Visualizing Impact of Weather on Traffic Congestion Prediction: A Quantitative Study

Shahrukh Hussain
Dept. of Computer Science
Forman Christian College (A Charterred University)
Lahore, Pakistan
shahrukh.1472@gmail.com

Usama Munir
Dept. of Computer Science
Forman Christian College (A Charterred University)
Lahore, Pakistan
usama-munir@live.com

Muhammad Salman Chaudhry
Dept. of Computer Science
Forman Christian College (A Charterred University)
Lahore, Pakistan
salmanchaudhry@fccollege.edu.pk
salmanchaudhry909@gmail.com

***Abstract*— A substantial amount of research has been done to develop improved Intelligent Transportation Systems (ITS) to alleviate traffic congestion problem. These include methods that incorporate indirect impact on traffic flow such as weather. In this paper we study the impact of weather changes on traffic congestion along with more spatial and temporal factors, such as weekday/time and location, which is a different angle to this problem. The proposed solution uses all these indicators to estimate the flow of traffic. We evaluate the level of congestion (LOC) based on the traffic volume grouped in certain regions of the city. The index for the defined LOC indicates the traffic flow from "free flowing" to "traffic jam". The data for the traffic volume count is collected from the Department of Transportation (DOT) for NYMTC. Weather conditions along with special and temporal information have an essential part in predicting the congestion level. We use supervised machine learning for this purpose. The prediction models are based on certain factors such as the volume count of the traffic at entry and exit point of each street pair, the day of the week, timestamp, geographical location, and weather parameters. The study is done on the major roadways of each of the four prominent boroughs in New York. The results of the traffic prediction model are established by using the Gradient Boosting Regression Tree (GBRT) which shows accuracy of 97.12%. Moreover, the calculation speed is relatively fast, and it has stronger applicability to the prediction of congestion conditions.**

Index Term- *Gradient Boosting, Decision Tree Algorithm, Supervised Machine Learning, Traffic Congestion*

I. INTRODUCTION

Urban traffic has seen enormous increase in recent times globally. The overall process of modernization is speeding up, leading to the rapid growth of vehicular traffic on the road. To cater the needs for huge surge in traffic, urban city road networks are becoming over complex. Consequently, urban traffic problems are getting serious and traffic congestion is one of them [1]. In metropolitan cities, if the factors leading to the congestion are neglected and congestion is not predicted properly and reported to the users in time, it can lead the road networks to be paralyzed. The early step to tackle the problem of congestion is to prevent it from happening. Therefore, the establishment of traffic flow, forecasting with respect to the day and time of the day is conducive to the preparation of targeted preventive measures which serve as early warning. The usage of Intelligent Transportation Systems (ITS) to predict traffic-related information has gained popularity in the field of smart transportation. A well-designed ITS can estimate and inform drivers of the locations and time frame of congested road sections, thus allowing them to avoid taking that route. Moreover, it can also provide significant information for authorities of large metropolitan areas to control the parameters of the traffic signal controls ahead time to reduce the level of congestion (LOC).

Supervised machine learning models are highly effective and fast when training structured data. However, model performance and accuracy is highly dependent on the dataset since its correct input features and labelling followed by minimum null values defines how well a model performs in a real world applications. These models are expected to generate adequate results with precision as the datasets become more diverse. We have used the supervised machine learning models to estimate the traffic flow and congestion as in recent times. The correlations between implicit traffic-related data and weather condition data define how much one value influences the other. A detailed exploratory analysis was performed over important weather features that impact congestion on the roads the most, to unveil individual impacts over the traffic flow within the given route at certain time and day of the week.

II. BACKGROUND

Researchers from different domains have studied the problem of traffic flow and recurrent congestion using various techniques in the past. Statistical analysis is based on a variety of features that lead to measuring the congestion of vehicles on road such as motion of the vehicle, stationary time of the vehicle, velocity of the vehicle or the cluster of the vehicle within the selected segment of the road network. Data collection is the first step to solve the problem of traffic. Various methods include using the GPS-based [2] and cellular-based [3] sensors installed in smart phones, vehicles, and roadsides, to gather data of geographical location and timestamp. While others have used images and videos from cameras on the road network of metropolitan areas and drones [9], to extract the vehicle data from intersections, highways, and freeway. Both the supervised and unsupervised machine learning algorithms have played an integral part in predicting the congestion based on the feature, labelled and unlabeled. In our study, we took the data from each entry and exit of each pair of streets to estimate the traffic congestion and measure the influence of the weather features on the overall pattern of traffic flow.

III. RESEARCH QUESTIONS

The following are the research questions that we tackled in our study:

- Can Supervised Machine Learning accurately predict the traffic flow?
- Does weather condition have any impact on the traffic congestion?
- How strong is the influence of weather condition on the traffic flow?

IV. RESEARCH METHODOLOGY

In this section we have addressed the approach and methodology we have used in the study to predict the traffic congestion on road network. The aim of the research is to study, analyze and use supervised machine learning to produce adequate results with minimal error.

A. Approach

Supervised machine learning is the approach we have used in this research to study the problem statement. With the experiments conducted on the dataset, we were able to make clear cut judgment for going with the supervised machine learning algorithms for fast and relatively accurate results. All the inputs (features) and outputs are labelled in the dataset that we want to train the model on. It is also important to note that the supervised machine learning is mainly used to deal two problem-sets: classification and regression. For accurate prediction of traffic congestion, we went with training the model based on continuous values of the traffic count. Congestion level of traffic count was varied with the help of the classification model. The accuracy and precision of scaled traffic count which is used in the research to define the level of congestion (LOC).

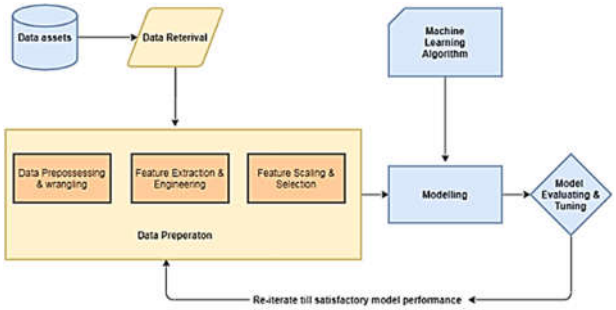

Fig. 1. Supervised Machine Learning Pipeline

The basic pipeline can be seen in Figure 1. The workflow for the proposed system in this research comprises of main steps, which involve the ingestion of raw data that has been obtained from the source and then apply the data processing techniques to wrangle, process and engineer meaningful features and attributes from this dataset. It helped us in evaluation of which machine learning algorithm is suitable for our data. The features selected from the data are used in the deployment of the model that we have selected. We have used the regression-based models to experiment with our dataset as data comprises of the continuous values. The regression model includes independent variables pertaining to month, hour, and a dummy variable for weekend, the geographical clusters variable, and direction, as well as weather data involving several of these variables.

B. Dataset Description

For the data acquisition process, we took the monthly historical traffic data for the month of March and year 2018 from the Department of Transportation (DOT) for New York Metropolitan Transportation Council (NYMTC). The New York City Dataset comprising of its four major Boroughs showed a decent distribution of values found within each Borough with no visible skewness or anomalies observed as stated in TABLE I.

TABLE I. BOROUGH DISTRIBUTION

No.	Borough	Number of values	Percentage of Dataset
1	The Bronx	1368	27.94%
2	Queens	1152	23.53%
3	Manhattan	1320	26.96%
4	Brooklyn	1056	21.57%

Due to unavailability of the Latitude and the longitude of each street within these Boroughs we had to use the external source to extract the exact geographical location of all of the 22 streets in the data set. We retrieved the geographical data using the Google Maps API, geocoding service was enabled to perform this task and results obtained were integrated into the dataset. We acquired the weather data from the external source as well for the same geographical location and timestamp in our dataset. Weather Data acquired for the same location have 9 attributes such as cloud cover, precipitation, dew point, relative humidity, precipitation cover, Temperature, visibility, conditions, and wind chill.

V. EXPLORATORY ANALYSIS

In this section, we have highlighted upon how we have processed the data, analyzed given features to derive relations. Exploratory data analysis (EDA) of the dataset is conducted in this chapter of the report. Various methods and approaches were used for scaling, standardization, and normalization of the dataset for model training and testing to obtain the satisfactory outcomes. Furthermore, we evaluated the performance of the selected model in this research for best model selection, validation of selected model and finally its evaluation is analyzed and discussed in this section.

A. Scaling Traffic Count

Traffic Count, being a continuous value within the data comprises of values ranging from 0 to 3000 that need to be normalized to a better form to relate with various features and extract useful relations with them. For vigilant representation of co-relations, Traffic Count Label was scaled and grouped by the respective starting and end locations. Values were scaled using min-max scaler and scaled into 4 equal divisions from 0 to 3 and termed as Congestion Divisions. The approach behind the min-max scaler is to subtract the minimum value in the feature and then divides by the range. The range is the difference between the original maximum and original minimum as shown in the Equation 1. The advantage of using a scaler is that the shape of the original distribution is preserved.

$$x_{sc} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Using this min-max scaler we have defined the level of congestion (LOC) based on traffic count and grouped into its respective borough, which is also normalized. LOC is categorized into the 4 discrete values: 0 represents low congestion level, 1 represents mild congestion level, 2 represents slightly high congestion level and finally 3 represents the high congestion level. The distribution of overall traffic count is shown in the described range from 0 to 3.

B. *Relations with weather features*

Based on domain knowledge, four relevant weather features (Conditions, Precipitation, Cloud Cover, Visibility) were selected that could influence traffic count. Their values were scaled to obtain a convenient severity level between 0 and 3, same as that of Scaled Traffic Count. The scaled divisions were made based on domain knowledge related to that feature. Finally, values were compared with Scaled Traffic Count to observe strongest intersections of similar severity values as a means to depict influence. The greater the number of overlapping of similar severities, greater will be the influence.

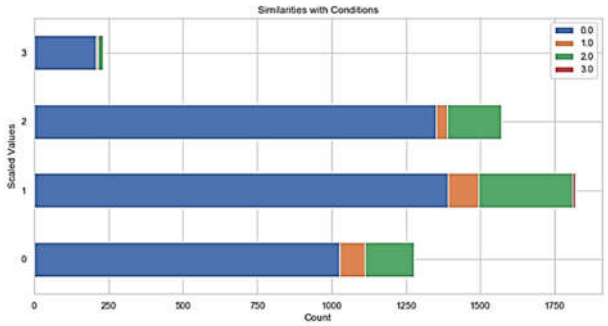

Fig. 2. Relation Count Distribution of Conditions and Scaled Count

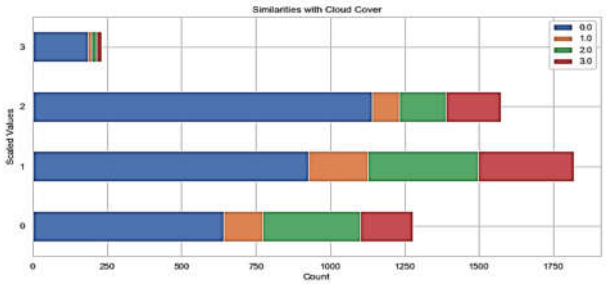

Fig. 3. Relation Count Distribution of Cloud Cover and Scaled Count

Comparison of scaled severity values of weather features with Traffic Count showed that greatest intersections arose with Conditions and Cloud Cover as shown in Fig. 2. And Fig. 3. Further evaluation done in the Feature Analysis section.

VI. Model Selection

A. *Base Accuracy*

Since our target label i.e. Traffic Count was continuous in nature, Regression Models were implemented as a means to provide best accuracies. Tested models comprised of Linear Regression, Lasso Regression, Decision Trees, Random Forest Regression (RF) [6], and Gradient Boosting Regression Tree (GBRT). Among all, Random Forest showed best results with an accuracy of 96.14%.

B. *Hyperparameter Optimization*

After optimizing the hyperparameters of all the regression models we have selected for this research, it was observed that the base accuracy of the Gradient Boosting Regression Tree was 92.39% increased to 97.12% using the Grid Search CV and validation of the model was evaluated by the shuffle spilt validation. Overall, we observed the improvement of 5.12% in the model accuracy. The gradual increment can be observed in TABLE II.

TABLE II. PARAMETER-TUNING ACCURACY

Accuracy	n_ estimators	max_ depth	min_ sample_leaf	max_sample _split
92.39%	100	3	1	2
94.38%	200	3	1	2
95.90%	100	5	1	2
92.83%	100	3	2	2
92.83%	100	3	1	5
96.22%	200	5	2	5
96.42%	400	5	2	5
96.62%	400	7	2	5
97.12%	400	7	5	5
96.67%	600	7	5	10

VII. Model Evaluation

Although after rigorous experimenting with optimizing the hyperparameters of the selected models, we were to generate correct validation results to evaluate the performance of the model on training data. However, fit of a proposed regression-based model should therefore be better than the fit of the mean model, so all models were evaluated via R^2 error in which the ratio of the variance of the model and the total variance of target variable was taken and put forward a value between 0 and 1 with 1 being the best one. MAE was also used to identify the difference between the forecasted value and the actual value. The general definition of the R^2 score can be seen in the Equation 2.

$$R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}} \tag{2}$$

Where,

$$SS_{res} = \sum_i (y_i - \hat{y}_i)^2 \tag{3}$$

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \tag{4}$$

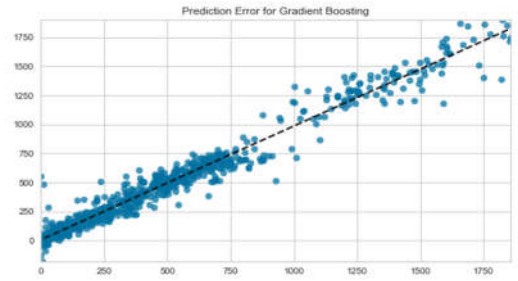

Fig. 4. R2 Accuracy for Gradient Boosting

VIII. Feature Analysis

To interpret an ensemble of our gradient boosting model, we used feature importance. These can be interpreted as the variables which are most predictive of the target. Trained model was tested for co-relations among all existing features which showed vibrant relations of Traffic Count with Conditions as standing out among all other weather features with the greatest feature importance, also observed in Fig 5, similar to our prior exploratory analysis in the section. Relations with Weather features above. Yellowbrick's feature importance method utilized the feature_importances_ parameter from Random Forest to give us clear relations. Extracted features were stored in a data frame and visualized.

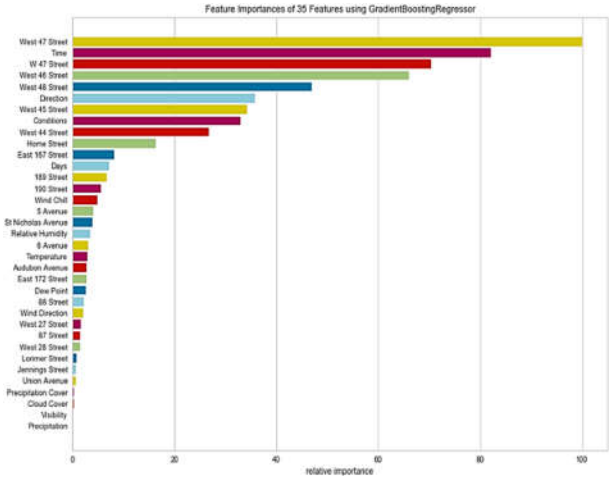

Fig. 5. Feature Importance

IX. Results

On successfully recording the highest accuracy for our model GBRT, the predicted results were visualized via a heatmap through the Folium library. The visualized points indicate the coordinates of the start and end of street locations of the designated routes within the data. Furthermore, the intensity of the color of the marker of that location expresses the intensity of traffic flow in that particular route. It ranges into 4 severity level of congestion with color palette of blue, green, yellow. The red color indicating a relatively high traffic flow and a darker blue color represents a relatively low traffic flow. Visualized results for time durations 7:00-8:00 a.m. describe as the peak hour for the traffic flow, leading to congestion, it can be seen in the Fig 6. With help of the visualization we can study and understand the traffic pattern.

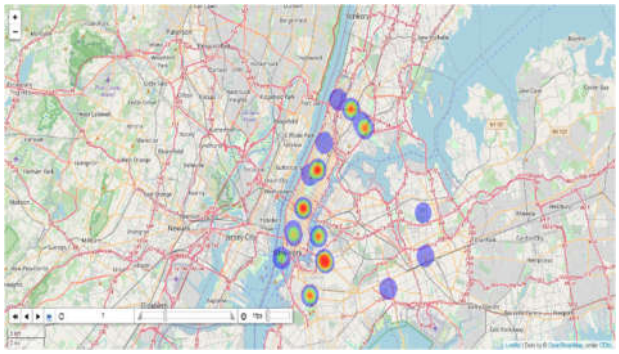

Fig. 6. Visualized Results

X. Conclusion and Future Scope

After detailed analysis, we have come to the conclusion that external factors like weather (in our case) do impact traffic congestion as a whole for the given dataset, and that our Gradient Boosting Regression Tree model records the best accuracy score for predicting Traffic Count after inducing the best parameters i.e. 97.12% considering all relevant features as mentioned in TABLE III. On the contrary, the greatest feature importance examined among all the weather features was of Weather Conditions, followed by Wind Chill.

TABLE III. Model Accuracy

Model	Base Accuracy	After Hyperparameter Tuning
Linear Regression	75.23%	71.60%
Lasso Regression	75.22%	71.50%
Decision Tree	93.24%	94.10%
Random Forest	96.14%	96.80%
Gradient Boosting	92.39%	97.12%

In this paper, the model is based on regular day predictions. However, we look forward to implementing a more robust model that will also consider the Planned Special Events (PSEs), like festival holidays, social events like concerts, sporting events like cricket and football matches and so on. Moreover, seasonal change may also affect the traffic flow largely because of the ambiguities in weather they may bring. Therefore, we also look forward to working with this aspect. We believe our research will pave the way for greater opportunities in the field of data gathering and will help in developing a more stabilized road network of the city which is less prompted to traffic congestion.

Acknowledgment

Our deepest gratitude to the Department of Computer Science at Forman Christian College, Lahore for providing us with the necessary resources needed to adequately conduct our research without any constraints.

References

[1] Sweet, M. "Traffic Congestion's Economic Impacts: Evidence from US Metropolitan Regions". *2013 SAGE journals-Urban Studies*, vol. 51, pp. 2088-2110, 2013

[2] Thianniwet, Thammasak & Phosaard, Satidchoke & Pattara-atikom, Wasan. "Classification of Road Traffic Congestion Levels from GPS

Data using a Decision Tree Algorithm and Sliding Windows". *2009 World Congress on Engineering*, vol. 1, 2009

[3] Jahangiri, A., & Rakha, H. A. "Applying Machine Learning Techniques to Transportation Mode Recognition Using Mobile Phone Sensor Data". IEEE *Transactions on Intelligent Transportation Systems*, vol. 16, pp. 2406–2417, 2018

[4] Jayapal, C., & Roy, S. S. "Road traffic congestion management using VANET", 2016 *International Conference on Advances in Human Machine Interaction (HMI)*, 2016

[5] P. Chhatpar, N. Doolani, S. Shahani, and R. Priya, "Machine learning solutions to vehicular traffic congestion," 2018 *International Conference on Smart City and Emerging Technology (ICSCET)*, 2018.

[6] Y. Liu and H. Wu, "Prediction of road traffic congestion based on random forest," 2017 *10th International Symposium on Computational Intelligence and Design (ISCID)*, vol. 2, 2017

[7] Chowdhury, B.,Kinhikar, M. & Alleema, N. N. "Road Traffic Prediction using Machine Learning". *International Research Journal of Engineering and Technology (IRJET)*. vol. 06, 2019

[8] M. M. Chowdhury, M. Hasan, S. Safait, D. Chaki, and J. Uddin, "A traffic congestion forecasting model using cmtf and machine learning," *2018 Joint 7th International Conference on Informatics Electronics and Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision and Pattern Recognition (icIVPR)*, 2018

[9] Huang, F.-R., Wang, C.-X., & Chao, C.-M. "Traffic Congestion Level Prediction Based on Recurrent Neural Networks". *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2020

[10] Jia Lu, & Li Cao. "Congestion evaluation from traffic flow information based on fuzzy logic". *2003 IEEE International Conference on Intelligent Transportation Systems*., 2003