Mental Health In the Pandemic Era: A data-driven study on Race, Sex, Education, and Disability Status

Chang Zhou
Department of EECS
University of California, Berkeley
chang_zhou@berkeley.edu

Yinyin Liu
Department of EECS
University of California, Berkeley
jessica_liu@berkeley.edu

Yefan Zhou
Department of EECS
University of California, Berkeley
yefan0726@berkeley.edu

1 Introduction

COVID-19 has imposed a huge challenge on people's health, both physically and mentally. In our project, we explored dataset 1B - COVID-19: Impact on Mental Health to analyze the mental health conditions of US citizens during pandemic. After research on the data, we are particularly interested in two questions:

First, we are interested in what factors are highly correlated with depression and anxiety rate. The datasets provide data including insurance coverage, reduced access to healthcare, timeliness of healthcare and so on. In our project, we build models to predict the level of depression and anxiety using those factors. Then we run significance tests to figure out which factors are most significant. By asking this question, even though we don't have the capacity to conduct causal inference, we hope our prediction models could still bring some insights into how to possibly lower anxiety and depression rates.

Second, another question we are interested in is how people in different gender/race/geographical/ disability status groups differ in terms of anxiety level. The pandemic has been a challenge to the society as a whole, but some subgroups of our society may have suffered more: Asians may have gone through Asian hate in early 2020, people with lower educational levels might have their job stability threatened, different genders may react to the pandemic differently. We believe efforts should not only be made to improve the mental health status of society as a whole, but more attention needs to be paid to different subgroups. In this study, we hope to bring this issue to light by utilizing the data science knowledge we have gained in class.

In the first research question, we found that insurance coverage and intermediate access to healthcare are most correlated with the anxiety and depression level. In the second research question, we found that females,, hispanic and disabled people have higher levels of depression and anxiety, while male, asian/white people have lower levels of depression and anxiety. Those results may provide statistical insights to government policy makers on how to enhance public mental health, and to ensure that vulnerable groups are taken care of during the pandemic.

2 Related Work

There have been lots of studies on the impact of COVID-19 on public mental health. [6] provides a systematic study of the mental impact of COVID-19 and identifies several key factors across different social groups. However, it focuses on making comprehensive exploratory data analysis while our

Data 200 Graduate Project.

work includes building models to predict anxiety symptoms. [5] studies the global prevalence of anxiety disorder symptoms due to the pandemic and provides insight into the difference between countries. On the contrary, our work focus on analyzing the anxiety symptoms in the United States and looking deeper into the difference between states and social groups. [3] builds the linear model to predict the COVID-19 survivors' anxiety, but we emphasize predicting the anxiety of the public during pandemics, and we show that the ensemble supervised learning methods like Random Forest and Gradient Boosting perform better than the linear model.

3 Methodology

In the methodology section, we will introduce how we preformed data preprocessing and how we built models with our cleaned data. After we constructed the model, we also conduct inference on the model coefficients.

3.1 Data Cleaning

In order to provide training data for different models and to solve our research questions, we performed the following data preprocessing steps:

- Step 1: Merging the data from the first four datasets:
  The data frames are merged based on the tuple (date, group, subgroup).
- Step 2: Converting the responses into feature columns:
  Indicator columns and value columns are converted to feature columns.
- Step 3: One-hot encoding of the categorical data:
  Some features are categorical, such as subgroups (races, genders, age groups, and educational levels). To use those features in our model, one-hot encoding is needed to convert the data to numeric data.
- Step 4: Missing data
  We drop rows with more than 3 NaN, and then Impute the rest of the missing values with the mean value of the column.
- Step 5: Normalize the data
  Since we are planning to use OLS for the task, numeric features need to be normalized in order to not bias the l2 loss.

3.2 Modeling

From the exploratory analysis, we observed that there is a large difference in anxiety level among different races, genders, age groups, and educational levels. The differences are significant on the plots; thus we want to see if it is also significant statistically.

We want to figure out if we could use reduced access to medical care, health insurance coverage, and other factors to predict the anxiety level of a group. To answer the above two research questions, our approach is to build several linear regression models with and with lasso regularization on different subgroups and evaluate the model performance. We choose lasso regularization because it can get rid of some redundant features and improve model stability. We also tried some other models such as random forest and gradient boost as alternatives to see if they can improve the prediction performance. We find that the gradient boost method is sensitive to its hyperparameter learning rate so we use cross-validation to do the hyperparameter tuning on train set to find the optimal setting.

To analysis the performance of our models, we use R^2, root mean squared error (RMSE) and mean squared error (MSE) as our metrics. We use R^2 as one of our metrics to identify the proportion of the variance for our dependent feature, which in our case, is the percentage of people who suffered from either anxiety and depression, that's explained by independent feature variables in our regression models. We use RMSE as the other metric in order to identify the standard deviation of prediction errors; in order word, to know how close to the ground truth our models' predictions are. We use MSE as the criterion to do cross-validation in the hyperparameter tuning.

3.3 Inference

To answer the other question what are some factors that are related to the high public anxiety rate, we conducted significance tests with the null hypothesis:

H_0: race/gender/education level has no impact on the anxiety level;

In other words, the coefficients corresponding to the relevant features are 0. We conduct the significance test by bootstrapping the samples, and then fit 1000 models to get the confidence interval for the coefficients, if the confidence intervals don't contain zero, then we could confidently reject the null hypothesis. By analysing the coefficients of our linear regression model, it helps us answers our research question that which groups suffer the anxiety problem most and demand treatment.

4 Exploratory Data Anaylsis

This section provides the results of exploratory data analysis to give a preliminary insight into the two questions we consider. After aggregating four datasets in dataset-1B, we have a total of 10 features and one label, as listed in the Table 1. Section 4.2 answers the first question by identifying the high-correlated factors with anxiety levels and visualizing the correlations' regional (state) distribution. In Section 4.2, we provide insights to the second question by providing data visualizations for private insurance coverage rate, reduced access to general health care rate, reduced access to mental health care rate, and anxiety or depression levels for people in different groups and compare and contrast those data visualizations.

4.1 Preliminaries

We assign abbreviations to the names of the features for simplicity. The chart below shows the name-abbreviation pair.

Full Name of Feature/Label	Abbreviation
Delayed Medical Care, Last 4 Weeks	Delayed Care
Delayed or Did Not Get Care, Last 4 Weeks	Delayed or Not Care
Did Not Get Needed Care, Last 4 Weeks	Not Care
Private Health Insurance Coverage	Private Insurance
Public Health Insurance Coverage	Public Insurance
Uninsured at the Time of Interview	Uninsured
Needed Counseling or Therapy But Did Not Get It, Last 4 Weeks	Not Therapy
Received Counseling or Therapy, Last 4 Weeks	Get Therapy
Took Prescription Medication for Mental Health And/Or Received Counseling or Therapy, Last 4 Weeks	Took Prescription/Counseling
Took Prescription Medication for Mental Health, Last 4 Weeks	Took Prescription
Symptoms of Anxiety Disorder or Depressive Disorder	Anxiety Symptoms

Table 1: List of features and labels and assigned abbreviation.

4.2 Question 1: Highly-Correlated Factors to Predict Depression and Anxiety Level

We show two EDA results in this section, correlation matrix visualization and the similarity of state distribution of the features.

In Figure 1, we calculate the pairwise Pearson coefficient between features and label and visualize the matrix using heatmap. We focus on finding the highly correlated features with **Anxiety Symptoms**

in the second last row. We can tell that the coefficients of features **Delayed care**, **Delayed or Not Care**, **Not Care**, and **Not Therapy** are larger than 0.5, meaning these features positively correlate with anxiety levels. We can draw **our first finding** that the high-correlated factors are the failures and delays in getting Medical Care. This finding indicates that, people are more likely to feel anxious if they feel sick but fail to get immediate care and treatment during the pandemic. However, we hypothesize collinearity between the four features because we can see that the coefficients between **Not Therapy** and the three features regarding failures to get medical care are nearly 0.3. In addition, the information these features represent is similar, and they could stand for the duplicate sampling records.

Another type of extraordinary features is medical insurance. **Private insurance** is negatively correlated with **Anxiety Symptoms** (-0.3), while **public insurance** and **uninsured** are weakly positively correlated (0.2). We also find that public/no insurance is positively correlated to failure/delay in getting medical care/therapy (about 0.1), and no insurance has a significant negative coefficient (-0.5) with **Get Therapy**. In contrast, **private insurance** has a negative coefficient with failure and delay and is positively correlated with **Get Therapy**.

These results lead to **the second interesting finding** that public insurance is ineffective in providing people with immediate medical care and therapy. Therefore, people even view having public insurance as having no insurance, and public/no insurance identically result in high anxiety levels. On the contrary, private insurance effectively reduces anxiety by providing reliable and timely medical support.

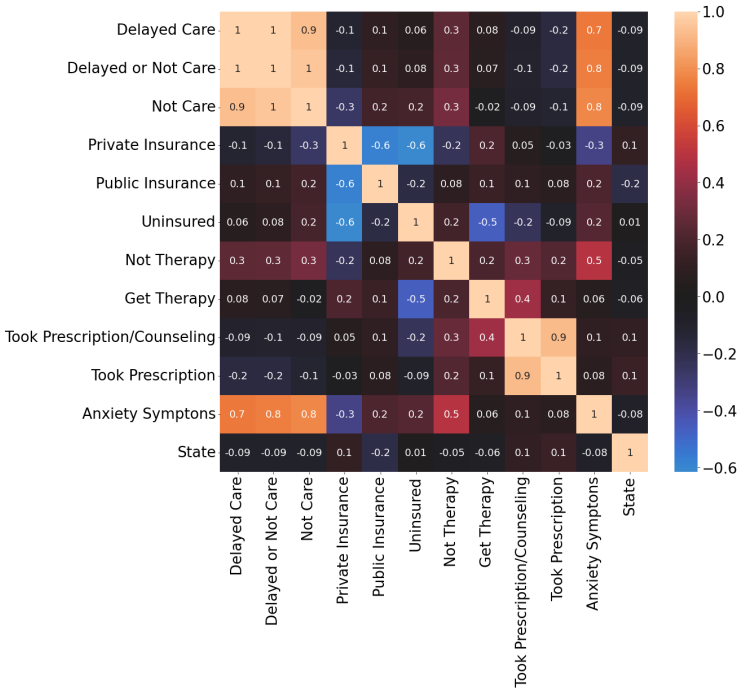
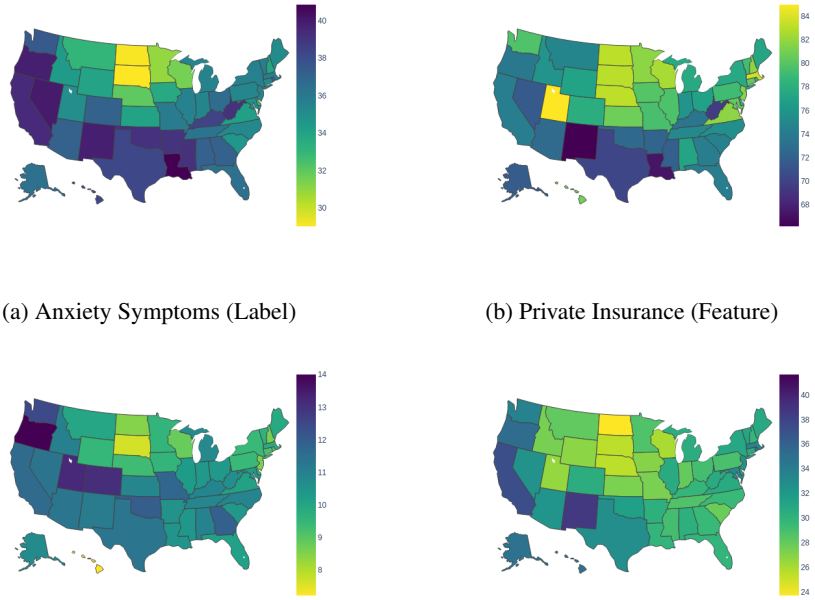


Figure 1: Heatmap of pairwise Pearson correlation matrix on features.

In Figure 2, We visualize the regional distributions of anxiety symptoms and three medical care and insurance features on a United States Map. We can see the similarity in color brightness distribution between the four sub-figures. For example, the states along the coast and south central states like Texas, New Mexico and California generally have a dark color among the four figures. It indicates that these states have high anxiety levels, low private insurance coverage rates, and a high failure rate to get timely therapy or medical care. We draw the **the third finding** from this result that the state factors influence the distribution of anxiety levels. A hypothesis is that the state's medical policy or insurance price influences the rate of private insurance coverage. The low coverage rate directly decreases the possibility of getting timely medical care/therapy which finally leads to a high anxiety rate.

We further investigate the causes of low private insurance in California. [2] finds that private health insurance in California is expensive due to a lack of provider competition. There are fewer medical groups and competing hospitals in California, so even if the monopoly provider sets a high price for the basic plan, people have no choice but to pay it. The potential treatment is to use policy to regularize the market and restore the competition. The cause of low private health insurance coverage in some states is an open research question for future work.



(a) Anxiety Symptoms (Label)

(b) Private Insurance (Feature)

(c) Not Therapy (Feature)

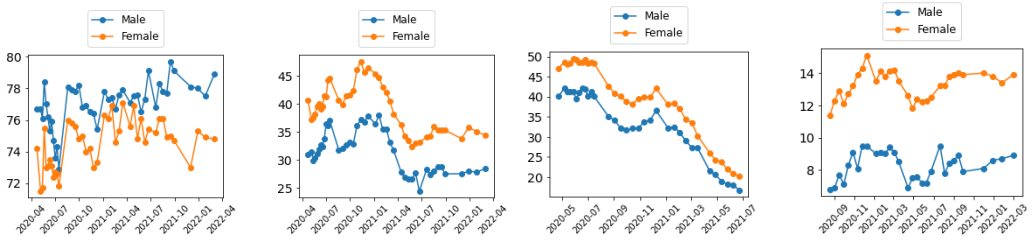
(d) Delayed or Not Care (Feature)

Figure 2: Comparing the regional distributions of Anxiety Label and Medical Care, Insurance features.

4.3 Question 2: How People in Different Groups Differ in Terms of Depression and Anxiety Level

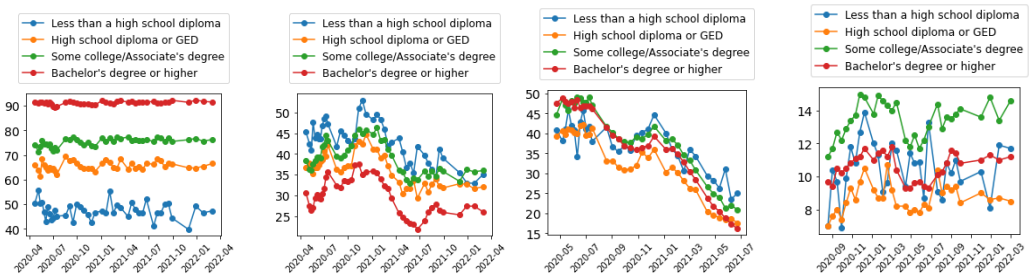
The data visualizations for private health insurance coverage rate, reduced access to general healthcare rate, reduced access to mental healthcare rate, and anxiety or depression levels for people in different groups are shown in Figure 3, Figure 4, Figure 5, and Figure 6. We observed that males, Asians/whites, non-disabled and more educated people are more likely to be covered by private insurance; males, Asians/white, and non-disabled, are less likely to undergo reduced access to healthcare, both generally and mentally. Possibly as a consequence, they have lower depression and anxiety level.

Besides, we also noticed that as the pandemic evolves, people tend to have more access to general healthcare over time, yet their access to mental health support remains the same throughout the pandemic. This could possibly be a warning sign that the mental health problem has been omitted.



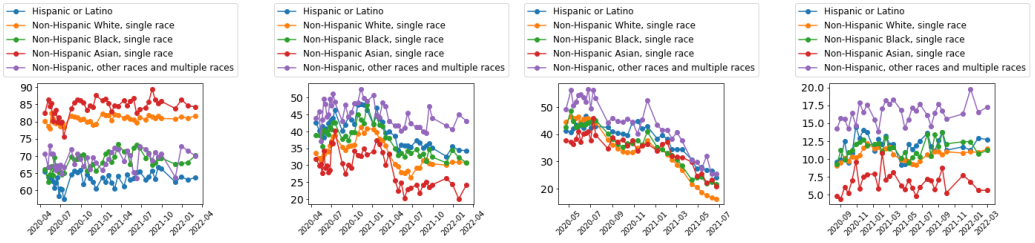
(a) (b) (c) (d)

Figure 3: Data Visualizations for Different Groups of People in the US by Sex: (a) The percentage of people who have private insurance in the US by sex. (b) The percentage of people who suffered from anxiety or depression disorder in the US by sex. (c) The percentage of people who received delayed general health care or did not get health care in the US by sex. (d) The percentage of people who needed counselling or therapy for mental health but did not received it in the US by sex.



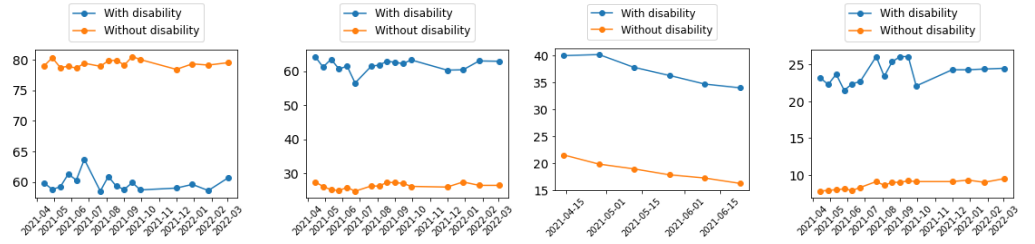
(a) (b) (c) (d)

Figure 4: Data Visualizations for Different Groups of People in the US by Education: (a) The percentage of people who have private insurance in the US by education. (b) The percentage of people who suffered from anxiety or depression disorder in the US by education. (c) The percentage of people who received delayed general health care or did not get health care in the US by education. (d) The percentage of people who needed counselling or therapy for mental health but did not received it in the US by education.



(a) (b) (c) (d)

Figure 5: Data Visualizations for Different Groups of People in the US by Race: (a) The percentage of people who have private insurance in the US by race. (b) The percentage of people who suffered from anxiety or depression disorder in the US by race. (c) The percentage of people who received delayed general health care or did not get health care in the US by race. (d) The percentage of people who needed counselling or therapy for mental health but did not received it in the US by race.



(a) (b) (c) (d)

Figure 6: Data Visualizations for Different Groups of People in the US by Disability Status: (a) The percentage of people who have private insurance in the US by disability status. (b) The percentage of people who suffered from anxiety or depression disorder in the US by disability status. (c) The percentage of people who received delayed general health care or did not get health care in the US by disability status. (d) The percentage of people who needed counselling or therapy for mental health but did not received it in the US by disability status.

5 Modeling, Inference and Analysis

5.1 Question 1: Highly-Correlated Factors to Predict Depression and Anxiety Level

We build models to predict the anxiety level based on the medical care and health insurance features listed in Table 1. The dataset size is 991, and the number of features is 10. We split two-thirds of the dataset as a

	Test R^2	Test RMSE	Train R^2	Train RMSE
Linear Regression	0.6970	3.1629	0.7070	3.0235
Random Forest	**0.7139**	**3.0734**	0.9569	1.1590
Gradient Boosting	0.7136	3.0750	0.8314	2.2938

Table 2: Comparing three regression approaches on predicting anxiety levels.

train set and the rest as a test set. We select three approaches for modeling and report their train/test RMSE and R^2 in Table 2. We can find that random forest has the best R2 and RMSE on the test set. The superior performance of random forest is due to the ensemble learning method for prediction. It can avoid overfitting better because of the randomness of multiple decision tree models. However, we find that random forest has a suboptimal generalization gap [1], defined as the difference between test and train RMSE in this case. It means the random forest model is very likely to have poor performance if we can collect more data to evaluate. The Linear Regression model has a good generalization gap, but the test RMSE is relatively high due to the model's limited number of parameters (10). The Gradient Boosting model performs well on the generalization gap and tests RMSE, and we propose to select this model as the final optimal model for future application. We present the details of the three models about the confidence interval of feature coefficient, hyperparameter tuning, and cross validation in the following.

Linear Regression and Bootstrapped Confidence Interval For Linear Regression model, we show its residual plot in Figure 7. We can tell that there is no apparent visual patterns in the residual distribution, which means our prediction is unbiased.

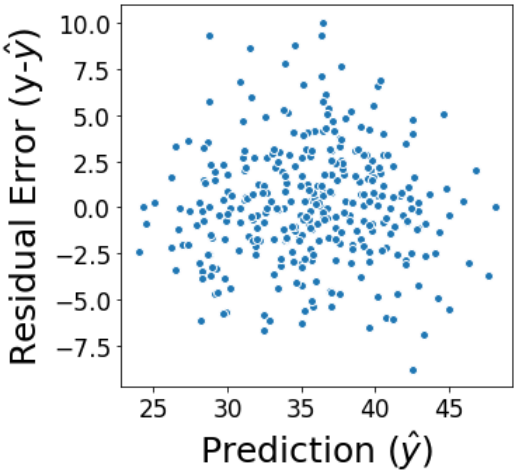


Figure 7: Residual Plot of the Linear Regression model.

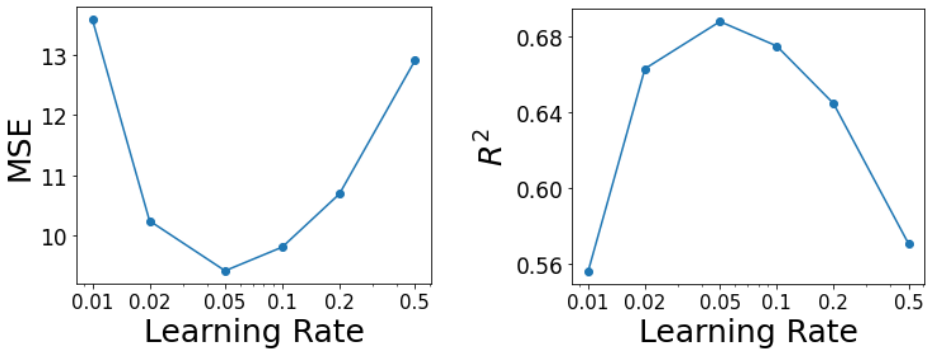
We show the bootstrapped confidence interval of the model coefficients in Table 3. For the bootstrap sampling, we randomly sample the original dataset for 100 times and build 95% confidence interval based on the 100 models. First, we identify the features **Not Care**, **Uninsured** and **No Therapy** reject the null hypothesis, indicating they are significant in affecting the prediction labels. This observation of these significant features is aligned with the correlation experiment finding in Section 4.2. Second, we find the collinearity between features, e.g., **Delayed Care** fails the significance test because it might represent the same information with **Delayed or Not Care**. Similarly, **private insurance** also fails because it is linearly dependent on **public insurance** and **no insurance** because the sum of the three insurance features could be the total population. The observation also applies to the features **Not Therapy** or **Get Therapy**, **Took Prescription/Counseling**. Therefore, we can partition the 10 features into four types, namely Care, Insurance, Therapy and Prescription. Future modeling can select the one feature from each group to improve the generalization of the model and avoid overfitting.

	Coefficient	Lower of CI	Upper of CI
Delayed Care	1.191	-0.117	0.313
Delayed or Not Care	**1.181**	**0.055**	**0.498**
Not Care	**1.512**	**0.046**	**0.343**
Private Insurance	-0.146	-0.111	0.133
Public Insurance	0.383	0.047	0.231
Uninsured	**0.780**	**0.083**	**0.331**
Not Therapy	**1.321**	**0.495**	**0.646**
Get Therapy	-0.012	-0.117	0.212
Took Prescription/Counseling	0.183	-0.285	0.211
Took Prescription	**0.739**	**0.027**	**0.497**

Table 3: Confidence Interval (CI) of the Linear Regression model coefficients.

Random Forest For the Random Forest Regression model, we use 100 tree estimators to ensemble the prediction and use squared error as the criterion for evaluating the quality of one split. For each estimator, we expand the tree until each leaf is pure without setting the maximum depth. This could lead to the results we get in Tabel 2 that the train RMSE is much smaller than the test RMSE, and the model may not be robust to out-of-distribution [4] data.

Gradient Boosting and Cross Validation For the Gradient Boosting method, we use 100 tree estimators, use squared error as the criterion for evaluating the quality of one split, and set the maximum depth as 3. We consider all the features during the split. We find that the performance of this approach is quite sensitive to one hyperparameter learning rate, so we do the cross validation to find the optimal learning rate. We scan the learning rate in the span of $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$ and apply 5-fold cross validation to the train set and these models. We report the mean squared error (MSE) and R^2 versus different learning rates in Figure 8 and find that the optimal learning rate is 0.05. Then we use this learning rate to train the final model and report the result in Table 2.



(a) (b)

Figure 8: Hyperparameter Tuning by Cross Validation.

5.2 Question 2: How people in different groups differ in terms of anxiety level

Modeling From the exploratory data analysis for question 2, we observed that people in different sex, education, race, and disability groups seem to go through different levels of anxiety. In order to statistically check this hypothesis, we first extract the data about different groups by sex, education, race, and disability. We then build separate linear regression models with and without l1 regularization to predict anxiety levels using features such as insurance, access to medical care, and most importantly, for each of the different groups, we included the one-hot encoded subgroups as our features. We normalized the feature columns, and dealt with missing

values. We shuffle the data and use 67% of the data as the training data set and 33% of the data as the testing set. We fit the data into the basic linear regression model with and with our l1 regularization. We choose to use l1 regularization to remove redundant features that have collinearities and to improve model stability.

The results of our different models are summarized in the following tables: Table 4, Table 5, Table 6, and Table 7. There is no significant difference in r-squared and root mean squared error with and without l1 for each groups. However, with l1 regularization do have impacts in the significant tests which we will discussed right after.

Model (By Sex)	Train R-Squared	Test R-Squared	Train RMSE	Test RMSE
Linear Regression (without L1)	0.976	0.958	5.15	4.99
Linear Regression (with L1)	0.956	0.937	5.22	4.77

Table 4: Modelling Results for People in the US Grouped by Sex

Model (By Education)	Train R-Squared	Test R-Squared	Train RMSE	Test RMSE
Linear Regression (without L1)	0.932	0.883	5.73	5.78
Linear Regression (with L1)	0.905	0.900	5.82	5.75

Table 5: Modelling Results for People in the US Grouped by Education

Model (By Race)	Train R-Squared	Test R-Squared	Train RMSE	Test RMSE
Linear Regression (without L1)	0.951	0.878	5.91	5.91
Linear Regression (with L1)	0.942	0.878	5.94	5.97

Table 6: Modelling Results for People in the US Grouped by Race

Model (By Disability Status)	Train R-Squared	Test R-Squared	Train RMSE	Test RMSE
Linear Regression (without L1)	1.00	0.990	6.14	4.33
Linear Regression (with L1)	0.998	0.996	6.18	5.21

Table 7: Modelling Results for People in the US Grouped by Disability Status

Inference From the modeling part, we observe that sex, education, race and disability features indeed obtain non-zero coefficients in the linear regression models. We wonder whether the coefficients are indeed non-zero, or it was due to sample noise.

To answer this question, we perform significance test on the coefficients for each model. The null hypothesis is H_0: the coefficients are zero. To perform the t-test, We first bootstrapped 1000 models, and calculate the 95% confidence interval for the coefficients. If the confidence interval doesn't contain 0, we can confidently reject the null hypothesis.

Model	Sex(Male)	Education(< high school)	Race (other races)	Disability (Without)
LR (without L1)	(1.73, 9.48)	(-1.40, 10.6)	(-3.46, -0.254)	(-19.7, 13.0)
LR (with L1)	(0, 2.05)	(0, 1.65)	(-2.03, -0.323)	(-13.8, 0)

Table 8: Significance test results for sex, education, race and disability status

One thing that we could observe from the significance test is that most of the features cannot pass the significance test when we are using ordinary least square (OLS) model, and we think the reason is because there are colinearity

between the features. If we apply L1 regularization (Lasso), we could pass the significance tests. This is mainly because Lasso forces some of the coefficients to zero when they don't add enough predictive power to the model. Therefore some of the colinearity was removed. Note that when we apply Lasso to the model, the model are no longer unbiased. This violent the assumption of significance test (the model is unbiased) and could introduce some inaccuracy in the test process. However, since we use a relatively small regularization coefficients, we assume the bias is small enough to be neglected.

Another thing we can observe is most of the feature coefficients align with our EDA, for instance, people with lower education level, from minority race group, with disability tend to have higher depression and anxiety level. However, one feature that surprises us is the sex feature. In our EDA, male should has lower depression and anxiety level compared to female, yet in the t-test with Lasso regression, we observe that the confidence interval of the feature Sex(Male) is strictly positive. We explain that again by feature colinearity, since male has -0.96 correlation with public healthcare and 0.73 correlation with private healthcare, we think the results of the significance test might be distorted by the colinearities. When we increase the regularization factor to around 0.1, the categorical feature 'Sex=Males' could be dropped and the model still performed well in making predictions. Thus, we cannot conclude that males are naturally less anxious or depressed than females. One possible reason for males having lower anxiety or depression rate is that male could have a higher possibility of accessing private health care, and people who have access to private health care are generally less anxious. Also, men are negatively correlated with reduced access to both general healthcare and mental healthcare while women are positively correlated to reduces healthcare.

6 Implementation of peer review feedback

We received many valuable suggestions in the peer review session, and we have incorporated many of the in our final analysis.

- **data collection and sampling** For data collection and sampling, we have removed the long and detailed introduction from the report to our analysis notebook.
- **data cleaning** We standardized our data cleaning procedure (merging data frames, drop/impute missing data, normalization etc), so that the whole pipeline is more coherent.
- **EDA** We used to put all the exploratory analysis in the report. Our classmates suggested that we only keep the plots that are most closely related to our research questions. Thus we only kept the most informative plots.
- **modelling** Previously we only used R square score to evaluate the model, and our classmates suggested using more different metrics for the models. Therefore we also included RMSE.
- **inference** We added more details to our inference so that people with less experience with those techniques could also read our report comfortably.

7 Discussion

7.1 Social Impacts and Ethical Concerns

Health Insurance Coverage One of the most important finding is that the health insurance coverage is highly correlated with the anxiety level, and it varies in different states. Most states along the coastline have low private health insurance coverage due to the high price of the insurance plan. The high price of the insurance is caused by the lack of provider competition. Our proposal to resolve the issue is to introduce policies to restore the competition and regularize the market. It could potentially be a fundamental solution to reducing social anxiety levels.

General Healthcare and Mental Healthcare From our EDA section, we observed that the reduced access to general healthcare rate is decreasing for all groups; however, the reduced access to mental healthcare rate remains. We strongly suggest that the US government should provide more sufficient mental healthcare to people in all group.

Sex We find a big difference between the anxiety levels of males and females in the exploratory data analysis. However, the significance test results suggested otherwise. When we look deep into the feature space, we noticed that this is because males have better access to healthcare and are more likely to be covered by private insurance compared to females. The colinearity distorted the significance test. This analysis gives insight into the social gender inequality of private health insurance and healthcare access during the pandemic. We suggest that the government should take action to provide sufficient healthcare to female and improve the quality of public healthcare.

Race From the exploratory and model inference analysis we could confidently conclude that people from other races other than Asian and whites have higher anxiety levels compared to other race groups. This might also

due to their low private health insurance rate and high reduced access to healthcare rate. We suggest that the government should take action to provide sufficient healthcare to minority groups.

Disability From the plots we made in the analysis, disabled groups seem to constantly have significant higher anxiety/depression levels, which may be because they are facing more challenges/inconveniences during the pandemic. Our suggestion to the government would be to take action to provide more accessible healthcare for those who are physically challenged.

Education level Groups with higher educational levels generally show lower anxiety than low educational levels. One of the possible reasons is that the high educational group has more information and is more confident about surviving the pandemic. However, as seen in the gender section, there still exists collinearity between education level and access to mental health care. Therefore, the educational level could be a director factor in the healthcare access, and then the healthcare access directly affects the anxiety level.

7.2 Future Work

7.2.1 Question 1: Highly-Correlated Factors to Predict Depression and Anxiety Leve

Modeling In this work, we build a model with all the available features and identify the collinearity between features through significant test. A meaningful future work includes using less features for modeling to see if the performance can be boosted.

Inference We propose an interesting hypothesis that low private insurance coverage is one of the fundamental factors to the social anxiety levels. We also propose a question about the causes of the low coverage in the states along coast. We will explore these questions with more experiments and data analysis in the future.

7.2.2 Question 2: How people in different groups differ in terms of anxiety level

Modeling Besides LR models, our team have also establish models like random forest. However, the R^2 and RMSE metrics did not improve and the models seemed to be over-fitting; thus we did not include them in the report. In the future with more data, we could possibly apply more complicated models with hyper-parameter tuning to improve the testing R^2 and RMSE. We could also use clustering to figure our the relationship between different subgroups.

Inference Our work provide insights on the correlation between subgroups and depression/anxiety level. However, we couldn't draw any causal conclusions like 'Reduced access to healthcare is one cause of high depression/anxiety level', thus we couldn't say 'If we increase access to healthcare, the depression and anxiety level could be reduced'. In the future more causal inference need to be done before taking actions based on those insights.

References

[1] Hoffer, E., Hubara, I., and Soudry, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 30, 2017.

[2] Los Angeles Times. Letters to the editor: Health insurance is expensive, but it's not the insurers' fault. `https://www.latimes.com/opinion/story/2020-11-15/insurance-isnt-why-healthcare-is-expensive`, 2020.

[3] Mazza, M. G., De Lorenzo, R., Conte, C., Poletti, S., Vai, B., Bollettini, I., Melloni, E. M. T., Furlan, R., Ciceri, F., Rovere-Querini, P., and Benedetti, F. Anxiety and depression in covid-19 survivors: Role of inflammatory and clinical predictors. *Brain, Behavior, and Immunity*, 89:594–600, 2020. ISSN 0889-1591. doi: https://doi.org/10.1016/j.bbi.2020.07.037. URL `https://www.sciencedirect.com/science/article/pii/S0889159120316068`.

[4] Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., and Lakshminarayanan, B. Likelihood ratios for out-of-distribution detection. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/1e79596878b2320cac26dd792a6c51c9-Paper.pdf`.

[5] Santomauro, D. F., Herrera, A. M. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., Abbafati, C., Adolph, C., Amlag, J. O., Aravkin, A. Y., et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic. *The Lancet*, 398(10312): 1700–1712, 2021.

[6] Xiong, J., Lipsitz, O., Nasri, F., Lui, L. M., Gill, H., Phan, L., Chen-Li, D., Iacobucci, M., Ho, R., Majeed, A., and McIntyre, R. S. Impact of covid-19 pandemic on mental health in the general population: A systematic review. *Journal of Affective Disorders*, 277:55–64, 2020. ISSN 0165-0327. doi: https://doi.org/10.1016/j.jad.2020.08.001. URL https://www.sciencedirect.com/science/article/pii/S0165032720325891.