COMPANY

Heart-Double Associates

An analytics consulting firm with talent and expertise in big data analytics of health data.

OVERARCHING GOAL

The Heart-Double Associates data science team has created predictive models and developed insights to identify an individual's risk for health problems (e.g. high cholesterol, heart disease, etc.) based on key user information. These predictive models will help to quantitatively inform users to change lifestyle choices or consult a doctor before health problems become real.

Possible Clients: Company specializing in smarter exercise equipment (e.g. stationary bike), smartphone application developers, health-insurance companies, hospitals, etc.

Data sources

IHIS
UCI Heart Disease Data 1, Data 2

MEET OUR TEAM

Health gurus and health specialties

- ***Praveen:*** health and effects of sleep, heart attacks and occupation
- ***Rajini***: body-mass index (BMI) and effects of diet, health, and work habits
- ***Aaron:*** heart disease and effects of smoking and cholesterol
- ***Dara:*** cholesterol and effects of yoga and vitamins
- ***Eric:*** heart disease and effects of gender, age, cholesterol, max heart rate during exercise, and menopause.

PREFACE

The following chapters detail and illustrate the importance of collecting specific user information in order to better predict health problems such as heart disease or high cholesterol.

-Heart Double Associates

TABLE OF CONTENTS

Executive Summary

Sleep and Health

In order to identify behavioral features that can be added into a predictive health model, we look at sleep's relation to overall health. We observe that individuals who sleep between 7 and 9 hours on average have a lower risk profile when looking at other health-related variables.

Occupation and Heart Attacks

We chart the relative rates of heart attacks within individuals in each occupational group, ranging from the riskiest (Construction and Extraction) to least risky (Computer and Mathematical).

BMI

Developed a model to predict on an individual's body mass index (BMI) is affected by his/her daily eating and drinking habits. The developed model can be built into an app. that specifically calculates how the user's BMI is affected by his/her age, alcohol consumption, diet (e.g. sweets and vegetables), working hours, and travel habits.

Cholesterol, Yoga, and Vitamins

We show that an individual's use of yoga and vitamins has a noticeable effect on our ability to predict cholesterol levels.

Heart disease and Smoking

We built a model linking heart disease to smoking. While this model provides some predictive insight, the data from which it was derived severely restricts its predictive power. The graphs demonstrate this fact in a convincing manner.

Heart disease between genders: age and cholesterol

Men aged 40 and up are at an increased risk for heart disease compared to women of the same age. This statistically significant effect does not appear to be the result of increased cholesterol levels.

Heart disease risk model: effects of gender, age, cholesterol, and heart rate during exercise

We built a model that predicted an individual's risk for heart disease based on his/her gender and age, cholesterol level, and maximum heart rate attained during exercise. Specifically the model predicted that being older, male, having high cholesterol, and reaching a low maximum heart rate during exercise increased the likelihood of heart disease.

Heart Disease and Menopause

40-42 year old women with menopausal symptoms have a significantly higher likelihood of heart disease compared to 40-42 year old women who have never experienced menopausal symptoms. Future work can predict a woman's likelihood of having menopause (if not available) from other key information to see if it can better predict a woman's risk of heart disease.

Sleep and Health

We considered sleep as one of our behavioral features, and its relationship with various health risk-factors. To do this, we used survey data where individuals reported how many hours they slept per night. After some exploration, it became clear that sleep within the normal range [7-9 hours] is related to other features which indicate general health. However sleep on either extreme, is positively correlated with riskier health indicators:

Hours of Sleep vs. Heart Attacks

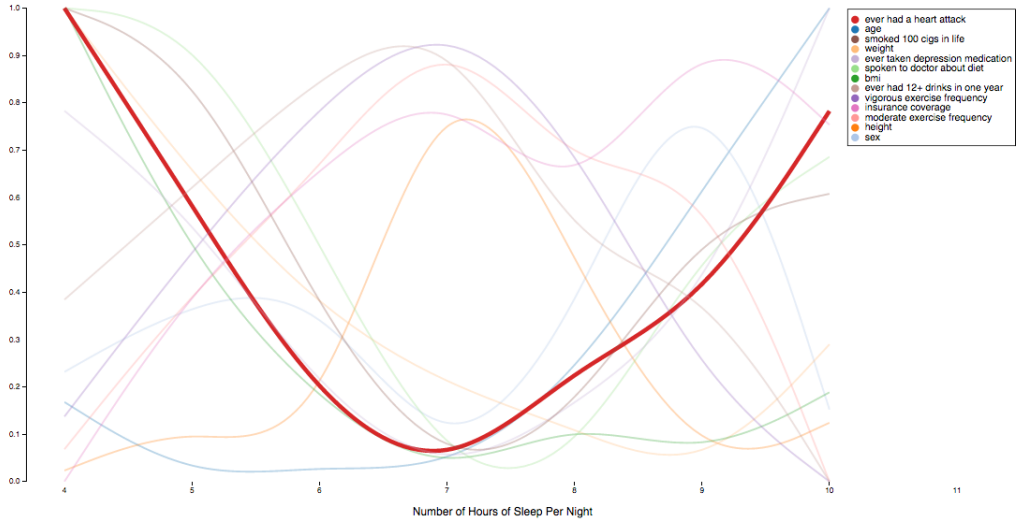

FIGURE 1.

Hours of Sleep vs. Regular Exercise

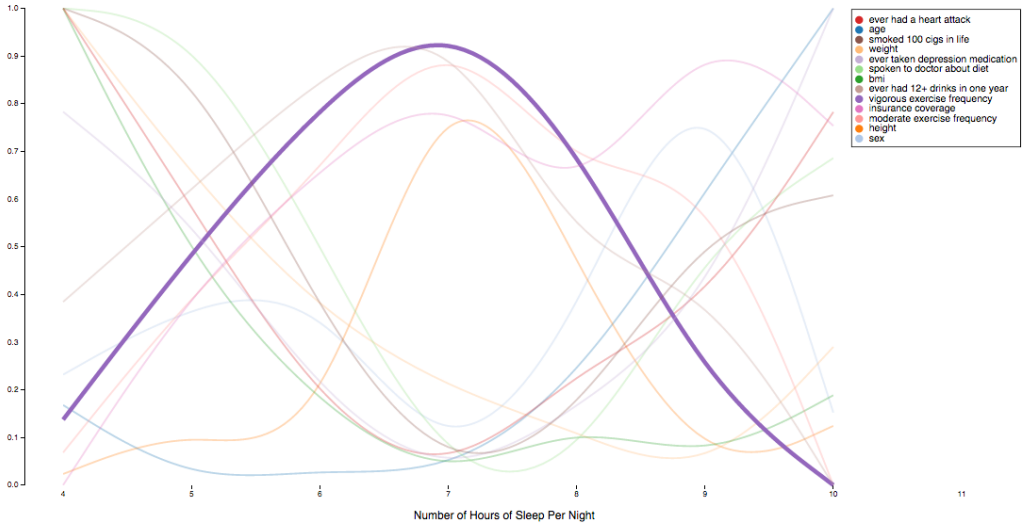

FIGURE 2.

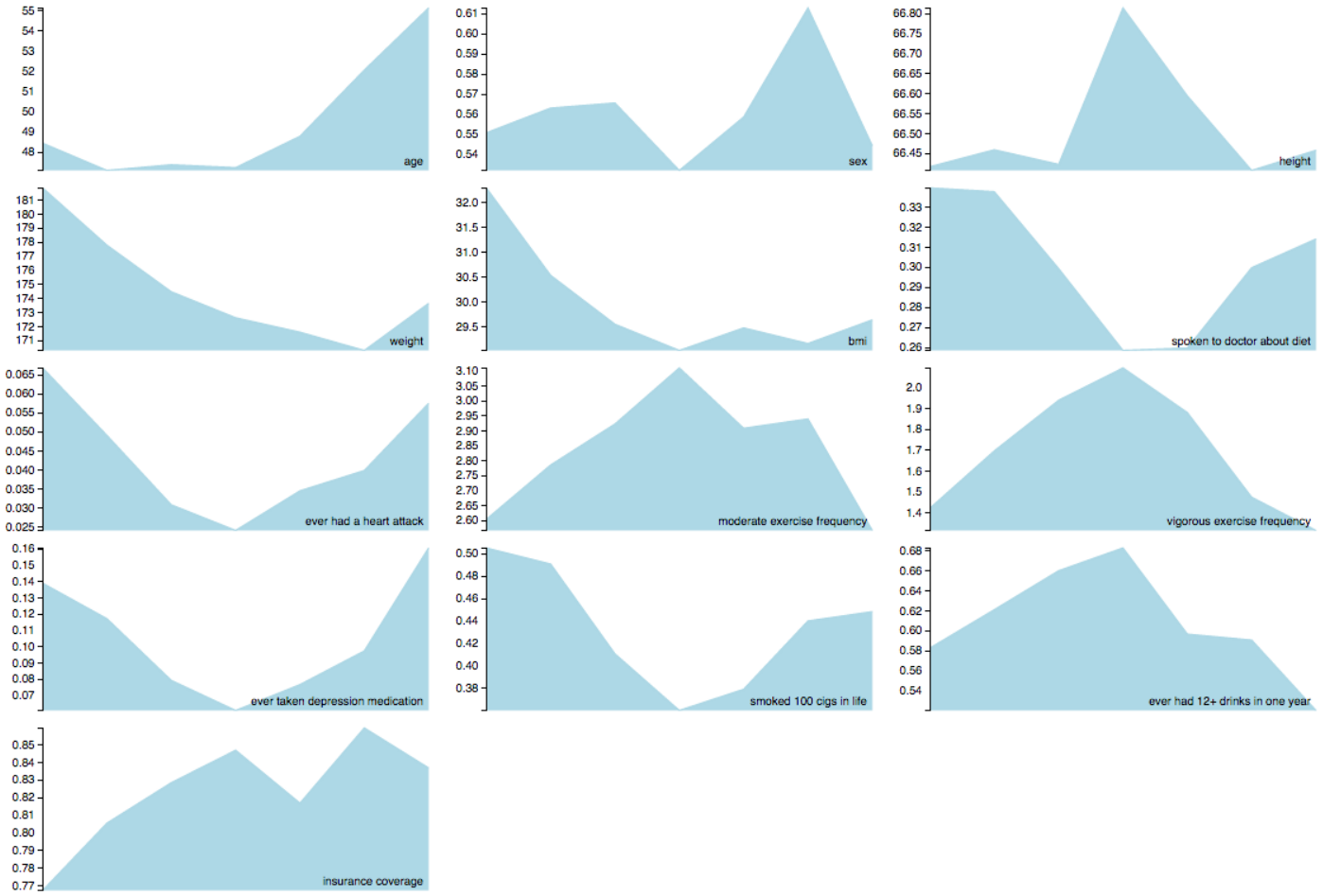

FIGURE 3.

Quantity of sleep on its own cannot function as a predictor of health, particularly heart health, but could enhance the quality of the models detailed in selections below, as one of many features.

Additionally, we built a model that attempted to predict whether an individual was getting a “healthy” amount of sleep (7-9 hours) based solely on Sex, Age, body mass index (BMI), and Occupation. The model accurately classified about half of the population, with 51% of individuals classified as unhealthy sleepers, actually being one (measured by precision). The recall shows what percentage of each group was correctly classified.

Healthy Sleeper	Precision	Recall
No	0.51	0.43
Yes	0.48	0.56

TABLE 1.

Occupation and Heart Health

Similar to one's sleeping behavior, occupation has notable relationship with heart health. Occupation would work well as an addition to models with more heart health-specific features. Specifically, individuals working in Construction and Extraction have the highest heart attack rates and individuals in Computer and Mathematical fields have the lowest heart attack rates.

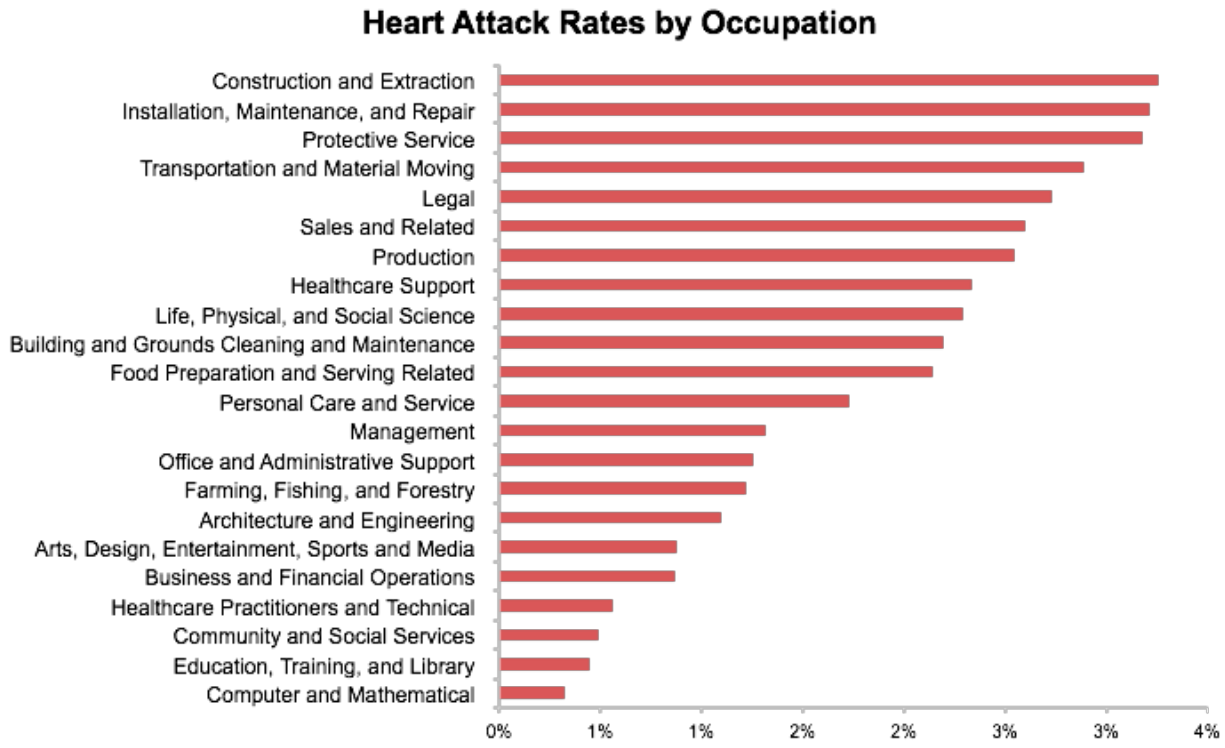

FIGURE 4.

Predicting BMI

Data Used is from year 2010
Number of samples collected : 84,000
After Clean up, samples used for model prediction : 2424
Features Started with: 169
Features after clean up and Elimination : 105
Actual features used for Model selection around 60

IHIS data analysis was analyzed to predict heart disease using BMI and other obesity related information. The poor quality of the data resulted in very poor predictions (99% accuracy but recall very close to zero) hence the focus was shifted to predicting BMI. The idea is to predict the BMI based on daily health, diet, and work activities such as food intake, drinking habits, work environment, etc.

Logistic regression model was used to analyze the features and eliminate the features which did not contribute toward predicting a user's BMI. A significant number of features were eliminated before finally arriving at the features of highest significance toward this prediction. The bubble chart below shows the contribution of each feature towards BMI and includes features which correlate with an increase in BMI and features which correlate with a decrease the BMI. Diameter of each circle will gives the relative contribution of each feature toward predicting a user's BMI (the larger the diameter, the larger the impact).

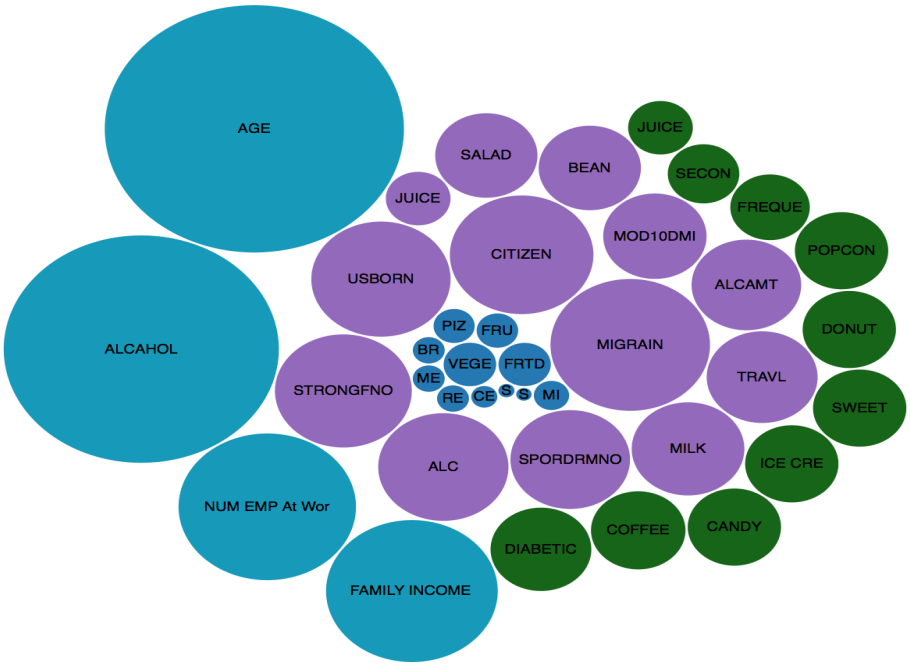

FIGURE 5.

Figure 6 below describes the difference among the precision and accuracy values among all the models used for prediction.

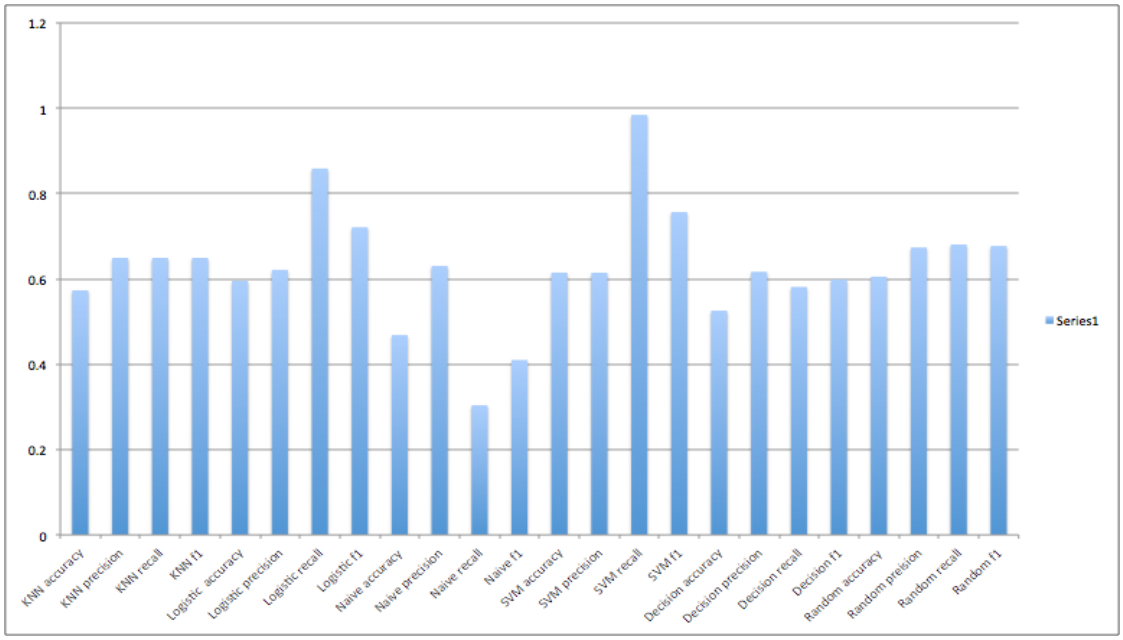

FIGURE 6.

Cholesterol, Yoga, and Vitamins

We also analyzed the effect of doing yoga and/or taking vitamins and supplements on cholesterol levels. In this analysis, the features used were binary (did the patient do yoga in the past year, and did the patient take vitamins or supplements in the past year), as was the target variable (did the patient have high cholesterol in the past year).

There was a total of 415 patients included in this data set (collected in 2007), with the following breakdowns on their yes/no answers:

Total Yoga and Vitamins/Supplements Breakdown:

	Yoga	**Vitamins/Supplements**
No	220	36
Yes	195	379

TABLE 2.

Table 2 (above) breakdown is displayed in Figure 7 below:

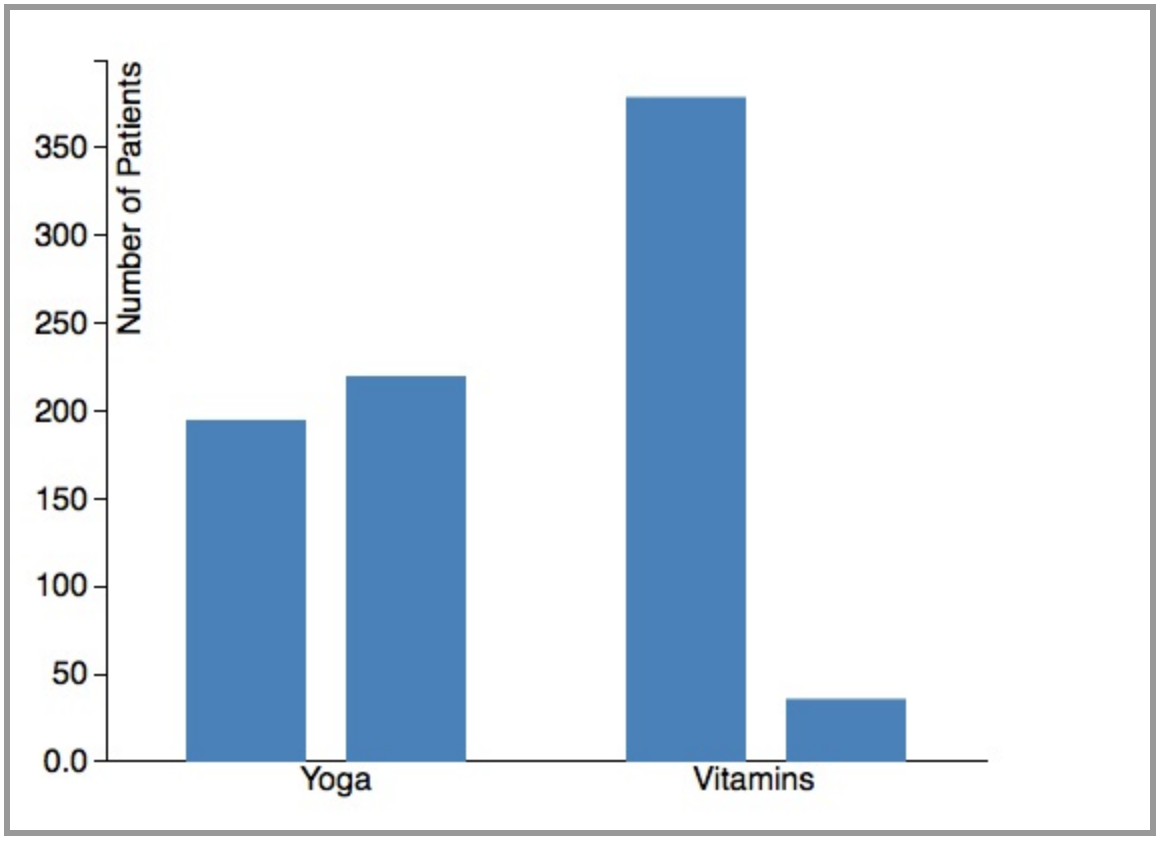

FIGURE 7.

Total Yoga and Vitamins/Supplements Breakdown (Table 3):

	Yoga: Yes	**Yoga: No**
Vitamins: Yes	183	196
Vitamins: No	12	24

TABLE 3.

Using a K-Nearest Neighbors analysis (with 4 notes), we obtained the following precision and recall rates (Table 4 below):

High Cholesterol	**Precision**	**Recall**
No	0.37	0.55
Yes	0.73	0.56

TABLE 4.

This means that of all our predictions for patients **without** high cholesterol in the past year, we were accurate 37% of the time, and for patients **with** high cholesterol, we were 73% accurate. Furthermore, we predicted 55% of the patients without high cholesterol as not having high cholesterol and predicted 56% of the patients with high cholesterol as having high cholesterol in the past year.

Heart disease and Effects of Smoking and Cholesterol

We attempted to use daily cigarette use and resting blood pressure to predict incidence of heart disease. We took a sample of 400 individuals and ran an SVM classifier. The first graph (Figure 8) implements the classifier using scaled values with Matplotlib. Unfortunately, its predictive power is less than ideal. The second graph (Figure 9) reproduces that data in an unscaled manner using D3. The third chart (Figure 10) represents the accuracy of the model. Over repeated simulations, the accuracy varied between 55%-60%, only slightly better than would be achieved by chance. The third graph plots the incidence of heart disease with resting blood pressure and cigarette use in three dimensions. That plot was created using the R package, scatterplot3d.

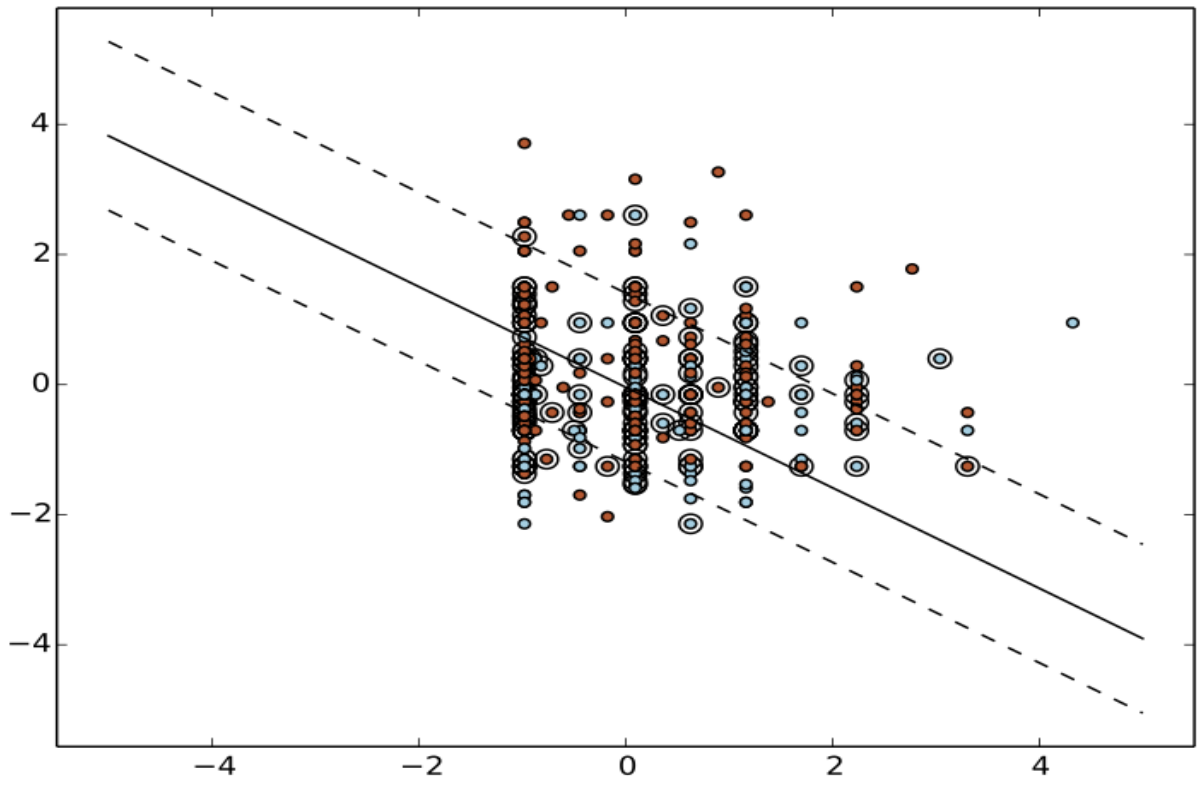

FIGURE 8.

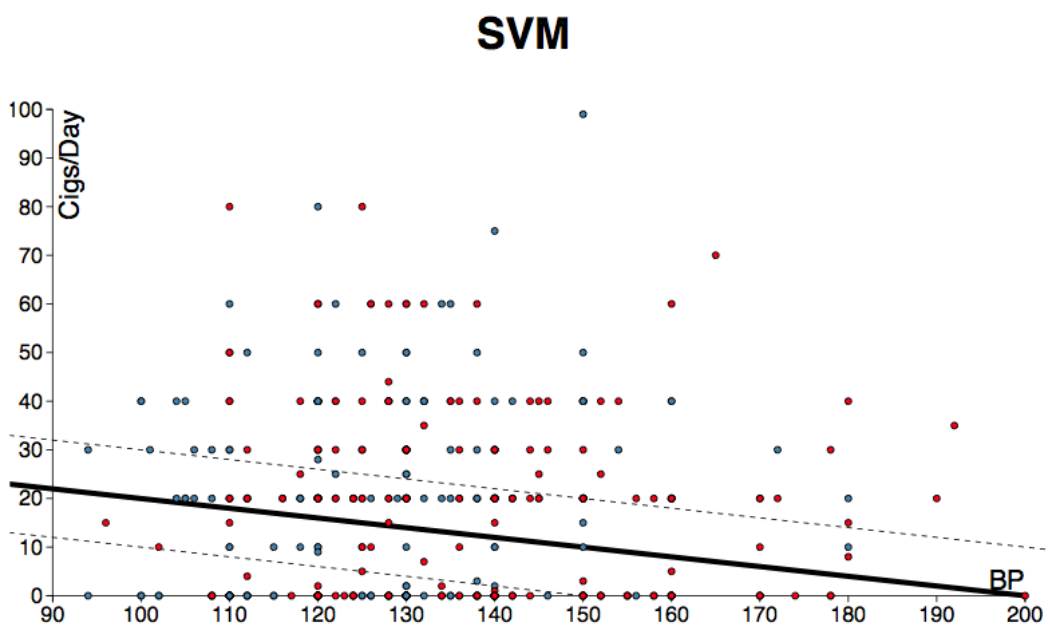

FIGURE 9.

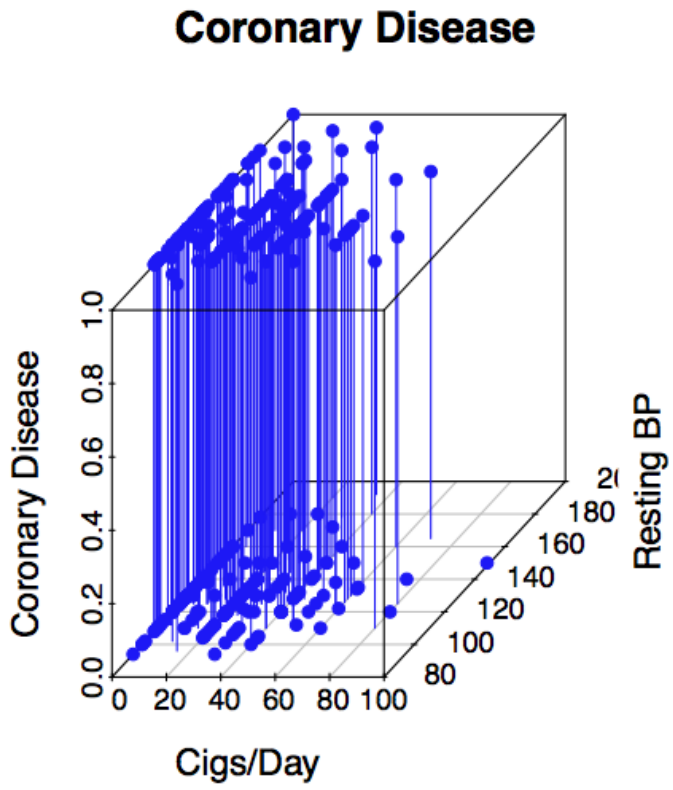

FIGURE 10.

Heart Disease Between Genders: Age and Cholesterol

Using IHIS data from 2000-2013, men aged 40+ were found to be at a significantly (*t-test, p*<0.05) increased risk for heart disease compared to women of the same age (see Figures 11 and 12).

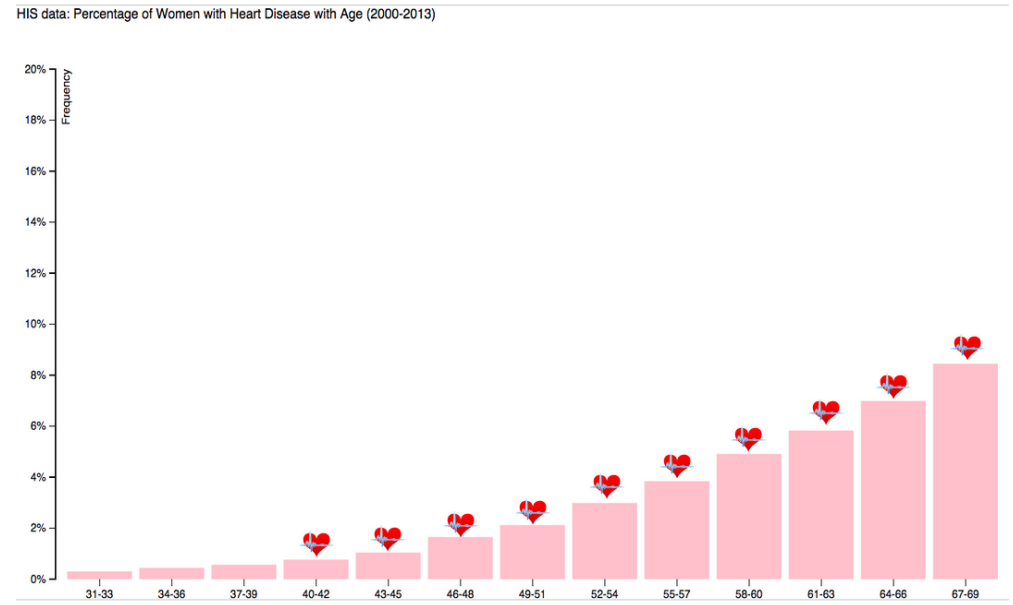

FIGURE 11.

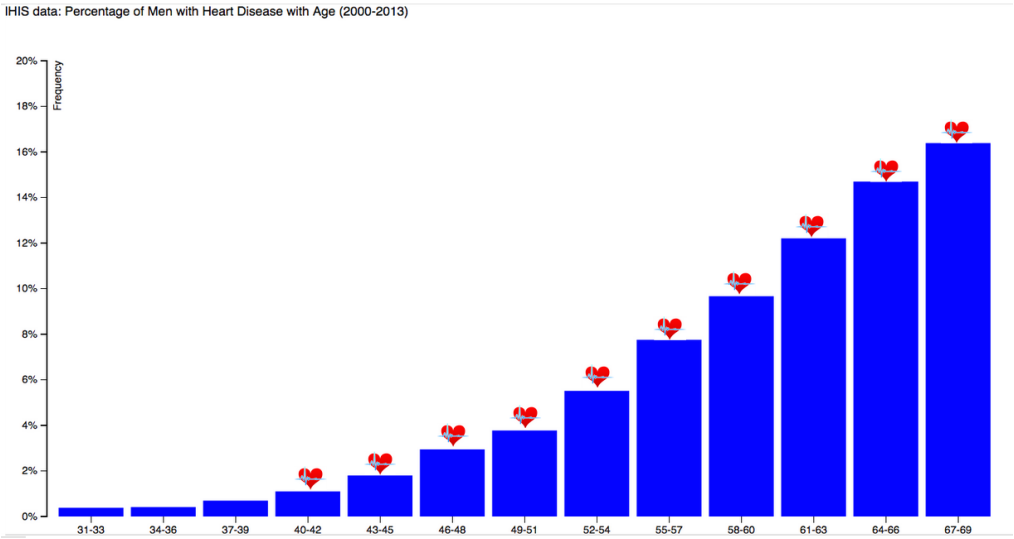

FIGURE 12.

Assuming cholesterol level has a positive relation to risk of heart disease, the increased risk of older men compared to older women does not appear to be the result of increased cholesterol levels in older men (see Figures 13 and 14). The percentage of women with high cholesterol in the past year is statistically higher ($p<0.05$) between 40-42 years of age as well as above 57 years of age compared to their male counterparts. Unless cholesterol has a negative relation with heart disease, it appears that the increased risk of heart disease from being an older man is largely independent of cholesterol.

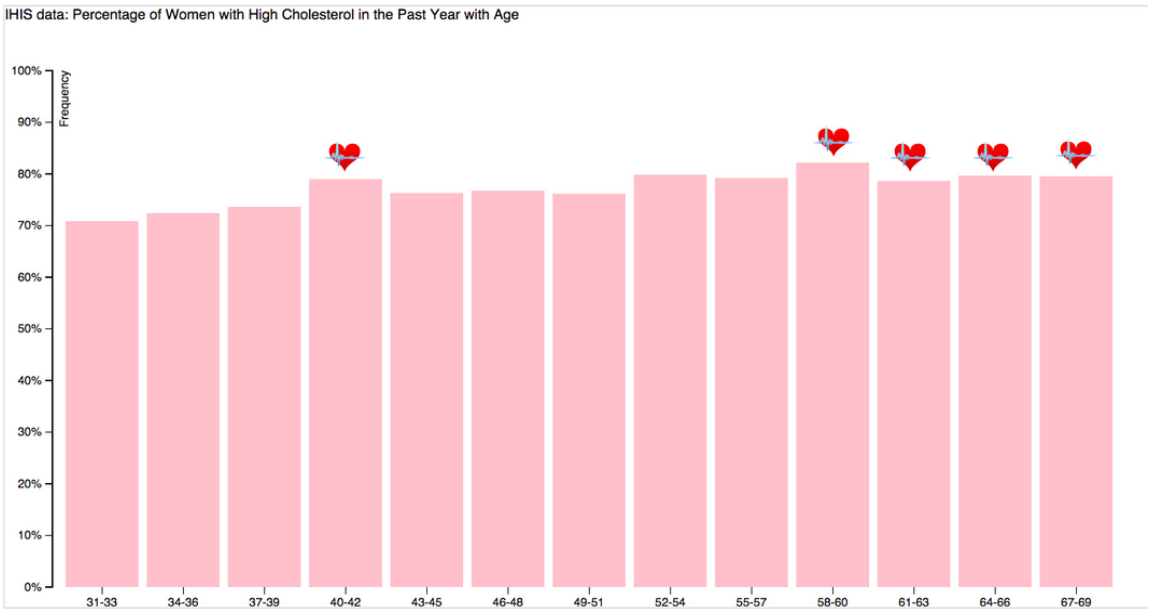

FIGURE 13.

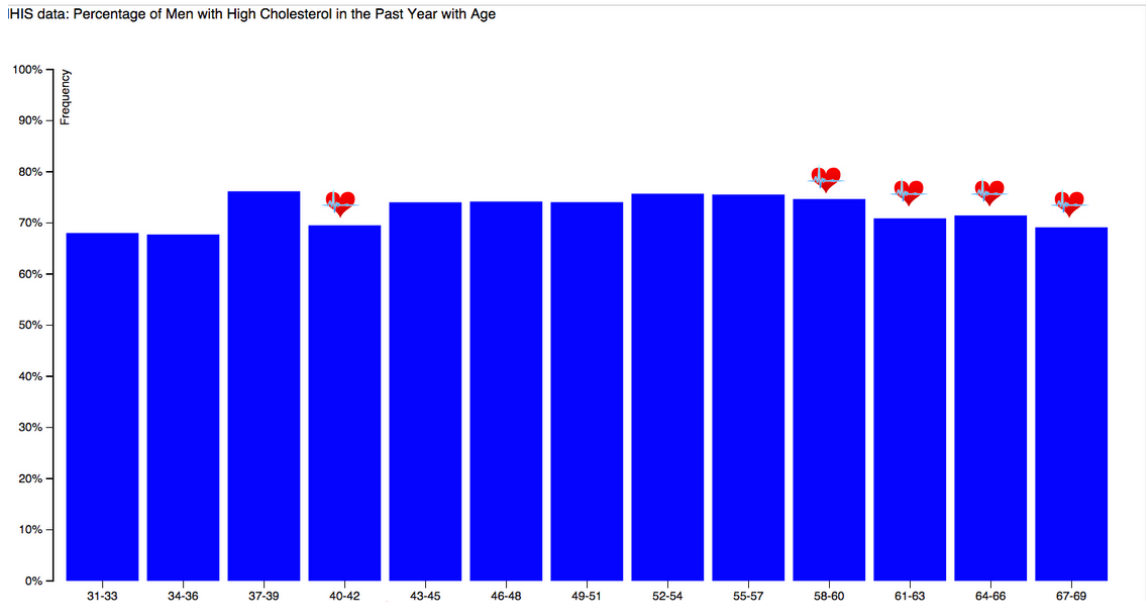

FIGURE 14.

Heart disease risk model

Using the previous insight of the significant combined effect of age and gender to model the UCI datasets, we built a logistic regression model that predicted an individual's risk for heart disease (P(heartDisease)) using three highly significant ($p<0.001$) features:

1. age*gender
2. cholesterol level [cholesterol]
3. maximum heart rate achieved during exercise [max exercise HR]

$$P(heartDisease) = \frac{1}{1+e^{-(b_0+b_{age*gender}X_{age*gender}+b_{cholesterol}X_{cholesterol}+b_{\max exercise HR}X_{\max exercise HR})}}$$

Of the models tested, this model had the highest prediction accuracy (74%) with a precision and recall of 70% and 68%, respectively. More generally, this model predicts that being an older male, having high cholesterol, and achieving a low maximum heart rate during exercise increases the likelihood of heart disease.

As quantitative examples of this model, if a 42 year old man who achieves 142 max beats per minute (bpm) during exercise reduces his cholesterol level from 250 mg/dL to 240 mg/dL (keeping all other features constant), he will have reduced his risk for heart disease by 1.3%. Compared to a man, a woman with these exact same stats will have a 22.4% reduced risk of heart disease. And finally, if this woman increases her max heart rate during exercise from 142 bpm to 152 bpm (keeping all other features constant), she will have reduced her risk for heart disease by 3.6%.

Heart Disease and Menopause

We used 1994 and 1998 IHIS data to determine the relationship of a woman's risk for heart disease with her menopausal status. As shown in Figure 16 and 17, we determined that 40-42 year old women with menopausal symptoms have a significantly (t-test, $p<0.05$) higher likelihood of heart disease compared to 40-42 year old women who have never experienced menopausal symptoms. Thus it appears that a younger woman's likelihood of heart disease may be increased if she has menopause. Future work can predict a woman's likelihood of having menopause (if status is unknown) from other key information such as smoking, diabetes status, age, and other factors. This can then be incorporated in the heart disease model to determine whether it can more accurately predict a woman's risk of heart disease.

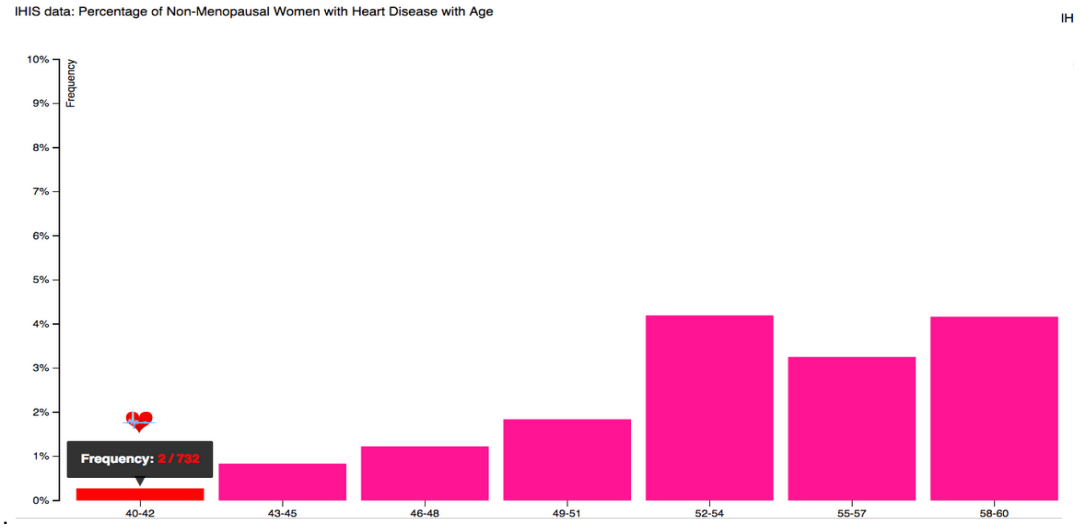

FIGURE 15.

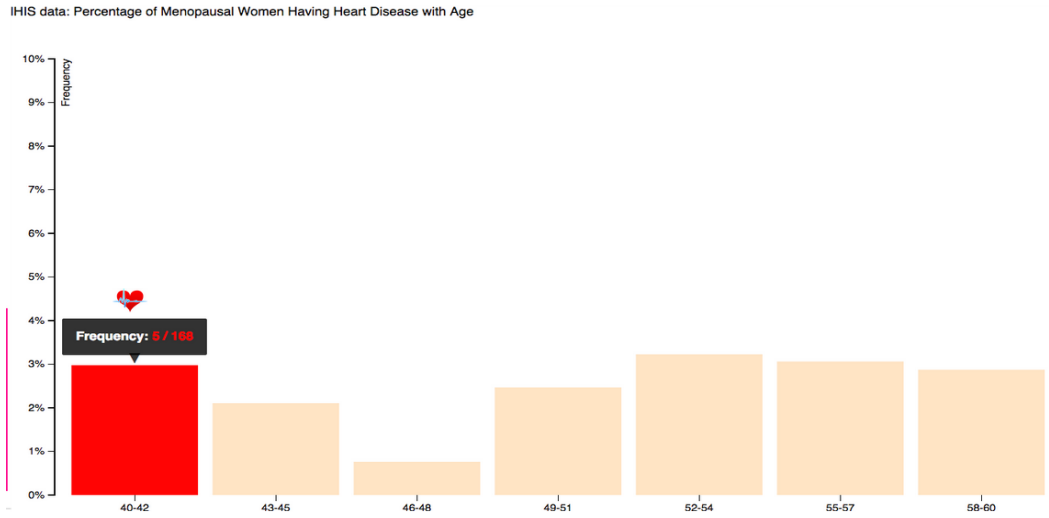

FIGURE 16.