Capstone I Report: Recommending Pain Relievers

Daniel Y Abebe

1. INTRODUCTION

The purpose of this study is to recommend pain relievers for specific conditions based on the satisfaction level of patients after pain relief medication along with the effectiveness of the medicine and its ease of use for different conditions.

1.1 Problem Statement

Both prescribed and non-prescribed use of pain relievers, as well as rates of opioid-related mortality and admissions to emergency departments, have increased in the last few years. There is no exact figure on how many people are affected by this problem but the number has increased according to 'National Center for Biotechnology Information'. There are many different pain relievers, and each one has advantages and risks. Some types of pain respond better to certain medicines than others. Each person has a different response to a pain reliever for the same type of condition. This is mainly because the conditions that cause pain are very complicated, which is one of the reasons why pain management is so difficult. This difficulty decreases the quality of life of patients that, in many cases, can suffer from constant pain with little to no relief. General-purpose pain-relieving medication is often ineffective even the best painkiller like morphine can sometimes fail to inhibit pain in cancer patients. Painkillers are working by activating a receptor cell call REV_ERBs that sends chemical signals inside the cell to block the production of certain genes regulates pain-causing and inflammatory molecules inside the body. The goal of this project is to build a model that can predict painkiller selected conditions and for different age groups.

1.2 Stakeholders

The possible stakeholders of this project includes pharmaceutical companies, physicians and clinicians, insurance companies and patients. Knowing the consumers' response for the product is crucial in many majors. In addition to the available research and development in the field will help physicians and clinicians while subscribing to these drugs. Pharmaceutical companies and insurance companies will also identify their focus drugs for their consumers.

1.3 Description of Dataset

The dataset used in this project was acquired from kaggle where uploaded on kaggle by scraping the WebMD site

(https://www.kaggle.com/rohanharode07/webmd-drug-reviews-dataset). The dataset provides user reviews on specific drugs along with related conditions, side effects, age, sex, and ratings reflecting overall patient satisfaction. The dataset contains 12 features with mixed variables including categorical, numerical and date. There are above 360 thousands of rows of unique reviews and are updated till Mar 2020.

2. DATA WRANGLING

The inspiration of this dataset was intended to answer following questions:

- Identifying the condition of the patient based on drug reviews?
- How to predict drug rating based on patient's reviews?
- How to visualize drug rating, kind of drugs, types of conditions a patient can have, sentiments based on reviews

Which is not exactly what we are about to do, but not too far from the author's intention. Hence there is a little more data wrangling work here following almost all of the six core activities of data wrangling.

Table 1 The first five observations before the data is cleaned

	Age	Condition	Date	Drug	DrugId	EaseofUse	Effectiveness	Reviews	Satisfaction	Sex	Sides	UsefulCount
0	75 or over	Stuffy Nose	9/21/2014	25dph-7.5peh	146724	5	5	I'm a retired physician and of all the meds I ...	5	Male	Drowsiness, dizziness , dry mouth /nose/thro...	0
1	25-34	Cold Symptoms	1/13/2011	25dph-7.5peh	146724	5	5	cleared me right up even with my throat hurtin...	5	Female	Drowsiness, dizziness , dry mouth /nose/thro...	1
2	65-74	Other	7/16/2012	warfarin (bulk) 100 % powder	144731	2	3	why did my PTINR go from a normal of 2.5 to ov...	3	Female	NaN	0
3	75 or over	Other	9/23/2010	warfarin (bulk) 100 % powder	144731	2	2	FALLING AND DON'T REALISE IT	1	Female	NaN	0
4	35-44	Other	1/6/2009	warfarin (bulk) 100 % powder	144731	1	1	My grandfather was prescribed this medication ...	1	Male	NaN	1

I. **Discovering**: looking at the table 1, there are 12 features in the dataset but not all are useful for the intended plan. 'Date' and 'DrugId' features are not important so I will get rid of them. Since we are not intended to do sentiment analysis, 'Review' and 'Sides' features are used. The last feature that is considered not impactful is 'UsefulCount', hence it is dropped. There are over 1800 unique conditions in the dataset ranging from birth control to depression. Since our focus is only on drugs that are used for pain reliefs, we select conditions that are only related to pain and deal with it. The remaining dataset contains 53649 rows which is about 14.79% of the total dataset rows.

II. **Structuring:** The shape and order of the data doesn't need restructured.

III. **Cleaning:** Data cleaning starts with missing values. Pandas provides isnull(), isna() functions to detect missing values. Both of them do the same thing. Using

isna().sum() returns the number of missing values in each column. Looking at the output of isna().sum() bellow, Age and Sex has missing value. Drop these null values using dropna() function.

The second issue is managing the outliers. There are different methods of finding the outliers including via visualization such as boxplot and scatter plot and through mathematical functions such as Z-Score and interquartile range (IQR). But in our case; it can be detected simply by looking at the maximum and minimum values since the numerical features are fixed lower and upper bounds. The describe() function computes a summary of statistics pertaining to the DataFrame columns. Using this function, we can see whether the values are out of range and we couldn't find any so there are no outliers.

```
#Check if null values exist in the dataset
print('Do null values exist in the dataset? ' + str(df_data.isnull().values.any()))
#if null values exist calculate the total null vaues in each column:
df_data.isna().sum()
```

```
Do null values exist in the dataset? True
Age             1905
Condition          0
Drug               0
EaseofUse          0
Effectiveness      0
Rating             0
Sex             3698
dtype: int64
```

Figure 1. Missing value counts with the corresponding feature

IV. **Enriching:** we don't need to drive data from existing data. However, the predictors under feature 'Condition' are 50 unique variables. It is a better idea to have fewer predictor variables rather than having many of them; which leads to redundancy/irrelevance, overfitting, productivity and understability. For this project we chose one condition with the corresponding top five frequently used drugs selected among 606 target variables.

V. **Validating:** by taking a deeper look at the data values to make sure they make sense statistically using describe() function to the correct business context.

VI. **Publishing:** at this point our dataset is well structured and cleaned and ready to use for analytics processes.

3. EXPLORATORY DATA ANALYSIS

We conducted the data exploration on the cleaned data. EDA is the process of performing initial investigations on data so as to discover patterns to spot anomalies, to test hypotheses and to

check assumptions with the help of summary statistics and graphical representations. We start exploring from frequency counts and distribution of patient population.

Frequency counts and population distribution

Gender: to see the gender proportion of the patient we constructed a pie chart. As shown in the chart, The majority of the patients (not necessarily the patients but the patients who rate the drug after they utilize it) are females with 61% and 39% of males.

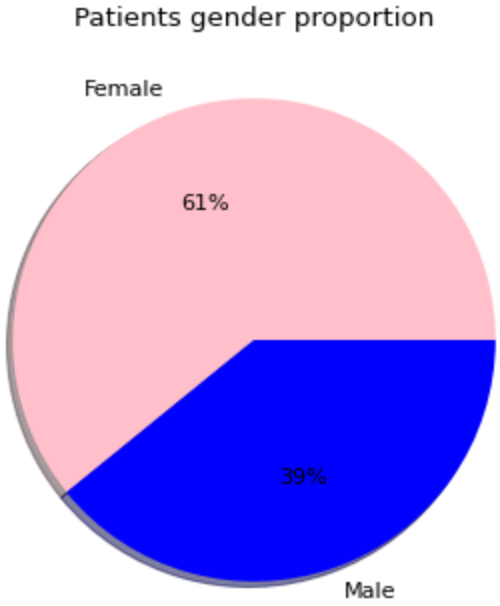


Figure 2. Pie chart for patients' gender proportion

Age: the age distribution looks uniform with a skewness to the left. Most patients are middle aged between 40-60. When we further divide the age distribution based on the gender, the distribution remains the same except the variation in number of counts.

EaseofUse, Effectiveness, Satisfaction: we plotted the frequency of each of these features to see the trend of rating across age groups. Looking at the plots in Fig 4, only a few drugs not are easy to use. The same is true for effectiveness as well even though the number of ineffective drugs are not as small as ease of use. We noted an interesting thing in the satisfaction trend that most people are either lenient or harsh raters in both genders.

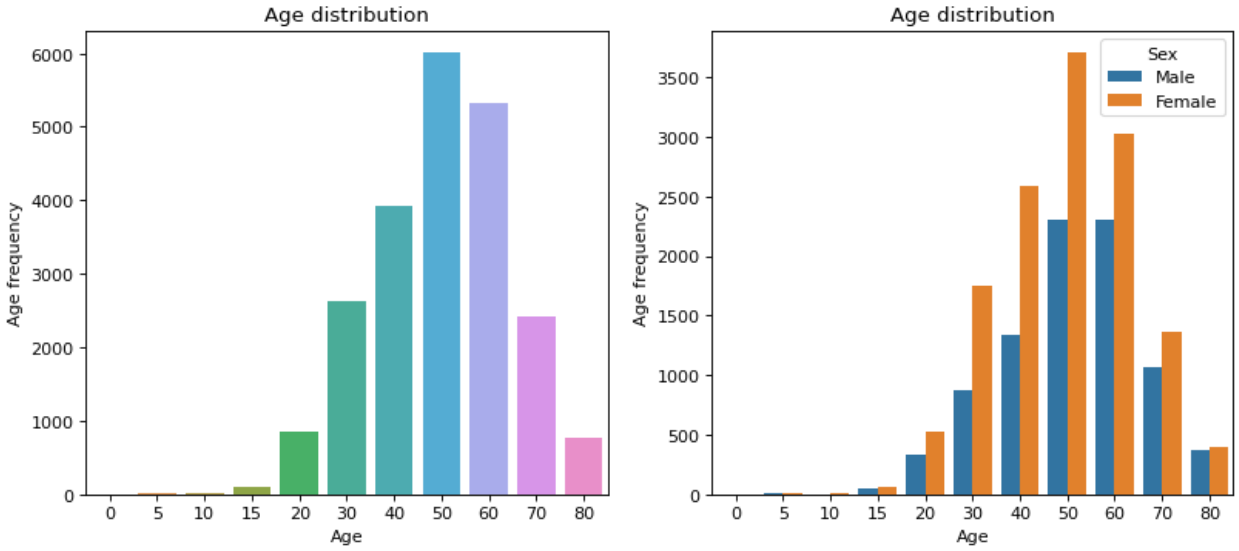


Figure 3. Comparison of age distribution

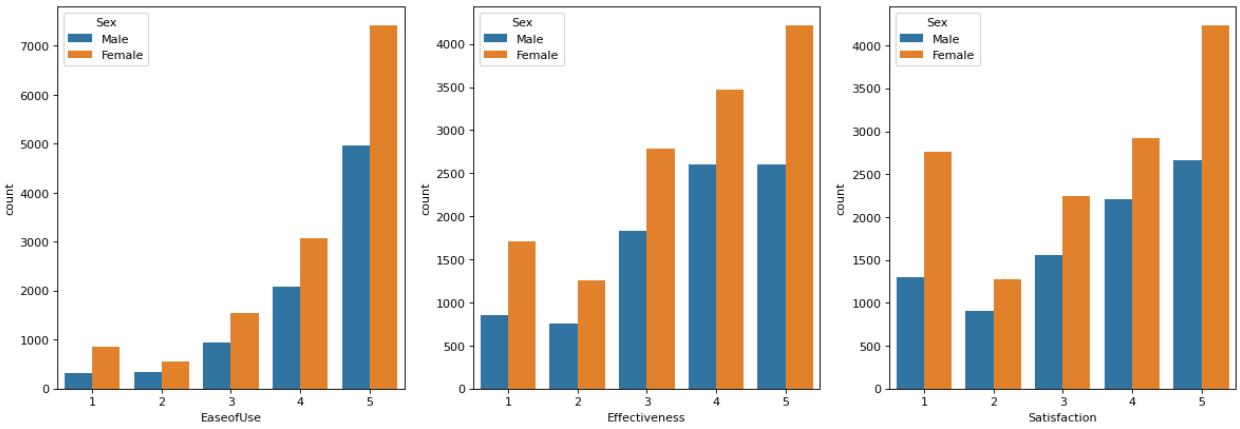


Figure 4. Comparison of EaseofUse, Effectiveness and Satisfaction

Further observation on EaseofUse, Effectiveness and Satisfaction

In order to find out the relationship between these three features, we calculated their mean and plotted with respect to Age. As shown in the plot, Fig. 5, the mean value of effectiveness and satisfaction showed strong association compared to the mean value of EaseofUse and Satisfaction except for the age group 0-10. In addition, the mean of satisfaction level is less than both the mean of effectiveness and ease of use. The mean of EaseofUse is greater than the mean of Effectiveness and Satisfaction.

Correlation Matrix before and after balanced dataset

Fig. 6 presents the correlation matrix heat map plotted before balancing the dataset. As shown in the figure, there is a positive correlation between EaseofUse, Effectiveness and Satisfaction. The correlation between Effectiveness and Satisfaction is stronger with correlation value of 0.82 compared to EaseofUse and Satisfaction correlation with correlation value of 0.53 which proves the plot on the mean of these three features in Fig 5. The age of the patients has negative correlation with effectiveness and satisfaction with values -0.048 and -0.034 respectively. Surprisingly, the correlation between age and ease-of-use is positive with value 0.031. Looking at these values of Age and (EaseofUse,Effectiveness and Satisfaction) correlation, they are very close to zero which is an indication that these rating features are independent of the age of the patients.

The frequency of our target clases (the drugs) are not balanced (oxycodone: 7412, hydrocodone: 4681, tramadol: 4020, gabapentin:3369, neurontin: 2592). By taking the approximate median frequency which is hydrocodone class: 4681, we down sampled the major class oxycodone class and upsampled minor classes tramadol, gabapentin and neurontin classes. The correlation after resampling (balanced dataset) is shown in Fig. 7. The correlation between EaseofUse, Effectiveness and Satisfaction remain the same but the correlation between age and other numerical features changes a bit. The correlation between the age of the patients with effectiveness and satisfaction after balancing the classes are -0.017 and 0.0016 respectively. The recent correlation between age and ease-of-use is 0.036.

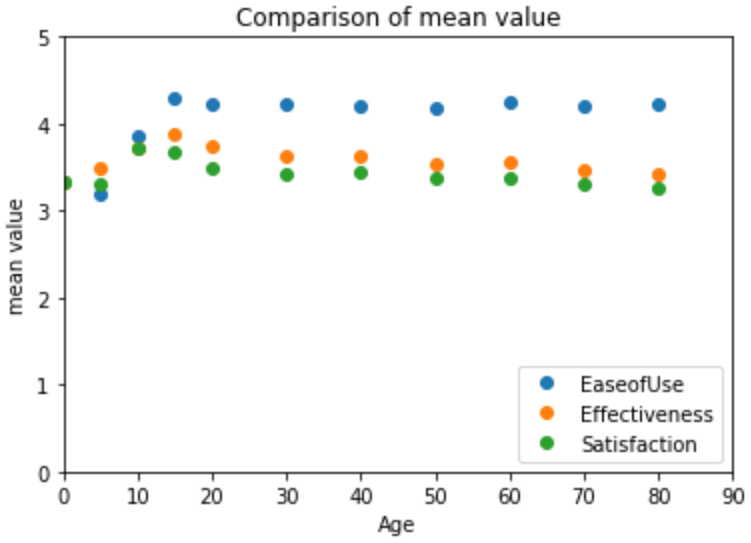


Figure 5. Comparison of mean value of EaseofUse, Effectiveness and Satisfaction

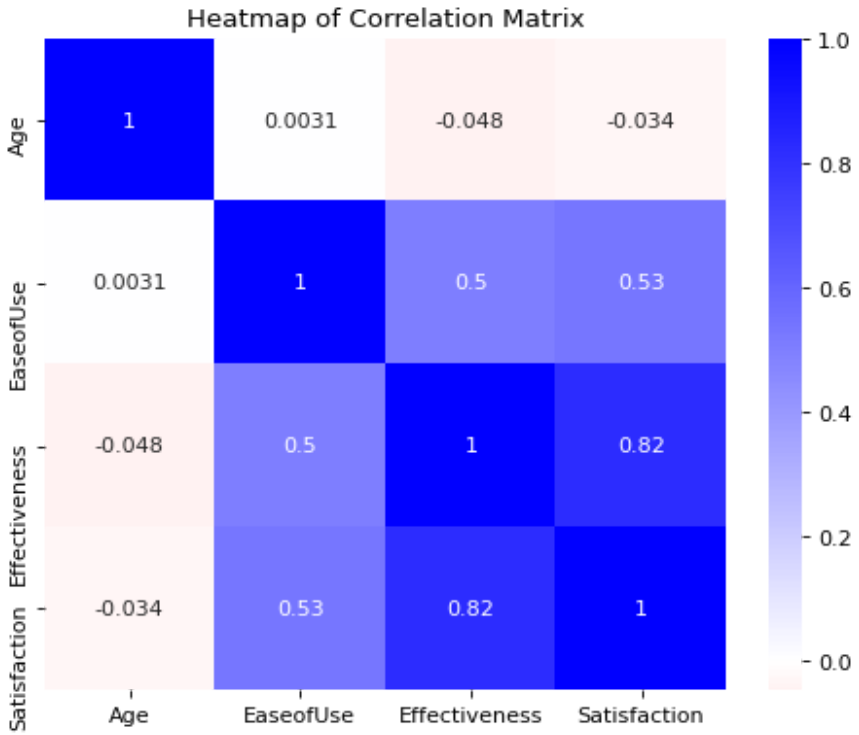


Figure 6. Correlation matrix heatmap plot before balancing the data

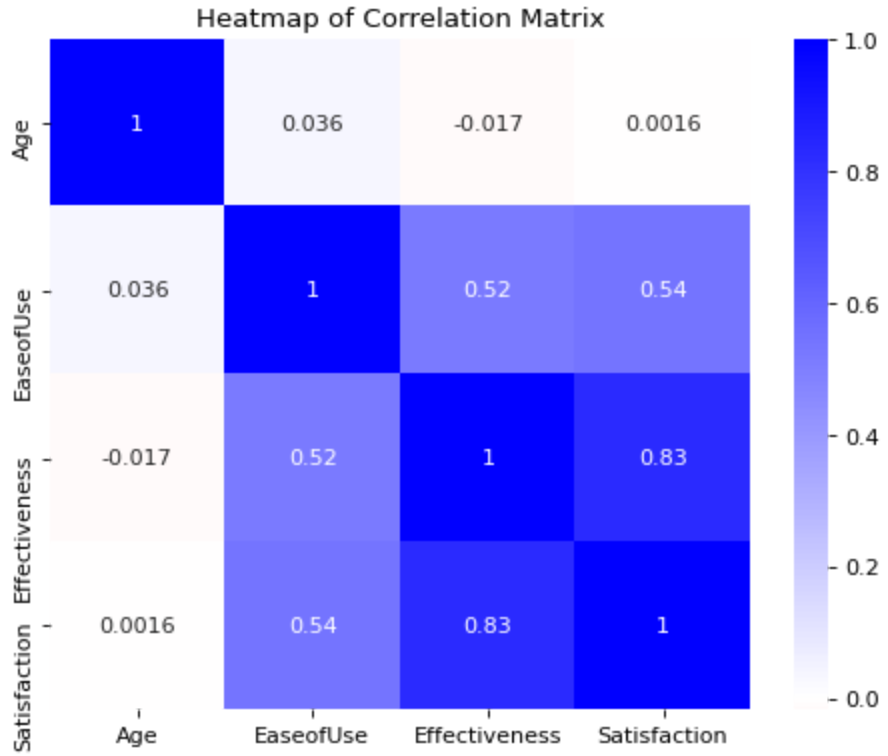


Figure 7. Correlation matrix heatmap plot after balancing the data

Statistical Test

In order to further find the relationship between features by raising basic questions such as:

1. Is there a significant difference between the patient's response towards EaseofUse and Satisfaction?
2. Is there a significant difference between the patient's response towards Effectiveness and Satisfaction?
3. Is there a significant difference between the patient's Satisfaction depending on their age?

All the statistical test evaluated using paired sample t-test. The paired sample t-test is a univariate test that tests for a significant difference between 2 related variables. For instance: for the last question we proposed the null hypothesis as:
there is no difference in patients' Satisfaction depending on their age. And the alternative hypothesis is: there is a difference in patients' Satisfaction depending on their age. After analysis we failed to accept the null hypothesis as the calculated p_value for 95% confidence interval is less than 0.05.

Multicollinearity

Multicollinearity refers to predictors that are correlated with other predictors. Multicollinearity occurs when a model includes multiple factors that are correlated not just to the response variable, but also to each other. In other words, it results when we have factors that are a bit redundant. Sklearn does not have a built-in way to check for multicollinearity. Looking at the result of the variance inflation factor, all numerical features have high multicolinearity as shown in Table 2. However, the VIF of Effectiveness is highest, so we will remove it from the model.

Table 2 variance inflation factor result

	VIF	Features
0	7.277050	Age
1	15.196056	EaseofUse
2	23.901570	Effectiveness
3	18.684840	Satisfaction

4. MACHINE LEARNING

Create dummy variable

To include the categorical data in the regression, we create dummy variable encoding on nominal features using a very 'convenient method called: get_dummies' which does that seemlessly. It is extremely important that we drop one of the dummies, alternatively will introduce multicollinearity.

Split input features and output label

Before partitioning the input features, we standardize the features by scaling the features. This helps the features to centered around 0 with a standard deviation of 1 is not only important if we are comparing measurements that have different units.The output variable is categorical so we encoded using LabelEncoder() function. Then using split_train_test function partition original data into training and test sets for training with size of 75% and 25% respectively. We used stratify to split the parameter so that the proportion of values in the sample produced will be the same as the proportion of values provided to the parameter.

Data Modeling

We modeled the data using different classification: Decision Tree, Random Forest, Gradient Boosting, XGBoosting, Extra tree and KNeighbors classifiers and the best model performance out of six will be selected.

Tune Hyperparameters for Classification

Machine learning algorithms have hyperparameters that allow us to tailor the behavior of the algorithm to our specific dataset. In order to find out the best hyperparameters, we modeled cross-validation on the training dataset and create hyperparameter options then selected the best hyperparameter.

Result and Discussion

It is necessary to quantify the strength of the relationship between the predictors and the outcome. Ranking the predictors depending on this manner is important especially for large amounts of data. Looking back at our exploratory data analysis there are no irrelevant features in our dataset, since our data is not large. But we computed important features using features importance modes for selected models, as shown in Fig. 8, to see the most impactful features. In most cases the feature ‘Satisfaction’ is of high importance and the gender of the patient is of low importance in predicting the outcome.

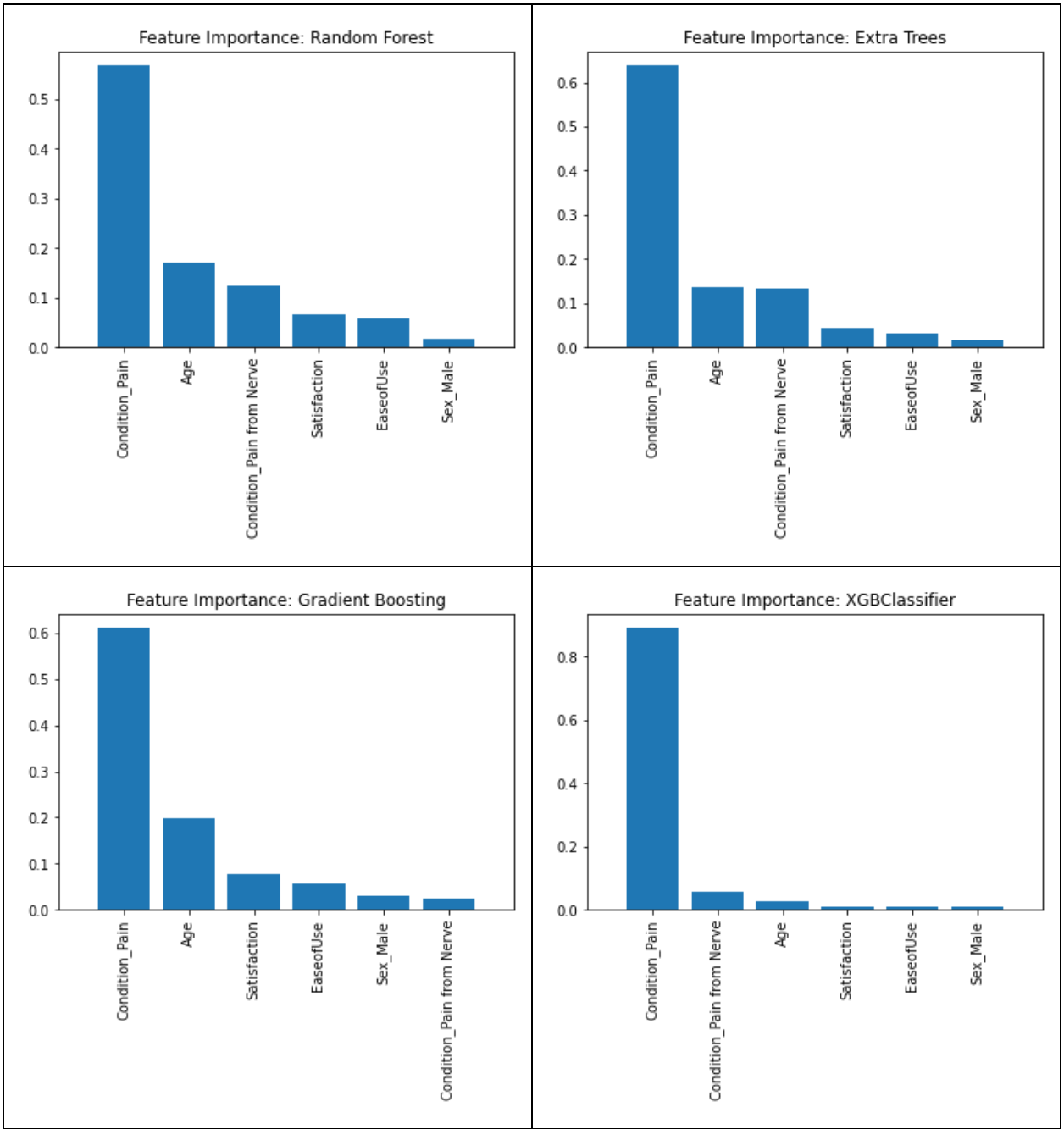


Figure 8. Feature Importance of selected models

Evaluation of Models

Accuracy score: The following table summarizes the accuracy score results of the models having five classes. The accuracy of Random forest, XGBoost, Extra Tree and Bagging 0.54 which is bigger than Logistic Regression, Support Vector Machine, Decision

Tree, Gradient Boosting, Gradient Boosting and KNeighbors. Hence, Random forest, XGBoost, Extra Tree and Bagging performs well in predicting pain relief with accuracy of 0.54 and ROC score of 0.712 for five drugs (5 classes).

Confusion Matrix: the other method to gauge the performance of the classification model and confusion matrix shows where the model went wrong and offers us guidance to correct our path. Looking at the confusion matrix plot shown in Fig. 9, the true positive above 0.20 is in acceptable range since our classes are five and dividing 100% by 5 is 20% or normalized 0.2. The true positive values of each class are different for instance 'gabapentin' has high true positives than the rest in all models. The last two classes 'oxycodone' and 'tramadol' have the lowest true positive.

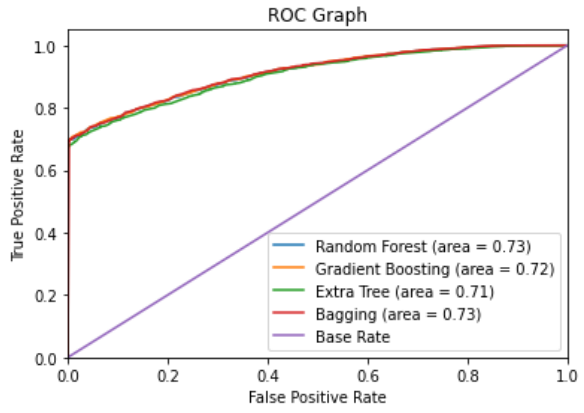
The possible reason for this imbalanced result output is that there is a big difference in the number of instances for each age group. A small difference often doesn't matter i.e small imbalance is not only common but it is expected. However, in our dataset, some age groups have no records at all for the last two classes.
Random Forest and XGBoost perform the same.

Table 3 Evaluations summary of models

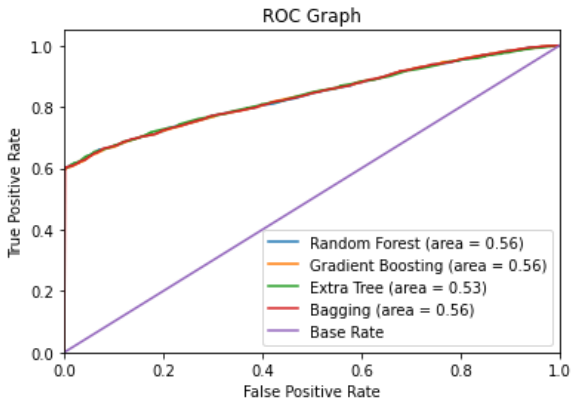
Model	Accuracy Score	ROC Score
Logistic Regression	0.435	0.647
Support Vector Machine	0.500	0.687
Decision Tree	0.438	0.649
Random Forest	0.538	0.711
Gradient Boosting	0.485	0.678
XGBoost	0.539	0.712
Extra Tree	0.539	0.712
KNeighbors	0.526	0.704
Bagging	0.539	0.712

Figure 9. Comparison of confusion matrix of different models

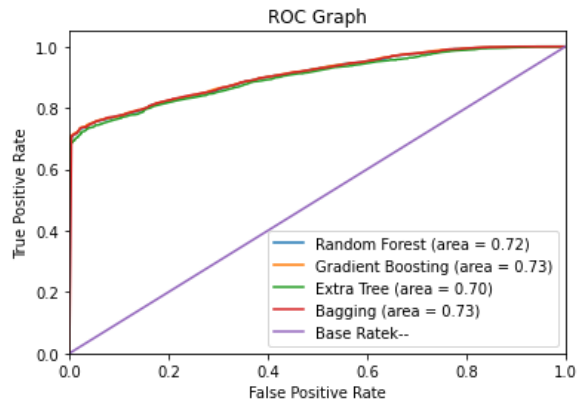
ROC Curve: The AUC-ROC curve is another performance measurement for classification problems at various thresholds setti ngs. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much a model is capable of distinguishing between classes.



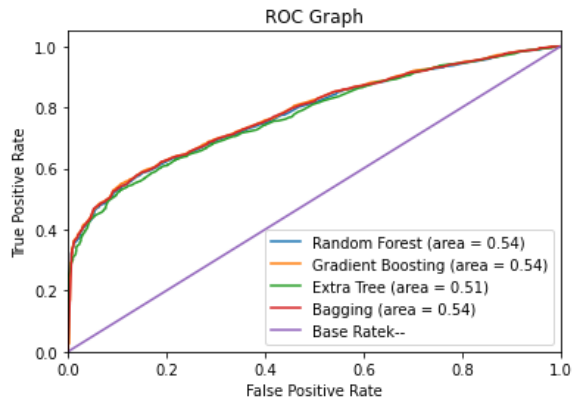
a. oxycodone



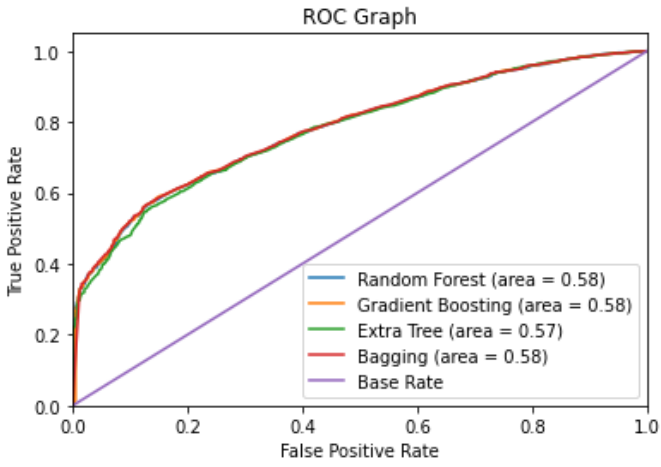
b. hydrocodone



c. neurontin



d. gabapentin



e. tramadol

Figure 10. ROC curve of each classes

Summary

Looking at the comparison of model performance and feature importance plots, we are able to find important findings on recommending pain relievers. The first thing one can note here is that the ‘Condition_pain’ is the most important feature in recommending the pain relievers and ‘Age’ of the patient is the second most influential feature. It is obvious that the rank of feature importance varies from model to model. However, the rank of importance was different from model to model. Among the top four models used in predicting pain relievers, Random Forest, Extra Tree, Gradient Boosting the second most important feature is ‘Age’ but for XGBoost, it is the other kind of condition.

To grab the best model out of all models we used majority voting, and Random forest, SGBoost, Extra Tree and Bagging have the equal accuracy and ROC score which is 0.539 and 0.712 respectively. The maximum AUC from the plotted ROC graph is 0.73 for the first two classes and 0.58 for the last three classes. 0.73 AUC means 73% of time the model predicts correctly for the specific condition and selected age and other requirements.

Recommendations

The accuracy and performance of the model is recommending the pain relievers considering the condition and the age of the patient is fair but not superb. One of the main reasons is the limitation of an unbalanced dataset and small size of dataset. The unbalance is higher for some of the features which are high impact in model performance like the age of the patient. Hence, we need to collect more data to train the model which the author believes will recommend pain relievers for patients having different conditions and age groups.