UC DAVIS
GRADUATE SCHOOL
OF MANAGEMENT

Machine Learning for Early Alzheimer's Risk Detection:

A Data-Driven Approach Using Biometric and Biographical Features

Anakin Liu, Jerry Xia, Ruiyuan Yang

Executive Summary

This study investigates how machine learning techniques can be used to predict the risk of Alzheimer’s disease by utilizing easily accessible biographical and biological metrics. By utilizing the Alzheimer’s Prediction Dataset (Global) (Panday, 2023), which includes over 74,000 samples with 23 features, the goal was to create a predictive model that supports early risk assessment without the need for extensive neurological examinations. The dataset was preprocessed to address missing values and normalize continuous variables as needed. Categorical variables were encoded, and feature selection was performed using Lasso Regression to pinpoint the most significant predictors while reducing multicollinearity.

The research utilized several Decision Tree-based classifiers, such as Random Forest and Boosting, with Logistic Regression serving as a baseline model. Each model was trained and fine-tuned through hyperparameter optimization to improve predictive performance. The models were evaluated using AUC-ROC, accuracy, and the Precision-Recall Curve to measure classification effectiveness. The top-performing model achieved a Precision of 100%, an overall accuracy of about 60%, showcasing its capability to prevent falsely labelled Alzheimer’s cases. Feature-importance analysis indicated that Age, APOE-ε4 allele, and family history were the key factors influencing the model’s decisions.

The results of this study underscore the promise of machine learning in improving early Alzheimer’s risk assessment through non-invasive and readily available patient data. By incorporating predictive analytics into standard healthcare evaluations, medical professionals can identify individuals at risk sooner, facilitating timely interventions and personalized care strategies.

Background

Alzheimer’s disease is a progressive neurodegenerative disorder that impacts millions of people around the globe, creating significant challenges for healthcare systems and society. Early detection is crucial for effective intervention, but traditional diagnostic methods often involve costly and time-consuming neurological evaluations, such as brain imaging or cerebrospinal fluid analysis

(Jack et al., 2018). There is increasing interest in using machine learning (ML) to create predictive models based on readily available patient information, which could facilitate initial risk assessments without the need for specialized medical tests. Research has shown that ML techniques, especially decision tree-based models like Random Forest and Boosting, are effective in classifying neurodegenerative diseases (Falahati et al., 2014). Moreover, Lasso Regression has been commonly employed for feature selection, aiding in the identification of the most important predictors while minimizing noise and multicollinearity (Tibshirani, 1996). The main goal of this study is to evaluate the performance of these models in differentiating between individuals at a higher risk of developing Alzheimer's and those with a lower likelihood. If successful, it could act as early warning systems, allowing healthcare professionals to suggest further neurological evaluations and preventive measures before irreversible cognitive decline occurs.

Methodology

a. Data Preprocessing

Given the dataset's completeness and balance, as shown in Figure 1, no data cleaning process is needed. Lasso was applied for feature selection, identifying the most relevant predictors while eliminating redundant variables.

b. Model Selection and Training

A combination of classification algorithms was employed to assess predictive performance. Logistic Regression was used as a baseline model due to its interpretability and well-established performance in binary classification tasks. Decision tree-based models, including Random Forest and Boosting algorithms, were then implemented to capture complex interactions among features. Random Forest was chosen for its robustness in handling high-dimensional data, while Boosting methods were applied to iteratively refine weak classifiers and enhance overall prediction accuracy.

c. Hyperparameter Optimization

To achieve optimal model performance, hyperparameter tuning was conducted using Bayesian optimization. Since misclassification of non-Alzheimer's cases as positive could lead to unnecessary, costly, medical procedures for patients, minimizing false positives was a primary objective. The best model selection decision threshold was adjusted to 0.8 accordingly. Conservatively, only cases with a high probability of Alzheimer's were classified as positive, significantly enhancing model precision in the cost of poor recall. The impact of this adjustment was evaluated through confusion matrices and precision-recall metrics, ensuring that the model remained clinically viable while reducing false alarms.

Results and Analysis

During model development, our initial strategy for feature selection employed Lasso Regression as in Figure 2. However, we discovered that Lasso aggressively filtered out over 30 variables—primarily due to the presence of numerous one-hot encoded country features. The correlation between them rendered Lasso less effective for our dataset, potentially discarding informative predictors.

Subsequently, we established a baseline using Logistic Regression. As shown in Figure 3 and Figure 4, the model achieved an overall accuracy of 71.43%, with similar Precision and Recall. While the baseline provided interpretability, its performance lagged behind tree-based methods. We then pivoted to Decision Tree–based algorithms, starting with a Random Forest model. Despite extensive hyperparameter tuning, including the application of *GridSearch Cross-Validation* and *Randomized Cross-Validation*, the Random Forest model's accuracy plateaued at around 73%, as shown in Figure 5, with minimal gains in other key metrics.

During the course of the training process, our focus shifted toward reducing, even eliminating false positives (FP). The main rationale behind this shift was 1. We noticed a plateau in accuracy no matter the approach we take in hyperparameter optimizations or alternative model usage, and 2. We recognized that there is a strong incentive for clinicians to filter out Alzheimer's patients from the rest with maximum confidence. This will help both the hospital to save cost on diagnostic tests, while helping patients significantly reduce spending. Note that the reverse approach would not have the same effect: if we devised a model that eliminates false negatives, the direct outcome is that the model will filter out patients that are not inflicted with Alzheimer's disease with 100% precision. However, given the premise that all subjects tested by the model are present at the hospital likely because they show neurodegenerative symptoms, they will likely still seek other diagnostic tests even if the model categorizes them as non-Alzheimer's with 100% precision. This FN-eliminating model will therefore not achieve the same impact as a FN-eliminating model. To this end, we investigated four distinct strategies to eliminate FP:

1. **Custom Threshold Adjustment:** We increased the decision threshold (e.g., 0.8) so that only instances with a very high probability of Alzheimer's were flagged positive. This adjustment effectively reduced FP by making the classifier more conservative.
2. **Custom Cost Function:** We designed a cost function that heavily penalized FP, thus steering the optimization process toward configurations that minimized FP.
3. **Class Weight Adjustments:** By increasing the weight of the negative class during training, the model was encouraged to avoid false alarms.
4. **Systematic Threshold Tuning:** We performed an exhaustive search over a range of thresholds to identify the point where FP was eliminated while retaining as many true positives (TP) as possible.

The iterative exploration that mixed and matched the above methods revealed that while the Random Forest model struggled to push accuracy beyond 73%, a fine-tuned RF configuration using a decision threshold of 0.8 showed promise in reducing FP. Building on these insights, our champion model was ultimately developed using a Gradient Boosting classifier optimized with Bayes' search. This model was tuned with a decision threshold of 0.86 and applied a 2× sample weight on the negative class to penalize misclassification of negative results. As shown in Figure 6, this configuration achieved a 0 FP result on the test set, thereby enhancing the model's clinical viability for early Alzheimer's risk detection.

Recommendations

Model Enhancement & Performance Improvement

1. **Feature Selection Enhancement**：**SHAP (Shapley Additive Explanations):** Unlike traditional methods such as Lasso, SHAP enables both global and local feature importance analysis. **Recursive Feature Elimination (RFE):** This technique removes the least significant features, refining the feature space while maintaining the most informative variables.
2. **Classification Performance Enhancement**：**XGBoost:** Implements efficient handling of missing values, L1 and L2 regularization to prevent overfitting, and parallel processing to improve training speed. **LightGBM:** Uses a leaf-wise split mechanism instead of a level-wise approach, making it computationally efficient and highly scalable. This is particularly beneficial for high-dimensional data, reducing training time while maintaining accuracy.

   Deep Learning Approaches (MLPs, CNNs): Multi-Layer Perceptrons (MLPs) can capture complex patterns in tabular data, while Convolutional Neural Networks (CNNs) are well-suited if medical imaging data (e.g., MRI scans) is incorporated in the future.
3. **Threshold Optimization**：**Precision-Recall Curve Analysis:** Instead of relying on a default threshold (0.5), a precision-recall curve can help determine the ideal probability threshold that maintains 100% precision while maximizing recall. **Calibrated Probability Adjustments:** Techniques such as Platt Scaling and Isotonic Regression can be used to refine probability outputs.

Clinical & Practical Integration

1. **Bridging Machine Learning & Medical Practice: Collaborate with neurologists, radiologists**, and clinical experts to refine feature selection and improve model interpretability. Moreover, **integrating clinical trial datasets** (e.g., ADNI, AIBL) for external validation and to ensure model generalizability across diverse populations is important as

well.

2. **Deploying as a Clinical Decision Support Tool:** Develop a **web-based or API-driven deployment** to allow hospitals and healthcare institutions to integrate the model into existing systems. As well as create **mobile health applications** that provide early screening tools for at-risk patients, enabling proactive intervention before symptoms progress significantly.

Conclusion

Optimized Gradient Boosting Model: The best-performing model was an optimized Gradient Boosting (GB) classifier with a probability threshold of 0.86, effectively reducing false positives.

Limitations & Constraints

1. **Data Limitations: One-Hot Encoding Challenges:** The presence of numerous categorical variables, when one-hot encoded, led to feature explosion and misleading feature importance. **Lack of External Validation:** The model's performance has only been tested on the given dataset, and external datasets are required to ensure generalizability.
2. **Model Constraints: False Positive/False Negative Trade-off:** No model is perfect in minimizing both FP and FN simultaneously. Precision-Recall balancing remains a challenge. **Feature Interaction Complexity:** Deep learning approaches may be needed to better model complex feature dependencies.

Reference

Falahati, F., Westman, E., & Simmons, A. (2014). Multivariate data analysis and machine learning in Alzheimer’s disease with a focus on structural magnetic resonance imaging. *Journal of Alzheimer's Disease, 41*(3), 685-708.

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., & Holtzman, D. M. (2018). NIA-AA research framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia, 14*(4), 535-562.

Panday, A. (2023). Alzheimer’s Prediction Dataset (Global). Kaggle. Retrieved from https://www.kaggle.com/datasets/ankushpanday1/alzheimers-prediction-dataset-global

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288.

Appendix

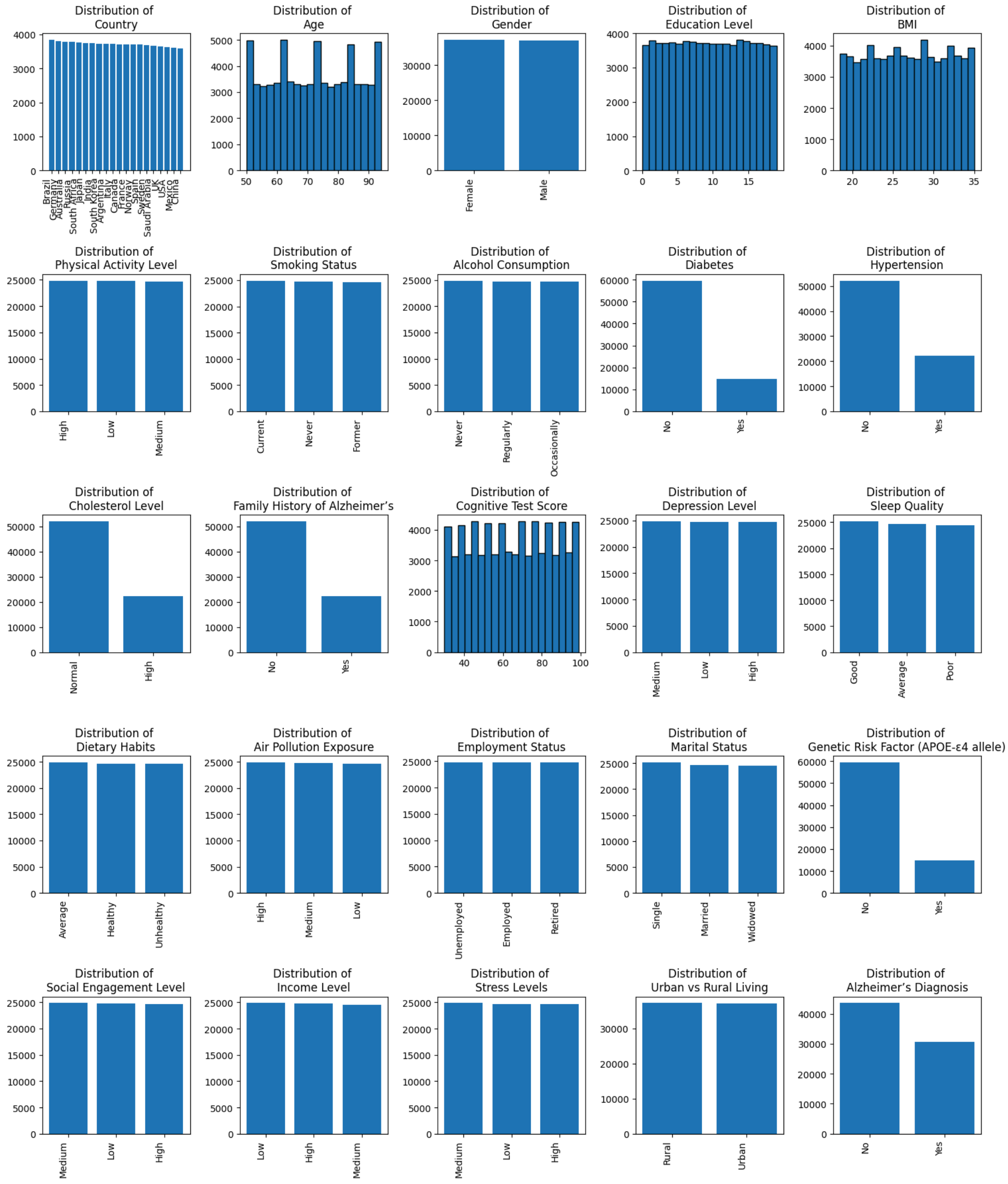


Figure 1. Dataset feature visualization

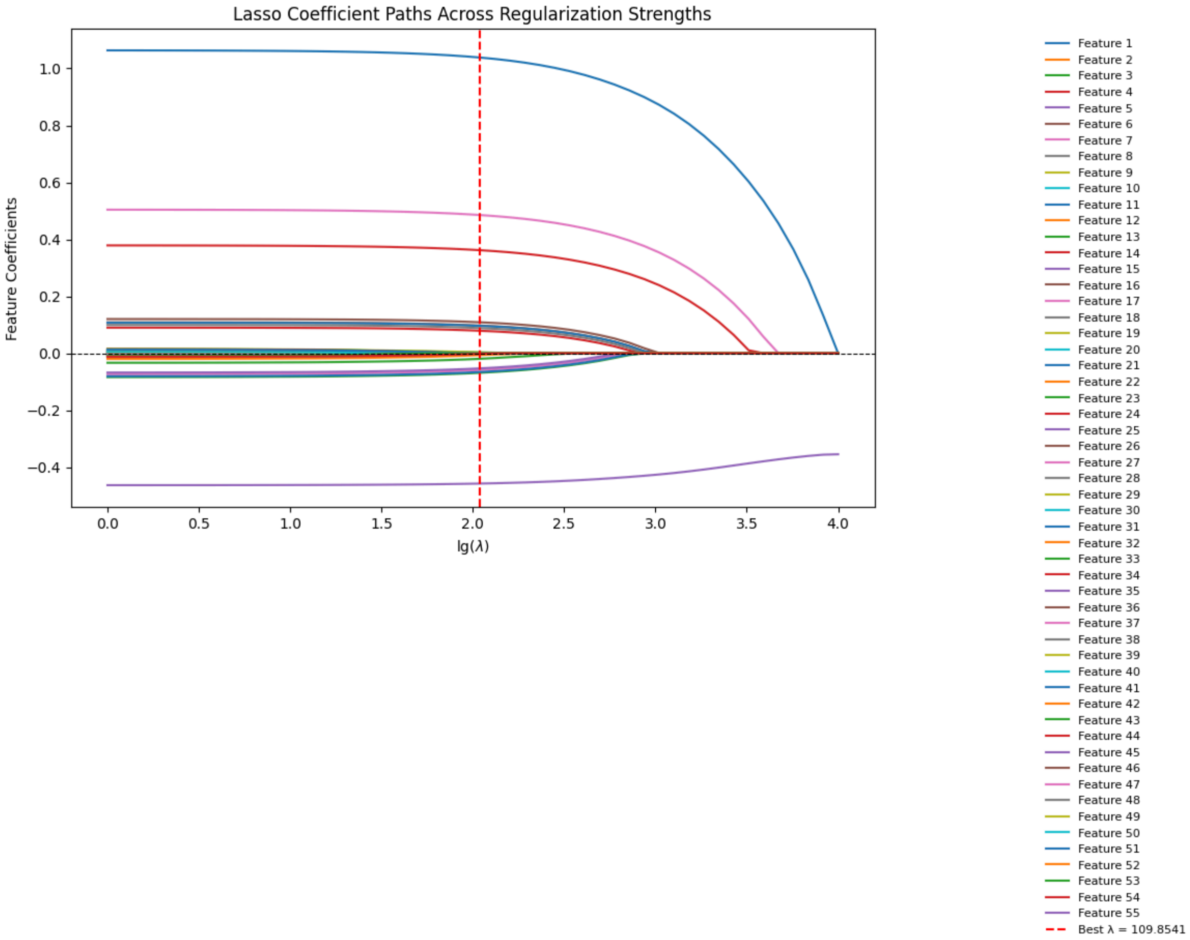
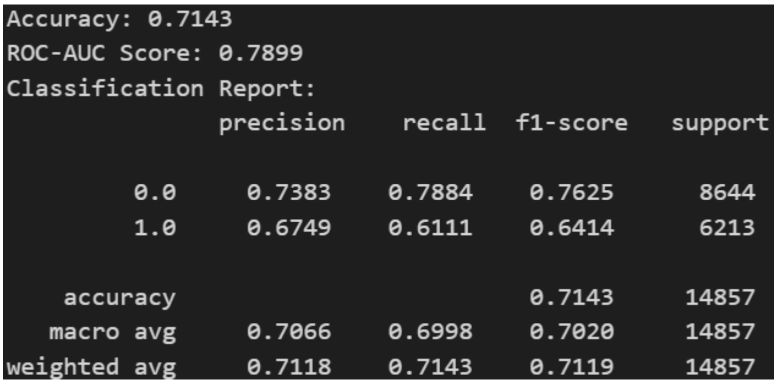


Figure 2. Lasso coefficient paths



```
Accuracy: 0.7143
ROC-AUC Score: 0.7899
Classification Report:
              precision    recall  f1-score   support

         0.0     0.7383    0.7884    0.7625      8644
         1.0     0.6749    0.6111    0.6414      6213

    accuracy                         0.7143     14857
   macro avg     0.7066    0.6998    0.7020     14857
weighted avg     0.7118    0.7143    0.7119     14857
```

Figure 3. Metrics of logistic regression

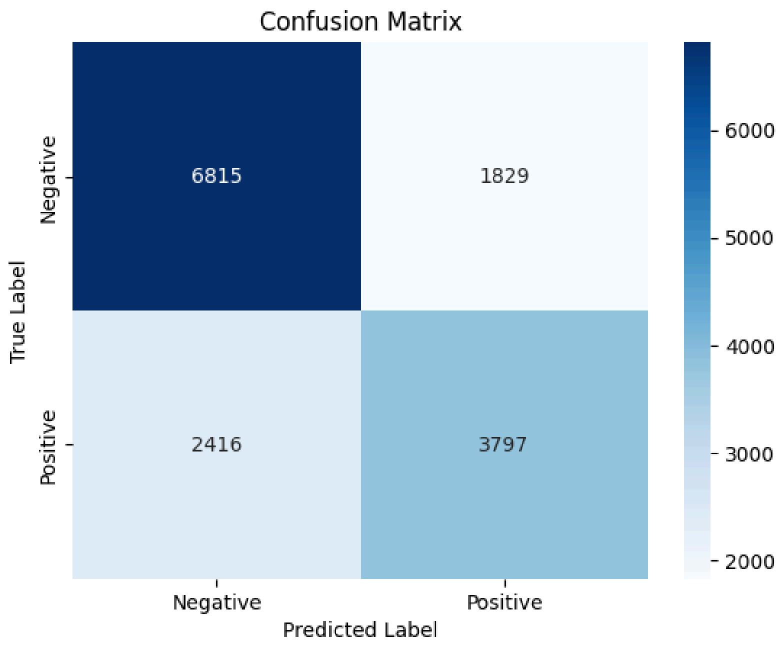
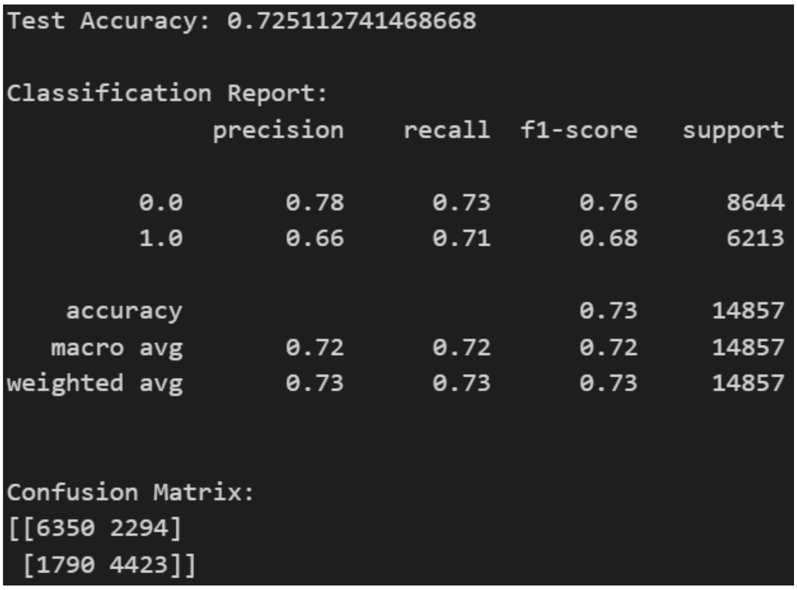
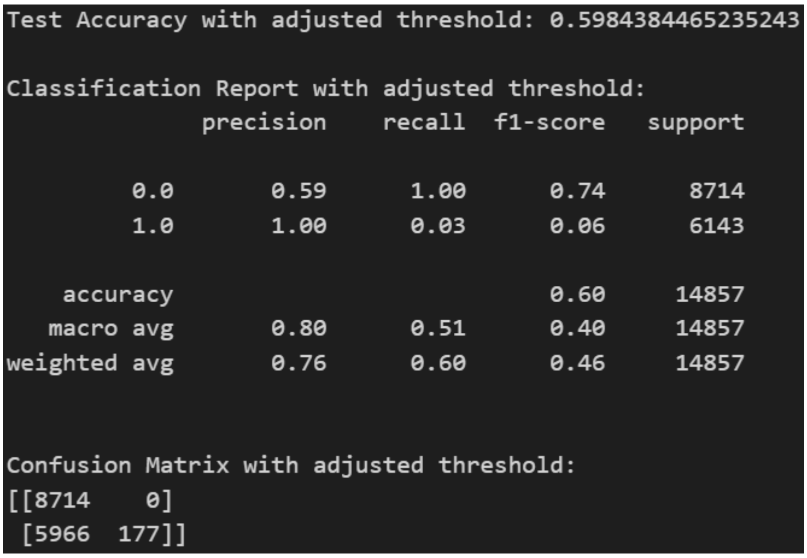


Figure 4. Confusion matrix of logistic regression



```
Test Accuracy: 0.725112741468668

Classification Report:
              precision    recall  f1-score   support

         0.0       0.78      0.73      0.76      8644
         1.0       0.66      0.71      0.68      6213

    accuracy                           0.73     14857
   macro avg       0.72      0.72      0.72     14857
weighted avg       0.73      0.73      0.73     14857


Confusion Matrix:
[[6350 2294]
 [1790 4423]]
```

Figure 5. Metrics of fine-tuned random forests



```
Test Accuracy with adjusted threshold: 0.5984384465235243

Classification Report with adjusted threshold:
              precision    recall  f1-score   support

         0.0       0.59      1.00      0.74      8714
         1.0       1.00      0.03      0.06      6143

    accuracy                           0.60     14857
   macro avg       0.80      0.51      0.40     14857
weighted avg       0.76      0.60      0.46     14857


Confusion Matrix with adjusted threshold:
[[8714    0]
 [5966  177]]
```

Figure 6. Metrics of conservative boosting classifier