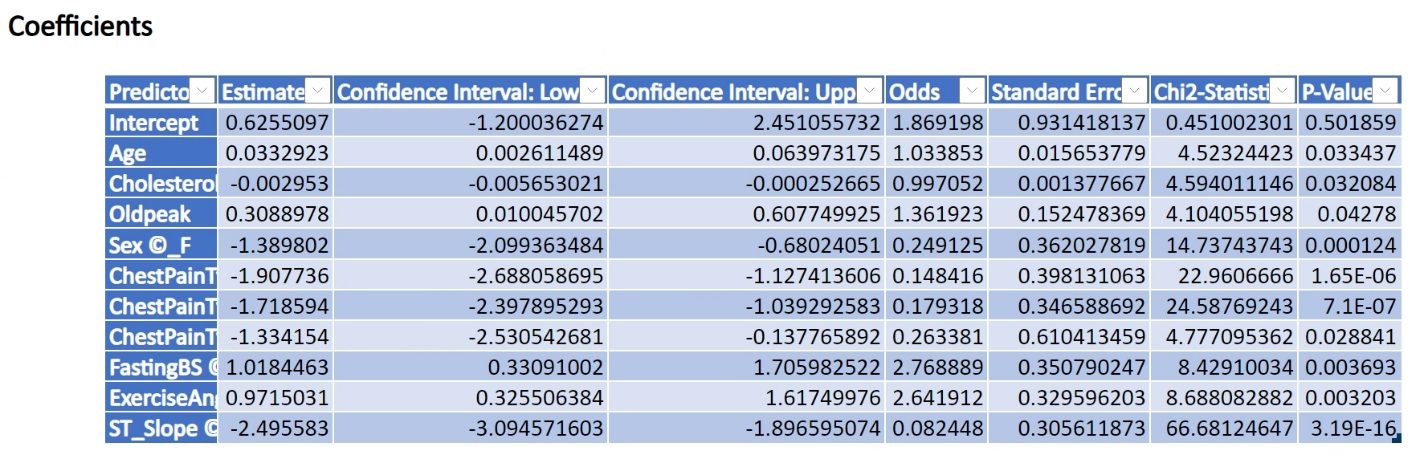
**Project Milestone 3**

**Logistic Regression:**

We ran the logistic regression and ran a feature selection step with the best subset, and these are the significant variables we landed on:  
Age, Cholesterol, Oldpeak, Sex ©\_F, ChestPainType ©\_ATA, ChestPainType ©\_NAP, ChestPainType ©\_TA, FastingBS ©\_1, ExerciseAngina ©\_Y, ST\_Slope ©\_Up

The new cutoff that we chose after analyzing the initial decile chart is **0.5133**. So after running the logisitic regression again with significant variables and the new cutoff value, the derived equations would be:



The model’s equations are

logit (Heart\_Disease=1) = 0.63 +0.033(Age) -0.003(Cholestrol) +0.31(Oldpeak) -1.39(Sex\_F) -1.91(ChestPainType\_ATA) -1.72(ChestPainType\_NAP) -1.33(ChestPainType\_TA) +1.02(FastingBS\_1) +0.97(ExerciseAngina\_Y) -2.5(ST\_Slope\_Up)

Odds (Heart\_Disease=1) = 1.87 \*1.03(Age) \*0.99(Cholestrol) \*1.36(Oldpeak) \*0.25(Sex\_F) \*0.14(ChestPainType\_ATA) \*0.18(ChestPainType\_NAP) \*0.26(ChestPainType\_TA) \*2.77(FastingBS\_1) \*2.64(ExerciseAngina\_Y) \*0.08(ST\_Slope\_Up).

**Classification Tree:**

Similarly, we ran a classification tree analysis and after having analyzed the decile chart, we chose a **new cutoff value of 0.222**, following which we re-ran the analysis. These are our results:

The results **without setting a limit to the number of records** in the terminal:

|  |  |
| --- | --- |
| Full Tree | Best Tree |
|  |  |

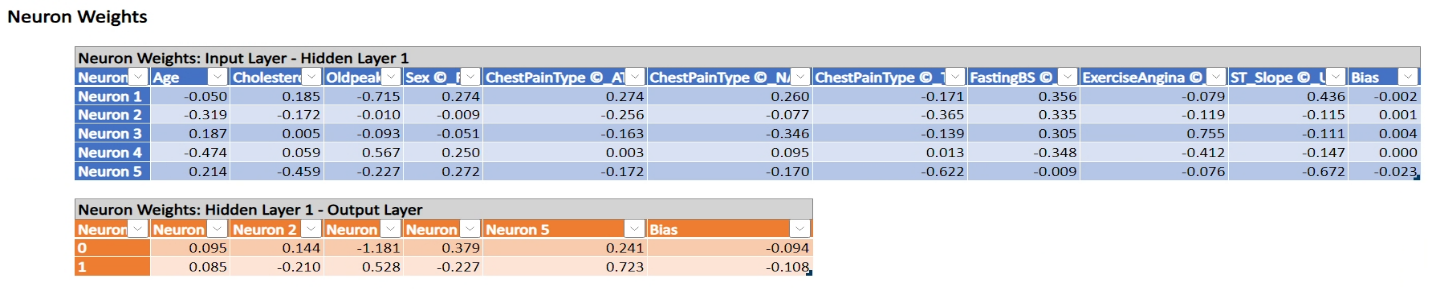
The results after **limiting the number of records in the terminal to “46” (default random value)**:

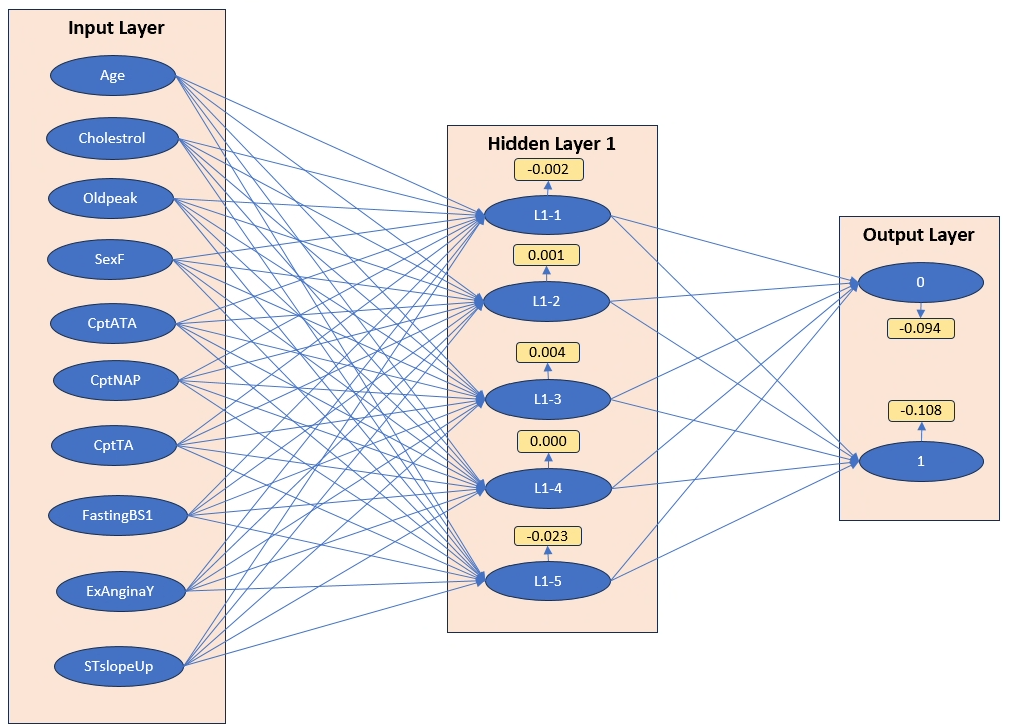
|  |  |
| --- | --- |
| Full Tree | Best Tree |
|  |  |

**Neural Network:**

We first performed an automatic neural network analysis using significant variables. Based on the output, we chose to run network 6 (highest sensitivity), network 2 (highest precision) and network 5 (optimum sensitivity and precision). All the 3 networks had a good ROC curve and AUC values (0.89, 0.90 and 0.85 respectively). So, we decided to go with network 5 since it had a good balance between sensitivity and precision (higher sensitivity though). For further steps, we analyzed the decile chart for network 5 and determined a **new cutoff of “0.571”** to run the network again.

These were the resultant outputs:

Data Analysis (**first model- Logistic Regression**)

* + **What are the business questions? (in case you changed your questions after the first two milestones)**

When attempting to see what influences CVDs, can we model the individual’s health indicators to analyze what combination of parameters makes an individual more likely to have a CVD?

What is the most influential factor of the listed health indicators that has the strongest correlation in determining whether an individual has a CVD?

What factors have the least impact or influence on determining whether an individual has a CVD?

* + **The model you use? Why?**

The first model we used was logistic regression as the model predicts the probability that a given input belongs to certain independent variables or how they may be influenced by them. In this case, we can use the model to look at given health indicators to analyze their effect on the likelihood of having a CVD, whether they are impactful or not.

* + **Which variables are used? Why?**

The variables used are Age, Cholesterol, Oldpeak, Sex ©\_F, ChestPainType ©\_ATA, ChestPainType ©\_NAP, ChestPainType ©\_TA, FastingBS ©\_1, ExerciseAngina ©\_Y, ST\_Slope ©\_Up. These variables were selected through feature selection step with the best subset, leading use with the variables being the significant.

* + **Any variable selection techniques used?**

The variable selection technique used was a feature selection step with the best subset.

* + **Report the model output, the equation (if applied), the explanation of coefficients (if applied)**

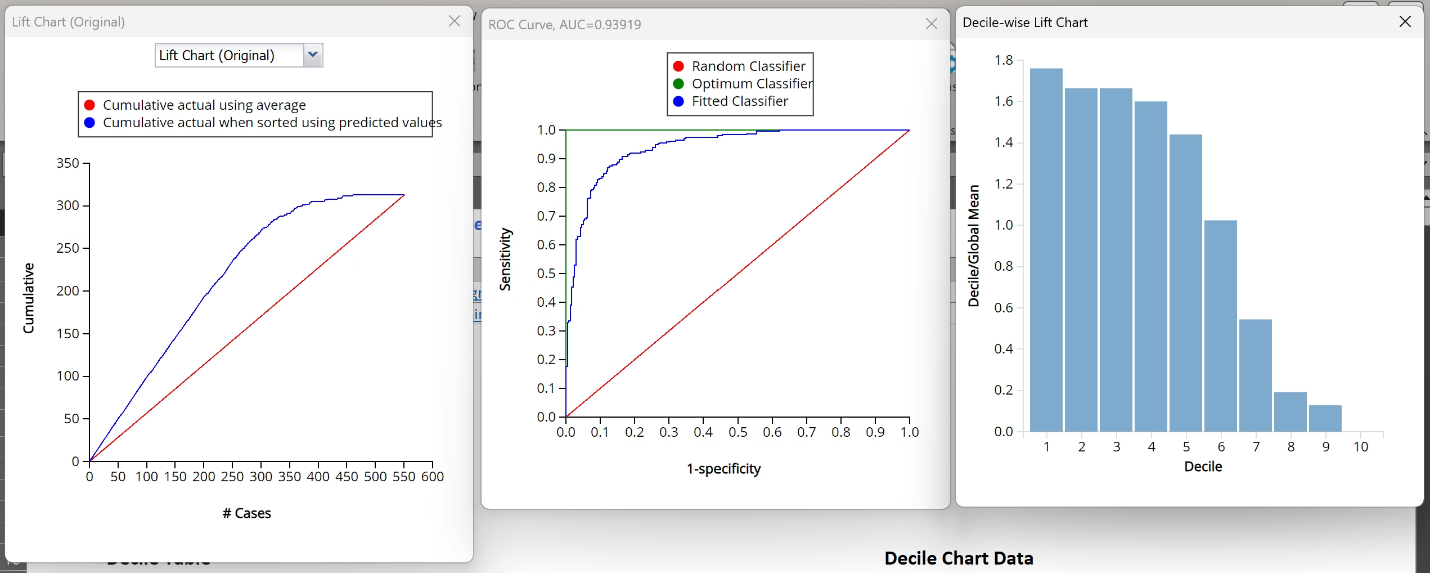
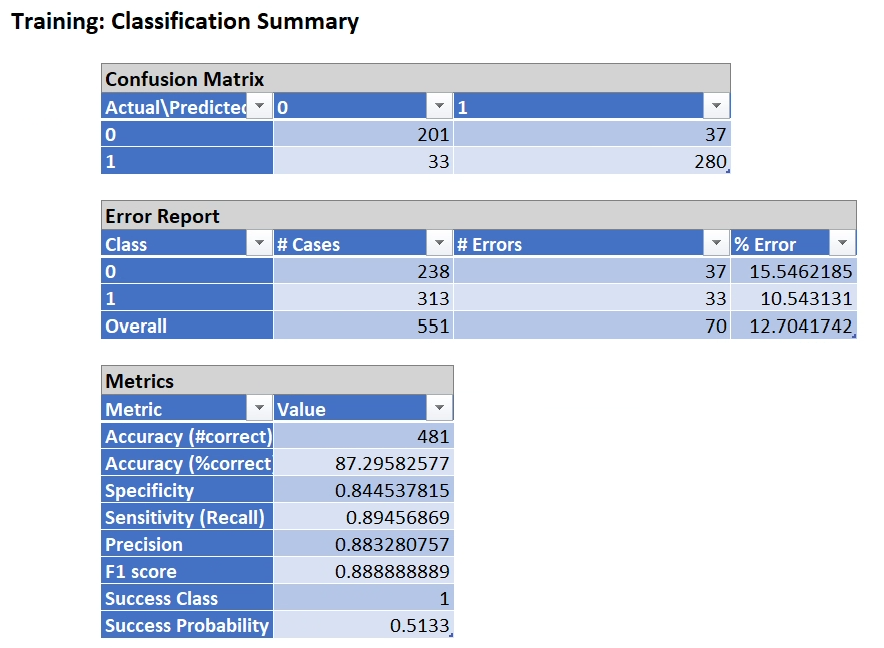
The model’s equations are

logit (Heart\_Disease=1) = 0.63 +0.033(Age) -0.003(Cholestrol) +0.31(Oldpeak) -1.39(Sex\_F) -1.91(ChestPainType\_ATA) -1.72(ChestPainType\_NAP) -1.33(ChestPainType\_TA) +1.02(FastingBS\_1) +0.97(ExerciseAngina\_Y) -2.5(ST\_Slope\_Up)

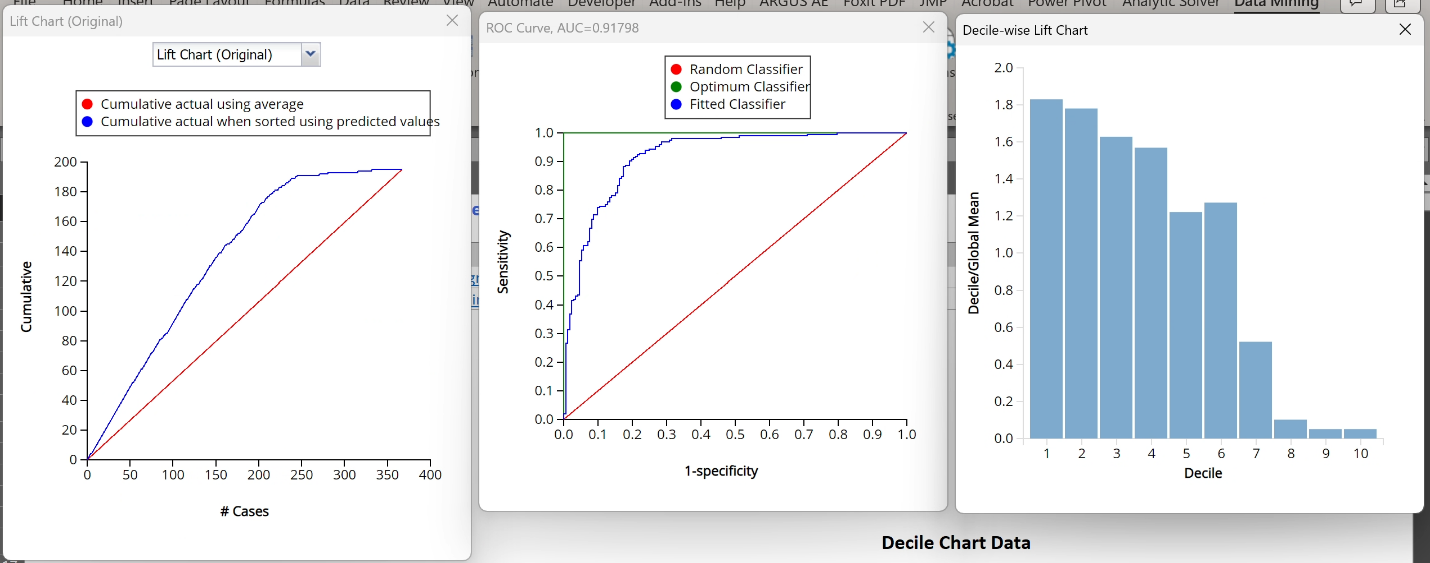
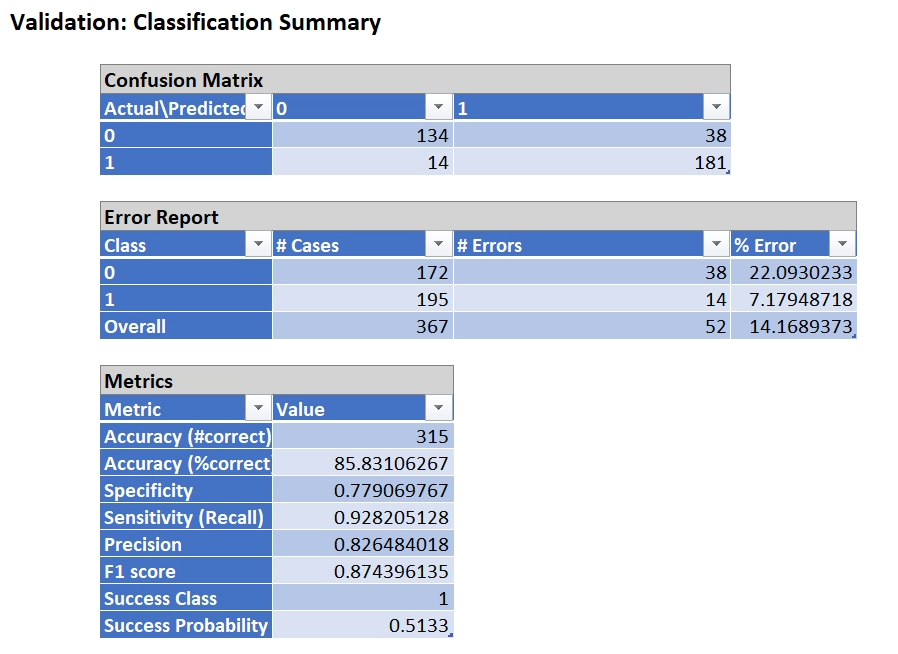
Odds (Heart\_Disease=1) = 1.87 \*1.03(Age) \*0.99(Cholestrol) \*1.36(Oldpeak) \*0.25(Sex\_F) \*0.14(ChestPainType\_ATA) \*0.18(ChestPainType\_NAP) \*0.26(ChestPainType\_TA) \*2.77(FastingBS\_1) \*2.64(ExerciseAngina\_Y) \*0.08(ST\_Slope\_Up).

* + **Report Training and validation (and test) data summary report and lift charts. How’s the model work?**

**FOR TRAINING**



**FOR VALIDATION**



The validation data had an error rate of 0 of 22.09% and 1 of 7.17%, for a total error rate of 14.16%. Ther accuracy is 85.83%, specificity of .779 and recall of .08264 and an F1 score of .874 with a success probability being .5133.

The validation data lift chart has a steady lift and indicates that the models predicitions are consistently better than random health indicators or guesses. It demonstrates the model is accurate and able to provide reliable information and insights.

The training data had a 0-error rate of 15.54% and a 1 error rate of 10.54% for a total of 12.7%. 481 records were accurate for 87.295%. .84453 is the specifity and the recall is .894. The precision is .8832 and the F1 score is .888.

The training data lift chart also indicates that the model’s predicitions are correct and going to be more accurate than random guesses. It shows the model is accurate and you can extract insights from it.

* + **If answering a classification question, based on your results, how do you choose the cutoff value?**

To choose a new cutoff value based on the results and analyzing the initial decile chart you choose the cut of value (in this case .5133) by looking at how closely the predicted values match the deciles.

* + **How do the results answer your questions?**

With these answers, we can model the individual’s health indicators to analyze what combination of parameters makes an individual more likely to have a CVD. Also, we can determine the most influential factor of the listed health indicators that has the strongest correlation in determining whether an individual has a CVD. The most overall influential factor is ST\_slope\_up with –2.5 and the most positive factor is having a fasting blood sugar of 1>120mg/dl. The least influential health indicator is Cholesterol with it being equal to .003.

* + **Provide a hypothetical example of one new record and make prediction/classification.**

Characteristics include:

* Age: 50
* Cholesterol 220 mg/dL
* Oldpeak: 2.5
* Sex: Male
* Chest pain Type: ATA = 1
* Fasting Blood Sugar: Normal
* Exercised-induced Angina: = Yes/1
* ST Segment Slope: Upward

logit (Heart\_Disease = 1) = 0.63 + 0.033(Age) - 0.003(Cholesterol) + 0.31(Oldpeak) - 1.39(Sex\_F) - 1.91(ChestPainType\_ATA) - 1.72(ChestPainType\_NAP) - 1.33(ChestPainType\_TA) + 1.02(FastingBS\_1) + 0.97(ExerciseAngina\_Y) - 2.5(ST\_Slope\_Up)

Probability (Heart\_Disease =1) .614

* + **The conclusions that you have drawn/suggestions/conclusions to the firm (e.g., the actions that one can take in response to the findings). Please write a full paragraph of conclusions and assume your audience does not have any statistical background.**

Based on the model we used, this one being logistic regression, there are identifiable health indicators that can elevate risk for heart disease. These include age, execrise amounts, the level of your blood sugar while fasting, and exercised-induced angina or chest pain, as well as various other types of chest pain. Older individuals with an upward sloping ST segment or exercise amount are at increased risk. Also, those who experience exercise-based chest pain also show higher risks of heart disease. This shows that modifications to your lifestyle such as regular monitoring of blood sugar levels, and timely medical assessments assessing these health indicators can have positive effects on mitigating heart disease overall. A good way to counteract and use some of the findings from this is to increase awareness among high-risk groups to be monitored. Also, for all populations this can be useful as individuals who are aware of this can make life modifications to reduce the risk of heart disease.

Data Analysis (**Second model-Classification Trees**)

* + **What are the business questions? (in case you changed your questions after the first two milestones)**

When attempting to see what influences CVDs, can we model the individual’s health indicators to analyze what combination of parameters makes an individual more likely to have a CVD?

What is the most influential factor of the listed health indicators that has the strongest correlation in determining whether an individual has a CVD?

What factors have the least impact or influence on determining whether an individual has a CVD?

* + **The model you use? Why?**

The second model that we use is the **classification tree**. We use this to generate rules to classify the patient based on the health indicator inputs. One of the other reasons is that it is easy to understand and requires minimal calculation.

* + **Which variables are used? Why?**

The variables used are Age, Cholesterol, Oldpeak, Sex ©\_F, ChestPainType ©\_ATA, ChestPainType ©\_NAP, ChestPainType ©\_TA, FastingBS ©\_1, ExerciseAngina ©\_Y, ST\_Slope ©\_Up.

These variables are chosen based on the stepwise feature selection in case of linear regression to only include inputs that are significant and to minimize the risk of overfitting.

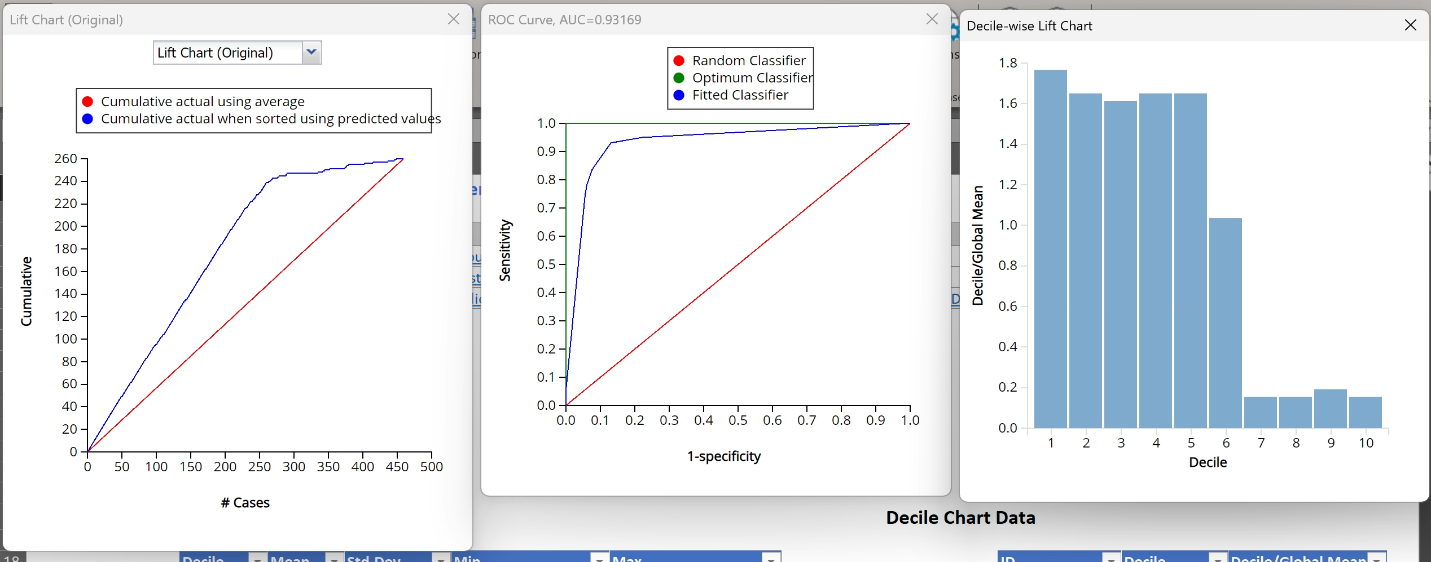
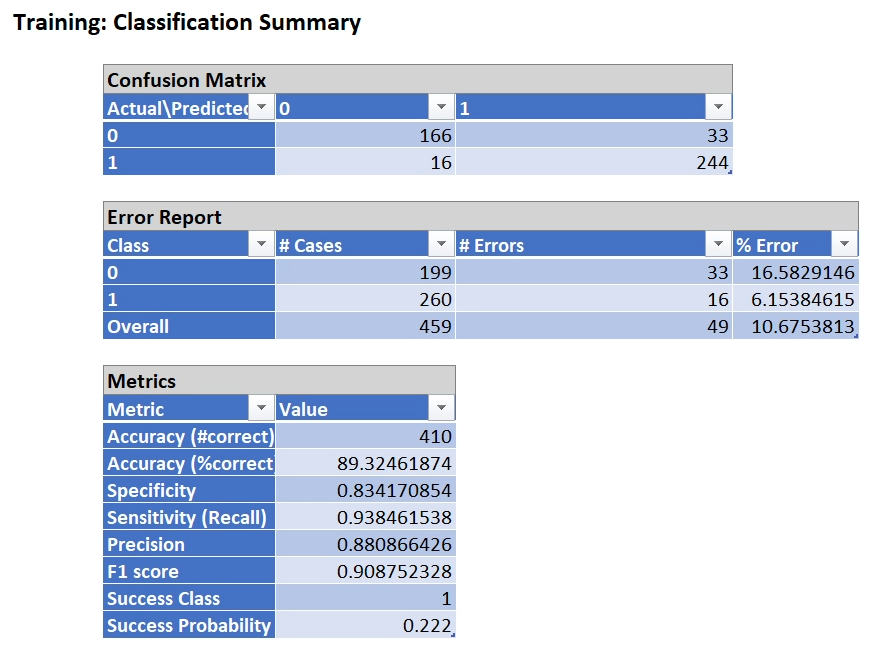
* + **Any variable selection techniques used?**

The variable selection technique used was a Stepwise feature selection with the best subset on the linear regression model.

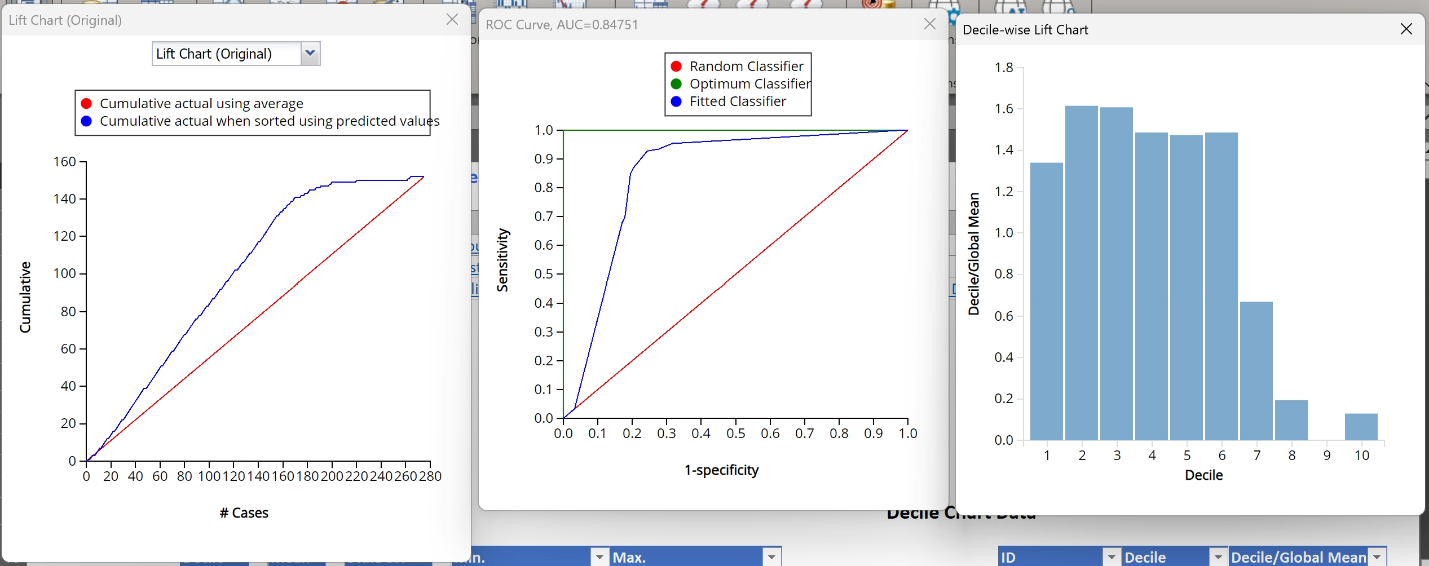
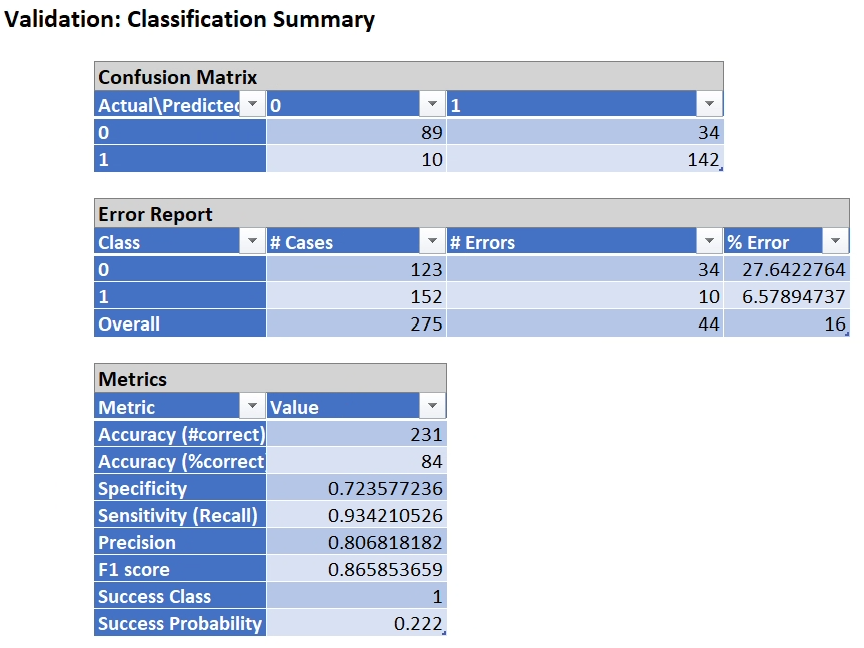
* + **Report the model output, the equation (if applied), the explanation of coefficients (if applied)**

The model’s rules for the Best pruned tree are:

1. IF **ST\_Slope ©\_Up > 0.08**AND **Choesterol < -1.25** Then **Heart Disease = 1**
2. IF **ST\_Slope ©\_Up > 0.08** AND **Choesterol > -1.25** AND **ExerciseAngina ©\_Y < 0.2** Then **Heart Disease = 0**
3. IF **ST\_Slope ©\_Up > 0.08** AND **Choesterol > -1.25** AND **ExerciseAngina ©\_Y > 0.2** Then **Heart Disease = 1**
4. IF **ST\_Slope ©\_Up < 0.08** AND **Sex ©\_F = 0** AND **Age > -0.91** Then **Heart Disease = 1**
5. IF **ST\_Slope ©\_Up < 0.08** AND **Sex ©\_F = 0** AND **Age < -0.91** AND **ChestPainType ©\_TA < 0.63** Then **Heart Disease = 1**
6. IF **ST\_Slope ©\_Up < 0.08** AND **Sex ©\_F = 0** AND **Age < -0.91** AND **ChestPainType ©\_TA > 0.63** Then **Heart Disease = 0**
7. IF **ST\_Slope ©\_Up < 0.08** AND **Sex ©\_F = 1** AND **Age < 0.5** AND **Cholestrol < 0.74** Then **Heart Disease = 0**
8. IF **ST\_Slope ©\_Up < 0.08** AND **Sex ©\_F = 1** AND **Age < 0.5** AND **Cholestrol > 0.74** Then **Heart Disease = 1**
9. IF **ST\_Slope ©\_Up < 0.08** AND **Sex ©\_F = 1** AND **Age > 0.5** Then **Heart Disease = 1**
   * **Report Training and validation (and test) data summary report and lift charts. How’s the model work?**

**FOR TRAINING**

**FOR VALIDATION**



The % error for class 0 is 27.64% in the validation set and 6.57% in the case of class 1. The overall % error is 16% with a recall of 0.934 and a precision of 0.806.

The validation lift chart has an ROC value of 0.847 which suggests that the model works well with the data and can predict the results accurately. The high recall is a good sign as it implies the model classifying people with heart disease correctly.

The % error for class 0 is 16.58% in the training set and 6.15% in the case of class 1. The overall % error is 10.67% with a recall of 0.938 and a precision of 0.88.

The training lift chart has an ROC value of 0.931 which suggests that the model works well with the data and can predict the results accurately.

The % error for class 0 is 36.36% in the training set and 12.5% in the case of class 1. The overall % error is 23.91% with a recall of 0.875 and a precision of 0.724.

The test lift chart has an AUC value of 0.769 which shows that the model is moderately accurate in the test case. This basically means when we take two people – one with and one without heart disease, there is a 76.9% chance that the person with disease has a higher score than the one without disease.

* + **If answering a classification question, based on your results, how do you choose the cutoff value?**

From the decile chart for the validation set, we see for 60% of the records in the validation data the decile mean is greater than 1 so we choose the new cut off value based on these results (0.222 in this case).

* + **How do the results answer your questions?**

From the rules generated, we can classify people with heart disease and people without heart disease based on the given input parameters.

* + **Provide a hypothetical example of one new record and make prediction/classification.**
* Age: 50
* Cholesterol 220 mg/dL
* Oldpeak: 2.5
* Sex: Male
* Chest pain Type: ATA = 1
* Fasting Blood Sugar: Normal
* Exercised-induced Angina: = Yes/1
* ST Segment Slope: Upward

From the given rules, the classification is 1

* + **The conclusions that you have drawn/suggestions/conclusions to the firm (e.g., the actions that one can take in response to the findings). Please write a full paragraph of conclusions and assume your audience does not have any statistical background.**

We have generated classification tree models which provided us with a set of rules which can be used to classify the inputs. The primary role of these rules is to make it easier for the readers to use as they don’t require any extensive calculations to generate an output. We also found out various parameters to determine the working capability of the model like recall, precision, accuracy, ROC curve, etc. From our findings, we can see that one of the primary distinctions for classification is whether an individual has an upward sloping segment. There are also many subcategories that the classification depends on. A good way to counteract and use some of the findings from this is to increase awareness among high-risk groups to be monitored. Also, for all populations this can be useful as individuals who are aware of this can make life modifications to reduce the risk of heart disease.

Data Analysis (**Third model – Neural Network**)

* + **What are the business questions? (in case you changed your questions after the first two milestones)**

When attempting to see what influences CVDs, can we model the individual’s health indicators to analyze what combination of parameters makes an individual more likely to have a CVD?

What is the most influential factor of the listed health indicators that has the strongest correlation in determining whether an individual has a CVD?

What factors have the least impact or influence on determining whether an individual has a CVD?

* + **The model you use? Why?**

The third model we use is Neural Network. We use this to understand any complex relationship between the input parameters.

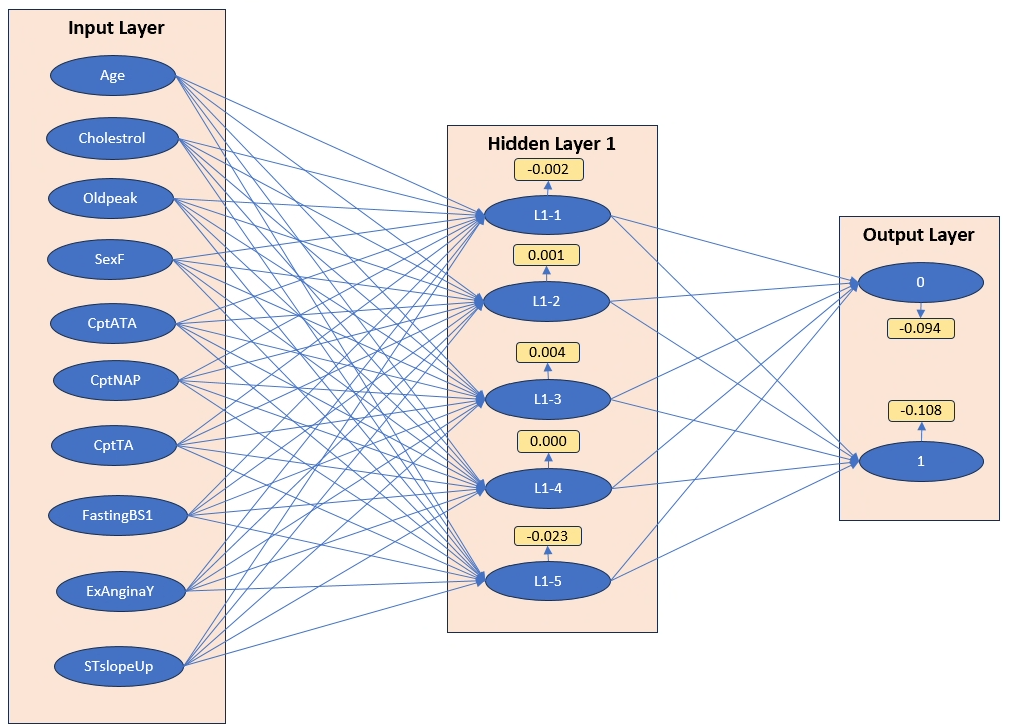
* + **Which variables are used? Why?**

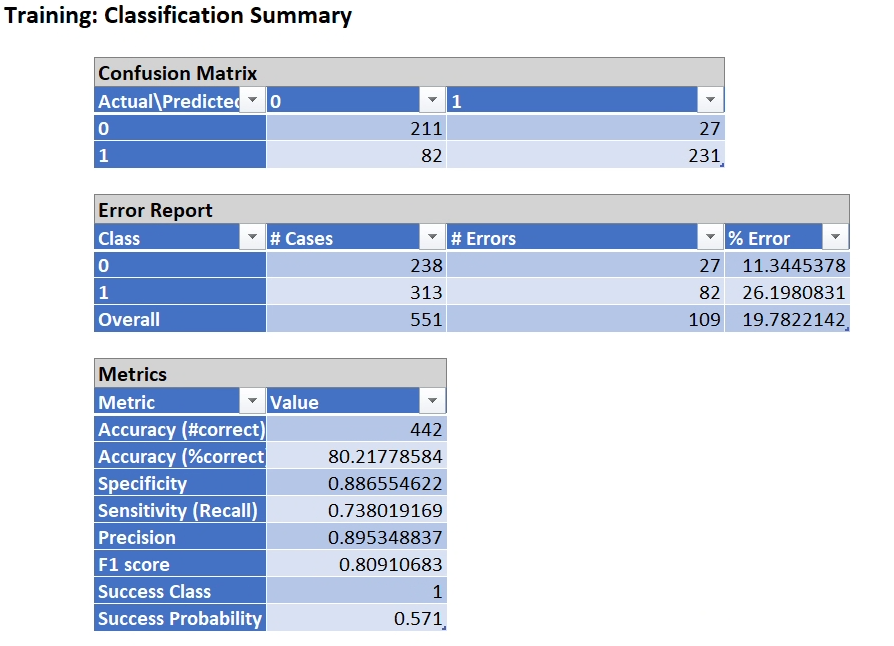
The variables used are Age, Cholesterol, Oldpeak, Sex ©\_F, ChestPainType ©\_ATA, ChestPainType ©\_NAP, ChestPainType ©\_TA, FastingBS ©\_1, ExerciseAngina ©\_Y, ST\_Slope ©\_Up.

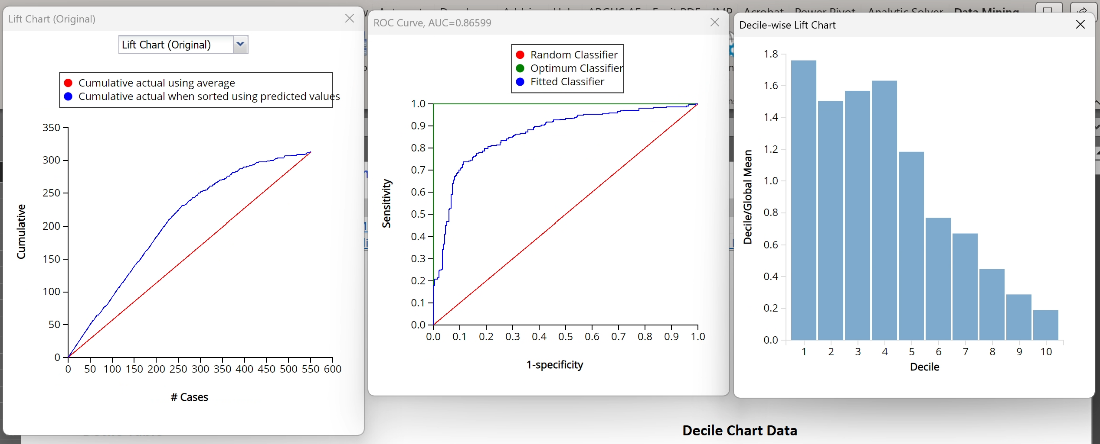
These variables are chosen based on the stepwise feature selection in case of linear regression to only include inputs that are significant and to minimize the risk of overfitting.

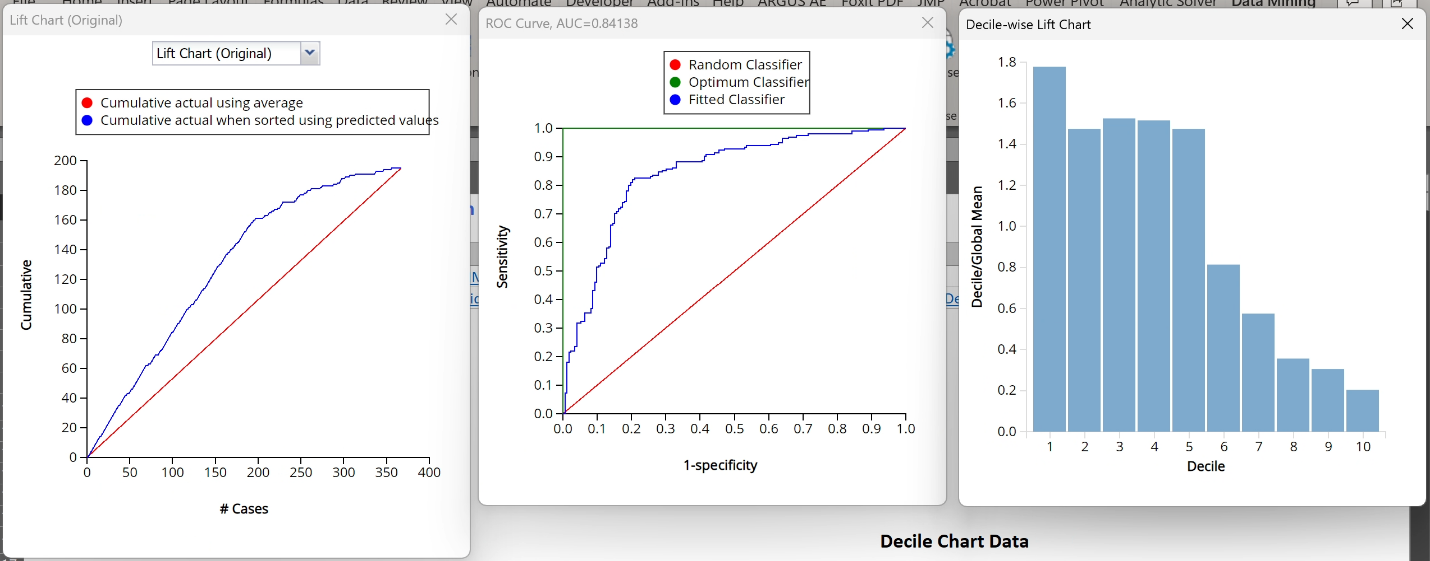
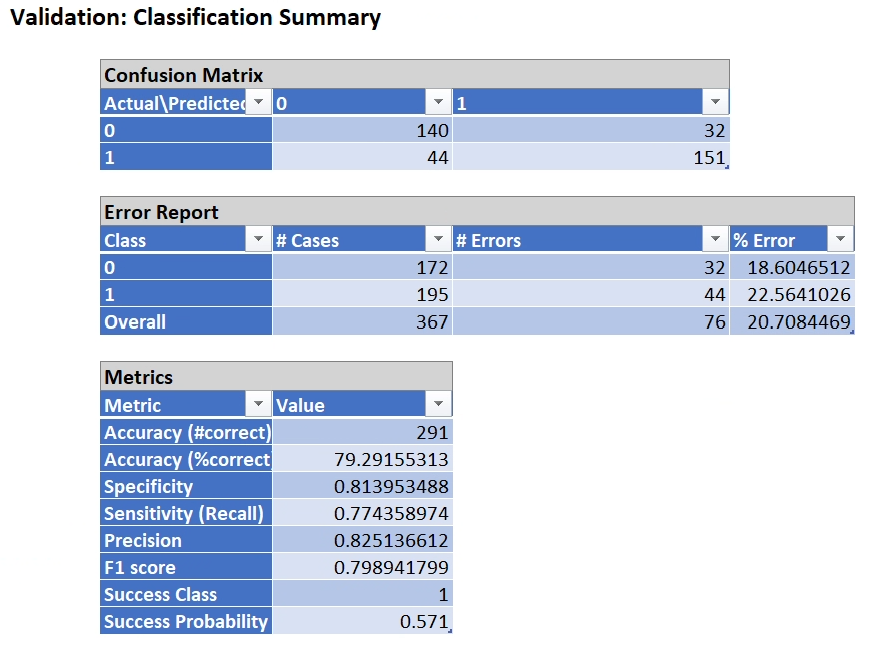
* + **Any variable selection techniques used?**

The variable selection technique used was a Stepwise feature selection with the best subset on the linear regression model.

* + **Report the model output, the equation (if applied), the explanation of coefficients (if applied)**
  + Report Training and validation (and test) data summary report and lift charts. How’s the model work?

**FOR TRAINING**



**FOR VALIDATION**

The % error for class 0 is 18.6% in the validation set and 22.56% in the case of class 1. The overall % error is 20.7% with a recall of 0.774 and a precision of 0.825.

The validation lift chart has an ROC value of 0.841 which suggests that the model works well with the data and can predict the results accurately. This basically means when we take two people – one with and one without heart disease, there is an 84.1% chance that the person with disease has a higher score than the one without disease.

The % error for class 0 is 11.34% in the training set and 26.19% in the case of class 1. The overall % error is 10.67% with a recall of 0.738 and a precision of 0.895.

The training lift chart has an ROC value of 0.865 which suggests that the model works well with the data and can predict the results accurately.

* + **If answering a classification question, based on your results, how do you choose the cutoff value?**

From the decile chart for the validation set, we see for 50% of the records in the validation data the decile mean is greater than 1 so we choose the new cut off value based on these results (0.571 in this case).

* + **How do the results answer your questions?**

From the equations produced by the model, we can use the input parameters to classify the patients. Also, we can change the parameters like the learning rate to govern the pace at which the model updates or learns the values of the parameters.

* + **Provide a hypothetical example of one new record and make prediction/classification.**
* Age: 50
* Cholesterol 220 mg/dL
* Oldpeak: 2.5
* Sex: Male
* Chest pain Type: ATA = 1
* Fasting Blood Sugar: Normal
* Exercised-induced Angina: = Yes/1
* ST Segment Slope: Upward

Output (L1-1) = 1

Output (L1-2) = 1

Output (L1-3) = 0.99

Output (L1-4) = 0.000036

Output (L1-5) = 1.35 x

Output (0) = 0.264

Output (1) = 0.571

From this we can see the classification is 1.

* + **The conclusions that you have drawn/suggestions/conclusions to the firm (e.g., the actions that one can take in response to the findings). Please write a full paragraph of conclusions and assume your audience doesn’t have any statistical background.**

Through our analysis using a neural network model, we've uncovered significant insights into the factors influencing cardiovascular diseases (CVDs). Key indicators like age, cholesterol levels, and symptoms such as chest pain and exercise-induced angina emerge as critical determinants of CVD occurrence. This analysis facilitates patient classification based on these factors, enabling targeted interventions and personalized healthcare approaches. To mitigate CVD risk, close monitoring and management of cholesterol levels, particularly in older individuals, are advised. Encouraging regular exercise and adopting healthy lifestyle habits also emerge as effective strategies for reducing CVD incidence. Overall, our findings emphasize proactive, tailored healthcare measures for optimal outcomes.