Project Design

Topic Selection - Data

For this project, I wanted to find data that was large and had a discrete variable to predict. I was interested in mental health outcomes, so I sought datasets that might let me answer one of a few questions: what factors tend to predict diagnosis for mental health problems? Which characteristics indicate which diagnoses? Who tends to be the main clientele for mental health services? I found a few survey datasets that had information I could use to answer similar queries, so I started by choosing the dataset that had the largest sample and variety of questions – the National Survey on Drug Use and Health (2016). From there, I worked to define a few questions I sought to answer. My initial project focused on the question: from the general populace, who tends to utilize mental health resources?

As time went on and I came closer to presenting a deliverable, I felt uncomfortable and unhappy with what this information could be utilized for. A thought that came to mind was changing insurance pricing for groups based on this modeling, so I instead changed the scope of my project. From there I sought to answer a different question: from the subset of the population who are depressed, which groups actually make use of mental health services? This question was actionable with positive ramifications: we could see what populations are routinely underutilizing mental health resources, and potentially work on policies/campaigns to raise awareness and increase access for those groups.

Data Cleaning and Manipulation

After downloading and importing my data, I needed to do a lot of work to get it into a usable format. The survey had over 2600 columns with each being a different question asked, and with over 55,000 rows I couldn't read the whole dataset into my computer at once. Instead, I started by defining which questions I would be using in my model, and proceeded to manually found the column index for each of these questions. Then, I pulled only those columns for every row in the dataset, which gave me all participants' answers for the questions I was interested in.

Once I had this basic dataset, I began to clean and organize it. I dropped any participants who hadn't provided an answer to each of the survey questions I was using. Then, I had to decode the survey answers. Most of them were what I would call discrete particulars: the questions and answers were along the lines of What's your marital status? 1 = married, 2 = widowed, 3 = separated, 4 = never married, 99 = skip. For all of the survey questions I had to create dummies that capture effects on each group. Much of the survey data seemed as if it might be continuous (e.g. Age) when they were in fact bucketed oddly or were discrete particulars, and as such I spent much time making sure that each question had been decoded into dummies.

The three most important variables for me were whether the participant had suffered from an MDE (Major Depressive Episode) in their past, whether they had experienced another MDE in the past year, and whether they had sought treatment (inpatient or outpatient) in the past year.

Modeling

At the onset, my project aimed to determine which people from the general populace utilized mental health services. To this end, I ran multiple model types to predict usage as output: Logistic Regression, KNN, Random Forest, Gradient Boosting, and SVM. My dataset was highly imbalanced as about only 8% of the participants had utilized mental health services, so models trained on accuracy simply guessed that no on had used mental health services and had quite a high accuracy. As such I instead trained them to score on AUC, where it would maximize the TPR and minimize the FPR.

In my initial modeling Random Forest was by far the best modeling tool. However, I then began to gridsearch among each of my classification models. After fine-tuning the hyperparameters, almost all of my classification models had an extremely similar AUC (shown in the chart below).

	Log Reg	KNN	RF	GB	SVM
AUC	.799	.738	.797	.799	.705

As I neared upon choosing a model, I decided to change the topic of my project, and shifted towards the second question I proposed.

My new project aimed to determine what groups of people *from a subset of people who could greatly benefit from mental health services* actually utilized those services. While a subjective topic, I decided that people who had had MDEs in their past and had one in the previous year likely could benefit from mental health services. Depression is marked by having multiple MDEs throughout one's life, so these people were likely to have depression, and given that they had an MDE in the previous year it seems counselling would likely have been helpful. This obviously does not capture all the people who could benefit from mental health services, but I think it highly effectively captured a group that would benefit from utilization (i.e. it didn't grab everyone, but the people it did grab were likely to be in that group; in statistical terms, it was highly precise even if the recall wasn't great).

After taking the same course for this project as I had for my initial topic, I found extremely similar results: most models gave me virtually identical AUCs. I settled upon the Logistic Regression model because it gives Beta values that are easy to interpret, and as such is a highly actionable model.

I also tried changing my threshold value for my predictions on who utilizes mental health services. I wanted to ensure that if someone hadn't utilized mental health services they were coded as such; in other words, I wanted to minimize false negatives and have a very high recall. This is because those are people who could use help in accessing mental health resources, but would fall through the cracks. This came at the cost of having a higher false positive rate, but that simply means that people who were already utilizing mental health services would also be included in our initiative to increase access to mental health – a redundancy with little cost.

Model Results

I was extremely happy with the insights my model provided. It demonstrated that certain groups of people are far more likely to utilize mental health services; particularly, Caucasian people, straight people, and women are far more likely to utilize mental health services. Minority groups – both racially and of sexual orientation – routinely underutilized mental health services, so there is a gap that needs to be bridged. The next step in working

through this problem would be to identify why these groups aren't using mental health services: is it a lack of awareness (and as such we ought work on advertising)? Is it a lack of access (and as such should expand services)?

A further finding of my data was that people below the poverty line were more likely to utilize mental health services compared to people living above it. This served as a good access check financially – it seems that in my dataset, financial barriers were not the main inhibitor for usage of mental health services.

Tools

Python
Pandas
Scikitlearn
GridSearchCV

Algorithms

KNN
Logistic Regression
SVM
Gradient Boosting
Random Forest

What I'd do differently next time

I spent way too much time modeling something that I didn't have a good action case for. I think I need to define that far earlier in my projects, and make sure that I'm making progress towards my goal. Had I done that I could've spent more time bringing in data from earlier years, making my findings more robust. Further, I then could've added interaction terms, which I really wanted to evaluate in my model.