Data Science and Machine Learning Essentials

By Christopher Cerritos, Tam Do, Mike Avelar
Completed by Christopher Cerritos, Tam Do, Mike Avelar on 12/14/2019

Objectives

In this lab, you will learn how to import a raw dataset onto Azure ML, split data, train a dataset, evaluate the model and visualize results.

What You'll Need

To complete this lab, you will need the following:

- An Azure ML account.
- A web browser and Internet connection.

Platform Specifications

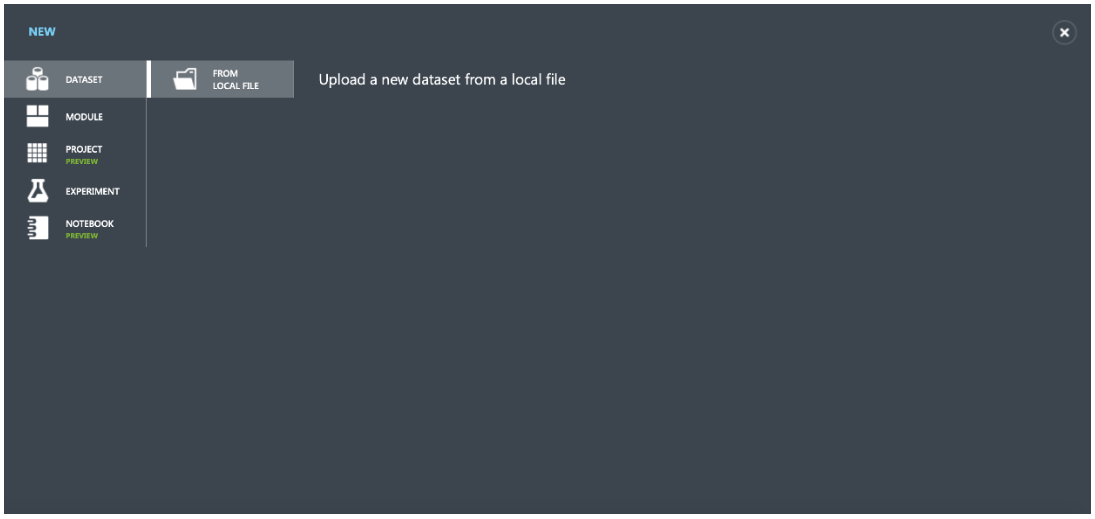
Low priority Azure ML on a single processor core.

- 10GB

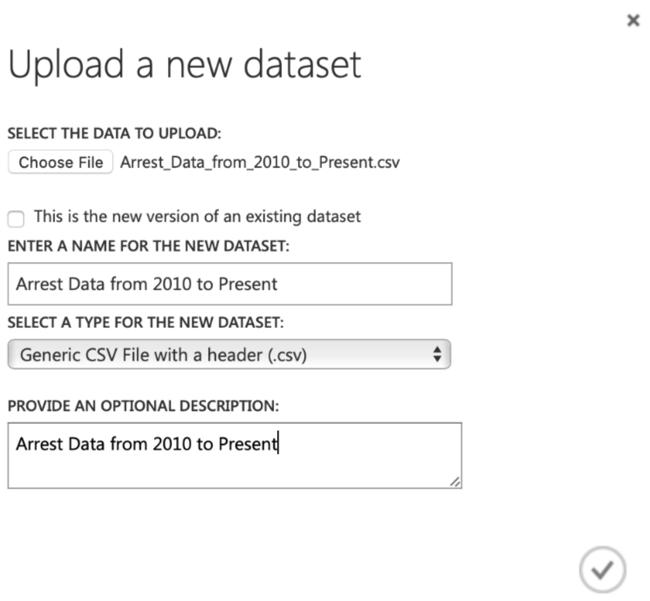
Note: This resulted in longer wait times.

Upload the Data File to Create a New Dataset in Azure ML

1. Get raw dataset from the Los Angeles Open Data.
2. Export the data in CSV for Excel format.
3. Open and login to Azure ML (http://studio.azureml.net).
4. Go to Datasets and select **NEW**.



5. Click **FROM LOCAL FILE**. Then in the **Upload a new dataset** dialog box, browse to select the **Arrest_Data_from_2010_to_Present.csv** file from the folder where you downloaded on your local computer and enter the following details as shown in the image below, and then click the OK icon.
   - This is a new version of an existing dataset: Unselected
   - Enter a name for the new dataset: Arrest Data from 2010 to Present
   - Select a type for the new dataset: Generic CSV file with a header (.csv)
   - Provide an optional description: Arrest Data from 2010 to Present



Upload a new dataset

SELECT THE DATA TO UPLOAD:

Choose File Arrest_Data_from_2010_to_Present.csv

□ This is the new version of an existing dataset

ENTER A NAME FOR THE NEW DATASET:

Arrest Data from 2010 to Present

SELECT A TYPE FOR THE NEW DATASET:

Generic CSV File with a header (.csv)

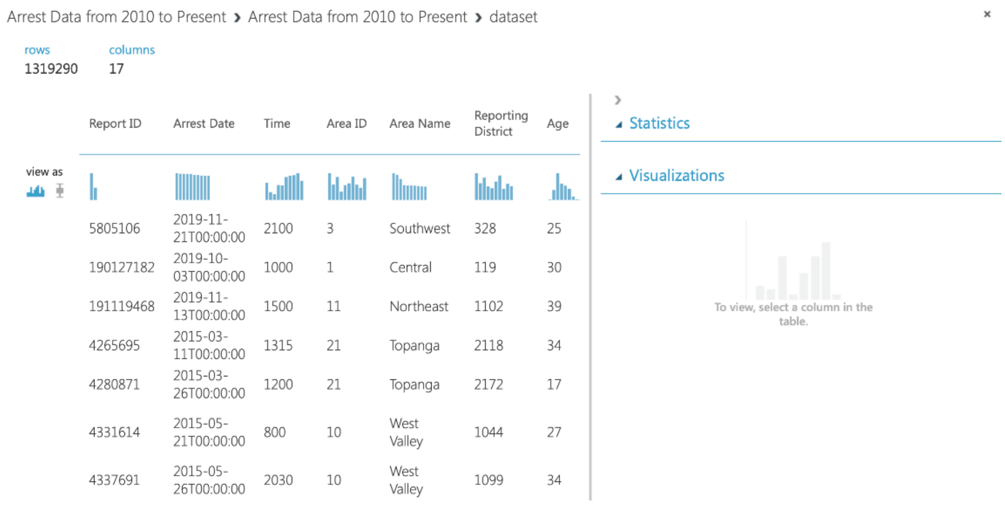
PROVIDE AN OPTIONAL DESCRIPTION:

Arrest Data from 2010 to Present

6. Wait for the upload of the dataset to be completed, and then on the experiment items pane, expand **Saved Datasets** and **My Datasets** to verify that the **Arrest Data from 2010 to Present** dataset is listed.

Visualize the Dataset in Azure ML

1. Drag the **Arrest Data from 2010 to Present** dataset to the canvas for the **Arrest Data from 2010 to Present** experiment.
2. Right-click the output port for the **Arrest Data from 2010 to Present** dataset on the canvas and click **Visualize** to view the data in the dataset.



Arrest Data from 2010 to Present › Arrest Data from 2010 to Present › dataset

rows	columns
1319290	17

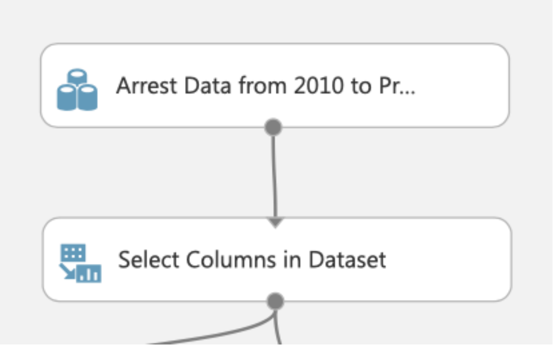
Report ID	Arrest Date	Time	Area ID	Area Name	Reporting District	Age
5805106	2019-11-21T00:00:00	2100	3	Southwest	328	25
190127182	2019-10-03T00:00:00	1000	1	Central	119	30
191119468	2019-11-13T00:00:00	1500	11	Northeast	1102	39
4265695	2015-03-11T00:00:00	1315	21	Topanga	2118	34
4280871	2015-03-26T00:00:00	1200	21	Topanga	2172	17
4331614	2015-05-21T00:00:00	800	10	West Valley	1044	27
4337691	2015-05-26T00:00:00	2030	10	West Valley	1099	34

To view, select a column in the table.

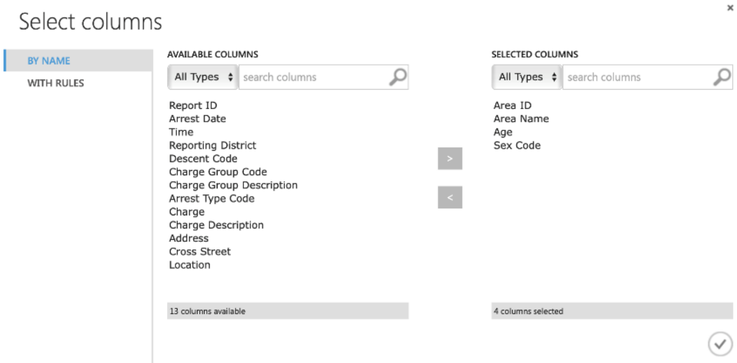
3. Verify that the dataset contains the data you viewed in the source file (Excel), and then close the dataset.

Add and Train a Model

1. Open the **Arrest Data from 2010 to Present** experiment you have already created.
2. On the left again, in the search experiment items, find **Select Columns in Dataset**.
3. Drag it to the middle and connect dataset port from **Arrest Data from 2010 to Present** to the dataset port of **Select Columns in Dataset**.



4. Click on select columns and in properties, click on **Launch column selector**.
5. Select **BY NAME** and choose only *Area ID, Area Name, Age, and Sex Code* from the left columns to the select columns.



Select columns

BY NAME
WITH RULES

AVAILABLE COLUMNS

All Types | search columns

Report ID
Arrest Date
Time
Reporting District
Descent Code
Charge Group Code
Charge Group Description
Arrest Type Code
Charge
Charge Description
Address
Cross Street
Location

13 columns available

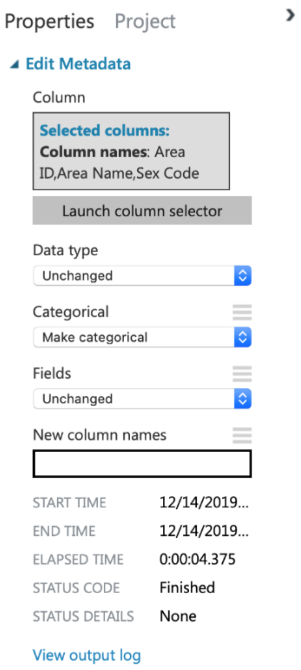
SELECTED COLUMNS

All Types | search columns

Area ID
Area Name
Age
Sex Code

4 columns selected

6. We will then add **Edit Metadata** from the modules on the left side. Click and drag it to the middle.
7. Connect dataset result port from **Select Columns in Dataset** to the **Edit Metadata** dataset port.
8. Click on **Edit Metadata** and go to its properties to the right.
9. In properties click **Launch column selector** and choose columns name *Area ID, Area Name, Sex Code*.
10. Again, the properties are as followed:
    - Data Type: Unchanged
    - Categorical: Make categorical
    - Field: Unchanged
    - New column names: Blank.



Properties Project

Edit Metadata

Column

Selected columns:
Column names: Area ID,Area Name,Sex Code

Data type: Unchanged

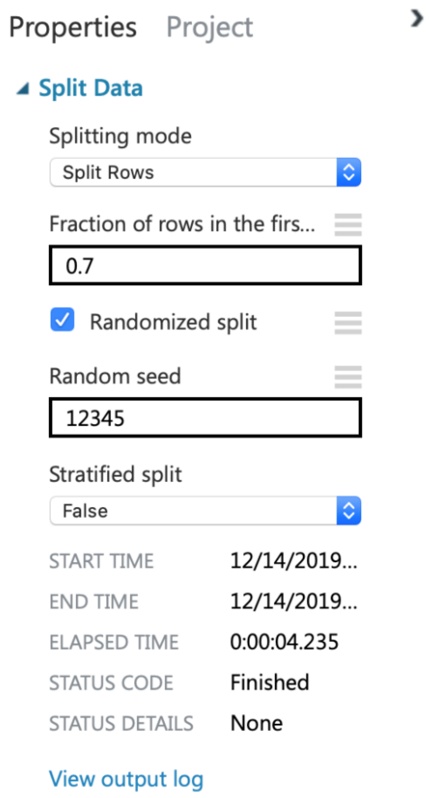
Categorical: Make categorical

Fields: Unchanged

New column names:

START TIME 12/14/2019...
END TIME 12/14/2019...
ELAPSED TIME 0:00:04.375
STATUS CODE Finished
STATUS DETAILS None

11. On the left again, in the search experiment items find **Split Data**. Click and drag it to the middle.
12. Select dataset result port from **Edit Metadata** to **Split Data** dataset port.
13. Click on **Split Data** and change its properties as followed:
    - Splitting mode: Split Rows
    - Fraction of rows in the first output dataset: 0.7
    - Randomized split: Checked
    - Random seed: 12345
    - Stratified split: False



Split Data

Splitting mode: Split Rows

Fraction of rows in the firs...: 0.7

☑ Randomized split

Random seed: 12345

Stratified split: False

START TIME 12/14/2019...

END TIME 12/14/2019...

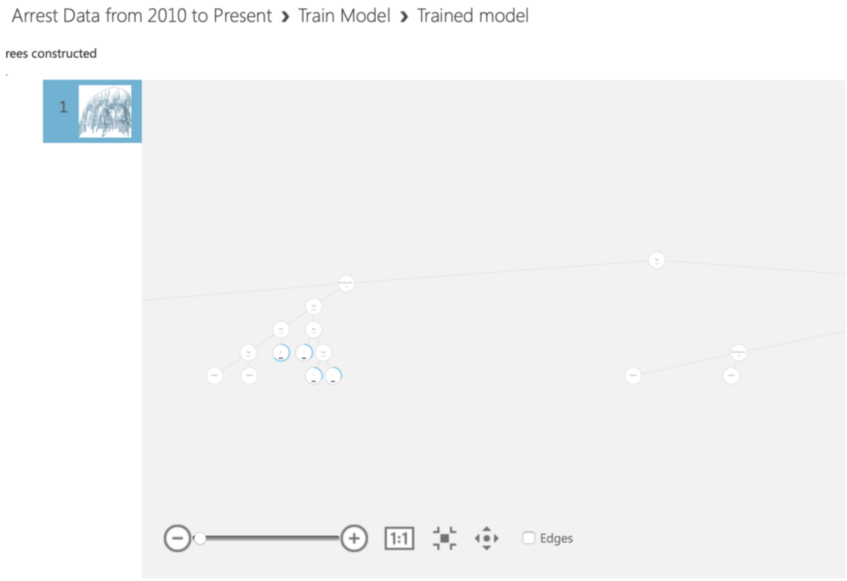
ELAPSED TIME 0:00:04.235

STATUS CODE Finished

STATUS DETAILS None

Visualization

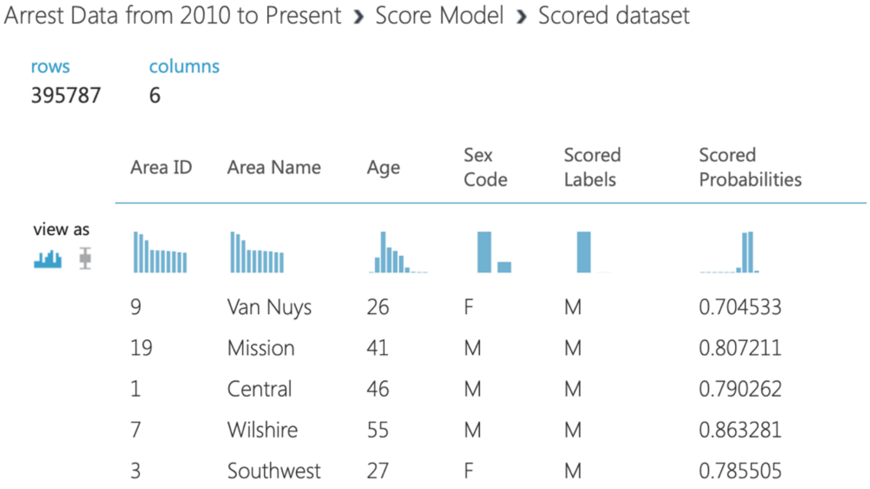
1. In this step we will use a **Train Model** for us to train our models in classification or regression.
2. In the search experiment items find **Train Model**. Click and drag it to the middle.
3. Connect result dataset port of **Split Data** and connect to **Train Model** dataset port.
4. Click on **Train Model** and on its properties click **Launch column selector**.
    - Column names: *Sex Code*
5. In the search experiment items find **Two-Class Decision Forest**. Click and drag to the middle.
6. Select untrained model port from **Two-Class Decision Forest** to untrained model port in **Train Model**.
7. Then click on **Two-Class Decision Forest** and changes it properties as followed:
    - Resampling method: Bagging
    - Create trainer mode: Single Parameter
    - Number of decision trees: 1
    - Maximum depth of the decision trees: 32
    - Number of random splits per node: 128
    - Minimum number of samples per leaf node: 1
8. Save and run the **Arrest Data from 2010 to Present** experiment. Then, when the experiment has finished with “Finished running” at the top of the pane, visualize the Results dataset output of the Train Model module, and view the data. Then close the results dataset.



Arrest Data from 2010 to Present › Train Model › Trained model

rees constructed

9. In the search experiment items find **Score Model**. Click and drag it to the middle.
10. Connect **Score Model** with the **Train Model**.
11. Save and run the experiment. Then, when the experiment has finished with "Finished running" at the top of the pane, visualize from the results from the **Score Model** module. Then close the results dataset.



Arrest Data from 2010 to Present › Score Model › Scored dataset

rows 395787

columns 6

view as

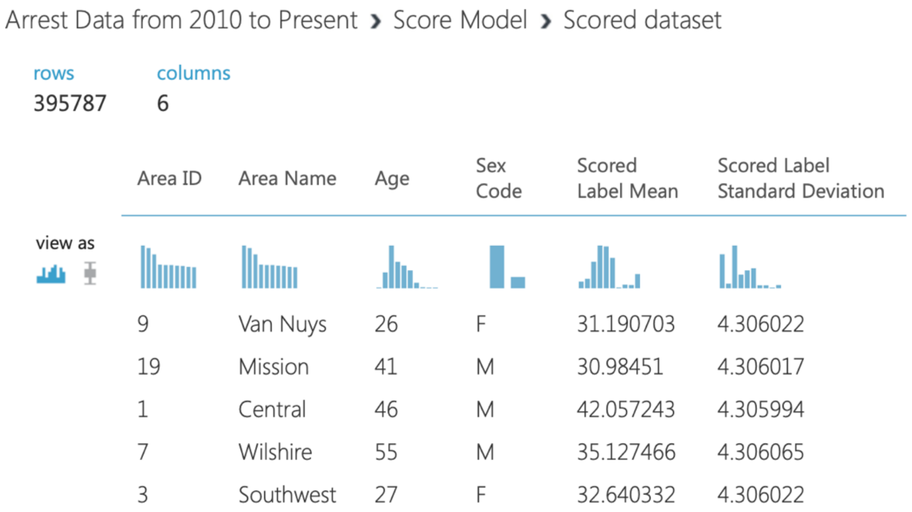
Area ID	Area Name	Age	Sex Code	Scored Labels	Scored Probabilities
9	Van Nuys	26	F	M	0.704533
19	Mission	41	M	M	0.807211
1	Central	46	M	M	0.790262
7	Wilshire	55	M	M	0.863281
3	Southwest	27	F	M	0.785505

Train Model for Ages

1. In the search experiment items find **Train Model**. Click and drag it to the middle.
2. Connect the result dataset port of **Split Data** and connect it to the **Train Model** dataset port.
3. Click on **Train Model** and on its properties click **Launch column selector**. Adjust the properties to the following:
    - Column names: Age
4. In the search experiment items find **Bayesian Linear Regression**. Click and drag it to the middle.
5. Select the untrained model port from the **Two-Class Decision Forest** to the untrained model port in **Train Model**.
6. Then click on the **Bayesian Linear Regression** module and changes its properties to the following:
    - Regularization weight: 1

    - Allow unknow category: Checked

7. In the search experiment items find the **Score Model** module. Click and drag it to the middle.
8. Connect the **Score Model** module with the **Train Model**.
9. Save and run the experiment. Then, when the experiment has finished with “Finished running” at the top of the pane, visualize from the results from the **Score Model** module. Then close the results dataset.



Arrest Data from 2010 to Present › Score Model › Scored dataset

rows 395787

columns 6

Area ID	Area Name	Age	Sex Code	Scored Label Mean	Scored Label Standard Deviation
9	Van Nuys	26	F	31.190703	4.306022
19	Mission	41	M	30.98451	4.306017
1	Central	46	M	42.057243	4.305994
7	Wilshire	55	M	35.127466	4.306065
3	Southwest	27	F	32.640332	4.306022

10. Lastly, from the search experiments items, find **Permutation Feature Importance**. Click and drag it to the canvas.
11. Connect the output ports of the **Split Data** and **Train Model** to the input ports of the **Permutation Feature Importance** module.
12. Save and run the experiment. You will see a “Failed” message at the top of pane. The model we tried to produce is not effective at this time. We needed more time to look into this issue to produce a working model.

Save and Close the Experiment

1. Click a blank area of the experiment canvas to select the experiment. Then in the **Properties** pane, enter the **Summary** *Arrest Data from 2010 to Present* and the **Description** *A simple experiment to import raw data onto Azure ML, split data, train a dataset, evaluate the model and visualize the results*. Then save the experiment.
2. In the Azure ML Studio page, on the left side, click the **Experiments** icon and note that your experiment is listed. You can return to it at any time from here.

Summary

This lab was designed to help you become familiar with the basic process of importing a raw dataset onto Azure ML, split data, train a dataset, evaluate the model and visualize the results. The model you produced is not particularly effective at this time. Clearly some iterative work would be required to further cleanse the data, identify the most meaningful features to include in the model, and compare the results when using a range of different algorithms.

References

- Arrest Data from 2010 to Present (Azure ML experiment)
- Dataset: https://data.lacity.org/A-Safe-City/Arrest-Data-from-2010-to-Present/yru6-6re4
- GitHub: https://github.com/ccerritoss/cis-3200