Christopher Antos

Springboard 3.4

Project Proposal

Don’t Over Fit! II

Overfitting is common issue that can occur within many types of prediction models. When a data scientist over fits a model, they are training the model to work with one specific dataset. This dataset that may not be representative what occurs in the world. This is due to either their training data being too small or not necessarily representative of the whole population. At the same time, some groups such as start-ups only have small amounts of data that they can collect and must make inferences on the responses of large groups to evaluate if their idea is worth it. The goal of this project is to demonstrate some techniques that can help avoid overfitting especially in a classification sense.

The data is coming from the Don’t Over Fit! II Kaggle Competition. The data has already been cleaned for missing values. There are 250 observations for 300 continuous variables and 1 binary target variables. Various forms of decision trees, support vector machines, and gradient boosting models will be used to analyze and classify results. Model diagnostic plots can give insight to what variables are import. Other important visualizations can include histograms and boxplots to determine any sort of skew within the data.

TMBD Movie Data

Movies throughout the past century have been source of entertainment for audiences around the world. Despite their entertainment value, they also cost large amounts of money to produce. To justify these costs, the production company must be able estimate how much the movie will make. TMBD provided data for 7000 movies for us to analyze to determine revenue.

To analyze the data, the following steps must be taken. Missing values either need to be dropped or inputted with a reasonable assumption. From there, strings are changed to categorical variables for ease of analysis. Categorical variables from there will be used to group various variables for box plots. For the prediction itself, gradient boosting, random forest or regression model can be used. Models will be trained and tested. Some will use k fold validation. Roc curves and model influence correlation plots will help determine what variables are import.