STATION BASED BIKE SHARING DEMAND PREDICTION

By

Xin Zhao, BCom, Ryerson University, 2018

A Major Research Project

presented to Ryerson University

in partial fulfillment of the requirements for the degree of

Master of Science

in the Program of

Data Science and Analytics

Toronto, Ontario, Canada, 2019

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A MAJOR RESEARCH PROJECT (MRP)

I hereby declare that I am the sole author of this Major Research Paper. This is a true copy of the MRP, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this MRP to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this MRP by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my MRP may be made electronically available to the public.

STATION BASED BIKE SHARING DEMAND PREDICTION

Xin Zhao

Master of Science 2019

Data Science and Analytics

Ryerson University

ABSTRACT

Bike-sharing have been increasing popularity in recent years due to its usage flexibility, reduction in traffic congestion and carbon footprint. Being able to accurately predict each bike-sharing station's demand at any given hour is crucial for inventory management. This report first manipulated Bike Share Toronto ridership data with Toronto City Center weather data from 2016 Quarter 4 to 2017 Quarter 4, then implemented machine learning algorithms in particular Regression Trees, Random Forest, and Gradient Boosting Machine (GBM) to forecast station based hourly bike-sharing demand in the City of Toronto. The results indicated that Random Forest based prediction model was the most accurate model by comparing Root Mean Square Error (RMSE) of all bike-sharing stations.

Key words:

Bike-Sharing Demand Prediction, Regression Trees, Random Forest, Gradient Boosting Machine

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Bike-sharing system is a service that allows users to rent bike for a short amount of time from one location to another location [1]. There are two main types of bike-sharing system, with docking stations and dockless. For a docked bike-sharing system, bikes are locked at each dock, and users need to pay for one time rental fee or annual membership fee to unlock the bike. After each usage, users need to return the bike to any bike-sharing station. For a dockless bike-sharing system, users leave bikes at their destination and the next user will locate the bike through an app's GPS [1]. This system creates convenience for users, however it requires city's corporation on bike-sharing planning and monitor how each bike is placed in the city to reduce illegal parking.

Bike-sharing systems have been increasing popularity in recent years due to its flexibility and being an alternative of transportations with fewer emissions. Based on data compiled by MetroBike's Bike-Sharing Blog, as of May 2018 more than 1,600 bike-sharing programs were in operation worldwide compared to 855 in 2014 [2]. In just 4 years, the number of bike-sharing programs has increased by 88%.

This report focuses on Toronto docked bike-sharing system, Bike Share Toronto. As of 2019 August, the bike-sharing system has 470 bike-sharing stations, and 5000 bikes. There had been significant increase ridership from 2016 to 2018¸ there were 834,235 ridership in 2016, 1,510,802 ridership in 2017 and 1,975,384 ridership in 2018 [3]. It is import to accurately predict each station's bike-sharing demand pattern by implementing machine learning algorithms. To get detailed insights, each bike-sharing station's demand had been divided into hours and weather data had been added to find any correlations between weather and bike-sharing demand. The result of this report can be used for bike-sharing inventory management to balance each bike-sharing station's demand and supply.

2. LITERATURE REVIEW

The research on predicting bike-sharing demand is vast. Before the era of big data, linear regressions were widely used for prediction models. As more public data became available, machine learning algorithms such as Random Forests, Support Vector Machine (SVM), Recurrent Neural Network (RNN) and Convolutional Graph Neural Network have been implemented for prediction models. Deep learning algorithms like the ones presented in Table 1 require large amounts of data to train the models. Since this project only contained 1.4 million bike-sharing data of the City of Toronto, the literature review focused but not exclusively on Regression Tree based machine learning algorithms which required less training data than others.

In the study of Singhvi et al. [4], linear regression models were created to predict pairwise bike-sharing demand by using New York City's Citi Bike and Taxi public dataset from April 2014 to July 2014. The dataset contained bike usage between the time periods of 7 AM to 11 AM and included the following information: (a) start station id, (b) end station id, and (c) station geographic location for each bike trip. There were other variables used as covariates for prediction including precipitation and spatial variables (e.g. population and housing units of each neighbourhood). By making scatter plots, the authors found that by assigning each bike-sharing station to the corresponding neighbourhood showed a higher correlation between bike and taxi usage. This hypothesis was further proved when they ran the regression models. They created two linear regression models to predict bike-sharing demand, and used step wise backward selection to select significant independent variables. The regression model based on station-level only achieved 24% of the bike-sharing demand. Whereas the regression model based on neighbourhood level leverages the spatial information of each neighbourhood to achieve 74% accuracy on the training set. When they ran the model on the test set, the neighbourhood regression model achieved 0.42

RMSE. The limitation of this paper was that it only covered the bike-sharing demand from 7 AM to 11 AM, which lacked the accuracy of hourly bike-sharing demand for the whole day.

One of the popular machine learning techniques used for bike-sharing demand prediction is Random Forest. Patil et al. [5] used 17,000 bike-sharing data from Capital Bike Share in Washington DC between 2011 January 1st to 2012 December 31st to predict bike-sharing demand by implementing Random Forest machine learning algorithm. The dataset contained *time, season, holiday, working day, weather, temperature, humidity, wind speed, casual, registered, and count*. This dataset contained several categorical data (e.g. *season and weather*) besides continuous variables. In order to train the model, they converted the categorical variables into dummy variables. For example in the *season* variable, they converted spring as 1, summer as 2, fall as 3 and winter as 4. They used Root Mean Squared Logarithmic Error (RMSLE) to validate their model and achieved 0.5 on the test set. The limitation of this model was inaccuracy when predicting outliers in the dataset. For example there might be an unusual amount of bike demand for a large group of tourist who show up on a random day.

In a study done by Feng & Wang [6], they compared the performance of bike-sharing demand prediction by using multiple linear regression and Random Forest on the same Capital Bike Share public dataset in Washington DC. They first created a multiple linear regression model with seven variables *weather, temp, attempt, humidity, wind speed, date time* and *season* to predict the bike-sharing demand. Unfortunately, the model only received 32.7% accuracy on the test set because the categorical variables are converted to dummy variables. The second prediction model they created was using Random Forest with 5000 trees. The result increased the accuracy of bike-sharing demand prediction to 80% on the test set.

Besides Random Forest, a research done by Sachdeva & Sarvanan [7] compared the performance of several machine learning based prediction models including SVM, Neural Network, Poisson Regression, Random Forest, and GBM with the same dataset as the previous two studies. The results showed that SVM, Neural Network, and Poisson Regression were not good at predicting bike-sharing demands because of the categorical variables in the dataset. Whereas Random Forest and GBM achieved RMSLE of 0.39 and 0.37 respectively on the test set. In order to achieve better results, they combined Random Forest based model with GBM based model by using a linear combination with weights of their relative performance and achieved RMSLE of 0.36.

In recent years, many researches on bike-sharing demand predictions have incorporated more complex machine learning techniques. Pan et al. [8] conducted a research on predicting bike-sharing demand by training a Deep Long Short Term Memory (LSTM) Recurrent Neural Network (RNN). The dataset they used to train the model was the bike usage data of New York City's Citi Bike System in 2017. Along with bike-sharing data, weather and time data were also considered in the prediction model. There were a total of more than 16 million trips, between 360 bike-sharing stations in the entire dataset. During the research, the authors clustered stations based on neighbourhoods by using the community detection method proposed by Rosvall et al. [9] and chose the top two communities with respect to the number of bike trips to filter low-quality data. Each bike-sharing station contained the number of rents and returns in different time periods of a day. The data was then organized as a matrix consists of N column features and M column time periods to feed into the RNN model. The output of the model were the number of rents and returns in each time period. The net demand was computed by finding the difference between bike rents and

returns. The results showed that their LSTM model achieved RMSE of 3.0203 on the training set and 1.9323 on the test set.

More literature reviews on bike-sharing demand prediction using deep learning algorithms are summarized in Table 1.

Authors	**Time Span**	**Number of Records**	**Number of Bike-Sharing Stations**	**Methods**	**Results (Error Metric)**
Chai, Wang & Yang [10]	51 months	49 million	827	Multi Graph Convolutional Neural Network	4.75 (RMSE)
He, Lin, & Peeta [11]	36 months	28 million	272	Graph Convolutional Neural Network – Data Driven Graph Filter	2.12 (RMSE)
Xu et al. [12]	11 months	18 million	721	Back-Propagation Neural Network	2.053 (RMSE)
Zhou et al. [13]	24 months	5 million	167	Markov Chain Models	10.23 (Average Absolute Error)

Table 1: Summary of Deep Learning Literature Review

3. EXPLORATORY DATA ANALYSIS

Data pre-processing is a major process in the presented project. It provided an overview of the dataset by using different data visualizations. Afterwards, the data was transformed to the necessary input shape of the bike-sharing demand prediction models.

The original bike-sharing dataset was downloaded from Open Data - City of Toronto website [14]. Due to data availability, the website only provides data from 2014 Quarter (Q) 4 – 2015 Q3 and 2016 Q3 - 2017 Q4. Therefore this report analyzed Bike Share Toronto Ridership data from 2016 Q4 to 2017 Q4 with a total of 1.7 million ridership records for 268 bike-sharing stations. Each ridership data was in *CSV* format and contained attributes of *trip start time, trip end*

day and time, trip duration, trip start station, trip end station, and *user type*. Bike Share Toronto bike-sharing station information was also collected from Open Data - City of Toronto website [14] to fill the missing station IDs for 2016 Q4, 2017 Q3, and 2017 Q4 dataset. During this process, it was found that the station names were not consistent in different quarters. For example, in 2016 Q4 dataset, station names of Queens Park/ Bloor St W, 424 Wellington St. W were recognized as Queen's Park/Bloor St W" and 424 Wellington St W in the bike-sharing station file.

In order to make the data consistent, a Python package called *SequenceMatcher* was used to find the similarity ratio of the station's names in different files. After comparing the result, it was found that any station names with a ratio of 0.88 for 2016 Q4, 0.85 for 2017 Q3 and 2017 Q4 were qualified to be replaced. Since this report was predicting hourly bike-sharing demand, all *trip_start_time* was converted to hourly based time, and created new variables to represent year, month, day, and hour. Variable *day_of_week* was created in a range of 0-6 to represent Monday to Sunday. Variable *season* was also created to represent Spring (March to May) as 1, Summer (June to August) as 2, Fall (September to November) as 3, and Winter (December to February) as 4.

Besides inconsistent bike-sharing station names, there were some abnormal records in *trip_duration_seconds* attribute. In Figure 1, there are two box plots of bike-sharing trip duration in seconds, and the box plot on the left shows there are a lot of outliers in the original dataset. Interquartile Range (IQR) was used to detect the outliers in *trip_duration_seconds* attribute [15]. IQR is calculated as follows, firstly sort the values in ascending order, and find first (25^{th} percentiles) and third quartiles (75^{th} percentiles). Then use Q3-Q1 to find IQR, and lower bound is calculated as q1 - (1.5 * IQR), upper bound is calculated as q3 + (1.5 * IQR). The data with (1) trip duration lower than upper bound, (2) greater than lower bound and (3) 60 seconds were kept for further analysis. The box plot on the right in Figure 1 shows the distribution graph of bike-

sharing trip duration in seconds after outlier removal, and it clearly shows a better normal distribution of the dataset.

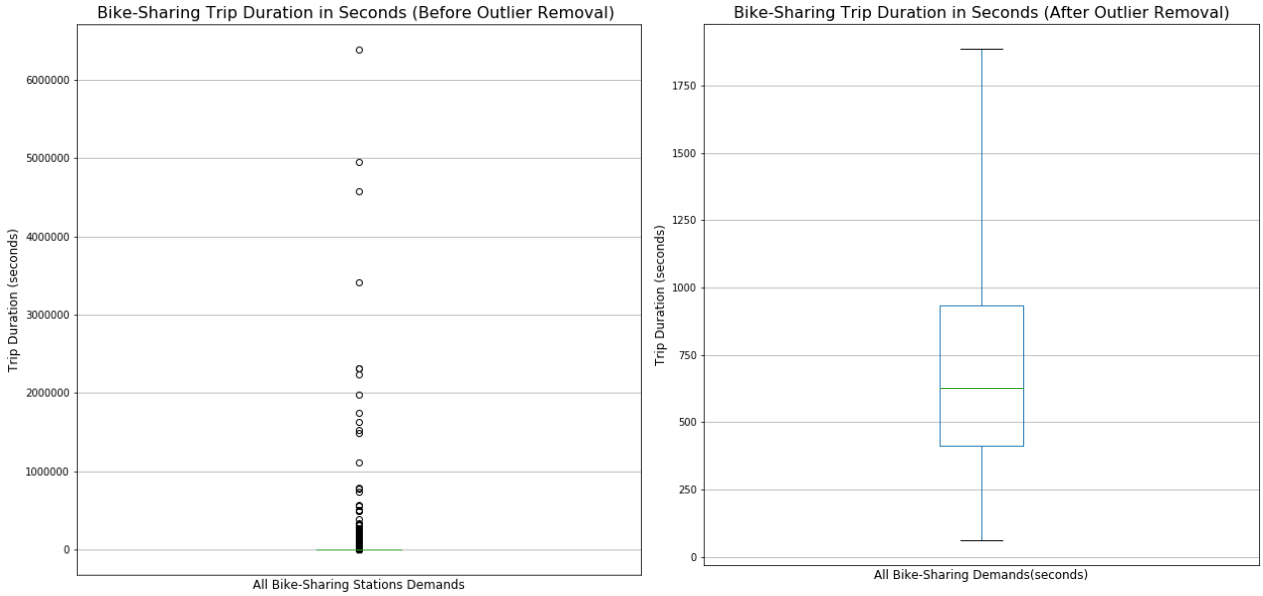

Figure 1: Bike-Sharing Trip Duration in Seconds (Before Outlier Removal & After Outlier Removal)

Before looking into each bike-sharing station, an exploratory data analysis was conducted for bike-sharing demand of all stations. Figure 2 shows the number of bike-sharing demands in different seasons. The bar chart shows that most of the demand was in fall and summer. The peak in demand might be caused by school seasons which starts in September. Time was further divided into months, days of week and hours to find any insights. From the data visualization of Figure 3, 4, and 5, it was found that the most bike-sharing demand was during October, Wednesday, and 17:00. Whereas the least bike-sharing demand was during January, Sunday, and 4:00. In order to see if membership type has a strong impact on the demand, a pie chat was created. On the right side of Figure 5 shows more than 80% of the demand had a membership.

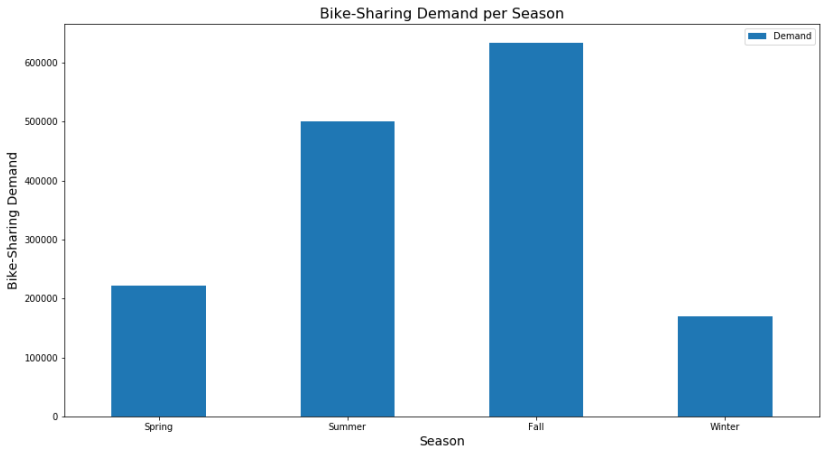

Figure 2: All Stations Bike-Sharing Demand per Season

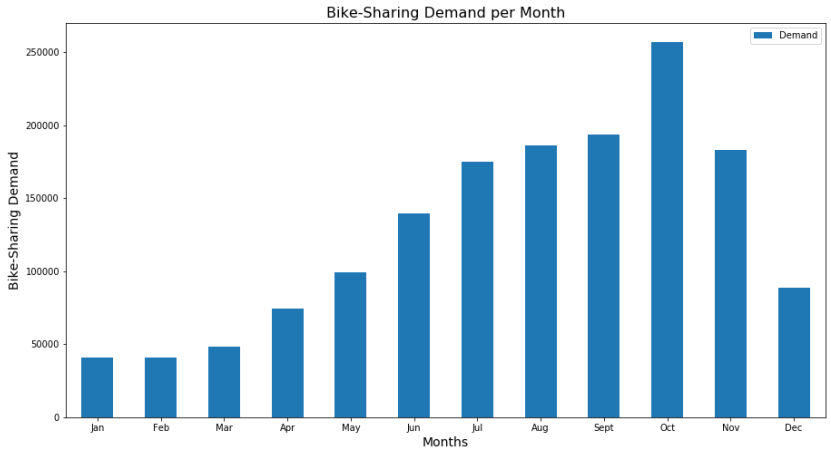

Figure 3: All Stations Bike-Sharing Demand per Month

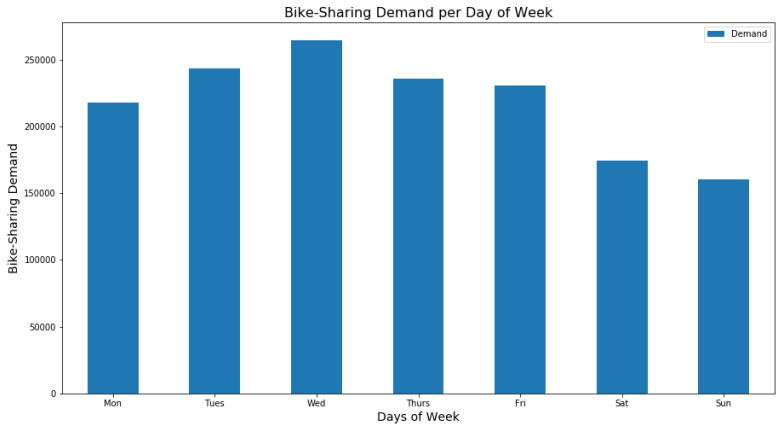

Figure 4: All Stations Bike-Sharing Demand per Day of Week

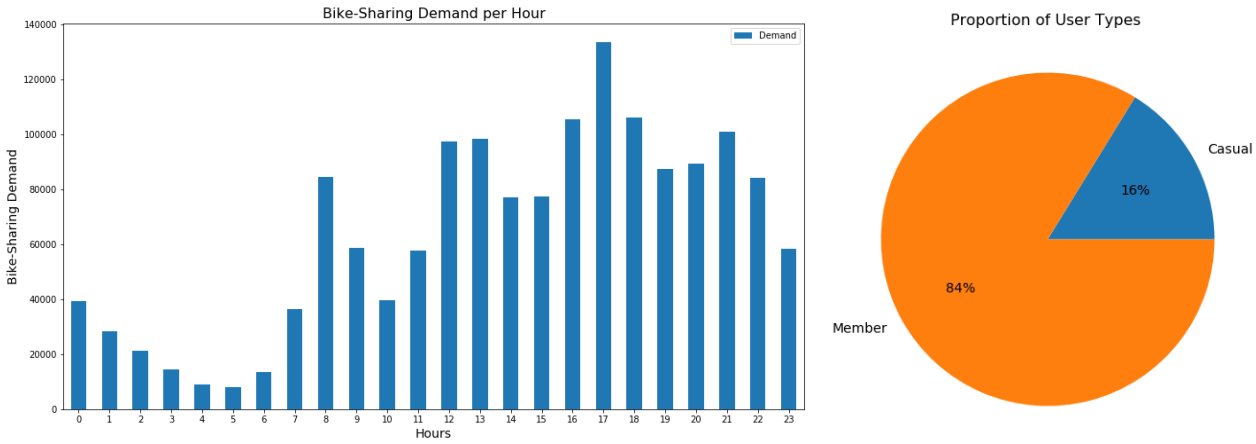

Figure 5: All Stations Bike-Sharing Demand per Hour & Proportion of User Type

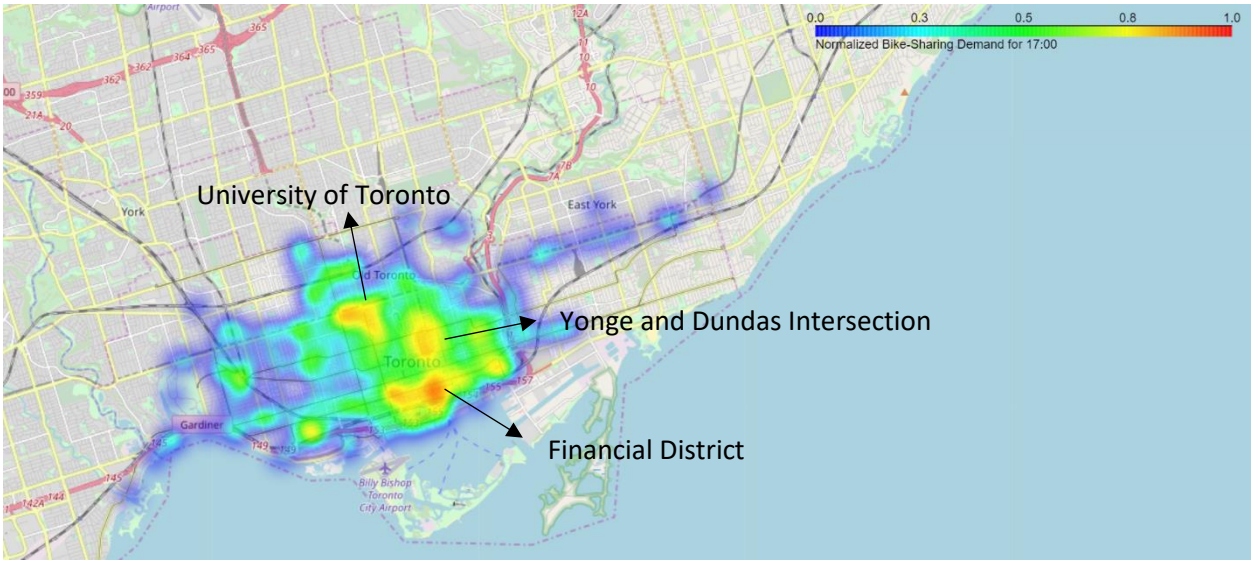

Figure 6: Normalized Bike-Sharing Demand for 17:00 Time Period

Figure 6 is a heat map that shows the normalized Bike-sharing demand for each station at 17:00 (highest demand hour). From the heat map, the most bike-sharing demand at 17:00 was near the financial district, Yonge and Dundas intersection, and near University of Toronto.

By going through the original bike-sharing dataset, three new boolean attributes were created to represent *non-member*, *member*, *and holiday*. If the *user_type* is Casual, then assign 1 to *non-member*, and 0 to *member*, whereas if the *user_type* is Member, then assign 0 to *non-member*, and 1 to *member*. *Holiday* attribute was created by verifying the bike-sharing demand date was included in Canada's official holidays. In order to find out the total hourly bike-sharing

demand of members and non-members a *groupby* function was used on *non-member* and *member* columns. For the bike-sharing stations with zero demand in some hours from 2016 Q3 to 2017 Q4, an empty hourly demand timestamp data frame was created in the time period and left joined with the original bike-sharing data frame on *time* attribute. Then assigned one demand to *member* and *Total_Demand* column in order to predict hourly bike-sharing demand consistent with all bike-sharing stations.

Besides bike-sharing information, hourly weather data of Toronto City Center [16] from 2016 Q4 to 2017 Q4 was also added to help with a better prediction on bike-sharing demand. During data exploration, it was found that there were a lot of missing values in some attributes (Figure 7). Therefore, the weather attribute with the least number of missing values were kept for further analysis were *Temp(°C), Dew Point Temp (°C), Rel Hum (%), Wind Dir (10s deg), Wind Spd (km/h), Visibility (km) and Stn Press(kPa).* For each of these attributes, the mean value was used to fill the missing values.

In order to avoid multicollinearity in the prediction model, each pair of the remaining weather attribute and total sharing-bike demand were calculated by the Pearson correlation coefficient formula (see Figure 8).

	Attribute	# of Missing Values
0	Date/Time	0
1	Year	0
2	Month	0
3	Day	0
4	Time	0
5	Temp (¡ãC)	12
6	Temp Flag	10966
7	Dew Point Temp (¡ãC)	12
8	Dew Point Temp Flag	10966
9	Rel Hum (%)	11
10	Rel Hum Flag	10967
11	Wind Dir (10s deg)	619
12	Wind Dir Flag	10646
13	Wind Spd (km/h)	19
14	Wind Spd Flag	10959
15	Visibility (km)	21
16	Visibility Flag	10957
17	Stn Press (kPa)	13
18	Stn Press Flag	10965
19	Hmdx	9787
20	Hmdx Flag	10968
21	Wind Chill	9072
22	Wind Chill Flag	10968
23	Weather	9212

Figure 7: Number of Missing Values of Weather Attributes

$$r = \frac{\Sigma(X_i - \overline{X})\,(Y_i - \overline{Y})}{[\Sigma(X_i - \overline{X})^2\,\Sigma(Y_i - \overline{Y})^2]^{1/2}}.$$

X_i: individual X point at index i

Y_i : individual X point at index i

$\overline{X}$: sample mean of X variable

$\overline{Y}$: sample mean of Y variable

Figure 8: Pearson Correlation Coefficient Formula [17]

Figure 9 shows the correlation matrix of weather attributes with total bike-sharing demand. *Temp(°C)* and *Real Hum (%)* had high correlation with *Total_Demand*. *Dew Point Temp (°C) and Visibility (km)* had been removed to avoid multicollinearity, since they have high correlation with *Temp(°C)* and *Real Hum (%)* and lower coorelation with *Total_Demand*.

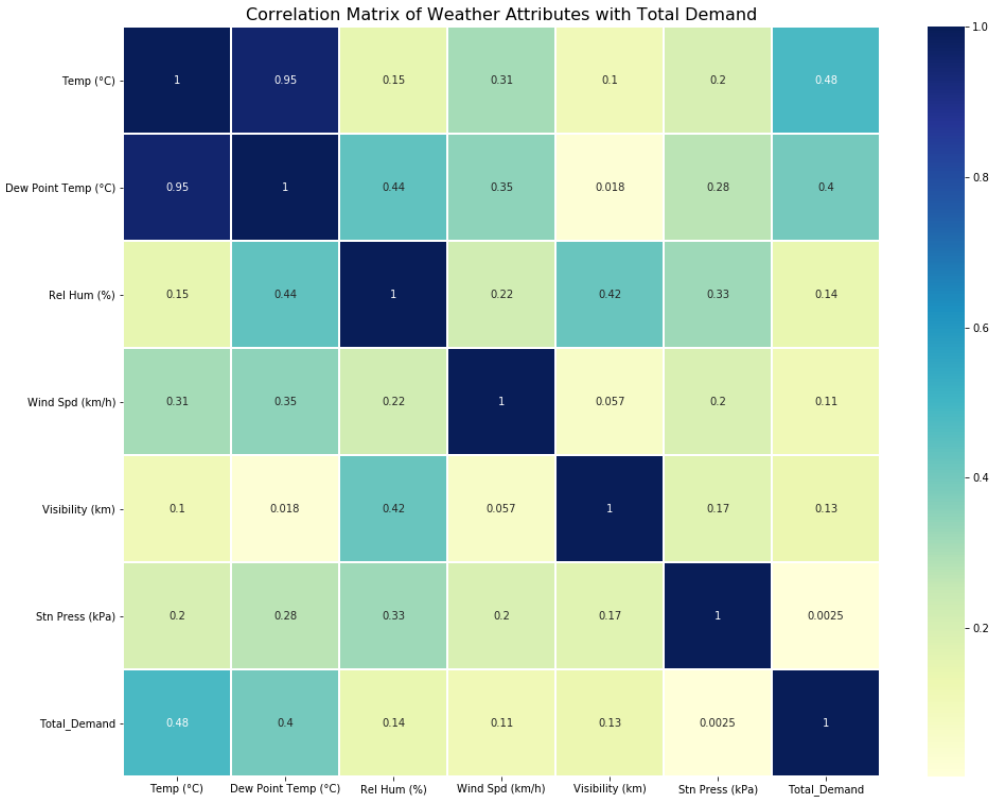

Figure 9: Correlation Matrix of Weather Attributes with Total Bike-Sharing Demand

After data exploration and pre-processing, a summary of all attributes that were used for each bike-sharing station demand prediction model is shown in Table 2.

Variable Name	Variable Type	Variable Range
Time	Time stamp	2016-10-01 00:00:00 to 2017-12-31 23:59:00
Temp (°C)	Float	-19.3 to 30.2
Rel Hum (%)	Percentage	18 to 100
Wind Spd (km/h)	Float	0.0 to 74.0
Stn Press (kPa)	Float	97.7 to 103.3
Day_of_Week	Integer	0 to 6
Year	Integer	2016 to 2017
Season	Integer	1: Spring 2: Summer 3: Fall 4: Winter
Month	Integer	1 to 12
Day	Integer	1 to 31
Hour	Integer	0 to 23
Holiday	Boolean	0 or 1
Non-Member_Demand	Integer	0 to 458
Member_Demand	Integer	0 to 900
Total_Demand	Integer	0 to 1101
Station_ID	Integer	7000 to 7268

Table 2: Summary of All Pre-Processed Attributes

4. METHODOLOGY AND EXPERIMENTS

During the experiments, different approaches were used to treat categorical variables *year, season, month, day, hour* and *holiday*. At first, one hot encoding method was used for each categorical variable by creating new binary variables for each category; however the results showed a much higher error rate than treating each category as an integer. Therefore *year, season, month, day, hour* and *holiday* variables were kept as integers.

This project aimed to use machine learning algorithms to predict hourly bike-sharing demand for each bike-sharing station in the City of Toronto. In order to find if there was any significant variance between user types, the variables: (1) total bike-sharing demand, (2) non-member bike-sharing demand and (3) member bike-sharing demand were set as the dependent variable for prediction.

The machine learning algorithms used to predict hourly bike-sharing demand for each station were Regression Trees, Random Forest, and GBM. The reason why Regression Tree based machine learning algorithms were chosen was because they work better with smaller dataset and categorical variables compared to the Neural Network based algorithms. A Regression Tree is a type of supervised learning method and it works by splitting the dataset into regions, and then for each region finds the regression that minimizes the Mean of Squared Error (MSE) (see Figure 10) between each predicted and actual values. This process is recursively repeated until the entire data is covered [18].

Regression Trees can be visualized as a flow chart diagram (see Figure 11). Regression Trees start from the root node which contains the entire dataset and then it splits the dataset by testing every value in all independent variables to find the one that minimizes the MSE. The terminal nodes contain the final values that best fit the regression. However, the Regression Tree based prediction model had problems of low prediction accuracy and high variance because this model used greedy algorithms to minimize MSE and had high correlations in the predicted results.

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \widehat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

Figure 10: Mean of Squared Error Formula [19]

y_i: i th actual value

$\hat{y}_i$: i th predicted value

n: number of values

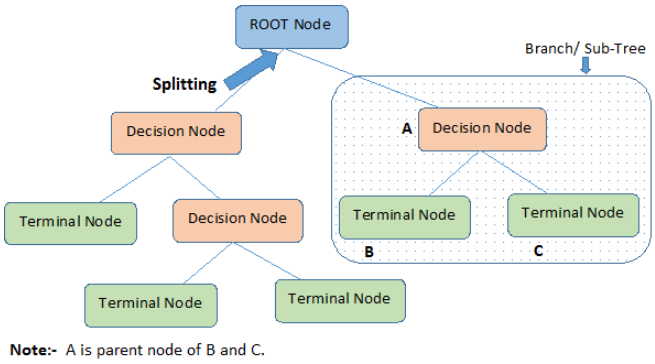

Figure 11: Regression Tree Diagram [20]

There were two approaches to overcome the problems of the Regression Tree model. The first approach was to implement the Random Forest ensemble learning method (see Figure 12). The Random Forest method works by combing multiple Regression Tree predictions from random subsets of the dataset and use the mean value of predictions as the final result. For each tree, it limits to a random sample of features therefore the prediction are less correlated, and the model is less likely to over fit on the training data [21].

The second approach was to use GBM (see Figure 13). This is a method to ensemble weak prediction models. The main difference between GBM and Random Forest is that each tree is grown sequentially by using the errors of the previous tree to produce a Regression Tree with a lower error value. GBM starts off by calculating the mean value of the dependent variable, then builds a tree based on the errors of the previous tree. Afterward, GBM combines the original leaf with the new tree to predict values. To prevent overfitting, GBM uses a user-defined learning rate to scale the contribution from the new tree. Each time there is a new tree added to the prediction model, the errors get smaller. GBM continues to build trees until adding additional trees does not significantly reduce the errors [22].

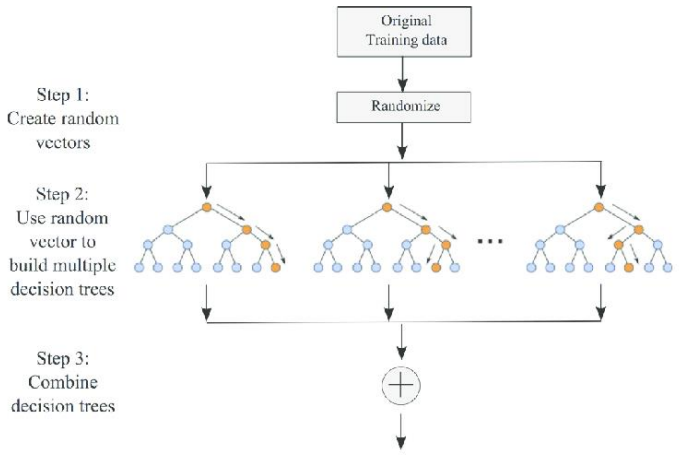

Figure 12: Random Forest Diagram [21]

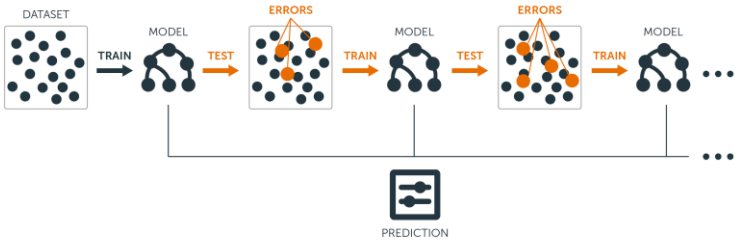

Figure 13: GBM Diagram [22]

In order to compare and evaluate each model's performance, cross-validation was used by randomly splitting 70% of the entire dataset to training dataset and 30% of the entire data to testing dataset. RMSE was used as the evaluation metric (see Figure 14). RMSE is computed as the standard deviation of the residuals. A lower RMSE value indicates the model predicts closer to the actual values [24] which is a more accurate model.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{pred,i} - y_i)^2}{n}}$$

Figure 14: RMSE Formula [24]

y_i: i th actual value

$y_{pred,i}$: i th predicted value

n: number of values

The programming language used for this experiment was Python, and the whole experiment was built in *Jupyter* environment. *Sklearn* packages were used to assist with creating the prediction models, including *DecisionTreeRegressor*, *RandomForestRegressor* and *GradientBoostingRegressor*. For Random Forest and GBM based prediction models, *GridSearchCV* package was used to find out the number of estimators between 50, 100 and 500 that can achieve the lowest RMSE value.

5. RESULTS AND DISCUSSIONS

The objective of this project is to predict hourly bike-sharing demand for each bike-sharing station in the City of Toronto. Three machine learning algorithms were used for the prediction models. The first model was based on Regression Trees, the second one was based on Random Forest and the last one was based on GBM. For evaluation, each model's error rate was calculated based on test set's average RMSE for all 268 bike-sharing stations. Random Forest based prediction model achieved the best result with 0.639 RMSE, followed by GBM based prediction model's 0.678 RMSE. The worst model was Regression Tree based which had 0.883 RMSE.

In order to see if there are any significant difference between user types, non-member and member bike-sharing demand were then predicted separately. Figure 15 shows that all three prediction models achieved a lower RMSE with non-member demand than member demand. This is because non-member users rent bike with more patterns and were more affected by weather. For example, 40% of non-member users rent bikes during weekends; the variance of temperature of non-member is 36 compared to member's 68.

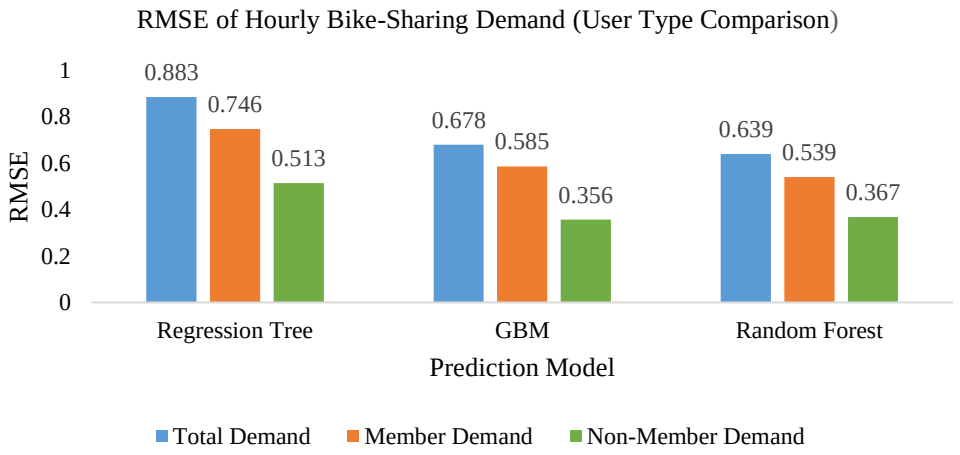

Figure 15: RMSE of Hourly Bike-Sharing Demand (User Type Comparison)

Further analysis was taken by selecting the top 5 highest demand bike-sharing station to compare with the RMSE for all stations. Results (see Figure 16) show that for the top 5 highest demand bike-sharing stations, the Random Forest based prediction model still achieved the best result with 1.753 RMSE compared to 2.123 and 2.451 RMSE for the GBM and the Regression Tree based models.

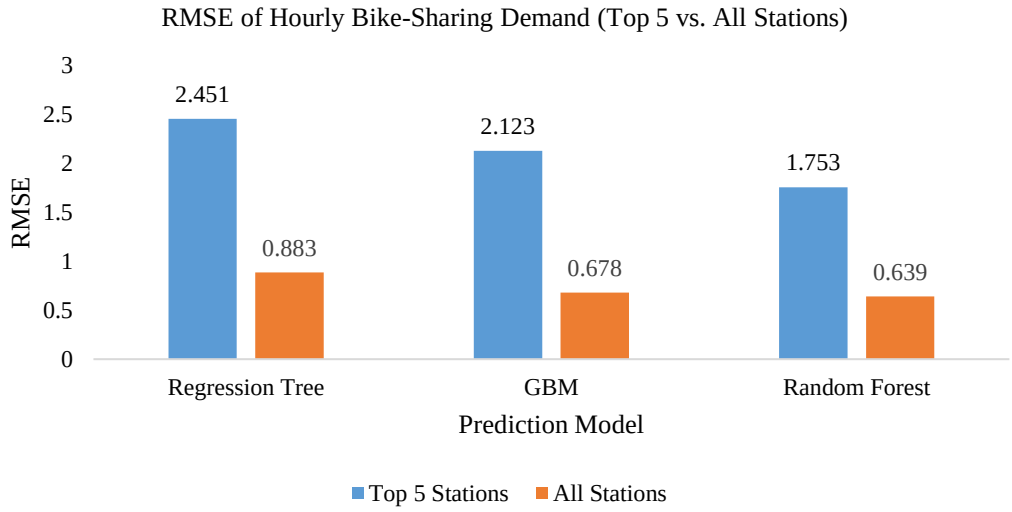

Figure 16: RMSE of Hourly Bike-Sharing Demand (Top 5 vs. All Stations)

Out of three algorithms, Random Forest achieved the best RMSE result, therefore Random Forest prediction model was chosen for more detailed analysis. Figure 17 shows the seasonal comparison of average hourly bike-sharing demand RMSE for all bike-sharing stations. The result indicates winter had the lowest RMSE, whereas summer had the highest RMSE. The reason behind this is during winter less people rented bikes due to the cold and snowy weather in the City of Toronto, therefore hourly bike-sharing demand in winter was more influenced by weather and achieved better prediction results. During summer season there were higher bike-sharing demand and some non-member users rented bikes for specific outdoor events such as music festivals or parades which is not considered in this model, therefore it increased prediction difficulty in the summer season.

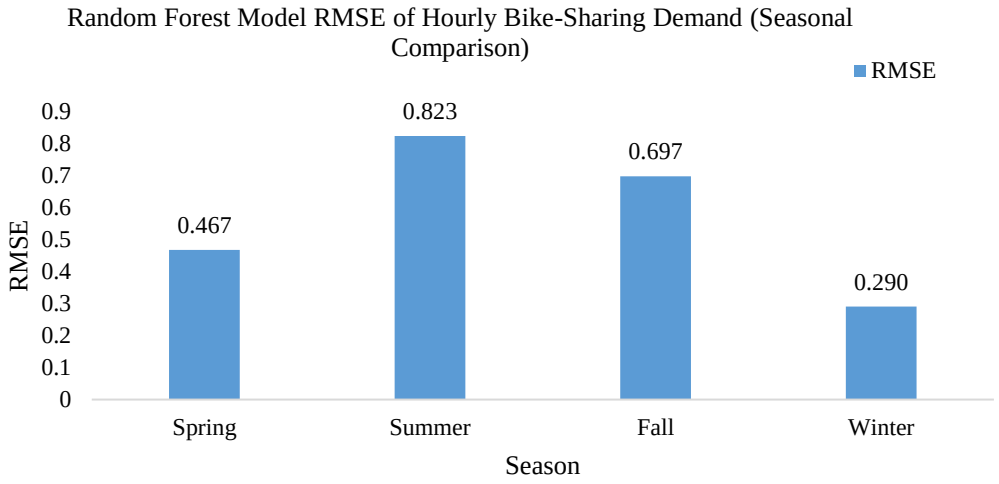

Figure 17: Random Forest Model RMSE of Hourly Bike-Sharing Demand (Seasonal Comparison)

Figure 18 shows that the Random Forest model achieved lower RMSE value during the weekends than weekdays. The top 5 bike-sharing demand stations during weekdays are Cherry Beach, Union station, Bay St/ Wellesley St W, Bay St/ College St (East Side), Simcoe St / Wellington St. These stations are close to the financial district and Union station which is one of the largest transportation terminal in the city of Toronto. This means during weekdays people h ave to go to work and rent bikes with less influence on the weather variables. Whereas the top 5b ike-sharing demand stations during weekends are Cherry Beach, Queen St W / Portland St, Q ueens Quay W / Lower Simcoe St, York St / Queens Quay W, King St W / Spadina Ave. Thesest ations are closer to the entertainment district, China town and Habourfront Ferry. This means th e bike-sharing demand during weekends were more correlated to the independent variables in th e prediction model.

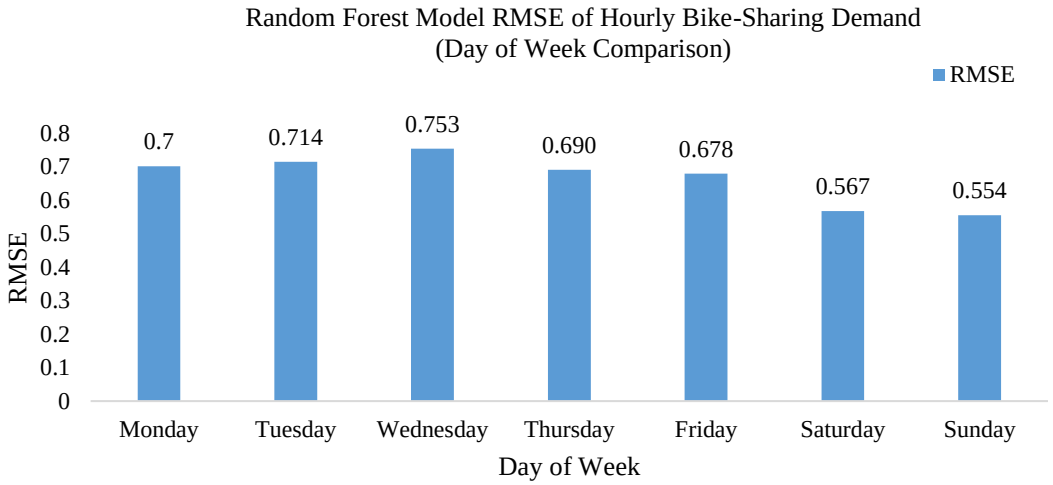

Figure 18: Random Forest Model RMSE of Hourly Bike-Sharing Demand (Day of Week Comparison)

Three bike-sharing stations were chosen from different areas of the City of Toronto to demonstrate the predicted hourly total bike-sharing demand. Figure 19 shows the hourly bike-sharing demand for Union station. Although there are some differences between the predicted and actual demand, the Random Forest based model still predicted the pattern of the demand. Union station is a transportation hub and commuters from the suburbs arrive at this station. Therefore the

bike-sharing demand was the highest during 7:00 to 8:00. During lunch hours 11:00-13:00, the bike-sharing demand was at the second highest. During rush hour 17:00, there were not many people rented bikes from Union station as people were arriving to Union station and take other transportations to leave the city.

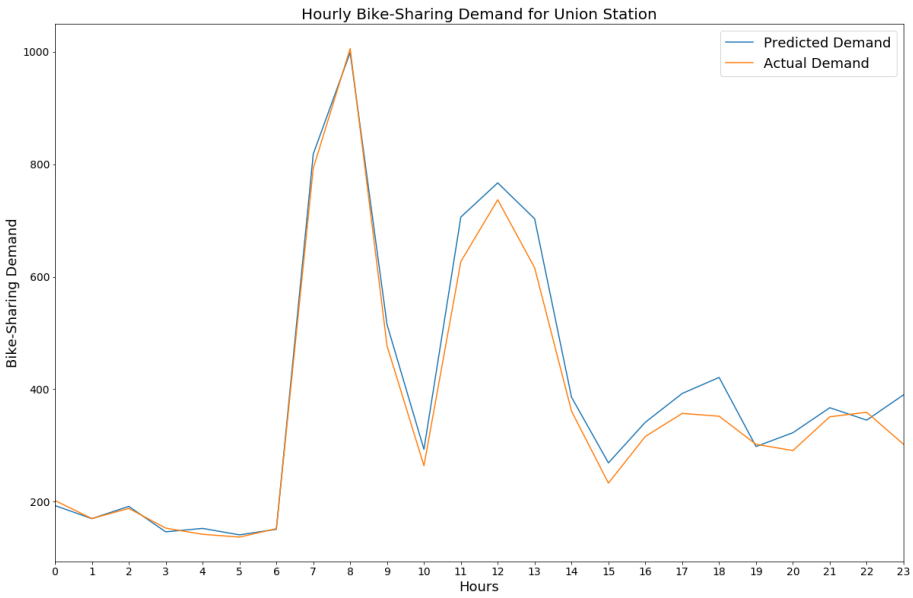

Figure 19: Hourly Bike-Sharing Demand for Union Station

Figure 20 shows the hourly bike-sharing demand for HTO Park station, the most bike-sharing demand occurred after 18:00. This is because HTO Park is one of Toronto's iconic place to relax. The park is located near a long sand beach with amazing views. Therefore people rent bikes at the station and ride along the beach after work hours.

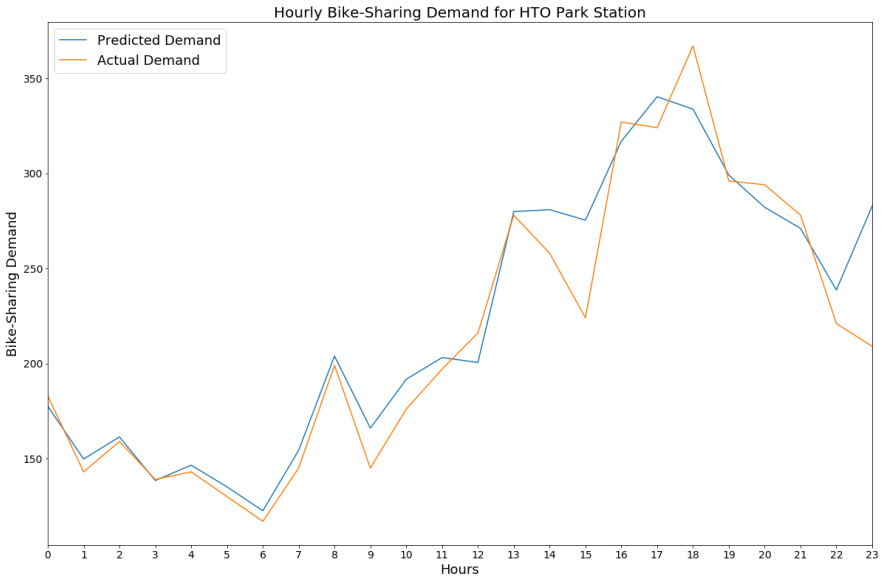

Figure 20: Hourly Bike-Sharing Demand for HTO Park Station

Figure 21 shows the hourly bike-sharing demand for Victoria and Dundas station. The pattern is quite different than the previous ones. The time period which had the most bike-sharing demand was between 20:00 to 21:00. This is because this bike-sharing station is located near Ryerson University and Eaton shopping mall. The mall closes at 21:30, and the latest lecture at Ryerson University finishes at 21:00. Therefore people rent bikes at this station the most between 20:00 to 21:00 and the prediction model perfectly matches this pattern.

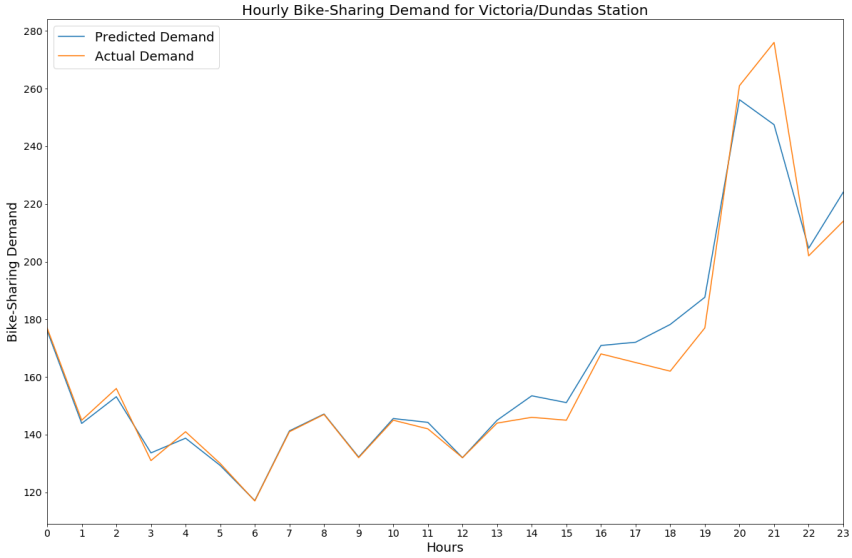

Figure 21: Hourly Bike-Sharing Demand for Victoria/Dundas Station

Simcoe St/Wellington St South Station (see Figure 22) shows multiple bike-sharing demand spikes. There were high bike-sharing demand occurred at this station during 8:00 - 9:00, 12:00 – 13:00, 16:00 – 18:00 and 20:00 to 22:00. The reason behind this is because there are multiple large office buildings and restaurants near this bike station. People rent bikes from this station to work, during lunch break, get to subway stations and have dinner after work. The concert hall Roy Thomson Hall is also located close to this bike station, which attracted bike-sharing demand after the performance during 20:00 to 22:00.

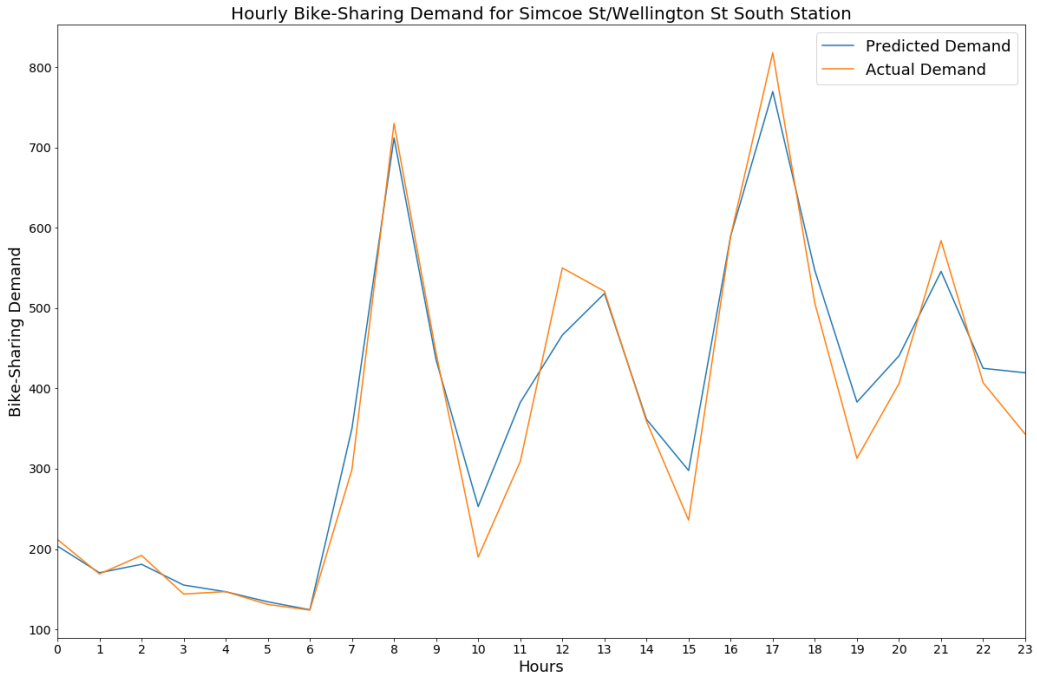

Figure 22: Hourly Bike-Sharing Demand for Simcoe St/Wellington St South Station

6. CONCLUSION AND FUTURE WORKS

This project explores models to predict station based hourly bike-sharing demand in the City of Toronto based on ridership data and weather data. Regression Trees, Random Forest, and GBM were implemented. Out of three algorithms, Random Forest based model achieved the best results with 0.638 RMSE. When the hourly bike-sharing demand was separated by user types, and top 5 demand bike stations, the Random Forest model still achieved the best accuracy with the

lowest RMSE. More importantly, the Random Forest model predicted demand pattern of each bike-sharing stations which can provide more detailed insights and assist with balancing each bike-sharing station's demand and supply.

As there is increasing popularity of bike-sharing in the City of Toronto, a future analysis with more ridership records can be used to train station based hourly bike-sharing demand prediction model with more complex deep learning algorithms to achieve better accuracy. Future studies can be conducted by using *from_station* and *to_station* to predict the most travelled bike routes of each hour. These bike routes can be used to suggest future bike lane development. This Random Forest based prediction model can also be improved by adding more independent variables to analysis more detailed insights and unusual bike-sharing demand. For example there could be a Boolean variable to indicate whether there is a major event including music festival, food festival, parade, etc. Demographics variables including population, household income, age, education level can also be added as well while the bike-sharing demand prediction is based on the nearest neighbourhood.

7. APPENDIX A – PROJECT GITHUB LINK:

https://github.com/xinzhao-datascience/Major-Research-Project

8. REFERENCES:

[1] Wikipedia contributors. (2019). Bicycle-sharing system. In *Wikipedia, The Free Encyclopedia*. Retrieved August 12, 2019, from https://en.wikipedia.org/w/index.php?title=Bicycle-sharing_system&oldid=909808109

[2] Richter, F. (2018). Infographic: Bike-Sharing Clicks Into Higher Gear. Retrieved August 12, 2019 from https://www.statista.com/chart/14542/bike-sharing-programs-worldwide/

[3] Wikipedia contributors. (2019). Bike Share Toronto. In *Wikipedia, The Free Encyclopedia*. Retrieved August 12, 2019, from https://en.wikipedia.org/w/index.php?title=Bike_Share_Toronto&oldid=910817077

[4] Singhvi, D., Singhvi, S., Frazier, P. I., Henderson, S. G., O'Mahony, E., Shmoys, D. B., & Woodard, D. B. (2015). Predicting Bike Usage for New York City's Bike Sharing System. *AAAI Workshop: Computational Sustainability*. 110-114

[5] Patil, A., Musale, K., & Rao, B. P. (2015). Bike Share Demand Prediction using RandomForests. *IJISET International Journal of Innovative Science, Engineering & Technology*, *2(4)*

[6] Feng, Y., & Wang, S. (2017). A forecast for bicycle rental demand based on Random Forests and multiple linear regression. *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*. IEEE doi:10.1109/icis.2017.7959977

[7] Sachdeva, P., & Sarvanan, K. N. (2017). Prediction of Bike Sharing Demand. *Oriental Journal of Computer Science & Technology*, *10*(1), 219-226

[8] Pan, Y., Zheng, R. C., Zhang, J., & Yao, X. (2019). Predicting bike sharing demand using recurrent neural networks. *Procedia Computer Science,147*, 562-566. doi:10.1016/j.procs.2019.01.217

[9] Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences, 105* (4), 1118-1123. doi:10.1073/pnas.0706851105

[10] Chai, D., Wang, L., & Yang, Q. (2018). Bike Flow Prediction with Multi-graph Convolutional Networks. *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL 18.* doi:10.1145/3274895.3274896

[11] Lin, L., He, Z., & Peeta, S. (2018). Predicting Station-level Hourly Demand in a Large-scale Bike-sharing Network: A Graph Convolutional Neural Network Approach. *Transportation Research Part C: Emerging Technologies*, 97, 258-276. doi:10.1016/j.trc.2018.10.011

[12] Xu, X., Ye, Z., Li, J., & Xu, M. (2018). Understanding the Usage Patterns of Bicycle-Sharing Systems to Predict Users' Demand: A Case Study in Wenzhou, China. *Computational Intelligence and Neuroscience*, 2018, 1-21. doi:10.1155/2018/9892134

[13] Zhou, Y., Wang, L., Zhong, R., & Tan, Y. (2018). A Markov Chain Based Demand Prediction Model for Stations in Bike Sharing Systems. *Mathematical Problems in Engineering,2018*, 1-8. doi:10.1155/2018/8028714

[14] City of Toronto. (2018). Transportation - Data Catalogue. Retrieved August 12, 2019 from https://www.toronto.ca/city-government/data-research-maps/open-data/open-data-catalogue/transportation/

[15] Upton, G. J., & Cook, I. (1996). Understanding Statistics. Oxford University Press.

[16] Climate Change Canada. (2019). Historical Climate Data. Retrieved August 12, 2019 from http://climate.weather.gc.ca/

[17] Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. The American Statistician,42(1), 59. doi:10.2307/2685263

[18] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Regression Trees. *Classification And Regression Trees,* 216-265. doi:10.1201/9781315139470-8

[19] FreeCodeCamp.org. (2018, October 08). Machine learning: An introduction to mean squared error and regression lines. Retrieved August 12, 2019 from https://www.freecodecamp.org/news/machine-learning-mean-squared-error-regression-line-c7dde9a26b93/

[20] Decision Trees in R. (2018). Retrieved August 12, 2019 from https://www.datacamp.com/community/tutorials/decision-trees-R

[21] Ho, T. K. (1995). Random Decision Forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition 1(1), 278*. doi:10.1109/icdar.1995.598994

[22] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics, 29(5).* 1189-1232

[23] Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via Random Forests. *Expert Systems with Applications, 42*(10), 4621-4631. doi:10.1016/j.eswa.2015.02.001

[24] Aggelen, A.. (2018). Linear Regression and Gradient Descent. Retrieved August 12, 2019 from https://blog.arinti.be/linear-regression-and-gradient-descent-89938783cb59