Predicting the 10-year Risk of Future Heart Disease

Andrew Mashhadi and Ajay Patel

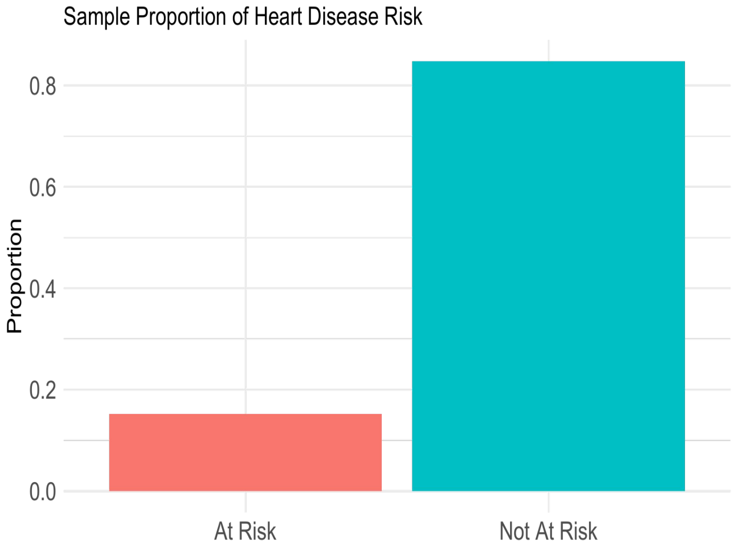
Agenda

- Introduction and Data
- Exploratory Data Analysis
- Feature Engineering
- Methodology
- Results
- Conclusions

Introduction and Dataset

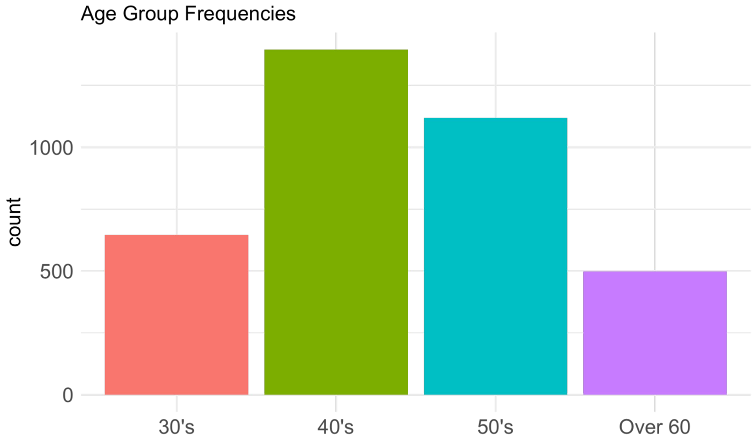
- Every year, cardiovascular disease accounts for 18 million lives (WHO)
- More than ⅘ cardiovascular deaths are due to heart attacks and strokes (WHO)
- We explored how well demographics, medical history, genetics, and behavioral risks can predict for the “10-year risk of future heart disease”
- Our dataset is an ongoing cardiovascular study on the residents from Framingham, Massachusetts
    - Publicly available on Kaggle
- We have nearly 4000 observations and 15 different predictor variables

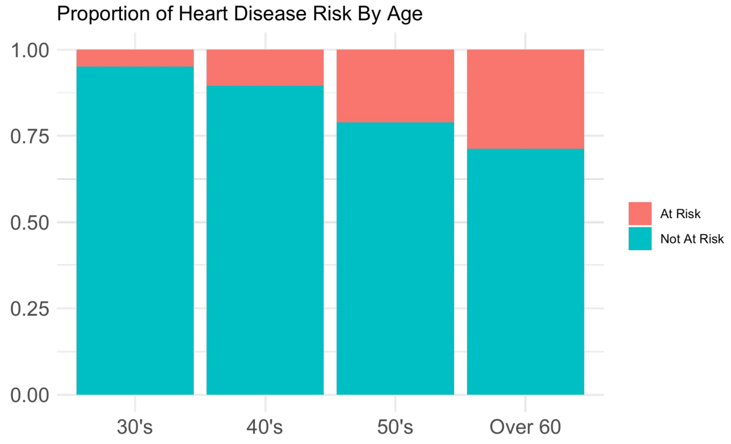
Exploratory Data Analysis - Response Variable



- Nearly 80% of patients in our study are ***not*** at risk for 10-year heart disease

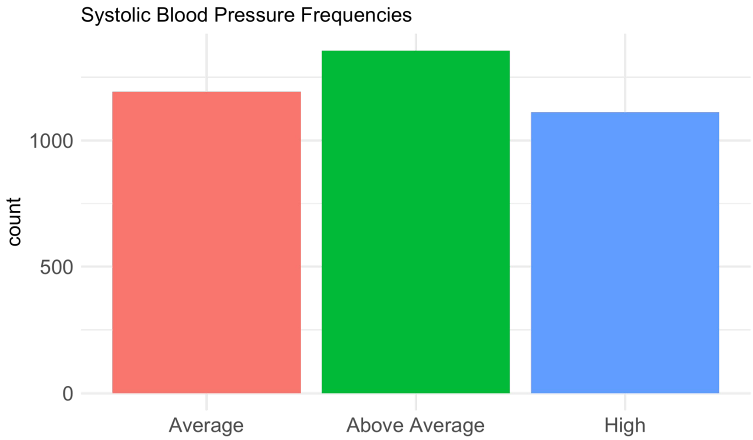
Exploratory Data Analysis - Age

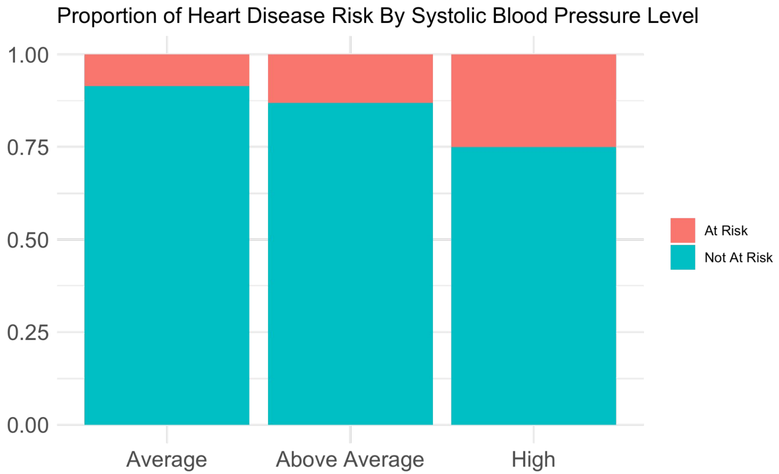




- Most patients in our study are in their 40's and 50's
- As age increases, the proportion of patients at risk of heart disease increases

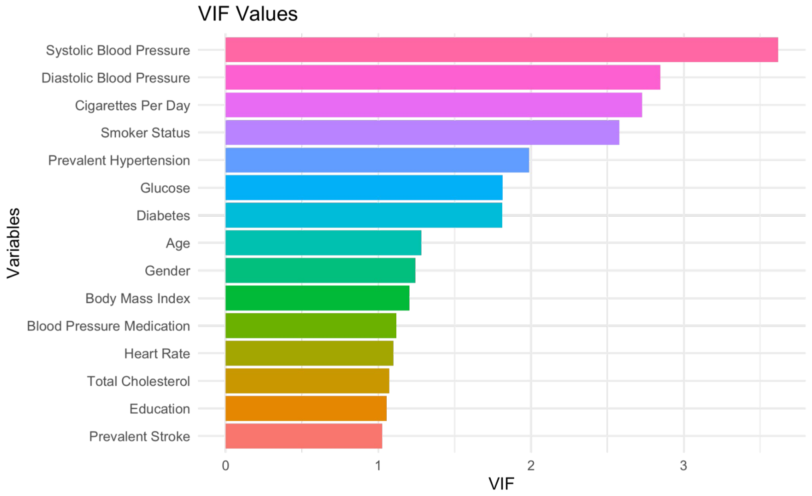
Exploratory Data Analysis - Systolic Blood Pressure

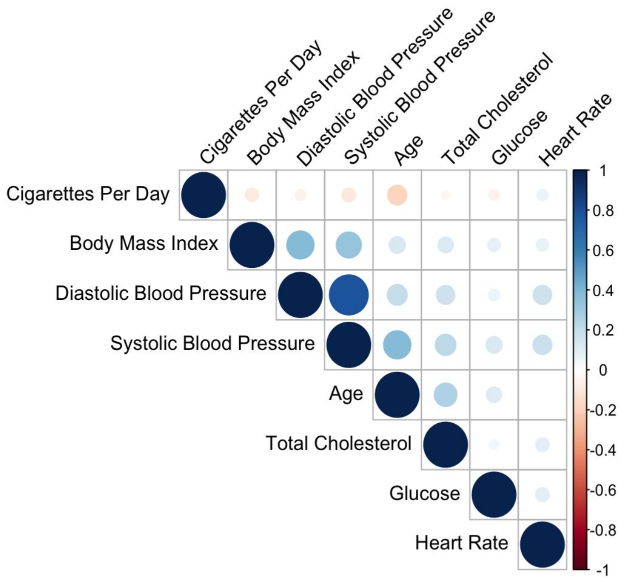




- There is nearly an equal number of patients in each systolic blood pressure group
- As systolic blood pressure increases, the risk of 10-year heart diseases increases

Exploratory Data Analysis - VIF and Correlations





- All VIFs are under 4
- Only Diastolic Blood Pressure and Systolic Blood Pressure have a strong correlation

Feature Engineering

Interaction Terms

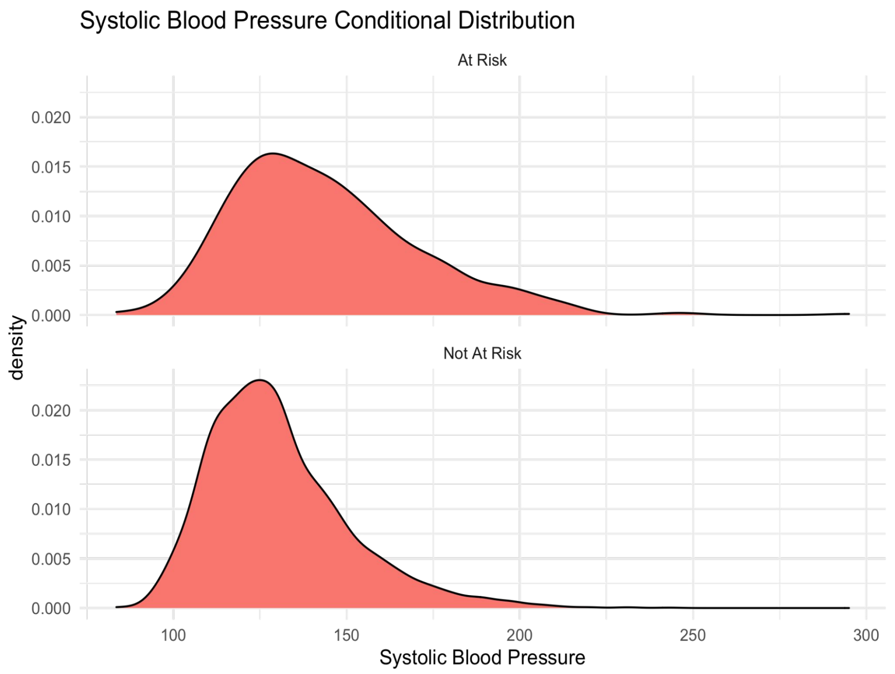
- Age * Cigarettes Per Day
- Systolic Blood Pressure * Cigarettes Per Day
- Systolic Blood Pressure * Glucose

Quadratic Terms

- $(\text{Systolic Blood Pressure})^2$
- $(\text{Diastolic Blood Pressure})^2$
- $(\text{Glucose})^2$

Feature Engineering

- Explored normalized conditional distributions
- If the distributions appear normal but have different variances
- Then, the log-odds is a quadratic function of variable
- So, we add the variable’s quadratic term
- At Risk Variance = 727
  Not At Risk Variance = 416



Methodology - Overview

- 4 models (3 different logistic regressions, 1 classification model)
- Randomly split 80% of the data into a training set; Other 20% is test set
    - Partitioned additional 20% of training data into a validation set
    - Scaled both sets using training set means and standard deviations
- Given the class imbalanced, we randomly oversampled training data to better detect when patients are at risk for 10-year heart disease
- Used cross validation on the validation set to tune each model's hyper-parameters
- Evaluated each model with the ROC-AUC performance metric
- Geometric mean, SquareRoot(Sensitivity * Specificity), was used to find threshold for proper classification (e.g. confusion matrices and test metrics)

Methodology - Models

Logistic Regression + Backwards Elimination

- Started with all variables
- Eliminated variables one-by-one using AIC criterion
- Continued until the AIC can no longer decrease from removing an individual predictor

Logistic Regression + PCA

- Applied PCA to the training data
- Found the optimal number of principal components (***n***) with validation set and scree plots
- Found the optimal logistic regression model using the first ***k*** principal components where ***1*** <= ***k*** <= ***n***

Methodology - Models

Logistic Regression + ElasticNet Regularization

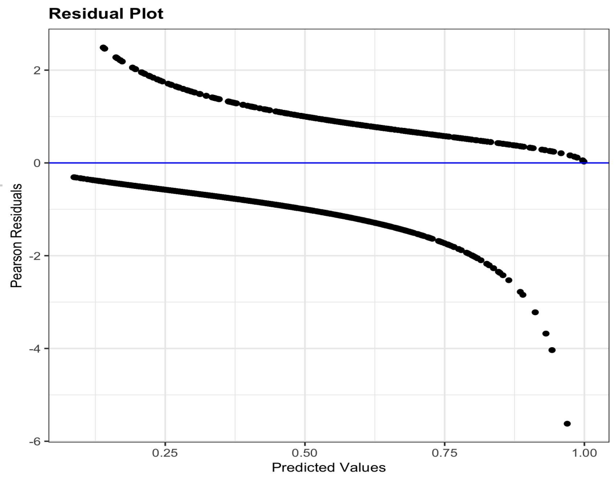
- ElasticNet combines LASSO and Ridge regularization
- Started with all variables and let algorithm determine the best set
- Tuned the mixing and regularization terms (alpha and lambda) with the validation set

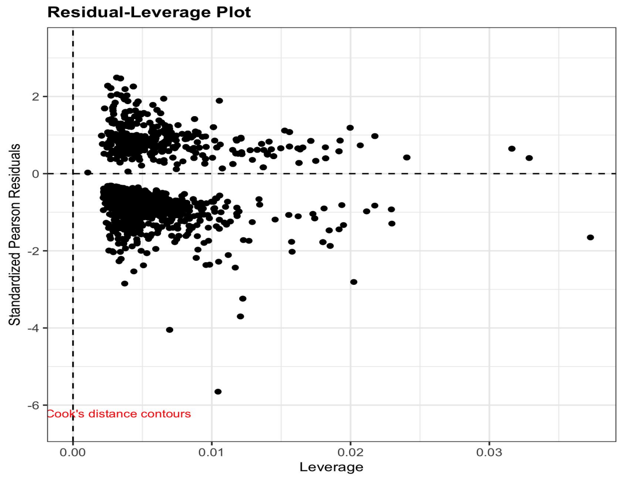
eXtreme Gradient Boosting (XGBoost)

- XGBoost is a decision-tree based algorithm
- Started with all variables and let algorithm determine the best set
- Tuned the number of trees and the maximum depth of each tree with the validation set

Logistic Regression + Backwards Elimination Results

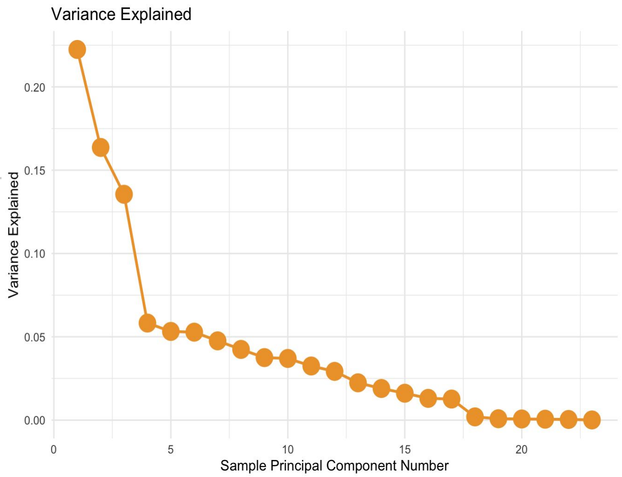
- Reduced number of variables from 23 to 13
- The AIC criterion dropped from 2851.2 to 2836.6
- This model is adequate compared to the full model (p-value = 0.864)
- Pearson's Goodness of Fit Test has a p-value = 0.22
- No observations with large leverages or large residuals

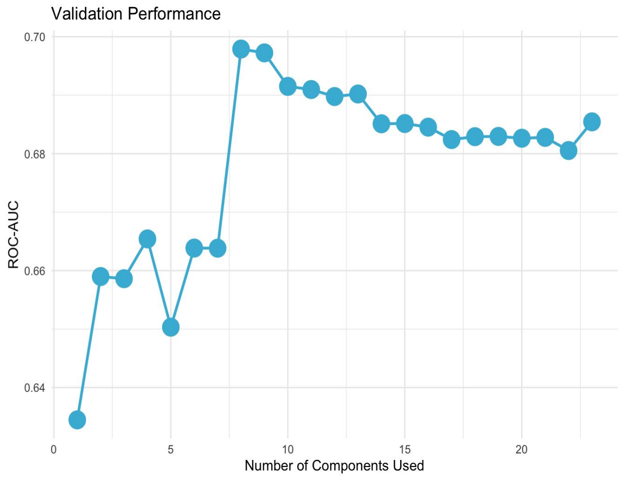




Logistic Regression + PCA Results

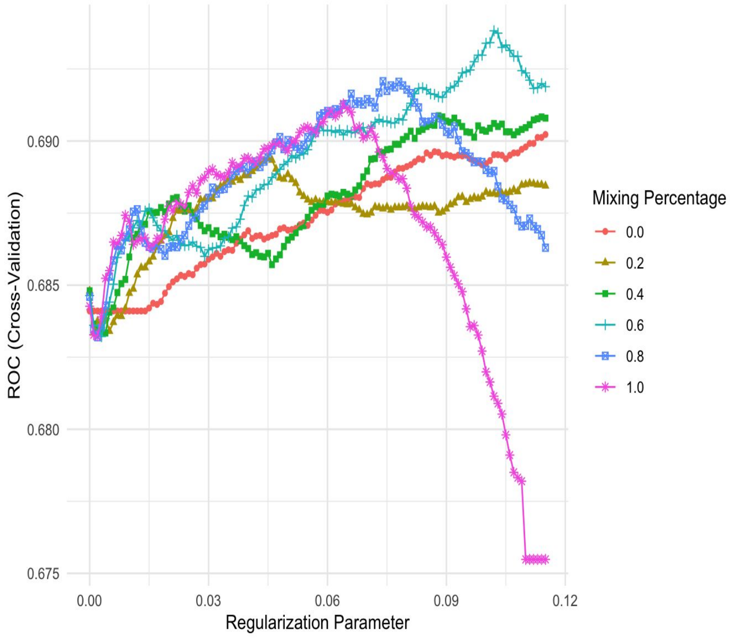
- Found that the first 9 principal components explained over 80% of the sample variance
- Validation set indicated that the first 8 principal components yielded best model
- Pearson's Goodness of Fit Test has a *p*-value = 0.30
- Again, no observations with large leverages or large residuals





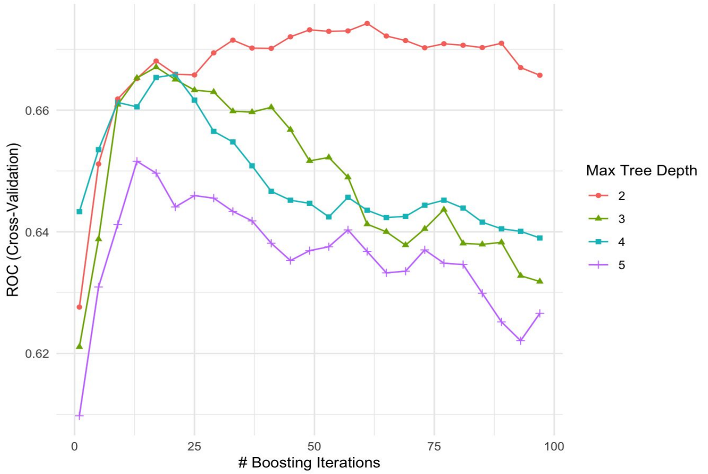
Logistic Regression + ElasticNet Results

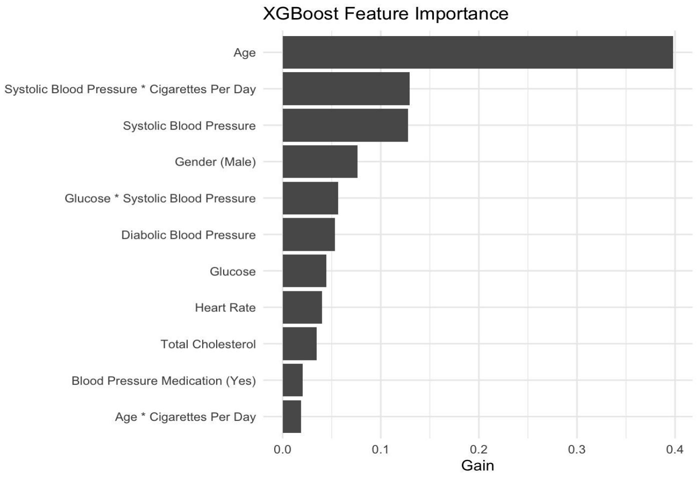
- Reduced number of variables from 23 to 6
- Tuned alpha between [0, 1] and tuned lambda between [0, 3]
- Found optimal alpha and lambda is 0.6 and 0.098 respectively
- Variables remaining: gender, age,
  systolic blood pressure, (systolic blood pressure)^2,
  age * cigarettes per day,
  systolic blood pressure * cigarettes per day



XGBoost Results

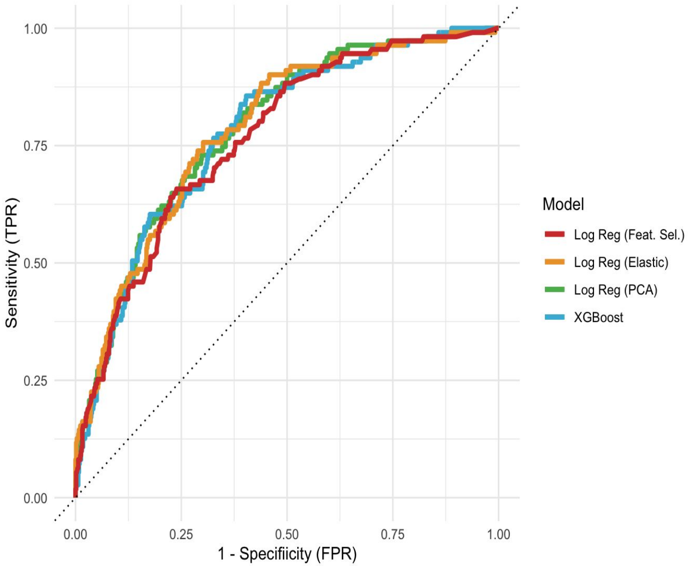
- Optimal number of trees is 13
- Optimal maximum tree depth is 2
- Age is the most important variable from the feature importance plot
- The 3 most important variables here are also in the Logistic Regression + ElasticNet model





Test Set Results

Model	Sensitivity	Specificity	Balanced Accuracy
Logistic Regression (Backwards Selection)	0.802	0.608	0.705
Logistic Regression (PCA)	0.739	0.700	0.719
Logistic Regression (Elastic Net)	0.775	0.661	0.718
XGBoost	0.703	0.673	0.688



- All four models demonstrated "good" performance
- The table indicates that the PCA and ElasticNet models performed the best at optimal threshold
- The ROC-AUC indicates that the Backwards Selection model performed the best across all thresholds

Conclusions

- Successfully modeled the 10-year risk of heart disease with demographics, medical history, genetics, and behavioral risks
- We think the variables in our models can be the starting point to diagnosing the 10-year risk of heart disease
- Quite remarkable that we can obtain these results with machine learning and feature selection without a medical degree
- The variables we used can be obtained in an annual physical
    - Perhaps there are more advanced metrics worth exploring