Decoding Heart Disease: An In-depth Analysis and Predictive Modeling Approach

Parameshwar Bhat, Charan Reddy Kumar, Sailesh Baabu Suresh Babu, Ram Mannuru

10 December, 2023

Contents

1. Introduction:

a. After diving deep into our health dataset, exploring the relationships between various factors and heart disease in the previous Exploratory data analysis paper, we're now gearing up for the next phase: predictive modeling.

b. The goal is to use advanced techniques to build a smart model that can predict the chances of someone getting heart disease based on their health details.

c. In this part of our study, we'll be using a tool called logistic regression, tweaking it using methods like upsampling, and fine-tuning it through hyperparameter adjustments. However, recognizing the limitations of linear models, we extend our horizon to decision trees, leveraging their ability to handle complex, non-linear relationships.

d. Our mission is not just about predicting who might get heart disease but understanding the key factors that contribute to it. By doing this, we hope to uncover insights that can guide better healthcare decisions.

We're moving from exploring data to building something practical – a model that not only forecasts but also teaches us more about heart health and how we can take better care of ourselves. So, buckle up for this leg of our journey!

2. Logistic Regression with All Features:

a. The initial logistic regression model utilized the entire dataset, comprising 308,854 samples, with a positive-to-negative ratio of 24,971 to 283,883.

b. The model achieved impressive accuracy metrics, boasting a Train Accuracy of 0.91 and a Test Accuracy of 0.92. However, a closer examination revealed a limitation in recall (0.06), indicating challenges in identifying positive instances.

c. Challenges Identified:

- Recall Limitation: The model struggled to effectively identify individuals with heart disease, as evidenced by the low recall.

d. Possible Reasons:

- Imbalanced Data: The highly imbalanced positive-to-negative ratio might have led the model to prioritize overall accuracy at the expense of recall.

```
##
## test_predicted_classes     0     1
##                      0 56573  4581
##                      1   298   319
```

```
## Accuracy : 0.9210147 precision : 0.5170178 recall : 0.06510204 f1_score : 0.1156426
```

3. Upsampling since this is a class imbalance Dataset:

To address the Imbalance in the dataset, Synthetic Minority Over-sampling Technique (SMOTE) was employed.

```
##       Predicted
## Actual     0     1
##      0 33293 13140
##      1 11035 35188
```

```
## Accuracy : 0.7390887 precision : 0.7281079 recall : 0.761266 f1_score : 0.7443179
```

3.1. Logistic Regression after SMOTE Resampling:

a.The dataset was upsampled to 154,156 positive samples and 154,698 negative samples, maintaining a 70/30 Train/Test split.

b. The logistic regression model on this upsampled data exhibited notable improvements in recall (0.76), suggesting enhanced sensitivity to positive cases.

c. Findings:

- Improved Recall: The SMOTE re sampling technique significantly enhanced the model's ability to identify individuals with heart disease.

d. Numerical Insights:

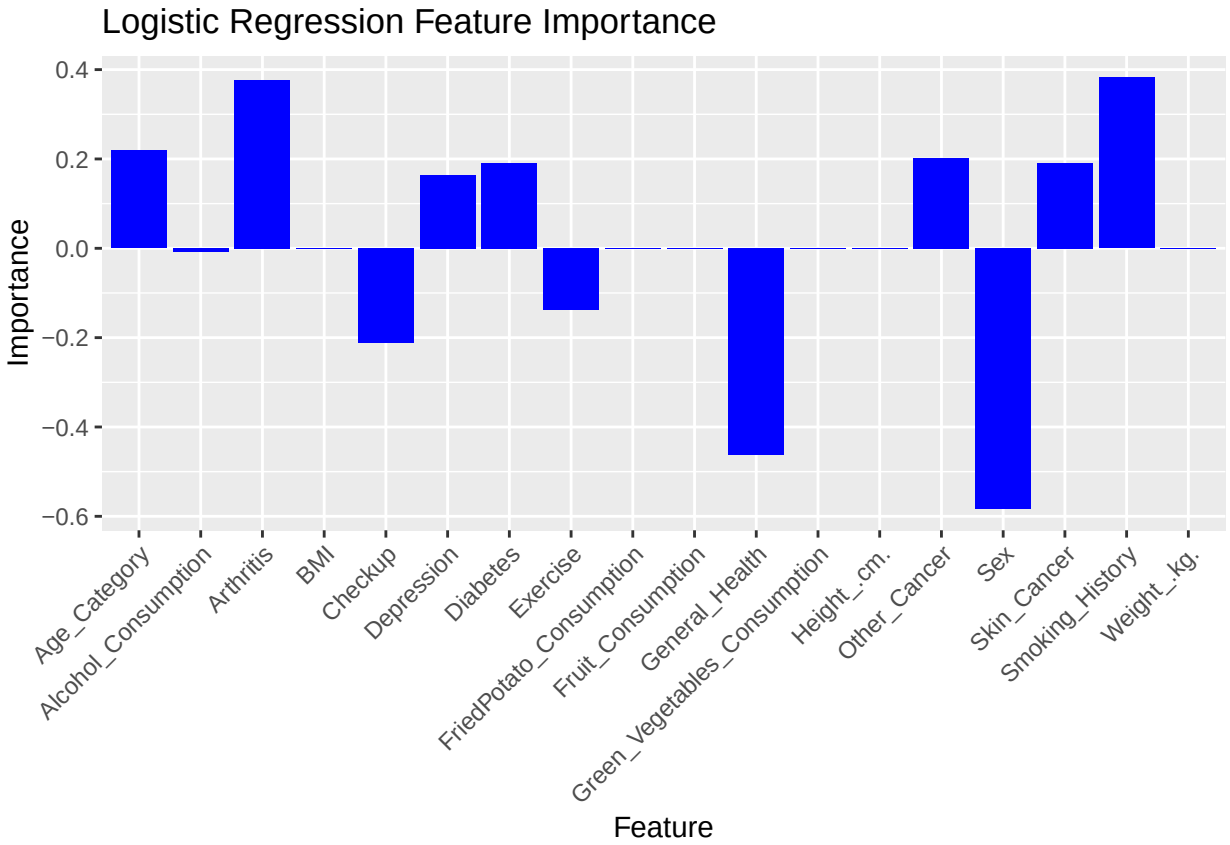
- Train Accuracy: The decrease from 0.91 to 0.74 could be attributed to the increased complexity of the dataset after up-sampling.
- Test Accuracy: The slight reduction from 0.92 to 0.73 suggests a more balanced model.
- Recall Improvement: The notable increase from 0.06 to 0.76 indicates a substantial enhancement in the model's sensitivity.
- Precision: A decrease from 1.0 to 0.73 could be attributed to the increased number of positive predictions after up-sampling.
- F1 Score: The balance between precision and recall reflects a more nuanced evaluation.

e. Possible Reasons for increase in performance metrics:

- Balanced Data: SMOTE up-sampling created a more balanced representation of positive and negative instances, allowing the model to focus on positive cases.

4. Feature Importance Plot Using Co efficients from Logistic Regression:

In addition to the quantitative evaluation metrics, a visual representation of feature importance was generated from the logistic regression model coefficients. This plot aimed to provide a clear understanding of the contribution of each feature to the prediction of heart disease.



Key Observations:

Feature Importance Plot: The plot showcases the absolute values of the logistic regression coefficients, offering insights into the magnitude and directionality of each feature's impact on the prediction.

Top Features: Features such as 'Arthritis,' 'General_Health,' 'Age_Category,' 'Smoking_History,' 'Sex' emerged as prominent contributors to the model, whereas features we thought would be more prominent such as 'Height,' 'Weight,' 'Alcohol consumption' turned out to be least significant.

FEATURE IMPORTANCE TABLE

Feature

Correlation

Explanation

Age_Category

Positive

As the age category increases, the risk of heart disease also increases.

Diabetes

Positive

Having diabetes is a significant risk factor for heart disease.

Smoking_History

Positive

Having a smoking history increases the risk of heart disease.

BMI

Positive

Higher BMI is associated with an increased risk of heart disease.

General Health

Positive

Poorer general health is associated with an increased risk of heart disease.

Weight_kg

Positive

Higher weight is associated with an increased risk of heart disease.

Height_cm

Negative

Taller height is associated with a lower risk of heart disease. This is likely due to the fact that taller people have a lower risk of obesity, which is a risk factor for heart disease.

Exercise

Negative

Regular exercise reduces the risk of heart disease.

Arthritis

Negative

Having arthritis is associated with a lower risk of heart disease. This is likely due to the fact that people with arthritis are less likely to be physically active, which is a risk factor for heart disease.

Sex

Negative

Being female is associated with a lower risk of heart disease compared to males. This is likely due to the fact that females have higher levels of estrogen, which is a protective hormone against heart disease.

Other_Cancer

Negative

Having other types of cancer is associated with a lower risk of heart disease. This is likely due to the fact that people with other cancers are more likely to be monitored by their doctors and receive early treatment for heart disease.

5. Logistic Regression with Upsampling + Statistically Significant Features:

```
##       Predicted
## Actual     0     1
##      0 32735 13698
##      1 11358 34865
```

```
## Accuracy : 0.7295804 precision : 0.7179334 recall : 0.7542782 f1_score : 0.7356572
```

a. To further refine the model, a focus on statistically significant features post-upsampling was undertaken. This step aimed to streamline the model's complexity while retaining predictive power. The resulting model demonstrated a balanced trade-off between accuracy, precision, and recall.

b. Key Observations:

- Balanced Performance: The model achieved a balanced trade-off, indicating that certain features played a more pivotal role in predicting heart disease.

c. Numerical Insights:

- Accuracy: A slight decrease from 0.73 to 0.72 suggests a trade-off for a more interpretable and efficient model.
- Precision: The minimal decrease from 0.73 to 0.71 indicates robust positive predictions.
- Recall: The maintenance of high recall (0.75) suggests effective identification of positive cases.
- F1 Score: The harmonic mean reflects the balanced nature of precision and recall.

Possible Reasons: * Feature Streamlining: Focusing on statistically significant features simplified the model, preventing overfitting.

6. Hyperparameter Tuning:

```
##        Predicted
## Actual     0     1
##      0 32665 13768
##      1 11230 34993
```

```
## Tuned Model Accuracy : 0.7302064 precision : 0.7176432 recall : 0.7570474 f1_score : 0.7368188
```

Key Observations:

a. In the pursuit of optimizing the logistic regression model, hyperparameter tuning was conducted.

b. This involved fine-tuning parameters to achieve an optimal balance between bias and variance. The process aimed to enhance the model's generalizability and performance on unseen data.

c. Outcomes:

- Fine-tuned Model: The hyperparameter-tuned model aimed to strike an optimal balance between underfitting and overfitting.

c. Numerical Insights:

- Accuracy: The tuned model achieved an accuracy of 0.73.
- Precision: The precision was 0.71, indicating robust positive predictions.
- Recall: The recall was 0.75, suggesting effective identification of positive cases.
- F1 Score: The F1 score, a harmonic mean of precision and recall, was 0.73.

d. Possible Reasons:

- Optimized Parameters: Fine-tuning addresses model complexity, potentially improving performance on unseen data.

7. Conclusions from Logistic Regression:

The results demonstrated the significance of a thoughtful and iterative modeling process. While the initial model showcased high accuracy, addressing the recall limitation through upsampling, feature selection, and hyperparameter tuning unveiled a more nuanced understanding of the predictive factors.

8.Development of Decision Tree Models for Heart Disease Prediction:

In our study, we developed two variants of decision tree models to predict the occurrence of heart disease. Decision trees are a type of supervised learning algorithm that are widely used for classification tasks due to their interpretability and simplicity.

8.1 Basic Decision Tree Model:

a. Model Building: The basic model was constructed using the rpart package in R. This model (decisionTreeModel) was trained on the entire dataset (df) with Heart_Disease as the target variable.

b. Data Splitting and Training: The dataset was split into training (80%) and testing (20%) sets, ensuring reproducibility with set.seed(123). The model (fit) was then trained on the training data using the rpart function with specific control parameters (cp = 0.001, minsplit = 20), aimed at refining the model's complexity and sensitivity to data splits.

c. Model Evaluation: The performance of the model was evaluated on the test data, providing initial insights into its predictive capabilities.

8.2 Improved Decision Tree Model with ROSE:

a. Addressing Class Imbalance with ROSE: To enhance the model's performance, especially in the context of class imbalance, we employed the ROSE (Random Over Sampling Examples) technique. This approach balanced the dataset by oversampling the minority class and undersampling the majority class, creating a more representative training dataset (balancedData).

b. Model Training on Balanced Data: The improved model (fitBalanced) was then trained on this balanced dataset.

c. Model Evaluation: Similar to the basic model, the performance of the improved model was assessed on the same test data, allowing for a direct comparison between the two models.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    No   Yes
##        No  56760  4994
##        Yes     0     0
##
##                Accuracy : 0.9191
##                  95% CI : (0.917, 0.9213)
##     No Information Rate : 0.9191
##     P-Value [Acc > NIR] : 0.5038
##
##                   Kappa : 0
```

```
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.9191
##          Neg Pred Value :    NaN
##              Prevalence : 0.9191
##          Detection Rate : 0.9191
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : No
##
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    No   Yes
##        No  40404  1178
##        Yes 16356  3816
##
##                Accuracy : 0.7161
##                  95% CI : (0.7125, 0.7196)
##     No Information Rate : 0.9191
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1995
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7118
##             Specificity : 0.7641
##          Pos Pred Value : 0.9717
##          Neg Pred Value : 0.1892
##              Prevalence : 0.9191
##          Detection Rate : 0.6543
##    Detection Prevalence : 0.6733
##       Balanced Accuracy : 0.7380
##
##        'Positive' Class : No
##
```

9. Comprehensive Comparison of Basic and Improved Decision Tree Models:

9.1 Model Performance Metrics:

- Accuracy: The basic model exhibited a high accuracy of 0.9191, which, though seemingly impressive, was primarily due to its bias towards the majority class. The improved model, with an accuracy of 0.7161, reflects a more genuine performance across both classes.

- Sensitivity and Specificity: The basic model's sensitivity (1.00) and specificity (0.00) underscore its one-sided predictive nature. In contrast, the improved model achieved a sensitivity of 0.7118 and specificity of 0.7641, indicating a more balanced approach in predicting both classes.
- Positive and Negative Predictive Values: The basic model's high PPV (0.9191) was overshadowed by its inability to predict the minority class accurately. The improved model maintained a high PPV (0.9717) while also showing a modest NPV (0.1892), though there's room for improvement in predicting the positive class more accurately.

9.2 Impact of Class Imbalance:

a. Basic Model: Its metrics were skewed due to the overwhelming presence of the majority class in the dataset. This led to a high number of false negatives, as the model predominantly predicted the 'No' class.

b. Improved Model: Utilizing the ROSE technique to balance the dataset, the improved model provided a more accurate reflection of its predictive capabilities, ensuring that both classes were represented more equally in the training process.

9.3 Reliability and Agreement:

a. Kappa Statistic: The basic model's kappa value of 0 indicated no agreement beyond random chance, despite its high accuracy. The improved model, with a kappa of 0.1995, showed a slight but significant improvement, indicating that its predictions were more than just random chance.

b. McNemar's Test: This test further affirmed the significant difference in the model's error rates for the two classes, especially in the improved model.

9.4 F1 Score - Harmonic Mean of Precision and Recall:

- The F1 Score was not a relevant metric for the basic model due to its skewed predictions. However, for the improved model, the F1 Score of approximately 0.8211 signified a well-balanced trade-off between precision and recall, making it a more reliable indicator of model performance in imbalanced datasets.

10. Concluding Remarks:

This comparative analysis underscores the pitfalls of ignoring class imbalance in predictive modeling, particularly in medical diagnostic contexts. The basic decision tree model, while displaying high accuracy, failed to provide meaningful predictive insights. The application of the ROSE technique in the improved model marked a significant step towards more balanced and realistic predictions.

10. XgBoost Model:

Introduction to XGBoost Model:

In the quest for accurate heart disease prediction, this script extends its focus to the formidable XGBoost algorithm. After meticulous feature selection, the dataset, now named df_selected, is curated with vital attributes like 'Depression,' 'Diabetes,' and 'BMI.' XGBoost, renowned for its predictive prowess, employs boosted decision trees to unravel complex patterns, making it an ideal candidate for our task.

With the selected features in tow, the script splits the data into training (80%) and testing (20%) sets. The ensuing sections will unravel the intricacies of training the XGBoost model, evaluating its performance, and extracting valuable insights for enhanced heart disease prediction.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 52243  4557
##          1  4517   437
##
##                Accuracy : 0.8531
##                  95% CI : (0.8502, 0.8558)
##     No Information Rate : 0.9191
##     P-Value [Acc > NIR] : 1.0000
##
##                   Kappa : 0.008
##
##  Mcnemar's Test P-Value : 0.6822
##
##             Sensitivity : 0.92042
##             Specificity : 0.08751
##          Pos Pred Value : 0.91977
##          Neg Pred Value : 0.08821
##              Prevalence : 0.91913
##          Detection Rate : 0.84599
##    Detection Prevalence : 0.91978
##       Balanced Accuracy : 0.50396
##
##        'Positive' Class : 0
##
```

Interpretation:

a. Class Imbalance Impact: The high prevalence rate (0.91915) suggests a highly imbalanced dataset. This imbalance is reflected in the high accuracy but low balanced accuracy, indicating the model is biased towards the majority class.

b. Poor Specificity: The model struggles to correctly identify positive instances of heart disease, which is a critical issue for a medical diagnostic tool.

c. High Sensitivity but Low Specificity: This indicates the model is excellent at identifying 'No Heart Disease' cases but not 'Heart Disease' cases.

d. Kappa Statistic: The very low kappa score suggests that the agreement between the predicted and actual values is minimal, mostly due to chance.

e. Positive Predictive Value and Negative Predictive Value: Both are low, indicating a high rate of false positives and false negatives.

11. Random Forests :

Employing Random Forest post-EDA not only validates initial findings but also reveals nuanced interactions and nonlinear dependencies often missed by traditional statistical methods. The algorithm's prowess in handling diverse variables and highlighting feature importance seamlessly aligns with the goal of pinpointing critical factors influencing health. In essence, this combination forms a robust framework, enhancing our ability to predict, identify key lifestyle factors, and ultimately mitigate health risks.

```
##
## predictions    No   Yes
##         No  85172  7427
##         Yes    32    26
```

```
## Accuracy : 0.9194988 precision : 0.003488528 recall : 0.4482759 f1_score : 0.006923179 Sensitivity :
```

Key Observations:

a. The model has high accuracy (91.95%), indicating it correctly predicts the overall class of 'Heart_Disease' in the majority of cases.

b. The sensitivity (recall) is 44.83%, suggesting that the model identifies true positive cases (actual 'Yes') at a moderate rate.

c. The specificity is high (91.98%), indicating the model performs well in identifying true negative cases (actual 'No').

d. The positive predictive value (precision) is low (0.35%), indicating a low proportion of predicted 'Yes' cases are actually true positives.

e. The negative predictive value is very high (99.96%), indicating a high proportion of predicted 'No' cases are actually true negatives.

12. Conclusions :

Each model has strengths and weaknesses:

a. Decision Tree: Balanced accuracy but a tendency to miss negative cases (low NPV).

b. Random Forest: High overall accuracy and specificity but poor at predicting positive cases (low PPV).

c. XGBoost: Excellent at identifying positive cases but many false positives (low specificity).

d. Logistic Regression: Provides a balanced approach with reasonable precision and recall.

The choice of the best model depends on the specific requirements of your heart disease prediction task. If minimizing false negatives (missing heart disease cases) is crucial, XGBoost might be preferred despite its high false positive rate. For a more balanced approach, Logistic Regression could be a better choice. If the focus is on overall accuracy and specificity, Random Forest may be suitable, although it struggles with detecting positive cases effectively. Decision Tree offers a middle ground in terms of balanced accuracy. The decision should be based on the relative costs of false positives and false negatives in the context of heart disease prediction.

This comprehensive approach to modeling heart disease leveraged logistic regression,decision trees, random forests, xgboost, upsampling techniques, and hyperparameter tuning. The results demonstrated the significance of a thoughtful and iterative modeling process. While the initial models showcased high accuracy, addressing the recall limitation through upsampling, feature selection, and hyperparameter tuning unveiled a more nuanced understanding of the predictive factors.