Lucas Reynolds
Springboard Data Science Career Track
January 2020 Cohort
Capstone Project 1

Machine Learning

Overview

My goal for the machine learning step was to evaluate multiple algorithms and choose the one with the best performance for our model. The model needs to best predict whether a review would have a "good" rating (8 or above out of 10) based on drug review text.

Models

I tested three algorithms and the highest performing was a Random Forest Classifier. The model using this algorithm was able to achieve a score of about 0.93 while the others both scored around 0.75. This means the model has about a 93% probability of correctly distinguishing whether a drug review is for a good rating or not. A model choosing at random would score around 0.5.

The Random Forest Classifier performs well for this situation because it builds multiple decision trees. Each decision tree within the model branches randomly at different points and the algorithm then averages the decision of each tree. This averaging of multiple decisions helps improve the accuracy while also helping the model better predict on new data moving forward.