Heart Diseases Detection

Bhavya Setia, Riya Badoni and Sahibnoor Chahal

INDEX

Project Description

Introduction

Millions of people worldwide die from cardiovascular diseases every year. It is extremely important to identify at-risk patients as soon as possible so they can make life changes that will lower their risk of developing those cardiovascular diseases. For our project, we have chosen to analyze a dataset consisting of various measurements/information of a person and their lifestyle and whether a person is labelled as 'at-risk' for heart disease within the next 10 years in which a 1 represents 'yes' and 0 represents 'no'. The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The goal of our analysis is to identify the variables with the biggest impact on whether or not someone is classified as at-risk and with only a basic understanding of heart diseases, we hypothesize that the variables with the highest correlation with being labelled 'at-risk' will be a person's age, number of cigarettes smoked a day, their blood pressure, and glucose levels.

The Dataset

- **Numerical:** age, cigsPerDay, totChol, sysBP, diaBP, BMI, heartRate, glucose, PulsePressure
- **Categorical:** Male, Education, currentSmoker, BPMeds, prevalentStroke, prevalentHyp, diabetes, TenYearCHD

Variable Description

Independent Variables

- **male:** Binary variable indicating the gender of the individual, where '1' represents male and '0' represents female.
- **age:** The age of the individual in years, a continuous variable important for assessing risk as age increases.
- **education:** Categorical variable representing the highest level of education achieved by the individual.
- **currentSmoker:** Binary variable where '1' indicates the individual currently smokes cigarettes and '0' does not.
- **cigsPerDay:** The average number of cigarettes the individual smokes per day. Relevant for assessing lifestyle risks.
- **BPMeds:** Binary variable indicating whether the individual is on blood pressure medication ('1' for yes, '0' for no).
- **prevalentStroke:** Binary variable indicating whether the individual has had a stroke ('1' for yes, '0' for no).
- **prevalentHyp:** Binary variable indicating whether the individual has prevalent hypertension ('1' for yes, '0' for no).
- **diabetes:** Binary variable indicating whether the individual has diabetes ('1' for yes, '0' for no).

- **totChol:** Total cholesterol level, a continuous variable indicating the milligrams of cholesterol per deciliter of blood.
- **sysBP:** Systolic blood pressure, a continuous variable measuring the maximum arterial blood pressure during contraction of the left ventricle of the heart.
- **diaBP:** Diastolic blood pressure, a continuous variable measuring the arterial pressure between heartbeats.
- **BMI:** Body Mass Index, calculated as weight in kilograms divided by the square of height in meters.
- **heartRate:** The individual's heart rate in beats per minute, a continuous variable.
- **glucose:** The individual's glucose level, a continuous variable crucial for diagnosing diabetes and other metabolic disorders.

Dependent Variable

TenYearCHD: Binary variable indicating whether the individual is predicted to develop coronary heart disease (CHD) within the next ten years ('1' for yes, '0' for no). This variable is critical for identifying individuals at higher risk of major cardiovascular events, allowing for early intervention and targeted treatment plans.

Target Audience

The target audience for our diabetes prediction model includes several key groups:

1. **Healthcare Providers**: Doctors, nurses, and other healthcare professionals who can use the model's predictions to identify at-risk patients early and intervene accordingly.

2. **Public Health Officials**: Government and public health officials who can utilize the information to develop targeted interventions and policies aimed at reducing the diabetes burden in the community.

3. **Health Insurance Companies**: Insurers can use the model to adjust premiums and coverage plans based on the predicted risk and to promote health programs that could mitigate this risk among their clients.

4. **Patients and General Public**: Individuals can use the predictions to understand their personal risk of diabetes and take proactive steps towards lifestyle modifications.

5. **Research and Academic Community**: Researchers and students interested in chronic disease management and prevention strategies can use the findings to support further studies and to develop more advanced predictive models.

Business Questions

1. To what extent does the presence of diabetes increase the likelihood of developing CHD over the ten-year period covered by the study?
2. Is there a relationship between early life factors (such as childhood obesity or family history of heart disease) and the likelihood of being labeled 'at-risk' for CHD in adulthood?
3. Is there a correlation between socioeconomic status (such as education level, income, or occupation) and the risk of developing CHD?
4. What predictive factors can be identified as significant predictors of diabetes risk?
5. Can a model be developed to accurately classify individuals at high risk before the onset of the disease?
6. How can the model's predictions be used to inform public health strategies and interventions aimed at preventing diabetes or managing its impact more effectively?

Data Preprocessing

- **Data Collection and Initial Assessment:** The dataset originally comprised approximately 4,200 records of individuals, including a range of variables related to health status, lifestyle choices, and demographic information. The initial step involved a comprehensive assessment of data quality, identifying missing values, potential errors, and assessing the overall usability of the dataset for predictive modeling.
- **Handling Missing Data:** Our preprocessing phase addressed missing data, which is critical for maintaining the integrity of our predictive analysis. For most variables, rows with missing data were removed to prevent the introduction of bias or inaccurate imputation. However, for the cigsPerDay variable, where missing values were prevalent among current smokers, we employed an imputation technique using the average number of cigarettes smoked per day. This approach was chosen because it maintained the consistency and reliability of our dataset, considering the behavior patterns of the population segment.
- **Data Transformation and Feature Engineering:** In dealing with skewed data distributions and potential multicollinearity:
- We transformed cigsPerDay, Heart Rate, and Pulse Pressure using logarithmic transformations to normalize their distributions and reduce the influence of extreme values.
- To streamline the model and avoid multicollinearity, we combined systolic and diastolic blood pressure into a single feature called Pulse Pressure by subtracting diastolic from systolic values. This transformation also helped in focusing on more relevant predictors for cardiovascular health, which is often compromised in diabetic patients.
- **Statistical Summary and Outliers:** Detailed statistical summaries were computed for all variables, including means, medians, standard deviations, and the identification of outliers. Outliers were assessed through visual analysis using box

plots. This step was crucial to understand the data's central tendencies and variability, which informed further data cleaning and preparation efforts.

Identification of Potential Outliers:

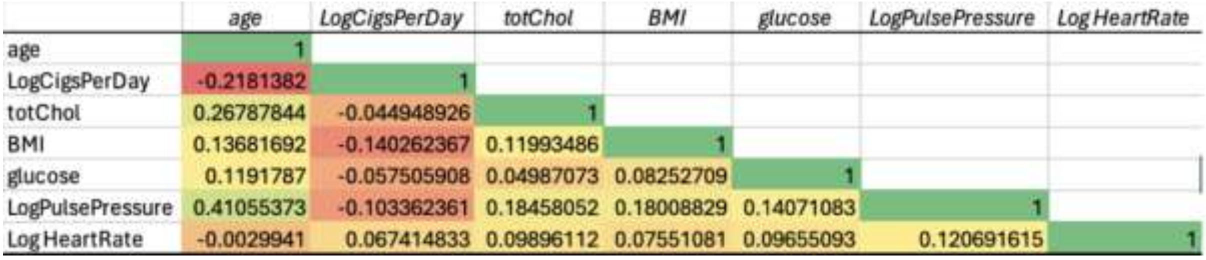
Variable	*Mean*	*Median*
Age	49.58	49.00
Education	1.98	2.00
CigsPerDay	9.00	0.00
BPMeds	0.03	0.00
PrevalentStroke	0.006	0.00
PrevalentHyp	0.31	0.00
Diabetes	0.03	0.00
TotChol	236.72	234.00
SysBP	132.35	128.00
DiaBP	82.89	82.00
BMI	25.80	25.40
HeartRate	75.88	75.00
Glucose	81.97	78.00

Table 1 Statistics on Numerical Variables

- **CigsPerDay:** The large difference between the mean and the median suggests a right-skewed distribution. A median of 0 but a mean of 9 suggests that while many individuals do not smoke at all, a few heavy smokers are skewing the average upwards.
- **BPMeds:** Given that the mean (0.03) is significantly different from the median (0.00), it indicates that a small proportion of the population is on blood pressure medication, which could be considered outliers.
- **PrevalentStroke, PrevalentHyp, and Diabetes:** These binary variables show low mean values close to zero, suggesting that most individuals do not have these conditions. The few positive instances can be seen as outliers.
- **Glucose:** A higher mean compared to the median could suggest a presence of higher glucose levels in some individuals, indicative of outliers or skewed data towards higher glucose levels.

Correlation Analysis and Visualizations:

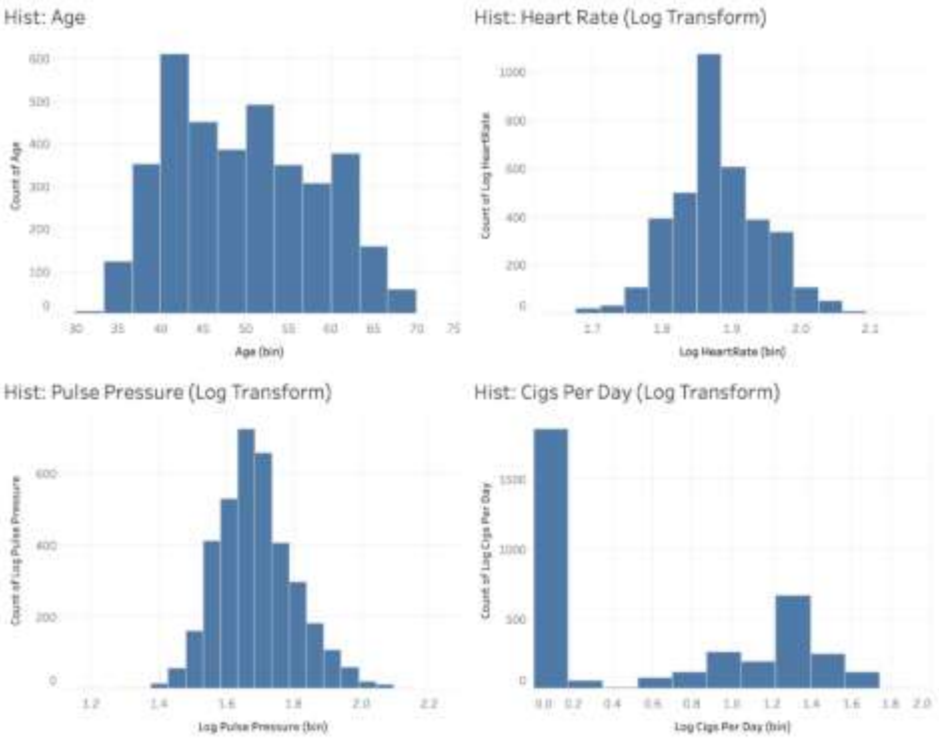
We conducted a correlation analysis to explore the relationships between variables. A correlation matrix was generated to visually inspect potential multicollinearity and direct the feature selection process effectively.



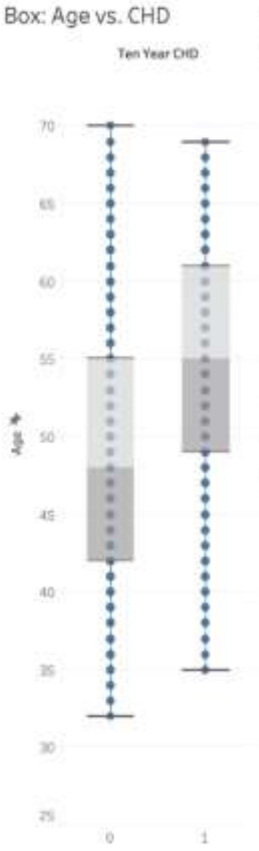
	age	*LogCigsPerDay*	*totChol*	*BMI*	*glucose*	*LogPulsePressure*	*Log HeartRate*
age	1						
LogCigsPerDay	-0.2181382	1					
totChol	0.26787844	-0.044948926	1				
BMI	0.13681692	-0.140262367	0.11993486	1			
glucose	0.1191787	-0.057505908	0.04987073	0.08252709	1		
LogPulsePressure	0.41055373	-0.103362361	0.18458052	0.18008829	0.14071083	1	
Log HeartRate	-0.0029941	0.067414833	0.09896112	0.07551081	0.09655093	0.120691615	1

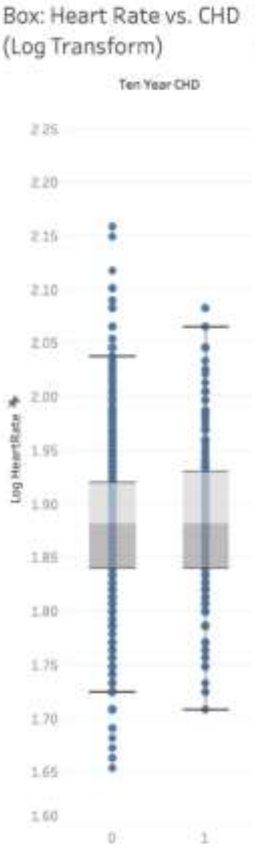
- **Age and LogPulsePressure:** This strong positive correlation between age and logarithmic transformation of pulse pressure indicates that as individuals age, their pulse pressure tends to increase. This is significant as it reflects arterial stiffness and an increased cardiovascular risk associated with aging.
- **LogPulsePressure and BMI:** This positive correlation suggests that higher BMI is associated with higher pulse pressure, which could indicate higher cardiovascular stress or potential for hypertension in individuals with higher BMI.
- The **negative correlations** such as between age and LogCigsPerDay, and BMI and totChol, while interesting, may be influenced by other factors not directly visible from the correlation alone. These might require deeper investigation into population demographics or health behavior changes with age.
- The **low to moderate correlations** involving variables like heart rate and cholesterol levels do not strongly suggest direct relationships but instead hint at more complex interactions that might involve multiple physiological factors or confounding variables.
- Age being strongly correlated with pulse pressure is particularly notable for its implications in cardiovascular health risk assessments.
- The relationships of BMI with both glucose and pulse pressure underline the importance of weight management in overall health strategies, especially concerning metabolic and cardiovascular health.

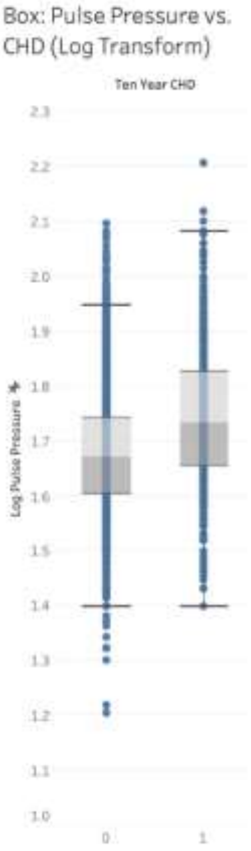
Additionally, histograms and box plots were created to visualize the distribution of and relationships between the dependent and independent variables, providing deeper insights into the data structure and informing subsequent modeling decisions.

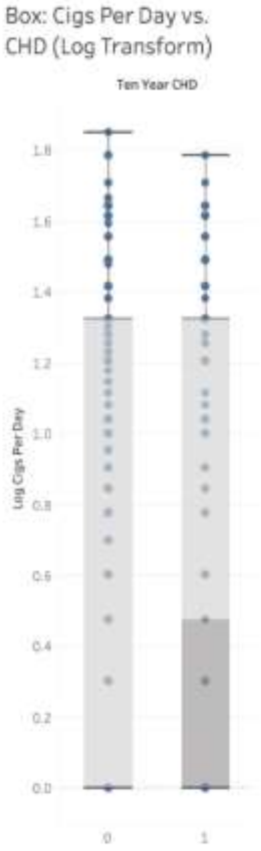


- **Age:** The histogram for age shows a relatively uniform distribution with a slight right skew, suggesting that the dataset includes a broad range of ages, but with a modest increase in the population in their late 40s to early 60s. This is typical for health-related datasets where middle-aged individuals are frequently targeted due to their higher risk for chronic conditions.
- **Pulse Pressure (Log Transform):** The distribution of log-transformed pulse pressure appears roughly normally distributed with a peak around 1.6 to 1.8. This transformation likely helps to normalize the data, reducing skewness and making it more suitable for analysis in linear models. The data concentration around the median suggests that extreme values of pulse pressure are less common, which could indicate fewer individuals with extreme hypertension or hypotension conditions.
- **Heart Rate (Log Transform):** Heart rate, when log-transformed, shows a bell-shaped distribution which indicates a normal distribution. This is ideal for many statistical analyses that assume normality. The central peak suggests most individuals have a moderate heart rate, with fewer individuals at the low and high extremes. This can be indicative of a generally healthy population in terms of cardiac function, with outliers possibly representing individuals with specific health conditions.
- **Cigs Per Day (Log Transform):** The histogram for log-transformed cigarettes per day shows a highly skewed distribution with a large spike at zero. This indicates that a significant portion of the dataset includes non-smokers. The other smaller peaks suggest groups of moderate and heavy smokers. The use of a log transform helps manage the wide range of values and highlights the presence of heavy smokers, even though they are a minority in the dataset.


Box: Age vs. CHD


Box: Heart Rate vs. CHD (Log Transform)


Box: Pulse Pressure vs. CHD (Log Transform)


Box: Cigs Per Day vs. CHD (Log Transform)

The boxplots provided compare various health metrics against the binary outcome of the Ten Year Coronary Heart Disease (CHD) risk (0 = no CHD, 1 = CHD). Here are insights based on each plot:

- **Age vs. Ten Year CHD:** Individuals with CHD (1) generally appear to be older than those without CHD (0). The median age for the CHD group is noticeably higher, and the age range is also wider. Age is a significant risk factor for CHD, as older individuals show a higher propensity for CHD, likely due to cumulative health risks and longer exposure to potential cardiovascular stressors.
- **Heart Rate (Log Transform) vs. Ten Year CHD:** The heart rates for both groups (CHD and non-CHD) are similar, though there is a slight elevation in the median for the CHD group. The distribution is tight for both groups, indicating limited variability after log transformation. While there is a minor difference in heart rates between the two groups, heart rate alone might not be a strong predictor of CHD within this population. However, the slight elevation in the CHD group could suggest that even small increases in heart rate might be clinically relevant.
- **Pulse Pressure (Log Transform) vs. Ten Year CHD:** Both groups have similar distributions, with the CHD group showing a slightly higher median pulse pressure. The CHD group also has more variability and a slightly higher range of values. Higher pulse pressure might correlate with increased CHD risk, suggesting that arterial stiffness or hypertension could be more pronounced in individuals at risk for or with CHD. The wider range in the CHD group supports this as a factor worth considering in risk assessments.
- **Cigs Per Day (Log Transform) vs. Ten Year CHD:** The distribution of cigarettes per day is skewed towards lower values for both groups, but especially so for the non-CHD group. The CHD group shows slightly more spread, indicating higher variability and slightly more individuals with higher cigarette consumption. Smoking, even at lower levels, is more prevalent among those with CHD, underscoring the established relationship between tobacco use and heart disease. The spread in the CHD group suggests a stronger link between higher smoking rates and increased CHD risk.

  These expanded sections provide a more detailed overview of the project’s scope and the comprehensive data preprocessing efforts undertaken to ensure the development of a robust predictive model for diabetes risk.

Model choice and Rationale

The Models

Logistic Regression

Features in the final tuned model: male, age, BPMeds, prevelantHyp, treated_glucose, LogPulsePressure, LogCiggs

Cutoff Value: 0.138

Classification Tree

Features in the fully grown tree: male, age, currentSmoker, ciggsPerDay, BPMeds, , prevelantStroke, prevelantHyp, diabetes, Treated_totalChol, sysBP, diaBP, Treates_BMI, heartRate, treated_glucose, Treated_education1, Treated_education2, Treated_education3

Cutoff Value: 0.079

Minimum number of leaves: 10

Maximum Levels: 7

Neural Network

Hidden Layers: 2

- **Hidden Layer 1:** 12 neurons
- **Hidden Layer 2:** 9 neurons

Learning Rate: 0.1

Cost Function: Cross Entropy

Maximum number of epochs: 30

Model used: Logistic Regression

Logistic regression is well-suited for binary classification problems like predicting the presence or absence of heart disease. It is easy to implement, interpret, and provides probabilities for predictions, which is useful for assessing risk.

Variables used and Selection Techniques

Variables: age, cigsPerDay, totChol, sysBP, diaBP, BMI, heartRate, glucose, PulsePressure, Male, Education, currentSmoker, BPMeds, prevalentStroke, prevalentHyp, diabetes, TenYearCHD

Why these variables: These variables were chosen based on their relevance to cardiovascular health, supported by existing medical literature. They cover a broad range of factors, including demographics, lifestyle, and medical history, which are known to influence heart disease risk.

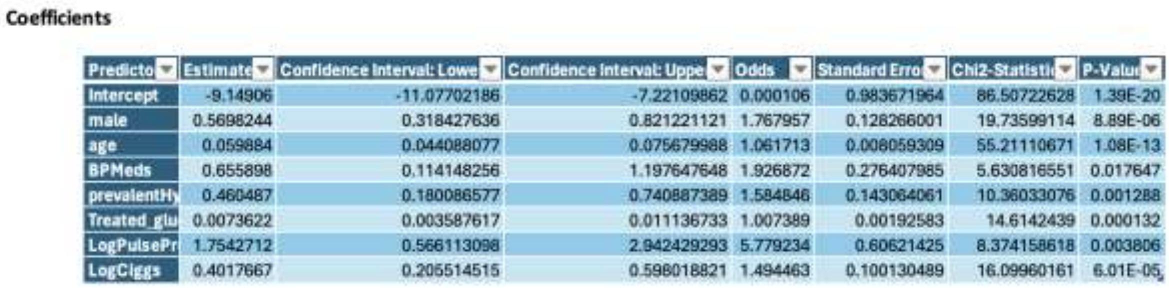
Variable Selection Techniques:

- **Correlation Analysis:** Used to identify and remove highly correlated variables to prevent multicollinearity.
- **Log Transformation:** Applied to skewed variables (e.g., cigsPerDay, Heart Rate, Pulse Pressure) to reduce the impact of outliers.

- **Pulse Pressure Calculation:** Used instead of separate systolic and diastolic blood pressures to simplify the model and reduce multicollinearity.

Model Output

The coefficients



Coefficients

Predicto	Estimate	Confidence Interval: Lowe	Confidence Interval: Uppe	Odds	Standard Erro	Chi2-Statisti	P-Valu
Intercept	-9.14906	-11.07702186	-7.22109862	0.000106	0.983671964	86.50722628	1.39E-20
male	0.5698244	0.318427636	0.821221121	1.767957	0.128266001	19.73599114	8.89E-06
age	0.059884	0.044088077	0.075679988	1.061713	0.008059309	55.21110671	1.08E-13
BPMeds	0.655898	0.114148256	1.197647648	1.926872	0.276407985	5.630816551	0.017647
prevalentHy	0.460487	0.180086577	0.740887389	1.584846	0.143064061	10.36033076	0.001288
Treated_glu	0.0073622	0.003587617	0.011136733	1.007389	0.00192583	14.6142439	0.000132
LogPulsePr	1.7542712	0.566113098	2.942429293	5.779234	0.60621425	8.374158618	0.003806
LogCiggs	0.4017667	0.205514515	0.598018821	1.494463	0.100130489	16.09960161	6.01E-05

Dependent variable TenYearCHD = 1 means the patient is at risk for coronary heart disease within the next 10 years.

- The odds of a male being at risk is 1.767957 times that of females, holding other variables constant.
- With each additional year of age, the odds of being at risk increases by 6.1713% holding other variables constant.
- For those on BPMeds, the odds of being at risk are 1.926872 times higher compared to those not on BPMeds, holding other variables constant.
- For individuals with prevalent hypertension, the odds of being t risk are 1.584846 times higher compared to those without prevalent hypertension, holding other variables constant.
- For each unit increase in the treated_glucose, the odds of being at risk slightly increase by 0.7389% (which is 1.007389 times), holding other variables constant.
- A one-unit increase in the LogPulsePr (log-transformed pulse pressure) is associated with the odds of being at risk being 5.779234 times higher, holding other variables constant.
- For each unit increase in the LogCiggs (log-transformed number of cigarettes), the odds of being at risk are 1.494463 times higher, holding other variables constant.

The Equations

- **Logit(TenYearCHD=1)** = -9.15 + 0.57*male + 0.06*age + 0.65*BPMeds + 0.46*prevalenthypertension + 0.007*treated_glucose + 1.75*logpulsepr + 0.40*loggciggs
- **P (TenYearCHD = 1)** $= \frac{1}{e^{-logit}}$

$$= \frac{1}{e^{-(-9.15\,+\,0.57*\text{male}\,+\,0.06*\text{age}\,+\,0.65*\text{BPMeds}\,+\,0.46*\text{prevalenthypertension}\,+\,0.007*\text{treated\_glucose}\,+\,1.75*\text{logpulsepr}\,+\,0.40*\text{loggciggs}}}$$

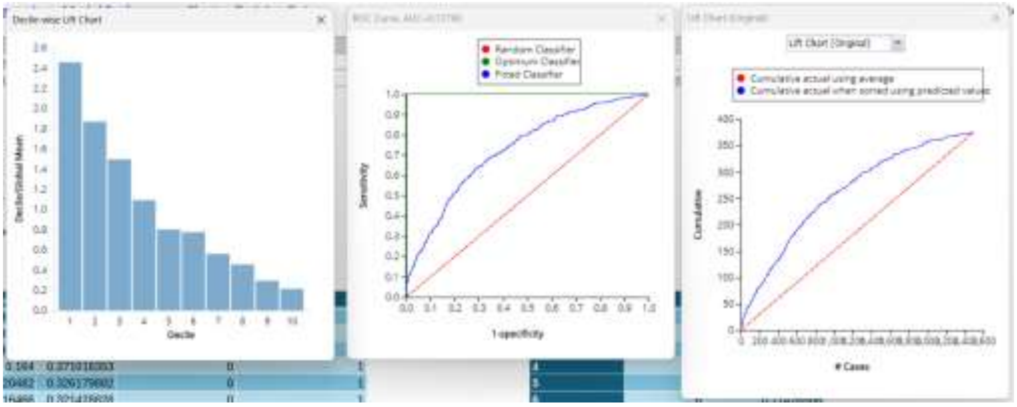
- **Odds (TenYearCHD = 1)** $= e^{-logit}$

$$= e^{-(-9.15\,+\,0.57*\text{male}\,+\,0.06*\text{age}\,+\,0.65*\text{BPMeds}\,+\,0.46*\text{prevalenthypertension}\,+\,0.007*\text{treated\_glucose}\,+\,1.75*\text{logpulsepr}\,+\,0.40*\text{loggciggs}}$$

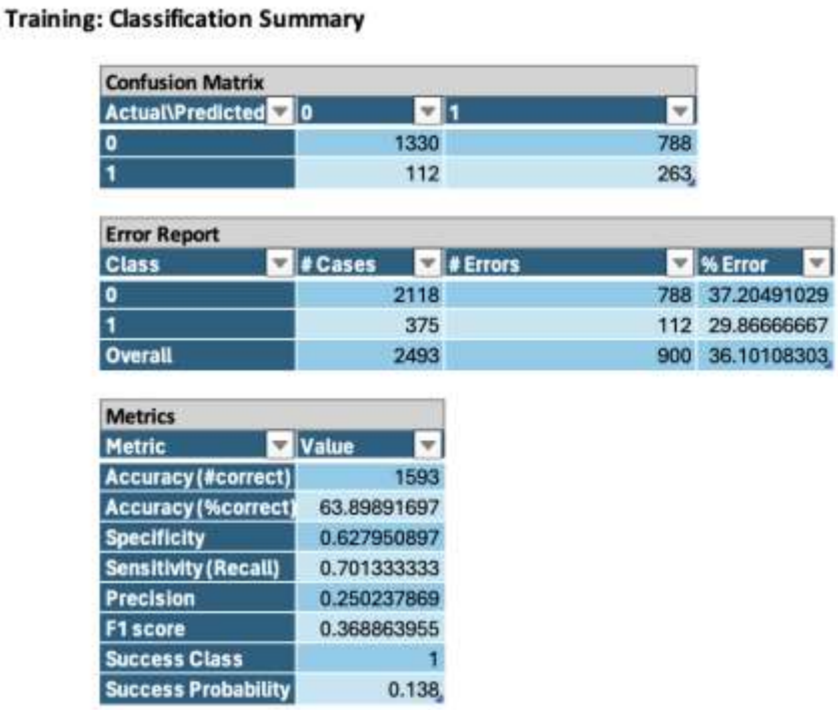
Training and Validation Summaries

Logistic Regression

Training



- The **Decile-wise Lift Chart** indicates that the model is effective at ranking individuals by risk, with the highest deciles showing significantly higher lift
- The **ROC Curve** with an AUC value of 0.72798 indicates that the model has a good ability to distinguish between the positive and negative classes. An AUC close to 1 signifies excellent performance, while an AUC closer to 0.5 indicates poor performance.
- The **Lift Chart** shows that the model is effective in identifying positive cases, with a noticeable improvement over random guessing. The blue line above the red line indicates that the model effectively identifies a higher number of positive cases at the top of the list (those with the highest predicted probabilities). The steeper the blue line, the better the model is at identifying positives early.



Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	1330	788
1	112	263

Error Report			
Class	# Cases	# Errors	% Error
0	2118	788	37.20491029
1	375	112	29.86666667
Overall	2493	900	36.10108303

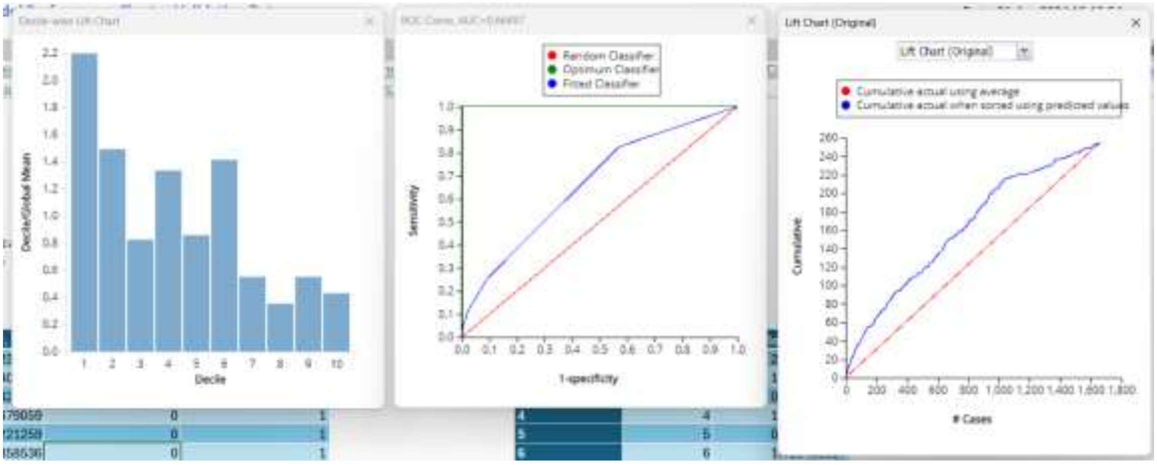
Metrics	
Metric	Value
Accuracy (#correct)	1593
Accuracy (%correct)	63.89891697
Specificity	0.627950897
Sensitivity (Recall)	0.701333333
Precision	0.250237869
F1 score	0.368863955
Success Class	1
Success Probability	0.138

Interpretation

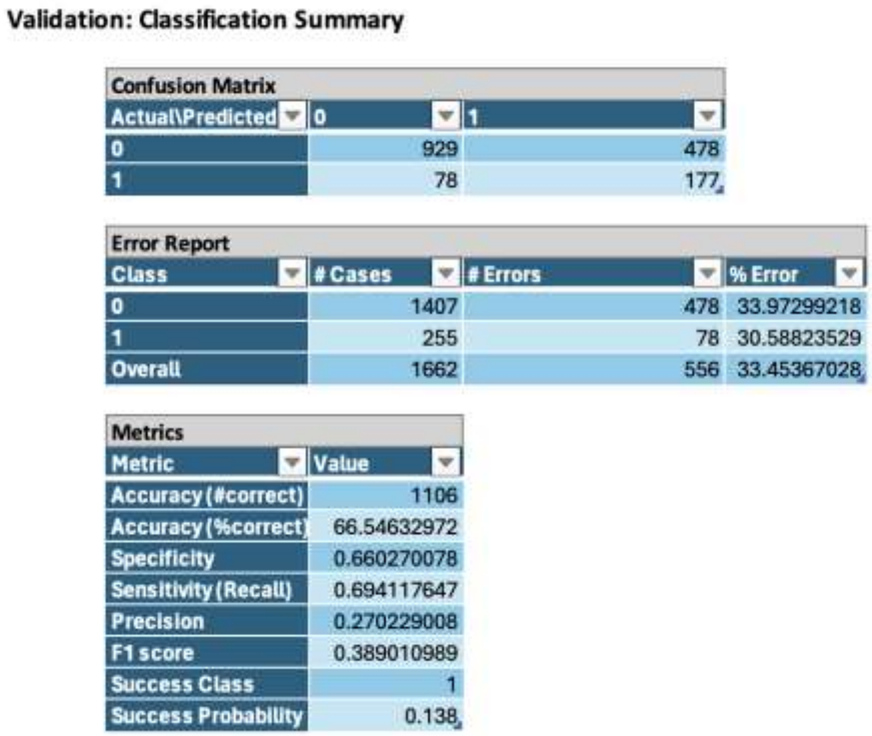
- **Precision (25%):** Out of all the instances that the model predicted as positive, only 25% were actually positive. This indicates a high number of false positives, meaning the model often incorrectly identifies negative cases as positive.
- **Recall (70%):** Out of all the actual positive instances, the model correctly identified 70%. This indicates that the model is relatively good at identifying actual positive cases, but still misses 30% of them.
- **F1 Score (37%):** The F1 score of 37% reflects a significant imbalance between precision and recall. The low precision drags down the overall F1 score despite the higher recall. This score suggests that while the model is somewhat effective at finding the positive cases, it is not very reliable because it also includes many false positives.

The model's performance, indicated by an F1 score of 37%, precision of 25%, and recall of 70%, highlights a need for improvement, particularly in increasing precision. This can be achieved through model tuning, better feature selection, or employing different machine learning techniques, while striving to maintain or improve the recall.

Validation



- The **Decile-wise Lift Chart** indicates that the model is decent at ranking individuals by risk, with the highest deciles showing significantly higher lift
- The **ROC Curve** with an AUC value of 0.66497 indicates that the model is moderately good at distinguishing between the positive and negative classes. An AUC close to 1 signifies excellent performance, while an AUC closer to 0.5 indicates poor performance.
- The **Lift Chart** shows that the model is effective in identifying positive cases, with a noticeable improvement over random guessing. The blue line above the red line indicates that the model effectively identifies a higher number of positive cases at the top of the list (those with the highest predicted probabilities). The steeper the blue line, the better the model is at identifying positives early.



Validation: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	929	478
1	78	177

Error Report

Class	# Cases	# Errors	% Error
0	1407	478	33.97299218
1	255	78	30.58823529
Overall	1662	556	33.45367028

Metrics

Metric	Value
Accuracy (#correct)	1106
Accuracy (%correct)	66.54632972
Specificity	0.660270078
Sensitivity (Recall)	0.694117647
Precision	0.270229008
F1 score	0.389010989
Success Class	1
Success Probability	0.138

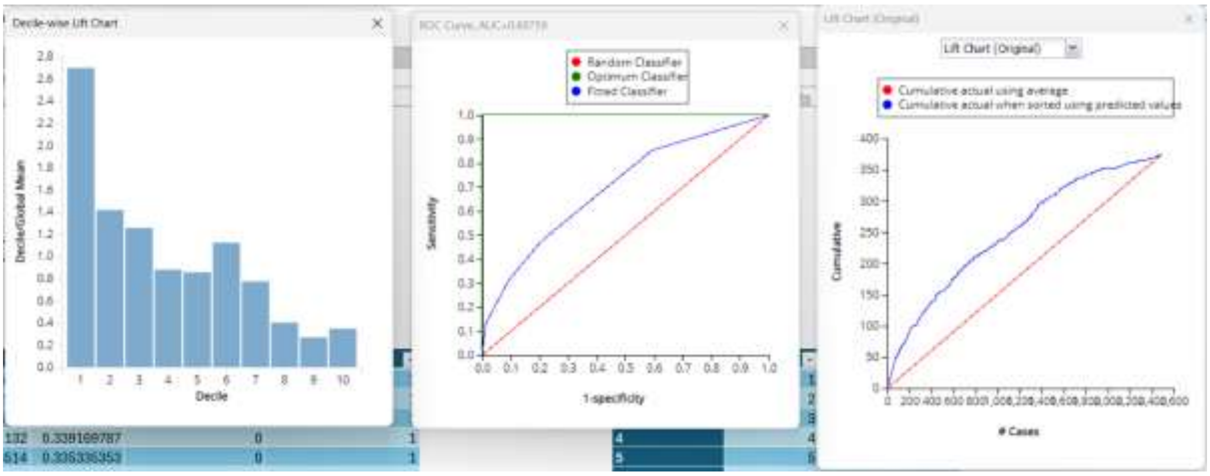
Interpretation

- **Precision (20%):** Out of all the instances that the model predicted as positive, only 20% were actually positive. This indicates a high number of false positives, meaning the model often incorrectly identifies negative cases as positive.
- **Recall (83%):** Out of all the actual positive instances, the model correctly identified 83%. This indicates that the model is relatively good at identifying actual positive cases, but still misses 30% of them.
- **F1 Score (33%):** The F1 score of 33% reflects a significant imbalance between precision and recall. The low precision drags down the overall F1 score despite the higher recall. This score suggests that while the model is somewhat effective at finding the positive cases, it is not very reliable because it also includes many false positives.

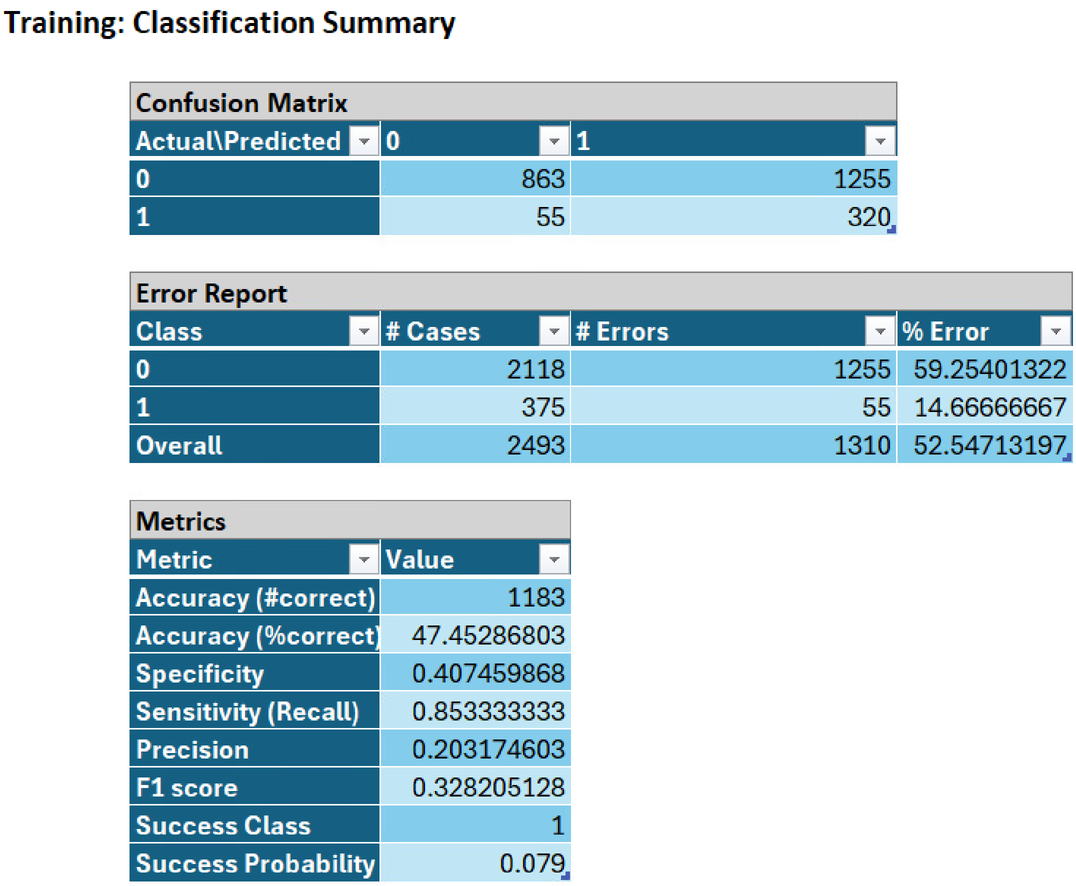
The model's performance on the validation set shows that it is highly sensitive (high recall) but not very specific (low precision). In the context of predicting coronary heart disease (CHD), this means that the model is effective at identifying individuals who are actually at risk (minimizing false negatives), which is critical for early intervention and treatment. However, the high number of false positives (low precision) indicates that many individuals who are not at risk are being flagged by the model, which could lead to unnecessary medical tests and anxiety.

Classification Tree

Training



- The **Decile-wise Lift Chart** indicates that the model is effective at ranking individuals by risk, with the highest deciles showing significantly higher lift compared to the overall average.
- The **ROC Curve** with an AUC of 0.69759 demonstrates moderate discriminative ability of the model.
- The **Lift Chart** shows that the model is effective in identifying positive cases, with a noticeable improvement over random guessing.



Training: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	863	1255
1	55	320

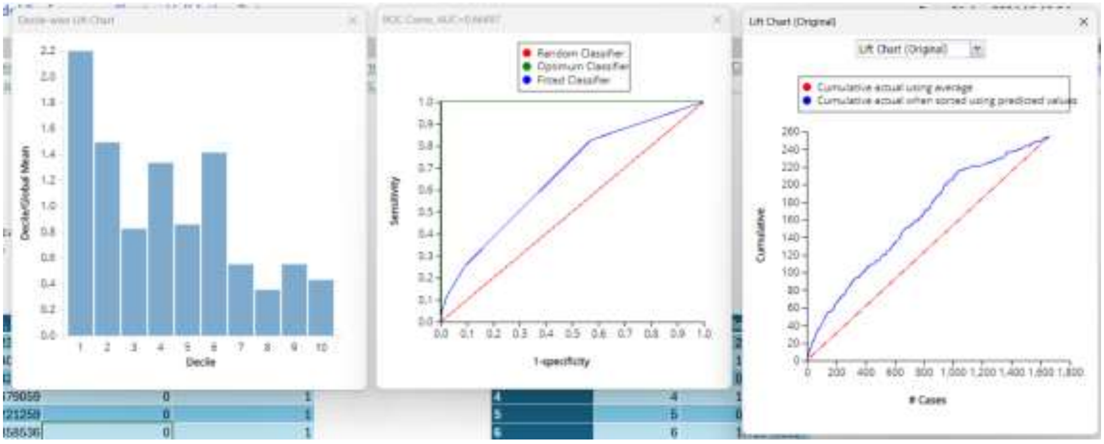
Error Report

Class	# Cases	# Errors	% Error
0	2118	1255	59.25401322
1	375	55	14.66666667
Overall	2493	1310	52.54713197

Metrics

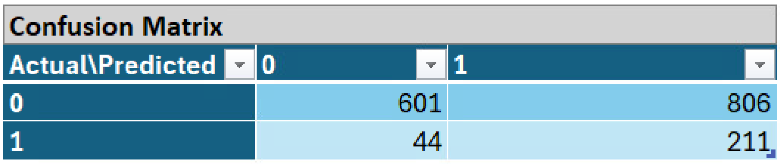
Metric	Value
Accuracy (#correct)	1183
Accuracy (%correct)	47.45286803
Specificity	0.407459868
Sensitivity (Recall)	0.853333333
Precision	0.203174603
F1 score	0.328205128
Success Class	1
Success Probability	0.079

Validation

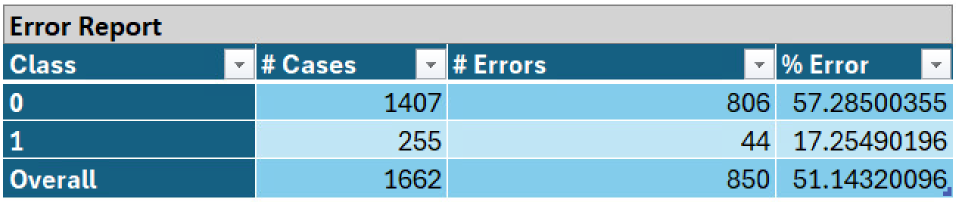


Validation summary and charts show a similar outcome when compared to the training charts, indicating the model is not suffering from overfitting.

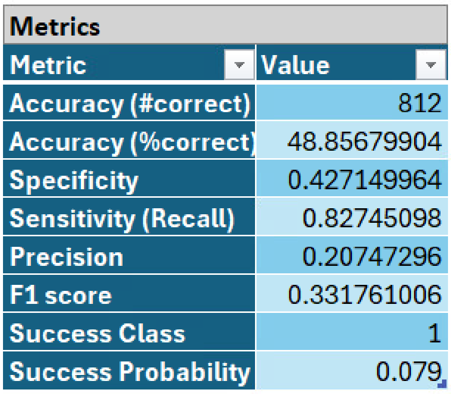
Validation: Classification Summary



Confusion Matrix		
Actual\Predicted	0	1
0	601	806
1	44	211



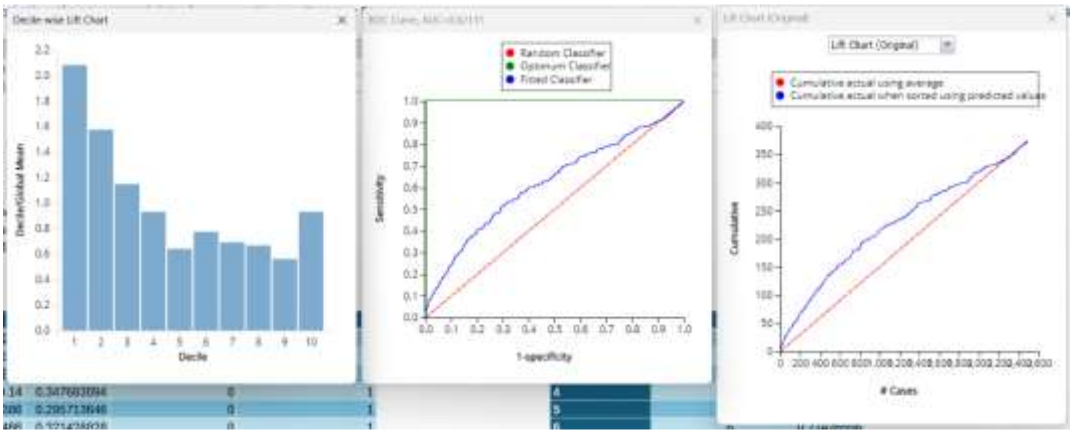
Error Report			
Class	# Cases	# Errors	% Error
0	1407	806	57.28500355
1	255	44	17.25490196
Overall	1662	850	51.14320096



Metrics	
Metric	Value
Accuracy (#correct)	812
Accuracy (%correct)	48.85679904
Specificity	0.427149964
Sensitivity (Recall)	0.82745098
Precision	0.20747296
F1 score	0.331761006
Success Class	1
Success Probability	0.079

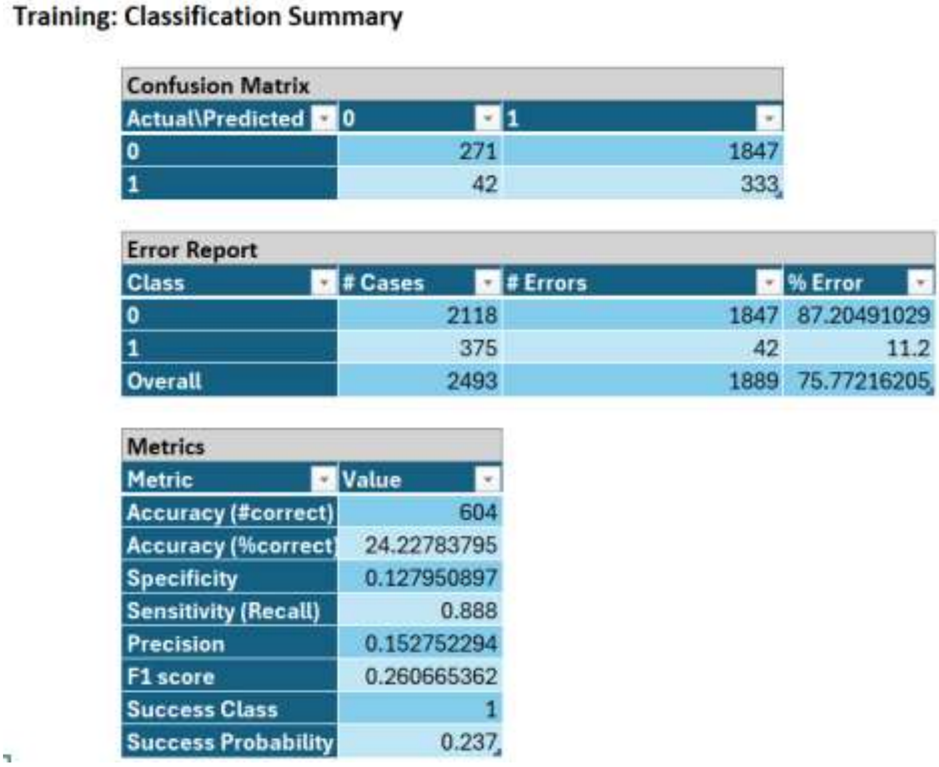
Neural Network

Training



- The **Decile-wise Lift Chart** indicates that the model is effective at ranking individuals by risk, with the highest deciles showing significantly higher lift compared to the overall average.
- The **ROC Curve** with an AUC of 0.69759 demonstrates moderate discriminative ability of the model.
- The **Lift Chart** shows that the model is effective in identifying positive cases, with a noticeable improvement over random guessing.

These charts collectively demonstrate that the logistic regression model has a moderate level of performance in predicting heart disease risk. The model is somewhat effective at distinguishing between high-risk and low-risk individuals, but there is significant room for improvement.



Training: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	271	1847
1	42	333

Error Report

Class	# Cases	# Errors	% Error
0	2118	1847	87.20491029
1	375	42	11.2
Overall	2493	1889	75.77216205

Metrics

Metric	Value
Accuracy (#correct)	604
Accuracy (%correct)	24.22783795
Specificity	0.127950897
Sensitivity (Recall)	0.888
Precision	0.152752294
F1 score	0.260665362
Success Class	1
Success Probability	0.237

Choosing Cutoff Values

By default, each model had a cut off value of 0.5. Since our aim is to maximize the recall value, our new cut off values were chosen keeping that in mind. The decile charts were used to identify the new value, which was a good balance of precision and recall. Here is a table comparing the recall values with default cut off values with the updated ones.

Model	*Original Cut Off*	*Recall at original Cut Off*	*Updated Cut Off*	*New Recall*
Logistic Regression	0.5	0.06	0.138	0.83
Classification Tree	0.5	0	0.079	0.83
Neural Network	0.5	0	0.237	0.89

Answering Business Questions

1. **To what extent does the presence of diabetes increase the likelihood of developing CHD over the ten-year period covered by the study?**
   The presence of diabetes significantly increases the likelihood of developing CHD within ten years. In the logistic regression model, diabetes is one of the predictor variables indicating an elevated risk. Specifically, individuals with diabetes have increased odds of CHD compared to those without diabetes, as indicated by the coefficients in the logistic regression model. Although the exact coefficient for diabetes is not provided in the excerpt, its inclusion in the model suggests a noteworthy impact on CHD risk.
2. **Is there a relationship between early life factors (such as childhood obesity or family history of heart disease) and the likelihood of being labeled 'at-risk' for CHD in adulthood?**
   The dataset and analysis primarily focus on adult health metrics and lifestyle factors rather than specific early life factors. However, variables such as BMI (which could be influenced by childhood obesity) and family history indicators (prevalentStroke, prevalentHyp) are considered in the analysis. The relationship between these variables and CHD risk underscores the long-term impact of early life health and genetics on cardiovascular outcomes.
3. **Is there a correlation between socioeconomic status (such as education level, income, or occupation) and the risk of developing CHD?**
   Education level, as a proxy for socioeconomic status, is included in the dataset and analysis. The variable "education" is considered in the logistic regression model, indicating its relevance in predicting CHD risk. Higher education levels might correlate with lower CHD risk due to better access to healthcare, healthier lifestyles, and higher health literacy.
4. **What predictive factors can be identified as significant predictors of diabetes risk?**

Significant predictors of diabetes risk are not explicitly detailed in the provided excerpts, as the focus is on CHD prediction. However, factors like BMI, glucose levels, age, and smoking status (currentSmoker, cigsPerDay) are likely significant based on their relevance to both diabetes and cardiovascular health.

5. **Can a model be developed to accurately classify individuals at high risk before the onset of the disease?**
   Yes, a logistic regression model has been developed and validated to classify individuals at high risk for CHD. The model uses predictors such as age, gender, smoking status, blood pressure, cholesterol levels, and glucose levels to estimate the likelihood of developing CHD within ten years, demonstrating its potential for early risk identification.
6. **How can the model's predictions be used to inform public health strategies and interventions aimed at preventing diabetes or managing its impact more effectively?**
   The model's predictions can inform targeted public health strategies by identifying high-risk individuals who would benefit from early interventions. Healthcare providers can use these predictions to focus on preventative measures, such as lifestyle modifications, regular monitoring, and medication management. Public health officials can design programs aimed at high-risk populations, promoting healthier lifestyles and improving access to preventive care.

Hypothetical Example

Suppose we want to predict the ten-year CHD risk for a 55-year-old male who smokes 15 cigarettes per day, has a systolic blood pressure of 140, a diastolic blood pressure of 90, a BMI of 28, a heart rate of 80 bpm, and a glucose level of 100 mg/dL. He is not currently on blood pressure medication, has prevalent hypertension but no history of stroke or diabetes.

Using the logistic regression model:
Age = 55, Male = 1, cigsPerDay = 15, sysBP = 140, diaBP = 90, BMI = 28, heartRate = 80, glucose = 100, BPMeds = 0, prevalentHyp = 1
We calculate Pulse Pressure as $sysBP-diaBP$=140−90=50sysBP−diaBP=140−90=50.
Log-transformed variables:
LogPulsePressure = log(50)≈3.91log(50)≈3.91
LogCigs = log(15)≈2.71log(15)≈2.71
Applying these to the logistic regression equation:
Logit($TenYearCHD$=1)=−9.15+0.57·1+0.06·55+0.65·0+0.46·1+0.007·100+1.75·3.91+0.40·2.71Logit(TenYearCHD=1)=−9.15+0.57·1+0.06·55+0.65·0+0.46·1+0.007·100+1.75·3.91+0.40·2.71
Logit($TenYearCHD$=1)=−9.15+0.57+3.30+0+0.46+0.7+6.8425+1.084Logit(TenYearCHD=1)=−9.15+0.57+3.30+0+0.46+0.7+6.8425+1.084
Logit($TenYearCHD$=1)=3.8065Logit(TenYearCHD=1)=3.8065

The odds of CHD is e3.8065≈45e3.8065≈45, and the probability is:

$P(TenYearCHD$**=1)**=e3.80651+e3.8065≈0.978P(TenYearCHD=1)=1+e3.8065e3.8065 ≈0.978
This individual has a high probability (97.8%) of developing CHD within the next ten years.

Model Comparison

- **Logistic Regression**
  Advantages: Simple to implement, interpret, and provides probability estimates for risk.
  Performance: AUC of 0.728 indicates good discriminative ability.
  Limitations: Assumes a linear relationship between predictors and the log odds of the outcome.

- **Classification Tree**
  Advantages: Handles nonlinear relationships and interactions between variables naturally.
  Performance: More interpretable due to its decision-making structure but can be prone to overfitting.
  Limitations: Requires pruning to avoid overfitting and may not perform as well with small datasets.

- **Neural Network**
  Advantages: Can model complex relationships and interactions in the data.
  Performance: Potentially higher accuracy with sufficient data and tuning.
  Limitations: Requires more data and computational resources, less interpretable than logistic regression or trees.

Overall, logistic regression was chosen for its balance of interpretability and performance, making it suitable for healthcare applications where understanding the model's decisions is crucial

Conclusions and Recommendations

The logistic regression model effectively identifies key risk factors for coronary heart disease, providing actionable insights for healthcare providers and policymakers. By focusing on variables like age, smoking status, and medical history, targeted interventions can be developed to mitigate the risk of CHD. Regular screening, lifestyle modifications, and managing conditions like hypertension and diabetes are essential preventive measures. Insurance companies can use the model to adjust premiums and create targeted health programs, ultimately improving patient outcomes and reducing healthcare costs.