Predicting Depression: A Comparative Analysis of Logistic Regression and Random Forest Models

Jonathan Ma

Introduction

Depression is a pervasive mental health disorder that significantly impacts the global population, influencing quality of life, productivity, and overall well-being. Characterized by persistent sadness, lack of interest, and a plethora of cognitive and physical symptoms, depression can severely impair an individual's daily functioning. Despite its prevalence, diagnosing depression remains a complex challenge due to its multifaceted nature and the variability of its manifestations among individuals. Traditional diagnostic methods rely heavily on subjective assessments and patient self-reporting, which are susceptible to biases and inaccuracies. Consequently, there is a pressing need for more objective, reliable, and scalable diagnostic tools.

Recent advancements in machine learning (ML) offer promising solutions to these challenges. Machine learning, a branch of artificial intelligence, utilizes statistical techniques to give computer systems the ability to "learn" from data without being explicitly programmed. In the context of depression, ML techniques can be employed to analyze vast arrays of data—from genetic profiles to behavioral signals collected via wearable technology—thus providing new insights into patterns that may not be visible to human observers.

This paper explores various machine learning techniques used to classify depression, focusing on the efficacy and applicability of different models. We discuss logistic regression and random forest classifiers in detail, as they represent both traditional and ensemble machine learning approaches, respectively. Logistic regression, known for its simplicity and interpretability, offers a probabilistic framework for modeling depression based on a linear combination of input features. In contrast, the random forest approach leverages an ensemble of decision trees to handle complex interactions between features, providing robustness against overfitting and enhancing predictive accuracy.

The selection of these models is guided by their distinct characteristics: logistic regression's transparency facilitates understanding and explaining the model's decisions, which is crucial in clinical settings, while random forests provide high accuracy and handle non-linear relationships effectively, making them suitable for complex diagnostic scenarios where multiple factors influence depression.

In reviewing these methodologies, the paper aims to highlight how machine learning can complement traditional diagnostic methods, leading to improvements in the early detection and classification of depression. Such advancements could revolutionize treatment approaches, personalize patient care, and ultimately improve outcomes for individuals suffering from this debilitating condition.

Literature Review

Depression is a major global health issue with significant societal impacts, affecting millions worldwide. The development of predictive models using machine learning (ML) offers promising avenues for early detection and personalized treatment strategies. This review synthesizes findings from recent studies that apply ML techniques to predict depression across various contexts and populations. A common thread among the studies is the diverse methodological approaches employed to harness the predictive power of ML. Zulfiker et al. (2021) and Dinga et al. (2018) exemplify the use of clinical, psychological, and biological data to train algorithms capable of identifying depressive episodes. The inclusion of broad data types, ranging from electronic health records (Nemesure et al., 2021) to social media activity (Liu et al., 2022), underscores the versatility and breadth of data that can inform ML models. Predictive accuracy is notably enhanced through the integration of multimodal data. For instance, Hong et al. (2022) demonstrated how smartphone sensor data could be utilized to predict depressive moods effectively, pointing to the potential of real-time, non-invasive monitoring systems. Several studies have focused on specific populations, enhancing the relevance and applicability of their models. For example, Haque et al. (2021) targeted child depression, while Shin et al. (2020) focused on postpartum depression, each adapting ML models to the unique characteristics and needs of these groups. Similarly, Su et al. (2021) tailored their approach to predicting depression in elderly Chinese populations, reflecting the importance of demographic-specific factors in ML accuracy. Despite the advancements, challenges such as data heterogeneity, model interpretability, and ethical concerns regarding privacy and consent remain prevalent. Addressing these issues, Gao et al. (2018) and Narayanrao & Kumari (2020) discussed the importance of ethical ML practices and the need for transparent, interpretable models that stakeholders can trust and understand. A significant advancement in the field is the use of ML to predict treatment outcomes. Lee et al. (2018) and Chekroud et al. (2016) explored how ML models could predict therapeutic outcomes, aiding in the personalization of treatment plans. This is crucial for improving treatment efficacy and patient outcomes in clinical settings.

Methods

The study utilized a comprehensive dataset targeting a diverse demographic cross-section. The data is a modified data set originally obtained from the NSDUH. The data is modified by approximately equalizing the two-subgroups (50% of depression vs. 50% of depress-free cases) to avoid the imbalanced data problem (i.e., there is about 10% of depression cases in the original data; when this is the case, most of the machine learning approaches may not work well). This dataset included variables such as age, gender, income, race, and several psychosocial factors, which previous research has linked to mental health outcomes. Data was anonymized and standardized to ensure privacy and consistency across measures. Prior to analysis, the data underwent several preprocessing steps to optimize its suitability for the logistic regression model. Categorical variables were encoded using one-hot encoding to transform them into a binary format, suitable for logistic regression analysis. Observations were balanced so that depressive cases and non depressive cases are equally weighted. The core of our analysis involved developing a logistic regression model to predict the probability of depression among individuals.

Logistic regression was chosen due to its efficacy in binary classification tasks and its ability to provide probabilities that an observation falls into a particular category, which is crucial for clinical decision-making. A random forest classifier was employed to leverage its robustness against overfitting and its capability to handle high-dimensional data. The model consisted of multiple decision trees, each trained on a random subset of the data and features, and the final output was determined by majority voting among all trees. Both models were evaluated using accuracy and recall to determine the overall effectiveness of each model in classifying depression correctly. Given the critical nature of diagnosing depression, a high recall was prioritized to minimize false negatives. For seeing which threshold value to choose, the geometric mean was chosen as a balancing metric, one that would find a decent value of recall without sacrificing too much accuracy.

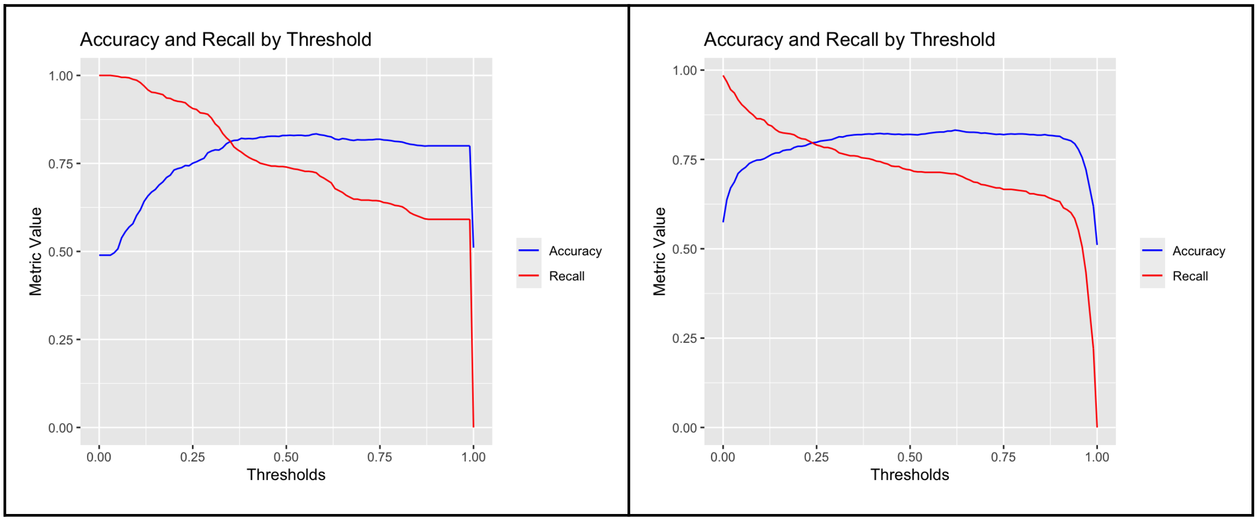
Findings

The initial phase of the analysis involved a detailed visual inspection of the covariates with respect to the major depressive episode Symptom Inventory (mdeSI). This visual inspection aimed to identify any clear patterns or anomalies that might influence the reliability and interpretation of subsequent statistical analyses. Through the use of bar graphs, we were able to assess the distribution and potential impact of these variables on depressive symptoms, ensuring that the data adhered to expected patterns without extreme outliers or evident biases. Following the visual assessment, we applied Chi-Square tests to confirm the relationships between each of the covariates and the incidence of depressive symptoms as measured by the mdeSI. The Chi-Square test, being particularly suited for categorical data, was utilized to ascertain if the observed distributions of depressive episodes across different categories were due to chance or if they exhibited significant associations.

The results from the Chi-Square tests provided a strong statistical foundation for the relationships observed during the visual inspections. Several low significance variables were added to provide noise to the model and ensure against overfitting

Variable	Chi-Square	p-value
Gender	622.36	2.2e-16
Age	241.95	2.2e-16
Race	45.164	3.675e-09
Income	15.683	0.001317

A decision threshold was decided to start at 0.01 and increment by 0.01 intervals until 1.0. This threshold would be iterated through all accuracy and recall rates, and the balance between accuracy and recall was determined by the geometric mean that balanced the highest recall rate. In this stage, for the logistic model, a threshold was chosen to be 0.29 and for the random forest, a threshold was chosen as 0.07, the figures are shown below.

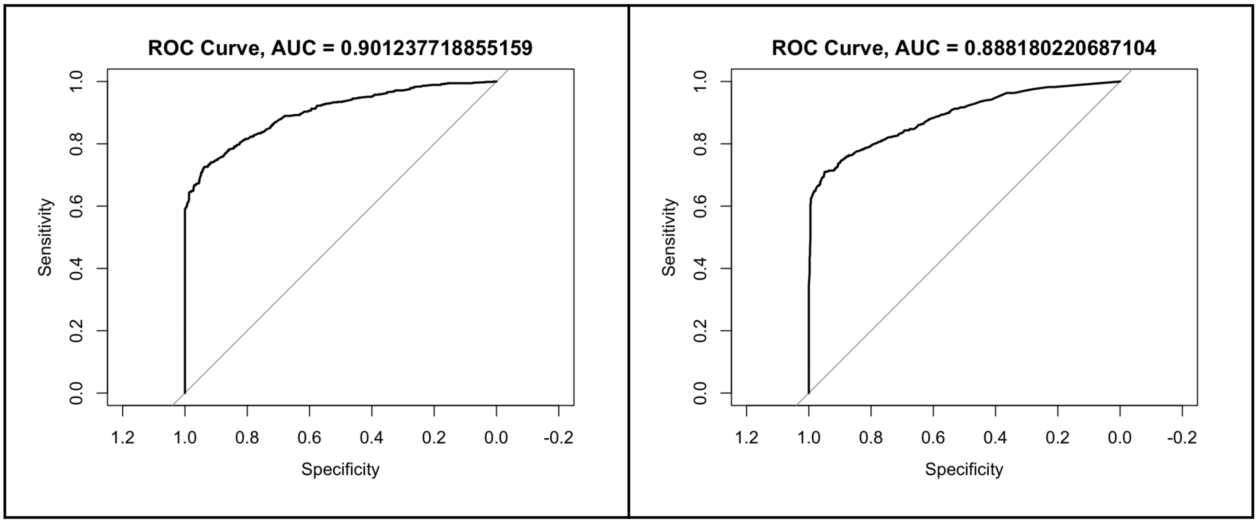


The logistic classifier model correctly predicts whether an individual is depressed or not about 78% of the time. The relatively moderate accuracy indicates that there might be significant overlap in the characteristics of the classes, or the model may not capture all the complexities of the data. The model has a very high recall, which means it successfully identifies about 88.96% of all actual cases of depression. High recall is particularly valuable in medical or critical fields where missing a case (false negative) can have severe implications, such as failing to diagnose a patient who is actually depressed. The AUC value is above 0.9, which indicates a good ability to distinguish between the classes (depressed and not depressed). A higher AUC value generally suggests that the model has a better overall performance at various threshold settings.

The random forest model shows a slightly lower accuracy (73.87%) compared to the logistic model, indicating worse overall prediction correctness. This suggests that the random forest might be considering too many patterns in the data. While slightly lower than the logistic model, the recall rate (88.28%) of the random forest is also high, indicating it effectively identifies a large majority of actual depression cases. Although it misses slightly more true cases than the logistic model, it still provides a robust performance. The AUC (0.8882) is slightly lower than that of the logistic model but still indicates good discriminatory ability. The difference in AUC between the models is minimal, suggesting that both models are nearly equally capable of distinguishing between the classes. The following information has been summarized below:

	Threshold	Accuracy	Recall	AUC
Logit Model	0.29	78%	88.96%	0.9012
Random Forest	0.07	73.87%	88.28%	0.8882

The two ROC curves are shown as well, left being the logistic model and right being the random forest model.



Discussion

Both models demonstrate good recall, which is crucial for applications where missing an actual positive case has serious consequences. However, there is a slight trade-off between recall and accuracy. Ultimately, the logistic model was chosen due to the importance of higher recall and area under the curve. For critical health applications where the priority is to minimize false negatives, both models do well, but the logistic model does slightly better in terms of recall. If the application also needs to consider the balance of model complexity and interpretability, logistic regression might be preferred. The generalizability of the models to unseen data or different populations is crucial, particularly if the training data doesn't represent the broader population. The quality and diversity of the data used to train the models significantly impact their accuracy and reliability. If the dataset lacks diversity or has biased sampling, the model predictions could be skewed, affecting individuals from underrepresented groups. The choice of features and their interactions might not have been fully explored, potentially omitting important predictors of depression. Judging from the aspect of specific number of trees

Exploring more complex feature interactions and incorporating domain expertise in mental health could uncover more significant predictors of depression. Techniques like Principal Component Analysis (PCA) for dimensionality reduction or exploring non-linear feature interactions might reveal hidden patterns. For logistic regression, refine the threshold selection by using techniques such as the ROC curve analysis to identify a threshold that balances sensitivity and specificity based on the clinical or practical significance of false positives versus false negatives. Techniques like Youden's Index or cost-sensitive learning could be more systematically employed. Techniques such as cross-validation could be used to determine the optimal number of trees and depth. Grid search or random search can be used to experiment with different configurations (like max depth, min samples split, and max features) to find the best settings for the random forest.

These models can be deployed in a controlled environment where their predictions can be monitored over time against actual outcomes. This continuous monitoring can help in detecting drifts in model performance and can be crucial for recalibrating or retraining models to adapt to changes in data patterns.

Conclusion

In this study, we evaluated the effectiveness of logistic regression and random forest models in predicting depression, an undertaking that highlights the intricate balance required in model performance metrics. Our findings demonstrate that both models perform commendably, with each showing distinct advantages that cater to different operational or clinical needs.

The logistic regression model, with its accuracy of 78%, proves especially valuable in scenarios where the cost of missing a true case of depression is high. Its ability to correctly identify a significant majority of depressive instances, coupled with a remarkable AUC of 0.9012, underscores its utility in clinical settings where high sensitivity is paramount. This model's relative simplicity and interpretability further enhance its appeal, providing clear insights into the factors influencing its predictions, which is crucial for clinical decision-making and patient communication.

The random forest model exhibits a slightly worse overall accuracy and a comparable AUC, suggesting its proficiency in handling more complex patterns within the data. Although it slightly trails the logistic regression model in AUC, its performance remains robust, making it suitable for applications where a balance between sensitivity and specificity is necessary. The random forest model's ability to manage non-linear relationships and interactions among features makes it a potent tool for complex diagnostic challenges, albeit at the cost of increased computational demand and reduced interpretability.

Both models demonstrate that machine learning can significantly aid in the early detection and diagnosis of depression, paving the way for timely and more effective interventions. Future work should focus on refining these models through advanced feature engineering, optimized model tuning, and continuous validation on diverse datasets to enhance their accuracy, reliability, and applicability across different populations. This approach not only improves the models' predictive power but also their adoption in real-world settings, ultimately contributing to better healthcare outcomes in the realm of mental health.

References

- Zulfiker et al. (2021) - Explored six different machine learning models to detect depression. Link
- Dinga et al. (2018) - Used clinical, psychological, and biological data to predict depression courses. Link
- Haque et al. (2021) - Employed Pearson correlation to predict child depression from survey data. Link
- Nemesure et al. (2021) - Developed stacked machine learning models to predict depression from EHR data. Link
- Cho et al. (2021) - Created a predictive model for community dwellers' depression using machine learning. Link
- Hong et al. (2022) - Proposed features based on multimodal sensor data for predicting depressive mood using smartphones. Link
- Shin et al. (2020) - Aimed to develop predictive models for postpartum depression using machine learning. Link
- Liu et al. (2022) - Reviewed machine learning methods for predicting depression from social media data. Link
- Su et al. (2021) - Used machine learning to predict depression in the elderly in China. Link
- Gao et al. (2018) - Reviewed machine learning applications in major depression classification and prediction. Link
- Narayanrao & Kumari (2020) - Analyzed different machine learning algorithms for depression prediction. Link
- Mariñelarena-Dondena et al. (2017) - Conducted a comparative study of machine learning approaches based on language usage for depression prediction. Link
- Kessler et al. (2016) - Employed machine learning methods to predict the persistence and severity of major depressive disorder. Link
- Lee et al. (2018) - Used machine learning to predict therapeutic outcomes in depression. Link
- Chekroud et al. (2016) - Utilized machine learning to aid prediction of clinical remission for antidepressants. Link
- Priya et al. (2020) - Predicted anxiety, depression, and stress using machine learning algorithms. Link
- Choudhury et al. (2019) - Used machine learning to predict depression among Bangladeshi undergraduates. Link
- Yadav et al. (2020) - Explored various algorithms for predicting depression from routine survey data. Link
- Sajjadian et al. (2021) - Reviewed machine learning in predicting depression treatment outcomes. Link
- Sau & Bhakta (2017) - Developed a predictive model to diagnose anxiety and depression in the elderly using machine learning. Link