**Research Paper**

**Myocardial Infarction Complications**

Jijith Vasudevan Nair

Colorado State University Global

MIS581: Capstone: Business Intelligence and Data Analytics

Dr. Justin Bateh

02/05/2023

**Abstract**

Chronic illness has become a growing problem in each developed and developing country. For getting accurate results, the selection of parameters is an essential process in the data mining process. Predictive analysis needs to be conducted to understand the likelihood of complications in myocardial infarction. Myocardial infarction is caused by sudden damage of myocardial tissue, causing organs to fail due to insufficient oxygen supply. When the myocardial tissue is ruptured, it blocks the coronary vessel, causing plaque in the arteries. This rupturing causes ischemia, where the muscle cells start to damage and will die sooner or later. It is also known as a heart attack when the arteries are blocked or suddenly experience prolonged blood flow. Predicting the complications based on information from the patient's hospital records at the time of admission and hospital period will help the clinic to help reduce risk factors and thereby save a lot of time, effort, and money for the patient. The Myocardial Infarction (MI) dataset has 124 variables, from which 2-112 variables could be used as input features for predicting and 113-124 variables could be used for possible complications. The purpose of the research is to find the input feature responsible for the complications and to test whether the age or gender of the patient has any influence. The null hypothesis for the two business questions was rejected because there was statistical significance using p-values of less than 0.05. Two input features, age, and nr\_03, were identified as significant in explaining the complications. This shows that the dataset can be used to predict complications from myocardial infarction to a certain extent. Only one input feature showed its statistical significance with the model and the patient's age.

**Introduction**

Myocardial infarction is caused by sudden damage of myocardial tissue, causing organs to fail due to insufficient oxygen supply. When the myocardial tissue is ruptured, it blocks the coronary vessel, causing plaque in the arteries. This rupturing causes ischemia, where the muscle cells start to damage and will die sooner or later. It is also known as a heart attack when the arteries are blocked or suddenly experience prolonged blood flow. The sudden blockage is caused in the coronary artery is usually because of a blood clot formation called a thrombus. The common symptoms include chest pain or discomfort that leads into the shoulder, jaw, or neck (Čulić, 2007). It often occurs in the center of the chest or left side of the chest and lasts for some time.

The prolonged blood flow in the coronary artery occur when there is faster heart beat or the patient suffers from low blood pressure. Heart attack can occur without blood clot formation if there is a shortage in oxygen supply when the demand is more significant. Patients with atherosclerosis are very likely to experience a heart attack. Studies show that every 40 seconds, someone in the US has a heart attack. Approximately 790,000 Americans have a heart attack every year. A general topic explored in the medical world today is a heart attack. It is assumed that an individual in their older age who suffers from a heart attack has a higher risk of death. Even if chest pain is a significant warning sign of a heart attack, it is often confused with other conditions like indigestion, pneumonia, and heartburn. In addition to looking at an individual's ECG for an early diagnosis, other factors can help predict or detect heart attacks earlier, which is the primary objective of this project. Clinical research suggests that the risk factors like physical activity, stress, eating, temperature, coffee, or alcohol should also be considered. Modifications in lifestyle, physical activity, and adequate care may defer the condition's occurrence.

**Objectives**

The dataset proposed for this project contains a CSV file for the Myocardial Infarction complication prediction database collected in Krasnoyarsk. Predictive models are one of the most common analytic tools to identify at-risk patients. Forecasting these risks accurately will help effective patient care with cost-effective treatment (Cai et al., 2020). Tailoring and combining different models and investigating if these models can be improved by extending or adding new predictors. Various predictive analytics can be used in solving the heart disease data problem. For this assignment, a decision tree is chosen as the model-building technique to make the model easier to interpret and understand. A decision tree is a supervised automatic learning technique used for classification tasks. A tree is constructed by an algorithm that identifies a path that splits the dataset based on certain conditions. The goal here is to create a model that predicts if a patient has a risk of heart failure by learning decision rules inferred from the patient's health records.

Healthcare industries worldwide generate, collect, and store a large amount of data. With the advanced technologies, the approach to determining the factors causing health issues has now increased based on machine learning techniques. Predictive analysis needs to be conducted to understand the likelihood of complications in myocardial infarction. Chronic illness has become a growing problem in each developed and developing country. For getting accurate results, the selection of parameters is an essential process in the data mining process. Research in the field of cardiovascular using data mining and prediction is an ongoing effort. Many organizations, such as Cleveland heart clinic, conduct studies and surveys to collect data for cardiovascular diseases.

This project aims to discuss and attempt predictive analytics for complications in Myocardial infarction using the supervised learning method. This is done by stating the identified data problem, determining what factors or variables are responsible for the analysis, and building a model using predictive analytics. Decision trees are the selected model for the project because they can visually represent decisions and decision-making that can be interpreted easily by the readers.

**Overview of study**

Myocardial infarction can lead to heart failure, arrhythmia, and death. Clinical studies have shown that early identification and timely intervention for acute MI can significantly reduce mortality. The risk assessment models are subjective, and the data that go into them are challenging to obtain. Generally, the assessment is only conducted among high-risk patient groups. With cardiovascular disease increasing, substantial research has focused on developing prediction tools. Using machine learning, Myocardial infarction prediction provides an individualized assessment of the likelihood of myocardial infarction. It can be used to identify low- and high-risk patients and help their clinical decisions.

According to Mythili et al. (2013), around 17 million people die annually due to cardiovascular disease. Heart disease is much easier to treat if detected in the earlier stages. Common heart diseases include coronary heart disease, hypertension, cardiomyopathy, heart failure, etc. Heart disease prediction is a significant problem in health care as it is one of the leading issues among patients with unhealthy lifestyles. Early prediction of these diseases will be beneficial for aversion or cure.

Myocardial infarction can have complications during treatment that may not worsen the prognosis during the term. Considering that, most patients in their acute and semi-acute periods may experience complications leading to a worsening of the course diagnosis or even patient’s death. Often a vey experienced heart specialist cannot always consider or prepare for the development of these complications. Predicting myocardial infarction complications in a timely manner can carry out the necessary preventive measures wherever necessary.

Predicting the complications based on information from the patient's hospital records at the time of admission and hospital period will help the clinic to help reduce risk factors and thereby save a lot of time, effort, and money for the patient (Golovenkin et al., 1997). These complications could affect decision-making efforts by the doctor or the caregiver. Patient death due to these complications could also be avoided if the problem is solved with better accuracy. The data analysis will help solve the prediction of complications by reducing the risk factors during the treatment and providing them with the necessary care and measures to prevent heart attacks on time.

**Research questions and hypotheses**

The research question is related to an issue that a specific organization or entity faces. A research project could be initiated to address an issue an organization wants an analyst to investigate for the company. The issue and question for this course are more leaning towards organization or industry-specific, and due to that, it is not generalizable.

The most challenging issue in modern healthcare is Myocardial infarction. Acute myocardial infarction is usually correlated with higher mortality rates in the initial years and after. Myocardial infarction remains high in all countries (Heather et al., 2022). This is true for the population in urban areas, where they are exposed to stress and imbalanced nutrition. For example, millions of people suffer from heart attacks in the United States every year, and many die before arriving at the hospital.

MI could occur without or with complications that usually do not worsen the long-term forecast of the likely course. Half of the patients simultaneously in the acute and sub-acute periods could have complications that lead to serious implications and even death. The following questions will help get insights.

1. Which input feature is most responsible for in-hospital deaths due to myocardial infarction?
2. Is gender an influencing factor in complications due to Myocardial infarction?

**Overview of hypotheses**

Hypothesis testing is an act of inferential statistics used widely by analysis to test an assumption. It starts with sample data and hypothesis tests; Analysts test a statistical statement to see if it supports the argument (Paneth & Susser, 1995). Based on the best evidence, hypotheses are educated guesses that variables like age, gender, STENOK\_AN., influence the complications from MI.

**Hypothesis testing:**

***Question 1***

Determine if any input features contribute to in-hospital deaths due to myocardial infarction.

Ho: There is no significant contribution by the predictors towards the complications from myocardial infarction

Ha: There is a significant contribution by the predictors towards the complications from myocardial infarction.

***Question 2***

Research and compare the effect of (i) gender; and (ii) age on in-hospital complications due to myocardial infarction.

Ho: There is a correlation between gender and the complications from myocardial infarction

Ha: There is no significant correlation between gender in complications from myocardial infarction.

Age

Ho: There is a correlation between Age and the complications from myocardial infarction

Ha: There is no significant correlation between Age in complications from myocardial infarction.

**Literature review**

According to the study by Guy S (1995), MI complications are uncommon. Their importance is emphasized because of their ability to fix them with early diagnosis and suitable treatment. The prolonged blood flow in the coronary artery occurs when there is a faster heartbeat, or the patient suffers from low blood pressure. Heart attack can occur without blood clot formation if there is a shortage in oxygen supply when the demand is more significant. Patients with atherosclerosis are very likely to experience a heart attack. Studies show that every 40 seconds, someone in the US has a heart attack. Approximately 790,000 Americans have a heart attack every year. A general topic explored in the medical world today is heart attack. It is assumed that an individual in their older age who suffers from a heart attack has a higher risk of death.

According to McCrath et al. (2005), around 17 million people die annually due to cardiovascular disease. Heart disease is much easier to treat if detected in the earlier stages. Common heart diseases include coronary heart disease, hypertension, cardiomyopathy, heart failure, etc. Heart disease prediction is a significant problem in health care as it is one of the leading issues among patients with unhealthy lifestyles. Early prediction of these diseases will be beneficial for aversion or cure.

Myocardial infarction can have complications during treatment that may not worsen the prognosis, considering that most patients in their acute and semi-acute periods may experience complications leading to a worsening course diagnosis or even patient death. Often a very experienced heart specialist cannot always consider or prepare for the development of these complications. Predicting myocardial infarction complications on time can carry out the necessary preventive measures wherever necessary.

Predicting the complications based on information from the patient's hospital records at the time of admission and hospital period will help the clinic to help reduce risk factors and thereby save a lot of time, effort, and money for the patient (Golovenkin et al., 1997). These complications could affect decision-making efforts by the doctor or the caregiver. Patient death due to these complications could also be avoided if the problem is solved with better accuracy. The data analysis will help solve the prediction of complications by reducing the risk factors during the treatment and providing them with the necessary care and measures to prevent heart attacks on time.

**Research design**

**Methodology**

The Myocardial Infarction (MI) dataset has 124 variables, from which 2-112 variables could be used as input features for predicting and 113-124 variables could be used for possible complications. Complications for prediction include the admission time to the hospital from variables in columns (2-112) except columns 93 to 95, 100 to 105, and the first day. All input variables except 94 and 95, 101 and 102, and 104 and 105, the second day, can be used for prediction. The dataset contains 1700 observations, 111 input features, and 12 complications. Myocardial Infarction (MI) dataset contains 7.6% of missing values too.

Some of the analytical techniques used are data mining which involves importing a dataset and getting access to the dataset from <https://archive.ics.uci.edu/ml/datasets/Myocardial+infarction+complications>. The next step will be data cleaning which processes the data by replacing non-numeric values using numeric values and selecting required variables from columns and observations in each dataset.

SAS on Demand will be used for summary statistics to retrieve information and details about the Myocardial Infarction dataset. The histograms created with SAS SGPLOT will help summarize the data set quantitatively. Boxplot provides data distribution on a five-number summary from minimum to maximum with the first quartile, median, and third quartile. The outliers and the skew will be easier to understand using the box blots for each variable. Histograms and Vertical box plots are created using PROC UNIVARIATE, PROC GPLOT, and PROC BOXPLOT procedures in SAS(SAS Help Center, 2021).

There are binary attributes and numeric attributes in the Myocardial Infarction (MI) dataset. The binary or categorical attributes are mentioned. There are 32 numeric variables, and the remaining variables are binary types. All complications are represented as binary or categorical attributes. Numerical attributes include some Ordinal attributes. The variables with a sequence that are meaningful as a ranking are represented as ordinal values. The dataset contains many such variables where the actual numeric values are unknown. THE SAS INPUT function will be used for the conversion. SAS will automatically try to convert variables into numeric variables in some cases.

**Methods**

The data dictionary is mainly used for enhanced database assessments. Data analysis allows for building representations in the way of data design, commonly known as conceptual. Conceptual information explains the data and the procedures for using the data. Data analysis equals the establishment of data nature, while functional analysis equals the establishment of data usage. However, both data and functional analysis are integrated. The following are some of the key variables that will be used for the data analysis.

**Complications**

|  |  |
| --- | --- |
| Atrial fibrillation (FIBR\_PREDS) | Binary |
| Supraventricular tachycardia (PREDS\_TAH) | Binary |
| Ventricular tachycardia (JELUD\_TAH) | Binary |
| Ventricular fibrillation (FIBR\_JELUD) | Binary |
| Third-degree AV block (A\_V\_BLOK) | Binary |
| Pulmonary edema (OTEK\_LANC) | Binary |
| Myocardial rupture (RAZRIV) | Binary |
| Dressler syndrome (DRESSLER) | Binary |
| Chronic heart failure (ZSN) | Binary |
| Relapse of the myocardial infarction (REC\_IM) | Binary |
| Post-infarction angina (P\_IM\_STEN) | Binary |
| Lethal outcome (cause) (LET\_IS) | Binary |

**Feature**

|  |  |
| --- | --- |
| Age | Numeric |
| Gender | Numeric |
| Serum sodium content (Na\_BLOOD) | Numeric |
| Serum AlAT content (ALT\_BLOOD) (IU/L) | Numeric |
| Serum AsAT content (AST\_BLOOD) (IU/L) | Numeric |
| Serum CPK content (KFK\_BLOOD) (IU/L) | Numeric |
| White blood cell count (billions per liter) (L\_BLOOD) | Numeri |
| ESR (Erythrocyte sedimentation rate) (ROE) (мм) | Numeric |

The Myocardial Infarction (MI) data is imported using the SAS import, selecting columns from 1 to 124 and then displaying the summary of attributes to get a feel of the dataset. During the data cleaning process, the non-numeric characters were replaced with numeric characters. we are saving our dataset into a CSV format while giving it a new header.

The correlation will show how related the features are to each other and the target variable. Distribution analysis will show how sales campaigns influenced sales under the different clusters. Box plots will also help to identify what variable has impacted the target variable most. Fitting into a Logistic regression model and predicting sales using the available variables is an integral part of the study.

**Limitations**

Some of the limitations of Myocardial Infarction (MI) have non-numeric characters, which had to be substituted with numeric characters and thus might result in the variance of the results. The number of fields should be limited in the Myocardial Infarction (MI) analysis to avoid losing out on vital overall information, which might change analysts’ perceptions (Golovenkin et al., 2020).

**Ethical Considerations**

EHRs improve the quality of care, encourage patient, collaborative efforts, aid in disease diagnosis, increase practice efficiency, and provide constant access to patient health information. Additionally, how we communicate has changed due to smartphones and other web-enabled smart devices. These gadgets enable users to quickly and conveniently access the online services offered by various businesses. As a result, both individuals and organizations are now very concerned about data privacy and confidentiality. Accountability for healthcare data breaches is crucial because these types of data are much more sensitive than other types of data, making tampering with them potentially fatal and irreversible for patients. Healthcare data should therefore have increased security and be breach-proof. All the data used in the project will follow HIPAA regulations, and PHI data is not included in the analysis.

**Findings**

**The binary logistic regression model**

PROC LOGISTIC can be used to run logistic regression on binary response data. The response, Y, of the analysis, can take two possible values, the patient has complications from MI (1), or the patient does not have any complications (0) (SAS Help, 2021). The logistic regression model will help explain the effects of input features variables on the complications from Myocardial Infarction. The LOGISTIC procedure specifies categorical and explanatory variables as model statements to fit a model.

**Interpretation of the results of the logistic regression model**

PROC LOGISTIC returns background information about the fitted model first. The information included in the results is the name of the data set, the response variable, the number of observations, and the function. The Response Profile table (Figure 1) lists the categories of outcome (complication and no complication) and their total frequencies for the current sample. The class-level information explains the categories available in the explanatory variable.

The results shown in Figure 2 with Model Fit data (Figure 2) consist of AIC and BIC data for the model. Comparing the results from AIC and SC are very helpful, and the smaller values, the better. The Results display that the AIC and SC are better than the baseline, which isn't always ideal.

The Testing Null Hypothesis table explains that the null hypothesis tests that AGE and nr\_03 are significant from the selected input features. The Analysis of Maximum Likelihood Estimates table's estimate column contains the Wald Chi-Square statistics, with Wald Chi-Square tests and the p-value of the test. The Wald chi-square test explains that the selected explanatory variables (AGE, nr\_01, nr\_02, and nr\_03) are statistically significant, indicating that the null hypothesis can be rejected for all hypotheses.

**Predictive Modeling**

Predictive modeling is an important data modeling technique that creates a statistical model demonstrating regression and data classification. If the Response variable is nominal, predictive modeling is called classification. PROC GLMSELECT utilizes stepwise selection with options SELECT=SBC and STOP=SBC. Stepwise selection terminates by adding or dropping any effect that increases the SBC statistic. The model at step 3 minimizes SBC and is selected.

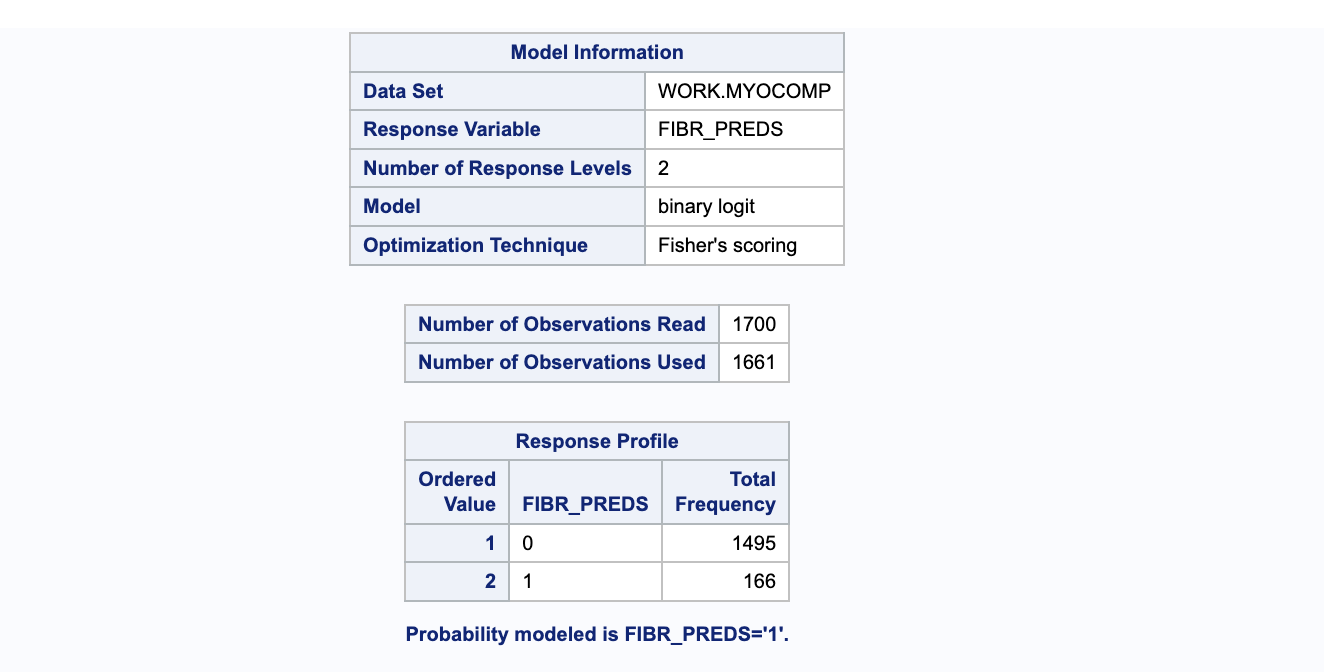
The chi-squared can compare explanatory variables' power in regression models containing multiple predictors. The chi-square value is 50.71, for nr\_03, and AGE is 39.02. Both the R-squared and adjusted R-squared are moderately acceptable. The coefficient estimates can be significant because their p-values are less than 0.05.

**Recommendations for further analysis**

Business analytics techniques often require vast data sets much larger than what is available from medical experiments. The Myocardial Infarction data sets with less than 1700 records are small and not the best for stable performance for prediction models. Based on the analysis and R-Squared value, increasing the data points would be ideal for more accurate results. To better fit the data outliers, we need to have removed them to improve the R-Squared and adjust the R-squared value.

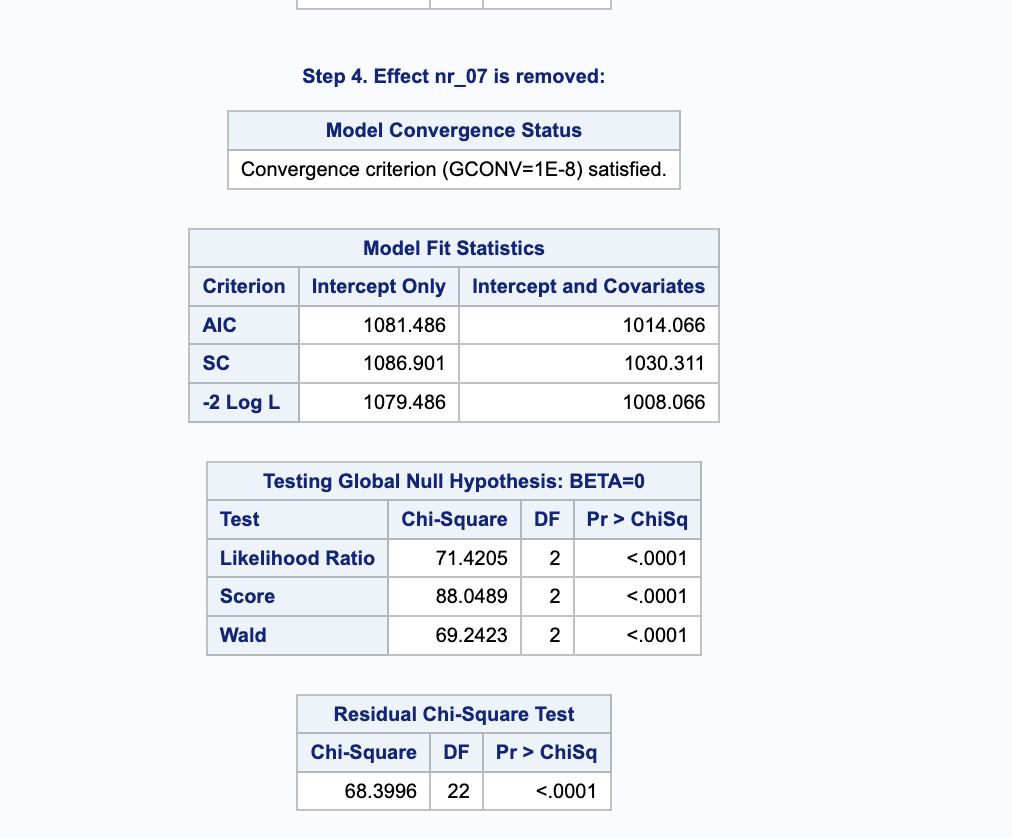
**Figure (Screenshot) 1**

*Response profile*

**

**Figure (Screenshot) 2**

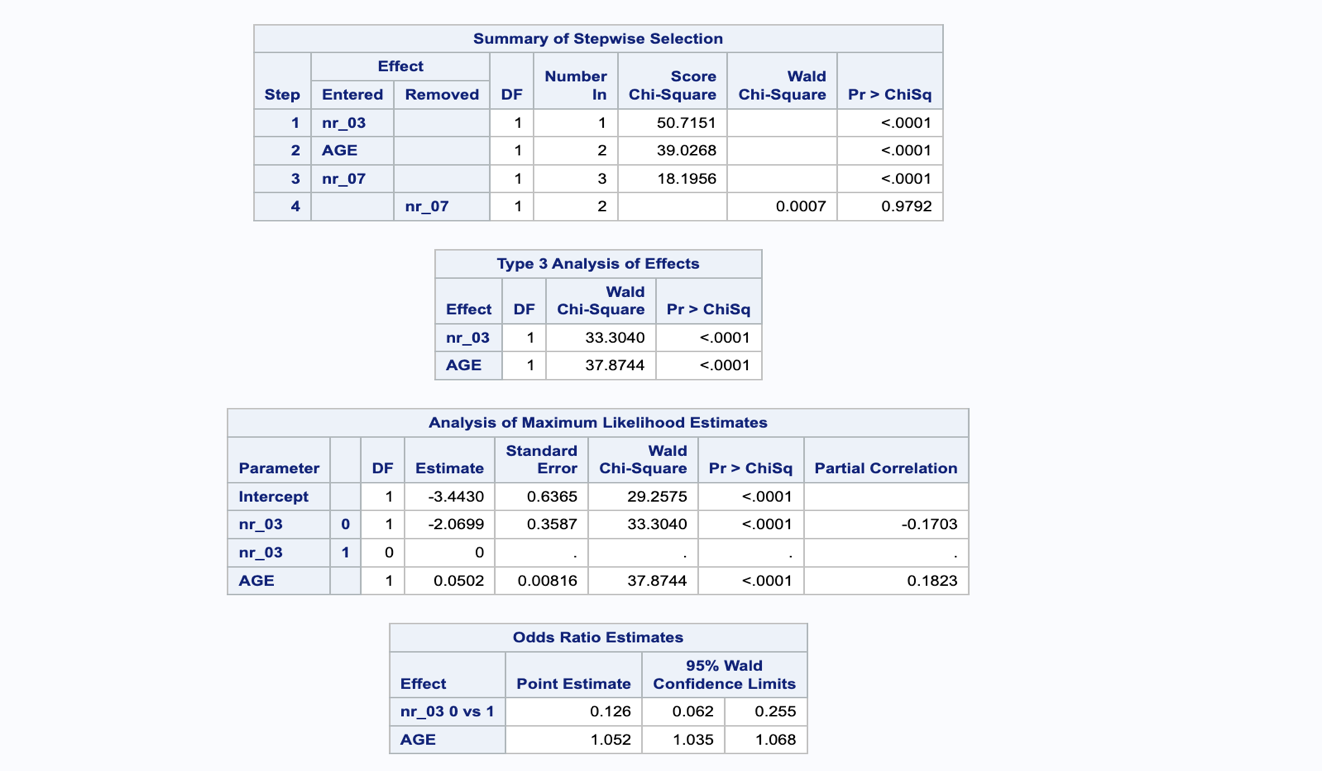
*Stepwise selected model*

**

Stepwise regression builds a model by removing or adding predictor variables using F and T-tests. The removal or addition is usually done based on the statistical tests and estimated coefficients. The chi-square test results show the value is based on the ability to predict the complications values with and without input features. Logistic regression assumes that the resulting logit transformation is linear and that the logarithmic curve in the result doesn’t include outliers.

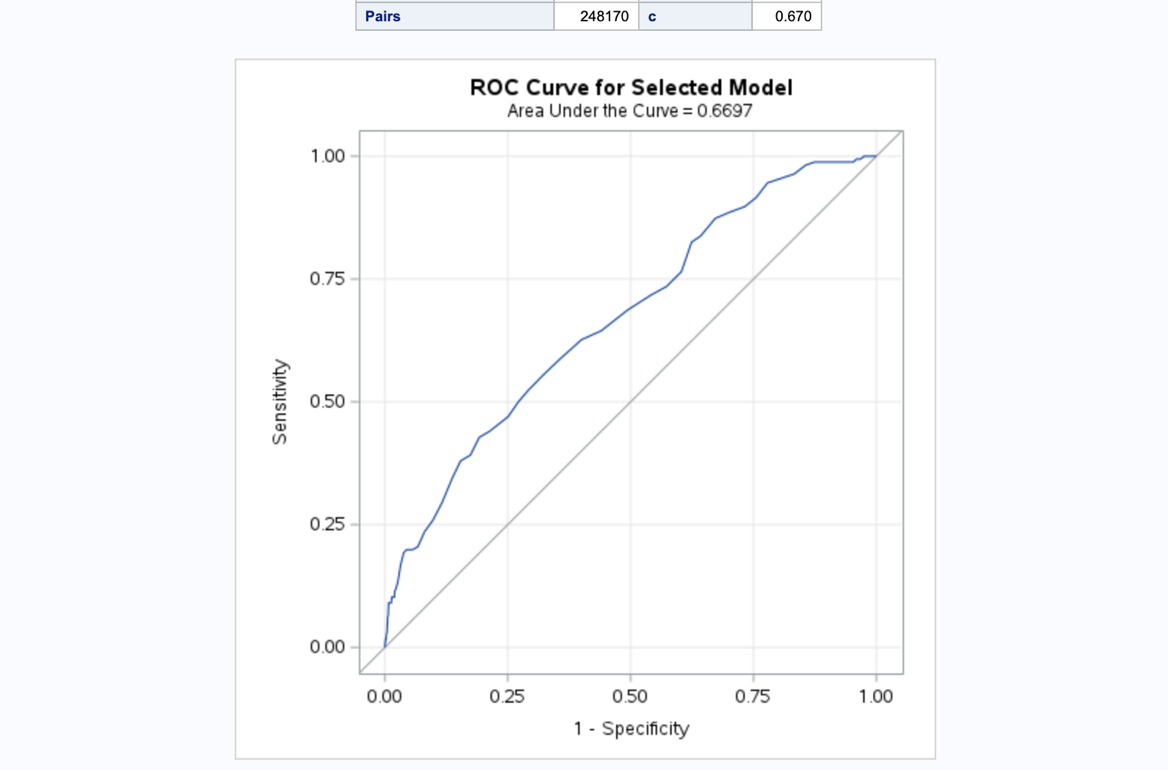
**Figure (Screenshot) 3**

*Logistic regression model*

**

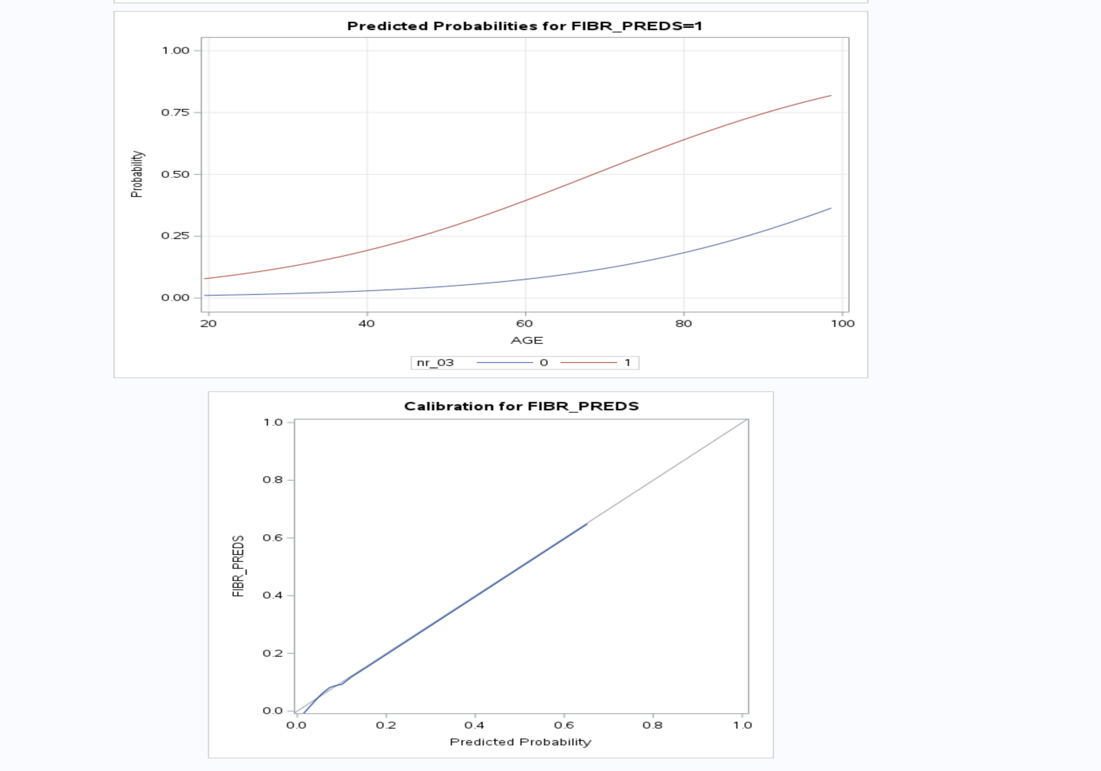
**Figure (Screenshot) 4**

*ROC Curve for selected model*

**

The ROC curve explains the model's performance with the AUC value and the aggregate value of the performance. This tells how much the model is capable of class distinction. If the AUC is higher, the model can distinguish the patients with complications with the input features. The value .6697 shows that the predictions are 66% correct.

**Figure (Screenshot) 5**

*Predicted probabilities*

**Conclusion**

The dataset proposed for this project contains a CSV file for the Myocardial Infarction complication prediction database collected in Krasnoyarsk. The goal of the portfolio project is to analyze and process the data and forecast these risks accurately will help effective patient care with cost-effective treatment. Two business questions were identified, and the null and alternate hypotheses were stated. Descriptive statistics and visual analysis were performed to examine the data. Correlation analysis with scatter plots was performed to identify the relationship between explanatory variables and the target variables. Using the selected explanatory variables, a logistic regression model was built to predict heart disease. The null hypothesis for the two business questions was rejected because there was statistical significance using p-values of less than 0.05. Two input features, age, and nr\_03, were identified as significant in explaining the complications.

**References**

Cai, R., Wu, X., Li, C., & Chao, J. (2020). Prediction models for cardiovascular disease risk in the hypertensive population: a systematic review. *Journal of Hypertension, 38*(9), 1632-1639.

<https://journals.lww.com/jhypertension/Abstract/2020/09000/Prediction_models_for_cardiovascular_disease_risk.2.aspx>

Čulić, V. (2007). Critical risk factors for myocardial infarction. International journal of cardiology, 117(2), 260-269.

Golovenkin, S. E., Shulman, Gorban, A. N., V. A., & Rossiev, D. A. (1997). Complications of myocardial infarction: a database for testing recognition and prediction. Preprint.

Hauck Jr, W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the american statistical association, 72*(360a), 851-853.

Heather, L. C., Hafstad, A. D., Halade, G. V., Harmancey, R., Mellor, K. M., Mishra, P. K., ... & Ussher, J. R. (2022). Guidelines on models of diabetic heart disease. American Journal of Physiology-Heart and Circulatory Physiology, 323(1), H176-H200.

Mythili, T., Padalia, N., Mukherji, D., & Naidu, A. (2013). A heart disease prediction model using SVM-decision trees-logistic regression (SDL). *International Journal of Computer Applications, 68*(16).

<https://www.researchgate.net/profile/Mythili-Thirugnanam/publication/273261237_A_Heart_Disease_Prediction_Model_using_SVM-Decision_Trees-Logistic_Regression_SDL/links/5cd3c7bf92851c4eab8c563b/A-Heart-Disease-Prediction-Model-using-SVM-Decision-Trees-Logistic-Regression-SDL.pdf>

Paneth, N., & Susser, M. (1995). Early origin of coronary heart disease (the “Barker hypothesis”). Bmj, 310(6977), 411-412.

# Rahim Rasool (2019, Nov 4). *Logistic Regression in R: A Classification Technique to Predict*

# *Credit Card Default.* <https://online.datasciencedojo.com/blogs/logistic-regression-in-r-a-classification-technique-to-predict-credit-card-default>

# Ryan Kelly (2014, June 23). Rpubs. *Classification* <https://rpubs.com/ryankelly/ml_logistic>

SAS Help Center. (2021). [Base SAS Procedures Guide](https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/proc/titlepage.htm): Base SAS® 9.4 Procedures Guide, Seventh Edition.

https://documentation.sas.com/doc/en/pgmsascdc/9.4\_3.5/proc/titlepage.htm

Tsien, C. L., Fraser, H. S., Long, W. J., & Kennedy, R. L. (1998). Using classification tree and logistic regression methods to diagnose myocardial infarction. In *MEDINFO'98* (pp. 493-497). IOS Press.

https://ebooks.iospress.nl/volumearticle/19224