Group Project Report

6-033-11A Data Mining Techniques
Semester: Fall 2018

BY

Quan Hao	11248609
Gabriel Lainesse	11189782
Chaoyang Zheng	11249259

Presented

to

Jian Tang

Due date: 7/12/2018

Future Trip Volume Prediction

Quan Hao 11248609
Gabriel Lainesse 11189782
Chaoyang Zheng 11249259
Course: Data Mining Techniques

Abstract

Currently, more than 1,000 cities around the world have established or plan to establish a bike sharing system, spreading from campus, subway stations and residential areas to commercial center, grand parks and organizations. As a great application of sharing economy and the first large-scale bike sharing system in North America, Bixi Montréal plays an essential role in promoting travel flexibility and transportation efficiency, while also positively impacting the environment and quality of life.

The primary goal of this project is to establish accurate models that will predict the number of bikes required in different areas of the city at different points in time, as the way to anticipate demand. To do so, we take into consideration a number of features, such as weather conditions, weekday daytime and fuel price, through which pertinent suggestions about the demand could be provided to Bixi service to help them increase the operational efficiency. We apply and compare results obtained from linear regression and random forest models. Our result shows that the random forest with staring neighborhood gave us the best result.

I. Introduction

In this paper, by using several features, such as weather conditions, weekday daytime, statutory holiday and festival data, city of Montréal geographical feature data and fuel price, we proposed a regression model to predict the number of Bixi bikes demand levels in different areas of Montréal (such as Outremont, Ville-Marie, and Le Sud-Ouest, etc.). The model was trained with Bixi trip data, which contains details of the trips via the BIXI Montréal self-service bike network in 2017 [1], resulting in the processing of approximately 4,740,357 records.

The ability to accurately predict the count of Bixi trips in the different neighbourhoods on future data could bring abundant of valuable insights. Firstly, the standard issue observed within bike-sharing systems are the imbalances in the distribution of bikes among stations., Hence understanding the fluctuating demand demanding levels of Bixi bicycles could provide insights to Bixi on how many bicycles they should provide in each region depending on different conditions (i.e. weather conditions, weekday, and temperature, etc.), and it, therefore, enables them to arrange the bicycle distribution across Montréal better, and increase bike usage, efficiency and possibly membership count.

Moreover, our study enables city officials to gain a better understanding of the number of Bixi bikes used in each region, which provides a compelling blueprint for future work into this type of model as a tool for traffic analysis and infrastructure planning.

Finally, there is no doubt that this paper could be convincible evidence to motivate the government to invest more in either Bixi or the sharing economy, and the analytical models that can be built around them, which is consistent with the concept of sustainable development and healthy living habit.

II.Related work

In the last decade, the increasing number of European and North American cities that established bike-sharing systems has triggered more and more people in the academic community to focus on this topic of research, accompanied by emerging literature that analyzed the system through different perspectives.

Using time series models to engage state predictions is one of the most popular types of in this field. One of the representative research made by Borgnat et al. [3] focuses on analyzing what are the dynamics of movements in Lyon at various hours of the day by using the trip data of the *Vélo'v* program (a sharing bicycle system). The paper by Jensen et al. [12] analyzed the travel speeds of bikers using the Lyon bike sharing system and concluded that bicycles outstrip cars in downtown Lyon regarding speed.

Another hot type of research in this field makes use of data mining techniques, such as factor analysis and clustering, to search the underlying patterns and correlations within the bike-sharing data. For example, in the research by Froehlich et al. [5], data from Barcelona's shared bicycling system called *Bicing* was analyzed to uncover patterns of human behaviour, such as daily routines and cultural influences. Additionally, research by Vogel et al. [13] derived bike activity patterns by introducing extensive operational data from bike-sharing systems.

Further literature focused on improving the efficiency of bike sharing systems. For example, the paper written by Nair et al. [11] developed critical insights on the efficiency and functioning of a particular type of biking system, which allows passengers to lend and return the bikes at any place (i.e. without fixed docks). Moreover, the research made by Raviv et al. [8] sought to address these asymmetries by optimizing bike repositioning operations.

In the context of the city of Montréal bike sharing system, the paper by Wenger et al [16] attempted to predict the number of bicycle rides on each one of ten different streets in Montréal in a given day, introducing a couple of features such as the day of the year, the day of the week, weather, air pollution, holidays, festivals, hockey and football games. This paper differentiates from the previous research in different ways. Firstly, we are focusing on regional (among each district) differences of bicycle use, while the recent research was engaged under a biking lane perspective (i.e. comparing usage in 10 different lanes).

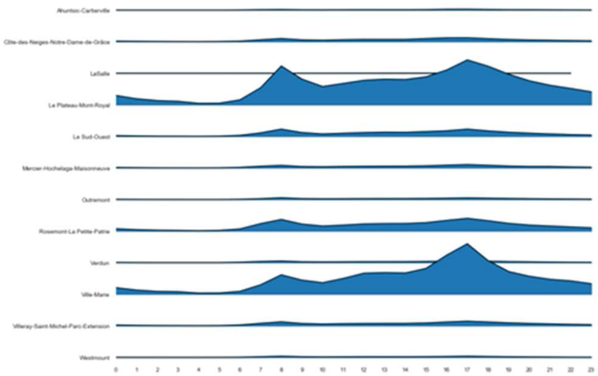
Moreover, based on the regional differences, our project introduced both time series data (DateTime data) and regional data., Therefore we can predict the count of trips in each given neighbourhood at a specific time (the hour of the day, the day of the week, the month of the year, etc.). Furthermore, compared with previous research which mostly engages in short-term predictions, our models could be applied in longer-term predictions, as our features are generalized (i.e. weather conditions and weekday). Of course, retraining the model with the newest data would be paramount to keeping our models accurate.

III. Problem Definition and Data description

This project aims at predicting the number of future Bixi bicycle and demanding level in a given region. To perform this task, we have used multiple datasets, which are presented in this section.

A: Bixi dataset

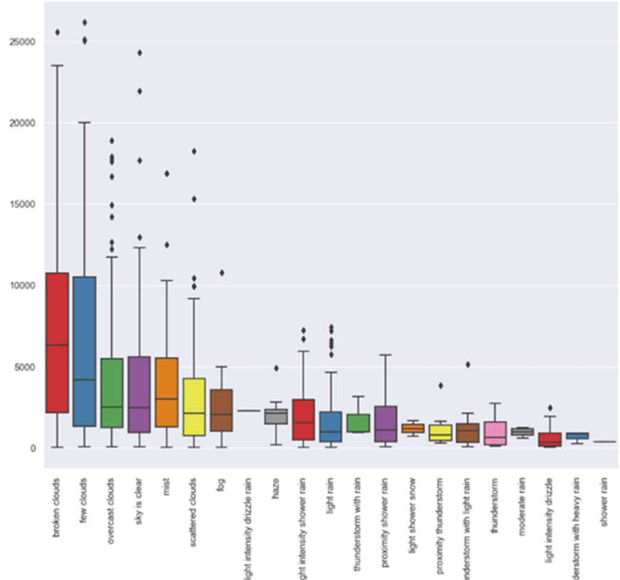
The first dataset is made of two parts: the Bixi trip history from 2014 to 2018 (only including data from April to November for each year since Bixi only operates during this time of the year) and the list of Bixi stations. Both datasets were obtained from the Bixi open data web page [1][2]. The trip history contains one record for each trip, including the station IDs for the start and end stations, the timestamp at the start and the end of the trip, the duration of the trip and whether the user was a Bixi member or not. The Bixi stations data includes coordinates (latitude and longitude) and the name of each station.



Graph1: Ridgeline Plot of Trip Count per Hour of the Day per Neighborhood

B: Weather dataset for Montréal

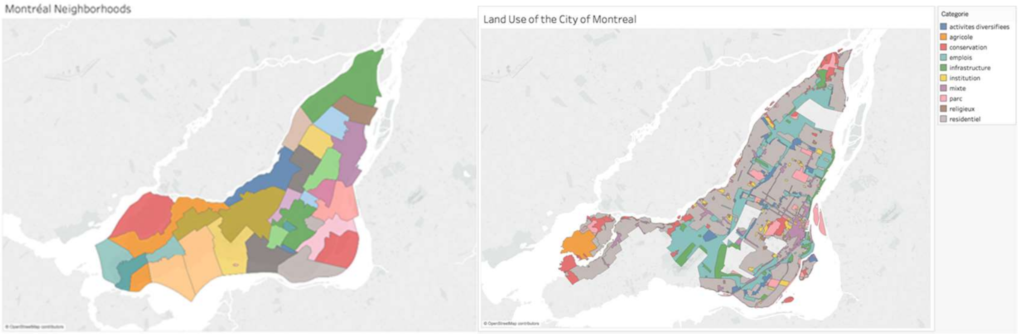
The second dataset consists of historical hourly weather data for Montréal. We had the intuition that weather patterns would be efficient in predicting Bixi rental usage, as rain should discourage most people from getting onto a bike. We first settled with data downloaded from Weatherstats.ca [15], which is a site that aims to archive data published from Environment and Climate Change Canada. This dataset included exciting metrics, such as the cloud cover. However, it did prove incomplete regarding actual weather observed (such as "light rain" or "thunderstorm"). We had to infer these ourselves from the numerical data (e.g. "cloud cover" and "precipitation amount"). To resolve this problem, we found on Kaggle a second historical weather dataset which included precisely what we were missing: descriptions of observed weather. Even though data from 2018 was missing from this new dataset, the quality of data and its usefulness convinced us to use this one instead.



Graph2: Box plot for the number of Bixi trips in Montréal under each weather condition in a given day

C: City of Montréal geographical features dataset

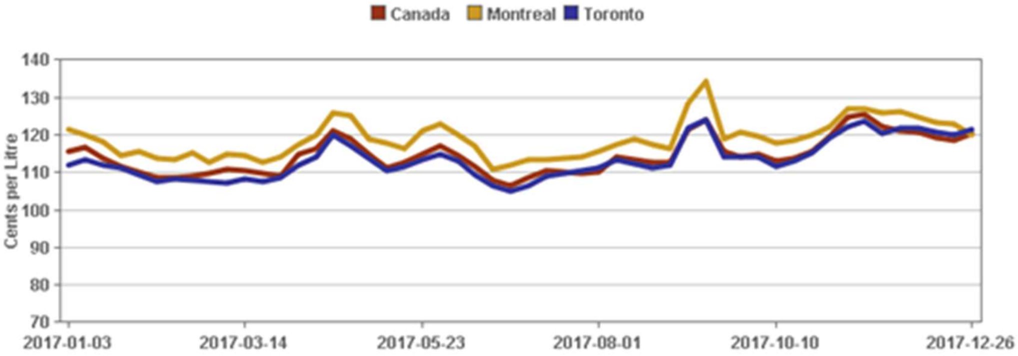
The third dataset we found consists of geographical features from the City of Montréal Open Data portal. The city made available multiple JSON and GeoJSON files that allowed us to perform feature engineering on Bixi stations (more on this later). The datasets we ended up using contain data on "Affectations" (Land Use) (residential, mixed, diversified, employment, institutional, etc.) [8] and on neighborhoods [9].



Graph3 : Montréal neighborhood and landusued information

D: Fuel price dataset

The fourth dataset used for our project is merely the weekly history of Toronto fuel prices, taken from the Ontario Government website. [6] Because we could not get the daily price in Montréal or for the Province of Québec, we decided to use oil prices for the Toronto area instead. When we compare the price of natural gas in Montréal over the past few years, we find that Montréal is generally more expensive than Toronto, but the absolute value was not of much value to us. We need to study the trend of oil prices to predict the potential relationship between oil prices and Bixi, which is why this dataset was selected.



Graph4:: Average retail price for for regular gasoline in 2017

E: Statutory Holidays and Festival dataset

Finally, the last dataset that we ended up using is a list of statutory holidays observed in the province of Quebec, along with dates for the five most prominent events that occur in Montréal during the summer, namely: the *Just For Laughs Comedy Festival*, the *Montréal International Jazz Festival*, the *L'International des Feux Loto-Québec* fireworks event, the *Montréal F1 Grand Prix*, and the *Osheaga* music festival[10].

IV.Feature Engineering & Data Pre-Processing

We have performed multiple feature engineering tasks in order to add features to the dataset we wanted to use to build our models. The limited content of the Bixi trip history data justified this quest for increased dimensionality.

The first type of feature engineering we did consisted of mapping stations to neighbourhoods, land use, and great parks. To do so, we first read the polygons contained in the JSON files from the City of Montréal using the *geopandas* library and used the *contains()* function from the *shapely* library. This function returns *True* when a given coordinate is found within the boundaries set by a given polygon. For each station, we iterated over all polygons contained in the various JSON files and retrieved the name of the first polygon which contained that station. The features we obtained from this process include the land use around the Bixi station, the neighbourhood that the station is located in, and if the station is located within a great park.

The second type of feature engineering we did consist of transforming trip datetime values into dimensional features, such as day of the week (Monday, Tuesday, etc.), week of the year, month, hour of the day, and period of the day (categories for hour of the day, such as "early morning", "noon", etc.).

The third type of feature engineering we did involve the mapping of other features using the start date from the individual trips. These features include weather values (temperature, humidity and pressure levels, as well as wind speed), descriptions of weather observed, current fuel prices and whether or not there was a festival or a statutory holiday on a given day.

We also divided the route affectation model into two different groups. For the shorter route affectation feature, we only recorded the land use surrounding the starting station. For the longer route affectation, we recorded both the land use around the starting station and the ending station and combined them.

VI. Algorithm selection and optimization

To predict the Bixi future trip volume, two regression models, a multiple linear regression and a random forest regression were developed. For shorter route affectation, we record both the starting station and ending station. Before we started to build these models, we calculated the number of rows and the mean of the count of trips for each group of features, in order to get a sense of the quality of the grouping and to make sure our target variable had meaning on its own. The count of trips is 75,986 for the starting neighbourhood models, 135,384 for long route affectation models and 120,127 for the short route affectation models. The average trip count per group is 62.88 for the starting neighbourhood models, 35.014 for the long route affectation models and 39.46 for the short affectation models. The modelling and optimization based on validation data could be seen as follows:

A. Optimizing Feature Representation

Because it does not make sense to directly convert categorical data into continuous variable [i.e. Using 1 to represent for Ville-Marie (neighbourhood) and 2 for Outremont], we used one hot encoding to convert all of our categorical variables into binary variables. For example, we chose to use 7 binary features to represent the day of the week., If a trip happened on Monday, the feature vector for the day of the week would look like this: [1,0,0,0,0,0,0]. One hot encoding brings more flexibility and is more accurate in building the regression model.

B. Multiple linear regression

The first model we implemented is the multiple linear regression model. We attempted to model the relationship between variables and the count of trips (which is our response variable) by fitting a linear function with the observed data. It can be represented by the formula below:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \epsilon.$$

The training dataset was randomly selected to represent 70% of the total dataset. For multiple linear regression models, we got scores of 0.4517 with the starting neighborhood model, 0.46629 with long route affectation model and 0.3349 for the count of trips with short route affectation model.

C.Random forest regression model for route affectation

Random forests can be seen as the aggregation of decision trees which are not correlated among each other. We used a forest with 300 decision to avoid the issue of overfitting, while still testing forests with 10, 20 and 50 decision trees. Results appeared to converge as early as with 10 trees only. We finally set the estimator as 300 and got a score of 0.83554 with starting neighbourhood model, 0.8253 with the shorter route affectation model, and 0.7032 with the longer route affectation model.

Prediction

We were successful in making predictions using our models. Examples of predictions can be seen in the table below:

Month	Day of the Week	Hour	Festival?	Holiday?	Avg. Temp °C	Avg. Wind Speed (km/h)	Avg. Pressure (kPa)	Avg. Humidity (%)	Fuel Price	Weather Condition	Neighbourhood/ Affectation	Predicted Usage
May	Sunday	8 AM	No	No	15	5	1000	75	120	Clear Skies	Ahuntsic Cartierville	3.15
July	Sat.	12 PM	Yes	No	25	5	1000	75	120	Clear Skies	Ville-Marie	192.15
Sept.	Mon.	8 AM	No	No	15	5	1000	75	120	Overcast	Residential Mixed	376.10
Sept.	Mon.	8 AM	No	No	15	5	1000	75	120	Overcast	Residentia Residential	150.96

Graph5 : Predictions results

It would be reasonable to predict that there would be low demand for bikes on a Sunday morning of May, in the Ahuntsic-Cartierville neighbourhood (which is far from the centre of the city of Montréal). It would also be reasonable to predict that there would be high demand for bikes on a Saturday, the day of a festival, at 12 PM in downtown Montréal (Ville-Marie neighbourhood). For both cases, this is what we observe in our predicted values of 3.15 bikes and 192.15 bikes.

Another example was made to compare demand on the same Monday morning of September for trips going from residential areas to work areas (mixed, in this example) (which would be "commute" trips) against trips going from residential areas, but also ending in residential areas ("non-commute" trips). We expected the demand to be higher for commute trips than for non-commute trips at that time of the day. This is the result we have obtained, with 376.10 bikes needed for commute trips throughout Montréal at 8 AM opposed to 150.96 for non-commute trips.

Conclusion

Due to non-linear features, we failed to predict the count of trips using the multiple linear regression model. The multiple linear regression models we used to have a lousy representation and none of the score is over 0.5. Random forests seem to be the best fit for our research question. The best score we got is from the random forest with starting neighbourhoods, which was 0.8355.

VII. Result summary

A. Algorithm comparison

To test the performance of models, we used R-square and the Mean Square Error (MSE) as our approach to measuring how well our models were performing.

The value of R-squared (between 0 and 100%) represents how much of the total variation of the dependent variable can be explained by the independent variables from the models. In this case, we use it to measure the fitness of our models on the testing dataset.

The Mean Square Error represents the squared difference between our model's predictions and the ground truth. The lower it is, the better.

Performance metrics are detailed in the table below. All performance metrics were calculated on the test dataset, using the train-test split methodology for a 70% train / 30% test split of the data.

	Linear Regression Neighbourhood	**Linear Regression Affectation (Long)**	**Linear Regression Affectation (Short)**	**Random Forest Neighbourhood**	**Random Forest Affectation (Long)**	**Random Forest Affectation (Short)**
R2	0.4517	0.8298	0.4662	0.6963	0.3349	0.8129
MSE	9416.46	2922.17	3620.59	2060.18	5096.96	1433.78

Graph6: Performance metrics

We can see in the table above that linear regression models do not perform well, while random forest models do a much better job. The MSE is much lower for the random forest neighbourhood model than the random forest affectation model, probably because the long form of the land affectation attribute was not a good predictor of Bixi traffic. However, the MSE is still lower with the short form of the land affectation attribute, indicating that the broader categories were better at predicting traffic, even more so than neighbourhoods (2060.18 vs 1433.78).

VIII. Discussion and Conclusion (What we have learned through the project)

A. Working with a big dataset

We first started to build the model based on four years. The dataset became too big to handle once we added all the features (i.e. the weather conditions, fuel price, neighbourhood, route etc.), as around 9 Gigabytes were required to store the data, which took 15 minutes or more to read from the disk at the beginning. Initially, we tried to randomly divide the dataset into several small samples (each one being no more than 200 MB), and we would train models using these samples. However, the performance of the models against the test dataset was always wrong, and so this approach was deemed inefficient.

After deep consideration and searching online, we ended up storing our dataset in the HDF format, which is designed to store vast amounts of data. This new approach allows for reading the data from the disk in only 30 or 40 seconds.

Training the models using the whole dataset proved to be slow as well. We have ended up building models using only the data from 2017. Once our models are tested and deemed worthy of further use, we could retrain them using data from the previous (and later) years as well.

B.Treating categorical data carefully during modelling

One hot encoding should be applied when no ordinal relationship among categorical variables exists. We first implemented the integer encoding and allowed the model to assume a natural ordering between categories, but we would end up with unexpected results. To fix this, we removed the integer encoded variables and performed one-hot encoding of categorical variables, which provided us with a lot of binary variables to train the models with. With that approach, the performances of our multiple linear regression models and our random forest models were significantly improved.

C. Model comparison

In our research of Bixi usage, we found that random forest models outperformed the multiple linear regression models. We tried different hyperparameters to avoid the issue of overfitting, such as the tree depth limit and the number of trees within a forest. However, the score we got each time were converging on the same score and MSE as obtained with the 300 trees forest, which gave us confidence in our final models.

Appendix:

[1] Aubert Sigouin, BIXI Montréal (public bicycle sharing system)
https://www.kaggle.com/aubertsigouin/biximtl

[2] Bixi Open Data. URL: *https://www.bixi.com/en/open-data*

[3] Borgnat, Pierre, et al. "Shared Bicycles in a City: A Signal Processing and Data Analysis Perspective." 26 2010. Scientific Commons. 1 February 2010 <http://www.scientificcommons.org/58104633>

[4]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[5]Froehlich, J., J. Neumann and N. Oliver. "Measuring the pulse of the city through shared bicycle programs." International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems- UrbanSense08. 2008.

[6] Government of Ontario – Fuels price survey information. URL : *https://www.ontario.ca/data/fuels-price-survey-information*

[7] Kaggle - Historical Hourly Weather Data. URL : *https://www.kaggle.com/selfishgene/historical-hourly-weather-data*

[8] Montréal Données Ouvertes – Affectation du Sol. URL : *http://donnees.ville.montreal.qc.ca/dataset/affectation-du-sol*

[9] Montréal Données Ouvertes – Arrondissements. URL : *http://donnees.ville.montreal.qc.ca/dataset/polygones-arrondissements*

[10]Montréal Festival and events for everyone. URL:https://www.mtl.org/en/what-to-do/festivals-and-events

[11]Nair, Rahul & Miller-Hooks, Elise & Hampshire, Robert & Busic, Ana. (2012). Large-Scale Vehicle Sharing Systems: Analysis of Vélib'. International Journal of Sustainable Transportation - INT J SUSTAIN TRANSP. 7. 10.1080/15568318.2012.660115.

[12]Pablo Jensen, Jean-Baptiste Rouquier, Nicolas Ovtracht, Céline Robardet. Characterizing the speed and paths of shared bicycles in Lyon. Transportation Research Part D: Transport and Environment, Elsevier, 2010, 15 (8), pp.522-524. <10.1016/j.trd.2010.07.002>. <hal-00541307>

[13]Patrick Vogel, Torsten Greiser, Dirk Christian Mattfeld, Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns, Procedia - Social and Behavioral Sciences, Volume 20, 2011, Pages 514-523, ISSN 1877-0428.

[14]Rahul Nair, Elise Miller-Hooks, Robert C. Hampshire & Ana Bušić (2013) Large-Scale Vehicle Sharing Systems: Analysis of Vélib', International Journal of Sustainable Transportation, 7:1, 85-106, DOI: 10.1080/15568318.2012.660115

[15] Weatherstats.ca. URL : *https://montreal.weatherstats.ca/download.html*

[16] Wenger Robert, Haomin Zheng, Stefan Dimitrov Robert, Biking Lane Usage Prediction URL: http://rl.cs.mcgill.ca/comp598/fall2014/comp598_submission_102.pdf