Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

An Analytical Study on Indian Transportation using Machine Learning

SHREYA VAIRAVA SUBRAMANIAN
Sri Ramachandra Faculty of Engineering and Technology
Sri Ramachandra Institute of Higher Education and Research
Chennai, India

PUVVADA YASASWINI
Sri Ramachandra Faculty of Engineering and Technology
Sri Ramachandra Institute of Higher Education and Research
Chennai, India

ABSTRACT – The deficit of analytics and management of data leads to an improper planning and construction of transport networks thereby leading to a high degree of traffic signal breaches and accidents. The idea presented in the paper is to give an insight on course of a route from its source to the destination. A machine learning approach integrated with descriptive statistics and regression analysis was utilized to detect the information of a route including an individual study on every parameter. Since a simple descriptive study might not generalize a notion, an additional regression analysis was performed to help the policy makers of transportation industry in carrying a planned outcome on traffic management and accident control.

KEYWORDS – Descriptive statistics, Regression analysis, Machine Learning

I. INTRODUCTION

In roadways, a proper understanding of the structure of the road, the ability to withstand large volumes of vehicles, the inclusion of speed breakers and road intersections are crucial for transportation planning. The primary step in this approach is to develop an end-to-end model that gives an insight on routing. In contrast to the technological advancements that airways have through air traffic controls and railways have through rail traffic controls in the coordination of transport, roadways lack a proper traffic control medium. Manual traffic regulations prevail in India. The power of data in the contemporary world has a larger quantum when compared to a human regulator. This is similar to what Dimah Dera said in his book that the transportation system is evolving from a technology-driven independent system to a data-driven integrated system of systems [1].The construction of roads after planning which involves basic factors like lighting on roads, regulation of traffic through signals, lane markings such as road marking for movement of vehicles and zebra crossing for pedestrians, dividers for separation of incoming and outgoing traffic and road signs are to be taken into consideration in order to overcome road safety violations including breach of traffic signals and occurrences of road accidents. Despite having these factors responsible for planning and construction of roadways, an improper coordination can sometimes have an expensive impact on economy.

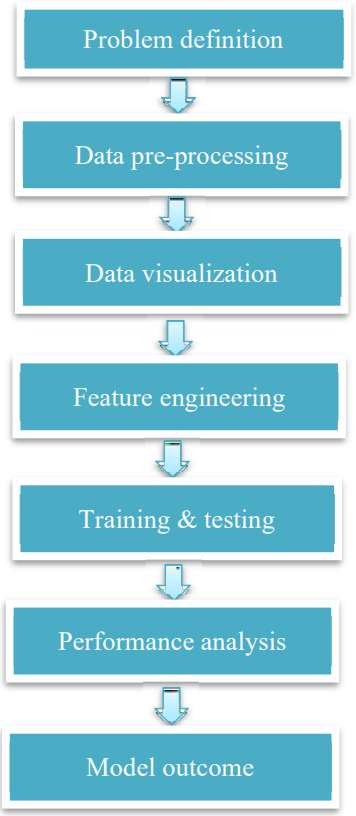
Sometimes, in railways, poor construction of rail tracks often leads to train mishaps which have become very frequent in India. Despite having digitalized mechanisms to control rail traffic, traffic controllers commit negligence sometimes. There is a lack of proper coordination and management. It is essential to have an analytical study on transportation data in terms of statistical testing and forecasting to regulate the movement and prevent any unforeseen mishaps. Another drawback is the utilization of outdated tools and equipments used for construction of railways. With the advancements in technology, there is no proper skill utility facilities. An account to justify these issues is the train accident in Odisha, which occurred due to clash between three trains coming towards one another. This was due to lack of proper coordination and negligence of railway traffic control systems. If there was a digitalized format of data which can be made available to locomotive drivers at any point of time, such collisions could have been avoided. This is one account when in comparison, there are multiple accidents which have become very common in India. Instead of addressing these fundamental issues in planning and construction of railways, there is much more emphasis given on building modern locomotives. This indicates lack of timing which is also a major concern.

In comparison to roadways and railways, airways are much more advanced in terms of developments taking place in transportation sector. There is a revised planning and implementation of movement of aircrafts. The construction of air vehicles involves a very deep planning and a modern yet an effective approach to prevent collisions as much as possible. Unfortunately, this development is not booming in the management of railways and roadways. In this regard, a machine learning model was built which analyzed the possible parameters involved in determining the routing information such as time taken to travel, distance traveled by the mode of transport and expenses shared through the entire journey. In terms of regression analysis, the data was analyzed with multiple measures such as normality testing, hypothesis testing and correlation testing. A hybrid machine learning model integrated using various techniques was implemented to detect which model gave the best result. The results were finally compiled to check for various vulnerabilities. There were many such vulnerabilities found. In order to overcome these vulnerabilities, some advanced ML algorithms were built on top of these vulnerabilities. The reason for taking regression analysis is to ensure that it helps in forecasting the trends likely to happen in the future. This can be an optimal solution to avoiding mishaps. Another advantage of this analysis is to determine the risk factors associated with planning and construction of vehicles. In terms of optimization, an effective and an efficient model can be built which can overcome the existing vulnerabilities in this industry. Such prevailing issues were taken as the base to develop a cost-effective model which could possibly eradicate any breach of road safety protocols. This process can collectively be defined as Transportation Data Analytics which can help in making enhanced decisions regarding planning, construction and management of transportation networks.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

II. RESEARCH METHODOLOGY

The importance of data in a model is to have an overall impact on the performance of the model. The dataset was taken from an available source. Many performance evaluation metrics can be used to validate the efficiency of a machine learning model [2]. There is clear trade-off between the performance of a machine learning model and its ability to produce explainable and interpretable predictions [3]. The pipeline used in our model is:



Given a problem definition, the goal would be to provide an optimal solution. In our case, the definition is to identify the routing information of a journey. In order to pre-process the data, descriptive statistics were used to measure the scale of the dataset. Methods available in python such as describe() to describe the summary of the dataset, info() to provide information about the structure of the data frame object, mean() to print the mean of the data frame objects, std() to print the standard deviation of data frame objects and memory_usage() to detect the memory allocation of index and columns.

During the data cleaning process, the first step is to check for null values or missing values in the dataset. There are various methods to check for missing values. One method is by adding individual data frames together. The second method is to implement python libraries such as numexpr and bottleneck to enhance the numerical computations. The third method is to detect for outliers using a box plot model. In our case, the first two methods mentioned above were implemented. It was found that there were many missing values in the dataset. In order to overcome those missing values, a method called fillna(0) was implemented to replace the missing values by zero. Before proceeding to data visualization, standardization was performed for every file present in the dataset.

STANDARDIZATION:

Standardization is a method of scaling the features of a model in order to have an easier medium for comparison. The mean is set to zero and the standard deviation is set to one. This is an efficient method to detect the outliers particularly in dealing with large datasets.

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

Here, there are four parameters. The parameters:

Z = Standardized value
x = Original value
μ = Common mean
σ = Standard Deviation

The next step to standardization is data visualization. In our model, the libraries used for visualization were matplotlib, seaborn and plotly. Matplotlib and seaborn are common libraries for static visualization whereas plotly is an advanced library developed for an exclusive dynamic as well as an interactive visualization. For every problem statement identified, there were multiple data visualizations to support the model.

Feature engineering was performed where features were selected from the data set to be given for training and testing. Out of the eight data frame objects, seven objects were given as a feature of X and eighth object was given as a feature i.e., target variable to Y. In order to train and test, train_test_split() was called from sklearn.model_selection .

In order to evaluate the performance of model, various ML algorithms were employed to enable a comparative study and analyze the best model with the result generated. The models used for performance analysis primarily were:

LEARNING ALGORITHMS:

- Linear Regression
- Random Forest
- Decision Tree
- Neural Networks
- Ridge Regression (Tikhonov Regularization)
- Least Absolute Shrinkage and Selection Operator (Lasso) Regression (L1 – Regularization)
- ElasticNet Regression

ENSEMBLING TECHNIQUES:

- Bootstrap aggregating (Bagging)
- Stacking
- Boosting (Gradient Boosting)

A hybrid model using the above listed learning algorithms was created that was integrated with ensembling algorithms. A comparison was performed on which model gave the best result. It is important to understand the significance behind these algorithms.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

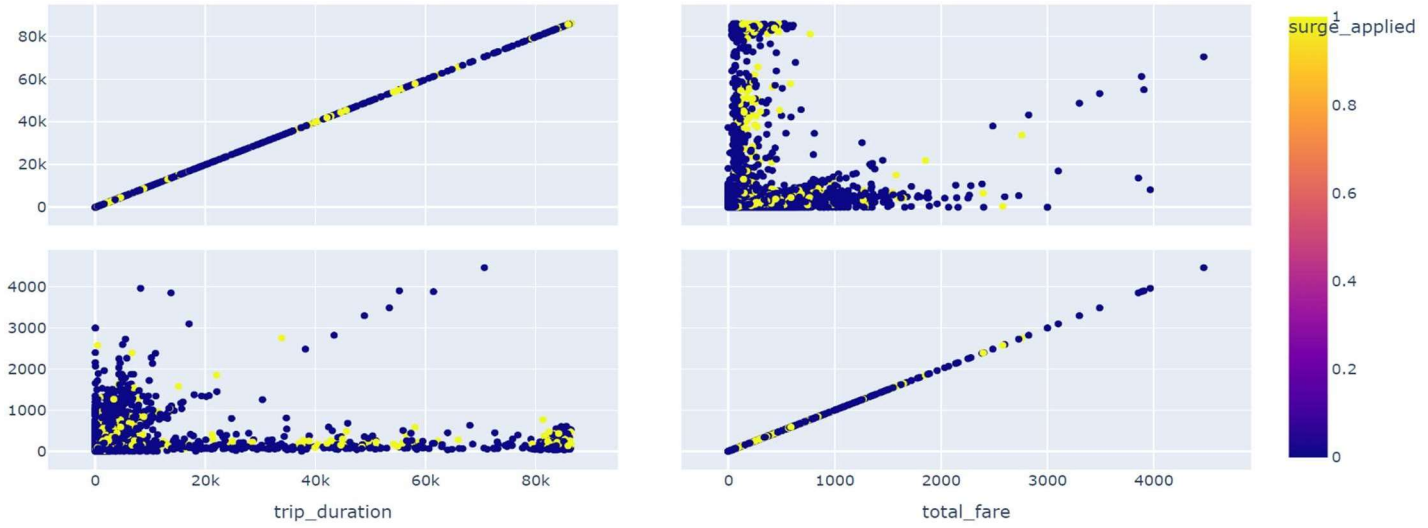


Fig (1)

Fig (1) shows the relationship between two variables with a scatter plot i.e., trip_duration and total_fare with respect to surge applied to the total fare.

LINEAR REGRESSION:

Linear regression is an algorithm applied to give the relationship between a dependent variable in the form of Y and an independent variable in the form of X.

$$Y_i = f(X_i, \beta) + e_i \quad (2)$$

The parameters:

Y_i = Dependent variable (Y)
X_i = Independent variable (X)
β = Value of the intercept
e_i = Error rate

Linear regression can be used in transportation data analytics for various applications such as forecasting. It can prevent collision detection and assist in the optimization of transportation planning and management. It is often used as a predictive tool, and it helps to explain the relationship between independent variables (X1, X2 . . . Xk) and the tested dependent variable (Y)[4].

RANDOM FOREST:

Random forest is an algorithm by which the root node calculates the output of a model by combining the individual outcomes coming from multiple decision trees which acts as leaf nodes. The trees have no dependency over one another. Random forest is a supervised ensemble learning method that acts based on the decision tree [5].

ALGORITHM FOR RF:

The algorithm for random forest is as follows:

- Decide the number of trees to be chosen for the model.
- A new bootstrap sample can be created for the previously decided 'N' trees using a method called resampling.
- Train the decision tree.
- For every leaf node present in the tree, select 'R' features randomly.
- Compute the entropy and information gain for those featured selected previously.
- Iterate until the final node is reached for computation.

In our case, since the model is a regression model, we would simply calculate the Root mean squared error and R2 score for evaluation. For classification models typically, the entropy and information gain using GINI and CART can be computed.

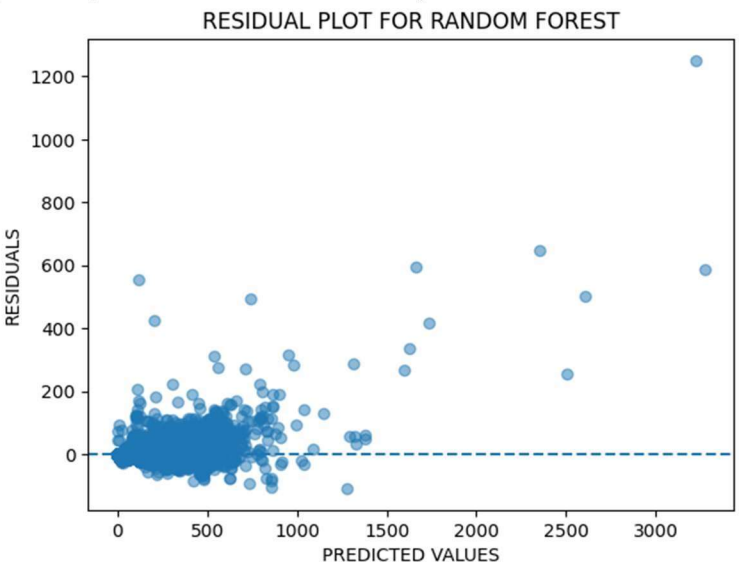


Fig (2)

Fig (2) indicates the residual plot for random forest in regression. Since 95% of the residuals are scattered and accumulated around 0, this is an indication that the relationship between the predicted values and actual values are linear. Thus, it is observed that the plot works best for this regression model.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

DECISION TREE:

Decision tree is an algorithm used to evaluate the model based on individual outputs given by leaf nodes. Its working is similar to that of random forest. Thus, we can say that Random Forest is another sub-division of decision tree. In this model, Decision Tree Regressor is implemented due to the model being a regression model. The difference between a decision tree and decision tree regressor is that the former can be used for dealing with regression and classification problems while the latter is specifically designed to deal with regression problems. Also, in comparing RF and DT, RF separates the nodes by equal division through binary search whereas DT recursively partitions the leaf nodes to form subsets of the root node.

The name of this method is inspired from the shape of a tree, where the class labels are the leaves and the features (or conditions) are the branches[6]. By analyzing the route characteristics, the policy making cell of transportation industry can make improved decisions on public transit management.

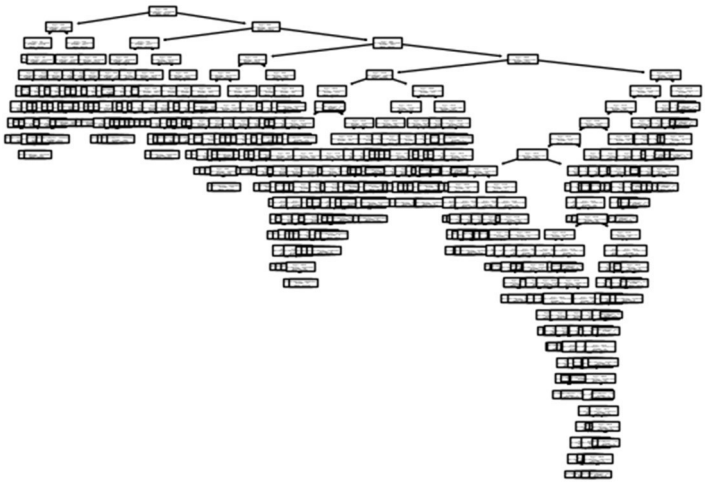


Fig (3)

Fig (3) indicates the relationship between the feature variable “tip” and the class name given as “total_fare”. Due to large data, the graph is condensed for visualization.

NEURAL NETWORKS:

Neural networks are a class of machine learning but also a deep learning network. A neural network consists of interconnected neurons which processes an input and finally produces the output. A neural network consists of Input signals which in our case is the objects of the data frame. Those eight objects are given as input samples. The input signals have to pass through the input layers which consists of eight neurons. Every neuron from the input layer will be connected with every other neuron present in the hidden layer of the neural network. The number of hidden layer in our model is one. Similar to input layers, every neuron in the hidden layer will be connected to every other neuron in the output layer. The signals after passing through the output layer will combine to operate with an activation function and a summation processor, both of which would function simultaneously.

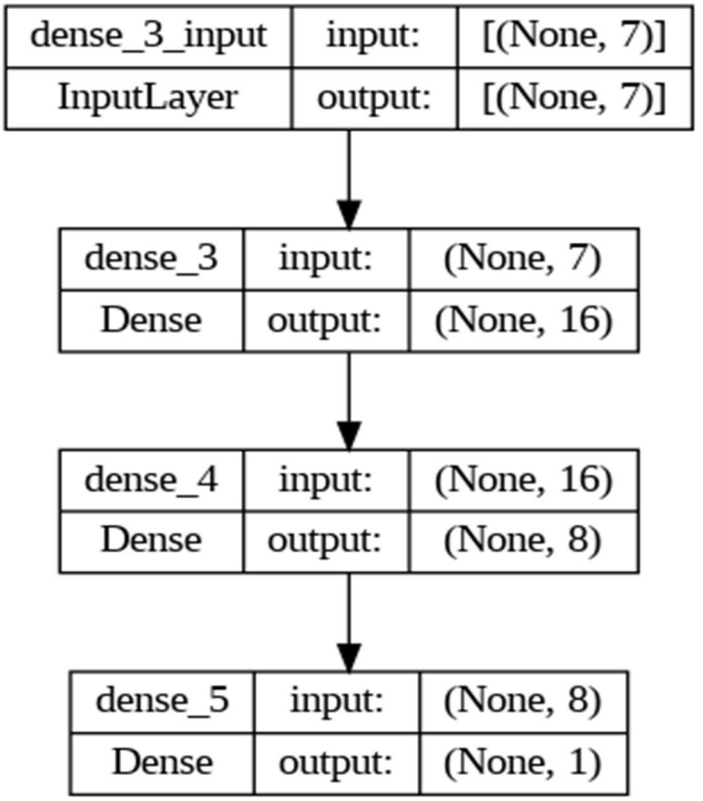


Fig (4)

Fig (4) gives the flow of the neural network which is created using a Sequential model. The model has three dense layers given as input layers. The dimensions reduce from 16 x 1 to 1 x 1.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_3 (Dense)             (None, 16)                128

 dense_4 (Dense)             (None, 8)                 136

 dense_5 (Dense)             (None, 1)                 9

=================================================================
Total params: 273
Trainable params: 273
Non-trainable params: 0
_________________________________________________________________
```

Fig (5)

Fig (5) displays the model summary for the neural network trained using the sequential model created. It is observed that there are 273 trainable parameters.

MULTI – COLLINEARITY ANALYSIS:

Multi-collinearity analysis is performed in machine learning to detect high correlations between any two individual variables present in the model. In our model, multi-collinearity was tested with the following algorithms:

- Ridge Regression (Tikhonov Regularization)
- Least Absolute Shrinkage and Selection Operator (Lasso) Regression (L1 – Regularization)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

- ElasticNet Regression
- Variance Inflation Factor (VIF)
- Tolerance

RIDGE REGRESSION:

Ridge regression, also known as L2-Regularization is a regularization technique used to prevent the model from getting overfitted. Although ridge regression analysis is a biased estimation method, it does not need to eliminate explanatory variables [7].

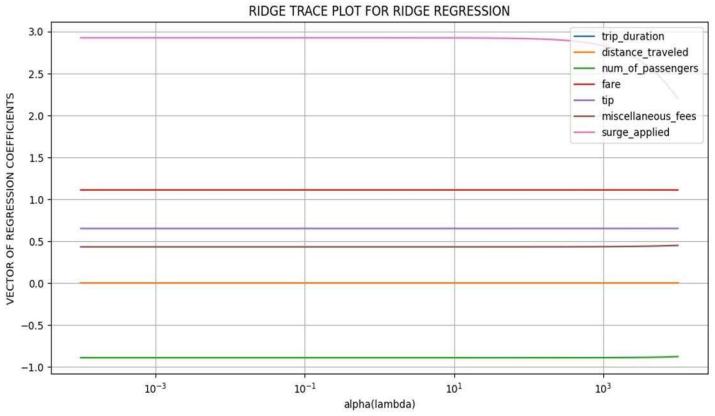


Fig (6)

Fig (5) represents a ridge trace plot of ridge regression. As the parameters increases from left to right in the axes of the plane, they eventually will shrink to zero at one point. The parameters "num_of_passengers", "distance_traveled", "miscellaneous fees", "tip" and "fare" indicate stability of coefficients. But the parameter "surge_applied" indicates an instability where the value eventually decreases from left to right.

The formula for Ridge Regression is:

$$SSE_{L_2} = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \lambda \sum_{j=1}^{P} {\beta_j}^2$$

Here, (3)

SSE = Sum of squared errors
n = Number of terms or samples
yi = Actual value
ŷi = Predicted value
λ = Tuning parameter
β = Vector of regression coefficients
P = Penalty parameter

LASSO REGRESSION:

It is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical regression model [7]. The goal of lasso regression is to minimize the squared errors between the actual values and the predicted values. It is particularly more efficient for our model because it works well for data with larger dimensions. The formula is:

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 + \lambda \sum_{j=1}^{m}|\hat{\beta}_j|$$

Here, (4)

N = Number of terms or samples
Yi = Actual value
Xi^T = Predicted value
β = Vector of Regression coefficients
m = Dimensionality of feature space

The main difference between Lasso and Ridge Regression is the addition of the penalty parameters. In ridge regression, the penalty parameter is added through squared sum of coefficients multiplied with alpha. But in lasso regression, the penalty term to the loss function is added through the sum of absolute value of coefficients multiplied with lambda.

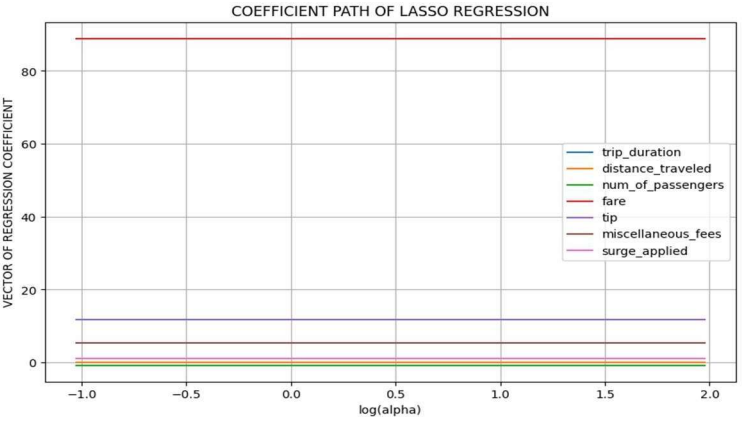


Fig (7)

Fig (7) shows a coefficient path plot of lasso regression. It indicates the behavior of "vector of regression coefficients" against "log(alpha)" value. It can be observed that the parameters "num_of_passengers", "distance_traveled", "surge_applied" are shrunk to zero. The parameters like "miscellaneous_fees" and "tip" are close to zero. The parameters which are zero or are closer to zero indicates that they are less influential on other features in the data. But the parameter "fare" is selective.

ELASTICNET REGRESSION:

ElasticNet Regression is a result of combination of the penalty parameters of L1-Regularization and L2-Regularization. Compared to lasso, the elastic net approach performs better with data of the kind $p \gg n$ with several co-linearities between variables [8].

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}{\hat{\beta}_j}^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|\right)$$

(5)

Here,

SSE = Sum of squared errors
n = Number of terms or samples
yi = Actual value
Xi^T = Predicted value
λ = Tuning parameter
β = Vector of regression coefficients

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

P = Penalty parameter
m = Dimensionality of feature space

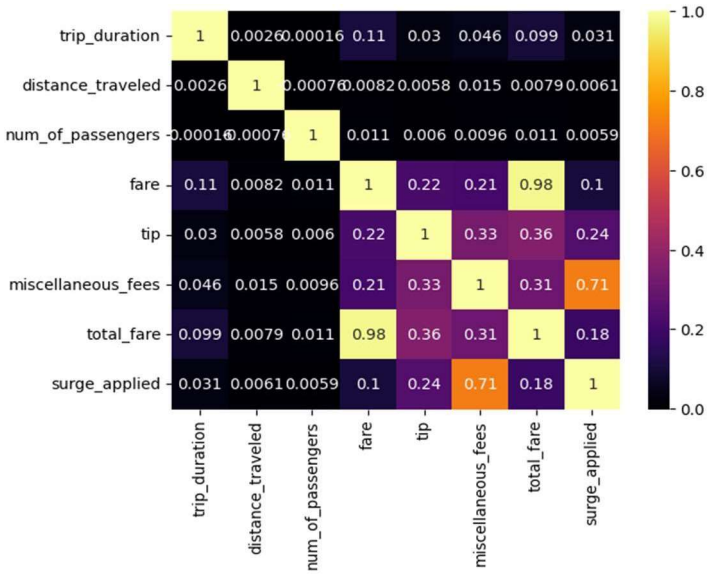


Fig (8)

Fig (8) shows the correlation matrix or the feature importance plot for ElasticNet Regression.

From the methods mentioned above, Ridge regression, Lasso regression and ElasticNet regression help in analysis of routing which could help make an advanced planning and management of transportation.

VARIANCE INFLATION FACTOR:

It provides an index to indicate the increase of VIF due to collinearity [9]. It displays the spread of regression coefficient. The formula for VIF is:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Here, (6)

R_i^2 = Regression Coefficient

The interpretation of the VIF present in a model can be done as follows:

- If the value of VIF > 5, it is an indication that there is high multi – collinearity in the data between two variables.
- If the value of VIF = 1, it is an indication that there is no multi – collinearity in the model.
- If the value of 1 < VIF < 5, it is an indication that there is moderate multi – collinearity present in the model.

VIF shows best results for regression models in comparison to classification models. VIF can be used in traffic analysis to help assess the level of multi – collinearity prevailing in transportation system. It can take any two parameters for comparison and can help in the final assessment.

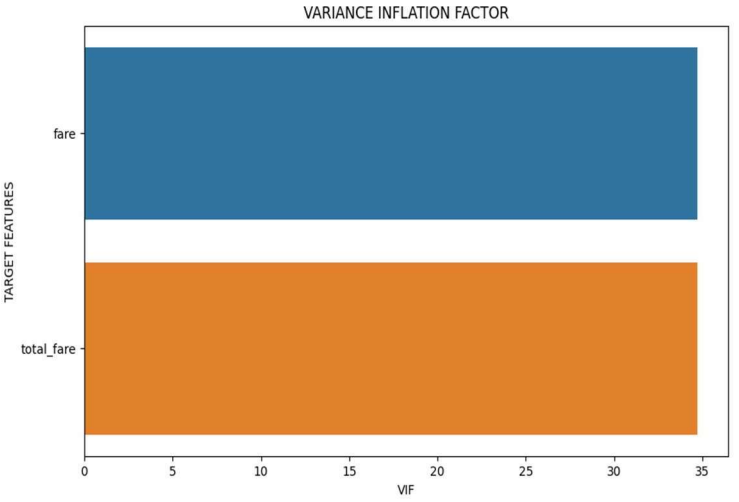


Fig (9)

Fig (9) shows a plot of variance inflation factor (VIF) against the target features. It is observed that the multi – collinearity between "fare" and "total_fare" is nearly 35. Since the VIF > 5, it suggests a very high multi – collinearity in the model. In order to reduce the multi – collinearity, the model was trained previously with ridge regression and lasso regression.

TOLERANCE:

Multicollinearity was assessed via calculating tolerance values with values smaller than a certain threshold value being considered as a sign of multicollinearity [10].

$$\textbf{\textit{Tolerance (T) = 1 / VIF}}$$

The interpretation of Tolerance can be done as follows: (7)

- If the value of Tolerance is closer to 0, it indicates a high multi – collinearity.
- If the value of Tolerance is closer to 1, it indicates a low multi – collinearity.

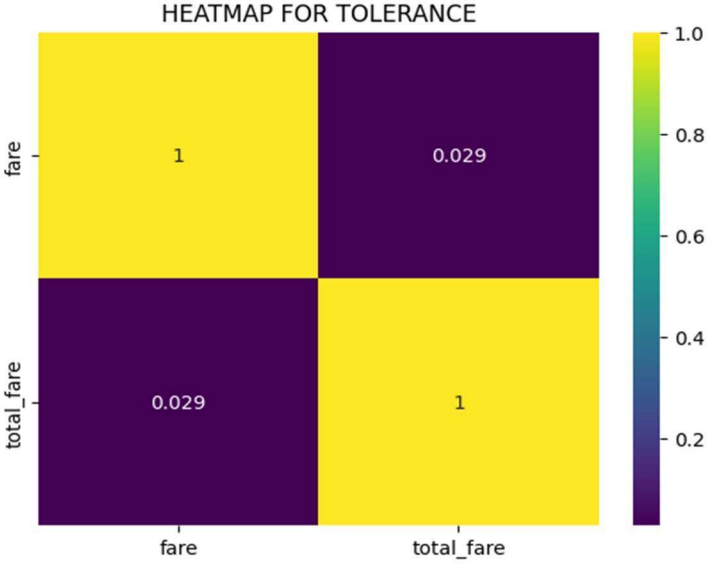


Fig (10)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

Fig (10) shows the heatmap for tolerance. There were two parameters that were taken for the multi – collinearity assessment using tolerance. They were "fare" and "total_fare". From the interpretation given above, it can be understood that any tolerance measures closer to 0 indicates a high multi – collinearity. So, our model having a tolerance value of 0.029 (nearly closer to zero) indicates a high multi - collinearity.

ENSEMBLING TECHNIQUES:

BOOTSTRAP AGGREGATING (BAGGING):

The purpose of bagging is to decrease variance while retaining the bias of a decision tree and preventing overfitting [11]. The algorithm of bagging is as follows:

- Create a new bootstrap sample by selecting some data points in tandem. Random forest may be employed to select the data points.
- Individual model is trained for every set of bootstrapped samples.
- The final prediction is based on sum of class probabilities where the class with the highest sum probability is chosen.

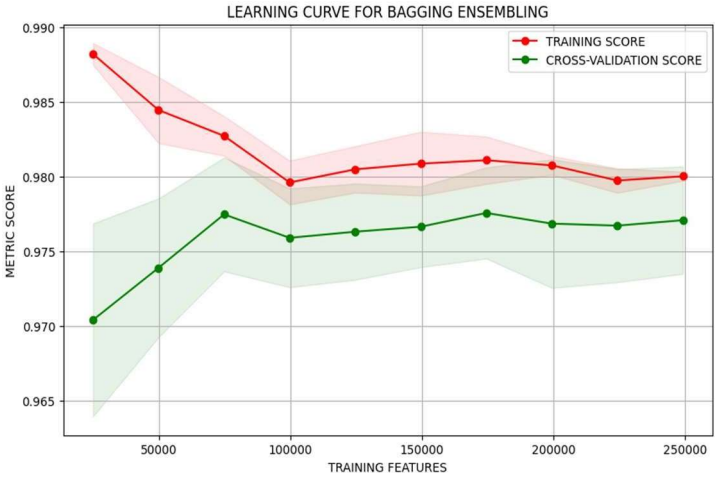


Fig (11)

Fig (11) indicates the learning curve plotted between training features on X – axis such as "num_of_passengers", "tip", "fare", "surge_applied", "miscellaneous_fees", "trip_duration" and "distance_traveled" against a metric score on the Y – axis. Since there is no significant gap between the training score and the cross-validation score, it is an indicated that our model has not been overfitted. The points (0.9180, 100000) and (0.9176, 100000) indicate where the variance reduction had begun.

STACKING:

Stacking builds its models using different learning algorithms and then a combiner algorithm is trained to make the ultimate predictions using the predictions generated by the base algorithms [12]. Stacking combines the individual inputs to form the output. Stacking makes the final prediction of the model based on various methods such as boosting, decision tree etc. Stack architectures can be visualized using various plots. One such method of visualization can be the stack architecture which can be plotted to combine the values of any two algorithms in the form of a base model and a meta model.

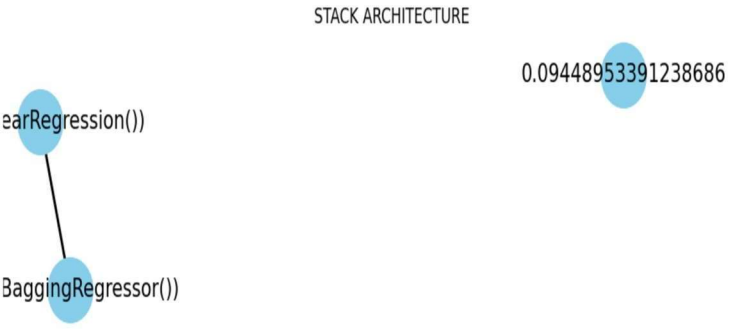


Fig (12)

Fig (12) displays the stack architecture of our model. It is called a stack comparison model. There were two models implemented in the architecture out of which the base model was taken to be BaggingRegressor () and the meta model was LinearRegression (). It is a graph created with two nodes representing each type of the models mentioned above. The final output after the combination of individual inputs was computed to be 0.09 (approx..). It is a good score in comparison to a regression metric like R2 – Score.

BOOSTING (GRADIENT BOOSTING):

Boosting is similar to bagging, but with one conceptual modification. Instead of assigning equal weighting to models, boosting assigns different weights to classifiers, and derives its ultimate result based on weighted voting. In case of regression a weighted average is usually the final output [12].

.

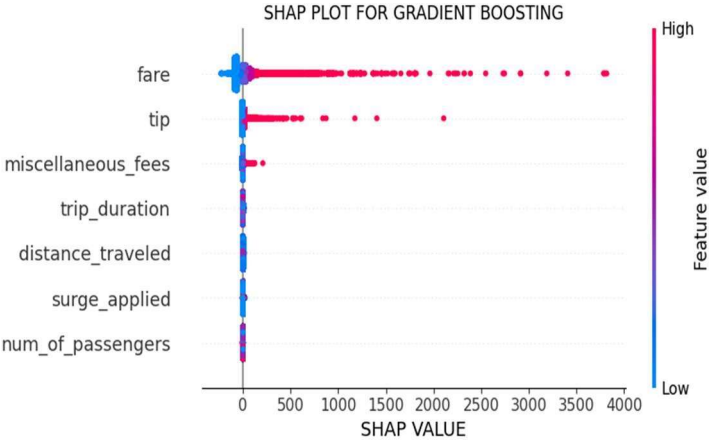


Fig (13)

Fig (13) displays a Shapley Additive Explanation (SHAP) plot of gradient boosting. The features of the model are plotted on Y- axis against the SHAP value plotted on X – axis. From the plot, it can be observed that the features "fare", "miscellaneous_fees" and "tip" plotted on the grid extends from left to right. It can be concluded

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

that the feature "fare" contributes more towards the model output. The SHAP plot is considered as one of the most advanced plots. The algorithm for gradient boosting is as follows:

- The base model has to be chosen for gradient boosting.
- In order to calculate the loss function, stochastic gradient descent can be employed.
- The residual has to be computed by subtracting the Y_test from the predicted values.
- The output can be computed by combining the values of base model.

CAUSAL INFERENCE:

A method employed to determine the cause of certain features and its effects on the model and its performance is called causal inference analysis. There are various kinds through which causal inference of a model can be found. One such type that we are going to be using in our model is called Instrumental Variable Analysis (IV Analysis).

INSTRUMENTAL VARIABLE ANALYSIS (IV ANALYSIS):

Instrumental Variable analysis is primarily used to check for endogeneity in a model. In simpler words, there is a target variable to analyze whether it has any effects on the model performance.

Indirectly, the target variable might not be influential on other variables of the model, but have an effect on the outcome of the model during performance analysis. A method to understand such indirect relationships between a target variable and the model outcome is called as IV Analysis. Similar to how an overfitted or an underfitted model might carry noise that affects the performance of the model, this is a much more advanced version of those methods which estimates the model. For any model, it is important to assess its impact with possible factors. If any factor is omitted, it can lead to a biased result. In order to detect such omitted variable and its effect, IV Analysis is used.

IV-2SLS Estimation Summary

Dep. Variable:	total_fare	R-squared:	0.9874
Estimator:	IV-2SLS	Adj. R-squared:	0.9874
No. Observations:	389395	F-statistic:	7.047e+06
Date:	Sat, Jul 08 2023	P-value (F-stat)	0.0000
Time:	05:11:42	Distribution:	chi2(7)
Cov. Estimator:	robust		

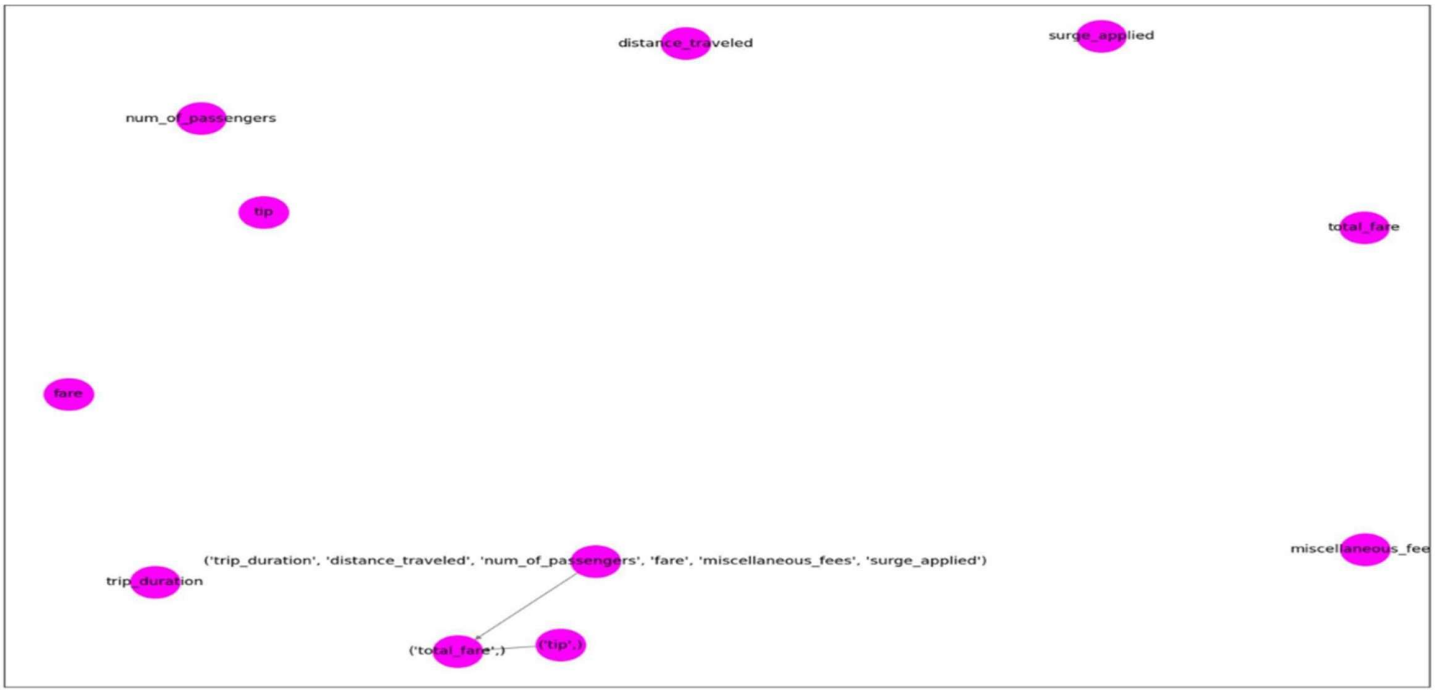
Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
trip_duration	-0.0002	9.82e-06	-22.525	0.0000	-0.0002	-0.0002
distance_traveled	-0.0004	0.0001	-3.1488	0.0016	-0.0007	-0.0002
num_of_passengers	-1.4883	0.0551	-26.997	0.0000	-1.5963	-1.3802
fare	1.1089	0.0027	405.67	0.0000	1.1036	1.1143
miscellaneous_fees	0.3907	0.0181	21.587	0.0000	0.3552	0.4262
surge_applied	3.3068	0.2626	12.591	0.0000	2.7920	3.8215
tip	0.6479	0.0260	24.959	0.0000	0.5970	0.6988

Fig (14)

Fig (14) indicates the IV-2SLS Estimation summary of IV Analysis. The endogenous variable was "tip". The exogenous variable was ["trip_duration", "distance_traveled", "num_of_passengers", "fare", "miscellaneous_fees", "surge_applied"]. The instrument variable was "total_fare". The variable "tip" is also an instrument used in the model. It was found that tip was endogenous on "total_fare" and "num_of_passengers" more. Instrument is significant on the model.

Fig (15) shows the causal diagram for IV Analysis. It shows that there is a dependency of "tip" on "total_fare" and on the variable "num_of_passengers". There is a connected edge between these three variables that shows a relationship. This shows that there is an influence of one variable over the other variables present in the model. Thus, it can be concluded that there is a fine presence of endogeneity in our model as some variable influences the other variables.
Fig (15)



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

INFLUENTIAL ANALYSIS:

Similar to how AutoML is used for feature extraction in order to select the features, Influential Analysis is employed to determine the influential features which have a direct impact on the model performance. This is especially useful for individual feature analysis. Its first function is for outlier detection. When a particular model has missing values or null values, the model performance may be greatly affected. In order to detect such values and prevent the model performance from getting affected, Influential analysis are crucial. There are various types of influential analysis available. But we are going to perform the influential analysis using:

- Cook's Distance
- Hat Matrix

COOK'S DISTANCE:

Developed in 1977, Cook's distance is a measure primarily used to detect features which have a significant impact in the performance of a model. The formula for Cook's distance is:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{(p+1)\,\hat{\sigma}^2}$$

The interpretation of Cook's distance (CD) is: (8)

- Any measure of CD > 1 is considered influential.
- A measure of CD <= 1 is considered non – influential.

It is important to note that any CD > 1 can range to infinity for some models indicating high influence.

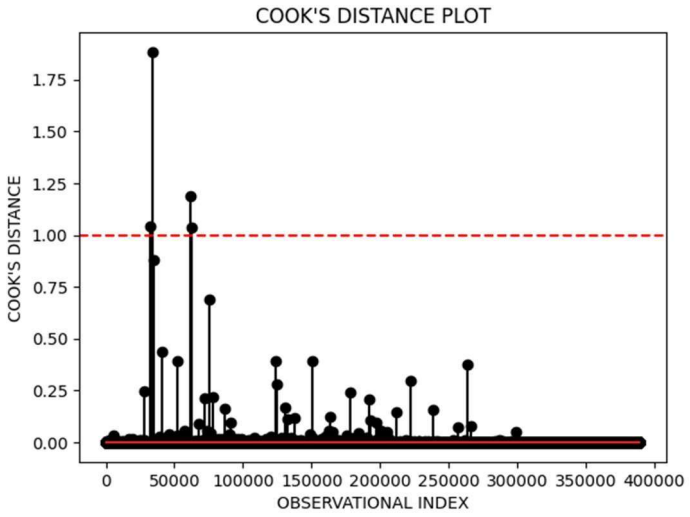


Fig (16)

Fig (16) indicates the Cook's Distance Plot. The observational index on the X – Axis is plotted against Cook's distance on the Y – Axis. The threshold line is denoted by the dotted line. In the plot, the threshold line is touching one exactly. It indicates that there is no potential influence of one feature over any other features in our model. Thus, our model performs well for Cook's distance.

HAT MATRIX:

The Hat Matrix also known as the Projection matrix is used to determine the influence of data points i.e., "hat value" of one feature over the other features present in the model. The only difference between Hat Matrix and Cook's Distance is the calculation of influence of data points over one another. The formula for Hat Matrix is as follows:

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \tag{9}$$

Here,

H = Hat Matrix
X = Matrix I
$\boldsymbol{X}'$ = Matrix II

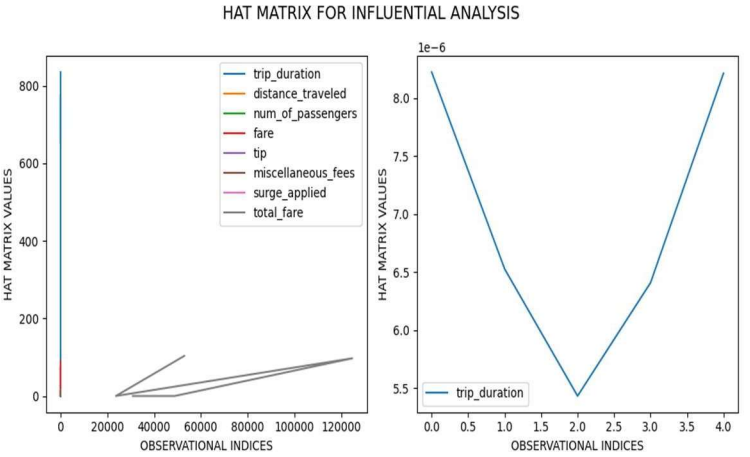


Fig (17)

Fig (17) shows a side-by-side plot of Hat Matrix for Influential Analysis. From the first plot, it is evident that the quantity "trip_duration" from X_train and "total_fare" from Y_train have a significant influence on the model. The plot on the right shows the influence of X_train quantity. It can be concluded that "trip_duration" and "total_fare" have a great influence.

HYPOTHESIS TESTING:

The method of accepting or rejecting a hypothesis based on various tests performed on a model is called as hypothesis testing. Certain kinds of hypothesis testing are dependent on regression analysis and classification analysis respectively. On the other hand, certain kinds of hypothesis testing might be independent of regression and classification analysis and are used just for the purpose of comparison between any quantities present in the model. If any relationship is found on comparison, it would be displayed. Also, there are different kinds of hypothesis testing for regression. The tests which we are going to perform are:

- Durbin – Watson Test
- Breusch – Pagan – Godfrey test

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

DURBIN – WATSON TEST:

Durbin – Watson test is a kind of hypothesis testing employed to determine the level of autocorrelation present in a model. The error terms in our model are Root Mean Squared Error (RMSE) and R2-Score. The Residual is calculated by finding out the difference between Y_test and predicted values. The aim of Durbin – Watson test is to find the correlation between error distributions and the residuals of the model.

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} \quad (10)$$

et = residual at time T

The range of Durbin – Watson (DW) is as follows:

- If DW < 2, it indicates a positive correlation
- If DW > 2, it indicates a negative correlation
- If DW = 2, it indicates no autocorrelation

In our model, the null hypothesis is assumed as "There is no autocorrelation between the error terms and residuals". The alternative hypothesis is assumed as "There is an autocorrelation between the error terms and residuals". Let's check computationally to determine the level of autocorrelation in our model.

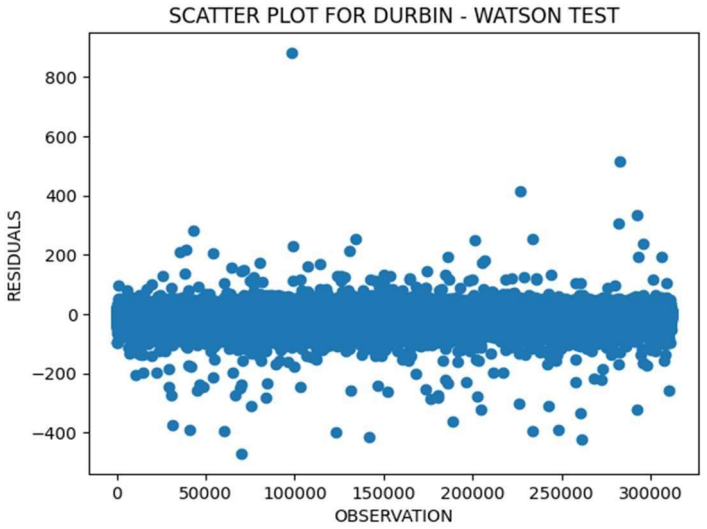


Fig (18)

Fig (18) shows the scatter plot to visualize the Durbin – Watson test. It can be observed that there is no significant pattern exhibited in the plot. The clusters are formed evenly throughout the plane. The clusters do not accumulate at one point. Thus, it can be concluded that there is no autocorrelation. The null hypothesis is accepted. The value of Durbin – Watson measure for our model was found to be 2.004046734474755 which can be estimated to 2.00. Since 2.00 is nearly equal to 2, the null hypothesis is met which states that there is no auto correlation. In order to investigate further, another test is employed the details of which are specified.

BREUSCH -–PAGAN -–GODFREY TEST:

Breusch – Pagan – Godfrey test or the Breusch – Pagan test or the Breusch – Godfrey test is a kind of hypothesis testing used to determine the presence of homoskedasticity or heteroskedasticity in a model. Homoskedasticity refers to the presence of an equal distribution of the residuals. Heteroskedasticity refers to an unequal distribution of the residuals in a model. The null hypothesis is assumed as "There is homoskedasticity in the model". The alternative hypothesis is assumed as "There is heteroskedasticity in the model".

The range of Breusch – Pagan (BP) is as follows:

- If BP > 0.05 where 0.05 is the level of significance, it indicates a heteroskedasticity in the model.
- If BP < 0.05, It indicates a homoskedasticity in the model.

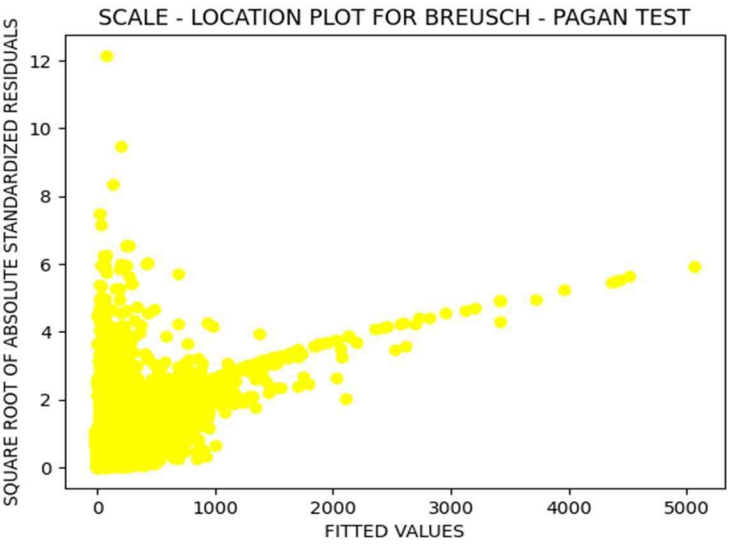


Fig (19)

Fig (19) shows a Scale – location plot for Breusch – Pagan test. It can be observed that there is a horizontal line formed after traversing through the plane from left to right. This indicates that the distribution of residuals is even. It states that there is homoskedasticity in the model. Some points are scattered randomly which also supports the presence of homoskedasticity in the model. Also, the Breusch – Pagan measure for our model was 0.0. Since 0.0 < 0.05, the null hypothesis can be rejected. The alternative hypothesis can be considered. It can be concluded that our model is homoskedastic.

From both of these methods taken for hypothesis testing, we could find that the model was:

- No autocorrelation
- Homoskedasticity, which are good results for a model.

Usually, it is quite difficult to obtain such clean results for any model because there might be some or the other vulnerability which can create an impact in the model performance. But, with these two tests, it can be concluded that the model is free from any external factors that might affect the model during hypothesis testing.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

CORRELATION TESTING:

Like how some analysis such as multi – collinearity analysis and IV Analysis are used to assess the relationship (or in other terms can be termed as correlation) indirectly, Correlation Testing is employed specifically to determine the correlation between any two features. It is usually associated more with skewness and kurtosis. There are various types of correlation testing. But for our model, since it is primarily a regression model, two methods are employed. They are:

- Point – Biserial Correlation
- Pearson's Correlation Coefficient

POINT – BISERIAL CORRELATION:

Point – Biserial Correlation (P – B Correlation) is used to assess the correlation between a binary variable and a continuous variable primarily. In our model, the feature "surge_applied" is the only binary column and "total_fare" is among the many continuous columns. So, these two featured are taken for correlation testing using P -B Correlation.

The P – B Correlation can be interpreted as follows:

- Any P – B Correlation measure that is closer to +1 indicates a positive correlation.
- Any P – B Correlation measure that is closer to -1 indicates a negative correlation.

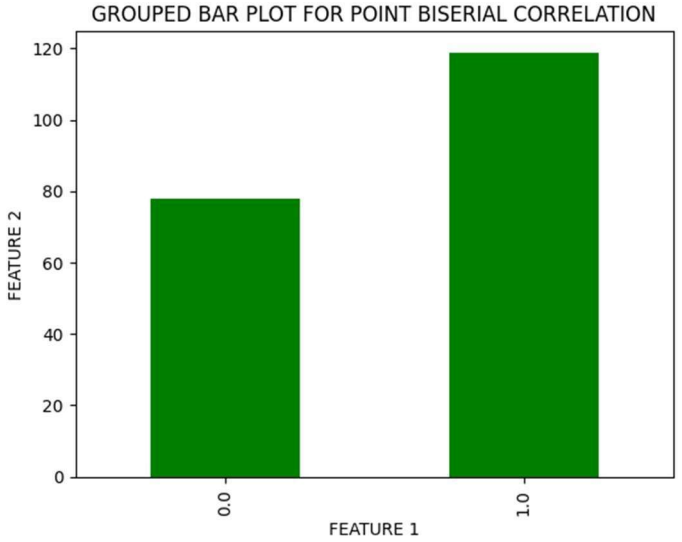


Fig (20)

Fig (20) shows a grouped bar plot for P – B Correlation. As it was earlier mentioned that one feature is a binary feature. They are represented by 0.0 and 1.0 in the X – Axis. Since there is a significant difference between the bars plotted for every group, it is an indication that there is a positive correlation. Not just visually, the correlation measure was also tested statistically. The P – B Correlation measure obtained was 0.2722528955527508 (nearly estimated to 0.27). Since 0.27 is nearly equal to +1, it is an indication of a stronger positive correlation. Further investigations can be done to discover any other significance about the model.

PEARSON'S CORRELATION COEFFICIENT:

Different from how the P – B Correlation analyzed between a binary and a continuous variable, the Pearson's Correlation Coefficient analyzes a linear correlation between any two continuous features in a model. The formula for Pearson Correlation coefficient is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (11)$$

r = Coefficient of Correlation
$\bar{y}$ = Common Mean of Yi
y_i = No. of observations of Yi
x_i = No. of observations of Xi
$\bar{x}$= Common Mean of Xi

The Pearson's correlation coefficient can be interpreted as follows:

- Any Pearson's Correlation Coefficient closer to -1 indicates a negative linear correlation.
- Any Pearson's Correlation Coefficient closer to +1 indicates a positive linear correlation.

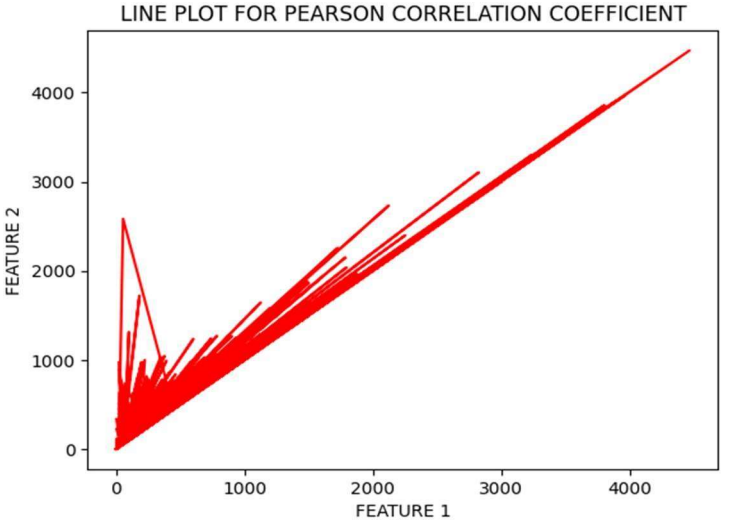


Fig (21)

Fig (2) shows the line plot for Pearson's Correlation Coefficient. It can be observed that the data points are accumulated around the straight line passing from left to right in the plane. This indicates a positive correlation. Statistically, the Pearson's Correlation Coefficient was found to be 0.9790577995022397 (nearly estimated to 0.97). Since 0.97 is nearly closer to 1, it indicates a positive correlation.

Thus, it can be concluded that our model has a presence of positive correlation with respect to certain features. Usually, a low level of positive correlation exists in any model. But, the Coefficient of Correlation measure for P – B Correlation and that of Pearson's Coefficient test indicate s a stronger positive correlation. This type of correlation can be overcome by causal inferences or influential analysis which has been performed before.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

NORMALITY TESTING:

The final step for any regression analysis is the test for normality. The test for normality is defined as the number of residuals found in a model. The residuals can be determined by subtracting the predicted values from actual values.

$$\textit{Residuals} = \textit{Actual Value} - \textit{Predicted Value} \quad (12)$$

There are several kinds of normality testing methods available in Python. But the methods which are used for regression analysis are limited. We are going to perform the normality testing only with two methods. Those are:

- Shapiro – Wilk Test
- Jarque – Bera Test

SHAPIRO – WILK TEST:

A test used to determine the distribution of data in a model is called as the Shapiro – Wilk Test. While dealing particularly with residuals, the purpose of Shapiro – Wilk is to assess the normality distribution of residuals over a model.

The Shapiro – Wilk (SW) measure can be interpreted as follows:

- If the SW is near to 1, it indicates that the distribution of residuals is normal.
- But if the SW is near to 0, it indicates that the distribution of residuals in a model is abnormal.

The null hypothesis for SW test is assumed as "There is a normal distribution of residuals in the model". The alternative hypothesis is assumed as "There is an abnormal distribution of residuals in the model".

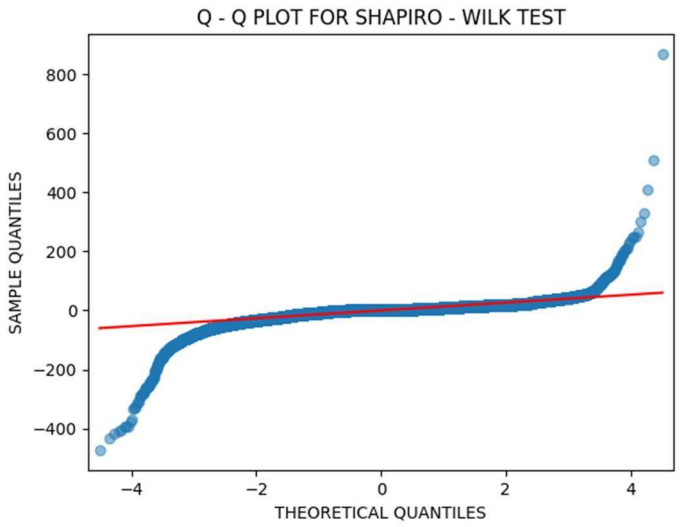


Fig (22)

Fig (22) shows a Quantile – Quantile plot for Shapiro – Wilk test. It can be observed that some data points align over the straight line while some does not. From the plot, the points towards the end raise upwards which shows a negative skewness. Both of these assertions specify that the distribution of residuals is abnormal. Thus, the null hypothesis can be rejected.

JARQUE – BERA TEST:

A test used to predict whether the skewness and kurtosis is normal for the distribution of residuals is called the Jarque – Bera test. Similar to how the SW test is used to determine the skewness, the Jarque – Bera (JB) test is used to additionally give the value of kurtosis. Skewness is used to determine the direction of the residual distribution towards the end of a plane. Kurtosis is used to measure the heavy tails and its distribution on a plane.

The null hypothesis of JB test can be assumed as "The data is normally distributed". The alternative hypothesis can be assumed as "The data is abnormally distributed".

The formula for JB test is:

$$JB = n(S^2/6 + (K-3)^2/24) \quad (13)$$

The JB test can be interpreted as follows:

- If P – Value is extremely high, it indicates a normal distribution.
- If P – Value is extremely low, it indicates an abnormal distribution.

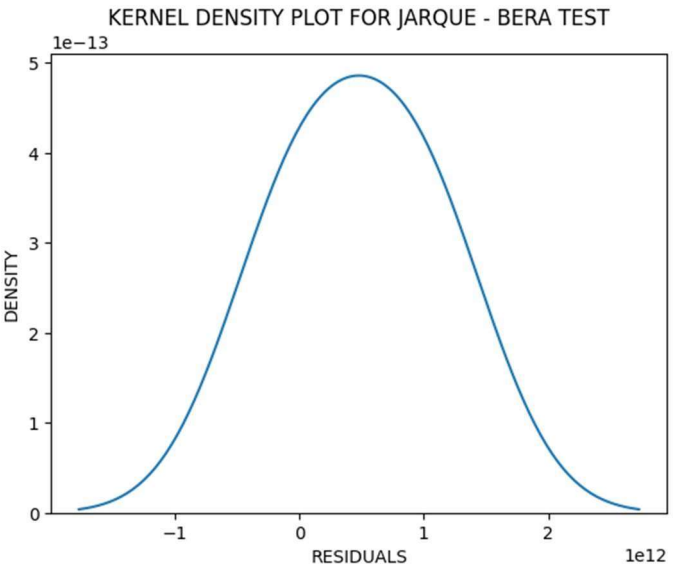


Fig (23)

Fig (23) shows the Kernel density estimation plot for Jarque – Bera test. It can be observed that the kernel density curve is bell – shaped and symmetrical. There are no crests or troughs in the plot. Despite the graph being symmetrical and bell – shaped, there might be little amount of skewness and kurtosis depicted as the graph does not incline anywhere towards the end. The P – Value on calculation was found to be (0.0 < 0.05). A lower P – Value indicates an abnormal distribution. Thus, the null hypothesis can be rejected. Even if the plot came in favor, statistical significance contradicted.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

III. RESULTS AND DISCUSSIONS

The model was successfully tested with various analysis and testing methods. Whilst being vulnerable to certain analysis, the model performed well in others. The goal of this paper is to design an efficient model that would help in the planning and management of transportation systems in India. In order to prevent traffic safety violations and accident occurrences, the model is analyzed in depth to give a productive and a cost – effective solution. The results for data pre – processing and data visualization are as follows:

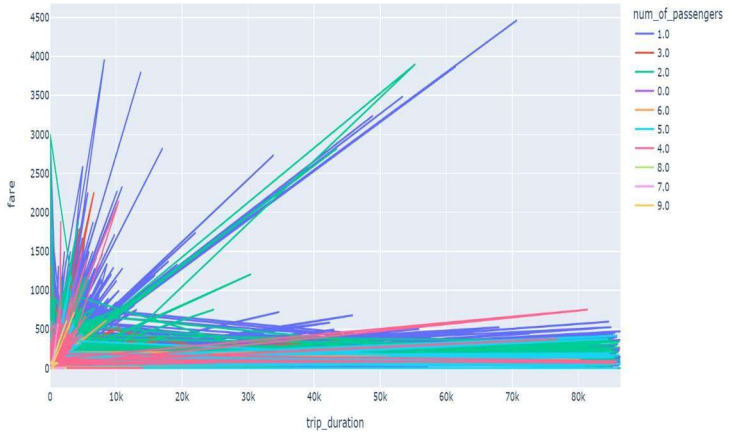


Fig (24)

Fig (24) shows a line plot with "trip_duration" in X – Axis and "fare" in Y – Axis. The "num_of_passengers" is taken as a third parameter based on which the trip has been divided. From the graph, it can be observed that Passengers of Group 1.0 who had travelled for around 70000 kms had paid the highest fare ranging up to Rs. 4500.

Fig (25) shows a pie chart that indicates the time taken by the passengers to travel from source to the destination. It can be observed that the Passengers from group 1.0 consume the maximum time to travel because of the long distance. In contrast, Passengers of Group 9.0 had paid the lowest fare due to their shorter distance covered during the journey.

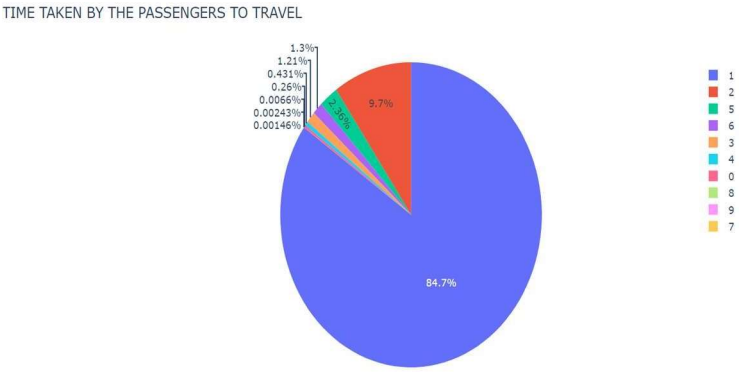


Fig (25)

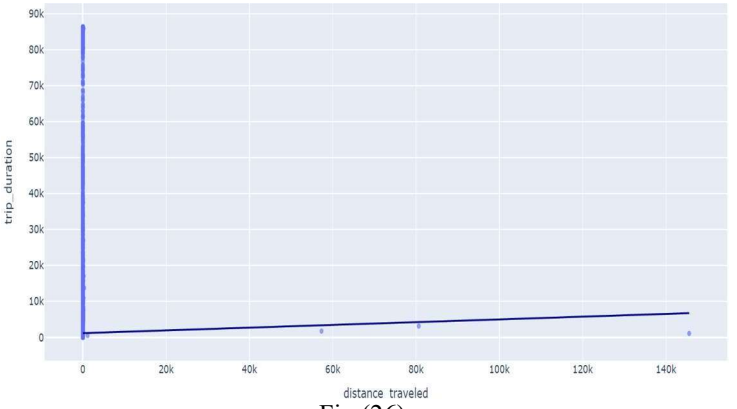


Fig (26)

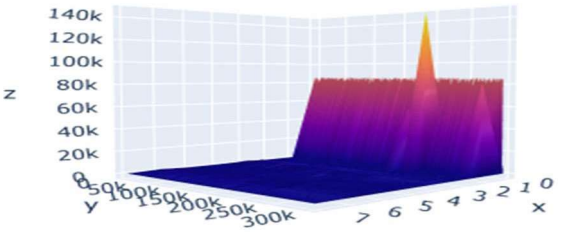


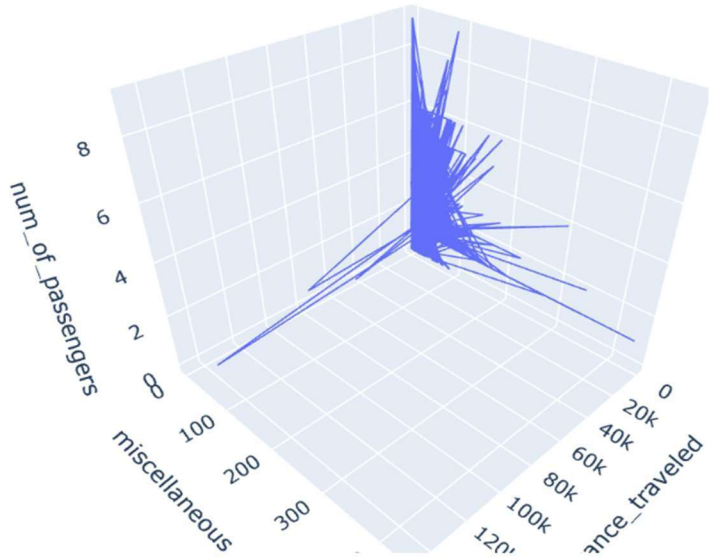
Fig (27)

Fig (26) shows the relationship between "trip_duration" and "distance_traveled" through a scatter plot. It can be observed that the duration of a trip extends up to 90,000 minutes. There are only four data points which are scattered randomly that create a non – linear relationship between the two parameters.

Fig (27) shows a Camera Control Graph plotted between three quantities which are "num_of_passengers", "distance_traveled" and "total_fare". It can be observed that the purple color emerging through the graph represents Group 1.0 that had travelled the maximum in terms of distance and time taken.

Fig (28) shows a 3D line plot that takes the variables "distance_traveled" on X – Axis, "miscellaneous_fees" on Y – Axis and "num_of_passengers" on Z – Axis. It was found that Passengers of Group 1.0 had paid a miscellaneous fee of Rs. 72 for

Fig (28)



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

a total maximum distance of 145. 5176 kms.

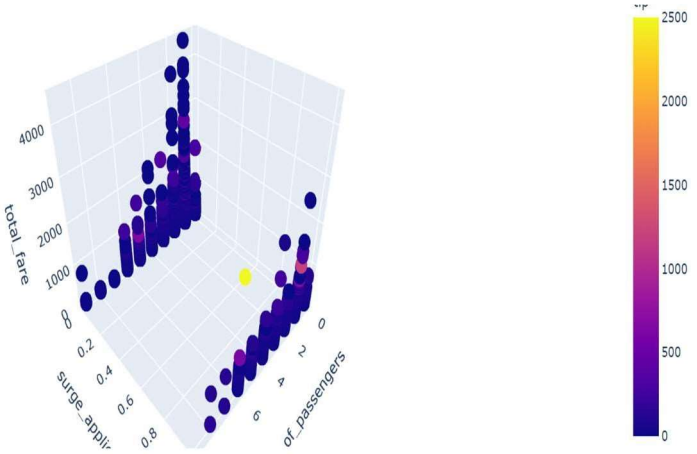


Fig (29)

Fig (29) shows a 3D scatter plot which takes the variables "num_of_passengers", "surge_applied" and "total_fare". The color bar shown in the right shows the range of tip applied to the trip from (0, 2500). It can be observed that Passengers of Group 6.0 had given the maximum tip of Rs. 2500. The Passengers of Group 1.0 had given the least tip amounting to Rs.0.

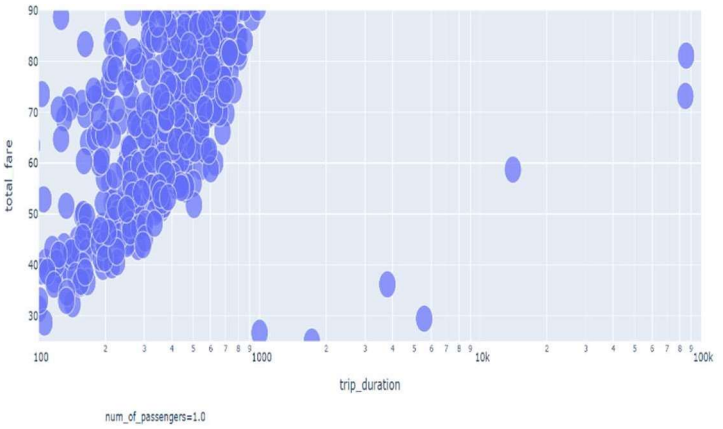


Fig (30)

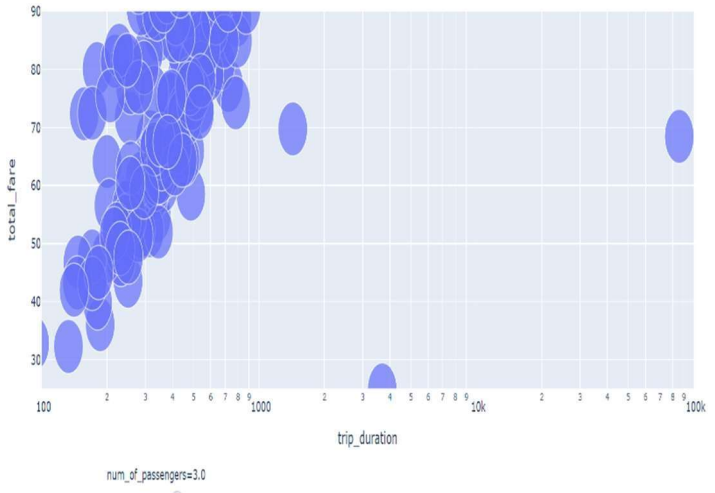


Fig (31)

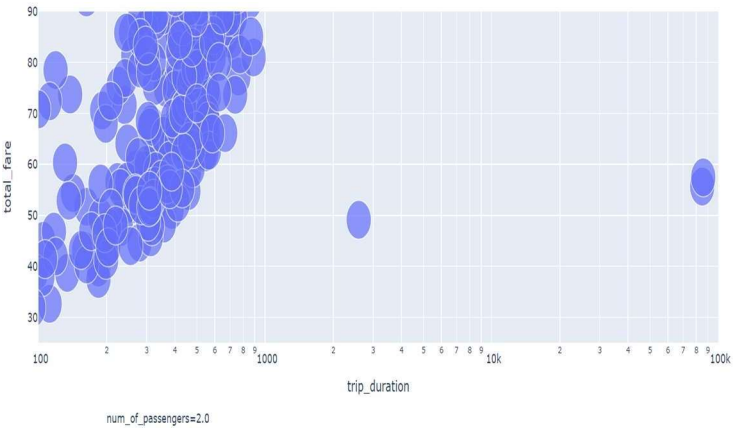


Fig (32)

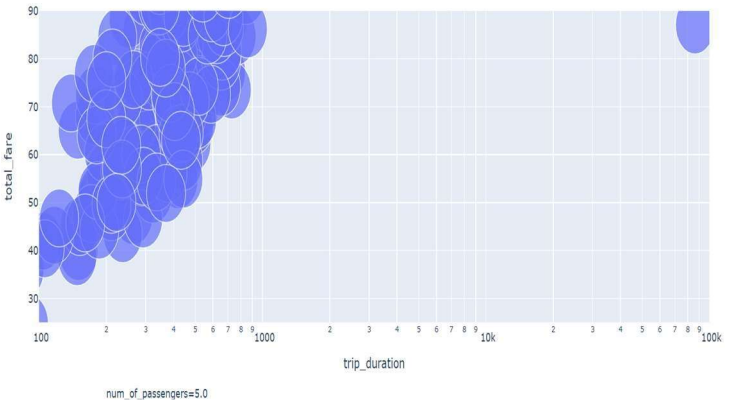


Fig (33)

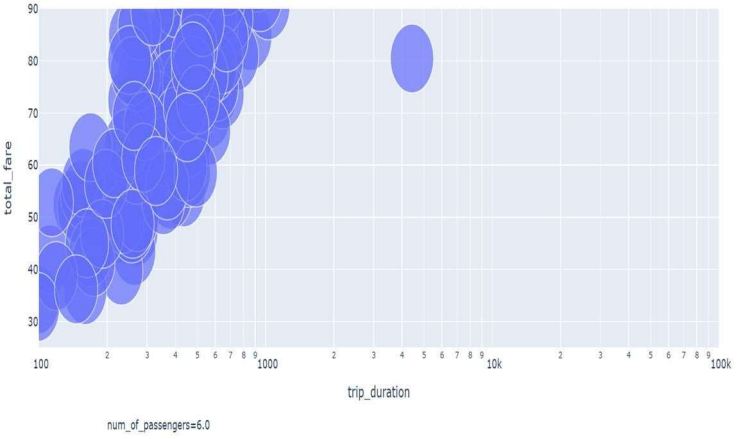


Fig (34)

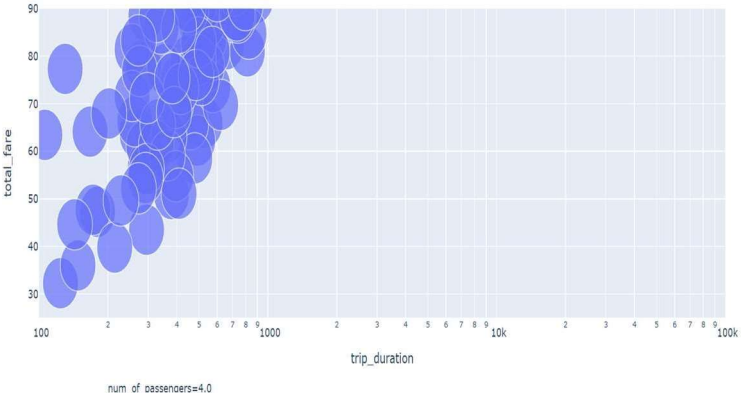


Fig (35)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

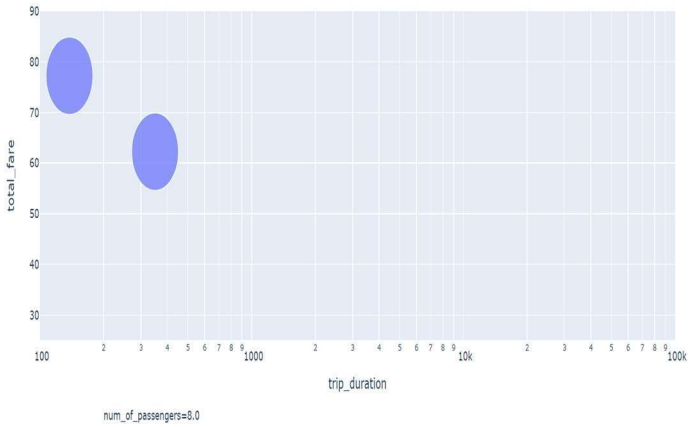


Fig (36)

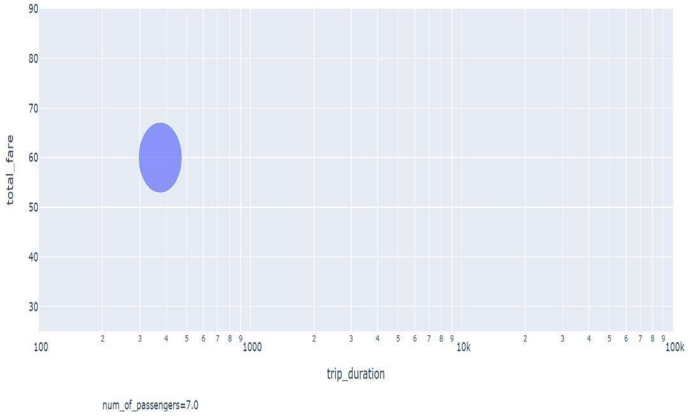


Fig (37)

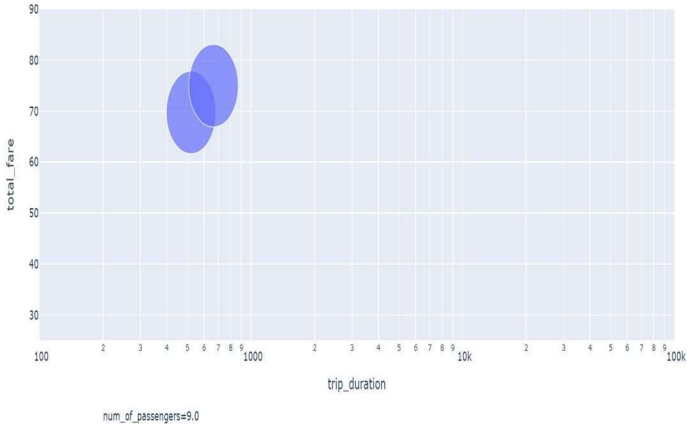


Fig (38)

Fig (30), (31), (32), (33), (34), (35), (36), (37) and (38) shows a decomposed bubble plot indicating the relationship between "num_of_passengers", "trip_duration" and "total_fare". The number of passengers in each group starting from Fig (28) to Fig (36) starts decreasing in the order of Group 1.0 to Group 9.0. From the plots, it can also be observed that the number of passengers of Group 1.0 were maximum who preferred to travel by one particular mode of transport. In contrast, there were only two passengers of group 9.0 who preferred to travel by another mode of transport.

The model was initially trained with four learning algorithms primarily. Those were Linear Regression, Random Forest, Decision Tree and Neural Networks. Let's check the table given in Fig (37) to determine which technique has given the best result.

LEARNING ALGORITHM	RMSE	R2 - SCORE
LINEAR REGRESSION	12.96	0.98
RANDOM FOREST	12.31	0.98
DECISION TREE	4.66	0.99
NEURAL NETWORKS	37.5	NIL

Fig (39)

Fig (39) gives the comparison of model metrics such as Root Mean Squared error (RMSE) and R2 – Score. Any RMSE score which is closer to 0 indicates a good score. If the score is away from zero by a random X value, it means that the performance is X points away from the actual value. Eventually, this will create a big difference between the actual values and the predicted values. On the other hand, the best R2 - Score is any score that is closer to one or is equal to one. From the figure, it can be understood that the best performing model is "Decision Tree". This is because the predicted value has less deviation from the actual value in terms of RMSE Score. In terms of R2 – Score, the value i.e., 0.99 is nearly closer to one which indicates a good score. Since Decision Tree performed well in both the metrics in terms of comparison, it is regarded as the best algorithm for the model.

Similar to how the model comparison was done for learning algorithms, let us check the comparison for ensembling algorithms.

ENSEMBLING TECHNIQUES	RMSE	R2 - SCORE
BAGGING ENSEMBLING	4.33	0.99
STACKING ENSEMBLING	4.5	0.99
BOOSTING ENSEMBLING	4.25	0.99

Fig (40)

Fig (40) shows the results of model performance with ensembling techniques. It can be observed that "Boosting Ensembling" has given the best result. This is because it shows less deviation of predicted values from actual value in terms of RMSE and an R2 – Score nearly touching one which is a best score. Thus, boosting ensembling has performed well.

MULTI - COLLINEARITY ANALYSIS TECHNIQUES	RMSE	R2 - SCORE
RIDGE REGRESSION	12.96	0.98
LASSO REGRESSION	13.74	0.98
ELASTICNET REGRESSION	12.99	0.98

Fig (41)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

Fig (41) shows the tabular representation of the multi – collinearity analysis techniques performed in our model. Only the three above mentioned techniques were taken for the first evaluation. It can be observed that the technique that has given the best result was Ridge Regression. This is because it shows less deviation in comparison of predicted values from actual values in terms of RMSE score. It gives an R2 – Score of 0.98 i.e., nearly equal to one. Thus, it is considered as a best fit model. Overall, the best performing model amongst the three techniques taken for Evaluation I was "Ridge Regression".

MULTI - COLLINEARITY ANALYSIS TECHNIQUES	SCORE
VARIANCE INFLATION FACTOR	34.518047
TOLERANCE	0.02897

Fig (42)

Fig (42) indicates the VIF score and the tolerance score for multi – collinearity assessment. It can be observed that the results generated from the analysis techniques given in Fig (41) are contradicting to the Fig (42). There were no vulnerabilities found in those three techniques. But when VIF and tolerance were analyzed, it was found that VIF for two features "fare" and "total_fare" was 34.518047 which is significantly higher. This indicates a high level of multi – collinearity. In order to justify the same, tolerance was employed as an additional measure. It can be observed that the level of tolerance is 0.02897 which also indicates a high level of multi – collinearity. In order to overcome this vulnerability of multi – collinearity, Ridge regression and Lasso regression were employed. On employing, the results generated were satisfactory.

The next step of analysis was the Instrumental Variable Analysis (IV Analysis). In the course, there was an endogeneity found in our model. The instrumental variable that was responsible to introduce the endogeneity was "tip". It had effect on two features such as "total_fare" and "num_of_passengers". The summary of the IV analysis can be found in Fig (14). Fig (43) shows the instrumental variable "tip" on "num_of_passengers" and "total_fare". It can be concluded that there is a dependency of tip on the number of passengers and the total fare during the estimation.

INSTRUMENTAL VARIABLE ANALYSIS	INSTRUMENTAL VARIABLES
CAUSAL INFERENCE	['tip', 'num_of_passengers', 'total_fare']

Fig (43)

INFLUENTIAL ANALYSIS	SCORE
COOK'S DISTANCE	1.0

Fig (44)

After IV Analysis, Influential Analysis was performed using two methods i.e., Cook's Distance and Hat matrix. Fig (43) shows the value of Cook's distance obtained through statistical testing. A Cook's distance value of 1.0 indicates a good measure. It can also be observed in the threshold line given in Fig (16). Thus, it can be concluded that the there is no potential influence of any feature over any other features.

Fig (45)

INFLUENTIAL ANALYSIS	INFLUENTIAL HAT VALUES
HAT MATRIX	['trip_duration', 'fare', 'total_fare']

But this was not the same case with the hat matrix. Hat matrix shows a contradictory result to what was investigated in Cook's distance. Fig (45) shows that influential data points in our model through hat matrix. The hat values that influenced the other features present in our model were "trip_duration", "fare" which were values of the X_train. Similarly, "total_fare" was the feature of Y_train. In these influential hat values list, "trip_duration" upon analysis was found to have a greater influence on "total_fare" when compared to the influence of "fare" and "total_fare".

After the analysis, the last step to any machine learning project would be testing. Our model was first given for hypothesis testing using two prominent tests which were Durbin – Watson and Breusch – Pagan – Godfrey tests. The results of the hypothesis testing can be found in Fig (46). A DW Measure of 2.00 (nearly to 2) indicates the absence of autocorrelation in our model. Similarly, a BPG measure of 0.0 indicates the presence of homoskedasticity in our model.

HYPOTHESIS TESTS	TEST MEASURES
DURBIN - WATSON	2.0
BREUSCH - PAGAN - GODFREY	0.0

Fig (46)

It is very crucial to check for the skewness and kurtosis is any model. It can be checked either using correlation testing or normality testing. Let's discuss first about the correlation testing.
Fig (47) gives a table for the results obtained in the correlation testing. The model was tested using P – B Correlation and Pearson's Correlation Coefficient. A value of 0.27 in P -B Correlation and a value of 0.97 indicates the presence of a positive correlation. In order to investigate this vulnerability, the model was tested for IV Analysis and Influential analysis.

CORRELATION TESTS	TEST MEASURES
POINT - BISERIAL CORRELATION	0.27
PEARSON'S CORRELATION COEFFICIENT	0.97

Fig (47)

The last part of our model testing is the normality testing. The normality test is primarily used to determine the distribution of residuals in a model but it is also a good measure for skewness and kurtosis mentioned earlier. In order to check for the distribution, the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

model was tested using two methods. Those were Shapiro – Wilk test and Jarque – Bera test. Let's see the results of the normality testing.

NORMALITY TESTS	TEST MEASURES	P - VALUE
SHAPIRO - WILK TEST	0.03	0.0
JARQUE - BERA TEST	230186797	0.0

Fig (48)

Fig (48) shows the table of results obtained. It can be seen that the test measure for Shapiro – Wilk test is 0.03 (nearly equal to zero) which is an indication of abnormal distribution of residuals. Similarly, for Jarque – Bera test, the P – Value is taken more than the JB test measure. The P – Value is 0.0 (< 0.05) indicating an abnormal distribution of residuals. From testing with both the methods, we could conclude that the pattern of residuals follows an abnormal distribution. The entire model can be summarized for regression analysis.

OLS Regression Results

Dep. Variable:	total_fare	R-squared:	0.981
Model:	OLS	Adj. R-squared:	0.981
Method:	Least Squares	F-statistic:	2.283e+06
Date:	Tue, 11 Jul 2023	Prob (F-statistic):	0.00
Time:	03:39:20	Log-Likelihood:	-1.2484e+06
No. Observations:	311516	AIC:	2.497e+06
Df Residuals:	311508	BIC:	2.497e+06
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.9583	0.038	-51.327	0.000	-2.033	-1.883
x1	-0.0002	5.87e-06	-36.101	0.000	-0.000	-0.000
x2	-0.0011	0.000	-6.676	0.000	-0.001	-0.001
x3	-0.8909	0.026	-34.691	0.000	-0.941	-0.841
x4	1.1111	0.000	3370.337	0.000	1.110	1.112
x5	0.6586	0.001	457.571	0.000	0.656	0.661
x6	0.4298	0.003	141.381	0.000	0.424	0.436
x7	2.9299	0.086	34.040	0.000	2.761	3.099

Omnibus:	223320.589	Durbin-Watson:	2.004
Prob(Omnibus):	0.000	Jarque-Bera (JB):	230186797.002
Skew:	-2.067	Prob(JB):	0.00
Kurtosis:	136.106	Cond. No.	1.52e+04

Fig (49)

Here, the independent variables [x1, x2, x3, x4, x5, x6, x7] represents the columns ["trip_duration", "distance_traveled", "num_of_passengers", "fare", "tip", "miscellaneous_fees", "surge_applied]. The column "total_fare" is a dependent variable. The model used to generate the regression summary is OLS. The proportion of variance put by the independent variables on the dependent variable comes up to 98.1%. The F – Statistic is very big. Its corresponding P – Value is 0.0. This indicates that our model is a significant model. The P – Value corresponding to T – statistic accounts to 0.0 which indicate that the model is significant. The Omnibus value is 223320.589 which states that there is an abnormal distribution of residuals. Similar to the result observed in Shapiro – Wilk test, the skewness which is -2.067 indicates a negative skewness where the tail is towards the upward direction. There is a high kurtosis observed from the summary. Overall, after testing the model for multiple analysis and tests, it can be concluded that certain factors have an influence either directly or indirectly on the model impacting the performance.

IV. CONCLUSION

The model was successfully tested with various analysis and techniques. The goal of the research paper was complete with various factors identified that came either in favor or against the idea. The factors that came in favor can be accepted while those that came against the proposed idea were overcome by applying various techniques to reduce the vulnerability caused.

V. FUTURE ENHANCEMENTS

In the future, the model can be extended to a deep learning and a reinforcement model enabling to undergo various tests to determine the optimal solution. The data can be thoroughly tested with real time tools either in terms of hardware of software for practical usage.

VI. REFERENCES

[1] Parth Bhavsar, Ilya Safro, Nidhal Bouaynaya, Robi Polikar, Dimah Dera. Machine Learning in transportation data analytics. Retrieved from: https://doi.org/10.1016/B978-0-12-809715-1.00012-2

[2] Ashokkumar Palanivinayagam, Claude Ziad El-Bayeh, Robertas Damaševičius. Twenty Years of Machine-learning-Based Text Classification: A Systematic Review. Retrieved from https://doi.org/10.3390/a16050236

[3] Pantelis Linardatos, Vasilis Papastefanopoulos, Sotiris Kotsiantis. Explainable AI: A Review of Machine Learning Interpretability Methods. Retrieved from https://doi.org/10.3390/e23010018

[4] Józef Gorzelany, Justyna Belcar, Piotr Kuźniar, Gniewko Niedbała, Katarzyna Pentoś. Modelling of Mechanical Properties of Fresh and Stored Fruit of Large Cranberry Using Multiple Linear Regression and Machine Learning. Retrieved from: https://doi.org/10.3390/agriculture12020200

[5] Mehdi Zamani Joharestani, Chunxiang Cao, Xiliang Ni, Barjeece Bashir and Somayeh Talebiesfandarani. PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. Retrieved from: https://doi.org/10.3390/atmos10070373

[6] Hamed Dabiri, Visar Farhangi, Mohammad Javad Moradi, Mehdi Zadeh Mohamad and Moses Karakouzian. Applications of Decision Tree and Random Forest as Tree-Based Machine Learning Techniques for Analyzing the Ultimate Strain of Spliced and Non-Spliced Reinforcement Bars. Retrieved from: https://doi.org/10.3390/app12104851

[7] Frank Emmert-Streib and Matthias Dehmer. High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection. Retrieved from: https://doi.org/10.3390/make1010021

[8] Wentao Wang, Jiaxuan Liang, Rong Liu, Yunquan Song, Min Zhang. A Robust Variable Selection Method for Sparse Online Regression via the Elastic Net Penalty. Retrieved from: https://doi.org/10.3390/math10162985

[9] Bahareh Kalantar, Husam A. H. Al-Najjar, Biswajeet Pradhan, Vahideh Saeidi, Alfian Abdul Halin, Naonori Ueda, Seyed Amir

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 07, July-2023

Naghibi. Optimized Conditioning Factors Using Machine Learning Techniques for Groundwater Potential Mapping. Retrieved from: https://doi.org/10.3390/w11091909

[10] Karoline S. Sauer, Stefanie M. Jungmann, Michael Witthöft. Emotional and Behavioral Consequences of the COVID-19 Pandemic: The Role of Health Anxiety, Intolerance of Uncertainty, and Distress (In)Tolerance. Retrieved from: https://doi.org/10.3390/ijerph17197241

[11] Himan Shahabi, Ataollah Shirzadi, Kayvan Ghaderi, Ebrahim Omidvar, Nadhir Al-Ansari, John J. Clague, Marten Geertsema, Khabat Khosravi, Ata Amini, Sepideh Bahrami, Omid Rahmati, Kyoumars Habibi, Ayub Mohammadi, Hoang Nguyen, Assefa M. Melesse, Baharin Bin Ahmad and Anuar Ahmad. Flood Detection and Susceptibility Mapping Using Sentinel-1 Remote Sensing Data and a Machine Learning Approach: Hybrid Intelligence of Bagging Ensemble Based on K-Nearest Neighbor Classifier. Retrieved from: https://doi.org/10.3390/rs12020266

[12]Federico Davina, Aude Gilson, Francisco Goméz-Vela, Miguel García Torres and José F. Torres. Stacking Ensemble Learning for Short-Term Electricity Consumption Forecasting. Retrieved from: https://doi.org/10.3390/en11040949