Auditing an XGBoost Automated Decision System (ADS) for Stroke Prediction Data

Responsible Data Science

Section 001

Jon Dinh, Yash Jha

I. Background

A stroke, or a cerebrovascular accident, is a medical emergency wherein blood flow to the brain is interrupted or blocked. This sudden drop in oxygen and nutrients will cause brain cells to start dying within just a few minutes. It can cause lasting brain damage, long-term disability, and possible death.

According to the World Health Organization, roughly 15 million people per year suffer a stroke globally. Of those 15 million, ⅓ or 5 million die. Another ⅓ are permanently disabled (WHO). In addition, the WHO states that the most significant and in-control variables that contribute to a stroke are high blood pressure (HBP) and tobacco use. Although in developed countries the incidence of stroke has decreased, the total number of cases have an increasing trend due to increased longevity and the aging population.

The purpose of this automated decision system is to classify whether a patient is likely to suffer a stroke from provided features. The goal of this ADS is a binary classification problem, which will be reflected in the methodology of the system.

II. Input/Output

The data used in this ADS is a combination of personal and medical data. The data contains 5110 entries and 12 features. Individuals' personal features include their gender (male/female), age, whether they were ever married, their work type (private, self-employed, etc.), their residence type (urban, rural, etc.) and smoking status (never, formerly, currently, etc.). The remaining features are medical data points, which include presence of hypertension, presence of heart disease, average glucose level, and body mass index (BMI). The output variable is stroke, a binary with '1' indicating that the individual suffered a stroke. The output variable is stroke, a binary with '1' indicating that the individual suffered a stroke.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 5110 non-null   int64
 1   gender             5110 non-null   object
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64
 4   heart_disease      5110 non-null   int64
 5   ever_married       5110 non-null   object
 6   work_type          5110 non-null   object
 7   Residence_type     5110 non-null   object
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object
 11  stroke             5110 non-null   int64
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

Figure 1: column data types

The value distributions for numerical and categorical features are shown below in Figure 2 and Figure 3. The distributions of the numerical features in Figure 2 indicate that 'age' is fairly evenly distributed while 'avg_glucose_level' and 'bmi' are right skewed with the majority of values on the lower end.

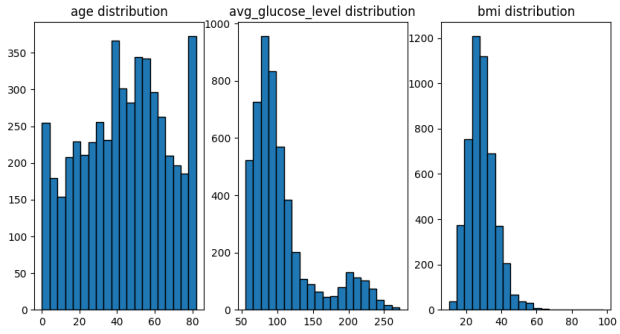

Figure 2: numerical distributions

The distributions of the categorical features in Figure 3 show that the majority of individuals are female and the majority have been or are married. Most individuals also work in the private sector. The distribution of residence type is roughly evenly split between urban and rural. Finally, most people have never smoked while a significant number of individuals have their smoking status recorded as unknown.

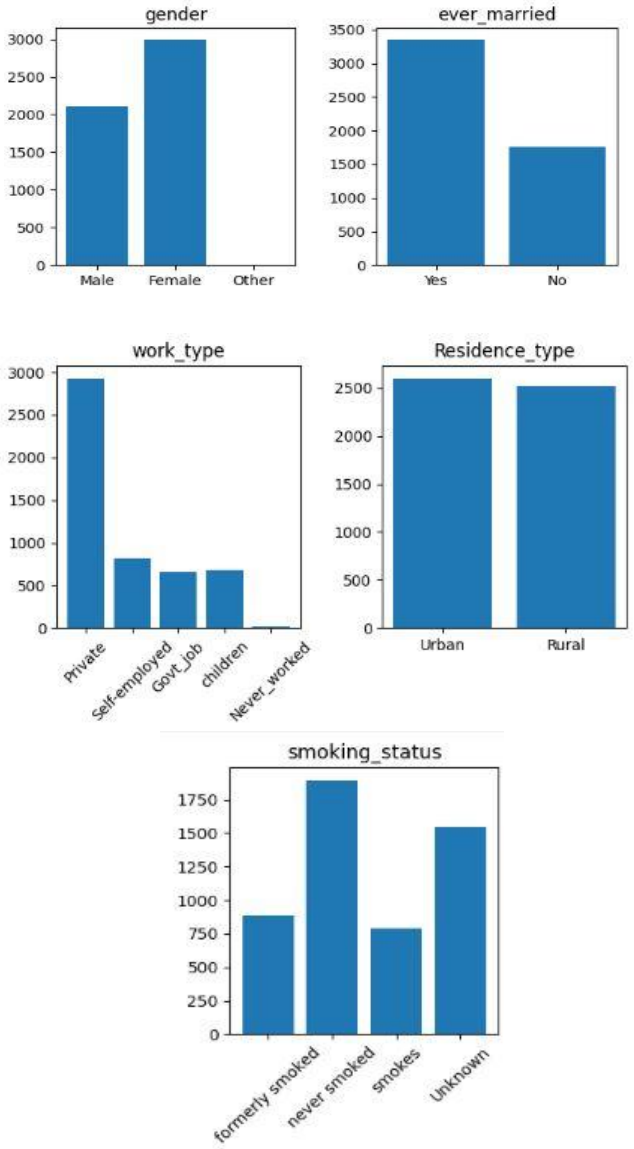

Figure 3: categorical distributions

III. Implementation and Validation

In the original automated decision system, missing data imputation was the first task to be completed. From EDA, the only feature with missing values was BMI (Figure 4). The chosen method of imputation was feature mean, as the median and median of the BMI is relatively close.

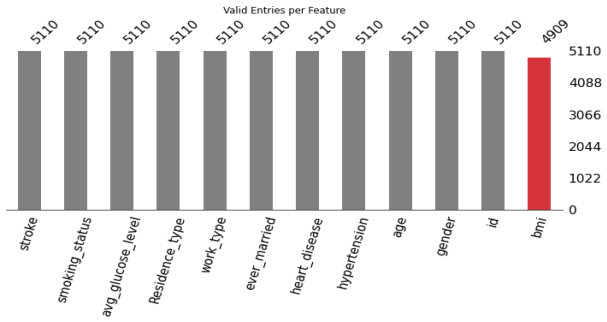

Figure 4: Distribution of missing values

Next, the categorical features were encoded according to the order of appearance of unique values per feature column.

The class distribution was originally heavily imbalanced with a 19:1 ratio in favor of individuals labeled ‘no stroke’ (shown in Figure 6). To correct this imbalance, a hybrid data sampling method consisting of Random Under Sampler and SMOTE was used to balance the total class distribution to a 2490:2490, or 1:1 ratio.

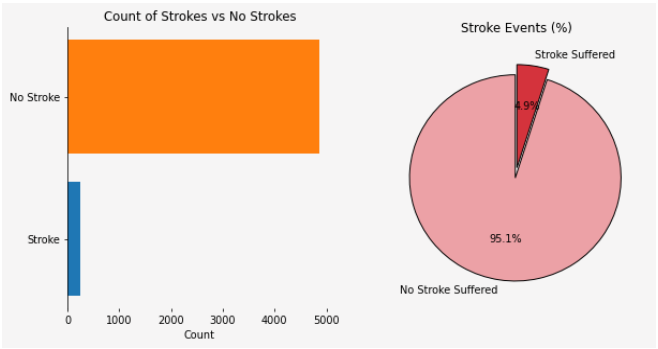

Figure 5: class distribution

Next, feature selection was done by calculating chi-squared for categorical features and ANOVA metrics for numerical features. From this, the following features were dropped: (1) smoking status (2) heart disease (3) hypertension (4) BMI. The resulting tabular data was then normalized and standardized according to features that are not normalized, and those that are normalized but have a large range.

An XGBoost model was then hard coded with learning_rate = 0.01, max_depth = 3, and n_estimators = 1000. The model was cross validated with k = 5 folds, using ROC accuracy as the scoring. The original metrics chosen to evaluate the system implementation was (1) CVS (2) ROC accuracy (3) precision (4) recall (5) f1-score (6) support (7) TNR (8) FPR (9) TNR (10 TPR.

We then observed that there are points where a mix of data leakage occurs. The original ADS chose to split the data into train and test sets after artificial resampling, feature selection and scaling. However, generating artificial observations from the entire dataset (including the test data) increases the risk of overfitting, and decreases robustness to new observations. The same effects are induced by scaling the entire dataset from future test predictions. In addition, the resampling method is a combination of two undersampling and oversampling. It provides no reflexive adaptation between the two processes, compared to other available methods such as ADASYN. Furthermore, the model itself is not tuned, and rather hard-coded with hyper parameters.

IV. Outcomes

We ran the automated decision system as received and recorded its performance metrics. Performance metrics were based on the classic classification metrics, utilizing the confusion matrix for imputation. Accuracy, precision, recall and f1 scores were computed and recorded. Precision and recall were computed to assess the classifier's ability to classify strokes correctly and find all strokes, which is especially important when the positive class is heavily outnumbered. F1 score was calculated to compare the impacts of precision and recall, which is important to gauge the model's tendencies. Accuracy metrics were calculated based on their average weighted by support, where support is defined as the the number of true instances for each label. This was done to counteract any imbalances in class distribution. Furthermore, accuracy was computed for 'male' and 'female' gender groups.

However primarily, we are concerned with the False Negative Rate of the decision system in regards to performance. The goal of the system is to predict the occurrence of a stroke. In essence, this is an assistive intervention, and we are concerned with ensuring predicting quality for those with higher needs. In other words, False Negatives are those with need or have higher risk of stroke, but are predicted negative. The ADS is recorded with an accuracy, precision, recall, and f1 score of approximately 0.839. The FNR is measured at 0.12.

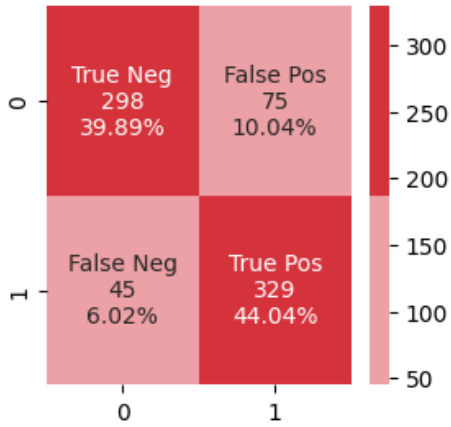

Figure 6: original ADS confusion matrix

Accuracy	0.839
Precision	0.842
Recall	0.839
F1	0.839
False Positive Rate	0.21
True Positive Rate	0.88
True Negative Rate	0.799
False Negative Rate	0.12

Figure 7: overall model performance metrics

With respect to fairness metrics, demographic parity ratio was calculated to compare selection rates across gender groups. Because demographic parity ratio does not take true values into account, equalized odds ratio was also computed to compare true positive and true negative rates across gender groups. Finally, FNR, FPR, and selection rate were computed for ‘male’ and ‘female’ gender groups. Below are the basic fairness metrics with gender grouping.

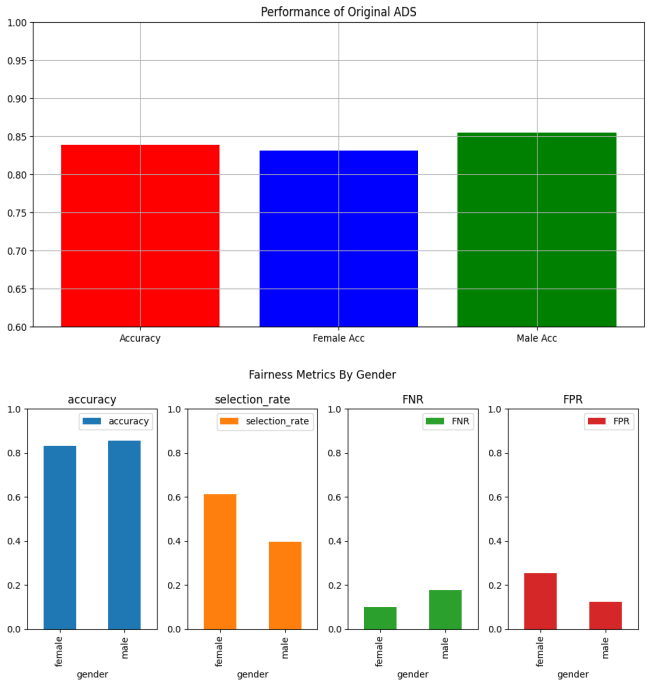

Figure 8: accuracy and fairness for model and gender

The primary metrics to focus on in terms of fairness are EOR and FNRP. Ideally, a good EOR score indicates the model performs well regardless of sensitive group membership, as well as similar FPR and TPR. Similar to performance metrics, FNRP measures the ratio between all disjoint sensitive group memberships, in this case gender. A high FNRP indicates that the probability of someone needing assistance, but falsely predicted, is the same across subpopulations. Below are the recorded fairness metrics in parity ratio form.

False Negative Rate Parity Ratio	0.569
Equalized Odds Ratio	0.493
Demographic Parity Ratio	0.644
Accuracy Ratio	0.973
Selection Rate Ratio	0.644

Figure 9: fairness metrics in parity ratio form

We find the accuracy ratio between gender subpopulations is 0.973. This indicates that the model has similar overall performance in predictive power between genders. However FNPR, EOR, DPR and selection rate between subpopulations hovers in the mid range between 0 and 1. This indicates that although predictive power is similar and high overall, there exists a fair amount of bias and a substantial amount of difference in predictive quality.

From model examination, we see that the feature, gender, lies in the mid range of importance. This is supported by the feature importance extraction using inherent XGBoost attributes and SHAP values.

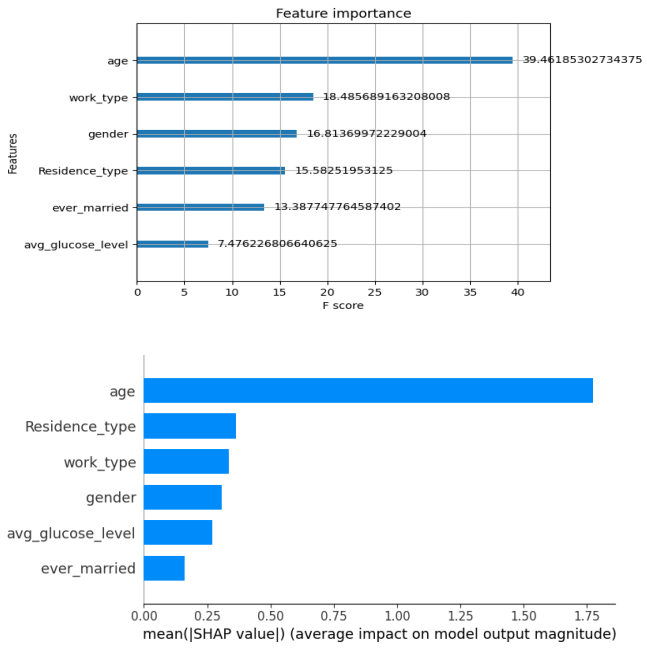

Figure 10: feature importance with XGBoostClassifer and SHAP values

Thus, we can conclude that the automated decision system does not rely heavily on the group membership of gender. The predictive strength of each feature is examined and interpreted with regards to both prediction outcomes using LIME.

We see in the first local explanation, that the top most contributing feature to the prediction class of "no stroke" is age. It indicates that at the local level, the model contributes those of a younger age to have a significantly decreased probability of suffering a stroke. In this case, the observation has a reported gender of Male, with an importance of middling value.

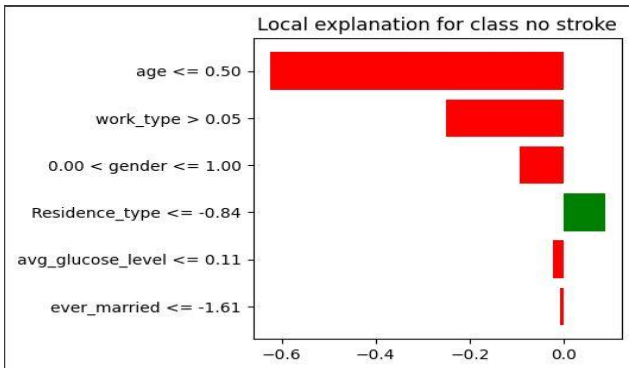

Figure 11: LIME explanation of predicted "no stroke"

Similarly, for an observation predicted as "stroke", we see that at a local level, the model contributes higher ages to have a massive increase in probability of suffering a stroke. For this observation, the gender feature is reported as Female, with a low importance.

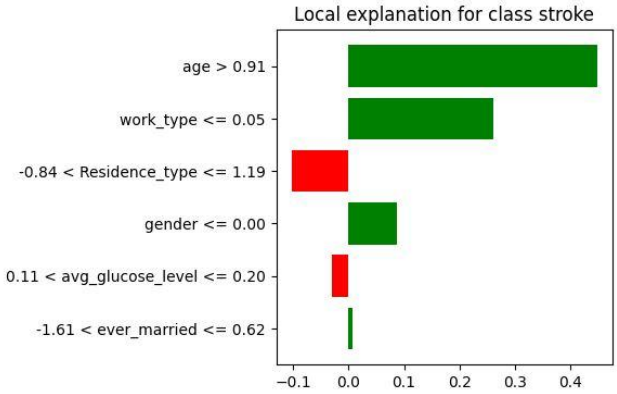

Figure 12: LIME explanation of predicted "stroke"

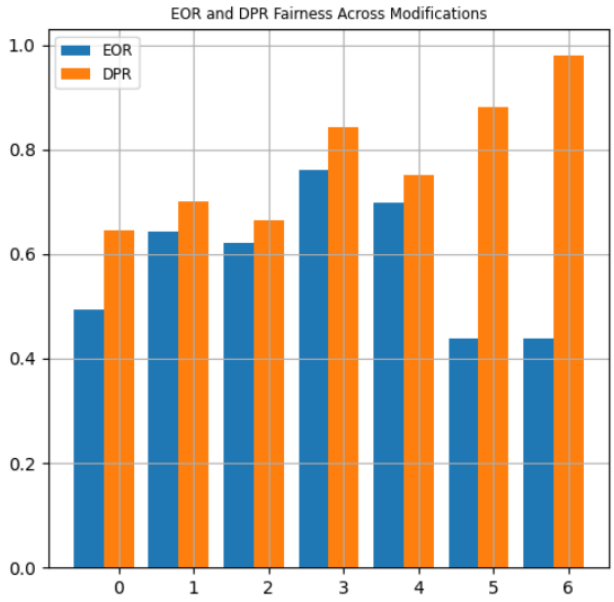

Figure 13: EOR and DPR with fairness modifications.

In an effort to improve the accuracy and fairness of the ADS with respect to gender, we made several modifications: (0) original ADS, (1) ADASYN , (2) SMOTE-Tomek, (3) SMOTE-ENN, (4) correlation remover, (5) hyperparameter tuning, (6) threshold optimizer. Modifications (1), (2), and (3) were sampling techniques used to balance class distribution during training; the original ADS used standard SMOTE. The correlation remover was fitted on training data with respect to gender and used on both train and test data. Hyperparameter tuning was done with grid search using varying learning rate (.001, .01, .05, .1), max_depth (2 to 10), and n_estimators (500, 1000, 1500, 2000). Finally, a threshold optimizer constrained by equalized odds ratio was applied.

V. Summary

We believe that this ADS had a good initial performance with an accuracy of .839 and a false negative rate of .12 (Figure 7). However, this is with the caveat that the class distribution was balanced, which is not realistic in the real world if the ADS is exposed to new data. In terms of fairness, the ADS was consistent with respect to predictive power across genders, yet other metrics (FNPR, EOR, DPR), indicated bias with how predictions were made across gender groups (Figure 9).

We believe the data was appropriate for this ADS because XGBoost is based on decision trees, which are appropriate for determining a binary variable based on some categorical features. In addition, some of the numerical features also have categories; for example, age might be split into young, middle-age, and senior people, so we can use decision trees to find such thresholds.

With respect to accuracy and fairness metrics, we believe that our analysis was relevant for both patients and doctors. The decision to focus on FNR in the accuracy metrics was primarily motivated by the assumption that doctors will likely be most interested in ensuring that all patients at risk of stroke are detected. In terms of fairness metrics, we chose to focus on gender as our sensitive attribute and used FNRP, EOR and DPR to assess biases during prediction. This is relevant for patients to assure them that gender, an attribute that has no logical correlation with any factors that would directly affect classification, is a minimal factor for risk of stroke. In particular, FNRP is crucial to assess whether or not the ADS is missing positive cases more often in one gender than another. Both our accuracy and fairness metrics are crucial for interested private parties such as hospital boards and equity firms because liability for patients who are labeled as 'false negative' is minimized and they would like to feasibly claim that the model is fair across gender groups.

The implementation was done well, especially considering that the main problem was imbalanced class distribution, yet the original ADS was able to remedy this with SMOTE. It was also very accurate at .839 as mentioned previously, with a low FNR of .12.

However, we observed an apparent disconnect in fairness with respect to gender with the original FNPR, EOR, and DPR values at .569, .493, and .644, respectively. And while the accuracy across gender groups was very similar, we also observed in our LIME analysis that gender was a decently important feature considered by the ADS (Figures 11 and 12). Because of this, we decided to apply modifications to the ADS shown in Figure 13 to reduce prediction bias, which resulted in improvements in DPR and declines in EOR. This shows that the model's classification is not dependent on gender, yet the chances of being classified as positive or negative are different across gender groups. Finally, we observed that the model was reasonably robust with respect to maintaining accuracy when fairness modifications were added (shown below in Figure 14). However, the model is not robust in terms of classifying brand new data; it maintains high accuracy, yet FNR and FPR remain high because the class distribution in the real world is still so heavily imbalanced despite it being corrected during training.

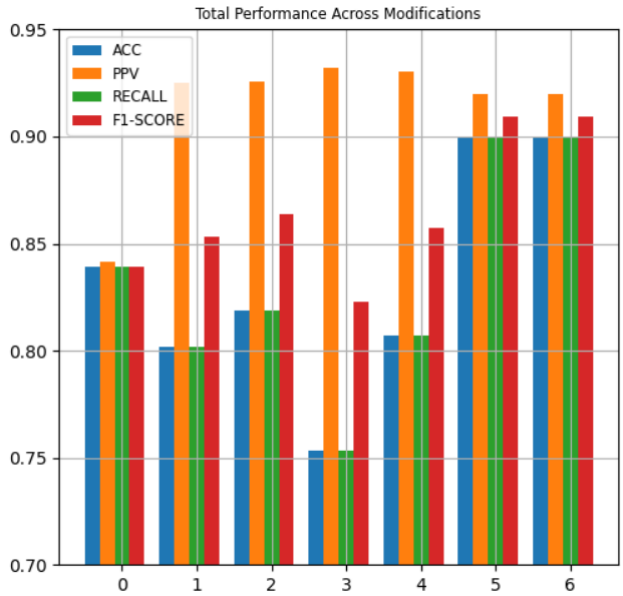

Figure 14. Accuracy metrics across fairness modifications: (0) original ADS, (1) ADASYN , (2) SMOTE-Tomek, (3) SMOTE-ENN, (4) correlation remover, (5) hyperparameter tuning, (6) threshold optimizer.

In terms of recommended improvements, we suggest using the same SMOTE-ENN to increase precision and F1 while still maintaining accuracy (Figure 14). We also observed that using correlation remover with hyperparameter tuning significantly improved accuracy metrics across the board. Finally, a threshold optimizer with equalized odds ratio as the constraint increases fairness while maintaining other metrics (Figures 13 and 14).

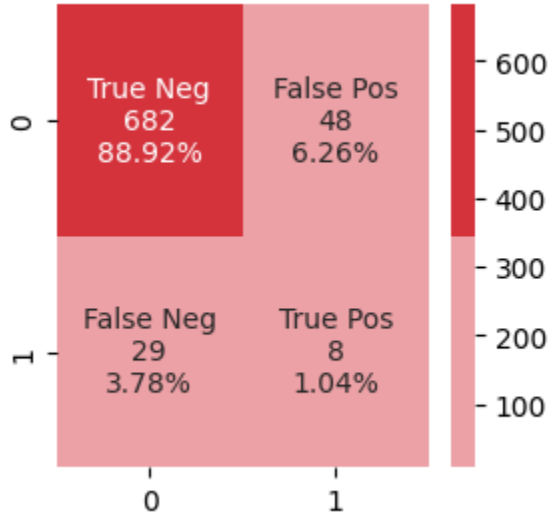

Figure 15. Confusion matrix for ADS modified by correlation remover, grid search, and threshold optimizer.

I believe we would be comfortable deploying this ADS publicly with the appropriate modifications, primarily because the accuracy is consistently much higher than the industry standard of 70%. It is a good starting point for classifying stroke predictions in terms of pure accuracy, yet the issue of lowering FNR and with respect to new data remains (Figure 15) . For this reason, we believe it is still far too risky to use in the industry because FNR remaining high when exposed to new data is extremely risky for patient diagnosis. In conclusion, the ADS uses relevant features and a sensible pipeline to classify stroke predictions, yet is not robust enough to new data primarily due to largely imbalanced class distribution, which seems to only be solvable through a more complex feature space or a stronger pipeline.

VI. References

A. World Health Organization. (n.d.). *Stroke, Cerebrovascular accident*. World Health Organization. Retrieved April 17, 2023, from https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html

B. Deshpande, T. (Contributor.). (2021, January 26). *Stroke Prediction Dataset*. Kaggle. Retrieved March 21, 2023, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

VII. Appendix (Results)

Performance Metrics	acc	ppv	npv	recall	f1	TN	FP	FN	TP	roc_acc	FPR	TPR	TNR	FNR
Original ADS	0.839	0.842	0.869	0.839	0.839	298	75	45	329	0.839	0.201	0.88	0.799	0.12
adasyn ADS	0.802	0.925	0.966	0.802	0.853	599	131	21	16	0.626	0.179	0.432	0.821	0.568
smote_tomek ADS	0.819	0.926	0.967	0.819	0.864	612	118	21	16	0.635	0.162	0.432	0.838	0.568
smote_enn ADS	0.751	0.932	0.974	0.751	0.821	554	176	15	22	0.677	0.241	0.595	0.759	0.405
Correlation Remover ADS	0.807	0.93	0.971	0.807	0.857	600	130	18	19	0.668	0.178	0.514	0.822	0.486
GridSearch ADS	0.9	0.92	0.959	0.9	0.909	682	48	29	8	0.575	0.066	0.216	0.934	0.784
Threshold Optimizer ADS	0.896	0.919	0.959	0.896	0.907	679	51	29	8	0.573	0.07	0.216	0.93	0.784

Table 1. Performance metrics across modifications.

Fairness Metrics	FNRP	EOR	demo_parity_diff	DPR	female_acc	female_sr	female_FNR	female_FPR	male_acc	male_sr	male_FNR	male_FPR
Original ADS	0.569	0.493	0.218	0.644	0.832	0.613	0.101	0.253	0.855	0.395	0.177	0.125
adasyn ADS	0.808	0.643	0.066	0.701	0.769	0.219	0.619	0.211	0.846	0.154	0.5	0.136
smote_tomek ADS	0.984	0.622	0.068	0.665	0.79	0.204	0.571	0.192	0.858	0.135	0.562	0.12
smote_enn ADS	0.656	0.762	0.048	0.829	0.724	0.278	0.476	0.266	0.788	0.231	0.312	0.207
Correlation Remover ADS	0.835	0.697	0.054	0.751	0.781	0.217	0.524	0.204	0.843	0.163	0.438	0.142
GridSearch ADS	0.816	0.438	0.009	0.88	0.903	0.077	0.714	0.067	0.895	0.068	0.875	0.065
Threshold Optimizer ADS	0.816	0.438	0.005	0.933	0.905	0.075	0.714	0.064	0.883	0.08	0.875	0.078

Table 2. Fairness metrics across modifications.