An Accurate Estimation of Interstate Traffic of Metro City Using Linear Regression Model of Machine Learning

Apurv Chandel, Shubham Sagar, Badavath Uday Kiran, Prabhas Prasad, Nidhi Lal

Dept. of Computer Science and Engineering
IIIT Nagpur, India

apurv.chandel@cse.iiitn.ac.in,shubhamsagar5567@gmail.com,badavathuday456@gmail.com,
prabhasprasad106@gmail.com,nidhi.lal@cse.iiitn.ac.in

***Abstract*— Nowadays Traffic is a big issue for everyone and it's a concern for everyone who goes through it in their day to day life. An increase in population slows down the traffic and make it even worse day by day. The Settlement of modern civilization does have a look at this but unable to act as to cover so people. By many methods and patterns, we can observe the traffic and get the data and predict the next and consecutive observations. Then observation is made by the observation agency and then required out and predictions are through that. Being in cosmopolitan city traffic is the most basic thing which occurs in one's life. Google Maps & Bing Maps collect the data through GPS and calculate and mark the areas having high congestion. The linear regression plays a vital role in this data analysis for this and make a prediction and make graphs through data analysis using TPOT Regressor, Random Forest Regressor and make_pipeline to this.**

***Keywords*— Linear Regression, TOPT Regressor, XGB Regressor, Random Forest Regressor.**

I. INTRODUCTION

Traffic congestion has increased dramatically in India [1]. Congestion and the associated slow urban mobility can have a huge adverse impact on both the quality of life and the economy. Every city in India is polluted? Are Delhi and Mumbai less or more congested than, say, Patna and Varanasi [2]? Is Congestion in different within cities across the centre, and at different times of the day? How Indian cities are different than US? What does the future hold [3]?

This data shows the mobility of vehicle in India. Slow Traffic are found even in the peak hours, and in both large and small cities. India's mean travel speed across cities is just 25.4 km per hour, much slower than the mean travel speed of 36.5 km per hour in metropolitan cities in the US [4]. There is difference in mobility in India across many cities. A factor of nearly two separates the fastest and slowest cities [5]. These differences are given by the differences in uncongested mobility [6].

Linear regression is simple approach for predicting a response using a single feature. There are Two variables which are linearly related. Hence, we try to find a linear function which predicted the response value(y) as accurately as possible as a function of the feature or independent variable(x).

II. RELATED WORK

Regression problem where related data of each employee represent one observation. We assume that the experience, education, role, and city are the independent features, while the salary depends on them [7]. In regression analysis, we usually consider some phenomenon of interest and take number of observations where each observation has one or more features [8]. Following the assumption that (at least) two of the features depends on the others, we try to establish a relation among them [9].

We need regression find some phenomenon influences the other or how several variables are related. Example, we can use it to determine if and to what extent the experience or gender impact Jobs [10]. Regression is very informative when we want to forecast a response using a new set of predictors. For example, we can predict electricity consumption of a household for the next month given the outdoor temperature, time of day, and number of people in that household [11].

Regression is used in many vast numbers of fields: economy, computer science, electronics, and so on. Its importance rises every day with the availability of large amounts of data and increased awareness of the practical value of data [12].While implementing linear regression for some dependent variable y in the set of independent variables $\mathbf{x} = (x_1, \ldots, x_r)$, where r is the number of predictors, we assume linear relationship between x and y: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon$. This equation is termed as the regression equation. β_0, β_1, …, β_r are here regression coefficients, and ε is the random error [13].

Linear regression which calculates the estimators of the regression coefficients are denoted with b_0, b_1, …, b_r. They define as the estimated regression function $f(\mathbf{x}) = b_0 + b_1x_1 + \cdots + b_rx_r$. This function should give the dependencies between the inputs and output [14].

The predicted response of $f(\mathbf{x}_i)$. for each observation $i = 1, \ldots, n$, should be as close to the corresponding actual response y_i. The differences $y_i - f(\mathbf{x}_i)$ for all observations $i = 1, \ldots, n$, are called the residuals [15]. To get the best results, we usually minimize the sum of squared residuals (SSR) for all observations $i = 1, \ldots, n$: $\text{SSR} = \Sigma_i (y_i - f(\mathbf{x}_i))^2$. This method is called the method of ordinary least squares [16].

III. Proposed Work

The project here contains a very vast description of linear regression. Which contain the sklearn module, TPOT Regression, Random Forest Regression. The graphs proposed have the general discussion over the explanation about the increase in the accuracy of the data set through various regression which we have used in it. A library like pandas, numpy, make_pipeline. We have been provided with the data of Metro_Interstate_Traffic_Volume Dataset from which we have extracted the information which can be used in further uses.

We have to use TPOT Regressor to optimize the code and to enhance the value of the R^2 score in the data set. We have initial R^2 of 0.855874068223368in the given set and after using the TPOT Regressor we have optimized it to the 0.9250189989233896**.** Which can be further used in any of the Information data sciences and analysis.

We have read the data set then categoric encoding is performed and it optimizes the data and all the strings are converted into float representation using pandas and converted date time column structure in multiple columns like the year, month, day, the hour. After which we have selected the value of x and y where y includes data of traffic volume and all others are included in x. then we used train_test_split to choose perfect and random data so that we can prove the accuracy of the predicted values. Then we have used TPOT Regressor to optimize the code by setting different values of time (MAX Time_MINS=60,420,840,1800) to optimize the code to get better accuracy and exported the code into metro_traffic.py file. After using TPOT Regressor we have calculated the R2 score and plotted the graph of predicted Traffic Volume vs True Traffic Volume.

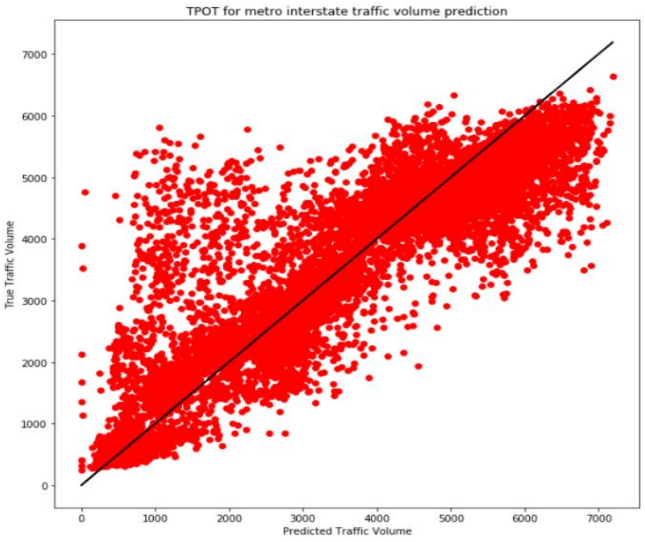

Fig 1: Initial Metro_Interstate_Traffic_Volume Prediction.

After 60 min of TPOT Regressor Optimisation we get optimized code in metro_traffic.py file using the code in the file we use the package Random_Forest_Regressor and make_pipeline functions to improve the accuracy but we have to set the file name into reading location of that of data set and we have set the “target” into “traffic_volume” and we have once again categorial encoding of that data set to convert string data into float64 data. After using the make_line function we have predicted the traffic volume using TPOT.predict and calculated the R2 score and plotted its graph.

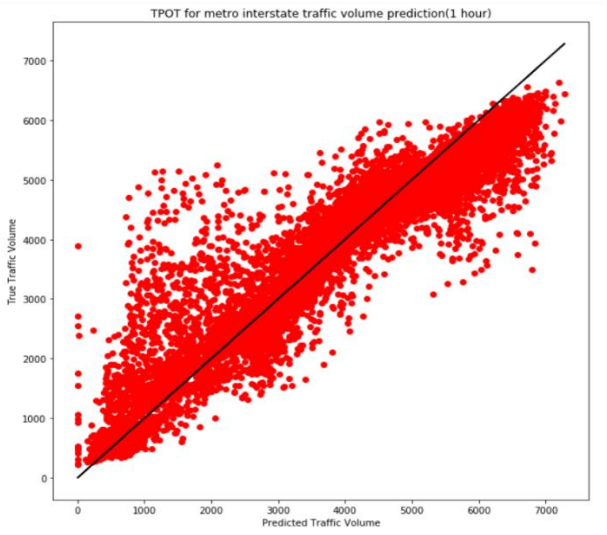

Fig 2: TPOT for Metro Interstate Traffic Volume Prediction(60 min).

After 420 min of TPOT Regressor Optimisation we get optimized code in metro_traffic.py file using the code in the file we use the package Random_Forest_Regressor functions to improve the accuracy but we have to set the file name into reading location of that of data set and we have set the “target” into “traffic_volume” and we have once again categorial encoding of that data set to convert string data into float64 data. After using the make_line function we have predicted the traffic volume using TPOT.predict and calculated the R2 score and plotted its graph.

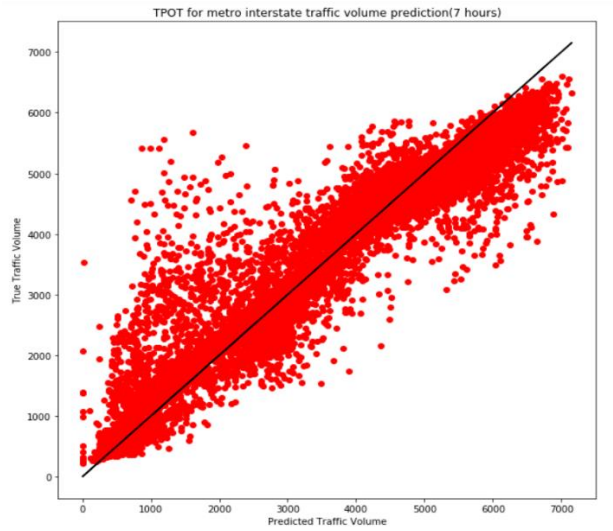

Fig 3: TPOT for Metro Interstate Traffic Volume Prediction (420 min).

After 840 min of TPOT Regressor Optimisation we get optimized code in metro_traffic.py file using the code in the file we use the package Extra_Trees_Regressor functions to improve the accuracy but we have to set the file name into reading location of that of data set and we have set the “target” into “traffic_volume” and we have once again categorial encoding of that data set to convert string data into float64 data. After using the make_line function we have predicted the traffic volume using TPOT.predict and calculated the R2 score and plotted its graph.

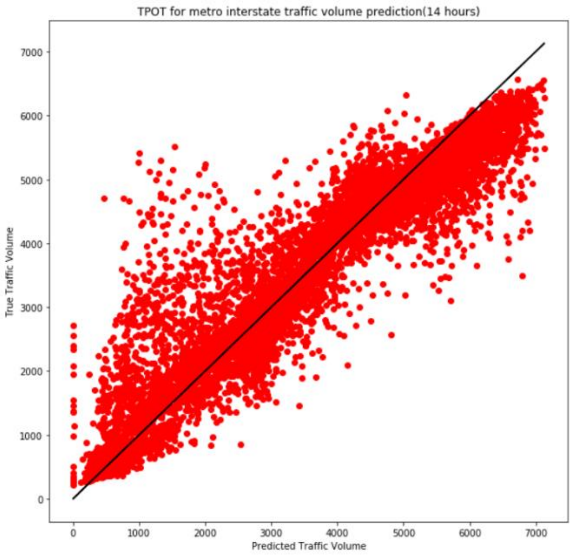

Fig 4: TPOT for Metro Interstate Traffic Volume Prediction (840 min).

After 1800 min of TPOT Regressor Optimisation we get optimized code in metro_traffic.py file using the code in the file we use the package XGB_Regressor functions to improve the accuracy but we have to set the file name into reading location of that of data set and we have set the “target” into “traffic_volume” and we have once again categorial encoding of that data set to convert string data into float64 data. After using the make_line function we have predicted the traffic volume using TPOT.predict and calculated the R2 score and plotted its graph.

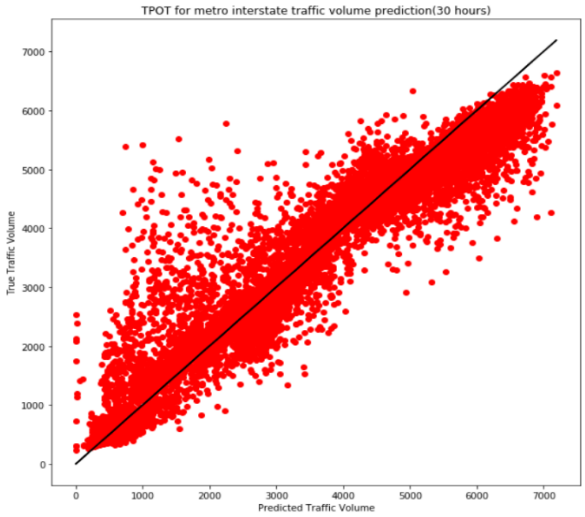

Fig 5: TPOT for Metro Interstate Traffic Volume Prediction (1800 min).

IV. Results

In all the figures 1,2,3,4,5 X-axis represents “Predicted Traffic Volume” and Y-axis represents “True Traffic Volume” and the line plotted is predicted traffic volume. According to previous model calculated Previous Maximum Coefficient of Determination R^2 = 0.855874068223368.

After testing and training we get the following results corresponding to random state 42 and max_mins = 60,420,840,1800. After 1800 minutes we get the Proposed Maximum Coefficient of Determination R^2 = 0.9250189989233896 which has improvement of 8.078867355281992% from the proposed work.

V. Conclusions

In this paper, we presented our methodology/method on extracting performance of the processor and using this information, we tried to develop a linear regression model to predict the performance of any processor and increase the accuracy of that model. We developed an extensive end to end simulation framework, which generally includes tools for data collection and also for measuring the accuracy of our performance. Although our experiment has shown good results, an extensive experimentation is required to verify all the aspects of our work as many independent variables are not considered. Besides conducting additional experiments, we are actively working on improving our methodology to increase the R^2 score. Main directions of our work are required to increase the coefficient of R^2 of prediction. Our models are also successful for estimating the future system performance and models when limited or large data is available for already built system and models solely on the basis of specifications of the processors.

References

[1] Abdulla, Dudekula, Gandikota Ramu, and N. Mamatha. "A survey on citywide traffic estimation techniques." In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pp. 3313-3318. IEEE, 2017. Chin, Woon Hau, Zhong Fan, and Russell Haines. "Emerging technologies and research challenges for wireless networks." IEEE Wireless Communications 21.2 (2014): 106-112.

[2] Weichenthal, Scott, Keith Van Ryswyk, Alon Goldstein, Scott Bagg, Maryam Shekkarizfard, and Marianne Hatzopoulou. "A land use regression model for ambient ultrafine particles in Montreal, Canada: A comparison of linear regression and a machine learning approach." *Environmental research* 146 (2016): 65-72.

[3] Polson, Nicholas G., and Vadim O. Sokolov. "Deep learning for short-term traffic flow prediction." *Transportation Research Part C: Emerging Technologies* 79 (2017): 1-17..

[4] Abdulla, Dudekula, Gandikota Ramu, and N. Mamatha. "A survey on citywide traffic estimation techniques." In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pp. 3313-3318. IEEE, 2017. Chin, Woon Hau, Zhong Fan, and Russell Haines. "Emerging technologies and research challenges for wireless networks." IEEE Wireless Communications 21.2 (2014): 106-112.

[5] Polson, Nicholas G., and Vadim O. Sokolov. "Deep learning for short-term traffic flow prediction." *Transportation Research Part C: Emerging Technologies* 79 (2017): 1-17.Lynggaard, Per, and Knud Erik Skouby. "Deploying -technologies in smart city and smart home wireless sensor networks with interferences." Wireless Personal Communications 81.4 (2015): 1399-1413. Davis, Lucas W. "The effect of driving restrictions on air quality in Mexico City." *Journal of Political Economy* 116, no. 1 (2008): 38-81.

[6] Davis, Lucas W. "The effect of driving restrictions on air quality in Mexico City." *Journal of Political Economy* 116, no. 1 (2008): 38-81.

[7] Davis, Lucas W. "The effect of driving restrictions on air quality in Mexico City." *Journal of Political Economy* 116, no. 1 (2008): 38-81.

[8] Moser, Bernhard, and Pius Loetscher. "Lymphocyte traffic control by chemokines." *Nature immunology* 2, no. 2 (2001): 123.

[9] Polson, Nicholas G., and Vadim O. Sokolov. "Deep learning for short-term traffic flow prediction." *Transportation Research Part C: Emerging Technologies* 79 (2017): 1-17.Lynggaard, Per, and Knud Erik Skouby. "Deploying -technologies in smart city and smart home wireless sensor networks with interferences." Wireless Personal Communications 81.4 (2015): 1399-1413. Davis, Lucas W. "The effect of driving restrictions on air quality in Mexico City." *Journal of Political Economy* 116, no. 1 (2008): 38-81.

[10] Polson, Nicholas G., and Vadim O. Sokolov. "Deep learning for short-term traffic flow prediction." *Transportation Research Part C: Emerging Technologies* 79 (2017): 1-17.

[11] Abdulla, Dudekula, Gandikota Ramu, and N. Mamatha. "A survey on citywide traffic estimation techniques." In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pp. 3313-3318. IEEE, 2017. Chin, Woon Hau, Zhong Fan, and Russell Haines. "Emerging technologies and research challenges for wireless networks." IEEE Wireless Communications 21.2 (2014): 106-112.

[12] Fitts, Paul M. "Human engineering for an effective air-navigation and traffic-control system." (1951).

[13] Gordon, Robert L., Warren Tighe, and I. T. S. Siemens. *Traffic control systems handbook*. No. FHWA-HOP-06-006. United States. Federal Highway Administration. Office of Transportation Management, 2005.

[14] Gartner, Nathan H. *OPAC: A demand-responsive strategy for traffic signal control*. No. 906. 1983.

[15] Erzberger, Heinz. "Automated conflict resolution for air traffic control." (2005).

[16] National Research Council. *Flight to the future: Human factors in air traffic control*. National Academies Press, 1997.