**Transport**

**Analysis**

Name: Saurabh Dharmadhikari

PGP-DSBA Online January’ 21

Date: 26/10/2022

Table of Contents:

Executive Summary……………………...................................................................................................................................3

Introduction ...................................................................................................................................................................3

Data Description ............................................................................................................................................................3

Questions:

1. Basic data summary, Univariate, Bivariate analysis, graphs, checking correlations, outliers and

missing values treatment (if necessary) and check the basic descriptive statistics of the dataset. ..............................7

2. Split the data into train and test in the ratio 70:30. Is scaling necessary or not? ...................................................10

3. Build the following models on the 70% training data and check the performance of these

models on the Training as well as the 30% Test data using the various inferences from the

Confusion Matrix and plotting a AUC-ROC curve along with the AUC values. Tune the models

wherever required for optimum performance. ………………………………………………………………………………………………………11

a. Logistic Regression Model ………………………………………………………………………………………………………………………………….11

b. Linear Discriminant Analysis ………………………………………………………………………………………………………………………….……12

c. Decision Tree Classifier – CART model ………………………………………………………………………………………………………….…….12

d. Naïve Bayes Model ...………………………………………………………………………………………………………………………………………….13

e. KNN Model …………………………………………………………………………………………………………………………………………………………13

f. Random Forest Model. …………………………………………………………………………………………………………...………………………….14

g. Boosting Classifier Model using Gradient boost. ……………………………………………………………......................................14

4. Which model performs the best? ……………………………………………………………………………………………………………………….15

5. What are your business insights? ……………………………………………………………………………………………………………………….15

Executive Summary

You work for an office transport company. You are in discussions with ABC Consulting company for

providing transport for their employees. For this purpose, you are tasked with understanding how do the

employees of ABC Consulting prefer to commute presently (between home and office). Based on the

parameters like age, salary, work experience etc. given in the data set ‘Transport.csv’, you are required to

predict the preferred mode of transport. The project requires you to build several Machine Learning

models and compare them so that the model can be finalised.

Introduction

We need to build various model using machine learning algorithms to predict if employees would prefer to use an office transport service if the option is available to them, to travel from their home to the office and back after completing their shifts. For this problem we have been provided with data such as age, gender, if the employee is an engineer or if they have done MBA, also their salary, work experience, how far they live and if they have license and what mode of transportation do they use currently. After we build various models, we need to check how well are we able to predict and compare each model with the others and access which model works best. The objective is to build various Machine Learning models on this data set and based on the accuracy metrics decide which model is to be finalised for finally predicting the mode of transport chosen by the

employee.

Data Description

Age: Age of the Employee in Years

Gender: Gender of the Employee

Engineer: For Engineer =1, non-Engineer =0

MBA: For MBA =1, non-MBA =0

Work Exp: Experience in years

Salary: Salary in Lakhs per Annum

Distance: Distance in Kms from Home to Office

license: If Employee has Driving Licence -1, If not, then 0

Transport: Mode of Transport

Questions:

1. Basic data summary, Univariate, Bivariate analysis, graphs, checking correlations, outliers and

missing values treatment (if necessary) and check the basic descriptive statistics of the dataset.

|  | **Age** | **Gender** | **Engineer** | **MBA** | **Work Exp** | **Salary** | **Distance** | **license** | **Transport** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 28 | Male | 0 | 0 | 4 | 14.3 | 3.2 | 0 | Public Transport |
| **1** | 23 | Female | 1 | 0 | 4 | 8.3 | 3.3 | 0 | Public Transport |
| **2** | 29 | Male | 1 | 0 | 7 | 13.4 | 4.1 | 0 | Public Transport |
| **3** | 28 | Female | 1 | 1 | 5 | 13.4 | 4.5 | 0 | Public Transport |
| **4** | 27 | Male | 1 | 0 | 4 | 13.4 | 4.6 | 0 | Public Transport |

Above is the sample head of the data frame in which we have our information stored. We can see various heads under which the information is stored like gender, engineer, MBA, transport etc.

Basic information:

RangeIndex: 444 entries, 0 to 443

Data columns (total 9 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Age 444 non-null int64

1 Gender 444 non-null object

2 Engineer 444 non-null int64

3 MBA 444 non-null int64

4 Work Exp 444 non-null int64

5 Salary 444 non-null float64

6 Distance 444 non-null float64

7 license 444 non-null int64

8 Transport 444 non-null object

dtypes: float64(2), int64(5), object(2)

We observe basic information provided by the above table. We have a total of 9 columns and 444 entries. 7 columns are of float and integer types whereas 2 are object type variables, i.e., Gender and Transport.

Description and summary of data:

|  | **count** | **mean** | **std** |  | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Age** | 444.0 | 27.747748 | 4.416710 |  | 18.0 | 25.0 | 27.0 | 30.000 | 43.0 |
| **Engineer** | 444.0 | 0.754505 | 0.430866 |  | 0.0 | 1.0 | 1.0 | 1.000 | 1.0 |
| **MBA** | 444.0 | 0.252252 | 0.434795 |  | 0.0 | 0.0 | 0.0 | 1.000 | 1.0 |
| **Work Exp** | 444.0 | 6.299550 | 5.112098 |  | 0.0 | 3.0 | 5.0 | 8.000 | 24.0 |
| **Salary** | 444.0 | 16.238739 | 10.453851 |  | 6.5 | 9.8 | 13.6 | 15.725 | 57.0 |
| **Distance** | 444.0 | 11.323198 | 3.606149 |  | 3.2 | 8.8 | 11.0 | 13.425 | 23.4 |
| **license** | 444.0 | 0.234234 | 0.423997 |  | 0.0 | 0.0 | 0.0 | 0.000 | 1.0 |

Through above chart we see the mean age is 27.5. Most have completed engineering, about 25% have done MBA. Average work experience is about 6 years. Salary is 16 lakhs per annum ranging from 6.5 lakhs to 57 lakhs per annum. Average distance is 11 km, ranging from minimum of 3.5 km to 23 km. most employees do not have a license.

Null values:

Age 0

Gender 0

Engineer 0

MBA 0

Work Exp 0

Salary 0

Distance 0

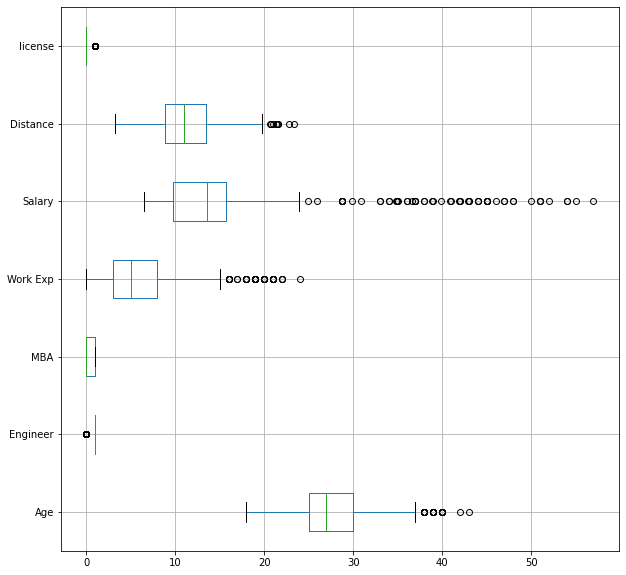
license 0

Transport 0

There are no null values in the given data and thus does not require any treatment.

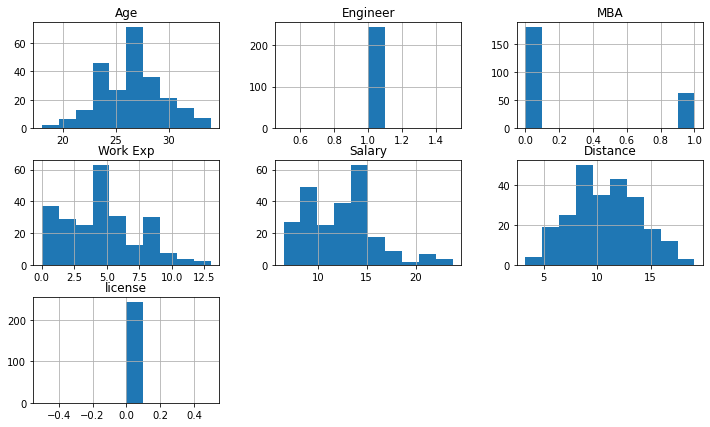
Univariate analysis:

Box plots:



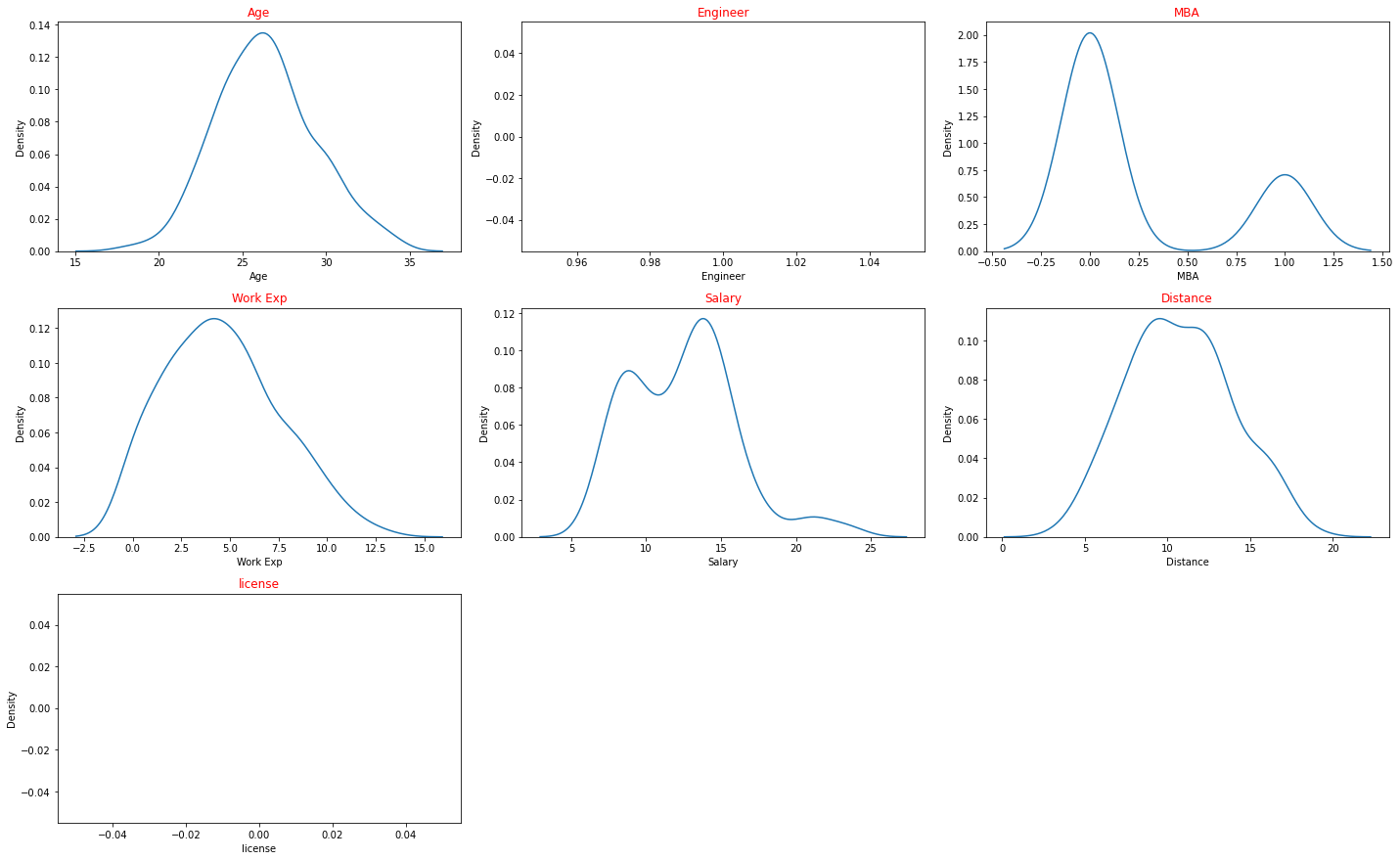
We find many outliers in salary, distance, a few in work exp and age but as this data is collected internally from the company, we believe this data to be right and will not be treating it for outliers.

Histograms:



Here also with the help of histogram we observe that maximum number of people working are of age 27 years. Most people are earning around 14 lakhs per annum.

Distribution graph:



We find that there is largely no normal distribution for the features we have as data.

Heatmap:



We also observe that engineer and MBA have no linear trend towards all other features in our dataset. Work exp, salary and age have a high positive correlation to each other.

RangeIndex: 444 entries, 0 to 443

Data columns (total 9 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Age 444 non-null int64

1 Gender 444 non-null int8

2 Engineer 444 non-null int64

3 MBA 444 non-null int64

4 Work Exp 444 non-null int64

5 Salary 444 non-null float64

6 Distance 444 non-null float64

7 license 444 non-null int64

8 Transport 444 non-null object

dtypes: float64(2), int64(5), int8(1), object(1)

We have converted gender column into and integer type feature to build models to predict transport. Male is now given a value of 1 and female is given a value of 0.

|  | **Age** | **Gender** | **Engineer** | **MBA** | **Work Exp** | **Salary** | **Distance** | **license** | **Transport** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 28 | 1 | 0 | 0 | 4 | 14.3 | 3.2 | 0 | Public Transport |
| **1** | 23 | 0 | 1 | 0 | 4 | 8.3 | 3.3 | 0 | Public Transport |
| **2** | 29 | 1 | 1 | 0 | 7 | 13.4 | 4.1 | 0 | Public Transport |
| **3** | 28 | 0 | 1 | 1 | 5 | 13.4 | 4.5 | 0 | Public Transport |
| **4** | 27 | 1 | 1 | 0 | 4 | 13.4 | 4.6 | 0 | Public Transport |

2. Split the data into train and test in the ratio 70:30. Is scaling necessary or not?

It is a good practice to scale the data before developing a machine learning model as it makes it easy for the machine to learn and solve the problem.

Train set:

Age Gender Engineer MBA Work Exp Salary Distance license

201 29 1 0 0 5 15.9 10.5 0

386 27 1 1 1 6 12.9 15.6 0

329 27 1 1 0 6 12.9 13.3 0

249 23 1 1 0 0 6.9 11.7 0

349 30 1 1 0 7 14.9 14.0 0

.. ... ... ... ... ... ... ... ...

255 29 0 0 0 7 13.6 11.7 0

72 29 0 0 0 7 14.6 7.7 0

396 25 1 0 0 3 9.9 15.9 0

235 24 1 1 1 0 7.7 11.3 1

37 25 0 0 0 3 9.6 6.7 0

[310 rows x 8 columns]

Above is the test set with 70% of the data with all the independent variables **after scaling**.

Here we have randomly picked the dependent variable. As shown below.

201 Public Transport

386 Public Transport

329 Public Transport

249 Private Transport

349 Public Transport

...

255 Public Transport

72 Public Transport

396 Public Transport

235 Public Transport

37 Private Transport

Name: Transport, Length: 310, dtype: object

Test set:

Age Gender Engineer MBA Work Exp Salary Distance license

247 26 0 1 0 8 14.6 11.6 0

179 27 1 0 1 5 13.9 10.0 0

186 35 0 1 0 16 28.7 10.2 0

31 24 1 1 1 2 8.6 6.4 0

218 33 1 1 0 11 16.7 10.9 1

.. ... ... ... ... ... ... ... ...

39 22 1 1 0 3 8.4 6.8 0

192 36 1 1 1 18 28.7 10.4 1

300 25 1 0 0 5 13.7 12.7 1

277 25 0 1 0 5 18.9 12.2 0

98 31 1 1 0 10 14.8 8.4 0

[134 rows x 8 columns]

Above is the randomly picked test set of independent variables with 30% of the data **after scaling**.

247 Public Transport

179 Public Transport

186 Public Transport

31 Public Transport

218 Private Transport

...

39 Public Transport

192 Public Transport

300 Private Transport

277 Private Transport

98 Public Transport

Name: Transport, Length: 134, dtype: object

We can also see test data of the dependent variable.

3. Build the following models on the 70% training data and check the performance of these

models on the Training as well as the 30% Test data using the various inferences from the

Confusion Matrix and plotting a AUC-ROC curve along with the AUC values. Tune the models

wherever required for optimum performance.:

a. Logistic Regression Model

b. Linear Discriminant Analysis

c. Decision Tree Classifier – CART model

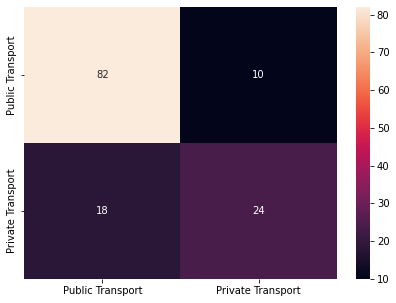
d. Naïve Bayes Model

e. KNN Model

f. Random Forest Model

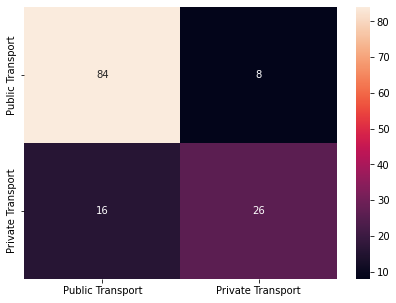
g. Boosting Classifier Model using Gradient boost.

1. **Logistic Regression Model:**



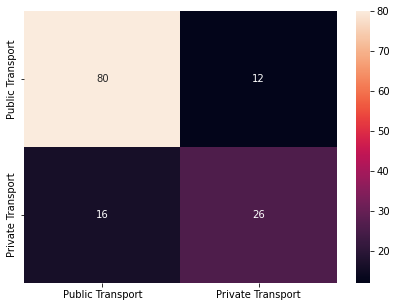
Logistic regression model has an accuracy of 0.7910447761194029.

1. **Linear Discriminant Analysis:**



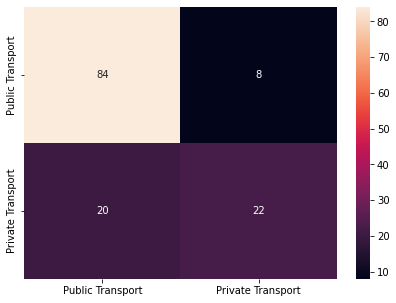
Linear discriminant analysis has an accuracy score of 0.8208955223880597.

1. **CART Model:**



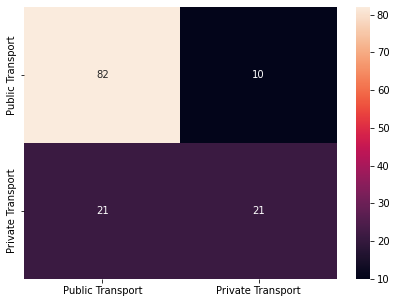
CART Model has an accuracy of 0.7910447761194029.

1. **Naïve Bayes Model:**



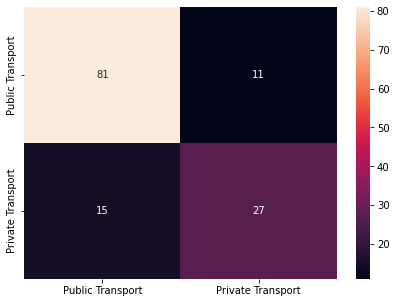
Naïve bayes model gives us an accuracy of 0.7910447761194029.

1. **KNN Model:**



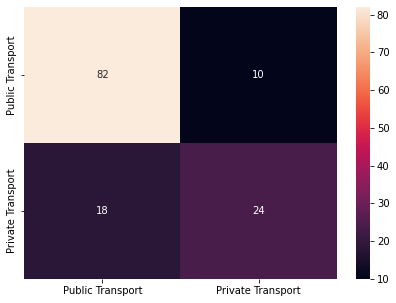
KNN gives an accuracy of 0.7686567164179104.

1. **Random Forest Model:**



Random forest gives us an accuracy score of 0.8059701492537313.

1. **Boosting Classifier Model using Gradient boost:**

****

Boosting classifier model has an accuracy score of 0.7910447761194029.

4. Which model performs the best?

Accuracy score for different model:

Logistic Regression Model:0.7910447761194029

Linear Discriminant Analysis: 0.8208955223880597

CART Model: 0.7910447761194029

Naïve Bayes Model: 0.7910447761194029

KNN Model: 0.7686567164179104

Random Forest Model: 0.8059701492537313

Boosting Classifier Model using Gradient boost 0.7910447761194029

As the above data shows Linear discriminant analysis model is predicting with the best accuracy score of 82%. Hence Linear discriminant performs best.

5. What are your business insights?

As we can see that there are many employees are using a public transport and thus, they would like to use a transport provided by the office.