Project Deliverable #3
Final Report

COMSW4995_022_2023_1 - Applied Machine Learning

Instructor: Vijay Pappu

Members: Ariel Goldman (apg2164), Yun-Yun Tsai (yt2781), Jason Jin (hj2602)
Yuhui Wang (yw3937), and Achmad Arviandito Caessara (aac2272)

Project Overview

Stroke is a leading cause of death and disability globally, with around 13.7 million new cases every year. In the United States, stroke is the fifth leading cause of death, and someone dies of stroke every four minutes. Early detection is crucial in limiting brain damage and improving patient outcomes. Machine learning has shown great promise in predicting stroke risk by analyzing various patient data, such as gender, age, BMI, lifestyle factors, medical history, and smoking history. For instance, a recent study used machine learning algorithms to predict stroke risk with an accuracy of 91.2%.

However, imbalanced data can introduce bias in the model, leading to low accuracy and high false negatives rate. This issue is particularly problematic in medical diagnosis where precision is crucial. In medical diagnosis, a false negative result can have severe consequences, such as delayed treatment and worsening of the patient's condition. To tackle this issue, various machine learning techniques can be employed to balance the data and improve model performance.

By accurately predicting stroke risk, healthcare providers can identify high-risk patients and intervene earlier, potentially reducing the number of strokes and improving patient outcomes. A recent study found that early detection of stroke and timely intervention could reduce the mortality rate by up to 50%. However, to ensure the accuracy and reliability of the models, it is crucial to address the imbalanced data issue using appropriate machine learning techniques. Therefore, we are conducting this project to tackle the problem of imbalanced data in tabular data for medical applications, aiming to obtain a reliable model with improved accuracy.

Approach

To approach this problem, we wanted to explore both a variety of data sampling techniques and a variety of machine learning models. In order to ensure the quality of the dataset and gain insights into the characteristics of the features, we start by performing data cleansing and exploratory data analysis. This involves checking for missing or invalid data, identifying outliers and anomalies, and assessing the distribution and variability of the features. In addition, we also employ data preprocessing techniques to further prepare the dataset for analysis.

We created four different sets of data to feed into our models by applying four different sampling techniques to our stroke dataset. First, we included a default sampling that simply used the data without doing anything to address the imbalance of positive and negative samples, in order to have a baseline. We then utilized undersampling, oversampling, and the synthetic minority oversampling technique (SMOTE), by applying each one to our stroke dataset to create three more options for training data. These three sampling techniques were all attempts to address the imbalanced nature of our dataset. Once we had our four training data options: default, undersampling, oversampling, and SMOTE, we were ready to try using each of these options to train each of our model options.

We experimented with various machine learning methods, including SVMs, Decision Trees, Random Forest, Boosting techniques (XGBoost, Hist Gradient Boosting, Gradient Boosting, and AdaBoost), Logistic Regression (LR), and Neural Network (MLP) with Stochastic Gradient Descent and Adam solvers. We tested all possible combinations of datasets and models, including conducting hyperparameter tuning to identify the optimal predictors for each model. Subsequently, we evaluated all of these combinations using metrics such as accuracy, precision, recall, AUC, and AP to determine the best model and data splitting method.

Model Comparison

Support Vector Machines (SVMs):

The Support Vector Machines (SVMs) showed the best overall performance with the Random Undersampling data, achieving 68.98% accuracy, 21.91% AUC, and 22.58% AP. The same data preprocessing technique also resulted in the highest recall of 76%, indicating that the model was able to identify most of the positive instances correctly. On the other hand, the highest precision of 11.45% was obtained with the weighted data, suggesting that the model made fewer false-positive predictions.

Decision Trees:

The Decision Trees showed consistent performance with 95.11% accuracy, but 0% precision and recall, indicating that the model did not predict any positive instances. However, the Random Oversampling, Random Undersampling, SMOTE, and Balanced Weight data produced the highest AUC (52.45%) and AP (4.89%), indicating that the model's ability to distinguish between positive and negative instances was the highest with these techniques.

XGBoosting Classifier:

The XGBoosting Classifier showed the best overall performance with the Random Undersampling data, achieving 70.45% accuracy, 20.26% AUC, and 20.91% AP. Similarly, the highest recall of 76% was also achieved with the Random Undersampling data. However, the highest precision of 27.03% was obtained with the Random Oversampling data, suggesting that the model made fewer false-positive predictions.

Hist Gradient Boosting Classifier:

The Hist Gradient Boosting Classifier also showed the best overall performance with the Random Undersampling data, achieving 70.74% accuracy, 20.80% AUC, and 21.88% AP. The highest recall of 76% was achieved with the Random Undersampling data, and the highest precision of 36.36% was obtained with the Baseline data.

Gradient Boosting Classifier:

The Gradient Boosting Classifier showed the best overall performance with the Random Undersampling data, achieving 68.00% accuracy, 19.79% AUC, and 20.57% AP. The highest recall of 78% was achieved with the Random Undersampling data, and the highest precision of 18.85% was obtained with the SMOTE data.

AdaBoost Classifier:

The AdaBoost Classifier showed the best overall performance with the Random Oversampling data, achieving 74.27% accuracy, 21.13% AUC, and 22.71% AP. The highest recall of 80% was achieved with the Random Oversampling data, and the highest precision of 13.65% was also obtained with the Random Oversampling data.

Logistic Regression (LR):

The Logistic Regression (LR) showed the best overall performance with the Random Undersampling data, achieving 73.68% accuracy, 28.73% AUC, and 30.08% AP. The highest recall of 82% was achieved with the Random Undersampling data, and the highest precision of 100% was obtained with the Baseline data.

Neural Network (MLP) with Stochastic Gradient Descent (SGD):

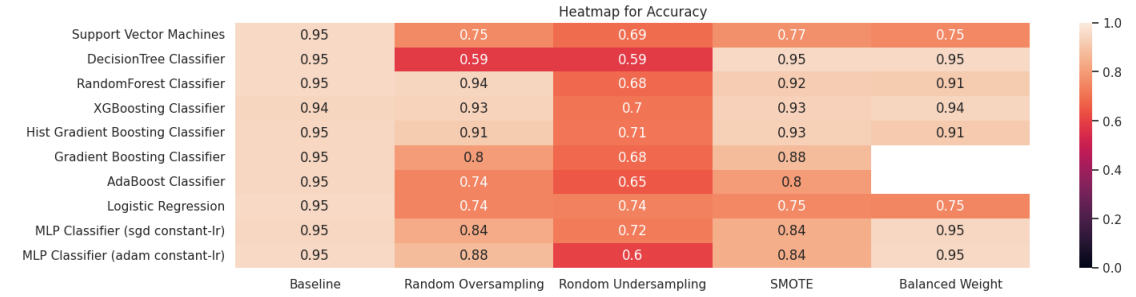
The Neural Network (MLP) with Stochastic Gradient Descent (SGD) showed the best overall performance with the Baseline data, achieving 94.91% accuracy, 18.97% AUC, and 20.01% AP. The highest recall of 76% was achieved with the Random Undersampling data, and the highest precision of 13.13% was obtained with the Random Oversampling data.

Neural Network (MLP) with Adam:

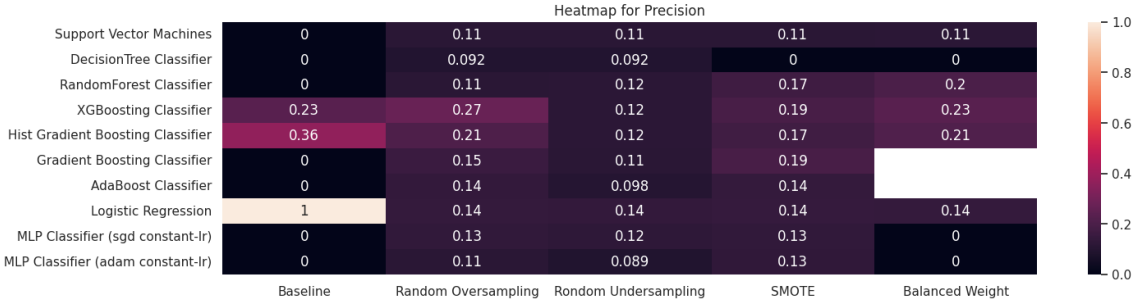
Finally, the Neural Network (MLP) with Adam showed the best overall performance with the Baseline data, achieving 95.11% accuracy, 40.09% AUC, and 14.28% AP. The highest recall of 78% was achieved with the Random Undersampling data, and the highest precision 13.33%.

Comparative Analysis:

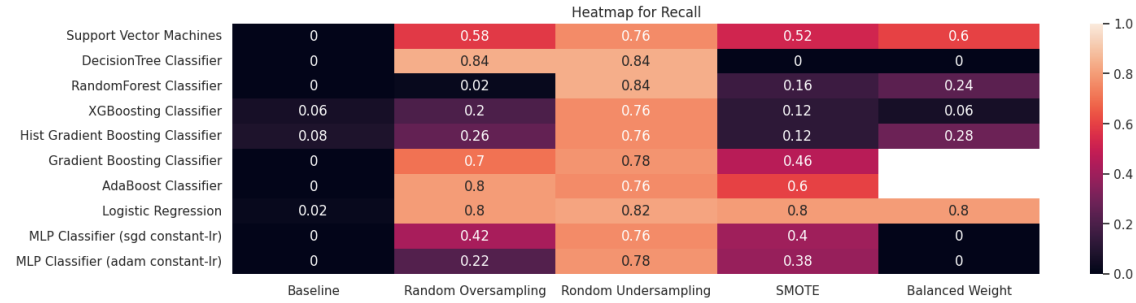
- Best recall (82%): Logistic Regression with Random Undersampling data achieved the highest recall value, making it the top choice for maximizing recall in predicting stroke cases.
- Best precision (100%): Logistic Regression with Baseline data achieved perfect precision, making it the best option for maximizing precision in predicting stroke cases.
- Highest AUC (40.09%): Neural Network (MLP) with Adam as the solver and constant learning rate using Baseline data had the highest AUC value
- Highest AP (30.08%): Logistic Regression with Random Undersampling data achieved the highest AP value, suggesting that this model provides the best balance between precision and recall in predicting stroke cases.



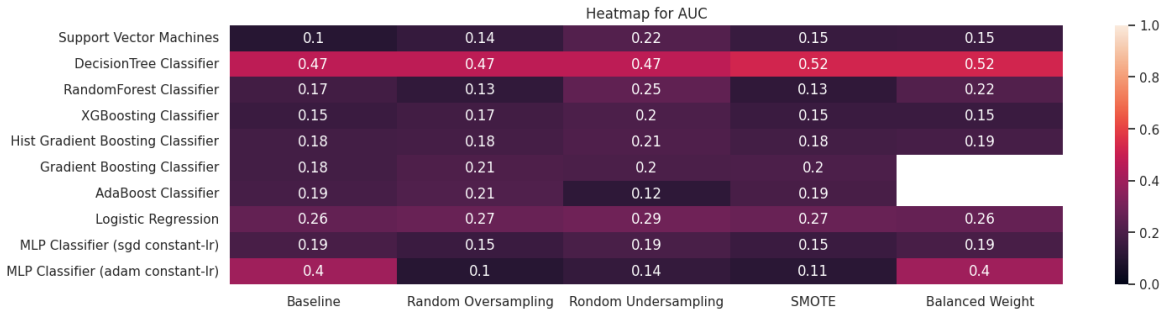
Model Comparison based on Accuracy metric



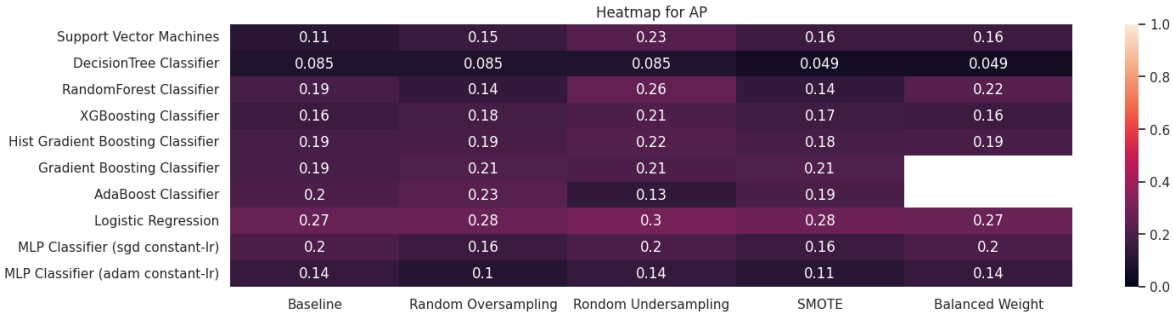
Model Comparison based on Precision metric



Model Comparison based on Recall metric



Model Comparison based on AUC metric



Model Comparison based on AP metric

Conclusion

We evaluated a variety of machine learning methods, including SVMs, Decision Trees, Random Forest, Boosting techniques (XGBoost, Hist Gradient Boosting, Gradient Boosting, and AdaBoost), Logistic Regression (LR), and Neural Network (MLP) with Stochastic Gradient Descent and Adam solvers, to analyze imbalanced data in tabular medical records related to strokes. We used four different resampling strategies like Random Oversampling, Random Undersampling, SMOTE, and Balanced Weight to address the imbalance issue. Our analysis focused on performance metrics such as accuracy, AUC, AP, recall, and precision.

Our analysis revealed that Logistic Regression demonstrated the most promising results for handling imbalanced data in stroke-related medical records. However, the Boosting techniques and MLP models also showed potential, although further optimization and experimentation may be required for specific applications. We also noted that MLP models with Adam solver exhibited the highest AUC values.

Overall, our findings suggest that careful selection of appropriate machine learning methods and resampling strategies can help mitigate the issues of imbalanced data in medical record datasets. The results of this study can be useful for future research in this area, as well as for practical applications in the medical field.

References:

- World Stroke Organization. (2019). Global Stroke Fact Sheet 2019.
- Eysenbach, G. (2021). Predicting risk of stroke from lab tests using machine learning algorithms: Development and evaluation of prediction models.