Predicting Heart Attack Risk Using Machine Learning and CRISP-DM Methodology

Satvik Atmakuri

November 1, 2024

Abstract

Predicting heart attack risk is vital for preventative healthcare. This research presents a model developed using the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, a structured, iterative approach designed for data mining projects. Using a machine learning model, we predict heart attack risks based on health data. Each phase of CRISP-DM, from understanding the business problem to model deployment, plays a critical role in ensuring the reliability of predictions. Through systematic data preparation, feature engineering, and model evaluation, the model achieves high accuracy in predicting heart attack risk.

1 Introduction

Heart disease is a leading cause of death worldwide, making early prediction crucial. Predicting heart attack risk using machine learning can improve preventive measures. This paper outlines a heart attack risk prediction model developed using CRISP-DM, which provides a clear, repeatable framework encompassing business understanding, data processing, and model evaluation. We implement machine learning techniques that utilize health indicators as input features, producing accurate predictions on heart attack risk.

2 The CRISP-DM Methodology

The CRISP-DM methodology consists of six phases:

1. **Business Understanding**: Define the problem's scope and objectives.

2. **Data Understanding**: Gather initial data, conduct exploratory data analysis (EDA), and assess data quality.

3. **Data Preparation**: Clean, preprocess, and transform data to create a dataset for modeling.

4. **Modeling**: Select and apply machine learning models and fine-tune them to achieve optimal results.
5. **Evaluation**: Validate model performance against objectives.
6. **Deployment**: Integrate the model into a usable system or provide actionable insights for decision-making.

3 Business and Data Understanding

3.1 Business Understanding

The project goal is to develop a predictive model for heart attack risk. Target users are healthcare professionals who may use this tool for early detection and intervention. A successful model should prioritize high recall to ensure maximum identification of high-risk individuals.

3.2 Data Understanding

The dataset comprises various health indicators (age, cholesterol levels, blood pressure, etc.) known to influence heart disease risk:

- **Data Collection**: Collected health data from multiple sources, ensuring variety and representativeness.
- **Exploratory Data Analysis (EDA)**: EDA was conducted to identify key patterns, relationships, and correlations among features. Initial findings indicated that features like age, cholesterol, and blood pressure had strong correlations with heart attack risk.

4 Data Preparation

Data preparation is essential to improve data quality and model performance. This phase involved:

- **Data Cleaning**: Addressed missing values by imputing mean values for continuous variables and mode values for categorical variables.
- **Outlier Detection and Removal**: Outliers in continuous features such as cholesterol were analyzed and removed to prevent skewing model results.
- **Feature Engineering**: Created new features (e.g., risk categories) and transformed continuous features (e.g., age) into discrete intervals, improving model interpretability.
- **Scaling and Encoding**: Standardized numerical features and one-hot encoded categorical variables to enhance model compatibility.

5 Modeling and Algorithm Selection

Several machine learning models were evaluated, including Logistic Regression, Decision Trees, and Random Forests:

- **Initial Model Selection**: Random Forest was selected for its ability to handle complex interactions and provide high accuracy. Logistic Regression was also used for its interpretability in understanding risk factors.
- **Cross-Validation and Hyperparameter Tuning**: Hyperparameters such as ‘max_depth‘ and ‘n_estimators‘ were fine-tuned using grid search and cross-validation to achieve optimal recall.

6 Evaluation Metrics and Analysis

The model was evaluated based on various metrics to ensure a balanced performance:

- **Accuracy**: Measured overall prediction accuracy, though limited in imbalanced datasets.
- **Recall**: Crucial for identifying high-risk individuals, recall was prioritized to minimize false negatives.
- **Precision**: Ensured that predictions of high-risk cases were likely accurate, reducing false positives.
- **F1-score**: Combined recall and precision to provide an overall performance metric.

6.1 Confusion Matrix Analysis

The confusion matrix provided detailed insights, allowing adjustments to achieve high recall and precision in heart attack risk predictions. Analysis indicated that the tuned model accurately identified high-risk cases with minimal false negatives.

7 Deployment and Integration

To ensure the model's utility, deployment was considered with an emphasis on user accessibility:

- **API Development**: Developed an API for easy integration with healthcare applications, allowing real-time predictions.
- **User Interface**: Designed a user-friendly dashboard displaying predictions, risk scores, and explanations to aid healthcare providers in decision-making.

8 Results and Discussion

The CRISP-DM methodology facilitated a structured approach, leading to a model with high predictive power:

- **High recall** reduced false negatives, effectively identifying high-risk individuals.
- **Balanced precision and recall** ensured that predictions were reliable for healthcare use.

This systematic approach demonstrates CRISP-DM's value in building machine learning models for sensitive applications such as heart disease risk prediction.

9 Conclusion

This study illustrates the successful application of CRISP-DM methodology in predicting heart attack risk. Each phase of CRISP-DM contributed to a robust, high-performing model. Random Forest and Logistic Regression were effective in handling complex health data, achieving high recall and precision. Future work could include exploring additional data sources or advanced algorithms to further enhance model accuracy.

References

[1] Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. Journal of Data Warehousing, 5(4), 13-22.

[2] Goyal, A., Yusuf, S. (2006). The burden of cardiovascular disease in the Indian subcontinent. Indian Journal of Medical Research, 124(3), 235-244.

[3] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.