Predicting 10-Year Coronary Heart Disease Risk

Andy Minga

December 9, 2025

Outline

Background

- Cardiovascular disease is a leading cause of death worldwide.
- CHD develops silently, making early detection essential.
- Predictive modeling helps identify high-risk patients for preventive care.
- This project follows statistical learning principles from ISLP (2023).

Goal of the Study

Objective: Build a predictive model for 10-year CHD risk.

- Compare linear and nonlinear supervised learning models.
- Understand which variables contribute most to CHD risk.
- Address the significant class imbalance.
- Recommend the most clinically useful model.

Data Source

Framingham Heart Study

- 4,240 observations
- Demographics, lifestyle habits, medical history, physiological measures
- Target: **TenYearCHD**

Class Imbalance Problem

Distribution of Target Variable:

- Negative (no CHD in 10 years): 85%
- Positive (CHD in 10 years): 15%

Why it's a problem:

- Models may learn to always predict "no CHD."
- High accuracy but terrible recall for CHD cases.
- Clinically dangerous because high-risk patients are missed.

Fixing the Class Imbalance

Techniques used:

- **SMOTE Oversampling**: synthetically increases minority CHD cases.
- **Class Weighting**: penalizes misclassification of CHD more heavily.

Why these work:

- SMOTE prevents the model from simply memorizing repeated minority samples.
- Class weighting forces the model to pay more attention to CHD prediction.
- Together, they improve recall without dramatically hurting precision.

Why These Models? (Based on Data Nature)

The dataset has:

- Mix of continuous and categorical variables
- Many correlated medical features
- Moderate size (4k rows)
- Nonlinear relationships (e.g., BP, glucose)

Therefore, we chose:

- **Logistic Regression** — baseline, interpretable, handles linear structure well.
- **SVM** — captures nonlinear boundaries without needing many features.
- **Random Forest** — handles interactions, robust to noisy medical data.
- **XGBoost / LightGBM** — strong performance for tabular data with nonlinearities.

Why Not Use Other Models?

- Neural networks need larger datasets and less multicollinearity.
- KNN performs poorly with standardized clinical variables and imbalanced labels.
- Naive Bayes assumes independence, which is unrealistic for medical variables.

Model Evaluation Metrics

Metrics Used:

- Accuracy
- Recall (most important clinically)
- Precision, F1-score
- ROC-AUC
- Confusion Matrix

Why recall matters most: Missing a patient at risk of CHD is far worse than flagging a false positive.

Model Comparison

Model	Acc.	AUC	Recall	F1
Logistic Regression	0.67	0.701	**0.60**	0.36
SVM	0.68	0.653	0.49	0.32
Random Forest	0.85	0.63	0.02	0.04
XGBoost	0.77	0.65	0.26	0.25
LightGBM	0.80	0.59	0.14	0.18

Reasons:

- Linear structure aligns well with medical risk factors.
- Less affected by class imbalance after weighting.
- Coefficients are interpretable $\rightarrow$ important for clinical decisions.
- Avoids overfitting, which tree models struggle with on imbalanced data.

Most important: It achieves the highest recall for CHD cases.

Why Tree-Based Models Underperform in Recall

- They maximize overall accuracy, not minority-class detection.
- Splitting criteria (Gini/Entropy) favor the majority class.
- Even with SMOTE, trees still lean heavily toward predicting "no CHD."
- High accuracy is misleading due to the 85% negative class.

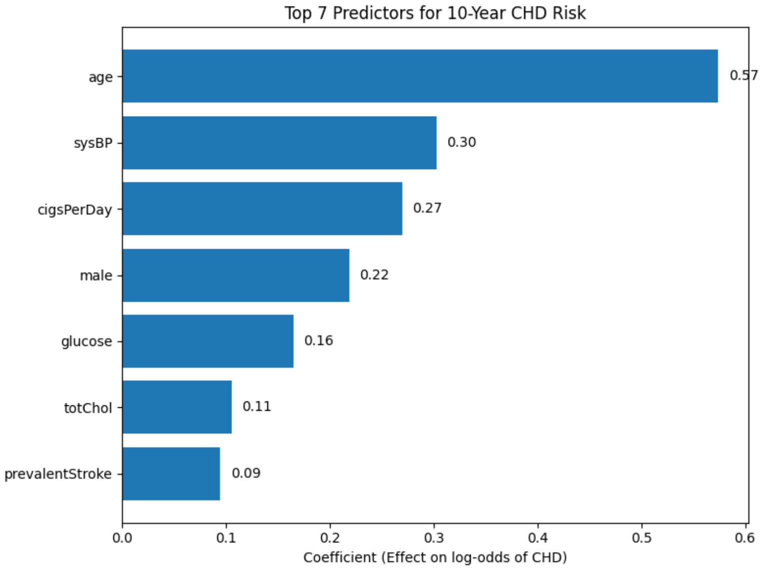
Why SVM Performs Moderately Well

- Captures nonlinear patterns better than Logistic Regression.
- Performs well on medium-sized datasets.
- But suffers from:
    - Sensitivity to scaling
    - Difficulty handling imbalanced data
    - Less interpretability

Top Predictors Across All Models

- Age
- SysBP
- BMI
- Glucose
- Prevalent Hypertension
- CigsPerDay
- Total Cholesterol

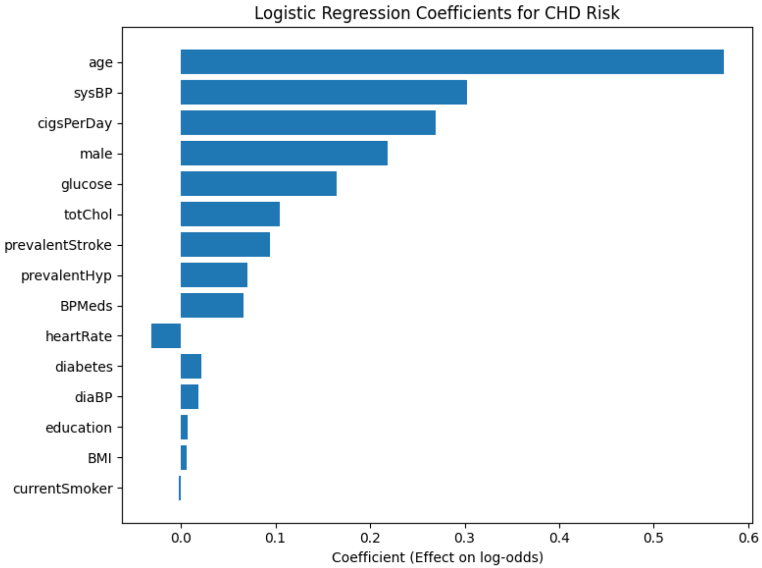
Visual Summary of Top Predictors



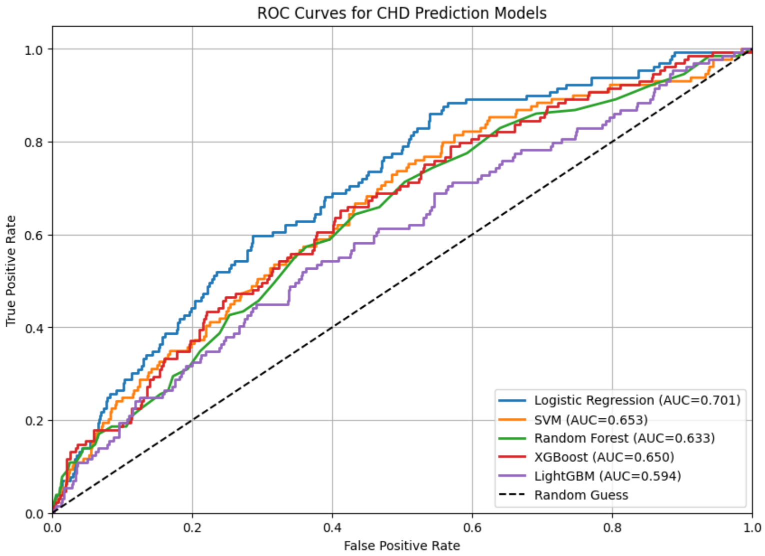
Logistic Regression Coefficients

- Positive coefficient $\rightarrow$ increased CHD risk
- Odds ratios show multiplicative effects
- Interpretation is clinically intuitive

Coefficient Plot

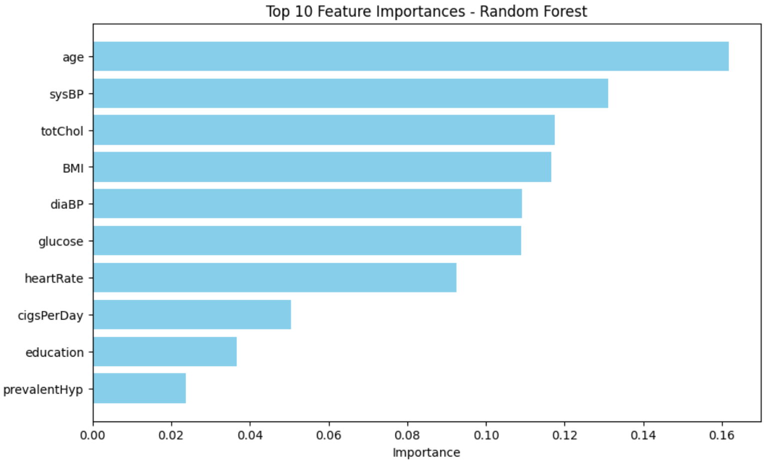


ROC Curves

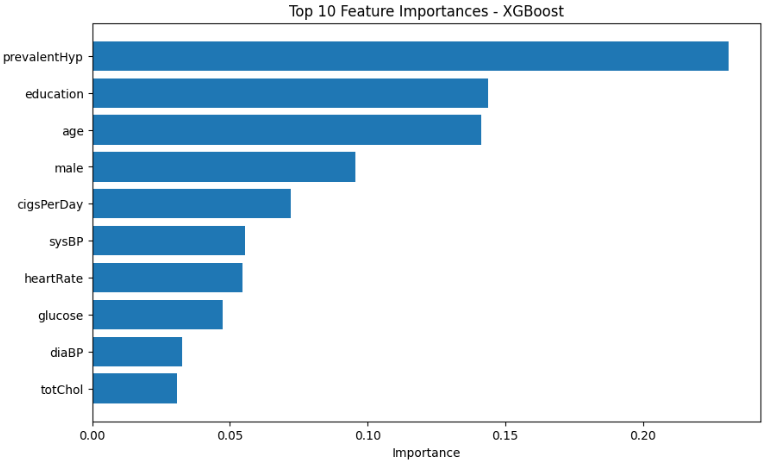


ROC Curves for CHD Prediction Models
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (AUC=0.701)
SVM (AUC=0.653)
Random Forest (AUC=0.633)
XGBoost (AUC=0.650)
LightGBM (AUC=0.594)
Random Guess

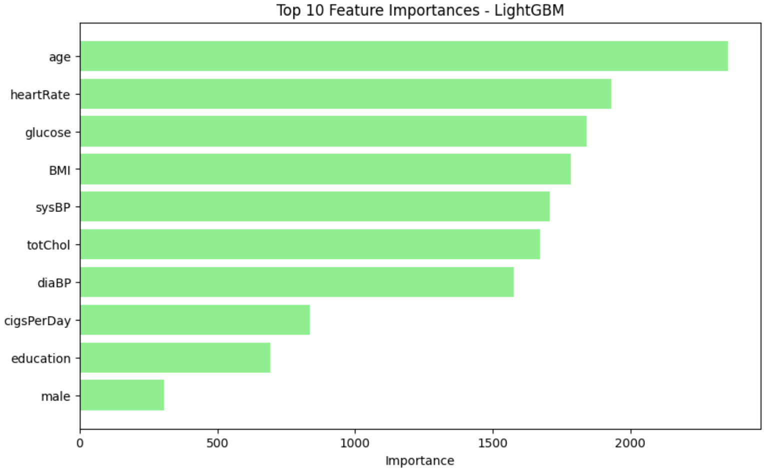
Random Forest Importance



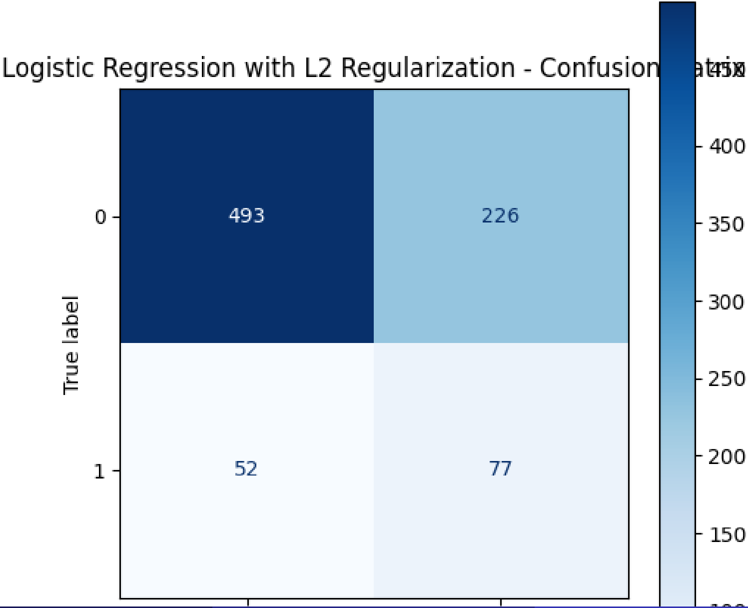
XGBoost Importance



LightGBM Importance



Confusion Matrix: Logistic Regression



Conclusion

- **Best Model: Logistic Regression**
  - Highest recall (most important in medicine).
  - Interpretable — doctors need explanations, not black-box outputs.
  - Works well with class weighting.
- SVM is acceptable but less interpretable.
- Tree-based models predict majority class too often.
- Key predictors: Age, SysBP, BMI, Glucose, PrevalentHypertension.

Future Directions

- Explore ensemble stacking or neural networks.
- Use cost-sensitive loss to boost minority-class recall.
- Apply threshold tuning and calibration for clinical use.

References

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

Python Software Foundation (2024). *Python Language Reference*.