Heart Disease Analysis and Prediction Report using Logistic Regression

Kacper Koźmin

June 1, 2025

1 Introduction

The aim of this project was to analyze heart disease data and create a predictive model based on logistic regression. The project consisted of two main parts: exploratory data analysis and building and optimizing a machine learning model.

2 Data Description

The utilized dataset `heart_disease_dataset.csv` originated from the UCI Machine Learning Repository and contained information about patients along with various medical and diagnostic parameters. The data was collected as part of cardiological studies conducted in several medical centers.

Data Source: UCI Machine Learning Repository - Heart Disease Dataset

The main variables included:

- **Disease** - target variable (True/False - presence of heart disease)
- **Age** - patient age
- **Sex** - patient gender (male/female)
- **Chest.pain.type** - type of chest pain
- **Serum.cholesterol.in.mg.dl** - serum cholesterol level
- **Maximum.heart.rate.achieved** - maximum heart rate achieved
- **Exercise.induced.angina** - exercise-induced angina
- **ST.depression.induced.by.exercise.relative.to.rest** - ST depression induced by exercise
- **Number.of.major.vessels** - number of major blood vessels

3 Data Processing

3.1 Categorical Variable Conversion

All categorical variables were converted to numerical format suitable for modeling:

- **Disease**: True=1 (disease present), False=0 (healthy)
- **Sex**: male=0, female=1
- **Fasting.blood.sugar**: False=0, True=1 (>120 mg/dl)
- **Exercise.induced.angina**: False=0, True=1

3.2 Correlation Analysis

Correlation analysis was performed between all numerical variables. The top 5 features most correlated with the target variable (Disease) were identified, which helped in feature selection for the model.

4 Methodology

4.1 Data Split

The dataset was divided into:

- **Training set**: 70% of data (used for model training)
- **Test set**: 30% of data (used for model evaluation)

`set.seed(100)` was used to ensure reproducibility of results.

4.2 Feature Selection

The model was built using the following features:

- ST.depression.induced.by.exercise.relative.to.rest
- Exercise.induced.angina
- Chest.pain.type
- Age
- Sex
- Maximum.heart.rate.achieved
- Number.of.major.vessels

Features were selected based on correlation analysis and domain knowledge about heart disease risk factors.

5 Results

5.1 Confusion Matrix and Basic Metrics

The model achieved the following results on the test set with a decision threshold of 0.5:

Table 1: Model performance metrics at threshold 0.5

Metric	Value
Accuracy	0.7568
Precision	0.7949
Recall (Sensitivity)	0.8387
Specificity	0.6486
F1-Score	0.8163

5.2 ROC Curve Analysis

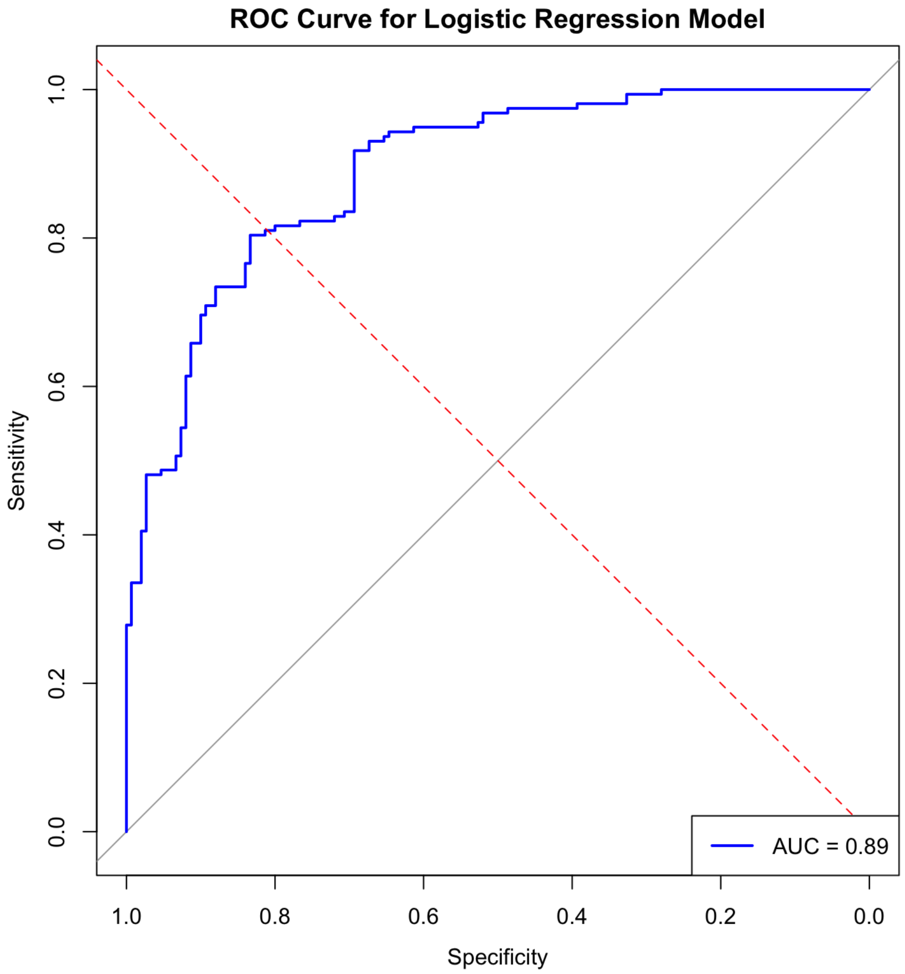

Figure 1: ROC curve for logistic regression model

The model achieved AUC = 0.89, indicating very good discriminatory ability. An AUC value above 0.8 is considered very good in the context of medical diagnostics.

5.3 Probability Distribution

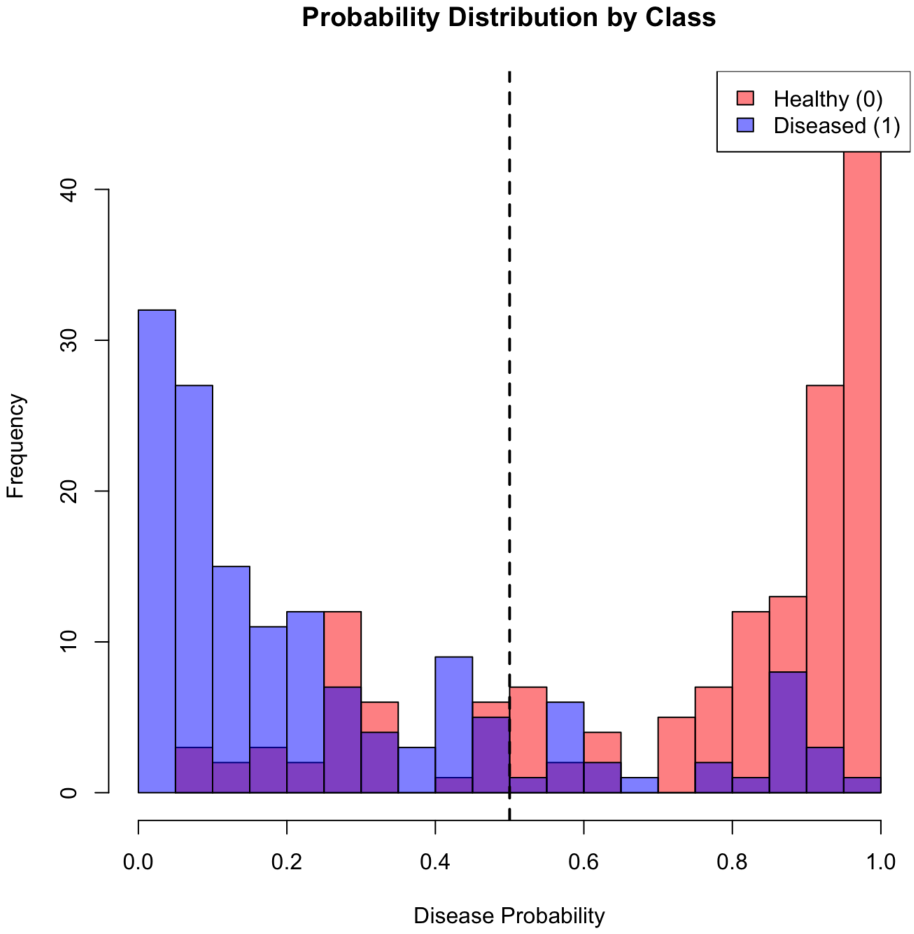

Figure 2: Probability distribution by class

The histogram shows clear separation of probabilities between classes of healthy and diseased patients, confirming good model quality. The dashed line at 0.5 represents the default decision threshold.

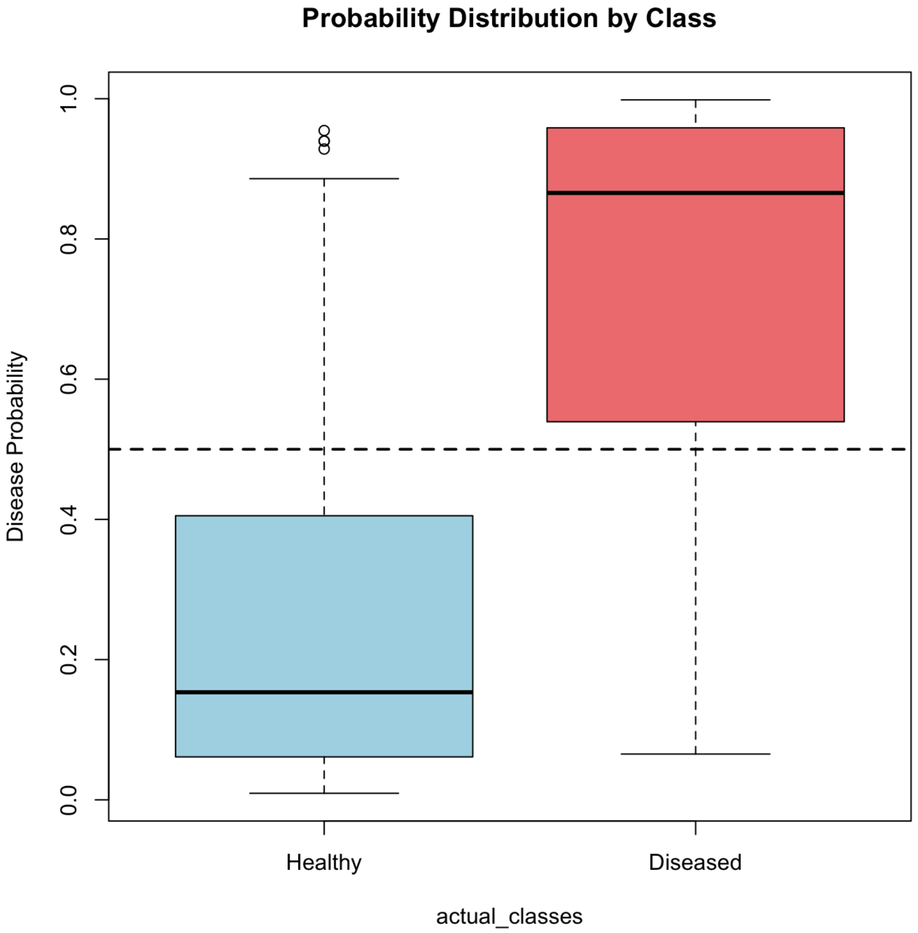

Figure 3: Box plot of probability distribution by class

The box plot confirms clear differences in probability distribution between groups, with the median for diseased patients significantly above 0.5.

6 Decision Threshold Optimization

6.1 Analysis of Different Thresholds

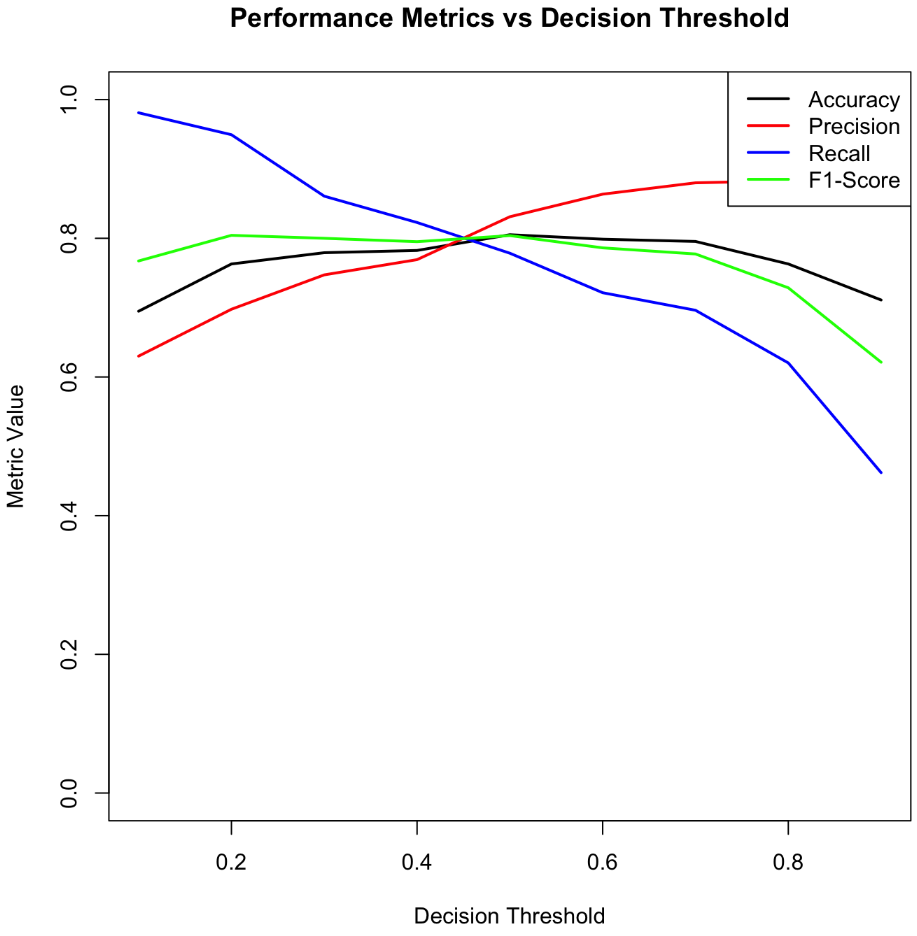

Figure 4: Performance metrics as a function of decision threshold

Performance analysis was conducted for different decision thresholds (0.1 - 0.9). Results show that:

- **Accuracy** reaches maximum around threshold 0.2-0.3
- **Recall** decreases with increasing threshold
- **Precision** increases with increasing threshold
- **F1-Score** reaches maximum at threshold 0.2

6.2 Optimal Decision Threshold

Table 2: Comparison of metrics for different decision thresholds

Metric	Threshold 0.5 (default)	Threshold 0.2 (optimal)
Accuracy	0.757	0.784
Precision	0.795	0.714
Recall	0.839	0.968
F1-Score	0.816	0.823

The optimal threshold is 0.2, meaning a patient is classified as diseased if the probability is $\geq$20%. This threshold was chosen based on F1-Score maximization.

7 Interpretation and Conclusions

7.1 Choice of Low Decision Threshold

Using a low threshold (0.2) instead of the standard (0.5) is justified in the medical context:

1. **Cost of errors**: False negative results (missing disease) have much more serious consequences than false positives
2. **High sensitivity**: At threshold 0.2, the model detects 96.8% of disease cases
3. **Screening**: The model can serve as a preliminary screening tool

7.2 Model Strengths

- High AUC (0.89) indicates very good discriminatory ability
- Good probabilistic interpretation of results
- Stable results across different decision thresholds
- Use of clinically relevant predictors

7.3 Limitations

- Relatively small dataset may limit generalization
- Model may require validation on an independent dataset
- Precision at low threshold is lower (71.4%)

8 Summary

The developed logistic regression model demonstrates very good predictive properties for heart disease detection. AUC equal to 0.89 and decision threshold optimization allow achieving high sensitivity (96.8%) with acceptable precision.

The key element of success was proper data processing, thoughtful feature selection, and decision threshold optimization considering the medical specifics of the model application.