Predictive Modeling: An Attempt at Predicting Travel Times In Bengaluru Accounting For Geographic And Economic Effects

Jaison Sophia[1], S Althaf[2] and Nambiar Gautham[3]

[1] Student, SMS CUSAT

[2]Assistant professor, NIT Calicut

[3] Student, NIT Calicut

Abstract: The transport system of a country reflects the efficiency and growth of the country. As population increases, the number of vehicles increase, congestion and traffic increase leading to increase travel times, Evolution comes about in the transport system of a country, to increase physical connectivity and economic development, to reduce congestion and travel times. This paper aims to use machine learning algorithm on big data to understand how the effects of rainfall, temperature and pollution help predict travel times. The four machine learning algorithms used include linear regression, ridge regression, random forest regression and elastic net regression. The predicted travel times obtained by all models were compared with the observed travel times in order to determine which model gives better prediction. From the predictive modeling algorithms run on these datasets it is observed that, random forest regression is best suited in predicting travel times in Bengaluru City from i^{th} zone to j^{th} zone in the p^{th} hour of weekdays and weekends after accounting for effects of pollution, temperature, rainfall and economic activity.

1. Introduction

Bengaluru is a city that has relatively higher congestion levels compared to other cities in spite of smaller population. This is due to the limitations of older public transport networks which are primarily road based along with a significant growth in private vehicles [1]. Various reasons can be suggested for congestion including increasing level of motor vehicles, level of infrastructure, psychological factors, policy gaps etc. the average journey speed in Indian cities is also low [3]. Time to travel can vary because of numerous reasons like traffic, poor infrastructure, climatic conditions, mode of travel etc. Travel time variability analysis has been used by policy makers because a high value is allocated by users to the level of services attributes [2].

This paper aims to understand the influence of various factors like precipitation, rainfall, temperature and economic development in uber travel times and tries to develop a predictive model of travel time after considering these factors. These factors were used as a parameter to

train the algorithms. For this, quarterly average travel times of two hours ie. 10 am in the morning and 5pm in the evening across weekdays and weekends were considered separately. Quarterly average datasets of Aerosol Optical Depth, precipitation, temperature and nightlight data across wards of Bengaluru are considered. The travel times obtained through the predictive models used in this paper were contrasted against the observed travel times in order to determine which model gives better prediction. The four machine learning algorithms used include linear regression, ridge regression, random forest regression and elastic net regression.

2.Literature Review

Rapid expansion of India's cities has caused the surge in travel demand that in turn has caused a staggering impact on the inadequate transport infrastructure. The escalation in the ownership and use of motor vehicles has led to disturbingly high levels of congestion, noise, air pollution and traffic danger. Mobility and accessibility have declined for segments of the population [4]. Congestion may impede the growth of those cities and have negative aggregate implications for development. Policy makers have used a variety of policies to attempt to address this problem. These policies include increased road provision, the development of new subway lines or bus rapid transit systems, and quantitative restrictions that prevent cars from operating at certain hours depending on their plate number [5]. Denser, more populated cities are slower and there is a hill-shaped relationship between cities per capita income and mobility plus a city's mobility is related to characteristics of its road network. Economic development also brings about better travel infrastructure which facilitates uncongested mobility [6]. Mass rapid transit systems are to be adopted by all cities with populations exceeding one million according to a recommendation made by the National Urban Transport Policy of the Government of India with a view of recognizing the critical need to establish sustainable urban mobility [7]. Accessibility increases total wealth and redistributes wealth and economic opportunities will be lost for those places that do not grow their accessibility as economic opportunities shift to take profit of the accessibility benefits. New infrastructure contributes to agglomeration economies thus increasing aggregate output. Metro rail contributes to greater economic opportunity by improving accessibility especially with regards to the economic growth potential obstructed due to the dense, extensive traffic congestion [8]. Since the Delhi Metro Rail started in 2002, a gradual reduction in vehicle demand and reduced consumption of petrol, diesel and CNG was witnessed, according to 2007 estimates [9].

When Stochastic Response Surface Method was used in a study to investigate the expansion of stochastic response surface of travel time variation under uncertain factors of rainfall intensity and traffic volume, it was found that there is a significant influence of traffic accidents, traffic volume, and amount of rainfall, in travel times on the Hanshin Expressway study area. The most effective indicator in measuring performance of a transportation system was pointed out as travel time distribution. Uncertainties caused by the supply factors, demand factors and other external factors of a transportation system have impacted the properties of travel time distribution. Regression analysis was considered, to formulate a functional relationship between the travel time and various uncertain parameters [10]. This paper controls for precipitation, aerosol optical depth and temperature. According to a study conducted on Korean freeway, increase in rainfall

in unit time leads to increased non recurrent traffic congestion and thus reduces travel speed. Average non recurrent congestion surged due to heavy snowfall. Rainfalls and snowfalls occurring from 4:31–8:00 p.m brought about the highest non recurrent traffic congestion. But the rainfall and snowfalls witnessed after 8:00 p.m. or before 7:00 a.m. led to minimum measures in non-recurrent traffic congestion [11]. Direct and indirect effects of atmospheric aerosols and suspended particles have brought about a distressing influence on the human health, visibility, environment and air quality [12]. Statistically significant correlation was found to exist between Aerosol Optical Depth and the RSP (respirable suspended particulates) measured at the air quality monitoring stations. This clearly suggests that the satellite data is reliable source for aerosol-related air pollution studies. Aerosol Optical Depth data covers a more expansive area and has superior spatial resolution when compared with concentrations extracted from the ground-based air quality monitoring networks [13].

Machine learning techniques have been used to predict short term traffic congestion using vehicle trajectories [14]. Change detection in environment can be done through machine learning algorithms specifically classifiers [15]. Random forest is used to work on a road traffic forecasting model because of its exceptional high robustness, performance and practicability [16]. Machine learning methods have also been compared as short term prediction models of travel times [17].

Data

For this analysis travel time data was obtained from Uber movement. Uber provides data by multi-hour period by day, or alternatively, by individual hour by calendar quarter. Uber Movement divides each city into small zones and provides a travel time between zone pairs for each hour of the day [6]. The consolidated data provides quarterly average hourly travel times data across weekends and weekdays between zones from 2016 Q_1 to 2020 Q_2.

Quarterly average data of pollution, temperature and rainfall levels of each zone and intermittent zones were extracted from datasets found in Google Earth Engine. The dataset ERA5 Monthly aggregates - Latest climate reanalysis produced by European Centre for Medium range Weather Forecasts / Copernicus Climate Change Service was used for collecting quarterly average temperature of the 198 Bruhat Bengaluru Mahanagara Palike wards, for the 17 quarters. NASA Earth Exchange Global Daily Downscaled Climate Projections were used to collect quarterly average precipitation across the wards for the 17 quarters. Quarterly average Aerosol optical depth was extracted from the optical depth 047 band of the Moderate Resolution Imaging Spectroradiometer Terra and Aqua combined Multi-angle Implementation of Atmospheric Correction (MAIAC) Land Aerosol Optical Depth gridded Level 2 product produced daily at 1 km resolution (MCD19A2 V6 data product). Nighttime light data are regularly used to measure the intensity of economic activities on the earth surface, and for the estimate socioeconomic parameters. Nighttime light composite data can be extracted from the Visible Infrared Imaging Radiometer Suite (VIIRS) Day-Night Band (DNB) carried by the Suomi National Polar-orbiting Partnership (NPP) Satellite. NPP-VIIRS is equipped with wider radiometric detection range and on-board radiometric calibration, thus giving a more error free nighttime light source for economic modeling [18]. Aerosol Optical Depth, precipitation, temperature and nightlight data

which acts as a scale of economic development, are considered as potential attributes affecting the mean travel times of Uber services.

3.Methodology

Supervised machine learning algorithms are those functions that do prediction on past data or out of a sample. They understand the mapping of the function that generates output variable Y from input variable and train it on a portion of the dataset (train dataset) to produce a model which is tested against another portion of the dataset (test dataset). Linear regression is one of the simplest linear methods for regression. Linear regression does not involve any tuning. Ridge regression is used to control complexity, by regularization which involves considering near zero coefficients. L2 regularisation ensures that the coefficients are shrunk by the same factor evicting the possibility of elimination of coefficients and chances of creating sparse models. In ridge regression, high values of alpha restrict the coefficients, thus decreasing generalization. Elastic net is a regularised linear regression combining penalties of both lasso and ridge. The methodology is to record two statistics for each model trained- the R squared value and the Mean squared error, followed by predicting and evaluating the model on the train dataset and test dataset consecutively. R squared value depicts the proportion of variance in the dependent variable/output variable that can be explained by the independent variable/input variable in the regression model. Mean squared error is the average of the square of the difference between the predicted values and the actual value (this difference is termed as error).

The Y dataset contains past travel time data and the X dataset contains the aerosol optical depth measures, nightlight measures, temperature and precipitation measures of source and destination wards separately. To ensure the accuracy of the models, the dataset is split into train and test dataset with the training dataset as 60% of random observations of the original data and 40% constituting the test dataset. After training the algorithms to each train dataset, R squared value, Mean squared error and root mean squared error are calculated following which the same is done in the test dataset too. Travel times are predicted after considering the influence of pollution, precipitation, temperature and economic development, in the source and destination areas of travel, across both the training sets and test sets. Predicted travel times are then plotted against the test data values. Initially linear regression is performed, and values predicted on test dataset, but the results were not satisfactory with R squared value below 1. Linear regression can be described as a classical parametric method that sometimes requires clear cut modeling of nonlinearities and interactions. In a scenario, where the number of observations ‘n’ is greater than the number of variables, linear regression breaks down but this can be easily remedied by shrinkage methods like ridge regression, the lasso or the elastic net and so linear regression is treated as plausibly robust. As a result we run ridge regression on the datasets. For ridge regression, alpha value taken is 4. Elastic net regression algorithm is also trained and used to predict on the test dataset. Unlike the other machine learning algorithms discussed above random forests can be described as nonparametric and give way to the learning of interactions and nonlinearities without having to explicitly model them. They also work better when data mining through data in scenarios having an obviously larger number of observations than the number of

variables p or vice verse [19]. Random forests address the overfitting of training data, the n_estimators i.e. the number of trees to build is set as 300, with random state = 0.

The R squared values and the mean squared error obtained from each machine learning algorithm across four datasets which include travel times at 10 am on a weekday, travel times at 10 am on a weekend, travel times at 5 pm on a weekday and travel times at 5 pm on a weekend are tabulated and examined. Along with random forest regression, feature importance was also calculated to understand which features contributed to the prediction process and on how to avoid overfitting.

4. Results

	HOD= 10 WEEKDAY			
	R Squared	MSE	R Squared	MSE
OLS	-0.001530863	1412032.392	-0.006546337	1399429.541
Ridge regression	0.004187284	1403970.525	0.000481759	1389658.182
Random forest	0.923284884	108158.6532	0.436946299	782829.3174
Elastic Net regression	0.009840496	1396000.207	0.001172658	1388697.606

	HOD= 10 WEEKEND			
	R Squared	MSE	R Squared	MSE
OLS	-0.058657531	1083200.873	-0.052503569	1065794.578
Ridge regression	0.005883243	1017163.822	0.005556663	1007001.161
Random forest	0.922610132	79184.03275	0.43422916	572915.3903
Elastic Net regression	0.009894879	1013059.182	0.011396898	1001087.176

	HOD= 16 WEEKDAY			
	R Squared	MSE	R Squared	MSE
OLS	-5.69E-05	1591008.606	-0.000118611	1609395.352
Ridge regression	0.004062491	1584455	0.003496937	1603577.196
Random forest	0.926584659	116797.7942	0.441573297	898622.7535
Elastic Net regression	0.007492507	1578998.126	0.005265886	1600730.596

	HOD= 16	WEEKEND		
	R Squared	MSE	R Squared	MSE
OLS	-1.04E-09	1283543.05	-0.001063487	1309748.35
Ridge regression	0.00510323	1276992.833	0.004065103	1303038.323
Random forest	0.925349218	95817.49252	0.467462553	696749.059
Elastic Net regression	0.009655601	1271149.668	0.009279947	1296215.446

According to the R squared value of each model used, linear regression and ridge regression fail at predicting travel times with R squared values falling below 0.05 and for the former the values are negative. Elastic net regression models also have no significant values, but there is a slight improvement in the values thus validating the statement that it is a 'regularized optimization' for linear regression bridging ridge and lasso regression [20]. The best predictor model is the random forest regression that provides 92-93% accuracy across training datasets and 43-46% accuracy on test sets. This gap between the accuracy of the training dataset and test sets implies that further tuning and cross validation is needed, but among the four predictive analytics model used, random forest regression suits the best. The below given graphs that plot the predicted travel times by each model against the actual values, best describe the accuracy of the models.

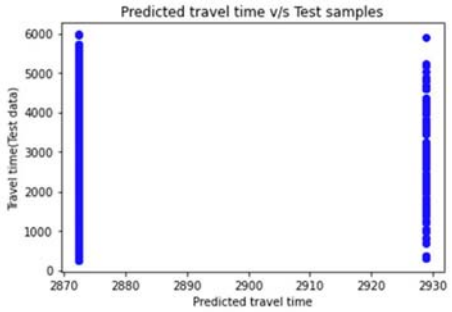

Fig 4.1 Plot of predicted travel times on test dataset against observed travel times using the Linear regression prediction model. It is clear that the travel times predicted were significantly different from observed travel times, thus making the model void for prediction of travel times with the given datasets and variables.

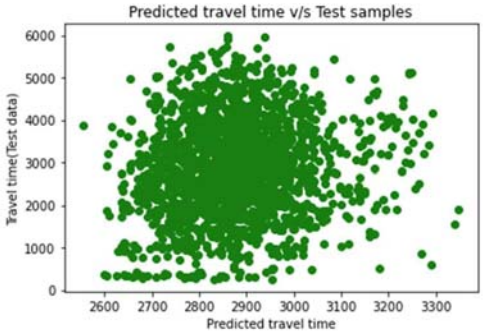

Fig 4.2 Plot of predicted travel times on test dataset against observed travel times using the ridge regression prediction model. It is clear that the travel times predicted were significantly different from observed travel times but it is a comparatively better at prediction compared to linear regression model, But the ridge regression model is void for prediction of travel times with the given datasets and variables.

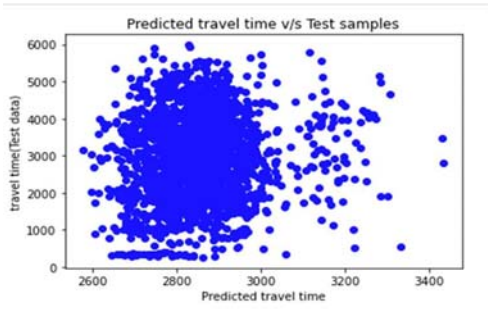

Fig 4.3 Plot of predicted travel times on test dataset against observed travel times using the Elastic Net regression prediction model. It is clear that the travel times predicted were significantly different from observed travel times but an improvement in prediction can be witnessed compared to the above two models, But the the model is still void for prediction of travel times with the given datasets and variables.

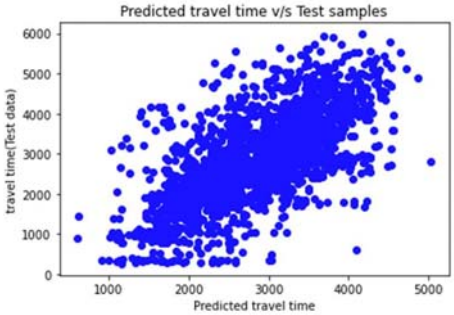

Fig 4.4 Plot of predicted travel times on test dataset against observed travel times using the Random forest regression prediction model. It is clear that the travel times predicted were closer to the observed travel times, thus making the model best ,out of the four models used, for prediction of travel times with the given datasets and variables.

On calculating feature importance of the input variables it was found that economic activity, precipitation and pollution at source has a more significant effect on the travel times compared to other variables used. Better accuracy on the test dataset can be gained by avoiding the other variables. Refer table 1.0

	cols	imp
3	viirs_source	0.343322
1	aod_source	0.299539
0	ppt_source	0.196098
4	ppt_dst	0.041708
5	aod_dst	0.037333
7	viirs_dst	0.035821
2	temp_source	0.029955
6	temp_dst	0.016224

Table 1.0 clearly shows that only three variables have a more significant effect on the prediction model.

5. Conclusion

This paper has identified random forest regression model as an apt model to predict travel times of Bengaluru City using past travel time data and variables like aerosol optical depth, precipitation, temperature and night light data (as a measurement of economic activity). In the event of growing usage of vehicles and increased urbanization this model can help predict the travel times and understand the effect of environmental factors and economic factors on a person's travel time. It can be helpful for new development and construction decisions. It can also influence decisions regarding other high-speed rail and mass rapid transit systems. But the Uber dataset does not strictly imply a person's travel time data. And considering the other modes of transport can also improve the applications of the prediction model. But there is a large difference in the R squared values gained from the train dataset and test dataset. This is because

of overfitting and maybe solved by excluding the variables with minimum feature importance and hyperparameter tuning.

References:

[1] Boston Consulting Group 2011 Unlocking cities: the impact of ridesharing across India.

[2] Duran-Hormazabal E & Tirachini A 2016 Estimation of travel time variability for cars, buses, metro and door-to-door public transport trips in Santiago, Chile *Research in Transportation Economics* , **59**

[3] Alam M A & Ahmed F 2013 Urban transport sysatems and congestion: a case study of Indian cities *Transport and Communications Bulletin for Asia and the Pacific*

[4] Pucher J, Korattyswaropam N, Mittala N and Ittyerah N 2005 Urban transport crisis in India. *Transport Policy* **12** pp 185-198

[5] Akbar P A & Duranton G 2017 Measuring the cost of congestion in highly congested city: Bogota

[6] Akbar P A, Couture V, Duranton G & Storeygard A 2018 Mobility and congestion in urban India .

[7] Ministry of Urban Development Government of India 2014 National Urban Transport Policy 2014

[8] Sharma R & Newman P 2018 Does Urban Rail Increase Land Value in Emerging Cities? Value Uplift from Bangalore Metro *Transportation Research Part A: Policy and Practice* **117**

[9] Sreedharan E 2007 Delhi Metro- the changing face of urban public transport in india *Indian Journal of Transport Management* **56**

[10] Chalumuri R S & Yasuo A 2013 Modelling travel time distribution under various uncertainties on Hanshin expressway of Japan *European Transport Research Review* .

[11] Chung Y 2012 Assessment of non-recurrent congestion caused by precipitation using archived weather and traffic flow data *Transport Policy* **19** pp 167-173.

[12] Gunaseelan I, Bhaskar B & Muthuchelian K 2014 The effect of aerosol optical depth on rainfall with reference to meteorology over metro cities in India *Environmental Science and Pollution Research* , **21**

[13] Lau Kai Hon LI C MAO , J & CHEN J.-C. 2003 A new way of using MODIS data to study air pollution over Hong Kong and the Pearl River Delta *Optical Remote Sensing of the Atmosphere and Clouds III;* **4891**

[14] Elfar A, Talebpour A, & Mahmassani H S 2018 Machine learning approach to short-term traffic congestion prediction in a connected environment *Transportation Research Record* **2672** pp 185-195

[15] Chan Jonathan Cheung-Wai, Chan Kwok-Ping & Yeh Anthony Gar-On 2001 Detecting the nature of change in an urban environment: A comparison of machine learning algorithms *Photogrammetric Engineering and Remote Sensing* **67.2** pp 213-226.

[16] Y. Liu and H. Wu 2017 Prediction of Road Traffic Congestion Based on Random Forest *2017 10th International Symposium on Computational Intelligence and Design (ISCID)* pp. 361-364

[17] D. Nikovski, N. Nishiuma, Y. Goto and H. Kumazawa 2005 Univariate short-term prediction of road travel times *Proceedings. 2005 IEEE Intelligent Transportation Systems, 2005* pp. 1074-1079

[18] Li X, Xu H, Chen X, & Li C 2013 Potential of NPP-VIIRS Nighttime Light Imagery for Modeling the Regional Economy of China *Remote Sens* **5 (6)** pp 3057-3081.
[19] Gromping U 2009 Variable Importance Assessment in Regression: Linear Regression versus Random Forest *The American Statistician* **63**
[20] Hans C 2011 Elastic Net Regression Modeling With the Orthant Normal Prior *Journal of the American Statistical Association* **106**