Statistical Modeling for Heart Disease Risk Prediction

Siddhant Joshi
Cognitive Science
A16878588

Jordan Nishi
Cognitive Science
A16201086

Abstract

This project aims to develop a deeper understanding of signs of heart disease. Despite being America's leading cause of death, taking 695,000 lives in 2021, according to the Center for Disease Control (CDC), a consensus on the specific predictors for this disease are an ongoing area of study for the research community. Our data were collected by the CDC on 253,680 individual's physiological information, daily activities, dietary intake, and presence of heart disease. To determine the top five contributors towards heart disease, Multiple Logistic Regression (MLR) and Principal Component Logistic Regression (PCLR), used 5-fold cross validation, and tuned their decision boundaries. The best model performed at an accuracy of 77.14%, reporting the top predictors as high blood pressure, general health, age, sex, and high cholesterol. Our findings will aid the scientific and medical field in identifying the onset of heart disease and initiating the necessary steps to prevent further development.

1 Introduction

Motivation

Heart disease is an umbrella term that describes any physiological issues that interfere with the cardiovascular system. These include stroke, coronary artery disease, congenital heart disease, and heart failure. According to the Center for Disease Control (CDC), heart diseases are the leading cause of death for people in the United States. Despite its prevalence, a consensus on the specific contributors to the development of this class of diseases is an ongoing endeavor of the scientific community. Our project attempts tackles this question by considering many factors, ranging from lifestyle choices, physiological well-being, and medical history, to extract reliable features with which to build an effective predictive model for classifying and quantifying a patient's risk for heart disease.

Dataset

The data we use for this project were collected from the CDC's (Center for Disease Control) 2015 Behavioral Risk Factor Surveillance System (BRFSS), which surveys Americans on health-related risk factors, use of preventative services, and history of health conditions.

Our data contains 21 varied predictors on patients' biological information such as their blood pressure, daily activities and dietary intake, as well as presence of medical complications such as diabetes and stroke. These predictors are either gathered directly from question responses or calculated based on their responses. In total, there are 253,680 participants (observations), with 23,893 having heart disease.

Due to the imbalance of individuals with and without heart disease in this dataset, we will balance the data when preparing it for model training and testing.

Predictor	Type	Value	Description
HighBP	Categorical	0, 1	1 = Has high blood pressure, 0 = does not
HighChol	Categorical	0, 1	1 = Has high cholesterol, 0 = does not
CholCheck	Categorical	0, 1	1 = Had cholesterol checked in the past 5 years, 0 = did not
BMI	Continuous	12 - 98	Patient BMI
Smoker	Categorical	0, 1	1 = Is a smoker, 0 = not a smoker
Stroke	Categorical	0, 1	1 = Has suffered a stroke, 0 = Has not suffered a stroke
Diabetes	Categorical	0, 1, 2	0 = No diabetes, 1 = Developing diabetes, 2 = Has diabetes
PhysActivity	Categorical	0, 1	Did physical activity in past month
Fruits	Categorical	0, 1	1 = Ate fruits within the past month, 0 = did not
Veggies	Categorical	0, 1	1 = Ate vegetables within the past month, 0 = did not
HvyAlcoholConsump	Categorical	0, 1	1 = Heavy drinker, 0 = does not drink
AnyHealthcare	Categorical	0, 1	1 = Has access, 0 = No access
NoDocbcCost	Categorical	0, 1	1 = Can see doc, 0 = cannot see doc
GenHlth	Categorical	1 - 5	General health ranking 1 = best, 5 = worst
MentHlth	Continuous	0 - 30	Number of days with poor mental health
PhysHlth	Continuous	0 - 30	Number of days with poor physical health
DiffWalk	Categorical	0, 1	1 = Patient has difficulty walking, 0 = does not have difficulty
Sex	Categorical	0, 1	1 = Male, 0 = Female
Age	Categorical	1 - 13	Patient age (higher rank = older patient)
Education	Categorical	1 - 6	Education level (higher rank = higher level of education)
Income	Categorical	1 - 8	Patient income (higher rank = more income)
HeartDiseaseorAttack	Categorical	1, 0	1 = Has heart disease, 0 = healthy

Figure 1: Description of predictors and target variable.

Hypotheses

For this project, we'd like to investigate the following four hypotheses:

1) The PCLR model will perform better than the bonehead model as it will frame predictions around highly-variable predictors, thereby "learning" relationships between afflicted and healthy patients compared to simply assigning the same outcome value to every input.

2) The logistic regression model will perform better than the bonehead model since forward subset selection and decision boundary tuning would ensure the model "learns" relationships between afflicted and healthy patients compared to simply assigning the same outcome value to every input.

3) We expect the following predictors to be the top contributors when driving predictions. In logistic regression, these predictors would be heavily weighted during prediction and in PCLR regression, the first couple PCs will be most influenced by these predictors. The predictors and their selection criteria are listed below:

- Age → Historically, elderly patients are more susceptible to due to aging and weakened bodily systems.
- Smoking → It is well-known that smoking compromises the cardiovascular and respiratory systems, prompting us to believe smokers would have weaker hearts and therefore increased chances of heart disease.
- High Cholesterol → The build-up of cholesterol hampers or blocks blood flow, putting more strain on the heart when it is pumping blood.
- High BP → Similar to high cholesterol, increases in blood pressure can put strain on the heart, increases chances for strokes and weakened cardiovascular processes.
- Diabetes → The long-term heart/blood vessel damage caused by high levels of glucose could pose a significant threat to heart health.

4) Furthermore, we hypothesize that these top five predictors will be quite similar between the two models since the predictors we hypothesized are based on biological research in the field, which should not deviate based on the model used to make the predictions.

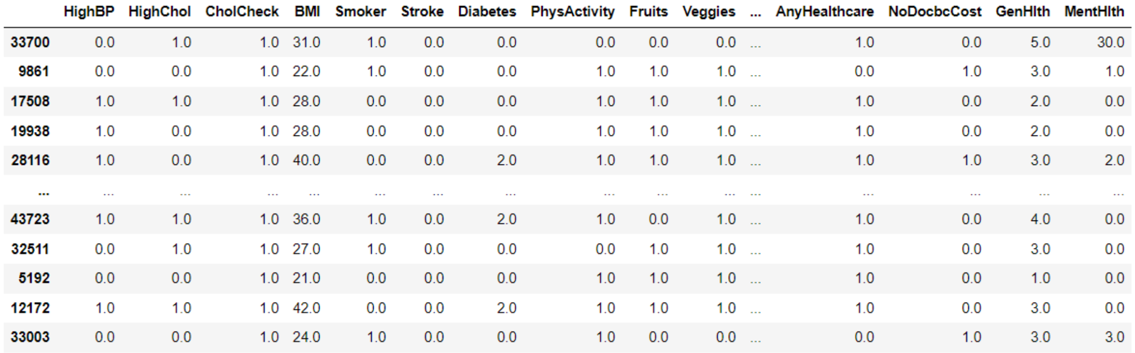
2 Method

Data Preparation

With only 9.41% of the sample having a history of heart disease, we will be performing a 50-50 split on these two samples, using 23,893 responses indicating no history of heart disease and all 23,893 responses with a history for a total of 47,786 observations. The 23,893 responses with no history of heart disease will be selected at random so as to remain unbiased across all predictors.

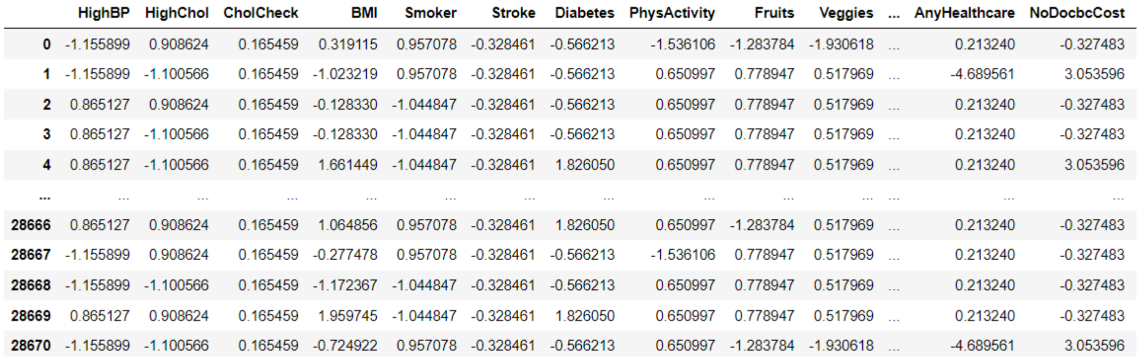
While truncating the data this way compromises the statistical power of our predictions, correcting for the over-representation of a single group would allow fitted models to better learn the difference between heart disease-afflicted patients and healthy ones.

Before using the data to train and test statistical models, we scaled it using the z-scoring methods provided by the sklearn Python library. This way, we balanced the impact of each predictor on the target variable.



	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	Veggies	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth
33700	0.0	1.0	1.0	31.0	1.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	5.0	30.0
9861	0.0	0.0	1.0	22.0	1.0	0.0	0.0	1.0	1.0	1.0	...	0.0	1.0	3.0	1.0
17508	1.0	1.0	1.0	28.0	0.0	0.0	0.0	1.0	1.0	1.0	...	1.0	0.0	2.0	0.0
19938	1.0	0.0	1.0	28.0	0.0	0.0	0.0	1.0	1.0	1.0	...	1.0	0.0	2.0	0.0
28116	1.0	0.0	1.0	40.0	0.0	0.0	2.0	1.0	1.0	1.0	...	1.0	1.0	3.0	2.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
43723	1.0	1.0	1.0	36.0	1.0	0.0	2.0	1.0	0.0	1.0	...	1.0	0.0	4.0	0.0
32511	0.0	1.0	1.0	27.0	1.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	3.0	0.0
5192	0.0	0.0	1.0	21.0	0.0	0.0	0.0	1.0	1.0	1.0	...	1.0	0.0	1.0	0.0
12172	1.0	1.0	1.0	42.0	0.0	0.0	2.0	1.0	1.0	1.0	...	1.0	0.0	3.0	0.0
33003	0.0	0.0	1.0	24.0	1.0	0.0	0.0	1.0	0.0	0.0	...	0.0	1.0	3.0	3.0

Figure 2: Pre-standardized data



	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	Veggies	...	AnyHealthcare	NoDocbcCost
0	-1.155899	0.908624	0.165459	0.319115	0.957078	-0.328461	-0.566213	-1.536106	-1.283784	-1.930618	...	0.213240	-0.327483
1	-1.155899	-1.100566	0.165459	-1.023219	0.957078	-0.328461	-0.566213	0.650997	0.778947	0.517969	...	-4.689561	3.053596
2	0.865127	0.908624	0.165459	-0.128330	-1.044847	-0.328461	-0.566213	0.650997	0.778947	0.517969	...	0.213240	-0.327483
3	0.865127	-1.100566	0.165459	-0.128330	-1.044847	-0.328461	-0.566213	0.650997	0.778947	0.517969	...	0.213240	-0.327483
4	0.865127	-1.100566	0.165459	1.661449	-1.044847	-0.328461	1.826050	0.650997	0.778947	0.517969	...	0.213240	3.053596
...	...	...	...	...	...	...	...	...	...	...	...	...	...
28666	0.865127	0.908624	0.165459	1.064856	0.957078	-0.328461	1.826050	0.650997	-1.283784	0.517969	...	0.213240	-0.327483
28667	-1.155899	0.908624	0.165459	-0.277478	0.957078	-0.328461	-0.566213	-1.536106	0.778947	0.517969	...	0.213240	-0.327483
28668	-1.155899	-1.100566	0.165459	-1.172367	-1.044847	-0.328461	-0.566213	0.650997	0.778947	0.517969	...	0.213240	-0.327483
28669	0.865127	0.908624	0.165459	1.959745	-1.044847	-0.328461	1.826050	0.650997	0.778947	0.517969	...	0.213240	-0.327483
28670	-1.155899	-1.100566	0.165459	-0.724922	0.957078	-0.328461	-0.566213	0.650997	-1.283784	-1.930618	...	-4.689561	3.053596

Figure 3: Standardized data

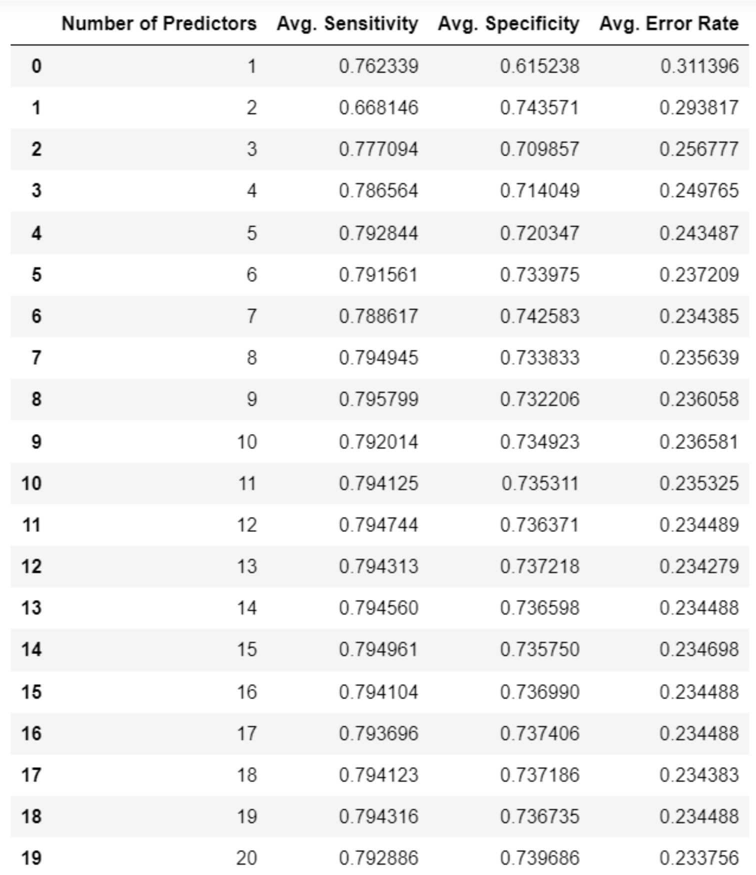
General Method

We used the forward subset selection algorithm to for model selection, devising the best combination of predictors. Then, we verified our findings with 5-fold cross-validations and calibrated the final model's decision boundary using Receiver Operating Characteristic (ROC) analysis to determine the most effective one.

We used the NumPy, Pandas, Statsmodels, Matplotlib, Seaborn, and Scikit-Learn libraries for Python to build two classification models that would generate predictions about whether or not a given individual has or is at risk for developing heart disease:

Model 1 - Multiple Logistic Regression

The first model will make use of a multiple logistic regression algorithm. The data will be shuffled and split into training, validation, and testing sets such that 60% of observations are used for training, and 20% each are used for validation and testing. We used forward subset selection with a Bayes' decision threshold (50%) to generate 21 models with the lowest error rate scores on the training set for each class of model. These models will then be **cross-validated** using 5-fold cross-validation using the validation set. Based on previous research, 5 folds were determined to offer a reliable trade-off between computational cost and model performance. The error rate scores will be averaged across each fold for each model and the model with the lowest test average on the validation set will be the optimal model.



	Number of Predictors	Avg. Sensitivity	Avg. Specificity	Avg. Error Rate
0	1	0.762339	0.615238	0.311396
1	2	0.668146	0.743571	0.293817
2	3	0.777094	0.709857	0.256777
3	4	0.786564	0.714049	0.249765
4	5	0.792844	0.720347	0.243487
5	6	0.791561	0.733975	0.237209
6	7	0.788617	0.742583	0.234385
7	8	0.794945	0.733833	0.235639
8	9	0.795799	0.732206	0.236058
9	10	0.792014	0.734923	0.236581
10	11	0.794125	0.735311	0.235325
11	12	0.794744	0.736371	0.234489
12	13	0.794313	0.737218	0.234279
13	14	0.794560	0.736598	0.234488
14	15	0.794961	0.735750	0.234698
15	16	0.794104	0.736990	0.234488
16	17	0.793696	0.737406	0.234488
17	18	0.794123	0.737186	0.234383
18	19	0.794316	0.736735	0.234488
19	20	0.792886	0.739686	0.233756

Figure 4: Performance metrics over set of best logistic regression models for each class during 5-fold cross validation. Error rates, sensitivities, and specificities have all been averaged.

Lastly, we tuned the decision threshold based on the ROC curve of the best model (Fig. 5). The decision threshold at which the accuracy is the highest will be selected. Since our data are already balanced, seeking to minimize the ratio of false positive results to true positive results can be given less priority compared to overall accuracy. To assess final performance, we will train the final model on the combined training and validation sets before generating performance metrics on the testing set (Fig. 6).

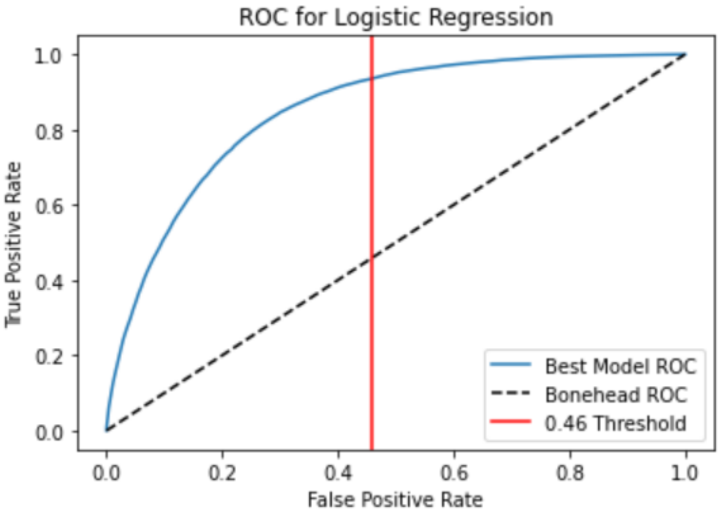
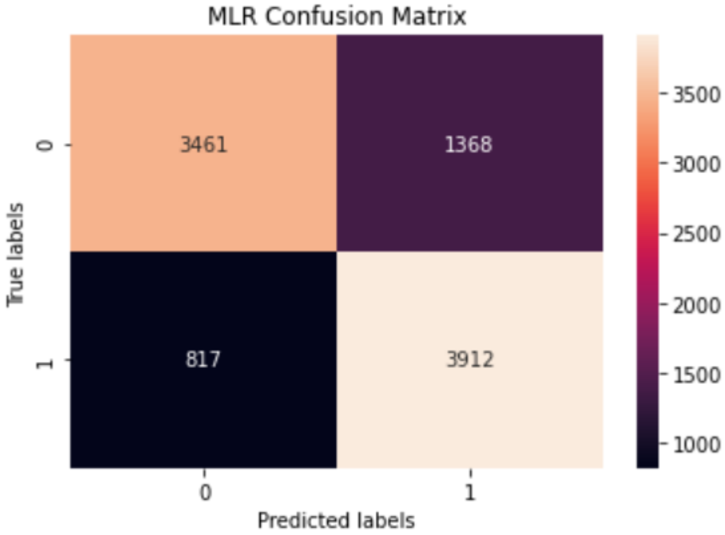


Figure 5: ROC Curve for Multiple Logistic Regression. Best threshold is 0.46

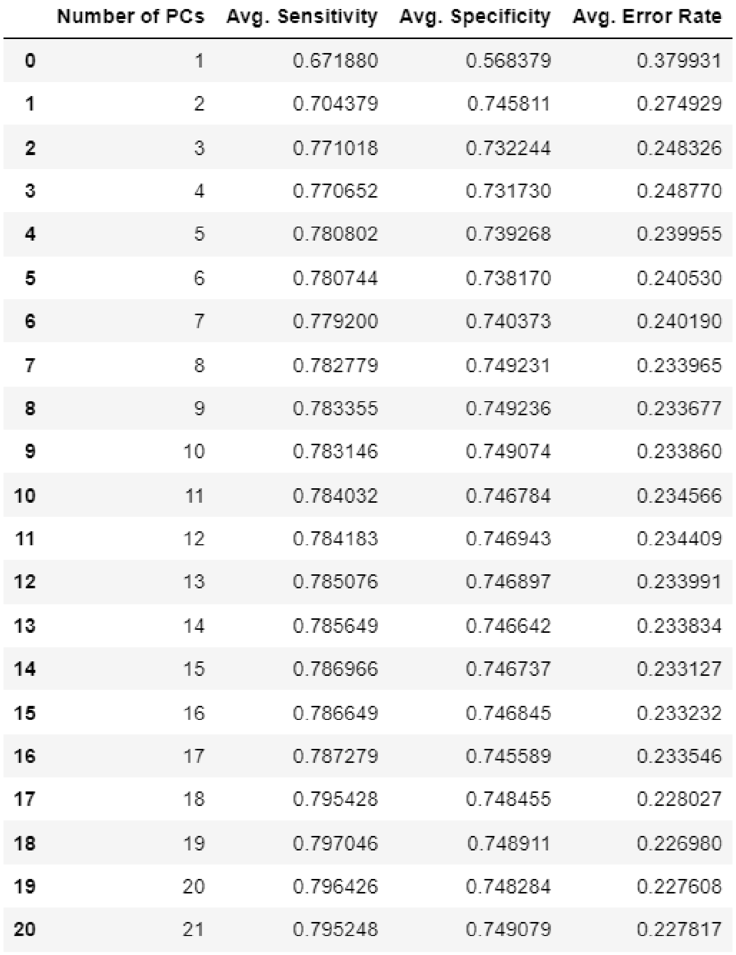


```
Specificity: 0.7167115344791882
Sensitivity: 0.8272362021569042
Accuracy: 0.7713956894747855
```

Figure 6: Multiple Logistic Regression Confusion Matrix

Model 2 - Principal Component Logistic Regression

The second model utilizes a principal component logistic regression (PCLR) algorithm described as follows. Using sklearn's train_test_split() method, the data will be shuffled and split into validation and testing sets such that 80% of observations are used for validation, and the remaining 20% are used for testing. We will convert the training set into a $n \times p$ matrix where p is the number of predictors and n is the number of observations. A singular value decomposition (SVD) will be run on this matrix to generate U, S, and V^T matrices. When interpreting the SVD result, the columns of U contain the principal components (21 PCs) with each value indicating the contribution of each observation to that principal component. The columns of U will be used in the model selection process, serving as the inputs that will be weighted by logistic regression parameters later on. These 21 models will be **cross-validated** using 5-fold cross-validation on validation set. The error rates of the test and train subsets of the validation set will be averaged across each class of model, with the model resulting in the lowest average being the most optimal.



	Number of PCs	Avg. Sensitivity	Avg. Specificity	Avg. Error Rate
0	1	0.671880	0.568379	0.379931
1	2	0.704379	0.745811	0.274929
2	3	0.771018	0.732244	0.248326
3	4	0.770652	0.731730	0.248770
4	5	0.780802	0.739268	0.239955
5	6	0.780744	0.738170	0.240530
6	7	0.779200	0.740373	0.240190
7	8	0.782779	0.749231	0.233965
8	9	0.783355	0.749236	0.233677
9	10	0.783146	0.749074	0.233860
10	11	0.784032	0.746784	0.234566
11	12	0.784183	0.746943	0.234409
12	13	0.785076	0.746897	0.233991
13	14	0.785649	0.746642	0.233834
14	15	0.786966	0.746737	0.233127
15	16	0.786649	0.746845	0.233232
16	17	0.787279	0.745589	0.233546
17	18	0.795428	0.748455	0.228027
18	19	0.797046	0.748911	0.226980
19	20	0.796426	0.748284	0.227608
20	21	0.795248	0.749079	0.227817

Figure 7: Performance metrics over set of best PCLR models for each class during 5-fold cross validation. Error rates, sensitivities, and specificities have all been averaged.

Again, we tuned the decision threshold based on the ROC curve of the best model (Fig. 8). The decision threshold at which the accuracy is the highest will be selected. Not only will the subset of PCs that generate the lowest testing error rate on the validation data be the best model, but since we are using principal components, this is effectively a kind of regularized procedure. Each successive PC contains increasing amounts of variance, thereby limiting the amount of noise introduced with each successive model class. Lastly, to assess final performance, we will train the final model on the validation sets before generating performance metrics on the testing set (Fig. 9).

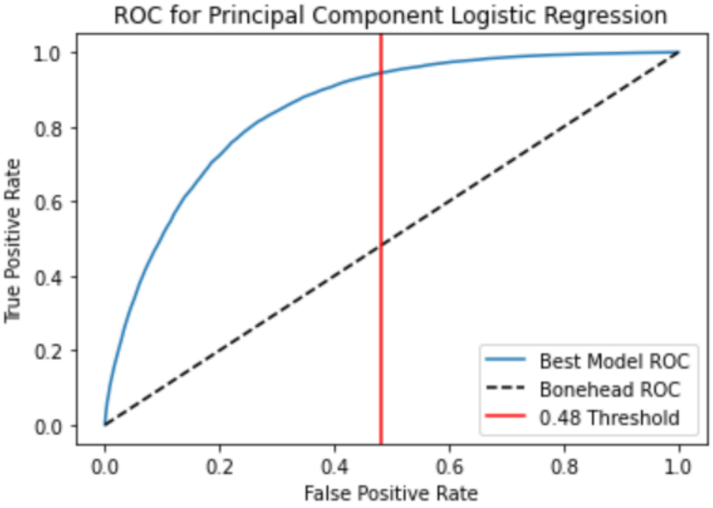
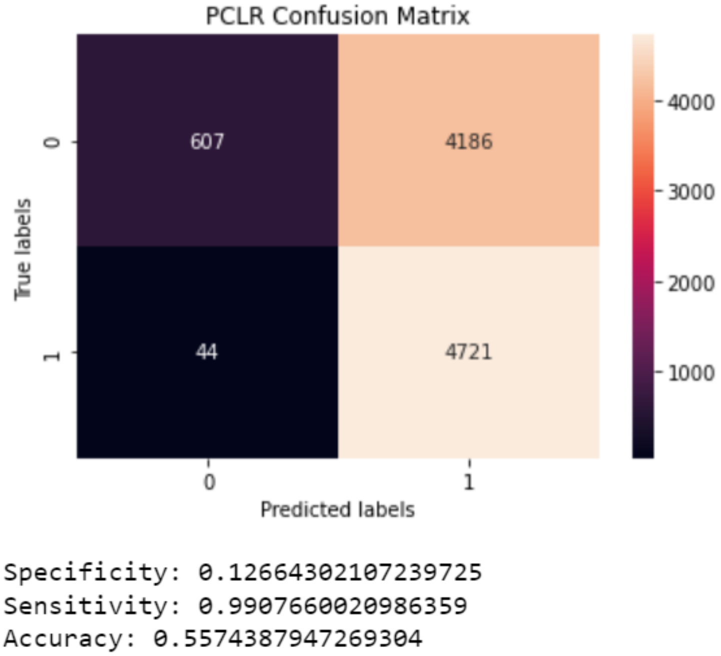


Figure 8: ROC Curve for Principal Component Logistic Regression. Best threshold is 0.48



```
Specificity: 0.12664302107239725
Sensitivity: 0.9907660020986359
Accuracy: 0.5574387947269304
```

Figure 9: Principal Component Regression Confusion Matrix

3 Results

Final Model Selection

After thoroughly comparing testing data of our Multiple Logistic Regression (MLR) and Logistic Regression (PCLR) models, we've concluded that our MLR more accurately predicts if an individual has heart disease or is at risk of it. Our MLR model has a testing accuracy rate of 77.14%, while our PCLR reached an accuracy score of of 55.74%.

Both models are quite complex, as the PCLR utilizes 19 out of 21 available PCs and the MLR utilizes all 21 predictors to make predictions. As a result, both models are aso quite flexible, since their consideration of majority of the predictors grants multiple degrees of freedom for decision-making.

MLR Regression Parameter Estimates

HeartDisease = 0.285(HighBP) + 0.580(GenHlth) + 0.793(Age) + 0.40(Sex) + 0.33(HighChol) + 0.311(Stroke) + 0.149(DiffWalk) + 0.191(Smoker) + 0.055(MentHlth) - 0.09(Income) - 0.046(HvyAlcoholConsump) + 0.09(CholCheck) - 0.003(Veggies) - 0.006(Education) + 0.023(PhysActivity) + 0.013(Fruits) - 0.021(AnyHealthcare) + 0.048(PhysHlth) + 0.073(NoDocbcCost) + 0.131(Diabetes)

PCLR Parameter Estimates

HeartDisease = (-1094.402)PC1 + (257.325)PC2 + (-139.376)PC3 + (-16.099)PC4 + (-77.170)PC5 + (18.891)PC6 (-16.212)PC7 + (53.874)PC8 + (9.632)PC9 + (-5.027)PC10 + (24.123)PC11 + (-3.335)PC12 + (-23.445)PC13 + (5.563)PC14 + (19.830)PC15 + (6.134)PC16 + (10.016)PC17 + (46.574)PC18 + (-26.289)PC19

4 Conclusion and Discussion

Hypothesis 1 - PC Logistic Regression Performs Better Than Bonehead

We hypothesized that our PC Logistic Regression model will perform better than the bonehead model, which would simply assign the same outcome to every input. While our PCLR only performed with an accuracy of 55.74% at an optimal threshold of 0.48, it is greater than the bonehead accuracy rate by 5.74%, thereby proving our hypothesis.

Hypothesis 2 - Logistic Regression Performs Better Than Bonehead

We hypothesized that our Logistic Regression model will perform better than the bonehead model, which would simply assign the same outcome to every input. The MLR model we developed reached an accuracy of 77.14%, with a optimal threshold of 0.46. This is 27.14% more accurate than the bonehead model, thereby proving our hypothesis.

Hypothesis 3 - Top Predictors

We hypothesized that Age, Smoking, HighChol, HighBP, and Diabetes would be top contributors when predicting the presence or onset of heart disease.

Through our multiple logistic regression, we found that the estimated top predictors of our best model were actually HighBP, GenHlth, Age, Sex, HighChol. We were correct in hypothesizing that HighBP, HighChol, and Age were the most influential, however, we did not account for GenHlth and Sex.

In regards to the PCLR, CholCheck, GenHlth, Age, Fruits, and Smoker were the estimates of top 5 predictors of heart disease.

Hypothesis 4 - Top Predictors Will Be Consistent Between Two Models

We hypothesized that the MLR and PCLR would both show us that Age, Smoking, High Cholesterol, High Blood Pressure, and Diabetes are the top predictors. After running our tests, we found that the two models shared predictions for HighBP, HighChol and Age, but differed in GenHlth, Sex, Smoking, and Diabetes.

Implications and Next Steps

Considering the overall performance of both models, the multiple logistic regression is a better option for producing generalizable and accurate predictions. By analyzing a comprehensive data set comprising various demographic, physiological, and lifestyle factors, the findings from this research provide a robust statistical model for predicting the presence or risk of heart disease in individuals. This can significantly aid in early symptom detection, allowing individuals at risk to seek medical attention and exercise treatment or preventative care against this deadly disease.