# Capstone Project 2

*Client Report*

## Introduction

Deaths caused by heart disease increased by 3% between 2011 and 2014. Factors which contribute to heart disease include high blood pressure, high blood cholesterol, age, obesity, whether or not a person smokes, and whether or not there is a family history of heart disease. The question which I wish to address in this report is: What features act as indicators of heart disease?

In this study we use logistic regression to examine how features suggest the presence of heart disease. I use logistic regression because it helps to explain the relationship of a variable (the odds of having heart disease) to one or more independent variables (age, weight, etc.).

The original Cleveland study examined the predictive quality of the following indicators:

- Age
- Gender
- Type of chest pain
- Resting pressure of the patient at the time of admission
- Blood cholesterol
- Fasting blood sugar
- Resting electrocardiographic results
- Maximum heart rate achieved
- Exercised induced angina

After reviewing, wrangling, and cleansing the data, I examine instead the predictive quality of the following indicators:

- Number of cigarettes smoked per day
- Number of years that the patient has smoked
- Resting heart rate of the patient
- Resting systolic blood pressure of the patient
- Resting diastolic blood pressure of the patient
- Peak exercise-induced heart rate of the patient
- Peak exercise-induced systolic blood pressure of the patient
- Peak exercise-induced diastolic blood pressure of the patient

I selected the data set from the University of California – Irvine's Machine Learning Repository. The initial dataset contained 899 records with 76 columns. I reduced it, through data cleansing, to just under 400 records with 8 columns. The shape of the data did not require any modification.

I selected these features because I believed that they could be easily measured in a doctor's office and might prove to be a quick test of the presence of heart disease. I am also hopeful that they may offer a different picture when compared.to the results of the other study.

# Capstone Project 2

*Client Report*

## Data Set Cleansing

In order to process the data, I needed to address the presence of multiple null values in numerical fields. After an initial exploratory data analysis, I did the following to cleanse the data:

- I dropped records which matched the following criteria:
  - The *History of Heart Disease* field was null or not either a zero or a one.
  - The *Number of Cigarettes Smoked per Day* field and the *Number of Years Smoked* field were both null.
  - The *Number of Cigarettes Smoked per Day* field was not null but the *Number of Years Smoked* field was null.
  - The *Number of Cigarettes Smoked per Day* field was null but the *Number of Years Smoked* field was not null.
  - The *Is Smoker* field, the *Number of Cigarettes Smoked per Day* field and the *Number of Years Smoking* field were all null.
  - Any of the heart rate or blood pressure fields were either null or zero.

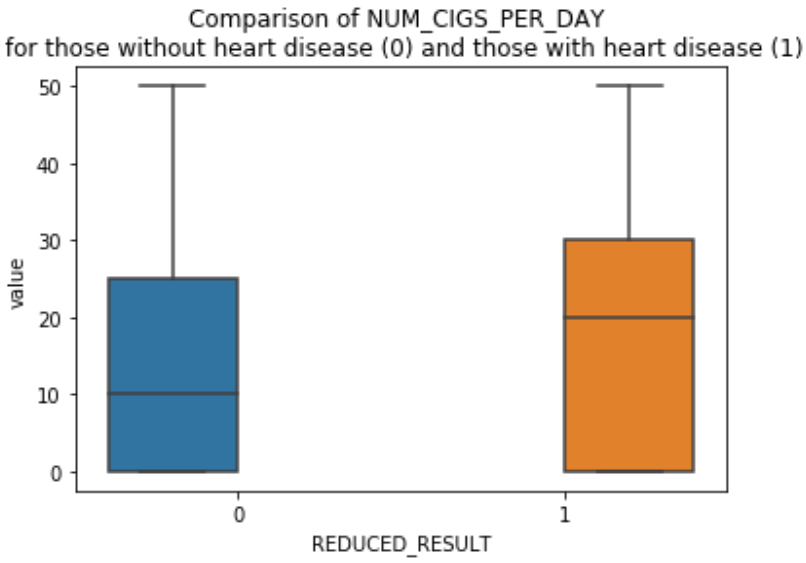
- Then, I checked to see if a record showed that a patient smoked, but did not report the number of cigarettes and/or the number of years which the patient smoked. I removed these records because it was impossible to back-fill the missing information.

Once the data cleanup was complete, I dropped all unnecessary fields.

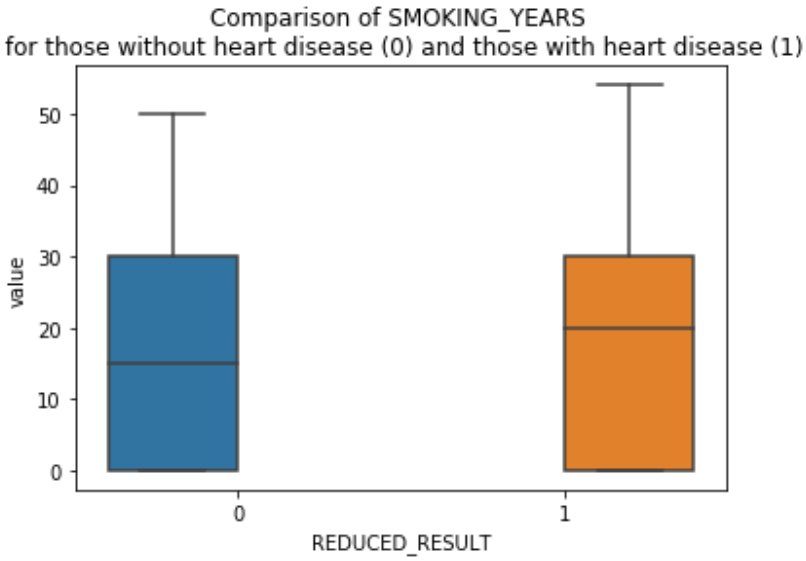
## Exploratory Data Analysis

Here, all plots intend to compare values for those with heart disease and those without heart disease.

Let's begin by examining the plot for the number of cigarettes smoked per day.




Then, let's examine the plot for the years of smoking.

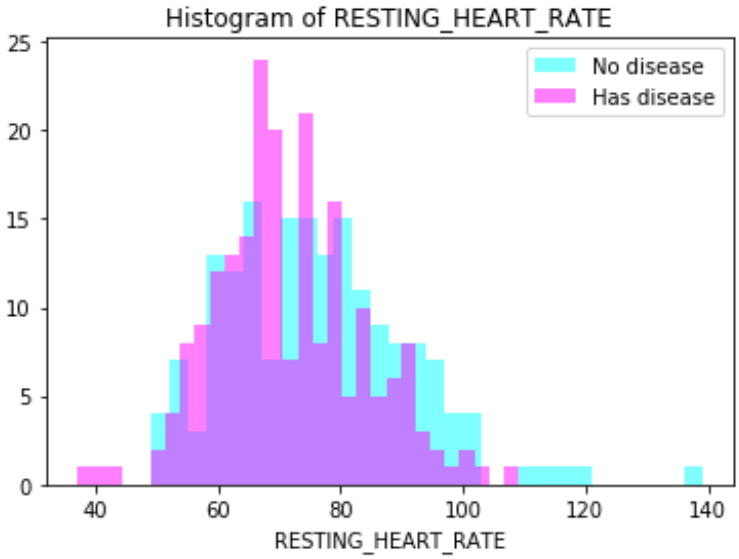


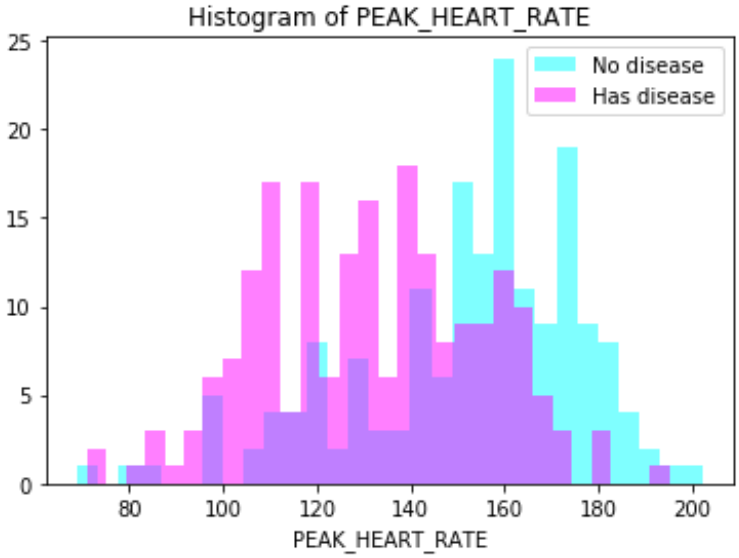

In these plots, we see significant overlap between the samples with heart disease and the samples without heart disease. To determine if the sample differences are real or if they are due to random fluctuations, I performed 10,000 permutation tests on the data, and got the following results:

| Sample set | Difference in means | Margin of error | p-value |
|---|---|---|---|
| Num. of Cigs. / Day | -25.43 | 1.46 | 1.000 |
| Smoking Years | -4.12 | 3.23 | 0.993 |

The large p-values shown above indicate that we should *not* reject a hypothesis that the difference in the sample means is equal to zero. In other words, any differences noted in the data are likely due to random fluctuations and not to actual differences.

Next, let's examine the distributions for resting and peak heart rates.






Again, to verify that distributions are truly different, I performed 10,000 permutation tests on the data. The results are shown here:

| Sample set | Difference in means | Margin of error | p-value |
|---|---|---|---|
| Resting Heart rate | 4.25 | 2.74 | < 0.01 |
| Peak Heart rate | 20.75 | 4.81 | < 0.01 |

Here we find evidence that indicates that we can reject a hypothesis which states that the difference in means of the two distributions is equal to zero. This implies that the resting and peak heart rates for those with heart disease and those without heart disease are truly different. The difference is statistically significant.

## Machine Learning

Now that we have a general idea of what the data looks like, we can use machine learning to build a model that helps us understand which features act as indicators of heart disease. I chose the logistic regression and random forest classifiers because they offer both an easily understood interpretation of the results and they are forgiving of less-than-perfect data.

I began by varying the tuning parameters of each classifier until I achieved each classifier's "best fit.[1]" Since the logistic regression classifier performed better than the random forest classifier, I chose to use it for the model.

This, however, led to an issue. The logistic regression classifier is sensitive to multicollinearity. For the model to make sense, I needed to address whatever multicollinearity existed among the features. To do this, I first checked the variance inflation factors[2] (VIFs) of the individual features. Here are the results from the first data fit:

| VIF | Feature |
|---|---|
| 4.275403 | Number of cigarettes smoked / day |
| 4.632674 | Number of years of smoking |
| 2.463650 | Family history of heart disease |
| 33.882437 | Resting heart rate |
| 99.328243 | Resting systolic blood pressure |
| 126.246305 | Resting diastolic blood pressure |
| 45.193992 | Peak systolic blood pressure |
| 73.596654 | Peak systolic blood pressure |
| 42.255892 | Peak diastolic blood pressure |

To improve the model's usefulness, I needed to eliminate as much of the multicollinearity as possible. So, I iteratively removed the feature with the highest VIF (the resting diastolic blood pressure) from consideration and re-fit the model. Then, I recalculated the features' VIFs to verify that the multicollinearity had been reduced.

[1] "Best fit" is measured by the model's *accuracy*, which the ratio of the number of correct predictions to the total number of predictions.

[2] Variance inflation factors measure how much the variance of the estimated regression coefficients are inflated as compared to when the features are not linearly related. It is used to explain how much multicollinearity (correlation between features) exists in the regression analysis.

I repeated this process, balancing the model's predictive power against the need to minimize the impact of the features' multicollinearity, until I achieved an optimal result. I found the best logistic regression fit includes the following coefficients:

| Feature | Coefficient |
|---|---|
| Number of cigarettes smoked per day | 0.007 |
| Number of Years smoking | 0.006 |
| Peak Heart Rate | -0.033 |
| Peak Diastolic Blood Pressure | 0.009 |

| Performance Metric | Value |
|---|---|
| Accuracy[3] | 0.76 |
| Recall[4] | 0.81 |
| Precision[5] | 0.74 |
| F1 score[6] | 0.77 |

Overall, this model performs with an accuracy of 76%. Noting that the precision is less than the recall implies that, when the model *misclassifies* a sample, it will more likely to classify a patient as having heart disease who does not have heart disease, rather than classify someone who has heart disease as not having heart disease.

Using the expression for logistic regression, we may write the result as:

$$\log\left(\frac{p}{1-p}\right) = 0.007x_1 + 0.006x_2 - 0.033x_3 + 0.009x_4$$

where

p = the probability of having heart disease

$x_1$ = number of cigarettes smoked per day

[3] The metrics show that the model performs with 76% accuracy which means that 76% of all the predictions were accurate.

[4] The recall provides the ratio of correct predictions to the actual number of correct predictions (true positives + false negatives). This may be thought of as a measure of thee predictor's completeness. A low recall may indicate many false negatives.

[5] The precision provides the ratio of correct predictions to the actual number of correct and incorrect predictions (true positives + false positives). Precision may be thought of as a measure of the predictor's exactness. A low precision may indicate a large number of false positives.

[6] The F score conveys the balance between precision and recall. It is defined as $2 * precision * recall/(precision + recall)$.

$x_2$ = number of years that the patient has smoked

$x_3$ = the peak heart rate

$x_4$ = the peak diastolic blood pressure

Or, solving for p, we find:

$$p = \frac{1}{1 + e^{-(0.007x_1 + 0.006x_2 - 0.033x_3 + 0.009x_4)}}$$

If we consider the case where the patient does not smoke, $x_1$ and $x_2$ are both equal to zero. We then have:

$$p = \frac{1}{1 + e^{\,0.033x_3 - 0.009x_4}}$$

If a person's peak heart rate is 120 beats per minute and the person's peak diastolic blood pressure is 80 mm Hg, then the probability that the person has heart disease is found to be:

$$p = \frac{1}{1 + e^{\,0.033*120 - 0.009*80}} = 0.04$$

On the other hand, if a person's peak heart rate is 180 beats per minute and the person's peak diastolic blood pressure remains 80 mm Hg, then the probability that the person has heart disease is found to be:

$$p = \frac{1}{1 + e^{\,0.033*180 - 0.009*80}} = 0.01$$

If we now consider the case where the patient smokes 10 cigarettes per day and has smoked for 10 years, we find, for a peak heart rate of 120 beats per minute and a peak diastolic blood pressure is 80 mm Hg, the probability that the person has heart disease to be:

$$p = \frac{1}{1 + e^{-(0.007*10 + 0.006*10 - 0.033*120 - 0.009*80}} = 0.04$$

And, if we consider the case where the patient smokes 10 cigarettes per day and has smoked for 20 years, we find, for a peak heart rate of 120 beats per minute and a peak diastolic blood pressure is 80 mm Hg, the probability that the person has heart disease to be:

$$p = \frac{1}{1 + e^{-(0.007*10 + 0.006*20 - 0.033*120 - 0.009*80}} = 0.05$$

## Conclusion

I analyzed a data set from the University of California – Irvine's Machine Learning Repository which contained heart disease data. I examined the data set to see if other features play a predictive role in identifying heart disease.

Using exploratory data analysis, I discovered that, surprisingly, there are no statistically significant differences between the number of cigarettes smoked per day and having heart disease or between the

number of years a person smoked and having heart disease. We also found that there were statistically significant differences between the resting heart rate and peak heart rate of those patients who have heart disease and those patients who do not.

Then, we used logistic regression to model the relationship between the odds of having heart disease and number of cigarettes smoked per day, the number of years of smoking, the peak heart rate, and the peak diastolic blood pressure.

Based on these results, I would recommend that a practitioner note not only the well-known predictors of heart disease, but also pay close attention to the peak cardiac functions. Of these, the peak heart rate has the greatest impact as a predictor of heart disease.

Also, we note that, where this model does have some explanatory power, it is still limited. There are, no doubt, many more factors which can be explored. For example, I would explore the interactions of the various predictors (i.e. how does peak heart rate influence peak diastolic blood pressure?).