Bike Sharing Demand Prediction

Sai Charan Talipineni
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15260
sat122@pitt.edu

Sai Rakesh Ghanta
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15260
sag163@pitt.edu

ABSTRACT

The goal of our project is to predict the demand of Healthy Ride; the public bike sharing system in Pittsburgh operated by Pittsburgh Bike Share. We aim to apply data mining techniques to examine two potential tasks. We gather the bike rental and stations data from Healthy Ride official website, weather data from National Climatic Data Center (NCDC) and Holiday data from the City of Pittsburgh Holiday Schedule. The dataset is skewed. We will use the classification algorithms to address the rebalancing problem and identify those conditions which indicates bikes would be rented from a specific station at a specific hour. We also apply classification methods to determine the rent and return behaviors. Finally, we will run the regression models to predict the quantity of bikes rented out from a specific station at a specific hour and find the significance of the predictors. Our prediction result capture the trend of hourly demand in the course of day of the period we cover, and our model will offer insight to study the balancing problem between supply and demand and the kinesis in the city of Pittsburgh.

Keywords

Machine Learning; Data visualization; Bike Sharing; Demand Prediction; Data Mining

1. INTRODUCTION

Healthy ride bike station:

In the modern day, health consciousness is increasing. In Pittsburgh, Healthy Ride has become a more and more important mode of transport preferred by the public, which offers a more convenient type of mobility, reduce the urban traffic and decrease the pollution caused by vehicles. Pittsburgh bike sharing is becoming popular day to day. Dynamic demand of bikes turns it out into an unbalanced system. The motivation behind this project is to help the bike rental company to pre-arrange the distribution of bikes in each station and minimize unnecessary waste of resources, and improve efficiency. The aim of the project is to predict the abundance/scarcity of the bikes to be rented from a specific station at a specific time given the information which is available. The scale of time is considered as hour. At first, we form a dataset merging the datasets of stations, bike rental records, weather data and city holiday schedule together, and then apply the machine learning methods to build a model to predict the demand of bikes.

The information considered in this project is identified with three distinct fields. The first is the bicycle rental records and stations information from Healthy Ride [1], and it incorporates the subtle elements of each bicycle trip and the area directions of each station. The second is the climate information accumulated from National Climatic Data Center (NCDC) [2], and it incorporates the most extreme and least temperatures recorded in every day. The third is the occasion plan information from the City of Pittsburgh Holiday Schedule [3][4], and it incorporates the data of weekdays and holidays. After the completion of data cleaning process, they are merged into one dataset. The schema of the dataset will be familiarized in detail in the next section. In this project, we focus on constructing a machine learning model predicting the hourly demand of bikes in each station, which is the key to study the balancing problem between supply and demand.

2. RELATED WORK

There are many studies about how bike sharing system functions and where the most popular stations are located. In order to deliver information on how our system is being used, many developers, researchers and everyone interested in the details of bike sharing system have examined the data for the purpose of analysis, visualization, development, or general curiosity.

Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns [8] by Patrick Vogel et.al. analyze extensive operational data from bike-sharing systems in order to derive bike activity patterns. A common issue observed in bike-sharing systems is imbalances in the distribution of bikes. They used Data Mining to gain insight into the complex bike activity patterns at stations. Activity patterns reveal imbalances in the distribution of bikes and lead to a better understanding of the system structure. A structured Data Mining process supports planning and operating decisions for the design and management of bike-sharing systems.

Data Analysis and Optimization for (Citi)Bike Sharing [9] by Eoin O'Mahony and David B. Shmoys analyze system data to discover the best placement of bikes to facilitate usage. They solve routing problems for overnight shifts as well as clustering problems for handling mid rush-hour usage. The tools developed from this research are currently in daily use at NYC Bike Share LLC, operators of Citi bike.

Bicycle-Sharing System Analysis and Trip Prediction [10] by Jiawei Zhang et.al. inferred the potential destinations and arriving time of each individual trip beforehand so that it can effectively help the service providers schedule manual bike re-dispatch in advance. They studied the individual trip prediction problem for bicycle-sharing systems. They studied a real-world bicycle-sharing system and analyze individuals' bike usage behaviors first. Based on the analysis results, a new trip destination prediction and trip duration inference model is introduced. Experiments conducted on a real-world bicycle-sharing system demonstrated the effectiveness of proposed model.

There were several other works published by data enthusiasts such as **Tableau: Healthy Ride PGH Bike Share Data** [7] by Lauren Renaud, **What's happening with Healthy Ride?** [5] by Jackson Whitmore and **A Study of Healthy Ride Pittsburgh Data** [6] by George Lejnine which gave us much deeper insights into the complex data.

3. ILLUSTRATION OF DATA

Our data sets include three kinds of data which are bike rental and station data, weather data, and holiday schedule data. After gathering the data, we select the useful features in each dataset at first. After that, we clean the data and encode some variables.

3.1 Data Collection

The datasets are available from Healthy Ride Pittsburgh official website. The first dataset consists of Rental records for 3rd and 4th quarters in 2015 and all the four quarters 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. The second dataset consists of Station information which has: Station ID and Station name, Lat/Long coordinates and Rack Quantity at each station. In order to improve our model, we decided to incorporate external attributes which we believe could help us determining the correlation. These datasets are City of Pittsburgh 2015 and 2016 Holiday Schedules which gives us the details of holidays. We also considered weather data obtained from official National Climatic Data Center (NCDC) website which include daily weather history and observations including: temperature, humidity, precipitation, atmospheric pressure, visibility, wind speed and several weather conditions. The data can be acquired from the website by providing required dates and ordering the data for free. Since we are using the past data, we will get huge amounts of data and we need to extract the important features.

3.2 Data Cleaning

The weather data is obtained from NCDC and it offers hourly weather data. The original dataset contains many features. However, we extract features TMAX and TMIN which are maximum and minimum temperatures as we believe these temperature has a significant effect on bike riders. We took the temperature as average of maximum and minimum temperatures TMAX and TMIN. We divided the average temperature TAVG into low (<45), medium (>45 and <70) and high (>70). Also, we have included Saturday and Sunday along with the holiday schedule. We have coded holiday = 1 and Non-holiday = 0 and added the attribute week of the year. We have joined the weather data same as above with the common attribute date. We merge all the above datasets into one dataset which contains the time and calendar data, ride record data and weather data. To prepare the dataset for training, we split the date attribute to hour, day, month and year. It might be the most significant factor to increase the model accuracy and reveal more hidden correlations.

3.3 Data Visualization

In order to better understand the correlation of various predictors in the data, exploratory data analysis must be used to examine the data more carefully. 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.

The below table gives the information about the trips of the bikes by the day in a week. As we can see in the table, Saturday has the highest count of the trips compared to the other days.

Table 3.1. Distribution of the count of bikes rented across different days in a week.

S. No.	Days	Count
1.	Monday	17251
2.	Tuesday	17344
3.	Wednesday	18850
4.	Thursday	17407
5.	Friday	19467
6.	Saturday	24620
7.	Sunday	23945

The below table gives the information about the trips of the bikes by the hour of a day for the whole dataset. As we can see from the table, evening 17:00 hour is very popular in the day.

Table 3.2. Distribution of the count of bikes rented across different hours in a day.

S. No.	Hour	Count
1.	0	1393
2.	1	902
3.	2	653
4.	3	243
5.	4	185
6.	5	723
7.	6	1336
8.	7	4152
9.	8	6451

S. No.	Hour	Count
10.	9	5768
11.	10	6957
12.	11	9131
13.	12	10652
14.	13	10390
15.	14	10354
16.	15	10342
17.	16	11876
18.	17	13357
19.	18	10510
20.	19	8316
21.	20	5918
22.	21	3980
23.	22	3090
24.	23	2205

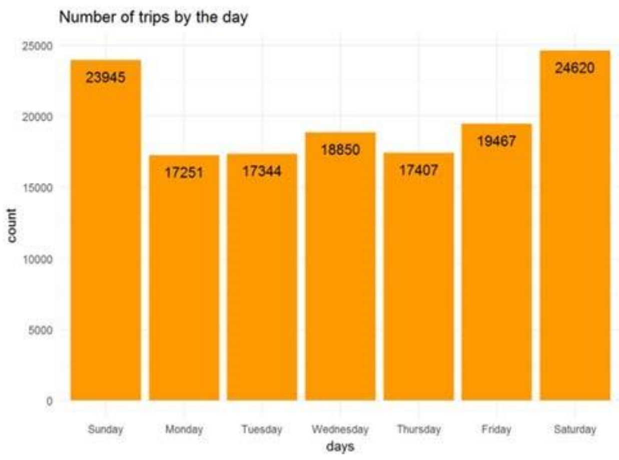
The below table gives the information about the trips of the bikes by the day in a week like whether it is holiday or a weekday. We coded weekday as the 0 and holiday as the 1. As the number of weekdays is five so we got more count for the weekdays compared to the holidays. But when we take holidays and weekdays as holiday hours and weekday hours, it is evident that holidays that the highest trips.

Table 3.3. Distribution of the count of bikes rented across holidays and week days.

S. No.	Type of the Day	Count
1.	0	86932
2.	1	51952

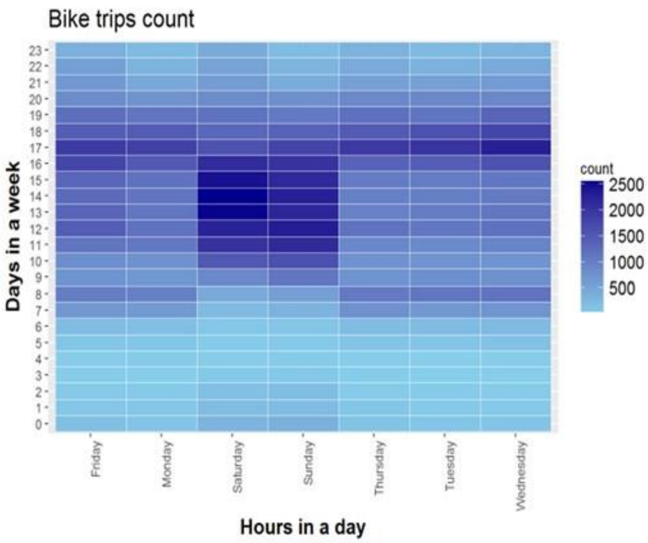
We also show a few plots to help picture the connection between the quantity of bicycles rented 'count' and the bike sharing related factors utilizing the final cleaned dataset. We use these plots to observe certain trends and patterns in the data.

Figure 1: The histogram plot of the count of bikes rented across different days of the week.



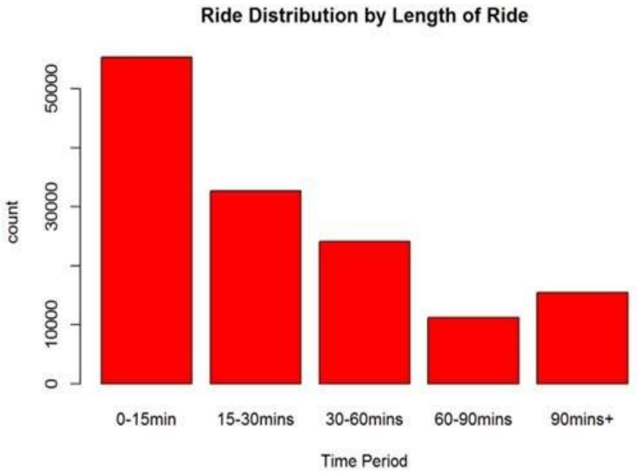
In Figure 1. we analyzed which days of the week are the most famous. Saturday by a wide margin is the most mainstream, trailed by Sunday. Monday is the most reduced of all days, with the quantity of rides getting increased as the week passes by.

Figure 2: The plot of the count of bikes rented across different hours in different days of the week.



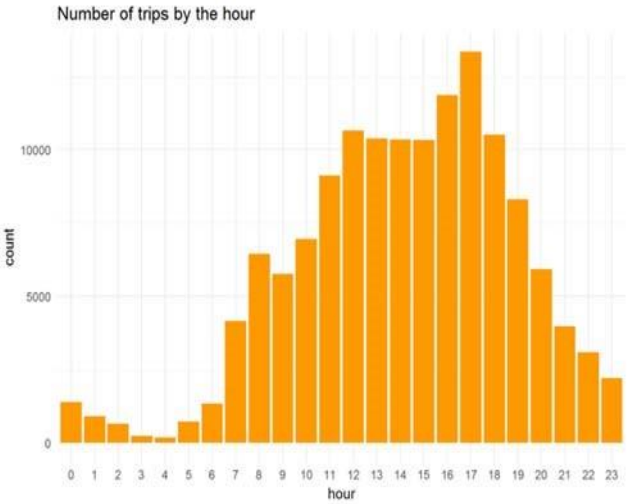
In Figure 2. the plot goes into more depth looking at what time of the day is most popular, broken down by day of the week. As we found in the diagram above, Saturday is the most mainstream, topping at 2 PM.

Figure 3: The plot of the count of bike rides in different time periods.



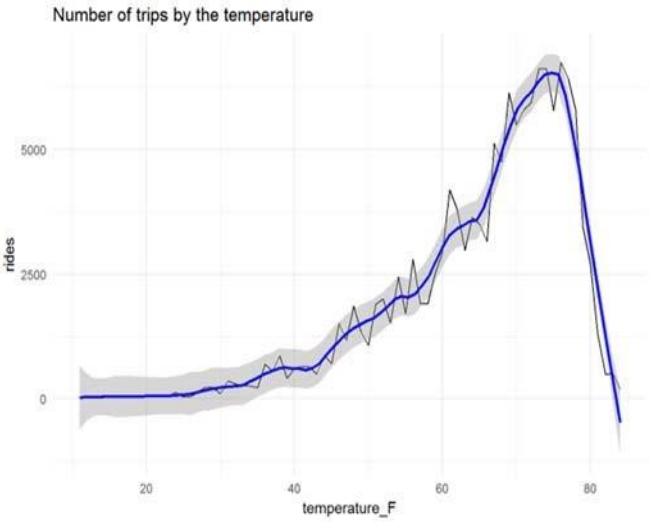
In Figure 3. the diagram takes a glimpse at to what extent bike rides last. Rides under 15 minutes represent most of all rides, while less number of rides lasts between 60 to 90 minutes.

Figure 4: The plot of the count of bike rides in different hours of a day.



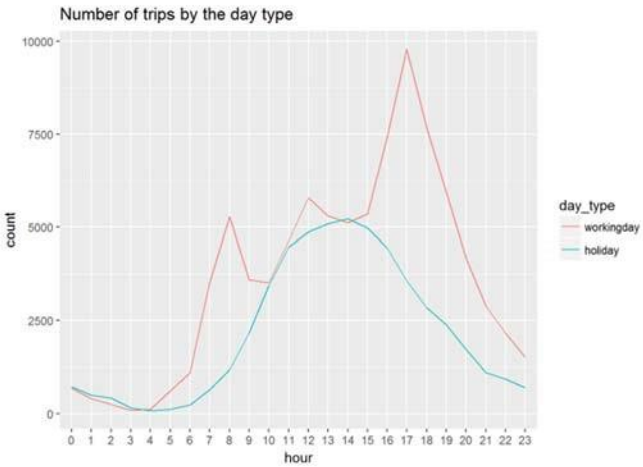
In Figure 4. we analyzed which hours of the days are the most popular. 5 PM is the peak hour where most number of the trips are observed. At 4 AM, least number of trips are observed. If we see the trend, the number of trips keep increasing from afternoon to evening.

Figure 5: The plot of the count of Bikes each day across the temperatures observed across each day.



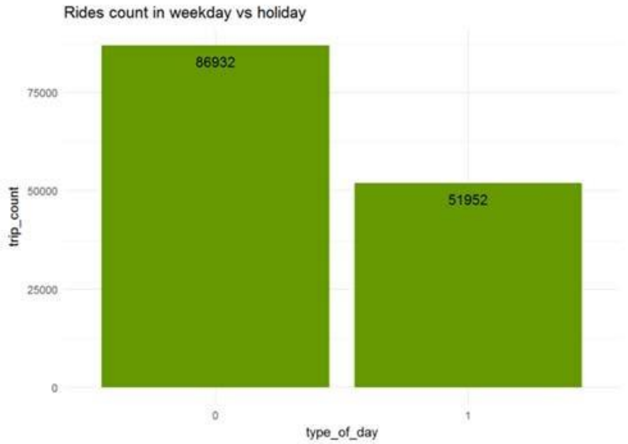
In Figure 5. In this chart we looked at how the daily temperature (provided by NCDC) compares to the number of daily rides. If we analyze the plot we can see that the peak is observed at 67 F. We can see that most of the trips are observed on a warm day whereas on the other hand least number of trips are observed on a cold day. So we can see that weather is one of the important features.

Figure 6: The plot of the count of trips across the week days and holidays.



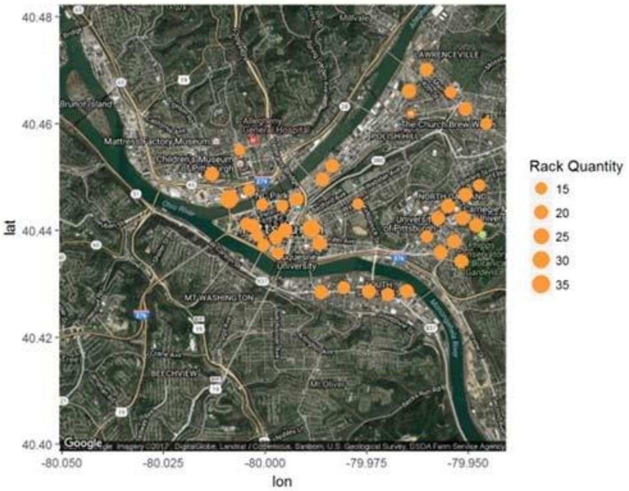
In Figure 6. we looked at how the number of trips vary from weekdays and holidays. We can see the count of the trips is more on weekdays when compared to the holidays.

Figure 7: The plot of the count of trips across different hours of working days and holidays.



In Figure 7. we looked at how the number of trips vary during different times of weekdays and holidays. We can see the count of the trips is in the peak during 6 PM of a working day and 2 PM of a holiday.

Figure 8: The plot of the spatial aspects of the trips.



In Figure 8. we analyze the spatial aspects of the trips. The Healthy Ride framework's ability is conveyed all through Pittsburgh however seems grouped in the Downtown zone and moderately portioned all through whatever is left of the Pittsburgh territory.

4. METHOD

At first, we divided the dataset attributes into two subsets corresponding to the bikes rented and returned. Then, we aggregated the count of bikes rented at a particular hour of a day in a month at a particular station. Again, we aggregated the count of bikes returned at a particular hour of a day in a month at a particular station. Difference of rent ~ return count of bikes forms the heart of rebalancing problem.

If the bikes are:

- Balanced (rent=return) it is coded as 0;
- else it is imbalanced and coded as 1.

4.1. Classification for the rebalancing problem

In this case, we applied binary classification techniques to address the rebalancing problem. Here, we considered the bikes count is balanced, if the number of bikes rented at a particular station at a particular hour of a day in a month is equal to the number of bikes returned at the same hour of a day in the same month at a same station. In addition, we considered the bikes count is imbalanced for the particular hour of a day in a particular month at a particular station, if the number of bikes rented is greater or less than the number of bikes returned at the same hour of a day in the same month at a same station (Balanced = 0, Imbalanced = 1). We divided the dataset into 19days training and remaining days testing data. We trained the model using the training dataset and we tested the model against testing dataset. We applied five different classification methods, including KNN, Logistic Regression, Naive Bayes, Dtree and Adaboost. We found the performance measures and ROC; they are shown in the Evaluation section. The model can be used to predict at a particular station at a particular hour of a day in a month whether the bike count is balanced or imbalanced. In the next classification, we took the next step to improve the project goal and to address the problem more efficiently by classifying data again.

4.2. Classification to determine the rent and return behaviors.

Again, we applied the binary classification techniques to the imbalanced situations. We took the subset of the original data where there is an imbalanced situation. We aggregated the count of bikes rented at a particular hour of a day in a month at a particular station. Again, we aggregated the count of bikes returned at a particular hour of a day in a month at a particular station. We found the difference of rent ~ return count of bikes. We coded the response variable for the particular hour of a day in a month at particular station as 0 if the number of bikes rented is less than the number of bikes returned at that particular hour. In addition, we coded the response variable for the particular hour of a day in a month at particular station as 1 if the number of bikes rented is greater than the number of bikes returned at that particular hour. Same as above classification we divided the dataset into 19days training and remaining days testing data. We trained the model using training data and tested against testing data. We found the performance measures and ROC, which are shown in the evaluation section. The model can be used to predict at a particular station at a particular hour of a day in a month whether rent > return or rent < return. In the next method, we took one more forward step to improve the project goal and to address the problem more efficiently by applying regression to the dataset and to find the exact number of bikes behavior at a particular station at a particular hour of a day in a month.

4.3. Regression

This is the next step in addressing the problem and more efficient method. This can help the company more efficiently to rebalance the bikes optimally between the stations. Here, we started with the baseline models for the future work. We divided the dataset into 19 days training and remaining days testing. We applied the linear regression and trained the model with training data. The model is tested against testing data and found the performance measures. The performance results are not so great as the data is not normally distributed and the data is very less for this problem to address efficiently. The performance measures are shown in the evaluation section. To find the exact bikes at a particular station at a particular hour of a day in a month more efficient algorithms need to be applied with much more data to train the model, which we considered it as future development as the bike sharing program has started in 2015.

5. EVALUATION

We organized the dataset for classification and we appended them to the equation to examine which stations are inclined to be imbalance. For this task we used different valuation performances such as accuracy, precision, recall, f-scores and AUC. Table 5.1 summarizes performance of all the models we used for binary classification to address the rebalancing problem;

Table 5.1. Performance Evaluation Table – Classification (1)

	KNN10	LR	NB	DTree	ADA
error	0.2598	0.4920	0.4886	0.5000	0.5495
accuracy	0.7401	0.5079	0.5113	0.5000	0.4504
precision	0.9059	0.9191	0.9215	0.9034	0.9257
recall	0.7958	0.5044	0.5062	0.5000	0.4277
fscore	0.8459	0.5966	0.5988	0.9492	0.5728
auc	0.5245	0.5456	0.5652	0.5000	0.5717

*We considered different variations of KNN (k=10,30,50) and showcased best out of the three variations.

Next we plotted receiver operating characteristic (ROC) curve, to illustrate the performance of a binary classifier system as its discrimination threshold is varied. Figure 5.1 showcases the ROC plot for classification methods applied to address the rebalancing problem.

Figure 5.1. ROC curve for Classification – (1)

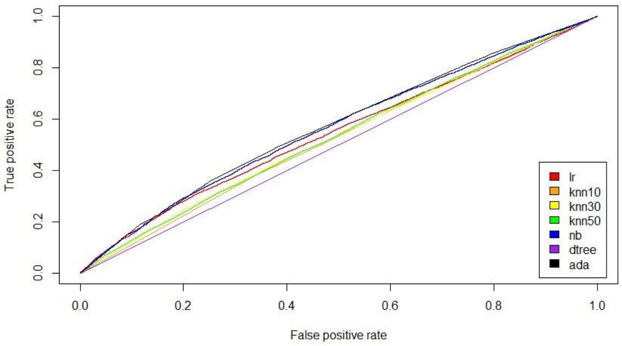


Table 5.2. summarizes performance of all the models we used for binary classification to determine the rent and return behaviors;

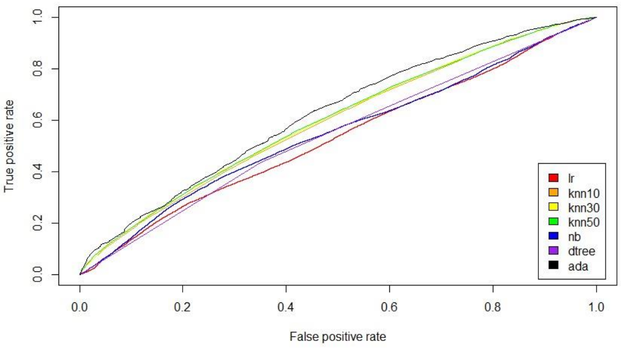
Table 5.2. Performance Evaluation Table – Classification (2)

	KNN50	LR	NB	DTree	ADA
error	0.4741	0.4856	0.4761	0.4830	0.4372
accuracy	0.5258	0.5143	0.5238	0.5169	0.5627
precision	0.5947	0.5259	0.5425	0.5252	0.5800
recall	0.5753	0.5143	0.5238	0.5288	0.6021
fscore	0.4560	0.4682	0.4788	0.5815	0.5510
auc	0.5988	0.5288	0.5477	0.5431	0.6175

*We considered different variations of KNN (k=10,30,50) and showcased best out of the three variations.

Next we plotted receiver operating characteristic (ROC) curve, to illustrate the performance of a binary classifier system. Figure 5.2. showcases the ROC plot for classification methods applied to determine the rent and return behavior.

Figure 5.2. ROC curve for Classification – (2)



Performance results of our regression model are RMSE = 1.30, ME = 0.09 , R2 = 0.07 and adjusted R2=0.07. This model is not so great because dataset is not enough and the dataset available is not normally distributed. The main drawbacks of the dataset and other discoveries are stated clearly in the discussion.

6. DISCUSSION

There were a lot of challenges faced during the project. The dataset itself is a big challenge. The dataset is skewed. It is not distributed normally. This causes the main challenge while performing data analysis. We ought to device more data visualization methods to provide much deeper insights into data. Classification improvement is needed. Since the dataset is skewed, the results were not as expected. We need to reduce time-consumption while implementing SVM. Parameter Tuning needs to be done. The factors hour and temperature affects greatly on the bike rentals.. Adding several new features like popular events in the city like Light up night, Steelers/Pirates/Pens Games etc. and geographical and temporal data might help our case. Since the bike sharing system started only in May 2015 we have less data to build a model and perform prediction. Having more data can enhance our results.

7. CONCLUSION

In this section, we showcase our key take away from the project. This project helped us gain insights into real-time data. We found SVM takes considerable time on very large datasets. Advanced Regression models need to be applied to predict the particular number of bikes at a particular station at a particular hour of a day more effectively. More data is required to train the model for better results. The factors hour and temperature affects greatly on the bike rentals.

8. ACKNOWLEDGMENTS

We would like to thank Dr. Yu-Ru Lin sincerely for her supervision and suggestions in each milestone of the project. We really appreciate her dedication and attention to detail in each step of the project. We also greatly appreciate the other groups for sharing their insights on the project in the group discussion.

9. REFERENCES

[1] Healthy ride data.
https://healthyridepgh.com/data/

[2] National climatic data center.
https://www.ncdc.noaa.gov/

[3] City of Pittsburgh holiday schedule. (2016)
http://apps.pittsburghpa.gov/pcsc/2016 Holiday Schedule.pdf.

[4] City of Pittsburgh holiday schedule. (2015)
http://apps.pittsburghpa.gov/pcsc/2015 Holiday Schedule.pdf.

[5] What's happening with Healthy Ride?
http://suds-cmu.org/2016/04/21/whats-happening-with-healthy-ride/

[6] A Study of Healthy Ride Pittsburgh Data
http://lejnine.com/bike analysis.html

[7] Tableau: Healthy Ride PGH Bike Share Data
http://www.laurenrenaud.com/blog/2016/2/1/healthy-ride-pgh-bike-share-data

[8] Vogel, Patrick, Torsten Greiser, Dirk C. Mattfeld. 2011. Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns. *Procedia - Social and Behavioral Sciences* 20 514–523. doi: 10.1016/j.sbspro.2011.08.058.

[9] O'Mahony, E., and D. B. Shmoys. 2015. "Data Analysis and Optimization for (Citi)Bike Sharing". *In Proceedings of the Twenty Ninth AAAI Conference on Artificial Intelligence*, January 25-30, 2015, Austin, Texas, USA., 687–694. AAAI.

[10] J. Zhang, X. Pan, M. Li, and P. Yu. Bicycle-sharing system analysis and trip prediction. *In MDM*, 2016.

10. AUTHOR CONTRIBUTIONS

In this project, both the authors contributed equally to the initial extraction and cleaning of data. We all contributed equally for providing insights into data and data exploration. ideas of model design and participated in the data collection and cleaning process. Sai Charan Talipineni was responsible for classification methods KNN, Logistic Regression (LR), Naïve Bayes (NB), Decision Trees and AdaBoost. Sai Rakesh Ghanta was responsible for regression analysis. Both the authors evaluated the performance of the model and reviewed the manuscript.