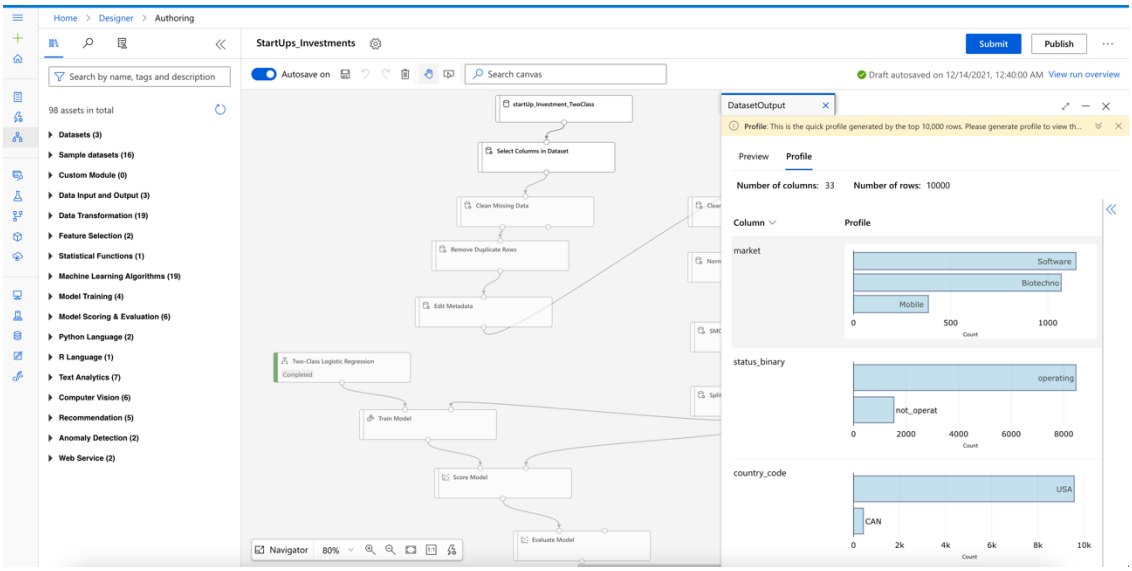
Configurations and parameters of the Modules

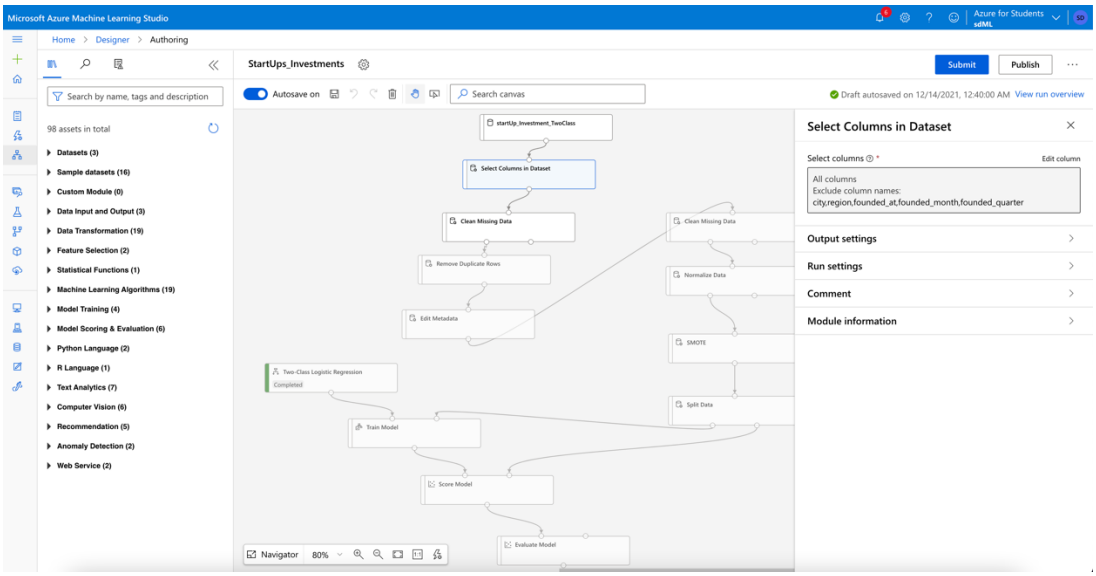
Step 1: Adding Dataset to Azureblobstorage

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”.

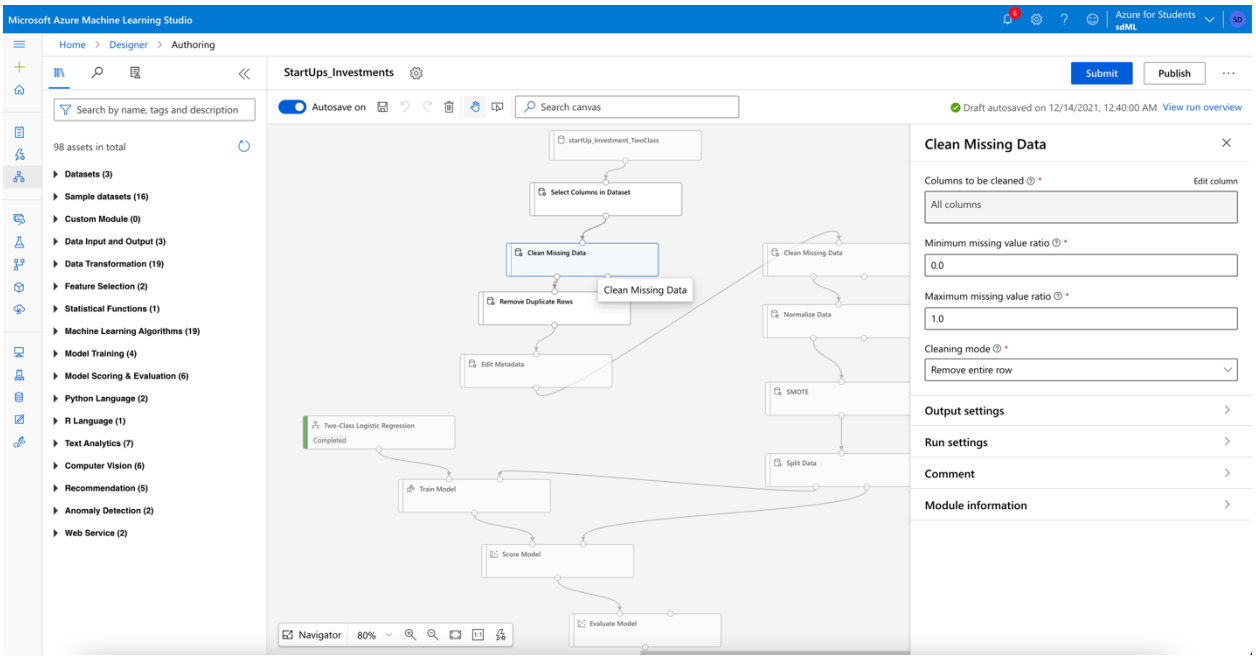
The file path for the dataset added to the designer is given below.



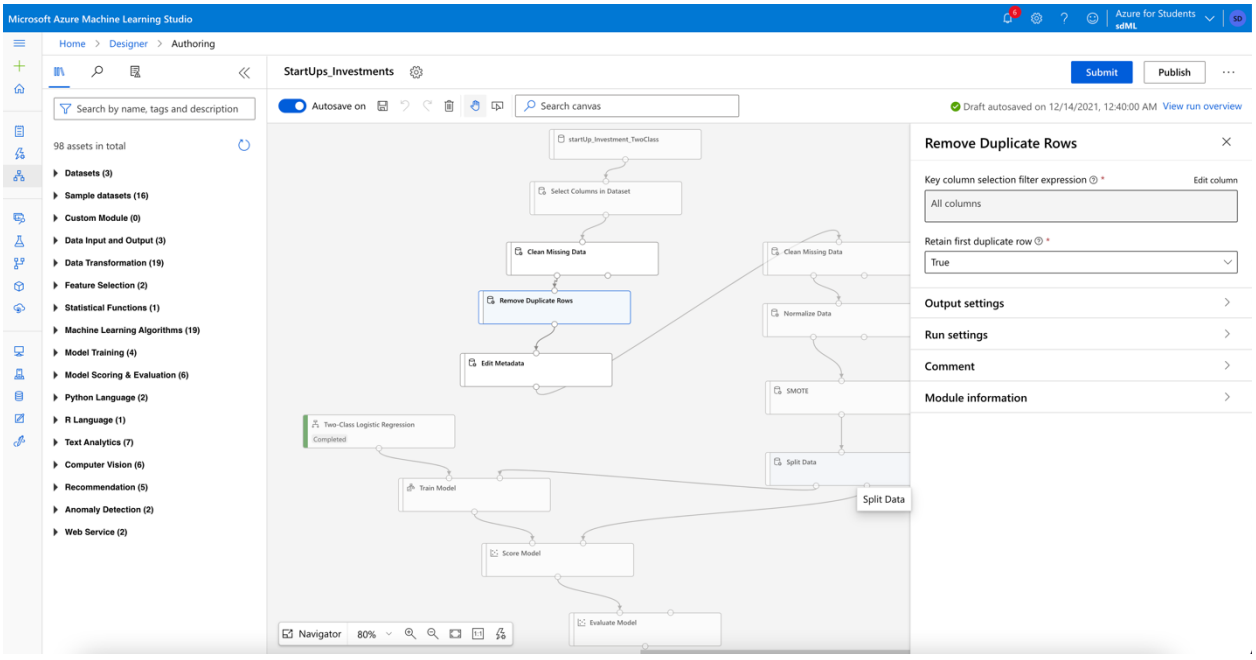
Step2: Select columns in the dataset(row_count = 19958, col_count = 28)



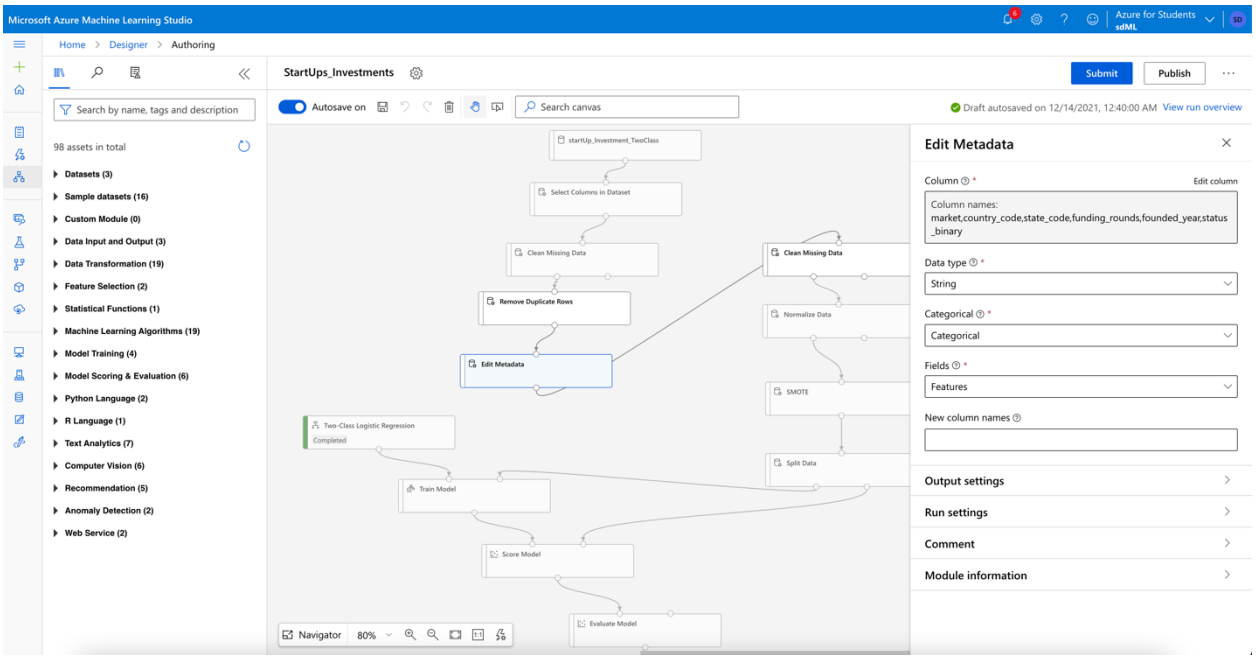
Step 3: Clean Missing Data**(row_count = 19958, col_count = 28)**



Step 4: Remove Duplicate Rows**(row_count = 19907, col_count = 28)**

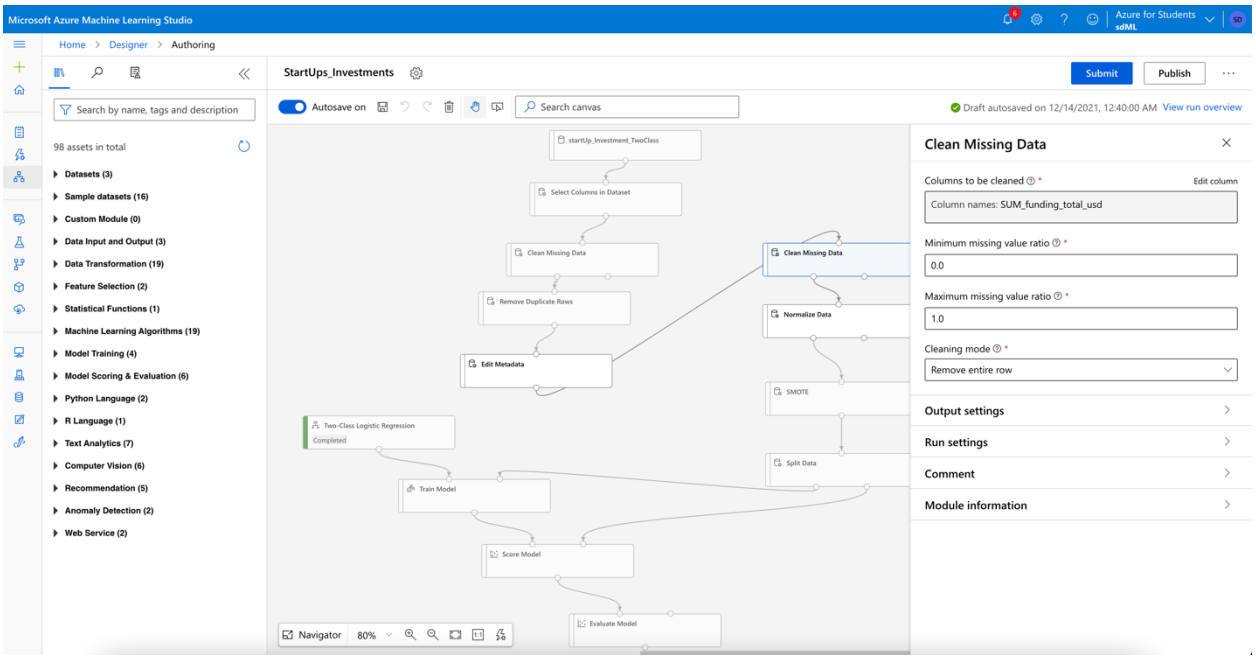


Step 5: Edit Metadata**(row_count = 19907, col_count = 28)**

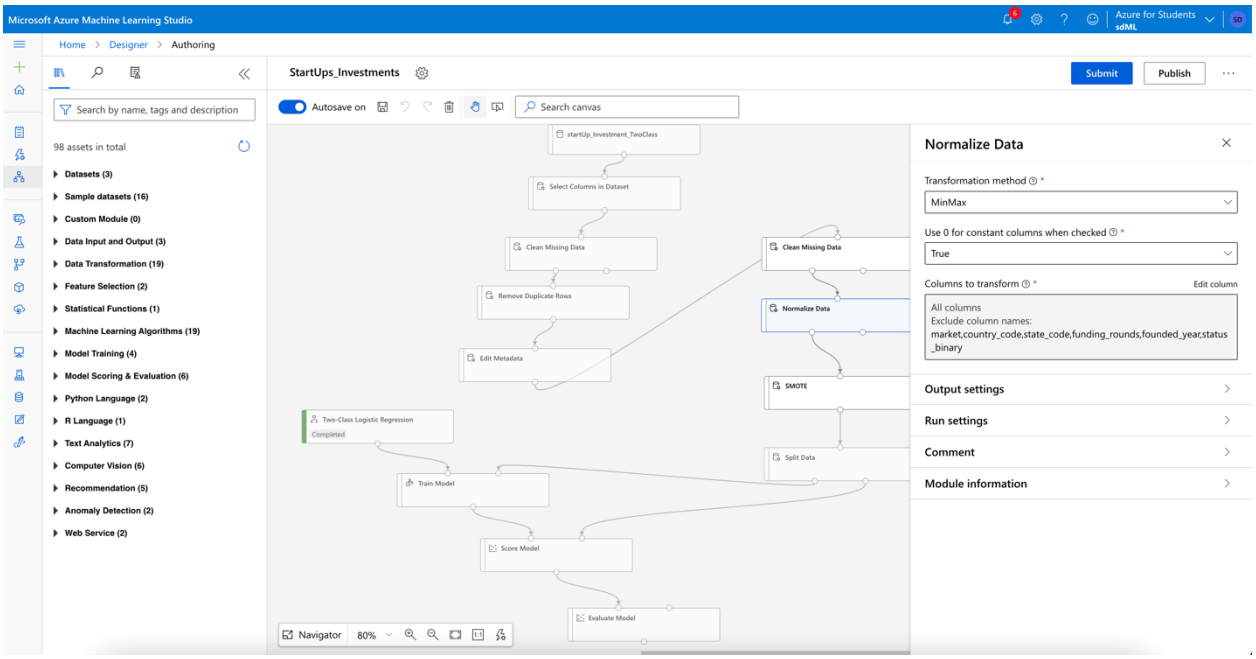


Step 5: Clean Missing Data(row_count = 19907, col_count = 28)

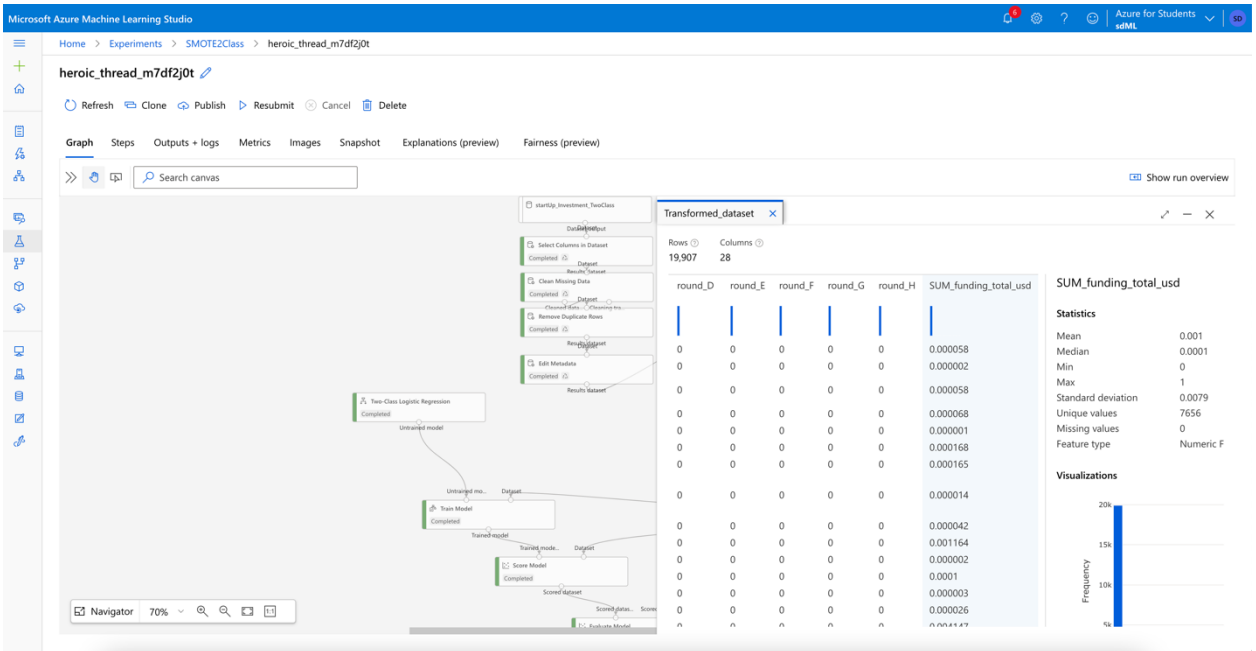
Usually should be put after “Clip Values” module for removing outliers in “SUM_funding_total_usd” column.



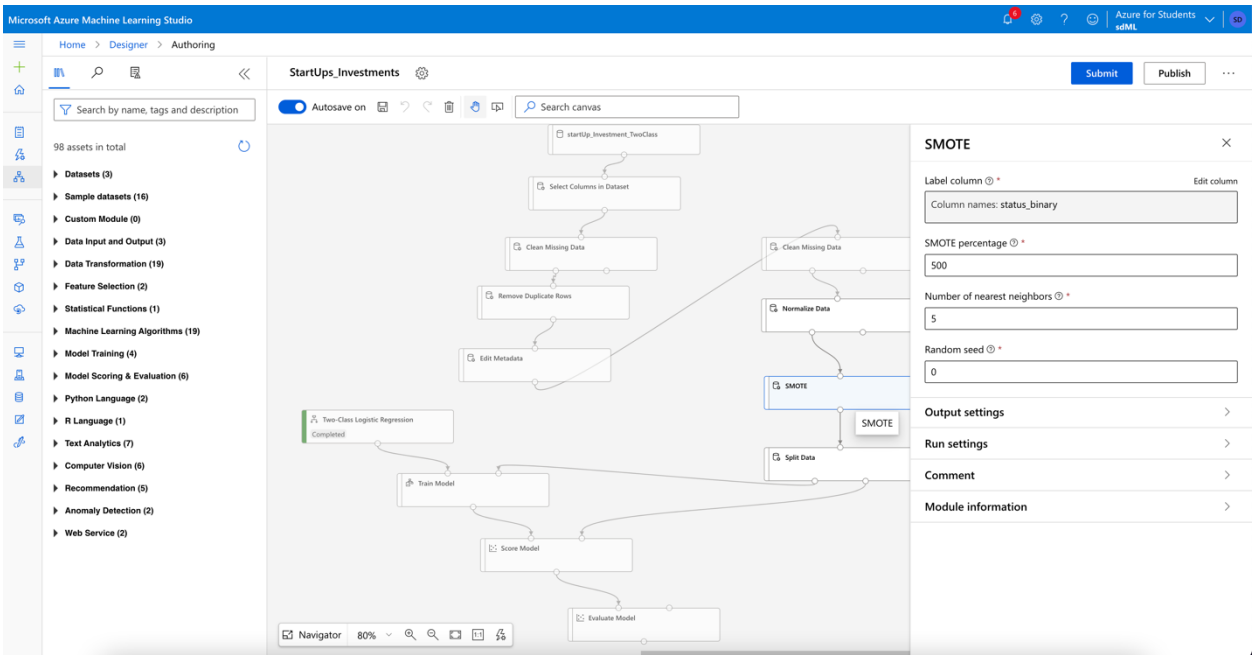
Step 6: Normalize Data(**row_count = 19907, col_count = 28)**



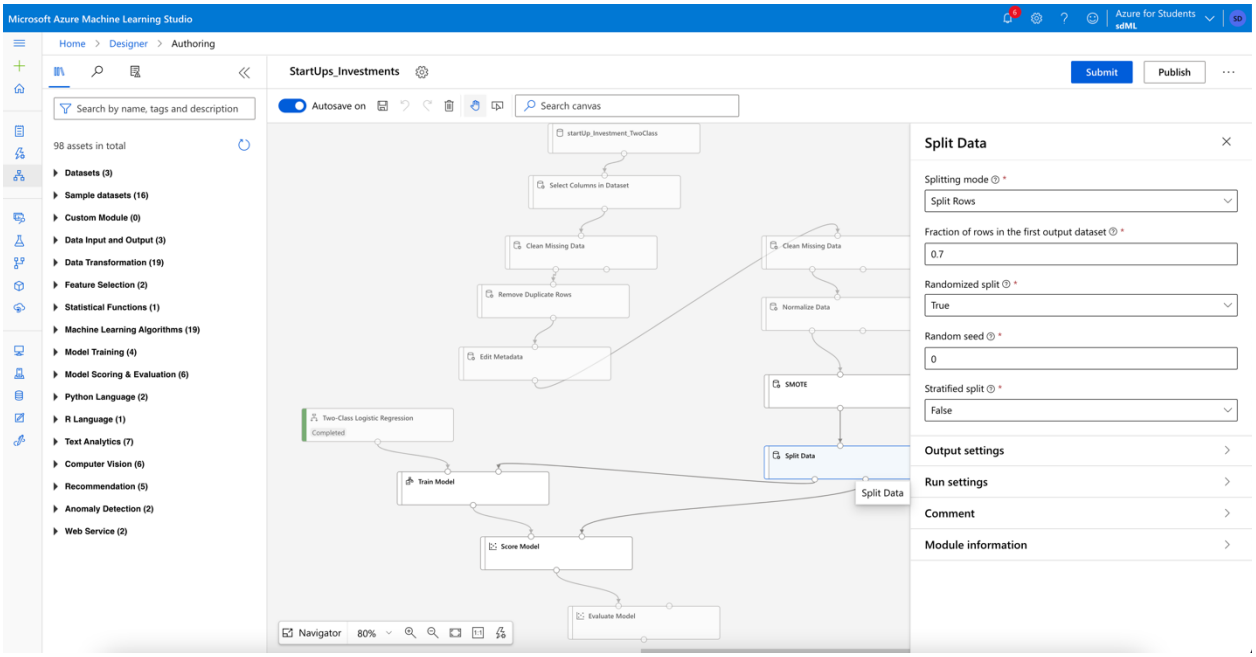
Transformed Dataset



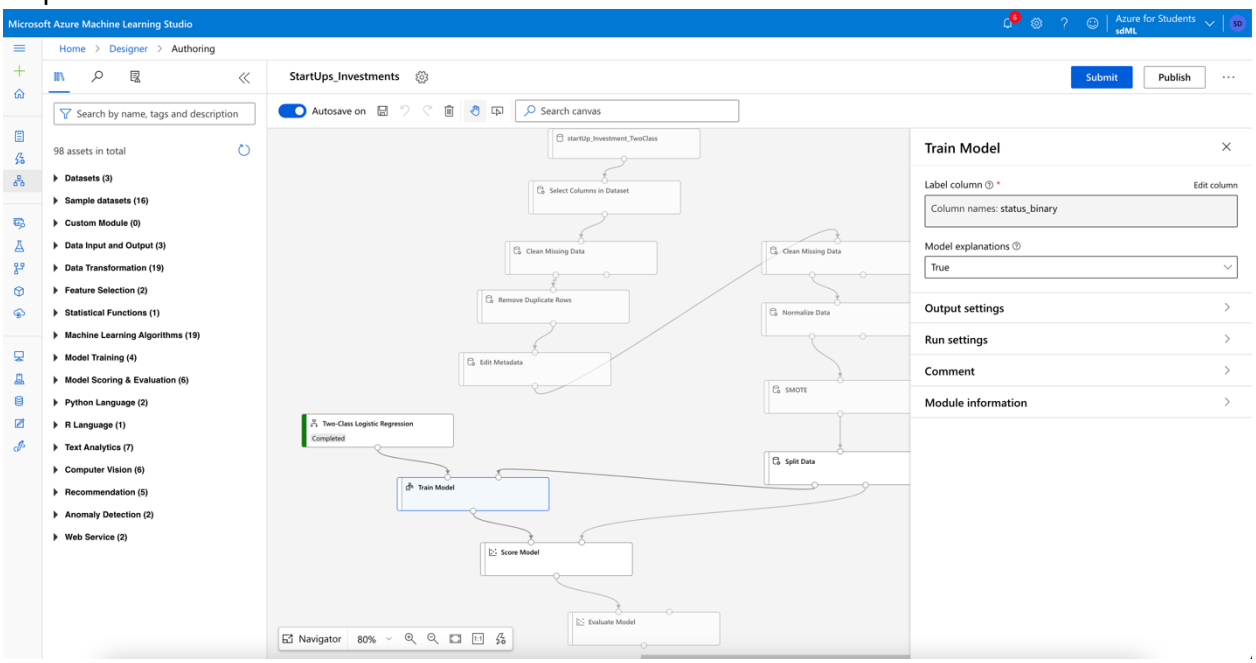
Step 7: SMOTE for handling imbalanced classes(**row_count = 35667, col_count = 28)**



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

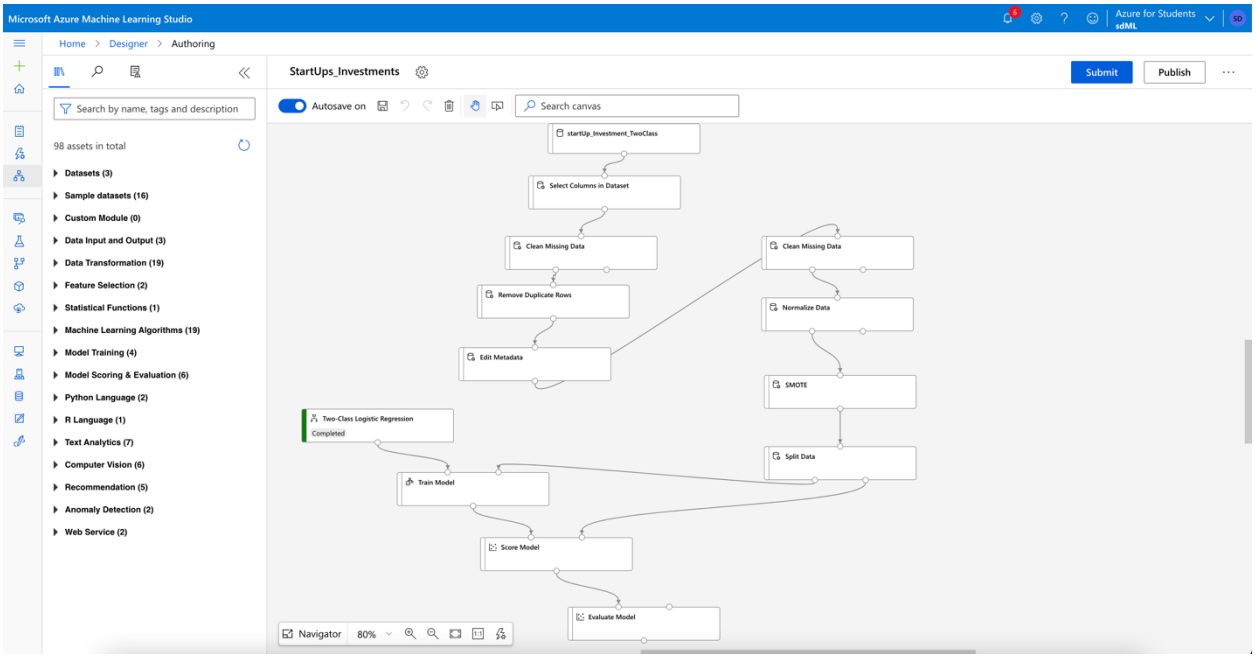


Step 9: Train Model



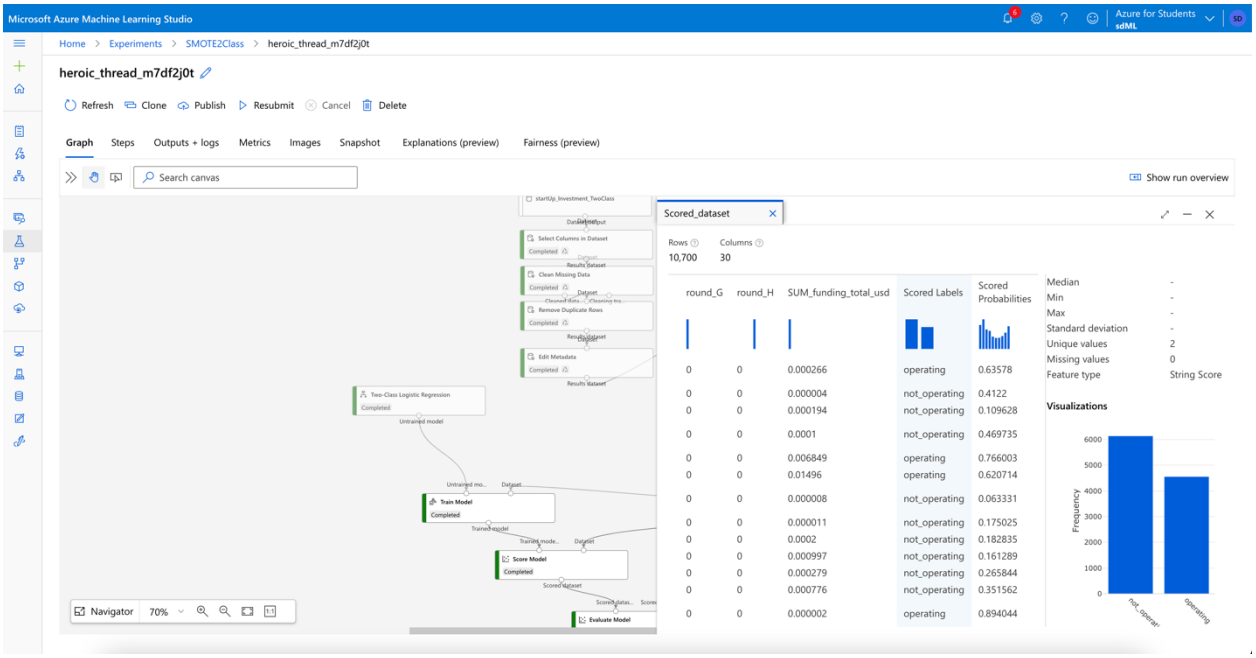
Step 10: Score Model & Evaluate Model **(row_count = 10700, col_count = 30)**
Not-operating : 6141, Operating: 4559

Screenshot of the complete Pipeline Execution

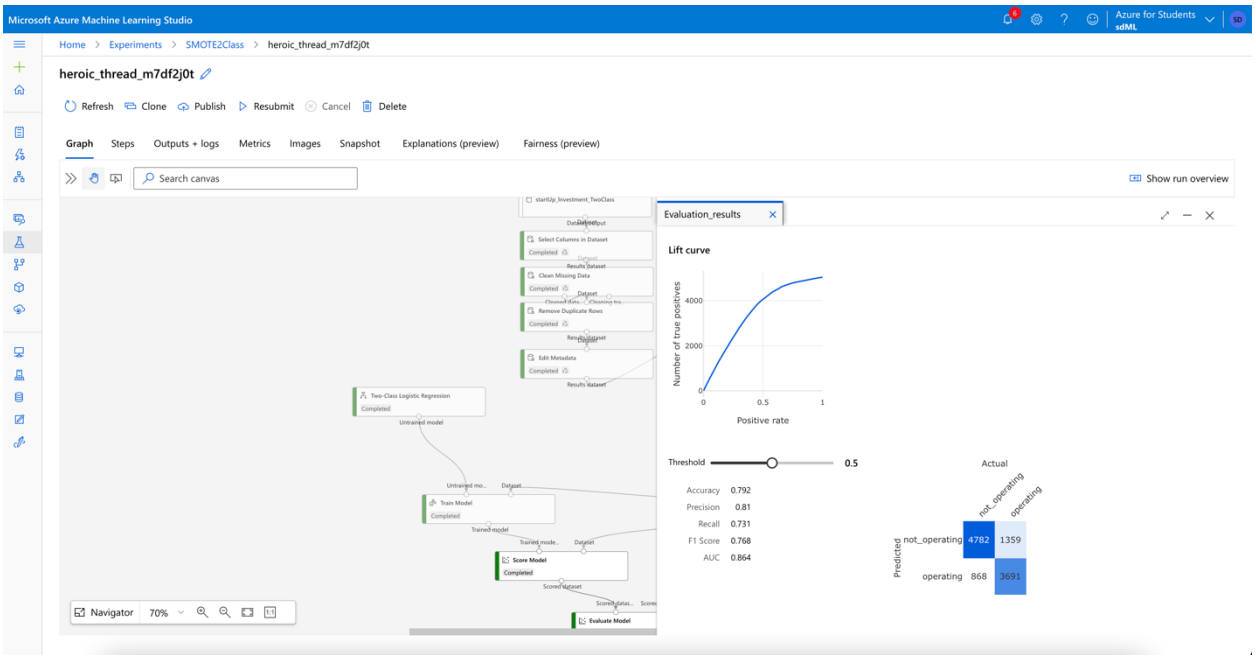


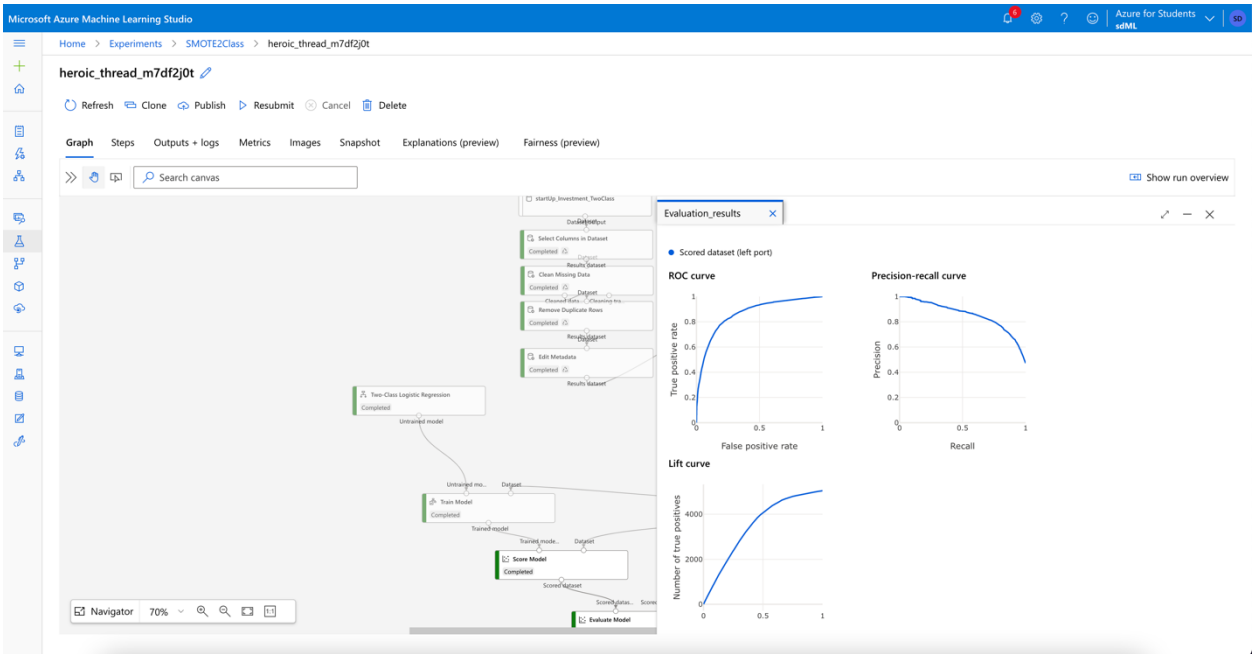
Results

Scored Label Statistics (Distribution of the two-classes)



Evaluation Metrics




Microsoft Azure Machine Learning Studio
Azure for Students
sdML
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
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%
Select Columns in Dataset
Clean Missing Data
Remove Duplicate Rows
Edit Metadata
Two-Class Logistic Regression
Train Model
Score Model
Evaluate Model