**Project – Cars**

**Table of Content**

|  |  |  |
| --- | --- | --- |
| **S.No.** | **Content** | **Page No.** |
| 1. | Objective | 3 |
| 2. | Exploratory Data Analysis | 3 |
|  | 2.1 Basic Data Summary | 3 |
|  | 2.2 Bivariate Analysis | 4 |
|  | 2.3 Summary of the Data Set | 4 |
|  | 2.4 Multi co-linearity | 6 |
|  | 2.5 Insights from Exploratory Data Analysis | 8 |
| 3. | Logistic Regression Model Building | 8 |
|  | 3.1 Interpretation from Logistic Regression Model | 8 |
|  | 3.2 KNN Model | 9 |
|  | 3.2.1 Interpretation from KNN model | 9 |
|  | 3.3 Naïve Bayes Model | 10 |
|  | 3.3.1 Interpretation from Naïve Bayes Model | 10 |
|  | 3.4 Actionable Insights and Recommendations | 11 |

1. **Objective**

The objective of this project work is to identify which mode of transport people prefer to commute to their office. This report work predicts whether the people will prefer to travel via car or not.

1. **Exploratory Data Analysis**
   1. **Basic Data Summary**

The employee data is given in the Cars.csv file. Data provided relates the employee transport with other attributes such as Age, Gender, Engineer, MBA, Work Exp., Salary, Distance and license. Basic structure of the data file is shown in the Table-1 below:

**Table-1: Basic data summary**

'data.frame': 418 obs. of 9 variables:

$ Age : int 28 24 27 25 25 21 23 23 24 28 ...

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

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

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

$ Work.Exp : int 5 6 9 1 3 3 3 0 4 6 ...

$ Salary : num 14.4 10.6 15.5 7.6 9.6 9.5 11.7 6.5 8.5 13.7 ...

$ Distance : num 5.1 6.1 6.1 6.3 6.7 7.1 7.2 7.3 7.5 7.5 ...

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

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

**2.2 Bivariate Analysis**

The data represents the Bivariate Analysis as the dependent variable (i.e. Transport) depends upon two or more variables.

**2.3 Summary of the data set**

**Table-2: Summary**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Age** | **Engineer** | **MBA** | **Work Exp.** | **Salary** | **Distance** | **License** |
| **Min** | 18.00 | 0.0000 | 0.0000 | 0.000 | 6.50 | 3.2 | 0.0000 |
| **1st Qu** | 25.00 | 1.0000 | 0.0000 | 3.000 | 9.60 | 8.6 | 0.0000 |
| **Median** | 27.00 | 1.0000 | 0.0000 | 5.000 | 13.00 | 10.9 | 0.0000 |
| **Mean** | 27.33 | 0.7506 | 0.2614 | 5.873 | 15.42 | 11.3 | 0.2038 |
| **3rd Qu** | 29.00 | 1.0000 | 1.0000 | 8.000 | 14.90 | 13.6 | 0.0000 |
| **Max** | 43.00 | 1.0000 | 1.0000 | 24.000 | 57.00 | 23.4 | 1.0000 |

**Gender:** Female = 120 Male = 297

**Transport:** 2Wheeler = 83 cars = 35 Public Transport = 299

1. **Checking and Treating multi co-linearity**

From the exploratory data analysis, it is observed that there exist multi co-linearity between variables Age, Work Exp. and Salary.

Multi co-linearity between Age and Work Exp. is found to be 92%

Multi co-linearity between Age and Salary is found to be 86%

Multi co-linearity between Salary and Work Exp. is found to be 93%

The correlation between different variables can be observed from the Fig.1 shown below:

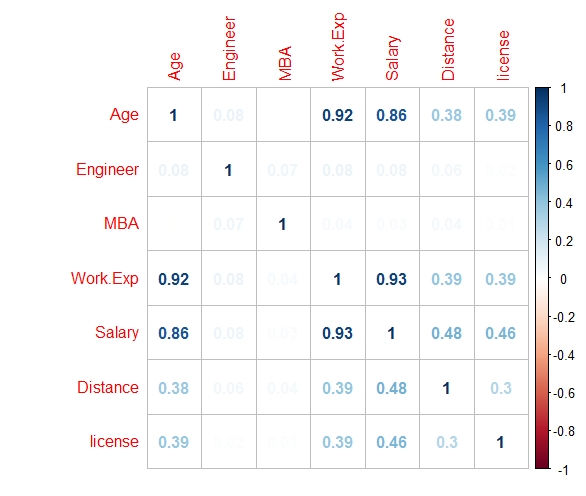


Fig.1: Correlation plot

**Treatment of Multi co-linearity:**

From Fig.1 it is evident that there exist a strong correlation between the variables Age, Work Exp. and Salary. Therefore in order to remove or treat the problem of multi co-linearity, the concept of dimensionality reduction is utilized. Variables Age, Work Exp. and Salary are grouped together into a single variable named as “Employee”.

Factor analysis method is utilized to group the variables and the results obtained are tabulated in Table-4 and Table-5 respectively.

**Table-3: Eigen Value**

|  |  |  |
| --- | --- | --- |
| 2.8099 | 0.1423 | 0.0478 |

**Table-4: Factor analysis solution without rotation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **PA1** | **h2** | **u2** | **com** |
| **Age** | 0.92 | 0.85 | 0.1450 | 1 |
| **Work Exp.** | 1.00 | 0.99 | 0.0054 | 1 |
| **Salary** | 0.93 | 0.87 | 0.1311 | 1 |

**Table-5: Factor analysis solution with rotation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **PA1** | **h2** | **u2** | **com** |
| **Age** | 0.92 | 0.85 | 0.1450 | 1 |
| **Work Exp.** | 1.00 | 0.99 | 0.0054 | 1 |
| **Salary** | 0.93 | 0.87 | 0.1311 | 1 |

* 1. **Insights from exploratory data analysis**
* Data set contains the categorical and continuous data.
* No oultiers exists in the data set.
* Existence of NA value is found.
* The data set represents the bivariate analysis as the dependent variable (Transport) depends on two or more variables.
* There is existence of multi co-linearity between the variables Age, Work Exp. and Salary.
* Correlation between Age and Work Exp. is 92%.
* Correlation between Age and Salary is 86%.
* Correlation between Work Exp. and Salary is 93%.
* Factor analysis method is employed for the treatment of multi co-linearity.

1. **Logistic Regression Model**

**Interpretation from Logistic Regression Model**

* Logistic regression (LR) model was build with considering Transport as the dependent variable and all other variables Age, Gender, Engineer, MBA, Work Exp., Salary, Distance and license as the independent variables.
* Following observations were made from the outcome of LR model

Accuracy = 70.67%

Sensitivity = 56.03%

Specificity = 12.35%

1. **KNN Model**

**Interpretation from KNN Model**

* Partitioning the data set into train and test data using the 70:30 rule. Which means 70% of the data is used to build the model and the remaining 30% of the data is to test the performance of the build model.
* Model performance was observed from the confusion matrix corresponding.
* Accuracy = 70.85%
* Sensitivity = 54.16%
* Specificity = 14.20%

1. **Naïve Bayes model**

**Interpretation from Naïve Bayes Model**

* Following observations are made from the Naïve Bayes Model

Accuracy = 70.95%

Sensitivity = 57.39%

Specificity = 15.47%

1. **Bagging**

**Interpretation from Bagging Model**

* Following observations are made from the Naïve Bayes Model

Accuracy = 78.77%

Sensitivity = 68.01%

Specificity = 10.47%

1. **Adaptive Boosting**

**Interpretation from Adaptive Boosting Model**

* Following observations are made from the Naïve Bayes Model

Accuracy = 78.86%

Sensitivity = 68.34%

Specificity = 12.47%

1. **XGBoosting**

**Interpretation from XGBoosting Model**

* Following observations are made from the Naïve Bayes Model

Accuracy = 78.87%

Sensitivity = 68.39%

Specificity = 11.27%

1. **SMOTE**

**Interpretation from SMOTE Model**

* Following observations are made from the Naïve Bayes Model

Accuracy = 78.24%

Sensitivity = 69.12%

Specificity = 15.47%

Comparison of all the model together is listed in the Table-6

**Table-6: Comparison of different models**

|  |  |  |  |
| --- | --- | --- | --- |
| **Prediction Summary** | | | |
| **Techniques** | **Accuracy** | **Sensitivity** | **Specificity** |
| Logistic Regression | 70.67% | 56.03% | 12.35% |
| KNN | 70.85% | 54.16% | 14.20% |
| Naïve Bayes | 70.95% | 57.39% | 15.47% |
| Bagging | 78.77% | 68.01% | 10.47% |
| Adaptive Boosting | 78.86% | 68.34% | 12.47% |
| XGBoosting | 78.87% | 68.39% | 11.27% |
| XGBoosting1 | 78.87% | 68.38% | 11.25% |
| XGBoosting2 | 78.87% | 68.38% | 11.23% |
| SMOTE | 78.24% | 69.12% | 15.47% |

1. **Actionable Insights**

* From the above build models performances as listed in Table-6, it is found that the XGBoosting model gives much better result.
* In the predictions of employees travelling via Car as mode of transport, the deciding variables are Age, Work Exp. and Salary.
* It is most likely that an employee with higher experience, age and salary would prefer to travel by car.