Capital Bikeshare: A Data Science Case Study

The Goal:

Use numerous data science techniques to understand the wealth of data at hand and develop tools and recommendations for how Capital Bikeshare can use this data.

Exploratory Data Analysis

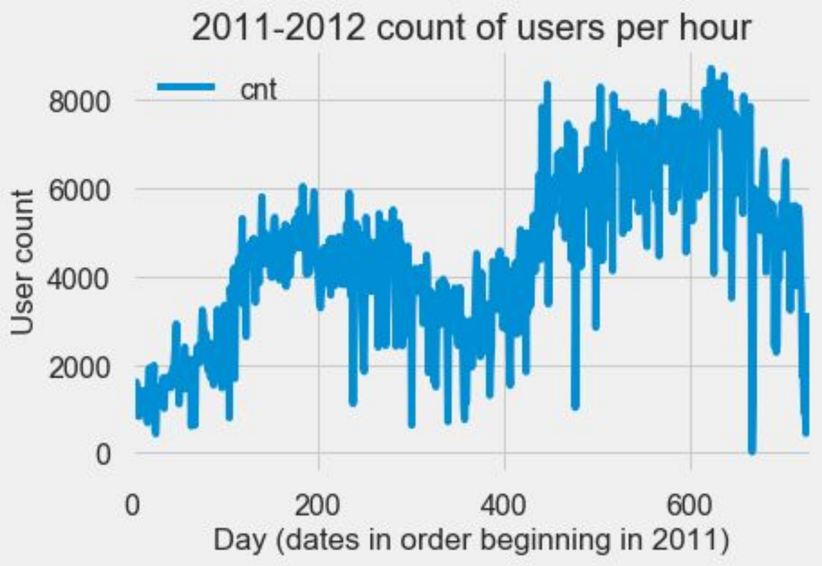
- View the data in its entirety
- Explore differences in user groups
  - Box and Whisker
  - Time Series
- Daily user counts
- Bootstrapped variance and mean plots
- Correlation Matrices

1

A first look at the data

This is a plot of 731 days (roughly two years) of rider count data from Capital Bikeshare in Washington D.C. The various ebbs and flows mean there are certainly patterns to be studied.

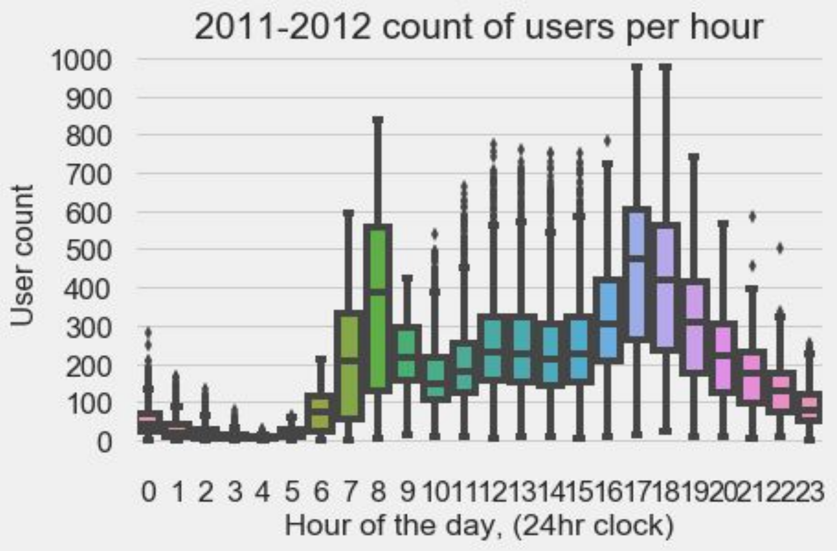
While the most basic causes of these patterns is the change of seasons but we delve deeper.



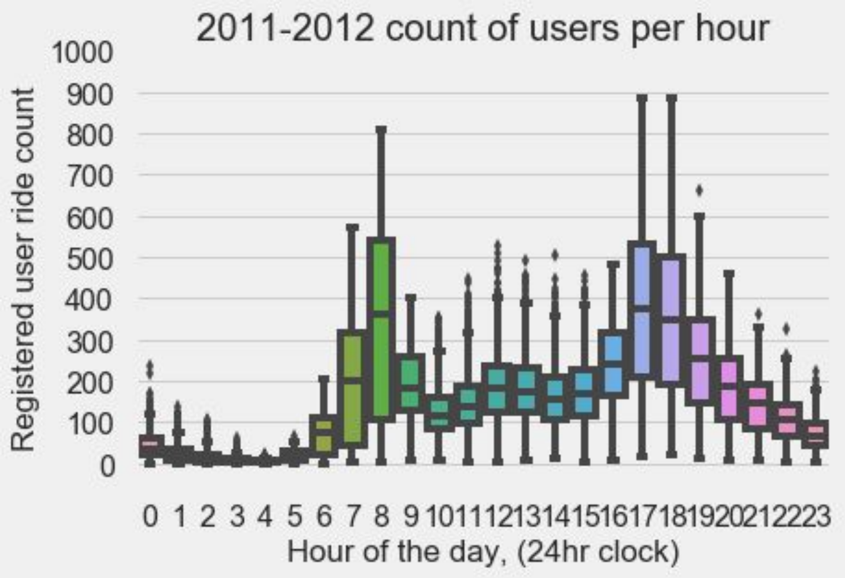
User counts by hour

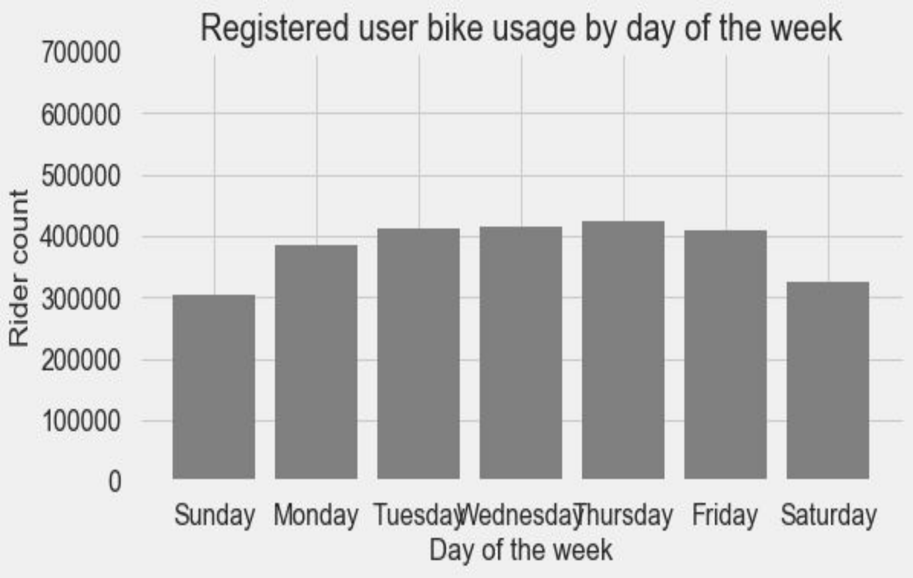
This plot shows how the entire user base waxes and wanes throughout the day peaking in the evening hours at the conclusion of the work day.

But do casual and registered riders use the system differently? Let’s investigate.

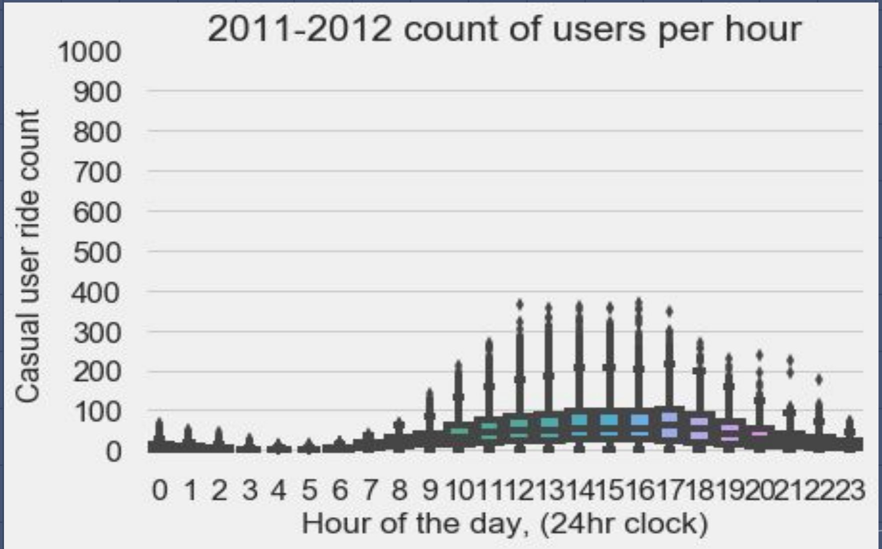


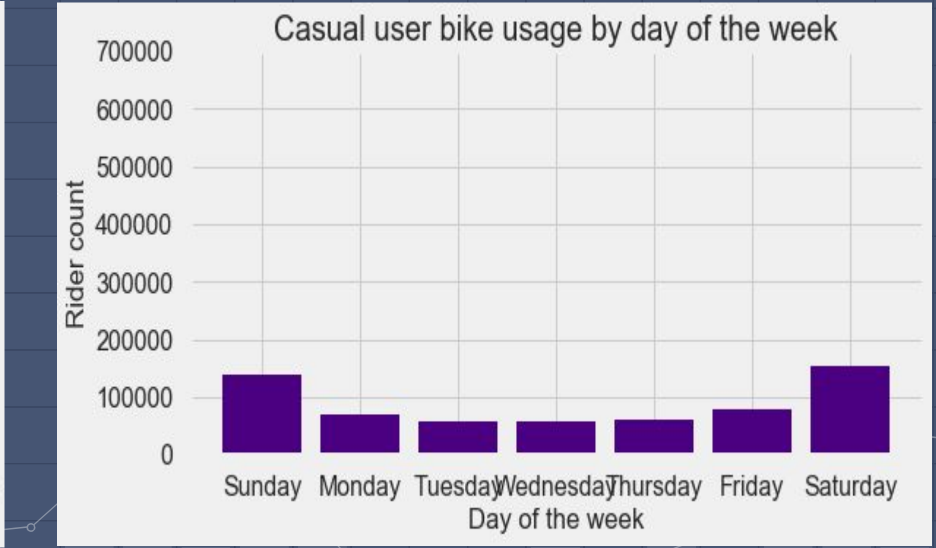
Registered riders use Capital Bikeshare to commute





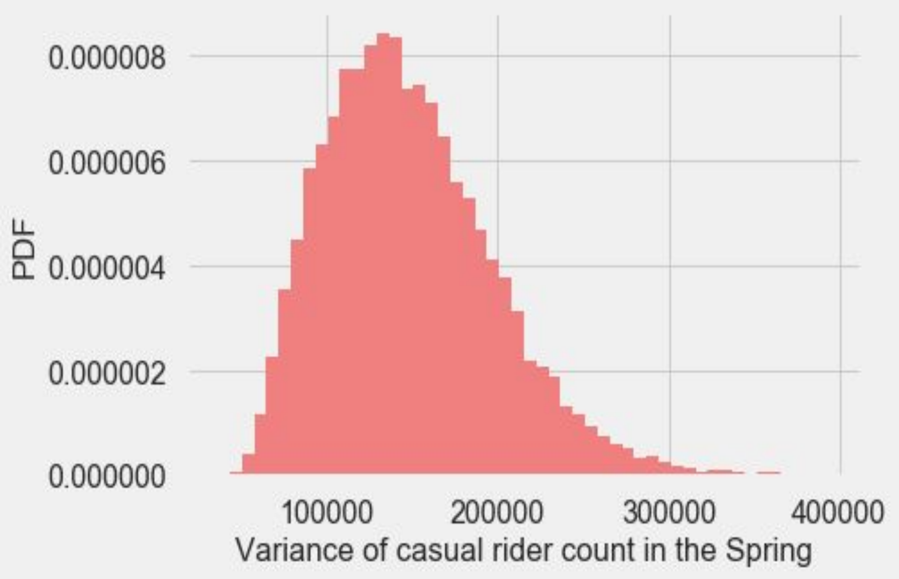
Casual riders use Capital Bikeshare for leisure, predominantly on weekends.

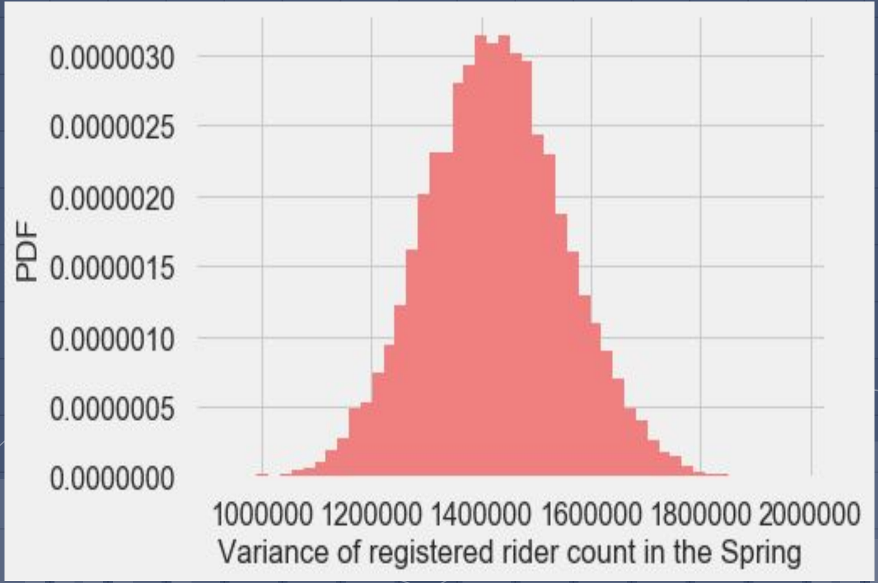




Casual Riders flock to the service in the spring

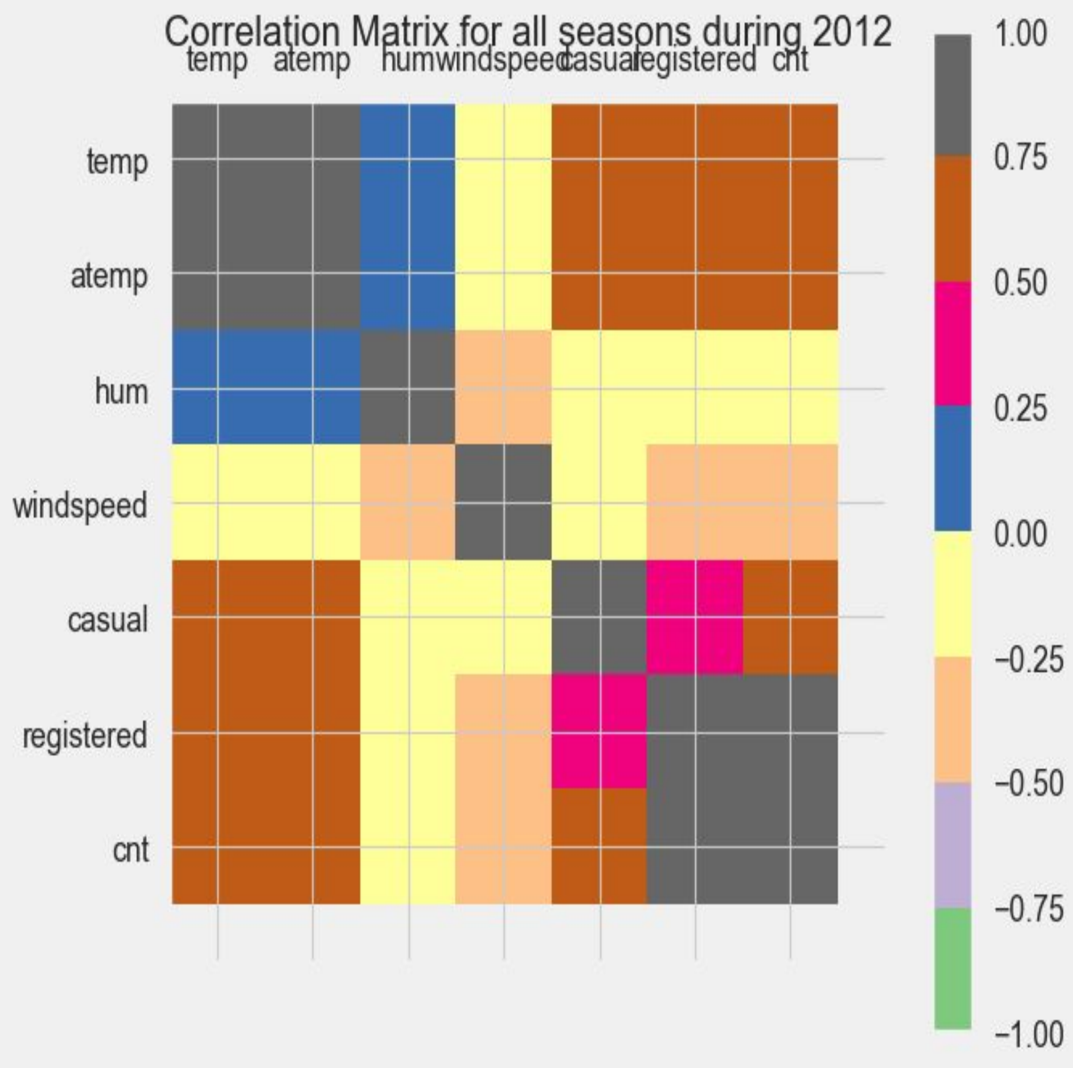
When compared to registered riders we observe an intense positive skew.





Correlation Matrix

An interesting visual that showcases the strong correlation between temperatures and rider count.



Machine Learning

- Problem Classification
  - Discussion of metrics for optimization
- Ridge Regression
- Lasso Regression
- Random Forest Regression
  - Randomized Search CV
- Prediction Experiment

2

Problem Classification

To address our problem in a Machine Learning sense we must understand whether or not it is a classification or regression problem.

Since we are attempting to predict the amount of users on any day given a set of parameters we are facing a **regression** problem. We will tackle this using: Ridge, Lasso, and Random Forest Regression and compare the resulting models.

Metrics for Optimization

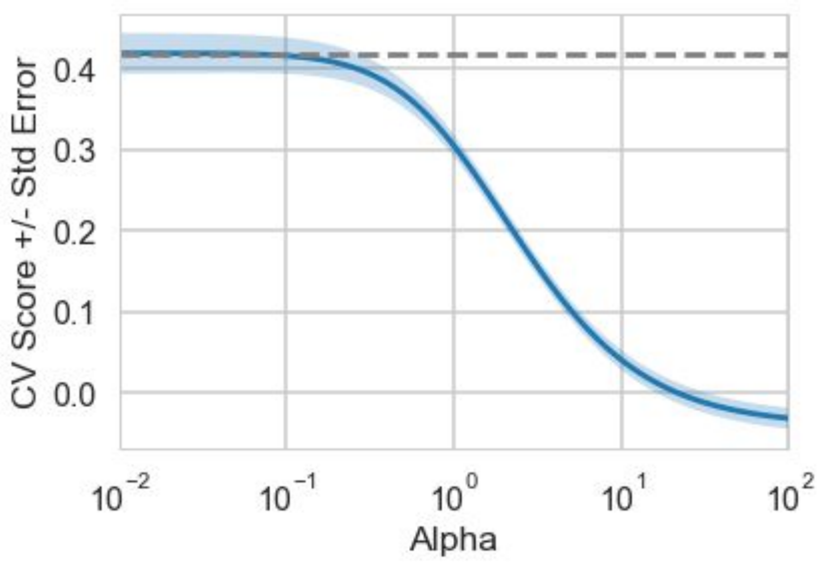
- MAPE: Since we our model only outputs positive results, minimizing MAPE will cause us to bias under-prediction.
- R^2: This is a strong metric as it shows how closely the model fits the data. If the r^2 value is too high we could risk overfitting. Thankfully the graphs allow us to view overfitting if it is occuring. (Good)
- MAE: A simple metric that gives us an idea of how much error our model has. (Good)
- RMSE: A slightly more complicated mean error metric that places a high weight on large errors. Minimizing this metric would bias a model that shys away from the occasional large error.

We will be looking at how the random forests improve in regards to their R^2, MAE, and RMSE metrics after Randomized Search Cross Validation. You'll find that there are some incredible improvements!

Ridge Regression

This plot shows the plot for determining optimal values of alpha. Alpha values at or below 0.1 should yield the best model available.

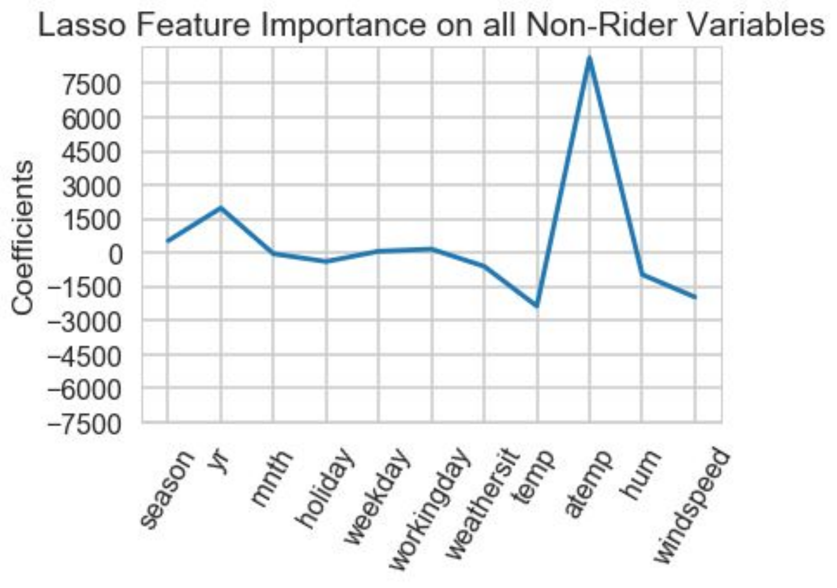
Unfortunately the model score poorly with an R^2 term below 0.5, and RMSE and MAE values far above the searched random forest values.



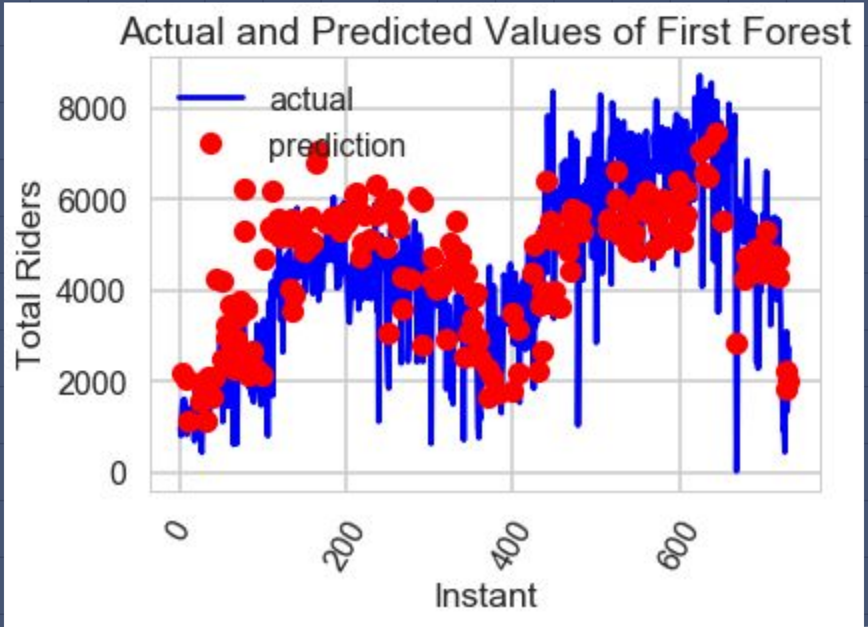
Lasso Regression

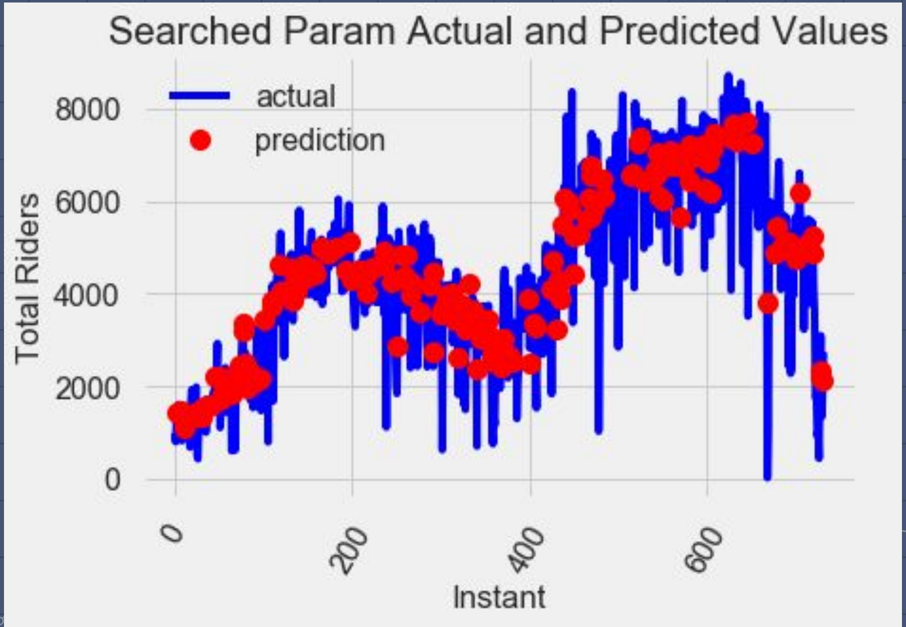
This plot shows the importances for of features in the highest scoring (R^2 0.81) Lasso model, but it sports a concerning feature. Atemp is a value constructed from temperature and humidity and dominates the model.

Removing atemp caused the model to plummet in accuracy, but this was not an issue in random forest regression, thus the later option was pursued.



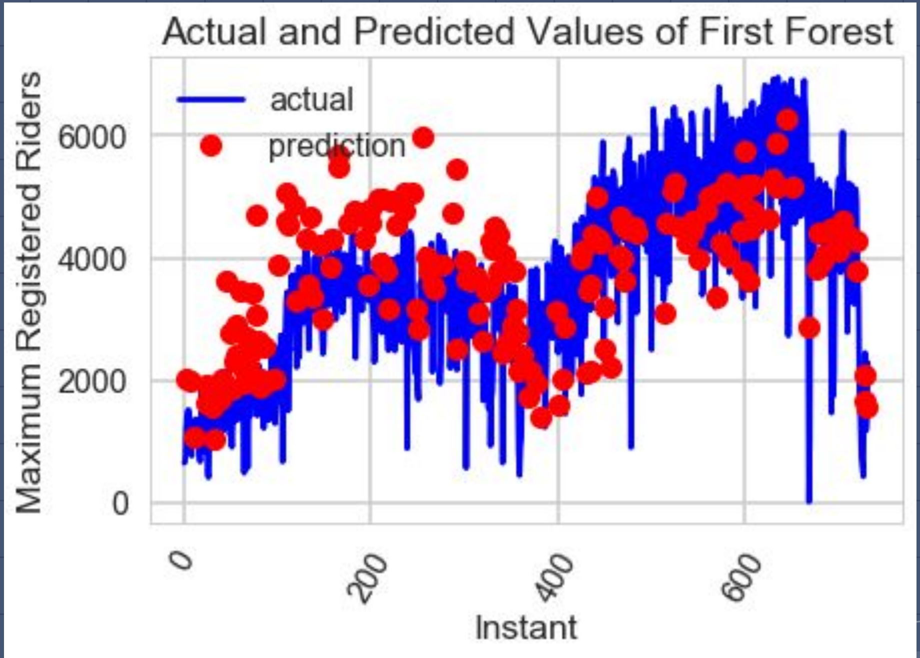
Total Riders Random Forest Regression ~ 0.30 R^2 improvement

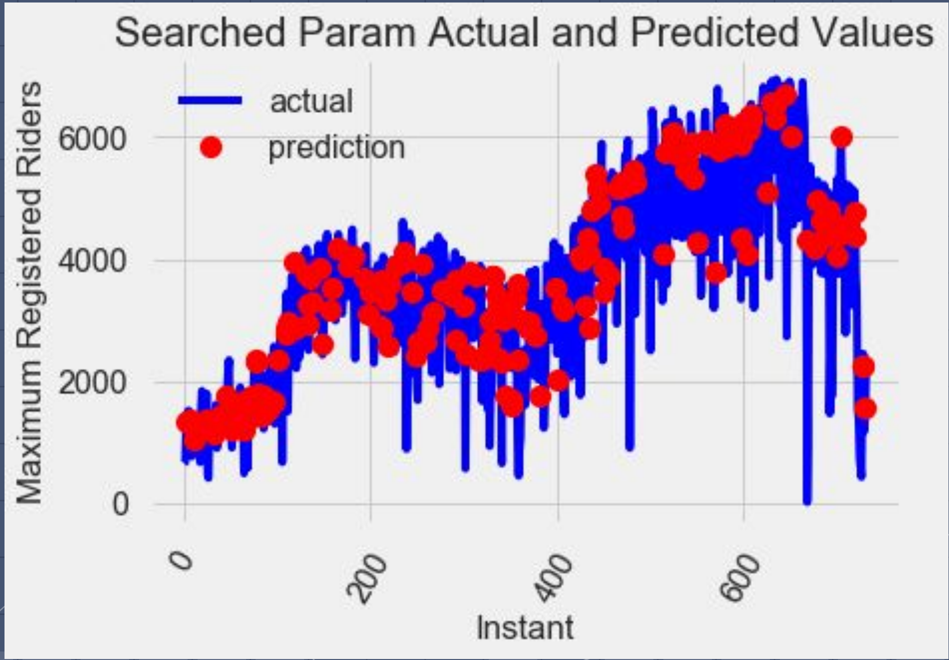




We see an incredible improvement after Randomized Search CV

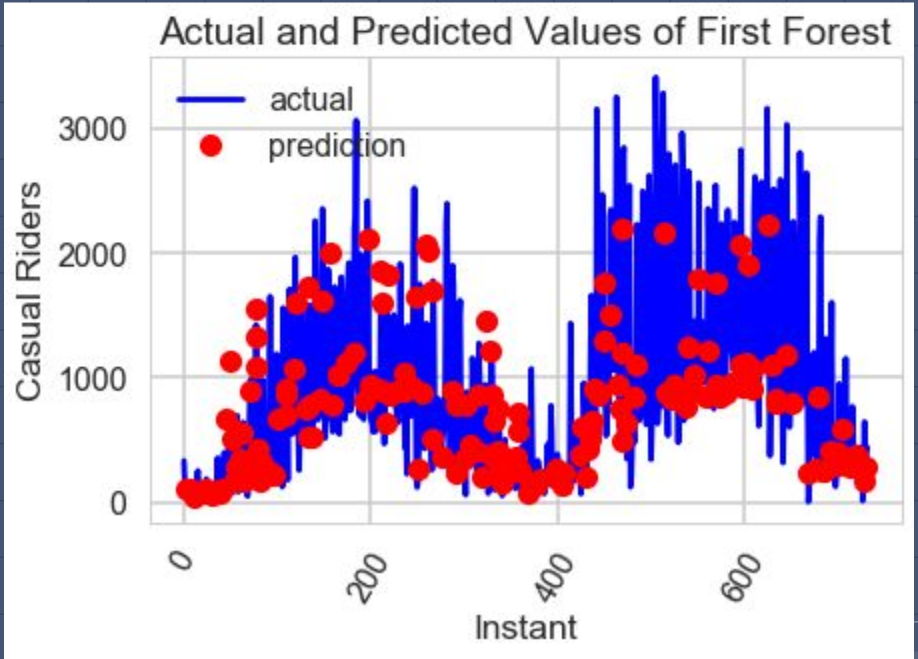
Registered Riders Random Forest Regression ~ 0.28 R^2 improvement

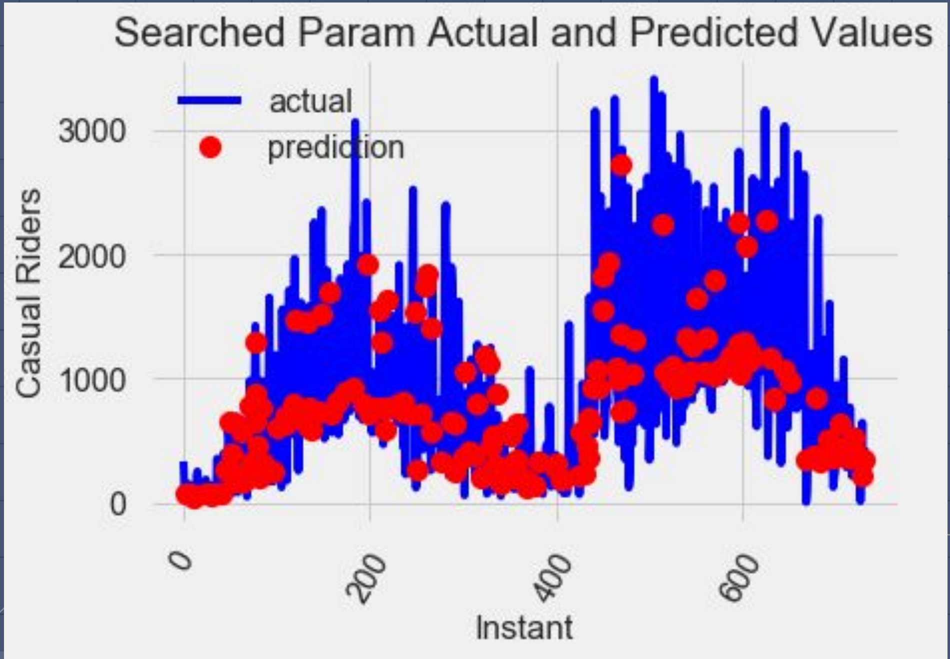

Actual and Predicted Values of First Forest


Searched Param Actual and Predicted Values

We see an incredible improvement after Randomized Search CV

Casual Riders Random Forest Regression ~ 0.06 R^2 improvement

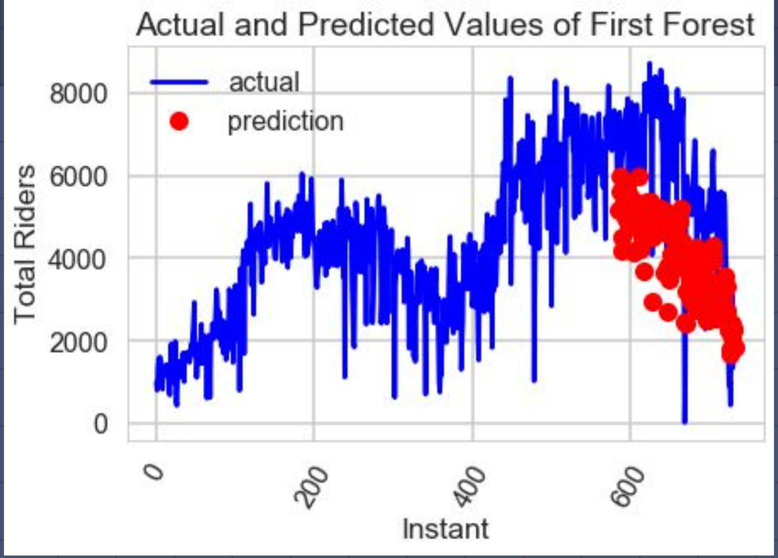


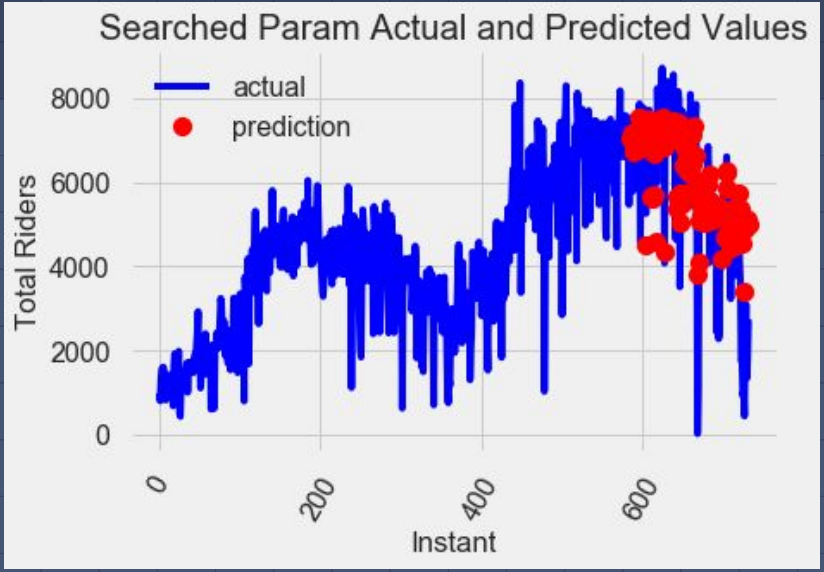


Randomized Search CV shows strong improvements

	Total User Count	Registered Users	Casual Users
First Forest R^2	0.602	0.583	0.785
Searched Forest R^2	0.900	0.864	0.849
First Forest RMSE	1238.04	1051.07	272.48
Searched Forest RMSE	619.04	600.987	228.67

Experimental Forest - Train on first 80% test on last 20%





Notice the underprediction before the search and the overprediction afterwards. this is primarily caused by temporal influences in the data set. That is to say, since our model has time components and patterns based on time of the year using all of that data to predict the next few weeks can cause problems with overprediction if the section we are predicting trends downwards, as it does towards the end of the year.

Recommendations

1. Suggested Implementations of Machine Learning for Capital Bikeshare.

2. Improving Financial Stability by Transitioning Users to, and Expanding the, Registered subcategory of the user base

3. What Should Capital Bikeshare Pursue Next in the Scope of Data Science and Data Analysis?
    a. Station by Station Data Collection
    b. Crowd-Sourced User Preference Data

3

Suggested Implementations of Machine Learning for Capital Bikeshare

- I suggest using this tool to forecast usage numbers over the next season or year.

- These predictions will inform them as to when they will need as many bikes on the road as possible, and when they can take more bikes off the road for preventative maintenance procedures.

Improving Financial Stability by Transitioning Users to, and Expanding the, Registered subcategory of the user base

Value of Transitioning the user base to registered users

The subscription models guarantees income monthly and smooths cash flow.

How to transition the casual user base

Advertise and/or partner with local universities to encourage new students to sign up for the annual subscription service.

What Should Capital Bikeshare Pursue Next in the Scope of Data Science and Data Analysis?

Station by Station Data Collection

Collecting data from each bike rack will give insight into the usership in different areas of the city. What kind of usership occurs at the Smithsonian bike rack? When will the demand there be highest? That data will answer those questions.

Crowd-Sourced User Preference Data

Understanding why their customers use the service could provide incredible insights into targeting subsections of the current and potential user base with different types of marketing. Asking the users directly, and perhaps compensating them with one free casual ride, should prove to be an excellent source of data.

About the author:

I am Jonathan Orr, a Data Science Career Track student at Springboard.

You can contact me at: https://www.linkedin.com/in/jondavidorr/

CREDITS

Special thanks to all the people who helped make this possible:

- Guidance from Springboard mentors including:
  - Vaughn DiMarco
  - Dipanjan Sarkar
  - Kenneth Gil-Pasquel
- Presentation template by SlidesCarnival