The Application of Machine Learning on Risk Classification of Heart Disease

Xiaoyu Lin Kyoka Ono

GitHub Link: https://github.com/kyokaono8811/biostat625finalproject.git

Contributions:

- Xiaoyu Lin — "Data description, Data cleaning, Write report"
- Kyoka Ono — "Model Training, Evaluation Metrics, Write report"

Abstract

Heart disease is a leading cause of death for adults in the U.S, and early detection of key risk factors is essential for prevention. In this study, we apply multiple machine learning methods, including logistic regression, random forest, and GAM to identify the significant risk factors of heart disease. We then evaluate model performance and compare variable importance across methods using evaluation Metrics.

Introduction

Heart disease affects millions of individuals in the U.S, making early identification of risk factors a high public-health priority. Traditional statistical approaches have identified several predictors, but machine learning methods can capture more complex, nonlinear relationships.

This project applies several different machine learning algorithms to the `heart_2022_no_nans.csv` dataset from the CDC Behavioral Risk Factor Surveillance System (BRFSS) on Kaggle. Our goal is to determine which model outputs the highest evaluation metric for predicting heart attack, and which variables significantly associated with having had a heart attack?

Methods

Data Source

The dataset contains 40 variables for over 200,000 survey participants (all complete cases).

```
Number of participants: 246022
```

```
Number of variables: 40
```

Preprocessing

Variable Selection

We choose 10 significant covariates out of 40 for our models based on literature review and use `HadHeartAttack` as the predicted variable

Outcome Variable:

- `HadHeartAttack`: Binary indicator (Yes/No) of whether a doctor diagnosed the respondent with a heart attack.

Predictor Variables:

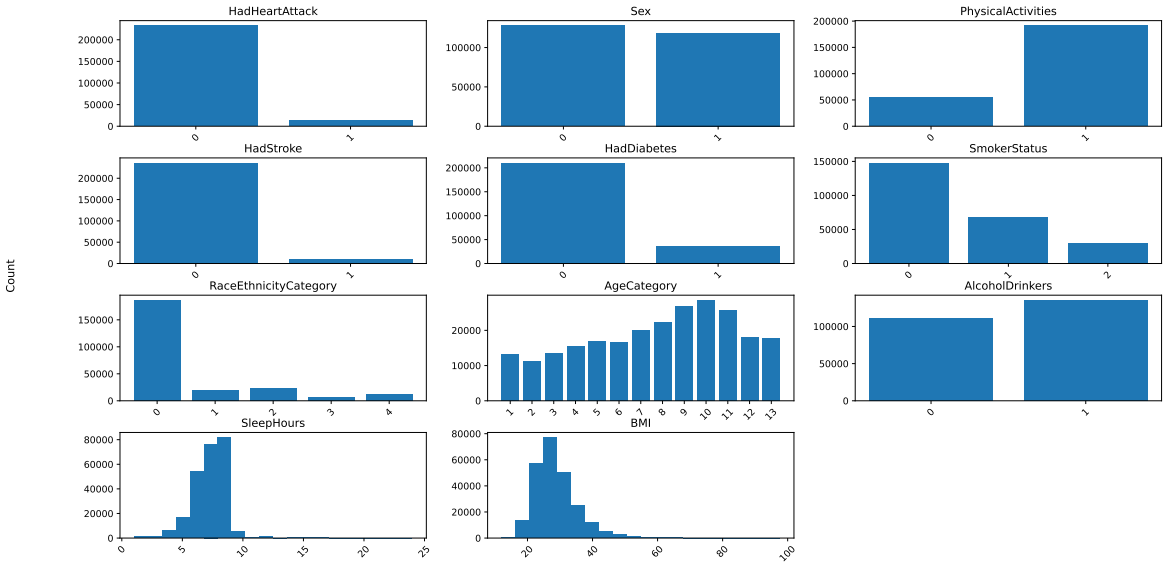
- `Sex`: Biological sex of the participant (Male/Female).
- `PhysicalActivities`: Whether the participant engaged in physical activities in the past month (Yes/No).
- `SleepHours`: Average number of hours of sleep per night (numeric).
- `HadStroke`: Whether had a stroke (Yes/No).
- `HadDiabetes`: Whether had a diabetes (Yes/No/Yes, but only during pregnancy (female)/No, pre-diabetes or borderline diabetes).
- `SmokerStatus`: Smoking status of the participant (Former smoker/Never smoked/Current smoker – now smokes every day/Current smoker – now smokes some days/No).
- `RaceEthnicityCategory`: (White only, Non-Hispanic/Black only, Non-Hispanic/Other race only, Non-Hispanic/Multiracial, Non-Hispanic/Hispanic)
- `AgeCategory`: Age group (18–24, 25–29, . . . , 80+).
- `BMI`: Body mass index (numeric).
- `AlcoholDrinkers`: Whether a participant is a heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per (Yes/No).

Data Cleaning

We would like to create dummy variables for categorical variables and merge equivalent levels.

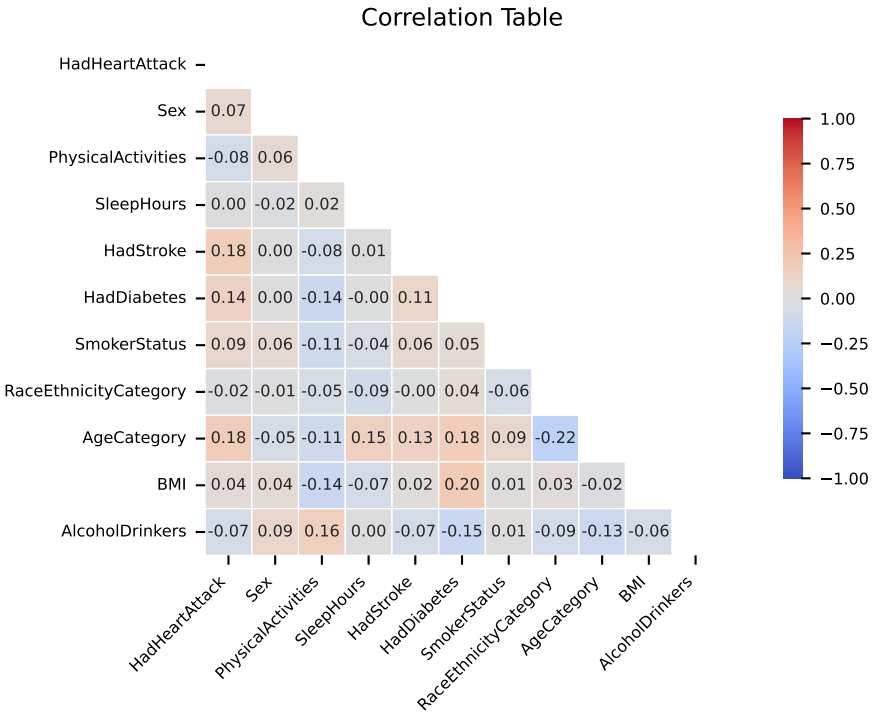
Data Description

Let's look at the distribution of the variables



From the graph above, we can see that some categorical variables such as `HadHeartAttack`, `HadStroke`, and `RaceEthnicityCategory` exhibit noticeable class imbalance. Other categorical variables are more balanced. Numerical variables like `SleepHours`, and `BMI` show skewed distributions.

Now Let us build a correlation table. We use spearman correlation because the dataset contains a mixture of categorical and numerical variables. Spearman does not assume linearity or normality, making it a better measure of association than Pearson for this dataset.



All pairwise correlations were relatively weak, indicating that no individual predictor shows a strong linear association with heart attack or with each other. This is expected in multi-factor health datasets, where the outcome is influenced by many small effects rather than a single dominant variable.

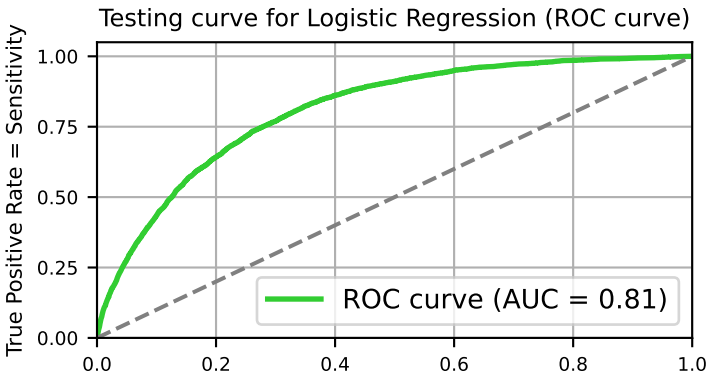
Low pairwise correlations do not imply weak predictive power for nonlinear effects, interactions, and combined contributions can still provide meaningful classification performance in multivariate models.

Models Applied

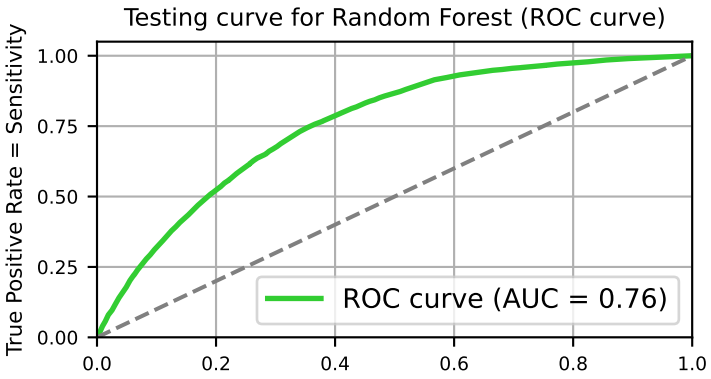
Highly imbalanced classes may affect model performance if not addressed, so we will use the method of undersampling or class weights during model training. And to improve model performance, the skewed numerical variables should be standardized or centered, especially for algorithms sensitive to scale (e.g. logistic regression).

Evaluation Metrics

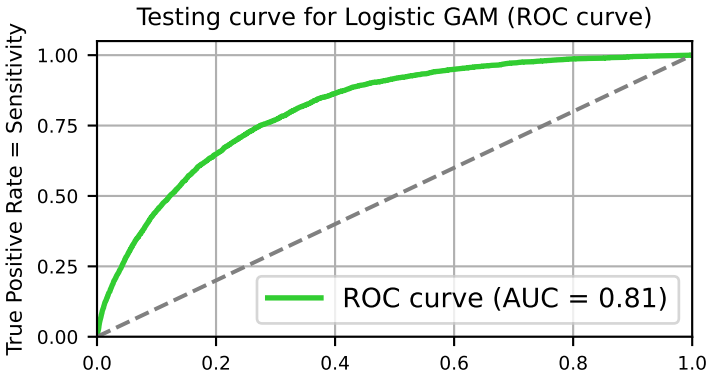
1. Logistic Regression



2. Random Forest



3. GAM



Results

	Model	Accuracy	Precision	Recall	F1	AUROC
0	Logistic Regression	0.713022	0.131658	0.760357	0.224452	0.809221
1	Random Forest	0.674069	0.111783	0.715207	0.193347	0.756188
2	Logistic GAM	0.706369	0.129873	0.767800	0.222166	0.812238

The logistic regression has an accuracy of 71%, precision of 13 %, recall of 76%, and F1 score of 22%. The AUROC is 81%. It took 0.5 seconds to run the model.

The random forest has an accuracy of 67%, precision of 11%, recall of 72%, and F1 score of 19%. The AUROC is 76%. The model training time was 1 second.

The Logistic Generative Additive Model has an accuracy of 71%, precision of 13%, recall of 77%, and F1 score of 22%. The AUROC is 81%. The time it took to train the model was 1 second.

Conclusion

Logistic Regression has the highest accuracy metric, indicating that it has the highest correction rate for evaluating patients who both did and did not have a heart attack history. Random Forest shows the highest precision score, meaning it has the highest correction rate for evaluating patients who had a heart attack in real life out of the patients who were predicted to have heart attack. However, all of these models have significantly low precision scores, making them unreliable for precision evaluation. This result is anticipated because the heart attack has already happened in the past, and the variables are collected to represent association with heart attack history and not intended for prediction purposes. Logistic Regression has the highest F1 score, and it implies that it may be the optimal model for evaluating heart attack prevalence among the three models because F1 score takes the balance between precision and recall into account. However, all of the F1 scores are low, due to the significant low values in precision.

Future Work

Currently, most of the models use the default hyperparameters from the scikit-learn package; therefore, we can conduct hyperparameter tuning for model enhancement using methods such Grid Search and Random Search. In order to retrieve results with higher F1 scores, the state of the art (SOTA) models such as RNN and XgBoost may be deployed. However, these models are computationally expensive, and hyperparameter tuning for these models is time consuming and requires parallel computing tools like GPUs. Additionally,

there is imbalance in predictor variables such as a history of diabetes patients, their smoking status, and race/ethnicity category. Therefore, this may lead the models to cause representation bias where minority groups are underrepresented while the model is going through the training phase.

References

- Pytlak, K. (n.d.). Personal key indicators of heart disease [Data set]. Kaggle. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data
- Centers for Disease Control and Prevention. (2024, December 2). Heart disease risk factors. https://www.cdc.gov/heart-disease/risk-factors/?CDC_AAref_Val=https://www.cdc.gov/heartdisease/risk_factors.htm
- Al-Zaiti, S.S., Martin-Gill, C., Zègre-Hemsey, J.K. et al. Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. Nat Med 29, 1804–1813 (2023). https://doi.org/10.1038/s41591-023-02396-3