Capstone 3
Final Report
Chris Coble
7/29/2024

Mental Health Disparities Amongst DIfferent Races During Economic Strife

Introduction

Mental health and systemic racism are pervasive issues facing all Americans at one moment or another, and during the height of the COVID-19 pandemic, many people were extremely challenged. Due to extreme health concerns, lack of social connection, loss of habits/routine, scarcity of resources or financial strain, many people were afflicted with acute stress due to the height of the COVID-19 pandemic. Because of systemic racism's deep integration into the U.S. economic system, minority races are more often lacking access to the resources in times of crisis; whether that be food, water, general health care, or even mental health care.

Without proper mental health care, there is the glaring issue of the suffering of the people's suffering, and there also is the larger impact, where families and groups are largely disrupted because of mental illness. Severe mental illness can lead to dropping out of society, halting work, stopping education, potentially suicide and homicide, and induction of more mental illness in those around the afflicted. So according to systemic racism, there would be higher incidence of mental health or lack of resources in minority groups. Thus, during the pandemic, minority races, specifically, Hispanic, Black, and Asian groups, were differently impacted than the majority group in the US.

In this project, the models constructed will attempt to predict the average mental health score of racial/ethnic groups in the US based on race/ethnicity, unemployment rate change, and the unemployment total within the racial group. The output mental health score will be on a continuous range from 0-48. The score is that range because of the datasets used to construct the model. One of the sources is a kaggle dataset compiling information from the National Center for Health Statistics (NCHS) with mental health scores from GAD-7 and PHQ-9 questionnaires given to various individuals throughout 2020-2022 (https://www.kaggle.com/datasets/subidit/indicators-of-anxiety-or-depression). The other dataset used in this project is unemployment statistics in the view of race/ethnicity from the Bureau of Labor Statistics (BLS) from 2019-2022 (https://www.bls.gov/cps/cps_aa2022.htm). This data was vital in finding the mental health pattern in the US in this most recent medical and economic crisis.

Data Wrangle

https://github.com/gisthuband/Capstone_3_Mental_Health_Score_Predictor/blob/main/data_wrangle/data_wrangle.ipynb

As previously mentioned, the two data sources came from the NCHS and the BLS, two government entities, that reported on mental health survey scores from 2020 to 2022 and unemployment statistics from 2019 to 2022. The NCHS dataset contained information, such as date, race/ethnicity, mental health survey score, scoring confidence intervals, and indicator of mental health test. With the BLS report, there was information from 2019 to 2022 containing information on the unemployment rate and total amount of unemployment for the four racial groups included in both the mental health report and the BLS report: Hispanic, Black, White, and Asian. Changes in unemployment rates and unemployment total numbers were calculated from 2019 to 2020, 2020 to 2021, and 2021 to 2022. These three year changes were aligned with the three years worth of mental health information form the NCHS. If the changes were positive, then that meant that the unemployment rate/ unemployment total were increasing and vice versa. The unemployment information was parsed out in gender and racial differences, but the genders were reconciled with, as the total rates were weight-averaged and the total amount were summed within each racial group. With this, the two datasets were merged, containing, date, score, mental health indication (anxiety, depression, anxiety or depression), racial group, change in unemployment of the group, change in the unemployment rate of the group.

Exploratory Data Analysis

https://github.com/gisthuband/Capstone_3_Mental_Health_Score_Predictor/blob/main/exploratory_data_analysis/exploratory_data_analysis.ipynb

The exploratory data analysis was split up into several steps: checking for missing data and imputation, visualizing mental health scores based on racial/ethnic groups, a Y-data profiling report, an attempt at time analysis of the mental health scores during the two year period, and correlation and relationship tests in heatmaps and z score tests.

In the first and second step, the missing data in the mental health scores were identified and imputed by using racial grouping yearly means. The important missing data was in the mental health scores. There were about 96 observations that were missing the scores, and there were around 700 observations in total for the two year period and the four racial/ethnic groups. And the missing values were split up evenly amongst the racial groups, so the distribution of the missing values were decided by the racial group and the year of the scoring. During the viewing of all of the scores, there were some disturbing patterns found within the data: Hispanic and Black Americans were sufficiently more afflicted with mental health issues based on higher distributions of scores (closer to 48, which would indicate higher indication of severe anxiety and depression). Interestingly, White Americans were lower than the two previously mentioned groups, but not as low as the Asian American group. Once the means of the specific racial/year groupings were isolated, the missing values were imputed.

Next, the Y-data profiling took place. Within the report, it was found that the data was tidy and the distribution of scores was Gaussian. There was also a lack of correlation amongst numerical features (unemployment rate change and unemployment change) with the survey scores. However, there was a collinearity found within the unemployment rate and unemployment amount change, as to be expected.

In the next step, a time series analysis was attempted. Using the data from 2020 to 2022, the dates were plotted against the mental health scores. The graphs showed that the scores were much higher in 2020 and the beginning of 2021, but then dropped as time progressed into 2022. The racial groups were introduced into the graph via adding a hue. The graphs confirmed what the scores distributions did previously: Hispanic and Black Americans are on average higher scoring on the surveys compared to White and Asian Americans from 2020-2022. During this step, there was an intention to try and decompose the data to see if there was seasonal or trend pattern within the data, however, because there was only one crest and trough, the seasonal pattern would not be able to be deduced in a generalizable way (would need more recessions to create a seasonal pattern), and the trend also was not able to be discerned due to lack of information. The decomposition step was passed over, but the visuals were valuable.

The last step was to check for relationships between the data. A heatmap of the Pearson correlations within the data was visualized. The map revealed what was previously revealed: not high correlation between mental health score and unemployment statistics, and high collinearities between the unemployment statistics. Next Chi-Squared test was considered, to see if there were significant relationships between categorical variables, however, since the output variable will be a continuous range of values, the z-tests were considered to be more important to the direction of the project. The z-test compared the means of the survey scores between the different racial groups and also compared means of the survey scores between different test indications and their survey scores. The z tests revealed that there were significant differences between Asian and Black or Hispanic scores, and anxiety of depression indications were significantly different than depression indications ($p < .05$, z statistic > 1.96). With this information in hand, modeling and predicting was just around the corner.

Preprocessing and Modeling

https://github.com/gisthuband/Capstone_3_Mental_Health_Score_Predictor/blob/main/preprocessing/preprocessing.ipynb

https://github.com/gisthuband/Capstone_3_Mental_Health_Score_Predictor/blob/main/modeling/modeling.ipynb

Since the output is expected to be a continuous range of scores, a regression model is the general direction of the model. The categorical data can be encoded into numerical data via one-hot encoding, but the already existing information can also be input into a regression model as well. Univariate regression will not be an option, as the variance in the employment and

employment rate variables is most likely not high enough in order for discovery of patterns in the data.

The preprocessing step was quite simple as it involved one main step, construction of a class that could input the data frame and produce the training and test data, one-hot encoded potentially or standardized, for the model to work with.

Eight models were deployed to examine this data: Ordinary Least Squares, Ridge Regression, Lasso Regression, Random Forest Regressor, K Neighbors Regressor, Linear Support Vector Machine, Gradient Boosting Regressor, and a Deep Neural Network. The test results are in the table below:

Model	R_Squared	Root_Mean_Squared_Error
Ordinary_Least_Squares	0.07267146928	6.034206296
Ridge	0.7311340333	3.285131356
Lasso	0.7621064319	3.344189157
Random_Forest	0.7223852013	3.418525669
K_Neighbors	0.7991674969	3.383203504
Gradient_Boosting	0.7573394553	3.136241371
Support_Vector_Machine	0.7517927508	3.442532139
Deep_Neural_Network	0.7567943633	3.260544893

From these results, you can see that the worst performing model was the Ordinary Least Squares algorithm, and Lasso Regression, K Neighbors Regression, and Gradient Boosting Regression were the best performing algorithms as far as R squared values and root mean squared error. However, the R squared values, typically centered around .75 showing that the data was limited as far as the patterns that could be detected. Also the root mean squared error was centered around 3, symbolic of getting the survey score wrong by one question (as one question would have a value of 0-3 depending on the severity of the answer). The Ordinary Least Squares algorithm produced a terrible R squared value in comparison, at .07, most likely because the data used to construct that line was only the strictly numeric data: the unemployment rate change and the unemployment amount. These numeric values contained a limited range of values, as they were generated from the BLS report for each year change and for each racial group (set of 12 values each to assign to each value). The other models were given one hot encoded categorical information (race, test indicator) that could help the algorithm understand the data. K Neighbors most likely worked very well because the data could be partitioned quite well according to racial group in the high dimensional space of the dataframe. The Lasso algorithm also performed quite well because the penalty term of the algorithm reduces the weight of unimportant features close to zero, similar to feature reduction.

The models chosen to move onto the prediction stage of the project were Lasso, K_Neighbors, and the Random Forest model. The Lasso and K_Neighbors were chosen because they were the best performing models in terms of R squared and mean squared error. Meanwhile, the Random Forest model was chosen because it was the worst performing of the decently performing models (any models performing in the range of .7 R squared or higher). By comparing the best performing models to the worst performing model, we can see how the data is being interpreted differently in the algorithmic dimension of the models. So if Lasso performs similarly in numerical prediction as K Neighbors, that means that the lasso algorithm is able to minimize features so much that it can create a solid, continuous range of predictions like the K Neighbors algorithm. If the Lasso produces similar results to the Random Forest model, then the Lasso's feature reduction is similar enough to the averaged decision trees in the forest model. Comparison is quite informative on how the data is being interpreted mathematically.

Results

The models: Random Forest, K Neighbors, and Lasso were selected for the results function; the function took type of indication, racial group, unemployment rate change and unemployment number change, and with this it constructed a mental health score of the average person in that demographic with a root mean squared error statistic as well. The function was run with three random inputs and the results are listed below:

Demographic	Increasing Unemployment Average Score	Decreasing Unemployment Average Score
Latino/Hispanic	31.49 (2)	31.84 (7)
White	34.35 (7)	29.18 (7)
Black	35.79 (5)	29.11 (4)
Asian	28.31 (4)	25.28 (4)

As one can see in the table, it is divided into a demographic column, increasing unemployment average score, and decreasing unemployment average score. These averages are based on the behavior of the three selected models together. The numbers in the row indicate the score predicted for the demographic, and the number of randomly generated numbers factored into the average. Ideally, there would be at least 6 or 7 data points for each demographic, but this is a nice start that still captures some behaviors within the data. Whether unemployment is increasing or not, Hispanic/Latino groups are consistently around 32 based on these numbers; again, since the data points for one of the averages is only two, it may not be accurate of the trend. For the White American group, the score goes down around 5 points (roughly equivalent to going down to 0 for two questions on the exams). For the Black American group, the score goes down to 29, like the White American group, and the score is going down almost 6 points. For the Asian American group, the scores go down 3 points (or one question) with the turn of the economy.

For a view at the individual model's performances, look below:

Model:	K Neighbors		Lasso		Random Forest	
Demographic	Increasing Unemployment Average Score	Decreasing Unemployment Average Score	Increasing Unemployment Average Score	Decreasing Unemployment Average Score	Increasing Unemployment Average Score	Decreasing Unemployment Average Score
Latino/Hispanic	29.30962302	31.389375	34.1943432	32.28036598	30.95265988	31.85393851
White	34.70034014	29.66832672	33.03765746	26.97539632	35.30451059	30.88601631
Black	36.51006945	28.55336227	36.74044565	28.94509449	34.89530425	29.8260867
Asian	32.9265377	30.96402778	26.06093523	21.37554107	25.94710045	23.49165406

Typically, Hispanic/Latino groups are showing very small changes in averages between increasing and decreasing unemployment. This is most likely due to the small number of data points for increasing unemployment. The White group seems to be highly variable with Lasso. THe Black group seems to be the most consistent across the three models. And the Asian group's prediction from the K Neighbors model is much higher than the other two. Based on the inconsistency in the individual model's performance, besides for Black American predictions, the other three groups are somewhat suspect, and most likely require more data points to capture the full extent of the model's abilities.

These results show that, on average, there is a positive impact on the predicted mental health scores, when the economy recovers. This symbolically translates to the importance that economic needs are met by different groups in America. Without resources, mental health suffers, extending a toll on America as a whole in terms on premature deaths by suicide, drug addiction, homicide, falling into trafficking, dropping out of society/education prematurely, etc.

More predicted data needs to be collected to see if the averages change for the Hispanic/Latino group. More data from other economic/health crises like the 2008 housing crisis plus the 2009 Influenza pandemic would be additional information that would be helpful in the construction of the data. With this kind of information, hopefully one of the models would be able to break the .8 R squared score, ensuring a better representation of the data. The most representative model would be extremely useful, as that kind of model could relay the scores to predict the amount of diagnoses, cost of medications on average, incidence of psychiatry and therapy, and hospitalization costs.

Conclusion

Mental health is extremely important to the success of any society, but mental health is under threat during economic/health crises like seen in COVID-19; furthermore, this kind of phenomenon can impact different racial groups more severely than others due to already-existing systemic racism. In a well functioning society, the members of the society have their basic needs met; when mental health degrades, people are less likely to have the mental capacity to retain enough labor for rent/housing, utilities, food, and water. Without these

resources, a human body is likely to degrade in mental health and physical health even further. Thus it is paramount that mental health is treated seriously at the start. It is up to the government or extremely philanthropic institutions to provide preventative and curative options for at-risk racial groups that are being restricted economically due to systemic racism. All people in unison, act like pillars to a societal structure; when their minds and in turn, bodies, collapse, the roof will cave in, leaving the structure broken and vulnerable.