“Drug/Medicine Review by Patients”

Jiawei Deng

May 6, 2021

Final Report

Abstract

In this project, I am discovering how effective different models can perform in an ordinal categorization/regression problem. For example, amazon have products reviews under each merchandise. For me in this project, I am selecting a drug review dataset to analyze the relationship between reviews and ratings.

Introduction

Objective/Goal

In this project, my objective is to use textual reviews to predict the ratings given by reviewers.

Dataset

The dataset I chose was originally published on “UCI Machine Learning Repository”, the following URL is my source of the dataset.

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

Generally speaking, this dataset contains reviewers or patients’ review information, including:

- UniqueID: their account ID
- drugName: name of drug used in their treatment
- condition: name of their physical conditions/reason of disease
- review: plain text of their reviews towards to their treatment and drug
- rating: how happy their feel towards their treatments (1 - 10)
- date: date they give the review entry
- usefulCount: number of users who found review useful

Statistics

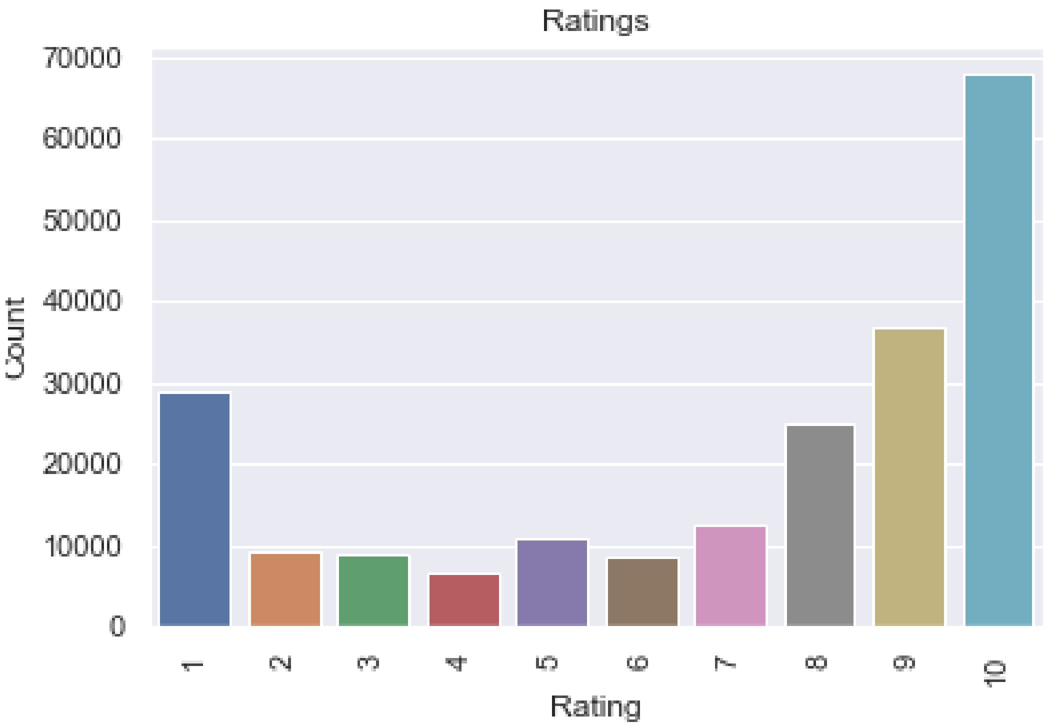
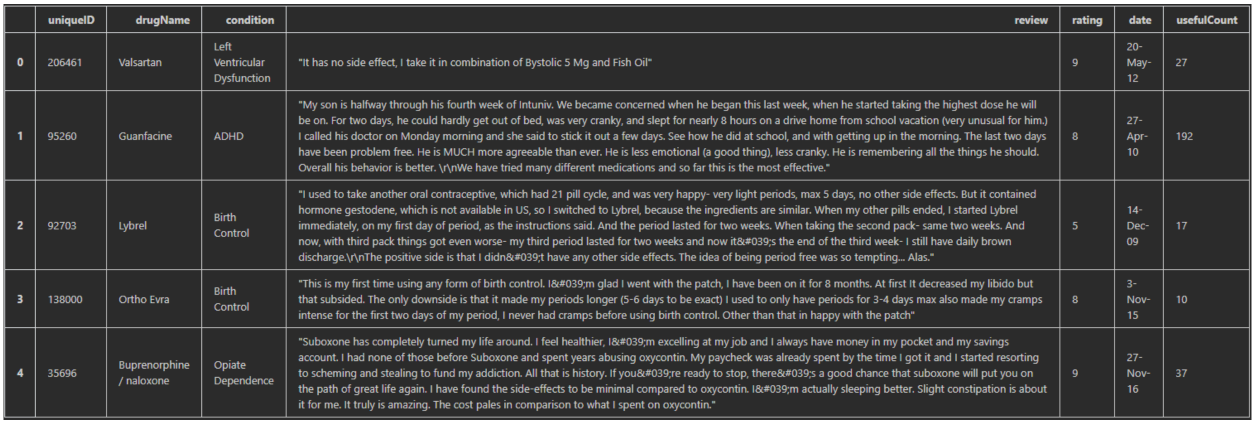


Fig 1. Distribution of Ratings

This is a distribution of the whole dataset of how ratings are distributed, we can find out that the majority of patients tend to give greater or equal to 8 if they feel positive to their treatment. While there are plenty of patients feels negative of the treatment tend to give 1. Ratings between 2 to 7 are far less than the others.

Examples



	uniqueID	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combination of Bystolic 5 Mg and Fish Oil"	9	20-May-12	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of Intuniv. We became concerned when he began this last week, when he started taking the highest dose he will be on. For two days, he could hardly get out of bed, was very cranky, and slept for nearly 8 hours on a drive home from school vacation (very unusual for him.) I called his doctor on Monday morning and she said to stick it out a few days. See how he did at school, and with getting up in the morning. The last two days have been problem free. He is MUCH more agreeable than ever. He is less emotional (a good thing), less cranky. He is remembering all the things he should. Overall his behavior is better. \r\nWe have tried many different medications and so far this is the most effective."	8	27-Apr-10	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, which had 21 pill cycle, and was very happy- very light periods, max 5 days, no other side effects. But it contained hormone gestodene, which is not available in US, so I switched to Lybrel, because the ingredients are similar. When my other pills ended, I started Lybrel immediately, on my first day of period, as the instructions said. And the period lasted for two weeks. When taking the second pack- same two weeks. And now, with third pack things got even worse- my third period lasted for two weeks and now it&#039;s the end of the third week- I still have daily brown discharge.\r\nThe positive side is that I didn&#039;t have any other side effects. The idea of being period free was so tempting... Alas."	5	14-Dec-09	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth control. I&#039;m glad I went with the patch, I have been on it for 8 months. At first It decreased my libido but that subsided. The only downside is that it made my periods longer (5-6 days to be exact) I used to only have periods for 3-4 days max also made my cramps intense for the first two days of my period, I never had cramps before using birth control. Other than that in happy with the patch"	8	3-Nov-15	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around. I feel healthier, I&#039;m excelling at my job and I always have money in my pocket and my savings account. I had none of those before Suboxone and spent years abusing oxycontin. My paycheck was already spent by the time I got it and I started resorting to scheming and stealing to fund my addiction. All that is history. If you&#039;re ready to stop, there&#039;s a good chance that suboxone will put you on the path of great life again. I have found the side-effects to be minimal compared to oxycontin. I&#039;m actually sleeping better. Slight constipation is about it for me. It truly is amazing. The cost pales in comparison to what I spent on oxycontin."	9	27-Nov-16	37

Fig 2. Example of my dataset

Methodologies

Preprocess

- Combined training set and test set into one large dataset.
- Delete URLs from the plain text.
- Remove numbers and punctuations, only alphabetical letters are allowed in text.
- Convert all letters to lowercase.
- Slightly remove stop words (97 stop words).
- Some stemming to assembly words with same stems.

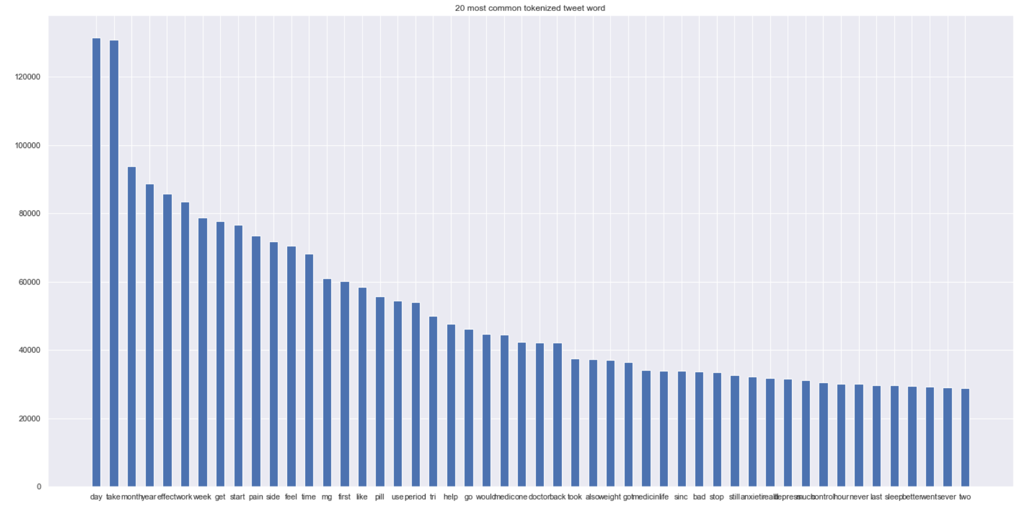


Fig 3. Top 50 Frequent Words After Preprocessing

After cleaning the reviews, I plot a frequency bar plot. From the plot we can see that they are gradually decreasing which means some meaningless high frequency words have been removed and the dataset is well prepared for next step.

Feature Extraction

My objective in this project is to figure out which method can make best connections between reviews and ratings, so I chose different feature extraction methods and compare their results.

- Unigram feature
- TF-IDF (Term Frequency – Inverse Document Frequency)
- Bag of Words (unigram, bigram, trigram, and their combinations)

Feature Selection

In the first, since I need to compare three feature extraction methods, so I used SelectKBest on each of them and set K to 1000 to observe the results. After getting the best performing feature extraction method and algorithm, I am tuning the parameter of K in order to improve the result.

Training Algorithm

I implemented different models to train the dataset, there are linear models, tree models, ensemble methods. In linear methods, I used two methods, logistic regression and ordered logistic regression. Because this is an ordinal regression task so I suppose ordered logistic regression can perform better.

- Linear models:
    - Logistic Regression
    - Ordered Logistic Regression
- Naïve Bayes:
    - Multinomial Naïve Bayes
- Vector Space:
    - SVM (Support Vector Machine)
- Neighbors:
    - KNN (K Nearest Neighbors)
- Ensemble Methods:
    - Random Forest
    - Extra Tree
- Tree Methods:
    - Decision Tree

Results

Models and Evaluation Results

	Unigram	TF-IDF	Bag of Words (1-3)
Naïve Bayes	0.37	0.37	0.37
Ordered Logistic Regression	0.39	0.39	0.39
Random Forest	**0.78**	0.61	**0.78**
SVC	0.41	0.38	0.41
Logistic Regression	0.42	0.32	0.42
KNN	0.34	0.33	0.34
Extra Tree	**0.76**	0.52	**0.76**
Decision Tree	**0.73**	0.29	**0.73**

Table 1. F1 score of each model and feature extraction method

Observations and Explanation

From the models and their evaluation results, random forest, Extra tree, and decision tree are the best performing models. Therefore, the first conclusion I made is that for this ordinal regression task, ensemble and tree methods are the most appropriate and accurate. For linear regression models, ordered logistic regression is not as good as logistic regression. Among all models, KNN has the lowest f1 score.

As unigram features have the same results as bag of words. It can be inferred that the 1000 best performing features are actually all unigram features. Therefore, the best performing feature extraction method is unigram.

Ensemble methods are performing well, I think there are mainly two reasons:

- Ensemble method is always better accuracy and performance than single model.
- Ensemble reduces the spread of dispersion when making predictions.

As for single models, decision tree has the best performance. Unlike linear models, tree models generate multiple branches and learn decision rules by inferring from the dataset.

For KNN, I think the reason why it has the lowest f1 score is because it depends on the relation of words and its neighbors. However, in this task, the connection behind words is not too important since we can approximately infer the probable rating from one or two sentimental words in the text.

Parameters

For all models I tested, I set K = 1000 to choose the 1000 best features from each, and selection method is chi-square.

Improvement strategies

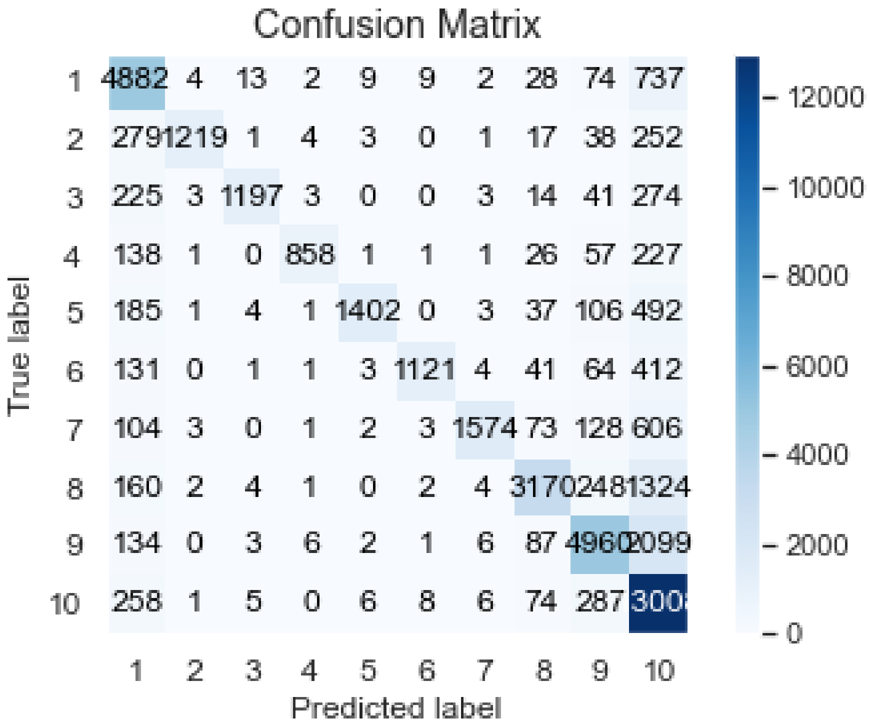
K	F1 Score
100	0.7290586566851882
500	0.7763001883151606
1000	0.7733475925882872
5000	0.7769279055169367

Table 2. F1 score of random forest using unigram features on different K value

By tunning the parameter of K we can see K value range in 500 – 5000 performs pretty well and stable, which means number of features between 500 – 5000 will provide the best results.

Analysis

Error analysis



	1	2	3	4	5	6	7	8	9	10
1	4882	4	13	2	9	9	2	28	74	737
2	279	1219	1	4	3	0	1	17	38	252
3	225	3	1197	3	0	0	3	14	41	274
4	138	1	0	858	1	1	1	26	57	227
5	185	1	4	1	1402	0	3	37	106	492
6	131	0	1	1	3	1121	4	41	64	412
7	104	3	0	1	2	3	1574	73	128	606
8	160	2	4	1	0	2	4	3170	248	1324
9	134	0	3	6	2	1	6	87	4960	2099
10	258	1	5	0	6	8	6	74	287	300

Fig 4. Confusion Matrix for Random Forest (K = 1000, n_estimator = 100, unigram features)

Insights

From the table in Models and Evaluation Results, ordered logistic regression performs not as good as logistic regression which is a little bit unexpected. Ensemble method contains different single models so it should perform the best. For single models, there is not a single model is competitive as decision tree for ordinal regression.

Interpretation

Since this dataset is actually an ordinal dataset, but the ordinal logistic regression didn't perform very well. I think the reason is that the reviews are not very corresponding to the ratings given by reviewers. Unlike sentiment analysis, sentiment analysis has a clearer difference between each sentiment. Because each reviewer will have different scale, so the ratings are more subjective to reviewers. I think this may be the reason why ordered logistic regression is worse than logistic regression.

Another analysis for the confusion matrix for random forest. This is the best performing model I have, and the result is highly acceptable for me. It performs pretty well on rating from 2 to 7, there are some misclassifications on right bottom corner. I think the reason is that for a positive review, the model cannot accurately predict what is the exact rating the reviewer want to give. But the overall mapping looks good.

References

UCI Machine Learning Repository: Drug Review Dataset (Drugs.com) Data Set. (2021). Retrieved 7 May 2021, from https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

Brownlee, J. (2020). Why Use Ensemble Learning?. Retrieved 7 May 2021, from https://machinelearningmastery.com/why-use-ensemble-learning/#:~:text=There%20are%20two%20main%20reasons,the%20predictions%20and%20model%20performance.

S. Baccianella, A. Esuli and F. Sebastiani, "Evaluation Measures for Ordinal Regression," 2009 Ninth International Conference on Intelligent Systems Design and Applications, 2009, pp. 283-287, doi: 10.1109/ISDA.2009.230.