Data Mining for Business (BUDT758T)

Project Title: _Hourly Demand Predicting and unbalanced stations classified for Capital Bikeshare_

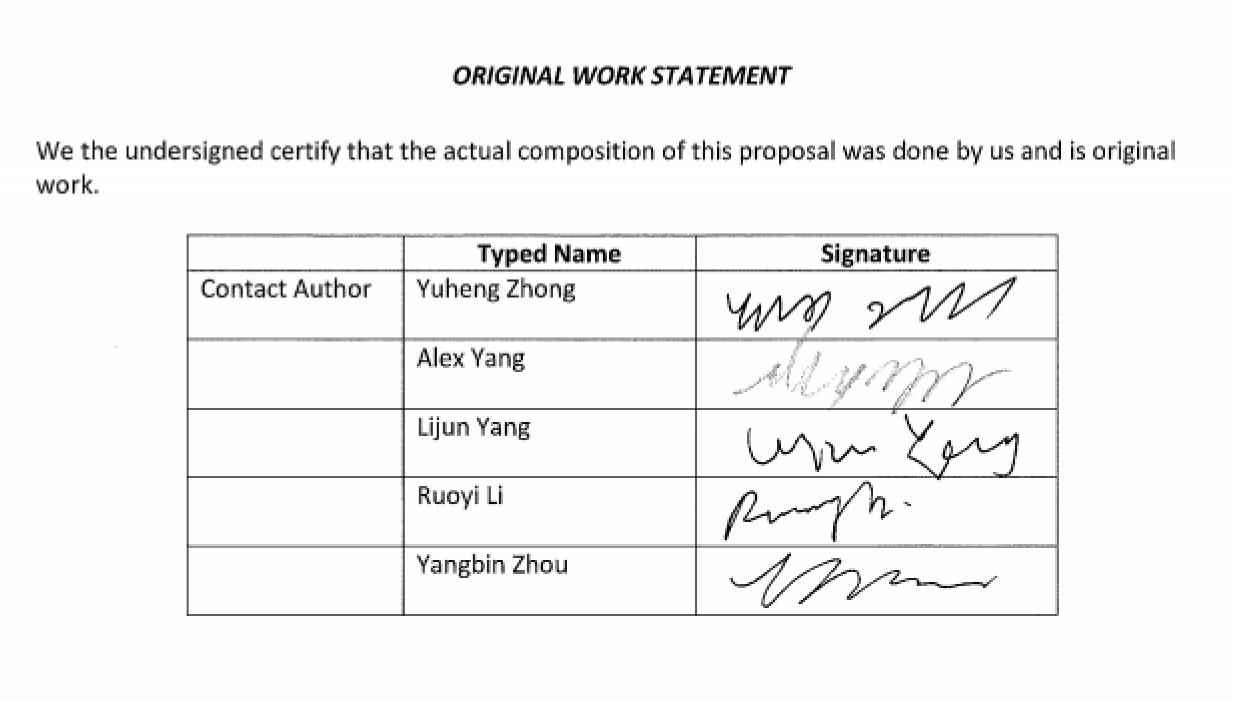
Team Members: ______________Alex Yang 113542701___________________________

________________Lijun Yang 115819542___________________________

________________Ruoyi Li 116232401___________________________

________________Yangbin Zhou 114872418______________________

________________Yuheng Zhong 116226189______________________



ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	**Signature**
Contact Author	Yuheng Zhong	
	Alex Yang	
	Lijun Yang	
	Ruoyi Li	
	Yangbin Zhou	

II. **Executive Summary**

According to the characteristics of sharing economy, data mining and predictive analysis would be great helpful and important for those sharing economy companies like Capital Bikeshare Inc, to generate strategies to maximize their profits. This research focus on analyzing and predicting the dynamic demand based on time, area, and other environmental factors for Capital Bikeshare Inc, and proposing recommends about how to maximize profits based on the demand analysis and predicted result. In addition, we will analyze nearby stations for return and provide return recommendation.

For bike sharing industry, there are two ways to increase the profit-- Reducing the cost by solving the unbalanced station issues and increasing the revenue by using dynamic price based on demand changing. Unbalanced bikeshare station is one of significant issues in bike sharing industry, which can limit the revenue and increase the cost. Either too full or too empty would make a bad influence on customer experience and result in extra cost for capital bikeshare company to rebalance these stations. In past years, about 55% of operational cost came from rebalancing stations in Capital Bikeshare INC. The idea is that when users are finishing their bike trips and going to return their bikes to a certain station, we would compare the utilization percentage of that station and nearby stations of that station. If the utilization percentage of the initial destination station is higher than a threshold, we just let users return their bikes. However, if the initial destination station's usage percentage is lower than the threshold, the system would analyze the utilization for all-nearby stations in a certain distance parameter and recommend the high utilization nearby stations to the users to return their bikes to get bonus discount on fare. The utilization of each station is calculated by *(demand - return)/total slot* and set a cutoff to classify those unbalanced **(In this business case, we focus on too empty)**. Area demand predicting model will be activated once the actual utilization exceeds the cutoff, those areas with high demand and low return predicted result will be marked as "Low price return area" on the bikeshare app to encourage customers to drop bikes into these areas with a low rate. New stations building plan can also be support by this predictive model. In additional, dynamic price strategy can be generated based on demand predictive model to increase revenue, just like what Uber did. When the predicted demand trend is decreasing, we could probably low the price in order to simulate bike usage, and when the predicted demand trend is increasing to a certain point, we could increase the price to maximize profit.

III. Data Description

Data source:

- https://www.capitalbikeshare.com/system-data
- http://opendata.dc.gov/datasets/capital-bike-share-locations/data
- https://i-weather.com/weather/washington/history/monthly-history/?gid=4140963&station=19064&month=1&year=2017&language=english&country=us-united-states
- https://dchr.dc.gov/page/holiday-schedules

Data description:

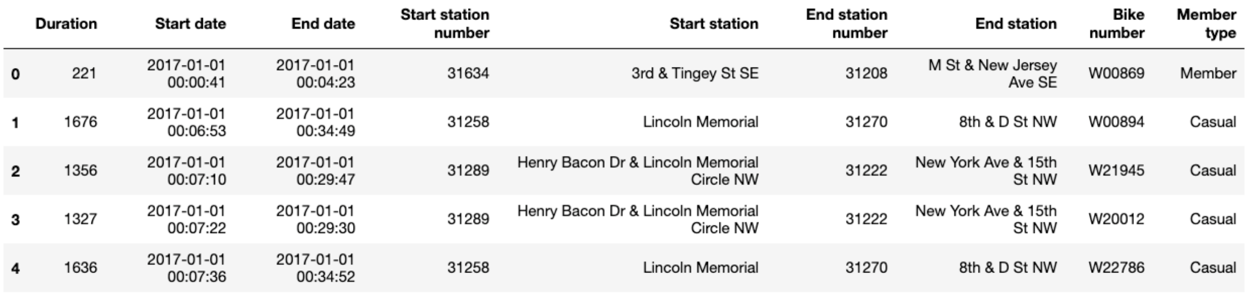
- *demand(numerical): count of total rental bikes including both casual and registered.*
- season(categorical) : 1:spring, 2:summer, 3:fall, 4:winter.
- date(categorical) : date and hour time, resampled by hour.
- holiday and weekend(categorical) : 1:holiday and weekend, 0:not holiday and weekend.
- weather(categorical) : brief description of weather.
- temperature(numerical) : average temperature in Celsius.
- humidity(numerical): humidity level, percentage.
- wind_speed(numerical): average wind speed, km/h.

Sample size and number of variables:

1. Capital Bikeshare data in 2017: 3757777 rows, 10 variables.
2. Weather data: 10678 rows, 10 variables.
3. Station data: 539 rows, 18 variables.

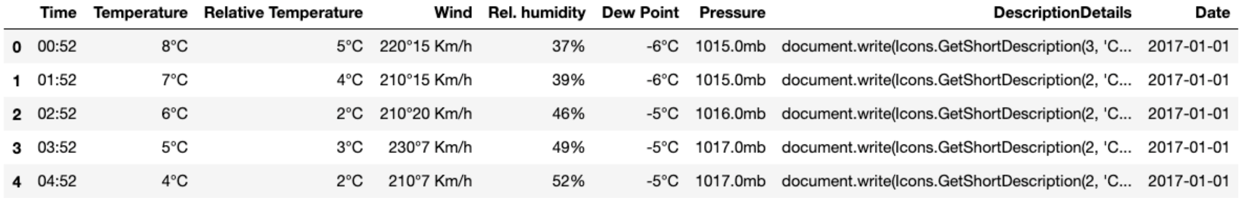
Sample

1. Capital Bikeshare original data in 2017



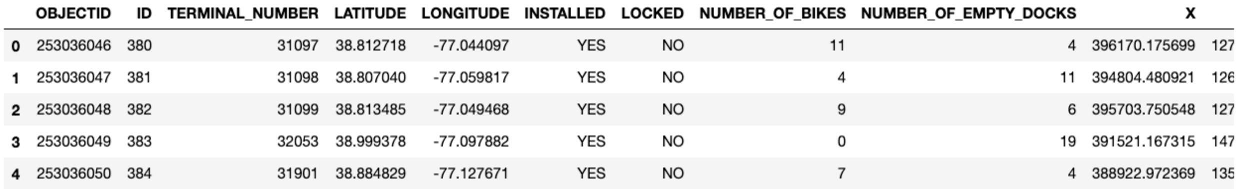
	Duration	Start date	End date	Start station number	Start station	End station number	End station	Bike number	Member type
0	221	2017-01-01 00:00:41	2017-01-01 00:04:23	31634	3rd & Tingey St SE	31208	M St & New Jersey Ave SE	W00869	Member
1	1676	2017-01-01 00:06:53	2017-01-01 00:34:49	31258	Lincoln Memorial	31270	8th & D St NW	W00894	Casual
2	1356	2017-01-01 00:07:10	2017-01-01 00:29:47	31289	Henry Bacon Dr & Lincoln Memorial Circle NW	31222	New York Ave & 15th St NW	W21945	Casual
3	1327	2017-01-01 00:07:22	2017-01-01 00:29:30	31289	Henry Bacon Dr & Lincoln Memorial Circle NW	31222	New York Ave & 15th St NW	W20012	Casual
4	1636	2017-01-01 00:07:36	2017-01-01 00:34:52	31258	Lincoln Memorial	31270	8th & D St NW	W22786	Casual

2. Weather data



	Time	Temperature	Relative Temperature	Wind	Rel. humidity	Dew Point	Pressure	DescriptionDetails	Date
0	00:52	8°C	5°C	220°15 Km/h	37%	-6°C	1015.0mb	document.write(Icons.GetShortDescription(3, 'C...	2017-01-01
1	01:52	7°C	4°C	210°15 Km/h	39%	-6°C	1015.0mb	document.write(Icons.GetShortDescription(2, 'C...	2017-01-01
2	02:52	6°C	2°C	210°20 Km/h	46%	-5°C	1016.0mb	document.write(Icons.GetShortDescription(2, 'C...	2017-01-01
3	03:52	5°C	3°C	230°7 Km/h	49%	-5°C	1017.0mb	document.write(Icons.GetShortDescription(2, 'C...	2017-01-01
4	04:52	4°C	2°C	210°7 Km/h	52%	-5°C	1017.0mb	document.write(Icons.GetShortDescription(2, 'C...	2017-01-01

3. Station data



	OBJECTID	ID	TERMINAL_NUMBER	LATITUDE	LONGITUDE	INSTALLED	LOCKED	NUMBER_OF_BIKES	NUMBER_OF_EMPTY_DOCKS	X	
0	253036046	380	31097	38.812718	-77.044097	YES	NO	11	4	396170.175699	127
1	253036047	381	31098	38.807040	-77.059817	YES	NO	4	11	394804.480921	126
2	253036048	382	31099	38.813485	-77.049468	YES	NO	9	6	395703.750548	127
3	253036049	383	32053	38.999378	-77.097882	YES	NO	0	19	391521.167315	147
4	253036050	384	31901	38.884829	-77.127671	YES	NO	7	4	388922.972369	135

Why the data are of interest

The goal of this project is to perform data analysis to help capital bikeshare INC, increasing their profits. As what we mentioned before, profits can be increased by reducing the cost--solving unbalanced stations issue and increasing the revenue--generating dynamic price strategy. By looking at Bikeshare transaction and station information data set, we could figure out the utilization of each station and area. Thus, they are great helpful for us to classify unbalanced/balanced station. According to the supply-demand curve, dynamic price strategy should be generated based on the demand changing. Unlike other public transportation, weather can be a key factor which affect the market demand of bike share a lot. For example, few people would like to ride bike in a heavy raining day.

In conclusion, market demand analysis is the key to support every business strategy and recommends in this research. Bikeshare transaction data and weather data can help us to understand how demand changing in a specific period. Bikeshare transaction data and bike station data can be used to classify balanced/unbalanced stations/area.

III. Research Questions

How do we figure out a dynamic pricing strategy by predicting demand in a specific period?

According to principles of the supply and demand curve and price discriminations, dynamic price strategy can be a good way to increase the total revenue based on the changing demand. Thus, a total demand predictive model needs to be built. Once we can predict the demand in different time period, we can generate a well-designed dynamic price strategy based on this model.

Market demand can be predicted within a specific period by looking at date fields like month, day, hour, weekend, and season. We expect that people may be more likely to travel during weekdays over weekends and summertime over winter. A much more accurate prediction could possibly be surmised from weather data. After considering these factors, the resulting prediction is the demand for all stations at a certain datetime. We may use this predictive model to design our dynamic pricing strategy. We can generate a general price for all stations at certain time periods to encourage people to use the bikeshare.

How could we classify balanced/unbalanced stations/areas?

Unbalanced is one of the significant issues in bikeshare industry. If the station is too empty, customers would be angry because they cannot find a bike to ride. If the station is too full, it is not good for company because the most of bikes are not used. On the other words, the most of bikes cannot make money from customers in the too full station. Unbalanced stations also cost company extra money to rebalance them. Thus, we should find out a way to classify which station/area is unbalanced and generate appropriate strategy and recommendations based on what we find. First of all, we should define a formula to create a new dummy variable called "Balanced" with 1 for balanced station and 0 for unbalanced station. And then we should find a reasonable cutoff.

The size of low price/zone return zone can be decided by looking at the average distance between each station. This average distance can be the start point to find out the appropriate zone size.

IV. **Methodology**

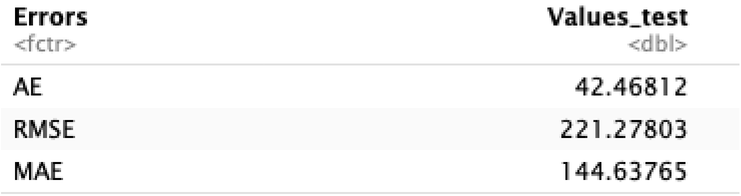
1. Linear regression is a basic and commonly used type of predictive analysis, and it is about to find a mapping function from input variables to a continuous output variable. The general idea of regression is to test two things: (1) Is a set of predictors doing well in predicting? (2) Which variables and how do they represent the magnitude and sign of the outcome variable? Thus, we might be able to use regression to explain the relationship between demand and one or more other independent variables in our case.
2. Random Forest is an ensemble learning method for classification and regression. The model is a type of additive model that makes predictions by combining decisions from a sequence of base models, and it operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. In our project, we might be able to use this method to solve regression problem.
3. Support vector machine (SVM) analysis is a popular machine learning tool for classification and regression. In SVM for regression problem, it is about to fit a model to predict a quantity for future and the data point would be as close as possible to the hyperplane unlike SVM for classification. Thus, this method might be applied to our project.
4. Employ Euclidean distance to calculate the distance between each station, and unify the length units of distance.
5. Panel fixed effect can be used to build a classification model to classify a Balanced/Unbalanced station in a specific time period. In this case, both “Station Number” and “Date” can be assigned to index to build the plm model. Thus, we can find out which variable is significant for predicting demand and return number for a specific station at a specific time period.
6. GLM model with Poisson distribution will be used to predict the demand and return number for a specific station at a specific time period
7. Usage can be calculated by (demand-return)/total slot, usage variable will be created to every history record. By looking at the summary of usage, we can decide a reasonable cutoff to classify balanced/unbalanced station for dataset.

V. Results and Findings

Total demand prediction model & Dynamic Price:

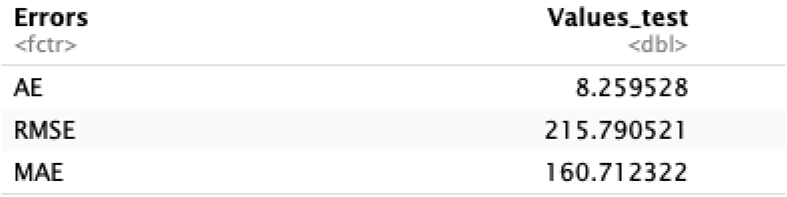
Model evaluation

Linear regression:



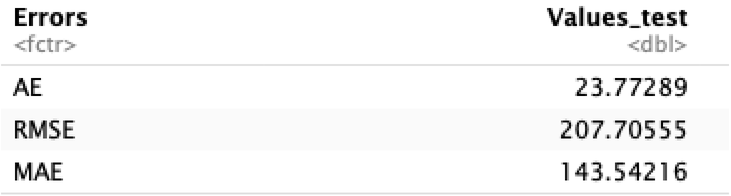
Errors	Values_test
<fctr>	<dbl>
AE	42.46812
RMSE	221.27803
MAE	144.63765

Linear regression with log transformation:



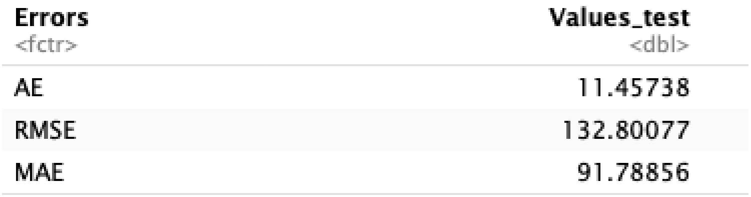
Errors	Values_test
<fctr>	<dbl>
AE	8.259528
RMSE	215.790521
MAE	160.712322

SVM:



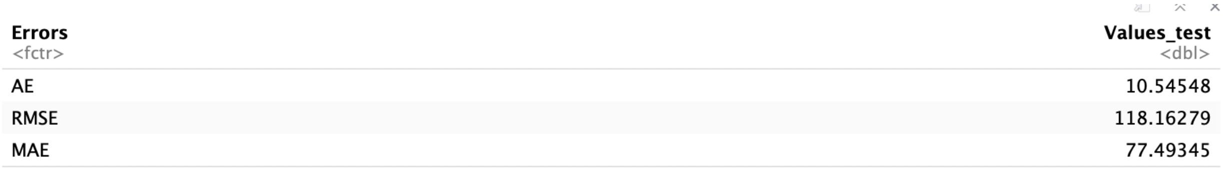
Errors	Values_test
<fctr>	<dbl>
AE	23.77289
RMSE	207.70555
MAE	143.54216

Random forest:



Errors	Values_test
<fctr>	<dbl>
AE	11.45738
RMSE	132.80077
MAE	91.78856

Tuned random forest:



Errors <fctr>	Values_test <dbl>
AE	10.54548
RMSE	118.16279
MAE	77.49345

The predictive model for the demand was built from using the predictors which included the datetime and weather data. Some of the more specific datetime fields included Month, Day, Hour, Weekend, Season, and Holiday. Weather data specifies whether it was cloudy, rainy, snowy, or clear. There was also wind speed in kilometers per hour, temperature in Celsius, and humidity as a percentage. Out of all the models tried, the best predictive model with regards to error measures for overall demand for a certain time period was the Random Forest model.

Low price return zone:

First, we need to calculate the distance matrix between each bikeshare station in Washington D.C. The matrix helps us to find the nearby stations of selected stations. Then we need to decide the distance parameters of each station. With the actual situation in Washington D.C that average block distance is 0.18 mile, we think that parameter being smaller than 0.54 miles makes sense for recommendation system. The output of distance matrix and nearby stations within 0.54 miles can be found in Appendix (Figure 1 & Figure 2)

Then, we define the unbalanced and balanced areas to determine the low-price return area. With employing formula: (demand - return)/number_of_slot, we can get the daily usage percentage of each station in the historical dataset. Then we managed to decide the cutoff value of usage percentage to differentiate stations into 'balanced' and 'unbalanced'. We think that usage percentage = 1 is a clear and easy-to-understand value for cutoff. In such situation, the total slot can be treated as demand buffer zone to satisfy demand of station. If percentage is equal to or larger than 1, we conclude these stations as unbalanced station, otherwise we define them as balanced.

After doing so, each station in each day is divided into two levels: 'balanced' and 'unbalanced'. We build the panel estimators for generalized linear models based on the dataset to find out which variable is significant to predict demand and return for a specific station at a specific time period. According to the output of panel data analysis (appendix-figure 3 & figure 4), daily max temperature, daily min temperature, max steady wind, total precipitation, pressure, holiday, and season are significant to predict demand in each station. Daily max temperature, daily min temperature, max steady wind, pressure, holiday, and season are significant to predict return in each station.

Based on the output of PGLM model, we built two glm model with Poisson distribution to predict demand and return in each station. We assigned predicted demand and return back to the test data set and calculated predicted “Balanced/Unbalanced” variable for each test dataset observation. According to the confusion table (Appendix-figure 5), we could find that the error rate of this model is 0.02%, which is good enough.

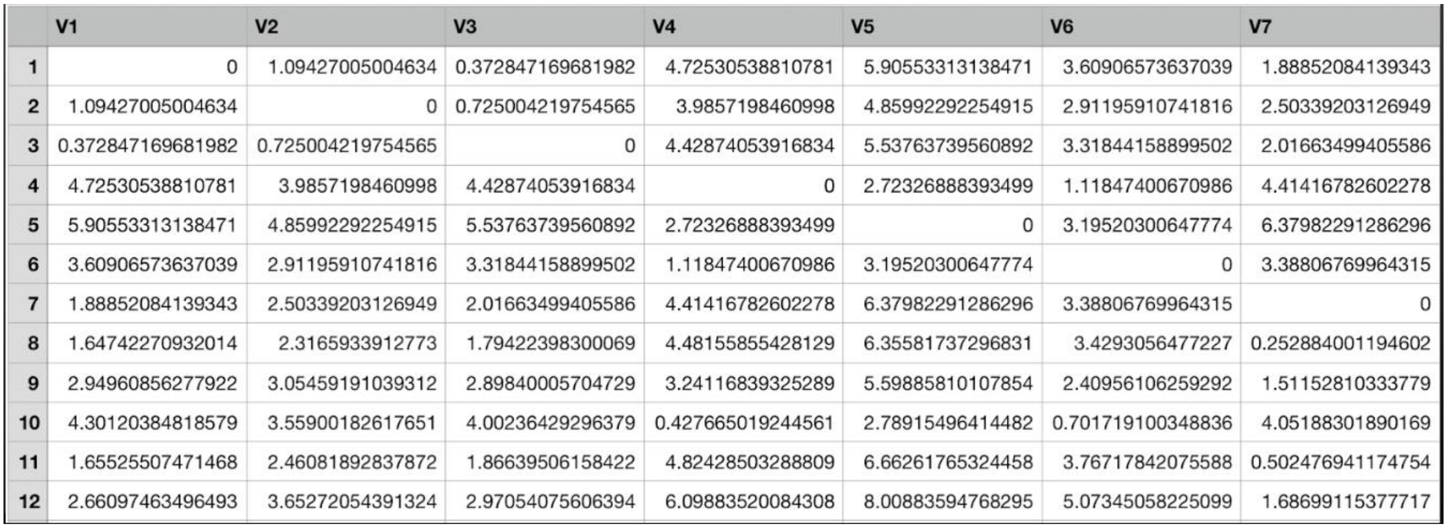
New stations distribution plan:

The panel data classification can also be used to support new stations distribution plan. We can predict the usage situation during a long time period. If we find some stations or some regions are always in the condition of unbalanced, we may consider build a new station nearby the stations or the regions.

VI. **Conclusion**

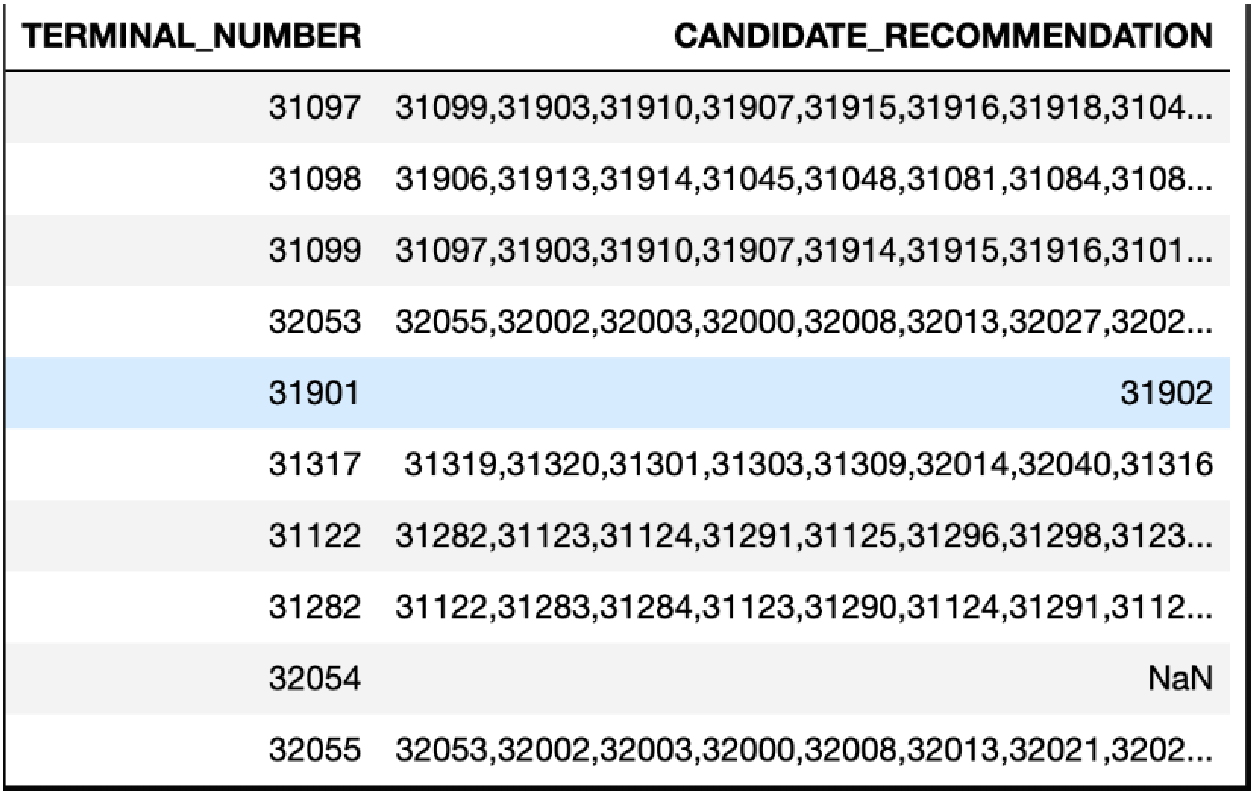
Data mining and predictive analysis can help sharing economy company like capital bikeshare a lot on generating business strategies to improve overall performance. In general, there are two ways to increase the profit, the one is increasing the revenue and the other one is reducing the cost. For bikeshare industry, dynamic price based on demand changing can be a good way to increase the total revenue. The low-price rate can encourage customer to spend money on their products and service when the market demand in low. High price can generate more revenue when the demand is high. For data analysis perspective, a total demand predict model can provide Capital Bikeshare INC, a direction to generate their dynamic price strategy. Unbalanced station is one of significant issues in bike sharing industry. Either too empty or too full will result in a bad influence in customer experience. This issue will also cost company extra money to rebalance these stations. Low price/Free return zone is the idea to encourage customers to help us rebalancing these stations. This is a win-win strategy, it helps company saving money in rebalancing stations and customer can also enjoy a low even free bike ride.

VII. **Appendix**



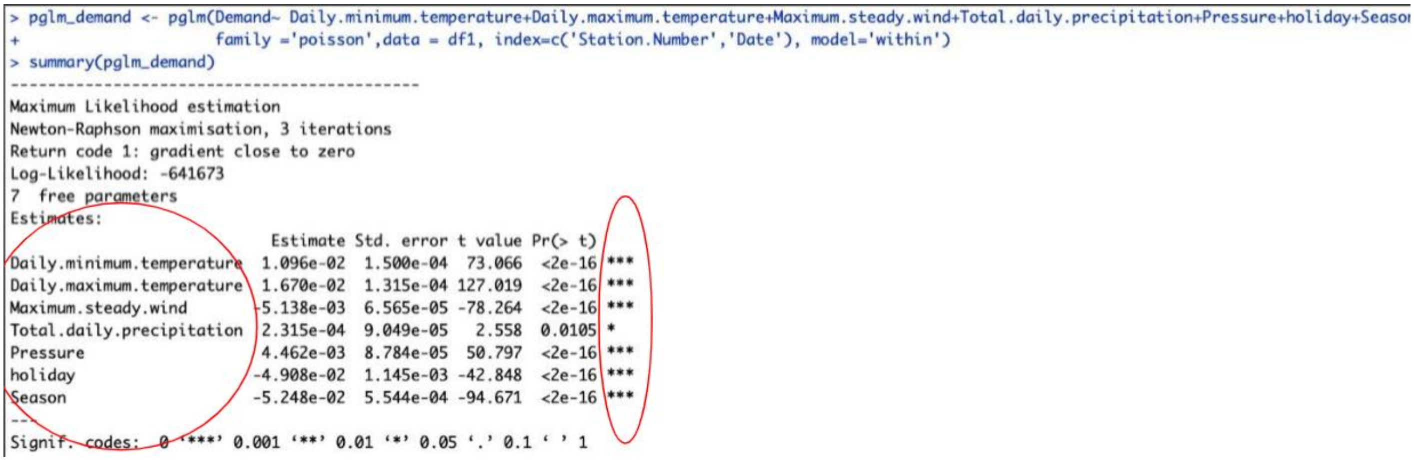
	V1	V2	V3	V4	V5	V6	V7
1	0	1.09427005004634	0.372847169681982	4.72530538810781	5.90553313138471	3.60906573637039	1.88852084139343
2	1.09427005004634	0	0.725004219754565	3.9857198460998	4.85992292254915	2.91195910741816	2.50339203126949
3	0.372847169681982	0.725004219754565	0	4.42874053916834	5.53763739560892	3.31844158899502	2.01663499405586
4	4.72530538810781	3.9857198460998	4.42874053916834	0	2.72326888393499	1.11847400670986	4.41416782602278
5	5.90553313138471	4.85992292254915	5.53763739560892	2.72326888393499	0	3.19520300647774	6.37982291286296
6	3.60906573637039	2.91195910741816	3.31844158899502	1.11847400670986	3.19520300647774	0	3.38806769964315
7	1.88852084139343	2.50339203126949	2.01663499405586	4.41416782602278	6.37982291286296	3.38806769964315	0
8	1.64742270932014	2.3165933912773	1.79422398300069	4.48155855428129	6.35581737296831	3.4293056477227	0.252884001194602
9	2.94960856277922	3.05459191039312	2.89840005704729	3.24116839325289	5.59885810107854	2.40956106259292	1.51152810333779
10	4.30120384818579	3.55900182617651	4.00236429296379	0.427665019244561	2.78915496414482	0.701719100348836	4.05188301890169
11	1.65525507471468	2.46081892837872	1.86639506158422	4.82428503288809	6.66261765324458	3.76717842075588	0.502476941174754
12	2.66097463496493	3.65272054391324	2.97054075606394	6.09883520084308	8.00883594768295	5.07345058225099	1.68699115377717

(Figure-1 Distance Matrix)



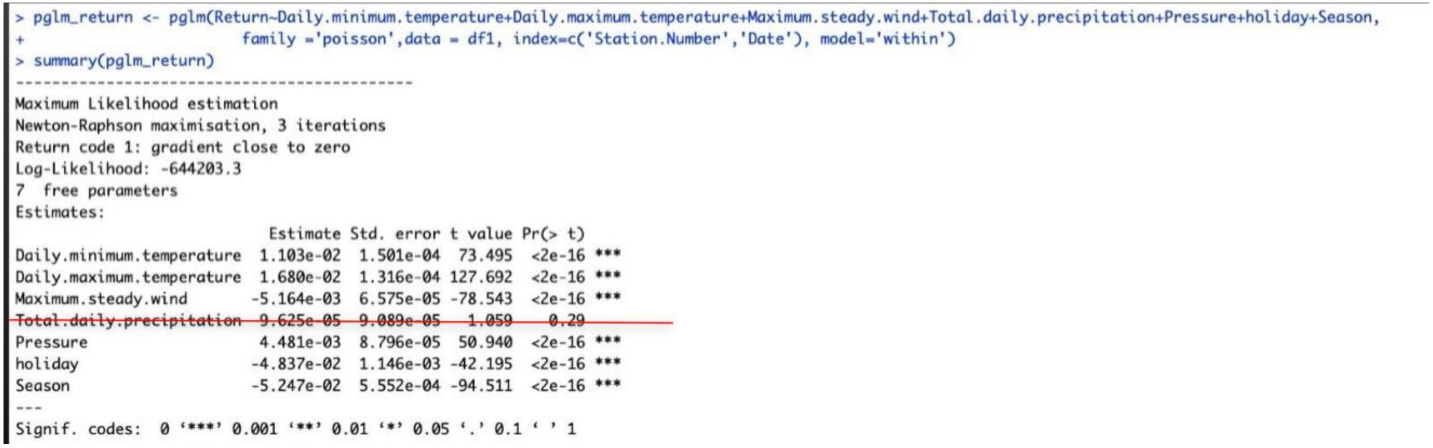
TERMINAL_NUMBER	CANDIDATE_RECOMMENDATION
31097	31099,31903,31910,31907,31915,31916,31918,3104...
31098	31906,31913,31914,31045,31048,31081,31084,3108...
31099	31097,31903,31910,31907,31914,31915,31916,3101...
32053	32055,32002,32003,32000,32008,32013,32027,3202...
31901	31902
31317	31319,31320,31301,31303,31309,32014,32040,31316
31122	31282,31123,31124,31291,31125,31296,31298,3123...
31282	31122,31283,31284,31123,31290,31124,31291,3112...
32054	NaN
32055	32053,32002,32003,32000,32008,32013,32021,3202...

(Figure-2 Nearby station of each station within 0.54 miles)



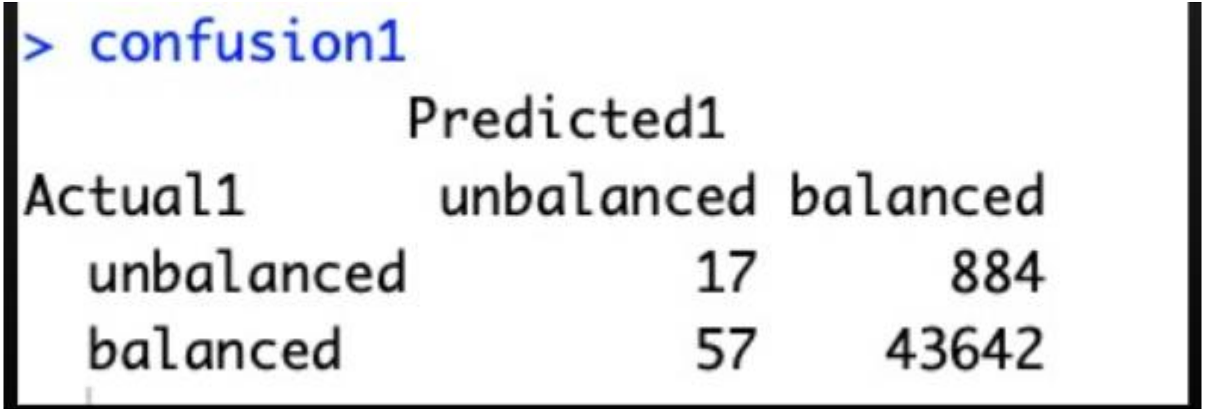
```
> pglm_demand <- pglm(Demand~ Daily.minimum.temperature+Daily.maximum.temperature+Maximum.steady.wind+Total.daily.precipitation+Pressure+holiday+Seaso
+                    family ='poisson',data = df1, index=c('Station.Number','Date'), model='within')
> summary(pglm_demand)
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 3 iterations
Return code 1: gradient close to zero
Log-Likelihood: -641673
7  free parameters
Estimates:
                           Estimate Std. error t value Pr(> t)    
Daily.minimum.temperature  1.096e-02  1.500e-04  73.066  <2e-16 ***
Daily.maximum.temperature  1.670e-02  1.315e-04 127.019  <2e-16 ***
Maximum.steady.wind       -5.138e-03  6.565e-05 -78.264  <2e-16 ***
Total.daily.precipitation  2.315e-04  9.049e-05   2.558  0.0105 *  
Pressure                   4.462e-03  8.784e-05  50.797  <2e-16 ***
holiday                   -4.908e-02  1.145e-03 -42.848  <2e-16 ***
Season                    -5.248e-02  5.544e-04 -94.671  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Figure 3--PGLM output of demand)



```
> pglm_return <- pglm(Return~Daily.minimum.temperature+Daily.maximum.temperature+Maximum.steady.wind+Total.daily.precipitation+Pressure+holiday+Season,
+                    family ='poisson',data = df1, index=c('Station.Number','Date'), model='within')
> summary(pglm_return)
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 3 iterations
Return code 1: gradient close to zero
Log-Likelihood: -644203.3
7  free parameters
Estimates:
                           Estimate Std. error t value Pr(> t)    
Daily.minimum.temperature  1.103e-02  1.501e-04  73.495  <2e-16 ***
Daily.maximum.temperature  1.680e-02  1.316e-04 127.692  <2e-16 ***
Maximum.steady.wind       -5.164e-03  6.575e-05 -78.543  <2e-16 ***
~~Total.daily.precipitation  9.625e-05  9.089e-05   1.059    0.29~~
Pressure                   4.481e-03  8.796e-05  50.940  <2e-16 ***
holiday                   -4.837e-02  1.146e-03 -42.195  <2e-16 ***
Season                    -5.247e-02  5.552e-04 -94.511  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Figure 4 --PGLM output of return)



```
> confusion1
            Predicted1
Actual1      unbalanced balanced
  unbalanced         17      884
  balanced           57    43642
```

(Figure 5--Confusion table)