Machine Learning:

Using the Logistic Regression Model to Predict Coronary Heart Disease

Wright Analytics

Introduction

- Logistic regression is a ***parametric*** learning method that can be used to used to model a ***qualitative*** variable with one or more ***continuous*** independent or ***indicator*** variables. Moreover, the dependent variable contains ***two levels***. Since this is the case, the levels are often represented with binary coding. Ones and zeroes

- So Y = 1 or Y = 0 when the qualitative levels are given a binary coding

Introduction

- The occurrence of Y = 1 usually represents the event of ***success*** in some process/phenomena the modeler is studying
- On the other hand, the occurrence of Y = 0 usually represents the event of ***failure*** in the same process/phenomena the modeler studying
- One can think of Y = 1 as a “success” response and Y = 0 as a “failure” response

Introduction to Logistic Regression

- The logistic regression model is ***similar*** to the linear regression
- The ***difference*** between models is mainly the modeling objective.
- Linear regression is used to model an ***unbounded*** continuous dependent variable whereas the logistic regression model attempts to model a dependent variable that contains categories

Introduction to Logistic Regression

- The dependent variable in linear regression is unbounded because it not constrained to some condition. It can take on any ***real value*****.**
- When the dependent variable is qualitative, or Y is coded in binary, the objective is to ***estimate*** the probability of Y = 1
- Probability always lies in the interval (0,1)

Introduction to Logistic Regression

- Since probability always lies in the interval (0,1), one can say that probability ***constrained*** to the interval
- In fact, the logistic regression model outputs ***conditional probabilities***
- The idea of conditional probability is like the ***conditional mean*** that we explain in our “Machine Learning: Learning from Data Using Linear Regression” presentation

Introduction to Logistic Regression

- The ***logit*** form of the logistic regression model is

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 * X$$

- The ***logit*** form of the multinomial logistic regression model is

$$\ln(\frac{p}{1-p}) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta_n * X_n$$

Introduction to Logistic Regression

- To obtain the probability form of the model, we exponentiate both equations and solve for ***p***

- The ***probability*** form of the logistic regression model is

$$p = \frac{e^{\beta_0 + \beta_1 * X}}{1 + e^{\beta_0 + \beta_1 * X}}$$

Introduction to Logistic Regression

- The ***probability*** form of the multinomial logistic regression model is

$$p = \frac{e^{\beta_0+\beta_1*X_1+\beta_2*X_2+\ldots+\beta_n*X_n}}{1+e^{\beta_0+\beta_1*X_1+\beta_2*X_2+\ldots+\beta_n*X_n}}$$

- We predict Y = 1 by using the probability form of the model. ***Remember, the probability form of the model outputs conditional probabilities***

Introduction to Logistic Regression

- To be brief, conditional probability is written as $\Pr(\mathrm{Y} \mid \mathrm{X})$
- Where X and Y are two different events where X ***has already occurred, or been observed,*** and Y ***has not yet occurred***
- We use conditional probability to answer questions like, “what is the probability Y occurs given we already know that X happened?”

Introduction to Logistic Regression

- Again, we use the probability form of the logistic regression model to determine whether Y = 1 or Y = 0 based on the values of the independent variable(s)

- If ***p > c*** when we use the model to estimate a probability for an observation X, where c is just a constant in the interval (0,1), then we conclude that Y = 1. If ***p*** is not greater than ***c***, then observation X's Y label is zero

- The number ***c*** is called a ***threshold***

Brief Summary of Logistic Regression

- The logistic regression model is a ***parametric*** learning method that assumes linearity between the qualitative variable, often coded in binary, and the independent variable(s)

- There are two forms of the model: logit form and probability form

- The logistic regression model outputs conditional probabilities based on the value of X and are used to label an observation X as either Y = 0 or Y = 1

Project Results: Overview

- In our project “Machine Learning: Using the Logistic Regression Model to Predict Coronary Heart Disease”, we analyze data from the Framingham Heart Study

- The data set we obtained contains medical information from Framingham, Massachusetts residents who participated in the ongoing study

Project Results: Overview

- The data set we obtained contained various variables that numerically represent major risk factors for ***absolute short-term coronary heart disease (CHD)***

- The binary dependent variable we modeled is ***Risk***. ***Risk = 1*** indicates that the patient is at risk of developing fatal or nonfatal CHD within ten years or less. ***Risk = 0*** if the patient is not at risk

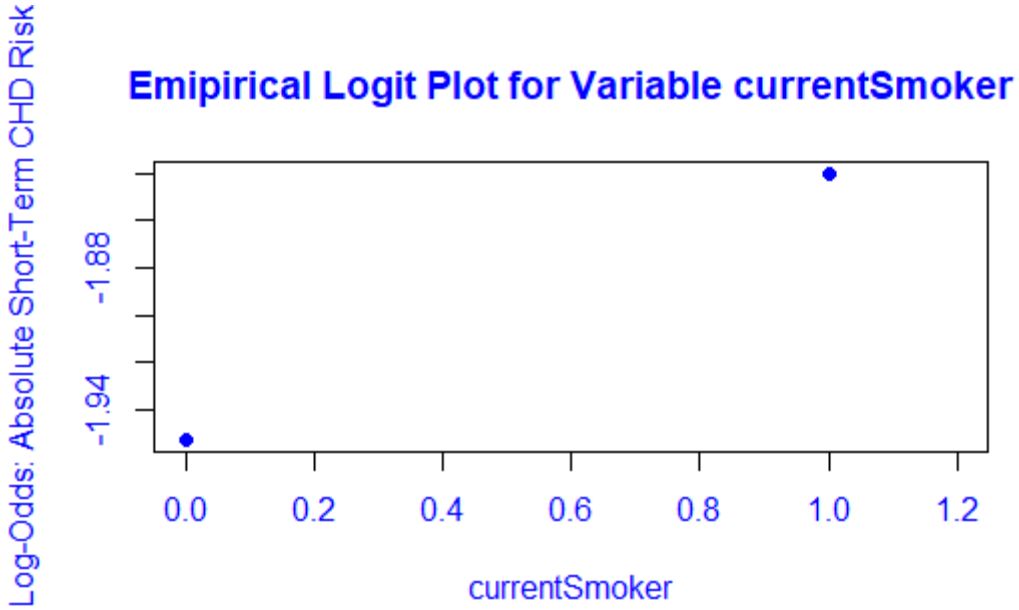
Project Results: Empirical Logit Plots

- ***Empirical logit plots*** are visualization tools that can be used to assess the linearity condition between the dependent and ***continuous*** independent variable(s). Additionally, dependent variable and independent ***indicator*** variables that represent categories in ***binary***

- ***It is not required to assess linearity for indicator variables because it is automatic***. However, they can show us the ***direction*** and ***magnitude*** between the dependent variable and indicator variable

Project Results: Empirical Logit Plots

- The variable ***currentSmoker*** indicates whether the patient was a current smoker at the time of their doctor visit or not. ***currentSmoker = 1*** if yes. ***currentSmoker = 0*** if not



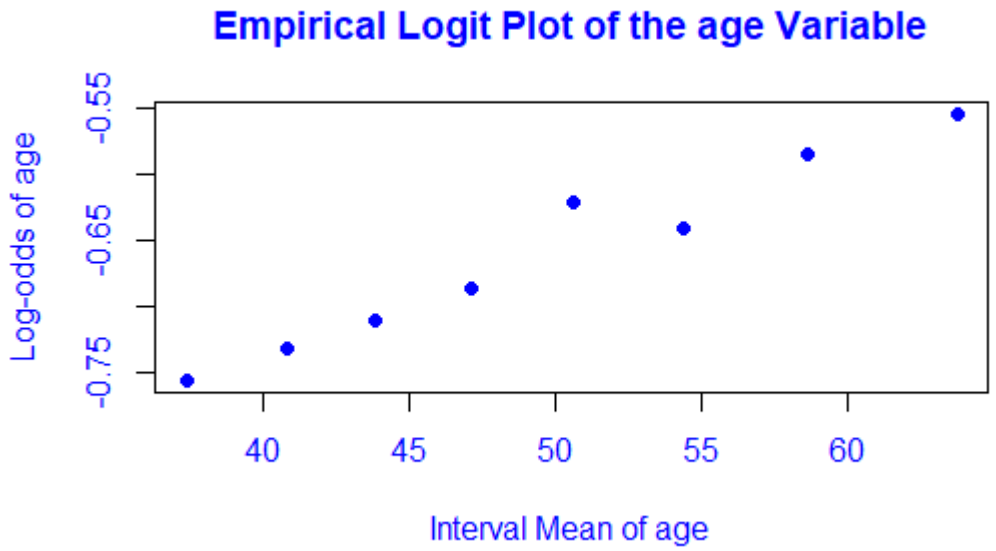
Project Results: Empirical Logit Plots

- This empirical logit plot shows that smokers may have a higher risk of developing fatal or nonfatal CHD within the next ten years as compared to non-smokers

- It is important to note that we ***cannot*** conclude that smoking ***causes*** an increased risk of the condition in people based on the plot. We can only conclude that the relationship between the logit form of the model is linear and ***positively associated*** with the act of smoking

Project Results: Empirical Logit Plots

- The variable ***age*** is a continuous variable that contains the ages of the patients in the study.



Project Results: Empirical Logit Plots

- The plot shows that advancing age may be a risk factor of developing fatal or nonfatal CHD within the next ten years

- Again, we cannot conclude that advancing age causes an increased risk of developing the condition in people based on the plot. We can only conclude that the relationship between the logit form of the model is linear and ***positively associated*** with growing older

Project Results: Major CHD Risk Factors

- In our research we found that ***advancing age, sex-specific advancing age, certain levels of body mass index (BMI), smoking, diabetes, and certain levels of systolic blood pressure*** are thought to be major risk factors that contribute to development of coronary heart disease in patients.

- Based on our research and intuition, we built a logistic regression model where the risk factors listed above were used as independent variables in the model.

Project Results: The Independent Variables

- X.1 indicates the sex of the patient. X.1 = 1 if the patient is male.

  X.1 = 0 if the patient is female.
- X.2 and is the integer age of the patient.
- X.3. represents males who are forty-five and older. This variable was ***engineered*** to represent ***sex-specific*** advancing age. X.3 = 1 if the patient is male and forty-five or older. X.3 = 0 if the patient is female or the patient is male but younger than forty-five

Project Results: The Independent Variables

- X.4. It represents women who are fifty-five and older. X.4 = 1 if the patient is female and fifty-five or older. X.4 = 0 if the patient is male or the patient is female but younger than fifty-five. This variable was ***engineered***
- X.5 and is the patient's body mass index
- X.6 indicates if the patient is diabetic or not. X.6 = 1 if the patient is diabetic. X.6 = 0 if they are not
- X.7 represents the number of cigarettes smoked by the patient per day on average

Project Results: The Independent Variables

- X.8 indicates the smoking status of the patient. X.8 = 1 if the patient is a current smoker. X.8 = 0 otherwise. It represents the act of smoking. ***NOT*** how many cigarettes the patient smokes
- X.9 is the patient's systolic blood pressure
- X.10 indicates if the patient previously had hypertension. X.10 = 1 if they did. X.10 = 0 otherwise
- X.11 is the patient's total cholesterol level
- X.12 indicates if the patient previously had a stroke or not. X.12 = 1 if they did. X.12 = 0 if they didn't

Project Results: The Model

- The model we obtained and use for predictions is

$$ln(\frac{\hat{p}}{1-\hat{p}}) = -7.652 + 0.355X.1 + 0.053X.2 + 0.361X.3 + 0.266X.4$$
$$+ 0.003X.5 + 0.713X.6 + 0.021X.7 + 0.047X.8 + 0.014X.9$$
$$+ 0.198X.10 + 0.002X.11 + 0.901X.12$$

Project Results: A Prediction

- In the conclusion of our project, we use the model to determine if a fifty-year-old male who has a body mass index of 30, is diabetic, smokes 20 cigarettes per day on average, has a systolic blood pressure reading of 150, has a history of stroke, is hypertensive and total cholesterol level of 128. Plugging these values in to the model yields

$$ln(\frac{\hat{p}}{1-\hat{p}}) = -7.652 + 0.355 * 1 + 0.053 * 50 + 0.361 * 1 + 0.266 * 0$$
$$+ 0.003 * 30 + 0.713 * 1 + 0.021 * 20 + 0.047 * 1 + 0.014 * 150$$
$$+ 0.198 * 0 + 0.002 * 128 + 0.901 * 1$$

Project Results: A Prediction

- The estimated logit model returns 0.241. Solving for ***p*** we obtain

$$\frac{\hat{p}}{1-\hat{p}} = 0.241$$

$$\hat{p} = e^{0.241} - \hat{p}e^{0.241}$$

$$\hat{p} = \frac{e^{0.241}}{1+e^{0.241}}$$

Project Results: A Prediction

- Which translates to a 56.0% chance the patient will suffer from fatal or nonfatal CHD.

- This probability was significantly higher than the threshold value of ***c*** that we used. This ***c*** optimized the accuracy of the model. Therefore, for this patient, it is highly likely he will develop CHD within the next years if he does not take or follow any risk mitigation steps