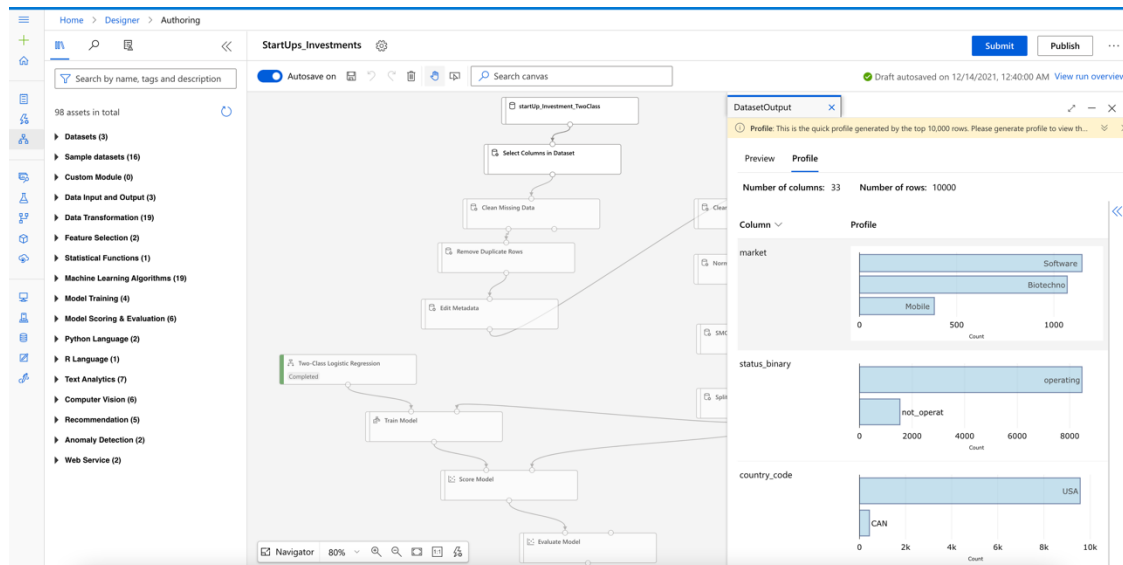


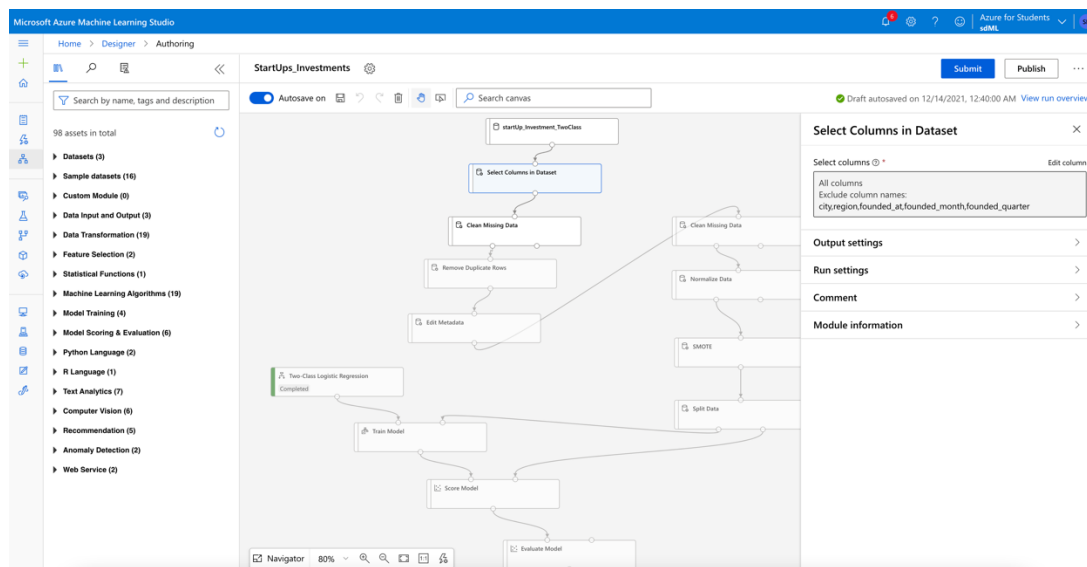
Configurations and parameters of the Modules

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 starting with 'startUp_Investment_TwoClass', followed by '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 configuration panel includes the following settings:

- 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. The central canvas shows a workflow starting with 'startUp_Investment_TwoClass', followed by 'Select Columns in Dataset', 'Clean Missing Data', 'Remove Duplicate Rows', 'Edit Metadata', 'Train Model', 'Score Model', and 'Evaluate Model'. The 'Remove Duplicate Rows' module is highlighted, and its configuration panel is open on the right. The configuration panel includes 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 central canvas displays a workflow for the 'StartUps_Investments' dataset. The 'Edit Metadata' panel is open on the right, showing the following configuration:

- 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 central canvas displays a workflow for the 'StartUps_Investments' dataset. The 'Clean Missing Data' panel is open on the right, showing the following configuration:

- 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. The main canvas displays a workflow for data normalization. The workflow starts with 'startups_investment_twoClass', followed by '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 settings are shown on the right.

Normalize Data

Transformation method: MinMax

Use 0 for constant columns when checked: True

Columns to transform: All columns
Exclude column names: market_code, state_code, funding_rounds, founded_year, status_binary

Output settings: >

Run settings: >

Comment: >

Module information: >

Transformed Dataset

The screenshot shows the Microsoft Azure Machine Learning Studio interface. The main canvas displays a workflow for data normalization. The workflow starts with 'startups_investment_twoClass', followed by '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 'Transformed_dataset' output is shown, and its statistics are displayed on the right.

Transformed_dataset

Rows: 19,907 Columns: 28

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

Statistics

Statistic	Value
Mean	0.001
Median	0.0001
Min	0
Max	1
Standard deviation	0.0079
Unique values	7656
Missing values	0
Feature type	Numeric F

Visualizations

Frequency

Step 7: SMOTE for handling imbalanced classes(**row_count = 35667, col_count = 28**)

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The central canvas shows a workflow for the 'StartUps_Investments' dataset. The workflow includes modules for 'startUp_Investment_TwoClass', 'Select Columns in Dataset', 'Clean Missing Data', 'Remove Duplicate Rows', 'Edit Metadata', 'SMOTE', 'Split Data', 'Train Model', 'Score Model', and 'Evaluate Model'. The 'SMOTE' module is highlighted, and its configuration panel is open on the right. The configuration panel shows the following settings:

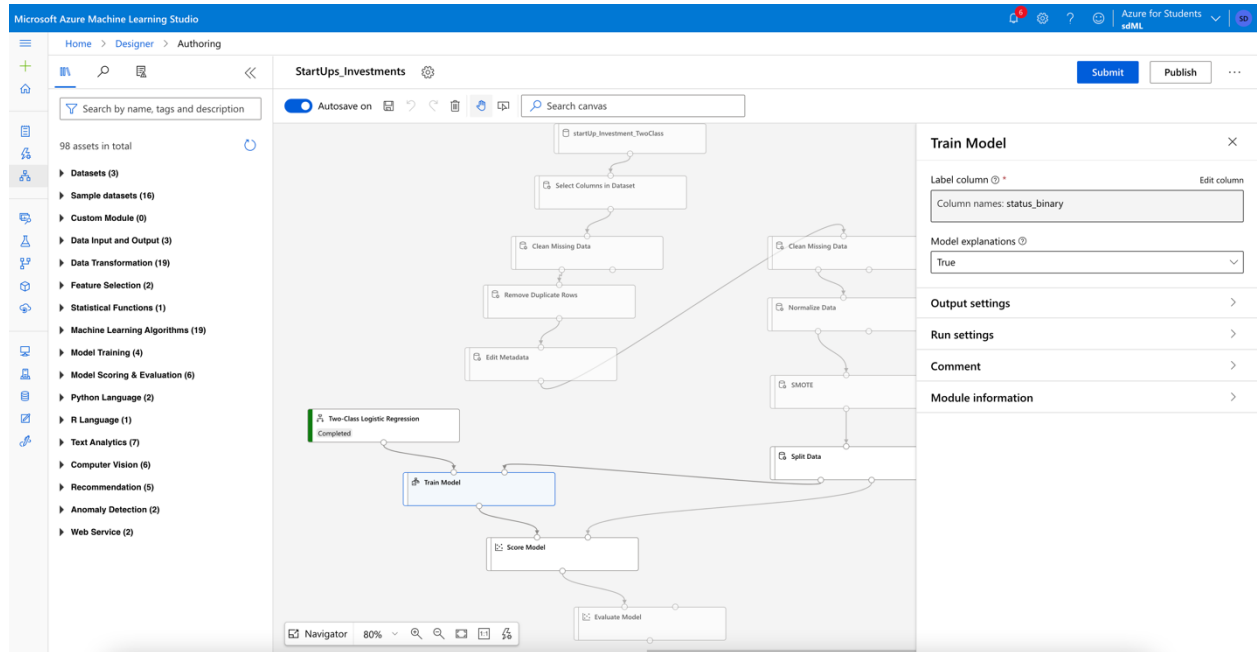
- Label column: status_binary
- SMOTE percentage: 500
- Number of nearest neighbors: 5
- Random seed: 0
- Output settings: >
- Run settings: >
- Comment: >
- Module information: >

Step 8: Split Data
(Result Dataset1: (**row_count = 24967, col_count = 28**),
Result Dataset2: (**row_count = 10700, col_count = 28**)

The screenshot displays the Microsoft Azure Machine Learning Studio interface, showing the same workflow as the previous step. The 'Split Data' module is highlighted, and its configuration panel is open on the right. The configuration panel shows the following settings:

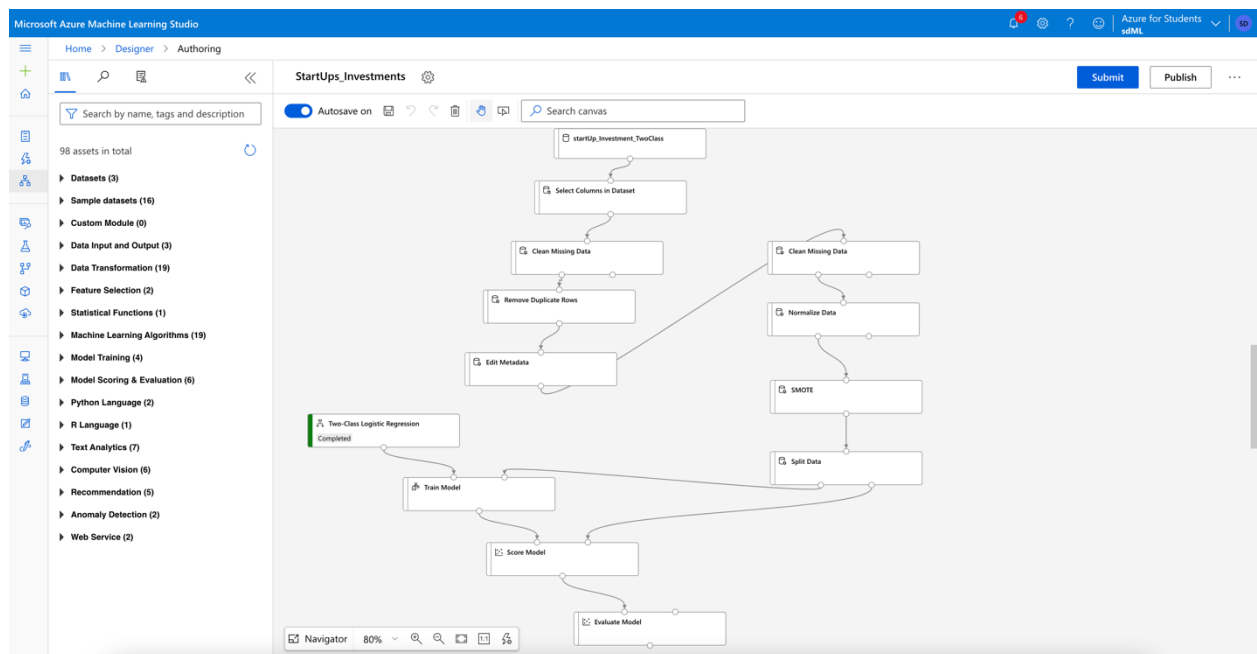
- Splitting mode: Split Rows
- Fraction of rows in the first output dataset: 0.7
- Randomized split: True
- Random seed: 0
- Stratified split: False
- Output settings: >
- Run settings: >
- Comment: >
- Module information: >

Step 9: Train Model



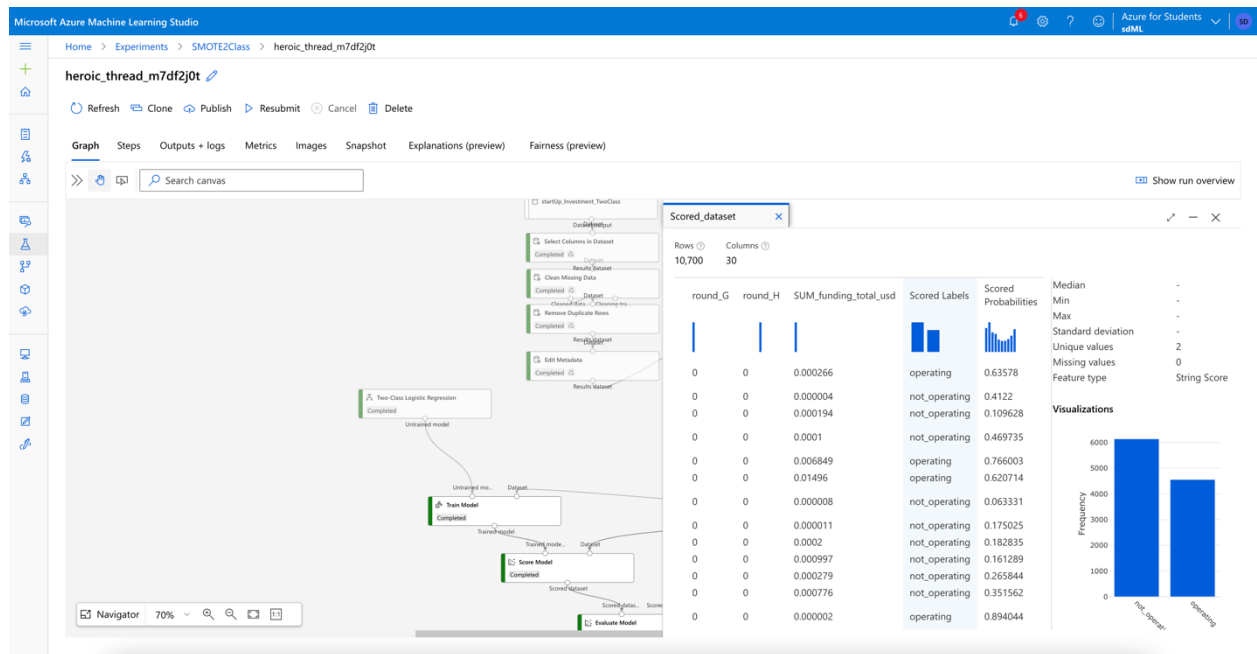
Step 10: Score Model & Evaluate Model (**row_count = 10700, col_count = 30**)
Not-operating : 6141, Operating: 4559

Screenshot of the complete Pipeline Execution

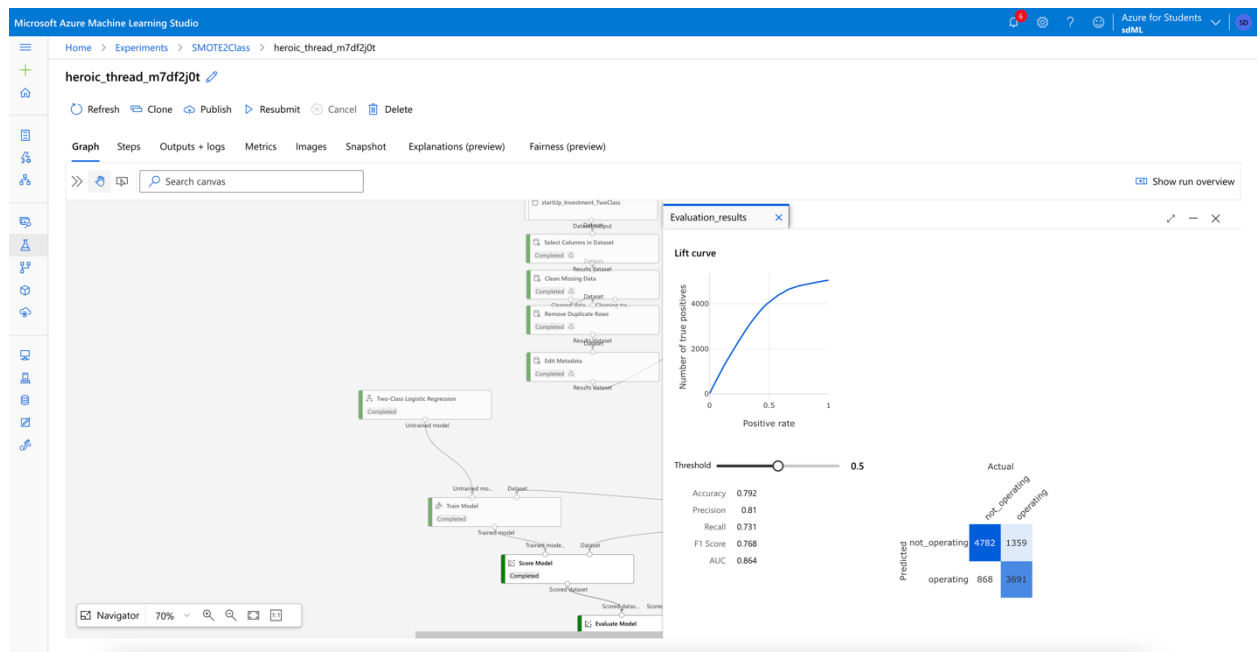


Results

Scored Label Statistics (Distribution of the two-classes)



Evaluation Metrics



Microsoft Azure Machine Learning Studio

Home > Experiments > SMOTE2Class > heroic_thread_m7df2j0t

heroic_thread_m7df2j0t

Refresh Clone Publish Resubmit Cancel Delete

Graph Steps Outputs + logs Metrics Images Snapshot Explanations (preview) Fairness (preview)

>> Search canvas

Show run overview

Identity, Investment, TwoClass

Dataset

Completed Select Columns in Dataset

Completed Results Dataset

Completed Clean Missing Data

Completed Dataset

Completed Remove All Rows

Completed Remove Duplicate Rows

Completed Partition Dataset

Completed Split Metadata

Completed Results Dataset

Completed Two-Class Logistic Regression

Untrained model

Untrained model

Completed Train Model

Completed Trained model

Completed Training model

Completed Score Model

Completed Scored dataset

Scored dataset

Scored dataset

Completed Evaluate Model

Evaluation_results

Scored dataset (left port)

ROC curve

True positive rate

False positive rate

Precision-recall curve

Precision

Recall

Lift curve

Number of true positives

Navigator 70%