

Credit risk prediction in Azure ML Studio

COMP 4211 - Tutorial 09

Chun-Kit Yeung

Hong Kong University of Science and Technology

2018-04-20

In this tutorial, I am going to introduce how you can do a machine learning project without coding, by using the Azure ML Studio. Moreover, I will introduce other evaluation measures other than accuracy.

The example I am going to use in this tutorial is **credit risk prediction**. It is an important task because loans to individuals become the most vulnerable segment of commercial banks' investment in a volatile financial environment. Searching safe methods for modelling refund loans reliability is one of the methods for credit losses risk reduction in commercial banks.

Create an experiment

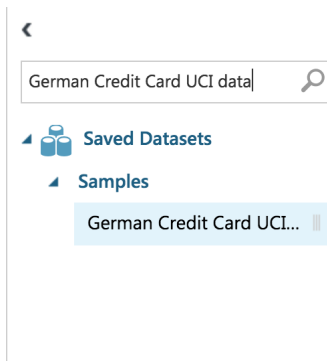
Let's create an experiment using Azure ML Studio.

Create Experiment

- 1 Create a new experiment by clicking **+NEW** at the bottom of the Machine Learning Studio window, select **EXPERIMENT**, and then select **Blank Experiment**.
- 2 The experiment is given a default name that you can see at the top of the canvas. Select this text and rename it to something meaningful, for example, **Credit Risk Prediction**. The name doesn't need to be unique.

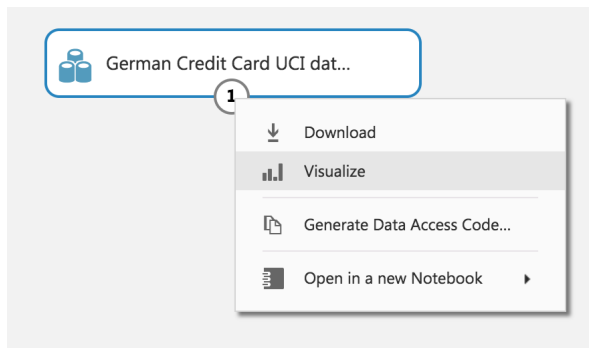
Choose the dataset

- ③ To the left of the experiment canvas is a palette of datasets and modules. Type **German Credit Card** in the Search box at the top of this palette to find the dataset labeled **German Credit Card UCI dataset**. Drag this dataset to the experiment canvas.



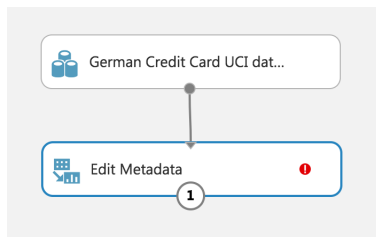
Visualize the dataset

Let's see what the data looks like using Azure. Click the output port at the bottom of the dataset, and then select **Visualize**.



Rename the columns

After visualizing, it is found out that the column names are not really meaningful. Here we are going to give the column names by **Edit Metadata**. Search this module in the search canvas, and drag it to the main canvas.

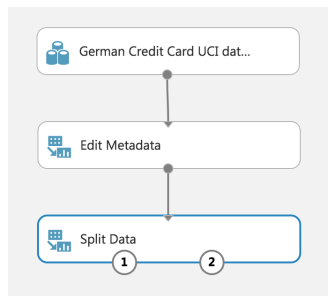


Rename the columns

- Select all the available columns in the “Launch column selector”.
- In the “new columns name”, fill it with the following text:
Status of checking account, Duration in months, Credit history, Purpose, Credit amount, Savings account/bond, Present employment since, Installment rate in percentage of disposable income, Personal status and sex, Other debtors, Present residence since, Property, Age in years, Other installment plans, Housing, Number of existing credits, Job, Number of people providing maintenance for, Telephone, Foreign worker, Credit risk

Train-test split

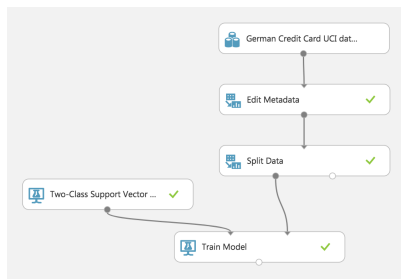
Then, we are going to split the whole dataset into training set and test set. To do so, search “Split Data” in the search canvas, and drag it to the main canvas. Let’s leave the default setting and you could try to modify them.



Afterwards, run the experiment now for once now, and have a look at the training set and the test set by using the “Visualize” in the “Split Data”.

Choose and train the model

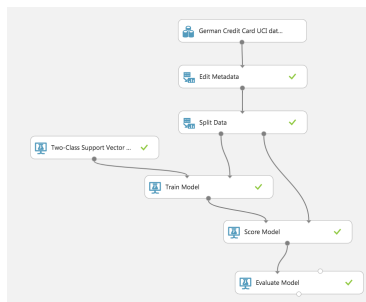
In this tutorial, we are going to use the support vector machine. Search it on the search canvas, choose “Two-Class Support Vector Machine”. Moreover, search “Train Model” in the search canvas and drag it to the main canvas. Eventually, connect them as follows:



By doing so, “Train Model” will train the machine learning model of “Two-Class Support Vector Machine” with the data in the first port of “Split Data”.


Evaluate the model

To evaluate the trained model in Azure ML Studio, “Score Model” and “Evaluation Model” are useful. The score model takes the trained model and make a prediction on the test dataset, whereas the evaluation model provides a thorough analysis on the result of score model. Search and drag them to the main canvas, and connect them as follows:



Evaluation measure in evaluation model

When you visualize the “Evaluation Model”, you will find that there are many evaluation measures provided.

True Positive	False Negative	Accuracy	Precision	Threshold		AUC
74	76	0.738	0.574	0.5		0.760
False Positive	True Negative	Recall	F1 Score			
55	295	0.493	0.530			
Positive Label	Negative Label					
2	1					

What are these?

Precision, Recall and F1-score

CONFUSION MATRIX

TP = True Positives

TN = True Negatives

FP = False Positives

FN = False Negatives

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

AUC

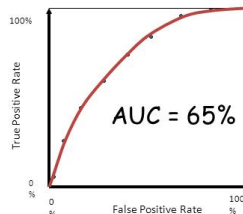
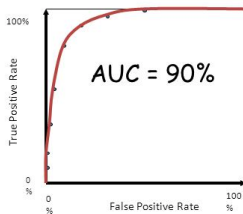
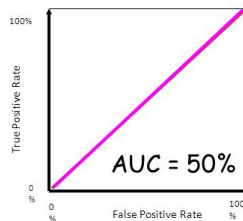
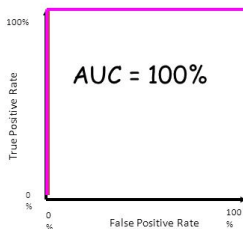
AUC stands for area under the ROC curve. The ROC curve depends on the true positive rate (TPR) and false positive rate (FPR) by varying the threshold of the prediction. The TPR and FPR are defined as:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$


AUC

AUC for ROC curves



Evaluation measure in evaluation model

After knowing the evaluation measures provided by the Evaluation Model, how would you comment on the model proposed? Is it a good or bad model?

True Positive	False Negative	Accuracy	Precision	Threshold		AUC
74	76	0.738	0.574	0.5		0.760
False Positive	True Negative	Recall	F1 Score			
55	295	0.493	0.530			
Positive Label	Negative Label					
2	1					

If it is good, how good is it?

If it is bad, how bad is it? What are the potential ways to improve it?