Prediction and Analysis of Psychiatric Medication Effectiveness

By: Kaitlin Bustos

Introduction:

According to Salzer, Brusilovskiy, and Townley (2018) only one third of those with serious mental illnesses have reported being in remission for at least one year. This indicates that 70% of those with mental illnesses are not recovering from their received diagnoses. If an individual undergoes treatment which does not work and therefore is ineffective it may consequently lead to one's psychiatric disorder going untreated. These individuals who have not been adequately treated are in the majority producing this low recovery rate. As pharmaceutical treatments are popular among the various treatment options for individuals suffering from mental illnesses it may be useful to begin to investigate treatment patterns associated with psychiatric medication. Employing data science techniques within the field of mental health can reveal patient patterns that otherwise go unnoticed and lead to more predictable medical outcomes. Determining factors contributing to psychiatric medication ineffectiveness and predicting pharmaceutical treatment outcomes for those with psychiatric disorders may give rise to a methodological approach to treatment recommendations within mental health, and potentially increase the low recovery rate associated with mental illnesses.

Data:

The source of the data was the ICPSR which is a shared data platform encompassing various data archives from the social and behavioral sciences. The data used in my analysis was a subset of variables from the Collaborative Psychiatric Epidemiology Survey (2001-2003).

The data was in .csv format and the data points represented an individuals survey responses. Columns were label encoded and a codebook was provided.

Factors involved in the analysis were individual characteristics taken from survey screening responses and medication details taken from survey responses regarding the use and effects of ones prescribed psychiatric medication. Medication success was the target of the analysis.

Cleaning and EDA:

Missing values were converted into NaN’s and rows were removed where medication success (target) and medication ID were NaN. Rows and columns with greater than 80% missing values were removed. The remaining missing values were label encoded to represent “unknown” or filled with the median where it was appropriate. Columns were renamed respectively to each

variable's appropriate label and it was ensured columns contained the correct data type (float or int). Medication effectiveness was converted into a binary categorical variable where ‘1’ was considered successful if it was rated very effective by an individual, otherwise it was rated unsuccessful as ‘0’.

The distribution of successful vs unsuccessful medication was explored to determine if the target had any class imbalance. A function was created and figures were produced for a given column to determine the likelihood of medication being successful or unsuccessful in order to gauge any interesting trends relating to medication success. To measure correlation between variables a correlation heatmap was produced.

Modelling:

Each categorical variable that was originally label encoded and had no inherent order was converted into dummy variables to ensure models interpret the data correctly. Since medication ID had over 200 categories within the column, the top 20 most frequently mentioned medications were considered for dummy variables. The factors and target were set as X and y respectively and the data was split 80/20 into train and test sets. A logistic regression model was fit and scored on both the train and test set. The top 20 most predictive factors of medication being successful/unsuccessful were determined by logistic regression coefficients. In order to optimize logistic regression hyperparameters 5 fold cross validation was performed on the C parameter. A logistic regression with the C parameter producing the highest cross validation score was fit and scored on the train and test set of the data. A confusion matrix was constructed from the logistic regression model predictions on the test set.

To allow for comparison between machine learning models a random forest model was fit and scored on the train and test sets of X and y. Hyperparameter optimization was carried out by performing 5 fold cross validation on the number of estimators and tree depth parameters. With the highest scored parameters from cross validation a second random forest model was fit and scored on the test and train data. A confusion matrix was constructed from the random forest model predictions on the test set. The random forest model was the selected predictive model since the initial confusion matrix yielded more promising results than the logistic regression confusion matrix. Though since the random forest model confusion matrix still had a low precision score and the cost of false positives are high within the context of my analysis; the y threshold was adjusted to increase the precision score which consequently decreased the F1 score. The results are shown in the table below.

Model	Train Accuracy	Test Accuracy	F1 Score
Logistic Regression	65%	65%	0.75
Random Forest	70%	65%	0.70

Findings and Conclusions:

It was determined from the analysis that 40% of individuals who are prescribed psychiatric medication do not find their medication to be very effective which suggests practitioners may be prescribing pharmaceutical treatments that are not successful and consequently their patient may fail to recover from their mental illness.

From the logistic regression coefficients several top factors contributing to medication ineffectiveness relate to depression such as: i) If the reason for taking medication is sadness, ii) Med type: Zoloft, iii) Med type: Celexa, iv) Med type: Wellbutrin. These antidepressants have differing mechanisms of actions versus Doxepin which was considered in the analysis and did not show up alongside the other three antidepressants which may suggest certain antidepressants may be more effective compared to others. A Chi Squared test was carried out to determine if there was a relationship between taking antidepressants versus other medication and medication success, this relationship was considered significant, p=.006.

Among the top predictors of medications being unsuccessful is if an individual experiences difficulties carrying out daily activities due to their condition. A Chi Squared test was employed to determine if there is a relationship between the amount of days (> 15) one is limited in daily activities due to their condition and medication success. This relationship was considered significant, p= .005. This may suggest pharmaceutical treatments may be less effective the greater debilitation one experiences due to their mental illness.

The random forest model selected as the predictive model determines with 70% precision and recall those individuals who will likely have psychiatric medication be successful and those who will likely not have psychiatric medication be successful.

In a clinical setting, utilizing the predictive model can better gauge patient prognosis and may avoid practitioners admistersting pharmaceutical treatments that are not likely to be successful for a given individual. In the future the notion of predicting patient prognosis can be extended across other treatment options such as therapy. Further research investigating mood disorders may enhance the understanding of what is necessary for pharmaceutical treatments to be more successful for mental illness' such as depression. These directions are ultimately steps toward potentially increasing the low recovery rate associated with mental illnesses.