CARS PROJECT

By: Omar Ossama

Contents

1. EDA

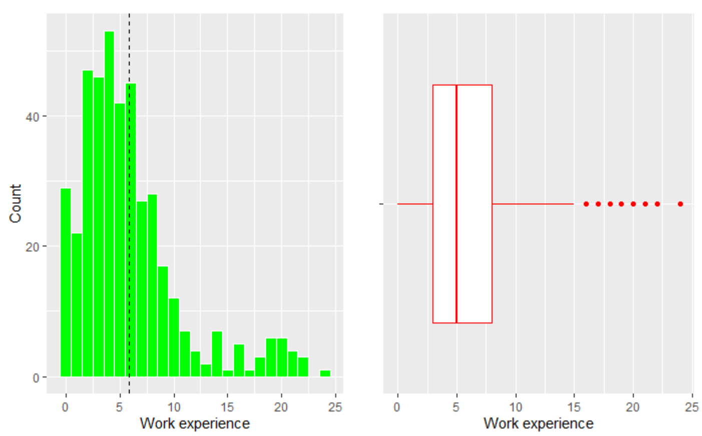
1.1 Univariate Analysis:

1.1.1. Age:

Minimum age is 18.
Maximum age is 43.
Average age is 27.33.
Outliers are present above and below the upper and lower limits.
Data slightly skewed to the right.

1.1.2. Work Experience:

Some employees have no work experience.
Maximum work experience is 24 years.
Average is 5.873 years.
Outliers present above the upper limit.
Data skewed to the right.

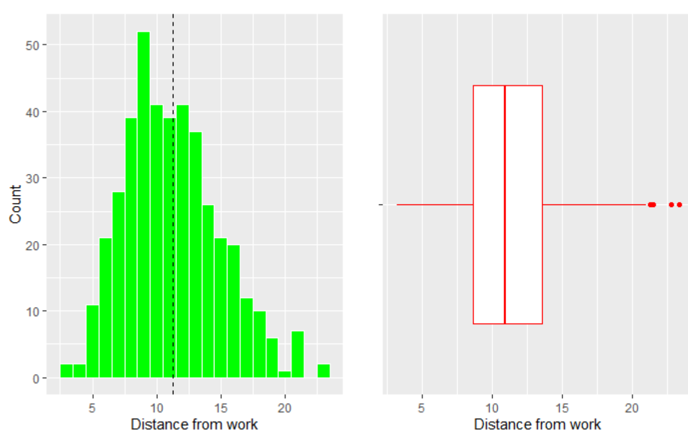


1.1.3. Salary:

Minimum salary is $6,500/year.
Maximum salary is $57,000/year.
Average is $15,418/year.
Outliers present above the upper limit.
Data skewed to the right.

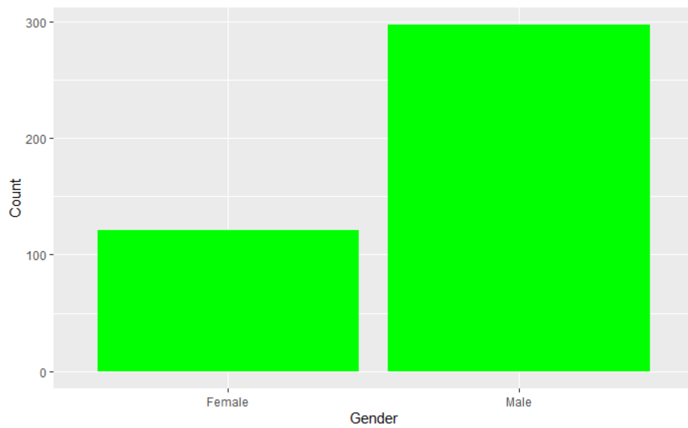
1.1.4. Distance from work:

Minimum distance is 3.2 KM.
Maximum distance is 23.40 KM.
Average is 11.29 KM.
Outliers present above the upper limit.
Data skewed to the right.



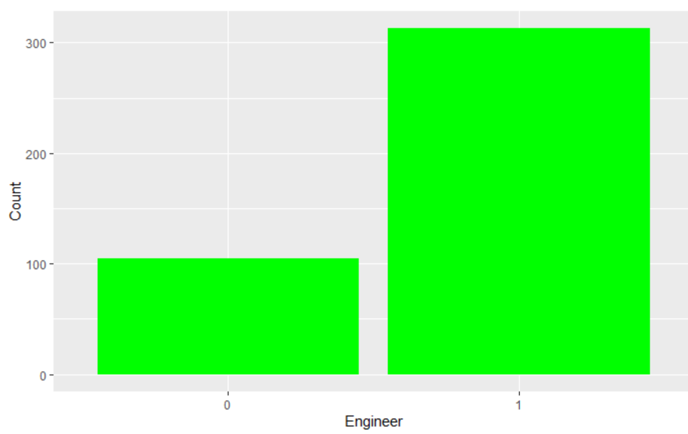
1.1.5. Gender:

Employees are mostly male with 71% while females make up the remaining 29%.



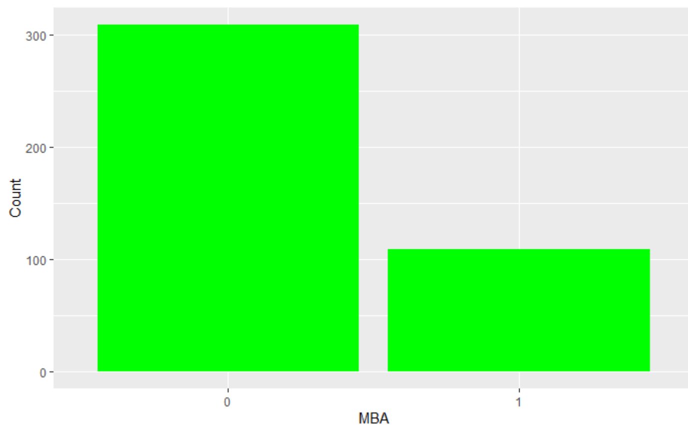
1.1.6. Engineering degrees:

75% of the employees have an engineering degree while the remaining 25% do not.



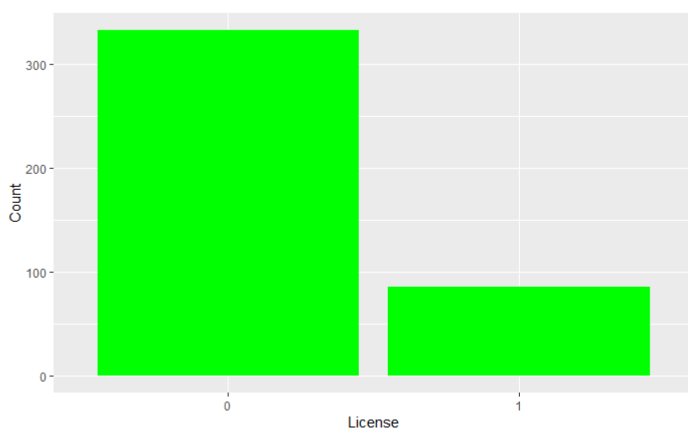
1.1.7. MBA degrees:

74% of the employees do not have an MBA degree. 26% do.



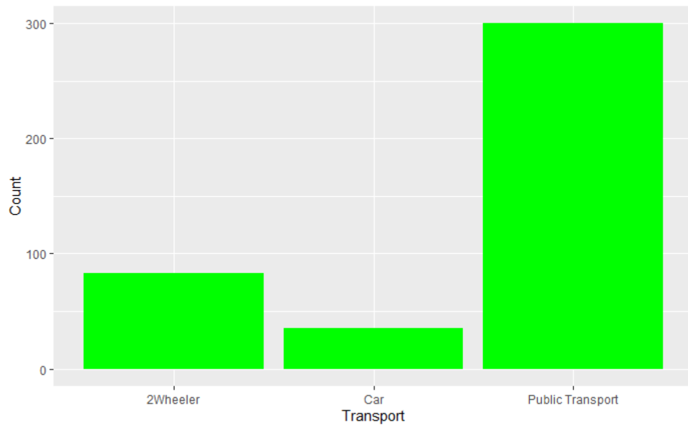
1.1.8. License:

With about 80%, most of the employees do not have a license.



1.1.9. Mean of transport:

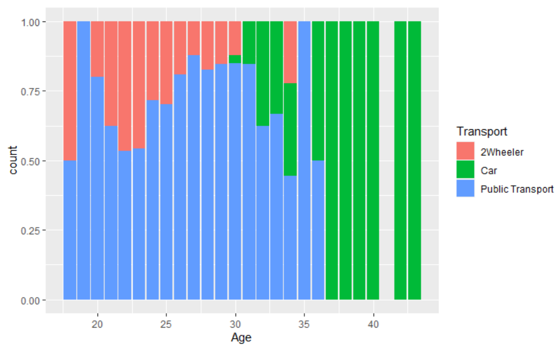
Most of the employees prefer taking public transport to go to work. About 20% use a 2 wheeler. With only 8% preferring to use their cars.

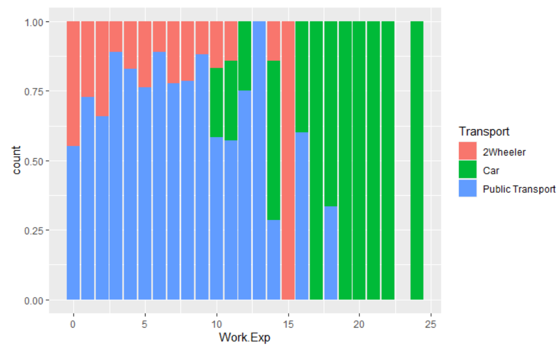


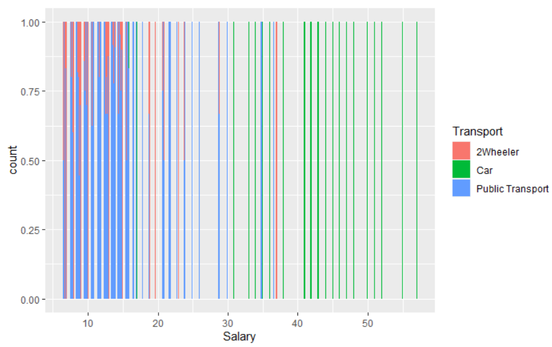
1.2. Bivariate Analysis:

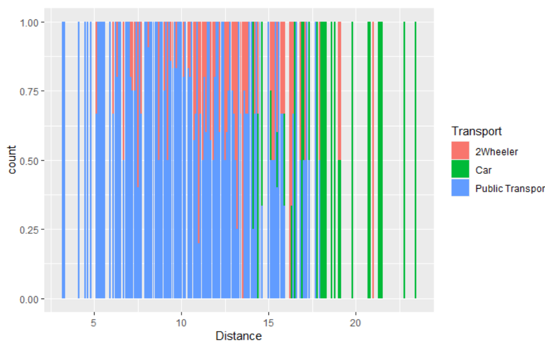
1.2.1. Correlation between method of transport (the dependent variable) and the other variables:

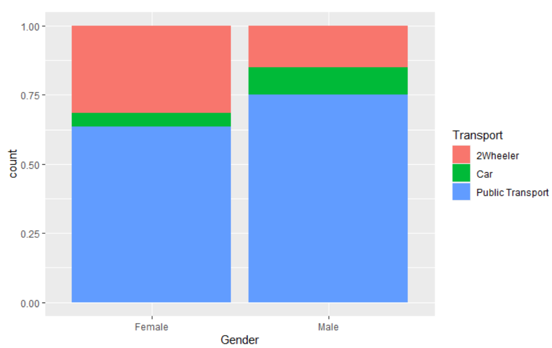
All the variables appear to be correlated with the “Transport” variable except for the “Engineer” variable.

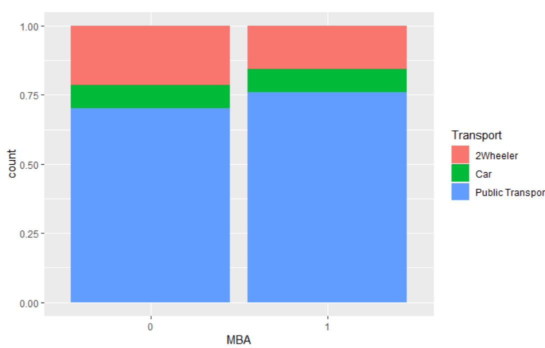


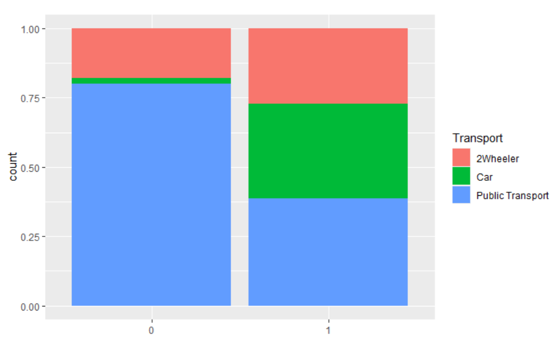


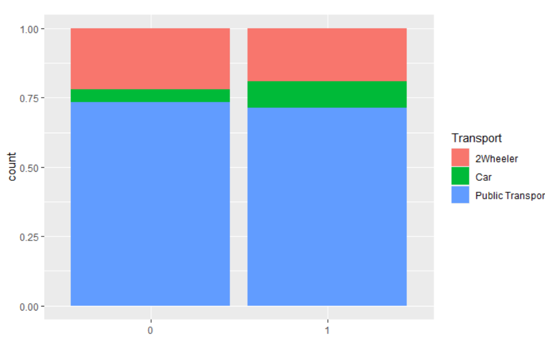






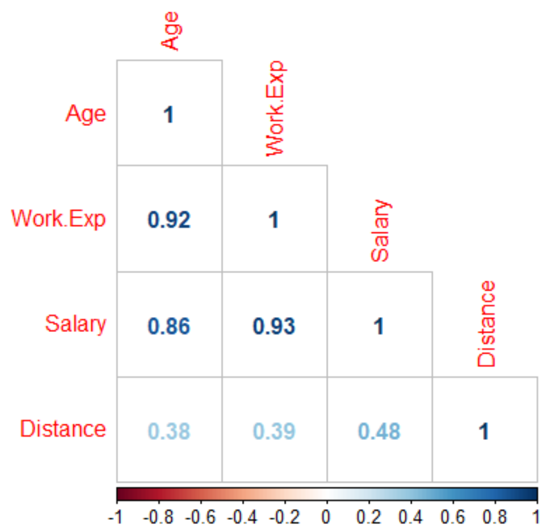






1.2.2. Correlation between numerical variables

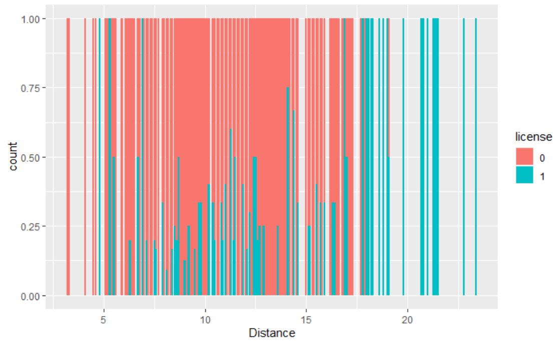
Age, Work Exp and Salary variables have a high correlation. Distance variable is not correlated with the others

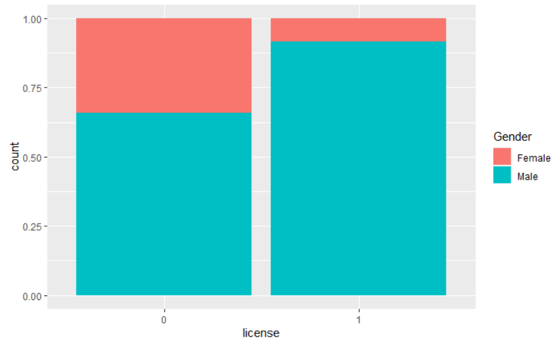


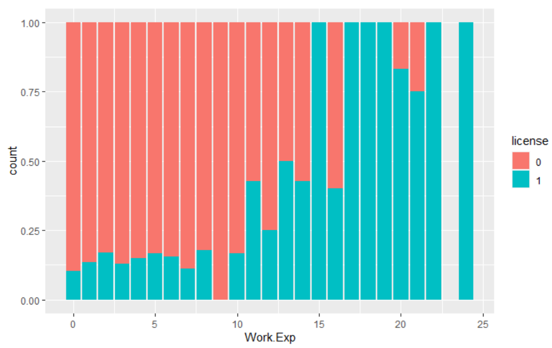
1.2.3. Important correlations between the remaining variables

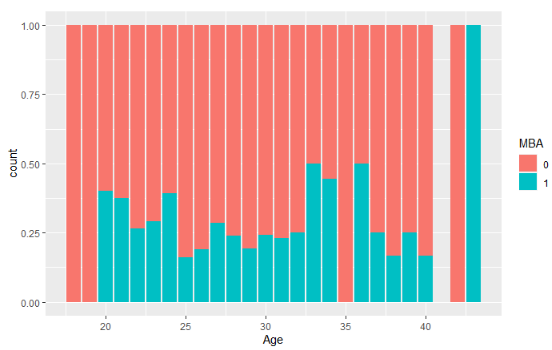
License variable appear to have a significant correlation with Distance, Gender, Work Experience and Age variables.

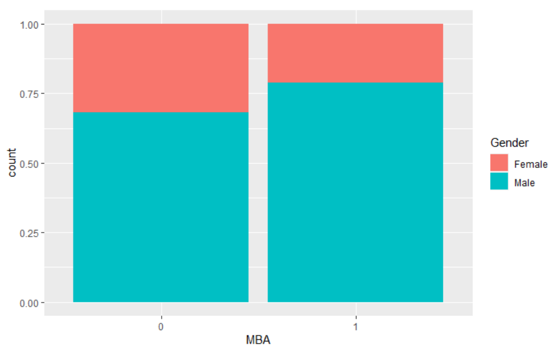
While MBA variable seems to have a correlation with the Gender variable, it does not appear to have a correlation with the Age variable.











1.3. The most challenging aspects and methods used to deal with them

1) Dataset is highly imbalanced. SMOTE will be used to artificially oversample the minority
2) A new dependent variable will be created with only 2 levels, employees who use a car and those who do not. As the company is only interested in those who arrive by car.
3) Outliers are present in the numerical variables. So, the values will be capped.
4) Conversion of the Gender variable from a factor to a numerical variable by converting its values to 1s for males and 0s for females.

2. Modeling

2.1. Logistic regression model performance (Without SMOTE)

After removing variables with multicollinearity and insignificant variables.

2.1.1. Against train data:

Accuracy: 98.63%
Sensitivity: 99.25%
Specificity: 91.67%

2.1.2. Against test data:

Accuracy: 97.62%
Sensitivity: 97.39%
Specificity: 100%

2.2. Logistic regression model performance (With SMOTE)

2.2.1. Against train data:

Accuracy: 99.37%
Sensitivity: 99.23%
Specificity: 99.65%

2.2.2. Against test data:

Accuracy: 96.83%
Sensitivity: 96.52%
Specificity: 100%

2.3. Naïve Bayes model performance (Without SMOTE)

2.3.1. Against train data:

Accuracy: 99.32%
Sensitivity: 99.63%
Specificity: 95.83%

2.3.2. Against test data:

Accuracy: 95.24%
Sensitivity: 94.78%
Specificity: 100%

2.4. Naïve Bayes model performance (With SMOTE)

2.4.1. Against train data:

Accuracy: 98.11%
Sensitivity: 99.76%
Specificity: 94.91%

2.4.2. Against test data:

Accuracy: 96.83%
Sensitivity: 96.52%
Specificity: 100%

2.5. KNN model performance (Without SMOTE)

2.5.1. Against train data:

Accuracy: 97.95%
Sensitivity: 98.88%
Specificity: 87.50%

2.5.2. Against test data:

Accuracy: 96.03%
Sensitivity: 98.26%
Specificity: 72.73%

2.6. KNN model performance (With SMOTE)

2.6.1. Against train data:

Accuracy: 99.84%
Sensitivity: 99.76%
Specificity: 100%

2.6.2. Against test data:

Accuracy: 96.03%
Sensitivity: 97.39%
Specificity: 81.82%

2.7. Applying Random Forrest as a bagging technique (Without SMOTE)

2.7.1. Against train data:

Accuracy: 100%
Sensitivity: 100%
Specificity: 100%

2.7.2. Against test data:

Accuracy: 97.62%
Sensitivity: 97.39%
Specificity: 100%

2.8. Random Forrest (With SMOTE)

2.8.1. Against train data:

Accuracy: 100%

Sensitivity: 100%

Specificity: 100%

2.8.2. Against test data:

Accuracy: 100%

Sensitivity: 100%

Specificity: 100%

2.9. Applying Extreme Gradient Boosting as a boosting technique (Without SMOTE)

2.9.1. Against train data:

Accuracy: 100%

Sensitivity: 100%

Specificity: 100%

2.9.2. Against test data:

Accuracy: 98.41%

Sensitivity: 98.26%

Specificity: 100%

2.10. Extreme Gradient Boosting (With SMOTE)

2.10.1. Against train data:

Accuracy: 100%

Sensitivity: 100%

Specificity: 100%

2.10.2. Against test data:

Accuracy: 100%

Sensitivity: 100%

Specificity: 100%

3. Conclusion

From the results we can see that:

1) With a 100% accuracy, sensitivity and specificity, the Extreme Gradient Boost and the Random Forrest models preformed perfectly after applying SMOTE.
2) Logistic regression and Naïve-Bayes models performed very well. But were worse than the Extreme Gradient Boost and Random Forrest models.
3) KNN model had a bad performance even after using SMOTE.
4) Most of the variables have a correlation with the Transport variable. But Age and Distance were found to be the best predictors.