DEFAULT OF CREDIT CARD CLIENTS

Problem Identification

- For banks, risk management and default detection has always been a crucial part in issuing credit card.
- Default on credit card bills can result in great financial loss
- To reduce or even prevent loss of this kind, banks need to determine whether appropriate people are given credit based on their banking habits and information.

Summary

Here we explore the possibility of predicting the case of customers default payments in Taiwan using various factors of the customers such as age, gender, history of past payment, education, Amount of bill statement, Amount of previous payment.

Description

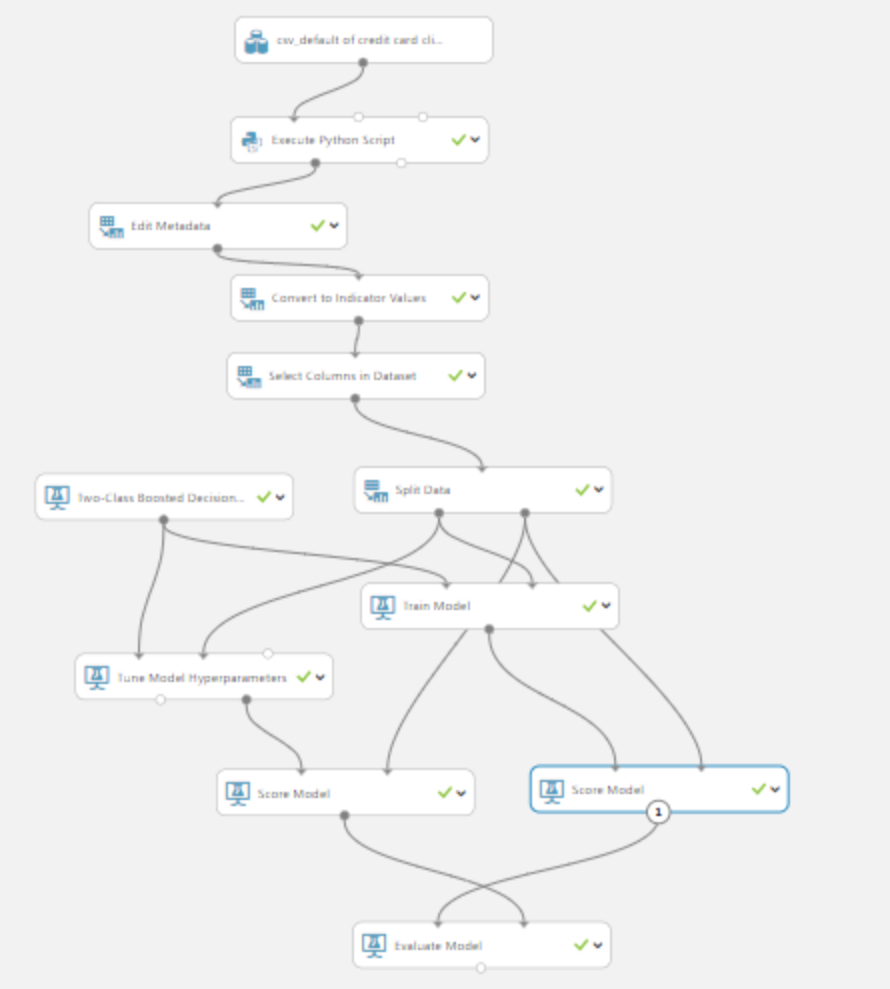
The purpose of this experiment is to explore the possibility how machine learning can be used to predict whether a customer can be given credit based on his banking habits. This will greatly help banks to prevent financial losses. we will be using a machine learning technique called Classification for this task. Classification is considered an instance of supervised learning, i.e., learning where a training set of correctly identified observations is available.

Dataset

	ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4
1	1	20000	2	2	1	24	2	2	-1	-1	-2	-2	3913	3102	689	0
2	2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272
3	3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331
4	4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314
5	5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940

BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	default payment next month
0	0	0	689	0	0	0	0	1
3455	3261	0	1000	1000	1000	0	2000	1
14948	15549	1518	1500	1000	1000	1000	5000	0
28959	29547	2000	2019	1200	1100	1069	1000	0
19146	19131	2000	36681	10000	9000	689	679	0

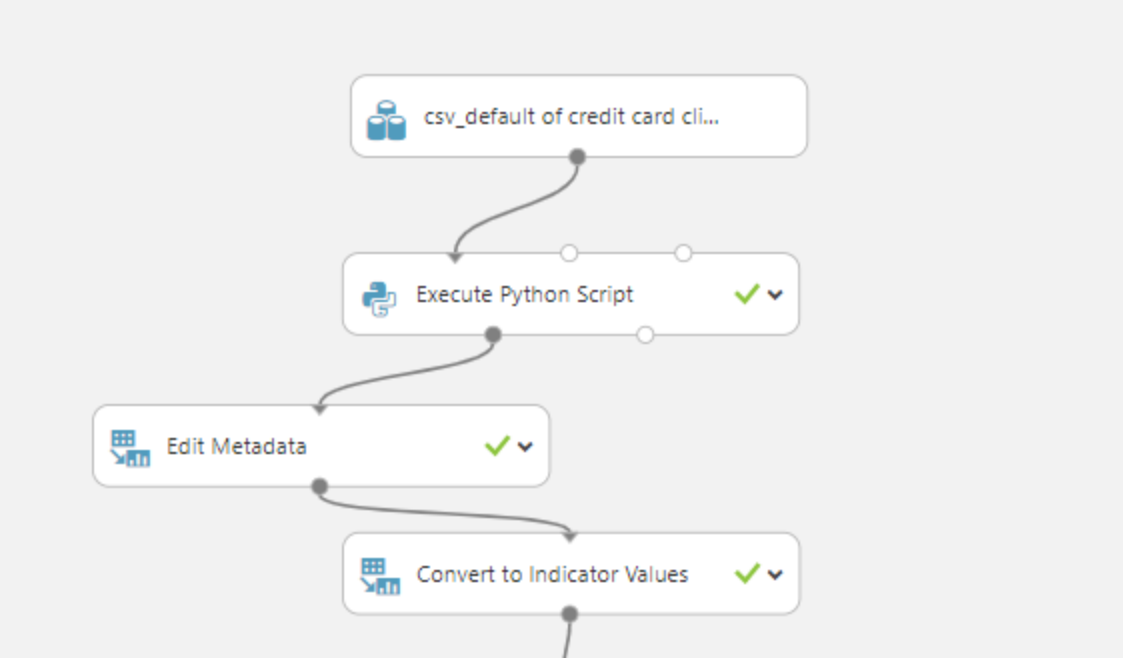
The datasets consist of 30000 rows and 24 columns. Every row in our table represents a unique customer and every customer is represented by 23 different factors such as age, gender, history of past payment, education, Amount of bill statement, Amount of previous payment etc. The default payment next month column indicates whether a customer has been a defaulter in the next month. The original data and the description of individual columns can be found in the UCI data repository.



Model

Pre-process Data

The first few steps consists of preparing the data for the downstream modules.



1. After importing the data from the source, we remove duplicates in the data if any. This step is optional.
2. Execute python script module allows us to write python code and it runs that code on the data frame.
3. Edit metadata cell helps us to change the datatype of our attributes, convert them to categorical etc.
4. Next, we will convert the attributes which was convert to categorical type in the previous step to one hot encoded value. The name one hot encoding comes from the way we design the values where one bit is hot/on and rest are set off.

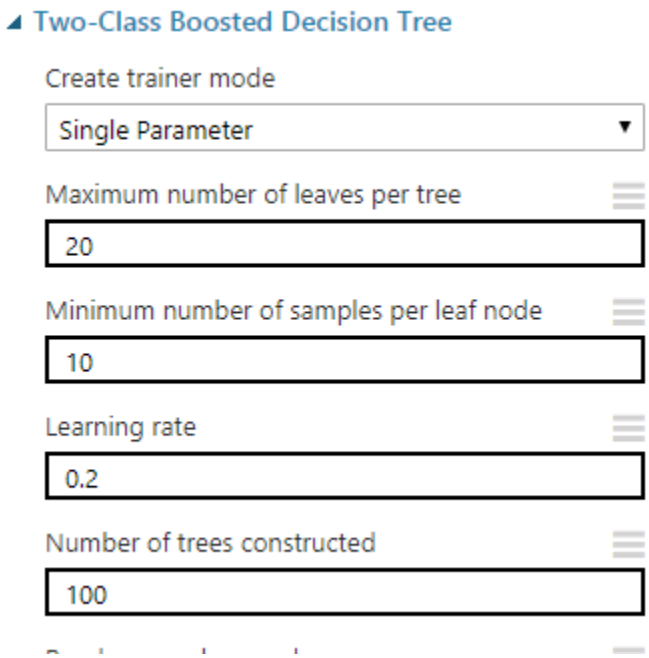
Learners

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.

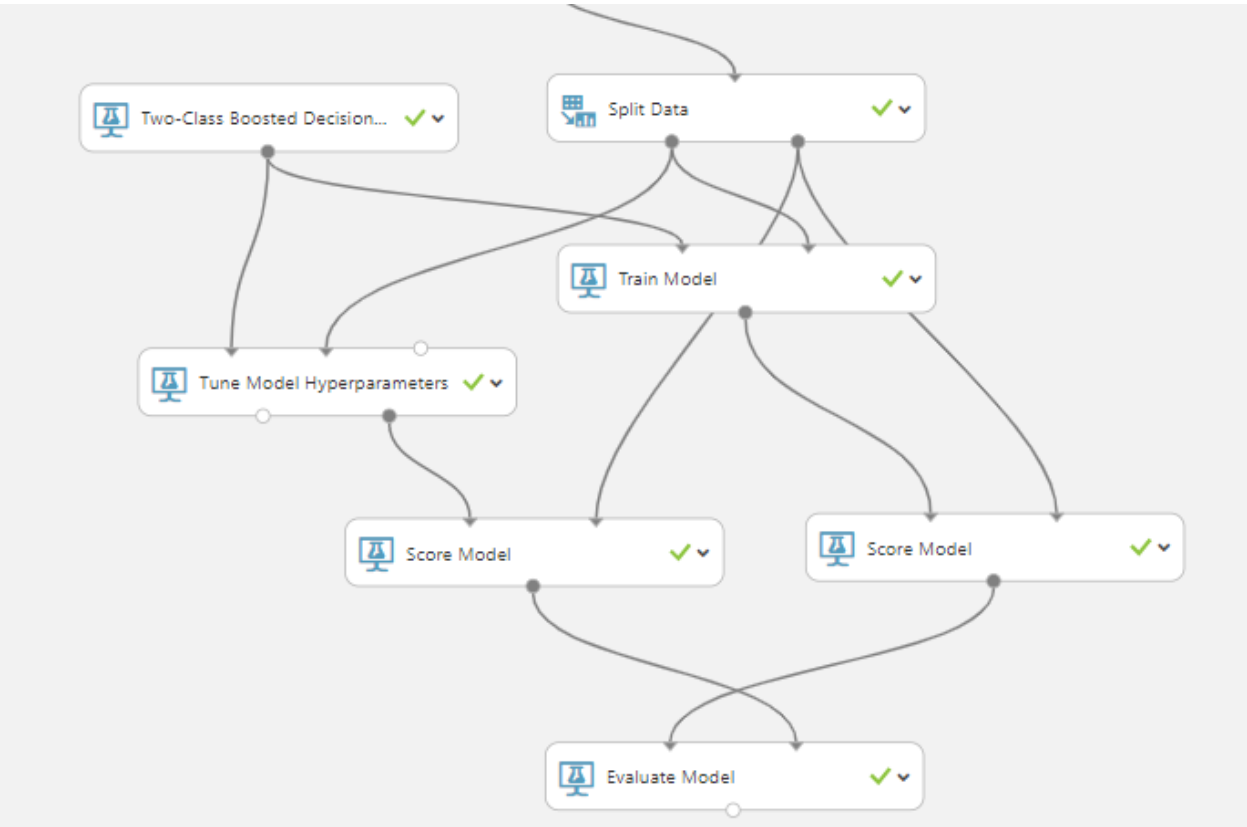
Generally, when properly configured, boosted decision trees are the easiest methods with which to get top performance on a wide variety of machine learning tasks

Learner parameters

The parameters that are specific to our learner is called the learner parameters. These are the knobs and handles of our learner which need to tune to extract the best out of our learner. These parameters vary with the learner.



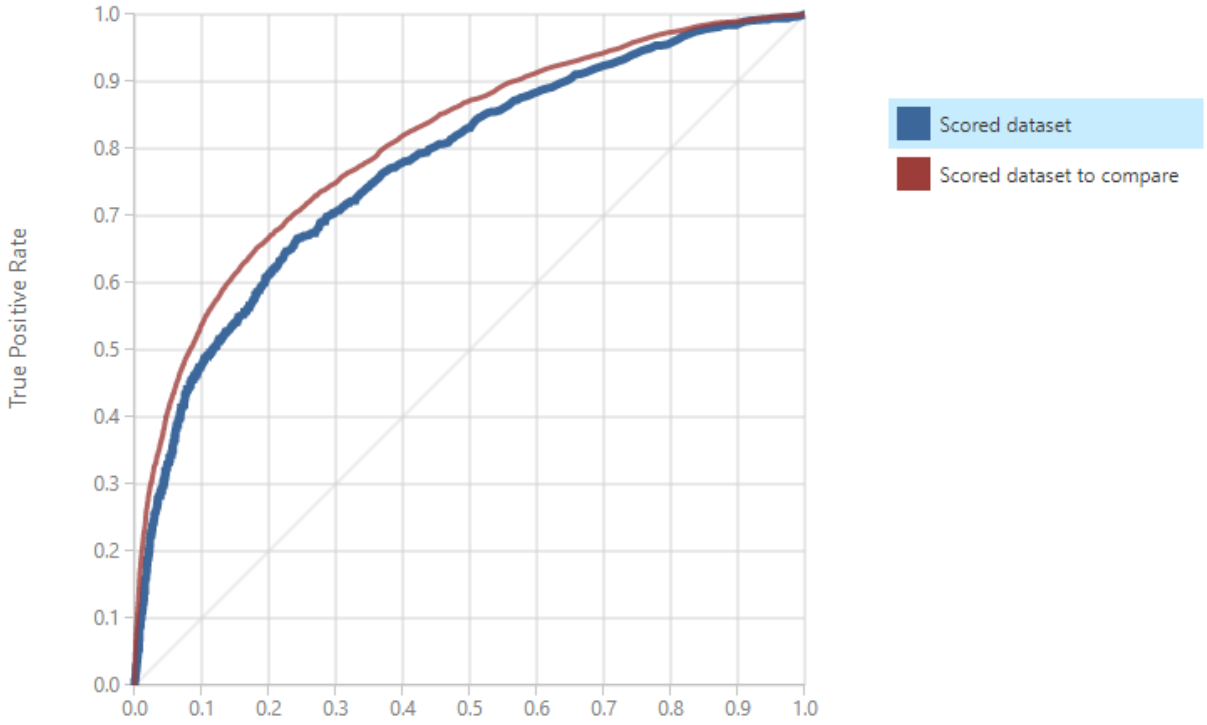
1. **Maximum number of leaves per tree**, indicate the maximum number of terminal nodes (leaves) that can be created in any tree.
2. **Minimum number of samples per leaf node**, indicate the number of cases required to create any terminal node (leaf) in a tree.
3. The **learning rate** determines how fast or slow the algorithm converges on the optimal solution.
4. **Number of trees constructed**, indicate the total number of decision trees to create in the ensemble.



Results

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represent degree or measure of separability. It tells how much model can distinguish between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between classes.

The below graph shows the ROC curve for our learner. The AUC Score of our learner is 77.4% which is not bad. Blue line represents the learner's performance before tuning hyperparameters. Red line represents the learner's performance using the best hyperparameters. We clearly notice the performance has improved after tuning the hyperparameters.



The below table is called the confusion matrix. It represents the number of True positives, True negatives, False positives and False negatives counts of our model on the test set. Ideally, the number of True positives and True negatives should be maximum. Confusion matrix is one of the important classification metrics.

True Positive	False Negative
561	735
False Positive	**True Negative**
351	4273

Informally, accuracy is the fraction of predictions our model got right. Precision refers to the percentage of your results which are relevant. On the other hand, recall refers to the percentage of total relevant results correctly classified by your algorithm.

Accuracy	Precision
0.817	0.615
Recall	**F1 Score**
0.433	0.508

Credits

- https://www.wikipedia.org/
- https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/tune-model-hyperparameters
- https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-boosted-decision-tree
- https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
- https://gallery.azure.ai/Experiment/Anomaly-Detection-Credit-Risk-5
- https://towardsdatascience.com/precision-vs-recall-386cf9f89488