Forecasting Bike Rental Demand

**Introduction:**

The bike-sharing services have been under usage in many countries in the recent years. Bike sharing service attracted considerable amount of people especially those who commute regularly to their work place or schools/colleges to avoid delays due to heavy street traffic in busy cities and also a desire for an environmentally friendly transportation. This paper examines the Capital Bike share program implemented in Washington D.C.

**Objective:**

This project aims to predict the total number of bikes rented on an hourly basis from the historic data spanning two years provided by Capital Bike share.

**Data:**

The dataset in this project is provided by Kaggle. The data includes rental and usage data of bike renting spread across two years. The trading data has 10866 observations of 12 variables, while the test data has 6493 observations of 9 variables. The training set consists of rental data for the first 20 days of each month, while testing data consists of the remaining 10 days. Our aim is to predict the demand for the remaining 10 days.

**Date time**: Date & Time   
**Season**: 1 = spring, 2 = summer, 3 = fall, 4 = winter  
**Holiday**: Either weekend or public holiday  
**Working day**: Neither a weekend nor holiday  
**Weather**:

1. Clear

2. Cloudy

3. Light snow/rain

4. Heavy Rain/snow

**Temp**: Temperature in Celsius  
**Atemp**: Temperature in Celsius   
**Humidity**: relative humidity  
**Wind speed** – wind speed  
**casual**: Non-registered users count

***Registered*** – Registered users count  
***count*** – Total rentals

**Feature Engineering:**

Some of the data from the given dataset could influence the accuracy of the prediction. For instance, the temperature of the day could impact the usage of bikes on a particular day. Just by looking at the raw data when the temperature is more than 10 degree Celsius, the number of bike sharing seems to be high. Likewise the usage of bikes could be considered as high if it morning and evening peak hours.

* There isn’t any missing value in the data hence we are good to proceed with the existing data
* Parsing the Date time
* Generated new variables like hour, month, day and year by splitting the Date time Colum.
* I presume that atemp would be more relevant for an individual’s decision of renting a bike as it is actually described as the “feels like temperature “ in the dataset
* Some hours clearly have more demand than other hours, so we looked at the data to identify peak hours for bike rental. We noticed that on weekdays, the peak hours for bike rentals were between 7-9 am and 5-7 pm. On weekends, the peak hours were anywhere between 10 am - 6 pm. Then we classified each observation depending on whether it fell into the weekday peak hours, the weekend peak hours, or none of them.

**Algorithms Used:**

**XGBoost & Gradient Boosting:**

Boosting methods work iteratively to create a new learner at every stage; these new learners are then trained on the error residuals at a current iteration to produce new learners which are stronger than the previous stage. Applied to decision trees, every decision tree is works on the error residuals of the previous iteration to produce a better decision tree. The collection of these decision trees is then used as the overall model for predicting values.

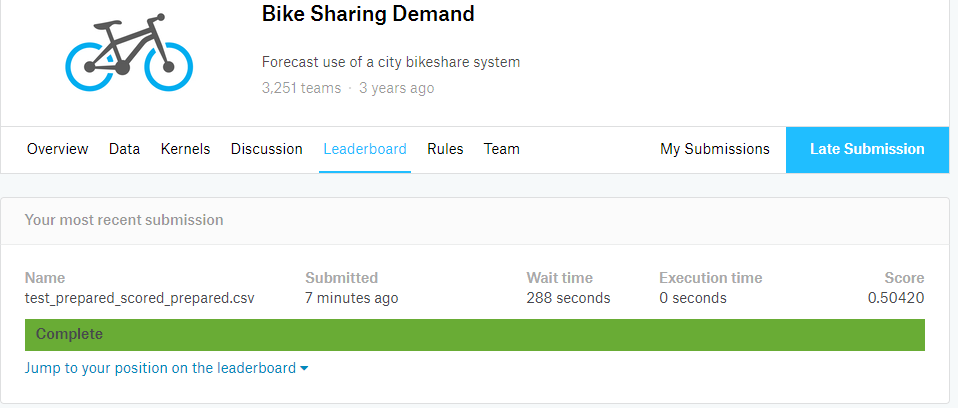
Gradient Boosting is an ensemble learning method which uses multiple weak learners which are combined to form a strong learner. Gradient Boosting as the name suggests uses boosting.

**Random Forests:**

The fundamental idea underlying Random Forests is that training a decision tree repeatedly on a dataset produces a new decision tree every time and multiple such trees reduce the overall error of the model. In this problem, the decision of using Random Forests was driven considering the weak performance of Regression As I presume that the problem is not amenable to them.

**Conclusion:**

Using the proposed models of XGboost and Randon forest, I was able to predict the accuracy of the whereas I used Random Forest as I have got the high score of -0.930. I submitted my result to Kaggle page and I have got the below score 0.50420 which ranked in the range of 1760.



**Suggestions for Future improvement.**

There are likely endless tweaks that I could do to make my model better.

* If I have the traffic data, I could have used it to improve the prediction.
* As discussed in class during the presentation, I tried to include the new feature with employment details which didn’t give me any improvement on the score. Hence I didn’t include it into my feature engineering.