Heart Attack Prediction Model

Will Wu

Apr 4th, 2025

Project Overview– Heart Attack Prediction

Motivation

- **Personal Experience**

Problem

- **Over 800,000 heart attacks occur annually in the U.S.**
- **Limitations in Existing Models:** 1. Age limitation: 30-79. 2. Require blood test

Solution

Early Detection System

- Preventive medications
- Respond quickly

Machine Learning Models

- Predict risk of heart attack (Yes/No)
- Less limitations

Even 1% Reduction in Heart Attack

PEOPLE

8000

Less heart attack incidents

MONEY

161

Million USD saved from healthcare expenditures annually

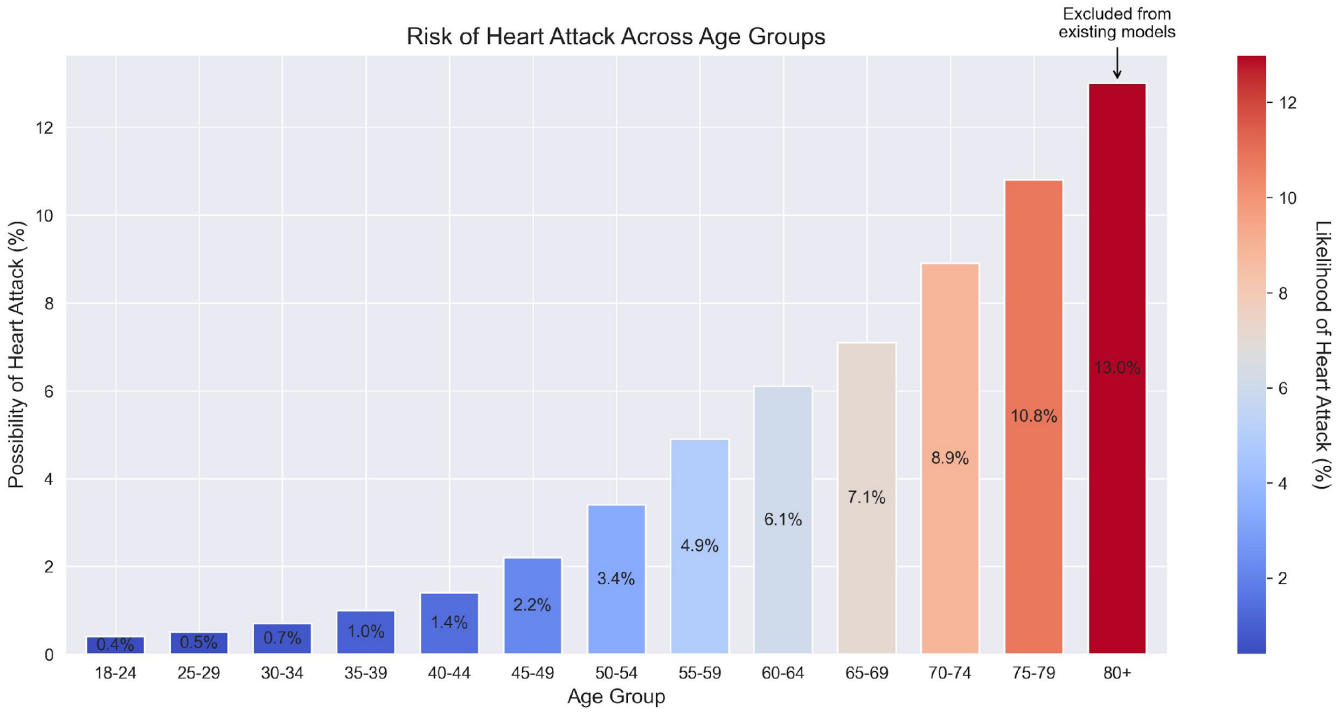
Dataset Overview

Overview

- CDC Behavioral Risk Factor Surveillance System (BFRSS) Survey Data (2022 + 2023)
- Target Variable: Had Heart Attack
- 850,000 observations
- 37 features

Issue	Solution
Missing Values: • 75% observations contain missing values • 92% features contain missing values	**Missing Values:** • Retain 86% observations • Retain all features
Class Imbalance: • 5% positive vs. 95% negative	**Resampling:** • 50% – 50% balance train data

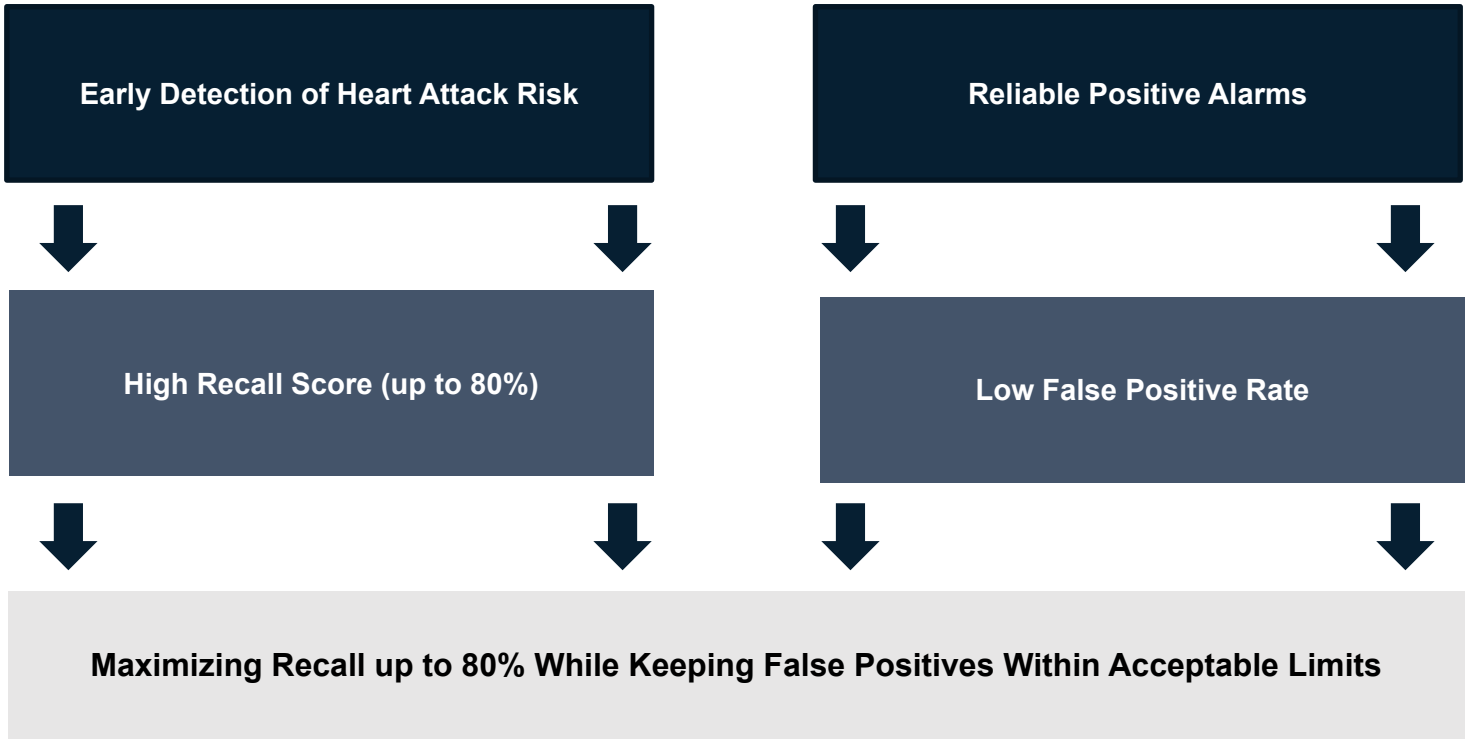
Exploratory Data Analysis (EDA)



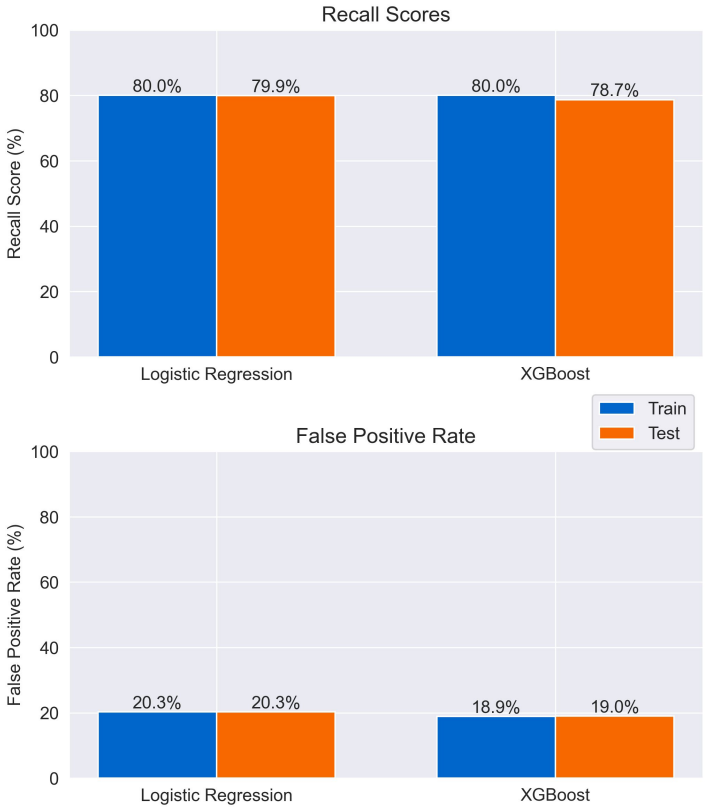
Age Group:

- Heart attack risk increases exponentially after age 45.
- Most existing models exclude individuals aged 80+.

The Metrics of Success



Modeling



Trained Models:

- 6 Models (Logistic Regression, XGBoost, Naïve Bayes, Decision Tree, Random Forest, Neural Network)

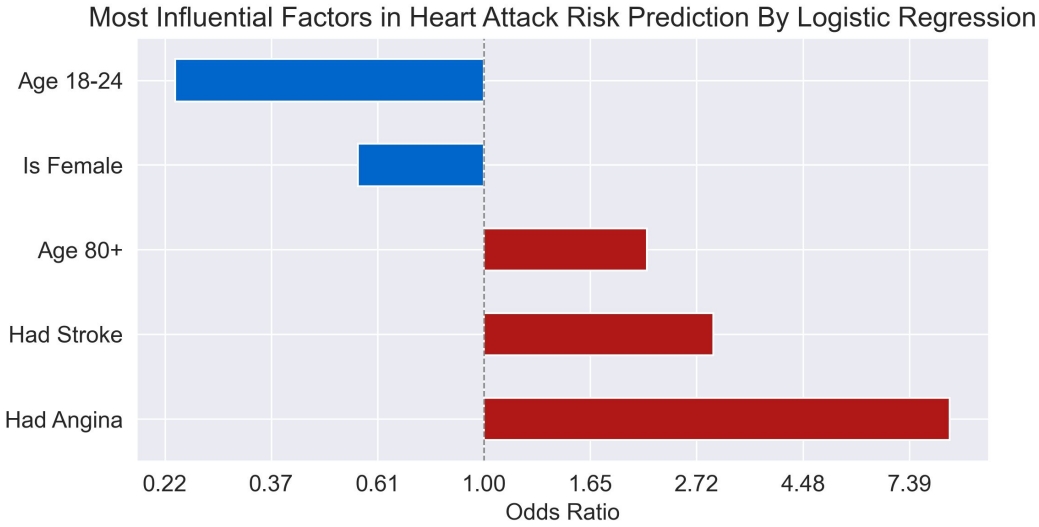
Tune Threshold:

- Meet 80% Recall target on train data

Final Model:

- Logistic Regression
- 1.2% higher in recall score = identified 9600 additional heart attacks
- Recall: The model successfully flagged 79.9% of users who actually had heart attack risk.
- FPR: For users who do not have heart attack risk, 20.3% of them receive a false alarm.

Interpretations of Logistic Regression

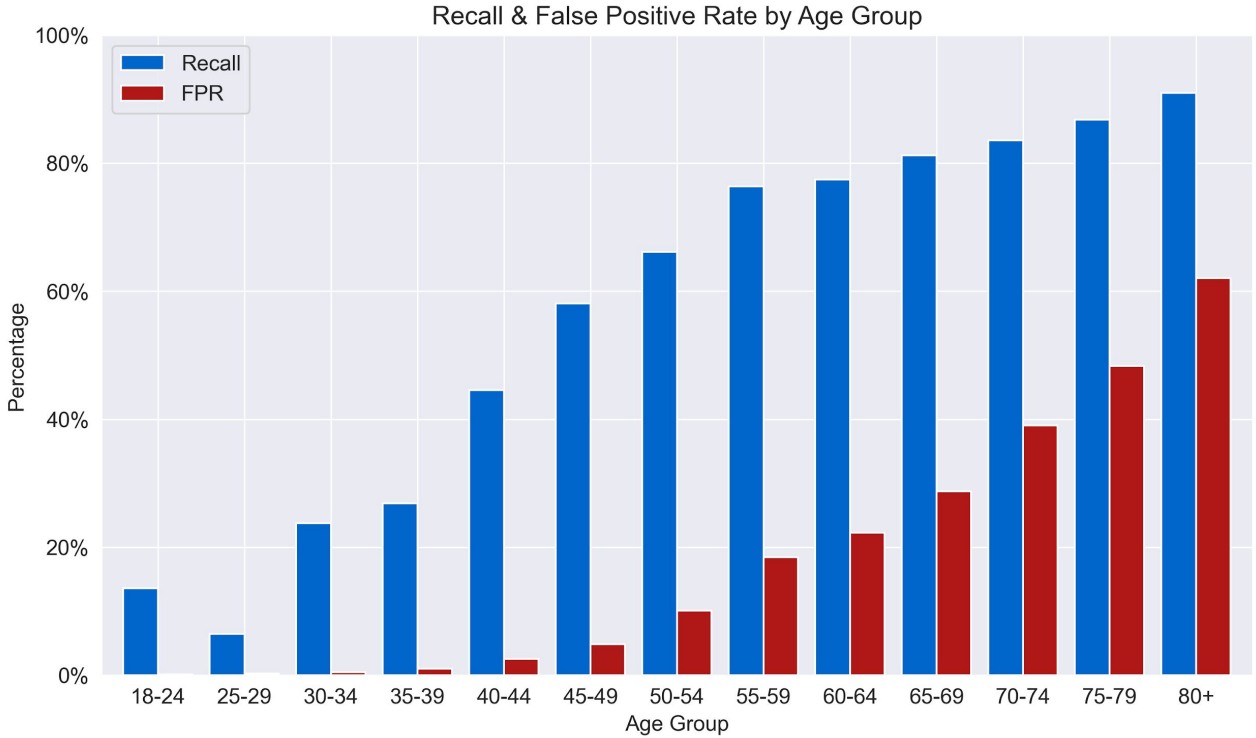


Age:
- 80+: higher odds (2.15x higher)
- 18-24: lower odds (0.23x lowers)

Angina & Stroke:
- The odds of heart attack are significantly higher when a user has any of these medical histories

Gender:
- Female: lower odds (0.55x lower)

Limitations



Trade-offs: Recall vs. FPR

- Higher recall score comes at the cost of higher false positive rate

Age 18-44:

- Caution with ‘Negative’ prediction.
- Action:
- Seek clinical evaluation with presence similar symptoms.

Age 70+:

- Caution with ‘Positive’ alarm.
- Action:
- Consult healthcare professionals
- Use different type of model for cross-validation.

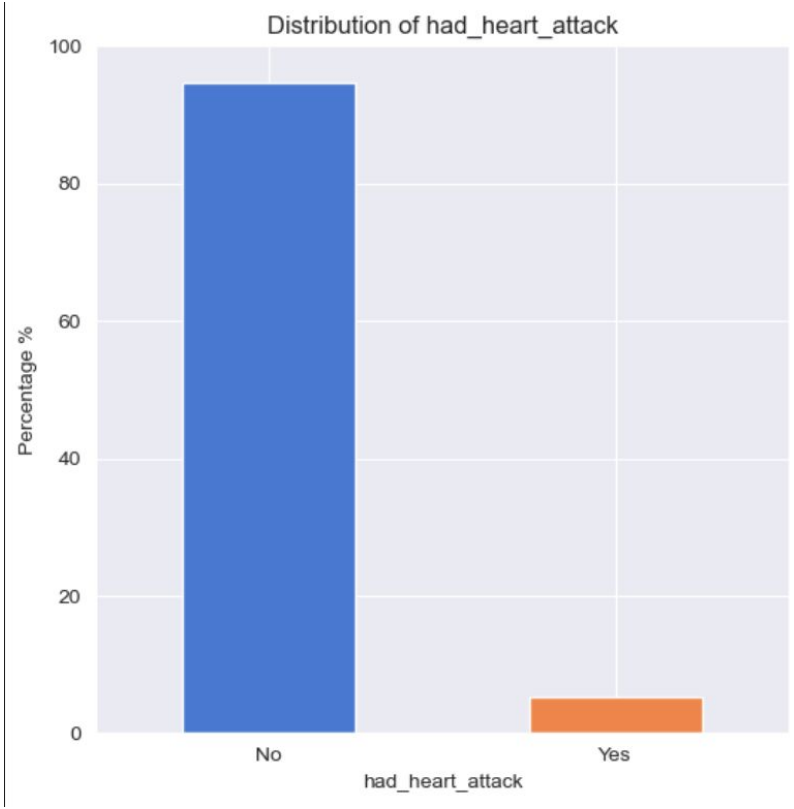
Thank You

github.com/willwu29

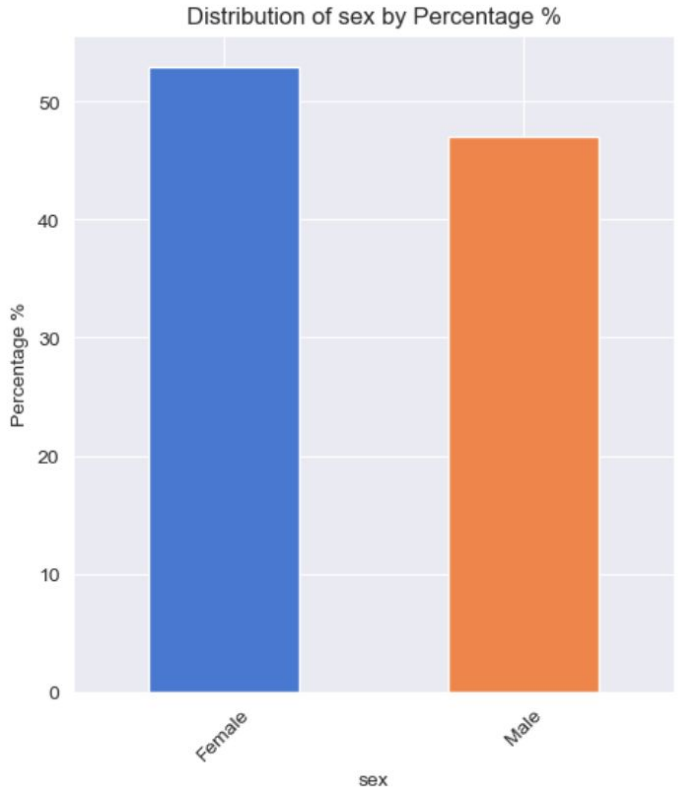
Repositories: heart-attack-prediction-model

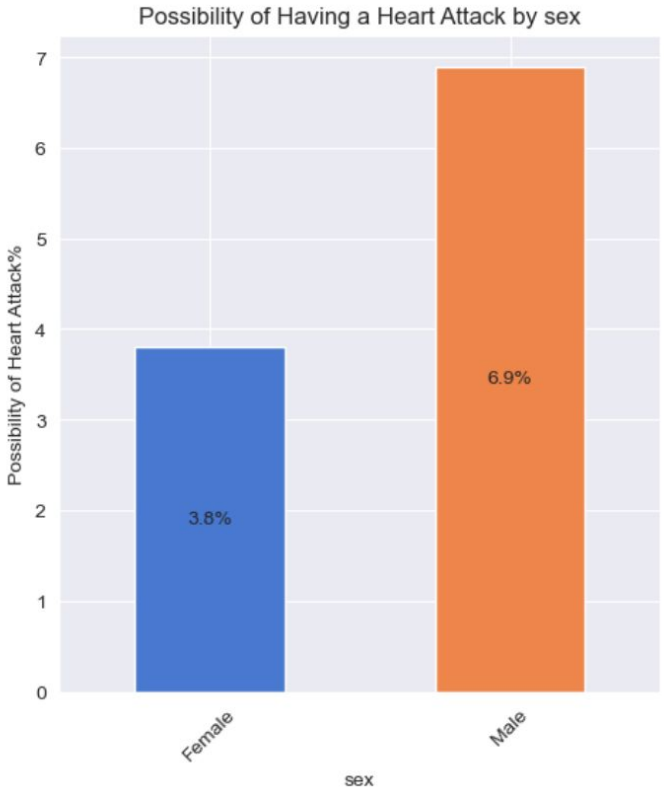
Appendix

EDA Insight: target variable: imbalanced distribution

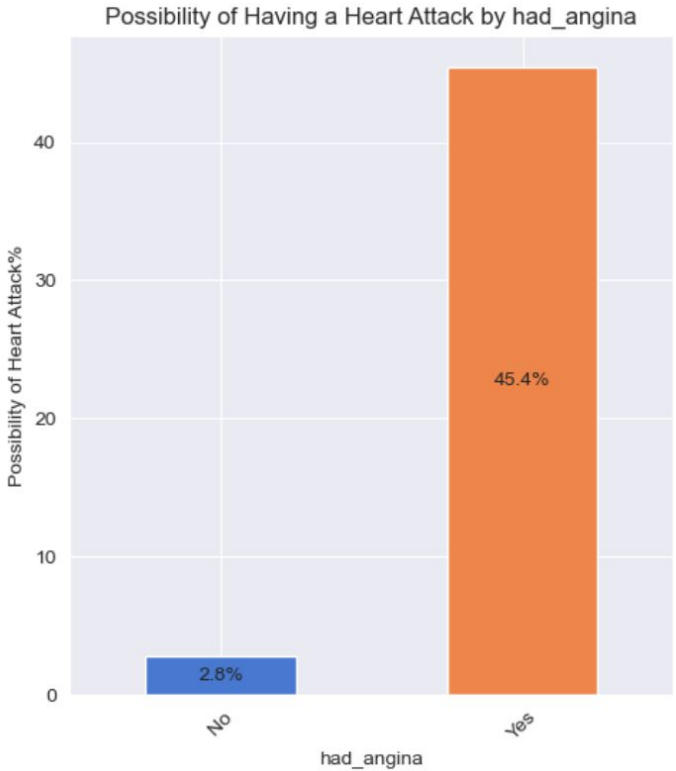


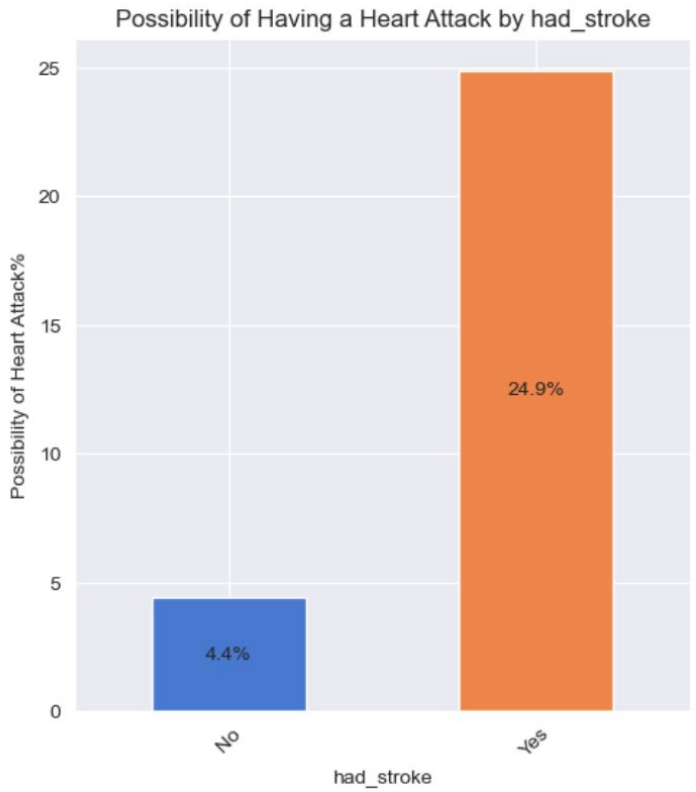
EDA Insight: gender



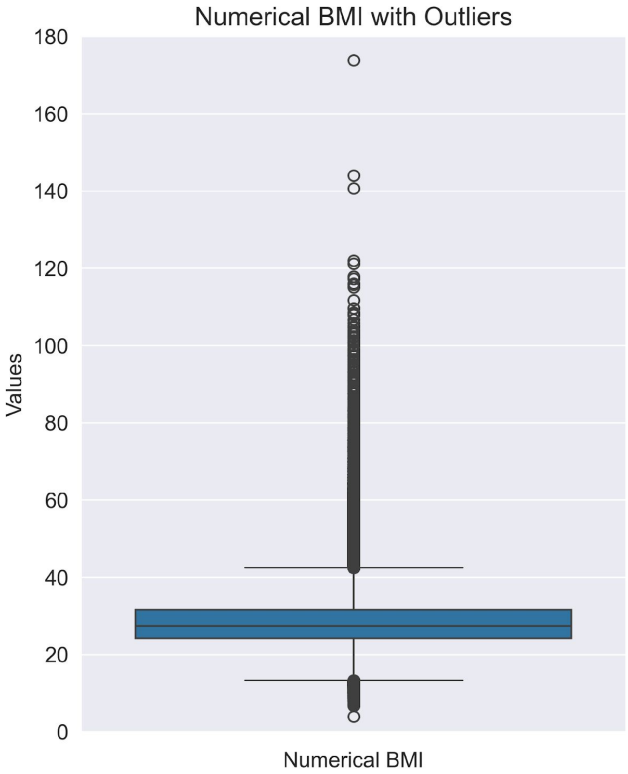


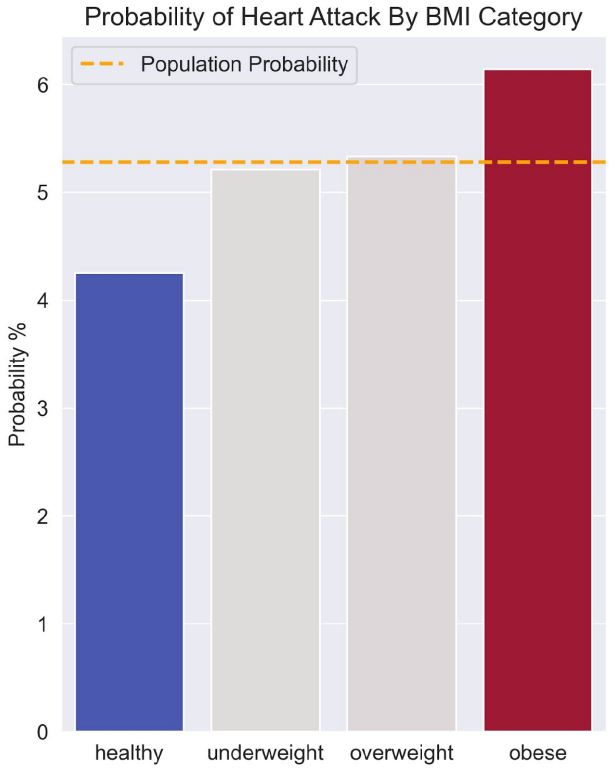
EDA Insight: had_angina, had_stroke





EDA Insights and Feature Engineering

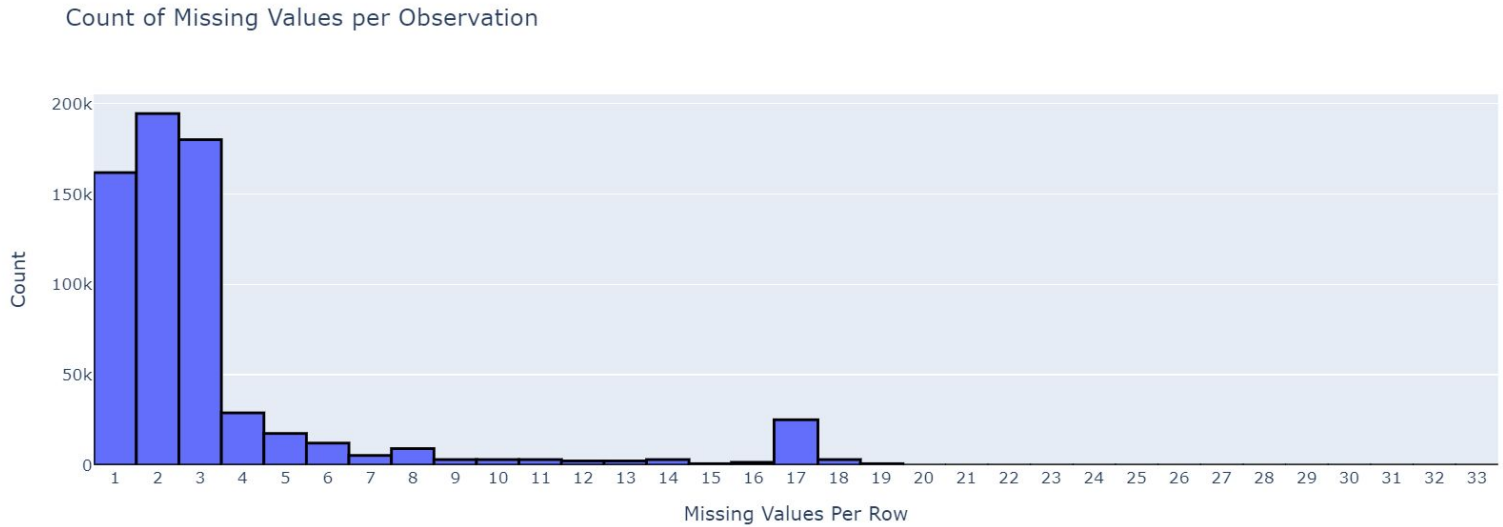




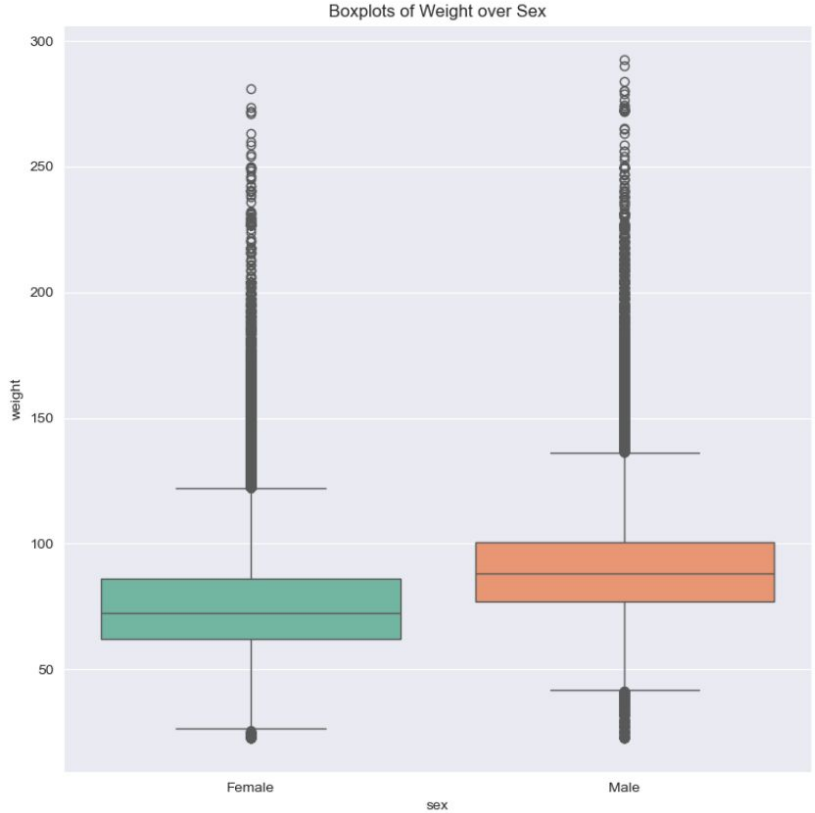
Numerical BMI (Left):
- Outliers

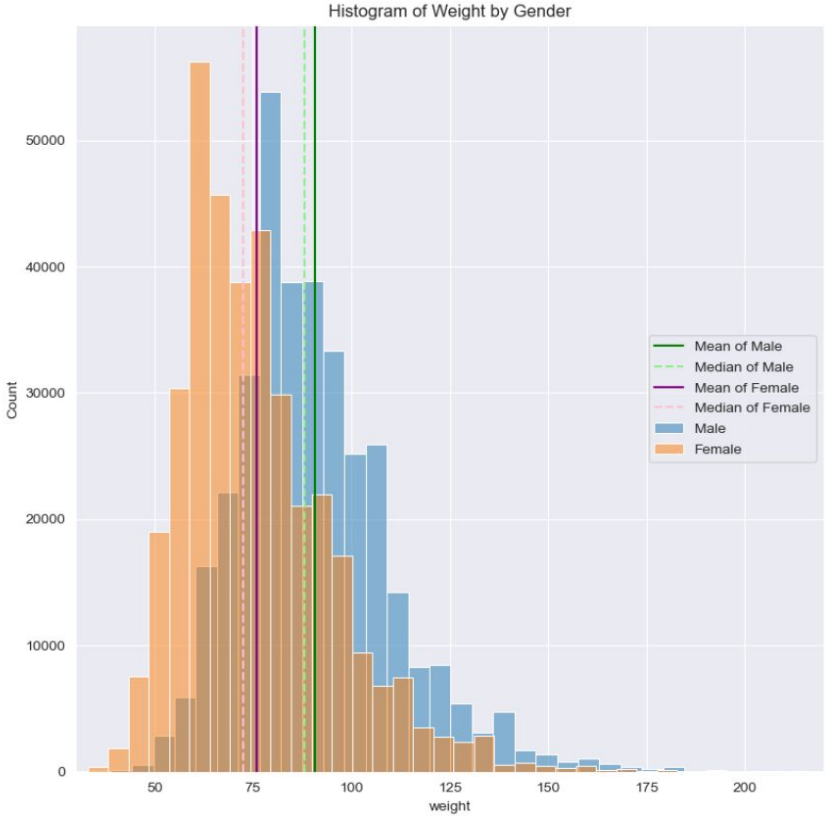
Bin BMI into categories (Right):
- Remove outliers
- Healthy group: Less likely
- Obese group: More likely

Data cleaning: missing values

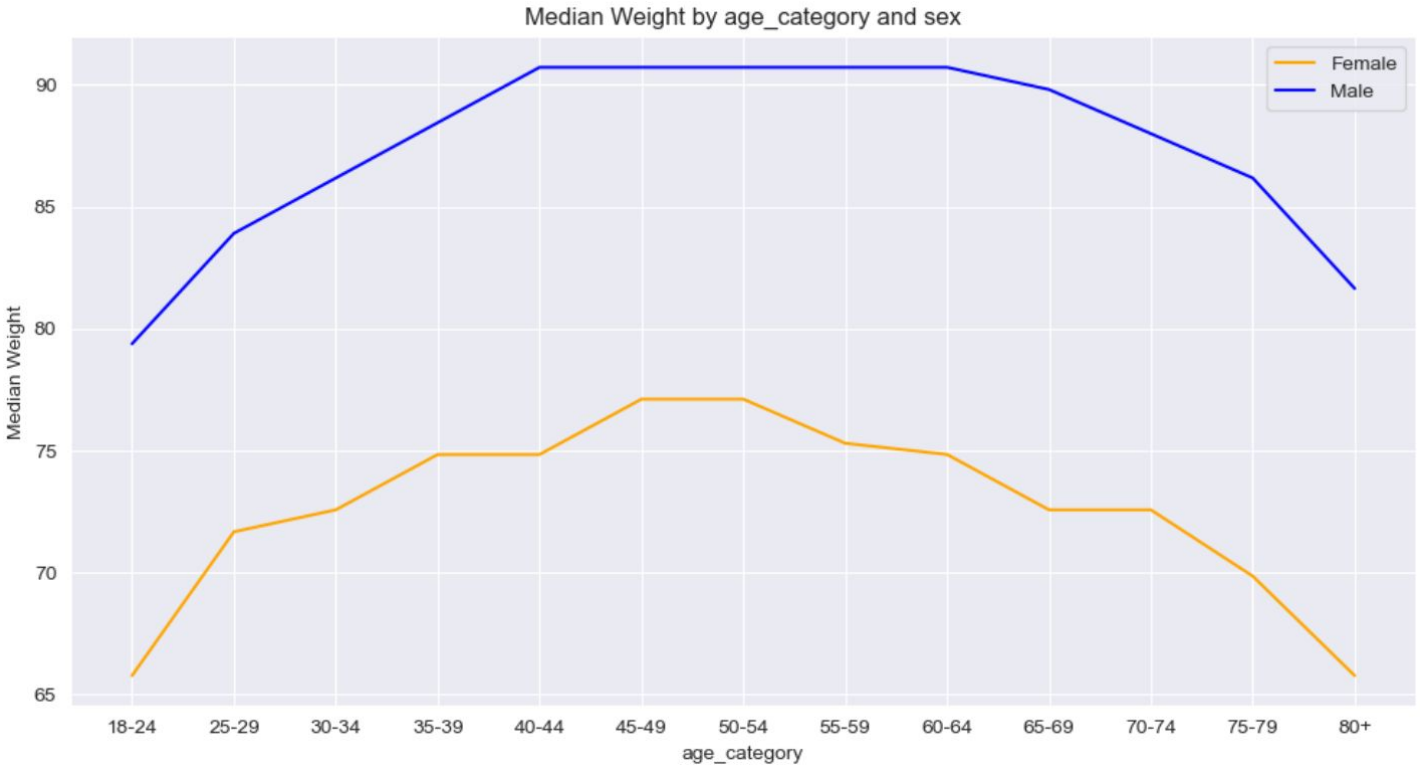


Impute missing values of weight - median of weight by gender and age





Impute missing values of weight - median of weight by gender and age



Baseline Models Metrics

Metrics Comparison Across Different Models (using test data):
random_state=42 is applied to all models that incorporate randomness.

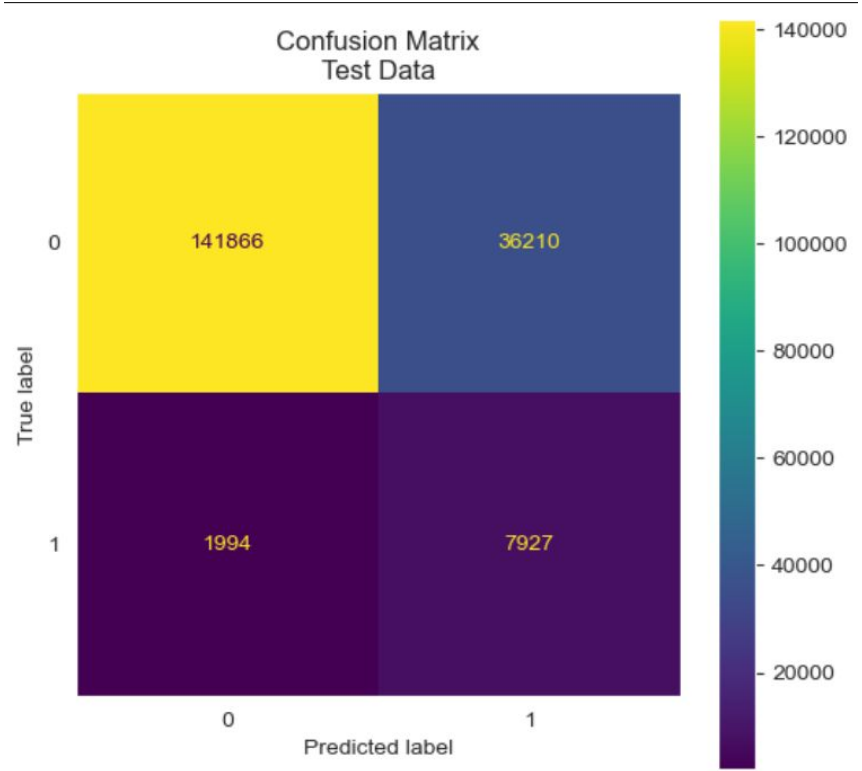
Model	X_train	Hyperparameters	PCA	Recall	FPR	Accuracy	Precision	F1 Score	AUC	Notes
Logistic Regression	Upsampled	max_iter=1000	No	0.763	0.164	0.832	0.206	0.325	0.885	
Naive Bayes	Upsampled		No	0.780	0.260	0.742	0.143	0.242	0.822	
Decision Tree	Upsampled	max_depth=7	No	0.744	0.171	0.824	0.195	0.309	0.856	
Random Forest	Upsampled	n_estimators=50, max_depth=7	No	0.766	0.188	0.809	0.185	0.298	0.871	
XGBoost	Downsampled		No	0.781	0.187	0.811	0.189	0.304	0.881	slightly overfitting
Neural Network	**Downsampled**	3 hidden layers, 10 neurons in each layer	No	**0.800**	**0.209**	0.791	0.176	0.288	0.882	

Advanced Models Metrics

Metrics using test data:

Model	Recall	FPR	AUC	Accuracy	Precision	F1	Hyperparameters	Threshold
Logistic Regression	**0.799**	**0.203**	0.881	0.797	0.180	0.293	C=0.001, penalty=l2, solver=saga	0.4500
XGBoost	0.787	0.190	0.883	0.808	0.187	0.303	n_estimators=40, learning_rate=0.12, max_depth=11, min_child_weight=10, subsample=0.7, reg_alpha=1, reg_lambda=10	0.4926
Neural Network	0.787	0.190	0.883	0.808	0.187	0.302	3 hidden layers (64, 32, 16 neurons), Dropout: 0.3, Early Stopping: Enabled, Batch Size: 256	0.4914
Random Forest	0.791	0.207	0.878	0.793	0.176	0.288	max_depth=16, min_samples_leaf=80, min_samples_split=3, n_estimators=100, bootstrap=False	0.4878
Naive Bayes	0.801	0.259	0.835	0.745	0.147	0.249	var_smoothing= 0.001	0.2458
Decision Tree	0.786	0.211	0.870	0.789	0.172	0.282	max_depth=11, min_samples_split=4, min_samples_leaf=80, criterion=entropy	0.4615

Final Model: Logistic Regression - Confusion Matrix



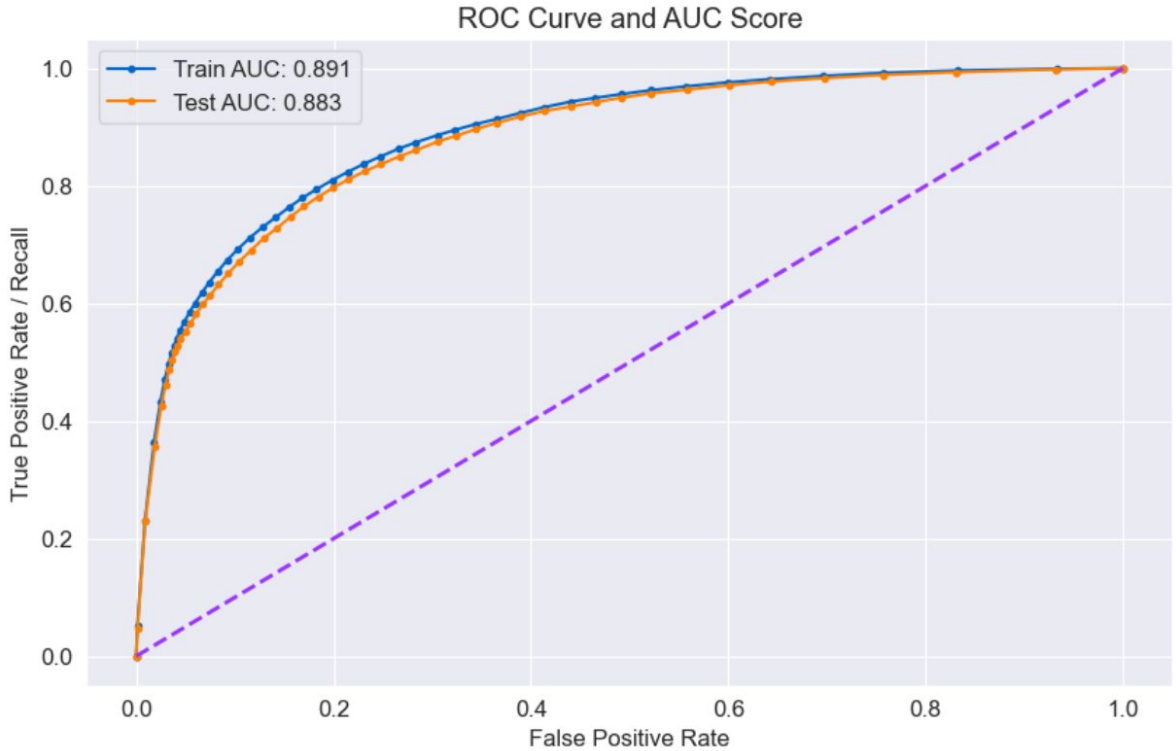
Recall:

- 7927 / (1994 + 7927) = 79.9%
- True Positives / (Total Positives)

False Positive Rate

- 36210 / (36210 + 141886) = 20.3%
- False Negatives / (Total Negatives)

Final Model: Logistic Regression - ROC Curve of Logistic Regression



Recall, False Positive Rate of Age

prediction_outcome	False Negative	False Positive	True Negative	True Positive	Recall	FPR
age_category						
18-24	38	20	11780	6	0.136364	0.001695
25-29	58	25	9518	4	0.064516	0.002620
30-34	61	55	10930	19	0.237500	0.005007
35-39	87	119	11845	32	0.268908	0.009947
40-44	87	318	12184	70	0.445860	0.025436
45-49	112	578	11273	155	0.580524	0.048772
50-54	163	1390	12370	319	0.661826	0.101017
55-59	173	2730	12064	560	0.763984	0.184534
60-64	269	3936	13725	925	0.774707	0.222864
65-69	264	5474	13587	1140	0.811966	0.287183
70-74	292	6815	10651	1488	0.835955	0.390187
75-79	207	6346	6792	1366	0.868404	0.483026
80+	183	8404	5147	1843	0.909674	0.620176

Age 18-24
Recall:

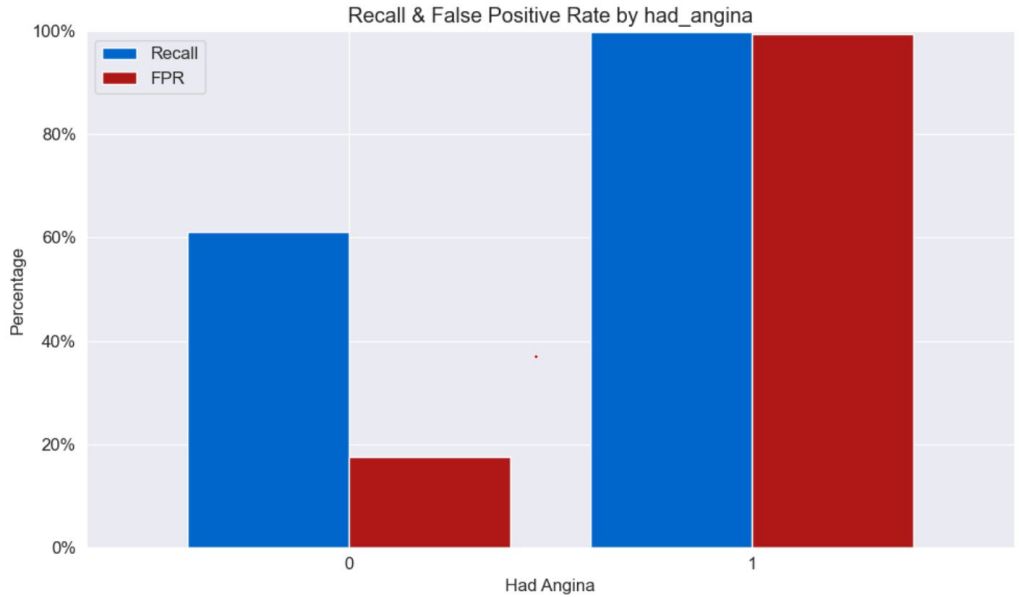
- True Positives / (Total Positives)
- 6 / (38 + 6) = 13.4%

Age 80+
False Positive Rate

- False Negatives / (Total Negatives)
- 8404 / (8404 + 5147) = 62.0%
-

Recall, False Positive Rate of had_angina

prediction_outcome had_angina	False Negative	False Positive	True Negative	True Positive	Recall	FPR
0	1990	30273	141831	3123	0.610796	0.175899
1	4	5937	35	4804	0.999168	0.994139



Supporting studies: Framingham Risk Score

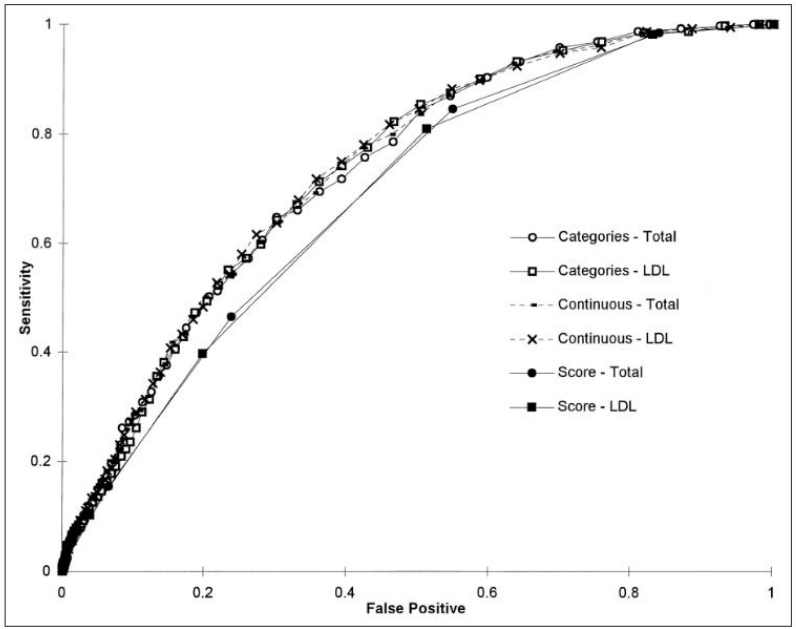


Figure 1. Receiver operating characteristic curves for prediction of CHD in Framingham men over a period of 12 years. Separate plots were used for continuous, categorical, and risk factor sum models, according to whether TC or calculated LDL-C was used.