PREDICTING EMPLOYEE MODE OF TRANSPORT

Group#8

Abstract

Predicting Employee Mode of transport for a Company leveraging Machine Learning Techniques

Jithesh / Vinod Kumar /Tarique / Shivkumar

Contents

This project requires us to understand what mode of transport employees prefers to commute to their office. The dataset Cars_edited.csv 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 an employee will use Car as a mode of transport or not. Also, which variables are a significant predictor behind this decision?

Exploratory Data Analysis

Data Summary

The dataset consists of 9 Variables with a total of 444 Observations. Following are the characteristics of the variables: -

```
$ 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 ...
```

Professional Qualifications of the employees in the base dataset (Engineer, MBA) are of the type Integer and need to be converted to Factors. Employees having a License or not also needs to be converted to the type Factor.

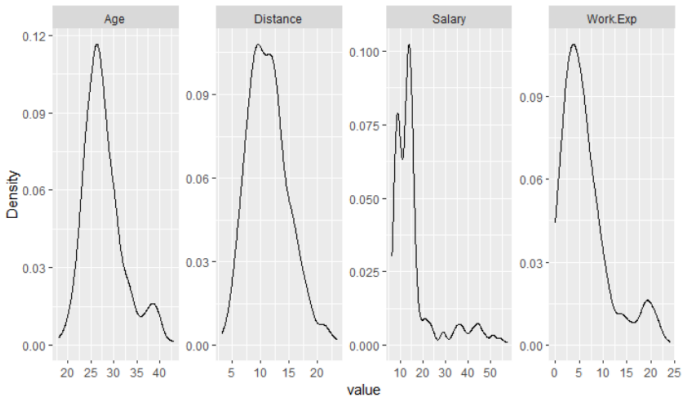
MBA records has ONE missing value present in the dataset.

Univariate Analysis

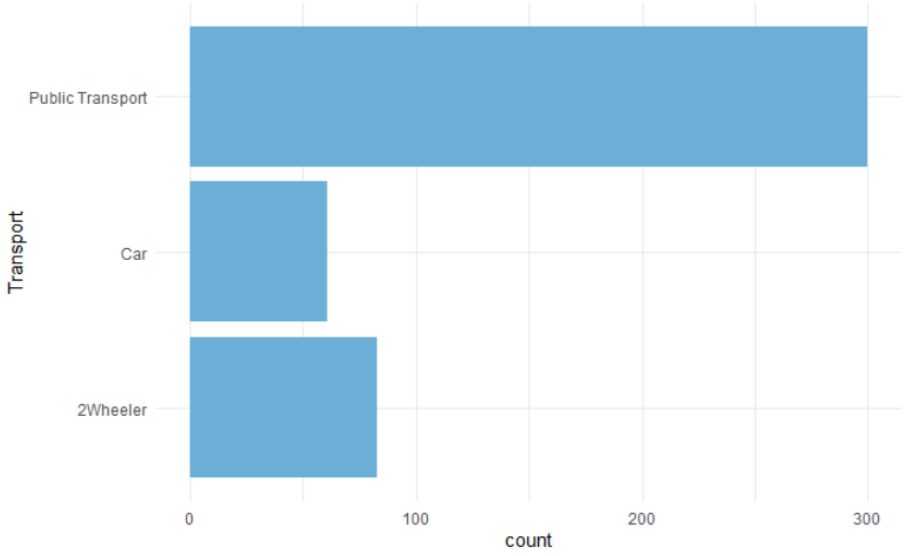
Following table presents the summary statistics of the dataset: -

```
      Age          Gender    Engineer    MBA         Work.Exp        Salary          Distance      license
 Min.   :18.00   Female:128   0:109    0   :331   Min.   : 0.0   Min.   : 6.50   Min.   : 3.20   0:340
 1st Qu.:25.00   Male  :316   1:335    1   :112   1st Qu.: 3.0   1st Qu.: 9.80   1st Qu.: 8.80   1:104
 Median :27.00                        NA's:  1   Median : 5.0   Median :13.60   Median :11.00
 Mean   :27.75                                   Mean   : 6.3   Mean   :16.24   Mean   :11.32
 3rd Qu.:30.00                                   3rd Qu.: 8.0   3rd Qu.:15.72   3rd Qu.:13.43
 Max.   :43.00                                   Max.   :24.0   Max.   :57.00   Max.   :23.40
          Transport
 2Wheeler       : 83
 Car            : 61
 Public Transport:300
```

Following represents the distribution of the data with continuous variables. The data related to Age and Distance is close to a normal distribution, while the data distribution related to Salary and Work Experience is right skewed as the Mean is greater than the Median. This leads to outliers in the data related to salary and work experience.

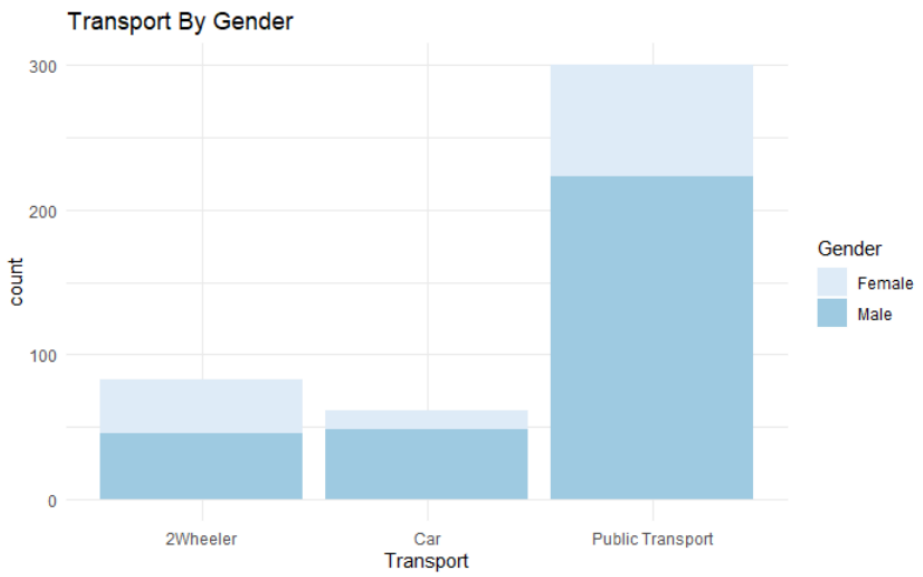


Public Transport users constitute 67.6% of the data followed by Two-Wheeler users (18.7%) and Car users (13.7%)

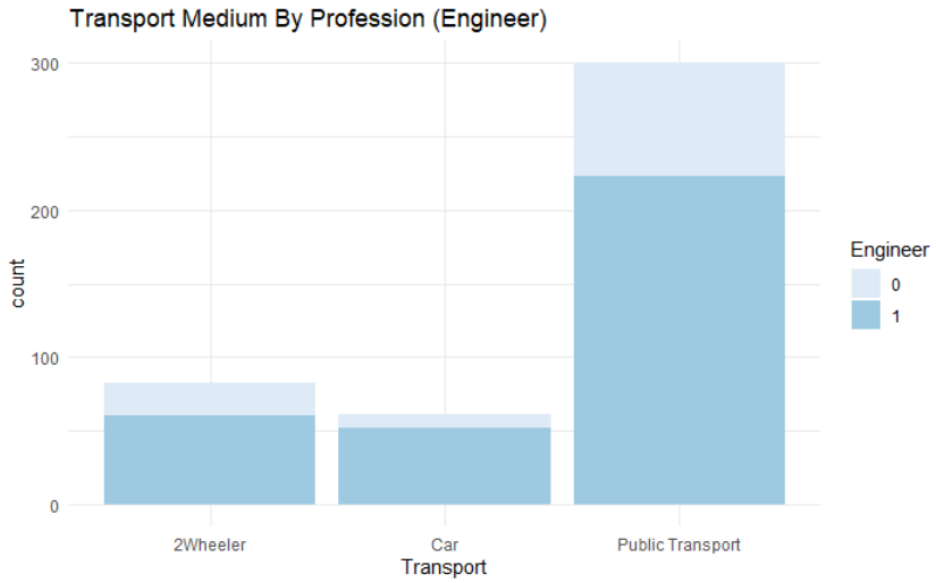


Bivariate Analysis

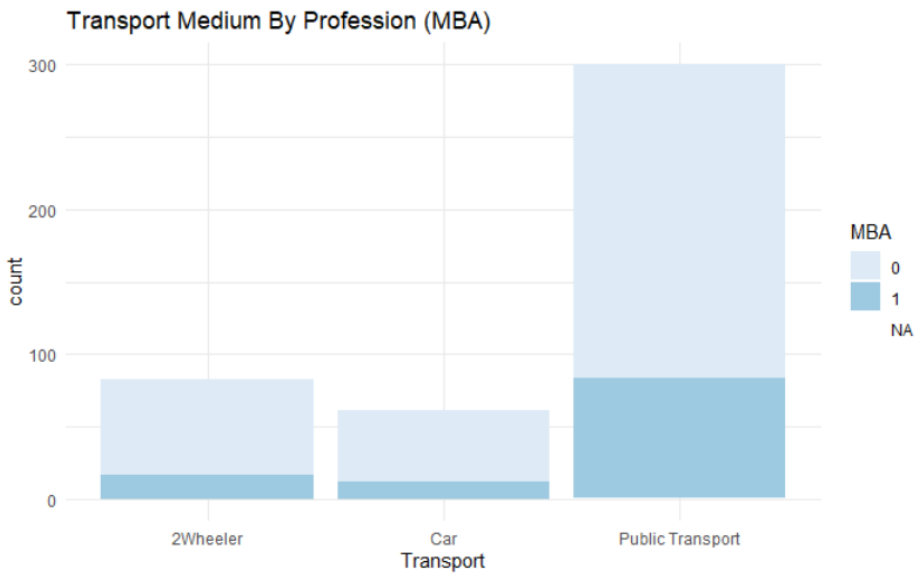
Male Employees which constitute 71% of the data are dominant users of all the three modes of transport.



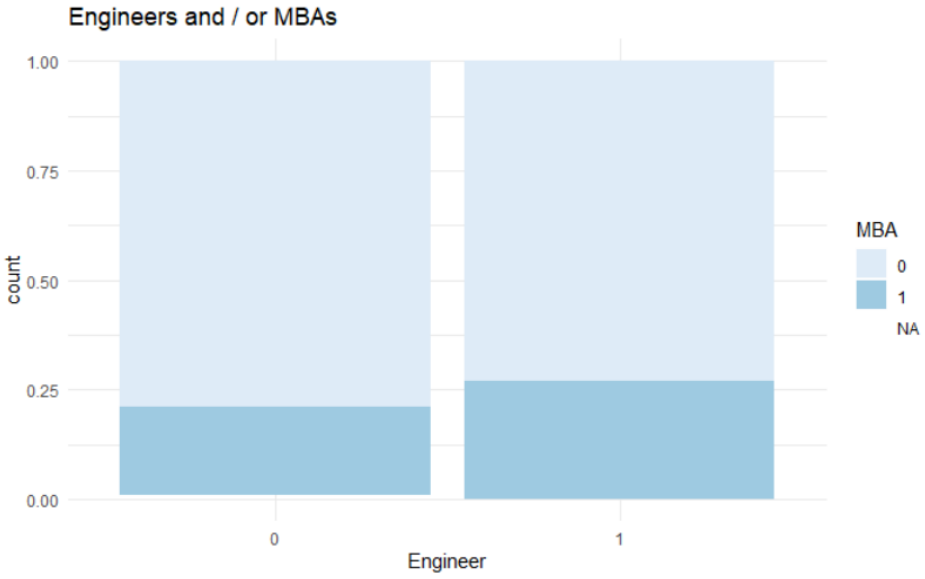
Engineers who constitute 75% of the data are dominant users of all the three modes of Transport.



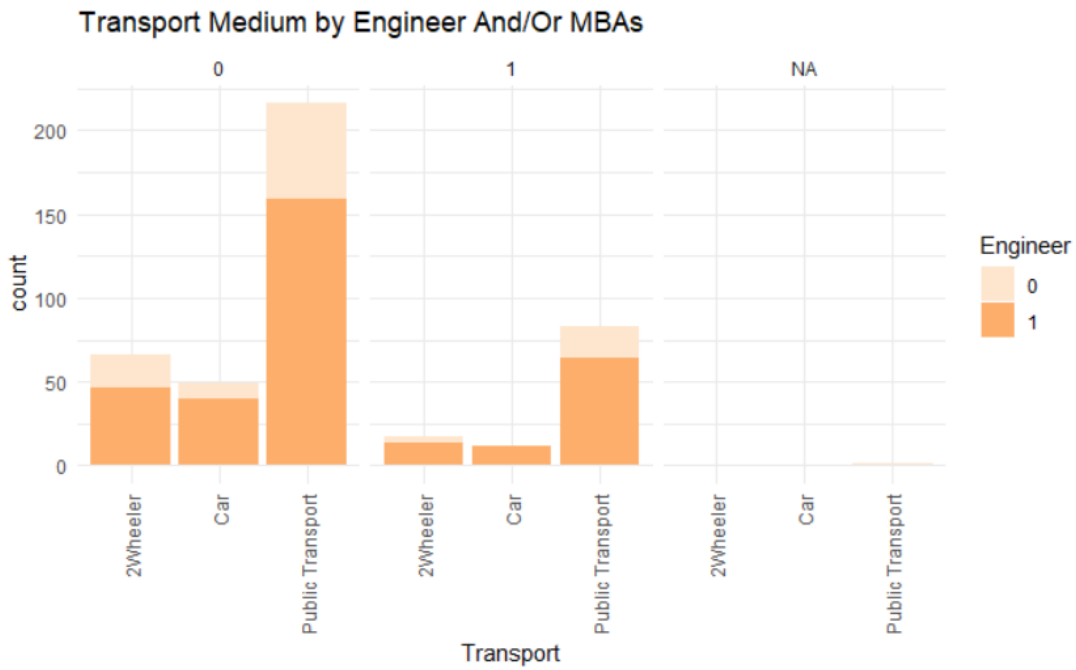
Non-MBAs who constitute 75% of the data are dominant users of all the three modes of transport.



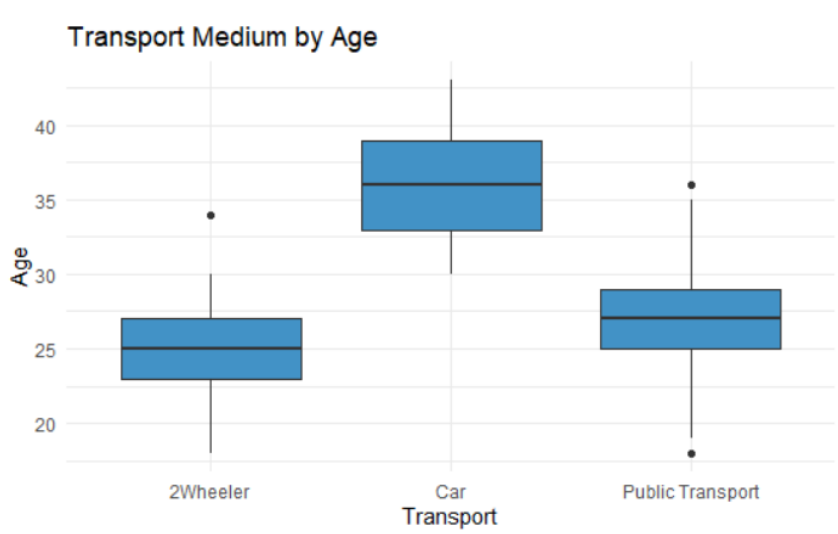
Engineers who are also MBAs constitute close to 25% of the Engineering population.



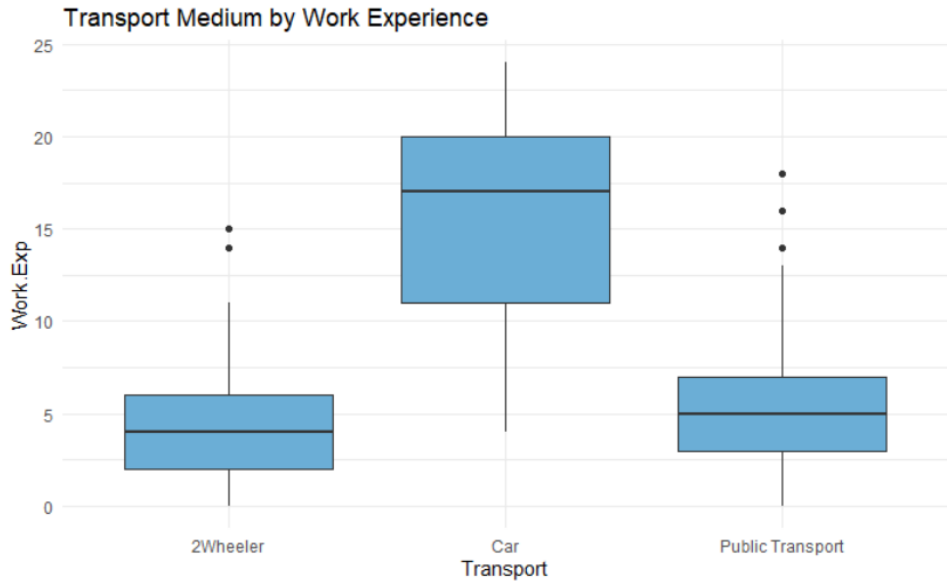
Engineers who are non-MBAs are dominant users of all the three modes of transport than the Engineers who are MBAs as well. Public Transport is being used by a larger number of employees who are Engineers and non-MBAs



The median Age of employees travelling by Car is the highest as compared to other two modes of transport.

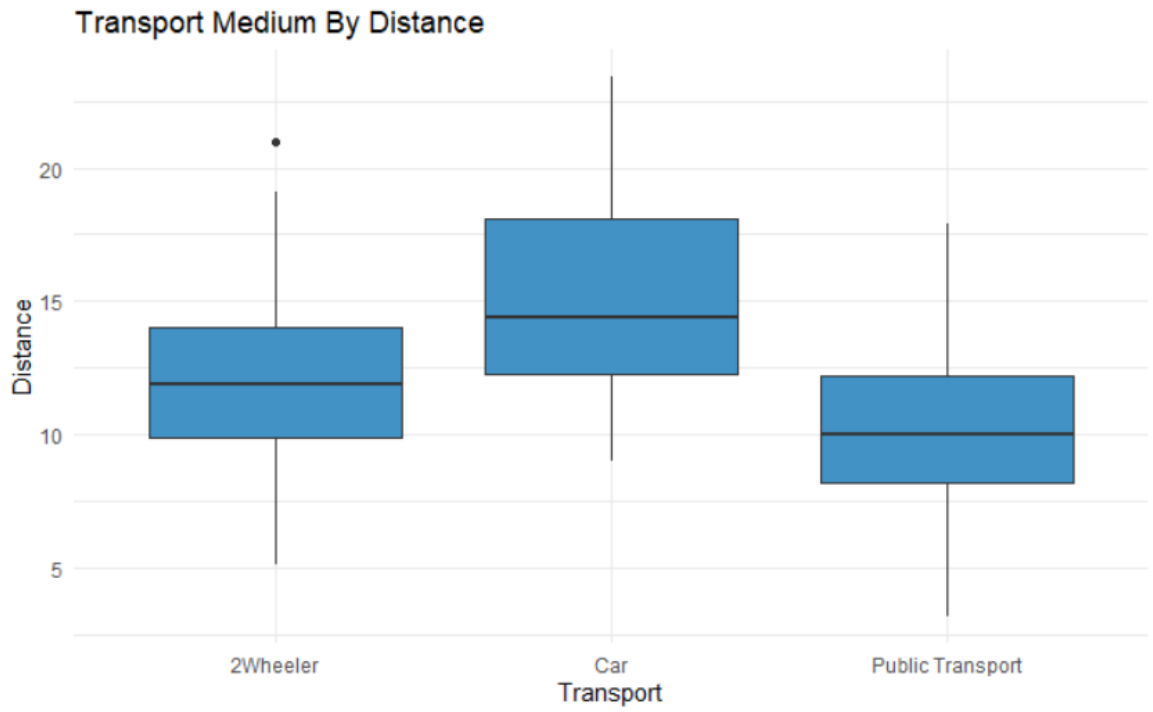


Median Work Experience of Employees using Car as a mode of transport is higher than the other two.



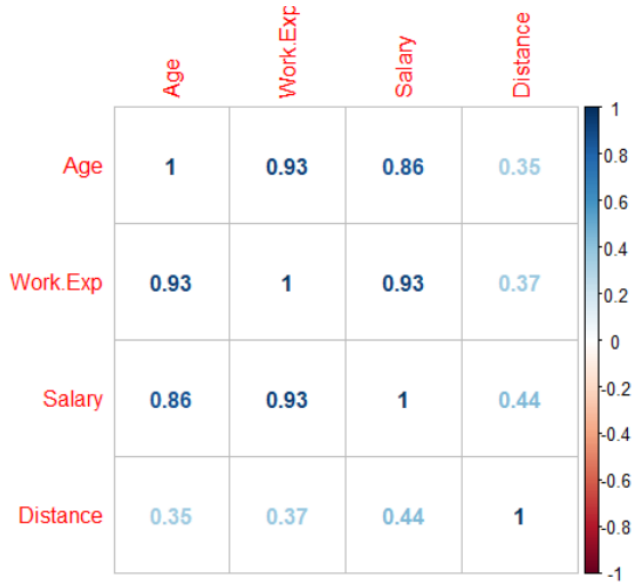
The Median Salary of Employees who travel by car is higher than the that of the other two modes of transport.

Median Distance travelled by Employees who use a car is higher than the other two modes of transport.



Multi-Collinearity

Employee Age, Work-Experience and Salary are highly correlated with each other and this could be a scenario of Multi-Collinearity



Statistical Significance of Multi-Collinearity

Variable 1	Variable 2	P-Value
Age	Work Experience	< 2.2e-16
Age	Salary	< 2.2e-16
Salary	Work Experience	< 2.2e-16

Observations

Thus, Employees who travel by Car exhibit the following characteristics: -

- High Work Experience
- High Salary / Income
- Reside at a location far away from office

Engineers who are non-MBAs are dominant users of all the three modes of transport than the Engineers who are MBAs as well. Public Transport is being used by a larger number of employees who are Engineers and non-MBAs

Employee Age has a high degree of multi-collinearity with Work Experience and Salary respectively. So does Work Experience which has a high degree of multi-collinearity with Salary. Thus, variables Age, Work Experience and Salary are not completely independent if each other. Distance could be independent of the remaining variables

Data Preparation

Following steps were taken to prepare the data: -

- As we need to predict whether an employee will use Car as a mode of transport, the Transport column in the dataset needs to be converted to a Binary classification column
    - 1 - Employees who use car as a medium of transport
    - 0 - Employees who do not use car as a medium of transport
- There is ONE missing record in the MBA column which has been replaced by the mode of the data
- Gender column has been converted to a binary classification column
    - 1 – Male
    - 0 – Female
- The dataset has been scaled for KNN based algorithms
- The dataset has been split in train and test based on 70:30 ratio

Modelling

Logistic Regression

LR#1 - Full Model

Following is the output of a Full Model based on Logistic Regression: -

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -89.6295    24.3590  -3.680 0.000234 ***
Age           2.8854     0.8138   3.546 0.000392 ***
Gender1      -1.1219     1.2852  -0.873 0.382720
Engineer1     0.2427     1.0477   0.232 0.816824
MBA1         -1.0421     1.1101  -0.939 0.347856
Work.Exp     -1.6518     0.5782  -2.857 0.004280 **
Salary        0.2096     0.1063   1.972 0.048553 *
Distance      0.6858     0.2237   3.066 0.002173 **
license1      2.3149     1.1121   2.082 0.037375 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significant variables based on p-values are as follows: -

- Age
- Work Experience
- Salary
- Distance
- License

Odds Ratio based on the Full Model

```
  (Intercept)          Age      Gender1    Engineer1         MBA1     Work.Exp       Salary     Distance     license1
1.186888e-39 1.791129e+01 3.256747e-01 1.274650e+00 3.527174e-01 1.917126e-01 1.233161e+00 1.985386e+00 1.012393e+01
```

For a unit increase in Age, the odds of an employee to travel by car *increases by 17.91 times* versus travelling by other modes of transport. An Employee having License increases the odds of travelling by car as well.

However, Age and Work Experience have a high VIF value and this indicates multi-collinearity between the two variables. Variable Inflation Factor results based on the Full Model are as follows: -

```
      Age    Gender  Engineer       MBA  Work.Exp    Salary  Distance   license
19.711926  1.517249  1.250535  1.480806 29.472571  5.719133  2.478930  1.914832
```

LR#2 - Model after dropping Work Experience

Following is the output of the Model after dropping Work Experience which has a high degree of multi-collinearity with Age: -

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -38.19685    7.83092  -4.878 1.07e-06 ***
Age           1.01246    0.23200   4.364 1.28e-05 ***
Gender1      -0.45797    1.04287  -0.439 0.660554
Engineer1     0.38655    0.93057   0.415 0.677859
MBA1         -1.69095    0.97763  -1.730 0.083693 .
Salary       -0.04465    0.05475  -0.816 0.414764
Distance      0.45996    0.13790   3.336 0.000851 ***
license1      1.59939    0.81664   1.959 0.050170 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Variables considered for this model are NOT multi-collinear as shown by the below VIF output

```
     Age   Gender Engineer      MBA   Salary Distance  license
2.290377 1.195850 1.111478 1.442990 2.055122 1.309078 1.306927
```

Odds Ratio based on the Model without Work Experience

```
  (Intercept)          Age      Gender1    Engineer1         MBA1       Salary     Distance     license1
2.578206e-17 2.752364e+00 6.325644e-01 1.471888e+00 1.843442e-01 9.563340e-01 1.584007e+00 4.950018e+00
```

The following variables have an impact on using car as a transport mode in the descending order based on the odd ratio

- License
- Age
- Distance
- Engineer
- MBA
- Gender
- Salary

LR#3 – Model considering significant variables from LR#2

Following is the output of the Model based on significant variables and close to significant variables from LR#2: -

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -35.4993     6.8122  -5.211 1.88e-07 ***
Age           0.9080     0.1875   4.842 1.29e-06 ***
MBA1         -1.8899     0.9271  -2.038 0.041501 *
Distance      0.4445     0.1341   3.315 0.000915 ***
license1      1.2527     0.7109   1.762 0.078030 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Odd Ratio

```
 (Intercept)          Age         MBA1     Distance     license1
3.827003e-16 2.479295e+00 1.510892e-01 1.559669e+00 3.499847e+00
```

The following variables have an impact on using car as a transport mode versus other modes in the descending order based on the odds ratio

- License
- Age
- Distance
- MBA

A unit increase in Age leads to increase in the ODDS of using a car as a mode of transport versus other modes by 2.5 times. A unit increase in the distance leads to an increase in the ODDs of using a car a mode versus other modes by 1.6 times.

LR#5 – Model based on ‘UP Sampling

Employees travelling by car forms 13.7% of the dataset, this could be a scenario of class imbalance. Following is the output of a Logistic Regression Model created by UPSAMPLING the dataset to remove class imbalance. Work-Experience variable is not considered on account of multi-collinearity with Age.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -37.65686    4.63781  -8.120 4.68e-16 ***
Age           1.06588    0.13888   7.675 1.66e-14 ***
Gender1      -0.76251    0.62455  -1.221    0.2221
Engineer1     0.49176    0.59591   0.825    0.4092
MBA1         -0.93702    0.56371  -1.662    0.0965 .
Salary       -0.03685    0.03514  -1.049    0.2943
Distance      0.42091    0.08735   4.819 1.44e-06 ***
license1      1.41089    0.57685   2.446    0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic Regression Model Performance Measures based on Train Data

LR Model	Sensitivity	Specificity	Accuracy
LR#1	83.7%	98.88%	96.7%
LR#2	76.7%	99.25%	96.14%
LR#3	76.7%	99.25%	96.14%
LR#5	95.35%	94.4%	94.53%

Logistic Regression Model Performance Measures based on Test Data

LR Model	Sensitivity	Specificity	Accuracy	Wald Test (p-value)
LR#1	67%	98.2%	93.98%	0.035
LR#2	83.33%	98.2%	96.24%	0.00019
LR#3	77.78%	98.2%	95.49%	8.4e-06
LR#5	94.44%	95.65%	95.49%	

Best Model for Logistic Regression is LR#5 which was Up Sampled to remove class imbalance.

KNN

Data Preparation

- The dataset was split into Train and Test Data with a 70:30 ratio
- For KNN#1, Train Data and Test Data were scaled independently, to include the continuous and categorical variables. This ensures that the test data does not inherit characteristics of train data
- For KNN#2, only the numeric variables were scaled
- For KNN#3, Test data continuous variables were centered & scaled on train data
- K value considered was 17 (square root of number of sample)

Following are the model performance measures based on k-value of 17: -

KNN Model	Sensitivity	Specificity	Accuracy
KNN#1	69.23%	100	81.2%
KNN#2	76.92%	99.09%	94.89%
KNN#3	72.22%	100%	95.45%

Naïve Bayes

Naïve Bayes may not apply for this dataset as Age, Work Experience and Salary are not independent of each other (key assumption in Naïve Bayes). There is a high degree of multi-collinearity as explained in the previous Section (Exploratory Data Analysis)

However, a Naïve Bayes model has been created and following are the results: -

Naïve Bayes Model	**Sensitivity**	**Specificity**	**Accuracy**
NB#1 – Full Model	83.33%	98.26%	96.24%
NB#2 - Model without Work Experience	72.22%	99.13%	95.48%

Confusion Matrix Interpretation & Model Validation

		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

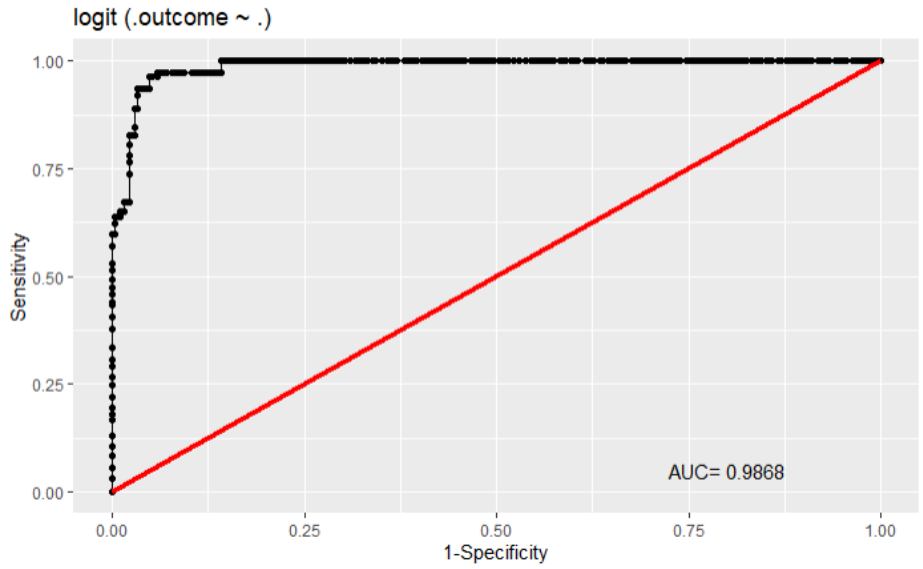
- 0 – Those Employees who do not use cars as a transportation mode (Negative Class)
- 1 – Those Employees who use car as a transportation mode (Positive Class)

The models for which the confusion matrix has been created are for

- Logistic Regression
- KNN
- Naïve Bayes

Model	**Sensitivity**	**Specificity**	**Accuracy**
NB#1 – Full Model	83.33%	98.26%	96.24%
KNN#2	76.92%	99.09%	94.89%
LR#5	94.45%	95.65%	95.49%

ROC Plot for Logistic Regression Model LR#5



The model which predicts those employees who use the car the best are Logistic Regression (LR#5) and Naïve Bayes(NB#1). Naïve Bayes (NB#1) considers all the variables including the multi-collinear variables and thus defies the need for variables to be Independent. Thus Logistic Regression(LR#5) has the best Sensitivity values and this model does not consider Work Experience as this has a high multi-collinearity with Age

Bagging and Boosting Models

Bagging

Following are the results of the random forest model for the train data

Random Forest	Hyperparameters	Sensitivity	Specificity	Accuracy
RF#1	Mtry = 4 Ntree = 200	90.69%	98.88%	97.7%
RF#2	Mtry = 3 Ntree=150	88.37%	98.89%	97.43%

Following are the results of the random forest model for the test data

Random Forest	Hyperparameters	Sensitivity	Specificity	Accuracy
RF#1	Mtry = 4 Ntree = 200	77.78%	98.2%	95.48%
RF#2	Mtry = 3 Ntree=150	94.45%	98.26%	97.74%

Boosting

Gradient Boosting (GBM)

Three gradient boosting models were run and following is the summary based on relative variable importance scores

Gbm#1: Cv folds 10, Ntrees 200

var <chr>	rel.inf <dbl>
Age	47.85921670
Salary	45.78224087
Distance	2.60162683
Work.Exp	2.59220870
license	0.95665272
MBA	0.15422830
Gender	0.05382589
Engineer	0.00000000

Gbm#2: cv folds: 20, ntrees 200

var <chr>	rel.inf <dbl>
Age	50.1456723
Salary	39.1915447
Work.Exp	5.0674265
Distance	3.3106586
license	1.3448940
MBA	0.5533669
Gender	0.3864370
Engineer	0.0000000

Gbm#3: repeatedcv 10, using caret library

var <chr>	rel.inf <dbl>
Age	47.3828105
Salary	37.2359219
Distance	5.6031595
Work.Exp	4.8150130
license1	3.2195369
MBA1	1.3597904
Gender1	0.3837678
Engineer1	0.0000000

XGBoost

Following gives the Variable Importance based on XGBoost:-

	Overall <dbl>
Salary	100.0000000
Work.Exp	69.3767946
Age	39.4468041
Distance	33.1598307
license1	6.0618748
MBA1	4.5795407
Engineer1	0.1369253
Gender1	0.0000000

AdaBoost

Following are the variable importance based on Adaboost:-

	Importance <dbl>
Age	100.000000
Salary	97.837435
Work.Exp	96.187547
license	65.827740
Distance	64.895600
Gender	10.607755
Engineer	8.296048
MBA	0.000000

Following are the results of all models including Bagging & Boosting on test data:-

Model	**Sensitivity**	**Specificity**	**Accuracy**
GBM#1(Gradient Boost)	92.85%	95.79%	95.49%
GBM#2(Gradient Boost)	92.85%	95.79%	95.49%
GBM#3(Gradient Boost)	93.33%	96.61%	96.24%
XGB#1(XG Boost)	88.89%	98.26%	86.47%
ADB#1(Ada Boost)	77.8%	99.13	96.24
NB#1 (Naïve Bayes)	83.33%	98.26%	96.24%
KNN#2 (K Nearest Neighbor)	76.92%	99.09%	94.89%
LR#5 (Logistic Regression)	94.45%	95.65%	95.49%
RF#1 (Random Forest)	77.78%	98.2%	95.48%
RF#2 (Random Forest)	94.45%	98.26%	97.74%

Final Model Validation

Best models amongst all are

- Logistic Regression with Up Sampling
- Random Forest

As they are giving a high sensitivity which predicts the employees travelling by Car the best. Logistic regression would be a good model to leverage on account of the explanatory power of the variables that the model provides. Age, Distance, License & Salary are the variables that explain the best why employees choose car as the mode of transport.

Actionable Insights and Recommendations

The Company decision to invest in a parking space for its Employees will depend on the following criteria, as per the explanatory power of the variables provided by the above models: -

- Average Age of Employees
- The resident address of the Employees
- Average Salary of the Employees

Employees between 32 to 37 Years, travelling for a distance above 14km and having a salary above 17 Lakhs, are likely to use a car to travel to office. The number of Employees having these characteristics can be determined to decide on the need and size of parking space required.

The Company decision to invest in providing transportation services for its Employees will also depend on the same criteria, as per the explanatory power of the variables provided by above models. The Company will be able to decide on the span of the transportation services. Transportation services can be provided to the younger population along with those who belong to the lower salary band (less than 12 Lakhs) and travelling less than 14 km. This decision can serve as a perk to improve employee satisfaction.