Shayaan Hussain **19K41A04E6**

Artificial Intelligence

Assignment 4

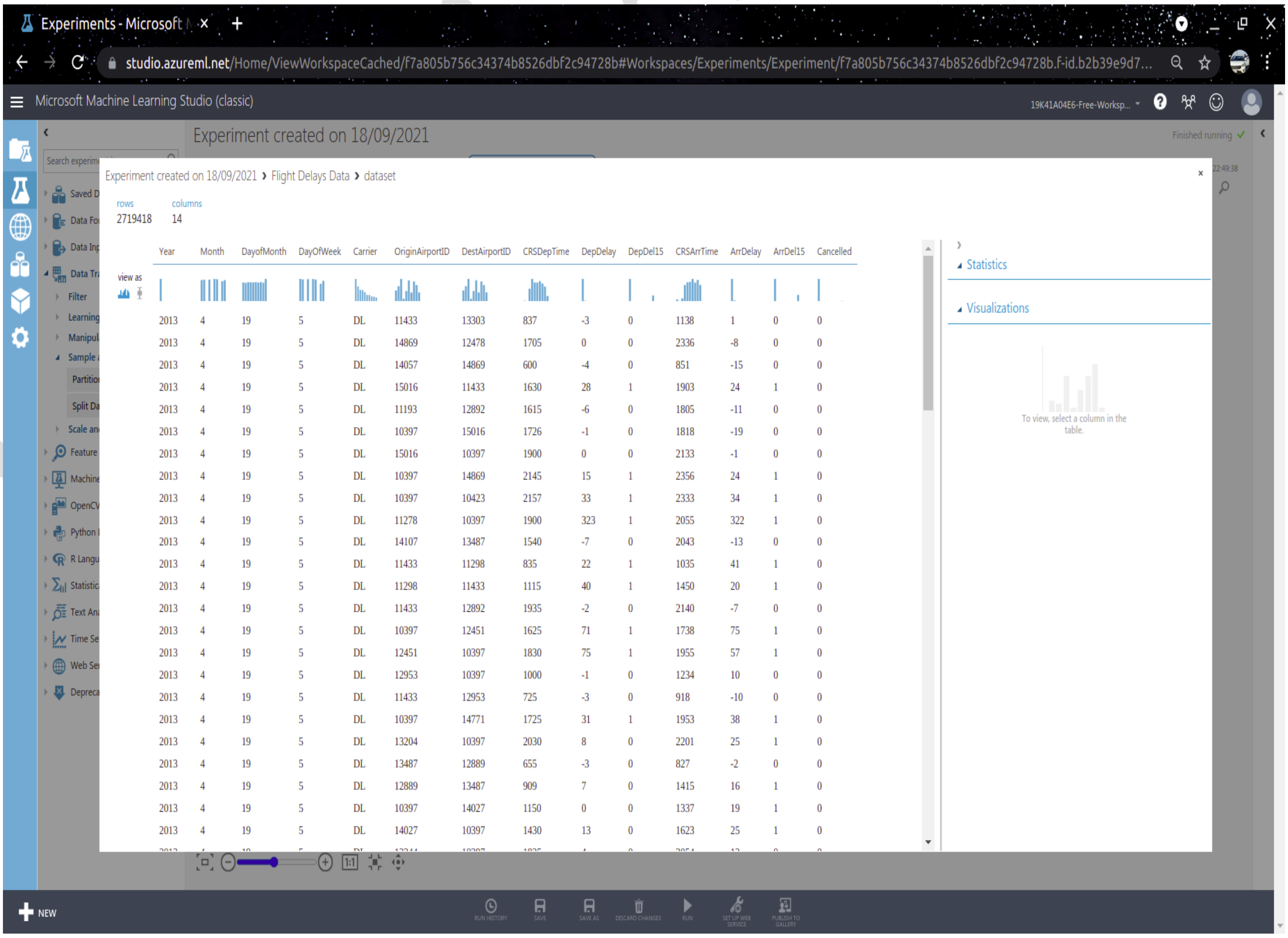
Software Used:- Microsoft Azure Machine Learning Studio

About the assignment:- The model used in this assignment is Two Class Logistic Regression Model. The trained model finds the probability of a flight to be delayed by more than 15 minutes based on certain parameters. The steps followed in the workflow are:

1. Load the data
2. Study the data and find out the necessary variables
3. Pre Process the data
4. Choose the model to be used (Two Class Logistic Regression Model in this case)
5. Split the data for training and testing
6. Train the model with the training data
7. Score the model using the testing data
8. Evaluate the model based on the results

Workflow:-

- Study the data



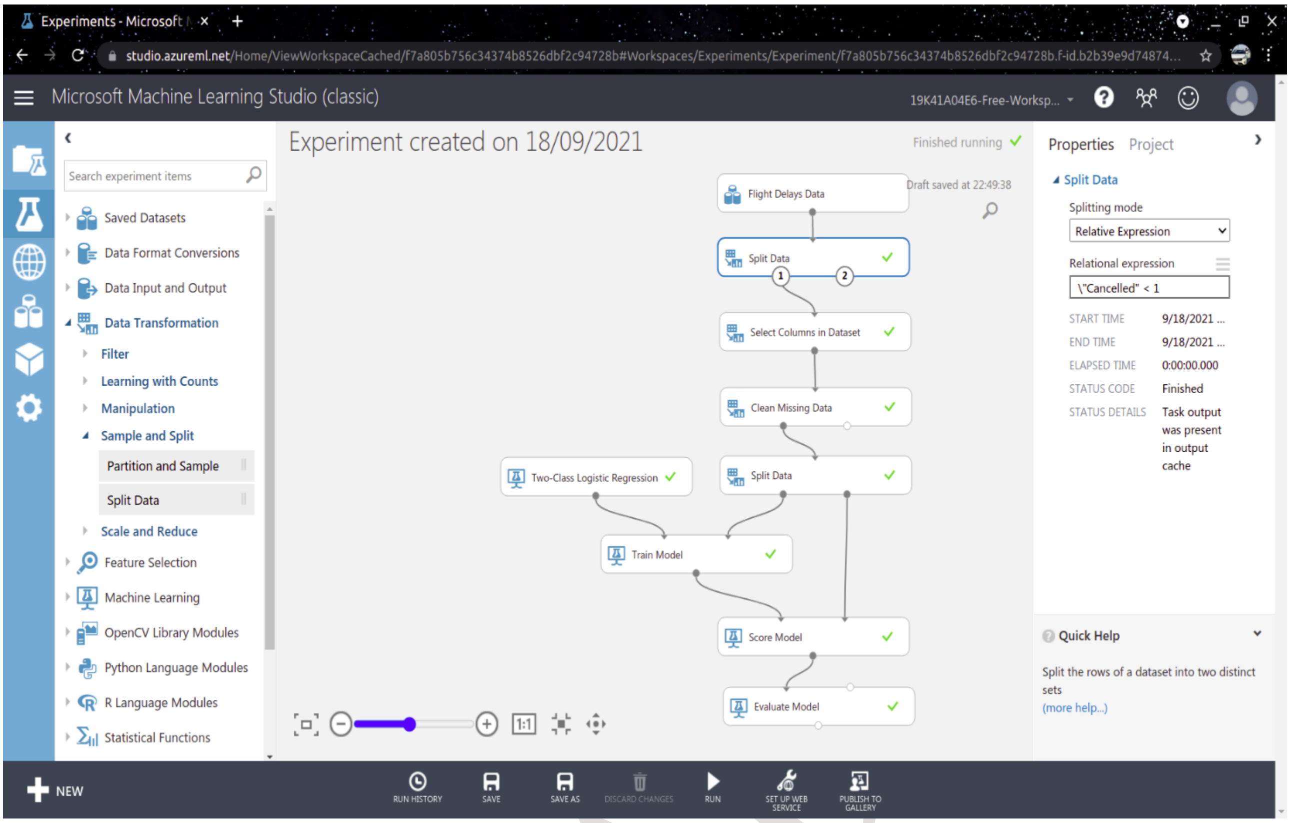
Experiment created on 18/09/2021 › Flight Delays Data › dataset

rows 2719418 columns 14

Year	Month	DayofMonth	DayOfWeek	Carrier	OriginAirportID	DestAirportID	CRSDepTime	DepDelay	DepDel15	CRSArrTime	ArrDelay	ArrDel15	Cancelled
2013	4	19	5	DL	11433	13303	837	-3	0	1138	1	0	0
2013	4	19	5	DL	14869	12478	1705	0	0	2336	-8	0	0
2013	4	19	5	DL	14057	14869	600	-4	0	851	-15	0	0
2013	4	19	5	DL	15016	11433	1630	28	1	1903	24	1	0
2013	4	19	5	DL	11193	12892	1615	-6	0	1805	-11	0	0
2013	4	19	5	DL	10397	15016	1726	-1	0	1818	-19	0	0
2013	4	19	5	DL	15016	10397	1900	0	0	2133	-1	0	0
2013	4	19	5	DL	10397	14869	2145	15	1	2356	24	1	0
2013	4	19	5	DL	10397	10423	2157	33	1	2333	34	1	0
2013	4	19	5	DL	11278	10397	1900	323	1	2055	322	1	0
2013	4	19	5	DL	14107	13487	1540	-7	0	2043	-13	0	0
2013	4	19	5	DL	11433	11298	835	22	1	1035	41	1	0
2013	4	19	5	DL	11298	11433	1115	40	1	1450	20	1	0
2013	4	19	5	DL	11433	12892	1935	-2	0	2140	-7	0	0
2013	4	19	5	DL	10397	12451	1625	71	1	1738	75	1	0
2013	4	19	5	DL	12451	10397	1830	75	1	1955	57	1	0
2013	4	19	5	DL	12953	10397	1000	-1	0	1234	10	0	0
2013	4	19	5	DL	11433	12953	725	-3	0	918	-10	0	0
2013	4	19	5	DL	10397	14771	1725	31	1	1953	38	1	0
2013	4	19	5	DL	13204	10397	2030	8	0	2201	25	1	0
2013	4	19	5	DL	13487	12889	655	-3	0	827	-2	0	0
2013	4	19	5	DL	12889	13487	909	7	0	1415	16	1	0
2013	4	19	5	DL	10397	14027	1150	0	0	1337	19	1	0
2013	4	19	5	DL	14027	10397	1430	13	0	1623	25	1	0

Experiments - Microsoft
studio.azureml.net/Home/ViewWorkspaceCached/f7a805b756c34374b8526dbf2c94728b#Workspaces/Experiments/Experiment/f7a805b756c34374b8526dbf2c94728b.f-id.b2b39e9d7...
Microsoft Machine Learning Studio (classic)
19K41A04E6-Free-Worksp...
Experiment created on 18/09/2021
Finished running ✓
view as
Statistics
Visualizations
To view, select a column in the table.
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY

- Remove all the cancelled flights from the dataset



Split Data

Splitting mode: Relative Expression

Relational expression: \"Cancelled" < 1

START TIME 9/18/2021 ...

END TIME 9/18/2021 ...

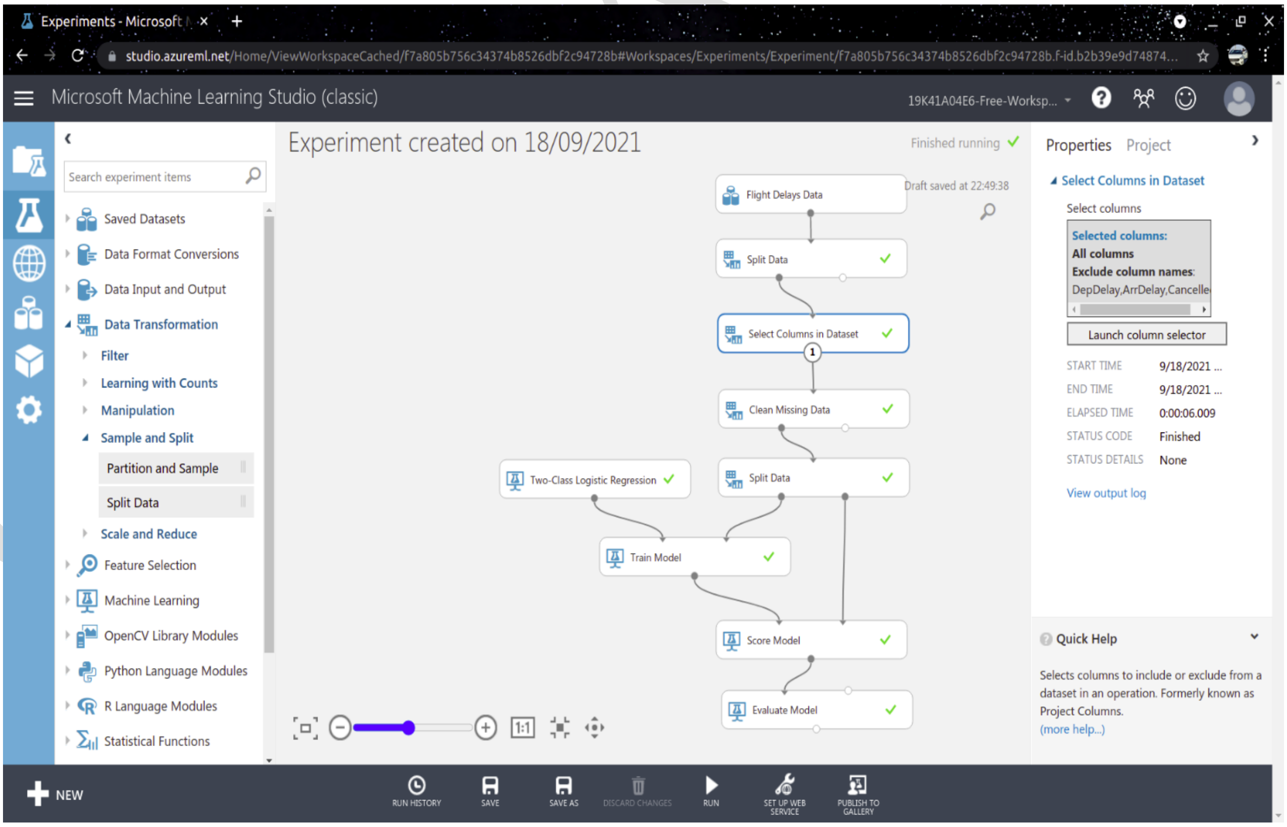
ELAPSED TIME 0:00:00.000

STATUS CODE Finished

STATUS DETAILS Task output was present in output cache

Quick Help: Split the rows of a dataset into two distinct sets (more help...)

- Drop the columns that are not needed for the training and testing



Select Columns in Dataset

Select columns

Selected columns:
All columns
Exclude column names: DepDelay,ArrDelay,Cancelle

START TIME 9/18/2021 ...

END TIME 9/18/2021 ...

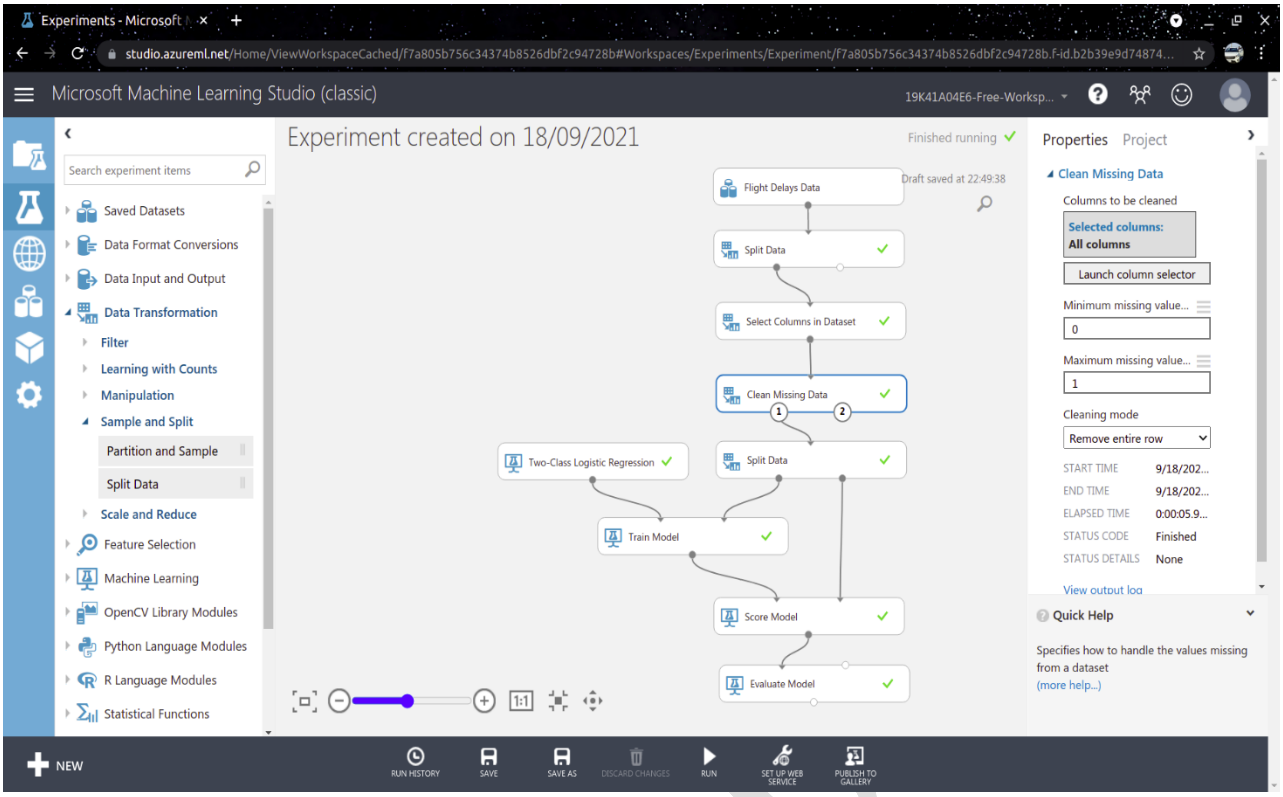
ELAPSED TIME 0:00:06.009

STATUS CODE Finished

STATUS DETAILS None

Quick Help: Selects columns to include or exclude from a dataset in an operation. Formerly known as Project Columns. (more help...)

- Remove rows with any NULL or missing values



Clean Missing Data

Columns to be cleaned

Selected columns:
All columns

Launch column selector

Minimum missing value...
0

Maximum missing value...
1

Cleaning mode
Remove entire row

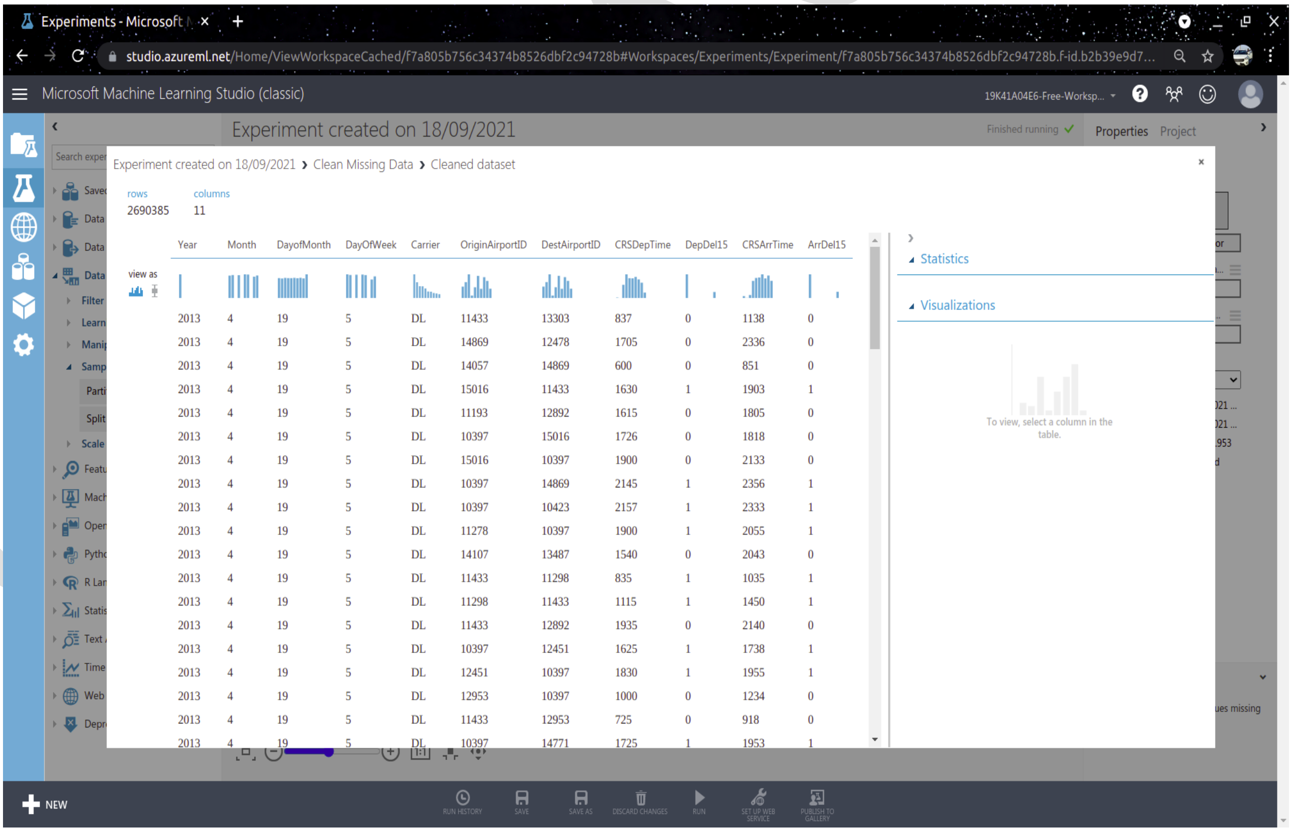
START TIME	9/18/202...
END TIME	9/18/202...
ELAPSED TIME	0:00:05.9...
STATUS CODE	Finished
STATUS DETAILS	None

View output log

Quick Help

Specifies how to handle the values missing from a dataset
(more help...)

- The dataset before splitting for training and testing is shown below

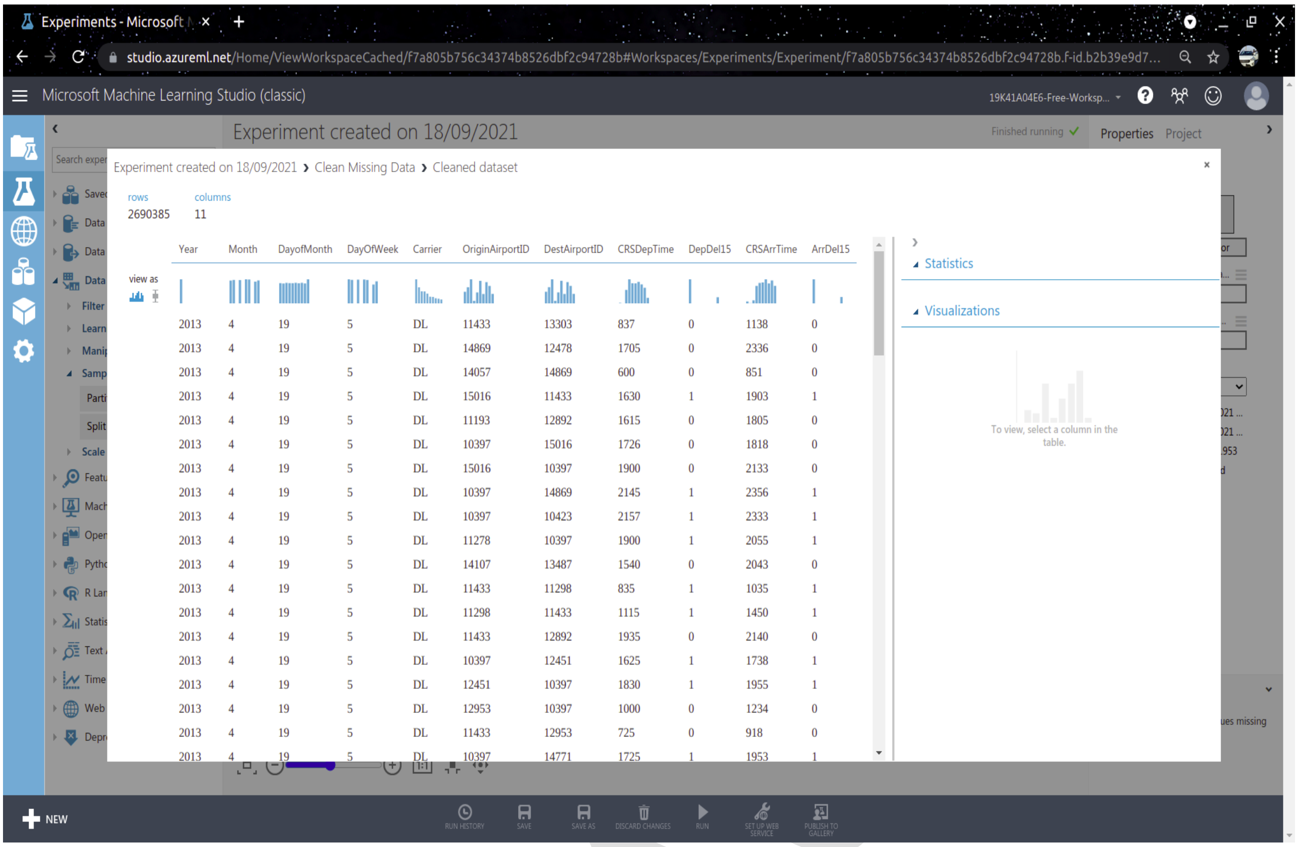

Experiments - Microsoft
studio.azureml.net/Home/ViewWorkspaceCached/f7a805b756c34374b8526dbf2c94728b#Workspaces/Experiments/Experiment/f7a805b756c34374b8526dbf2c94728b.f-id.b2b39e9d7...
Microsoft Machine Learning Studio (classic)
19K41A04E6-Free-Worksp...
Experiment created on 18/09/2021
Finished running ✓
Properties
Project
view as
Statistics
Visualizations
To view, select a column in the table.
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY

Experiment created on 18/09/2021 › Clean Missing Data › Cleaned dataset

rows 2690385
columns 11

Year	Month	DayofMonth	DayOfWeek	Carrier	OriginAirportID	DestAirportID	CRSDepTime	DepDel15	CRSArrTime	ArrDel15
2013	4	19	5	DL	11433	13303	837	0	1138	0
2013	4	19	5	DL	14869	12478	1705	0	2336	0
2013	4	19	5	DL	14057	14869	600	0	851	0
2013	4	19	5	DL	15016	11433	1630	1	1903	1
2013	4	19	5	DL	11193	12892	1615	0	1805	0
2013	4	19	5	DL	10397	15016	1726	0	1818	0
2013	4	19	5	DL	15016	10397	1900	0	2133	0
2013	4	19	5	DL	10397	14869	2145	1	2356	1
2013	4	19	5	DL	10397	10423	2157	1	2333	1
2013	4	19	5	DL	11278	10397	1900	1	2055	1
2013	4	19	5	DL	14107	13487	1540	0	2043	0
2013	4	19	5	DL	11433	11298	835	1	1035	1
2013	4	19	5	DL	11298	11433	1115	1	1450	1
2013	4	19	5	DL	11433	12892	1935	0	2140	0
2013	4	19	5	DL	10397	12451	1625	1	1738	1
2013	4	19	5	DL	12451	10397	1830	1	1955	1
2013	4	19	5	DL	12953	10397	1000	0	1234	0
2013	4	19	5	DL	11433	12953	725	0	918	0
2013	4	19	5	DL	10397	14771	1725	1	1953	1

- The data is split 50-50 for training and testing


Experiments - Microsoft
studio.azureml.net/Home/ViewWorkspaceCached/f7a805b756c34374b8526dbf2c94728b#Workspaces/Experiments/Experiment/f7a805b756c34374b8526dbf2c94728b.f-id.b2b39e9d7...
Microsoft Machine Learning Studio (classic)
19K41A04E6-Free-Worksp...
Experiment created on 18/09/2021
Finished running
Properties
Project
Statistics
Visualizations
To view, select a column in the table.
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY

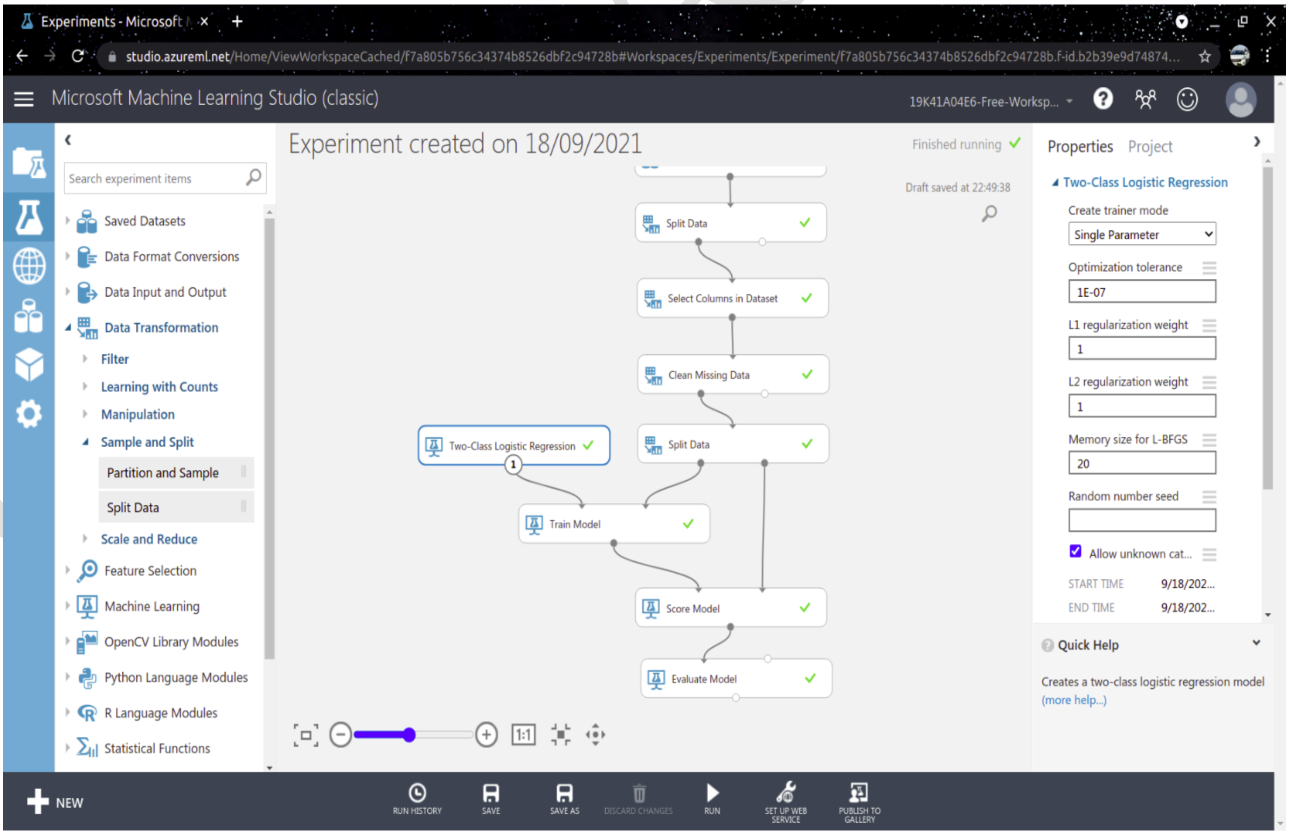
Experiment created on 18/09/2021 › Clean Missing Data › Cleaned dataset

rows 2690385

columns 11

Year	Month	DayofMonth	DayOfWeek	Carrier	OriginAirportID	DestAirportID	CRSDepTime	DepDel15	CRSArrTime	ArrDel15
2013	4	19	5	DL	11433	13303	837	0	1138	0
2013	4	19	5	DL	14869	12478	1705	0	2336	0
2013	4	19	5	DL	14057	14869	600	0	851	0
2013	4	19	5	DL	15016	11433	1630	1	1903	1
2013	4	19	5	DL	11193	12892	1615	0	1805	0
2013	4	19	5	DL	10397	15016	1726	0	1818	0
2013	4	19	5	DL	15016	10397	1900	0	2133	0
2013	4	19	5	DL	10397	14869	2145	1	2356	1
2013	4	19	5	DL	10397	10423	2157	1	2333	1
2013	4	19	5	DL	11278	10397	1900	1	2055	1
2013	4	19	5	DL	14107	13487	1540	0	2043	0
2013	4	19	5	DL	11433	11298	835	1	1035	1
2013	4	19	5	DL	11298	11433	1115	1	1450	1
2013	4	19	5	DL	11433	12892	1935	0	2140	0
2013	4	19	5	DL	10397	12451	1625	1	1738	1
2013	4	19	5	DL	12451	10397	1830	1	1955	1
2013	4	19	5	DL	12953	10397	1000	0	1234	0
2013	4	19	5	DL	11433	12953	725	0	918	0
2013	4	19	5	DL	10397	14771	1725	1	1953	1

- The model is initialized



Properties Project

Two-Class Logistic Regression

Create trainer mode: Single Parameter

Optimization tolerance: 1E-07

L1 regularization weight: 1

L2 regularization weight: 1

Memory size for L-BFGS: 20

Random number seed:

☑ Allow unknown cat...

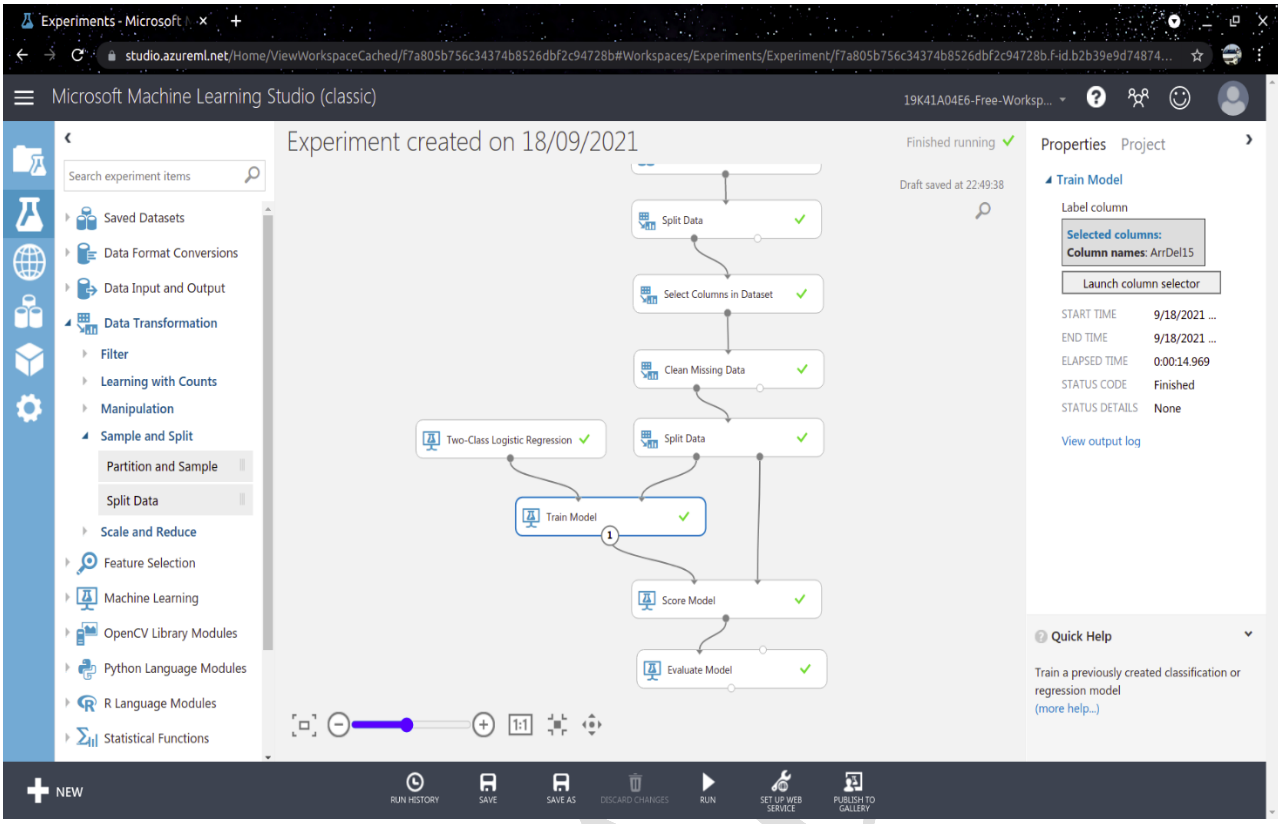
START TIME 9/18/202...

END TIME 9/18/202...

Quick Help

Creates a two-class logistic regression model (more help...)

- Train the model using the training data



Experiment created on 18/09/2021

Finished running ✓

Draft saved at 22:49:38

Properties Project

Train Model

Label column

Selected columns:
Column names: ArrDel15

Launch column selector

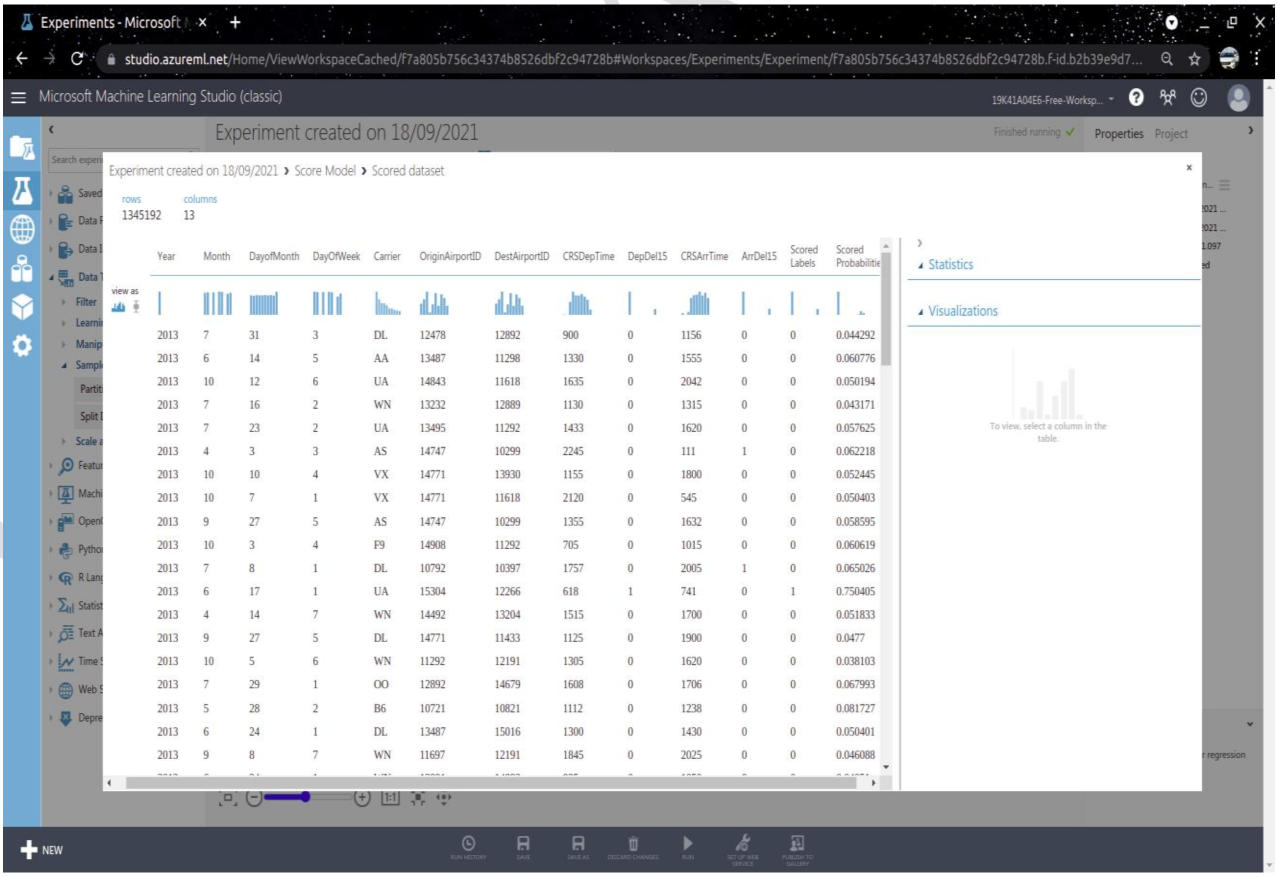
START TIME	9/18/2021 ...
END TIME	9/18/2021 ...
ELAPSED TIME	0:00:14.969
STATUS CODE	Finished
STATUS DETAILS	None

View output log

Quick Help

Train a previously created classification or regression model
(more help...)

- The scored dataset is shown below



Experiment created on 18/09/2021 › Score Model › Scored dataset

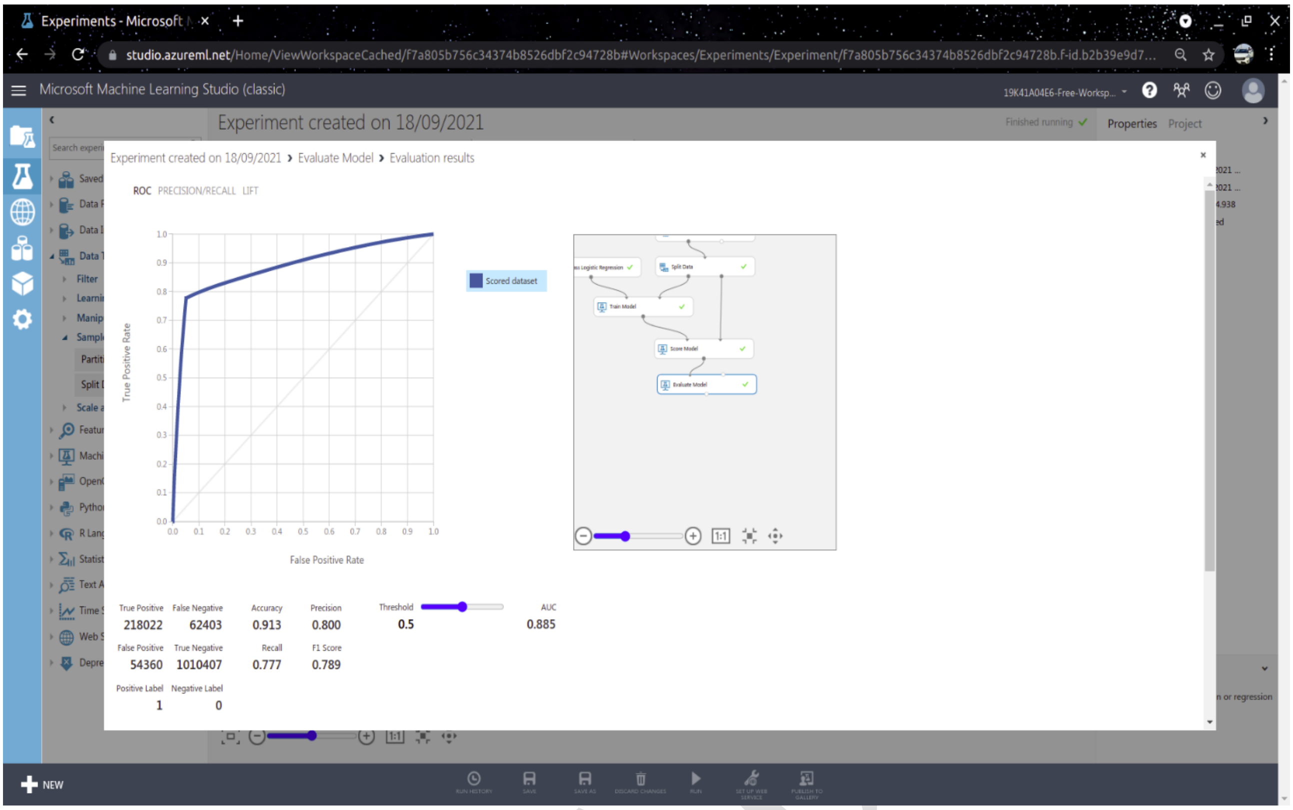
rows	columns
1345192	13

Year	Month	DayofMonth	DayOfWeek	Carrier	OriginAirportID	DestAirportID	CRSDepTime	DepDel15	CRSArrTime	ArrDel15	Scored Labels	Scored Probabilitie
2013	7	31	3	DL	12478	12892	900	0	1156	0	0	0.044292
2013	6	14	5	AA	13487	11298	1330	0	1555	0	0	0.060776
2013	10	12	6	UA	14843	11618	1635	0	2042	0	0	0.050194
2013	7	16	2	WN	13232	12889	1130	0	1315	0	0	0.043171
2013	7	23	2	UA	13495	11292	1433	0	1620	0	0	0.057625
2013	4	3	3	AS	14747	10299	2245	0	111	1	0	0.062218
2013	10	10	4	VX	14771	13930	1155	0	1800	0	0	0.052445
2013	10	7	1	VX	14771	11618	2120	0	545	0	0	0.050403
2013	9	27	5	AS	14747	10299	1355	0	1632	0	0	0.058595
2013	10	3	4	F9	14908	11292	705	0	1015	0	0	0.060619
2013	7	8	1	DL	10792	10397	1757	0	2005	1	0	0.065026
2013	6	17	1	UA	15304	12266	618	1	741	0	1	0.750405
2013	4	14	7	WN	14492	13204	1515	0	1700	0	0	0.051833
2013	9	27	5	DL	14771	11433	1125	0	1900	0	0	0.0477
2013	10	5	6	WN	11292	12191	1305	0	1620	0	0	0.038103
2013	7	29	1	OO	12892	14679	1608	0	1706	0	0	0.067993
2013	5	28	2	B6	10721	10821	1112	0	1238	0	0	0.081727
2013	6	24	1	DL	13487	15016	1300	0	1430	0	0	0.050401
2013	9	8	7	WN	11697	12191	1845	0	2025	0	0	0.046088

To view, select a column in the table.

Experiments - Microsoft
studio.azureml.net/Home/ViewWorkspaceCached/f7a805b756c34374b8526dbf2c94728b#Workspaces/Experiments/Experiment/f7a805b756c34374b8526dbf2c94728b.f-id.b2b39e9d7...
Microsoft Machine Learning Studio (classic)
19K41A04E6-Free-Worksp...
Experiment created on 18/09/2021
Finished running ✓
Properties Project
view as
Statistics
Visualizations
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY

- The evaluation of model is done and the results are as follows



Experiment created on 18/09/2021 › Evaluate Model › Evaluation results

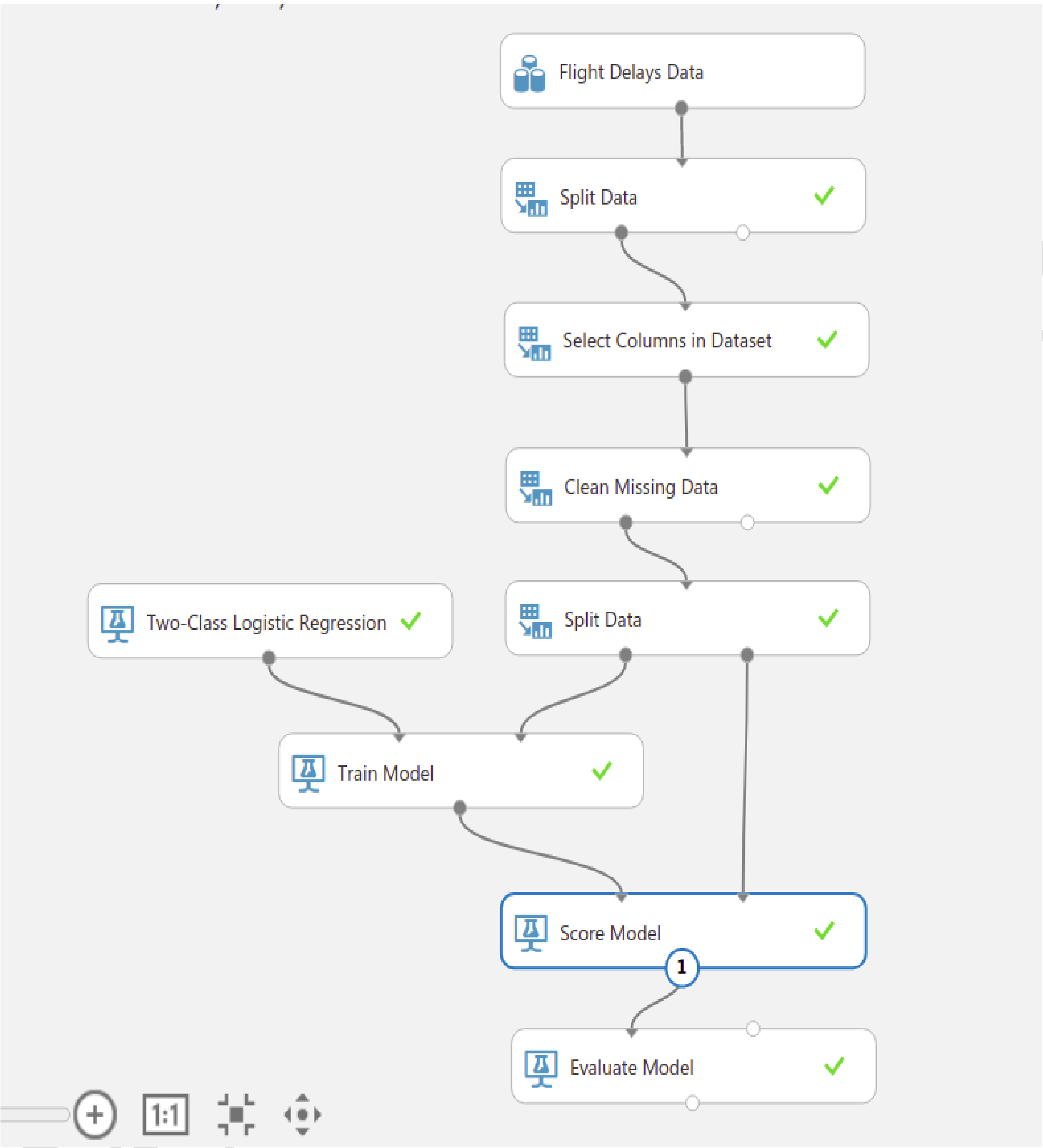
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
218022	62403	0.913	0.800	0.5	0.885
False Positive	**True Negative**	**Recall**	**F1 Score**		
54360	1010407	0.777	0.789		
Positive Label	**Negative Label**				
1	0				

Conclusion: From the evaluation, we can see that the accuracy of the model is 91.3% and the model is able to predict the outcome properly in most cases. The parameters used to predict the delays are:

1. Year of flight
2. Month of flight
3. Date of flight
4. Day of the week of the flight
5. The Carrier/Airline
6. Origin Airport
7. Destination Airport
8. Scheduled Departure Time
9. Departure Delay (atleast 15 minutes)
10. Scheduled Arrival Time
11. Arrival Delay (atleast 15 minutes)

This means that on giving the first 10 parameters as input, the model can predicts the probability of the flight getting delayed by more than 15 minutes for arriving at the destination airport.

The final workflow is as follows:


Flight Delays Data
Split Data
Select Columns in Dataset
Clean Missing Data
Two-Class Logistic Regression
Split Data
Train Model
Score Model
1
Evaluate Model
1:1