##### **Heart Failure Risk Analysis in Support of Health/Life Insurance Underwriting**

*business objective*

Group 2 Quant Consulting (G2QC). has been asked to provide an analytical review and quantitative method in support of the client, Mike Lee, and Associates Insurance LLC (M.L.&A.), for the purposes of supporting risk based underwriting decisions of health and life insurance application approval and pricing. The underlying data set to support this endeavor is publicly available and can be found at <https://www.kaggle.com/fedesoriano/heart-failure-prediction>.

*Data overview*

The data set is comprised of 11 predictor features and a binary target (heart disease) with 918 observations. The objective of this analysis is to classify individuals into a group that is expected to suffer from heart failure or those that are not. Each observation is a cross section of medical records belonging to a unique individual. Table 1.1 provides feature names and definitions.

Table 1.1

| **Target Variable** | **Definition** |
| --- | --- |
| Heart Disease | Has the patient suffered from heart failure? |
| **Predictor Variables** | **Definition** |
| Age | Age of Patient |
| Sex | Gender of Patient (F/M) |
| Chest Pain Type | ATA (Atypical Angina), NAP (Non Anginal Pain), TA (Typical Angina), ASY (Asymptomatic) |
| Resting BP | Resting Blood Pressure of the patient |
| Cholesterol | Cholesterol of the patient |
| Fasting Blood Sugar | Binary Response indicating if a patients’ fasting blood sugar level is greater than 120 mg (1/0) |
| Resting ECG | Normal (no abnormalities), ST (patient exhibits ST abnormality), LVH (patient exhibits left ventricular hypertrophy) |
| Max Heart Rate | Maximum tested heart rate of the patient |
| Exercise Angina | Does the patient exhibit exercise induced angina? (Y/N) |
| Old Peak | The patient's previously recorded ST level peak |
| ST Slope | Is the patient’s ST level increasing, decreasing, or level? (up, down, or flat) |

*Data pre-processing & cleaning*

The data in its raw state contained five variables (‘Sex’, ‘Chest Pain Type’, ‘Resting ECG’, ‘Exercise Angina’, and ‘ST Slope) that were coded as alpha-numeric strings. To prepare the data for modeling (which requires all predictors to be recorded as either continuous or categorical numeric values), we have transposed these features into sets of categorical variables. For example, the ‘Sex’ feature in the raw data was recorded as ‘F’ for female patients and ‘M’ for male. We transposed that into two categorical features ‘F’ (1=Female, 0=Male) and ‘M’(1=Male, 0=Female). We then dropped one of the features from the ‘Sex’ parent group to avoid perfect multicollinearity[[1]](#footnote-0). We followed this methodology for all five alpha-numeric features to produce a usable data set from which to build our classification models covered later in this summary.

A preliminary review of the data found that of the 918 observations, 172 (18.7%) of the entries for the cholesterol feature were coded as ‘0’. Though, as cholesterol (in living patients) must be a positive value, we have treated these observations as incorrectly coded. To address this miscoding and retain the otherwise reliable data for all other features, we have imputed values by replacing observations where cholesterol equals zero with the mean value of cholesterol grouped by age and gender.

We have confirmed that cholesterol (aside from the miscoded values) is normally distributed and is therefore a good candidate for imputing missing values without introducing bias to our analysis. Additionally, as both age and gender (male) are positively correlated with heart disease we can control for some of the variance within the distribution of cholesterol. The resultant distribution following the imputation of cholesterol observations yielded a normally distributed feature, comparable to the distribution (excluding miscoded values).

*preliminary analysis*

An exploratory review of the data finds that Exercise Induced Angina and Old ST Peak are the features that are most positively correlated with heart disease, while ST Slope Up, Max Heart Rate, and Atypical Angina are the features most negatively correlated with heart disease. Table 1.2 has been included in our review to provide visual confirmation that the features are correlated with heart disease in the manner that one would expect.

Table 1.2: Feature Correlation with Heart Failure

| **feature** | **+** | **feature** | **-** |
| --- | --- | --- | --- |
| Exercise Induced Angina | 0.49 | ST Slope Up | -0.62 |
| Old Peak | 0.40 | Max HR | -0.40 |
| Male | 0.31 | Atypical Angina | -0.40 |
| Age | 0.28 | Non Anginal Pain | -0.21 |
| Fasting Blood Sugar | 0.27 | Typical Angina | -0.05 |
| ST Slope Down | 0.12 |  |  |
| Resting Blood Pressure | 0.11 |  |  |
| ST Abnormality | 0.10 |  |  |
| Cholesterol | 0.09 |  |  |
| Left Ventricle Hypertrophy | 0.01 |  |  |

*model development process and selection*

G2QC analysts developed three unique classification models, which were each peer reviewed for methodological soundness and performance. Classification methods were chosen by G2QC as the preferred technique since the target variable (heart disease) is binary. The three models employed for review were Decision Tree, Naïve Bayes, and Logistic Regression.

Prior to model implementation, the data set was split into training and test sets. Of most importance is how the model(s) perform against the test set, as this will provide the closest comparison to real life circumstances. The performance metrics of interest are Accuracy, Precision, and Recall[[2]](#footnote-1). Table 1.3 provides those performance scores for each model against the test population.

Table 1.3

|  |  | No Heart Disease | | Heart Disease | |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | Precision | Recall |
| Decision Tree | 82.2% | 82.7% | 75.2% | 81.9% | 87.7% |
| Naïve Bayes | 84.1% | 81.3% | 82.6% | 86.2% | 85.1% |
| Logistic Regression | 86.0% | 82.0% | 87.0% | 90.0% | 86.0% |

G2QC understands that M.L.&A. will need to balance the profit motive faced by any firm within a market economy against the ethical concerns of providing a vital service to as many interested parties as possible. In an ideal setting, accessibility to health care would not be based on ability to pay. However, the realities of the day are that access to health care is directly related to ability to pay. Therefore, G2QC evaluated the three proposed models in a head-to-head ‘horse race’ to identify the model that best suits M.L.&A. in its effort to optimize profit maximization and social conscientiousness.

With these competing priorities in mind, G2QC proposes that M.L.&A. is best served by taking on conservative underwriting principles that places the highest weight on accurately predicting the precision of individuals that will not suffer from heart failure (precision of those without heart disease), coupled with the accurate prediction of those that will suffer from heart failure (precision of those with heart disease). In so doing, M.L.&A. will set itself on a path to fiscal responsibility, which G2QC believes, will yield the best results in attempting to optimize the profit maximizing motive of all firms against a socially conscientious constraint to serve as many clients as fiscally responsible.

*findings*

As seen in Table 1.3, the Decision Tree model rank ordered the highest in terms of precision when predicting those that will not suffer from heart failure, but only by .7 percentage points and was the strongest of performers predicting recall of those with heart failure. The logistic regression was the strongest performer in terms of precision of those expected to suffer from heart disease (8.1 percentage points greater than the decision tree) as well as the recall of those expected to not suffer heart failure, and overall accuracy. The Naïve Bayes model did not score as the preeminent model in any performance metrics.

Figure 1: Logistic regression coefficients

Within the logistic regression model, the features that most contributed to the accuracy of the predictions are: Old peak, chest pain type (ATA and NAP), and ST slope.

*Conclusion*

Based on the overall performance of the logistic regression, G2QC will provide M.L.&A. a production modeling environment and end user interface to support the real time evaluation of Health and Life insurance application approval and pricing decisions based on the classification of the candidate as an individual that is expected to suffer from heart failure or not. To process the evaluation, M.L.&A. will need the applicant to provide current information on the 11 predictor features outlined in Table 1.1. Furthermore, this model can also be used to monitor existing customers to determine if they are at risk so that early preventative actions can be recommended. By doing so, G2QC and M.L.&A. are charting a path to a profitable partnership for both parties that each can be confident that the profits are complementary to an ethically sound business model.

1. Perfect Multicollinearity occurs when one feature is a perfect linear function of another, which leads to unreliable prediction estimates [↑](#footnote-ref-0)
2. Accuracy = (True Positive + True Negative) / (True Positive + False Positive + True Negative + False Negative)

   Precision = True Positive / (True Positive + False Positive)

   Recall = True Positive / (True Positive + False Negative) [↑](#footnote-ref-1)