Drugs Review

G2T5
Jose Tan
Neo Jia Ying
Nor Aisyah
Wong Wei Ling

TABLE OF CONTENTS

01

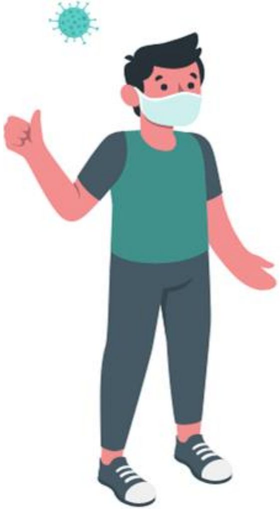
Introduction

Project Title
Problem Statement
Motivation

Project Title

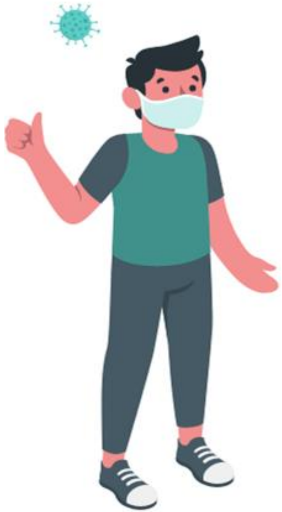
Discovering insights from patient's reviews and recommend most suitable drug using Regression and Classification models

Description



- Drug reviews play a significant role in providing crucial medical care information
- Patients now are more health conscious
- Increasingly using the Internet to gather information in managing their own health
- Looking for stories from patients online, which they might not be able to find among their friends and family.

Problem



- An overwhelming number of over-the-counter drug reviews online.
- Patients have to go through them manually and individually to find the most suitable drug for their condition
- **Highly inefficient** and **time consuming**

Motivation

Aid patients in self prescription of drugs

- Gather insights on patient’s reviews & **recommend the top drugs** based on other patients’ **reviews**, **ratings** and **sentiment scores**
- Improve the **effectiveness** of the review sites and **efficiency** in the time taken to find the best drug.

02

Literature Review

Research Paper 1

Disease Prediction and Drug Recommendation Android Application using Data Mining (Virtual Doctor)

- Predict disease by analyzing user's symptoms using machine learning algorithms (Decision Tree, Naive Bayes, KNN etc.)
- Recommend drugs using weighted average method (Useful count + Rating)
- Top 5 rated drugs will be recommended for each disease in the Android Application

Improvements

- Find out **sentiment score** of drug reviews
- Predict **rating of each review** for test data with regression models (sentiment score, useful count,rating of train data).
    - Strengthen reliability
        - People who have taken the drug and find that it is good will review it positively

Research Paper 2

Detecting Side Effects and Evaluating Effectiveness of Drugs from Customers' Online Reviews using Text Analytics and Data Mining Models

- Classified reviews into meaningful attributes to provide helpful recommendation to users in selecting best drug (Neural Network, Logistic Regression etc)
  - Considers the side effects of a drug
  - Determine if the benefits can outweigh the side effects
  - Compare with similar drugs

Improvements

- All **rounded approach** instead of just looking at side effects
- Take into account:
    - Rating of drug
    - Useful count of review
    - Sentiment score for review
    - Number of users who rated the drug
- Derive an overall score for each drug and recommend to users based on 4 factors.

03

Dataset

Dataset

ID (Numerical)
Index of review entry (Column is renamed during data cleaning)

drugName (Categorical)
Name of drug

condition (Categorical)
Name of medical condition

review (Text)
Patients' review

rating (Numerical)
10 star patient rating

date (Interval)
Date of review entry

usefulCount (Numerical)
Number of users that found review useful

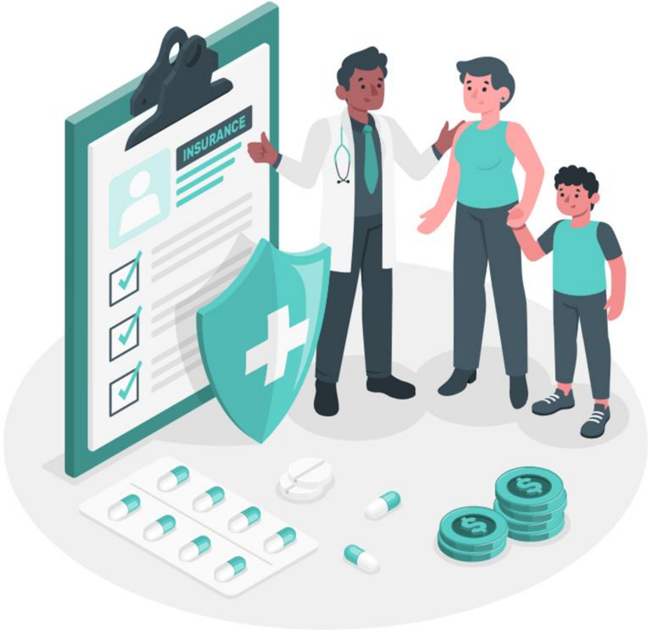
Dataset

	ID	drugName	condition	review	rating	date	usefulCount
0	206461	Valsartan	Left Ventricular Dysfunction	"It has no side effect, I take it in combinati...	9	2012-05-20	27
1	95260	Guanfacine	ADHD	"My son is halfway through his fourth week of ...	8	2010-04-27	192
2	92703	Lybrel	Birth Control	"I used to take another oral contraceptive, wh...	5	2009-12-14	17
3	138000	Ortho Evra	Birth Control	"This is my first time using any form of birth...	8	2015-11-03	10
4	35696	Buprenorphine / naloxone	Opiate Dependence	"Suboxone has completely turned my life around...	9	2016-11-27	37

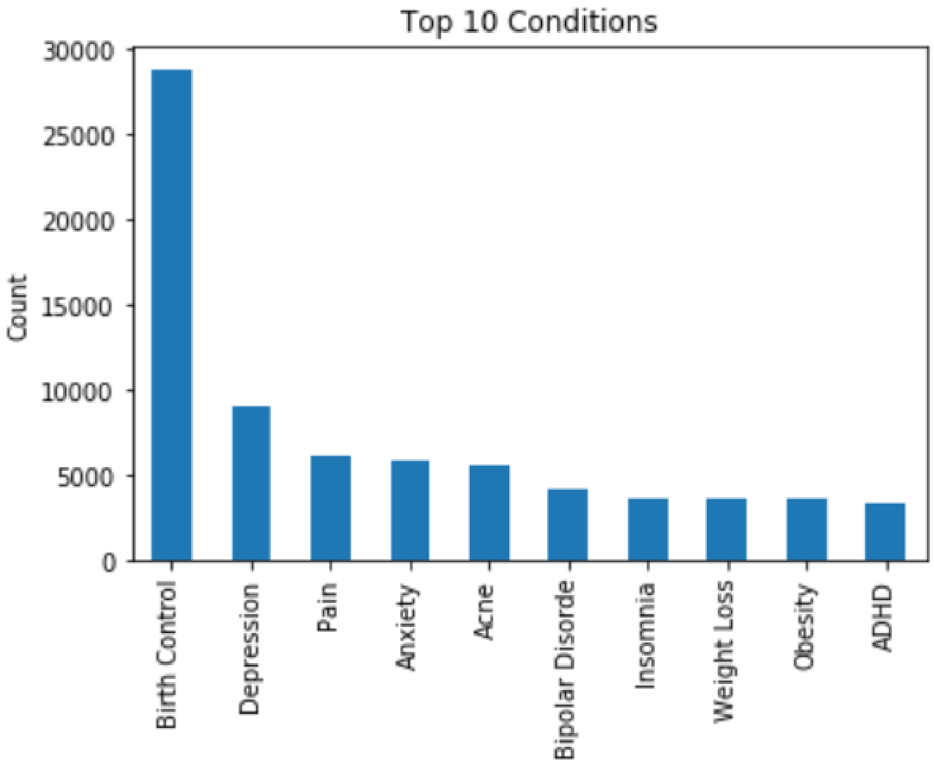
- 2 datasets downloaded from UCI Machine Learning Repository (Train & Test)
- Proportion of test data to train data is approximately 33.33%

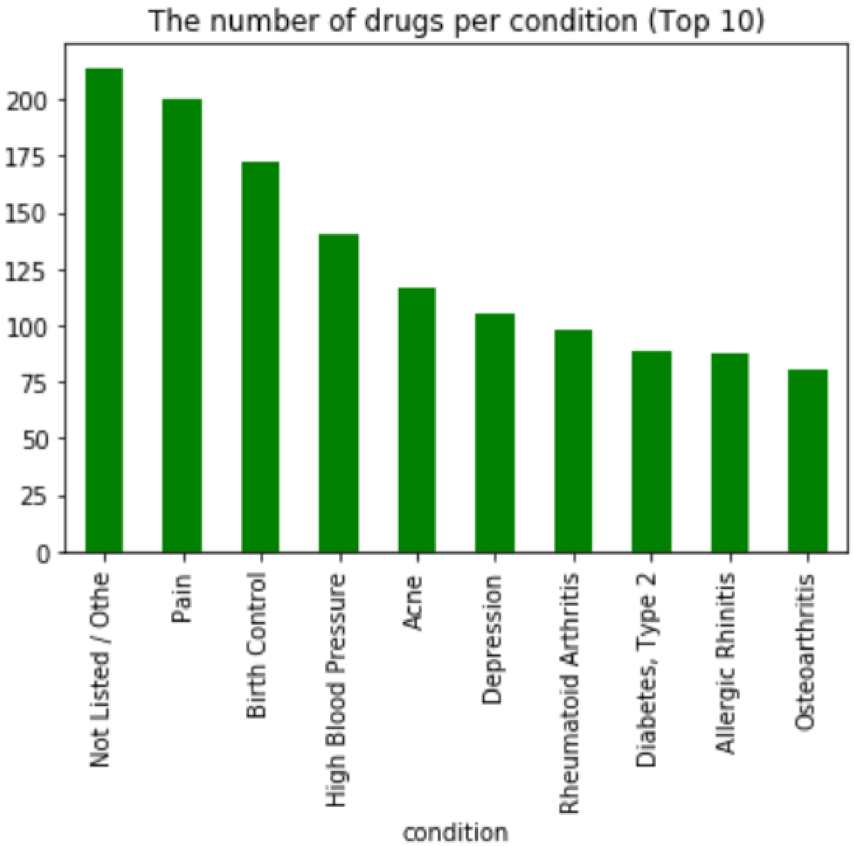
04

EDA

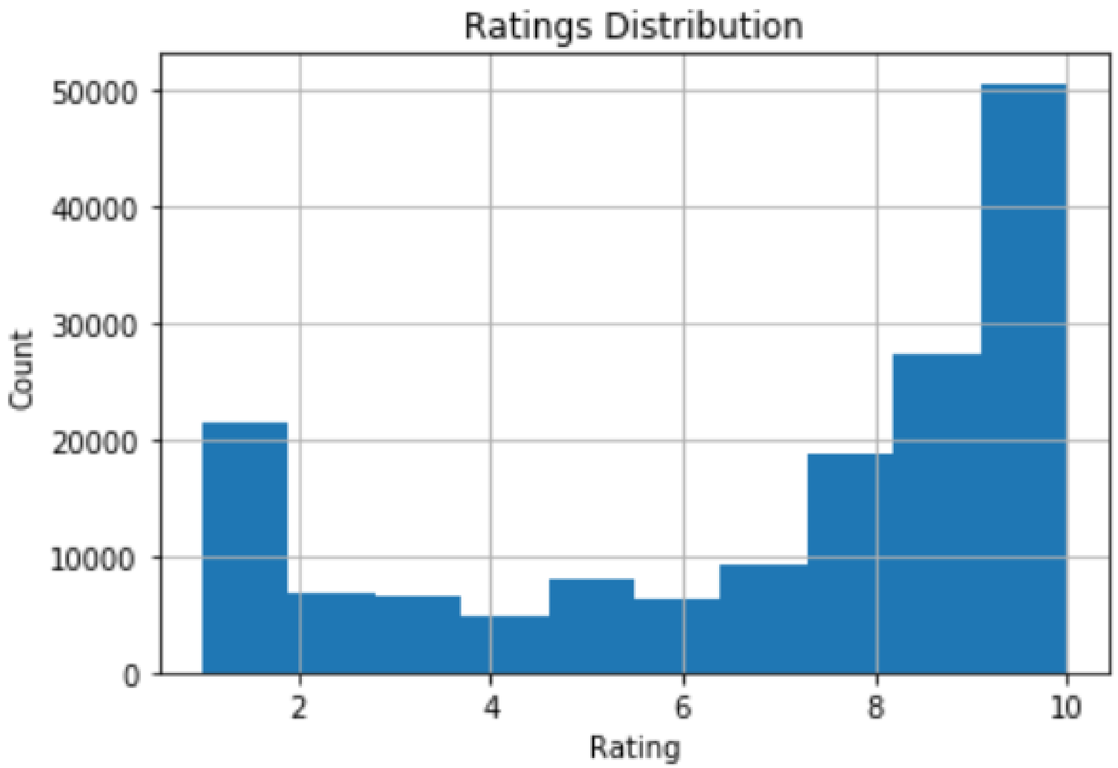


EDA

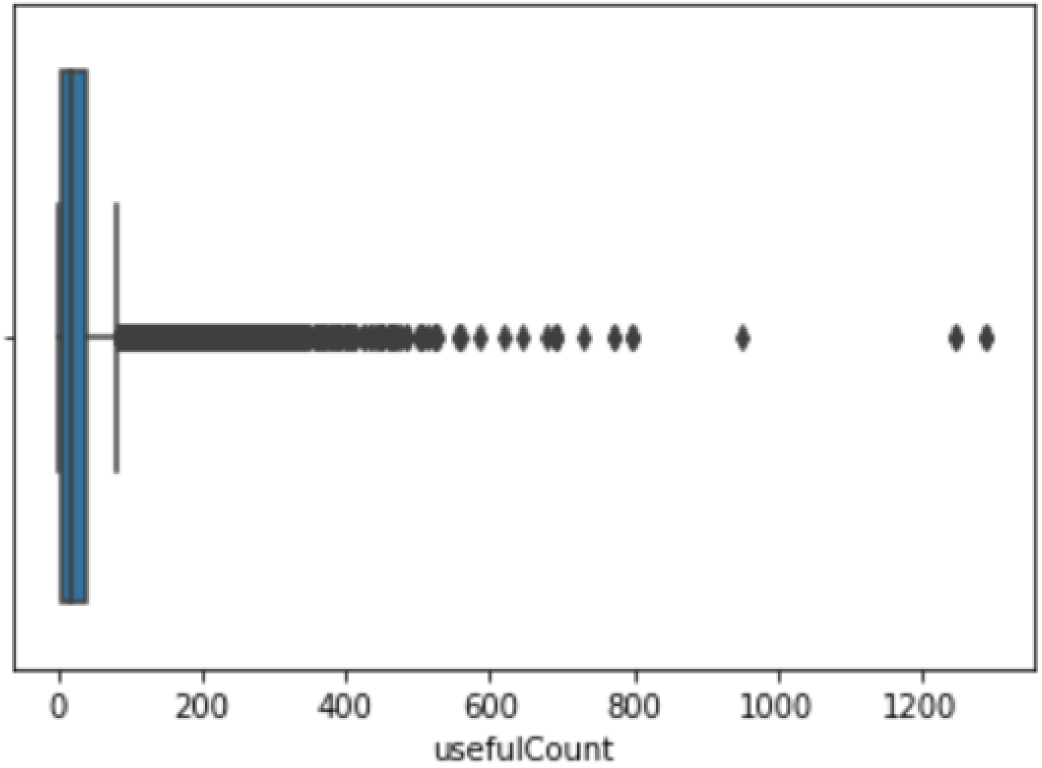

Top 10 Conditions
Count
30000
25000
20000
15000
10000
5000
0
Birth Control
Depression
Pain
Anxiety
Acne
Bipolar Disorde
Insomnia
Weight Loss
Obesity
ADHD


The number of drugs per condition (Top 10)
200
175
150
125
100
75
50
25
0
Not Listed / Othe
Pain
Birth Control
High Blood Pressure
Acne
Depression
Rheumatoid Arthritis
Diabetes, Type 2
Allergic Rhinitis
Osteoarthritis
condition

Skewed 'ratings' distribution



Presence of Outliers

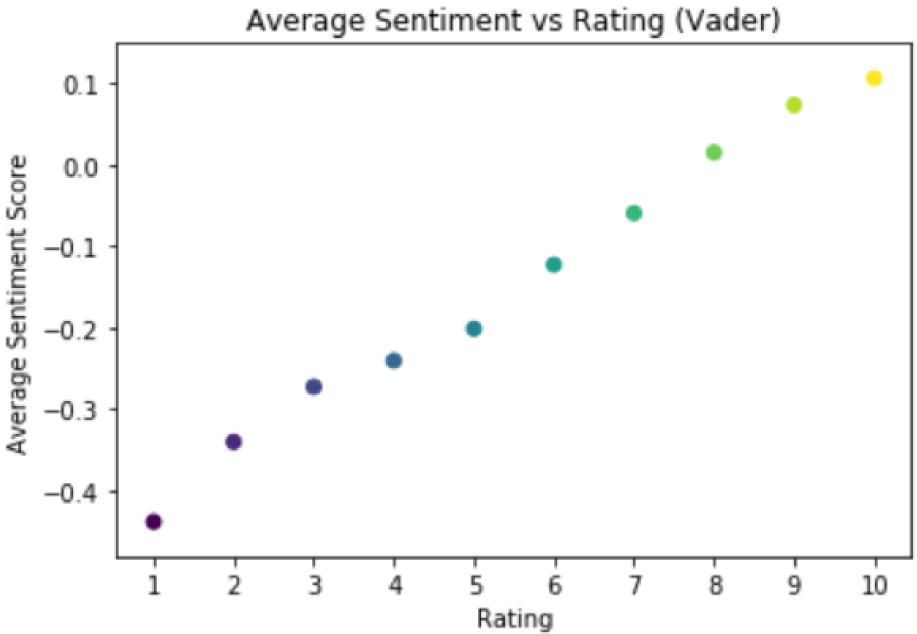


Preliminary Selection of Tool

Predicting Sentiment Score of Review with Vader and TextBlob

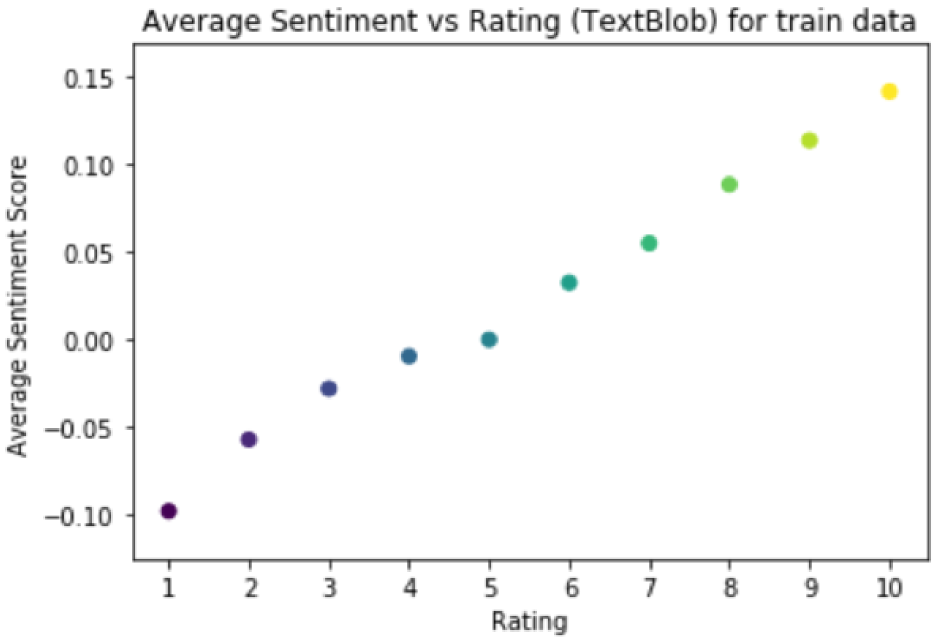
Sentiment	Vader	Textblob
Positive	Score >= 0.05	Score > 0
Neutral	-0.05 < Score < 0.05	Score = 0
Negative	Score <= -0.05	Score < 0

Vader



- Neutral sentiment: ratings 7 & 8
- Positive sentiment: ratings 9 & 10
- Negative sentiment: ratings < 7

TextBlob



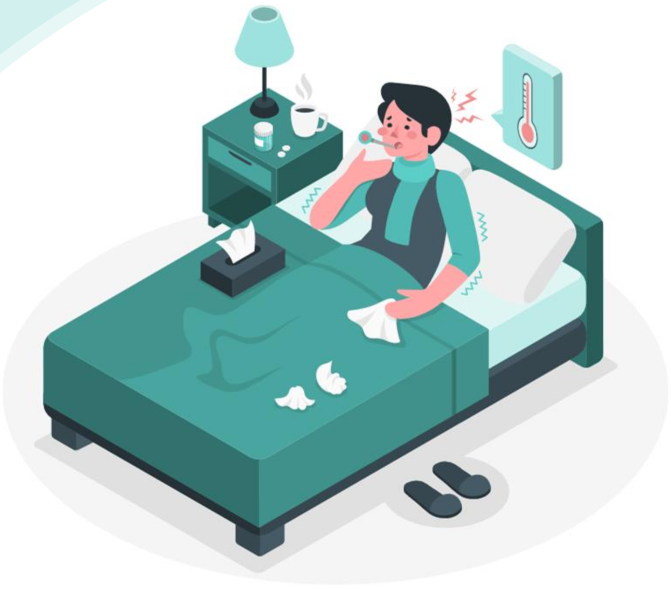
- Neutral sentiment: rating 5
- Positive sentiment: ratings >= 6
- Negative sentiment: ratings < 5

Vader or TextBlob?

- Vader work better with slang, emojis, etc.
- TextBlob performs better with more formal language usage
- Since our reviews are written in more formal language, **TextBlob** will be better.

05

Methodology & Results



Objectives

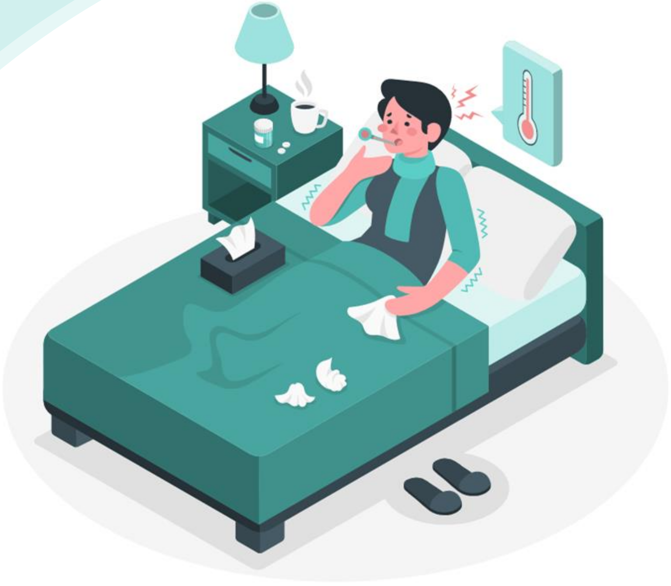
1) *Predict Sentiment Score of Review and Classify the Reviews based on Sentiment*

2) *Predict Ratings of each review on Test Data*

3) *Determine Overall Score of Each Drug to Recommend to Patients*

05a.

Predict Sentiment Score of Review and Classify the Reviews Based on Sentiment



Predict Sentiment Score of Review and Classify the Reviews Based on Sentiment

- From our preliminary results, we used **TextBlob** to predict sentiment scores.
- Used Term Frequency - Inverse Document Frequency (**TF-IDF method)** to convert the raw text into numerical format so that it can be processed by classifiers.
- **6** Classification models used: Logistic Regression, Naive Bayes, Random Forest, Decision Tree, KNN and Adaboost

Predict Sentiment Score of Review and Classify the Reviews Based on Sentiment

Results

- Used **macro average F1-score** to determine the best performing model as our data is imbalanced (13355 vs 37173)

- **Random Forest** has the highest macro average F1 score, thus it is our best performing model

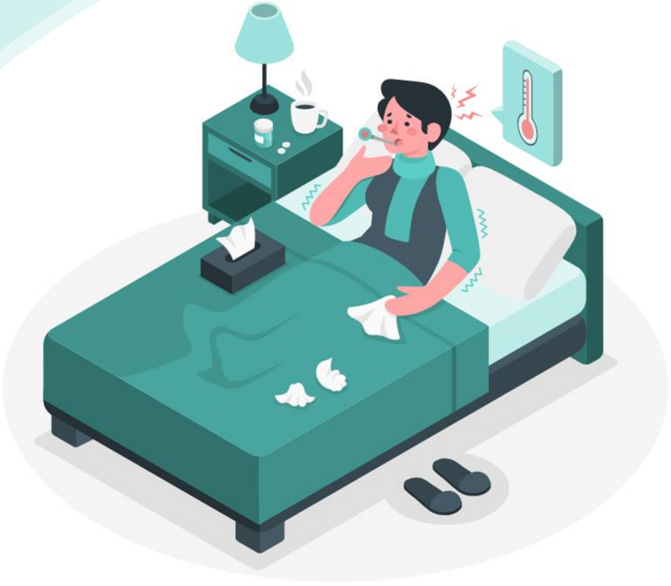
Random Forest

Accuracy: 0.9185797973400887

	precision	recall	f1-score	support
False	0.98	0.71	0.82	13355
True	0.90	0.99	0.95	37173
accuracy			0.92	50528
macro avg	0.94	0.85	0.88	50528
weighted avg	0.92	0.92	0.91	50528

05b.

Predict Ratings of each review on Test Data



Predict Ratings of Each Review on Test Data

- Used **rating of drug**, **useful count** & **sentiment score of reviews** to train regression models
- Scaling is done because useful count and sentiment scores are on different scales
  - Used **RobustScaler** as outliers are present in our dataset
- **5** Regression models used: KNN, Multiple Linear Regression, Decision Tree, Lasso, Ridge and ElasticNet Regression

ID	drugName	condition	review	rating	date	usefulCount	sentiment_textblob	usefulCount_scaled	sentiment_textblob_scaled	rating_scaled
206461	Valsartan	Left Ventricular Dysfunction	"It no side effect, I take combination Bystoli...	9	2012-05-20	27	0.000000	0.354839	-0.223283	0.2

Predict Ratings of Each Review on Test Data

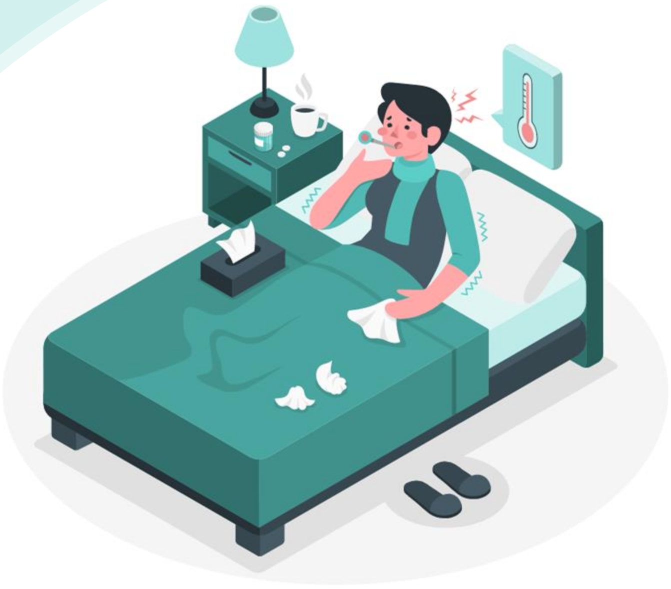
Results

- Used **RMSE** as:
  - It is a good measure for how **accurately** a model predicts the response
  - Accuracy is the most important criterion for fit if the model is used for prediction

```
Comparison of rmse for the 6 models
KNN: 0.5734837441858722
MLR: 0.5022879110257089
Decision Tree: 0.7251606278176802
LASSO: 0.5022619755414564
Ridge: 0.5022877882160374
ElasticNet: 0.5485629829195381
```

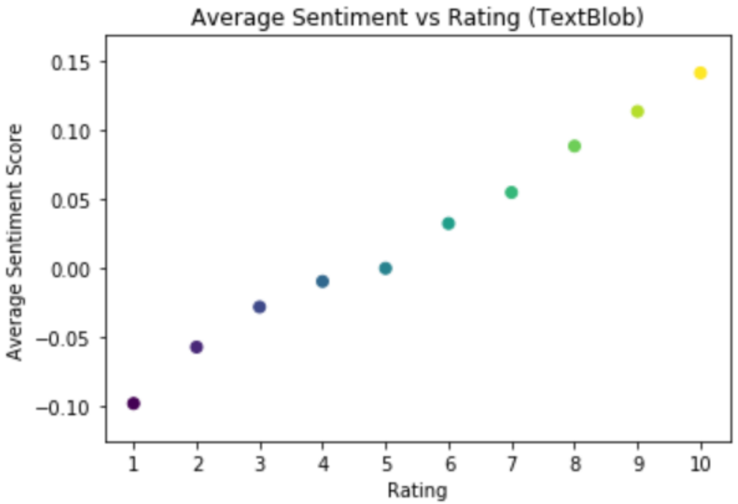
05c.

Determine Overall Score of Each Drug to Recommend to Patients



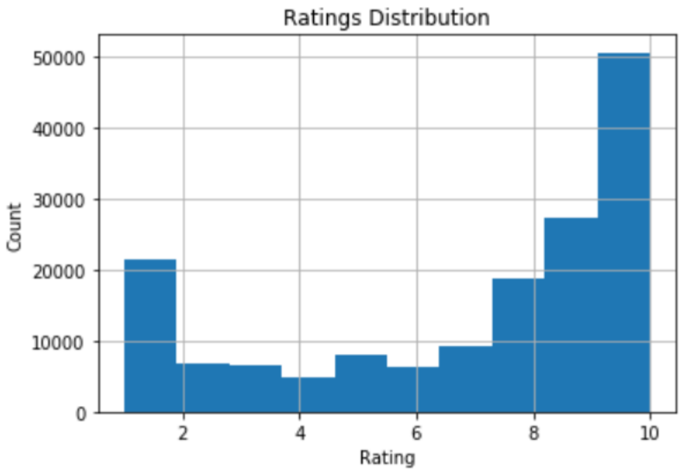
Determine Overall Score of Each Drug to Recommend to Patients

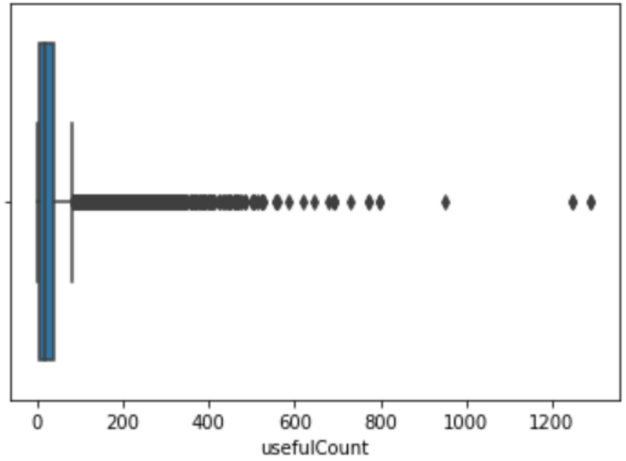
- Initially we wanted to determine overall score of drug with:
  - Rating of drug
  - Useful count of review
  - Sentiment score for review
  - Number of users who rated the drug
- However, sentiment score is **omitted** as it is correlated to rating



Determine Overall Score of Each Drug to Recommend to Patients

- Also, the dataset has **outliers** and **highly skewed**, hence we decided to use:
  - Median rating of drug
  - Median useful counts of reviews





Determine Overall Score of Each Drug to Recommend to Patients

- Final variables used to **determine overall score**:
  - Median rating of drug
  - Median useful count of review
  - Number of users who rated the drug

Equation:

score = x%*(Median usefulCount) + y%*(Number of users who rated the drug) + z%*(Median rating of drug)

Determine Overall Score of Each Drug to Recommend to Patients

- To determine **weightage** of each variables in the equation:
  - Used **ExtraTrees Classifier** to determine **feature importance** of the 3 variables (ratings, usefulCount and no. of users who rated drug)
  - Extra Trees randomly selects split point
    - Allows for lesser correlation between the decision trees in the ensemble

Determine Overall Score of Each Drug to Recommend to Patients

Results

```
feature_impt.describe()
```

	Number of users	Median rating	Median usefulCount
count	5.000000	5.000000	5.000000
mean	0.038632	0.927970	0.033398
std	0.006627	0.010485	0.003958
min	0.030446	0.913440	0.029917
25%	0.034064	0.922586	0.030000
50%	0.039393	0.928251	0.032356
75%	0.041861	0.935936	0.035553
max	0.047395	0.939637	0.039165

In order of importance:

Median rating > Number of users > Median Useful Count

Determine Overall Score of Each Drug to Recommend to Patients

Drug Recommender System

- Users are required to manually type in their condition
- Results of all drugs available for that condition will be returned
- Overall score arranged in descending order.

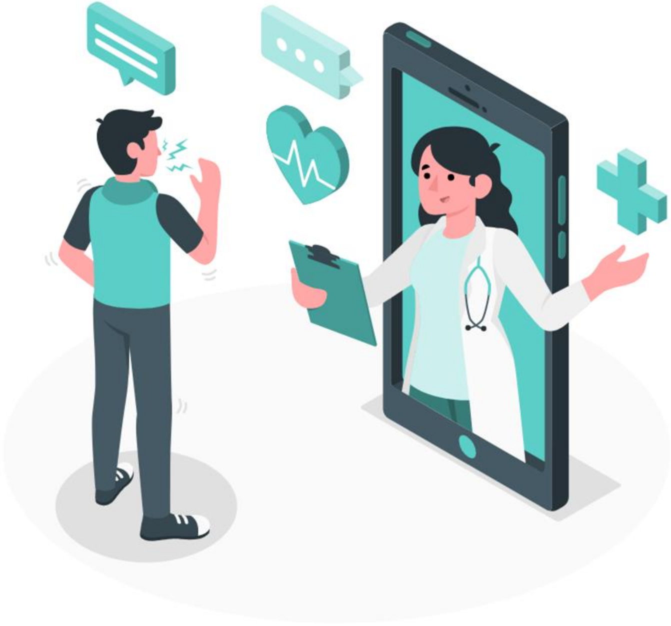
```
Please input your condition: hiv infection
Results:
29 records found
```

drugName	condition	Number of users	Median rating	Median usefulCount	score
Cobicistat / elvitegravir / emtricitabine / te...	HIV Infection	23	10.0	18.0	10.769397
Efavirenz / emtricitabine / tenofovir	HIV Infection	23	10.0	12.0	10.569007
Odefsey	HIV Infection	1	10.0	35.0	10.487275
Abacavir / dolutegravir / lamivudine	HIV Infection	18	10.0	13.5	10.425947
Stribild	HIV Infection	16	10.0	11.0	10.265188
Triumeq	HIV Infection	15	10.0	11.0	10.226556
Emtricitabine / tenofovir	HIV Infection	2	10.0	17.5	9.941435

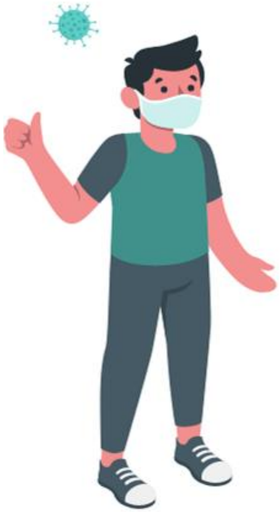
**Higher overall score = more suitable drug*

06

Future Work



Cross Validation



- Use **cross-validation** to evaluate our models
- To **flag problems** such as overfitting since it reserves a sample of the dataset and trains the model using the remaining dataset
- Not done due to time constraints.

SMOTE

- Final results show that we have an **imbalance of positive and negative sentiments**.
- Used **macro average F1-score** to choose best performing model
- After doing further research, **Synthetic Minority Oversampling Technique (SMOTE)** can be used to tackle imbalance data.
- Not done due to time constraints

References

- https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664
- https://machinelearningmastery.com/extra-trees-ensemble-with-python/#:~:text=Unlike%20random%20forest%2C%20which%20uses,a%20split%20point%20at%20random.&text=The%20random%20selection%20of%20split,the%20variance%20of%20the%20algorithm
- https://www.mwsug.org/proceedings/2019/IN/MWSUG-2019-IN-064.pdf
- https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/
- https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29
- https://monkeylearn.com/blog/what-is-tf-idf/
- https://www.researchgate.net/publication/343064584_Disease_Prediction_and_Drug_Recommendation_Android_Application_using_Data_Mining_Virtual_Doctor
- https://towardsdatascience.com/sentiment-analysis-vader-or-textblob-ff25514ac540

Thank you