Capstone Project: Heart Attack

Hio Wa Mak

11/02/2024

Problem Identification

- Heart disease remains the leading cause of death in the United States (CDC, 2021)
- Someone experiencing a heart attack every 40 seconds
- In the US, approximately 805,000 people have a heart attack each year
- Heart disease costs a huge burden of $239.9 billion annually

The Big Question

How can the Centers for Disease Control and Prevention (CDC) develop an algorithm to accurately classify whether individuals have had heart attack(s) with over 80% sensitivity by year-end, using data on demographic characteristics, medical histories, and a range of health and lifestyle factors?

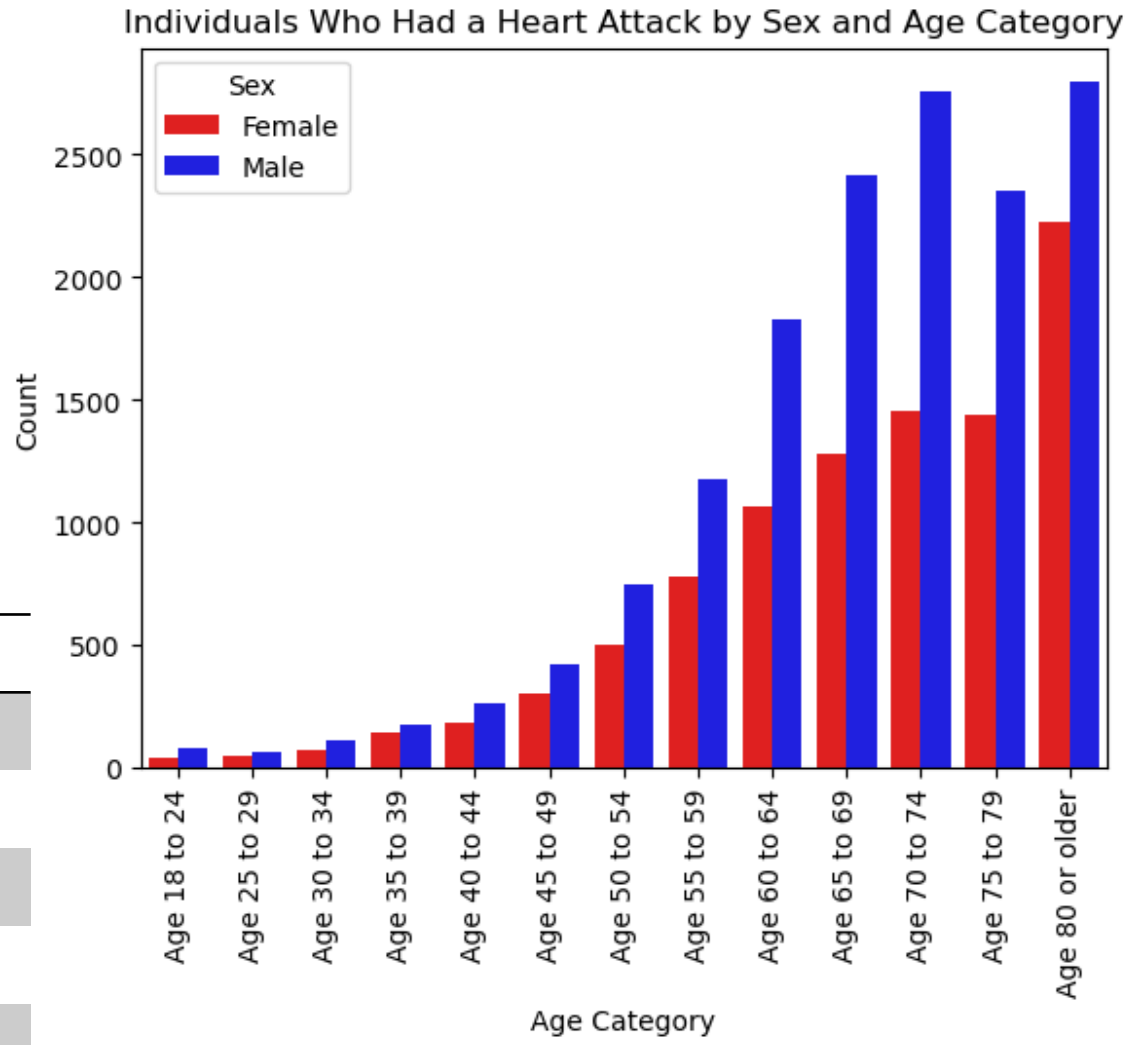
Data

- Data originates from the CDC's 2022 Behavioral Risk Factor Surveillance System (BRFSS)
- It was sourced from Kaggle
- The dataset was refined to 40 variables deemed relevant to heart attack risk, with data processing and selection documented by Kamil Pytlak on a GitHub repository
- The BRFSS conducts annual telephone surveys with over 400,000 U.S. adults, gathering information on their health status and behaviors
- **Goal**: Use supervised machine learning techniques to classify individuals with a history of heart attack

Exploratory Data Analysis

- Age and sex appeared to be important demographic predictors
- Other related variables are

“HadAngina” (r = .44)	**“DifficultyWalking” (r = .17)**
“HadStroke” (r = .19)	**“RemovedTeeth” (r = .17)**
“HadCOPD” (r = .14)	**“ChestScan” (r = .17)**
“HadKidneyDisease” (r =.12)	**“GeneralHealth” (r = -.19)**
“HadArthritis” (r = .12)	**“PhysicalHealthDays” (r .14)**
“HadDiabetes” (r = .15)	**“AgeCategory” (r = .18)**
“DeafOrHardOfHearing” (r = .10)	



Modeling Steps

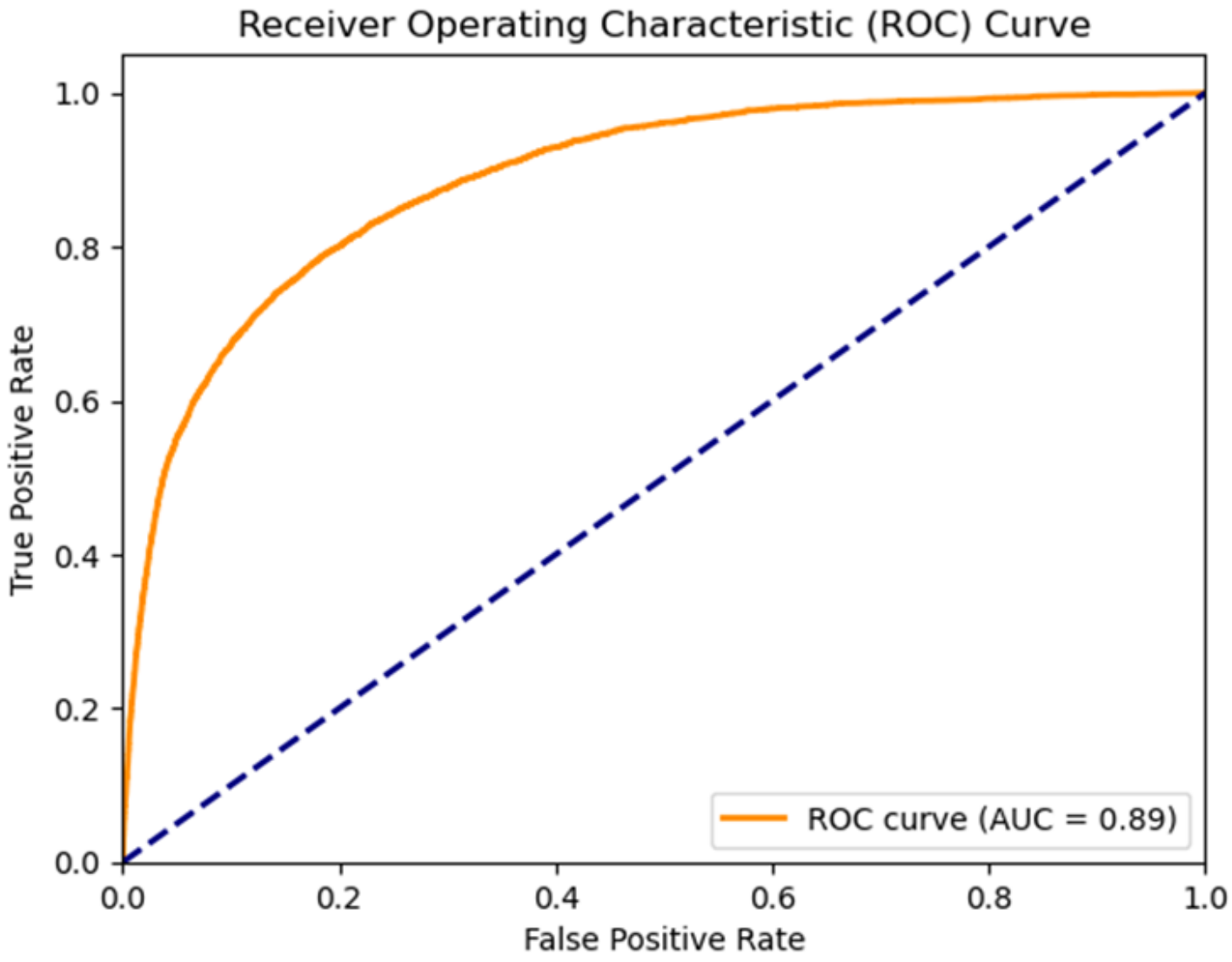
1. Run multiple models (Logistic Regression, Random Forest, XGBoost, and Naive Bayes) and use cross-validation to fine-tune hyperparameters.
2. Evaluate the models using metrics such as accuracy, precision, recall, and F1-score.
3. Select the best model based on recall (sensitivity), prioritizing the minimization of false negatives — the risk of incorrectly identifying a person as not having had a heart attack when they actually have.
4. Analyze the feature importance of the final model.
5. Assess the model's performance on the test set.
6. Refit the best model using the entire dataset.
7. Save the final model for future use

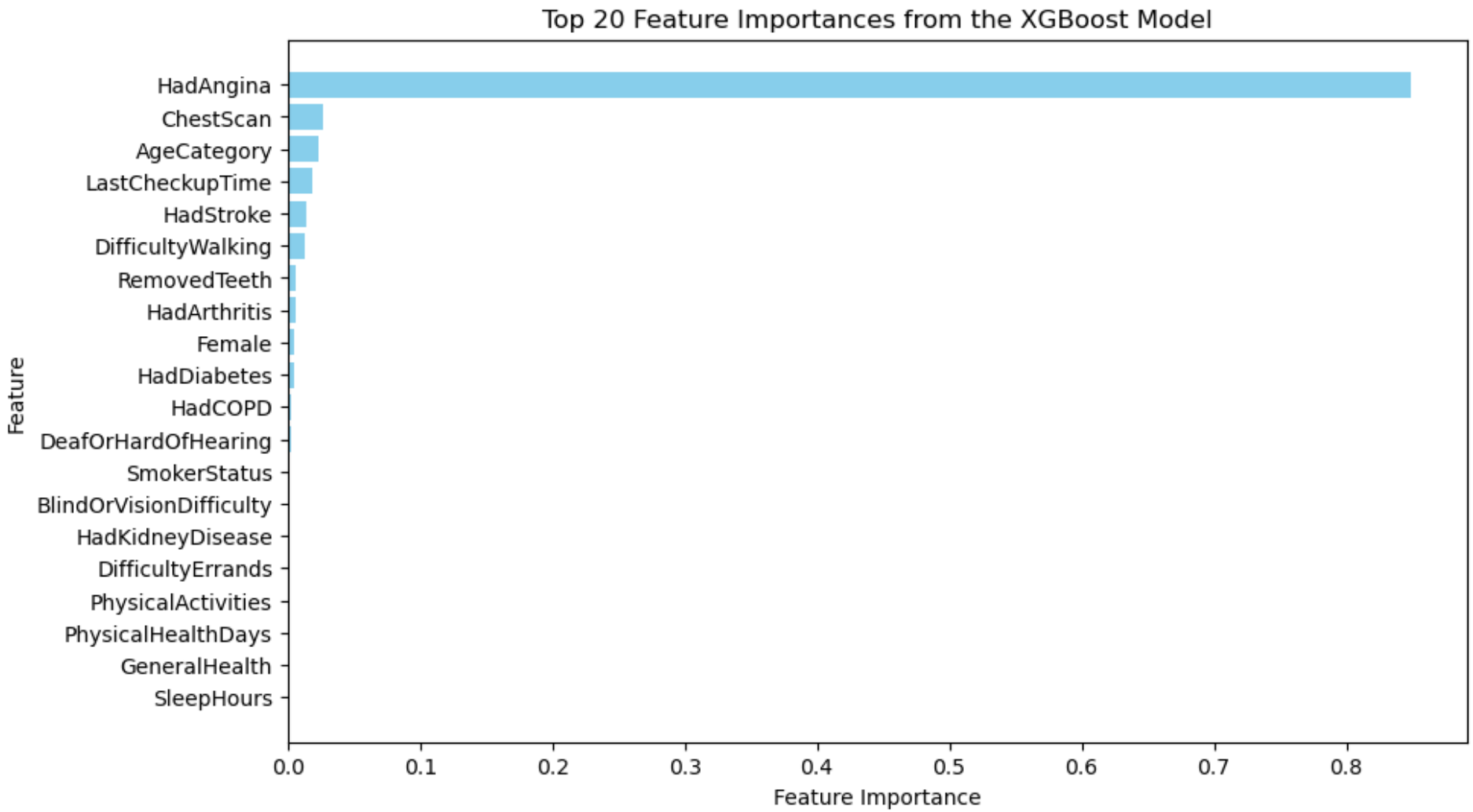
Model Performance Comparison on Training Data

	Logistic Regression	Random Forest	XGBoost	Naïve Bayes
For Class 1				
Sensitivity/Recall	0.76	0.78	0.84	0.71
Precision	0.22	0.20	0.17	0.19
F1-Score	0.34	0.32	0.28	0.30
Macro Average				
Sensitivity/Recall	0.80	0.80	0.80	0.76
Precision	0.69	0.59	0.58	0.58
F1-Score	0.62	0.60	0.57	0.60
Accuracy	0.83	0.81	0.75	0.81

Best Model

- XGBoost is the best classifier
- Model performance on testing data
  - Sensitivity/Recall: 0.85
  - Precision: 0.17
  - F1-Score: 0.28
  - Accuracy:0.75




Top 20 Feature Importances from the XGBoost Model
HadAngina
ChestScan
AgeCategory
LastCheckupTime
HadStroke
DifficultyWalking
RemovedTeeth
HadArthritis
Female
HadDiabetes
HadCOPD
DeafOrHardOfHearing
SmokerStatus
BlindOrVisionDifficulty
HadKidneyDisease
DifficultyErrands
PhysicalActivities
PhysicalHealthDays
GeneralHealth
SleepHours
Feature
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Feature Importance

The top 10 most important features

- Comorbidities:
  - HadAngina
  - HadStroke
  - Had Arthritis
  - HadDiabetes
- Demographic Characteristics:
  - Age
  - Female
- Health-Related Factors:
  - ChestScan
  - LastCheckUpTime (recent)
  - RemovedTeeth
- Impairment of Daily Living:
  - DifficultyWalking

Recommendations for CDC

- Identify high-risk individuals and provide targeted interventions
  - Individuals with a history of heart attack: secondary prevention
    - Medication adherence, regular follow-ups, cardiac rehabilitation programs, and diet and lifestyle coaching
  - Individuals without a history of heart attack: primary prevention
    - Those flagged by the model as *false positives* (high-risk individuals)
    - Heart disease and lifestyle education, management of risk factors (e.g., high blood pressure), preventive health screenings, physical activity programs, dietary counseling, and stress reduction workshops

Recommendations for CDC

- Enhancing predictive power and development of risk assessment tool
  - Refine model using future annual survey data to improve utility and reliability
  - Heart attack risk assessment tool
- Trends in emerging risk factors
  - Monitor changes in importance of risk factors over time
  - Revise preventive strategies

Summary and Conclusion

- The XGBoost model offers the most effective classification for determining whether individuals have had heart attack(s)
- Identified important comorbidities, demographic characteristics, health-related factors, and impairment of daily living related to heart attacks
- Limitation: data imbalance, low precision and f-1 score
- Public health impact
  - CDC or other public health organizations can proactively identify risk groups and allocate resources accordingly to improve health at the population level
  - Develop risk assessment tool
  - Track changes in risk factors over time