Heart Attack Capstone Project Final Report

Hio Wa Mak

10/31/2024

Heart disease remains the leading cause of death in the United States (CDC, 2021), with someone experiencing a heart attack every 40 seconds. Each year, approximately 805,000 people have a heart attack each year in the United States. Heart disease costs a huge burden of $239.9 billion annually. Given this impact, it is crucial to identify specific demographic, health, and lifestyle risk factors associated with heart attacks to inform more targeted prevention strategies and interventions.

The central question for this project is: How can the Centers for Disease Control and Prevention (CDC) develop an algorithm that predicts the likelihood of a heart attack with over 80% sensitivity (recall) by the end of the year, based on an individual's demographic profile, medical history, and other health and lifestyle factors?

The dataset used in this analysis was sourced from Kaggle and originates from the CDC's 2022 Behavioral Risk Factor Surveillance System (BRFSS). Initially comprising over 300 variables, 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.

I conducted data wrangling by dropping irrelevant variables and removing rows with missing data on the target feature, "HadHeartAttack", which indicates whether the person has a history of heart attack. Then, I reviewed the distributions of the continuous variables and identified potential outliers. After closer examination, these outliers did not appear to be data entry errors and were therefore retained in the dataset. Next, I analyzed the descriptive statistics of all categorical variables in the dataset, noting that many of them were highly imbalanced (e.g., 5.7% of individuals had experienced a heart attack). I further examined participants' demographic characteristics, including sex, race, age, and their intersections. Among all participants, 53% were female, 74% were Non-Hispanic White, and the entire sample consisted of adults, with a higher proportion falling in the 55 to 75 age range (Figure 1). Most participants reported their health as "good" (32%) or "very good" (34%) (Figure 2). Males also had higher rates of heart attack history across all ages and sex differences is especially pronounced after Age 50 (Figure 3).

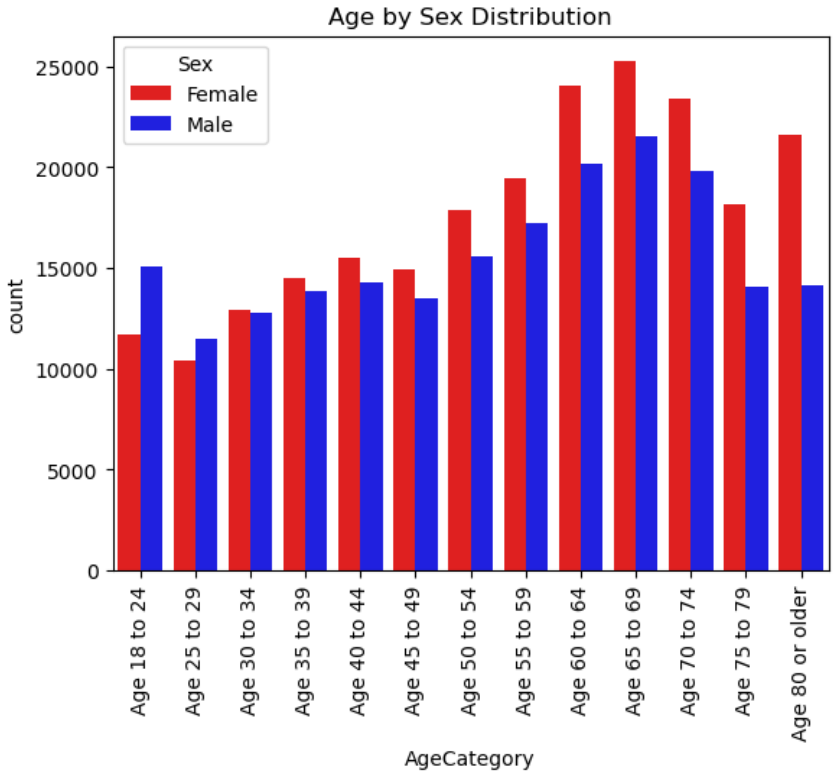


Figure 1. Participant Distribution by Age and Sex.

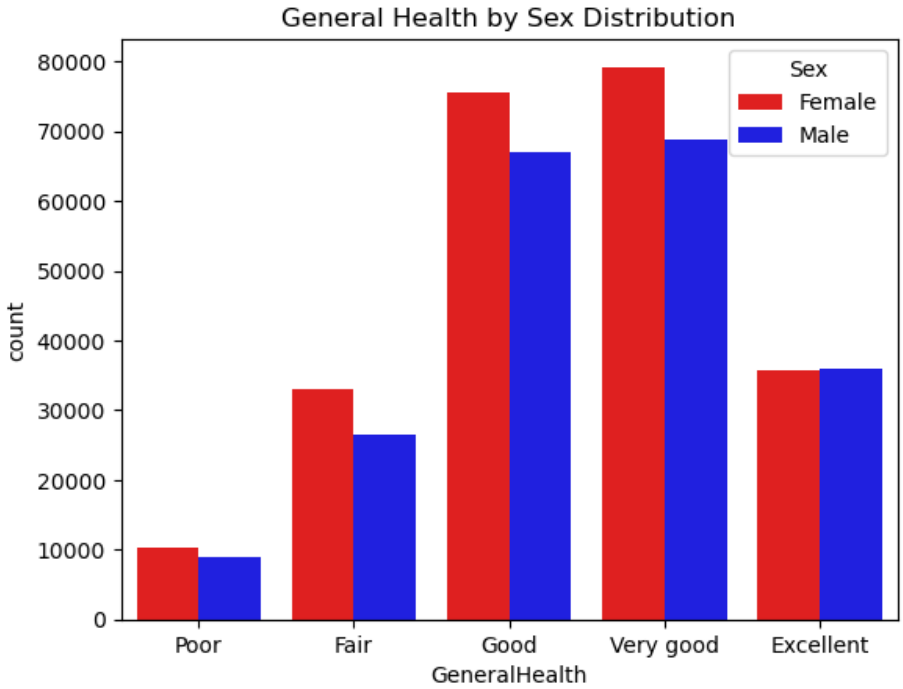


Figure 2. Perceived General Health by Sex.

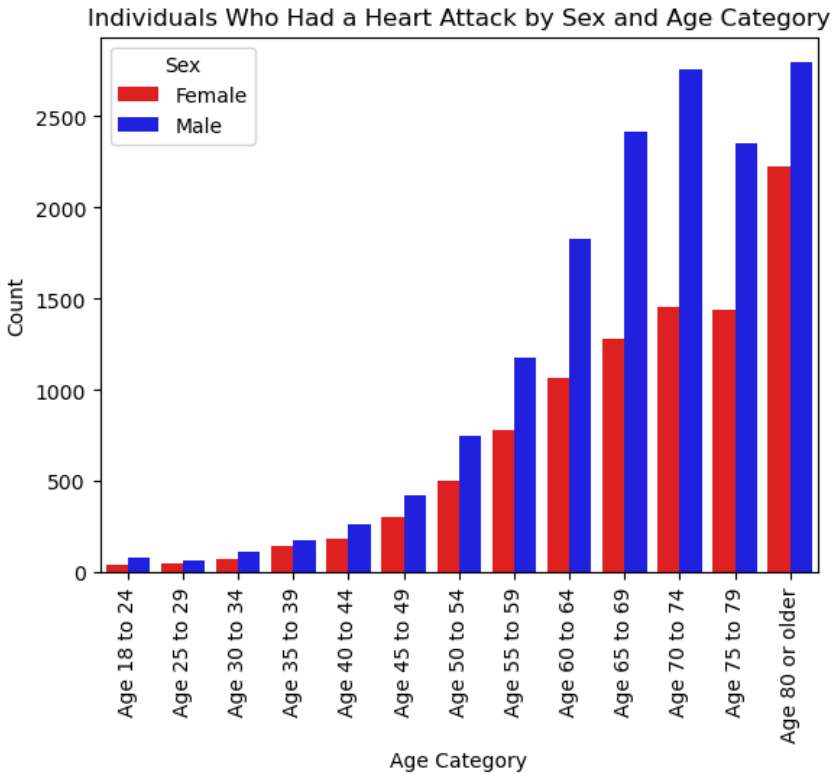
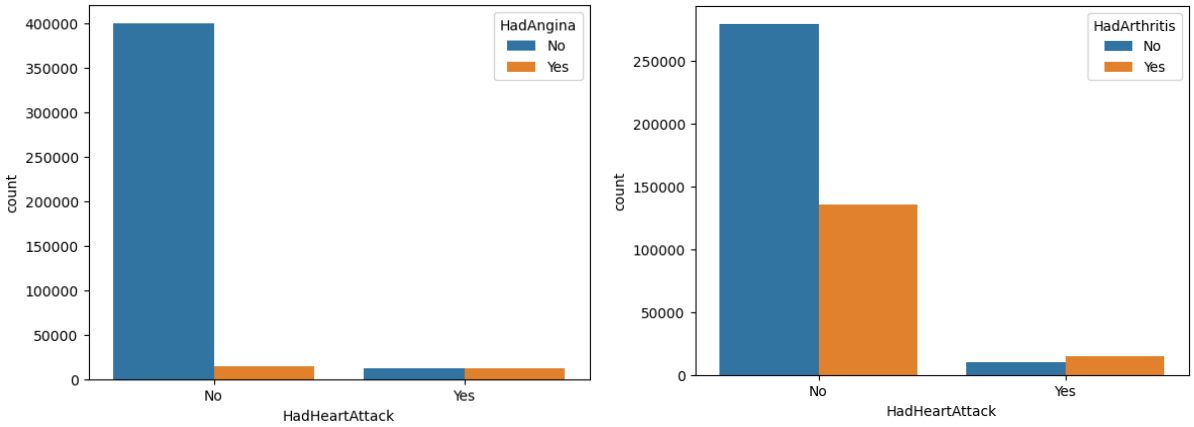


Figure 3. Heart Attack Rates by Age and Sex.

I explored the relationship between heart attack history and other categorical and ordinal variables, including conditions like angina, demographic characteristics, additional health conditions, and lifestyle factors. Exploratory data analysis indicated that variables such as “HadAngina” (history of angina), “HadArthritis” (history of arthritis), “ChestScan”(history of chest scan), older age, and male sex were particularly associated with heart attack history. Additionally, a higher number of self-reported physically unhealthy days appeared to correlate with heart attack history.



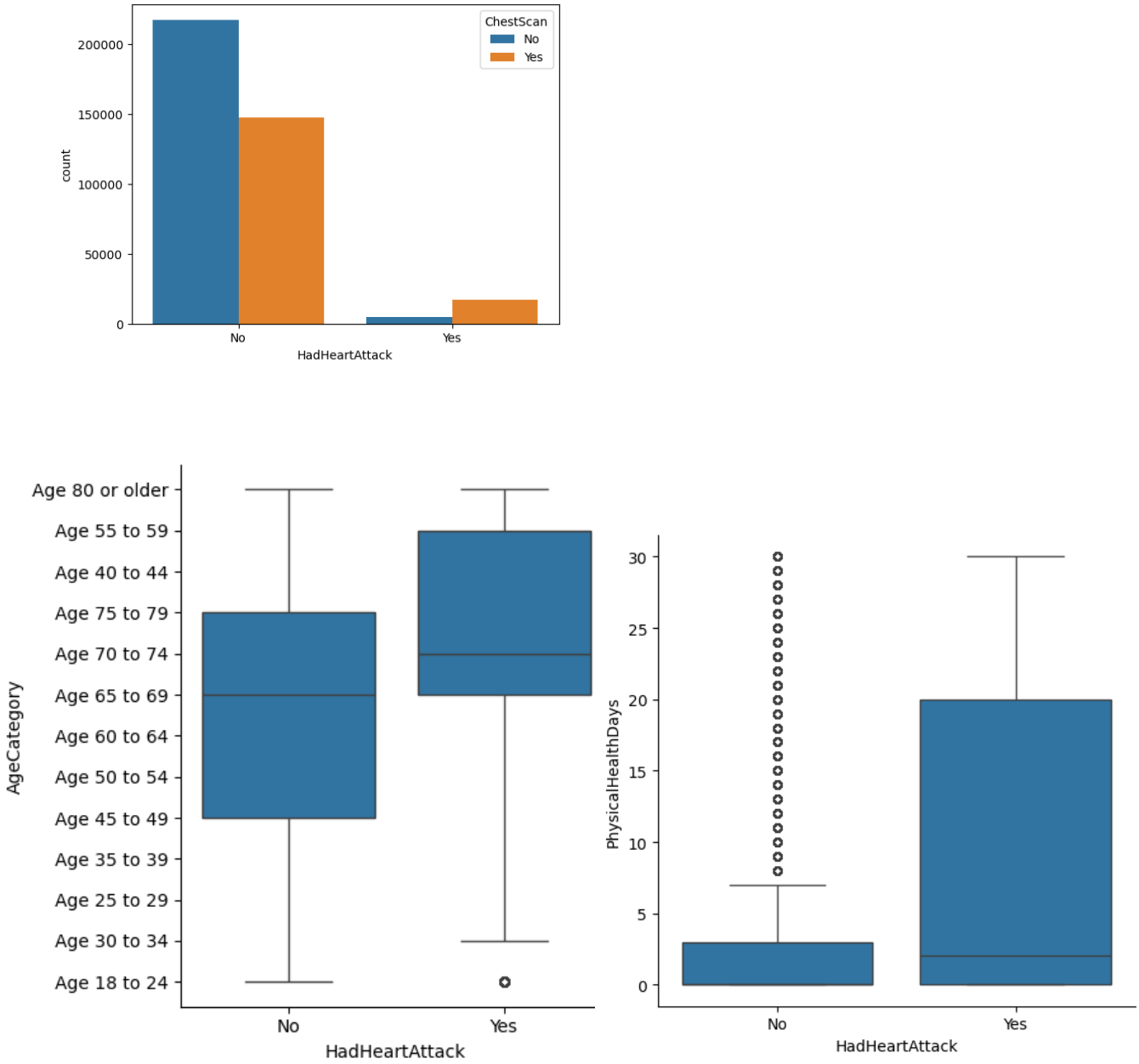


Figure 4. Health and Age Differences Between Individuals With and Without a History of Heart Attack

Then, I recoded the categorical variables into numbers. Binary variables were converted to 0 and 1, while most ordinal categorical variables were encoded as integers, except for state and race/ethnicity. I also used coded BMI into ordinal categories called “BMI_category” and subsequently dropped the original weight, height, and BMI columns to minimize redundancy. Race/ethnicity was recoded using dummy encoding for later modeling. Additionally, I removed data from non-U.S. states and eliminated columns with duplicated information after recoding.

I further explored the associations between heart attack and other predictors using a heatmap (see Figure 5). Results suggest that a “HadAgina”, “ChestScan”, “RemovedTeeth”, “PhysicalHealthDays”(number of physically unhealthy days), “GeneralHealth”, and “AgeCategory” were all associated with heart attack history.

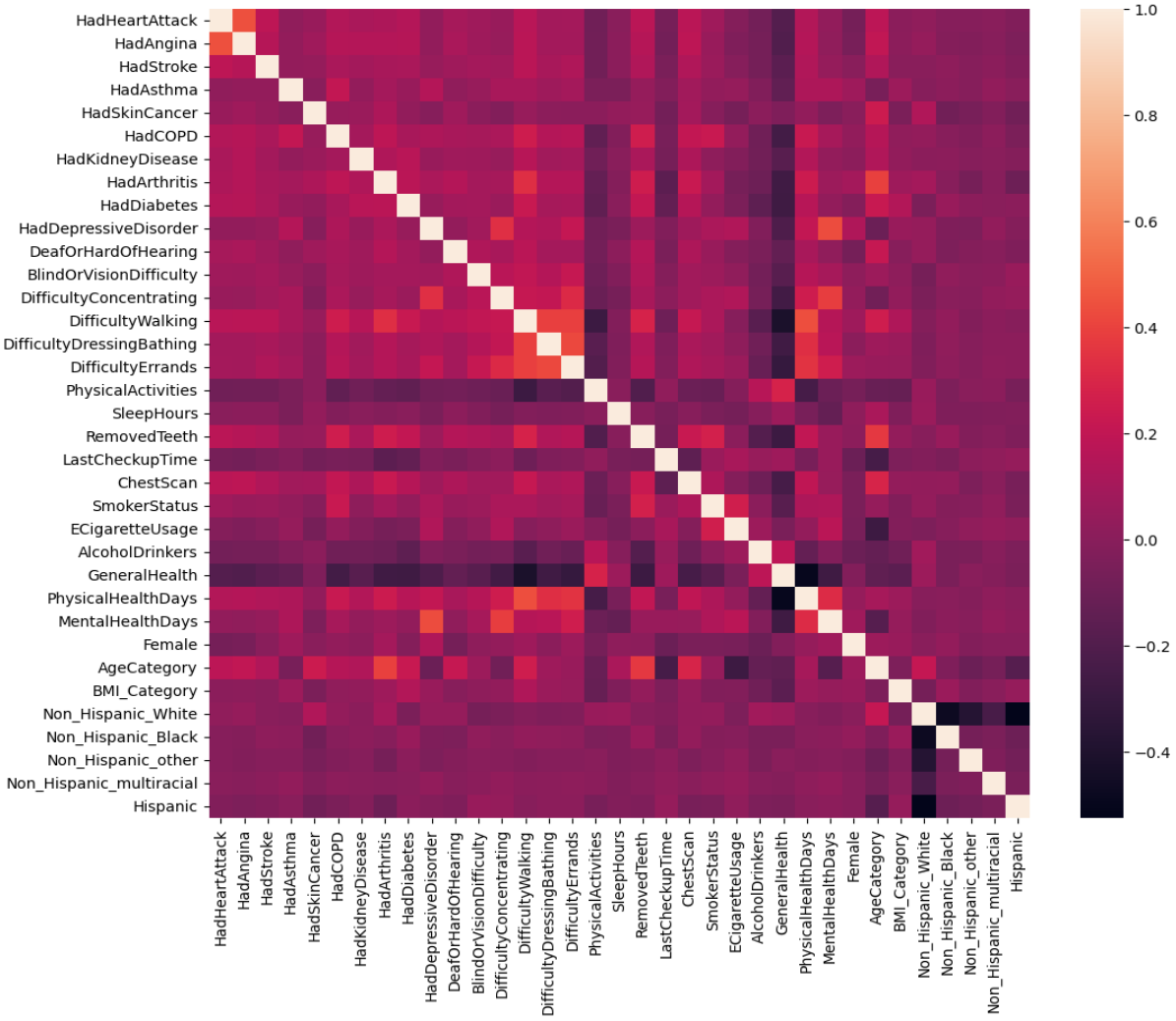


Figure 5. Heatmap Showing Correlations with Heart Attack History.

I defined the relevant features predicting heart attack history as those having correlations >= |0.1|. These variables were “HadAngina” (r = .44), “HadStroke” (r = .19), “HadCOPD” (r = .14), “HadKidneyDisease” (r = .12), “HadArthritis” (r = .12), “HadDiabetes” (r = .15), “DeafOrHardOfHearing” (r = .10), “DifficultyWalking” (r = .17), “RemovedTeeth” (r = .17), “ChestScan” (r = .17), “GeneralHealth” (r = -.19), “PhysicalHealthDays” (r = .14), “AgeCategory” (r = .18).

I split the data into training and testing sets. To address missing values, I applied mean imputation for numeric features and mode imputation for categorical features. Once the missing values were imputed, I combined the categorical and numeric features back into a single training set and a single testing set. Finally, I standardized all features to ensure they are on the same scale, a crucial step for many machine learning models. Note that missing data imputation and feature standardization were performed after splitting the data to avoid data leakage. The processed were saved for subsequent modeling.

The primary objective of this project is to identify the best classifier and optimal model parameters to predict whether an individual has had a heart attack. The process involves the following 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.

Given the imbalanced nature of the data (5.7% of participants experienced a heart attack), specific strategies were employed to address this challenge. First, sensitivity/recall was chosen as the primary scoring criterion for cross-validation and final model selection to prioritize sensitivity. Second, class weights were added as a hyperparameter to help the model better account for the imbalance. Importantly, oversampling and undersampling techniques were not used, as the imbalance reflects the natural prevalence of heart attacks in the population, not a selection bias. Applying these techniques could lead to overfitting on the training data, reducing generalization to the test data.

Table 1 presents model performance results. Among the four models evaluated, XGBoost performed the best, achieving the highest sensitivity/recall for the positive class (0.84) and a macro-average sensitivity/recall of 0.80. While precision and F1-scores were relatively low (ranging from 0.17 to 0.34 for the positive class), this was expected due to the data imbalance and the focus on optimizing sensitivity/recall. Feature importance analysis highlights “HadAngina”, “ChestScan”, “AgeCategory”, “LastCheckUpTime”, “HadStroke”, and “DifficultyWalking” as the key predictors in this model (Figure 6).

Table 1. Comparison of Model Performance

	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

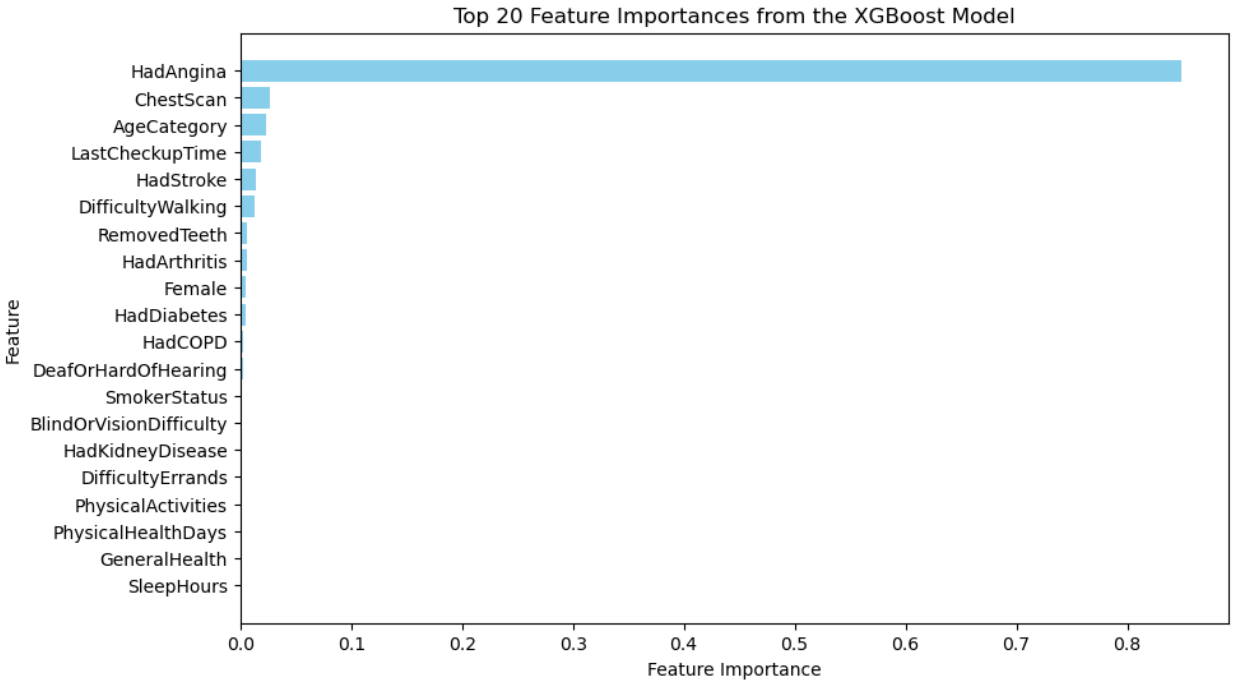


Figure 6. Feature Importance Based on XGBoost Classifier.

When evaluated on the test data, the model's accuracy, precision, sensitivity/recall, F1-score, and AUC score closely matched the training results, indicating strong generalization to unseen data (Figures 7 and 8). Lastly, the model was refitted on the entire dataset using the optimal hyperparameters, and the final model was saved for future deployment.

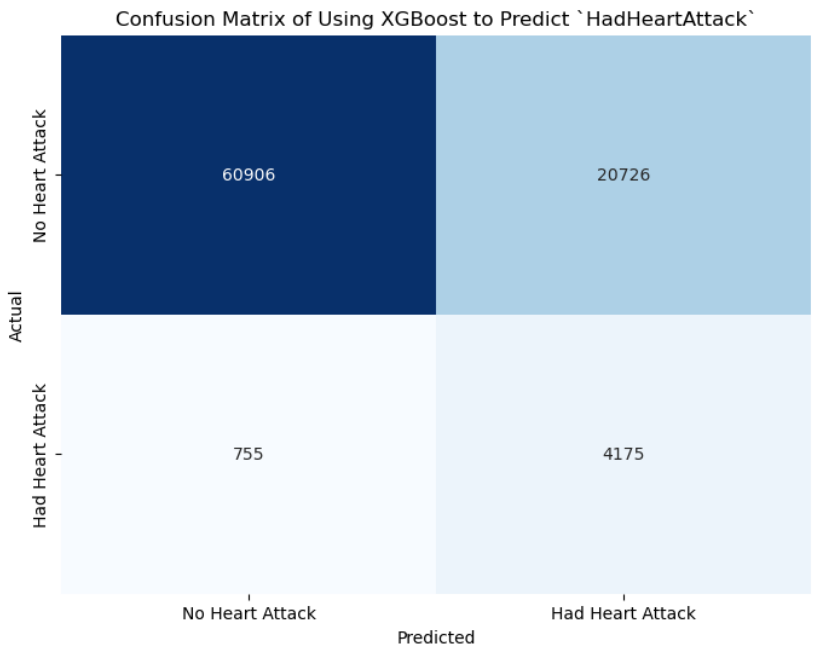


Figure 7. Confusion Matrix Using XGBoost to Predict Heart Attack.

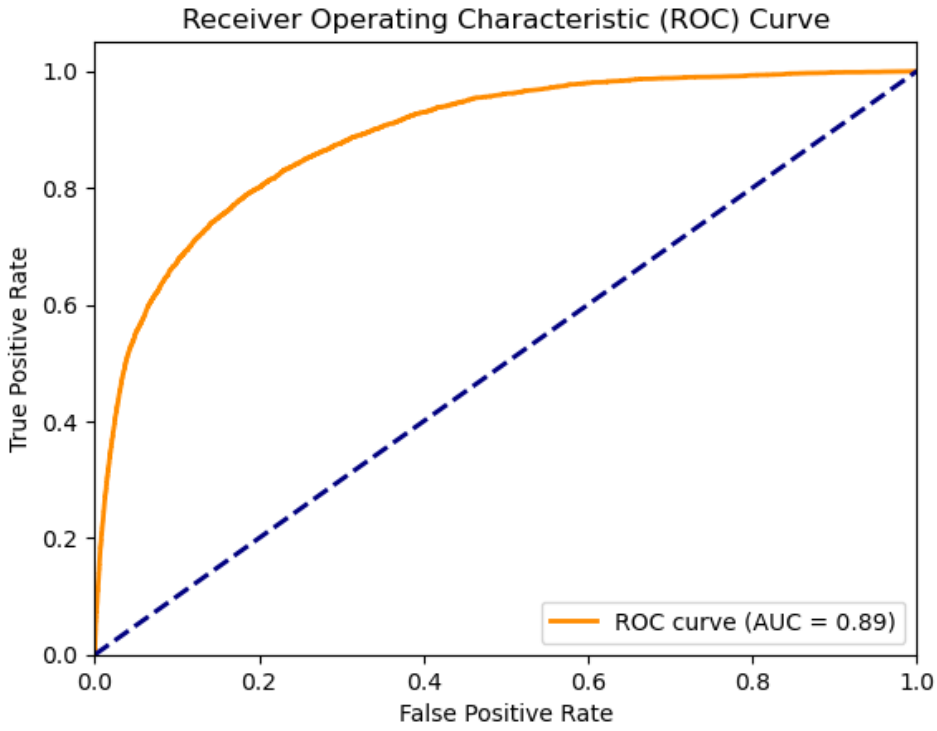


Figure 8. ROC Curve Using XGBoost to Predict Heart Attack.

Recommendations and Future Research

Identify High-Risk Individuals and Provide Targeted Interventions

The CDC can leverage this model to identify individuals at risk and provide targeted interventions. For individuals who reported a history of heart attack, the focus should be on secondary prevention to prevent recurrence, improve health outcomes, and support lifestyle changes. Recommended interventions include reminders for medication adherence, regular follow-ups, information on cardiac rehabilitation programs, and diet and lifestyle coaching. For high-risk individuals flagged by the model as false positives (i.e., predicted to have a heart attack history but have not experienced one), the focus shifts to primary prevention. These individuals could benefit from 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.

Enhancing Predictive Power and Development of Risk Assessment Tool

By applying this model to predict outcomes in next year's survey, the CDC can validate and enhance the model's performance. Over time, this approach strengthens the model's utility and reliability, making it a robust tool for heart attack risk assessment.

Trends in Emerging Risk Factors

Using the model with annual survey data allows the CDC to monitor changes in the importance of risk factors over time. By observing shifts in feature importance, the CDC may identify emerging risk factors for heart attacks, as well as track evolving trends in existing risk factors, providing valuable insights for preventive strategies.