Angie Oehler
Final Paper

Patient Characteristic Survey in Predictive Analytics

The dataset:

The patient Characteristic Survey is collected and released by the state of New York every two years. This dataset is gathered and assembled from all regions and health services throughout the state and provides the mental health status for patients along with demographic information such as, diagnosis, additional diagnosis, patient sex, age, transgender and/or LBGTQ status, and veteran status. It also provides information about race, employment, education, and other health factors.

My plan with this project is to predict the mental health diagnosis of veterans seeking treatment. I plan to focus on individuals who are eligible to serve in the armed forces. While it is possible to enlist prior to the age of 18, active service does't begin until the person reaches the age of majority.

Predictive goals

I want to know if we can take a patient's variables and predict if they will be diagnosed with a severe mental illness versus general mental illness. Doing so could improve treatment options, and help to get the patient better treatment faster. I'm specifically interested in doing this for veterans because their experiences and needs can differ from the general population. There are also differences between male and female patients, and I want to know if this applies to veterans as well. My end goal with this is to predict what the patient's end diagnosis will be so that appropriate intervention can be applied as soon as possible.

Identified risks

The greatest risk with this proposal is that the veteran segment in the data might be too small to be meaningful in predictive analytics. That being said, veterans only make up a small

portion of the US population, so this might not be a factor and would actually mirror what happens in the real world. According to the US census, veterans make up only about 7% of the US adult population (Vespa). If the need arises, I can shift focus to patients as a group rather than just a subset of patients.

The other risk with this dataset is that it is all categorical. This might limit some of the models I could use. I've also never worked with all categorical data, so this could be more of a challenge to work with than I was expecting.

Visualizations and preliminary analysis

This dataset is all categorical data without a lot of options per category. This kind of dataset lends itself very well to histograms and bar charts when performing exploratory data analysis.

	Program	Sex	Orientation	Race	Veteran	Mental_Illness	TBI	SeriousIllness	Principal Diagnosis Class	Additional Diagnosis Class
0	OUTPATIENT	FEMALE	BISEXUAL	BLACK	NO	YES	NO	NO	MENTAL ILLNESS	MENTAL ILLNESS
1	OUTPATIENT	FEMALE	BISEXUAL	BLACK	NO	YES	NO	NO	MENTAL ILLNESS	MENTAL ILLNESS
2	INPATIENT	FEMALE	LESBIAN OR GAY	BLACK	NO	YES	NO	NO	MENTAL ILLNESS	MENTAL ILLNESS
3	OUTPATIENT	FEMALE	LESBIAN OR GAY	BLACK	NO	YES	NO	NO	MENTAL ILLNESS	MENTAL ILLNESS
4	OUTPATIENT	FEMALE	LESBIAN OR GAY	BLACK	NO	YES	NO	NO	MENTAL ILLNESS	MENTAL ILLNESS

I explored the variables I thought could be important against the patient program and if they were diagnosed with having a serious mental illness. More detail and examples are included in the next section.

Data adjustments

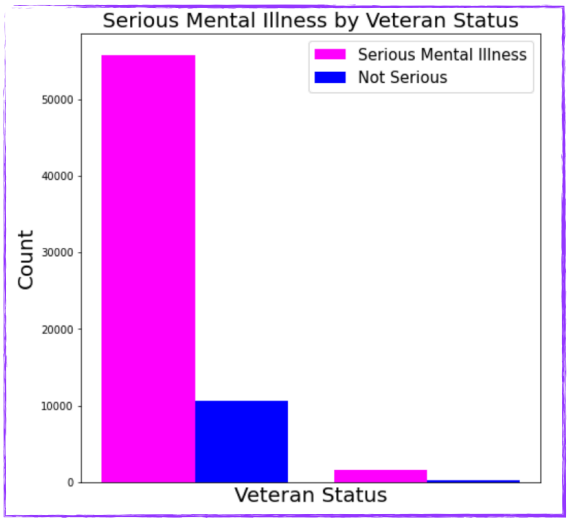
I had to make some adjustments to the dataset in order to remain centered on adults with mental illness. This will let me focus on those that are in a defined program, are receiving treatment, and are eligible to serve in the military. I have also removed individuals with both substance abuse disorders combined with a mental disorder because that is considered a dual

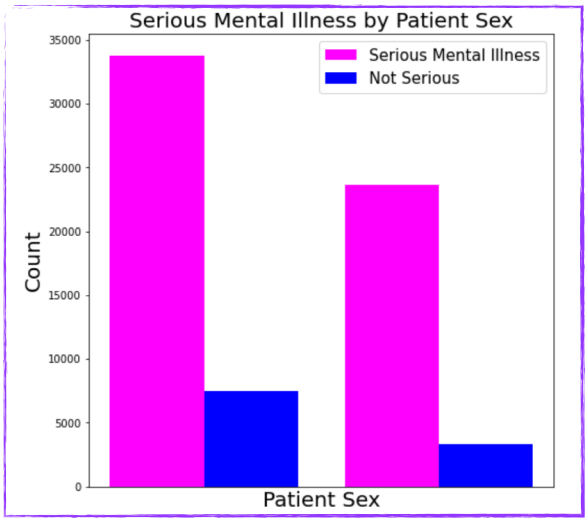
diagnosis (Kollasch - Parker). Eliminating these patients allows me to maintain single diagnosis disorders as a focal point.

After the initial analysis, I changed my predictive question from focusing on a patient's veteran status to general patient status. I considered using gender, but decided to leave it at patient status. Based on the counts of each within the dataset, this will give me a more balanced study as can be seen in the below graphics.

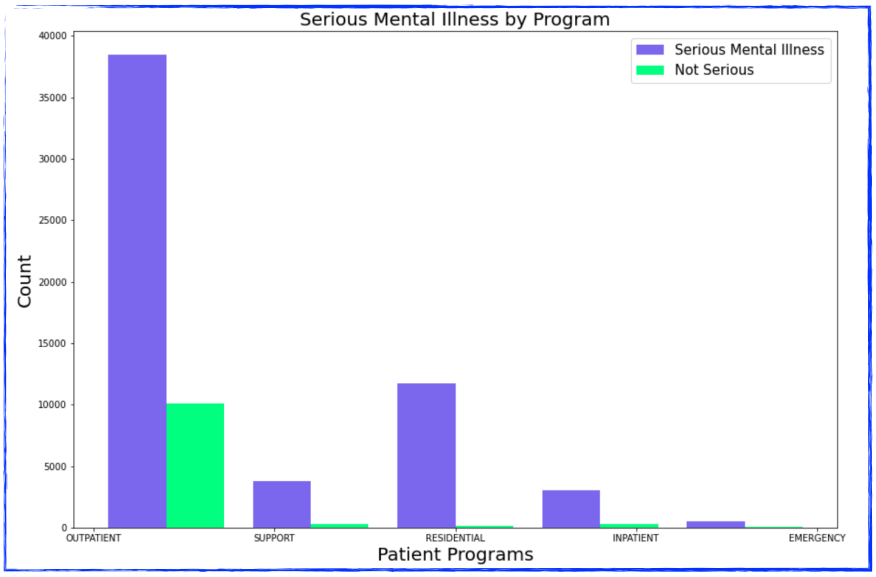
```
Veteran Totals
 NO     66372
YES      1872
Name: Veteran, dtype: int64

Gender Totals
 FEMALE    41301
MALE       26943
Name: Sex, dtype: int64
```

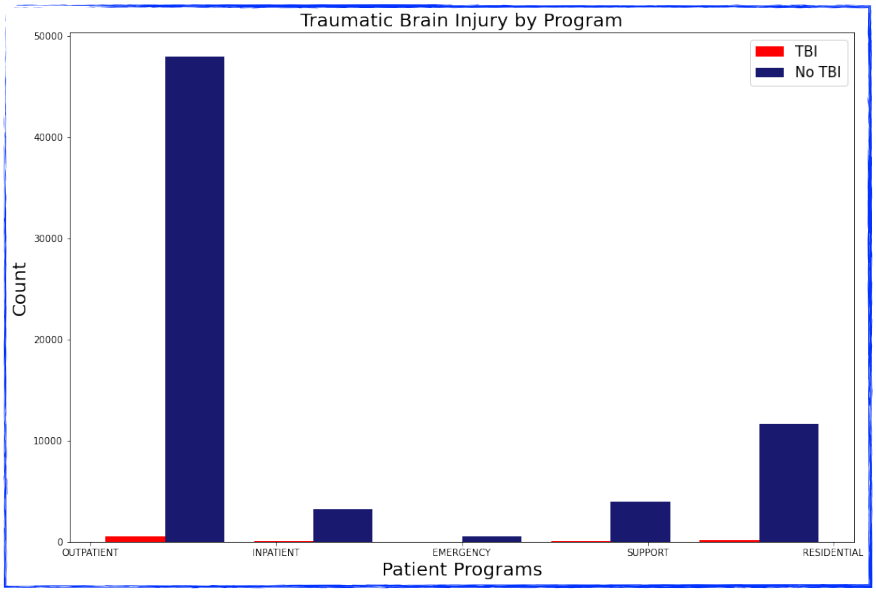
Below are the graphs showing veteran status and patient gender. The left graphic displays the count of a patient's veteran status and whether or not they have a serious mental illness compared to the population. The right figure shows the patient's sex and whether or not they have been diagnosed with serious mental illness.

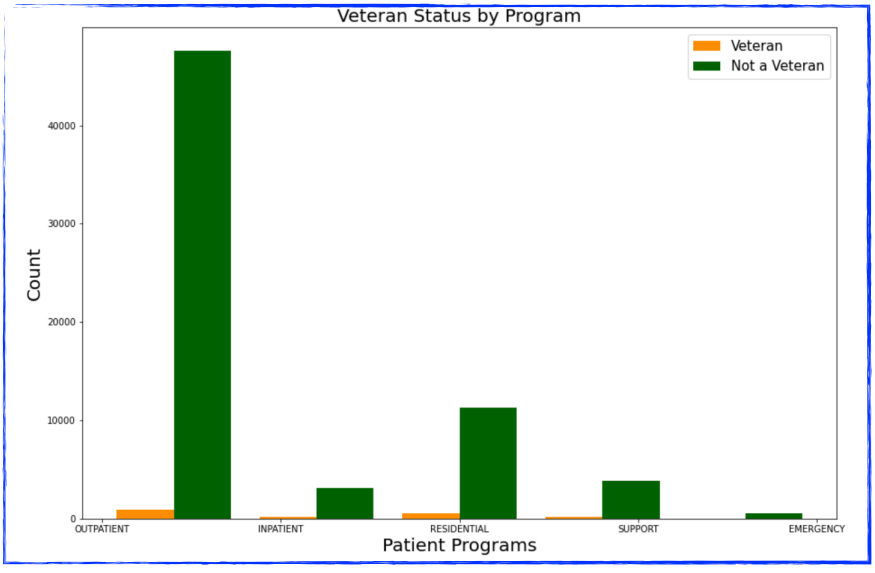




I also examined how the data variables of traumatic brain injury, veteran status, and patient gender are reflected according to the patient's treatment program as can be seen below. The treatment options are: Outpatient, Inpatient, Emergency, Support, Residential.


Serious Mental Illness by Program
Serious Mental Illness
Not Serious
Count
Patient Programs
OUTPATIENT
SUPPORT
RESIDENTIAL
INPATIENT
EMERGENCY


Traumatic Brain Injury by Program
TBI
No TBI
Count
Patient Programs
OUTPATIENT
INPATIENT
EMERGENCY
SUPPORT
RESIDENTIAL


Veteran Status by Program
Veteran
Not a Veteran
Count
Patient Programs
OUTPATIENT
INPATIENT
RESIDENTIAL
SUPPORT
EMERGENCY

Based on the above charts, it looks like patient sex could be a likely variable that deserves consideration for the predictive analytics project; however, the patient group as a whole looks like a better option.

Model selection

Because of the nature of my dataset, I believe my best choice is to use a classification model. They work by assigning data into classes, with examples including binary classification or multi-label classification (Brownlee). Since I want to know if the patient will be diagnosed with severe mental illness or not, binary classification models will be a good fit.

The two models I gave the most consideration to are decision trees and logistic regression. I like decision trees because they can be very effective in helping to find significant variables and identify the relationships between variables (Mitchell). I also like logistic regression for this dataset because it uses a binary target variable and doesn't require a normal distribution the data. As discovered during the EDA portion of this project, this data is not normally distributed.

My two favorite methods of model evaluation are confusion matrix or mean square error. I really like a confusion matrix because they are simple with easy to read and easy to explain results.

Additional data cleaning and prep

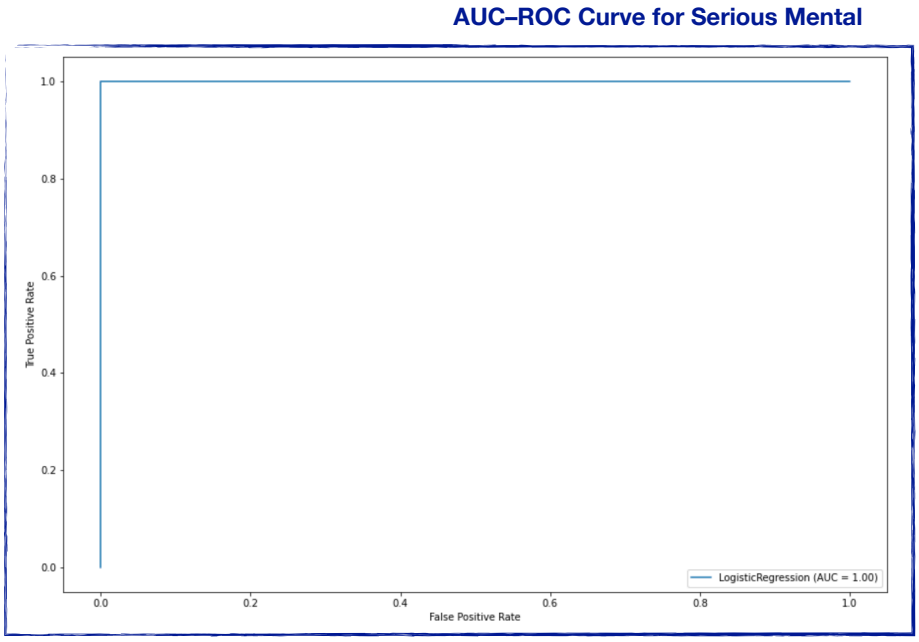
I performed the following steps to finalize and prepare my data for modeling:

- Checked the data for null values – there were not any.
- Removed anyone not eligible to serve in the armed forces to maintain a dataset of adults.
- Limited the number of variables to the features I am most interested in and are more likely to be used in predicting a serious mental illness.
- As discussed previously, I created visuals to explore mental illness rates and program categories by both patient veteran status and patient gender.

- Converted all the categorical variables to numeric variables in order to run the logistic regression model.

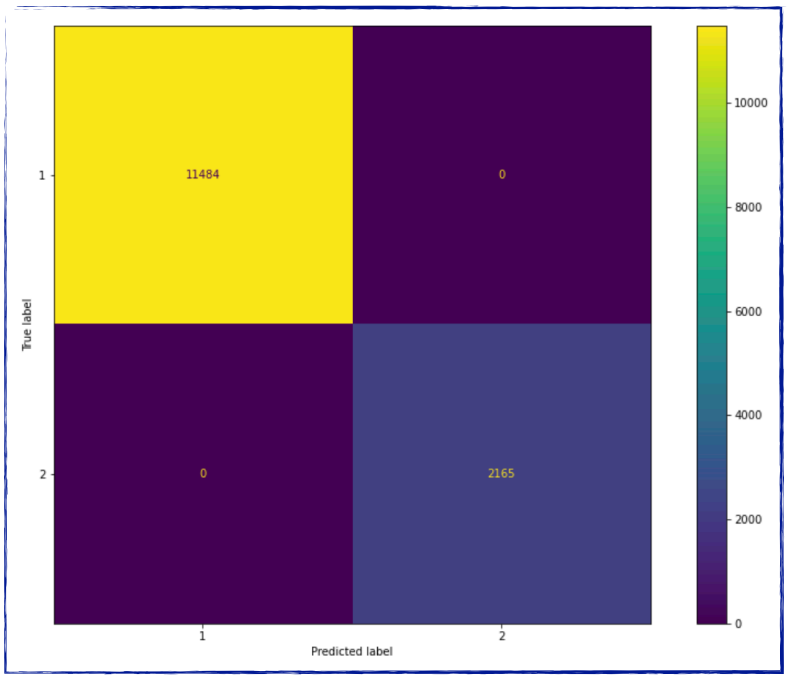
Model building and evaluation

I performed logistic regression on my data and used the AUC–ROC Curve to measure the results. This is a great way to check how well the classifier was able to distinguish between the classes in my data set. I like using the AUC–ROC Curve to visualize just how well my model did or didn't perform. I think it was the perfect choice for this model since the serious mental illness variable is a binary feature that is either yes or no. In this case, the AUC is 1, which means the model was able to distinguish between the positive and negative variables. To the side is the graphic showing these results.



I chose to further evaluate my model with a confusion matrix for a couple of reasons. First, when the AUC-ROC Curve is combined with the confusion matrix, I can really see how accurately the classifier was able to predict the true positives and true negatives. And second, they clearly display the results and are easy to understand. In this case, the confusion matrix reinforces the encouraging results presented by the AUC–ROC Curve. This model has done very well to predict whether or not the person has a serious mental illness. According to the

Confusion Matrix



confusion matrix, the model accurately predicted the positive and negative results, while both type 1 and type 2 errors were mitigated.

The classification report's precision, recall, and F1 scores were all one, which means the model was able to correctly predict results.

- Precision of 1 shows the predictions were correct.
- Recall of 1 shows the classifier was able to catch the positive instances.
- F1-score of 1 shows the model was able to correctly predict the positive results.

Classification Report

	precision	recall	f1-score	support
1	1.00	1.00	1.00	11484
2	1.00	1.00	1.00	2165
accuracy			1.00	13649
macro avg	1.00	1.00	1.00	13649
weighted avg	1.00	1.00	1.00	13649

Conclusion/recommendations

Based on the above results and the classification report, it looks like I will be able to use logistic regression to help predict if an individual will be diagnosed with a serious mental illness. The appropriate intervention will be recognized and incorporated into the patient's treatment program at an earlier point. This will allow the patient to receive better treatment, and hopefully, avoid a prolonged serious mental illness diagnosis.

Sources Referenced:

Aniruddha. (2020). *AUC-ROC Curve in Machine Learning Clearly Explained*. Retrieved from https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

Brownlee, J. (2020). *4 Types of Classification Tasks in Machine Learning.* Retrieved from https://machinelearningmastery.com/types-of-classification-in-machine-learning/.

Lisa Kollasch-Parker, DNP, APRN, FNP-C. Nebraska Board of Nursing, Member.

Mitchell, M. (2019). *Selecting the Correct Predictive Modeling Technique.* Retrieved from https://towardsdatascience.com/selecting-the-correct-predictive-modeling-technique-ba459c370d59.

New York State. (2021). *Patient Characteristics Survey (PCS): 2019*. Retrieved from https://catalog.data.gov/dataset/patient-characteristics-survey-pcs-2019.

Vespa, J. (2020). *Those Who Served: America's Veterans From World War II to the War on Terror.* Retrieved from https://www.census.gov/library/publications/2020/demo/acs-43.html.