Travel Time Prediction using Machine Learning and Weather Impact on Traffic Conditions

Inspiring Excellence

BRAC UNIVERSITY

Authors:

Bilash Deb	13310009
Salehin Rahman Khan	14101197
Ashikul Haque Khan	14101001
Khandker Tanvir Hasan	14201061

Department of Computer Science and Engineering,
BRAC University

Supervised by:

Dr. Md. Ashraful Alam

Assistant Professor,
Department of Computer Science and Engineering
BRAC University

Declaration

We hereby declare that this thesis is based on results found from our own work. Materials of work found by other researcher are mentioned by reference. This thesis, neither in whole or in part, has been previously submitted for any degree. We carried out research under the supervision of Dr. Md. Ashraful Alam.

Signature of Supervisor

Dr. Md. Ashraful Alam
Assistant Professor
Dept. of Computer Science and Engineering
BRAC University

Signature of Authors

Bilash Deb
Author

Salehin Rahman Khan
Author

Ashikul Haque Khan
Author

Khandker Tanvir Hasan
Author

ACKNOWLEDGEMENTS

Thanks to Almighty Allah, the creator, and the proprietor of this universe for providing us direction, guidance and confidence to finish this research within time.

We are particularly thankful to Dr. MD. Ashraful Alam, our thesis supervisor, for his complete support and guidance in completing our research.

We are also grateful to the BRAC University Faculty Staffs of the Computer Science and Engineering, who have been a constant mentor for us in our entire undergraduate study period at BRAC University, especially for building our basics in education and improving our insights.

Lastly, we want to express our genuine appreciation to our parents and siblings for their support and care. We are thankful to all of our friends who helped us directly or indirectly by their support and encouragement to finish our thesis in time.

Table of Contents

List of Figures

List of Tables

ABSTRACT

The growth of Intelligent Traffic System (ITS) have recently been quite fast and impressive. Analysis and prediction of network traffic has become a priority in day to day planning in social, economic and more widespread set of areas. With a vision to further contribute to this vast field of research, we propose an approach to forecast level of traffic congestion on the basis of a time series analysis of collected data using machine learning. Moreover, the proposed approach allows us to find a correlation between varying parameter of weather and level of traffic congestion. Traffic data collected from Uber Movement for the city of Mumbai, India was fed to multiple of pre assessed machine learning algorithm. We have then analyzed the results of the different machine learning algorithms and see which algorithm efficiently provides us with the optimum accuracy of the prediction which is 85%. Thus, in this study, we attempt to find new knowledge between traffic congestion and weather by using big data processing technology. Changes in traffic congestion due to the weather is evaluated to create a long term prediction model and forecast traffic congestion on a daily basis. Hence a new prediction model that is an extension of the time-varying prediction model has been proposed as the part of this thesis that incorporates the change of occupancy caused due to weather conditions.

Keywords: Machine Learning, Traffic congestion, Forecasting, Weather, Intelligent Transport System (ITS), Support Vector Machine (SVM), Linear Regression, Correlation.

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Congestion can happen due to various conditions such as road work, peak hour traffic, accidents, and inclement weather conditions [1]. In this work, the traffic congestion caused by weather conditions is studied and the effect of weather conditions on mean travel time between different nodes of Mumbai city is analyzed and a prediction model is proposed based on the result of some learning algorithms to accurately forecast traffic congestion. Two ways can be considered to reduce congestion on urban freeways; one is to increase the total freeway capacity by expanding the number of lanes on the existing roads or new roads, but this requires extra lands and enormous expenditure on the infrastructure which is often not viable in many urban areas. Another solution is to use various traffic control strategies in order to efficiently use the existing freeways. The control strategies often involve predicting congestion levels and proactively managing the traffic before congestion is reached.

1.2 MOTIVATION

The increase of population density and of the relative amount of car owners makes traffic jams an important problem of modern societies. Traffic jams area major source of discomfort of drivers, but also the cause of an increased number of traffic accidents, especially in large cities. According to *The New Indian Express* A study has pegged the avoidable social cost of traffic congestion in Bengaluru at 38,000 crore Indian rupees annually [2]. The cost covers time delays, man-hours lost, extra fuel consumed, vehicle wear and tear, traffic accidents and environmental damage. The study, commissioned by taxi aggregator Uber and done by Boston Consulting Group, claimed India loses about 1.5 lakh crore Indian rupees annually due to traffic congestion in Delhi, Mumbai, Bengaluru and Kolkata. To add more to the list, it's not only cities with stupendous population that are facing this dilemma. Developed cities all around the globe are spending this extraneous cost just by wasting time on road. Below is a figure that shows the hidden cost of traffic congestion of various developed cities.

Figure 1.2.1: the hidden cost of traffic congestion in various cities [3]

With little being done to provide efficient transport solutions, people are getting used to spending more and more time commuting from one point to another. In appreciation of this problem, we wanted to create a model which will help to accurately predict these congestions and can be aided in various sectors from government planning to more personal daily basis planning.

1.3 CONTRIBUTION SUMMARY

The whole target was to build a forecasting model that would also take the weather conditions as a contributing factor to predict the traffic congestion. The results from this model can be interpreted and used in a different of ways as per the user's point of view. It can be used to improve the ITS by traffic management system of individual cities by analyzing their data through our system and generalizing a pattern of traffic movement to take action beforehand accordingly. This paper also attempts to find a correlation between several variables of weather and traffic congestion and outlines the effect of weather on traffic. To analyse the pattern of traffic flow we have fed our secondary data to different machine learning algorithms to compare the performance. So we have also outlined the prediction analysis of our selected algorithm that works best on our collected data from Uber Movement. Thus the contribution of this paper can be summarized as follows.

- A model is proposed for long term prediction of traffic congestion to be used from urban planning to be used by common people.
- Impact of several variable of weather on the traffic congestion is studied.
- Performance analysis of different machine learning algorithm on our secondary data from Uber Movement is outlined.

1.4 THESIS OUTLINE

The rest of the thesis is organized as following chapters:

- Chapter 2 is Background Information which consists of two sections. Firstly, it reviews the kind of data that have been used in our research and secondly "Literature Review" indicates our information collection repository.

- Chapter 3 is Methodology and it deals with the process and tells us about the algorithms that we have used for implanting the model.

- Chapter 4 is Result Analysis explaining the result that we have obtained and our implementation of the outcome that we have obtained.

- Chapter 5 is the Conclusion. It summarizes the overall research and portrays our vision for the future regarding this research.

Chapter 02

Background Information

2.1 UNDERSTANDING THE DATA SET

All of the data used in our research are secondary data. That is the data has been collected by other organization or entity that might have been collected in their own way for some other purpose [4]. Traffic data of Mumbai city from 2016 till date have been collected from Uber through a project they call Uber Movement. Data for weather for the city of Mumbai is collected from Wunderground for the same time period.

2.1.1 Traffic Data

This January, Uber unveiled "Uber Movement", a tool intended for use by city planners and researchers looking into ways to improve urban mobility. It provides anonymized data from over two billion Uber trips in the cities of Bogota, Boston, Johannesburg, Manila, Paris, Sydney, Washington D.C., Mumbai and are adding more cities to the list to help urban planning around the world. These data are open sourced and was targeted so that city officials can measure the impact of road improvements, major events and transit lines. So that planners and policy makers can analyze transportation patterns and make smart investments on future infrastructure projects and to power breakthrough insights and ideas with open data for all, specifically, it includes the arithmetic mean, geometric mean, and standard deviations for aggregated travel times over a selected date-range between every zone pair in each of these cities. Uber Movement is open to the public and can be download in .csv [comma separated value] format directly from [Uber Movement's Website]. Below are some tables that depicts the type of data that we have used from the Uber Movement site.

Table 1 shows an example of the mean time taken and its upper and lower bound for each day between two particular nodes of Mumbai city during a specific date range. This helps in classifying a pattern between each day of the week between two nodes for the whole year.

Day of Week	Origin Movement ID	Origin Display Name	Destination Movement ID	Destination Display Name	Date Range (M/D/Y)	Mean Travel Time (Seconds)
Mon	541	Mantralay Rd	108	R.B.I Branch	12/1/2017 - 12/31/2017	2340
Tues	541	Mantralay Rd	108	R.B.I Branch	12/1/2017 - 12/31/2017	2365
Wed	541	Mantralay Rd	108	R.B.I Branch	12/1/2017 - 12/31/2017	2466
Thu	541	Mantralay Rd	108	R.B.I Branch	12/1/2017 - 12/31/2017	2623
Fri	541	Mantralay Rd	108	R.B.I Branch	12/1/2017 - 12/31/2017,	2764

Table 2.1.1.1: Mean Travel Times day of the week

Each day is also segmented into 5 stages such as AM Peak, Midday, PM Peak, Evening, early Morning to study the pattern of the rush hour. Table 2 shows an example of the mean time taken and its upper and lower bound for each day for each of the aforementioned segments between two particular nodes of Mumbai city during a specific date range.

Time Of Day	Origin Movement ID	Origin Display Name	Destination Movement ID	Destination Display Name	Date Range (M/D/Y)	Mean Travel Time (Seconds)
Daily Average	541	Mantralay Rd	108	R.B.I Branch	12/1/2017 - 12/31/2017	
AM Peak	541	Mantralay Rd	108	R.B.I Branch	12/1/2017 - 12/31/2017	
Midday	541	Mantralay Rd	108	R.B.I Branch	12/1/2017 - 12/31/2017	2642
PM Peak	541	Mantralay Rd	108	R.B.I Branch	12/1/2017 - 12/31/2017	2677
Evening	541	Mantralay Rd	108	R.B.I Branch	12/1/2017 - 12/31/2017	2446
Early Morning	541	Mantralay Rd	108	R.B.I Branch	12/1/2017 - 12/31/2017	

Table 2.1.1.2: Mean Travel times time of the day

Table 3 shows the hourly aggregated data that further helped us to analyze the hourly pattern.

sourceid	dstid	hour of the day	mean travel time	standard deviation travel time	geometric mean travel time	geometric standard deviation travel time
1	3	18	4825.54	836.44	4753.35	1.19
1	4	12	4154.69	627.68	4109.5	1.16
1	5	6	1093	538.85	995.61	1.5
1	6	0	2860.52	611.29	2802.77	1.22
1	7	19	4526.77	772.83	4464.62	1.18
2	1	19	6936.63	1179.43	6844.01	1.18
2	3	7	4215.85	884.97	4130.23	1.22
2	6	14	2703.1	427.93	2670.2	1.17
2	9	21	1384.31	314.54	1349.61	1.25
3	1	8	3359	1128.74	3179.96	1.39
3	4	15	2855.21	481.24	2814	1.19
3	5	9	4318.09	1729.94	4061.49	1.39

Table 2.1.1.3: Hourly aggregated data between several combinations of nodes

Similarly, the data was also aggregated monthly to study the seasonal variation and the pattern in the mean travel time between selected nodes.

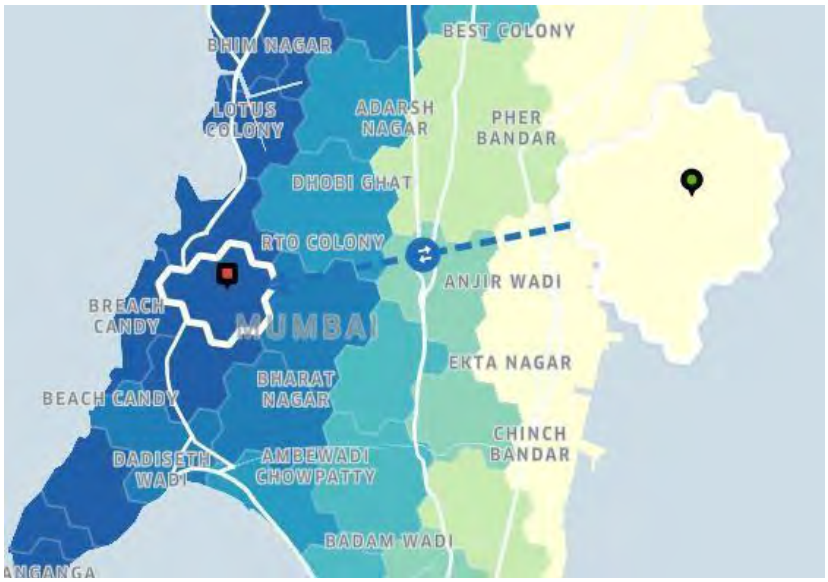

Figure 2.1.1.1: The Uber Web interface colors cells in the city grid based on the average travel time to them from the specified pin

2.1.2 Weather Data

Data of weather around Mumbai is collected from the year 2016 till date to match with the timeline. There are many factors of weather that have a combinatorial effect on different regions of a country. So far, we have narrowed down few key factors such as average temperature, humidity, dew point, wind speed, pressure and precipitation collected from Wunderground. Weather Underground or Wunderground is a commercial weather service providing real-time weather information via the Internet.

Weather Underground provides weather reports for most major cities across the world on its website, as well as local weather reports for newspapers and websites. Its information comes from the National Weather Service (NWS), and over 250,000 personal weather stations (PWS)[5]. The table below shows a portion of the weather data that we have collected from Wunderground.

Date (D/M/Y)	Average Temperature (°F)	Average Dew Point (°F)	Maximum Humidity (%)	Maximum Wind Speed (mph)	Maximum Pressure (Hg)	Average Precipitation (in)
13/09/16	80	76	89	13	29.83	0.04
14/09/16	82	76	89	13	29.83	0
15/09/16	78	77	100	13	29.8	3.15
16/09/16	79	77	94	8	29.74	0.35
17/09/16	78	77	100	14	29.69	1.54
18/09/16	77	77	100	23	29.78	1.38
19/09/16	79	76	100	14	29.83	0.87
20/09/16	78	76	100	14	29.83	5.16
21/09/16	76	76	100	15	29.83	2.68

Table 2.1.2.1: Parameters of Weather Data

2.2 LITERATURE REVIEW

Predicting travel time is difficult for various reasons. To predict travel time for a section of a road, traffic conditions or the speed changes of the vehicles along the section must be estimated. It is difficult to accurately estimate the traffic conditions since they can widely vary in spatial and time domain quite good amount of research has been done in the related areas of this field. To outline some of the relevant works, Bauza R. et.al, [6] proposed a cooperative traffic congestion detection based upon vehicle to vehicle communication for road traffic congestion prediction and got congestion detection probabilities of 90%. Manish R.Joshi[7] and Theyazn Hassn Hadi[7] did an intensive research on Different prediction

techniques and reviewed on network traffic analysis. Eric Horvitz et.al, [8] did a study on deployed traffic forecasting service. Their research has led to the deployment of a service named JamBayes that is being actively used by over 2,500 users via smartphones and desktop versions of the system. Jerome Treboux et.al, [9] did a short term prediction with more than 99% accuracy where they collected the data using sensors that they placed on different locations of Santander City. This is one of the few papers that has the weather constraint included. A research on a prediction model Jiwan Lee, Bonghee Hong and Kyungmin Lee [10], used 48 weather forecasting factors and attempted to correlate this data with multiple linear regression analysis. Mainly researches have been carried out in the following three categories.

2.2.1 Data Mining Research

Several studies have systematically reviewed data collecting methodologies, in particular collecting section based data such as travel time [11]. In [12], the authors have proposed a model on video based data collection. Recently, the proliferation of wireless communication infra-structures and navigation technologies have enhanced data collection and data coverage. These technologies (i) collect vehicle positions, (ii) infer relevant information concerning vehicular kinematic characteristics and congestion, and (iii) provide congestion information to drivers [13].

2.2.2 Traffic Flow Prediction Research

Historically, several authors have claimed that the configuration of a city's street network plays an important role in vehicular flow and, hence, used centrality measures of a street graph to model and predict traffic. Specifically, authors such as Turner [14] proposed betweenness centrality as a good predictor of traffic flow. Although Gao, et al. [15], criticized this approach and proposed a new model of traffic flow based on the non-uniform distribution of human activity and the distance-decay law.

2.2.3 Research with other taxicab data from GPS

Zheng, et al. [16] provide an interesting framework for analyzing taxicab data, which consists of linking pairs of regions (i, j) to three key features: (1) the number of taxis going from region i to region j, (2) the average speed these taxi drives when commuting from region i to region j, and (3) the ratio between the actual travel distance and the distance between the centroids of these two regions. By mapping taxi trajectory data from 30,000 taxis driving in Beijing from March to May in 2009 and 2010 onto this framework Zheng et al. seek flaws in current urban planning.

Chapter 3

Methodology

3.1 UBER MOVEMENT: TRAVEL TIME CALCULATION METHODOLOGY

Although the data collected comes with the mean travel time between nodes, it is worth mentioning in this paper how the travel times are actually calculated. The Uber Partner app, while on trip, records latitude, longitude, and a timestamp (Date/time) every 4 seconds. These GPS trace pings are commonly used to provide navigational routing, fare calculations, match partners with riders, and user experience elements, such as plotting the position of the car in the Uber Rider app. When aggregated, these GPS trace pings can also be used to derive average travel times between the zones in a given region. Uber Movement processes these GPS trace pings using the following high-level steps:

STEP 1 - Zone Assignment: For each trip, unsorted GPS trace pings are assigned an appropriate zone as defined by a shape file.

STEP 2 - Mean Epoch: For each zone a trip passes through, the mean GPS ping within that zone is computed. After this step, the overall trajectory is lost but we do know the average timestamp within each zone a trip passed through.

STEP 3 - Zone to Zone Travel Time: The elapsed time from each mean GPS ping to all subsequent GPS pings is measured, thereby providing zone-to-zone travel times for each trip.

STEP 4 - Aggregate Trips: Zone-to-zone travel times are aggregated from all trips. After this step, trip level information is lost and we only know statistical measures of Zone-to-zone travel times aggregated from many trips.

STEP 5 - Privacy Constraints: Travel time statistics are removed for zone pairs that either a) do not meet a minimum number of trips or b) the minimum count of unique riders necessary to preserve rider privacy. (This step is implemented in tandem with Step 4 but listed as a separate step for ease of understanding)

STEP 6 - Release: Zone-to-zone travel time averages are made available via Movement's interactive travel time's solution, including several available CSV export options.

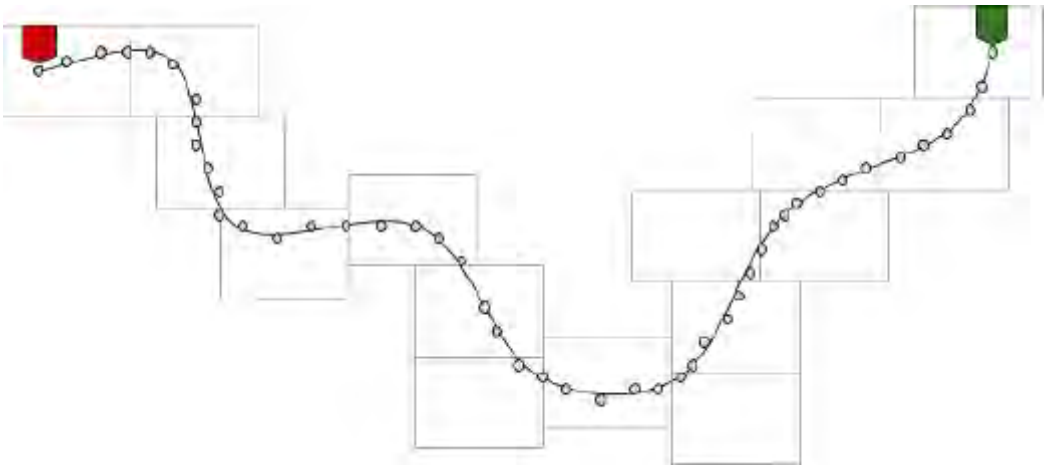

Figure 3.1.1: The zones passed by a Uber trip between the colored starting and end points.

3.2 LEARNING MODELS

The data in the format of .csv file has been read as input and selected classifiers called upon from an array have been used to train upon this data set. Data was split into two parts; one the training set and the other the testing set using train_test_split function in Sklearn. The *test_size=0.2* inside the function indicates the percentage of the data that should be held over for testing. The ratio was altered to see the performance of each algorithm upon different ration of training and testing data. Below is a figure of the workflow of our methodology.

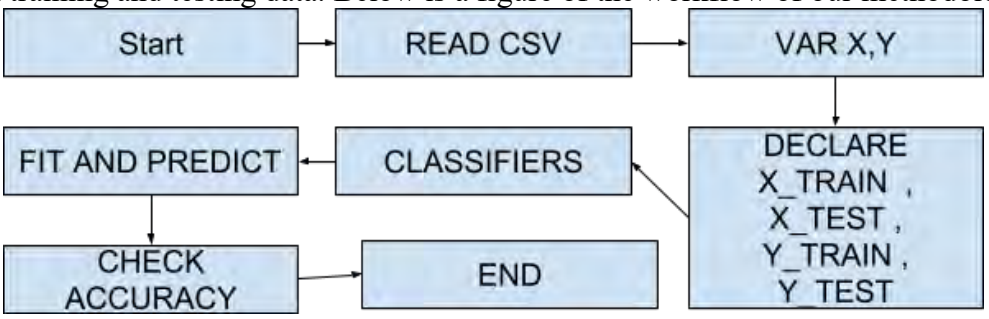

Figure 3.2.1: Workflow of the system

The selected classifiers for our model are described in the section on the next page. They were selected on the basis of their ability to recognize pattern in data set

3.2.1 Decision Trees.

A decision tree is a predictor $h: X \rightarrow Y$, that predicts the label associated with an instance x by travelling from a root node of a tree to a leaf. At each node on the root-to-leaf Path, the successor child is chosen on the basis of a splitting of the input space. The main purpose of Decision Tree is to shrink the training dataset in the smallest tree [17].

Pseudo code:

INPUT: training set S, feature subset $A \subseteq [d]$
if all examples in S are labeled by 1, return a leaf 1
if all examples in S are labeled by 0, return a leaf 0

if $A = \emptyset$, return a leaf whose value = majority of labels in S
else :

Let j = argmax$i \in A$ Gain(S, i)
if all examples in S have the same label
Return a leaf whose value = majority of labels in S
else

Let $T1$ be the tree returned by ID3($\{(\mathbf{x}, y) \in S : x\,j = 1\}$, $A \setminus \{j\}$).

Let $T2$ be the tree returned by ID3($\{(\mathbf{x}, y) \in S : x\,j = 0\}$, $A \setminus \{j\}$). Return
the tree

Different algorithms use different implementation of $Gain(S, i)$. The simplest definition of gain is the decrease in training error. One of such gain measure is information gain, the equation for information gain is given below,

$$IG(A) = I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - remainder(A) \qquad (1)$$

Where

$$remainder(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I\left(\frac{p_i}{p+n}, \frac{n_i}{p+n}\right) \qquad (2)$$

3.2.2 Random Forests

A random forest is a classifier consisting of a collection of decision trees, where each tree is constructed by applying an algorithm A on the training set S and an additional random vector, θ, where θ is sampled independently and identically distributed from some distribution. The prediction of the random forest is obtained by a majority vote over the predictions of the individual trees. It is very efficient on large data sets. Random forest uses

Gini index for deciding the final class of each tree. If data set T contains examples from n classes Gini index, Gini (T) is defined as

$$Gini(T) = 1 - \sum_{j=1}^{n} (P^2)_j \qquad (3)$$

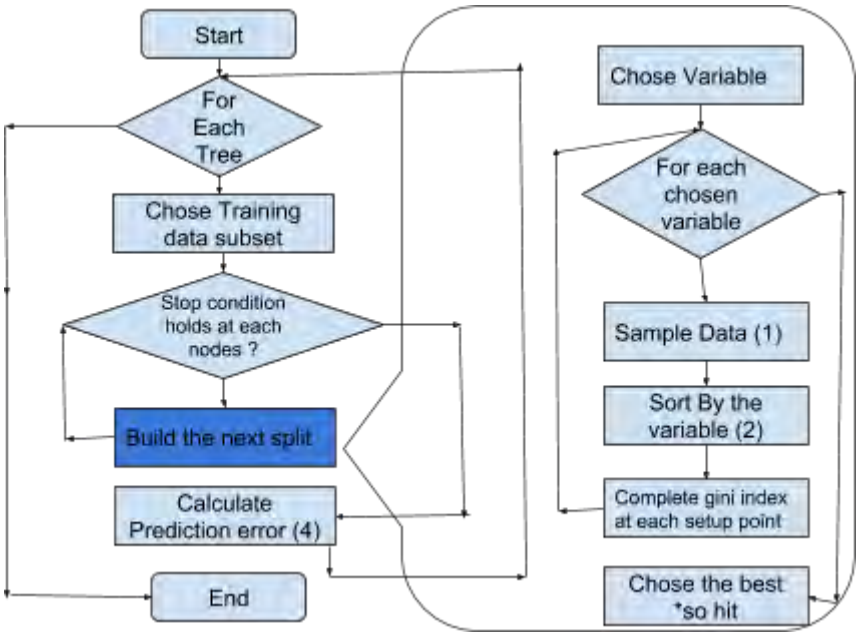

Figure 3.2.2.1: Flowchart of Random Forests [18]

3.2.3 Linear Regression for polynomial regression task

The hypotheses of the multivalued regression analysis are

$$h_0(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n (4)$$

What linear regression does is that it tries to plot a best fit line through a scatter diagram of recorded points, the linear equation of the best fit line is the linear squared regression equation where the value of dependent variable can be found out from one or more independent variables.

The best fit line is found out by decreasing the average distance of original value to the points on the linear equation. This distance is called the cost function which is calculated by the formula

$$Costfunction(J_\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \qquad (5)$$

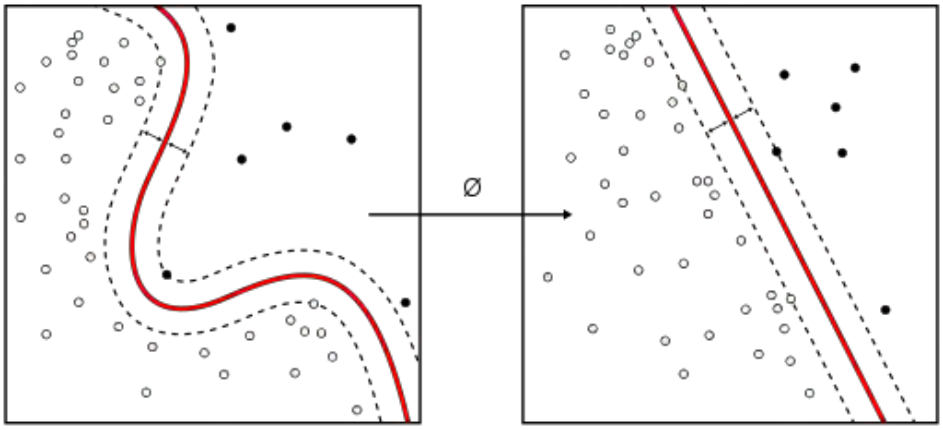

Figure 3.2.3.1: Linear Regression (Source: Wikipedia)

In a regression analysis, after first including all the variables, the variables that are not thought to be important are removed as the regression model is developed .This practice is necessary for preventing an increase in $R-$ squared value because of large number of independent variables and for removing the independent variables that are highly correlated. Therefore, it is necessary to use the variable removal method in order to leave the optimal variables only

3.2.4 Logistic Regression

Logistic Regression searches the whole datasets to find the hyper plane which fits the most for identifying the classes. The core of logistic regression is “logistic function”. Logistic function is also called the sigmoid function. This Function was mainly developed for describing the properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It is a ‘S’ shaped curve which can take real-valued number and map it into a value between 0 and 1. The function given below:

$$\emptyset_{sig}(z)=\frac{1}{1+\exp(-z)} \qquad (6)$$

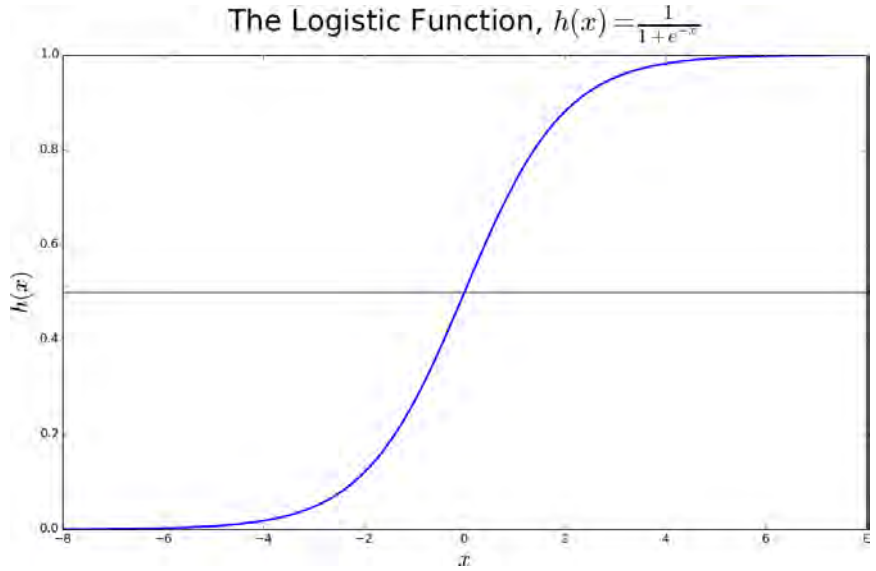

Figure 3.2.4.1: Logistic Function Curve

3.2.5 Support Vector Machine(SVM)

Support vector machines (SVM) are kernel machines that implement maximum margin methods. The maximum margin is generated by the kernel using a set of weighted vectors of training data called support vectors. Basic concept of this algorithm is finding a hyper plane in order to classify the datasets. There are two kinds of SVM classifier –

a) SVM Linear Classifier

b) SVM non-linear Classifier.

The flowchart of SVM classifier is given below:

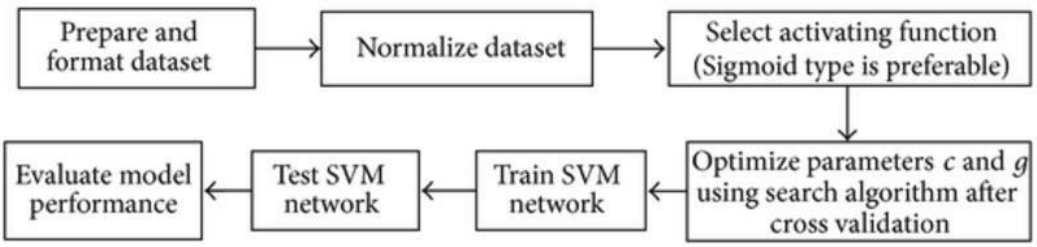

Figure 3.2.5.1: SVM Workflow

SVM uses the quadratic approach to define the problem of maximizing separability between classes. The margin is subject to constraint of the smoothness of the solution.

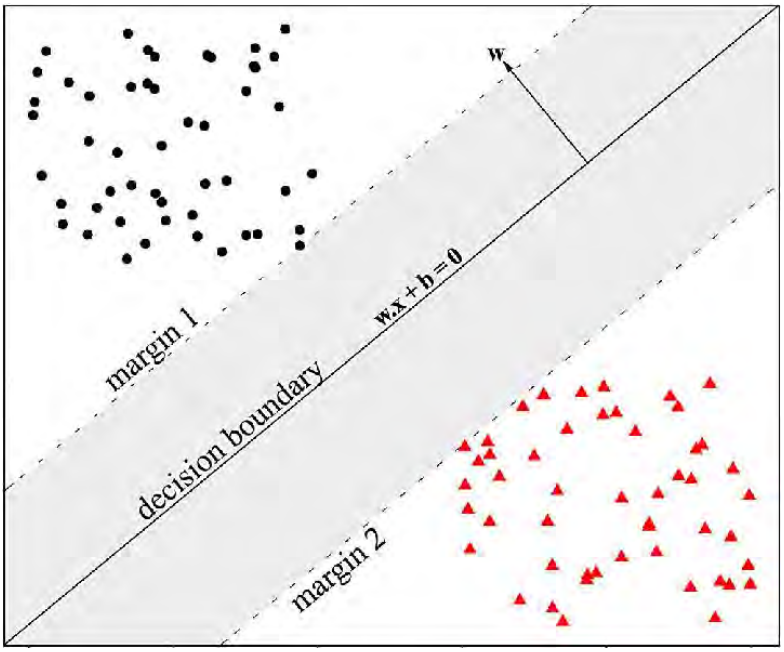

Figure 3.2.5.2: Main Components of a Binary Support Vector Machine [19]

3.3 METHODOLOGY USED TO ANALYZE WEATHER IMPACT ON TRAFFIC

In order to analyze how the weather conditions, affect traffic congestion, the product moment correlation coefficient between each of the parameters of weather discussed in Chapter 2 and the mean travel time taken between two nodes in different days in varying weather conditions is calculated.

3.3.1 Correlation Coefficient

Correlation coefficients for two variables signify the degree of linearity between them. For sufficient amount of data, the degree of linearity can be measured as strong, positive, negative or no correlation. Correlation coefficient ρ for two variable x and y is defined as:

$$v = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1\,i}^{n} x^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1\,i}^{n} y^2 - (\sum_{i=1}^{n} y_i)^2}} \qquad (7)$$

The value of ρranges from $-1 < \rho < 1$, where $\rho = 1$ represents a perfectly positive correlation, that is, the sample data points of (x, y) lie on a straight line with a positive slope and $\rho = -1$indicates a perfectly negative correlation.

Months used for analysis was chosen based on months with typical winter, rainy and typical summer to observe seasonal effects.

3.4 DATA USED IN LEARNING MODELS

If we look back to the type of data shown in the tables in Chapter 2, we would see that data from Uber Movement consists of data from one node to all possible nodes of Mumbai city. The source ID(sourceid) and destination Id(dstid) were labeled by Uber Movement. That is a sourceid of 541 will always represent the location Mantranalay Road and the dstid 108 will always represent the location R.B.I Branch of Mumbai city. So for the machine learning part, the feature set for the training included the source id, the destination id, time of the day, day of the week for every combination of origin and destination between nodes that is available. The data were subdivided into four .csv files, one for each quarter of the year. The accuracy is calculated, for testing data in the same quarter of the same year.

To analyze the pattern, we have selected a single mother node and have analyzed the data from that node to 10 other nodes over the period of 2016 till date. The choice of these nodes were not random but rather were particularly selected from prior knowledge as they require to travel through busy highways and intersections to travel from one node to the other. The algorithms were trained upon three models of data

Model 1: Using data from table 1, Travel time day of the week to analyze overall mean time to travel between the mother node to the other nodes in each day of the week.

Model 2: Using data from table 2, Travel time hour of the day to see a pattern of congestion during the rush hour.

Model 3: Using monthly aggregated data to see a seasonal variation in the pattern.

The data set were split into two parts, i) the training set which was used to train the learning models and ii) the testing set, to figure out the accuracy of our prediction. The ratio of training set to testing set were also varied to study the learning curve of each of the machine learning algorithms.

Weather data was not directly used as a parameter to train the algorithms but rather our study was merely to find a correlation between weather variables and mean travelling time. The correlation for each of the factors of weather and the mean travel time for the same day were measured using the method described in section 3.3

Chapter 4

Result Analysis

4.1 LEARNING PERFORMANCE

Throughout these models we are predicting the mean travel time between nodes. As the output is a numerical value the performance of the algorithms was evaluated using the root mean squared error. Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

The formula is shown below:

$$RMSE = \left[\sum_{i=1}^{N} \frac{(z_{f_i} - z_{o_i})^2}{N}\right]^{\frac{1}{2}} \qquad (8)$$

Where, $z_{f_i} - z_{o_i}$ = difference between original and predicted value and N is the sample size.

The table below shows the percentage accuracy that we have obtained for the learning algorithms that we have used for different size of training set.

	Percentage of Data in Test Set			
	40%	50%	60%	80%
Algorithm Name	**Accuracy**			
Decision Tree Regression	42%	54%	68%	73%
Random Forest	56%	62%	74%	83%
Multivalued Linear Regression	48%	69%	75%	84%
Logistic Regression	48%	63%	78%	85%
SVM	35%	60%	68%	70%

Table 4.1.1: Percentage accuracy of algorithms for different size of data sets

Percentage accuracy was measured according to the formula below:

$$Percentage\ accuracy = (1 - error) \times 100\% \qquad (9)$$

Below is a chart that shows the percentage accuracy for each of the machine learning algorithms for different ratio of testing and training set.

Figure 4.1.1 : Percentage accuracy for different algorithms

The learning rate of multivalued linear regression for different size of training and testing set is shown below.

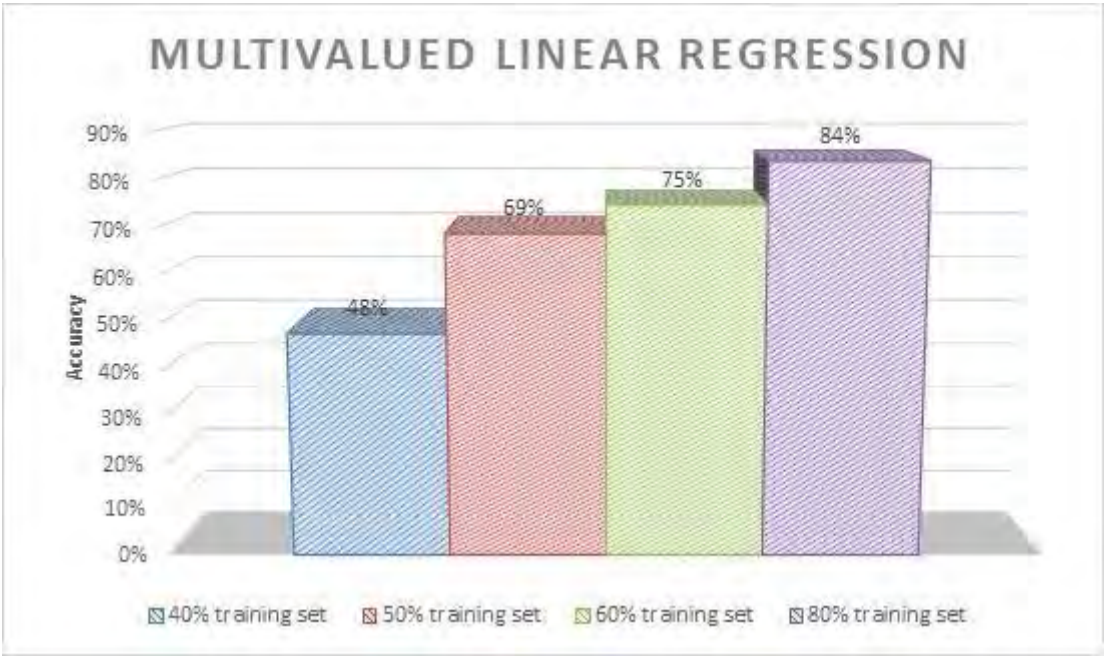

Figure 4.1.2 : Multivalued linear regression learning rate.

4.2 TREND ANALYSIS

Some expected pattern was found in the behavior of the commuters of the Mumbai city. Predictably, the mean travel time during rush hour is upper bounded to be 11.9% longer than travelling in early morning. The pattern is the same for all of the routes between the mother node to the other nodes. The algorithms generated some charts that depicts the pattern in our study which are shown below. Figure 4.2.1 to 4.2.4 shows the average travel times by day of the week for four quarters of the year 2017

Figure 4.2.1: Mean travel time by day for quarter 1[January 2017 – March 2017]

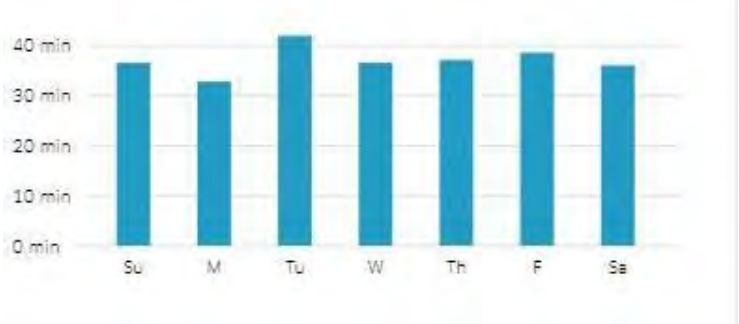

Figure 4.2.2: Mean travel time by day for quarter 2[April 2017 – June 2017]

Figure 4.2.3: Mean travel time by day for quarter 3[July 2017 – September 2017]

Note that the mean travel time increased for almost every day for quarter 3 than quarter 2 as that quarter is subjected to heavy rainfall in the region of Mumbai

Figure 4.2.4: Mean travel time by day for quarter 4[October 2017 – December 2017]

Figure 4.2.5 and 4.2.6 shows the average travelling time by period for each day for two of the quarters in the year 2017

Figure 4.2.5: Mean travel time by period for quarter 2[April 2017 – June 2017]

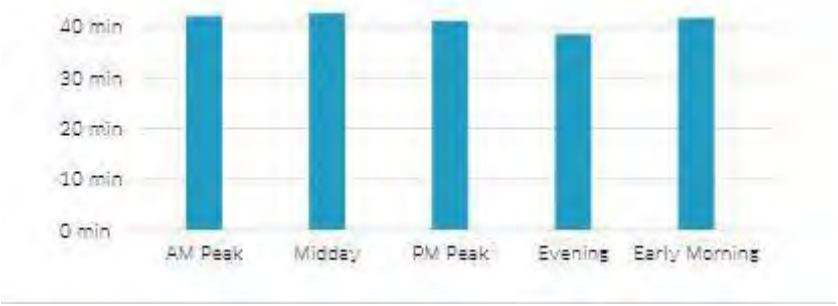

Figure 4.2.6: Mean travel time by period for quarter 3[July 2017 – September 2017]

Again figure shows that the average travelling time throughout the days in the third quarter were longer than those of the second.

4.3 CORRELATION ANALYSIS

For our study, we have selected average temperature, humidity, dew point, wind speed, pressure and precipitation as the factors of weather as mentioned earlier with which tried to find out a correlation with travelling time. Except for the precipitation factor none of the other parameters had any effect on the mean travelling time in between the nodes. Figure

shown below tells us that there is almost no correlation between temperature and mean travel time. The data points are too scattered to draw a best fit line.

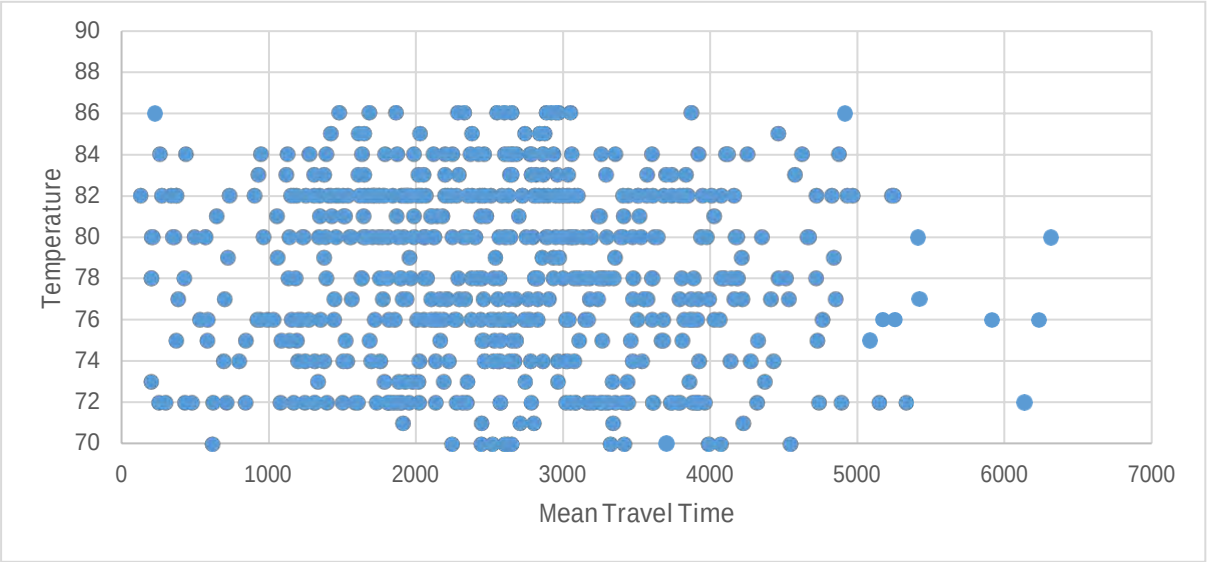

Figure 4.3.1: Scatter diagram of temperature against mean travel time.

On the other hand, the data shows, there is an increase in travel time of around 12.6% for 1-inch increase in precipitation. Although it cannot be claimed that precipitation has a direct correlation on traffic demand, traffic flow but it surely causes other mishaps like road blockage, water clog, that lead to traffic congestion. Below is a figure that shows the positive correlation between precipitation and mean travel time between two nodes of Mumbai city. The correlation coefficient found from the data collected between the two randomly selected node is 0.78.

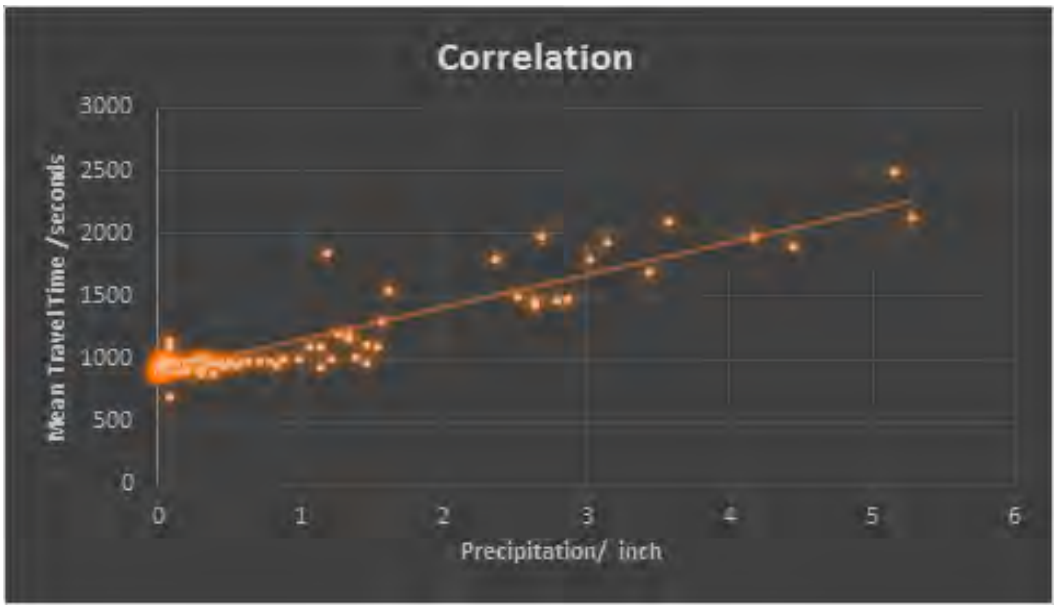

Figure 4.3.2 : Positive correlation between mean travel time and precipitation

Chapter 5

Conclusion

5.1 CONCLUSION

In this thesis with we presented using a model using machine learning algorithms to forecast the mean travelling time with an accuracy up to 85% trained on the collected data from Uber Movement. The study is categorized into three components. 1) The performance analysis of different machine learning algorithm trained on different size of training set. It is seen that the regression analysis worked best on the data that we had collected. 2) Recognizing pattern of daily commuters and analyzing data on a quarterly basis to study seasonal variation. This showed us that travelling in evening can sometimes take longer than PM peak period and certain holidays can cause a particular day to have an irregular pattern than usual 3) the impact of weather events on the travel time prediction was investigated. The third part of our research showed us that the other factors of weather does not affect travelling time as much as precipitation. Although it cannot be said that precipitation has a direct link to traffic congestion but in the case of Mumbai city. One possible cause might be that precipitation might have reduced the freeway capacity but the trip demand to use the freeway has not decreased accordingly

5.2 LIMITATIONS

The limitation to our system comes mainly from the data that we have obtained. Uber Movement data does not contain all the necessary information that could have made our predictive model more realistic, or could have outputted result more precise to urban planning. Uber Movement does not contain the following data that we think could have helped our cause better.

- The route taken by the Uber trips.
- The width and the traffic flow and density.
- The number of trips made between every nodes per day

The model learned from data obtained only from Uber Movement. Uber mainly works as a taxi aggregator in the city of Mumbai where majority of the vehicles are sedan cars. So

the prediction model is subjective specifically to sedan cars and taxi like models explicitly. Other means of transport might have different mean travelling times than the result obtained from our model

5.3 FUTURE WORK

In the future we would like to come up with a way to mine our own data as per our need instead of relying on secondary sources. We would like to include further parameters such as traffic flow, traffic density, average speed, etc. We would like to conduct a similar experiment in another city where there might be other factors of weather having an impact on the flow of traffic.

Instead of just analyzing and forecasting somewhat predictable pattern of data, we would like to look for solutions to reduce the level of traffic congestion. Geospatial and geo temporal data can be analyzed to identify busy intersections using betweeness centrality to help urban planners to come up with ways to liquidate flow of traffic further. In the past few years ITS has covered huge steps but there is more way to go in this area of research.

REFERENCES

- Chin, Kwai-Sang & Tummala, V & P. F. Leung, Jendy & Tang, Xiaoqing. (2004). A study on supply chain management practices: The Hong Kong manufacturing perspective. International Journal of Physical Distribution & Logistics Management. 34. 505-524. 10.1108/09600030410558586.
- "The money lost on our Roads" *The New Indian Express,* N.p.,20 April 2018. Web.[Online] http://www.newindianexpress.com/opinions/editorials/2018/apr/20/the-money-lost-on-our-roads-1803847.html
- The Data Team, "The Economist," 28 February 2018. [Online]. Available: https://www.economist.com/blogs/graphicdetail/2018/02/daily-chart-20.
- Salkind, N. J. (2010). *Encyclopedia of research design* Thousand Oaks, CA: SAGE Publications Ltd doi: 10.4135/9781412961288
- Wikipedia contributors. (2018, July 11). Weather Underground (weather service). In *Wikipedia, The Free Encyclopedia*. Retrieved 20:22, July 17, 2018, from https://en.wikipedia.org/w/index.php?title=Weather_Underground_(weather_service)&oldid=849770444
- Bauza, R. and Gozálvez, J., 2013. Traffic congestion detection in large-scale scenarios using vehicle-to-vehicle communications. Journal of Network and Computer Applications, 36(5), pp.1295-1307.
- Joshi, Manish & Aldhayni, TheyaznTheyazn. (2015). A Review of Network Traffic Analysis and Prediction Techniques. .
- Horvitz, Eric et al. "Prediction, Expectation, and Surprise: Methods, Designs, and Study of a Deployed Traffic Forecasting Service." *UAI*(2005).
- Treboux, Jérôme & Jara, Antonio J. & Dufour, Luc & Genoud, Dominique. (2015). A predictive data-driven model for traffic-jams forecasting in smart santader city-scale testbed. 64-68. 10.1109/WCNCW.2015.7122530.
- Michael Bolt,J. Craig Prather, Haley Harrell, Tyler Horton, John Manobianco, Mark L. Adams, "Design and Testing of Novel Airborne Atmospheric Sensor Nodes", *Geoscience and Remote Sensing Letters IEEE*, vol. 15, pp. 73-77, 2018, ISSN 1545-598X.
- J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu and C. Chen, "Data-driven intelligent transportation systems: A survey.," IEEE Trans. Intelligent Transportation Systems, vol. 12(4), pp. 1624-1639, 2011.

- N. Buch, S. A. Velastin and J. Orwell, "A review of computer vision techniqu es for the analysis of urban traffic," IEEE Trans. Intelligent Transportation Systems, v ol. 12(3), pp. 920-939, 2011.
- G. Marfia, M. Roccetti and A. Amoroso, "A new traffic congestion prediction model for advanced traveler information and management systems," Wireless Commu nications and Mobile Computing, vol. 13(3), p. 266–276, 2013.
- Alasdair Turner. From axial to road-centre lines: a new representation for space syntax and a new model of route choice for transport network analysis. Environment and Planning B: Planning and Design, 34(3):539–555, 2007.
- Song Gao, Yaoli Wang, Yong Gao, and Yu Liu. Understanding urban traffic- flow characteristics: a rethinking of betweenness centrality. Environment and Planning B: Planning and Design, 40(1):135–153, 2013.
- Yu Zheng, Yanchi Liu, Jing Yuan, and Xing Xie. Urban computing with taxicabs. In Proceedings of the 13th international conference on Ubiquitous computing, pages 89–98. ACM, 2011.
- N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, “SMOTE:synthetic minority over-sampling technique”, Journal of artificial intelligence Research, 2002, pp. 321-357.
- Texture Feature Extraction and Classification of SEM Images of Wheat Straw/Polypropylene Composites in Accelerated Aging Test - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Flow-chart-of-SVM-algorithm-on-predicting-classification_fig6_283469343 [accessed 18 Jul, 2018]
- Finding the Most Significant Elements for the Classification of Organic Orange Leaves: A Data Mining Approach - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Decision-boundary-margins-and- parameters-of-a-support-vector-machine_fig1_317928807 [accessed 18 Jul, 2018]