

Predicting Stroke

...

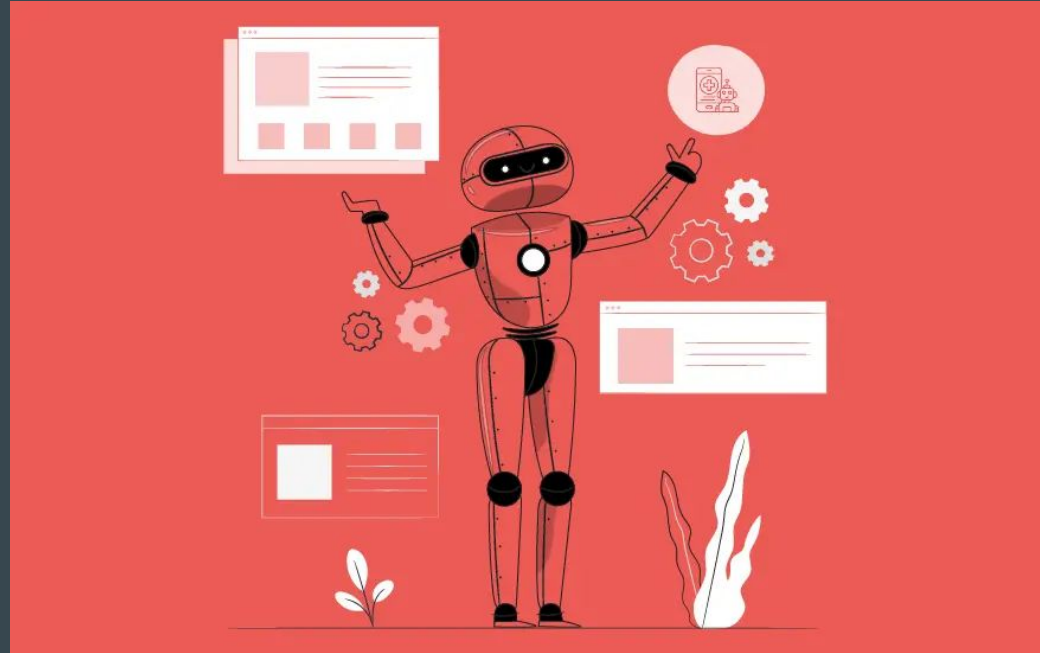
Leveraging Machine Learning Techniques in
Electronic Health Records

Introduction

Final Project: Machine Learning

Outcome Prediction:
Stroke

Predictors: Electronic
Health Records (EHR)



Research Focus

Objective:

Predict stroke using machine learning on electronic health records (EHR).

Key Questions:

1. Can machine learning models accurately predict stroke outcomes?
2. How do top risk factors affect predictive power?
3. Can gender-specific models generalize across genders?

Data Description

Source: Synthea Stroke Synthetic Patient Data Series for Risk Prediction ML [Chen, AJ \(2022\)](#)

Sample Size: 32,034 Patients

Outcome: Stroke

Stroke: 8,197 (25%)

No Stroke: 23,837 (75%)

Dimensions:

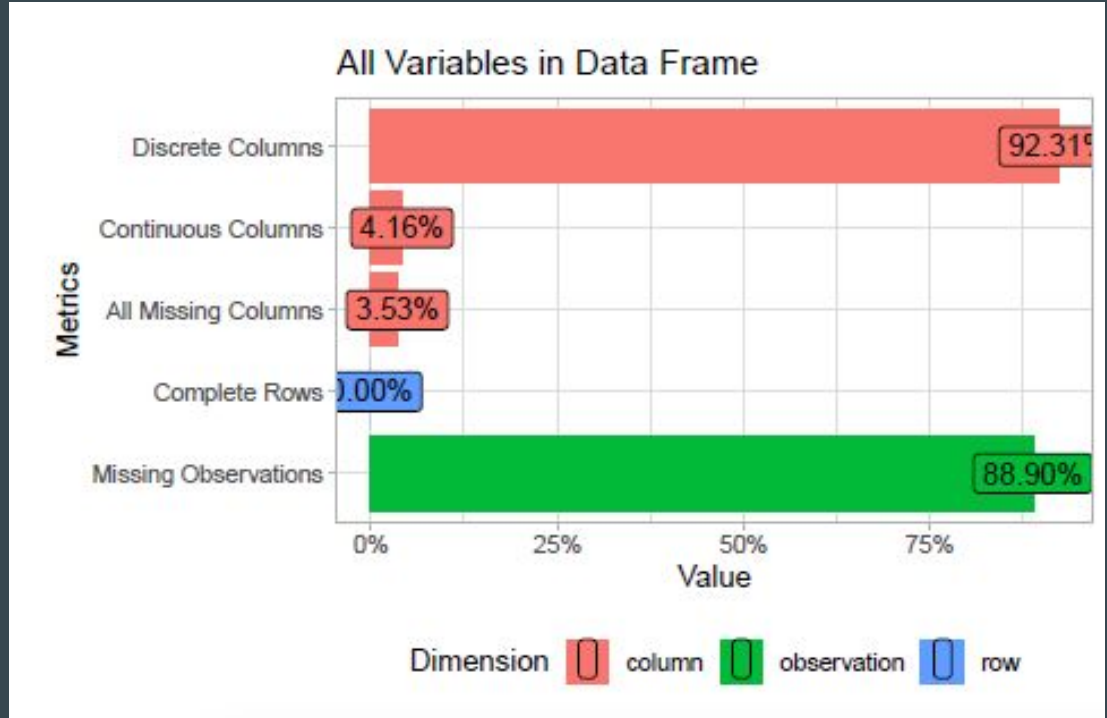
- Patient Number
- Stroke Outcome Label
- 800 EHR Variables
 - Lab Results
 - Survey Findings
 - Procedures
 - Diagnosis
 - Medical History
 - Medications
 - Immunizations
 - Allergies

High Dimensional, Sparse Data

High Dimensionality: 800 Variables

Row Sparsity: 0% of rows are complete


Overall Sparsity: 88.9% of data is blank



Top Risk Factors for Stroke

1. High blood pressure
2. Heart disease
3. Diabetes
4. Smoking
5. Birth control pills
6. History of TIAs (mini-strokes)
7. High red blood cell count
8. High blood cholesterol and lipids
9. Lack of exercise
10. Obesity
11. Excessive alcohol use
12. Illegal drugs
13. Abnormal heart rhythm
14. Cardiac structural abnormalities
15. Older age
16. Race
17. Gender

Top Risk Factors for Stroke

1. High blood pressure
2. Heart disease
3. Diabetes
4. Smoking
5. Birth control pills
6. History of TIAs (mini-strokes)
7. High red blood cell count
8. High blood cholesterol and lipids
9. Lack of exercise
10. Obesity
11. Excessive alcohol use
12. Illegal drugs
- 13. Abnormal heart rhythm** 
14. Cardiac structural abnormalities
15. Older age
16. Race
17. Gender

- [1] atrial fibrillation
- [2] catheter ablation of tissue of heart
- [3] insertion of biventricular implantable cardioverter defibrillator
- [4] 3 ml amiodarone hydrochloride 50 mg/ml prefilled syringe
- [5] atropine sulfate 1 mg/ml injectable solution
- [6] digoxin 0.125 mg oral tablet
- [7] electrical cardioversion

800 Variables
-Original Dataset-

124 Variables
-Top Risk Factors-

20 Risk Groups

Highly
Likely
Variables

Possibly
Likely
Variables

1

TRUE

0.25

0

NA

0

1

ABNORMAL

0.25

0

NORMAL

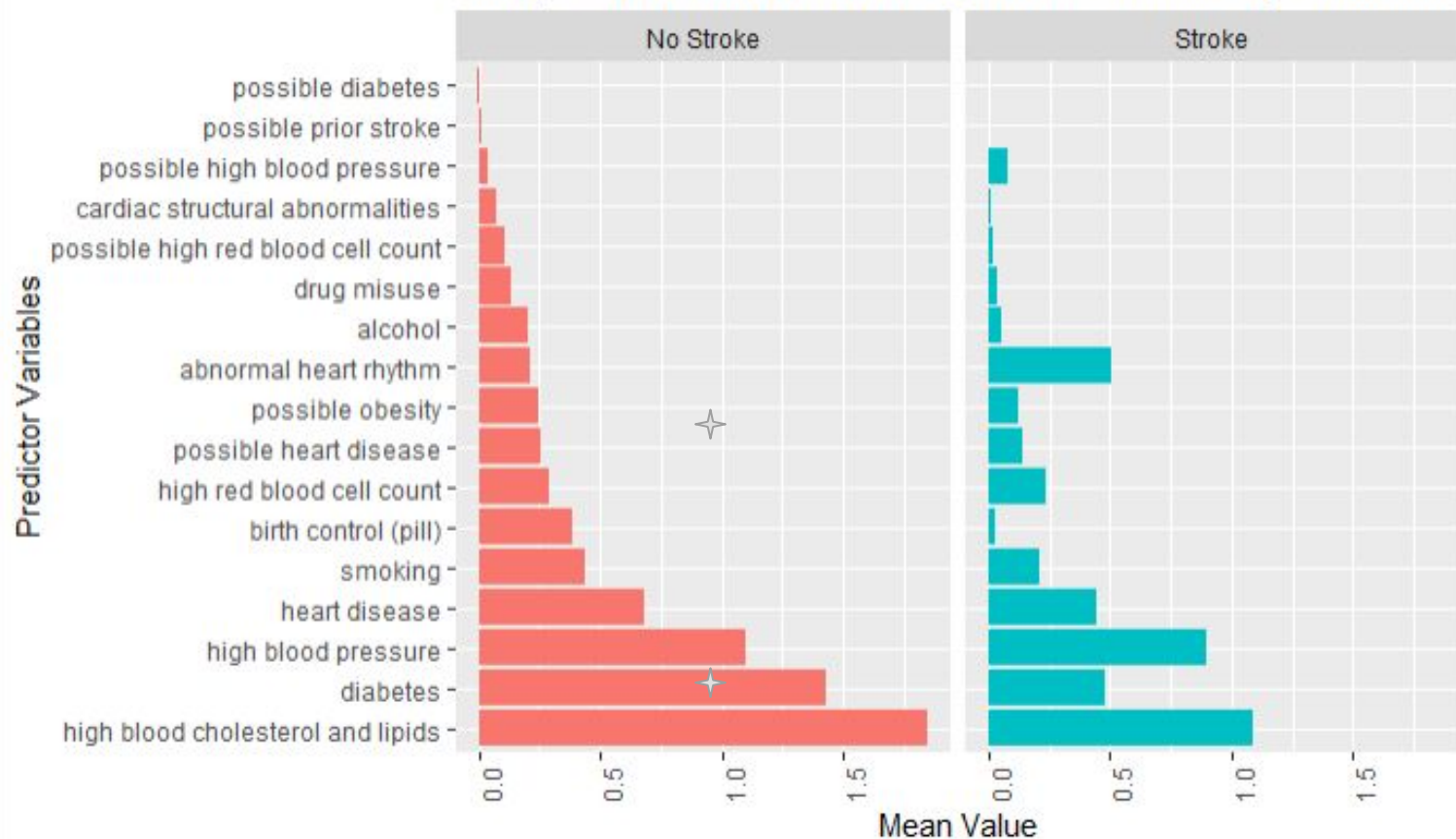
NA

0

Total

Total

Comparison of Mean Values for Each Variable by Outcome



Methodology

Preprocessing: Dimensionality reduction

Training 80%: Testing 20%

Feature Engineering:

- One-hot encode - character variables
- Impute mean (except XGBoost models)

Model Evaluation:

- Cross-validation: 5-fold with 3-repetitions

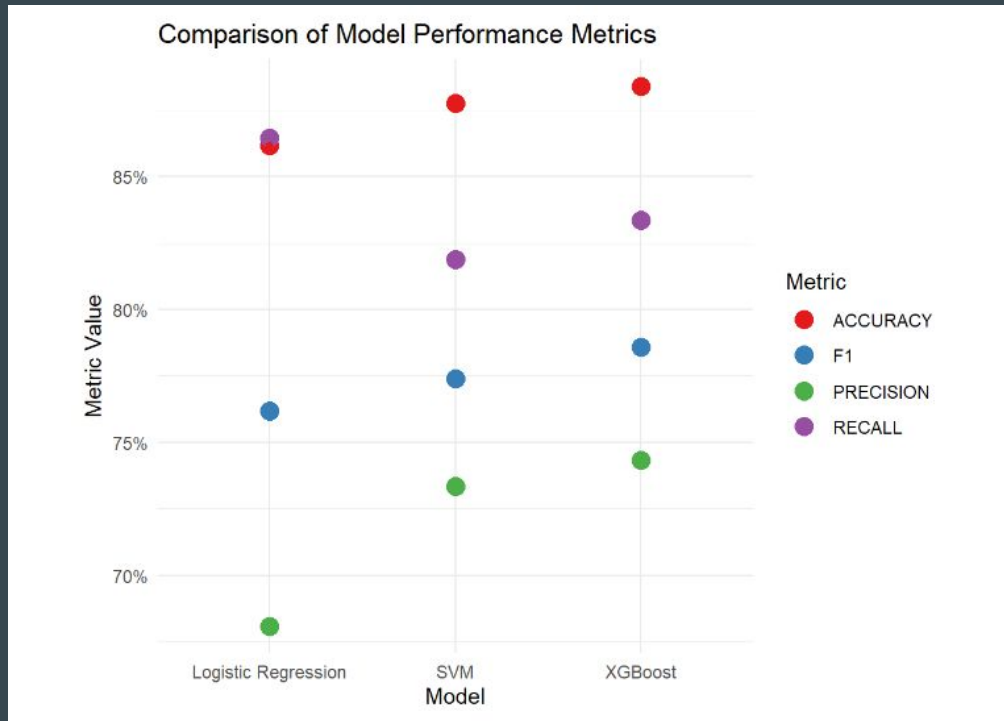
Machine Learning Models

Logistic
Regression

Support
Vector
Machines

XGBoost

Results



Logistic Regression:
Achieved the highest recall

SVM: Underperformed
compared to XGBoost.

XGBoost: Delivered the best
overall performance across
metrics.

Interpretability Matters:
Prioritized over purely
achieving the best metrics.

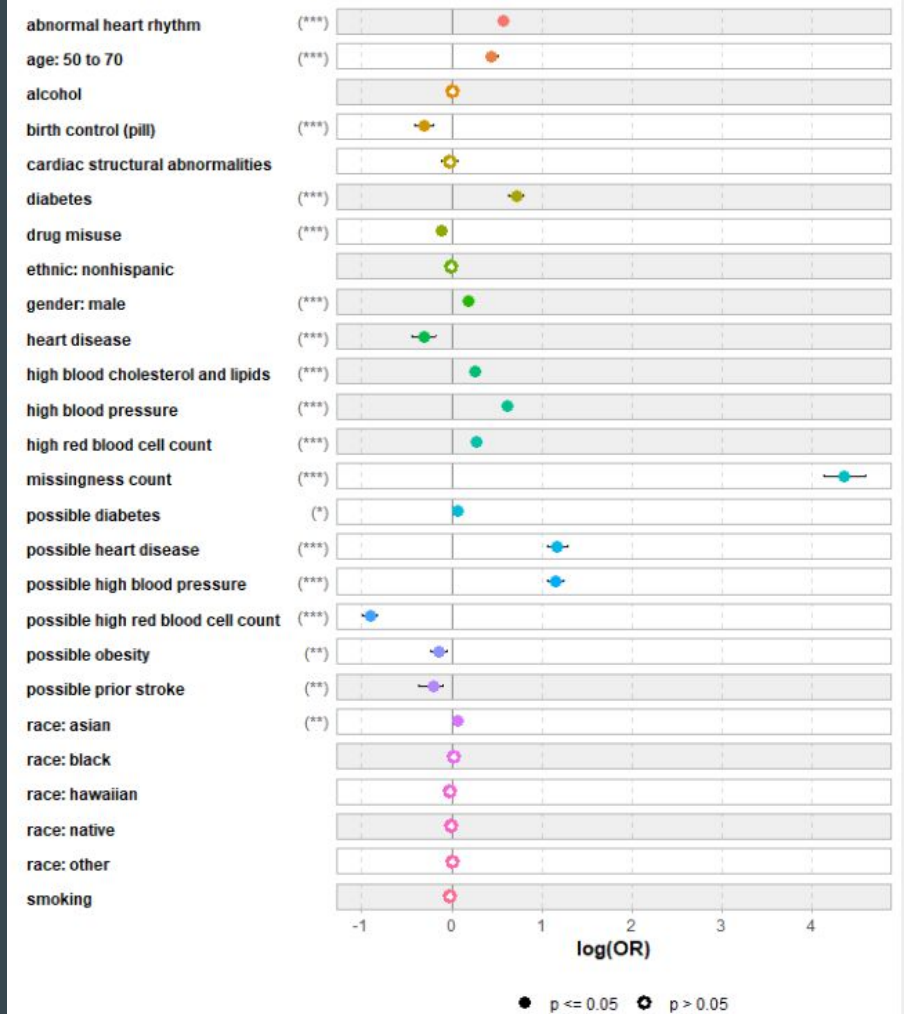
Key Findings

Top Predictors:

- Missingness Count
- High blood pressure
- Heart disease
- Diabetes
- Abnormal heart rhythm

Challenges:

- Addressing missing data
- Balancing interpretability and accuracy in medical contexts



Ethical Implications

Importance of transparency in predictive models

Ensuring equitable predictions across demographics

Mitigating potential biases in synthetic data

Further validation required with real-world EHR datasets

Conclusion

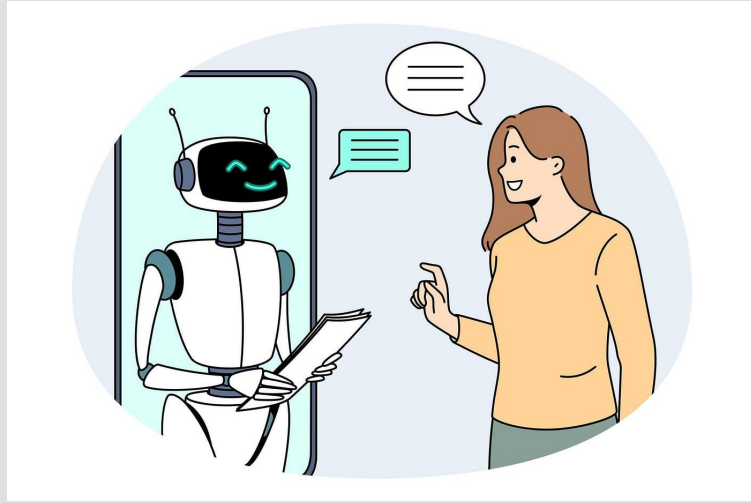
Machine learning can predict stroke.

But, Data Analysts need to work with Field Experts on data design.

Future work:

Test dataset without top risk variables.

Explore additional features like lifestyle and socioeconomic factors.



Thank You.