DSC 630 Milestone 4

Brian Reppeto

Bellevue University

DSC 630

Frank Neugebauer

October 27, 2024

1. Data Preparation:

* Loaded the data.
* Retrieved basic information about the dataset.
* Missing Values: Checked for missing values and handled them by filling with the column mean.
* Correlation Matrix: I plotted a heatmap of the correlation matrix to understand the relationships between the features.
* Feature Selection: Used SelectKBest to choose the top 5 features based on ANOVA F-statistic. The selected features were extracted and used for modeling. ['age', 'prevalentHyp', 'sysBP', 'diaBP', 'glucose']

2. Model Building and Evaluation:

* Train-Test Split: Split the data into training and test sets (80-20 split).
* Model Selection: Built two models - Logistic Regression and Random Forest.
* Training: Both models were trained using the selected features.
* Evaluation:
  + I generated classification reports for each of the models.
  + ROC AUC Score was calculated to evaluate the models' abilities to discriminate between classes.
  + Understood that refining needed to happen.
  + Applied SMOTE to help balance the dataset.

3. Results Interpretation:

* Logistic Regression Results: Classification report and ROC AUC scores were printed. The classification report included precision, recall, F1-score, and support for each class.
* Random Forest Results: Similar evaluation metrics were printed for the Random Forest model.

Logistic Regression with SMOTE Results:

----------------------------------------

Classification Report:

precision recall f1-score support

0 0.66 0.65 0.66 745

1 0.63 0.64 0.63 694

accuracy 0.65 1439

macro avg 0.65 0.65 0.65 1439

weighted avg 0.65 0.65 0.65 1439

ROC AUC Score: 0.6453

Random Forest with SMOTE Results:

---------------------------------

Classification Report:

precision recall f1-score support

0 0.88 0.89 0.88 745

1 0.88 0.87 0.87 694

accuracy 0.88 1439

macro avg 0.88 0.88 0.88 1439

weighted avg 0.88 0.88 0.88 1439

ROC AUC Score: 0.8793

4. Conclusions and Recommendations:

The performance of both models was evaluated using SMOTE to handle class imbalance in the dataset, and the results show clear differences in their effectiveness:

1. Logistic Regression with SMOTE:
   * The model achieved an accuracy of 65% with a ROC AUC score of 0.6453. The classification report shows moderate performance, with precision, recall, and F1-scores around 0.63 to 0.66 for both classes.
   * The results indicate that Logistic Regression was able to provide some discrimination between patients at risk and those not at risk of heart disease, but its overall performance is limited, potentially due to its linear nature not fully capturing the complex relationships within the data.
2. Random Forest with SMOTE:
   * The Random Forest model significantly outperformed Logistic Regression, with an accuracy of 88% and a ROC AUC score of 0.8793. Both classes showed high precision and recall (0.88 for class 0 and 0.87 for class 1), indicating that the model can effectively identify patients at risk of heart disease with a high degree of reliability.
   * Random Forest’s superior performance suggests it can capture the complexities and interactions between features, providing a more accurate prediction for heart disease risk.

Recommendation

* Given the significantly better performance of Random Forest compared to Logistic Regression, it is recommended to use Random Forest as the primary model for predicting the 10-year risk of coronary heart disease.
* The high precision and recall make Random Forest a more reliable choice for healthcare use cases, where minimizing false negatives (missing at-risk patients) and false positives (unnecessary concern or tests for low-risk patients) are both crucial.

DSC 630 Milestone 4

1. **Will I be able to answer the questions I want to answer with the data I have?**

Yes, the Framingham dataset provides enough features (demographics, behavioral, and medical) to predict coronary heart disease risk (CHD), which is the main objective. The dataset includes relevant predictors such as age, BMI, smoking habits, blood pressure, cholesterol levels, and diabetes status, which are essential for heart disease prediction.

1. **What visualizations are especially useful for explaining my data?**

**Histograms**: To show the distribution of continuous variables like age, BMI, cholesterol, etc.

**Box plots**: To detect outliers in the dataset (cholesterol, BMI, systolic blood pressure).

**Correlation heatmap**: To display relationships between features, like systolic/diastolic blood pressure and smoking/cigarettes per day.

**Scatter matrix**: Useful for identifying patterns among continuous features.

**ROC Curve**: To visualize the performance of your machine learning models, especially for classification problems like predicting CHD.

1. **Do I need to adjust the data and/or driving questions?**

**Data** Yes, the Framingham dataset contains missing values in several key variables (education, cigsPerDay, BPMeds, totChol, BMI, and glucose) I will handle these with the Mean.

**Driving questions**: Additionally, I will need to address the class imbalance in the target variable (CHD), as the dataset is heavily skewed towards non-cases. Using **SMOTE** or other resampling techniques can help balance the classes.

1. **Do I need to adjust my model/evaluation choices?**

I will experiment with models like Logistic Regression, Random Forest, and Support Vector Machine, and apply techniques like cross-validation and AUC to assess performance. As previously mentioned, I will handle the class imbalance with SMOTE. This will likely improve the model performance

1. **Are my original expectations still reasonable?**

Yes, my expectation is to accurately predict coronary heart disease using the Framingham dataset. Currently, I do not see any issues moving forward with this prediction.

**Milestone 2: Predicting the 10-Year Risk of Coronary Heart Disease**

**Introduction**

**Problem Statement:** Cardiovascular disease, particularly coronary heart disease (CHD), remains one of the leading causes of death worldwide, accounting for millions of fatalities each year. The burden of CHD not only affects individuals but also places an immense strain on healthcare systems due to the costs associated with treatment and long-term care. This project aims to address the critical issue of predicting CHD using the Framingham heart disease dataset. The goal is to develop a predictive model that can accurately identify individuals at high risk, enabling early intervention strategies that can mitigate the risk of CHD and improve patient outcomes.

**Importance of the Problem:** The ability to predict the likelihood of CHD is crucial for several reasons. Firstly, early detection allows for timely preventive measures, such as lifestyle modifications or pharmacological interventions, which can significantly reduce the risk of CHD. Secondly, accurate risk prediction models can help prioritize healthcare resources, ensuring that high-risk individuals receive the attention and care they need. Moreover, understanding the factors that contribute to CHD can help inform the public, and help create health strategies and policies aimed at reducing the incidence of heart disease at the population level.

**Target Audience:** The outcome of this project is relevant to a wide range of stakeholders. Healthcare providers, cardiologists and primary care physicians, can use the predictive model to identify patients who are at an elevated risk of CHD and tailor their treatment plans accordingly. Public health officials may also find the model useful for designing targeted interventions and educational campaigns that address the risk factors associated with CHD. Insurance companies could leverage the model to better assess the health risks of their clients, leading to more accurate premium calculations and risk management strategies.

**Data Source:** The Framingham heart disease dataset used in this project originates from the Framingham Heart Study, a study that began in 1948 in Framingham, Massachusetts. The study's primary objective was to identify the common factors or characteristics that contribute to cardiovascular disease (CVD). Over the years, the study has collected extensive data on several generations of participants, making it one of the most comprehensive sources of data on cardiovascular health. The dataset used in this project includes over 4,000 records with 15 attributes, such as age, cholesterol levels, systolic and diastolic blood pressure, smoking status, diabetes status, and family history of heart disease. These variables are crucial for assessing an individual's risk of developing CHD, and their inclusion in the dataset makes it an invaluable resource for predictive modeling.

**Data Relevance:** The Framingham dataset provides a rich source of data on the risk factors associated with CHD, many of which are well documented within medical literature. The dataset's lengthy 10-year study also allows for the analysis of trends and changes in risk factors over time, which is essential for accurately predicting the risk of CHD. Furthermore, the dataset includes a diverse range of participants, which helps ensure that the predictive model developed in this project will be applicable to a broader population. Moreover, the use of this dataset allows for the creation of a model that is not only accurate but also grounded in real-world clinical data.

**Model Selection and Rationale**

**Planned Models:** For this project, I plan to employ several machine learning models to predict the risk of CHD, each chosen for its specific strengths and suitability for the task at hand.

1. **Logistic Regression:** Logistic regression is a traditional and widely used statistical method for binary classification problems. In the case of this project where the outcome variable is categorical as to whether an individual will develop CHD or not. Logistic regression is particularly valuable because it provides interpretable coefficients that indicate the strength and direction of the relationship between each predictor variable and the outcome. This transparency is essential in healthcare, where understanding the factors driving the model's predictions is critical for clinical decision-making.
2. **Random Forest:** Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests are robust to overfitting, particularly when dealing with a large number of predictors, as they reduce variance through the averaging of multiple trees. This makes them well-suited for handling the complex interactions between variables in the Framingham dataset. Additionally, random forests can provide feature importance scores, helping to identify which risk factors are most predictive of CHD.
3. **Support Vector Machine (SVM):** SVMs are powerful supervised learning models used for classification tasks. By employing kernel functions, SVMs can model complex, non-linear relationships in the data, making them an excellent choice for predicting CHD. The flexibility of SVMs in handling high-dimensional spaces is also a significant advantage.
4. **XGBoost:** XGBoost, or Extreme Gradient Boosting, is an advanced implementation of gradient boosting algorithms. It is known for its efficiency and speed. XGBoost is particularly effective in dealing with unbalanced datasets, which may be a concern in this project if the number of CHD cases is significantly lower than non-cases. Its ability to handle missing data and optimize decision trees through regularization makes it a strong candidate for this project.

**Evaluation Metrics:** The primary evaluation metric for this project will be the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The AUC-ROC is a measure of the model's ability to distinguish between positive (CHD) and negative (non-CHD) classes. A higher AUC-ROC value indicates better model performance, with a value of 1.0 representing perfect discrimination and a value of 0.5 indicating no better than random chance. The choice of AUC-ROC is driven by its robustness and its ability to capture the trade-offs between sensitivity (true positive rate) and specificity (true negative rate) across different thresholds.

In addition to AUC-ROC, other metrics such as accuracy, precision, recall, and the F1-score will be used to provide a more comprehensive evaluation of the models’ performance. Accuracy will measure the proportion of correct predictions, while precision and recall will help assess the model's effectiveness in identifying true positive cases without producing too many false positives. The F1-score, which is the harmonic mean of precision and recall, will provide a balanced measure of the model's performance, particularly in scenarios where there is an imbalance between the classes.

**Expected Outcomes**

**Learning Objectives:** This project aims to uncover the key risk factors contributing to CHD and determine which machine learning model is most effective in predicting the risk of CHD. Through the analysis of the Framingham dataset, I hope to gain insights into the relative importance of different predictors, such as age, cholesterol levels, and blood pressure. Additionally, I seek to understand the trade-offs between model complexity and interpretability.

**Risks and Ethical Considerations:** Several risks and ethical considerations must be addressed in this project. One significant risk is the potential for bias in the predictive model, particularly if the dataset does not adequately represent certain demographic groups. This could lead to disparities in the accuracy of predictions across different populations. To mitigate this risk, I will conduct thorough bias checks and ensure that the model is tested on diverse subgroups within the dataset.

Another ethical concern is the potential misuse of the model's predictions. For example, insurance companies might use the model to justify higher premiums for individuals identified as high-risk, leading to discrimination based on health status. To address this, I will emphasize that the model should be used to support, not replace, clinical judgment and that it should be deployed with appropriate safeguards to prevent misuse.

Privacy is also a critical ethical consideration, given that the dataset contains sensitive health information. Although the dataset has been anonymized, it is essential to ensure that any findings or model outputs do not inadvertently compromise the privacy of individuals. All data handling and analysis will adhere to strict privacy and confidentiality standards.

**Mitigation Strategies:** To address the risks and ethical concerns identified above, I will implement several mitigation strategies. First, I will use fairness-aware machine learning techniques, such as re-sampling or re-weighting, to reduce bias in the model. Second, I will ensure that the model's predictions are accompanied by explanations that make it clear how the predictions were derived. This will help users understand and trust the model's outputs and ensure that the model is used appropriately.

**Contingency Plan**

**Alternative Approaches:** If the initial models do not perform as expected or fail to meet the required standards of accuracy and fairness, I will explore alternative approaches. One possibility is to employ ensemble methods, such as stacking or boosting, which combine the predictions of multiple models to improve overall performance.

References

Ashish Bhardwaj. (2022). Framingham heart study dataset [Data set]. Kaggle. <https://doi.org/10.34740/kaggle/dsv/3493583>