Bike Sharing Prediction

Springboard

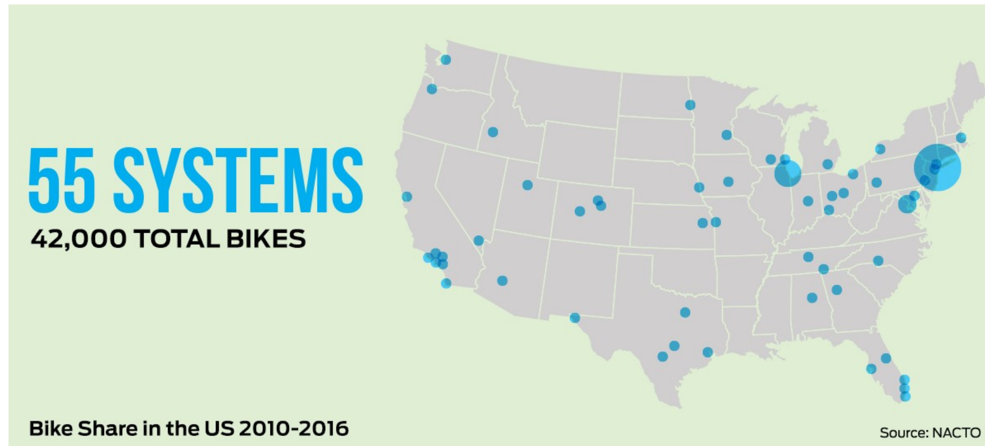
Shashank V. Maiya

Data Science Intensive Capstone Project

October 1, 2018 Cohort

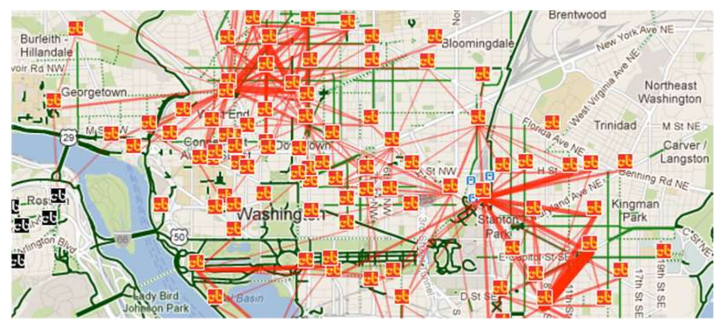
Bike Sharing Systems

- Very prevalent in major metropolitan cities – Washington D.C., Chicago, New York City, Boston, Miami, San Diego, San Francisco, etc.



- Bike Sharing Systems – Facilities which let people borrow bikes from a 'dock' or a bike rack and return it back at another 'dock' belonging to the same system

- Used for short distance commutes
- Mostly used by commuters for daily office commutes, by tourists for short distance travel



Prediction Problem

Over 35 Million trips made in the year of 2017

- Number of bikes rented out at a particular time of the day 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

DECOBIKE SAN DIEGO

NICE RIDE MINNESOTA

DIVVY

Mobile Apps

Government Bodies

- Parking Facilities
- Bike Lanes

Data Overview

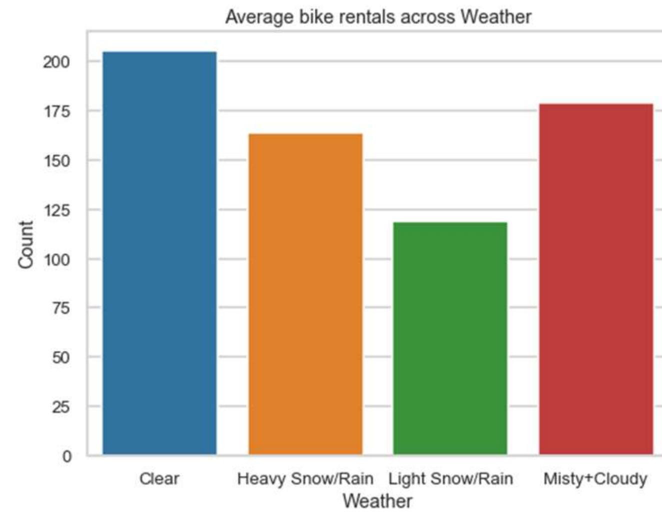
- Data set obtained from Kaggle
- Provided feature set
  - Weather conditions – Temperature, Humidity, Windspeed
  - Day – Working day or not
  - Time of the day

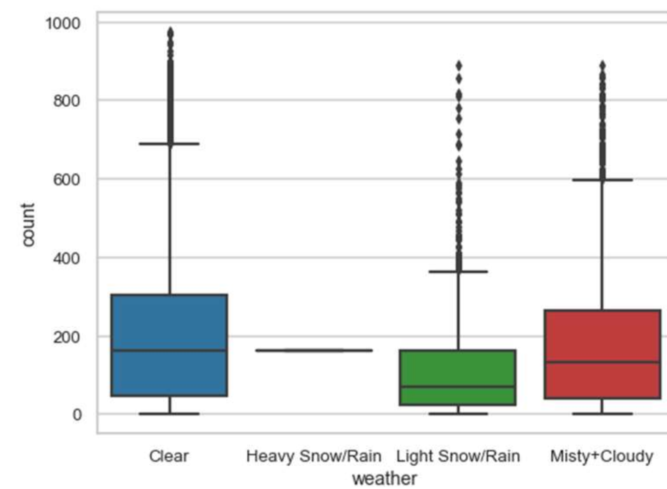
datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
2011-01-01 00:00:00	Spring	0	0	Clear	9.84	14.395	81	0.0	3	13	16
2011-01-01 01:00:00	Spring	0	0	Clear	9.02	13.635	80	0.0	8	32	40
2011-01-01 02:00:00	Spring	0	0	Clear	9.02	13.635	80	0.0	5	27	32

Exploratory Data Analysis (EDA)

EDA – Weather

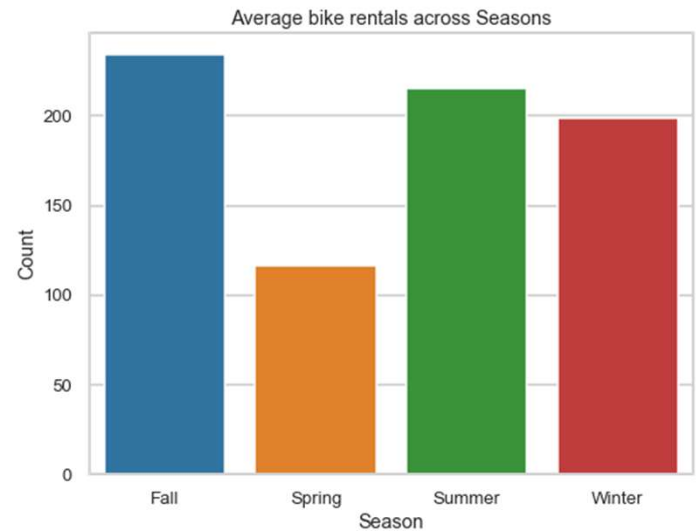
- Higher bike rental when weather is more clear and sunny
- Single instance of a Heavy Snow/Rain condition   Changed to Light Snow/Rain condition

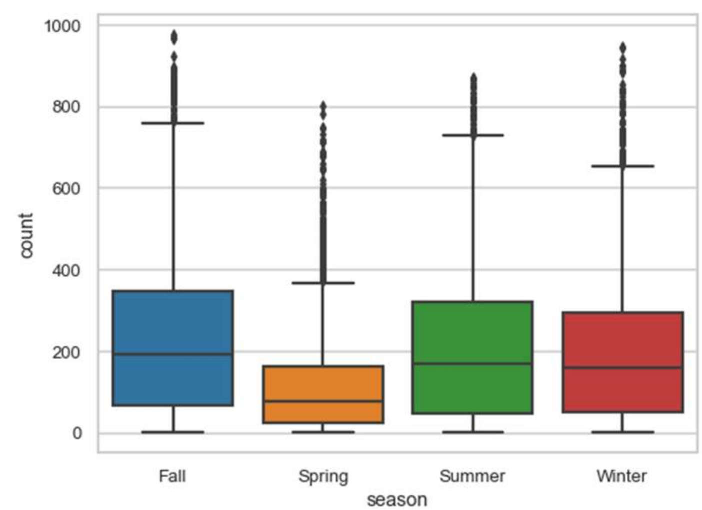




EDA – Season

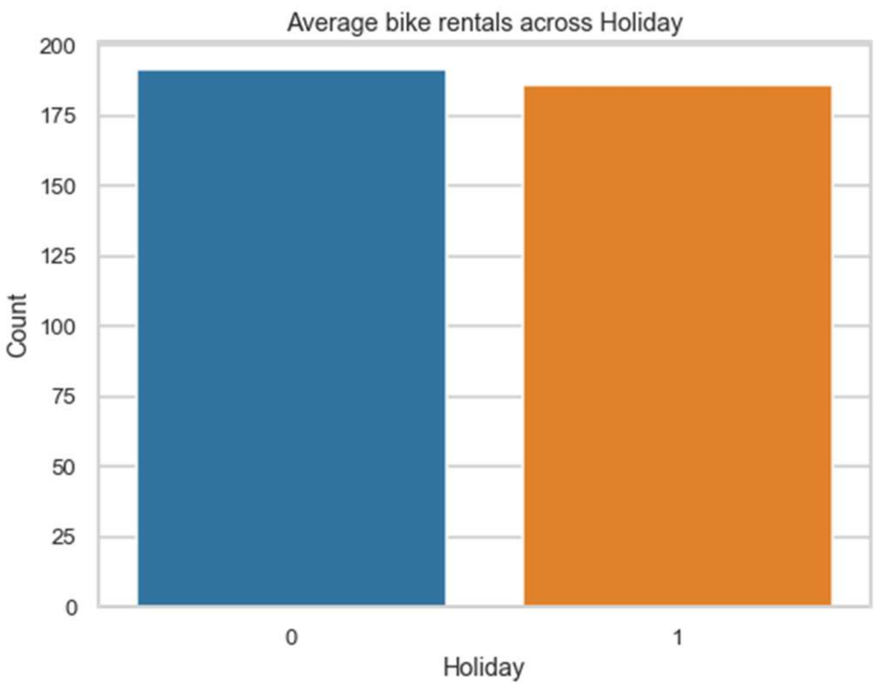
- Highest bike reservations during Summer (April to June) and Fall (July to September) and lowest in Spring (January to March)

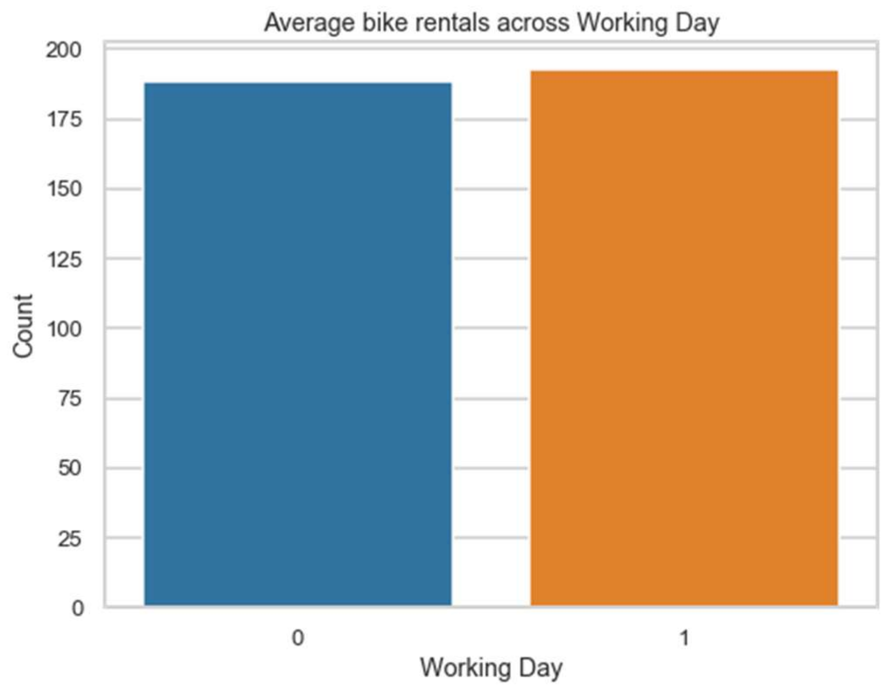




EDA– Working Day

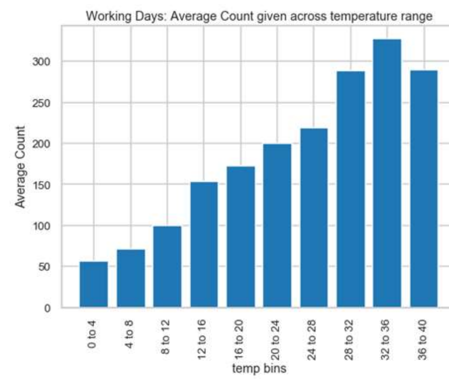
Overall average bike rental count on a Working day or Non-working day are sa

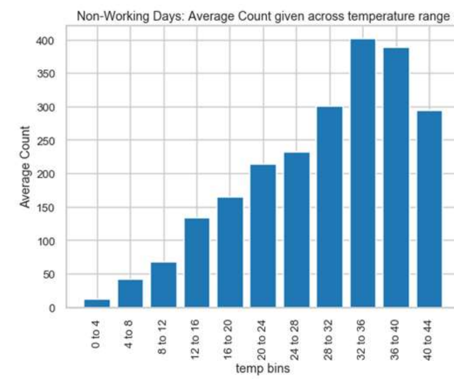


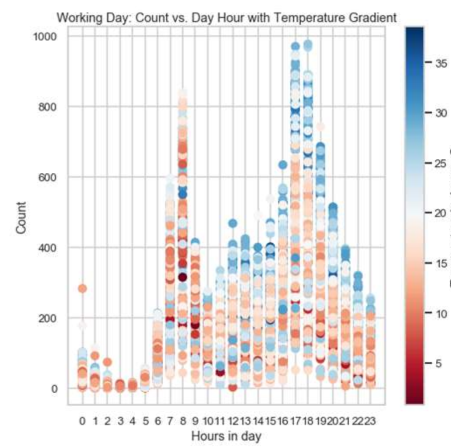


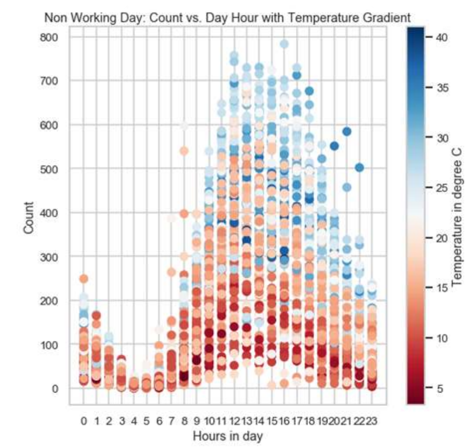
EDA – Temperature

- Steady increase in biking count with temperature
- Ideal temperature for biking is between 32 and 36 degree Celsius



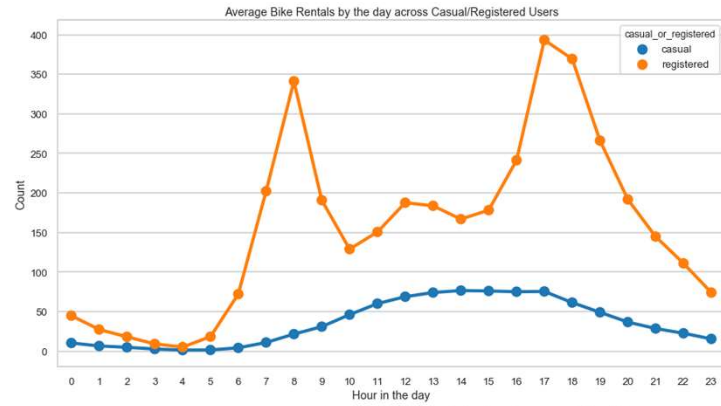


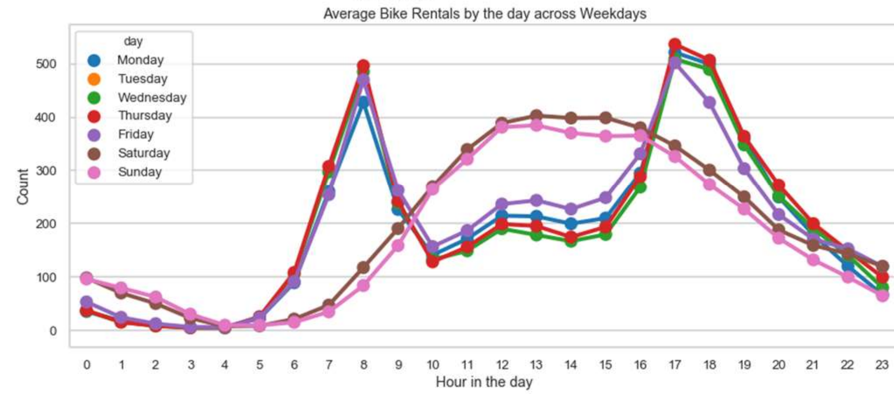


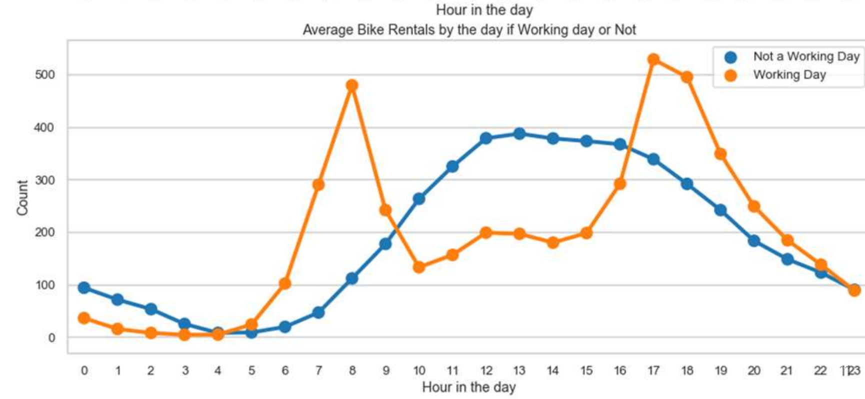


EDA – Hourly Distribution

- 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

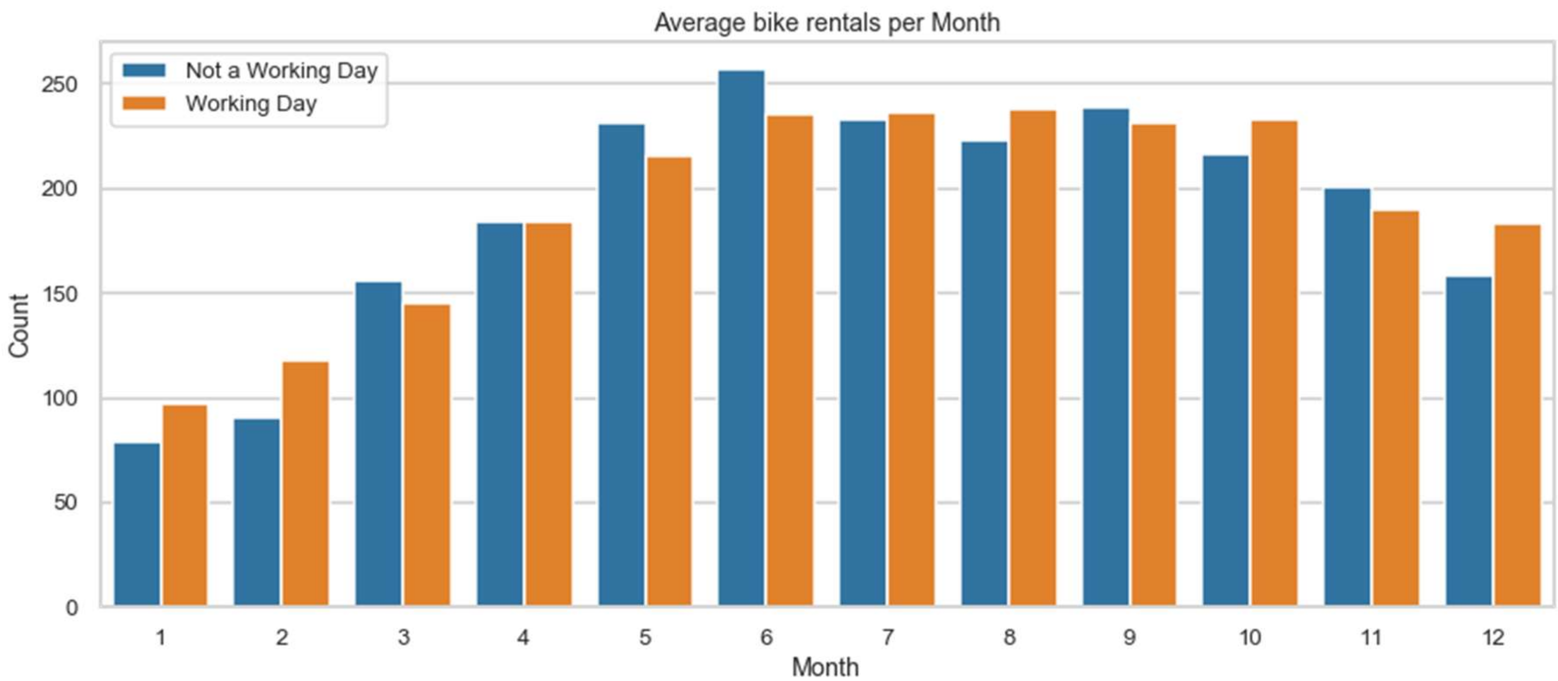






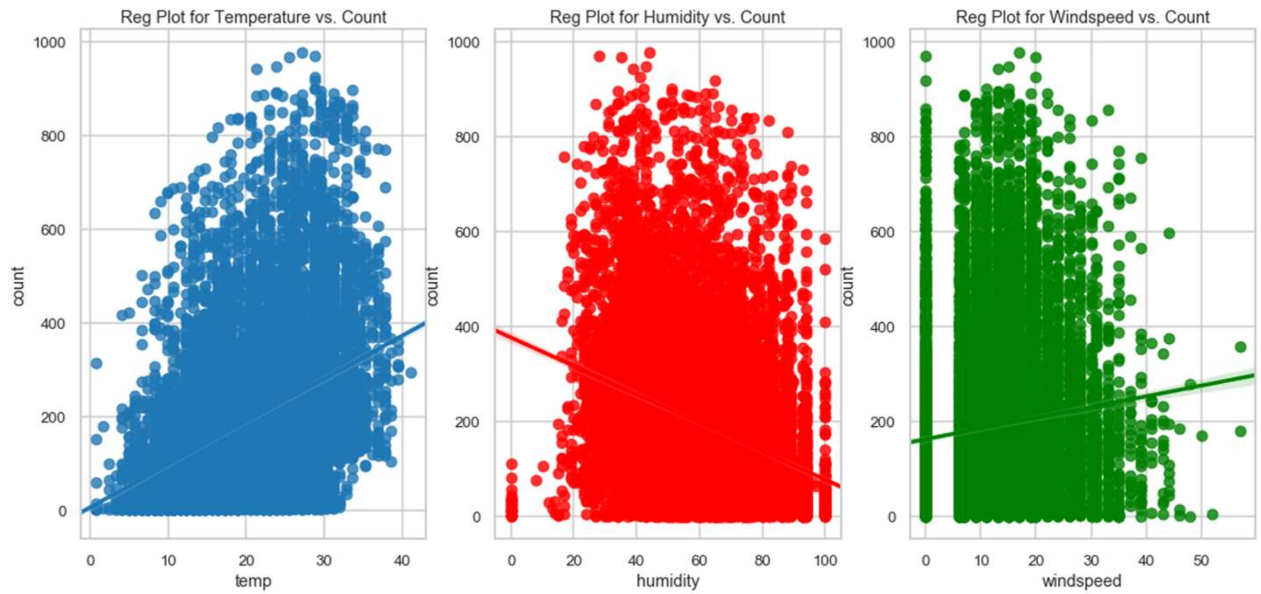
EDA – Monthly Distribution

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



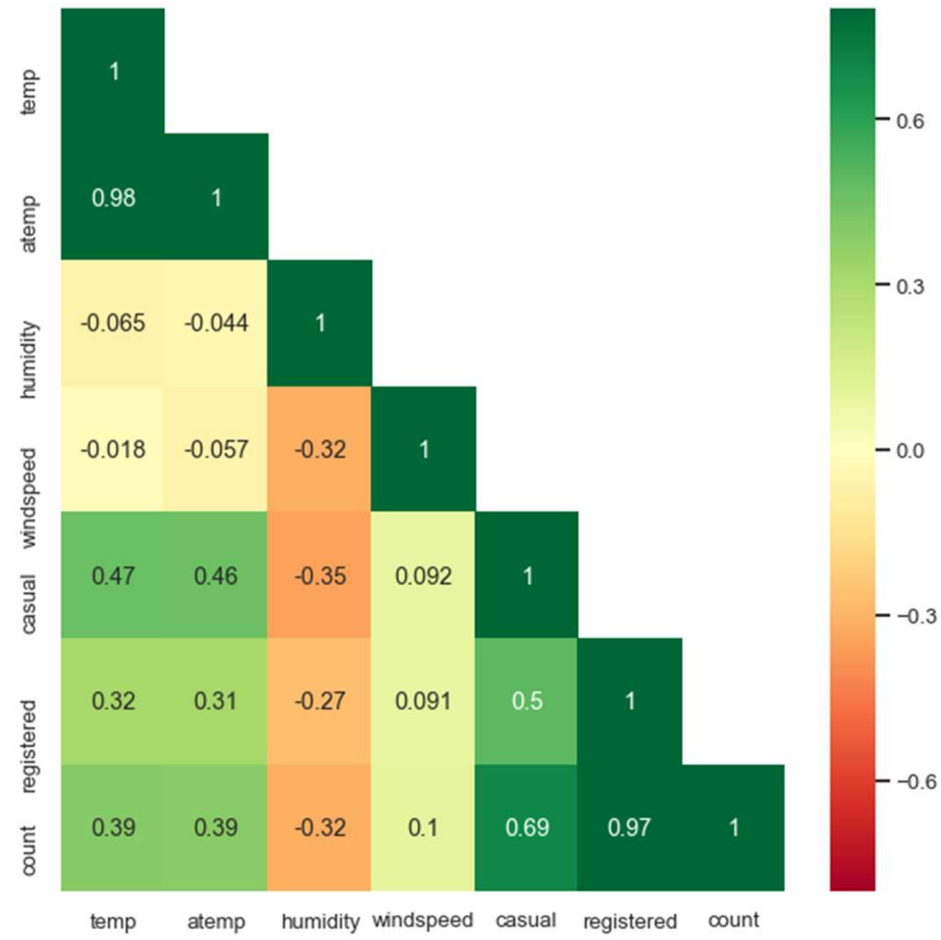
Regression Plots

- We see a strong positive correlation between count and temperature
- We see a strong 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

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

Dropped Feature

Retained Feature

Target Feature

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

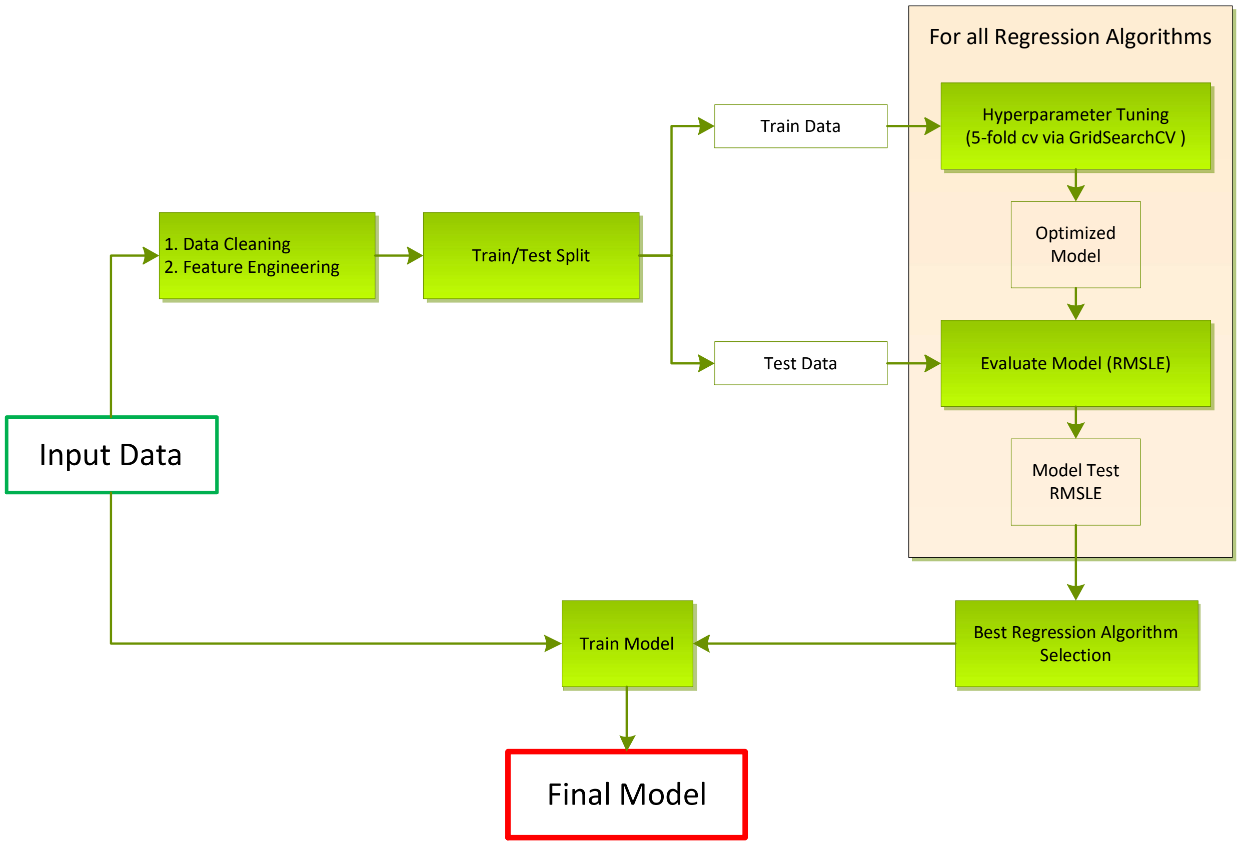
datetime	month_1	month_2	month_3	...	month_9	month_10	month_11
2011-01-01 00:00:00	1	0	0	...	0	0	0
2011-01-01 01:00:00	1	0	0	...	0	0	0
2011-01-01 02:00:00	1	0	0	...	0	0	0

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

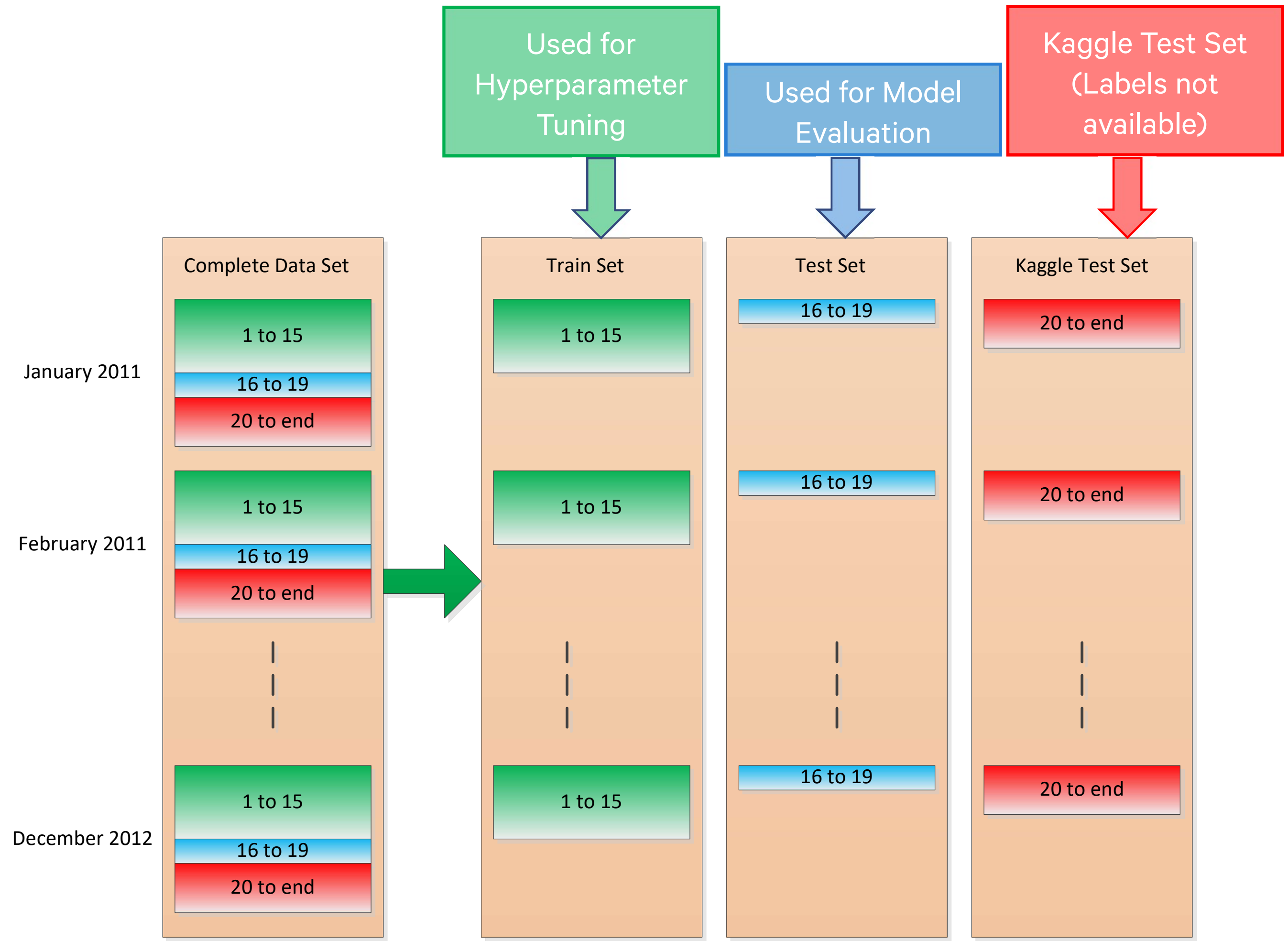
Modeling

Modeling Overview

- Type: Supervised Learning
- Regression Problem: Possible Target values [0, ∞)



Train/Test Split



Evaluation Metric - RMSLE

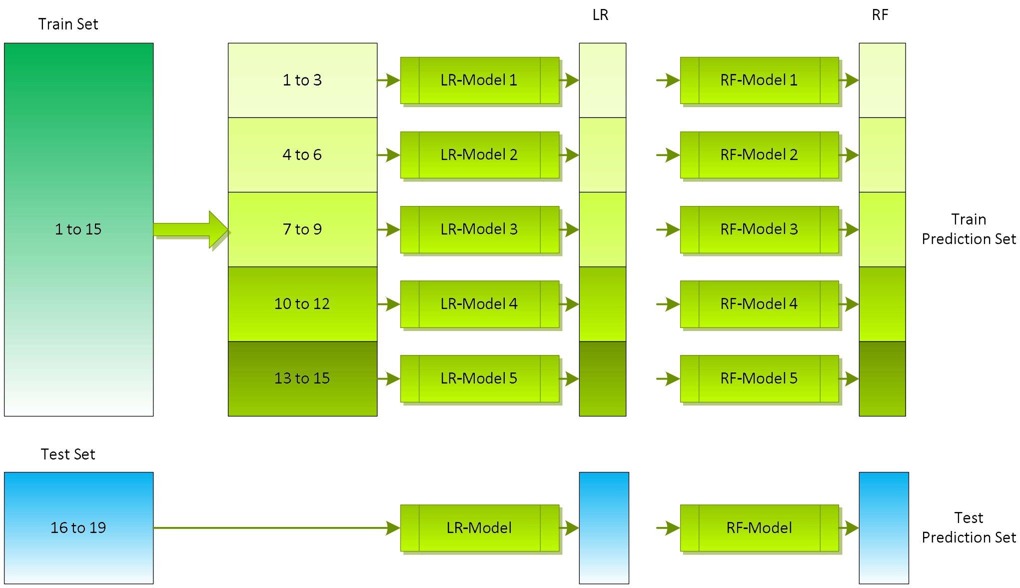
- RMSLE = Root Mean Square Log Error
- RMSLE = $\sqrt{\frac{1}{n}\sum_{i}^{n}(\log\ (p_i + 1) - \log(a_i + 1))^2}$
  - n is the number of hours in the test set
  - p_i is the predicted count
  - a_i is the actual count
  - $\log(x)$ is the natural logarithm

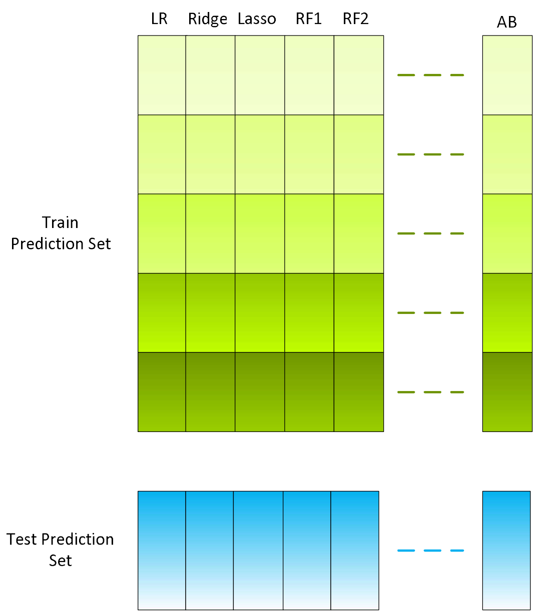
Regression Algorithms Used

- 3 Categories
  - Linear Algorithms
  - Ensemble Algorithms
  - Stacking Algorithms where predictions from Linear and Ensemble methods were used to make final predictions

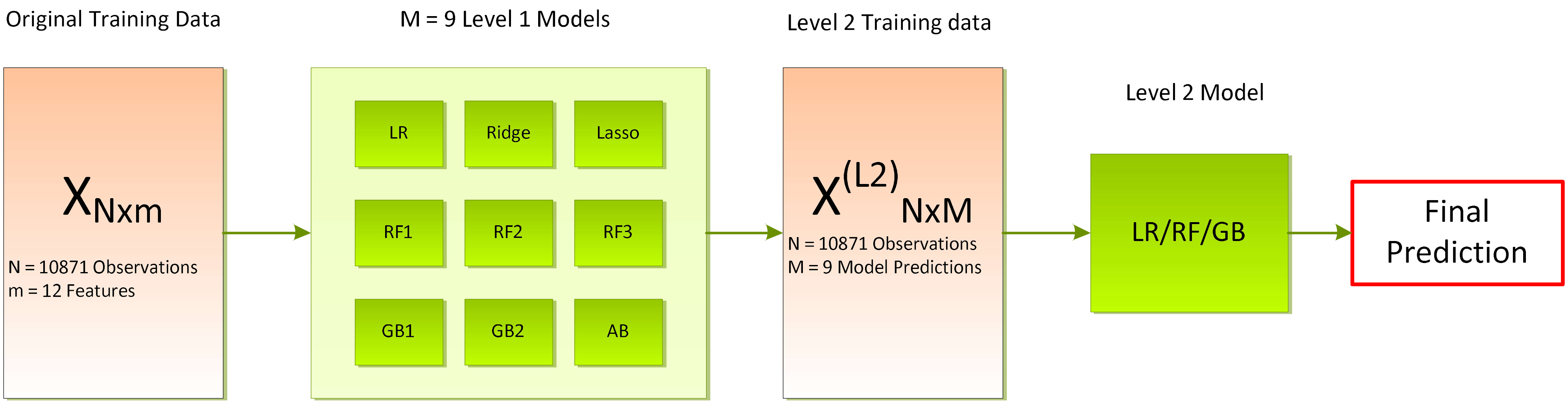
Category	Regression Algorithms
Linear	Linear Regression
	Ridge
	Lasso
Ensemble	Random Forest – 1
	Random Forest – 2
	Random Forest – 3
	Gradient Boost – 1
	Gradient Boost – 2
	Adaboost
Stacking	Linear Regression
	Random Forest
	Gradient Boost

Stacking Model Details

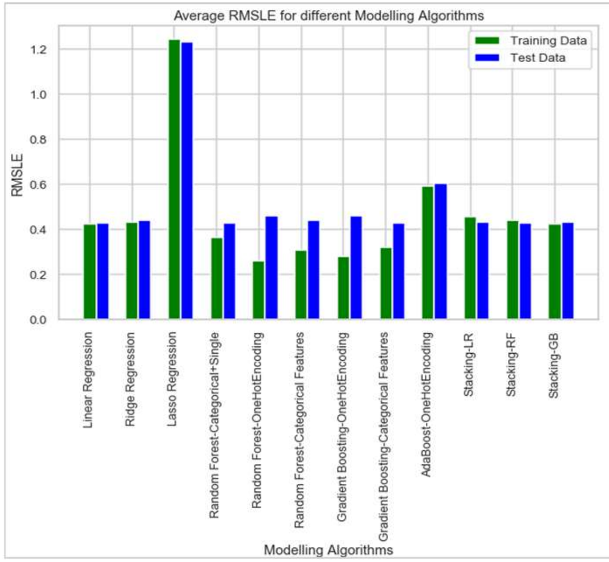


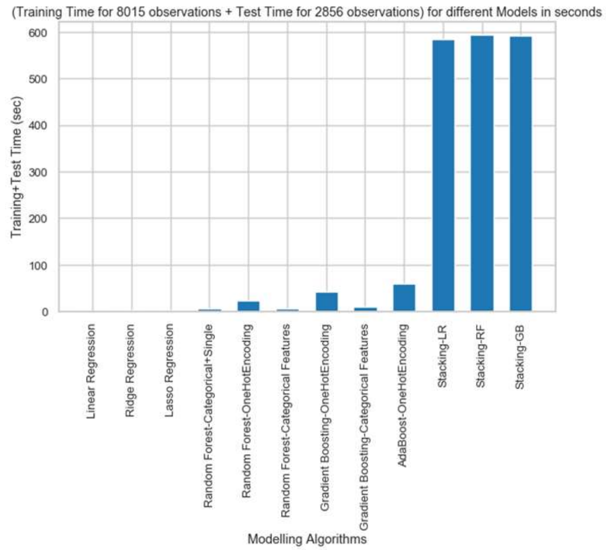


Stacking Model Summary



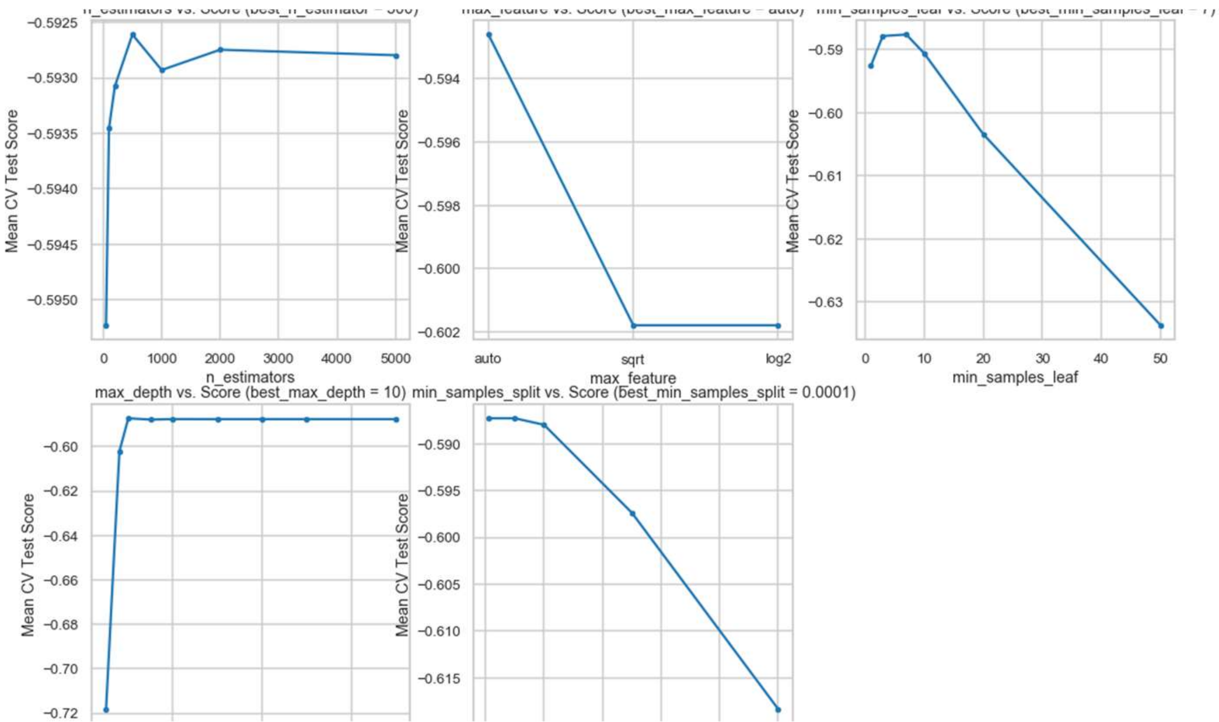
RMSLE & Modeling Time Summary





- Stacking Models didn't provide any substantial gains in prediction accuracy (RMSLE)
- Very high Train+Test times for Stacking Models
- Random Forest Modeling Algorithm used as our Final Model
    - Uses Categorical Features (No OneHotEncoding)
    - Uses Single Model for Working and Non-Working Days

Final Model – Random Forest Regressor

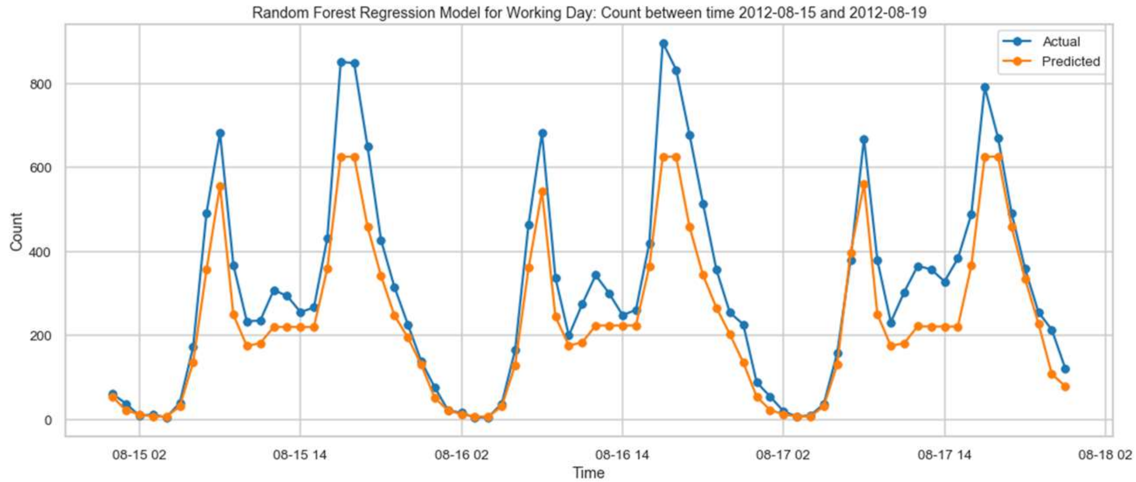


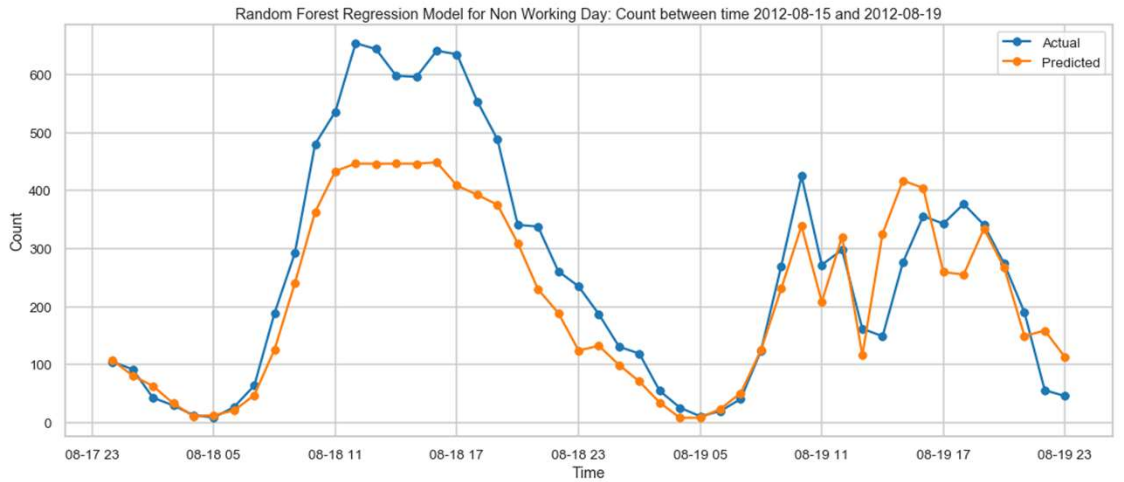
Hyperparameter Tuning

- 5 Hyperparameters Tuned
  - N_estimators = number of trees in the forest = 500
  - Max_features = max number of features considered for splitting a node = ‘auto’ = all features used
  - Min_sample_leaf = min number of samples allowed in a leaf node = 7
  - Max_depth = max number of levels in each decision tree = 10
  - Min_samples_split = min number of data points placed in a node before the node is split ~ 2

Model Performance

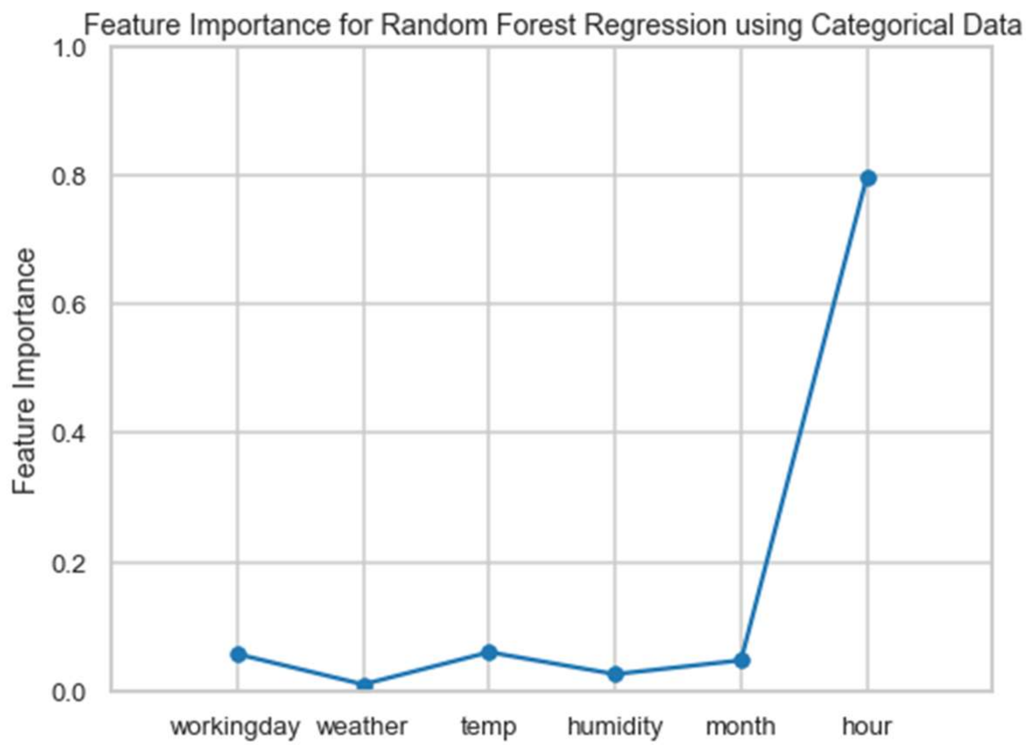
Actual vs. Predicted Bike Rental Count on Test Data





Feature Importance

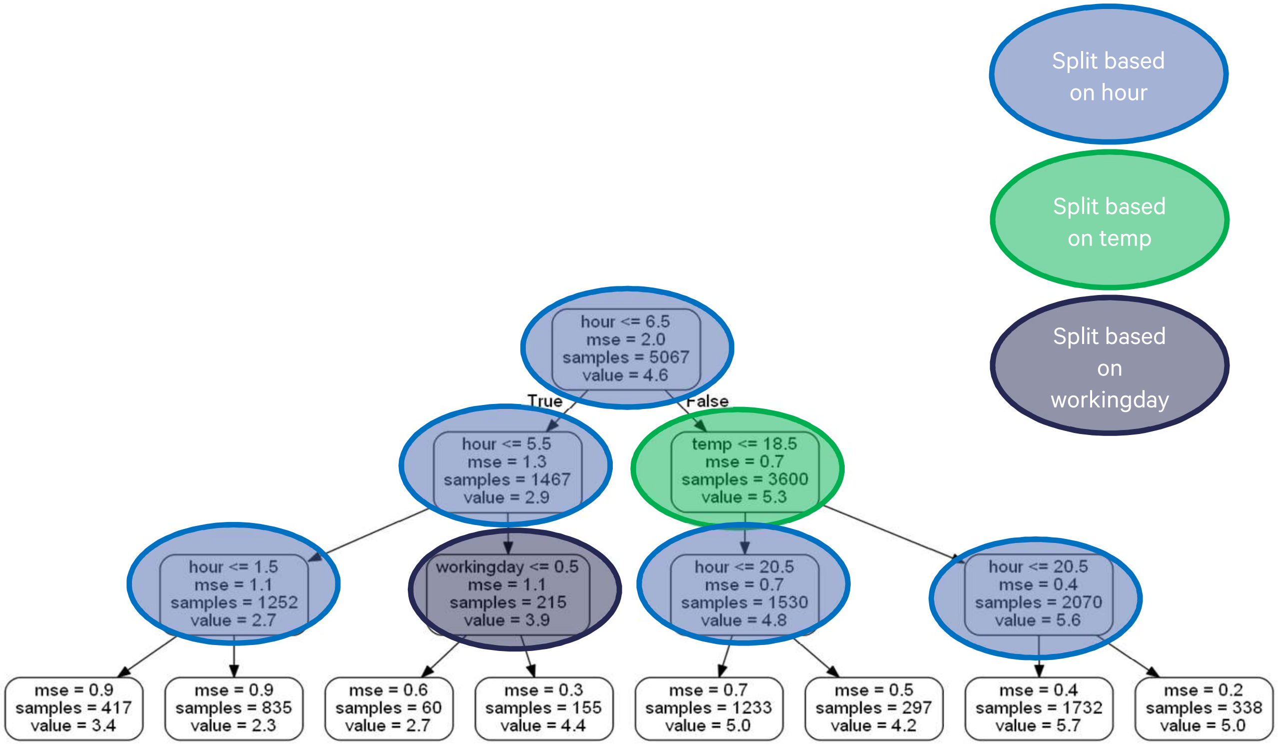
'Hour' feature has the highest importance by far



One Sample Decision Tree

Visualizing One Sample Decision Tree with max_depth = 3

First few splits are mostly based on ‘hour’ feature (indicating the relative importance of ‘hour’ feature)

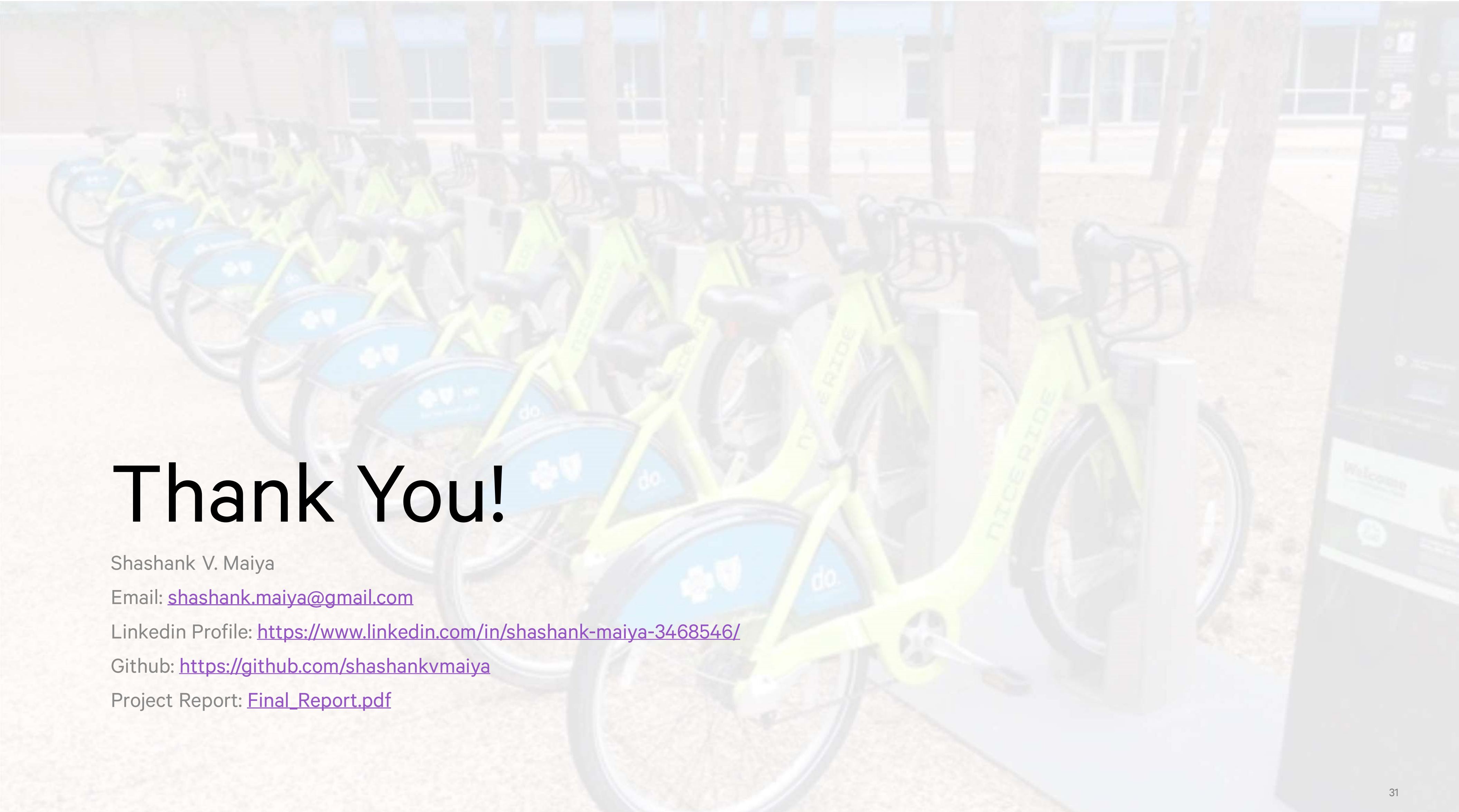


Conclusions

- Out of the 12 models tried, Random Forest yielded best prediction accuracy (lowest RMSLE) with a RMSLE score of 0.42 on Test Data
- Stacking individual models doesn't provide any improvement in RMSLE score
- 'Hour' of the day holds the most importance in prediction
- We see two rental patterns across the day
  - Working Day pattern – peak bike counts at 8am & 5pm peak hours
  - Non-working day pattern – Steady pattern with peak bike count at ~12 noon

Limitations and Ideas for Model Improvement

- Model Limitations
  - Lack of extreme weather condition (weather = 'Heavy Snow/Rain'). Hence, we cannot predict bike rental counts accurately under those constraints
- Model Improvement Ideas
  - Use of 'casual' and 'registered' users data. Estimate these two separately and add them to get the final count
  - Using Windspeed data. First Predict windspeed (for all the instances where we have 0) using other columns and then use it for prediction.



Thank You!

Shashank V. Maiya

Email: shashank.maiya@gmail.com

Linkedin Profile: https://www.linkedin.com/in/shashank-maiya-3468546/

Github: https://github.com/shashankvmaiya

Project Report: Final_Report.pdf