STROKE PREDICTION

ALEX MILLER, JESICCA GAO, AYESHA SAEED
“THE REGRESSION ROCKSTARS”

OVERVIEW

DATASET

BRFSS 2015 Dataset of health indicators

VARIABLES

- HIGHBP
- HIGHCHOL
- CHOLCHECK
- BMI
- SMOKER
- STROKE
- HEARTDISEASEORATTACK
- PHYSACTIVITY
- FRUITS
- VEGGIES
- HVYALCOHOLCONSUMPTION
- ANYHEALTHCARE
- NODOCBCCOST
- GENHLTH
- MENTHLTH
- PHYSHLTH
- DIFFWALK
- SEX
- AGE
- EDUCATION
- INCOME

OUTCOME - STROKE

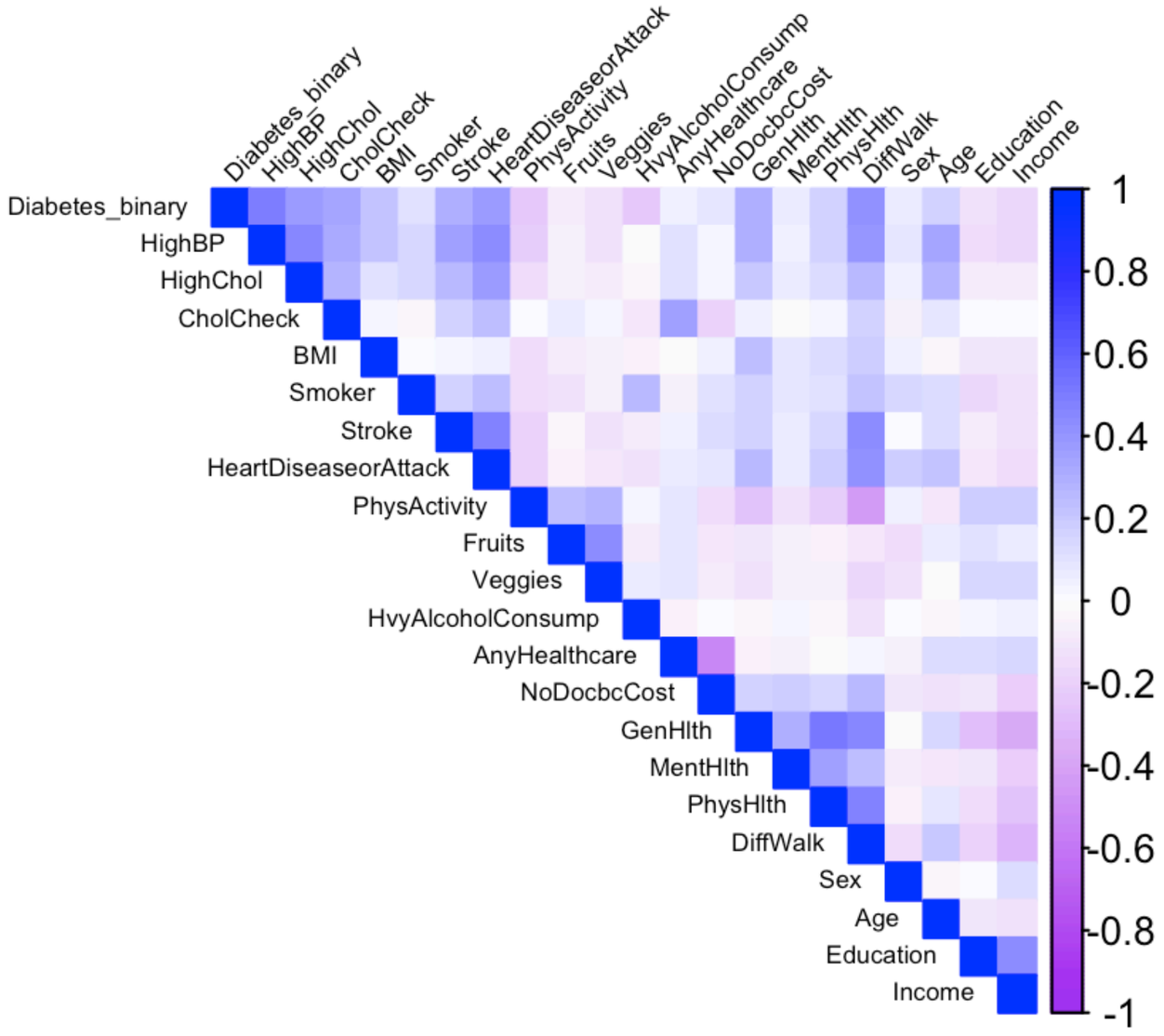
- Strokes are the fourth leading cause of death in the United States, and the pathogenesis of this disease is complex and multi-faceted.
- There are numerous medical and environmental risk factors that can increase the likelihood of an individual developing a stroke, including high cholesterol, lack of physical activity, and poor diet.
- These risk factors are among others from the dataset our team will use to build models for prediction of strokes given those factors.
- Stroke prediction using existing data can help individuals and healthcare providers better understand the pathogenesis of the disease.

TARGETS

Which modeling approach provides the highest accuracy in predicting stroke occurrence?

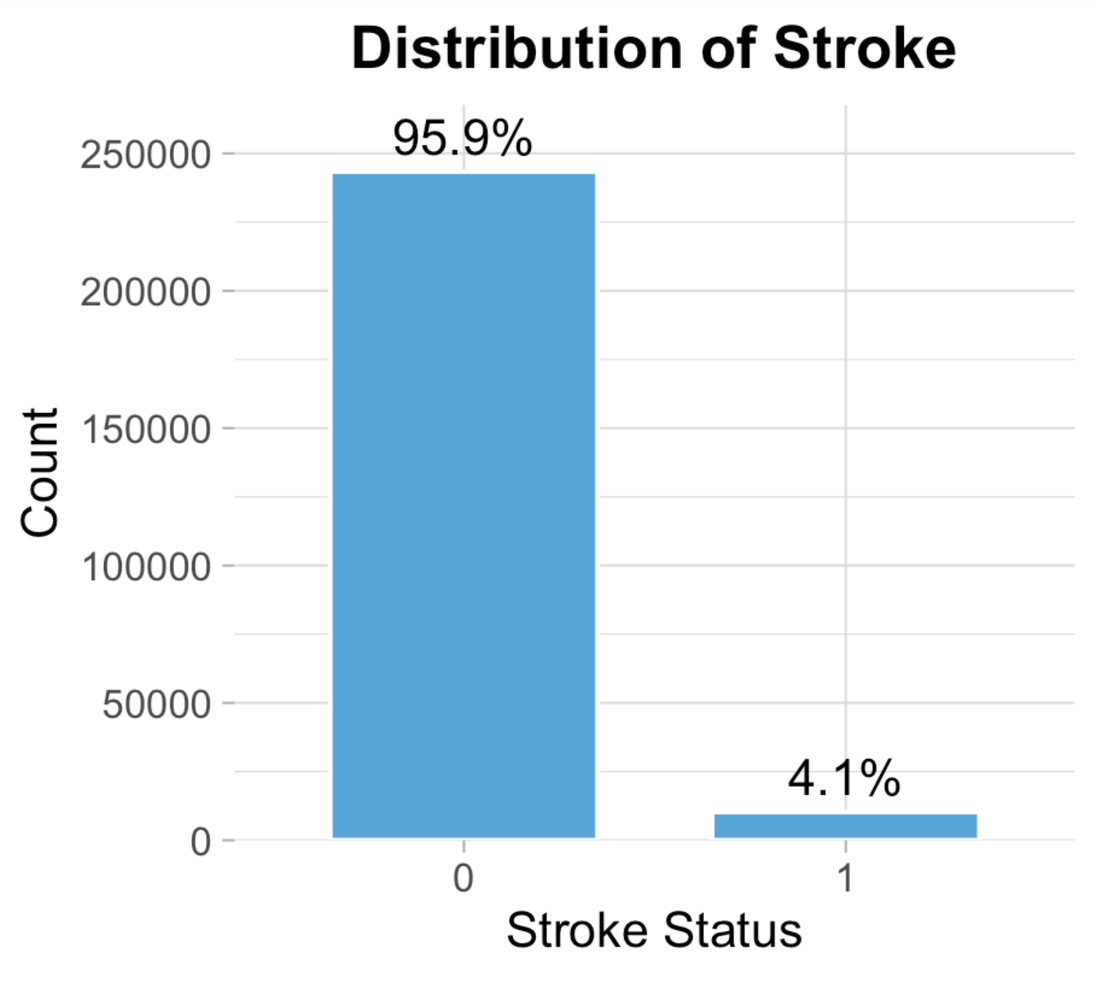
Based on the model results, which predictors are most influential in determining stroke risk?

EXPLORATORY DATA ANALYSIS



From the correlation heatmap, we can see that the outcome variable, Stroke, has the strongest correlations with the predictors **DiffWalk** and **HeartDiseaseorAttack**.

These two predictors could be significant in multiple of our models.



Our outcome variable, Stroke, has very few cases (4%). Because our data is unbalanced, we could encounter some problems downstream when training and testing our models.

LOGISTIC REGRESSION

Using Balanced Data
Training & Testing Data

Steps

1. Create initial multivariable logistic model using training data
2. Use backwards elimination to select predictors significant at the 0.05 level
3. Test the model using the test data
4. Determine the model's predictive accuracy, sensitivity, and specificity

Original model had high specificity but extremely low sensitivity

→ We created a new dataset consisting of all cases and twice as many controls as the total number of cases

Backward Elimination to Select Final Model

Initial model

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -5.252924   0.419831 -12.512  < 2e-16 ***
Diabetes_binary          0.201416   0.087635   2.298 0.021542 *
HighBP                   0.585230   0.089262   6.556 5.51e-11 ***
HighChol                 0.166242   0.079908   2.080 0.037488 *
CholCheck               -0.323765   0.223516  -1.449 0.147474
BMI                     -0.021817   0.006155  -3.545 0.000393 ***
Smoker                   0.153407   0.076220   2.013 0.044149 *
HeartDiseaseorAttack     0.989491   0.084640  11.691  < 2e-16 ***
PhysActivity            -0.132260   0.082477  -1.604 0.108803
Fruits                   0.007739   0.079087   0.098 0.922051
Veggies                 -0.163922   0.090129  -1.819 0.068951 .
HvyAlcoholConsump       -0.259600   0.202386  -1.283 0.199599
AnyHealthcare            0.409863   0.204205   2.007 0.044737 *
NoDocbcCost              0.311272   0.114065   2.729 0.006355 **
GenHlth                  0.275013   0.044876   6.128 8.89e-10 ***
MentHlth                 0.009638   0.004189   2.301 0.021404 *
PhysHlth                 0.006258   0.004072   1.537 0.124358
DiffWalk                 0.512923   0.092985   5.516 3.46e-08 ***
Sex                     -0.029917   0.078326  -0.382 0.702490
Age                      0.143558   0.016687   8.603  < 2e-16 ***
Education                0.051819   0.038495   1.346 0.178262
Income                  -0.092536   0.019671  -4.704 2.55e-06 ***
```

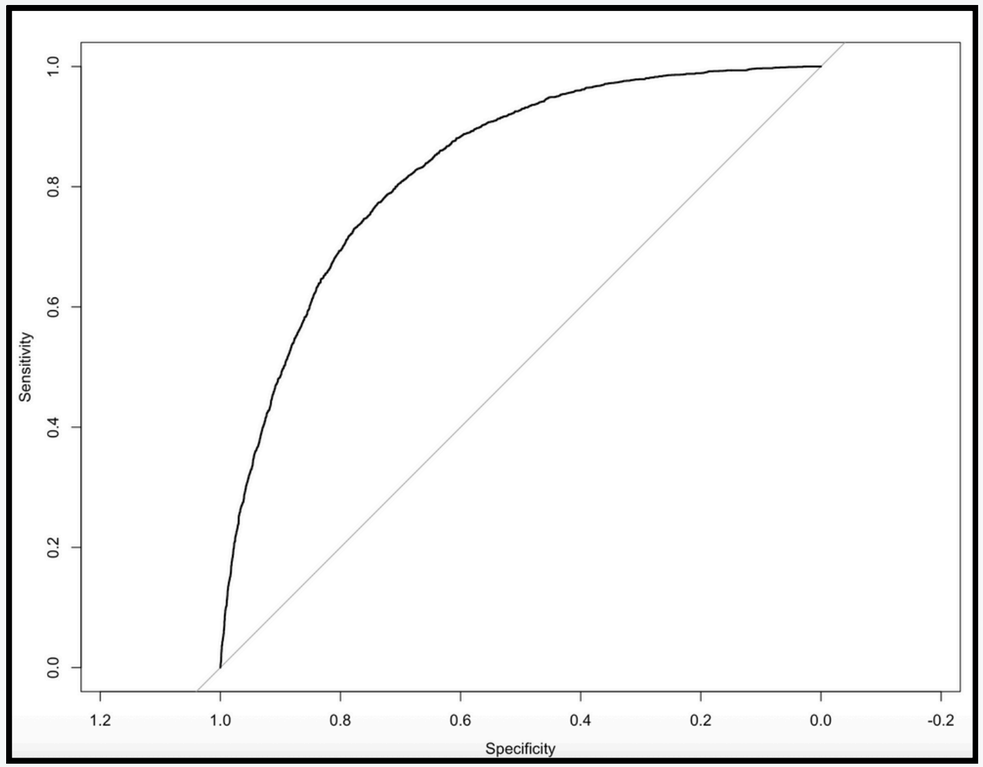
Final mode (0.05 significance)

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -5.327611   0.296555 -17.965  < 2e-16 ***
Diabetes_binary          0.199239   0.087164   2.286 0.022266 *
HighBP                   0.564140   0.088849   6.349 2.16e-10 ***
HighChol                 0.166794   0.079637   2.094 0.036221 *
BMI                     -0.020975   0.006135  -3.419 0.000628 ***
Smoker                   0.147574   0.074722   1.975 0.048271 *
HeartDiseaseorAttack     0.992059   0.083647  11.860  < 2e-16 ***
NoDocbcCost              0.286227   0.111685   2.563 0.010383 *
GenHlth                  0.306792   0.040512   7.573 3.65e-14 ***
MentHlth                 0.011601   0.004065   2.854 0.004324 **
DiffWalk                 0.584489   0.088028   6.640 3.14e-11 ***
Age                      0.148885   0.016381   9.089  < 2e-16 ***
Income                  -0.086986   0.017808  -4.885 1.04e-06 ***
```

*checked VIF & tolerance values and found no multicollinearity issues

Evaluating Model (using test data)

ROC Curve



AUC = 0.8307

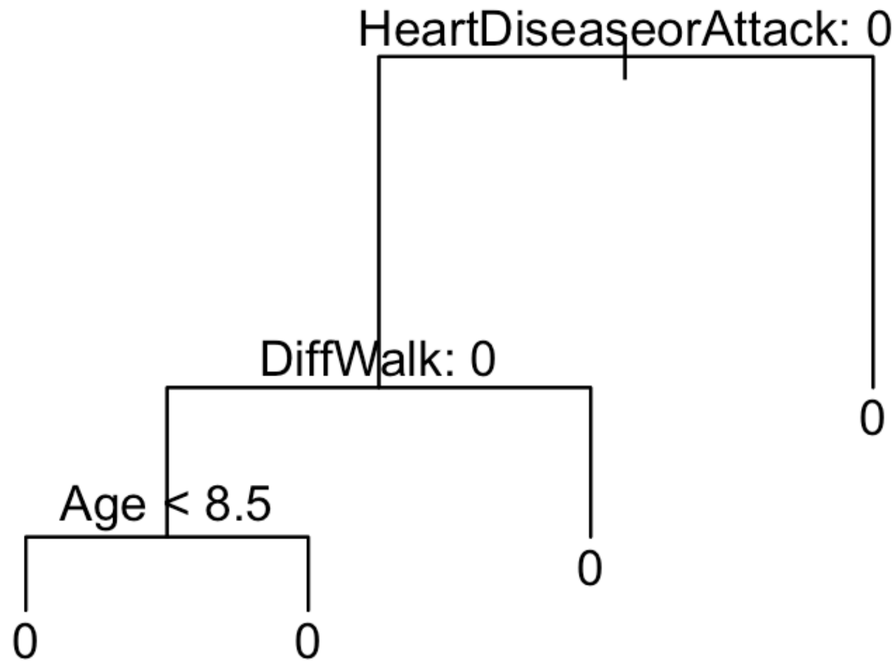
Sensitivity: 0.557

Specificity: 0.872

Accuracy: 0.765

CLASSIFICATION TREE

CLASSIFICATION TREE - UNBALANCED



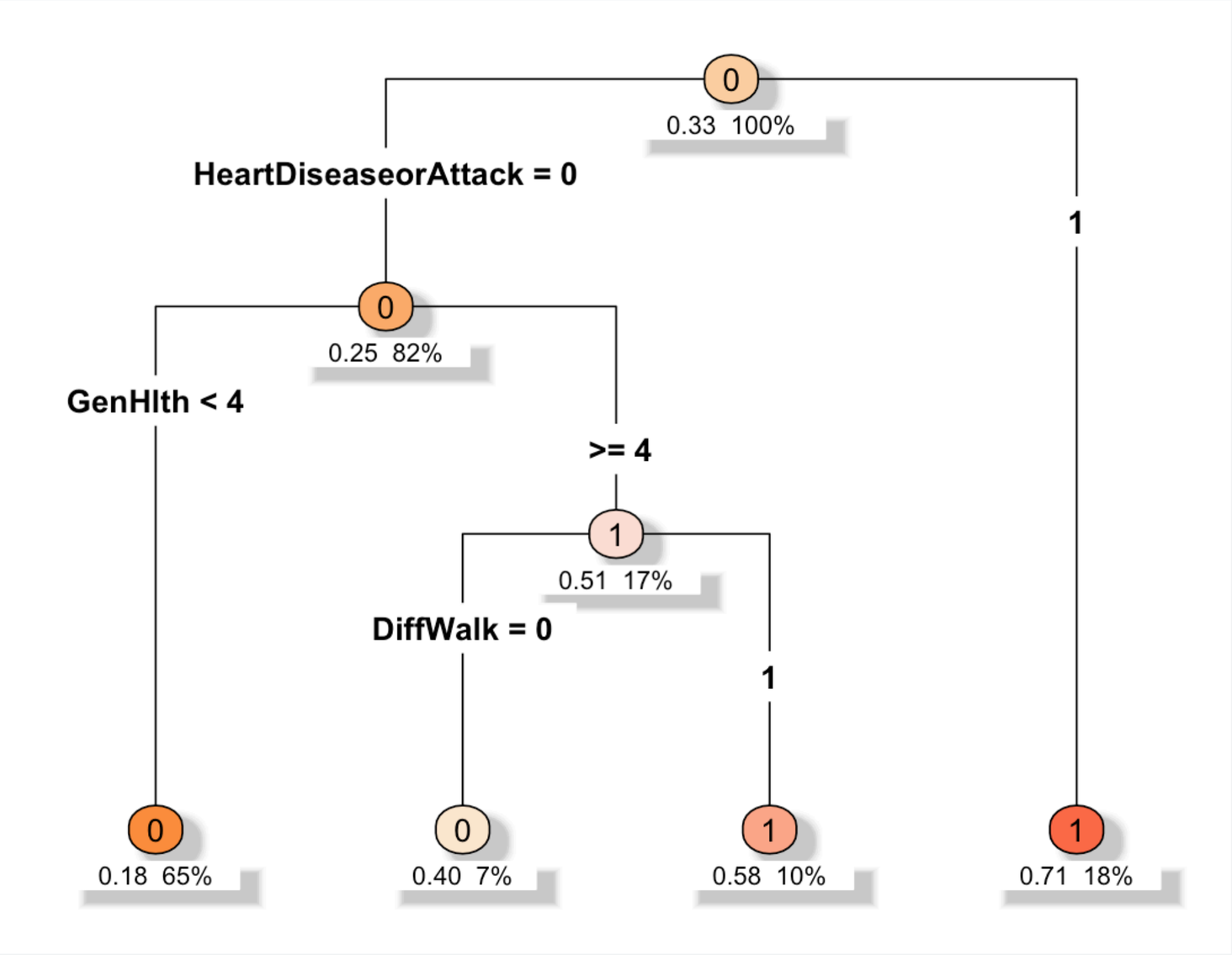
Unbalanced dataset resulted in the following tree

- All outcomes predicted stroke = 0
- First tree was also the best tree using cross-validation

Solution?

- Balance dataset

CLASSIFICATION TREE - BALANCED



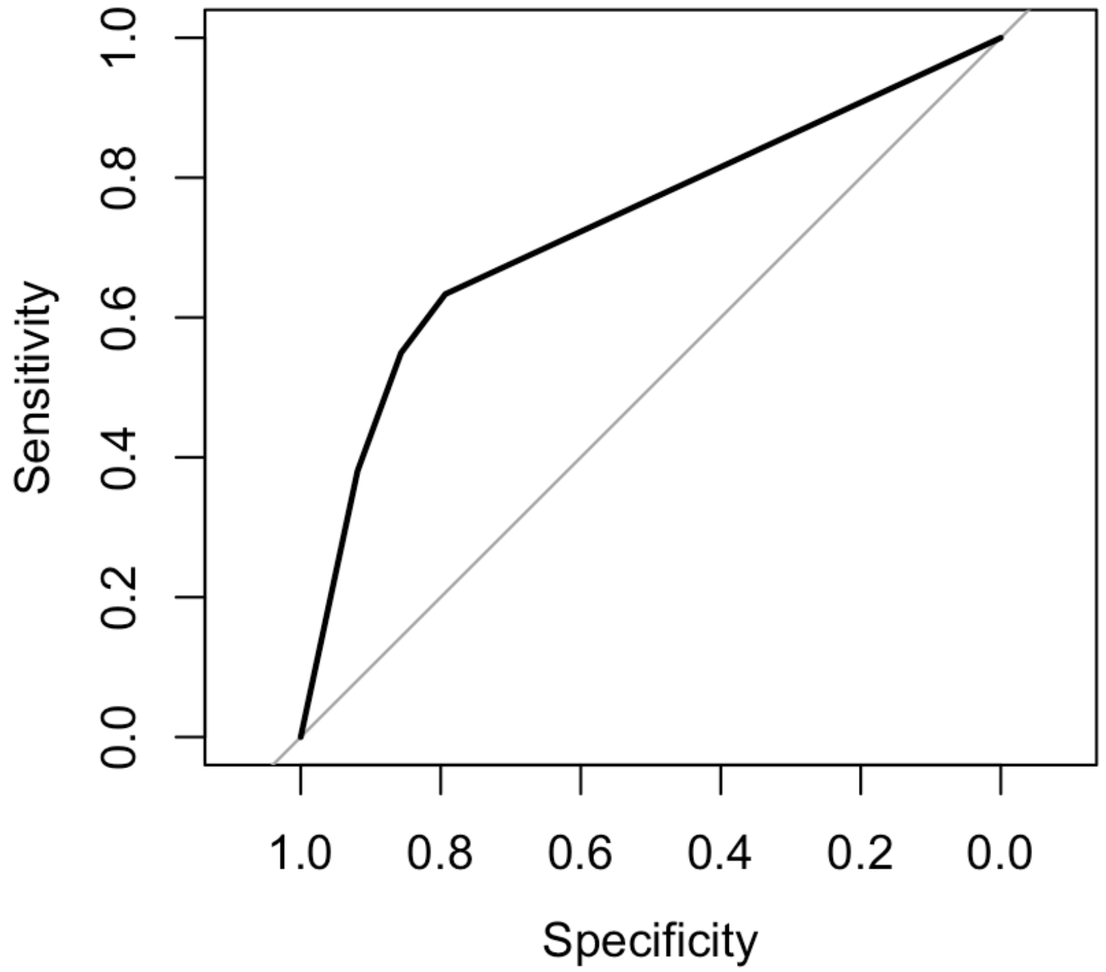
We balanced the dataset to have a case:control ratio of 1:2.

Our tree now predicts stroke outcomes as both 0 and 1 instead of just 0.

Predictors it uses:

- HeartDiseaseorAttack
- GenHlth
- Diffwalk

CLASSIFICATION TREE - ROC CURVE

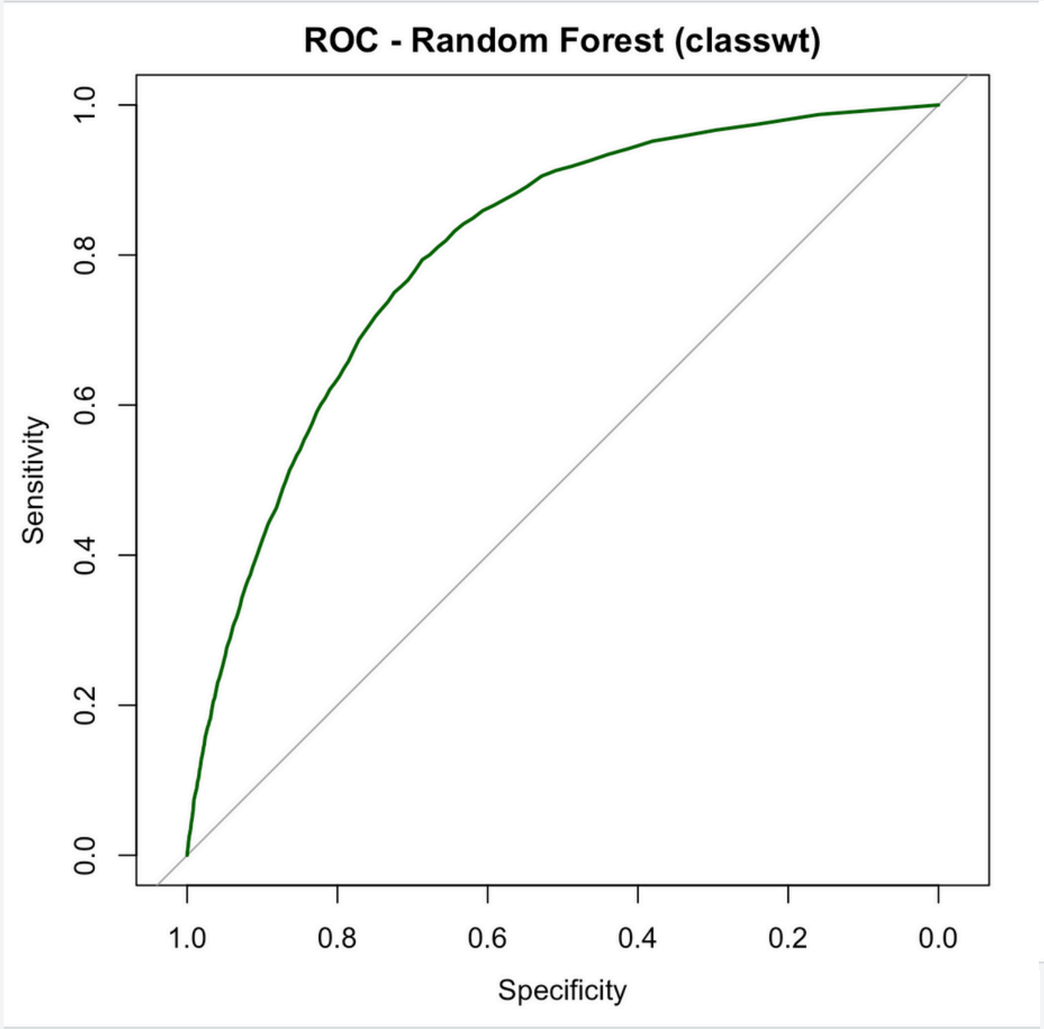


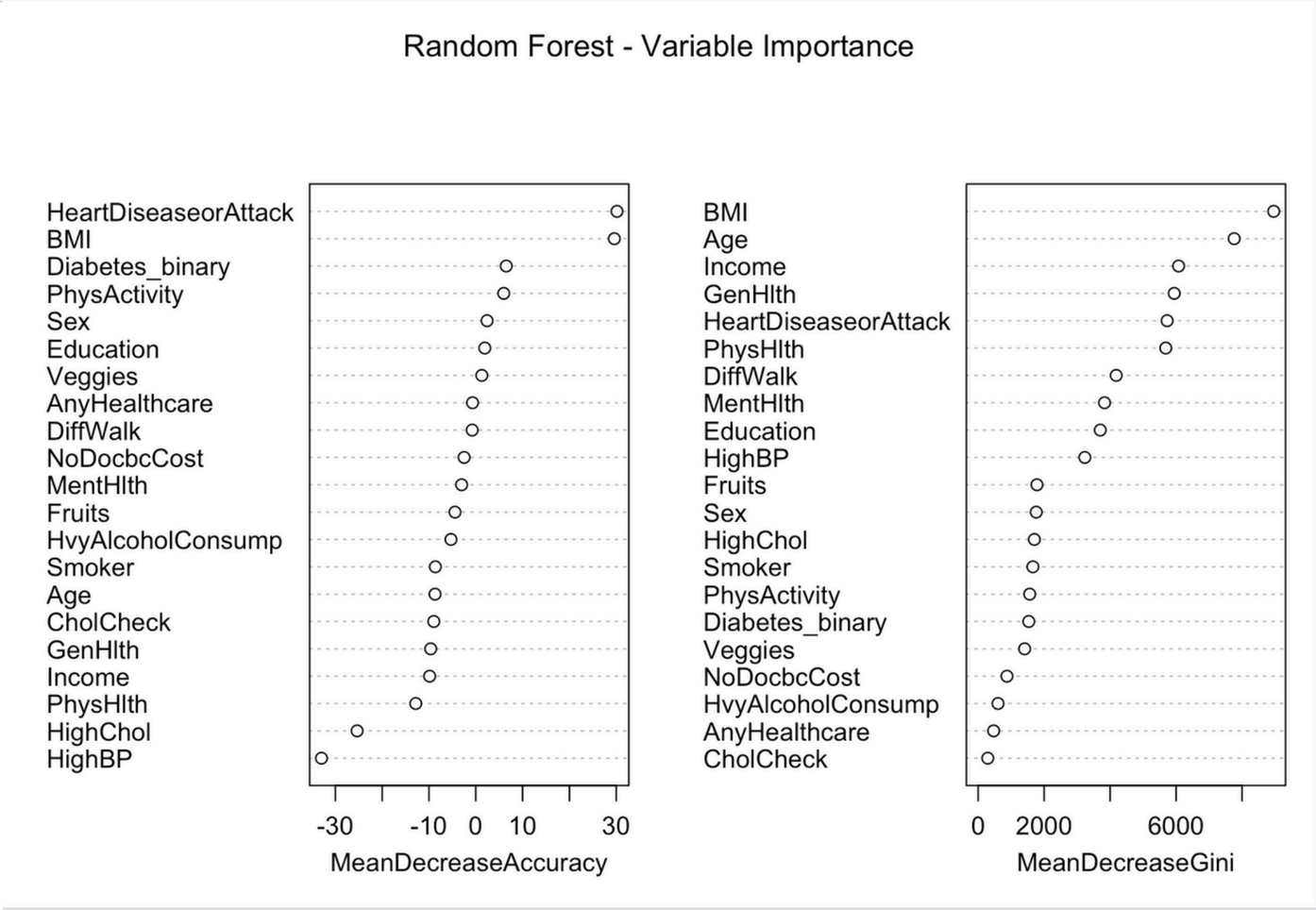
AUC: 0.7298
Accuracy : 0.8444
Sensitivity : 0.54930
Specificity : 0.85674

Note: our train/test split was done before balancing the dataset, and then training test was balanced

RANDOM FOREST

FEATURE IMPROTANCE





ACCURACY : 0.9583

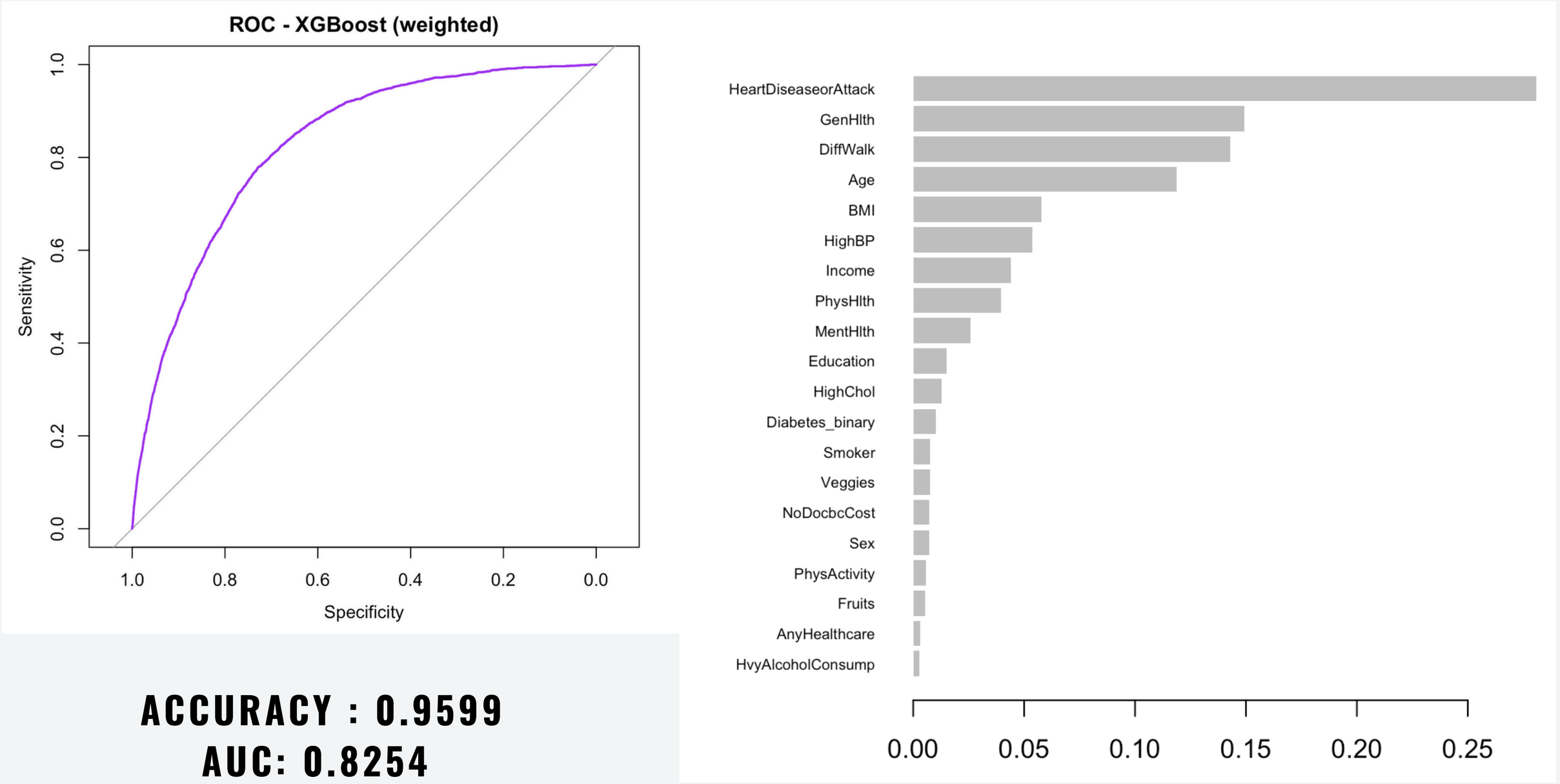
AUC: 0.8055

Mean Decrease Accuracy
Measures the drop in model accuracy when a feature is removed (higher = more important).

Mean Decrease Gini
Measures how much a feature improves node purity in decision trees (higher = stronger splitter).

XGBOOST

FEATURE IMPROTANCE



ACCURACY : 0.9599

AUC: 0.8254

TOP 10 MOST IMPORTANT FEATURES

Feature	Random Forest (RF)	XGBoost
HeartDiseaseorAttack	Most important	Most important
GenHlth	Medium-low (RF)	High importance
DiffWalk	Top 3	Top 3
Age	Important	Important
BMI	Important	Important
HighBP	Medium (Gini importance)	Clearly important
Income / Education	Medium importance	Strong importance
MentHlth	Medium-low	Moderate importance
Lifestyle (Fruits, Activity)	Low in both	Low in both

RESULTS

Model	Accuracy	AUC
Multivariable Logistic Regression	0.7650	0.8307
Classification Tree	0.8444	0.7298
Random Forest	0.9583	0.8055
XGBoost	0.9599	0.8254

KEY FINDINGS

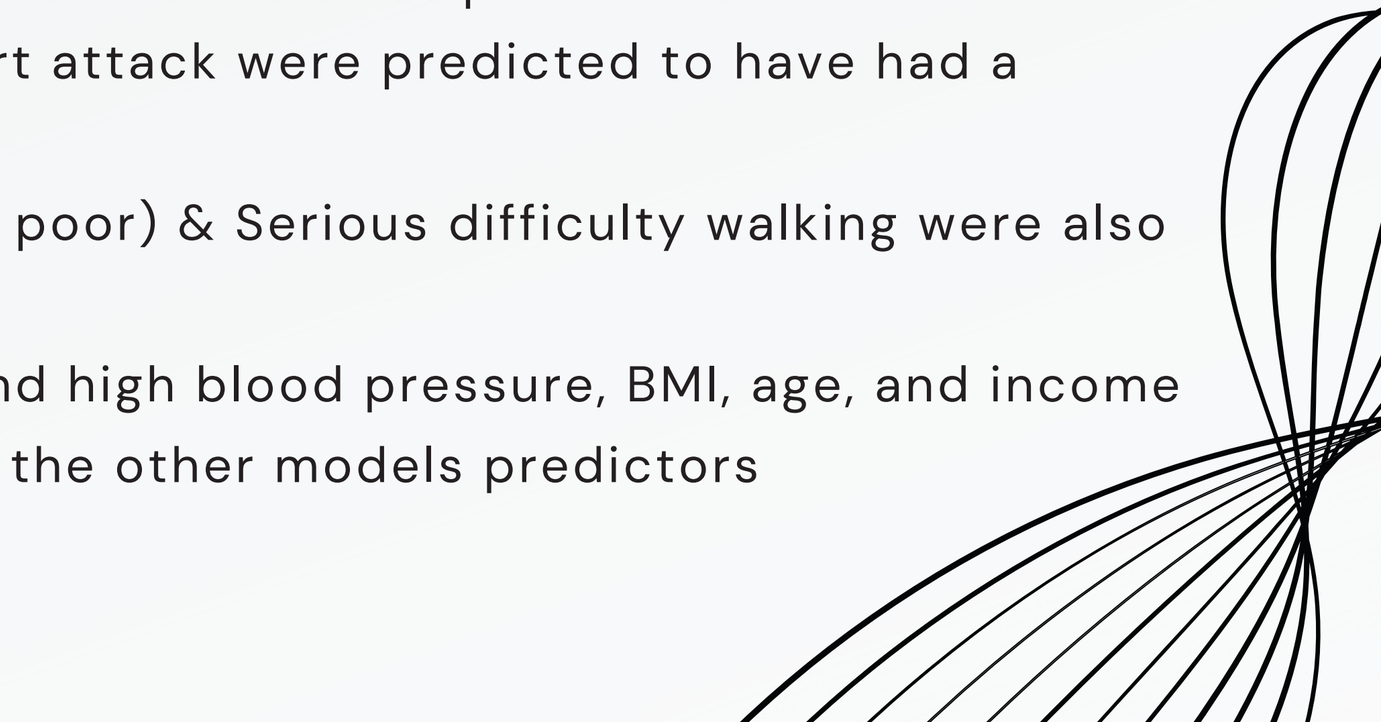
1.**XGBoost** outperformed all models in AUC and accuracy

2.**Feature Importance ---->Consistent key risk factors:**
HeartDiseaseorAttack history
Self-rated general health (GenHlth)
Difficulty walking (DiffWalk)
Body Mass Index (BMI)
Physical activity and smoking behavior

3.Even **simple models** (e.g., Logistic Regression) revealed clear directional relationships for stroke risk factors.

WHAT HAVE WE LEARNED?

- We faced difficulties developing statistical models to predict such a **rare outcome** ~4% of our sample had a stroke
  - Initial classification trees had all 0s at terminal nodes
  - Logistic regression **predicted 0** for nearly every outcome, as this yielded the **highest accuracy -----> Resample for a balanced dataset**
- **Classification tree** highlighted **only a few** variables as predictors:
  - All who had heart disease or a heart attack were predicted to have had a stroke
  - General health (1 = excellent to 5 = poor) & Serious difficulty walking were also used in the classification tree
- **Multivariable logistic regression** found high blood pressure, BMI, age, and income significant in the model in addition to the other models predictors



IMPLICATIONS

1.Policy Insight: Promoting access to **routine check-ups and cholesterol screening (CholCheck)** may reduce stroke risks.

2.Behavioral Campaigns: Encourage **physical activity and stroke symptom awareness**, especially in high-risk groups.

3.Targeted Interventions: Predictive modeling helps flag vulnerable individuals using low-cost, scalable methods.

BIG PICTURE

1. **Preventable risks matter**: Lack of physical activity, obesity, and diabetes all play major roles.

2. **Survey data is powerful**: No lab tests or imaging—just behavior & health self-reporting.

3. **Health equity insight**: Variables like income, education, and cost barriers to care hint at systemic disparities.

GROUP MEMBER CONTRIBUTIONS

- Regular project meetings - All group members
- Background research - All group members
- Logistic regression - Alex
- Classification tree - Ayesha
- Random Forest & Xgboost - Jessica
- Write-up - All group members (methods & results for the section we researched)

THANK YOU

We welcome any questions and/or comments!