**Health care: Heart attack possibility in Cleveland**

Date:08/10/2020 Group: 5

Cassidy Fitzpatrick | Hardeep Gill | Immaculate Chinonso Ezeogbo | Muhammad Junaid Saumtally | Xiaoxi Man | Yasser Khan

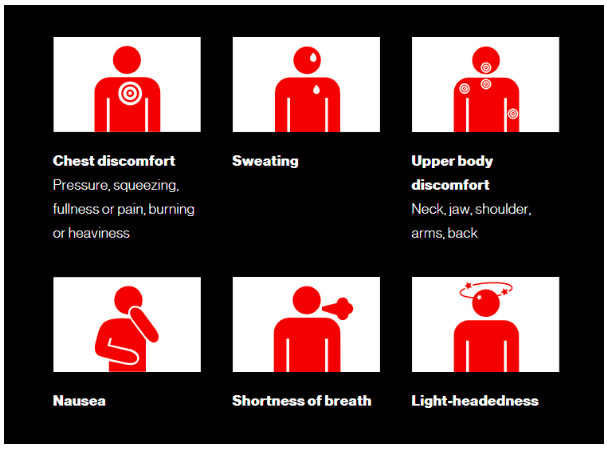
1. **Introduction and objectives**

A heart attack happens every 40 seconds in the United States, which makes it a leading cause of death among patients [1]. More than 805,000 Americans have a heart attack every year and in most cases, patients suffer from recurrent heart attacks [2]. Therefore, it is important to understand the leading factors of heart attack, in order to detect and prevent heart attacks.

The major causes and signs of heart attack have been studied extensively. The causes of heart attack include [3]:

* Excess weight, unhealthy food habits and lack of physical activity
* Diabetes causes damage to blood vessels so diabetes is a major factor in developing cardiovascular disease
* Smoking is one of the biggest causes of cardiovascular disease

The signs of a Heart Attack could be [4]:



Here we are using a dataset which was donated by Cleveland Clinic Foundation on 01-07-1988 for analysis and research, to create a regression formula to predict the possibility of a person having a heart attack.

1. **Data Preparation**

<https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility/data>

The database for the analysis contains 76 attributes but all published experiments refer to using a subset of 14 of them. Extracted from the dataset ‘heart.csv’ have no null values and contain 303 rows of data. The attributes used in the analysis are as follows:

*Table 1: List of all variables included in the data and their description*

|  |  |
| --- | --- |
| **Field Name** | **Medical Name and description** |
| Target (0 = unlikely or 1 = likely) | Heart Disease - likelihood of patient experiencing a heart attack |
| Age | Age of patient |
| Sex | Sex of patient |
| Cp (Value 0: typical angina, Value 1: atypical angina, Value 2: non-anginal pain, Value 3: asymptomatic) | Chest pain type |
| Trestbps (mm Hg on admission to the hospital) | Patients resting blood pressure |
| Chol (mg/dl) | Serum cholesterol |
| Fbs (more than 120mg/dl) | Fasting blood sugar |
| Restecg (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable left ventricular hypertrophy by Estes' criteria) | Resting electrocardiographic results of patient |
| Thalach | Maximum heart rate achieved |
| Exang (1 = yes; 0 = no) | Exercise induced angina |
| Oldpeak | ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot) |
| Slope | The slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping) |
| CA (0-3) | Number of major vessels coloured by fluoroscopy |
| Thal (0 = normal; 1 = fixed defect; 2 = reversible defect) | A blood disorder called thalassemia |

1. **Analysis**

**3.1 Data Visualization**

The dependent variable "target" is binary (Heart disease:0 = no, 1 = yes) and logistic regression has been used for prediction and also calculating the probability of success of the trained model.

The general Logistic regression equation is:



The first step in the analysis is to pre-process the data to identify and remove duplicate rows and any rows containing missing values. We used the box and whisker plots to detect outliers in each target variable(i.e. values that are outside the range of -3δ and +3δ). The outliers found in the target variables are trestbps, chol, thalach, and oldpeak and were removed from the data so as to not skew our results. Two rows were identified as duplicates and the entirety of those rows were removed.

All variables were plotted in the correlation heat map to understand the strength of relationship between different variables. As can be seen in the correlation plot (figure 1 a), cp, thalach and slope are highly correlated with the target. It was also found 'fbs' and 'chol' are the lowest correlated with the target variable. All other variables have some correlation with the target variable. Furthermore, a bar chart (figure 1 b) was plotted to count how many subjects had heart attacks. There are 131 patients with no heart attack and 162 patients who experienced a heart attack.

|  |  |
| --- | --- |
|  |  |

*Figure 1a) On the left a correlation matrix b) on the right a bar chart with heart attacks count*

**3.2 Individual Variable Regression Analysis**

Individual variable regression analysis between variables and the target were conducted to determine if each one of them can be used to indicate heart attack in patients. Furthermore, visualization on each one of the variables was also completed to better understand the relationships between variables and the target.

**3.2.1 Age and 3.2.2 Sex**

A regression analysis was conducted between age and target and sex and target. It was found that age is not a good predictor for heart attacks. Regression between sex and target did show that the sex variable can be used to predict heart attacks in patient. During the analysis, as shown in (figure 2 a) below, it was discovered that patients between the ages of 40-65 years are at a higher risk for heart attack. It was also determined, from (figure 2 b), that the probability of heart attack in females is 78% and in males is 45%.

|  |  |
| --- | --- |
|  |  |

*Figure 2 a) On the left, a line graph showing heart attack count over age, b) On the right, a bar chart showing probability of heart attack based on gender with blue being women and red being men.*

**3.2.3 Chest Pain (CP)**

An individual regression between different types of chest pains and target was completed. There are four different types of chest pains; typical angina, atypical angina, non-anginal pain and asymptomatic. As shown in figure 3b, it was determined 47% have typical angina, 17% have atypical angina, 29% have non-anginal and 8% asymptomatic pain. It was found chest pains are a good indicator for predicting heart attack. From (figure 3 a), based on the counts, it can be concluded that it is atypical angina and non-anginal chest pains that are more likely to induce a heart attack.

|  |  |
| --- | --- |
|  |  |

*Figure 3 : a) on the left, a bar chart showing different chest pain counts with respect to the target, b) a pie chart showing distribution between chest pains that are susceptible to heart attacks*

**3.2.4 Resting Blood Pressure (trestbps) and 3.2.5 Cholesterol Measurement (chol)**

The individual regression analysis conducted between resting blood pressure (trestbps) and the target and cholesterol measurement and the target. The analysis did not show a strong relationship between the two variables and the target. During the analysis it was determined (figure 4 a) that patients with resting blood pressure among 120 to 150 have a higher risk of getting heart attack. It was also found (figure 4 b), that cholesterol level among 200 to around 275 gives a higher possibility of heart attack.

|  |  |
| --- | --- |
| **A picture containing colorful, game  Description automatically generated** |  |

*Figure 4 a) on the left, a line graph showing resting blood pressure against target, b) on the right, line graph showing cholesterol level against target*

**3.2.6 Fasting Blood Sugar (fbs) and 3.2.7 Resting electrocardiographic measurement (restecg)**

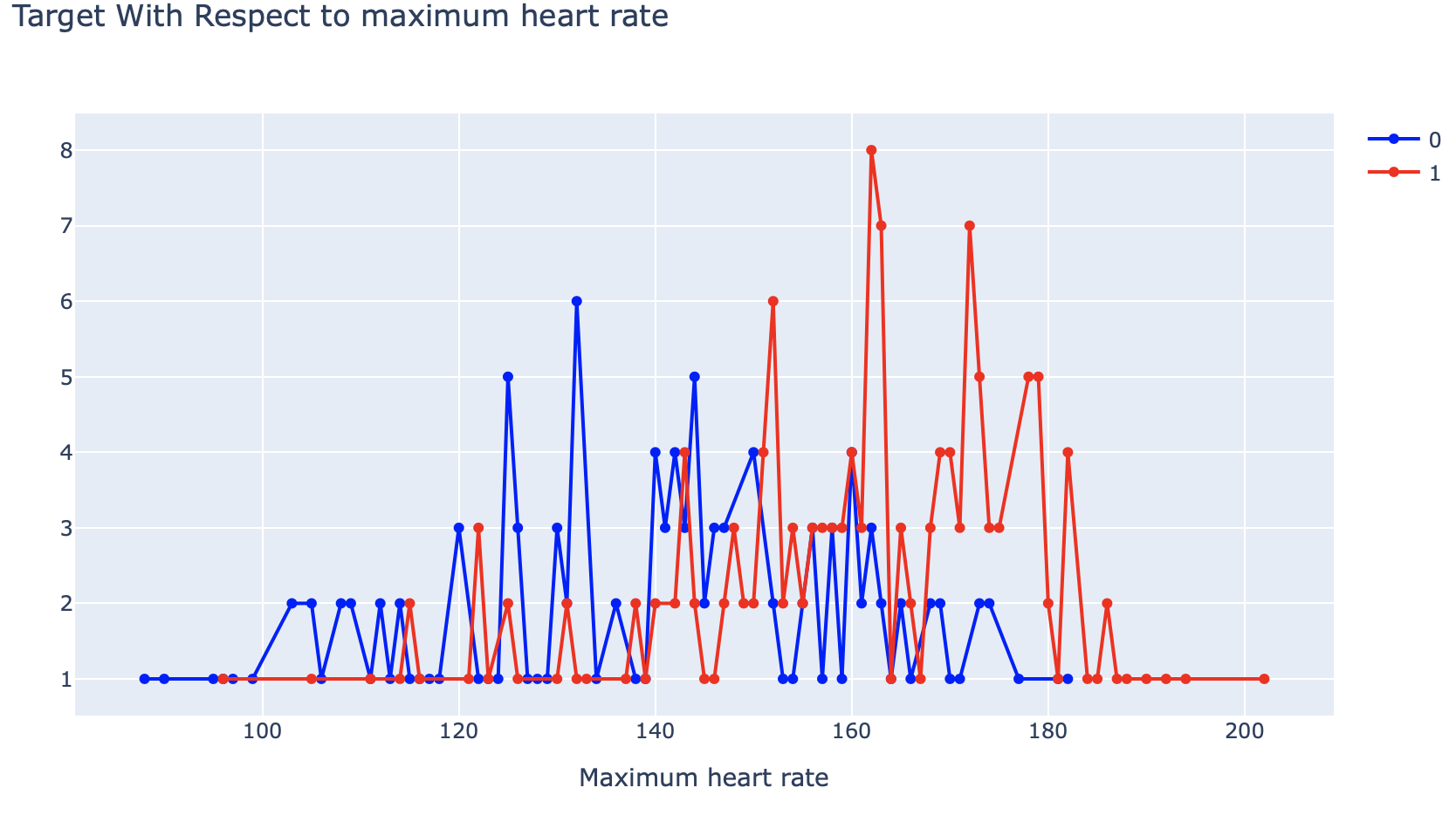
An individual regression was completed between fasting blood sugar and the target and resting electrocardiographic measurement and the target. It was found both of these variables were not good indicators for predicting heart attacks. Upon visual analysis, as shown in figure 5a, it was determined that 51% of patients have fasting blood sugar rate over 120 mg/dl. Furthermore, as shown in figure 5b, having ST-T wave abnormality is more likely to lead to a heart attack.

|  |  |
| --- | --- |
|  |  |

*Figure 5 a) on the left, a bar graph showing fasting blood sugar against target, b) on the right, bar graph showing electrocardiographic measurement level against target*

**3.2.8 Maximum Heart Rate (thalach)**

The individual regression conducted between maximum heart rate and the target indicated that it is a reliable variable to indicate risk of heart attack. From further analysis, as shown in figure 6, it was found that a heart rate higher than 140bpm shows a higher possibility of heart attack.

****

*Figure 6 showing maximum heart rate against target*

**3.2.9 Exercise Induced Angina (exang) and 3.2.10 Exercise induced depression (oldpeak)**

The individual regression between exercise induced depression and the target and exercise induced depression revealed that both of these variables are strong indicators for predicting the risk of heart attack. From data visualization, it was found (as shown in figure 7a) that very low to no exercise induced depression increases the risk of patients having a heart attack. In addition, as seen in figure 7b, the rate of heart attack for exercise is 70% and no exercise angina is 23%.

|  |  |
| --- | --- |
|  |  |

*Figure 7 a) on the left, a bar graph showing exercise induced angina against target, b) on the right, a plot showing target with respect to exercise induced depression*

**3.2.11 The slope of the peak exercise ST segment (slope) and 3.2.12 The number of major vessels (ca)**

The regression between the slope of the peak and the target indicates this variable is not a strong indicator to predict heart attacks among patients. The number of major vessels was found to be a strong indicator of increased risk in heart attacks based on the regression analysis conducted. From chart visualizations, it was determined that among different slopes of the peaks, the downsloping peak has the highest risk of heart attack as shown in figure 8a. Furthermore, it was determined, as shown in figure 8b, patients with ‘ca’ equal to 0 are more likely to have a heart attack while those with ‘ca’ of 3 are least likely to experience a heart attack.

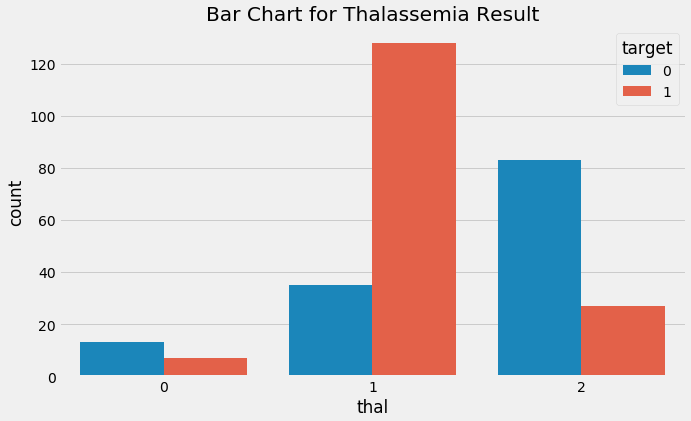
|  |  |
| --- | --- |
|  |  |

*Figure 8 a) on the left, a bar graph showing the number of blood vessels colored by a fluoroscopy, b) on the right, a bar chart showing target with respect to the number of vessels coloured*

**3.2.13 Thalassemia (thal)**

The individual regression analysis conducted between thalassemia and the target indicated that it is a strong variable to predict the risk of a heart attack. Thalassemia had three characteristics, normal, fixed defect, and reversible defect.(Note: There was an error on this dataset since we found four variables instead of the described three. Correction was made by subtracting one from each row of that column)

From figure 9, it can be seen that the frequency of heart attack depends a great deal on thalassemia (thal). It is clear that patients with 'thal' value equal to 1 (fixed defect) are more likely to have heart disease.

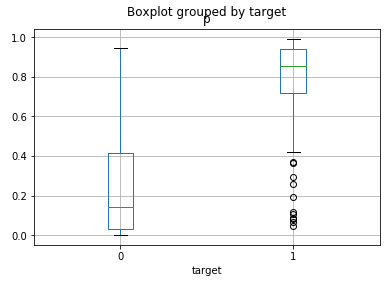
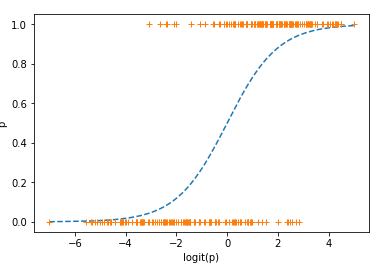


*Figure 9, a bar graph showing defect rates due to thalassemia against the target.*

**3.3 Model Regression Analysis**

Logistic regression is a type of regression analysis in statistics used for prediction of outcome of a categorical dependent variable from a set of predictor or independent variables. The target variable is binary and therefore, we shall use this technique for prediction and also calculate the probability of success. The summary from this analysis shows 13 attributes with p-values higher than the significance level of 5%, thereby showing low or no statistically significant relationship with the probability of a heart attack in these patients. Hence, the applied backward elimination approach is applied to remove features with p values greater than 5%.

Once completed, it was noticed that there is a strong evidence to prove that the features highlighted in the table all have a strong positive (cp, thalach,slope) and negative (sex, exang, thal, ca) associations with the target variable. This means that these features can be used to predict whether or not a patient is at risk of possible heart attack. Below (figure 10 b) we can see the boxplot of the distribution of the estimated odds against the value of the actual response.



*Figure 10 a, on the right, a logit plot of the analysis against the target, and figure 10 b, on the left, a boxplot of the distribution of the estimated odds against the value of the actual response*

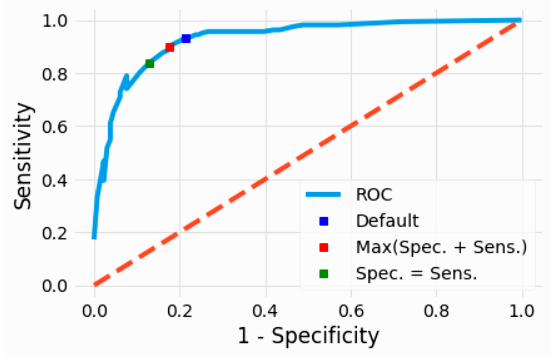
Next, evaluation of the logistic regression model was completed using the pseudo r-squared, statistical accuracy, MCC, precision, recall, specificity, f-score, confusion matrix, Receiver Operating Characteristic (ROC) and Area Under Curve (AUC).

The confusion matrix (figure 11 a) shows 101+148=249 correct predictions and 14+30=44 incorrect ones. This includes 148 true positives, 101 true negatives, 30 false positives, and 14 false negatives. Next, we extracted the TP as 101, the FP as 30, the FB as 14 and the TP at 148. Furthermore, the accuracy of our model was calculated at 85% which is quite high with a Mathews correlation coefficient between the estimates to have a true response of 70%. Our model’s precision was 83% and its recall was 91%. Additional stats include the model’s f-score of 87%, specificity of 77%, and sensitivity at 91%. From these statistics it is clear that the model is more sensitive than specific. The positive values are predicted more accurately than the negatives.

|  |  |
| --- | --- |
|  |  |

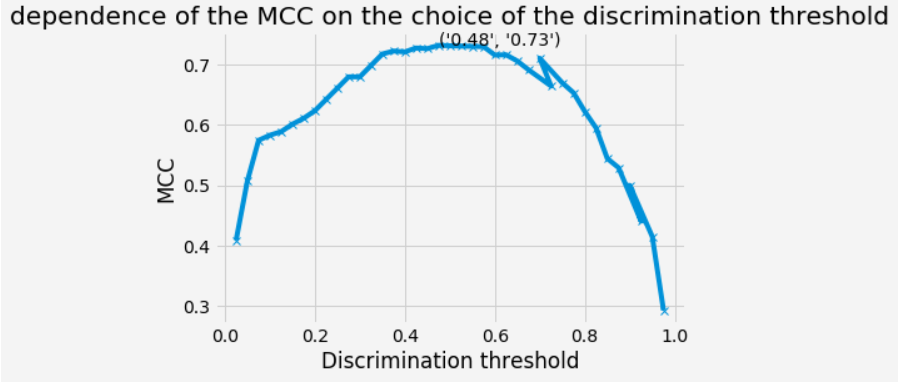
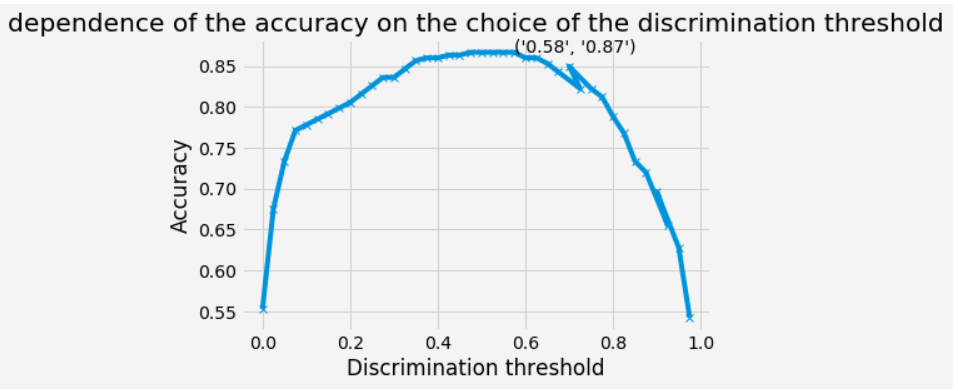
*Figure 11 a, on the left, the confusion matrix and 11 b, on the right, shows the ROC curve.*

Next, the ROC curve (figure 12) was generated with the set of pre-specified discrimination threshold of 0.5. The curve to the right shows the plot of true positive rate against false positive rate. Ideally, it is preferable to have a model that predicts lesser false positives and more true positives. For instance, if a solution that accepts a 20% false positive rate, can generate a total of almost 90% true positives. Area under the ROC curve was shown to be 0.92. The area under the ROC curve quantifies model classification accuracy; An area of 1 is ideal. The closer the AUC to 1 the better. This model's AUC is 0.93, which is highly desirable and another indicator model is reliable.



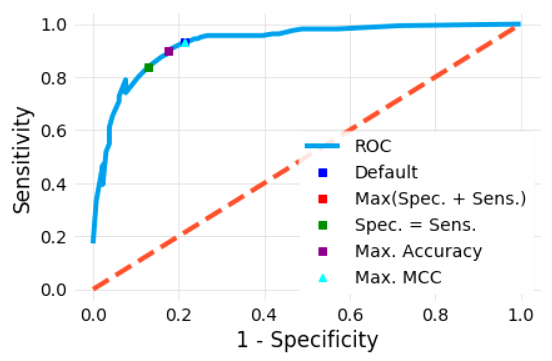
*Figure 12, the ROC curve with a discrimination threshold of 0.5*

The classifier made a total of 293 predictions (i.e., 293 patients for the heart attack). Out of those 293 patients, the classifier predicted that 163 patients are susceptible to heart attack, and 130 patients are not. But in reality, 162 patients are susceptible to heart attack, and 131 patients are not. From figure 13 a, it can be seen that the point where the obtained accuracy reaches the maximum is 0.87. The model's MCC is 0.73 at an optimal threshold of 0.58 as seen in figure 13 b. Additionally, from the above right plot, it can also be seen that the point where the obtained accuracy reaches the maximum is 0.73.



*Figure 13 a, on the left, shows the dependence of the accuracy on the choice of the discrimination threshold, and figure 13 b, on the right, shows the dependence of the MCC on the choice of the discrimination threshold.*

Finally, the discrimination threshold was adjusted to give us most accurate logistic mode. The classifier made a total of 293 predictions (i.e., 293 patients for the heart attack). Out of those 293 patients, the classifier predicted that 163 patients are susceptible to heart attack, and 130 patients are not. In reality, 162 patients are susceptible to heart attack, and 131 patients are not. Some additional statistics include the model's precision at 86% and it's recall at 86% as seen in figure 14. From these statistics it is clear that the model is a little more sensitive than specific. The positive values are predicted a little more accurately, than the negatives.

**

*Figure 14 shows the model sensitivity against specificity*

1. **Conclusion**

All attributes selected after the elimination process show p-values lower than 5% and thereby suggesting a significant role in the Heart disease prediction. In this project, it was established that the following signs/factors are strong predictors of heart attack; sex, chest pain, exercise induced angina (exang), the number of major vessels (ca), thalassemia (thal), maximum heart rate (thalach) and the slope of the peak exercise ST segment (slope). The model predicted with 87% accuracy. Furthermore, the model has a highly, almost balanced sensitivity and specificity. The Area under the ROC curve is 0.93 which is very satisfactory.

1. **References**

[1] Heron, M. [Deaths: Leading causes for 2017 pdf icon [PDF – 3 M]](https://www.cdc.gov/nchs/data/nvsr/nvsr68/nvsr68_06-508.pdf). National Vital Statistics Reports;68(6). Accessed November 19, 2019.

[2] Fryar CD, Chen T-C, Li X. [Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010 pdf icon [PDF-494K]](https://www.cdc.gov/nchs/data/databriefs/db103.pdf). NCHS data brief, no. 103. Hyattsville, MD: National Center for Health Statistics; 2012. Accessed May 9, 2019.

[3] <https://www.nhs.uk/conditions/heart-attack/causes/>

[4] Image Reference: <https://www.heartandstroke.ca/heart/conditions/heart-attack>