HEART DISEASE STATISTICAL ANALYSIS

By: Kristine Dinh

INTRODUCTION

- Retrieved data from Kaggle
- Use logistic regression for model development
- Use bootstrapping method for model selection and validation
- Introduce simple R syntax for statistical modeling

OUTLINE

- Logistic Regression
- Data Cleaning
- Data Exploratory Analysis
- Modeling Bootstrap
- Limitation/Future Research

INTRODUCTORY TO LOGISTIC REGRESSION

1) General form of simple logistic regression model

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

- $E(Y|x)$ - Conditional mean of Y given x
- $\pi(x)$– average probability of the occurrence of an event

$$0 \leq \pi(x) \leq 1$$

2) Transformation to linear regression model:

$$g(x) = log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

- $g(x)$ – continuous linear function
- x - Independent variable
- β_0 - Intercept
- β_1 - Slope

3) Fitting the model:

Logistic Likelihood function

$$L(\beta) = \log\big(l(\beta)\big) = \sum_{i=1}^{n}\{y_i log[\pi(x_i)] + (1 - y_i)log[1 - \pi(x_i)]\}$$

4) Testing for Significant:

i. Deviance (D statistics) ~ Sum of squares residual

$$D = -2\sum_{i=1}^{n}\left[y_i log\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)log\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right], \text{where } \hat{\pi}_i = \hat{\pi}(x_i)$$

ii. G statistics

$$G = 2\left\{\sum_{i=1}^{n}[y_i \log(\hat{\pi}_i) + (1-y_i)\log(1-\hat{\pi}_i)] - [n_1\log(n_1) + n_0\log(n_0) - nlog(n)]\right\} \sim \chi_1^2$$

5) Confident Interval

- Wald-based CI
  - Positive square root of the variance estimator
- Venzon and Moolgavkar method
  - Plot log-likelihood function vs coefficient values
  - Maximizing log-likelihood

DATA CLEANING

1. Add description factor variables
2. Add reference level for factor variables
3. Add bins for continuous variables

`data.frame 303 obs. of 14 variables`

DATA EXPLORATORY ANALYSIS

Gender

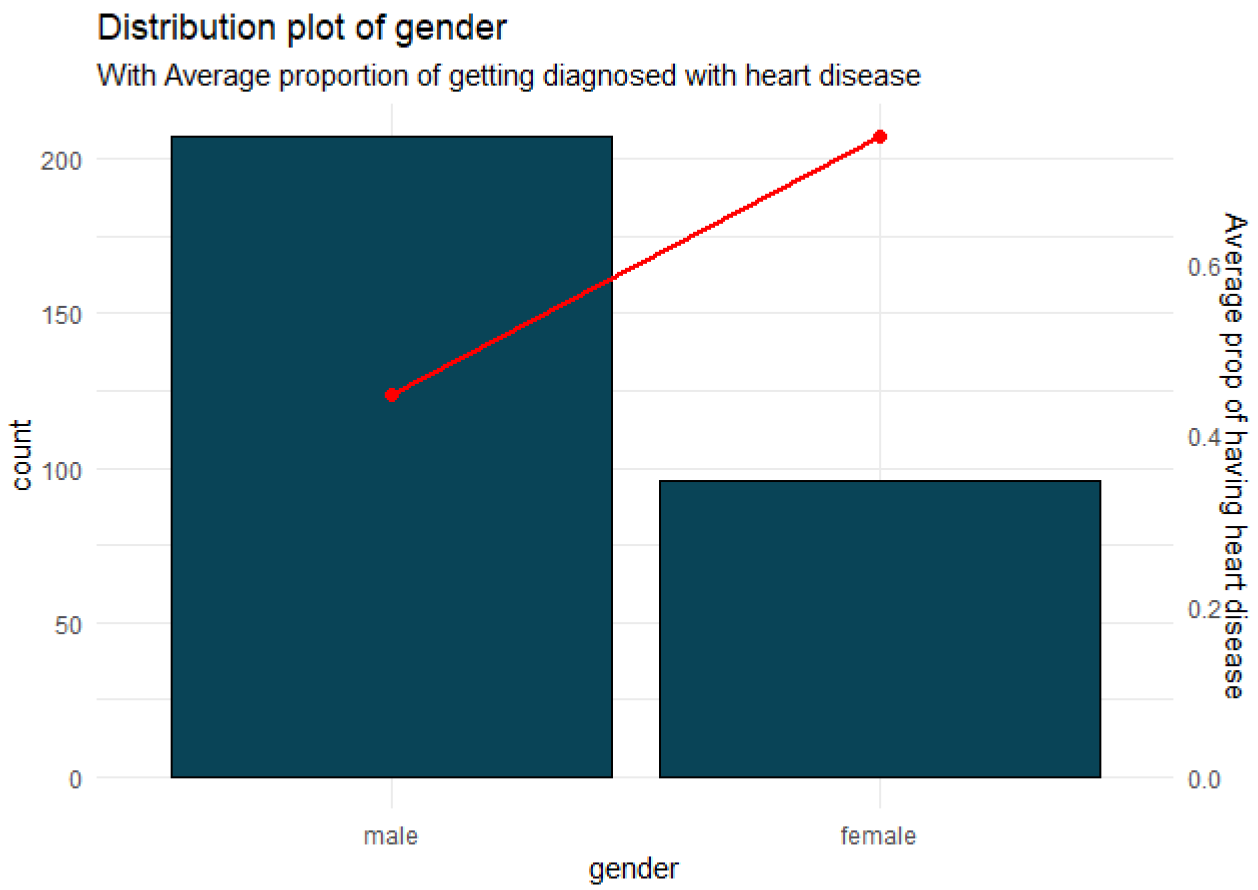
Reference level: male

distribution table

xvar	count	prop
male	207	0.4493
female	96	0.7500

summary statistics

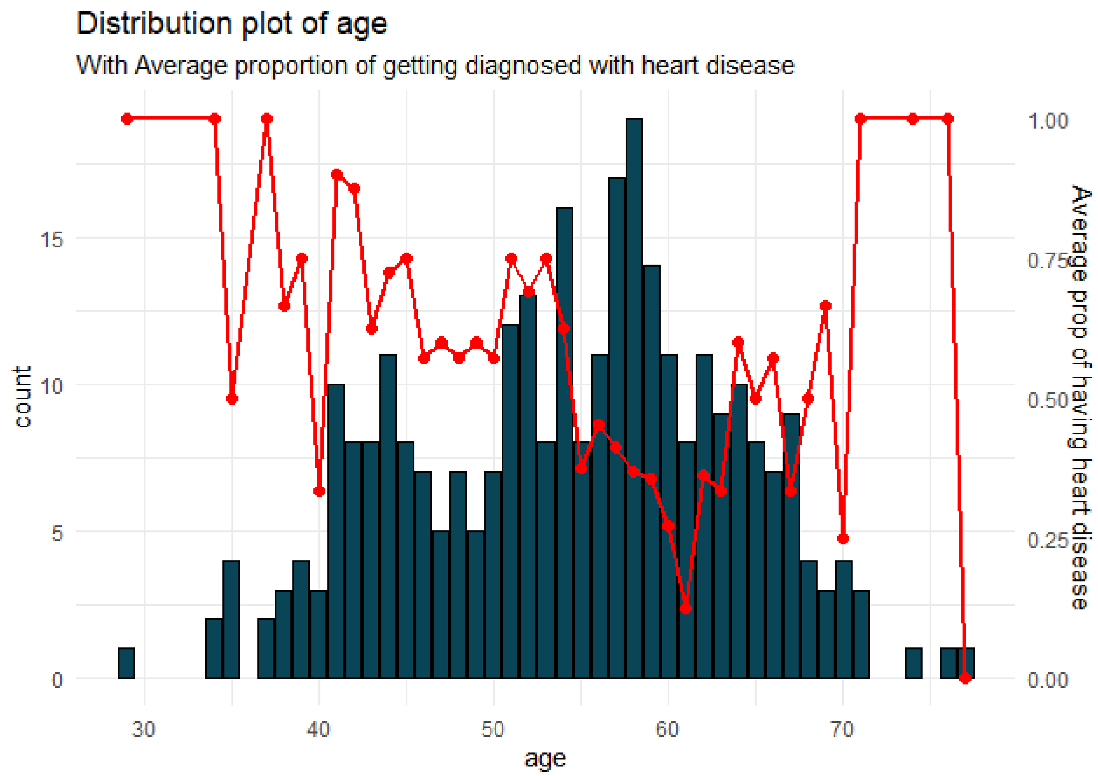
term	estimate	std.error	statistic	p.value	2.5 %	97.5 %
(Intercept)	-0.204	0.140	-1.457	0.145	-0.479	0.069
xvarfemale	1.302	0.274	4.752	0.000	0.777	1.855



Age

summary statistics

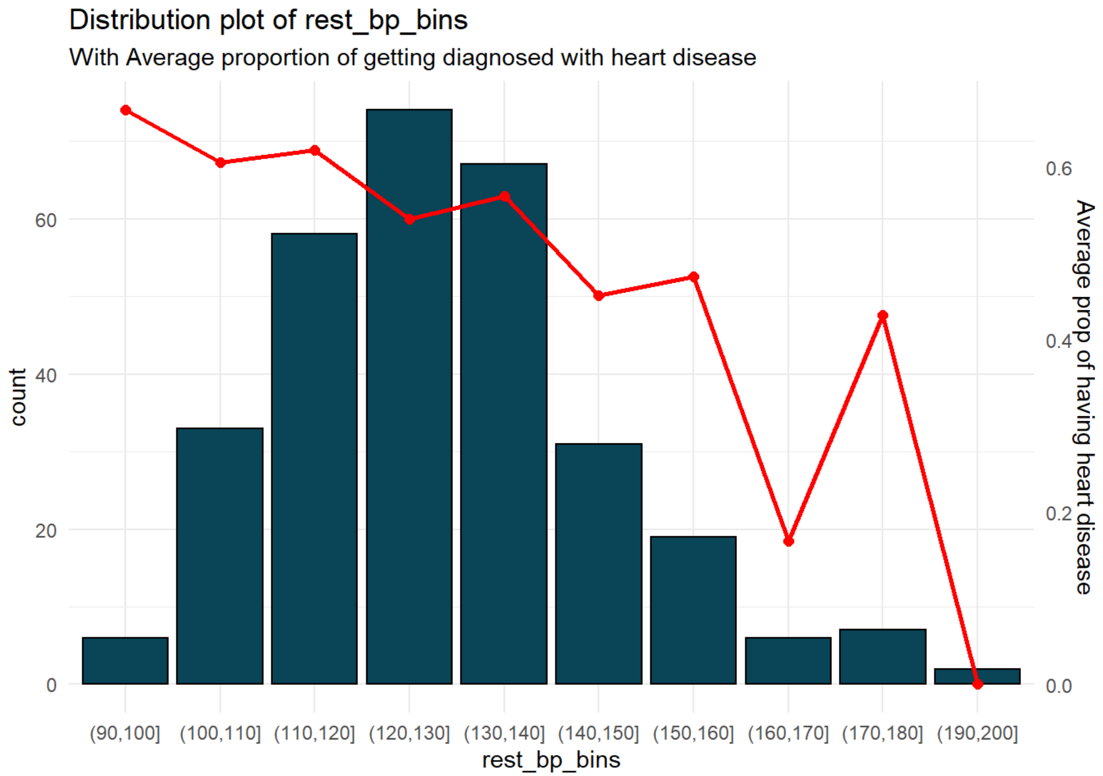
term	estimate	std.error	statistic	p.value	2.5 %	97.5 %
(Intercept)	3.036	0.756	4.014	0	1.585	4.557
xvar	-0.052	0.014	-3.841	0	-0.080	-0.026



Resting Blood Pressure

summary statistics

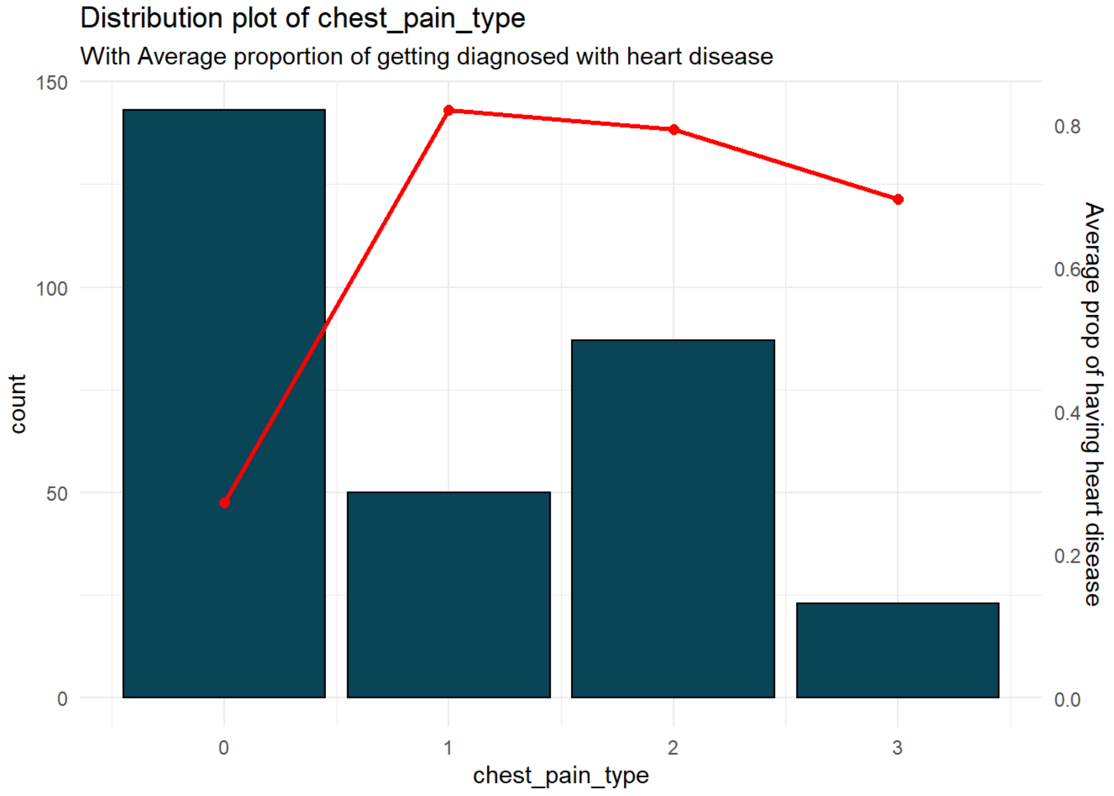
term	estimate	std.error	statistic	p.value	2.5 %	97.5 %
(Intercept)	2.409	0.904	2.665	0.008	0.663	4.219
xvar	-0.017	0.007	-2.489	0.013	-0.031	-0.004



Chest Pain Type

summary statistics

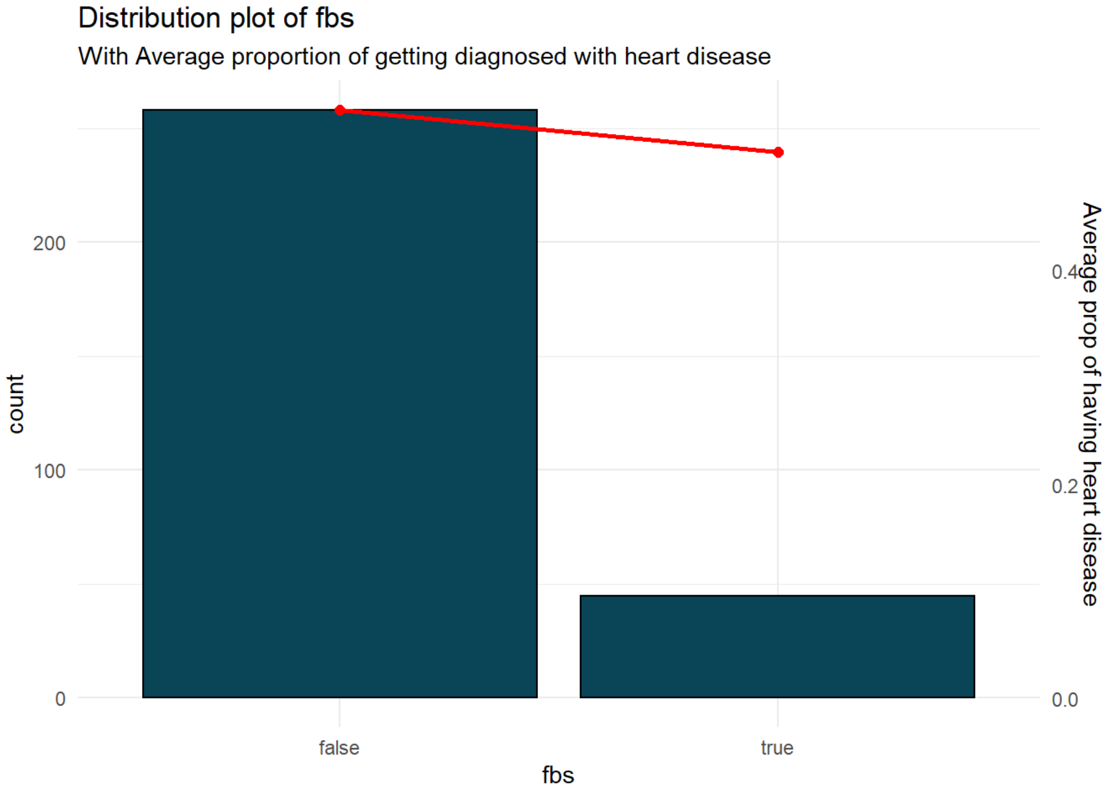
term	estimate	std.error	statistic	p.value	2.5 %	97.5 %
(Intercept)	-0.695	0.169	-4.122	0	-1.031	-0.369
xvar	0.984	0.139	7.079	0	0.720	1.266

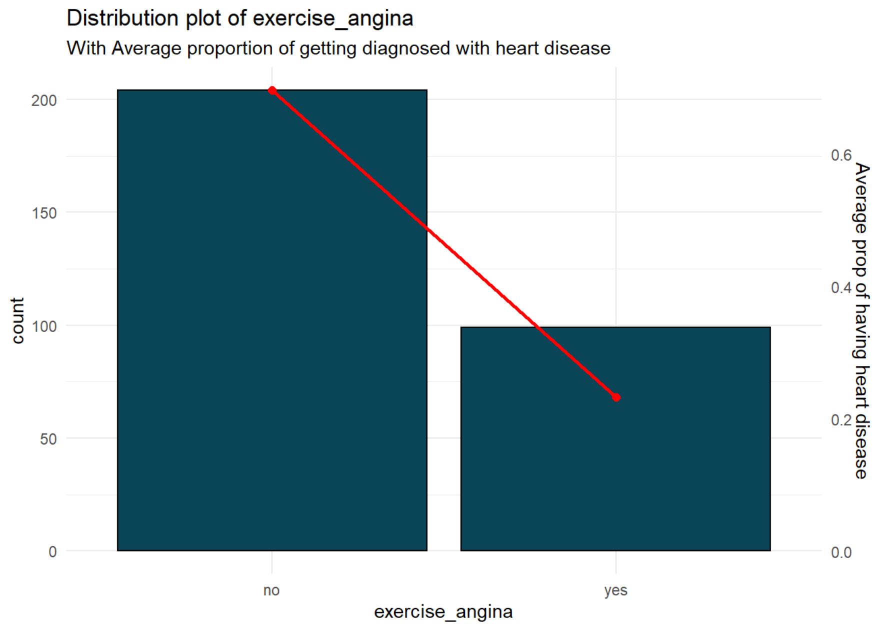


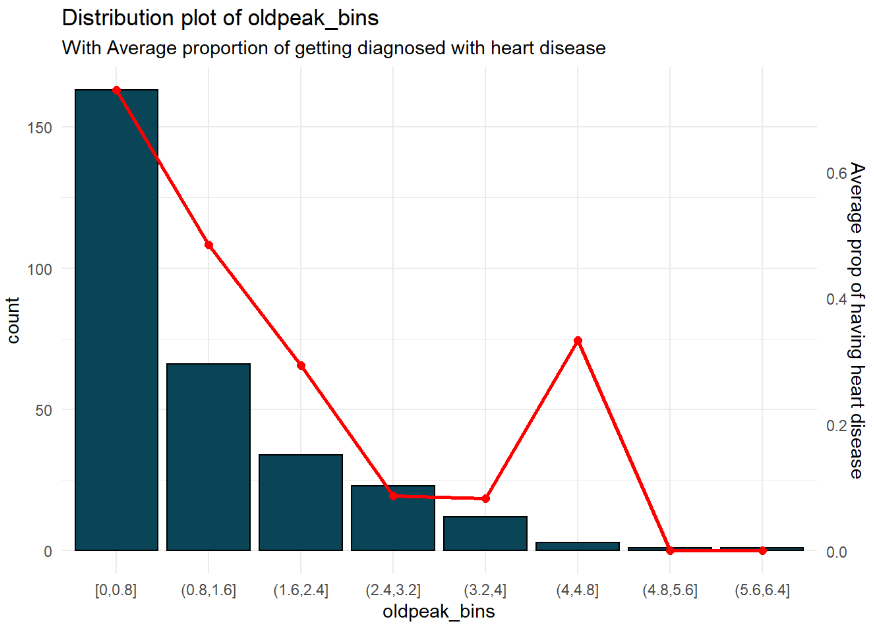
Fasting Blood Sugar > 120 mg/dl

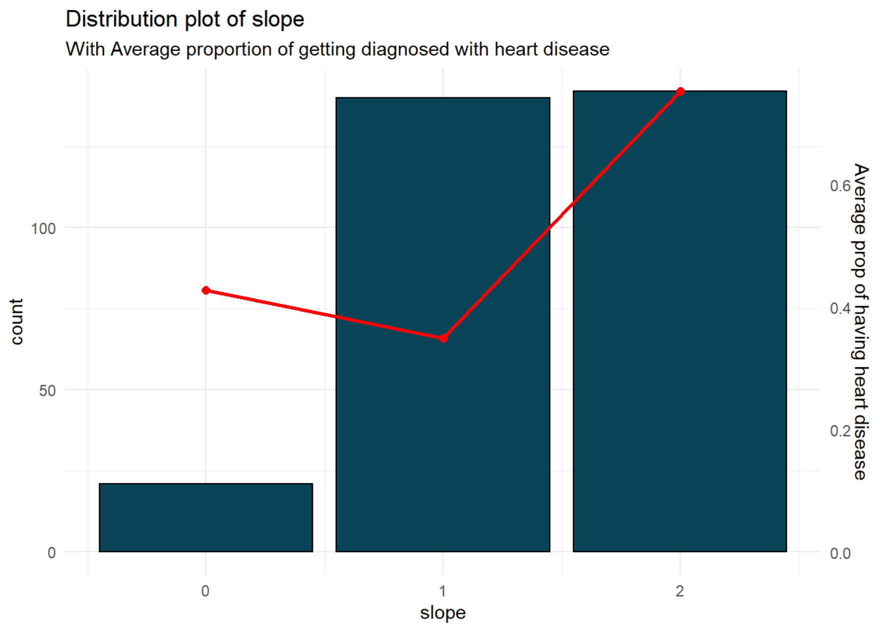
summary statistics

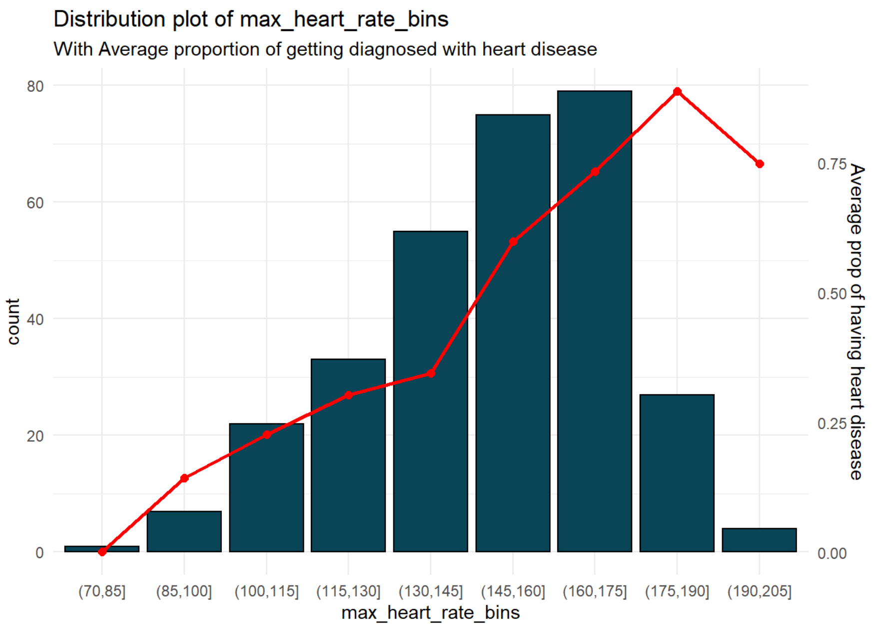
term	estimate	std.error	statistic	p.value	2.5 %	97.5 %
(Intercept)	0.202	0.125	1.616	0.106	-0.042	0.449
xvartrue	-0.158	0.323	-0.488	0.626	-0.794	0.480

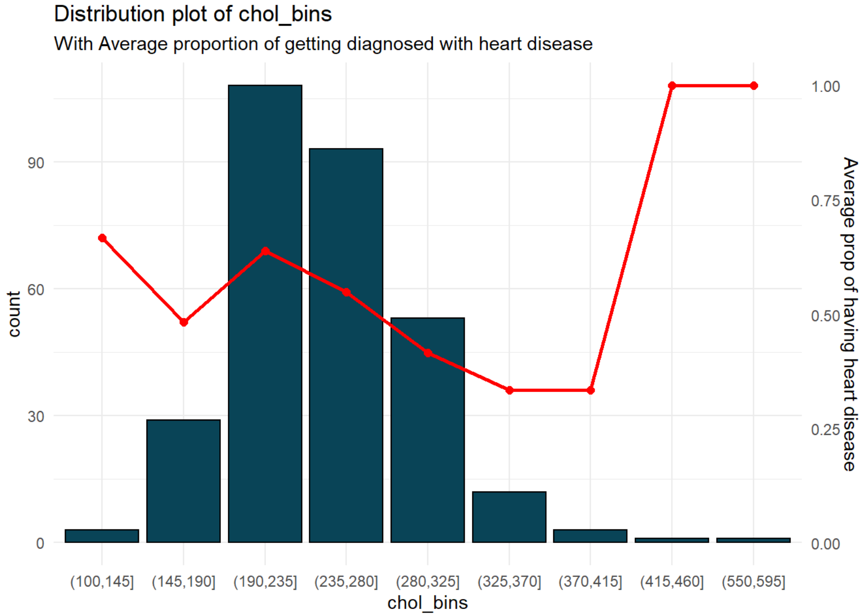


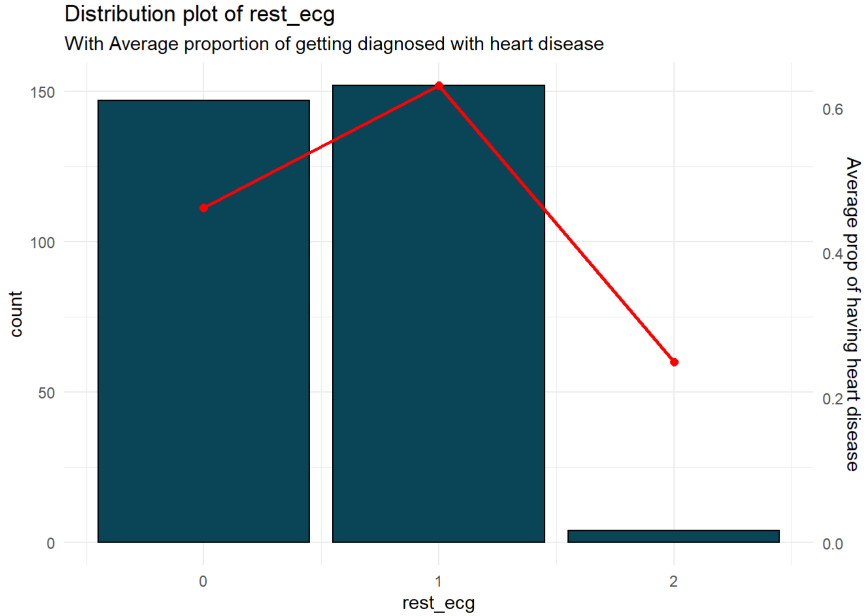

Distribution plot of exercise_angina
With Average proportion of getting diagnosed with heart disease
count
Average prop of having heart disease
200
150
100
50
0
0.6
0.4
0.2
0.0
no
yes
exercise_angina

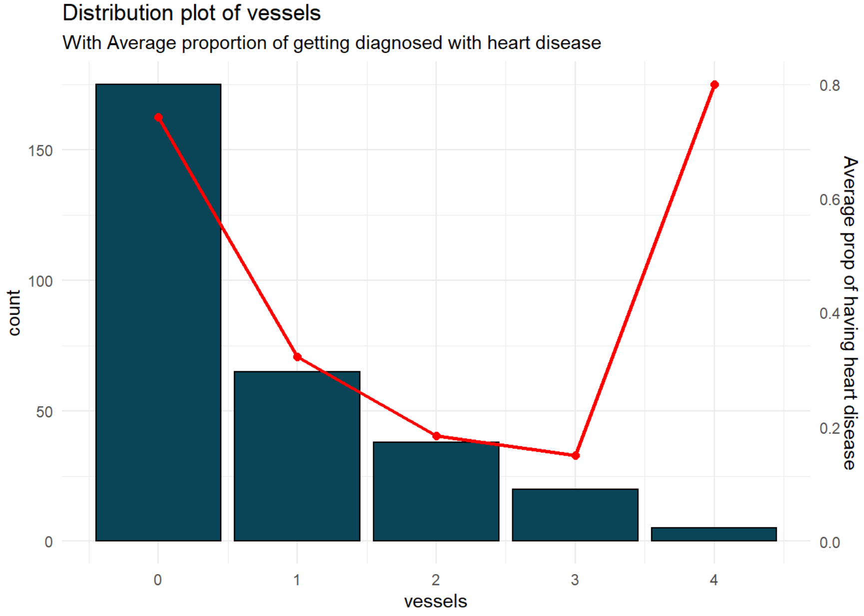

Distribution plot of oldpeak_bins
With Average proportion of getting diagnosed with heart disease
count
Average prop of having heart disease
150
100
50
0
0.6
0.4
0.2
0.0
[0,0.8]
(0.8,1.6]
(1.6,2.4]
(2.4,3.2]
(3.2,4]
(4,4.8]
(4.8,5.6]
(5.6,6.4]
oldpeak_bins


Distribution plot of slope
With Average proportion of getting diagnosed with heart disease
count
Average prop of having heart disease
100
50
0
0.6
0.4
0.2
0.0
0
1
2
slope


Distribution plot of max_heart_rate_bins
With Average proportion of getting diagnosed with heart disease
count
Average prop of having heart disease
80
60
40
20
0
0.75
0.50
0.25
0.00
(70,85]
(85,100]
(100,115]
(115,130]
(130,145]
(145,160]
(160,175]
(175,190]
(190,205]
max_heart_rate_bins


Distribution plot of chol_bins
With Average proportion of getting diagnosed with heart disease
count
Average prop of having heart disease
chol_bins
(100,145]
(145,190]
(190,235]
(235,280]
(280,325]
(325,370]
(370,415]
(415,460]
(550,595]


Distribution plot of rest_ecg
With Average proportion of getting diagnosed with heart disease
count
Average prop of having heart disease
rest_ecg


Distribution plot of vessels
With Average proportion of getting diagnosed with heart disease
count
Average prop of having heart disease
vessels


Distribution plot of thal
With Average proportion of getting diagnosed with heart disease
count
Average prop of having heart disease
thal

BOOTSTRAPPING

```
fit <- glm(y ~ x, data = train, family = "binomial")
```

`y` - response variable
`x` - one or more explain variables
`data` - training data
`family` - distribution

- Random partition: 0.60, 0.70, 0.80, 0.90, and 0.99
- Repeated: 10 times
- Features selection: from 1 to all variables
- Validation metric: AIC and AUC

```
for (i in xvars){
  for (j in nrounds){
    for (k in rp){
       modeling functions
    }
  }
}
```

FEATURES SELECTION

random_partition	nrow_train	nrow_test	nrounds	target	n_features	features	AIC	auc
0.99	299	4	10	dx_heart	4	chest_pain_type,oldpeak,vessels,thal	211.0654	0.8981895
0.90	272	31	10	dx_heart	4	chest_pain_type,oldpeak,vessels,thal	198.3173	0.8789868
0.60	181	122	10	dx_heart	4	gender,max_heart_rate,exercise_angina,vessels	177.1497	0.8771337
0.90	272	31	10	dx_heart	4	gender,max_heart_rate,exercise_angina,vessels	216.3574	0.8759481
0.99	299	4	10	dx_heart	4	age,gender,chest_pain_type,oldpeak	222.8361	0.8753062
0.70	212	91	10	dx_heart	4	gender,max_heart_rate,exercise_angina,vessels	189.4515	0.8744709
0.99	299	4	10	dx_heart	4	chest_pain_type,exercise_angina,vessels,thal	225.5231	0.8743765
0.80	242	61	10	dx_heart	4	chest_pain_type,oldpeak,vessels,thal	184.4960	0.8737395
0.70	212	91	10	dx_heart	4	chest_pain_type,oldpeak,vessels,thal	171.1942	0.8734447
0.99	299	4	10	dx_heart	4	chest_pain_type,rest_bp,max_heart_rate,oldpeak	233.8244	0.8734262
0.70	212	91	10	dx_heart	4	age,chest_pain_type,oldpeak,vessels	184.8035	0.8732091
0.80	242	61	10	dx_heart	4	age,chest_pain_type,oldpeak,vessels	198.1735	0.8729822
0.99	299	4	10	dx_heart	4	gender,exercise_angina,slope,vessels	227.7914	0.8726583
0.90	272	31	10	dx_heart	4	age,chest_pain_type,oldpeak,vessels	211.2985	0.8718788
0.99	299	4	10	dx_heart	4	age,gender,chest_pain_type,vessels	239.0499	0.8716321
0.90	272	31	10	dx_heart	4	chest_pain_type,max_heart_rate,oldpeak,vessels	201.9927	0.8716053
0.70	212	91	10	dx_heart	4	age,gender,chest_pain_type,oldpeak	184.0670	0.8713181

FEATURES SELECTION

random_partition	nrow_train	nrow_test	nrounds	target	n_features	features	AIC	auc
0.99	299	4	10	dx_heart	4	chest_pain_type,oldpeak,vessels,thal	211.0654	0.8981895
0.90	272	31	10	dx_heart	4	chest_pain_type,oldpeak,vessels,thal	198.3173	0.8789868
0.60	181	122	10	dx_heart	4	gender,max_heart_rate,exercise_angina,vessels	177.1497	0.8771337
0.90	272	31	10	dx_heart	4	gender,max_heart_rate,exercise_angina,vessels	216.3574	0.8759481
0.99	299	4	10	dx_heart	4	age,gender,chest_pain_type,oldpeak	222.8361	0.8753062
0.70	212	91	10	dx_heart	4	gender,max_heart_rate,exercise_angina,vessels	189.4515	0.8744709
0.99	299	4	10	dx_heart	4	chest_pain_type,exercise_angina,vessels,thal	225.5231	0.8743765
0.80	242	61	10	dx_heart	4	chest_pain_type,oldpeak,vessels,thal	184.4960	0.8737395
0.70	212	91	10	dx_heart	4	chest_pain_type,oldpeak,vessels,thal	171.1942	0.8734447
0.99	299	4	10	dx_heart	4	chest_pain_type,rest_bp,max_heart_rate,oldpeak	233.8244	0.8734262
0.70	212	91	10	dx_heart	4	age,chest_pain_type,oldpeak,vessels	184.8035	0.8732091
0.80	242	61	10	dx_heart	4	age,chest_pain_type,oldpeak,vessels	198.1735	0.8729822
0.99	299	4	10	dx_heart	4	gender,exercise_angina,slope,vessels	227.7914	0.8726583
0.90	272	31	10	dx_heart	4	age,chest_pain_type,oldpeak,vessels	211.2985	0.8718788
0.99	299	4	10	dx_heart	4	age,gender,chest_pain_type,vessels	239.0499	0.8716321
0.90	272	31	10	dx_heart	4	chest_pain_type,max_heart_rate,oldpeak,vessels	201.9927	0.8716053
0.70	212	91	10	dx_heart	4	age,gender,chest_pain_type,oldpeak	184.0670	0.8713181

FINAL MODEL

```
fit <- glm(dx_heart ~ gender + max_heart_rate + exercise_angina + vessels,
           data = data,
           family = "binomial")
```

- gender – sex (1 = male; 0 = female)
- max_heart_rate – maximum heart rate achieved
- exercise_angina - exercise induced angina (1 = yes; 0 = no)
- vessels - number of major vessels (0-3) colored by fluoroscopy

- dx_heart $= \begin{cases} 1 & \textit{has heart disease} \\ 0 & \textit{doesn't have heart disease} \end{cases}$

Summary Statistics

term	estimate	std.error	statistic	p.value	2.5 %	97.5 %
(Intercept)	-4.105	1.198	-3.426	0.001	-6.533	-1.819
genderfemale	1.363	0.336	4.060	0.000	0.721	2.042
max_heart_rate	0.033	0.008	4.396	0.000	0.019	0.049
exercise_anginayes	-1.687	0.335	-5.037	0.000	-2.360	-1.042
vessels	-0.828	0.161	-5.137	0.000	-1.159	-0.525

LIMITATION & FUTURE RESEARCH

- AUC is too high, need more data
- Explore more variables from different dataset
  - Other diseases the patient have
  - What type of work does the patient do
- Explore different types of model to compare
  - XGboost
  - Ridge regression
  - Random forest

MORE INFORMATION

- Data Set Information: https://archive.ics.uci.edu/ml/datasets/Heart+Disease
- Data Set Download: https://www.kaggle.com/ronitf/heart-disease-uci
- Applied Logistic Regression Textbook: https://www.google.com/books/edition/Applied_Logistic_Regression/bRoxQBIZRd4C?hl=en&gbpv=1&printsec=frontcover
- GLM R function: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm
- Bootstrapping Method: https://en.wikipedia.org/wiki/Bootstrapping_(statistics)

?

EMAIL: KDINH@SDSU.EDU