Project Report for Predicting Heart Attack Based on Clinical Data

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

May 30 2024

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 the CDC, diabetes, obesity, diet, disability, and alcohol consumption will influence the risk of getting heart disease. However, the CDC did not state which factor will have more effects. In this project, the first question that needs to be answered is how would these factors cause heart attacks. 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.

Data Preparation

The dataset was first loaded from a CSV file, which contains patient information from the year 2022. This dataset contains missing values and focuses on a variety of health indicators including categorical and numerical features. There are 13 categorical features which are 'RaceEthnicityCategory', 'AgeCategory', 'GeneralHealth', 'PhysicalActivities', 'SmokerStatus', 'HadSkinCancer', 'HadStroke', ''Sex', 'DifficultyWalking', 'AlcoholDrinkers', 'HadAsthma', 'HadDiabetes', and 'HadKidneyDisease'. The 4 numerical features are 'PhysicalHealthDays', 'BMI', 'MentalHealthDays', and 'SleepHours'.

Once the dataset is loaded, 109993 observations with missing values which are 25.8% of the original dataset are removed to ensure that the analysis is based on reliable and complete data. Following this, the dataset is addressed for imbalance based on the number of records reporting heart attacks. The dataset was imbalanced, with a majority of instances representing patients without a heart attack, where 316586 cases have no heart attack while 18553 have. To address this imbalance, upsampling of the minority class (patients with a heart attack) was performed to match the number of instances in the majority class. After the upsampling, 298033 observations have been added to the data set, so there are total of 633172 observations .

The dataset was split into training and testing sets, maintaining the proportion of the target variable using stratified sampling. The training subset has 427391 observations while the validation and test subsets have 47488 and 158293 respectively. Categorical features were one-hot encoded to transform them into a format suitable which has 47 columns for machine learning algorithms. Numerical features were scaled to have a mean of zero and a standard deviation of one. After that, the training data was further split into training and validation sets which respectively have 226218 and 25136 observations to tune the model and prevent overfitting.

Modeling

Baseline Model

We started with an initial basic neural network architecture, which included an input layer corresponding to 12 neutrons and 47 variables we chose from the dataset with ReLU activation, one hidden layer with 8 neurons and ReLU activation, and an output layer with a sigmoid activation function. The initial model was trained using standard hyperparameters and evaluated on the validation set.

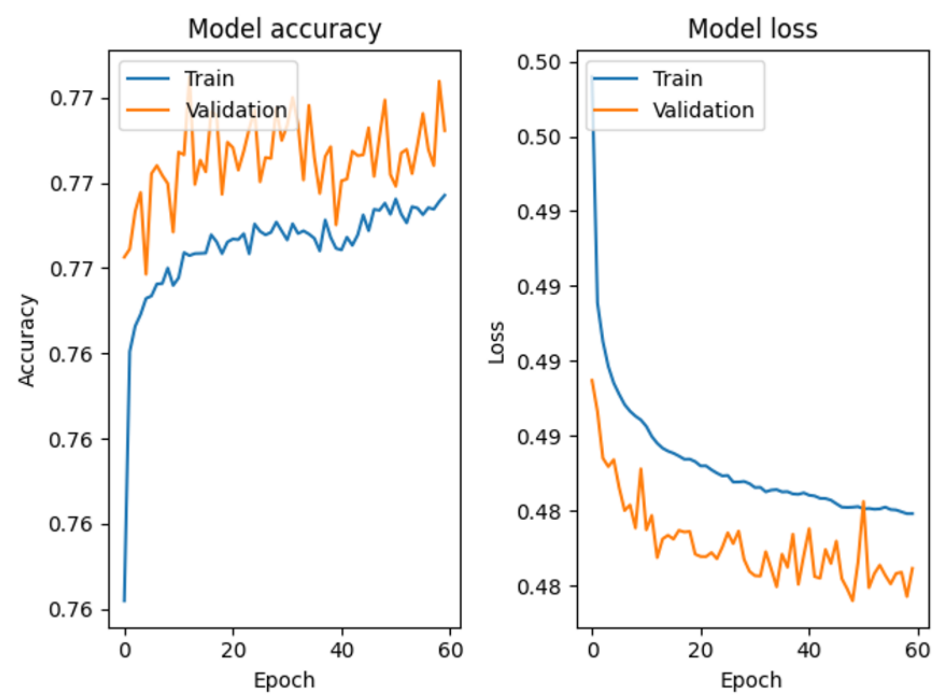

Figure 1: Baseline Model Train and Validation

Figure 1 illustrates the training and validation accuracy and loss of the baseline neural network model over 60 epochs. The left plot displays the model accuracy, while the right plot shows the model loss. The training accuracy starts at around 0.745 and gradually increases to approximately 0.775 by the end of 60 epochs. This steady increase indicates that the model is learning from the training data over time. The validation accuracy starts higher than the training accuracy at around 0.76 and fluctuates throughout the training process. It shows a general upward trend but with noticeable variability, ending slightly higher than the training accuracy at around 0.78. However, the fluctuation in validation accuracy suggests that the model is generalizing to the validation set but with some instability. This could be due to the inherent variability in the validation set or a sign of overfitting where the model learns specific patterns in the training data that do not generalize well. The initial rapid decrease in validation loss followed by stabilization suggests that the model quickly learns the main patterns but struggles to improve further, possibly due to overfitting or limited complexity in the model architecture.

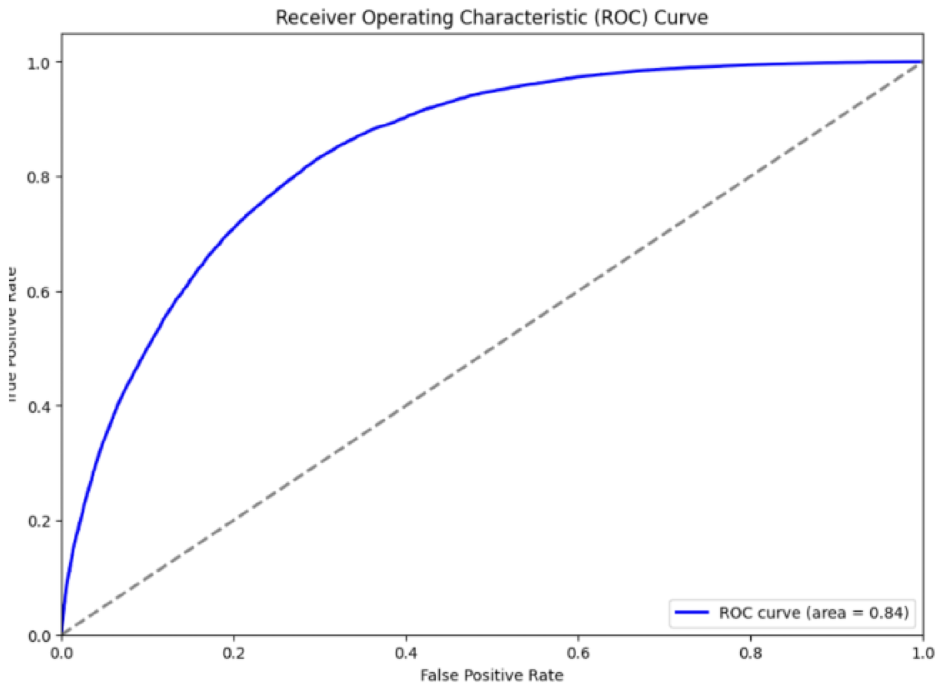

Figure 2: Baseline Model ROC-AUC

After that, we evaluate the performance of the baseline model on testing data. The confusion matrix shows that the model correctly identified 55,458 instances of 'No Heart Attack' (True Negatives) and 65,865 instances of 'Heart Attack' (True Positives). However, it also misclassified 23,688 instances as 'Heart Attack' (False Positives) and 13,282 instances as 'No Heart Attack' (False Negatives). The Area Under the Curve (AUC) is 0.84, suggesting that the model has a strong ability to differentiate between the two classes.

In terms of precision, the model demonstrates that 81% of the instances predicted as 'No Heart Attack' are correct, while 74% of the instances predicted as 'Heart Attack' are correct. This indicates that the model is slightly more accurate in predicting non-heart attack cases compared to heart attack cases. The recall for 'No Heart Attack' is 70%, suggesting that the model correctly identifies 70% of the actual non-heart attack cases. In contrast, the recall for 'Heart Attack' is higher at 83%, indicating that the model is better at correctly identifying heart attack cases.

The F1-scores, which balance precision and recall, are 0.75 for 'No Heart Attack' and 0.78 for 'Heart Attack', reflecting a moderate performance. Overall, the model achieves an accuracy of 77%, meaning that 77% of all predictions made by the model are correct. The macro averages for precision, recall, and F1-score are all 0.77, suggesting that the model performs similarly across both classes when averaged without considering class imbalance. The weighted average, also 0.77, indicates the model's overall performance while taking into account the number of instances for each class.

While the baseline model demonstrates reasonable accuracy and performs moderately well in predicting both 'No Heart Attack' and 'Heart Attack' cases, there is room for improvement. Specifically, the model's recall for 'No Heart Attack' cases is lower, indicating a need for better identification of non-heart attack cases. Enhancing the model through techniques such as hyperparameter tuning, adding regularization, or using more complex architectures could potentially improve its overall performance.

Optimal Model

Bayesian hyperparameter optimization was carried out to enhance the performance of the initial neural network model. Using Keras Tuner, we explored a range of configurations, including the number of units in the hidden layers and various learning rates. The optimal model architecture consisted of three hidden layers and one output layer. The first hidden layer had a tunable number of neurons ranging from 2 to 24, employing the activation function chosen from "relu", "tanh", and "sigmoid". The other two layers also had tunable neurons within the same range as the first hidden layer. The output layer had one neuron with a sigmoid activation function to handle binary classification.

In addition to configuring the layers, the learning rate was also treated as a tunable hyperparameter with potential values of 0.01, 0.001, and 0.0001. This comprehensive tuning allowed us to search for the optimal set of hyperparameters that would maximize the model's performance on the validation set. The hyperparameter search revealed the best configurations for the number of units in each layer, dropout rates, and learning rate, resulting in an optimized neural network model.

For the tuning process, the `max_epochs` is set to 10 which makes a maximum of 10 epochs for training each model during the tuning process. Additionally, there is an early stop function to stop the running if the validation loss is not decreasing for 5 consecutive epochs. To find the best hyperparameter, we use the function which randomly selects the value for the hyperparameter.

After applying the best hyperparameter in the optimal model and fitting the model on the training data with 60 epochs, the optimal model has 3 Dense layers. The model has [18, 24, 16] neurons in the three layers.

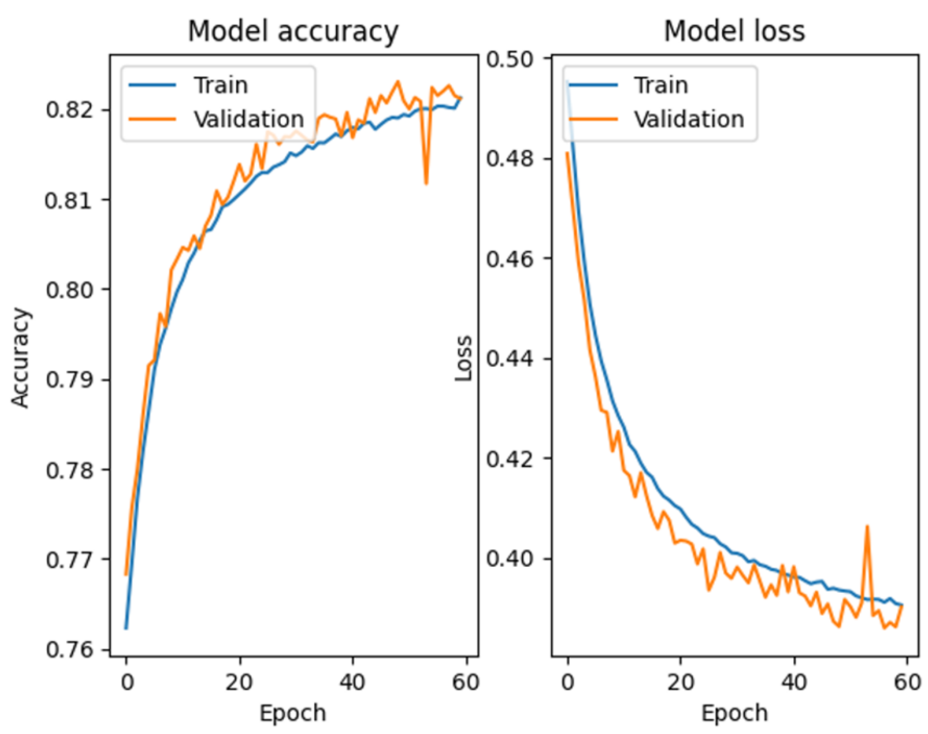

Figure 3: Optimal Model Train and Validation

Figure 3 illustrates the training and validation accuracy and loss of the optimized neural network model over 60 epochs, highlighting the model's improved performance. The training accuracy starts at around 0.76 and steadily increases to approximately 0.82 by the end of the training period, indicating effective learning from the training data. Similarly, the validation accuracy begins at 0.77 and closely tracks the training accuracy curve, also reaching about 0.82 by the end of 60 epochs. This close alignment between the training and validation accuracy curves suggests that the model generalizes well to unseen data, avoiding overfitting.

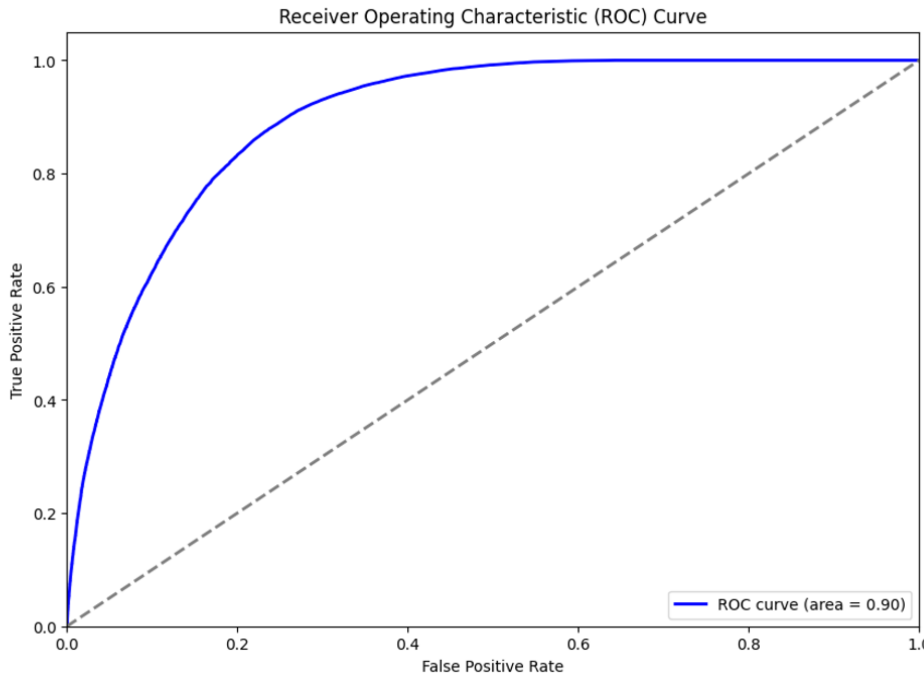

Figure 4: Optimal Model ROC-AUC

The model achieves an overall accuracy of 82% on the test set, correctly predicting 82% of instances. The precision for predicting 'No Heart Attack' is 0.89, meaning that 89% of these predictions are correct, while the precision for 'Heart Attack' predictions is 0.77, indicating that 77% of these predictions are accurate. The model demonstrates a recall of 0.72 for 'No Heart Attack' and an impressive 0.92 for 'Heart Attack,' suggesting it successfully identifies 92% of actual heart attack cases. This high recall for heart attacks is critical in medical applications, ensuring most positive cases are detected. The F1-scores, which balance precision and recall, are 0.80 for 'No Heart Attack' and 0.84 for 'Heart Attack,' further illustrating the model's robust performance. The macro and weighted averages for precision, recall, and F1-score, all around 0.82, highlight the model's balanced and reliable performance across both classes. These metrics confirm that the optimization process significantly enhanced the model's predictive capabilities, making it a valuable tool for predicting heart attacks.

Summary of Model Performance on the Test Set

Table 1: Model Performance Summary

Models	Architecture	Precision	Recall	F1-score	ROC-AUC
Optimized MLP	3 hidden layers [18,24,16]	0.83	0.82	0.82	0.90
Baseline MLP	2 hidden layers [12,8]	0.77	0.77	0.77	0.84
Logistic Regression	17 predictors	0.75	0.77	0.76	0.84

The Optimized MLP Model with 3 hidden layers (18, 24, and 16 neurons) demonstrates superior performance across all metrics (Precision, Recall, F1-score, and ROC-AUC) compared to the Baseline MLP Model and Logistic Regression. This suggests that the added complexity in the Final MLP Model effectively improves its predictive capability for heart attack prediction.

The optimized model outperforms the initial model across all major performance metrics. The overall accuracy of the optimized model is 0.82, compared to 0.77 for the initial model. The precision for 'No Heart Attack' in the optimized model is 0.89, compared to 0.81 in the initial model, and for 'Heart Attack,' it is 0.77 compared to 0.74. The recall for 'Heart Attack' in the optimized model is significantly higher at 0.92, compared to 0.83 in the initial model. The F1-scores for both classes are also improved in the optimized model. The AUC score of the optimized model is 0.90, indicating better discriminative ability compared to the initial model's AUC score of 0.84.

Comparison across those three models, including the logistic regression model which is our best model during our previous machine learning process. The optimized MLP, with three hidden layers consisting of 18, 24, and 16 neutrons respectively, outperforms the baseline MLP and logistic regression across all metrics. It achieves a precision of 0.83, recall of 0.82, F1-score of 0.82, and an ROC-AUC of 0.90. This makes the optimized MLP a more reliable and accurate tool for predicting heart attacks, highlighting the effectiveness of hyperparameter optimization and deeper network architectures.

In contrast, the baseline MLP, which has two hidden layers with 12 and 8 neurons, achieves a precision, recall, and F1-score of 0.77, and an ROC-AUC of 0.84. The logistic regression model, using 47 predictors, performs similarly to the baseline MLP with a precision of 0.75, recall of 0.77, F1-score of 0.76, and a ROC-AUC of 0.84.

Error Analysis

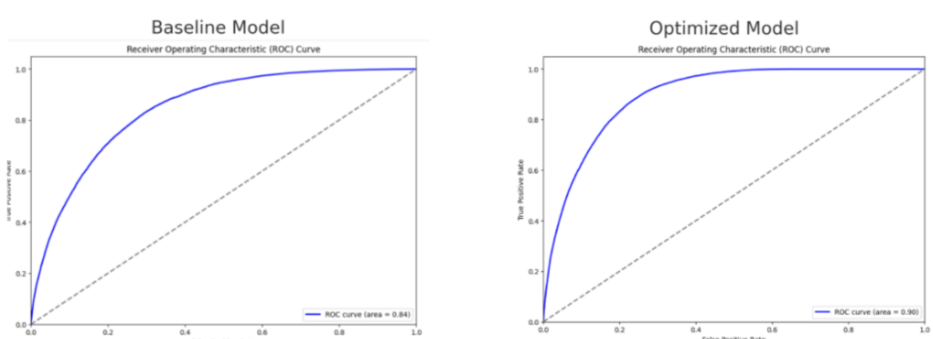

Figure 5: ROC-AUC Comparison of Baseline Model and Optimized Model

The baseline model, depicted on the left, has an ROC-AUC of 0.84. The ROC curve shows a steady rise, indicating moderate discriminative ability. However, the optimized model, shown on the right, demonstrates a higher ROC-AUC of 0.90. The ROC curve for the optimized model rises more steeply towards the top-left corner, reflecting better sensitivity and specificity. This improvement in the ROC-AUC indicates that the optimized model is more effective at distinguishing between heart attack and non-heart attack cases, further validating the enhancements made through hyperparameter optimization and model architecture refinement.

To identify the misclassified observation types and patterns, we used data drift detection to see the difference between misclassified and correctly classified test data. After seeing the distribution analysis in misclassified and correctly classified test data, we found their statistical distributions are statistically the same, meaning that the model is already well-trained 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.

Discussion

The results from our study indicate that our optimized MLP model significantly enhances the predictive capability for heart attacks compared to the baseline MLP model and logistic regression. By employing a deeper network architecture and comprehensive hyperparameter tuning, we achieved improvements across all major performance metrics, including precision, recall, F1-score, and ROC-AUC. Notably, the optimized model exhibited a substantial increase in recall for heart attack cases, ensuring a higher rate of correct identifications, which is critical for medical applications. The process of addressing data imbalance and removing missing values proved essential in enhancing the model's reliability and accuracy. Despite the overall improvement, the model still faces challenges in distinguishing between certain cases, as evidenced by the similar statistical distributions of correctly and incorrectly classified instances. This suggests that future work could explore more advanced techniques, such as incorporating additional data sources, employing ensemble methods, or refining feature engineering, to further improve the model's robustness. Additionally, our findings emphasize the importance of hyperparameter optimization and the potential of neural networks in medical prediction tasks, offering valuable insights for future projects in the healthcare domain.

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.