Predicting Heart Attack Risk using CRISP-DM and Classification

Neel Desai

September 29, 2023

Abstract

In this research, we aim to develop a predictive model that accurately determines the risk of heart attacks in individuals based on various health indicators. Leveraging the CRISP-DM methodology and utilizing a dataset sourced from Kaggle, our primary objective is to construct a binary classification model that achieves over 75% in key performance indicators: accuracy, recall, and precision.

1 Introduction

Heart disease remains one of the leading causes of death worldwide. With advancements in technology and data analytics, there is a growing need for accurate prediction models that can assess the risk of heart attacks in individuals. In this study, we adopt the CRISP-DM methodology, a structured approach to data mining that ensures the rigorous and systematic development of models. Classification, a subset of data mining, is particularly suitable for our task as it allows for the categorization of data into predefined groups. In our case, we aim to classify individuals into two categories: high risk and low risk of suffering a heart attack.

2 Methodologies

2.1 CRISP-DM Methodology

Cross Industry Standard Process for Data Mining (CRISP-DM) is a widely accepted methodology that provides a structured approach to planning and executing data mining projects. It comprises six phases:

1. **Business Understanding:** This phase involves understanding the project objectives, requirements, and translating these into a data mining problem definition. It also includes identifying key performance indicators (KPIs) to evaluate the success of the project.

2. **Data Understanding:** Here, the data is collected, described, and explored to understand its structure, quality, and potential for answering the data mining question.

3. **Data Preparation:** This phase involves tasks like data cleaning, transformation, and feature engineering to make the data suitable for modeling.

4. **Modeling:** In this phase, various algorithms and techniques are applied to the prepared data to build models. It may involve selecting the best algorithm, tuning parameters, and validating models against a training dataset.

5. **Evaluation:** The performance of the model is assessed against a test dataset using the KPIs defined in the business understanding phase. The results are then reviewed to determine if the model meets the business objectives.

6. **Deployment:** Once a satisfactory model is developed, it is deployed in a real-world environment to make predictions or decisions based on new data.

Each phase of CRISP-DM is iterative, allowing for revisiting earlier stages based on findings or challenges encountered in subsequent phases.

3 Implementation

3.1 Alignment with CRISP-DM

1. **Business Understanding:**

   - *Objective Definition:* Our primary aim was to predict if an individual is at high risk of suffering from a heart attack.
   - *KPIs Identification:* The key performance indicators for our model were accuracy, recall, and precision, all set above a threshold of 75%.

2. **Data Understanding:**

   - *Data Collection:* We utilized a dataset from Kaggle containing various health-related features.
   - *Data Exploration:* Preliminary exploration provided insights into feature distributions and the nature of target variables.

3. **Data Preparation:**

   - *Data Cleaning:* With the aid of PyCaret, many data cleaning tasks are automated. For instance, it can handle missing data imputation. An example might be filling missing values in a column using a statistical method like mean or median.

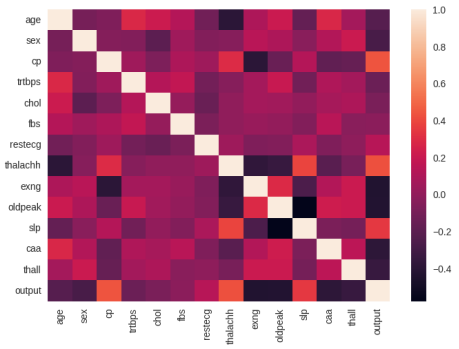

Figure 1: Correlation Matrix

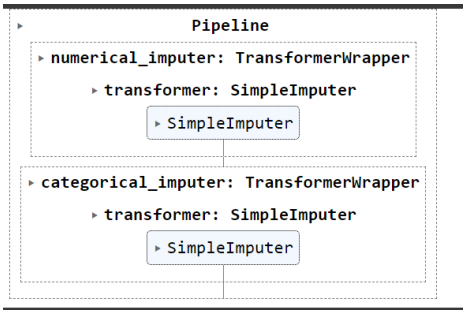

Figure 2: Preprocessing Pipeline by PyCaret

- *Feature Engineering:* PyCaret also aids in generating new features that can enhance model performance. For example, it can automatically create polynomial features based on existing ones to capture non-linear relationships.
- *Data Transformation:* PyCaret streamlines data transformation tasks, ensuring the data conforms to requirements of specific algorithms. A typical transformation is normalization, where feature values are scaled between 0 and 1.

4. **Modeling:**

- *Algorithm Selection:* We employed PyCaret, an AutoML library, to evaluate a range of models, with Naïve Bayes emerging as the top performer.
- *Model Training:* The Naïve Bayes model was trained using the training dataset.

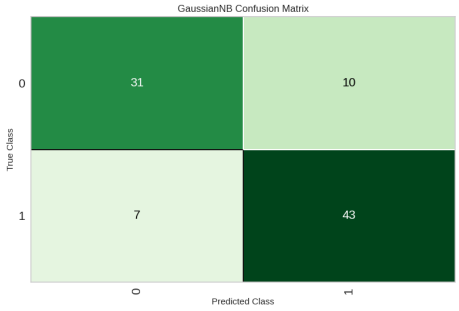

Figure 3: Confusion Matrix

- *Model Tuning:* We made attempts to further tune the model using PyCaret. However, the original model's performance surpassed the tuned versions.

5. **Evaluation:**

   - *Validation Set Evaluation:* The model achieved an accuracy of 82%, precision of 81.8%, and recall of 85.8% on the validation set.
   - *Test Set Evaluation:* On a separate test set, the model yielded an accuracy of 81.32%, precision of 81.13%, and recall of 86%.

3.2 Choosing Evaluation Metrics

For our specific use-case, precision and recall are of greater importance than mere accuracy. While accuracy gauges overall correctness, it might be misleading, especially in datasets with class imbalances. Precision, highlighting the correctness of positive identifications, is pivotal to avert false alarms. Recall indicates the proportion of actual positive cases we captured, ensuring high-risk individuals are not missed.

3.3 Model Finalization

Given the consistent and satisfactory performance across validation and test datasets, we finalized the Naïve Bayes classifier for our predictions regarding heart attack risk.

4 Conclusion

In our endeavor to predict the risk of heart attacks based on health indicators, we adopted the CRISP-DM methodology, ensuring a systematic and comprehensive approach to the problem. By leveraging the capabilities of the PyCaret

AutoML library, we efficiently navigated through the complexities of data preparation, modeling, and evaluation. The Naïve Bayes model emerged as the top performer, achieving our predefined KPIs. Its performance on both validation and test datasets affirmed its robustness and reliability. The emphasis on precision and recall as key evaluation metrics ensured that our model is both cautious and thorough in its predictions, a critical aspect given the life-altering implications of the predictions. As healthcare continues to integrate with data science, methodologies like CRISP-DM and tools like PyCaret will undoubtedly play pivotal roles in driving innovations and improving patient outcomes.

5 References

1. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

2. PyCaret. (n.d.). PyCaret: An open source, low-code machine learning library in Python.

3. Code https://github.com/neel26desai/cmpe255$_{a}ssignment3$

4. Dataset https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.