CSC 483: Machine Learning Fundamentals

Assignment 1: Feature Cleaning, Analysis, and Selection

For this assignment, you will be performing exploratory data analysis similar to what we did in class. This includes looking at summary statistics, graphing, identifying outliers, finding correlations between features, and finally, feature selection.

I have provided a dataset for you to use, and the directions are written with it in mind. That said, if you can find a dataset that you find more interesting, you may use it if it meets the following criteria:

- There are a minimum of 50 observations.
- There is *at least* one feature of numeric values with a ride range of values.
- There is *at least* one feature of categorical values.
  - This can be given as strings, i.e. sedan, van, truck.
  - This may be given by discrete integers, i.e. 1 = sedan, 2 = van, 3 = truck, etc.
- There is a minimum 5 features total in the dataset.
- Identify a feature as the dependent feature. That is, the one you would want to predict or classify.

If you don't wish to provide your own dataset, please use the "auto.csv" file provided. A description of the features can be found in "auto_descriptions.txt". The dependent variable will be **highway-mpg**.

For a passing grade for this assignment, you must do the following as a minimum:

1. Determine if there are any missing values; if so, determine a way to replace any missing values for each feature that has missing values. Use a text cell to me in words how you filled missing values for each feature that needed it.
2. Use one-hot encoding to transform any string nominal features.
3. Create a correlation matrix of all numeric features.
   a. In a text cell, state which features are high correlated with the dependent variable.
   b. Also, make note of any features that are correlated with each other. Highly correlated features will have values below -0.8 and greater than 0.8.
4. Choose a minimum of 3 features to make crosstabs *and* plots versus the highway-mpg. Make two of them the most correlated features and make one of them the least correlated.
   a. Describe in plain English what the plots tell us about the features' relationship to the dependent variable highway-mpg.

Above and Beyond Ideas

1. Features that are not highly correlated with the dependent variable often aren't useful when creating models. If the correlation is between [-.2,.2], identify and remove those features from the dataframe.

2. Features that are highly correlated with each other, regardless of their correlation to the dependent variable can also cause noise hurting our model. Identify independent features that are highly correlated with each other and remove the ones that are the less correlated with the dependent variable of the two. Assume two features are highly correlated if their value is less than -0.8 or greater than +0.8.
3. Go ahead and train a model, fine tune hyperparameters, report error rates.