Unmasking Cardiovascular Disease

Early Detection of Heart Disease
via the Power of
Machine Learning

Agenda Overview

- Defining the Main Objective of the Analysis
- Overview of the Heart Disease Dataset
- Data Exploration, Cleaning, and Feature Engineering
- Summary of Model Training Process
- Key Findings From the Predictive Model
- Recommended Next Steps for Future Work

Defining the Main Objective

Shining a Light on a Silent Killer

The Scope and The Solution

The Impact

Cardiovascular diseases remain the leading cause of death worldwide, responsible for approximately 17.9 million deaths annually and accounting for 31% of all global fatalities. Heart attacks and strokes cause 80% of these deaths, with one-third occurring prematurely in individuals under 70 years old.

Early Risk Identification

Detecting patients at high risk for cardiovascular disease is essential for early intervention, which can greatly enhance treatment success and lower the risk of complications.

Predictive Tool Development

We will create a tool that can reveal hidden patterns and complex connections among risk factors to not only forecast heart disease but also pinpoint the key contributors to that prediction.

Data Overview

The Evidence That Remains

UCI Heart Disease Dataset

The UCI Heart Disease dataset is part of the UCI Machine Learning Repository and is designed to assist researchers and practitioners in diagnosing heart disease. Primarily derived from a Cleveland-based database, it contains 12 key features that are crucial for analysis and predictive modeling of heart disease.

The original dataset was contained 918 rows and 12 columns. Of which 5 were object or categorical values, 6 integer values, and 1 float value.

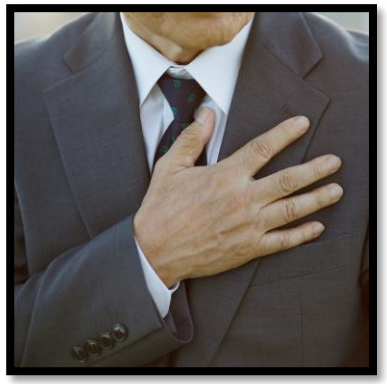
There were no duplicated or missing values in the dataset.

By exploring the dataset, we hope to derive insights to help us in data cleaning, feature engineering, and ultimately successfully predicting heart disease optimized for reducing missed diagnosis (recall).

Data Dictionary: Core Variables

Age:
Age, measured in years.

Sex:
The gender of the patient.
Male = 1, Female = 0.



Chest Pain Type:
Categorized into four types:
typical angina, atypical angina,
non-anginal pain, asymptomatic.

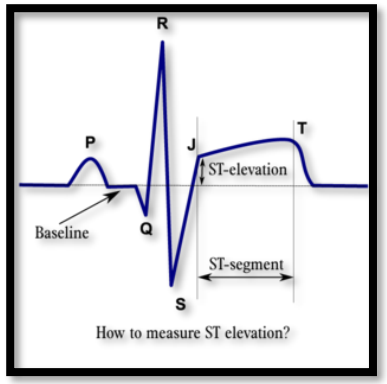
Resting Blood Pressure:
Systolic pressure
measured in mm Hg at the time
of hospital admission.

Cholesterol:
Serum cholesterol levels in mg/dL.

Fasting Blood Sugar:
Indicates whether fasting blood
sugar is >120 mg/dL
(1 = true; 0 = false).

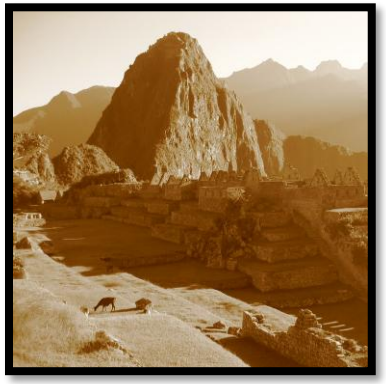
Core Variables cont.

*



Resting ECG:
Electrocardiographic results: Normal, ST-T wave abnormalities, or left ventricular hypertrophy.

**



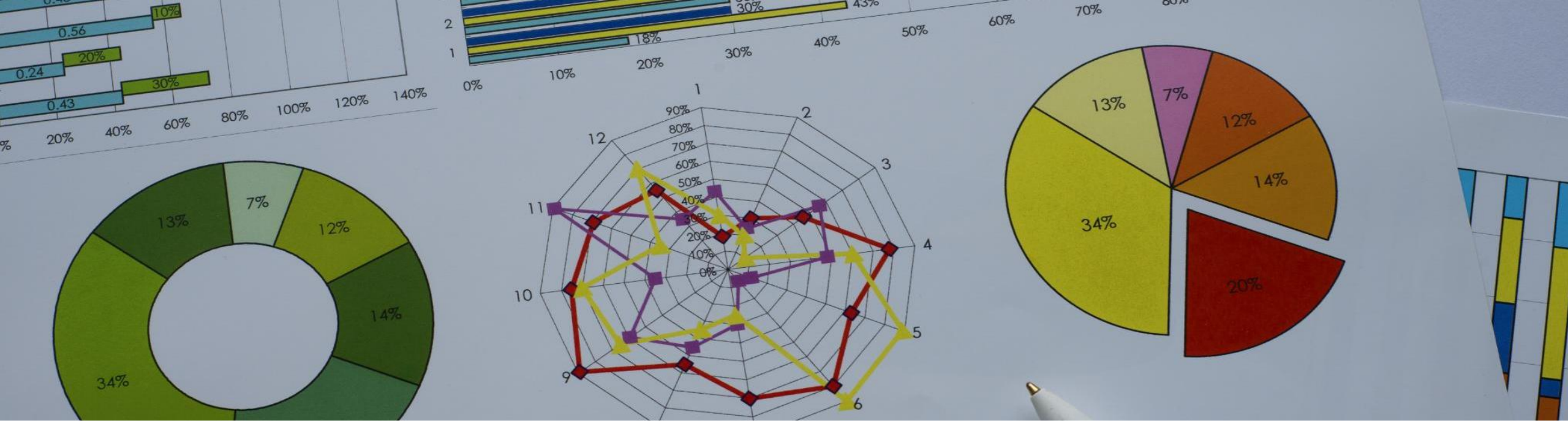
Old Peak:
ST depression induced by exercise relative to rest.

Maximum Heart Rate:
Electrocardiographic results: Normal, ST-T wave abnormalities, or left ventricular hypertrophy.

ST Slope:
The slope of the peak exercise S T segment (upsloping, flat, down sloping).

Exercise Angina:
Presence or absence of exercise-induced angina (1 = yes; 0 = no).

Heart Disease:
Our target variable. (1 = heart disease, 0 = Normal).



Statistical Summary

Numerical Features:

•**Age**: 28-77 years (mean: 53.5)
•**Resting BP**: 0-200 mmHg (median: 130); ~50% elevated/high; 0 = missing data
•**Fasting Blood Sugar**: Binary (median: 0); most patients non-diabetic/prediabetic
•**Max Heart Rate**: 60-202 bpm (mean: 136.8); normally distributed
•**Oldpeak**: -2.6 to 6.2 (median: 1.0); 75% ≤1.5 indicating good to moderate ST depression
•**Disease Prevalence**: ~55% diagnosed with heart disease

Categorical Features:

•**Sex**: 75% male
•**Chest Pain Type**: ~50% asymptomatic (most common)
•**Resting ECG**: Slightly over 50% normal
•**Exercise Angina**: ~50% no angina
•**ST_Slope**: ~50% flat (suspicious for disease)

Data Exploration, Cleaning, and Feature Engineering

Close Behind the Suspect's Path

EDA Methodology

Distribution Analysis

We began by plotting the distributions of each of the variables in the dataset individually to look for levels of anomalous values, skewness and other contributing factors that would inform our decisions in further data cleaning.

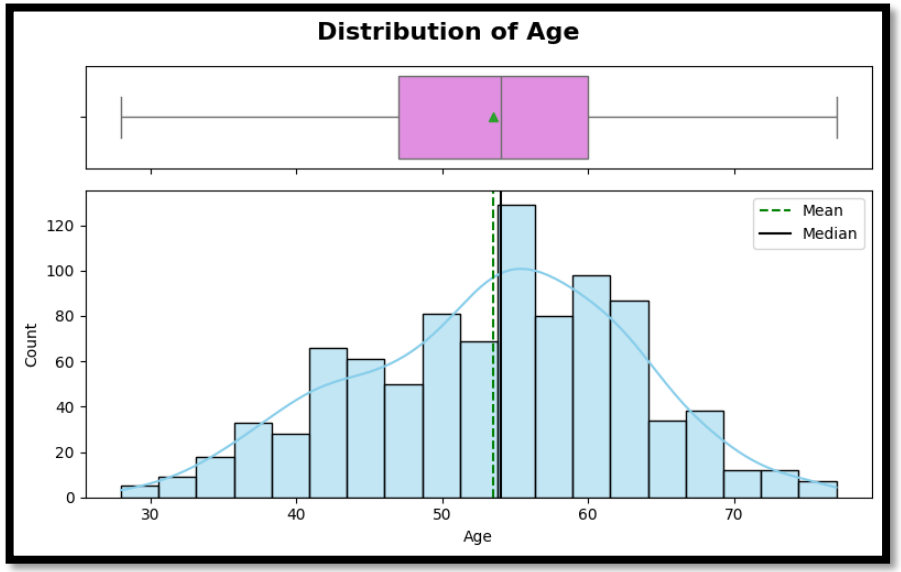
Data Cleaning

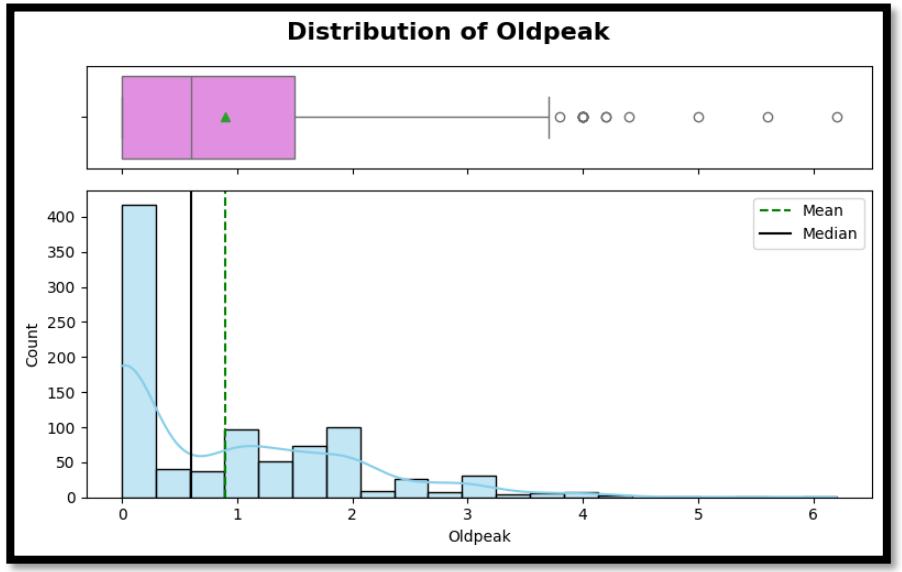
We looked closely for anomalous values or inconsistent data and uncovered one row that ultimately had to be dropped. We also found multiple values within the Cholesterol variable that later had to be imputed for the model to work well.

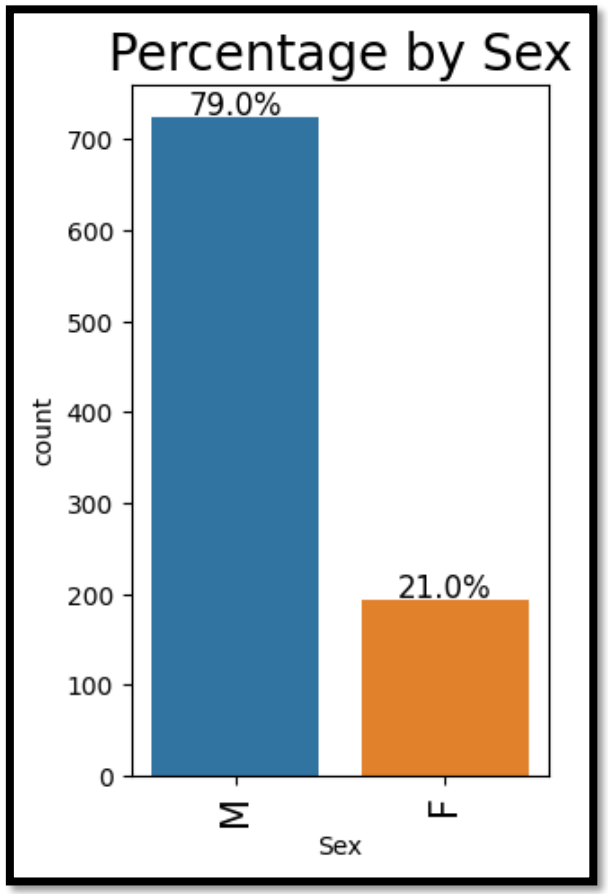
Feature Engineering

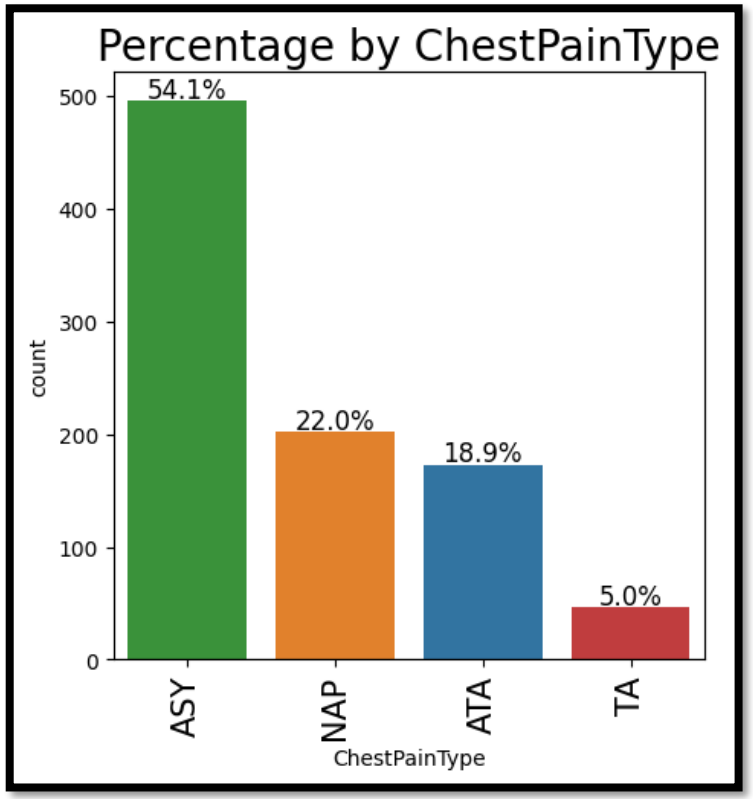
After a thorough review of the bivariate analysis, we were able to identify key correlations and variable interactions that led to the development of three new variables to assist in accurate predictions.

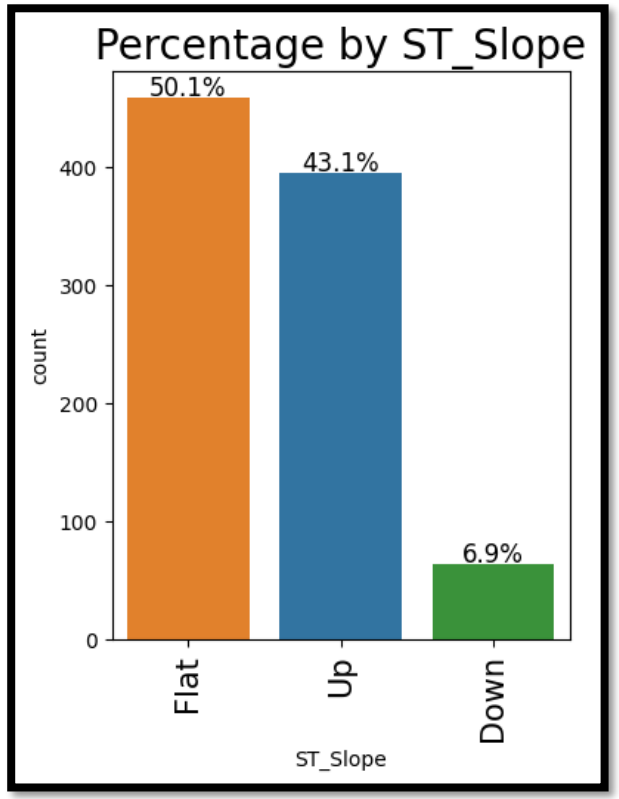
Key Variable Data Distributions











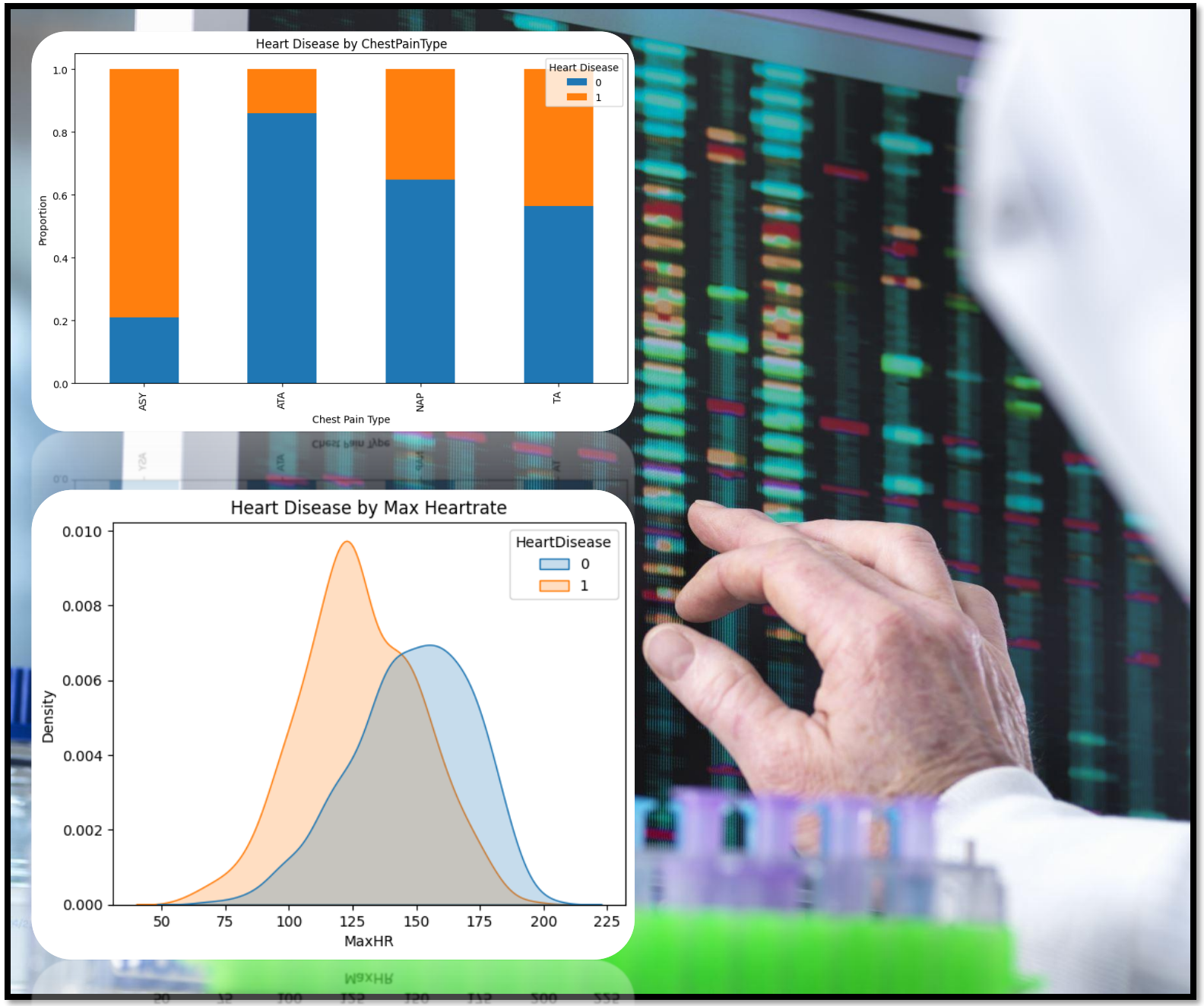
Handling Missing Values and Data Inconsistencies

Cholesterol Imputation

We found 171 rows in the dataset in which Cholesterol value was “0”. To fill the “missing” values, we determined to group the dataset by Age, Sex, and RestingBS as those are key variables for estimating cholesterol levels. We calculated the means of those groups and imputed over the missing values.

Consistency Checks

We identified only one row in the dataset that was missing multiple values. Since it was an anomalous entry (meaning we couldn’t find any other patients in the dataset with similar values in the other variable fields), we ultimately made the decision to drop that row from the dataset.



Feature Engineering Strategy

Correlation Analysis

Through our Bivariate analysis, we determined the top six most highly-correlated variables to be ST_Slope, Cholesterol, ChestPainType, MaxHR, ExerciseAngina, and Oldpeak.

Predictors in Area of Overlap

Patients who are asymptomatic for chest pain were highly likely to have heart disease, which we found perplexing. Diving deeper, we noted that Oldpeak, ST_Slope, and FastingBS values were significantly different within asymptomatic patients with and without heart disease. We also found the same to be true within the average value range of the MaxHR variable.

Feature Engineering

As we noted earlier, Oldpeak, FastingBS and ST_Slope are all very helpful in predicting heart disease among those that are asymptomatic for Angina and have average MaxHR. As such, we created three new features: Oldpeak x FastingBS (OPxBS) to cancel out the risk associated with elevated Oldpeak unless the patient is showing diabetic BS. “Oldpeak_High” for those with Oldpeak values of >= 1.20, and “OPxBSxST” in which we'll multiply OPxBS and ST_Slope after we giving ST_Slope an ordinal value that signifies increasing risk (upsloping = 0, flat = 1, down = 2).

Model Training and Evaluation

Putting the Right Detective on the Case

Selection and Justification of Modeling Techniques

Model Selection Criteria

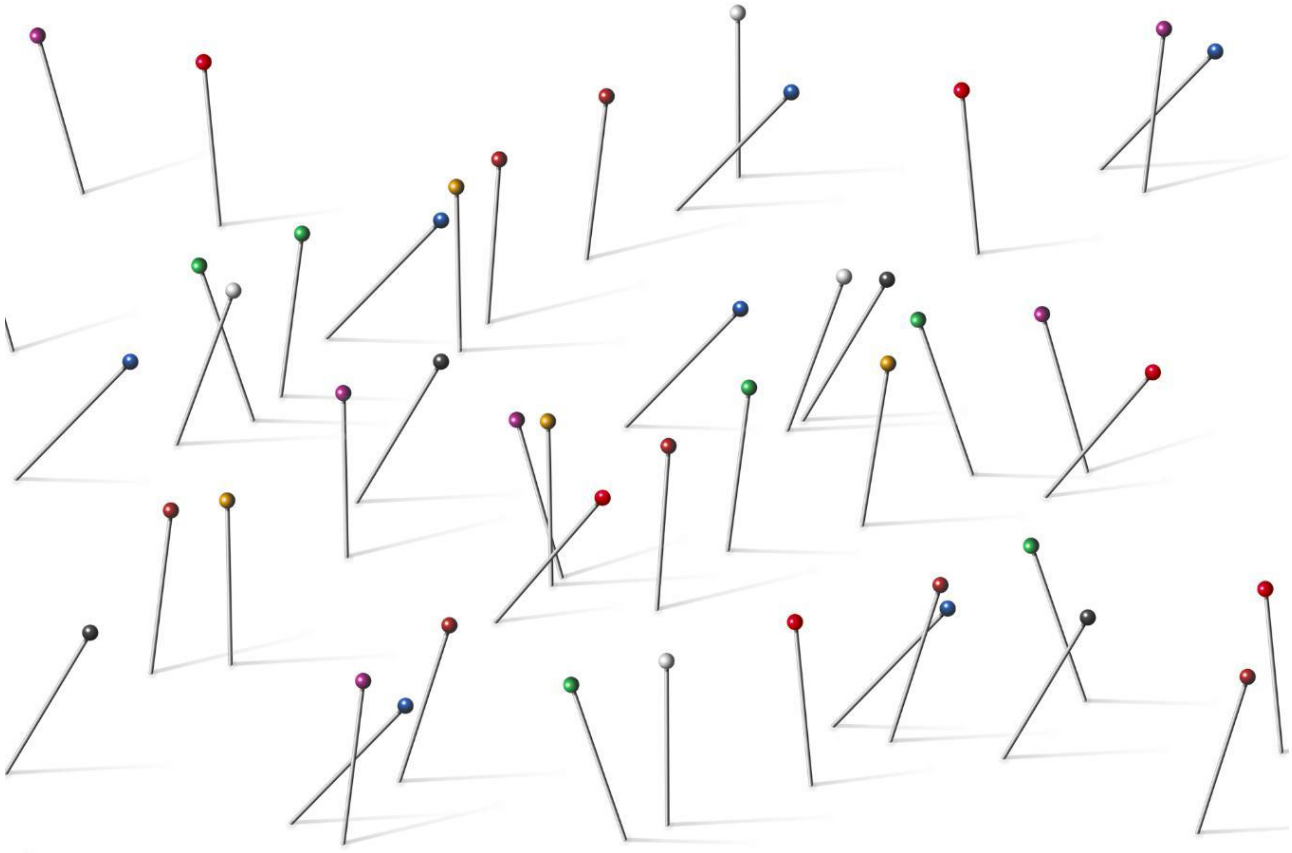
Five models were selected to address both linear and non-linear patterns within the data, emphasizing not only prediction accuracy but also interpretability. Each model will be optimized to prioritize recall, given the critical importance of minimizing missed diagnoses.

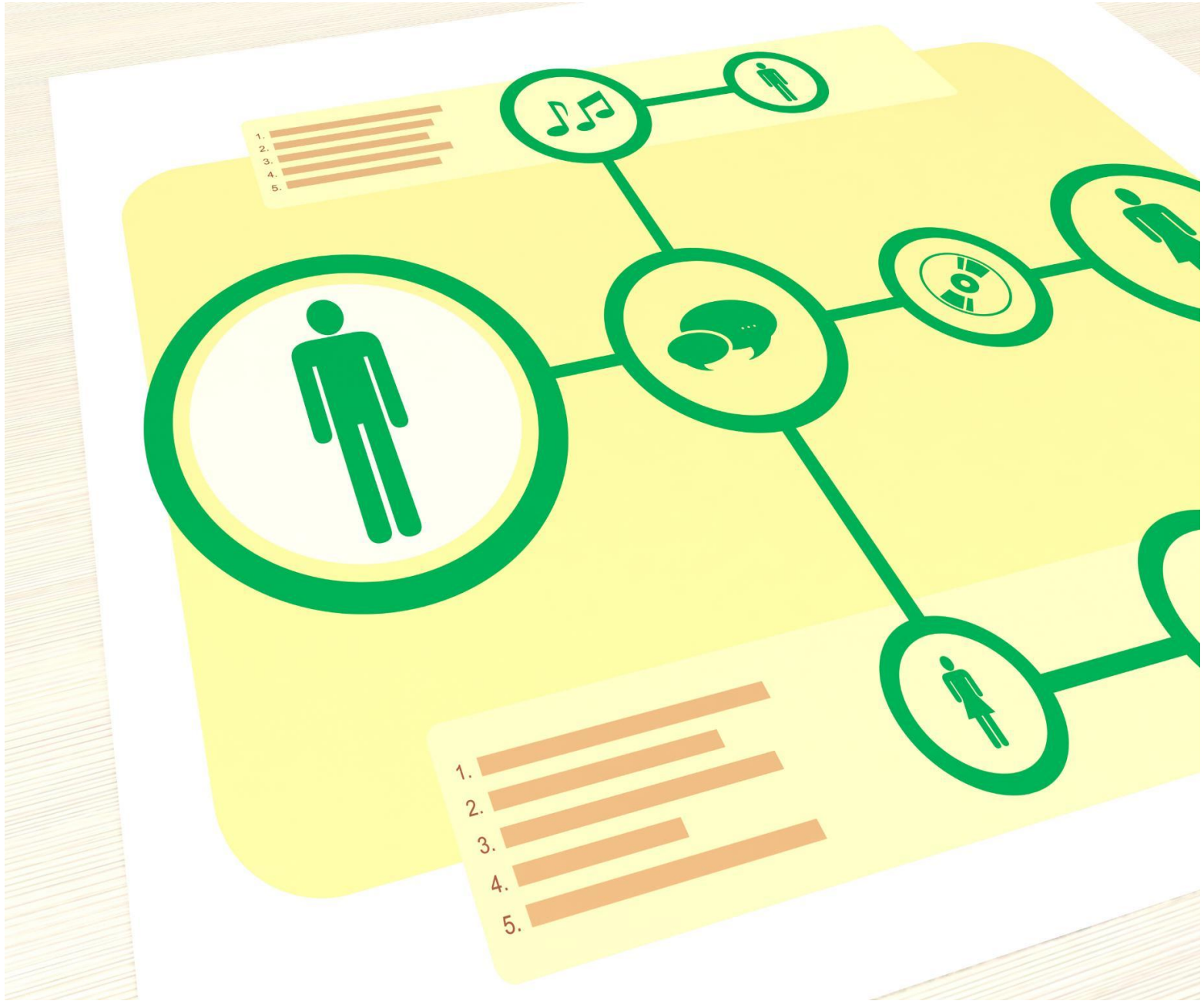
Linear-Oriented Models

Logistic Regression and Support Vector Classifier were included to capture linear relationships. Additionally, Logistic Regression can serve as a surrogate model to interpret more complex black-box models if needed.

Tree-Based and Ensemble Models

Decision Tree, Random Forest, and XGBoost were selected for their strength in identifying intricate patterns within the data.





Training, Validation, and Testing Methodology

Base Model Training

We started by applying a log transformation to skewed variables (Cholesterol and Oldpeak) and scaling the data for linear models. After one-hot encoding the categorical variables, we trained all five models with their default settings. The Decision Tree model performed the poorest and was excluded following initial training and evaluation.

Tuning and Validation

The four remaining models underwent tuning with GridSearchCV, focusing on optimizing recall by using an fbeta_score with a beta value of 2. Their metrics were compared and a final model chosen.

Performance Evaluation Metrics

Accuracy

Accuracy measures the overall correctness of a model by calculating the proportion of true results among total cases.

Precision

The measure of the proportion of positive predictions (heart failure) that are actually correct. It focuses on reducing false positives – in this case not diagnosing a patient as having heart disease and increasing medical costs and emotional stress.

Recall – Our primary metric

Recall, also known as the true positive rate, is our primary metric for model evaluation. It measures the proportion of actual positives correctly identified by the model and minimizes false negatives (misdiagnosing an unhealthy patient as healthy).

F1 Score – our secondary metric

The harmonic mean of precision and recall.

Final Model Selection

Our **SVC** model emerged as the recommended classifier based on performance metrics of a 97% true positive rate (recall score) – 2% higher than any other model. It was also the least overfit meaning it will generalize well to unseen data.

Model	Tuned LR Train	Tuned LR Test	Tuned SVC Train	Tuned SVC Test	Tuned RF Train	Tuned RF Test	Tuned XGB Train	Tuned XGB Test
Accuracy	0.825375	0.836957	0.856753	0.831522	0.885402	0.858696	0.85266	0.826087
Precision	0.774257	0.795082	0.802419	0.779528	0.837895	0.822034	0.826374	0.801724
Recall	0.965432	0.95098	0.982716	0.970588	0.982716	0.95098	0.928395	0.911765
F1 Score	0.859341	0.866071	0.883463	0.864629	0.904545	0.881818	0.874419	0.853211

Key Findings and Feature Importances

Obtaining a Confession

Important Predictive Features

ST_Slope is the most significant indicator of heart disease

Among all features, downsloping and flat ST segments during stress tests are strong markers of heart disease, while upsloping segments are typically associated with healthy individuals.

Patients without chest pain symptoms (ASY) face greater risk

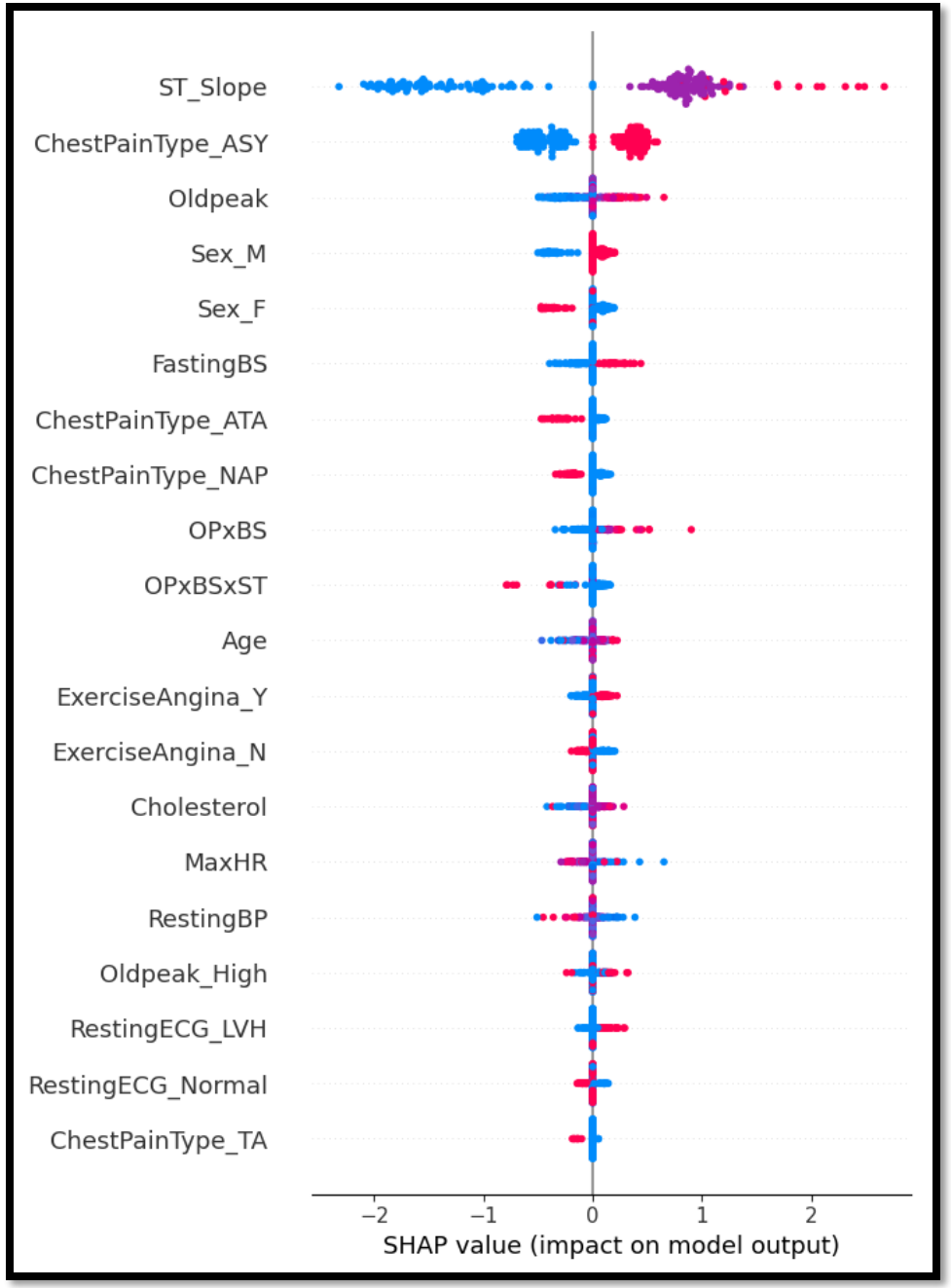
This highlights the unexpected threat of “silent” heart disease, which lacks clear warning signs. Conditions such as diabetes, chronic heart disease, or severe arterial blockages can reduce the sensation of chest pain.

Using multiple measurements together enhances understanding

A combination of elevated Old Peak and Resting Blood Sugar levels provides moderate predictive value, indicating that the interaction between ST depression and metabolic factors offers additional insight, especially in cases where individual measures alone overlap between those with and without heart disease.

Traditional risk factors have limited impact

Changes observed during exercise are far more effective at predicting heart disease compared to conventional risk factors like cholesterol, age, and blood pressure, underscoring the greater relevance of dynamic stress test results over static laboratory measures.



Minimizing Risk – Features that Reduce Recall

Gender: Males Need a Second Look

Sorry, Guys. Men are much more likely to develop heart disease, and gender is the number 1 feature in making sure we catch as many cases of heart disease as we can.

Age: Older Patients are more Prone

It's intuitive, and makes sense, but removing age from our model increases the chance of missed diagnosis. This means age plays a strong role in prediction and older patients are more at risk.

Old Peak x Fasting Blood Sugar x ST_Slope

The unique interaction is essential in ensuring we don't miss any diagnosis of heart disease. While patients with high Old Peak values, diabetic symptoms, or flat or downsloping ST slope are indicative of heart disease, the unique combinations of their values can help to diagnose heart disease in patients who might otherwise be predicted healthy using their individual values alone.

Getting Ready for Prosecution

Potential Improvements to Model Performance

Model Stacking

Although the SVC model achieved the highest recall, its precision was only around 80%. The Random Forest model, with a 95% recall, outperformed by about 2% in both F1 score and precision. Combining these two models, or potentially fine-tuning another model to emphasize F1 or precision and averaging all three results, could improve prediction accuracy.

Further Feature Engineering

The features used in this model were developed based mainly on the data scientist's intuition. Collaborating with medical experts and applying deeper domain knowledge to analyze the data could enhance feature engineering, potentially increasing the model’s overall accuracy and pushing recall closer to 98 or 99%.

Revisit the Cholesterol Values

Since a significant portion of the dataset (18%) contains imputed Cholesterol values that may not reflect true measurements, it would be beneficial to review the data collection methods. Obtaining the actual values and retraining the model with updated data could lead to improved performance.

Conclusion

Early Detection Benefits

Predictive modeling provides valuable tools that enable early detection of heart disease, improving patient prognosis.

Model Refinement Importance

Ongoing refinement and validation ensure the reliability and accuracy of predictive models in clinical settings.

Clinical Impact

Validated models support clinical applications that can enhance treatment plans and patient outcomes.