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

Milestone Report

Overview

Drug reviews provide important information to drug manufacturers, such as levels of customer satisfaction, successful outcomes, or negative side effects. Several websites exist with drug reviews that could potentially be scraped and used for analysis or product research. However, the rating scales and information required often differ from site to site. If we could predict whether or not a rating would be good based on the review text, we would be able to label reviews with missing ratings, standardize ratings across review sources, and better understand the factors that most influence a drug's rating.

Data Wrangling

Dataset

Drugs.com provides a database of peer-reviewed information on various medicines as well as user ratings and reviews gathered from across the web. A dataset including 215,063 reviews from Drugs.com was available via the UCI Machine Learning Repository.[1]

The dataset contained 161,297 rows and the following 7 columns: *index* (review ID number), *drugName* (name of the drug reviewed), *condition* (diagnosed condition for which the drug was prescribed), *review* (given text of the review), *rating* (rating from 1.0 to 10.0), *date* (review data), *usefulCount* (number of website visitors who found the review useful). The dataset was imported as a pandas dataframe and previewed to gain an understanding of overall structure and context.

Missing/Incorrect Data

899 rows were missing condition data and an additional 900 rows had a comment about the number of usefulness votes the review received rather than the actual condition (e.g. "2 users found this comment helpful"). The total of these rows with missing or incorrect values was only around 1% of all rows. To avoid removing these rows and the reviews and rating data they contained, I filled the condition values with "NoCondition."

[1] https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

Rating Context

Over 30% of reviews had a rating of 10 and ratings 1, 8, and 9 each held over 10% of reviews. I had some expectation that most reviews would be at one end or the other of ratings, but was surprised that the large majority had a rating of 10. Ratings 2 through 7 respectively held 6% of the overall reviews.

To understand general sentiment and context, I sampled 10 reviews from each rating score. Reviews sampled for ratings 1 through 4 generally had an overall negative sentiment, reviews sampled for ratings 5 through 7 were more neutral or contained a mixture of positive and negative sentiment, and reviews sampled for ratings 8 through 10 had an overall positive sentiment.

Cleaning Text

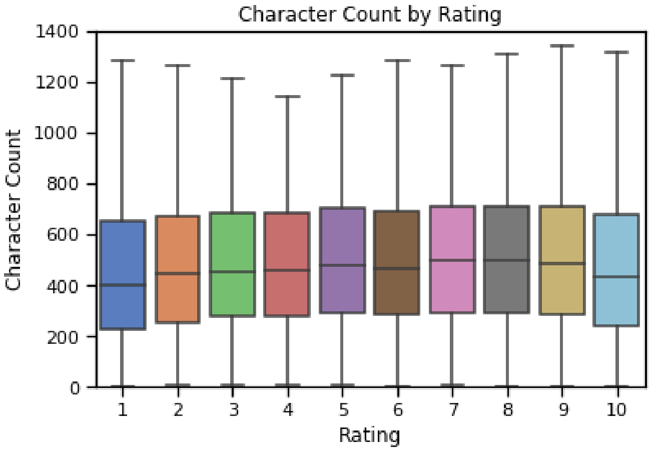
A bag-of-words approach was used for analyzing the review text. Essentially, reviews are separated by words and a matrix or table is created with a column for each word and a row for each review. The matrix is filled with the number of times each word appeared in each review. This allows us to evaluate the text quantitatively.

Before creating the matrix, text was cleaned and standardized. All words were made lowercase, any punctuation, characters, or words that weren't essential to the meaning of the text were removed. To avoid treating the same word with different tenses or endings as different words, the text was stemmed (e.g. aching, ached, and ache are all stemmed to ach). These steps of standardization not only help to reduce the size of our matrix and time it would take to train and run a model, but simplifying our data also will improve our model's performance.

Exploratory Data Analysis

Review Length

The length of reviews was calculated using character count. Boxplots were used to visually explore any relationship between review length and rating. There didn't seem to be a strong relationship. I had anticipated reviews with the highest and lowest ratings (1 and 10) would be the longest, but they actually tended to be slightly shorter.



To better evaluate a relationship between review length and rating, I calculated Pearson's correlation coefficient. This coefficient helps us quantify if there's a linear relationship (e.g. longer reviews have higher ratings). The coefficient was around 0.02 which is very weak, essentially the two variables are not linearly correlated.

Drug Names

To determine drugs with the largest effect on rating, I first determined which drugs had significant average differences in ratings. A p-value of 0.05 was used to determine significance. 233 drugs had statistically significant average differences in rating ranging from an average decrease in rating of 4.87 to an average increase of about 3.00. Below are the top 5 drugs with largest significant average increase and largest significant average decrease in rating.

5 Drugs with Largest Avg. Increase in Rating		
	Drug Name	**Sig. Avg. Diff. in Rating**
1	Privine	3.00
2	Zinc Oxide	3.00
3	Astelin	3.00
4	Acetaminophen / pseudoephedrine	3.00
5	Chlorpheniramine / phenylephrine	3.00

5 Drugs with Largest Avg. Decrease in Rating		
	Drug Name	**Sig. Avg. Diff. in Rating**
1	Systane	-4.87
2	Succinylcholine	-4.62
3	Trimethoprim	-4.25
4	Chloraseptic Sore Throat Spray	-4.00
5	Monistat 7	-3.97

The 5 drugs with the largest average increase in rating are all over the counter (OTC) options for common conditions like nasal congestion, allergies, pain, and rashes. The 5 drugs with the largest average decrease in rating included some OTC options for eye drops, sore throat, and yeast infection along with a muscle relaxant and an antibiotic for urinary tract infection that requires a prescription.

Medical Condition

Similar to drug names, I explored which medical conditions had a significant average difference in rating. To minimize the effect of outliers and conditions with small sample size, conditions that appeared less than 50 times were not included. Again, a p-value of 0.05 was used. There were 75 conditions with significant average differences ranging from an average decrease in rating of about 3.25 to an average increase in rating of about 3.00. Below are the top 5 conditions with the largest significant average increase and the top 5 conditions with the largest significant average decrease.

5 Conditions with Largest Avg. Increase in Rating		
	Medical Condition	**Sig. Avg. Diff. in Rating**
1	Costochondritis	3.00
2	Menopausal Disorders	3.00
3	B12 Nutritional Deficiency	3.00
4	Von Willebrand’s Disease	3.00
5	Gingivitis	3.00

5 Conditions with Largest Avg. Decrease in Rating		
	Medical Condition	**Sig. Avg. Diff. in Rating**
1	Bronchospasm Prophylaxis	-3.25
2	Sore Throat	-3.11
3	Herpes Zoster, Prophylaxis	-3.00
4	Macular Edema	-2.85
5	Pancreatic Exocrine Dysfunction	-2.50

The 5 conditions with the largest significant average increase in rating include pain-inducing inflammation of the rib cage, menopausal disorders, a nutritional deficiency that results in tiredness and weakness, hemophilia, and gingivitis. The 5 conditions with the largest significant average decrease in rating include include asthma, sore throat, shingles, build up of fluid in the retina, and a deficiency of pancreatic enzymes that leads to an inability to digest food properly

Words

I explore the 10 most frequent words in reviews for each rating. At first there was a lot of overlap of certain words across the top 10 words of each rating. I revisited my stop words and removed words with the most overlap. The top 10 words for each rating are listed below.

Using a p-value of 0.05, I compared ratings with and without each word and determined words with a significant average difference in rating. There were 342 words with significant average differences in rating. The significant average differences ranged from an average decrease in rating of about 2.15 to an average increase in rating of about 1.76. Below are the 5 words with the largest significant average increase in rating and the 5 words with the largest significant average decrease in rating.

	5 Words with Largest Avg. Increase in Rating	
	Words	**Sig. Avg. Diff. in Rating**
1	best	1.76
2	love	1.53
3	great	1.41
4	life	1.40
5	year	1.39

	5 Words with Largest Avg. Decrease in Rating	
	Words	**Sig. Avg. Diff. in Rating**
1	wors	-2.15
2	horribl	-1.50
3	stop	-1.48
4	switch	-1.43
5	extreme	-1.43

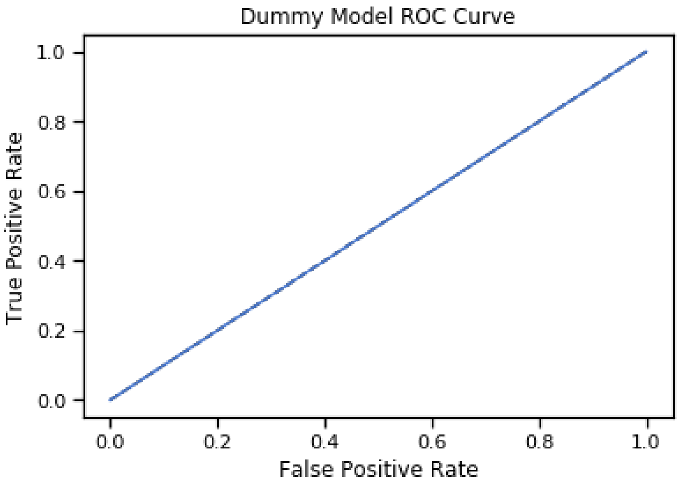
Machine Learning

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

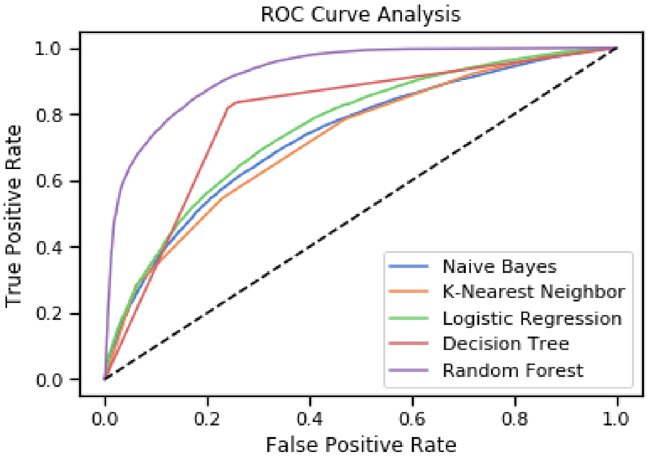
Understanding the performance of a weak model helps with distinguishing strong models. As a baseline, I used a model that randomly predicts whether or not a review has a positive rating. The random model achieved an AUC score of 0.50 and the ROC curve is plotted to the right. This means the model has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

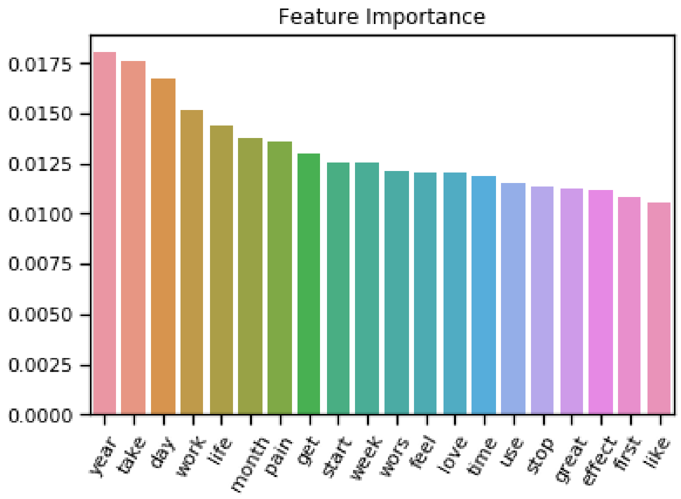
Model Tuning

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.