Consider “heart.excl” dataset to answer the following questions. This database contains 14 attributes. The "target" field refers to the presence of heart disease in the patient. It is integer-valued 0 = no/less chance of heart attack and 1 = more chance of heart attack

**Question 1 :** Is there any difference between the presence of heart disease (no/less chance / more chance) of female patients and male patients? (use a statistical test)?

**Answer :**

To compare the presence of heart disease between female and male patients, you can use a statistical test like the chi-square test for independence. This test assesses whether there is a significant association between two categorical variables. In this case, the variables would be gender (female or male) and the presence of heart disease (no/less chance or more chance).

H0 : there is no difference between the presence of heart disease of female patients and male patients.

H1 : there is difference between the presence of heart disease of female patients and male patients.

|  |  |  |  |
| --- | --- | --- | --- |
| Observed Frequency | | | |
|  | Target | | |
| Sex | 0-less risk | 1-risk | Grand Total |
| 0-Female | 24 | 72 | 96 |
| 1- Male | 114 | 93 | 207 |
| Grand Total | 138 | 165 | 303 |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Expected Frequency | | | |
|  | Target | | |
| Sex | 0-less risk | 1-risk | Grand Total |
| 0-Female | 43.72 | 52.28 | 96 |
| 1- Male | 94.28 | 112.72 | 207 |
| Grand Total | 138 | 165 | 303 |

The p-value (1.00716E-06) indicates the probability of observing the data if the null hypothesis (no association between gender and heart disease risk) were true. In this case, the very small p-value suggests that there is a significant association between gender and heart disease risk. Since the p-value is less than the typical significance level of 0.05, we reject the null hypothesis and conclude that there is evidence of an association between gender and heart disease risk.

Overall, this chi-square test indicates that there is a significant difference in the distribution of heart disease risk between males and females.

**Question 2:** Do you think normalization or standardization techniques would help to draw meaningful insights from the dataset? If so, what are those?

**Answer :**

Normalization or standardization techniques can indeed help in drawing meaningful insights from the dataset, especially if the features have different scales or units. Normalization scales the values of features to a range between 0 and 1, while standardization scales the values to have a mean of 0 and a standard deviation of 1. These techniques can make the data more comparable and help algorithms converge faster during training. For instance, you can use Min-Max scaling for normalization or z-score standardization.

**Question 3**: Perform suitable regression analysis and develop a predictive model. Further, interpret the model parameters.

**Answer :**

For regression analysis and predictive modeling, we can use techniques like logistic regression since the target variable is binary (presence of heart disease: 0 or 1). Logistic regression models the probability that a given input belongs to a particular class (in this case, the probability of having heart disease). We can use various features from the dataset as input variables to predict the likelihood of heart disease.

|  |  |  |
| --- | --- | --- |
|  | Coefficients | Odds |
| Intercept | 0.664556627 | 1.943629 |
| cp\_3 | 2.427629555 | 11.33199 |
| cp\_2 | 1.9995717 | 7.385892 |
| thal\_fixed defect\_1 | 1.705225169 | 5.502625 |
| thal\_reversable defect\_2 | 1.450450356 | 4.265035 |
| ca\_4 | 1.363794614 | 3.911006 |
| cp\_1 | 0.861151586 | 2.365884 |
| slope\_2 | 0.72498222 | 2.064694 |
| restecg\_1 | 0.466220198 | 1.593958 |
| thalach\_std | 0.46285603 | 1.588605 |
| fbs | 0.419236445 | 1.5208 |
| age\_std | 0.254451542 | 1.289754 |
| chol\_std | -0.22177476 | 0.801096 |
| restecg\_2 | -0.432267 | 0.649036 |
| oldpeak\_std | -0.45742054 | 0.632914 |
| trestbps\_std | -0.45955219 | 0.631566 |
| thal\_normal\_0 | -0.64350485 | 0.525448 |
| slope\_1 | -0.74578433 | 0.474362 |
| exang | -0.77658218 | 0.459975 |
| sex | -1.86368015 | 0.155101 |
| ca\_3 | -2.2577014 | 0.104591 |
| ca\_1 | -2.34206849 | 0.096129 |
| ca\_2 | -3.47871033 | 0.030847 |

These coefficients and odds ratios represent the impact of each predictor variable on the likelihood of heart disease. Let's interpret them:

**Intercept:** The intercept term represents the baseline log-odds of heart disease when all predictor variables are zero. The corresponding odds ratio (1.9436) suggests that the odds of heart disease are approximately 1.94 times higher when all predictor variables are at their reference levels.

**Chest Pain Types (cp\_1, cp\_2, cp\_3):**

These coefficients represent the effect of different types of chest pain on the log-odds of heart disease.

The odds ratios indicate that individuals with chest pain type 3 have the highest odds of heart disease compared to the reference category.

**Thalassemia Types (thal\_fixed defect\_1, thal\_reversible defect\_2, thal\_normal\_0):**

These coefficients represent the effect of different types of thalassemia on the log-odds of heart disease.

Individuals with a fixed defect thalassemia have the highest odds of heart disease among the thalassemia categories.

**Number of Major Vessels (ca\_4, ca\_3, ca\_2, ca\_1):**

These coefficients represent the effect of the number of major vessels colored by fluoroscopy on the log-odds of heart disease.

As the number of major vessels increases, the odds of heart disease also increase.

**Other Variables:**

Positive coefficients indicate an increase in the log-odds of heart disease, while negative coefficients indicate a decrease.

Odds ratios greater than 1 indicate an increase in the odds of heart disease for a one-unit increase in the predictor variable, while odds ratios less than 1 indicate a decrease.

For example, being male (sex) is associated with lower odds of heart disease compared to being female, as indicated by the odds ratio of 0.1551.

Overall, these coefficients and odds ratios provide insights into how each predictor variable influences the likelihood of heart disease in the logistic regression model.

**Question 4 :** Assess the model's predictive power using appropriate evaluation tools.

**Answer :**

To assess the predictive power of the model, you can use evaluation tools such as confusion matrix, accuracy, precision, recall. These metrics will help us understand how well the model is performing in terms of correctly predicting instances with and without heart disease. Additionally, we can use techniques like cross-validation to ensure the robustness of the model evaluation.  
  
The cut-off value is .5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classification Report | | | | |
|  |  | Predict | Predict |  |
|  |  | 0 | 1 |  |
| Actual | 0 | 115 | 115 | 230 |
| Actual | 1 | 115 | 115 | 230 |
|  |  | 230 | 230 | 460 |

|  |  |
| --- | --- |
| Accuracy | 88.45% |
| Precision | 86.93% |
| Recall | 92.73% |
| TPR | 92.73% |
| FPR | 16.67% |

These metrics are commonly used to evaluate the performance of a binary classification model, such as logistic regression, in predicting the likelihood of heart disease. Let's break down each metric:

**Accuracy:** Accuracy measures the proportion of correctly predicted instances (both true positives and true negatives) out of all instances. In this case, the accuracy of the model is 88.45%, indicating that it correctly predicts heart disease status for approximately 88.45% of the cases in the dataset.

**Precision:** Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It is calculated as the ratio of true positives to the sum of true positives and false positives. In this case, the precision of the model is 86.93%, indicating that among all instances predicted as having heart disease, approximately 86.93% actually have heart disease.

**Recall (True Positive Rate - TPR):** Recall, also known as sensitivity or true positive rate (TPR), measures the proportion of true positive instances that were correctly identified by the model out of all actual positive instances. It is calculated as the ratio of true positives to the sum of true positives and false negatives. In this case, the recall of the model is 92.73%, indicating that the model correctly identifies approximately 92.73% of all individuals with heart disease.

**False Positive Rate (FPR):** FPR measures the proportion of false positive instances (instances wrongly predicted as positive) out of all actual negative instances. It is calculated as the ratio of false positives to the sum of false positives and true negatives. In this case, the FPR of the model is 16.67%, indicating that approximately 16.67% of all individuals without heart disease were wrongly classified as having heart disease.

In summary, these metrics provide a comprehensive assessment of the model's performance in predicting heart disease. A high accuracy, precision, and recall, along with a low false positive rate, indicate that the model performs well in distinguishing between individuals with and without heart disease.

|  |  |  |  |
| --- | --- | --- | --- |
| Cut-off | Accuracy | Precision | Recall |
| 0 | 0.544554 | 0.544554 | 1 |
| 0.1 | 0.80198 | 0.737557 | 0.987879 |
| 0.2 | 0.834983 | 0.780488 | 0.969697 |
| 0.3 | 0.867987 | 0.830688 | 0.951515 |
| 0.4 | 0.881188 | 0.848649 | 0.951515 |
| 0.5 | 0.884488 | 0.869318 | 0.927273 |
| 0.6 | 0.877888 | 0.9 | 0.872727 |
| 0.7 | 0.867987 | 0.92517 | 0.824242 |
| 0.8 | 0.818482 | 0.958333 | 0.69697 |
| 0.9 | 0.745875 | 0.958333 | 0.557576 |
| 0.92 | 0.726073 | 0.955556 | 0.521212 |

From this table we can see that for each cut-off point how the model will perform. At cut-off = .6 Accuracy, Precision & Recall value concides.

|  |  |  |
| --- | --- | --- |
| Cut-off | TPR | FPR |
| 0 | 1 | 1 |
| 0.1 | 0.987879 | 0.42029 |
| 0.2 | 0.969697 | 0.326087 |
| 0.3 | 0.951515 | 0.231884 |
| 0.4 | 0.951515 | 0.202899 |
| 0.5 | 0.927273 | 0.166667 |
| 0.6 | 0.872727 | 0.115942 |
| 0.7 | 0.824242 | 0.07971 |
| 0.8 | 0.69697 | 0.036232 |
| 0.9 | 0.557576 | 0.028986 |
| 0.92 | 0.521212 | 0.028986 |

Here we can see the False Positive rate is close to zero which is also a very good indicator.