Predicting the number of bikes required in the facilities in the Pittsburgh stations three hours in advance

A Rebalancing Problem

Damodar Padubidri Bhat
School of Information Science
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15260
dab249@pitt.edu

Sanket Sameerkumar Bagewadi
School of Information Science
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15260
sab258@pitt.edu

Abstract — **Bike sharing, as an alternate means of transport, has attracted widespread attention in sustainable transportation research community. However, the competence of the rebalancing operations in a bicycle sharing system determines its success. This paper deals with the problem of rebalancing of bike sharing system. The two main problems of the bike sharing system are underutilization of a station or when the number of bikes available are less than the demand. Both these issues lead to inefficiency in the system. This paper tries to solve these problems by predicting number of bikes that will be available three hours in advance. It provides details of the data cleaning process and offers an insight into the dataset by analyzing the dataset and providing results in form of various data visualization techniques.**

Keywords—Rented Bikes; Pittsburgh; Data Mining; Utilize; Regression

1. INTRODUCTION

A new transit system on two wheelers can be experienced if this is available in all corners of the city. This system is present in more than 500 cities across the world [4]. Citizens can come out of the congested business areas and a search for parking space is eliminated saving a lot of time. Also, when we have a look at the public transport in many of the cities, considering a travel from point A to point B the passenger should change the buses multiple times. This can be eliminated. Accessibility and affordability of the service is much easier and cheaper than the traditional bus and rental car services. The spending on travel within the city drops exponentially with the use of rental bikes. From one of the papers we found that, increase in physical activity improves the health, hence saving lives. As the facility is much needed for any city, there arises a problem of how do we maintain the facility so that the utilization is high in most of the stations. Here, by high utilization we mean that having sufficient number of bikes need at the station also looking at the stations where the count should be decreased due to underutilization. We are trying to predict the bikes required at a station three hours in advance. If this can be successfully predicted, the administrators of this facility can easily move the bikes from one location to another. Thus, making the system more accessible to the public and efficient.

2. RELATED WORK

Bicycle-Sharing System Analysis and Trip Prediction [1] in this paper, the author discusses about how to predict the number of bikes required in each station to avoid the problem of lack of docking space for the arrived bikes and moving bikes to the stations were docks are not utilized. Metrics of regression problems like MSE (Mean Square Error) and R^2 (i.e., Coefficient of Determination) are used as evaluation metrics.

Bike-Sharing Prediction System [2] in this paper, the authors discuss the hierarchical prediction model to predict the number of bikes. They also use gradient boosting algorithm to predict the entire traffic. Geo-Space Contrary Prediction model is used compare with same period prediction datasets to improve the results.

Predicting Bike Usage for New York City's Bike Sharing System [3] in this paper, the authors discuss about analyzing the bike rental trends during the start of business hours (i.e., 7:00 AM to 11:00 AM during weekdays). Regression analysis is used for the predictions of bikes during the morning rush hours.

Predicting Bikeshare System Usage Up to One Day Ahead [4] in the paper, the author use regression techniques like Support Vector Regression, Random Forest Regression and Gradient Tree Boosting for prediction. Here the prediction is done for next twenty-four hours. Weather information and previous bike usage data has been used in the analysis.

3. DATA DESCRIPTION

3.1 Data Source

Data for this analysis has been collected from Healthy Ride website. Healthy Ride is non-profit shared bike facilitator in Pittsburgh. The mission is to provide affordable, easy-to-use transit for the public. Data is available for public use on the website. Datasets are available from quarter-2 of Year 2015 up to quarter-4 of Year 2016. Every quarter has two files. The first consists of rental information which includes fields like Trip ID, start time, stop time, bike ID, usage duration etc. The second dataset consists of Station ID, Station Name, latitude, and longitude and rack quantity. The datasets are collected in such a way that the personal information of the users cannot be found out. All the trips are recorded by using the key Trip ID and not using the details of the users' ID or information. The datasets clearly indicate the number of racks present in each station. This information will be handy in deciding the number status of the station like if it the busiest or has empty slots in the racks. We also use the weather information to understand the renting trends during different seasons. This data is collected from the National Centers for Environmental Information. The dataset consists of weather records for all the dates that are present in the rental dataset. In addition, the average temperature, snow fall indicator, rain indicator is present in the dataset.

3.2 Data Cleaning

This is one of the most important steps in the process of analysis of the data and predicting the result. Having improper data or noise leads to improper prediction. The steps that have been followed in data cleaning process is as follows. The datasets of all the quarters are merged together. Missing values in the dataset were removed. Insignificant columns in the dataset were removed. Undefined values in the dataset is are removed. The date and time were in a single column in the dataset and needed to be split into day, month, and year. The weather data obtained also need cleaning. The unwanted columns from the weather dataset were discarded. The date format in the weather dataset were not in the format required and hence needed to be changed so that the weather dataset could be merged with the bike dataset.

To understand the bike renting trend in all the seasons, we added a column to the dataset which indicates the climate of the day. This was done using the columns present in the initial weather dataset which gave specific indications on if the day was "Sunny", "Rain" or "Snow". Further, holidays were considered to know the trend on those days. The records present in the weather dataset were identified as either a business day (Monday to Friday) or weekend (Saturday and Sunday) by creating a column for holiday indication.

3.3 Data Visualiazation

To make proper analysis and prediction, it is important to understand and investigate the data at hand more thoroughly. We have added a new column to our existing data set which will give us the count of the bikes which are rented. We have some merged the cleaned weather data to understand if there is a relationship between the weather and bikes rented and if yes, then to understand this relationship.

In the figure 1, we plot a graph of number of bikes rented per hour in year 2015 and 2016

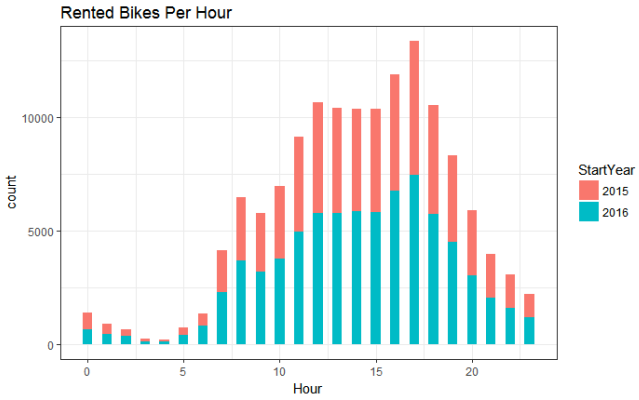

Fig. 1. Bar plot for bikes rented per hour and year.

We also wanted to understand which is the busiest months in year in a year are and hence we have plotted a graph of number of bikes rented each month in the figure 2.

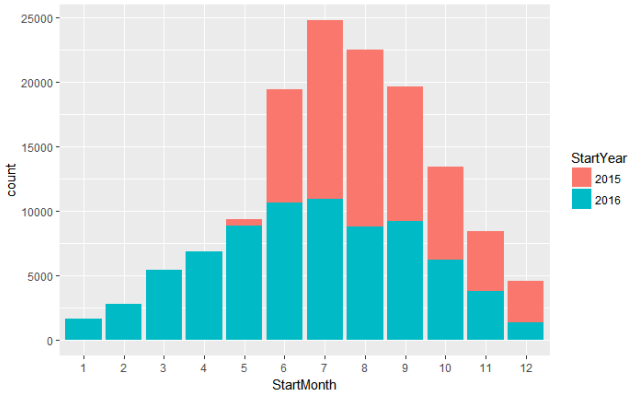

Fig. 2. Bar plot for bikes rented per month and year.

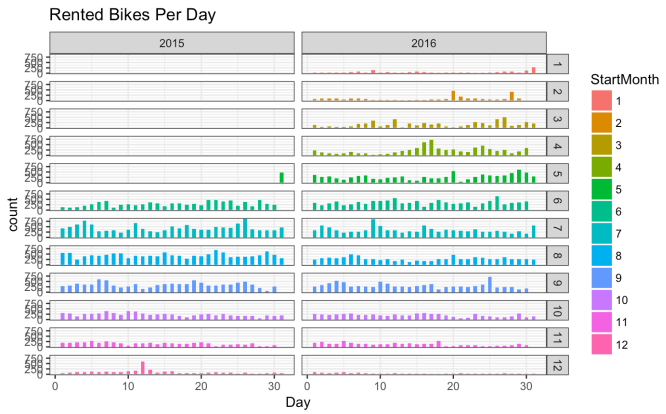

Fig. 3. Grid plot for bikes rented per day each month for both years.

These results are important since it tells us which are hours of the days and months of the year we need to focus on.
We found that between 11 A.M – 7 P.M the bikes are rented the most and between 2 A.M – 5 A.M bikes are rented the least. We also found that between May – September bikes are rented the most and least between December – February.

We also divided our dataset in terms of climate. We have three categories, sunny day, rainy day, or snowy day. We wanted to find out the impact climate has on number of bikes rented. As we can see from the Table below people hire bikes significantly more in summers than in other two conditions.

Climate	Count
Sunny	113461
Rainy	1610
Snowy	22332

Table 1. Climate wise bikes rented.

We also wanted to find number of bikes that are rented on a weekday and on a weekend.

Day	Count
Weekday	93748
Weekend	43655

Table 2. Weekday/weekend wise bikes rented.

Number of bikes rented on the weekend are almost half of number of bikes rented on weekday. Considering weekend consists of only 2 days but it is constituting almost half of weekday bikes, people use bikes more often on weekend.

We also found out which are the busiest stations each hour of the day with respect to bikes rented as well as returned.

For bikes rented, the top three busiest station and their corresponding hour is given in Table 3. Similarly, for the bikes returned, the top three busiest station and their corresponding hour is given in Table 4.

Station Name Pick up	Hour of the day	Count
10th St & Penn Ave	5 P.M - 6 P.M	848
Liberty & Stanwix	4 P.M – 5 P.M	694
S 27th & Tunnel Blvd	12 P.M – 1 P.M	693

Table 3. Busiest Rented Station each hour of day.

Station Name Drop off	Hour of the day	Count
S 27th & Tunnel Blvd	2 P.M - 8 P.M	6258
Forbes Ave & Market Sq	12 P.M – 1 P.M	674
21st St & Penn Ave	10 P.M – 11 P.M	332

Table 4. Busiest Rented Station each hour of day.

We found the busiest station for bikes rented and returned for each hour of the day.

We have also mapped the stations on google maps using the latitude and longitude attribute and plotted the count of bikes rented per station. This is shown in figure 4

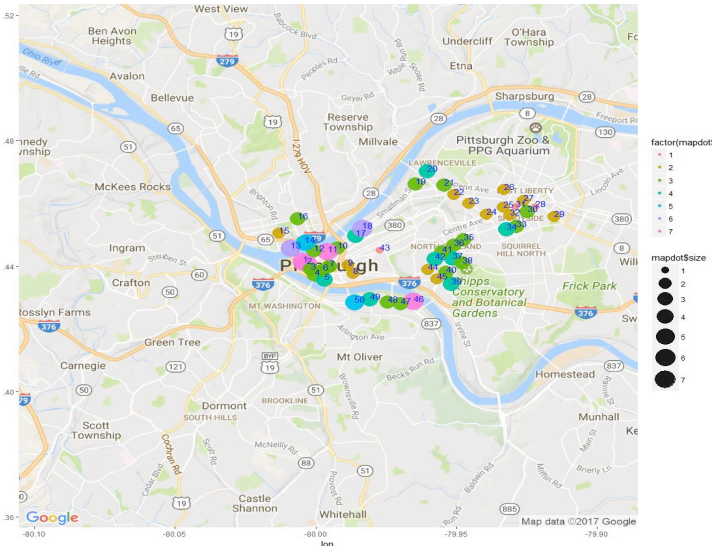

Fig. 4. Google map of count of bikes rented per station

The magnitude and color of the circle represents the different number of bikes rented each station

4. DATA MODELLING

Regression Techniques

We divided our dataset into training set and testing set. The training set consist of first 19 days of the month while the testing set consist of the next 10 days. This was done to avoid overfitting. We want to predict the number of bikes per hour in the stations. For this reason, we are using regression techniques. We have started our prediction model using the most basic Linear Regression.

4.1 Linear regression. We have used the linear regression model where we have included all the attributes to fit the model and then predict the result based on the testing dataset. From the results of the prediction on the testing data, we found that the linear regression model did not fit the data well. Most of the data points were located away from the line of best fit. The predictors that are used here are Climate, Weekday, Maximum Temperature, Latitude, Longitude, PickupHour, PickupMonth and PickupYear.

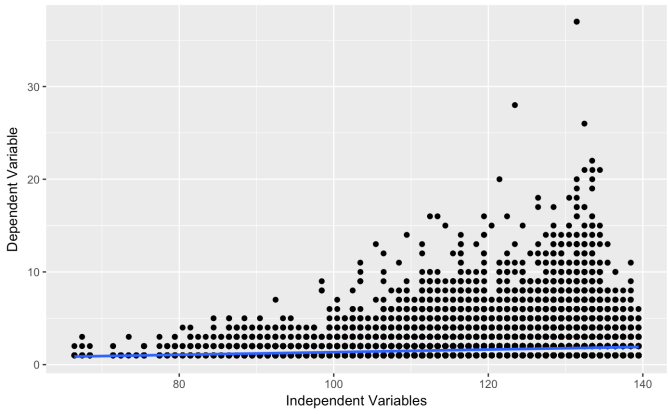

Fig.5. Linear regression plot of dependent vs independent variable.

For linear regression, the RMSE value is 1.28.

4.2 Polynomial Regression. Since the result of the linear regression did not fit the data well, we have tried to implement the polynomial regression using the training and testing dataset as mentioned above. Here we have used fourth degree evaluation for Climate, Latitude, Longitude; degree three for maximum temperature. From the Q-Q plot we can understand that the data is rightly skewed. For Polynomial regression, the RMSE value is 1.319.

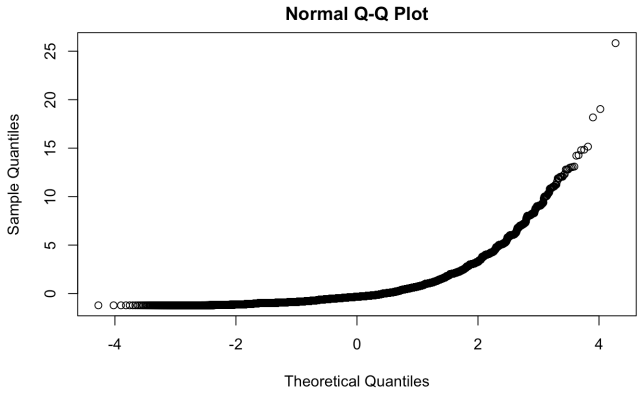

Fig.6. Normal Q-Q plot for polynomial regression

4.3 Random Forest Regression. Here, for random forest regression we have used the package random forest in R. We used the predictors Climate, Maximum Temperature, Weekday, PickupHour, The pickup station ID. We used 500 trees in this modelling.

In the figure which represents error vs number of trees, we can understand that as the number of trees increases, the error rate decreases. We tried to work on this modelling with number of trees as 1500. But it was taking an extended computation time and the error rate improvement was not significant. We confirmed this from few sources online that once the threshold has been crossed there will be no significant improvement in the error, So, we have resorted to using 500 trees for random forest evaluation. From the figure representation for feature extraction, we find that From.Station.ID is an important feature in random forest evaluation. The top three features extracted are From.Station.ID, PickupHour (The bike pickup hour) and maximum temperature of the day. The node purity for From.Station.ID stood out at 3870 and for weekday it was 495.

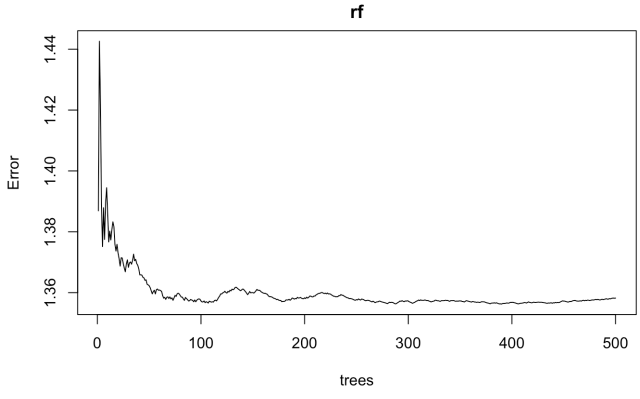

Fig.7. Error rate vs Number of Trees

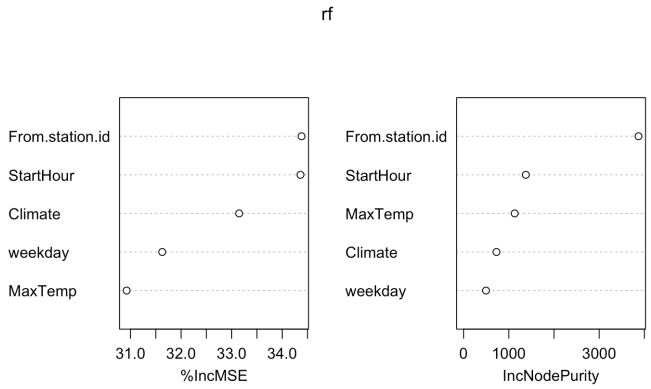

Fig.8. Feature Extraction for Random Forest

For random Forest regression, we obtained an RMSE of 1.29

4.4 Gradient Boosting. In the gradient boosting model, we have used the number of trees to be 20000. The interaction depth is 3. The predictors are Climate, Maximum Temperature, Weekday, PickupHour, pickup station ID. The relative influence for From.Station.ID stood out to be the highest with 57.37. Climate predictor had the least relative influence among the features. The figure shows us the relative influence of different predictors.

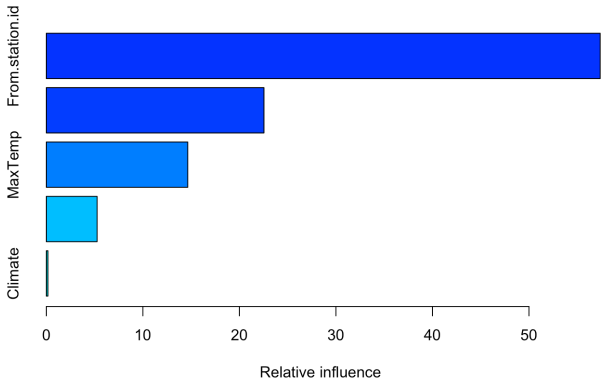

Fig.9. Relative influence graph for gradient boosting.

4.5 XG Boost. Here, we have used the package 'xgboost' in R. We have selected eta to be 0.5 and objective to be reg:linear. The RMSE value for XG boost model was 1.63.

4.6 Ensemble Model. Here, we selected two model with good performance measured using the RMSE values. We have selected Random Forest and Gradient Boosting for this model. Ensemble model combines two different models into one model to give a good output and hence reducing the error. The RMSE value for this model is 1.274.

RMSE Observations for all the models.

Model	**RMSE**
Linear Regression	1.288
Polynomial Regression	1.319
Random Forest	1.290
Gradient Boosting	1.275
XG Boost	1.632
Ensemble Model	1.274

From the above RMSE observations for different models, ensemble model performs the best. We tried improving random forest model by increasing the trees, but there was not significant improvement in the RMSE value. Addition of predictors to the model did not support the processing with a large time consumption. In gradient boosting model, the increasing the number of trees did not result to gathering a better RMSE value for the model. The ensemble model resulted to be the best among the models that were tested.

5. DISCUSSION

In this section, we would like to discuss various suggestion which may help algorithms perform more efficiently. First, the dataset consists of data from 2015 and 2016 and we felt that this is comparatively small dataset to make accurate prediction. Second there were a lot of errors in the dataset like missing values and in correct values which made it difficult to analyze and apply the algorithm. The only way we could deal with them was to remove these values which further reduced our dataset. Third there was no data with regards to the customers or various events happening in Pittsburgh. Having this information could lead to better prediction. Fourth we have considered start point for applying the algorithms which alone is not sufficient for good prediction. Drop station also play an important role as the additional bikes need to be moved to the stations with high demand.

6. CONCLUSION

This project has helped us to gain in-depth knowledge of data mining. We were able apply all the knowledge gain from the class. From data cleaning, to different visualization techniques to applying various algorithm. We also observed that the rack quantity provided by the Bike share company is not sufficient. There were cases

where number of bikes rented or returned in an hour exceeded the rack quantity. Weather, weekday, Pickuphour were the most important parameters overall for all the algorithms. Most of the algorithm performed at the same level except for XG Boost which performed significantly worse. The Ensemble model performed the best amongst all.

7. FURTHER WORK

Predictions which will consider other external factors like events happening over city, customer details may be able to predict more accurately. Further work also includes training the XG boost more efficiently so that it performs much better. We would also like to implement other techniques like Ridge regression and Neural Network.

8. ACKNOWLEDGEMENT

We would like to thank Prof Dr. Yu-Ru Lin and the GSA for guiding and helping us throughout the project. We appreciate the commitment and attention shown by them.

9. CONTRIBUTIONS

Sanket was responsible for cleaning and preparing the rental dataset. He was also responsible for Linear, Polynomial, and Random forest regression technique. Damodar was responsible for cleaning the preparing the weather dataset and applying the Gradient, XG boosting techniques and Ensemble models. Both contributed equally for the various data visualizations.

REFERENCES

[1] Zhang, J., Pan, X., Li, M., & Philip, S. Y. (2016, June). Bicycle-sharing system analysis and trip prediction. In Mobile Data Management (MDM), 2016 17th IEEE International Conference on (Vol. 1, pp. 174-179). IEEE.

[2] Cai, Q., Xue, Z., Mao, D., Li, H., & Cao, J. (2016, April). Bike-Sharing Prediction System. In International Conference on Technologies for E-Learning and Digital Entertainment (pp. 301-317). Springer International Publishing.

[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] Giot, R., & Cherrier, R. (2014, December). Predicting bikeshare system usage up to one day ahead. In Computational intelligence in vehicles and transportation systems (CIVTS), 2014 IEEE symposium on (pp. 22-29). IEEE.