Identifying Risk Factors for Major Depressive Disorder Through Predictive Modeling and Data Analysis

AMS 595 Project Report

Abby Bindelglass Jane Condon Nicholas Tardugno Sydney Walters-Diaz
Department of Applied Mathematics and Statistics, Stony Brook University

Abstract

Major Depressive Disorder (MDD) is a widespread mental health condition whose early detection is critical for improving clinical and public-health outcomes. Using the 2023 Mental Health Client-Level Data (MH-CLD) dataset, which contains over seven million client records, this project identifies key demographic, socioeconomic, and clinical predictors associated with MDD and evaluates a suite of statistical and machine-learning models for MDD risk classification. Following extensive preprocessing—including variable recoding, imputation, removal of redundant labels, and one-hot encoding—we conducted exploratory data analysis to characterize data quality, distributional patterns, and associations among predictors. To support interpretability and reveal overarching patterns, we used a variety of data visualizations such as heatmaps and histograms to highlight trends across demographic and clinical variables. Logistic and probit regression models revealed that co-occurring anxiety disorders, personality disorders, and substance abuse problems are strong positive predictors of MDD, while certain mental health diagnoses such as bipolar disorder and schizophrenia are associated with lower odds of an MDD diagnosis. To improve predictive accuracy, we implemented Random Forest, Stochastic Gradient Descent, LightGBM, and CatBoost classifiers. Comparative model assessment using ROC curves, confusion matrices, and Brier scores showed that the Random Forest classifier provides the best balance of predictive accuracy, calibration, and interpretability for this dataset. Feature importance results highlight that co-occurring mental health conditions, age, region, and employment status play the most substantial roles in predicting MDD. This study demonstrates that ensemble learning methods, paired with rigorous data preparation, can effectively support early identification of individuals at elevated risk for MDD, offering insights that may inform screening strategies and mental-health policy.

1. Introduction

1.1. Background Information

Major Depressive Disorder, also referred to as clinical depression, is a mood disorder characterized by feelings of persistent sadness and/or hopelessness. An individual diagnosed with this disorder may also experience sleep disturbances, lack of interest in normal activities, lack of energy, trouble concentrating, and suicidal thoughts or actions. There are many “risk factors” that may influence the likelihood of an individual developing this disorder, such as homelessness, unemployment, or being diagnosed with other mental health disorders. It is important for mental health professionals to be aware of these risk factors so that they can efficiently diagnose and provide these individuals with treatment, such as therapy or medications, before it is too late.

1.2. Problem Statement and Objectives

We are interested in examining what factors influence the likelihood that a client in the MH-CLD 2023 dataset receives a diagnosis of Major Depressive Disorder (MDD). We will explore demographic and socioeconomic factors in addition to co-occurring disorders as possible influences. Ultimately, we will compare the level of influence of the different explanatory factors on MDD in this population. This can serve to inform mental health professionals of the “risk factors” for an individual to develop MDD before symptoms become severe. It can also raise awareness and influence public policy decisions to reduce systemic problems, such as homelessness, high unemployment rates, and substance abuse, that may cause an individual to develop MDD.

2. Data

The dataset we will use is the Mental Health Client-Level Data (MH-CLD) 2023 public use files from the Substance Abuse and Mental Health Services Administration (SAMHSA). It includes demographics and mental health characteristics for clients using mental health and support services through state mental health agencies in the U.S. The dependent variable is a binary indicator showing whether the client has a MDD diagnosis or not. Some independent variables to be explored are age, sex, race, employment status, residential stability, co-occurring mental health diagnosis, education level, location (such as Midwest, Northeast, etc.), and substance use diagnosis.

3. Data Preprocessing

To prepare the data for statistical modeling and predictive modeling, we first preprocess the data to ensure that it is interpretable, numerically stable, and ready for model training.

3.1. Handling Missing Data and Special Codes

The MH-CLD dataset uses SAMHSA-specific numeric codes to indicate missing, unknown, or inapplicable responses, (e.g., -9, -8, 7, etc.). All such codes will be replaced with NAN and will later be imputed.

Co-occuring mental disorder flags (such as ANXIETYFLG, ADHDFLG, etc.) were preserved as binary numeric variables. Missing entries were assigned a value of 0, indicating ”disorder not reported.” Substance use variables (SUB and SAP) were also coded in a similar fashion.

3.2. Mapping Variables for Easier Readability

In general, the readability of large survey data is very poor. To make the data easier to interpret and understand, we will map our variables to the values/explanations provided in the codebook. The following variables were mapped to human-readable labels: age group, sex, education level, marital status, residential status, veteran status, employment status, ethnicity and race, region, substance use diagnosis, and co-occurring mental health disorder flags. These labeled columns are used for visualization and then removed from the numerical modeling dataset.

3.3. Creating Binary Indicators

To simplify interpretation and support multiple modeling approaches, we can construct the follow variables into binary indicators:

- **MDD diagnosis flag (MDD)**: takes the value 1 if a depressive disorder was reported.
- **HAS_SAP**: takes the value 1 if any substance–use problem was reported.
- **HAS_SUBSTANCE_USE**: takes value 1 if the `SUB` diagnosis category indicated an alcohol or substance–use disorder.
- **ANY_OTHER_MH_DISORDER**: takes value 1 if any co–occurring mental–health disorder flag was reported.
- **IS_HOMELESS**, **IS_VETERAN**, **IS_MARRIED**: binary indicators derived from their respective categorical variables.

3.4. Feature Selection and Cleaning

As the MH-CLD dataset is high-dimensional and contains a high number of unnecessary columns, we will only use the columns that are relevant to our analysis. We retain the following predictors:

- Demographic factors (age, sex, race, ethnicity)
- Socioeconomic factors (education level, employment status, marital status, residential status, and region)
- Clinical factors (substance use co-occurring mental health diagnosis)

3.5. One-Hot Encoding of Categorical Variables

All categorical variables were one-hot encoded using

```
pandas.get_dummies(..., drop_first=True)
```

to avoid collinearity. All human-readable label columns (ending in LABEL) were removed after encoding to retain only numeric features.

3.6. Imputation of Missing Values

To handle missing values and to ensure a complete feature matrix for model training, we impute the missing values as following:

- Categorical variables: Missing values are imputed with the most frequent category (mode imputation)
- Numerical variables: Missing values are imputed with the median value (median imputation)

3.7. Output of Data Preprocessing Function

We wrap all of this into a data preprocessing function which returns two different dataframes:

- `clean_df` : Used for descriptive statistics and data visualization
- `model_df` : Used for statistical modeling and for training all machine-learning models.

This preprocessing pipeline ensures that the data is clean, consistent, and analytically ready, and balances interpretability with the numerical precision required for modern machine-learning methods.

4. Exploratory Data Analysis

The exploratory data analysis (EDA) provides an initial understanding of the MH-CLD 2023 dataset and helps identify potential factors associated with Major Depressive Disorder (MDD).

4.1. Data Overview and Data Quality

After running the dataset through our data preprocessing function, we summarized our cleaned dataset to contain 7,035,641 client records and 57 variables. It should be noted that the dataset includes duplicates in both coded numerical variables and their corresponding label versions.

All in all, diagnostic variables and constructed binary indicators (e.g., HAS_SAP, HAS_SUBSTANCE_USE, MDD) have no missing values, indicating reliable reporting of disorder-related information. In contrast, several demographic and socioeconomic variables display a high percentage of missing values: educational level (EDUC) with about 3.6 million missing records (51%), employment status (EMPLOY) with about 4.3 million (61%), and veteran status (VETERAN) with about 3.9 million (56%). Despite these gaps, core predictor and outcome variables remain largely complete, ensuring the reliability of later analyses. The uneven pattern of missingness suggests that administrative fields related to socioeconomic status are less reliably captured, an important consideration for modeling and interpretation.

4.2. Descriptive Statistics of Predictors

The MH-CLD dataset contains a wide range of demographic, socioeconomic, and clinical predictors, most of which are encoded as ordinal or binary variables. Age is broadly distributed across adult ranges, with females slightly outnumbering males (53% vs. 47%). Clinically, the population exhibits substantial diagnostic burden: 76% have at least one additional mental-health condition, 47% have an anxiety-related disorder, and 36% present with substance-use problems. In contrast, disorders such as ADHD, conduct disorder, and dementia appear only rarely. Major depressive disorder (MDD) itself is present in approximately 27% of clients.

These distributions highlight the heterogeneity of the client population, characterized by high comorbidity and considerable variation in demographic and socioeconomic backgrounds. The strong representation of anxiety and other mental-health conditions aligns with elevated MDD rates, while disorders such as schizophrenia or bipolar disorder appear inversely associated with MDD coding, likely reflecting prioritization of a primary diagnosis. Together, these patterns establish the foundational structure of the dataset and motivate subsequent modeling decisions regarding which demographic, socioeconomic, and clinical features are most predictive of MDD.

4.3. Association Between Predictors and MDD

To evaluate how demographic, socioeconomic, and clinical factors relate to Major Depressive Disorder (MDD), we conducted Chi-Square tests, examined row-wise percentage distributions, and calculated Cramer's V effect sizes.

Across all categorical predictors, Chi-Square tests indicated statistically significant associations with MDD (all $p < 0.001$). However, given the extremely large sample size, statistical significance must be interpreted cautiously. Even very small differences between groups result in extremely small p-values. Therefore, we focus on effect sizes (Cramer's V, odds ratios) and practical magnitude rather than statistical significance alone.

The percentage distributions showed clear patterns. Females had a notably higher prevalence of MDD than males (32% vs. 21%). Certain age groups (particularly 18–24, 50–64) showed elevated MDD rates around 32–34%, compared to only about 5% in the 0–11 group. Individuals with higher education, full-time employment, or divorced/widowed marital status also displayed higher MDD proportions. Several co-occurring disorders—especially conduct disorder, bipolar disorder, schizophrenia/psychotic disorders, and trauma-related disorders—were linked to higher MDD prevalence (33–38%). Substance use problems were similarly associated with elevated MDD rates.

Cramer's V values showed that although all associations were statistically significant, most were small in magnitude. Age (0.196), education (0.148), and sex (0.123) showed the strongest effects, while variables like veteran status, race, and employment class exhibited very weak associations.

Overall, these results suggest that while many predictors differ across MDD status, only a subset—particularly age, sex, education, and certain clinical diagnoses—show meaningful effect sizes that may contribute more substantially to modeling and interpretation.

5. Data Visualization

5.1. Histograms and Bar Charts

In order to visualize the dataset, various methods of presentation were used. The most extensive were histograms and bar charts. Histograms were created in order to better visualize the proportion of MDD diagnosis for different groups. Interestingly, features with the highest importance in later models did not always have the highest proportion of MDD diagnoses. Stacked bar charts were used for non-binary variables and highlighted the proportion of MDD diagnoses for each category. This also provided a visual for the number of people in each category.

5.2. Heatmaps

Like stacked bar charts, heatmaps provide information about how proportions change over different categories. This is especially helpful for categorical variables that are also numerical, like age ranges and educational ranges. These visualizations provide a good background before using specific models in order to analyze the data. Specific methods of data analysis are used for different models as well. These include odds ratio graphs, ROC Curves, confusion matrices, and calibration curves.

6. Regression

We implement logistic and probit regression models to explore the relationship between demographic, socioeconomic, and mental health factors and the likelihood of MDD. Logistic and probit regression models are used because they are well-suited for binary outcomes. They give coefficient estimates to quantify the effect that a predictor has on the probability of diagnosis of MDD.

The models are estimated with maximum likelihood and use the same set of predictors so they can be directly compared.

6.1. Logistic Regression

Logistic regression was estimated to model the logarithmic odds of being diagnosed with MDD as a linear function of covariates. The model converged after seven iterations and produced a pseudo-R^2 value of 0.1573, displaying moderate explanatory power of the data.

Co-occurring mental health conditions were strongly associated with MDD. Anxiety disorder was the most influential predictor, with an odds ratio of 2.6. This means that people diagnosed with anxiety are more than twice as likely to be diagnosed with MDD, compared to those without anxiety, keeping the other variables constant. Diagnoses related to personality disorders, OR = 1.6, and substance abuse problems (SAP), OR = 1.3, are also positively associated with MDD risk. In contrast, some mental health diagnoses were associated with lower odds of MDD diagnosis. Bipolar disorder and schizophrenia corresponded to the most reduced odds of MDD, with ratios of 0.19 and 0.20, respectively.

Demographic and socioeconomic characteristics also contributed to MDD risk. Female sex and homelessness were associated with increased odds of MDD, with odds ratios of 1.5 for sex and 1.2 for homelessness. Age, marital status, and region showed a smaller positive association with MDD risk. Employment and veteran status showed negative associations with MDD risk.

The sample size was extremely large, so even small associations between predictors and MDD were detected as statistically significant, leading to very small p-values. Therefore, when interpreting the results, we consider the size and direction of the estimated effects, in addition to statistical significance. The findings indicate that both the possession of other mental health conditions and socioeconomic characteristics are associated with the likelihood of MDD diagnosis.

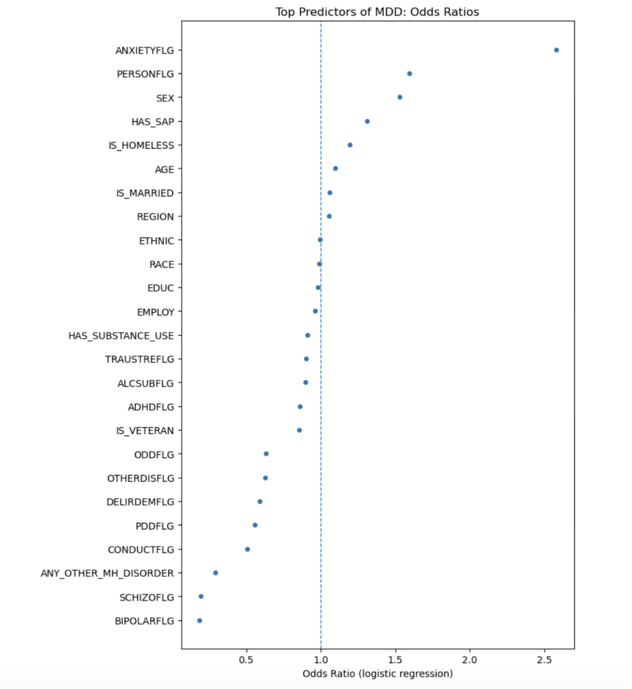

Figure 1. Odds Ratios for Predictors of MDD

6.2. Probit Regression

A probit regression model was fit with the same predictor set as the logistic model for robustness. The probit model converged after 6 iterations and achieved a pseudo-R^2 value of 0.1566, nearly the same as the logistic model. The results of the probit model were consistent with those of the logistic model. The direction of the estimated effects was the same across the models, indicating that the findings are not just due to the choice of model. Anxiety disorder, personality disorders, substance abuse problems, and sex were strongly associated with increased likelihood of MDD diagnosis. Bipolar disorder, schizophrenia, and the presence of other co-occurring mental health disorders were associated with decreased likelihood of MDD diagnosis.

6.3. Model Comparison and Predictive Performance

To compare the predictive performance of the two models, we used Receiver Operating Characteristic (ROC) curves. The Area Under the Curve (AUC) was very similar for the two models.

- Logistic regression AUC = 0.7671143499944019
- Probit regression AUC = 0.767049165121894

An AUC of approximately 0.77 suggests a good ability to distinguish between individuals with and without an MDD diagnosis and shows a substantial improvement over random classification. The ROC curves are nearly identical, meaning that the two models perform the same in terms of

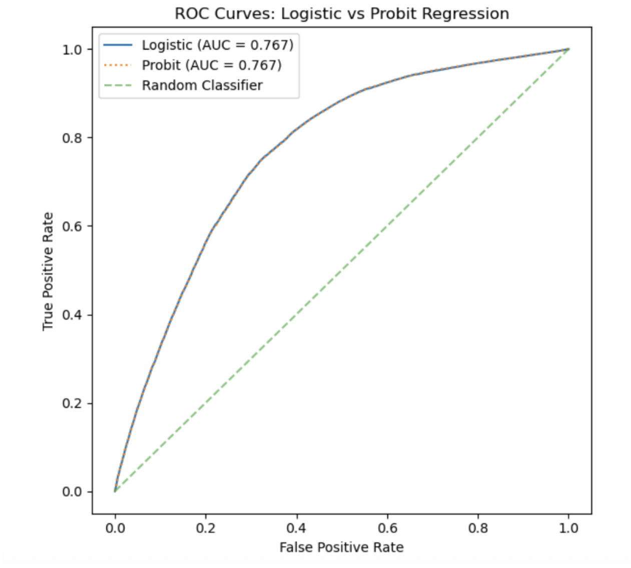

Figure 2. ROC Curves

prediction. Odds ratios make logistic regression easier to interpret, so it is our primary model, while probit regression is used for a robustness check.

7. Machine Learning Models

In addition to logistic regression, we will also use a variety of machine learning models to predict whether a patient will be diagnosed with Major Depressive Disorder based on clinical, socioeconomic and demographic factors. We deploy tree-based models such as Random Forest, Light Gradient Boosting Machine (LIGHTGBM), and CatBoost, as well as linear models such as Stochastic Gradient Descent (SGD) Classifier. We will compare each model based on several different performance metrics to determine which model would best assist mental health professionals in predicting whether an individual will develop Major Depressive Disorder.

7.1. Random Forest

Due to the fact that we are working with a large, high-dimensional dataset with various types of data and nonlinear interactions present, we will begin our task of predictive modeling with a Random Forest classifier. Random Forest is an ensemble learning method that constructs many decision trees during the training process and aggregates their predictions through majority voting. For noisy clinical datasets with millions of observations and with high potential for interactions (such as substance use and homelessness, homelessness and employment status, etc.), Random Forest may be superior to Logistic Regression in terms of capturing non-linear relationships between predictors and making more accurate predictions. We use the following parameters in our Random Forest Classifier:

- **n_estimators = 200**: Specifies the number of decision trees in the forest. Although our model may benefit from increasing the number of trees to 500 in terms of performance, it would be computationally expensive and thus, we must limit the number of decision trees to 200 for the sake of efficiency.
- **max_depth = 15**: This limits how deep each individual tree can grow, helping to prevent overfitting by restricting overly complex trees.
- **max_features = "sqrt"**: At each split, the model considers the square root of the total number of features, which allows the model to generalize better.
- **n_jobs = -1**: The model utilizes all available CPU cores to train trees in parallel, which can significantly speed up computation.
- **class_weight = "balanced"**: Since we have imbalanced data, this automatically adjusts class weights inversely proportional to class frequencies, which can help the model learn from imbalanced data.
- **random_state = 42**: We ensure reproducibility by setting the seed.

After training the model, we make predictions on the test data and obtain the following results:

Metric	**Value**
Accuracy	0.735
Precision	0.504
Recall	0.810
F1 Score	0.621
AUC	0.836
Brier Score	0.170

Table 1. Performance Metrics for the Random Forest Model

The model achieves 73% accuracy, which is reasonable for this type of dataset. This indicates that the model will correctly predict whether an individual has Major Depressive Disorder 73% of the time. As mentioned previously, we have limited the complexity of the model for the sake of efficiency, which may have limited our model accuracy. Thus, we must look at other performance metrics as well. The model achieves a much higher recall score, which is generally considered to be one of the most important performance metrics in clinical prediction tasks. The model correctly identifies 81% of real MDD cases, indicating that the model is proficient at minimizing false negatives. While the precision score (i.e., the rate of true positives) is relatively low, it is crucial to identify all positive cases, as it would be significantly more harmful to fail to detect a positive case than it would be to misdiagnose an individual with MDD. This is further illustrated using a confusion matrix:

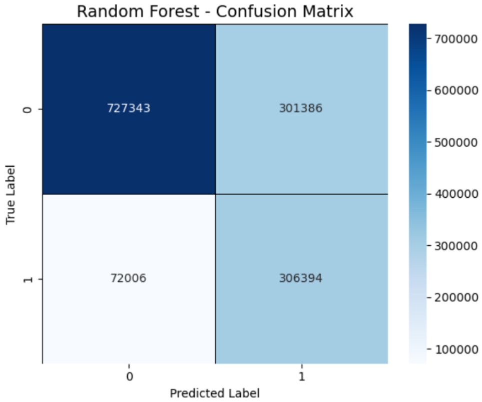

Figure 3. Calibration Curve - Random Forest Model.

We also evaluate calibrated probabilities using the Brier score and obtain a value of 0.17, indicating that this is a well-calibrated model. We can also look at the calibration curve, which further illustrates that this is a well-calibrated model:

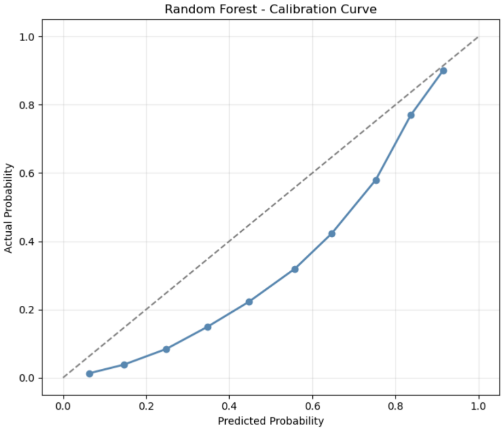

Figure 4. Calibration Curve - Random Forest Model.

7.2. Stochastic Gradient Descent Classifier

Unfortunately, Random Forest is not always ideal for handling sparse data with several encoded features, as we have in the MH-CLD dataset. Thus, we will also implement a Stochastic Gradient Descent (SGD) Classifier model. SGD is also significantly more efficient than both Random Forest and the traditional Gradient Descent algorithm, as it trains using one sample, rather than computing gradients over the entire dataset. The algorithm initializes model parameters, shuffles the dataset randomly, and iterates through each data point, where it computes a prediction, calculates the gradient of the loss function, and updates model parameters. Compared to models such as Logistic Regression and Random Forest, SGD is better equipped to deal with high-dimensional datasets containing millions of rows. We fit an SGD model using the following parameters:

- ○ **loss = “log_loss”**: Uses logistic regression loss.
- ○ **max_iter = 1000**: Maximum number of iterations (epochs) over the training data.
- ○ **tol = 1e-3**: Training stops early if improvement in the objective function is below this tolerance.
- ○ **random_state = 42**: We ensure reproducibility by setting the random seed.

After training the model, we make predictions on the test data and obtain the following results:

Metric	Value
Accuracy	0.755
Precision	0.571
Recall	0.352
F1 Score	0.435
AUC	0.776
Brier Score	0.160

Table 2. Performance Metrics for the SGD Model

While this model achieves a slightly higher accuracy than the Random Forest model, it obtains a poor recall score. While the precision score is slightly higher, it is more important to obtain a higher recall score in clinical prediction tasks. This can be further illustrated using a confusion matrix:

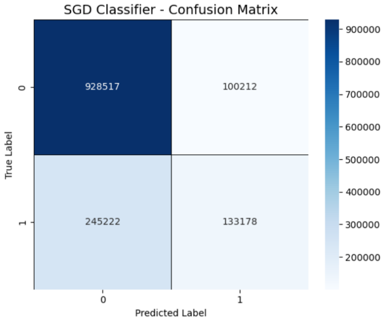

Figure 5. Confusion Matrix - SGD Model

As shown in the confusion matrix, the model is proficient at detecting individuals who do NOT have MDD, but it fails to detect many individuals who actually have MDD. This indicates that the model is biased toward predicting the negative class, which is the majority class in this dataset. Based on this information, this model is not ideal for our prediction task. SGD performs poorly on this dataset due to a variety of reasons, such as the presence of non-linear relationships between variables, class imbalance, and complexity of clinical data. This leads to problems such as underfitting and poor recall. A linear model such as this one is simply not complex enough to capture interactions between variables and is sensitive to multicollinearity, which is often present in datasets of this nature. To improve this model, we may want to manually include polynomial terms, pairwise interactions, and/or splines.

7.3. Light Gradient Boosting Machine (LIGHTGBM)

Since SGD has proven itself to be ineffective at handling data with complex nonlinear interactions and class imbalance, we also implement a Light Gradient Boosting Machine model, also known as LIGHTGBM. Unlike other tree-based models such as Random Forest, LIGHTGBM uses histogram-based gradient boosting, parallelized tree construction, and leaf-wise growth, making it significantly more efficient than other tree-based models. It is also highly optimized for sparse matrices and uses class-weighted boosting to reduce bias towards the majority class. We implement a LIGHTGBM model with the following parameters:

- **num_leaves = 64**: Maximum number of leaves in each tree. We choose 64 to improve accuracy without overfitting.
- **learning_rate = 0.05**: Determines how much each tree contributes to the final model. We choose 0.05, as we would like to maintain a balance between accuracy and efficiency.
- **n_estimators = 500**: Number of boosting iterations (trees). We limit this to 500 for the sake of computation time, but we could increase this number to improve accuracy.
- **min_child_samples = 50**: Minimum number of samples required to create a leaf. We set this to 50 to prevent overly specific, high-variance splits.
- **subsample = 0.8**: Fraction of training observations randomly sampled for each tree. We do this to introduce randomness, to reduce overfitting.
- **colsample_bytree = 0.8**: Fraction of features randomly sampled for each tree. This encourages diversity among trees and improves generalization.

After fitting and training the model, we make prediction on the test data and obtain the following peformance metrics:

Metric	Value
Accuracy	0.803
Precision	0.682
Recall	0.503
F1 Score	0.579
AUC	0.850
Brier Score	0.134

Table 3. LightGBM Model Performance Metrics

While this model has a high accuracy compared to the other models, accuracy alone is misleading due to the fact that the dataset is imbalanced. We obtain a reasonable precision score, but a poor recall score, indicating that the model is too conservative in predicting MDD. This means that the number of false positives is reasonable, but the number of false negatives is excessive. While it is an improvement from the SGD model, the model is missing approximately half of the individuals who actually have MDD, which is a clinical concern. This is further illustrated using a confusion matrix:

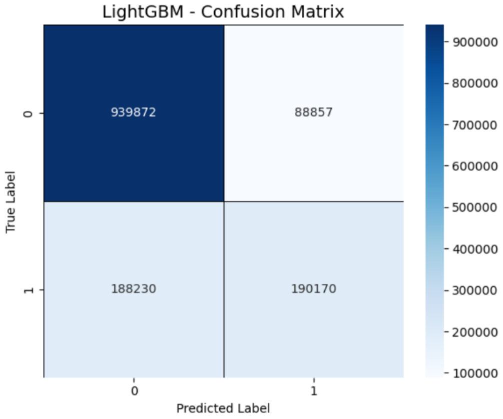

Figure 6. Confusion Matrix - LIGHTGBM Model

Although we obtain a poor recall score, it is important to note that this model is very well calibrated in comparison with the Random Forest model as exemplified by the Brier score and the calibration curve:

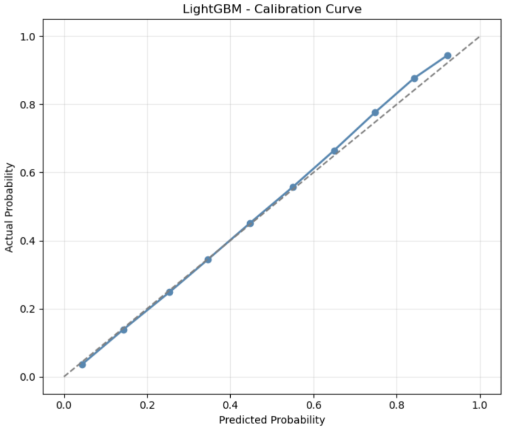

Figure 7. Calibration Curve - LIGHTGBM Model

This model's poor performance may be attributed to the fact that it uses a boosting algorithm, rather than bagging. Its objective function prioritizes accuracy, often at the detriment of other performance metrics such as recall, unlike models such as Random Forest. Models which use boosting create high-variance, high-confidence predictors, which means that if the model isn't very confident that an individual has MDD, the probability stays low and the predicted class will be non-MDD. In order to improve the recall of this model, we could lower the classification threshold to a smaller value such as 0.3, increase `class_weight` for the minority class, and construct deeper, less constrained trees. We could also use a custom loss function, such as focal loss.

7.4. CatBoost Model

Since our dataset contains a large amount of categorical predictors, we implement a CatBoost model. CatBoost is also more robust to noisy, sparse, and/or correlated features due to its use of symmetrical trees and gradient-based feature selection. When categorical features dominate, CatBoost often achieves a higher accuracy compared to LIGHTGBM. We implement a CatBoost model with the following parameters:

○ **iterations = 2000**: Number of boosting iterations (trees). We limit this to 2000 for the sake of reducing computation time.
○ **learning_rate = 0.03**: Step size for updating the model. We choose 0.03 to improve stability and prevent overfitting, without sacrificing efficiency.
○ **depth = 6**: Maximum depth of each decision tree. We set this to a moderate value of 6 to capture complex relationships without overfitting.
○ **od_type = "Iter"**: Enables early stopping when improvement is stagnant between consecutive iterations.
○ **od_wait = 40**: Number of iterations to wait for improvement before stopping early.
○ **task_type = "CPU"**: Model uses CPU training.
○ **verbose = 100**: Prints training progress every 100 iterations for monitoring.

After training the model, we make predictions on the test data and obtain the following performance metrics:

Metric	Value
Accuracy	0.801
Precision	0.676
Recall	0.502
F1 Score	0.576
AUC	0.848
Brier Score	0.135

Table 4. CatBoost Model Performance Metrics

The results are nearly identical to those of the LIGHTGBM model. We also obtain a similar confusion matrix:

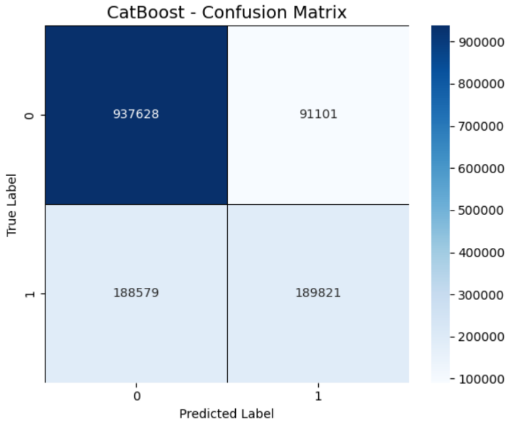

Figure 8. Confusion Matrix - CatBoost Model

Similar to the LIGHTGBM algorithm, the CatBoost algorithm also prioritizes accuracy, often sacrificing other metrics such as recall. Based on these results, it is evident that models which use a boosting algorithm are not ideal for scenarios where we would like to obtain a high recall score, such as clinical prediction tasks. However, the CatBoost model is very well calibrated, as shown by the Brier score and the calibration curve:

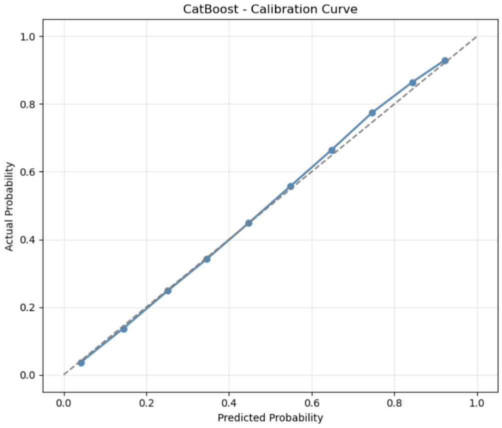

Figure 9. Calibration Curve - CatBoost Model

To improve recall, it may be beneficial to tune hyperparameters such as tree depth, learning rate, number of iterations, and regularization parameters, as CatBoost is known to be sensitive to hyperparameters.

7.5. Final Model and Results

Determining which model is superior depends on the scenario in which it is used. As SGD is clearly not an ideal model for this dataset, we must choose between Random Forest, LIGHTGBM, and CatBoost. Random Forest is the ideal model to use when:

- ○ Obtaining a high recall score is more important than obtaining a high accuracy score.
- ○ Stability and interpretability is important.

A Random Forest would be the ideal model for clinical decision support, where it is dangerous to miss a positive case, as that would lead to a patient not receiving the proper care. Random Forest could be used for tasks such as creating screening tools (to detect potential MDD cases early) and hospital readmission prediction.

A LIGHTGBM model would be the ideal model to use when:

- ○ Speed and efficiency is important, especially when working with large datasets.
- ○ Accuracy is prioritized over other performance metrics such as recall, precision, etc.

A LIGHTGBM model would be the ideal model for tasks such as real-time risk scoring and updating large EHR systems quickly.

A CatBoost model would be the ideal model to use when:

- ○ Obtaining well-calibrated probabilities and conservative estimates is important.
- ○ Categorical features dominate.
- ○ Overall accuracy is more important than other performance metrics such as recall, precision, etc.

A CatBoost model could be used for tasks such as insurance or policy risk scoring or public health population predictions.

Since our objective is to assist mental health professionals in detecting MDD early and to avoid false negatives, we choose the Random Forest model as our final model. From this model, we obtain the following feature importance plot:

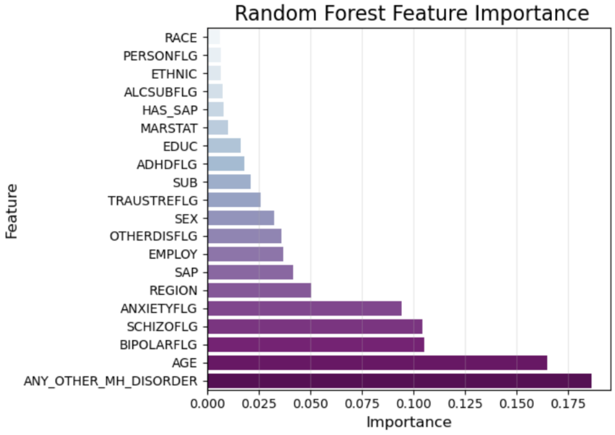

Figure 10. Feature Importance - Random Forest Model

Based on this model, the following predictors have the strongest effect on whether or not an individual will be diagnosed with MDD:

- ○ Being diagnosed with another mental health condition, such as bipolar disorder, schizophrenia, or anxiety.
- ○ Age
- ○ Region
- ○ Employment Status

8. Conclusion

In this project, we set out to identify the most influential risk factors associated with Major Depressive Disorder and to determine which predictive modeling approach is best suited for early detection within a large, complex, and highly imbalanced mental-health dataset. Through extensive preprocessing and exploratory data analysis, we established a clearer understanding of the demographic, socioeconomic, and clinical characteristics of the MH-CLD population. The data revealed substantial heterogeneity across diagnostic categories and meaningful patterns of missingness, emphasizing the importance of thoughtful feature engineering and imputation in producing reliable results.

Across both logistic and probit regression models, co-occurring mental health conditions—particularly anxiety disorders—emerged as the strongest predictors of MDD. Personality disorders and substance abuse problems also showed meaningful positive associations, while bipolar disorder and schizophrenia displayed negative associations after controlling for other covariates. These findings suggest that the clinical presentation of MDD is intertwined with a broader constellation of mental health concerns, supporting existing literature on diagnostic comorbidity. Socioeconomic factors such as employment status, homelessness, and marital status additionally contributed to MDD risk, underscoring the role of structural influences on mental health.

Performance comparisons across machine-learning methods demonstrated that ensemble tree-based models substantially outperform linear classifiers for this application. While the Stochastic Gradient Descent classifier struggled to capture nonlinearities and interaction effects, both LightGBM and CatBoost achieved strong predictive accuracy and calibration. However, the Random Forest model provided the most balanced performance, excelling in recall, AUC, and model stability while maintaining interpretable measures of feature importance. Because our research emphasizes early detection and minimizing missed cases, the Random Forest model is the most appropriate choice for clinical decision support and screening applications.

The feature importance analysis from the Random Forest model highlights several predictors with substantial influence on MDD classification: the presence of other mental health diagnoses, age, region, and employment status. These findings reinforce the complex interaction between clinical and socioeconomic factors in shaping mental-health outcomes. Ultimately, this study demonstrates the value of combining rigorous statistical modeling with modern machine-learning techniques to better understand—and more accurately predict—risk for Major Depressive Disorder. Future work may extend this analysis by incorporating temporal patterns, exploring causal inference methods, or integrating additional clinical text data to further refine predictive performance.

References

Bhuva, L. (2025, February 24). Stochastic Gradient Descent (SGD): A comprehensive guide to faster machine learning. Medium. https://medium.com/@lomashbhuva/stochastic-gradient-descent-sgd-a-comprehensive-guide-to-faster-machine-learning-b3afaf496a52

GeeksforGeeks. (2025, October 31). Random Forest Algorithm in Machine Learning. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/

Mayo Clinic Staff. (2022, October 14). Depression (major depressive disorder) – Symptoms and causes. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/depression/symptoms-causes/syc-20356007

Substance Abuse and Mental Health Services Administration. (2025, February 13). Mental Health Client-Level Data (MH-CLD). SAMHSA. https://www.samhsa.gov/data/data-we-collect/mh-cld-mental-health-client-level-data