# Project Report : Mode of Transport Prediction

**Objective :**

The objective of the project is to understand what mode of transport employees prefers to commute to their office. The attached data '[Cars.csv](https://olympus.greatlearning.in/courses/7362/files/737590/download?verifier=d51upLRepA9fEuQTEqkZhdlZKgs8Ce1eY3fwjfU2&wrap=1)' includes employee information about their mode of transport as well as their personal and professional details like age, salary, work exp.

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?

**Data Dictionary**

|  |  |
| --- | --- |
| Age | Age of the Employee in Years |
| Gender | Gender of the Employee |
| Engineer | For Engineer =1 , Non Engineer =0 |
| MBA | For MBA =1 , Non MBA =0 |
| Work Exp | Experience in years |
| Salary | Salary in Lakhs per Annum |
| Distance | Distance in Kms from Home to Office |
| license | If Employee has Driving Licence -1, If not, then 0 |
| Transport | Mode of Transport |

**Hypothesis formation**

The data file contains one Dependent (Transport) and 8 Predictor variables.

The assignment aim is to predict if a person will use car or not and also the significant variables behind this decision

**Null Hypothesis (Ho)** – No predictor is able to predict the Mode of transport

**Alternate Hypothesis (Ha)** – At least one of the predictors is able to predict the Mode of transport .

**Data Understanding and transformation**

**Data Overview**

Let us check for Variables names, Five Point Summary, and their data type. This is the first opportunity to explore our data.

In the data set there are 444 rows, each with 9 different variables(columns).

A quick inspection of the data sheet shows that there is only one missing value for one of the rows in variable MBA variable. We see that 75% of the data has MBA = 0 , so we have imputed this missing value as MBA = 0 .

Some of the important insight about the data set :

1: The dependent variable, Mode of Transport, is categorical variables with 3 values – Car, 2 wheeler and Public Transport. Since we are only interested to know if a person will be using Car or not , we will convert this to a 2 level categorical variable – car and NoCar . This will simplify our dataset and also some of the modelling algorithms only accept 2 level dependent variable.

2. The below variables have only 2 levels of values:

* Gender
* MBA
* Engineer
* License

3. There are some continuous variables in the data set too:

* Age
* Salary
* Experience
* Distance

4. Looking at the dataset, we see some of the variables have quite a huge range. We will check for the outliers later in detail.

**Exploratory data analysis**

1. There is no missing value in the data set
2. The initial structure of the data is as below:

$ Age : int 28 23 29 28 27 26 28 26 22 27 ...

$ Gender : Factor w/ 2 levels "Female","Male": 2 1 2 1 2 2 2 1 2 2 ...

$ Engineer : int 0 1 1 1 1 1 1 1 1 1 ...

$ MBA : int 0 0 0 1 0 0 0 0 0 0 ...

$ Work.Exp : int 4 4 7 5 4 4 5 3 1 4 ...

$ Salary : num 14.3 8.3 13.4 13.4 13.4 12.3 14.4 10.5 7.5 13.5 ...

$ Distance : num 3.2 3.3 4.1 4.5 4.6 4.8 5.1 5.1 5.1 5.2 ...

$ license : int 0 0 0 0 0 1 0 0 0 0 ...

$ Transport: Factor w/ 3 levels "2Wheeler","Car",..: 3 3 3 3 3 3 1 3 3 3

1. The distribution of dependent variable – Mode of Transport

2Wheeler car Public Transport

83 61 300

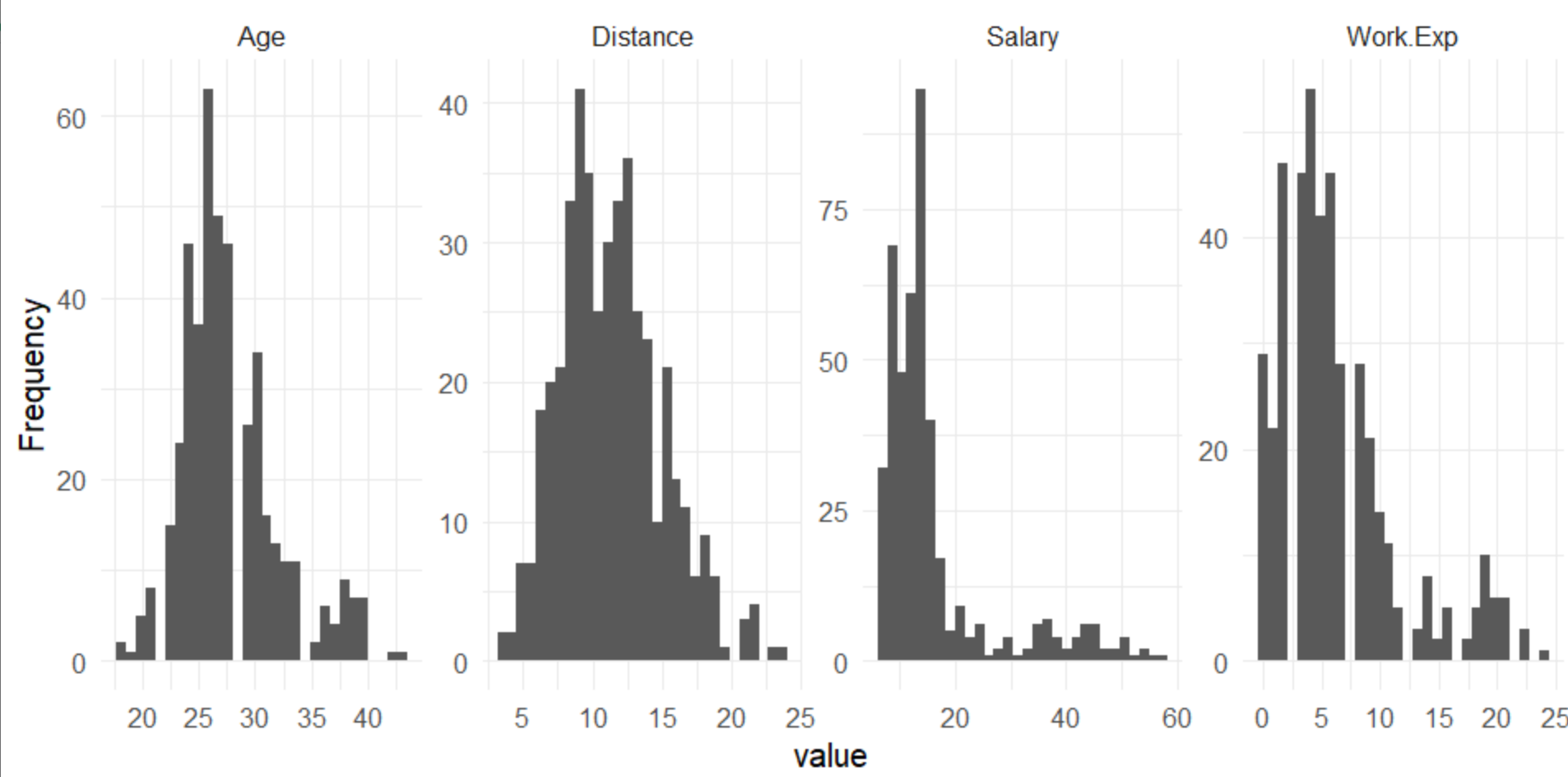
So only 61/444 = 13:7% of the dataset has positive value. We can say our data set in **imbalanced**.

**The 5 point analysis of the data :**

|  |
| --- |
| Age Gender Engineer MBA Work.Exp Salary |
| Min. :18.00 Female:128 Min. :0.0000 Min. :0.0000 Min. : 0.0 Min. : 6.50 |
| 1st Qu.:25.00 Male :316 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.: 3.0 1st Qu.: 9.80 |
| Median :27.00 Median :1.0000 Median :0.0000 Median : 5.0 Median :13.60 |
| Mean :27.75 Mean :0.7545 Mean :0.2523 Mean : 6.3 Mean :16.24 |
| 3rd Qu.:30.00 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.: 8.0 3rd Qu.:15.72 |
| Max. :43.00 Max. :1.0000 Max. :1.0000 Max. :24.0 Max. :57.00 |
| Distance license Transport |
| Min. : 3.20 Min. :0.0000 2Wheeler : 83 |
| 1st Qu.: 8.80 1st Qu.:0.0000 Car : 61 |
| Median :11.00 Median :0.0000 Public Transport:300 |
| Mean :11.32 Mean :0.2342 |
| 3rd Qu.:13.43 3rd Qu.:0.0000 |
| Max. :23.40 Max. :1.0000 |

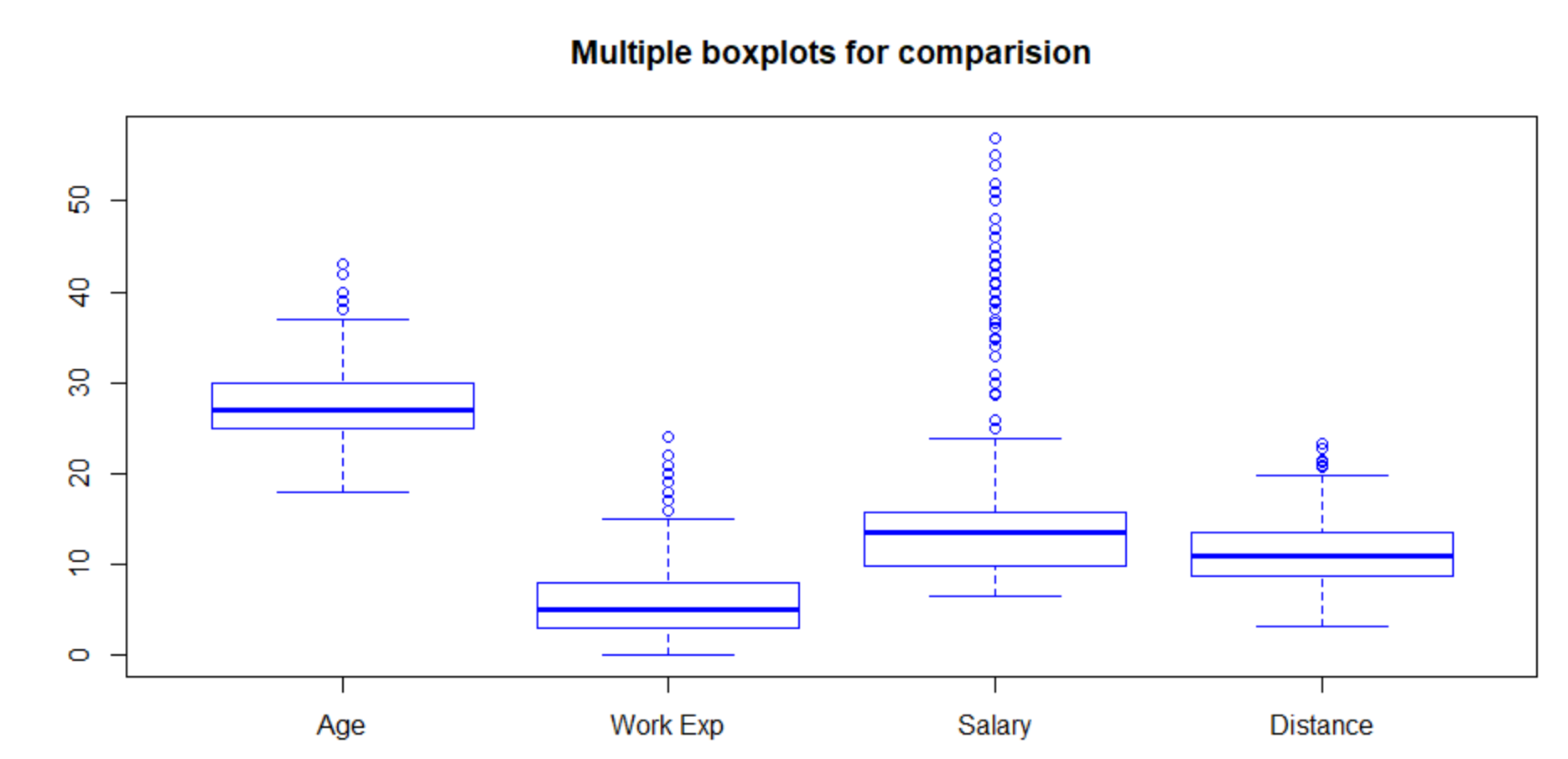
**Univariate Analysis**

Let us take a look at the continuous variables and how they are distributed.

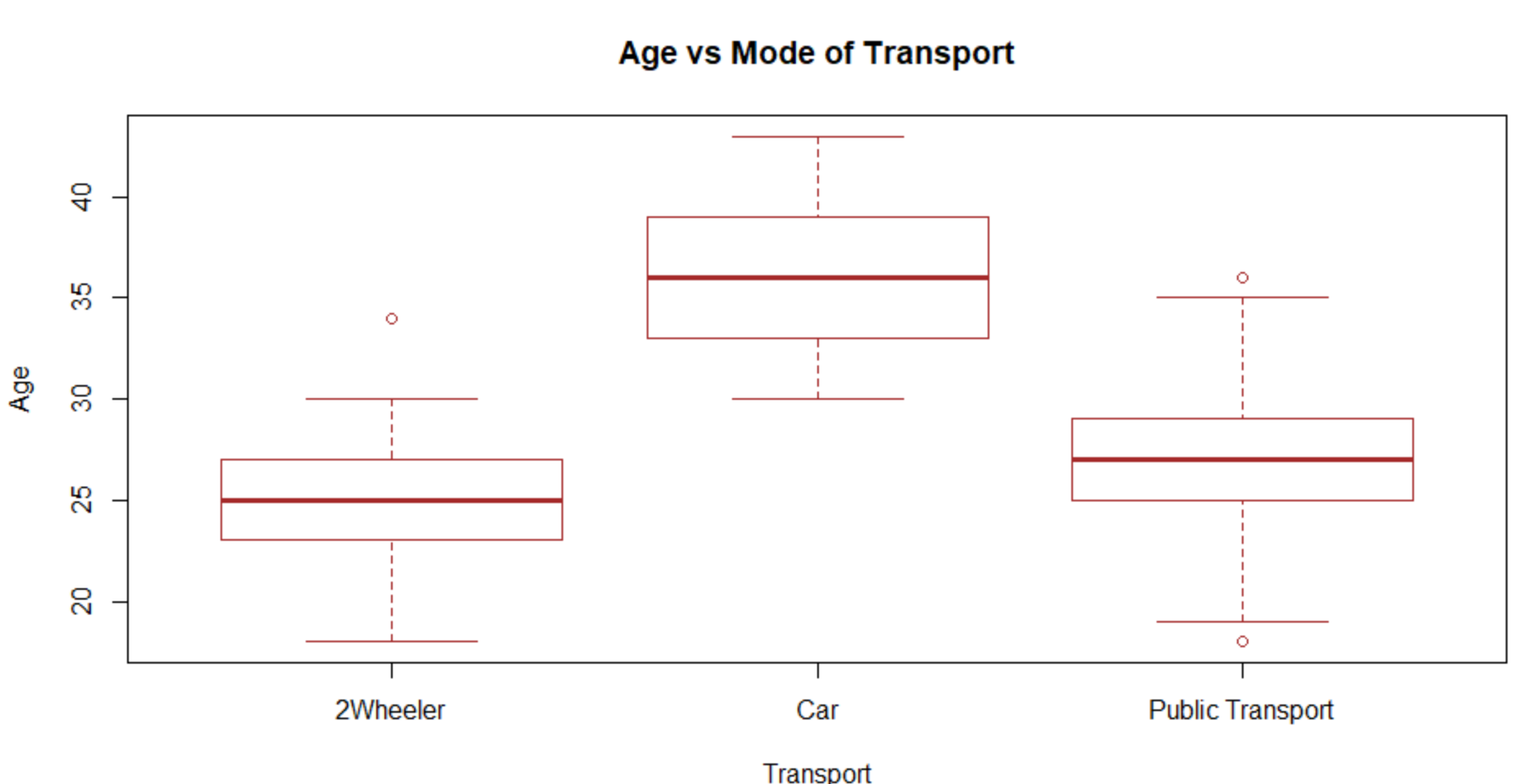


We see here age and distance are normally distributed, however both Salary and Work Experience variables are right skewed.

Let us take a look at the boxplot for these variables:

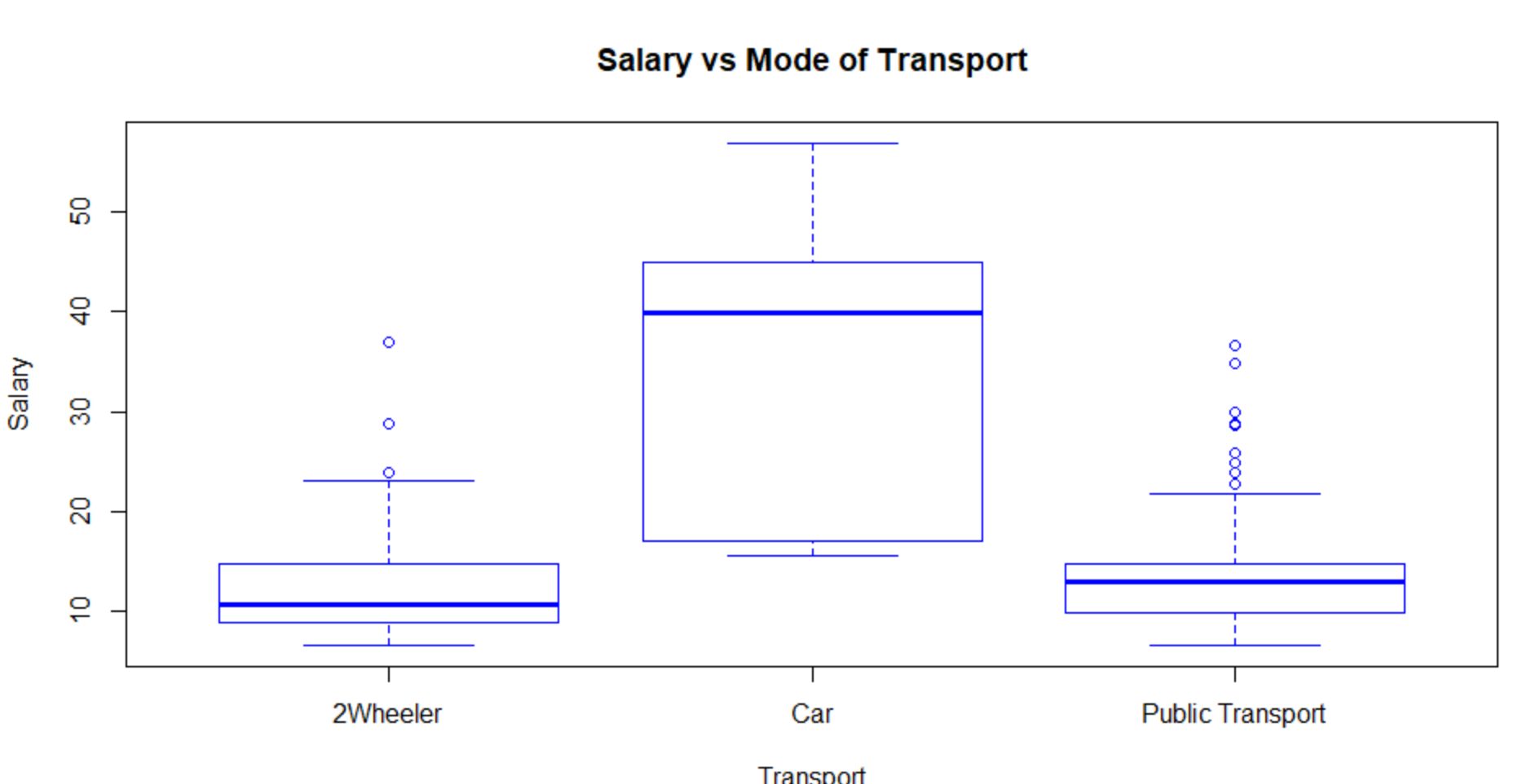


We see there are outliers in each variable but for Salary and Work experience there are many outliers.



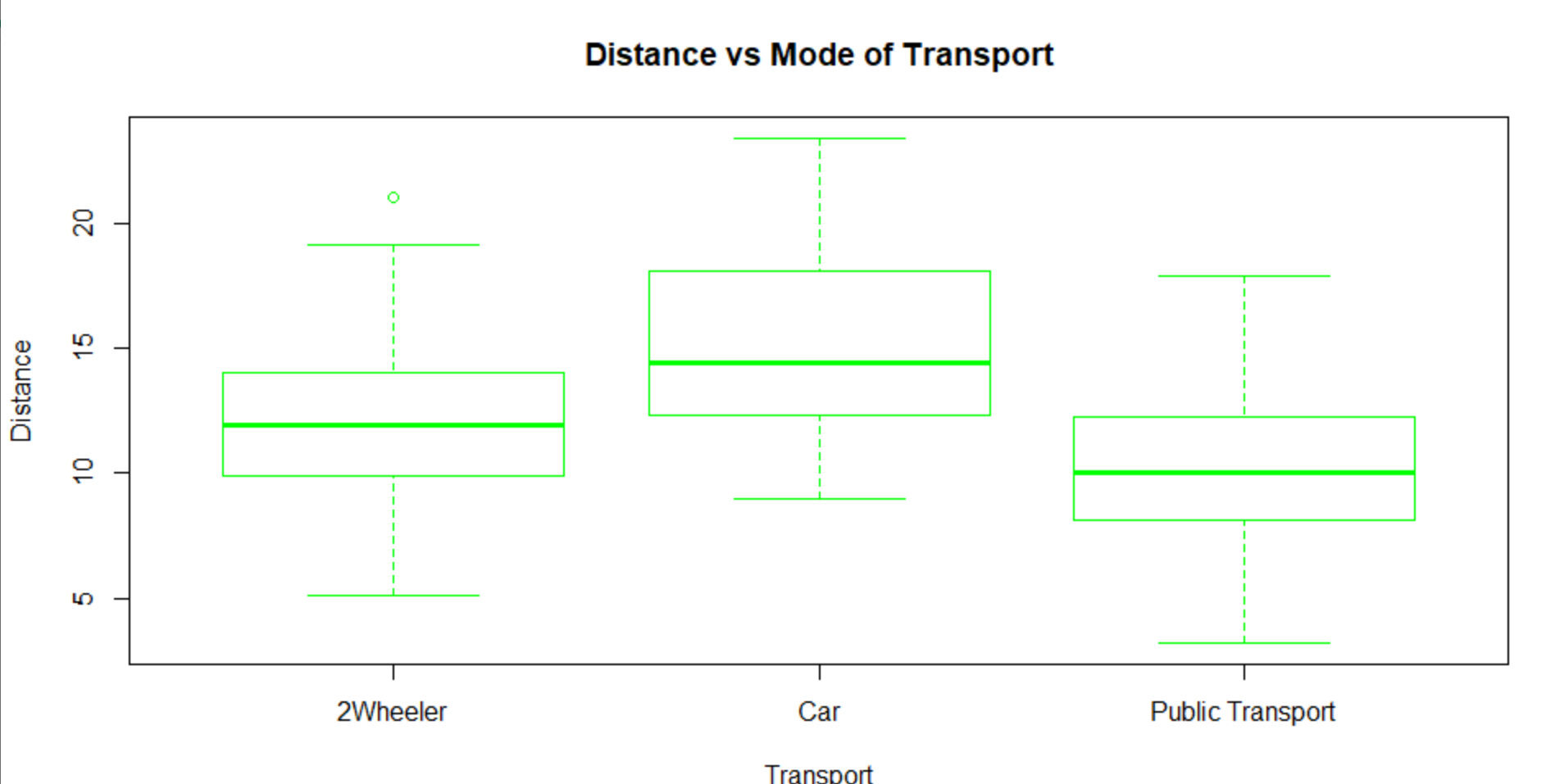
It is quite evident that:

* Among them who have their own vehicle
  + people with age group 30-40 are only using car with majority of them around age 35
  + People with age group below 30 are mainly using 2 wheeler
* People who are using public transport have a wider age range but mostly they are below 35.



Now this is quite interesting. We have already seen lot of outliers in boxplot for Salary . But when we do this bivariate analysis for Salary vs Mode of transport, we find that all the high range salaried people are within the IQR in terms of using car. This tells us that most people with high salary are using car.

We validate this from the data sheet and we find there are 52 people with more than 30 lac annual salary and 46 of them use car .We see similar observation for age and work experience variables also

.

**So we conclude that there is no need of outlier treatment for any of the variables and they are valid values in the data set .**

Let us now see how the other variables like Gender, MBA, Engineer are related to Mode of transport

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Transport** | | |  |
| Gender | **2 Wheeler** | **Car** | **Public Transport** | **Total** |
| Female | 38 | 13 | 77 | 128 |
| Male | 45 | 48 | 223 | 316 |
|  |  |  |  |  |
| Non MBA | 66 | 49 | 217 | 332 |
| MBA | 17 | 12 | 83 | 112 |
|  |  |  |  |  |
| Non Engineer | 23 | 9 | 77 | 109 |
| Engineer | 60 | 52 | 223 | 335 |

The above summary gives a easy insight to the data set. Our population is

* 71% male , 29% female
* 25% MBA , 75% non MBA
* 75% Engineer , 25% Non Engineer

Car usage is more for engineers and males.

**Check for Multicollinearity - Plot the graph based on Multicollinearity & treat it.**

It is very evident and logical that age and experience and salary are highly corelated and proportional. Let us draw the correlation plot.

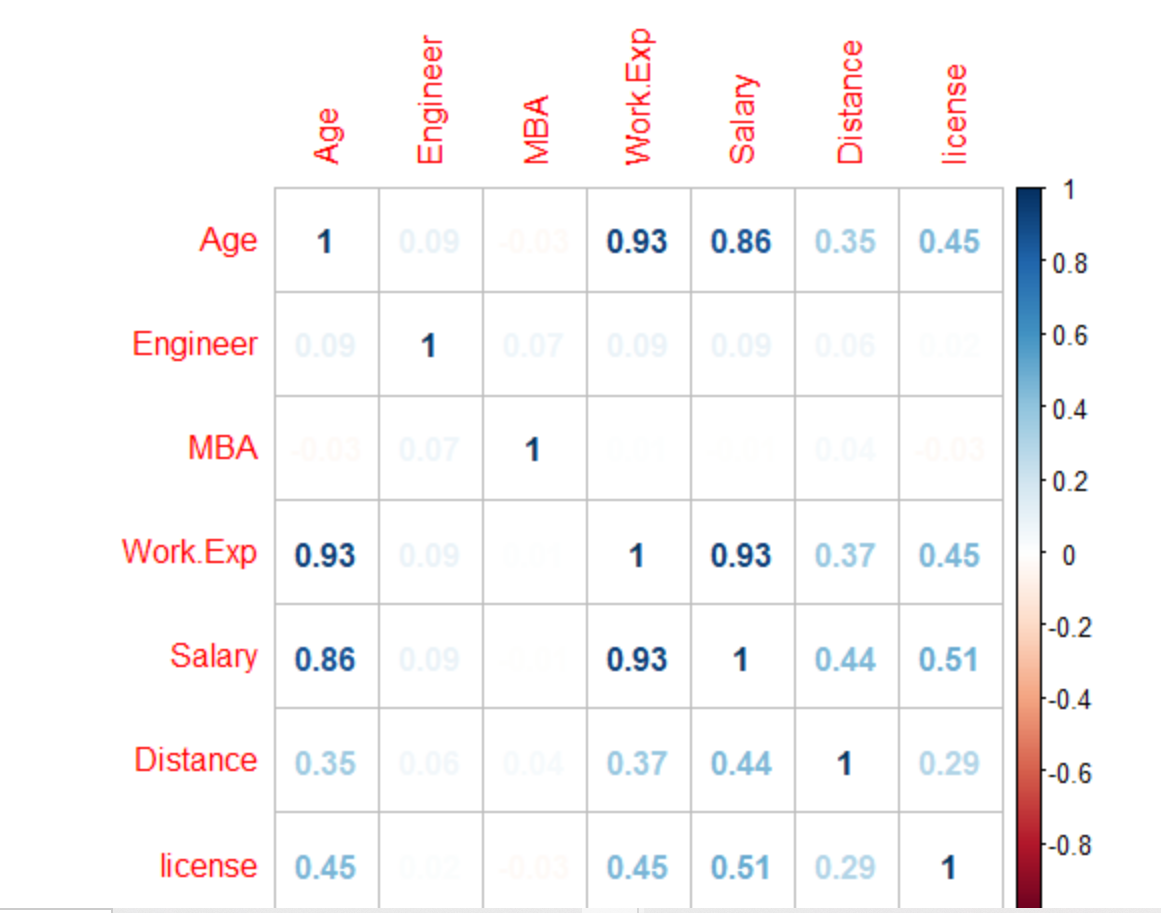
So from the below plot we see that there are 3 pairs of variables which are highly (positive) corelated :

* Age , Salary
* Age, Work Experience
* Work Experience, Salary

For examining the patterns of multicollinearity, it is required to conduct t-test for correlation coefficient.

ppcor package helps to compute the partial correlation coefficients along with the t-statistics and corresponding p values for the independent variables.

As expected the correlation between age, salary, work exp is highly significant;



We will build a simple linear model with all predictors.

***model0 = lm(Transport~., mydata2)***

***summary(model0)***

*Coefficients:*

*Estimate Std. Error t value Pr(>|t|)*

*(Intercept) -0.859940 0.148678 -5.784 1.40e-08 \*\*\**

*Age 0.021808 0.006321 3.450 0.000615 \*\*\**

*Gender 0.032144 0.022701 1.416 0.157500*

*Engineer 0.019843 0.023247 0.854 0.393822*

*MBA -0.028908 0.023235 -1.244 0.214115*

*Work.Exp -0.004513 0.007712 -0.585 0.558741*

*Salary 0.014745 0.002832 5.207 2.97e-07 \*\*\**

*Distance 0.011636 0.003112 3.739 0.000209 \*\*\**

*license 0.138636 0.028208 4.915 1.26e-06 \*\*\**

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 0.2092 on 435 degrees of freedom*

*Multiple R-squared: 0.6382, Adjusted R-squared: 0.6316*

*F-statistic: 95.93 on 8 and 435 DF, p-value: < 2.2e-16*

Here we see the adjusted R square is 63% which is average but only 4 ( out of 8 ) predictors is significant .

We will now check the VIF values for each predictor

*vif(model0)*

*Age Gender Engineer MBA Work.Exp Salary Distance license*

*7.889832 1.072856 1.015667 1.033233 15.735670 8.872136 1.275106 1.448074*

.

Note the high VIF value for Work experience which is causing the multicollinearity.

We will need to preform a step regression method to reduce the unwanted variables.

***# Performing Step regression to remove insignificant variables***

***ols\_step\_both\_p(model0, pent = 0.05, prem = 0.3, details = TRUE)***

Stepwise Selection Summary

--------------------------------------------------------------------------

Added/ Adj.

Step Variable Removed R-Square R-Square C(p) AIC RMSE

--------------------------------------------------------------------------

1 Salary addition 0.584 0.583 60.0840 -70.4657 0.2225

2 license addition 0.608 0.606 33.6520 -94.4555 0.2163

3 Age addition 0.622 0.619 18.4760 -108.9260 0.2126

4 Distance addition 0.634 0.631 6.0410 -121.2564 0.2094

--------------------------------------------------------------------------

So we have only 4 significant predictors:

* Salary
* License
* Age
* Distance

We will build a model with these 4 variables only and check the VIF .

***model1 = lm(Transport~Salary+Age+license+Distance, mydata2)***

***summary(model1)***

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.786267 0.104210 -7.545 2.63e-13 \*\*\*

Salary 0.013623 0.002007 6.787 3.73e-11 \*\*\*

Age 0.019407 0.004437 4.374 1.52e-05 \*\*\*

license 0.132235 0.027367 4.832 1.87e-06 \*\*\*

Distance 0.011742 0.003094 3.795 0.000169 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2094 on 439 degrees of freedom

Multiple R-squared: 0.6341, Adjusted R-squared: 0.6307

F-statistic: 190.2 on 4 and 439 DF, p-value: < 2.2e-16

In this model all the variables are significant, however we did not have much improvement in the R square value but we get excellent VIF for all variables.

vif(model1) # VIF is around 1 for all variables

Salary Age license Distance

4.446987 3.877891 1.359842 1.257421

So we can conclude that we have removed the multicollinearity from the data set.

**Data Preparation :Prepare the data for analysis (SMOTE)**

As we have seen in the beginning, our data set is quite imbalanced as our majority class( Mode of transport = Car ) is only about 13%. A problem with imbalanced classification is that there are too few examples of the minority class for a model to effectively learn the decision boundary. To tackle this imbalance issue , we can use SMOTE technique.

**SMOTE** stands for Synthetic Minority Oversampling Technique. This is a statistical technique for increasing the number of cases in your dataset in a balanced way. The module works by generating new instances from existing minority cases that you supply as input

To use SMOTE () we will use the DMwR package in R.

We first split the data in train(75%) and test(25%) sets. Since this is imbalanced data, it is very important to maintain the minority class ratio during splitting. In our case , we split based on Transport variable so that both the sets have similar ratio.

***split <- sample.split(cleandata$Transport, SplitRatio = .75)***

***smote.train<-subset(cleandata, split == TRUE)***

***smote.test<-subset(cleandata, split == FALSE)***

table(smote.train$Transport)

0 1

287 46 🡺 46/(46+287) = 13.81%

> table(smote.test$Transport)

0 1 🡺 15/(15+96) = 13.6%

96 15

So we see the minority class ratio is maintained in both sets.

**balanced.traindata <- SMOTE(Transport ~., smote.train, perc.over = 350, k = 5, perc.under = 150)**

perc.over means that 1 minority class will be added for every value of perc.over

Now let us see the distribution of minority class in SMOTE data set ( **balanced.traindata)**

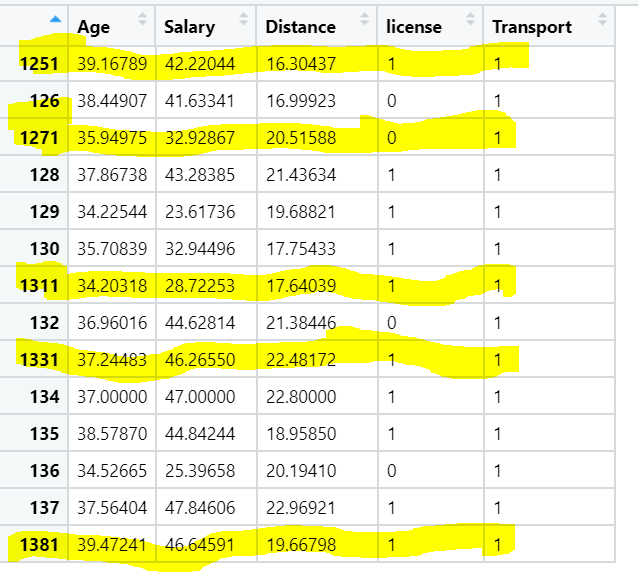
table(balanced.traindata$Transport)

0 1

207 184

By this the total number of observation is close to the original but the percentage of 1’s ( car) has increased now and the current ratio is 47%

Here are some of the augmented values after SMOTE.



In our further model building, we will use both this SMOTE data set and original data set and try to compare the model accuracy to understand the advantage of SMOTE.

**Modeling :** Create multiple models and explore how each model perform using appropriate model performance metrics

* + KNN
  + Naive Bayes
  + Logistic Regression

We will use the original data set as well as SMOTE data set for these 3 model building and will evaluate the model performances.

**Logistic Regression :**

**LogitModel1 = glm(Transport ~ .,**

**data = TrainData,**

**family = binomial(link="logit"))**

Deviance Residuals:

Min 1Q Median 3Q Max

-2.28231 -0.11800 -0.04665 -0.01215 2.91489

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -31.95688 5.95245 -5.369 7.93e-08 \*\*\*

Age 0.85309 0.18847 4.526 6.00e-06 \*\*\*

Salary -0.04842 0.04935 -0.981 0.32644

Distance 0.29735 0.10905 2.727 0.00640 \*\*

license 2.07056 0.78147 2.650 0.00806 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 267.445 on 332 degrees of freedom

Residual deviance: 58.576 on 328 degrees of freedom

AIC: 68.576

We see Age , Distance and license have been identified as significant variables in the model .

Just to ensure the absence of multicollinearity , we test the VIF values which confirms the multicollinearity has been removed from the dataset.

|  |
| --- |
| vif(LogitModel1)  Age Salary Distance license  2.099726 2.170907 1.084611 1.258092 |
|  |
|  |

We use this model on the test dataset and observer the confusion matrix.

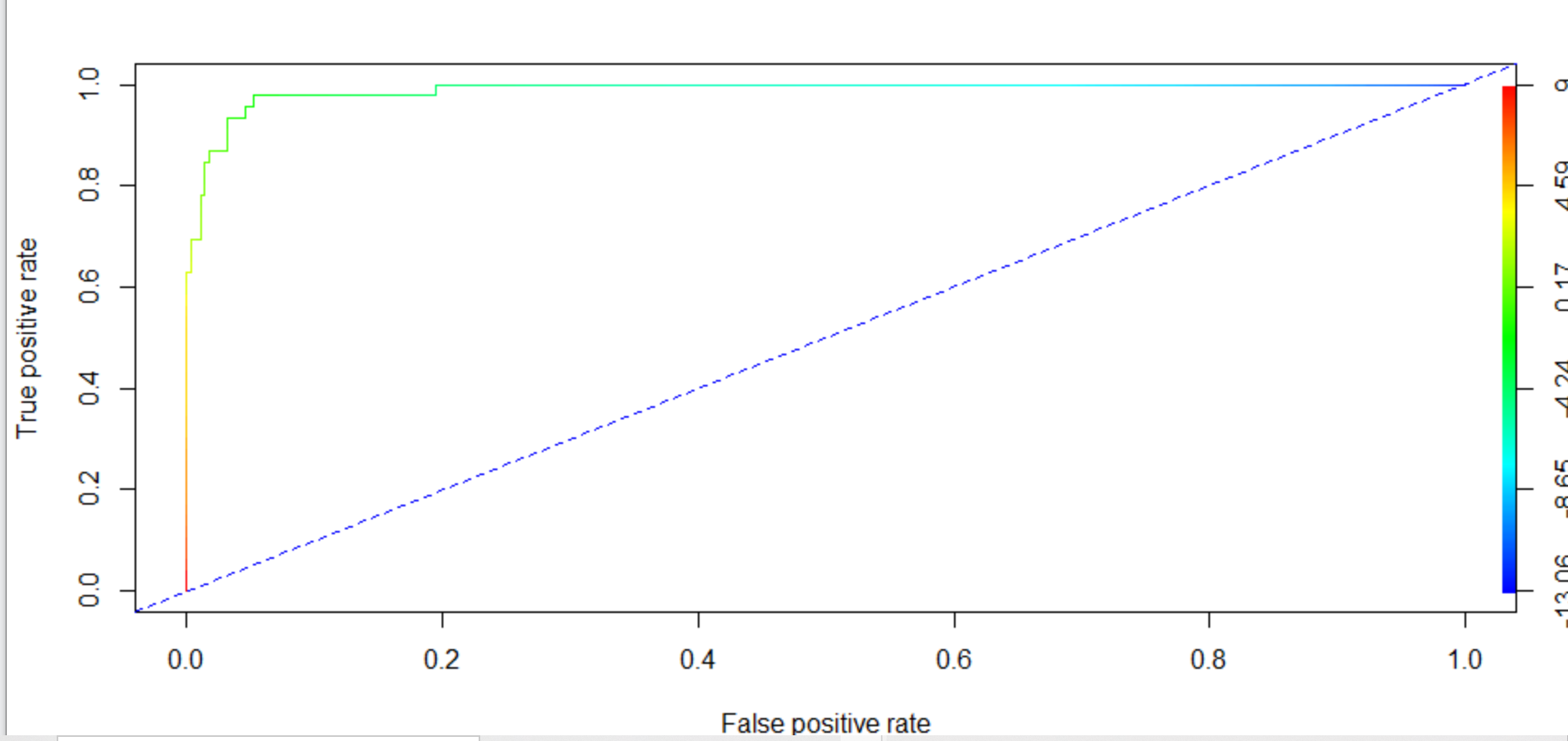
FALSE TRUE

0 96 0

1 7 8

This tells that we have predicted all the False ( no car ) values correctly , however out of actual 15 positive cases we could predict only 8 . so overall accuracy of the model is great but we fail to achieve a good prediction for our most important target employees who use car.

**ROC curve on the Train dataset**



Now let us run Logit on SMOTE dataset.

Confusion matrix looks like below:

FALSE TRUE

0 91 5

1 1 14

Here we have missed to identify few True Negatives but most importantly we have identified lot more true positives which is most important for our business case.

|  |  |  |
| --- | --- | --- |
|  | **Logistic Regression** | |
| Parameters | Original DataSet | SMOTE Dataset |
| Accuracy | 93.70% | 94.60% |
| Sensitivity | 53% | 93% |
| Specificity | 0% | 5% |
| AUC | 0.98 | 0.98 |
| KS | 0.92 | 0.92 |
| GINI Index | 0.97 | 0.97 |

So looking at the various model validation parameter , it is clear data logistic regression is giving better result with SMOTE dataset.

**KNN Model**

As a prerequisite for KNN modelling , the data set needs to be scaled to reduce error.

# Normalize variables

scale = preProcess(TrainData, method = "range")

train.norm.data = predict(scale, TrainData)

test.norm.data = predict(scale, TestData)

knn\_fit<- knn(train = train.norm.data[,1:4], test = test.norm.data[,1:4], cl= train.norm.data$Transport,k = 3,prob=TRUE)

Let us see the confusion matrix :

knn\_fit

FALSE TRUE

0 96 0

1 6 9

Just like initial LOGIT model , we have predicted all the False ( no car ) values correctly , however out of actual 15 positive cases we could predict only 9 . so overall accuracy of the model is great but we again fail to achieve a good prediction for our most important target employees who use car.

Let us run this on SMOTE dataset and analyse the confusion matrix.

knn\_fit1

FALSE TRUE

0 92 4

1 1 14

So in the SMOTE data set we have few true negative but we are able to predict all the positive cases (14 out of 15)

|  |  |  |
| --- | --- | --- |
|  | **KNN** | |
| Parameters | Original DataSet | SMOTE Dataset |
| Accuracy | 86.00% | 95.00% |
| Sensitivity | 60% | 93% |
| Specificity | 0% | 4% |

**Naïve Bayes Model :**

We run NB model on SMOTE data set and the confusion matrix is as below :

FALSE TRUE

POSITIVE 95 1

Negative 2 13

So we get a very good matrix here with accuracy = 97% and sensivity = 87%

**Apply both bagging and boosting modeling procedures to create 2 models and compare its accuracy with the best model of the above step**

**BAGGING**

We will now use Bagging technique on our SMOTE data set to see how our model works . Bagging (aka Bootstrap Aggregating): is a way to decrease the variance of your prediction by generating additional data for training from your original dataset using combinations with repetitions to produce multisets of the same cardinality/size as your original data.

To use Bagging , we will use **ipred** and **rpart** library in R.

We first need to convert all factor variables to numeric ones like License and Transport in the data set .

bagging1 <- bagging(Transport ~.,

data=balanced.traindata,

control=rpart.control(maxdepth=5, minsplit=4))

Now when we use this model on the tests SMOTE set , we get the below matrix.

FALSE TRUE

0 92 4

1 2 13

So we see a very good result here. The model provided great accuracy and

sensivity.

**Boosting**

We will try another ensemble method called boosting where we train the

weak learners sequentially to get better results.

xgb.fit <- xgboost(

data = features\_train,

label = label\_train,

eta = 0.001, #this is like shrinkage in the previous algorithm

max\_depth = 3, #Larger the depth, more complex the model; higher chances of overfitting. There is no standard value for max\_depth. Larger data sets require deep trees to learn the rules from data.

min\_child\_weight = 3,#it blocks the potential feature interactions to prevent overfitting

nrounds = 10000,#controls the maximum number of iterations. For classification, it is similar to the number of trees to grow.

nfold = 5,

objective = "binary:logistic", # for regression models

verbose = 0, # silent,

early\_stopping\_rounds = 10 # stop if no improvement for 10 consecutive

trees

)

We need to play around with the control parameters to get a better model

output .

FALSE TRUE

0 94 2

1 2 13

This model performs slightly better than bagging. We have rightly

predicted 2 more True negative cases. We will now try to find the best model using XGBOOST() .

An initial result of XGBOOST() is as below :

FALSE TRUE

0 91 5

1 3 12

Now , we will run this model for number of control parameters like below.

lr <- c(0.0001, 0.001, 0.01, 0.1, 0.3, 0.5, 0.7, 1)

md <- c(1, 3, 5, 7, 9, 15) // Maximum depth

nr <- c(2, 50, 100, 1000, 10000) // Number of rounds

We record the true positive rate for each of the iteration and finally get the best model as below :

##Best XGBoost Model

xgb.fit <- xgboost(data = features\_train,

label = label\_train,

eta = 0.7,

max\_depth = 5,

min\_child\_weight = 3,

nrounds = 50,

nfold = 5,

objective = "binary:logistic",

verbose = 1,

early\_stopping\_rounds = 10)

With the below confusion matrix

FALSE TRUE

0 91 5

1 1 14

Let us take a look at the comparative model performance:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | Sensivity | Specificity |
| Bagging Model | 94.50% | 86.60% | 4% |
| Boosting Model | 96.40% | 86.60% | 2% |
| Best XGBoost Model | 94.60% | 93.30% | 5% |

Looking into the model performance matric , we can concur that XGBoost model is giving the best performance in terms of providing higher higher sensivity .Please note this model performance is similar to our Logit model that we built on SMOTE data set.

**Actionable Insights & Recommendations:**

After completing the model building exercise and going through the various model performances , we understand that :

1. In our data set, most significant predictors to predict whether an employee will use car or not is:
   1. Age of the employee
   2. Whether the employee has license to drive or not
   3. Distance of the work location from home
   4. Salary of the employee
2. We have also seen that most of the employees below the age of 30 either use 2-wheeler or public transport.
3. Employees with higher salary (greater than 30lac) and aged more than 30 are most likely to use car.
4. So if an automobile company is planning for a sales promotion event , they should definitely target those employees with higher salary band and age above 30 years.
5. There is a good population of people who are at medium salary range but they are more than 30 and have to travel more in the public transport. They may not be using car because of financial constraints. But all other factors are favourable for them to use cars. We can also target these set of population with attractive discounts to acquire some new customer as converting then from 2 wheeler and public transport to car is easier.