Analysis of Hubway Bike Rental Data

Subhabrata Mukherjee

Dec 2017

Introduction

The objective of this analysis:

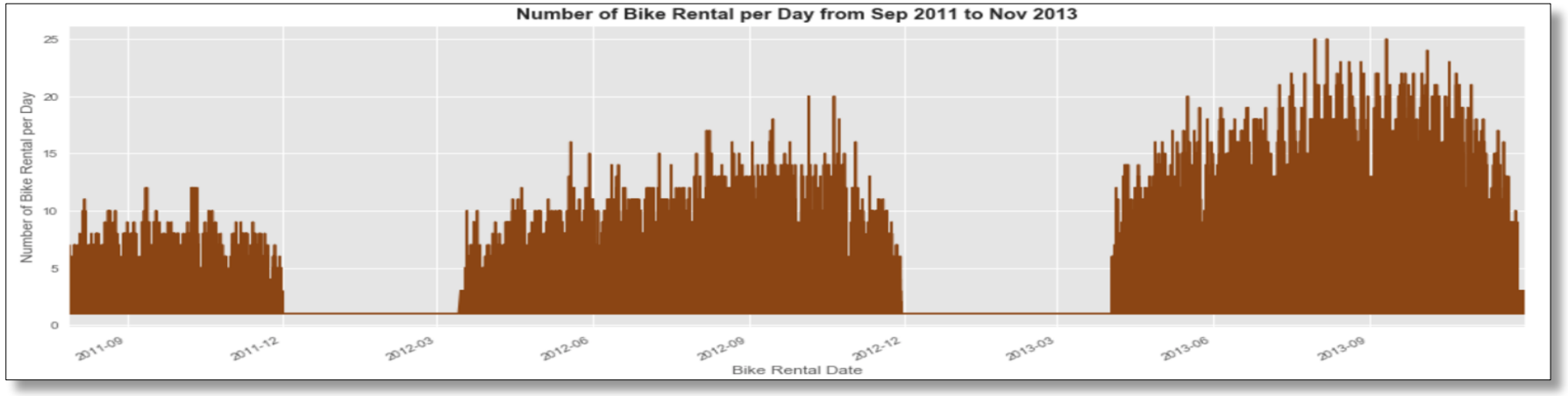
- Analyze bike rental data to know the demand for bike on weekdays
- Identify most prominent time of the day having highest demand for bike
- Create predictive model of Hubway bike demand using machine learning

Data sets

- Hubway Trip Data from 2011 to 2013
  - Date, time, origin and destination stations, plus the bike number
- Weather Data from 2012 to 2013
  - Daily temperature, humidity, wind speed

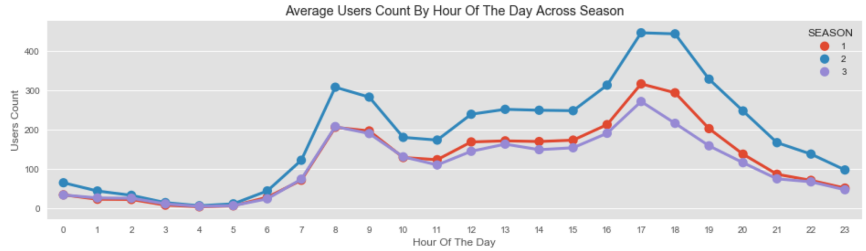
DATA EXPLORATION

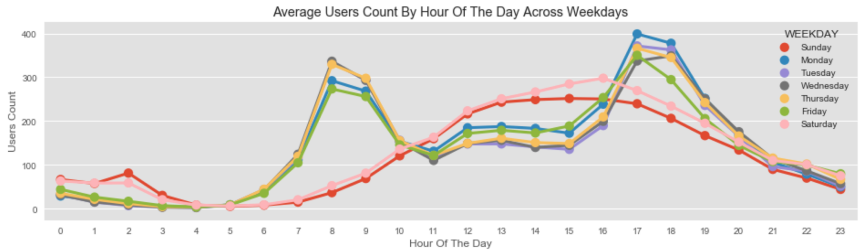
Daily Bike Rental by Year

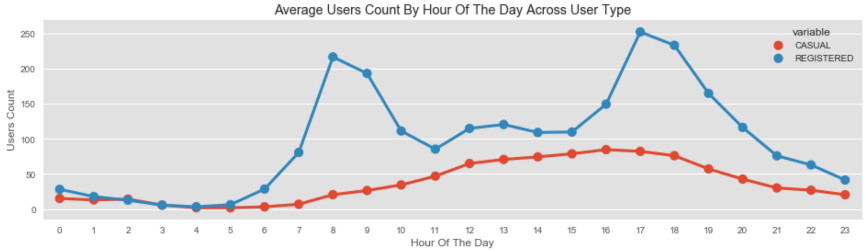


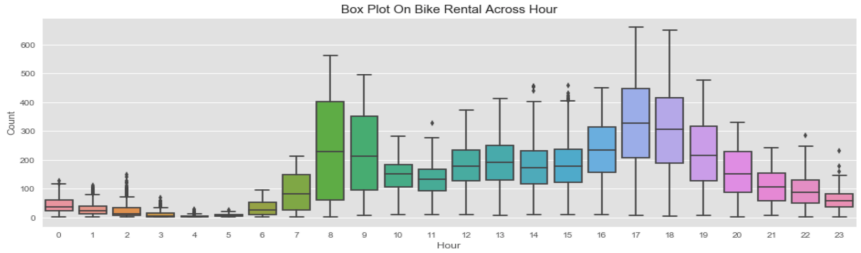
- The bike share program is gaining popularity over time.
- Weather is an important factor in daily demand of bike rental

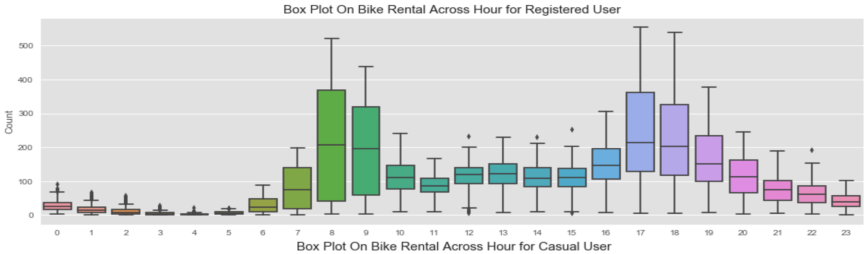
Average User Count By Hour Of The Day

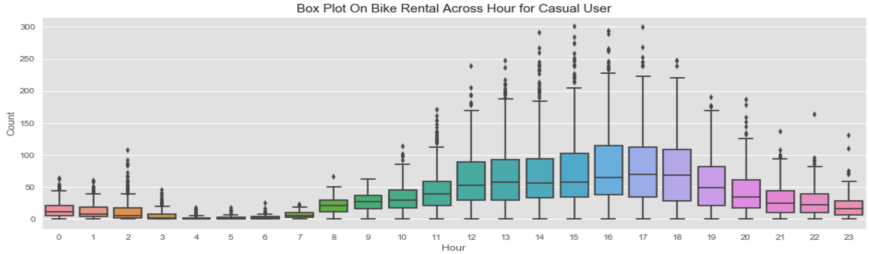








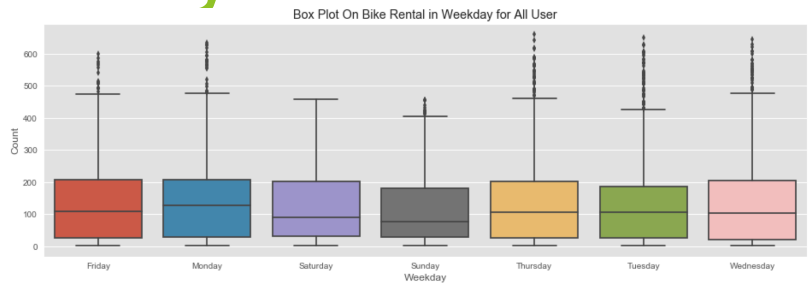


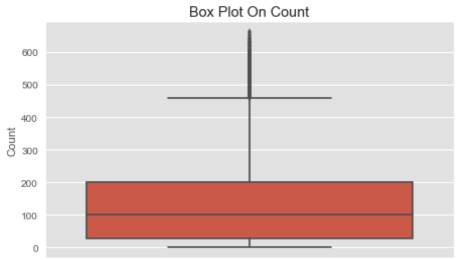


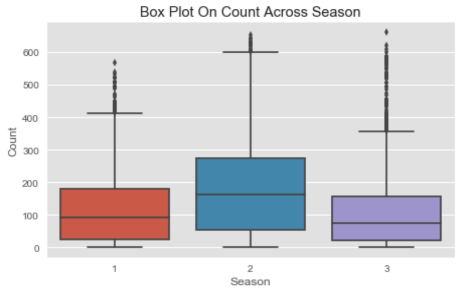
The trend of bike demand over hours has three categories:

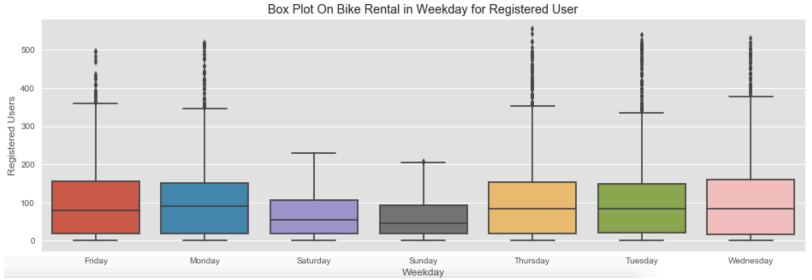
- ❑ High : 7-9 and 17-20 hours
- ❑ Average : 10-16 hours
- ❑ Low : 0-6 and 21-24 hours

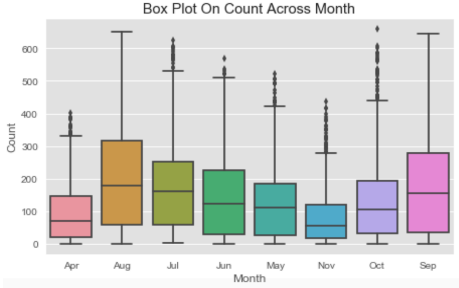
Daily Bike Rental Distribution

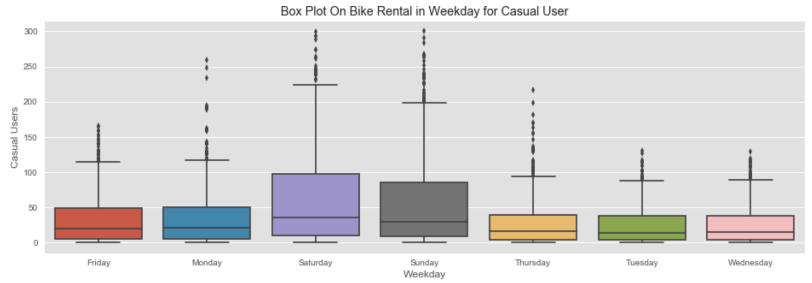


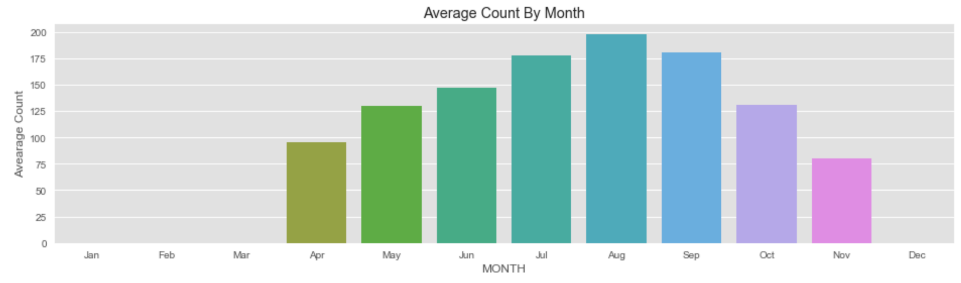






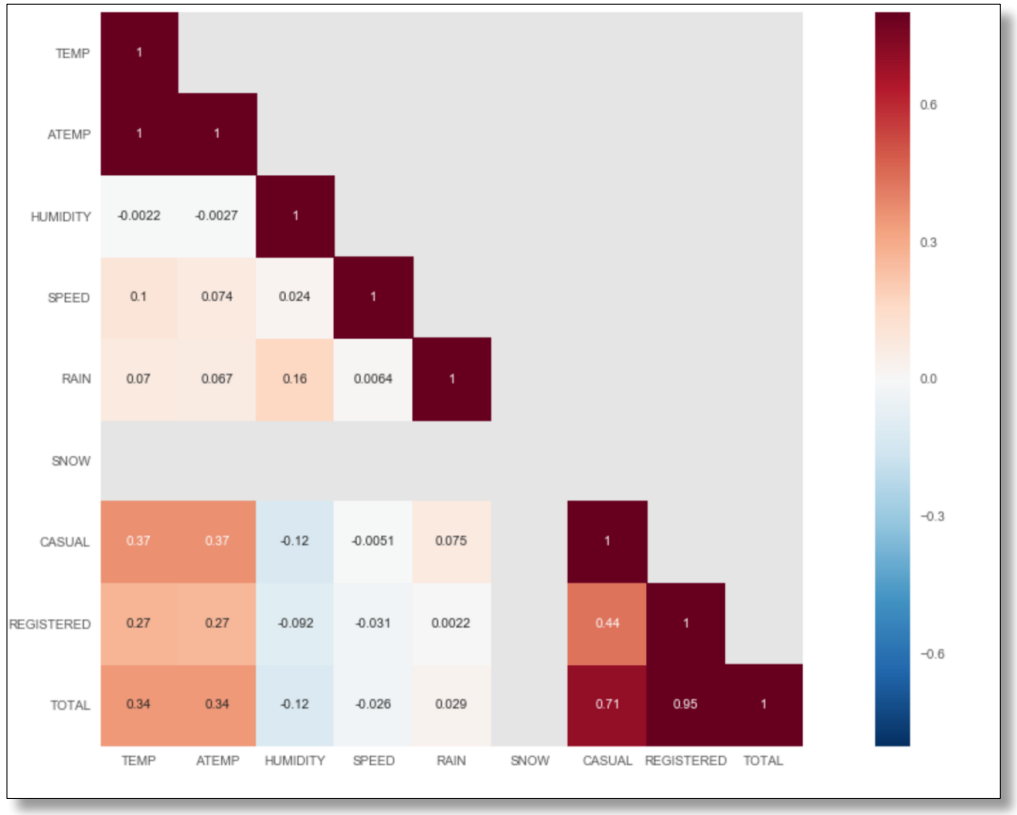






- ❑ Registered users has more demand in weekday
- ❑ Casual users has more demand in weekend
- ❑ The demand in more in Summer
- ❑ August and September has more demand

Daily Bike Rental – Other Factors



Selected 12 features	
TEMP	ATEMP
HUMIDITY	SPEED
RAIN	SNOW
CASUAL	REGISTERED

- Variable temp is positively correlated
- Feels-like-temp is correlated with bike rental
- Humidity is less correlated compared to temp
- Wind speed is less correlated compared to temp and humidity

Correlation matrix shows some relationships but none strong enough to skew the model

PREDICTIVE MODELS

Model Selection

- Following models were used to predict bike demand
  - Linear Regression
  - Random Forest
  - Gradient Boosting
- Analysis done on different scenario
  - Only Registered User
  - Only Casual User
  - Both Registered User and Casual User
  - All User with a “User Type” feature

Feature Engineering

Independent Variables

- datetime: date and hour in "mm/dd/yyyy hh:mm" format
- season: Four categories-> 1 = spring, 2 = summer, 3 = fall, 4 = winter
- holiday: whether the day is a holiday or not (1/0)
- workingday: whether the day is neither a weekend nor holiday (1/0)
- weather: Four Categories of weather

    1. Clear, Clouds
    2. Mist, Drizzle, Fog
    3. Rain, Haze
    4. Snow, Thunderstorm, Squall

- temp: hourly temperature in Fahrenheit
- atemp: "feels like" temperature in Fahrenheit
- humidity: relative humidity
- windspeed: wind speed

Dependent Variables

- Registered: number of registered user
- Casual: number of non-registered user
- Casual: number of total rentals (registered + casual)

```
Data columns (total 56 columns)

TEMP          6696 non-null int64    07PM          6696 non-null uint8
ATEMP         6696 non-null int64    08PM          6696 non-null uint8
HUMIDITY      6696 non-null int64    09PM          6696 non-null uint8
SPEED         6696 non-null int64    10PM          6696 non-null uint8
RAIN          6696 non-null float64  11PM          6696 non-null uint8
SNOW          6696 non-null float64  FRIDAY        6696 non-null uint8
SPRING        6696 non-null uint8    MONDAY        6696 non-null uint8
SUMMER        6696 non-null uint8    SATURDAY      6696 non-null uint8
FALL          6696 non-null uint8    SUNDAY        6696 non-null uint8
12AM          6696 non-null uint8    THURSDAY      6696 non-null uint8
01AM          6696 non-null uint8    TUESDAY       6696 non-null uint8
02AM          6696 non-null uint8    WEDNESDAY     6696 non-null uint8
03AM          6696 non-null uint8    APR           6696 non-null uint8
04AM          6696 non-null uint8    AUG           6696 non-null uint8
05AM          6696 non-null uint8    JUL           6696 non-null uint8
06AM          6696 non-null uint8    JUN           6696 non-null uint8
07AM          6696 non-null uint8    MAY           6696 non-null uint8
08AM          6696 non-null uint8    NOV           6696 non-null uint8
09AM          6696 non-null uint8    OCT           6696 non-null uint8
10AM          6696 non-null uint8    SEP           6696 non-null uint8
11AM          6696 non-null uint8    WORKDAY       6696 non-null uint8
12PM          6696 non-null uint8    HOLIDAY       6696 non-null uint8
01PM          6696 non-null uint8    WEEKEND       6696 non-null uint8
02PM          6696 non-null uint8    WEEKDAY       6696 non-null uint8
03PM          6696 non-null uint8    CLEAR         6696 non-null uint8
04PM          6696 non-null uint8    DRIZZLE       6696 non-null uint8
05PM          6696 non-null uint8    RAIN          6696 non-null uint8
06PM          6696 non-null uint8    STORM         6696 non-null uint8
```

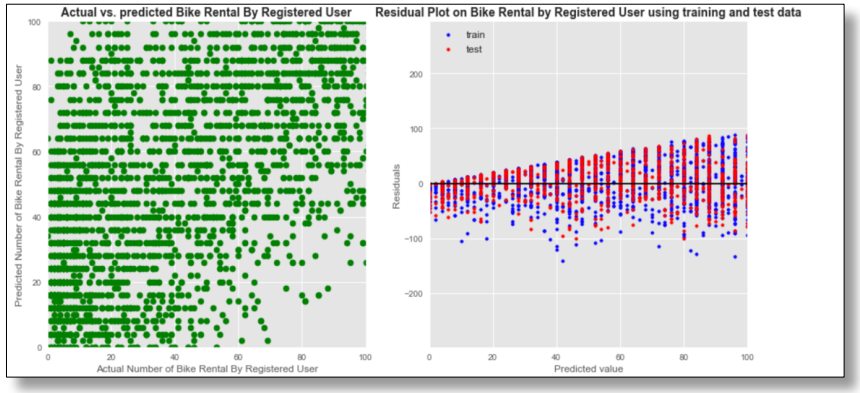
Result - All Model

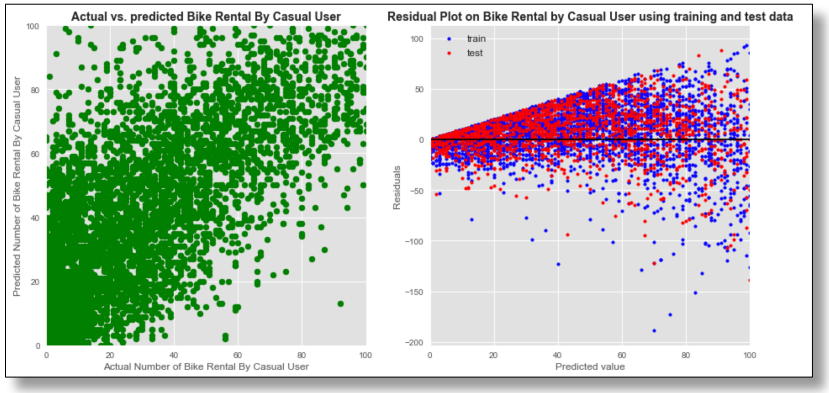
Model	Registered User (M1)	Casual User (M2)	All User (M3)	All User (Combined)
Linear Regression	0.6	0.58	0.63	0.5
Random Forest	0.81	0.76	0.81	0.81
Gradient Boosting	0.86	0.84	0.86	0.86

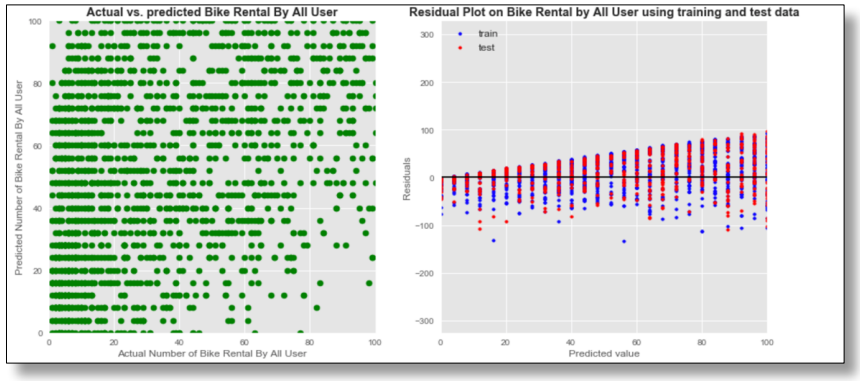
Model	Registered User (M1)	Casual User (M2)	All User (M3)	All User (Combined)
Linear Regression	3776.83	830.7	5714.19	3490.64
Random Forest	2138.91	536.22	3338.04	1401.63
Gradient Boosting	1603.58	357.43	2443.98	1034.96

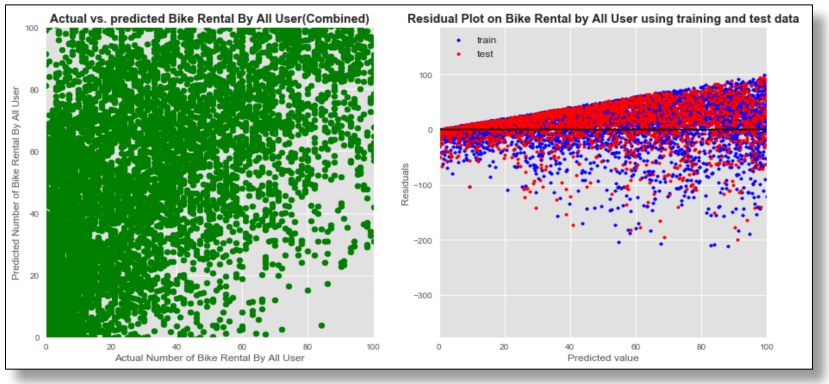
After comparing R2 & MSE the All User (Combined) model has the best prediction

Linear Regression



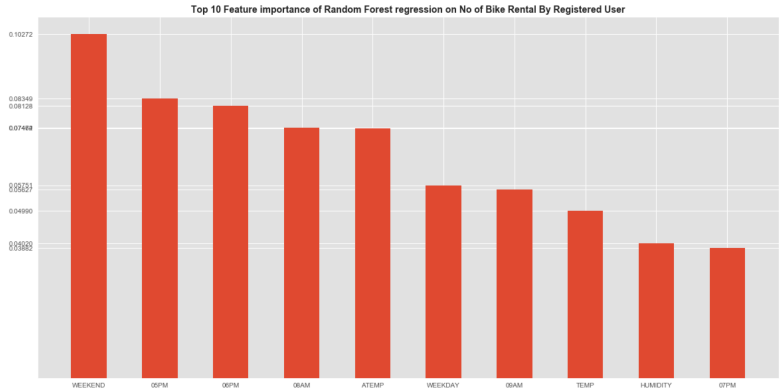


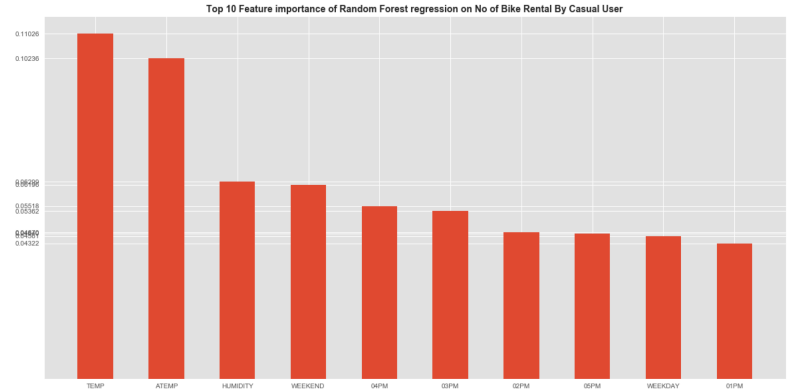


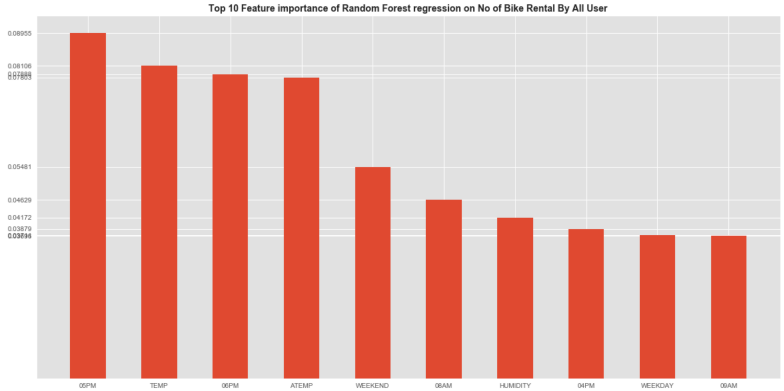


- ❑ Registered User Model explains 60% of the variance in the data with MSE on test data is 3776.83
- ❑ Casual User Model explains 58% of the variance in the data with MSE on test data is 830.70
- ❑ All User Model explains 63% of the variance in the data with MSE on test data is 5714.19
- ❑ All User Model (Combined) explains 50% of the variance in the data with MSE on test data is 3490.64

Random Tree Regression

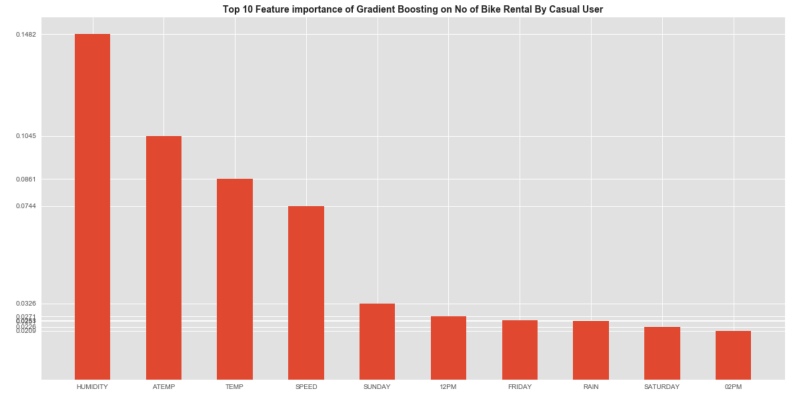


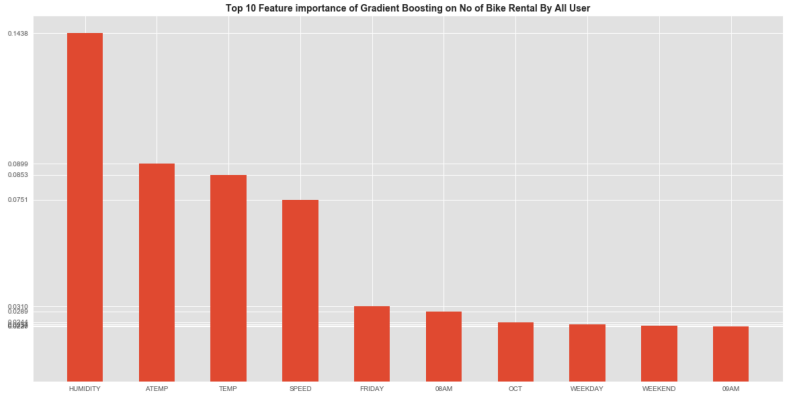


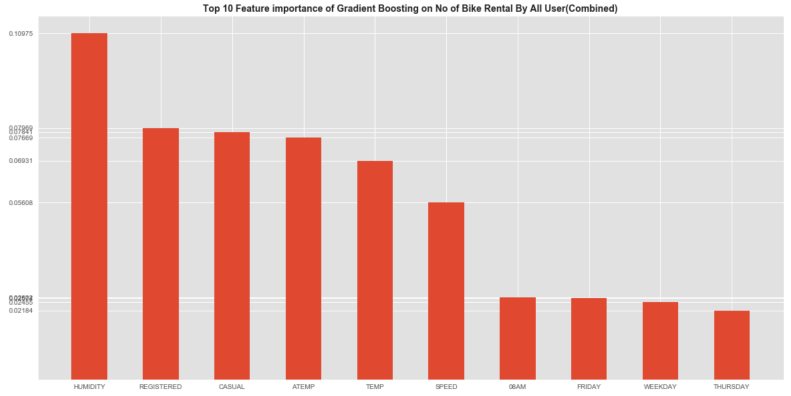


- ❑ Registered User Model explains 81% of data's variance; Weekend and 05PM are most important features during training
- ❑ Casual User Model Model explains 76% of data's variance; Temp and Feels-Like-Temp are most important features during training
- ❑ All User Model explains 81% of data's variance; 05PM and Temp most important features during training
- ❑ All User Model (Combined) explains 81% of data's variance; Registered and Temp most important features during training

Gradient Boosting Regression







- ❑ Registered User Model explains 86% of data's variance; Humidity and Feels-like-Temp are most important features during training
- ❑ Casual User Model explains 84% of data's variance; Humidity and Feels-Like-Temp are most important features during training
- ❑ All User Model explains 86% of data's variance; Humidity and Feels-Like-Temp are most important features during training
- ❑ All User Model (Combined) explains 86% of data's variance; Humidity and Registered are most important features during training

Recommendations and future work

Based on the performance of All User (Combined) model, the recommendation is to use this model to predict bike rental demand on a given day.

- For Registered User:
  - Conduct a study to determine the root cause if there is any relationship between a stating station and ending station and then predict the demand for such combination
- For Casual User:
  - Investigate if the number of casual user can be converted in to registered user so that the demand can be prediction with more accuracy .

Future opportunities to continue this work include further training of the regression model on similar city with more accurate dataset to determine if the demand for bike rental holds good.

Thank You