

Predicting of Impact on Re-admission Rates for Patients Hospitalized with Heart Disease.

GANESH KASTURI

IMS Proschool Institute for Data Science

Andheri Batch, Mumbai(W)

1 INTRODUCTION.

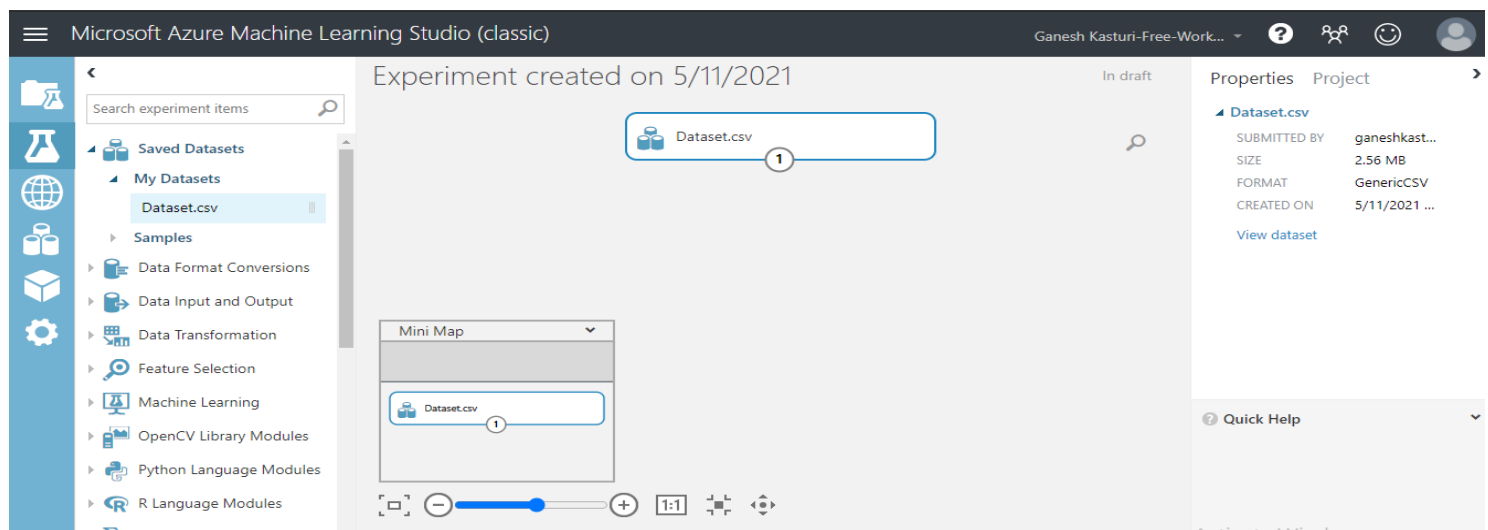
A healthcare organization together with a couple of government hospitals in a city has collected information about the vitals that would reveal if the person might have a coronary heart disease in the next ten years or not. This study is useful in early identification of disease and have medical intervention if necessary. This would help not only in improving the health conditions but also the economy as it has been identified that health performance and economic performance are interlinked.

As a data scientist, you are required to construct a classification model based on the available data and evaluate its efficacy. Your activities should include - performing various activities pertaining to the data such as, preparing the dataset for analysis; checking for any correlations; creating a model; evaluating the performance of the classification model.

2 METHODOLOGY

2.1 Data Set

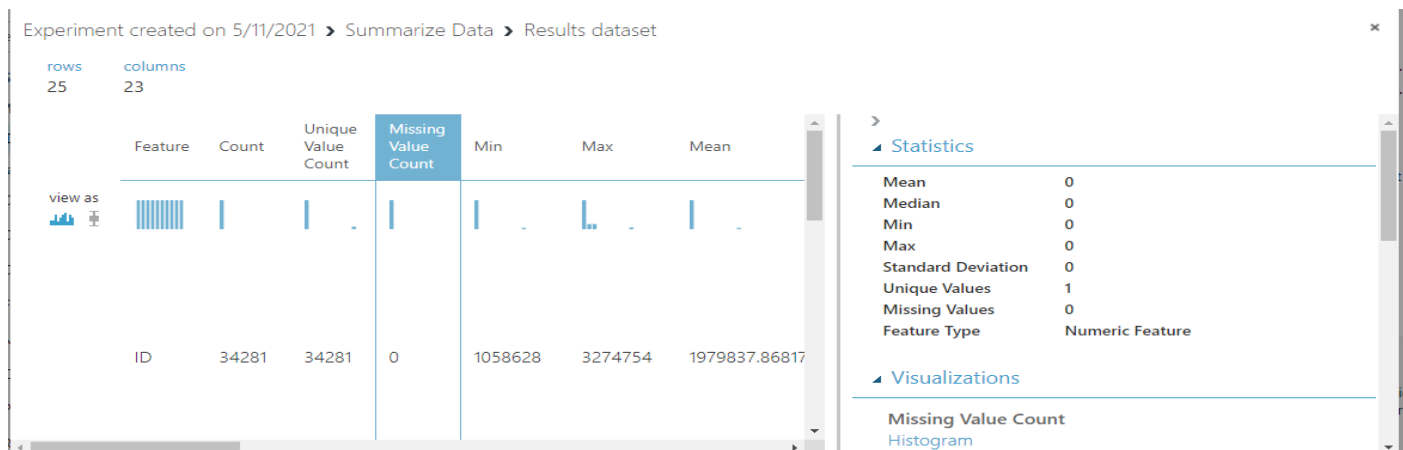
To explore this problem, we have taken a dataset and upload it into Microsoft Azure Machine Learning Studio(Classic).



2.2 Summarize Dataset.

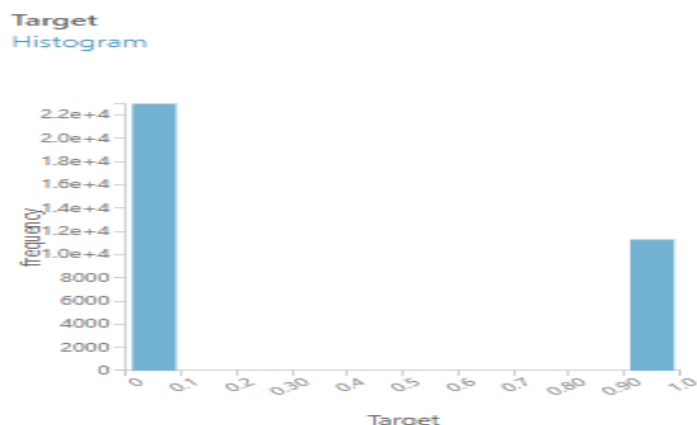
From the Summary of Dataset, we have find some explanation about the dataset.

1. There are Total **23** columns and Count of the dataset is around **34291**.
2. All the columns as of **Numeric Feature!** We must do less pre-processing because we do tree-based algorithm.
3. From the Summary we have Predict there are **No Missing values** Present in the Data Set.

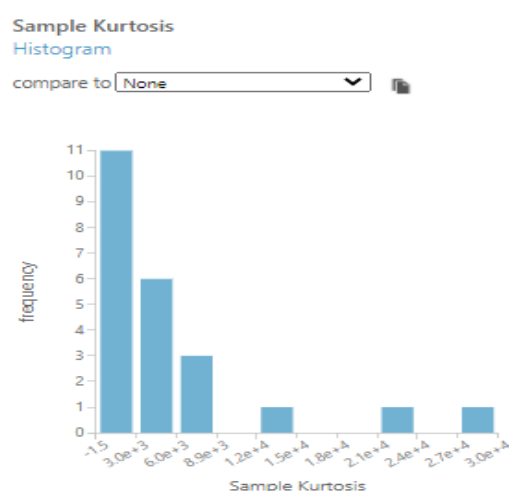
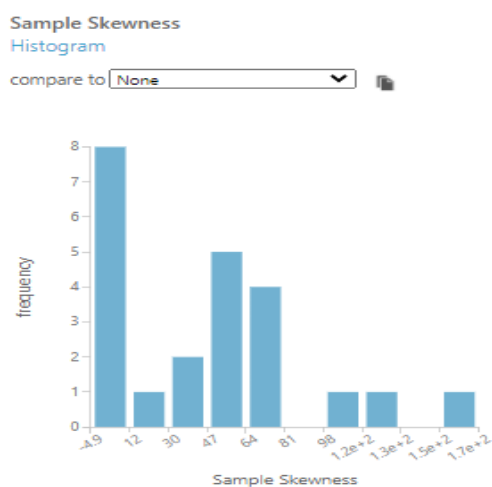


3 Exploratory Analysis

3.1 Prior to performing any analysis, we conducted exploratory analysis to preview the data type, attributes, and overall patterns of the data. We are interested in the class label “Target”.



3.2 We have Also Visualize the Skewness and Kurtosis to see how the data is Actually positively or negatively Skewness!

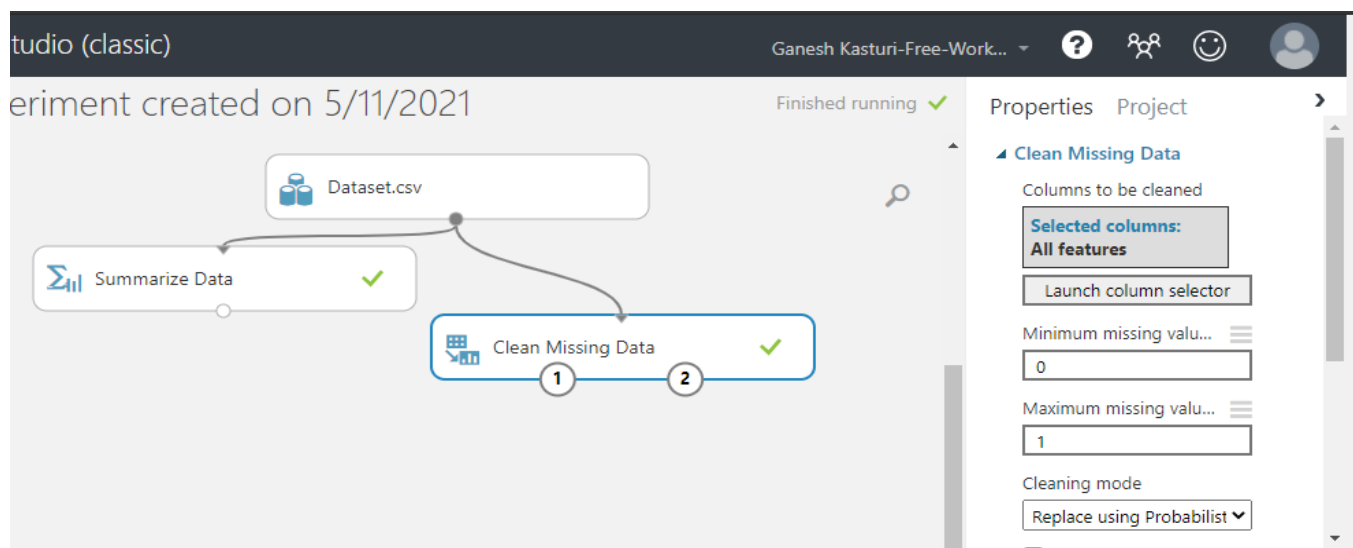


4 Data Pre-Processing

After the exploratory analysis, we found several challenges lies in the original dataset, and thus some data wrangling tasks such as data cleaning, dealing with missing values, creating new variables, and data transformation needs to be addressed before modelling.

4.1 Dealing with Missing Data.

We discovered no missing values; this dataset has 23 variables which contain no missing values. But for our requirement we have check by use the Tools and lets the Azure deal with the Missing values.



4.2 Normalize Data

Data Normalization is a common practice in machine learning which consists of transforming numeric columns to a common scale.

The screenshot shows the Azure Machine Learning Studio interface. The main workspace displays a workflow starting with a 'Dataset.csv' file. This file is processed by a 'Summarize Data' node, followed by a 'Clean Missing Data' node, and finally a 'Normalize Data' node. The 'Normalize Data' node is highlighted with a blue border and a green checkmark. The right-hand pane shows the 'Properties' tab for the 'Normalize Data' node. It includes a 'MinMax' dropdown, a checked 'Use 0 for constant ...' option, and a 'Columns to transform' section. Under 'Selected columns', it lists 'All columns' and 'Exclude column names: A2'. A 'Launch column selector' button is also present. Below this, a table shows the experiment's status: START TIME (5/11/20...), END TIME (5/11/20...), ELAPSED TIME (0:00:02.4...), STATUS CODE (Finished), and STATUS DETAILS (None).

5 Split the Dataset.

We have Split the dataset into 2 parses. Training 70% and to test the model 30%.

The screenshot shows the Azure Machine Learning Studio interface. The main workspace displays a workflow starting with a 'Dataset.csv' file. This file is processed by a 'Select Columns in Dataset' node, followed by a 'Clean Missing Data' node, then a 'Normalize Data' node, and finally a 'Split Data' node. The 'Split Data' node is highlighted with a blue border and a green checkmark. The right-hand pane shows the 'Properties' tab for the 'Split Data' node. It includes a 'Splitting mode' dropdown set to 'Split Rows', a 'Fraction of rows in the first...' input field set to '0.7', a checked 'Randomized split' option, a 'Random seed' input field set to '0', and a 'Stratified split' dropdown set to 'False'.

6 Train Model

We have Train the Model in Microsoft -Azure. We have a Prediction Variable "Target"

Select a single column

The screenshot shows the configuration for the 'Select Columns in Dataset' node. It features a 'BY NAME' tab and a 'WITH RULES' tab. The 'WITH RULES' tab is active, showing a configuration with 'Include' selected from a dropdown, 'column names' selected from another dropdown, and a text input field containing 'Traget' (with a red 'x' icon next to it).

The screenshot shows the Azure Machine Learning Studio interface. The main workspace displays a workflow starting with a 'Dataset.csv' file. This file is processed by a 'Summarize Data' node, followed by a 'Select Columns in Dataset' node, then a 'Clean Missing Data' node, then a 'Normalize Data' node, then a 'Split Data' node, and finally a 'Train Model' node. The 'Train Model' node is highlighted with a blue border and a green checkmark. The right-hand pane shows the 'Properties' tab for the 'Train Model' node, which is currently empty.

7 Two – Class Boosted Decision Tree

Since, we have Summarize, and from the EDA prediction we have seen that this problem we have to be predicted using classification algorithm.

Learning Studio (classic) Ganesh Kasturi-Free-Work... ?

Experiment created on 5/11/2021 In draft

Properties Project

Two-Class Boosted Decision Tree

Create trainer mode
Single Parameter

Maximum number of leaves per tree
20

Minimum number of samples per leaf node
10

Learning rate
0.1

Number of trees constructed
100

Dataset.csv Draft saved at 6:22:06 PM

Summarize Data ✓

Select Columns in Dataset ✓

Clean Missing Data ✓

Normalize Data ✓

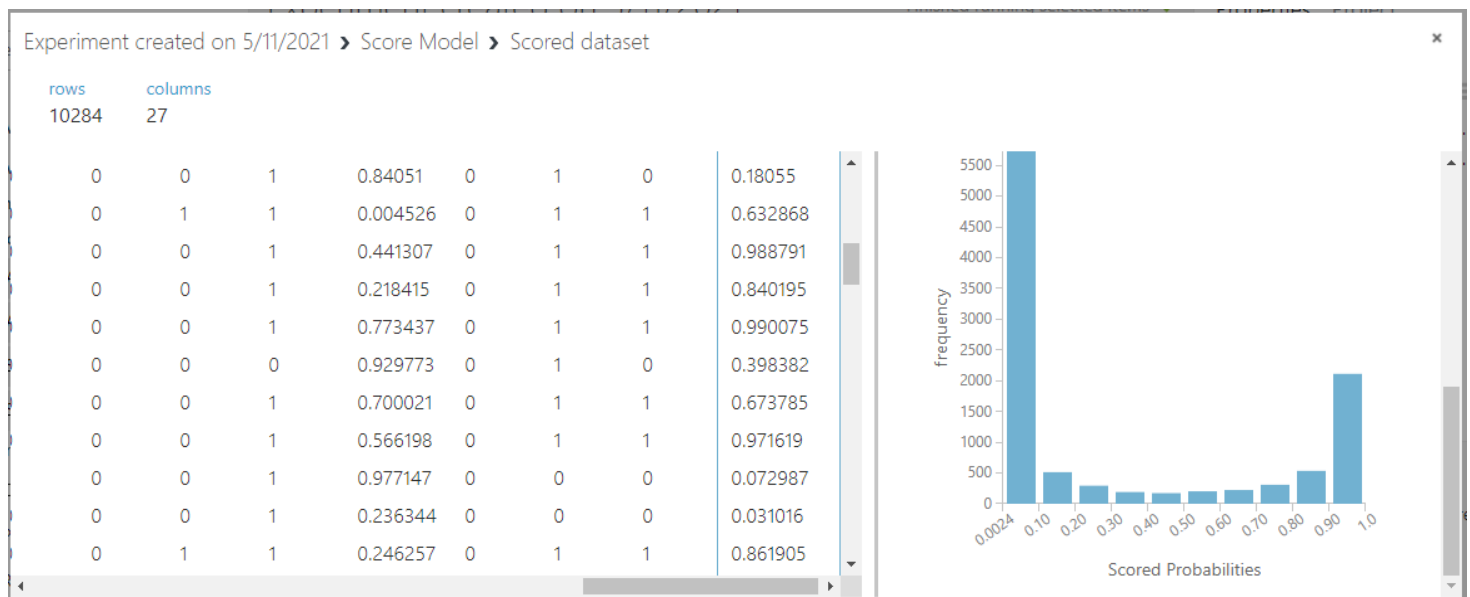
Split Data ✓

Train Model

Two-Class Boosted Decision... 1

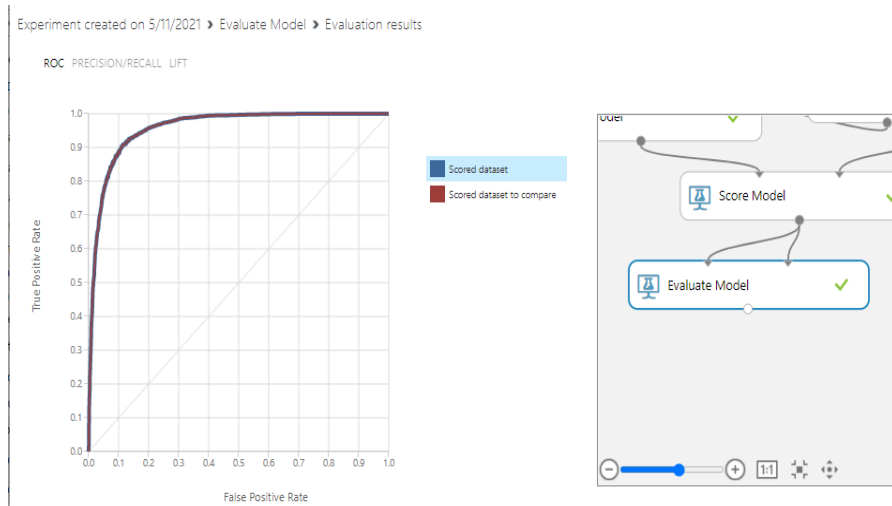
8 Score Model

We must Train the Model on testing data, so we have to perform the Score Model.



9 Evaluate Model:

Model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out-of-sample) data.

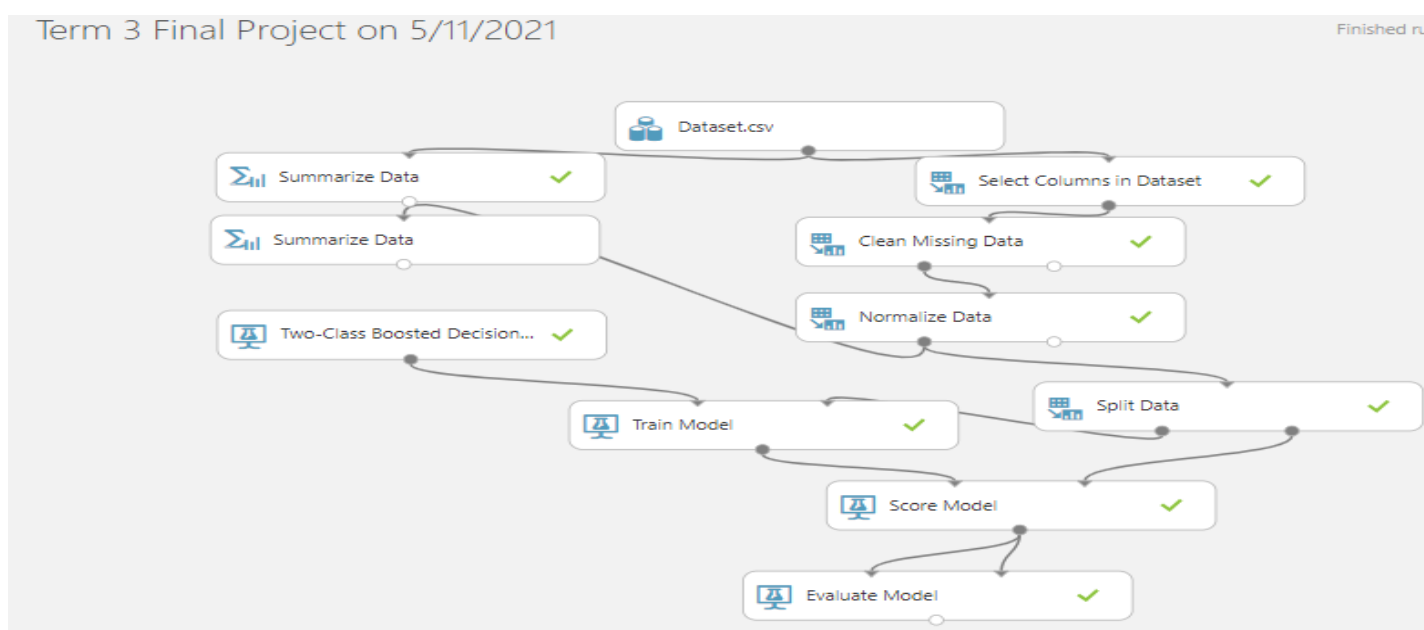


False Positive Rate

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
2854	517	0.898	0.843	0.5	0.958
False Positive	True Negative	Recall	F1 Score		
532	6381	0.847	0.845		
Positive Label	Negative Label				
1	0				

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	1951	148	0.204	0.848	0.713	0.929	0.579	0.827	0.979	0.008
(0.800,0.900]	428	118	0.257	0.878	0.791	0.899	0.706	0.870	0.962	0.019
(0.700,0.800]	225	85	0.287	0.891	0.823	0.881	0.772	0.895	0.949	0.028
(0.600,0.700]	137	89	0.309	0.896	0.837	0.862	0.813	0.911	0.936	0.038
(0.500,0.600]	113	92	0.329	0.898	0.845	0.843	0.847	0.925	0.923	0.049
(0.400,0.500]	85	87	0.346	0.898	0.848	0.826	0.872	0.936	0.910	0.060
(0.300,0.400]	79	116	0.365	0.894	0.847	0.804	0.895	0.946	0.894	0.075
(0.200,0.300]	93	198	0.393	0.884	0.839	0.769	0.923	0.958	0.865	0.101
(0.100,0.200]	108	419	0.444	0.854	0.811	0.704	0.955	0.973	0.804	0.158
(0.000,0.100]	152	5561	1.000	0.328	0.494	0.328	1.000	1.000	0.000	0.958

10 Final Model.



11 Suggest ways of improving the model.

After running the decision tree, we decided to use a boosting method by the relatively new algorithm tree-boosting for model improvement. Boosting is an ensemble method that create a strong classifier based on weak classifiers, according to how correlated are the learners to the actual target variable. The errors of the previous model are corrected by the next predictor, by adding models on top of each other iteratively until the training data is accurately predicted or a maximum number of models are added.

We applied and tuned the algorithm for better performance. e.g. learning rate to prevent overfitting (etc=0.01).

12 Any interesting observations.

- We have seen data is Imbalanced dataset from the Target variable visualization.
- Dataset is Normally Distributed.

13 CONCLUSIONS.

In this work we adopted machine learning methods using Microsoft Azure Studio to identify high risk patients and evaluated machine learning algorithms. Study achieved high accuracy due to the sophisticated pre-processing procedure. The Two-class Based Tree Boosting Algorithm method is reported to be the best method for prediction of the readmission rate for Heart Diseases.