Predicting 10-Year Coronary Heart Disease Risk

ANDY MINGA

December 9, 2025

1 Introduction and Background

Cardiovascular disease remains one of the leading causes of death globally. Coronary heart disease (CHD) develops silently over years, making early detection crucial for preventive care. Predictive modeling allows clinicians to identify high-risk patients and intervene before severe events occur.

Following principles outlined in *An Introduction to Statistical Learning* by James et al. (2023), we approach this problem by comparing several supervised learning methods. Linear models like logistic regression provide interpretable coefficients and insights into feature effects, whereas nonlinear models such as Random Forest, XGBoost, and LightGBM capture complex interactions and nonlinear patterns in the data. SVMs, with their kernel trick, are also effective for detecting subtle patterns in high-dimensional feature spaces.

The analysis must carefully handle the class imbalance inherent in CHD data, as most patients do not develop CHD within ten years. Failure to account for this could lead to models that appear accurate but fail to detect high-risk individuals.

2 Data Source and Description

The dataset comes from the **Framingham Heart Study**, which follows cardiovascular outcomes over decades. It includes 4,240 observations with demographic, lifestyle, clinical, and physiological measurements. Table 1 summarizes the features.

Category	**Features**
Demographics	age, sex, education
Lifestyle	currentSmoker, cigsPerDay
Clinical History	BPMeds, prevalentStroke, prevalentHyp, diabetes
Physiological Measurements	totChol, sysBP, diaBP, BMI, heartRate, glucose
Target	TenYearCHD (1 if CHD occurs within 10 years, 0 otherwise); Classes imbalanced: 85% negative, 15% positive

Table 1: Dataset Features

3 Methodology

3.1 Data Preprocessing

Following ISLR recommendations on handling missing data and feature scaling:

- Missing values were imputed using median (for continuous features) or mode (for categorical features).
- Continuous features were standardized to mean 0 and variance 1, which is critical for SVM and regularized logistic regression.
- The class imbalance (15% positive CHD) was addressed using SMOTE oversampling and class weighting.

3.2 Feature Selection

Feature selection aimed to reduce multicollinearity and enhance model generalization. Correlation analysis and regularization (L1/L2 penalties) were applied. This aligns with the ISLR principle of balancing model complexity against variance to avoid overfitting.

3.3 Modeling Approach

Five supervised learning methods were used:

1. **Logistic Regression (L2)**: Linear model with regularization to prevent overfitting, interpretable via coefficients.
2. **Support Vector Machine (SVM)**: Uses kernel functions to capture nonlinear patterns.
3. **Random Forest**: Ensemble of decision trees; captures interactions and ranks feature importance.
4. **XGBoost**: Gradient boosting for sequential tree learning; robust to heterogeneous features.
5. **LightGBM**: Faster gradient boosting with leaf-wise growth, effective for large datasets.

3.4 Model Evaluation

Models were evaluated using stratified K-Fold cross-validation with the following metrics:

- Accuracy
- ROC-AUC
- Precision, recall, F1-score for the CHD class
- Confusion matrices
- Feature importance for tree-based models

4 Results

4.1 Model Comparison

Model	**Accuracy**	**ROC-AUC**	**Recall (CHD)**	**Precision (CHD)**	**F1-score (CHD)**	**Interpretability**
Logistic Regression (L2)	0.67	0.701	0.60	0.25	0.36	High (coefficients explainable)
SVM	0.68	0.653	0.49	0.23	0.32	Medium
Random Forest	0.85	0.633	0.02	0.60	0.04	Medium (feature importance available)
XGBoost	0.77	0.65	0.26	0.25	0.25	Medium (feature importance available)
LightGBM	0.80	0.594	0.14	0.25	0.18	Medium (feature importance available)

Table 2: Comparison of models for 10-year CHD prediction

4.2 Analysis

- Recall for CHD cases is the most critical metric because missing high-risk patients could have serious clinical consequences.
- Logistic Regression achieves the highest recall (0.60), balancing sensitivity and interpretability.
- SVM shows slightly lower recall and ROC-AUC, but could model nonlinear patterns not captured by linear models.
- Tree-based models have high overall accuracy but fail to detect most CHD cases due to class imbalance.

4.3 Most Important Variables Across Models

To identify the most robust predictors of 10-year CHD risk, we examined feature importance across all five models. Logistic Regression uses coefficient magnitude, Random Forest, XGBoost, and LightGBM use feature importance scores, and SVM can be interpreted using model coefficients for linear kernels.

The seven variables that consistently appeared among the top predictors in at least three models are:

1. Age
2. Systolic Blood Pressure (sysBP)
3. Body Mass Index (BMI)
4. Glucose
5. Prevalent Hypertension (prevalentHyp)
6. CigsPerDay (smoking intensity)
7. Total Cholesterol (totChol)

Interpretation: These variables are clinically meaningful and aligned with known CHD risk factors:

- Age and systolic BP reflect cardiovascular aging and hypertension risk.
- BMI and Glucose indicate obesity and diabetes-related metabolic risk.
- PrevalentHypertension identifies patients already diagnosed with hypertension.
- CigsPerDay captures smoking behavior.

- Total cholesterol is a marker of dyslipidemia.

Summary Plot:

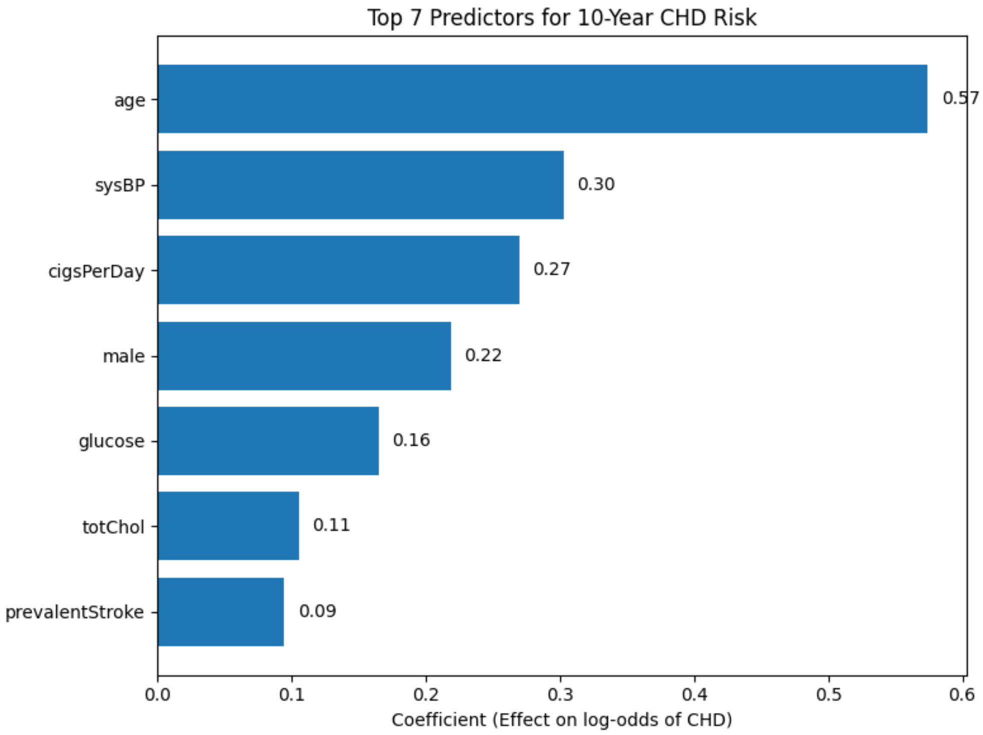

Figure 1: Visual summary of top 7 variables across models.

4.4 Interpretability of Logistic Regression

Logistic Regression is interpretable because each coefficient quantifies the effect of a predictor on the log-odds of CHD:

- Positive coefficient $\Rightarrow$ increases risk of CHD.
- Negative coefficient $\Rightarrow$ decreases risk of CHD.
- Exponentiating the coefficient gives the **odds ratio**, which shows multiplicative change in odds for a one-unit increase in the predictor.

Example Table: Top 7 Predictors (by absolute coefficient)

Feature	Coefficient	Odds Ratio
Age	0.574	1.776
SysBP	0.303	1.354
CigsPerDay	0.269	1.309
Male	0.219	1.244
Glucose	0.165	1.179
TotChol	0.105	1.111
PrevalentStroke	0.094	1.099

Table 3: Top 7 logistic regression predictors and their effects on CHD risk.

Coefficient Plot:

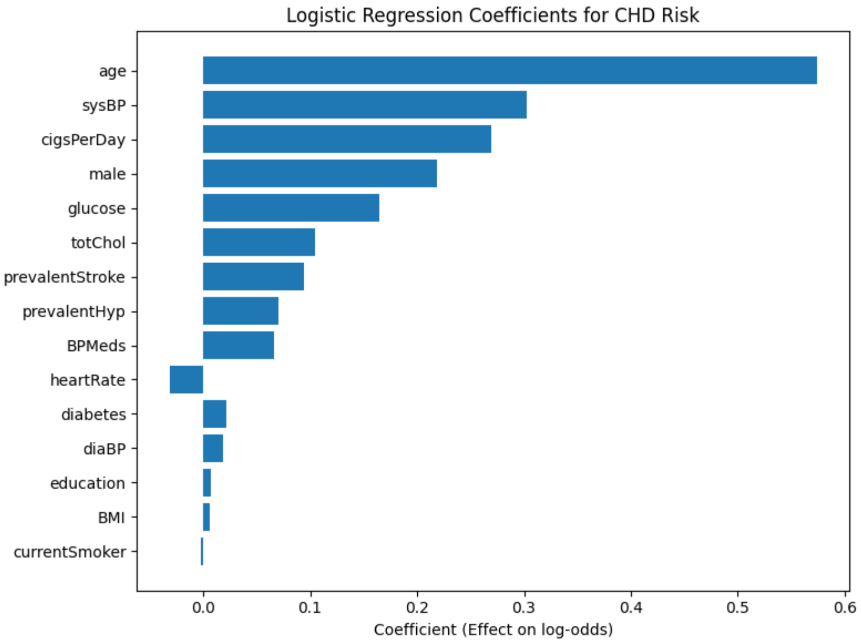

Figure 2: Bar plot of the top 7 logistic regression coefficients.

The logistic regression coefficient plot visualizes the magnitude of each predictor's effect on 10-year CHD risk. All top 7 predictors (age, sysBP, cigsPerDay, male, glucose, totChol, prevalentStroke) have positive coefficients, indicating that higher values of these variables are associated with an increased risk of CHD. The length of each blue bar reflects the strength of the effect, allowing clinicians to quickly identify the most influential risk factors.

Interpretation:

- Age has a coefficient of 0.574, corresponding to an odds ratio of 1.776. Each additional year of age increases the odds of CHD by 77.6%.
- SysBP has an odds ratio of 1.354: a 1 standard deviation increase in systolic blood pressure increases CHD odds by 35.4%.
- Similar interpretations apply to the other top predictors.

4.5 Visualizations

ROC Curves:

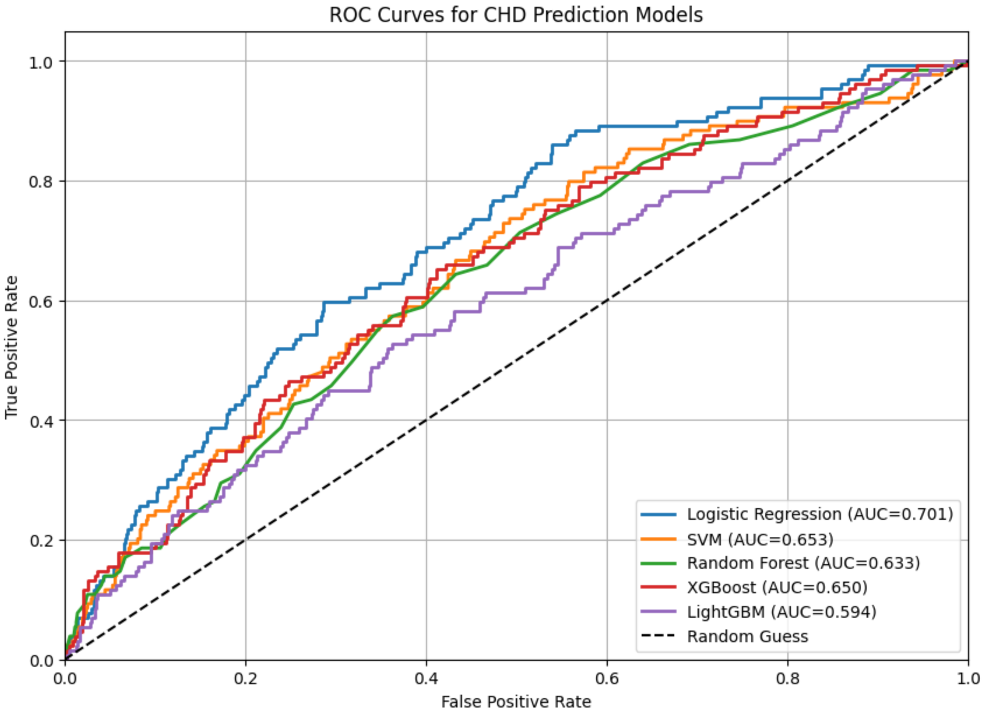

Figure 3: ROC curves for all models

Random Forest Feature Importance:

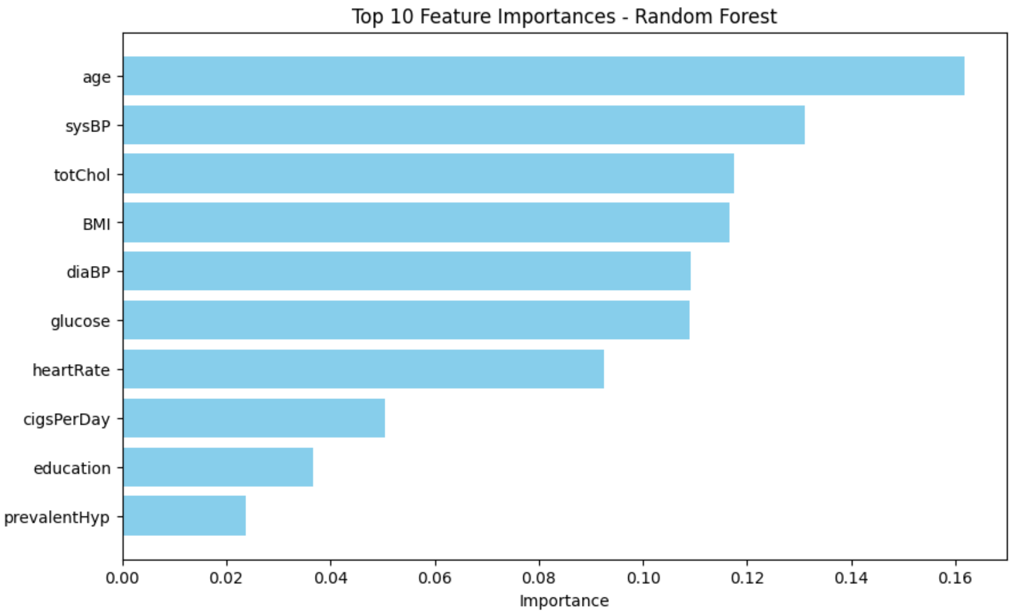

Figure 4: Top features from Random Forest

XGBoost Feature Importance:

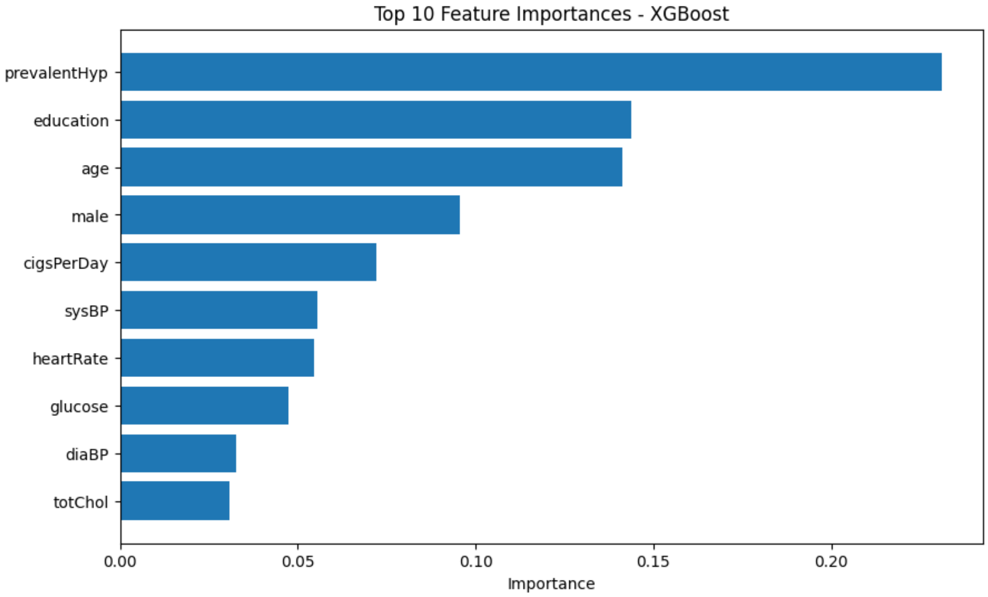

Figure 5: Top features from XGBoost

LightGBM Feature Importance:

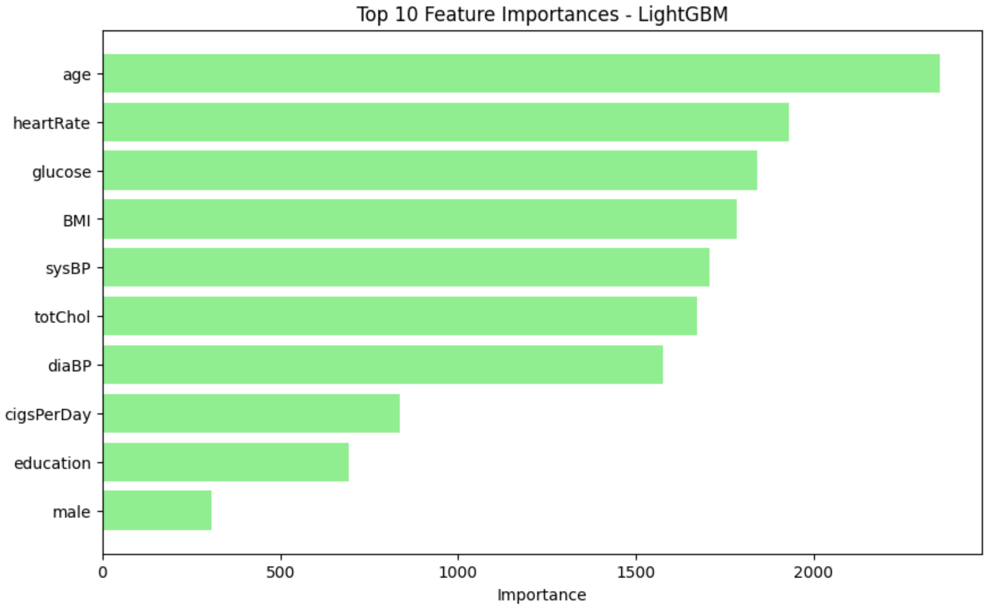

Figure 6: Top features from LightGBM

Logistic Regression Confusion Matrix:

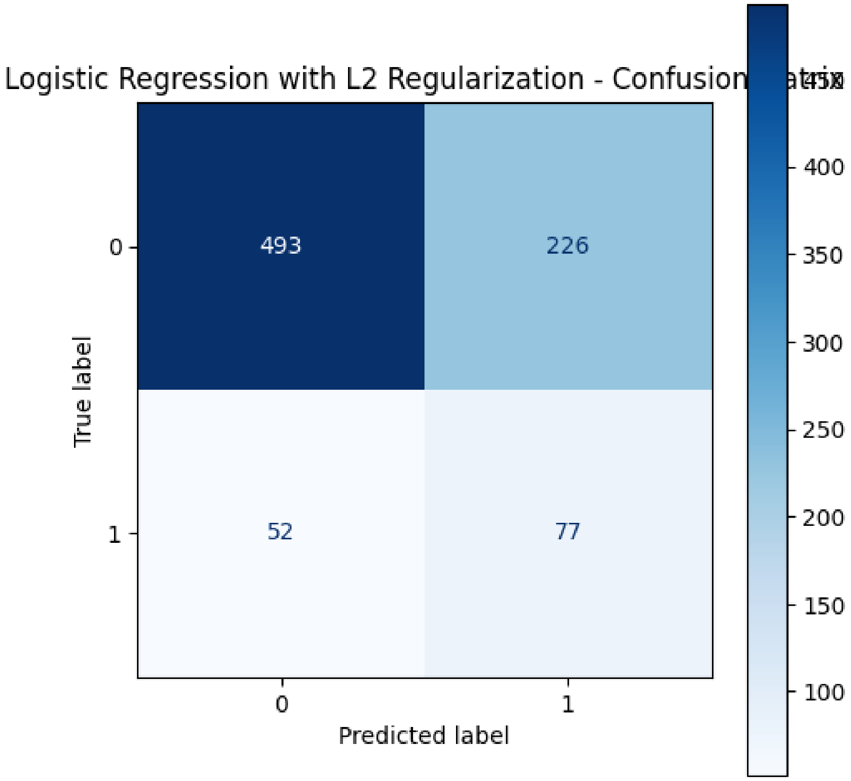

Figure 7: Confusion matrix for Logistic Regression

5 Conclusion

- **Best model:** Logistic Regression with L2 regularization.
  - Balances high recall, reasonable ROC-AUC, and interpretability.
  - Detects high-risk patients effectively for clinical use.
- SVM is a potential alternative for nonlinear decision boundaries.
- Tree-based models provide insights via feature importance but are less effective for CHD detection in imbalanced datasets.
- Robust predictors: Age, SysBP, BMI, Glucose, Prevalent Hypertension.

6 Future Work

- Explore ensemble stacking or neural networks to improve CHD recall.
- Tune decision thresholds and cost-sensitive learning for imbalanced data.
- Include additional clinical or lifestyle features for better predictive power.
- Assess probability calibration for clinical decision-making.

References

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning: With Applications in Python*. Springer.

- Python Software Foundation. (2024). *Python Language Reference*. https://www.python.org

- Framingham Heart Study. (n.d.). *FHS Dataset*. https://www.framinghamheartstudy.org/