Machine Learning

Jorge Londono

February 23, 2019

# Model Selection

Since the purpose of this project is to find out how all of the variables in the data set, such as climate and weather, date and time, and holidays affect the ridership of Uber, I will be using a linear regression model to show the relationship of the variables. To be more precise, the dependent variable (amount of rides) is a continuous variable, therefore the most apt model would be a linear regression which would be supervised since the variables to be tested will be specifically selected from the data set.  
  
The main features of the model as mentioned earlier are: the weather conditions (e.g. is it clear, is it overcast, is it snowing, is it raining, etc.), the amount of precipitation (if any), the climate (a measurement of the temperature), the date and the hour of the day the observation took place, whether the observation takes place on a weekend, and whether the observation takes place on a Holiday.  
  
When graphed on a histogram, it is apparent that the ridership variable is not normally distributed, but right skewed. Therefore I will transform the dependent variable with a log function in order to make it normal. This log variable will be used in both the training set and testing set.  
The linear regression model that I will be training will be selected based on the highest adjusted R-square value. When I am looking at which features will be ultimately be selected, the corresponding coefficients for the model will all have to be significant.

# Evaluation of Success

Once the model has been created with my training set, I will test it out using the “predict” function in R on the testing set. The testing set is Uber Ridership data from the months of April through July in 2015. Once my model has been used on the testing set, I will use the R-squared function to find out the value of how my linear regression model preformed. The R-squared function is: 1 – sum of squared errors/total sum of squares.

# Results

With the training set, my linear regression model had an adjusted R-squared value of 0.7874. With the testing set, my linear regression model had an R - Squared value of 0.8756. This is value shows a fairly strong relationship between the features of my model and Uber Ridership.