Predicting the demand of a bike sharing system

A binary classification of the volume of demand of BIXI stations

Team ZMS

Maryam El Arfaoui
maryam.el-arfaoui@polymtl.ca
1939476

Sofien Ben Ayed
sofien.ben-ayed@polymtl.ca
1971009

Zhuofei Kang
zhuofei.kang@polymtl.ca
1939634

***Abstract*—A bike sharing system as Bixi can benefit from an accurate prediction of the demand in bikes at a specific station and at specific time, enabling a faster recovering of the network equilibrium. Using data about the weather and the Bixi stations, we implement a binary classification of the volume of the demand through 3 techniques : logistic regression, random forest and multi-layer perceptron. The evaluation is done through the calculation of the F1-score as precision and recall are as important for our task. The multi-layer perceptron outperformed the other models and approaching the problem as a regression one, thus predicting the number of withdrawals before concluding about the volume lead to better results than a direct binary classification.**

***Index Terms*—binary classification, bike sharing, logistic regression, multi-layer perceptron**

I. Introduction

Bike sharing system is popular in many cities all over the world where people can rent a bike in a station to go to the other station of their desired destination. However, the balance of bike sharing is a problem since people tend to go to the same station at the same period. For example, people are more likely to go to Downtown. Therefore, the operators should take much time to recover the network equilibrium. Thus, the operators could use machine learning tools to help them keep the network balanced. In this paper, we tested several models to help operators address this problem. For this project, we are given a dataset, including for each date in 2015 and 2016, the temperature (C), the drew point (C), the relative humidity, the wind direction (10s deg), the wind speed (km/h), the visibility (km), the visility indicator, the pressure at the station (kPa), the hmdx, the wind chill, a brief description for the weather, the indication of public holiday, the station code and of course the target feature wich is the number of withdrawals. Based on these information, we develop a machine learning approach to predict the volume of the demand : high if the number of withdrawals is at least 8 and low if it is not.

In our project, in order to develop a machine learning model, at first, we had to preprocess the data, impute missing values, combine features to make it more meaningful, add other known information about Bixi stations and select the most important features for our model. Only then, we try four machine learning models based on the datasets : logistic regression, random Forest, multi-Layer perceptron classifier and multi-layer perceptron regressor. Finally, we evaluate these models using the F1-score metric. We've found that the multi-layer perceptron outperformed the other models and approaching the problem as a regression one first worked better than immediately trying to classify the volume (binary). Our final score on the test set of Kaggle is 0.57625.

In our report, we first describe briefly our exploration of the data, as well as our choices for the preprocessing of the data, and then show how we select the features for the model. After that, we present four prediction models and the decisions we have taken about training/validation split, our strategy to deal with the unbalancement of the data and the tuning of the hyper-parameters for the best model. We finally present the experimental results and discuss about what can still be improved.

II. Exploring the data

Before preprocessing the data, we must know exactly what is in it. In the dataset, we could find that not all the data has been included, while there are some missing values in some data columns. In the TableI, we show the number of recordings for every feature, where we could find that the total number of recordings is 1546454. There are some missing values in Temperature, Drew point, Relativite humidity, wind direction, Wind speed, Visibility,Pressure at the station,hmdx, Wind Chill and Weather, and there is no data in Visility indicator. Meanwhile, considering the limited data of Wind Chill (only 9646, 0.62%) and the non-existant data for the Visility Indicator, these features won't be considered in our model. In fact, we recalculated Wind Chill using the temperature and the wind speed (discussed in the preprocessing section). Also, there is less than 0.7% of the recordings missing all the features related to the weather. Finally, the Weather feature is missing for 40% recordings.

In exploring the data, the Fig 1 shows the total number of Withdrawals by hour, we can find that there are two peaks in 8:00am and 5:00pm, because these are the times where people move between places (going or coming back from work, university...) and are more likely to rent a Bixi bike. We could also find out that at 4:00pm and 6:00pm, the number of withdrawals is much higher than other time periods. The total number reaches the minimum at 4:00 am.

It was important for us also to visualise the variation of withdrawals by days of the week as shown in Fig 2, and

Data columns	Number of data
Date/Hour	1546454
Temperature (C)	1545362
Drew point (C)	1545362
Relativite humidity	1545362
wind direction (10s deg)	1544270
Wind speed (km/h)	1545362
Visibility (km)	1545362
Visility indicator	0
Pressure at the station (kPa)	1545362
hmdx	495768
Wind Chill	9646
Weather	1545362
Public Holiday	1546454
Station Code	1546454
Withdrawals	1546454
Volume	1546454

TABLE I
THE NUMBER OF DATA FOR EACH FEATURE

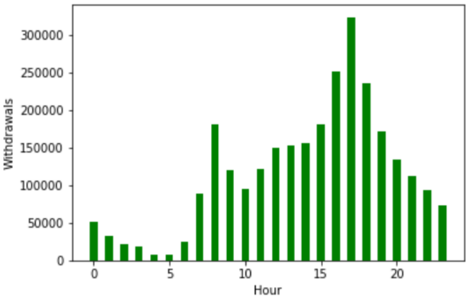

Fig. 1. Total number of Withdrawals by hour

by months as shown in Fig 3. The variation between the days of the week is not great, but it remains interesting to say that the demand in the middle of the week increases compared to the beginning and the weekend. However, the variation between months is remarkable, the total number of withdrawals between May and August reaches a maximum, which is quite normal since it is the period when the weather is nice in Canada, unlike other months in which the weather is unpleasant. We do not forget to mention that December, January, February and March are not represented on the graph because it's the time of winter in Quebec, so Bixi company stops its activity.

we checked the distribution of each feature in the seek of extreme values that could be outliers but fortunately none was found.

III. PREPROCESSING THE DATA

Before developing machine learning models, we should first preprocess the data and we start by filling the missing values that were mentioned in the previous section. Then, in order to avoid a useless complexity, some features and attributes of features were combined. It was important also to transform some types of data to numerical values and apply a normalisation as this can have a significant impact on how the models perform. Using another dataset from Bixi, we accessed the latitude and longitude of Bixi stations and worked on it to extract the most important knowledge for our models.

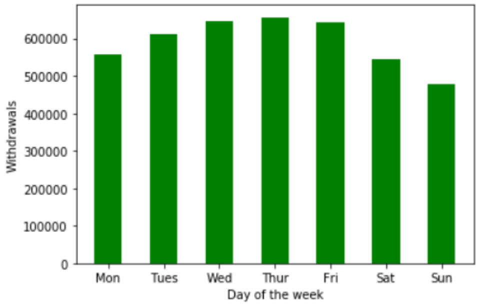

Fig. 2. Total number of Withdrawals by day of the week

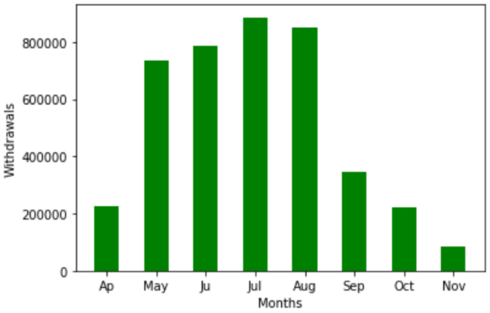

Fig. 3. Total number of Withdrawals by month

A. Hour and date

In order to lead to better computation of similarities, we transform the hour and date to polar coordinates, because it could help more easily understand the continuity of the original time and the continuity of the head and tail. Following are the equations that show the transformations we did to deal with circularity.
$Month\_Cos = cos(month * \frac{2*\Pi}{12})$. month goes from 0 to 11.
$Month\_Sin = sin(month * \frac{2*\Pi}{12})$
$Hour\_Cos = cos(hour * \frac{2*\Pi}{24})$. hour goes from 0 to 23.
$Hour\_Sin = sin(hour * \frac{2*\Pi}{24})$
Fig 4 shows the hour and date distribution after transformation.

B. Weather

In the data column of weather, we find there are 15 distinct categories. Considering the similarity of some classes, we gather the weather in only 7 categories, Table II shows the categories of classification. If the value of the weather for an entry is "ND" which means that the value is missing, we look for the nearest entry in time with a non-missing value for weather and impute by this value. We do not proceed if the nearest entry is separated by more than 2 hours.

C. Combining Temperature, Drew Point and Wind Speed

Meteorologists have created different indexes in order to grasp the feeling of hot and cold. These indexes can be more informative than using raws temperatures. We calculate two of them :

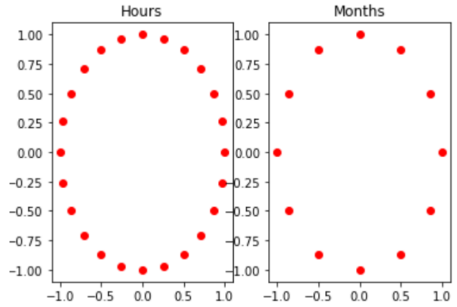

Fig. 4. Hour and date in polar coordinates

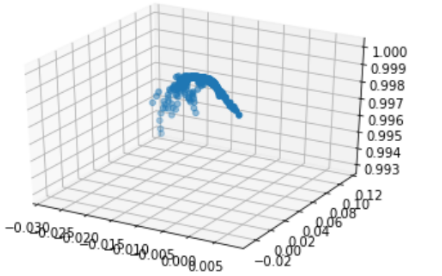

Fig. 5. Cartesian coordinates of position of stations

categories	subcategory
Nuageux	Nuageux, Gnralement Nuageux
Dgag	Dgag, Gnralement Dgag
Pluie	Pluie, Averses de pluie, Pluie modre, Averses de pluie modre
Pluie forte	Pluie forte, Averses de pluie forte
Brouillard	Brouillard, Bruine
Orages	Orages
Neige	Neige, Averses de neige

TABLE II
CATEGORIES OF WEATHER

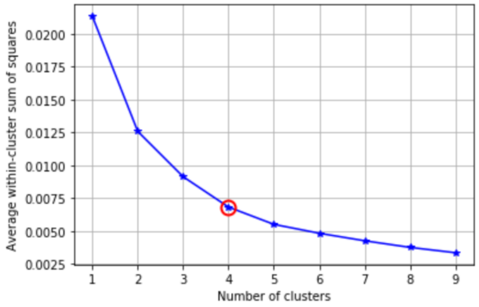

Fig. 6. elbow curve of K-means clustering

- The humidex which is an index used by Canadian meteorologists [1] [2] to describe how hot the weather feels to the average person. It is calculated using the air temperature and the drew point.

$$H = T_{air} + 0.555[6.11e^{5417.7530(\frac{1}{273.16} - \frac{1}{273.15+T_{dew}})} - 10]$$

- The windchill is the lowering of body temperature due to the passing-flow of lower-temperature air. It is calculated [1] [2] using the air temperature and the wind speed.

$$T_{wc} = 13.12 + 0.6215T_a - 11.37\nu^{+0.16} + 0.3965T_a\nu^{+0.16}$$

These indexes replace Temperature and Drew Point. In fact, these are recalculations as the dataset doesn't include enough values for these features.

D. The position of BIXI stations

In order to give the model more significant information, and because we thought that the position of Bixi stations is a relevant data as two stations situated near each other should experience a similar demand, we used the trip history in the open data [3] provided by Bixi company, where we could access the code, name, latitude and longitude of stations. Latitude and longitude are not taken as they are. We first transform them to get Cartesian coordinates :

$$x = \mathbf{R}.cos(latitude).sin(longitude)$$

$$y = \mathbf{R}.cos(latitude)$$

$$z = \mathbf{R}.sin(latitude)$$

where **R** should be the Earth radius but we use $\mathbf{R} = 1$ as we do not need the real coordinates but just the positions between the stations. Fig 5 shows the result of transformation in 3D coordinates.

Injecting these coordinates directly did not improve the quality of the models. Thus, we tried instead to identify some 'regions' in Montreal. If two stations are in the same region, there is a higher chance of having similar demand. Therefore, we make clusters of BIXI stations using their positions. When we develop the machine learning model, the "regions" (clusters) will be added as a new feature. For that, we use K-means on the Cartesian coordinates. We try different number of clusters and select the best one to apply. Fig 6 shows the Average within-cluster sum of squares by number of clusters. Considering the degree of declining, we select 4 clusters as the best choice. Then, the positions of BIXI stations will have a feature 'cluster', in the range of 0 to 3.

E. Features selection

To reduce the dimension of the data, we chose to train the models with less features than provided and reduce the number of values of some of them, to do so, we had to:

- Remove features : the features wind direction (10s deg) and Pressure at the station (kPa) were removed because they were irrelevant for the rent of bicycles, and it was verified because when we trained the model using these features and without them, nothing changed so we decided that it is better to remove them. We also removed Visility indicator as it contains many missing values (more than 99%).

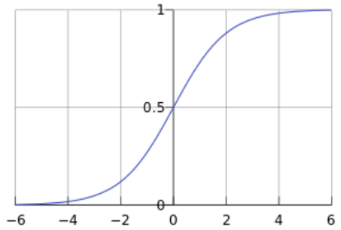

Fig. 7. sigmoid function

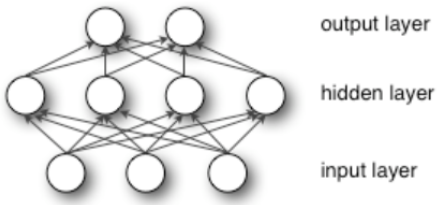

Fig. 8. Structure of Multi Layer Perceptron

- Calculate features: As they were almost empty, Wind Chill and hmdx features were calculated by combining other features as explained before, then they were normalised in addition to other features (Wind speed (km/h), Visibility (km), Relative humidity (%)) in which the missing values were imputed by the mean value. Finally, the features Year, Station Code, Weather (with only 7 categories) and Cluster were One-Hot encoded.

IV. Machine learning models

In our project, we try four popular machine learning models to predict the volume of BIXI stations. Here two paths were taken in order to tackle the problem :

- Binary classification - Direct prediction of the volume : low or high. For binary classification, three algorithms were tested: 1)Logistic Regression, 2)Random Forest, 3)Multi-Layer Perceptron.
- Regression - Prediction of the value of withdrawals and then conclusion about the volume using a threshold. Regression was only tested with Multi-Layer Perceptron as it obtained far better results than the others.

A. Logistic Regression

Logistic Regression [4] is a machine learning method used to solve a two-class (0 or 1) problem. Logistic regression is supported by linear regression theory, but logistic regression introduces nonlinear factors through the Sigmoid function, so it is easy to deal with the 0/1 classification problem. The logistic function is an S-shaped curve which maps the probability in range in 0 and 1 while not include 0 and 1. The figure 7 shows the curve of the Sigmoid function. In the following equation, $\mathbf{b}$ and $\mathbf{W}$ are the parameters we want to learn in the model.

$$\mathbf{y} = sigmoid(\mathbf{W}^T\mathbf{x} + \mathbf{b}) \tag{1}$$

where

$$sigmoid(x) = \frac{1}{1 + \exp(-x)} \tag{2}$$

Where $\mathbf{x}$ is input data and $\mathbf{y}$ is the input data's class probability vector. When the class probability $\mathbf{y}$ is greater than threshold, the output will be 1, otherwise it will be 0.

B. Random Forest

Random Forest [5] is a method based on decision tree. ”Forest” means a set of trees, ”Random” represents the construction of every decision tree in the random forest carries random sampling with bootstrap, as well as random sampling of features. The steps of establish random forest are following:

- 1) If there are N samples, there are N samples randomly selected (each time one sample is randomly selected, and then returned to continue selection). These selected N samples are used to train a decision tree as a sample at the root of the decision tree.
- 2) When each sample has M attributes, when each node of the decision tree needs to be split, m attributes are randomly selected from the M attributes, satisfying the condition m $\ll$ M. Then some strategy is used (such as information gain) from these m attributes to select one attribute as the split attribute of the node.
- 3) During the decision tree formation process, each node must be split according to step 2 until it can't be split again. In order to avoid overfitting, we set a maximum depth determined through experiment.
- 4) Create a large number of decision trees according to steps 1 to 3, which constitutes a random forest.

C. Multi-Layer Perceptron

In machine learning, the perceptron [6] is a two-category linear classification model and belongs to the supervised learning algorithm. Enter the feature vector for the instance and the output as the category of the instance (take +1 and -1). The perceptron corresponds to a separate hyperplane that divides the instances into two categories in the input space. Multi Layer Perceptron (MLP) [7] represents the perceptron which includes at least one hidden layer (except for one input layer and one output layer). Single-layer perceptrons can only learn linear functions, while multi-layer perceptrons can also learn nonlinear functions. The layers in Multi Layer Perceptron are all fully connected layer. Fig 8 shows a simple structure of Multi Layer Perceptron. In our case, the activation function chosen in Multi Layer Perceptron is ReLU. Compared to other activation functions, ReLU has the following advantages: for linear functions, ReLU is more expressive, especially in deep networks; for nonlinear functions, ReLU has a constant gradient due to non-negative intervals, so there is no vanishing gradient problem, which keeps the convergence speed of the model at a stable state. The equation of ReLU is:

$$ReLU(x) = \begin{cases} x & if x > 0 \\ 0 & if x \leq 0 \end{cases}$$

The multi-layer perceptron was originally designed for classification, but it can also handle regression problems, as long as the softmax and other classifiers are changed to sigmoid regression. We tried both with MLP : classification and regression.

V. Experiments

A. Data split

80% of the data was used for training and 20% for validation. As the classes (Volume) are unbalanced (ratio 1:14), we use a stratified split. Moreover, we upsample the minority class (1) for better predictions of our models.

B. The impact of data preparation and selection

The decisions we made to increase the performance of the models we used was based on experimental results:

- preprocessing and features selection: In order to see their impact, we did the test on logistic regression and MLP, When using these techniques, the results (F1 score) of logistic regression pass from 0,46 to 0,49 and those of the Multi-Layer Perceptron pass from 0,56 to 0,59. And hat was enough to prove the positive impact of preprocessing and features selection. It is important to mntion also that the circular way of encoding month and hour only works well with MLP and the clustering was also just added for MLP.
- upsampling: Without upsampling the minority class, the results were terrible : logistic regression classified everything as the majority class while random forest gave 0.25

C. Experiment results

When we measure the result of prediction, we always choose F1 score as the metric. The equation of F1 score is:

$$F1score = 2 * \frac{precision \cdot recall}{precision + recall}$$

Where

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$TP : TruePositive$$

$$FP : FalsePositive$$

$$FN : FalseNegative$$

Table III shows the F1 score of Logistic Regression, Random Forest, Multi-Layer Perceptron Classifier and Multi-Layer Perceptron Regressor models. The following results were not obtained through the exact pipeline. Indeed, the circular way of encoding hour and month improved the performance of the multi-layer perceptron of 1.2% but significantly decreased the performance of logistic regression by 6%. For this model, the best performance was obtained through a regular One-Hot encoding of hour and month.

Models	**F1 score**
Logistic Regression	0.4861838596991516
Random Forest	0.5087069096816774
Multi-Layer Perceptron Classifier	0.5899244853625738
Multi-Layer Perceptron Regressor	0.606580576684946

TABLE III
F1 SCORE OF THE FOUR MODELS

According to the table III, it is clear that the Multi-Layer Perceptron regressor gives better results than the three classification models, this model has a considered benefit as it allows playing with the threshold in order to get better prediction results. It is important to mention that Through a set of experiments on various splits of the data, 8 was the threshold chosen for MLP regressor, and it's perfectly coherent as it is exactly the way a high volume is defined.

VI. Discussion

A. Pros of our approach

Before developing the machine learning models, we started by preprocessing the data . We have removed some less relevant data columns, such as wind direction and Pressure at the station. If we train the models based on these features, the results could not be improved while the computation cost increased. Meanwhile, taking advantage of the temporal aspect of our dataset to impute the large amount of missing values for the weather feature (40%) helped the models perform better. But this had to be done with caution. By only taking the nearest entry without fixing any limit of time between the two harmed the quality of the models. Indeed, weather can change within a few hours.

When we consider hour and date features, we have transformed hour and month to polar coordinates, where it could help more easily understand the continuity and periodic of the original time and the continuity of the head and tail. The circular way of encoding month and hour worked well only for the Multi-Layer Perceptron.

The Multi-Layer Perceptron regressor gives better results than the three classification models, which can be explained by the fact that we give the regressor more information in the training phase by feeding the model with exact numbers of withdrawals. Sometimes, A 0 (low) may correspond to a number of withdrawals of 7 while a 1 (high) corresponds to 9. By giving those information, it allows it to better discriminate with near border values.

B. Future work

When we train the machine learning models, we consider BIXI stations clusters as a feature, this could increase significantly the performance of the models. However, we think that a better way would be to train the models on each cluster and choose the model with higher performance for each model separately. An interesting feature of the station that we could add is the altitude as people surely tend to rent a bike more to go down than to go up.

VII. Conclusion

In our project, we have predicted the demand of bike sharing system based on Logistic Regression, Random Forest, Multi-Layer Perceptron Classifier and Multi-Layer Perceptron Regressor models. Before developing these models, we explored and preprocess the data to remove the not relevant features and fill in missing values. After that, we trained the models mentioned above and test them on the datasets. Appropriate preprocessing and features selection improve the results of machine learning models. We measured the results by F1 score metric, and we found out that Multi-Layer Perceptron Regressor performs better than other models with a F1 score up to 0.60658. And our model also has a good performance on the Kaggle testset, with a final score of 0.57625.

References

[1] A. Lebel, "Quest-ce que le refroidissement olien?" Societe Radio-Canada, 28 dcembre 2017.

[2] S. mtorologique du Canada, "Comment calcule-t-on l'humidex?" Environnement Canada.

[3] "Bixi open data," https://montreal.bixi.com/en/open-data.

[4] F. E. Harrell, "Ordinal logistic regression," in *Regression modeling strategies*. Springer, 2015, pp. 311–325.

[5] M. Pal, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[6] Y. Freund and R. E. Schapire, "Large margin classification using the perceptron algorithm," *Machine learning*, vol. 37, no. 3, pp. 277–296, 1999.

[7] M. Riedmiller, "Advanced supervised learning in multi-layer perceptronsfrom backpropagation to adaptive learning algorithms," *Computer Standards & Interfaces*, vol. 16, no. 3, pp. 265–278, 1994.