EFFECTIVENESS OF STD DRUGS (HACKEREARTH ML CHALLENGE)

Approach:

- ➔ Use effectiveness rating, number of times the drug was used and a combination of reviews, drug name and use case to predict the base score
- ➔ Use the combination of review, drug name and drug use case to find a sentiment score for both training and testing data
- ➔ Clean the text using NLP techniques, build up a vocabulary
- ➔ Use word2Vec mappings to generate weight matrices using the embeddings
- ➔ Fetch the features into a Bidirectional LSTM model (built using Keras and Tensorflow backend) to predict the sentiment score
- ➔ Use the generated sentiment score along with effectiveness rating, number of times the drug was used and other generated features into the Gradient Boosting model to predict the base score
- ➔ Training score -> 0.999, Testing score -> 0.916

Feature Engineering

- ➔ Studied the general behaviour of the data by plotting various graphs
- ➔ Divided the date approved by UIC into 3 new features -> day, month and year
- ➔ Created classes using the months to identify the pattern diseases and drug used in specific months. Created 4 classes, one each for Winter, Spring, Summer and Autumn
- ➔ Cleaned the review text by patients using NLP techniques, removed stopwords and then studied the behaviour by finding the 1-gram, 2-gram and 3-grams.
- ➔ Came up with a feature of sentiment score using Name, Use case and review of the drug to get a numerical field to fit the Regressor
- ➔ Tried using Label encoding and One hot encoding to encode the drug names and use cases -> not so effective in this case

Tools Used

- ➔ Numpy
- ➔ Pandas
- ➔ Sklearn
- ➔ Matplotlib and Seaborn (for data visualization)
- ➔ NLTK
- ➔ WordCloud
- ➔ Keras with Tensorflow backend
- ➔ Bi-LSTM
- ➔ Word2Vec embeddings
- ➔ GradientBoostingRegressor

Results:

- ➔ Training Score -> **0.9999**
- ➔ Testing Score -> **0.916**