Predict Employees' Preferred Mode of Transport

PROJECT REPORT

Ruchi Toshniwal | Machine Learning | August 12, 2019

Table of Contents

1. Project Objective

- Explore the dataset Cars.csv, that contains employee information about their mode of transport as well as their personal and professional details like age, salary, work experience etc.
- Build models to predict employees' preferred mode of commute to office.
- Determine which variables are significant predictors in determining the preferred mode of transport.
- Create multiple models (KNN, Naïve Bayes, Logistic Regression) and explore how each model performs using appropriate model performance metrics.
- Apply both bagging and boosting modeling procedures to create 2 models and compare their accuracy with the best model of the above step.
- Provide actionable insights and recommendations from the best model.

2. Assumptions

2.1 Logistic Regression

- Binary Logistic Regression requires the dependent variable to be binary/dichotomous.
- For Logistic Regression, predictor variables should be independent. There should be little or no multicollinearity among the predictor variables.
- Logistic regression assumes linear relationship between each of the independent variables and logit of the outcome.
- There should be no influential values (extreme values or outliers) in the continuous predictors.

2.2 Naïve Bayes

- Naïve Bayes works under the assumption that all the predictor variables are independent(probabilistically) of each other.

2.3 k Nearest Neighbour

- kNN is a non-parametric algorithm, it makes no assumptions on the distribution but it performs best when all independent variables are at a comparable scale.

3. Exploratory Data Analysis

3.1 Environment Setup and Data Import:

3.1.1 Install required packages and libraries

Please refer the accompanying file source_code_project_5.R for all the related code for this project.

Packages required

- Dplyr
- mice
- ggplot2
- gridExtra
- corrplot
- ppcor
- caTools
- DMwR
- car
- class
- e1071
- gbm
- xgboost
- caret
- ipred
- rpart
- ROCR
- ineq

3.1.2 Set up working directory

Set the working directory location to folder containing the data

3.1.3 Import the dataset

Use **read.csv** command to read the file Cars.csv into R.

3.2 Descriptive Statistics

3.2.1 Dimensions of the dataset and variable type

Use the str command to look at the dimensions and structure of the data.

Observations:

- Number of Rows in the dataset: 418
- Number of columns: 9
- Dependent Variable 'Transport' is a ternary variable. To make binary predictions for the employee mode of transport i.e. to predict if employee uses car or not, we will have to make a new categorical variable where, the value is 1 if employee uses car and 0 if employee uses other mode of transport.
- Number of Predictor Variables: 8
- Gender, Engineer, MBA, License and Transport are categorical variables.
- Age, Work.Exp, Salary and Distance are continuous variables
- Gender is a factor variable with levels – female and male.

3.2.2 Data Dictionary and feature type

Variable	*Description*	*Continuous/Categorical*
Age	Age of employee	Continuous
Gender	Male and Female	Categorical
Engineer	1 if employee is an engineer, 0 if not	Categorical
MBA	1 if employee is an MBA, 0 if not	Categorical
Work.Exp	Work Experience of the employee	Continuous
Salary	Salary of the employee	Continuous
Distance	Distance between employee home & office	Continuous
License	1 if employee has license, 0 if not	Categorical
Transport	2Wheeler, Car and Public Transport	Categorical

3.2.3 Check for Missing Values

- There is 1 missing value in the column MBA.

- Impute this missing value using the **mice** package, using method = 'logreg' since 'MBA' is a categorical variable and maxit = 50.
- The value imputed thus puts MBA = 0 in place of NA.
- After imputation, the data is complete and there are no missing values.

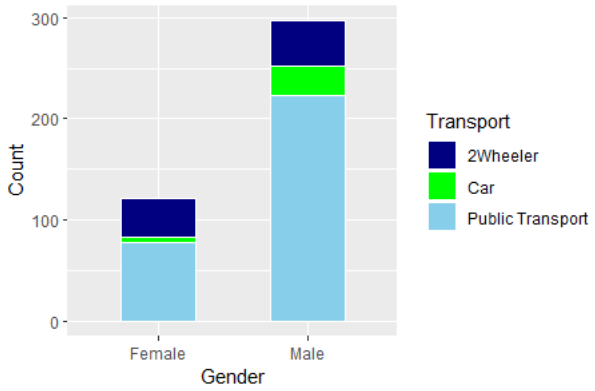
3.2.4 Analysing Continuous predictor variables

Variable	*Mean*	*Median*	*Min*	*Max*	*Range*	*1st Quartile*	*3rd Quartile*	*IQR*	*SD*
Age	27.3	27.0	18.0	43.0	25.0	25.0	29.0	4.0	4.2
Work.Exp	5.9	5.0	0.0	24.0	24.0	3.0	8.0	5.0	4.8
Salary	15.4	13.0	6.5	57.0	50.5	9.6	14.9	5.3	9.7
Distance	11.3	10.9	3.2	23.4	20.2	8.6	13.6	5.0	3.7

- Mean and median of all the variables are close. Mean is higher than the median in all the cases suggesting outliers towards the higher end.
- The max values are much higher than the 3rd quartile values for all variables. Strong possibility of outliers.
- Inter Quartile Range for all the variables is much lesser than half of their Range. This also alludes to the presence of outliers in the data.

3.2.5 Analysing Categorical predictor variables

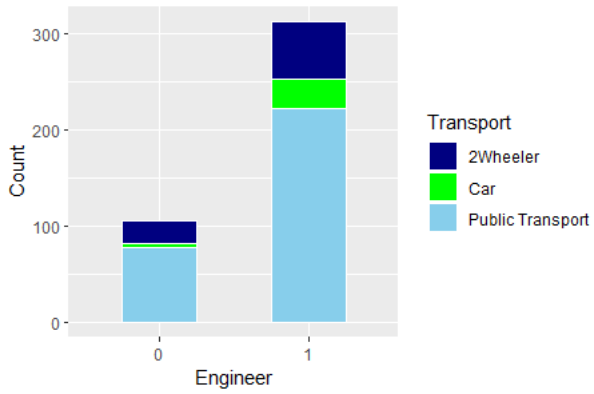
1. Gender



Observations:

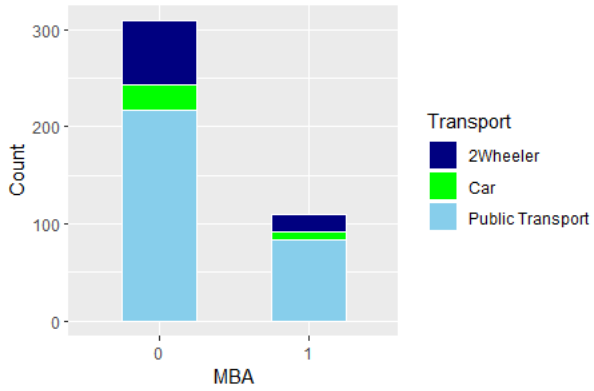
There are 121 females and 297 males in our data.

2. Engineer



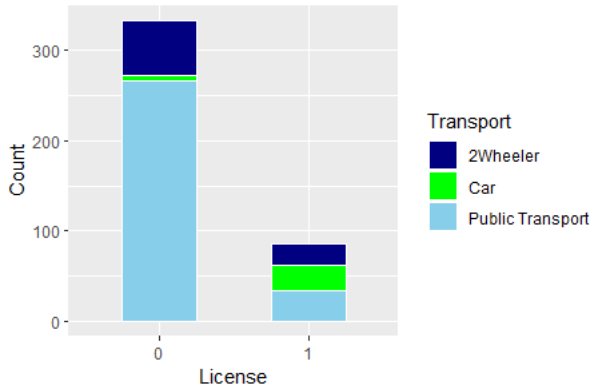
There are 313 Engineers in the data and 105 non-engineers.

3.MBA



There are 109 MBAs in the data and 309 non-MBAs.

4. License



85 people have a license while 333 do not have a license.

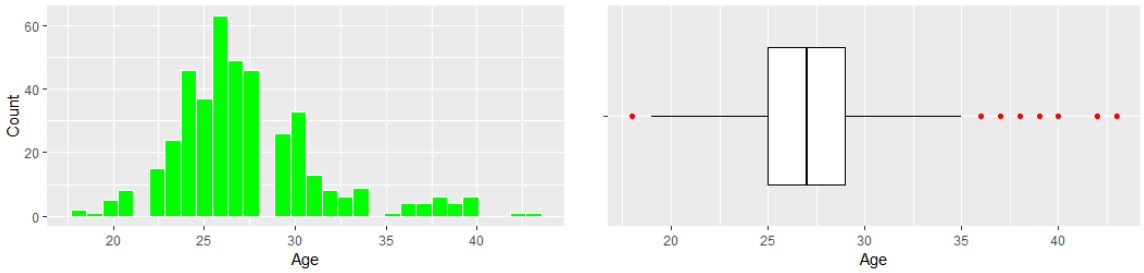
The chances of people with a license driving a car looks significantly higher than the ones without a license.

3.3 Data Visualisation

3.3.1 Univariate Analysis - Histogram and Boxplots of the variables

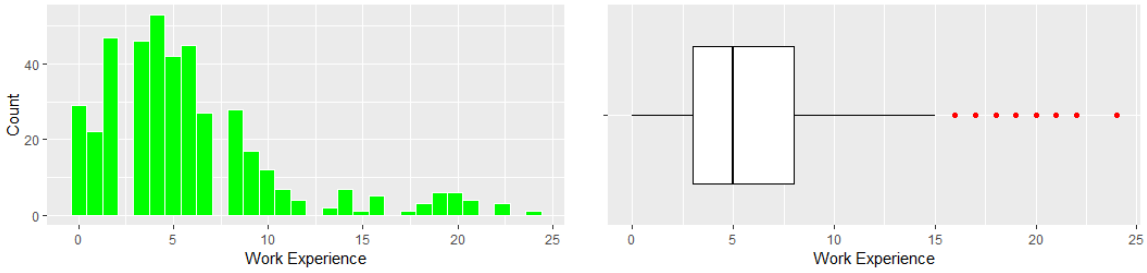
Let us have a look at the distribution and spread of the Continuous variables in our data

1. Age



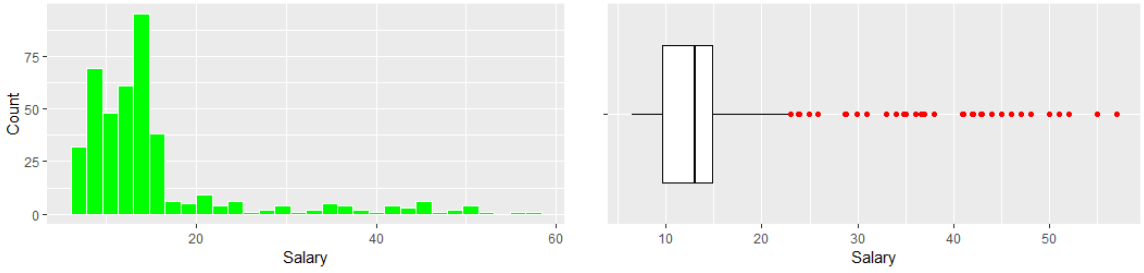
- The distribution is slightly left skewed and consequently there are a few outliers.
- These outliers are well within the expected values for employee age and there is nothing odd about them.

2. Work Experience



- This distribution is left skewed and there are outliers towards the higher end.
- These outliers too are well within the expected range. The senior employees in the company are generally have more work exp.

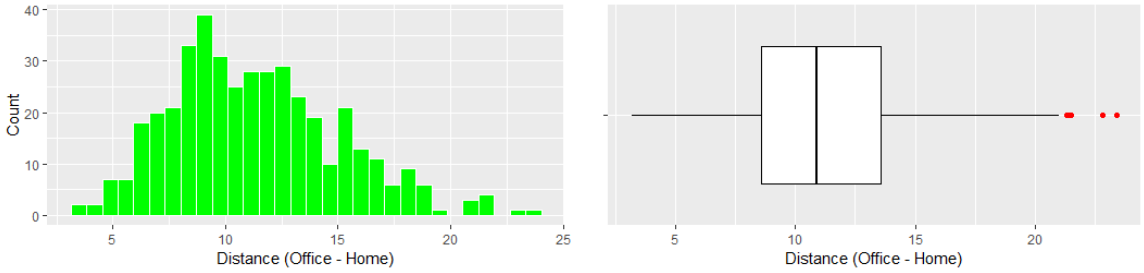
3. Salary



- The salary distribution is left skewed in most organisations and hence these outliers do not indicate anything unusual.

- None of the values is abruptly high or low.

4. Distance (Office - Home)



- The distribution of distance between employees' homes from the office is fairly normal.
- There are a few outliers, but they aren't abruptly high or low and hence do not require any action.

3.3.2 Outlier Treatment

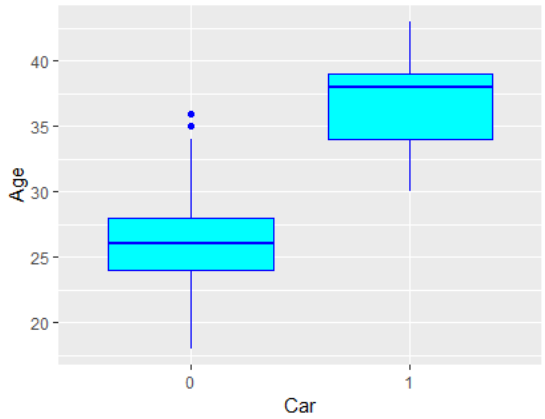
Although all of the continuous variables in our data have outliers, none of these values seem abruptly high or low. Therefore, at this point we are not going to treat these outliers. However, we will keep in mind, the presence of these outliers during our subsequent analysis.

3.3.3 Bivariate Analysis

Create a new dependent variable 'Car'. This would be a factor variable with 2 levels such that it is 1 if mode of transport is Car and 0 otherwise.

3.3.3.1 Relationship of dependent variable Cars (categorical) with other continuous variables

1. Age vs Car

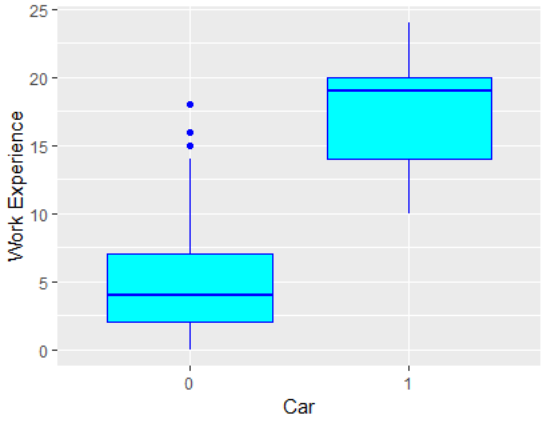


Observations:

The average age of people coming to office using a Car is significantly higher than those not using a car.

In fact, the lowest age of Car riders lies beyond the third quartile of non-Car riders.

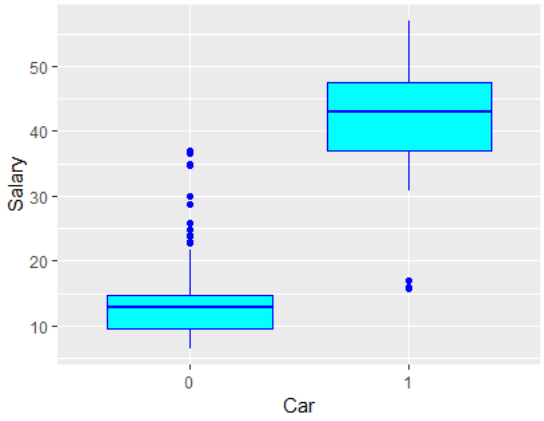
2. Work Experience vs Car



The average work experience of Car riders is nearly 15 years more than non–Car riders .

Higher the work experience, higher are the chances of travelling by car.

3. Salary vs Car

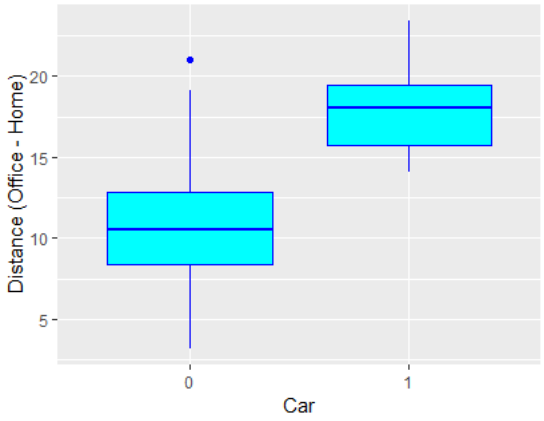


Similar story, the average salary of Car riders is much higher than the average salary of non-Car riders

There are some outliers where although the salary is on the lower end, they travel by car.

These 3 outliers all have 10 years of work exp and their distance from office is almost similar. It is possible that this is a group of friends who carpool.

4. Distance (Office - Home)



The average distance travelled by car riders is higher than the average distance travelled by non-car riders .

Observations:

1. More the age of the employee, more are their chances of using a Car for commute.
2. Higher the salary of the employee, higher their chances of using Car.
3. Higher the work exp of the employee, higher their chances of using Car.
4. Clearly there is multicollinearity in our data because usually older employees have higher work exp and higher salaries. It is possible that high positive correlation between all these variables and dependent variable indicates the same.
5. We need to get rid of the multi collinearity.
6. When distance between Office and home is higher, employees prefer Cars over Public Transport or 2 Wheelers.
7. We need to obtain t-statistics to see if these apparent relationships are statistically significant.

3.3.3.2 Relationship of dependent variable Car (categorical) with other categorical variables

Gender

Car	Female	Male
0	115	268
1	6	29

Engineer

Car	0	1
0	100	283
1	5	30

MBA

Car	0	1
0	283	100
1	26	9

		License	
		0	1
Car	0	327	56
	1	6	29

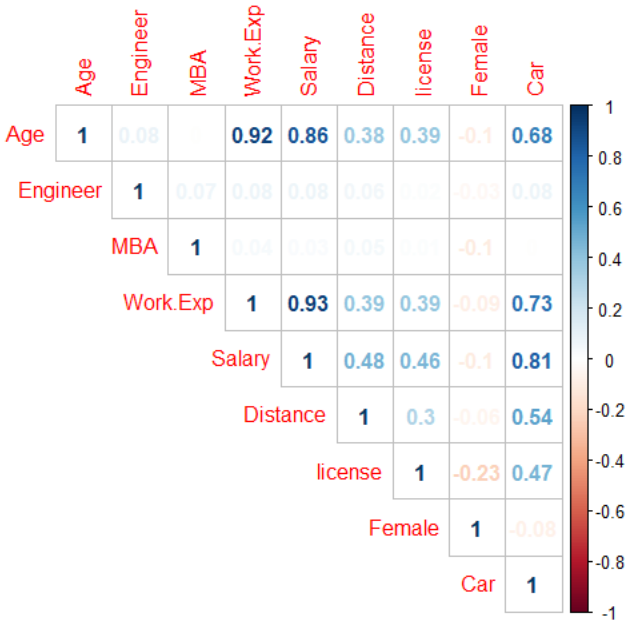
Observations:

- People with a license are more likely to come to office by car than the ones without license.
- Variables Gender, Engineer and MBA, on their own so not seem to have much impact on mode of transport.

3.3.3.3 Correlation Plot

Since our categorical variables are binary, we can use them along with our continuous variables to create a correlation plot.

Update the variable Gender to female such that that value 1 indicates a female and 0 indicates male.



Observations:

- Dependent Variable 'Car' has high correlations with Salary, Work.Exp, Age, Distance and license.
- But Age, Salary, Work.Exp, Distance and license have high correlations amongst themselves.
- Work.Exp and Salary are very highly correlated.
- There is a slightly negative correlation between Female and license.

3.3.3.4 Correlation Plot Significance

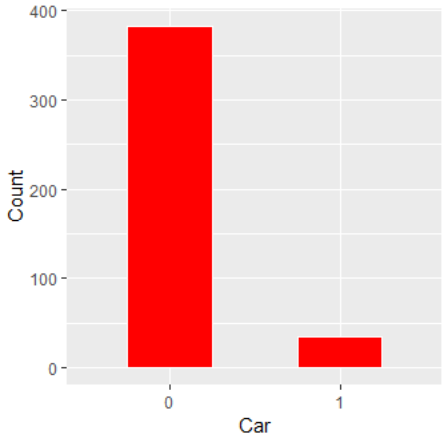
Let us look at the p-values of the t-statistics of the variables to understand the significance of the correlation

- High-Correlation between the following pairs is significant at 0.05 level - Car-Work.Exp, Car-Salary, Car-Distance, Car-license, Age-Work.Exp, Work.Exp-Salary, Work.Exp-Distance, Work.Exp-License, Salary -Distance, Salary -license.
- High-Correlation between the following pairs is NOT significant at 0.05 level - Car-Age, Car-Work.Exp, Age-Salary, Age-Distance, Age-license, Distance-license.
- The Pearson correlation coefficient is a measure of the strength of the linear relationship between two variables. The significant correlation coefficients >0.3 indicate that the variation of the 2 variables is correlated.
- High-Correlation of 'Car' with other variables is good for prediction purposes. However, collinearity between dependent variables needs to be treated.

3.3.4 Evidence for Multicollinearity

- Correlation plot and correlation significance values indicates the presence of collinearity between Age, Work-exp, Salary and license.
- Correlation Plots give us correlation between any 2 variables. To determine multicollinearity (linear relationship between more than 2 variables), we will use VIF values when building Logistic Regression model.
- Multicollinearity does not affect a kNN model

4. Data Preparation



The figure on the left indicates that our data is highly **imbalanced** with respect to the dependent variable 'Car'.

Car = 0 : 383

Car = 1 : 35

The minority class i.e. Car = 1 is only 8.4% of the data.

To fix this imbalance in the data, we will use the SMOTE technique.

Before using SMOTE, split the data into train and validation set (80:20)

4.1 Split the data into Train and Validation Set

Set Seed= 1000.

Split the data 80:20. Use the larger set as Train and smaller set of data as Validation Set.

The distribution of the dependent variable should be comparable in the original and the split sets.

	Number of Rows	*Number of Columns*	*Car == 1*	*Car == 0*	*Minority Class %*
Original	418	9	383	35	8.37%
Train Set	334	9	306	28	8.38%
Validation Set	84	9	77	7	8.33%

4.2 Balance Train Data using SMOTE

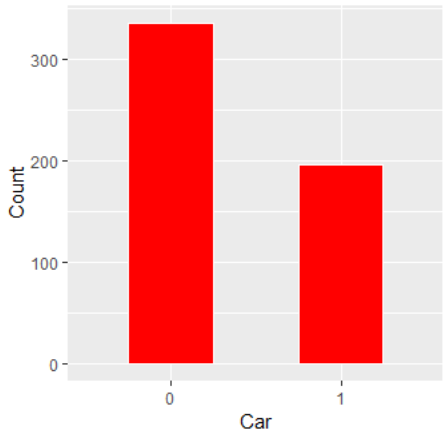
- Use SMOTE function of the DMwR package
- The dependent variable should be a factor.
- Take perc.over = 600 and perc.under = 200
- Before using SMOTE the train sample distribution was

Car = 0	**Car = 1**
306	28

- After SMOTE the (balanced) train sample looks like the following distribution

Car = 0	**Car = 1**
336	196

- The minority sample proportion has been synthetically improved, using k = 5. That is, 168 new samples are generated using 5 nearest neighbours.



5. Predictive Modelling

- Predicting if an employee commutes using car or not is a classification problem; hence classification models must be built.
- Our dependent variable 'Car' is a binary variable.
- We will use the balanced data created using SMOTE for training.
- For validation, the original validation set created via splitting will be used.
- Use Logistic Regression, kNN and Naïve Bayes modelling techniques to build predictive models, to predict customer churn.
- Use bagging and boosting to see if they improve the predictions.
- Use Model Performance Metrics to compare the three models' performance on the validation set

5.1 Logistic Regression Model

5.2.1 Treating Multicollinearity

- Start building a logistic model with all variables and gradually drop the variables that cause multicollinearity or are insignificant predictors. (Refer code for details.)
- Check the VIF scores of the model.
- Drop the variable with high VIF (>5). Use judgement and previous analysis before dropping any variable.
- Repeat the first 2 steps till we obtain a model in which all variables have low VIF values (<5).
- Remove the variables that are insignificant.
- We drop Work.Exp because it has high VIF. After it is dropped, the VIF of new model variables is in acceptable range. The model is now free of multicollinearity.

5.2.2 Model Optimization

- MBA is insignificant variable for our logistic model; hence we drop it to improve model AIC.
- Dropping them improves AIC from 71.5 to 70.5.

5.2.3 Model

```
Call:
glm(formula = Car ~ ., family = "binomial", data = logit.train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.90614  -0.01255  -0.00094   0.00553   2.44447

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -43.14864    8.24781  -5.232 1.68e-07 ***
Age           0.62755    0.19472   3.223  0.00127 **
Engineer      1.85010    0.91783   2.016  0.04383 *
Salary        0.15746    0.05413   2.909  0.00362 **
Distance      1.05348    0.21573   4.883 1.04e-06 ***
license       2.18522    0.84051   2.600  0.00933 **
Female        2.71206    0.88699   3.058  0.00223 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 700.229  on 531  degrees of freedom
Residual deviance:  56.509  on 525  degrees of freedom
AIC: 70.509

Number of Fisher Scoring iterations: 10
```

5.2.4 Model Interpretation

- Age, Engineer, Salary, Distance, license and Female are significant in predicting employee mode of commute.
- Log odds of using Car, while all the other variables are 0 is -5.901214
- Coefficient of Age = 0.62755. Odds Ratio is e^0.62755 = 1.873
- Coefficient of Engineer = 1.8501. Odds Ratio is e^1.8501 = 6.36
- Coefficient of Salary = 0.15746. Odds Ratio is e^0.15746 = 1.17
- Coefficient of Distance = 1.05348. Odds Ratio is e^1.05348 = 2.8676
- Coefficient of license = 2.18522. Odds Ratio is e^2.18522 = 8.8926
- Coefficient of Female = 2.71206. Odds Ratio is e^2.71206 = 15.06
- Odds of driving a Car to the office increase by 87% as age increases by 1 year.
- Odds of driving a Car are higher for an engineer than a non-engineer.
- Odds of driving a Car increase by 17% with every one unit increase in Salary.
- Odds of driving a Car increase by 180% with every one unit increase in Distance.
- Odds of driving a Car increase if you have a license.

- Odds of driving a Car increase if you are a female.

5.2.5 Model Acceptability

- The **accuracy** of our LR model on in sample data is 97.7% and on the validation set is **98.8%**
- The accuracy of both train and validation is within 10%. This is an acceptable model and does not overfit the train data.
- Logistic Regression model predicts Mode of transport with high Accuracy.

Confusion Matrix – Training data

		Predicted	
		0	1
Actual	0	**330**	**6**
	1	**6**	**190**

Accuracy = 97.7%
Sensitivity = 0.97
Precision = 0.97

Confusion Matrix – Validation data

		Predicted	
		0	1
Actual	0	**76**	**1**
	1	**0**	**7**

Accuracy = 98.8%
Sensitivity = 1
Precision = 0.875

5.3 k-Nearest Neighbours Model

5.3.1 Variable Normalization

- Variable scaling is very important in distance-based algorithms like kNN, to make sure that variables with higher numeric values do not bias the model against variables with lower numeric values.
- All the variables must be brought to a similar scale/normalised.

5.3.2 Model Optimization

- We use different values of k to arrive at a model that gives the best results in terms of model accuracy.
- Luckily, k=3 gives us very good results on the validation data. We will use k=3 for our kNN model.

5.3.3 Model Interpretation

- A kNN model cannot provide us much insight into the data, it can only help us with final predictions.
- Model accuracy on validation data is 100%.
- kNN model predicts Mode of transport with high Accuracy.

Confusion Matrix – Validation data

		Predicted	
		0	1
Actual	0	77	**0**
	1	**0**	7

Accuracy = 100%

Sensitivity = 1

Precision = 1

5.4 Naïve Bayes Model

5.4.1 Variable Transformation

- Naïve Bayes algorithm needs the dependent variable to be converted to a factor variable.

5.4.2 Model Interpretation

- The dependent variable in our data is a categorical variable with 2 levels. All our predictor variables are numeric. A Naïve Bayes algorithm is applicable in such a case. However, it is not very easy derive insights from such a model.
- Naïve Bayes model predicts Mode of transport with high Accuracy.

5.4.3 Model Acceptability

- The accuracy of our NB model on in-sample data is 99% and on the validation set is 100%
- The accuracy of both samples is within 10%. This is an acceptable model and does not overfit the train data.
- Naïve Bayes is applicable here because the dependent variable is a categorical variable. Naïve Bayes cannot provide us much insight into the data and predictor variables.

Confusion Matrix – Training data

		Predicted	
		0	1
Actual	0	**330**	**6**
	1	**1**	**195**

Accuracy = 98.7%

Sensitivity = 0.995

Precision = 0.97

Confusion Matrix – Validation data

		Predicted	
		0	1
Actual	0	**77**	**0**
	1	**0**	**7**

Accuracy = 100%
Sensitivity = 1
Precision = 1

6. Bagging and Boosting ensemble models

Ensemble modelling techniques like bagging and boosting help to make better predictions for imbalanced datasets.

6.1 Bagging

Confusion Matrix – Training data

		Predicted	
		0	1
Actual	0	**306**	**0**
	1	**2**	**26**

Accuracy = 99.7%
Sensitivity = 0.964
Precision = 1

Confusion Matrix – Validation data

		Predicted	
		0	1
Actual	0	**77**	**0**
	1	**0**	**7**

Accuracy = 100%
Sensitivity = 1
Precision = 1

6.1.1 Model Interpretation

- We use decision tress for our bagging algorithm.
- Bagging models do not provide insights into the data.
- The predictions provided by bagging algorithms are accurate even with unbalanced data.
- We do not have to use synthetic sampling techniques like SMOTE when using bagging. We can use the original unbalanced data. Bagging does sampling internally and creates many random samples where the minority class gets decent representation.
- Bagging model predicts Mode of transport with high Accuracy.

6.1.2 Model Acceptability

- The accuracy of our Bagging model on in-sample data is 99% and on the validation set is 100%
- The accuracy of both samples is within 10%. This is an acceptable model and does not overfit the training data.

6.2 Boosting

- We will use xgboost() as or boosting algorithm.
- Xgboost works with matrices
- All variables must be numeric for xgboost
- We also need to separate training data and the dependent variable

Confusion Matrix – Training data

		Predicted	
		0	1
Actual	0	**306**	**0**
	1	**3**	**25**

Accuracy = 99.1%
Sensitivity = 0.893
Precision = 1

Confusion Matrix – Validation data

		Predicted	
		0	1
Actual	0	77	0
	1	0	7

Accuracy = 100%

Sensitivity = 1

Precision = 1

6.2.1 Model Interpretation

- We use the xgboost algorithm for boosting our prediction model.
- Boosting algorithms do not provide much insight into the data and variable significance.
- Although balancing the data helps the boosting model, even without balancing the performance of the boosting algorithm is better than most other algorithms.
- Boosting model predicts Mode of transport with high Accuracy.

6.2.2 Model Acceptability

- The accuracy of our xgboost model on in-sample data is 99% and on the validation set is 100%
- The accuracy of both samples is within 10%. This is an acceptable model and does not overfit the training data.

7. Model Comparison

Following is the result we get by running our 3 models on the Validation (out of sample) dataset.

Model Name	Accuracy	Sensitivity	Specificity	Precision	AUC
Logit Train	0.977	0.969	0.982	0.969	0.976
Logit Test	0.988	1	0.987	0.875	0.994
kNN Test	1	1	1	1	1
Naïve Bayes Train	0.987	0.995	0.982	0.970	0.989
Naïve Bayes Test	1	1	1	1	1
Bagging Train	0.997	0.964	1	1	0.982
Bagging Test	1	1	1	1	1
Boosting Train	0.991	0.893	1	1	0.998
Boosting Test	1	1	1	1	1
Baseline model (Car =0 for all)	0.916	0	1	-	-

7.1 Key Observations

- All the models predict mode of transport with high accuracy.
- Generally, boosting and bagging give us better predictors. Especially when the data is unbalanced. In our present case however, kNN and Naïve Bayes with SMOTE and Bagging and Boosting without SMOTE, all give us high accuracy and sensitivity values.
- Except logistic regression, all the other models have perfect predictions on validation data.
- More than Accuracy, here we should focus on Sensitivity and Precision when we encounter such unbalanced data.
- All the algorithms are able to detect the 7 observations were Car = 1.
- Logistic regression makes a mistake when it picks 1 false positive.
- According to these metrics, we can use any of these algorithms in production. However, only Logistic Regression gives us insight into the data.

- Key variables that help to predict if an employee travels in Car are
    - Distance, Age, Salary, Gender and License

8. Interpretations and Recommendations from the models

- If the sample is a true representation of the population, fewer employees commute using a car, most prefer other means of transport.
- Distance of travel is a major factor in deciding the mode of transport. If the distance is higher (>14), employees are more likely to use cars.
- Another very important factor is salary. Only people with higher salaries (>30) travel by car.
- One small exception to the above criteria is gender. Females are more likely to travel by car than males even when the salary is mid-range.
- Older employees are more likely to travel by car.
  (Since age of the employees and their salaries and work experience have high positive correlation, we can focus on either one of these variables at a time.)