Cars case-study solution in Predictive Modelling

Created by - Snehal Gawand

May 2020

Objectives

- Business problem:

To understand and identify preferred mode of transport for employees to commute to their office. As well as understand the factors usage of cars as mode of transport.

- Modelling objectives:

Create multiple models and exploring the performance of each model using model performance metrics. Application of bagging and boosting modeling procedures to identify the accuracy with best model.

Models would be built using

- KNN
- Naïve Bayes
- Logistic regression

Exploratory Data Analysis

- Dataset provided has the following characteristics:
  - No. of records : 418. → Dataset has very few records to build the model
  - Most preferred means of transport by the employees is Public transport, followed by 2 Wheeler and then Car.

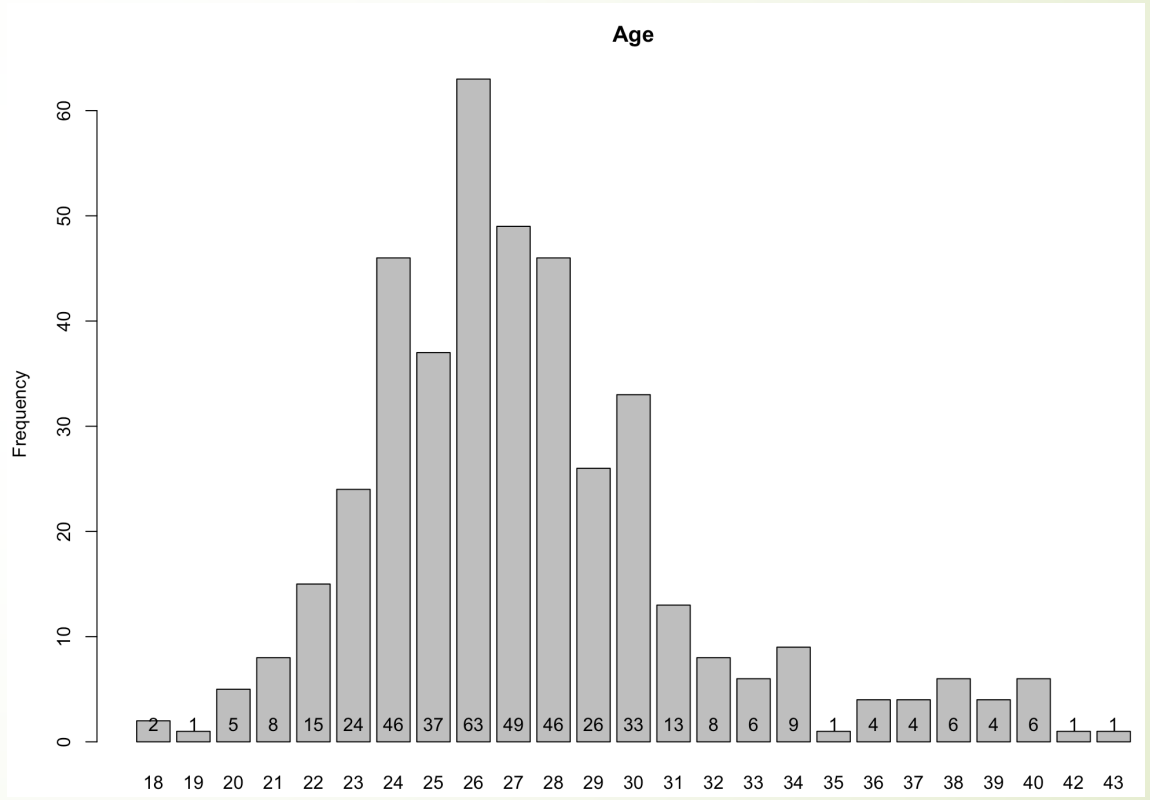
2Wheeler	Car	Public Transport
83	35	300

It indicates the usage of car is only 8.3%

  - Out of total 418 employees, 313 are engineers, 109 are MBA and only 85 employees have licence

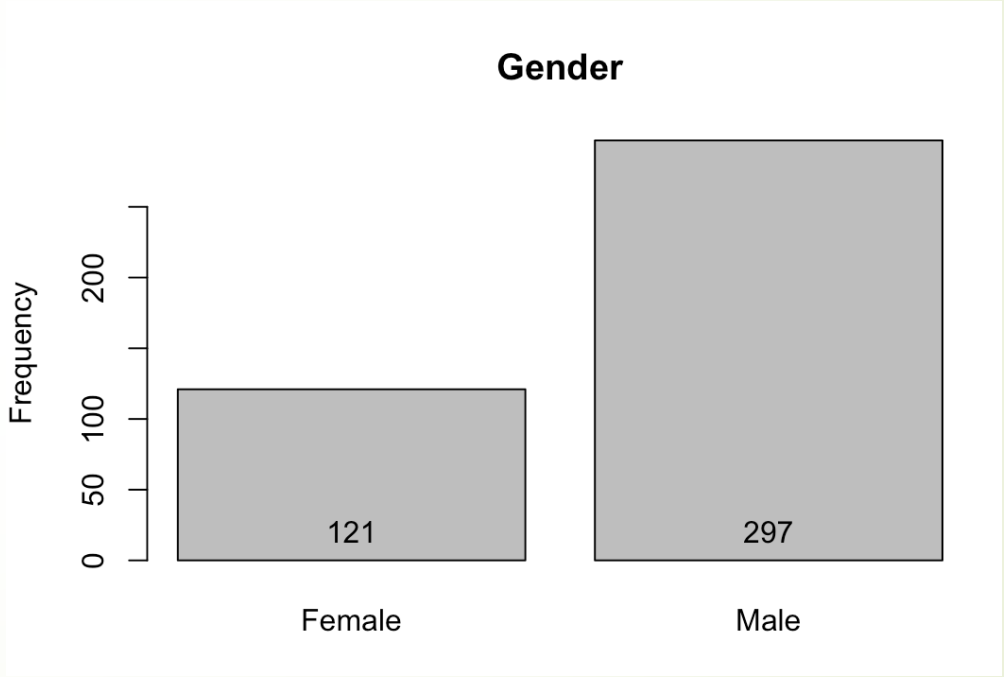
Univariate Data Analysis

- Majority of the employess are within the age group 24 to 31, with maximum if 63 employees with age 26 years



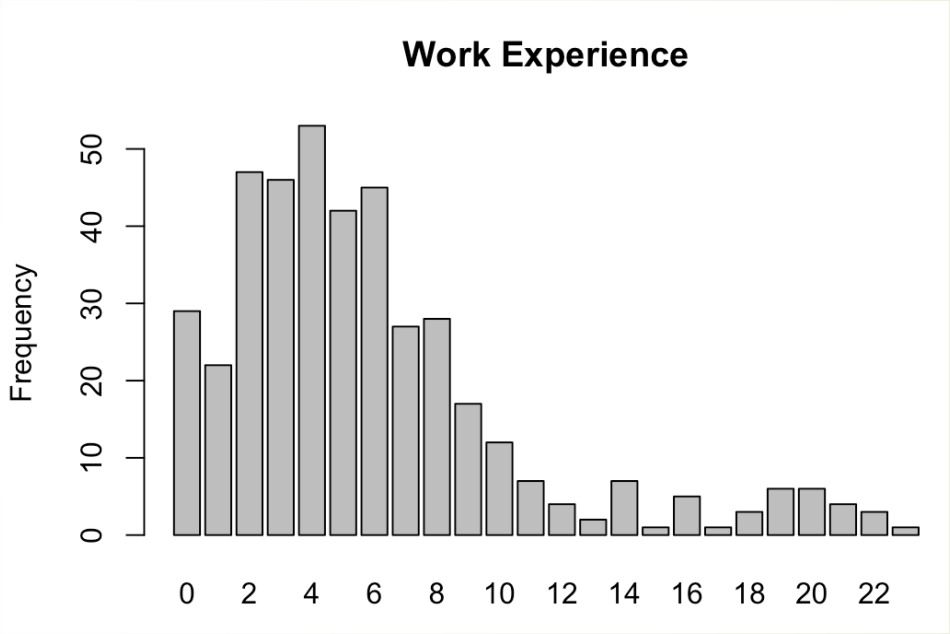
Univariate Data Analysis

- Amongst 418 employees, 121 are Female whereas 297 are Male



Univariate Data Analysis

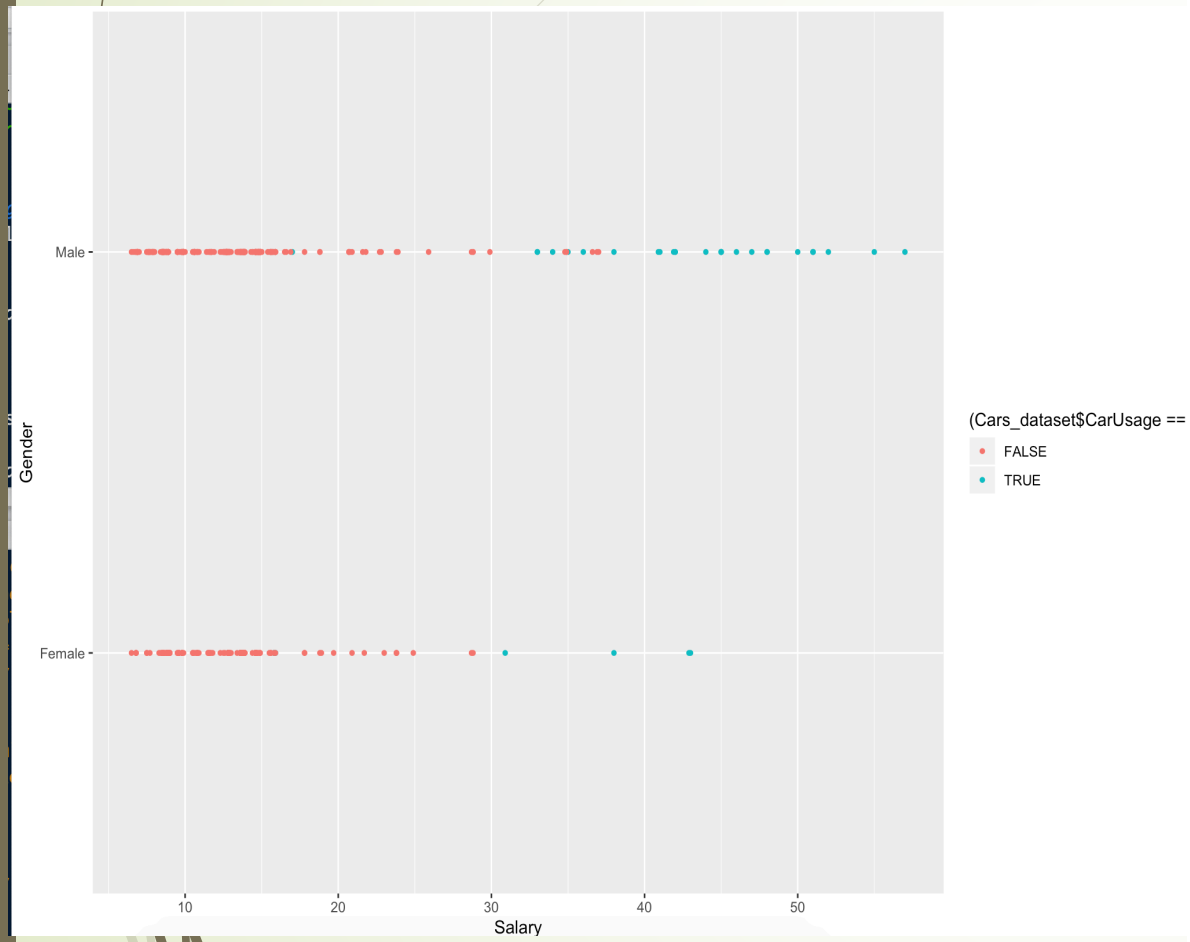
- Most of the employees have work experience between 2 to 8 years.
- Maximum number of employees have 5 years of experience

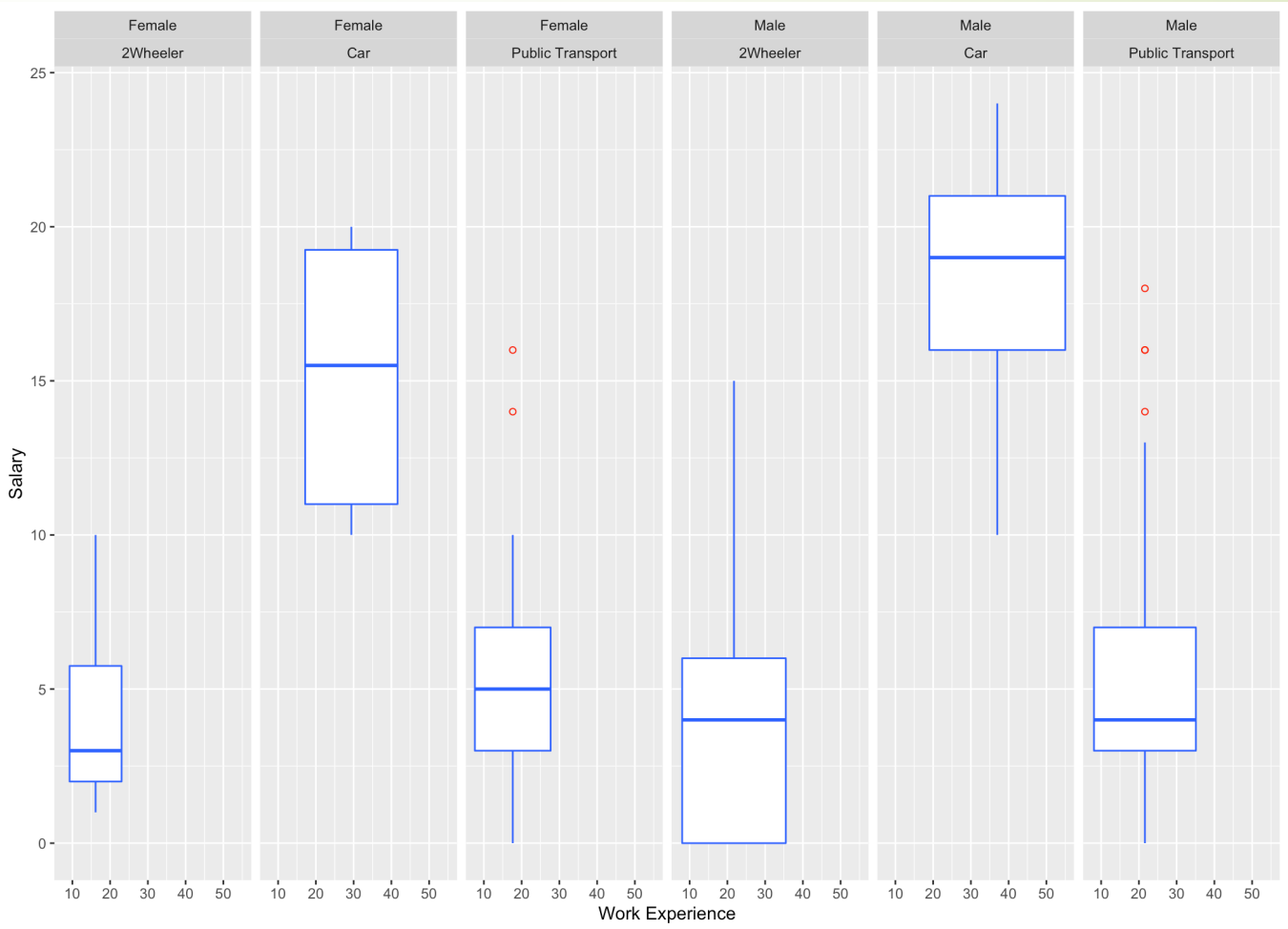


Bivariate Data Analysis

- Female employees with work experience below 20 years have less salary as compared to female employees having expereince in range of 20 to 40 years.
- These employees with higher experience and salary prefer Car as their mode of transport
- Male employees with experience between 30 to 50 years prefer Car as mode of transport and have higher salary
- This clearly indicates the choice of using car as travel mode is more biased towards salary and years of experience

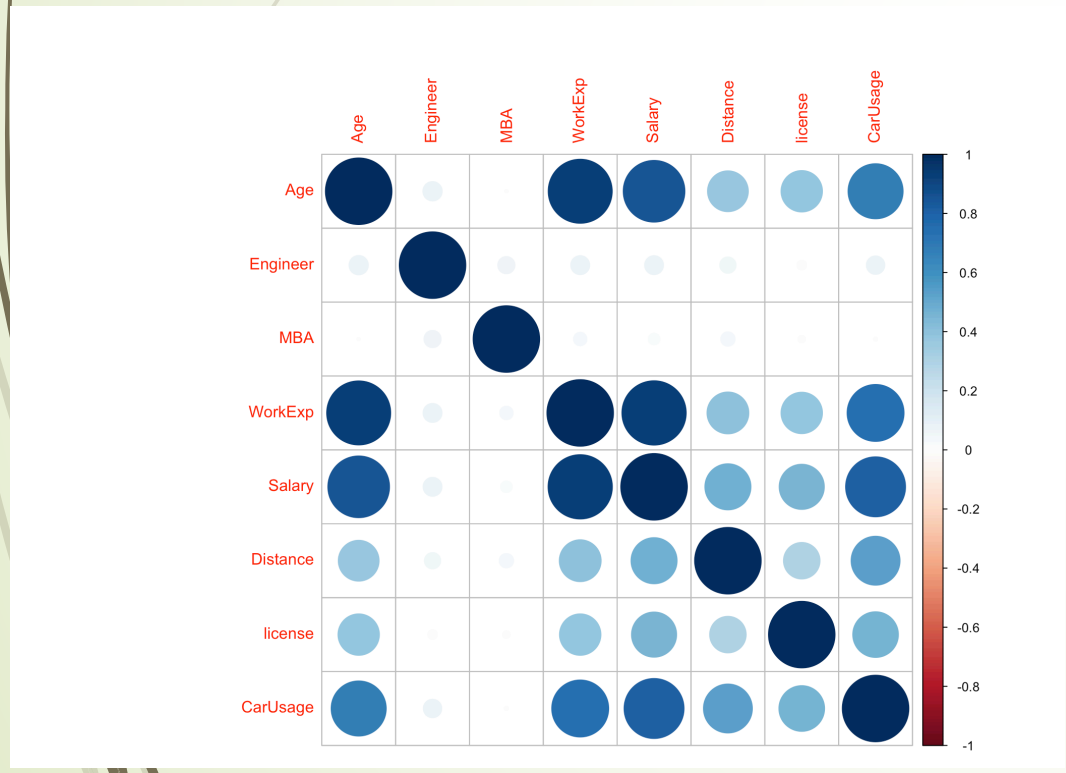
Bivariate Data Analysis

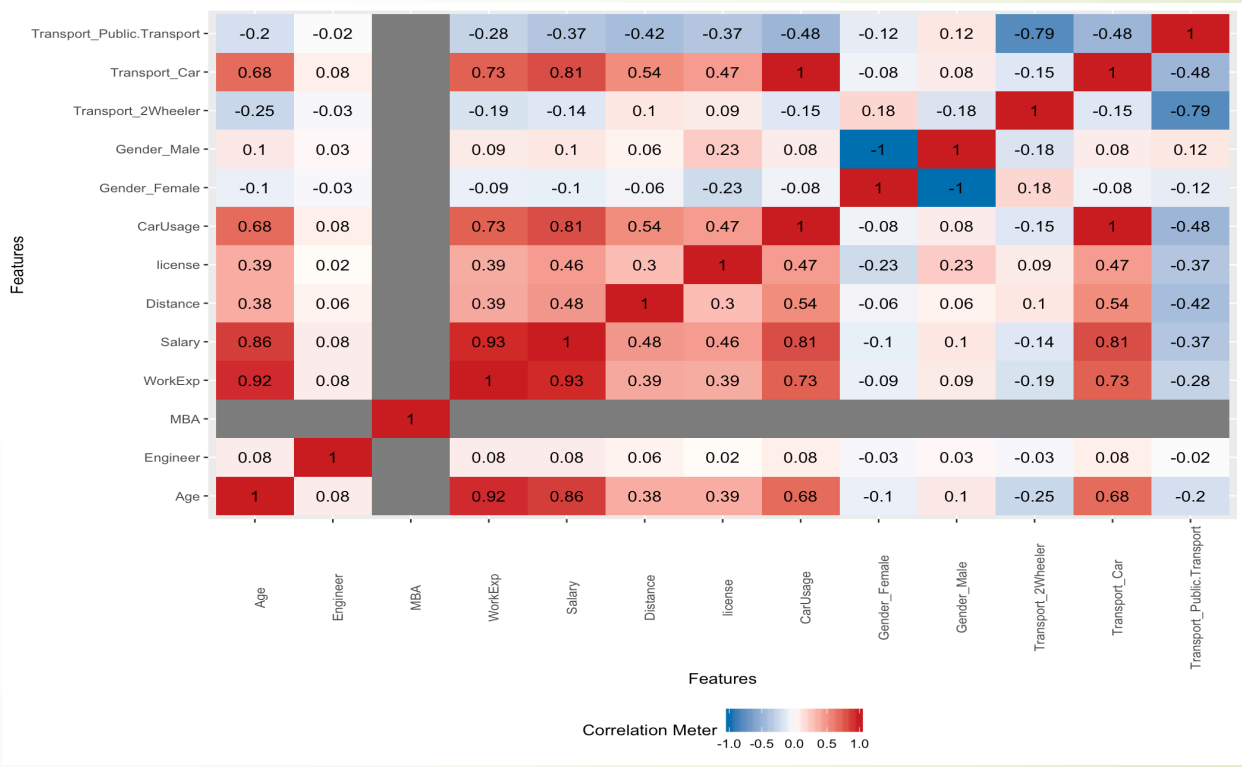




Multivariate Data Analysis

- There is high degree of correlation between employee age, work experience and salary
- Moreover, employees with high salary tend to use car as their mode of transport.





EDA Conclusion

- As the employess with preferance towards using car is low and the car usage data seems to be biased towards Age, Salary and WorkExp we need to use appropriate sampling techniques so as to predict the correct employees which will use car in future.

Data Preparation

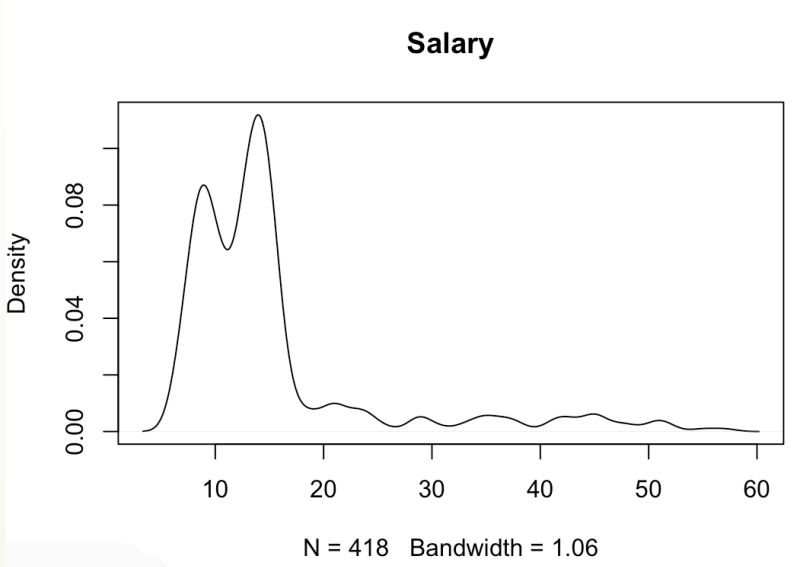
- Missing value treatment
  - The input dataset does not have any missing values.
- Columns such as Gender, Engineer, MBA, license converted to factor for ease of use
- New column CarUsage added for help in predicting employees with car usage while creating models
- Outlier treatment

  Salary has few outliers observations.

  Normality test is performed to check for skewness.

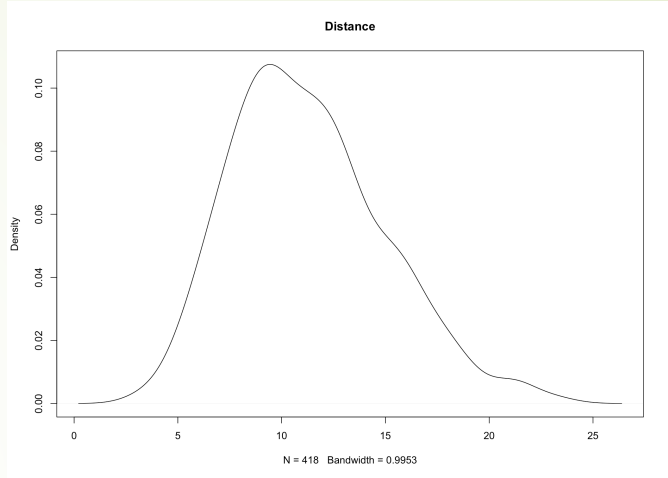
  Graph shows data in Salary column is skewed

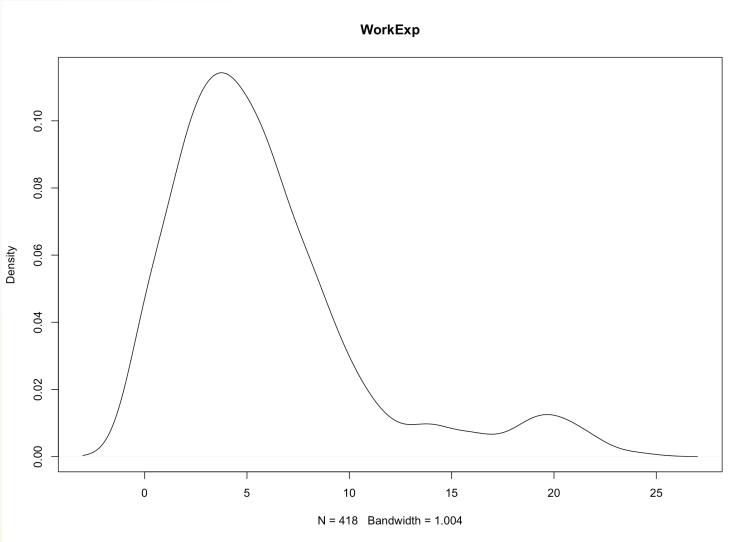
  towards right



Data Preparation

- Distance column does not have any outliers. There is no significant difference in the percentage distribution.

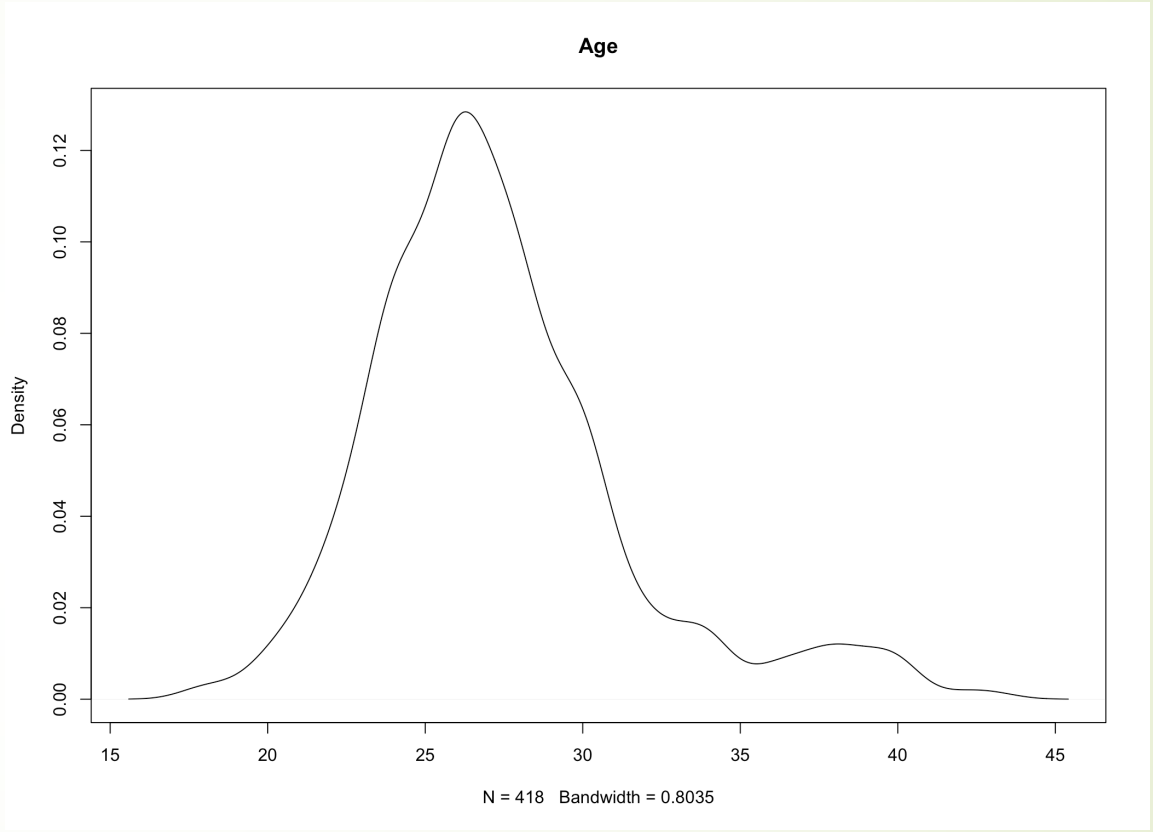




For the next continous variable ‘WorkExp’, there exists few outliers. The normality test via graphical way shows data skewed towards right.

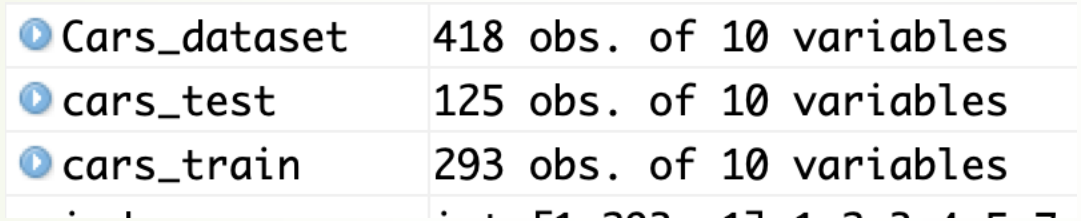
Data Preparation

- The data in Age column seems to be
- normally distributed with no outliers



Data Modeling

- Splitted the input data into train and test datasets in 70:30 ratio. The train and test dataset have same percentage of CarUsage as base data



- Removed unused variables from the newly created train and test datasets
- Using SMOTE equally splitted train dataset in car users and non-car users

0	1
0.5	0.5

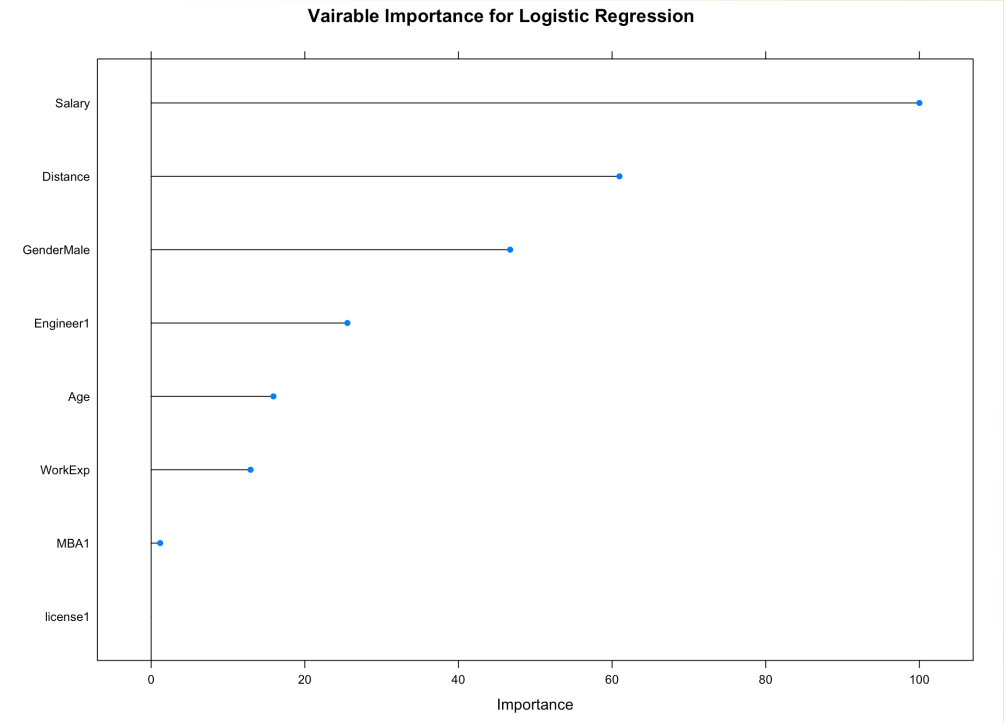
Logistic Regression

- Built logistic regression model on the train dataset.
- For every one unit change in Age, the log odds of car usage vs non car usage changes by -3.022
- For every one unit change in Salary, the log odds of car usage vs non car usage increases by 5.035
- The most important variables

  for car usage are Salary and Gender.

  Male employees are more inclined

  towards usage of car



Logistic Regression

- Prediction on the train dataset using the logisitic regression model

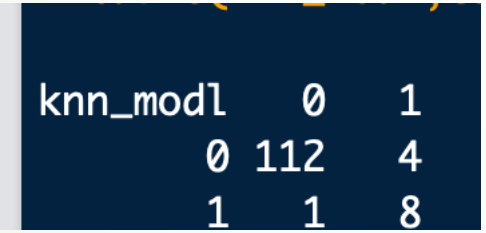
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 260   0
         1  10  23
```

- Accuracy of the model is 96.59%
- Sensitivity is 100%. This means, the model is good enough to accurately predict the employees using cars as mode of transport
- Specificity of model is 96.3%. This indicates, the model needs to be improved for idenfitication of employees not using car as mode of transport

- Conclusion : Overall it is a good model

KNN Algorithm

- KNN algorithm is used for matching a point with its closest k neighbours in multi-dimensional space.
- In the dataset, MBA_imp new logical column is created and it has one value set as true. This means one null value has been imputed
- Splitted the imputed dataset into train and test dataset
- We have 293 observations in training dataset, square root of 293 is around 17.11. Hence, we will create model with k value 17
- Accuracy of this model is 96%
- Confusion matrix



- Conclusion : Overall the model is good

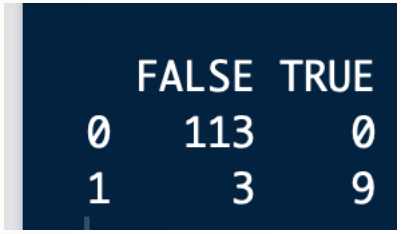
Naïve Bayes

- Naive Bayes model is created with set of independent variables. However we observed that Age, WorkExp and Salary are highly dependent on each other.

  Hence, Naïve Bayes model cannot be created with the given dataset

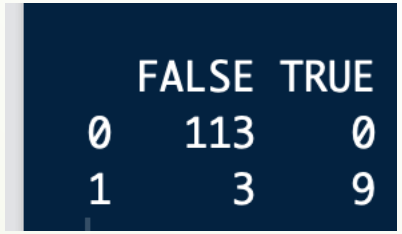
Bagging and Boosting

- The dataset is again splitted into test and train to apply the bagging and boosting techniques
- Confusion matrix generated via Bagging



	FALSE	TRUE
0	113	0
1	3	9

- Confusion matrix generated via Boosting



	FALSE	TRUE
0	113	0
1	3	9

Bagging and Boosting

- XGBoost works with matrices that contain all numeric variables. Applying XGBoost technique generates below confusion matrix

	FALSE	TRUE
0	111	2
1	3	9

Conclusion

- Employee data seems to be biased towards use of car, when there is increase in age, WorkExp and Salary
- Model built via both logistic regression and KNN technique derive approximately similar results. However, as the data provided in the dataset is very less, it would not be wise to make predictions only based on the finite amount of data.
- Bagging and boosting helped improving the model, however generated similar kind of summary in terms of predicting sensitivity and specificity via confusion matrix

THANK YOU