General Assembly SF 17 – Data Science Class Project

Project Milestone 2. Due Date Oct 26, 2015.

Submitted by Anil G.

--

* The project goal is to predict the bike rental demand in Washington DC.
* The project is based on the completed Kaggle competition – “Bike Sharing Demand - Forecast use of city bikeshare system”
* This is supervised regression machine learning problem.
* The competition provides training and test dataset. On initial exploration the data seems to be good quality. The dataset is quite small and can be processed easily on a laptop.
  + Training Set
    - Two years of hourly data from Jan 2011 till Dec 2012
    - Contains the predictor values for first 19 days
    - Contains the output for the first 19 days
    - 10866 rows of observations
    - No missing values
* Test Set
  + Two years of hourly data from Jan 2011 till Dec 2012
  + Contains the predictor values for last days of the month (from 20th till 28, 29, 30 or 31)
  + Predict the bike demand for the test dataset *only based on the training data available till that time.*

Input or Independent variables

* datetime - this is date and time of the day for the observation in YYYY-MM-DD 24HR:MI:SS format.
* season – This data is in numeric format with 1 = spring, 2 = summer, 3 = fall, 4 = winter
* holiday – whether the day is considered a holiday. Numeric format 0 – means not a holiday and 1 means holiday
* workingday - whether the day is neither a weekend nor holiday. Numeric format 0 – means not a working day, 1 means it is a working day.
* weather – data in numeric format where 1 means: Clear, Few clouds, Partly cloudy, Partly cloudy; 2 means: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist, 3 means: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered cloud; 4means: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
* temp - temperature in Celsius
* atemp - "feels like" temperature in Celsius
* humidity - relative humidity
* windspeed - wind speed

Output or dependent variables

There are actually two sets of out variables.

* casual - number of non-registered user who initiate rentals
* registered - number of registered user who initiate rentals
* count - number of total rentals (i.e. casual + registered)

Since these two dependent variable may have different models to accurately predict the outcome.

Based on the initial data analysis, there seems to be of good quality with no missing values in the data.

Initial scatter plots were drawn against the individual independent variable and the total rentals (*count).*

(Please see the project-ag1.py source code). Note - no analysis was done against the datatime variable as I need to extract the hour of the day from the input so that the plots reflect the bike demands against the time of the.

Following are initial set of observations.

* Surprisingly there are more rentals in fall and winter?
* more rentals on non-holiday
* more rentals on working days - seems like rentals are more often used to commute to work and back
* As expected the rentals drop when weather is really bad. However light snow or light rain or clouds does not affect rental demand as much
* most rentals when the outside temperature of 20 and 30 Celsius when weather is mild
* most rentals when the outside temperature feels like 30 and 40 Celsius. Seems like humidity and weather have strong influence on rental demand and not just the temperature.
* Unless humidity is really low (<20%) or high (>80%) the rental demand is not much impacted by humidity.
* high windspeed reduces the bike rental demand