**BIKE SHARING MINI PROJECT**

**Introduction:**

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return

back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return

back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of

over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic,

environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration

of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into

a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important

events in the city could be detected via monitoring these data.

**About the Data Set :**

Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions,

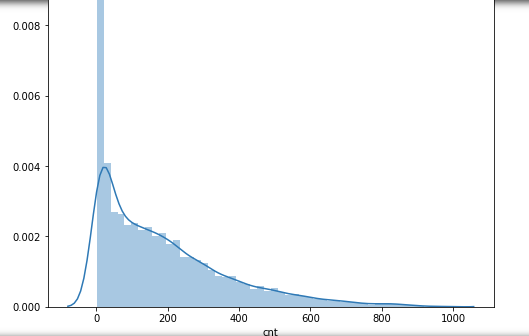
precipitation, day of week, season, hour of the day, etc. can affect the rental behaviors. The core data set is related to

the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is

publicly available in http://capitalbikeshare.com/system-data. We aggregated the data on two hourly and daily basis and then extracted and added the corresponding weather and seasonal information. Weather information was extracted from http://www.freemeteo.com.

**Data Exploration:**Data exploration was performed on the raw dataset (hourly dataset) we have 17379 rows and 17 columns.

And we plotted the graph to check the mean value of the predictor variable y which is (cnt) in this case.



So for this given task we had to perform a Linear Regression model to predict the variable ‘cnt’.

**Linear Regression and parameter selection:**

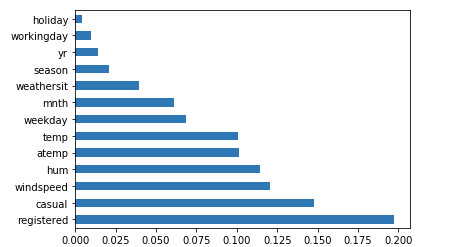
Initially we have a lot of columns to work with but on closer inspection most of them are not required. The casual and registered columns together from the cnt column for which we need to perform prediction.

Initially we take the parameters we think can mostly affect the cnt feature.

So we select the columns "weekday","holiday","workingday","temp","atemp","hum","windspeed"

And we created a Linear Regression model based off all these parameters. Upon fitting and prediction for the predictor variable y we get a Linear Model score of 0.24 which is actually less. Then we try to eliminate or add features that we think are useful.

So for selecting these appropriate features we are going to use the Extra Tree Classifier which will give us a list of all the parameters that we need to consider in our model to improve our score.



We can see that apart from registered and casual the other parameters that affect the model are

Shown in the above graph so we create a new Linear Regression model from the features given to us. On creating this model the score only mildly shifts to 0.26 which is still quite low.

So we can say that the Linear Regression Model is not decisive enough to make any conclusive predictions for the column ‘cnt’.

So features like "temp","atemp","hum","windspeed" are the maximum contributors to have any significant effect on the predictor variable.

**Random Forest Regression:**

We went a step further and performed Random Forest Regression on the hourly dataset to predict y(cnt) with the help of the features.

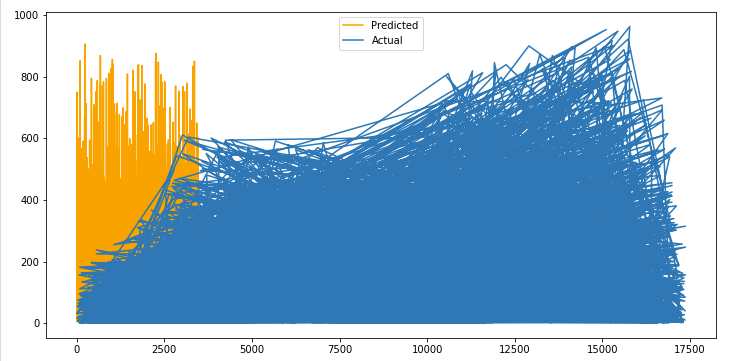
So we are going to create a function for dummy columns and we dropped the columns that have no meaning or relationship with the predictor variable..."season", "yr", "mnth", "hr", "weekday", "weathersit" are the selected features.

We split the data into a 75:25 ratio and randomly take values from the dataset to avoid bias.

Then we use the RandomForest Regressor to predict y for the test values of X

We reported a mean absolute value of 31.24 which shows that the data is scarce and the model accuracy is ~ 91.8%

We plot the predicted value of y we got from the Random Forest and also y test values that we originally split in the test set.



We can see the yellow shaded area representing the predicted values and how sparse it is in comparison to the Actual data.

**Recommendations:**

From the graphs and the data exploration and based on all the models we have created we know that the data is scattered and has a lot of points that are not being predicted properly.

We know for a fact that features like "weekday","holiday","workingday","temp","atemp","hum","windspeed" heavily influence the prediction of our y predictor ‘cnt’

So for any related operations on the hourly data and to gain more insights on the effects of these features on the entire bike share ecosystem we need to consider the above mentioned features.

**Conclusion :**

Random Forest Model Accuracy score is close to 91.8% and we have observed a 31.24 Mean absolute value.

Linear Regression is not the correct approach to predict for the predictor variable y (cnt)