

# KEY TO THE HEART

## AN EXPLORATION OF CARDIOVASCULAR DISEASE

Anonymous

December 2020

Through statistical methods including Principal Component Analysis, logistic regression, and boosting, a deeper understanding of the underlying factors that may lead to cardiovascular disease have been explored. Identifying these factors early on could reduce an individual's chance of falling victim to the United States' leading cause of death.

CARDIOVASCULAR  
RESEARCH





## **1.) Introduction**

Cardiovascular disease is the leading cause of death in the United States according to the CDC. While many factors will affect an individual's chance of getting heart disease, some stand out more than the rest. This research project was oriented at determining these most significant factors and the resulting effects of reducing or increasing them. In addition, from a more statistical standpoint, this project will also explore the relatively new topic of boosting with logistic regression.

Anyone can look retrospectively at how a patient contracted heart disease. The goal, or hope, of this project is to proactively identify individual's at risk and focus on certain factors to optimize one's chance of avoiding the disease.

Boosting is a modern technique that is best used with linear regression. However, logistic regression can still be used with some minor alterations on the technique. The effectiveness of boosting a logistic model will be analyzed.

Given a dataset with 70,000 patients and characteristics ranging from age and weight to glucose and cholesterol levels, two statistical models were created to identify factors that influence the likelihood of an individual to get cardiovascular disease and to experiment with modern methods to find new meaning in healthcare analytics.

After completing this project, I believe with more patients and more variables influencing likelihood these models could have a large impact on reducing the number of deaths from cardiovascular disease.



## 2.) Data Description

The chosen data set contains eleven characteristics from 70,000 patients screened for cardiovascular disease. I used the binary variable *cardio* as the target variable, 1 indicating a patient has the disease and 0 if not. The patients are split with approximately half with heart disease. The following variables are the independent variables: *age*, *gender*, *height*, *weight*, *systolic blood pressure (ap\_hi)*, *diastolic blood pressure (ap\_lo)*, *cholesterol*, *glucose*, *smoking*, *alcohol intake*, and *physical activity*.

Grouping variables on type we have three categories: objective features, examination features, and subjective features.

### **OBJECTIVE FEATURES:**

*Age*, *gender*, *height*, and *weight* are factual information for each patient.

- *Age* was given in days and converted into years for easier interpretation
- *Height* is measured in centimeters
- *Gender* is denoted by 1 for female and 2 for male

*Age*, *height*, and *weight* all resemble a normal distribution which is to be expected since we have many samples (Appendix A). This effect is described by the Central Limit Theorem, a famous result that essentially says that as a sample size increases it will approach a normal distribution.

Using Principal Component Analysis, a technique that reduces the complexity of our variables to capture variability of observations, I observed that the objective feature *age* is the most correlated with *cardio*. *Gender* and *height* are highly collinear so I decided against using these together in the model (Appendix B). I chose to keep *height* because I wanted more continuous variables in my model.



## EXAMINATION FEATURES:

*Systolic* and *diastolic blood pressure*, *cholesterol*, and *glucose* are the results from a medical examination.

- *Systolic* and *diastolic blood pressure* are measures of blood coming in and out of the heart using integer values ranging from around 50 to 200.
- *Cholesterol* and *glucose* were measured using factors from 1 to 3 indicating normal, above normal, and well above normal, respectively.

Values of *systolic* and *diastolic blood pressure* outside the range of approximately 50 to 200 are either singular, extreme cases or incorrectly entered data and thus outliers were identified and removed using a Bonferroni outlier test. The Bonferroni test identifies patients with blood pressure at least 3 standard deviations from the mean.

For *systolic blood pressure*, patients with values above 130 are considered hypertensive (having high blood pressure). For *diastolic blood pressure*, patients with values above 80 are considered hypertensive. These two values are measured concurrently and, as observed in PCA analysis, they are collinear (App. B).

The large majority of patients have *cholesterol* and *glucose* levels that are normal, with the remaining portion split evenly between above normal and well above normal.

## SUBJECTIVE FEATURES:

The amount that a patient smokes (*smoke*), intakes alcohol (*alco*), and exercises (*active*) are subjective features given by patients.

- Over 90% of patients reported “No” when asked whether or not they smoke.
- Almost 95% of patients reported intaking a small amount of alcohol on average.
- Approximately 80% of patients claim to be active in their lives.




### **3.) Model Selection and Methodology**

A model and a heart both take something in and pump something out in one way or another. Both must work proficiently to satisfy their purpose. The purpose of this paper is to gain an understanding rather than strict prediction. Of course, prediction offers valuable information but the prediction ultimately relies on the structure of the model. Hence, a multiple logistic regression model was chosen in addition to an exploration using boosting with logistic regression trees.

#### **MULTIPLE LOGISTIC REGRESSION**

Given that the target variable is binary, choosing a logistic regression was a sensible choice to explore the predictive abilities of a model. Since many variables were used in the model it became a multiple logistic model. A multiple logistic model uses given variables to build a model that will predict the likelihood of an observation having a target variable value of '1'. The model uses Maximum Likelihood Estimation which is a method that uses given data to yield the probability of a positive result that most closely resembles the data.

It is important to note here that when selecting a multiple logistic model I was more interested in qualitative results over the quantitative ones. That is to say that predicting whether or not someone has cardiovascular disease is not an especially useful ability since there are blood tests and electrocardiograms that will diagnose a patient much more accurately. Supposing that prediction was a valuable asset, the steps forward following the prediction would be unclear from just a probability value. The desired information from the multiple logistic regression was



the magnitude and sign of the coefficients of each variable term. Once the model is produced it is quite simple to analyze the coefficients to determine positive and negative effects on the probability and respective degrees of intensity.

## BOOSTING

Boosting essentially creates many simple trees that learn from each previous tree. In order to make a subsequent tree, the algorithm analyzes the previous tree to find which part had the most error. Then the subsequent tree is artificially biased to improve that specific part of the tree to make sure it predicts well. This process is repeated until there is a full forest of trees that have learned from one another. Each tree is quite simple and are called weak learners. However, having a large forest of weak learners is similar to the wisdom of the crowd and becomes very powerful. Having the simpler trees avoids overfitting our model, something that would happen if we used many very complex trees made to reduce error in the training dataset. Boosting goes further than random forests by introducing a simple learning algorithm to strengthen the model.

Logistic regression yields probabilities, so the accuracy when compared with actual outcomes can be limited. The goal with boosting logistic regression was to determine if using probabilities split at 0.5 would be accurate indicators of outcome. In other words, explore the notion of whether the boosting algorithm will be accurate in predicting an outcome given a somewhat simple method of converting probabilities to binary.

## 4.) Results

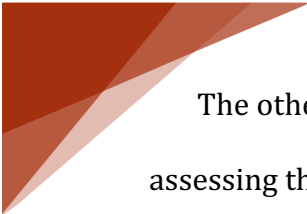
After completing the multiple logistic regression model, it was time to analyze the coefficients to gain an understanding into how each factor affects the probability of contracting heart disease. An individual has a higher likelihood of contracting the disease if the individual:

- Is older
- Weighs more
- Is shorter
- Has high systolic and diastolic blood pressure
- Has low levels of glucose
- Has high cholesterol
- Doesn't smoke
- Doesn't drink alcohol
- Is not active

To much surprise, factors that would typically be considered healthy for an individual (activity, not smoking, not drinking, etc.) may not lower an individual's likelihood. Very interestingly, the logistic model predicts that individuals who smoke and drink alcohol may

actually reduce their chances of heart disease. However, the caveat of this result is that smoking and alcohol variables are subjective features and therefore are unclear as to the degree of consumption since they are binary. Additionally, height has a significant effect on the likelihood as well as age. Thus we can conclude that height and age, as they increase, can increase an individual's likelihood of contracting cardiovascular disease. Appendix D confirms the results in visual plots.

<i>Dependent variable:</i>	
	Cardio
Age	0.05201*** (0.00133)
Weight	0.01260*** (0.00067)
Height	-0.00444*** (0.00113)
Systolic	0.05384*** (0.00067)
Diastolic	0.00012** (0.00005)
Glucose	-0.12151*** (0.01741)
Cholesterol	0.50617*** (0.01534)
Smoke	-0.14800*** (0.03295)
Alcohol	-0.18490*** (0.04132)
Active	-0.22034*** (0.02154)
Constant	-10.11667*** (0.21137)
Observations	69,982
Log Likelihood	-39,784.13000
Akaike Inf. Crit.	79,590.25000
<i>Summary of regression</i>	



The other half of this paper, and the arguably more interesting aspect, is assessing the effectiveness of creating a boosted forest of logistic regression trees. While it may sound like an obscure statistical method that won't have any real world application, boosting is a very powerful technique as mentioned priorly. It is well known that boosting linear regression trees can greatly reduce the mean square error so naturally we investigate the accuracy of a boosted logistic forest.

Age, systolic blood pressure, weight, and height are the most influential variables in predicting the likelihood from the boosted forest (Appendix C).

After creating the boosted forest we can predict the probability of contracting heart disease and then convert probabilities with values above 0.5 to be positive cases (cardio values of '1') and below to be negative (cardio values of '0'). Comparing the predicted values to the actual the boosted forest yielded an error rate of approximately 21%. While this is not a very low error rate it performed better than the multiple logistic regression that had an error rate of 27%.

While this is still a high error rate, it would be premature to claim that logistic regression and boosting do not work well together. There is a very large assumption made when converting probabilities to 1s and 0s, something that will artificially inflate the error rate. The next step in studying boosting with logistic regression would be alternative techniques to reduce the error rate given probabilities for each patient. Additionally, ten variables do not encompass all influences on an individual. A plethora of factors are all intertwined, so by the nature of medical data it can be extremely difficult to accurately model.



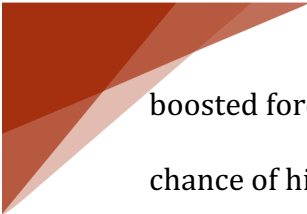


## **5.) Conclusion**

For the first objective of this paper, many previously considered obvious ideas were challenged. For instance, the factors associated with a healthy lifestyle may not have reduce the likelihood of contracting cardiovascular disease. It is important to note that this was research strictly on heart disease and thus does not suggest that smoking and alcohol consumption always have beneficial properties.

As previously mentioned, predicting the likelihood was not the main focus of this project. The analysis gathered from the regression as well as using Principal Component Analysis allowed for a deeper understanding of the factors inherent in heart disease. The next step for healthcare professionals would be to use their expertise to filter through the variables to select ones most easily altered on a patient to patient basis. For example, an individual's information could be run through the model to get a preliminary probability. This is only to be used as a flag of sorts to make the patient aware of their condition. The important step is to identify the factors that could most reduce the individual's likelihood, whether that be increased activity, medication to reduce blood pressure, or losing weight.

The exploration of boosting a logistic regression model was more of passion project, one that did not yield especially encouraging results at first glance but does have potential in the healthcare field. In terms of prediction it performed better than the multiple logistic model and provided valuable information regarding relative influence of each variable. With more time and data, different methods could be used to reduce the error rate and create a more accurate and powerful

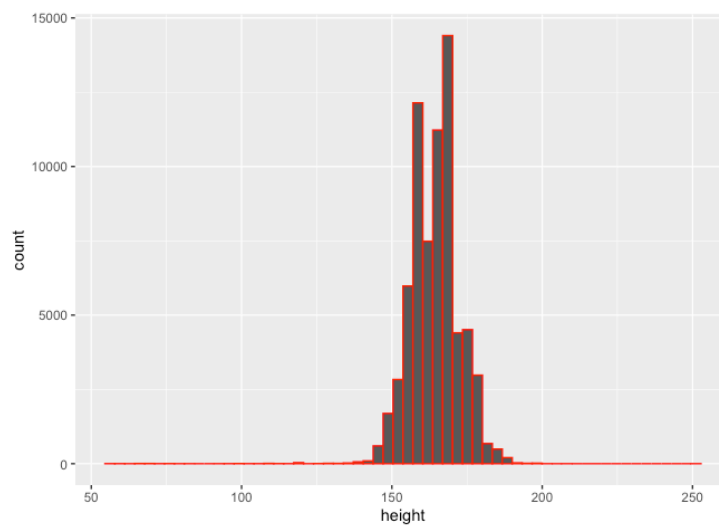
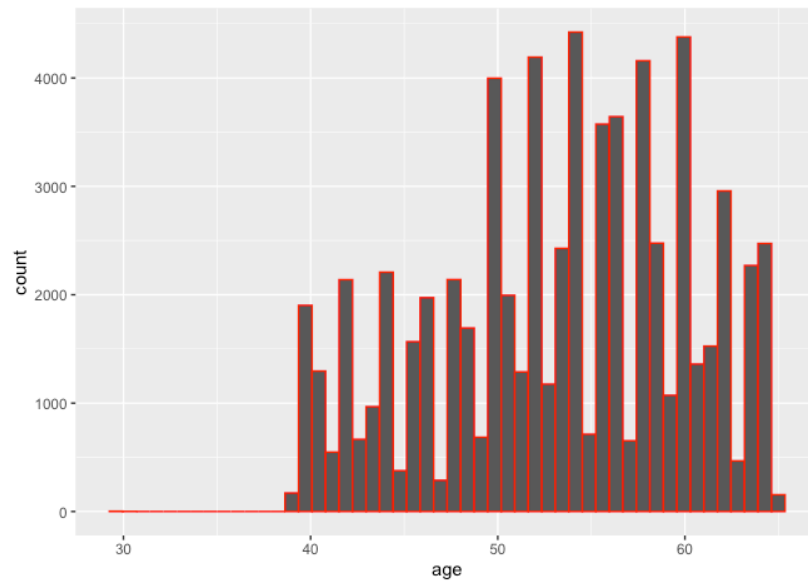


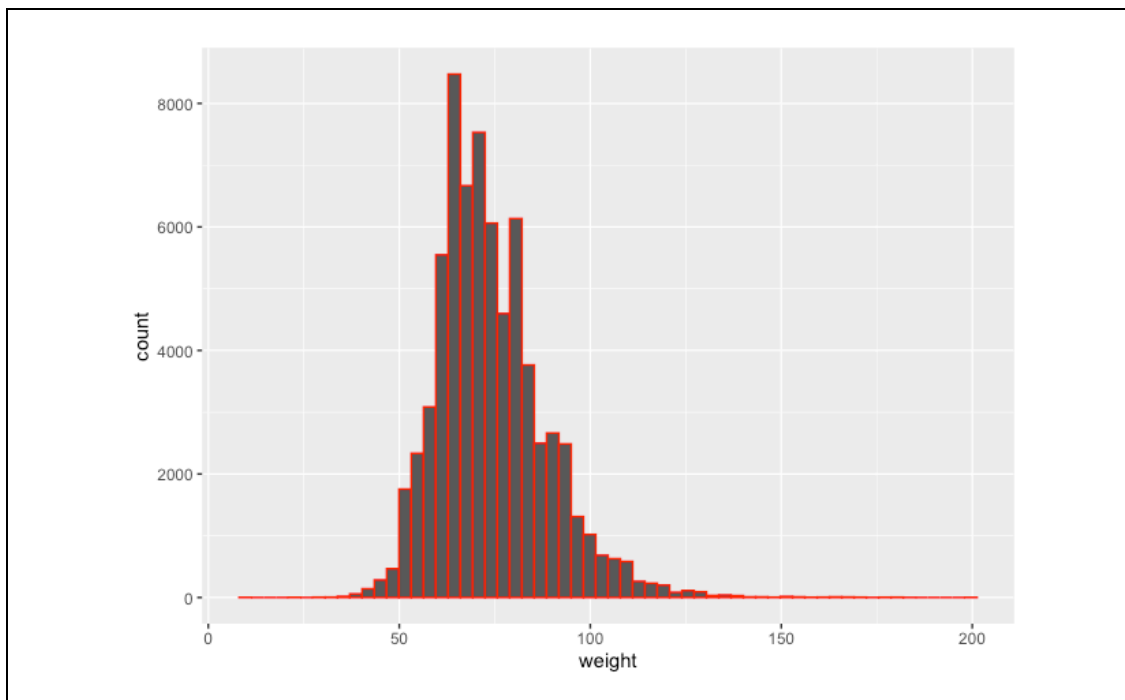
boosted forest. Since logistic regression yields probabilities there will always be a chance of high probabilities having negative outcomes and vice versa so 1s and 0s don't give the full story. Given the sometimes unpredictable nature of medical data a higher error rate is almost to be expected, especially with a low number of variables.

It is my hope that cardiovascular disease research will continue to improve and find innovative approaches to reduce the fatalities each year. Although this project only scratched the surface, I am a firm believer in “every bit counts”.

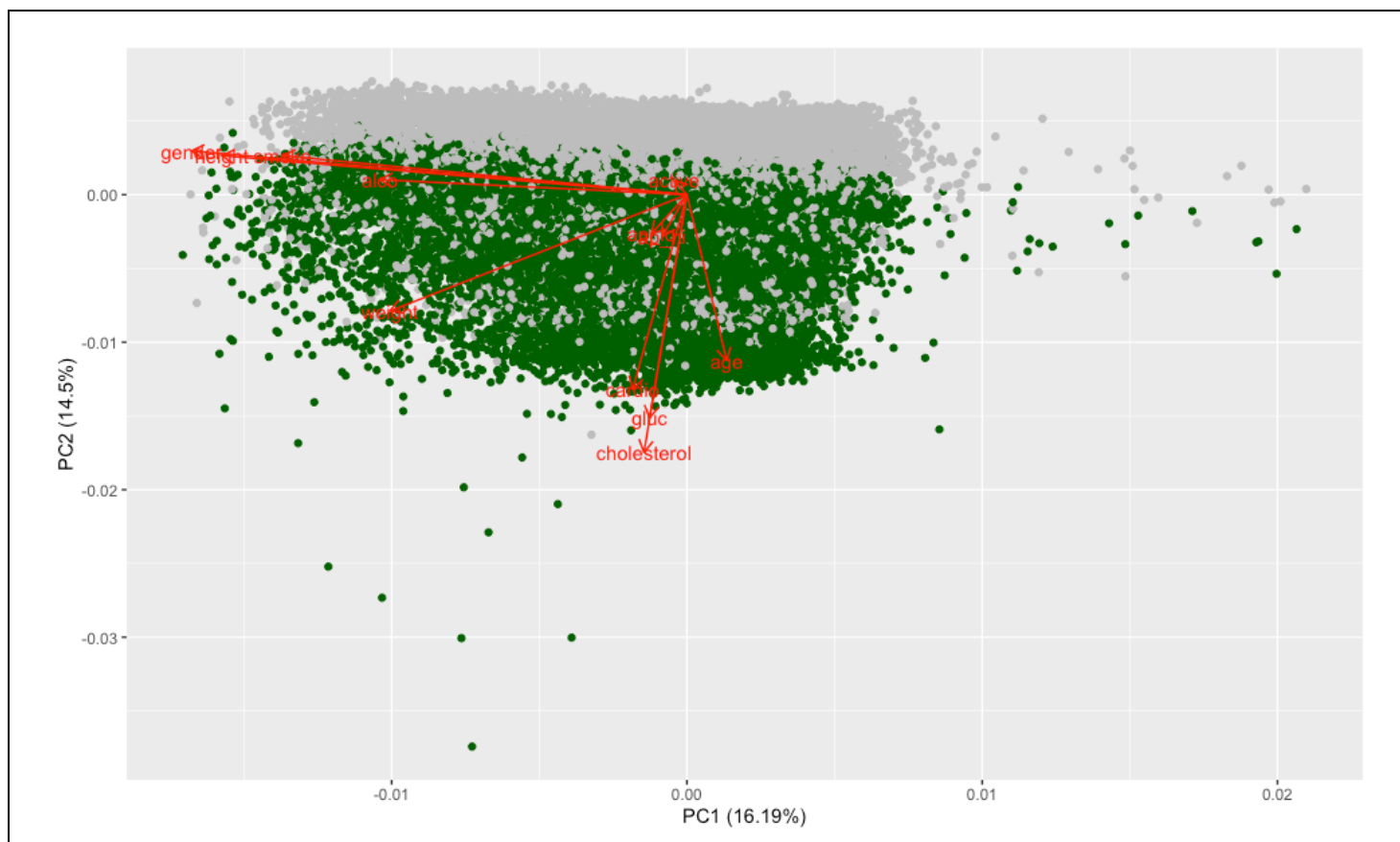
# Appendix

## A.) Histograms of age, height, and weight



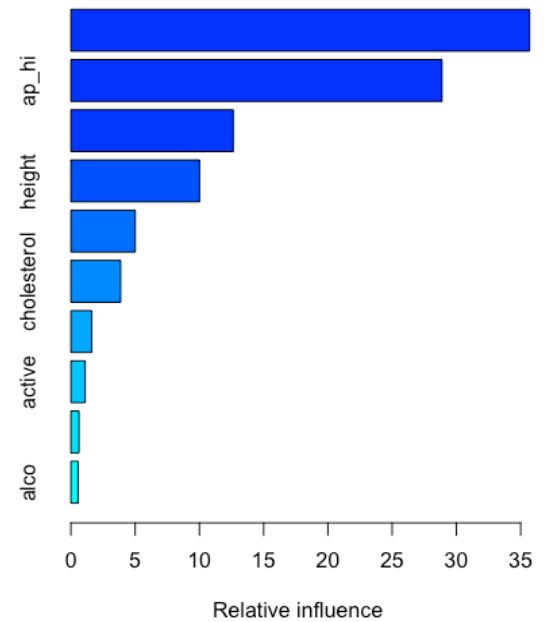


## B.) Principal Component Analysis



### C.) Boosting summary

	var	rel.inf
age	age	35.6903693
ap_hi	ap_hi	28.8767790
weight	weight	12.6322039
height	height	10.0076587
ap_lo	ap_lo	5.0086737
cholesterol	cholesterol	3.8707801
gluc	gluc	1.6185805
active	active	1.1082166
smoke	smoke	0.6295590
alco	alco	0.5571791



#### D.) Logistic regression plot per selected variable

