Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 4, 2021, Pages. 11682 - 11691
Received 05 March 2021; Accepted 01 April 2021.

Analysis of traffic flow in different weather conditions

Rohit Singh,

Computer Science & Engineering, SRM Institute of Science and Technology, Kattankulathur,

Aarzu,

Computer Science & Engineering, SRM Institute of Science and Technology, Kattankulathur,
Mr. Sibi Amaran,
Asst. professor, Computer Science & Engineering,, SRM Institute of Science Technology,
Kattankulathur
Dr. K Sree Kumar,
Asst. professor, Computer Science & Engineering, SRM Institute of Science Technology,
Kattankulathur

ABSTRACT

Analyzing and predicting the flow of traffic has always been a subject of great importance. The governments and traffic police can use this information to direct the traffic and facilitate the flow of traffic more freely and effectively. The information obtained can prevent future traffic congestion and save hours of commute time and thus resources. It can also save lives since the emergency services such as the ambulance can move faster to the hospitals. Predicting the flow of traffic is of paramount importance to effectively overcome these challenges. This paper presents several regression machine learning techniques that combine the performance of various algorithms and compares it with existing Machine Learning Algorithms using the mean square error, root mean square error and other performance metrics.

Index Terms: Traffic flow, Machine Learning, Linear regression, Decision Tree, Random Forests, AdaBoost, Root mean square error(rmse), Gradient Boosting

I. INTRODUCTION

Traffic flow information is of great importance for planning the transport activities and related research activities. The economy of a country is greatly affected by the amount of time wasted in traffic jams, which would otherwise have been spent working. Commuters in the top fifteen most-congested cities spent 83 hours on average stuck in traffic in the year 2017. The estimation of traffic flow is of paramount importance in deciding the flow of cars and diverting the traffic to alternate routes to accommodate the increased traffic. Using past records of the amount of traffic and employing machine learning methods on it, we can predict the amount of traffic

Traffic congestion degrades ambient air quality and increases vehicle emissions, and also recent studies have shown morbidity and mortality rate for drivers, commuters and also individuals living near major roadways and busy places. Our present understanding of the air pollution and its impacts like congestion on roads is not very comprehensive. Thus, the study of traffic prediction methods is of paramount importance in easing these difficulties. This can help the traffic controllers to manage the traffic. The problem of traffic congestion is caused mainly by wealthy car drivers who prefer to use their own vehicles for transport instead of using the public transport, lengthens the journey times and also forces the fares of public transport to increase significantly.

Traffic data first needs to be collected to be analyzed by the algorithms. Various methods are available for the collection of traffic data. Sensors are installed on the streets and traffic lights to measure the number of vehicles passing the roads in a given time period.

There are as of now different techniques to gather traffic data, for example, manual road studies, test vehicles or floating vehicle information (FCD), street-side finders, shut circuit TV (CCTV) camera video pictures,

among which circle identified information and FCD are commonplace fixed. Yearly traffic registration information gathered by street side locators have higher accuracy because of less influence of outside elements and the number of a wide range of vehicles passing the sensor-introduced area can be tallied. Different investigations have devoted to the assessment and forecast of traffic dependent on traffic registration information.

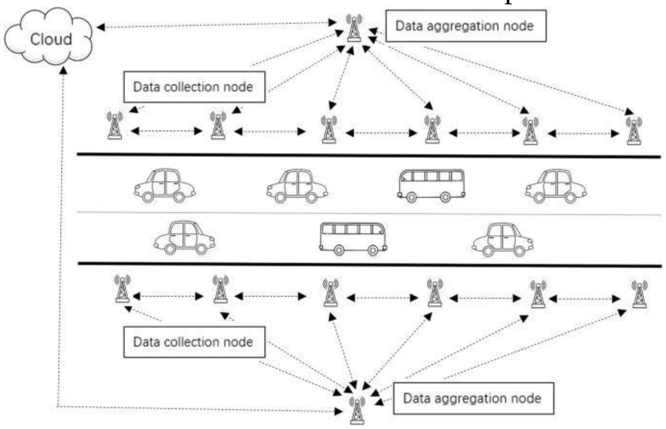


Figure: Sample traffic data collection

\

There is a need to analyze the flow of traffic to make informed decisions. The government can use this data to plan the construction of new roads and decide if the roads need to be widened. The roads prone to accidents need to be assessed over the risks and checked of alternate roads can be constructed to ease the burden of traffic on that road.

II. RELATEDWORK

Different solutions have been proposed over the years to accurately predict the volume of traffic flow. LILIAN PUN[1] and co proposed a multiple regression approach . They have taken different parameters of traffic and used a traffic data set from hong kong and their proposed approach is able to accurately predict the traffic using the different corelated measures

BIN FENG and co[2] have proposed a model based on time based combined traffic flow. The data is collected at route intersections from where the necessary prediction is based. SHANMEI LI [3] and co have used a method to predict the air traffic flow using k means clustering algorithm. GUOWEN DAI and co[4] have worked on a model that combines spatio temporal analysis with gated recurrent unit. The GRU is then used to predict the traffic using the spatio temporal information.

JIYAO AN et al [5] have worked on a system that uses novel based fuzzy neural network and it divides the block into smaller blocks.

RONGHAN YAO and co[6] have proposed to use a weighted markov model to predict the traffic volumes during different periods Lizong Zhang and co [7] proposed work uses a hybrid forecasting framework that uses support vector regression and random forest algorithm to select the best information subset to determine the optimum forecasting parameters.

SAIF EDDIN et al [8] have used a vehicle trajectory dataset and applied a mesoscopic traffic flow model and described the spatio-temporal probability distributions of vehicle trajectories. YIXUAN MA and co[9] have proposed a convolutional LSTM neural network model to predict the multi lane traffic.

CHUN AI et al [10] have proposed a hybrid neural network algorithm that uses radial basis function to predict the road speed and congestion in traffic. Xinqiang Chen and co[11] have worked on a model that makes use of data denoising schemes to eliminate the outliers and smooth the data and uses the LSTM neural network for predicting the traffic. Azzedine Boukerche et al [12] have proposed a model that can help intelligent transportation systems make more informed decisions like road routing, traffic congestion control etc. They have used two models, one being statistics based and the other based on machine learning. This paper has given an overall survey of the different approaches to predict the short term traffic.

Huakang Lu et al [13] have proposed a novel long short term memory LSTM network which has been enhanced by TCC (temporal aware convolutional context) blocks. They have used a new loss switch mechanism for this. The whole network was trained in an end to end way.

Saiqun Lu and co [14] have proposed a combined model that uses autoregressive integral moving average (ARIMA) and a long short term memory neural network model. Their versatile model which combined both these models gave better prediction results.

Linjiang Zheng et al [15] have proposed a spatial temporal method for feature optimization to predict the traffic for the short term. They have employed the gradient boosted regression tree for their model. Their experimental results have demonstrated that their DSTO GBRT model can provide efficient and adaptive prediction of the traffic flow.

III. PROPOSED SOLUTION

There are various approaches that are used for predicting the flow of traffic. Different ML models can give different results. The results also depend on the type of dataset used and the preprocessing steps involved. WE have to take into account the resources required for the models and their complexities. Not all ML models can be applied to every type of dataset.

We have taken into consideration these points and implemented the ML models and analyzed their performance. We have compared the accuracy score for these models in order to arrive at the most optimum algorithm and then proceed to build our model.

A. Dataset

The dataset used in this study is Metro Interstate Traffic Volume Data Set This dataset is an improvement on its predecessor and contains different features including the weather data. It measures Hourly traffic data along Minneapolis-St Paul, MN for westbound I-94. It also includes the weather and holiday features from the years 2012 to 2018. The dataset has 48204 number of instances and 9 attributes.

B. Data Preprocessing

The preprocessing is done in various stages. In order to apply the machine learning models the dataset needs to be properly preprocessed and inconsistent data needs to be rectified. In order to do this, the following steps are performed:

1. Drop unnecessary columns from data set. After checking the correlation between features, the columns that are not required and will not help for prediction are dropped.

2. Convert values to numerical format: In order to properly apply the machine learning algorithms, the columns are transformed into numerical values so that the model can use it to predict the values, including the day and weather columns.
3. Replace null and infinity values: The dataset consists various row as null or infinity values. This is addressed by replacing the infinity values as maximum values and missing values as average value.

C. Architecture

The architecture diagram in Fig. 2 consists of applying preprocessing methods to the dataset followed by the data being fetched into the regressor.

These algorithms are then compared based on their accuracy score to determine the best performing model.

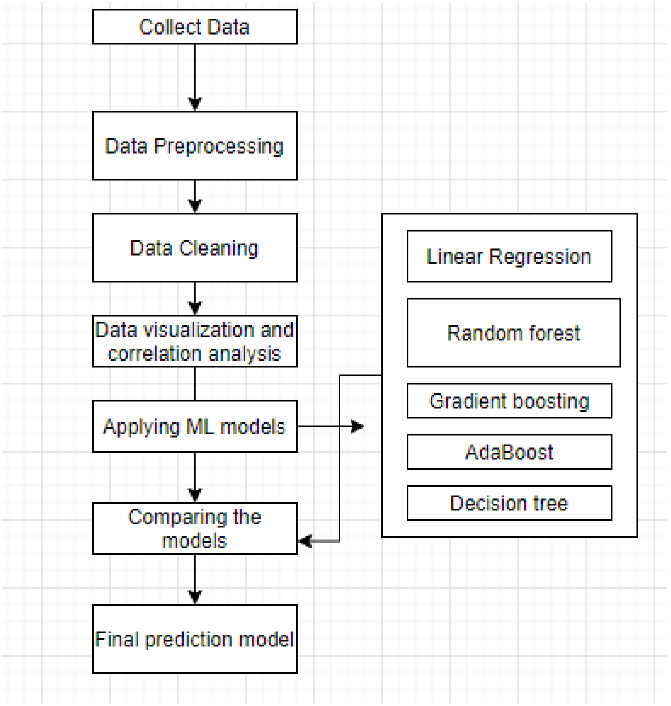


Figure2–Architecture Diagram

D. MachineLearning Algorithms

- ***Linear Regression**:* It is a supervised machine learning algorithm that is used to solve regression problems. It plots the points in a two dimensional axes and finds the best fit line that will pass through the maximum number of points. It is a really simple algorithm

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \ldots + \beta_i X_i}$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

- ***Decision Tree**:* This is a supervised learning algorithm which can be used for classification as well as regression problems. It sorts the tree from root node to leaf node. The nodes after sorted can be then used to predict the values. It is a model that predicts the value of a target variable by simply learning the simple decision rules that is inferred from the data features.
- ***RandomForest**:* It is a set of decision trees that are randomly selected from a training set and then a vote is aggregated from all the decision trees randomly and final object tested is given. The Random Forest Algorithm merges the output of multiple Decision Trees to generate the final output.
- ***Gradient Boosting:*** Gradient Boosting trains many models in a gradual, additive and sequential manner**.** This model identifies the shortcomings by using gradients in the loss function (y=ax+b+e, e needs a special mention as it is the error term). The loss

function is a measure indicating how good are model’s coefficients are at fitting the underlying data.

- ***AdaBoost:*** The AdaBoost Algorithm begins by training a decision tree in which each observation is assigned an equal weight. After evaluating the first tree, it increases the weights of those observations that are difficult to classify and then lowers the weights for those that are easy to classify. The second tree is as a result grown on this weighted data. Thus, it is able to improve upon the predictions of the first tree. The new model is therefore Tree 1 + Tree 2.

IV. EVALUATIONMETRICS

The various parameters used for evaluating the different ML models are:

- Root mean square error: It gives the measure of closeness to as specific value. The formula for accuracy score is mentioned below:

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

In this equation, y is the true label, y’ is the predicted label which is given after prediction. For multilabel classification, The accuracy score give us the accuracy of the subset. When the entire set of predicted values matches the true label the accuracy score is given as 1.0,otherwise 0.0.

- Coefficient of Determination: The coefficient of determination (R^2 or r-squared) is a measure of statistics that is used in a regression model to determine the proportion of variance of the dependent variable that on the independent variable. The coefficient of determination is a good measure of how well the data fits the model (the goodness of fit) and how well it matches the actual data. It can take values between 0 and 1.

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

V. RESULTS AND ANALYSIS

The different ML models were implemented and the results were compared to find the best performing model. Linear regression gave a root mean square error value of 1845.63. The graph for the actual values vs the predicted values is drawn in fig 3. It gave the Coefficient of determination 0.144.

The decision tree regressor gave an accuracy score of 93 percent and a rmse value of 509.91 which is better than the linear regression model. The graph for the actual values against predicted values is shown in fig 4. The Coefficient of determination is 0.93 for this model.

Table2. RMSE Score for various models.

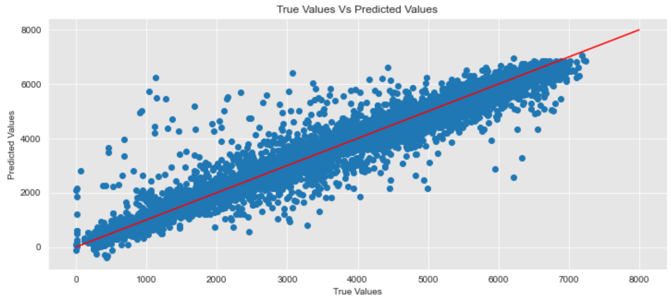
Linear Regression	1844.005636
Decision Tree	493.413445
Random Forest	428.853892
Gradient Boosting	371.460229
Ada Boost	439.732726

The Random Forest Regressor model gave an accuracy score of 95 percent with a rmse value of 428.85 which is still better than the predecessor. The Coefficient of determination is 0.95 for the random forest regressor model.

In the gradient boosting algorithm, we set the number of estimators to 600 and the number of estimators to 11 and this achieves an accuracy score of 96 percent and a rmse value of 371.46. The Coefficient of determination is 0.965 for the gradient boosting model which is the highest among all the models.

We also implement the AdaBoost algorithm and set the number of estimators to 60 and the learning rate to 0.005 and it produced a model with 95 percent accuracy and a rmse value of 439.73. The Coefficient of determination is 0.951 for this model.

Thus, after comparing all the models we see that the gradient boost algorithm gives the best accuracy score and the least rmse value. Thus, it can be used for predicting the flow of traffic for the future use cases too for the given dataset.



True vs predicted values Gradient boosting

This is the graph for the selected model that plots the predicted values against the true values after taking the gradient boosting algorithm. We can infer that that the line is really close to the points and fits maximum amount of data hence the high accuracy score.

Also, the given dataset can be used to draw interesting insights about the pattern of traffic and when the traffic is at its peak. In fig 6, we can see that the traffic is at its highest in the July month. Thus, we can say that the traffic is heaviest after the summer months. In figure 7, we plot the traffic against the years. The traffic was the heaviest in the year 2017. From figure 8, we can see that the traffic count is maximum for clear and cloudy weather conditions and when the weather is adverse the traffic is low. This is to be expected since people will prefer to travel on clear weather conditions only.

VI. CONCLUSION

The traffic flow analysis and prediction was conducted by using different algorithms and different algorithms were compared based on their accuracy score and their root mean square error value. Based on our findings, the gradient boosting algorithm performed best with a 97 percent accuracy and a root mean square error value of 371.46. We also analyzed the traffic data to draw some insights about the most active time periods and months.

Thus, after the research work and analyzing the accuracy score and the Coefficient of determination of the different machine learning models, we come to the conclusion that for our purpose the gradient boosting algorithm performed better than the other algorithms.

The analysis and prediction of traffic based on the data given was successfully implemented which was the main aim. For future work, we can gather more insight by taking into account more traffic conditions using the sensors and traffic monitoring device. A better model can be implemented if we can get real time traffic data and thus predict the congestion in roads based on real time data.

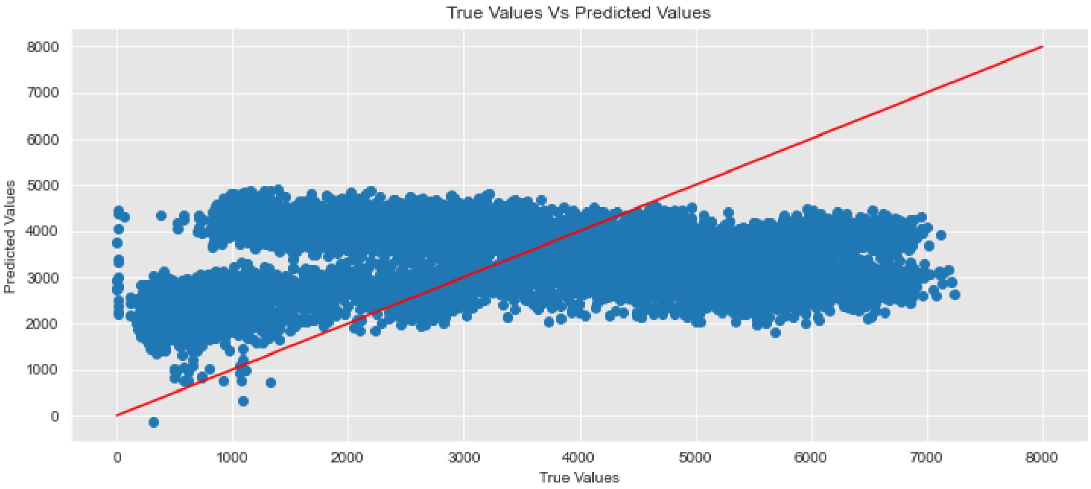


Fig:2 Linear regression true vs predicted values

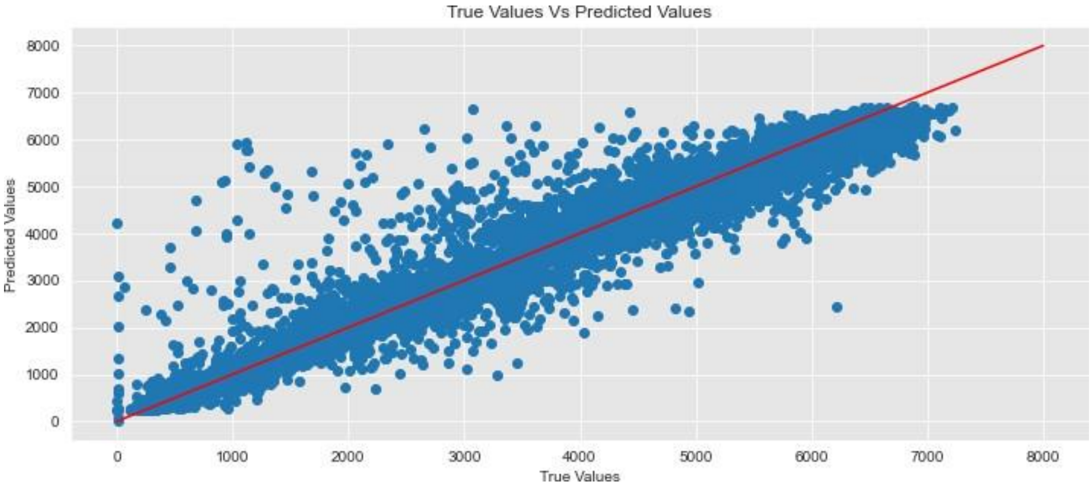


Fig:3 Random forest true vs predicted values

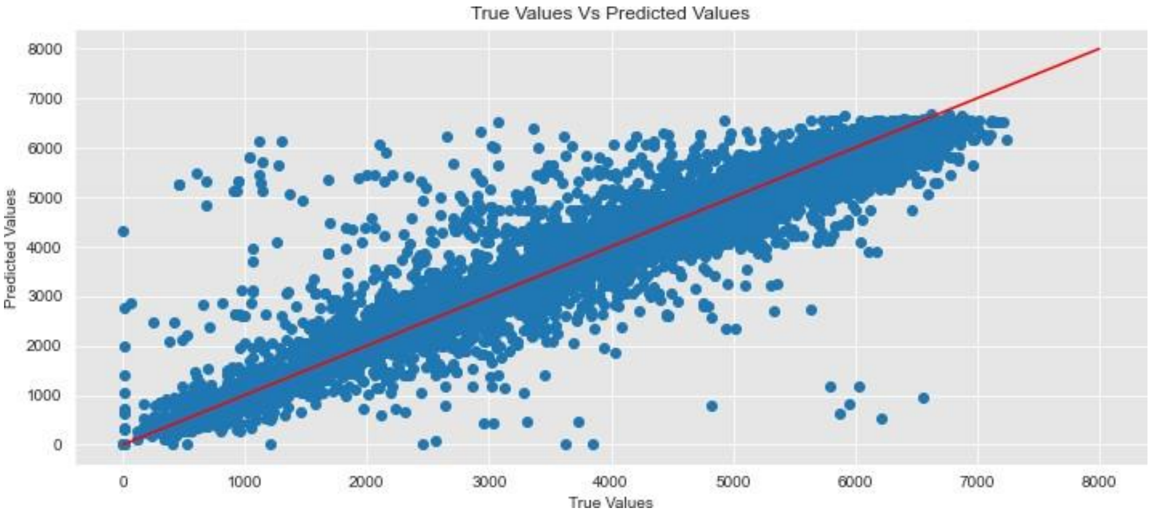


Fig:4 Decision tree true vs predicted values

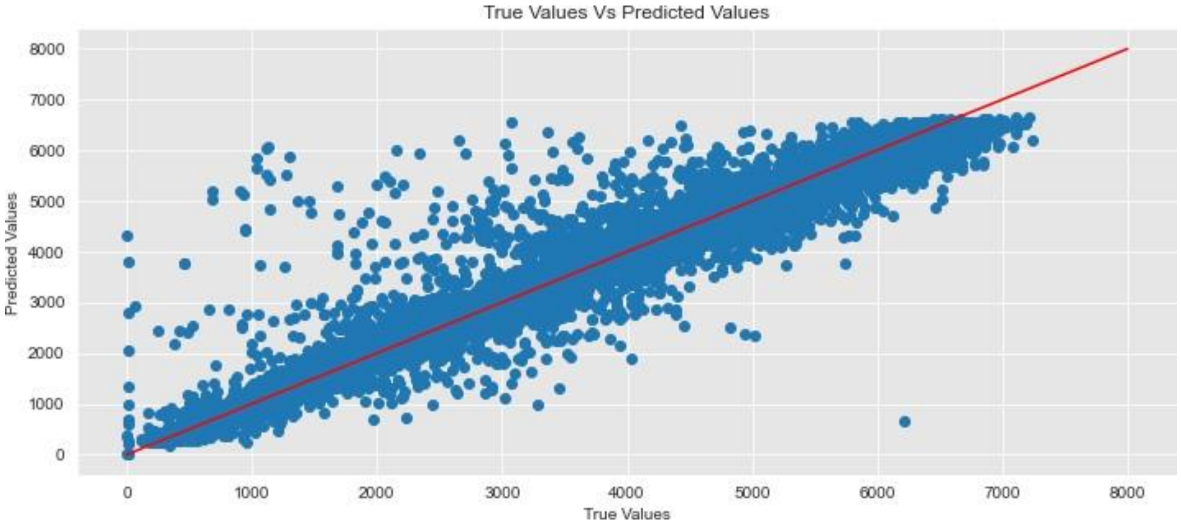


Fig 5: AdaBoost true vs predicted values

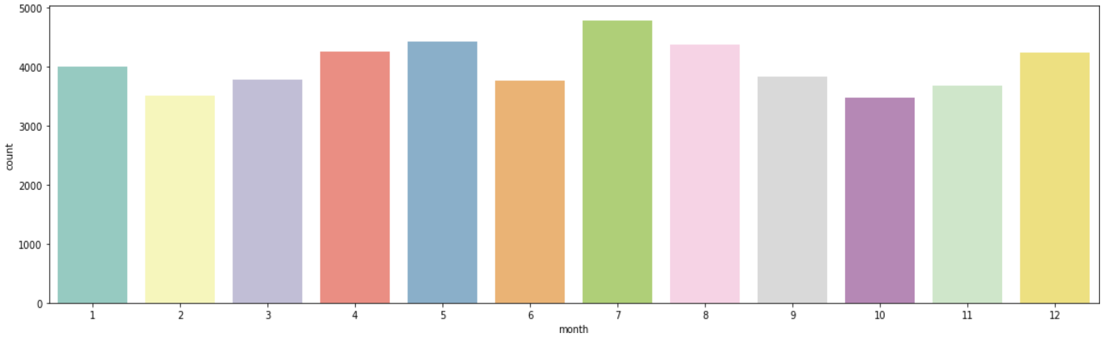


Fig 6: Traffic flow vs months

Received 05 March 2021; Accepted 01 April 2021.

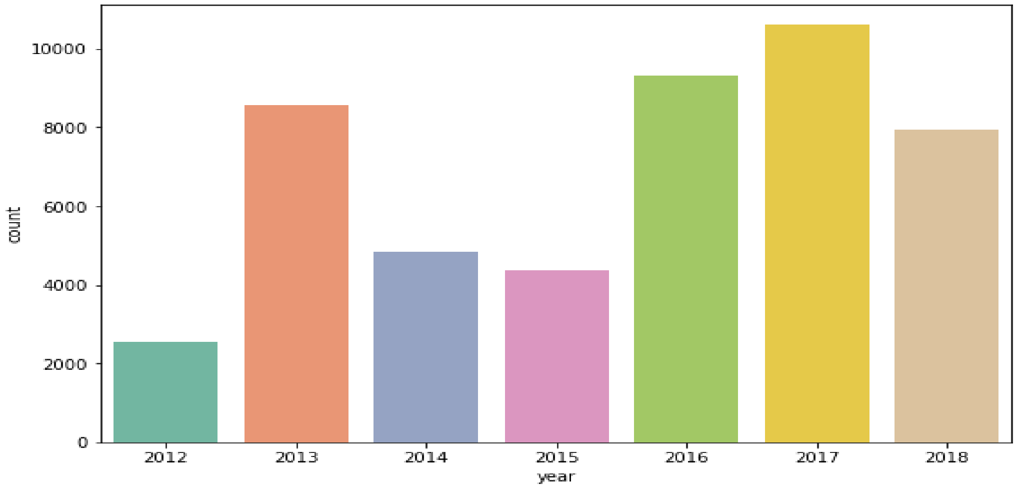


Fig 7: Traffic flow vs years

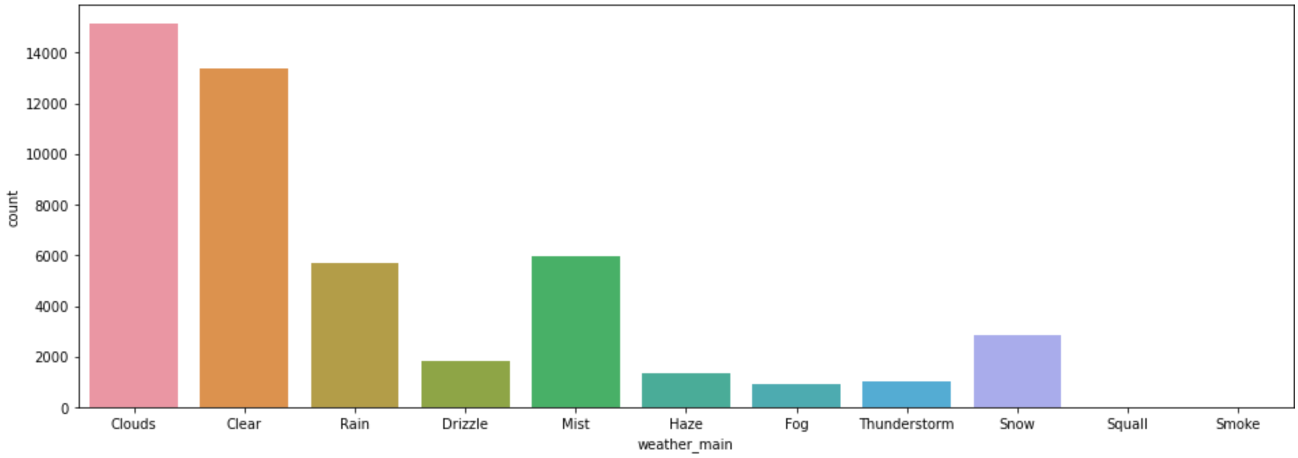


Fig 8: Traffic flow for different weather

REFERENCES

[1] LILIAN PUN1, PENGXIANG ZHAO 2, AND XINTAO LIU1, “A Multiple Regression Approach for Traffic Flow Estimation “

[2] BIN FENG 1, JIANMIN XU 1, YONGJIE LIN 1, (Member, IEEE), AND PENGHAO LIA Period-Specific Combined Traffic Flow Prediction Based on Travel Speed Clustering .”

[3] SHANMEI LI , CHAO WANG , AND JING WANG “Exploring Dynamic Characteristics of Multi-State Air Traffic Flow A Time Series Approach “

[4] GUOWEN DAI1, CHANGXI MA 1, AND XUECAI XU “Short-Term Traffic Flow Prediction Method for Urban Road Sections Based on Space Time Analysis and GRU Air Traffic Flow: A Time Series Approach Prediction Based on Travel Speed Clustering .

[5] JIYAO AN , (Member, IEEE), LI FU , MENG HU , WEIHONG CHEN, AND JIAWEI ZHAN, “A Novel Fuzzy-Based Convolutional Neural Network Method to Traffic Flow Prediction With Uncertain Traffic Accident Information”

[6] RONGHAN YAO , WENSONG ZHANG , AND DONG ZHANG, “Period

Division-Based Markov Models for Short-Term Traffic Flow Prediction."

[7] Lizong Zhang, Nawaf R Alharbe_, Guangchun Luo, Zhiyuan Yao, and Ying Li "A Hybrid Forecasting Framework Based on Support Vector Regression with a Modified Genetic Algorithm and a Random Forest for Traffic Flow Prediction"

[8] SAIF EDDIN G. JABARI 1,2, DEEPTHI MARY DILIP DIANCHAO LIN AND BILAL THONNAM THODI2 s, "Learning Traffic Flow Dynamics Using Random Fields."

[9] YIXUAN MA , ZHENJI ZHANG AND ALEXANDER IHLER,"Multi-Lane Short-Term Traffic Forecasting With Convolutional LSTM Network."

[10] CHUN AI , LIJUN JIA , MEI HONG , AND CHAO ZHANG "Short-Term Road Speed Forecasting Based on Hybrid RBF Neural Network With the Aid of Fuzzy System-Based Techniques in Urban Traffic Flow."

[11] XinqiangChen ,Huixing Chen , Yongsheng Yang , HuafengWuc, Wenhui Zhang , Jiansen Zhao , Yong Xiong "Traffic flow prediction by an ensemble framework with data denoising and deep learning model."

[12] AzzedineBoukerche, Yanjie Tao, Peng Sun, "Artificial intelligence-based vehicular traffic flow prediction methods for supporting intelligent transportation systems."

[13] Huakang Lu, Zuhao Ge, Youyi Song, Dazhi Jiang, Teng Zhou, Jing Qin, "A Temporal-aware LSTM Enhanced by Loss-switch Mechanism for Traffic Flow Forecasting"

[14] SaiqunLu ,Qiyan Zhang , Guangsen Chen , Dewen Seng "A combined method for short-term traffic flow prediction based on recurrent neuralnetwork"

[15] Linjiang Zheng ,Jie Yang , Li Chen , Dihua Sun , Weining Liu "Dynamic spatial-temporal feature optimization with ERI big data for Short-term traffic flow prediction"