GitHub Issue

Multi-class Classification

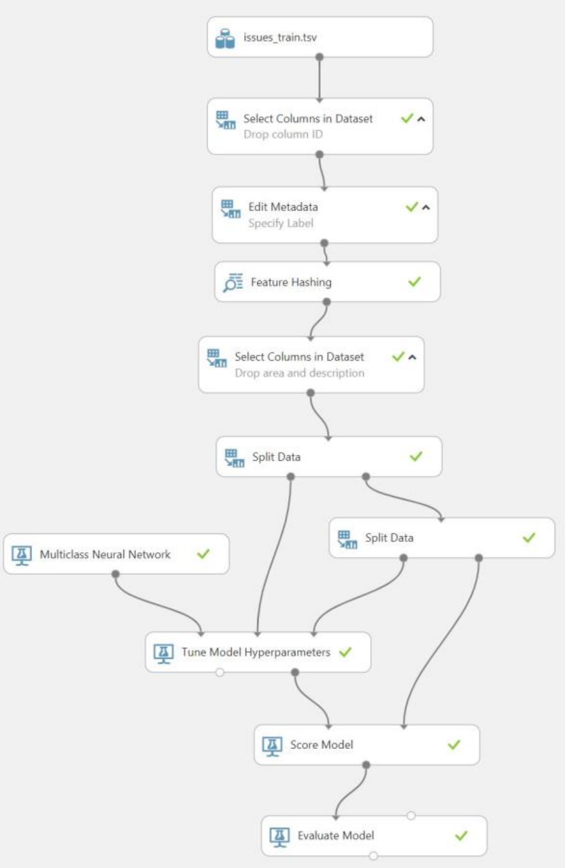
in Azure ML

In this session

- Question and Data
- Dataset description
- Create Data Set
- Place dataset
- Drop column ID
- Specify Label
- Hash feature
- Drop area and description
- Split Data
- Train, Score, Evaluate
- Metrics for multiclass classification Evaluation

The finished model

https://raw.githubusercontent.com/laploy/ML.NET/master/GitHub-Issue/github-issue-azureML-model.JPG


issues_train.tsv
Select Columns in Dataset
Drop column ID
Edit Metadata
Specify Label
Feature Hashing
Select Columns in Dataset
Drop area and description
Split Data
Multiclass Neural Network
Split Data
Tune Model Hyperparameters
Score Model
Evaluate Model

Question and Data

Question: what is the category of this issue?

Dataset:

issues_train.tsv

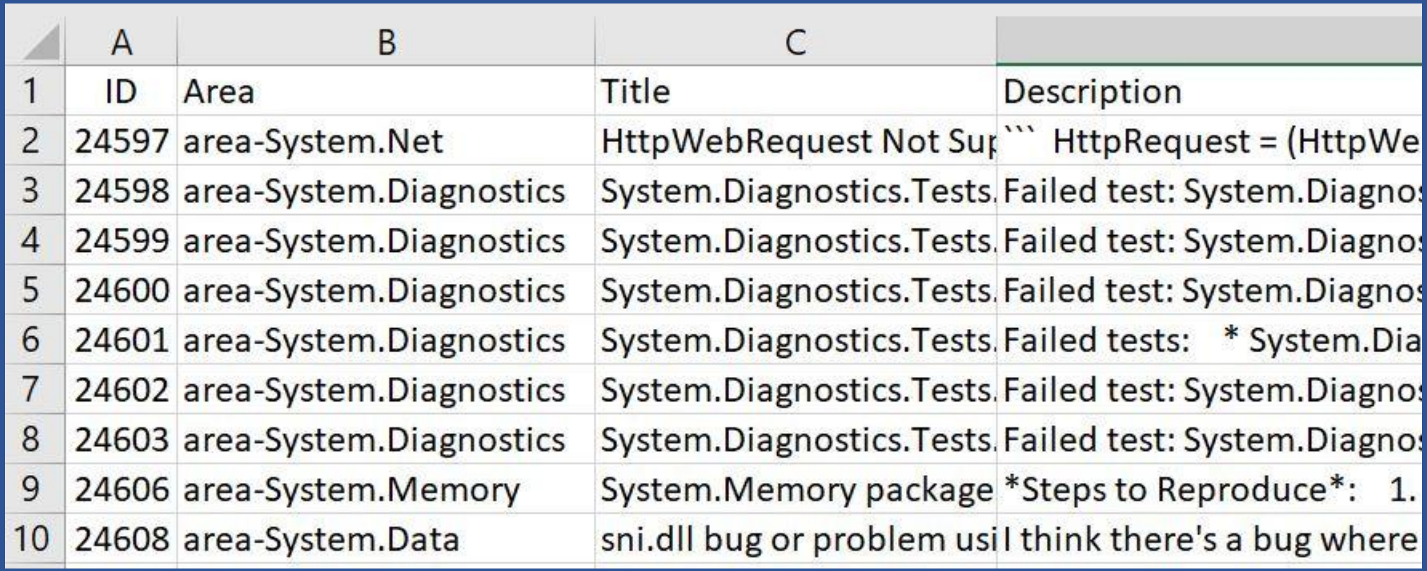
https://raw.githubusercontent.com/laploy/ML.NET/master/GitHub-Issue/issues_train.tsv

issues_test.tsv

https://raw.githubusercontent.com/laploy/ML.NET/master/GitHub-Issue/issues_test.tsv

Dataset description

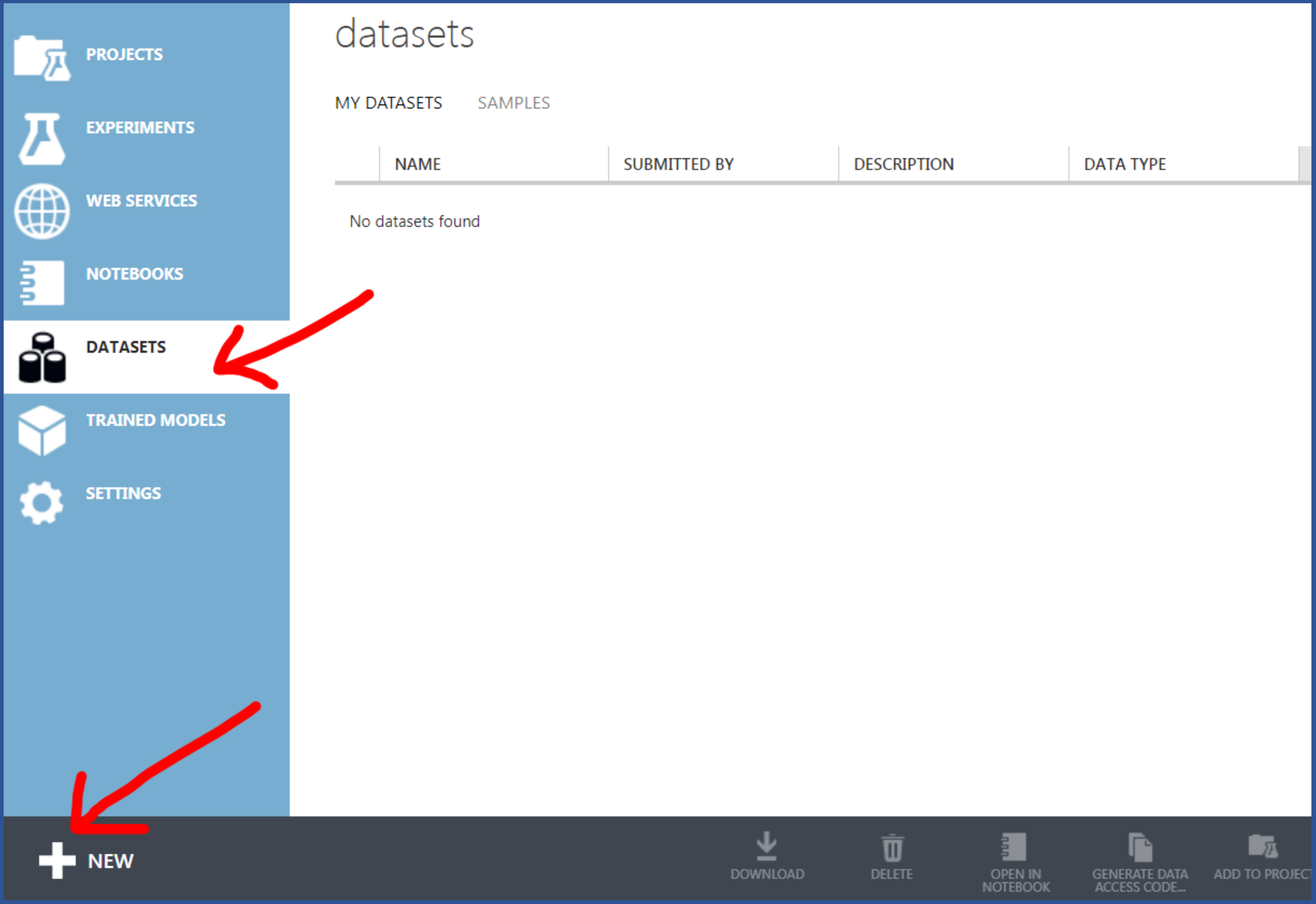
ID:	Issue Identification Number	Must be dropped
Area:	Issue area	This is the label
Title:	Issue title	This is the first feature
Description:	Issue description	This is the second feature



	A	B	C	
1	ID	Area	Title	Description
2	24597	area-System.Net	HttpWebRequest Not Sup	``` HttpRequest = (HttpWe
3	24598	area-System.Diagnostics	System.Diagnostics.Tests.	Failed test: System.Diagnos
4	24599	area-System.Diagnostics	System.Diagnostics.Tests.	Failed test: System.Diagnos
5	24600	area-System.Diagnostics	System.Diagnostics.Tests.	Failed test: System.Diagnos
6	24601	area-System.Diagnostics	System.Diagnostics.Tests.	Failed tests: * System.Dia
7	24602	area-System.Diagnostics	System.Diagnostics.Tests.	Failed test: System.Diagnos
8	24603	area-System.Diagnostics	System.Diagnostics.Tests.	Failed test: System.Diagnos
9	24606	area-System.Memory	System.Memory package	*Steps to Reproduce*: 1.
10	24608	area-System.Data	sni.dll bug or problem usi	I think there's a bug where

Create Data Set

Click DATASET / + NEW / import -> issues_train.tsv



datasets

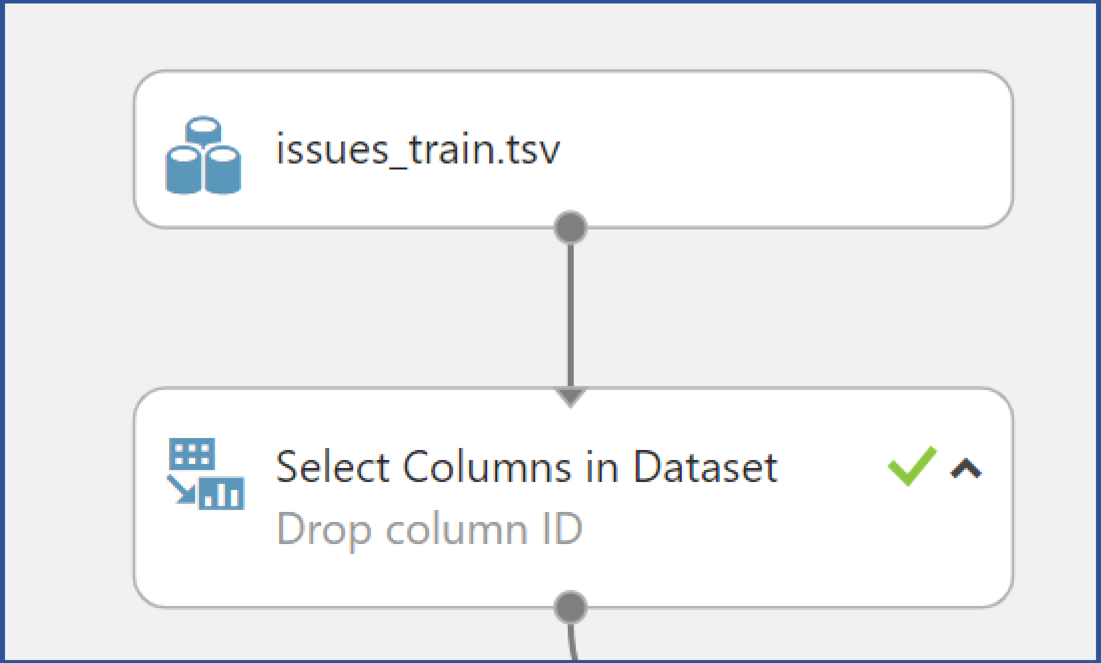
MY DATASETS SAMPLES

NAME	SUBMITTED BY	DESCRIPTION	DATA TYPE

No datasets found

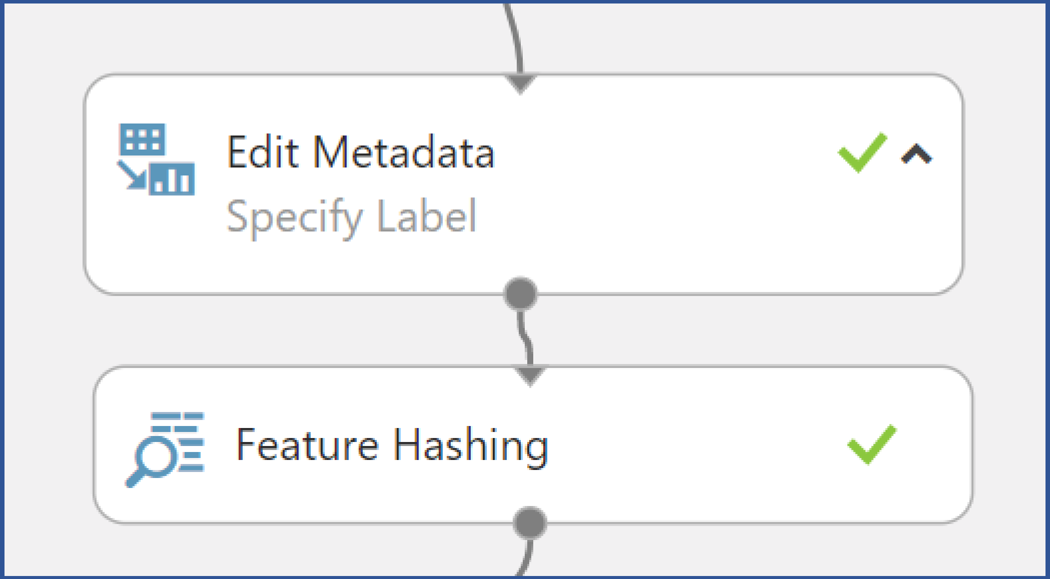
Place dataset

Drop column ID



Specify Label

Hash feature



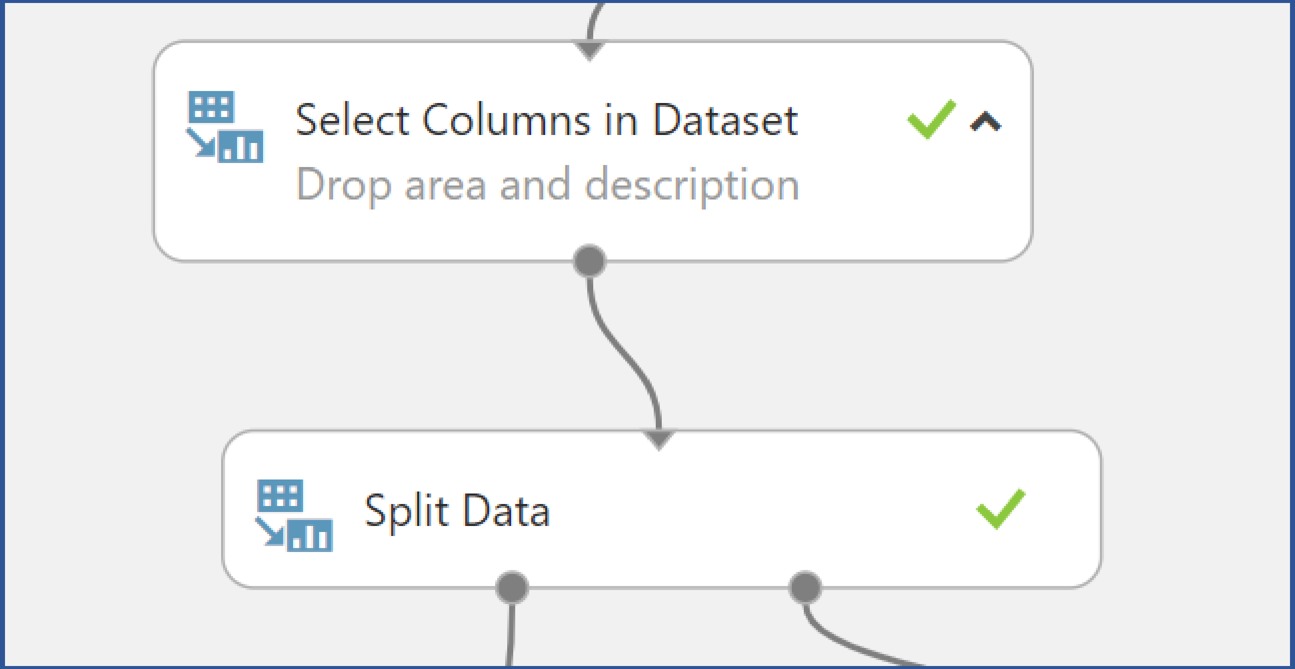
Edit Metadata ✓

Specify Label

Feature Hashing ✓

Drop area and description

Split Data

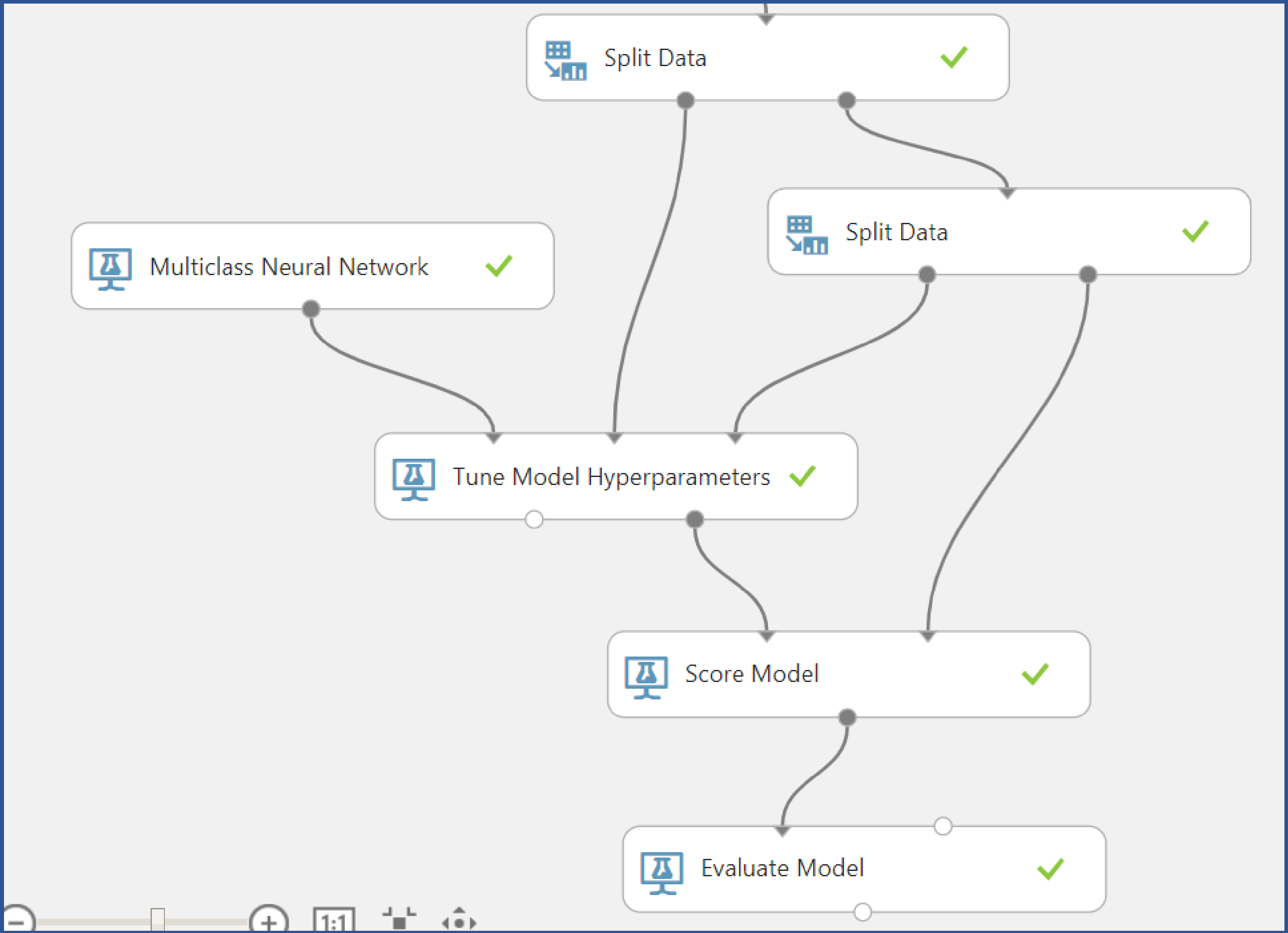


Select Columns in Dataset ✓

Drop area and description

Split Data ✓

Train, Score, Evaluate



Metrics for multiclass classification Evaluation

- Micro Accuracy - Every sample-class pair contributes equally to the accuracy metric. You want Micro Accuracy to be as close to 1 as possible.
- Macro Accuracy: Every class contributes equally to the accuracy metric. Minority classes are given equal weight as the larger classes. You want Macro Accuracy to be as close to 1 as possible.
- Log-loss: You want Log-loss to be as close to zero as possible.
- Log-loss reduction - Ranges from [-inf, 100], where 100 is perfect predictions and 0 indicates mean predictions. You want Log-loss reduction to be as close to zero as possible.

Next step

Create AutoML of GitHub issue prediction