Bike Sharing Prediction

Shashank V. Maiya

Data Science Intensive Capstone Project

October 1, 2018 Cohort

Springboard

The Problem

- Bike Sharing Facilities very prevalent in major metropolitan cities
- Used by commuters for daily office commutes
- Used by tourists for short distance travel
- For example in Washington D.C, Number of bikes rented out at a particular time varies from <10 to 1000

- What factors affect Bike Sharing rental count?
- How many Bikes will be required at a given time of the day?

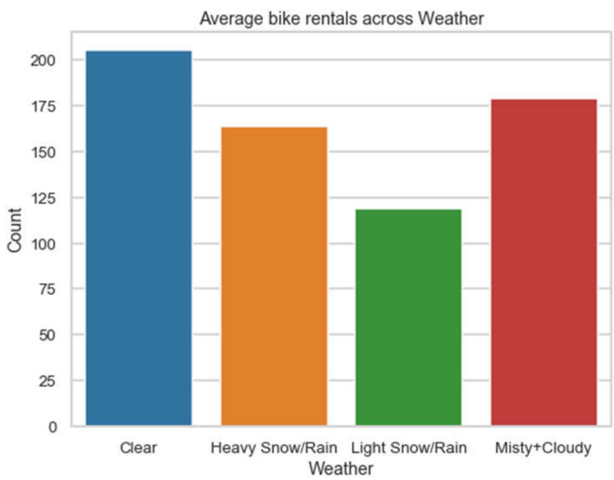
Who might care?

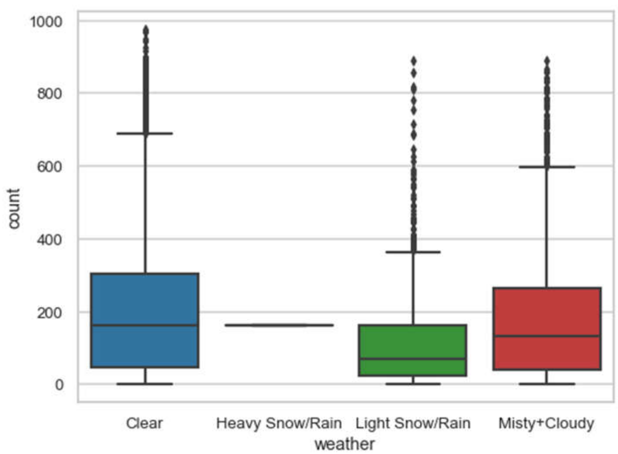
- Bike Company Vendors
  - Capital BikeShare
  - Citi Bike
  - Bird
- Mobile Apps
- Kiosks/Bike stores
- Government Bodies – Parking Facilities

Data Overview

	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
datetime											
2011-01-01 00:00:00	Spring	0	0	Clear	9.84	14.395	81	0.0	3	13	16
2011-01-01 01:00:00	Spring	0	0	Clear	9.02	13.635	80	0.0	8	32	40
2011-01-01 02:00:00	Spring	0	0	Clear	9.02	13.635	80	0.0	5	27	32

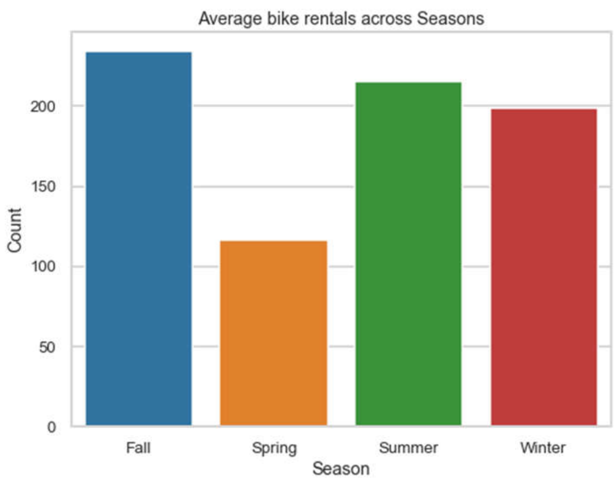
- Data set obtained from Kaggle
- Factors that affect Bike Sharing count
  - Weather conditions – Temperature, Humidity, Windspeed
  - Day – Working day or not
  - Time of the day

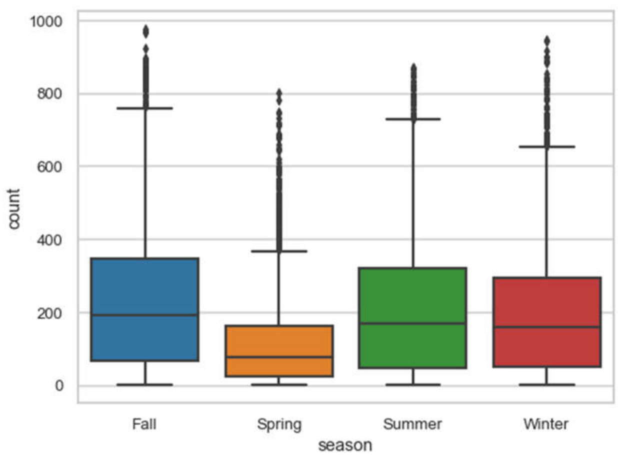




Exploratory Data Analysis – Weather

- Higher bike rental when weather is more clear and sunny
- Single instance of a Heavy Snow/Rain condition → Changed to Light Snow/Rain condition



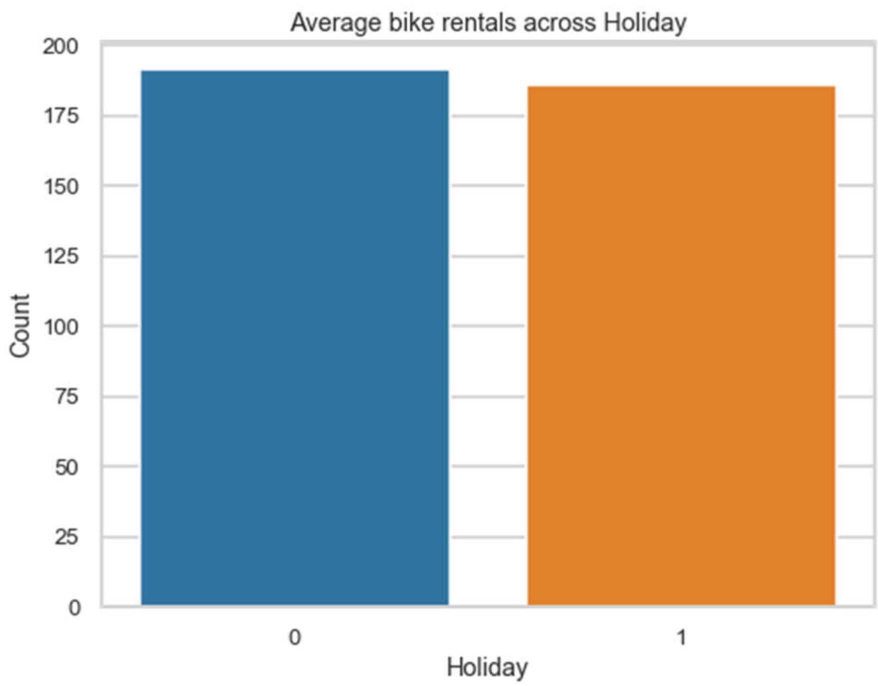


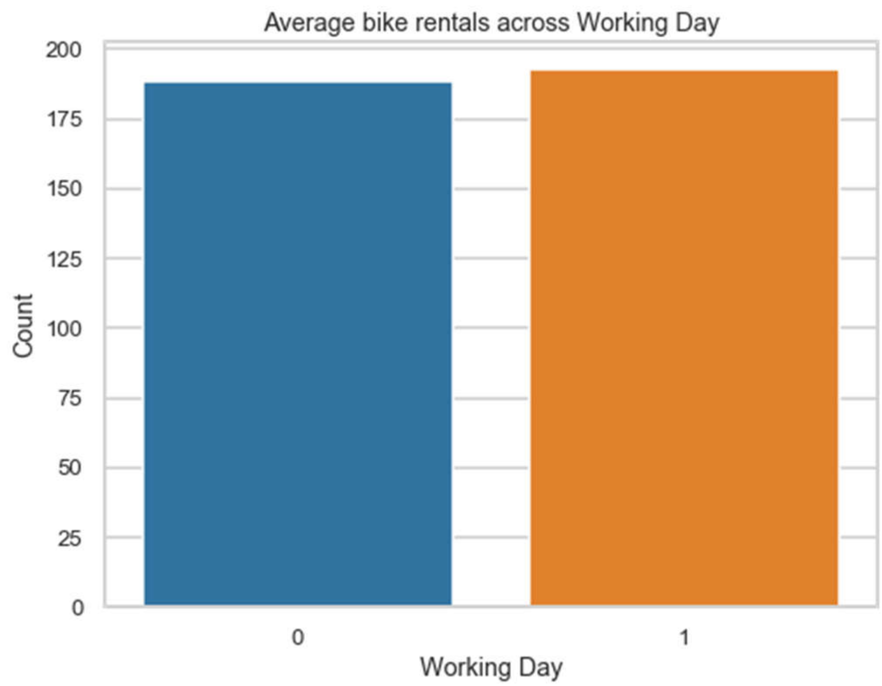
Exploratory Data Analysis – Season

- Highest bike reservations during Summer (April to June) and Fall (July to September) and lowest in Spring (January to March)

Exploratory Data Analysis – Working Day

Overall average bike rental count on a Working day or Non-working day are sa

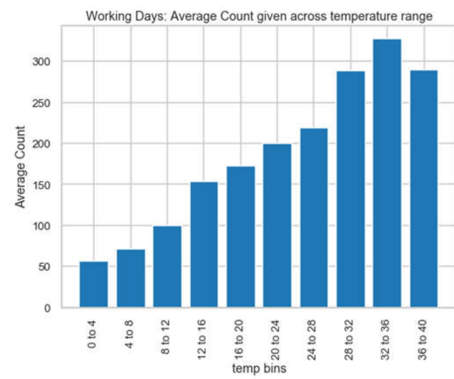


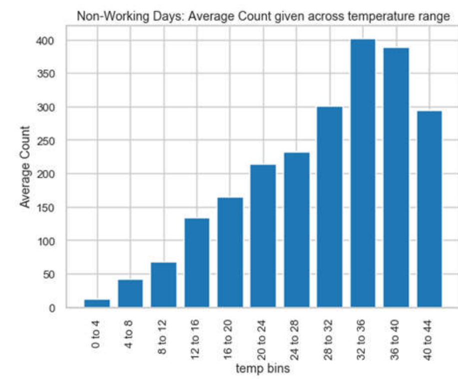


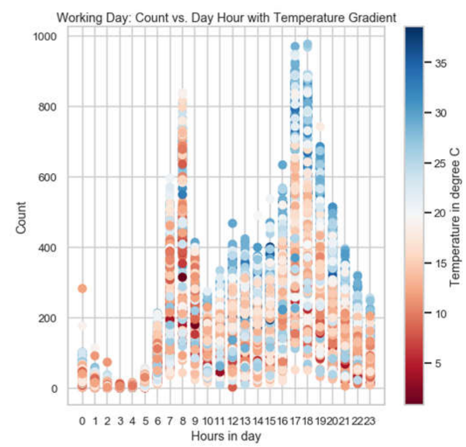
Exploratory Data Analysis - Temperature

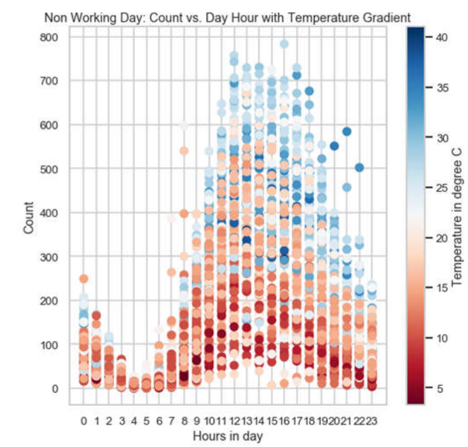
Steady increase in biking count with temperature

Ideal temperature for biking is between 32 and 36 degree Celsius

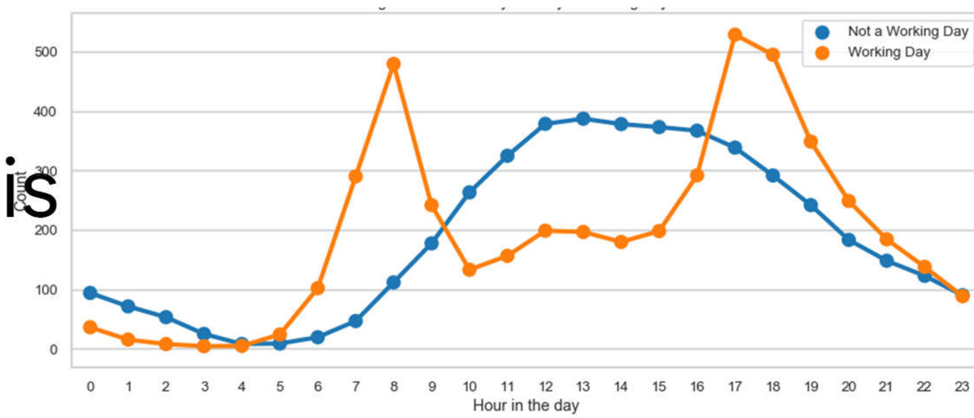




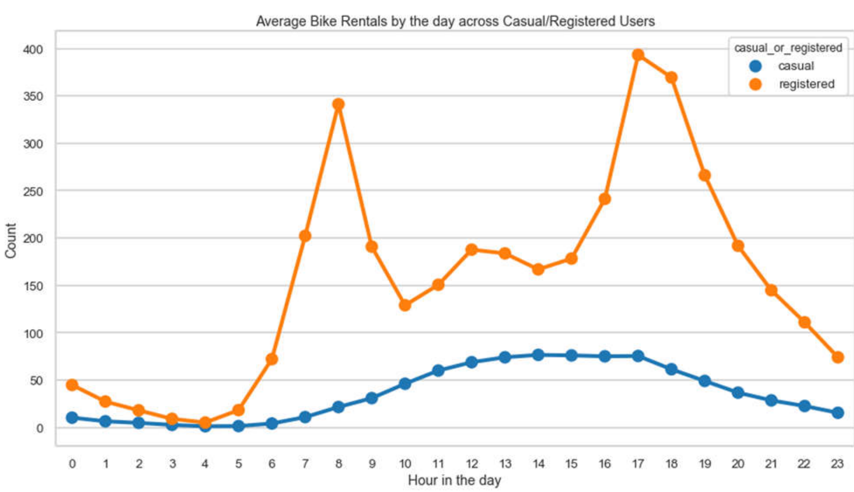


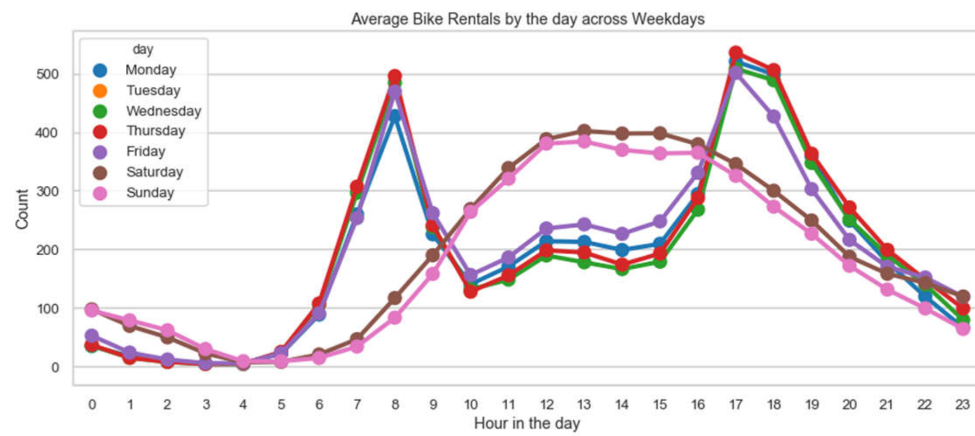


Exploratory Data Analysis



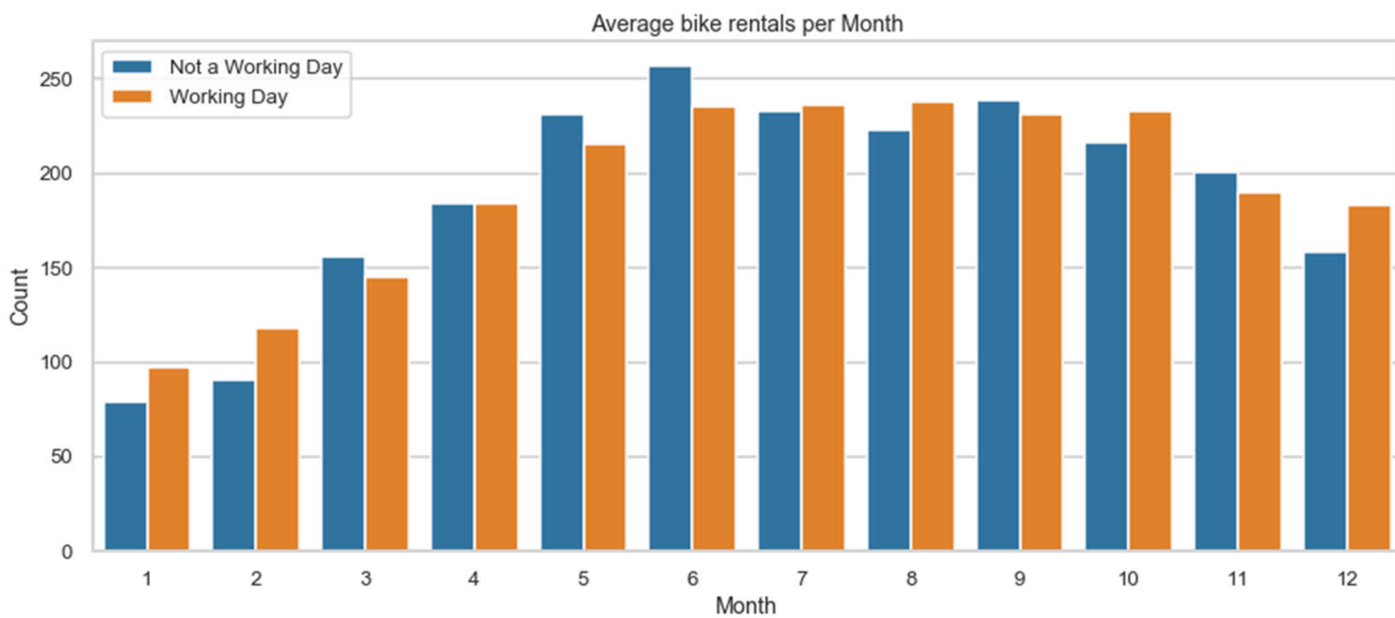
- Two biking patterns
  - Working Day Pattern: Registered Users + Working daily Commuters + 8am & 5pm peak hours
  - Non-Working Day Pattern: Casual Users + Tourists on Holidays + Steady pattern with ~12 noon peak count



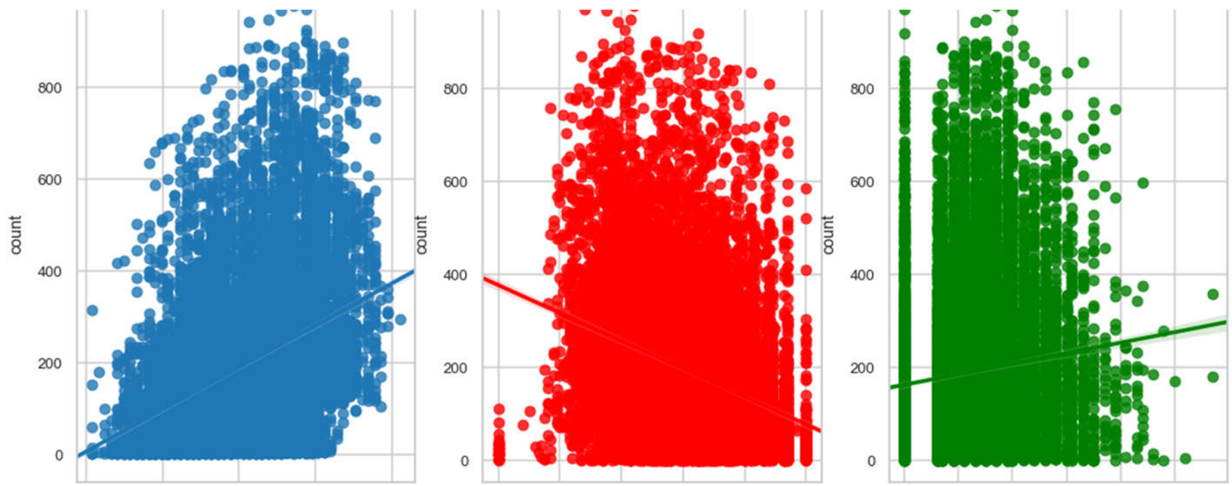


Exploratory Data Analysis – Monthly Distribution

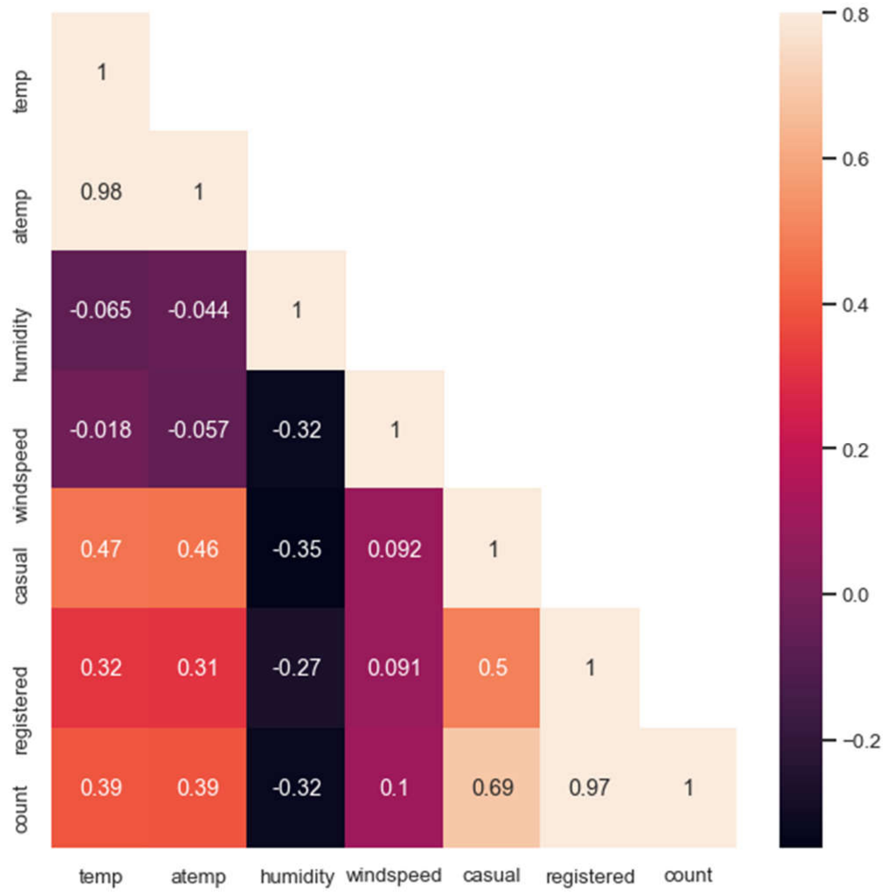
- Most rentals are in the months of June and May while least are on January and February.



Regression Plots



- We see a positive correlation between *count* and *temperature*
- We see a negative correlation between *count* and *humidity*
- *Count* has a weak dependence on *windspeed and several missing (or erroneous) data points (labeled as 0s)*



Correlation Analysis – Heatmap

- *temp* (true temperature) and *atemp* (feels like temperature) are highly correlated
- count = casual + registered ➔ count is highly correlated with casual and registered

Feature Engineering

Feature Engineering – 1

datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count	month	date	hour	day
2011-01-01 00:00:00	Spring	0	0	Clear	9.84	14.395	81	0.0	3	13	16	1	1	0	Saturday
2011-01-01 01:00:00	Spring	0	0	Clear	9.02	13.635	80	0.0	8	32	40	1	1	1	Saturday
2011-01-01 02:00:00	Spring	0	0	Clear	9.02	13.635	80	0.0	5	27	32	1	1	2	Saturday

datetime	weather	month	hour
2011-01-01 00:00:00	1	1	0
2011-01-01 01:00:00	1	1	1
2011-01-01 02:00:00	1	1	2

OneHotEncoding

datetime	weather_1	weather_2
2011-01-01 00:00:00	1	0
2011-01-01 01:00:00	1	0
2011-01-01 02:00:00	1	0

datetime	month_1	month_2	month_3	...	month_9	month_10	month_11
2011-01-01 00:00:00	1	0	0	...	0	0	0
2011-01-01 01:00:00	1	0	0	...	0	0	0
2011-01-01 02:00:00	1	0	0	...	0	0	0

datetime	hour_0	hour_1	hour_2	...	hour_20	hour_21	hour_22
2011-01-01 00:00:00	1	0	0	...	0	0	0
2011-01-01 01:00:00	0	1	0	...	0	0	0
2011-01-01 02:00:00	0	0	1	...	0	0	0

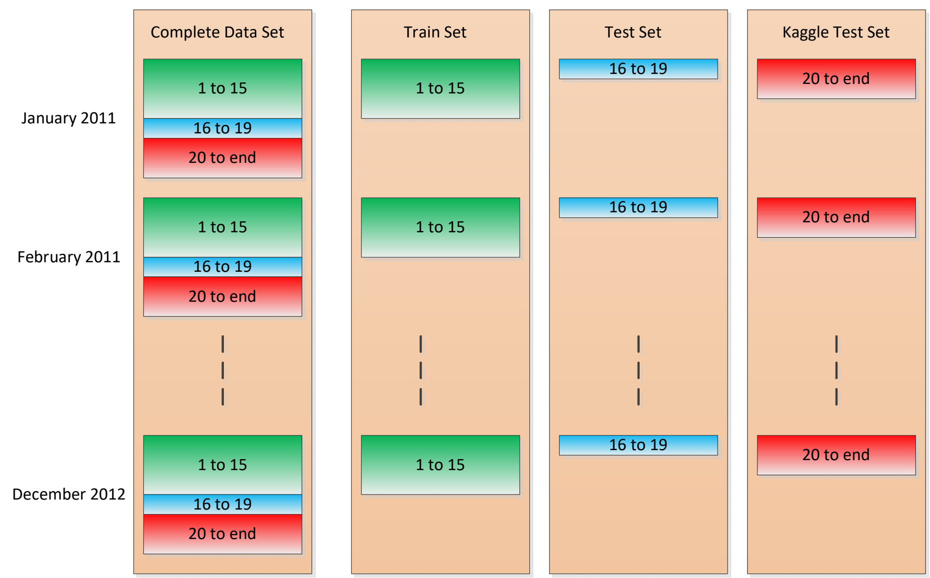
Modeling Overview

Modeling Steps

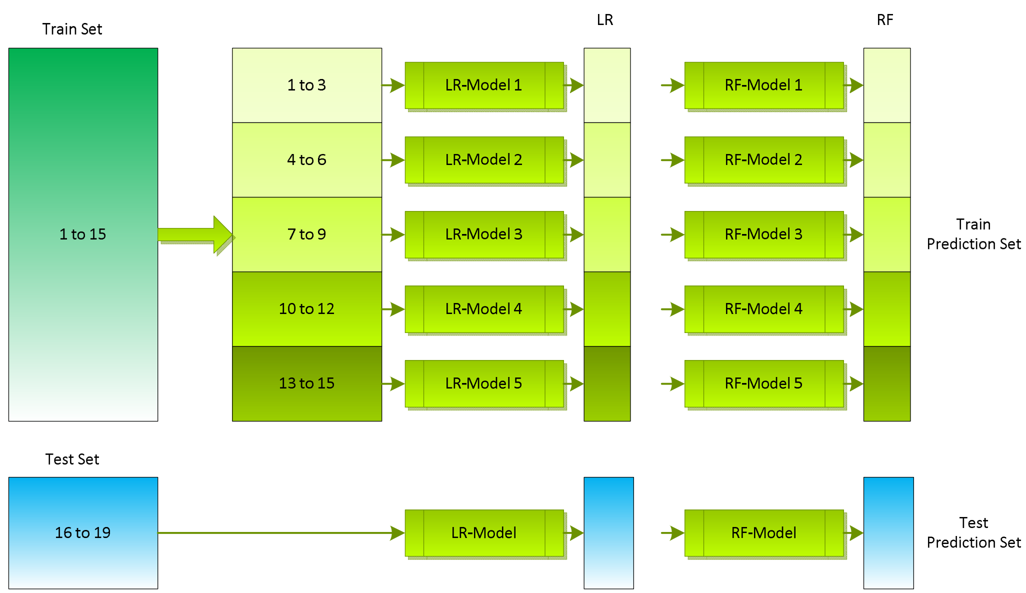
Evaluation Metric - RMSLE

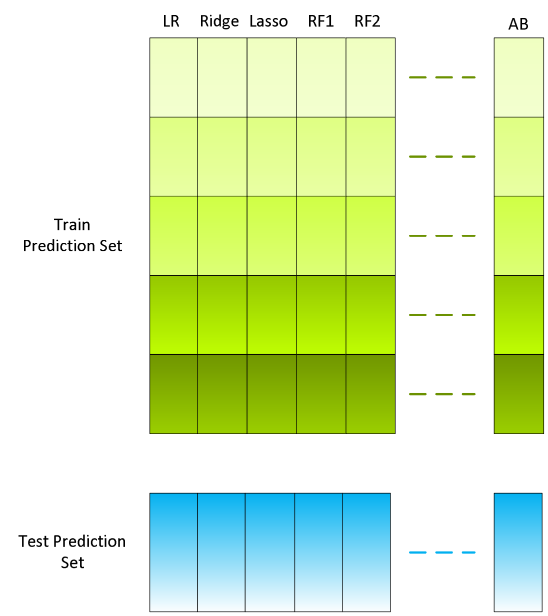
- $\sqrt{\frac{1}{n}\sum_i^n(\log\ (p_i + 1) - \log(a_i + 1))^2}$
- n is the number of hours in the test set
- p_i is the predicted count
- a_i is the actual count
- $\log(x)$ is the natural logarithm

Train-Test Split

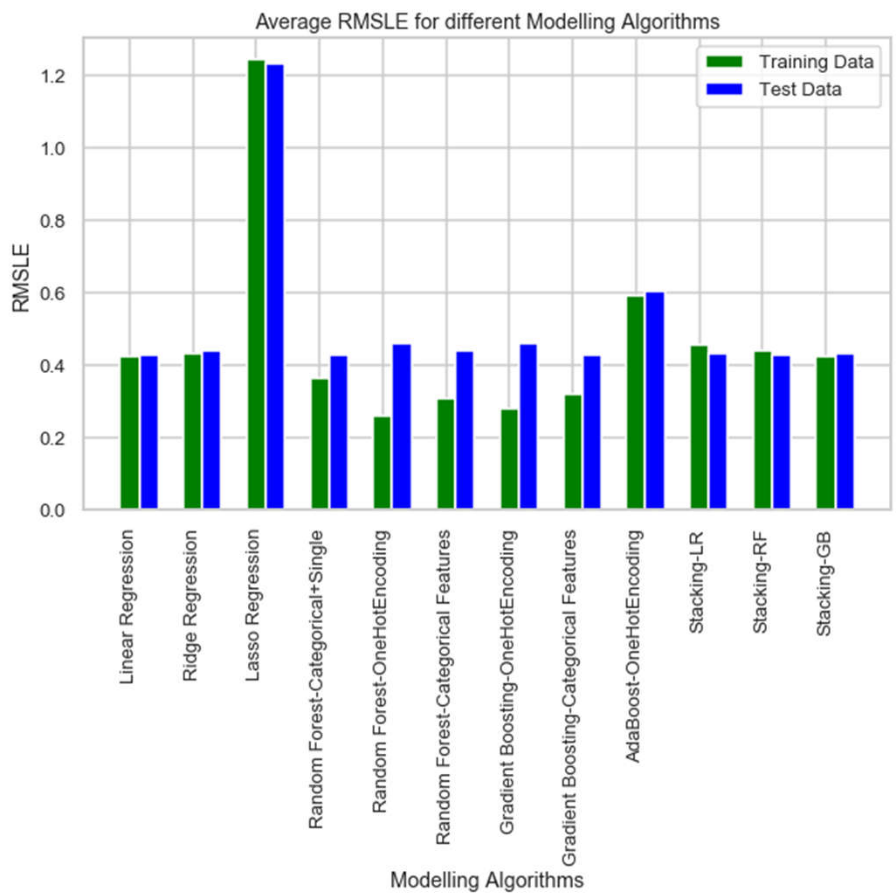


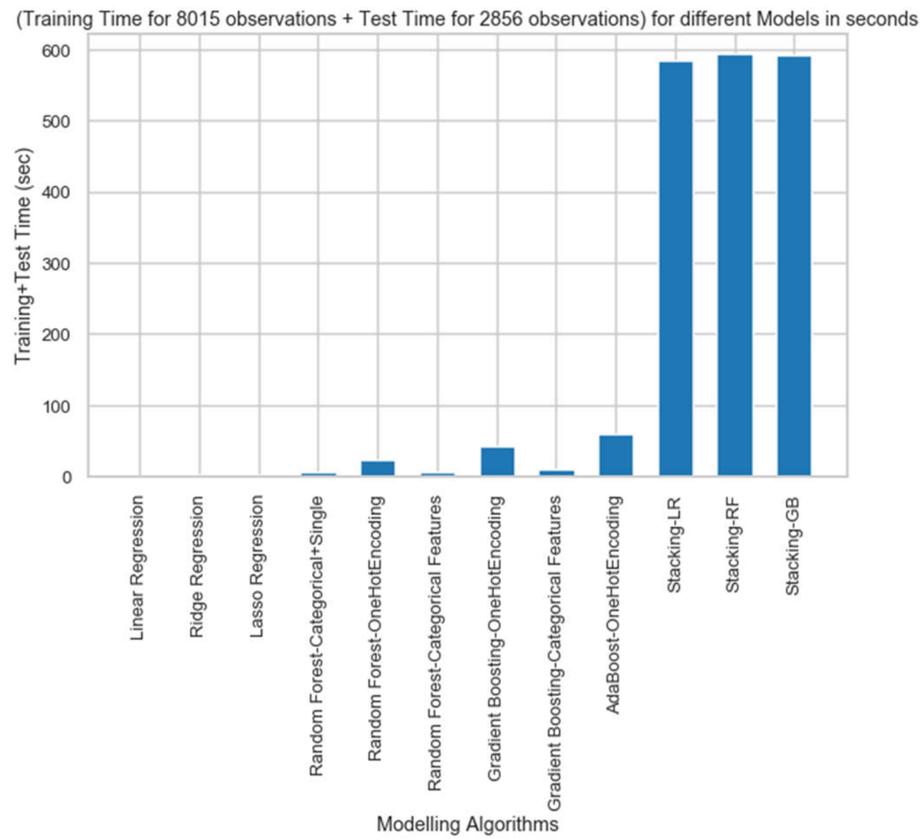
Stacking Modeling Procedure



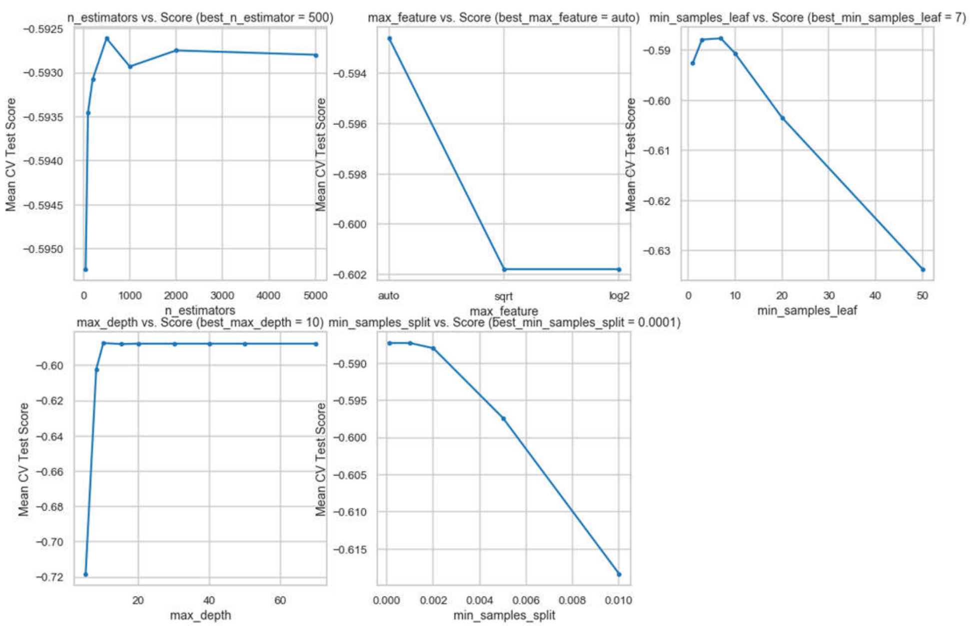


RMSLE & Modeling Time Summary

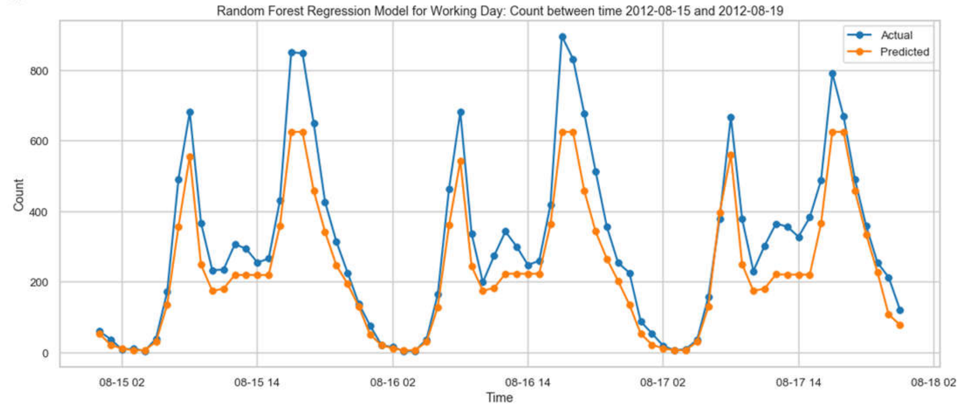


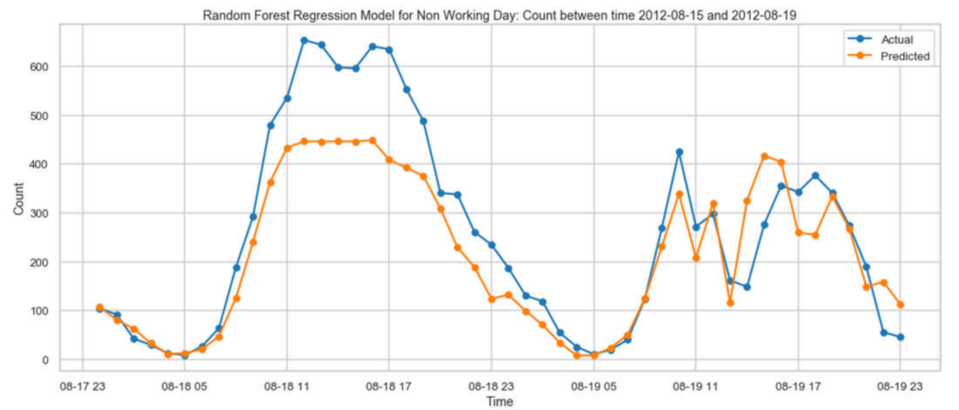


Random Forest Regression Hyperparameter Tuning

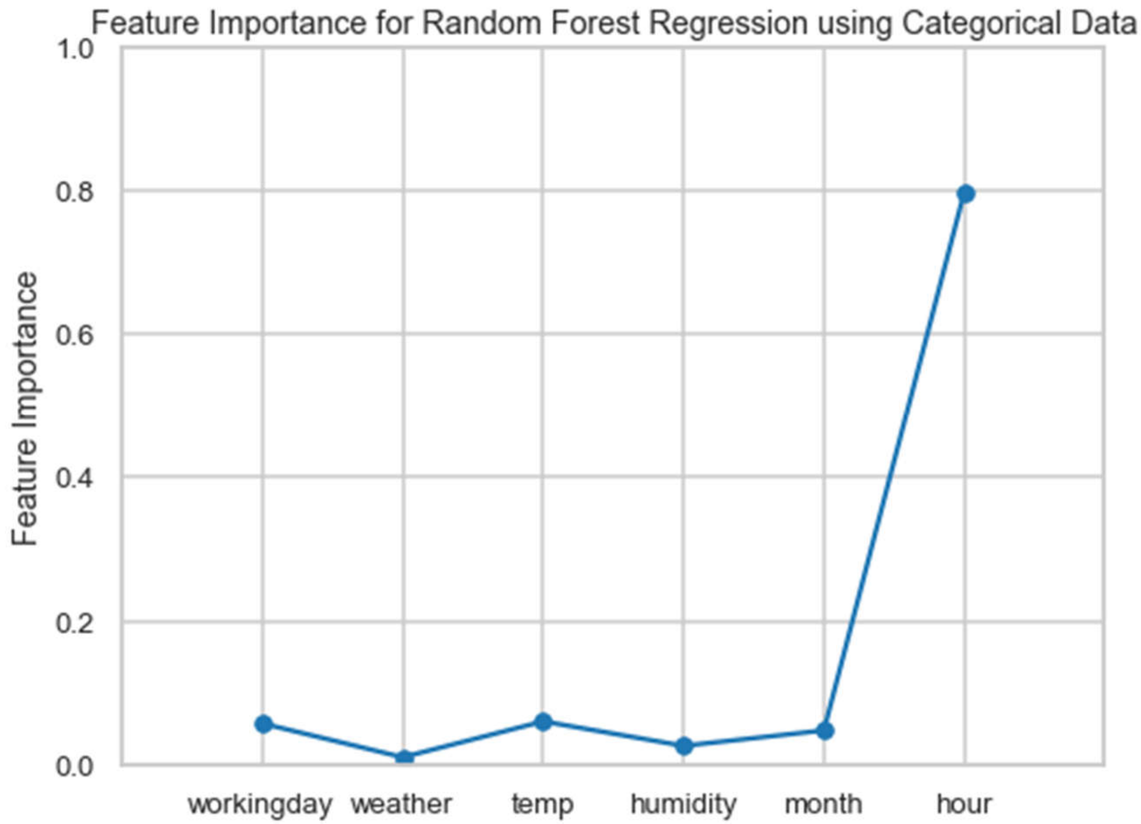


Random Forest Regression Model Performance

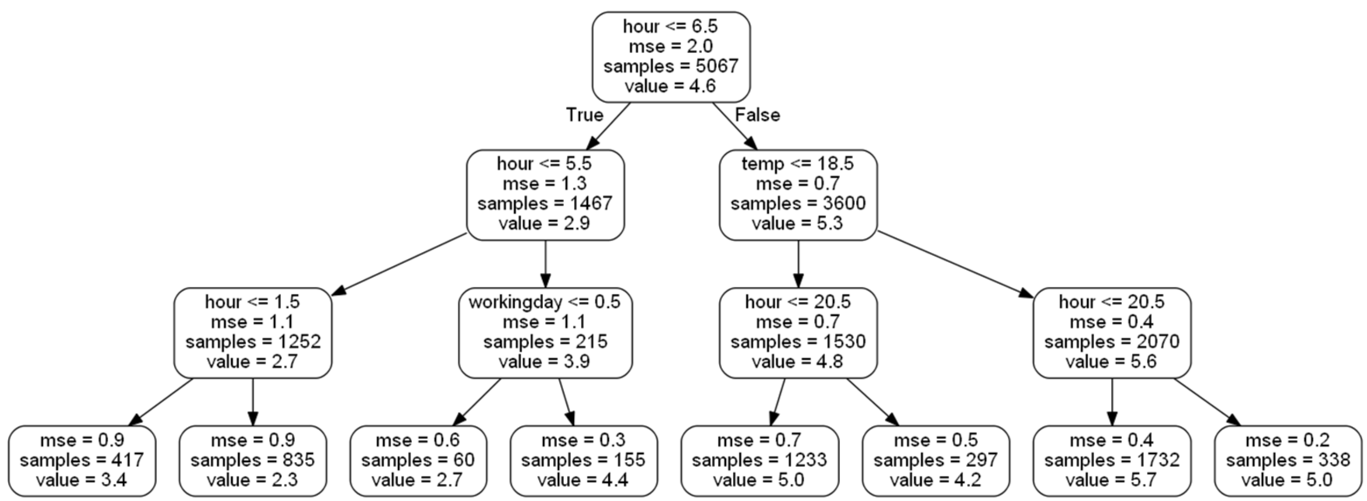




Random Forest Regression Feature Importance



Random Forest Regression
One Sample Decision Tree



Limitations and Ideas for Model Imporvement

Conclusions