# High Stakes, Hidden Signals: ML-Based Stroke Risk Assessment

Wil Sheffield

# Why Stroke Prediction?

- Stroke is a leading cause of death and disability worldwide (Approx. 140,000 deaths annually in US)
- Identifying high-risk individuals can enable preventative care and intervention
- ML approaches allow us to improve predictive accuracy over rule-based approaches

# What Makes Stroke Risk Complex?

- Medical data often has non-linear relationships, letting tree based models like XGBoost or Random Forests to be more suitable
- Socio-demographic features can also have unpredictable effects which encourages the use of complex models

# Objective

Identify the best binary classifier model and most important feature(s) in predicting the presence of stroke based on patient data

- Compare multiple models to identify the best performing one
- *Understand which feature(s) contribute most significantly to prediction accuracy*

# State of the Art

**Common Approaches:**

- Individual Models: Logistic Regression & LightGBM
- Multiple Classifier Models: XGBoost & CatBoost

**Our Take:**

- Compare diverse set of classifiers head-to-head
- Focus on healthcare-critical metrics (Recall, AUC)
- Identify importance of features via SHAP

# Research Questions & Hypotheses

**RQ1:** Which model performs best for stroke prediction?

- **H1:** XGBoost will outperform other models due to its handling or nonlinearity and imbalance among features

**RQ2:** Which features matter most in terms of prediction?

- **H2:** Known, directly relevant features like age, BMI, and glucose levels will most strongly predict

# Dataset Overview

**Source:** Kaggle "Tabular Stroke Prediction Dataset"

**Size:** 15,304 samples

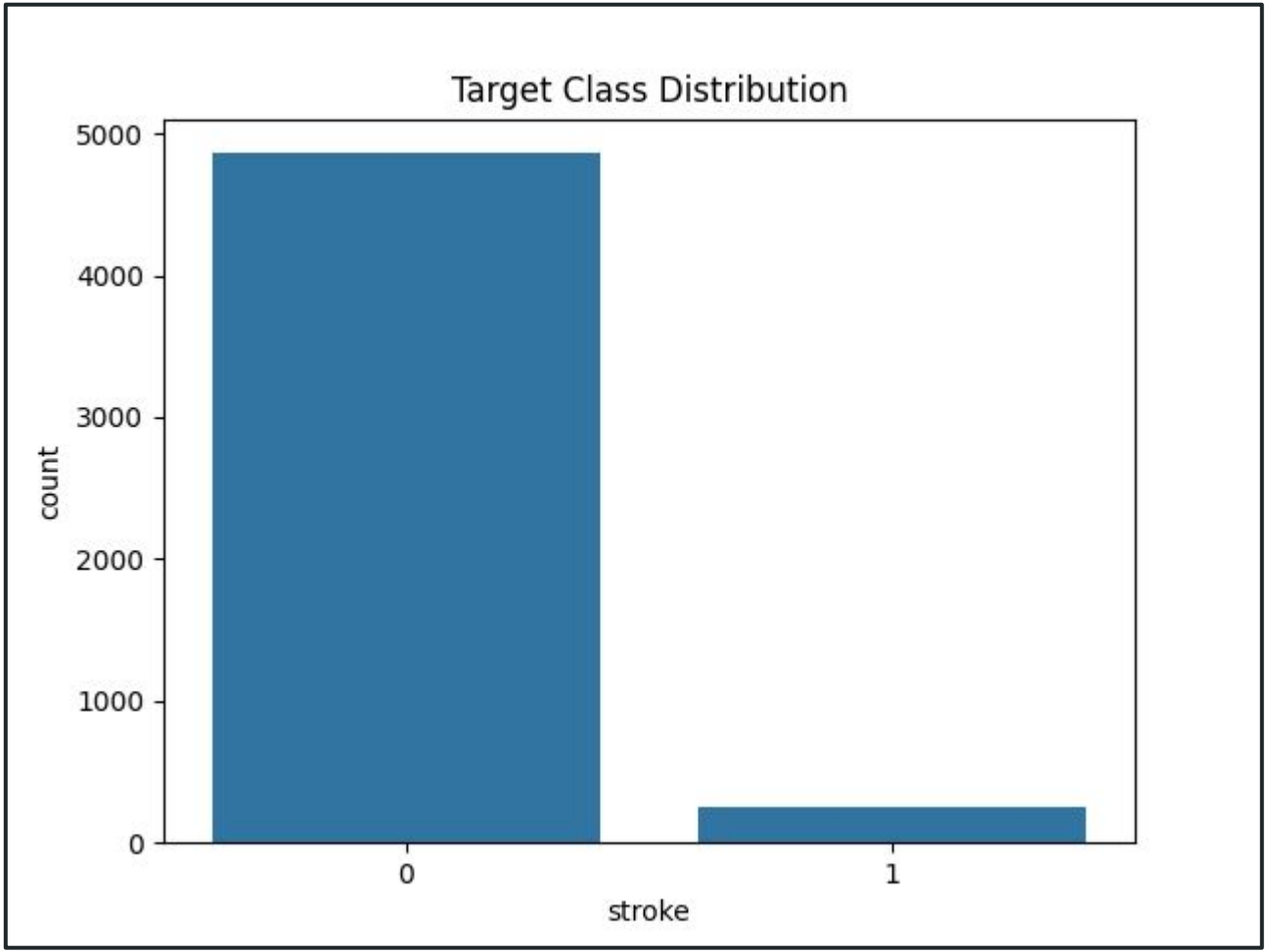
**Numerical Features:** age, avg_glucose_level, bmi

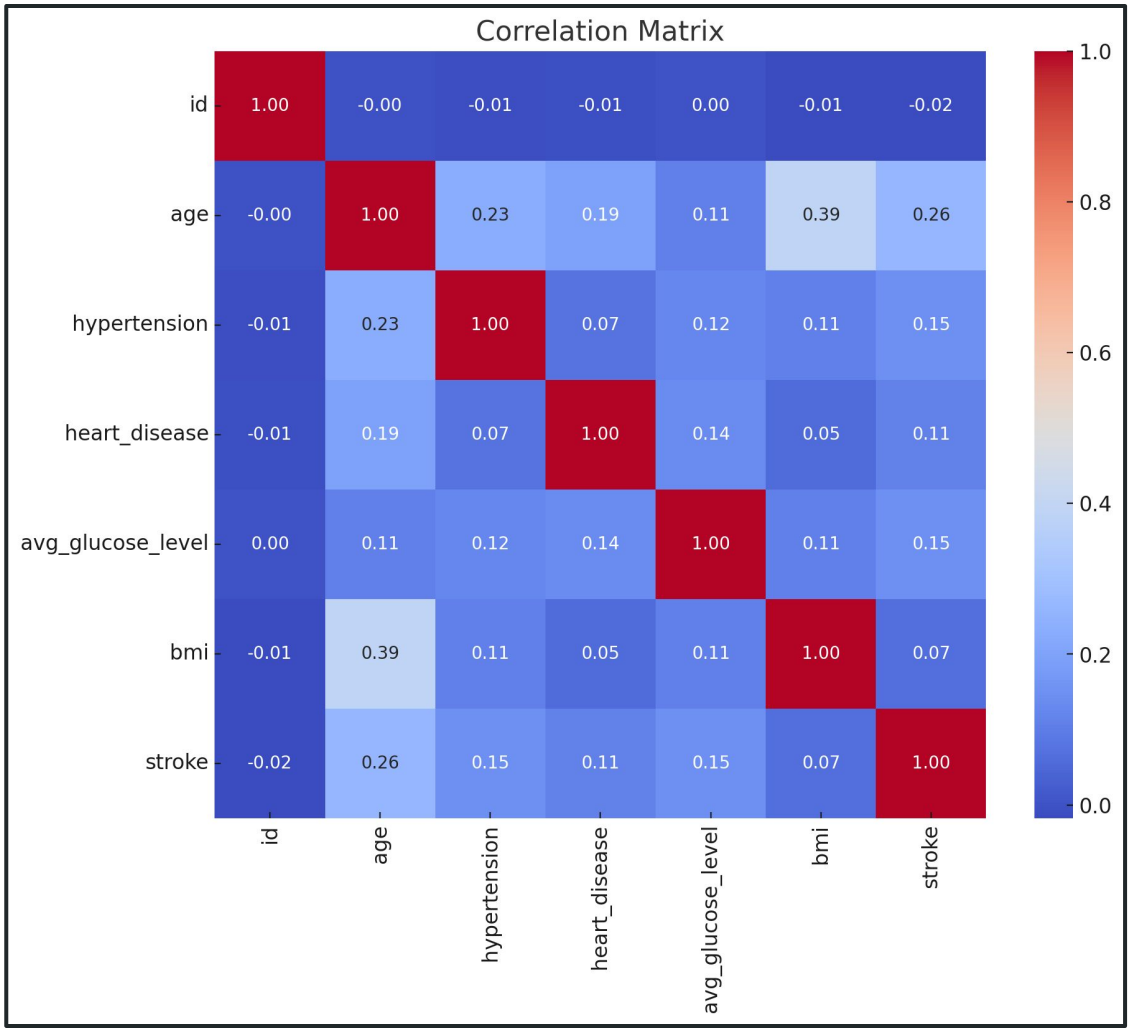
**Categorical:** gender, married, work type, Residence type, smoking status

**Target Variable:** stroke (binary: 0 or 1)

# EDA: Data Exploration

- Class imbalance is present as stroke cases are rare compared to non-stroke cases
- Missing values were handled in bmi with mean imputation
- Categorical features were one-hot encoded, preserving info without assumptions on ordinal relationships


Target Class Distribution
5000
4000
3000
2000
1000
0
count
0
1
stroke


Correlation Matrix
id
age
hypertension
heart_disease
avg_glucose_level
bmi
stroke
1.00 -0.00 -0.01 -0.01 0.00 -0.01 -0.02
-0.00 1.00 0.23 0.19 0.11 0.39 0.26
-0.01 0.23 1.00 0.07 0.12 0.11 0.15
-0.01 0.19 0.07 1.00 0.14 0.05 0.11
0.00 0.11 0.12 0.14 1.00 0.11 0.15
-0.01 0.39 0.11 0.05 0.11 1.00 0.07
-0.02 0.26 0.15 0.11 0.15 0.07 1.00
1.0
0.8
0.6
0.4
0.2
0.0

# Methodology: Modeling Approach

**Models Used:**

- Logistic Regression
- Decision Trees
- Random Forest
- XGBoost
- k-Nearest Neighbors

**Validation Strategy:**

- Train/Test Split
- 5-Fold Cross Validation
- Metrics: Accuracy, Precision, Recall, F1, AUC

| Model | Accuracy | Precision | Recall | F1 | ROC AUC |
|---|---|---|---|---|---|
| XGBoost | 0.7808219178 | 0.157480315 | 0.8 | 0.2631578947 | 0.8463168724 |
| Decision Tree | 0.9060665362 | 0.140625 | 0.18 | 0.1578947368 | 0.5617078189 |
| Random Forest | 0.9500978474 | 0.4285714286 | 0.06 | 0.1052631579 | 0.7859979424 |
| Neural Net (MLP) | 0.9491193738 | 0.375 | 0.06 | 0.1034482759 | 0.8016666667 |
| Logistic Regression | 0.9510763209 | 0 | 0 | 0 | 0.8413168724 |
| k-NN | 0.9432485323 | 0 | 0 | 0 | 0.5966872428 |

For stroke prediction, it is best to focus on **recall, F1, and AUC**

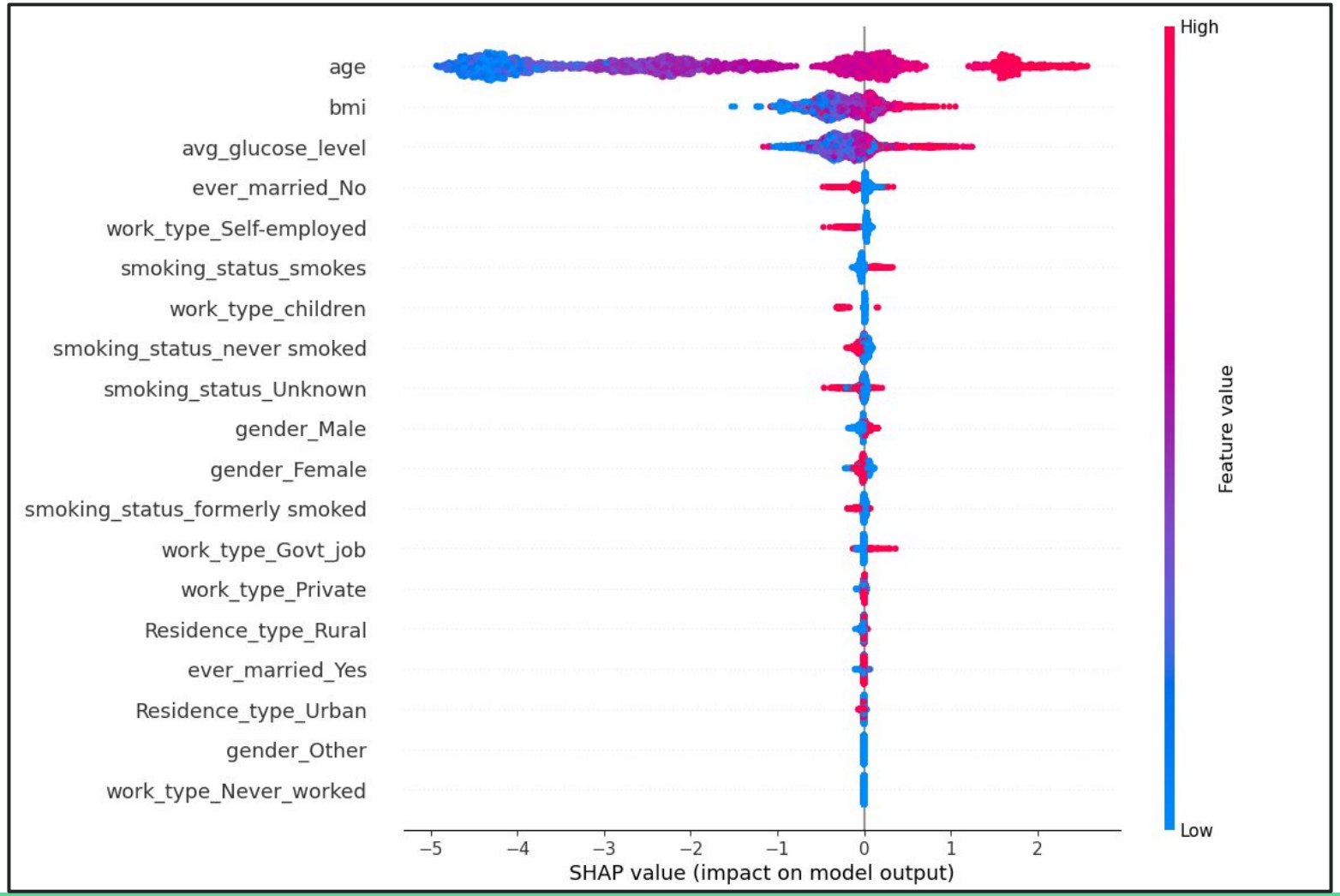
# Results: XGBoost

- Highest recall of 0.80: Most effective at identifying actual stroke cases
  - Despite having lower accuracy, it is more sensitive to the minority class
  - ROC AUC = 0.84, indicating strong predictive power
- Other models had incredibly low recall, indicating failure to identify stroke cases
- Random Forests and MLP fared worse due to class imbalance

# Results: Feature Significance

**Top Predictors (XGBoost)**: age, avg_glucose_level, & bmi

- SHAP values confirm that aforementioned numerical features dominate stroke prediction in terms of significance


age
bmi
avg_glucose_level
ever_married_No
work_type_Self-employed
smoking_status_smokes
work_type_children
smoking_status_never smoked
smoking_status_Unknown
gender_Male
gender_Female
smoking_status_formerly smoked
work_type_Govt_job
work_type_Private
Residence_type_Rural
ever_married_Yes
Residence_type_Urban
gender_Other
work_type_Never_worked
−5
−4
−3
−2
−1
0
1
2
SHAP value (impact on model output)
High
Feature value
Low

# Implications

In healthcare prediction tasks, recall matters more than accuracy as false negatives could mean missed early intervention

- XGBoost’s high recall makes it a strong candidate for real-world deployment or further clinical testing
- While simpler models like Logistic Regression are more interpretable, it may not offer life-critical predictions

Supports the use of tree-based ensemble methods in health risk modeling

# Potential Next Steps

- Larger, more diverse dataset & EHR integration
  - Include feature mapping instances to region
- More neural network experimentation
- Feature pruning and domain knowledge-based feature engineering

# Thank You!
# Any Questions?