https://doi.org/10.48009/3_iis_2020_53-58
Issues in Information Systems
Volume 21, Issue 3, pp. 53-58, 2020

PREDICTING VOLUME OF VEHICULAR TRAFFIC USING MACHINE LEARNING

Ernest Fountain, Texas A&M University-San Antonio, Email: efountain@jaguar.tamu.edu
Lo'ai Tawalbeh, Texas A&M University-San Antonio, Email: Ltawalbeh @tamusa.edu
Jeong Yang, Texas A&M University-San Antonio, Email: jeong.yang@tamusa.edu

ABSTRACT

Transport and communication are some of the most critical aspects of society. Various key sectors depend entirely on how goods are transported from one region to another. Therefore, having a clear picture of how transport routes are an essential consideration for any logistic company that intends to avoid inconveniences brought about by traffic congestion inroads. Besides, the comprehension of local traffic helps the local authorities in determining how appropriately they can plan urban centers to reduce congestion. In this study, a machine learning model has been developed with the primary purpose of predicting the volume of vehicles in a given area. The research employs three specific learning models, the Logistic Regression, XGBoost regressor, and Decision Tree Classfier. The performance of these three is compared, and the best algorithm is then used to predict the volume of cars given test data. From the comparison, the XGBoost regressor was found to have a better performance score as compared to the Logistic Regression and Decision Tree Classfier models.

Keywords: Traffic volume prediction, machine learning, logistic regression, performance evaluation

INTRODUCTION

Vehicle transportation is an integral part of our daily lives. There is, therefore, absolutely need to have an elaborate transport system that caters to the needs of the society in the most suitable manner. In the present age, various forms of transport have been introduced to supplement the commonly used roads for transportation. However, for relatively short distances, transporting goods and individuals by road is still the most popular mode embraced by individuals. In spite of its effectiveness, road transport has its downsides.

In urban areas, for instance, there is the constant problem of traffic congestion along the roads. This problem has various causes, among them being poor urban design as well as inadequate traffic control systems in place. The consequences of traffic jams are also significantly detrimental, especially when considering the economic impact that they leave on national economies. In 2018 alone, it estimated that traffic congestion resulted in a net loss of US$ 87 billion in productivity (LeBeau, 2019). The consequences of traffic congestion are not only felt in the national economy but also personal lives as well. There have been numerous reported instances of individuals losing their lives due to delays caused by traffic congestion. That is especially true when such people are in dire need of medical attention.

To resolve the issue, solutions have stayed put in place. These include proper urban planning, where roads are designed in patterns aimed at minimizing congestion. Traffic systems have also been put in place and designed to work in a manner that allows for the volume of traffic to be fixed in a way that reduces congestion. The introduction of mass transit systems has also been put in place to ensure that the number of vehicles on roads is condensed. The solution works on the basis that rather than using private cars, many individuals can use a single mass transit vehicle, eventually leading to a considerable decline in the number of cars on the roads. Although these efforts have proven to be successful in various ways, it is essential to consider the fact that there are still instances where the proposed methods do not perform as expected due to various reasons. It is from such failures that modern technology is being leveraged to help in combating the problem of traffic congestion. One such technique is machine learning and artificial intelligence (Tawalbeh 2017).

Machine Learning aims to teach the machines to act like humans. In other words, the machines learn to analyze, predict, and react by performing data analysis and training. This leads us to what is called data mining where tons of data is collected, processed and analyzed. Even though the specific technical functioning of data mining algorithms is quite complex -- they are a black box unless you are a professional statistician or computer scientist the uses and capabilities of these approaches are, in fact, quite comprehensible and intuitive. Data mining tells me about very large and complex data sets, the kinds of information that would be readily apparent about small and simple things. Data mining emerged as a result of trying to understand all of the information available to us. Since the beginning, people have used algorithms, mathematical models and statistical techniques to collect and analyze data. One of the most popular and useful technologies in this field is the automation of processes, which goes hand in hand with artificial intelligence.

Then, by searching for patterns and examining data to establish new unexpected relationships and correlations among elements will be key part of my process improvements initiatives. By studying these patterns, it will open a new area of understanding and provide deeper insights into transportation systems and planning to better serve the community (Dabiri 2020).

RELATED WORK

The bigdata analysis is getting more important in the last few years. The collected data is studied and analyzed to help the stakeholders and decision makers to plan and make better decisions. The data analysis techniques can be applied in many different aspects of our life such as education, transportation, cyber security research (Al-haija 2019), manufacturing, environment, and healthcare (Jararweh 2017).

Several studies devoted to solving the fundamental question of traffic congestion using machine learning and neural networks. Ata et al. conducted research and proposed a machine learning-based system that would solve the problem of congestion using Artificial Neural Networks. In the study, the authors utilized backpropagation techniques and artificial neural networks to inform drivers of possibly congested routes, thus asking them to seek alternative ways (Ata, 2019). Input parameters were fed to their model, which was then passed on to an artificial neural net. Backpropagation techniques were then conducted on the data. Finally, the results of the neural net would then be transmitted to drivers using RFID devices, and they would be projected using LCD screens. Fundamentally, this approach divided the neural net model into a fitting model and a time series model, and after that, the performance of the two models was evaluated. It was observed that the time series model had a better accuracy as compared to the fitting model.

Xing et al. studied on the Large-Scale Traffic Congestion Prediction based on the Symmetric Extreme Learning Machine Cluster Fast Learning Method with the sole purpose of contributing to existing solutions to the problem. With the motivation of the fact that the neural networks that were being used were prone to over-fitting, and such systems typically took a long time to train (Xing, 2019), the authors developed asymmetric extreme learning machine clusters whose performance was significantly faster than the conventional neural networks. The speed of this new technology arose from the fact that the approach would transform the presented data from large-scale data to small and medium scaled data, which could be easily managed. The application of this technology in traffic congestion control is that the model could be applied to different sections of a road along all the major roads in a given region by utilizing the transformation process of the technology. The researchers employed this technique due to the associated advantages such as increased training speed, boosted accuracy, effortless parallel acceleration of the training parameters, and the fact that the model employs a significantly reduced number of parameters in the training process.

In yet another attempt aimed at improving on the current neural network approach, Devi and Neetha explained how machine learning can be used to reduce traffic congestion in an IoT based smart city (Devi, 2017). The approach taken by the duo is built on the fact that the ideal city has a road network that is fitted with sensors that are connected to a centralized system. The data collected from the sensors, which are strategically located injunctions, are then categorized into T1 and T2. The T1 data forms the training set of the proposed algorithm, while the T2 information forms the test set of the algorithm. The researchers, after that, used five different algorithms to train their models. These included the Support Vector Machine, the Logistic Regression Classifier, the Random Forest Regressor, the Multilayer Perceptron, and the Decision tree classifier. Of these five, the Logistic Regression

classifier was found to have higher performance metrics with a precision of 100%, a recall of 99.5%, and an accuracy of 99.9%.

TRAFFIC CONGESTION: PROBLEM DEFINITION

In this section, the problem of traffic congestion is outlined, as well as the previously implemented solutions and the proposed solutions. An overview of the data provided is also discussed in this section.

Traffic Congestion in Urban Areas

It is crucial to consider the current situation of traffic congestion in urban areas. In a study conducted on traffic congestion concerning Polish cities, it was discovered that congestion arose mostly due to increased business entities in the Polish cities and the surge in passenger cars. The study revealed that the dynamic development of such cities prompted individuals to travel frequently on purposes ranging from business activities to recreational activities. These activities cumulatively resulted in a net increase in the number of automobiles in the cities, subsequently resulting in more traffic congestion (Koźlak, 2018).

Though based on Polish cities, the findings from the study can be applied to almost every city globally. Urban centers have become multi-purpose centers, thus attracting individuals from different parts of the globe to conduct their transactions. Without proper prior planning, this accumulation of individuals eventually leads to traffic congestion, subsequently leading to significant losses in terms of wastage of productive time.

Solutions to Traffic Congestion

Several solutions to traffic congestion have been implemented in various urban regions. One standard solution to this problem is the utilization of traffic lights providing the appropriate time scheduling for various junctions across different major cities. However, in spite of the popularity, this technique fails inadequately. The fixed periods, which are common in most traffic light systems, do not take into consideration the volume of vehicles present at any given time. Mass transit options such as buses and other subways systems have also been put in place in an attempt to reduce instances of traffic congestion. This approach has proven to be successful in cities with elaborate plans that can accommodate such mass transit systems. However, for cities with no such designs, this concept has not been implemented. Improving the currently existing infrastructure is also another way through which traffic congestion has been minimized in urban regions. By increasing the number of lanes of major roads as well as designing roads in innovative ways, municipal planning authorities have been able to come up with designs that can accommodate more vehicles (Reducing Traffic, 2019).

PROPOSED SOLUTION

In this study, a solution based on machine learning to the problem of traffic congestion is proposed. Here, data obtained from https://www.data.gov/ was used to design a predictive machine model that attempts to predict the volume of vehicles on a given highway. From the expected vehicle volume, directives can then be issued to direct the cars towards an appropriate route using either GPS techniques or the regular city traffic lights. The original data set initially consisted of 9 columns and 1279 rows. The data attributes in the columns were ID, Traffic Volume Count Location Address, Street, Date of Count, Total Passing Vehicle Volume, Vehicle Volume By Each Direction of Traffic, Latitude, Longitude, and Location. Table 1 defines each of those attributes. The process of training the machine learning almost takes the same process. The raw data is fed into the model for training to train the machine learning model. The trained model is tested for its accuracy against test samples, where its efficiency can be improved through progressive training of the model.

Table 1. Data Attributes Used in Analysis

Column	Data Type	Description: This field provides
ID	Int	ID of the count
Traffic Volume Count Location	Object	Street name of the count
Address, Street	Object	Address of the field.
Date of Count	Object	Date of the survey
Total Passing Vehicle Volume	Int	Total number of vehicles counted.
Vehicle Volume By Each Direction of Traffic	Object	Total number of vehicles headed towards a specific direction.
Latitude	float64	Latitude of the location to six significant figures.
Longitude	float64	Longitude of the location to six significant figures
Location	Object	Both the longitude and latitude to six significant figures.

They were proceeding to develop the model using the data as described above proved to be inappropriate. Various preliminary operations had to be conducted on the data to make it ideal for the development of the model. The first process to be undertaken was checking out for any missing values, where none was found. After that, the date of the survey was converted into a date-time data type rather than the object notation in which it was in. The total volume of vehicle count was then converted to an integer, and this was set aside as the target variable on which the machine learning model would have to predict. Next, the columns' location address, date of count, and street name were transformed to numerical values using Sklearn's label encoder.

The use of geographical data such as Latitude and longitude was discarded based on the restrictions and difficulties of accessing the Geopy API. The conversion of the features mentioned earlier into numerical values made it possible for one to identify the correlation between different values. The relationship was obtained using both the Pearson method and using the Seaborn library. Outliers in the data were identified and then removed to ensure that no bias was introduced in the model. Finally, the numerical values that were previously label encoded were the one-hot encoded using Sklearn's one hot encoder functionality. The primary objective of this step was to make the data features categorically so that they could be easily inputted into the model.

Finally, after conducting the preprocessing steps above, the training features were identified, and the target variable was set aside. Three specific predictive models were then initialized. These were the Logistic Regression model, the XGBoost model, and the Decision Tree Classifier model. To evaluate the performance of each of these models, the cross-validation score of the individual models was considered. This approach was taken due to its simplicity in showing how each model would perform in unseen data sets.

RESULTS AND DISCUSSION

Table 2 below shows the Mean validation score error for each of the three predictive models used, which represents the performance of each model. The result shows that the Decision Tree classifier gives the best results and has the least error.

Table 2. Mean Validation for Predictive Models Used

Predictive Model	Mean Cross-Validation Score
Logistic Regression	0.021563835645273928
XGBoost	0.3101373256586906
Decision Tree Classifier	0.009128006606564057

From the results, the XGBoost model was used to generate predictions for vehicle volume from the testing set as the model turned out to have the highest validation score with minimum recorded by decision tree classifier. During model training, there are chances that the model will not work with real data as its accuracy may deviate a lot from the expected results. This may result in noise introduced into the model during the training step; the cross-validation technique is used to feed the correct pattern. The training procedure involves reserving part of the data as the sample data set, which is later used for testing the model. The second step is using the rest of the data as a data-set to train the model. The trained model is checked against the private data. Model training using this method requires about 50

percent to be used as the data set and the rest as reserved data for testing. This technique is associated with high biasing levels (LeBeau, 2019).

Cross-Validation, also commonly known as rotation estimation, is a widely used statistical procedure in machine learning whose primary purpose is to investigate the performance of a given model. The use of cross-validation is basically to identify any possible instances of over-fitting that may occur as a result of using a specific model. The fundamental working of a cross-validation score model is based on the separation of provided data into two distinct categories, the training set, and the validation set. In typical situations, the training and validation sets are frequently crossed over subsequently to ensure that every single data set is evaluated in the model. In practice, k-fold validation is commonly used due to its simplicity and effectiveness. Here, the data is split into k equal parameters commonly known as folds. The model then retains every single fold for validation purposes while the remaining folds, k-1, are fed into the training process for learning purposes. The procedure is repeated until every fold in the data is used as a validation set. For small to medium data sets, the most commonly preferred number of k folds is 10. Other cross-validation processes widely employed are modified from this necessary operation of the k-fold cross-validation.

In other machine learning exercises such as classification problems, evaluation techniques such as confusion matrices are commonly employed. However, based on the nature of the issue tackled in this exercise, the use of confusion matrices was found to be inappropriate. Relying on the primary accuracy results from a classification model may provide one with the wrong perspective of the model in the sense that any given model may behave differently when presented with unseen new data (Cambridge (2016). Therefore, the accuracy score by itself cannot be considered to be sufficient for the evaluation of a model's performance. Typical disadvantages of using the cross-validation method include elevated type 1 error for comparison, underestimated performance variance, and overestimated the degree of freedom for comparison (Park, 2005).

CONCLUSION AND FUTURE WORK

This study practiced three machine learning models, the Logistic Regression, XGBoost regressor, and Decision Tree Classifier, to predict the volume of vehicles in a given area. The XGBoost model produced the best performance score with the testing data set. However, the model has been found to have a cross-validation score of 0.31, which is considerably low for practical machine learning purposes. The low value can be attributed to factors such as small data set for training and the failure to utilize all features optimally. The geographical data provided in the data set is significant in the determination of vehicle volume, but due to limitations outlined previously, they could not be put into practice. Future studies can, therefore, be conducted by incorporating the features above into the model for better results. To improve the accuracy of the prediction module, extensive data is needed since the accuracy of a model is directly proportional to the size of the data used.

The implementation of the proposed solution would yield the desired results, and it will be the right move towards automating transport processes. Application of machine learning in all economic activities, including the transport sector, would ensure timely delivery of goods and services. Initially, traffic control used to be carried out by human beings whose ability to make decisions is prone to errors. The speed of task execution by a machine is many times faster than if executed by human beings. Time is a very critical resource in modern business and saving time would guarantee high returns. Diverting traffic based on real-time monitoring system proposed would ensure that congestion on the roads is minimized and also lives would be saved. Time and resources usually lost in transit would be used in other resourceful activities.

Acknowledgment: This research is supported by the Expanded Chancellor Research Initiative (CRI) grant awarded to Texas A&M University-San Antonio, TX, USA. Grant awarded in 2019.

REFERENCES

LeBeau, P. (2019). Traffic jams cost the US $87 billion in lost productivity in 2018, and Boston and DC have the nation's worst. Retrieved from https://www.cnbc.com/2019/02/11/americas-87-billion-traffic-jam-ranks-boston-and-dc-as-worst-in-us.html.

Tawalbeh, H., Hashish, S., Tawalbeh, L. and Aldairi, A., 2017. Security in Wireless Sensor Networks Using Lightweight Cryptography. Journal of Information Assurance & Security, 12(4).

Dabiri, S., Marković, N., Heaslip, K. and Reddy, C.K., 2020. A deep convolutional neural network based approach for vehicle classification using large-scale GPS trajectory data. Transportation Research Part C: Emerging Technologies, 116, p.102644.

Al-Haija, Q.A., 2019, June. Autoregressive modeling and prediction of annual worldwide cybercrimes for cloud environments. In 2019 10th International Conference on Information and Communication Systems (ICICS) (pp. 47-51). IEEE.

Jararweh, Y., Al-Ayyoub, M. and Song, H., 2017. Software-defined systems support for secure cloud computing based on data classification. Annals of Telecommunications, 72(5-6), pp.335-345.

Ata, A., Khan, M. A., Abbas, S., Ahmad, G., & Fatima, A. (2019). Modeling SMART ROAD TRAFFIC CONGESTION CONTROL SYSTEM USING MACHINE LEARNING TECHNIQUES. Neural Network World, 29(2), 99-110. doi:10.14311/nnw.2019.29.008.

Devi, S., & Neetha, T. (2017). Machine Learning-based traffic congestion prediction in an IoT based Smart City. International Research Journal of Engineering and Technology (IRJET), 4(5).

Xing, Y., Ban, X., Liu, X., & Shen, Q. (2019). Large-Scale Traffic Congestion Prediction Based on the Symmetric Extreme Learning Machine Cluster Fast Learning Method. Symmetry, 11(6), 730. doi:10.3390/sym11060730.

Koźlak, A., & Wach, D. (2018). Causes of traffic congestion in urban areas. Case of Poland. SHS Web of Conferences, 57, 01019. doi:10.1051/shsconf/20185701019.

Park, D. (2005). Robust Cross-Validation Score. Communications for Statistical Applications and Methods, 12(2), 413-423. doi:10.5351/ckss.2005.12.2.413.

Cambridge (2016) Reducing Traffic Congestion and Pollution in Urban Areas. (2016, December 12). Retrieved from https://www.smartertransport.uk/smarter-cambridge-transport-urban-congestion-enquiry/.