CARNEGIE MELLON UNIVERSITY, AFRICA

FINAL PROJECT REPORT

MOBILE BIG DATA ANALYTICS AND MANAGEMENT
(04-800 X)

IMPROVING ROADS EFFICIENCY BY REDUCING TRAFFIC CONGESTION USING MACHINE LEARNING ALGORITHMS

By:

Kevin Shyaka

Moise Uwimana

Gabriel Ntwari

April 25, 2022

1. Abstract

Congestion control is a major issue in smart cities, affecting numerous difficulties that inhabitants encounter, including air pollution, fuel consumption, violations of traffic regulations, noise pollution, accidents, and time waste, as well as a range of other, missed opportunities. The availability of traffic data in various cities with a combination of machine learning models to predict traffic congestion can be used in traffic control and allocation of resources based on the predicted congestions. This work proposes a method of evaluating the congestion on different roads by classifying cars' speeds and the number of cars on a given road, or street at a given time. The method classifies traffic congestion into three classes: No congestion, mild and serious congestion. This is done by categorizing the average speed of each vehicle into mentioned classes. The prediction of classes was done by various classification models such as the k-nearest neighbor's algorithm, Decision Tree algorithm, support vector machine algorithm, and Random Forest algorithm. The models were evaluated using the following performance metrics: accuracy, precision, recall, and f1-score. K-nearest neighbor's algorithm provides the best accuracy in predicting congestions with 83.50% of accuracy on tests data in the data set. The models were deployed over San Francisco as a case study in order to estimate probable future traffic events in the city, in which a person at a given time can choose a road to pass in order to avoid traffic congestion in a specific direction.

1. Background and problem statement

Large cities are the most impacted by car traffic; every day, thousands of people go from their homes to their workplaces, schools, and activity centers, overloading the city's main highways, increasing air pollution, and difficulties associated with traffic congestion, such as automobile accidents[1]

One of the primary difficulties in developing nations is traffic congestion, which has an impact on people's everyday lives and harms economic and societal growth. The biggest issues of traffic congestion at the moment are traffic monitoring and predictions[2]. Congestion has a direct and indirect influence on a country's economy and the health of its citizens. According to Ayesha Ata et al[3], traffic congestion costs more than $1200 million per year in the US in terms of opportunity cost and fuel use. Time lost, particularly during peak hours, mental stress, and increased pollution to global warming are all key consequences of traffic congestion[4].

Traffic congestion management contributes to the country's development ensuring economic growth and the comfort of road users, which is unachievable without efficient traffic flow[5]. Authorities are focusing more on traffic congestion monitoring as the transportation industry evolves via the collection of traffic data. Therefore, traffic congestion forecasting provides authorities with the necessary time to arrange the allocation of resources to ensure that travelers' journeys are as smooth as possible.

The main goal of this project is to improve the efficiency of the roads by the provision of less congested routes to pass in to reduce traffic congestion. This in return will positively impact

human activities by reducing delays caused by traffic congestion, reducing accidents due to the high rate of collisions that occurred in traffic congestion, and others.

The best solution to address the identified problem is to predict accurate and timely traffic flow in all possible alternative roads to reach the same destination and recommend the road with low traffic jams to pass through which in return helps the driver to know the appropriate road to pass in. To achieve this, distinct techniques will be employed as the project proceeds. Dataset will be explored and analyzed to see different patterns in our data. Alternatively, some machine learning techniques will be used to predict the traffic flow. Comparing machine learning algorithms accuracies in predictions while predicting traffic in order to evaluate which type of model accurately gives the best result in terms of predicting traffic flow. In our project, we will mainly user defined functions in python to achieve our goal. After reaching our algorithm that predicts the least congested road to arrive at the same destination, it will be available to the users through a user-friendly system where it will predict the efficient road to pass the previous data.

2. Methodology

The information was obtained from the San Francisco Municipal Transportation Agency. The data are historical vehicle location information for San Francisco in 2020, including date and time, vehicle identifier, latitude, longitude, heading, and speed. The data provided depicts vehicle routes and city-wide traffic patterns at a given location and time. The data set contains approximately three million observations and six features. The first feature is "vehicle position date time," which provides the precise date and time of a vehicle with a data type of floating Timestamp; the second feature is "vehicle id," which provides a unique identification of a vehicle with a data type of numerical; and the third and fourth features, which are "Longitude" and "Latitude," respectively, provide a vehicle's location where the data type is numerical; The fifth feature displays the vehicle's heading or direction (in decimal degrees) from North. The final feature is "Average Speed," which provides data on the vehicle's instantaneous speed. (in mph)[6]

After obtaining the appropriate dataset for our projects, descriptive and predictive analytics were computed, with most of the tasks being completed in Python. Starting with descriptive analytics, tasks such as data cleaning, pre-processing, and explanatory data analysis were carried out. After being loaded into our working environment, which in our case is a jupyter notebook, the data was cleaned by handling missing values and anomalies and pre-processed using various python packages such as pandas and others. The story of traffic congestion on San Francisco roads, created using some python scripts given the longitude and latitude, was analyzed using interactive plots showing the location of each vehicle on distinct roads using Explanatory data analysis techniques. Furthermore, different histograms were plotted to thoroughly analyze the volume of vehicles at various time intervals such as hours of the day and days of the week.

Given the large volume of our dataset, which may cause computational issues during modeling, 5000 observations were chosen at random from the entire dataset. Given the average speed of the

vehicles, three categories of congestion were created: serious congestion, mild congestion, and no congestion [7].

Average Speed	Description of the speed	Traffic State level
<20	Low average speed	Serious congestion
(20,50)	Medium average speed	Mild congestion
>50	High average speed	No congestion

These categories were used to create multiple classes for the dependent variable. Some pre-modeling tasks were completed, such as standardization, feature engineering, and target imbalancement, in which the min-max scaler was used to normalize values, other columns were created and added to existing ones, and the SMOTE function was used to balance our target. When it comes to predictive analysis, the decision tree model was used to fit the data and predict the results. To assess the model's performance, evaluation metrics such as the classification report were used. To improve the results, we used grid search to tune various multi-class classification models such as decision tree, Knn neighbor, random forest, and support vector machine with their parameters. Finally, the best model with the best parameters was used to fit and predict the predictions.

3. Results and Discussions

After retrieving the data needed, the data was visualized on the map to see its distribution in the road as shown on the figure below.

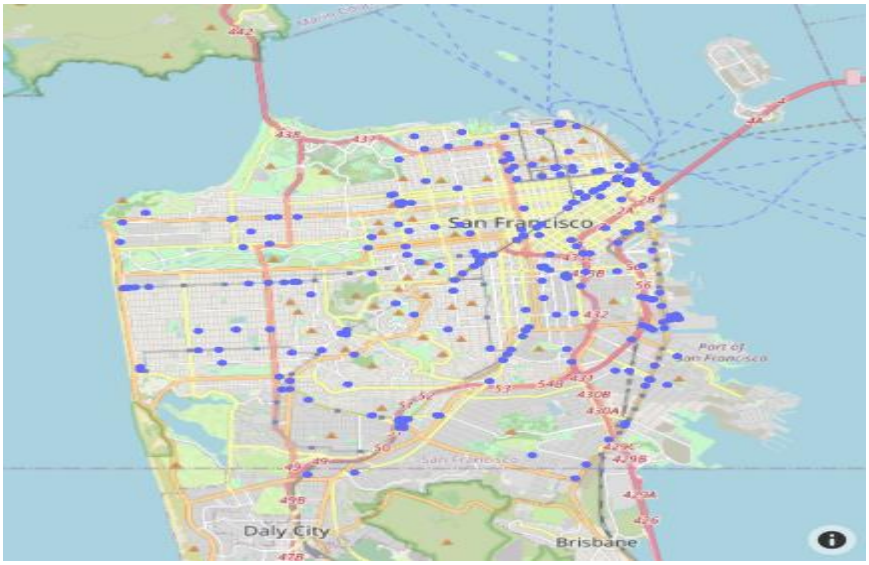

Figure 1: Traffic points in San Francisco

From the data obtained we saw that most of the road were in seriously congested state.

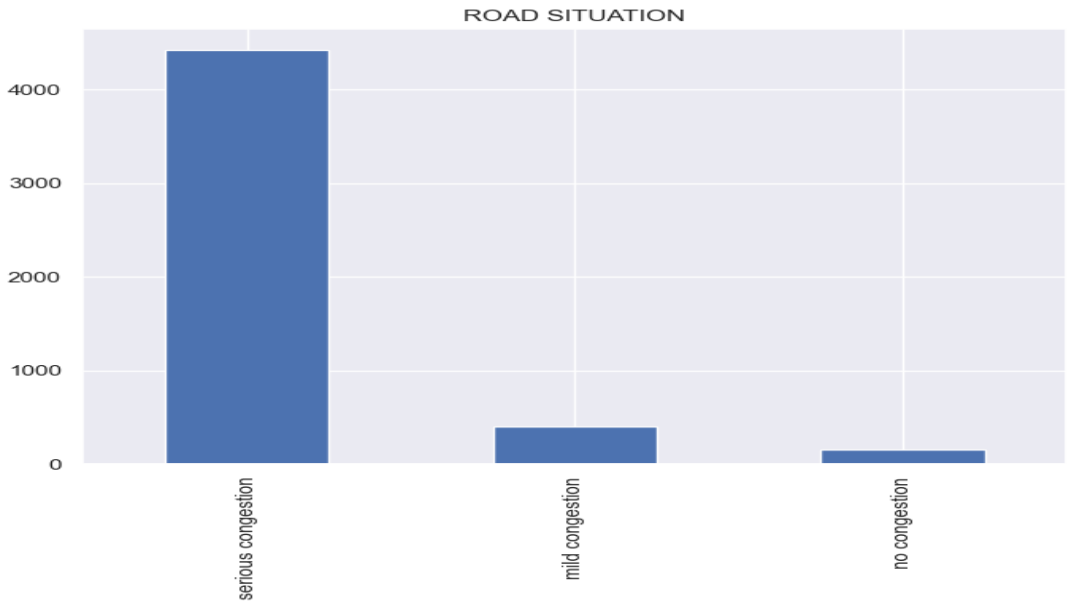

Figure 2:Congestion distribution in San Francisco.

By further exploring the dataset we saw that in rush hours, when people are going to work and coming from work, are the most serious congested hours.

Selecting the Best Model

In selecting the best model and its corresponding best parameter, a tuning technique that calculate the optimum values of hyperparameter called grid search was used. The result of each model is shown in the table below

	model	best_params	best_score
0	decision_tree	{'criterion': 'entropy', 'max_depth': 20}	0.816231
1	knn	{'metric': 'manhattan', 'n_neighbors': 3, 'weights': 'distance'}	0.822538
2	r_forest	{'bootstrap': False, 'criterion': 'gini', 'max_depth': 20, 'n_estimators': 60}	0.867916
3	svm	{'decision_function_shape': 'ovo', 'degree': 4, 'gamma': 'scale', 'kernel': 'poly'}	0.614009

Random forest model was identified as the best model after hyperparameter tuning.

Road Recommendation Results

After obtaining the best model, it was used to predict the best road to pass through at a given time in a particular direction.

For instance, a user might want to know the situation of the roads in the direction he/she is heading to at a particular hour, so he input in the python programming environment the following information.

Example of inserted data:

- Direction “west”
- Hour “8”

Recommendation in the form of graph

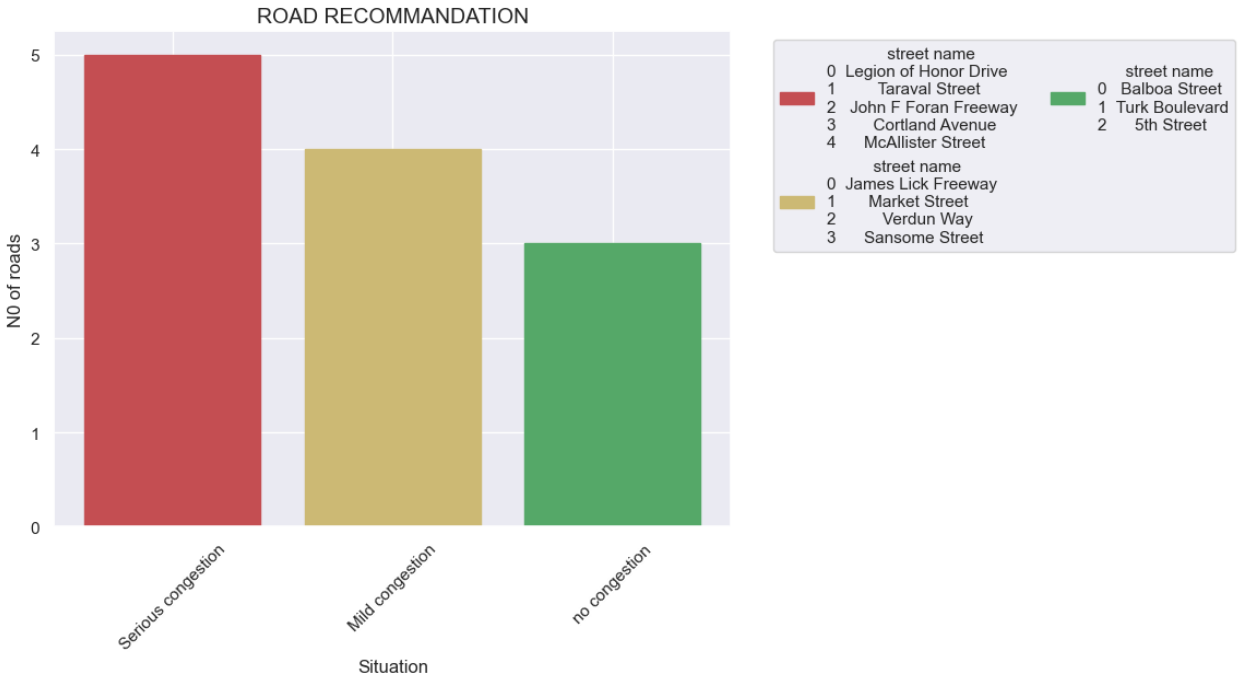

Figure 3: Predicted Roads suggestions

Interpretation: There are 5 serious congested road; however, he/she is recommended to pass in any three no congested road as indicated or at least in four mild congested road.

4. Reason learnt and Recommandation

Several lessons were learned as a result of this project. The first lesson we learned is that proper explanatory data analysis with a thorough examination of historical data leads to good modeling results. The second lesson is about the impact of a large amount of data. While running codes, we encountered computational issues. A lesson was also learned about the effects of target imbalancement on the model and how to handle that imbalancement using SMOTE oversampling. We learned how to provide street names using geographical python libraries given location coordinates such as longitude and latitude. A good recommendation would be to increase the number of roads in San Francisco, as descriptive and predictive analytics revealed significant congestion on more roads. More research is needed to determine how the model can be deployed so that the project can benefit drivers. Also, other models should be used such as deep learning to improve the accuracy of the results.

5. CONCLUSION

With the results obtained we were able to partially achieve the main goal of predicting the roads with and without congestions with the accuracy of 86% . With millions of data we have, we sampled around 5000 data for in order to reduce computation time. Despite trying several classification models , Knn neighbor's algorithm was obtained to perform better than the remaining on models. The results are logical since Knn neighbor's algorithm perform better for datasets with small number of features as we have. Additionally, a function that return all roads with predicted congestion level at a given time based on given direction have been constructed to facilitate the user with congestion .

Reference

[1] M. Saldana-Perez, M. Torres-Ruiz, and M. Moreno-Ibarra, “Geospatial Modeling of Road Traffic Using a Semi-Supervised Regression Algorithm,” *IEEE Access*, vol. 7, pp. 177376–177386, 2019, doi: 10.1109/ACCESS.2019.2942586.

[2] A. Elfar, A. Talebpour, and H. S. Mahmassani, “Machine Learning Approach to Short-Term Traffic Congestion Prediction in a Connected Environment,” *Transp. Res. Rec.*, vol. 2672, no. 45, pp. 185–195, 2018, doi: 10.1177/0361198118795010.

[3] A. Ata, M. A. Khan, S. Abbas, M. S. Khan, and G. Ahmad, “Adaptive IoT Empowered Smart Road Traffic Congestion Control System Using Supervised Machine Learning Algorithm,” *Comput. J.*, vol. 64, no. 11, pp. 1672–1679, 2021, doi: 10.1093/comjnl/bxz129.

[4] M. Akhtar and S. Moridpour, “A Review of Traffic Congestion Prediction Using Artificial Intelligence,” *J. Adv. Transp.*, vol. 2021, 2021, doi: 10.1155/2021/8878011.

[5] P. Goswami and D. Bhatia, “Congestion prediction in fpga using regression based learning methods,” *Electron.*, vol. 10, no. 16, 2021, doi: 10.3390/electronics10161995.

[6] “SFMTA - Transit Vehicle Location History (2020) | DataSF | City and County of San Francisco.” https://data.sfgov.org/Transportation/SFMTA-Transit-Vehicle-Location-History-2020-/48aa-8sj9 (accessed Apr. 25, 2022).

[7] T. Afrin and N. Yodo, “A survey of road traffic congestion measures towards a sustainable and resilient transportation system,” *Sustain.*, vol. 12, no. 11, pp. 1–23, 2020, doi: 10.3390/su12114660.