Heart Disease Predictor for Health Insurance Sales Agents

Cheng Yu Chen, Jiayu Zhang, Rohit Soans, Vinni Guan
chen4154@purdue.edu; zhan4358@purdue.edu, rsoans@purdue.edu , guan64@purdue.edu ;

Abstract

We have built a Shiny app to predict the probability of heart disease in a given individual. We believe that this app will be an asset in the hands of sales teams for Health Insurance Companies. Sales agents can use the app when they are pitching products that cover heart disease. On hearing what is the probability that they may get heart disease, most people are likely to be garner strong interest in Health Insurance products. We found a Cardiovascular dataset on Kaggle with 70,000 observations. We explored and transformed our data and build a model to predict the probability of heart disease.
Link to the app :https://chengyuchen.shinyapps.io/Team3Project/
YouTube : https://www.youtube.com/watch?v=bvZbR4dr_as
GitHub: https://github.com/zhangjiayuyu/R-Shiny-CardiovascularDisease
Shiny Template referenced from: https://mogali.shinyapps.io/attritionanalysis/

Business Problem

The UnitedHealthcare Consumer Sentiment Survey showed that only 9% of Americans surveyed "showed an understanding" of four basic health insurance terms — health plan premium, health plan deductible, out-of-pocket maximum and co-insurance. Due to the lack of understanding of Insurance Products, most people are less likely to switch their scheme and end up simply renewing their scheme with their current provider, even though those products maybe mispriced or unsuitable. This is one of the key challenges sales teams of health insurance companies face. Hence new customer acquisition in this industry is both difficult and expensive.

Analytics Problem

To calculate the predicted probability of heart disease based on key lifestyle indicators such as Age, Height, Weight, Cholesterol, Systolic blood pressure, Diastolic blood pressure, Glucose, Smoking, Alcohol and Physical activity.

Data

The dataset consists of 70,000 records of patient's data in 12 features, such as age, gender, systolic blood pressure, diastolic blood pressure, and etc. The target class "cardio" equals to 1, when patient has cardiovascular disease, and if it is 0, the patient does not have heart disease.

Data Dictionary			
Feature	**Variable Type**	**Variable**	**Value Type**
Age	Objective Feature	age	int (days)
Height	Objective Feature	height	int (cm)
Weight	Objective Feature	weight	float (kg)
Gender	Objective Feature	gender	categorical code
Systolic blood pressure	Examination Feature	ap_hi	int
Diastolic blood pressure	Examination Feature	ap_lo	int

Cholesterol	Examination Feature	cholesterol	1: normal, 2: above normal, 3: well above normal
Glucose	Examination Feature	gluc	1: normal, 2: above normal, 3: well above normal
Smoking	Subjective Feature	smoke	binary
Alcohol intake	Subjective Feature	alco	binary
Physical activity	Subjective Feature	active	binary
Presence or absence of cardiovascular disease	Target Variable	cardio	binary

Methodology Selection

The data was examined for missing values and outliers. We further transformed the data using min-max normalization so that there was no single predictor that dominated our results. We looked for variables that had a strong correlation (>0.85) with each other and dropped one of the said variables. We looked for variables that had very little variation as this would not help us explain he overall variance of our model. We have utilized descriptive analysis to understand the distributions of our predictors (lifestyle indicators) as well as our target variable (whether a person gets heart disease or not). We utilized predictive analytics to predict the probability of a heart disease based on given key lifestyle indicators.

Model Building

For our analysis, we used a Logistic Regression model for classification, which provided the greatest accuracy among other models tested. Another benefit of the logistic regression model is its interpretability. Clients would be able to understand the impact of each predictor/lifestyle indicator on their cardiovascular health.

Functionality

The sales agent can input the lifestyle indicators such as Age, Height, Weight, Cholesterol, Systolic blood pressure, Diastolic blood pressure, Glucose, Smoking, Alcohol and Physical activity. Based on these parameters, the app will predict the probability of heart disease for that individual. The user can also explore various visuals such as boxplots plotting the numeric variables against the target variable (1-heart disease 0- no heart disease) as well as bar plots of categorical variables. These plots will help the user understand the distribution of these variables within the existing dataset. If we had more time and experience, we would have liked to build an optimization model. By using the probability of heart disease as an objective function, we could guide a given client on how he or she can reduce their chances of heart disease below a certain threshold by reducing certain parameters by a defined amount.

GUI Design and Functionality