Work Trend Analysis based on Bicycle count

Saem Abdulhamid Sheth and Priya Saini

Department of Electrical and Computer Engineering, University of Waterloo
Waterloo, ON, Canada
sasheth@uwaterloo.ca
p25saini@uwaterloo.ca

***Abstract*— In this project, we aim to analyze the work trend based on bicycle count of Bloor St., Toronto, which can provide some useful insights in traffic data analysis. We correlate the bike count data with weather data to determine the changes in the pattern of bike count. In this, we did three different regression models to predict the bicycle count based on various factors affecting it. We also did various data analysis from this information and came up with some meaningful insights which can help to improve the traffic data mining and road infrastructure design and management.**

***Keywords*— Bicycle count, Regression models, traffic data mining, the work trend pattern**

1. *INTRODUCTION*

Toronto city council designed the pilot bike lanes on Bloor St. and installed the sensors to count the bikes passing through it, in the year 2016. The sensors data is then stored in a database, which is available on the Toronto Open data website. The data has 15-minute counts of bikes passing through it. This dataset was meant to send to the council to determine the changes in these lanes and decide whether the lanes should be removed, broadened, mended or continued to be used in the same way. In addition to this, it can be helpful in putting more cycle parking points to the nearby hotspots of this street. These sensors were never removed and this data can, therefore, be taken under consideration to determine the working pattern and traffic data mining purposes.

We obtained this dataset from Toronto's Open Data Catalog, which has data of bike counts on two intersections of Bloor St., Markram St. and Huron St. from February 2018 to February 2019. Also, the data has counts of the bikes in Eastbound and Westbound directions, which played a major role in determining the work patterns. We linked this data with corresponding weather data as weather conditions largely affect the bike riders. We also take consideration of weekends and holidays to determine the variations in riding patterns on these days as compared to the working days.

The aim here is to take the detailed approach of analyzing this data. Hereby using three different regression models, i.e. the Linear regression model, the Random Forest regression model, and the Multiple output regression model, we get to know that bike count is largely dependent on time of the day as well as different weather conditions, namely temperature, total rain, total snow, total precipitation, snow on the ground and wind speed.

Based on the regression model, we designed a program which predicts the total number of counts on both the intersections given the time and different weather conditions as input. We also observed some of the very meaningful insights such as what are the peak hours of bike traffic, when do people generally prefer to ride a bike, how seasonal change affects the bike counts, when people generally go to work and when they come back, how holidays and weekend days affects the bike count etc.

2. *RELATED WORK*

In the year 2014 one statistician of the United States of America did a similar survey on the Fremont bridge of Seattle city. He gathered some very interesting insights such as determining the weekly bike trips, relationship in bike trips on holidays and weekends, seasonal increment or decrement in bike counts, hourly patterns of bike counts[3].

In the same year, another data scientist from the University of Washington science institute did the analysis on the same data as the statistician did on the Fremont bridge of Seattle city. In addition to previous work done by the statistician, he did linear regression to predict the number of bike counts based on the available bike data and weather data and determined how bike counts vary according to the different weather conditions and the time of day[6].

In our project, we did the regression to predict the bike counts using three different models i.e. linear regression model, Random Forest regression model, and the MultiOutput regression model. Based on the R2 values and accuracies of these models, we decided the best suitable model for our application is the MultiOutput regression model. In addition to that, we did K-fold cross-validation, by which we checked the accuracies of these models on the unseen data set.

3. *METHODOLOGY*

The first part of our project involves data collection of bicycle count at different time intervals of the day at Bloor Street situated in Toronto. Another data set required for the analysis is the weather data corresponding to the days of which the bicycle data is available. After retrieving the data, both the data sets are combined to fetch the meaningful insights out of the data sets. Each step of data set gathering is discussed in detail:

1) *Bicycle Data set*: The required data for the bicycle count is available on Toronto's official website of Open Data Catalogue [4]. The given data set on the website contains the bicycle count of Huron Street and Markham Street and has both eastbound and westbound bicycle count of every 15 minutes for the period of February 10, 2018, to February 9, 2019, and has more than 35,000 rows of data. For the project, the data set is combined on an hourly basis as it is more beneficial to gather insights from hourly data rather than every 15-minute data. Further, the hourly data set is combined on a daily basis to analyze the daily traffic trend on the streets for both eastbound and westbound traffic. The program used for combining the data set is provided.

2) Weather Data set: The weather data of Toronto for the given period of February 10, 2018, to February 9, 2019, can easily be found on any weather forecast website and is easily available in CSV format. For our project, the weather data is gathered from Toronto's official weather website[5]. The data was available in different files for the year of 2018 and 2019. We merged the data in a way that the final data belongs to the desired Bicycle data period. The weather data available on the website contained column information for Date/Time, Year, Month, Day, Data Quality, Max Temp (°C), Max Temp Flag, Min Temp (°C), Min Temp Flag, Mean Temp (°C), Mean Temp Flag, Heat Deg Days (°C), Heat Deg Days Flag, Cool Deg Days (°C), Cool Deg Days Flag, Total Rain (mm), Total Rain Flag, Total Snow (cm), Total Snow Flag, Total Precip (mm), Total Precip Flag, Snow on Grnd (cm), Snow on Grnd Flag, Dir of Max Gust (10s deg), Dir of Max Gust Flag, Spd of Max Gust (km/h), Spd of Max Gust Flag. But for the analysis, we do not need all these data columns as bicycle count cannot be related to some of the data and moreover most of the values of certain column data were missing, so we deleted some of the columns and kept Date/Time, Year, Month, Day, Mean Temp (°C), Total Rain (mm), Total Snow (cm), Total Precip (mm), Snow on Grnd (cm), Spd of Max Gust (km/h). The data set thus chosen was almost complete but did consist of some missing values. Only 30-40 rows had the missing entries in Snow on Grnd column which was filed with 0s as all the entries corresponded to the summer season. Otherwise, the dataset was complete, which was beneficial and helped achieve the accuracy of the model.

3) *Weekend and Holiday Data*: The bicycle trend based on weather data can be related to the working days and non-working days around the calendar of the Ontario region in Canada. Meaningful understanding can be gained from the data excluding weekends and holiday data from the data set so generated. Different working habits and user patterns can be analyzed.

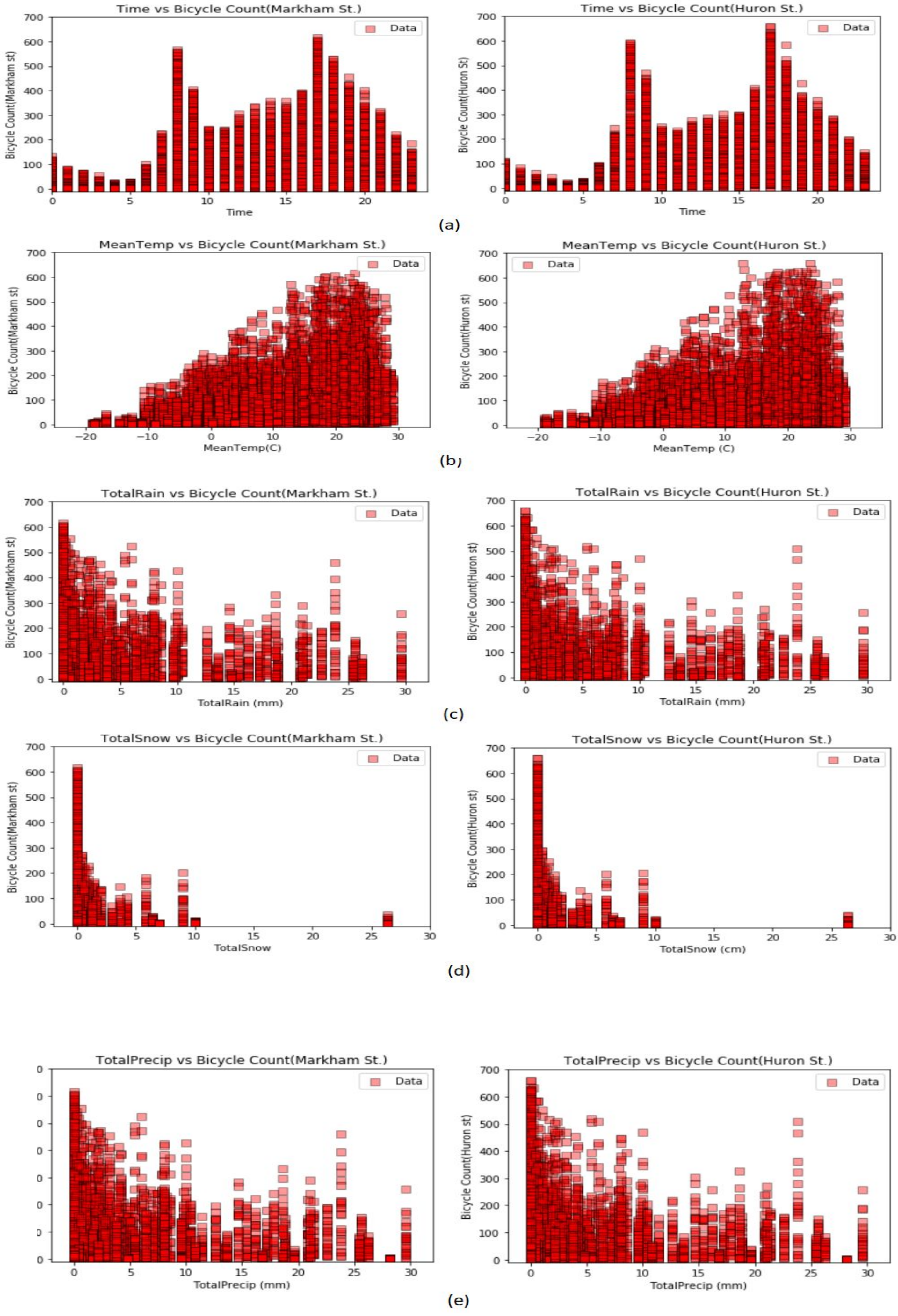
4) *Data Merge:* After successfully fetching bicycle data, weather data and non-working days data (includes weekends and holidays of the region), the data sets are combined. Both the weather data and bicycle data are merged into a single file. Further files are processed to combine the data sets on a daily basis, which is another way in which the problem statement needs to be achieved. Again, more processing is done to exclude the weekend and holiday data from the data set generated after merging the data sets. All the data sets, generated in this step, are kept in different files for easier processing while making the machine to learn the trend.

After properly cleaning and merging data according to project requirements, the parameters so obtained were compared against each other to check the relationship between them. The relationships analyzed are discussed in detail in the Result section. Based on their relationships, the explanatory variables were chosen. For the project, Time, Mean temperature of the region in degree Celsius, total rain in mm, total snow in cm, total precipitation in mm, snow on the ground in cm and speed of the wind in km/hr are considered. According to these variables, the total bicycle count on both Huron St and Markham St can be predicted. Not only the total bicycle count, eastbound and westbound bicycle count results for the streets can also be predicted.

Three machine learning models were used to gather useful insights from the data sets. The choice of the models was based on the relationship analysis of the parameters and suggestions provided by the authors who already worked in this domain. We used Linear regression, Random Forest regressor, and MultiOutput regressor as learning models. Detail insights based on each model is provided in the next section. We calculated the R^2 values and K-fold cross-validation accuracies for each of the models to decide the best model. Multiple experiments were conducted by changing the random state value of the random forest and MultiOutput regressor to achieve the best possible result. For learning the models, the dataset was separated into testing and training datasets by 80-20% ratio. Models were trained on training datasets and tested on the testing data. Further R^2 values were calculated and the output generated is plotted for deeper analysis and comparing the models against each other. We decided to carry on the project with MultiOutput regressor based on the highest value of the K-fold accuracy and R^2 values. For the value of k equal to 4, we got accuracy varying from 50% - 60%, details of which are discussed further.

Multiple graphs are plotted to see the different trend patterns of eastbound and westbound traffic on both the streets which clearly shows how the incoming and outgoing traffic trend differs with the time of the day based on various parameters supporting the count of the bicycles on the streets.

Besides this, daily analysis of the traffic has been done and plotted wisely. The trend helps to predict how the count of bicycles is distributed throughout the year based on the season


Time vs Bicycle Count(Markham St.)
Time vs Bicycle Count(Huron St.)
Bicycle Count(Markham st)
Bicycle Count(Huron St)
Time
Data
MeanTemp vs Bicycle Count(Markham St.)
MeanTemp vs Bicycle Count(Huron St.)
MeanTemp(C)
MeanTemp (C)
TotalRain vs Bicycle Count(Markham St.)
TotalRain vs Bicycle Count(Huron St.)
TotalRain (mm)
Bicycle Count(Huron st)
TotalSnow vs Bicycle Count(Markham St.)
TotalSnow vs Bicycle Count(Huron St.)
TotalSnow
TotalSnow (cm)
TotalPrecip vs Bicycle Count(Markham St.)
TotalPrecip vs Bicycle Count(Huron St.)
TotalPrecip (mm)
0
100
200
300
400
500
600
700
−20
−10
0
5
10
15
20
25
30

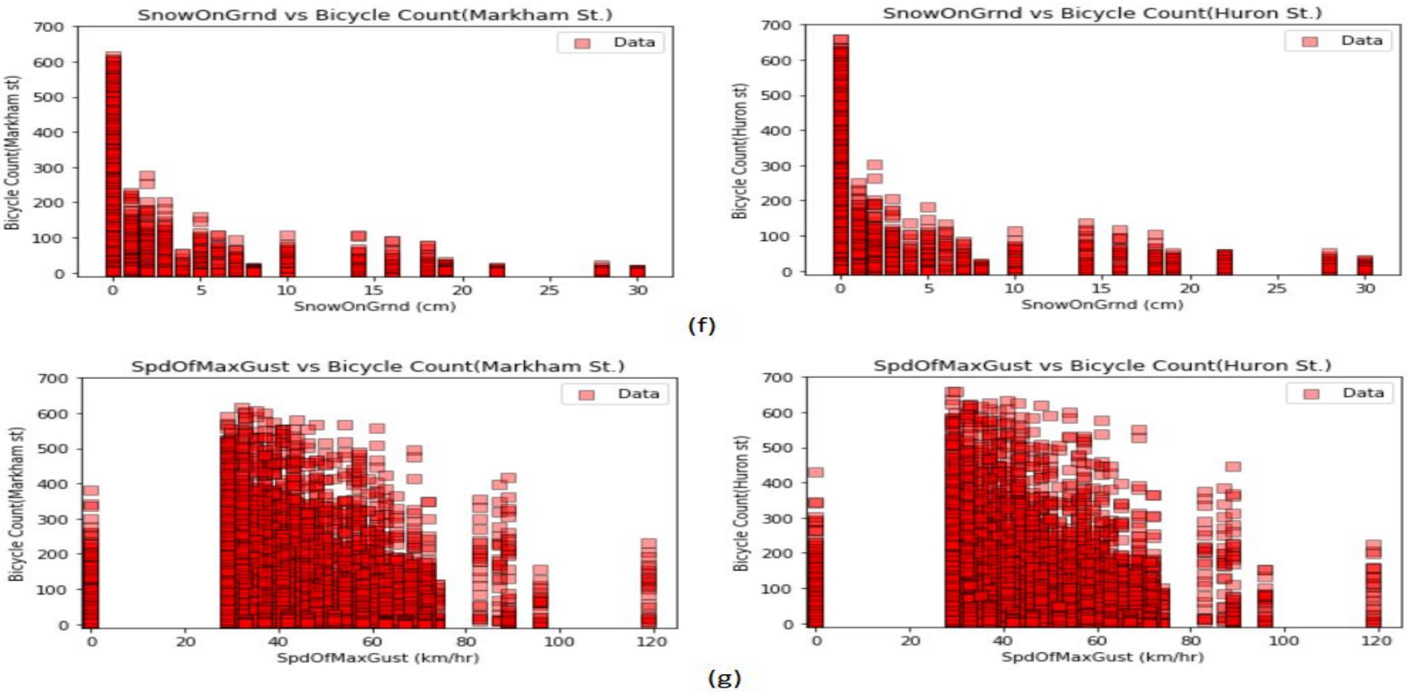
(a)

(b)

(c)

(d)

(e)



(f)

(g)

Fig 1. Explanatory Variable Analysis

patterns. Separate trend patterns were observed when plots were drawn based on working and non-working days.

A deeper insight into each graph plotted for the project has been discussed in detail in the next section.

4. RESULTS

A. Parameter analysis

The project involves various steps from explanatory variables selection to selecting the regression model that best fits the model proposed and predict the data accurately not only on the training data set but also on any unseen data. In order to achieve the goal, the first step taken is to analyze the bicycle count on Markham Street and Huron Street with respect to individual explanatory variables.

Figure 1 describes each explanatory variable trend with the bicycle count on both the streets. Figure 1(a) takes Time parameter into account and the graph plotted explains that bicycle count follows a particular pattern suggesting the rush hours and working patterns of the area. As the time approaches 5 A.M., the count tends to increase and rapidly increases till 8 A.M. and reaches its peak at 8 A.M., after which it decreases. It can be inferred that most people start leaving early for work. Again there is a sudden peak at 5 P.M.. which is usually the returning time of the people from their work and as time passes the traffic decreases. An obvious trend can be observed from this graph which indicates the traffic trend based on time.

Figure 1(b) plots the mean temperature of the city with respect to the bicycle count. An increasing relationship can be seen between both the parameters, which can easily be interpreted as people prefer to ride bicycles during summer rather than winter. A significant increase in the count can be observed as temperature increases beyond 10 °C.

The plot between the total rain and count as shown in Figure 1(c), shows that more bicycles are observed on the days with no rain or less rain than the days having heavy rain. Although from the graph we can say people ride bicycles on the day of rain but it cannot be said whether the rain was on a particular instant of time on which the count has been measured. It may be possible that at rush hours, there was no rain to affect the traffic much. But a generalized analysis would be the count decreases with an increase in rain.

Figure 1(d) shows that snow greatly affects the bicycle count. More bicycles are used when there is no snow or snow in centimeters is closer to 0. Some uncertain pattern can be observed when snow is nearly 25cm which probably accounts for some error values in the data set.

Figure 1(e) tells the relationship between the total precipitation and traffic. Precipitation in weather data accounts for anything that is falling from the sky, which involves both rain, hail, sleet, or snow. It can be seen that precipitation affects negatively to the count. The bicycle count decreases as precipitation increases. Data points get sparse as the precipitation increase in the value. These cases could be explained in a way that there may be no precipitation during the hours when people preferably commute.

Another factor that is taken into account to predict the bicycle traffic trend is snow on the ground in centimeters, as shown in Figure 1(f). A visible decrease in the pattern can be seen as snow on the ground increases. Some missing and error values account for uncertain graph predictions that is near to 25cm to 30cm of snow, as it is difficult for any rider to ride a bicycle having that much snow on the ground.

Again a decreasing trend can be visualized from the plot between the speed of gust (in km/hr) versus bicycle count, as shown in Figure 1(g). As the wind speed increases, people tend to avoid riding a bicycle on the streets. Which is obvious as it requires more effort to ride the bicycle in the wind.

Clearly, the count is greatly affected by all the parameters chosen and thus are taken into consideration for making our machine learning model and predicting results based on these variables.

B. Choosing Regression Models

i) *Linear regression* is the machine learning algorithm used for predicting the continuous value output (Total bicycle counts on Huron and Markram St.) based on the combination of continuous-valued input variables (input Features).

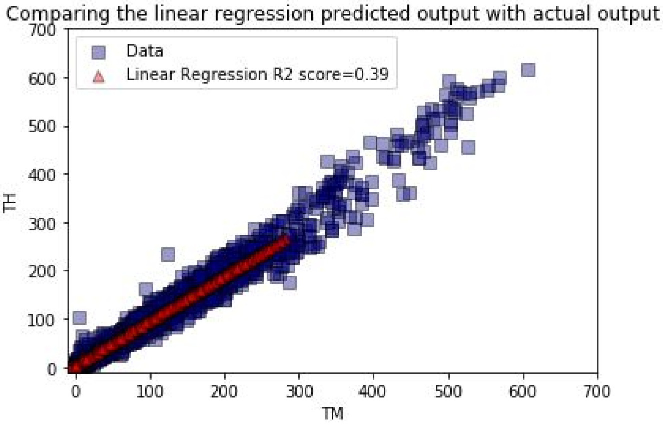


Fig. 2

Figure 2 represents the comparison between the predicted output and the actual output of the linear regression model. It indicates the R2 score of the model is 0.39. Since R2 score of the linear regression model is so low, it is not a good model for the application.

ii) *Random forest regression* is an ensemble machine learning algorithm used for predicting the output, which generates the multiple decision trees at the training time and predicts the output, at last the average value of outputs is considered as final output.

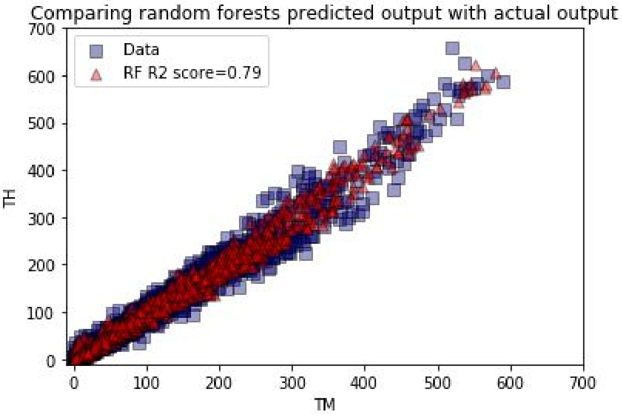


Fig. 3

The comparison between the predicted and actual output using Random Forest regressor is shown in Figure 3. The R2 score calculated for the model is 0.79. R2 score of Random Forest regressor is relatively better than the linear regression model, it can be considered to predict the output parameters for the proposed model i.e. total bicycle count on Markram St. and Huron St.

iii) *MultiOutput regressor* is used to predict multiple outputs at the same time(total bicycle Counts on both the streets).

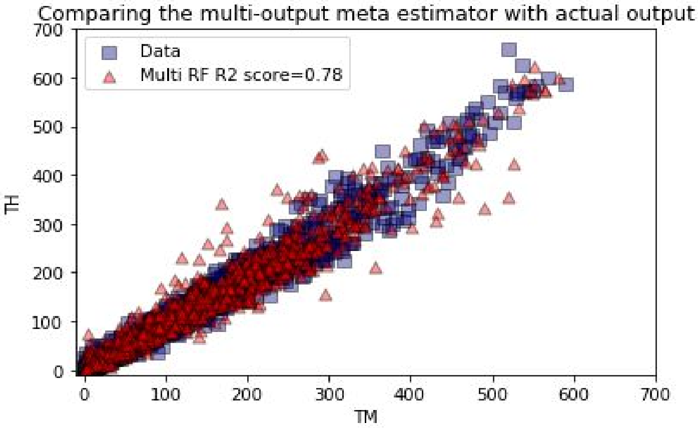


Fig. 4

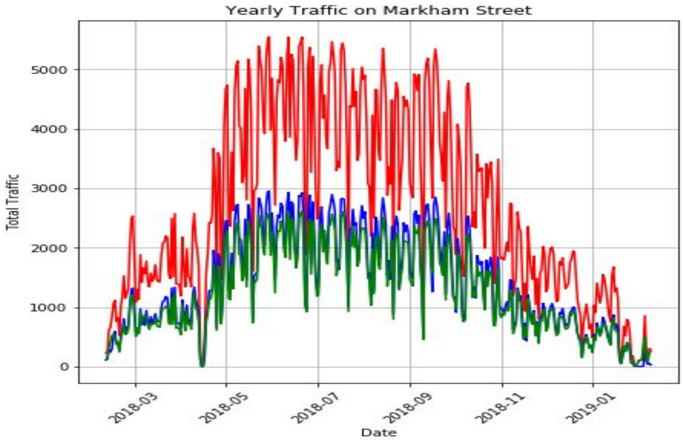
Figure 4 represents the comparison between the MultiOutput regressor based predicted and actual output. The R2 score of the model indicated in the graph is calculated to be 0.78. Since the R2 score of the model is good, it can also be taken into account for analysis and prediction of bicycle count on both the streets.

The following table compares the K-fold and R^2 accuracies of these 3 models.

Model	**K-fold Accuracies**	**R^2 scores**
Linear Regression	0.056	0.37
Random Forest regression	0.5514	0.79
MultiOutput regression	0.5517	0.79

C. *Traffic pattern analysis*

Figure 5 depicts the yearly change in the number of bikes on Markram St. and Huron St. respectively. The blue line in the figure represents the traffic in the Eastbound direction and the green line represents the traffic in the Westbound direction while the red line represents the total bike counts. From the graphical analysis, it can be said that bike counts are more in Spring and Fall seasons than in Winter season. It can also be said that bike count starts to increase in April ending and tends to fall in November beginning. From the plot, it can be inferred that comfortable weather for bike riding is between May and September, and bikers do not prefer riding bicycles from October till April ends. The effect of seasons on the bicycle count can clearly be seen using the graph.



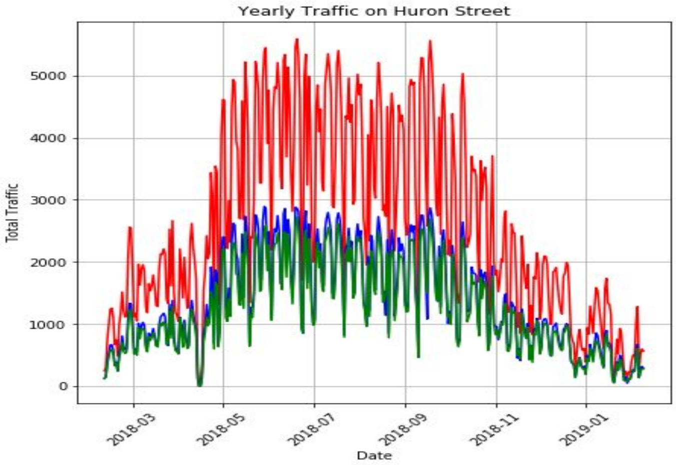


Fig. 5

Figure 6 represents the bike counts on Markram St. and Huron St. on a working day. The trend of the bicycle count remains similar to the overall pattern shown in Figure 5.

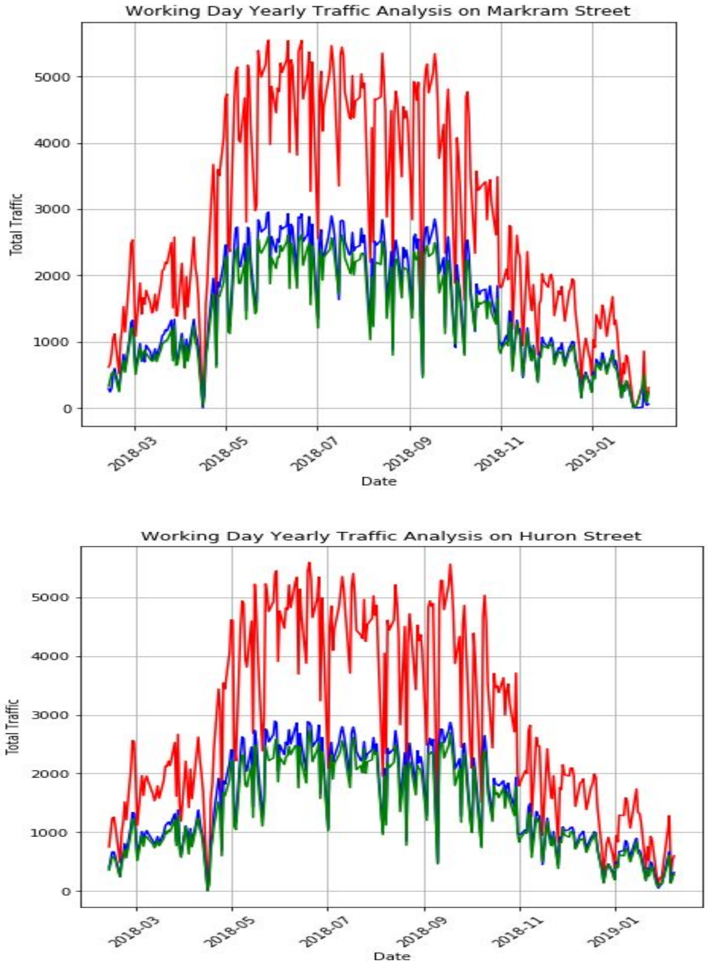


Fig. 6

The bike count of Markram St. and Huron St. on any non-working day is shown Figure 7. From the comparison between the figure 6 and 7, it can be said that the number of bikes significantly decreases on the non-working day as compared to the count of bicycles on working days. Obvious insight can be made as many people use lesser bicycles on a non-working day.

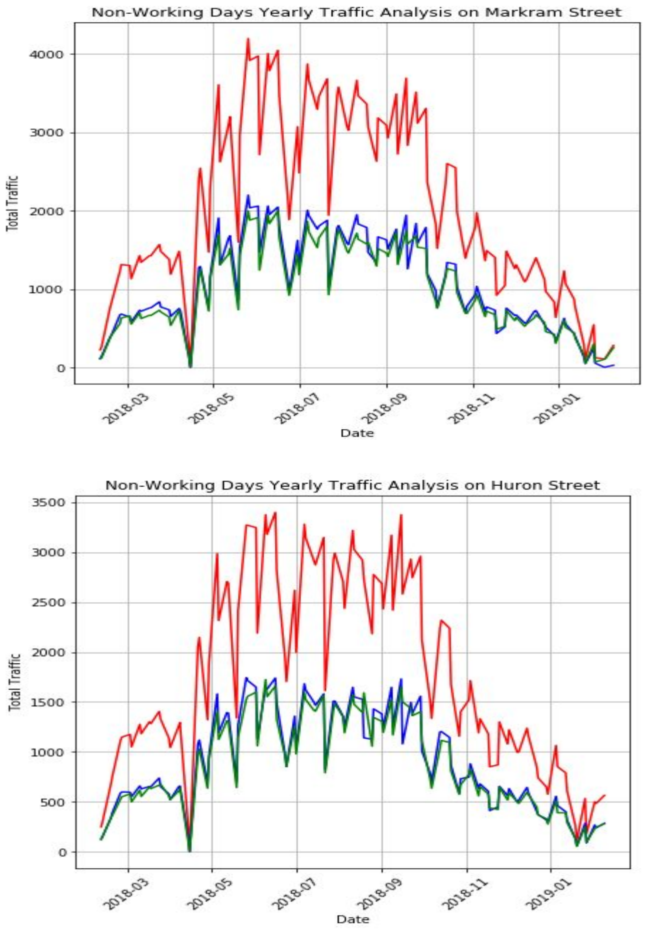
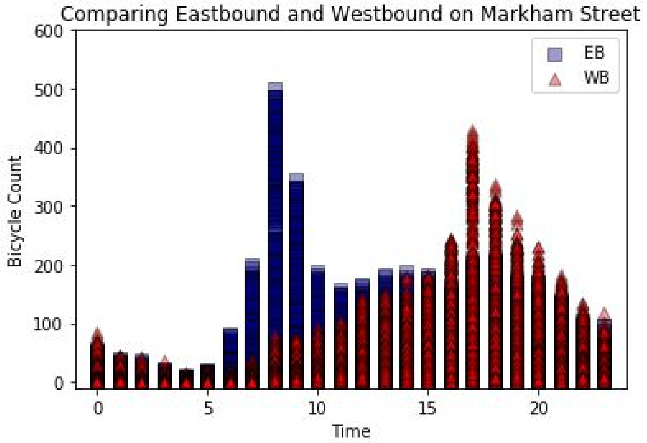


Fig. 7



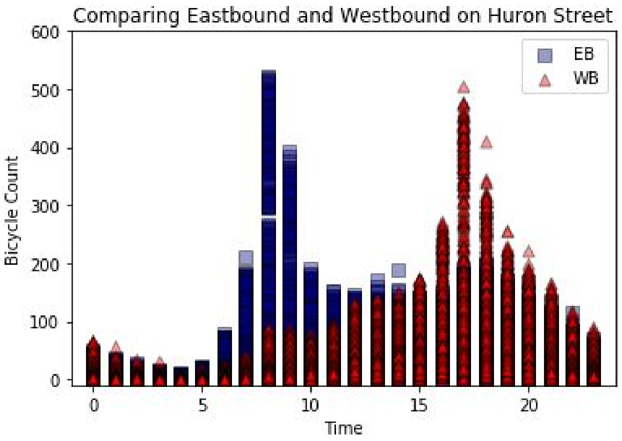
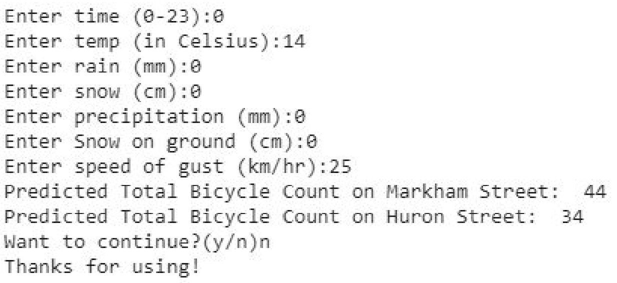


Fig. 8

Figure 8 depicts the change in bike counts in Eastbound and Westbound directions with respect to the time parameter. Eastbound traffic can be visualized using blue square boxes in the given figure and little red triangle denote the westbound traffic. A clear insight that can be drawn is that at 8 a.m. people leave for work and at 5 in the evening people come back from work. Also, the peak traffic hours are 8 o'clock in the morning and at 5 o'clock in the evening.

D. PREDICTION



```
Enter time (0-23):0
Enter temp (in Celsius):14
Enter rain (mm):0
Enter snow (cm):0
Enter precipitation (mm):0
Enter Snow on ground (cm):0
Enter speed of gust (km/hr):25
Predicted Total Bicycle Count on Markham Street:  44
Predicted Total Bicycle Count on Huron Street:  34
Want to continue?(y/n)n
Thanks for using!
```

Fig. 9

Figure 9 represents the output of the program which is developed to predict the total bicycle count given the input parameter values. Here, the input parameters are our explanatory variables i.e. Time and weather values and the output parameters are bike counts on Markram and Huron St.

5. CONCLUSION AND FUTURE WORK

In this work, we analyzed the traffic pattern of bicycle riding. This analysis was based on two unique datasets i.e. weather dataset and bike counts during the day timing dataset. Based on which we came up with some interesting and useful insights on how different weather situations affects the bike riding. In addition to this, we did some comparison between bicycle counts on working days and non-working days (holidays and weekends). Moreover, the comparison of traffic on opposite sides (Eastbound and Westbound) of the streets was done, which became very useful in inferring the working pattern of the city and peak traffic hours during the day. At last, we predict the bike counts depending on different situations, which is helpful in anticipating future traffic.

By evaluating and learning patterns from these datasets many insights are obtained, which can be helpful in traffic analysis and infrastructure maintenance problems.

In future, the same analysis can be done on different streets of the city and comparing the results with some additional insights such as working patterns comparison of different parts of the city and infrastructure improvements required in these parts can be made.

6. REFERENCES

[1] Chan, M., Gapski, G., Hurley, K., Ibarra, E., Pin, L., Shupac, A. & Szabo, E. (November 2016). *Bike Lanes, On-Street Parking and Business in Parkdale: A study of Queen Street West in Toronto's Parkdale Neighbourhood.* Toronto, Ontario.

[2] Ledsham, T., Liu, G., Watt, E. & Wittmann, K. (2014). *Mapping Cycling Behaviour in Toronto*. Toronto: Toronto Cycling Think & Do Tank.

[3] Mike Logsdon (June 2014). *A statistical analysis of biking on the Fremont Bridge, Part 1: Overview,* retrieved from https://www.seattlebikeblog.com/2014/06/09/a-statistical-analysis-of-biking-on-the-fremont-bridge-part-1-overview/

[4] Toronto. Open Data Catalogue, retrieved from https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/

[5] Government of Canada. *Historical Climate Data*, retrieved from http://climate.weather.gc.ca/

[6] Jake VanderPlas (June 2014). I*s Seattle Really Seeing an Uptick In Cycling?,* retrieved from https://jakevdp.github.io/blog/2014/06/10/is-seattle-really-seeing-an-uptick-in-cycling/

[7] Jake VanderPlas (July 2015). *Learning Seattle's Work Habits from Bicycle Counts.,* retrieved from https://jakevdp.github.io/blog/2015/07/23/learning-seattles-work-habits-from-bicycle-counts/

[8] Sztabinski, F. (2009). *Bike Lanes, On-Street Parking and Business: A Study of Bloor Street in Toronto's Annex Neighbourhood*. Toronto: Clean Air Partnership.