Analysis of Hubway Bike Rental Data

Subhabrata Mukherjee

Dec 2017

Introduction

The objective of this analysis:

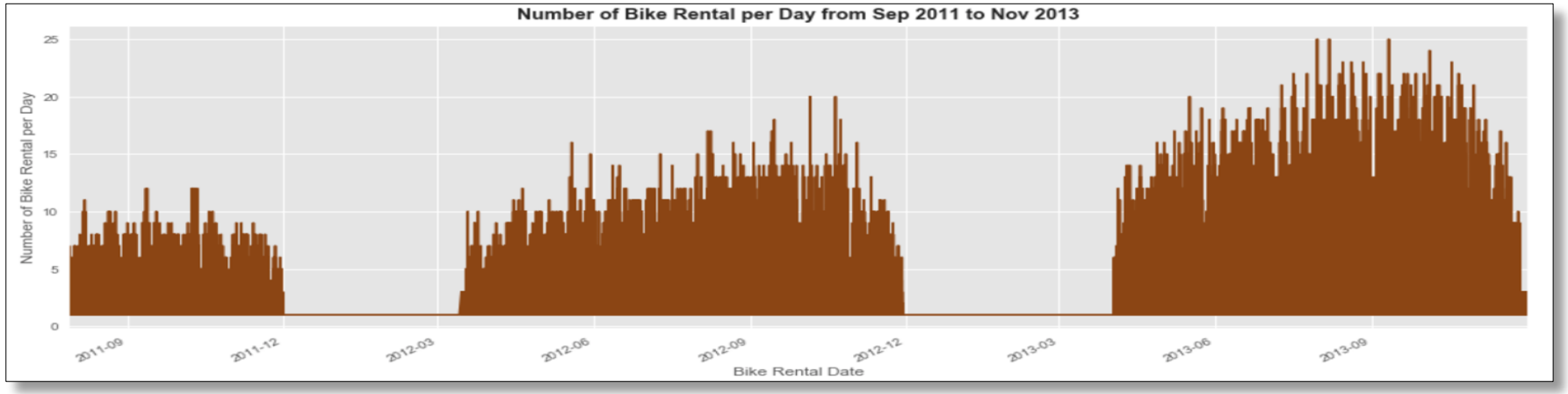
- Analyze bike rental data to know the demand for bike on weekdays
- Identify most prominent time of the day having highest demand for bike
- Create predictive model of Hubway bike demand using machine learning

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

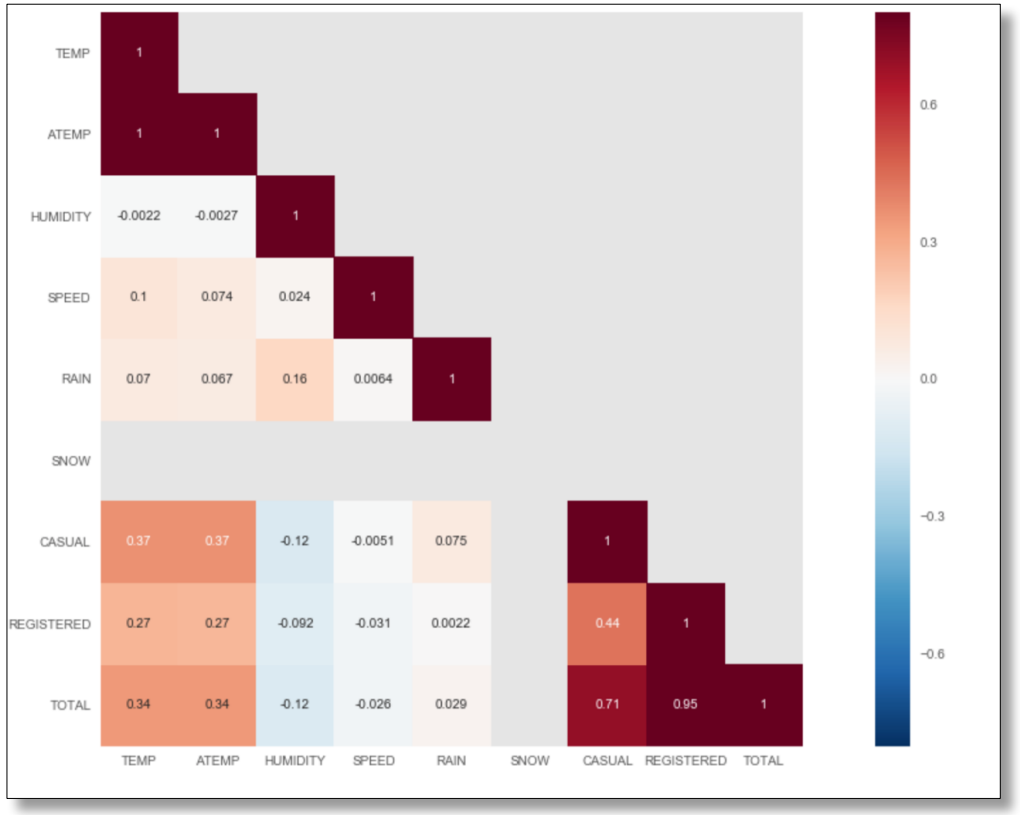
Daily Bike Rental by Year



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

PREDICTIVE MODELS

Linear Regression

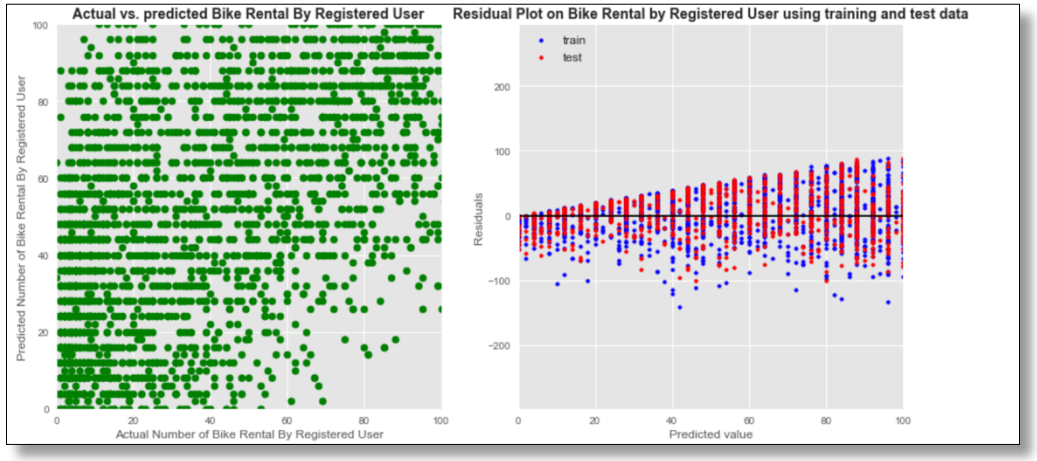


Selected 12 features	
TEMP	ATEMP
HUMIDITY	SPEED
RAIN	SNOW
CASUAL	REGISTERED

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

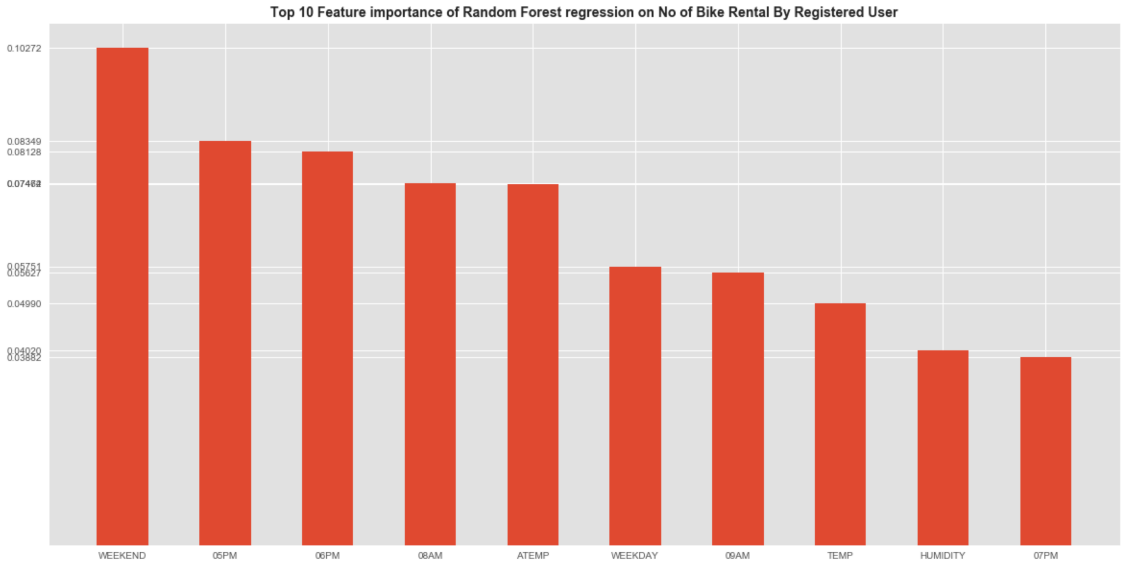
Correlation matrix shows some relationships but none strong enough to skew the model

Linear Regression – Registered User



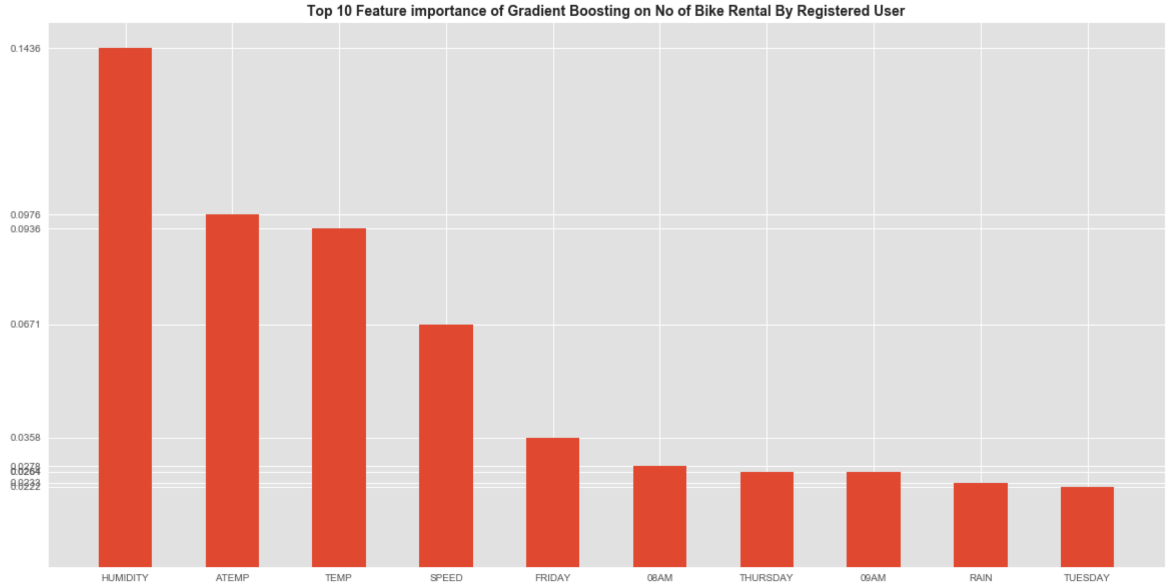
- Model explains 60% of the variance in the data – a moderate value for fit
- Training data has a MSE = 3720.20 and test data has a MSE = 3776.83
- Linear regression is not a good choice for this data set

Random Tree Regression – Registered User



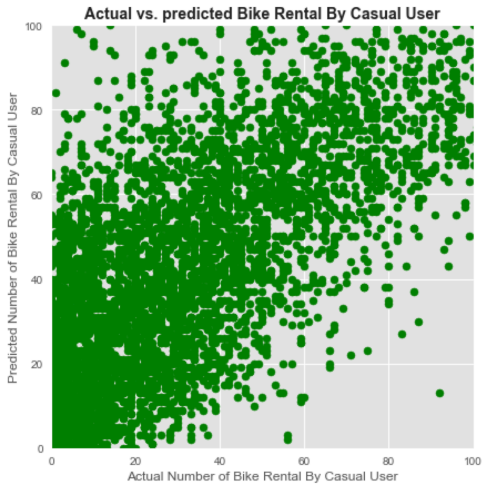
- Model explains 81% of data's variance but does not generalize well to the test set
- Weekend and 05PM are most important features during training

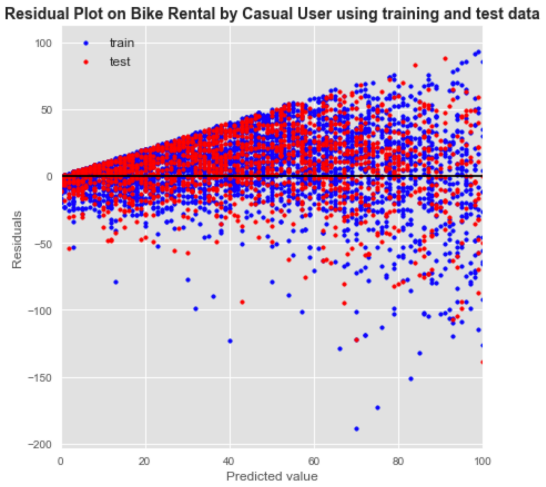
Gradient Boosting Regression – Registered User



- Model explains 86% of data's variance but does not generalize well to the test set
- Humidity and Feels-like-temperature are most important features during training

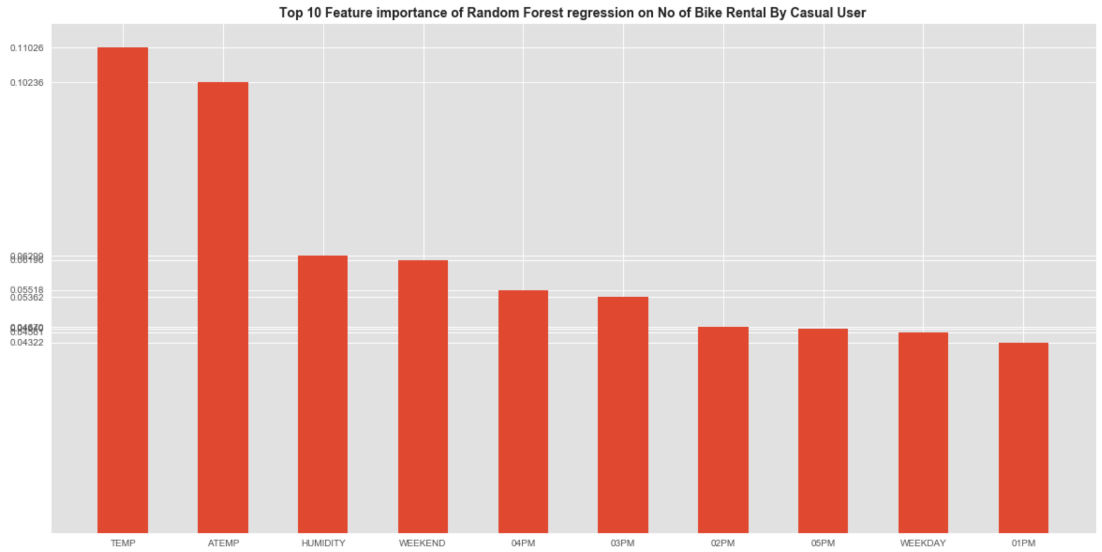
Linear Regression – Casual User





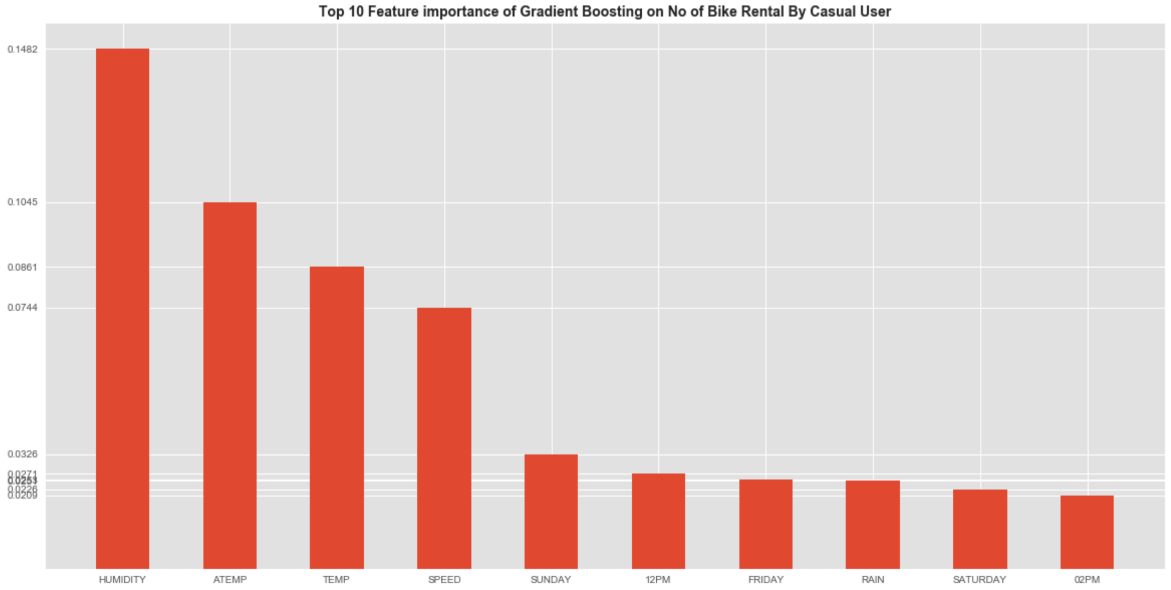
- ❑ Model explains 58% of the variance in the data – a moderate value for fit
- ❑ Training data has a MSE = 897.87 and test data has a MSE = 830.70
- ❑ Linear regression is not a good choice for this data set

Random Tree Regression – Casual User



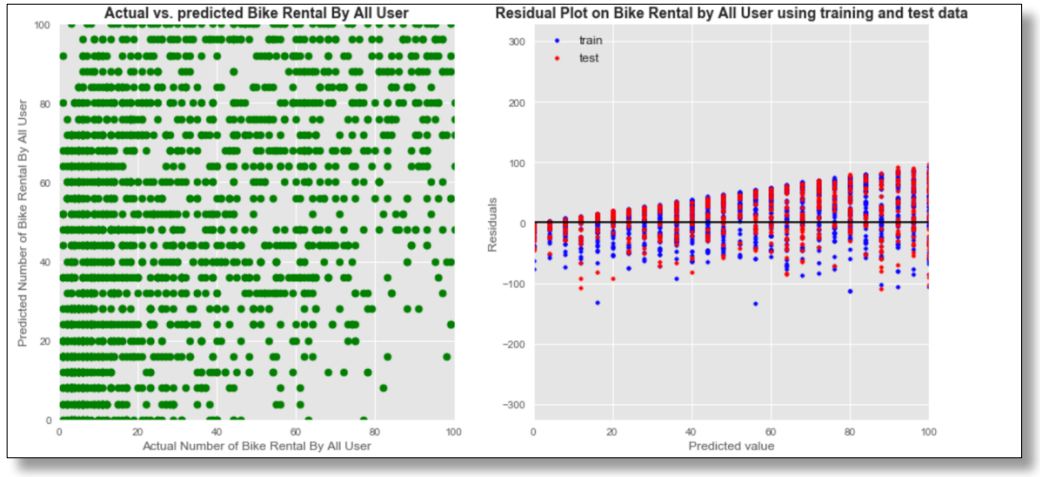
- Model explains 76% of data's variance but does not generalize well to the test set
- Temperature and Feels-Like-Temperature are most important features during training

Gradient Boosting Regression – Casual User



- Model explains 84% of data's variance but does not generalize well to the test set
- Humidity and Feels-Like-Temperature are most important features during training

Linear Regression – All User

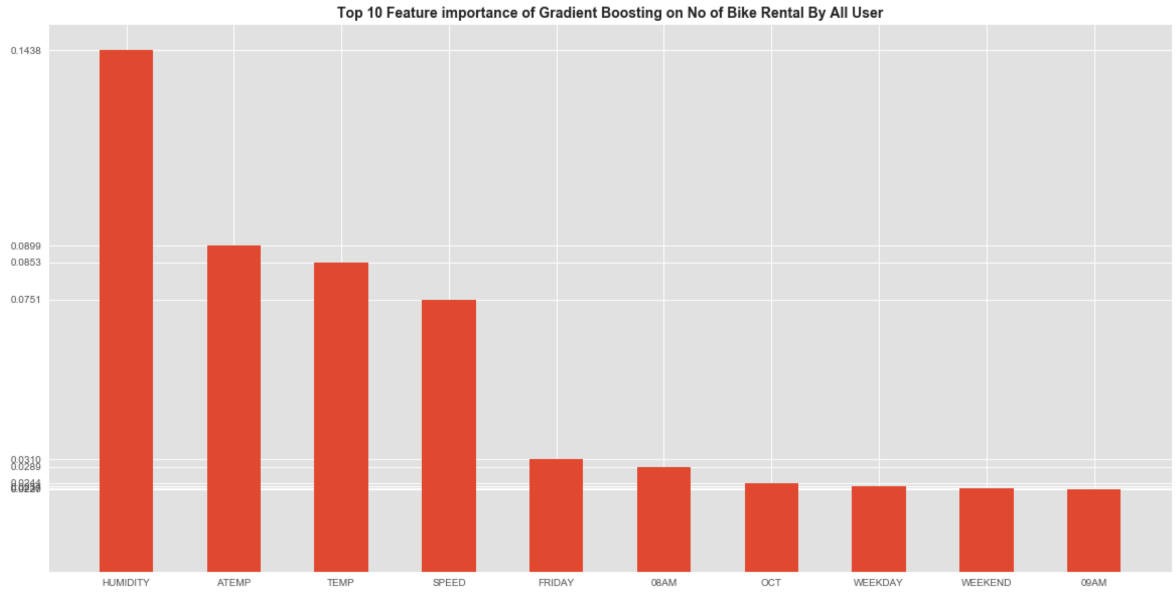


- Model explains 63% of the variance in the data – a moderate value for fit
- Training data has a MSE = 5756.49 and test data has a MSE = 5714.19
- Linear regression is not a good choice for this data set

Random Tree Regression – All User

- Model explains 81% of data's variance but does not generalize well to the test set
- 05PM and Temperature most important features during training

Gradient Boosting Regression – All User



- Model explains 86% of data's variance but does not generalize well to the test set
- Humidity and Feels-Like-Temperature are most important features during training

Compare R2 & MSE - 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

All User (Combined) model has similar R2 compare to M3 but MSE is less.

Recommendations and future work

- 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.