Heart Disease Data Analysis

BIOS 735 Group 2 - Mingwei, Di, Wanting, Eunchong, & Andrew

Github repository

Introduction

Introduction - Background

- In 2020, heart disease was the leading cause of death in the United States with 696,962 deaths attributed (followed by cancer & COVID-19) according to the final 2020 U.S. mortality data from the CDC.
- Many health status indicators are found related to heart disease such as high blood pressure, high cholesterol, smoking, diabetes, obesity (BMI > 30.0), insufficient physical activity, & excessive alcohol consumption.
- Understanding and detecting the factors that have the greatest impact on heart disease occurrence in populations is crucial in healthcare to improve the length and quality of life.

Introduction - dataset

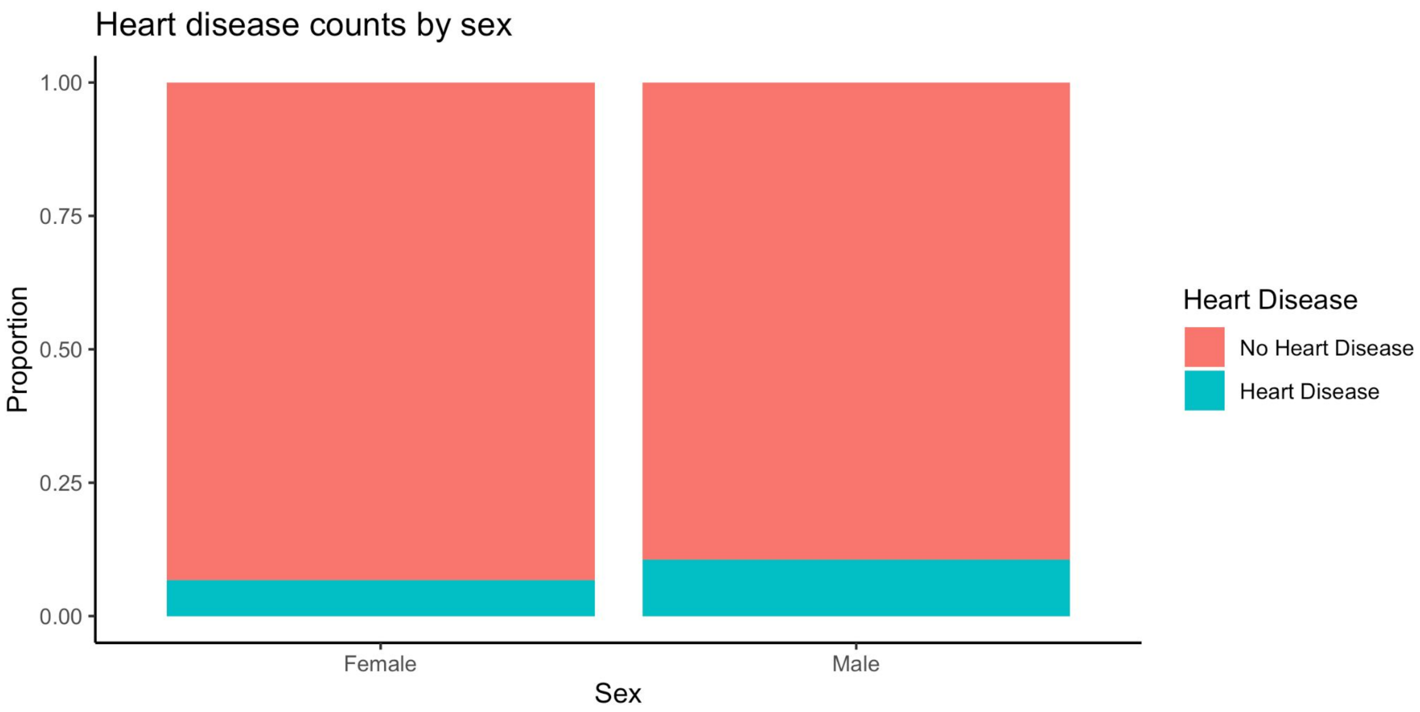
- Personal key indicators of heart disease (kaggle)
  - https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
- 319795 observations of 18 variables (1 outcome & 17 covariates)
- Primary Outcome of interest: Heart disease occurrence
  - Subject experienced either coronary heart disease or myocardial infarction (binary)
- Other factors:
  - BMI, Smoking status, alcohol consumption, stroke occurrence, poor physical health, poor mental health, walking difficulty, sex, age, race, diabetes, physical activity, general health status, average sleep time, asthma, kidney disease occurrence, skin cancer occurrence

Introduction - project aims/research questions

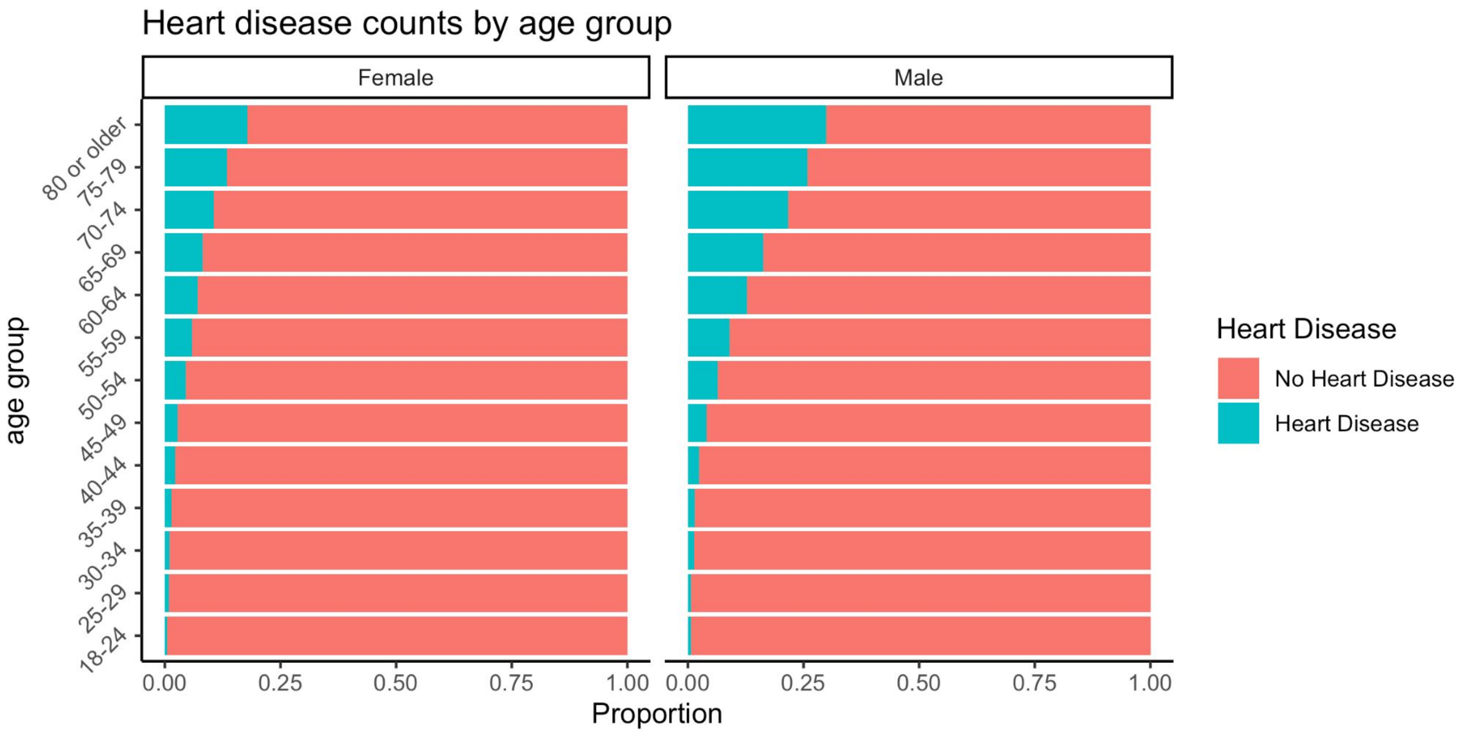
- Aim 1: determine relationship between heart disease prevalence and all other factors
- Aim 2: determine relationship between BMI & heart disease
- Aim 3: determine relationship between risky behaviors (smoking & drinking) with heart disease or stroke (poor health outcomes)
- Aim 4: Compare the model fitting performance between logistic regression model and other machine learning methods. And propose predictions for occurrence of heart disease given health condition indicators

Exploratory figures

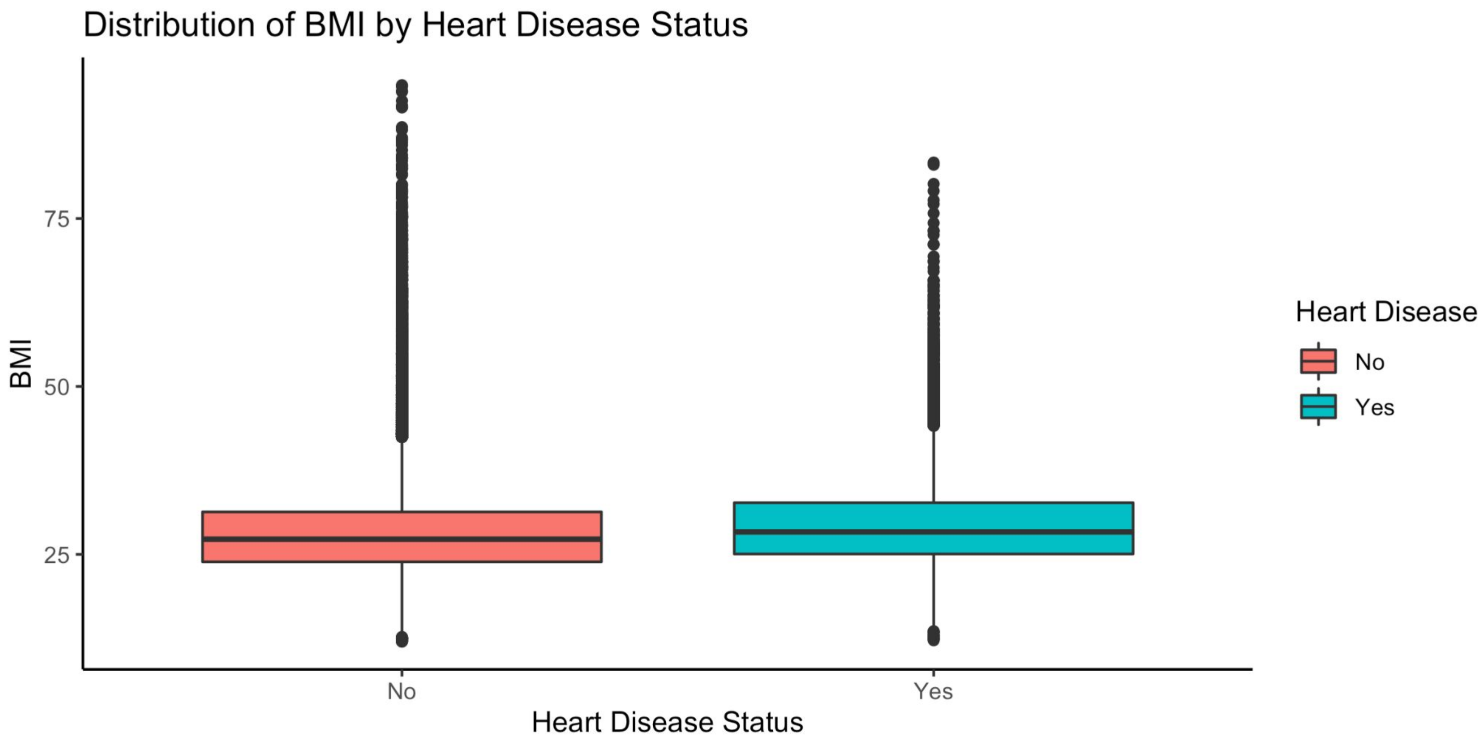
Heart disease by sex



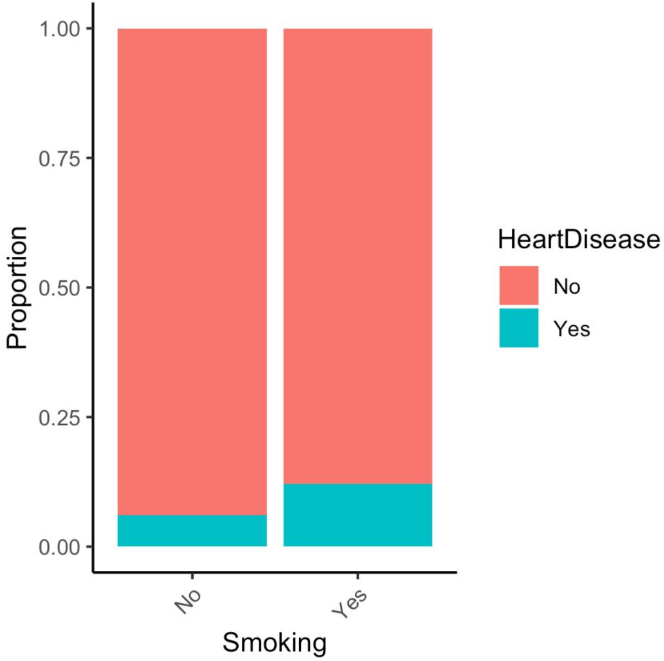
Heart disease by age & sex

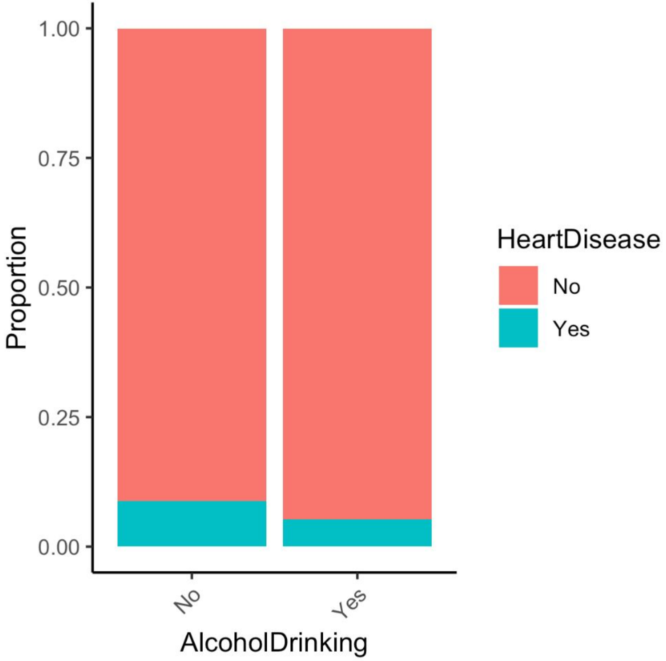


Heart disease by BMI

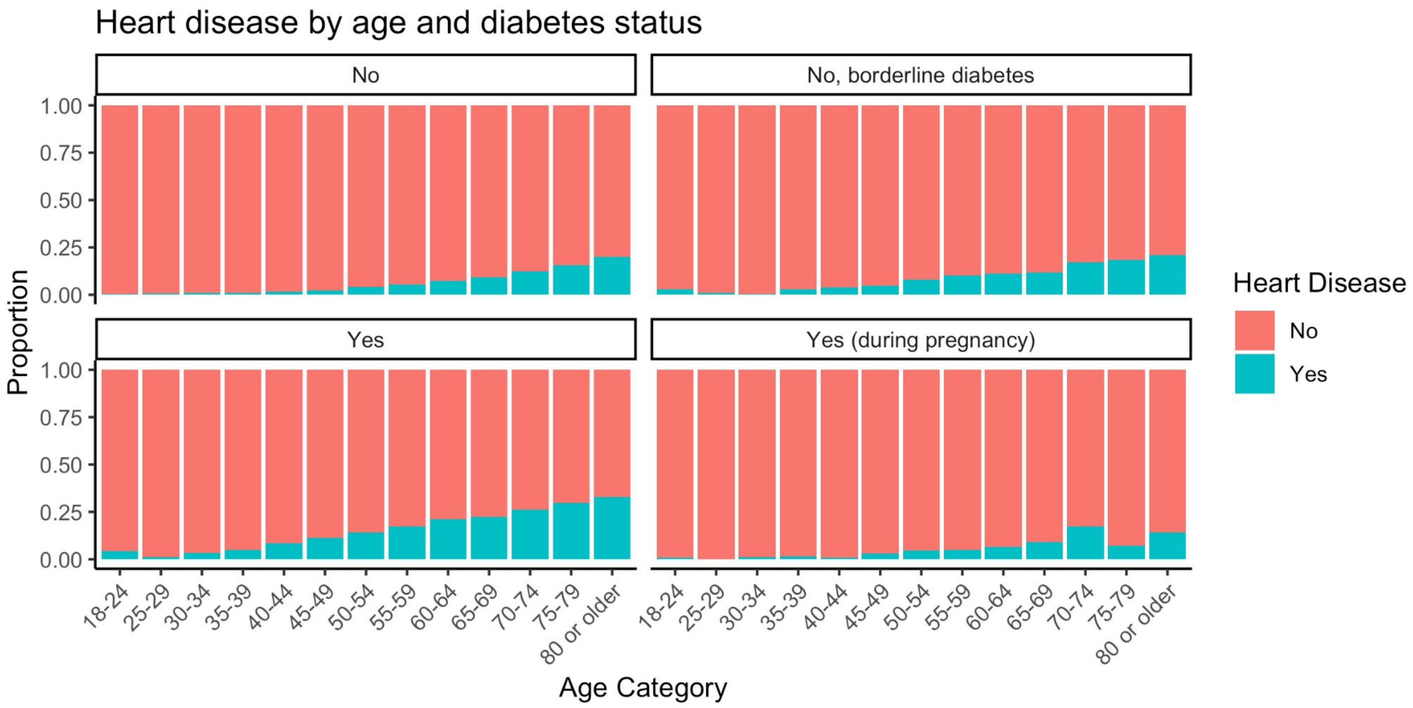


Heart disease by risky behaviors





Heart disease by diabetes status & age



Data Pre-process

- Data preparation
  - Feature scaling (continuous variable transform to stand normal)
  - Missing imputation (KNN)
  - Create dummy variables for categories
- Train/Test split
  - Split the dataset based on heart disease outcome (4:1)
    - Training: 255837 observations
    - Testing: 63958 observations
  - Cross validation on Train dataset
  - Report test data performance

Methods

Methods - Decision tree

- Decision trees are a versatile machine learning method with incredibly high levels of interpretability.
- All 17 covariates available for tree construction
- Adjusted for imbalance data by sampling approximate 3:1 ratio of negative to positive cases of heart disease.
- Accuracy, sensitivity, and specificity are recorded and reported
- Gini impurity measure is used for node splitting (higher coefficient indicates more differences in a node)

Methods - logistic regression

The logistic regression model is defined as follows:

$$y_i|x_i \sim Bernoulli(\pi_i) \;\; \text{for } i = 1, ..., n,$$

where $\;logit(\pi_i) = \log(\frac{\pi_i}{1-\pi_i}) = x_i^T\beta = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$

The log-likelihood of β for the logistic regression model is

$$l_n(\beta) = \sum_{i=1}^{n}\{y_i x_i^T\beta - \log(1 + \exp(x_i^T\beta))\}$$

Methods - logistic regression (IRLS Algorithm)

- Training set (80% of total dataset)
  - The number of observations in the training set: 255,837
  - The number of covariates: 17
- Gradient Descent or Stochastic Gradient Descent (briefly or not)
  - The update is $\beta^{k+1} = \beta^k - \alpha f'(\beta^k)$
  - When α=0.0001 and the initial beta is the zero vector, it took more than 1 hour with over 30,000 iterations (2/3 of them was around its convergence)
  - SGD: Mini-batch with the size of 0.01 * 255,837
    - It took about 30 mins
- Iterative Reweighted Least Squares (IRLS)
  - Pros: Faster, 1st and 2nd derivatives are known
  - Cons: Sensitive to the starting value, Hessian matrix (size:38*38)

Methods - logistic regression (IRLS Algorithm)

The log-likelihood in matrix form is

$$l_n(\beta) = Y^T X\beta - b(\beta)^T J_n,$$

where $Y = (y_1, ..., y_n)^T$, $X = (x_1^T, ..., x_n^T)^T$, $\beta = (\beta_0, ...\beta_p)^T$, $b(\beta)^T = (log(1+exp(x_1\beta)), ..., log(1+exp(x_n\beta))$, and $J_n = (1, ..., 1)$.

The first derivative of the log-likelihood is

$$l_n'(\beta) = X^T(Y - \mu),$$

where $\mu = (\frac{exp(x_1\beta)}{1+exp(x_1\beta)}, ..., \frac{exp(x_n\beta)}{1+exp(x_n\beta)})^T$,

The Fisher information is

$$I_Y(\beta) = E(-l_n''(\beta)) = X^T W X,$$

where $W = diag(\frac{exp(x_1\beta)}{(1+exp(x_1\beta))^2}, ..., \frac{exp(x_n\beta)}{(1+exp(x_n\beta))^2})$.

Methods - logistic regression (IRLS Algorithm)

The NR algorithm is

$$\beta^{k+1} = \beta^k + (X^T W^k X)^{-1} X^T (Y - \mu^k).$$

From the NR algorithm, one can derive IRLS algorithm for β^{k+1} such that

$$\begin{aligned}\beta^{k+1} &= (X^T W^k X)^{-1} X^T W^k z^T \\ &= \left(X^T \begin{bmatrix} w_1^k x_1 \\ \vdots \\ w_n^k x_n \end{bmatrix} \right)^{-1} X^T \begin{bmatrix} w_1^k z_1 \\ \vdots \\ w_n^k z_n \end{bmatrix}\end{aligned}$$

where a response vector is $Y = (Y_1, ..., Y_n)$, a design matrix is X having X_i as row i, $z = (z_1, ..., z_n)^T$ with $z_i = x_i\beta + e_i = x_i\beta + \frac{(1+exp(x_i\beta))^2}{exp(x_i\beta)}\left(y_i - \frac{exp(x_i\beta)}{1+exp(x_i\beta)}\right)$.

Methods - logistic regression (IRLS Algorithm)

- optim.IRLS(X, Y, beta)
  - X: design matrix, Y: response vector, beta: the starting beta
  - Convergence criterion: Absolute change of log-likelihood
  - Tolerance: 10^-10
- Speed?
  - About 5 seconds for Convergence
  - 9 Iterations
- Convergence?
  - Starting beta with all elements being 0
    - Converged
  - Starting beta with all elements being 0.5
    - Did not converge
- Result compared to glm function
  - Maximum difference between the two parameters: 3.046436*10^-7

Methods - logistic regression (BFGS Algorithm)

- Quasi-Newton Method – Avoid calculating the inverse of Hessian matrix at each iteration
- Update the step length a_k using line search satisfying Wolfe conditions:

$$f(\beta^k + a_k d_k) \leq f(\beta^k) + \rho a_k d_k^T \nabla f(\beta^k)$$
$$d_k^T \nabla f(\beta^k + a_k d_k) \geq \sigma d_k^T \nabla f(\beta^k)$$

- Update the approximation of the inverse of Hessian matrix by

$$H_{k+1} = \left(I - \frac{s_k y_k^T}{sk^T y_k}\right) H_k \left(I - \frac{y_k s_k^T}{sk^T y_k}\right) + \frac{s_k s_k^T}{s_k^T y_k}$$

Methods - logistic regression (BFGS Algorithm)

Algorithm 2 BFGS

Require: Given starting point β_0, convergence tolerance $\epsilon > 0$, and inverse Hessian approximation

$H_0 = I$;

$k \leftarrow 0$

while $|f(\beta^k) - f(\beta^{k+1})| > \epsilon$ **do**

$d_k = -H_k \nabla f(\beta^k)$; ▷ Compute the search direction

$\beta^{k+1} = \beta^k + a_k d_k$ ▷ where a_k is computed from a line search procedure to satisfy the Wolfe conditions

$s_k = \beta^{k+1} - \beta^k$

$y_k = \nabla f(\beta^{k+1}) - \nabla f(\beta^k)$

$H_{k+1} = \left(I - \frac{s_k y_k^T}{sk^T y_k}\right) H_k \left(I - \frac{y_k s_k^T}{sk^T y_k}\right) + \frac{s_k s_k^T}{s_k^T y_k}$ ▷ Update inverse Hessian approximation

$k \leftarrow k + 1$;

end while

Methods - logistic regression (BFGS Algorithm)

Algorithm 3 Line search satisfying Wolfe Conditions

Require: Set $\rho = 10^-4$, $\sigma = 0.9$,$a = 0$ $a_k = 1$ and $b = N$

while $a < b$ **do**

  if $f(\beta^k + a_k d_k) > f(\beta^k) + \rho a_k d_k^T \nabla f(\beta^k)$ **then**

    set $b = a_k$ and $a_k \leftarrow (a+b)/2$

  else if $d_k^T \nabla f(\beta^k + a_k d_k) < \sigma d_k^T \nabla f(\beta^k)$ **then**

    set $a \leftarrow a_k$

    if b = N **then** $a_k \leftarrow 2a$

    else $a_k \leftarrow (a+b)/2$

    end if

  end if

end while

Methods - logistic regression (glm)

- *logit*(p(Heart Disease)) = $X'\beta$, where X is the full design matrix of covariate values
- Models fit
  - Heart Disease ~ all covariates
  - Heart Disease ~ BMI
  - Heart Disease ~ Smoking + Drinking
  - Stroke ~ Smoking + Drinking
  - Full model: backward & forward variable selection based on AIC
- Heart disease classification predictions made on test set

Methods - R package implementation

- Package name: “glmLogistic”
- Functions
  - Loglik - computes value of log-likelihood
  - D1.loglik - 1st derivative of log-likelihood
  - Beta.updater - iterative estimates of β
  - optim.IRLS - compute parameter estimates with IRLS algorithm
  - optim.BFGS - compute parameter estimates with BFGS algorithm
- Output
  - Parameter estimates, standard error, log-likelihood, # iterations, final absolute change in log-likelihood

Methods - random forest & support vector machine

- Down-sample the training data set to obtain balanced classes
- Random Forest
  - R ranger package
  - 5-fold cross-validation to tune parameter – mtry
- Support Vector Machine
  - R caret package
  - Linear kernel
  - 5-fold cross-validation with ‘tunelength = 10’
  - Transfer categorical variable into dummy variables and center and scale before fitting
- Prediction performance evaluated using testing data set

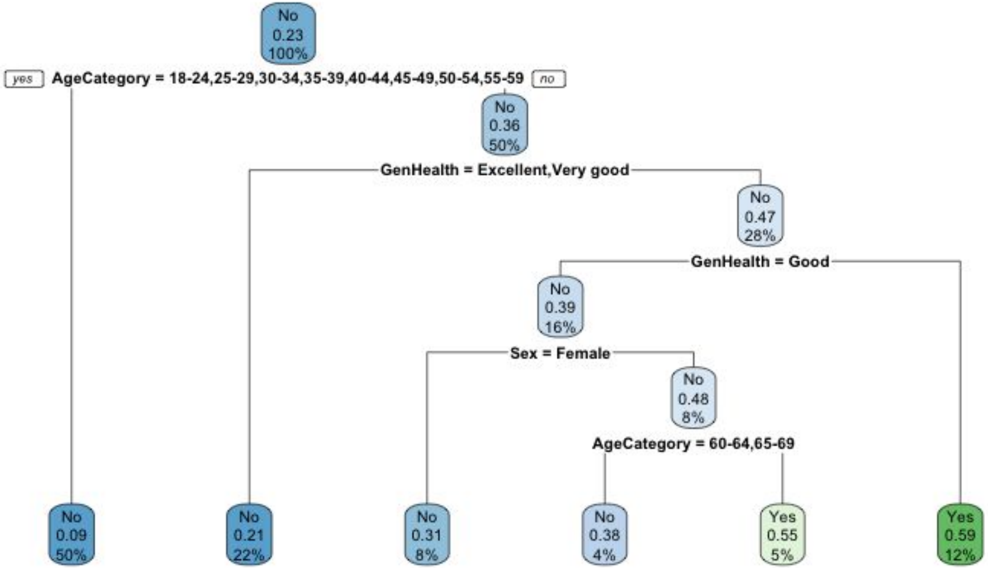
Machine Learning Methods

Classification Problem:

- Logistic Regression with Elastic net
- Random Forest/Boosting
    - Ranger function more suitable for random forest in large dataset
- SVM
    - Slow due to iterations to get distance of one sample with all others
- Principle Component
    - Used often when most are continuous variables, assuming linearity
    - Features need to be correlated
- K- means
    - Used in pre-process of data if missing, reduce dimensions

Decision Tree

- Age, general health category, & sex used for tree construction (out of 17 covariates)
- Subjects who are old and in worse than average health are more likely to have heart disease & Younger patients in good health do not have heart disease
- Accuracy: 0.866
- Sensitivity: 0.420
- Specificity: 0.908



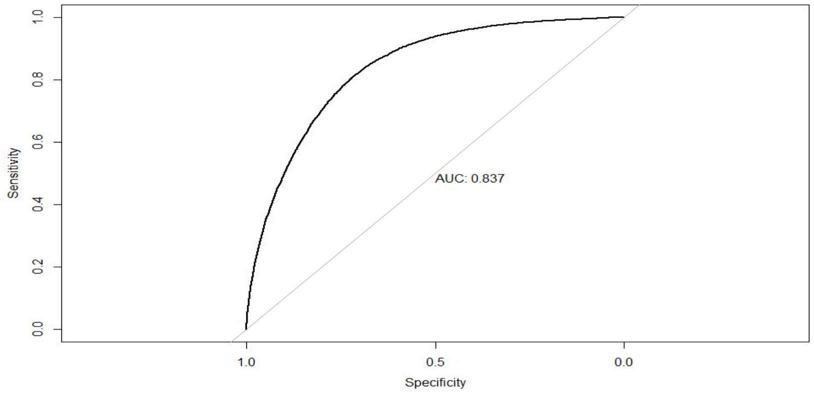
Results

Logistic regression (IRLS, BFGS, & glm)

Parameter	Estimate by function			Parameter	Estimate by function		
	glm	IRLS	BFGS		glm	IRLS	BFGS
Intercept	-6.3586	-6.3586	-6.3586	Race			
BMI	0.0086	0.0086	0.0086	Asian	-0.4976	-0.4976	-0.4977
Smoking: Yes	0.3546	0.3546	0.3546	Black	-0.2774	-0.2774	-0.2774
Alcohol: Yes	-0.2371	-0.2371	-0.2371	Hispanic	-0.2254	-0.2254	-0.2254
Stroke: Yes	1.0578	1.0578	1.0578	Other	-0.0125	-0.0125	-0.0124
PhysicalHealth	0.0022	0.0022	0.0022	White	-0.0221	-0.0221	-0.0221
MentalHealth	0.0051	0.0051	0.0051	Diabetic			
DiffWalking: Yes	0.2121	0.2121	0.2121	Borderline diabetes	0.1683	0.1683	0.1683
Sex: Male	0.7036	0.7036	0.7036	Yes	0.4924	0.4924	0.4924
Age Category				Yes (pregnancy)	0.1356	0.1356	0.1357
25-29	0.1483	0.1483	0.1483				
30-34	0.5313	0.5313	0.5313	PhysicalActivity: Yes	0.0203	0.0203	0.0203
35-39	0.5324	0.5324	0.5324	General Health			
40-44	1.0137	1.0137	1.0137	Fair	1.5346	1.5346	1.5346
45-49	1.3002	1.3002	1.3002	Good	1.0598	1.0598	1.0598
50-54	1.711	1.711	1.711	Poor	1.9333	1.9333	1.9333
55-59	1.9758	1.9758	1.9758	Very good	0.4788	0.4788	0.4788
60-64	2.2102	2.2102	2.2103				
65-69	2.4665	2.4665	2.4666	SleepTime	-0.0247	-0.0247	-0.0247
70-74	2.751	2.751	2.751	Asthma: Yes	0.2759	0.2759	0.2759
75-79	2.9461	2.9461	2.9461	Kidney Disease: Yes	0.5823	0.5823	0.5823
80 or older	3.1999	3.1999	3.2	Skin Cancer: Yes	0.1203	0.1203	0.1203

Machine Learning

- Random forest
  - Mtry = 6
- SVM
  - C = 1

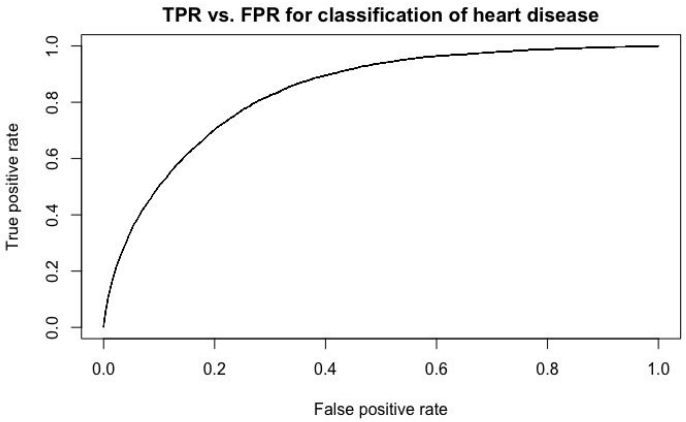


	Accuracy	Kappa	AUC
Random Forest	0.727 (0.724, 0.731)	0.229	0.832
Linear SVM	0.753 (0.750, 0.757)	0.248	0.837

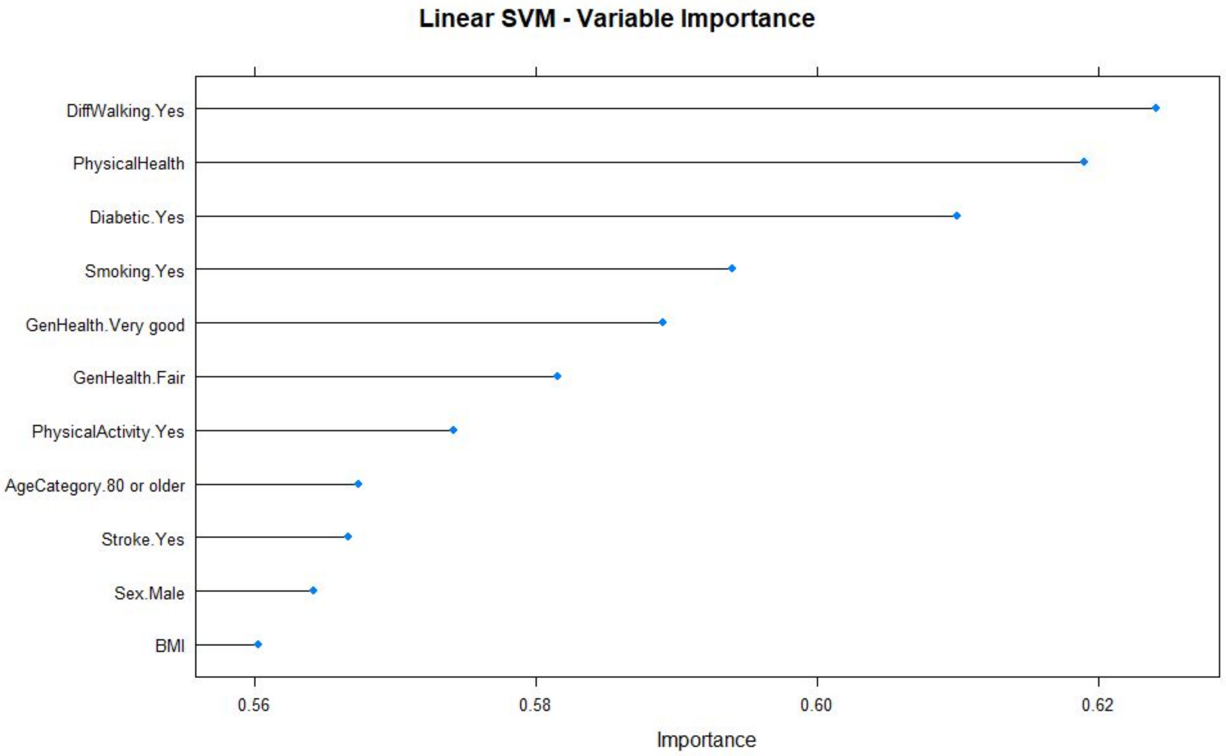
Logistic regression (GLM)

- Performance metrics
  - Accuracy: 0.916 (0.9139, 0.9182)
  - Sensitivity: 0.109
  - Specificity: 0.992
  - Positive predictive value: 0.545
  - Cohen's kappa: 0.157
- Trained model includes 38 coefficients
  - BMI = 0.0086
  - Smoking = 0.355
  - Alcohol Drinking = -0.237
  - Age (80+) = 3.20
  - Diabetes (yes) = 0.492
- exp(β)-fold multiplicative change in the odds of heart disease occurrence for a unit change in continuous covariates or presence of binary/categorical covariates (or β-fold change in log-odds

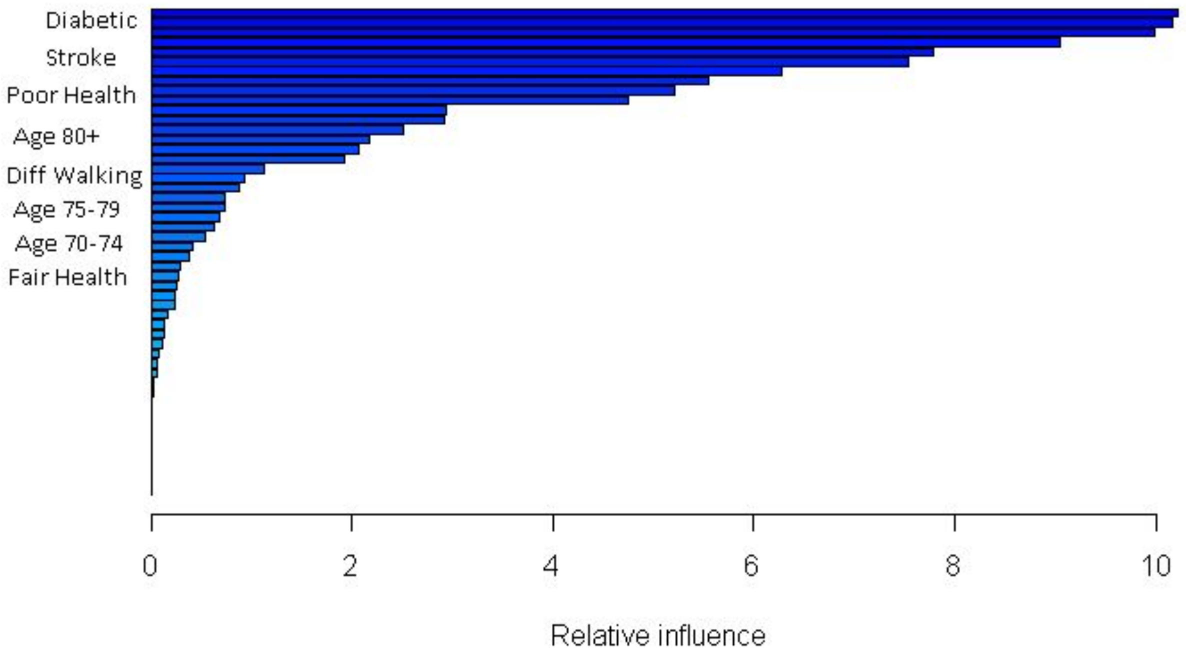
	Reference	
Prediction	No Heart Disease	Heart Disease
Heart Disease	57996 (91%)	4880 (1%)
No Heart Disease	488 (7%)	594 (1%)



Top 10 most important variables from SVM model



High Risk Factors



Results - parametric vs. non-parametric approaches

- Logistic regression (parametric) achieved accuracy of 0.916 (0.9139, 0.9182) with sensitivity of 0.109
- Non-parametric approaches don't have specific estimates for covariates in the data (hence the name)
  - Decision tree is quite interpretable (accuracy = 0.866 / sensitivity = 0.420)
  - Random forest & SVM are able to make accurate predictions as well as give estimates for relative variable importance.
- Both types of methods provide a balance of prediction capabilities & interpretability but sensitivity is a struggle with unbalanced data.
  - Feature weighting & addressing repeated observations may improve sensitivity

Conclusion / Discussion

Conclusion

- Developed R package (glmLogistic) to compute logistic regression parameter estimates with BFGS & IRLS
  - Estimates agree closely with glm()
  - Accuracy: 0.916 & Sensitivity: 0.109
- Other models: decision tree, random forest, support vector machine
  - Best accuracy: 0.866 & sensitivity: 0.420
- Influential factors
  - General health, age, smoking status, diabetic status, sex
- Interpretation of aims
  - BMI has 1.09-fold change in odds of heart disease
  - Smoking doubles risk of heart disease, drinking decreases overall risk
  - risk of stroke also increased by smoking, decreased by drinking

Discussion - clinical relevance

- Factors relevant for presence of heart disease can be used to emphasize surveillance efforts
  - ie. higher prevalence in subjects of old age
- Targeting of healthy habit campaigns in the public health sector based on statistical evidence of risk factors like smoking, diabetes, etc.
- Individuals with worse than average general health are at a substantially higher risk of heart disease relative to subjects in good health.

Discussion - concerns & limitations

- Unbalanced data (presence vs. absence of heart disease) leads to reduced model sensitivity in all cases
  - positive predictive values are approximately 50%
- Categorical factors like age are segmented into less refined classes. These may have more influence in estimation if recorded on a continuous scale.
- Confounders
  - Alcohol associated with reduced risk but heavy drinkers tend to be young