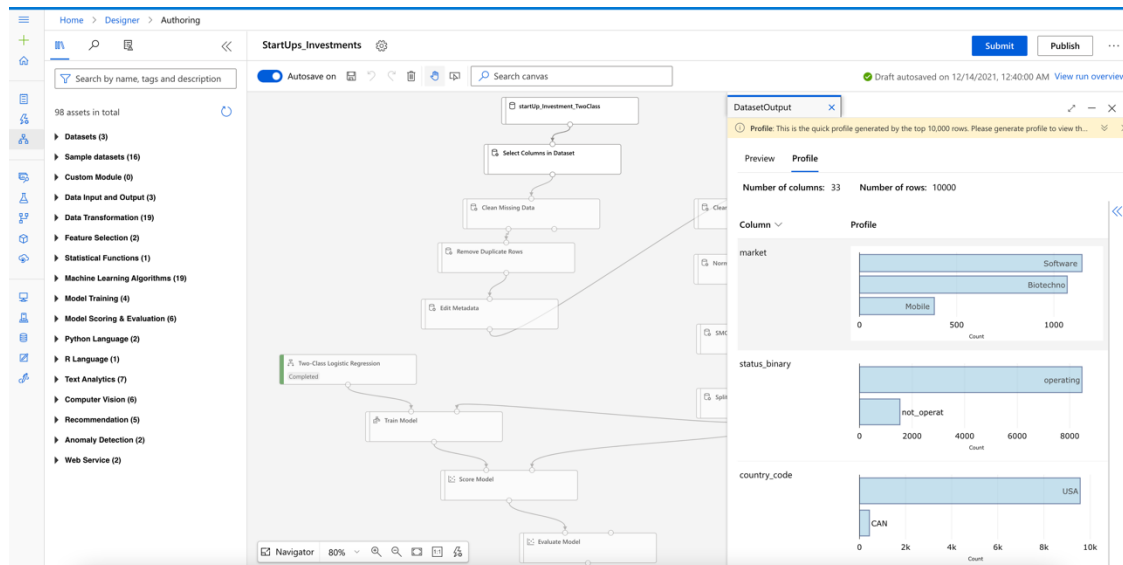


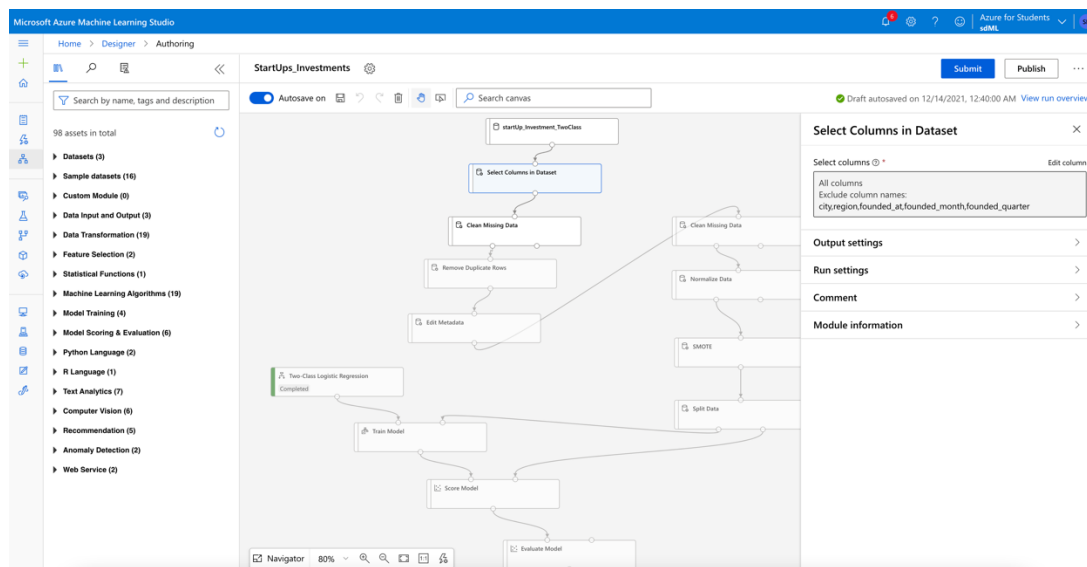
# Configurations and parameters of the Modules

## Step 1: Adding Dataset to Azureblobstorage

The status column originally has 3 classes / labels. A new column was created “status\_binary” that contains two categorical values i.e., (a)operating and (b) non-operating. ‘Operating’ values are the rows for which the original status was either “Operating” or “acquired”. The file path for the dataset added to the designer is given below.



## Step2: Select columns in the dataset(row\_count = 19958, col\_count = 28)



### Step 3: Clean Missing Data(row\_count = 19958, col\_count = 28)

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The central canvas shows a workflow for 'StartUps\_Investments' with modules including 'startUp\_Investment\_TwoClass', 'Select Columns in Dataset', 'Clean Missing Data', 'Remove Duplicate Rows', 'Edit Metadata', 'Train Model', 'Score Model', and 'Evaluate Model'. The 'Clean Missing Data' module is highlighted, and its configuration panel is open on the right. The panel settings are as follows:

- Columns to be cleaned:** All columns
- Minimum missing value ratio:** 0.0
- Maximum missing value ratio:** 1.0
- Cleaning mode:** Remove entire row
- Output settings:** (expandable)
- Run settings:** (expandable)
- Comment:** (expandable)
- Module information:** (expandable)

### Step 4: Remove Duplicate Rows(row\_count = 19907, col\_count = 28)

The screenshot displays the Microsoft Azure Machine Learning Studio interface, similar to the previous one, but with the 'Remove Duplicate Rows' module highlighted. The configuration panel for this module is open on the right, showing the following settings:

- Key column selection filter expression:** All columns
- Retain first duplicate row:** True
- Output settings:** (expandable)
- Run settings:** (expandable)
- Comment:** (expandable)
- Module information:** (expandable)

## Step 5: Edit Metadata(row\_count = 19907, col\_count = 28)

The screenshot shows the Microsoft Azure Machine Learning Studio interface. The main canvas displays a workflow for 'StartUps\_Investments'. The 'Edit Metadata' panel is open on the right, showing the following settings:

- Column: `market.country_code.state_code.funding_rounds.founded_year.status_binary`
- Data type: String
- Categorical: Categorical
- Fields: Features
- New column names: (empty)
- Output settings: (expandable)
- Run settings: (expandable)
- Comment: (expandable)
- Module information: (expandable)

## Step 5: Clean Missing Data( row\_count = 19907, col\_count = 28)

Usually should be put after “Clip Values” module for removing outliers in “SUM\_funding\_total\_usd” column.

The screenshot shows the Microsoft Azure Machine Learning Studio interface. The main canvas displays a workflow for 'StartUps\_Investments'. The 'Clean Missing Data' panel is open on the right, showing the following settings:

- Columns to be cleaned: `SUM_funding_total_usd`
- Minimum missing value ratio: 0.0
- Maximum missing value ratio: 1.0
- Cleaning mode: Remove entire row
- Output settings: (expandable)
- Run settings: (expandable)
- Comment: (expandable)
- Module information: (expandable)

## Step 6: Normalize Data (row\_count = 19907, col\_count = 28)

The screenshot shows the Microsoft Azure Machine Learning Studio interface. On the left, a sidebar lists various assets and datasets. The main canvas displays a workflow for 'StartUps\_Investments'. The workflow includes steps such as 'Select Columns in Dataset', 'Clean Missing Data', 'Remove Duplicate Rows', 'Edit Metadata', 'Normalize Data', 'SMOTE', 'Split Data', 'Train Model', 'Score Model', and 'Evaluate Model'. The 'Normalize Data' step is highlighted, and its configuration panel is open on the right. The configuration panel shows the 'Transformation method' set to 'MinMax', 'Use 0 for constant columns when checked' set to 'True', and 'Columns to transform' set to 'All columns'. The 'Exclude column names' list includes 'market\_code', 'country\_code', 'state\_code', 'funding\_rounds', 'founded\_year', 'status', and 'binary'.

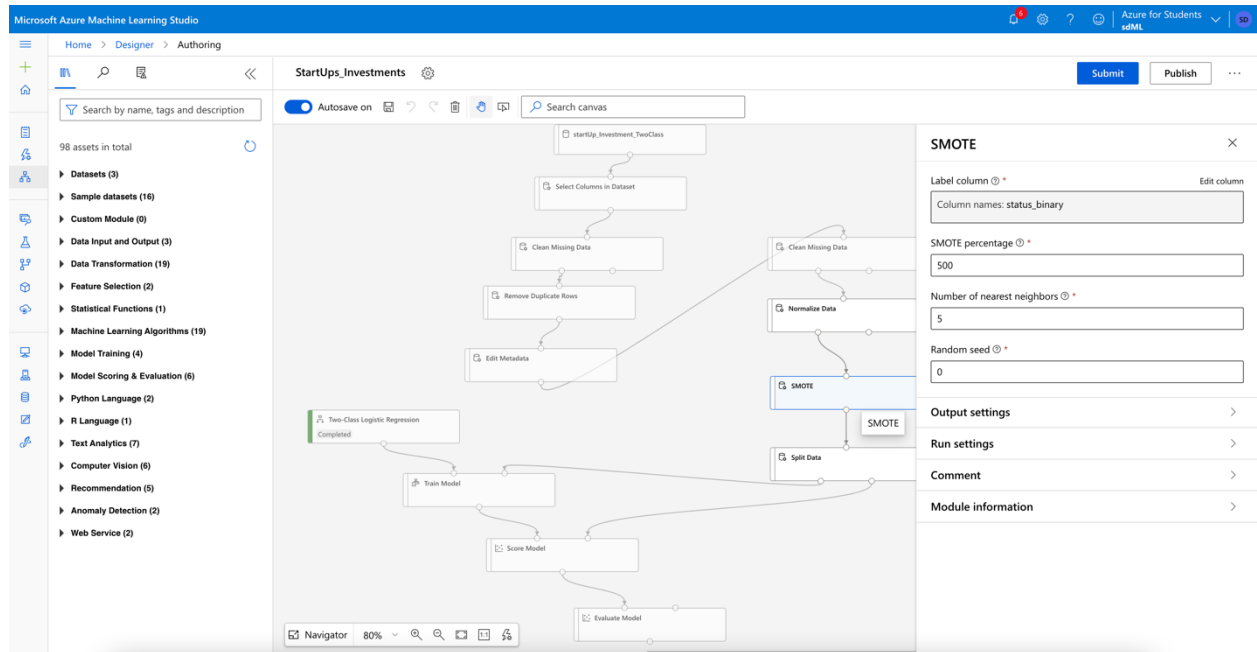
## Transformed Dataset

The screenshot shows the Microsoft Azure Machine Learning Studio interface. The main canvas displays a workflow for 'heroic\_thread\_m7df2j0t'. The workflow includes steps such as 'Select Columns in Dataset', 'Clean Missing Data', 'Remove Duplicate Rows', 'Edit Metadata', 'Train Model', 'Score Model', and 'Evaluate Model'. The 'Train Model' step is highlighted, and its configuration panel is open on the right. The configuration panel shows the 'Transformation method' set to 'MinMax', 'Use 0 for constant columns when checked' set to 'True', and 'Columns to transform' set to 'All columns'. The 'Exclude column names' list includes 'market\_code', 'country\_code', 'state\_code', 'funding\_rounds', 'founded\_year', 'status', and 'binary'.

round_D	round_E	round_F	round_G	round_H	SUM_funding_total_usd
0	0	0	0	0	0.000058
0	0	0	0	0	0.000002
0	0	0	0	0	0.000058
0	0	0	0	0	0.000068
0	0	0	0	0	0.000001
0	0	0	0	0	0.000168
0	0	0	0	0	0.000165
0	0	0	0	0	0.000014
0	0	0	0	0	0.000042
0	0	0	0	0	0.001164
0	0	0	0	0	0.000002
0	0	0	0	0	0.0001
0	0	0	0	0	0.000003
0	0	0	0	0	0.000026
0	0	0	0	0	0.000147

Step 7: SMOTE for handling imbalanced classes ( **row\_count = 35667**, **col\_count = 28**)

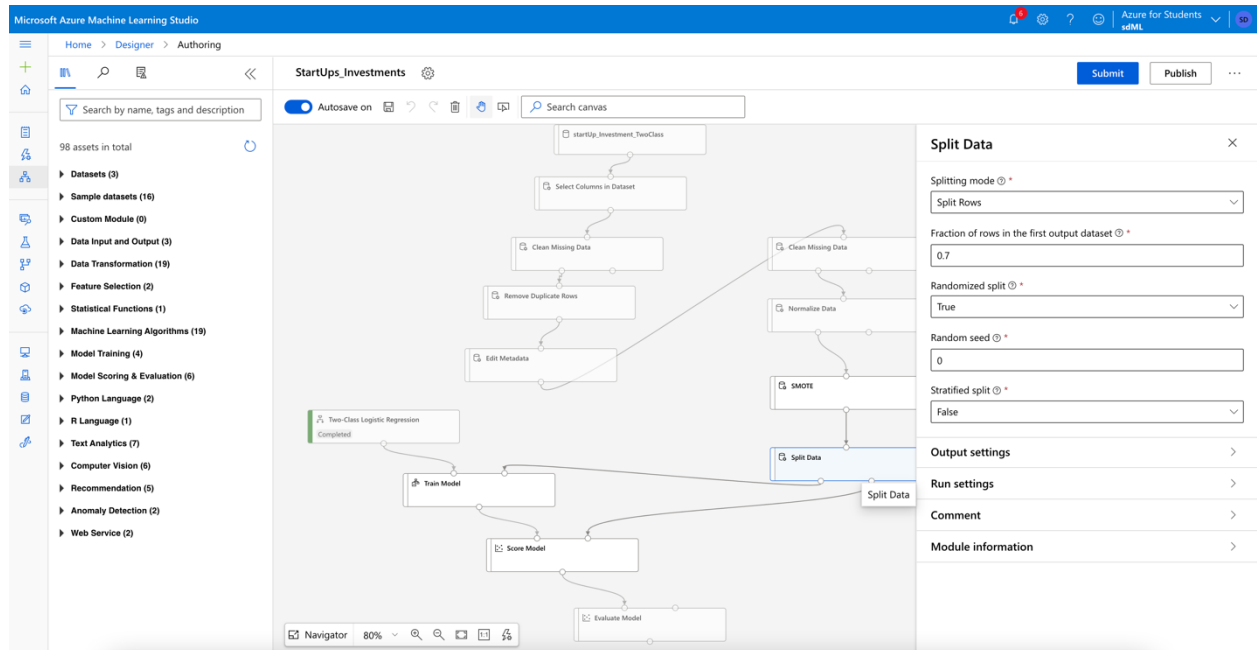
**SMOTE** – Synthetic Minority Oversampling Technique is a type of data augmentation for the minority class. Using this new examples can be synthesized from the existing examples for balancing the data.



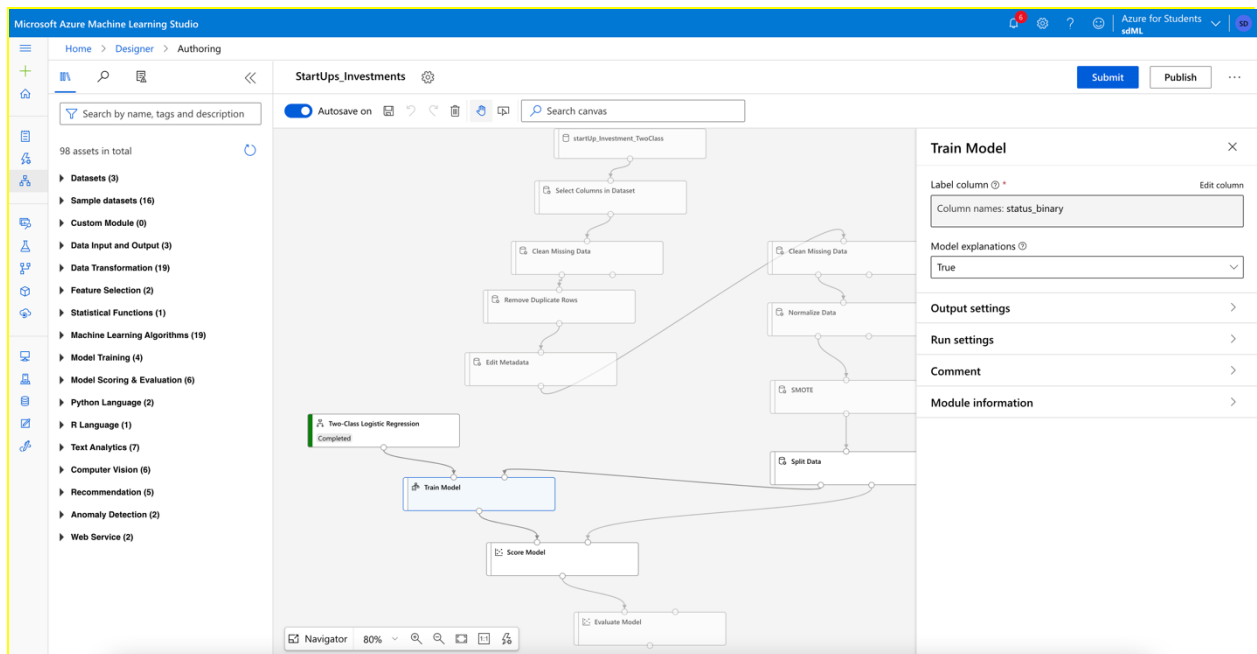
Step 8: Split Data

(Result Dataset1: ( **row\_count = 24967**, **col\_count = 28**),

Result Dataset2: ( **row\_count = 10700**, **col\_count = 28**)

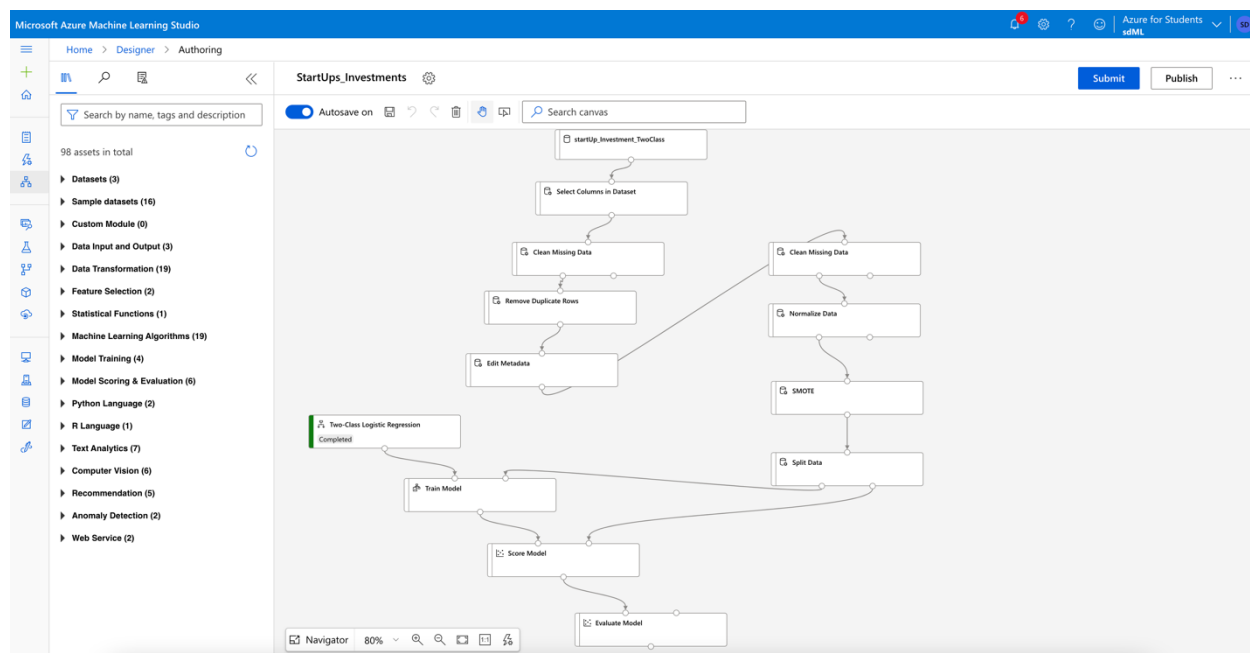


## Step 9: Train Model



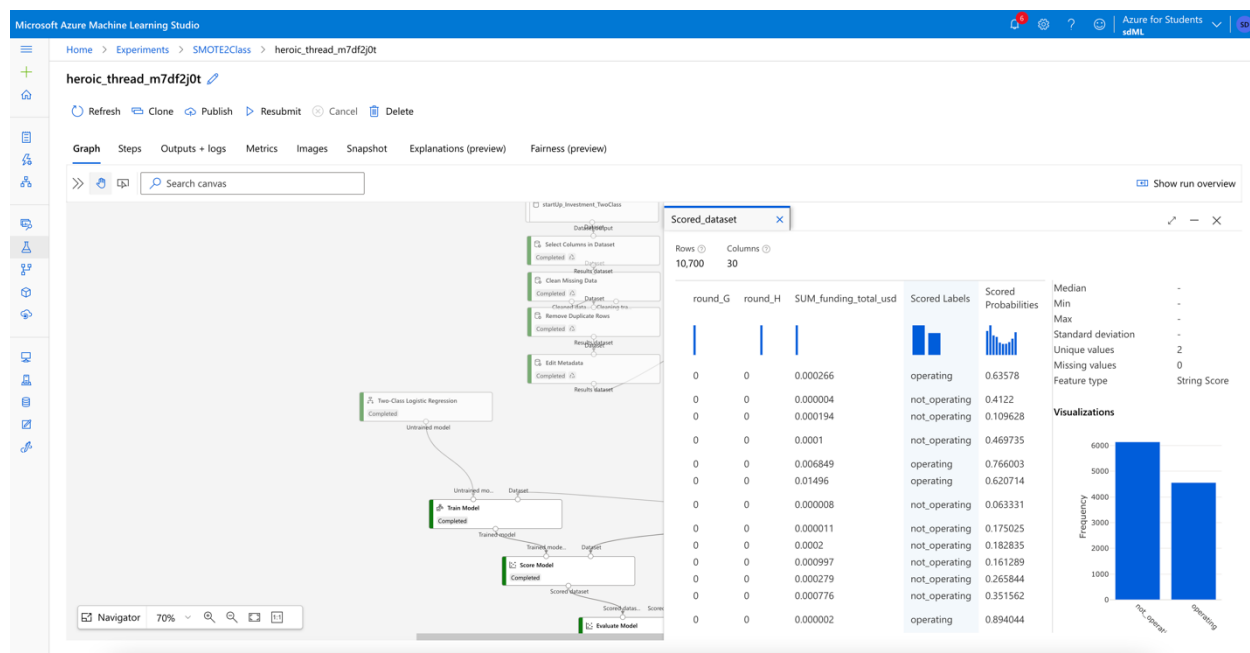
Step 10: Score Model & Evaluate Model (**row\_count = 10700, col\_count = 30**)  
**Not-operating : 6141, Operating: 4559**

**Screenshot of the complete Pipeline Execution**



## Results

### Scored Label Statistics (Distribution of the two-classes)



## Evaluation Metrics

