Predicting Heart Disease Risk

MGT 6203 - FINAL PROJECT REPORT – TEAM#13

Daniel Smith | Ankur Asthana | Jimmie Jain

Predicting Heart Disease Risk

Introduction

The objective of this study is to develop an optimized model for predicting heart disease and recognize relevant and influential factors. We are extracting significant and relevant factors from the data, which can hopefully be used as insight for healthcare providers, insurance companies, or any other industry that can apply our findings for business needs.

The original data is sourced from The Behavioral Risk Factor Surveillance System (BRFSS), which is an annual health survey conducted by the CDC via *telephone, gathering information* on health-related risk behaviors from over 400,000 Americans each year (Teboul 2022). This study explores the analysis of 21 important risk factors for heart disease from 2015 survey data to determine which factors are most significant, influential, and relevant for predicting heart disease for a given population.

Justification

Our aim is to develop a fairly accurate model along with significant heart disease predictors can be vital for business decisions. For example, companies can use findings on factors such as cholesterol, fruits, and vegetables to target consumers in marketing and advertising. Business impact extends to healthcare, public health, and insurance companies. For instance, healthcare providers can optimize resource allocation by prioritizing services for high-risk individuals. Public health organizations can utilize the predictive model to identify individuals at risk of heart disease and tailor personalized preventative care strategies such as lifestyle counseling or recommended health screenings. Insurance providers can use the model to assess the risk of heart disease among policyholders. Results can be incorporated into policy prices and incentives can be offered to adopt healthier behaviors.

Approach

As we need to predict our observation into binary outcomes, we need to use decision-based machine learning techniques. Our approach is to develop a logistic regression with all variables and analyze the results. Based on the results, we can try using other ML technique to develop more models (decision-tree based models), fine-tune the model and be able to choose one of them based on metrics and more balanced result for our business objective.

Data Overview

Data Source and Exploratory Data Analysis (EDA)

A cleaned dataset was made by Alex Teboul prior to the analysis of this study. This dataset consists of 253,680 observations across 22 factors/variables. 15 of the factors are binary while 7 are non-binary. During EDA, we found that number of positive cases are 9.42%, which suggests imbalance in the dataset and motivating us to pursue oversampling techniques later in the analysis. We found that there were no missing records (Fig. 2) or outliers (using cooks' distance > 5 and Fig. 3). Fig 1 shows a snippet of all variables.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	HeartDiseas	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	Veggies	HvyAlcoholC	AnyHealthca	NoDocbcCo	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Ir
2	0	1	1	1	40	1	0	0	0	0	1	0	1	0	5	18	15	1	0	9	4	
3	0	0	0	0	25	1	0	0	1	0	0	0	0	1	3	0	0	0	0	7	6	
4	0	1	1	1	28	0	0	0	0	1	0	0	1	1	5	30	30	1	0	9	4	
5	0	1	0	1	27	0	0	0	1	1	1	0	1	0	2	0	0	0	0	11	3	
6	0	1	1	1	24	0	0	0	1	1	1	0	1	0	2	3	0	0	0	11	5	
7	0	1	1	1	25	1	0	0	1	1	1	0	1	0	2	0	2	0	1	10	6	
8	0	1	0	1	30	1	0	0	0	0	0	0	1	0	3	0	14	0	0	9	6	
9	0	1	1	1	25	1	0	0	1	0	1	0	1	0	3	0	0	1	0	11	4	
10	1	1	1	1	30	1	0	2	0	1	1	0	1	0	5	30	30	1	0	9	5	

Fig. 1 Example of Dataset

HeartDiseaseorAttack	HighBP	HighChol	CholCheck
0	0	0	0
BMI	Smoker	Stroke	Diabetes
0	0	0	0
PhysActivity	Fruits	Veggies	HvyAlcoholConsump
0	0	0	0
AnyHealthcare	NoDocbcCost	GenHlth	MentHlth
0	0	0	0
PhysHlth	DiffWalk	Sex	Age
0	0	0	0
Education	Income		
0	0		

Fig. 2 Missing Values

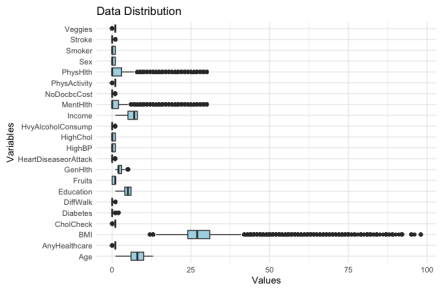

Fig. 3 Data Distribution

Exploratory Data Analysis

The data was imported from CSV to a dataframe, and a summary was made. From the summary, we observed that the dependent variable has a mean of 0.0942. This implies that 9.42% of the respondents experienced heart related issues. Additionally, we observed that most people say they had their cholesterol checked in the last 5 years, the mean BMI is 28.38, 15.8% of respondents has/had diabetes most people claimed they are not heavy drinkers, and the mean/median age group of respondents is 8, corresponding to 55-59 years of age. Next, each variable was checked for collinearity by using VIF analysis (VIF > 5 indicating collinearity). Collinearity was not found among any of the predictors.

According to initial analysis and correlation matrix (Fig. 4), HeartDiseaseorAttack shows positive correlations with variables like HighBP, HighChol, Stroke, Diabetes, GenHlth, Age, and Income. This suggests that individuals with higher blood pressure, cholesterol levels, history of stroke, diabetes, lower general health ratings, older age, and lower income are more likely to have heart disease or a heart attack. While PhysActivity shows a negative correlation with HeartDiseaseorAttack, indicating that individuals who engage in more physical activity are less likely to have heart disease or a heart attack.

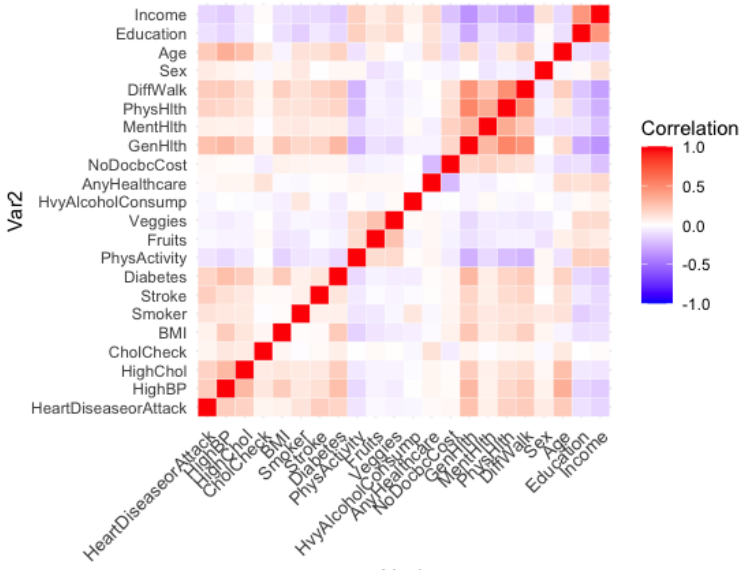

Fig. 4 Feature Correlation Matrix

Feature Engineering

We used Boruta algorithm to iteratively evaluate the importance of each feature in our dataset. It began by creating a shadow feature set through duplication and shuffling of the original data. Using a machine learning model, it assesses the importance of each feature. It then examined the discrepancy between the importance of original features and their shadow counterpart to distinguish between important and unimportant features. At the

end of the process, it comes up with a mean importance and decision results for every feature. The below Fig. 5 shows all the features with their mean importance.

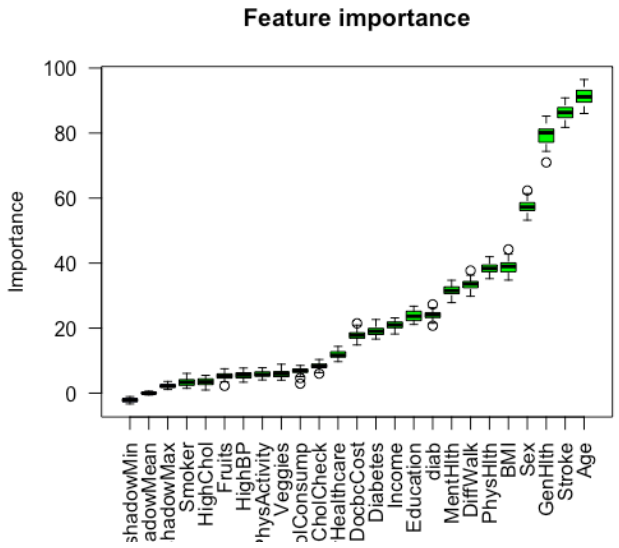

Fig. 5 Boruta Feature Importance

During our initial model development, we prepared the first model with all parameters and then truncated the parameter to include only top important/confirmed features. We confirmed that the model results were stable when the other features were removed.

We found the following predictors to be the most significant statistically – *Age, Stroke, GenHlth, Sex, BMI, DiffWalk, MentlHelth, diab, HighBP, HighChol, CholCheck, and Smoker.* **All our subsequent analysis is based on the top 12 important/confirmed features.**

Model Development

As our problem statement requires us to predict our observation into binary outcomes, we thought that model development using decision-based machine learning techniques could be useful. We started with base logistic regression and XGBoost model with a standard threshold of 0.5 and compared their performance. Based on the results, we selected the better performing model and improved the efficiency by either oversampling (as we know dataset is highly imbalanced and has less than 10% positive cases) and/or parameter tuning. The code is available in the Code folder at GitHub.

Logistic Regression

After importing the dataset, we preprocessed it to consider borderline diabetic individuals as diabetic. We then split the data into training and testing sets (80:20) with a seed of 123 and applied logistic regression to model the relationship between various predictors and the occurrence of heart disease. As noted, we used a threshold of 0.50 to classify the test set outcomes and evaluate the model's performance using a confusion matrix.

In our assessment, we found that the model is good in identifying people who don't have heart disease (specificity ~ 98%) but struggles to catch those who have heart issues (sensitivity ~ 12%). Despite 90% accuracy, there are cases where the model fails to accurately predict whether someone has heart disease or does not have heart disease. This indicates that we need to fine-tune the model to improve spotting people who are at risk of heart problems.

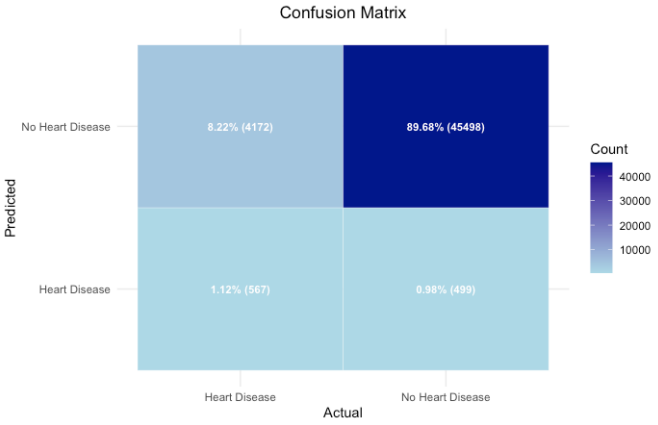

Fig. 6 Confusion Matrix: GLM

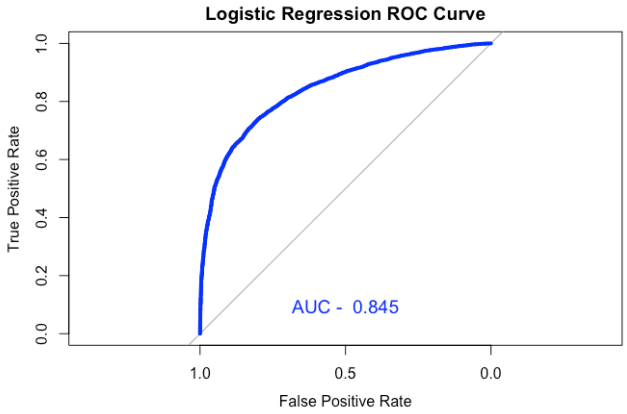

Fig. 7 ROC Curve: GLM

XGBoost

In the second approach, we imported the dataset and followed the same preprocessing step, considering borderline diabetic individuals as diabetic. We partitioned the data into training and testing sets (80:20) for subsequent analysis. We ensured that we are using the same seed as used in Logistic Regression. In this approach, we used gradient boosting algorithm XGBoost to capture the relationships between predictors and the likelihood of heart disease. Utilizing a series of decision trees (in this case, 100), the model learns iteratively from the data to improve its performance with each iteration. After training, we used a threshold of 0.50 to classify the outcomes of the test set and evaluated the model's performance using a confusion matrix.

The resulting XGBoost model exhibited a high specificity (~91%). This shows that it is effectively identifying individuals without heart disease. However, the XGBoost model was much better as compared to logistic regression model in detecting individuals with heart disease (sensitivity ~ 53%). The overall accuracy remained around 90% and the positive detection prevalence improved to 2%. This shows that XGBoost is a little better but is still affected by imbalanced data.

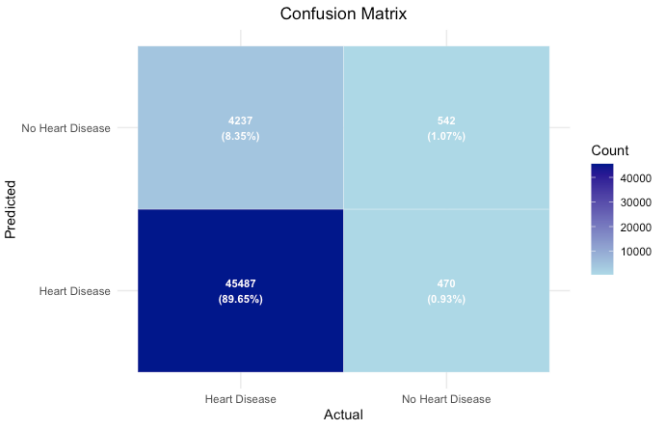

Fig. 8 Confusion Matrix: XGBoost

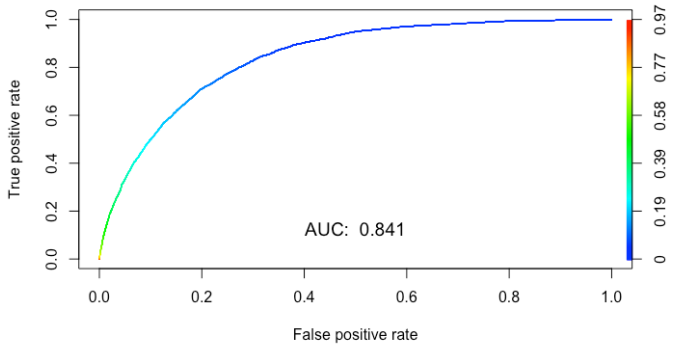

Fig. 9 ROC Curve XGBoost

Model Comparison

Below is the tabular comparison of the two base models –

Metric	XGBoost	Logistic Regression (GLM)
Accuracy	90.72%	90.79%
Sensitivity	53.55%	11.96%
Specificity	91.48%	98.91%
Positive Predictive Value (PPV)	11.34%	53.19%
Negative Predictive Value (NPV)	98.87%	91.60%
Prevalence	1.99%	9.34%
Balanced Accuracy	72.58%	55.44%
AUC	84.09%	84.51%

Table 1 Performance - XGBoost vs GLM w/ 0.5 threshold

Comparing the two models, both XGBoost and logistic regression achieved similar accuracy. The XGBoost model achieved an accuracy of 90.72% with a 95% confidence interval of (0.9047, 0.9097), while the logistic regression model achieved an accuracy of 90.79% with a 95% confidence interval of (0.9055, 0.9106). Below are some key differences in their performance based on other metrics –

- Sensitivity: XGBoost has a much higher sensitivity (53.55%) compared to logistic regression (11.96%). ***This implies that XGBoost is better at identifying patients with heart disease, even though it might also classify some healthy patients as having the disease (lower specificity).***
- Specificity: Logistic regression has a higher specificity (98.91%) compared to XGBoost (91.48%). ***This indicates logistic regression is better at correctly identifying patients without heart disease.***

Choosing the right model depends on our business objective. If we want to **detect heart disease early**, *XGBoost* might seem preferable because of its higher sensitivity. However, if we are looking to **identify patients without heart disease**, *logistic regression* seems to be a better choice. Additionally, both models have struggled with imbalanced data, as reflected by the lower prevalence of the positive class (heart disease) in the dataset (between 2% to 9%).

Interesting Findings

- **Age Impact:** The coefficient for Age is 0.253, which indicates that for each one-unit increase in Age, the log odds of having heart disease increases by about 2.53%. This suggests that as individuals age, they are more likely to develop heart disease
- **Stroke:** Based on coefficient of 0.99, individuals with a history of stroke have a substantially higher likelihood of heart disease
- **Gender:** The coefficient for Sex (0.73) suggests that males may have a higher probability of heart disease or attack compared to females, all else being equal.
- **BMI:** While the coefficient for BMI (0.0004) seems small, this indicates that higher BMI is associated with increased odds of heart disease or attack.
- **High Cholesterol:** The coefficient for HighChol (0.613) suggests that respondents with a higher cholesterol have a high probability of heart disease or attack keeping all independent variables constant.

Model Optimization (Oversampling and Parameter Tuning)

As we are looking to **identify patients who are at risk of heart disease**, we are choosing the **XGBoost** model to further explore techniques to improve the sensitivity of the XGBoost model while maintaining balance with specificity and accuracy. We used ROSE (Random Over-Sampling Examples) to handle class imbalance problems by oversampling the minority class (patients with heart disease) and potentially improve overall model performance for the minority class along with parameter turning.

Recall that only 9.42% of the respondents have indicated that they have experienced heart attacks or heart disease. Although our dataset has a large number of observations, it suffers from a class imbalance. This can force the model to prioritize learning from the more frequent negative class (no heart disease), which can lead to a situation where the model performs well overall on the majority class but struggles to accurately identify individuals with heart disease. For a doctor, it may be more important and crucial to be able to identify patients with a risk of heart disease earlier rather than predicting better that patient may not be at a risk of heart disease. We used oversampling techniques to address the imbalance. Oversampling techniques replicate instances from the minority class (heart disease) to create a more balanced dataset. It allows the model to learn from both classes effectively and potentially improve overall performance, especially sensitivity (true positive cases of heart disease). Fig. 10 shows the distribution of target variable – HeartDiseaseorAttack in original dataset and compares it with the balanced dataset. It shows how skewed the target variable was before oversampling.

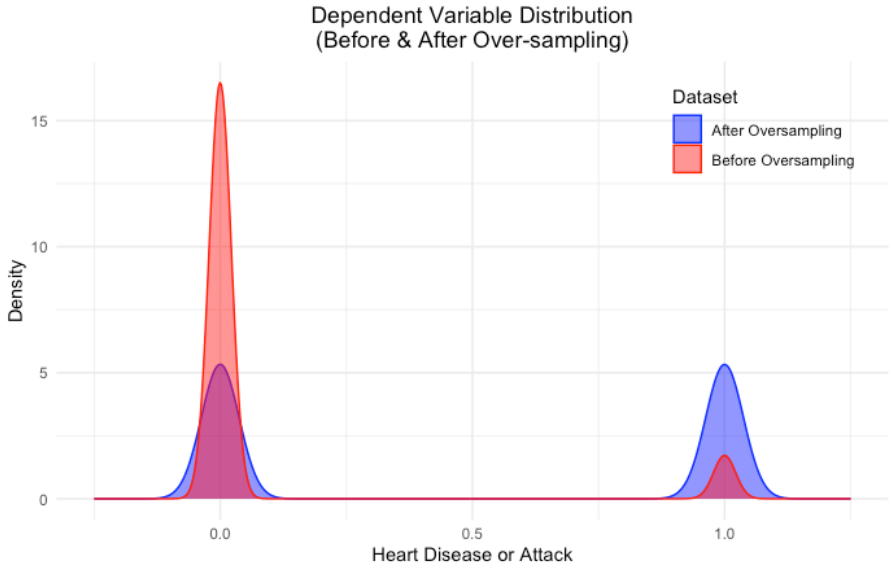

Fig. 10 Comparison: Before and After Oversampling

Note – The same source dataset, seed, threshold, and split ratio was used throughout the project while training and testing the model performance.

Key Result of Optimized Model

Below are some key results from the optimized model on the test dataset –

- **Accuracy:** 83.77% – Model successfully classifies around 84% of responses
- **Sensitivity (Recall):** 84.25% – Accurately identified individuals with heart disease (true positives)
- **Specificity:** 83.31% – Accurately identified individuals without heart disease (true negatives)
- **Balanced Accuracy:** 83.78% - Performed well in classifying both positive and negative cases
- **Positive Predictive Value (PPV):** Predicted individuals with heart disease with 83.06% accuracy
- **Negative Predictive Value (NPV):** Predicted individuals without heart disease with 85.01% accuracy

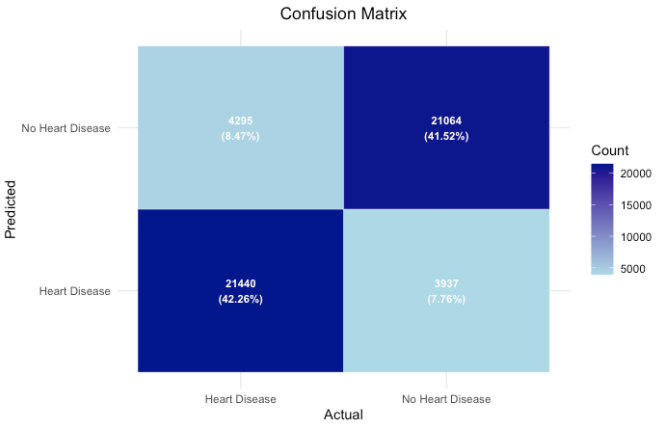

Fig. 11 Confusion Matrix Optimized XGBoost

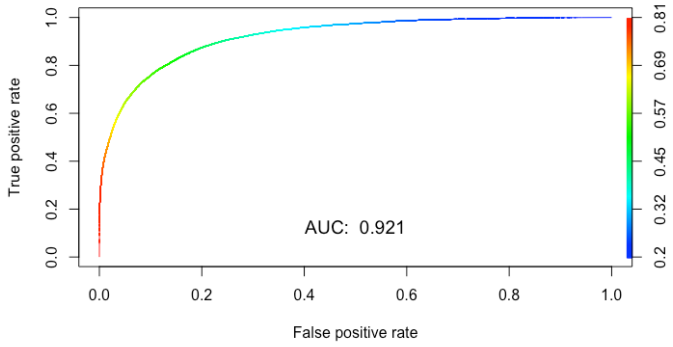

Fig. 12 ROC Curve Optimized XGBoost

Comparison of Optimized XGBoost Model

Table 2 shows comparison between the original and optimized XGBoost models. Below are the key improvements and changes in the tuned model –

- **Accuracy:** The original XGBoost model has a higher accuracy. It can be attributed to the class imbalance as it is better in identifying patients without heart disease. The accuracy of optimized model is slightly lower however it reflects a more balanced result
- **Sensitivity:** Optimized model has a significant improvement in sensitivity, which is critical for early detection or identifying patients who are likely to have heart disease
- **Specificity:** Both the models have similar specificity, but the optimized model is more balanced in terms of other metrics
- **PPV and NPV:** The new model has improvement in PPV and some reduction in NPV
- **Prevalence:** The new model has a significant improvement in prevalence from 1.99% to 49.28%, indicating the oversampling has helped us improve the data imbalance

Metric	XGBoost	Optimized XGBoost
Accuracy	90.72%	83.77%
Sensitivity	53.55%	84.25%
Specificity	91.48%	83.31%
Positive Predictive Value (PPV)	11.34%	83.06%
Negative Predictive Value (NPV)	98.87%	84.49%
Prevalence	1.99%	49.28%
Balanced Accuracy	72.58%	83.78%
AUC	84.09%	92.01%

Table 2: Performance - XGBoost vs XGBoost Tuned w/ 0.5 threshold

The optimized XGBoost model demonstrated a more robust performance for predicting heart disease. It is more effective in identifying individuals with heart disease while maintaining good accuracy (~84%) as compared to the original XGBoost or Logistic Regression models (~90%).

Conclusion

As noted, the optimized XGBoost model is more effective in identifying individuals with heart disease while maintaining good accuracy as compared to the original XGBoost or Logistic Regression models. The team recommends developing the XGBoost model further beyond this study.

We propose below recommendations to improve the model's capability –

- Exploring the addition of interaction terms and evaluating potential of including or excluding other variables (we have used top twelve variables based on Boruta). For example, an interaction terms between high blood pressure, age, cholesterol can offer more insight
- Improving the reliability/objectiveness of the source data. The current data source is based on a survey result and is subjective in nature
- Exploring other booster options such as gblinear and dart
- Optimizing booster parameters such as eta, gamma, and other hyperparameters can help improve the model's performance. Furthermore, identifying an optimal cut-off/threshold based on business objectives – *to prioritize detection of heart disease early or identifying patients without heart disease,* can help refine the model further
- Combining refined XGBoost and Logistic Regression models through blending can produce better predictions as compared to individual models

By pursuing the above avenues for improvements, we anticipate further advancements in the predictive accuracy.

Appendix

- GitHub Repo – GitHub Repo
- More figures available in GitHub Plots
- edX ID –
    - Daniel Smith – danielsmithche
    - Ankur Asthana – ankurasthana2k6
    - Jimmie Jain – jimmy987

Works Cited

Dane, S., & Centers for Disease Control and Prevention. (2017, August 24). Behavioral risk factor surveillance system. Kaggle. https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system

Klatsky A. L. (1999). Moderate drinking and reduced risk of heart disease. Alcohol research & health : the journal of the National Institute on Alcohol Abuse and Alcoholism, 23(1), 15–23 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6761693/

Teboul, A. (2022a, March 10). Heart disease health indicators dataset. Kaggle. https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset/data

Teboul, A. (2022b, March 10). Heart disease health indicators dataset notebook. Kaggle. https://www.kaggle.com/code/alexteboul/heart-disease-health-indicators-dataset-notebook/notebook

Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019, September 19). Building risk prediction models for type 2 diabetes using Machine Learning Techniques. Centers for Disease Control and Prevention. https://www.cdc.gov/pcd/issues/2019/19_0109.htm