

Forecasting Washington D.C.'s



Hourly Bikeshare Demand:

A Kaggle Competition

Bennet Voorhees

The Problem and Question

The goal of this Kaggle competition is to forecast hourly bike share demand in the the entire Capital Bikeshare region. Or, in Kaggle's words :

“In this competition, participants are asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.”

...it may not look explicit in the description, but Kaggle actually wants folks to forecast demand using information from the past. In other words (in Kaggle's Words:

“Your model should only use information which was available prior to the time for which it is forecasting.”

So there are really two complimentary questions to be answered:

- **Which of the given data determine Capital Bikeshare demand?**
- **What is the hourly demand of rentals given those factors?**

Hypothesis

From personal experience, I believe bike share is most popular during the evening commute when the temperature outside is warm; or during the weekends when the temperature outside is also warm, but after everyone is done with brunch or crawling out of bed after a rough night at the bars (11AM-3PM).

The Data

The Raw Data:

The dataset provided by Kaggle contains information about time, weather, and demand only. The data are already very tidy; there are also no missing observations for any of the variables. A [codebook](#) is provided at the end of the document.

Table 1: “Raw Data” provided by Kaggle

Time	Weather	Demand
Year/Month/Date/Hour	Season	Registered
Holiday?	Conditions	Casual
Workday?	Temperature (2)	
	Windspeed	
	Humidity	

Processing the raw data:

The only processing the data really need is breaking apart the time variable into its individual components (year, month, date, weekday, and hour). Some of these variables will be turned into categorical dummy variables (which I’ll discuss below).

Additionally, per the Kaggle competition’s rules, I have created lagged variables for each of the weather variables above, as well as some rolling averages.

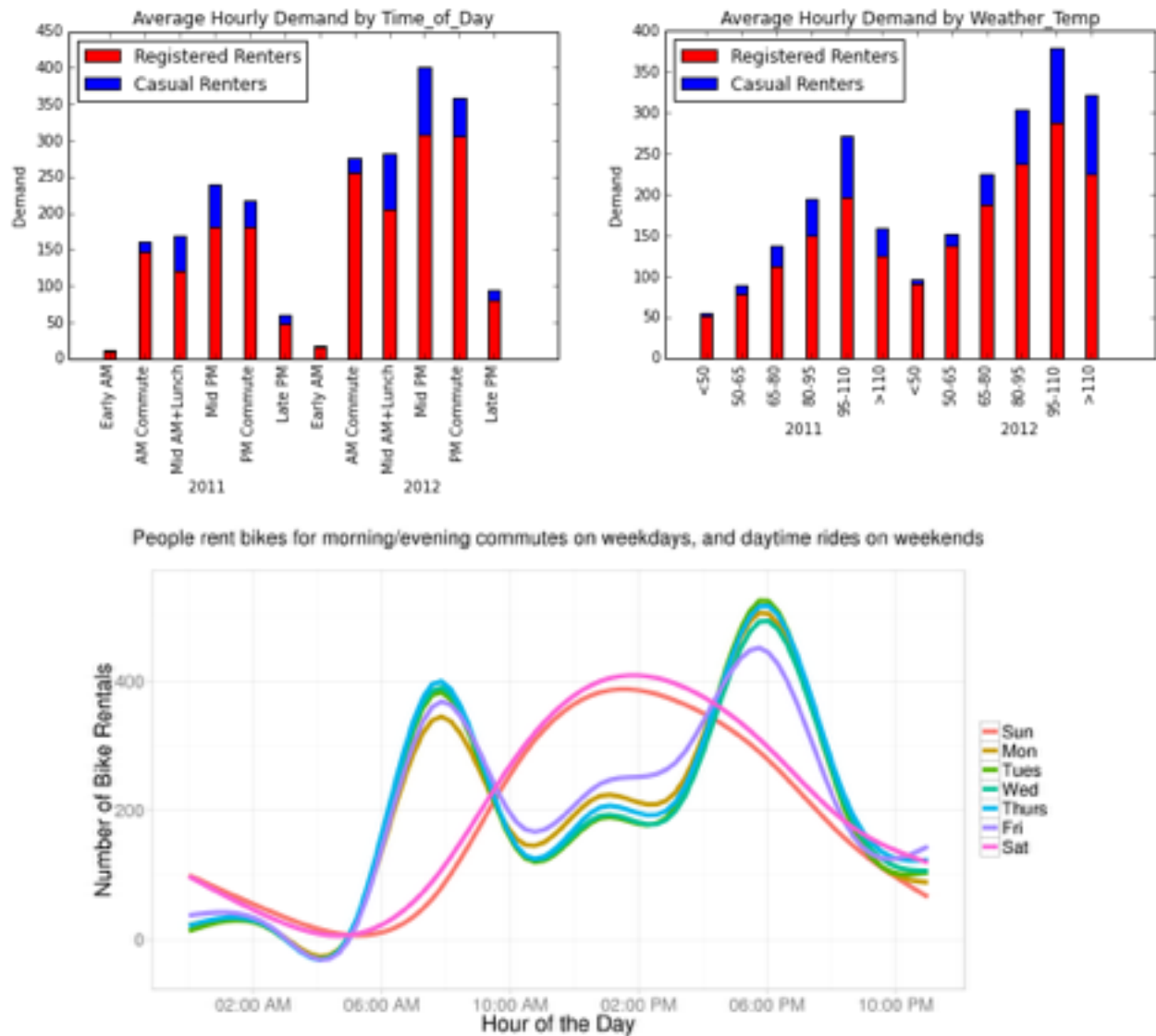
Exploratory Analysis

After feature engineering, I’ve taken a look at all of the variables in the dataset vis-a-vis the two demand variables (casual rider rentals and registered rider rentals). Figure 1 shows a breakdown of—what I think—are the most intuitive relationships.

Average hourly demand indeed is high during the weekday commuting hours (or shortly after brunch time during the weekends) and when the weather is particularly warm. Further, there seems to be a “word of mouth” or “availability” effect across time, as average hourly 2012 demand is higher than in 2011.

Figure 1: Demand vis-a-vis Time of Day and Temperature

(Note: bottom center chart was produced by Kaggle)



Feature Choices, Model Specification, and Results

Since I estimating continuous response variable (demand), I used a regression model.

Linear Regression:

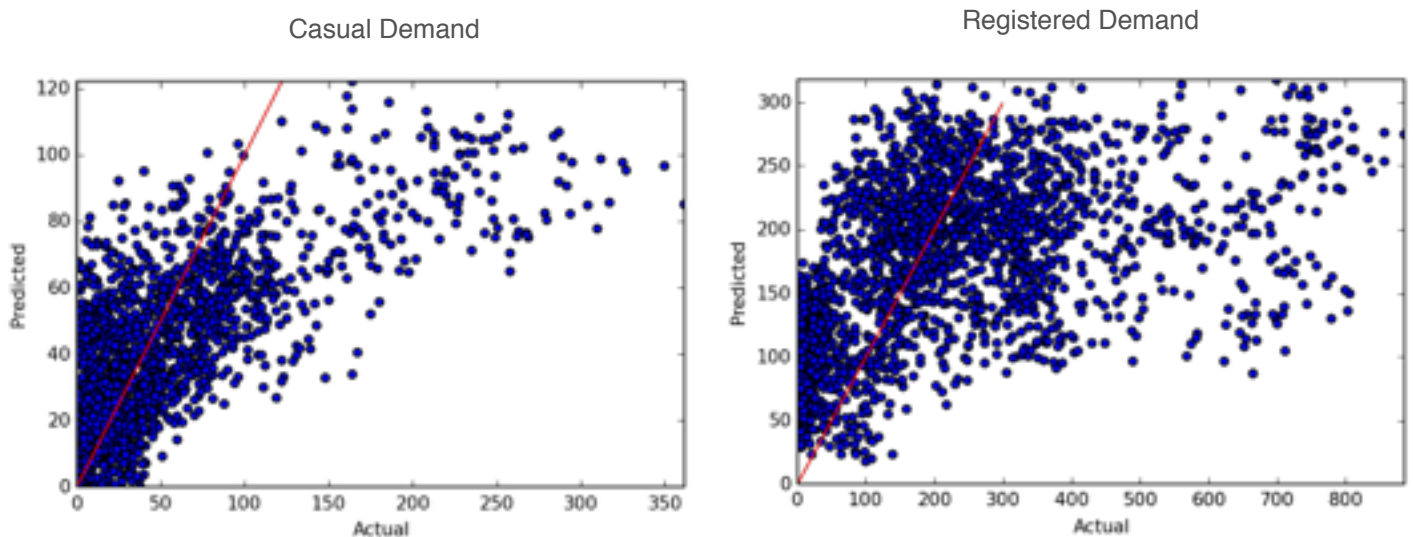
To start, I wanted to see the explanatory power of every non-lagged variable. Table 2 shows each variable's explanatory power for explaining variation in the two types of demand. Given that, on average, hourly casual demand is 36 and hourly rental demand is 155, not one of these variables alone has good particularly good standalone explanatory power.

However, I ended up arbitrarily selecting the top 5 variables ranked by explanatory power that were related (so to avoid multicollinearity) and plugged them into a multi-variable regression model. These variables are bolded in table 2.

Table 2: Variable explanatory power in single-

RMSE (Casual Users)	Variable	RMSE (Registered Users)	Variable
41.2	temp	138.4	Hour
41.4	atemp	140.1	temp
46.6	Workday	140.2	atemp
47.0	Hour	143.2	Year
47.0	humidity	144.5	humidity
47.5	Weekday	147.6	Workday
47.8	Weekday_5	147.7	weather
48.1	Weekday_6	147.8	weather_3
48.3	weather	148.1	Weekday
48.5	Weekday_2	148.1	Weekday_6

Figure 2: Performance of Multivariable Linear Regression



This improved the casual demand RMSE to 35, and the registered demand RMSE to 111. These errors were not stellar, considering that average hourly demand is about 150 and 40 for registered and casual users, respectively. The models also underpredicted high-demand periods (figure 2).

Decision Trees and Random Forests:

I decided put the linear regression model on the back burner, given its performance and technical difficulties in employing lagged variables (e.g., auto serial correlation and multicollinearity).

The next step was to use a decision tree regressor to better understand the data so to build a better predictive model. I used a decision tree and random forest in compliment to tune various parts of my model. This included using a grid search for both casual and registered demand to tune:

- The number of estimators in the forest
- The number of features to consider at each split of the tree
- The depth of the trees

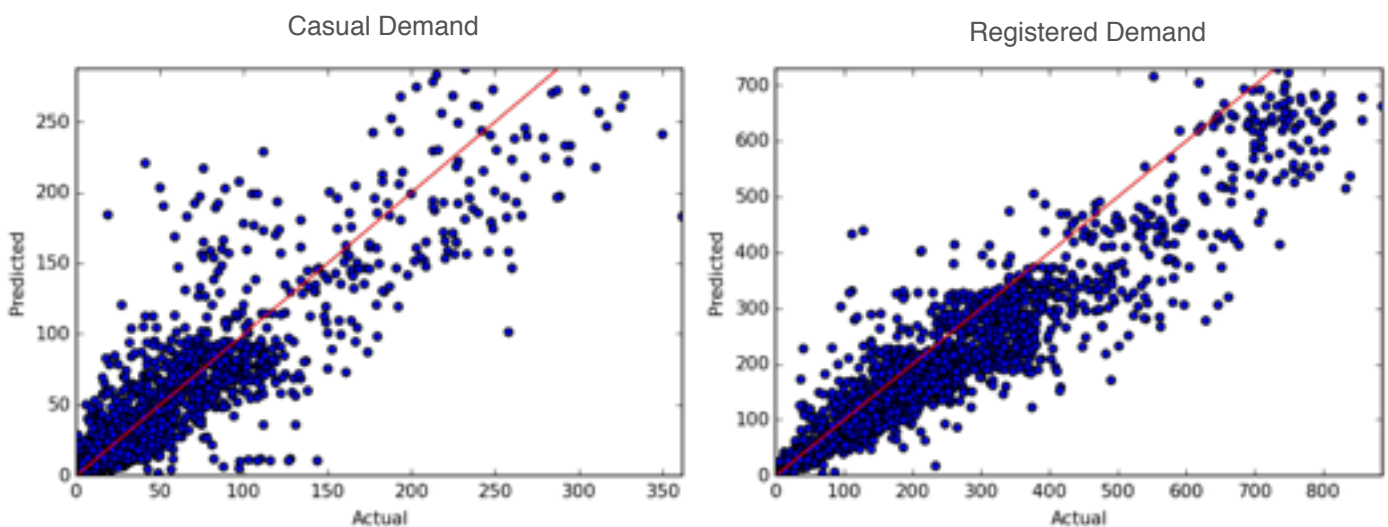
Table three shows that the feature importances were not astoundingly different than the single linear regression model results shown in table 2. The random forest estimator did a better job at picking out the year as an important feature—which was clearly something that marked an uptick in average hourly demand (as shown in the exploratory data analysis section of this paper).

Table 3 - Top 10 Random Forest Variable Importances

(Casual Users)	Variable	(Registered Users)	Variable
0.356	Hour	0.586	Hour
0.180	Workday	0.127	Workday
0.114	atemp_ravg_1	0.072	Year
0.094	atemp_l_1_sq	0.044	atemp_ravg_2
0.051	atemp_l_1	0.023	atemp_ravg_1
0.046	Year	0.022	atemp_l_1
0.022	atemp_ravg_2	0.019	atemp_l_1_sq
0.015	humidity_ravg_2	0.012	weather_ravg_2
0.013	windspeed_ravg_2	0.011	atemp_l_2
0.012	humidity_l_1_sq	0.011	atemp_l_2_sq

The random forest did a much better job of predicting casual and registered demand (figure 3). On average, the casual model had an root mean squared error of 23 , whereas the registered model was 60 . This is a marked improvement over the multi-variable linear model. However, it still represents a large margin of error in terms of average hourly demand for both registered and casual users.

Figure 3: Performance of Random Forest Regressor



Possible Improvements and Extensions

There are probably many other factors that influence bike-share demand at both the system-wide level, and at the station-level. As such the first step would be to get more data.

For example, the WMATA has data on metro closures and delays, which could be used to see if there is an uptick in demand during metro or bus bottleneck periods. Getting disaggregated data from Kaggle on station demand could allow the WMATA data to have more explanatory power for bikeshare stations that are close to (or even far away from) bus stops or metro stations.

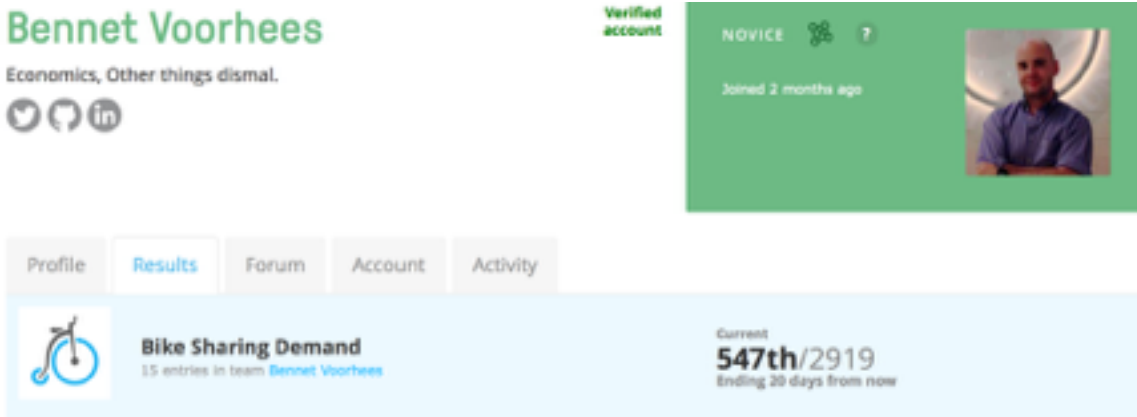
Conclusions, Challenges, Successes, and Takeaways

Challenges:

1. Learning what makes a decision tree/random forest classifier different than a regressor.
 - 1.1. We had covered in class the decision tree classifier, but not the regressor. I had to use online resources to find out how the algorithm decided to make a split. I also had to dive into scikit learn documentation to understand what arguments were different in the regressor method from the classifier method.
2. Learning how to tune a machine learning model
 - 2.1. I had to do a lot of review of the basics of tuning a decision tree or random forest. Once I reviewed the theory behind the algorithm and consulted stack overflow for advice, I made use of the gridsearchCV function to tune my trees and forests.
3. Learning about hardware limitations
 - 3.1. I learned computing takes a long time, and considering CPU and memory resources is important.
 - 3.2. The `n_jobs` argument allowed me to use all of the cores of my CPU to do cross validation and parameter tuning.
 - 3.3. This small dataset created a big computational task for my computer. I need to learn more about Amazon Web Services.

Successes:

Given this, I decided to make a submission to Kaggle with the random forest model— and it seems that the results weren't terrible relative to other competitors (I'm in the top 20%)!




Bennet Voorhees
Economics, Other things dismal.
Twitter GitHub LinkedIn

Verified account

NOVICE 7
Joined 2 months ago

Profile Results Forum Account Activity

 **Bike Sharing Demand**
15 entries in team [Bennet Voorhees](#)

Current
547th/2919
Ending 20 days from now