Final Project: Coronary Heart Disease Analysis

S670 Exploratory Data Analysis - Fall 2022

Akhil Venu Gopal, Meenakshi Sundaram Ganapathy,

Sasidev Mahendran, Sumitha Vellinalur Thattai

Introduction:

Need for the Project:

According to the World Health Organization, heart disorders are thought to be the cause of 17.9 million deaths worldwide each year. About 50% of the mortality in the United States and other developed nations is due to cardiovascular diseases, primarily heart attacks. In the US, about 697,000 people died from heart disease in 2020 - that is 1 in every 5 deaths. Coronary heart disease (CHD) is the most common type of heart disease, killing approximately 382,820 people annually in the US. Heart disease cost the US about $229 billion from 2017 to 2018. This includes the cost of health care services, medicines, and lost productivity due to death. Early cardiovascular disease prognosis can help high-risk individuals make decisions about lifestyle adjustments that will lessen problems.

Goals:

The objective of the project is to understand the major factors that influence the 10-year risk of Coronary Heart Disease (CHD). Additionally, we will use these factors to predict the odds of CHD risk. We primarily focused on the below research question:

Can we predict the CHD odds by a simple logistic Regression model, or is there a requirement for a more complicated model?

Data Description:

The dataset, which is publicly accessible on the Kaggle, comes from ongoing cardiovascular research. Men and women between the ages of 32 and 70 from the town of Framingham, Massachusetts, who had not yet developed overt symptoms of cardiovascular disease or suffered a heart attack or stroke were selected for the study. This dataset is from one of the cohorts of the study and has records of ~4200 people and includes 16 attributes (9 categorical and 7 numerical attributes) - demographic, behavioral, and medical risk factors. These attributes will be used to predict the ten-year risk of CHD (TenYearCHD).

The variables in the dataset are as follows:

- **Male -** Binary values for Male and Female [0 - Female, 1 - Male]
- **Age -** Age of the person [years]
- **Education -** Education level of the person classified into 4 levels (1,2,3,4)
- **CurrentSmoker -** Indicates whether the person is a current smoker or not [0 - nonsmoker, 1 - smoker]

- **CigsPerDay -** Indicates how many cigarettes a person smokes per day
- **BPMeds -** Indicates if the person is taking BP Medications or not [0 - not taking meds, 1 - taking meds]
- **PrevalentStroke -** Indicates if the person had a prevalent stroke or not [0 - did not have a stroke, 1 - had a stroke]
- **PrevalentHyp -** Indicates if the person was hypertensive or not [0 - does not have hypertension, 1 - has hypertension]
- **Diabetes -** Indicates if the person has Diabetes or not [0 - does not have Diabetes, 1 - has Diabetes]
- **TotChol -** Total cholesterol level of the person [in mg/dL]
- **SysBP -** Systolic Blood Pressure of the person [in mmHg]
- **DiastolicBP -** Diastolic Blood Pressure of the person [in mmHg]
- **BMI -** BMI of the person [lb/m^2]
- **Heart Rate -** Heart Rate of the person [bpm]
- **Glucose -** Glucose level of the person [mmol/L]
- **TenYearCHD -** Indicates if the person has a 10-year risk of Coronary Heart Disease [0 - no risk of CHD, 1 - risk of CHD]

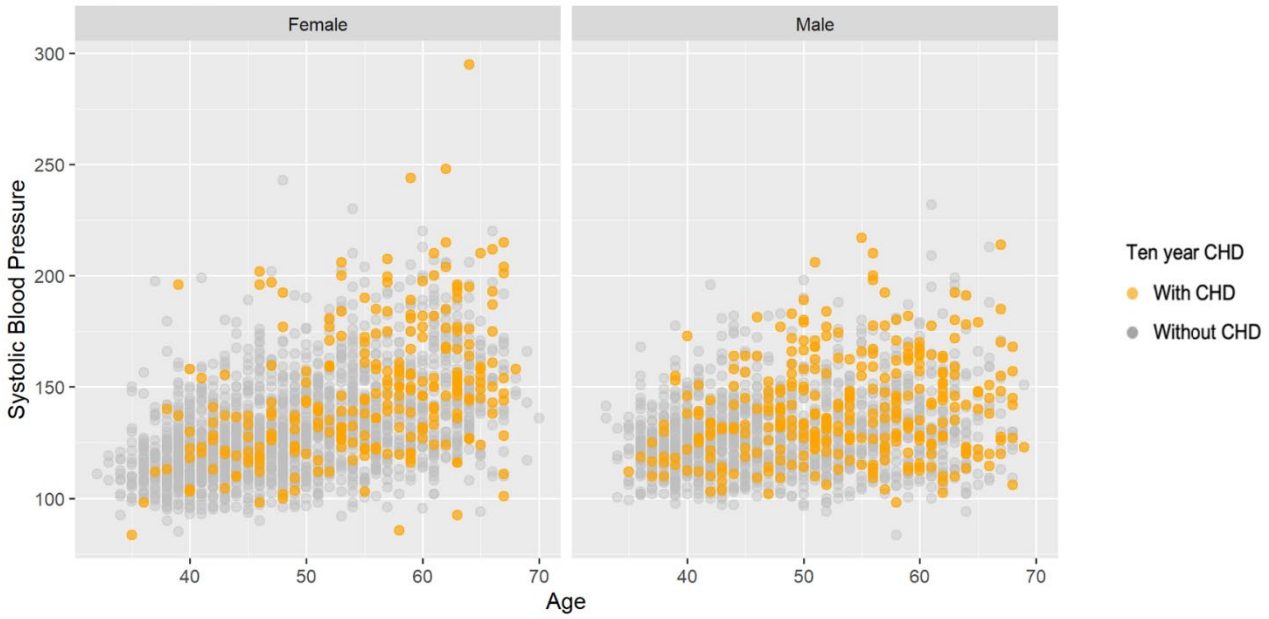
Data Cleaning and Preprocessing:

We started by cleaning and preprocessing the dataset - removing missing values and scaling the data. We saw that there are missing values in multiple columns - BMI, BPMeds, cigsperday, education, glucose level, heart rate, and total cholesterol. In total, there were 645 missing values in the dataset, and we excluded them from the analysis. Our final data set after removing null values contains 3656 records.

From figure 1, we can observe that age and systolic blood pressure are highly related to the ten-year risk of CHD. As age and systolic BP increase, the risk of CHD also increases. This trend holds for both men and women (from figure 2).

Figure 1: Age vs Systolic Blood Pressure

Figure 2: Systolic Blood Pressure vs Age vs CHD faceted by Gender



We start the analysis by looking at the summary statistics of the data. Below are the statistics of all the numerical variables in our dataset:

Variable	Mean	Std. Dev.	Min	Pctl. 25*	Pctl. 75*	Max
age	49.55	8.56	32	42	56	70
cigsPerDay	9.02	11.91	0	0	20	70
totChol	236.87	44.10	113	206	263.25	600
sysBP	132.37	22.09	83.5	117	144	295
diaBP	82.91	11.98	48	75	90	142.5
BMI	25.78	4.07	15.54	23.08	28.04	56.8
heartRate	75.736	11.98	44	68	82	143
glucose	81.86	23.91	40	71	87	394

* Percentile

From summary statistics, we observe that numerical variables are in different scales. For example, age ranges from 32 to 70, whereas systolic blood pressure ranges from 83 to 295. Hence, we scaled the variables to mean 0 and standard deviation 1 before building the model.

Modeling:

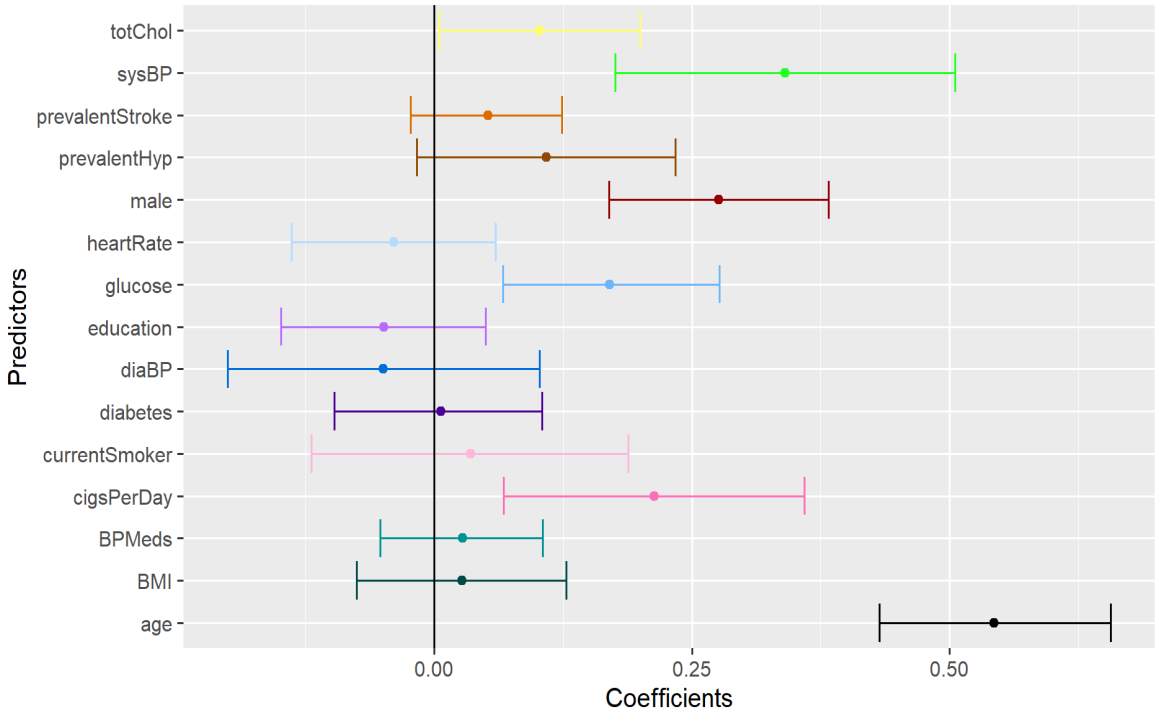
Baseline Model - Logistic Regression:

For our baseline model, we used logistic regression with all the features available in the dataset and looked at the confidence interval of each feature's coefficient estimates. The metric we have used to compare different models is AIC. Generally, the lower the AIC, the better the model. The AIC of the model with all the predictors included is 2786.2.

Model Formula:

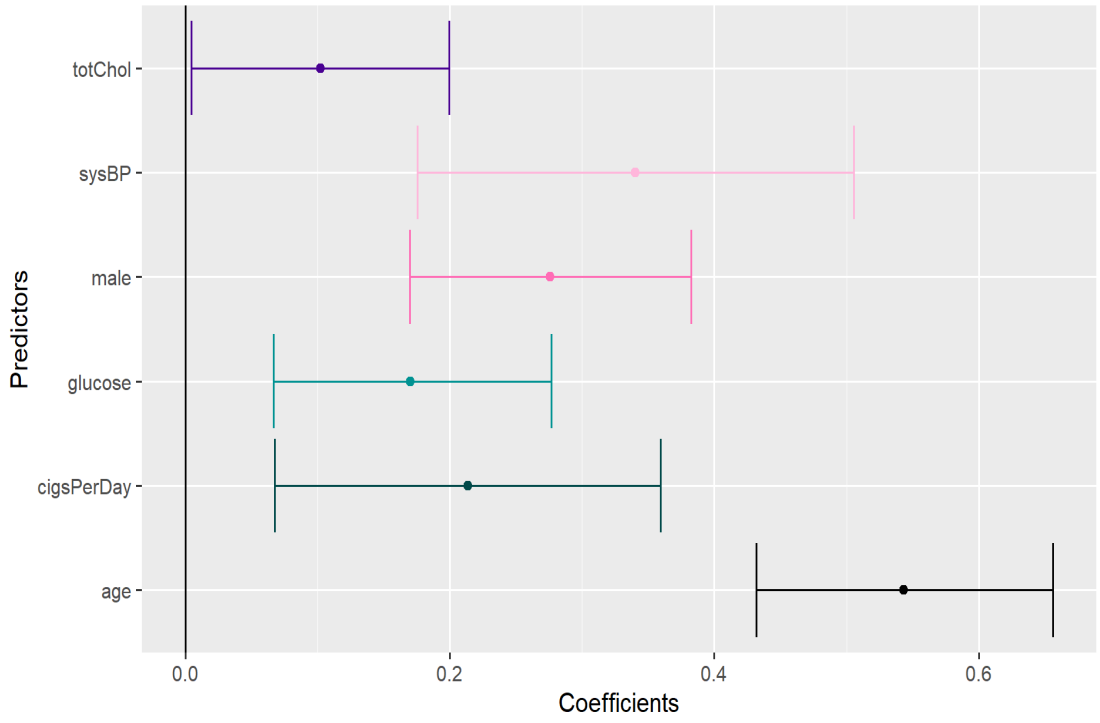
glm(formula = target ~ predictors, family = "binomial")

Figure 3: TIE Fighter Plot of Coefficients



From Figure 3, we can observe that some of the coefficients are very small compared to others. We should thus consider a model with fewer predictors. Below figure 4 focuses on the important features - Total Cholesterol, Systolic Blood Pressure, Gender, Glucose, Cigarettes per day, and Age. We will focus on these features to build the model.

Figure 4: TIE Fighter Plot for the Most Important Coefficients



Model Features	AIC Score
All features	2786.2
Total Chol, Sys BP, Gender, Glucose, Cigs per day, and Age	2776.5

When looking at the interaction terms (the effect of one variable on one or more variables), there was a strong relationship between age and systolic blood pressure. However, including them in the model increased the AIC score. Hence, the final model does not include the interaction term.

The below table shows the comparison of different models:

Model Features	AIC Score
Total Chol, Sys BP, Gender, Glucose, Cigs per day, and Age + Sys BP * Age + Sys BP * Cig per day	2779.4
Total Chol, Sys BP, Gender, Glucose, Cigs per day, and Age + Sys BP * Age	2777.9

Final Model:

When comparing the different model performances, the AIC score with predictors Total Cholesterol, Systolic Blood Pressure, Gender, Glucose, Cigarettes per day, and Age was the lowest. Hence, we will use this as our baseline model.

Model Formula:

glm(formula = df$TenYearCHD ~ age + cigsPerDay + sysBP + glucose + male + totChol , family = "binomial", data = predictors)

Model Coefficients:

Coefficients	Estimate	Std. Error
Intercept	-1.99	0.06
Age	**0.56**	**0.06**
CigsPerDay	0.23	0.05
SysBP	**0.39**	**0.04**
Glucose	0.17	0.04
Gender	**0.28**	**0.05**
TotChol	0.10	0.05

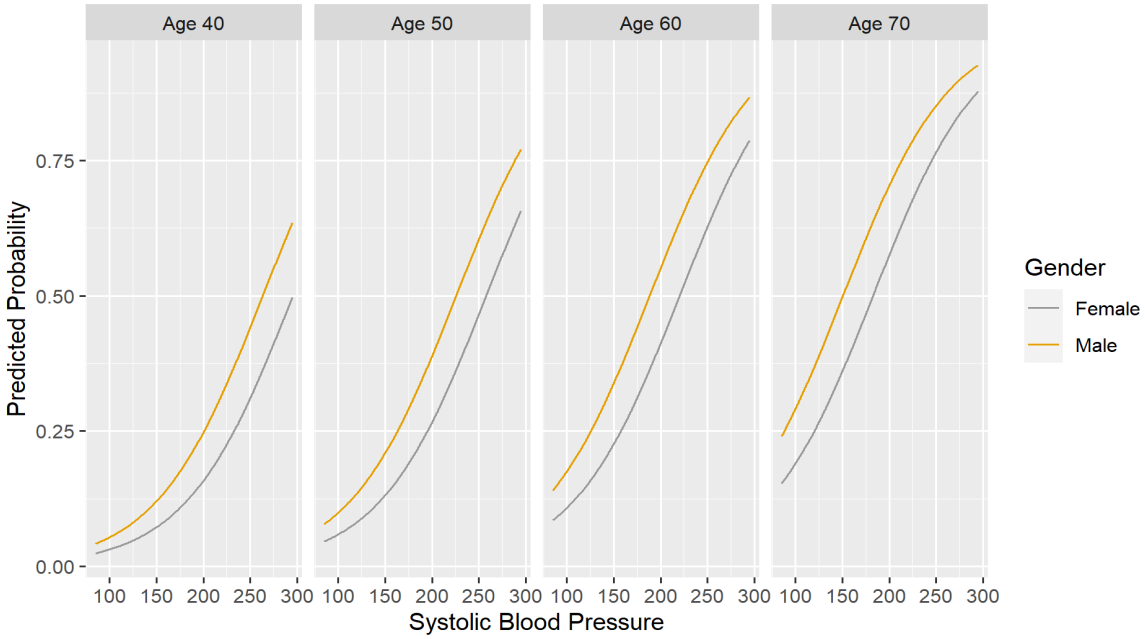
AIC of the above model - 2776.5

From the above model coefficients, we observe that age is the most important predictor followed by systolic blood pressure and gender.

Model Results:

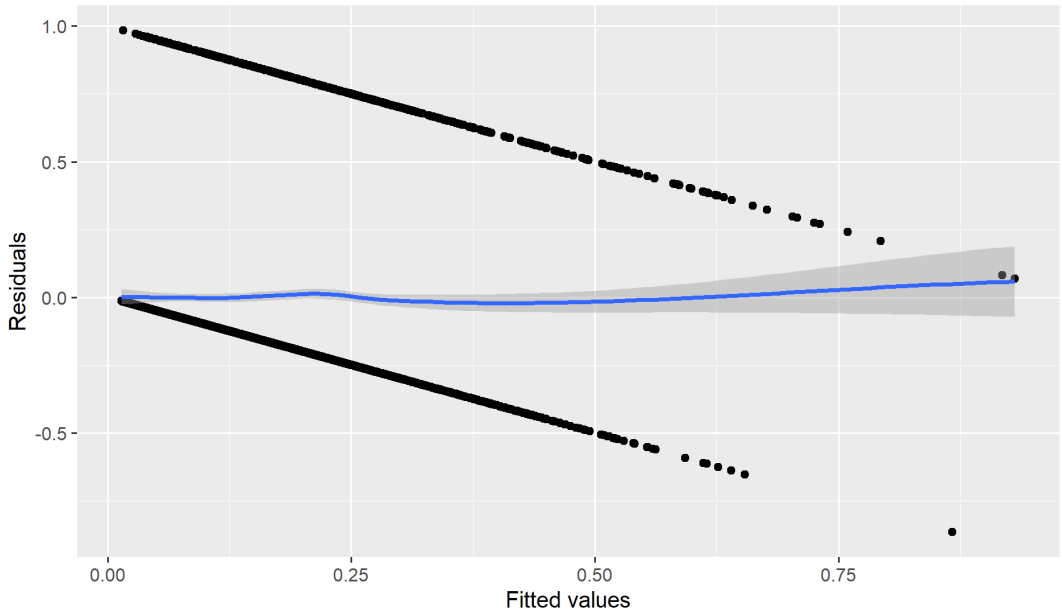
Figure 5 shows the prediction results for the test grid using the final model. The plot shows how the risk of ten-year CHD varies among men and women for different age groups and systolic blood pressure. The values for cigarettes per day and glucose were fixed at the median value. We can clearly observe that men are at a higher risk of ten-year CHD compared to women and the risk of CHD increases with an increase in age and systolic blood pressure.

Figure 5: Predicted Probabilities using Logistic Regression Model



When fitting a loess curve on the fitted vs residual value, it is almost a straight line at 0, indicating the model works well for the dataset. However, there is a small deviation at the right end of the plot, and we will build advanced models like GAM to see if there is any improvement in the model prediction.

Figure 6: Fitted vs Residual Plot for Baseline Model



Advanced Models:

a. GAM (Generalized Additive Model)

GAM is an adaptation of a Linear Model that allows us to model non-linear data while maintaining explainability. To improve the prediction of our model, we used GAM with the same set of features as our baseline model (Total Chol, Sys BP, Gender, Glucose, Cigs per day, and Age). Additionally, we tested the performance of the model by adding interaction terms. Furthermore, we added smoothing terms to the most important features - age and systolic blood pressure to check if the AIC scores reduced.

Below are the AIC scores of the different models with which we experimented:

Model Features	AIC Score
Total Chol, Sys BP, Gender, Glucose, Cigs per day, and Age	2776.5
Total Chol, Sys BP, Gender, Glucose, Cigs per day, and Age + Sys BP * Age	2777.9
Total Chol, Sys BP, Gender, Glucose, Cigs per day, and s(Age) + Sys BP * Age	2778.4
Total Chol, Gender, Glucose, Cigs per day, and s(Age, Sys BP)	2776.5

The above table clearly shows that the AIC scores of all the models are the same or greater than the AIC scores of the baseline logistic model. Below is the best GAM model with smoothing terms added for age and systolic blood pressure. The model has an AIC score of 2776.5, which is the same as the baseline logistic regression model.

Model Formula:

gam(formula = df$TenYearCHD ~ s(age, sysBP) + cigsPerDay + glucose + male + totChol, data = predictors, family = "binomial", method = 'REML')

Model Coefficients:

Coefficients	Estimate	Std. Error
Intercept	-1.99	0.06
CigsPerDay	0.23	0.05
Glucose	0.17	0.04
Gender	0.28	0.05
TotChol	0.10	0.05

Edf of smooth terms s(age, sysBP) - 2.01

AIC of the above model - 2776.5

Below is the prediction result using the GAM model. The prediction results are very similar to the baseline logistic regression model for all age groups.

Figure 7: Predicted Probabilities using GAM Model

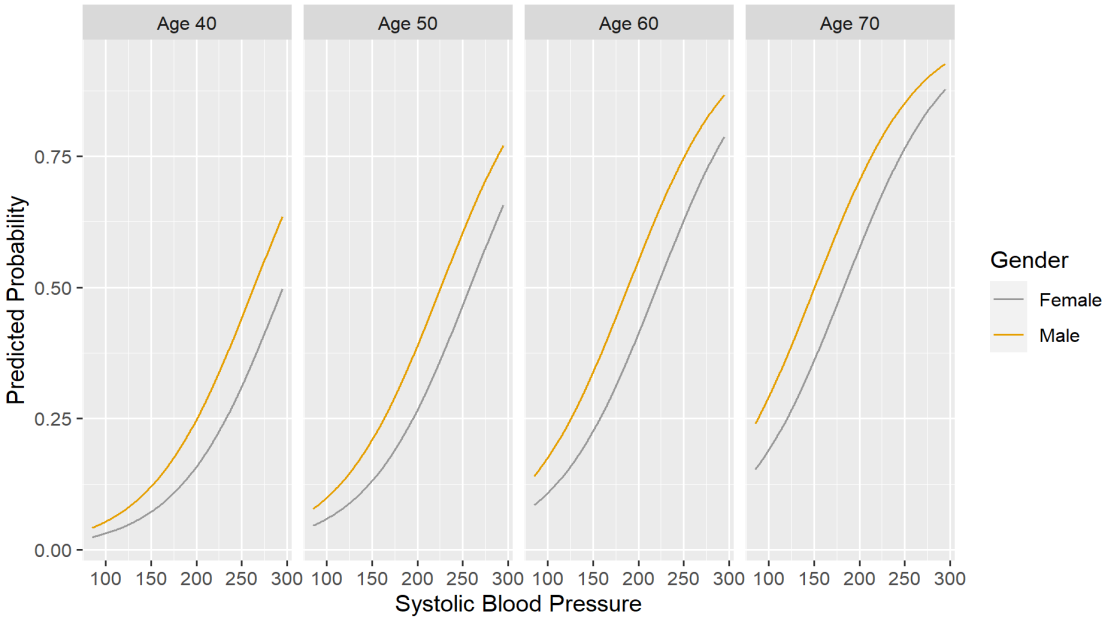
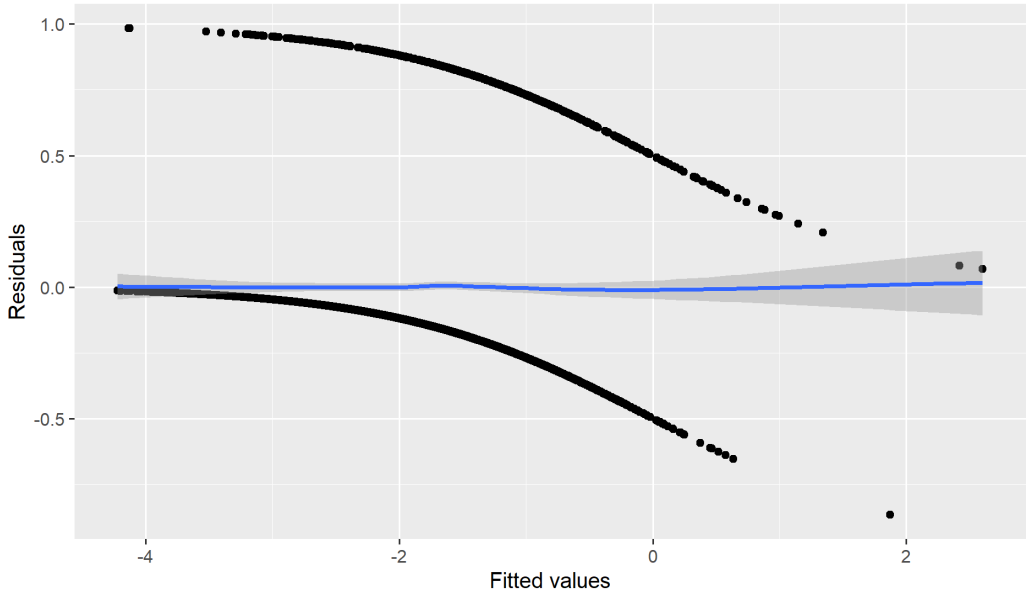


Figure 8 shows that the fitted vs residual plot is similar to the baseline logistic regression model. From the prediction results and the residual plot, we see that the performance of logistic regression and GAM are very similar. GAM produces similar results as a logistic model with additional complexity.

Figure 8: Fitted vs Residual Plot for GAM Model

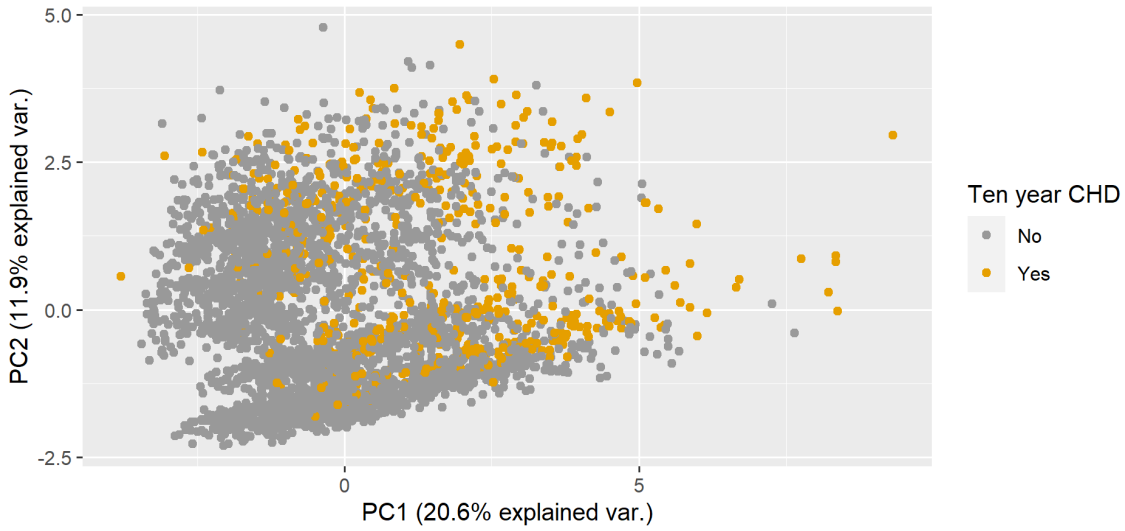


b. Principal Component Analysis (PCA)

The variables in the dataset are of different scales, hence we centered and scaled the data when performing PCA. The output of the PCA is a rotational matrix and principal components. We focused on the principal components to see if there is any clear pattern between the two groups of people.

The top two principal components explained about one-third of the variance. However, from the plot below with two principal components, we observe that there is no clear separation between people with and without a ten-year risk of CHD.

Figure 9: PCA of 2 Important Components (explains 32.5% variance)



Conclusion:

Through this project, we have gained insights into the various factors that potentially have effects on the ten-year risk of Coronary Heart Disease (CHD). We did variable selection using the TIE fighter plot and as per our analysis, the three major factors which influence the risk of CHD are the age of a person, systolic blood pressure, and gender. We observed that men are at a higher risk of ten-year CHD compared to women and the risk of CHD increases with an increase in age and systolic blood pressure. We experimented with a simple logistic regression model, GAM, PCA, and Lasso models. Using PCA was not very useful as there was no clear pattern observed. Also, the Lasso model could not be compared with the rest of the regression models as calculating AIC and residuals was not very straightforward. Comparing the logistic regression and GAM models, the AIC score, residual plot, and prediction results were very similar. Also, the logistic regression model did not have any smoothing terms and it was much more interpretable. Thus, we conclude that a simple logistic regression model is enough to predict the ten-year risk of CHD rather than any advanced models.

Limitations:

Some of the limitations that we faced in this project are as stated below:

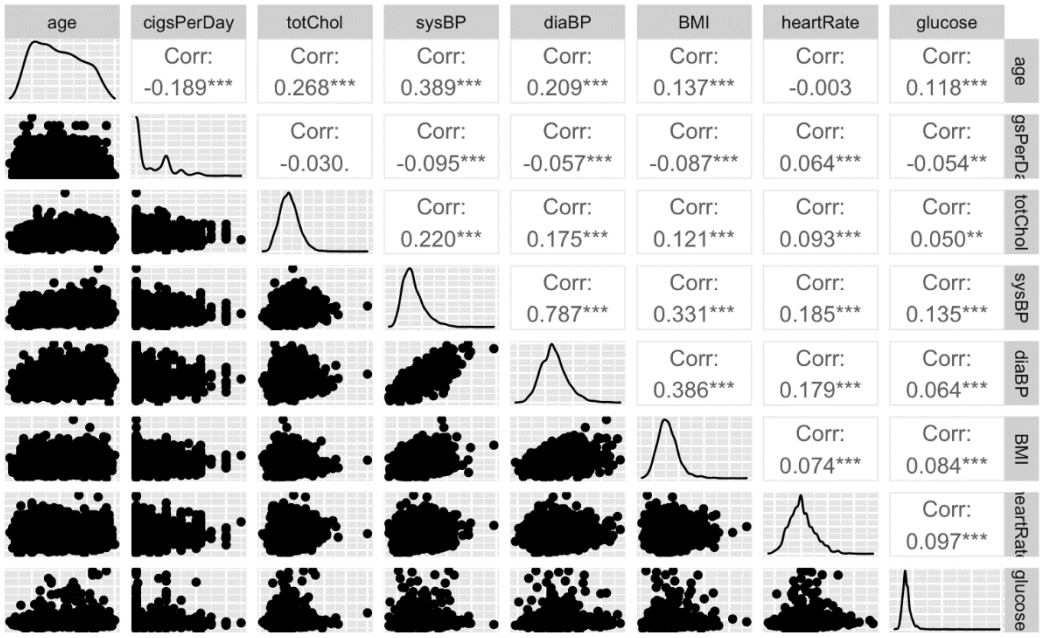
- The dataset had many features impacting the target variable (TenYearCHD**)**, so we had to experiment a lot to find the best combination of features for our model. Though the TIE fighter plot is one way to get important variables, there are other methods like regsubsets that can get us better combination results
- We were not able to find the AIC score for the Lasso model, so we could not compare the model's performance with other models
- 10% of the Glucose variable was null in the dataset. Since glucose is one of the top six variables in predicting the ten-year risk of CHD as per our analysis from the Tie Fighter plot, non-null values in this column can help us improve the model performance

Future Works:

- The Framingham dataset which we used is a subset of the ongoing cardiovascular research, so we would like to use a bigger dataset with a greater number of features to predict the CHD odds
- Only 15% of people in the dataset have a ten-year risk of CHD indicating a class imbalance. We plan to add more data to improve the class imbalance.
- We also plan to experiment with advanced models like bagging and boosting that can improve the prediction results

Appendix:

a. Pair Plot



b. Lasso Model

We used the lasso model to find the list of important features that can be used to improve the model's performance.

```
Call:  glmnet(x = predictors, y = df$TenYearCHD, family = "binomial",      alpha = 1, lambda
= cv.lasso$lambda.1se)

  Df %Dev  Lambda
1  6 9.31 0.02284
```

The left-side image shows the lasso model with the maximum lambda value and the right-side image shows the model with the minimum lambda value. From the minimum lambda value model, we see that 9 of the feature coefficients were reduced to zero. 5 of the 6 non-zero coefficients from the lasso model were used in our baseline model to make predictions.

```
16 x 1 sparse Matrix of class "dgCMatrix"
                           s1
(Intercept)     -1.9572239104
male             0.2472370790
age              0.5258224908
education       -0.0274905185
currentSmoker    0.0027817434
cigsPerDay       0.2032567769
BPMeds           0.0191648544
prevalentStroke  0.0392941402
prevalentHyp     0.0905834019
diabetes         0.0016313000
totChol          0.0739319556
sysBP            0.2993532786
diaBP            .
BMI              0.0005047691
heartRate       -0.0021609796
glucose          0.1560594447
```

```
16 x 1 sparse Matrix of class "dgCMatrix"
                           s1
(Intercept)     -1.82124053
male             0.10362553
age              0.37381992
education        .
currentSmoker    .
cigsPerDay       0.04292364
BPMeds           .
prevalentStroke  .
prevalentHyp     0.01717346
diabetes         .
totChol          .
sysBP            0.25700652
diaBP            .
BMI              .
heartRate        .
glucose          0.06199789
```