SHARING BIKE

MGMT 6160 Final project

Xuyang Bai, Mengyan Zhu

Table of Content

1. Initial Purpose and Introduction

Bike sharing system introduces a modern bike rental method. It enables users to rent and return a bike whenever they need and wherever they are. We are greatly impressed by the successful dock-less shared bikes system operated in China. Since 2014, when this new sharing bike program came into market, there is a rapid growth in the number of bike users. To illustrate, in Beijing, the usage of bicycle experienced a growth from 5.5% to 11.6 % with these new sharing bikes (“Bicycle-sharing system”, n.d. para. 4).

From our personal experiences, this new system is very convenient, cheap and easy to use. We are able to rent a bike at anywhere and anytime due to the dockless design. In recent year, however, it is obviously that the supply of the sharing bikes is far more than the demand. And only a small percentage of users would like to ride a bike during the winter season. Thus, predicting the demand of the sharing bike is important and necessary for local government, bike sharing companies, and city planning and administration department. Providing appropriate number of docks can help bike sharing companies to maximize its revenue and minimize cost.

While in USA, the major type of bike sharing is docked public bike system. There are many factors that determine whether people would like to use sharing bikes. For example, people are more likely to ride bikes when the weather condition is good. Thus, in spring and fall when the temperature is neither too low nor too high, the usage of the bike rental may be higher than that of summer and winter. Also, the location where people could rent and return the bike is also an important factor. Usually, the location near subway stations and shopping malls are in high demand. People love to rent a bike for a short trip to save time and effort. What’s more, we need to take the network effect into consideration. Current customers using sharing bikes may easily bring more and more future customers to the market, if they find it convenient. Thus, we are looking for a dataset that might include information about weather, the number of rental bikes in different time period, and various types of bike users. Hopefully, using our analysis in this paper, we can help relevant organizations to have a clearer view of the demand of sharing bike.

2. Data Source Description

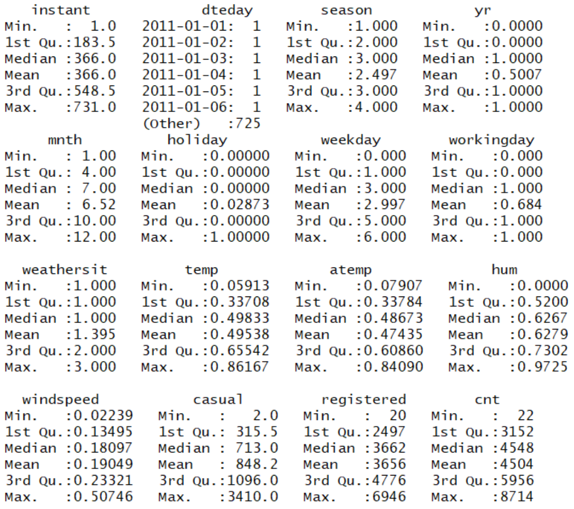
The bike sharing data is available on the UCI machine learning repository (Hadi Fanaee-T, 2013). The first dataset describes two aspects: daily count of bike rentals with different user groups in Capital bike-share system; and the corresponding weather condition from 2011 to 2012. Second dataset records the hourly count of bike rentals, and the rest information is same as the first dataset.

Since the information in the two datasets are similar, we mainly use the daily record dataset to explore the relationship between counts of bike rentals and various factors. And using the hourly record dataset only when we analyze the data by the hour level.

The data contains information about the seasonal and time related variables. To illustrate, it has columns including “Date”, “Month”, “Year”, “Season”, “Workingday”, “Weekday”, “Holiday” and “Hour”, which enable us to analyze the number of bike rental by different time measurement levels and uncover the pattern of bike usage over time. Also, the dataset records the weather condition such as temperature and feeling temperature in Celsius after applying normalized transformation, the normalized humidity and normalized wind speed, and a categorical variable named “weathersit” that indicates the characteristics of the weather, including cloudy, rainy, snow etc. Thus, we think the data is sufficient for us to discover the usage pattern and predict the demand of bike rental.

3. Exploratory Analysis

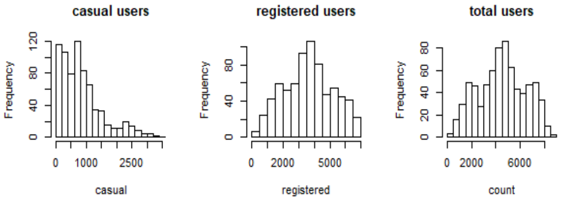
3.1 Overall Understanding



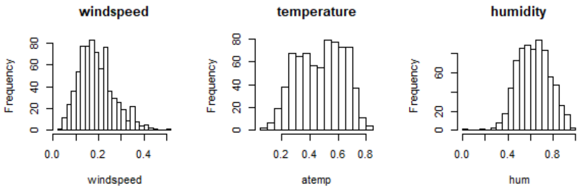
```
    instant            dteday        season           yr
 Min.   :  1.0   2011-01-01:  1   Min.   :1.000   Min.   :0.0000
 1st Qu.:183.5   2011-01-02:  1   1st Qu.:2.000   1st Qu.:0.0000
 Median :366.0   2011-01-03:  1   Median :3.000   Median :1.0000
 Mean   :366.0   2011-01-04:  1   Mean   :2.497   Mean   :0.5007
 3rd Qu.:548.5   2011-01-05:  1   3rd Qu.:3.000   3rd Qu.:1.0000
 Max.   :731.0   2011-01-06:  1   Max.   :4.000   Max.   :1.0000
                 (Other)   :725
      mnth          holiday           weekday        workingday
 Min.   : 1.00   Min.   :0.00000   Min.   :0.000   Min.   :0.000
 1st Qu.: 4.00   1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.000
 Median : 7.00   Median :0.00000   Median :3.000   Median :1.000
 Mean   : 6.52   Mean   :0.02873   Mean   :2.997   Mean   :0.684
 3rd Qu.:10.00   3rd Qu.:0.00000   3rd Qu.:5.000   3rd Qu.:1.000
 Max.   :12.00   Max.   :1.00000   Max.   :6.000   Max.   :1.000

   weathersit         temp              atemp              hum
 Min.   :1.000   Min.   :0.05913   Min.   :0.07907   Min.   :0.0000
 1st Qu.:1.000   1st Qu.:0.33708   1st Qu.:0.33784   1st Qu.:0.5200
 Median :1.000   Median :0.49833   Median :0.48673   Median :0.6267
 Mean   :1.395   Mean   :0.49538   Mean   :0.47435   Mean   :0.6279
 3rd Qu.:2.000   3rd Qu.:0.65542   3rd Qu.:0.60860   3rd Qu.:0.7302
 Max.   :3.000   Max.   :0.86167   Max.   :0.84090   Max.   :0.9725

   windspeed           casual          registered        cnt
 Min.   :0.02239   Min.   :   2.0   Min.   :  20   Min.   :  22
 1st Qu.:0.13495   1st Qu.: 315.5   1st Qu.:2497   1st Qu.:3152
 Median :0.18097   Median : 713.0   Median :3662   Median :4548
 Mean   :0.19049   Mean   : 848.2   Mean   :3656   Mean   :4504
 3rd Qu.:0.23321   3rd Qu.:1096.0   3rd Qu.:4776   3rd Qu.:5956
 Max.   :0.50746   Max.   :3410.0   Max.   :6946   Max.   :8714
```

From the descriptive statistics, we can find that the dataset includes 16 variables. The data type of most variables are continuous. And there is only one categorical variable which is date of a day (dteday). The target variable for prediction is the count of rental bikes including both casual and registered riders. Since month, holiday, weekday, and working day are categorical variables, we change their data types into factor for further analysis.

3.2 Possible Transformation

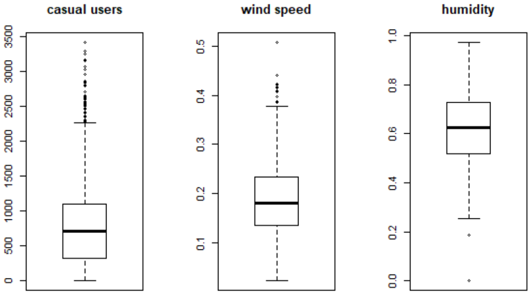


From the histogram of counts, we find that the distribution of casual users is highly right skewed, thus there might be some outliers in “casual” variable. We can apply transformation on the “casual” column, if we want to predict the demand of casual users. Compared to log, square root and exponential transformation, we find that square root can make the casual group become more normally distributed.



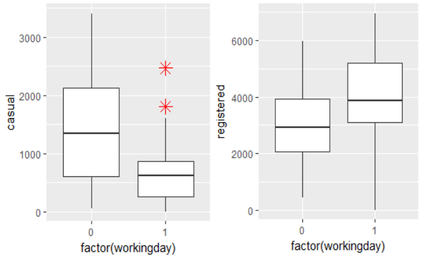
And from the three graphs above, we can find that the distribution of wind speed and humidity are slightly skewed. Thus, we can detect the outliers in these skewed data using boxplot.

3.3 Missing value and outliers



First two graph has more outliers, compared to the third one. From the boxplot of casual users, we can see that the count of rental bikes for casual riders is regarded as outliers, when the number exceeds 2300. In second graph, when wind speed higher than 0.38, the speed is considered to be outliers.

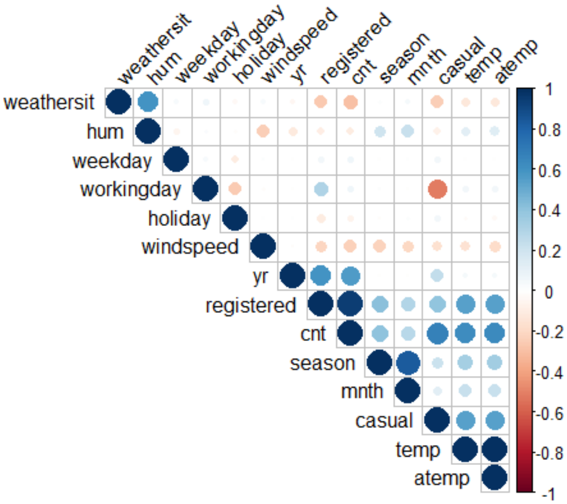
Since the count of casual bike users could be used as dependent variable for prediction. The outliers may have some reasonable explanations, for example, an abnormal high demand of casual bike rental may due to some important events happened in the city. And the outliers of wind speed may indicate some bad weather. Thus, we cannot delete all the outliers in our data, for the outliers may contain some useful information and unique insights.



These two boxplots show the comparison of the distribution of different types of users on working day (1) and non-working days (0). Casual users mainly rent sharing bikes on weekends or holidays, and they seldom use a bike on working days, because the median of usage is only about 600. And outliers in casual users group on working days may due to some special events. On the other hands, there is no big difference between the usage of bikes for registered group in working days and non-working days. But registered users would like to rent more bike on working days than weekday, which indicates that registered users select sharing bike as a way to commute. Overall, the average usage of registered group is higher than the casual users, and this result makes sense.

Both "day" and "hour" datasets do not have any missing value, so we don't need to take any further action in this part.

3.4 Find correlations between variables

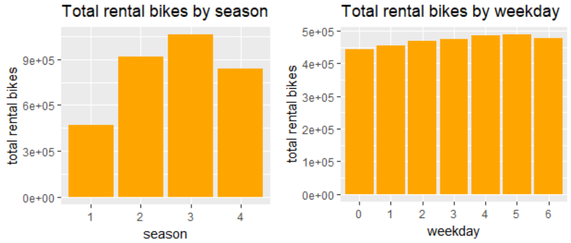


From the correlation plot, it is not surprising that the total count is highly positive correlated with the number of registered users (0.95) and casual users (0.67). Season is strongly related to month (0.83), and temperature has a strong relationship with feeling temperature (0.99). We can also find that the number of casual users has a negative correlation with working day (-0.52), which also indicates that casual users would like to rent the bikes more frequently on non-working days. "Weathersit" and humidity also have strong correlation (0.59). These highly correlated pairs may cause multi-collinearity problem and thus greatly decrease model accuracy. So we need further explore the relationship between these variables.

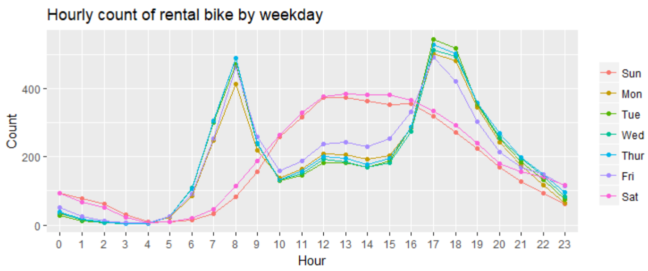
We find that many features have high correlation with total count, for example, the number of rental bikes is highly correlated to temperature and season. We can explore more on the relationship between independent and dependent variables.

We classify the independent variables into two categories: factors associated with time and factors associates with weather conditions. So we analyze the correlation between variables separately.

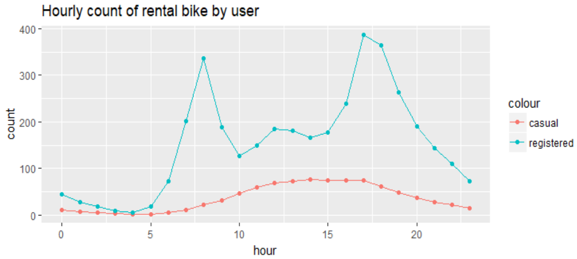
3.4.1 Detail examine of histograms and uncover potential relationships



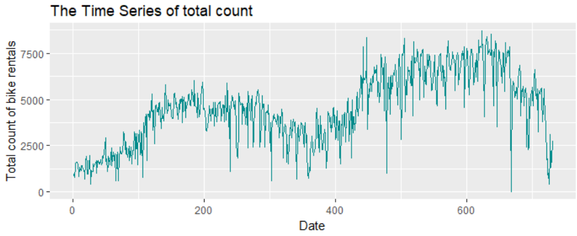
From the above histograms, we can learn that the peak usage of the rental bike happens in fall (season=3). It is quite surprising that the usage in spring (1) is lower than winter (4). And the usage of bike is almost same in each weekday, and it is slightly higher on Thursday and Friday.



Using the additional hourly bike rental information in "hour" dataset, we can learn that in working days, the peak time of using a bike is during 7-9 am and 5-6 pm, which are the regular timeframes for commute. Thus, the peak usage in these time periods indicates there is a high demand of rental bike among office staff and students. During daytime, the hourly count of rental bike is also about 200, which shows that the demand of rental is stable throughout the day. On weekends, the number of rental bikes is higher during 11am – 6pm, indicating that people mainly ride bike in the afternoon.

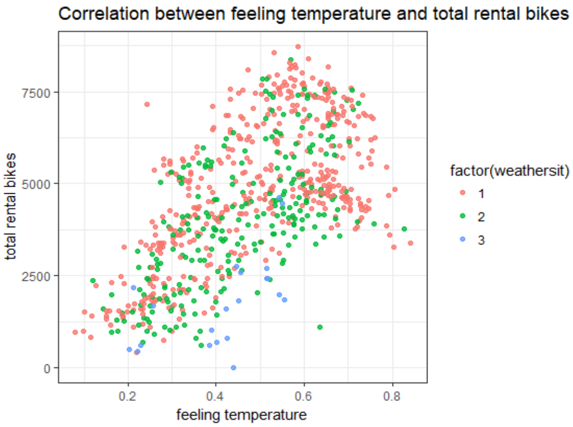


Here comes something interesting. The blue line in the above plot is the number of bike rentals from registered user. The trend looks very similar to weekday total bike rentals in last graph(the blue, green, purple, and brown lines). The red line in the above graph is bike rentals from casual users. The number has high level of usage from 11am to 6pm, which is similar to the trend of counts for all user types on weekend (the two pink lines from last graph). Thus, we can conclude that the registered users are more likely to use bike for commute, and casual users tend to use the sharing bike mainly to meet their needs in daytime.

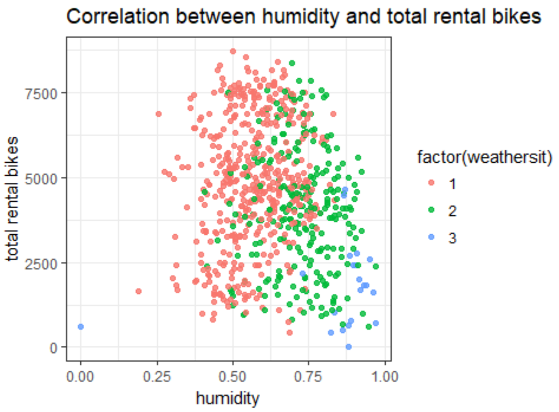


Since the timeframe of our data includes two years, we can plot the total count over time to detect the trend and seasonality of total usage. From the above time series plot, the average usage of the rental bikes in the second year is more than that of the first year. And the usage increases in summer and fall in each year, and decreases in spring and winter. Thus, it is possible to apply a time series model on our data and make prediction of bike rental in the future.

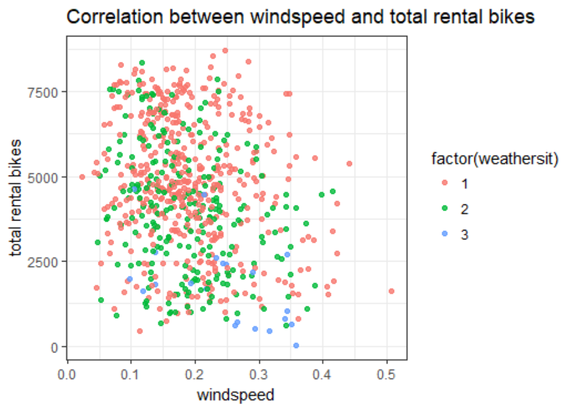
3.4.2 The count of rental bikes under different weather condition



In the graph above, most dots sit around the diagonal of y=x. It is apparent that total rental bikes and feeling temperature have linear relationship. When the weather is bad (weather sit=3), the feeling temperature is usually low, and only few people could like to rent a bike. And it seems that the numbers of rental bikes are almost the same in clear days (weathersit=1) and cloudy days (weathersit=2). So we can deduce that weathersit does not have a strong correlation with either temperature and total counts of rental bike.



When examining the correlation between total rental bikes and humidity, we can find that there is no obvious relationship between these two variables. When the “weathersit” equals to 1, the scatter point is concentrate more on the left-top corner, indicating that humidity will be low in a clear and few clouds day. When the weather is good, more people are willing to ride a bike. On the other hand, in snow days (weathersit=3), the humidity level is high, and only few people use a bike in this situation. Thus, we can conclude that there is a correlation between "weathersit" and humidity.



From the above plot, we can detect a weak negative correlation between total rental bikes and wind speed. When the wind speed is low, it is more likely to be a clear day, and more rental bikes.

3.5 Using Post Double Lasso Selection to resolve omitted variable bias

We first put all of the predictors into a simple linear regression model. We define the cnt (count of total bike rental) as dependent variable y and atemp (feeling temperature) as independent variable x. We capture all correlations as interaction effects between predictors. If we examine our model in detail, we find that it exists serious omitted variable bias. Almost all other independent variables can affect both x and y. For example, season has correlation with both feeling temperature and total number of bike rentals. Fall has lower feeling temperature, and it is more appropriate to rent a bike. Then, the change of y can either from change of an independent variable or its impact on x. For example, the increase in total number of bike rentals can either due to fall season, or due to chill feeling temperature at fall season. So we apply a post double Lasso model to apply the same set of predictors to predict both x and y. Then, use the combination of selected variables to predict y. This way, we control omitted variable, and we become more confident of the causal effect of x on y.

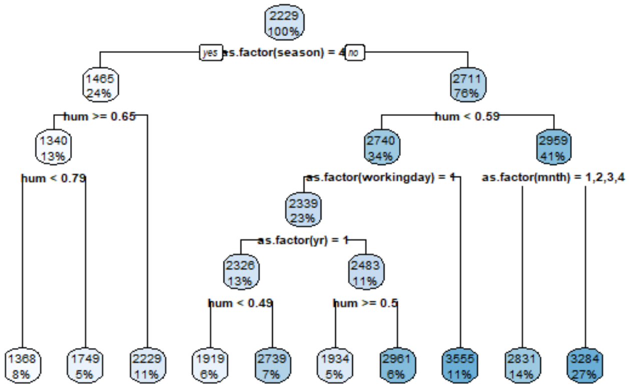
There are 605 sets of control variables in our linear regression model, while there are only 135 sets of control variables after the post double Lasso model. Thus, the Lasso model helps us filter out some uncorrelated control variables and only keep the variables selected in each Lasso model.

From the outcome of linear regression model, we can find that the coefficient of atemp(x) is 4034 and it has a significant effect on the total number of bike rentals (y) at 1% level of significance. While in post double Lasso model, the coefficient of atemp(x) is 4157, and its effect on total number of bike rentals (y) is also statistically significant. Thus, we can conclude that feeling temperature indeed has larger positive impact on the number of bike rentals, after we select the appropriate set of control variables in the model.

3.6 Using Causal Tree to find Treatment Effect of Feeling Temperature and Working Day

Then, we apply a Casual Tree model using those selected predictors to estimate heterogeneous causal effect in our data. And this can be viewed from two aspects: weather condition and time.

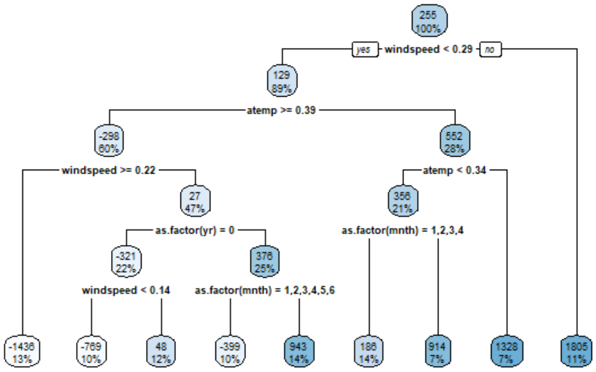
3.6.1 Select feeling temperature as treatment variable.



Originally, feeling temperature is continuous variable. In order to use feeling temperature as treatment effect, we calculate the median of feeling temperature. Label the temperature higher than median as 1 and lower than median as 0. We want to see the effect of hot day on number of bike rentals.

Compared to a low felling temperature, a high feeling temperature will increase 2229 bike rentals on average, but the specific casual effect may be slightly different due to other factors. To illustrate, during the winter season and when humidity is between 0.65 and 0.79, a higher felling temperature will increase 1368 counts of bike rentals. In working day, seasons except winter, and humidity is lower than 0.59, a higher temperature will increase 3555 counts of total bike rentals. Generally, higher temperature will make people rent more sharing bikes.

3.6.2 Select working day as treatment variable



Working day originally is a dummy variable. From this graph, we know that in a working day, the number of total bike rentals could be 255 more bikes than that in weekend. To be specific, when the wind speed is between 0.22 and 0.29, feeling temperature is higher than 0.39, in working day bike rentals will reduce by 1436, compared to a non-working day. And when the wind speed is higher than 0.29, the bike rentals of a working day is 1805 more than that of a non-working day.

3.6.3 Hypothesis

Based on the analysis and exploration in the data exploration part, we make two hypotheses:

1) The patterns of bike rental for casual users and registered users should be different.

2) Compared to time related factors, factors associated with weather conditions account more for the number of bike rentals.

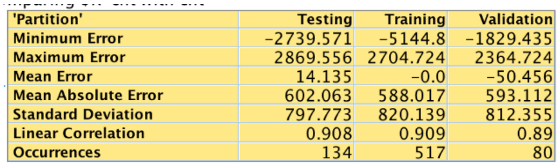
4. Modeling

4.1 Predicting total number of bike rental (cnt)

4.1.1 Regression Tree with Boosting Method

The team wants to use all independent variables to predict total number of bike rental (cnt). This way, bike sharing companies such as Citi Bank or department of transportation can have a better understanding of the change of the rental number and the predicted future number of bike rentals. As total number of bike rental is composed of registered users and casual users, Citi Bank can focus on registered users, while DOT can focus on casual users. The team first uses simple regression tree and realize that the accuracy rate is too low. In order to increase the accuracy, the team adds boosting tree to the simple tree.

Boosting method grows new trees based on current existing trees. After growing the first tree, next tree is fitted to the residuals. This way, we can increase the performance of the original tree on area where it did not perform well. The size of each individual tree can be controlled using shrinkage parameter lambda. This method learns from its own mistakes and generally increase the performance of a simple tree. Compared to not using any model, boosting method got 82.6% of accuracy rate, which is very high.

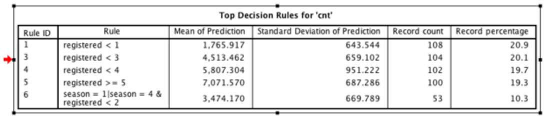


'Partition'	Testing	Training	Validation
Minimum Error	-2739.571	-5144.8	-1829.435
Maximum Error	2869.556	2704.724	2364.724
Mean Error	14.135	-0.0	-50.456
Mean Absolute Error	602.063	588.017	593.112
Standard Deviation	797.773	820.139	812.355
Linear Correlation	0.908	0.909	0.89
Occurrences	134	517	80

From the table above, we can see the mean absolute error for testing is 602, which is slightly larger than train and validation error.

4.1.2 CHAID algorithm

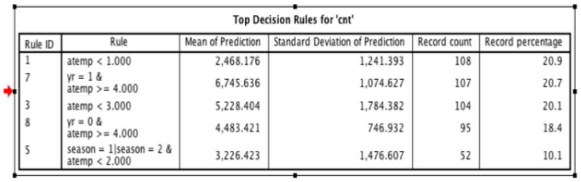
Then, we move to CHAID method to get more understanding of the predictors and the key decision rules. The R squared is 0.858, and tree depth is 3.



Top Decision Rules for 'cnt'

Rule ID	Rule	Mean of Prediction	Standard Deviation of Prediction	Record count	Record percentage
1	registered < 1	1,765.917	643.544	108	20.9
3	registered < 3	4,513.462	659.102	104	20.1
4	registered < 4	5,807.304	951.222	102	19.7
5	registered >= 5	7,071.570	687.286	100	19.3
6	season = 1|season = 4 & registered < 2	3,474.170	669.789	53	10.3

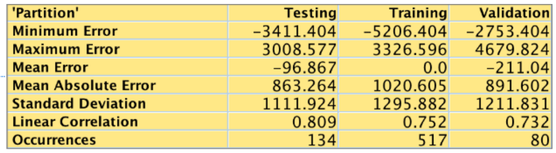
From the decision rules chart above, the red arrow shows the top decision rule. If number of bike rental from registered users is less than 3, the mean of predicted total number of bike rental is 4513.462. All of the decision rules above uses "registered" as a crucial classifier. This is reasonable, because a large portion of the total number of bike rentals is registered bike rentals. Then, we drop "registered" and "casual" in order to solely see the effect of direct independent variables. This time, R squared dropped to 0.565. There are nine nodes and the tree depth is three. From the decision rule plot, we can see the top decision rule is predicting total number of bike rentals as 6745.636 for "atemp" greater than or equal to 4.000.



Top Decision Rules for 'cnt'

Rule ID	Rule	Mean of Prediction	Standard Deviation of Prediction	Record count	Record percentage
1	atemp < 1.000	2,468.176	1,241.393	108	20.9
7	yr = 1 & atemp >= 4.000	6,745.636	1,074.627	107	20.7
3	atemp < 3.000	5,228.404	1,784.382	104	20.1
8	yr = 0 & atemp >= 4.000	4,483.421	746.932	95	18.4
5	season = 1|season = 2 & atemp < 2.000	3,226.423	1,476.607	52	10.1

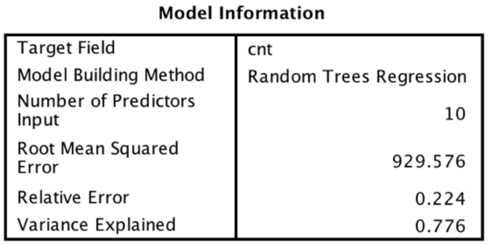
From the accuracy chart below, we can see the mean absolute error is 863, which is much larger than that of boosting method. The testing error for CHAID algorithm is smaller than that of training and validation set. This indicates the method has less over fitting issues than boosting method.



'Partition'	Testing	Training	Validation
Minimum Error	-3411.404	-5206.404	-2753.404
Maximum Error	3008.577	3326.596	4679.824
Mean Error	-96.867	0.0	-211.04
Mean Absolute Error	863.264	1020.605	891.602
Standard Deviation	1111.924	1295.882	1211.831
Linear Correlation	0.809	0.752	0.732
Occurrences	134	517	80

4.1.3 Random Forest

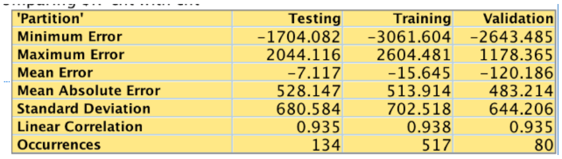
The team then uses random forest for prediction. Season is the most important predictor for total number of bike rentals, which is very reasonable. The model can explain 77.6% of variances.



Model Information

Target Field	cnt
Model Building Method	Random Trees Regression
Number of Predictors Input	10
Root Mean Squared Error	929.576
Relative Error	0.224
Variance Explained	0.776

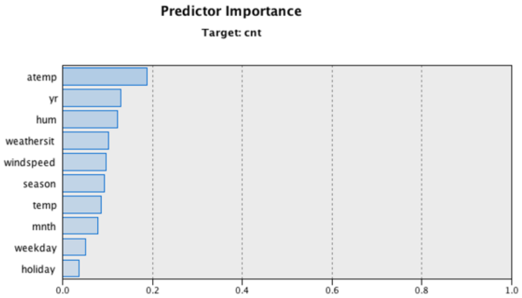
The mean absolute error for testing set is 528.147, which is smaller than boosting method and CHAID method. The error rate for validation set is even smaller than training set. In general, the testing set error is very similar to training and validation set.



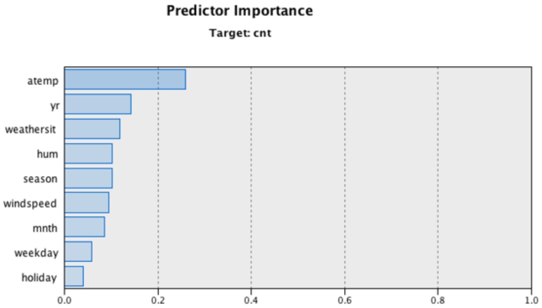
'Partition'	Testing	Training	Validation
Minimum Error	-1704.082	-3061.604	-2643.485
Maximum Error	2044.116	2604.481	1178.365
Mean Error	-7.117	-15.645	-120.186
Mean Absolute Error	528.147	513.914	483.214
Standard Deviation	680.584	702.518	644.206
Linear Correlation	0.935	0.938	0.935
Occurrences	134	517	80

4.1.4 Neural Network

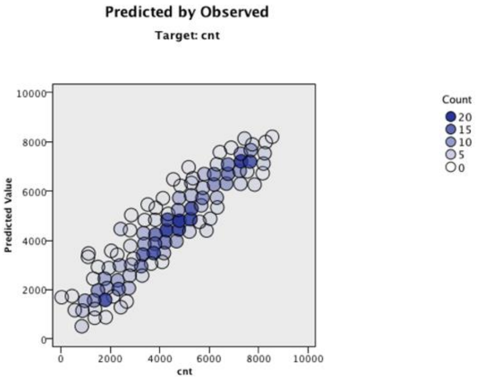
We also use neural network to try to increase accuracy and interpretability. From the predictor importance graph below, we can see that normalized feeling temperature in Celsius ("atemp"), year, humidity, weather situation, and wind speed are the most important predictors. Feeling temperature is much more important than real temperature. Feeling temperature is highly correlated with real temperature. We can drop real temperature and redo the analysis. Workday is not in the predictor importance graph, while weekday is the second to last importance predictor. Weekday is much more important than workday, and weekday has all information that workday have. We can also drop workday in our model.



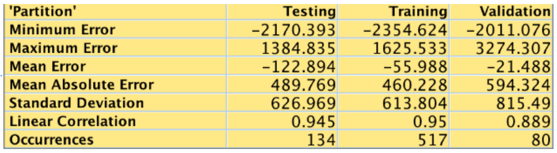
This time, we get the predictor importance graph below. Now, feeling temperature, year, weather situation, humidity, season, and wind speed are the most important predictors.



In general, the prediction is very accurate. If the observed is equal to predicted value, it should be a y=x line. From the predicted by observed dotted graph, we can see the dark blue dots are sitting in the middle of the dot group. Only very few light blue dots sit at the both side of the y=x line. The number of dots higher than the line is more than the number of dots sits lower than the line.



From the accuracy table, we can see the mean absolute error is 489.769, which is lower than the method 1 to 3. And it is very similar to training and validation error.



'Partition'	Testing	Training	Validation
Minimum Error	−2170.393	−2354.624	−2011.076
Maximum Error	1384.835	1625.533	3274.307
Mean Error	−122.894	−55.988	−21.488
Mean Absolute Error	489.769	460.228	594.324
Standard Deviation	626.969	613.804	815.49
Linear Correlation	0.945	0.95	0.889
Occurrences	134	517	80

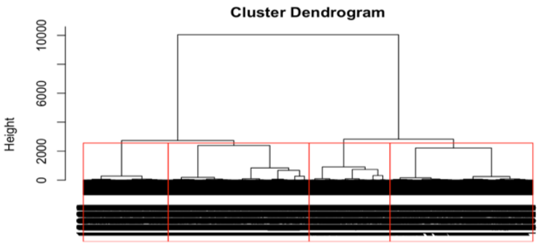
Then, we use the model without workday and temperature to test model 1-3. The test error for model 1, 2, and 3 respectively are 606.185, 863.264, and 547.854. They all changed a little bit. The best model for predicting total count is still neural network with error 489.769.

4.2 Clustering

4.2.1 Using Hierarchical Clustering to choose the right number of clusters

From our analysis in 4.1, we know that "atemp", year, weather situation, humidity, and season are the most important five factors. From these five factors, we choose "atemp", weather situation, humidity, and season to represent the overall weather predictor. We want to figure out the natural cluster from the hour dataset using these four predictors.

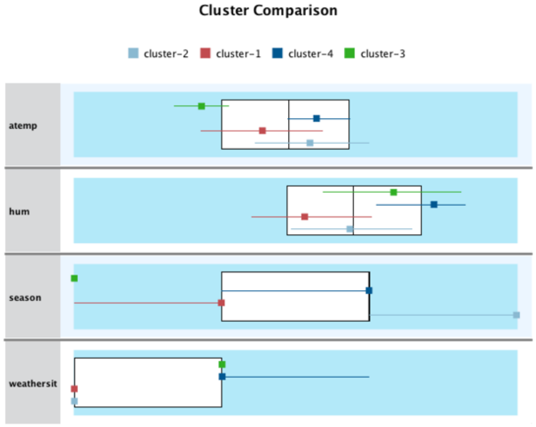
We tried to use two, three, four, and five clusters to fit the data for math class. Comparing the dendrogram for these four types, I think four clusters is the most appropriate choice for this dataset. Seeing from top to bottom, one split at height 10000 gives two set of clusters. One set of cluster split at height 2500, and the other split also at 2500. Four clusters will describe these four groups.



4.2.2 Using Non Hierarchical Clustering to examine each cluster.

4.2.2.1 K means

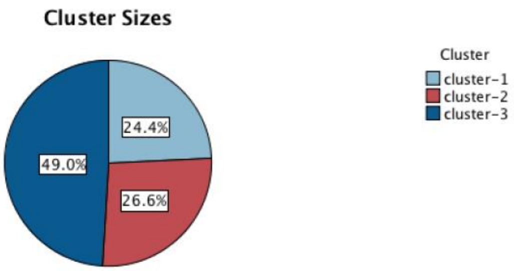
The model is able to do 20 iterations with error from 0.556 to 0. The cluster quality is higher than 0.25 and but lower than 0.5. From the cluster comparison table, we can see that weather situation separate cluster 1 2 and cluster 3 4. Season can separate all four clusters.



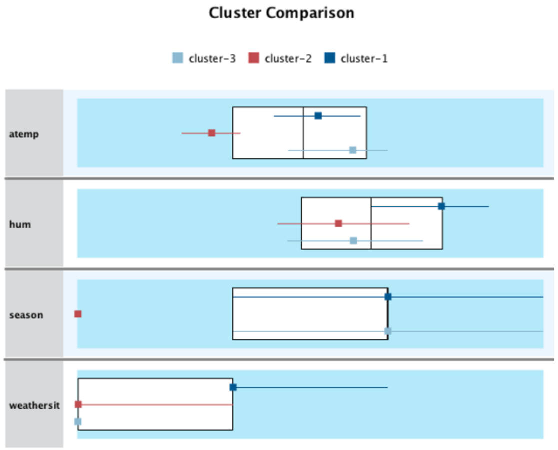
The size of the cluster is very different. Cluster 3 takes 9.4%, and cluster 4 takes 16.5%. Cluster 1 takes 32.9%, and cluster 2 takes 41.3%.

4.2.2.2 Two Step Model

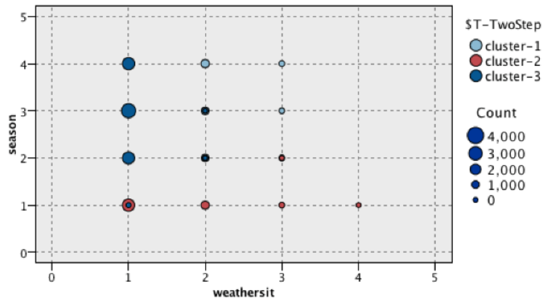
The two step model calculate the number of clusters from 2 to 15. It chooses three clusters. The cluster size seems more reasonable than the four clusters in k means. Thus, we choose three clusters.



From the cluster comparison graph, we know that season can separate cluster 2 from cluster 1 and 3. Weather situation can separate cluster 1 from 2 and 3.



From the graph below, we can clearly see that large number of bike rentals happens at weather situation equals 1. The number of bike rentals decreases as weather situation increases from 1 to 4.



Cluster 1 is medium temperature, high humidity, cloudy to light snow/ light rain/ thunderstorm/ scattered clouds, and summer/ fall/ winter. Cluster 2 is low humidity, low temperature, clear/ few clouds/ partly cloudy spring. Cluster 3 is high temperature, medium humidity, clear/ few clouds, and summer/ fall/ winter.

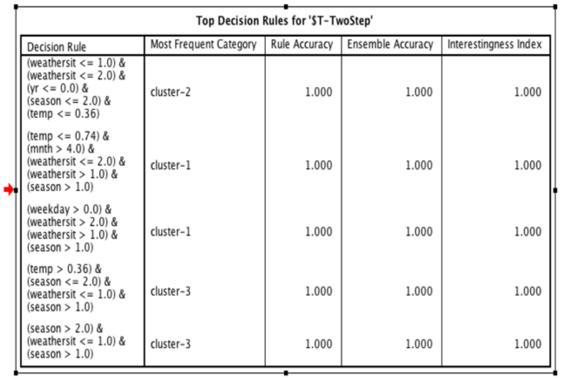
4.2.3 Predicting three clusters and understanding important predictors

4.2.3.1 Random forest

From Step 2, we have three clusters. Now, we use random forest to predict the classification. The model also predict season as the most important predictor, same as our prediction before. The accuracy rate is very high. The model can predict 99.46% data correct in the test set. Thus, this model is pretty reliable.

'Partition'	Testing		Training	
Correct	8,669	99.46%	8,649	99.84%
Wrong	47	0.54%	14	0.16%
Total	8,716		8,663	

From the decision rule table, we can know that both cluster 1 and 3 have two decision rules, while cluster 2 has 1 rule. The strongest rule is when temperature less than 0.74, month is 5-12, weather situation is 3-4, season is summer/ fall/ winter, the data belongs to cluster 1.



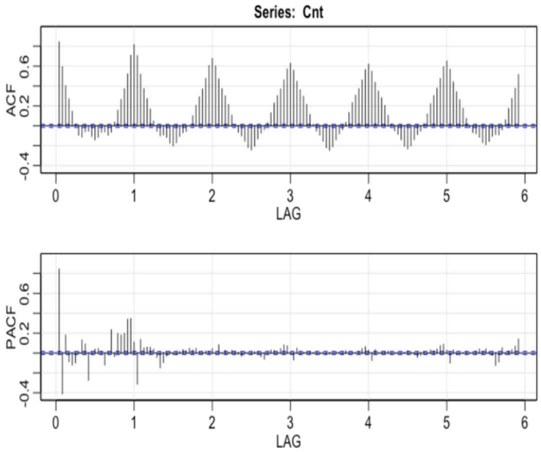
Top Decision Rules for '$T-TwoStep'

Decision Rule	Most Frequent Category	Rule Accuracy	Ensemble Accuracy	Interestingness Index
(weathersit <= 1.0) & (weathersit <= 2.0) & (yr <= 0.0) & (season <= 2.0) & (temp <= 0.36)	cluster-2	1.000	1.000	1.000
(temp <= 0.74) & (mnth > 4.0) & (weathersit <= 2.0) & (weathersit > 1.0) & (season > 1.0)	cluster-1	1.000	1.000	1.000
(weekday > 0.0) & (weathersit > 2.0) & (weathersit > 1.0) & (season > 1.0)	cluster-1	1.000	1.000	1.000
(temp > 0.36) & (season <= 2.0) & (weathersit <= 1.0) & (season > 1.0)	cluster-3	1.000	1.000	1.000
(season > 2.0) & (weathersit <= 1.0) & (season > 1.0)	cluster-3	1.000	1.000	1.000

4.3 Time Series Analysis

In "hour" dataset, we have the number of bike rentals for "casual", "registered", and "total" of every hour. Using time series algorithm, we can learn a lot about the seasonality of the current trend. We can also forecast the next 24-hour bike rentals.

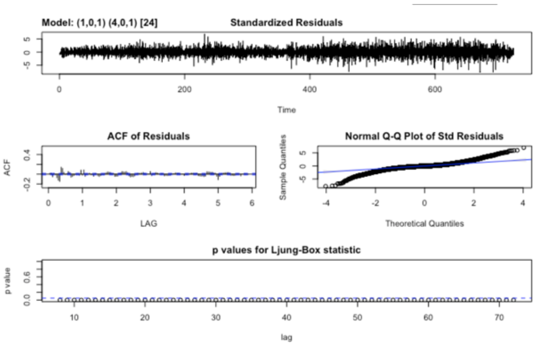
We first plot ACF and PCAF of the hourly total rental data. First, we look at ACF plot. The data shows clear seasonality traits, as it decreases first and then increase. It has a peak at 0, 1, 2, 3, 4, 5, and 6. This indicates the data peaks at day 0, 1, 2, 3, 4, 5, and 6. The seasonality frequency is every 24 hours, which is very reasonable. People tend to rent more bikes at certain time of a day and the pattern rotate for next day. The height of the peak for each day gradually decreases from 0.8 in day 0 to less than 0.6 at day 6. ACF plot shows exponential decay, and has auto correlation components. We need to capture seasonality and AR in our model.



Then, we look at the PCAF plot at the bottom. The first bar has a large positive value, and the second bar has a large negative value. And then it increases and decreases up and down. When lag grows from 1 to 6, the PCAF value goes close to zero. It shows moving average traits in the plot. The peak of ACF does not shows continuously stable pattern, and PCAF has value not equal to zero at lag 2 to lag 6. The data is not random walk. Thus, we can try SARIMA model to fit the current data and forecast future data.

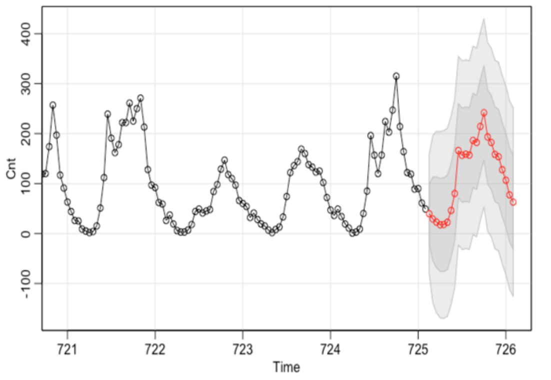
SARIMA model has p, d, q, P, D, Q, S parameters. P is AR order, d is difference order, q is MA order, P is SAR order, D is seasonal difference, Q is SMA order, and S is seasonal period. From ACF and PCAF, we know that the data is not random walk, thus we can specify d as 0, and D as 0. Also, the seasonality frequency is 24, so S is 24. We are not sure about the magnitude of other parameters. As we know, smaller BIC indicates that a model has higher prediction accuracy. Thus, we write a loop to determine the size of other parameters with the smallest BIC. It turns out that SARIMA 100(1,0,1,4,0,1,24) has the smallest BIC.

We first see the "Normal Q-Q Plot of Std Residuals" graph. If all dots sit on the blue line, the residual is normal distributed. From the graph, we know that many data from -4 to -2 and 2 to 4 are not on the blue line, which indicates the residual is not normal. From the ACF of residual plot, we can see the residual is auto correlated.

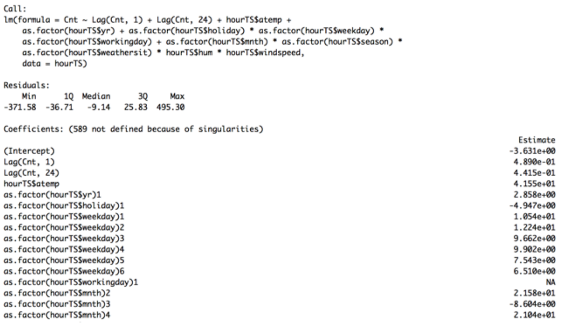


Then, we use this model to forecast next 24-hour bike rentals, which is the 725th day. The black dots and lines are observed data, and the red dots and lines are forecast data. It has a lower peak height than last day (724th day), but a higher peak height than the day before last day (723th day). On day 725, we have the first 3 hour observations (the first three black dot at the beginning of day 725). From 3am to 6am, the number of total bike rentals continuously goes down. 6am is the trough. Starting from 6am, it goes up again. The second derivative of 10 am is very high. 6pm is the peak of the day. Then, it gradually goes down. Day 725 refers to 1/1/2013, which is a Tuesday.

The dark shaded area indicates 80 percent confidence interval area, and the light shaded area indicates 95 percent confidence interval area. That is, each predicted value is possible to lie in the shaded area with a probability of 80% or 95%.

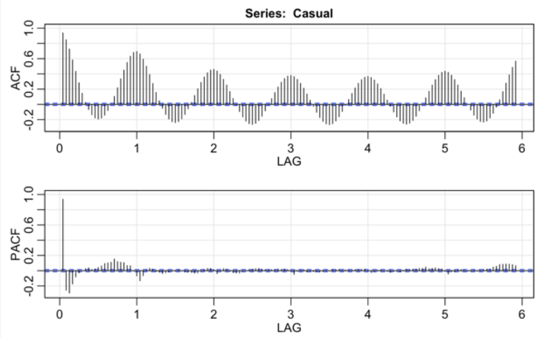


We know that next hour bike rental number is correlated with this hour, which is Lag(1). And it is also correlated with the number of bike rental happened 24 hour ago, which is Lag (24). Thus, we can add Lag (1) and Lag(24) to our linear regression model containing all other predictors and interaction terms. Then, we got the following model. The number of bike rentals of one our is positively correlated with lag(1) and lag(24) for a factor of 0.48 and 0.41, respectively.

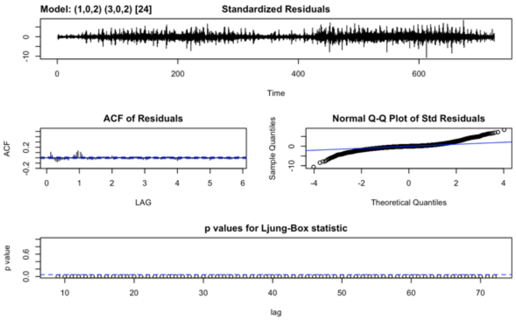


```
Call:
lm(formula = Cnt ~ Lag(Cnt, 1) + Lag(Cnt, 24) + hourTS$atemp +
    as.factor(hourTS$yr) + as.factor(hourTS$holiday) * as.factor(hourTS$weekday) *
    as.factor(hourTS$workingday) + as.factor(hourTS$mnth) * as.factor(hourTS$season) *
    as.factor(hourTS$weathersit) * hourTS$hum * hourTS$windspeed,
    data = hourTS)

Residuals:
    Min      1Q  Median      3Q     Max
-371.58  -36.71   -9.14   25.83  495.30

Coefficients: (589 not defined because of singularities)
                                                                       Estimate
(Intercept)                                                          -3.631e+00
Lag(Cnt, 1)                                                           4.890e-01
Lag(Cnt, 24)                                                          4.415e-01
hourTS$atemp                                                          4.155e+01
as.factor(hourTS$yr)1                                                 2.858e+00
as.factor(hourTS$holiday)1                                           -4.947e+00
as.factor(hourTS$weekday)1                                            1.054e+01
as.factor(hourTS$weekday)2                                            1.224e+01
as.factor(hourTS$weekday)3                                            9.662e+00
as.factor(hourTS$weekday)4                                            9.902e+00
as.factor(hourTS$weekday)5                                            7.543e+00
as.factor(hourTS$weekday)6                                            6.510e+00
as.factor(hourTS$workingday)1                                                NA
as.factor(hourTS$mnth)2                                               2.158e+01
as.factor(hourTS$mnth)3                                              -8.604e+00
as.factor(hourTS$mnth)4                                               2.104e+01
```

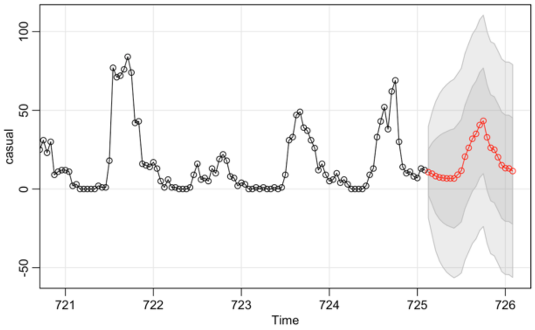
Then, we also build time series analysis for casual riders. We repeat the same process to get ACF and PCAF graph for casual. The second derivative of ACF of casual bike riders is smaller than that of total bike riders. The peak of ACF decreases from 1 to 0.4 from lag 0 to lag 3. Then, the peak of ACF increases from 0.4 to 0.6 from lag 4 to lag 6. The peak is not stable. The ACF graph shows autocorrelation, and seasonality. The PACF graph shows moving average. Thus, similar to total bike rental graphs, we also choose SARIMA model for forecasting. From the algorithm selection of the smallest BIC, SARIMA 1,0,2,3,0,2,24 is the best choice.



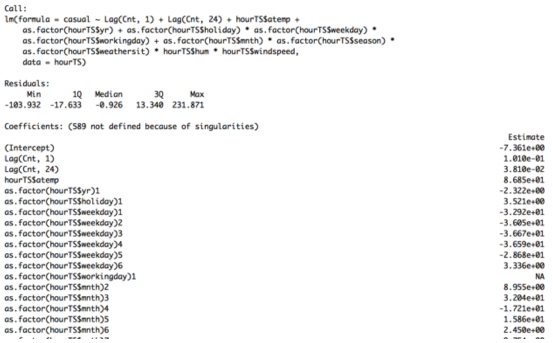
From the residual evaluation plot, the best SARIMA for casual has similar problem. The residual is not entirely normally distributed, and the residual has auto correlation.



Then, we use this model to forecast next 24-hour (day 725) casual user bike rental number. The number of casual riders goes down from 3am to 9am. 9am is the trough. Starting from 9am, it goes up and reach the first peak at 6pm. Then, it gradually goes down. The height difference between peak and trough is smaller, compared to that of the total count forecast. 6pm is the peak for both casual and total number of bike rentals.



Then, we add lag(1) and lag(24) into linear model. One day number of bike rental is correlated with lag(1) and lag(24) of a factor of 0.1 and 0.3, respectively. The estimate is not significant.



```
Call:
lm(formula = casual ~ Lag(Cnt, 1) + Lag(Cnt, 24) + hourTS$atemp +
    as.factor(hourTS$yr) + as.factor(hourTS$holiday) * as.factor(hourTS$weekday) *
    as.factor(hourTS$workingday) + as.factor(hourTS$mnth) * as.factor(hourTS$season) *
    as.factor(hourTS$weathersit) * hourTS$hum * hourTS$windspeed,
    data = hourTS)

Residuals:
     Min       1Q   Median       3Q      Max
-103.932  -17.633   -0.926   13.340  231.871

Coefficients: (589 not defined because of singularities)
                                                                        Estimate
(Intercept)                                                           -7.361e+00
Lag(Cnt, 1)                                                            1.010e-01
Lag(Cnt, 24)                                                           3.810e-02
hourTS$atemp                                                           8.685e+01
as.factor(hourTS$yr)1                                                 -2.322e+00
as.factor(hourTS$holiday)1                                             3.521e+00
as.factor(hourTS$weekday)1                                            -3.292e+01
as.factor(hourTS$weekday)2                                            -3.605e+01
as.factor(hourTS$weekday)3                                            -3.667e+01
as.factor(hourTS$weekday)4                                            -3.659e+01
as.factor(hourTS$weekday)5                                            -2.868e+01
as.factor(hourTS$weekday)6                                             3.336e+00
as.factor(hourTS$workingday)1                                                 NA
as.factor(hourTS$mnth)2                                                8.955e+00
as.factor(hourTS$mnth)3                                                3.204e+01
as.factor(hourTS$mnth)4                                               -1.721e+01
as.factor(hourTS$mnth)5                                                1.586e+01
as.factor(hourTS$mnth)6                                                2.450e+00
```

ACF and PACF of registered group looks almost the same as total count. As registered group takes most of the total count, its pattern can be greatly represented by total count pattern. Thus, we did not do separate analysis for registered group.

Some factors that can be concerned in future development is separating training, testing, and validation set of time series regression analysis for a higher accuracy score. We can also use dimension reduction methods such as lasso, ridge, or principal component regression to delete some less relevant predictors. Also, the regression does not incorporate the moving average pattern of the time series. In future development, we need to capture this trait in the model.

5. Conclusion

The paper begins with data exploration and correlation between variables. And we can learn that both factors associated with weather condition and time have strong correlation with the counts of bike rentals, while the bike rental patterns for registered users and causal users are totally different. Then we apply a double post Lasso model and causal Tree on the data to estimate the heterogeneous causal relationship between variables. And we find that both feeling temperature and working day are very powerful predictors.

Then, we apply three types of models to analyze the data from different perspectives. First, we ignore the time part and focus on using the rest factors to select important features that affect the demand. Among the four regression models, the Neural Network model performs best for it has the least absolute error. And the result shows that the most important factor is feeling temperature, which supports our initial hypothesis, and the top five predictor also include year, weather situation, humidity, and season. And the second type of model is clustering, and here we use the outcome from the regression model and focus on the top predictors for analysis. It turns out that the two-step method has more reasonable outcome with three clusters. And we can find that the largest cluster contains almost half of the total users, which indicates that majority of people would like to rental bikes in days with high temperature, medium humidity, clear/few clouds, and during summer/fall/winter season.

Then, we take the time factor into consideration and apply time series models to predict the future demand. Based on the data exploration result, we can learn that the trend of bike rentals for registered users is similar to the rental for total users, but it is significantly different from the casual users. So we make predictions for each group separately. We use SARIMA model to forecast next 24-hour bike rentals, and the result indicates that in terms of total users, 6 am is the trough and 6 pm is the peak of the day. The demand of bike rental increases during the daytime but there is a slightly decline after 10 am. The prediction of demand for casual users is similar, but what different from the register users is that there is no fluctuation during the daytime and the demand is keep increasing from 6 am to 6 pm. We also fit two of the six parameters in SARIMA model and other factors into a linear regression model, and it turns out that none of the predictors has significant effect on the counts of bike rental. While going back to our initial hypothesis, the analysis of time series data supports our hypothesis that the bike rental pattern of registered users is different from casual users.

The analysis of the paper could serve as a reference for organizations that develop bike sharing system. They can learn the change in demand in different time periods during a day, in different seasons, and under various kinds of weather condition, thus they could be able to adjust the arrangement of the bikes to meet people’s need in a specific day. Also, the analysis is also helpful to identify potential location of the docked bikes, since for registered users, they could like to use the bike during the office and school time, which indicates that the locations near schools and office buildings are worth consideration.

If we could get more data in future, we can improve our time series model by splitting training, testing, and validation set to secure a higher accuracy of prediction. And we would like to improve our model by applying dimension reduction methods or principal component regression to better select predictors for analysis. For future analysis, we would like to combine our data with the locations of bike rentals data, to prediction the demand of bike rentals for specific areas. And in this way, the analysis and prediction will be more straightforward and helpful for organizations to develop the bike sharing system.

Reference

Bicycle-sharing system. (n.d.). In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Bicycle-sharing_system

Chad, Stecher. “post-double-LASSO Estimator.” ECON 4961/6961: Econometric Methods For Big Data, 22 Apr. 2018, Rensselaer Polytechnic Institute, New York.

Chad, Stecher. “Heterogeneous Treatment Effects” ECON 4961/6961: Econometric Methods For Big Data, 17 Apr. 2018, Rensselaer Polytechnic Institute, New York.

Hadi Fanaee-T, (2013, December 20), Bike Sharing Dataset Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset

Code Reference

Creating plots in R using ggplot2 - part 10: boxplots. Retrieved from http://t-redactyl.io/blog/2016/04/creating-plots-in-r-using-ggplot2-part-10-boxplots.html

Davo, (2013, May 22). Using aggregate and apply in R. Retrieved from https://davetang.org/muse/2013/05/22/using-aggregate-and-apply-in-r/

Galit Shmueli, Peter C. Bruce, Inbal Yahav, Nitin R. Patel, Kenneth C. Lichtendahl Jr.-Data Mining for Business Analytics_ Concepts, Techniques, and Applications in R-Wiley (2017)

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*, Springer Science+Business Media, New York. http://www-bcf.usc.edu/~gareth/ISL/index.html

Note: We use both R and SPSS for modeling.

Code

Xuyang Bai, Mengyan Zhu, Github Repository https://github.com/Dodobb/Bike-sharing