Project Report for Predicting Heart Attack Based on Clinical Data

Team Name: Stupid Birds
Team Members: Zijie Chen, Yuhui Wang, Yangyijia Zhang, Zehan Hu

Apr 30 2024

Team Name: Stupid Birds

Team Members: Zijie Chen, Yuhui Wang, Yangyijia Zhang, Zehan Hu

Background

According to the Centers for Disease Control and Prevention (CDC), every 33 seconds, one person in the United States dies. About 1 of 5 deaths in the United States is because of heart disease. Each year, the government spends a lot of money and time researching heart disease. Our project focuses on the early detection of Heart Attack, also called a myocardial infarction, which happens when a part of the heart muscle doesn't get enough blood.

Clinical questions

According to CDC, diabetes, obesity, diet, disability, and alcohol consumption will influence the risk of getting heart disease. However, CDC did not state which factor will make more effects. In this project, the first question needed to be answered is how would these factors cause heart attack. Since the dataset includes several variables related to these factors such as BMI, HadDiabetes, and DifficultyWalking, the project will explore these factors in more detail. Besides factors stated by CDC, the paper of A.Judson Wells shows passive smoking is a cause of heart disease (1994). With the smoking status in the dataset, the project can add smoking level as a variable in the research. The project may answer more questions during the research process.

Exploratory Data Analysis

1.Data description

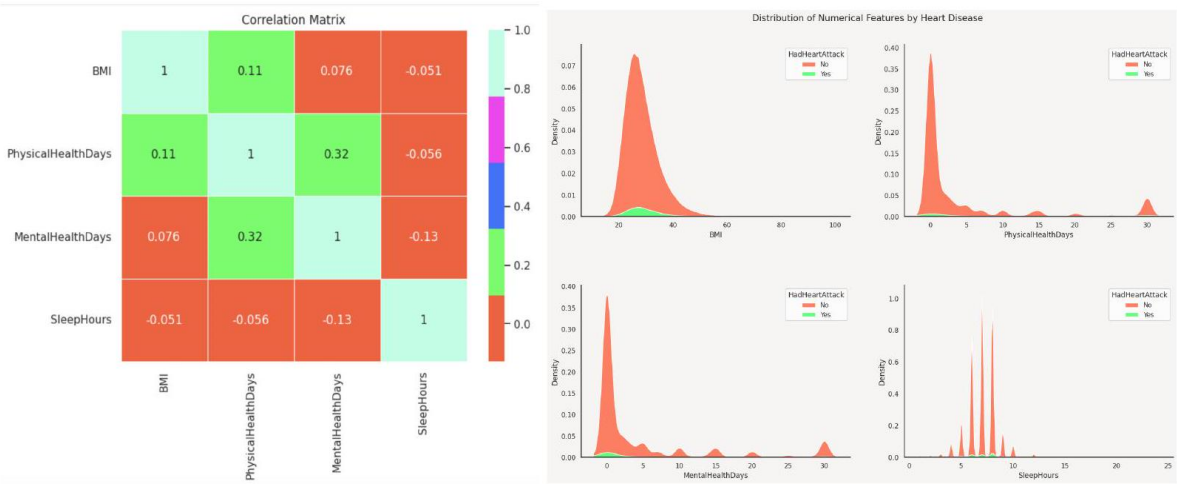
This dataset originates from the CDC and plays a significant role in the Behavioral Risk Factor Surveillance System (BRFSS). This system annually carries out phone surveys to gather information regarding the health conditions of residents in the U.S. The dataset contains 445132 lines of data and 18 variables including one outcome variable. HadHeartAttack is the outcome variable with 'Yes' indicating the patient has a heart attack and 'No' indicating the patient doesn't have a heart attack.

Among these 18 predictors, there are 4 continuous variables and 14 categorical variables. Continuous variables including PhysicalHealthDays, BMI, MentalHealthDays and Sleep hours. The average BMI among respondents is approximately 28.53, with a standard deviation of about 6.55, indicating a moderate variability in body weight relative to height. On average, respondents report about 4.35 days with a standard deviation of 8.69. Similarly, the mean number of days reported in mental health days is 4.38, with a standard deviation of 8.38. The range is the same for both of them, from 0 - 30. For the sleeping hours, the average sleeping number is roughly 7.02, with a relatively low standard deviation of 1.50, suggesting that most people's sleep durations cluster around the mean.

Below is a correlation matrix generated to demonstrate the 4 continuous predictors, PhysicalHealthDays, BMI, MentalHealthDays, and Sleep hours. If the correlation index between columns is too high, it means that the two variables are highly correlated and one of them shall be removed. On the graph, Physical Health Days and Mental Health Days show a moderate positive correlation of 0.32 and have the highest correlation. This indicates that individuals who report more days of poor physical health also tend to
report more days of poor mental health. BMI is slightly negatively correlated to sleeping hours and mental health days, physical health days are slightly negatively correlated to sleeping hours, and sleeping hours are slightly negatively correlated. Overall, most of them are very low connected, meaning most are independent of each other.

The provided density plots depict the distribution of four key variables-BMI, PhysicalHealthDays, MentalHealthDays, and SleepHours- segmented by individuals who have and have not experienced a heart attack. These plots offer critical insights into potential risk factors associated with heart attacks.

These plots emphasize the significance of lifestyle and health factors in the context of heart disease. They suggest that interventions aimed at improving lifestyle habits such as achieving a healthy weight, managing health conditions effectively, and ensuring adequate sleep could be crucial in reducing heart attack risk. Additionally, these visualizations underscore the need for continuous monitoring and analysis of these factors in both clinical and research settings to better understand and mitigate heart attack risks.



Besides numerical variables, there are 13 categorical variables, "Age", "Race", "General Health", "Physical Activities", "Smoker Status", "HadSkinCancer", "HadStroke", "Sex", "DifficultyWalking", "Alcohol Drinkers", "HadAsthma", "HadDiabetes" and "HadKidney-Disease". The "Age" variable has 14 distinct levels indicating different age segments. The general health status variable, with categories 'Excellent', 'Very Good', 'Good', 'Fair', and 'Poor', was analyzed to determine the health distribution with the population. The smoking status categorized into Current Smoker, Former Smoker, and Never Smoked, was analyzed to understand the smoking prevalence among the respondents, with nearly a half of them never smoked and a quarter of them used to smoke. For the race, there are dominantly white people recorded on the list. Other categorical variables starting with "Had" show whether the patient has this disease or not.

2.Feature Engineering and Data Cleaning

We use all features as predictors which ensures that any initial model includes the full breadth of available data. This approach can capture complex relationships and interactions. In addition, In the context of predicting heart attacks, even variables with a small effect size could be clinically significant. Variables that might seem minor could be part of important interactions or be significant for specific patient subgroups.

Since the raw dataset has a large amount of observations, we will drop all missing values. Dropping missing values preserve the integrity of the dataset compared to imputation. With a large number of observations, the loss of data due to dropping missing values may be offset by the volume of complete data that remains. Since the integrity of the data is paramount, especially in medical datasets where imputation can lead to incorrect inferences. Our previous part also indicates this dataset is unbalanced. There are 316586 observations (94.5%) are patients with no heart attack and 18553 observations (5.5%) are patients with heart attack. The unbalanced nature of the dataset, with a significant disparity between the number of observations for patients with and without heart attacks, poses a challenge for predictive modeling. It can lead to a model that is biased towards predicting the majority class and may not perform well in identifying the minority class.

We apply a downsampling method to make the dataset balanced. We first divide the dataset

into two groups: one for the majority class (patients without heart attack) and one for the minority class (patients with heart attack). Then we randomly select a subset of the majority class that is equal in size to the minority class. After that, combine the downsampled majority class subset with the original minority class dataset to create a new balanced dataset.

3.Data Splitting

We split the balanced data into training and testing sets, where the training set contains 70% of the balanced data and the testing set contains 30% of the balanced data.

4.Data Drift Detection

To make sure our models are valid in the training and testing process, we conduct a data drift detection to see if there is any difference on training distribution and test distribution. Here we can see that all features are following the same statistical distribution and there is no data drift.

	Feature	Type	Reference Distribution	Current Distribution	Data Drift	Stat Test	Drift Score
>	RaceEthnicityCategory	cat			Not Detected	Jensen-Shannon distance	0.005365
>	Sex	cat			Not Detected	Jensen-Shannon distance	0.005164
>	PhysicalActivities	cat			Not Detected	Jensen-Shannon distance	0.003991
>	MentalHealthDays	num			Not Detected	Wasserstein distance (normed)	0.003634
>	HadSkinCancer	cat			Not Detected	Jensen-Shannon distance	0.001465
>	HadStroke	cat			Not Detected	Jensen-Shannon distance	0.000757
>	HadAsthma	cat			Not Detected	Jensen-Shannon distance	0.000467

Modeling of the data

1.Introduction

We use the PyCaret package to tune several models using k-fold cross validation. After that, we use our top models to build our ensemble models. Finally, we use our top models to do predictions on the test set and analyze the result.

2.K-Fold Cross Validation Results

2.1.Gradient Boosting

Based on our results, the mean accuracy across all 10 folds is approximately 75.86%, Mean AUC is about 83.64%, confirming the model's good classification ability noted in the ROC curves. The model has a balanced recall and precision with both around 75.86% and 75.92% respectively, indicating a balanced trade-off between the two metrics. The mean F1-score, which is the harmonic mean of precision and recall, is approximately 75.84%, suggesting a good balance between precision and recall.

2.2.Logistic Regression

Based on our results, the model has a mean accuracy of about 75.98%. This metric indicates the proportion of the total number of predictions that were correct. The mean AUC score is 83.58%. The mean recall is 75.98%. Recall measures the model's ability to identify true positives from the data. The mean F1 score is 75.97%, which is a balance between precision and recall.

2.3.SVM model

The SVM model finds a mean accuracy and recall of 71.62%, complemented by a mean precision of 75.27%, a mean F1 score 70.18% and an impressive mean AUC of 82.52% The accuracy means that 71.62% of total numbers of predictions are predicted correctly.

2.4.Random Forest model

The model exhibits a mean accuracy and recall of 74.52%, complemented by a mean precision of 74.63%, a mean F1 Score 74.49% and an impressive mean AUC of 81.71% The accuracy shows that it correctly predicts the 74.52% of total numbers of predictions correctly.

2.5.Decision Tree mode

Decision tree model demonstrates a mean accuracy and recall of 67.27%, alongside a mean precision of 66.28%, an F1 score 67.27% and boasts a notably high mean AUC of 67.31%. The accuracy means that 67.27% of the total numbers of predictions are predicted correctly.

2.6.XGBoost

After building the XGBoost model, we have the results with accuracy 0.7524, AUC 0.8270, and F1-score 0.7521.

2.7.Ensemble model

According to previous results, the best-fitting model is logistic regression. We will build it for the bagging classifier. Then we will combine our top models such as logistic regression and SVM to build our voting classifier.

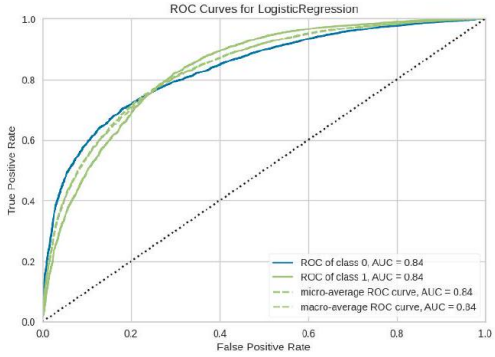
We used Logistic Regression to implement the Bagging ensemble model. Both accuracy and F1-score are 0.70 for the bagging model. From the result, we see that the bagging model accuracy and F1-score are lower than individual logistic regression. This is because the data bagging uses may be not as efficient as the whole data since some observations can be the same.

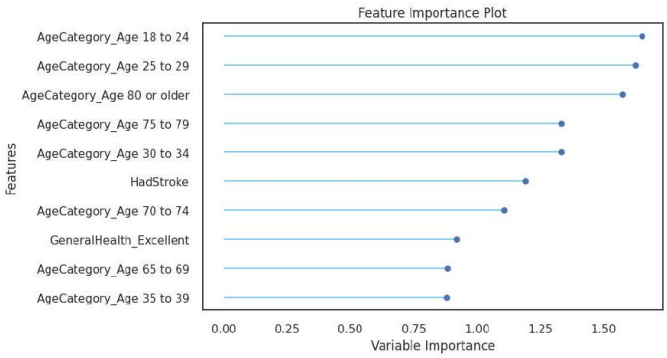
For the voting classifier, since logistic regression is the model with the best prediction accuracy and the model sets voting as hard, the accuracy result is very similar to the logistic regression. Besides accuracy, other metrics such as F1-score and recall are also similar to logistic regression.

3.Top Models Fitting Results On Test Set

After comparing the cross-validation results, we choose the top 3 models, which are logistic regression, XGBoost, and random forest. Here we will use these three models on test datasets.

3.1.Logistic Regression





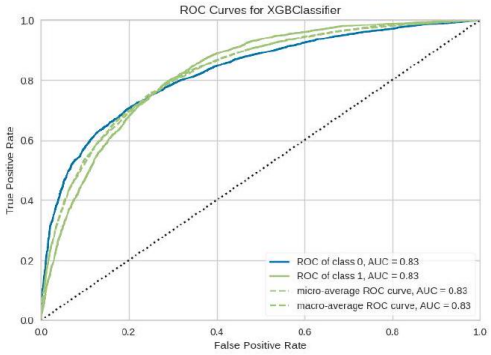
From our testing ROC Curves, the AUC for both classes is 0.84, indicating a good predictive performance. This suggests that the model has a strong ability to differentiate between patients who will have a heart attack and those who will not, with an 84% probability that the model will correctly discriminate between a randomly chosen pair of patients.

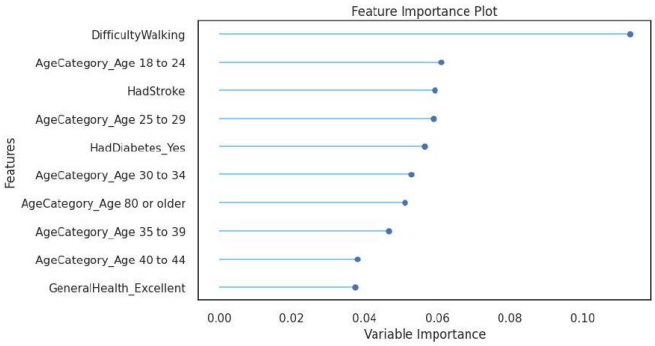
The feature importance plot shows that different age categories significantly impact the model, 'Age 18 to 24 ' has the highest predictive power followed by other age groups. The variable "HadStroke" also shows significant importance, reinforcing the medical insight that patients who have had a stroke are at a higher risk of heart attacks.

From the Logistic Regression Confusion Matrix, we can have a precision of 0.7597, recall of

0.7590, and F1-score of 0.7588, meaning that logistic regression predicts evenly positive values and negative values. Since the age category is dominant in the feature importance plot, the misclassified observations are greatly influenced by the patient's age. In other words, the logistic regression will classify the patient's heart attack greatly based on the patient's age category.

3.2.XGBoost

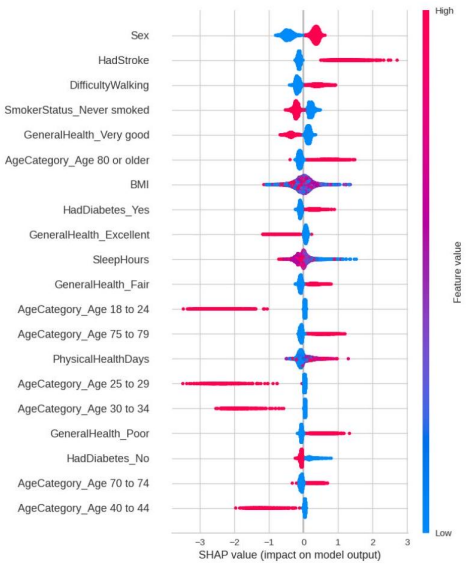




From our testing ROC Curves, the AUC for both classes is 0.83. This indicates a good level of discrimination ability of the model, with an 83% chance that the classifier will be able to distinguish between a positive and a negative instance.

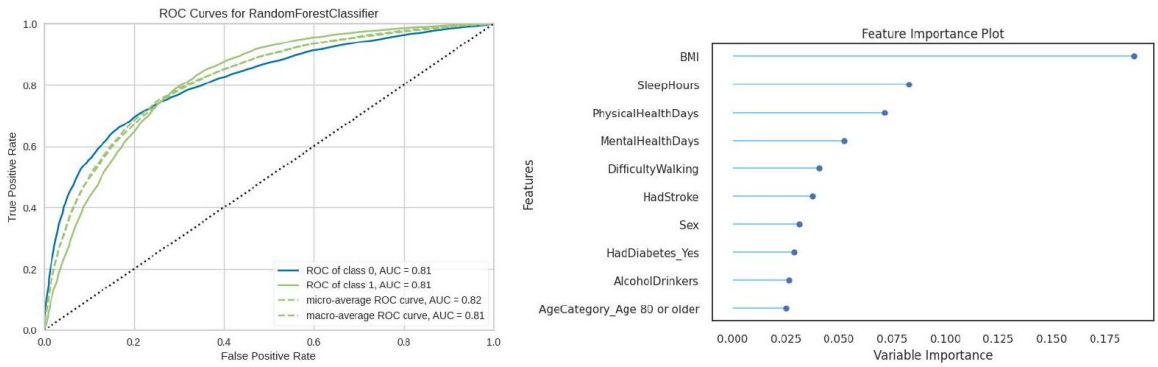
The feature importance plot shows that difficulty is the top predictive feature, suggesting a strong link between mobility issues and the risk of heart attacks. Several age groups are important features as they were in logistic regression.

From the XGBoost Confusion Matrix, we can have a precision of 0.7540, recall of 0.7528, and F1-score of 0.7525, meaning that XGBoost also predicts evenly positive values and negative values, which is similar to logistic regression. Compared to logistic regression, XGBoost predicts heart attack mostly based on the patient's DiffcultyWalking, then the patient's age category, HadStroke, and HadDiabetes. Therefore, the misclassified observations are greatly influenced by DifficultyWalking, which is following medical commonsense, as mobility issue is a crucial factor of having a heart attack.



From the interpretation plot, in the XGBoost model, we can conclude that a high value of HadStroke will significantly increase the probability of having a heart attack. In addition, after age 70, the patient will have a higher and higher probability of having a heart attack with the increase of age; in contrast, age under 30 will significantly reduce the probability of having a heart attack.

3.3.Random Forest



For both class 0 and class 1, the AUC is 0.81, which means the model correctly distinguishes between a person with heart disease and a person without heart disease 81% of the time. This shows the ability to identify whether a person would get heart disease is good.

In the random forest model, BMI shows leading importance. After that, Sleep Hours and PhysicalHealthDays show substantial importance. Noticeably, different from the other 2 models in which different age groups become the most important features, only the age group 80 or older is shown in the feature importance plot of the random forest model.

The random forest classifier shows a relatively balanced performance between precision 0.73 and recall 0.785, which is crucial for medical diagnostic models where both identifying as many positive cases as possible (high recall) and ensuring those identified are truly positive (high precision) are important.

The random forest has a good recall rate, which is particularly important in medical settings where missing a positive case (heart attack) can have severe consequences. This means it captures a substantial proportion of the actual positive cases.

The weakness of the model is the false positive rate is substantial, meaning a significant number of patients would be incorrectly flagged as having a heart attack. This could lead to unnecessary anxiety or treatments, which can be costly and potentially harmful.

The model could benefit from further tuning, possibly by adjusting the decision threshold to decrease the number of false positives or using techniques like feature engineering, sampling methods, or ensemble techniques to improve overall accuracy and precision.

4.Comparison of test results and training results

	Training Error		Testing Error	
Models	AUC	F1 Score	AUC	F1 Score
Logistic Regression	0.8358	0.7597	0.8355	0.7588
Xgboost	0.8364	0.7584	0.8270	0.7525
Random Forest	0.8171	0.7449	0.8133	0.7468

From the table, Logistic Regression is likely to be the most stable and effective model for this particular dataset, given its high and consistent AUC and F1 scores across both training and testing phases. Xgboost shows a higher AUC in training, but it does not perform as well as logistic regression in testing. Random Forest, although it does not perform as well as other models, shows good robustness in the F1 Score from training to testing.

Logistic Regression and Random Forest show excellent generalization capabilities. Logistic Regression, with nearly identical training and testing scores, would be recommended for scenarios where stability across different datasets is critical. While XGBoost shows potential in training, its performance dip in testing could be a concern. Techniques such as parameter tuning, pruning, or more advanced regularization might be necessary to improve its generalization. Random Forest's performance enhancement in testing for the F1 Score makes it suitable for applications where recall might be more important than precision, or where the cost of false negatives is high.

5.Error Analysis

To identify the misclassified observation types and patterns, we created boxplots for important numerical variables of misclassified and correctly classified test data respectively, and we created histograms for important categorical variables of misclassified and correctly classified test data respectively. After seeing the important feature distribution in misclassified and

correctly classified test data, we found their statistical distributions are generally the same, meaning that the model is already fine-tuned and there are no obvious patterns for misclassified observations, which is reasonable since we have balanced the raw data to avoid significant bias on classification.

Conclusion

After analyzing logistic regression, XGBoost, and random forest, we finally choose random forest as the model to predict the future unseen data. This is because the random forest model has excellent medical interpretability even though the AUC is slightly lower than the other two models. In medical practice, we would hope to predict the potential disease based on scientific measurements, rather than simply use the patient's age to judge.

Discussion section

In the beginning, the project expected some features like BMI, HadDiabetes, and DifficultyWalking would contribute to predictions of heart disease. However, models with the best prediction accuracy such as logistic regression heavily rely on age groups to determine the result. Random forest, on the other hand, uses features like BMI and HadStroke to predict the result, which is more suitable in the medical area. In the future, more work on refining the random forest model needs to be done to increase the accuracy and AUC of the model. Additionally, in the dataset the project uses, there are many missing values and the dataset is unbalanced. Finding a dataset with higher quality may help a lot in the prediction.

References

Wells AJ. Passive smoking as a cause of heart disease. J. Am. Coll. Cardiol. 1994;24:546-554. doi:

10.1016/0735-1097(94)90315-8. - DOI - PubMed