**PROJECT REPORT**

**Heart Failure Prediction and Analysis**

***Group Members:***

*Divya Rajesh Kannan*

*Aditya Chaganti*

*Shobha Ojha*

*Atharv Joshi*

**1. Executive Summary**

**1.1 Project Overview**

This project is aimed at developing a robust classification model for identifying individuals at risk of experiencing a heart attack. Cardiovascular diseases, including heart attacks, continue to be a major global health concern. Early detection and targeted intervention are key to reducing heart attack-related morbidity and mortality. This project utilizes R, a powerful statistical programming language, to analyze a comprehensive dataset containing the relevant medical variables.

**1.2 Objectives**

● To build a classification model using R that accurately assesses whether an individual is at elevated risk of experiencing a heart attack or not

● To identify and quantify the most significant risk factors associated with heart attacks

● To contribute valuable insights to the field of cardiovascular disease prevention through data-driven analysis

**1.3 Methodology**

Statistical techniques for classification models will be employed using R, including Logistic Regression and Decision Trees, to analyze the chosen dataset. The dataset encompasses twelve medical attributes, including Age, Sex, RestingBP, ChestPainType, Cholesterol, FastingBS, RestingECG, ExerciseAngina, Oldpeak and ST\_Slope. Extensive data preprocessing, feature selection and model evaluation techniques will be applied to ensure the model's accuracy and robustness.

**2. Background**

**2.1** **Domain**

Healthcare Industry - Cardiovascular Health and Risk Assessment

**2.2**  **Brief description of the scenario**

In the field of cardiovascular health and risk assessment, the Heart Attack Prediction Project is a critical initiative that aims to use advanced data analytics and machine learning techniques to improve the early detection and prediction of heart attacks. Timely identification of those at risk and those with early symptoms is crucial to prevent heart attacks and reduce mortality. The project uses the power of this data to develop predictive models to identify people at high risk of heart attack in the near future. The scope of the data analysis further extends to:

* *Data Collection:* Extensive and diverse data from various sources, including patient records, medical images and genetic data
* *Machine learning models:* developing and training machine learning algorithms to analyze collected data. These models learn patterns and correlations in data to identify individuals at risk of heart attack
* *Real-time monitoring:* Integrating predictive models into healthcare systems to enable real-time monitoring of patients' health status
* *Patient education:* development of educational materials and tools that inform patients about heart attack risk factors and prevention strategies

Overall, the Heart Attack Analysis and Prediction Project has the potential to save lives by enabling healthcare professionals to proactively identify people at risk of heart attack, ultimately reducing the burden of cardiovascular disease and improving public health and quality of life.

**2.3** **Decisions of Interest**

The decisions of interest for this project mainly include the business problems we are aiming to solve, and the consequent data-driven decisions we can make using them in the healthcare sector:

* Identifying individuals who are at an elevated risk for heart failure
* Analyzing the factors that could lead to heart failure

These decisions play a key role in the success of the project, ensuring that it not only provides accurate predictions but also translates predictions into data-driven decisions such as effective health interventions and improved patient outcomes.

**2.4**  **Decision makers**

When making key decisions, it is important to consider the various factors that may influence both the choice of machine learning models and the interpretation of dataset variables.

* Understanding the clinical significance of variables and risk factors is important for making informed decisions about feature selection, model development, and thresholds for risk stratification
* Data cleaning and preprocessing decisions, such as handling missing values ​​or values, can significantly affect a model
* Decision makers should prioritize characteristics historically known to be strong predictors of heart attack risk, such as age, gender, blood pressure, cholesterol, smoking status, family history and medical history
* In addition to raw data, decision makers should evaluate whether derived traits or transformations of variables (e.g., body mass index of height and weight) can improve the predictive power of the model
* Statistical tests and techniques should be used to determine the significance of each variable with respect to the target variable (heart attack risk)
* Validation strategy: Decisions about dividing the data into training, validation and test sets and the choice of cross-validation methods can affect model reliability and validity
* Model Interpretability: Interpretability of the chosen model is an important consideration, especially in healthcare, where understanding why the model makes certain predictions is crucial for clinical application
* Ethical and legal considerations: Decision-makers must ensure that model predictions and recommendations are consistent with ethical principles and laws related to patient privacy, informed consent, and discrimination

**3. Business Understanding**

**3.1 Business Objectives**

● *Early Intervention:* When it comes to heart disease, early detection is key, so an accurate prediction model can prevent heart attacks, repeat events, severe complications and reduce fatalities

● *Reduced Hospital Readmissions:* Hospitals can reduce readmission, which is beneficial to both the patient and hospital, as they can avoid penalties due to high readmission rates

● *Optimized Resource Allocation:* By focusing on the needs of patients at high risk, hospitals can allocate their specialized equipment, staff, and facilities optimally yet efficiently

● *Enhanced Patient Trust:* Proactive rather than reactive approaches to healthcare can increase the trust and loyalty of a patient, which subsequently enhances the reputation of the hospital

● *Cost Savings:* Prevention of heart failure, its complications and readmission lead to significant cost saving for hospitals due to reduced Emergency Room visits and extended procedures

**3.2 Situation assessment:**

According to the World Health Organization, heart disease is the leading cause of death globally, with almost 17.9 million deaths per year due to cardiovascular disease. It is reported that the number of deaths from cardiovascular disease has grown by 60% in the last 30 years. Whether this increase is purely due to a population boom, or other underlying factors is worth investigating. Furthermore, in America, the estimated direct cost of cardiovascular disease treatment is $226.2 billion, as of 2017-2018. This is a huge chunk of the annual healthcare costs of the country, so more accurate diagnostic procedures are needed to detect it earlier, thereby reducing costs and improving the patient’s quality of life.

**3.3 Data Mining Goals:**

● The major goal is to use the different medical attributes to try to identify any patterns in the data that would explain an underlying relationship to cardiovascular disease causation

● Using a combination of these medical attributes, an accurate and precise prediction model would be developed to assess whether a patient is prone to cardiovascular disease or not

**4. Data Understanding**

**4.1 Data Description**

We have one primary dataset for our business analytics project which has information about people with cardiovascular disease or who are at high cardiovascular risk. It has a total of 12 attributes which includes patient’s information like age, sex, and information about parameters such as restingBP, Cholesterol, MaxHR, etc. These parameters when trained in a machine learning model will help us to measure the risk factors such as hypertension, diabetes, hyperlipidaemia, or already established disease and detect which patients in the dataset are prone to a heart attack.

**4.2 Data Dictionary**

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Data Type** | **Data Description** |
| Age | num | Age of the patient (in years) |
| Sex | chr | Sex of the patient (M: Male, F: Female) |
| ChestPainType | chr | Chest pain type [TA: Typical Angina - chest pain, ATA: Atypical Angina - not normal signs, NAP: Non-Anginal Pain - not caused by angina, but pain from other issues, ASY: Asymptomatic - no chest pain] |
| RestingBP | num | Resting Blood Pressure [mm Hg] (normal is 120/80) |
| Cholesterol | num | Serum Cholesterol [mm/dl] (<120 is desired) |
| FastingBS | num | Fasting Blood Sugar [1: if FastingBS > 120 mg/dl, 0: otherwise] (1 is bad) |
| RestingECG | chr | Resting Electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria] |
| MaxHR | num | Maximum heart rate achieved [Numeric value between 60 and 202] (normal HR varies with age) |
| ExerciseAngina | chr | Exercise-induced Angina [Y: Yes, N: No]  (Chest pain from exercise) |
| OldPeak | num | ST [Numeric value measured in depression]  0 = baseline, >0 means there is a depression (not good) |
| ST\_Slope | chr | The slope of the peak exercise ST segment [Up: up sloping, Flat: flat, Down: down sloping] |
| HeartDisease | num | Output Class [1: heart disease, 0: Normal] |

**4.3 Sources and other information**

**Data source:** Kaggle - it's CC0: Public Domain. Note that the data is taken from Kaggle, but it was collected and reviewed by UCI(University of California Irvine)

**URL:**

<https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction?select=heart.csv>

* **Total observations:** 1190
* **Duplicated observations:** 272
* **Final dataset**: 918 observations
* **Usability rating:** 10.0

**4.4 Data quality**

The success and credibility of any predictive model depend on the quality of the data used for training and testing. In this section, we evaluate the data quality of the heart attack prediction dataset based on key parameters:

1.Accuracy:

● The accuracy of the data is vital for reliable predictions. The dataset was collected and reviewed by UCI, known for its expertise in healthcare data collection and curation

2. Completeness:

● A complete dataset is essential to develop a robust predictive model

● The heart attack prediction dataset is comprehensive, including a wide range of relevant features, such as RestingBP, FastingBS, MaxHR

● Appropriate actions will be taken to handle missing values too

3. Consistency:

● Consistency in data format and structure simplifies data preprocessing and analysis. The dataset maintains consistent formatting and units for all variables

4. Reliability:

● Data reliability is crucial for the model's trustworthiness. The heart attack prediction dataset has a usability rating of 10.0 and is taken from kaggle which is a trusted platform for finding reliable datasets

**5. Data Preparation**

**5.1 Data Selection**

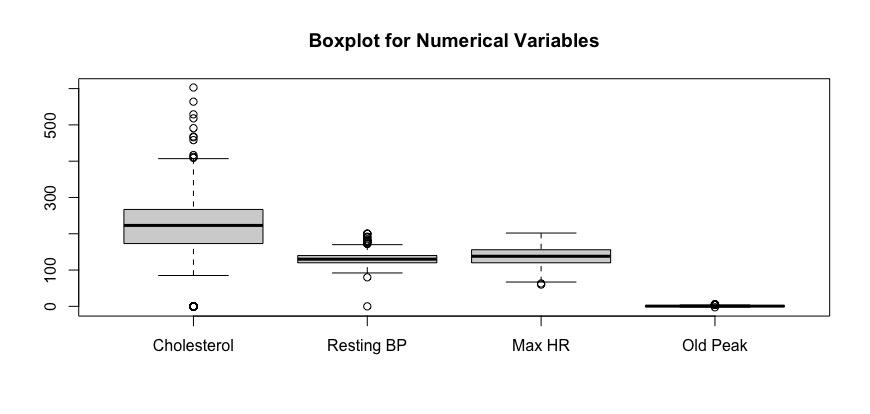
Data selection is a critical step in the data preparationprocess, with a goal of choosing and preparing the data that is most relevant to the business problem to be solved. In this case it is, identifying individuals at a risk of heart failure. This dataset from Kaggle is chosen as it consists of combining the common attributes of 5 different datasets, which has made it the largest heart disease dataset available so far for analysis.

**5.2 Data Preparation**

* ***Target Variable Definition:*** The target variable is a binary variable ‘Heart Disease’. A value of 1 indicates that the individual is at elevated risk for a heart attack while a value of 0 indicates that they are not at elevated risk for a heart attack
* ***Attribute Datatype Identification:*** The target variable, Heart Disease amongst others such as Sex, Chest Pain Type, Fasting Blood Sugar, Resting ECG, Exercise Angina, ST\_Slope are categorical variables that are misrepresented as a numeric or character data type
* ***Handling Categorical Variables:*** The datatypes of these variables are converted to the ‘factor’ data type as classification models require categorical variables for proper encoding

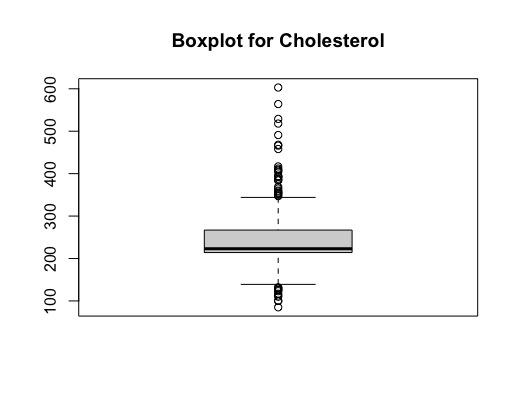
**5.3 Data Cleaning**

* ***Checking for missing values:*** The dataset is checked for missing values using the is.na() function but none are found
* ***Checking for duplicate values:***The dataset is also checked for duplicate values using the unique() function, but none are found
* ***Checking for Outliers:*** A side-by-side boxplot is constructed using all the numerical variables, except Age, to identify whether any outliers are present in the data. From initial analysis, it is seen that the Cholesterol variable has a high number of outliers. On closer analysis, it is seen that there are an unusually high number of 0’s. The normal range for cholesterol is <120 mm/dl, but it can be inferred from that instead of using NA’s, missing values have been placed with 0’s. To fix this, the outliers are replaced with NA’s and the imputation strategy is done using the median of the non-missing values (median = 223)



***Figure 5.1 Boxplot of all the numerical Variables***

Usually, data with any outliers would be treated using the imputation or omission strategy. However, since it is medical data, it is important to take the nature and context of the data into account. In medical data, the outliers can represent valid and abnormal health conditions that need medical attention, hence treating it as an outlier would not make sense as valuable data would be lost. For cholesterol, the normal value is less than 120 mm/dl. Anything greater than this is a sign of high cholesterol. Resting blood pressure is in the normal range when it is 120/20 (systolic = 120, diastolic = 80), so variations are a sign of high/low BP. Similarly, the normal range for MaxHR varies with age, so it is not correct to treat the range of MaxHR in this dataset as outliers, since age varies from 28 to 77 in it. Old Peak is a term for ST Depression. The baseline is 0 which indicates a healthy heart, but anything greater than 0 i.e., a deflection from the baseline, shows ST depression, which is a sign of the muscle not receiving enough oxygen, amongst other factors.



***Figure 5.2 Boxplot of Cholesterol without invalid values***

**6. Modeling**

**6.1 Data Description**

The final cleaned dataset ‘heart.csv’ obtained from Kaggle has a total of 918 observations and 12 attributes. The target variable chosen is Heart Disease, a binary response variable. There are 11 predictor variables, with 5 numerical variables and 6 categorical variables.

**6.2 Decision-Making Models**

* The problem is a classification problem as the target variable is a binary categorical variable as there are only two possible responses - 1: risk of heart failure, 0: not at risk of heart failure
* With this in mind, the chosen classification models are Logistic Regression and Classification Tree

**6.3 Model Choice Rationale**

**Logistic Regression:**

* Simple and interpretable
* Efficient in small datasets
* Probabilistic output that provides probabilities of class membership and respective confidence
* Computationally efficient and can train models quickly
* Widely used and accepted in the medicine field

**Classification Tree:**

* Provides a measure of feature importance
* Robust to outliers and missing values
* Handles both numerical and categorical data
* Ease of handling non-numeric data
* No assumption of linearity

Though both classification methods have advantages, the final choice of model for this dataset will depend on performance measures such as accuracy, sensitivity, and specificity.

**6.4 Model Development**

**Model-1: Logistic Regression**

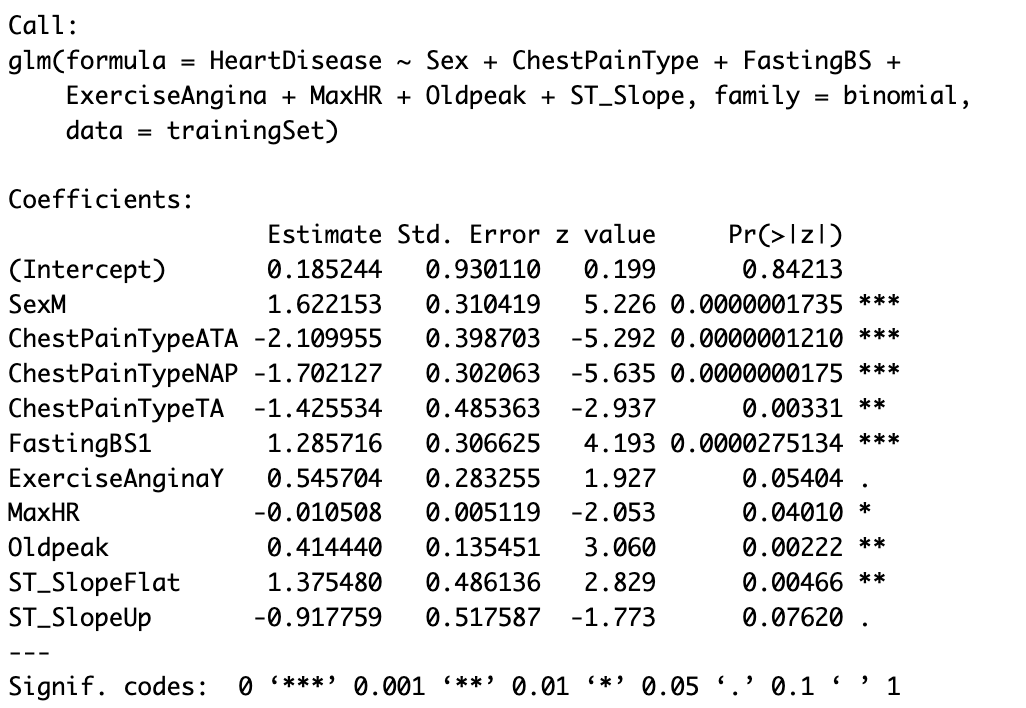
* The training set for this model contains 70% of total observations in the dataset, while the validation set contains the remaining 30% of observations
* This results in 643 observations in the training set and 275 observations in the validation set
* An initial regression model is run using all 11 attributes to find those that are statistically significant. Backward variable selection is then done to choose the statistically significant variables with p-values less than 0.05
* It is found that there are 6 significant variables: including Sex, Chest Pain Type, Fasting Blood Sugar, Exercise Angina, Max HR, Old Peak and ST Slope

A screenshot of a computer

Description automatically generated

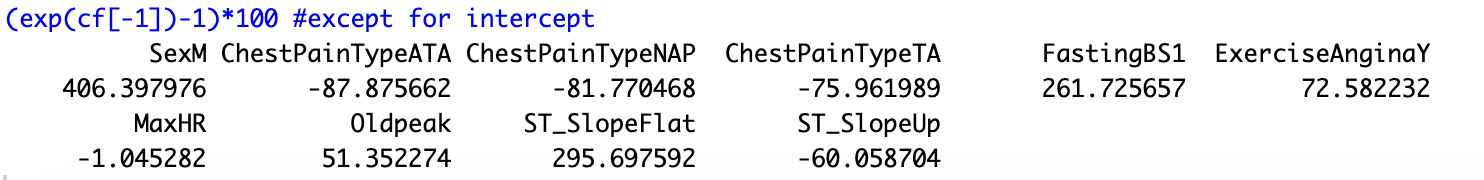
***Figure 6.1 Statistically Significant variables at p < 0.05 using all predictor variables***

* A logistic regression model is then run using only the significant variables. It is verified that all the variables used in this model are statistically significant at p < 0.05



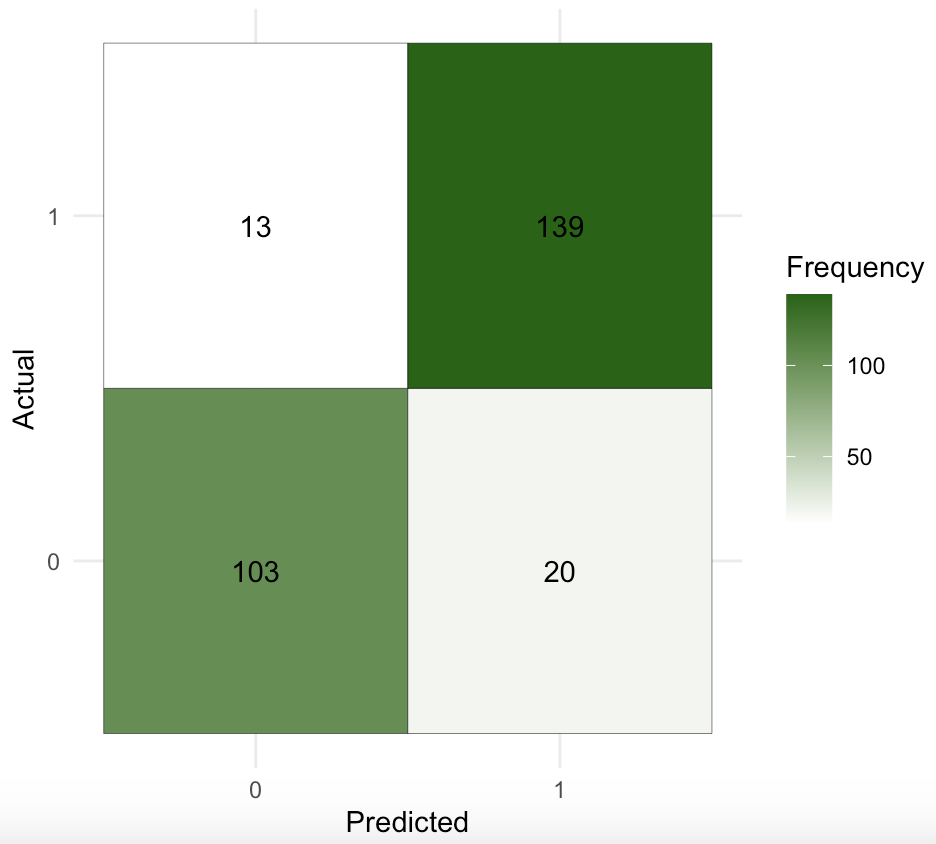
***Figure 6.2 Logistic Regression Model-1 with only statistically significant variables***

* The coefficients of the logistic regression model are found and the partial effect of each predictor variable on the odds is interpreted.
  + Males are over 406% more susceptible to heart disease than females
  + Having high fasting blood sugar (i.e greater than 120 mg/dl) gives a 261% more chance to have heart disease than those with normal fasting BS
  + Those with Exercise Angina, which is chest pain due to exercise, have a 72% more chance to have heart disease than those who do not
  + Those with OldPeak, or ST Depression > 0, which is abnormal, have 51% more likelihood to be at risk to heart disease
  + Those with a flat ST\_Slope are at 295% more risk to heart disease



***Figure 6.3 Partial Effect of Each Predictor Variable***

* Accuracy, sensitivity & specificity of the training set are:
  + Accuracy: 86.63%
  + Sensitivity: 89.61%
  + Specificity: 82.92%
* The model is then run on the validation set to identify the performance measures of the model
  + Accuracy is 88.00%: this means approx 88.00% of observations are classified correctly
  + Sensitivity is 91.45%: this means approx 91.45% of target class cases are classified correctly
  + Specificity is 83.74%: this means approx 83.74% of non-target class cases are classified correctly
* A confusion matrix is created by comparing the predicted probabilities of the logistic regression model to a default cut-off value of 0.5

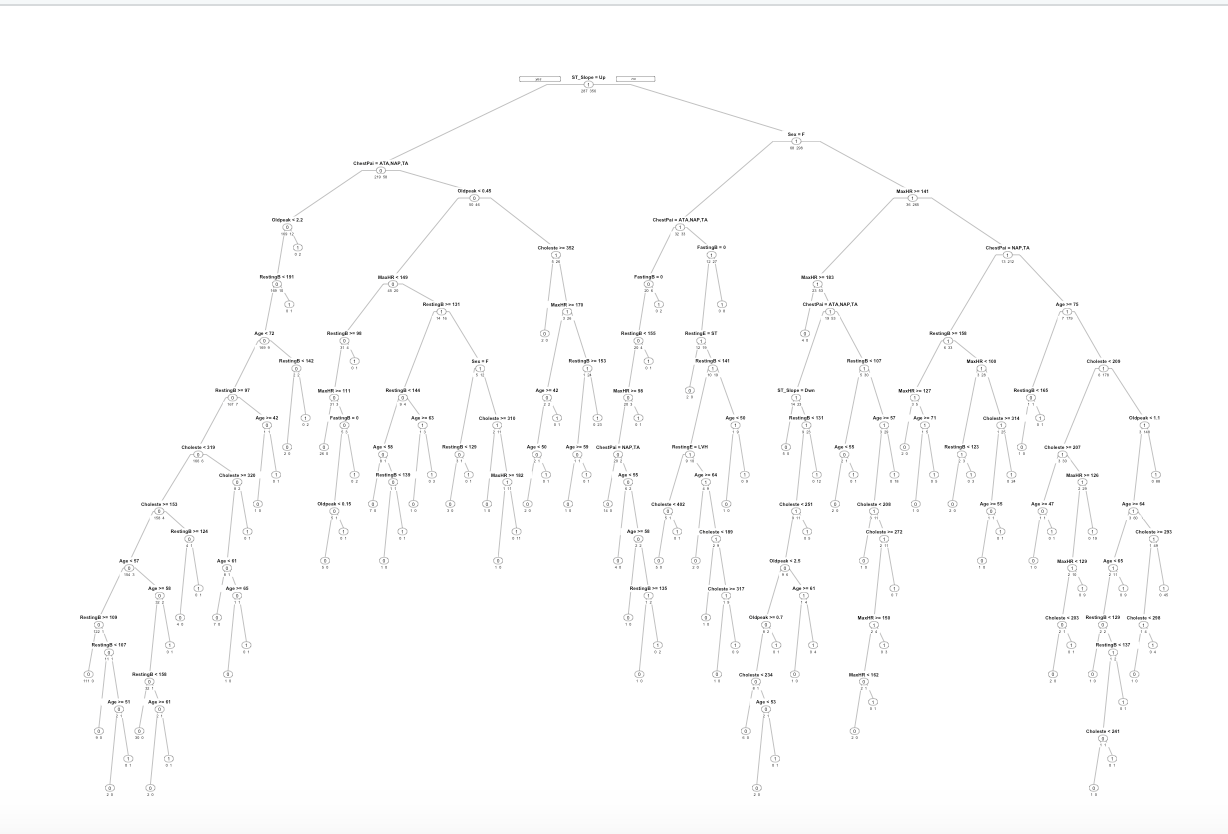


***Figure 6.4 Confusion Matrix of Model-1***

* + Positive Pred Value, also called Precision is 0.8374: this means that approx 83.74% of predicted target case classes belong to the target class
  + From the confusion matrix: TP = 139, TN = 103, FP = 20, FN =13
  + Misclassification rate is 0.12: this means that 12% of observations are incorrectly classified

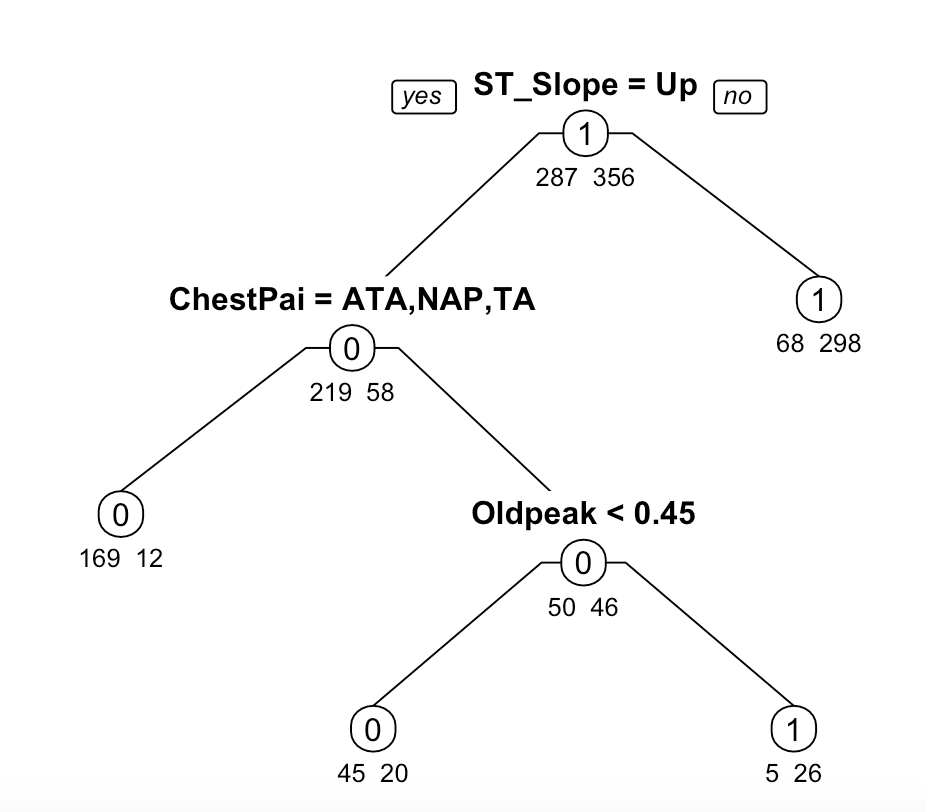
**Model-2: Classification Tree**

* The training set for this model will contain 70% of total observations in the dataset, while the validation set will contain the remaining 30% of observations
* This results in 643 observations in the training set and 275 observations in the validation set
* It is run using all 11 attributes, since classification models are commonly used for feature selection and to measure feature importance
* The full tree is run and continues to split nodes until each leaf node has minimum number of instances until a maximum depth is reached



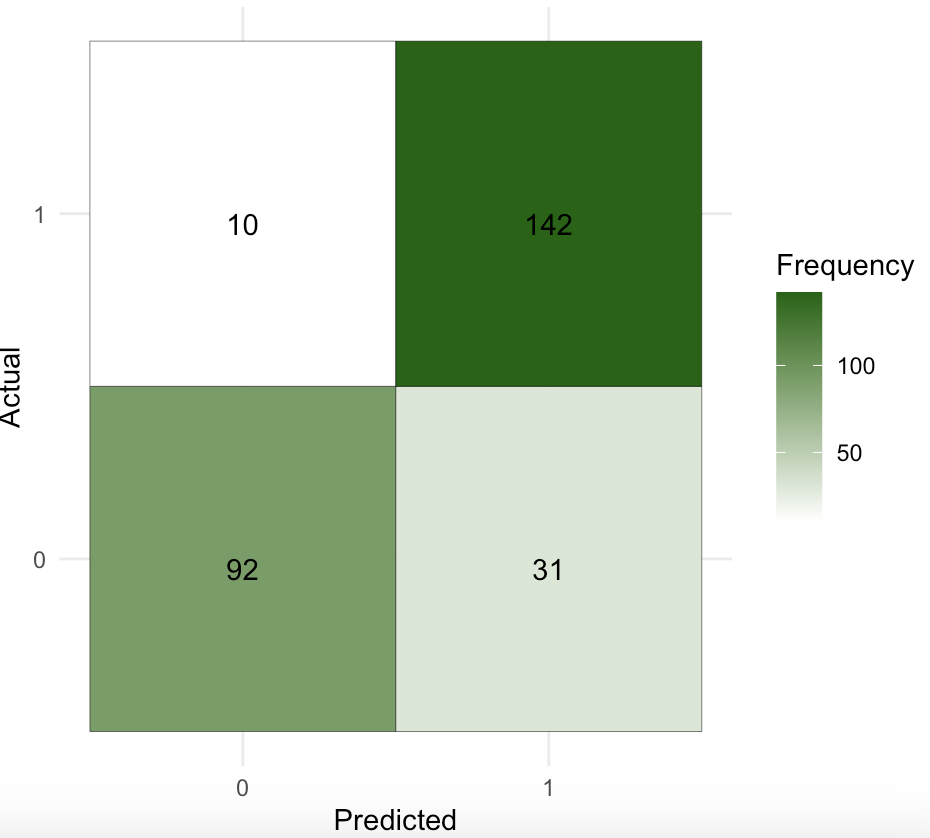
***Figure 6.5 Full Tree of Model-2***

* The resultant full tree has 12 possible complexity parameters, with the minimum error tree having CP = 0.0243902 and xerror = 0.37631 with 3 splits
* For this model, the pruned tree is the same CP = 0.0243902



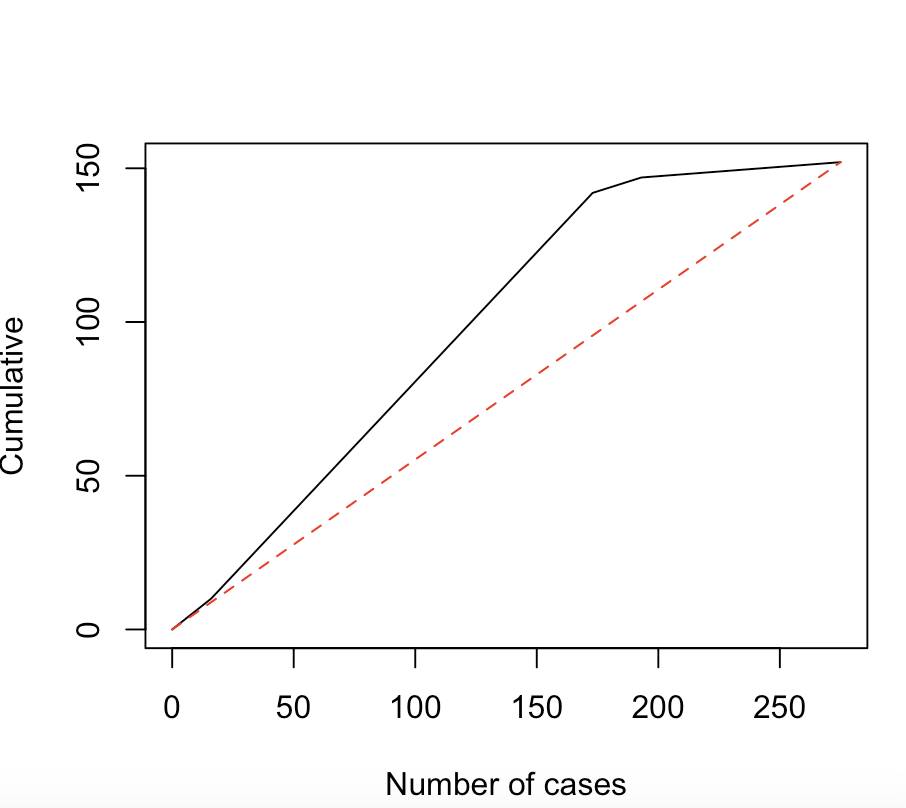
***Figure 6.6 Pruned Tree of Model-2***

* The nodes on the pruned tree are:
  + There is one Root Node: ST\_Slope
  + There are three Interior Nodes: Chest Pain Type, OldPeak,
  + There are eight Leaf Nodes: HeartDisease (1,0), note that the leaf nodes are not pure subsets
* Performance measures are calculated to assess model performance by creating a confusion matrix in the validation set:
  + Accuracy is 0.8509: this means that approx 85.09% of observations are classified correctly
  + Sensitivity is 0.9342: this means that approx 93.42% of target class cases are classified correctly
  + Specificity is 0.7480: this means that approx 74.80% of non-target class cases are classified correctly
  + Positive Pred Value, also called Precision is 0.8208: this means that approx 82.08% of predicted target case classes belong to the target class
* From the confusion matrix: TP = 142, TN = 92, FP = 31, FN = 10
  + Misclassification rate is 0.1490909: this means that approx 14.91% of observations are incorrectly classified



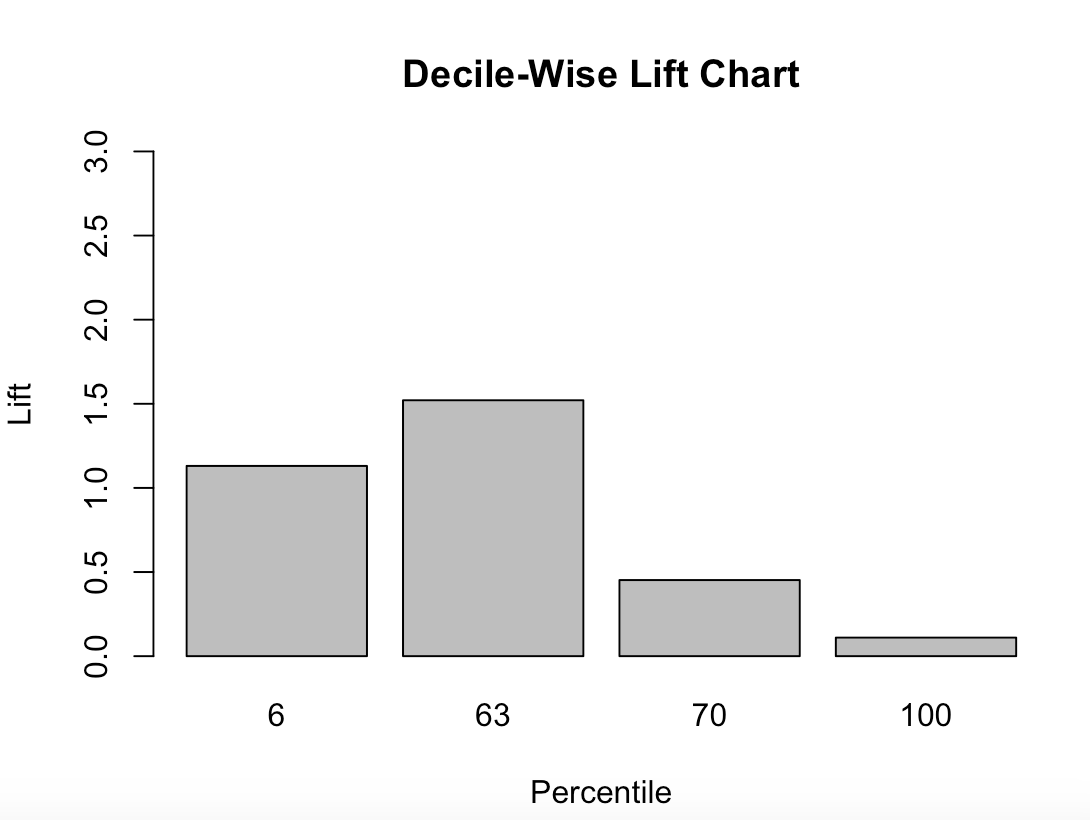
***Figure 6.7 Confusion Matrix of Model-2***

* Since the sensitivity in the model is quite high, the default cutoff of 0.5 is sufficient for finding the predicted probabilities
* To evaluate model performance independent of the cutoff value, the cumulative lift chart, decile chart and ROC Curve are used
  + *Cumulative Lift Chart:* The classification model shows superior predictive power when compared to the baseline model



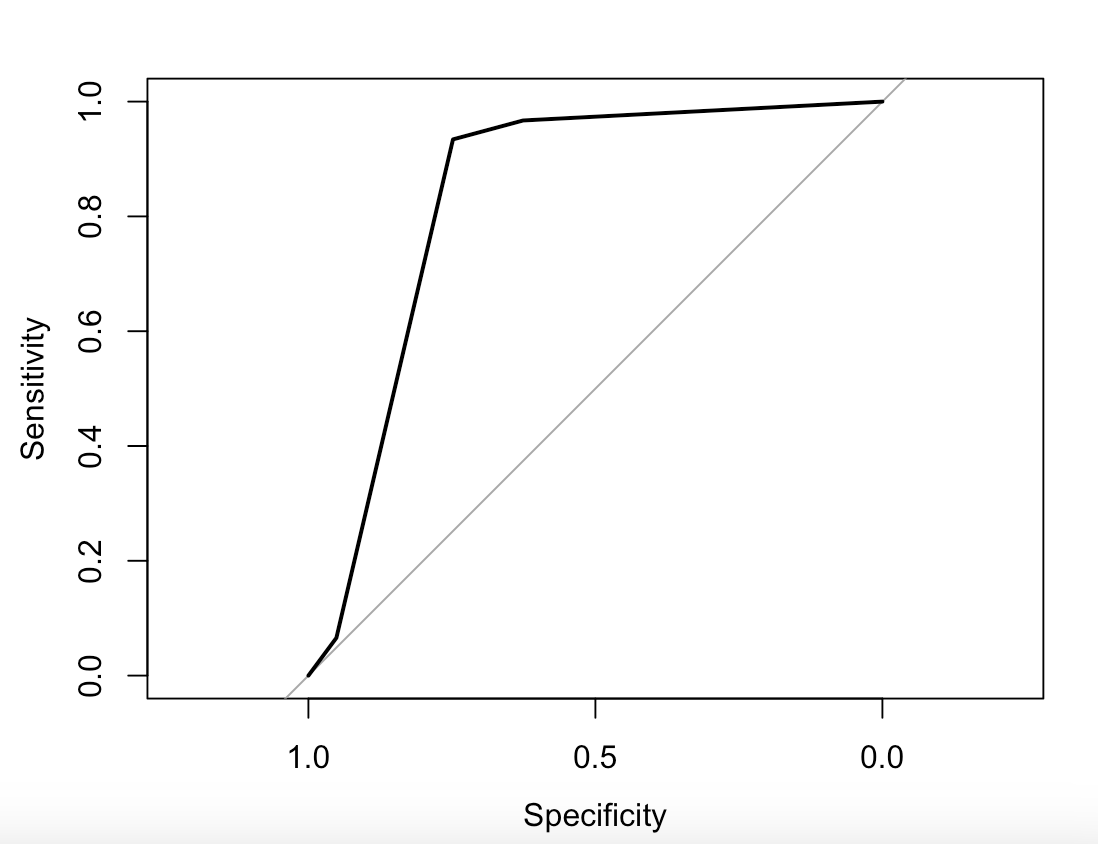
***Figure 6.8 Cumulative Lift Chart of Model-2***

* + *Decile Chart:* The top 6% of individuals with the highest predicted probability of having heart disease can be correctly captured as having heart disease by over 1.2 times as compared to if 6% of individuals are randomly selected



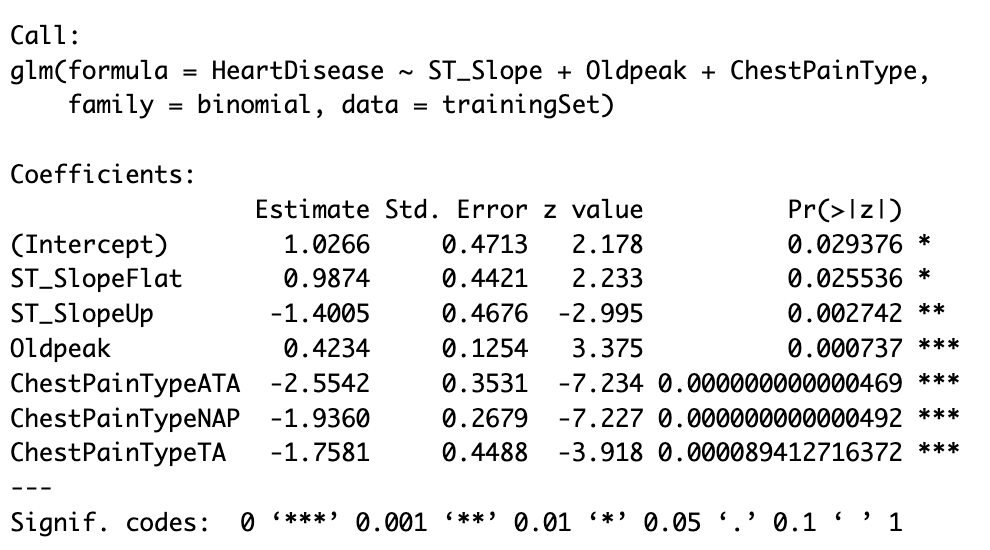
***Figure 6.9 Decile Lift Chart of Model-2***

* + *ROC Curve:* Again, this curve shows that the classification model outperforms the baseline model in terms of sensitivity and specificity across all cutoff values. Area under the curve = 0.8349 also reinforces this finding

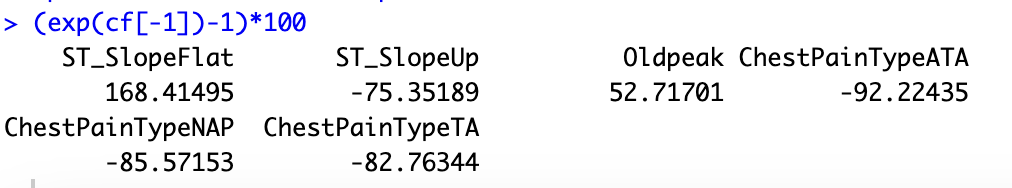


***Figure 6.10 ROC Curve of Model-2***

**Model-3: Logistic Regression**

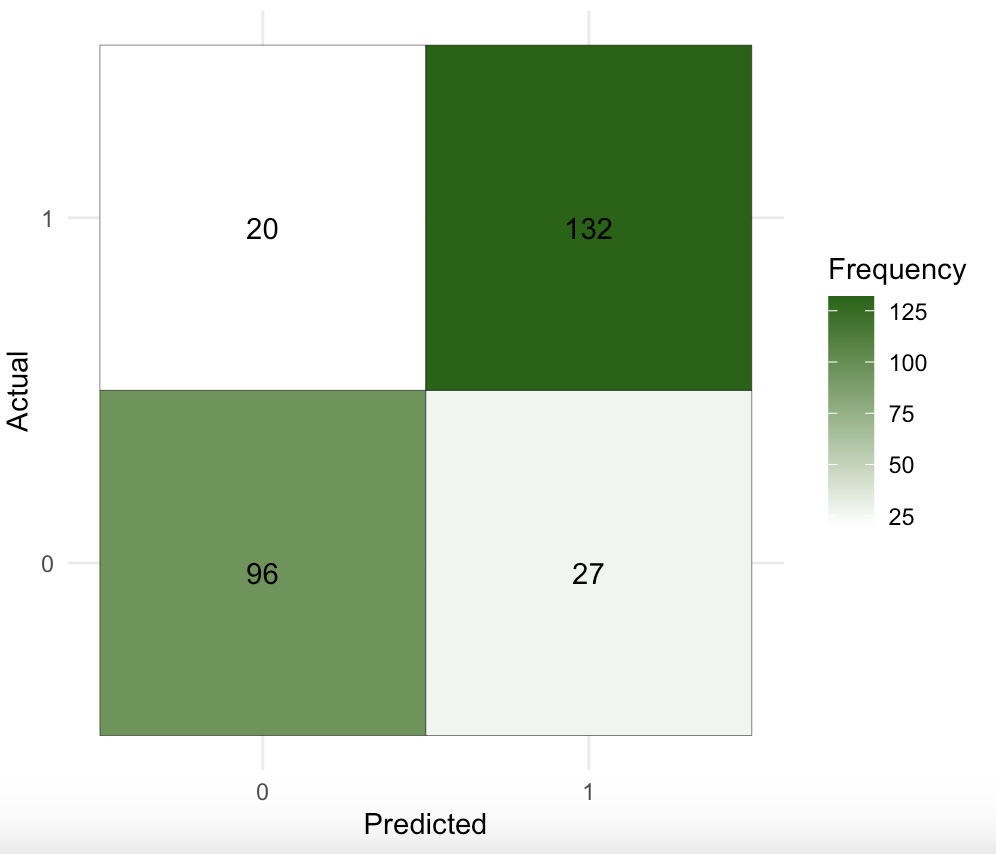
* Looking at the results from model-2, we decided to use the same attributes as shown by the pruned tree in a logistic regression model to see if there is any improvement in the performance measures
* The training set for this model will contain 70% of total observations in the dataset, while the validation set will contain the remaining 30% of observations
* This results in 643 observations in the training set and 275 observations in the validation set
* The selected attributes are all statistically significant at p<0.05 except Resting BP in the training model
* 

***Figure 6.11 Logistic Regression Model-3 with statistically significant variables***

* The coefficients of the logistic regression model are found and the partial effect of each predictor variable on the odds is interpreted:
  + Those with OldPeak, or ST Depression > 0, which is abnormal, have 52% more likelihood to be at risk to heart disease
  + Those with a flat ST\_Slope are at 168% more risk to heart disease

***Figure 6.12 Partial Effect of Each Predictor Variable***

* Accuracy, sensitivity & specificity of the training set are:
  + Accuracy: 83.20%
  + Sensitivity: 86.80%
  + Specificity: 78.75%
* The model is then run on the validation set to identify the performance measures of the model
  + Accuracy is 82.909091%: this means that approx 82.91% of observations are correctly classified
  + Sensitivity is 86.842105%: this means that approx 86.84% of target class cases are correctly classified
  + Specificity is 78.048780%: this means that approx 78.04% of non-target class cases are correctly classified
* A confusion matrix is created by comparing the predicted probabilities of the logistic regression model to a default cut-off value of 0.5

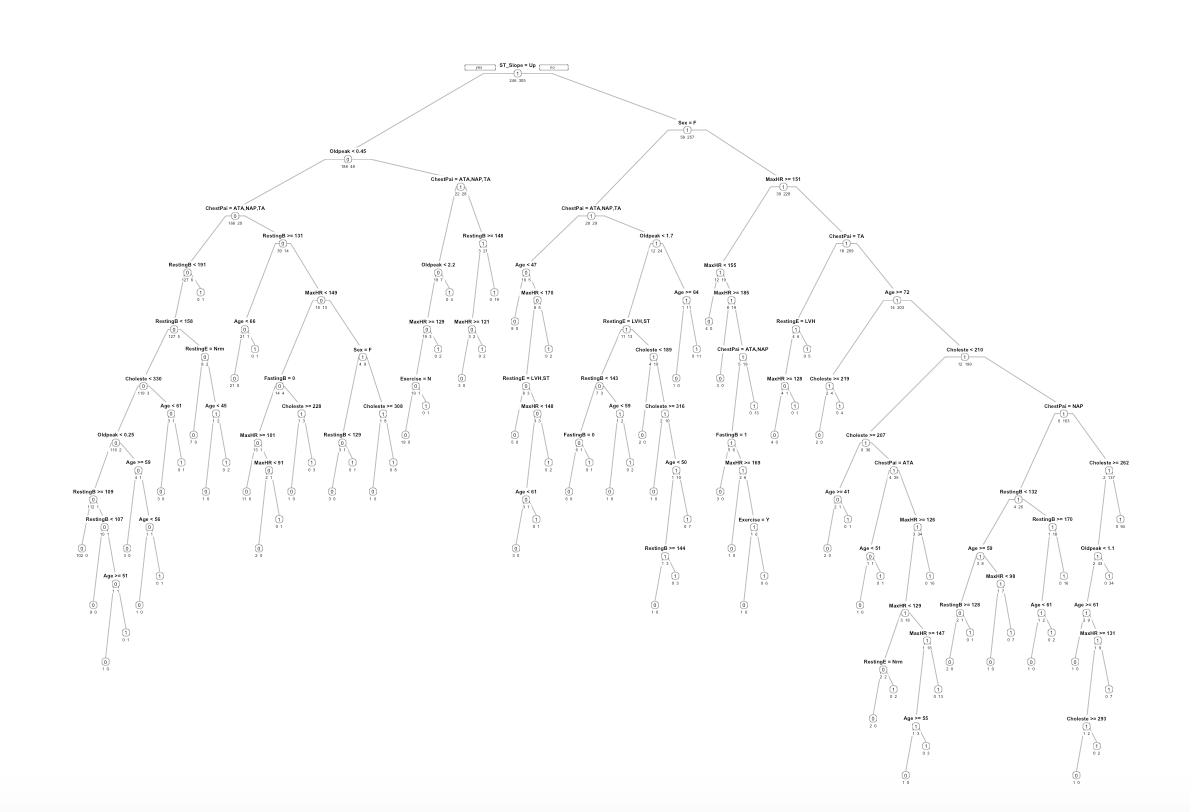


***Figure 6.13 Confusion Matrix of Model-3***

* From the confusion matrix: TP = 132, TN = 96, FP = 27, FN = 20
  + Positive Pred Value, also called Precision is 0.8276: this means that approx. 82.76% of predicted target case classes belong to the target class
  + Misclassification rate is 0.1709091: this means that approx. 17.09% of observations are incorrectly classified

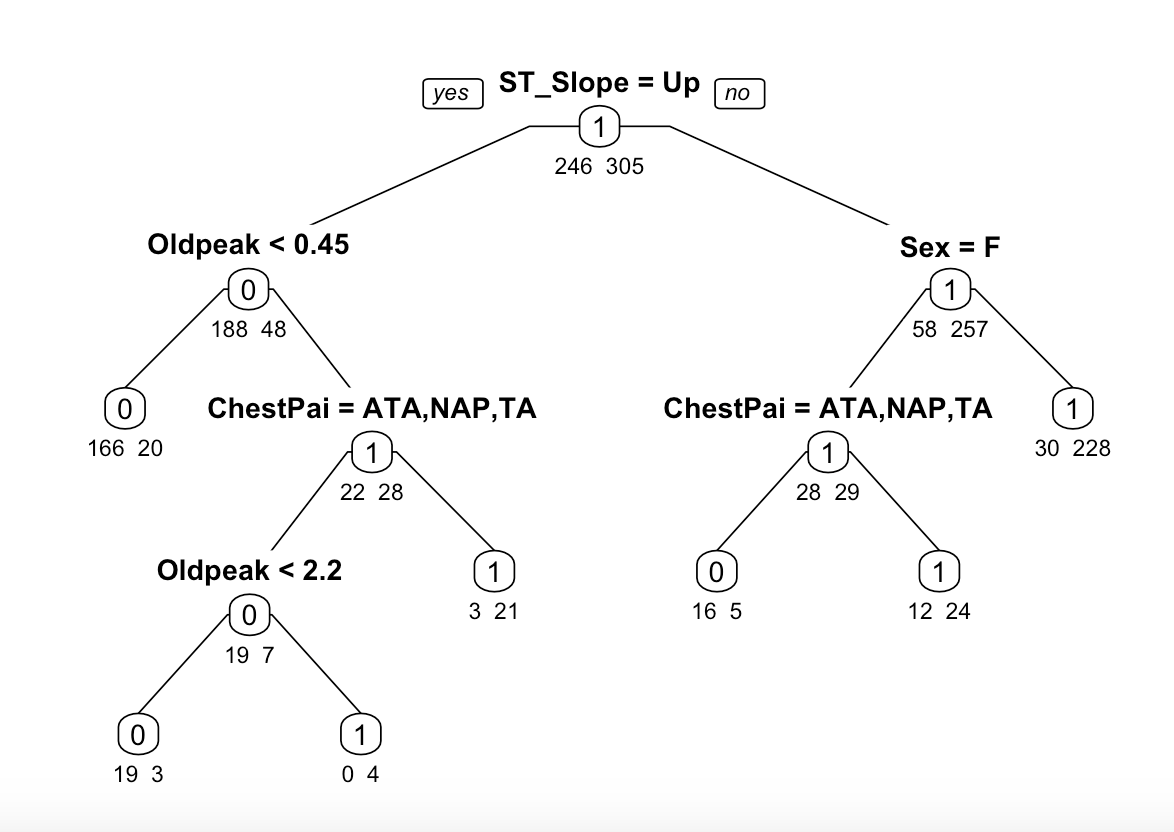
**Model-4: Classification Tree**

* Sincethe dataset has under 1000 observations, and classification trees are usually more efficient with larger number of observations, we decided to Increase the proportion of observations in the training to validation set by 10%, to see if the performance measures can be improved
* The training set for this model will contain 60% of total observations in the dataset, while the validation set will contain the remaining 35% of observations
* This results in 551 observations in the training set and 367 observations in the validation set
* It is run using all 11 attributes, since classification models are commonly used for feature selection and to measure feature importance
* The full tree is run and continues to split nodes until each leaf node has minimum number of instances until a maximum depth is reached



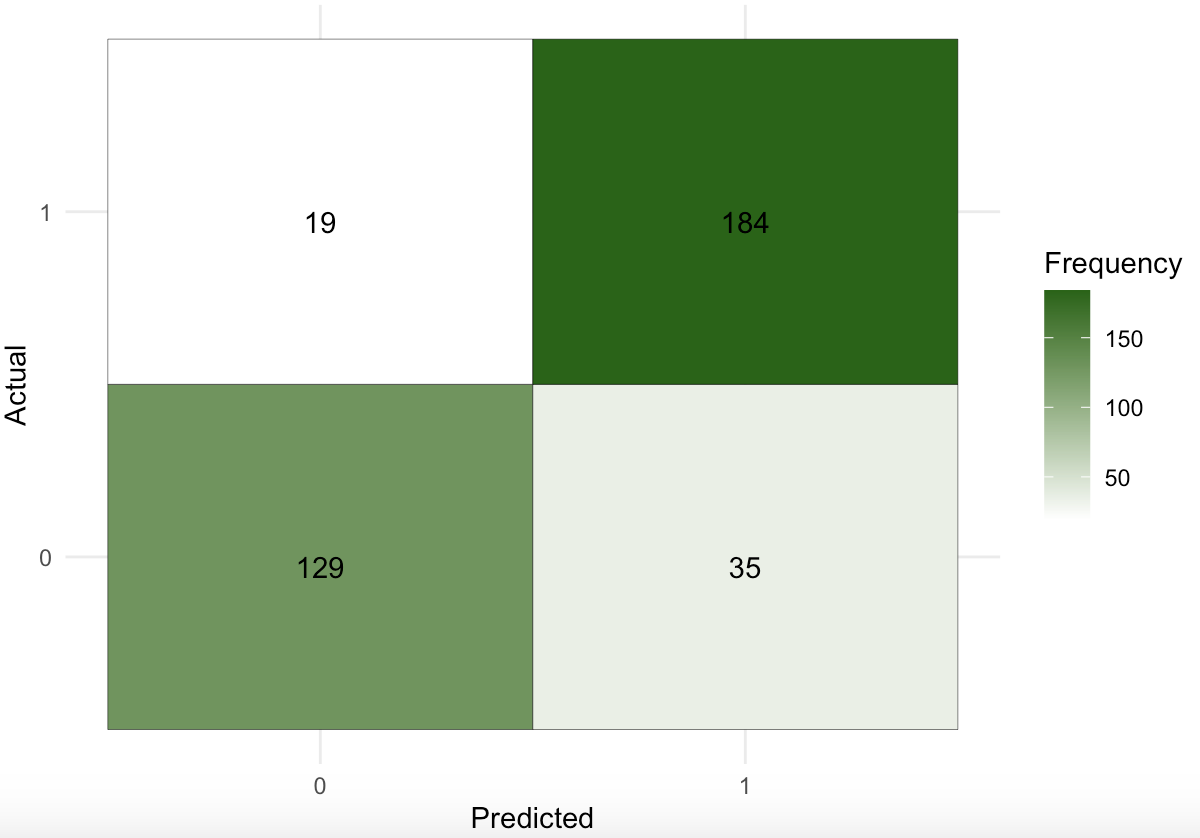
***Figure 6.14 Full Tree of Model-4***

* The full tree has 13 possible complexity parameters, with the minimum error tree having CP = 0.0081301 and xerror = 0.39024 and 6 splits
* For this model, the minimum error tree is also the pruned tree



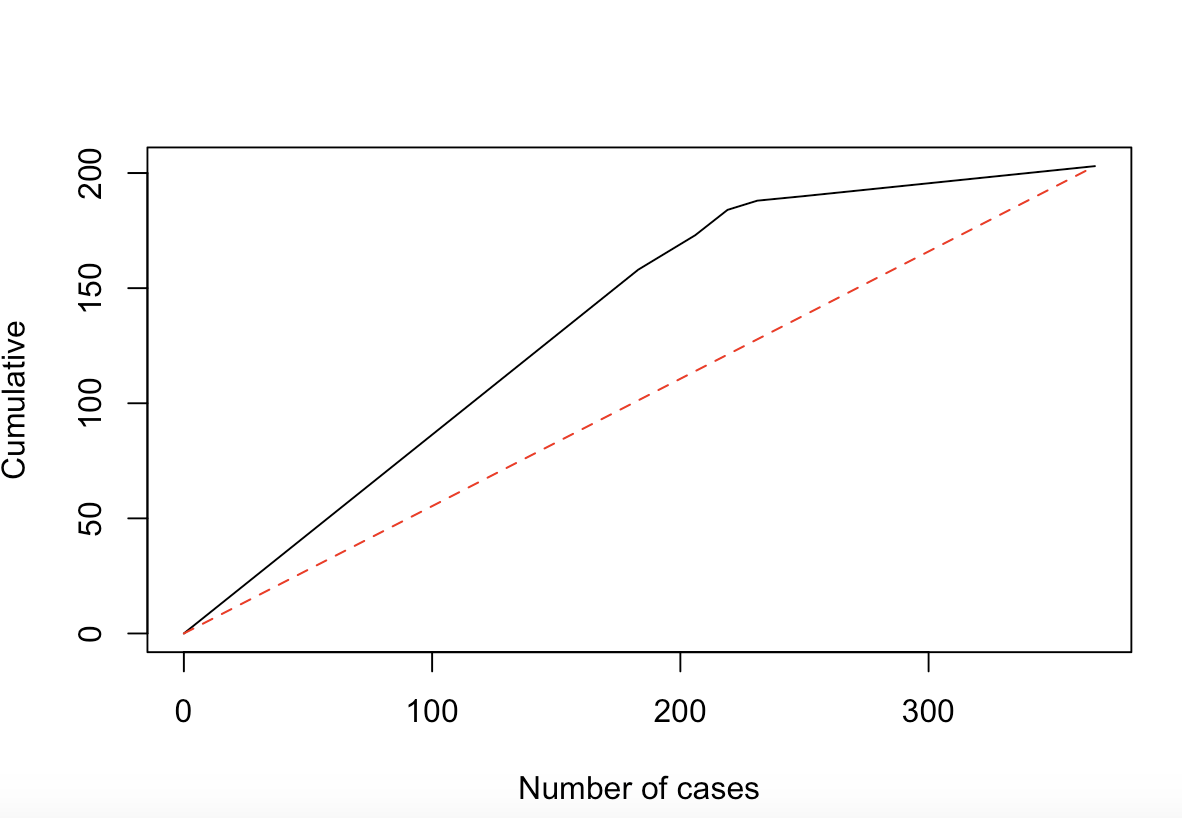
***Figure 6.15 Pruned Tree of Model-4***

* The nodes on the pruned tree are:
  + There is one root node: ST\_Slope
  + There are five interior nodes: ChestPainType, OldPeak, Sex
  + There are seven leaf nodes: HeartDisease (1,0) but note that the leaf nodes are not pure subsets
* Performance measures are calculated to assess model performance by creating a confusion matrix in the validation set:
  + Accuracy is 0.8529: this means that approx 85.29% of observations are classified correctly
  + Sensitivity is 0.9064: this means that approx 90.64% of target class cases are classified correctly
  + Specificity is 0.7866: this means that approx 78.66% of non-target class cases are classified correctly
  + Positive Pred Value, also called Precision is 0.8402: this means that approx 84.02% of predicted target case classes belong to the target class
* From the confusion matrix: TP = 184, TN = 129, FP = 35, FN = 19
  + Misclassification rate is 0.175: this means that approx 17.50% of observations are incorrectly classified

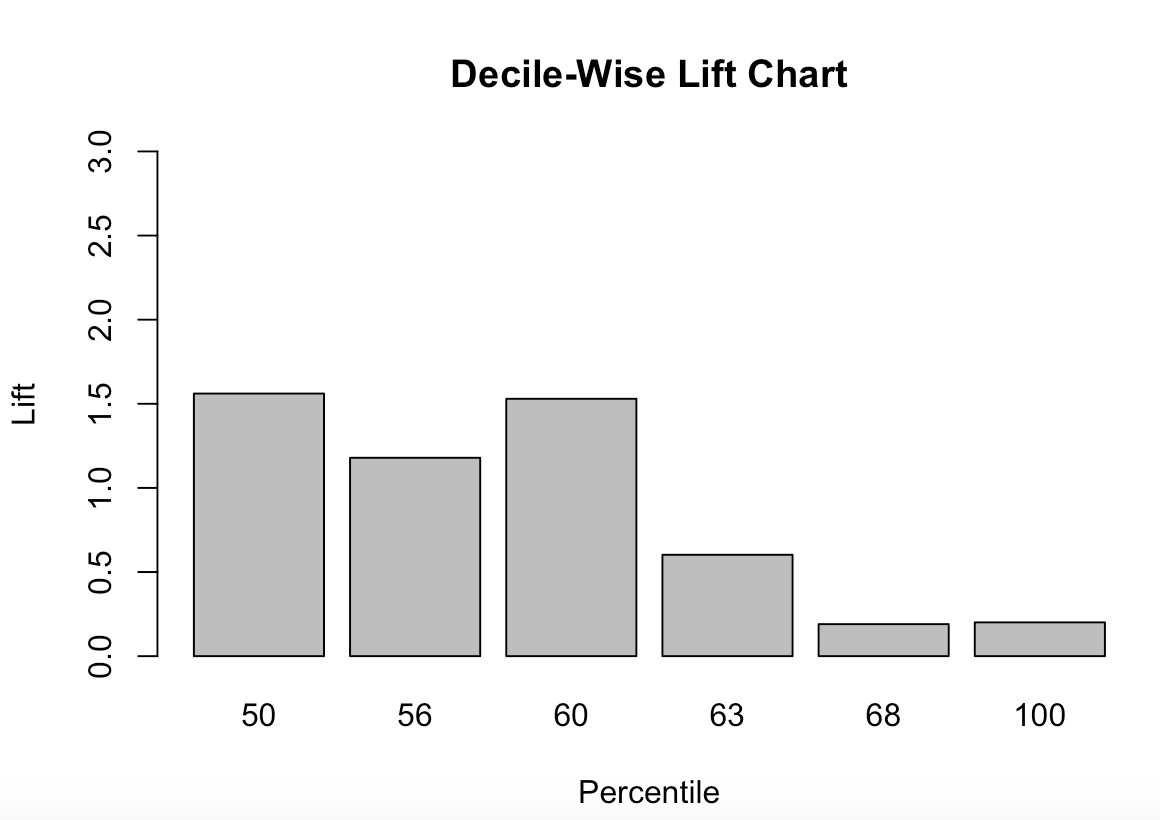


***Figure 6.16 Confusion Matrix of Model-4***

* To evaluate model performance independent of the cutoff value, the cumulative lift chart, decile chart and ROC Curve are used
  + *Cumulative Lift Chart:* The classification model shows superior predictive power when compared to the baseline model

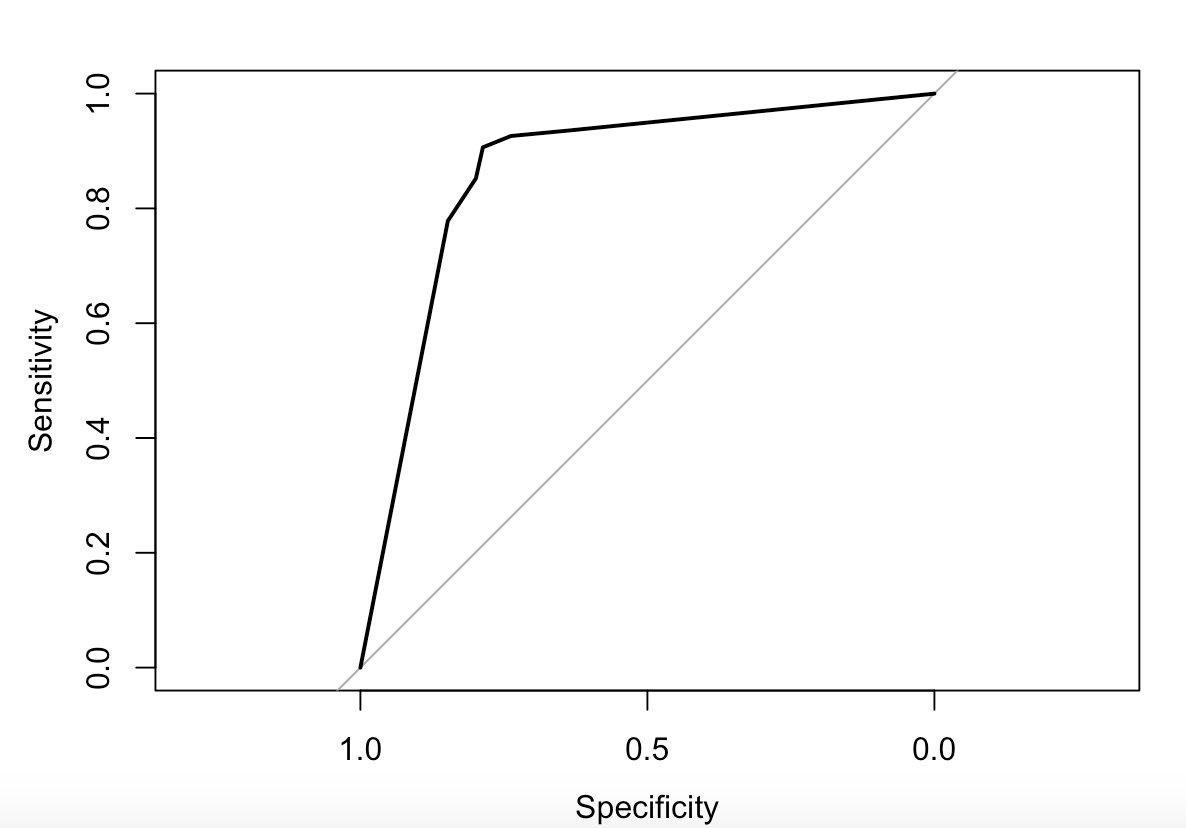


***Figure 6.17 Cumulative Lift Chart of Model-4***

* + *Decile Chart:* The top 50% of individuals with the highest predicted probability of having heart disease can be correctly captured as having heart disease by about 1.5 times as compared to if 50% of individuals are randomly selected.
  + 

***Figure 6.18 Decile Lift Chart of Model-4***

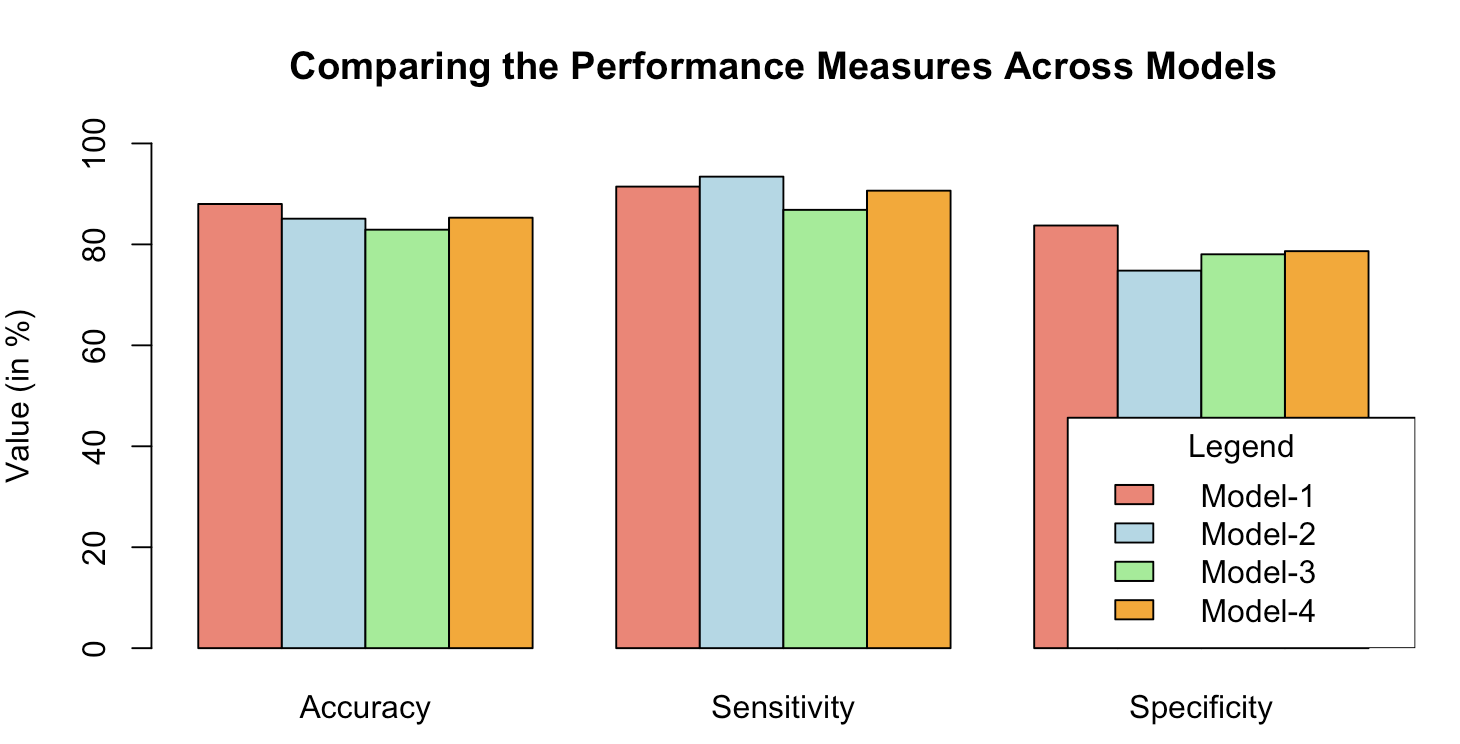
* + *ROC Curve:* Again, this curve shows that the classification model outperforms the baseline model in terms of its sensitivity and specificity across all cutoff values. Area under the curve = 0.8649 also reinforces this finding.



***Figure 6.19 ROC Curve of Model-4***

**7. Model Evaluation**

A bar plot is used to visualize the major performance measures - accuracy, sensitivity, specificity - across the 4 models to determine which is the most preferred model.



***Figure 7.1 Bar plot of Model Evaluation***

* Accuracy is highest in Model-1 at 88%
* Sensitivity is highest in Model-2 at 93.42%
* Specificity is highest in Model-1 at 83.74%

**8. Discussion**

1. **Model Recommendation:**

Based on the models built, I would recommend ***Model-1*** as the recommended model. Accuracy of predicting in the validation set is the most important measure of a robust model, and Model-1 has the highest accuracy at 88%. The sensitivity and specificity of Model-1 are also relatively high at 91.45% and 83.74% respectively. Keeping in mind that the business problem is to identify those with an elevated risk of heart disease, sensitivity is a very important measure, as it measures the proportion of correctly classified target cases i.e., those who are at risk for heart disease. Thus, Model-1 is recommended.

1. **Model Limitations:**

* *Dataset size:* With just over 900 observations, the model does not have as many observations as conventional datasets, which may affect performance
* *Outliers:* The significance of outliers can only be validated completely by a medical professional, so in a business scenario, validating the assumption of using the outliers would be done during the data-cleaning process
* *Potential Bias:* There is no mention of the demographics of the individual’s data is collected from in the dataset, which can cause inaccurate predictions if blindly applied as it is an observational study and analysis

**c. Cognitive Bias & Mitigation:**

Cognitive bias can arise at any stage of model development, from data collection to model evaluation. Some of the common kinds of bias that we have tried to avoid during the building of the classification models and interpretation of the results are:

* *Confirmation Bias:* Preference to information from pre-existing beliefs, such as knowing that certain attributes like age, gender and blood pressure are historically relevant in the context of heart disease. To mitigate this, a diverse dataset was used and inferences were made directly from the data independently of pre-existing facts
* *Overfitting Bias:* Fitting the model too closely to the training data, which captures the noise of the model and affects the ability of the model to predict new observations accurately. To avoid this, we pruned the classification tree and also tried to change the ratio of training to validation data to check whether performance improved
* *Anchoring Bias:* Giving undue importance to the order in which data is placed. To avoid this bias, we used the set.seed() function to randomly choose observations for the training and validation set

**d. Enhancements for Better Decision Support:**

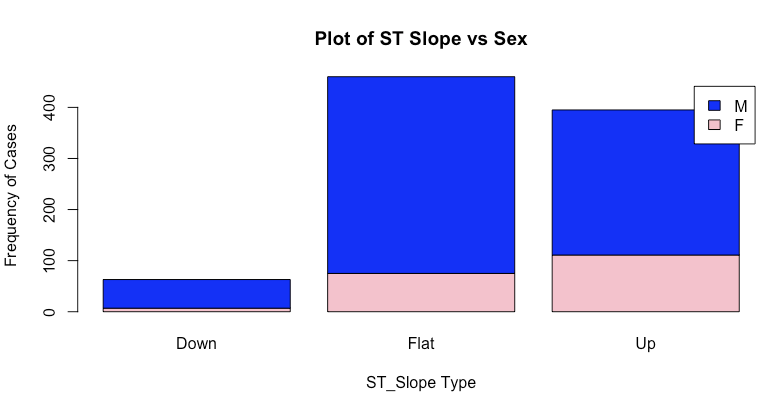
* *Comprehensive Data Integration:* Integrating more diverse datasets to gain a more holistic view of the decision context, in this case, individuals at risk for heart failure
* *Explainable AI:* Using AI to enhance the interpretability of the model, to increase trust in the decision support system, since decisions regarding medical advice have to be highly accurate to avoid severe consequences
* *Interactive Visualization:* Adding visualizations and tools to allow the user to interact with the data, decision-making process and better understand complex information
* *Real-time Insights:* Developing an interface to aid users getting accurate real-time insights could be potentially lifesaving
* *Collaboration with Medical Professionals:* To validate the findings of the models with historical data and to understand the intricacies of the medical data, collaboration with medical professionals is necessary
* *Data Governance & Ethical Considerations:* Patient data is highly sensitive, so proper data governance and ethical considerations for fair decision support to avoid bias is critical

To use this model in a real-world context, beyond an observational study, regular assessments and subsequent updates are essential to ensure that the decision support systems continue to meet the evolving needs of patients and medical professionals.

**Appendix:**

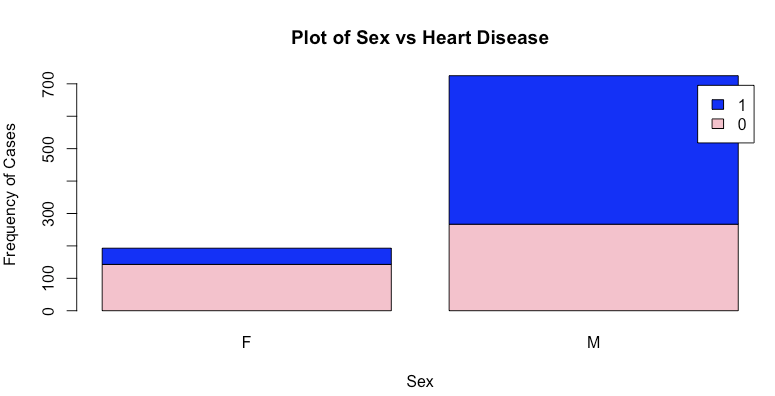
Some of the visualizations found from plotting the different statistically significant attributes against each other:

***Stacked column chart with sex v ST slope***



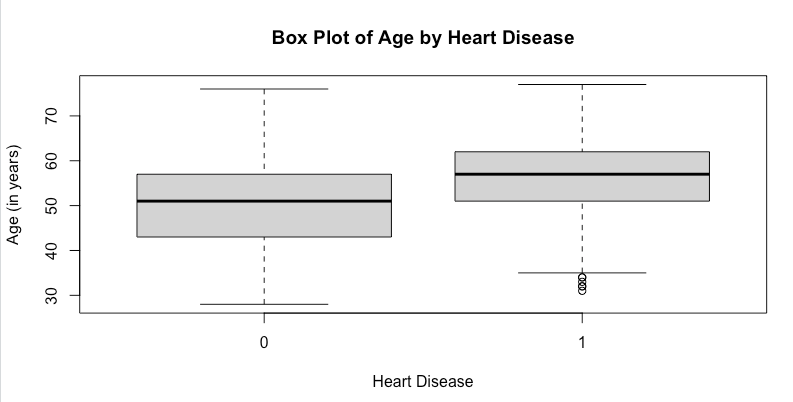
* Of the 57.1% of individuals with abnormal ST\_Slopes i.e., down sloping & flat, 48.03% of them are male

***Stacked column with heart disease v sex***



* 49.8% of people having heart disease are males, in comparison to 5.44% females - almost 10 times as likely!
* Males are more susceptible to heart disease

***Box plot of age v heart disease***



* The mean age of those with heart disease is higher than those without any heart disease.
* Those who have heart disease but are younger (less than 35) are shown as outliers, which agrees with the historical trend of older people usually being at risk of heart attacks

**References:**

1. <https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1>
2. <https://world-heart-federation.org/news/deaths-from-cardiovascular-disease-surged-60-globally-over-the-last-30-years-report/#:~:text=Search%20for%3A%20Search-,Deaths%20from%20cardiovascular%20disease%20surged%2060%25%20globally,the%20last%2030%20years%3A%20Report&text=GENEVA%2C%2020%20May%202023%20%E2%80%93%20Deaths,World%20Heart%20Federation%20(WHF).>
3. <https://www.ahajournals.org/doi/full/10.1161/CIR.0000000000001052>
4. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1906611/>