Predict and Prevent Mood Disorders: Project Write Up

Michelle Brier, Vincent Chen, Flore Perillat, Sophia Ouyang, Romain Kakko-Chiloff

Project Overview and Problem Addressed

Depression is one of the top three leading causes of disability in industrialized countries, imposing an annual toll on U.S. businesses that amounts to billions of dollars in lost productivity, medical expenditures, and other costs and contributing to over two-thirds of reported suicides in the U.S. every year. What can we do to predict and prevent such a major mental health issue affecting millions of people around the world? Specifically, how can data science help with this issue?

Over the course of the semester, our team worked with Professor Roberto Zicari from the University of Frankfurt in Germany and Professor Maurice Ohayon from the Stanford Center for Sleep Sciences and Medicine to address this issue. Professor Zicari provided access to electronic medical records from Geisinger, a physician-led healthcare system, with clinical data on over 185,000 patients from Pennsylvania and throughout the semester both mentors offered advice on what to focus on as well as medical journals and papers to use as resources.

After initial research on the profiles of depressed patients and typical overlapping mood disorders, we narrowed our focus to analyzing the intersection between depression and two other mood disorders, insomnia and anxiety, within the dataset we had access to, seeking to address the following points:

1. What are the profiles of patients with depression, patients with sleep disorder, and patients with anxiety?
2. How has the patient profile evolved from the first to last clinical encounter?
3. What are the correlates of depression in terms of social habits?
4. Is there any overlap in correlates of depression and correlates of sleep disorder and/or anxiety disorder, i.e. do people with insomnia display similar social histories as people with depression, and do people with anxiety display similar social histories as people with depression?
5. Can we develop predictive models for depression based demographic features and whether or not a patient has anxiety or a sleep disorder?

Process

After receiving the dataset, we reviewed the attached data dictionary to get an idea of existing features and features we would need to engineer ourselves, then preprocessed the data using three main files that were relevant to our focus: Demographics, Problem List (a list of diagnoses), and Social History.

Demographics was the dataset containing all the patients that have/had mood disorders. From this dataset, we focuses on the id of the patients, their date of birth, their race, and their sex.

Social History was the dataset containing the information that patients gave at their clinical encounters, including their habits on smoking, drug consumption and alcohol drinking. The difficulty we faced while working on this dataset was that one patient can have answered many times these questions and that his/her answers evolve with time. Thus, we focused on the first appointment, the last appointment, and the most given answer.

Problem List included lists of the mental health issues faced by each patient as well as whether each problem was resolved or not. To find the patients diagnosed with depression, insomnia, and anxiety, we looked at the icd10 codes, which is a universal medical classification that gives one code for each type of problem: F32 for depression symptoms, F41 for anxiety, and G47 for sleep disorders. Using regular expression and Natural Language Processing, we were able to identify patients with these disorders and whether or not their disorders were ever resolved.

In the next preprocessing stage of the project, we merged features from these three datasets based on patient IDs, creating one main dataset for our analyses. We then created new data features indicating whether the patients' disorders were resolved, the duration of treatment, and whether patients were diagnosed with both mental and sleep disorders. Finally, we labelled all our categorical features with numbers, regrouping patients with same characteristics.

To address our five main questions, we analyzed the merged dataset for demographic insights on the general patient population as well as insights on specific subgroups of patients: those with depression, those with anxiety, and those with sleep disorder. We then searched for correlation in medical and social histories between these patient subgroups as well as correlation between specific social habits and depression, anxiety, and sleep disorder. For all our analyses, we created data visualizations. Finally, we generated models for predicting depression using logistic regression, kNN, Perceptron, XG Boost, and Random Forest classification.

Findings and Solution

General profiles and statistics

In analyzing the distribution of time it took to resolve respective disorders for the three subgroups of patients (looking only at patients with resolved cases), we found that it took the longest to resolve depression. Out of the patients diagnosed with depression, the resolve rate for depression was 11.8%, compared to a 15.2% resolve rate for patients diagnosed with sleep disorder and a 5.5% resolve rate for patients diagnosed with anxiety. 35.7% of patients diagnosed with depression were also diagnosed with anxiety, and 24.2% of patients diagnosed with depression were also diagnosed with sleep disorder. 10.7% of patients with depression had both anxiety and sleep disorder, although 50.8% of depressed patients had neither.

How has the patient profile evolved from first to last encounter?

Looking specifically at the social habits recorded in the dataset (smoker status (light/heavy/non-smoker), alcohol consumption (y/n), use of illegal drugs (y/n), and whether or not the patient is sexually active (y/n)), we found no change between first and last clinical encounter for all patients. There was no difference in amount of change between first and last appointment within subgroups. Our conclusion from this was that social habits for all mental disorders were unaffected by clinical treatment.

Correlates of depression, sleep disorder, and anxiety

Again, looking at the social habits recorded in the dataset, we searched for correlation between particular habits and each of these three mental disorders. For patients diagnosed with depression, we found that a slightly higher percentage use illegal drugs and heavily smoke compared to non-depressed patients. For patients diagnosed with sleep disorder, we found that a higher percentage consume alcohol compared to patients without sleep disorder. For patients diagnosed with anxiety, we found that a higher percentage consume alcohol and are sexually active compared to patients without anxiety disorder.

From these results, we concluded that there were no strong correlations between depression, sleep disorder, and anxiety when looking at smoker status, use of illegal drugs, alcohol consumption, and sexual activity. People with depression do not necessarily have similar social histories as people with sleep disorder or people with anxiety. There was also no strong correlation between depression and these four social habits.

Predicting Depression

In developing our predictive models for depression (logistic regression, kNN, Perceptron, XG Boost, and Random Forest), we first looked at the results of each model with only demographic features, which had accuracies around 49% - 55%, indicating that demographics (gender, age, and ethnicity) do not strongly predict depression. For models with sleep disorder diagnosis and these demographic features, we again saw low accuracies around 49% - 55%, indicating that also looking at whether a person has a sleep disorder does not improve depression prediction accuracy. For models with anxiety disorder diagnosis and demographic features, we saw much higher accuracies around 63% - 73%, indicating that anxiety may be a better predictor of depression than sleep disorders. Indeed, as most of the time anxiety comes before depression, we decided to work on anxiety as an input of our model.

Finally, when factoring in gender, age, ethnicity, smoker status, drug consumption, alcohol consumption, sleep disorder diagnoses, and anxiety diagnoses, our models had accuracies from 68% - 72% (logistic regression: 72.3% , kNN: 68.05%, Perceptron: 69.63%, XG Boost: 72.49%, and Random Forest: 72.36%). Including smoker status, drug consumption, alcohol consumption, and sleep disorder diagnosis in addition to anxiety diagnosis and demographics improved the accuracy of kNN specifically from 63% to 68.05%, though it did not have much effect on the other four models. From this, we concluded that anxiety is a better predictor of depression than a sleep disorder, and drug consumption, alcohol consumption, and smoker status do not improve depression prediction accuracy.

Value of Solution

Overall, our project solution provides a better understanding of depression and its relationship with sleep disorder and anxiety in addition to an understanding of the social habit correlates of each disease. Finally, our solution includes useful predictive models that show which factors are more important than others when predicting depression.

Future Efforts

This project could have a been at least a year-long project with the huge amount of data we received. Among the thirteen datasets that we could have worked on, we only used 3 of them (Demographics, Social History, and Problem List) for our focus in this project. In the future, it is very possible to deepen our work by incorporating the other datasets to take into account additional factors when analyzing depression, sleep disorders, and insomnia. Some of the most promising datasets from our preliminary analyses are the Medication Records dataset and the list of Encounters that patients had; however, working on these datasets would require a knowledge of Time Series Analysis. These particular datasets include a list of the medicines prescribed by doctors to patients and a list of all clinical encounters the patients had. Potential focus questions that could be answered with these additional files include: *"When somebody takes a medicine, for how long he/she is taking it? Is it compliant with the prescription or is it auto-prescription? How have each patient's symptoms evolved over time? If a patient is diagnosed with both depression and sleep disorder, which kinds of medicine does he/she take at first (antidepressants, hypnotic, or both)? Does this change over time?"*

There are also different types of antidepressants: stimulant and sedative. It is known that if a patient takes a stimulant, it is more likely that he begins to take additional hypnotic medicines, which are used to treat sleep disorders, later. Including information about medication and change over timespan of encounters would allow us to address these additional questions as well as possibly provide more insight for the questions we focused on in this project. As adding data to our analysis can only bring more insights to our study, working on medicines and encounters and using the other datasets could only be a good extension of our project.

Furthermore, we did not focus as much on the origins and consequences of each disease. Thus, it could be a good idea to add a study on the causal links between the diseases we worked on. We could also add features to take into account different degrees of depression. Not all of the patients diagnosed with depression have the same symptoms. Working on patients labeled by the degree of their depression symptoms could improve our understanding of the causalities between the diseases.

Contributions of Each Group Member

Data Selection, preprocessing, and labelling: Romain

Data Visualizations:

Study on time repartition and social habit correlate visualizations: Flore

Study on first appointments: Sophia

Study on last appointments: Michelle

Study on deceased/alive patients: Vincent

General demographic statistics for the three focus disorders and distributions of time to resolve each disorder: Vincent

Predictive models for depression: Sophia, Romain

Write-Up/Presentation Slides: Michelle, Romain

Our Mentor, Roberto Zicari

We had a great experience working with Roberto Zicari. He was always available and gave great feedback on our work with insights on where we should begin and the direction we should head. Moreover, he gave us to access the Geisinger dataset and introduced us to Professor Maurice Ohayon at Stanford so that we could get specific feedback on the medical conclusions we came to.

However, we received these datasets a long time after the beginning of the project so we did not have as much time for data analysis and had to keep our list of main focus questions shorter than originally planned.