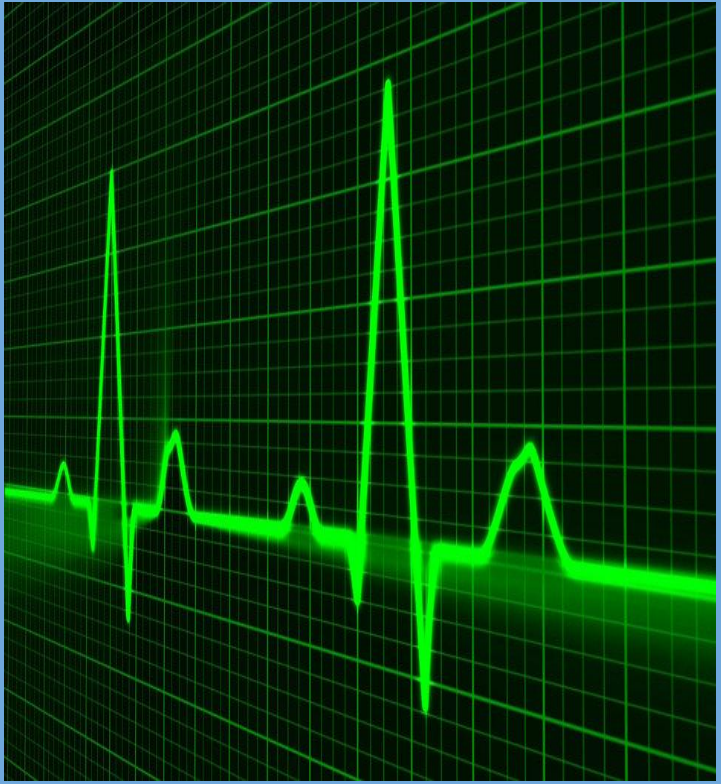
Pfizer Heart Disease Data Case Study

Lenin Kamma
October 2023



Case Study

Objective:

- Use a predictive model to identify the contributing factors toward heart disease
- Put together a mockup of visual to share the results of the analysis

Patient level dataset:

- Demographic, health outcomes and a Target Flag
- Heart disease patient data dictionary

Assumptions:

- The data in the dataset is assumed correct
- The dataset has only a limited observations

Key Terms:

Target, Output

Feature, Variable, Predictor, Input

Case Study High Level Steps

1. Data Import and Pre-processing
2. Exploratory Data Analysis
3. Model Selection and Implementation
4. Evaluation & Performance Measurement
5. Conclusions and Insights

1. Data import, cleaning, duplicate check, and initial investigation
2. Exploring data to understand its key characteristics, uncover patterns, visualizations and feature selection
3. Choosing the most appropriate ML algorithm and implementing it using training and test data
4. Evaluation, identifying metrics and performance measurement
5. Interpret the results and provide answers

1) Data Import and Pre-processing

Total Observations	303
Total Observations with Missing Values	0
Total Variables*	14
Total Binary Features	4
Total Numerical Features	5
Total Categorical Features	5
Duplicates Rows	0

*Target is included

- There are five instances where the major vessels colored by fluoroscopy (ca) have a value of 4, whereas the acceptable range is 0-3.
- Additionally, two instances have a Thalassemia value of zero, but the valid range is 1-3.
- In real world scenario Investigation is needed to understand why these values are not in the range

2) Exploratory Data Analysis - Data patterns

- Significant class imbalance with respect to gender

  96 - Female; 207- Male

- 23.7% of the Female and 30.7% of Male have heart disease.

  This shows Men are at higher risk

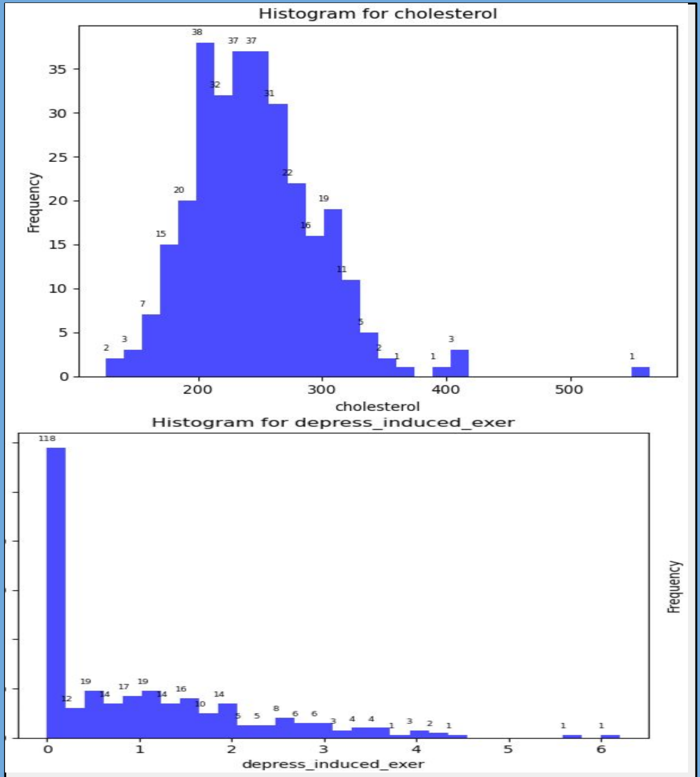
- Blood Cholesterol levels

  4 observations > 400 with confirmed heart disease

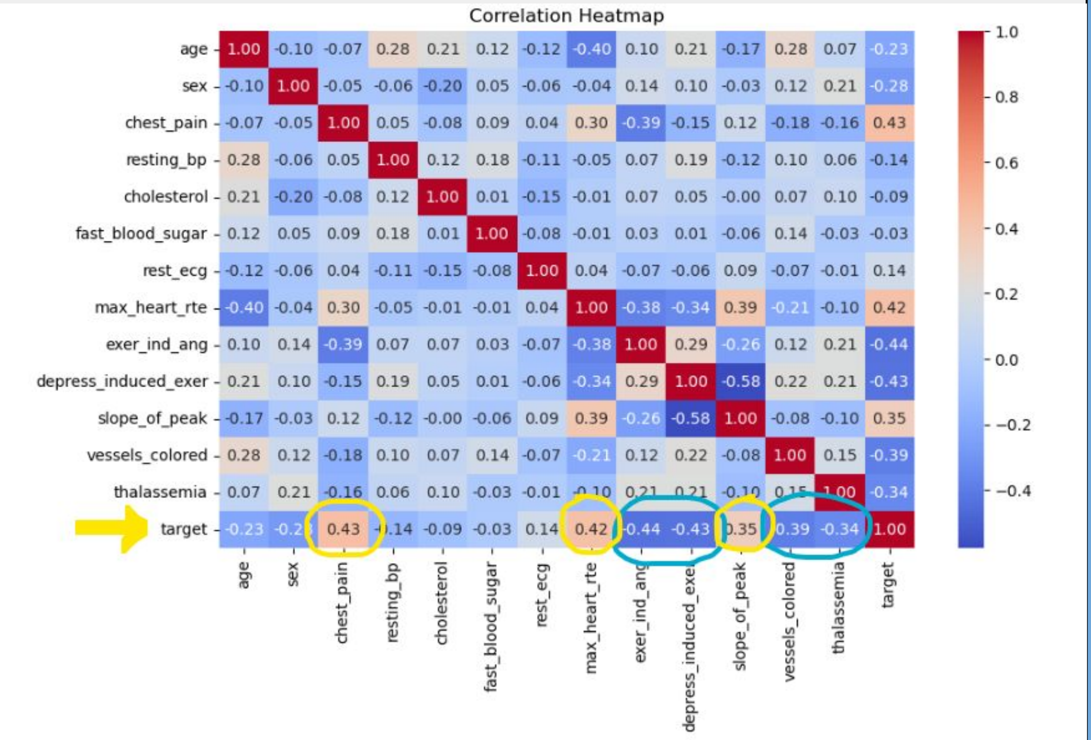
  Higher cholesterol levels can increase the risk of developing heart disease

- ST depression induced by exercise relative to rest is heavily right skewed as more values fell on left

  https://shorturl.at/rDSY0 - National Library of Medicine



2) Exploratory Data Analysis - Visualization



	Correlation
Chest_Pain	0.43
Max_Heart_Rate	0.42
Slope_of_Peak	0.35
Exercise_Angina	-0.44
ST depression induce	-0.43
Vessels_Colored	-0.39
Thalassemia	-0.34

2) Exploratory Data Analysis - Feature Selection

Identify and select a subset of the most relevant features

1. Recursive Feature Elimination (RFE):

- Rank the features based on their importance scores
- Remove the least important features and repeat the process

2. L1 Regularization (Lasso):

- This method removes some features by making their coefficients zero
- The non-zero coefficients correspond to selected features.

Standard Scaler is used to standardize the input

Original

chest_pain	resting_bp
3	145
2	130
1	130
1	120
0	120

Scaled

chest_pain	resting_bp
1.973123	0.763956
1.002577	-0.092738
0.032031	-0.092738
0.032031	-0.663867
-0.938515	-0.663867

2) Exploratory Data Analysis - RFE

Estimator Selection →

- Logistic Regression
- Gradient Boosting Machine
- Random Forest
- Decision Tree

→ Estimator is used to rank the important features

Classifier Selection →

- Decision Tree Classifier
- Logistic Regression Classifier

→ Classifier is then used to find the model accuracy by using the ranked feature

3) Model Selection and Implementation

Top two models with least # of input features are selected

	Estimator	Model	Total Features Selected	Selected Features	Ranking	Accuracy
0	Logistic Regression	DecisionTreeClassifier()	8	[sex, chest_pain, max_heart_rte, exer_ind_ang, depress_induced_exer, slope_of_peak, vessels_colored, thalassemia]	[2, 1, 1, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1]	1.000000
1	Logistic Regression	LogisticRegression()	8	[sex, chest_pain, max_heart_rte, exer_ind_ang, depress_induced_exer, slope_of_peak, vessels_colored, thalassemia]	[2, 1, 1, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1]	0.851485
2	Random Forest	DecisionTreeClassifier()	13	[age, sex, chest_pain, resting_bp, cholesterol, fast_blood_sugar, rest_ecg, max_heart_rte, exer_ind_ang, depress_induced_exer, slope_of_peak, vessels_colored, thalassemia]	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]	1.000000
3	Random Forest	LogisticRegression()	13	[age, sex, chest_pain, resting_bp, cholesterol, fast_blood_sugar, rest_ecg, max_heart_rte, exer_ind_ang, depress_induced_exer, slope_of_peak, vessels_colored, thalassemia]	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]	0.851485
4	Decision Tree	DecisionTreeClassifier()	13	[age, sex, chest_pain, resting_bp, cholesterol, fast_blood_sugar, rest_ecg, max_heart_rte, exer_ind_ang, depress_induced_exer, slope_of_peak, vessels_colored, thalassemia]	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]	1.000000
5	Decision Tree	LogisticRegression()	13	[age, sex, chest_pain, resting_bp, cholesterol, fast_blood_sugar, rest_ecg, max_heart_rte, exer_ind_ang, depress_induced_exer, slope_of_peak, vessels_colored, thalassemia]	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]	0.851485
6	Gradient Boosting	DecisionTreeClassifier()	13	[age, sex, chest_pain, resting_bp, cholesterol, fast_blood_sugar, rest_ecg, max_heart_rte, exer_ind_ang, depress_induced_exer, slope_of_peak, vessels_colored, thalassemia]	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]	1.000000
7	Gradient Boosting	LogisticRegression()	13	[age, sex, chest_pain, resting_bp, cholesterol, fast_blood_sugar, rest_ecg, max_heart_rte, exer_ind_ang, depress_induced_exer, slope_of_peak, vessels_colored, thalassemia]	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]	0.851485

3) Model Selection and Implementation

Decision Tree Classifier:

A decision tree classifier is like a flowchart that splits data by following a tree of yes-or-no questions

Max depth parameter:

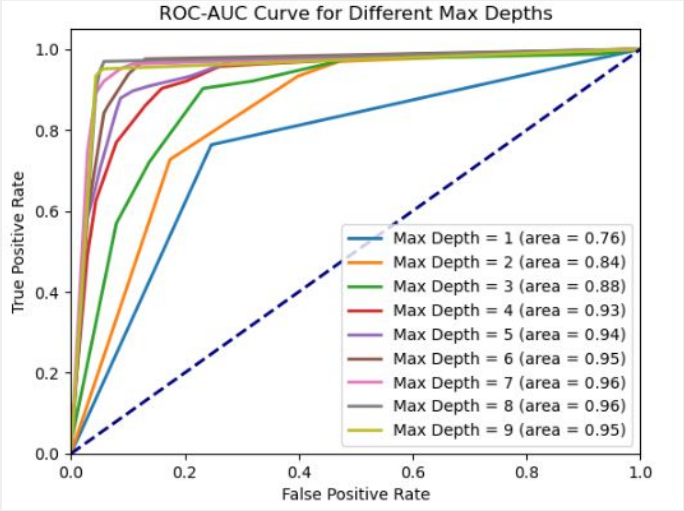
Controls how deep the tree can go, to prevent overfitting

ROC-AUC Curve:

The ROC curve shows the tradeoff between true positive rate and true negative rate at different thresholds

Max-Depth of 4 is selected

Max_Depth	1	2	3	4	5	6	7
Accuracy	0.8361	0.7869	0.8197	0.8525	0.8197	0.8197	0.7869



4) Evaluation and Performance Measurement

Decision Tree Classifier

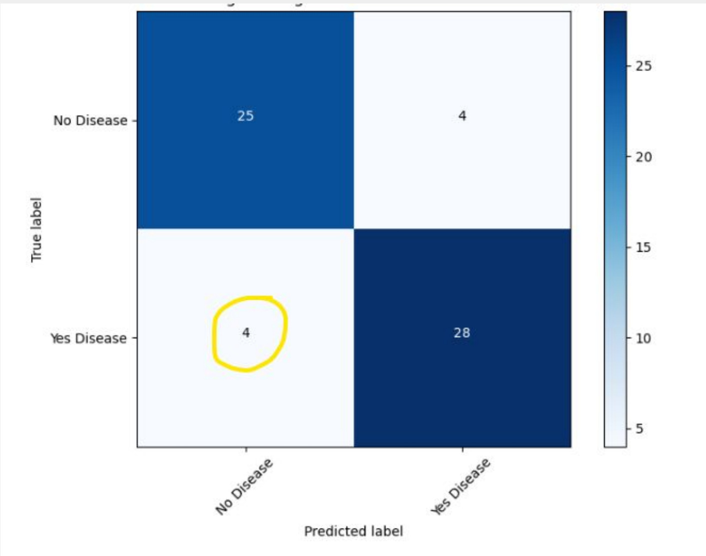
Logistic Regression Classifier:

Lasso Regularization Model:

All 3 have same

False Positive Rate (FPR) of 12.5%

False Negative Rate (FNR) of 13.7%



4) Evaluation and Performance Measurement

Model	Accuracy
Decision Tree Classifier	85
Logistic Regression Classifier	87
Logistic Regression (Lasso Reg)	87

Since All 3 Models have the same FPR/TNR rate, the Model with the highest Accuracy is the best choice

4) Evaluation and Performance Measurement - Conclusion

Model	Precision	
Decision Tree Classifier	81	90
Logistic Regression Classifier	86	88
Logistic Regression (Lasso Reg)	86	88

Model	Recall	
Decision Tree Classifier	90	81
Logistic Regression Classifier	86	88
Logistic Regression (Lasso Reg)	86	88

Blue - “No Disease” Class

Gray - “Yes Disease” Class

Any one of the Logistic Regression models can be selected because they have the Highest Recall Rate (88%)

Precision: Precision is the proportion of true positive predictions out of all positive predictions

Recall:the proportion of true positive predictions out of all actual positive instances

5) Conclusions

What are the contributing Factors for Heart Disease?

Features	Coefficients
chest_pain	0.797389
vessels_colored	0.781247
depress_induced_exer	0.749359
sex	0.651703
thalassemia	0.554705
exer_ind_ang	0.522403
slope_of_peak	0.403561
max_heart_rte	0.380143

Logistic Regression Classifier

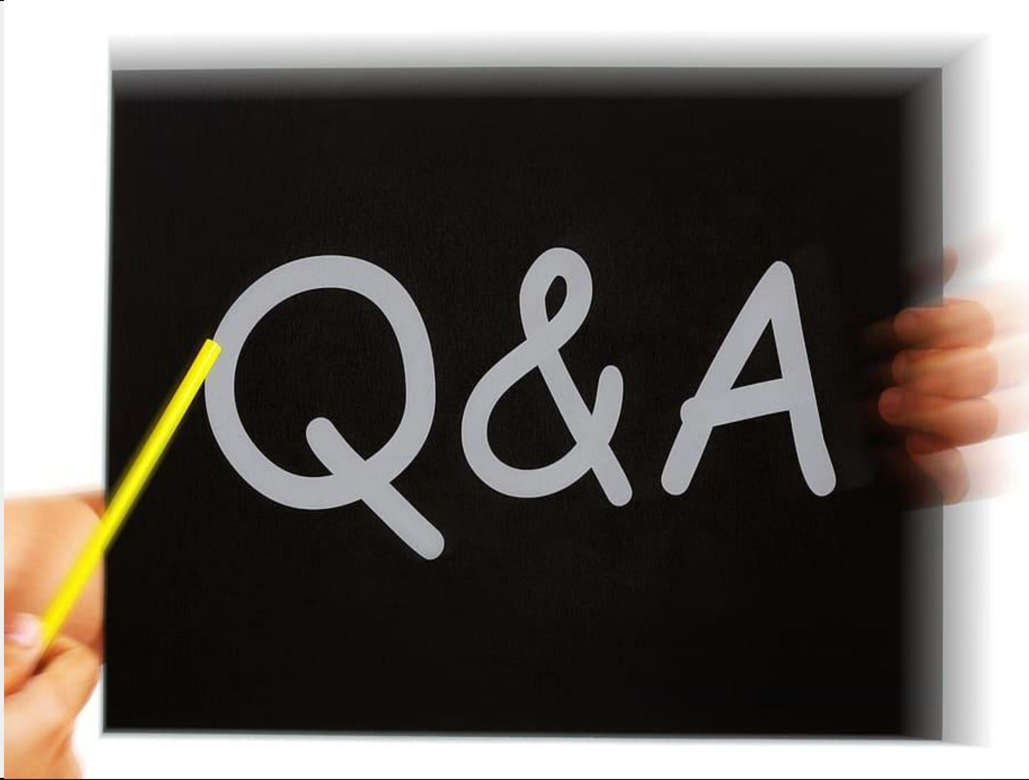
Features	Coefficients
chest_pain	0.815120
vessels_colored	0.781499
sex	0.721611
depress_induced_exer	0.718694
thalassemia	0.563693
exer_ind_ang	0.506319
slope_of_peak	0.402931
max_heart_rte	0.375146
rest_ecg	0.253354
resting_bp	0.246375
cholesterol	0.135303
age	0.065304
fast_blood_sugar	0.038718

Logistic Regression (Lasso)

All These are Contributing Factors

- Chest Pain
- Sex
- Vessels_Colored
- ST Depression induced by Exercise
- Thalassemia
- Exercise Induced Angina
- Slope of Peak Exercise ST Segment
- Max Heart Rate

Logistic Regression is the Better Choice


Q&A