




Predicting 10-Year CHD Risk Using Logistic Regression

A data-driven approach to identifying patients at risk for coronary heart disease using logistic regression modeling with a focus on clinical interpretability and addressing class imbalance challenges.

 **por Jose**



Project Objective

Build Predictive Model

Develop a logistic regression model to predict 10-year coronary heart disease risk using patient health data.

Ensure Medical Interpretability

Create a model that produces results healthcare professionals can understand and trust.

Handle Class Imbalance

Address the challenge of uneven distribution between positive and negative CHD cases.

Dataset Features



Demographics

- Age
- Sex



Behavioral

- Smoking status
- Cigarettes per day



Medical History

- Hypertension
- Diabetes
- Stroke



Clinical

- Blood pressure
- Cholesterol
- Glucose
- BMI





Data Preprocessing

Handle Missing Values

Missing data imputed using median for numerical variables and mode for categorical variables.

Manage Outliers

Extreme values clipped to 1st-99th percentile range to reduce model bias.

Standardize Features

All numerical features scaled to have mean=0 and standard deviation=1.

Split Dataset

Data divided into 80% training and 20% testing sets for proper validation.

Modeling Approach



Balanced Class Weights

Implemented
'class_weight=balanced'
parameter to address the
imbalance between CHD
and non-CHD cases.



Threshold Optimization

Compared default (0.5) and
optimal thresholds to
maximize model
performance.



Key Metrics Focus

Prioritized Recall and F1 Score to ensure high sensitivity for
at-risk patients.



Performance meting for heart disease whidolotional comfigmand

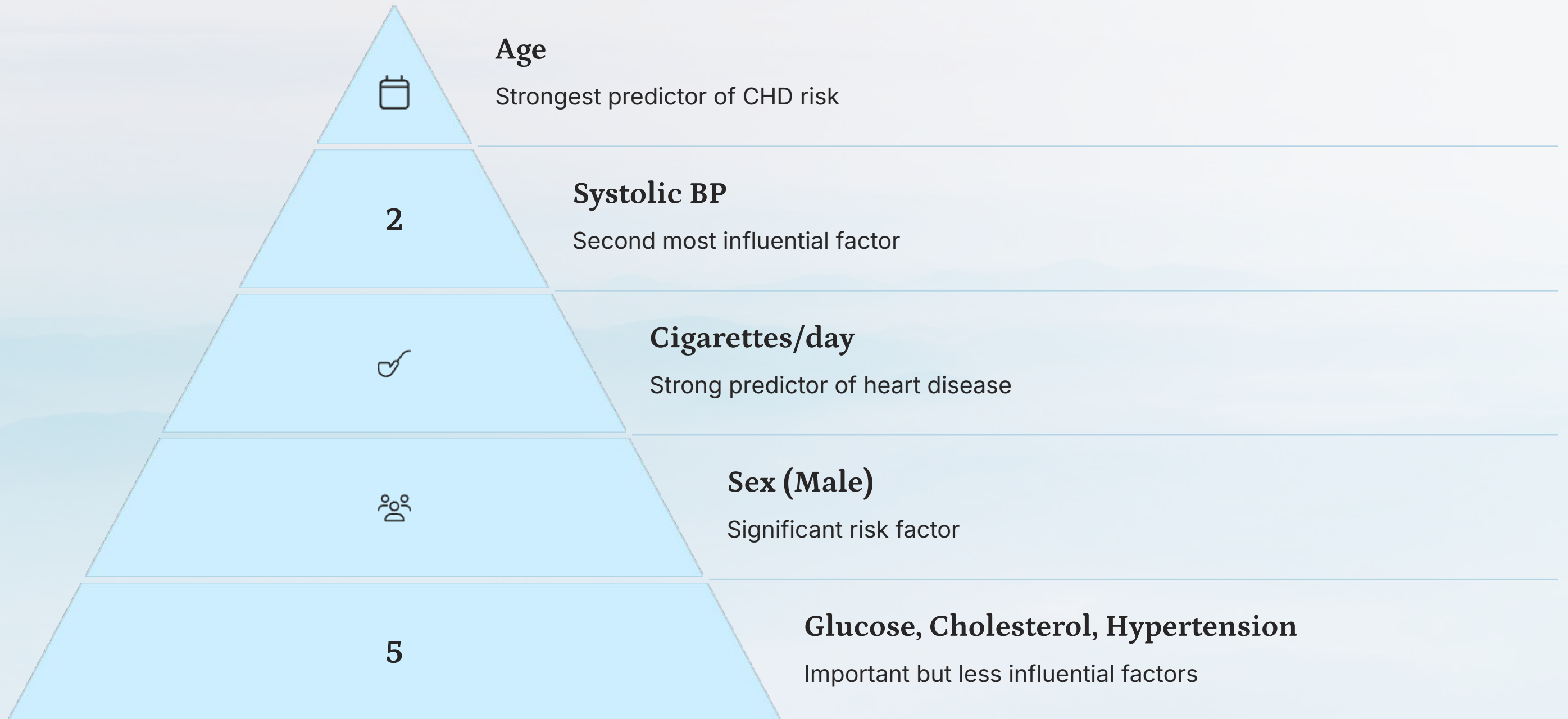


Performance Results

Metric	Default (0.5)	Optimal (~0.51)
Accuracy	71.5%	73.0%
Precision	29.8%	31.0%
Recall	65.7%	64.7%
F1 Score	0.41	0.42
ROC AUC	0.725	0.725

The model shows balanced performance with minimal threshold adjustment needed, indicating robust predictive capability.

Feature Importance



Clinical Interpretation

Balanced Model

Logistic regression with balanced class weights performed effectively

Clinical Utility

Ideal for integration into CHD screening protocols



Strong Detection

Achieved over 65% recall at default threshold

Medically Aligned

Key predictors match established clinical risk factors



Limitations and Next Steps



Advanced Models

Explore non-linear approaches including Random Forest and XGBoost for potentially higher accuracy.



Expanded Features

Incorporate longitudinal data and additional lab markers to improve predictive power.



External Validation

Test model performance on independent datasets from different patient populations.