Microsoft

Two Class Logistic Regression with Azure ML

Contents

Table of Contents

Goals and Requirements

Estimated time to complete lab is 40-45 minutes

Goals

1. Implement and design a model for automating loan approval.

2. Approach of using Two Class Logistic Regression

Requirements:

1. Access to an Azure Machine Learning Studio

Logistic Regression

Problem Statement

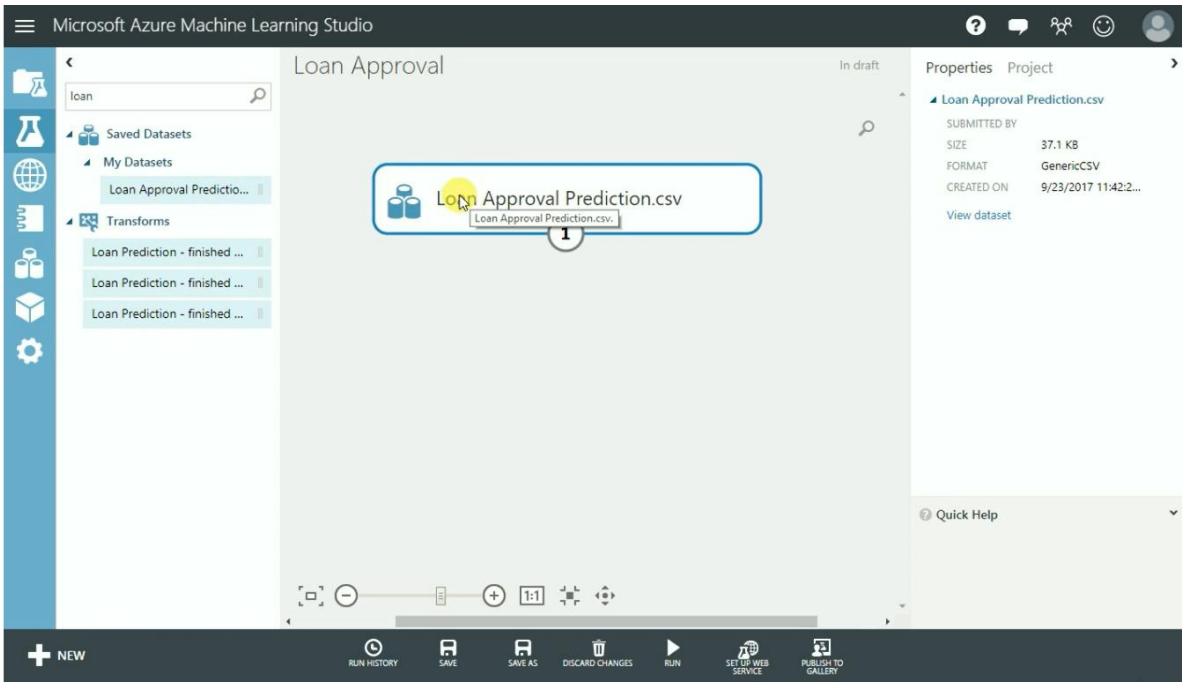
Problem Statement

- Automate loan eligibility process
- Identify customers whose loan will be approved

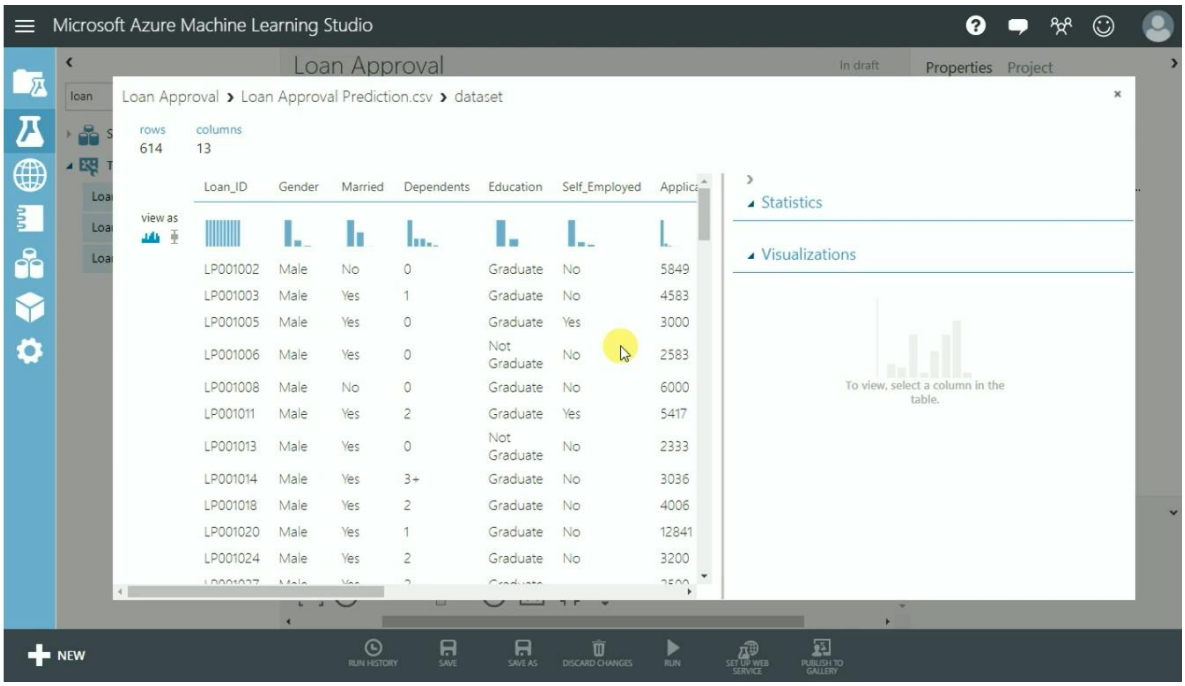
Loan_ID	Gender	Married	Dependents	Self_Employed	Income	LoanAmt	Term	CreditHistory	Property_Area	Status
LP001002	Male	No	0	No	$5,849.00		60	1	Urban	Y
LP001003	Male	Yes	1	No	$4,583.00	$128.00	120	1	Rural	N
LP001005	Male	Yes	0	Yes	$3,000.00	$66.00	60	1	Urban	Y
LP001006	Male	Yes	2	No	$2,583.00	$120.00	60	1	Urban	Y

Dataset for Loan Approval

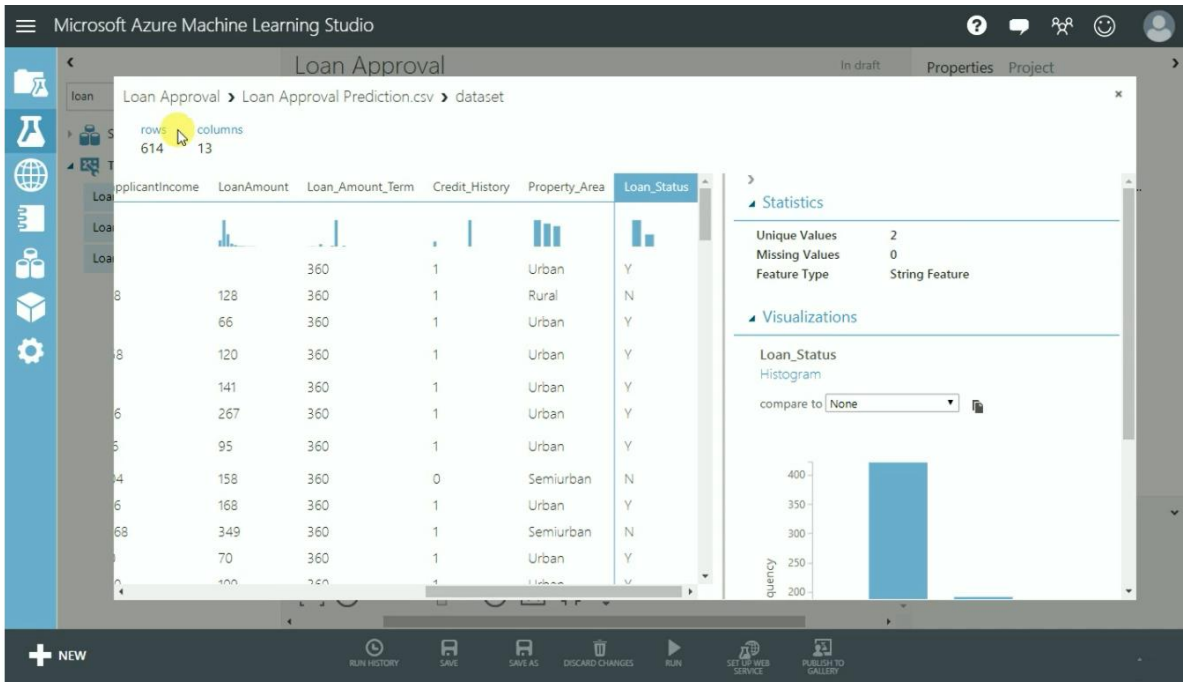
Place the dataset in canvas



Visualize the data that we need to deal which involved loan application

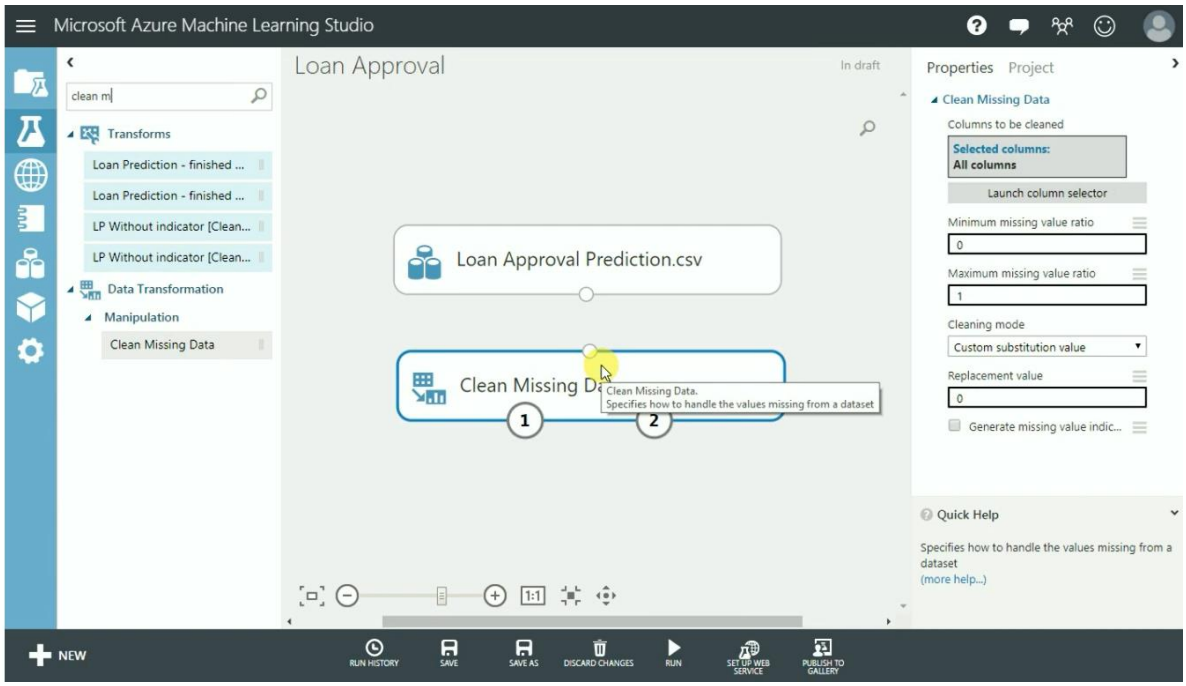


This dataset got 614 rows and 13 columns

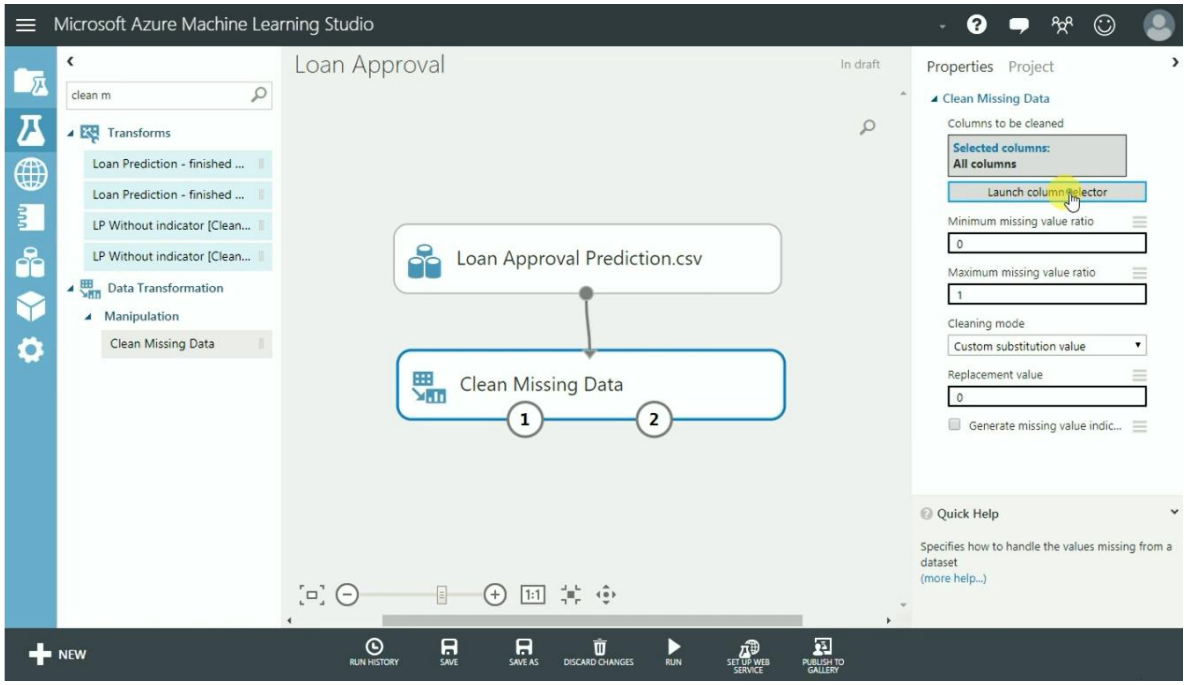


Cleaning the Data

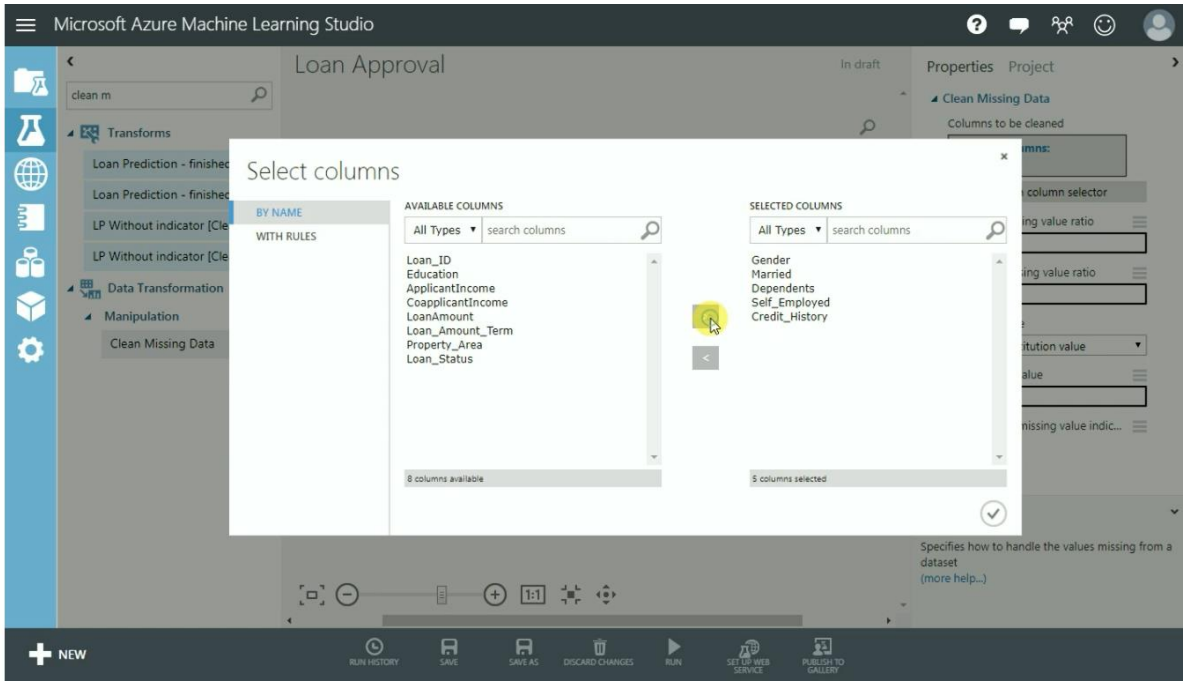
Visualize each column and check for missing values and perform clean missing data action



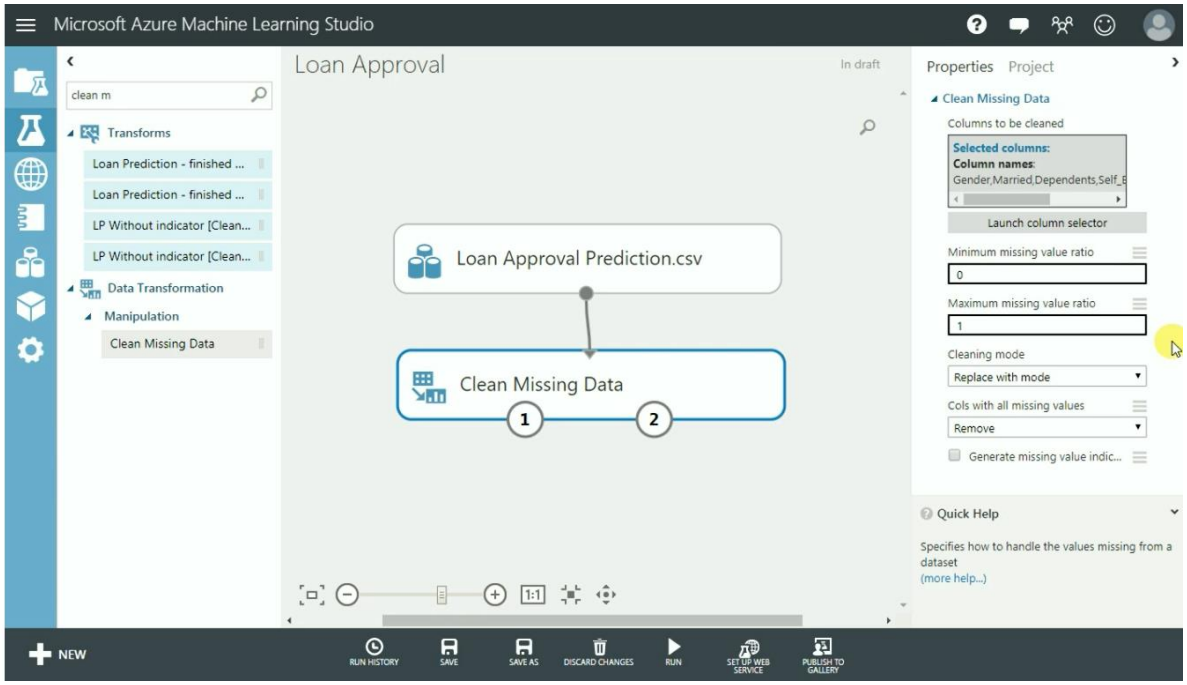
Attach both the datasets and launch column selector



Select the required types and click ok

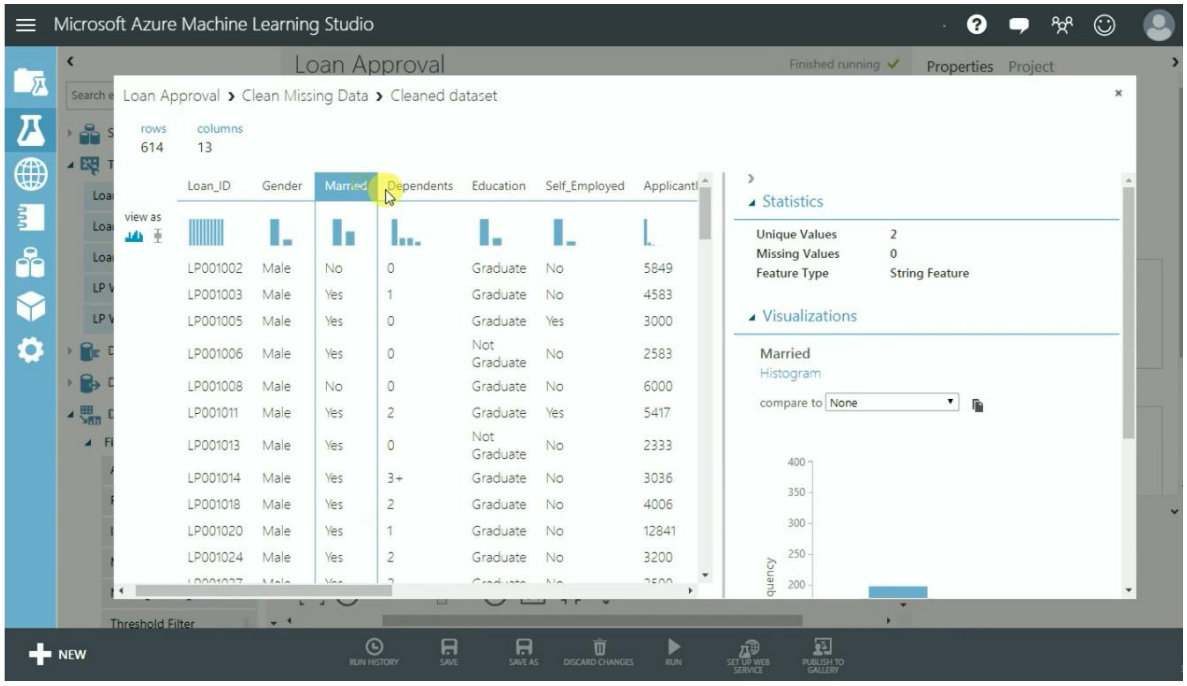


Select the required parameters to clean missing data for string features with replacement of mode



Run the module and visualize to check the result.

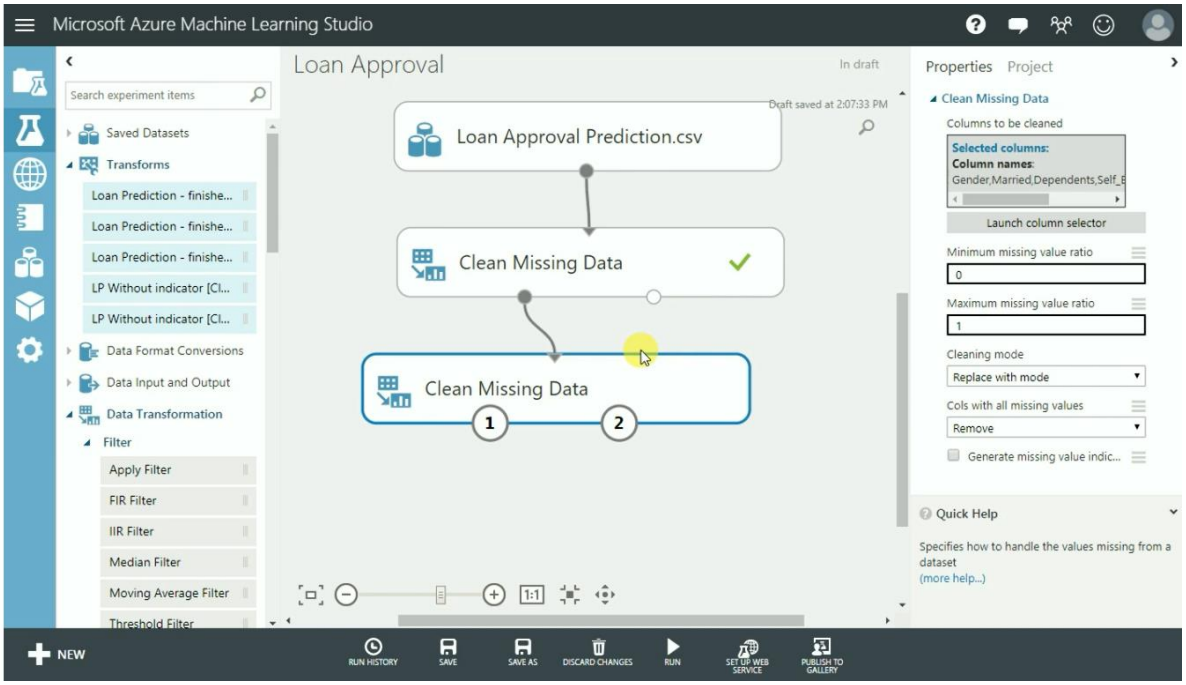
The result is successful



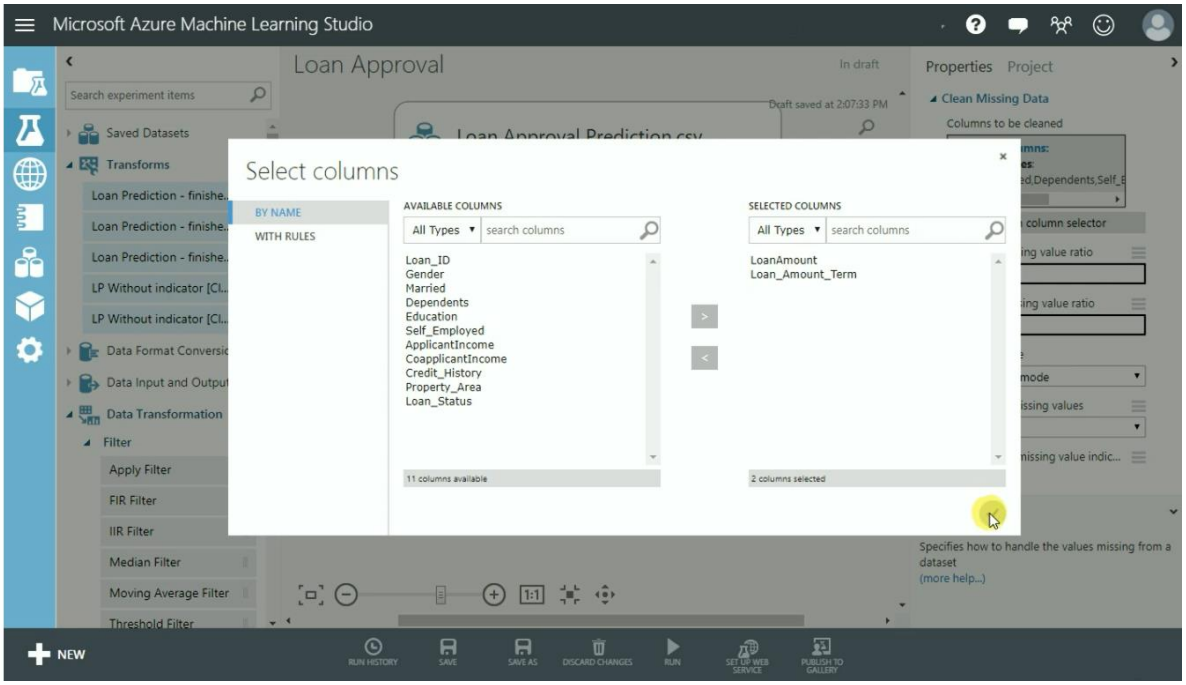
Cleaning Numeric Data

Now experiment for numeric features for clean missing data

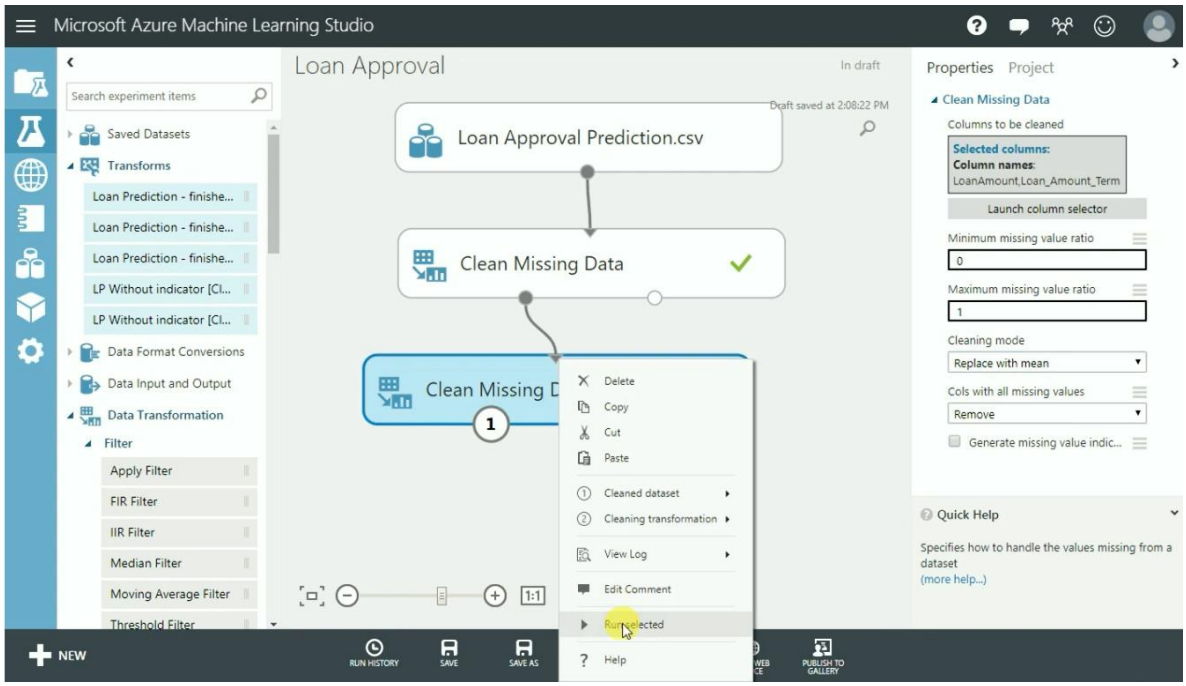
Simply Copy and paste clean missing dataset and connect the output node to input



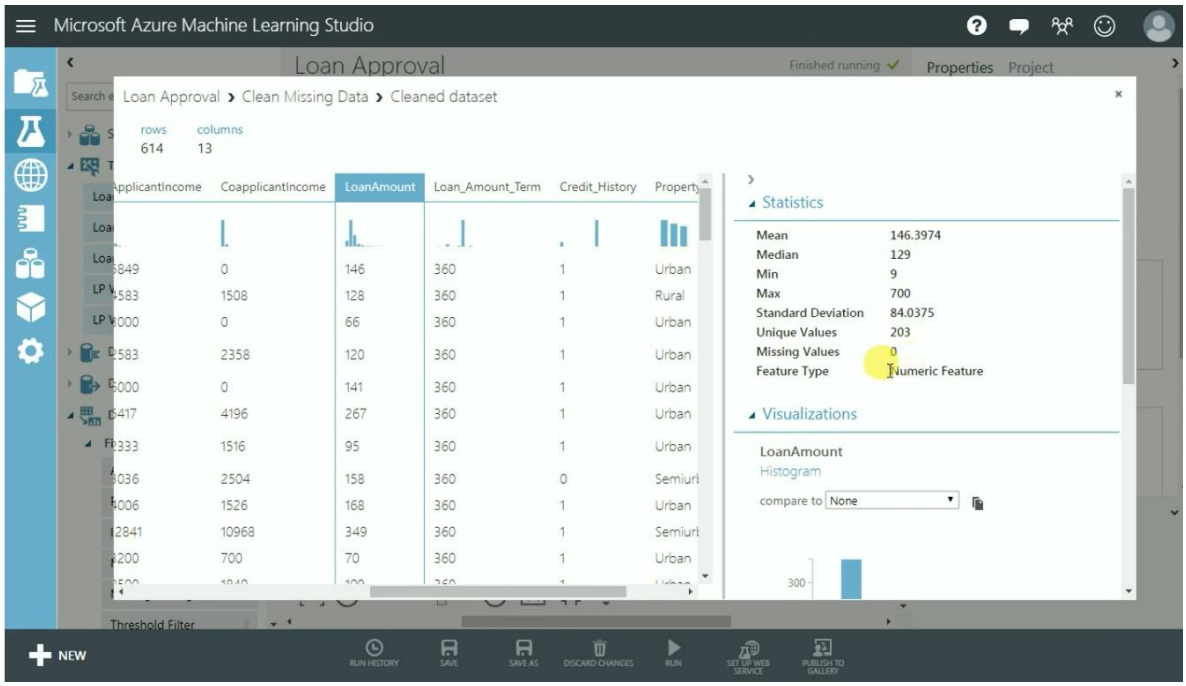
Launch column selector and select numerical features and press ok



Change the parameters cleaning mode by replace with mean and run the module

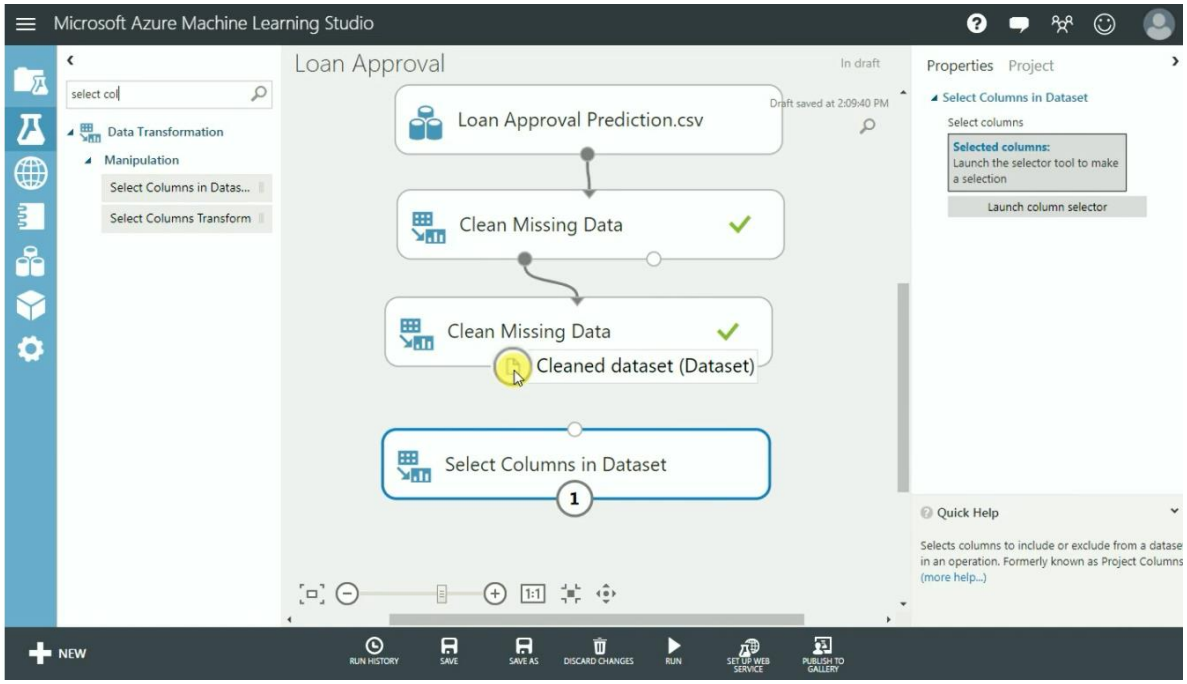


Visualize the module now and can see the numeric features showing missing values 0



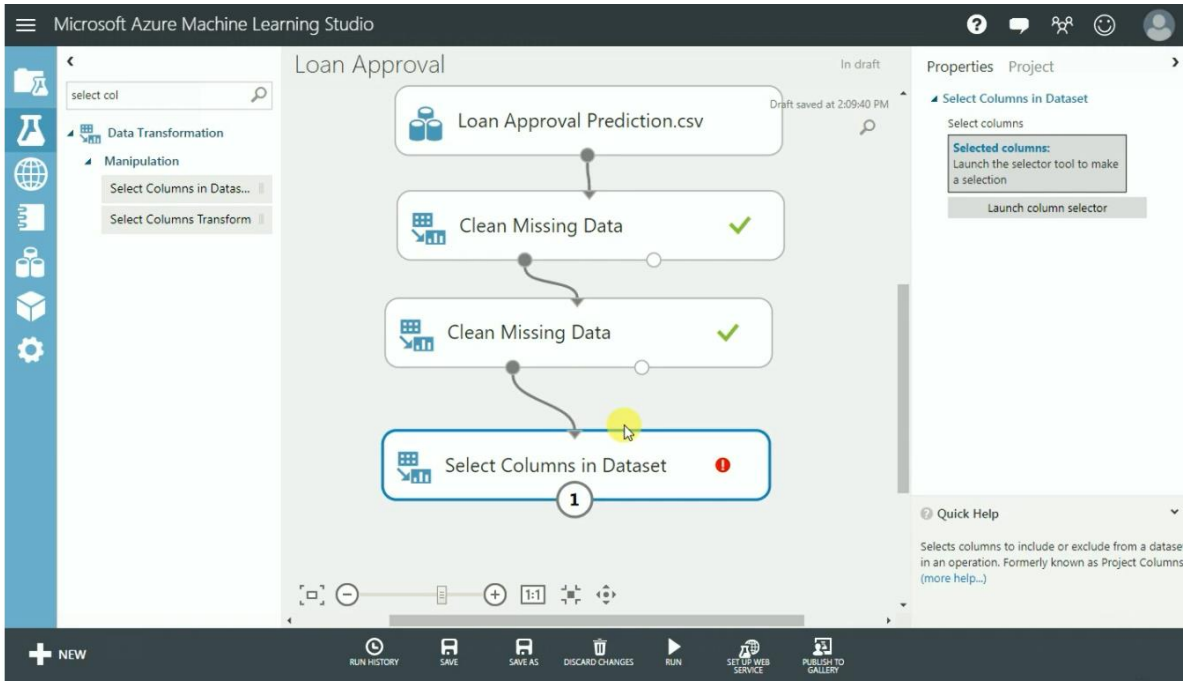
Select Columns in Dataset

Search columns dataset and drop in canvas



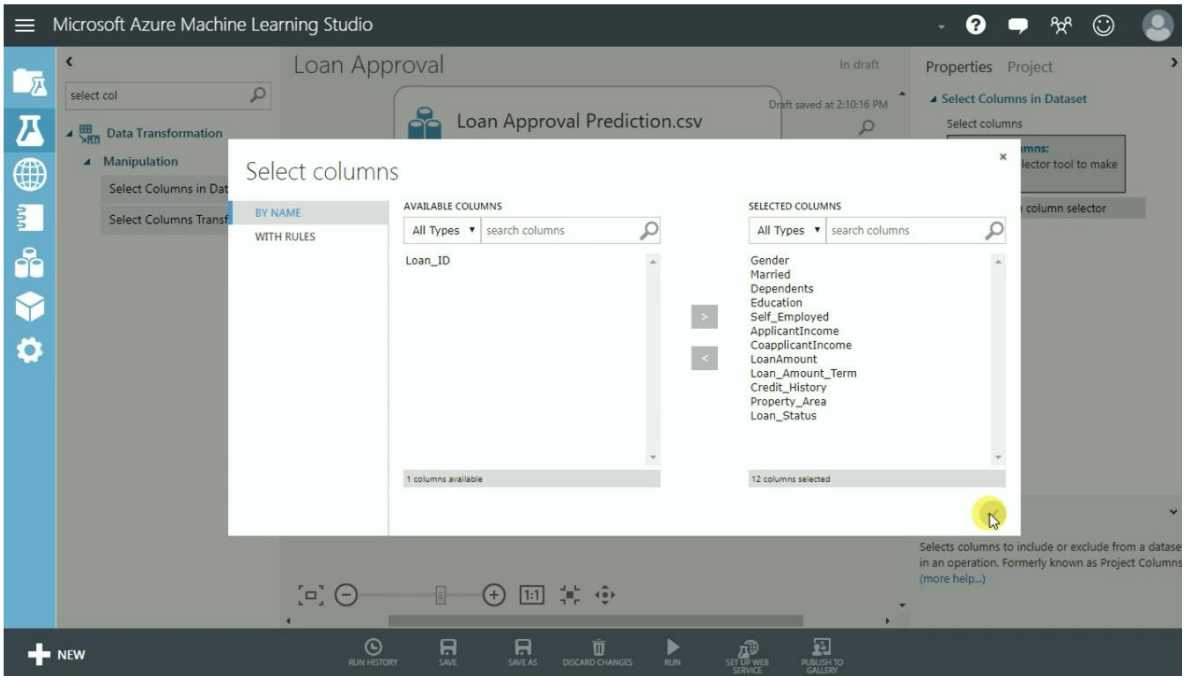
Connect the clean missing dataset output node to select columns in dataset input node

And launch column selector



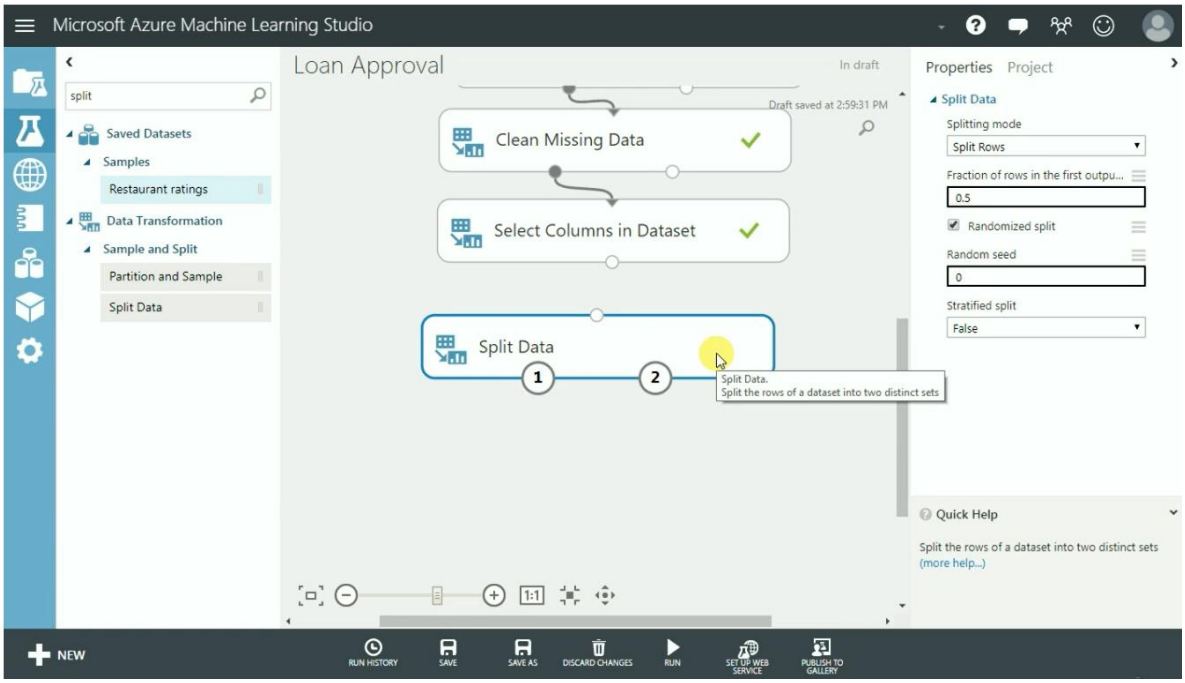
Select all the types except loan ID and press ok

Run and visualize

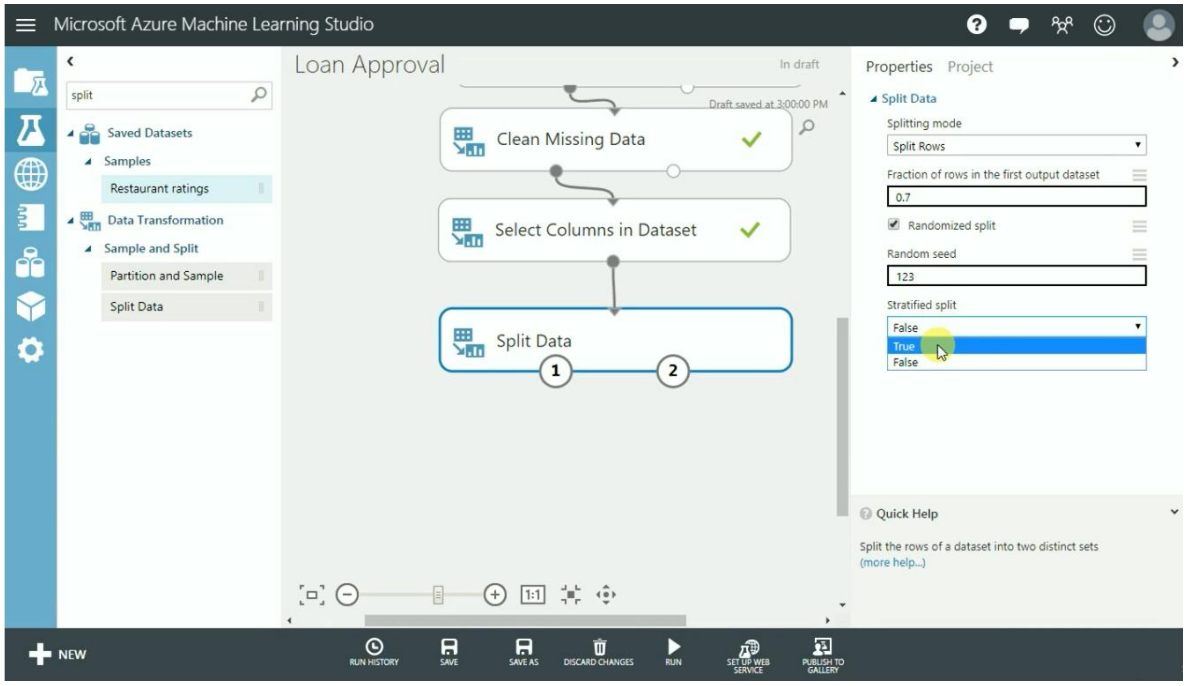


Split Data

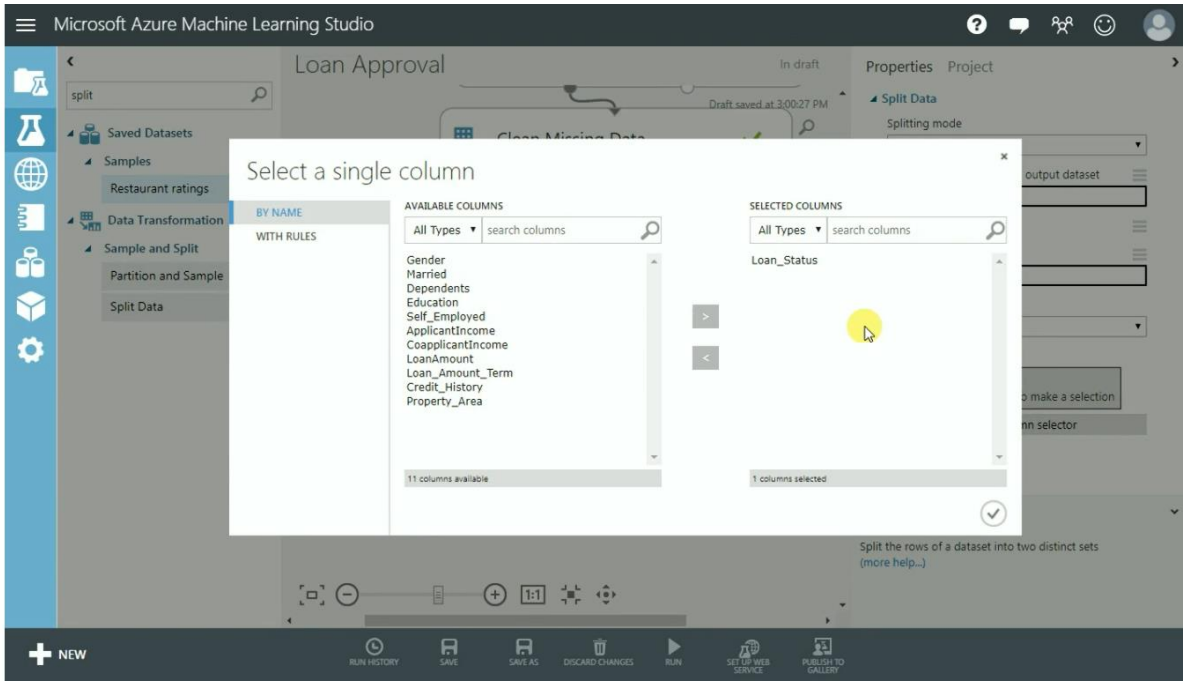
Search for split data and drop in canvas



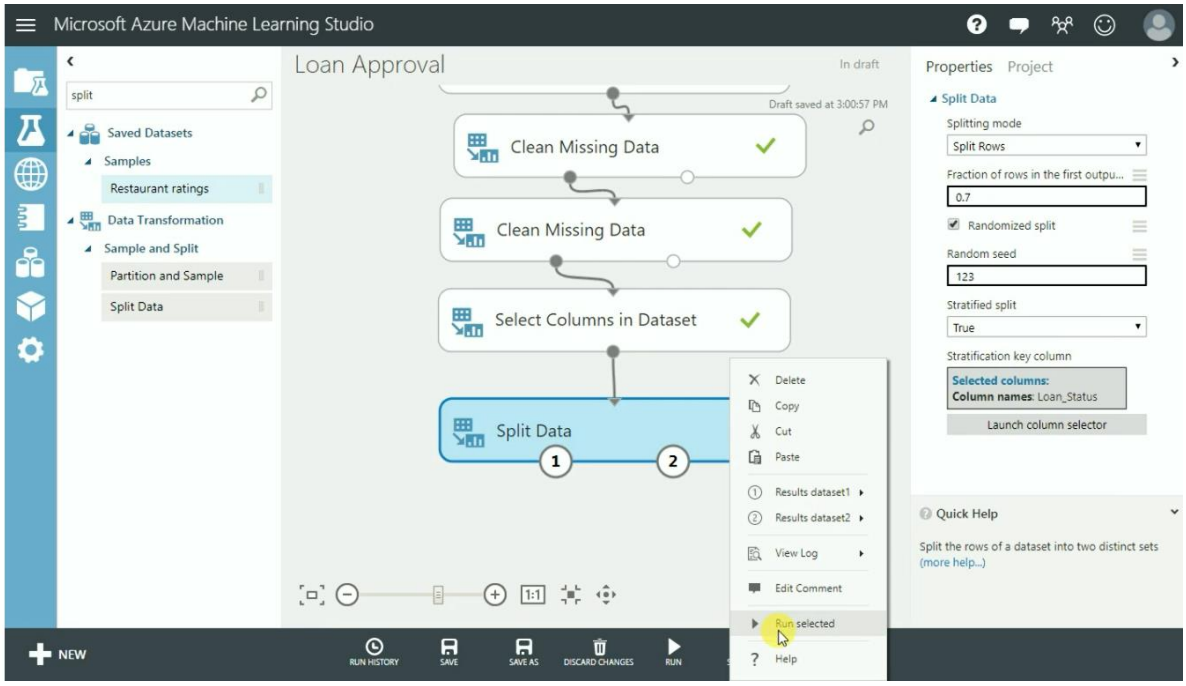
Attach both the dataset and input parameters as required



Now launch the column selector and select loan status and press ok

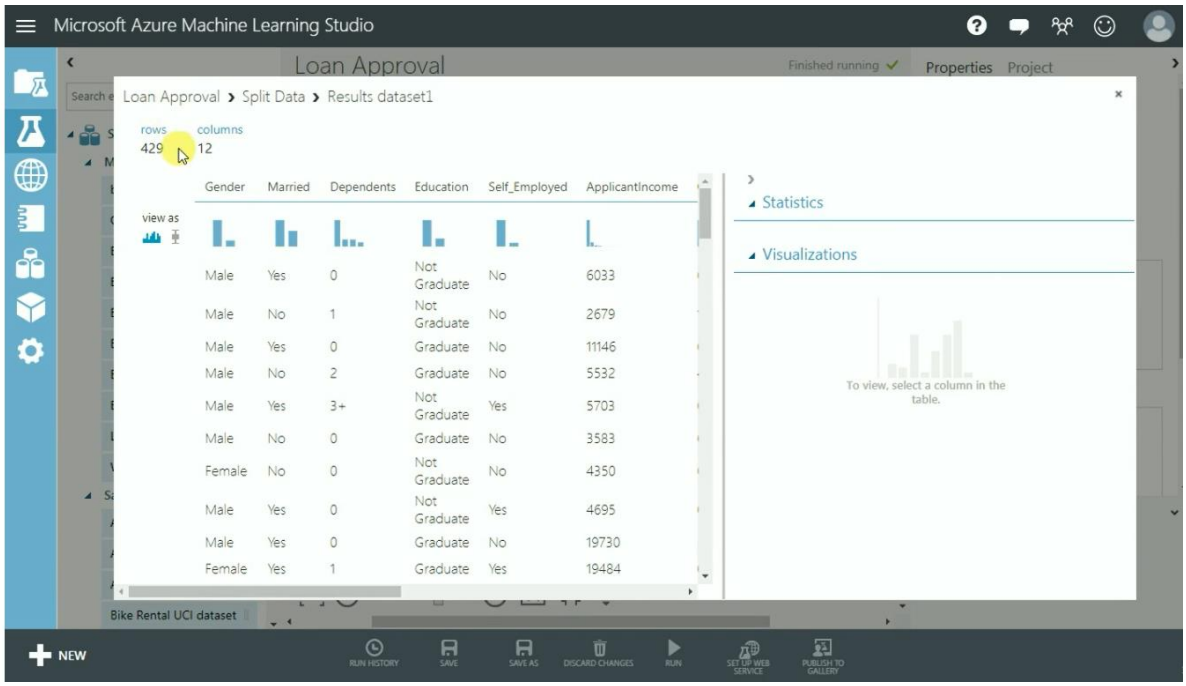


Run the module



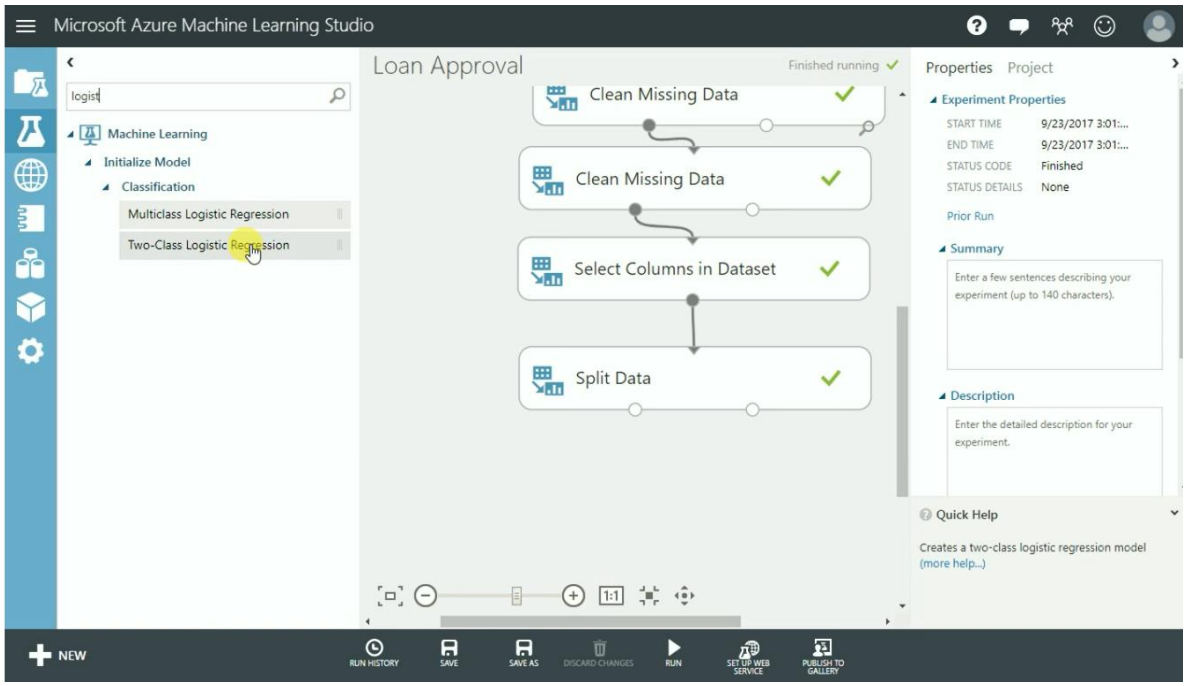
Visualize node 1 post execution. Now you can find 70 percent data here

And similarly, you can visualize node 2.

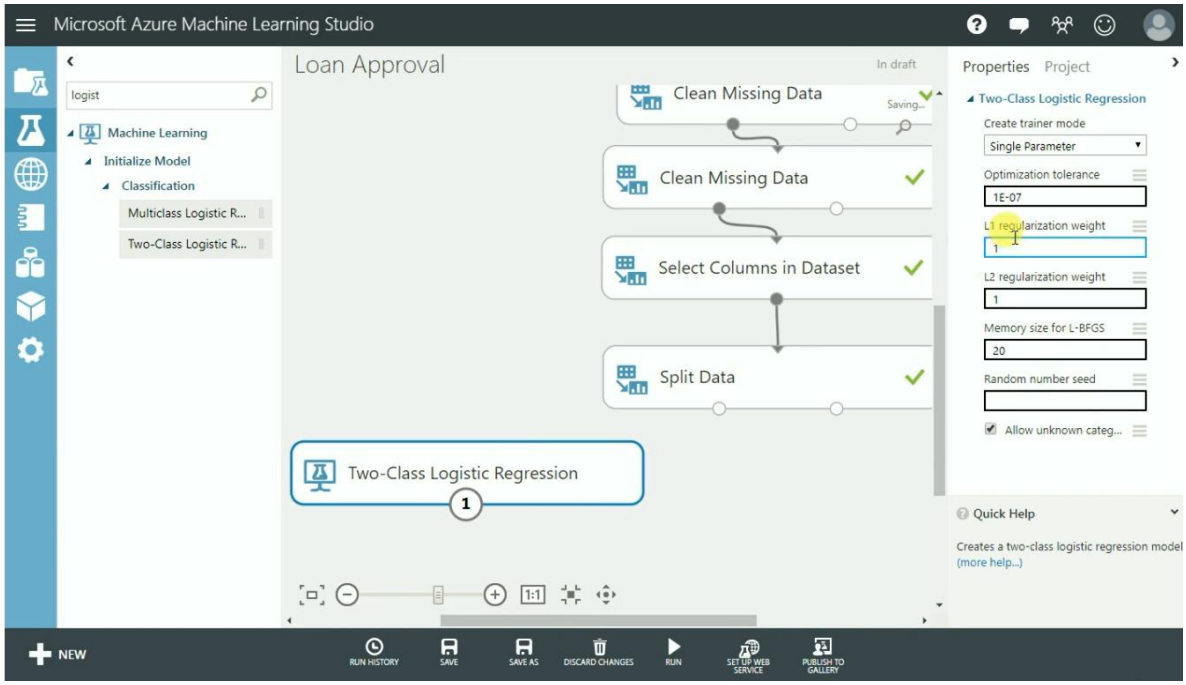


Usage of Two-Class Logistic Regression

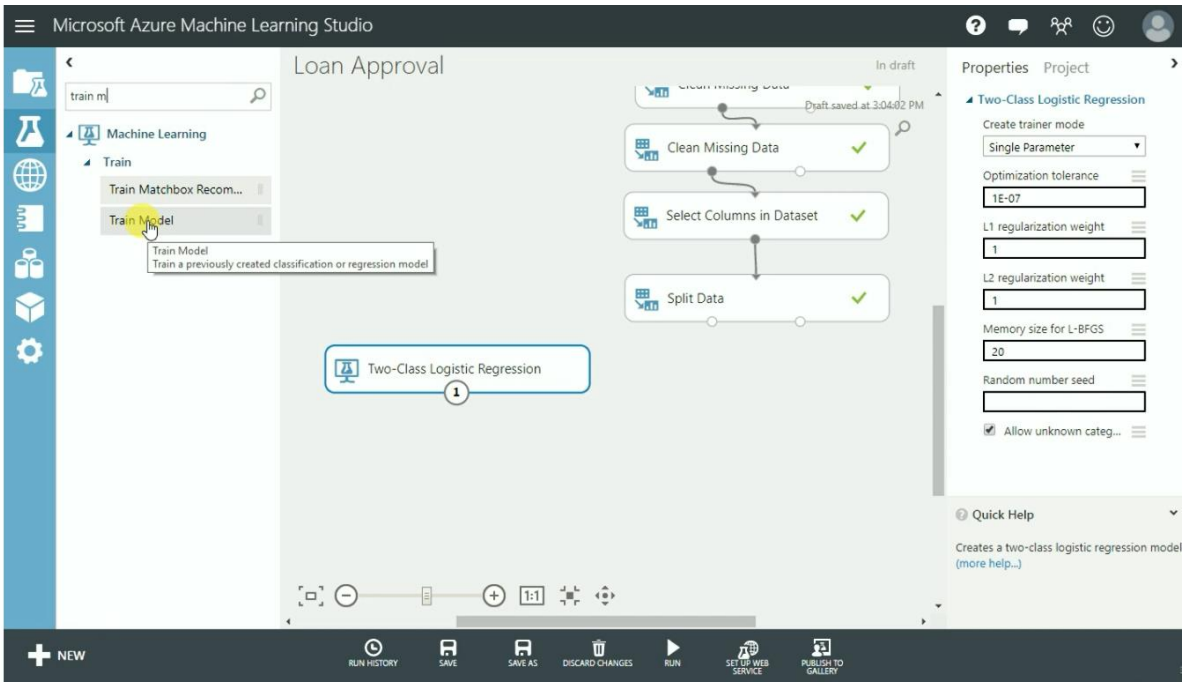
Search for two class logistic regression



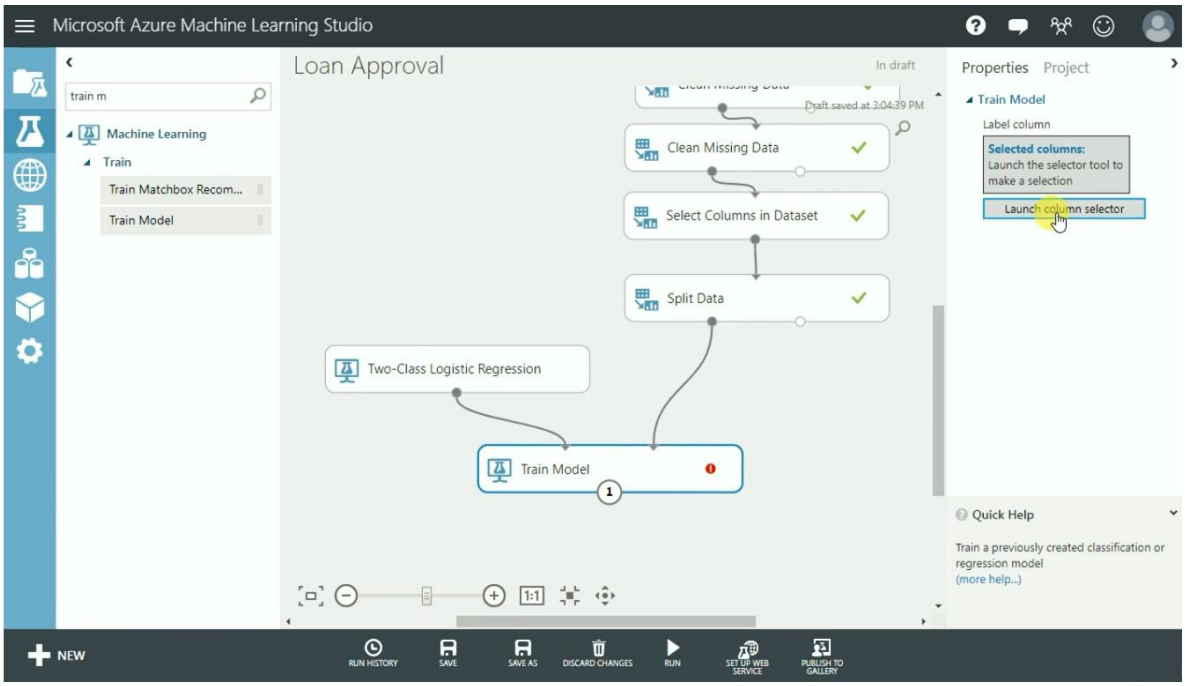
Drop the same in canvas and input required parameters



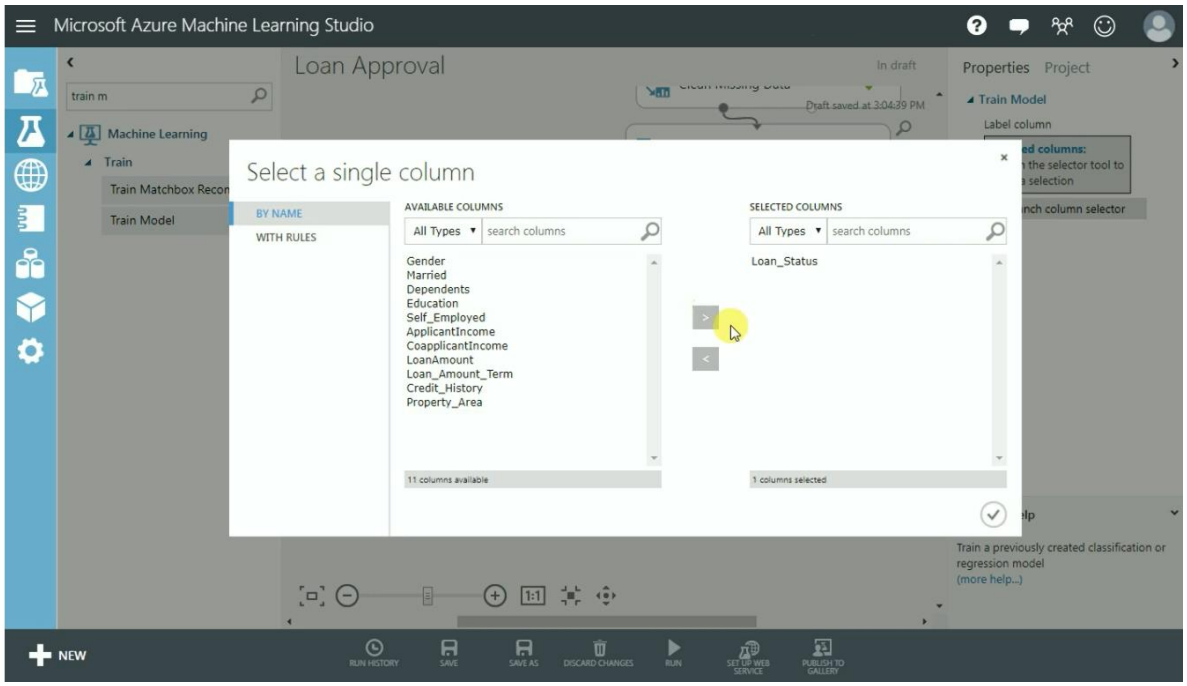
Search for model and drop in the canvas



Connect output node from split data and two class logistic regression to the train model input node

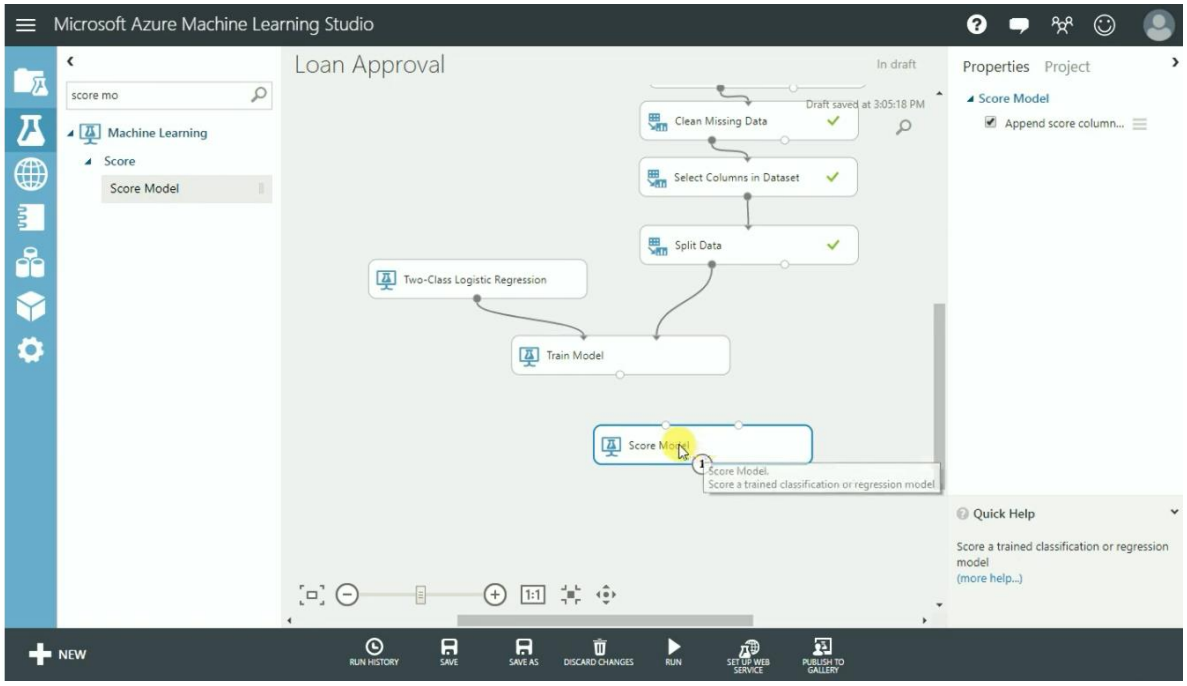


Now select launch column selector and select loan status and click ok

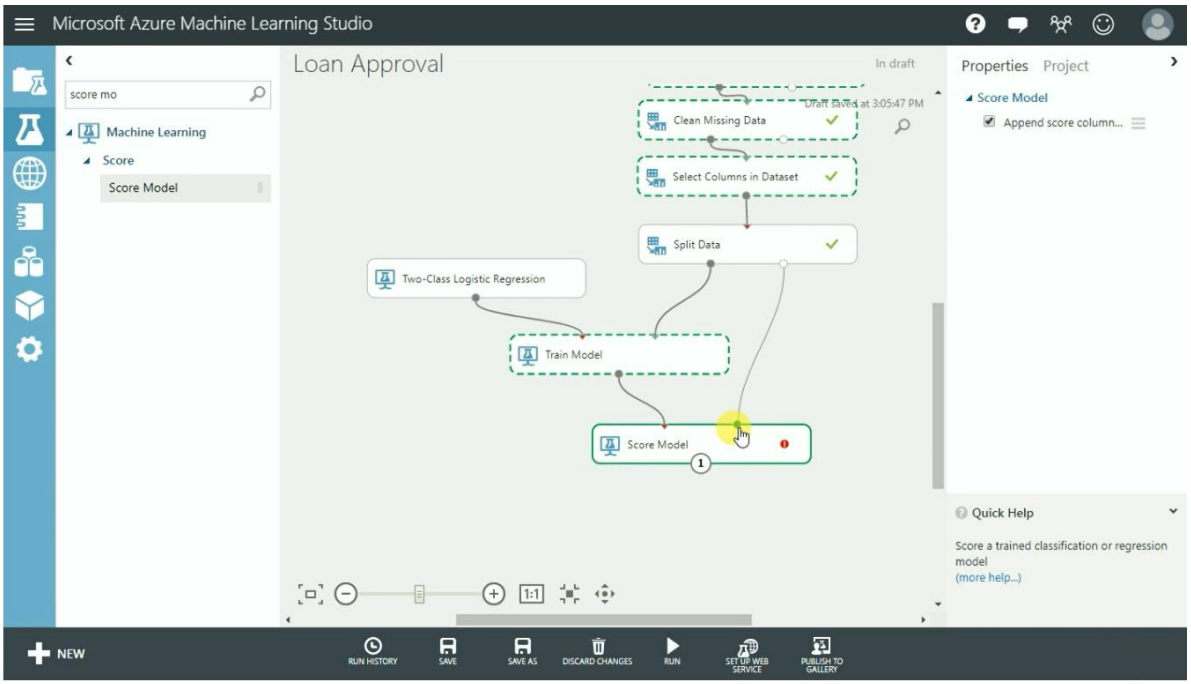


Score and Evaluate Model

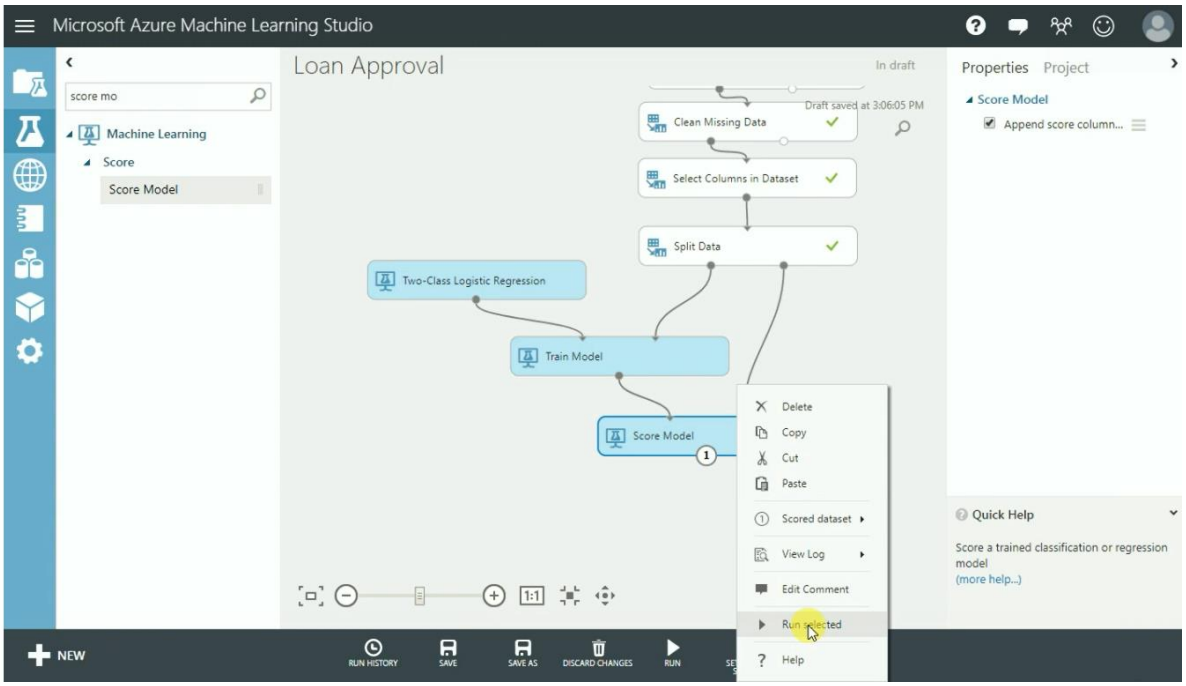
Search for score model and drop in canvas



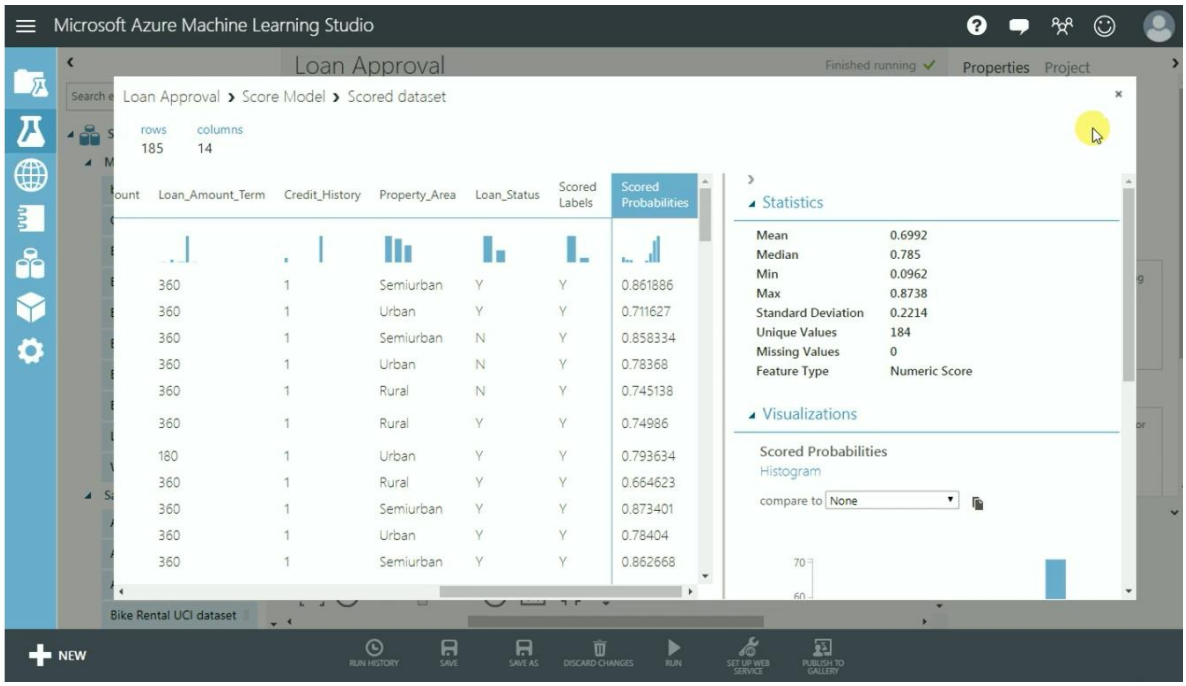
Connect output node from split data and connect to score model



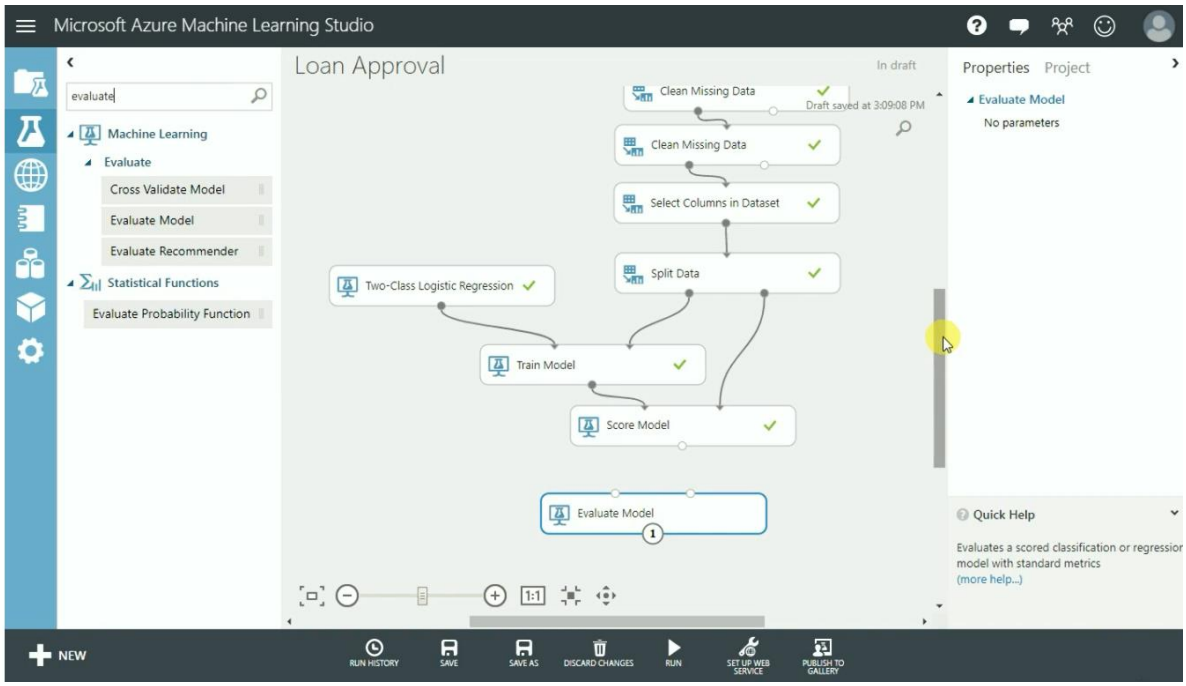
Right click and run selected



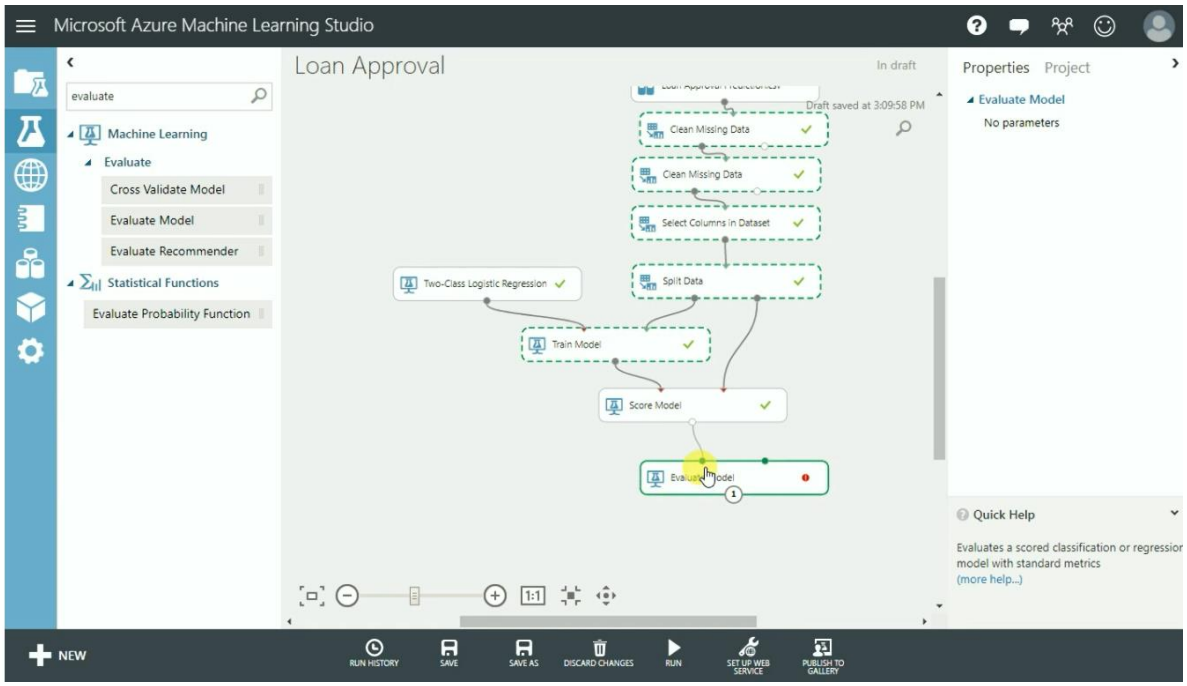
Visualize the result and check the accuracy



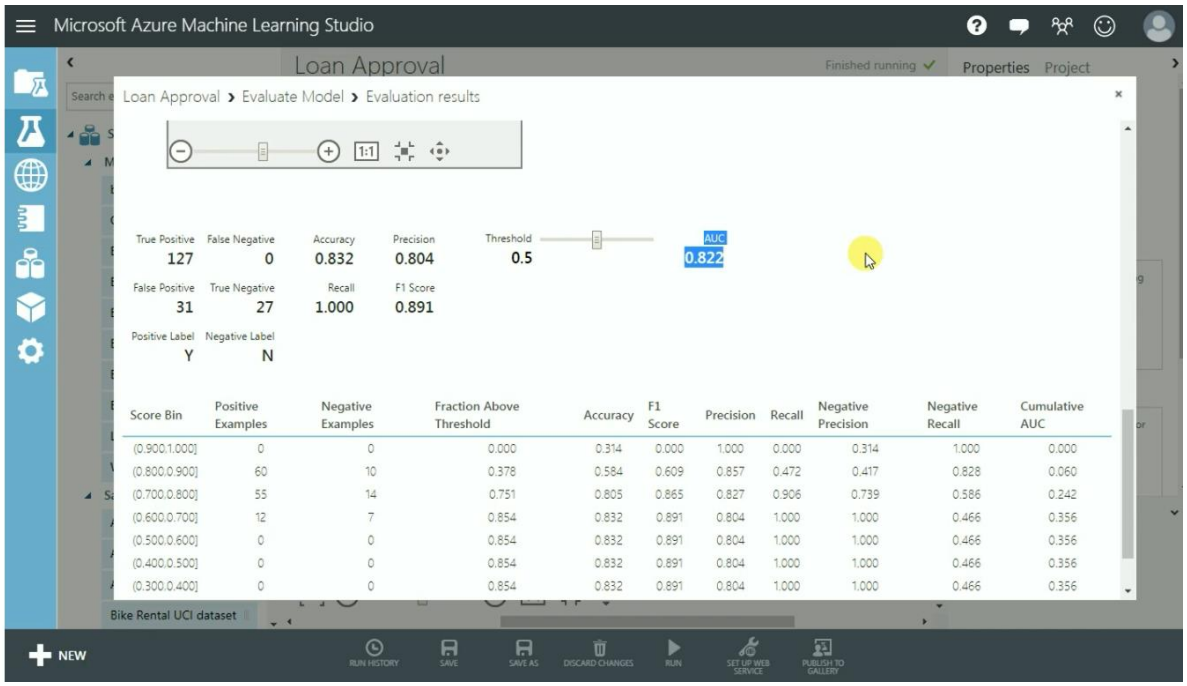
Since there is a compromise in accuracy search for evaluate model and drop in canvas



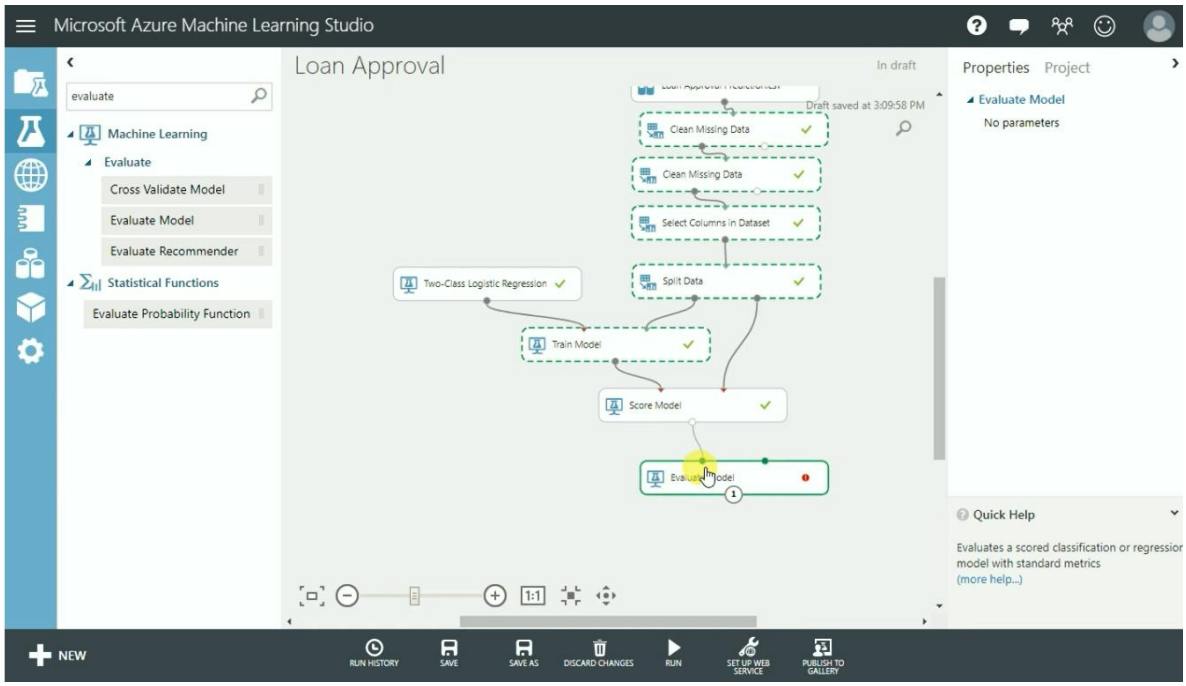
Connect the score model to evaluate model node1 and run selected



Visualize the output showing good accuracy

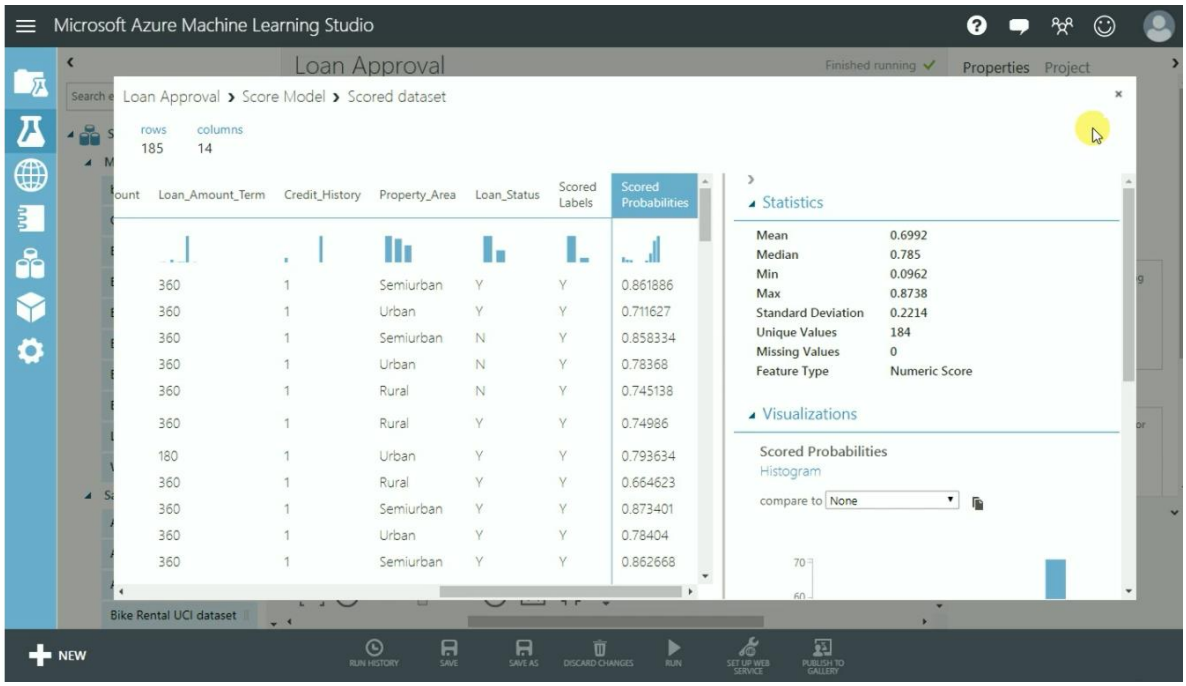


UNDERSTANDING THE RESULTS



Right click and visualize Evaluate model

Right click and visualize Score model

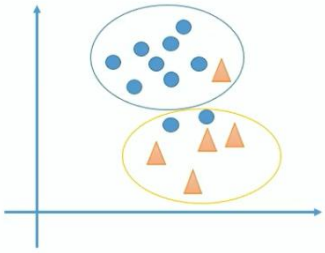


Understanding the Output

Prediction Outcome

	Predicted Positives	Predicted Negatives
Actual Positives	True Positives	False Negatives
Actual Negative	False Positives	True Negatives

	Predicted Positives	Predicted Negatives	
Actual Positives	8	2	10
Actual Negative	1	4	5
	9	6	



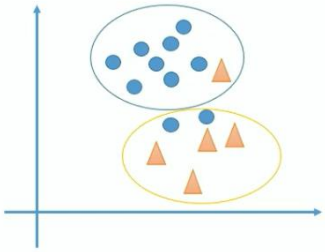
Accuracy – Proportions of total number of correct results

Accuracy = (8 + 4) / 15 = 0.8 or 80%

Prediction Outcome

	Predicted Positives	Predicted Negatives
Actual Positives	True Positives	False Negatives
Actual Negative	False Positives	True Negatives

	Predicted Positives	Predicted Negatives	
Actual Positives	8	2	10
Actual Negative	1	4	5
	9	6	



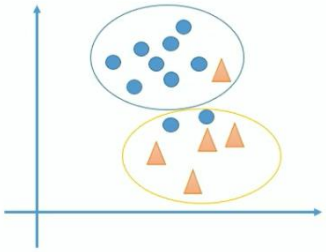
Precision – Proportion of correct positive results out of all predicted positive results

Precision = 8 / 9 = 0.889 or 88.9%

Prediction Outcome

	Predicted Positives	Predicted Negatives
Actual Positives	True Positives	False Negatives
Actual Negative	False Positives	True Negatives

	Predicted Positives	Predicted Negatives	
Actual Positives	8	2	10
Actual Negative	1	4	5
	9	6	



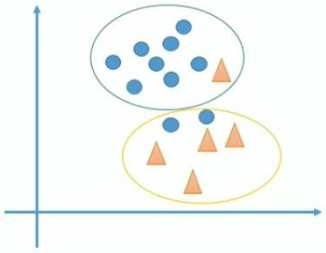
Recall – Proportion of actual positive cases

Recall = 8 / (8 + 2) = 0.8 or 80%

Prediction Outcome

	Predicted Positives	Predicted Negatives
Actual Positives	True Positives	False Negatives
Actual Negative	False Positives	True Negatives

	Predicted Positives	Predicted Negatives	
Actual Positives	8	2	10
Actual Negative	1	4	5
	9	6	



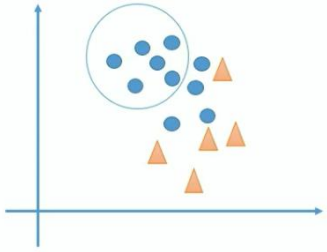
F1-Score – Weighted Average (Harmonic Mean) of Precision and Recall

F1Score = 2 * Precision * Recall / (Precision + Recall) = 0.84

Prediction Outcome

	Predicted Positives	Predicted Negatives
Actual Positives	True Positives	False Negatives
Actual Negative	False Positives	True Negatives

	Predicted Positives	Predicted Negatives	
Actual Positives	6	4	10
Actual Negative	0	5	5
	6	6	



In the Previous example

Precision = 6 / 6 = 1 or 100%

Recall = 6 / (6 + 4) = 0.6 or 60%

Average = 0.8

May lead to false interpretation

Precision = 0.889

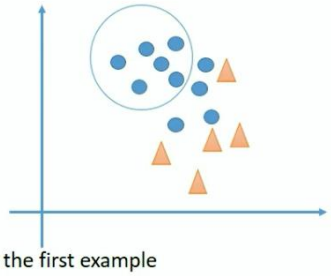
Recall = 0.8

Average = 0.84

Prediction Outcome

	Predicted Positives	Predicted Negatives
Actual Positives	True Positives	False Negatives
Actual Negative	False Positives	True Negatives

	Predicted Positives	Predicted Negatives	
Actual Positives	6	4	10
Actual Negative	0	5	5
	6	6	



In the first example

Precision = 6 / 6 = 1 or 100%

Recall = 6 / (8 + 2) = 0.6 or 60%

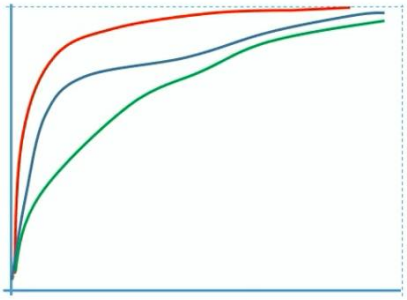
F1Score = 0.75

Precision = 0.889

Recall = 0.8

F1Score = 0.84

AUC ROC



True Positive Rate
TP / (TP + FN)

False Positive Rate
FP / (FP + TN)

AUC – Area Under the Curve

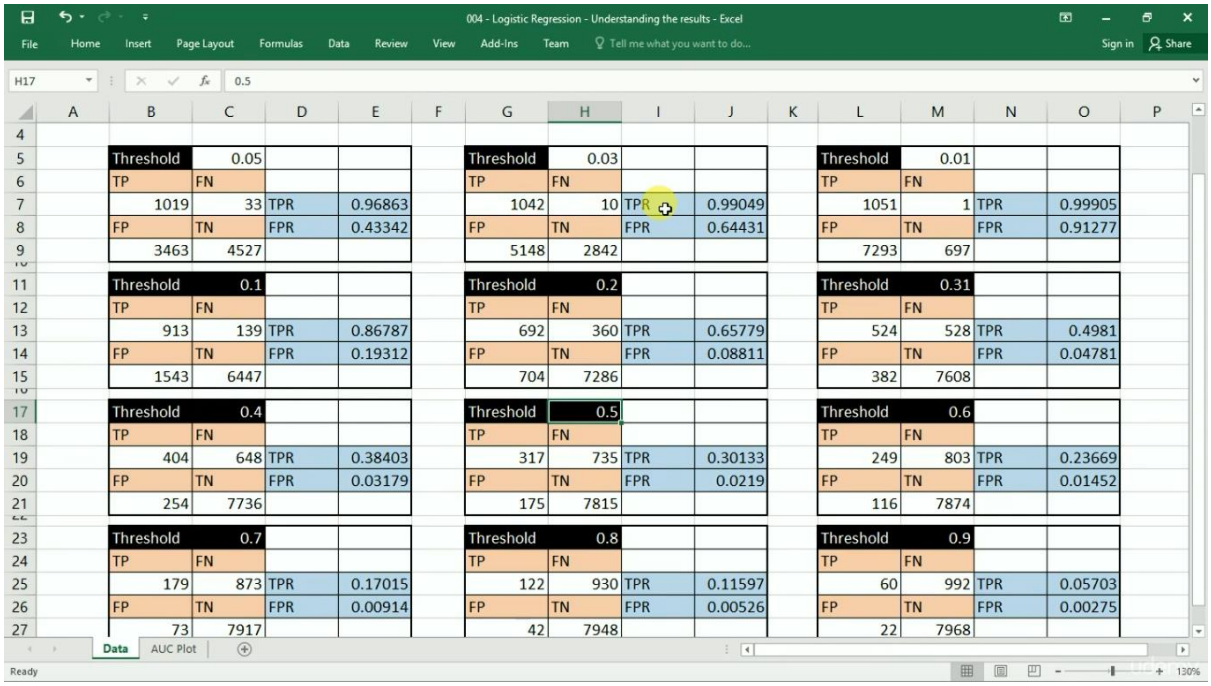
ROC – Receiver Operating Characteristics

First used during World War II for the analysis of radar signals

Following the attack on Pearl Harbor in 1941, the United States army began new research to increase the prediction of correctly detected Japanese aircraft from their radar signals.

For this purposes they measured the ability of radar receiver operators to make these important distinctions, which was called the Receiver Operating Characteristics

Provides a single number that lets you compare models of different types.



Threshold	0.05		
TP	FN		
1019	33	TPR	0.96863
FP	TN	FPR	0.43342
3463	4527		

Threshold	0.03		
TP	FN		
1042	10	TPR	0.99049
FP	TN	FPR	0.64431
5148	2842		

Threshold	0.01		
TP	FN		
1051	1	TPR	0.99905
FP	TN	FPR	0.91277
7293	697		

Threshold	0.1		
TP	FN		
913	139	TPR	0.86787
FP	TN	FPR	0.19312
1543	6447		

Threshold	0.2		
TP	FN		
692	360	TPR	0.65779
FP	TN	FPR	0.08811
704	7286		

Threshold	0.31		
TP	FN		
524	528	TPR	0.4981
FP	TN	FPR	0.04781
382	7608		

Threshold	0.4		
TP	FN		
404	648	TPR	0.38403
FP	TN	FPR	0.03179
254	7736		

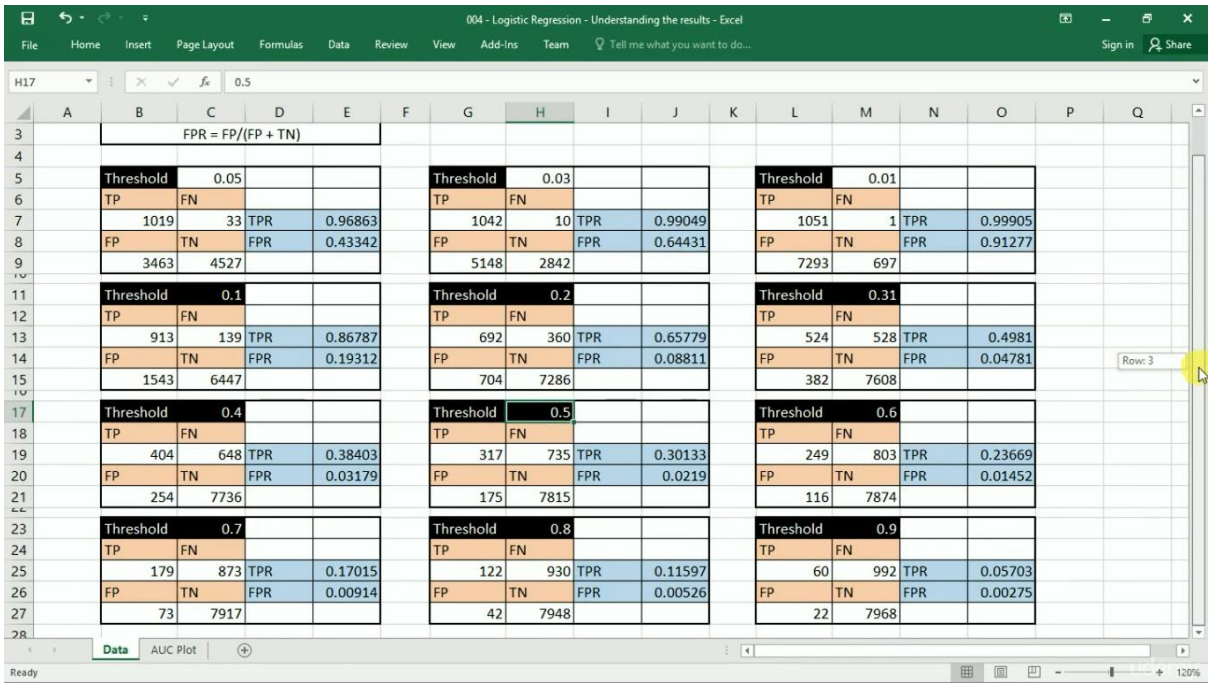
Threshold	0.5		
TP	FN		
317	735	TPR	0.30133
FP	TN	FPR	0.0219
175	7815		

Threshold	0.6		
TP	FN		
249	803	TPR	0.23669
FP	TN	FPR	0.01452
116	7874		

Threshold	0.7		
TP	FN		
179	873	TPR	0.17015
FP	TN	FPR	0.00914
73	7917		

Threshold	0.8		
TP	FN		
122	930	TPR	0.11597
FP	TN	FPR	0.00526
42	7948		

Threshold	0.9		
TP	FN		
60	992	TPR	0.05703
FP	TN	FPR	0.00275
22	7968		



004 - Logistic Regression - Understanding the results - Excel

FPR = FP/(FP + TN)

Threshold	0.05		
TP	FN		
1019	33	TPR	0.96863
FP	TN	FPR	0.43342
3463	4527		

Threshold	0.03		
TP	FN		
1042	10	TPR	0.99049
FP	TN	FPR	0.64431
5148	2842		

Threshold	0.01		
TP	FN		
1051	1	TPR	0.99905
FP	TN	FPR	0.91277
7293	697		

Threshold	0.1		
TP	FN		
913	139	TPR	0.86787
FP	TN	FPR	0.19312
1543	6447		

Threshold	0.2		
TP	FN		
692	360	TPR	0.65779
FP	TN	FPR	0.08811
704	7286		

Threshold	0.31		
TP	FN		
524	528	TPR	0.4981
FP	TN	FPR	0.04781
382	7608		

Threshold	0.4		
TP	FN		
404	648	TPR	0.38403
FP	TN	FPR	0.03179
254	7736		

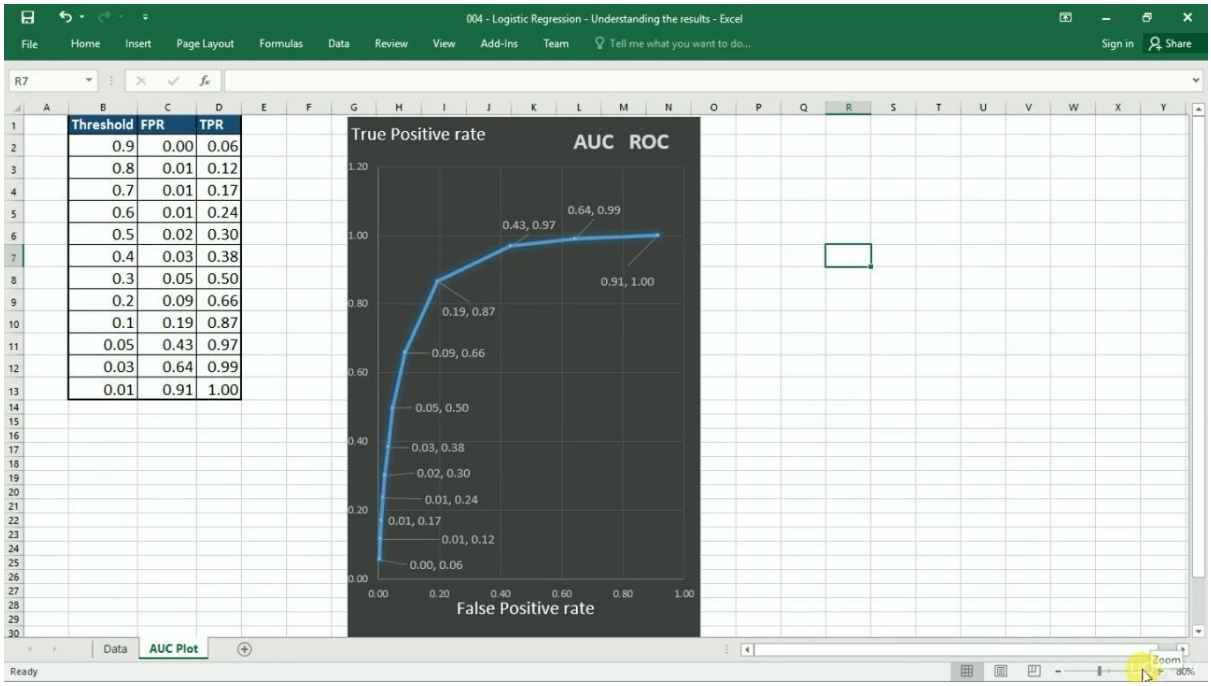
Threshold	0.5		
TP	FN		
317	735	TPR	0.30133
FP	TN	FPR	0.0219
175	7815		

Threshold	0.6		
TP	FN		
249	803	TPR	0.23669
FP	TN	FPR	0.01452
116	7874		

Threshold	0.7		
TP	FN		
179	873	TPR	0.17015
FP	TN	FPR	0.00914
73	7917		

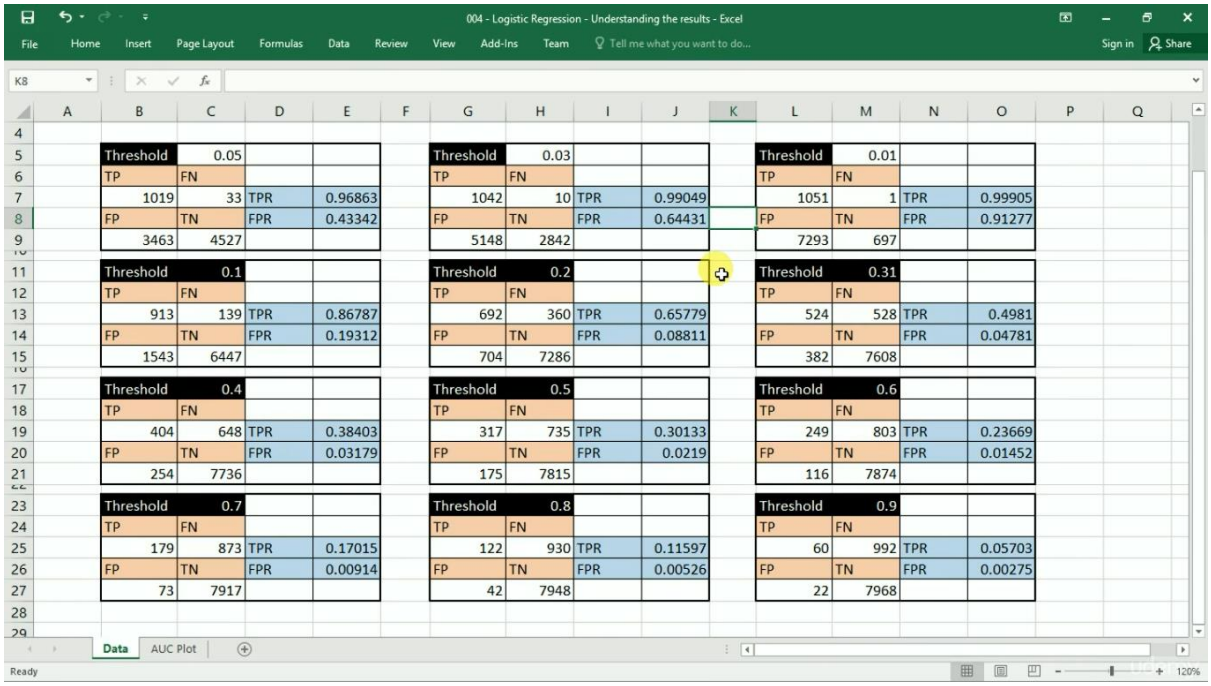
Threshold	0.8		
TP	FN		
122	930	TPR	0.11597
FP	TN	FPR	0.00526
42	7948		

Threshold	0.9		
TP	FN		
60	992	TPR	0.05703
FP	TN	FPR	0.00275
22	7968		



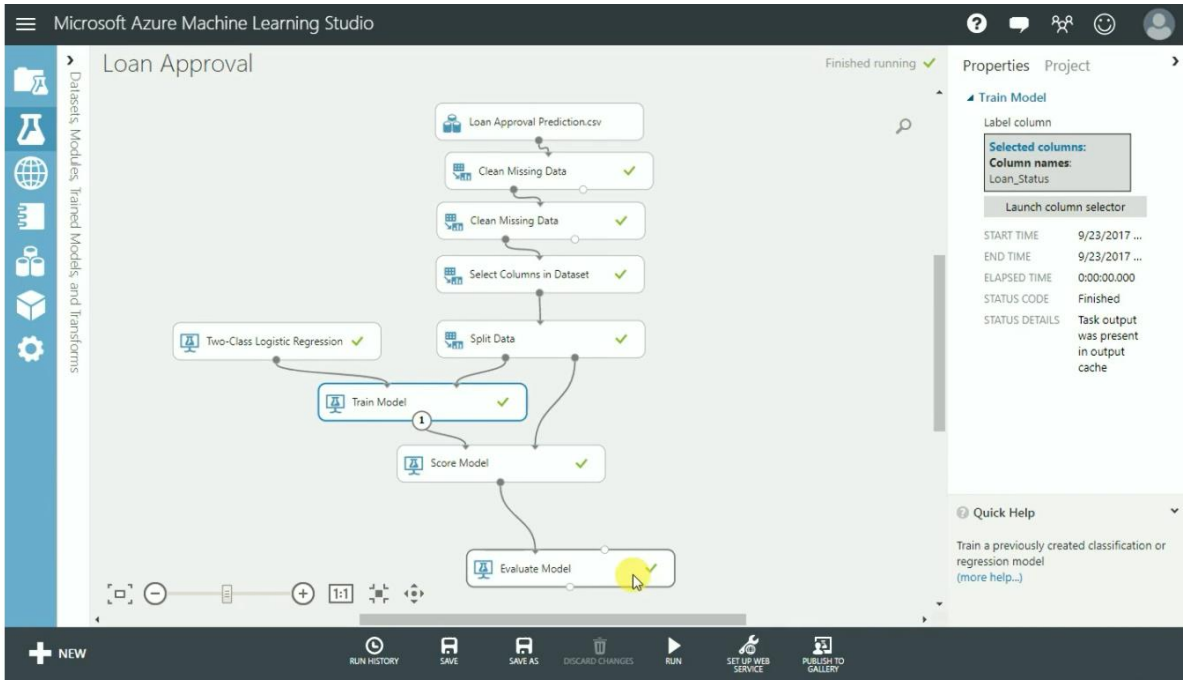
004 - Logistic Regression - Understanding the results - Excel

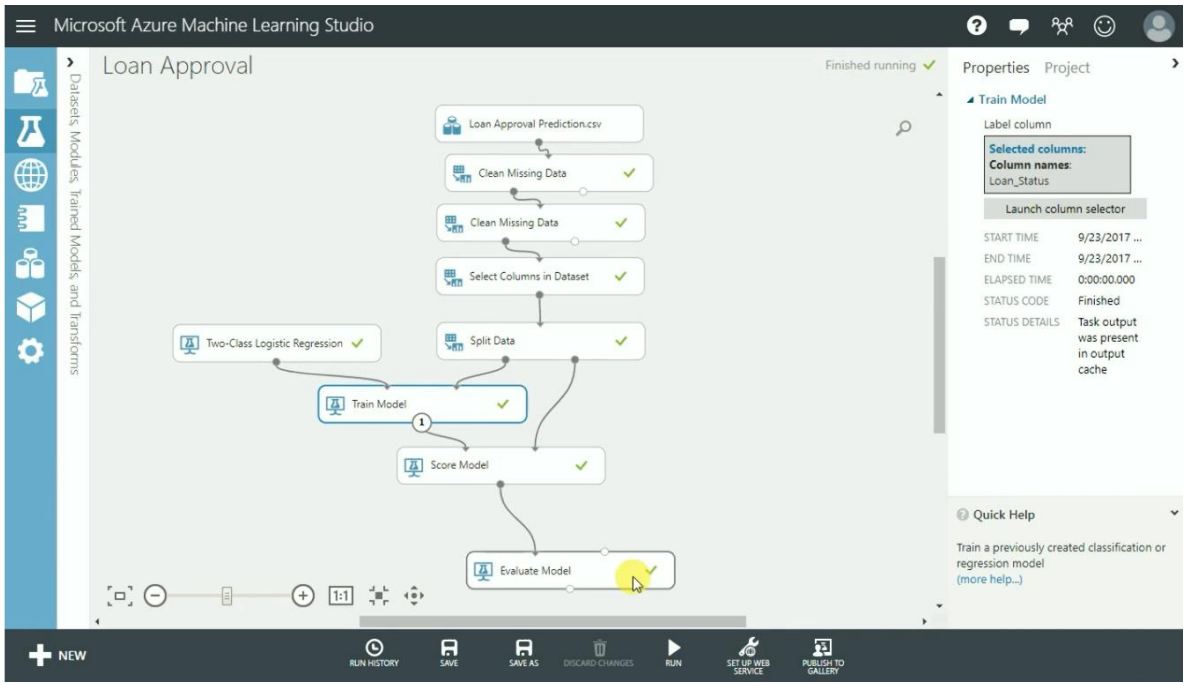
Threshold	FPR	TPR
0.9	0.00	0.06
0.8	0.01	0.12
0.7	0.01	0.17
0.6	0.01	0.24
0.5	0.02	0.30
0.4	0.03	0.38
0.3	0.05	0.50
0.2	0.09	0.66
0.1	0.19	0.87
0.05	0.43	0.97
0.03	0.64	0.99
0.01	0.91	1.00


004 - Logistic Regression - Understanding the results - Excel
File Home Insert Page Layout Formulas Data Review View Add-Ins Team Tell me what you want to do...
Sign in Share
K8
Threshold 0.05
TP FN
1019 33
FP TN
3463 4527
TPR 0.96863
FPR 0.43342
Threshold 0.03
TP FN
1042 10
FP TN
5148 2842
TPR 0.99049
FPR 0.64431
Threshold 0.01
TP FN
1051 1
FP TN
7293 697
TPR 0.99905
FPR 0.91277
Threshold 0.1
TP FN
913 139
FP TN
1543 6447
TPR 0.86787
FPR 0.19312
Threshold 0.2
TP FN
692 360
FP TN
704 7286
TPR 0.65779
FPR 0.08811
Threshold 0.31
TP FN
524 528
FP TN
382 7608
TPR 0.4981
FPR 0.04781
Threshold 0.4
TP FN
404 648
FP TN
254 7736
TPR 0.38403
FPR 0.03179
Threshold 0.5
TP FN
317 735
FP TN
175 7815
TPR 0.30133
FPR 0.0219
Threshold 0.6
TP FN
249 803
FP TN
116 7874
TPR 0.23669
FPR 0.01452
Threshold 0.7
TP FN
179 873
FP TN
73 7917
TPR 0.17015
FPR 0.00914
Threshold 0.8
TP FN
122 930
FP TN
42 7948
TPR 0.11597
FPR 0.00526
Threshold 0.9
TP FN
60 992
FP TN
22 7968
TPR 0.05703
FPR 0.00275
Data AUC Plot
Ready
120%

Impact Analysis & Stratification

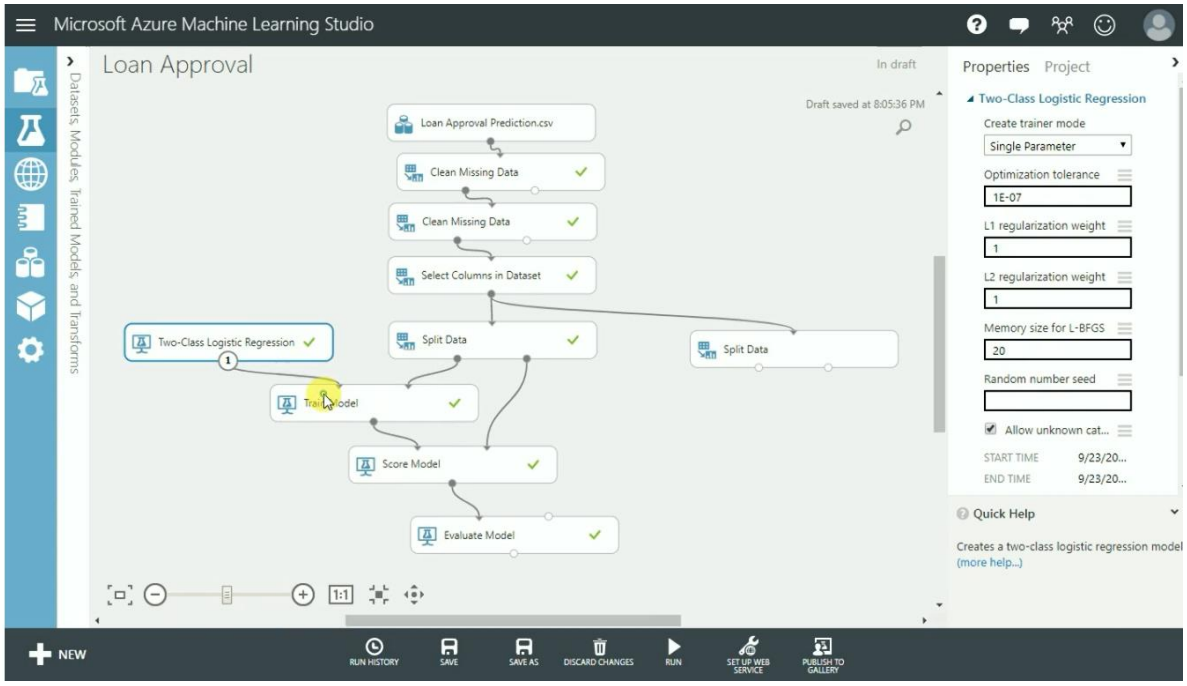
Select the Loan prediction model and visualize the evaluate model



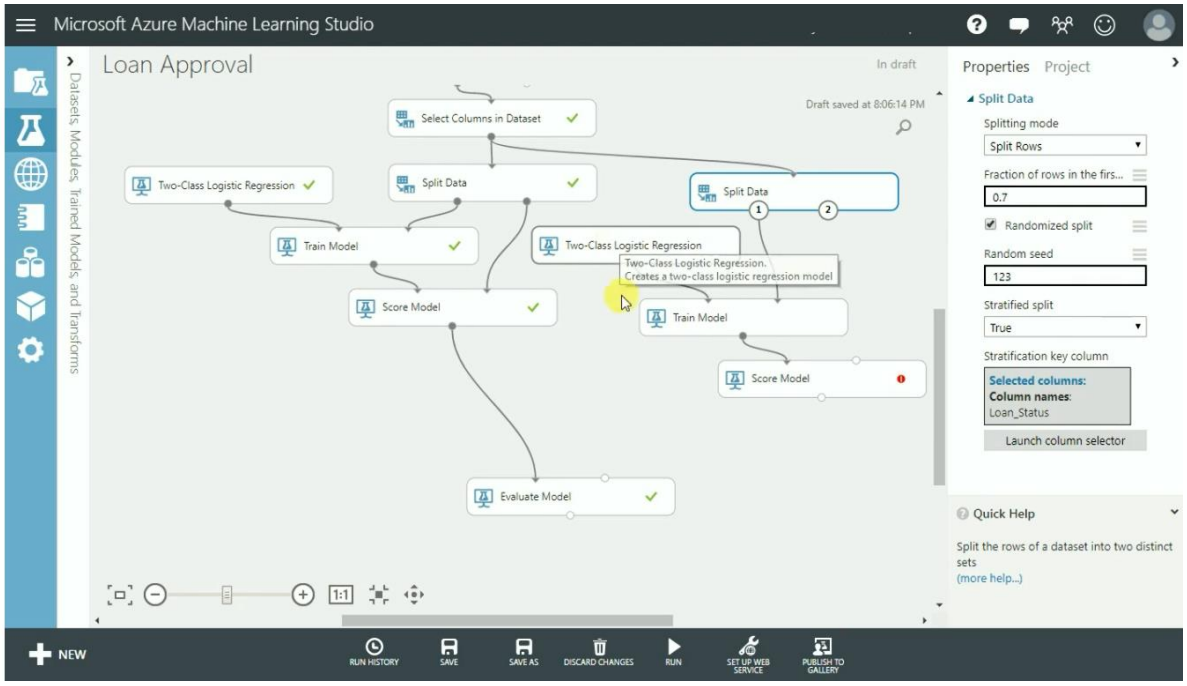


To analyse the impact of stratification which done at split data

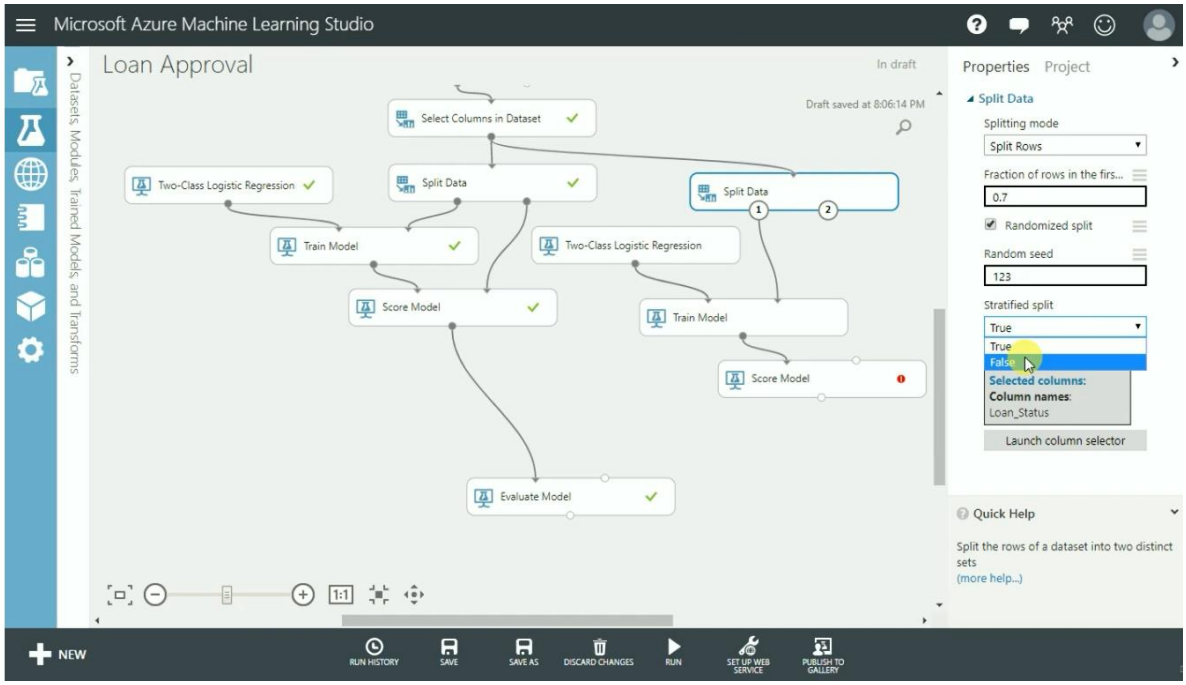
Repeat the datasets from split data again as below



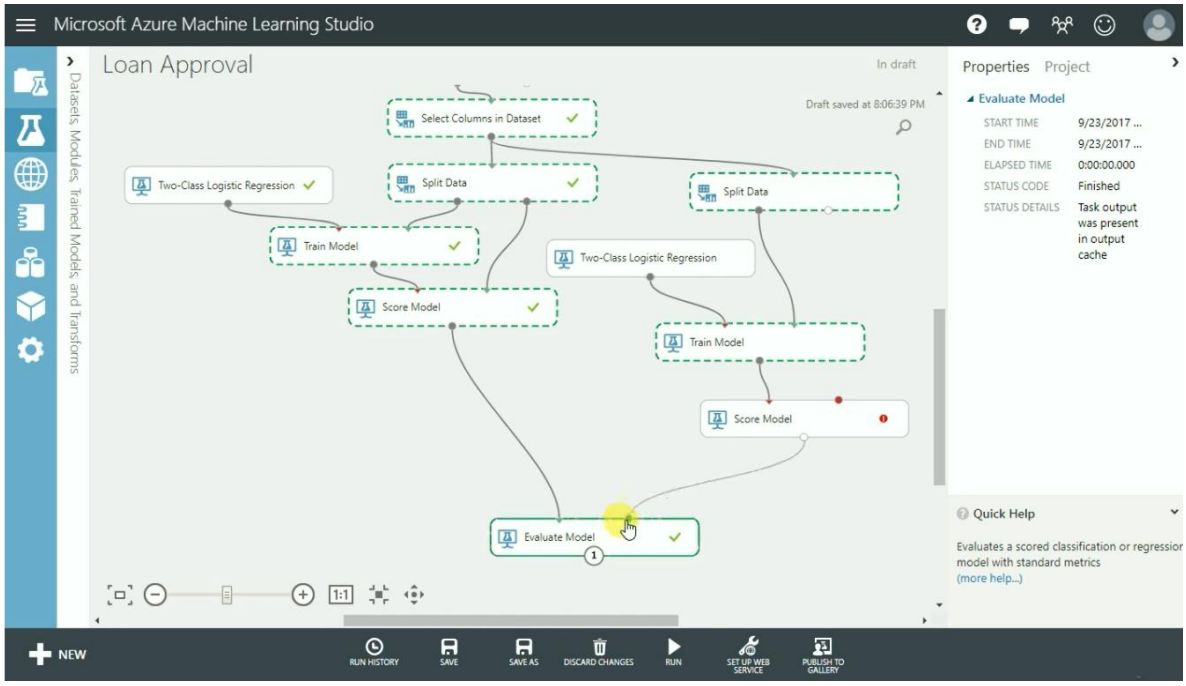
Copy and paste the data sets in split data as shown



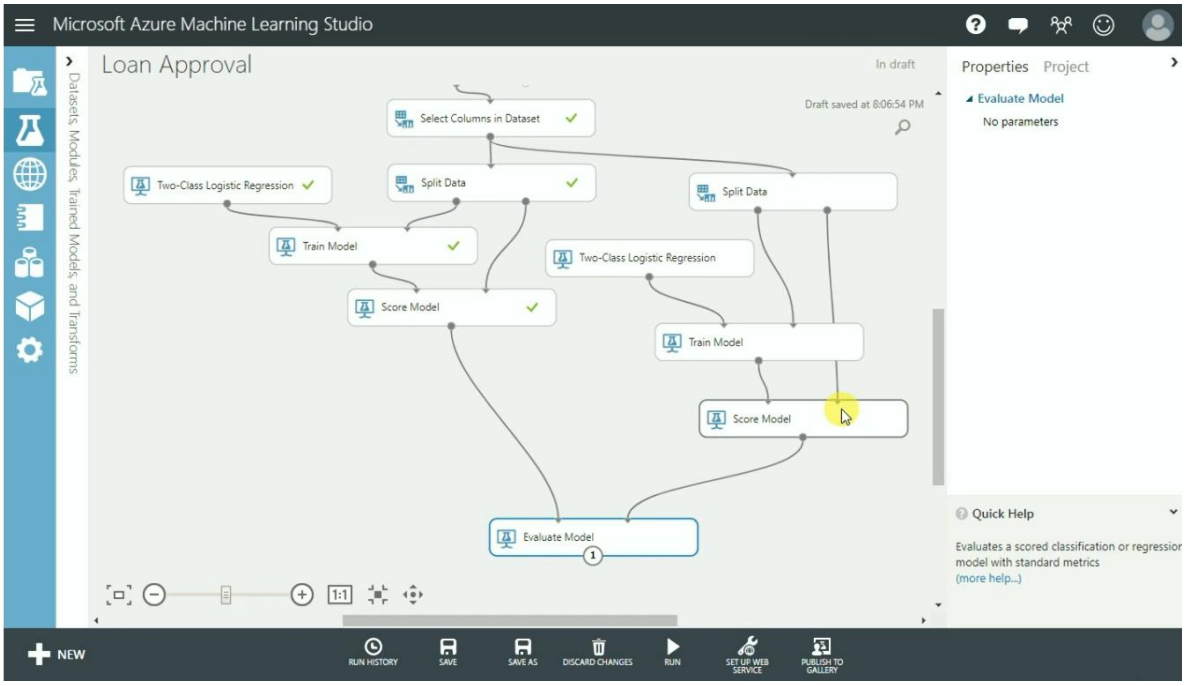
Change parameter as false in stratified split



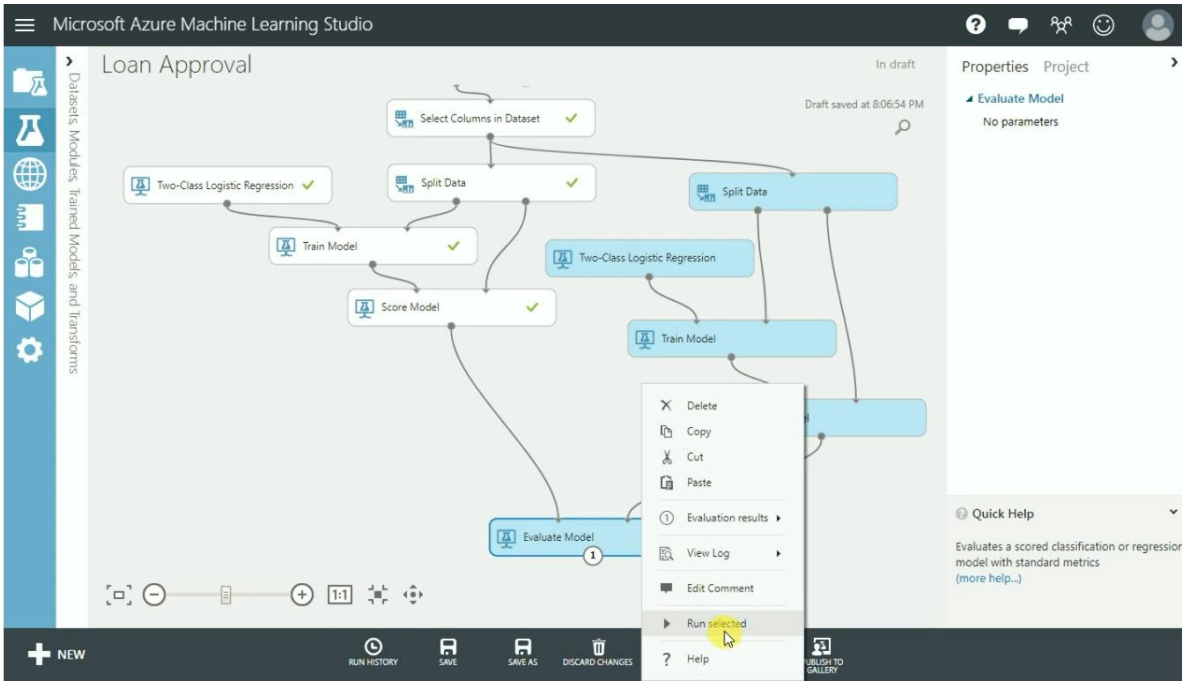
Connect output node from score model to input node of evaluate model



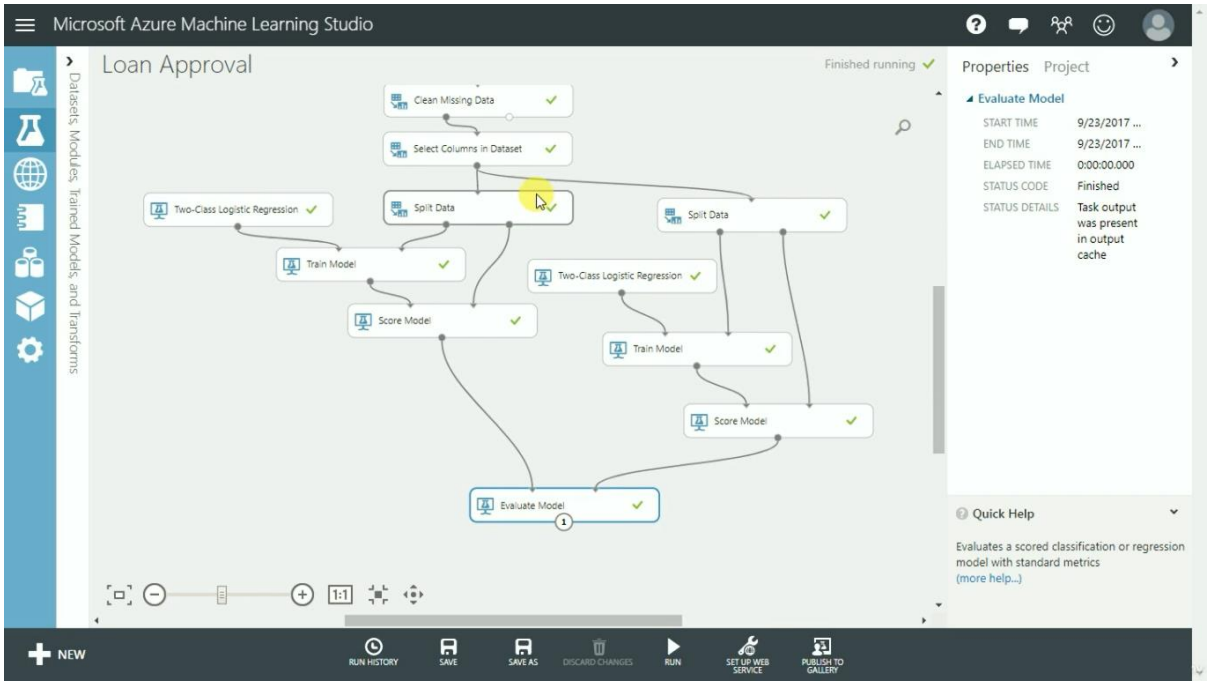
Connect split data output node 2 to score model input node 2



Now time to run the evaluate model



After successful execution visualize the evaluate model



After this action accuracy and auc has come down and thus found impact of stratification