DS-GA 1017 Responsible Data Science – Project Presentation

Evaluation of a Stroke Prediction System: Performance, Transparency, Fairness, and Beyond

Yuheng Yang (yy2597), Jialing Li (jl9716)

Background

- Stroke accounts for approximately 11% of all deaths globally
- Many survivors face long-term disabilities
- BUT: early detection helps a lot!
- Need an ADS that consumes health records and predict whether an individual will have a stroke in the future.

Kaggle User Saimon Dahal:

- Designed an ADS that uses Electronic Health Record (EHR) to extract features related to stroke and make preds.
- Binary classification. 0 for no stroke, 1 for stroke
- Optimization, data pre-processing, evaluation

Is the performance desirable? Is the system fair? What are the social implications?

System Being Audited: Input Data

Data Source

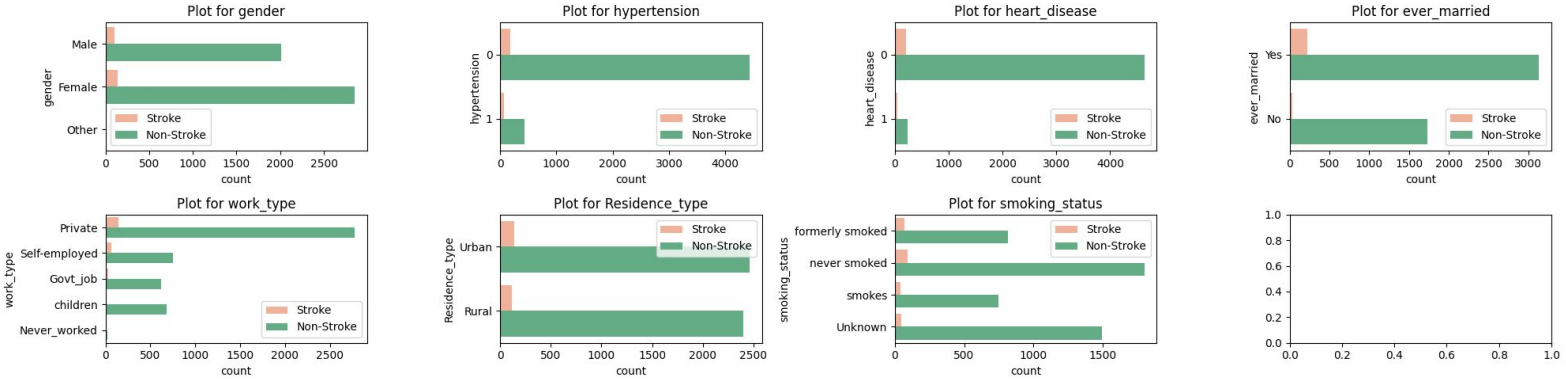
- **HealthData.gov**, United States Government Department of Health & Human Services
- Each record in the dataset is derived from the Electronic Health Records (EHR) of an individual
- 5110 entries and 12 columns

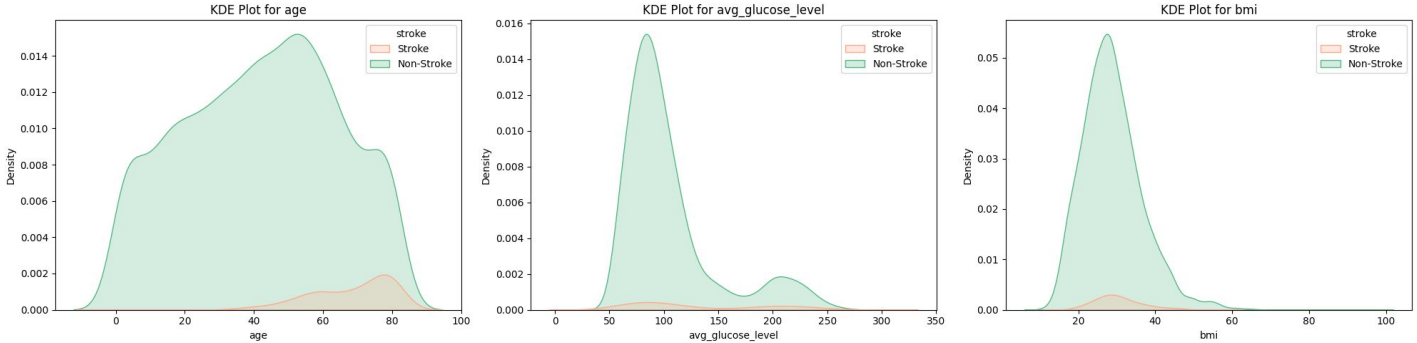
Data Profiling Summary

Column	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
Data Type	int64	object	float64	int64	int64	object	object	object	float64	float64	object	int64
Completeness	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete	Complete	201 Missing	Complete	Complete
Explanation	id	gender	age	whether hypertension	whether heart disease	if ever married	work type	area of residence	glucose level	BMI	whether smokes	whether stroke

System Being Audited: Input Data

Data Distribution



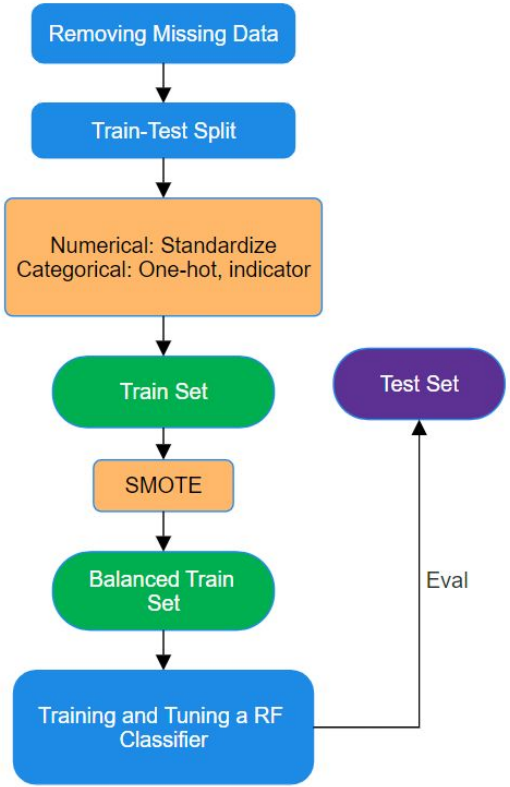


System Being Audited: Pipeline

Procedures

- Standard pre-processing
- Serious label imbalance may hurt model training
  - Solution: SMOTE
  - Upsampling minority class (stroke)
  - Balanced training set and leave test set untouched
  - **May creates serious problem**
- Experimented with multiple model choices and proposed a tuned Random Forest Classifier

- **In 5-fold cross-validation with balanced set, Acc=0.96**
- **Does it qualify a good design?**



Audit 1: Performance Overall

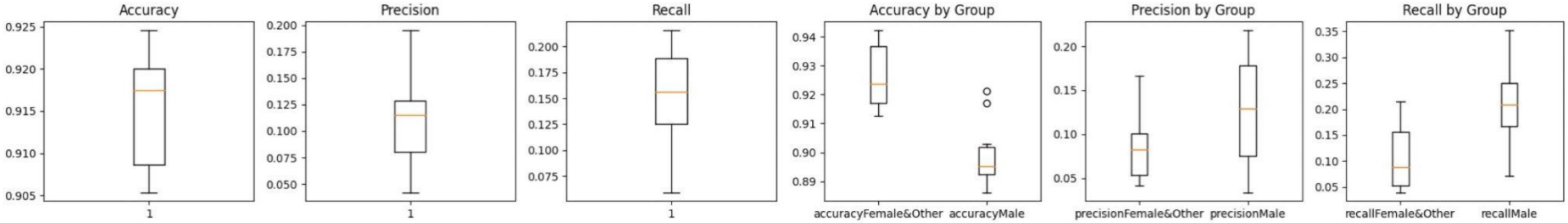
Healthcare Providers: We want conclusive results. Don't waste medical resources

Good Precision

Potential Patients: We want those truly in risk to be properly identified

High Recall

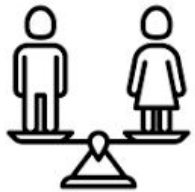
Experiments: 10 random runs



Audit 2: Fairness between Gender Groups

Problem Characterization

- Medical resource distribution: a limited resource allocation problem
- Goal: predict future possibility of stroke and ensure future performance–forward facing
- Want to ensure the ADS do not systematically underestimate the risk of a particular group



A good theoretical framework: Formal-Plus EO

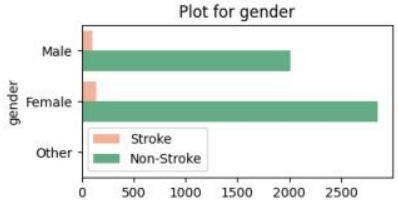
- Khan & Chouldechova:
- Error rate balance satisfies Formal-Plus EO
  - Equal FPR and FNR
- Equalized odds satisfies Formal-Plus EO
  - Error do not track group membership

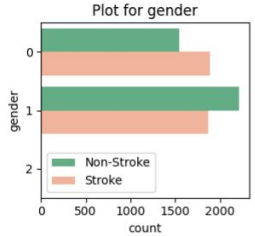
Choice of fairness metrics: Difference in FPR and FNR (absolute), FPR M-F&O, and FNR M-F&O (bias direction)

Audit 2: Fairness between Gender Groups

Another Problem: SMOTE

- SMOTE is a distance-based method.
- Susceptible to features dominating a group.
- Upsampling the minority class may overrepresent a particular gender group
  - So the model may deem the overrepresented group more associated with an outcome
- Exacerbates technical bias, introduces unfairness across gender groups through learning





Males positive instances are upsampled disproportionately after SMOTE

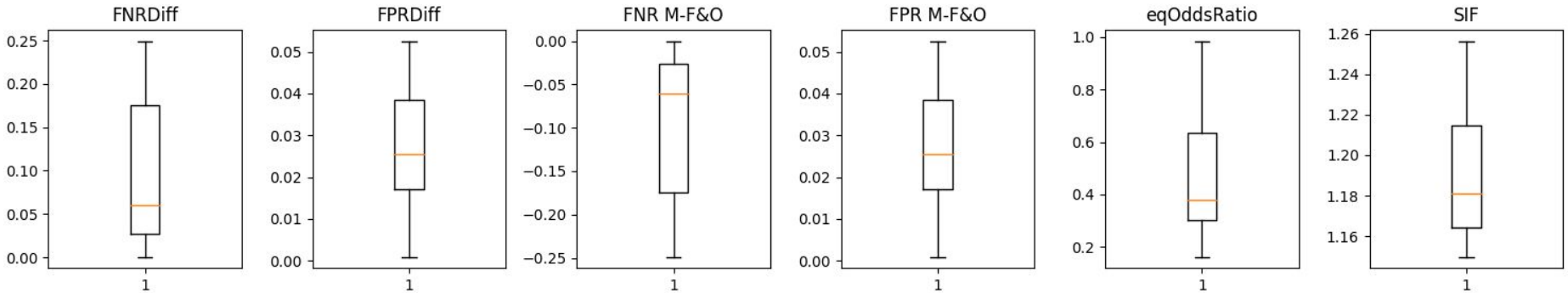
A Customized Fairness Metric: SMOTE Inflation Factor (SIF)

$$SIF = \frac{\text{proportion of stroke in male after SMOTE}}{\text{proportion of stroke in females and other after SMOTE}} \Big/ \frac{\text{proportion of stroke in male before SMOTE}}{\text{proportion of stroke in females and other before SMOTE}}$$

By recording the SIF at each run, we can monitor if the positive instances are upsampled in its original proportion across genders and if any subgroup is over-represented in the SMOTE process.

Audit 2: Fairness between Gender Groups

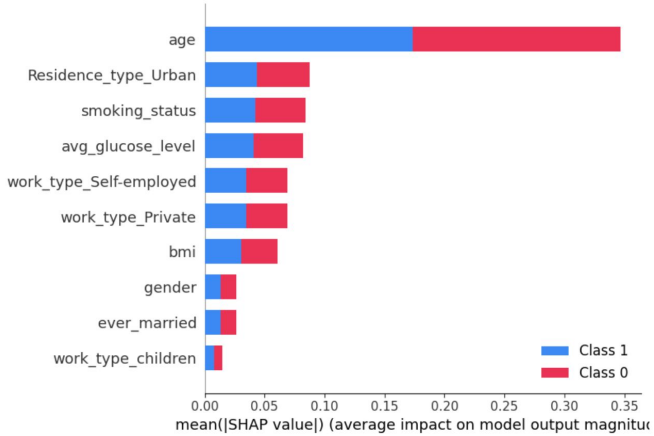
Experiments on 10 random runs



- Difference in FNR and FPR stably>0, equalized odds differs from 1: error rates track group membership, not fair
- FNR M-F&O stably<0 and FPR M-F&O stably>0: ADS underestimates the females and other group
  - Notably, more false negatives: disastrous disparity in a high-stake event
- SIF stably>1, SMOTE systematically amplifies the proportion male samples in the positive class than in the original dataset, distorting the true data distribution
  - Inflating the representativeness of males. It may introduce or exacerbate the system's bias

Audit 3: Interpretability and Transparency

Global Interpretability: Feature Importance



Local Interpretability: Forceplot

- Global:
  - Age is the most significant factor, followed by residence type, smoking status, and glucose levels.
- Local:
  - The results in individual instances are consistent with the feature importance scoring
- Bias Alarm: wrong representation of smoking_status during the encoding process.
  - ‘smokes’: 1, ‘never smoked’: 2, ‘formerly smoked’: 3, ‘Unknown’: 4.
  - Should adopt one-hot encoding

Conclusion

- We conclude that this ADS is inappropriate to be deployed in real-life setting
  - An imbalanced dataset that leads to an issue of predicting minority class
  - Unfair predictions for different gender groups
  - Wrong contribution of smoking status
  - Thus unable to provide valid and constructive information to the identified stakeholders
- Future ADS Improvements
  - Find a more balanced dataset: should involve a more diverse feature distribution for each class
  - Evaluate ADS before its implementation to optimize the chosen metrics
  - Define a customized metric (e.g., SIF in our case) to track the distribution change regarding the sensitive features to mitigate the technical bias.
  - Implement an appropriate encoding system (e.g. one-hot encoding for smoking status) to enhance model interpretability
  - Track the model performance and post-process the results in a timely manner to mitigate the bias

Thank you