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

Jon Dinh, Yash Jha

Background of Automated Decision System

Dataset: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

ADS: https://www.kaggle.com/code/tanmay111999/stroke-prediction-effect-of-data-leakage-smote/notebook

Purpose of the ADS: 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). The goal of this ADS is a binary classification problem, to classify whether a patient is will suffer a stroke from provided features.

Data Information

```
<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
```

Rows: 5110 observations

Columns: 10 features, 1 target feature

Source: confidential source for education purposes

Feature Names:

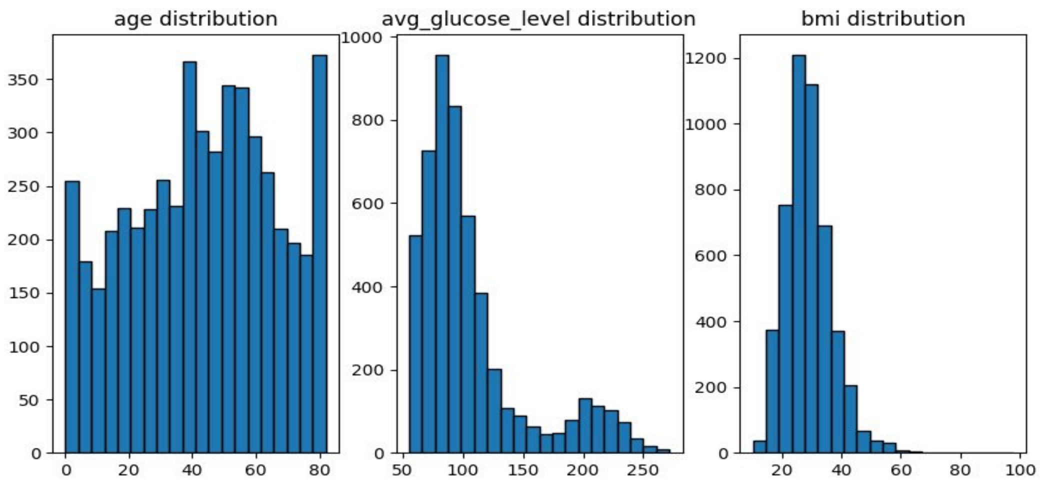
- **Categorical**: gender, hypertension, heart disease, residence type, work type, smoking status, marital status
- **Numerical**: age, average glucose level, BMI

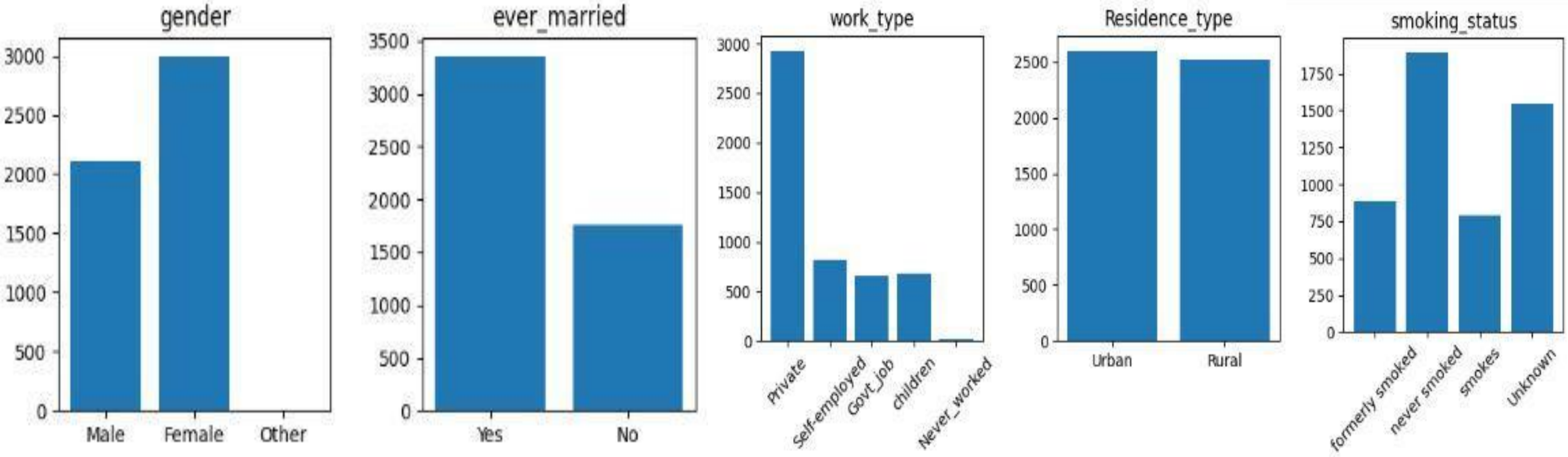
Missing values: only 201 missing values for BMI

Output: binary variable indicating stroke prediction (1), no stroke (0)

Distributions

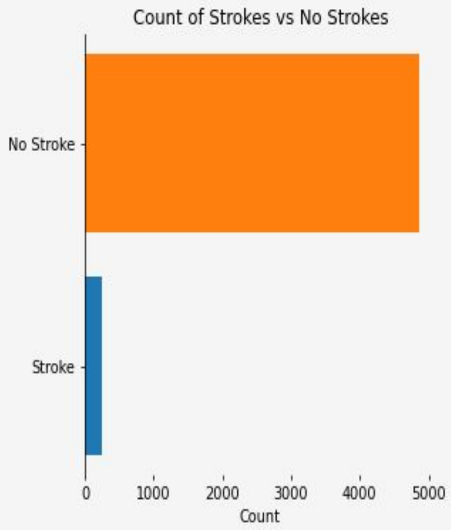
- Numerical on top
- Categorical on bottom

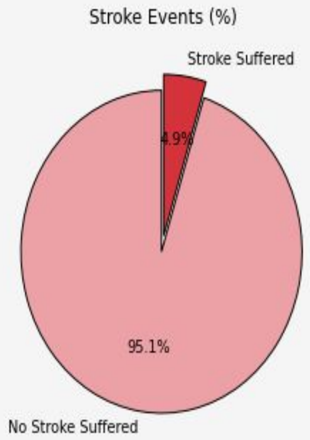


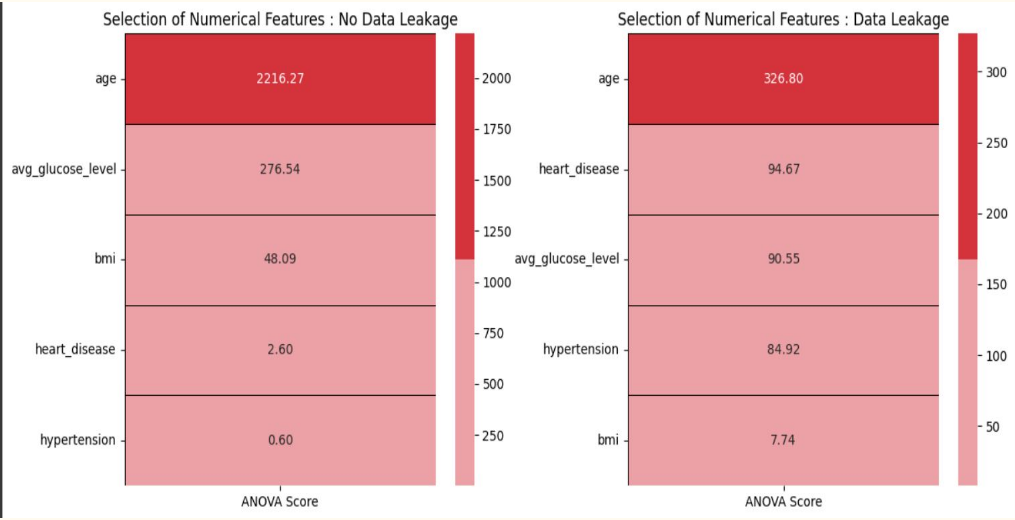


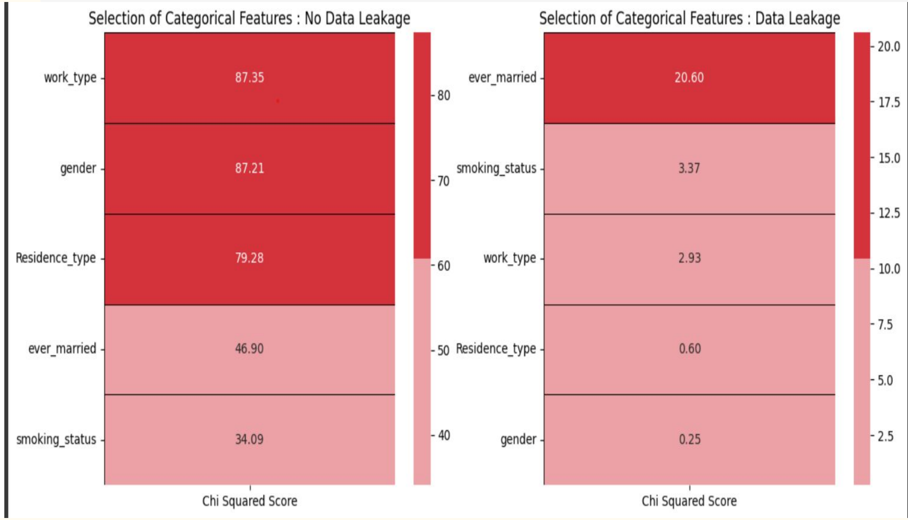
Implementation and Validation

- Class imbalance - SMOTE
- Features dropped: smoking_status, heart_disease, hypertension, BMI
- Missing value imputation - mean



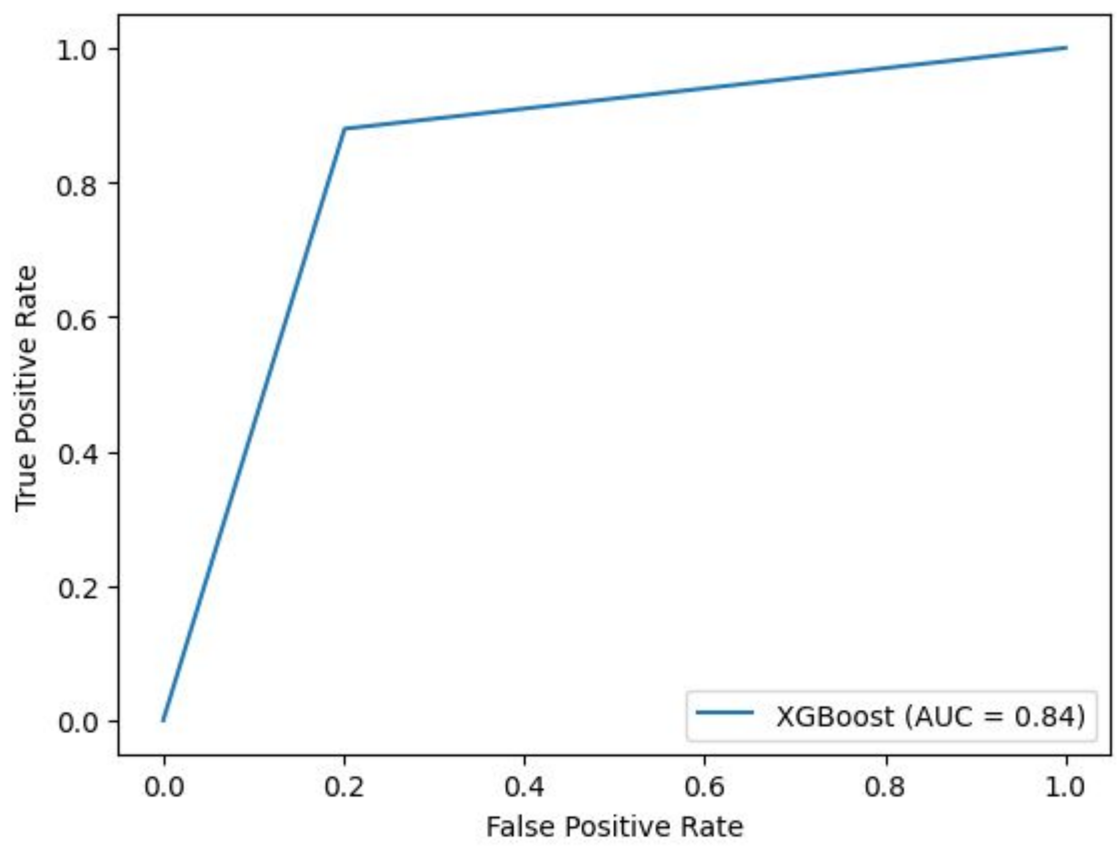






Implementation and Validation cont.

- **ADS:** XGBoost
    - learning_rate = .01
    - max_depth = 3
    - n_estimators = 1000
- **Cross val score:** 91.82%

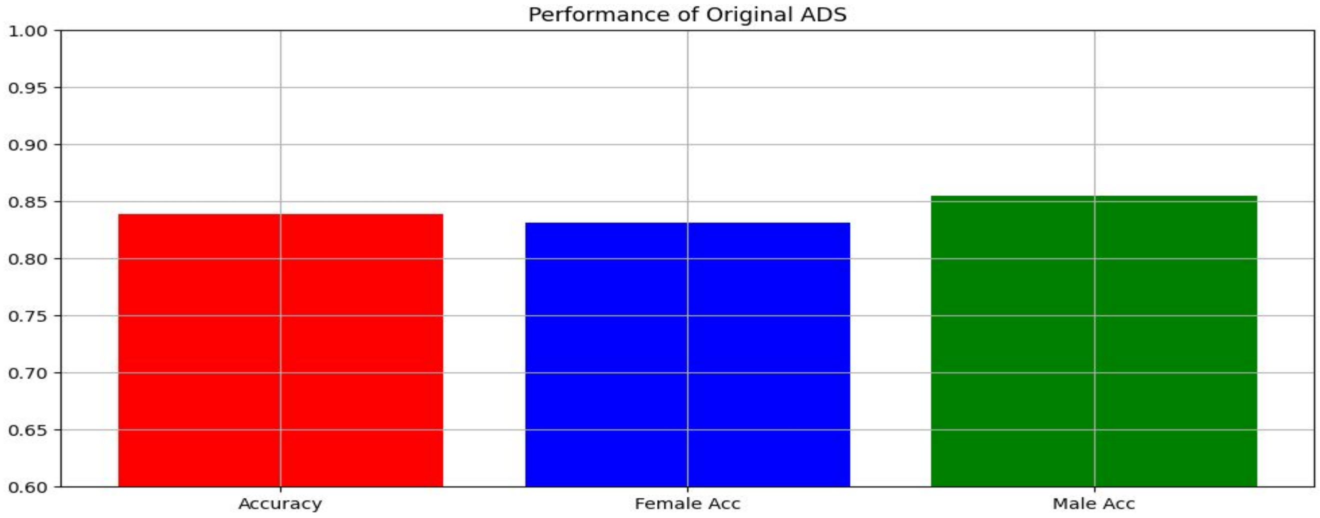


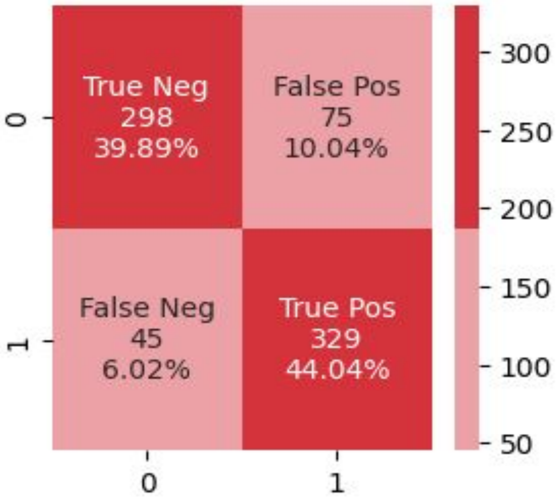
Outcome: Performance

- **Accuracy**: .839, **FNR** = .12
    - Consistent across gender groups
- **Precision, Recall, F1**
    - Assess ability to classify positive samples correctly

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

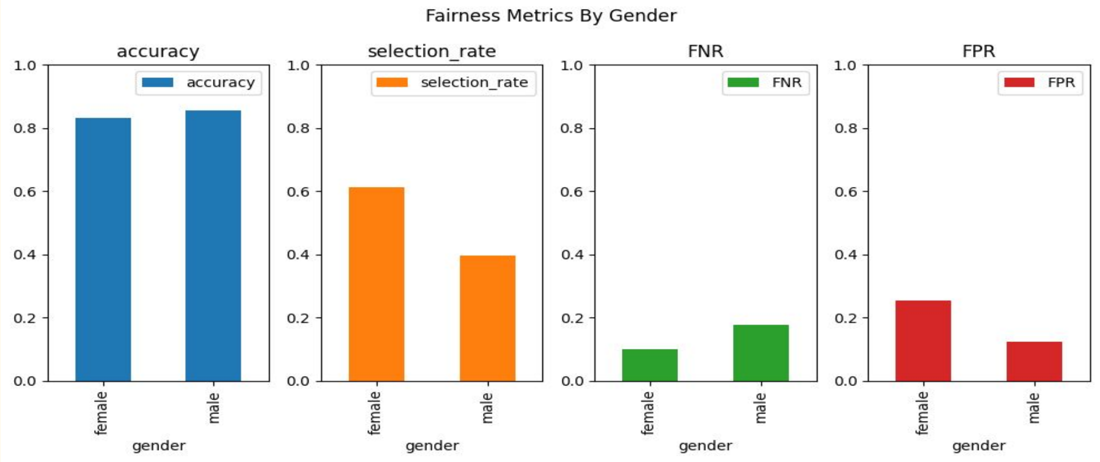




FNRP	equalized_odds_ratio	demo_parity_diff	demo_parity_ratio
0.569	0.493	0.218	0.644

Outcome: Fairness

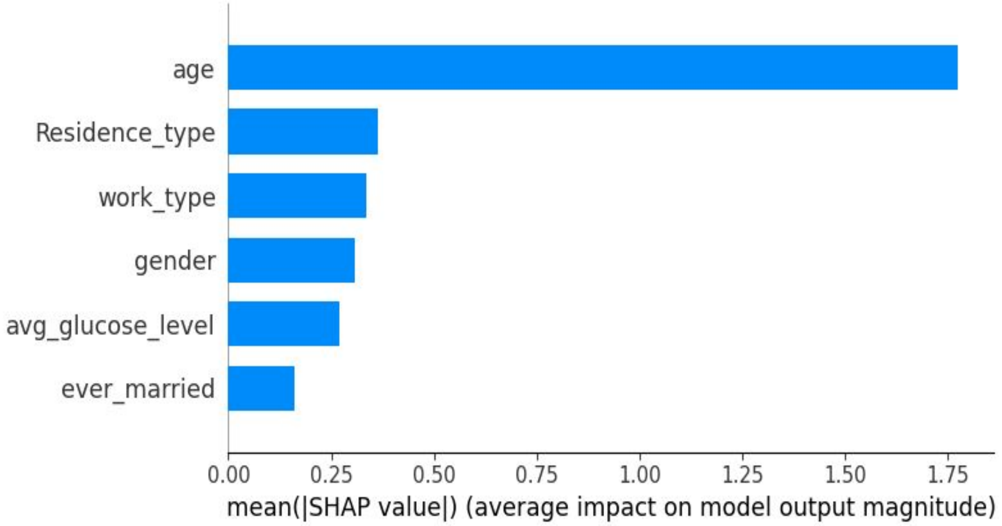
- **Fairness across gender groups:**
  - differing selection rates
  - bias with **FNR** and **FPR**
- **Other fairness metrics:** FNRP, EOR, DPR; relatively low

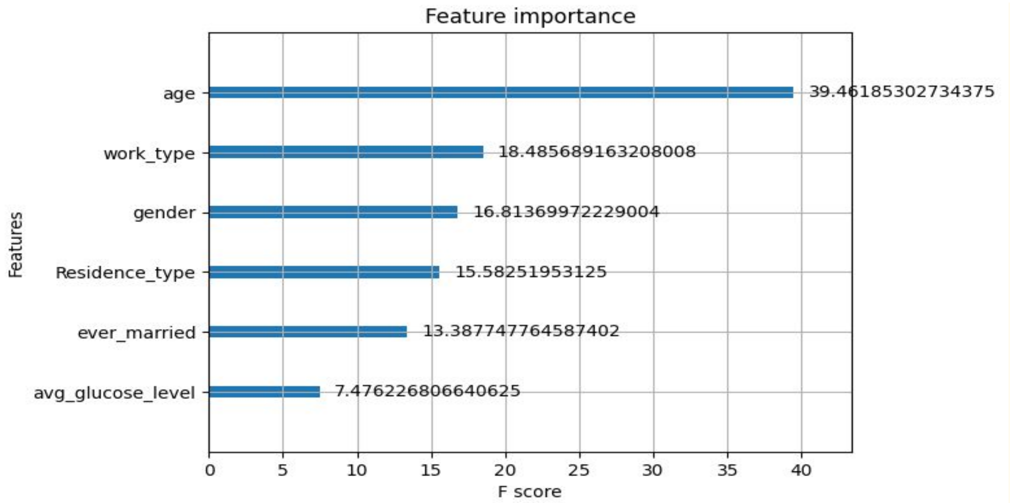


Interpretability of ADS

Feature Importance (all data):

- (1) SHAP
- (2) XGBoostClassifier

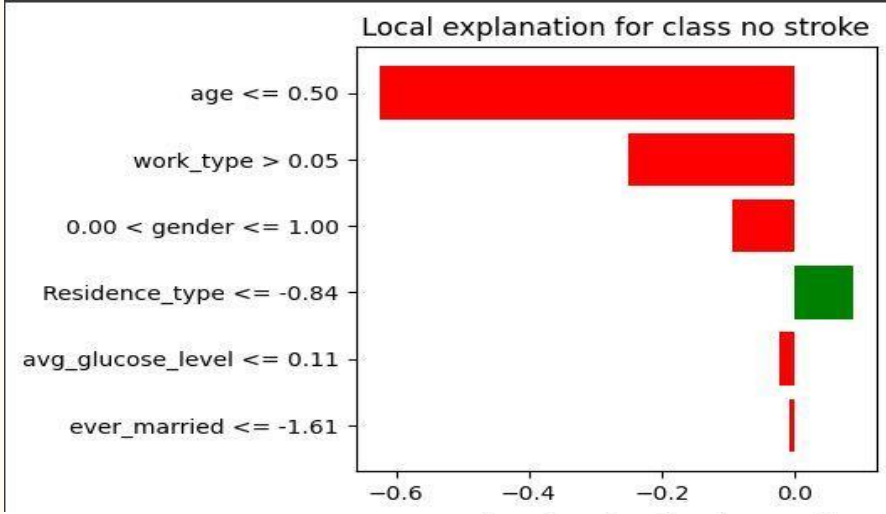


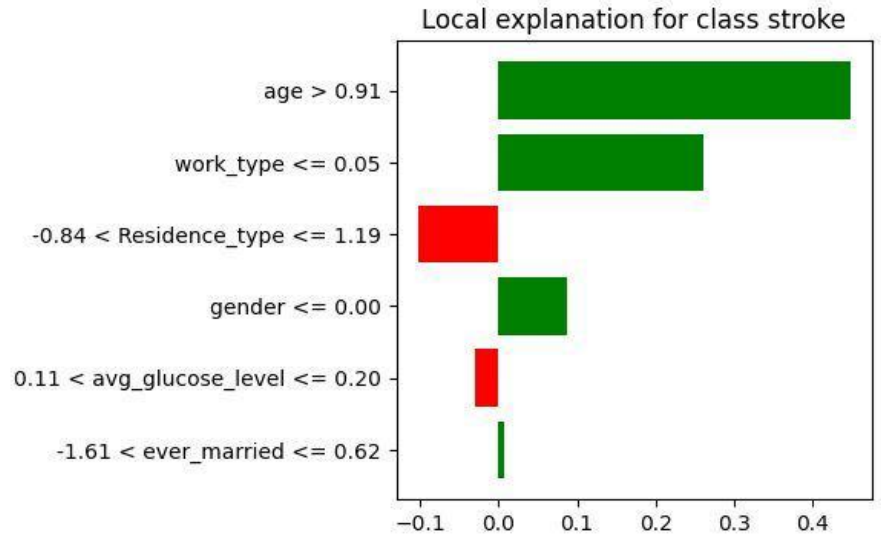


Interpretability of ADS cont.

Local Explanation of Features (LIME):

- Observation # 1
  - Actual: 0
  - Predicted: 0
  - Gender: Male
- Observation # 2
  - Actual: 1
  - Predicted: 1
  - Gender: Female





ADS Discussion

- **Accuracy**: .839, **FNR** = .12
  - Accurate, yet **FNR** is skewed by synthetic class distribution
- **Stakeholders:**
  - **Patients**: interested in fairness with respect to gender, should not bias stroke prediction. Interested in overall precision, recall and f1-score.
  - **Care-Givers, Staffing:** interested in accuracy, particularly **FNR** (can't miss positive cases)
  - **Third parties:** interested in both for marketing (**fairness**) and reliability (**accuracy**)
    - e.g. hospital boards, equity firms, etc.
    - can't be liable for misclassification
- **Optimization:**
  - **Accuracy**: FNR, F1
  - **Fairness**: FNPR, DPR, EOR

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

Deployment?

Modifications: (0) original ADS, (1) ADASYN , (2) SMOTE-Tomek, (3) SMOTE-ENN, (4) correlation remover, (5) hyperparameter tuning, (6) threshold optimizer

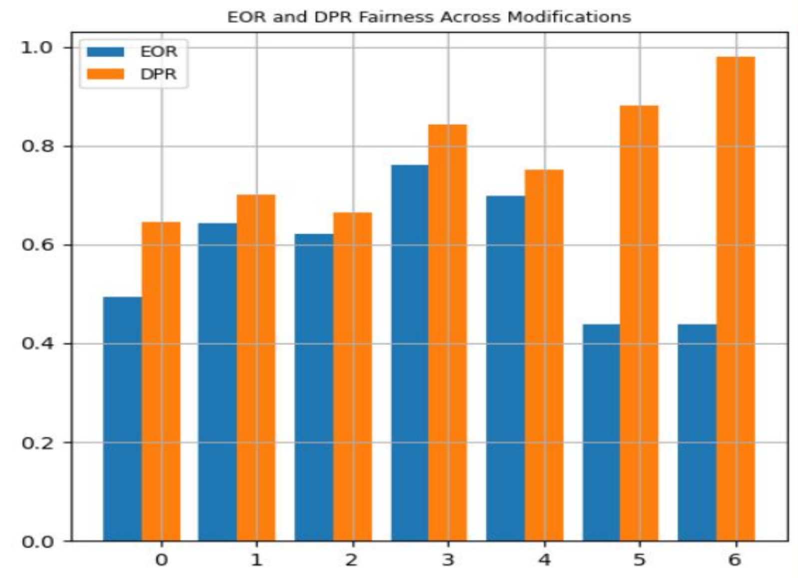
Would we feel comfortable deploying the model?

- **Publicly:** good starting point
- **Industry:** not really
  - Very vulnerable to new data
  - High **FNR** too risky

Focus: Higher EOR and DPR than the original ADS (idx 0)

Simple modifications to system process

- significant 10% increase in acc and f1 score
- significant increase in FNRP
- significant increase in DPR
- either same tight range or increase in EOR



Conclusion

- **Solid ADS:**
  - Data was appropriate
  - Good accuracy, great FNR/FPR for synthetic class distribution
- **Weaknesses:**
  - Not robust to new data; real world will have imbalanced class distribution
  - Fairness with respect to gender
- **Improvements**
  - Our fairness modifications (other SMOTE techniques, correlation remover, threshold opt)
  - Our accuracy modifications (hyperparameter tuning)
  - Potentially better feature space in the dataset that can more easily pick out positives
  - Stronger decision-tree ADS capable of finding thresholds