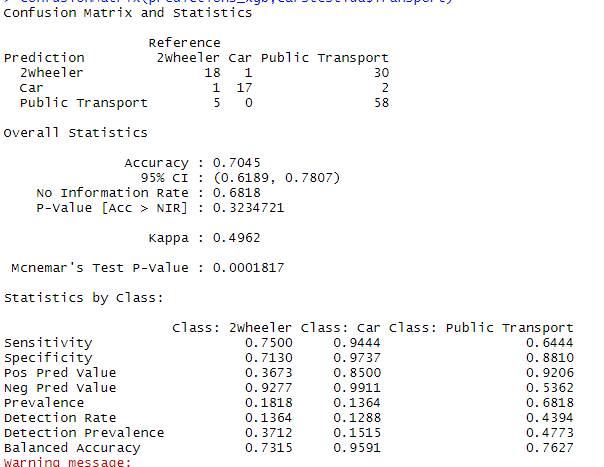


CART Conclusion:

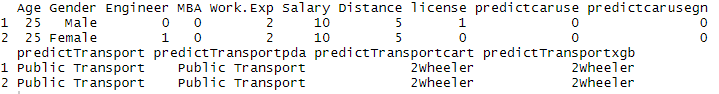
We get an overall accuracy of 61% and CART model also predicts Public transport as the means of transport for the both employees

### Boosting:



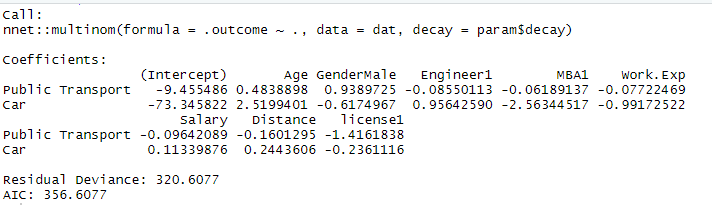


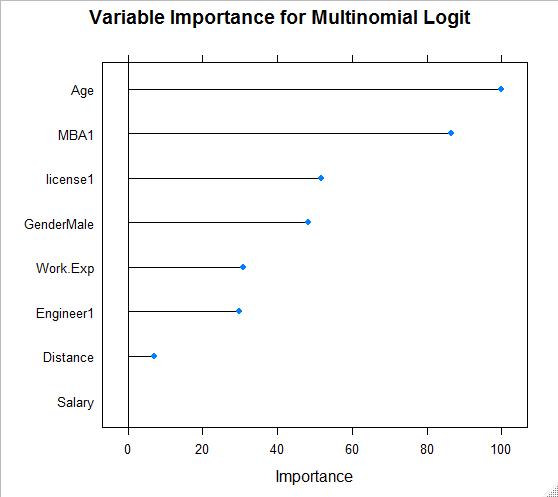
### **Boosting-summary:**

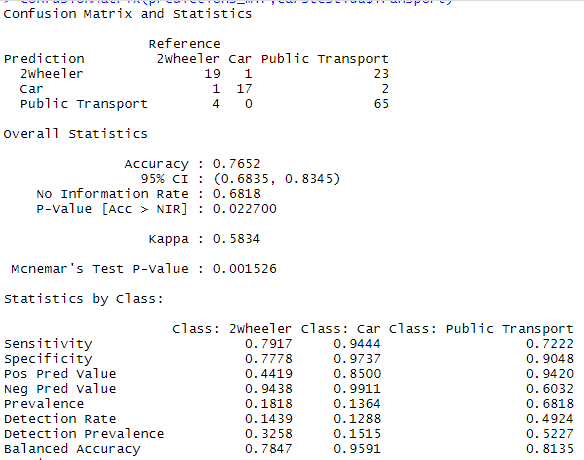


### **Multinomial Logistic regression:**

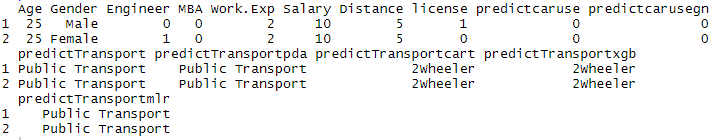
### 



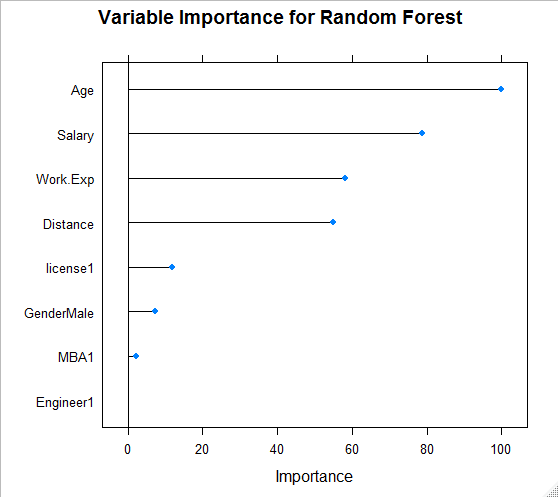


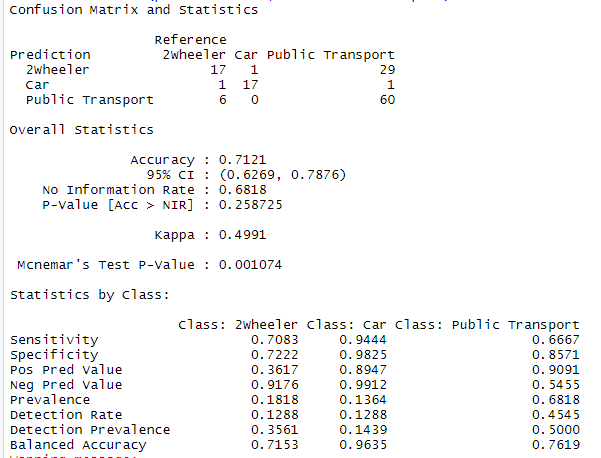


### **Multinomial Logit– Summary:**



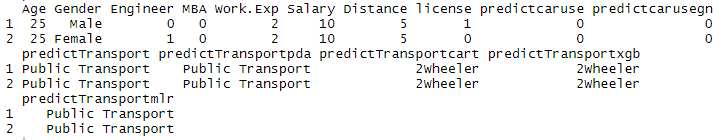
### **Random forest:**



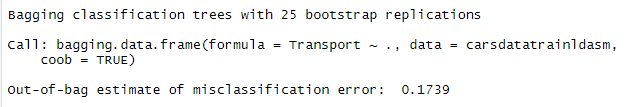


### 

### Random forest –Summary:



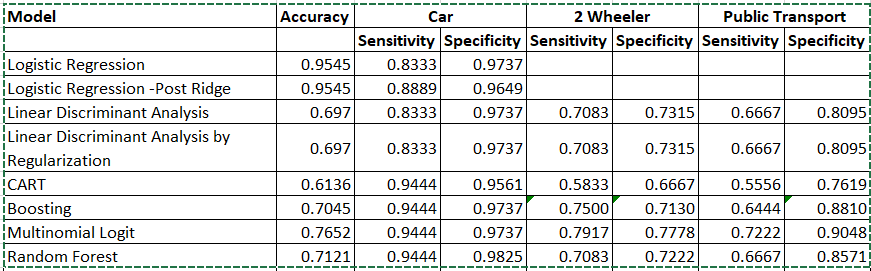
### Bagging:





### **Model comparison:**

### 



**Inference:** We have the highest accuracy with the Logistic Regression Model, but the increase of sensitivity (ability to predict true positives) in the other model values brings down their accuracy values. The specificity (ability to predict true negatives) of the models remains the same across multiple models and the varies little amongst the remaining.

Let us summarize the conclusions from analysis and models for employee’s decision whether to use car Or not:

* Important variables are Age, Work.Exp, Distance and License
* Age and Work.Exp are correlated hence we could use any one (prefer Work.Exp) here
* Hence employees with work exp of 10 and above are likely to use car
* Employees who must commute for distance greater than 12 are more likely to prefer car
* With license, we do see that 74% who commute through car have license and 89% who commute through bus don’t have license. But surprisingly 72% without a license use a 2-wheeler.
* Again, people with higher salaries (>20) are likely to use cars

### References:

<https://www.r-bloggers.com/identify-describe-plot-and-remove-the-outliers-from-the-dataset/>