Predicting a Heart Attack

Arpit Dang

Data Science Institute at Brown University

DATA 1030 Final Report

December 15th, 2024

GitHub Link: https://github.com/adang66/Data1030-Project

1,828 Words

Introduction

Heart attacks, a critical health concern, occur with alarming frequency in the United States. On average, someone has a heart attack every 40 seconds (Tsao et al., 2023). Approximately half of Americans have at least one of the three major risk factors for heart disease: high blood pressure, high cholesterol or smoking (Fryar et al., 2010). These statistics highlight the need for increased awareness, proactive prevention, and prompt medical intervention to combat this widespread health crisis.

The Centers for Disease Control and Prevention (CDC) have conducted annual Behavioral Risk Factor Surveillance System (BRFSS) surveys since 1984. BRFSS is a collaborative project between all the US states and US territories that collects data on health-related risk behaviours and chronic health conditions. This analysis focuses on the 2022 BRFSS survey (National Center for Chronic Disease Prevention and Health Promotion, 2023). The survey included 400,000+ American adults and collected various variables that may contribute to heart attacks. The survey provided valuable insights into the factors influencing heart attack risks.

The author removed the data points with missing values, leaving the dataset with 246,022 data points (Pytlak, 2023). The original survey by the CDC collected approximately 300 variables, narrowed down to 39 features deemed most relevant for heart attacks by the author. The target variable is 'HadHeartAttack', a binary outcome of either a 'Yes' or a 'No', framing this as a classification problem.

Numerous efforts have been made to develop machine learning models using this dataset in the discussion section on Kaggle, where the dataset was sourced from. However, these attempts have primarily been made by random individuals without established credibility in the field. Moreover, none of these attempts have been formally edited or published on any credible platform. However, many studies attempt to predict heart attack possibilities using different datasets. Tn et al. (2023) employed various machine learning algorithms on a separate dataset, and their findings suggested that chest pain (angina) and heart scans heavily correlate in predicting heart attacks.

Exploratory Data Analysis

The .describe() function was used to evaluate continuous features, while .value_counts() function was used for categorical and ordinal data. Interestingly, 23 out of the 39 features were binary categorical features, highlighting a limitation in the dataset. The predominance of binary features restricts the depth of information representation. Incorporating more ordinal features would've provided a more comprehensive data analysis with more significant variability.

Adams et al. (2020) conducted a study suggesting a strong association between higher BMI levels and an increased risk of heart attacks. In this analysis, comparing BMI between individuals who experienced a heart attack and those who did not also revealed a noticeable difference, with the median BMI being over 1 kg/m² higher among those who had a heart attack. Although not demonstrated in the Figure 1, the average BMI was also higher by 0.87 kg/m² in people who experienced a heart attack.

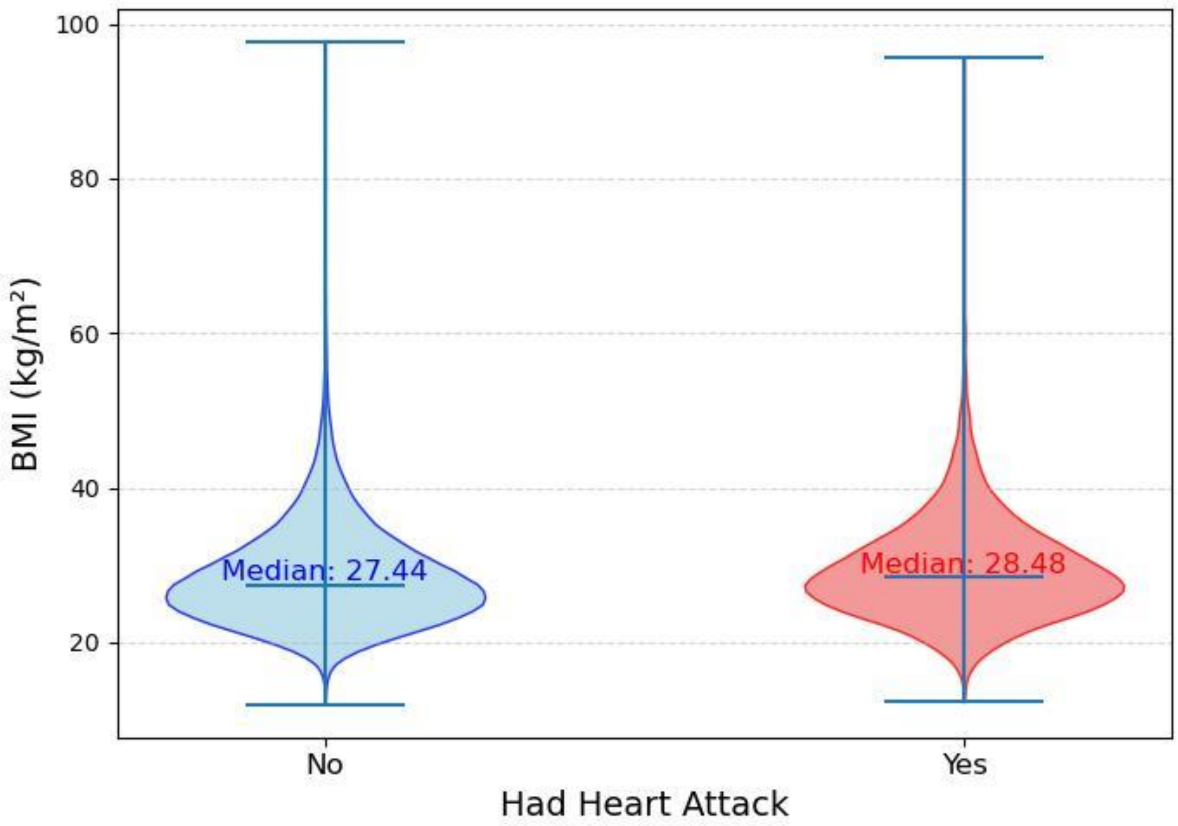

Figure 1. BMI Distribution by Heart Attack Occurrence

As seen in Figure 2, the dataset exhibits a notable lack of representation for most non-white racial groups, with 75.74% of participants identifying as white. Minority racial groups collectively account for less than a quarter of the participants. However, this distribution aligns with the demographic composition of the American population, where white individuals are the majority (United States Census Bureau, 2023). It is worth noting that heart attacks were most prevalent among the non-Hispanic multiracial population, with an occurrence rate of 6.5%, while the Hispanic population experienced the lowest rate at 4.0%.

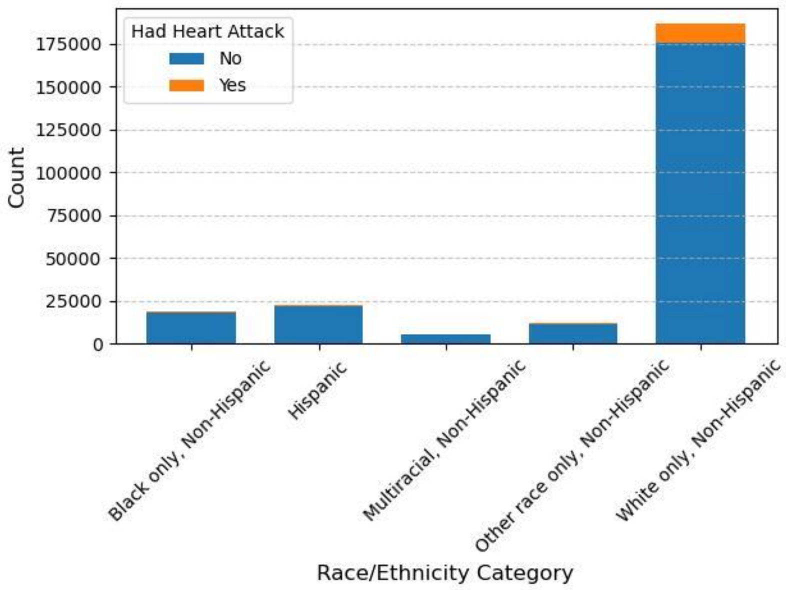

Figure 2. Heart Attack Occurrence by Ethnicity and Race

The survey showed that most adults surveyed were aged between 55 and 69, with heart attacks being more prevalent in older age groups. Age is likely to play a significant role in predicting heart attacks. However, the use of ‘AgeCategory’ as an ordinal feature is a limitation, as it could have been collected as a continuous variable for greater accuracy and detail.

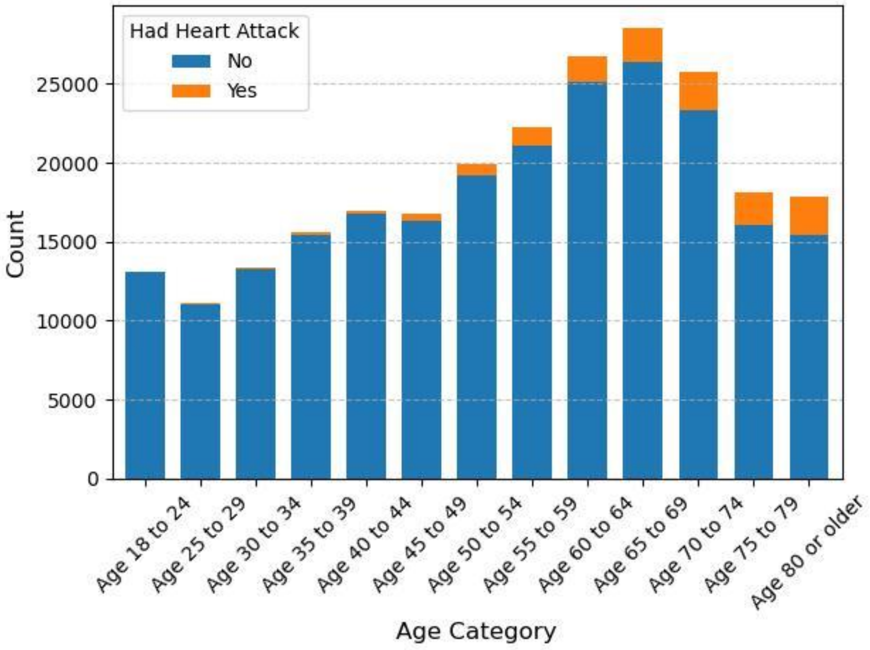

Figure 3. Comparison of Age Groups and Heart Attack Occurrence

The most notable characteristic of the dataset is a significant disparity in the distribution of the target variable, making it an imbalanced dataset as seen in Figure 4. Only 5.5% of data points are classified as ‘Yes’ (class 1) in ‘HadHeartAttack’, while most of the data points were in class 0 (‘No’ in ‘HadHeartAttack’).

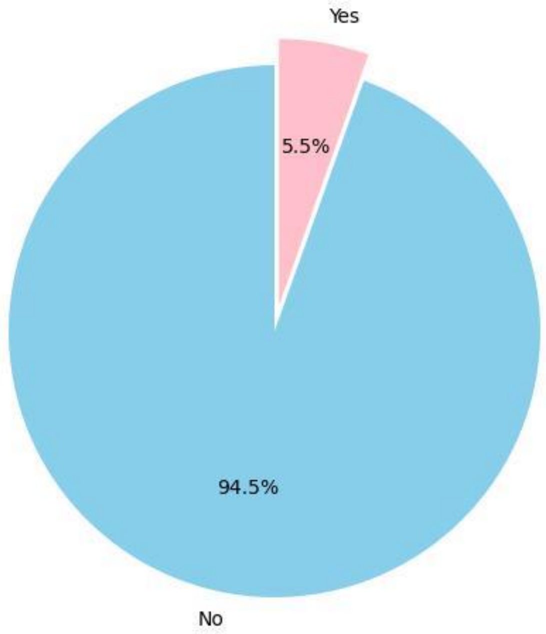

Figure 4. Distribution of the Target Variable ‘HadHeartAttack’ in the Dataset

The correlation matrix of continuous features illustrates the relationships between continuous variables. It reveals that none exhibit a strong correlation except for ‘WeightInKilograms’ and ‘BMI’. Despite their correlation of 0.86 (below the 0.9 threshold for strong correlation), both variables were retained in the dataset. In health-related analyses, it is important to preserve comprehensive information to capture all relevant aspects of an individual’s condition.

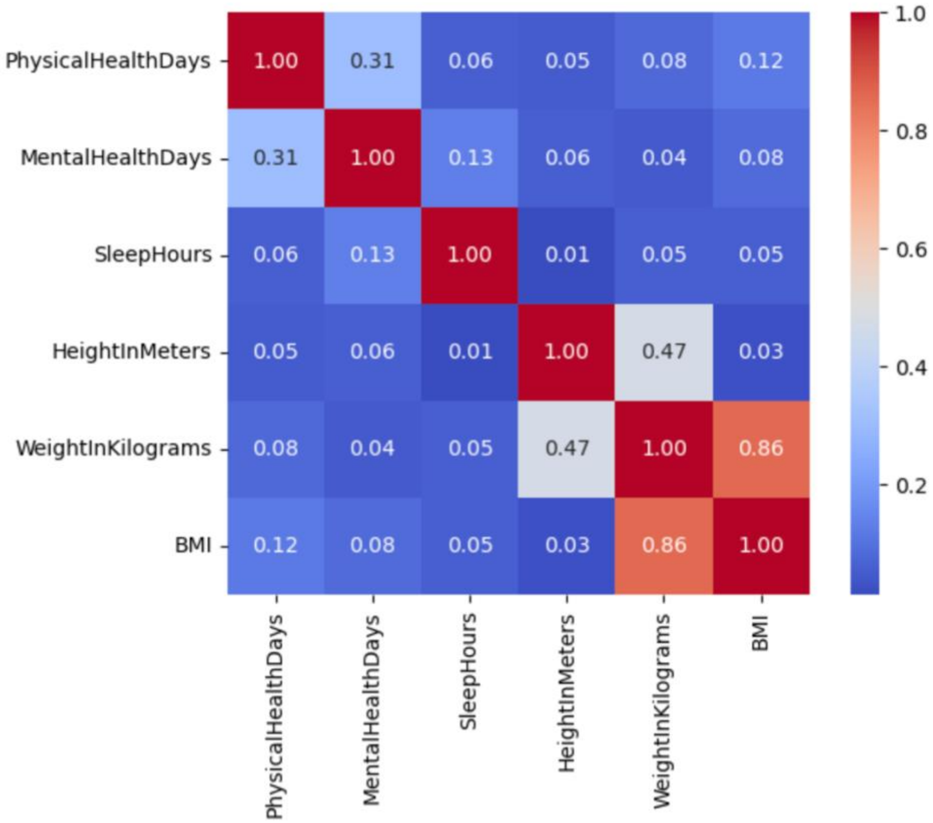

Figure 5. Heap Map of the Correlation Matrix of Continuous Features

Methods

The dataset was divided into a 60% training set, 20% validation set, and 20% test set. Due to the dataset's imbalance, with only 5.5% of data points belonging to class 1, a stratified split was employed to ensure proportional representation of classes across all subsets. Due to a large dataset with 246,022 points, a K-fold cross validation was not performed during splitting and training.

Table 1. Distribution of Heart Attack Cases across Training, Validation, and Test Sets

Type of Set	**Total Data Points**	**'No' to 'HadHeartAttack'**	**'Yes' to 'HadHeartAttack'**
Training Set	147 613	139 552	8 061
Validation Set	49 204	46 517	2 687
Test Set	49 205	46 518	2 687

A data preprocessing pipeline was developed to handle continuous, ordinal, and categorical features. For the six continuous features, a MinMaxScaler was applied, as the data fell within a defined range. OrdinalEncoder was applied to the six ordinal features after establishing their respective orders. Out of the 27 categorical features, 23 were binary with only two possible options. Initially, OneHotEncoder was used for preprocessing, but it unnecessarily created two columns for binary features where one column would suffice. To address this, OrdinalEncoder was used for binary features, encoding "No" as 0 and "Yes" as 1. As for the rest of the four categorical features, OneHotEncoder was used.

All data points were used since there were no missing values, and no features were excluded. The heatmap of variable correlations in Figure 5 shows that none exceeded the correlation coefficient threshold of 0.9.

Four machine learning algorithms (logistic regression, random forest classifier, support vector classifier, and XGBoost) were implemented to train the models with hyperparameters listed on table 2. The Support Vector Classifier was trained, validated, and tested on only 20% of the dataset due to the dataset's large size and limited computational resources.

Table 2. Hyperparameter Tuning and Class Balancing for Machine Learning Models

Machine Learning Algorithm	**Parameters**	**Balance Class Weights**
Logistic Regression	penalty: [l1, l2, elasticnet (l1_ratio = 0.5)] C: [0.001, 0.01, 0.1, 1, 10, 100]	class_weight = ‘balanced’
Random Forest Classifier	max_depth: [1, 3, 10, 30, 100, 300, None] max_features: [0.25, 0.5, 0.75, 1.0]	class_weight = ‘balanced’
Support Vector Classifier*	C: [0.01, 0.1, 1, 10, 100] gamma: [scale, auto]	class_weight = ‘balanced’
XGBoost	learning_rate: [0.001, 0.01, 0.1, 1, 10, 100] max_depth: [5, 10, 30, 100, 300, None]	scale_pos_weight = 17.31

*Trained, validated, and tested on only 20% of the dataset

The F2 score was chosen for this imbalanced dataset because it prioritizes recall over precision, making it well-suited for the problem at hand. Unlike other metrics, the F2 score does not consider true negative values, focusing instead on the trade-off between false negatives and false positives. Given that this is a medical diagnostic problem, where failing to predict a heart attack (false negative) is significantly more critical than incorrectly predicting one (false positive), the F2 score emphasizes recall to ensure fewer missed diagnoses. This makes it an ideal metric for such high-stakes scenarios. The baseline F2 score is 0.22, based on the assumption that all predicted points belong to class 1 for the 'HadHeartAttack' variable.

Each model was trained using 5 different random states, and the mean F2 score along with the standard deviation was calculated for evaluation. This approach eliminates any uncertainties caused by variations in the random state.

Results

Based on the mean F2 score of the models on the test set, the Logistic Regression performed the best at 0.512. All models had relatively similar F2 scores, but the logistic regression model was marginally better, as seen in Figure 6. All models were significantly better than the baseline F2 score of 0.22.

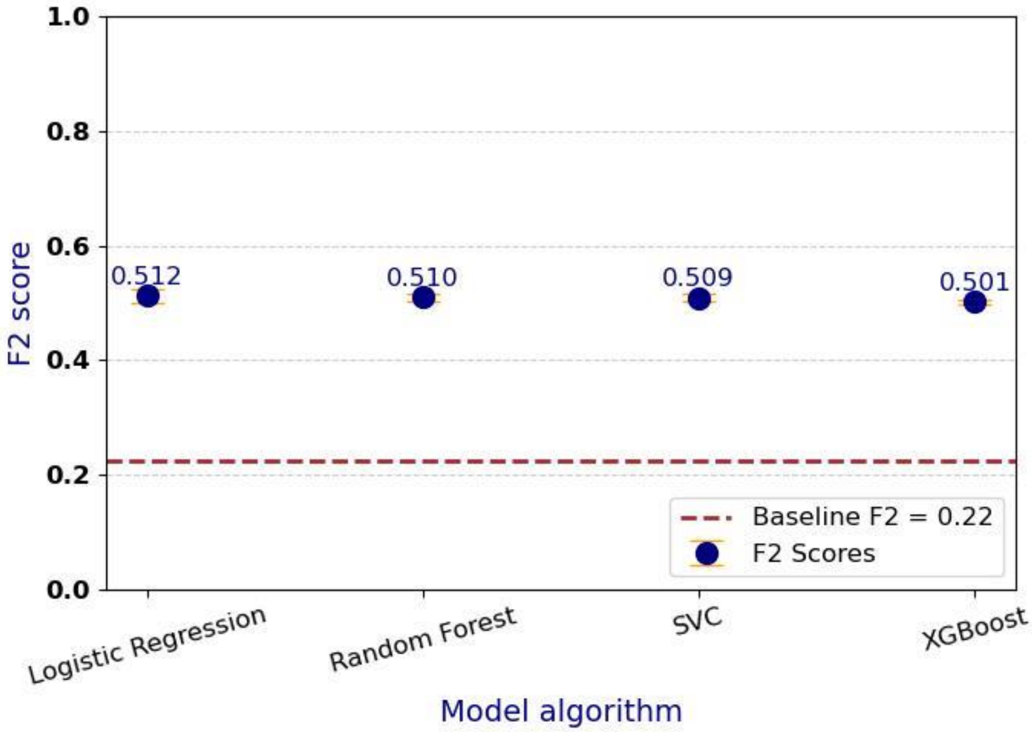

Figure 6. Comparison of F2 Scores Across Different Machine Learning Models

The logistic regression model achieved its highest F2 score of 0.531 using the following parameters: an elastic net penalty with an l1_ratio of 0.5 and a regularization parameter C = 10. The accuracy was 83.4%. Prioritizing a high recall score resulted in a significant drop in precision, as seen in the precision-recall curve shown in Figure 8. The F2 score was maximized when the recall is 0.768 with a precision of 0.214.

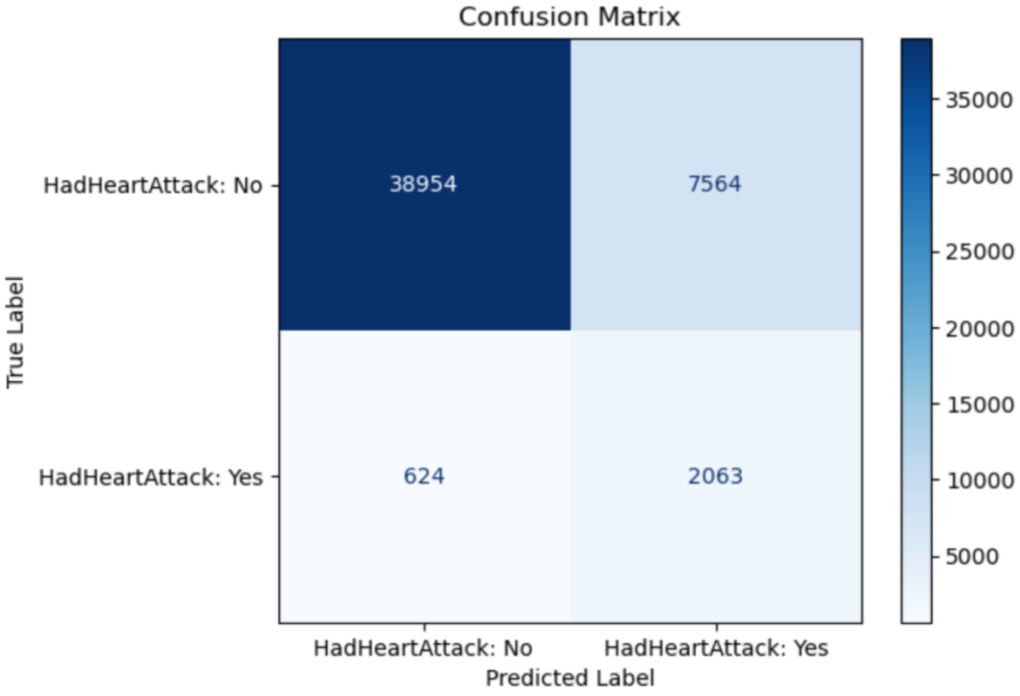

Figure 7. Confusion Matrix for the Best Performing Logistic Regression Model

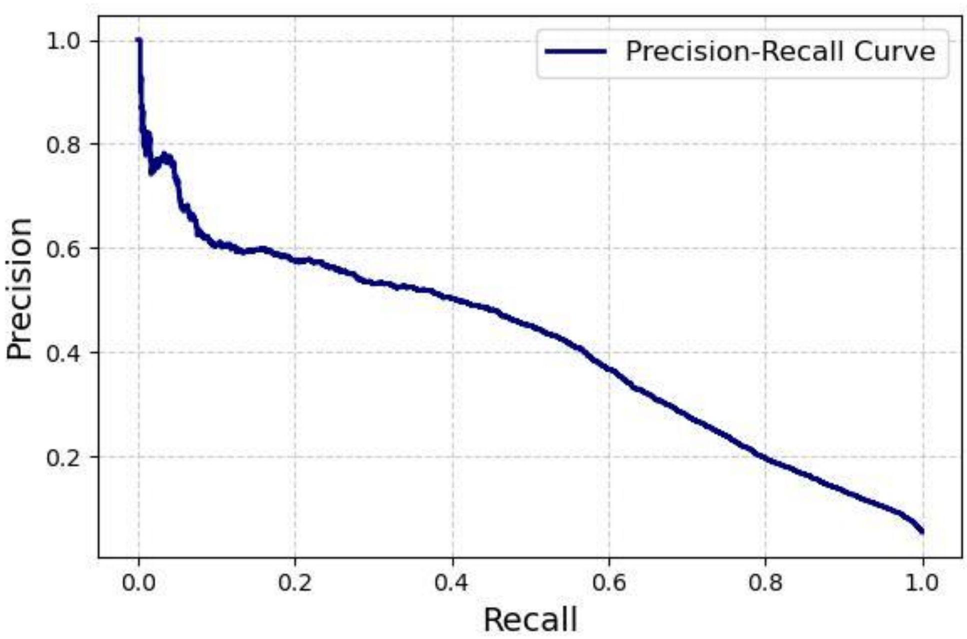

Figure 8. Precision-Recall Curve for the Best Performing Logistic Regression Model

Permutation importance, coefficient values, and SHAP values are used to identify the most influential features in the best-performing model. All features were scaled with a mean of 0, and a standard deviation of 1 when using coefficient values.

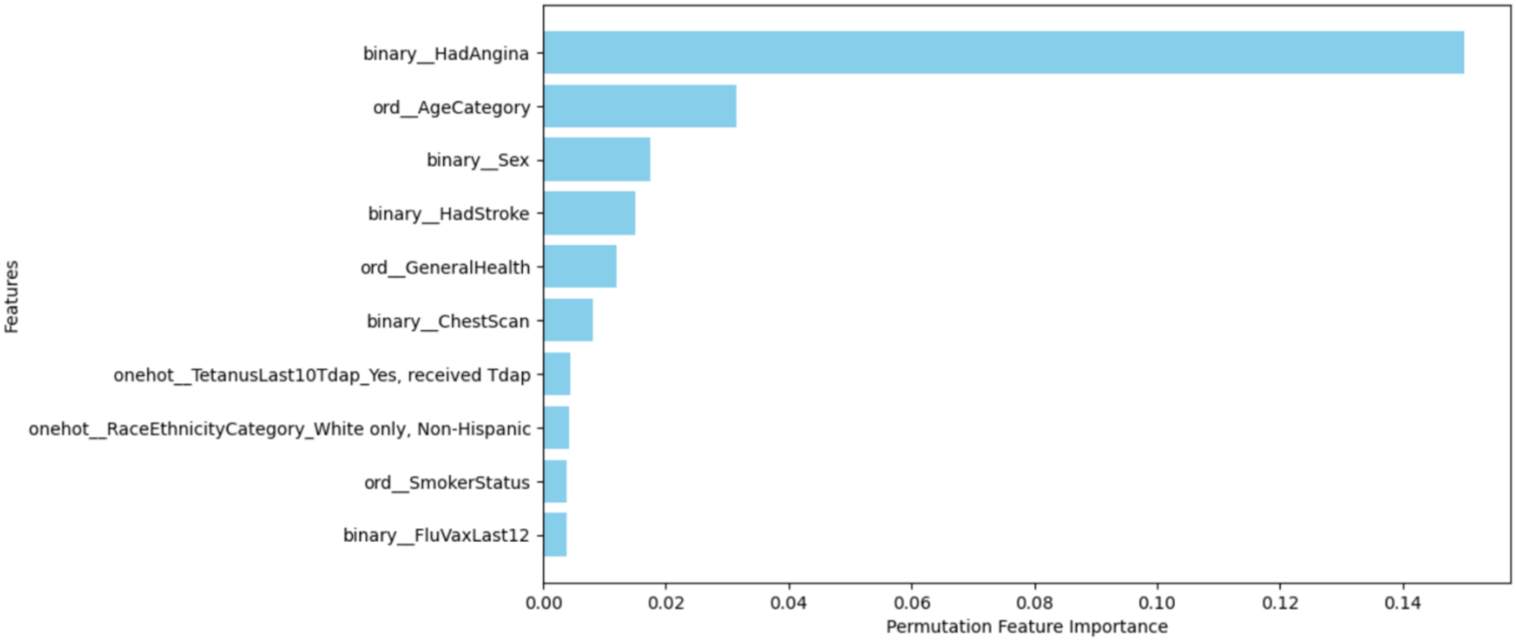

Figure 9. Top 10 Most Important Features according to Permutation Importance

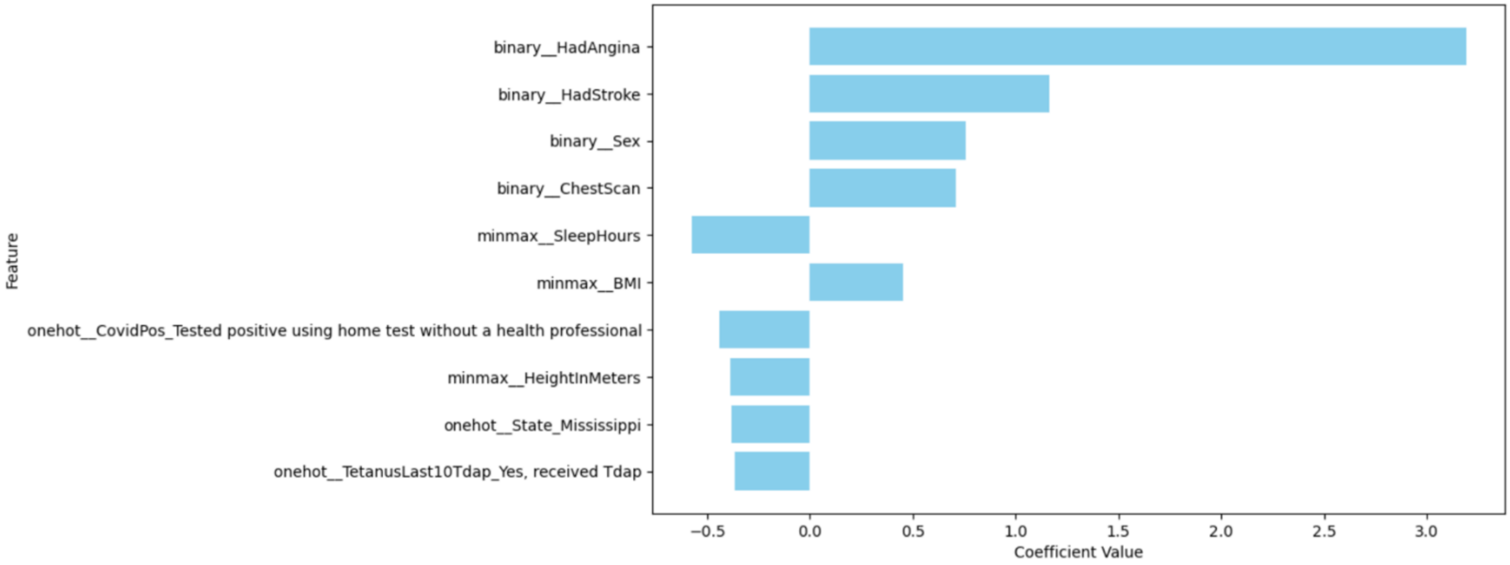

Figure 10. Top 10 Most Important Features according to scaled Coefficient Values

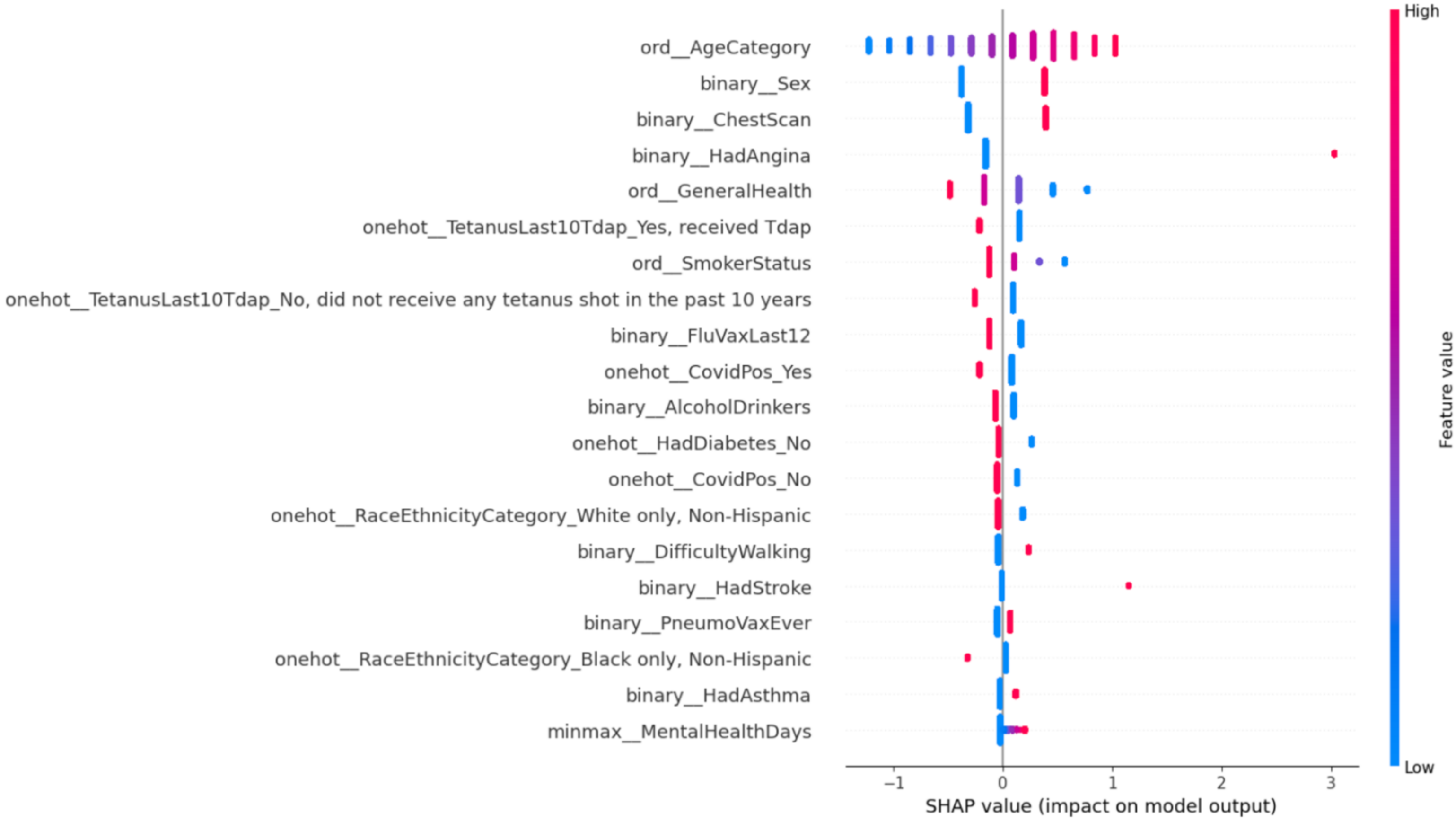

Figure 11. Feature Importance using SHAP Values

When analyzing the most important features using permutation importance, it is evident that ‘binary_HadAngina’ has the greatest impact on predicting heart attacks. If binary features had not been preprocessed using OrdinalEncoder, the top two features would have been ‘Onehot_HadAngina_Yes’ and ‘Onehot_HadAngina_No’. Using OrdinalEncoder for binary features prevented from creating unnecessary columns. This outcome is logical, as angina is a well-known early indicator of an increased risk of heart attacks (Mall, 2024). ‘binary_HadAngina’ is also the most influential feature when examining the coefficient values, and ranks among the top features based on SHAP values.

‘ord_AgeCategory’ emerges as one of the top features when evaluated through both permutation importance and SHAP values. This aligns with the prediction made during the exploratory data analysis, where it was anticipated that age would significantly influence heart attack predictions. As individuals age, their risk of experiencing a heart attack increases (Pallangyo et al., 2024).

‘binary_Sex’ is consistently identified as a key predictor of heart attacks across all three methods of feature importance analysis. The data showed that males experienced heart attacks more frequently than females. The observation was also validated when Man et al. (2020) studied sex as a biological variable in atherosclerosis. However, a significant limitation of the dataset was that the ‘binary_Sex’ variable only included two categories: male and female. Participants were required to select one of these options during the survey, with no provisions for non-binary or other genders.

Other notable features included ‘ord_GeneralHealth’, ‘binary_ChestScan’, and ‘ord_SmokerStatus.’

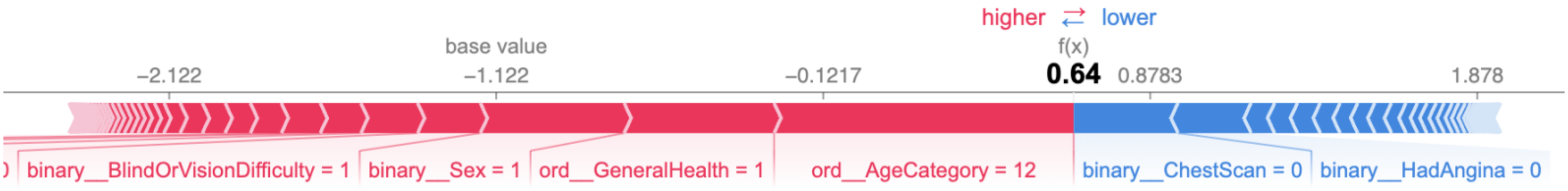

Figure 12. SHAP Force Plot: Contribution of Features to Model Prediction at Index 155

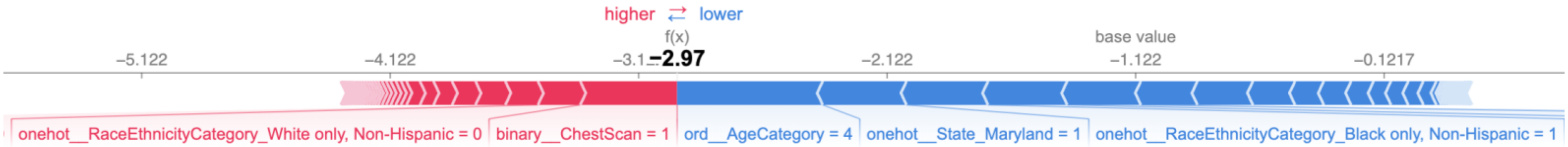

Figure 13. SHAP Force Plot: Contribution of Features to Model Prediction at Index 18,938

For a prediction of ‘Yes’ in ‘HadHeartAttack’, as illustrated in Figure 12, the outcome is influenced by factors such as older age, poor general health, and vision difficulties. These factors are consistent with commonly recognized health risks for heart attacks. Conversely, for index 18,938, the prediction of ‘No’ in ‘HadHeartAttack’ is primarily driven by demographic factors, including younger age, geographic location, and race, which reduce the likelihood of a heart attack.

Outlook

To improve the model's performance, we need to additionally try other machine learning algorithms like K-Nearest Neighbors (KNN), Naive Bayes, Bagging, and Boosting. These methods could help reduce false negatives and improve recall, which is critical in this case since there were 624 missed cases (1.27%) in the test set. To focus even more on catching positive cases, we could optimize for metrics like F3 or F5 scores, which heavily emphasize recall.

If more computational resources were available with more time, running an SVC on the entire dataset would be worth exploring. SVC is computationally expensive but can handle complex patterns, potentially improving recall and overall accuracy.

To better understand what drives the predictions, we could estimate feature importance using permutation methods or SHAP after dropping correlated variables. Pairwise permutations on two features at a time might also reveal key interactions. These steps would provide deeper insights into the model while helping fine-tune it for higher recall.

GitHub Link

https://github.com/adang66/Data1030-Project

References

Adams, B., Jacocks, L., & Guo, H. (2020). Higher BMI is linked to an increased risk of heart attacks in European adults: A Mendelian Randomisation study. *BMC Cardiovascular Disorders*, *20*(1). https://doi.org/10.1186/s12872-020-01542-w

Centers for Disease Control and Prevention. (2023, December 2). *CDC - 2022 BRFSS survey data and Documentation*. Centers for Disease Control and Prevention. https://www.cdc.gov/brfss/annual_data/annual_2022.html

Fryar, C. D., Chen, T.-C., & Li, X. (2012, August 1). *Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999-2010*. NCHS data brief. https://pubmed.ncbi.nlm.nih.gov/23101933/

Mall, S. (2024). Heart attack prediction using machine learning techniques. *2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 1778–1783. https://doi.org/10.1109/icacite60783.2024.10617300

Man, J. J., Beckman, J. A., & Jaffe, I. Z. (2020). Sex as a biological variable in atherosclerosis. *Circulation Research*, *126*(9), 1297–1319. https://doi.org/10.1161/circresaha.120.315930

Pallangyo, P., Mkojera, Z. S., Komba, M., Mfanga, L., Kamtoi, S., Mmari, J., Faraji, H. Y., Bhalia, S. V., Mayala, H. A., Matemu, G., Nkinda, A., Kifai, E., & Kisenge, P. R. (2024). Public knowledge of risk factors and warning signs of heart attack and stroke. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery*, *60*(1). https://doi.org/10.1186/s41983-023-00780-x

Pytlak, K. (2023, October 12). *Indicators of heart disease (2022 update)*. Kaggle. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Tn, K., P, S. C., S, M., Kodipalli, A., Rao, T., & Kamal, S. (2023). Prediction of early heart attack possibility using machine learning. *2023 2nd International Conference for Innovation in Technology (INOCON)*, 1–5. https://doi.org/10.1109/inocon57975.2023.10100993

Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Anderson, C. A. M., Arora, P., Avery, C. L., Baker-Smith, C. M., Beaton, A. Z., Boehme, A. K., Buxton, A. E., Commodore-Mensah, Y., Elkind, M. S. V., Evenson, K. R., Eze-Nliam, C., Fugar, S., Generoso, G., Heard, D. G., Hiremath, S., Ho, J. E., … Martin, S. S. (2023). Heart disease and stroke Statistics—2023 update: A report from the American Heart Association. *Circulation*, *147*(8). https://doi.org/10.1161/cir.0000000000001123

U.S. Census Bureau Quickfacts: United States. U.S. Census Bureau. (2023, July 1). https://www.census.gov/quickfacts/