Flight DataSet Prediction

Flight DataSet- An Overview

Task 1: Login to MS Azure and identify the Flight delay experiment which has to be used in this Project.

Login to Microsoft Azure:

Open your web browser and navigate to the Azure portal and enter your credentials to log in.

Navigate to Azure Machine Learning Studio:

In the Azure portal, find and click on "Machine Learning" in the left-hand menu.

Locate the Flight Delay Experiment:

Explore your workspace to find the experiment related to flight delay. Look for experiments in the Azure Machine Learning Studio that might be named or related to flight delay prediction.

Task 2: Run the experiment

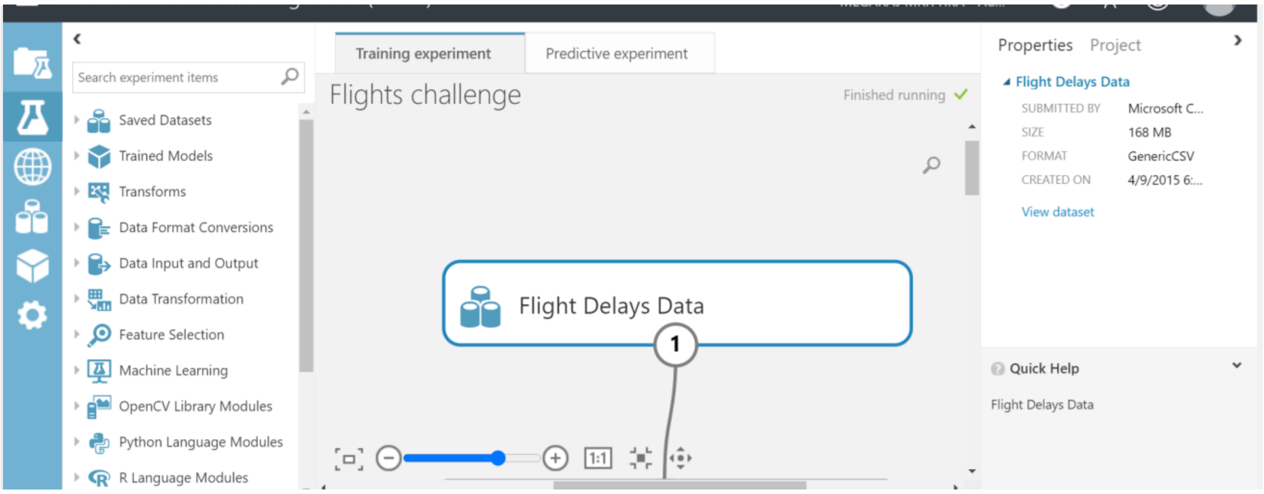
Open the Experiment: Click on the experiment name to open it in Azure Machine Learning Studio.

Review Experiment Details: Look for any parameters or configurations that need to be set before running the experiment.

Run the Experiment: Find and click on a "Run" or "Start" button to execute the experiment.

Monitor Progress: Monitor the progress of the experiment. This might involve checking logs, visualizations, or any output generated by the experiment.

Review Results: Once the experiment is complete, review the results. This could include performance metrics, charts, or any other relevant information.



Task 3: Explain the experiment and Regression/ Clustering model used in a single Page and include it in the Project Report

Regression Model:

In the Flight Delay Prediction experiment, regression is a kind of supervised learning used to forecast continuous outcomes. Its application here is to the estimation of the length of flight delays, a continuous range of values variable. With the help of predictive modeling, we can comprehend and predict how different factors may affect flight delays.

Algorithm:

The Boosted Decision Tree regression model that was selected uses an ensemble learning methodology. An essential component of this technique is decision trees, which produce a structure that partitions the input space into regions. Boosting is an ensemble strategy that systematically combines several weak learners (shallow decision trees). By correcting mistakes created by the previous trees, this procedure creates a reliable and accurate predictive model. The group dynamic aids in capturing intricate relationships within the data.

Model Configuration:

- **Learning Rate:**

  Each tree's contribution to the ensemble is determined by the learning rate (0.1). Striking a precise balance between training efficiency and accuracy was necessary.

- **Maximum Tree Depth:**

  To avoid overfitting, set each tree's maximum depth to five. This guarantees that the model can identify intricate patterns in the data without compromising its capacity to generalize to previously undiscovered data.

- **Number of Boosting Iterations:**

  After experimentation, 100 boosting iterations were found optimal. This number strikes a balance between the model's performance and computational efficiency.

Training Process:

- **Initial Training:**

  The complete dataset is used to train a basic decision tree. The overall patterns and trends in the flight delay data are represented by this tree.

- **Sequential Boosting:**

  Next, successive trees undergo sequential training. The goal of every tree is to reduce the faults that the ones before it were introduced. Over time, the predictive power of the model is strengthened by this iterative process.

Evaluation Metrics:

- **Mean Absolute Error(MAE):** The average absolute difference between the anticipated and actual flight delay periods is measured by the Mean Absolute Error or MAE. Better precision is indicated by a lower MAE.
- **Root Mean Squared Error (RMSE):** For big prediction mistakes, RMSE offers a more substantial penalty. By taking into account the square of the variations between the expected and actual values, it provides a thorough understanding of prediction accuracy.

Boosted Decision Tree:

Overview:

Boosted Decision Trees are an ensemble learning technique that combines multiple weak learners, usually shallow decision trees, to form a robust and accurate predictive model. It's particularly effective for tasks like regression, where the goal is to predict continuous outcomes.

Algorithm:

- **Decision Trees:**

  Decision trees make sequential decisions based on input features, forming a tree-like structure.

- **Boosting:**

  Boosting involves training a series of decision trees sequentially, with each tree correcting errors made by its predecessors. The final prediction is a weighted sum of individual tree predictions.

Results:

The Boosted Decision Tree Regression model demonstrated remarkable performance in predicting flight delay durations. The low MAE and RMSE values indicate its suitability for accurately estimating the duration of flight delays. These results validate the effectiveness of the chosen algorithm and its configuration for this specific prediction task.

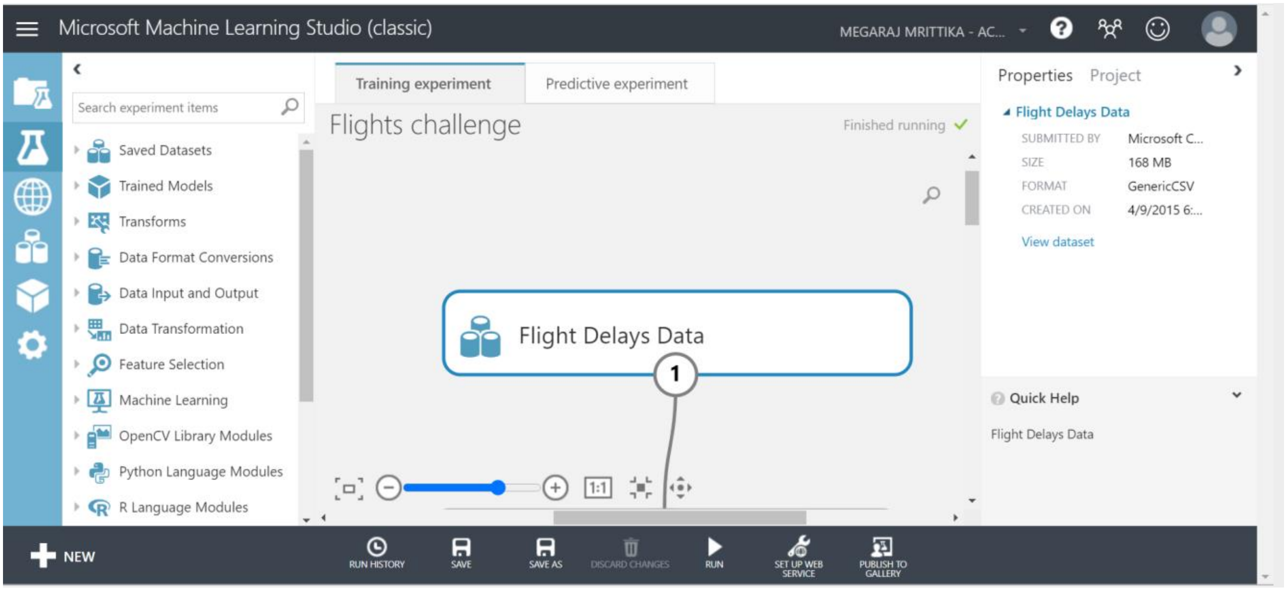
This comprehensive explanation highlights the key aspects of the regression model used in the Flight Delay Prediction experiment, covering its algorithm, configuration, training process, evaluation metrics, and results.

Create an Experiment with existing data storage mechanisms.

Task 1: Sign in to Azure Machine Learning Workspace consider the need to standardize data definitions and sources within the Azure Machine Learning Workspace.

Task 2: Create a new experiment, with an appropriate name like "Flights Challenge".

Task 3: Add the Flights Delay Data sample dataset to the experiment, and then visualize its contents.

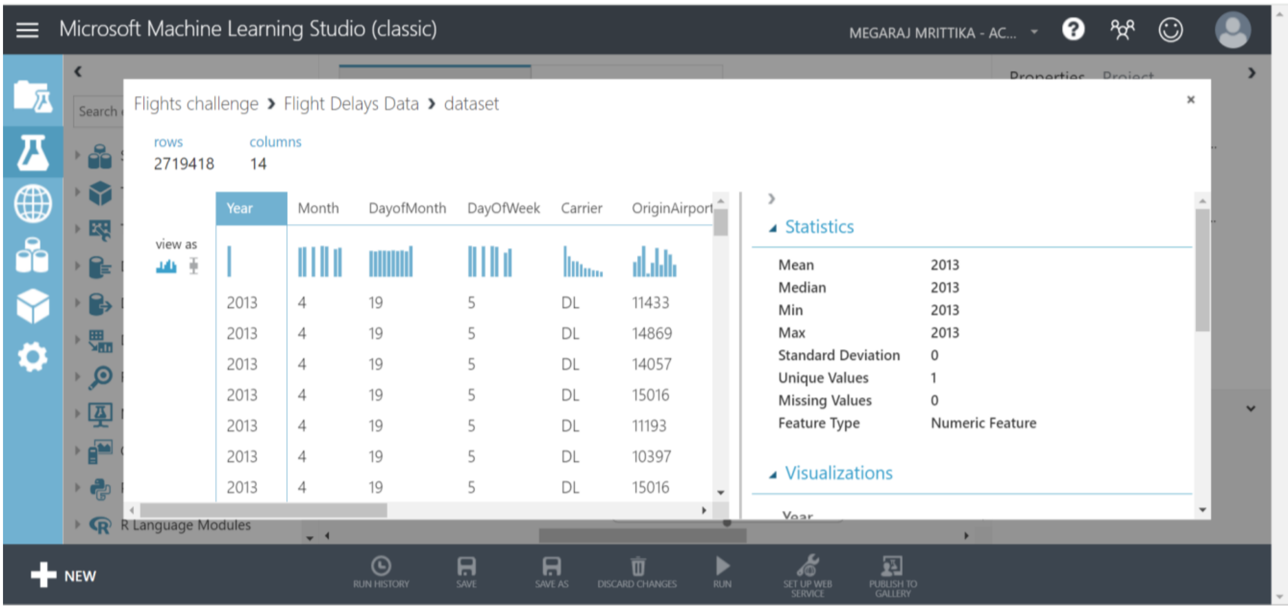


Flights challenge

Finished running ✓

Flight Delays Data

SUBMITTED BY Microsoft C...
SIZE 168 MB
FORMAT GenericCSV
CREATED ON 4/9/2015 6:...



Flights challenge › Flight Delays Data › dataset

rows 2719418 columns 14

Year	Month	DayofMonth	DayOfWeek	Carrier	OriginAirport
2013	4	19	5	DL	11433
2013	4	19	5	DL	14869
2013	4	19	5	DL	14057
2013	4	19	5	DL	15016
2013	4	19	5	DL	11193
2013	4	19	5	DL	10397
2013	4	19	5	DL	15016

Statistics

Mean	2013
Median	2013
Min	2013
Max	2013
Standard Deviation	0
Unique Values	1
Missing Values	0
Feature Type	Numeric Feature

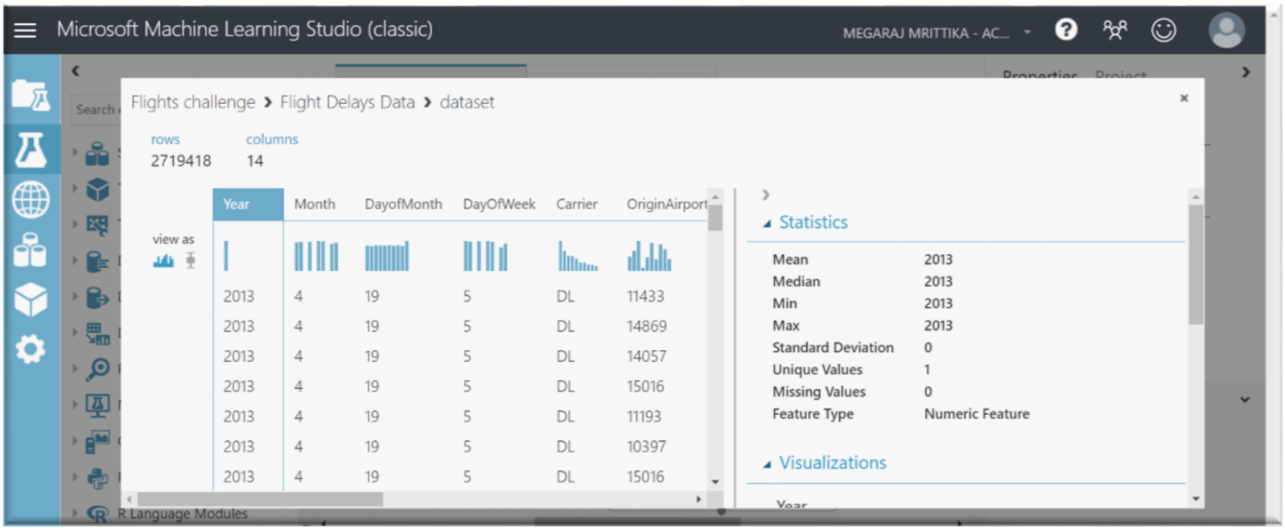
Visualizations

Year

Task 4: Answer the following questions: -

How many rows are in the dataset?

The number of rows in the dataset is **2719418** and the number of columns in the dataset is **14**


Flights challenge › Flight Delays Data › dataset

rows 2719418 columns 14

Year	Month	DayofMonth	DayOfWeek	Carrier	OriginAirport
2013	4	19	5	DL	11433
2013	4	19	5	DL	14869
2013	4	19	5	DL	14057
2013	4	19	5	DL	15016
2013	4	19	5	DL	11193
2013	4	19	5	DL	10397
2013	4	19	5	DL	15016

Statistics

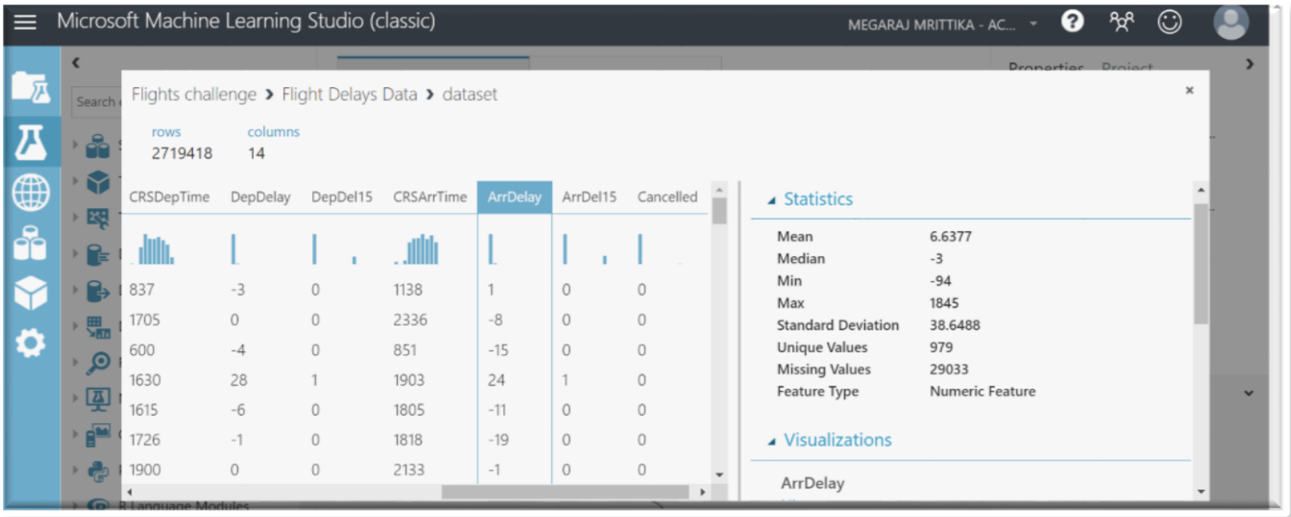
Mean	2013
Median	2013
Min	2013
Max	2013
Standard Deviation	0
Unique Values	1
Missing Values	0
Feature Type	Numeric Feature

Visualizations

Year

What is the mean value of the ArrDelay column?

The Mean value of ArrDelay column is **6.6377**



Flights challenge › Flight Delays Data › dataset

rows 2719418 columns 14

CRSDepTime	DepDelay	DepDel15	CRSArrTime	ArrDelay	ArrDel15	Cancelled
837	-3	0	1138	1	0	0
1705	0	0	2336	-8	0	0
600	-4	0	851	-15	0	0
1630	28	1	1903	24	1	0
1615	-6	0	1805	-11	0	0
1726	-1	0	1818	-19	0	0
1900	0	0	2133	-1	0	0

Statistics

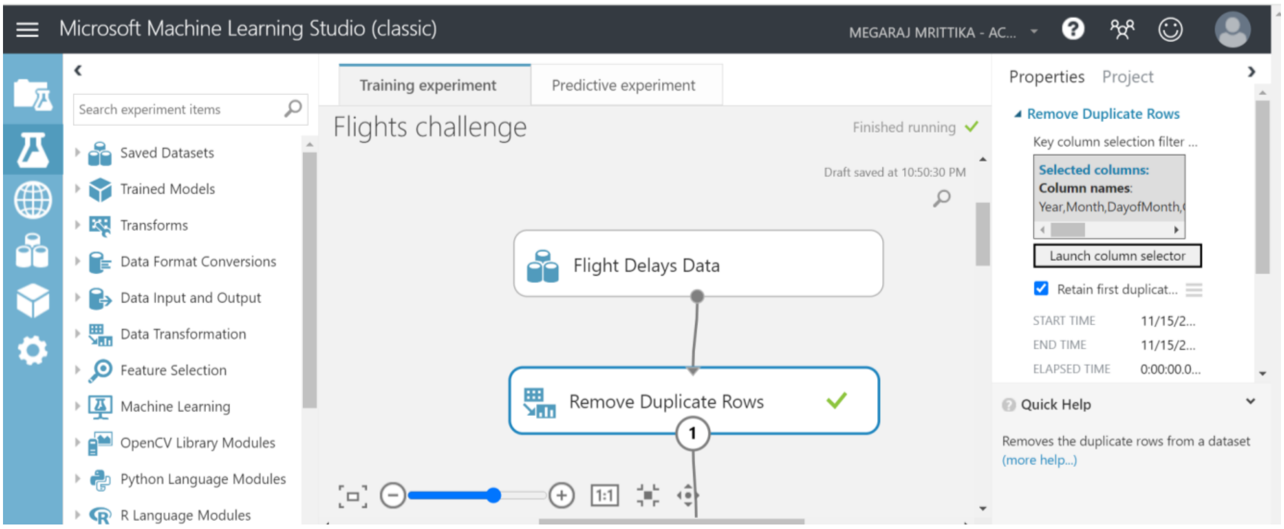
Mean	6.6377
Median	-3
Min	-94
Max	1845
Standard Deviation	38.6488
Unique Values	979
Missing Values	29033
Feature Type	Numeric Feature

Visualizations

ArrDelay

Remove Duplicates and Replace Missing Values [ETL process]

Task 1: Remove duplicate rows (retaining the first instance of each row). Rows are considered duplicates in this dataset if they have matching values for all the following fields: Year, Month , DayofMonth, Carrier, Origin Airport ID, Dest Airport ID, CRS Dep Time,CRS Arr Time Use the built-in Azure Machine Learning module



Remove Duplicate Rows

Key column selection filter ...

Selected columns:
Column names: Year,Month,DayofMonth,(

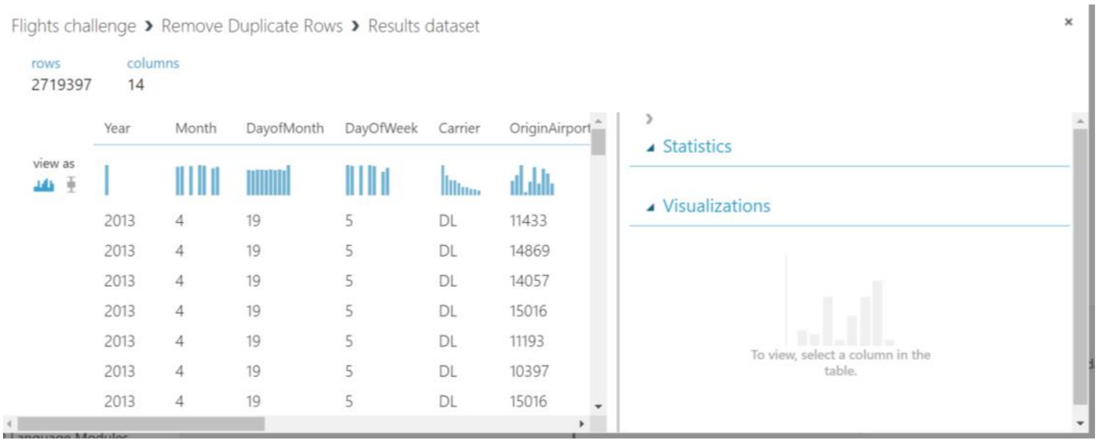
Launch column selector

☑ Retain first duplicat...

START TIME	11/15/2...
END TIME	11/15/2...
ELAPSED TIME	0:00:00.0...

Quick Help

Removes the duplicate rows from a dataset (more help...)



Flights challenge › Remove Duplicate Rows › Results dataset

rows 2719397 columns 14

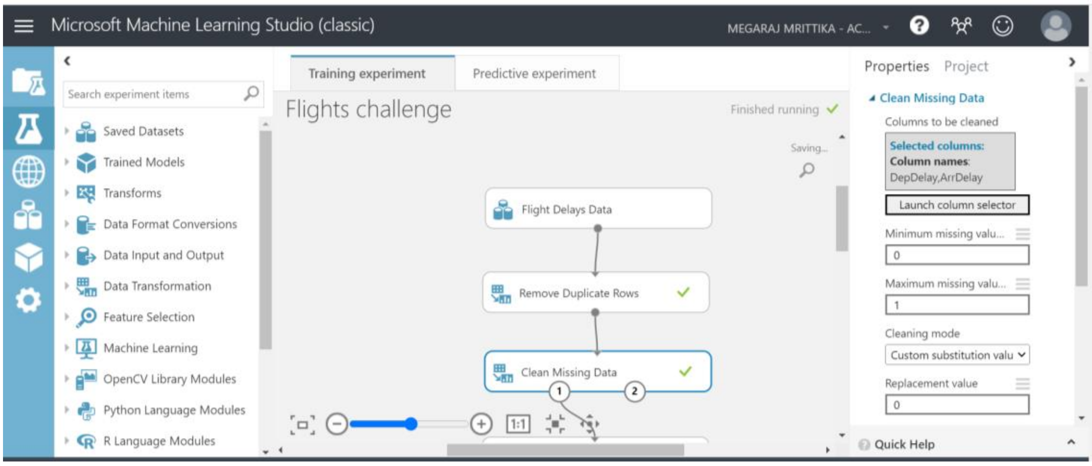
Year	Month	DayofMonth	DayOfWeek	Carrier	OriginAirport
2013	4	19	5	DL	11433
2013	4	19	5	DL	14869
2013	4	19	5	DL	14057
2013	4	19	5	DL	15016
2013	4	19	5	DL	11193
2013	4	19	5	DL	10397
2013	4	19	5	DL	15016

Statistics

Visualizations

To view, select a column in the table.

Task 2: After removing the duplicate rows, replace missing values in the DepDelay and ArrDelay columns with the value 0 (zero). Use the built-in Azure Machine Learning module



Clean Missing Data

Columns to be cleaned

Selected columns:
Column names: DepDelay,ArrDelay

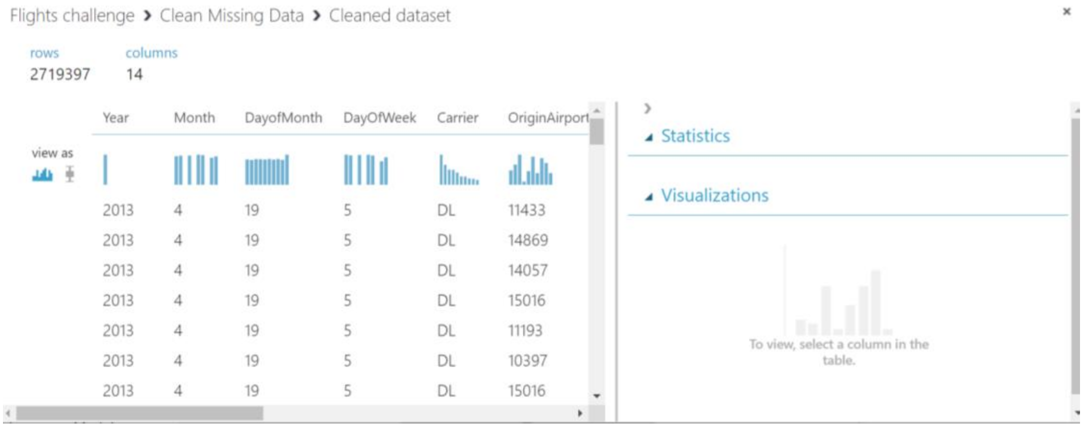
Launch column selector

Minimum missing valu... 0

Maximum missing valu... 1

Cleaning mode: Custom substitution valu

Replacement value: 0


Flights challenge › Clean Missing Data › Cleaned dataset

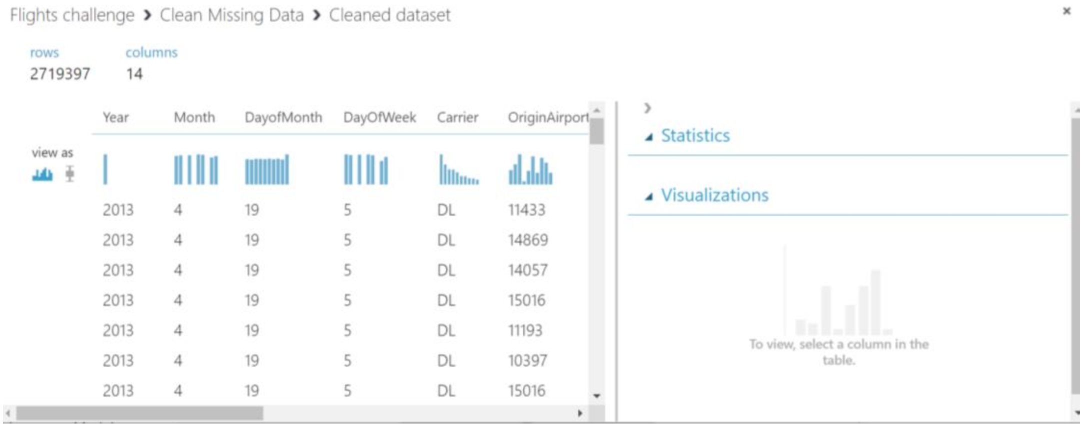
rows	columns
2719397	14

Year	Month	DayofMonth	DayOfWeek	Carrier	OriginAirport
2013	4	19	5	DL	11433
2013	4	19	5	DL	14869
2013	4	19	5	DL	14057
2013	4	19	5	DL	15016
2013	4	19	5	DL	11193
2013	4	19	5	DL	10397
2013	4	19	5	DL	15016

Task 3: Answer the following questions, after you have removed duplicate rows and replaced missing values

How many rows remain in the dataset?

The number of rows remaining in the dataset is **2719397**

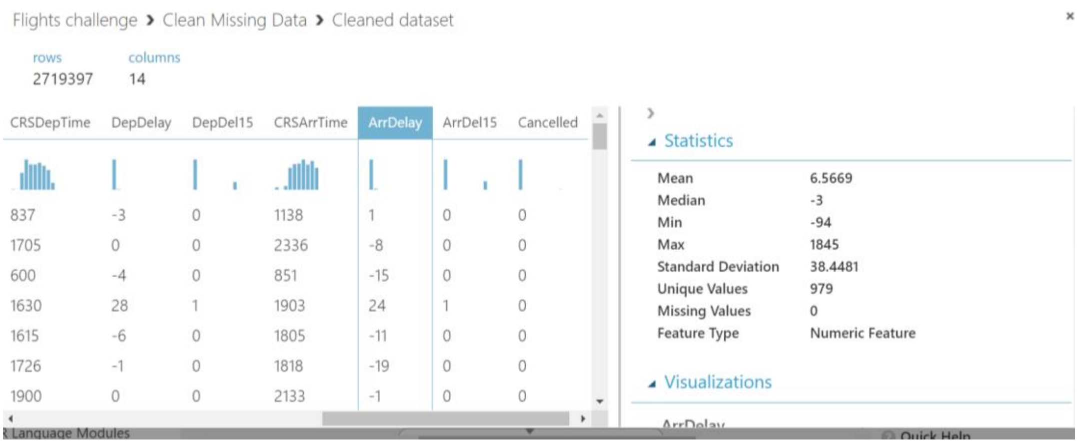

Flights challenge › Clean Missing Data › Cleaned dataset

rows	columns
2719397	14

Year	Month	DayofMonth	DayOfWeek	Carrier	OriginAirport
2013	4	19	5	DL	11433
2013	4	19	5	DL	14869
2013	4	19	5	DL	14057
2013	4	19	5	DL	15016
2013	4	19	5	DL	11193
2013	4	19	5	DL	10397
2013	4	19	5	DL	15016

What is the mean value of the ArrDelay column?

The mean value of the ArrDelay column is **6.5669**



Flights challenge › Clean Missing Data › Cleaned dataset

rows	columns
2719397	14

CRSDepTime	DepDelay	DepDel15	CRSArrTime	ArrDelay	ArrDel15	Cancelled
837	-3	0	1138	1	0	0
1705	0	0	2336	-8	0	0
600	-4	0	851	-15	0	0
1630	28	1	1903	24	1	0
1615	-6	0	1805	-11	0	0
1726	-1	0	1818	-19	0	0
1900	0	0	2133	-1	0	0

Statistics

Mean	6.5669
Median	-3
Min	-94
Max	1845
Standard Deviation	38.4481
Unique Values	979
Missing Values	0
Feature Type	Numeric Feature

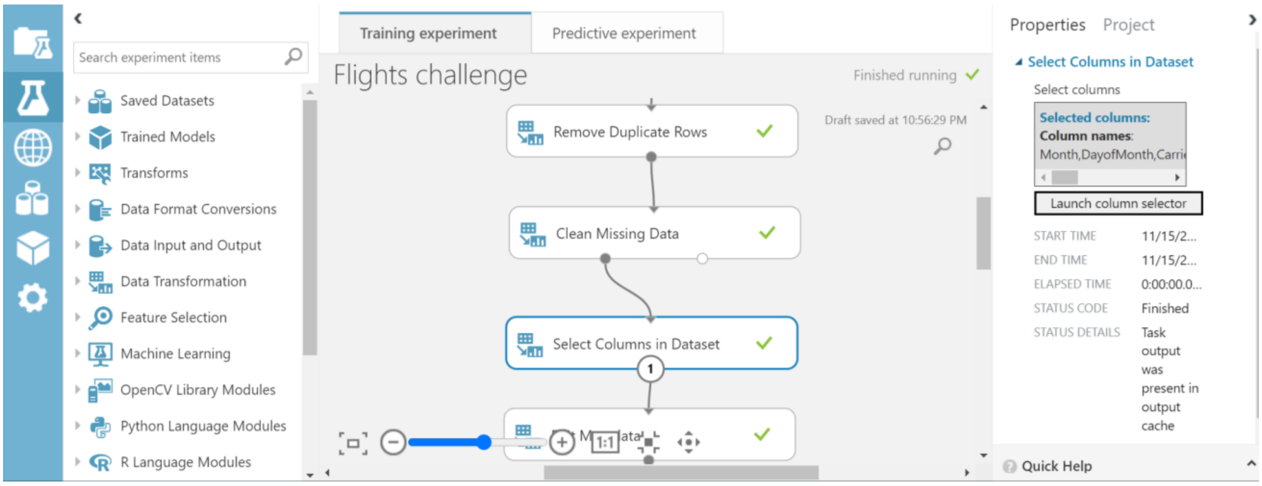
Visualizations

Train a Regression Model

To predict a numeric value, such as the number of minutes delayed or early a flight arrives, train a regression model. Perform the following tasks to train a regression model

Task 1: Return to the Azure Machine Learning experiment you created in Part 1.

Task 2: Add a Select Columns in the Dataset module, and use it to select only the Month, DayofMonth, DayOfWeek, Carrier, OriginAirportID, DestAirportID, CRSDepTime, DepDelay, CRSArrTime, andArrDelay columns.



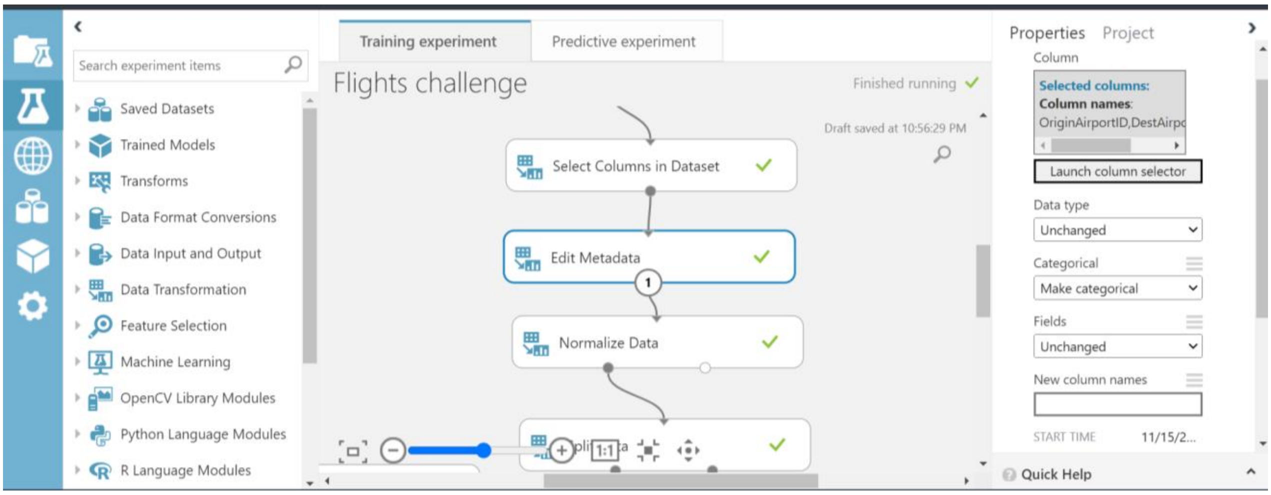
Select Columns in Dataset

Select columns

Selected columns:
Column names: Month,DayofMonth,Carri

START TIME	11/15/2...
END TIME	11/15/2...
ELAPSED TIME	0:00:00.0...
STATUS CODE	Finished
STATUS DETAILS	Task output was present in output cache

Task 3: Add an Edit Metadata module and use it to make the OriginAirportID, DestAirportID, and Carrier columns Categorical.



Column

Selected columns:
Column names: OriginAirportID,DestAirpo

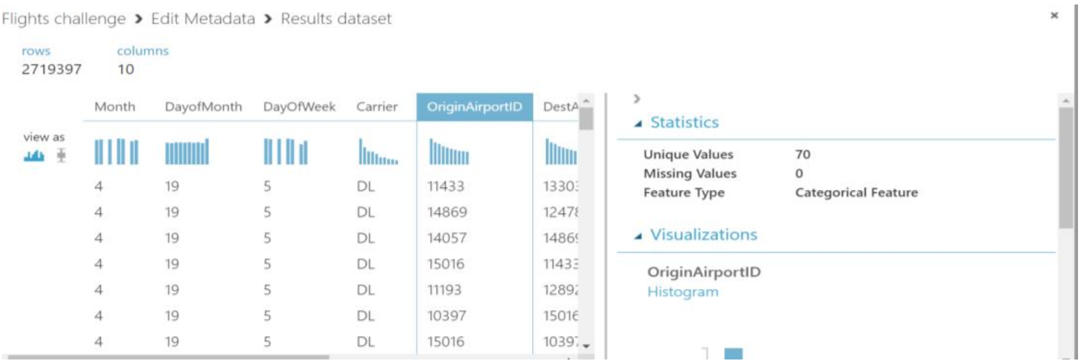
Data type: Unchanged

Categorical: Make categorical

Fields: Unchanged

New column names:

START TIME 11/15/2...



Flights challenge > Edit Metadata > Results dataset

rows 2719397 columns 10

Month	DayofMonth	DayOfWeek	Carrier	OriginAirportID	DestA
4	19	5	DL	11433	1330
4	19	5	DL	14869	1247
4	19	5	DL	14057	1486
4	19	5	DL	15016	1143
4	19	5	DL	11193	1289
4	19	5	DL	10397	1501
4	19	5	DL	15016	1039

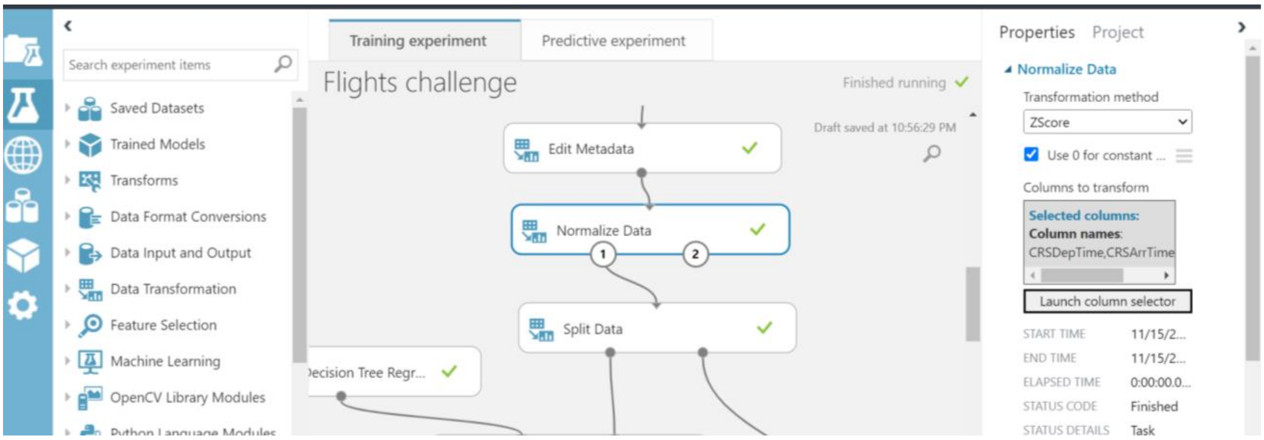
Statistics

Unique Values	70
Missing Values	0
Feature Type	Categorical Feature

Visualizations

OriginAirportID
Histogram

Task 4: Add a Normalize Data module and use it to standardize the CRSDepTime, CRSArrTime, and DepDelay columns using the ZScoretransformation method.



Normalize Data

Transformation method: ZScore

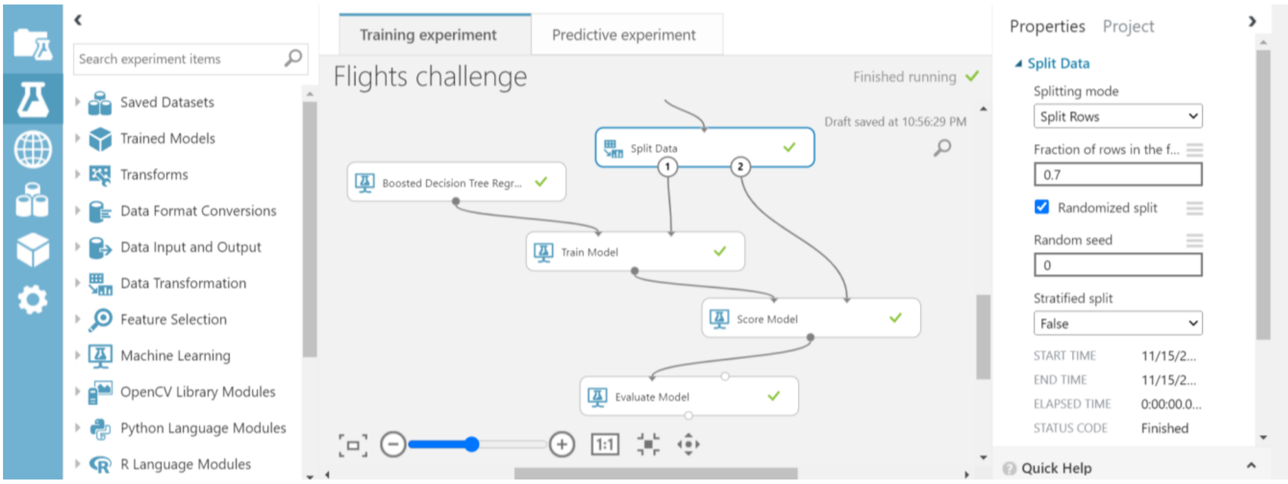
☑ Use 0 for constant ...

Columns to transform

Selected columns:
Column names: CRSDepTime,CRSArrTime

START TIME	11/15/2...
END TIME	11/15/2...
ELAPSED TIME	0:00:00.0...
STATUS CODE	Finished
STATUS DETAILS	Task

Task 5: Add a Split Data module and use it to split the rows into 70% / 30% subsets. Use a random seed value of 0.



Flights challenge

Finished running ✓

Draft saved at 10:56:29 PM

Split Data

Splitting mode: Split Rows

Fraction of rows in the f...: 0.7

☑ Randomized split

Random seed: 0

Stratified split: False

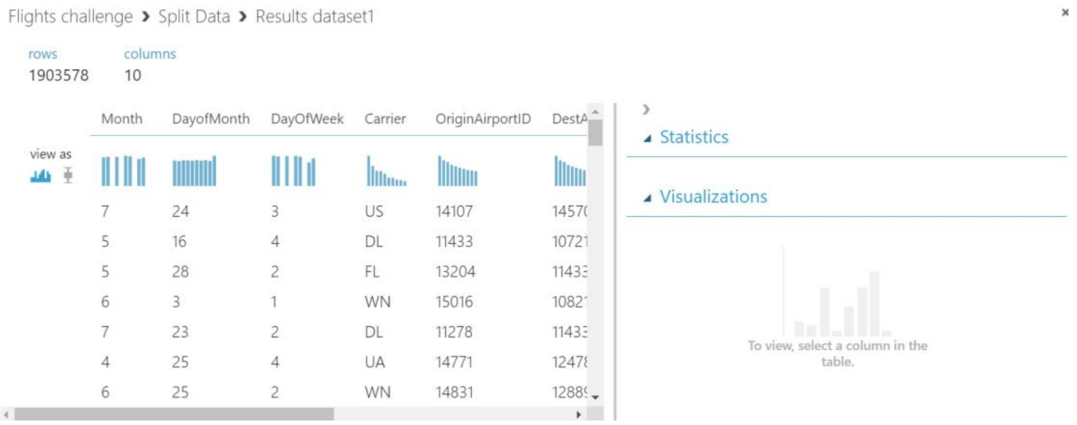
START TIME 11/15/2...

END TIME 11/15/2...

ELAPSED TIME 0:00:00.0...

STATUS CODE Finished

Train dataset:

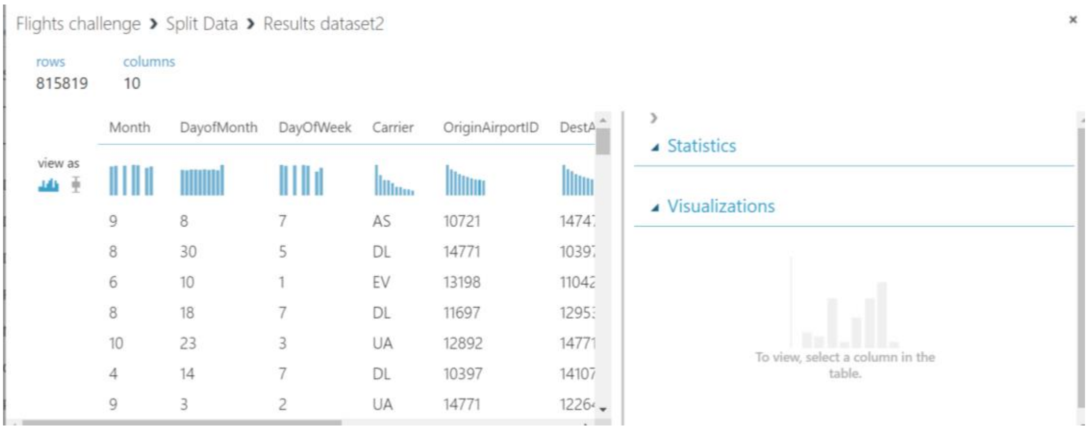


Flights challenge › Split Data › Results dataset1

rows	columns
1903578	10

Month	DayofMonth	DayOfWeek	Carrier	OriginAirportID	DestA
7	24	3	US	14107	14570
5	16	4	DL	11433	10721
5	28	2	FL	13204	11433
6	3	1	WN	15016	10821
7	23	2	DL	11278	11433
4	25	4	UA	14771	12478
6	25	2	WN	14831	12889

Test Dataset:



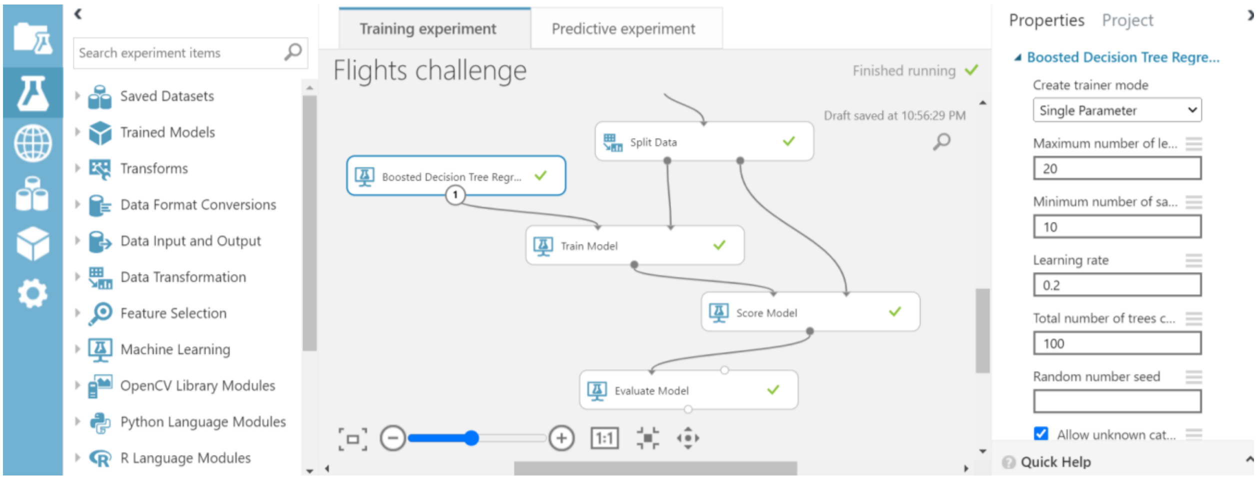
Flights challenge › Split Data › Results dataset2

rows	columns
815819	10

Month	DayofMonth	DayOfWeek	Carrier	OriginAirportID	DestA
9	8	7	AS	10721	14747
8	30	5	DL	14771	10397
6	10	1	EV	13198	11042
8	18	7	DL	11697	12953
10	23	3	UA	12892	14771
4	14	7	DL	10397	14107
9	3	2	UA	14771	12264

Task 6: Add a Boosted Decision Tree Regression module and a Train Model module. Then use the default settings to train the model with the 70% data split to predict the ArrDelay label column.

Boosted Decision Tree Regression module:



Boosted Decision Tree Regre...

Create trainer mode: Single Parameter

Maximum number of le...: 20

Minimum number of sa...: 10

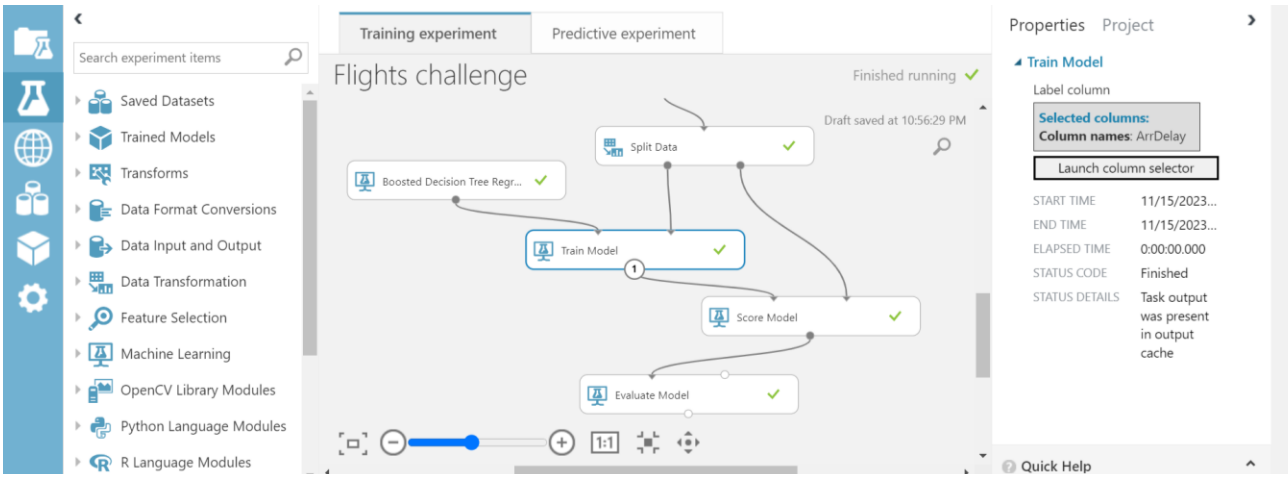
Learning rate: 0.2

Total number of trees c...: 100

Random number seed:

☑ Allow unknown cat...

Train Model Module:



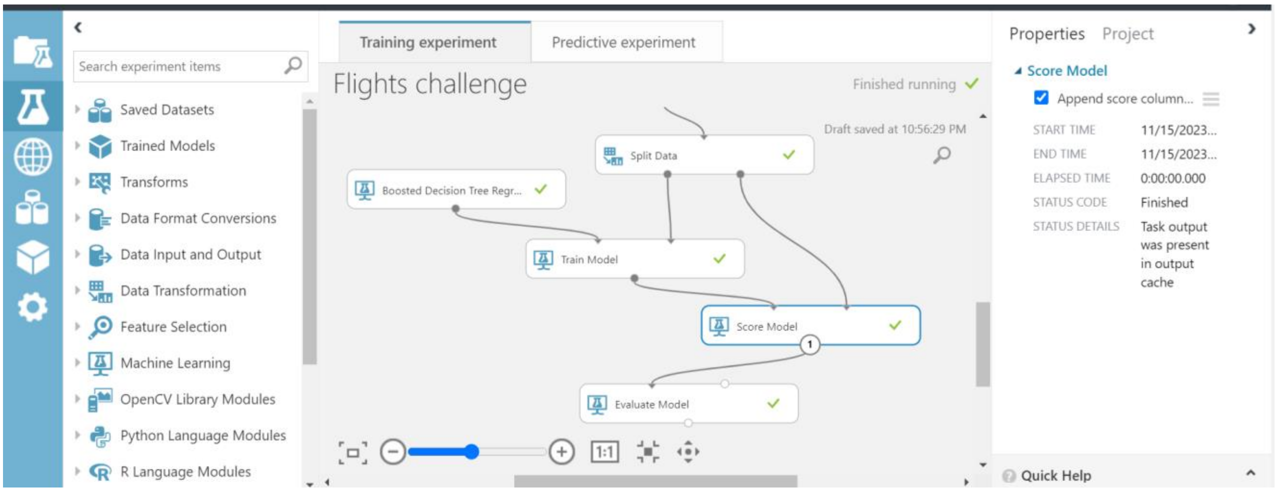
Train Model

Label column

Selected columns:
Column names: ArrDelay

START TIME	11/15/2023...
END TIME	11/15/2023...
ELAPSED TIME	0:00:00.000
STATUS CODE	Finished
STATUS DETAILS	Task output was present in output cache

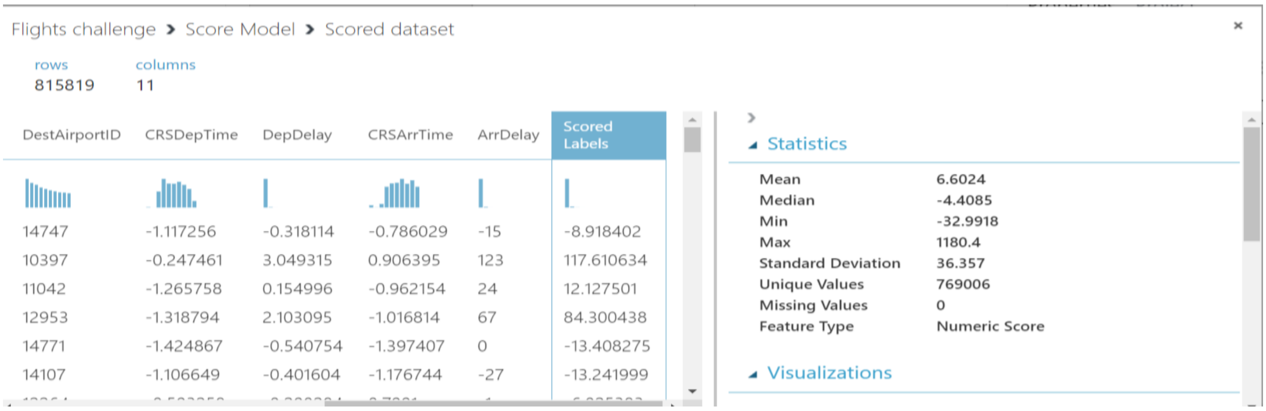
Task 7: Add a Score Model module, and use it to score the trained model using the 30% split of data.



Score Model

☑ Append score column...

START TIME	11/15/2023...
END TIME	11/15/2023...
ELAPSED TIME	0:00:00.000
STATUS CODE	Finished
STATUS DETAILS	Task output was present in output cache



Flights challenge › Score Model › Scored dataset

rows	columns
815819	11

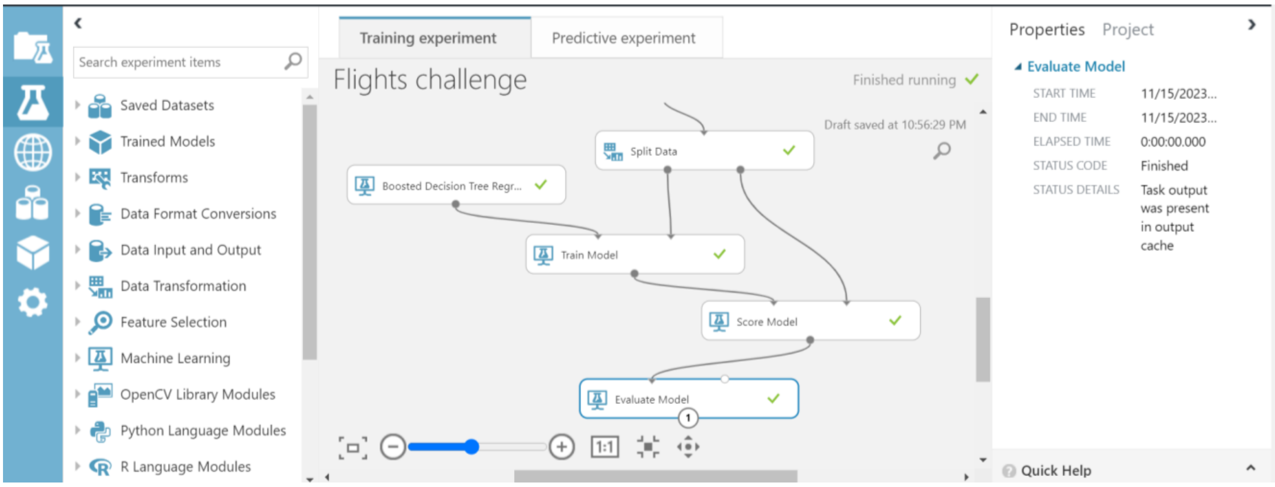
DestAirportID	CRSDepTime	DepDelay	CRSArrTime	ArrDelay	Scored Labels
14747	-1.117256	-0.318114	-0.786029	-15	-8.918402
10397	-0.247461	3.049315	0.906395	123	117.610634
11042	-1.265758	0.154996	-0.962154	24	12.127501
12953	-1.318794	2.103095	-1.016814	67	84.300438
14771	-1.424867	-0.540754	-1.397407	0	-13.408275
14107	-1.106649	-0.401604	-1.176744	-27	-13.241999

Statistics

Mean	6.6024
Median	-4.4085
Min	-32.9918
Max	1180.4
Standard Deviation	36.357
Unique Values	769006
Missing Values	0
Feature Type	Numeric Score

Visualizations

Task 8: Add an Evaluate Model module and use it to evaluate the results from the Score Model module.



Flights challenge

Finished running ✓

Draft saved at 10:56:29 PM

Properties Project

Evaluate Model

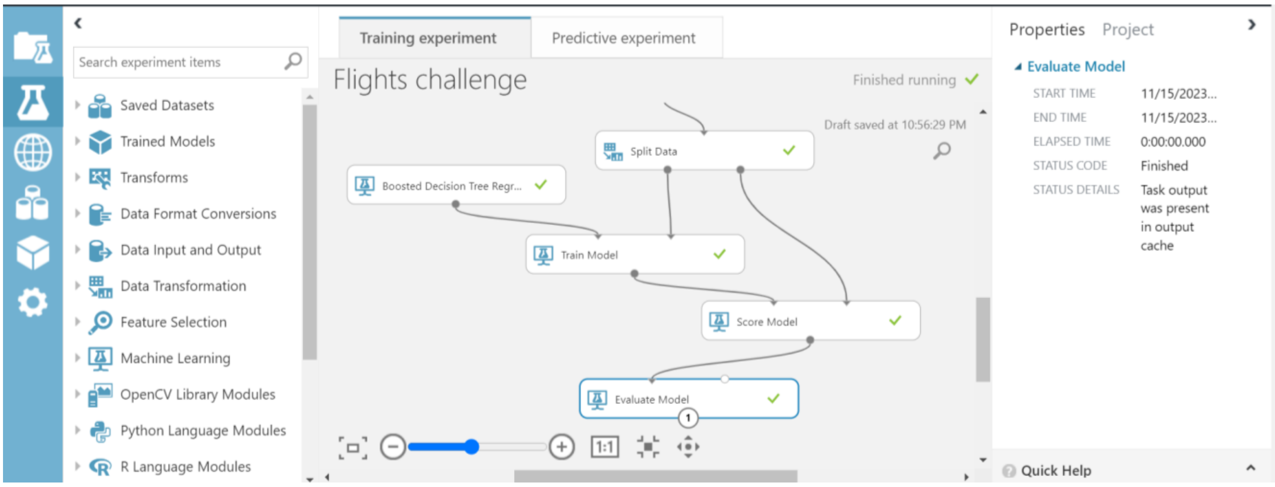
START TIME	11/15/2023...
END TIME	11/15/2023...
ELAPSED TIME	0:00:00.000
STATUS CODE	Finished
STATUS DETAILS	Task output was present in output cache

Test and Evaluate the Model

Perform the following tasks

Task 1: Run the experiment

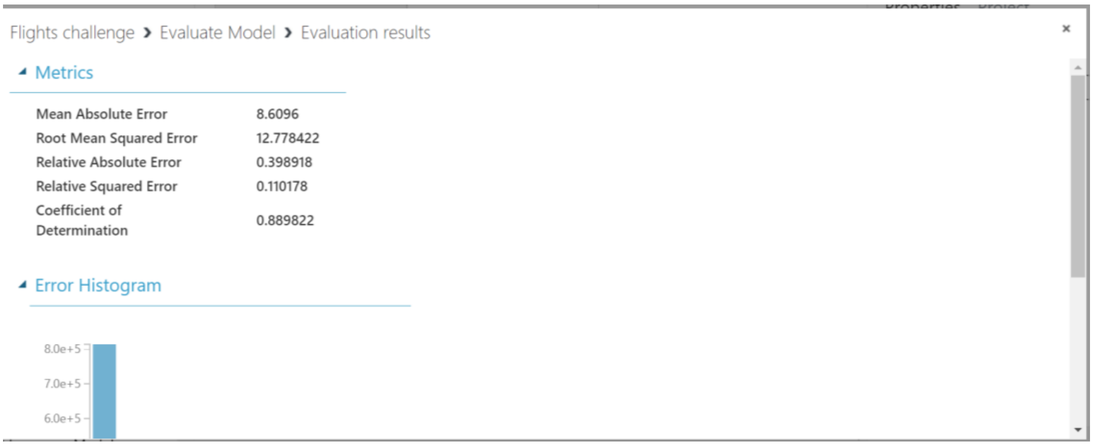
Task 2: When it has finished, visualize the output of the Evaluate Model module.



Evaluate Model

START TIME	11/15/2023...
END TIME	11/15/2023...
ELAPSED TIME	0:00:00.000
STATUS CODE	Finished
STATUS DETAILS	Task output was present in output cache

Evaluate result:



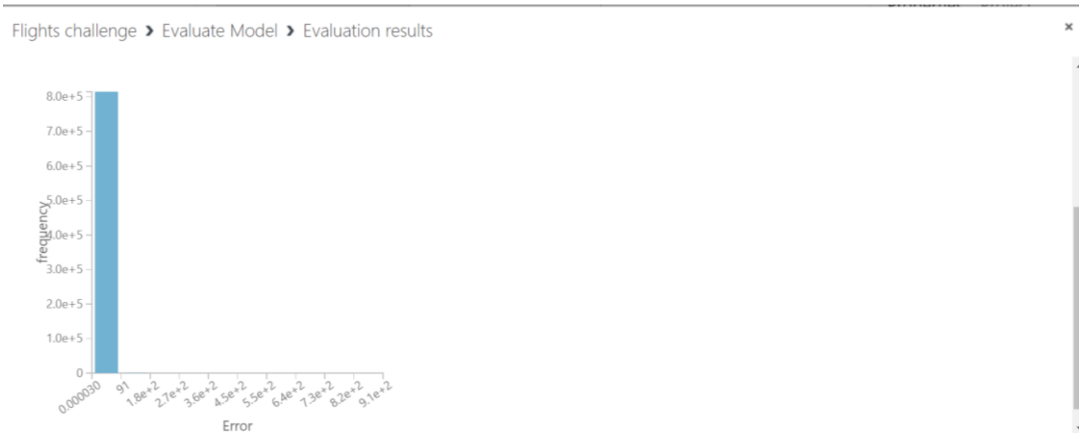
Flights challenge › Evaluate Model › Evaluation results

Metrics

Mean Absolute Error	8.6096
Root Mean Squared Error	12.778422
Relative Absolute Error	0.398918
Relative Squared Error	0.110178
Coefficient of Determination	0.889822

Error Histogram

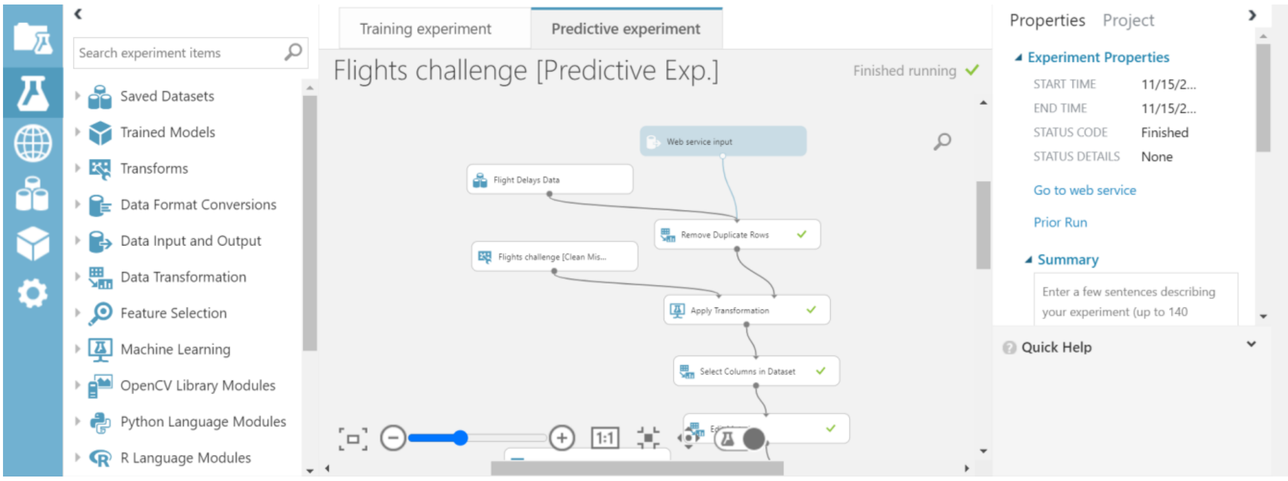
Error Histogram:



Flights challenge › Evaluate Model › Evaluation results

Publish and Use the Model

Task 1: Set up the experiment as a web service, creating a predictive experiment (if the option to do this is not available, save and re-run the experiment).

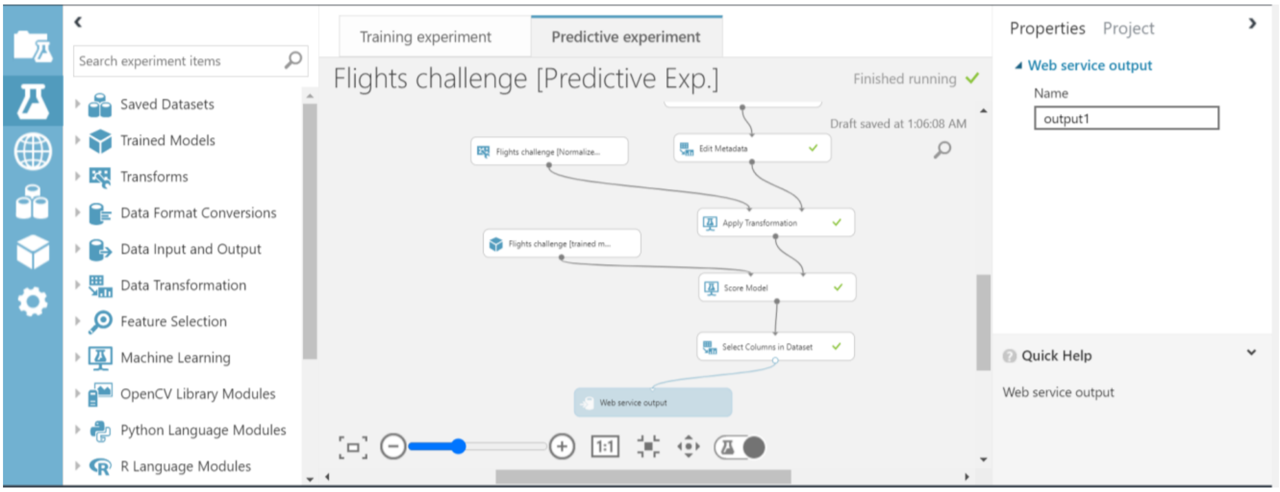


Flights challenge [Predictive Exp.]

Finished running ✓

START TIME	11/15/2...
END TIME	11/15/2...
STATUS CODE	Finished
STATUS DETAILS	None

Enter a few sentences describing your experiment (up to 140



Flights challenge [Predictive Exp.]

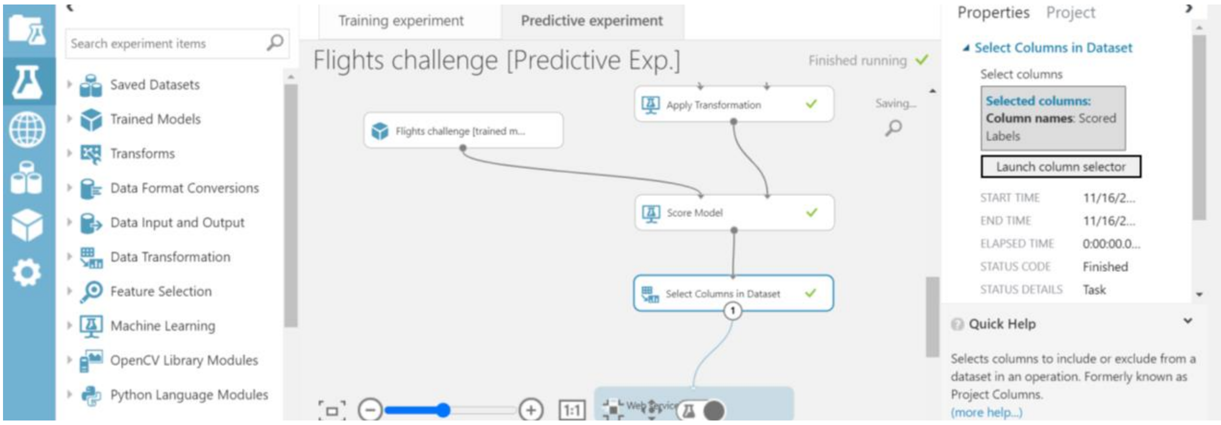
Finished running ✓

Draft saved at 1:06:08 AM

Name: output1

Web service output

Task 2: In the predictive experiment, add a Select Columns in the Dataset module and place it between the Score Model and Web service output modules. Use this to select only the Scored Labels column.



Flights challenge [Predictive Exp.]

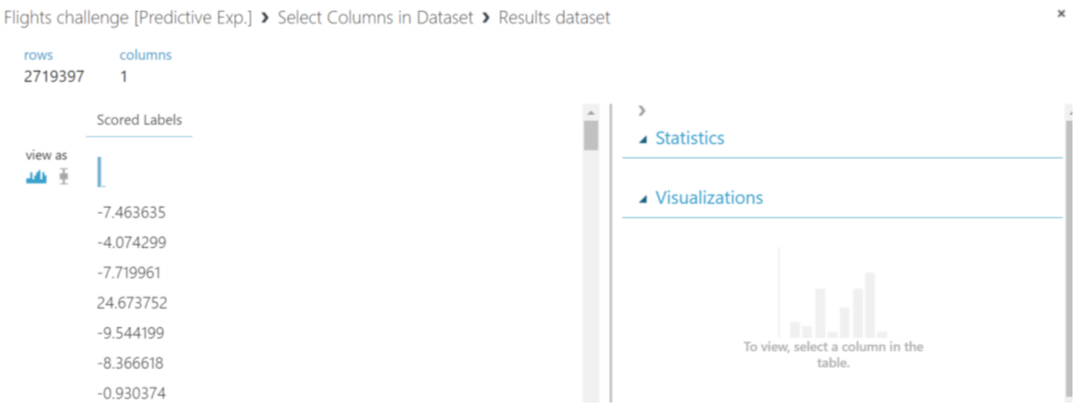
Finished running ✓

Select columns

Selected columns:
Column names: Scored Labels

START TIME	11/16/2...
END TIME	11/16/2...
ELAPSED TIME	0:00:00.0...
STATUS CODE	Finished
STATUS DETAILS	Task

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

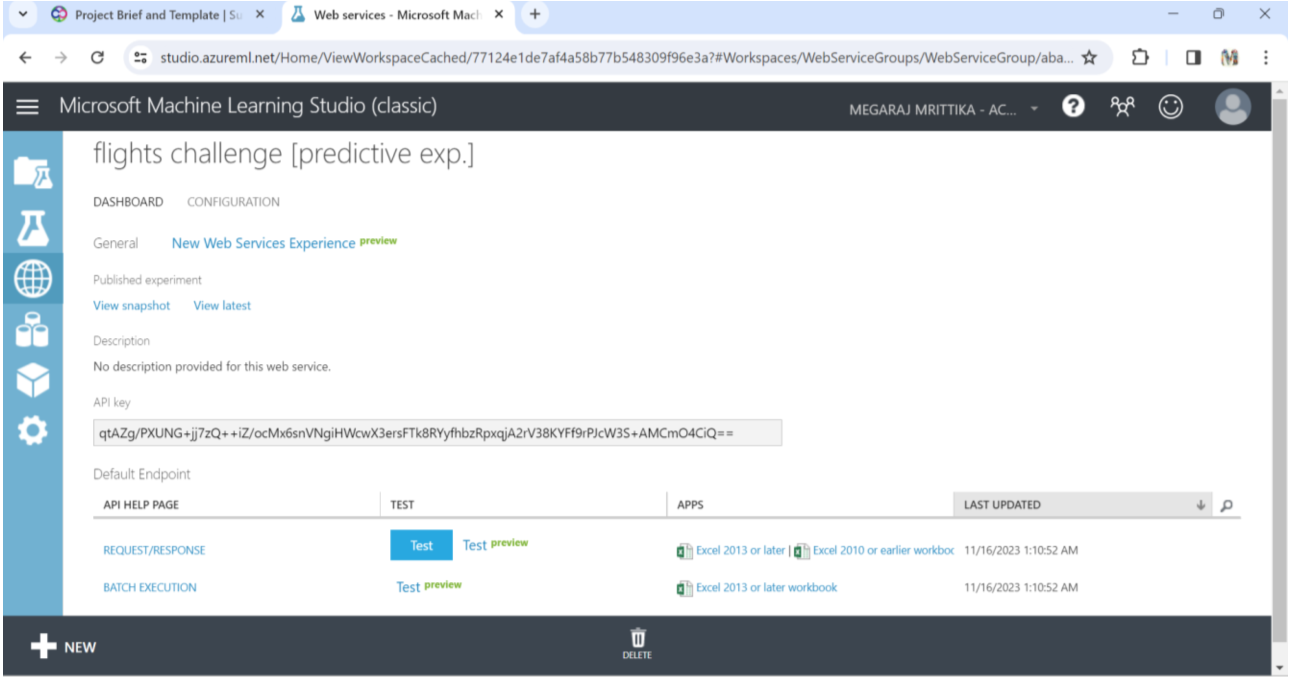


Flights challenge [Predictive Exp.] › Select Columns in Dataset › Results dataset

rows	columns
2719397	1

Scored Labels
-7.463635
-4.074299
-7.719961
24.673752
-9.544199
-8.366618
-0.930374

Task 3: Save and run the modified predictive experiment, and then deploy the web service.



Project Brief and Template | Su
Web services - Microsoft Mach
Microsoft Machine Learning Studio (classic)
MEGARAJ MRITTIKA - AC...
NEW
DELETE

studio.azureml.net/Home/ViewWorkspaceCached/77124e1de7af4a58b77b548309f96e3a?#Workspaces/WebServiceGroups/WebServiceGroup/aba...

flights challenge [predictive exp.]

DASHBOARD CONFIGURATION

General New Web Services Experience preview

Published experiment

View snapshot View latest

Description

No description provided for this web service.

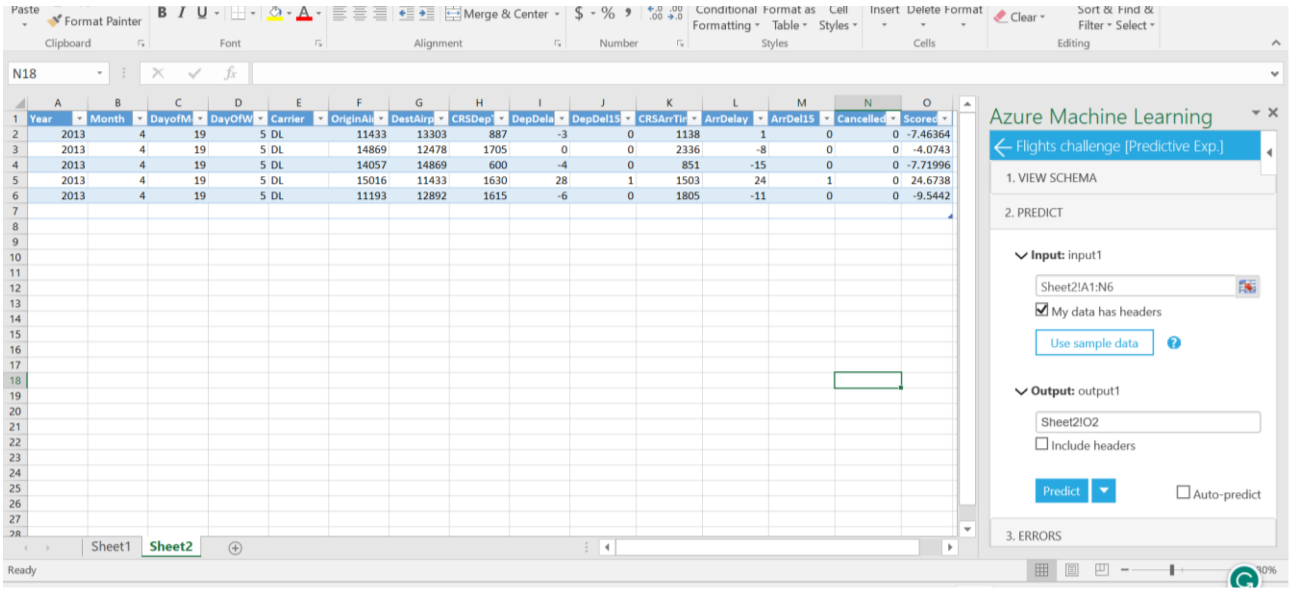
API key

qtAZg/PXUNG+jj7zQ++iZ/ocMx6snVNgiHWcwX3ersFTk8RYyfhbzRpxqjA2rV38KYFf9rPJcW3S+AMCmO4CiQ==

Default Endpoint

API HELP PAGE	TEST	APPS	LAST UPDATED
REQUEST/RESPONSE	Test Test preview	Excel 2013 or later | Excel 2010 or earlier workboo	11/16/2023 1:10:52 AM
BATCH EXECUTION	Test preview	Excel 2013 or later workbook	11/16/2023 1:10:52 AM

Task 4: Enter the values predicted by your model for each input row



Year	Month	DayofM	DayOfW	Carrier	OriginAi	DestAirp	CRSDep	DepDela	DepDel15	CRSArrTir	ArrDelay	ArrDel15	Cancelled	Scored
2013	4	19	5	DL	11433	13303	887	-3	0	1138	1	0	0	-7.46364
2013	4	19	5	DL	14869	12478	1705	0	0	2336	-8	0	0	-4.0743
2013	4	19	5	DL	14057	14869	600	-4	0	851	-15	0	0	-7.71996
2013	4	19	5	DL	15016	11433	1630	28	1	1503	24	1	0	24.6738
2013	4	19	5	DL	11193	12892	1615	-6	0	1805	-11	0	0	-9.5442

Azure Machine Learning

Flights challenge [Predictive Exp.]

1. VIEW SCHEMA

2. PREDICT

Input: input1

Sheet2!A1:N6

☑ My data has headers

Use sample data

Output: output1

Sheet2!O2

☐ Include headers

Predict

☐ Auto-predict

3. ERRORS