Prediction of the risk of a Myocardial Infarction or Heart Attack using different Machine Learning

Mimi Li Sangyi Su Haotian Zheng

2023-12-13

Github Link

https://github.com/sangyisu/625-project/tree/main

Abstract

Heart attack is associated with both mortality and economic costs. Early and accurate prediction of heart attack is crucial in medical care. Our study uses five machine learning models to predict heart attack risks, analyzing and comparing their accuracy. The research aims to enhance early detection of heart issues, potentially reducing health and financial burdens. Additionally, it offers insights into handling missing values in large and complex datasets, a common challenge in big data analysis within healthcare.

Introduction

Cardiovascular diseases (CVDs) stand as the leading cause of death worldwide, significantly contributing to global disability. The prevalence of CVDs has shown a marked increase, with cases rising from 271 million in 1990 to 523 million in 2019. Concurrently, fatalities due to these diseases have escalated from 12.1 million in 1990 to 18.6 million in 2019, underscoring the growing impact of these conditions on global health (Roth et al. 2020).

Heart attacks (Myocardial Infarction, MI), a severe form of cardiovascular disease (CVD), significantly contribute to mortality and disease burden (Salari et al. 2023). Post-MI heart failure (HF) is especially serious, affecting 6 million people and causing 300,000 deaths annually in the U.S (Members WG 2012). The economic impact of MI is also considerable, with over 1.1 million hospitalizations and an estimated direct cost of $450 billion in 2010 in the U.S (Weintraub et al. 2011). Despite the increasing trend in MI risk factors, the timely and early detection of heart issues is crucial for reducing further damage to patients and saving lives (Muhammad et al. 2020).

Artificial intelligence (AI) research in healthcare is advancing rapidly, demonstrating its potential in a wide range of medical fields. In our study, five machine learning models were used to predict the risk of heart attack and then the accuracy of each model was compared. Our research also provides possible ideas for dealing with missing values in big data.

Data Preparation

Study population

Our study utilizes data from the 15th cohort of the Medicare Health Outcomes Survey (HOS) spanning 1998-2014 (United States Department of Health and Human Services. Centers for Medicare and Medicaid Services 2016), to examine the physical and mental health of Medicare beneficiaries under managed care. This cohort includes participants surveyed in 2012, with a follow-up in 2014. However, our analysis is exclusively based on the 2012 data due to the incompleteness of the 2014 follow-up. The chosen dataset encompasses

296,320 subjects and 88 variables, offering significant insights for a detailed analysis of heart attack risk factors within this population.

Missing value and imputation

The dataset contains a substantial volume of missing values, which can be attributed to its extensive size and the nature of the survey data. The dataset notably comprises approximately 3.37 million missing entries (13% of the total), posing a significant challenge for analysis. To refine the dataset for more effective analysis, variables with over 20% missing data were omitted, leading to a more manageable set of 296,320 entries across 79 variables (without the ID and cohort columns).

In the refined dataset, approximately 2.1 million missing values, constituting 10% of the data, were identified. To address this, the Multivariate Imputation by Chained Equations (mice) function, using the Predictive Mean Matching (pmm) method, was used for imputation. It imputes missing values based on a sequence of linear regressions. Considering the function's compatibility, the dataset was separated into two subsets based on their data types (numeric or character) prior to applying the imputation process.

Optimization

The large scale of our dataset resulted in a notably slow imputation process. This was further exacerbated by the need to reinitialize the 'mice' function for each iteration in RStudio. To address this, parallel processing was applied using the multi-core capabilities of the Great Lakes cluster. The core strategy involved the use of the 'foreach' function to simultaneously execute the imputation functions across multiple cores. Then the results from each processing unit were combined into a single, cohesive dataset. This approach improved the speed and overall effectiveness of our data imputation.

Methods

Outcome variable

In our study, the primary outcome variable is "A Myocardial Infarction or Heart Attack." This variable is derived from the question in the survey that inquires about the participants' history of heart attacks, enabling us to identify those with a previous occurrence of this cardiovascular event.

Feature selection

Upon inspecting the data structure, we found that most variables in the dataset are categorical, with only four numeric variables: "C15HDACT," "C15HDPHY," "C15HDMEN," and "C15PCTCMP." This imbalance raised concerns about overfitting in the following machine learning models. To address this, Chi-Squared Feature Selection was applied, utilizing pairwise chi-squared tests. The results led to the selection of 78 features with p-values below 0.05, excluding only "C15PAOTLK" (p=0.1869).

Machine learning models

The study used five machine learning models—logistic regression, support vector machine, naive Bayes, random forest, and decision tree—to predict heart attack risks (Bhat 2023). The predictive accuracies were assessed for each mode and then were used to determine the most effective model. The dataset was divided, allocating 70% for training and 30% for testing purposes.

Logistic Regression Logistic Regression is a statistical method for binary classification, which is efficient in modeling the probability of binary outcomes. It functions by applying a logistic function to a linear combination of input features. The underlying equation for logistic regression is:

$$\log\left(\frac{P(\text{heart attack}_i = 1|\mathbf{x}_i)}{1 - P(\text{heart attack}_i = 1|\mathbf{x}_i)}\right) = \mathbf{x}_i^T\beta$$

Support Vector Method SVM is a powerful algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates different classes in the feature space. It aims to maximize the margin between classes. For this model, the training data comprises a randomly selected subset of 10,000 observations from the large training dataset, while the test data consists of 4,200 observations from the large testing dataset. In the case of support vector machines, the cost value is carefully adjusted through tuning, and the best-performing cost value is found to be 15 (Figure 1).

Naive Bayes Naive Bayes, a probabilistic classifier grounded in Bayes' theorem, operates under the assumption that features are independent. This model, despite its simplicity, is effective in text classification and other tasks. For our study, we fine-tuned the Laplace parameter of Naive Bayes and found that a value of 0 was optimal (Figure 2). This adjustment enhanced the accuracy of our results.

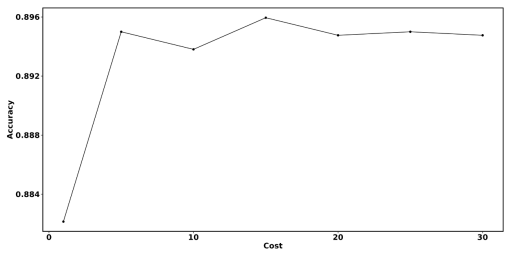

Figure 1: Accuracy vs Cost for SVM

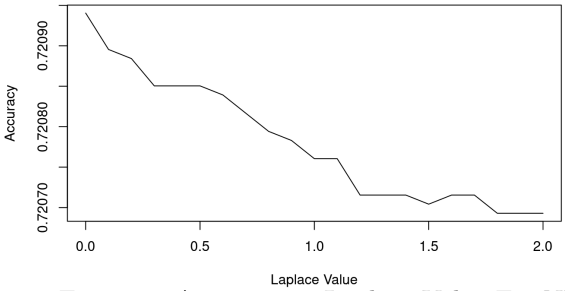

Figure 2: Accuracy vs Laplace Value For NB

Random Forest Random Forest, an ensemble learning method, combines multiple decision trees to enhance prediction accuracy and reduce overfitting. It trains each tree on a distinct random subset of the data and features. In our study, the 'ntree' (number of trees) and 'mtry' (number of features tried at each split) parameters were adjusted meticulously within the Random Forest model. The optimal configuration is ntree = 300 and mtry =8.831761.

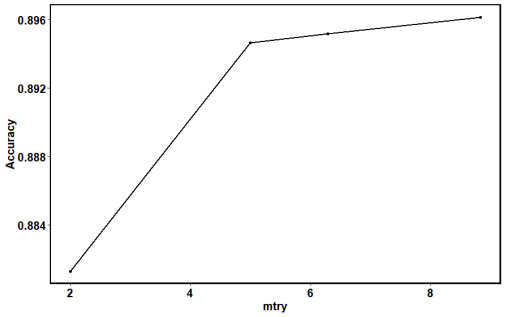

Figure 3: Accuracy vs ntree with ntree = 300

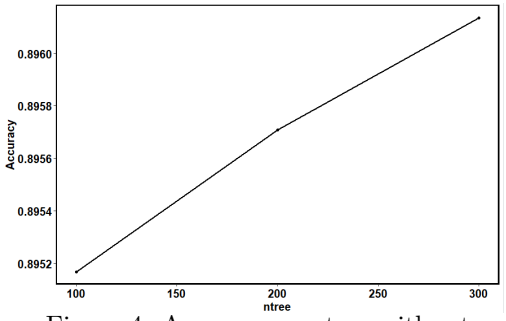

Figure 4: Accuracy vs ntree with mtry = 8.831761

Decision Tree The Decision Tree model operates like a tree structure, making decisions at each node based on a feature, leading to further nodes or outcomes (leaves). This model uses recursive dataset splitting for decision-making, known for its interpretability and visual clarity. In our study, the decision tree's optimal size was determined through cross-validation. This process entailed pruning the tree to its most effective dimensions, thereby enhancing predictive accuracy. The model was applied to the test set to evaluate its effectiveness.

Results

Post-optimization, the models' accuracy was evaluated using diverse machine learning techniques. As shown in table 1, the logistic regression model achieved an accuracy of 0.8967, while the SVM model reached 0.8960.

The Naive Bayes model attained 0.7209, the random forest model recorded 0.8961, and the decision tree model yielded an accuracy of 0.8764.

Comparison of Different Models	
Model	Accuracy
Logistic Regression	0.8967
Support Vector Method (cost=10)	0.8960
Naive Bayes	0.7209
Random Forest	0.8961
Decision Tree(tree size = 3)	0.8764

Table 1: Accuracy Comparison of Different Models

Discussion and Conclusions

The logistic regression model exhibited the highest accuracy in our results. However, it's important to consider that the relationship between the outcome and variables might not always be linear. This aspect should be carefully evaluated when predicting heart attacks, as it may influence the model's predictive effectiveness. Therefore, in the context of heart attack prediction, the random forest model with the second-highest accuracy should also be considered. It is also important to note that svm model was evaluated using a dataset of only 10,000 samples due to the inability of the complete dataset to fit within the model. This smaller sample size raises concerns about SVM's accuracy to larger datasets, so the comparability of SVM's accuracy with that of other models tested on more extensive datasets might be limited. A future research direction could focus on enhancing the SVM model's efficiency for handling larger datasets, thereby broadening its applicability and improving its predictive reliability in large-scale analyses. Additionally, considering the substantial number of features utilized in our analysis, overfitting might be another concern for the models.

Moreover, our study is based on the assumption that individuals who have heart attack history are categorized as high-risk for future heart attacks. However, this classification may not fully capture the complexity of risk factors. The current health conditions of these individuals could be either contributing factors to their initial heart attack or consequences of it. Without longitudinal data detailing the health trajectory of these individuals before and after the heart attack, it's challenging to differentiate between pre-existing risk factors and health issues that emerged as a result of the heart attack. This limitation is crucial as it impacts the accuracy and reliability of our risk prediction models, highlighting the need for future research to focus on datasets with a longitudinal span.

Another advantage of using long-term data is that it allows for a more comprehensive assessment of individuals who may be at high risk of heart attack but have no prior history of it. For instance, individuals without a history of heart disease but identified as high-risk would currently be classified as false predictions, reducing the accuracy of the model. However, if follow-up data show these individuals later developing heart disease, they should be included as true predictions.

Missing data and Optimization

The functionality of the 'mice' function is limited to handling datasets with a single data type. It is possible that 'mice' encodes categorical variables for the purpose of imputation and the critical issue arises in its inability to decode this information back successfully for imputation when dealing with varying data types. In cases where a dataset includes various types, like numeric and character, 'mice' will first impute missing values for one type (which is numeric in our study). In the second run, it will then address the other type. However, the exact mechanism of the 'mice' function in its two separate runs still remains unclear.

Given the large size of our dataset and the considerable amount of missing data, running 'mice' twice is extremely time-consuming and causes memory overflowing. To efficiently and successfully use 'mice', we divided our dataset into two parts based on the data types of the variables. Our method, however, potentially impacts the holistic understanding of the data's interrelationships. While it addresses the issue of time

consumption, it poses a new challenge of ensuring that the separation does not decrease the integrity of the overall imputation, especially considering the interaction between the two types of variables.

To evaluate the time efficiency of our methods, a function was created, revealing that the total execution time exceeded 6 hours. With the application of parallel processing, this duration significantly decreased (execution time=1.92hr), greatly enhancing the efficiency of the 'mice' function (about 70%). Additionally, another function was developed for imputation, which involved randomly filling missing values in each column based on the frequency of observations. While this method was faster, the 'mice' function, despite being more time-intensive, offers much greater plausibility and interpretability for missing value imputation.

Significance and Future Directions

While our study contends with the challenge of the nature of the dataset, we rigorously compare the accuracy of various machine learning models in predicting heart attack risks using our dataset. This comparison not only highlights the strengths and limitations of each model but also advances the understanding of machine learning's applicability in complex medical predictions. Despite the data's cross-sectional nature, our findings offer valuable insights into the predictive capabilities of these models, paving the way for future research that could integrate longitudinal data for more nuanced risk assessments.

Moreover, our study offers insights into efficient and effective data imputation techniques for large, mixed-type datasets, giving an idea of addressing the 'mice' function's limitations and time constraints. This contributes to more practical data handling methods. Future research should focus on enhancing the integrity of imputed data, with a particular emphasis on preserving the relationships between various types of variables. This direction is important for ensuring comprehensive and accurate data analysis in large-scale research scenarios.

Contribution

All members participat in all steps and contributed to the project

Missing data: Haotian Zheng

ML models: Sangyi Su

Report: Mimi Li

References

Bhat, Naresh. 2023. "Heart Attack Prediction Using Different ML Models." https://www.kaggle.com/code/nareshbhat/heart-attack-prediction-using-different-ml-models.

Members WG, et al. 2012. "Heart Disease and Stroke Statistics—2012 Update: A Report from the American Heart Association." *Circulation* 125 (1): e2–220.

Muhammad, Y., M. Tahir, M. Hayat, et al. 2020. "Early and Accurate Detection and Diagnosis of Heart Disease Using Intelligent Computational Model." *Scientific Reports* 10: 19747. https://doi.org/10.1038/s41598-020-76635-9.

Roth, GA, GA Mensah, CO Johnson, et al. 2020. "Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update from the GBD 2019 Study." *J Am Coll Cardiol* 76 (25): 2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010.

Salari, N., F. Morddarvanjoghi, A. Abdolmaleki, et al. 2023. "The Global Prevalence of Myocardial Infarction: A Systematic Review and Meta-Analysis." *BMC Cardiovascular Disorders* 23 (206). https://doi.org/10.1186/s12872-023-03231-w.

United States Department of Health and Human Services. Centers for Medicare and Medicaid Services. 2016. "Medicare Health Outcomes Survey (HOS), 1998-2014." Inter-university Consortium for Political; Social Research [distributor]. https://doi.org/10.3886/ICPSR23380.v3.

Weintraub, WS et al. 2011. "Value of Primordial and Primary Prevention for Cardiovascular Disease: A Policy Statement from the American Heart Association." *Circulation* 124 (8): 967–90.