Prediction of a Coronary Heart Disease

Khan Mukhtar

IMS Proschool Institute for Data Science. Andheri Batch, Mumbai(W)

1. Problem statement.

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 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.

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. Visualizations would be a value add.

2. Perform exploratory data analysis.

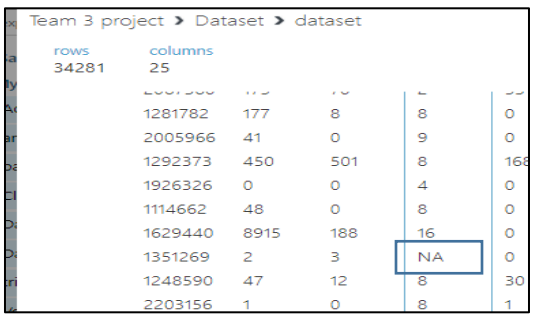
- **Summary of Dataset :**

I. There are Total **23** columns and Count of the dataset is around **34291**.

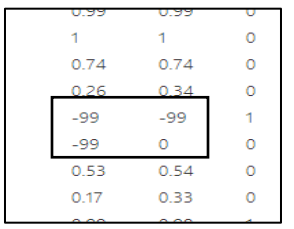
II. All variables are Numeric variables.

III. There are **missing/incorrect** values in columns i.e **A2,A15,A16.**

a) A2 column

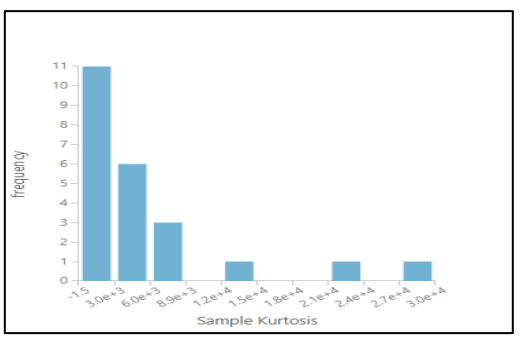


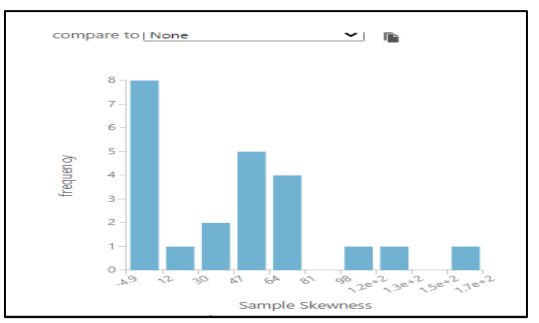
b) A15 ,A16 columns

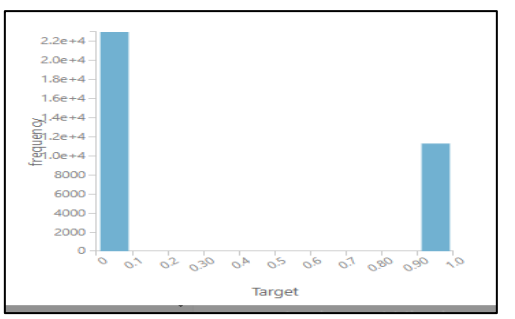


- **Exploratory Analysis**

A) We have Also Visualize the Skewness and Kurtosis to see how the data is Actually positively or negatively Skewness. And we can see there are more 0's than 1's in Target variable.

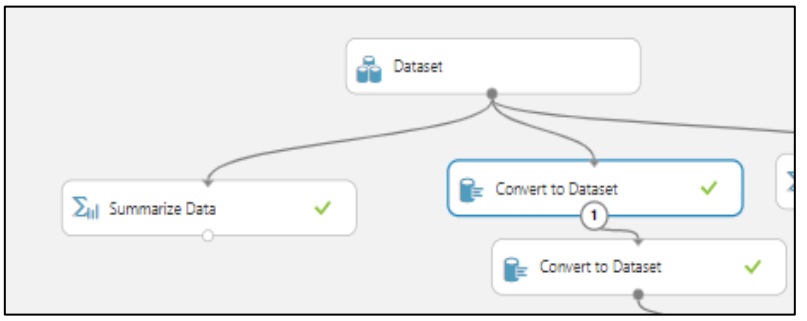


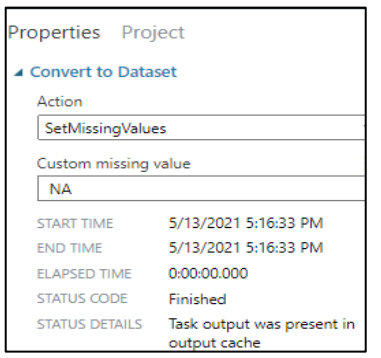


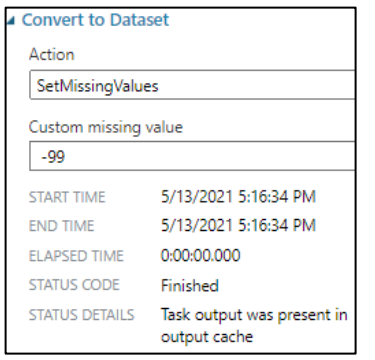


3. Data Pre-Processing.

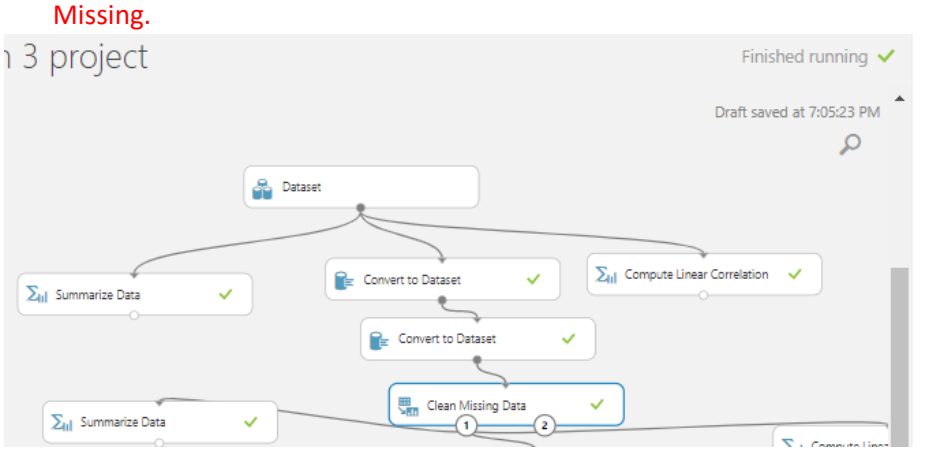
- **Missing Value Handling.**
  - AS we know that are dataset contain incorrect value & Outliers.
  - To handle them we had to use convert to Data module to set NA and -99 incorrect value as missing value because Azure didn't differentiate NA/-99 as missing value because it only identify blank rows as missing value. To replace with correct value we have make it Blank.
  - There are two same convert to data is used in first is for NA value & second for -99 value. See in below pic.

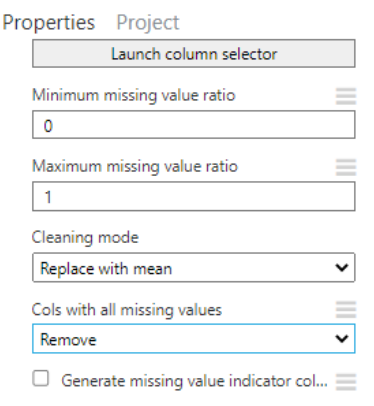






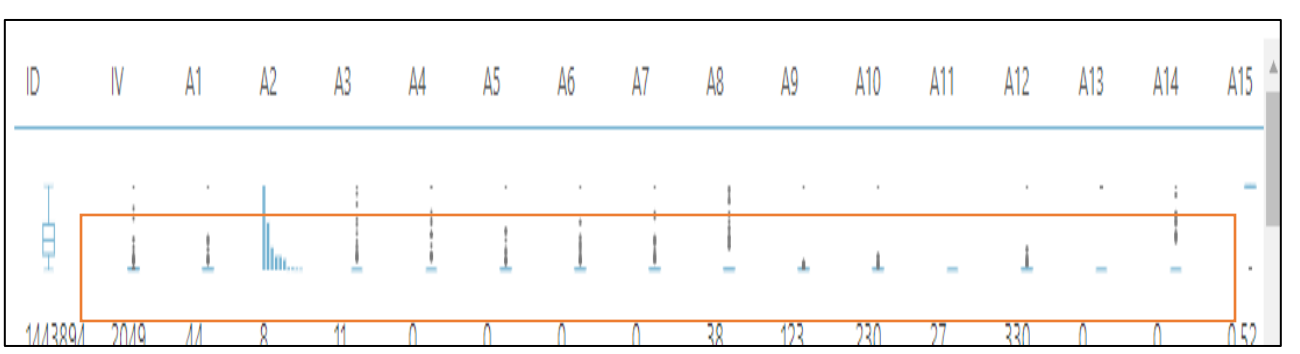
- So, After replace incorrect value with blank. Then we finally treat it with mean replacement technique. The module used is Clean Data Missing.

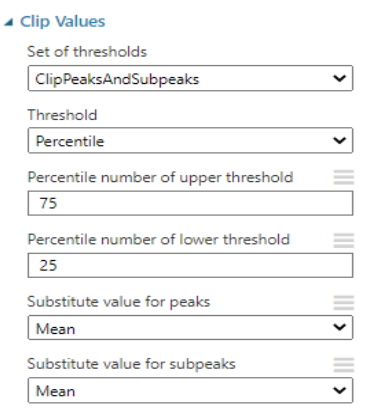




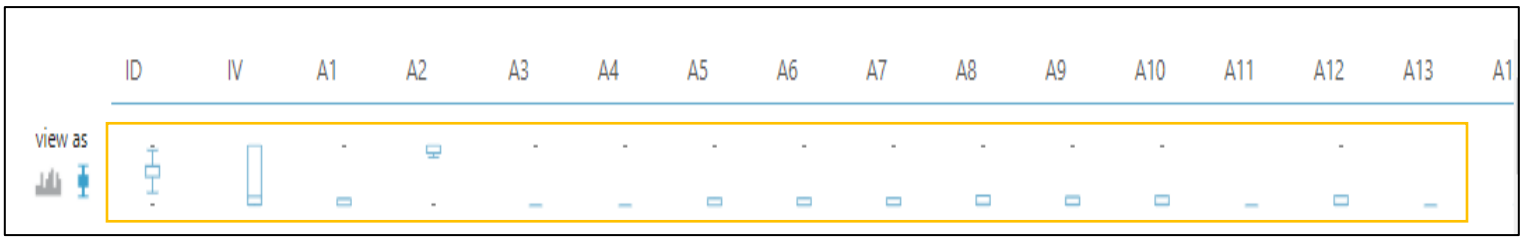
- **Outliers :-**
  - After treating missing value. Now we treat Outlier using Clip value module.
  - ClipPeaksandsubPeaks : It use to treat both upper & lower outliers. With Percentile as threshold .

  - **Before Handling outliers :-**

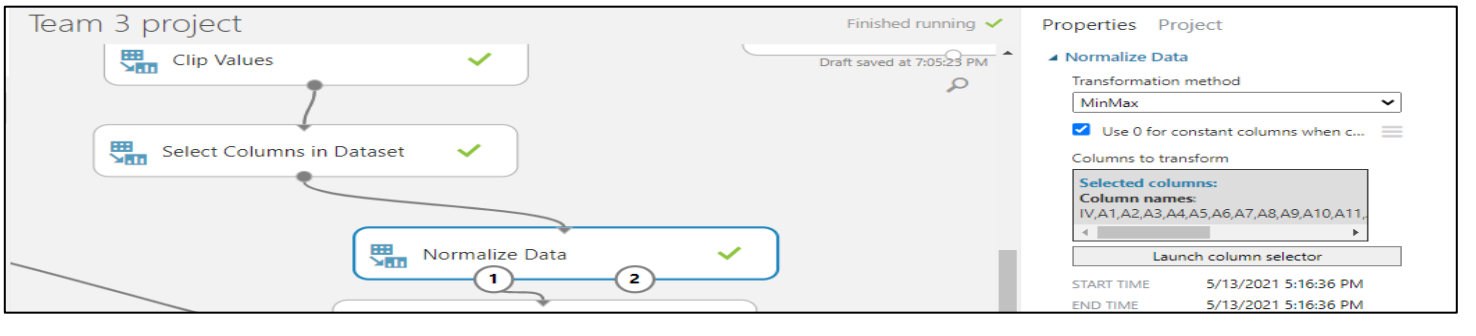




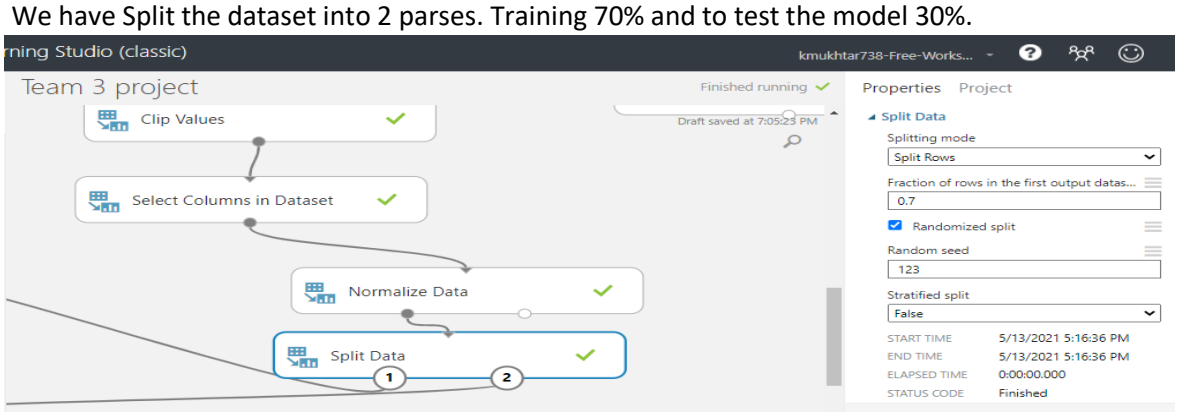
  - **After Treating Outliers :-**



- **Normalizing Data :**
  - All columns in dataset are numerical features but there larger gap between value.
  - Before making model on dataset, First Normalize data which help model to perform well.
  - Data Normalization is a common practice in machine learning which consists of transforming numeric columns to a common scale.
  - Select columns that want to normalize with select columns in Datasets



4. **Select Training data, test data**.
   - We have Split the dataset into 2 parses. Training 70% and to test the model 30%.

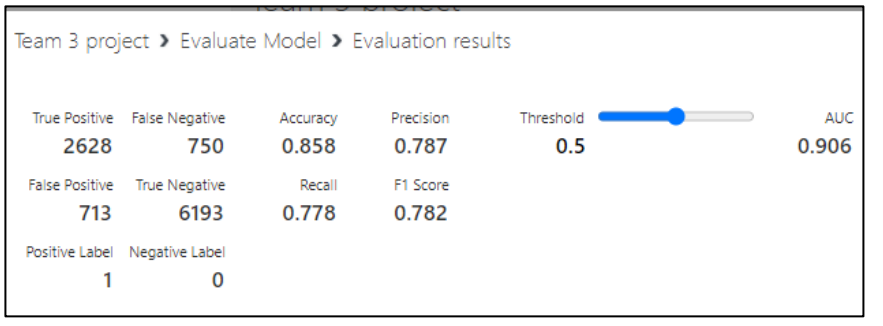


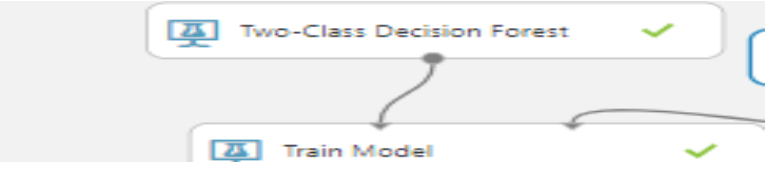
5. **Train the model.**

Model Comparison :

1) Two class Decision Forests model :

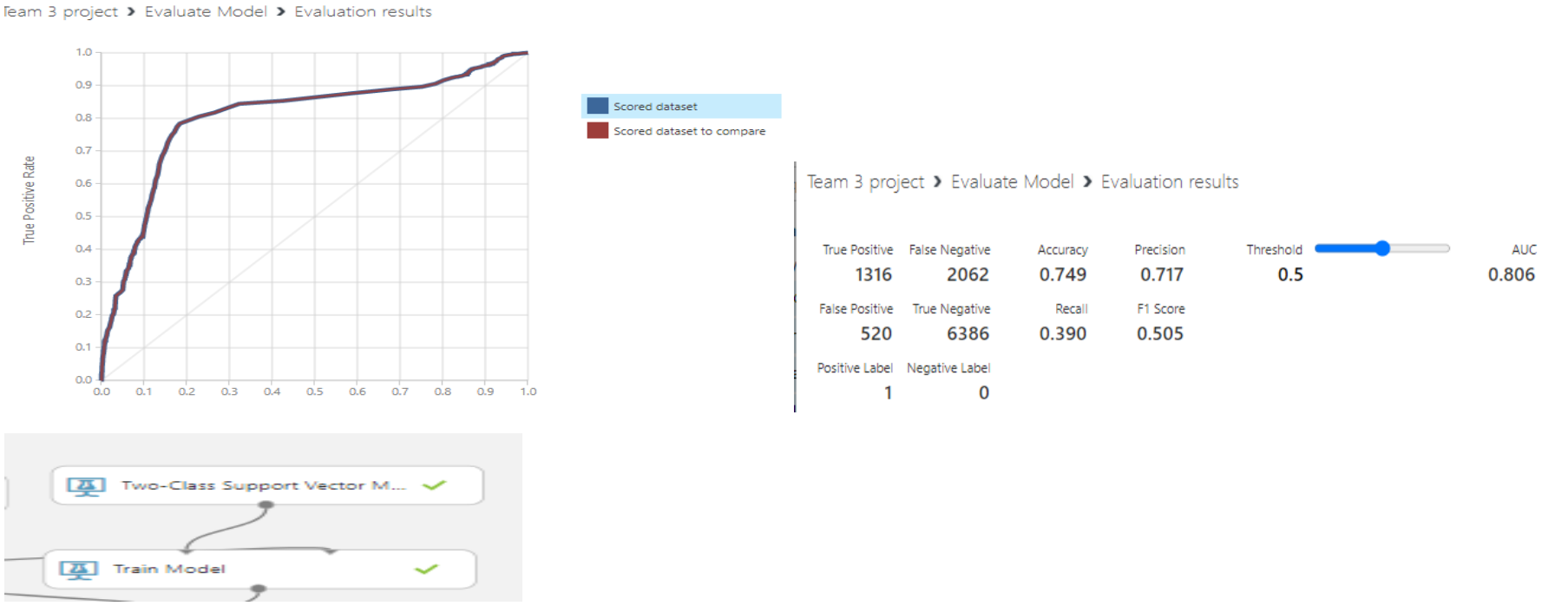
- Decision forests are fast, supervised ensemble models.
- This module is a good choice if you want to predict a target with a maximum of two outcomes.
- Ensemble methods are based on the general principle that rather than relying on a single model, you can get better results and a more generalized model by creating multiple related models and combining them in some way.
- After using this algorithm to create model.The preform of model is pretty good.
- As we see most of the area is under the curve i.e AUC more than 90 %.
- After doing hyperparameter tuning,we achieve **85% accuracy**





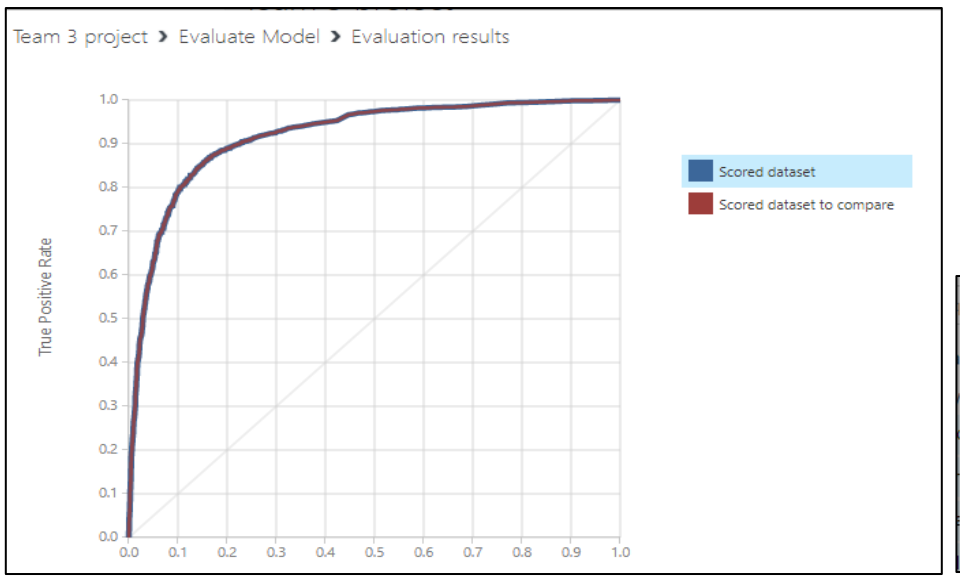
2) Two class Support Vector Machine model :

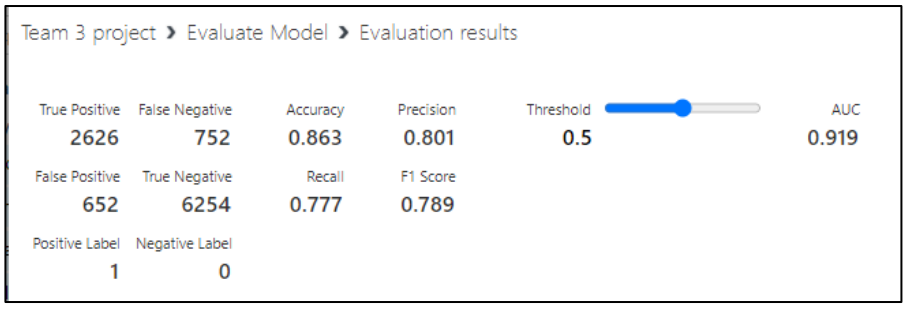
- Support vector machines are among the earliest of machine learning algorithms, and SVM models have been used in many applications, from information retrieval to text and image classification.
- SVMs can be used for both classification and regression tasks.
- It requires labeled data because it is supervised learning model.
- In the training process, the algorithm analyzes input data and recognizes patterns in a multi-dimensional feature space called the hyperplane.
- As we see, after implement of this model is good but not compare to Two class Decision Forests**.**
- ROC –AUC is less it is 80% And also Accuracy 74% which less than other model.
- We can improve it with using hyper meter tuning.

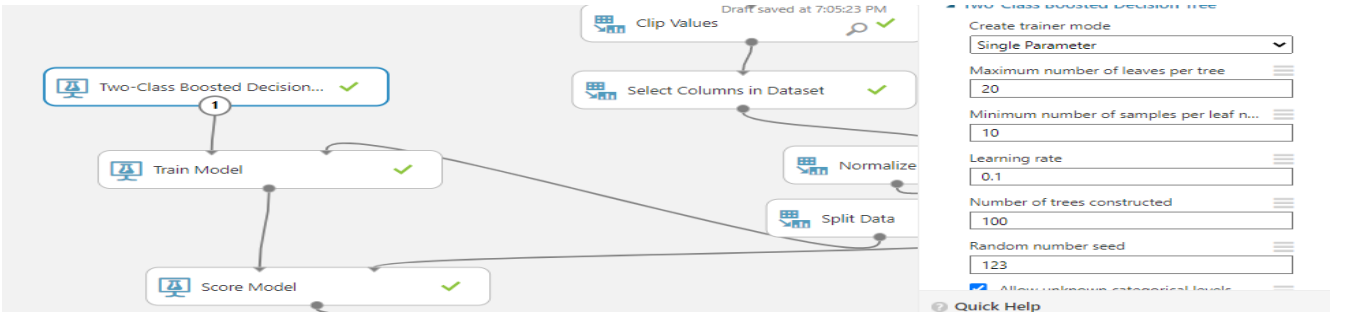


3) **Two-Class Boosted Decision Tree** :

- A boosted decision tree is an ensemble learning method in which the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the first and second trees, and so forth.
- Predictions are based on the entire ensemble of trees together that makes the prediction.
- Boosted decision trees are the easiest methods with which to get top performance on a wide variety of machine learning tasks.
- After implement, we get excellent perform with 86% accuracy & AUC with 91% with help of hyperpermeter tuning.
- This model can explain 86% of data from dataset.





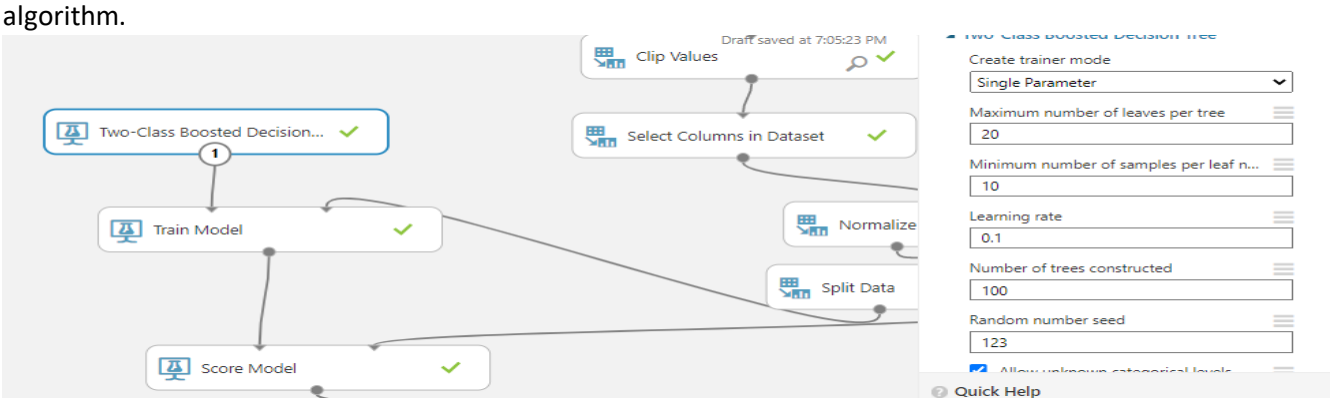


4) **Model Conclusion :**

- After see all model perform we can decide to use Two class Boosted Decision Forests.
- Cause it give best accuracy compare to other model.
- We can improve it with performing more hyperparameter tuning & learning rate.

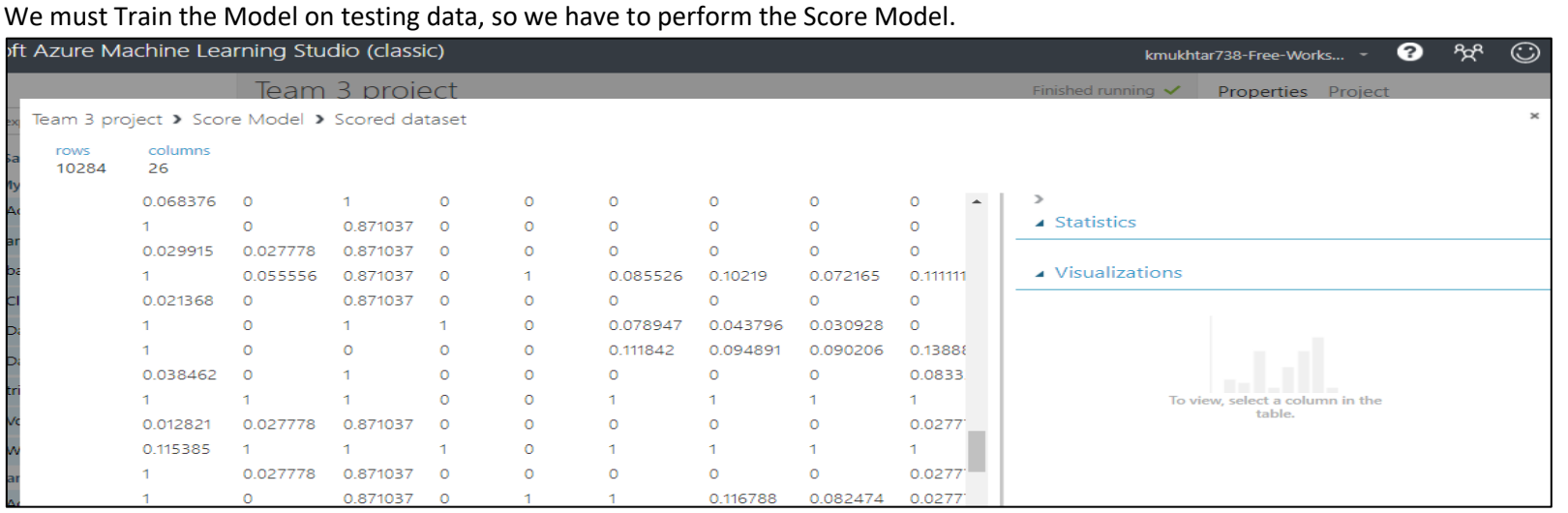
Model Implement : Two class Decision Forests :-

- To trained model we use Two – Class Boosted Decision Tree.
- We have Summarize, and from the EDA prediction we have seen that this problem we have to be predicted using classification algorithm.



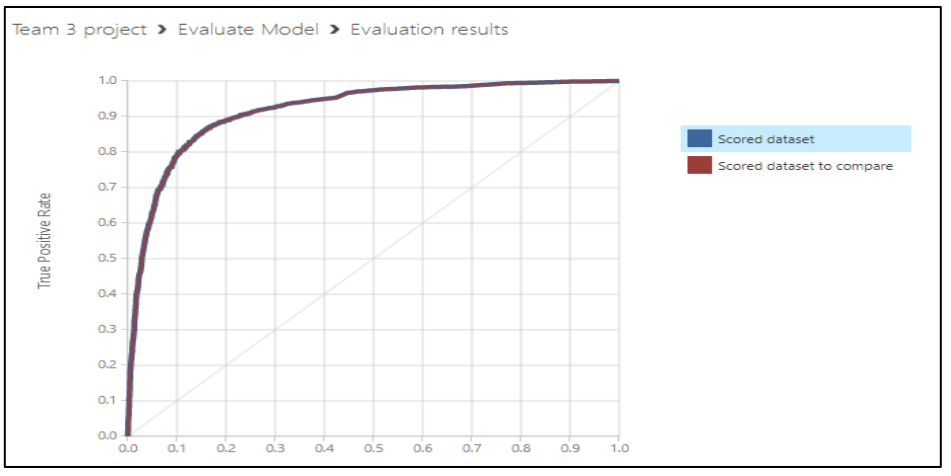
6. Score Model

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



7. Evaluate Model:

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



Team 3 project › Evaluate Model › Evaluation results

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
2626	752	0.863	0.801	0.5	0.919

False Positive	True Negative	Recall	F1 Score
652	6254	0.777	0.789

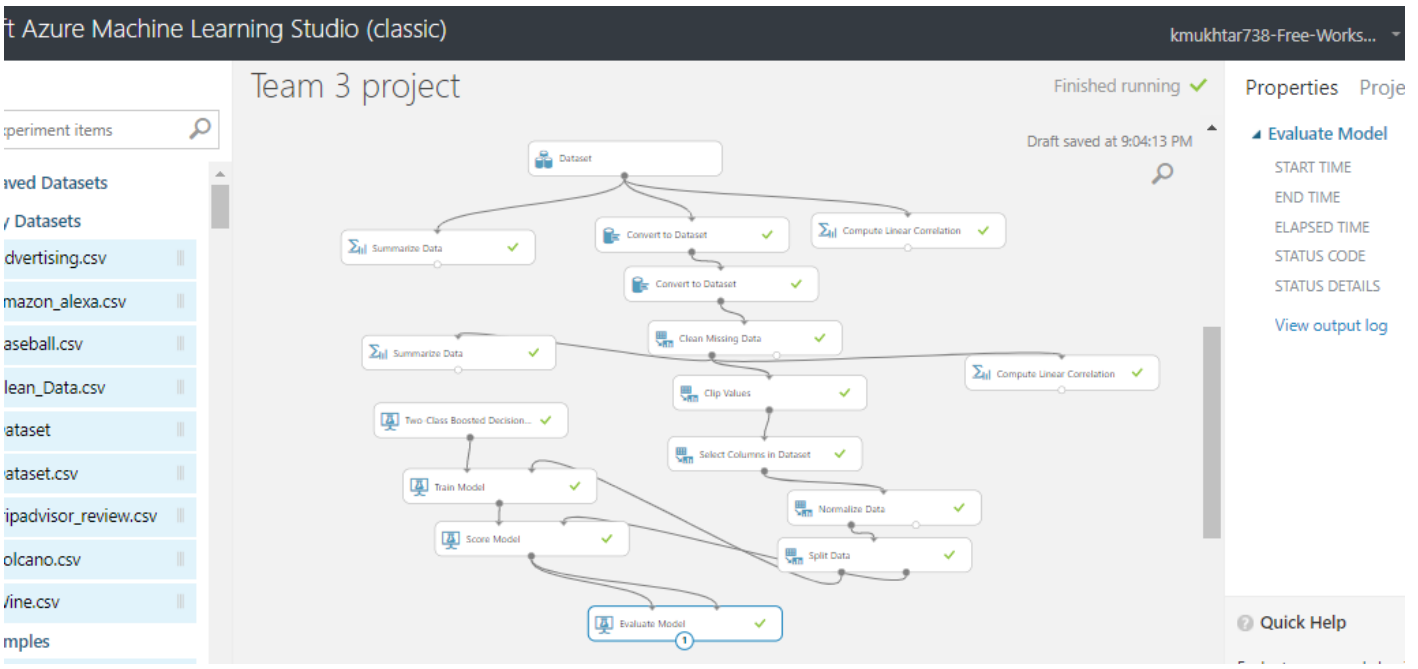
Positive Label	Negative Label
1	0

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	1250	113	0.133	0.782	0.527	0.917	0.370	0.761	0.984	0.004
(0.800,0.900]	591	125	0.202	0.827	0.675	0.886	0.545	0.813	0.966	0.012
(0.700,0.800]	362	140	0.251	0.849	0.739	0.854	0.652	0.847	0.945	0.024
(0.600,0.700]	244	134	0.288	0.860	0.772	0.827	0.724	0.873	0.926	0.038
(0.500,0.600]	179	140	0.319	0.863	0.789	0.801	0.777	0.893	0.906	0.053
(0.400,0.500]	140	184	0.350	0.859	0.793	0.768	0.819	0.908	0.879	0.074
(0.300,0.400]	126	211	0.383	0.851	0.790	0.734	0.856	0.923	0.848	0.100
(0.200,0.300]	101	297	0.422	0.832	0.776	0.690	0.886	0.935	0.805	0.137
(0.100,0.200]	105	483	0.479	0.795	0.746	0.629	0.917	0.948	0.735	0.201
(0.000,0.100]	280	5079	1.000	0.328	0.495	0.328	1.000	1.000	0.000	0.919

-

-

8. **Final Model.**



9. **Suggest ways of improving the model.**
    - After Testing all algorithm and see all accuracy, finally decided to use boosting method with Decision tree which give good accuracy compare with other algorithm.
    - It give good accuracy because of boosting method which works on creating strong classifier based on weak classifiers, according to how correlated are the learners to the actual target variable.
    - To improve model a model perform more we can use Hypermeter tuning & learning rate. It help to prevent overfitting.
    - We had previously use learning rate=0.1 , to improve we can try different learning rate & other parameter i.e (Number of tree, Number of leaves per tree, etc.)

10. **Any interesting observations.**
    - Dataset was imbalance & there can more pre processing can be done to make dataset more well for model creation.
    - The value was not proper.
    -

11. **Challenges faced and how you mitigated the challenges.**
    - Azure ML studio is new to me so at some point it hard to work with but with help of online source, it work.
    - Other challenges faced by me is while data pre-processing because data variable are unlabeled.

12. **Assumptions if any/ Conclusions.**
    - In this project, we have use machine learning in Azure ML studio to identify , if person will have Coronary Heart Disease in next ten year.
    - With help of Two class boosted decision tree we have make a model.
    - After, performing confusion matrix we can say that model have predict that most of the people do not Coronary Heart Disease in dataset.
    - The Two-class Based Tree Boosting Algorithm method is reported to be the best method for prediction of the Coronary Heart Disease.