Understanding what might potentially affect rented bikes and cause a “rebalancing” problem in Pittsburgh bike sharing system.

Abdulaziz Almuzaini
School of Information Science
University of Pittsburgh
135 N Bellefield Ave
Pittsburgh, PA 15260
aaa169@pitt.edu

Dimple Varma
School of Information Science
University of Pittsburgh
135 N Bellefield Ave
Pittsburgh, PA 15260
ddv1@pitt.edu

Abhishek Mukherjee
School of Information Science
University of Pittsburgh
135 N Bellefield Ave
Pittsburgh, PA 15260
abm84@pitt.edu

ABSTRACT

In this paper, we aim to use data mining techniques on bike sharing system in Pittsburgh to examine two potential tasks. The first one is to predict the number of bikes rented each hour by people in a particular bike station and we will test what kind of attributes that might affect the response variables, which is the count of rented bikes. The flexibility of the system that allow riders to rent and return the bikes to any stations caused what so called "imbalance problem" and this leads us to figure out what might cause this problem. To do the first task, a linear regression model will be used and it will be evaluated after that and examine the accuracy. For the second tasks, which is a classification problem, we will use logistics regression, KNN, Naïve Byes, SVM and decision trees models to classify what stations that might have the potential imbalance problem and compare the performance of the models.

Keywords

Pittsburgh; Bikes; Rented Bikes; Imbalance; Bike Stations; Bike Sharing; Data Mining;

INTRODUCTION

The bike sharing systems have been emerged around 20 years ago and this was related to the advanced technologies and health styles improvements hype [2]. The system have been used in more than 500 cities around the world and the system provides its users the flexibility of renting a bike from any stations nearby and return it back either to the same station or to any other station [1]. The increase of the demand of using this bikes instead of public or private transportation is might related to the increasing population in the major cities which makes using of cars or buses undesirable. From costly parking lots to bus delays, users have the opportunity not just saving money but also a better healthy style alternative and an ecological approach [1]. But the flexibility of the system usually lead to what is so called "Imbalance Stations" which is caused when a user can't find a bike to rent or she is not able to return a bike to a specific station because there is no a spot for the returned bike [3]. Therefore, the organizers have to rebalance stations manually every day by moving bikes from one stations to another, which seems not a practical solution especially for crowded cities [1]. By solving this problem both the users and the administrators will benefit from such a powerful prediction system when a user knows in advance where a station is empty or not. Also the organizers will be able to determine in advance the potential imbalanced stations and perform perfect actions [1][4].

In this paper, we aim to predict the number of bikes rented each hour by people in a particular bike station in Pittsburgh. More specifically, we will use the 20 days data to train the models and then for the testing phase we are predicting the number of bikes rented during last 10 days of each month using information of only before that day.

The paper is organized as follows. Section 1 discusses the literature reviews of different related papers. Section 2 provides the data explanation and the exploratory data analysis in general. Section 3 we use different useful models to test our prediction and their accuracies. Section 4 concludes the paper.

1. Related Work

Predicting Bikeshare System Usage Up to One Day Ahead by Giot and Cherrier [2] devices an algorithm that able to predict the number of bikes that will be rented in the next 24 hours to determine the potential imbalanced stations. They have used multiple regression methods but the Ridge Regression and Adaboost Regression have the high accuracy related to other algorithms.

Prediction of Bike Sharing Systems for Casual and Registered Users by Alhusseini [1] use the Support Vector Model (SVM) to predict the number of rented bikes and also he used a classification models such as Softmax Regression and SVM to classify demanded stations into 5 categories.

Predicting Bike Usage for New York City’s Bike Sharing System [3] in this paper, the authors were only focusing on the morning hours in NYC. They used regression models such as log-log model with incorporated attributes like population, weather and taxi usage. They proposed a method of analyzing the trips of bikes between neighborhoods that contain multiple stations and the result was promising and led to improved predictions compared to just examining the stations itself.

Incentivizing Users for Balancing Bike Sharing Systems [4] the authors proposed a method to encourage users solve the imbalance problems themselves by providing users with monetary incentives every time they contribute to rebalancing the system. Through smart app, the systems can use users target stations and suggest to them different stations that might become imbalanced soon.

2. DATA ANALYSIS

2.1 Data Source

The datasets are provided by Healthy Ride Pittsburgh website, a bike sharing system provider in Great Pittsburgh area. The first dataset is Rentals record files for 3rd and 4th quarter in 2015 and 1st quarter in 2016 which include: Trip ID, Bike ID, Trip start day and time, Trip end day and time, Trip duration (in seconds), Trip start station name and station ID, Trip end station name and station ID, and Rider type: Member (pay as-you-go customer); Subscriber (deluxe and standard monthly member customer); Daily (24-hour pass customer). Our second dataset is the Station information which has: Station ID and Station name, Lat/Long coordinates and Number of individual docking points at each station. In order to enhance our models we decided to incorporate external attributes which we believe they could help us determining the correlated effects. These datasets are a "City of Pittsburgh 2015 Holiday Schedule" dataset which gives us the holidays and a weather datasets provided by Underground Weather website which include daily weather history and observations including: temperature, wind chill temperature, dew point temperature, humidity, atmospheric pressure, visibility, wind direction, wind speed and weather condition. The website provide an API to make the data accessible to everyone. Since we are using historical data, we will get numerous amount of data and additional processes have to be used to extract the important features.

2.2 Data Cleaning

We have combined weather data with holiday data to generate a new csv file for 9 months, July 2015 to Feb 2016. Then merged it with our original dataset comparing with attribute ‘StartDate’. The dataset consisted of 1980 missing values. We decided to remove them as it’s an insignificant number as compared to the total number of records 61104. After comparing with Stations.csv we found that there were two stations which were not there in Stations.csv but were present in original merged dataset. These two stations ids, 1050 and 1051 were then removed.

2.2 Data Visualization

In order to understand the distribution of the data, the exploratory data analysis must be used to investigate the data more closely. After we cleaned the data and managed to make a new column that have the number of bikes rented which is a "count" column, we decided to see what relationships this column has with the previous attributes. In figure 2.1 we plot a density plot of a number of rented bikes divided by the year. It is shown that the 2016 is skewed and that might be related to the a few number of observations we have in the datasets for that year. For figure 2.2, the density plot shows what hours that have the demand bikes.

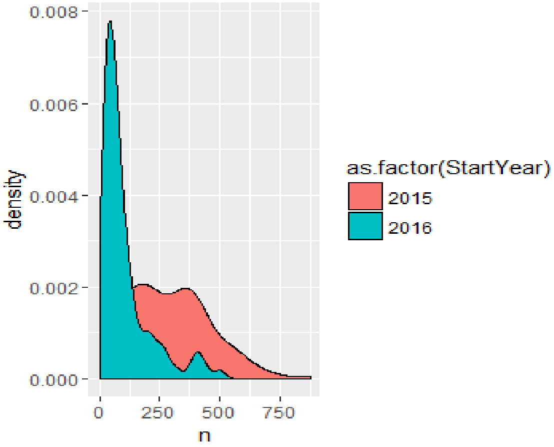

Figure 2.1. Density plot of factor ‘Start Year’.

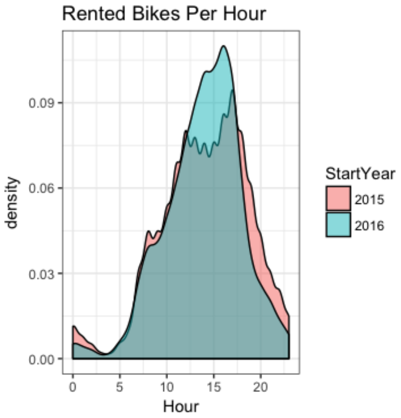

Figure 2.2. Density plot of factor ‘Start Year’.

Also, to make more exploratory graphs, we use ggplot functionality to split the data on months and years to see the distribution more easily. See figure 2.3. To determine what months are more likely to have more rented bikes we plot the months which is filled by the year to found out that July and August have the highest peek and that might be related to summer time in Pittsburgh whereas the coldest months like November, December and January have a few users. See 2.4.

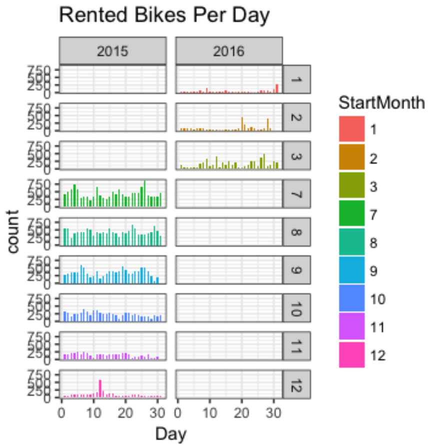

Figure 2.3. The distribution of rented bikes per months and years.

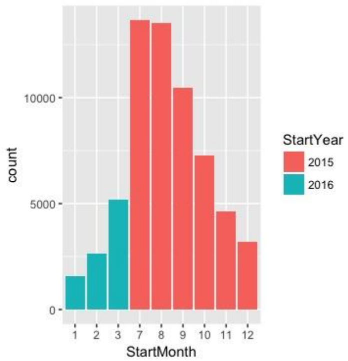

Figure 2.4. The bar plot of rented bikes per months and years.

The last part of visualization that is useful is that we aggregate the number of bikes per data and combined weather information on that days to see the correlation between the temperature and the number of counts reserved that day. Figure 2.5shows that there is a positive linear relationship between the two variables. As the temperature increases, users are more likely to rent bikes that day.

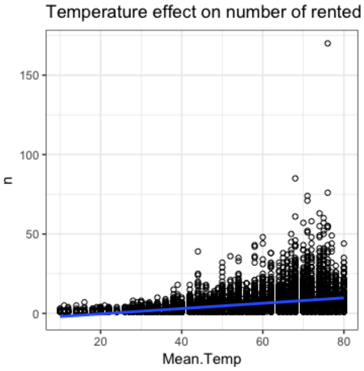

Figure 2.5. The correlation of rented bikes with the temprature.

2.3 Calculations

For the problem of imbalance, we calculated the number of bikes rented per station and then calculate the difference between the number of outgoing bikes and the number of incoming bikes at each station per day. We divide the difference by the rack quantity of each station which can be found in the stations file.

3. ALGORITHMS

3.1 Linear Regression

To predict the number of bikes rented, we used linear regression. Then we evaluated the model by using hold-out evaluation i.e, splitting up the data set into 19 days of training and last 10 days as test data for each month.

3.2 Classification

We plan to classify each station as balanced or not by applying classification algorithms. After calculating the imbalance at each station per day, classification algorithms like logistic, naïve bayes, knn, decision trees will be used to find the imbalance.

4. PROGRESS AND FUTURE WORK

Till now, we have done data processing, cleaning, summary statistics and applied linear regression model.

We plan to apply the classification algorithms like stated above. We will apply 10-fold cross validation for all algorithms and evaluate them by calculating F-score, Precision, Error, AUC and Recall.

5. REFERENCES

1. Alhusseini, M. 2014. *Prediction of Bike Sharing Systems for Casual and Registered Users*. Stanford University.

2. Romain Giot, Raphael Cherrier. Predicting Bikeshare System Usage Up to One Day Ahead. *IEEE Symposium Series in Computational Intelligence 2014 (SSCI 2014). Workshop on Computational Intelligence in Vehicles and Transportation Systems (CIVTS 2014)*, Dec 2014, France. pp.1-8, 2014.

3. Singhvi, D., Singhvi, S., Frazier, P. I., Henderson, S. G., O’Mahony, E., Shmoys, D. B., & Woodard, D. B. (2015, April). Predicting bike usage for new york city’s bike sharing system. *In AAAI 2015 Workshop on Computational Sustainability*.

4. Singla, A., Santoni, M., Bartók, G., Mukerji, P., Meenen, M., & Krause, A. (2015, January). Incentivizing Users for Balancing Bike Sharing Systems. In AAAI (pp. 723-729).