Report of DS 502 Final Project

Bikeshare Analysis

4/18/2017

Alexander Wald
Xuechun Li
Yiyuan Ma
Yan Zhao

Introduction

A bikeshare system makes a fleet of bikes available to people to ride during a short time period. Bikeshare systems are often located in cities and provide an efficient method of transportation for people to travel short distances within a city. A person can use the system by going to a bikeshare dock in a city, use a membership card or money to take out a bike, ride the bike, and then return the bike to a dock near the destination somewhere else in the city. Bikeshare systems provide different membership plans for people to become "registered users" and also allow "casual users" to pay each time that they use a bike. There are many factors that affect the demand (number of users) of bikeshare bikes on a given day or time. The following analysis explores ways in which the number of bikeshare users on a given day can be predicted using time and weather predictors.

The bikeshare dataset (https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset) that was used contains 731 data points and 16 attributes. Data was collected from the Capital Bikeshare System in Washington D.C. in 2011 and 2012. Before the data could be used, it needed to be preprocessed. The dataset consists of the following predictors

- Season (1-4), Year (0:2011 - 1:2012), Month (1-12), Weekday (0-6) whether today is a holiday, or working day
- Weather situation (Clear=1, Mist=2, Light Snow or Rain=3, Heavy Rain=4)
- Normalized Temperature, Perceived Temperature, Humidity, Windspeed
- Total # of bikeshare users, # of registered users, # of casual users

Methods and Results

Preprocessing of the Data Set

Before the dataset could be used it had to be preprocessed. Preprocessing involved adding additional columns to the dataset for the categorical variables and then later splitting the data into a training set (75% of all data) and a test set (25% of all data). For each of the categorical variables Season, Month, Weekday, and Weather Situation, the scalar predictor was split into a vector of binary predictors, one predictor for each category. For example, four additional predictors were added for season, Spring (0,1), Summer (0,1), Fall (0,1), Winter (0,1). The data for season was modified so that 1 would only appear in the one column for a certain data point's season. The other 3 columns were set to 0. In addition to this encoding for season, an additional encoding that captures the relationship between consecutive seasons was attempted:

	Winter Spring	Spring Summer	Summer Fall	Fall Winter
Winter	1	0	0	1
Spring	1	1	0	0
Summer	0	1	1	0
Fall	0	0	1	1

Even though the encoding in the table above captures the additional information of seasons being consecutive, the technique was unable to provide any improvement to linear regression. This is because the rows of the bit vectors for the season are not linearly independent. As a result, four predictors degenerated into two degrees of freedom resulting in higher training error, and, in this case, also higher test error.

Linear Regression

One of the methods that was used to answer some of the proposed questions is linear regression. To start, the predictors were divided into groups. The first group is time category predictors. This groups includes one binary predictor for each of the 7 weekdays, one binary predictor for each of the 12 months, 4 binary predictors for each of the seasons, another 4 predictors for consecutive season encoding, and 3 more binary predictors, one to indicate if the day is a holiday, one to indicate if the day is a weekday, and one to indicate if the day is a workday. The second group consists of weather predictors. This group contains 4 mutually exclusive binary predictors to describe the weather. Together the predictors indicate one of the following weather situations: clear, mist, light snow or rain, heavy snow or rain. The weather category also contains normalized temperature, perceived temperature, humidity, and windspeed. The third category is a long term time variable (Instant). Instant is n for the nth data point, and, in the Day dataset, is equivalent to the number of days elapsed since the beginning of data recording (January 1, 2011).

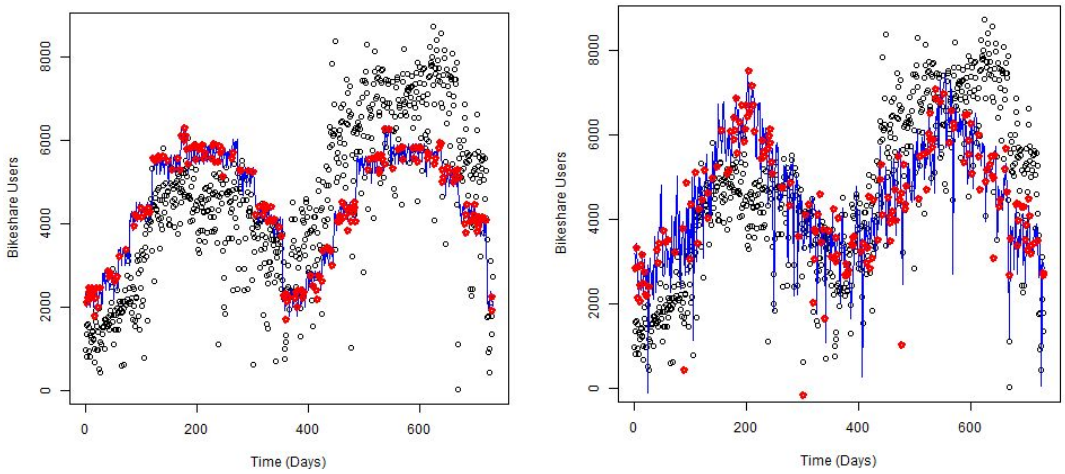
Some of the questions that were posed at the start of the linear regression analysis were

1. How well will linear regression perform in predicting the total number of bikeshare users given only time category variables? The question assumes no long term trends and also that someone does not have access to weather.
2. Is knowing the weather alone enough to accurately predict the number of bikers on a given day?
3. How well will linear regression perform in predicting the total number of bikeshare users given the time category and weather predictors? Because this questions does not include the Instant predictor, it assumes no long term trends.

To answer the first question, a linear regression model was trained using all of the time category predictors on 0.75 of the data and tested on the remaining 0.25 of the data. To answer the second question, only the weather predictors were used to predict the number of bikeshare users. The following graphs show the original data points in black, the models' predictions on training data in blue, and the predictions on test data in red. The graphs show a trend in test predictions that appears to be a compromise between the two years in the test set. Since this model was produced subject to the assumption that there are no predictable long term trends over the course of years, we cannot assume that there is any significance in the fact that the second year is biased above the first year. In addition, we do not have enough data to assume that an unrecorded year in the future would appear more like the first year (2011) or second year (2012) in this dataset. Therefore, the compromise between the two years is best for making predictions on years in general. After testing the time predictors model (graph a) on 20 different training and test set pairs, a mean test error for the of 2.23 million = 2235569.2 resulted. The weather predictors model (graph b) resulted in a test error of 2.00 million = 1997520.6. Switching from time predictors to weather predictors caused mean test error to drop by (2.23 - 2.00)/2.23=0.115 = 11.5%. Therefore, knowing only the weather conditions will likely enable one to more accurately predict the number of bikeshare users than only knowing the time category predictors.

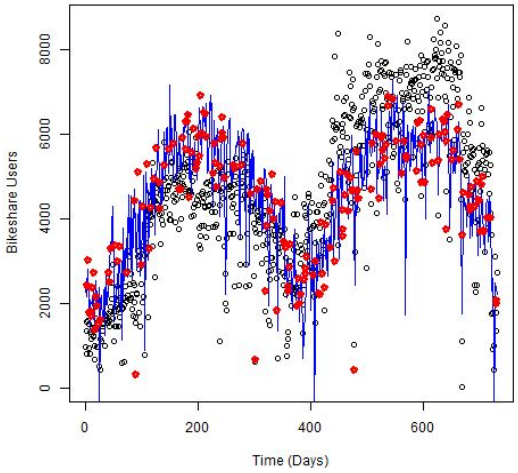
Total Number of Bikeshare Users on each Day during 2011 and 2012

(a) Time Predictors (b) Weather Predictors



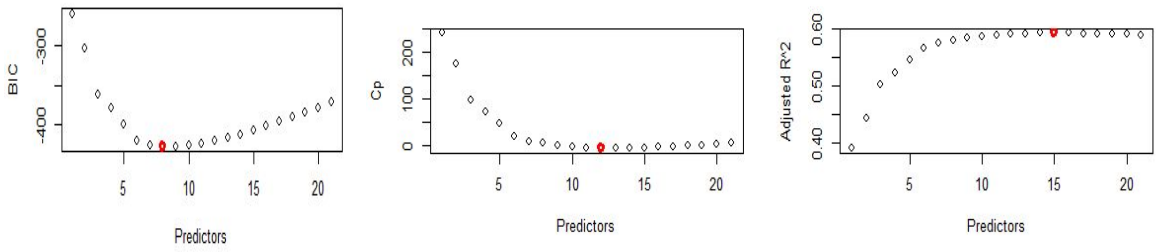
To answer the third question, both the time category and weather predictors were made available to linear regression. The following graph shows the resulting test predictions in red.

Total Number of Bikeshare Users on each Day during 2011 and 2012



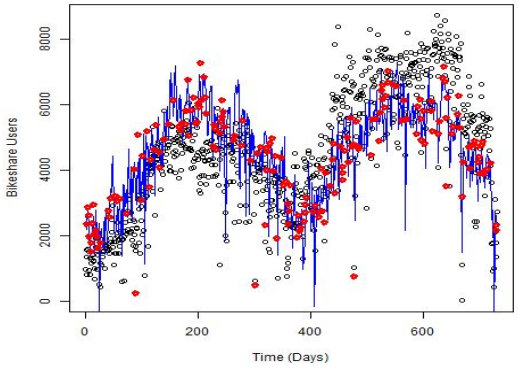
After supplying both time category predictors and the weather to the linear regression model, the average test error decreased to 1.66 million = 1657697. Therefore, making use of both time category and weather predictors is the best for predicting the total number of bikeshare users in a day with linear regression.

After answering the three questions above, it was then determined which predictors out of time category predictors and weather predictors form the best subset of predictors for producing a model that performs well on test data. The best subset technique was used to answer this question. The following graphs and the table show the results for the BIC, Cp and R2 statistics.



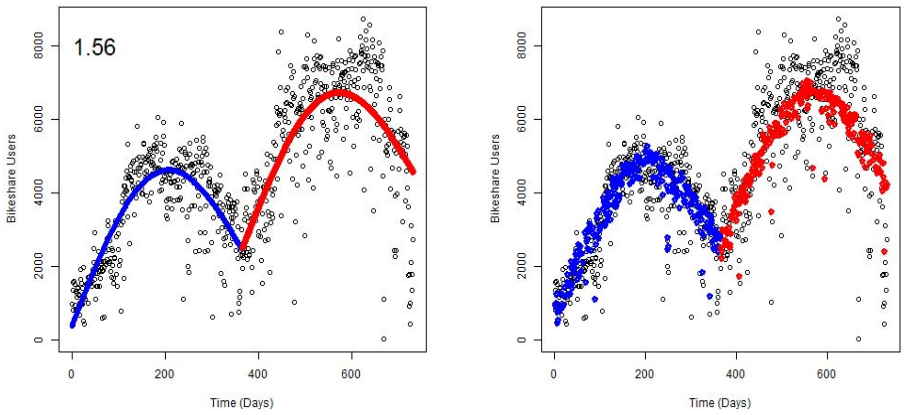
	BIC	Cp	R^2
Best Number of Predictors	8	12	15
Mean Test Error	**1603776**	**1596061**	**1596831**

The Mean Test Error, which was produced using 20 different training and test set pairs, is approximately equal for each of the three statistics. The BIC results, which contain the fewest number of predictors, are shown below.



The the predictors that were chosen are the binary predictors summer, winter, september, light snow or rain , and weekday, and the continuous predictors temperature, humidity, and windspeed.

After considering short term trends that can be predicted using the time category variables and the weather, an attempt to predict long term trends was made by considering the number of days since January 1, 2016. The first attempt at a long term model consists of a linear term to capture the increase in bikeshare users from 2011 to 2012. The model also consists of an abs(sin(c*time)) term to capture the yearly trend with number of users rising and falling according to the seasons. The model was trained on the first year and tested on the second year. The image on the left shows the predictions for the number of bikeshare users with the new model. This model resulted in a test error of approximately 1.56 million. The image on the right shows the model after weather has been added to the predictors. The average test error across 20 trials for the model with the weather added is 1.27 million = 1268299, which is smaller than the test error produced for the short term models. The fit to training data is shown in blue and the the test predictions are shown in red.



Because this long term model was created with knowledge of the test data (the second year), data snooping occurred. However, the model performed very well on the second year with a test error of only 1.27 million even though it was trained on just the first year. Therefore, the first year captures the assumed type of trend as it carries over to the second year very well. If there were an unrevealed third year in the data set, a good step to perform next would be to test the model on the third year. Since the apparent relationship between consecutive years that the number of bikeshare users tends to increase only occurs between one consecutive pair of years, we have little evidence that this model will be good for additional years. However, there is no evidence that this model will perform poorly on future years, so it would be worth testing this model on future years if the data were available.

Trees and Random Forest

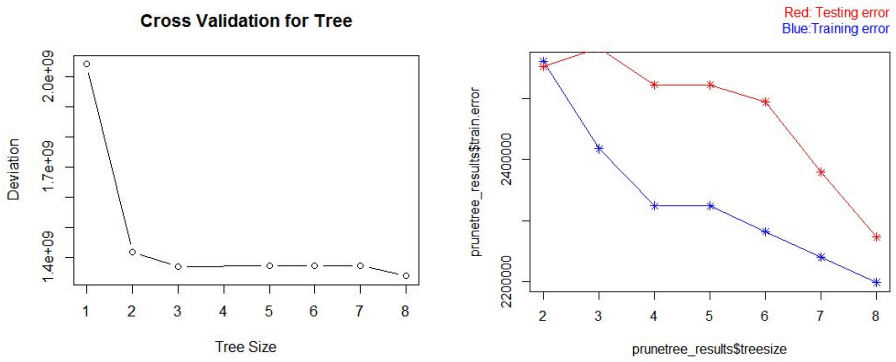
In this part, we try Tree and Random Forest to answer the similar questions mentioned in the Linear Regression part. We have following questions:

1. How well are regression tree's and random forest's performance in predicting the total user number (including register users and casual users) given only time category variables?
   For this question, we just use 12 binary predictors for months, 7 binary predictors for weekdays, 4 binary predictors for season, one binary predictor for holiday, one for working day and one for weekday. For all the binary predictors, 0 means "False" while 1 means "True".
2. How well are regression tree and random forest performance in predicting the total user number given by time and weather related variables?

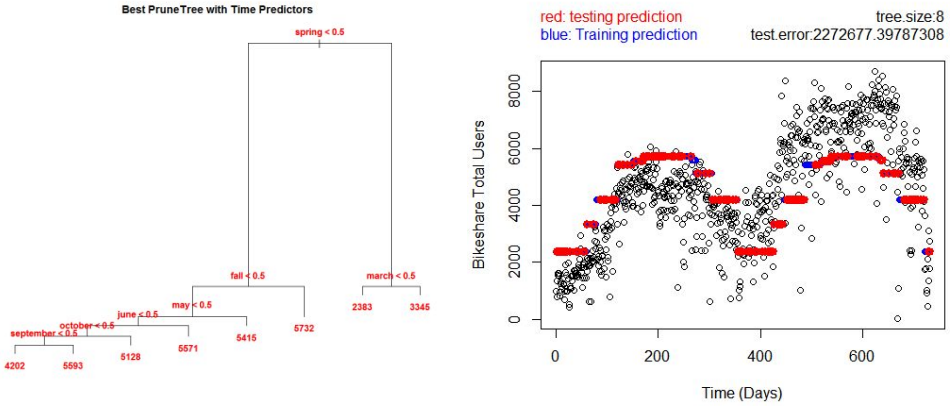
At first I will introduce the processing logic of using tree and random forest to analyzing those questions.

Following this logic, let's answer the first questions.
Here we fit a regression tree to the time category variables. First, we create 20 different randomly selected

training and testing dataset. For each data set, training data takes 3/4 data point of the whole dataset. We use the training data to fit the tree with cross validation to find out the "best" number of terminal nodes of each tree, and do tree pruning process. From the left figure below, we can see that the tree with 8 terminal nodes results in the lowest cross validation error rate. Then we can prune the tree to a 8-node tree. We use this tree to predict data and get the testing error is 2272677. The right picture shows the testing error and training error varying with the change of tree size.

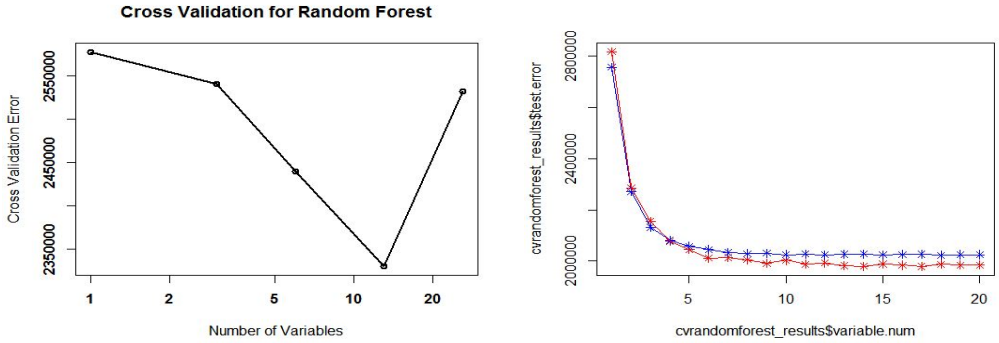


Then we get the pruned tree with 8 nodes. From the left figure we can see that, if we just have time related predictors, we can see that in September, we'll have most users. In spring, we will have less users, especially during the February and January. Fall also has pretty much users. And then, we plot the training prediction and testing prediction as well as the data set(black).



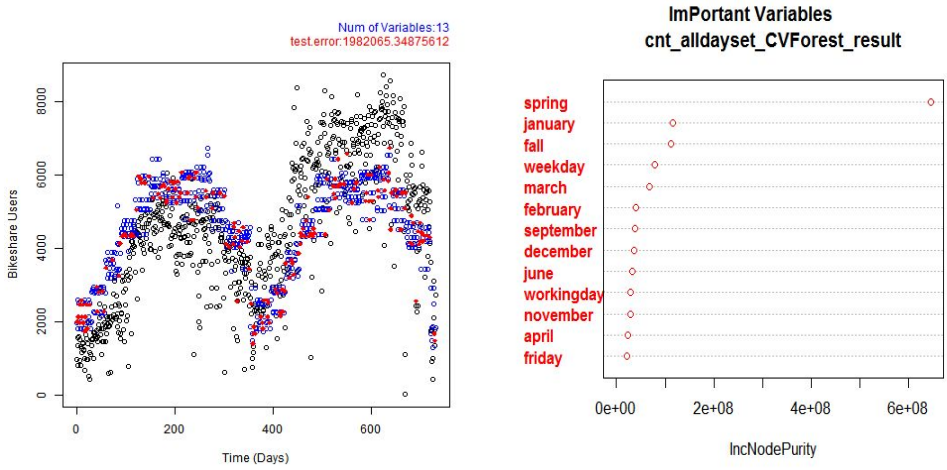
We also try to answer the first question with random forest. Following the logic above. In order to find how much predictors needed to be included, we use cross validation to determine that. The figure (a) shows that the cross validation result, so the best number of variables put in random forest is 13. The figure (b) shows that the testing error (red) and training error (blue) varying with the variables number from 1 to 20. We can see that the error becomes gradually smooth.

(a) Cross Validation result for Random forest (b) Testing error and Training error



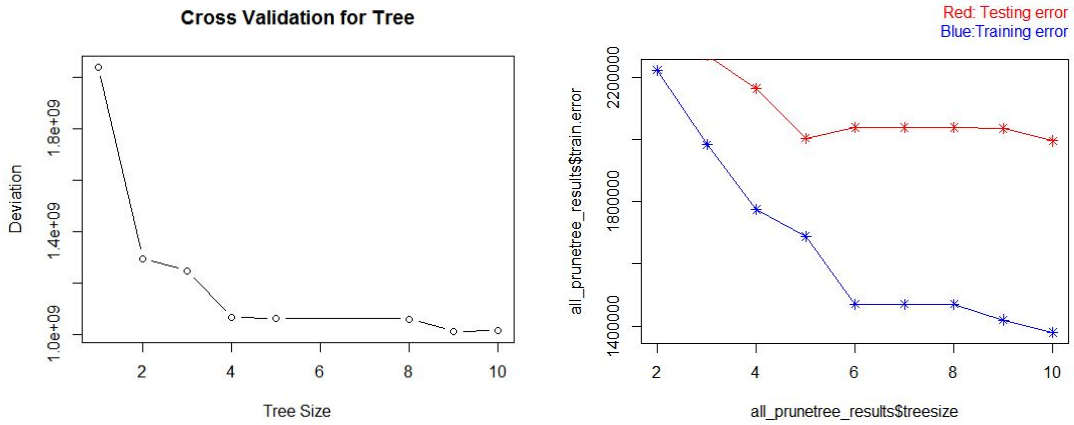
The figure (c) shows the random forest result with 13 variables contained in it and the error of 20 random selected datasets is 1.95 million. The red points are prediction with test data, the blue point are prediction with training data. The figure (d) give us the 13 most important variables, they are spring, January, fall, weekday, march, February, September, December, June, working day, November, April and Friday.

(c) Random Forest prediction Plot (b) Important Variables for Random Forest

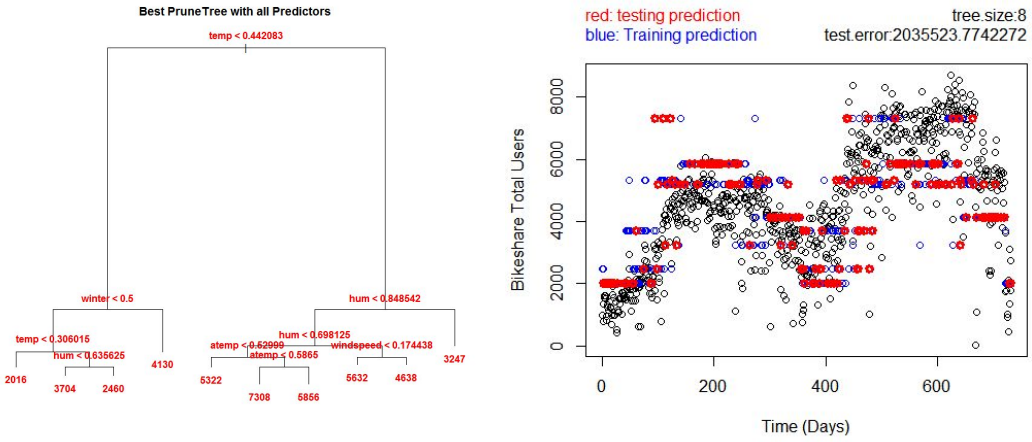


After answering the first questions, let's move to the second one. Let's see how tree and random forest perform with all predictors including the time related features and weather related features.

Similar, we use the training data to fit the tree with cross validation to find out the "best" number of terminal nodes of each tree, and do tree pruning process. From the left figure below, we can see that the tree with 9 terminal nodes results in the lowest cross validation error rate. Then we can prune the tree to a 9-node tree. We use this tree to predict data and get the testing error is 1995389. The right picture shows the testing error and training error varying with the change of tree size.

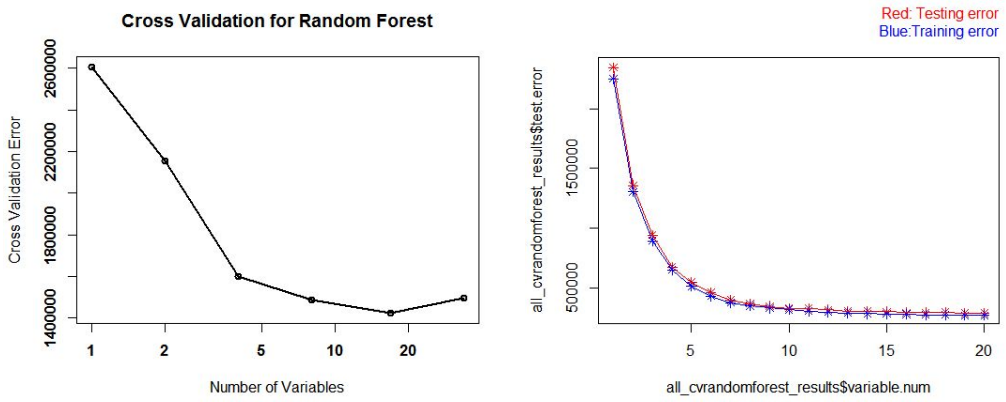


Then we get the pruned tree with 9 nodes. From the left figure we can see that, if we have all predictors, we can see that important predictors are temperature, humility, wend speed, feeling temperature (atemp) and winter, which means the weather condition is really important for bake sharing activity. When the temperature is higher than 18 ℃ and humility is lower than 69.81, as well as the feeling temperature is between the 23℃ and 25℃, the user is the most. And then, we plot the training prediction() and testing prediction as well as the data set(black).



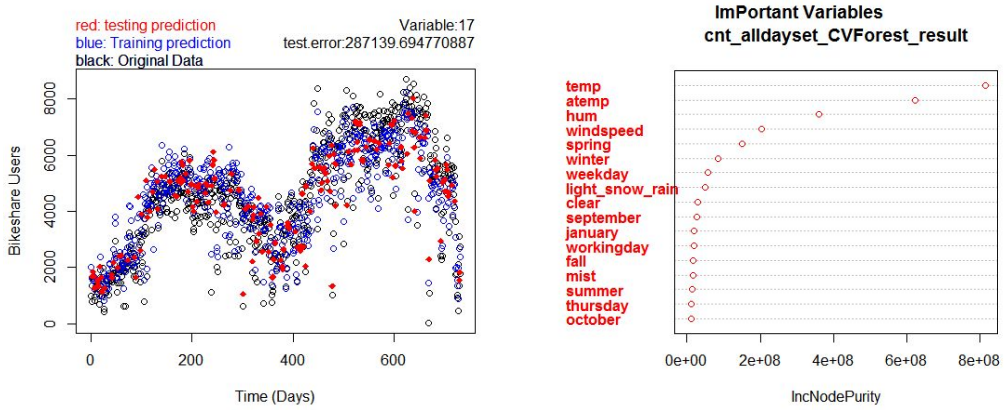
We also try to answer the first question with random forest. Following the logic above. In order to find how much predictors needed to be included, we use cross validation to determine that. The figure (a) shows that the cross validation result, so the best number of variables put in random forest is 17. The figure (b) shows that the testing error (red) and training error (blue) varying with the variables number from 1 to 20. We can see that the error becomes gradually smooth.

(a) Cross Validation result for Random forest (b) Testing error and Training error



The figure (c) shows the random forest result with 17 variables contained in it and the mean error of 20 random selected datasets is amazing 0.29 million. The figure (d) give us the 17 most important variables, they are temperature, feeling temperature, humility, windspeed, spring, winter, weekday, light_snow_rain, clear, September, January, mist, workingday, fall, summer, November, Thursday.

(c) Random Forest prediction Plot (b) Important Variables for Random Forest



Classification – LDA, QDA, Logistic and KNN

In this part, I will use four methods—LDA, QDA, Logistic and KNN together to do classification. By using cross validation, I can choose the best one from these four. And I can use my best results and the results from my teammate' classification tree to do cross validation again and get the final optimal results.

My part mainly consists of 4 sections:

1. Predicted the total number of users(cnt) using all of the predictors.
2. Tried different methods to modify the model and got improvements as well as failure experience.
3. Predicted the number of registered users and casual users using all of the predictors and came up with the solutions to increasing revenue for the Bikeshare Company.
4. Explored the influence from the specific weather terms (including temperature, perceived temperature, windspeed, humidity and weather situation) and time related terms (year, season, month, day, weekday, holiday, working day etc.)

Section 1: Predict the total number of users(cnt) using all of the predictors denoted as cnt01.

We set the original cnt as a binary variable cnt01:

Cnt01[cnt >= mean(cnt)] = 1

Cnt01[cnt < mean(cnt)] = 0

In other words, when cnt larger or equal to the average, I set it as 1. Otherwise, I set it as zero.

Taking LDA as an example, the code was written as :

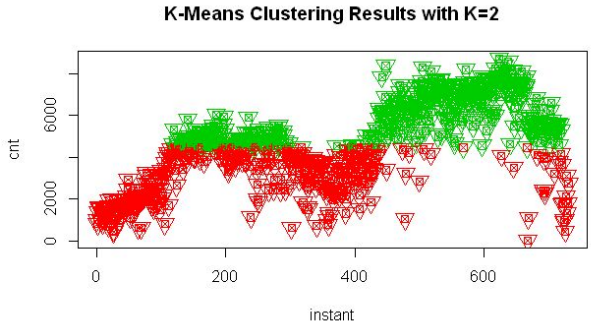
lda.fit=lda(cnt01~ yr + season + mnth + weekday + weathersit + holiday + weekday + workingday + temp + atemp + hum + windspeed ,data=dayset, subset=train)

Here I just use 'season', 'month', 'weekday' as predictors instead of binary those predictors like 'spring', 'summer', 'january', 'february', 'monday', 'tuesday' etc. because these two type of predictors have same result in dealing with classification. The best results are given by LDA, QDA and logistic with test error of 0.148 and type I error of 0.131 and type II error of 0.161. The confusion matrix is as the following:

	0	1
0	73	11
1	16	83

Section 2: Tried different methods to modify the model and get improvements.

Firstly, we considered changing the way of splitting cnt. We implemented clustering method k-means to split cnt to two groups. Cluster 1 stands for higher demand for bikes showing as green points and cluster 2 stands for lower demand of the bikes showing as red points in each day.



LDA, QDA and logistic gave the best result with test error of 0.137 and type I error as 0.123 and type II error as 0.147. The confusion matrix is as following:

	0	1
0	71	10
1	15	87

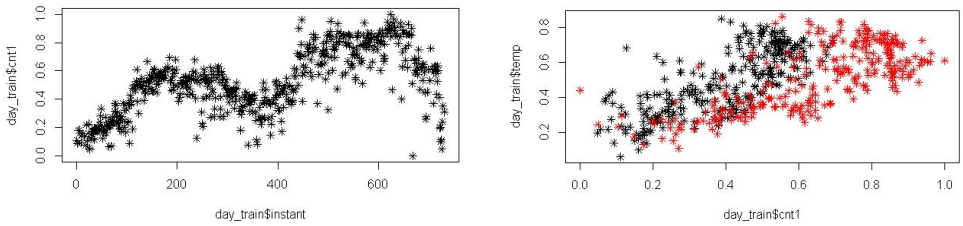
Comparing with the section one, the test error reduced as well as the type I error and type II error, which means the k-means method is a more reasonable way to separate the cnt.

Secondly, we consider changing the season variable with a clever method: We set the four seasons in coordinate system with spring equals to (1,0), summer (0,1), fall (-1,0) and winter (0,-1). If you plot these four points, you will get a diamond which means the four seasons are consisted with each other. So we change 'season' variable to 's1' and 's2' correlated with their coordinate. LDA and QDA give the best result with test error of 0.148 and type I error of 0.136 and type II error of 0.157. The confusion matrix is as following:

	0	1
0	70	11
1	16	86

The method of change of season leads to a higher test error (higher than the previous 'season' one), which means the clever method is not good. Probably the disturb of the negative coordinates lead to a higher test error rate. Finally I gave up the clever method in the later prediction models.

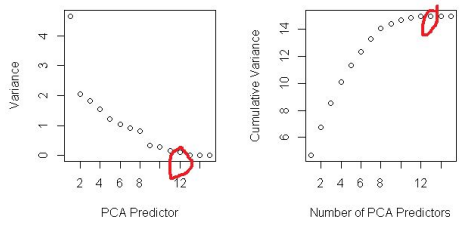
Then, by looking at the data I found another way to improve the classification.



In the first plot, x-axis is the change of time and y-axis is cnt. From tht plot, we can find the increasing trend for cnt. In the second plot, red plots stand for the 2nd years data while the black means the 1st year. It is obvious for us to say that the second year has more bike users than the first year and the two distribution seems isolated. If we just focus on the second year's data or the 1st year, we can get more precise conclusion of how the weather and specific time influence the number of total users. In other words, the difference of the two years results noise on other predictors. So we just focus on the 2nd year's data. By doing the training data and test data splitting processing and setting the binary cnt again, we get the best result of logistic model with test error of 0.087 and type I error of 0.074 and type II error of 0.092. The test error has been reduced a lot, which means focusing on the 2nd year's data leads to a better result. Confusion matrix:

glm.pred	0	1
0	25	2
1	6	59

Finally, we do PCA on predicting cnt. The PCA result are as follows:



We chose 12 as the optimal one because with 12 predictors we have lower damage of the data. Logistic gives us the optimal result with test error equals to 0.016 and type I error equals to 0.02 and type II error equals to 0.01. Comparing to the previous test error of 0.137, PCA decreased the test error to a large extent. Confusion matrix:

glm.pred	0	1
0	85	2
1	1	95

In sum, after PCA, we got the lowest test error in my section 2. PCA did a really good job.

Section 3: Predicted the number of registered users and casual users using all of the predictors.
Firstly, we looked at the plot of changing number of the registered users and the casual users with the change of time. The green point means higher demand while the red one means the lower demand.

I found that the number of registered users increased in the second year, while the casual users seem have a fixed pattern with densely distributed red points as well as the loosely distributed green points in summer and fall. By doing the prediction on registered users using all of the predictors, we get the optimal result from LDA with test error of 0.153, which is larger than the test error from predicting cnt (0.137).

	0	1
0	68	15
1	13	87

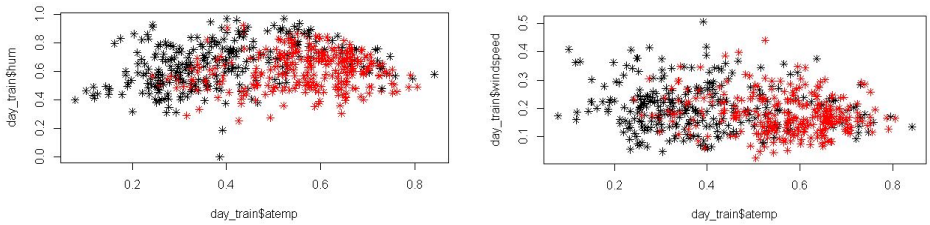
And for predicting the number of casual users, we got the optimal result from knn with k=23 and test error equals to 0.0765. The test error of predicting casual users is much smaller than predicting cnt (0.137).

In sum, the higher test error of predicting the number of registered users infers that it is harder to predict the number of registered users because the number of registered are changing year by year even day by day. The number of membership could increase according to discounted activities or other attractive activities and the number of users could increase at the same time. However, the lower test error of predicting the number of casual users makes it much easier to predict, which reflects the truth that the number of casual users have a fixed pattern just correlated with weather and date.

In order to make more profits for the Bike sharing company, they need to do more things to attract the number of registered users, which means they need to have more VIP customer.

Section 4: Explored the influence from the specific weather terms and time related terms.

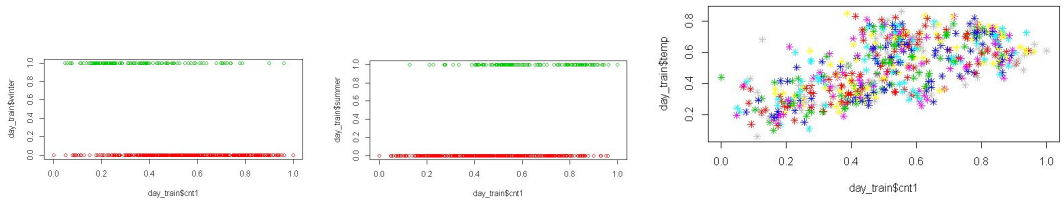
Weather related:



The red point means higher demand and the black point means the lower demand with x-axis means perceived temperature. We can capture the fact that higher perceived temperature correlated to higher demand.

Logistic model gave us the result with test error of 0.743, which means we have confidence of 0,743 to say that perceived temperature correlated with total number of users.

Time related:

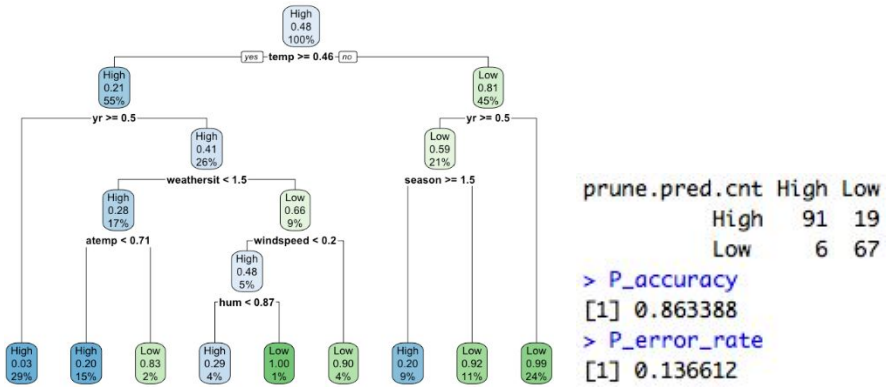


By using logistic, we can say that we have probability of 0.64 to have lower demand during winter while the probability of 0.61 to have higher demand during the summer.
The right plot shows the number of users on different day of the week (each color means each day from the week). It shows that almost all of the other time related predictors like day, month, and weekday have such a randomly distributed plot which means it is harder to make any specific predictions using time related predictors.

Classification Tree

To see how do these factors affect the number of total users, we also tried classification tree. In this part, we want to experiment with the other major CART function in R — ***rpart*** available in the *rpart library*. For comparison purposes, we applied the tree model both on raw data and improved data. And after compare with all the error rates, classification tree performed better with raw data —but only a bit. And our best results are shown below.

Pruned Classification Tree Based on CNT (raw data)




```
prune.pred.cnt High Low
          High   91  19
          Low     6  67
> P_accuracy
[1] 0.863388
> P_error_rate
[1] 0.136612
```

Figure 1. Classification Tree Based on CNT (raw data)

Figure 1 indicates that the most important predictors for predicts cnt are tempe-rature, year, season, weather, wind-speed, feeling temperature, and humidity.

On the whole, high demand occurs when temperature higher than 0.46, which is about 19°C or 66°F. To be more specific, in 2012 as long as temperature higher than 0.46, the number of bike-share users is all classified as "High". However, in 2011, high demand occurs when the weather is clear and the feeling temperature lower than 0.71, which is 29°C or 84°, otherwise the bike demand number will be classified as "Low". If the weather is not clear, "High Demand" also happens when humidity lower than 0.87 and wind-speed slower than 0.2, otherwise the cnt number will be classified as "Low".

on the contrary, low demand occurs when temperature higher than 0.46. But interestingly, in 2012, the demand of rental bikes in summer, fall even winter is classified as "High".
And for this Classification tree, the *test error* is 0.136612, which means the model is right about 86.3% of the time.

Expect settlement of above-mentioned subject matters, we are also interested in how do these factors affect the number of registered users and casual users. So, we did the same thing based on Registered user and Casual users respectively.

Pruned Classification Tree Based on Registered (raw data)

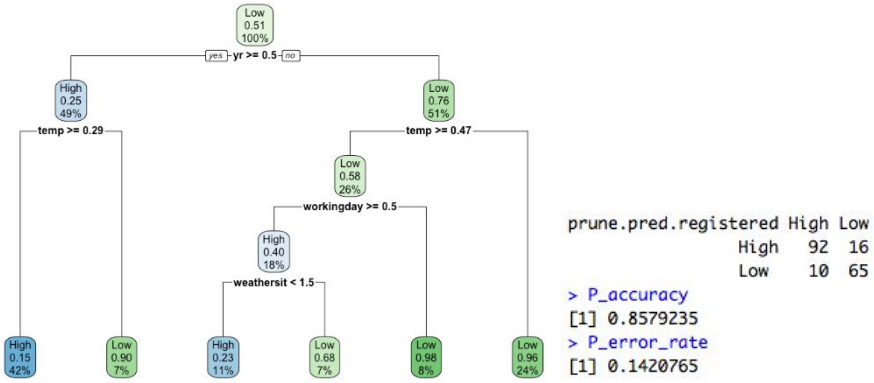


Figure 3. Classification Tree Based on Registered Users (raw data)

Form tree we can see that registered users in 2012 is generally more than the it was in 2011. In 2012, less register users rent bikes only when the temperature is higher than 11.89°C or 53°F. In 2011, the bike demand was low when the weather is clear, which is make sense because when temperature is higher 19°C or 66°F, ride bike in hot sunny day probably will not be comfortable.

Pruned Classification Tree Based on Casual (raw data)

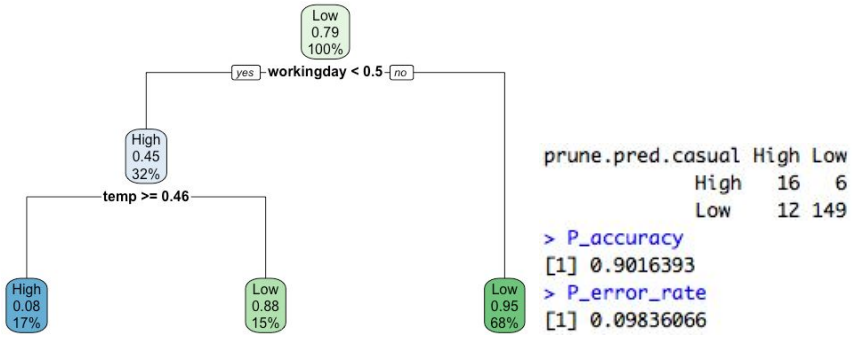
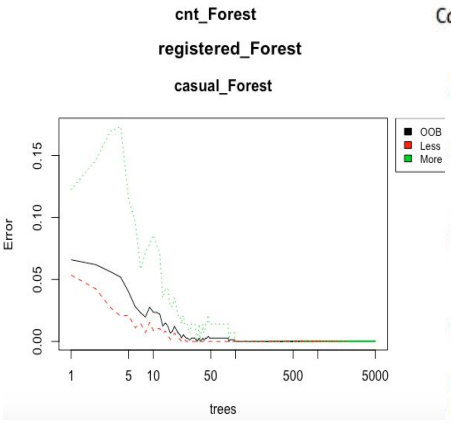


Figure 5. Classification Tree Based on Registered Users (raw data)

Figure 5 indicates that every year casual users rented more bikes in warm non-working days, rather than cold non-working days. And not a lot casual users rent bike during working-days.

Random Forest (Bagging)

To improve the results of classification tree, we also applied Random Forest to our two datasets.




```
Call:
 Call:
Call:
 randomForest(formula = cluster ~ ., data = dayset,
mtry = sqrt(11),      ntree = 5000)
               Type of random forest: classification
                     Number of trees: 5000
No. of variables tried at each split: 3

        OOB estimate of  error rate: 7.52%
Confusion matrix:
     Less More class.error
Less  570   19  0.03225806
More   36  106  0.25352113
```

Topper plot shows the bagging result based on cnt, the OOB error is shown on the right —11.35%, which is lower than the test error of classification tree(13.66%). The registered forest shows the bagging result based on registered, the OOB erroe equals 11.63%, which is also lower than the test error of classification tree (14.21%). Same for casual user, bagging is more accuracy than classification tree. The OOB error is 7.52%. Thus, we can say, bagging performs better than classification tree.

Conclusion

Using linear regression and random forest, we attempted to predict the number of users with different predictors. When we predict the user amount just with time category variables, including season, month, weekday, working day, and holiday, linear regression has a mean test error of 2.24 million and random forest had a single test error of 1.99 million. When weather was introduced as an available predictor, linear regression mean test error dropped to 1.66 million and 1.60 million when best subset was sued. Random forest test error dropped to only 0.27 million. Therefore, random forest is likely significantly better than linear regression at making use of weather to predict the total number of bikeshare users. However, linear regression has an advantage over random forest when predicting over long term trends. Linear regression resulted in an average test error of 1.29 million for predicting the trend of the second year. Enforcing a long term trend improved linear regression (but as data snooping occurred, we cannot be very confident of this trend holding true for a third year) for predicting number of users during the second year. Random forest is unable to extrapolate to future years since its advantages are restricted to the range of the training data. Future work could include combining linear regression and random forest techniques to leverage the advantages of both. This could be done by fitting a long term linear regression to the training time span of the data, subtracting the predictions away, and then training a random forest on the residuals to model the effects of the weather. The overall prediction of the number of bikeshare users would be the sum of the individual predictions of linear regression for long term trends and random forest for effects of weather.

Secondly, there is the summary of classification part. Cross validation result--PCA does the best work in predicting the demand (low or high) of the total number of bike users. Random forest has a better test error than trees, LDA, QDA and logistics method in most situations. Practical result--Weather related terms have larger influence on predicting the demand of number of users than time related terms. Higher temperature and season of summer attract more users. Lower temperature and the season of winter correlated with less users. Registered users are hard to predict which means they are more flexible with the company's policy while the number of casual users are more fixed with the change of time.