2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering (ICADEE) | 978-1-7281-9251-2/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICADEE51157.2020.9368922

Exploratory Data Analysis using Artificial Neural Networks

Sriram D[1], Kalaivani K[2*], Ulagapriya K[3], Saritha A[4], Sajeevram A[5]
[1,2,3,4]Department of Computer Science and Engineering, Vels Institute of Science, Technology and Advanced Studies, India
[5]Department of Information Technology, Sri Krishna College of Engineering and Technology, India
[*]kalai.se@velsuniv.ac.in

ABSTRACT: Data analysis helps travel organizations to provide better recommendations for investing in their future trips based on its business and personal trips. This paper presents the basic concepts, various types and levels of data analysis, predictive modeling techniques and appropriate performance measures. There are basically three types of algorithms for predicting such as *linear regression (machine learning model), analysis of Variance (statistical model) and artificial neural network (machine learning model).* Data Analysis is being used in many fields such as health care, manufacturing, information technology and so on. A travel dataset provided by the uber in Kaggle is used to study the performance of chosen predicting algorithms. The primarymethodology behind this study is to analyze and find the accuracy of the most frequent category of trip among all trips taken by a customer in a region using data analysis. The parameters which are taken into consideration are category, purpose, total distance and speed of the travel. The results of precision, recall, f1 score, Area Under Curve (AUC) and Receiver Operating Characteristic Curve (ROC) are evident that the Artificial neural network (ANN) based prediction is comparatively higher than other algorithms.

Keywords: Exploratory data analysis, linear regression, analysis of variance, hypothesis testing, artificial neural network, interquartile ratio, outliers.

1 INTRODUCTION

Travelling is unavoidable and it is a part of day to day activity for every human, there are a number of transportation support providers. They collect data from the commuters which helps to analyze and predict precise information about the customer satisfaction, number of trips taken, distance travelled, and frequency of trips to the same destination and so on. One of the most common ways of trips for near travelling is taken through roadways. The trips taken by the person could either be a personal or business trip for various purposes. Due to high mobility of people moving from one place to another it gave a start for companies to make travelling applications which made it easier for people to move from one place to another for various reasons.

The most known type of travelling for day to day situations are highly taken into consideration for the travelling companies to provide their customer to reach their destination on time which involve various things like start location, end location, distance and speed. Wide Range of companies started to develop applications which could be user friendly for customers. Companies like uber make their mobile applications in such a way that it helps their customers to book their trips whenever they require.

As a result, high volume of data generated in companies like Uber, which had to be managed and needed to be predicted for improving their profit based on their customer's trips in different places. A review has been done on exploratory data analysis related articles and the details are represented in Table 1.1.

Exploratory data analysis helps in providing all the required preprocessing techniques to be done and it's also used for finding the data type of each variable in the data set. Exploratory data analysis (EDA) is an

978-1-7281-9251-2/20/$31.00 ©2020 IEEE

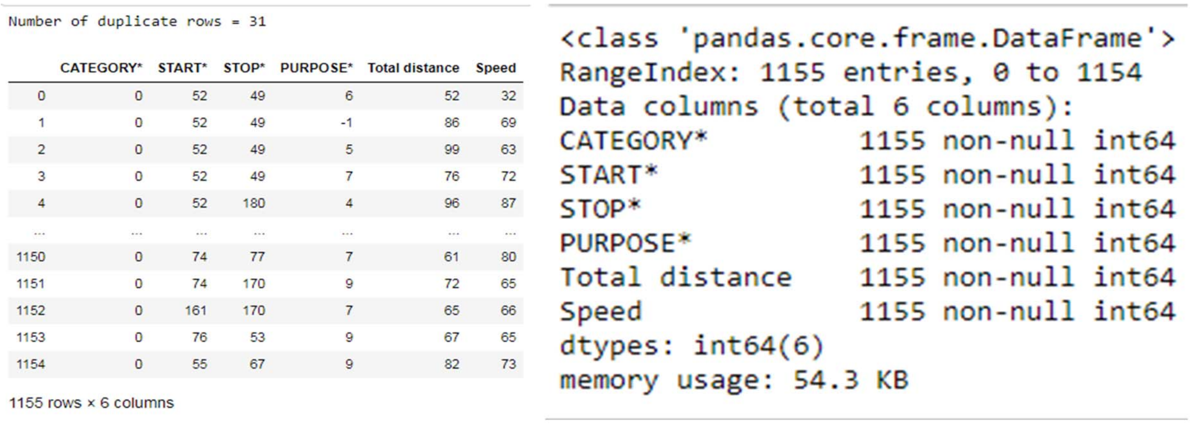
Table 1.1: Review of Exploratory data analysis

Ref	Algorithm / Technique	Metrics	Limitations
[1]	Density Strip Analysis	One Dimensional Heat map	Difficult to map color onto a continuous scale.
[2]	Exploratory Data Analysis	Air Quality Metrics	Not a Definite ratio of Numerical indication in visualization
[3]	Principal Component Analysis, K-means	Standardization	Difficult to predict the number of Clusters for its unit
[4]	Exploratory Data Analysis, Random forest Modelling	Rapid Miner	Model Tuning is insufficient
[5]	Neural Networks	Correlation Analysis	Outlier Analysis is not checked for better Correlation
[6]	ANN and Regression Analysis	Hash table Analysis	Classification report is not analysis for accuracy

analytical model which gives a summarized information about the variable that are present in the data set it is also known to be a helpful process in performing preprocessing techniques in each phase of its data analysis. DESCRIBE function, which helps in providing the in-depth description about its count, mean, minimum, maximum and its interquartile ratio is identified using this function call.INFO function helps in identifying if any missing function are present in the variables on the basis of the number of rows and columns of the data frame. DUPLICATE function is used for finding any duplicate values that are present in the data set if there are any duplicate values then it has to be dropped and then the information has to be provided with no duplicate values. This function call is represented in Fig 1.1.

2 LINEAR REGRESSION

In statistics, linear regression is a predictive modelling technique, which is used for finding the relationship that exists between two variables. It helps us to identify how the variables are dependent on each other. So, the linear modelling is performing with a helps of a dependentvariable with one or independent variables. There are two types of linear modeling techniques: single linear regression and multiple linear regression. Let X be the independent variable and Y be the dependent variable. A linear relationship between these two variables can be defined as follows, and this algorithm is represented in Fig. 2.1.



```
Number of duplicate rows = 31
```

	CATEGORY*	START*	STOP*	PURPOSE*	Total distance	Speed
0	0	52	49	6	52	32
1	0	52	49	-1	86	69
2	0	52	49	5	99	63
3	0	52	49	7	76	72
4	0	52	180	4	96	87
...	...	...	...	...	...	...
1150	0	74	77	7	61	80
1151	0	74	170	9	72	65
1152	0	161	170	7	65	66
1153	0	76	53	9	67	65
1154	0	55	67	9	82	73

1155 rows × 6 columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1155 entries, 0 to 1154
Data columns (total 6 columns):
CATEGORY*         1155 non-null int64
START*            1155 non-null int64
STOP*             1155 non-null int64
PURPOSE*          1155 non-null int64
Total distance    1155 non-null int64
Speed             1155 non-null int64
dtypes: int64(6)
memory usage: 54.3 KB
```

Fig. 1.1: Uber dataset Variables and its data type.

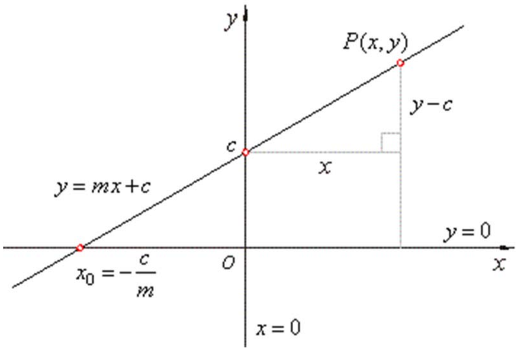


Fig. 2.1. Graphical representation of linear regression.

2.1 Least Square Method

This is performed by identifying the partial derivative of L, equating it to 0 and then identifying an expression for M and C. Finding the Error: So, to reduce the error function, a way to calculate the error in the first place is required. A loss function method in machine learning is simply a calculation of how different the predicted value is from the actual value on the Analysis. Quadratic Loss Function method is used in order to measure the loss or error in our model. Thus the model helps in providing a better predicted value from the actual value which we get from performing certain mathematical operations, these are the resulting equations which are obtained Quadratic Loss Function, Least square method as given below with its explanation in Fig.2.1.1 and 2.1.2**.**

$$L(x)=\sum_{i=1}^{n}(y_i-p_i)^2 \qquad m=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$c=\bar{y}-m\bar{x}$$

where L- partial derivative for minimum sum of squares, y –predicted value, p - Quadratic loss variable, m – Slope, x -Actual value, $\bar{x}$- mean of x, $\bar{y}$ - mean of yand c – Intercept.

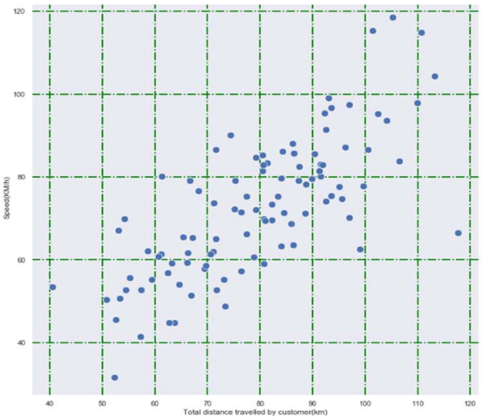


Figure 3a

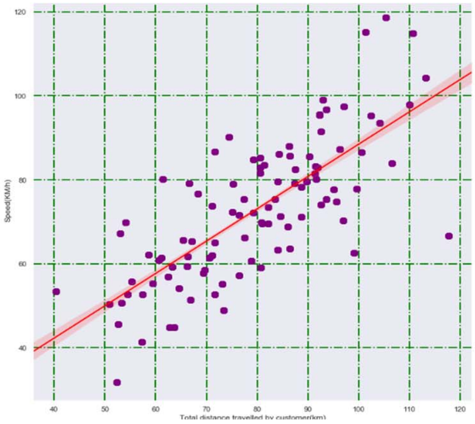


Figure 3b

Fig. 2.1.1. Least Square Linear Regression model of Actual value (independent variable-Speed, dependent variable-Distance), 2.1.2. Least Square Linear Regression model of Predicted value (independent variable-Speed, dependent variable-Distance)

2.2 Scikit-learn Linear Model

Scikit-learn is a linear modelling technique, which is used for performing training and testing of the data set to get much better and accurate results. In the data we take 70% training and 30% testing in this model and then we create a linear regression object which helps in predicting the model analysis. In order to get a constant accuracy ratio we keep the random state as 1000099 value. Thus the model runs in this random state which is encountered to be the highest value obtained of all. Based on the Scikit-learn prediction we could find that the accuracy of the model lies between 50-70%.From this inference the model seems to be performing moderately with the help of the R-Square Value which denotes the Accuracy or how well the model gets fit. The intercept represents the c value, slope represents the m value and Y_Pred represents the predicted value (Y_Pred = mX+C) to be taken into consideration and it's represented in Fig. 2.2.1, Fig 2.2.2 and Fig 2.2.3

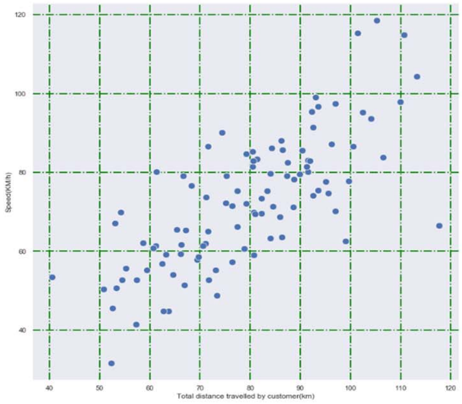
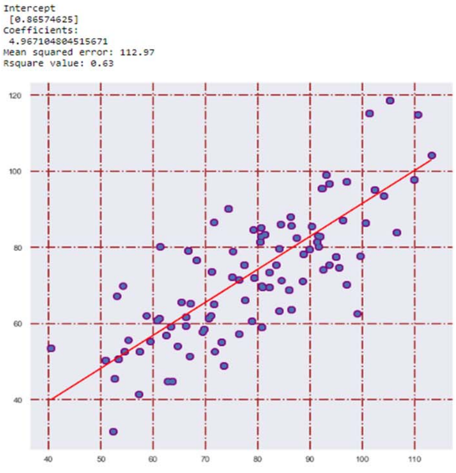


Figure 4a



```
Intercept
 [0.86574625]
Coefficients:
 4.967104804515671
Mean squared error: 112.97
Rsquare value: 0.63
```

Figure 4b

Fig. 2.2.1. Least Square Linear Regression model of Actual value (independent variable-Speed, dependent variable-Distance) Fig. 2.2.2. Predicted Linear regression model with mean square error, intercept, and coefficient with R-square (Accuracy).

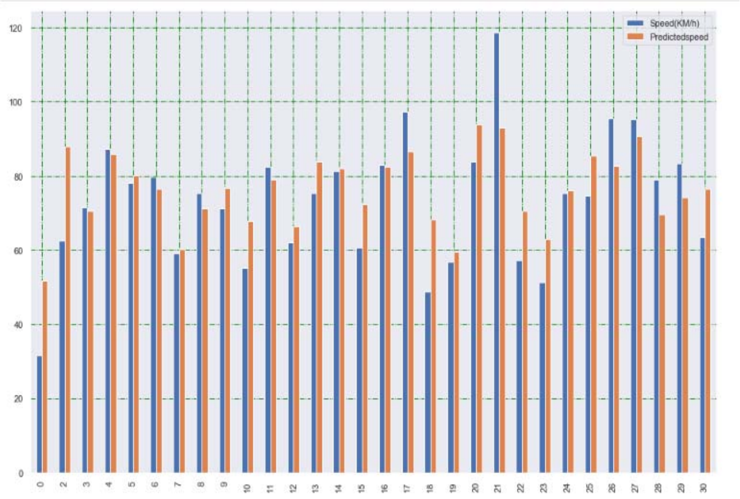


Fig. 2.2.3. Graphical Representation of Actual and predicted Value from the above types of linear regression models.

Based on the above inferences of two types of the linear regression methods we found that the Scikit-learn method performs slightly better than the least square method. Thus the Accuracy Score Obtained from the Scikit-learn is 63% and from least squares we got an accuracy score of about 56%.

3 ANALYSIS OF VARIANCE

Analysis of variance (ANOVA) is a statistical approach that is used for population mean differences that exist among two more factors. ANOVA checks the impact of one or more factors by comparing the means of different samples. In Analysis of Variance certain hypothesis testing is performed in order to provide inference of the results. Such type of Hypothesis Testing involves two types: Null Hypothesis and Alternate Hypothesis it checks by the alpha value using a defined significance level of 95%. There are two types of Hypothesis testing in Analysis of Variance:Null Hypothesis and Alternate Hypothesis.

3.1 One way ANOVA

The word treatment is generic and as such may denote different methods, machines, different advertisement copy platforms, different strategies, different brands and the like[12].

Total sum of the Squares = Treatment sum of the Squares + Error Sum of Squares

The variation in the sum of the squares of the response variable (dependent variables) is caused only by the treatment is attributed to error term. These inferences are shown in Table 3.1.1. Since the P value is greater than the significance level (0.05 alpha value), we accept the null hypothesis and state that the population mean of each customer with different Category (Business and personal) is same in each distance travelled by the customer.

Table 3.1.1 ANOVA output (dependent variable distance), (independent variable category).

	Data Frames	**Sum of the Squares**	**Mean of the Squares**	**Fvalue**	**P Value(> F)**
Design	1	338.3	338.3	1.30	0.25
Residuals	1153	298542.5	18.80	NAN	NAN

3.2 One way ANOVA with OLS(Ordinary least Square) Regression:

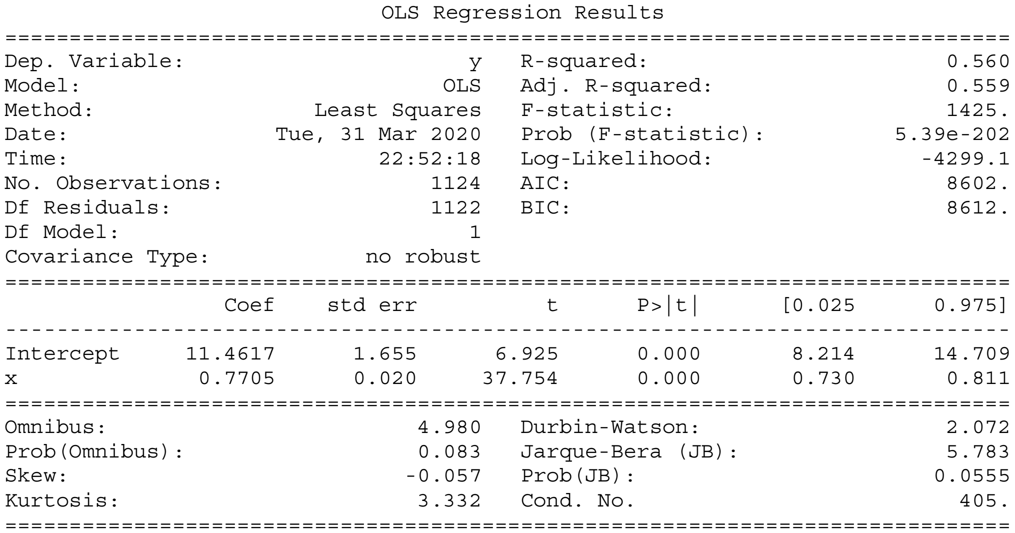
Analysis of Variance with OLS regression is performed if both the dependent and independent variables are continuous. Whereas, in normal one way analysis of variance it happens with a dependent variable as categorical and independent variable as continuous. This operation is performed to identify the mean difference that exists among two continuous variables. Dependent variable is distance, the independent variable is speed. These inferences are shown in Table 3.2.1 and Table 3.2.2.

Since the P value is greater than the significance level (0.05 alpha value), we Reject the null hypothesis and state that the population mean of each Speed travelled with different distance is not the same.

4 ARTIFICIAL NEURAL NETWORKS

Artificial Neural Network (ANN) is an inspiration from the Biological Neural Network algorithm that constitutes the animal brain. Moreover, ANN contains Active function, to perform well. Train and testing are also performed in this Algorithm by standardizing the variable into one unit using Standard Scalar function[9]. There are three layers which are present in the neural network that provide the root reason for predicting the data, they are the Input layer, Hidden layer and Output layer in each phase. Modeling Tuning is performed using the GridSearchCV package from the Scikit learner module. GridSearchCV helps in proving the parameter to be manually assigned. The hidden layer size, activation function, solver and the max iteration can be manually modified to get better accuracy which denotes how well the model performs. A Classification report is generated to provides its necessary values, they are Precision (positive predictive value), Recall (Sensitivity), F1 Score (Binary classification of 0's and 1's)and Support(it denotes how frequent the items appear in a data) with the overall Accuracy score in percentage.These inferences are shown in Table 4.1.

Table 3.2.1 OLS Regression Results of Ordinary least Square method



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.560
Model:                            OLS   Adj. R-squared:                  0.559
Method:                 Least Squares   F-statistic:                     1425.
Date:                Tue, 31 Mar 2020   Prob (F-statistic):          5.39e-202
Time:                        22:52:18   Log-Likelihood:                -4299.1
No. Observations:                1124   AIC:                             8602.
Df Residuals:                    1122   BIC:                             8612.
Df Model:                           1
Covariance Type:            no robust
==============================================================================
                 Coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     11.4617      1.655      6.925      0.000       8.214      14.709
x              0.7705      0.020     37.754      0.000       0.730       0.811
==============================================================================
Omnibus:                        4.980   Durbin-Watson:                   2.072
Prob(Omnibus):                  0.083   Jarque-Bera (JB):                5.783
Skew:                          -0.057   Prob(JB):                       0.0555
Kurtosis:                       3.332   Cond. No.                         405.
==============================================================================
```

Table 3.2.2. ANOVA output (dependent variable distance), (independent variable Speed)

	Data Frames	**Sum of theSquares**	**Mean of the Squares**	**Fvalue**	**P Value(> F)**
Design	1	1175570.4	1175570.4	1425.3	5.3e-2
Residuals	1122	138203.1	123.1	NAN	NAN

Classification Report of the Test Model in ANN:

Table 4.1 Accuracy Score of total number of business and personal trips in a category as 0's and 1's.

Classification	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.94	0.99	0.96	315
1	0.40	0.09	0.14	23
Macro Average	0.67	0.54	0.55	338
Weighted Average	0.90	0.93	0.91	338

From the above classification report the precision, recall, F1 score with its Macro Average and Weighted Average is also determined in percentile format. The testing model runs with an accuracy score of 93%, which tells how well the model, performs in this prediction of predictive modeling [8]. Support denotes the total number of items of 1's and 0's in the data. Thus this testing model gives us a strong classification report with better accuracy score when the random state is assigned to 65.Grid search Cross validation also helped in tuning the model. The Cross Validation is assigned to in the Grid search CV. Using test split function the model is separated as 70% training and 30% testing. These inference are shown in Table 4.2.

Classification Report of the Train Mode in ANN

Table 4.2 Denotes the Accuracy Score of total number of business and personal trips in a category as 0's and 1's.

Classification	**Precision**	**Recall**	**F1-Score**	**Support**
0	0.94	0.98	0.96	733
1	0.35	0.11	0.17	53
Macro Average	0.65	0.55	0.57	786
Weighted Average	0.90	0.93	0.91	786

From the above classification report the precision, recall, F1 score with its Macro Average and Weighted Average is also determined in percentile format. The training model runs with an accuracy score of 93% which tells how well the model performs in this prediction of predictive modelling. Support denotes the total number of items of 1's and 0's in the data.[7] Thus this training model gives us a strong classification report with better accuracy score when the random state is assigned to 65.Grid search Cross validation also helped in tuning the model. The Cross Validation is assigned to in the Grid search CV. Using the test split function the model is separated as 70% training and 30% testing.Based on the above inferences we find that both the training and testing are found to perform almost the same level with an accuracy score of 93%.

Confusion Matrix

Confusion matrix helps in providing the positive and negative values based on actual and predicted values combined to give a 2/2 matrix of total no of true positive, true negative, false negative and false positive numbers in the data set. These inference are shown in Table 4.3.

Table 4.3 Representation of Confusion Matrix

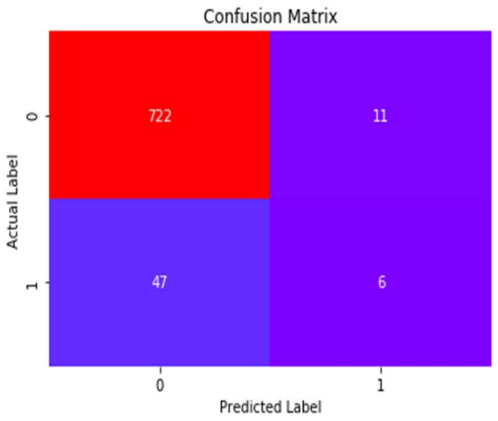
N (denotes the number of values)	**Predicted** Negative	**Predicte d** Positive
Actual Negative	True negative	False Positive
Actual Positive	False Negative	True positive

Thus, the true positive and true negative values are done based on the dependent and independent labels as X and Y in a table format. Based on training and testing the following figures confusion matrix of them are provided in **Fig. 4.1, Fig. 4.2.**

Based on the representation of the confusion matrix we come to a conclusion the true positive, true negative, false positive, false negative value of train data is slightly better than the test data.

Area under the Curve-:

AUC and ROC curve is a performance metrics measurement for classification problems at various threshold point settings. Rectifier characteristic curve is a probability curve and AUC represents degree or metrics of separation. It tells how much a model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting zeros as zeros and ones as ones. By The Rectifier operating characteristic curve is plotted with True Positive Rates against the False Positive Rates where True Positive Rates is on the Y axis and False Positive Rates is on the X axis. These inferences are shown in Fig.4.3.



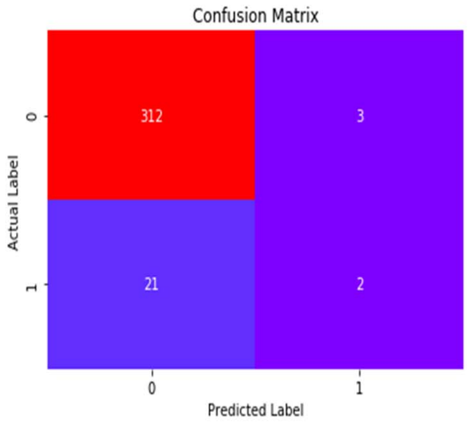


Fig.4.1. Representation of Confusion matrix of testing set with actual and predicted value. Fig. 4.2. Representation of Confusion matrix of training set with actual and predicted value.

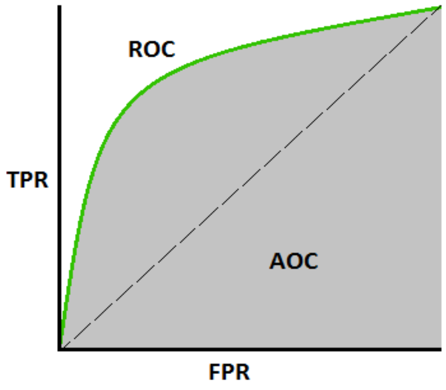


Fig. 4.3 Implementation of AUC-ROC curve

Methods used in AUC-ROC Curve:

True Positive Rate (TPR) / Recall / Sensitivity = TP/TP + FN.

Specificity = TN /TN + FP

False Positive Rate (FPR) =FP/TN + FN

AUC-ROC Curve of Training set and Test set:

Higher the AUC, better the model.

The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis. These inferences are shown in Fig.4.4 and Fig.4.5.

Based on the above inferences the AUC-ROC of training and testing data is identified using its true positive, true negative, false positive, false negative, sensitivity and recall.ROC is a probability curve and AUC represents degree or measure of separation. It tells how much a model is capable of distinguishing between classes.AUC on the training data is 87% and on test data is 72%. The precision and recall metrics are also almost similar between training and test set, which indicates no over fitting or under fitting has happened. Best grid model has better improved performance over the initial classifier model as the sensitivity was much lesser in the initial model. The Overall model performswell in predicting the category of trips travelled by a customer. These inferences are shown in Table 4.4.

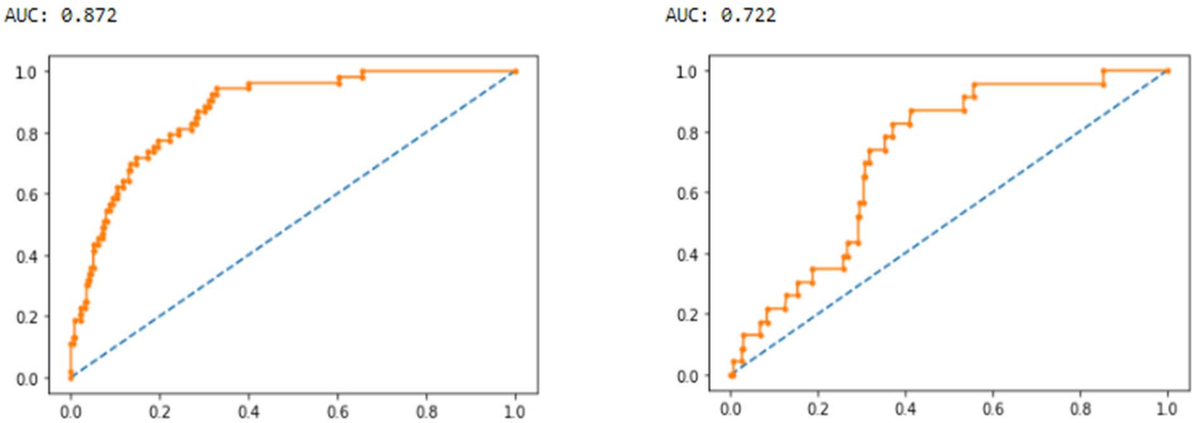


Fig.4.4. AUC curve train and test

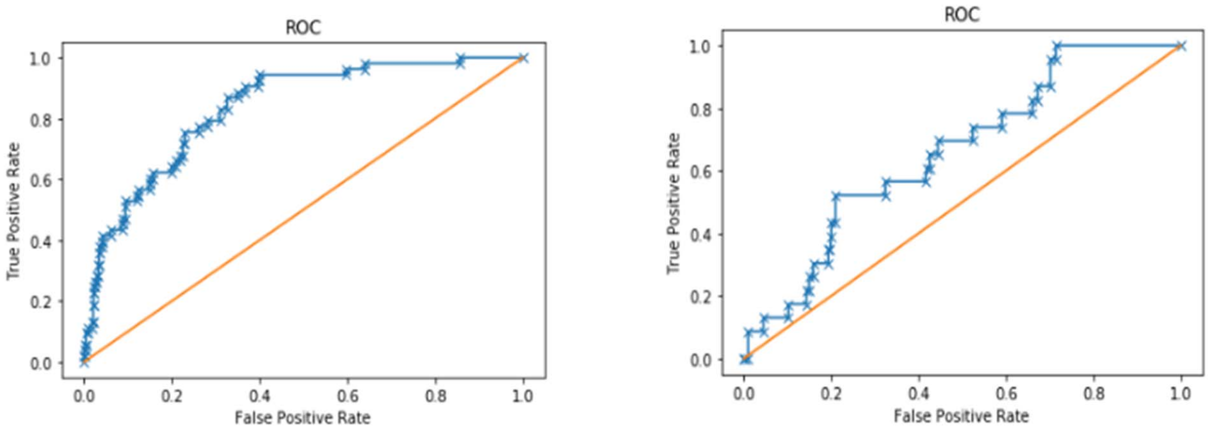


Fig.4.5 ROC Curve of train and test set.

Table 4.4 Accuracy Results in Each Algorithms

Algorithm	**Accuracy**
Linear Regression (R square value)	63%
ANOVA(OLS)	56%
Artificial Neural Networks	93%

5 CONCLUSION

Based on the three predictive models such as Linear Regression, Analysis of Variance, Artificial Neural Network, it is found that the Accuracy obtained in ANN is far better than Linear Regression and Analysis of Variance. The Accuracy Score of Linear Regression under least Square method is 55%, Scikit learn is found to be 64%, which is slightly better than least Square method. Training and testing were taken in a ratio of 70:30. Then standard scalar function was used which helped in providing or grouping all the variables into one unit and from the above inference it is found that ANN model seems to be performing much better than Linear regression with an Accuracy score of 93%. The Accuracy of the ANN model is found to be better performing than linear regression and ANOVA Ordinary least square method used in uber drive dataset.

6 REFERENCES

[1] Daniel Cheng,Tile based visual analytics for Twitter big data exploratory analysis,IEEE /2013.

[2]Changhui Yu, Research of time series air quality data based on exploratory data analysis and representation,IEEE/2016.

[3]Nasser ,D. Hamad , C. Nasr, Visualization Methods for Exploratory Data Analysis,IEEE/2016.

[4]Liangchen Guo , Yazhong Zhang , Chang Lu , A System for Exploratory Analysis in Cloud,IEEE / 2018.

[5]Lean Yu , Shouyang Wang , K.K. Lai,An integrated data preparation scheme for neural network data analysis,IEEE/2006.

[6]Bin Wang ; Yan Fang ; Jinfang Sheng , BTP Prediction Model Based on ANN and Regression Analysis,IEEE/2009.

[7] B.X.Wang, D.H.Zhang, J.Wang, et al, “Application of Neural Network to Prediction of Plate Finish Cooling Temperature”, Journal of Central South University of Technology, 2008,15(1):136−140.

[8] Demartines P. and J. Herault, "Curvilinear Component Analysis: A self-Organizing Neural Network for Nonlinear Mapping of Data Sets", IEEE Trans. on Neural Networks, vol. 8, no. 1, pp. 148-154, January 1997.

[9] Mao J. and A. K. Jain, "Artificial neural networks for features extraction and multivariate data projection", IEEE Trans. Neural Networks, vol. 6, no. 2, pp. 296-317, 1995. [

[10] Jolliffe I.T., "Principal Component Analysis". SpringerVerlag, 1986.

[11] Haykin, S., "Neural Networks. A comprehensive foundation". Prentice-Hall, Englewood Cliffs, NJ, 1999.

[12] Aggrawal C., "A Human-Computer Interactive Method for Projected Clustering", IEEE Trans. Knowledge & Data Eng., vol, 16, No 4, pp. 448-460, April 2004.