Use case 2: Predicting mental health

Goal: The goal of this project is to predict whether an individual will suffer mental illness by looking at their health, lifestyle and socio-economic status.

Data: Downloaded from Depression Dataset (kaggle.com) https://www.kaggle.com/datasets/anthonytherrien/depression-dataset/data.

Question 1:
Use methods of your choice (e.g. exploratory data analysis, statistical methods, visualisations etc.) to extract useful insights from the data.

For the exploratory data analysis (EDA), here are the following steps taken:

1. **Univariate Analysis**: I have inspected each variable individually to understand its distribution and potential outliers. This will provide insights into the characteristics of each variable and help identify any extreme values or anomalies.
2. **Bivariate Analysis**: I have inspected the relationship between each variable and the target variable i.e History of Mental Illness. This analysis will allow us to understand how each variable is associated with the presence or absence of a History of Mental Illness.
3. **Multivariate Analysis**: I have inspected the interactions between different variables and how they collectively relate to the target variables. This analysis will help us uncover complex relationships and patterns that may not be apparent in the univariate or bivariate analyses. Techniques such as scatter plots, correlation matrices were created to gain deeper insights into the data

Narratives for including/excluding variables of choice;

In all these tree based classifiers a feature is assessed based on how much impurity decreases when that feature is used to split the data at each node. Feature importance check against Target ('History of Mental Illness') was performed using algorithms like Random Forest Classifier, XGB classifier, CatBoost classifier , AdaBoostclassifier, Gradient Boosting Classifier , Spearman Rank Correlation and here is the result which depicts the relevant features. However creating models with these features below didn't show enough improvement in True positives rather True Negatives performance got reduced . These features were further evaluated with Chi-square feature selection was also used through Selectkbest from sklearn which showed that having minimum features reduced model's performance on predicting negative class

Technique Used to study Feature impact	Relevant Features to Mental Illness Prediction
Random Forest Classifier	All features show little to more importance where **Income** and **Age** shows strong relevance to target

AdaBoost Classifier	Features like Income, Employment Status,Education Level, Smoking Status and Dietary Habits shows importance. These (Income and Employment Status can be good indicators to predict target as they look very relevant to target)
Gradient Boosting Classifier	Features like Employment Status, Income, Smoking Status, Education level and Age shows little to high importance . Among these Employment Status and Smoking Status shows high importance to Target
XGB Classifier	Features like Employment Status, Smoking Status, Income and Education Level shows little to high importance. Among these Employment Status shows high importance to Target
Spearman Rank Correlation	Statistically significant differences exist between these two features(Age, Income)
Pearson Correlation	Age and Income shows no positive or negative relationship to Target . Therefore no linear relationship exist between Age and Target as well as Income and Target

Narrative supporting the model/s of choice,

The Primary reason for choosing the CatBoost algorithm is that this technique is well suited for varied hyperparameters thereby allowing more parameters on which we can perform hyperparameter tuning on and have more control on enhancing model interpretability. Also this is well suited to handle class imbalances just like other boosting algorithms also allows to add class weights, auto update class weights, pos_weights which can lead to improved performance on the minority set. Algorithmic adjustments can help the model penalize more on missing out on the minority class prediction. Bagging algorithms were good at identifying True Negatives(Predicting no Mental Illness but they were really worse at predicting the True Positives(Predicting Mental Illness). Apart from boosting algorithms like SGD and CatBoost observed decent F1-score with slightly below par performance on predicting True positives(recall). Apart from boosting, I did see a decent model performance while using Linear-SVM classifier and Logistic regression with regularization.

An assessment of model performance

Model performance can be analyzed in three ways

- Catboost model performed best when we look at improving True Positives more than True Negatives thats have achieved 59% recall on Positive class and 55% recall on negative class
- Linear SVM performed well when we look at F1-score of 62.19%. However the recall of positive class performance is about 48% correct as compared to negative class performance of 67%.

Please refer the excel sheet where i have listed down all experiments i have run w.r.t different algorithms/Sampling strategies used (failure experiments) and the impact it had on Target

Assessment_Model_Performance

What are the limitations of your chosen approach? How can you improve the performance of your model?

- Although I had tried handling outliers with different methods like IQR, Z-score, Winsorization but they did not really help as we have only few numerical features like Age and Income and rest are categorical . Having a way to handle the different frequencies within a categorical group could really help solve class imbalance further .
- This Model has not seen the full range of hyperparameters due to lack of High RAM computer. Aim to run hyperparameter tuning on all boosting algorithms using bayesian optimization techniques (Automate hyperparameter tuning process)
- Predicting mental illness is a complex and multidimensional problem so it is not surprising that it is difficult to model. Perhaps exploring more model types which could reveal a strategy that was just right for this task.
- Trying a neural net with custom loss function
- Understanding the exact business weightage on each class
- A neural network, for example, could open up a huge range of possibilities that have not been explored in this effort. The drawback of neural networks is of course the lack of transparency in what features the model is using to make predictions.
- Oversampling ,undersampling techniques did not make a big difference in the impact .
- An obvious opportunity in continuing this work would be to do more evaluation of features to remove those that are not helpful and perhaps experimenting if other features may be added and prove helpful
- Adding more patient entries may also be valuable as there can be a lot of variation in medical histories and having more entries may better reveal complex patterns for the models to improve predictions with less features needed

Assessment of potential model biases

- Model might be biased towards individuals with Sedentary and Moderate Lifestyles due to the dataset as it contains more data points under these sub categories than Active lifestyle . Therefore this might result in incorrect predictions upon trying to predict the group of individual with Active lifestyle
- Similarly there are more married individuals in the extracted dataset, therefore model predictions might be biased towards married individuals than Widowed, Divorced or Single. Model might predict correctly when the individual is Widowed and might treat other groups as widowed thus leading to biasness

What are the limitations of your chosen approach? How can you improve the performance of your model?

- Although I had tried handling outliers with different methods like IQR, Z-score, Winsorization but they did not really help as we have only few numerical features like Age and Income and rest are categorical . Having a way to handle the different frequencies within a categorical group could really help solve class imbalance further .
- This Model has not seen the full range of hyperparameters due to lack of High RAM computer. Aim to run hyperparameter tuning on all boosting algorithms using bayesian optimization techniques (Automate hyperparameter tuning process)
- Predicting mental illness is a complex and multidimensional problem so it is not surprising that it is difficult to model. Perhaps exploring more model types which could reveal a strategy that was just right for this task.
- Trying a neural net with custom loss function
- Understanding the exact business weightage on each class
- A neural network, for example, could open up a huge range of possibilities that have not been explored in this effort. The drawback of neural networks is of course the lack of transparency in what features the model is using to make predictions.
- Oversampling ,undersampling techniques did not make a big difference in the impact .
- An obvious opportunity in continuing this work would be to do more evaluation of features to remove those that are not helpful and perhaps experimenting if other features may be added and prove helpful
- Adding more patient entries may also be valuable as there can be a lot of variation in medical histories and having more entries may better reveal complex patterns for the models to improve predictions with less features needed