How Many Bikes are Needed?

Exploratory Analysis and Regression of a Bike Sharing Dataset

Shannon Crucy

June 10, 2016

Overview and Dataset

- These slides summarize an exploratory analysis of a bike-sharing dataset, including developing a prediction model for the total number of rented bikes.
- The dataset contains information on the rental of bikes per hour by Captial Bikeshare system, Washington D.C., USA in the years 2011 and 2012. It contains 17379 records.
- Outline:
  - Available Features
  - The Models
  - Conclusions

Available Features

- Information available in the dataset:

Date	Temperature
Season	Perceived Temperature
Year	Humidity
Month	Windspeed
Hour	Number of Weather situation
Holiday flag	Number of Registered Users
Workingday flag	Number of Casual Users
Day of Week	Total number of rented bikes

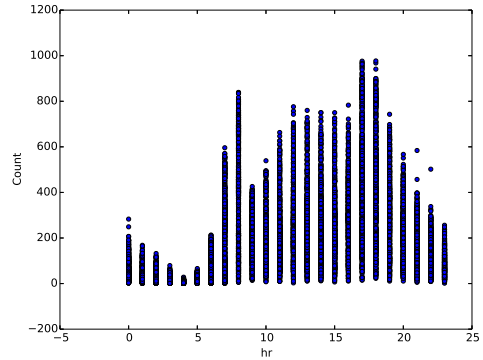
- Note: $Totalnumberofrentedbikes = RegisteredUsers + CasualUsers$. As the goal is to predict the total number of rented bikes, presumably somewhat in advance, the number of registered and casual users would not be available nor be a good input feature.
- The exact date will also be omitted as containing a high level of overlapping information with the year, month, and weekday fields.

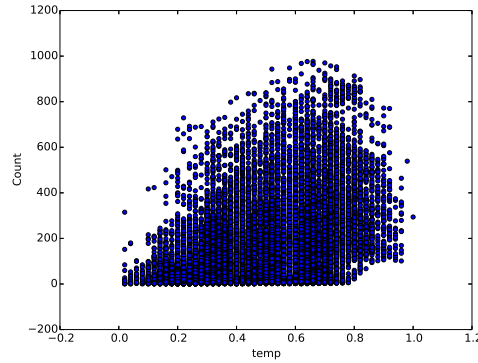
Feature Selection

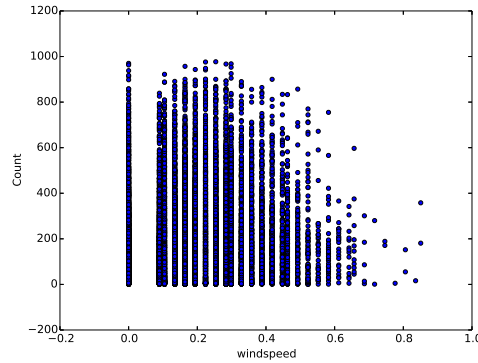
- The best features were determined by their linear correlation with the count of rented bikes (via sklearn's SelectKBest). Algorithms were built on groups of the best six, best nine, and all twelve available input variables, to check for possible over-fitting due to insufficient amount of input data relative to the features used.

Best 6	Best 9
Weekday	Weekday
Temperature	Temperature
Perceived Temperature	Perceived Temperature
Windspeed	Windspeed
Hour	Hour
Weather Situation	Weather Situation
	Humidity
	Working day flag
	Year

- Visually, the variables related to hour, temperature, and windspeed have the strongest correlation with number of rented bikes (additional scatter plots for other variables in the back-up).





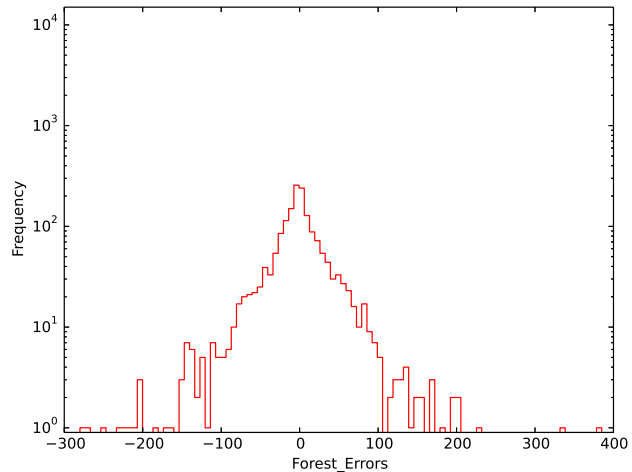


The Algorithms

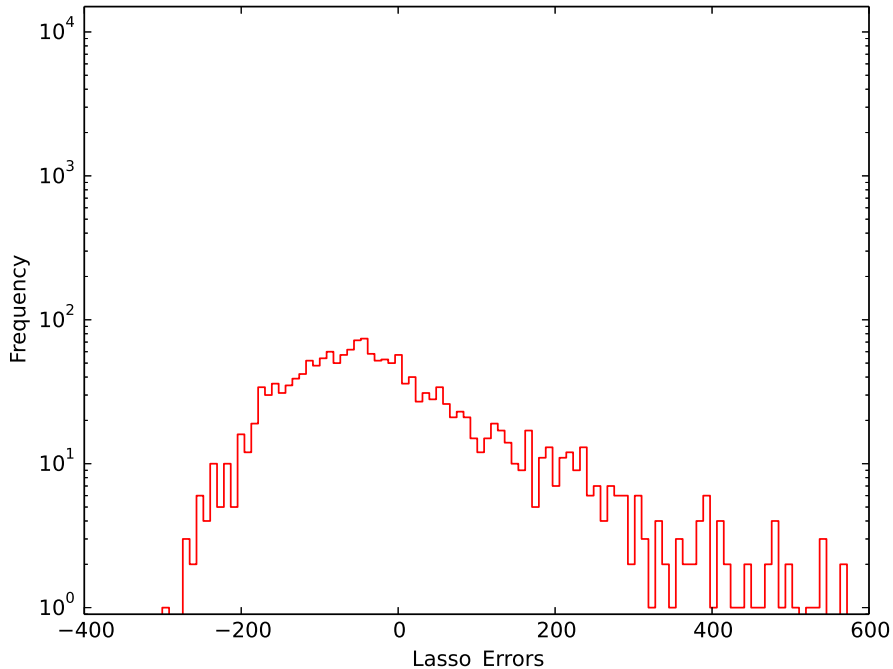
- Lasso and Random Forest regression algorithms (as implemented in sklearn) were built on each set of input variables; these algorithms were selected for their resilience against weak input variables. All features were scaled prior to training and a KFold of five was used. The mean absolute error was used for comparison, and 10% of the input data was reserved for evaluation.

'Best Algorithm': Random Forest

- 12 input features
- 100 estimators, minimum sample split of 6, maximum depth of 12
- Best KFold score: -43.68$\pm$5.54
- Global training score: -20.57
- Evaluation score: -28.98

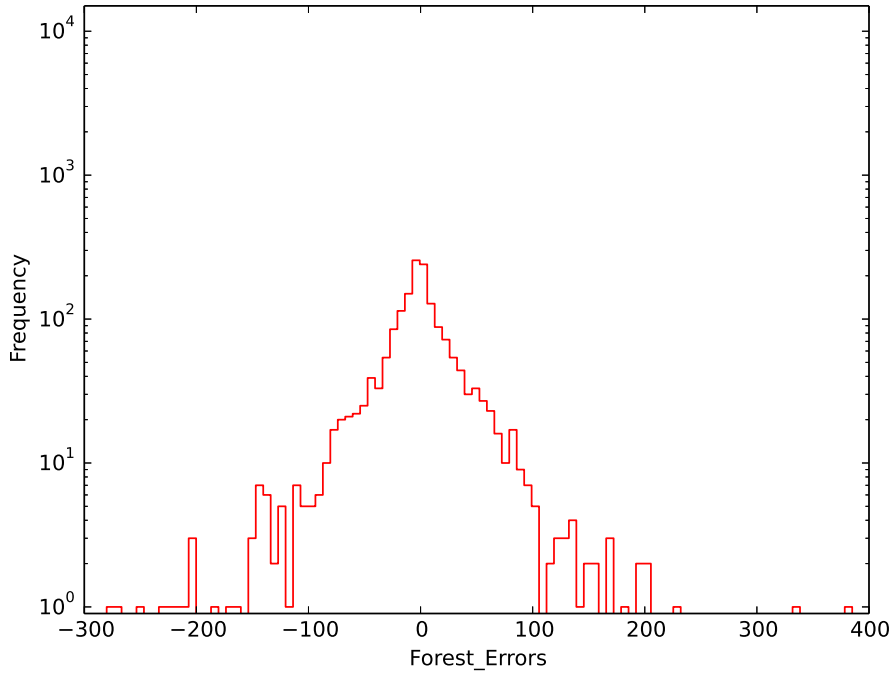


Lasso vs. Random Forest



- The Lasso algorithms showed greater consistency between KFold and evalution scores, but with worse scores and larger uncertainties than the Random Forest algorithms.
- They also seemed to show a distinct trend to overestimate bike usage (plot on right shows true-predicted).

Lasso vs. Random Forest



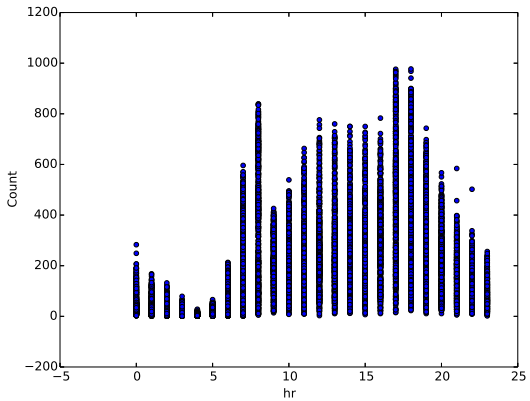
- The Random Forest algorithms produced better overall scores with smaller standard deviations and shows a symmetric error distribution (right). However, for a given algorithm the differences between best KFold, global, and evaluation scores were between 2-4 standard deviations, hinting at possible over-training.
- Both distributions hint that outliers might be present.
- Table for numeric comparison is in the backup slides.

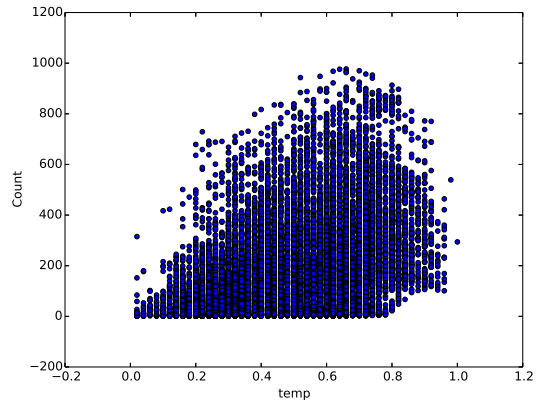
Conclusions

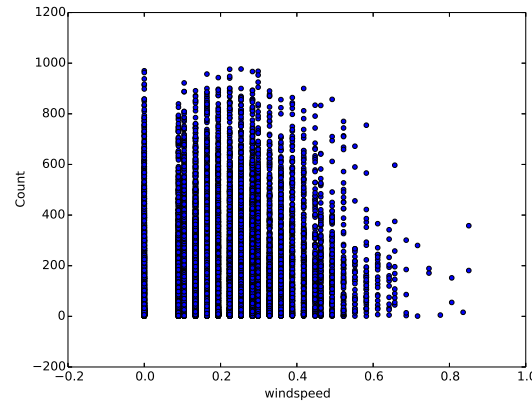
- Both the Lasso and Random Forest algorithms were used to build regressors to predict the number of rented bicycles, with an approximate error of 100 bicycles for Lasso and 50 for Random Forest. The random forest algorithms seemed to perform better but may have been over-trained.
- Expansions:
  1. Further exploration of input variables for correlations and redundencies. Was removing the registered users variable truely necessary? Assuming that users register somewhat in advance, that information would be suitable for inclusion as an input, and would almost certainly improve the performance.
  2. Outlier identification. These preliminary results indicate outliers might well be present, perhaps due to extreme weather conditions such as Hurricane Sandy. A method for identifying and possible removing these should be developed.
  3. Exploration of more algorithms and a larger parameter space would also be beneficial.

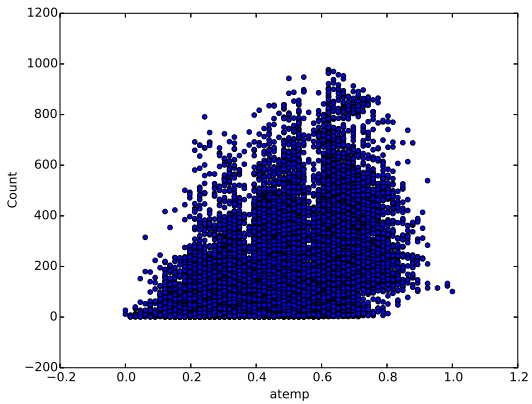
Back-Up

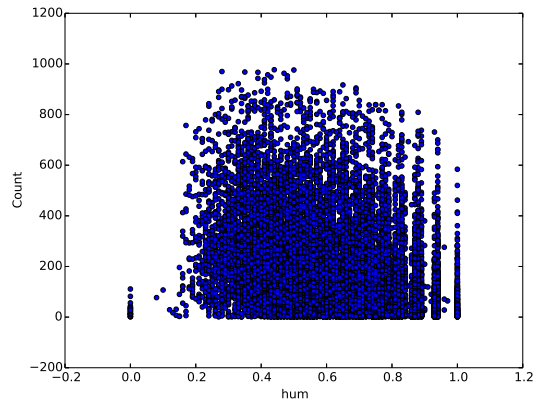
Variable Scatter Plots

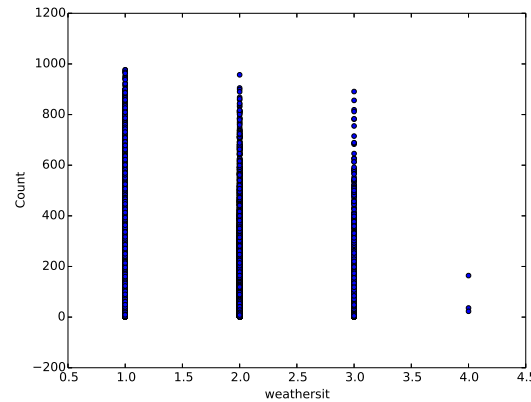




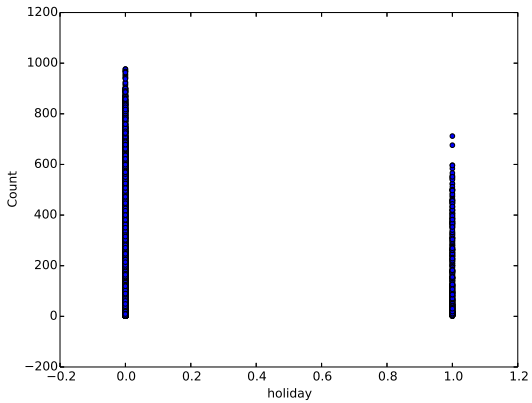


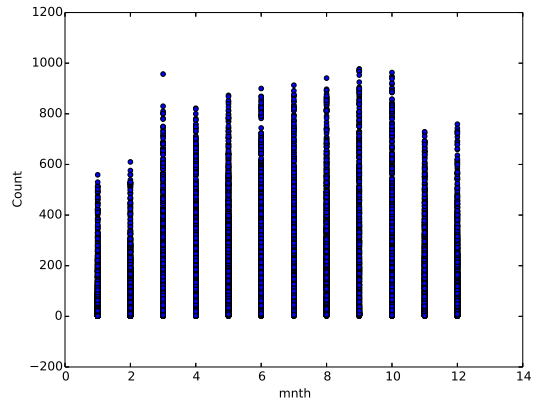


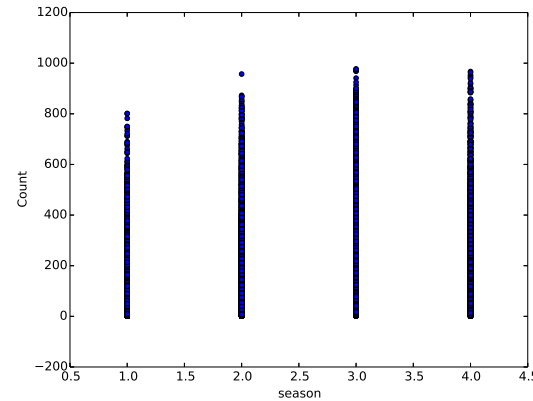


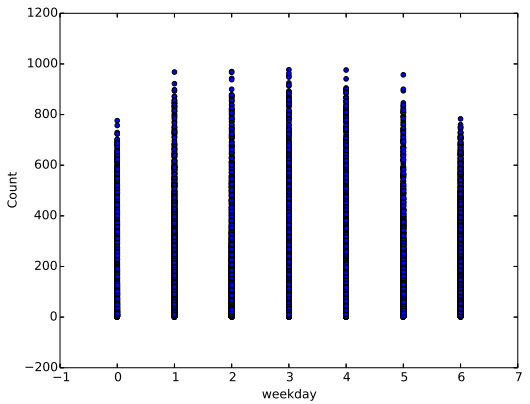


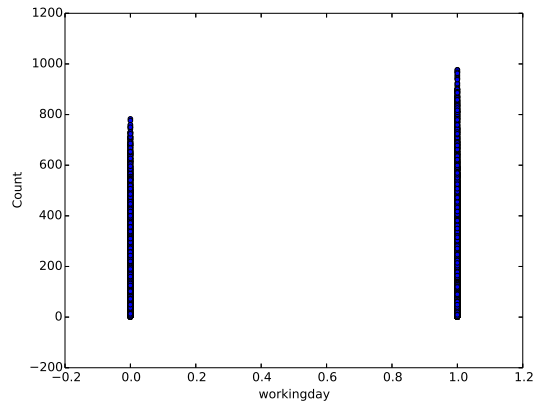
Variable Scatter Plots (page 2)

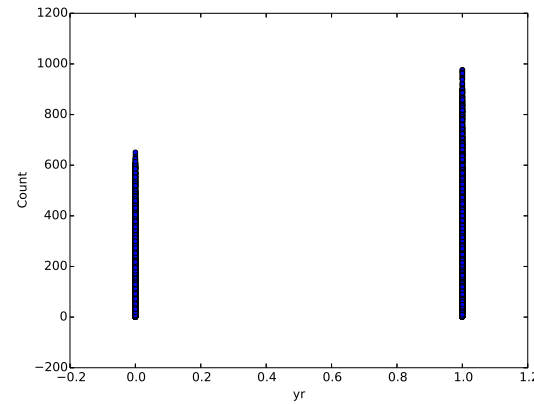








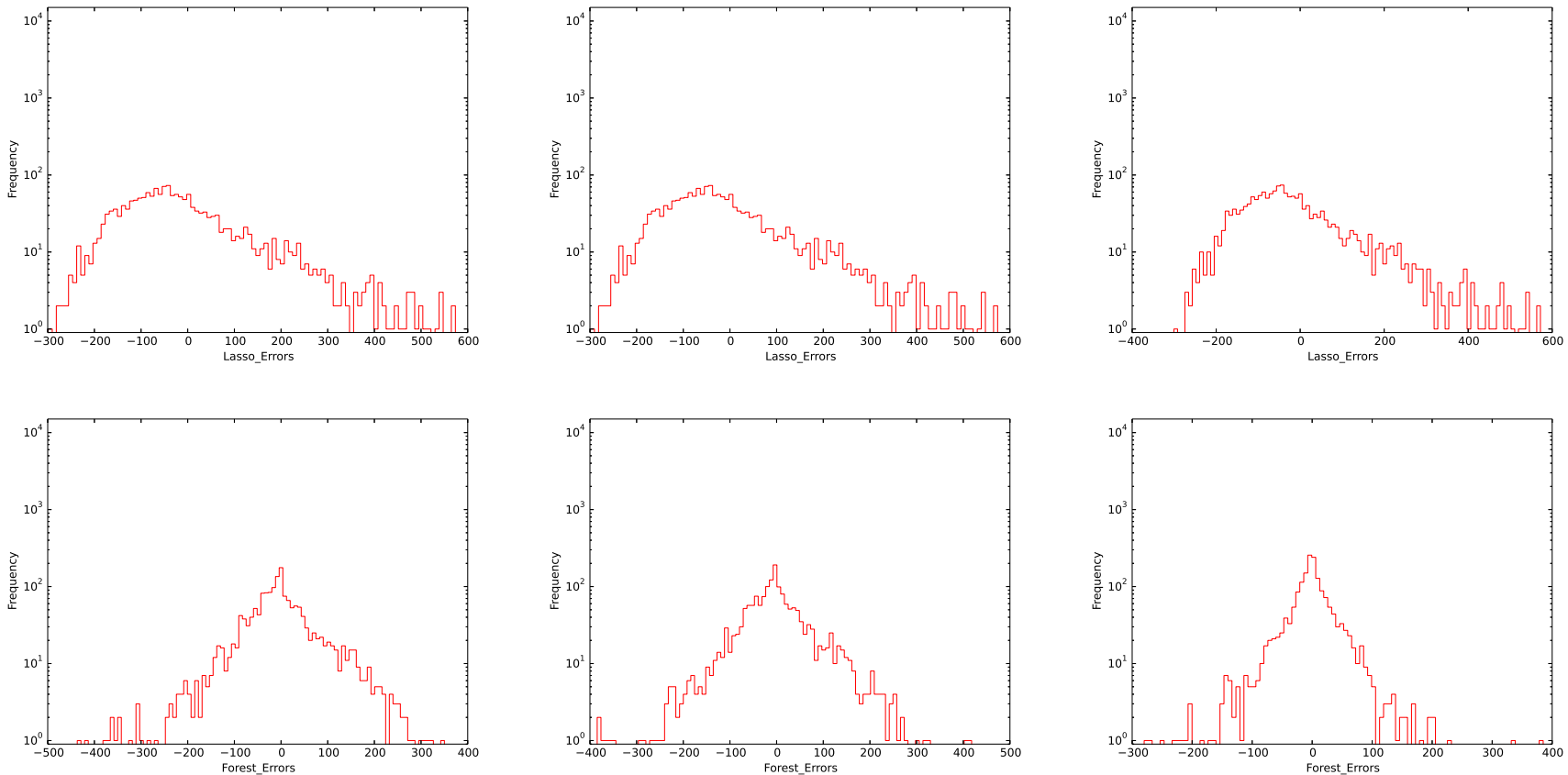




Algorithm Comparison

	Lasso Regression		
Variables	6	9	12
α	1.0	1.0	1.0
Best KFold Score	-109.38 $\pm$ 28.26	-109.51$\pm$28.37	-109.62$\pm$28.34
Global Score	-106.11	-106.11	-106.14
Evaluation Score	-105.42	-105.43	-105.48
	Random Forest Regression		
Variables	6	9	12
min_sample_split	14	10	6
estimators	100	100	100
max_depth	12	12	12
Best KFold Score	-66.87$\pm$15.05	-66.71$\pm$11.77	-43.68$\pm$5.54
Global Score	-49.90	-46.06	-20.57
Evaluation Score	-58.93	-56.84	-28.98

Error Comparison)



Left column is for algorithms with six input variables, center has nine, and right has all twelve.