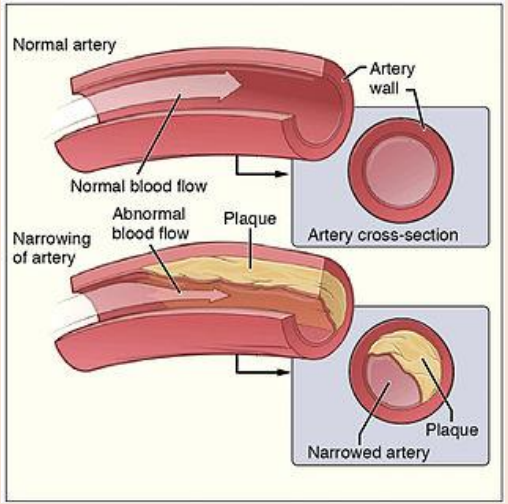
Risk of Heart Disease/Attack Classification

Hannah Kim
Project Classification

Heart Disease/ Attack Cause and Symptoms:

- Plaque build up causes arteries in blood vessel to shrink
  - Prevents blood flow
  - Causes further damage -blood clots

- Abnormal heartbeat
- Fatigue
- Pain in chest, arm, back
- Nausea
- Anxiety

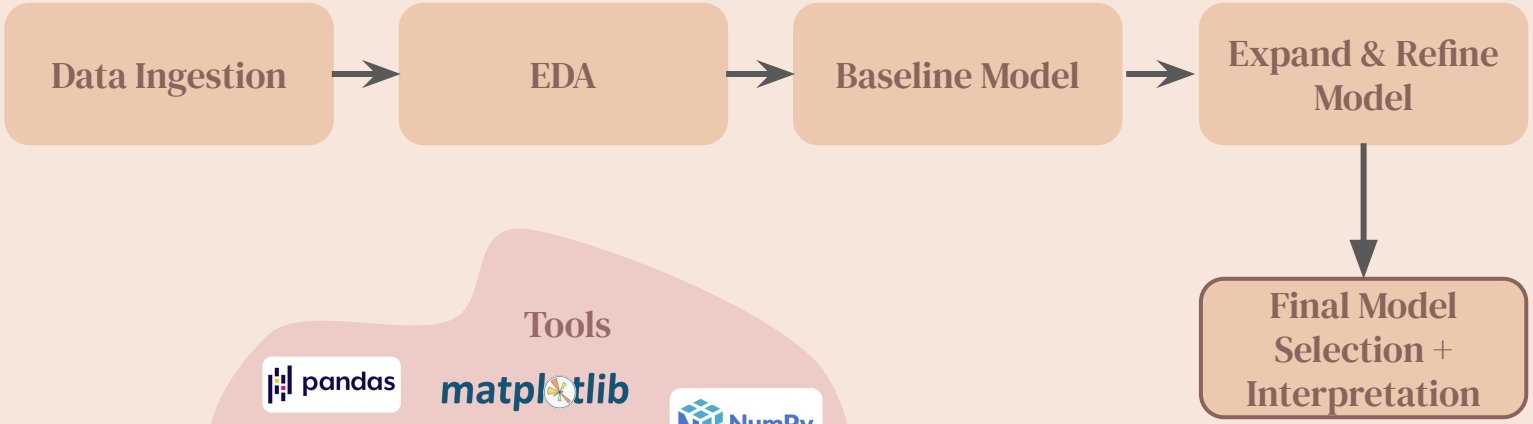


Heart Disease/ Attack in US

- One of the top leading causes of death in both women and men
- “1 in every 5 deaths in 2020” (CDC)
- “1 in 5 heart attack are silent” (CDC)
- Risk of heart disease/ attack can be prevented – lifestyle choices
- Notable risk factors
  - high blood pressure, high cholesterol, smoking (CDC)

- Goal: To predict individual’s risk to get heart disease/attack based on selected health indicators

Pipeline Overview

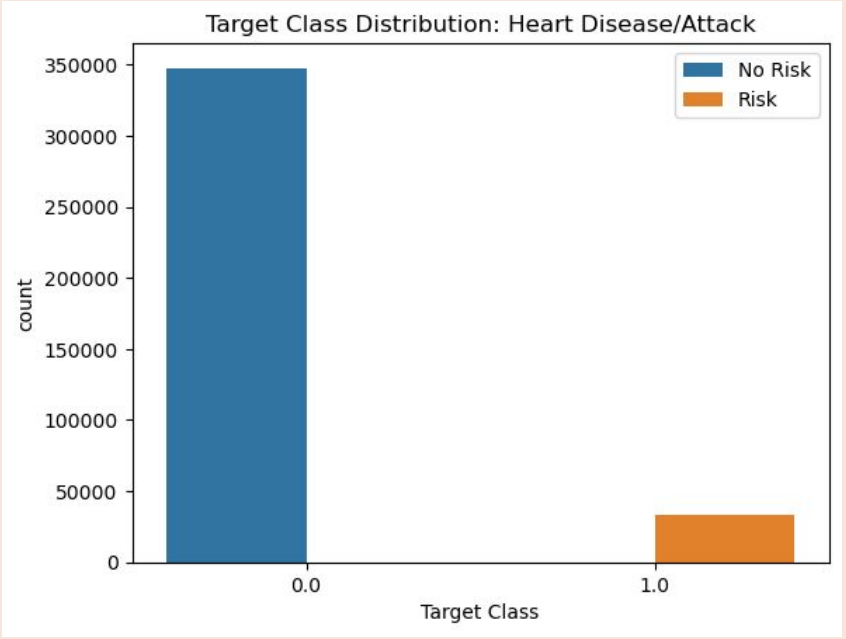


Data

- CDC's 2021 Annual Behavioral Risk Factor Surveillance System (telephone interview)
- Over 400k rows of data
  - Data cleaning: 381,147 interview responses w/ 15 features
  - Features: sex, age, bmi, overwt, diabetes, stroke, h_bp, h_choles, smoker, alcohol consumption, physcial exercise, fruit intake, vege intake [14]
  - Target: History of heart disease or attack

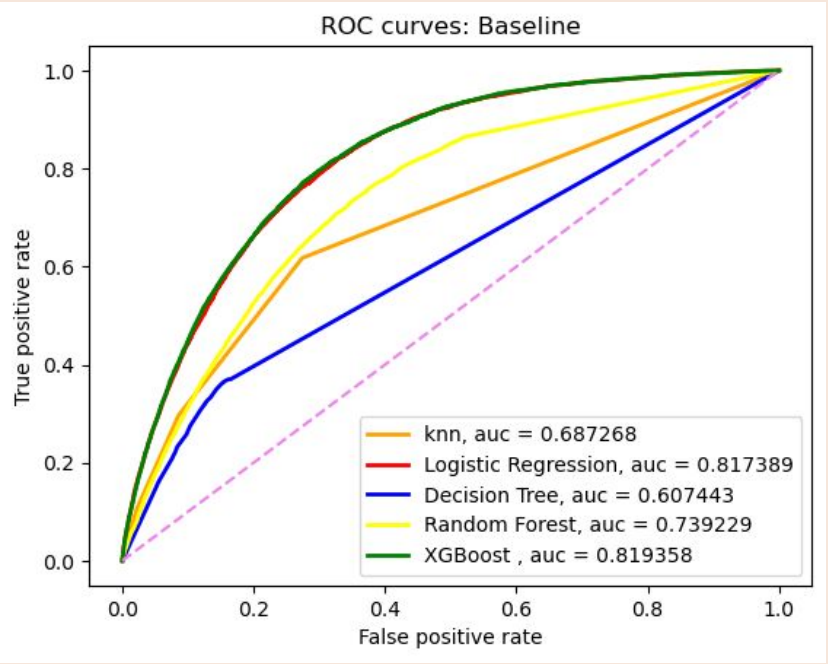
Data

- About 10% consists positive class
  - Risk of heart disease/attack
- Will need to handle class imbalance
  - Class weight balanced
  - Oversampling
  - SMOTE



Baseline Model

- Top three models based on ROC performance to further tune hyperparameters
  - Logistic Regression
  - Random Forest
  - XGBoost



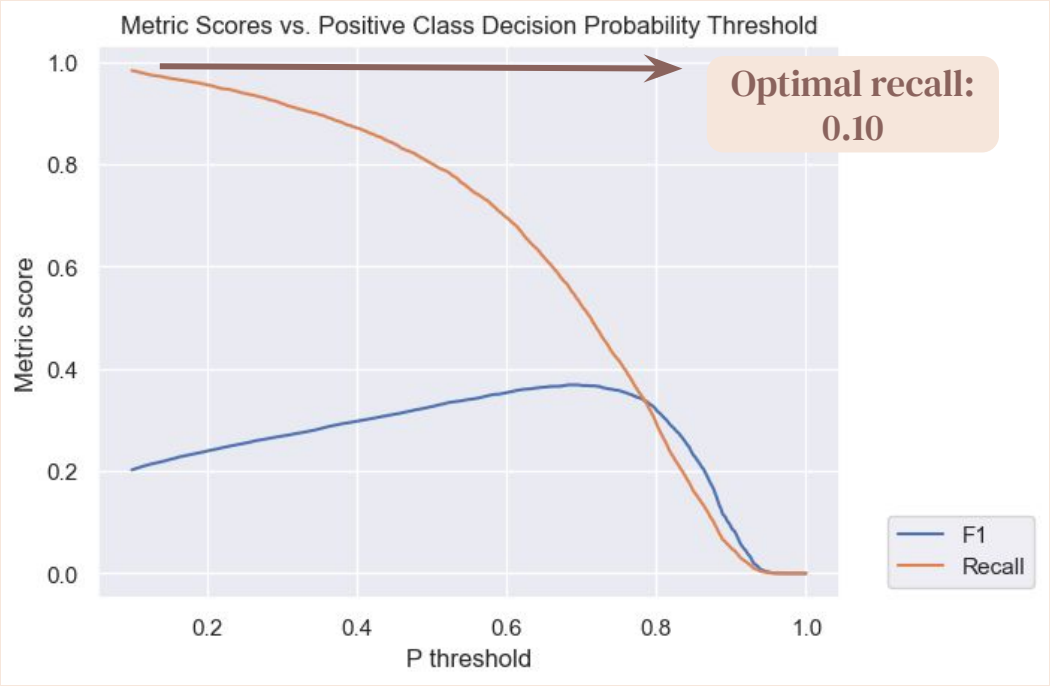
Model Comparison

	accuracy	precision	recall	f1	auc
Logistic Regression: Class Weight	0.711426	0.204148	0.786424	0.324151	0.817612
Random Forest: Class Weight	0.754384	0.221532	0.712483	0.337977	0.810336
XGBoost: Class Weight	0.708269	0.204545	0.801431	0.325910	0.822102
Logistic Regression: Oversampling	0.720408	0.207519	0.772411	0.327146	0.817695
Random Forest: Oversampling	0.777192	0.225056	0.627013	0.331225	0.794562
XGBoost: Oversampling	0.771945	0.213660	0.593818	0.314250	0.772993
Logistic Regression: Oversampling + Class Weig...	0.713901	0.205021	0.782349	0.324900	0.817697
Random Forest: Oversampling + Class Weight Bal...	0.772583	0.222110	0.633174	0.328860	0.795170
XGBoost: Oversampling + Class Weight Balanced	0.654747	0.163925	0.712980	0.266563	0.743006
Logistic Regression: SMOTE	0.721300	0.208242	0.773405	0.328133	0.817285
Random Forest: SMOTE	0.843526	0.270381	0.458159	0.340071	0.806094
XGBoost: SMOTE	0.889291	0.287861	0.175114	0.217759	0.790016
Logistic Regression: SMOTE + Class Weight Bala...	0.713166	0.204773	0.783641	0.324699	0.817282
Random Forest: SMOTE + Class Weight Balanced	0.841926	0.268263	0.460942	0.339147	0.806468
XGBoost: SMOTE + Class Weight Balanced	0.610311	0.169373	0.878155	0.283974	0.808637

- Interested in recall & ROC value
  - Secondary metric : F1 score
- Handle class imbalance
  - Adjust class weight
  - Oversampling
  - SMOTE
  - Oversampling + Class weight
  - SMOTE + Class Weight

Final Model : Adjusting Threshold

- XGBoost
  - Class imbalance technique used :
    - Class weight balanced
  - Threshold found w/ best recall score : 0.10



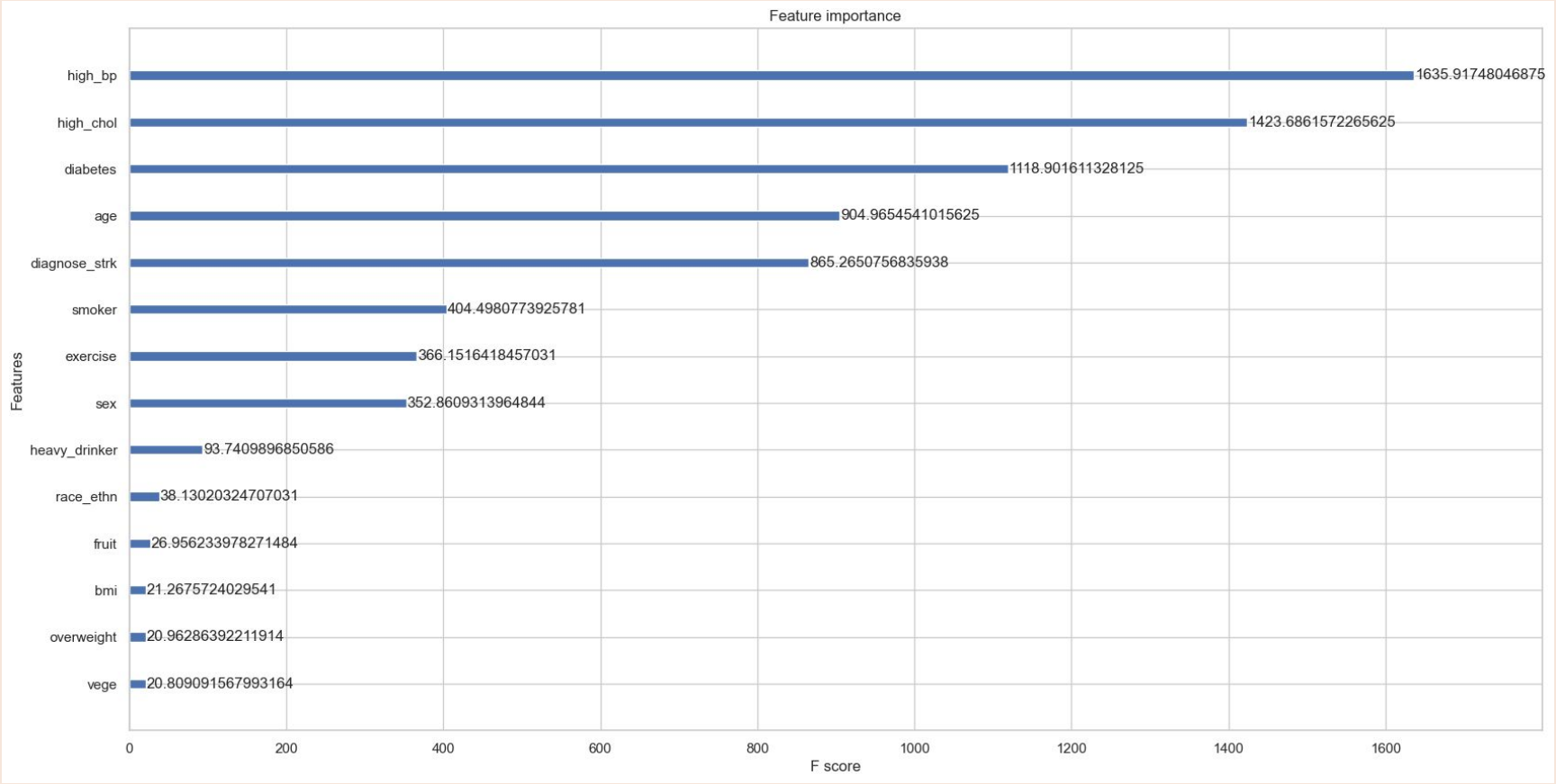
Final Model : XGBoost Comparing Thresholds

```
XGBoost (Test) Class Weight: Accuracy: 0.708269
XGBoost (Test) Class Weight: Precision: 0.204545
XGBoost (Test) Class Weight: Recall : 0.801431
XGBoost (Test) Class Weight: F1 : 0.325910
XGBoost (Test) Class Weight: Roc : 0.822102
```

```
XGBoost (Test) Class Weight + Adjusted Threshold: Accuracy: 0.317819
XGBoost (Test) Class Weight + Adjusted Threshold: Precision: 0.112787
XGBoost (Test) Class Weight + Adjusted Threshold: Recall : 0.983403
XGBoost (Test) Class Weight + Adjusted Threshold: F1 : 0.202364
XGBoost (Test) Class Weight + Adjusted Threshold: Roc : 0.618501
```

F1 Score & ROC value decreases w/ adjusted threshold

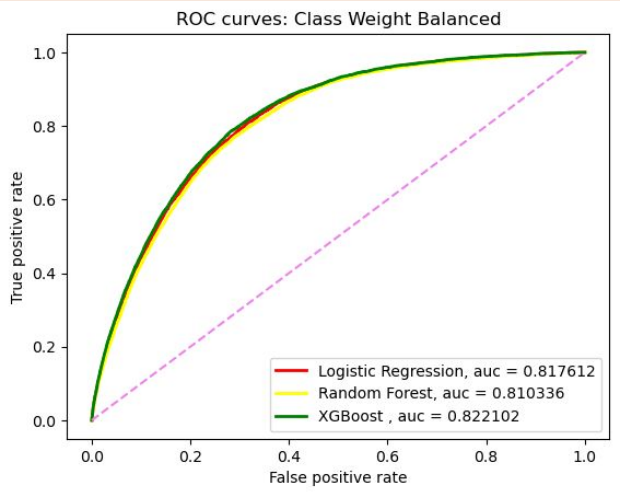
Feature Importance



Future Work

- Include more features (health indicators) related to heart disease/attack
- Do more feature engineering based on topic knowledge
- Deploy web app to display ML model

Appendix



Heart Disease/Attack Predictor Form

Sex:
Male

Age
18.0

Race/Ethnicity:
White, Non-Hispanic

BMI
12.0

Is your BMI over 25?
No

Have you ever been told you had Diabetes?
No

Have you ever been told you had a Stroke?
No

Have you smoked at lease 100 cigarettes in your life?
No

Have you been told you have high blood pressure?
No

Have you been told you have high cholesterol?
No

M: Had more than 14 drinks per week? /F:Had more than 7 drinks per week?
No

Have you exercised, other than your regular job during this past month?
No

Do you consume more than one fruit a day?
No

Do you consume more than one vegetable a day?
No

Predict