BUAN 6320.502 | Database Foundations for Business Analytics

Project 2: Azure ML

Date: December 3rd, 2021
Created by: The DB_Group
Created for: Database Foundations for Business Analytics course
No. of Pages: 19

Group Details

Group Members	UTD Email ID
Runjhun Sharma	runjhun.sharma@utdallas.edu
Sarthak Jain	sxj200036@utdallas.edu
Sarthak Khanna	sxk200151@utdallas.edu
Raksha Gujarathi	rxg210027@utdallas.edu
Srishti Patil	sxp210092@utdallas.edu
Sanika Jadhav	saj200004@utdallas.edu
Priyal Gupta	Pxg200016@utdallas.edu
Varun Bhavnani	vnb210000@utdallas.edu

Document Details

Document Name	Document Owner	Date of release	Version
The DB Group_Fundamentals of Database_Project 2	The DB_Group	3rd December 2021	V1.0

Contents

1. Problem Description

1.1. Problem Statement

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict possible heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia, or already established disease) need early detection and management wherein a machine learning model can be of great help.

1.2. Business requirements

Following are the business requirements

1. Perform exploratory data analysis and draw out key insights
2. Build predictive model:
   a. Build a two-class decision tree and evaluate the accuracy of the model
   b. Build two-class logistics regression and evaluate the accuracy of the model
   c. Build two-class Boosted Tree and evaluate the accuracy of the model
3. Deploy the best model and create the webserver
4. User Interface: Develop a user interface that will take the inputs from the end-user and predict heart disease in less than a minute

1.3. Data Description

1. AGE: Age of the individual
2. ANAEMIA: Anaemia is a deficiency in the number or quality of red blood cells in your body.
3. CREATININE_PHOSPHOKINASE: Creatine kinase or creatine phosphokinase is an enzyme chiefly found in the brain, skeletal muscles, and heart. An elevated level of creatine kinase is seen in heart attacks, when the heart muscle is damaged, or in conditions that produce damage to the skeletal muscles or brain.
4. DIABETES: People with diabetes are also more likely to have heart failure. Heart failure is a serious condition, but it doesn’t mean the heart has stopped beating; it means your heart can’t pump blood well. This can lead to swelling in your legs and fluid building up in your lungs, making it hard to breathe. Heart failure tends to get worse over time, but early diagnosis and treatment can help relieve symptoms and stop or delay the condition from getting worse.
5. EJECTION_FRACTION: A normal ejection fraction is more than 55%. This means that 55% of the total blood in the left ventricle is pumped out with each heartbeat. Heart failure with reduced ejection fraction happens when the muscle of the left ventricle is not pumping as well as normal. The ejection fraction is 40% or less.

6. HIGH_BLOOD_PRESSURE: If you have heart failure, there's a good chance you also have high blood pressure, or "hypertension." About two-thirds of people whose hearts can't pump enough blood because of the condition also have high BP or once did. Hypertension is a major risk factor for heart failure.
7. Platelets: Heart failure patients have increased whole blood aggregation,7 platelet-derived adhesion molecules. CHF patients also have higher mean platelet volume9 and soluble (and platelet-bound) P-selectin, regardless of the etiology.
8. SERUM_CREATININE: Patients with severe heart failure, particularly those on large doses of diuretics for long periods, may have elevated BUN and creatinine levels indicative of renal insufficiency owing to chronic reductions of renal blood flow from the reduced cardiac output.
9. SERUM_SODIUM: Hyponatremia or low serum sodium level is typically defined as a serum sodium concentration of <135 mEq/L and is one of the most common biochemical disorders featured in heart failure patients, with a prevalence close to 25% [2–4]. HF affects cardiac output by either decreasing heart rate or reducing the stroke volume
10. SEX: Gender of the individual
11. SMOKING: Smoking habits of the individual

1.4. Feature Selection

Feature selection methods are intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable. Some predictive modeling problems have a large number of variables that can slow the development and training of models and require a large amount of system memory. Additionally, the performance of some models can degrade when including input variables that are not relevant to the target variable.

One way to think about feature selection methods is in terms of supervised and unsupervised methods. The difference has to do with whether features are selected based on the target variable or not. Unsupervised feature selection techniques ignore the target variable, such as methods that remove redundant variables using correlation. Supervised feature selection techniques use the target variable, such as methods that remove irrelevant variables.

Another way to consider the mechanism used to select features that may be divided into wrapper and filter methods. These methods are almost always supervised and are evaluated based on the performance of a resulting model on a holdout dataset.

Wrapper feature selection methods create many models with different subsets of input features and select those features that result in the best performing model according to a performance metric. These methods are unconcerned with the variable types, although they can be computationally expensive. RFE is a good example of a wrapper feature selection method.

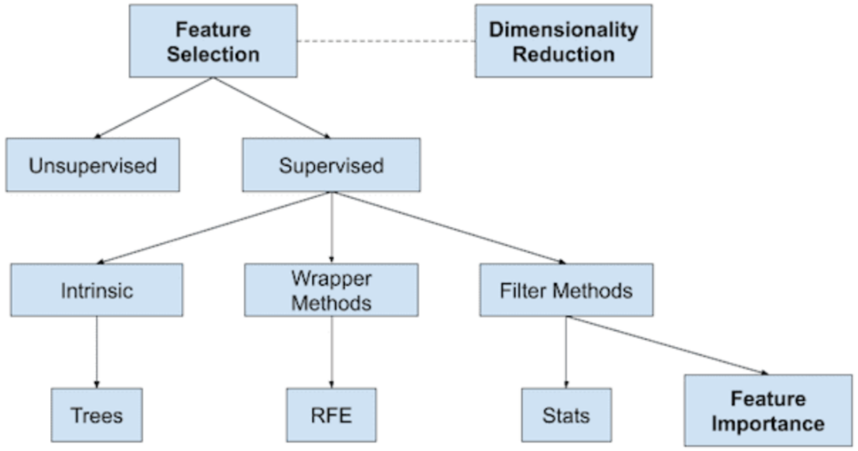
Finally, there are some machine learning algorithms that perform feature selection automatically as part of learning the model. We might refer to these techniques as intrinsic feature selection methods. This includes algorithms such as penalized regression models like Lasso and decision trees, including ensembles of decision trees like a random forest.

We can summarize feature selection as follows.

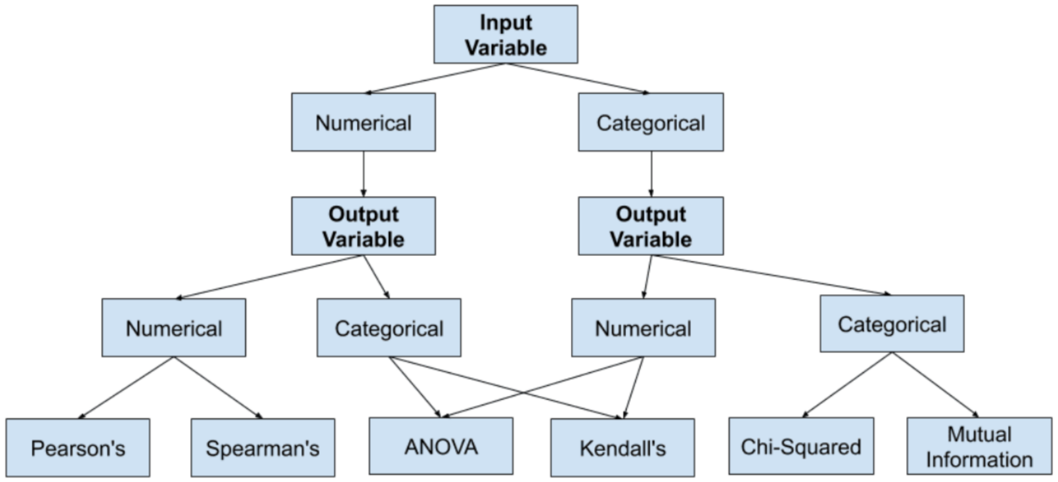
Feature Selection: Select a subset of input features from the dataset.

- Unsupervised: Do not use the target variable (e.g. remove redundant variables).
  - Correlation
- Supervised: Use the target variable (e.g. remove irrelevant variables).
  - Wrapper: Search for well-performing subsets of features.
    - RFE
  - Filter: Select subsets of features based on their relationship with the target.
    - Statistical Methods
    - Feature Importance Methods
  - Intrinsic: Algorithms that perform automatic feature selection during training.
    - Decision Trees
- Dimensionality Reduction: Project input data into a lower-dimensional feature space.

Overview of Feature Selection Techniques



How to Choose a Feature Selection Method

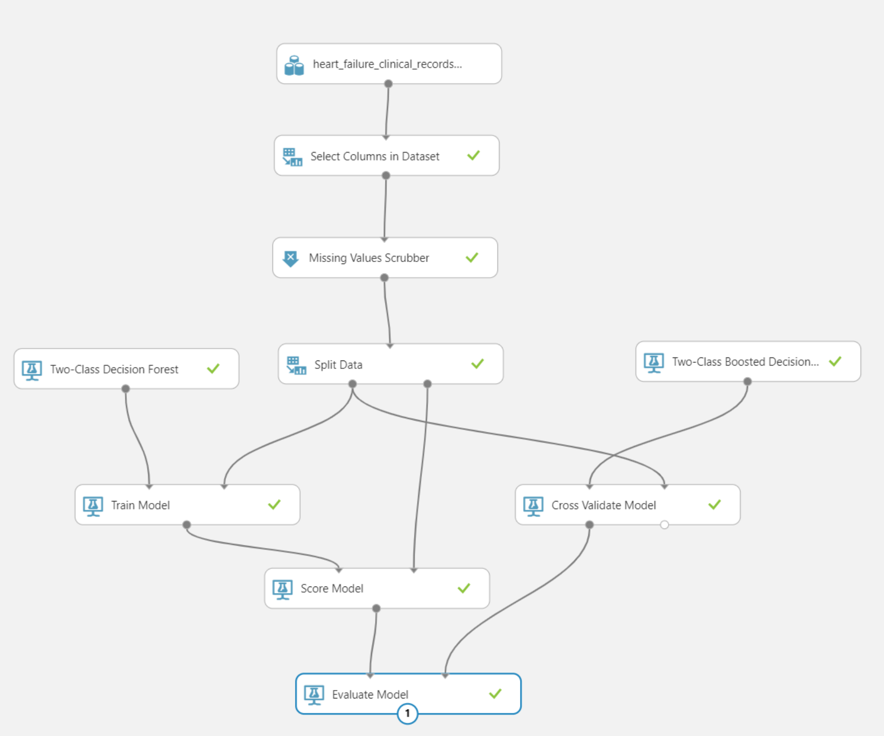


Feature selection for selected models

1. Classification Trees: As we will be building and assessing Decision Tree and Boosted Tree, we don't need to consider additional steps of feature selection as its intrinsic
2. Logistics Regression: As we will be dealing with categorical variables like DEATH_EVENT, we will try Chi-Squared and Mutual Information see assess the model

2. Azure ML Prediction Process Flow Diagram

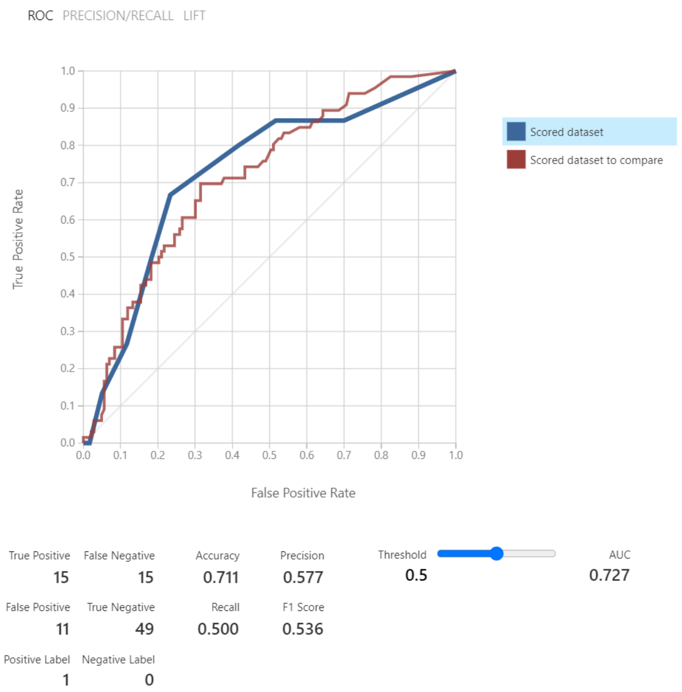
2.1. Two-class Decision Forest vs Two-class Boosted Tree


heart_failure_clinical_records...
Select Columns in Dataset
Missing Values Scrubber
Two-Class Decision Forest
Split Data
Two-Class Boosted Decision...
Train Model
Cross Validate Model
Score Model
Evaluate Model
1

Following are the steps we took in order to build the model

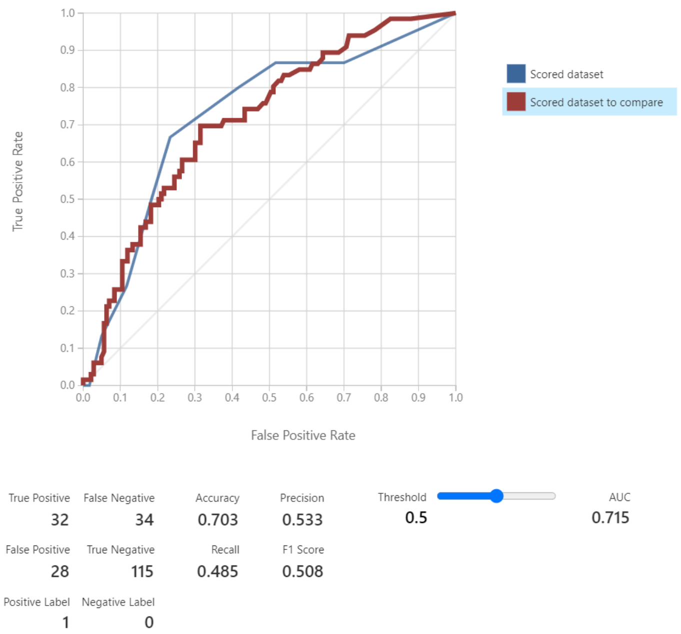
1. Selected the required columns
2. Cleaned the data set by using a missing value scrubber
3. Split the data into 70% training data and 30% validation data
4. Trained the model using two-class decision forest & two-class boosted tree
5. An extra step of cross-validation was applied for two-class boosted tree
6. Scored and evaluated the model

Two-Class Decision Forest



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
15	15	0.711	0.577	0.5	0.727
False Positive	**True Negative**	**Recall**	**F1 Score**		
11	49	0.500	0.536		
Positive Label	**Negative Label**				
1	0				

Two-Class Boosted Decision Tree



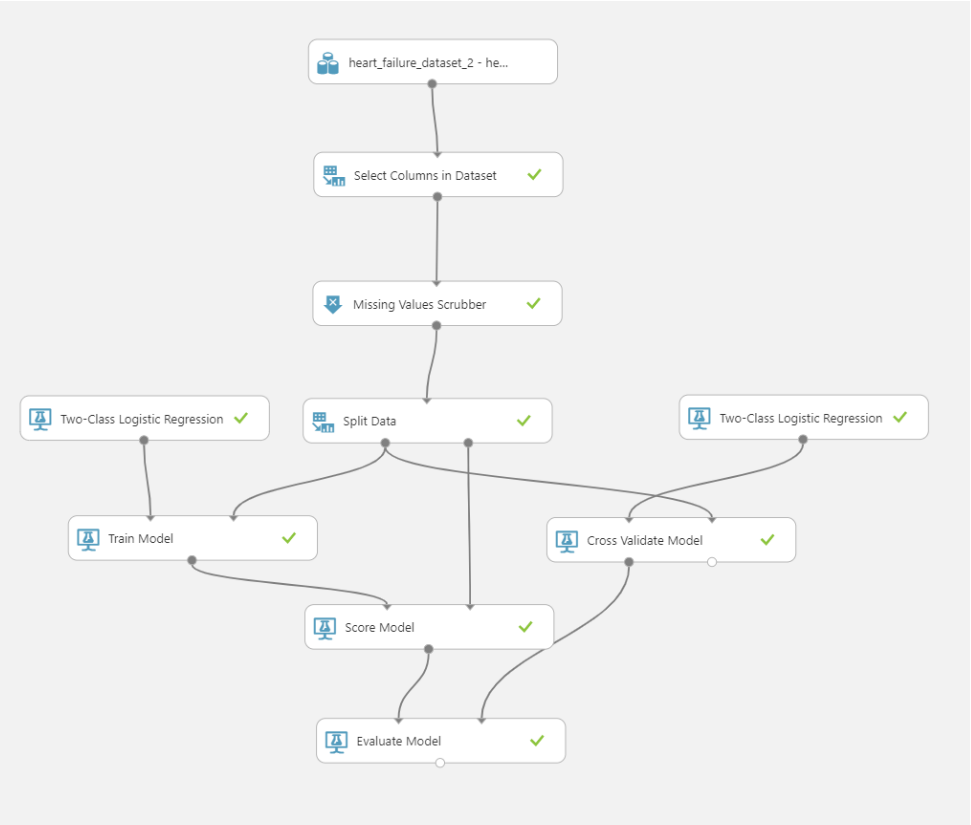
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
32	34	0.703	0.533	0.5	0.715
False Positive	**True Negative**	**Recall**	**F1 Score**		
28	115	0.485	0.508		
Positive Label	**Negative Label**				
1	0				

From the above graph, we can interpret that the area under the curve for

- Decision Forest: 0.727
- Boosted tree: 0.715

The higher the value of AUC, the better the model. From the values above, the **decision forest** is a better model

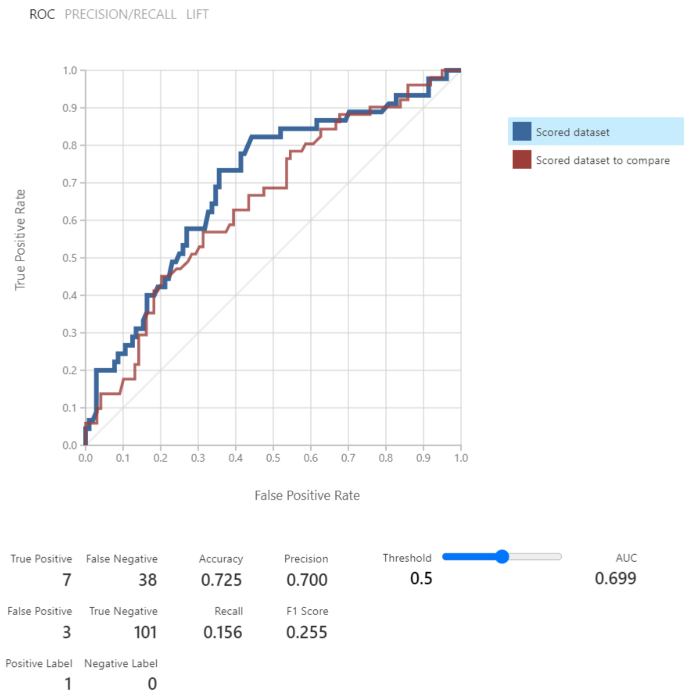
2.2. Two-class Logistics Regression


heart_failure_dataset_2 - he...
Select Columns in Dataset
Missing Values Scrubber
Two-Class Logistic Regression
Split Data
Two-Class Logistic Regression
Train Model
Cross Validate Model
Score Model
Evaluate Model

Following are the steps we took in order to build the model

1. Selected the required columns
2. Cleaned the data set by using a missing value scrubber
3. Split the data into 70% training data and 30% validation data
4. Trained the model using two-class logistics regression
5. An extra step of cross-validation was applied
6. Scored and evaluated the model

Two-Class logistics Regression w/o cross-validation



True Positive 7

False Negative 38

Accuracy 0.725

Precision 0.700

Threshold 0.5

AUC 0.699

False Positive 3

True Negative 101

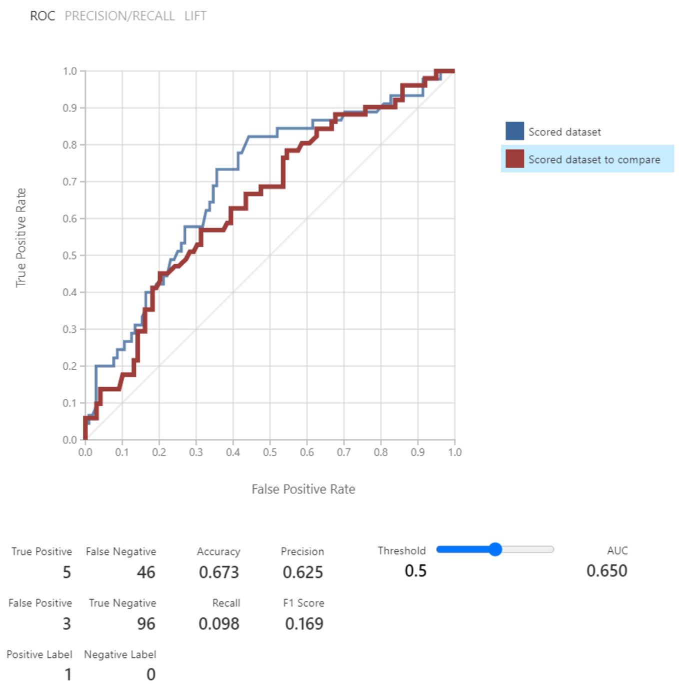
Recall 0.156

F1 Score 0.255

Positive Label 1

Negative Label 0

Two-Class Logistics Regression with cross-validation



True Positive 5

False Negative 46

Accuracy 0.673

Precision 0.625

Threshold 0.5

AUC 0.650

False Positive 3

True Negative 96

Recall 0.098

F1 Score 0.169

Positive Label 1

Negative Label 0

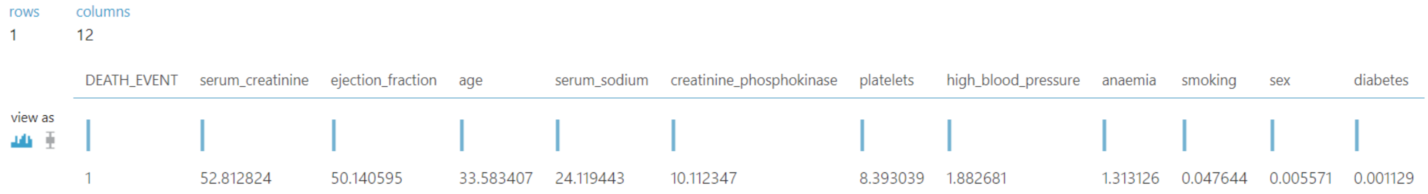
From the above graph, we can interpret that the area under the curve for

- Logistics Regression without cross-validation : 0.699
- Logistics Regression with cross-validation: 0.650

The higher the value of AUC, the better the model. From the values above, the **Logistics Regression without cross-validation** is a better model

2.3. Two-class Logistics Regression with Feature Selection

Feature Selection: Chi-squared

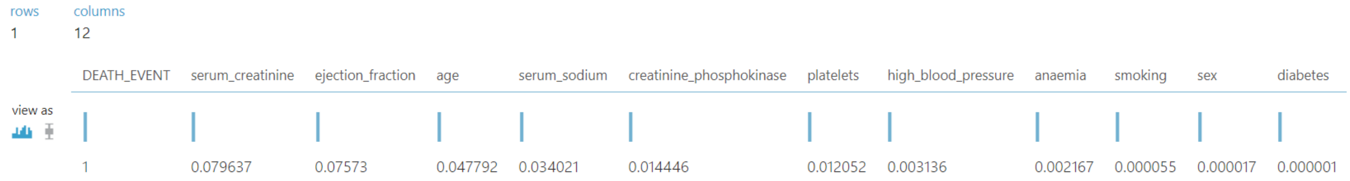


rows	columns
1	12

	DEATH_EVENT	serum_creatinine	ejection_fraction	age	serum_sodium	creatinine_phosphokinase	platelets	high_blood_pressure	anaemia	smoking	sex	diabetes
view as	1	52.812824	50.140595	33.583407	24.119443	10.112347	8.393039	1.882681	1.313126	0.047644	0.005571	0.001129

From the above Chi-squared result, we can interpret that columns serum_creatinine, ejection_fraction, age & serum_sodium have a statistically significant impact on the dependent variable: DEATH_EVENT. The reason behind this is that as per the values shown above, there is a sudden drop in the chi-squared values after serum_sodium

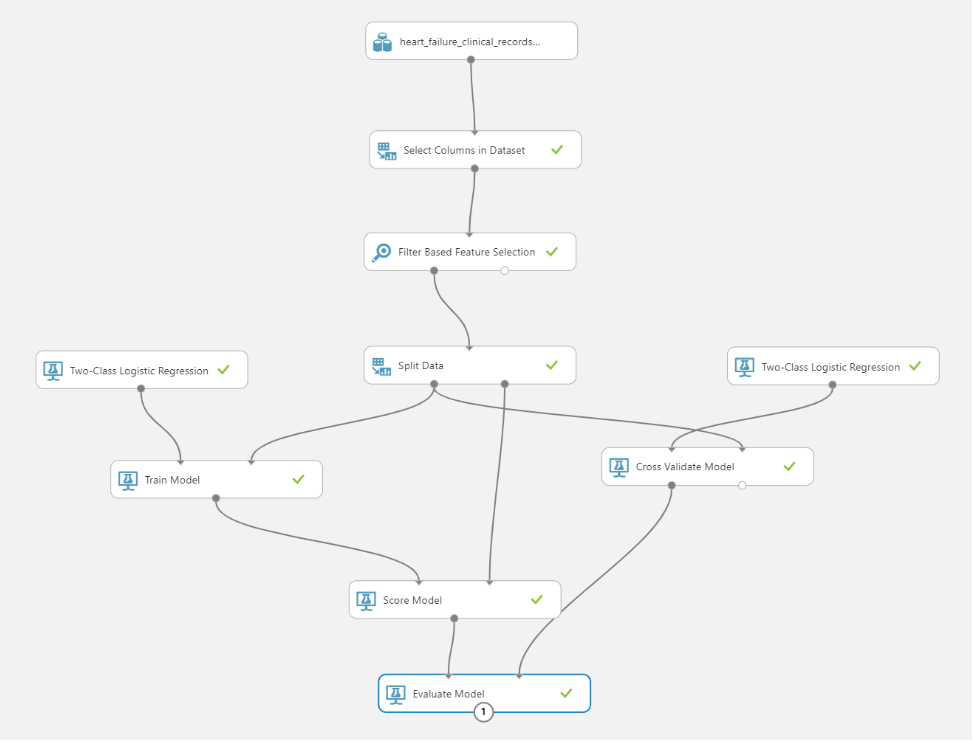
Feature Selection: Mutual Information



rows	columns
1	12

	DEATH_EVENT	serum_creatinine	ejection_fraction	age	serum_sodium	creatinine_phosphokinase	platelets	high_blood_pressure	anaemia	smoking	sex	diabetes
view as	1	0.079637	0.07573	0.047792	0.034021	0.014446	0.012052	0.003136	0.002167	0.000055	0.000017	0.000001

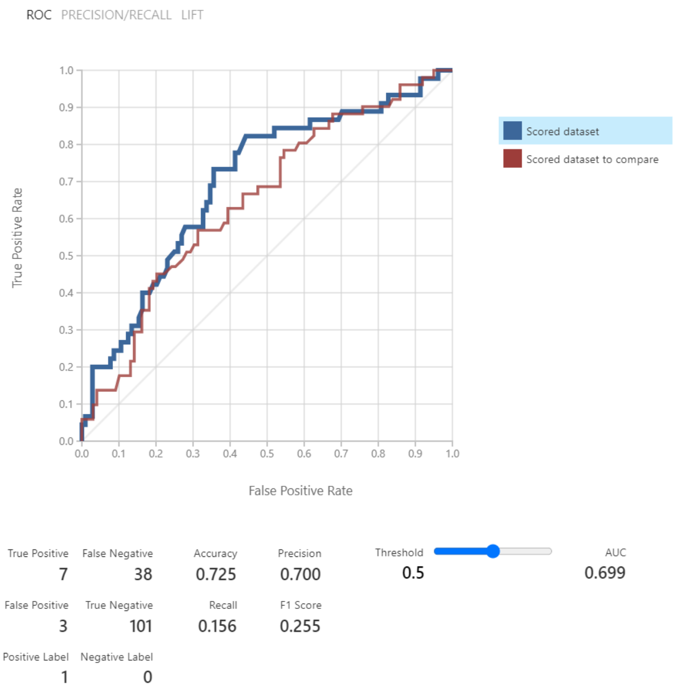
Mutual Information results in all low values and it's not clear to interpret. So we go ahead with our interpretation from chi-squared feature selection


heart_failure_clinical_records...
Select Columns in Dataset
Filter Based Feature Selection
Two-Class Logistic Regression
Split Data
Two-Class Logistic Regression
Train Model
Cross Validate Model
Score Model
Evaluate Model
1

Following are the steps we took in order to build the model

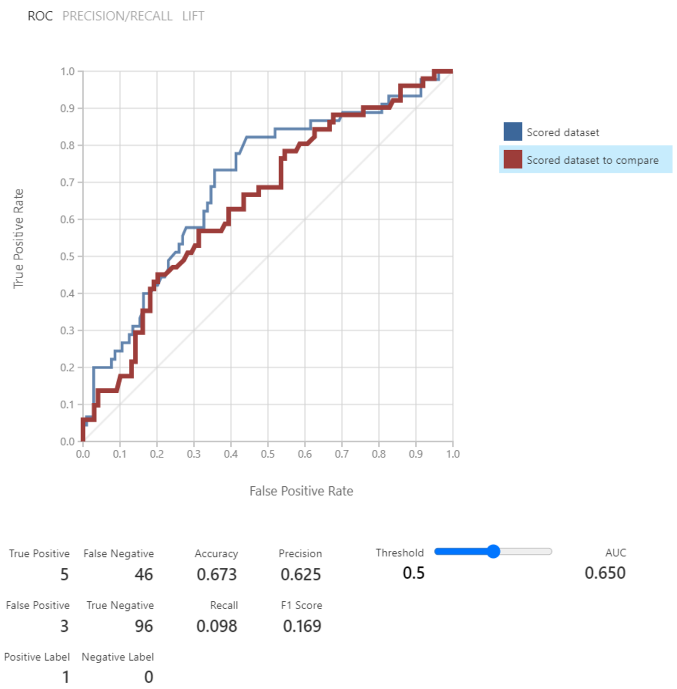
1. Selected the required columns
2. Cleaned the data set by using a missing value scrubber
3. Split the data into 70% training data and 30% validation data
4. Trained the model using two-class logistics regression
5. An extra step of cross-validation was applied
6. Scored and evaluated the model

Two-Class Logistics Regression w/o Cross-Validation



True Positive 7	False Negative 38	Accuracy 0.725	Precision 0.700	Threshold 0.5	AUC 0.699
False Positive 3	True Negative 101	Recall 0.156	F1 Score 0.255		
Positive Label 1	Negative Label 0				

Two-Class Logistics Regression with Cross-Validation



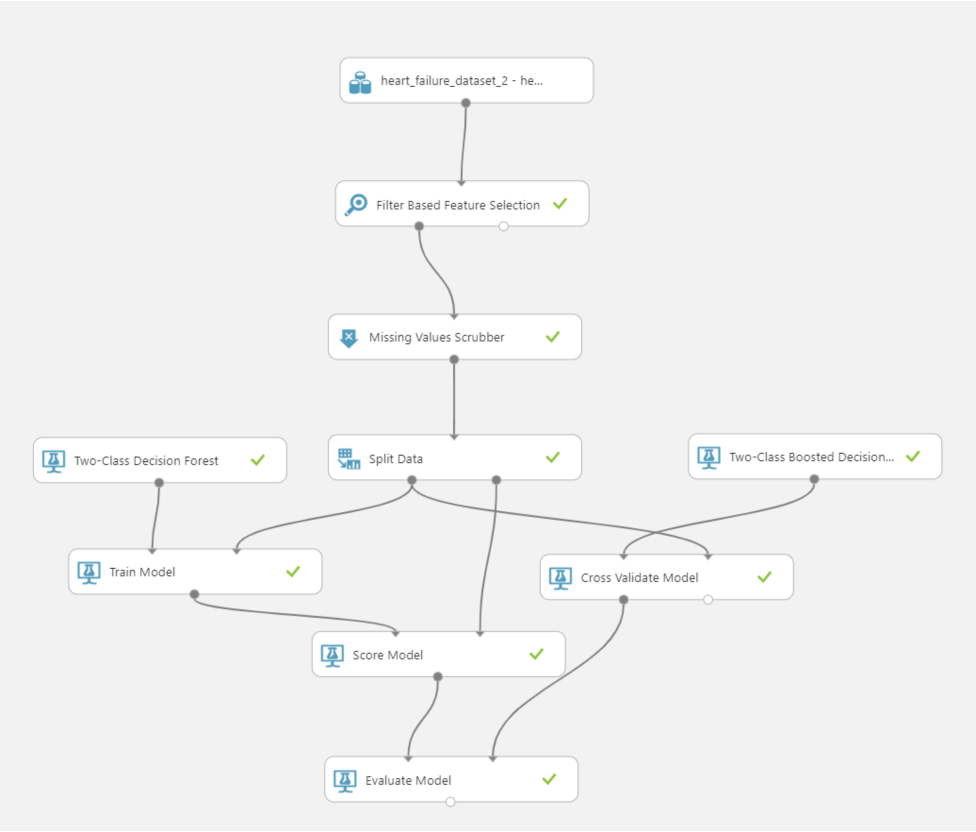
True Positive 5	False Negative 46	Accuracy 0.673	Precision 0.625	Threshold 0.5	AUC 0.650
False Positive 3	True Negative 96	Recall 0.098	F1 Score 0.169		
Positive Label 1	Negative Label 0				

From the above graph, we can interpret that the area under the curve for

- Without Cross-Validation: 0.699
- With Cross-Validation: 0.650

The higher the value of AUC, the better the model. From the values above, the **Logistics Regression without cross-validation** is a better model

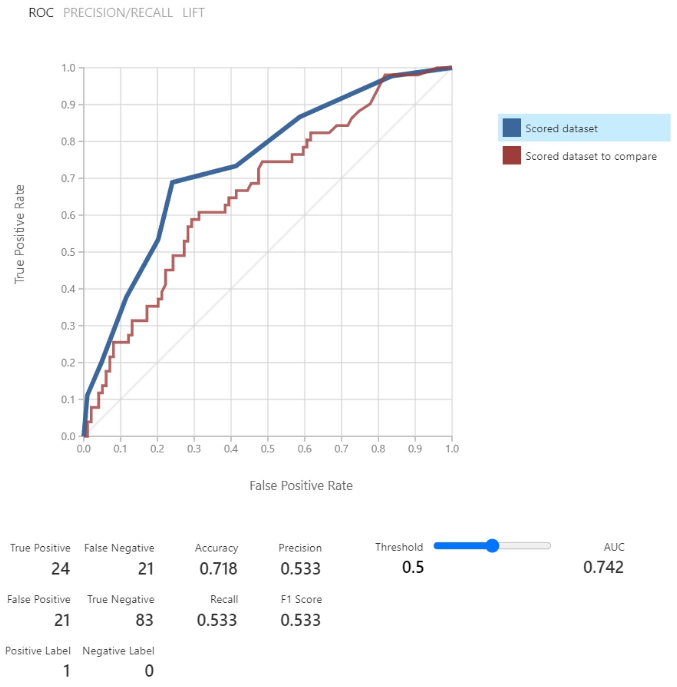
2.4. Two-class Decision Forest vs Two-class Boosted Tree with Feature Selection


heart_failure_dataset_2 - he...
Filter Based Feature Selection ✓
Missing Values Scrubber ✓
Two-Class Decision Forest ✓
Split Data ✓
Two-Class Boosted Decision... ✓
Train Model ✓
Cross Validate Model ✓
Score Model ✓
Evaluate Model ✓

Following are the steps we took in order to build the model

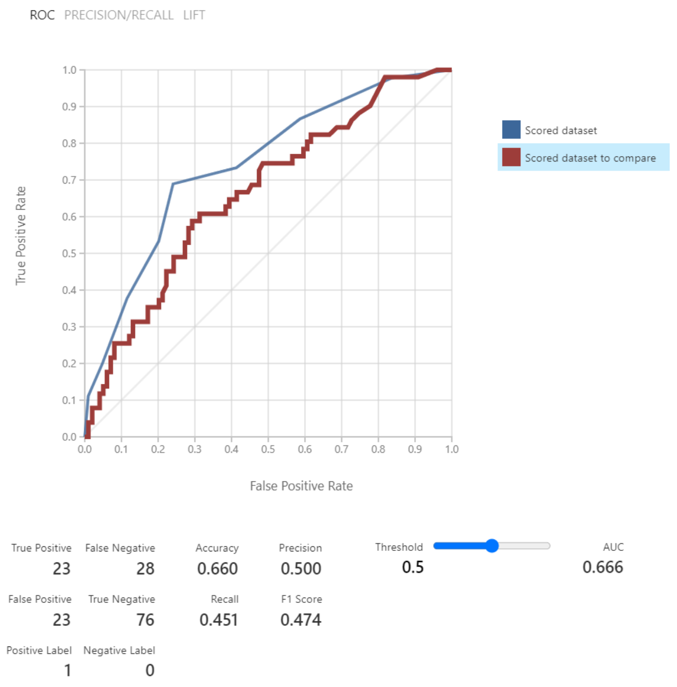
1. Selected the required columns via Feature selection
2. Cleaned the data set by using a missing value scrubber
3. Split the data into 70% training data and 30% validation data
4. Trained the model using two-class decision forest & two-class boosted tree
5. An extra step of cross-validation was applied for two-class boosted tree
6. Scored and evaluated the model

Two-Class Decision Forest



True Positive 24	False Negative 21	Accuracy 0.718	Precision 0.533	Threshold 0.5	AUC 0.742
False Positive 21	True Negative 83	Recall 0.533	F1 Score 0.533		
Positive Label 1	Negative Label 0				

Two-Class Boosted Decision Tree



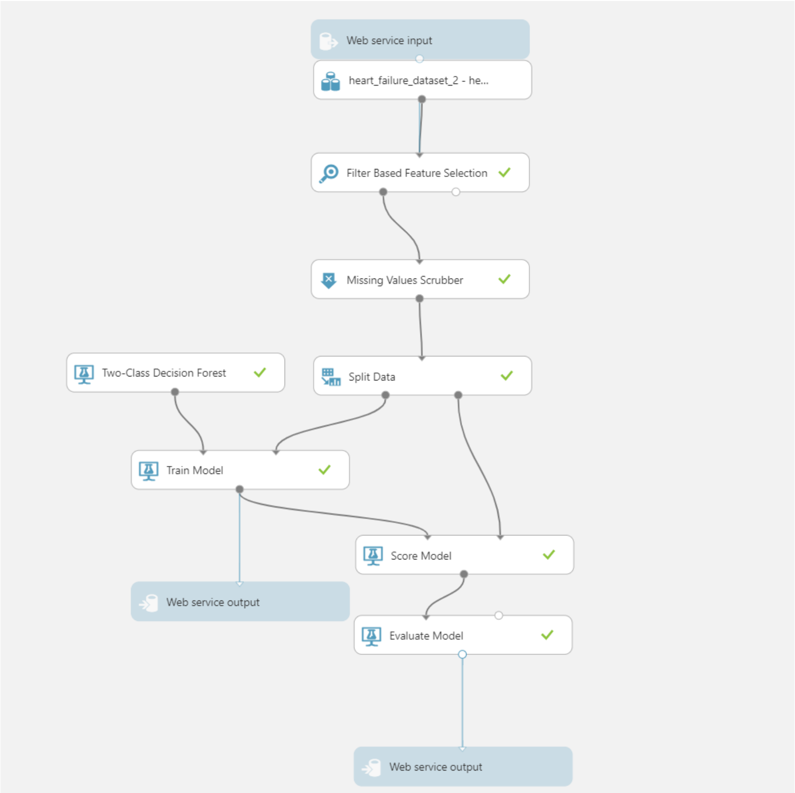
True Positive 23	False Negative 28	Accuracy 0.660	Precision 0.500	Threshold 0.5	AUC 0.666
False Positive 23	True Negative 76	Recall 0.451	F1 Score 0.474		
Positive Label 1	Negative Label 0				

From the above graph, we can interpret that the area under the curve for

- Decision Forest: 0.742
- Boosted tree: 0.666

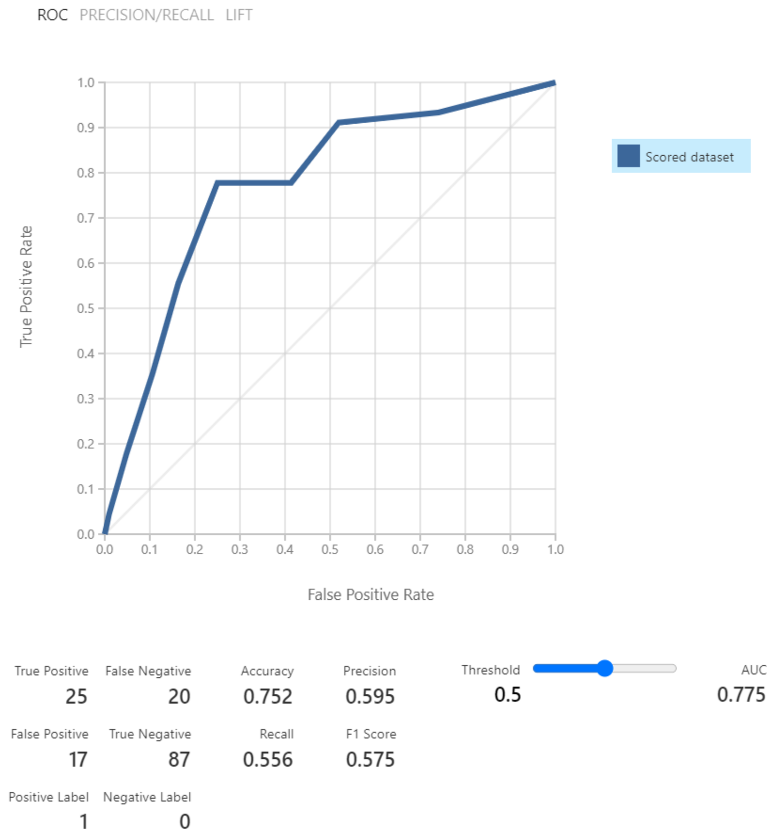
The higher the value of AUC, the better the model. From the values above, the **Decision Forest** is a better model

2.5. Final Model: Two-class Decision Forest with Feature Selection


Web service input
heart_failure_dataset_2 - he...
Filter Based Feature Selection ✓
Missing Values Scrubber ✓
Two-Class Decision Forest ✓
Split Data ✓
Train Model ✓
Score Model ✓
Web service output
Evaluate Model ✓
Web service output

Following are the steps we took in order to build the model

1. Selected the required columns via Feature selection
2. Cleaned the data set by using a missing value scrubber
3. Split the data into 70% training data and 30% validation data
4. Trained the model using two-class decision forest
5. Scored and evaluated the model



True Positive 25

False Negative 20

Accuracy 0.752

Precision 0.595

Threshold 0.5

AUC 0.775

False Positive 17

True Negative 87

Recall 0.556

F1 Score 0.575

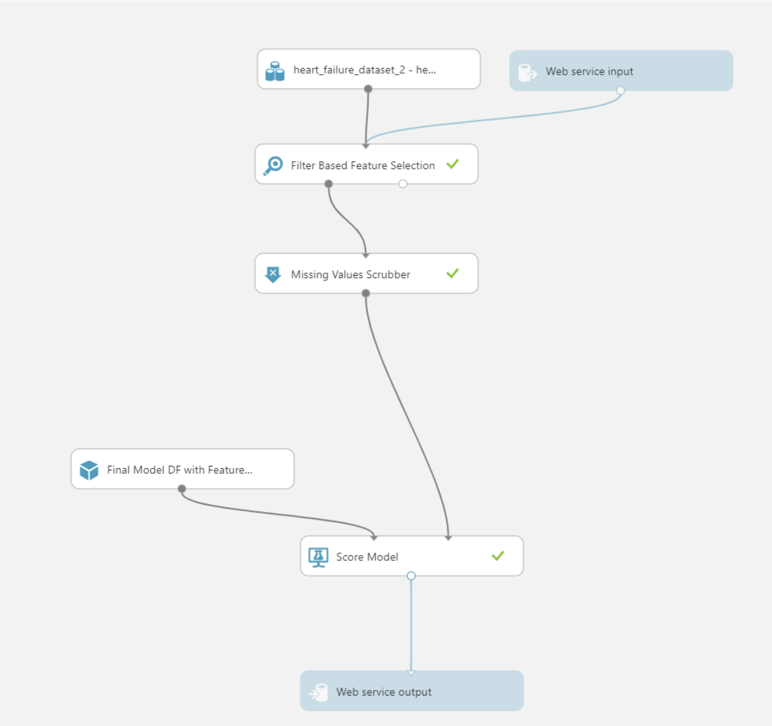
Positive Label 1

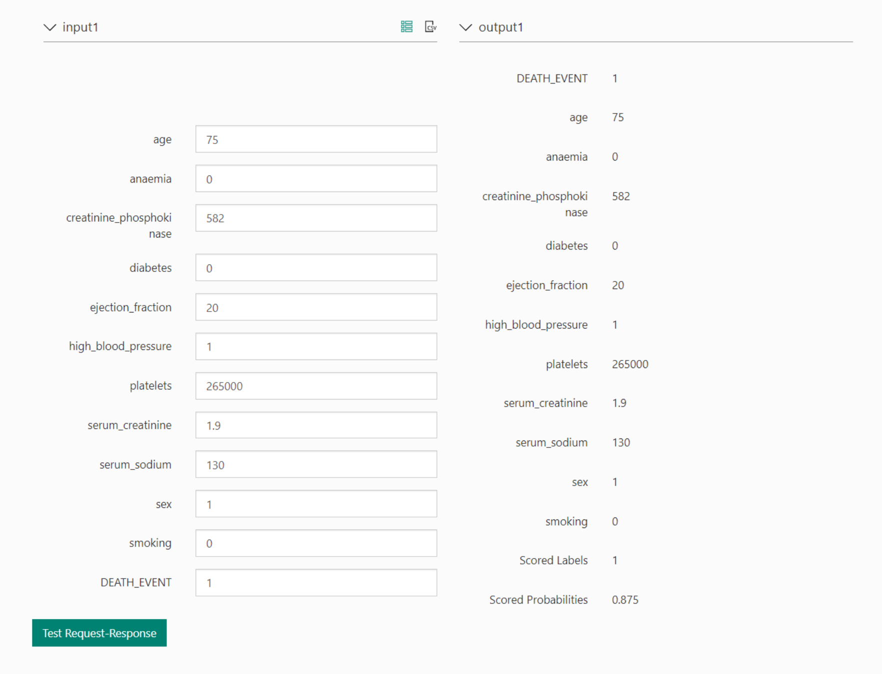
Negative Label 0

As we can see, the AUC has improved to **0.775**. We choose this model for our final deployment

3. Model Deployment using Azure Web Server

- Built-in feature in Azure ML allowed us to deploy a web server that creates a user interface with the feature of providing the input and the result
- After building the model, we deployed the model as a web service that will take the variables as input and provide the heart disease prediction as to the output. Please refer to the screenshots below


heart_failure_dataset_2 - he...
Web service input
Filter Based Feature Selection
Missing Values Scrubber
Final Model DF with Feature...
Score Model
Web service output



input1

age	75
anaemia	0
creatinine_phosphokinase	582
diabetes	0
ejection_fraction	20
high_blood_pressure	1
platelets	265000
serum_creatinine	1.9
serum_sodium	130
sex	1
smoking	0
DEATH_EVENT	1

output1

DEATH_EVENT	1
age	75
anaemia	0
creatinine_phosphokinase	582
diabetes	0
ejection_fraction	20
high_blood_pressure	1
platelets	265000
serum_creatinine	1.9
serum_sodium	130
sex	1
smoking	0
Scored Labels	1
Scored Probabilities	0.875

4. Limitations and Assumptions

4.1. Limitations

1. One of the limitations of decision trees is that they are largely unstable compared to other decision predictors. A small change in the data can result in a major change in the structure of the decision tree, which can convey a different result from what users will get in a normal event
2. decision trees are less effective in making predictions when the main goal is to predict the outcome of a continuous variable. This is because decision trees tend to lose information when categorizing variables into multiple categories
3. The smoking variable does not indicate whether or not a certain number of cigarettes per day is dangerous.
4. Blood pressure variables do not tell at what blood pressure level an individual is at risk of heart disease.
5. Diabetes level variables does not tell at what level an individual is at risk of heart disease

4.2. Assumptions

1. 300 sample infers to the population
2. Assumes that there is minimal or no multicollinearity among the independent variables
3. Sufficient variables and no further independent variables are required

5. Business Impact

1. A Diagnosis Tool in clinics:
    a. Integrate into the clinical work
    b. Help the cardiologist to detect cardiac arrest
    c. Improve the consistency of diagnosis
    d. decrease human error
    e. People at higher risk of Cardiovascular disease (prediction of Death with 1) can be further tested and given the care or lifestyle changes they require according to the most important factors that impact the same. This brings in revenue for the healthcare agent and can prevent cardiac disease or detect any risky activity.
2. Aid in various cancer detection R&D
    a. This model can be used as a reference to carry forward the R&D activities
    b. Various Medicines can be developed targeting the medical history which has a significant effect on the Tumor cell
3. Commercialize the Model for economic growth
    a. Compare the economic costs of using the predictive model and the current screening processes.
    b. Commercialize this model to local clinics

6. Members contribution

Topics	Members
Data Selection	Sanika Jadhav
Business Context and Problem definition	Sarthak Khanna, Runjhun Sharma
Model Development: Decision Forest vs Boosted Tree	Runjhun Sharma, Priyal Gupta
Model Development: Logistics Regression	Sarthak Khanna, Sarthak Jain
Model Development: Logistics Regression with Feature Selection	Srishti Patil, Raksha Gujarathi
Model Development: Decision Forest vs Boosted Tree with Feature Selection	Varun Bhavnani, Sanika Jadhav
Final Model Deployment	Varun Bhavnani
Documentation (limitation, Assumptions & Business Value)	All

- End of Document -