Application of Machine Learning Techniques on The Context of Predicting Upcoming Traffic Congestion and Providing The Best Preferred Path.

by

Muhammad Sadman Saquib
14101030
Mili Mohammad Ali
14101056
Marisha Tazmim
14101170
Faiyaaz Ahmad
18241021

A thesis submitted to the Department of Computer Science and Engineering in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
April 2019

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Muhammad Sadman Saquib 14101030	Mili Mohammad Ali 14101056
Marisha Tazmim 14101170	Faiyaaz Ahmad 18241021

Approval

The thesis/project titled ”Application of Machine Learning Techniques on The Context of Predicting Upcoming Traffic Congestion and Providing The Best Preferred Path.” submitted by

1.Muhammad Sadman Saquib (14101030)

2.Mili Mohammad Ali (14101056)

3.Marisha Tazmim (14101170)

4.Faiyaaz Ahmed (18241021)

Of Spring, 2019 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on April 25, 2019.

Examining Committee:

Supervisor:

Hossain Arif
Assistant Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:

Jia Uddin,PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:

Md. Abdul Mottalib,PhD
Professor and Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

Bangladesh being one of the most heavily populated country in the world, the traffic condition is worsening every year due to the increase of vehicle activities throughout the entire nation. Research claims that an average person in Dhaka spends up to 7 years of his/her lifetime just sitting in traffic. Without the proper knowledge of traffic situation, it is very hard to schedule our work and execute them accordingly. So, it is necessary to present a method that can predict the upcoming traffic congestion and by using that valuable information, it will help us to execute our chores in a way that will enable us to make the best and most effective use of our times. Since the growth of Intelligent Traffic System (ITS) has been upgrading in a quite effective way, it has enabled a vast areas and methods to analyze and predict traffic density. Our research proposes a way to predict the upcoming traffic density based on using different regression analysis techniques and using these prediction results, we can provide the best suited and less time consuming possible route for the user. The traffic dataset has been collected form Uber Movements for the city of Mumbai, India. We decided to choose this particular country since India is the closest and has the most similarities to our societal environment. We have processed the data and applied multiple machine learning techniques and from them we chose the best methods with the most optimum accuracy of over 75 percent. In this study, we attempted to find some better traffic prediction results by implementing Lin- ear Regression model, Logistic Regression model. We have initiated our works in an android based application which shows us the best possible route based on the upcoming traffic congestion according to different dates and routes. Hopefully, our research will be able to contribute to finding more enhanced predictions of day to day traffic congestion and enabling our users to plan their schedule ahead of time using these predictions results.

Keywords: Traffic congestion, congestion prediction, machine learning, Intelligent Transport System (ITS), regression analysis, correlation, Linear regression, Logistic regression

Table of Contents

List of figures

List of tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used withinthe body of the document

ε The permitivity of free space

θ theta is used to represent an angle

ALU Arithmetic Logic Unit

CD Contact Dynamics

FPU Floating Point Unit

GPU Graphics Processing Unit

ITS Intelligent Transport System

MAE Mean Absolute Error

MPM Material Point Method

MSE Mean Squared Error

PIC Particle-in-cell

RMSE Root Mean Square Error

RVE Representative Elemental Volume

SM Streaming Multiprocessor

VANET Vehicular Ad Hoc Network

Chapter 1

Introduction

1.1 Introduction

With the increase of urbanization and population in the cities around the world, traffic congestion has now become one of the biggest problem of this era. These congestions can happen due to different factors which are very common throughout the whole world. They are peak hour traffic, road construction, maintenance and accidents. Congestions can be handled in a few different ways. First of all, more constructions of roads, underpass, flyovers can be undertaken by the government which are very expensive and time consuming and for a certain time it can also increase the traffic congestion even more. But there are other smarter ways to deal with this problem using different technological strategies and making the most efficient use of the available routes in the city. One of these strategies is to make predictions of the upcoming congestion and suggest the best possible route which will take the lesser amount of time for a user to reach his/her destination. In our work, we have taken traffic congestion data from different points of the city Mumbai and applied multiple machine learning algorithms which has successfully provided us with an optimum accuracy.

1.2 Motivation

Traffic jam has become a terrible problem around the whole world especially in the urban areas which is leading to millions of wasted work hours and also a huge amount wasted gasolines every day. The rapid increase of population in urban areas, increase of the number of personal vehicle on the road with addition to more and more unskilled drivers are leading to traffic congestions[10]. Dhaka is the capital of Bangladesh and also one of the most densely populated metropolitan in the world consisting of a population of almost 20 million

people. According to the RSTP, the city dwellers make around 30 million trips every day. Of them, some 47 percent involve buses, 9 percent cars while 32 percent are made in rickshaws. Despite transportation in Dhaka being mostly road based, there are only 400km of roads in Dhaka and among them around 40 percent of the footpaths are occupied by street vendors, shops, garbage bins and construction materials. According to a government report, traffic congestion in the capital eats up around Tk 20,000 crore a year and some 3.2 million business hours are also lost to this problem every day. 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[3]. 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. These studies prove how terribly traffics congestion affect our lives in a negative way which has led us towards our work to handle this problem in a smarter and more efficient way

Chapter 2

Literature Review

1.3 Related Works

In this chapter, we've given a brief overview of different researches conducted in the past which are related to our topic. Some of these researches has been very effective and helpful on the context of predicting traffic throughout the world. An article about Ubiquitous Computing and Communications talked about an algorithm to predict traffic based on data analysis. Basically, it's a short prediction method to predict the traffic flow which is called SRBDP. Moreover, they designed the data characteristics and used clustering methods and a small amount of human intervention to determine the historical data congestion situation [4]. Another article suggests a new dynamic algorithm for Intelligent Transport System (ITS) which proposes a simple solution for all-to-one all departure time intervals shortest path problems with an optimal run time complexity [7]. Moreover, predicting traffic congestion can be difficult since the traffic condition in a given road can vary due to spatial and time domain of different roads and also due to vehicle speed changes. Different researches were conducted regarding this issue. One of the relevant work was done by Bauza R. et.al who proposed a traffic congestion detection system which is based on the communication of one vehicle to another and acquired a probability of 90 Percent [[2]]. Also, Eric Horvitz et.al, did a research based on deployed traffic forecasting service which has led to the deployment of a service called JamBayes which is being actively used by over 2500 users via smartphones and desktop computers [[5]]. The purpose of this paper is to propose a smart traffic management system using the Internet of Things and a decentralized approach to optimize traffic on the roads and intelligent algorithms to manage all traffic situations more accurately [8]. In this paper they proposed system is overcoming the flaws of previous traffic management systems. The system takes traffic density as input from cameras which is abstracted from Digital Image Processing technique and sensors data, giving output as signals management.

1.4 Research related to Traffic Prediction

Despite all these studies, a big number of authors suggested that the configuration of the street network of a particular city plays an important role in vehicular flow. For that reason, authors such as Turner suggested the usage of centrality measures on the basis of street graph in order to predict traffic flow [10]. Although Gao, et al. 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 [3]. We have introduced about one of concept of transit nodes, as a means for prepossessing a road network, with given coordinates for each node and a travel time for each edge, such that point-to-point shortest-path queries can be answered extremely fast. [1]. We have found out that in this paper [6] it discussed about the comparison of the efficiency of two algorithms, by estimation of their complexity. The problem discussed is how to find the shortest path in an environment with n states. Lastly authors in this paper used Krushkal's algorithm to find the shortest path. Finding fastest path is sort of routing system that provides instructions to users based upon "optimum" route solutions [9]. Research team propose a distributed, collaborative traffic congestion detection and dissemination system that uses VANET [11]. Each of the driver's smart phones is equipped with a Traffic App which is capable of location detection through Geographic Position based System (GPS). This information is relayed to a remote server which detects traffic congestion.

Chapter 3

Methodology

There are many researches done regarding traffic congestion and predictions in the past.A big number of these researches has included Logistic and Linear regression analysis and has been able to find optimum result for their predictions. Our research will take both testing and training data sets and produce upcoming prediction results using different machine learning algorithms such as Linear regression model, Logistic regression model. These are supervised machine learning techniques which takes various previous data set, trains itself to learn the correlation between different multiple features in the data set and use them to predict future results for the system. The outputs will be displayed in both graphical and tabular forms, where the tabular data will have the numerical properties explained, that the model could derive from the given data set. Alongside that, the system will have graphical representation, which will show the complete trend and provide an analysis for the hypothesis among the dependent and independent variables of our data set. The regression models will generate regression equation based on the analysis of the complete data set. From this equation, the end user will be able to derive the predicted outcomes of the dependent variable (expected number of seconds required to reach the destination), by providing the independent variables (Quarter of the day, date, higher bound, lower bound and mean travel time) as input variables. We have designed an application which will help the user by showing them the best possible routes via Google maps as a visual representation. This system will also enable them to choose the dates of the upcoming particular day they want to travel and also enable them to choose from multiple origin and destination points. After selecting the time and place, the system will show them the route that contains the least congestion and least number of times to travel based on our predictions results. Using these methods, our system will be able to provide an optimum prediction of upcoming traffic congestion throughout the year.

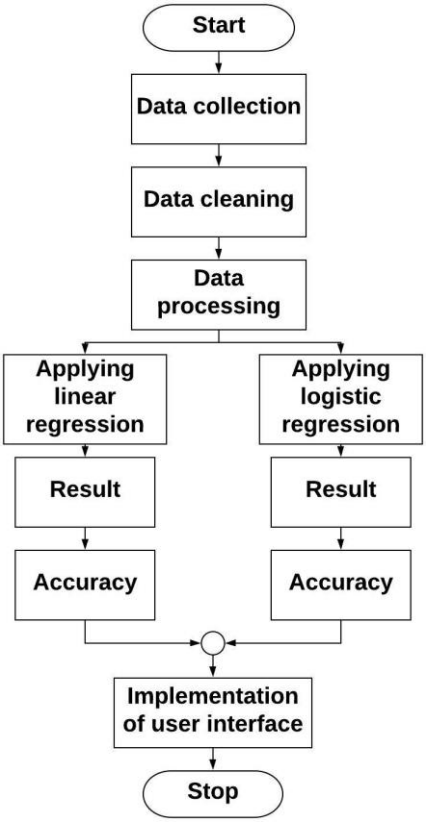

Fig. 3.1 Block Diagram of the work-flow in the system.

Our workflow started with the data collection. Then we had to clean the dataset as there was many empty entries in the dataset. Later we had to process the dataset according to our need and applied linear and logistic regression on the dataset. After getting the results of thetwo regression algorithms we tested the accuracy and implemented the algorithm with the higher accuracy into an user interface.

3.1 Understanding the data set

We have used secondary data for our research purposes. This data has been collected by Uber and has been later made open source in order to provide the researchers for their future traffic research works to make our cities more efficient. We have used the traffic data of Uber through their Uber Movement project. Our traffic data is based on city of Mumbai and the data are from 1st of January to 30th March 2018.

3.1.1 Traffic Data

Recently Uber has introduced an open source segment called “Uber Movement” which is basically a tool that enables the researchers to help improve urban mobility. This segment has trip data of over two billion anonymous trips of different cities like Bogota, Boston, Johannesburg, Manila, Paris, Sydney, Washington D.C., Mumbai and are adding more cities day by day. These data are fully accessible and open source for researchers around the whole world to analyze road conditions and also to work more efficiently on traffic systems. Uber movement data sets 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. Also, the data is characterized into different attributes which divides the data into different time ranges like whole daily travel time, 12 hour ranges like AM and PM and also into different quarters of the day like midday, early morning and evening time. All these data are collected and converted into .CSV format and is made available in their website. The Uber Partner app records the current latitude and a timestamp for every 4 second of their travel. 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. These GPS ping points provides the timestamps which can also be used to calculate a complete average travel time from the origin point to the destination points in a specific area. There are certain steps that Uber Movement uses to measure the travel time data using these GPS ping points. They are given below:

1. Certain GPS ping are assigned in different designated zones in every point of the Uber networks.

2. Whenever an Uber vehicle passes through the GPS ping points within a zone the travel time from one GPS point to another is calculated.

3. After the trip the has been completed, the mean of the elapsed time from each GPS point is calculated together and provides a complete mean travel time for one full trip.

4. These zone to zone travel time is measured from trips all around the network. After calculating the route travel time, the statistical time measurement are only kept as a record and other trip information is lost.

5. All these zone to zone travel time are later made available in .CSV formatted datasheets to be accessed by the users.

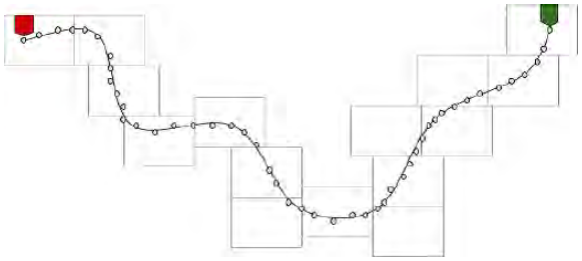

Fig. 3.2 The zones passed by an Uber trip between the colored starting and end points.

3.1.2 Android Interface

We have designed an application which will help the user by showing them the best possible routes via google maps as a visual representation. The interface will consist of a date box and a n option to choose their travelling routes where the user can input the date and route of the travel and the also enable the user to choose from multiple origin and destination points. After selecting the time and place, the system will show them the route that contains the least congestion and least number of times to travel based on our predictions results. There are different possible routes to travel from an origin point to a destination point. The system will basically search through our predicted results according to the route and date of the month the user has chosen. According to the selected date and route, the system will choose the possible route that takes the least amount of time to travel to the destination point and show that specific route in the user interface.

Fig. 3.3 Android interface with the options of choosing routes.

Here, the interface has the option of choosing the route of the travel the users wants to. After choosing the route the user will have to choose the date of the traveling time.

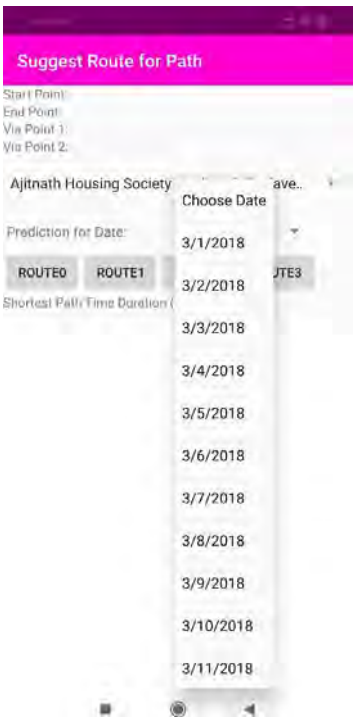

Fig. 3.4 Android interface with the options of choosing dates.

In this picture, After choosing the date and route both, the system will be able to start its prediction result.

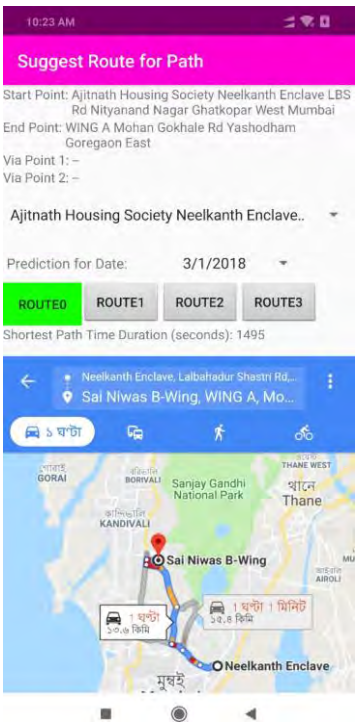

Fig. 3.5 Prediction of possible route with different dates including via points.

In this picture, the green button refers to the preferred possible path that will take the least amount of time to travel. Underneath the date the starting and ending point both are defined along with the option of via points if there is any. After tapping the green button, the best preferred path will be shown via google maps. Just above the map the estimated travel time is shown in seconds.

In table 3.1 provides a sample of the data-set collected from the Uber Movements where the daily mean travel time taken and the upper bound and lower bound for each certain days from one certain origin point to the destination point of Mumbai city. The type of data is available for three months of every day for the same origin and destination points. The data is also characterized into another different segment on the basis of AM and PM in table 3.2. The travel time of the whole day has been divided here into 12 hours for more accurate study of the traffic movements.

Table 3.1 Mean Travel Time Days of the month.

Date	Origin Move-ment ID	Origin Dis-play Name	Destination Move-ment ID	Destination Display Name	Daily Mean Travel Time	Daily Range Lower Bound Travel TIme	Daily Range Up-per Bound Travel TIme
01/01/2018	670	Ajitnath Soci-ety,	434	0/506, Western Express Hwy,	2617	1738	3940
01/02/2018	670	Ajitnath Soci-ety,	434	0/506, Western Express Hwy,	2334	1662	3279
01/03/2018	670	Ajitnath Soci-ety,	434	0/506, Western Express Hwy,	3910	2695	5674
01/04/2018	670	Ajitnath Soci-ety,	434	0/506, Western Express Hwy,	3808	2546	5967

Table 3.2 Mean Travel Time in a AM and PM Time Of The Specific Day.

Date	Origin Movement ID	Origin Display Name	Destination Movement ID	Destination Display Name	AM Travel Time(Seconds)	PM Travel Time(Seconds)
01/01/2018	670	Ajitnath Society,	434	0/506, Western Express Hwy,	2617	1738
01/02/2018	670	Ajitnath Society,	434	0/506, Western Express Hwy,	2334	1662
01/03/2018	670	Ajitnath Society,	434	0/506, Western Express Hwy,	3910	2695
01/04/2018	670	Ajitnath Society,	434	0/506, Western Express Hwy,	3808	2546

Chapter 4

Algorithm

The following discussion focuses on the techniques used to obtain expected output prediction in this study

4.1 Linear Regression

The hypotheses of the multivalued regression analysis are

$$h_0(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots . \theta_n x_n \quad (4.1)$$

Linear regression plots a straight linear line in a diagram of scattered recorded points. The equation for the straight line or the best fit line is where one or more independent variables can be used to calculate the value of a dependent variable. 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. The formula is given below:

$$Cost function(J_\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \quad (4.2)$$

To prevent the increase of R- squared value in a regression analysis, the unimportant variables are removed from the analysis with the development of the regression model. Therefore, the removal process is important in order to provide optimal and accurate results of the analysis.

4.2 Logistic Regression

Logistic regression looks through the entire datasets to discover the hyper plane which fits the most for recognizing the classes. The center of calculated relapse is "strategic capacity". Calculated capacity is additionally called the sigmoid capacity. This function was for the most part created for depicting the properties of populace development in biology, rising rapidly and maximizing at the conveying limit of nature. It is a 'S' molded bend which can take genuine esteemed number and guide it into an incentive somewhere in the range of 0 and 1. The function given below:

$$\varphi_{sig}(z) = \frac{1}{1 + exp(-z)} \tag{4.3}$$

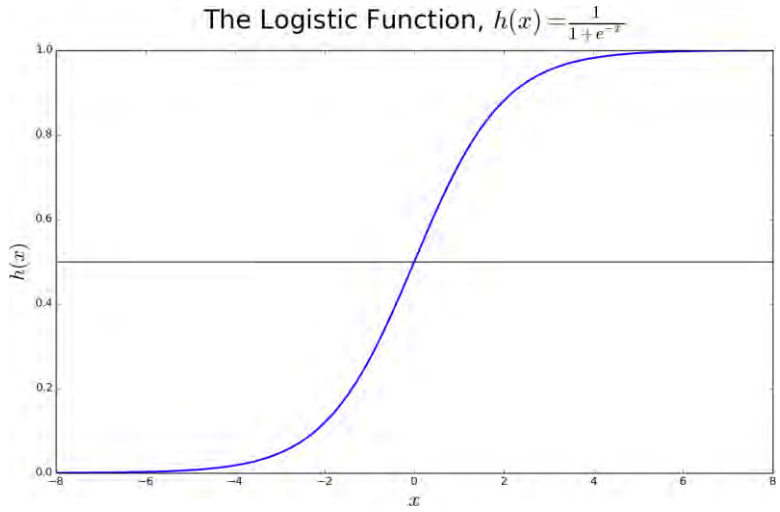

Fig. 4.1 Logistic Function Curve

Logistic regression uses an equation as the representation, very much like linear regression. Input values (x) are combined linearly using weights or coefficient values to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary values (0 or 1) rather than a numeric value [11]. Below is an example logistic regression equation:

$$e\Lambda(b0 + b1 * x)/(1 + e\Lambda(b0 + b1 * x)) \tag{4.4}$$

Where y is the anticipated yield, b0 is the predisposition or catch term and b1 is the

coefficient for the single information esteem (x). Every segment in your information has a related b coefficient (a steady genuine esteem) that must be gained from your preparation information. Logistic regression is a direct technique, however the expectations are changed utilizing the strategic capacity. The effect of this is the expectations can never again be comprehended as a straight mix of the contributions as we can with direct relapse, for instance, proceeding from over, the model can be expressed as:

$$p(X) = e\wedge(b0 + b1 * X)/(1 + e\wedge(b) + b1 * X)) \tag{4.5}$$

$$ln(p(X)/1 - p(X)) = b0 + b1 * X \tag{4.6}$$

This is valuable since it very well may be seen that the estimation of the yield on the privilege is straight once more (simply like linear regression), and the contribution on the left is a log of the likelihood of the default class.

This proportion on the left is known as the chances of the default class. Chances are determined as a proportion of the likelihood of the occasion isolated by the likelihood of not the occasion, for example 0.8/(1-0.8) which has the odds of 4. So all things considered it very well may be composed as:

$$ln(odds) = b0 + b1 * X \tag{4.7}$$

Since the odds of log are changed, this is called left hand side the log-odds or the probit. It is conceivable to utilize different sorts of capacities for the change, which is out of scope, yet thusly usually to allude to the change that relates the straight relapse condition to the probabilities as the connection work, for example the probit connect work. The type can be moved back to one side and compose it as:

$$odds = e\wedge(b0 + b1 * X) \tag{4.8}$$

Chapter 5

Implement and Result Analysis

In implementation for predicting output we used logistic and liner regression analysis over the dataset. A comparison is made between the logistic and liner regression analysis for predicting the outcomes, is generated. Therefore, the data sets need to be authentic and complete.

5.1 Data Processing

The process of collecting the dataset for the research and the standardization form of the raw data is given below.

5.1.1 Data Collection

Firstly, for prediction of upcoming traffic congestion the data online data sets for Bangladesh is not available. On searching multiple government websites, we could not find any digital data set and later we decided to go for the closest possible country's data which is India and we went with Mumbai city's data was collected from Uber movement. We have selected the data set consists of all possible nodes of Mumbai city from Uber movement websites. The data set is in the format of .csv file. From Uber movement we could download data of 3 months at a time for any route. To analyze the result more efficiently, we choose two destination points for two source points and to reach each destination we took 3 to 4 routes as options. For instance, from Ajitnath to Mohan Gokhale road we considered a direct route and 3 more routes with different via points. The table showing the average relationship between the diet plan and current weight is given below. The data has been split into two parts; one is training set and the other is testing set. The ratio was measured into 7:3 to see

the performance of each algorithm upon different values of training and testing data. The work flow we have followed is given below.

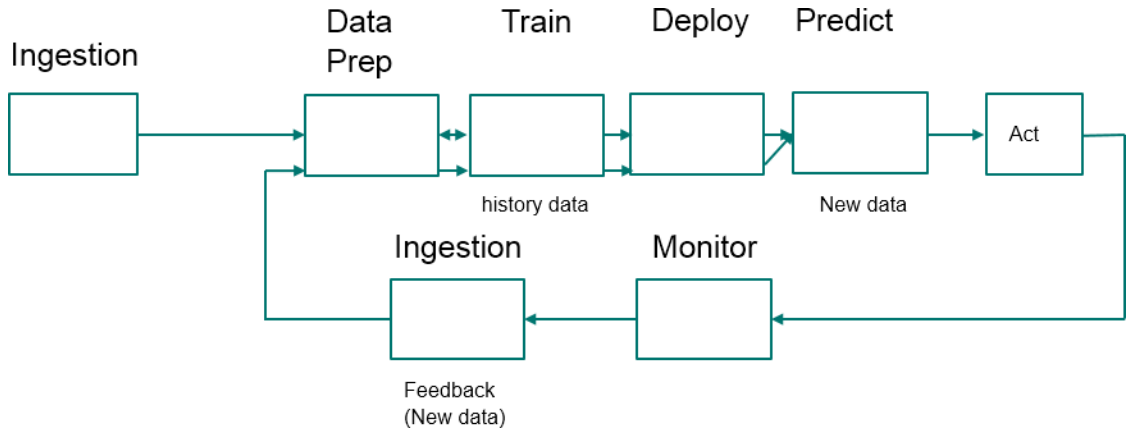

Fig. 5.1 Work flow of Implementation

A sample dataset of a direct route between source and destination is given below.

	A	B	C	D	E	F
1	Date	Origin Display Name	Destination Display Name	Daily Mean Travel Time (Seconds)	Daily Range - Lower Bound Travel Time (Seconds)	Daily Range - Upper Bound Travel Time (Seconds)
2	12/31/2017	Ajitnath	WING A, Mohan Gokhale Rd,	1021	675	1543
3	1/1/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1382	962	1984
4	1/2/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1253	789	1991
5	1/3/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1519	927	2490
6	1/4/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1468	921	2341
7	1/5/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1284	861	1915
8	1/6/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1080	752	1552
9	1/7/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1398	961	2032
10	1/8/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1332	926	1918
11	1/9/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1497	970	2310
12	1/10/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1519	1008	2289
13	1/11/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1613	997	2609
14	1/12/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1631	1078	2468
15	1/13/2018	Ajitnath	WING A, Mohan Gokhale Rd,	979	690	1389
16	1/14/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1322	842	2076
17	1/15/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1584	1089	2306
18	1/16/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1596	1115	2285
19	1/17/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1637	1057	2533
20	1/18/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1638	1086	2469
21	1/19/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1469	1009	2138
22	1/20/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1209	823	1777
23	1/21/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1495	950	2352
24	1/22/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1389	952	2025
25	1/23/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1541	1094	2170
26	1/24/2018	Ajitnath	WING A, Mohan Gokhale Rd,	1471	965	2241
27	1/25/2018	Ajitnath	WING A, Mohan Gokhale Rd,	946	695	1287

Fig. 5.2 Dataset of a direct path(Ajitnath to Mohan Gokhale Road)

Here in this table 5.2 a dataset of a direct path between Ajithnath and Mohan Gokhale Road is shown where in column A is the date of the travel, in column B is the source name, in column C is the destination name, in column D is the mean travel time, in column E is the highest travel time and in the last column is the lowest travel time from the source to destination.

5.1.2 Data Prepossessing

As we have used secondary data we needed to manually customize the data set according to our requirements. Data prepossessing needed the following steps:

1. Data cleaning : There was multiple null values in the data set which we needed to remove them in order to run the algorithms without any difficulties.

2. Data scaling : To get better accuracy of the trend analysis, the standardized data are ordered based on date of a particular month.

3. Feature Selection : As we could collect data from Uber movement and has limited attributes, all the input parameters were correlated to the output parameter. Therefore, we used linear regression and logistic algorithm on our data set.

4. Arranging data set : The data set we collected was made with only two nodes which was the source and the destination. For the via routes as we used multiple nodes we had to manually add the via points and arrange it in such a way so that we can run the algorithm. A figure is given below of data arrangement.

Table 5.1 Data set of a via path(Ajitnath to Mohan Gokhale Road using different node points)

Date	Origin Dis-play Name	Destination Display Name	Total time-Ajithnath to Mohan gokhale rd- (c+d)	(Aarey rd- m.g rd) travel time-a+b = c	(Ajithnath - Aarey rd)Daily Mean Travel Time(Second	(Aarey rd-pimpri-pada)Travel Time (Seconds)-dsa)-	(Pimpripad-a mohan gokhale rd)Travel Time (Seconds)- b
12/31/2017	Ajitnath	WING A, Mohan Gokhale Rd	4433	2007	2426	1692	315
01-01-18	Ajitnath	WING A, Mohan Gokhale Rd	4575	1790	2785	1474	316
01-02-18	Ajitnath	WING A, Mohan Gokhale Rd	4637	1754	2883	1545	209
01-03-18	Ajitnath	WING A, Mohan Gokhale Rd	5412	2277	3135	1920	357
01-04-18	Ajitnath	WING A, Mohan Gokhale Rd	4828	2228	2600	1892	336

In this table 5.1 where a via path between Ajithnath and Mohan Gokhale Road is shown where in 1st column is the date of the travel, in 2nd column is the source name, in 3rd column is the destination name, in the 4th column the mean travel time is given. Moreover, 5th column is the addition of the last two columns which is the distance from Ajithnath to Pimpripada and from Pimpripada to Gokhale road. And in the 6th column the distance between Ajithnath to Aarey road is given.

5.2 Result Analysis

5.2.1 Result analysis based on Logistic Regression Model

For the training and testing the predictor value was kept in y variable and the value that the algorithm will use for prediction is in the x variable.

$$x = trf - data.iloc[:, [4, 5]].values \tag{5.1}$$

$$y = trf - data.iloc[:, 3].values \tag{5.2}$$

trf-data is were the data set is saved as a variable and between the third brackets it is the column numbers of the data. The training and test dataset is divided in 70/30 percentage and the training set is randomly picked.

$$X - train, X - test, y - train, y - test = train - test - split(X, Y, test - size) = 0.25, random - state = 0) \tag{5.3}$$

Then the prediction values are returned as an array corresponding to the test set's array.

	A	B	C	D
1	ajitnath----mohan gokhale rd			
2	direct route – lower bound travel time	direct route – upper bound travel time	direct route -mean travel time(prediction)	actual value
3	789	1991	1495	1253
4	690	1989	1338	1278
5	913	1663	1368	1232
6	1071	2237	1056	1548
7	949	1927	1368	1352
8	859	2009	1368	1314
9	820	1946	1613	1263
10	958	2489	1613	1544
11	902	1696	1519	1237
12	961	2032	1371	1398
13	844	1557	1368	1147
14	925	2193	1371	1425
15	1002	2201	1613	1485
16	926	1918	1368	1332
17	1115	2285	1368	1596
18	965	2241	1613	1471
19	1003	2471	1519	1575
20	734	2294	1365	1297
21	928	2356	1519	1479
22	952	2025	1365	1389
23	752	1552	1519	1076
24	764	1632	1368	1179

Fig. 5.3 Result of direct route - Ajitnath to Mohan Gokhale road

Here in the fig 5.3 is a direct path between Ajithnath and Mohan Gokhale Road is shown where in column A is the highest travel time from the source to destination, in column B is the lowest travel time, in column C is the estimated travel time (our prediction) and in the last column is the actual mean travel time(from the dataset)

ajitnath----mohan gokhale rd		
Via 2	prediction	actual value
[690, 749],	1713	1439
[665, 838],	2111	1503
[845, 999],	1653	1844
[765, 748],	1711	765
[805, 933],	1653	1738
[830, 765],	1931	1595
[830, 846],	1711	1676
[786, 780],	1211	1566
[777, 1050],	2239	1827
[761, 936],	1653	1697
[847, 851],	1711	1698
[810, 846],	1711	1656
[982, 741],	1862	1723
[842, 1012],	1653	1854
[807, 989],	1653	1796
[836, 1036],	2071	1836
[798, 1018],	1711	1816
[939, 976],	1653	1915
[739, 884],	1653	1623
[813, 973],	1653	1786
[815, 980],	1653	1795
[681, 850],	1711	1531
[771, 761]]	1931	1532

Fig. 5.4 Result of via route 2 - Ajitnath to Mohan Gokhale road via Rani Sati Nagar

Here in the fig 5.4 Result of via route 2 - Ajitnath to Mohan Gokhale road via Rani Sati Nagar a via path between Ajithnath and Mohan Gokhale Road is shown where in 1st column is the travel time from Ajithnath to Rani Sati Nagar and from Rani Sati Nagar to Mohan Gokhale Road(in the array), in 2nd column is the estimated travel time (our prediction) from the source to destination and in the last column is the actual mean travel time(from the dataset).

ajitnath----mohan gokhale rd Via 3	prediction	actual value
[2334, 431],	4640	2765
[2206, 456],	2933	2662
[3135, 650],	4670	3785
[2419, 514],	2933	2933
[3425, 749],	4670	4174
[2799, 616],	3575	3420
[2450, 635],	3575	3085
[2615, 628],	4640	3243
[2894, 546],	2933	3440
[2312, 502],	4670	2814
[3095, 677],	4670	3772
[2426, 510],	3575	2936
[2432, 575],	3575	3007
[2857, 719],	3575	3576
[3314, 785],	3575	4099
[2938, 824],	3575	3575
[3014, 782],	3575	3796
[2503, 668],	2905	3171
[2512, 434],	3575	2946
[2886, 743],	3575	3629
[2735, 738],	2933	3473
[2409, 539],	2933	2948
[2227, 458]	3575	2685

Fig. 5.5 : Result of via route 2 - Ajitnath to Mohan Gokhale road via West Express and Vit Bhatti

Here in figure-5.5 a via path between Ajithnath and Mohan Gokhale Road is shown where in 1st column is the travel time from Ajithnath to Rani Sati Nagar and from Rani Sati Nagar to Mohan Gokhale Road(in the array), in 2nd column is the estimated travel time (our prediction) from the source to destination and in the last column is the actual mean travel time(from the dataset).

5.2.2 Result analysis based on Linear Regression Model

Throughout linear algorithm we are predicting the mean travel time between nodes. We have used Mean Absolute Error (MAE), Mean Square Error (RMSE) Root Mean Square Error (RMSE) to evaluate the performance of the algorithms as output is a numerical value MSE is more popular than MAE because MSE publishes larger errors. But, RMSE is even more popular than MSE because RMSE is interpret able in the "y" units. We have calculated the mean value in order to find standard deviation .The formula is shown below:

Mean Absolute Error (MAE) is the mean of the absolute value of the errors:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \text{..... 9}$$

Mean Squared Error (MSE) is the mean of the squared errors:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \text{..... 10}$$

Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \text{..... 11}$$

Fig. 5.6 : Squared Error Equation

$$\sigma = \sqrt{\frac{\Sigma\ (x - \bar{x})^2}{n}} \quad \text{..... 12}$$

σ = standard deviation

Σ = sum of

x = each value in the data set

$\bar{x}$ = mean of all values in the data set

n = number of value in the data set

Fig. 5.7 Standard Deviation Equation

Where, difference between original and predicted value and N is the sample size. The table below shows Descriptive Statistics for selected road on different times of day.

Table 5.2 Descriptive Statistics for arc of road type 4 on different times of day

Time of Day	**Mean**	**St. Dev**
1	73.27044	3.893379
2	74.71077	4.684149
3	74.60109	4.813776
4	71.86014	5.696717
5	72.37489	5.334079

Percentage accuracy was measured according to the formula below:

$$percentageaccuracy = (1 - error) * 100\% \qquad (5.4)$$

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

Fig. 5.8 Percentage accuracy for different algorithms

Table 5.3 represents the Dynamic routing table, which has the vertex, travel mean time between the nodes Table 5.4 gives the accuracy of linear regression algorithm. Figure 5.9 represents the junctions as the graphical nodes with travel time.

Table 5.3 Mean Travel Time Days of the week.

Date	Origin Move-ment ID	Origin Dis-play Name	Destination Move-ment ID	Destination Display Name	Daily Mean Travel Time	Lower Bound Travel Time	Upper Bound Travel Time
01/01/2018	489	Ajitnath,	146	Mohan Gokhale Road	1029	675	1543
01/02/2018	489	Ajitnath,	146	Mohan Gokhale Road	1382	962	1984
01/03/2018	489	Ajitnath,	146	Mohan Gokhale Road	1253	789	1991
01/04/2018	489	Ajitnath,	146	Mohan Gokhale Road	1519	927	2490
01/05/2018	489	Ajitnath,	146	Mohan Gokhale Road	1468	921	2341

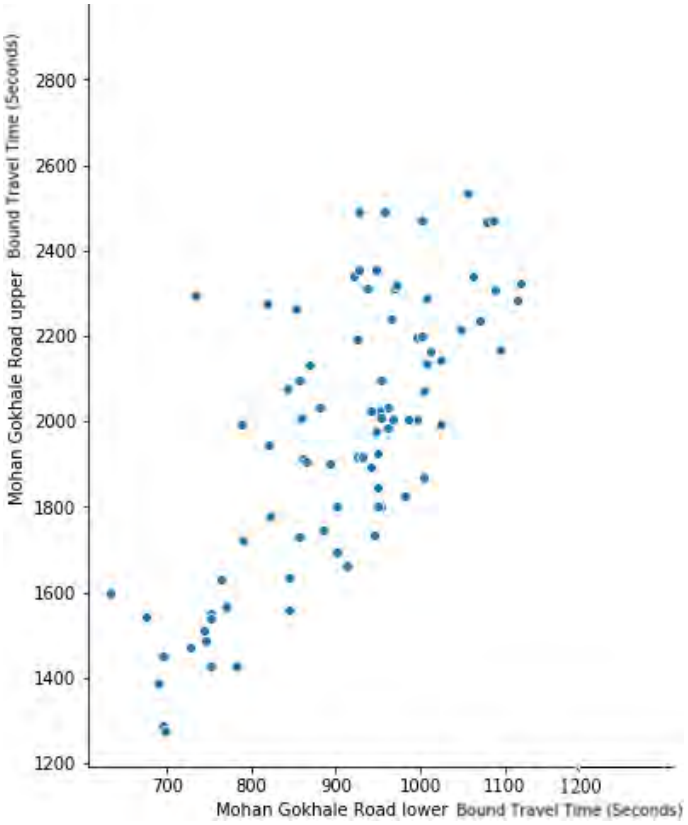

Fig. 5.9 The junctions as the graphical nodes with travel time

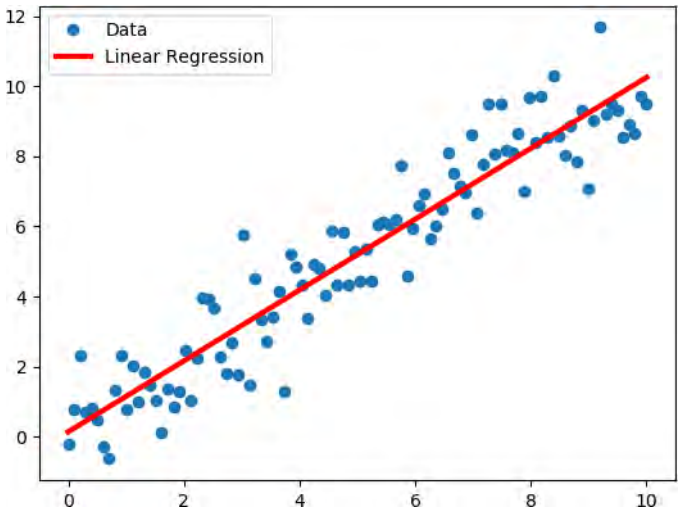

Fig. 5.10 Input dataset using Linear Algorithm

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 longer than traveling in early morning. The pattern is the same for all of the routes between the source nodes to the other nodes

From Figure 5.11 we can conclude that time of day effect particularly the travel time on the particular segment of the interstate. The Data around 2500 seconds from midday (between 1 and 4 pm) indicates that this is a route that gets comparatively low congested in the midday, and this leads to the problem of tting a theoretical distribution to thesepatterns.

Table 5.4 Mean Travel Time Days of the week.

Date Range(M/D/Y)	Time of Day	Origin Movement ID	Origin Display Name	Destination Movement ID	Destination Display Name	Daily Mean Travel Time	Lower Bound Travel Time	Upper Bound Travel Time
01/01/2018-1/31/2018	Daily Average	489	Ajitnath,	146	Mohan Gokhale Road			
01/01/2018-1/31/2018	AM Range	489	Ajitnath,	146	Mohan Gokhale Road			
01/01/2018-1/31/2018	Mid Day	489	Ajitnath,	146	Mohan Gokhale Road	1253	675	1543
01/04/2018-1/31/2018	PM Range	489	Ajitnath,	146	Mohan Gokhale Road			
01/05/2018-1/31/2018	Evening	489	Ajitnath,	146	Mohan Gokhale Road	1468	927	2341
01/05/2018-1/31/2018	Early Morning	489	Ajitnath,	146	Mohan Gokhale Road	1519	921	2490

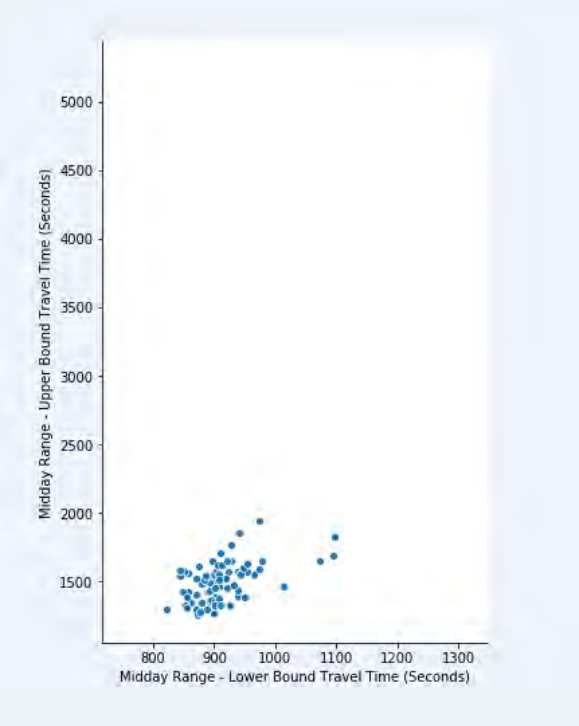

Fig. 5.11 Midday Range Travel Time

Time of day is another major factor that explains travel time on a given road segment. The rush-hour period illustrates clearly travel time variation during the day. By analyzing the results presented in Figure 5.12 and figure 5.13, we can realize that on a given day, there is usually a trafc congestion at selected Road in the evening which is mostly caused by high volume of trafc during that period. Midday (1pm - 4pm) is usually congestion free, which causes our method to display general keywords from the posts of that period. On the other hand in the morning selected area suffers most of its trafc problems due to high volume of ofce and school goers. Clearly, this is a direct consequence of volume capacity on the roads.

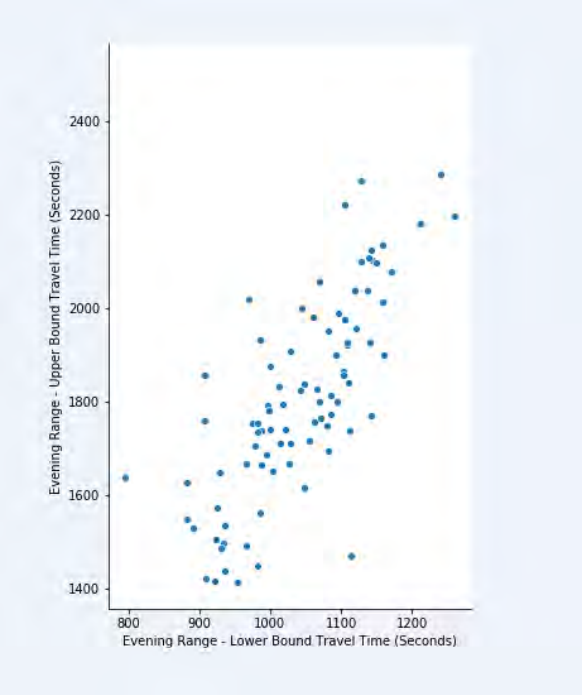

Fig. 5.12 Evening Range Travel Time

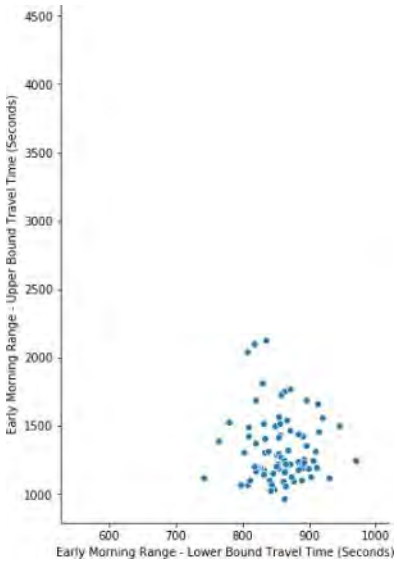

Fig. 5.13 Early Morning Range Travel Time

Mostly, a positive accuracy is interpreted as having no traffic congestion, whereas the accuracy result we have achieved is approximately is considered to indicate moderate traffic congestion. This scale can be perfected over time by analyzing the results. Further, we also have observed the causes of a traffic.

Table 5.5 Source and destination nodes with daily mean travel time and using the alter nodes (via)

Date	**Origin to Destination display name with Movement ID**	**Daily Mean Travel Time (Seconds**
1/1/2018-31/1/2018	Ajinath to Aeray road	2750
1/1/2018-31/1/2018	Aeray road to Pimpripada road	2250
1/1/2018-31/1/2018	Pimpripada road to Mohan gokhale road	2553

One of the main issue is addressed when manipulating the data is Universal Time in order to categorize it in its correct interval Data must be converted to local time so that it could help to compare travel times between links. To do this, the distance between two nodes is needed. However, since nodes are not on every link on the network, we would need to determine the exact path that connects the source node and destination node. Since there might be more

than one path, we have also considered that nodes between them. This has yield normalized time that is lower bound travel time and upper bound travel time, since the path chosen is the shortest path possible.

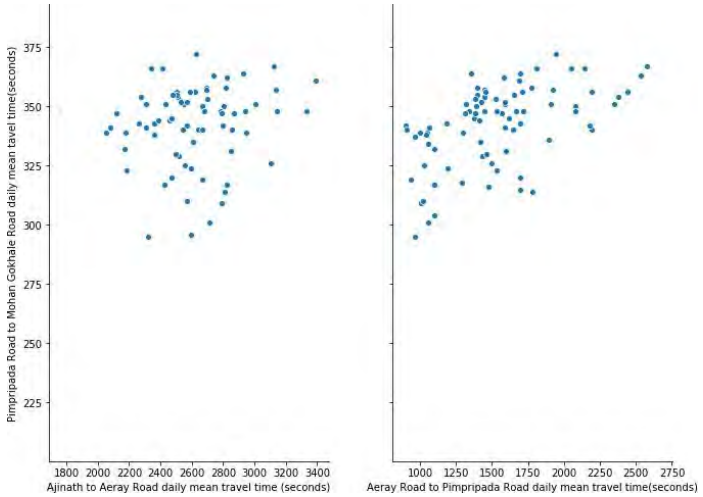

Fig. 5.14 Source and destination nodes with travel time and using the alter nodes (via)

However while searching on the same location we have found the interesting results of a historic traffic congestion. giving information that high volume of traffic), owing to the proximity of the football stadium arranged football event in that region) and cinema event are the major cause of traffic in that area. Similarly hospital area, amusement park and illegal parking also case of the traffic in that area. Instead of just analyzing and predicating pattern of data, we would like to look for solutions to reduce the level of traffic congestion.

Table 5.6 Accuracy test result

Algorithm	**Accuracy (Percentage)**
Linear regression algorithm	74.8%
Logistic linear algorithm	75.2%

Different researches were conducted regarding traffic congestion issue. One of the relevant work was done by Bauza R. et.al who proposed a traffic congestion detection system which is based on the communication of one vehicle to another and acquired a probability of 90% [3]. In our paper we have achieved approximately 89% probability.[3] The purpose of this paper is to propose a smart traffic management system to optimize traffic on the roads and intelligent algorithms to manage all traffic situations more accurately. In our paper we

have considered source node and destination node for primary and also considered connected node to identify the preferred path using uber movement website of Mubai city.[16]in this paper author used Krushkal's algorithm to find the shortest path. Finding fastest path is sort of routing system that provides instructions to users based upon "optimum" route solutions [12]. In our paper we have applied two machine learning algorithm. Linear regression and logistic regression.

Chapter 6

Conclusion and Future Works

This chapter draws a conclusion on the work that we have done till date and gives a picture of the future possibilities of our contribution in other sectors as well. It describes the functionalities, which can be additional functions of our operational application.

6.1 Conclusion

In this thesis we tried to show best possible route in a given source to destination points using the data collected from Uber movement for the city of Mumbai. In finding out the desired result in our research, we have achieved 75% accuracy using different type's machine learning algorithm on different size of data set which consists of different travel mean time in a different time of the day. Recognizing pattern of daily commuters and analyzing data we have learn traveling in evening time sometimes take more time than PM peak hours and beside in some occasional holiday time it gives unusual congestion pattern in routes which can be cause of sudden congestion of pedestrian or blockade in the road. Our thesis team have developed an android based map application to show our research output. In our application it takes input from the user as source, destination and date. Application process the input and shows the best possible route accordingly to the machine leaning algorithm and data set that we have used in our process. At the end we hope that our historical data analysis will be able to help commuters plan their future visit to their desired destination accordingly which is not available in real time analysis works which is done before us and this is the contribution we are adding for the people.

6.2 Future Works

In the future, we are planning to come up few other additional support in our system that would make our research more efficient and effective.

Firstly, data collection from third party would hamper your own system out of blue. As a consequence we would like to collect our own set of data to run our system smoothly. Secondly, there are many routes which are dangerous or accidental area to travel in different time of the day. In near future we would like to show the routes to user which are dangerous to travel and what will be the alternate route for them. We thing adding this feature in the future would make our system more useful to the user and make their life more easy in the regular life of commute.

References

[1] Bast, H., Funke, S., Matijevic, D., Sanders, P., and Schultes, D. (2007). In transit to constant time shortest-path queries in road networks. In *Proceedings of the Meeting on Algorithm Engineering & Expermiments*, pages 46–59. Society for Industrial and Applied Mathematics.

[2] 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):1295–1307.

[3] Gao, S., Wang, Y., Gao, Y., and Liu, Y. (2013). Understanding urban traffic-flow characteristics: a rethinking of betweenness centrality. *Environment and Planning B: Planning and Design*, 40(1):135–153.

[4] Gundlegård, D. and Karlsson, J. M. (2013). The smartphone as enabler for road traffic information based on cellular network signalling. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 2106–2112. IEEE.

[5] Horvitz, E. J., Apacible, J., Sarin, R., and Liao, L. (2012). Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. *arXiv preprint arXiv:1207.1352*.

[6] Joshi, M. and Hadi, T. H. (2015). A review of network traffic analysis and prediction techniques. *arXiv preprint arXiv:1507.05722*.

[7] Priyanka, U. and Nishadha, S. G. (2014). S-drive: A smart driving direction system. *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*.

[8] Shashikiran, V., Kumar, T. S., Kumar, N. S., Venkateswaran, V., and Balaji, S. (2011). Dynamic road traffic management based on krushkal's algorithm. In *2011 International Conference on Recent Trends in Information Technology (ICRTIT)*, pages 200–204. IEEE.

[9] Treboux, J., Jara, A. J., Dufour, L., and Genoud, D. (2015). A predictive data-driven model for traffic-jams forecasting in smart santader city-scale testbed. In *2015 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pages 64–68. IEEE.

[10] Turner, A. (2007). 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.

[11] Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., and Chen, C. (2011). Data- driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1624–1639.

[12] "Support Vector Machine," Hierarchical Clustering. [Online]. Available in http://www.saedsayad.com/support_vector_machine.html. [Accessed: 21-Jul-2018].

[13] "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.