Final Project

Claire Wang & Xeno Hu

1. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death in the United States, with one person dying of the disease every 34 seconds (Centers for Disease Control and Prevention, 2022). Previous research has highlighted male gender, old age, obesity, abnormal cholesterol and fasting blood glucose as important predictors for CVDs, among many others (Damen et al., 2016).

As CVD incidence continues to soar, the need for an effective predictive model cannot be overstated. In this report, we aim to construct a predictive model based on the "Heart Failure Prediction Dataset" by assessing the 11 possible predictors (including chest pain type, resting blood pressure, cholesterol levels, maximum heart rate, resting ECG measurements, etc.), and we explore any correlations that might exist between these various predictors and the outcome of CVD. Our research question is, which combination of these factors is most effective in predicting CVDs? In answering this question, we seek to enhance the scientific community's understanding of CVDs and their associated risk factors, ultimately contributing to better prevention, early identification, and management of this critical disease for both individuals and populations.

1.1 Data Description

This "Heart Failure Prediction Dataset" includes 11 characteristics that can be utilized to anticipate the potential risk of a CVD; it is called "Heart Failure Prediction Dataset" because it is not uncommon for a CVD to lead to heart failure (Velagaleti et al., 2007). The dataset was formed by merging five heart-related datasets (the Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog Heart Datasets) based on their 11 common features.This dataset is currently the largest available heart disease dataset for research purposes, consisting of 918 observations ("Kaggle Heart Failure Prediction Dataset, 2021). Amongst the predictors, Age, RestingBP, Cholesterol, MaxHR, and Oldpeak are continuous variables. Sex, ChestPainType, FastingBS, RestingECG, ExerciseAngina, ST_Slope, and HeartDisease are categorical variables. A detailed data dictionary can be found in the Appendix.

1.2 EDA

Missing Data & Complete Case Analysis

As visualized in Figure 1, we found that 173 out of the 919 observations had a serum cholesterol level of 0 mm/dl. Since this is physiologically impossible, we decided to drop these observations.

We decide to use complete case analysis because it is relatively easy to implement, and it can reduce bias in the estimates of the variable of interest when the missing data are missing completely at random or missing at random. Despite a 20% sample loss, we chose it over imputation due to the lack of accurate prediction for missing cholesterol levels.

CVD Distribution

As shown in Figure 2, the distribution of CVD appears to be relatively even between patients with/without the disease, suggesting that there is sufficient data available for both categories of the response variable to perform further statistical analysis and develop a predictive model.

Key Variables

Previous research has identified age, resting blood pressure, serum cholesterol, the presence of exercise-induced angina, and maximum heart rate as indicators of the presence of heart disease (Hajar, 2017). In addition, sex could be an important predictor of CVDs because men are at a higher risk of developing CVDs than women (Maas, 2010). Thus, we will examine these key predictors in our EDA. As shown in Figure 3, patients with CVD clearly have a higher median age. Due to this difference, we will consider age as one of our predictor variables. As shown in Figure 4, out of the patients in our datset, There is a nearly even split between men who do/do not have a CVD; however, females are much more likely to not have CVD (80% of them do not). Due to this apparent sex difference, we will consider including sex as a predictor variable as well. Figure 5 shows that on average, patients diagnosed with CVD have higher resting blood pressure, which is consistent with the aforementioned literature findings. Figure 6 shows that over 80% of patients with exercise-induced angina have CVD, whereas only around 25% of patients without exercise-induced angina have CVD. Therefore, exercise-induced angina might be an important predictor for CVDs as well.

2. Methodology

2.1 Model Assumptions

We use a logistic regression model because we are trying to predict the log-odds of getting a CVD, a binary response variable. The two assumptions that are important in logistic regression are independence and linearity.

- **Independence**: It is reasonable to assume that independence is satisfied, as every patient's physiological conditions are independent of anyone else's. There could be cases of violations, such as if multiple observations are taken from individuals within the same family or household, there may be dependencies between those observations. However, for the most part, it is safe to assume that the assumption is satisfied.
- **Linearity**: The 5 continuous variables that are involved in all models are Age, RestingBP, Cholesterol, MaxHR, and Oldpeak. As shown in all five plots in Figure 6, the dots are randomly scattered across the empirical logits plots. Therefore, the linearity assumption is satisfied for the model.

2.2 All Subset Selection

We plan to start with an all subset selection method because it allows for a comprehensive exploration of all possible combinations and predicts the "best" model based on Mallow's Cp values.

```
 [1] 255.73724 142.24583  93.94848  78.45767  42.99584  33.63583  23.26726
 [8]  17.63977  11.64749  10.81918  10.67854  10.93762  12.19263  14.09560
[15]  16.00000
```

As shown in the table above, the model with the lowest Cp score is model 11. Model 11 includes 8 predictors: Age, Sex, ChestPainType, RestingBP, FastingBS, ExerciseAngina, Oldpeak, and ST_slope.

term	estimate	std.error	statistic	p.value
(Intercept)	-4.9502150	1.2597480	-3.929528	0.0000851
Age	0.0320180	0.0135006	2.371599	0.0177113
SexM	1.7941249	0.3073380	5.837627	0.0000000
RestingBP	0.0124850	0.0072187	1.729548	0.0837111
FastingBS	0.3225325	0.3260654	0.989165	0.3225824
ChestPainTypeATA	-1.7002181	0.3492558	-4.868117	0.0000011
ChestPainTypeNAP	-1.5968666	0.2964066	-5.387420	0.0000001
ChestPainTypeTA	-1.6370724	0.4715080	-3.471993	0.0005166
ExerciseAnginaY	0.8882308	0.2624020	3.385001	0.0007118
Oldpeak	0.4124948	0.1398666	2.949201	0.0031860
ST_SlopeFlat	1.2980847	0.5165720	2.512883	0.0119749
ST_SlopeUp	-1.2244713	0.5592877	-2.189341	0.0285721

To validate our predictor selection approach, we also performed stepwise selection in both directions on the dataset, yielding the same 8 predictors as the all subset selection method.

Thus, these 8 predictors were deemed meaningful and subsequently used to fit a logistic model, which we will call Model 1.

2.3 p-value elimination

Using a $\alpha = 0.05$ significance level, the p-values of "RestingBP" and "FastingBS" are 0.084 and 0.32, respectively, which is larger than 0.05, suggesting the these 2 predictors might not be statistically significant. Hence, we considered a second logistic model, Model 2, with the remaining 6 predictors.

term	estimate	std.error	statistic	p.value
(Intercept)	-3.6008458	0.9467300	-3.803456	0.0001427
Age	0.0389149	0.0131564	2.957871	0.0030977
SexM	1.7915558	0.2995667	5.980491	0.0000000
ChestPainTypeATA	-1.6554956	0.3444960	-4.805558	0.0000015
ChestPainTypeNAP	-1.5793224	0.2958402	-5.338430	0.0000001
ChestPainTypeTA	-1.5108657	0.4633904	-3.260460	0.0011123
ExerciseAnginaY	0.9354101	0.2605893	3.589595	0.0003312
Oldpeak	0.4214819	0.1378234	3.058129	0.0022272
ST_SlopeFlat	1.2392333	0.5104693	2.427635	0.0151976
ST_SlopeUp	-1.2746532	0.5515375	-2.311091	0.0208278

Finally, we fit a logistic model with all predictors included in order to provide a baseline for comparing the models. We will call this Model 3.

term	estimate	std.error	statistic	p.value
(Intercept)	-5.4373046	1.7625169	-3.0849659	0.0020358
Age	0.0313784	0.0148105	2.1186575	0.0341194
SexM	1.8655490	0.3134065	5.9524904	0.0000000
ChestPainTypeATA	-1.6731804	0.3544226	-4.7208630	0.0000023
ChestPainTypeNAP	-1.5730121	0.3029404	-5.1924797	0.0000002
ChestPainTypeTA	-1.6332529	0.4838117	-3.3758029	0.0007360
RestingBP	0.0117792	0.0072988	1.6138654	0.1065566
Cholesterol	0.0024955	0.0019773	1.2620586	0.2069277
FastingBS	0.2923999	0.3311265	0.8830458	0.3772115
RestingECGNormal	-0.2297888	0.2842091	-0.8085204	0.4187911
RestingECGST	-0.1746017	0.3941671	-0.4429637	0.6577920
MaxHR	0.0005807	0.0057810	0.1004456	0.9199906
ExerciseAnginaY	0.9073515	0.2671360	3.3965895	0.0006823
Oldpeak	0.4108355	0.1406671	2.9206235	0.0034933
ST_SlopeFlat	1.3038217	0.5197574	2.5085195	0.0121238

.metric	.estimator	.estimate
roc_auc	binary	0.9344641

.metric	.estimator	.estimate
roc_auc	binary	0.9329264

.metric	.estimator	.estimate
roc_auc	binary	0.9349323

term	estimate	std.error	statistic	p.value
ST_SlopeUp	-1.2100372	0.5655279	-2.1396597	0.0323823

2.4 Performance Evaluation: ROC & AUC

We have chosen to evaluate the performance of the models using the ROC (Receiver Operating Characteristic) curve and the AUC score (area under the ROC curve). In the ROC curves, the x-axis plots the false positive rate (specificity), and the y-axis plots the true positive rate (sensitivity). The curves basically illustrate the trade off between the specificity and the sensitivity. As we can see, all of the models produced ROC curves that are closer to the top left corner, indicating relatively high true positive rate and low false positive rate. The AUC (area under the ROC curve) is a powerful single metric to summarize the performance of the classifying model. The closer to the AUC is to 1, the better the model is able to distinguish between positive and negative classes, as the AUC is basically the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

As displayed in the table above as well as Figure 8 below, the overlapping ROC curves and similar AUC scores suggest that these three models have similar predictive performance. If all three models' performances are comparable, we would prefer to use the simplest model, Model 2, which only has 6 predictors.

2.5 F-tests

To further elucidate whether or not Model 2 (6 predictors) indeed performs similarly to Model 1 (8 predictors) and Model 3 (11 predictors), we will conduct 2 F-tests. Each test will compare the fit of a pair of nested models and allow us to determine whether the additional terms in the more complex model significantly improve its fit compared to the simpler model.

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
736	490.7443	NA	NA	NA
734	486.1341	2	4.610217	0.099748

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
736	490.7443	NA	NA	NA
730	483.5804	6	7.163848	0.3059602

The p-values are 0.100 and 0.306 (>0.05 significance level) when comparing Model 2 to Model 1, and Model 2 to Model 3, respectively. This indicates that there is not enough evidence to reject the null hypothesis that all of the slopes corresponding to the eliminated terms (Cholesterol, RestingECG, MaxHR, RestingBP, and FastingBS) are zero. Therefore, there is not enough evidence to conclude that either Model 3 or Model 1 are significantly better than Model 2. We would prefer to use the simpler Model 2 (that we came up with through all subset selection and p-value elimination) over the more complex Model 3 (includes all of the predictors), as it has fewer variables and possibly clearer interpretations.

2.6 Simple Holdout Method

Finally, we decided to use the simple holdout cross validation method, in which 80% of our dataset was used to train the model, and 20% was used to test its performance, in order to compare the performances of Model 2 and Model 3. The resulting AUC scores are 0.907 and 0.903 for Models 2 and 3, respectively, pointing to the fact that Model 2 could have better predictive power than Model 3, affirming that our model selection was successful, to a certain extent.

```
[1] 746  12
```

```
Area under the curve: 0.9066
```

```
Area under the curve: 0.9027
```

3.Results

The equation for our final model is:
log(odds) = -3.60 + 0.039 (Age) + 1.792 (Sex Male) - 1.656 (Atypical Anginal Chest Pain) - 1.579 (NonAnginal Chest Pain) -1.511 (Typical Anginal Chest Pain) + 0.935(Exercise Angina) + 0.421(OldPeak) + 1.239(St_SlopeFlat) -1.275(ST_SlopeUp)

After evaluating and comparing the model performances using the ROC curve, the AUC metric, and an F-test (all three models are nested), we see that all three models achieved very similar performance in terms of AUC score (0.934 for model 1, 0.933 for model 2, and 0.935 for model 3). We selected Model 2 as our final model. Although it seems to have the lowest AUC score among all three models, because our F-tests show that Model 2 is not significantly different from Model 1 and Model 3. This suggests that there is not sufficient evidence indicating that it is helpful to additionally include the other five predictors Cholesterol, RestingECG, MaxHR, RestingBP, and FastingBS in the model.

The intercept for our model represents a newborn female with no chest pain, no exercise-induced angina, no ST depression (0 degrees) on their ECG upon exercise that remains flat throughout exercise. The intercept value corresponds to log-odds of having a CVD vs. not having a CVD being 0.027 ($e^{-3.6}$) in this population. Consistent with existing scientific literature, we see in our chosen model that as people get older, their log-odds of developing a CVD is expected to increase by $e^{0.039=1.04}$ with every passing year. We also found that if a patient has exercise angina, their likelihood of having a CVD is $e^{0.935} = 2.55$ times higher than that of a patient who doesn't have exercise angina, when holding all other variables constant. This makes sense because exercise angina has to do with discomfort when heart muscle receives insufficient blood and oxygen during physical activity, which is mostly due to narrowing of arteries and can lead to CVDs and heart failure. An unexpected discovery from our model is that people who are asymptomatic for chest pain are predicted to have the highest log-odds of developing CVD, while controlling for all other variables. This could be due to the fact that the duration of patients' CVD diagnosis is unknown, and some patients may no longer experience chest pain due to treatment, despite still being diagnosed with CVD.

Discussion

Our model identified the main predictors for cardiovascular disease in this dataset and quantified their impact on the likelihood of having the disease. We found that traditional factors such as cholesterol and blood pressure lose their predictive power when richer data, such as electrocardiogram results, presence of certain types of chest pain, and exercise tests, is available. It is important to emphasize that the list of 11 variables under consideration in this study has been narrowed down by researchers from a list of 76, involving several previous research studies to confirm this cut. As a result, further reduction of these variables from a model including all 11 variables could be somewhat challenging.

After identifying 172 data points with zero cholesterol levels, we opted to exclude them from our study. Nonetheless, we noticed that these eliminated observations had a greater probability of presenting CVD than the average in the complete dataset. Therefore, the exclusion of these data points may have decreased the reliability of our analysis and introduced bias into our results. The validity of our study might be compromised due to the exclusion of this relevant information.

Our study is also limited by the non-representative sample, which only includes individuals admitted to hospitals for heart disease-related conditions. Therefore, our findings cannot be generalized to the broader population. In order to develop generalizable predictive models for

CVDs, it would make more sense to expand the sample size and collect data from the entire general population to facilitate early detection and prompt timely treatment.

An idea for improvement on our current model is adding interaction terms. We had a couple of ideas, including age & sex, exercise angina & old peak, and chest pain type & sex. These options were chosen after reading through current literature, which suggested that women generally having a later onset of heart disease compared to men (Giardina, 2000), there may be differences in the way men and women experience chest pain during a heart attack (Granot, 2004), and that exercise-induced angina and ST depression during exercise are both markers of ischemia and may therefore be related to each other in predicting heart disease risk (Bekkouche, 2013).

References

Bekkouche, N. S., Wawrzyniak, A. J., Whittaker, K. S., Ketterer, M. W., & Krantz, D. S. (2013). Psychological and physiological predictors of angina during exercise-induced ischemia in patients with coronary artery disease. Psychosomatic medicine, 75(4), 413–421. https://doi.org/10.1097/PSY.0b013e31828c4cb4

Centers for Disease Control and Prevention, National Center for Health Statistics. About Multiple Cause of Death, 1999–2020. CDC WONDER Online Database website. Atlanta, GA: Centers for Disease Control and Prevention; 2022. Accessed May 1, 2023.

Damen, J. A., Hooft, L., Schuit, E., Debray, T. P., Collins, G. S., Tzoulaki, I., Lassale, C. M., Siontis, G. C., Chiocchia, V., Roberts, C., Schlüssel, M. M., Gerry, S., Black, J. A., Heus, P., van der Schouw, Y. T., Peelen, L. M., & Moons, K. G. (2016). Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ (Clinical research ed.), 353, i2416. https://doi.org/10.1136/bmj.i2416

Downs, J. R., Clearfield, M., Weis, S., Whitney, E., Shapiro, D. R., Beere, P. A., Langendorfer, A., Stein, E. A., Kruyer, W., & Gotto, A. M., Jr (1998). Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: results of AFCAPS/TexCAPS. Air Force/Texas Coronary Atherosclerosis Prevention Study. JAMA, 279(20), 1615–1622. https://doi.org/10.1001/jama.279.20.1615

Fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [May 2, 2023] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

Giardina E. G. (2000). Heart disease in women. International journal of fertility and women's medicine, 45(6), 350–357.

Granot, M., Goldstein-Ferber, S., & Azzam, Z. S. (2004). Gender differences in the perception of chest pain. Journal of pain and symptom management, 27(2), 149–155. https://doi.org/10.1016/j.jpainsymman.2003.05.009

Lim, Y. C., Teo, S. G., & Poh, K. K. (2016). ST-segment changes with exercise stress. Singapore medical journal, 57(7), 347–353. https://doi.org/10.11622/smedj.2016116

Maas, A. H., & Appelman, Y. E. (2010). Gender differences in coronary heart disease. Netherlands heart journal : monthly journal of the Netherlands Society of Cardiology and the Netherlands Heart Foundation, 18(12), 598–602. https://doi.org/10.1007/s12471-010-0841-y

Velagaleti, R. S., & Vasan, R. S. (2007). Heart failure in the twenty-first century: is it a coronary artery disease or hypertension problem?. Cardiology clinics, 25(4), 487–v. https://doi.org/10.1016/j.ccl.2007.08.010

Appendix

Data Dictionary

Age: The age of the patient in years.
Sex: The gender of the patient, identified as Male (M) or Female (F).
ChestPainType: The type of chest pain experienced by the patient, classified as Typical Angina (TA), Atypical Angina (ATA), Non-Anginal Pain (NAP), or Asymptomatic (ASY).
RestingBP: The resting blood pressure of the patient in mm Hg.
Cholesterol: The level of serum cholesterol in mm/dL.
FastingBS: The level of fasting blood sugar, indicated as 1 if FastingBS $>$ 120 mg/dL and 0 otherwise.
RestingECG: The results of the patient's resting electrocardiogram, classified as Normal, having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of $>$ 0.05 mV), or showing probable or definite left ventricular hypertrophy by Estes' criteria (LVH).
MaxHR: The maximum heart rate achieved by the patient, measured in beats per minute (bpm) and ranging from 60 to 202.
ExerciseAngina: Whether the patient experienced exercise-induced angina, identified as Yes (Y) or No (N).
Oldpeak: The degree of ST depression induced by exercise relative to rest, measured in depression.
ST_Slope: The slope of the peak exercise ST segment, classified as Up (upsloping), Flat (flat), or Down (downsloping).
HeartDisease: The output class indicating the presence (1) or absence (0) of heart disease.

Figure 1: Missing Data

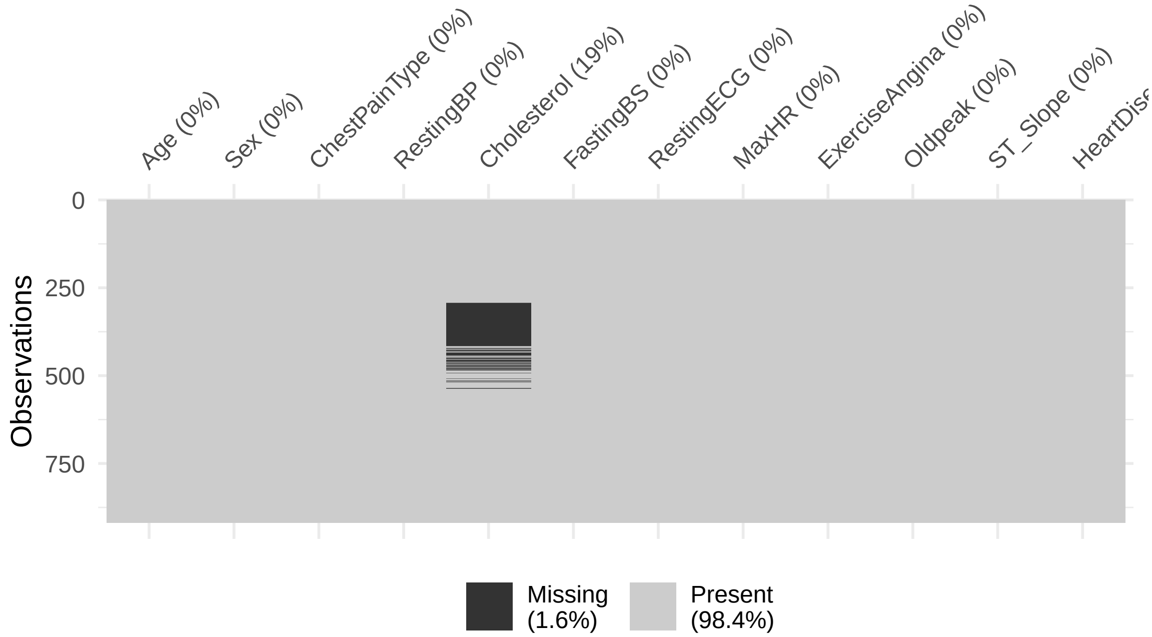


Figure 2: Distribution of Patients With/Without CVDs

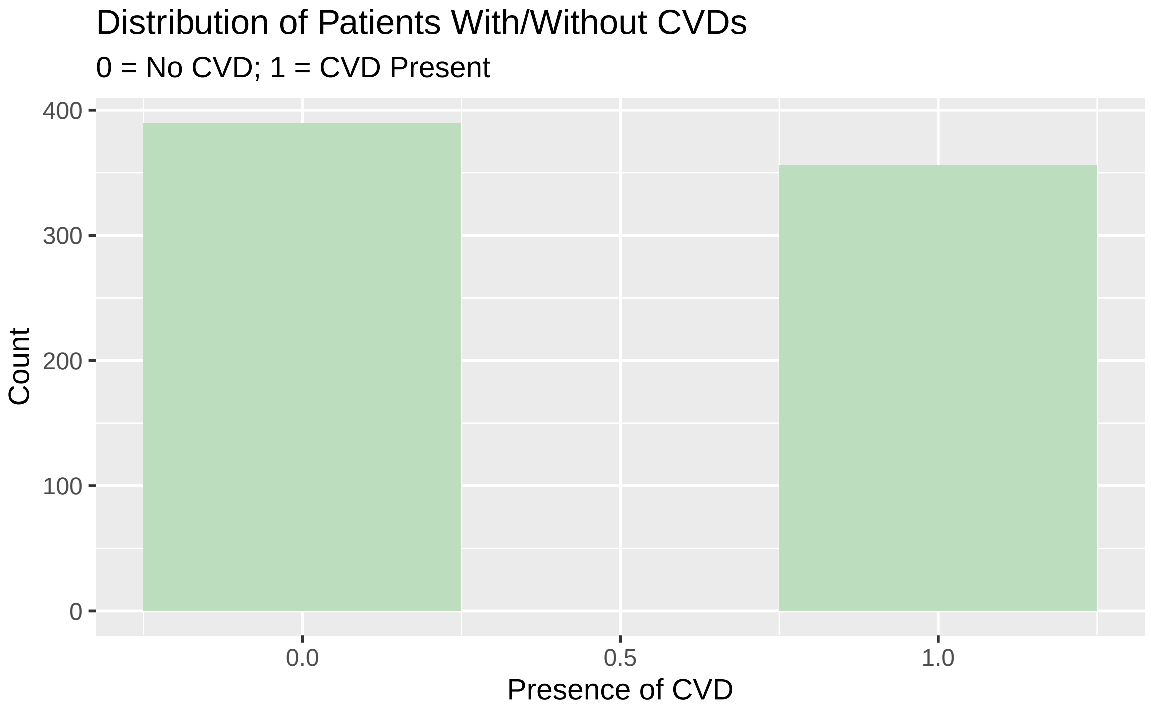


Figure 3: Box Plot for Age

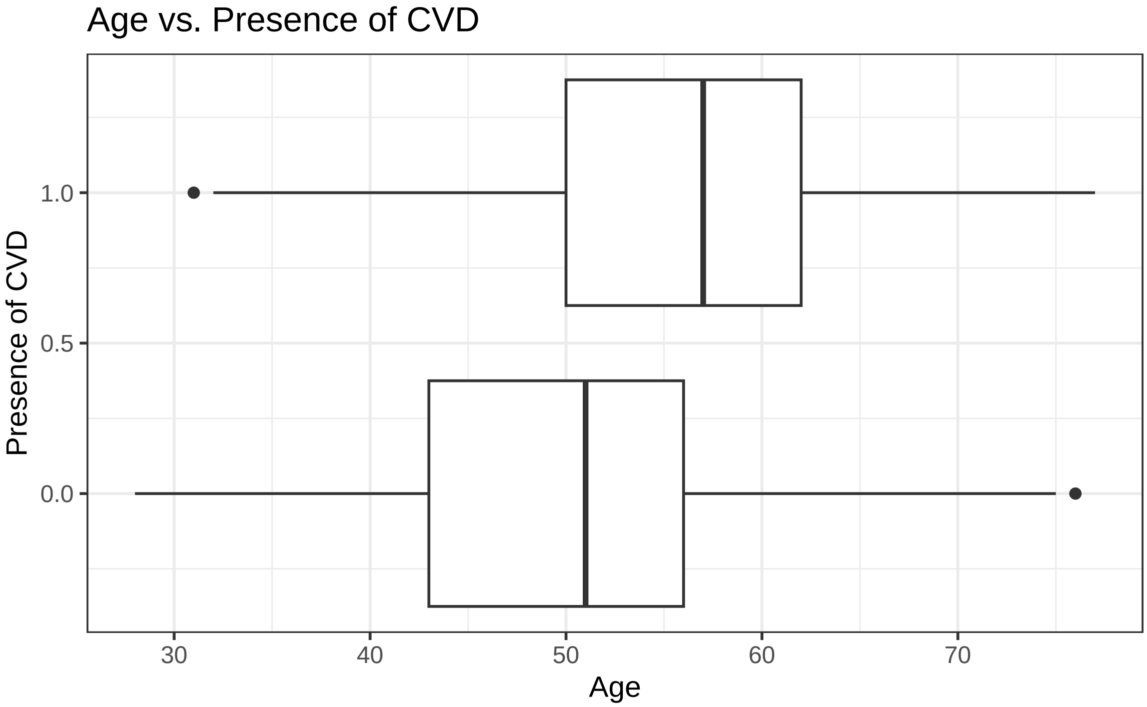


Figure 4: Proportion Plot for Sex

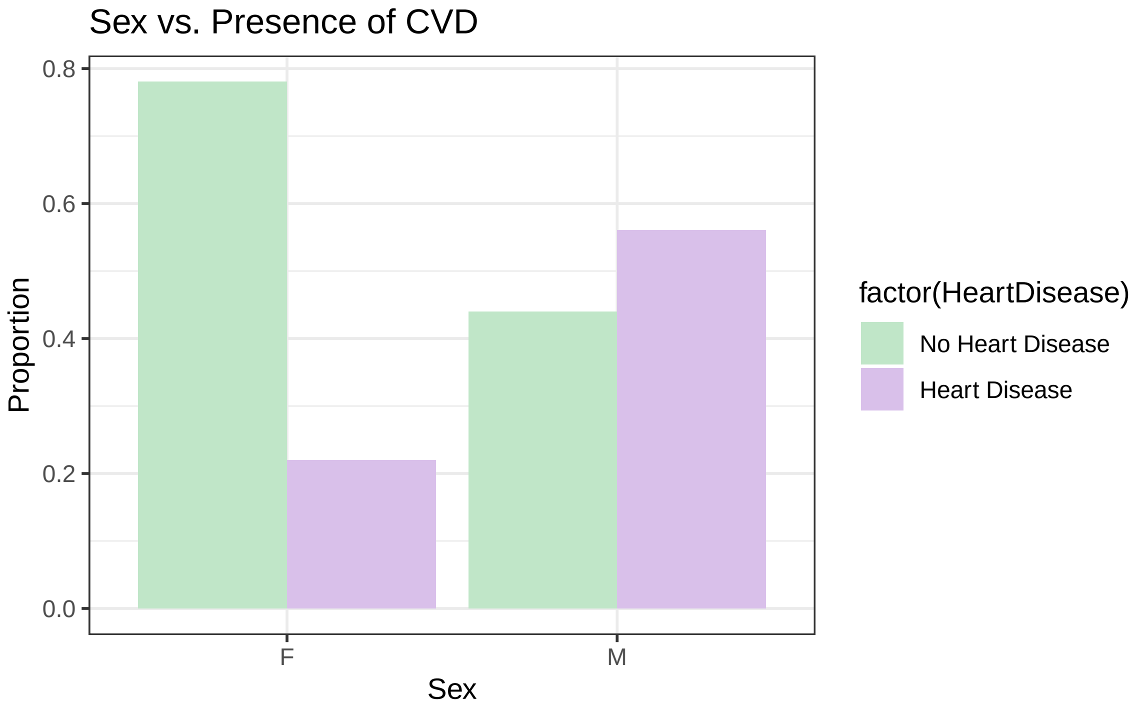


Figure 5: Box Plot for Resting Blood Pressure

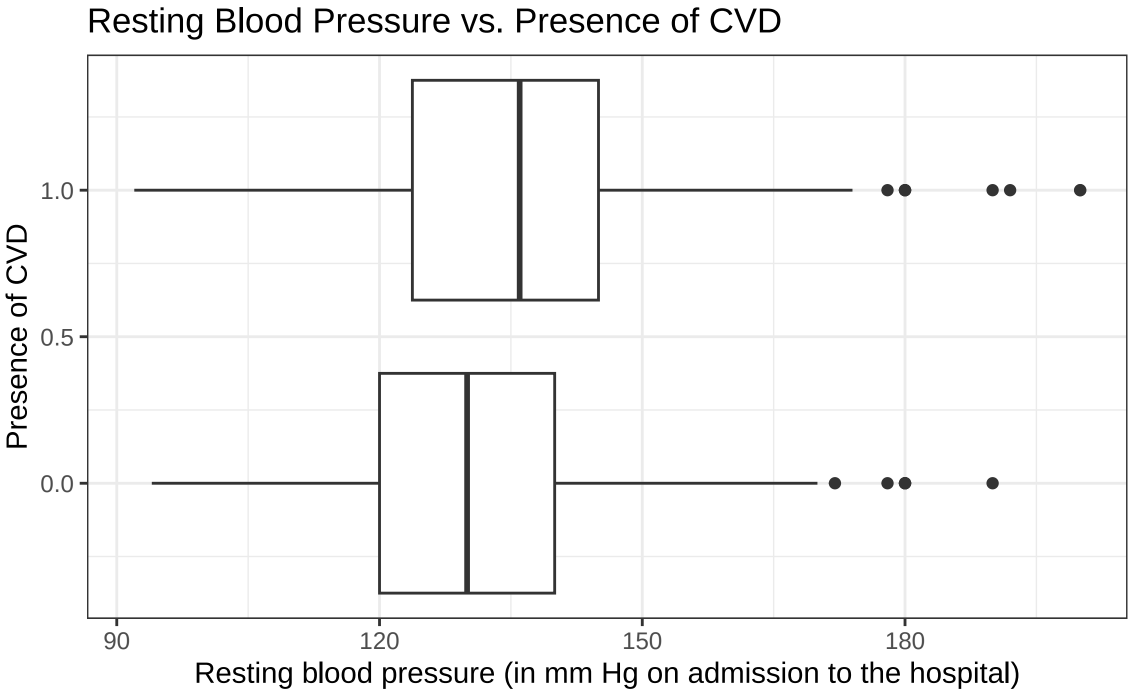


Figure 6: Proportion Plot for Exercise Angina

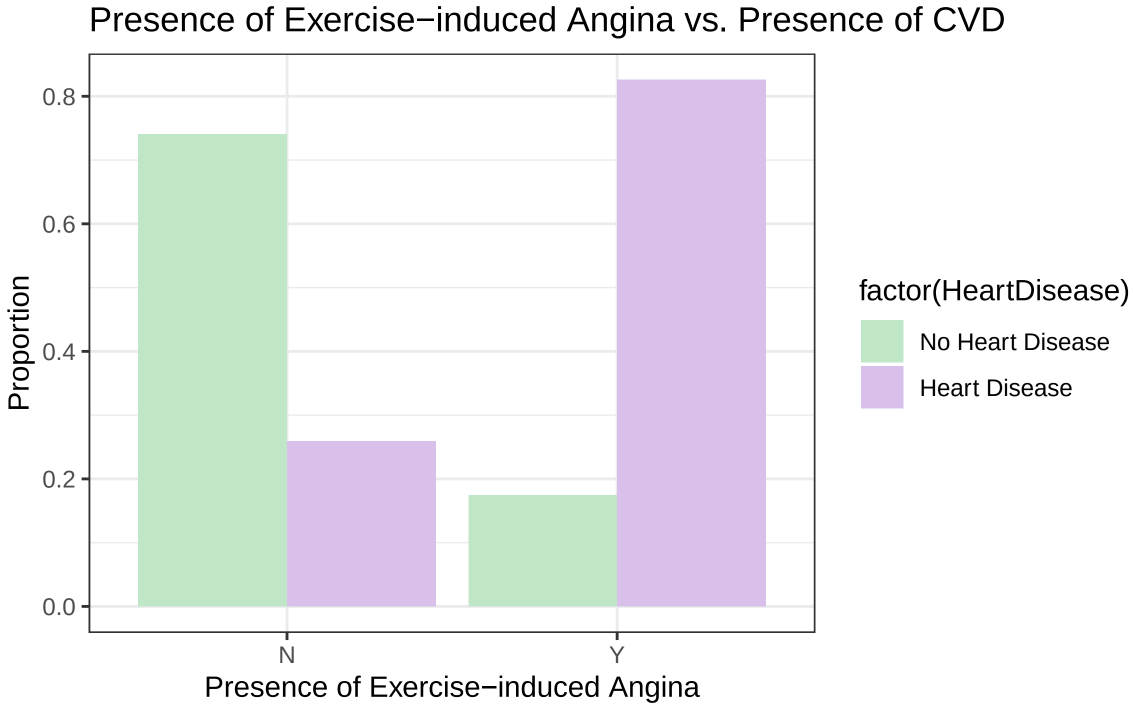


Figure 7: Linearity Assumption Diagnostic Plots for Continuous Variables

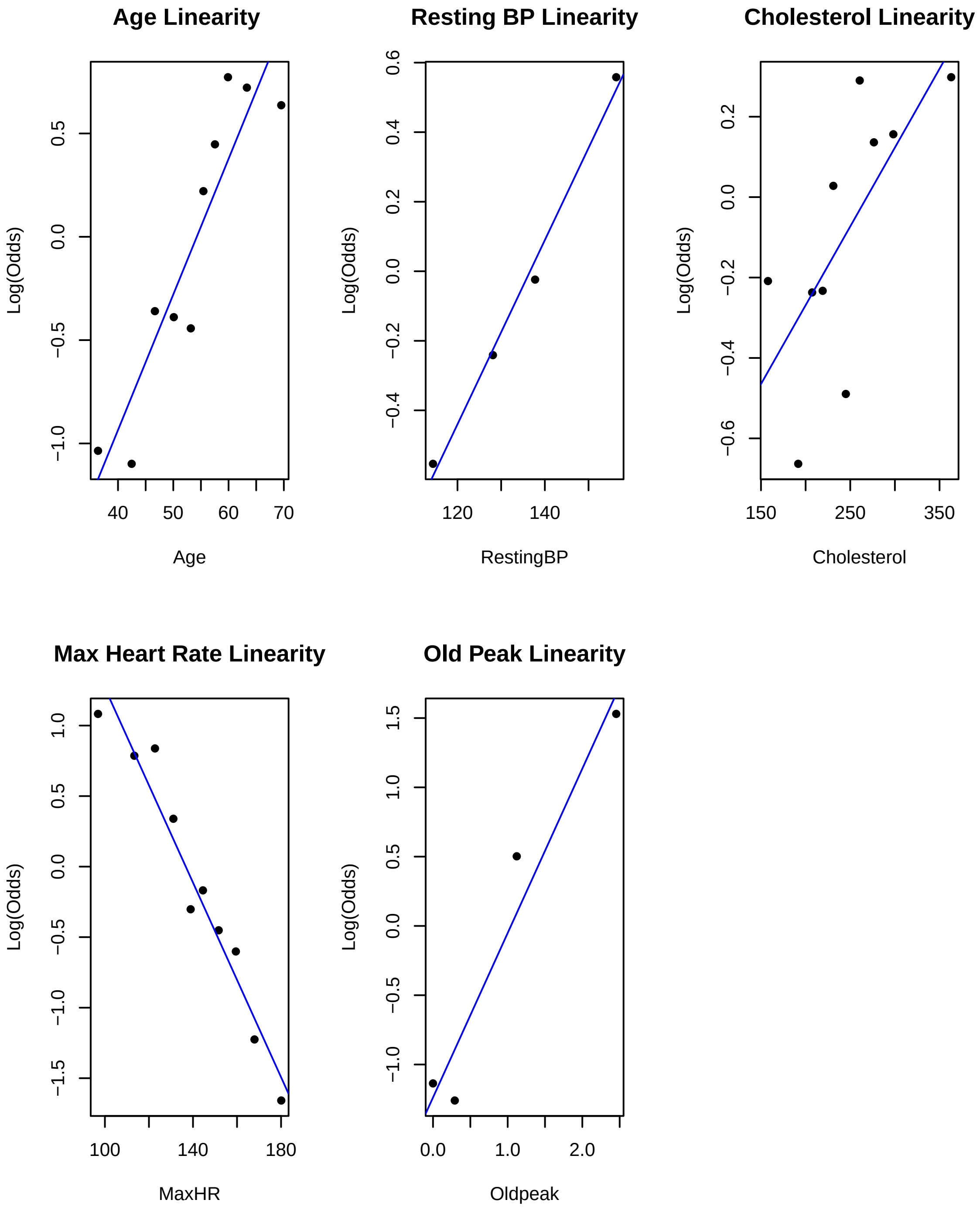
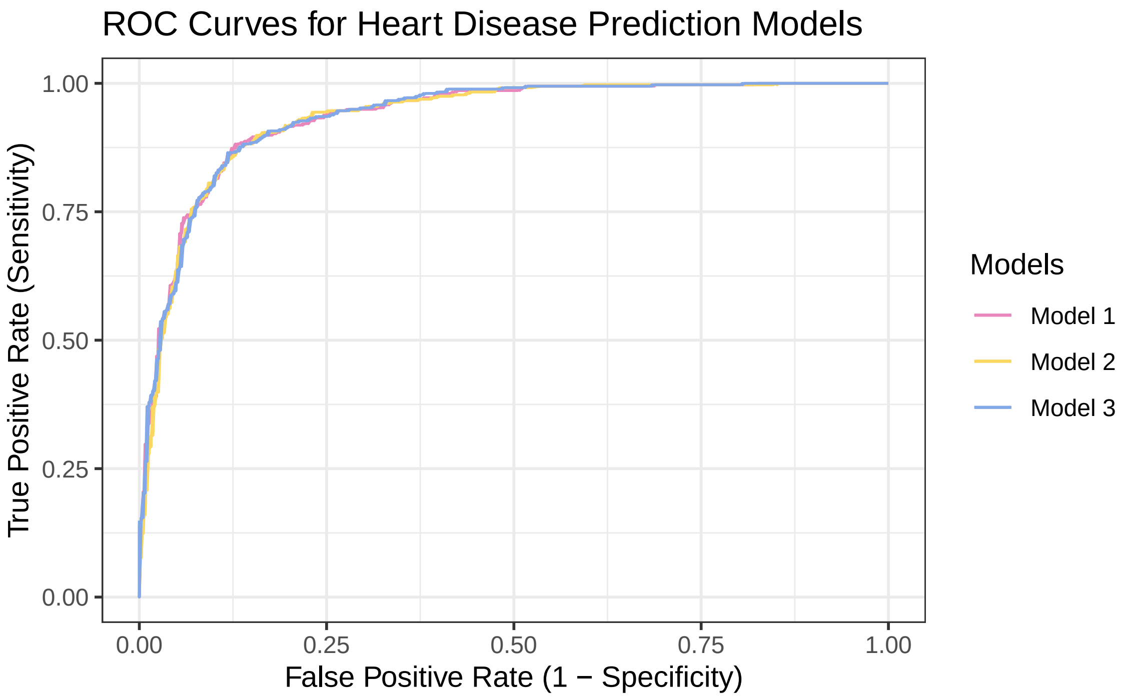


Figure 8: Overlapping ROC Curves for Models 1, 2, and 3


ROC Curves for Heart Disease Prediction Models
True Positive Rate (Sensitivity)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
False Positive Rate (1 − Specificity)
Models
Model 1
Model 2
Model 3