Data Science 100, Spring 2020

Investigating Causes of Heart Disease

Sophie Groenwold
perm #3424611

Priyanka Banerjee
perm #6409338

Abstract

In this work, we look at the Cleveland Heart Disease dataset, originally from the UCI Machine Learning Repository and sourced from Kaggle, to find an accurate predictor of heart disease. To accomplish this, we delve into the populations represented in our dataset, the features with the greatest influence on our prediction, and the extent to which our predictions are accurate. As a part of this endeavor, we utilize Principal Component Analysis (PCA), Logistic Regression Models, and a variety of exploration and visualization techniques. We found that the population of our sample data is predominantly male and in the range of 33-71 years. We discover that the main features influencing our prediction for heart disease are the number of vessels colored by fluoroscopy, presence of chest pain, ST depression induced by exercise relative to rest, sex, and the presence of thalassemia. With our methodology, we can accurately predict heart disease around 85% of the time given data points from the same population, though our model has a 2.632% false negative to 13.1518% false positive ratio, which is a good balance when dealing with inaccuracies in a healthcare domain as false negatives are potentially more dangerous for patients than false positives. We conclude that by using more sample data from populations that were not as prevalent in our original dataset and finding more relevant predictor variables, we can increase the accuracy of our model.

1 Introduction

Heart disease affects 30.3 million people in the United States annually, with life-changing consequences for each patient. It is the leading cause of death of most ethnicities for both men and women in the US, causing one in four deaths each year[1]. Finding which factors correlate with Coronary Artery Disease (CAD) can help many people get diagnosed early so that they can take mitigative measures and therefore prevent death from heart disease.

Some commonly known major risk factors of heart disease include high blood pressure, high blood cholesterol, smoking, obesity, diabetes, and physical inactivity [2]. Related factors whose combination creates a risk for cardiovascular disease are blood triglyceride and HDL cholesterol levels, age, gender, and psychosocial issues. Family history and alcohol intake have also been found to be risk factors for CAD [1]. The authors of this work have also found that in many datasets, human error in the measurement of blood sugar, blood pressure, and heart disease may need external methods for correction. This is helpful as it tackles an issue with variables present in our dataset.

To investigate the causes of heart disease, we use the Cleveland Heart Disease dataset [3], which we source from the UCI Machine Learning Repository[2]. From this, we use a cleaned version available on Kaggle[3]. The full dataset contains 76 attributes, most of which contain missing or extraneous values, so the publishers recommend the use of a subset of 14 attributes; this subset has been processed and contains no missing values. We have chosen this dataset because it contains a wide range of features, has been the choice for previous work that has looked into machine learning applications for health-related subjects and contains a presence attribute (whether or not the patient has heart disease) which is necessary for analysis with prediction tools.

[1] https://www.cdc.gov/heartdisease/facts.htm

[2] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[3] https://www.kaggle.com/ronitf/heart-disease-uci

Feature	Variable Type	Description
`age`	Continuous	Age of patient.
`sex`	Discrete	Sex of patient; binary value where 0 = female and 1 = male.
`cp`	Discrete	Presence of chest pain; value within $[0, 3]$, where typical angina = 1, atypical angina = 2, non-anginal pain = 3, and asymptomatic = 4.
`trestbps`	Continuous	Resting blood pressure; measured in mm Hg, upon admission to the hospital.
`chol`	Continuous	Cholesterol; measured in mg/dl.
`fbs`	Discrete	Whether or not fasting blood pressure is $>$ 120 mg/dl, binary value where 1 = true and 0 = false.
`restecg`	Discrete	Resting electrocardiographic results; value within $[0, 2]$, where 0 = normal, 1 = having ST-T wave abnormality, and 2 = showing probable or definite left ventricular hypertrophy.
`thalach`	Continuous	Maximum heart rate achieved.
`exang`	Discrete	Exercise induced angina; binary value where 1 = true and 0 = false.
`oldpeak`	Continuous	ST depression induced by exercise relative to rest.
`slope`	Discrete	Slope of the peak exercise ST segment; value within $[0, 2]$, where 1 = upsloping, 2 = flat, and 3 = downsloping.
`ca`	Discrete	Number of major vessels colored by flourosopy; value within $[0, 3]$.
`thal`	Discrete	Indication of thalassemia. Value from $[0, 3]$.
`target`	Discrete	Presence of heart disease in patient; binary value where no presence = 0 and presence = 1.

Table 1: Features of the Cleveland dataset, each with its type of variable and a description [3].

2 Questions of Interest

We have three primary questions of interest that are essential in understanding the causes of heart disease.

- What populations are represented in this dataset?
- To what extent can we accurately predict heart disease?
- Which features have the greatest influence in predicting whether or not an individual has heart disease?

3 Methods

As the demographic-specific attributes in our dataset are `age` and `sex`, our first question of interest is on dissecting them to gain insight on the balance of these populations. If there is an imbalance in either variable, this may create unintended bias in the model. Since the `sex` variable is binary, our goal is to see the count of each type; for this, we use a box plot. However, since `age` is a continuous variable, we opt to use a violin graph, to discern the distribution. For both, we use Altair[4]. Fewer samples of a particular age range would alert us to another potential bias in our model due to over- and under-representation.

For our next question of interest, we use a logistic regression model and the train_test_split capabilities provided by scikit-learn[5]. To measure accuracy, we use scikit-learn's `metrics.accuracy_score`,

[4] https://altair-viz.github.io/gallery/violin_plot.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

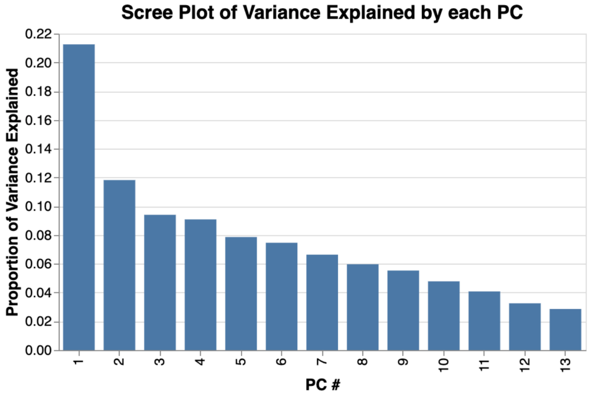

Figure 1: Variance explained by principal components. The first two PCs comprise of only 33.07% of the data.

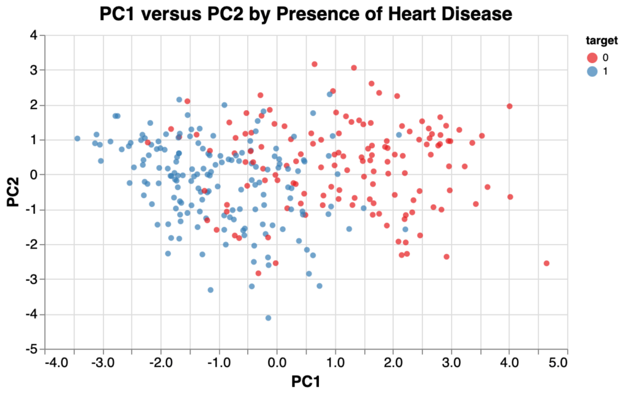

Figure 2: Scatter plot of PC1 versus PC2. Red indicates the presence of heart disease, and blue a lack thereof.

which returns the proportion of test samples that were correctly classified in our logistic model and we balance the model so that both the test prediction accuracy and the training prediction accuracy are maximized. Taking it one step further, since the dataset is regarding heart disease, getting a false negative score would be more harmful to the patients than a false positive score. Barring problems with medication when provided a false positive (though there are more tests given once a person is designated a heart risk, which may also help reveal a person's risk for cardiovascular disease), it is more dangerous for a patient if they are designated as not carrying heart disease when they do show the signs of cardiovascular problems. Therefore, we look at the proportions of false negatives and false positives in our inaccurately predicted points and fix the model so that it minimizes the number of false negatives by punishing those results in the data.

Lastly, to identify features that have the greatest influence in predicting the presence of heart disease, we could either identify the coefficients of our Logistic Regression Model or apply Principal Component Analysis (PCA) using Singular Value Decomposition (SVD)[6] to determine how high the rank of our data is, and see the which features contribute the most to each principal component. If the rank of our data is low, then we can proceed by using the principle components as our variables for our model. However, if the rank is high and each principal component utilizes a healthy portion of each variable, then it would be better to use the variables instead in the linear model. The PCA method helps get rid of non-contributing variables for our model and determine which ones have higher influences on the target variable.

4 Experimentation

4.1 Data

This Cleveland Heart Disease dataset (hereafter referred to as the Cleveland dataset) contains 303 patient instances and 14 features. The original dataset contained 76 features, but all published work with this dataset refers only to the subset of 14, and the dataset authors recommend that future work follows this direction. The dataset as retrieved from Kaggle has been preprocessed to only contain these 14 features and contains no missing values. A description of 14 features can be

[6] https://numpy.org/doc/stable/reference/generated/numpy.linalg.svd.html

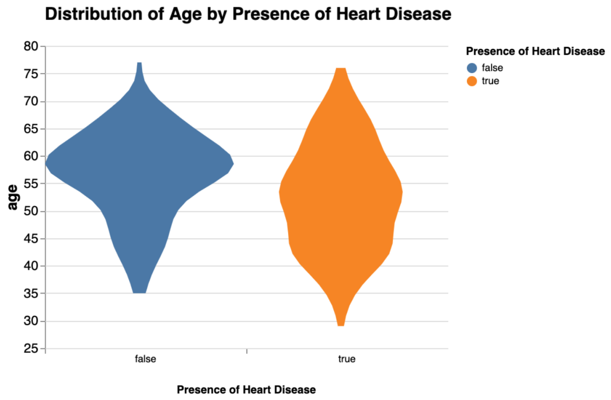

Figure 3: Distribution of age by presence of heart disease, as a violin graph.

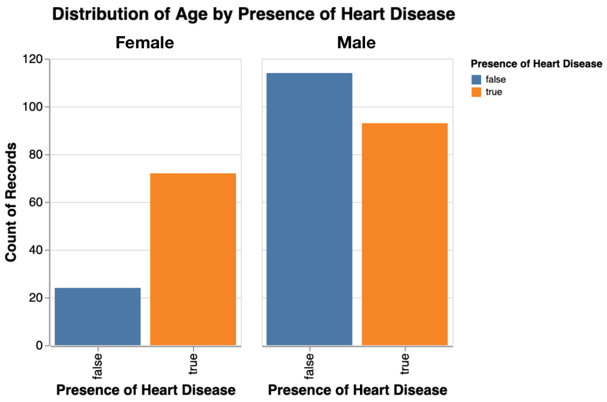

Figure 4: Distribution of sex by presence of heart disease, as a box plot.

seen in Table 1.

The patient instances of the Cleveland dataset were obtained from four medical facilities, with each with a primary collector: Andras Janosi, M.D. of the Hungarian Institute of Cardiology, Budapest; William Steinbrunn, M.D. of the University Hospital, Zurich, Switzerland; Matthias Pfisterer, M.D. of the University Hospital, Basel, Switzerland; and Robert Detrano, M.D., Ph.D. of the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. This dataset does not have a license, but the collectors ask to be cited any work that utilizes it [3].

Of the four principles of measurement, we primarily look into cost, since there are many kinds of cost associated with gathering medical data: financial, ethical, and privacy-related. In terms of monetary cost, the collection of the Cleveland dataset was sponsored by an independent donor, David W. Aha, working for the UCI Machine Learning Repository; no additional financial costs were placed on the patients for participation in the study. From an ethical perspective, the patients of this study were not subject to medical malpractice, or even treated differently than other patients due to their inclusion in the study; however, unequal proportions of demographic-specific information within our data (such as sex or age) may impact the results of our model and analysis. Within the context of privacy, the Cleveland dataset has redacted social security numbers and other identifying information upon release to the public.

As for the principle of distortion, previous work has noted that medical datasets are vulnerable to human error in data collection of fields such as blood pressure and blood sugar [1]; thus, this is a possible source of distortion for the Cleveland dataset, although there is no reason to believe that this dataset suffers from distortion any more than other datasets that have these same attributes.

4.2 Exploratory Analysis

4.2.1 Population Visualization

We investigate the distribution of instances across the two demographic-identifying attributes, `age` and `sex`, by creating two plots: a violin graph and a box plot, respectively. The violin graph for `age` can be seen in Figure 3; this graph shows that most patients of ages from roughly 58 - 62 are

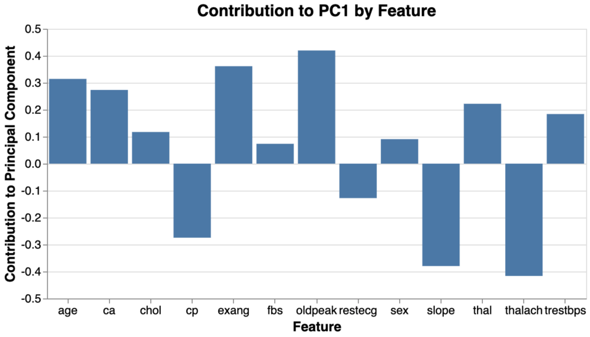

Figure 5: Weight of each feature in the first principal component (PC1).

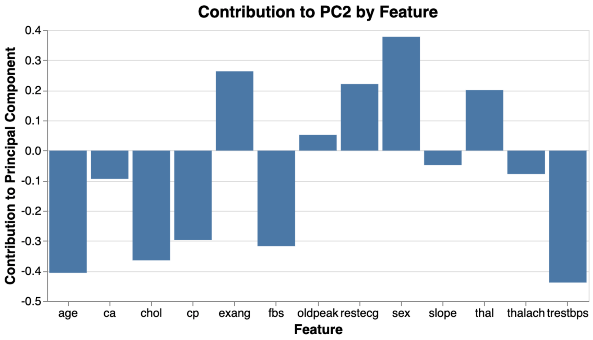

Figure 6: Weight of each feature in the second principal component (PC2).

diagnosed with heart disease. Figure 4, which plots `sex`, shows that there are unequal numbers of male and female patients from those who do have heart disease and from those who don't. We discuss the implications of these findings in Section 5.1.

4.2.2 Data Transformations and Models

To start our analysis, to make sure that disproportionate influence is not given to certain variables over others based on differences in scaling, we transform our data by centering and scaling the data using Z-score normalization. With this completed, we can use the data for our model. As the response variable in our dataset is categorical, the model we use is a Logistic Regression Model. We use test_train_split split the data to later be able to understand the accuracy of the model, making sure that the ratios for the target variable are equal in the training and test data. In our model itself, we use weights to incentivize false positives over false negatives as that would create a better model for our domain. Further analysis of this is related to our questions of interest, and as such are in Sections 5.2 and 5.3.

4.2.3 Discussion of Principal Component Analysis

Next, we utilize PCA to explore interesting patterns. We find that PCA does not significantly simplify the multidimensional data; a scree plot showing the variance each principal component is responsible for can be found in Figure 1. Despite this, we find that there is a noticeable degree of separation between the first and second principal components (see Figure 2). We also look into determining the relevance of our variables from how they contribute to each principal component. To do this, we plot the individual contributions of each dataset feature by taking the first row of V^T from our singular value decomposition for each feature; see Figures 5 and 6 for these plots regarding PC1 and PC2. We discuss the implications of these results later in Section 5.3.1.

5 Analysis, Results, and Interpretation

5.1 Dataset Population Representation

We note that the mode of selection for patients was likely not by random selection. The documentation for the Cleveland dataset was not clear in this aspect, but we anticipate several logistical constrains in collecting data for a diverse population. For instance, to justify collecting data for

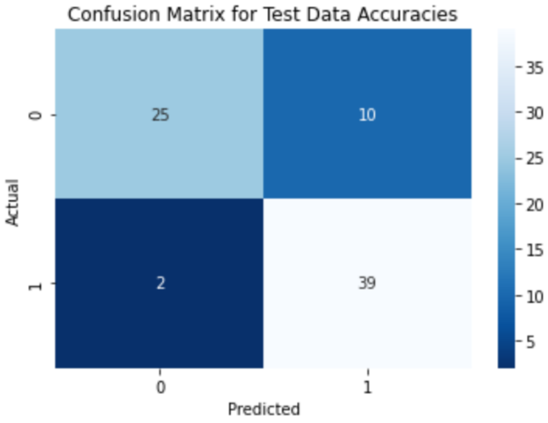

Figure 7: Confusion Matrix on the Test Data showing the distribution of the inaccuracies of the Logistic Model

patients with symptoms of heart disease, practitioners likely suspected that heart disease was a possibility; thus, there are more instances of `target` being 1/true (165 instances) versus 0/false (138 instances).

Returning to Figures 3 and 4, we see that the constrains likely influenced the unequal concentrations of patients based on age and sex. For instance, if there were an equal distribution across ages, the plots we currently see in Figure 3 would have the same width; however, they have most of their density from the age range of 40 - 65 years. In addition, the plot that shows the age distribution for no presence of heart disease displays a significant bump from ages 55 - 65 years, indicating that more people in this age range may think they have heart disease and thus look for an expert's diagnosis, but do not have heart disease after all.

Looking at statistics, we have determined that of the 303 patients included, 207 of them were men (68%) and 96 of them were women (32%). As sex is considered a predictor variable, having imbalanced proportions of sex would bias our model. Additionally, after looking at the age breakdown of the patients involved, we find that people under the age of 29 are not represented in the data, and there is only one data point regarding patients between 29-33 years old (29 yrs). Furthermore, people over 77 are not represented, and there are only 3 people sampled between the ages of 73 and 77 years old (74, 76, and 77 years). With disproportionate age group samples, the impact of the variable `age` can skew the algorithm. For example, if the one patient that is of age 29 years does not have a heart condition, then the algorithm could be biased against thinking that people of that age group could be having the disease. With random age group samples of equal proportion, this algorithmic bias could be resolved.

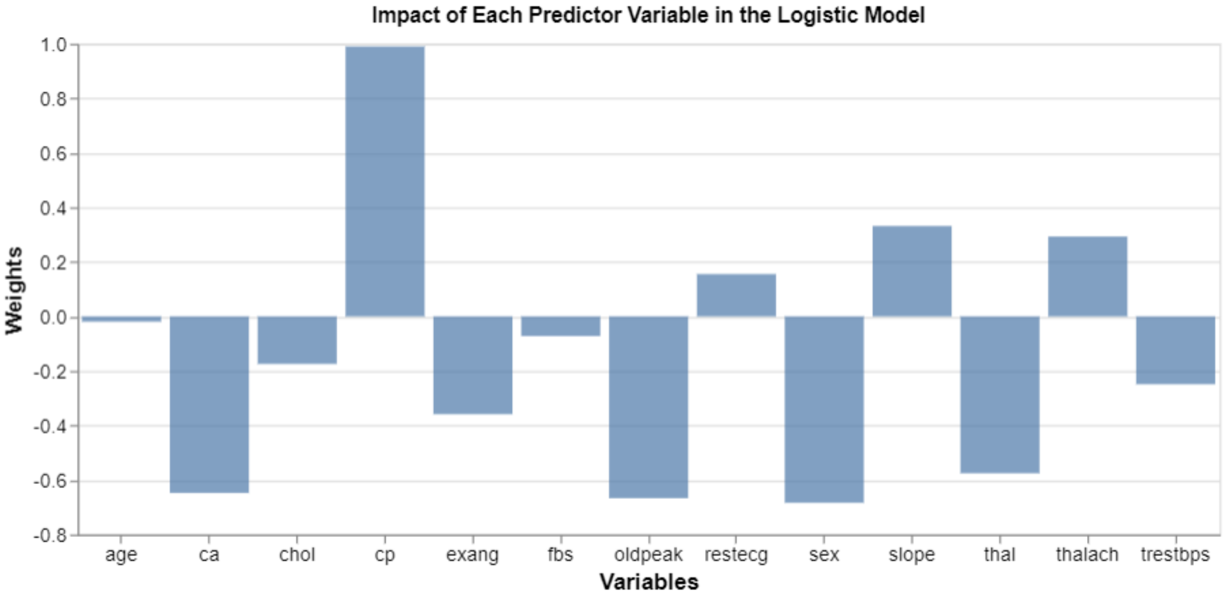

Figure 8: Weight of each feature in the final Logistic Model

5.2 Accurately Predicting Heart Disease

With the final Logistic Model, as discussed in Section 4.2.2, we look at the accuracy of the model in question. On a basic level, the test accuracy of our model is 84.2105% and the training accuracy of our model is 85.0220%. To understand the breakdown of the incorrect predictions, we create a confusion matrix, as shown in Figure 7. We find that balancing the ratio of the target values in the training data and test data does help not only increase the accuracy of the model but also reduces the instances of false negatives in the model. With a low false negative percentage and a high test/train accuracy, this model does help predict heart disease better than chance and is therefore a valid model. The extent that this model is accurate for our test data is 84.2105%, though, the accuracy could decrease if given points outside the age range of the data points used.

5.3 Influential Features

5.3.1 Principal Component Analysis

First, we discuss the results of our scree plot in Figure 1. As our first principal component is responsible for only 21.25% of the variance in the Cleveland dataset, and the second only 11.82%, we conclude that this data is high in rank and that each of the features in the Cleveland dataset is important in predicting the presence of heart disease.

However, we see from Figure 2 that there is a noticeable degree of separation when plotting a scatterplot of our two greatest principal components. We find that a value of 1.0 for PC1 is a rough cutoff for the separation between patients with and without heart disease. This indicates that PC1 and PC2 are describing different aspects of the data; therefore, even though the data has a high rank, there is the potential for high predictive accuracy in modeling tools, as seen from the previous section on Logistic Regression.

We also find that the distribution of each feature's importance is generally spread out, indicating that each feature is important in its contribution to the principal components. When looking at our PC1 and PC2 feature contribution plots, we notice that features that have a high contribution to PC1 generally have a low contribution for PC2: for instance, `age` has a weight of 0.3142 for PC1 but -0.4062 for PC2, and `trestbps` has a weight of 0.2220 for PC1 but 0.2007 for PC2. This supports the separation seen in 2, as although it is not clear what exactly each of our principal components is capturing, they are capturing different aspects of the data.

5.3.2 Logistic Regression

To find the more influential features through Logistic Regression, we look at the weights that the model places on each variable, as shown in Figure 8. We can see that though each variable has an impact, the variables `ca`, `cp`, `oldpeak`, `sex`, and `thal` have higher impacts on the model as shown by them being the farthest distance from the center zero line. This model is also surprising from the fact that `age` is not a higher impact predictor, with a weight of -0.01923556 in our model. Though our model is not 100% accurate on both our training and test data, we did hypothesize age to be a bigger factor than it ended up being.

6 Conclusion and Future Work

Overall, based on our exploration, we found that the populations represented are a 68%/32% split between male and female and the age range for the sample population is 29-77 years, though predominately the age range is 34-71 years. We can accurately predict heart disease with a test accuracy of 84.2105% and a false negative percentage of 2.632% with the model and data that we used. Finally, the variables `ca`, `cp`, `oldpeak`, `sex`, and `thal` have greater influence when predicting whether an individual has heart disease.

A challenge we faced was an attempt to merge the original Processed Cleveland Dataset from the UCI Machine Learning Repository with the dataset we sourced from Kaggle. We wanted to merge the sets as the Kaggle version had our response `target` variable, which determines the presence of heart disease, as a binary categorical variable, while the dataset in the UCI MLR had `target` as a categorical variable with a range of [0, 3]. The increase in data was appealing, though our endeavor was thwarted when we realized that there were inconsistencies in the data points of the two datasets. This alerted us to the fact that our dataset was further cleaned before being posted on Kaggle.

In the future, we could extend our analysis in many ways. In the original Cleveland dataset, there were 76 variables. We could add more of those variables or look at other potential predictors for heart disease, such as the age of first pregnancy for women, and see to what extent they influence our target variable. Additionally, though `age` did not turn out to be as impactful as a factor than previously believed, as the age sampling in our dataset is mostly between the ages of 34-71 years, we are less sure of the accuracy of our model if provided with patients that are outside that age range. Getting more balanced age range sampling and sex sampling for our training data would help make the model more accurate.

References

[1] Latha CBC and Jeeva SC. “Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques”. In: *Informatics in Medicine Unlocked* 16 (2019).

[2] Rachel Hajar. “Risk Factors for Coronary Artery Disease: Historical Perspectives”. In: *Heart views : the official journal of the Gulf Heart Association* 18 (2017), pp. 109–114.

[3] Andras Janosi et al. *UCI Repository of Machine Learning Databases*. 1988.

7 Appendix

7.1 Full page versions of Figures Shown

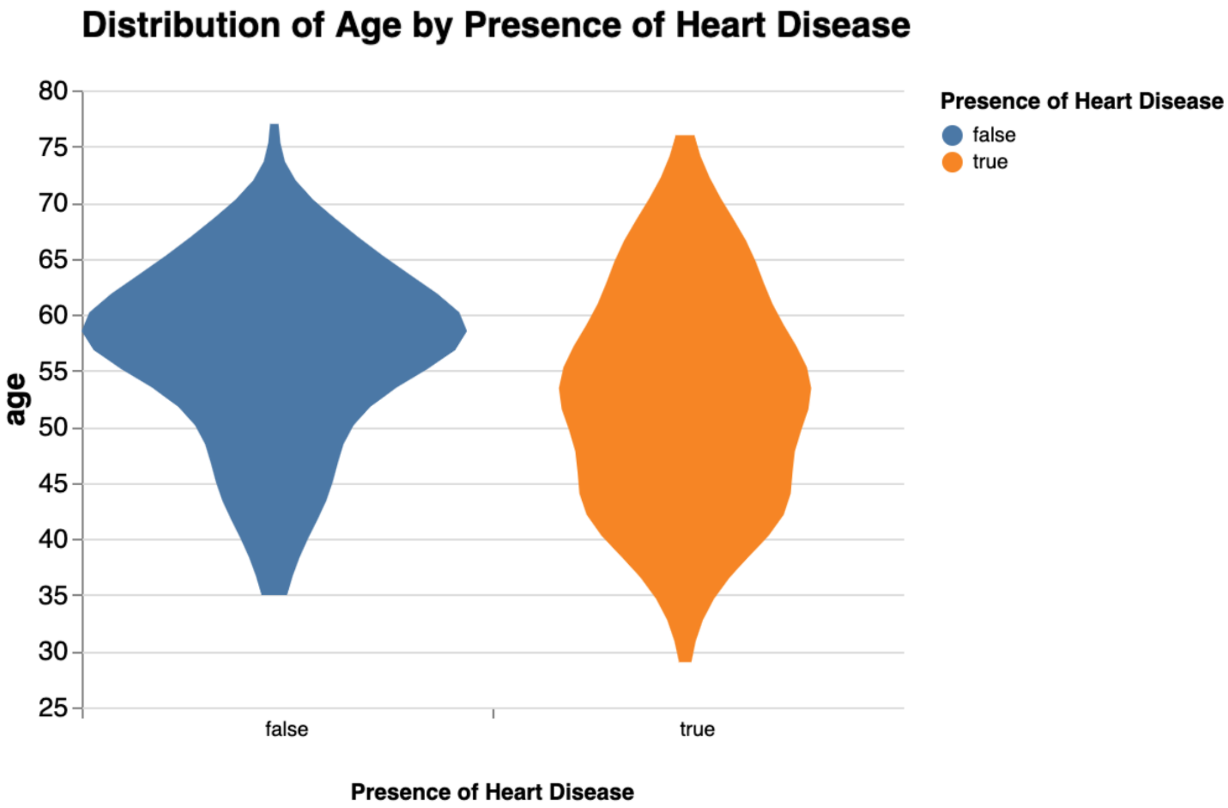

Figure 1: Distribution of age by presence of heart disease, as a violin graph.

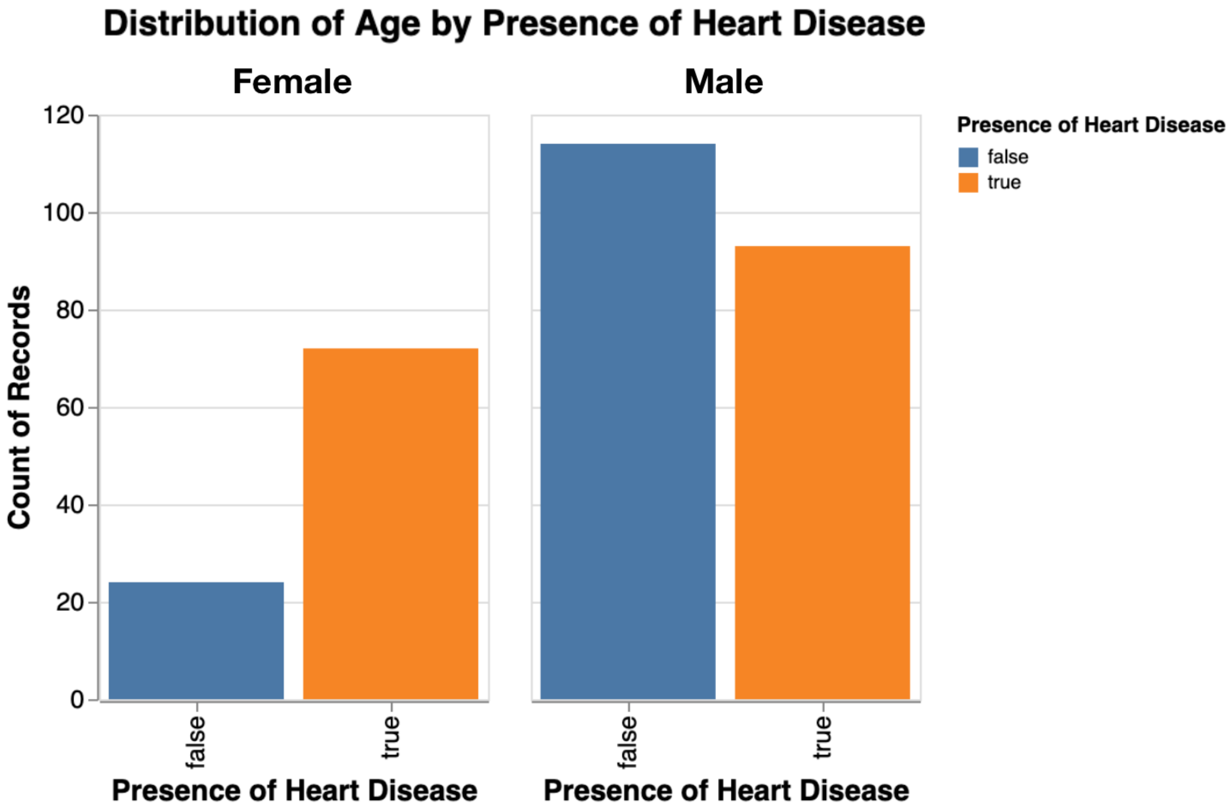

Figure 2: Distribution of sex by presence of heart disease, as a box plot.

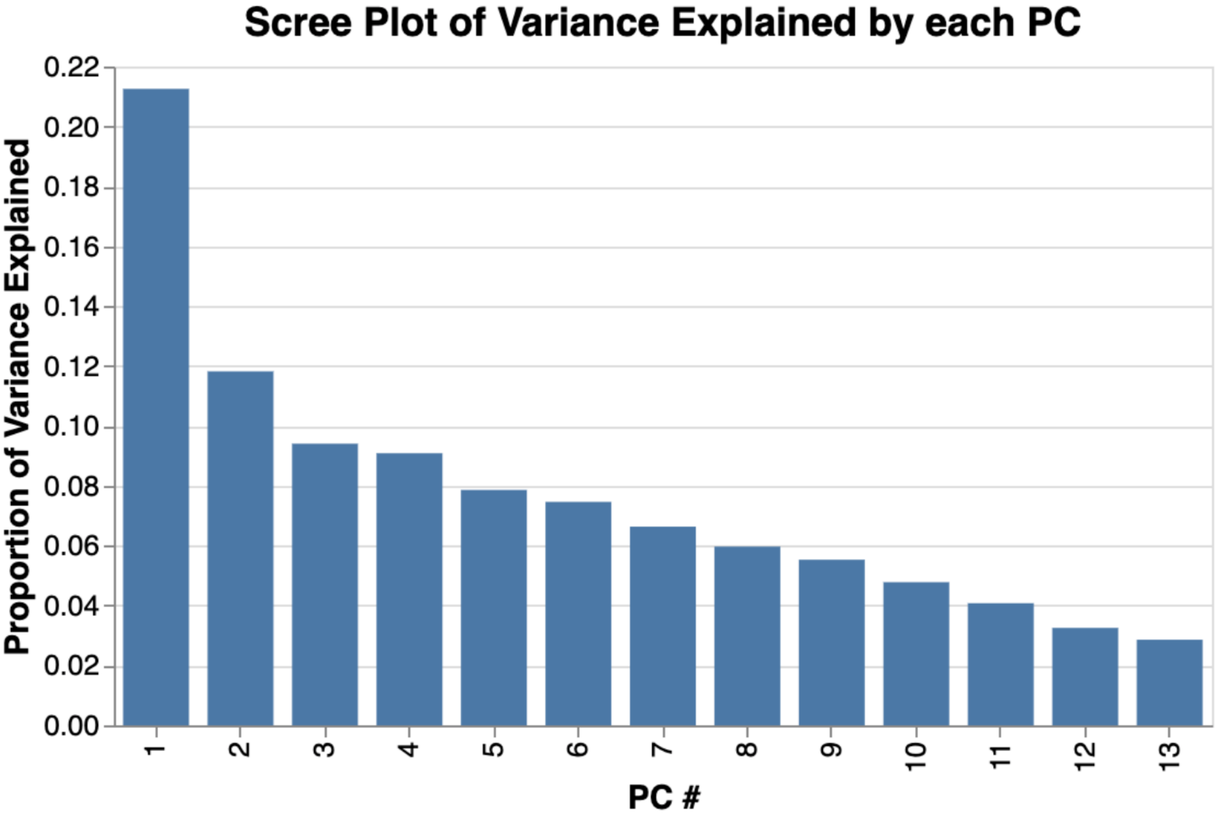

Figure 3: Variance explained by principal components. The first two PCs comprise of only 33.07% of the data.

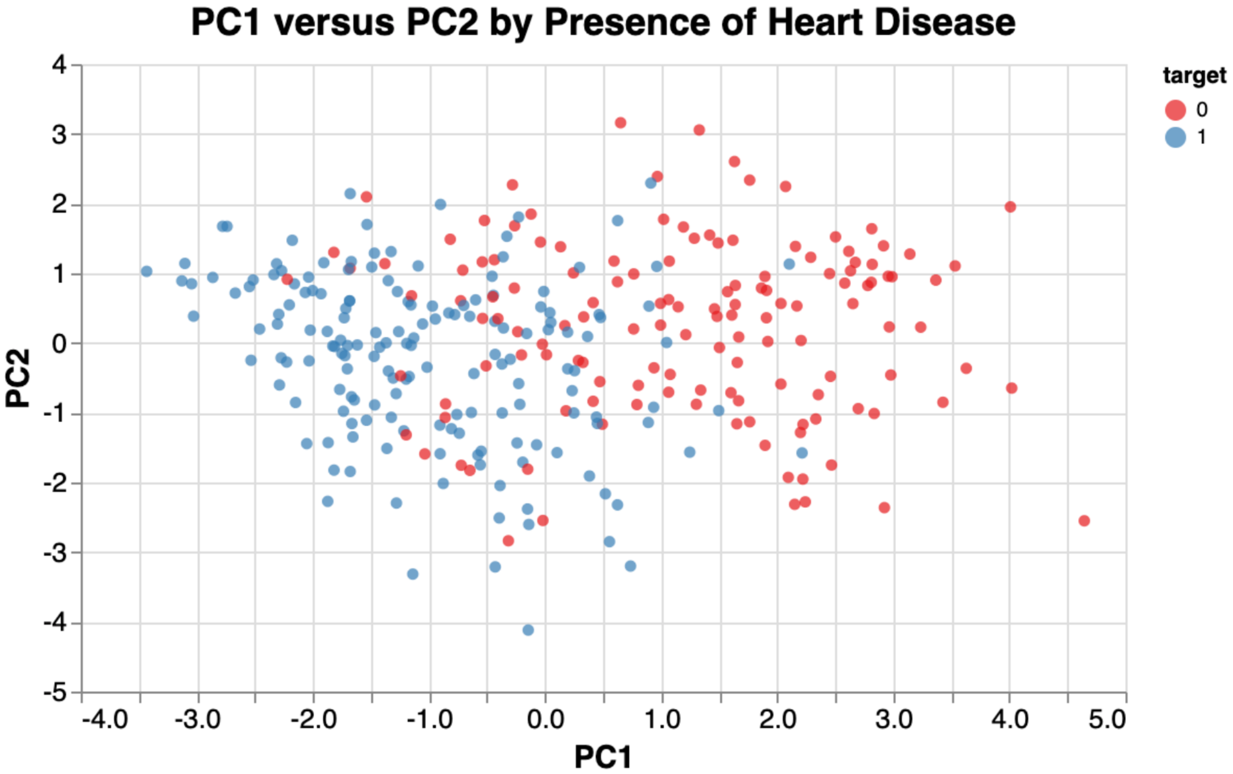

Figure 4: Scatter plot of PC1 versus PC2. Red indicates the presence of heart disease, and blue a lack thereof.

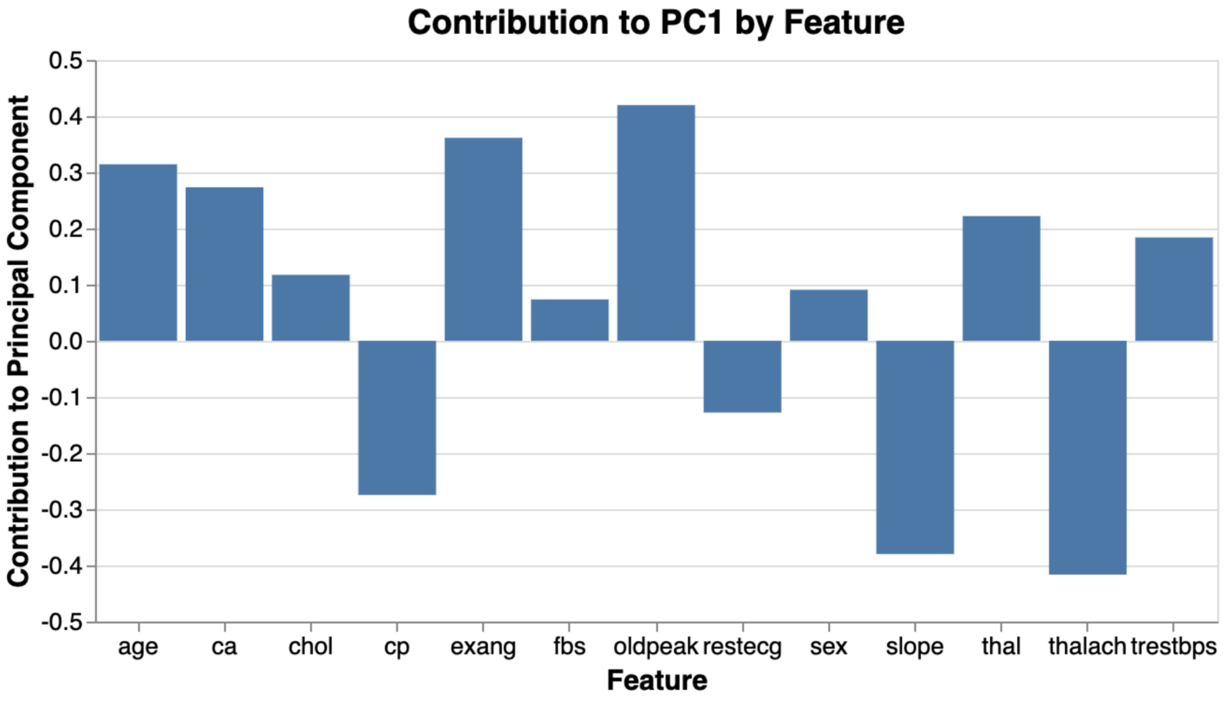

Figure 5: Weight of each feature in the first principal component (PC1).

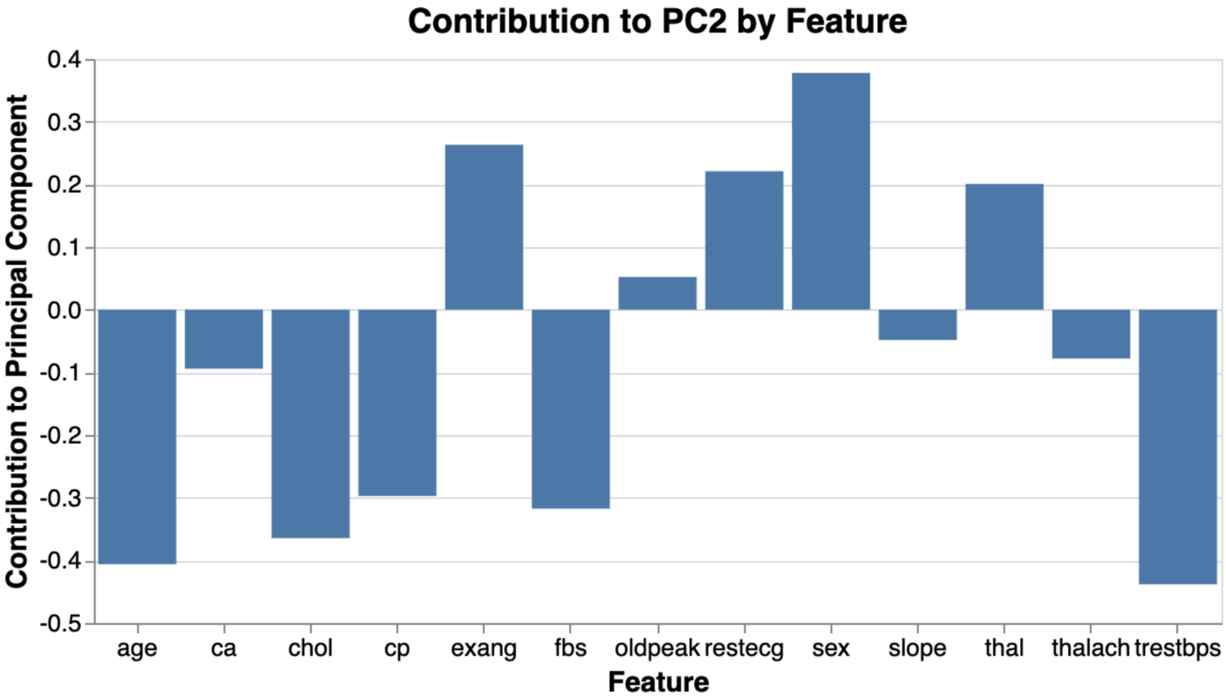

Figure 6: Weight of each feature in the second principal component (PC2).

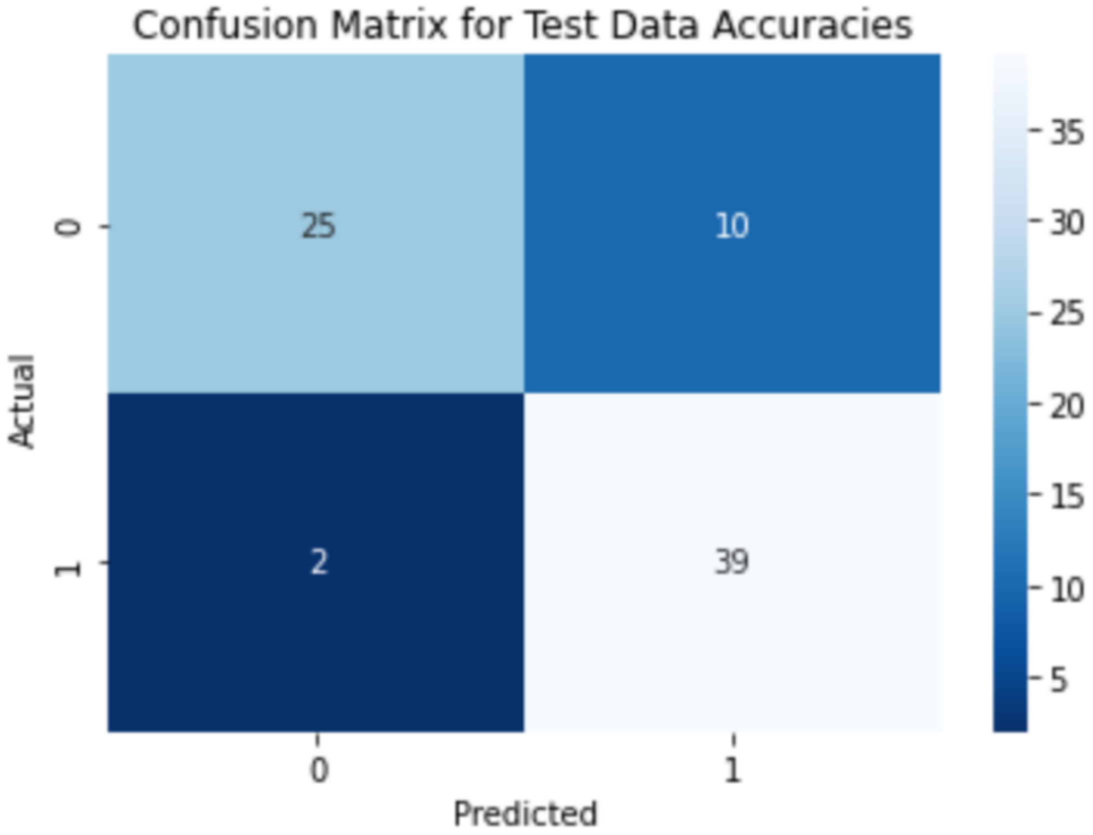

Figure 7: Confusion Matrix on the Test Data showing the distribution of the inaccuracies of the Logistic Model

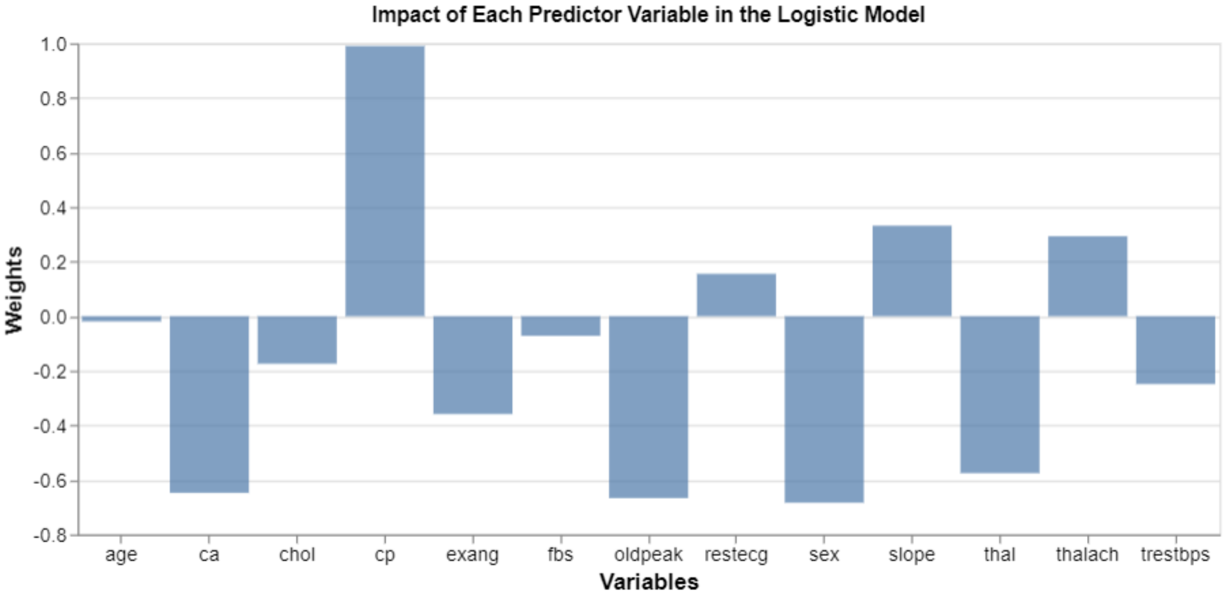

Figure 8: Weight of each feature in the final Logistic Model