Predicting 10-Year Risk of Coronary Heart Disease (CHD)

•••

Team 6: Alex Foster, Danae Vassiliadis, Natalie Kim, Mathew Spencer

March 4, 2024

Agenda

- Overview
- Exploratory Data Analysis
- Feature Selection
- Modeling Decisions
- Conclusion

Overview

Project Origin

Stakeholder

- Framingham General Hospital
- Chief of Cardiology: Patrick Fisher, MD, PhD
- Framingham, Massachusetts
- Population 71,265 (2021)

Challenge

- Framingham General's financial performance is struggling
- The hospital has identified an elevated rate of emergency care visits associated with heart attack, heart failure, and stroke victims as a key driver
- It wants to boost its mix of higher-margin preventive care business by targeting patients with elevated risk of coronary heart disease (CHD)

Analytical Plans & Goals

Our goal is to utilize a mix of statistical and machine learning models to **identify** patients with an elevated risk of CHD, **understand** which features are associated with that risk, and **enable** Framingham General to deploy targeted preventative care

Exploratory Data Analysis

Data Overview

- 4,238 rows
- 1 response variable
    a. 'TenYearCHD' = 10-year risk of coronary heart disease (0=no, 1=yes)
    b. 15.2% of raw data set is at risk
- 15 potential features

Sex	Current smoker	Prevalent stroke	Total cholesterol	BMI
Age	Cigarettes per day	Prevalent hypertension	Systolic blood pressure	Heart rate
Education	Blood pressure medication	Diabetes	Diastolic blood pressure	Glucose

Data source: https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression/data

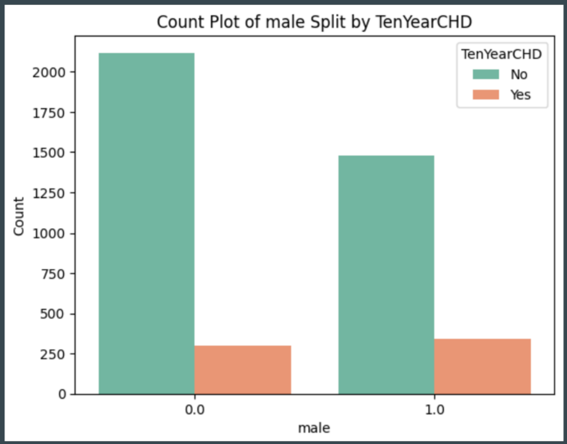
Data Pre-Processnig

- 7 columns with missing values
- Option 1: eliminate all rows with missing values (data reduction of ~9%)
- Option 2: impute missing values
  a. KNNImputer from Scikit-Learn
  b. Utilizes the k-nearest neighbors algorithm to replace missing values with the mean value of similar instances (as determined by distance in the feature space)

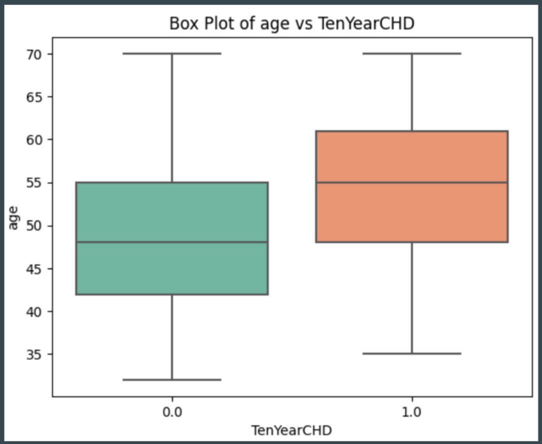
```
Columns with missing values:
education     105
cigsPerDay     29
BPMeds         53
totChol        50
BMI            19
heartRate       1
glucose       388
```

Documentation: https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html

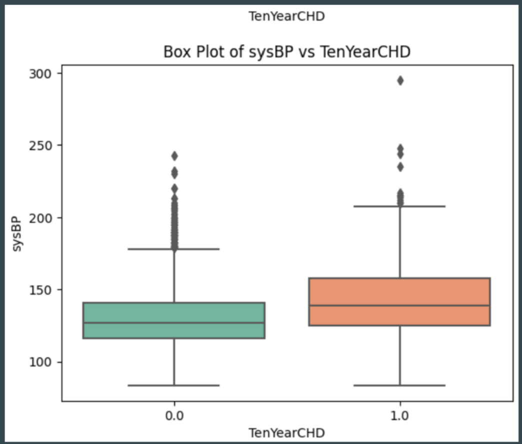
Patterns in the Data



53% of at-risk patients are male



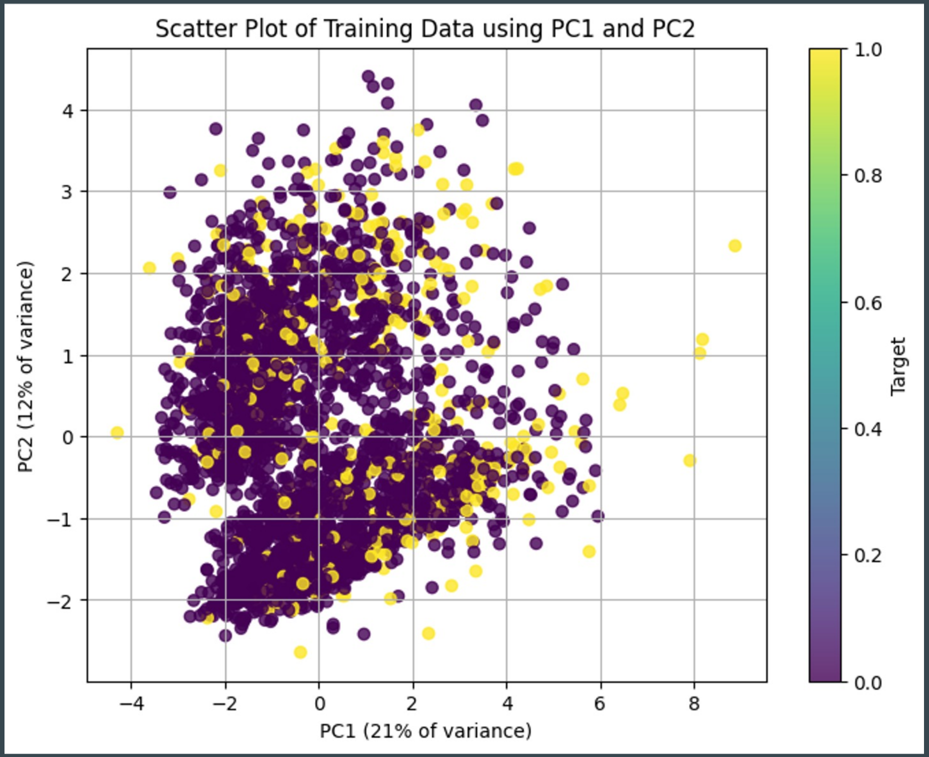
At-risk patients are 11% older on average



At-risk patients have 10% higher systolic blood pressure

Principal Components Analysis (PCA)

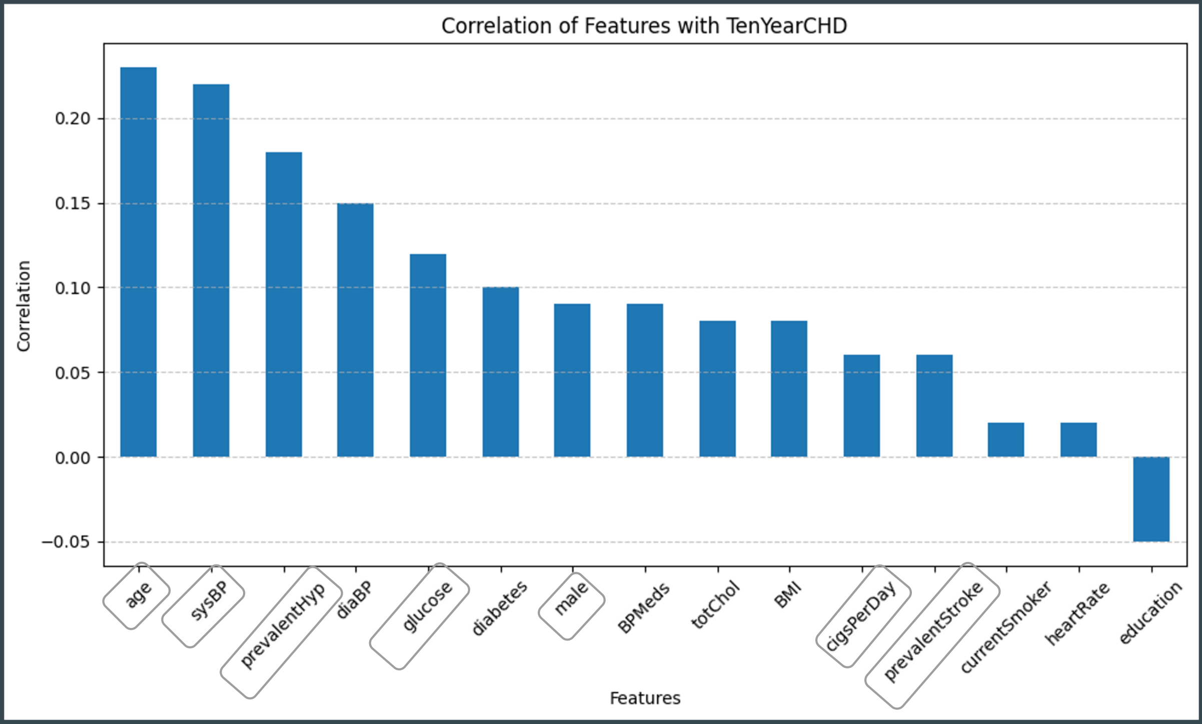
- PC1 and PC2 only explain 33% of variance in the data, but we can still extract insight
- PC1 has positive associations with age, BPMeds, prevalentHyp, sysBP, diaBP, and BMI.
- So patients that are older, take blood pressure medication, have hypertension, have higher blood pressure, and higher BMIs tend to have higher values for PC1.
- Yellow dots = at-risk patients
    a. (note: they're generally further to the right)



Feature Selection

Feature Selection

- First we analyzed a correlation matrix and isolated the relationship with our response variable
- Then we conducted recursive feature elimination (RFE) with a logistic regression estimator to isolate the best features



Documentation: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

Modeling Decisions

General Approach

Step 1 — Split data

Split data set into training (70%) and test (30%)

Step 2 — Build model

Fit independent classification models to training data

Step 3 — Optimize performance

Improve recall scores by tuning hyperparameters with GridSearchCV, evaluating scree plots, etc.

Step 4 — Validate results

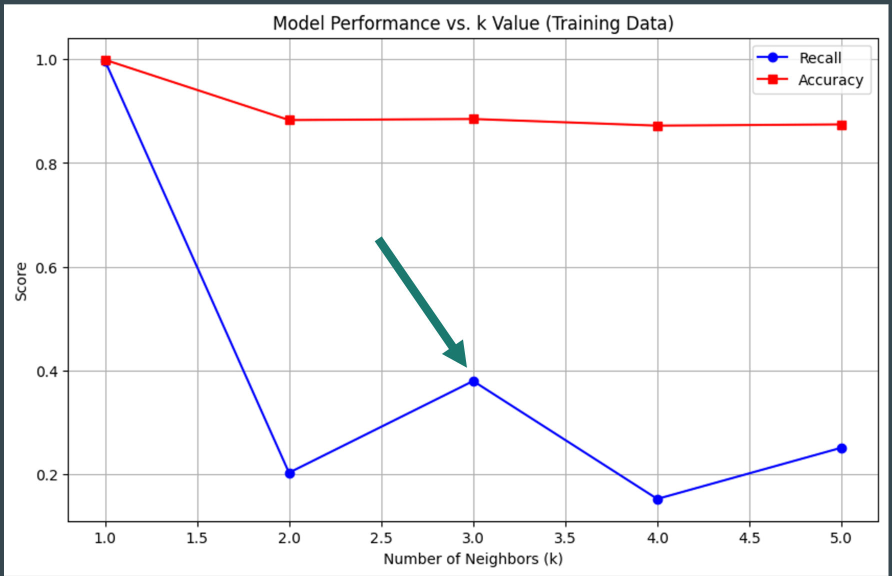
Apply model to test data

Step 5 — Iterate

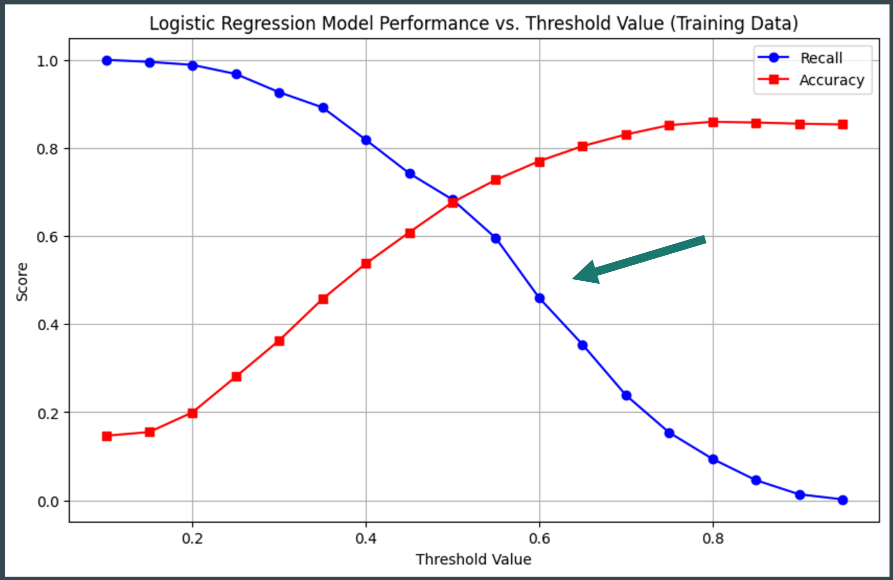
Repeat steps 2-4 as needed

Documentation: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Individual Model Optimization



Selecting optimal number of neighbors for KNN model



Applying appropriate threshold for logistic regression model

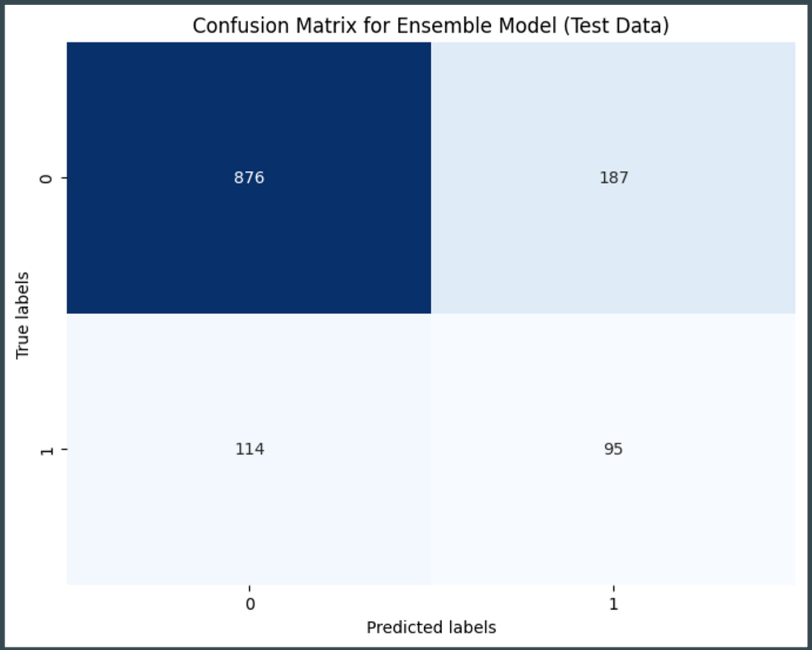
Individual Model Comparison

- Best accuracy: AdaBoost (83.9%)
- Best precision: AdaBoost (66.7%)
- Best recall: Radial Basis Function Kernel SVM (63.6%)
- Best F1: Random Forest (40.0%)
- Build Ensemble Model from individual models
    a. require recall >= 15% and accuracy >= 65%

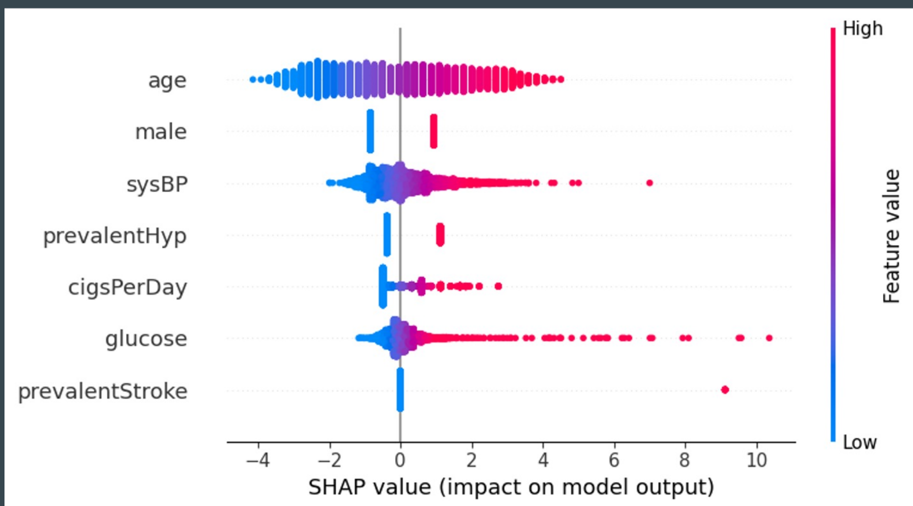
	Logistic Regression (test)	RBF SVM (test)	KNN (test)	Random Forest (test)	Gaussian Naive Bayes (test)	Linear SVM (test)	AdaBoost (test)
accuracy	0.762	0.678	0.812	0.764	0.801	0.833	0.839
precision	0.325	0.285	0.364	0.344	0.367	0.300	0.667
recall	0.416	0.636	0.191	0.478	0.292	0.014	0.038
f1	0.365	0.394	0.251	0.400	0.325	0.027	0.072

Ensemble Model

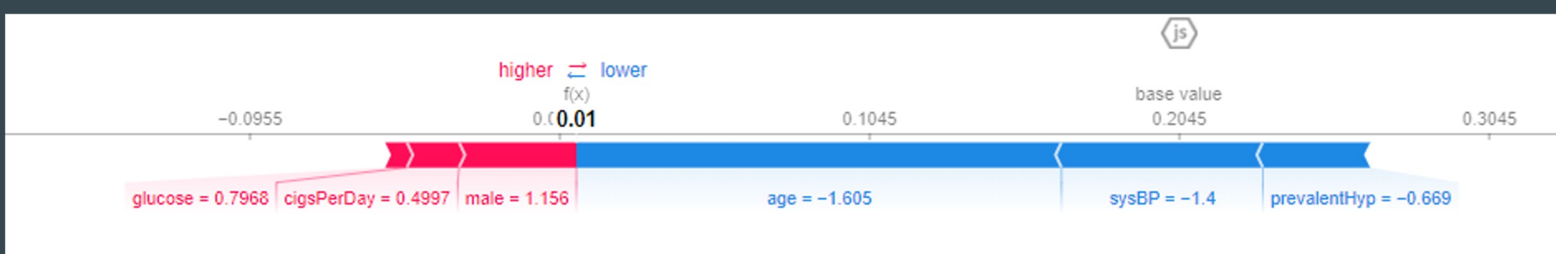
- Takes votes from best classification models:
    a. Logistic regression
    b. RBF SVM
    c. K-Nearest Neighbors
    d. Random forest
    e. Gaussian Naive Bayes
- Accuracy: 83.6% (2nd)
- Precision: 33.7% (5th)
- Recall: 45.5% (3rd)
- F1: 38.7% (3rd)



Explaining our Model’s Decisions



- Most important feature is ‘age’
- Least important feature is ‘prevalentStroke’
- Look at ‘glucose’: many red dots with positive SHAP values
    a. patients with high glucose readings are more likely to be at risk of CHD
- Predicted at-risk patient (below):
    a. low-risk attributes for this patient = age, sysBP, prevalentHyp
    b. High-risk attributes for this patient = glucose, cigsPerDay, and male



Conclusion

Benefits of Analysis

Deploy Preventive Measures

- Discourage smoking
- Prescribe medication
- Change dietary habits
- Encourage active lifestyle

Improve Patient Outcomes

- Reduced risk of CHD
- Stronger doctor-patient relationship
- Healthier lifestyle

Drive Operational Efficiency

- Emergency center visits pressure hospital margins, while preventive care drives margins
- Reduce CHD risk => less emergency care
- Proactive patients => more check-ups, lab work, etc.

Limitations & Areas for Further Exploration

- More thoughtful data pre-processing
  - Remove outliers (we maintained outliers because we thought it was important to capture all types of patients and we had no reason to believe the “outliers” were registered in error)
  - Exploring the limitations of data imputation with sklearn’s KNNImputer
  - Deploy other methods for feature selection (we studied correlation matrices, PC plots, and executed RFE)
- Different Modeling approaches
  - Creating a multi-stage model based on each observation’s feature values (rather than applying the same model to all observations)
  - Optimizing the weights of each model’s vote within the ensemble model (give more votes to the better sub-models)

Appendix

Store misc. Content and links here as needed

- content