Perfusion Prediction

Joe Krinke

12/10/2019

Summary

In this paper I present a model to determine if an individual has a heart perfusion defect based on physiological data. This model uses easily obtainable data such as an individual's sex, blood pressure, type of overall chest pain, and if they experience chest pain during exercise. Being male, having high blood pressure, and having exercise chest pain appear to increase the risk of having a perfusion defect. The effects of each type of overall chest pain on defect rate vary. Overall, this model performs fairly well, with a sensitivity of 74.3% and a specificity of 76.8%. The model's predicted perfusion status can be used to help medical professionals determine if additional testing is needed and reduce healthcare costs.

Introduction

A perfusion defect is when an area of the heart has reduced blood flow under increased stress. The decreased amount of blood flow can lead to damage to the heart muscles, which can cause heart failure and other medical problems. Additionally, the detection of perfusion issues can be used to determine if someone has coronary artery disease, if their stent is working, or if they are an especially risky patient to perform heart surgery on. Categories of perfusion issues include reversible defects and irreversible defects. The current diagnosis technique is a myocardial perfusion scan, an expensive test which requires an individual to exercise while connected to electrodes. However, this test is time-consuming, expensive, and may cause harm to patients. On top of this, symptoms for this ailment are vague, which can lead to late diagnosis and unneeded damage to the heart. The goal of this project is to build a logistic regression model to predict an individual's perfusion status using physiological data. Being able to diagnose the disease earlier could allow for early intervention and produce better outcomes for patients. Alternatively, one could use the results from this model to determine whether to test someone, which could reduce testing costs and expose patients to less potential harm.

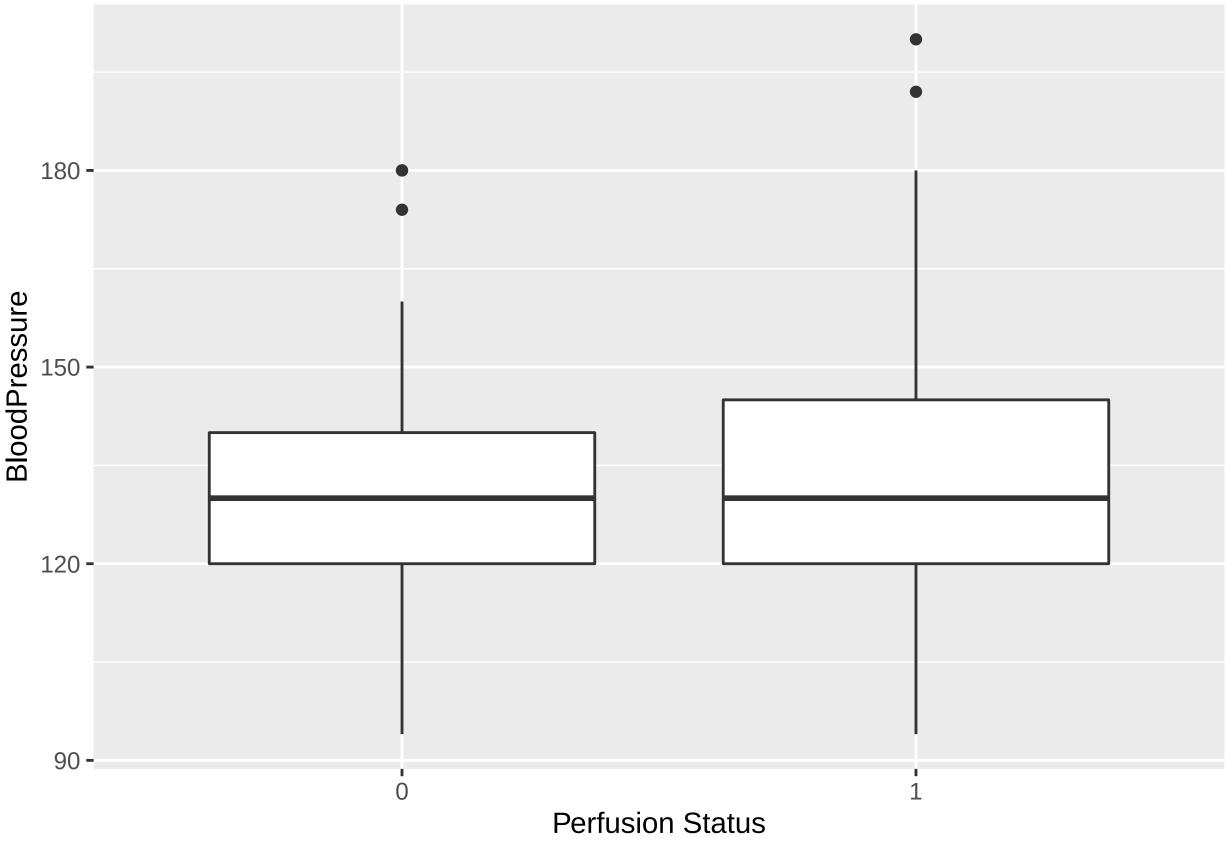
Data Used

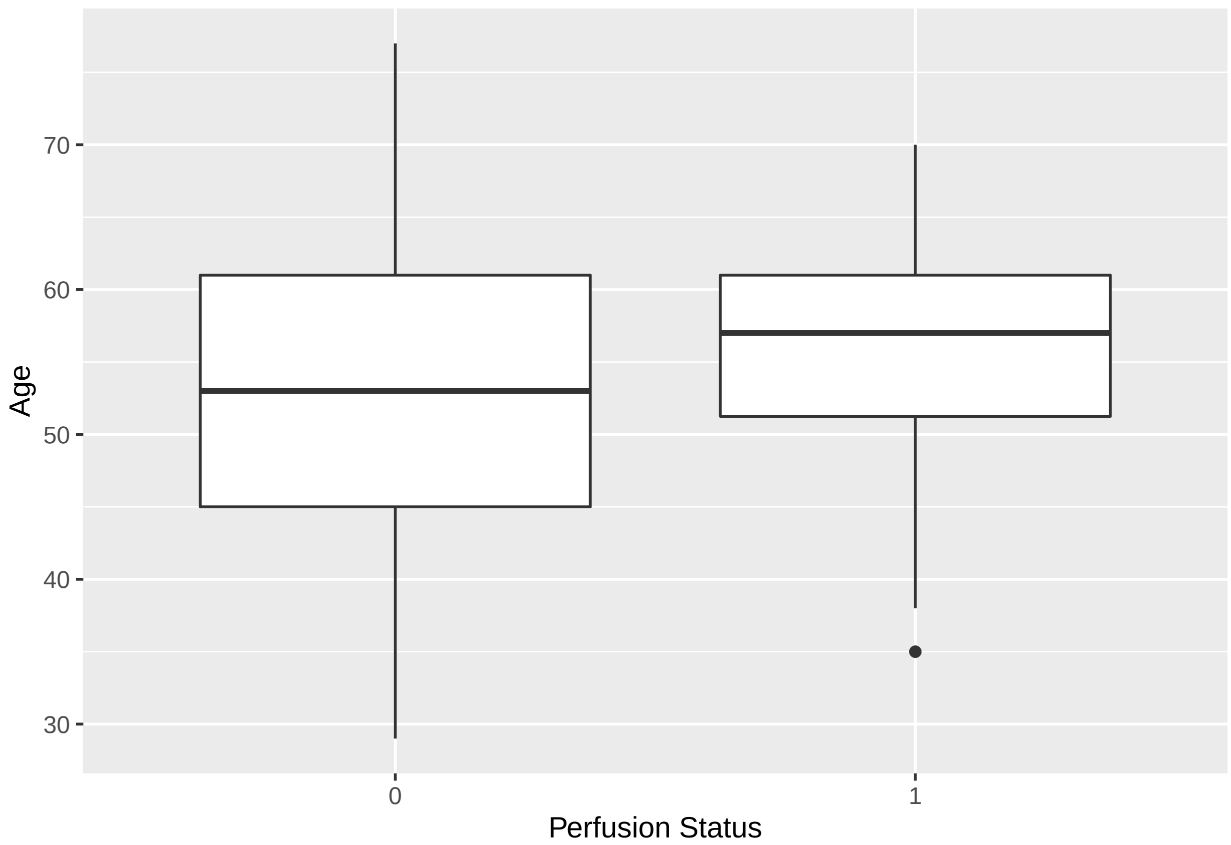
The data used was obtained from the HCI Machine Learning Repository. This dataset was collected from four hospitals: the Hungarian Institute of Cardiology, University Hospital, Zurich, University Hospital, Basil, and the V.A. Medical Center. This data included 76 attributes related to a person's demographics and heart health. However, the vast majority of research has been focused on the V.A. dataset, so I chose to use that data for my analysis. The subset of data I chose had 14 attributes and 270 observations; no missing data was observed. The variable for perfusion defects was originally separated into multiple categories corresponding to severity: no defect, reversable defect, and fixed defect. I recoded this variable into a binary defect/non-defect in order to deal with heavily imbalanced data (one class only had 14 observations). The variables Number of Colored Vessels, Slope of Peak ST, ST Depression, Heart Disease Status, and Maximum Heart Rate were excluded from the modeling and EDA process. This is because these variables were collected during the test for perfusion (or in later diagnostic stages). It would not make sense to include the previously mentioned variables in our model. A list of the final variables and their summary statistics are given below (table 1).

	Age	Sex	ChestPainType	BloodPressure	FastingBloodSugar	Resting Electrocardiograph	ExerciseAngina	BinaryPerfusion
	Min. :29.00	0: 87	1: 20	Min. : 94.0	0:230	0:131	0:181	Min. :0.000
	1st Qu.:48.00	1:183	2: 42	1st Qu.:120.0	1: 40	1: 2	1: 89	1st Qu.:0.000
	Median :55.00	NA	3: 79	Median :130.0	NA	2:137	NA	Median :0.000
	Mean :54.43	NA	4:129	Mean :131.3	NA	NA	NA	Mean :0.437
	3rd Qu.:61.00	NA	NA	3rd Qu.:140.0	NA	NA	NA	3rd Qu.:1.000
	Max. :77.00	NA	NA	Max. :200.0	NA	NA	NA	Max. :1.000

EDA

I conducted two types of exploratory analysis. For categorical variables I examined frequency tables and peformed chi-squared tests, and for continuous variables I examined binned plots and boxplots. There seemed to be no issues in the patterns of any of the binned plots (see appendix). The most promising continuous predictors were age and blood pressure. Below are boxplots of each against perfusion status (figures 1-2).





A number of categorical variables appeared to be potentially significant. The table below shows each categorical variable and its corresponding chi-squared value (table 2).

	P-Value
Sex	1.20980852656631e-10
ExerciseAngina	3.10990989490073e-07
FastingBloodSugar	0.485713505122641
ChestPainType	1.28531657847031e-06

Model Building Process

I began by centering the continuous variables. Next, based on the results of the EDA I built an initial logistic regression model using age, sex, chest pain, blood pressure, resting ECG, and exercise angina as predictors.I avoided using interaction terms, as their inclusion in the model would likely increase the total number of predictors included (you can't have an interaction without both variables) which would go against the study's goal of reducing needed testing. This model performed fairly well, with a sensitivity of .773, specificity of .716, and an overall accuracy of .748. However, since the goal of my analysis was to reduce unneeded testing, I performed backwards stepwise selection using AIC as my selection criterion. This reduction in needed variables should help minimize the amount of inion that physicians would have to collect from patients. Below is a table describing the final model selected (table 2). Note that the coefficients and confidence intervals have all been exponentiated.

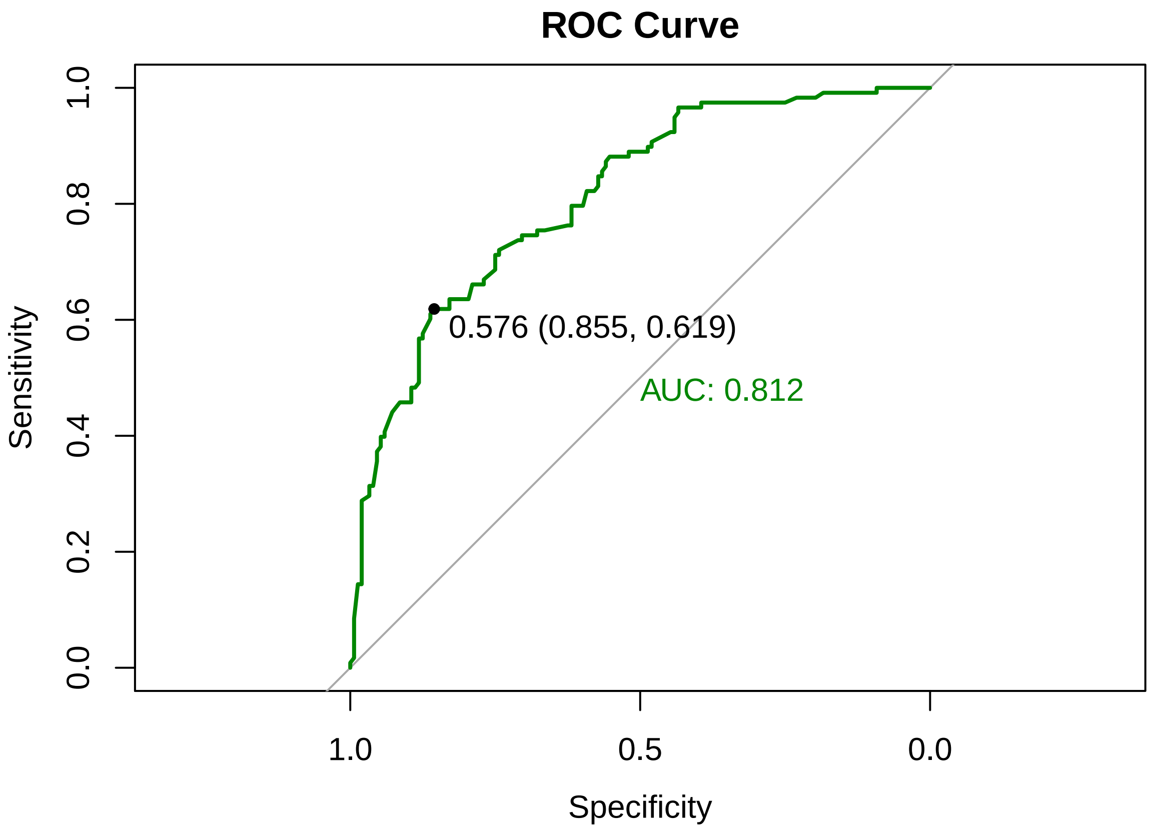
```
## Waiting for profiling to be done...
```

Variables	Exponentiated Coefficient	2.5 %	97.5 %	P-value
(Intercept)	0.0036193	0.0002112	0.0517658	0.0000568
BloodPressure	1.0226549	1.0063662	1.0400594	0.0000000
ChestPainType2	0.7148181	0.1993026	2.5528484	0.6027522
ChestPainType3	1.0931786	0.3592796	3.4264770	0.8758249
ChestPainType4	3.0623836	1.0451495	9.3213664	0.0427911
ExerciseAngina1	2.1011449	1.0983323	4.0432819	0.0073252
Sex1	9.1166277	4.4619531	20.1790471	0.0250291

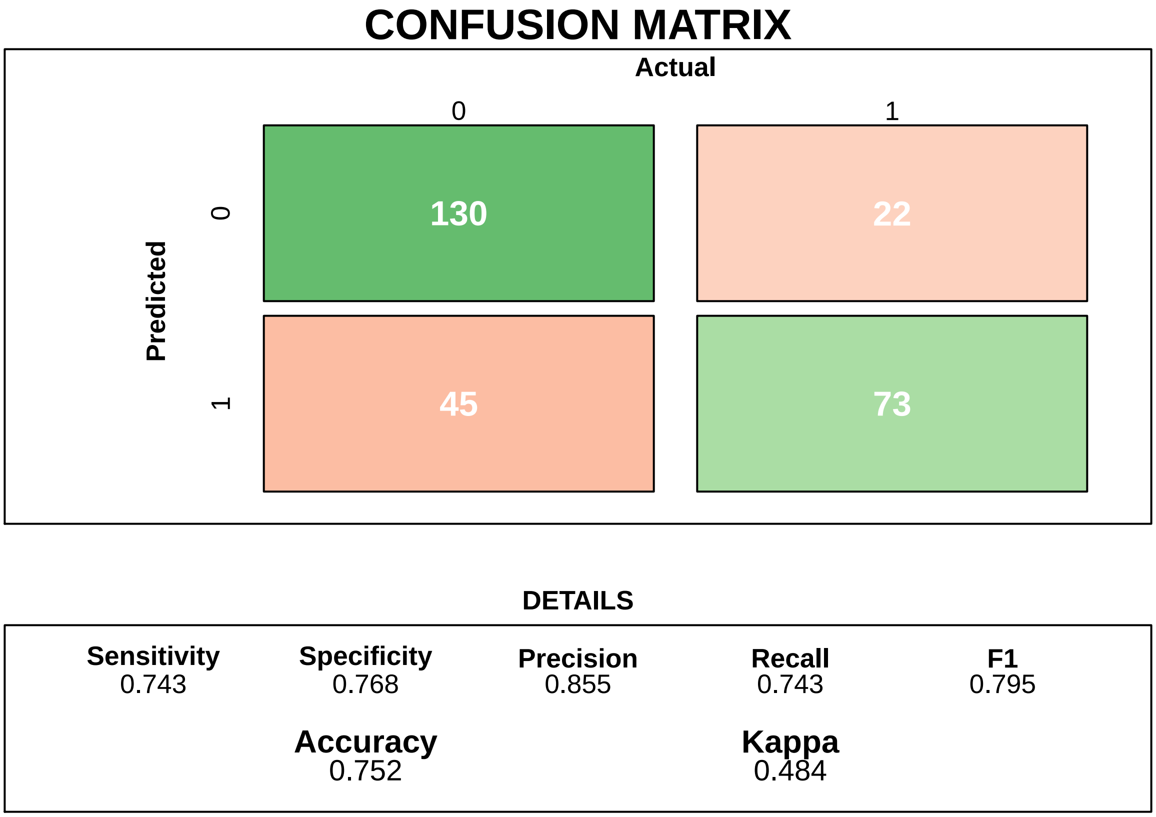
The optimal threshold for the final stepwise model is .576 with an AUC of .812 (figure 3).

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

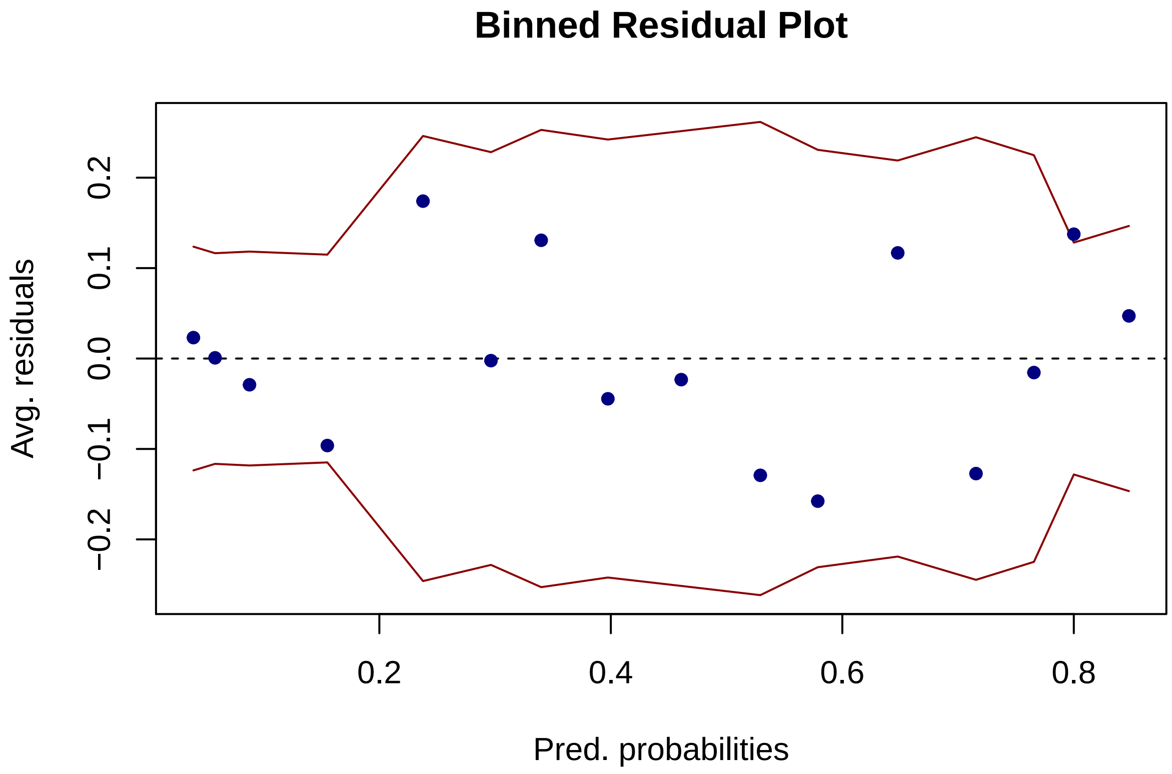


The reduced model actually had a slightly higher accuracy than the full model (figure 4). However, the increase in accuracy was accompanied by a small reduction in sensitivity. I consider this loss in sensitivity tolerable, as the reduced model no longer includes variables that require ECG testing or blood sugar testing. The savings associated with not having to run these tests likely would outweigh the costs associated with the sensitivity reduction.



In addition to examining the confusion matrix I also looked at various binned residual plots to evaluate the fit of the model. There seemed to be no issues in the overall binned residual plot apart from a single point that is on the edge of the SD lines (figure 5). The binned residual plots for individual predictors showed no problems and can be seen in the appendix. Additionally, I examined the VIFs and determined there were no mulitcolinearity issues. A table with the VIFs can be found in the appendix.

```
#Binned plot
binnedplot(fitted(model_final),residuals(model_final,"resp"),xlab="Pred. probabilities",col.int="red4",
```



Conclusion

In this paper I present a model to determine if an individual has a heart perfusion defect based on physiological data. This model uses easily obtainable data such as an individual's sex, blood pressure, type of overall chest pain, and if they experience chest pain during exercise. Being male, having high blood pressure, and having exercise chest pain appear to increase the risk of having a perfusion defect. The effects of each type of overall chest pain on defect rate vary. Overall, this model performs fairly well, with a sensitivity of 74.3% and a specificity of 76.8%. The model's predicted perfusion status can be used to help medical professionals determine if additional testing is needed and reduce healthcare costs.

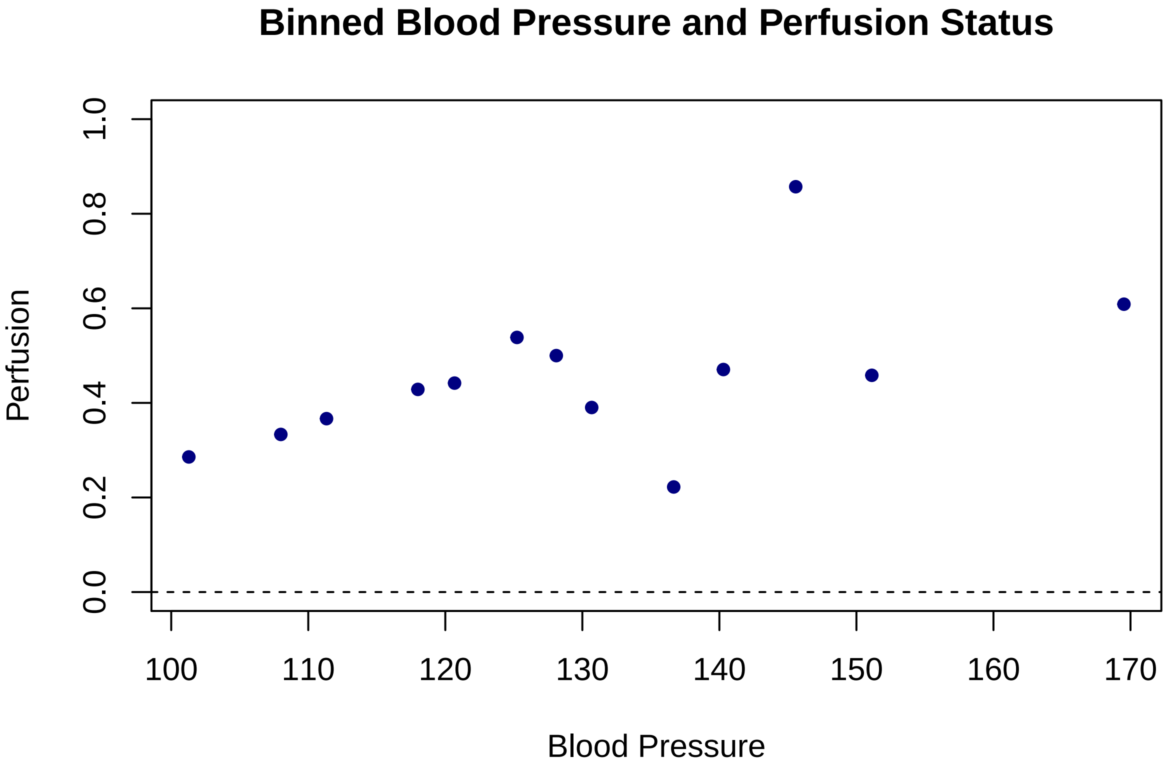
Limitations

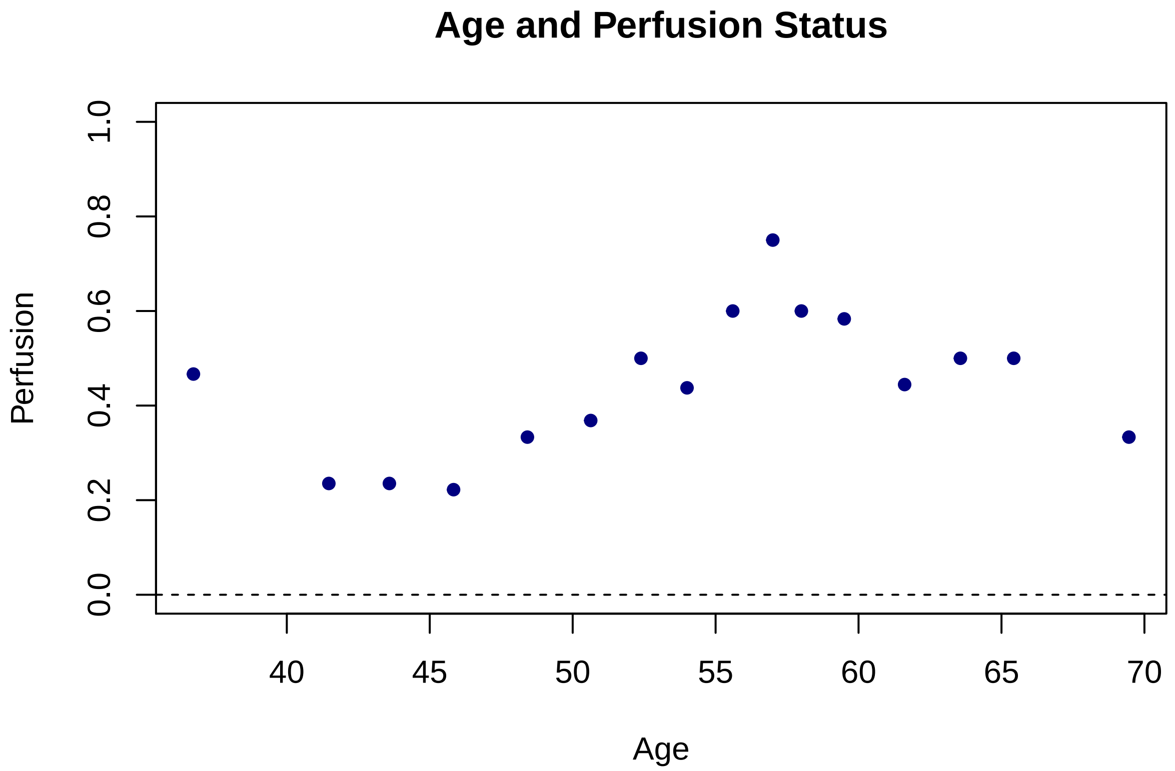
While the logistic model presented has its uses, it is not without limitations. In reality, heart perfusion is not typically measured in a binary fashion. There are varying degrees of perfusion that the model was unable to capture due to unbalanced data. Obtaining more data could allow one to build a multiclass model that would be able to provide more granular insights into the severity of a patient's perfusion defect. This more sensitive model could allow for opportunities for medical intervention, especially for those in the "reversable defect" stage.

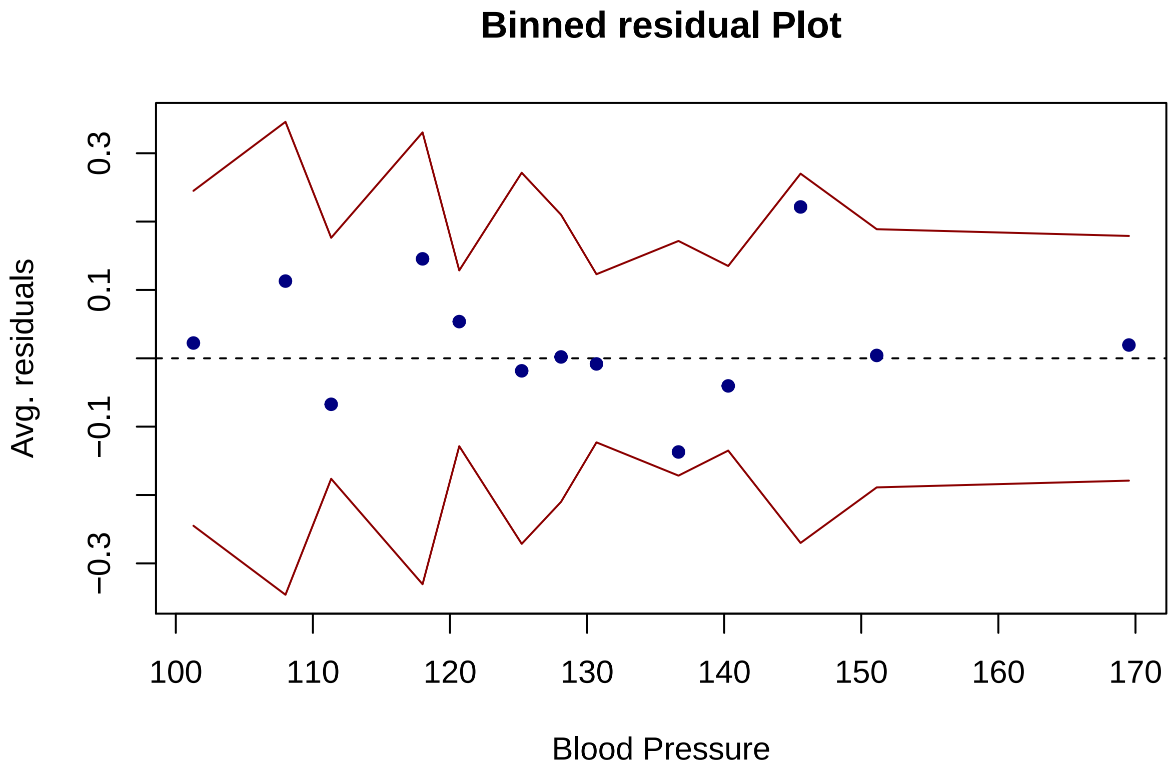
In addition, regression models as a whole assume that the effect of a variable is uniform across its entire range. Let's consider, for example, using a patient's blood pressure as a predictor. This model assumes that every one-unit increase in blood pressure increases the likelihood of a defect by the same amount. It is more likely, in reality, that having blood pressure below/under a certain threshold would be associated with illness. Future work could review the literature to determine such a threshold and use a binary blood pressure variable instead. This line of reasoning would also apply to other continuous variables as well.

It is also important to understand that the goal of this study was not to evaluate how each factor is associated with perfusion. The goal was to find a parsimonious model that could potentially reduce testing costs. Future research could include using more control variables that were excluded by this model (age, race, smoking, etc.) in order to better isolate each effect.

Finally, one should remember that the model produced should not be used to try to demonstrate direct causality. The model was not constructed with inference in mind and the human body's complexity makes drawing a direct line of causality difficult. Many systems in the body interact and many diseases stem from the same underlying causes. The logistic regression model should be viewed as demonstrating association and nothing more. ## Appendix




Age and Perfusion Status
Perfusion
0.0
0.2
0.4
0.6
0.8
1.0
40
45
50
55
60
65
70
Age



	GVIF	Df	GVIF^(1/(2*Df))
Sex	1.098140	1	1.047922
ChestPainType	1.230082	3	1.035116
BloodPressure	1.094401	1	1.046136
ExerciseAngina	1.162232	1	1.078069