Application of Logistic Regression Predictive Modeling in the Computational Diagnosis of Coronary Heart Disease

James Hopham, Raja Solanki, Prasanna Pawar

Introduction

Millions of individuals worldwide are afflicted by the dangerous and ubiquitous health condition known as heart disease, a problem associated with high morbidity and mortality rate. The World Health Organization estimates that heart disease contributes to 17 million annual deaths. Cardiovascular disease (CVD) is the leading cause of mortality in wealthy nations like the United States, accounting for 50% of all fatalities. By identifying the major risk factors that correspond to underlying mechanisms of heart disease, healthcare practitioners can decide on lifestyle modifications, and medications that can enhance patients' general health and wellbeing. Cardiovascular disorders come in a variety of forms; among the various forms of CVD, Coronary Heart Disease (CHD) (also referred to as Coronary Artery Disease) accounts for 41.2% of CVD related deaths in the United States. Due to the severity of this disease, physicians will diagnose certain patients as having a 10-year-risk for the development or worsening of CHD; this is similar in purpose to having a category for a pre-diabetic diagnosis for patients at risk of developing diabetes.

Coronary heart disease is a debilitating disease in which the coronary arteries supplying blood to heart tissue are blocked by plaque buildup. Decreased blood flow results in increased arterial pressure, decreased oxygen and nutrient supply to heart tissue, and potentially can result in cardiac arrest. Extensive research has shown that factors associated with plaque buildup and inflammation increase the risk of the development of CHD such as excessive levels of glucose in diabetic patients, elevated levels of cholesterol in a patient’s bloodstream, and the overall fitness of a patient.

The goal of our study is to use predictive modeling to identify the most significant factors within our dataset for the diagnosis of a 10-year-risk of CHD. The results of this study may contribute to greater developments in clinical practice and public health initiatives meant to increase precautionary measures against the progression of heart disease, and improve outcomes for those who are at risk of further complications.

Data Description

The dataset was sourced from the Framingham Heart Study from Boston Medical Center in which data from medical charts of patients were accumulated into an ongoing dataset that is being continuously added upon to generate models that better predict Coronary Heart Disease through retrospective chart review. A subset of the data was then cleaned and uploaded to Kaggle as part of a data science workshop challenge. The original dataset was 3390 rows and 17 columns. During the data wrangling processes, rows with NA values were removed as well as the column for patient id. No significant outliers were of significant concern within our dataset; all of our patients' health indicators fell within expected ranges. Among the predictor variables,

the columns containing the predictor variables for a diagnosis of diabetes, whether or not a patient was being prescribed blood pressure medication, and whether or not a patient had a history of stroke were removed from the dataset due to an overwhelmingly large ratio of zeros in the data that caused issues with overfitting in our initial model. This correction allowed us to reduce false negatives in our data. By removing these columns, we do not lose any significant information towards our predictive diagnosis as the indicator for the severity of a patient's diabetes can be understood through their glucose levels, the risk for a stroke or thrombosis can be attributed to cholesterol levels, and whether or not a patient is prescribed blood pressure medication can be explained with their diagnosis status for hypertension.

Our final dataset contains 2927 rows and 13 variables: age, education level, sex, smoking status, number of cigarettes per day, cholesterol level, systolic blood pressure, diastolic blood pressure, BMI, and glucose level. Our response classifier variable is whether or not a patient was diagnosed with a 10-year-risk of CHD. We further allocated our dataset into a 60% training and 40% test set ratio to perform cross validation on our models due to its superior performance to our initial split of 70% to 30% on this dataset.

Summary Statistics:

Table 1: CHD Risk Diagnosis, Sex and Smoking Status Ratios

Response Variable	Yes	No	Percent Diagnosed with Not Having Coronary Heart Disease Risk
#Patients Diagnosed (Y/N) with Coronary Heart Disease Risk in Dataset	444	2483	84.83088 %

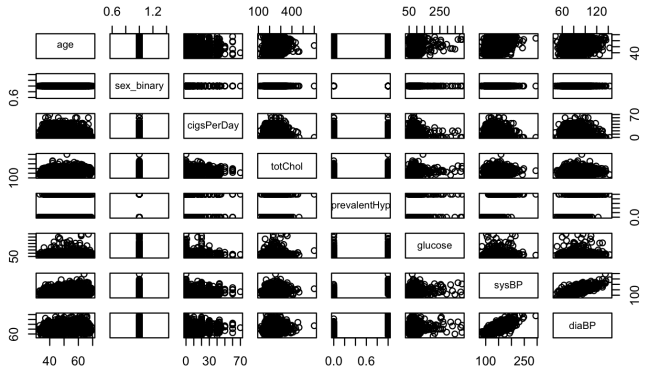
	Sex	Smoking Status
Percent	M: 44.65 % F: 55.35 %	Yes: 49.44 % No: 50.56 %

Table 2: Patient Health Indicator Summary

	Cholesterol Level	Systolic Blood Pressure (mmHg)	Diastolic Blood Pressure (mmHg)	BMI	Glucose Level
Mean	237.1	132.6	82.91	25.80	81.93

	Age	Education	# Cigarettes Per Day
Mean	49.51	1.965	9.113

Figure 1: Pairwise Assessment for Collinearity of Predictors



Methods and Results

Initial Model:

Through application of a backwards stepwise selection under AIC (Akaike Information Criterion) restrictions, we yield the following as our predictive model:

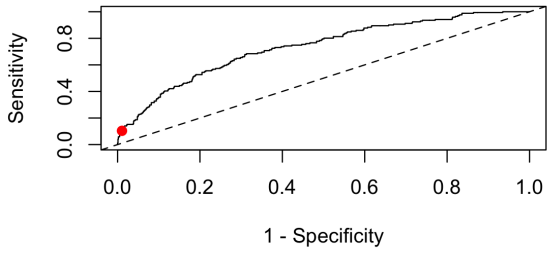
$$\widehat{p}\,(Diagnosis\ of\ Risk\ for\ Coronary\ Heart\ Disease)\ =$$

$$\frac{e^{\widehat{\beta}0+\beta 1(Age)+\widehat{\beta}2(Sex)+\widehat{\beta}3(\#Cigarettes\ per\ Day)+\widehat{\beta}4(Cholesterol\ Level)+\widehat{\beta}5(Systolic\ Blood\ Pressure)+\widehat{\beta}6(Diastolic\ Blood\ Pressure)+\widehat{\beta}7(BMI)\ +\widehat{\beta}8(Glucose)}}{1+e^{\widehat{\beta}0+\widehat{\beta}1(Age)+\widehat{\beta}2(Sex)+\widehat{\beta}3(\#Cigarettes\ per\ Day)+\widehat{\beta}4(Cholesterol\ Level)+\widehat{\beta}5(Systolic\ Blood\ Pressure)+\widehat{\beta}6(Diastolic\ Blood\ Pressure)+\widehat{\beta}7(BMI)\ +\widehat{\beta}8(Glucose)}}$$

Table 3: Initial Logistic Regression Confusion Matrix

	Actual		
Prediction	**0**	**1**	**Sum**
0	991	151	1142
1	9	20	29
Sum	1000	171	1171

Figure 2: Initial Logistic Regression ROC Curve



Our initial model yields an AUC (Area Under the Curve) of .734.

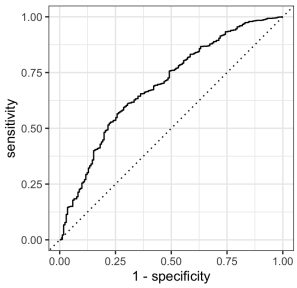
Application of Machine Learning Classification Models:

In an attempt to test other machine learning models, we applied machine learning classification models on our dataset with the same 60% to 40% split for training and test datasets. When conducting decision trees the XGBoost algorithm was used due to its integration of gradient

boosting to minimize loss, and because of its popular use in industry. Overall, the models performed worse than our initial logistic regression model in the following descending order: naive-bayes, random forest, decision trees, and k-nearest neighbors.

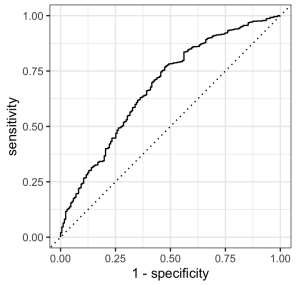
Due to these results, for the development of our final model we chose to further scrutinize our logistic regression model to better tune the model to improve its performance through checking pairwise associations among the predictors. Upon doing so we observe an understandable association between systolic and diastolic blood pressure.

Figure 3:Naive-Bayes



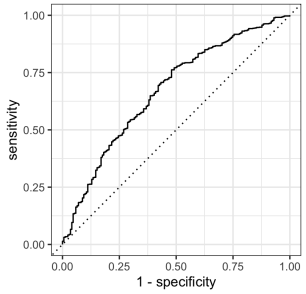
AUC: .6930117

Figure 4:Random Forest



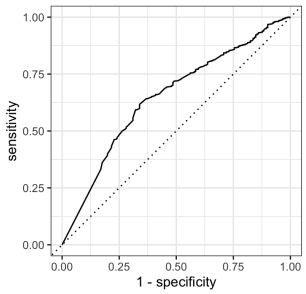
AUC: .68101

Figure 5: Decision Trees XGBoost



AUC: .6699298

Figure 6: K-Nearest Neighbors



AUC: .6480175

Conclusion

Ultimately diastolic blood pressure was removed since the conditions for increased systolic pressure inevitably affect diastolic blood pressure levels. We further tested accuracy of the logistic regression model when removing BMI and diagnosis of hypertension since these predictors should be heavily correlated with systolic blood pressure; however, we found that keeping diagnosis of hypertension as a predictor improved our model's accuracy. Our final model results in the following:

$$\widehat{p}\,(Diagnosis\ of\ Risk\ for\ Coronary\ Heart\ Disease) = \frac{e^{\widehat{\beta}0 + \widehat{\beta}1(Age) + \widehat{\beta}2(Sex) + \widehat{\beta}3(\#Cigarettes\ per\ Day) + \widehat{\beta}4(Cholesterol\ Level) + \widehat{\beta}5(Hypertension\ Diagnosis) + \widehat{\beta}6(Systolic\ Blood\ Pressure) + \widehat{\beta}7(Glucose)}}{1 + e^{\widehat{\beta}0 + \widehat{\beta}1(Age) + \widehat{\beta}2(Sex) + \widehat{\beta}3(\#Cigarettes\ per\ Day) + \widehat{\beta}4(Cholesterol\ Level) + \widehat{\beta}5(Hypertension\ Diagnosis) + \widehat{\beta}6(Systolic\ Blood\ Pressure) + \widehat{\beta}7(Glucose)}}$$

Table 4: Logistic Regression Summary

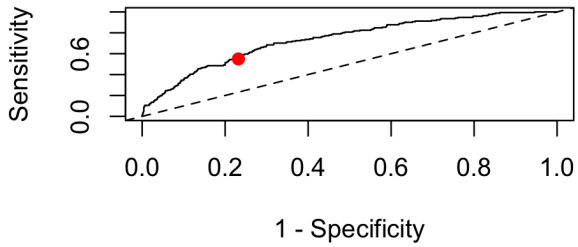
	Estimate	Std. Error	Z- Value	Pr(>|z|)

(Intercept)	-8.965780	0.823328	-10.890	< 2e-16
Age	0.062062	0.009570	6.485	8.88e-11
Sex	0.419018	0.154788	2.707	0.00679
CigsPerDay	0.027042	0.006052	4.469	7.87e-06
TotChol	0.003383	0.001549	2.183	0.02900
Hypertension	0.211847	0.192964	1.098	0.27226
SysBP	0.013152	0.004097	3.210	0.00133
Glucose	0.009973	0.003185	3.131	0.00174

Table 5: Final Logistic Regression Model Confusion Matrix

	Actual		
Prediction	**0**	**1**	**Sum**
0	768	77	845
1	232	94	326
Sum	1000	171	1171

Figure 7: Final Logistic Regression Model ROC Curve



Our final model performs marginally better than the initial model and yields an AUC of .7393 (an increase of .005); after testing other variations, this model performs the best among them all.

Limitations/Further Study

This project was limited with regards to not only the number of patients available, but also in the number of predictor variables available for analysis. Other markers with the potential to be highly significant in patient diagnosis can be accrued through methods such as genome wide association studies and genotype screening of relevant genes associated with an increased risk of heart disease. Other health indicators such as troponin levels, which can be used to assess the severity of heart tissue death, should also be collected and included with the patient data in order to produce a more robust model. Additionally, deep learning methods such as recurrent neural networks and convolutional neural networks can be used to analyze image data including angiograms to develop a diagnosis.

Works Cited:

1) https://www.kaggle.com/datasets/christofel04/cardiovascular-study-dataset-predict-heart-disea
2) https://www.bmc.org/stroke-and-cerebrovascular-center/research/framingham-study
3) https://pubmed.ncbi.nlm.nih.gov/31613540/
4) https://pubmed.ncbi.nlm.nih.gov/30790284/
5) https://pubmed.ncbi.nlm.nih.gov/32119297/

Appendix:

GitHub Repository