Predicting Depression, Anxiety, and Stress

Raghav Rastogi, Kartik Ullal, Kreena Totala

Summary:

Diagnosing depression, stress or anxiety can be extremely difficult, as it can affect different people in different ways. In addition, there is currently little understanding of the universal predictors of them. There are not enough clear risk factors for mental health professionals to look for when diagnosing a patient. This makes it difficult to tell mental illness apart from normal levels of stress, sadness, and anxiety.

Our dataset contains responses from the Depression Anxiety Stress Scale (DASS), as well as demographic information for each of the participants. The DASS consists of 42 questions about lifestyle habits. Each question is presented with a 4-point rating scale to indicate how strongly the participant relates to the question. We also have columns for age, gender, education, screen size, family size, orientation, race, marital status and some personality traits like enthusiasm, extroversion, introversion, organized or not, etc. We have around 40 thousand entries in the dataset.

Our goal is to use the data from the DASS responses, combined with demographic information, to find possible predictors of depression, anxiety and stress. We hope to add to the growing research surrounding mental illness by identifying possible risk factors.

A related study done in 2020 used similar machine learning-based analysis to look for behavioral differences between anxiety and depression (Richter 2020). They ended up with a good prediction accuracy, and identified specific behavioral patterns that are unique to each illness. They are essentially trying to target the same problem that we are: the current diagnostic process for depression, anxiety, and mental illness in general. Their results, as well as our results, show that machine learning may be a helpful addition to the current diagnostic standard for mental illness.

In our study, we have used Python to implement data import, cleaning, visualizations and prediction. Based on our visualizations, we decided which predictors are highly correlated with depression, anxiety and stress. Using the predictors with the highest correlation, we built a model that can predict these mental states.

We then focused on personality traits marked from the participants as well as demographic information to see if certain traits have a high correlation with mental illness. We also found out if any two of anxiety, depression and stress have influence on the other one.

Our results showed that we were able to predict depression, stress, and anxiety to some degree by using only demographic information and personality traits.

Methods:

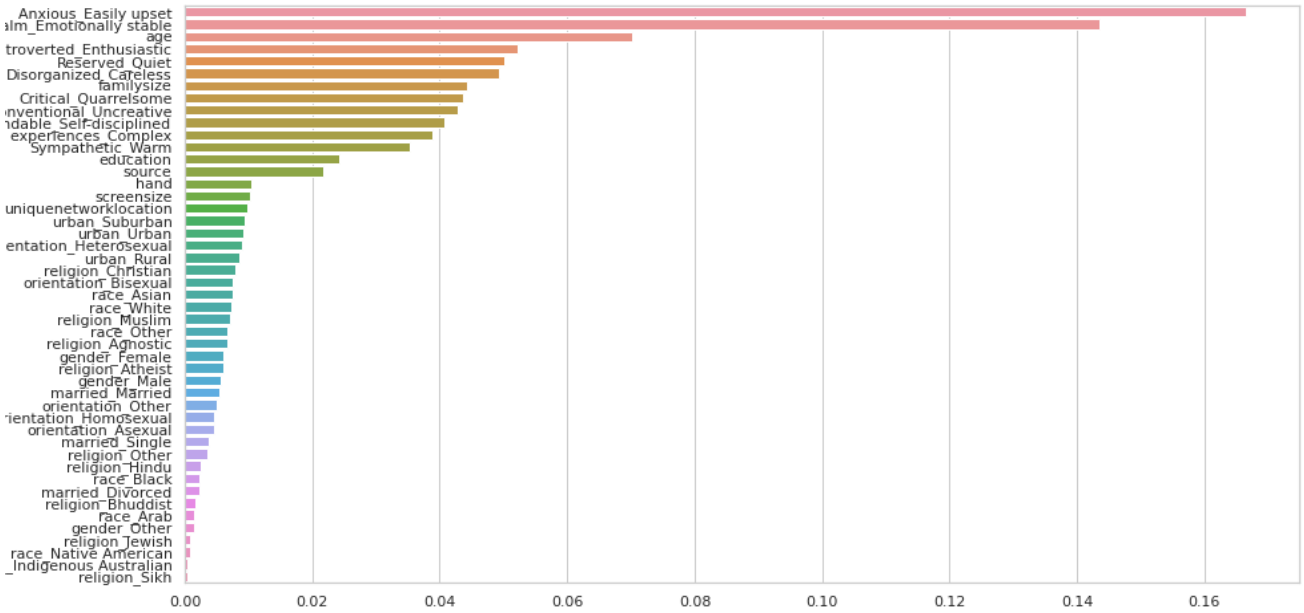
Data Cleaning and Transformation:

1. **Removing fake responses:** The DASS survey had a section to check the authenticity of, where 16 words were provided and every participant had to check words whose definitions they knew. However, 6 out of the 16 were made up words. We removed the rows of the participants that had checked for the fake words. Thus, we had a validity check on our dataset.
2. **Imputing Data:** Our dataset contained some NULL values in columns such as gender, race, orientation and religion. Since all these columns were nominal, we calculated the mode of each column and replaced the NULL values with the mode.
3. **Removed Outliers and Irrelevant Data:** We checked for outliers in our data. For instance, some participants had marked their age greater than 150, and family size greater than 100 etc. We removed rows with such values from our dataset. Moreover, the DASS survey website mentioned that the average time taken to complete the survey is 5-6 minutes. Since we had columns of the time taken to answer each question, we calculated the total time for the survey, and removed all rows with a total time less than 100 seconds. We also removed columns that did not contribute to the prediction. For example, time taken to finish the survey, voting status, and primary language.
4. **Created Separate Score Columns:** Our goal for this project is to predict depression, anxiety and stress scores. However, our dataset did not have separate columns for each participant. It only contains scores (1-4) for individual questions. According to the DASS paper, each score has a multiplier assigned to it. For example, score 1 was 0, score 2 was 2, score 3 was 4 and score 4 was 6. In addition, out of the 42 questions, there were 14 questions each for depression, stress and anxiety. According to that, we calculated total scores for depression, stress and anxiety for each participant. Once we calculated the total scores, we removed all columns with individual scores. So we were only left with the personality traits, demographic and bio-data and the scores for depression, anxiety and stress.
5. **One hot Encoding:** Most of our data was nominal. Using nominal data for a regression would not have given us desired results. To convert categorical data into numerical data, we used One hot Encoding. With one-hot, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector. All the values are zero, and the index is marked with a 1.

Machine Learning:

1) **Hold out set and standard training:** We removed 30% of the data at the start as a hold out data which will be used as a test set at the end. This will ensure that the RMSE error that we get at the end using the hold out data is reliable and the models are not overfitting the data. We then used the remaining 70% of the data and split the data to train (70%) and test(30%) which will be used to train and fine tune the models. [1] We also applied standard scaling to the data which converts the data to mean 0 and standard deviation as 1 for each attribute. This makes the model training more robust for prediction.

2) **Feature Extraction and dimensionality reduction**: Found the most important features using Random forest regressor (used forest.feature_importances) to reduce the dimension overload on the models and use only the most important attributes in the data. This helps to improve the accuracy of the models. Following are the important attributes that we found: Anxiety personality trait, Emotionally stable personality trait, age, Introversion and Extroversion, and so on as shown below. We only chose the attributes above "source" as they had considerable importance for prediction of the target variable. We discarded the other attributes.



3) **Fine Tuning the model:** We fine tuned the models using the following steps:

   a) We used parameter search in grid search cross validation (GridSearchCV) to find the optimum hyper-parameters for models - Lasso, ElasticNet, Ridge regression, Gradient Boosting regressor, LightGBM Regressor and XGBoost regressor.[2]

b) Following are the hyperparameters chosen for each model above which gave the least error on the GridSearchCV.[3]

   i) Lasso : Alpha = 0.001

   ii) Ridge : Alpha = 10

   iii) ElasticNet : Alpha = 0.001

   iv) GradientBoostingRegressor : learning_rate = 0.05, loss = 'huber', max_depth = 8, max_features =8, min_samples_leaf = 15, min_samples_split = 10, n_estimators =100

   v) LGBMRegressor: learning_rate = 0.1, n_estimators = 500, num_leaves = 4

   vi) XGBRegressor: learning_rate = 0.01, max_depth = 3, n_estimators = 3500

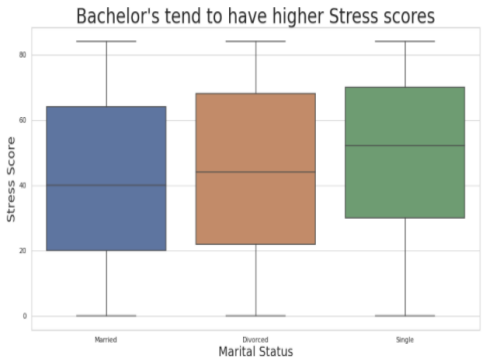
c) We used StackedCVRegressor which takes the predictions made by each of the models mentioned above, uses it as features and runs a meta regressor model (used XGBoost) to finally predict the target variable. This step makes predictions highly stable and reduces variance.[4]

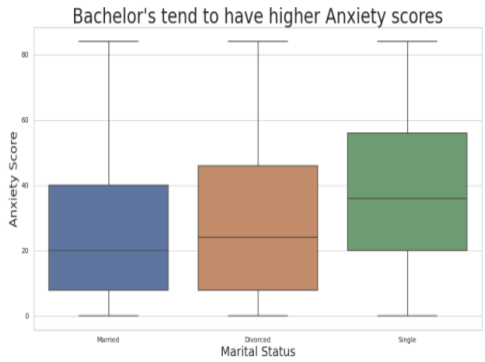
4) **Inverse-Standard Scaling and Prediction:** We applied inverse scaling to get the error which is scalable to the original range of scores from 0-84 for anxiety, stress and depression. We then fit the hold out data that we removed earlier to our final model and found the predicted target variable values. We then took these predicted values and found the RMSE error by comparing with the actual target variable values from the hold out set for each depression, anxiety and stress. [5]

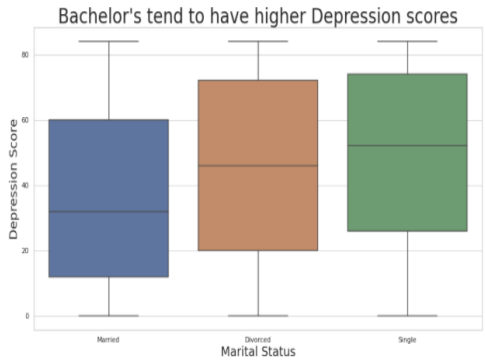
Results:

1) Exploratory Data Analysis

a) Marital Status : Married(blue), Divorced(orange), Single(green)

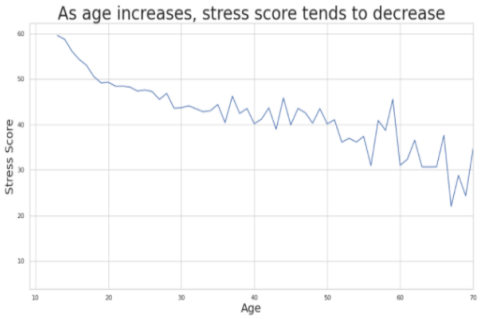


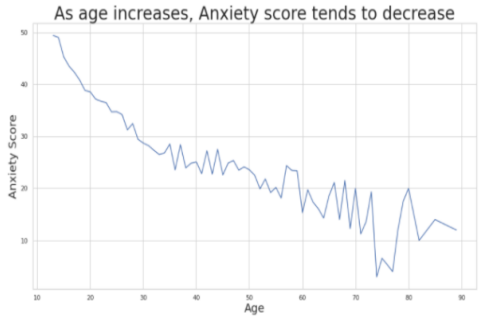


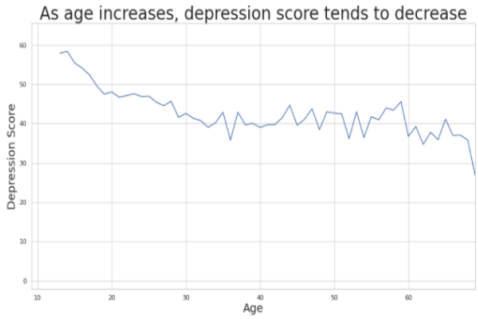


In the figure above, we can see that married people tend to have the least stress, anxiety, and depression while single people have the most.

b) Age : 10-90

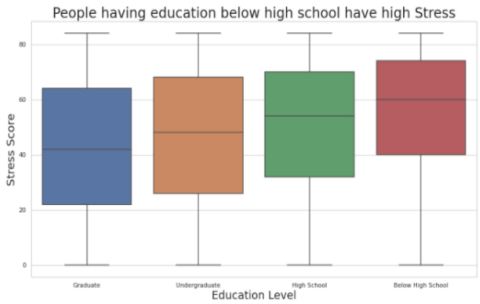


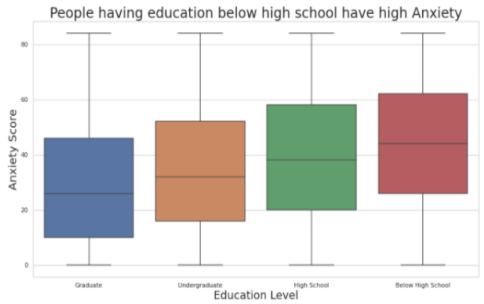


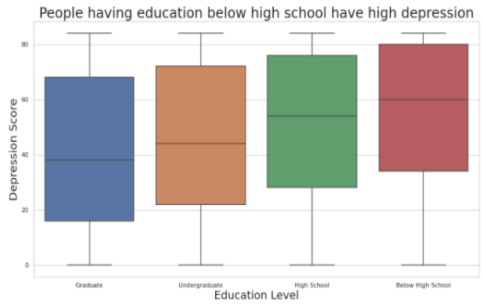


In the figure above, we can see that as age increases, stress, anxiety, and depression all decrease.

c) Education: Graduate(blue), Undergraduate(orange), High School(green), Below High school(red)

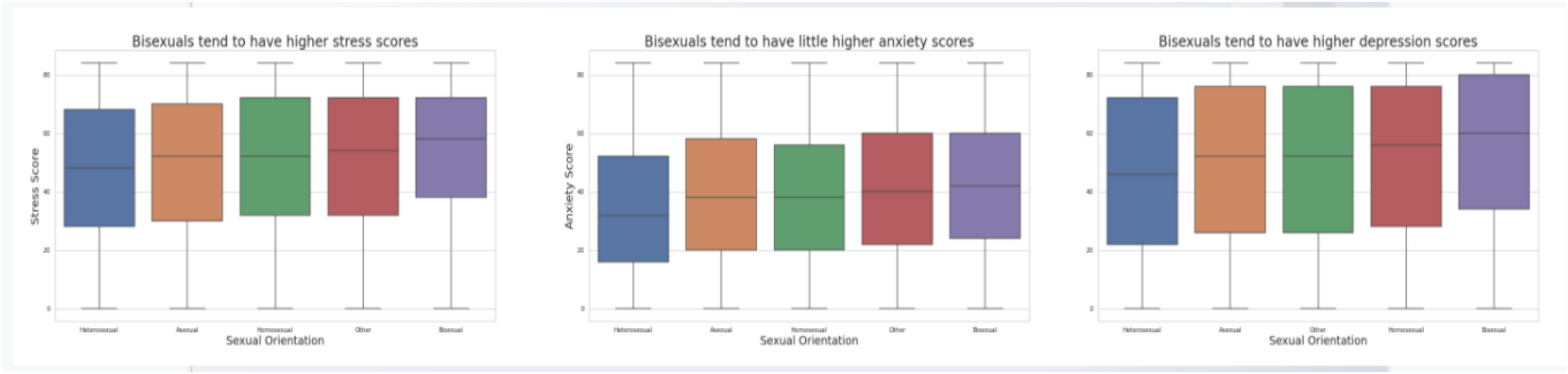






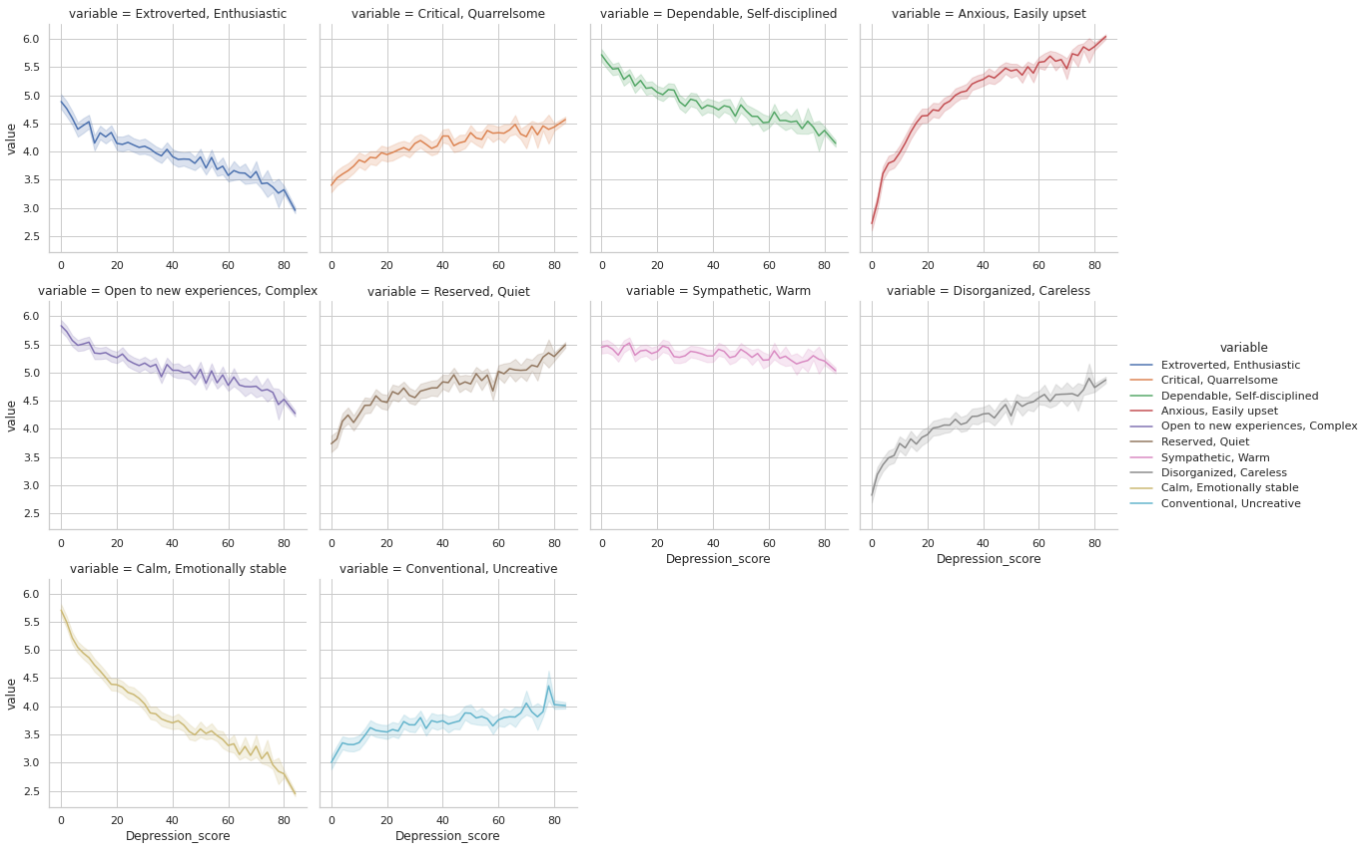
In the figure above, we can see that people with graduate level education tend to have lower stress, anxiety, and depression. We can see that there seems to be a pattern where higher education tends to lead people to have lower stress, anxiety, and depression.

d) Sexual Orientation: Blue(Heterosexual), Orange(Asexual), Green(Homosexual), Red(Other), Bisexual(Purple)



In the figure above, we can see that Bisexuals tend to have a higher stress, anxiety and depression scores. Also, heterosexuals rank the least among all with the lowest stress, anxiety and depression scores.

e) Personality traits:



The plot above shows correlations between various personality traits and depression scores. We can see the most extreme correlations seem to be the strong positive correlation with the anxious and easily upset personality type, and the strong negative correlation with the calm and emotionally stable personality type. The personality plots against anxiety and stress scores were very similar.

2) Final RMSE score:

Depression RMSE : 16.902
Stress RMSE : 11.602
Anxiety RMSE : 13.087

As we can see from above, using only the personality, emotional state and demographic information and background we were able to predict the score of people with good confidence on a 0-84 score range. This helps to assess people before asking very personal questions on a test which they might be reluctant to answer.

Discussion:

The results ultimately show that it is possible to predict depression, stress, and anxiety (with some error) based on just personality traits and demographic information. This is critical to the field of psychiatry and mental wellness in general, because it could lead to developing a better profile for those who suffer from mental illness. We are able to better identify potential "red flags" and hopefully help doctors in reaching the correct diagnosis. There is currently a lot of misdiagnosis in the mental health field, because mental illness is extremely difficult to diagnose. We are hoping that this project can help to further the search towards universal red flags for mental illness to help with the diagnostic process and ultimately lower misdiagnosis rates for these disorders.

In the future, this project can be expanded to include other mental illnesses, and their respective rating scores. For example, we can use scores from the mood disorder questionnaire (MDQ), as well as personality and demographic info of individuals with bipolar or borderline personality disorder. Additionally, we can look for trends specific to each of the disorders to make them easier to tell apart. A large issue with diagnosing mood disorders is that it can often be mistaken for a different mood disorder. We hope that these results can add to the recently growing field of using machine learning in clinical psychology.

Contributions:

Raghav Rastogi: Worked on Data Cleaning, Building the model for Depression, Data Search
Kartik Ullal: Worked on Data Cleaning and Visualizations, Building the model for Anxiety
Kreena Totala: Data Cleaning and Visualizations, Background Research, Build model for Stress

References:

Kaggle.com. 2021. *Predicting Depression, Anxiety and Stress*. [online] Available at: <https://www.kaggle.com/yamqwe/depression-anxiety-stress-scales> [Accessed 17 November 2021].

Lovibond, S.H. & Lovibond, P.F. (1995). *Manual for the Depression Anxiety Stress Scales*. (2nd. Ed.) Sydney: Psychology Foundation.

Richter, T., Fishbain, B., Markus, A., Richter-Levin, G. and Okon-Singer, H., 2020. Using machine learning-based analysis for behavioral differentiation between anxiety and depression. *Scientific Reports*, 10(1).

Appendix:

Code 1:

```
X_train, X_hold_out, y_train, y_hold_out = train_test_split(train, y, test_size = 0.3, random_state = 42)
y_hold_out_final = y_hold_out

X_train, X_test, y_train, y_test = train_test_split(X_train, y_train, test_size = 0.3, random_state = 42)
```

Code 2:

```
param_grid = [{'num_leaves':[4,40,100], 'learning_rate': [0.01,0.1,1], 'n_estimators':[5000,500]}]
grid_search = GridSearchCV(lightgbm, param_grid, cv=5,scoring='neg_mean_squared_error',return_train_score=True)
grid_search.fit(X_train,y_train)
print(grid_search.best_params_)
grid_search.score(X_test,y_test)
```

Code 3:

```
lasso = Lasso(alpha = 0.001)
ridge = Ridge(alpha = 10)
elasticnet = ElasticNet(alpha = 0.001)
gbr = GradientBoostingRegressor(learning_rate = 0.05, loss = 'huber', max_depth = 8, max_features =8,
                                min_samples_leaf = 15, min_samples_split = 10, n_estimators =100)
lightgbm = LGBMRegressor(learning_rate = 0.1, n_estimators = 500, num_leaves = 4)
xgboost = XGBRegressor(learning_rate = 0.01, max_depth = 3, n_estimators = 3500)
```

Code 4:

```
scvr = StackingCVRegressor(regressors=(ridge, lasso, elasticnet, gbr, xgboost, lightgbm),
                           meta_regressor=xgboost,
                           use_features_in_secondary=True)
```

Code 5:

```
y_pred = pd.DataFrame(y_pred, columns = ['target'])
y_pred_final = stdSc.inverse_transform(y_pred)
y_hold_out_final = y_hold_out
y_hold_out_final = stdSc.inverse_transform(y_hold_out_final)
np.sqrt(mean_squared_error(y_hold_out_final, y_pred_final))
```