Predicting Mode of Transport (ML)

Akalya KS

Table of Contents

[1 Project Objective 4](#_Toc43659658)

[2 Exploratory Data Analysis 4](#_Toc43659659)

[2.1 EDA - Basic data summary, Univariate, Bivariate analysis, graphs, Check for Outliers and missing values and check the summary of the dataset 4](#_Toc43659660)

[Exploratory Data Analysis: 4](#_Toc43659661)

[Univariate analysis: 5](#_Toc43659662)

[Bivariate analysis: 7](#_Toc43659663)

[Missing values and outliers: 8](#_Toc43659664)

[Multi Collinerarity: 8](#_Toc43659665)

[3.Data preparation and SMOTE: 9](#_Toc43659666)

[SMOTE: 10](#_Toc43659667)

[4.Building models: 10](#_Toc43659668)

[4.1.Logistic regression models: 10](#_Toc43659669)

[4.2.KNN Model: 12](#_Toc43659670)

[Interpretation of KNN Model: 13](#_Toc43659671)

[4.3. Applying Naive Bayes Model: 13](#_Toc43659672)

[Interpretation of Naïve Bayes model: 14](#_Toc43659673)

[4.4.Confusion matrix interpretation: 15](#_Toc43659674)

[Boosting and Bagging models: 15](#_Toc43659675)

[Applying bagging model: 15](#_Toc43659676)

[Applying Boosting: 16](#_Toc43659677)

[Actionable insights and recommendations: 17](#_Toc43659678)

# 

# 1 Project Objective

In this project ,we will have to study the preference of the transport which employees prefers to commute to their office.

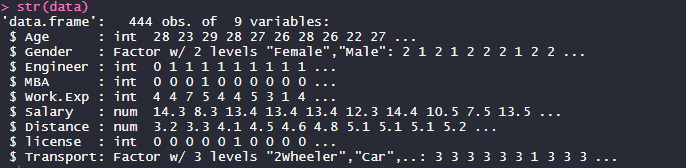
We need to predict whether or not an employee will use Car as a mode of transport. The objective is to build various Machine learning models to identify the preference.

# 2 Exploratory Data Analysis

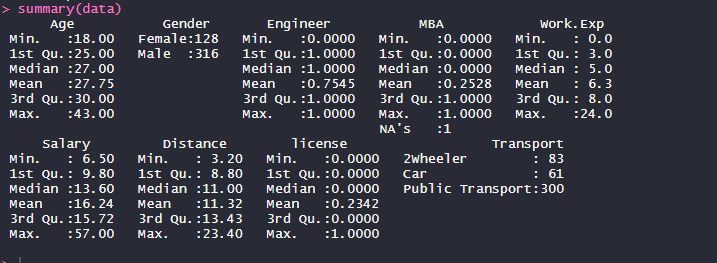
## 2.1 EDA - Basic data summary, Univariate, Bivariate analysis, graphs, Check for Outliers and missing values and check the summary of the dataset

### Exploratory Data Analysis:

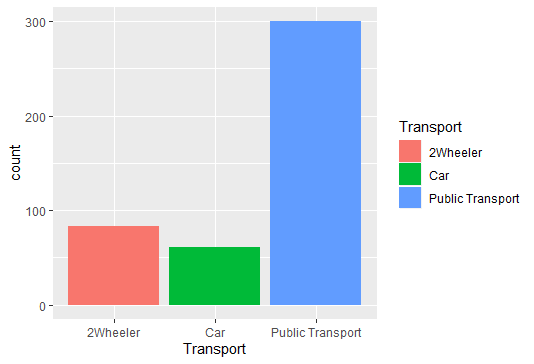
There are 9 variables in the dataset with 444 records.



Data summary:



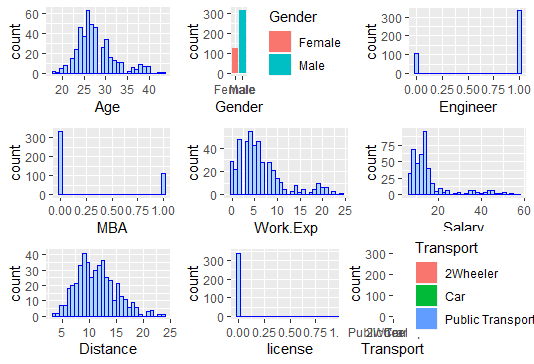
The summary of the data shows that target variable transport is a 3 class variable such as 2 wheeler,car and public transport.



The percentage distribution of transport variables as below:

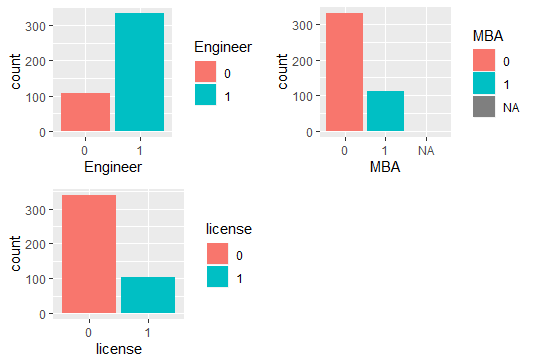


### Univariate analysis:

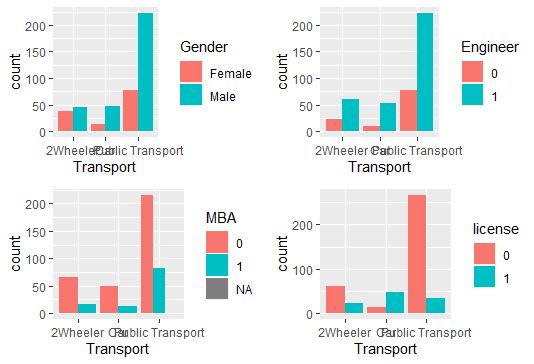


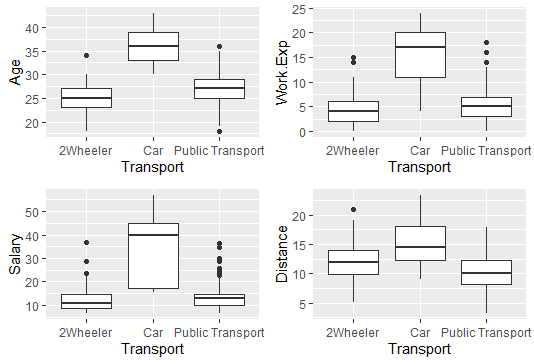
Insights:

Analysis shows that columns Engineer,MBA,license are behaving like categorical variables and hence can be converted to factors.



### Bivariate analysis:





Insights:

* Age & Transport: Plot shows that higher the age.mode of transport is car.
* Gender: Female prefer 2-wheeler more when compared to car and public transport.Very few female prefer car than public transport.Majorly 2 wheeler and public transport is used by female.
* Engineer: There is no significant difference due to engineer.
* MBA:Public transport is preferred by non-MBA when compared to MBA.
* License: People with no license are using 2-wheelers more than license people.Car is prefered by people with license more even though people without license is using both car and 2-wheeler.Public transport is dominated by people with no license.
* Work experience: People with work experience of more than 15 years is using cars.More experience leads to more usage of cars.
* Salary: Higher the salary,people prefer cars.2-wheeler and public transport is preferred by people with low salary.
* Distance: Car is preferred for longer distance.

Bivariate analysis shows age,salary,work experience and distance contributes to the usage of cars.They are the factors which will help in prediction.

### Missing values and outliers:

There was only one NA value in MBA column which was treated using Knn imputation.

Outliers are actually the real data collected which we will not treat since they will help in predicting models.

### Multi Collinerarity:

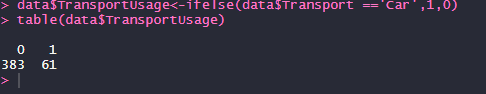


Insights:

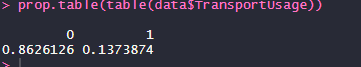
The multicollinerarity plot shows that work experience.age,salary are highly correlated.

## 3.Data preparation and SMOTE:

Since we are going to prepare models to undeestand the factors influence the car usage ,we will need to understand the proportion of cars being used in the data.Hence,we will convert the 3 class Tranpsort variable to 2- class variables where car will take 1 , 2-wheeler and public transport will take 0. We will store this in a new column as ‘Transport usage’.

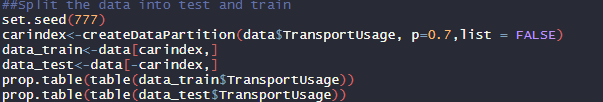


The publictransport and 2-wheeler is 86.2% and car is used at 13.7% in the given dataset.

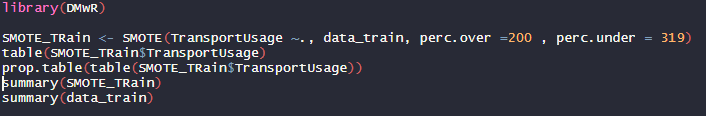


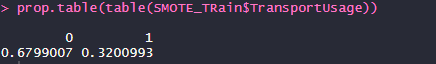
The proportion of car and other transport data is imbalanced and we will do SMOTE to balance the data before building models.

We will split the data into train and test dataset where SMOTE is applied to only train dataset.



### SMOTE:

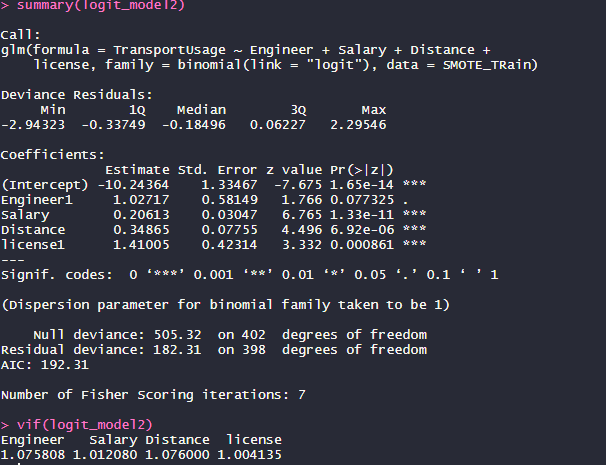
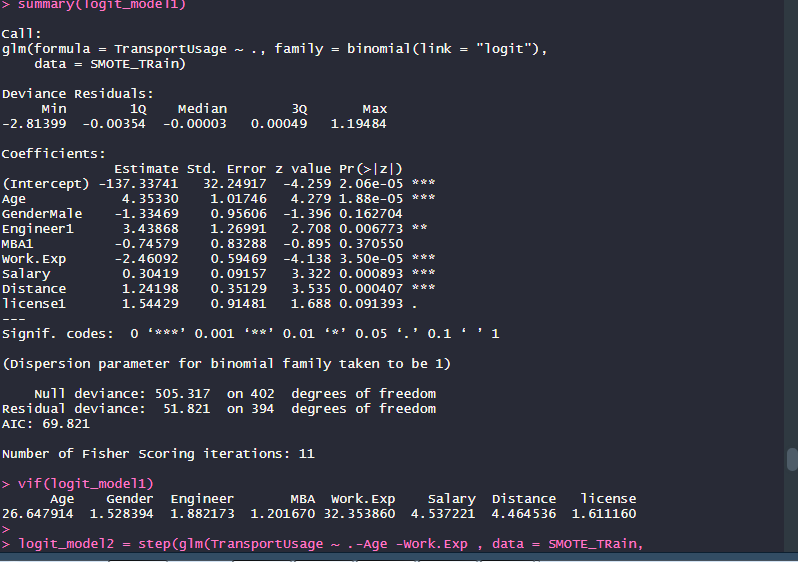




After balancing the data using SMOTE,we can see more than 10% increase in data which we will use for building models such as Logistic regression, Knn and Naïve Bayes model.

## 4.Building models:

### 4.1.Logistic regression models:



Confusion Matrix and Statistics

Reference

Prediction 0 1

0 260 29

1 14 100

Accuracy : 0.8933

95% CI : (0.859, 0.9217)

No Information Rate : 0.6799

P-Value [Acc > NIR] : < 2e-16

Kappa : 0.7471

Mcnemar's Test P-Value : 0.03276

Sensitivity : 0.7752

Specificity : 0.9489

Pos Pred Value : 0.8772

Neg Pred Value : 0.8997

Prevalence : 0.3201

Detection Rate : 0.2481

Detection Prevalence : 0.2829

Balanced Accuracy : 0.8620

'Positive' Class : 1

Applying Logistic regression shows that initially age,work experience highly significant and after removing them and performing vif,we can see the values are in range.

Interpretation:

The results show us the distribution of deviance residuals for the individual components used the dataset. We can summarize them as below:

1. Since maximum deviance is 2.29, It’s a good model. Lower is the deviance, better is the model.
2. The variables Age,work experience,alary,distance,engineer and license are significant.
3. Again, the difference between the residual and null deviance signifies that the model is a good once since the difference is high.
4. For Age, work experience the VIF value is greater than 5, which means the model has problem in estimating the coefficients.
5. The positive prediction value is 87.7% only and the sensitivity is 77.5%.The general model has an accuracy rate of 89% which is okay for the model prediction using the balanced data.

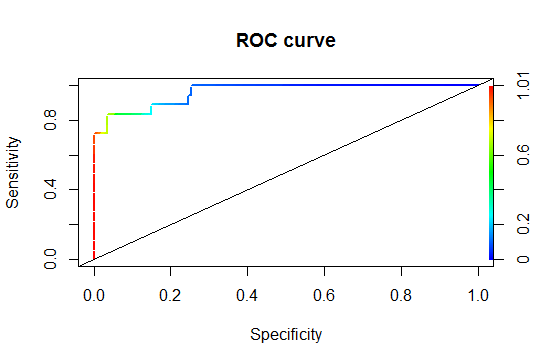
Using the balanced data,we got the AUC,ROC curve,KS and gini values.

AUC value:

> AUC

[1] 0.960039

ROC curve:



KS:

> train.ks

[1] 0.7982456

GINI value:

> train.gini

[1] 0.920078

### 4.2.KNN Model:

k-Nearest Neighbors

403 samples

8 predictor

2 classes: '0', '1'

Pre-processing: centered (8), scaled (8)

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 363, 362, 362, 363, 363, 363, ...

Resampling results across tuning parameters:

k Accuracy Kappa

5 0.9206697 0.8153524

7 0.9198364 0.8118234

9 0.9116046 0.7898279

11 0.9099776 0.7853529

13 0.9091646 0.7815184

15 0.9074573 0.7770144

17 0.9049359 0.7688212

19 0.9033109 0.7645682

21 0.8916839 0.7347938

23 0.8958312 0.7460219

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was k = 5.

> knn.CM\_train

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 264 13

1 10 116

Accuracy : 0.9429

95% CI : (0.9156, 0.9635)

No Information Rate : 0.6799

P-Value [Acc > NIR] : <2e-16

Kappa : 0.8681

Mcnemar's Test P-Value : 0.6767

Sensitivity : 0.8992

Specificity : 0.9635

Pos Pred Value : 0.9206

Neg Pred Value : 0.9531

Prevalence : 0.3201

Detection Rate : 0.2878

Detection Prevalence : 0.3127

Balanced Accuracy : 0.9314

'Positive' Class : 1

### Interpretation of KNN Model:

* Trained tuned model for k-NN gives 5 as the optimal value
* KNN model has the accuracy rate of 94.29 % which is higher than logistic regression model.
* The specificity is 96.35% and positive prediction value is 77.27%.

### 4.3. Applying Naive Bayes Model:

Naive Bayes Classifier for Discrete Predictors

Call:

naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:

Y

0 1

0.6799007 0.3200993

Conditional probabilities:

Age

Y [,1] [,2]

0 26.40146 3.122551

1 34.73942 3.028973

Gender

Y Female Male

0 0.3211679 0.6788321

1 0.3488372 0.6511628

Engineer

Y 0 1

0 0.2153285 0.7846715

1 0.1395349 0.8604651

MBA

Y 0 1

0 0.6861314 0.3138686

1 0.7519380 0.2480620

Work.Exp

Y [,1] [,2]

0 5.014599 3.089591

1 14.338148 4.443303

Salary

Y [,1] [,2]

0 12.88102 4.556363

1 32.11274 11.267554

Distance

Y [,1] [,2]

0 10.17518 3.050342

1 15.35171 3.133974

license

Y 0 1

0 0.8686131 0.1313869

1 0.4263566 0.5736434

* For continuous variables Naïve Bayes takes the mean and standard deviation or variability and treats it as cut off thresholds; say anything less than mean of distributed predictor values is 0 and more than mean is 1.

### Interpretation of Naïve Bayes model:

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 266 16

1 8 113

Accuracy : 0.9404

95% CI : (0.9127, 0.9615)

No Information Rate : 0.6799

P-Value [Acc > NIR] : <2e-16

Kappa : 0.8609

Mcnemar's Test P-Value : 0.153

Sensitivity : 0.8760

Specificity : 0.9708

Pos Pred Value : 0.9339

Neg Pred Value : 0.9433

Prevalence : 0.3201

Detection Rate : 0.2804

Detection Prevalence : 0.3002

Balanced Accuracy : 0.9234

'Positive' Class : 1

* Accuracy of NB model is 96.97% which is higher than both KNN and LR model.
* The positive prediction value is 93.3% ,specificity is 97.08%.

### 4.4.Confusion matrix interpretation:

In the business point of view,decision is made on positive rates for predicting the car usage.

Hence,we will evaluate models based on accuracy on test data,sensitivity to compare model performances.

|  |  |  |  |
| --- | --- | --- | --- |
| **Metrics** | **Logistic Regression** | **Naïve Bayes** | **KNN** |
| Accuracy | 93.94% | 97.73% | 95.45% |
| Specificity | 75.00% | 98.25% | 97.37% |
| Sensitivity | 97.32% | 94.40% | 83.30% |

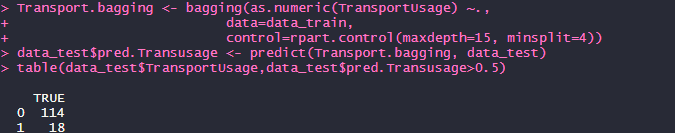
Interpretation:

Accuracy is higher for NaiveBayes model when compared to Lr and KNN model.But sensitivity is higher for LR model which proves that our models are not stable.

## Boosting and Bagging models:

Bagging and boosting are ensemble models where bagging uses random forests to train the data as multiple models using same algorithm and helps in creating the stronger model.

### Applying bagging model:



Interpretation:

Our bagging models is using the baseline approach calling everything as true,hence it’s in extreme.

### Applying Boosting:

For performing the boosting model,here we are using xgboost which will expect all the variables to numeric.Hence,we will convert variables to numeric.

features\_train = as.matrix(data\_train[,1:8])

> label\_train = as.matrix(data\_train[,9])

> features\_test = as.matrix(data\_test[,1:8])

> XGBmodel = xgboost(

+ data = features\_train,

+ label = label\_train,

+ eta = .001,

+ max\_depth = 5,

+ min\_child\_weight = 3,

+ nrounds = 10,

+ #nfold = 5,

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

+ verbose = 0, # silent,

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

+ )

> XGBpredTest = predict(XGBmodel, features\_test)

> tabXGB = table(data\_test$TransportUsage, XGBpredTest>0.5)

tabXGB

FALSE TRUE

0 111 3

1 3 15

Our xgboost model provides the accuracy rate of 95.45%.

#Accuracy: 95.45%

> sum(diag(tabXGB))/sum(tabXGB)

[1] 0.9545455

>

> #specificity : 83.33%

>

> 15/18

[1] 0.8333333

>

> #sensitivty :83.33% tp/p

>

> 15/18

[1] 0.8333333

Model comparison:

Using Smote train data ,we build Logistic regression,NB and Knn models and the accuracy using test data shows NB model performed better.Bagging models shows complete accuracy and boosting models shows 95.45% where our bagging has predicted 100% car users prediction.

## Actionable insights and recommendations:

* The variables like Age, Work.Experience, Distance and License are the important predictors for identifying transport preference.
* Age and Work.Exp are correlated hence we could use any one (prefer Work.Exp).
* Employees with work experience of 10 years and above are predicted 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. But surprisingly 72% without license use 2-wheeler.
* Again, people with higher salaries (>20) are likely to use cars