Enough Bikes Already for Sharing?

[Demand Prediction of Pittsburgh Bike Share]

Fengtao Wu
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15260
few14@pitt.edu

Yuwei Chen
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15260
yuc94@pitt.edu

Zhendong Wang
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15260
zhw65@pitt.edu

ABSTRACT

Pittsburgh's public bike sharing system Healthy Ride began the trail operation in summer in 2015 and the ambition of the system is "expanding access to public transit through easy-to-use, affordable active transportation opportunities". The goal of the project is to construct a machine learning model predicting the hourly demand of bikes in each station. We collect the bike ride and station data from Healthy Ride, weather data from National Climatic Data Center and calendar data from the City of Pittsburgh Holiday Schedule. The range of data is from July 1, 2015 to March 31, 2016.

We apply K-means clustering algorithm to identify and group the stations sharing the same temporal ride patterns in the course of day together at first. In general, the majority of hourly renting count in each station is zero. In order to handle the imbalanced problem, we combine the classification algorithms with the regression algorithms to predict the hourly demand of bikes in each station. Our prediction result capture the trend of hourly demand in the course of day well in the nine months we cover, and our model will offer insight to study the balancing problem between supply and demand and the mobility in the city of Pittsburgh.

Keywords

Data mining; Machine learning; Bike sharing

1. INTRODUCTION

The public bike sharing system offers a mobility service that people can rent and ride the public bikes for their journey in the city. The bikes are offered at stations that are distributed across the different areas in the city. The customer can rent a bike from the starting station, ride it for a journey, and return it to the destination station. Pittsburgh's public bike sharing system Healthy Ride[7] began the trail operation in summer in 2015 and the ambition of the system is "expanding access to public transit through easy-to-use, affordable active transportation opportunities".

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. However, the service provider faces the difficulty that they have to ensure that there are bikes available in the station whenever the customer starts and that there are parking lots available in the station whenever the customer arrives, and this problem is called the balancing problem between supply and demand. Healthy Ride releases the datasets quarterly in order to provide information on how the system is being used. 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.

We collect the bike ride record and station data from Healthy Ride, weather data from National Climatic Data Center (NCDC)[8] and calendar data from the City of Pittsburgh Holiday Schedule[6]. The range of data is from July 1, 2015 to March 31, 2016. We apply K-means clustering algorithm to identify and group the stations sharing the same temporal ride patterns in the course of day together at first. In general, the majority of hourly renting count in each station is zero. In order to handle the imbalanced problem, we combine the classification algorithms with the regression algorithms to predict the hourly demand of bikes in each station. The classification result is evaluated by accuracy, sensitivity and specificity, and the regression result is evaluated by the root mean square error (RMSE). Our prediction result captures the trend of hourly demand in the course of day well in the nine months we cover, and our model offers insights to study the balancing problem between supply and demand and the mobility in the city of Pittsburgh.

2. RELATED WORK

Patrick Vogel and Dirk C. Mattfeld[12] presented a data mining model to identify the typical usage patterns of bike-sharing systems and forecast the hourly bike demand of the system in Vienna. Patrick and Dirk incorporated the bike ride patterns and the effect of weather to the regression model which achieved a fairly high regression accuracy.

In this paper we present a hybrid machine learning model to predict the hourly bike demand of each station instead of the system. Though the schema of the datasets still looks similar, yet the distribution of the data instance changes significantly. Besides the fact that the number of data instances increases dramatically, the biggest challenge is that

© 2016 ACM. ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

the majority of hourly renting count in each station is zero and that the distribution of hourly renting count is highly imbalanced. Therefore, we combine the classification algorithms with the regression algorithms to predict the hourly demand of bikes in each station in order to handle the imbalanced problem.

3. DATASET

Our data sources includes three kinds of data which are calendar data, bike ride record and station data, and weather data. Then we select the useful features in each dataset at first. After that, we clean the data and encode some variables into a new data type.

3.1 Calender Data

The Calender Data is from Pittsburgh Holiday Schedule. We select the field *Holiday* indicating whether the given day is holiday, and confirm the day of week for each day.

3.2 Bike Ride Record and Station Data

The bike ride record data is from Healthy Ride website. The raw data contains the transaction data of bike rented from different stations. To fulfill our goal, we aggregate the data into hourly basis. Then we calculate the hourly bike renting count, the hourly bike returning count, the mean of bike ride duration less than one hour and the mean of bike ride duration less than two hours.

We also mark the date in the bike ride record data to show the day of week and whether a given day is on the weekend or not.

The station data is also from Healthy Ride website. We extract the geographical information (latitude and longitude) and the bike quantity of each station in the raw data.

3.3 Weather Data

The weather data is from NCDC API which offers hourly weather data. The raw dataset contains many features. However, we only select four features which are weather type, visibility, temperature and wind speed. We believe these features have significant effect on riding bikes. In the raw dataset, when some data at a hour is missing, we use the value of data at the nearest hour period to replace the missing value.

The weather type feature in the raw dataset is a list of weather type codes which are categorical, and it is hard to use them for the prediction. In order to handle the problem, we encode each original weather type code to a numerical value which represents how such a weather condition will affect the bike ride. For example, the weather like thunderstorm will have a higher value, while the weather like blowing will have a lower value. If multiple weather conditions appear in the same hour, we calculate the largest significance value among them to encode the weather conditions. The following is the detail encoding algorithm to calculate the significance of a given weather type:

First, we separate weather code into four main categories:

Level 0: the weather, such as sunny and clear, will not affect bike riding.

Level 1: the weather, such as mist and blowing, will have a slight effect on riding bike

Level 2: the weather, such as rain, fog and freezing, is not suitable for user to ride bike, but user can still ride bike.

Level 3: the weather, such as thunderstorm and snow, is so bad that user cannot ride bike.

Then, we use the following formula to calculate the significance value

$$weather.type = \begin{cases} 0 & \text{if level} = 0 \\ 3 \times level + significance - 1 & \text{if level} > 0 \end{cases}$$

The significance in the above formula represents the "+","-" in original field, and it indicates whether the weather condition is severe. "+" represents more severe, "-" represents less severe. Here, we let "+" be +1, "-" be -1 and if no "+-" let it be 0. The range for Level is from 0 to 3.

3.4 Data Merging

We merge all the above datasets into one dataset which contains the time and calendar data, ride record data and weather data. The merged dataset has 330,000 instances which means the join of the data of 24 hours, 50 stations and 275 days. It has 25 variables which belong to three kinds of data: time and calendar data, ride record data and weather data.

The variables related to time and calendar data include:

1. *Month*: the month of ride record instance;

2. *Day*: the day of ride record instance;

3. *Year*: the year of ride record instance;

4. *Hour*: the hour of ride record instance;

5. *Weekend*: indicated whether the day is on the weekend (1) or not (0);

6. *Holiday*: indicated whether the day is holiday (1) or not (0);

7. *DayofWeek*: indicated which day of the week the day is;

The variables related to ride record data include:

1. *StationId*: the station ID of the station where the bike was rented from in the ride record instance;

2. *Bike_out*: the count of bikes rented for pair of the station and the time;

3. *MeanOfTripDuration_1*: the mean of trip durations less than 1 hour for pair of the station and the time;

4. *MedianOfTripDuration_1*: the median of trip durations less than 1 hour for pair of the station and the time;

5. *MeanOfTripDuration_2*: the mean of trip durations less than 2 hours for pair of the station and the time;

6. *MedianOfTripDuration_2*: the mean of trip durations less than 2 hours for pair of the station and the time;

7. *Customer*: the count of bikes rented by the user whose usertype is "customer" for pair of the station and the time;

8. *Daily*: the count of bikes rented by the user whose usertype is "daily" for pair of the station and the time;

9. *Subscriber*: the count of bikes rented by the user whose usertype is "subscriber" for pair of the station and the time;

10. *Unknown*: the count of bikes rented by the user whose usertype is "unknown" for pair of the station and the time;

11. *Bike_in*: the count of bikes returned for pair of the station and the time;

12. *RackQnty*: the maximum of number of bikes in the station;

13. *Latitude*: the latitude of the station;

14. *Longitude*: the longitude of the station;

The variables related to weather data include:

1. *Weather.type*: the numeric score describing the weather condition for the hour;

2. *Visibility*: the visibility for the hour;

3. *Temperature*: the temperature for the hour;
4. *Wind.speed*: the wind speed for the hour;

Based on the above dataset, we have three important observations:

1. The distribution of hourly renting count depicted in Table 1 is highly imbalanced, and the majority of instances in the dataset have the hourly renting count equal to zero.

Table 1: Distribution of Hourly Renting Count

Hourly Renting Count	Percentage (%)
0	88.25
1	7.48
2	2.51
3	0.87
>3	0.89

2. The monthly overall renting count varies greatly through different months. Figure 1 indicates that customers use bikes much more often in summer than in winter.

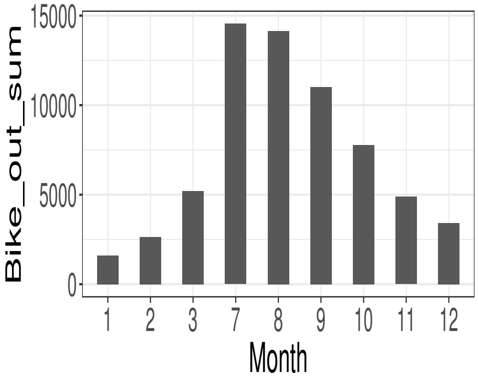

Figure 1: Distribution of Monthly Overall Renting Count.

3. Different stations have different temporal ride patterns in the course of day. In Figure 2, each line represents the total renting count in the course of day for each station. The figure demonstrates that customers tend to use bikes more often from 9 am to 9 pm generally, and that there exists some stations where customers prefer to renting bikes more often than the others.

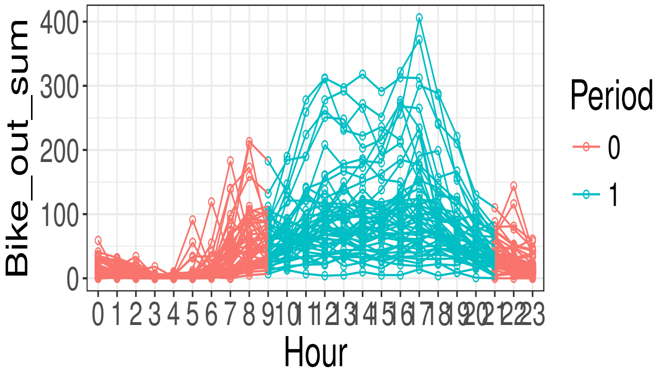

Figure 2: Temporal Ride Patterns in the Course of Day.

4. METHOD

According to the distribution of monthly overall renting count, the monthly overall renting count varies greatly through different months. Therefore, we build the prediction model for each month. We purpose a hybrid machine learning model including clustering, classification and regression algorithms to predict the hourly demand of bikes in each station. First, now that different stations have different temporal ride patterns in the course of day, we apply K-means algorithm to group the stations sharing the similar ride patterns into the same cluster. Second, since the distribution of hourly renting count is highly imbalanced, we conduct different classification algorithms to identify the records having non-zero renting count in each cluster. Third, we employ different regression algorithms to obtain the numerical result of renting count of which the records are classified to have non-zero renting count in each cluster.

4.1 Clustering

We use the dataset aggregated from the overall 330,000 ride records to conduct the clustering process. The dataset includes the average count of bikes rented, the average count of bikes returned, the average length of trip duration less than one hour and the average length of trip duration less than two hours in the course of day for each station. The information of rack quantity, latitude and longitude for each station is then added into the dataset, too. Therefore, the dataset for clustering has 50 rows which means 50 stations and 100 ($= 1 + 4 \times 24 + 3$) columns, and the dataset is normalized before running clustering algorithm.

The clustering algorithm we choose is K-means clustering algorithm. We evaluate the outcome of the algorithm with different validation indices that determine the cohesion and separation of clusters, and the indices are the Davies-Bouldin-Index[10], Dunn-Index[9], and Silhouette-Index[11]. High values for Dunn and Silhouette-Index indicated a proper clustering whereas the opposite holds for the Davies-Bouldin-Index. Results of the three indices are depicted in Table 2.

Table 2: Cluster Validation Results for K-means Algorithm

Number of clusters k	2	3	4	5
Davies-Bouldin	**1.1168**	1.7609	1.7441	1.8790
Dunn	**0.4345**	0.3273	0.2709	0.3169
Silhouette	**0.3920**	0.1569	0.1404	0.1246

According to Table 2, the drop in the Davies-Bouldin-Index and peaks for Dunn and Silhouette appears when $k = 2$, which indicates a proper clustering for 2 clusters. The MDS map is displayed in Figure 3.

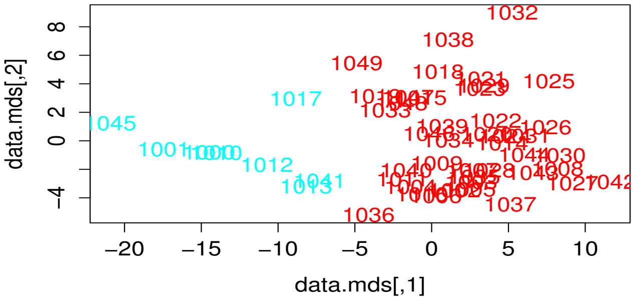

Figure 3: MDS Map for the Stations.

The results of the cluster analysis are temporally and spatially validated in order to identify the meaning of each cluster in reality. In Figure 3, The first cluster colored in red includes 42 stations and the second cluster colored in green includes 8 stations. We calculate and compare the average hourly renting and returning count in the course of day in each cluster. The result indicates that more bikes are rented from and returned to the stations belonging to the second cluster in the course of day. The comparison is displayed in Figure 4 and 5. Therefore, we call the second cluster the popular cluster and call the first cluster the ordinary cluster in the following sections. The station in the popular cluster is called the popular station, and the station in the ordinary cluster is called the ordinary station.

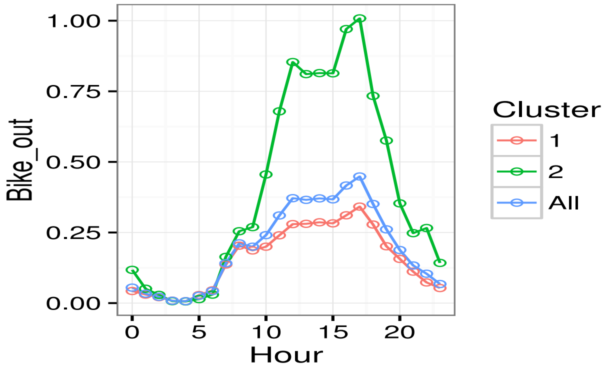

Figure 4: Average Hourly Renting Count in Two Clusters.

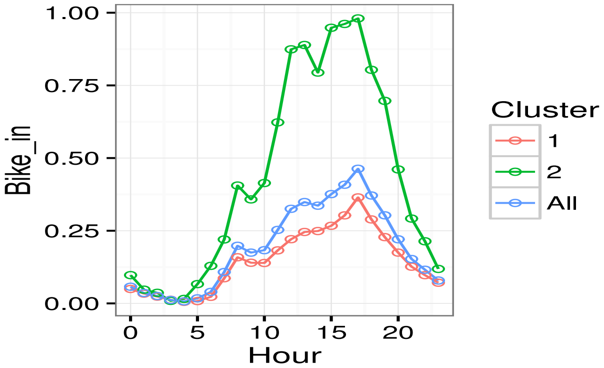

Figure 5: Average Hourly Returning Count in Two Clusters.

The clusters' geographical distribution is visualized in Figure 6 in order to identify spatial reasons for the activity patterns. Figure 6 demonstrates that 6 popular stations are located in the downtown area, 1 popular station is located in the campus of University of Pittsburgh, and 1 popular station is located in south side flats area. It may imply that bikes are the preferred mode of transport for the people working in the downtown area and the students studying in the University of Pittsburgh.

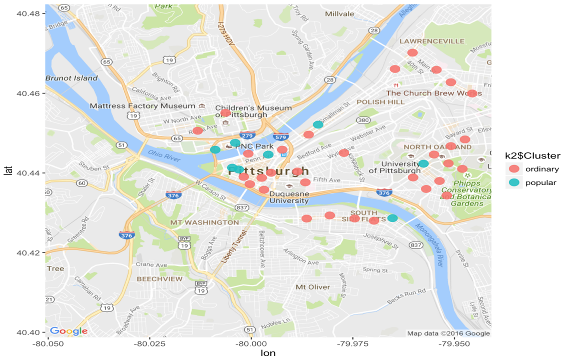

Figure 6: Clusters' Geographical Distribution in the Pittsburgh Map.

The above clustering result of stations will be used and kept the same for different months in the following classification and regression section, which means the status of the popular stations and the ordinary stations will not change for different months. The first reason is that the popularity of stations is found to be stable in general by the means of observing the top 5 popular stations through different months in the visualization system[5]. The second reason is that it simplifies the framework of classification and regression, and that makes the results in different months are comparable.

4.2 Autoregressive Model

After the clustering result is obtained, the prediction process includes another two stages which are classification stage and regression stage. The demand prediction is meaningful when the prediction is based on the information available.

During these two stages, we use the time and calendar data $\vec{TC}$, ride record data $\vec{R}$ and weather data $\vec{W}$ to predict the result. The prediction result B_t at hour t is based on the information at hour $t-i$, and this approach is similar to the concept of autoregressive model in the time series analysis. We choose the time lag i to be 1 and 2, and then call the corresponding models are AR(1) model and AR(2) model.

The equation of AR(1) model is

$$B_t = f(\vec{TC_t}, \vec{R_{t-1}}, \vec{W_{t-1}})$$

The equation of AR(2) model is

$$B_t = f(\vec{TC_t}, \vec{R_{t-1}}, \vec{W_{t-1}}, \vec{R_{t-2}}, \vec{W_{t-2}})$$

f represents the algorithms employed in the classification stage and regression stage. Both the classification and regression models are built for each month separately. The training dataset is the collection of records before 20th in each month, and the test dataset is the collection of records in the rest days in each month.

4.3 Classification

The distribution of hourly renting count is highly imbalanced in the overall 330,000 ride records, and the majority of ride records have the renting count to be zero. In the classification stage, we build the classification models for each month, and apply different classification algorithms including logistic regression, support vector machine[1], decision tree[3], naive Bayesian[1], adaptive boosting[4] and neural network[2] with no more than two hidden layers to predict whether the hourly renting count of the record is zero or not. Now that the stations have been grouped into the popular cluster and the ordinary cluster, the classification algorithm is conducted separately in the collections of records in the two clusters. The training dataset is the collection of records in each cluster before 20th in each month, and the test dataset is the collection of records in each cluster in the rest days in each month.

The response variable B_t in the classification stage is 0 or 1, which means the hourly renting count of the record is zero or non-zero. For those records which are classified to be 0, we call them zero records, and the hourly renting count is predicted to 0. For those records which are classified to be 1, we call them non-zero records. The records are marked and the hourly renting count will be predicted by the regression algorithms in the next stage.

Now that the distribution of hourly renting count is highly imbalanced, we consider the classification result detecting more true non-zero records is better, which implies that we regard the specificity more important than the sensitivity. Hence, we seek the proper cutoff by the means of 10-fold cross validation.

4.4 Regression

In the regression stage, we build regression models for each month, and apply different regression algorithms including linear regression, decision tree[3] and support vector regression[1] to predict the numerical results of the renting count of the records which are classified to be non-zero records in the classification stage. Similar to the classification stage, the regression algorithm is conducted separately in the collections of records in the two clusters. The training dataset is the collection of non-zero records in each cluster before 20th in each month, and the test dataset is the collection of records which are classified to be non-zero records in each cluster in the rest days in each month. We also calculate the ceiling, flooring and rounding value of the prediction result to convert it to be an integer. The results of different algorithms is evaluated by comparing the root-mean-square error.

5. EVALUATION RESULTS

We build the classification and regression models in AR(1) and AR(2) model framework separately for each month. Here we take the dataset of September 2015 as an example and follow AR(1) model framework to illustrate the classification and regression results.

5.1 Classification Result

Among all the classification algorithms, support vector machine algorithm is the least time-efficient. It takes a long time to train the model and use the model to predict, which indicates it will take more hours to obtain the cross validation result. Therefore, we use 10-fold cross validation to select an appropriate classification algorithm among other algorithms. The average metrics of classification algorithms in the training dataset are depicted in Table 3,4,5,6 and 7. As a result, logistic regression algorithm is the most time-efficient algorithm and also generates the satisfying performance.

Table 3: Average Metric of Logistic Regression if cutoff is from 0 to 1

	Ordinary Cluster	Popular Cluster
Accuracy	0.6289	0.6400
Sensitivity	0.5757	0.6888
Specificity	0.6389	0.6201
AUC	0.7243	0.8055

Table 4: Average Metric of Decision Tree if cutoff is from 0 to 1

	Ordinary Cluster	Popular Cluster
Accuracy	0.4966	0.6666
Sensitivity	0.4894	0.4698
Specificity	0.4980	0.7467
AUC	0.4861	0.6852

Table 5: Average Metric of Naive Bayesian if cutoff is from 0 to 1

	Ordinary Cluster	Popular Cluster
Accuracy	0.6219	0.6274
Sensitivity	0.5741	0.6957
Specificity	0.6309	0.5995
AUC	0.7148	0.7838

Table 6: Average Metric of Adaboost if cutoff is from 0 to 1

	Ordinary Cluster	Popular Cluster
Accuracy	0.6385	0.6888
Sensitivity	0.5827	0.6039
Specificity	0.6490	0.7233
AUC	0.7091	0.7959

Table 7: Average Metric of MLP if cutoff is from 0 to 1

	Ordinary Cluster	Popular Cluster
Accuracy	0.6237	0.6272
Sensitivity	0.5853	0.6444
Specificity	0.6310	0.6202
AUC	0.7163	0.7561

We seek the proper cutoff by the means of 10-fold cross validation and find cutoff = 0.2 is an appropriate cutoff value for the logistic regression algorithm. Metric Results of logistic regression in the test dataset when the cutoff is 0.5 and 0.2 are depicted in Table 8 and Table 9.

Table 8: Metric of Logistic Regression if cutoff = 0.5

	Ordinary Cluster	Popular Cluster
Accuracy	0.8439	0.7938
Sensitivity	0.9981	0.9472
Specificity	0.0121	0.3567

Table 9: Metric of Logistic Regression if cutoff = 0.2

	Ordinary Cluster	Popular Cluster
Accuracy	0.6849	0.6714
Sensitivity	0.6847	0.5939
Specificity	0.6802	0.8918

Hence, we choose logistic regression algorithm with the cutoff equal to 0.2 and obtain the classification result.

5.2 Regression Result

The root mean square errors (RMSE) of regression algorithms which are linear regression (LR), decision tree (DTREE) and support vector regression (SVR) in the test dataset are depicted in Table 10. The RMSE results which are ceiled, floored and rounded are also included. As a result, SVR generates the lowest RMSE and the flooring RMSE is the lowest among all four kinds of RMSEs. The root mean square errors of SVR in two clusters are depicted in Table 11.

Table 10: Metric of Regression Algorithm

	LR	DTREE	SVR
RMSE	0.9942	1.0143	0.8725
Ceiling RMSE	1.2415	1.2185	1.1947
Flooring RMSE	0.8433	0.8379	0.8255
Rounding RMSE	0.9987	0.9578	0.8504

Table 11: Metric of SVR in Two Clusters

	Ordinary	Popular	Combined
RMSE	0.7937	1.2040	0.8725
Ceiling RMSE	1.1453	1.4263	1.1947
Flooring RMSE	0.7443	1.1623	0.8255
Rounding RMSE	0.7625	1.2113	0.8504

We sum the hourly renting count of all 50 stations by hour in the course of day and compare the actual count and predict count in each cluster and combined test dataset during September 2015. As observed in Figure 7, 8 and 9, though the prediction overestimates the hourly demand on the bike-sharing system level, yet the prediction captures the trend of the hourly demand in the course of day quite well.

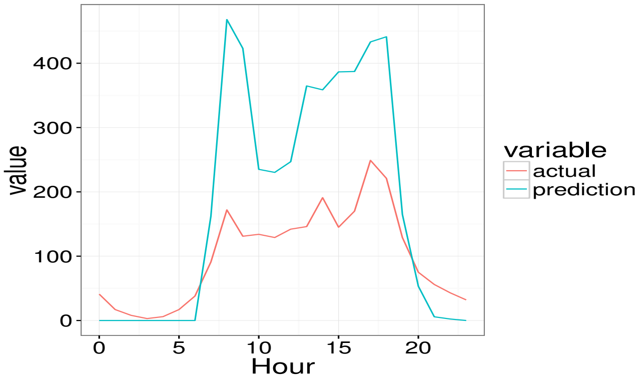

Figure 7: Comparison of actual renting count and predict renting count in the Course of Day in the Ordinary Cluster.

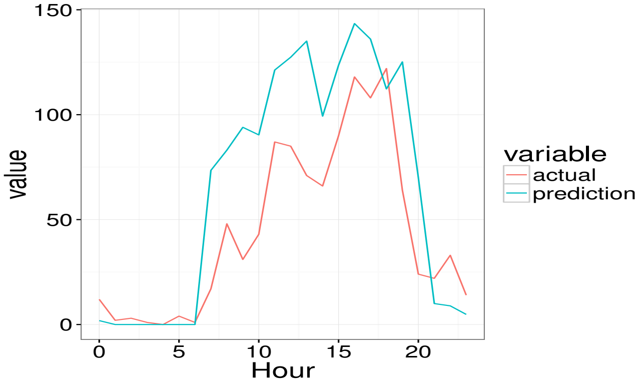

Figure 8: Comparison of actual renting count and predict renting count in the course of day in the Popular Cluster.

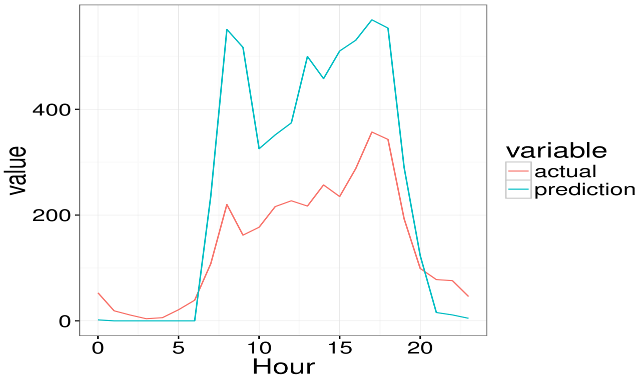

Figure 9: Comparison of actual renting count and predict renting count in the course of day in the combined test dataset.

5.3 AR Model Comparison

Following the scheme of AR(1) and AR(2) model, We apply the logistic regression algorithm with the cutoff equal to 0.2 and the support vector regression algorithm to predict the hourly demand of bikes for each station from July 2015 to March 2016. The root mean square errors of AR(1) and AR(2) models are depicted in Table 12,13 and 14. The RMSE is almost the same for AR(1) and AR(2) models, which may imply the information two hours ahead has insignificant influence on the hourly demand of bikes for each station.

Table 12: Metric of AR(1) and AR(2) Model in the Third Quarter in 2015

	Jul	Aug	Sep
RMSE: AR(1)	1.2460	1.0438	0.8725
RMSE: AR(2)	1.2629	1.0467	0.8842

Table 13: Metric of AR(1) and AR(2) Model in the Fourth Quarter in 2015

	Oct	Nov	Dec
RMSE: AR(1)	0.7261	0.4860	0.3740
RMSE: AR(2)	0.7275	0.4913	0.3680

Table 14: Metric of AR(1) and AR(2) Model in the First Quarter in 2016

	Jan	Feb	Mar
RMSE: AR(1)	0.3290	0.4778	0.6487
RMSE: AR(2)	0.3255	0.4881	0.6562

6. DISCUSSION

We refer to the concept of the autoregressive models to form the datasets for model training and testing, and construct the hybrid machine learning model including clustering, classification and regression techniques to predict the hourly demand of bikes in each station. The evaluation section indicates that our prediction result overestimates the actual demand in the course of day generally. It is mainly caused by that fact that we regard the specificity more important than the sensitivity in the classification stage, which a proportion of zero records are classified to be non-zero records. Therefore, in our future work, we will focus on the following enhancement:

1. Classification model enhancement: Facing the imbalanced problem, we will improve the classification results by tuning the parameters in the classification models. For example, we will seek a better cutoff to achieve a better tradeoff between sensitivity and specificity, tone the parameters in the kernel function after finding the solution to reduce the calculation time of SVM greatly, and apply random forest algorithm instead of decision tree. Also we will ensemble the classification results from different algorithms to boost the performance.

2. Autoregressive model enhancement: The current results of the AR(1) model and AR(2) model are very similar. It is mainly caused by the fact we have not analyzed the autocorrelation function (ACF) and partial autocorrelation function (PACF) in our data. We will conduct these analysis to figure out the most significant time lags. Another direction is to refer to the concept of other time series analysis model, such as Moving-Average model and Gauge model.

7. CONCLUSION

In summary, the project constructs a hybrid 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. Though the prediction overestimates the actual demand, our prediction result capture the trend of hourly demand in the course of day well in the nine months we cover, and our model will offer insights to study the rebalancing problem between supply and demand and the mobility in the city of Pittsburgh.

8. ACKNOWLEDGMENTS

We would like to thank Dr. Yu-Ru Lin for her instructions and suggestions in each milestone of the project sincerely. We also greatly appreciate the three groups sharing their thinking on the project in the discussion. The data employed in this project is from Healthy Ride, National Climatic Data Center and the government of the city of Pittsburgh. We appreciate for people and organization to record, organize and share those data.

9. REFERENCES

[1] e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien. https://cran.r-project.org/web/packages/e1071/index.html, 2015. [Online; accessed 12-December-2016].

[2] monmlp: Monotone multi-layer perceptron neural network. https://cran.r-project.org/web/packages/monmlp/index.html, 2015. [Online; accessed 12-December-2016].

[3] rpart: Recursive partitioning and regression trees. https://cran.r-project.org/web/packages/rpart/index.html, 2015. [Online; accessed 12-December-2016].

[4] ada: The r package ada for stochastic boosting. https://cran.r-project.org/web/packages/ada/index.html, 2016. [Online; accessed 12-December-2016].

[5] All you need to know about healthy ride in steel city. http://picso.org: 8889/˜classinfovis2016fall/projects/group-11/, 2016. [Online; accessed 12-December-2016].

[6] City of pittsburgh holiday schedule. http://apps. pittsburghpa.gov/pcsc/2016_Holiday_Schedule.pdf, 2016. [Online; accessed 12-December-2016].

[7] Healthy ride data. https://healthyridepgh.com/data/, 2016. [Online; accessed 12-December-2016].

[8] National climatic data center. https://www.ncdc.noaa.gov/, 2016. [Online; accessed 12-December-2016].

[9] J. Abonyi and B. Feil. *Cluster analysis for data mining and system identification*. Springer Science & Business Media, 2007.

[10] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[11] P.-N. Tan et al. *Introduction to data mining*. Pearson Education India, 2006.

[12] P. Vogel and D. C. Mattfeld. Strategic and operational planning of bike-sharing systems by data mining–a case study. In *International Conference on Computational Logistics*, pages 127–141. Springer, 2011.

10. AUTHOR CONTRIBUTIONS STATEMENT

In this project, all three authors contributed the ideas of model design and participated in the data collection and cleaning process. Yuwei Chen was responsible for clustering analysis. Fengtao Wu and Zhendong Wang were responsible for classification and regression analysis. Finally, all three authors evaluated the performance of model in each month. All authors reviewed the manuscript.