Mentorness Internship Project

On

Heart Attack Prediction

By,
Shobha Kumari Choudhary

TODAY'S AGENDA

- Problem Statement
- Dataset Preview
- Exploratory Data Analysis
- Modelling Approached Used
- Model Comparison
- Best Optimal Model
- Performance Metrices
- Business Insights & Recommendations

Problem Statement

Heart disease is one of the leading causes of death worldwide. Early prediction of heart attack risk can significantly improve patient outcomes by enabling timely medical intervention and lifestyle changes. By analyzing various medical and demographic factors, it is possible to develop a predictive model that can estimate the likelihood of a heart attack.

Objective

Develop a machine learning model that can predict the risk of a heart attack based on a set of medical and demographic variables.

Dataset preview

Overview of Datasets

Our dataset contain Patients information including various features that can help in Predicting the Risk of Heart Attack.

The Features Include

- AGE: AGE OF THE PATIENT
- SEX: SEX OF THE PATIENT (1 = MALE, 0 = FEMALE)
- CP: CHEST PAIN TYPE
- **TRESTBPS**: RESTING BLOOD PRESSURE (IN MM HG)
- **CHOL**: SERUM CHOLESTEROL (IN MG/DL)
- **FBS**: FASTING BLOOD SUGAR (1 = FASTING BLOOD SUGAR > 120 MG/DL, 0 = OTHERWISE)
- **RESTECG**: RESTING ELECTROCARDIOGRAPHIC RESULTS
- **THALACH**: MAXIMUM HEART RATE ACHIEVED
- **EXANG**: EXERCISE-INDUCED ANGINA (1 = YES, 0 = NO)
- **OLDPEAK**: ST DEPRESSION INDUCED BY EXERCISE RELATIVE TO REST
- **SLOPE**: SLOPE OF THE PEAK EXERCISE ST SEGMENT (0 = UPSLOPING, 1 = FLAT, 2 = DOWNSLOPING)
- **CA**: NUMBER OF MAJOR VESSELS (0-3) COLORED BY FLUOROSCOPY
- **THAL**: THALASSEMIA (3 = NORMAL, 6 = FIXED DEFECT, 7= REVERSIBLE DEFECT)
- **TARGET**: TARGET VARIABLE (1 = HEART ATTACK RISK, 0 = NO RISK)

Data Ingestion

Sample of the Dataset

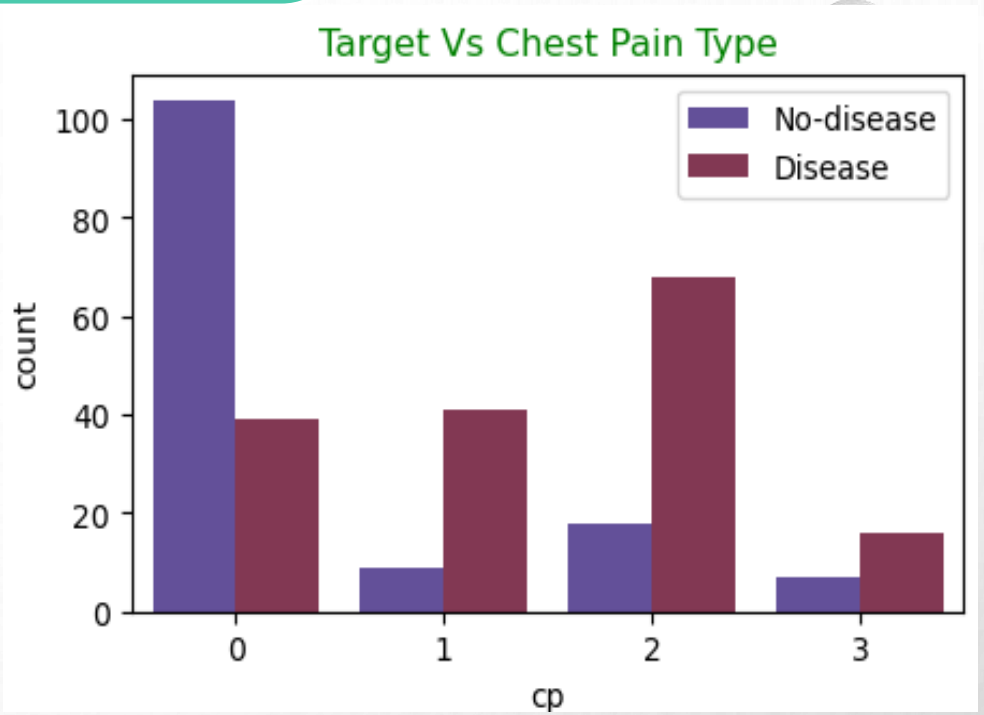
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

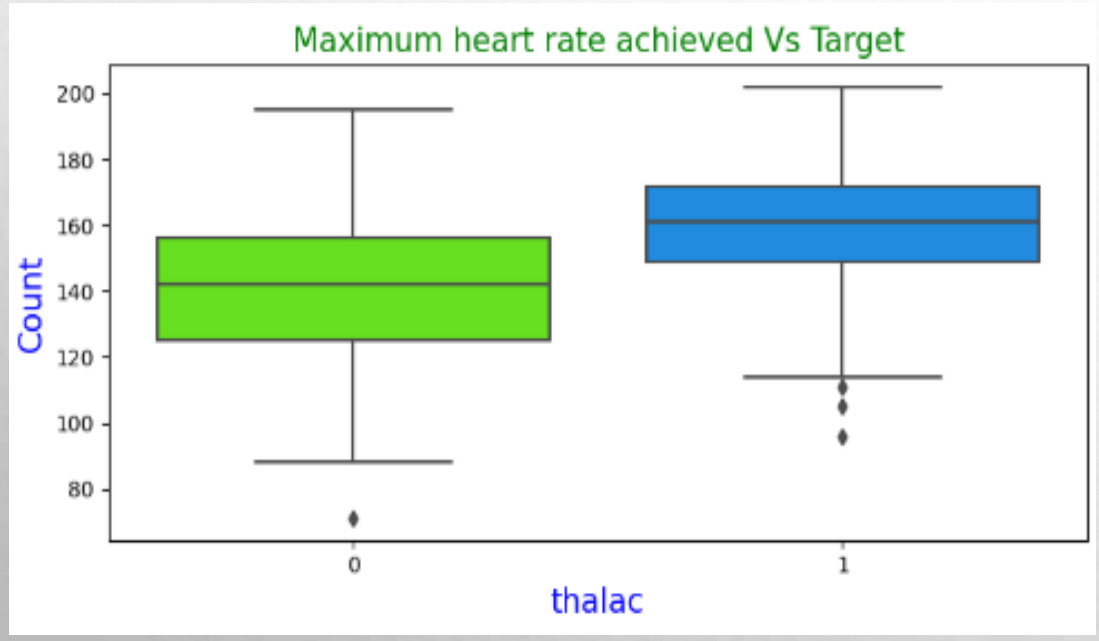
- ✓ Dataset contain 302 Rows & 14 Features.
- ✓ There are no NULL values present in the Dataset.
- ✓ There are no Duplicates.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 302 entries, 0 to 301
Data columns (total 14 columns):
 #    Column     Non-Null Count  Dtype
---   ------     --------------  -----
 0    age        302 non-null    int64
 1    sex        302 non-null    int64
 2    cp         302 non-null    int64
 3    trestbps   302 non-null    int64
 4    chol       302 non-null    int64
 5    fbs        302 non-null    int64
 6    restecg    302 non-null    int64
 7    thalach    302 non-null    int64
 8    exang      302 non-null    int64
 9    oldpeak    302 non-null    float64
 10   slope      302 non-null    int64
 11   ca         302 non-null    int64
 12   thal       302 non-null    int64
 13   target     302 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.2 KB
```

Exploratory data Analysis

- Maximum Patients suffer from type-0 chest pain. This is the most common chest pain type.
- Maximum Heart Disease occur with type-2 chest pain.

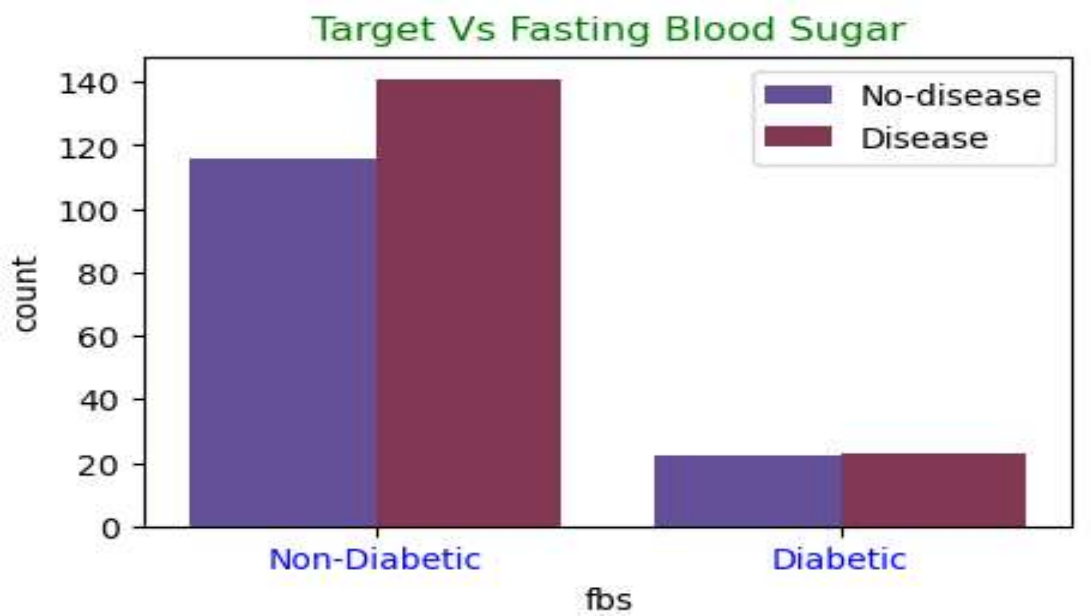


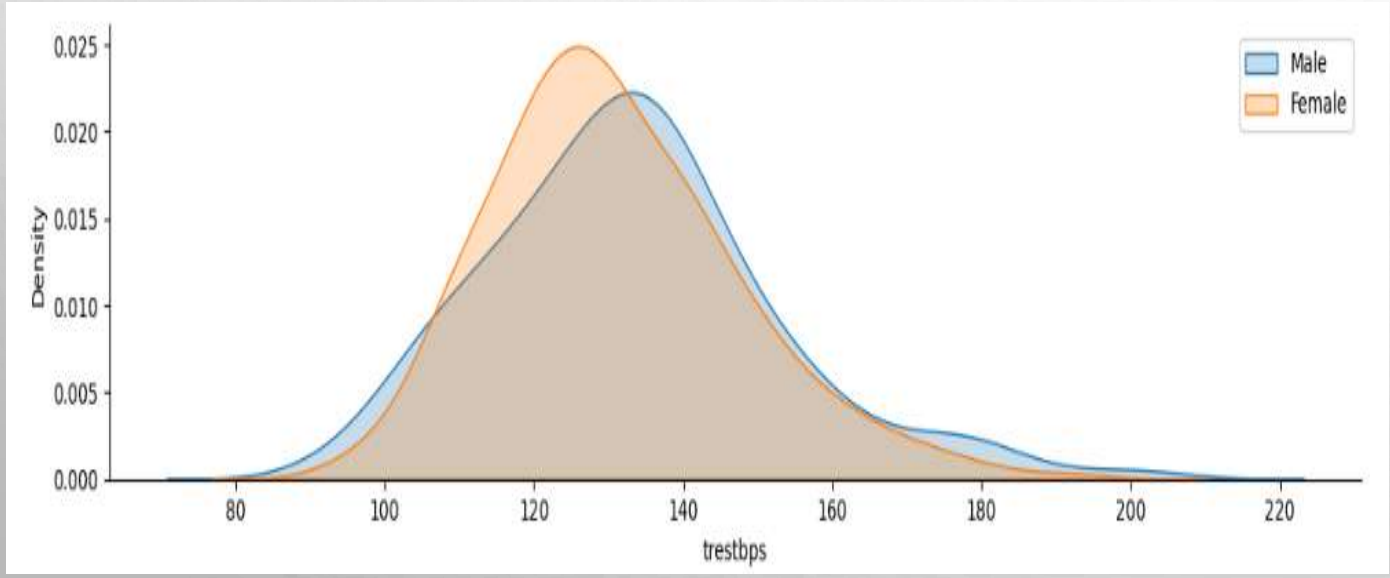


- Patients at a risk of heart disease achieve more rate.

Exploratory data Analysis

- Non-Diabetic patients are more at a Risk of Heart disease.
- It might be due to as Diabetic patients are already on a controlled diet, so their risk of Heart Disease is Low.





- Female has lower resting Bloop Pressure as compared to men.
- For female it is around 120 and for male it is around 140.

Modelling Approach Used

Data Cleaning

- Data is cleaned for any Anomalies present in the dataset.
- Null values, Outlier Treatment, incorrect Datatype mapping were all done using Appropriate methods.

Data Encoding

- All the Categorical columns were Encoded using Dummy Encoding.
- Model building Algorithm can't handle Categorical data, so we need to convert into encoded integers.

Data Transformation

- Features are Transformed (Scaled) using Standard Scalar.

Train – Test Split

- Train – Test split is done using the 70 : 30 ratio.

Models Built

- Built various models and checked Accuracy against the Test data.
- Done Hyper-parameter Tuning to get Better Accuracy and Generalized models.

Model Comparison

	Train_Accuracy	Test_Accuracy
Logistic Regression	0.871369	0.852459
Logistic Regression Tuned	0.871369	0.852459
LDA	0.867220	0.885246
LDA Tuned	0.850622	0.819672
Support Vector Classifier	0.858921	0.852459
SVC Tuned	0.879668	0.868852
KNN_4	0.858921	0.885246
KNN Tuned	0.842324	0.852459
Decission Trees	1.000000	0.803279
Decission Tree Tuned	0.921162	0.786885
Random Forest	1.000000	0.885246
RF Tuned	0.863071	0.885246
Bagging Classifier	0.987552	0.868852
Bagging Classifier Tuned	0.892116	0.819672
Ada Boost Classifier	1.000000	0.786885
Ada Boost Classifier Tuned	0.846473	0.885246

	Train_Accuracy	Test_Accuracy
Gradient Boosting Classifier	0.995851	0.819672
Gradient Boosting Tuned	0.867220	0.901639
XGBoost Classifier	1.000000	0.852459
XGBoost Tuned	0.879668	0.819672
Light GBM	1.000000	0.819672
LightGBM Tuned	0.896266	0.868852

Best Optimal Model

- Best Optimal model is Tuned Gradient Boosting Classifier model.
- The Accuracy in Train data is 86% and in Testing data Accuracy is 90%.

```
GridSearchCV
estimator: GradientBoostingClassifier
GradientBoostingClassifier
GradientBoostingClassifier(random_state=42)
```

Classification Report

Classification Report of the Training data:

	precision	recall	f1-score	support
0	0.88	0.83	0.85	110
1	0.86	0.90	0.88	131
accuracy			0.87	241
macro avg	0.87	0.86	0.87	241
weighted avg	0.87	0.87	0.87	241

Classification Report of the Test data:

	precision	recall	f1-score	support
0	0.96	0.82	0.88	28
1	0.86	0.97	0.91	33
accuracy			0.90	61
macro avg	0.91	0.90	0.90	61
weighted avg	0.91	0.90	0.90	61

Performance Matrices

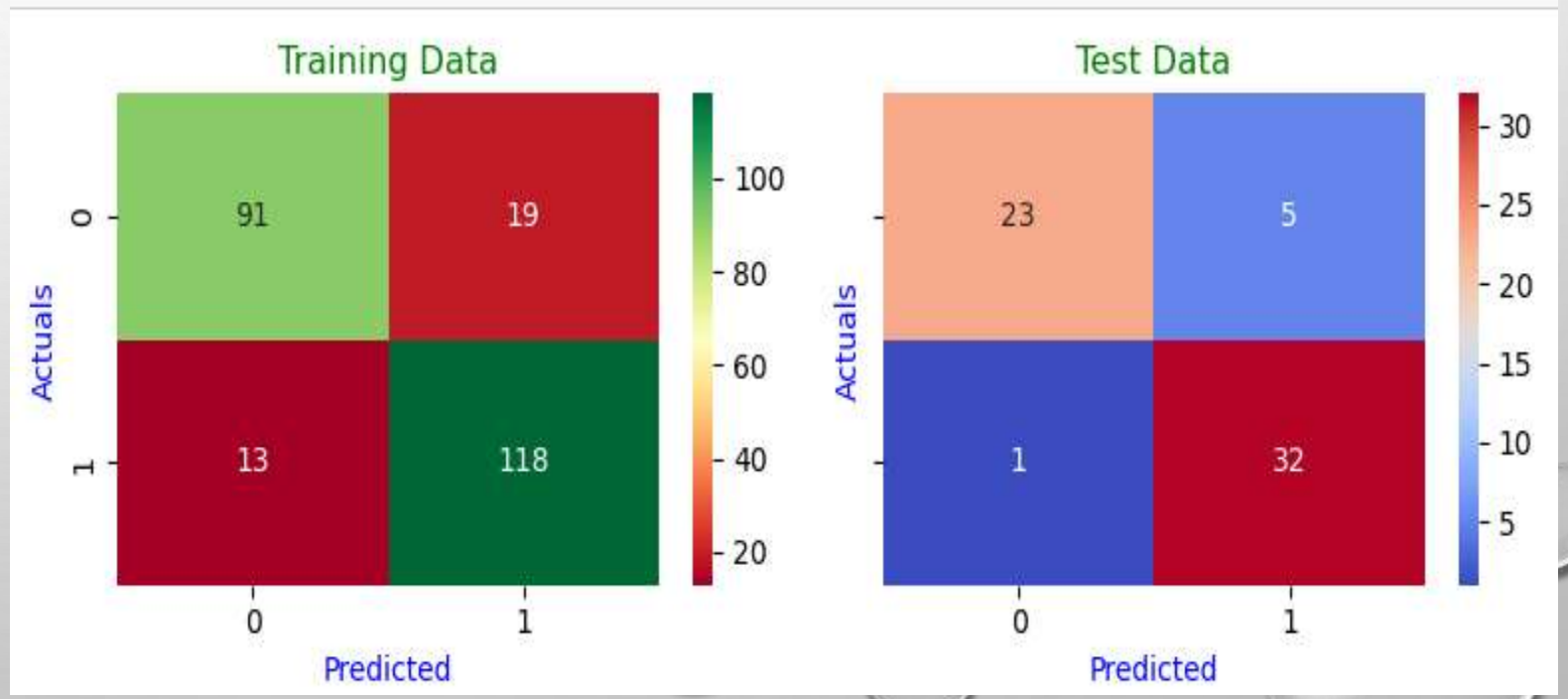
Confusion Matrix

True Positive : 32

True Negative : 23

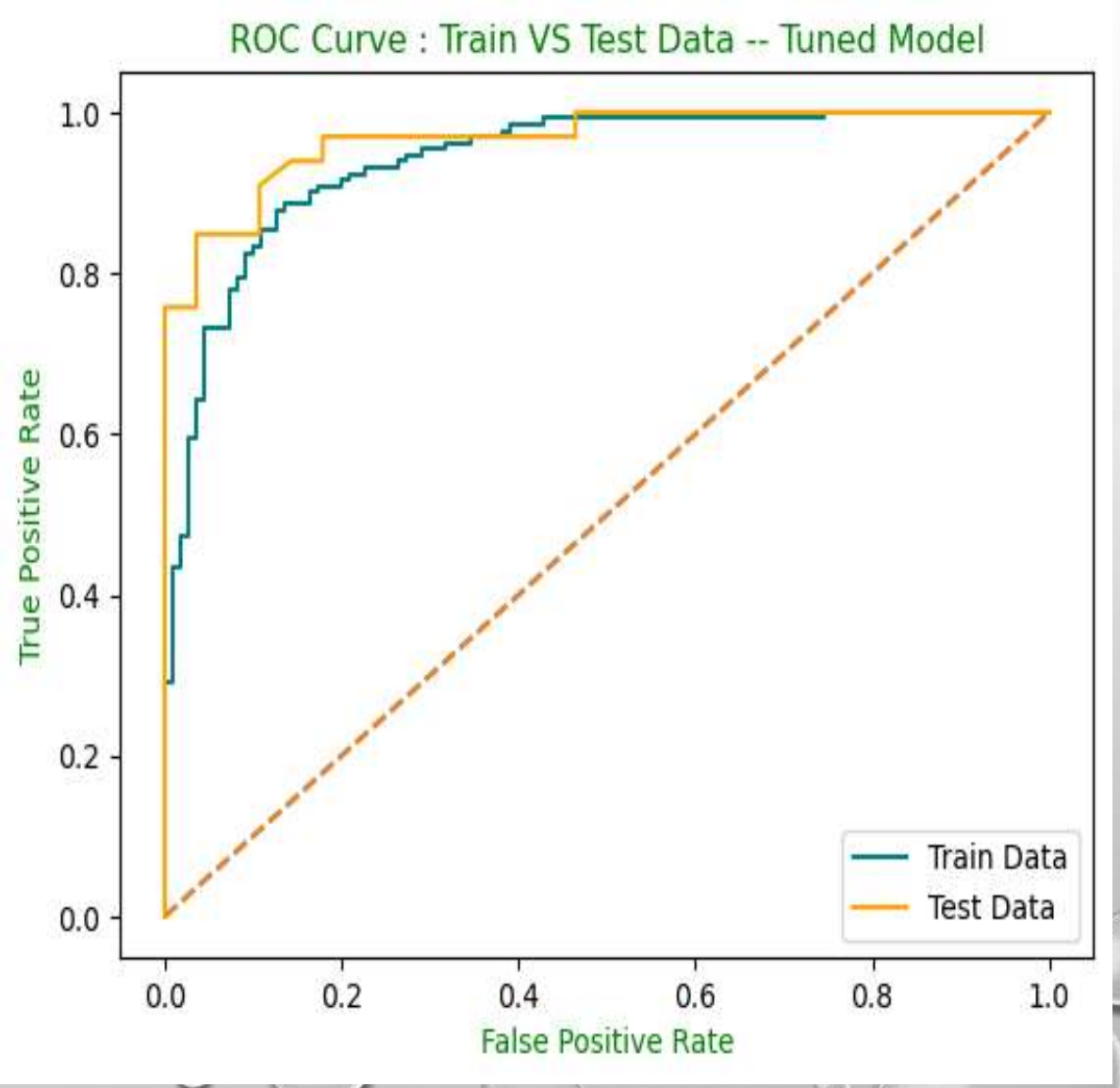
False Positive : 5

False Negative : 1



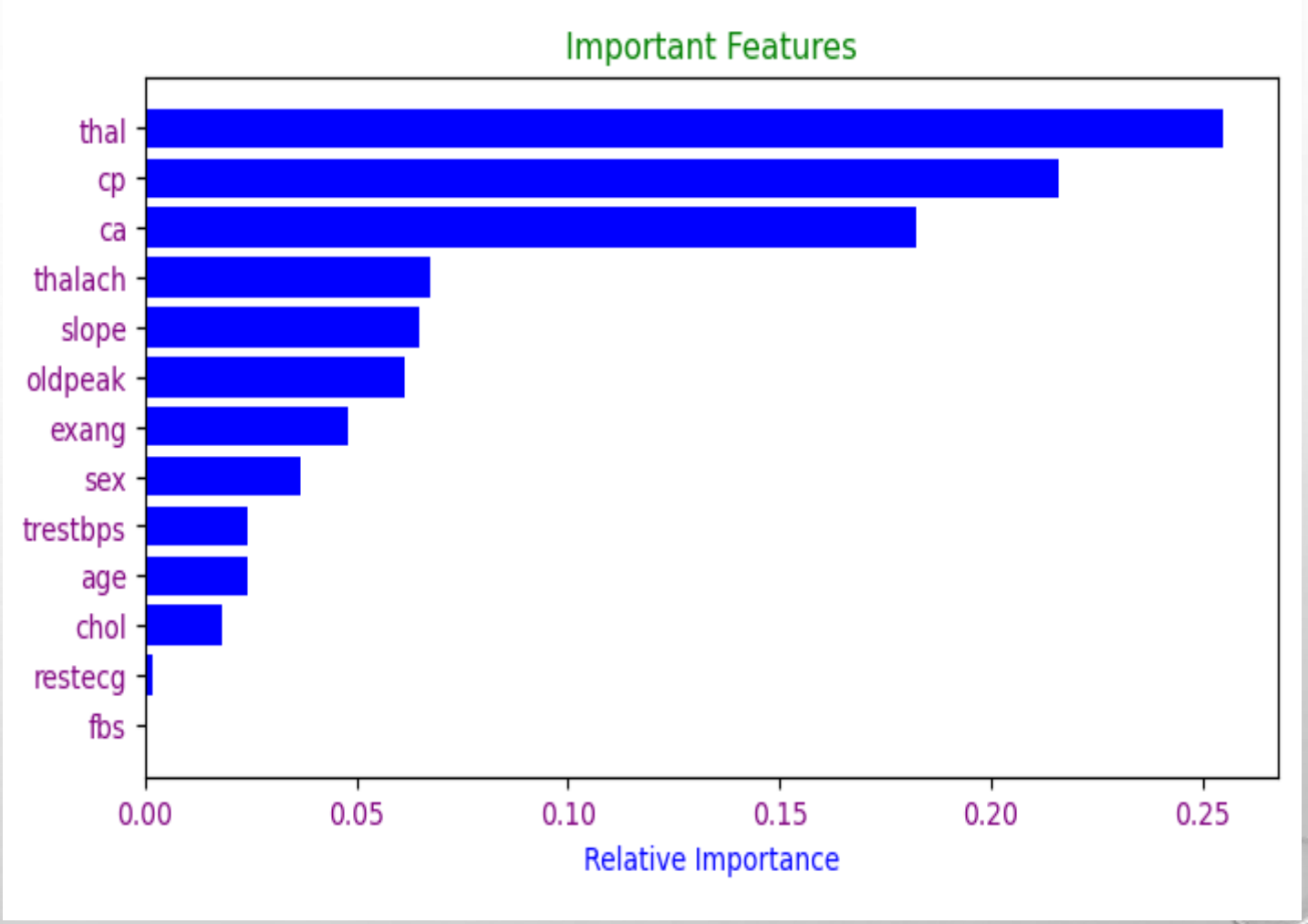
ROC – AUC Curve

- AUC for Train data is 93% and AUC for Test data is 96%.
- The AUC & ROC curve shows that it is covering almost the same area for Train and Test data. Therefore, this is considered a Generalized good model.
- Also, the AUC score is high and can distinguish between Positive and Negative classes vey well.



Feature Importance

The important features for predicting the Risk of heart attack are : thal, cp, ca, thalach & slope.



Business Insights & Recommendations

Insights

- Optimized Tuned GDB Model is performing better with highest AUC score of 96%.
- So, we can say that these Models are able to Separate between the Risk and No-Risk Classes Very well.
- These models can be considered a Good Generalized models.

Recommendations

- The 1st five Important features of Tuned GDB model are: thal, cp, ca, thalach & slope.
- These 5 features are Contributing most towards the Accuracy of the Model.
- Since these 5 features are contributing most towards the Accuracy, the business should ensure that these features should not contain any Anomalies, Null values or any kind of unwanted characters as a part of these columns in the future.
- Also, we should make sure these 5 features are made compulsory for :
  - Predicting Heart attack Risk
  - For model building and

THANK YOU