Traffic Flow Analysis Using Machine Learning

Mohammad Maaz Rashid
School of Computing
SRM Institute of
Science and
Technology
Chennai, India
mr8775@srmist.edu.in

Imaad Khan
School of Computing
SRM Institute of
Science and
Technology
Chennai, India
ik3746@srmist.edu.in

Dr T.K Sivakumar
Assistant Professor
School of Computing
SRM Institute of
Science and
Technology
Chennai, India
sivakumt2@srmist.edu.in

***Abstract*— Prediction and analysis of the traffic flow is a subject of utmost importance. The traffic department and Governments could make use of the crucial data and key-points to interchange the vehicular routes and expedite the traffic movement in a smooth and effective manner. This data can prevent congestions in the near future and help scale down the overall travelling time for an individual. Human lives can also be saved as the emergency facilities such as the ambulances, fire-fighting vehicles etc. would reach their respective emergency locations without any further delays. In order to control and eliminate the congestions and the gridlocks in a controlled and effective way, prediction and analysis of the traffic is paramount. This research paper will present various ML models such as Gradient boosting, Decision trees and a comparison will be drawn between them and the currently in-use primitive models like Linear Regression by utilizing different performance benchmarks such as the RMSE values.**

Keywords*— *Traffic-flow | Decision Tree| Linear regression | Gradient Boosting |Root mean square error (RMSE).

I. INTRODUCTION

Vehicular routes statistics and the related data to the road network is of utmost value for designing transport activities and other associated research activities. According to a study, commuters ordinarily spent over 79 hours caught in the traffic in 2017 alone in some parts of the world. Analyzing the vehicular movements helps in ascertain the commute of vehicles through various roads and re-routing them to different lanes and roads to reduce the occurrences of congestions. All the current and previous records and statistics pertaining to the vehicular traffic for a region can be utilized to identify the various road networks and the traffic patterns to predict the flow of the imminent and upcoming traffic.

Traffic congestion reduces air pollution in the area as well as increased traffic pollution, and recent studies have shown that the death toll for motorists, commuters and people living near highways and traffic congested areas is very high. Our current knowledge of air pollution and its impact on traffic congestion is insufficient. Therefore, the study of motor prediction methods is very important in alleviating this difficulty. This can help traffic controllers control traffic. This exchange increases the time required for the trip and thus forces the fare to rise.

The first step is to accumulate all the traffic-related data for analysis. There are numerous approaches for gathering the data. To collect the data, various detectors and equipments are instated throughout different road networks to estimate the volume of traffic on a road at a particular instance. Equipments such as personal road courses, test vehicles or floating vehicle data (FCD), sidewalk detectors, closed-circuit television (CCTV), camera, photographs, are commonly used. In this study, we attempted to gather relevant data needed to predict traffic flow and consider the type of traffic that exists in India.

II. LITERATURE SURVEY

A. Over the few decades, various discrete approaches have unfolded many solutions to predict the traffic flux precisely. Initially, a regression perspective was put-forth by Mr. Liliian. In this method, one was able to forecast the traffic flow precisely by taking into account various attributes from a traffic dataset of

Southern China. Another prototype introduced by Mr. Fieng suggested a representation using the overall summation of traffic flow. Data and useful insights were gathered at the road intersections using different devices. Yet another approach introduced by Mr. Shinmei implied forecasting the traffic flux using the K-Means model. Mr. Guowandai put-forth a framework that focused primarily on the temporal aspect along-with the "GRU". The G-recurrent unit utilizes the Spatial-Temporal to forecast the overall traffic flux.

B. Another interesting approach devised by Mr. Rong nyao suggested applying the Markov Method Technique to forecast the flux accurately. Mr. Li-Chang's research utilized a hybrid-prediction approach that consisted of a SVM along-with a random forest model to only use the most important and useful information to ascertain the most ideal and peak predictive-attributes.

Other models consist of applying a flux approach on a car-orientation dataset where the spatial-likelihood dispersal of the vehicles is delineated. Mr. Ylxvan prioritized using a LSTM framework to forecast the traffic flux along the different lanes of the roads. Mr. Chinai introduced another model consisting of radial Attributes to accurately forecast the velocities and the vehicular blockage. In order to remove the outliers and to even out all the noises, Mr. Xinqun presented a framework that utilized discrete techniques and schemes and also consisted of a LSTM framework to better forecast the obstructions on the roads.

C. The already existing ITS's (Intelligent Transport System) can be further tuned for them to make better choices in terms of vehicular routing and blockage reduction. Keeping this in mind, two frameworks have been devised by Mr. Bukersche. One is based on a statistical approach and the other is dependent on ML algorithms. LSTM frameworks can be further enhanced using Temporal convolutional context blocks. This was theorized by Mr. Hakaung in which he used the loss switches in the whole architecture. Frameworks integrating two models was first devised by Siaqumn. This model consisted of merging of the ARIMA and the LSTM frameworks. Obviously after the collaboration, the overall model gave improved outcomes. Feature enhancements can make predicting the flux for short intervals more accurate. Mr. Linjaingh exploited this fact and came up with a G-Boosting RT model to make it possible. The outcomes were far more efficient and gave improved traffic flow forecasting.

III. PROPOSED WORK

To accurately forecast the vehicular movements and predict the road blockages, many techniques have been proposed previously. We will be implementing certain Machine-learning algorithms in this project. For our model, we would utilize a dataset from "www.kaggle.com" with various significant attributes. We will only be implementing those algorithms that would be compatible with our dataset. The dataset that we have chosen and the pre-processing that will be performed on it will heavily influence the overall results of the prediction model.

Taking note of the points mentioned above, we would be implementing different ML-algorithms and compare and contrast their performances using different metrics. Eventually after comparing the algorithms, we will select the most favorable algorithm and continue to develop the prediction model.

A. Dataset

We will be utilizing a dataset from Kaggle.com. It contains the information regarding the Metro Interstate Traffic Volume. It contains a rich amount of significant information with various attributes pertaining to weather, day-offs, festivals, population etc. The raw data has been collected from the vehicles travelling from the area of Minneapolis towards Saint Paul. Around 48,000 instances and 9 attributes of the data have been specified in this dataset.

B. Data Pre-processing

Consequent to choosing a dataset, the pre-processing and data cleaning begins. For this purpose, we chose python and some of its significant modules and libraries. For instance, Numpy implements different mathematical computations throughout the algorithm and offers the functionality of implementing multi-dimensional arrays, Pandas assists in transformation and scanning of the dataset along-with the ability of representing data in the form of data-frames, Matplotlib creates immersive visualizations of the data and many more. Pre-processing is spread out in numerous phases. No dataset is completely flawless and inconsistencies are always present, hence pre-processing of the dataset initially is of paramount significance. Some of the steps implemented while pre-processing the data involve dropping of redundant and non-essential columns, discovering the missing values and deleting the rows and columns containing null values, checking the correlations amongst different attributes and dropping the columns that are non-significant in the forecasting.

Pre-processing also involves changing the formatting of the values of certain columns so that the ML algorithms can identify and read them. For instance, for our model, the data in the weather column had to be converted to a numerical format in order for the algorithm to utilize it.

Another instance where the pre-processing is significant is when certain data fields of the rows contain a null or an infinity value. Usually, these values are substituted with different entries. Null values have to be substituted with mean values and the infinity values are removed in favour of priorly-calculated maximum values.

C. Architecture

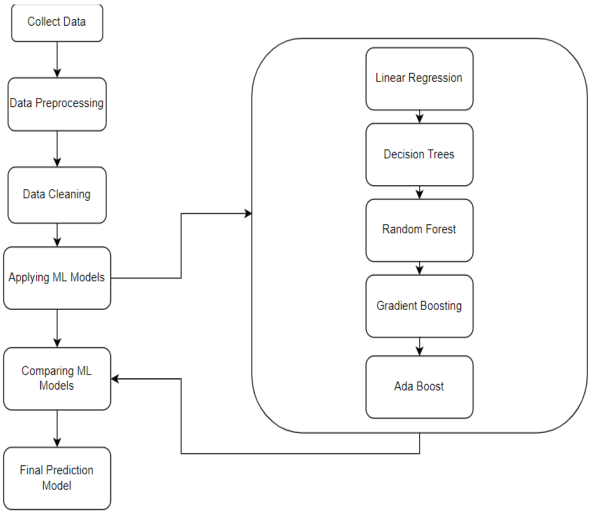


Fig 1 – Architecture Diagram

D. Machine Learning Algorithms

• ***Linear Regression****:* It is the most basic algorithm of supervised learning technique. In this algorithm, predictions are made for continuous attributes. Basically, the relationship between an independent and a dependent attribute is plotted on a 2-dimensional plane and the optimum fit-line that passes through maximum points is found. This line is ultimately used to predict different values.

• ***Decision Tree****:* Another type of supervised learning technique which uses a tree-structured distribution. The nodes illustrate the characteristics of a dataset, branching illustrate the resolution rules and the leaf nodes illustrate an outcome. The tree is traversed from the root to the leaf nodes and all the nodes are organized simultaneously. D-Trees forecasts the final attribute values by implementing discrete decision rules.

• ***Random Forest:*** Random Forest algorithm is based on the concept of ensemble learning. It is a type of supervised learning technique. It involves merging of all the individual decision trees and thus obtain the most optimum prediction model. All the votes from different decision trees are accumulated to achieve the final output. The average of all the outcomes from individual decision trees is taken to improve the overall predictive precision for a particular dataset.

• ***Gradient Boosting:*** Boosting is one of the most powerful algorithms in Machine-learning. Gradient Boosting, a type of boosting model works by minimizing the error in each and every step. Each predictor combines itself with its predecessor to minimize the overall residual error. Gradient boosting is a sequential ensemble technique where the overall accuracy is optimized over consecutive iterations. The gradient boosting algorithm consists of certain components like a loss function, weak learners and strong learners.

• ***AdaBoost:*** AdaBoost is another type of boosting algorithm where uniform significance is assigned to each of the observations at the start. Shortly afterwards assessing the 1st tree, the importance of other observations is incremented and for the observations which are simply classified, their importance is decremented. Now these latest sets of weighted information are used in constructing of a redesigned tree. Subsequently, the newer framework becomes an aggregate of the 1st and the 2nd tree. Hence this newer model is able to enhance the overall precision of the predictions produced by the 1st tree. Only certain crucial and dominant attributes are selected in this algorithm to generate convincing outcomes. Ada-boost functions mostly by transforming weak learners to strong learners.

IV. EVALUATION METRICS

The different parameters utilized for comparing and evaluating several algorithms:

- Root Mean Square Error (RMSE): RMSE provides approximation up-to a certain value. Accuracy score equation is depicted:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

In the above equation, (y_i) represents the true-label, ($y_i^{\wedge}$) represents the predicted label. In case of a multi-label classification, the RMSE value denotes the precision of the subset. The RMSE value is exactly 1.0 when the predicted values are equivalent to the true-value while as the value is 0 when they are not equivalent.

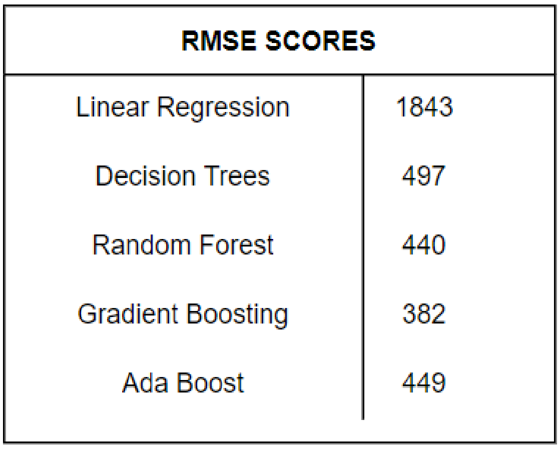
- Coefficient of Determination: Denoted by (R^2), it is a very essential statistical measurement utilized by various algorithms. It presents the relationship between the independent and the dependent variable by comparing their variances. This metric is a crucial measurement technique that informs of how effectively the figure fits the model and how effectively it equals with the real figures. The values it can accommodate varies from 0 and 1.

$$\text{Coefficient of Determination } (R^2) = 1 - \frac{SS_{\text{regression}}}{SS_{\text{total}}}$$

V. RESULT

Various machine learning models were implemented and their outcomes were collocated to discover the eminent working model. Root mean square error value of 1843 is given by Linear regression. The coefficient of determination is calculated as 0.144.

Accuracy score of 93 % is given by decision tree regressor and Root mean square error value of 509.9. Coefficient of determination is calculated as 0.93.



RMSE SCORES	
Linear Regression	1843
Decision Trees	497
Random Forest	440
Gradient Boosting	382
Ada Boost	449

Table 1: RMSE Scores for different models.

In gradient boosting, the total estimators were set around 595 and attained a 96 % accuracy score and RMSE value of 382. Coefficient of determination is calculated as 0.965 being highest in comparison with all the models.

Implementation of AdaBoost algorithm is done and the total estimators were around 60 and the learning rate was 0.005 thus producing a model with accuracy of 95 % and RMSE value of 439.72. Coefficient of determination is calculated as 0.951.

Gradient boosting algorithm gave the best accuracy score and the least RMSE value among the other models. Hence, it can be used in prediction of the traffic flow for later use-cases with the provided dataset.

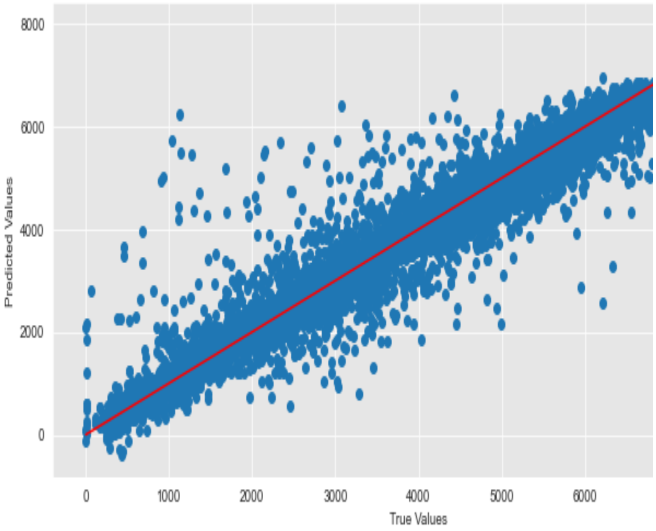


Fig 2: Predicted vs True values (Gradient-boosting)

Figure 2 clearly demonstrates how the straight line is in proximity to the points and fits most of the information and therefore has a less RMSE value.

The provided raw-data is analyzed to infer compelling understandings pertaining to the traffic movement arrangements and whether the traffic is at its peak at a given instance or not. Figure 3, shows that the traffic is at its peak in the month of July. Hence, it can be said that there is heavy traffic during the months of summer.

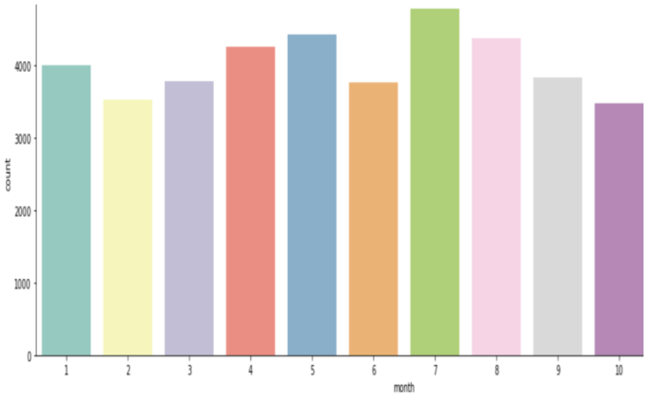


Fig 3: Traffic flow vs months

Figure 4, denotes the plot of traffic flow vs years. The traffic is highest in the year 2017.

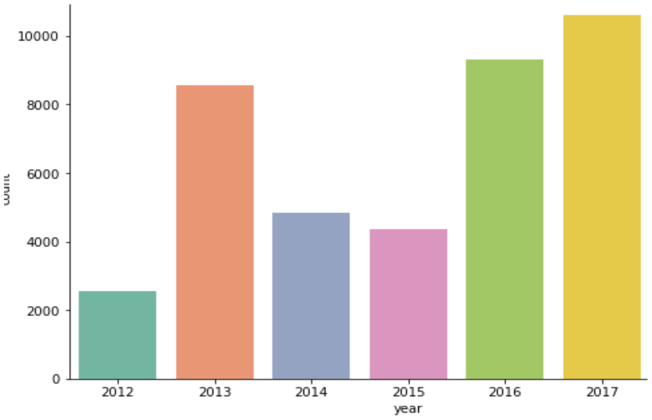


Fig 4: Traffic flow vs years

Through Figure 5, we can deduce that the traffic flow is weather dependent. Pleasant and bright climatic conditions indicate more traffic while as overcast conditions result in mediocre traffic. Hence, this shows people prefer travelling in clear weather rather in cloudy weather conditions.

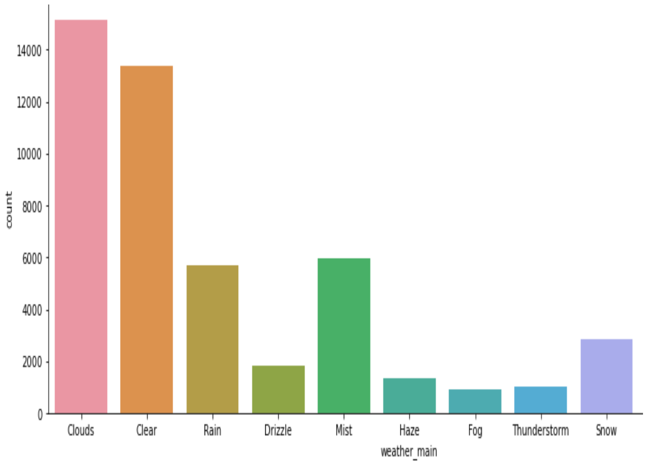


Fig 5: Traffic flow for different weather

VI. CONCLUSION

Analysis and prediction of the flow of traffic is done using various ML models and several models were collated formed on the accuracy score and the RMSE value. Based on the inferences, the best performing algorithm was gradient boosting algorithm with 97 % accuracy and RMSE value of 371.4. Analysis of the outcomes and the graphs was done to infer some understanding regarding the dependance of the traffic on certain attributes.

Following the research-work and after analysis of the coefficient of determination and accuracy score of the various ML models, it can be concluded that for our interest the gradient-boosting algorithm produced superior results among all other algorithms used.

Prediction and analysis of the traffic flow patterns utilizing this dataset is successfully implemented. For future learning, one can deduce more insights by taking into consideration more traffic situations making use of different sensors and vehicle monitoring devices. Superior models can be implemented if one can obtain real-time traffic flow statistics.

VII. ACKNOWLEDGEMENT

We are grateful to Dr T.K. Sivakumar (Associate Professor, School of Computing, SRM Institute of Science and Technology Chennai India) for his constant support and guidance throughout our project.

VIII. REFERENCES

1. Lizong Zhang, Nawaf Alharbe, Guangchun Luo, Zhiyuan Yao, and Ying Li's "A Hybrid Forecasting Framework Based on Support Vector Regression with a Modified Genetic Algorithm and a Random Forest for Traffic Flow Prediction"

2. A Survey on Big Data Based Vehicle Traffic Flow Prediction Using Deep Learning Algorithm.

3. Jiang and Adeli's "Dynamic Wavelet Neural Network Model for Traffic Flow Forecasting. Journal of Transportation Engineering".

4. Vlahogianni, Karlaftis, and J. C. Golias's "Temporal Evolution of Short-Term Urban Traffic Flow: A Nonlinear Dynamics Approach. Computer-Aided Civil and Infrastructure Engineering".

5. Amr Elfar, Alireza Talebpour, and Hani Mahmassani's "Machine Learning Approach to Short-Term Traffic Congestion Prediction in a Connected Environment".

6. Shanmei Li, Chao Wang's "Exploring Dynamic Characteristics of Multi-state Air Traffic Flow".

7. Lilian Pun, Peng Xiang Zhao, Xin Tao Liu's "A Multiple Regression Approach for Traffic Flow Estimation".