STAT 451 Project Report: *Predicting Stroke Events with Physical Health Data*

Group 22: Lucy Chen, Tina Cao, Grant Zhou, Judy Zhu

Introduction:

Stroke is the second leading cause of death globally, responsible for about 11% of total death. In this project, we implemented a panoply of machine learning techniques to explore whether a patient will have a stroke and the probability of having a stroke given measurements of a patient's health condition. From the Stroke Prediction Dataset published by Fedesoriano on Kaggle, we chose 10 numerical and categorical features like gender, age, and average glucose level as independent variables, and stroke events were labeled with 1 (had a stroke, about 95% of the entries) or 0 as the dependent variable. We implemented six major methods: imputation for missing data, one-hot encoding, oversampling on the unbalanced dataset, logistic regression for feature selection and model fitting, decision tree for classification, and GridSearchCV for model optimization. Our result showed that gender, age, bmi, and average glucose level have the largest effect on models for predicting stroke events. Performing a majority vote by combining logistic regression and decision tree, we achieved 0.849 accuracy, 0.507 recall, and 0.163 precision on the separated highly unbalanced test dataset.

Data Engineering and Exploratory Analysis:

Due to the unbalanced nature of our dataset, we divided our dataset into two parts. We performed oversampling on 70% of the dataset, which is used for train-test, cross-validation, and model fitting. For the rest 30% of the dataset, we left untouched as highly unbalanced test data to evaluate our model in more realistic scenario. Our dataset has N/A values, so we fixed this using data imputation with the median. Moreover, our dataset has several categorical features, for example, work type has two categories, private and self-employed. Hence, we performed one-hot encoding to separate these categorical features. Based on the dataset, we plotted the distributions of age and average glucose level for stroke and non-stroke groups (Figure 1). From the age distribution, we see that older people tend to experience a stroke event.

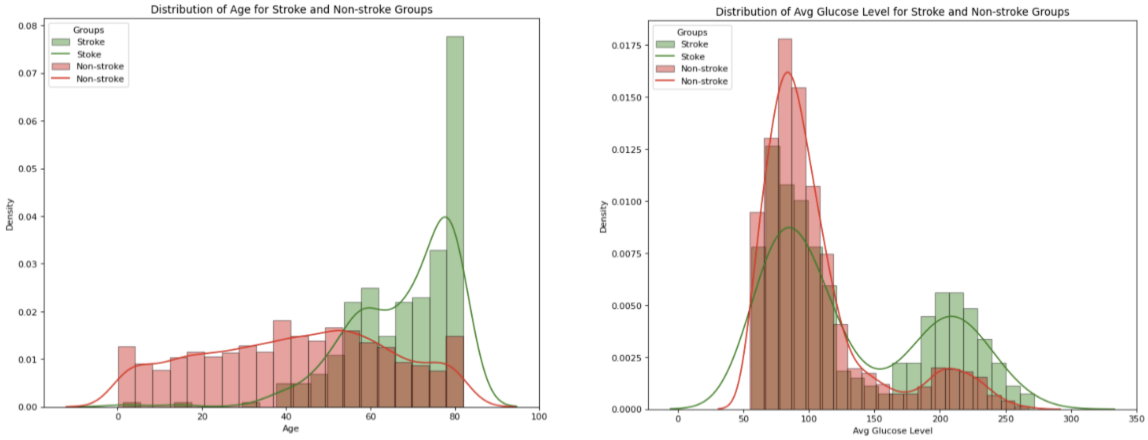

Figure 1: Distribution of Age and Avg Glucose Level for Stroke and Nonstroke Groups

Feature Selection & Model Training:

Firstly, we performed feature selection with logistic regression with L1 norm on the 19 categorical and numerical features after one-hot encoding. With C=0.01, we obtained three features with non-zero coefficients: **'age', 'gender_Male', 'avg_glucose_level'.**

On the oversampled training set, we performed GridSearchCV with fixed max_iter = 5000 and hyperparameters {'C': [0.01, 1, 100]} and obtained an optimized model with C=1. Then, with all features, we trained an ID3 decision tree model and performed GridSearchCV with the following hyperparameters: 'Max_depth': [1, 3, 5, 7], 'min_samples_split': [2,3,4,5], 'min_samples_leaf': [1,2,3,4]. The optimized model comes with hyperparameters: {'max_depth': 7, 'min_samples_leaf': 1, 'min_samples_split': 4}.

Both models' performances on the test data from the oversampled dataset are displayed below (Figure 2). Moreover, we also returned the importance score (related to a decrease in impurity) provided by the decision tree, and the top 3 features are **age, bmi, and avg_glucose_level,** which shows similar results as the feature selection with logistic regression (both **age** and **avg_glucose_level** appears in these two evaluation process).

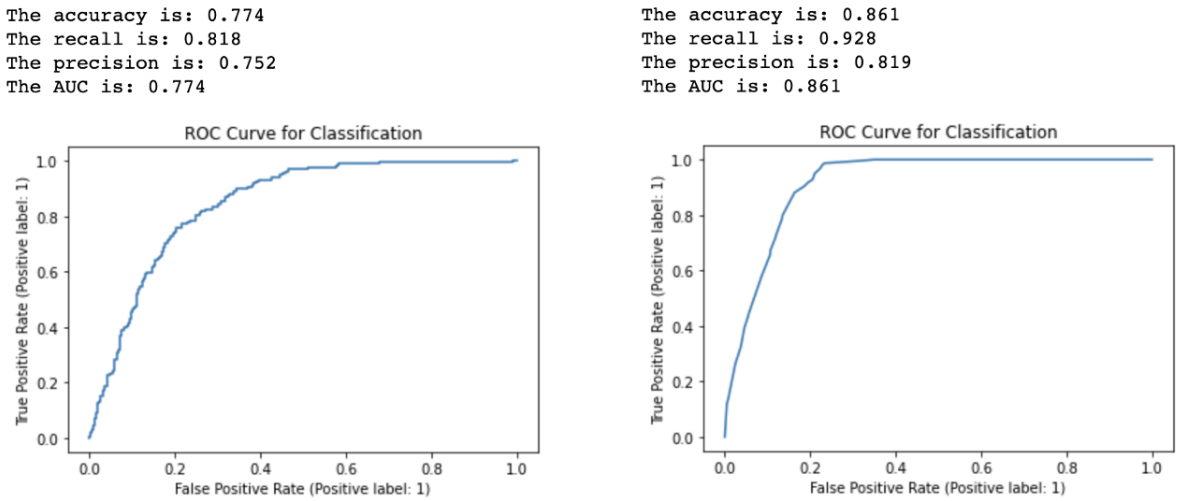

Figure 2: Performance of Logistic Regression (left) and Decision Tree (right) on Oversampled Test Data

Majority Vote Prediction:

Aside from two previous models, we adopted a "majority" vote prediction, where we only predict 1 for stroke condition if both logistic regression and decision tree models predict 1. There are two significant reasons for this. First, because we are predicting a serious matter considering health conditions, we need to be more cautious when giving a positive diagnosis. To avoid

misleading our patients and causing panic, we aim to minimize the possibility of generating false positive results.

The second reason is that data sets concerning stroke conditions are usually highly unbalanced. However, as shown below in Figure 3, two previous models demonstrate a weaker performance when predicting unbalanced data sets. Majority Vote Prediction, on the other hand, maintains good performance scores facing unbalanced datasets by eliminating false positives, and the performance metric is **accuracy: 0.849, recall: 0.507, and precision: 0.163 .**

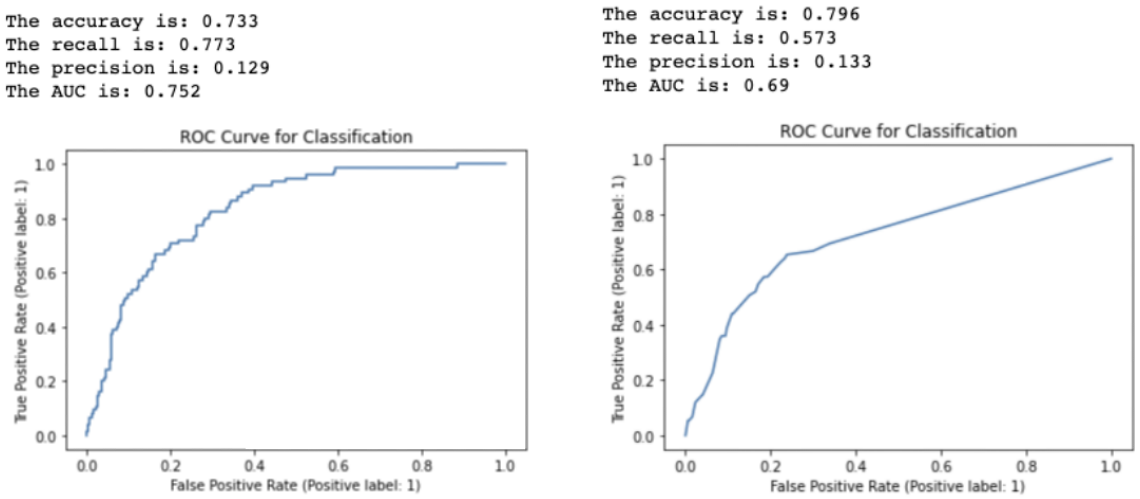

Figure 3: Performance of Logistic Regression (left) and Decision Tree (right) on Separated Unbalanced Test Data

Conclusion and Summary:

We built models that have high accuracy and precision on the oversampled datasets and relatively satisfying performance on the "untouched" dataset. Moreover, we identified features (age, bmi, avg glucose level) that are important in predicting stroke and provide certain information that can help design preventative measures. In the future, we might adopt other prediction models including PCA or random forest (refine the decision tree) to further increase our predictive power. Also, finding a more balanced stroke dataset can be extremely helpful.

Contribution:

Member	Proposal	Coding	Presentation	Report
Lucy Chen	0.8	1	1	0.9
Tina Cao	0.8	0.9	1	1
Grant Zhou	1	1	0.8	0.8
Judy Zhu	1	0.8	1	1