Regression Analysis of Heart Attack Risk

Executive Summary

Problem & Hypothesis

Using a dataset generated using ChatGPT that mimics a realistic experience from Kaggle, to what extent does cholesterol and diabetes affect heart attack risk?

The project topic is about exploring features relevant to heart health and lifestyle choices to determine if any of the features, including indicators such as cholesterol and diabetes, have a higher impact on the presence or absence of a heart attack risk than the other features through logistic regression analysis. Therefore, the hypothesis of the analysis is:

- Null Hypothesis- Cholesterol and diabetes do not statistically significantly affect the risk of a heart attack.
- Alternate Hypothesis- Cholesterol and diabetes statistically significantly affect the risk of a heart attack.

The null hypothesis means there is no significant effect on heart attack risk because of cholesterol and diabetes. This would mean any observed difference in heart attack rates is simply due to chance and not a real effect of the factors being studied. The goal is to reject the null hypothesis if the data provides sufficient evidence against it. The alternate hypothesis is the statement that is the opposite of the null hypothesis that I am trying to prove by performing the data analysis.

Data Analysis Process

The dataset has 8,763 global patient records consisting of 26 features related to heart health, lifestyle choices, lifestyle factors, and medical aspects to conduct a predictive analysis of the binary categorical feature, if the patient is at risk of a heart attack or not. The goal of data cleaning is to find any null or duplicated values in the dataset, correct any error or inconsistencies, and to get rid of any unnecessary variables that will not be used for the regression analysis.

Exploratory data analysis (EDA) is the initial step in inspecting data through visualizations to summarize its key characteristics and assess its suitability for modeling. EDA provides a summary of all variables in the dataset through univariate visualizations. Bivariate visualizations, on the other hand, display the relationship between independent variables and the target variable, helping to identify any correlations between them. Given that the analysis includes categorical variables, these were transformed into dummy variables. For logistic regression, one-hot encoding is essential. For variables with more than two categories, each category is represented as a separate column with binary values. This transformation effectively prepares categorical data for use in machine learning models.

An initial logistic model was constructed. To refine the model, the variance inflation factor (VIF) for each independent variable was calculated to identify those that should be removed due to elevated multicollinearity. After determining the VIFs, variables with values exceeding 10 were sequentially eliminated. Next, the backward stepwise elimination method was applied. This technique reduces the number of predictors and helps mitigate overfitting by removing features that do not significantly impact the target variable. Variables with p-values greater than 0.10 were eliminated, indicating a tolerance for a higher risk of false positives. This process reduced the number of independent variables from 20 to 1.

Outline of Findings

The regression equation for the reduced model: y = -0.713 + 0.0005(Cholesterol)

An interpretation of the coefficients of the reduced model:

- y represents risk of a heart attack.
- Keeping all things constant, for one unit of increase in cholesterol, the log odds of heart attack risk increase by 0.05%.

The removal process resulted in a decrease in the Log-Likelihood value from -5708.1 to -5715.2. It is crucial to compare models with the same number of predictors, as the Log-Likelihood tends to decrease regardless of statistical significance. In logistic regression, the Log-Likelihood Ratio (LLR) p-value indicates the model's significance, which fell from 0.5544 to 0.07021.

A low LLR p-value (typically < 0.05) allows us to reject the null hypothesis, while a high p-value (typically > 0.05) means we fail to reject it. Despite dropping diabetes, the LLR p-value for the model that included both cholesterol and diabetes was 0.05068. Both values exceed 0.05, indicating a failure to reject the null hypothesis. This suggests that the model is not significantly better than one using only the intercept, meaning the predictors may not contribute much to explaining the outcome.

Limitations

Based on the results of the logistic regression analysis, a key limitation of the study is the removal of features that may be crucial in real-world scenarios. Features are eliminated by calculating their variance inflation factors and removing those with excessively high values, followed by backward stepwise elimination to further refine the model. Additionally, logistic regression has limitations in modeling continuous variables or nonlinear relationships; it may struggle with complex underlying relationships and is prone to overfitting when analyzing large datasets.

Summary of Proposed Actions

The inability to reject the null hypothesis raises questions about whether cholesterol and diabetes significantly impact heart attack risk. That said, the model does serve as a useful starting point and the data summary generally makes sense concerning the order in which variables were removed during the analysis. A recommended course of action based on the research question stemming from the model results would be to encourage patients to closely monitor their cholesterol levels. Individuals with diabetes must also be vigilant, as they are at a significantly higher risk of heart attacks compared to those without the condition.

Overall, the features included in the analysis remain important, as they are commonly assessed by doctors to evaluate a patient's risk of a heart attack. For future studies, I would suggest incorporating additional CSV files with the same variables since logistic models tend to perform better with larger datasets. In real-life applications, logistic regression analysis might seem like a suitable approach. However, it may be worth exploring a different machine learning approach instead of sticking with logistic regression, as they could potentially yield a better-fitting model with higher accuracy.

Expected Benefits

This study helps healthcare professionals identify the key factors that increase the risk of a heart attack. By understanding how different risk factors relate to each other, they can make better prevention and treatment plans based on each patient's unique situation. The study uses logistic regression to provide coefficients for each risk factor, making it easy to see how they affect the chance of a heart attack. Healthcare providers can focus on the most important risks for each patient, leading to more targeted care.

Assessing heart attack risk has many benefits. It allows for early detection of problems, encourages healthy lifestyle changes, creates targeted treatment plans, and reduces the overall risk of heart issues. This proactive approach can help individuals take steps to prevent a heart attack, potentially saving lives.