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

Machine Learning

Project Objective

Our objective is to create a model that can predict whether or not a drug review will have a good rating based solely on the review text.

Machine Learning Objective

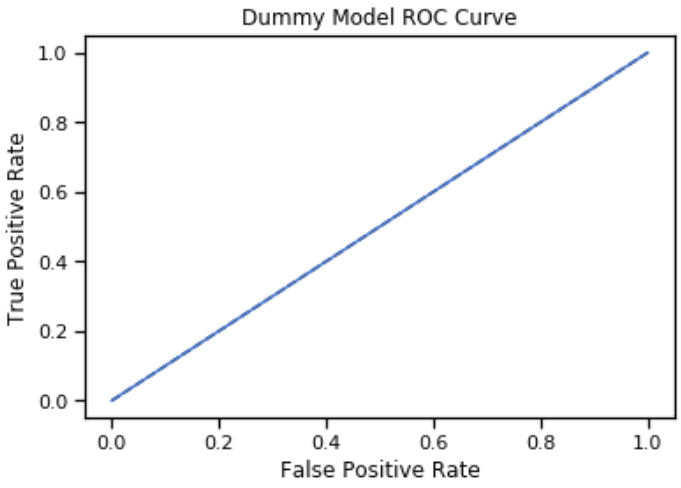
My goal for the machine learning step was to evaluate multiple algorithms and choose one with the best performance for our model. The model needs to best predict whether a review would have a "good" rating (8 or above out of 10) based on the review text.

Measuring Performance

For this project, Receiver Operating Characteristic (ROC) curve and the area under the ROC curve (AUC) are the most appropriate measures for model performance. The ROC curve is a visual representation of true positive and false positive rates. The AUC is the area of the plot under this curve. Better models will have ROC curves closer to the top left corner and larger AUC scores. The model performance examples below will illustrate this more clearly.

Weak Model

Understanding the performance of a weak model helps with distinguishing strong models. As a baseline, I used a model that randomly predicts whether or not a review has a positive rating. The random model achieved an AUC score of 0.50 and the ROC curve is plotted to the right. This means the model has the same rate of false positives as it does true positives.

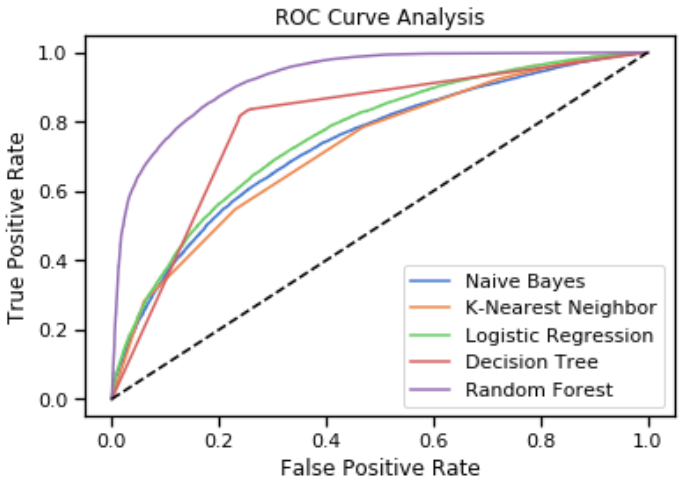


Model Selection

I tested some of the algorithms that are most commonly used for a classification problem like ours. These algorithms are Naive Bayes, K-Nearest Neighbor, Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. All of these models achieved an AUC score between 0.72 to 0.79, except for

the Random Forest Classifier (RFC). RFC outperformed the others with an AUC score of 0.92 (see below).

Algorithm	AUC Score
Random Forest Classifier	0.92
Decision Tree	0.79
Logistic Regression	0.76
Naive Bayes	0.73
K-Nearest Neighbor	0.72



Note how Random Forest has the ROC curve furthest to the top left and the largest AUC.

Model Differences

Random Forest performs so highly because of its ensemble approach. The algorithm creates a "forest" of decision trees with each tree branching randomly at varied points. The final prediction is then based on the average of each tree. This averaging of multiple trees increases the model's ability to overcome biases or overfitting of the data.

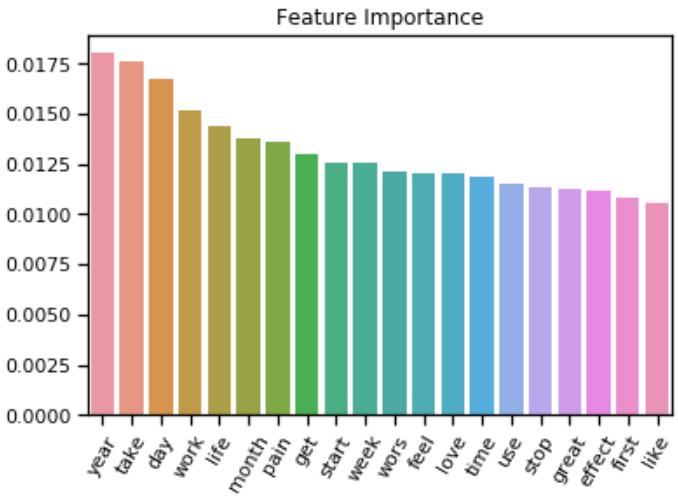
Model Tuning

The Random Forest Classifier has parameters that can be adjusted to improve the model's performance. Two parameters that have the largest impact are the number of trees and the maximum depth of each tree. The defaults for these two parameters are 100 trees and no max depth. With a function called Grid Search Cross Validation we're able to test multiple combinations of these parameters and evaluate which combination achieves the best performance.

150, 200, 250, and 300 trees with max depths of 40, 60, 80, and 100 were tested. The top performing combination was 300 trees and a max depth of 80. The AUC score with these adjusted parameters was minimal, from about 0.9248 to about 0.9279. For this project I would use the default parameters. The default settings performed close to the tuned parameters and with 200 less trees will be more efficient.

Feature Importance

Feature importance measures how much each word influences the model's prediction. Below are the 20 words with the largest feature importance. There were words like pain, wors, love, great, and like that you may expect, but I'm surprised that words like year, take, day and month are near the top.



Possible Next Steps

If I were to revisit or expand this project I'd do the following. I first would like to see how the model performs on additional drug reviews, particularly reviews from different sites. I'd also like to see how a more complicated model performed on the data, for example a recurrent neural network.