CIS3200 Term Project Tutorial

Authors: **Andrew Tran, Chris Alipio, Fiona Tang, Jeffrey Lam, Sean Pattarateranon**

Instructor: **Jongwook Woo, Ph.D.**

Date: 12/16/2019

Lab Tutorial

Andrew Tran (atran80@calstatela.edu)

12/16/2019

Determining Flight Delay Factors via Azure ML

Objectives

List what your objectives are. In this hands-on lab, you will learn how to:

- Grab datasets from ww.kaggle.com
- Condense the dataset to a smaller file size
- Train/Score a Linear Regression model
- View results and extrapolate meaning from it

Platform Spec

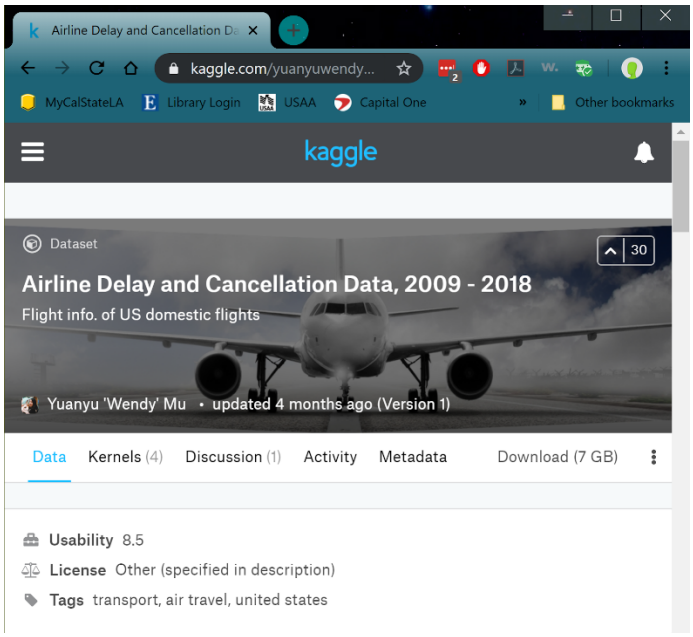
- Microsoft Azure ML
- Hardware: Any usable computer with internet access

Step 1: Get data manually from Kaggle

Explain what this step is for. This step is to get data manually….

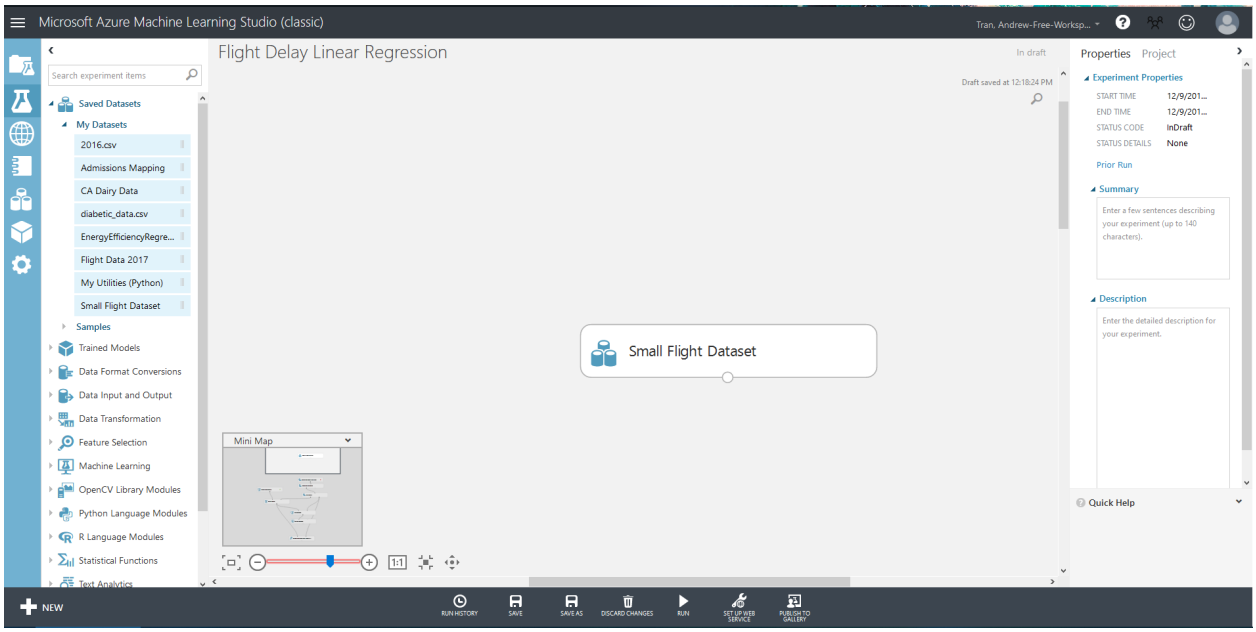
1. Navigate to https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018

2. Create an account if needed and click the download button.

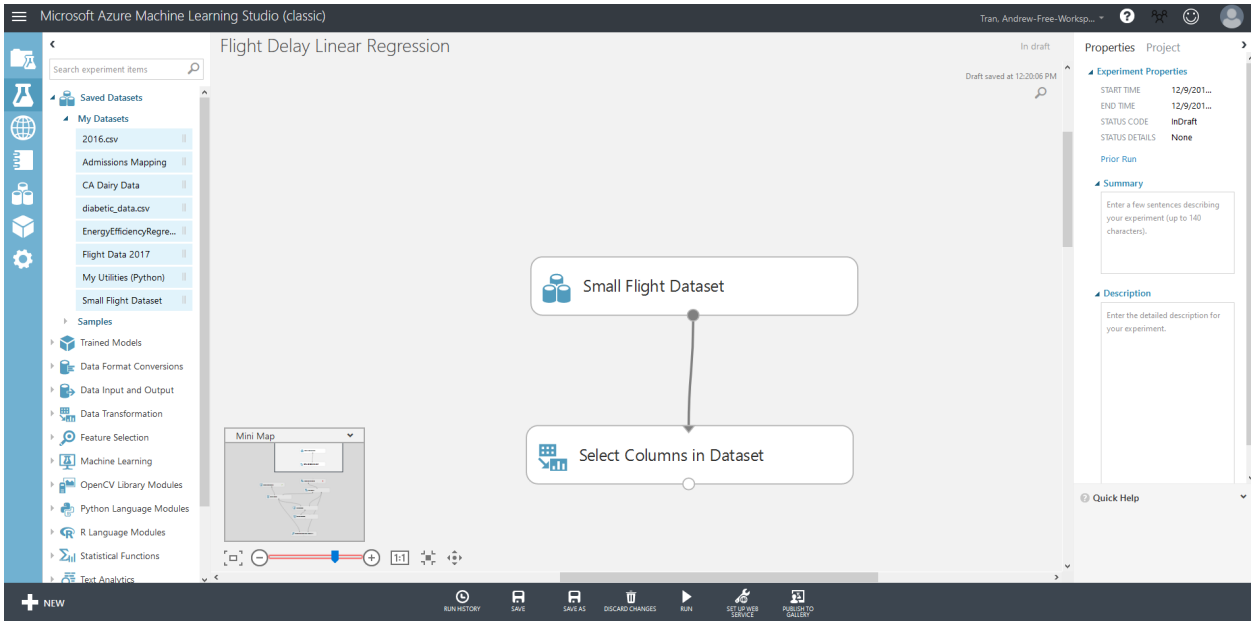


Step 2: Create a Linear Regression Model in Azure ML

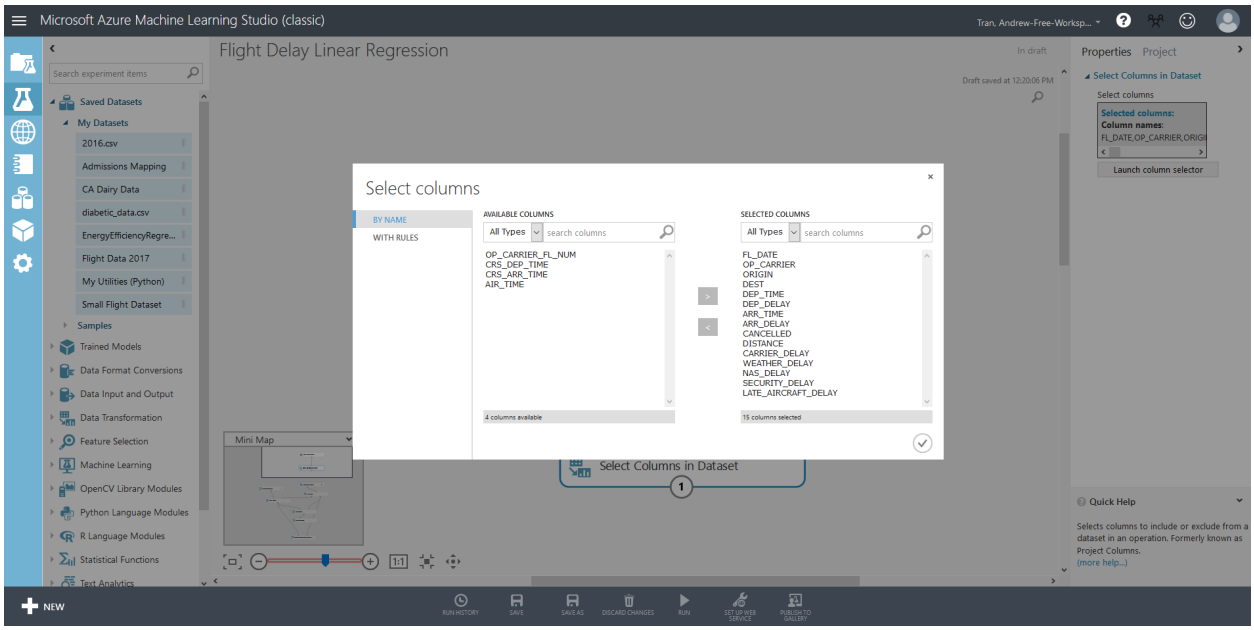
1. Upload the data set: Click New -> Dataset -> From Local File
2. Create the experiment: Click New -> Experiment -> Blank Experiment
3. In the search experiment items bar, look up the **Flight dataset** and drag it onto the canvas.



4. Search for a **Select Columns** module and drag it onto the canvas. Connect the output of the dataset to the new module's input.

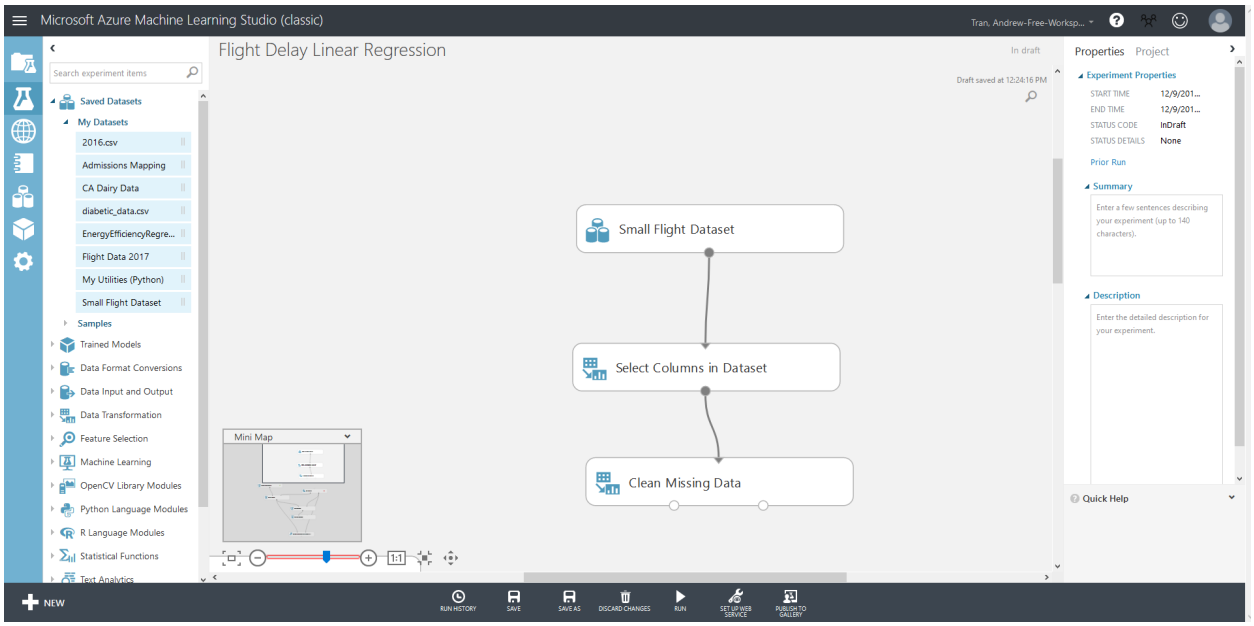


5. Select the columns to include in the experiment via the column selector. The ones we have chosen are depicted below.



6. We will need to clean the dataset up since there are missing values. Search for a **Clean missing data** module and drag it onto the canvas. Connect the **Select Columns**' output and connect it to the

new modules input.



7. We will be splitting the data into training and testing sets. Search for a **Split Data** module and connect the **Clean Missing Data** module's output into the new modules input. Click on the **Split Data** module and configure it as follows:

   Splitting Mode: **Split rows**

   Fraction of rows in the first output dataset: **.8**

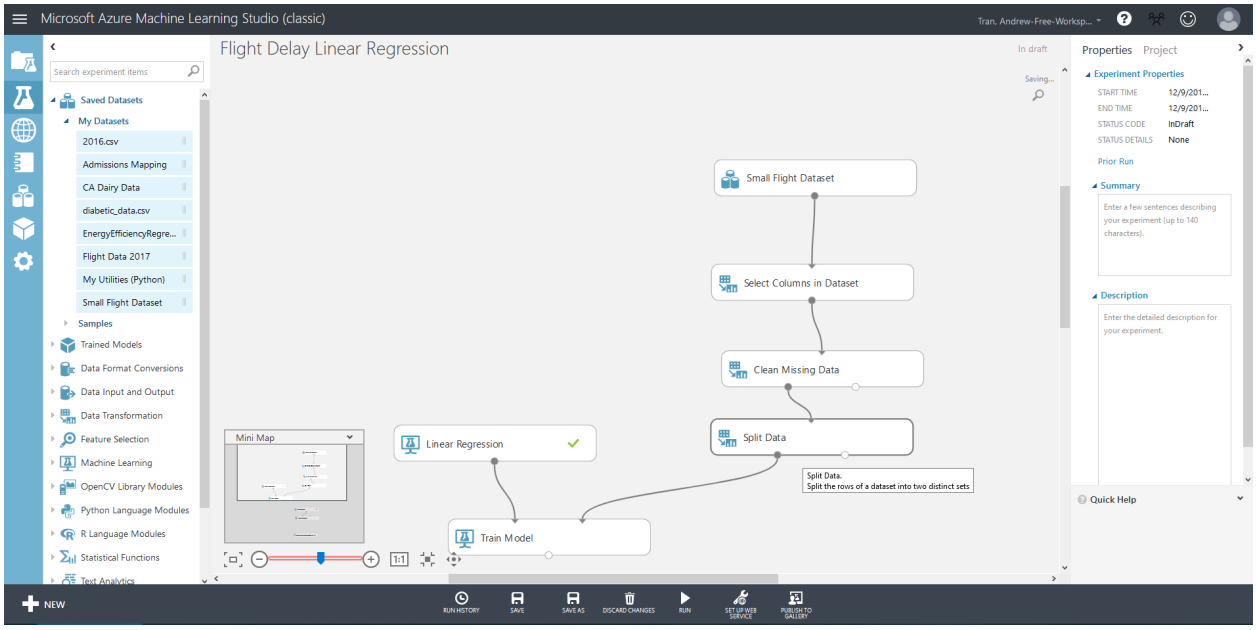
   Randomized split: **Checked**

   Random Seed: **5432**

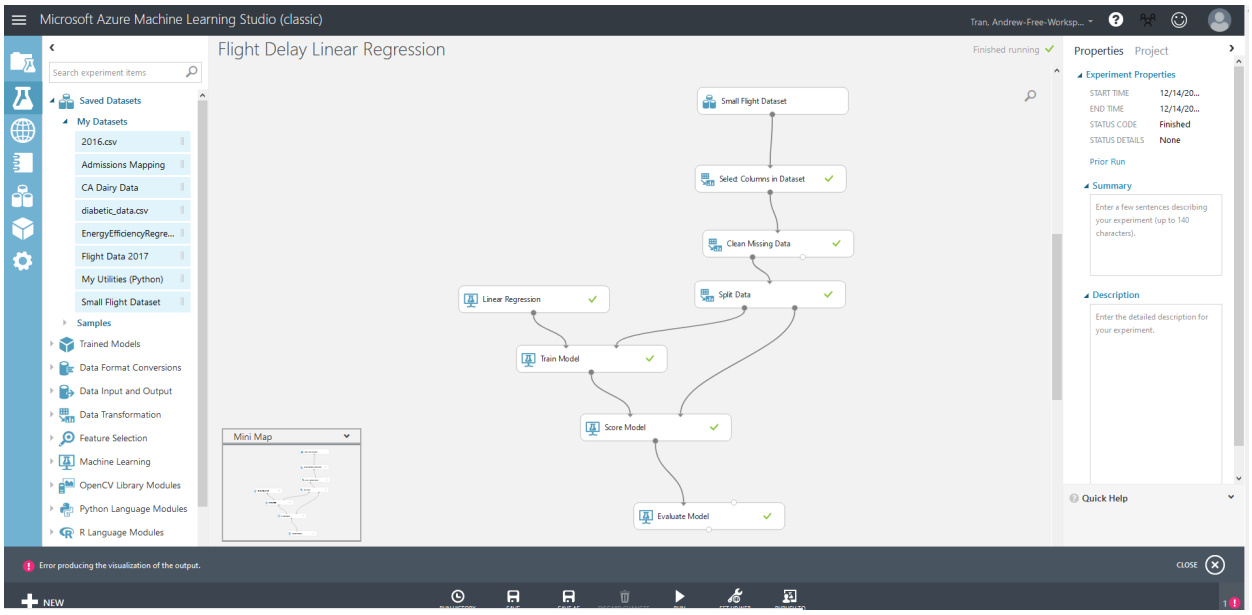
   Stratified Split: **false**

8. Search for the **Linear Regression** module and place it next to the **Split Data** module.
9. Search for the **Train Model** module and place it underneath the **Linear Regression** module; connect the **Linear Regression** module's output into the **Train model**'s input. Select the **Train Model** module and include only **DEP_DELAY** via the column selector.

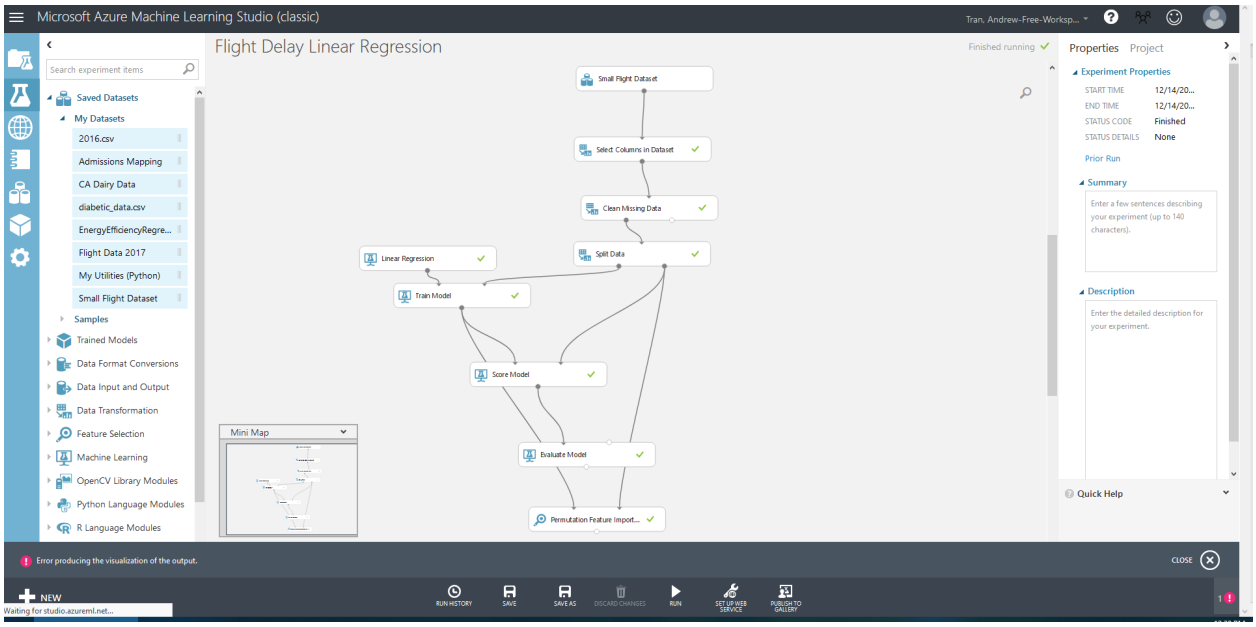
10. Connect the leftmost **Split Data** module's output to the **Train model**'s input. Your experiment should be similar to what is shown so far.



11. Search for and drop in a **Score Model** module and connect the **Train Model** module's output into the leftmost input of the new Module. Then connect the rightmost output of the **Split Data** module to the rightmost input of the **Score Model** module.
12. Search for and drop in an **Evaluate Model** module and connect the **Score Model** module's output into the leftmost input of the new module. It should look like what is show below.



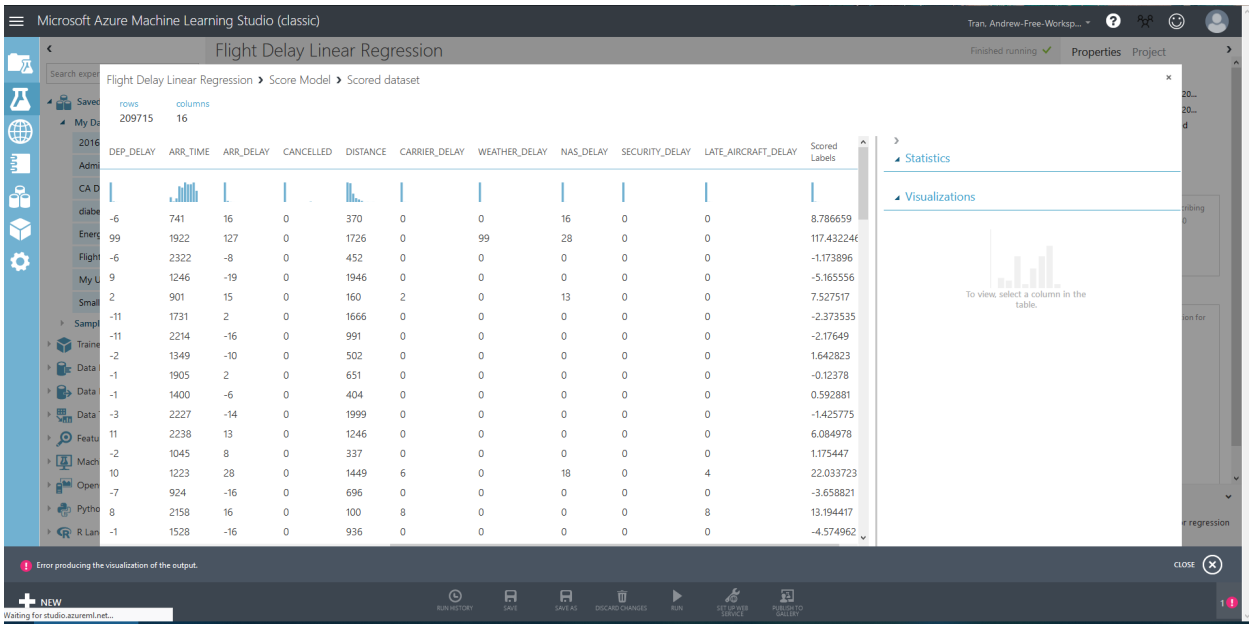
13. Search for a **Permutation Feature Importance** module and connect the **Train Model** module's output to the new module's left input. Then connect the **Split Data** module's rightmost output to the new module's right most input. Click on the module and select **Regression – Coefficient of Determination** under **Metric** for measuring performance.
14. We are finished with setting up the experiment. Name, save, and run the experiment. It should look similar to what is shown below.



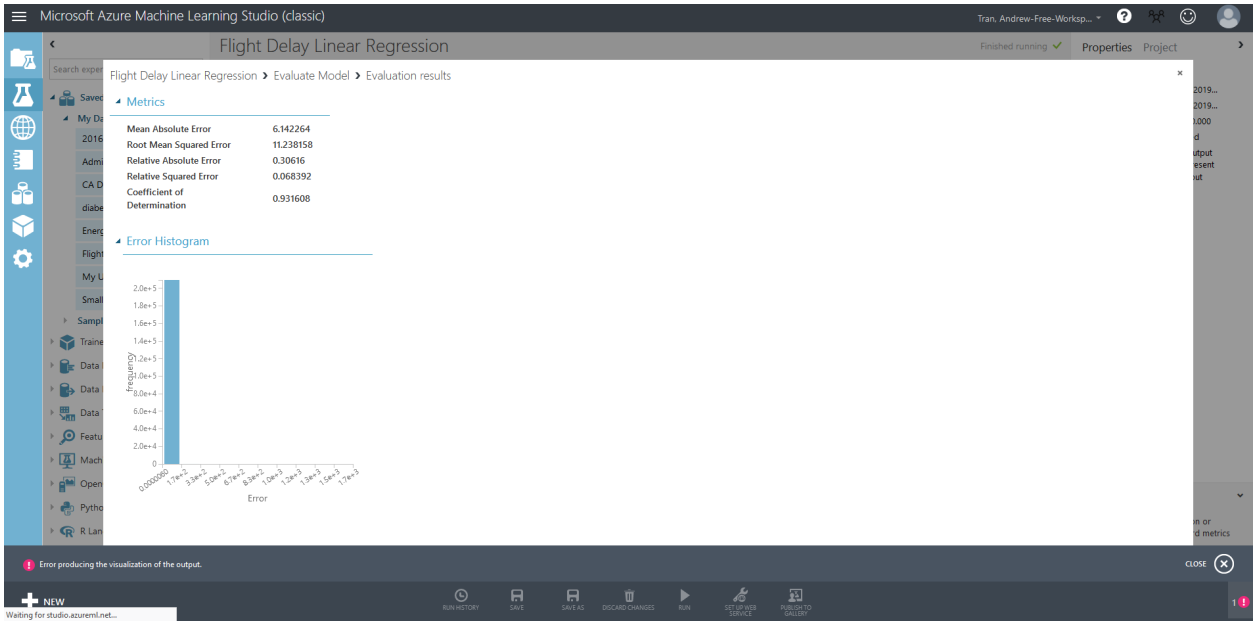
Step 3: Interpret Results

After running the experiment, we can visualize the results. Click on the **Score Model**'s output and select visualize. You should see something like this. We can see **the Scored Labels** as the model's attempt to

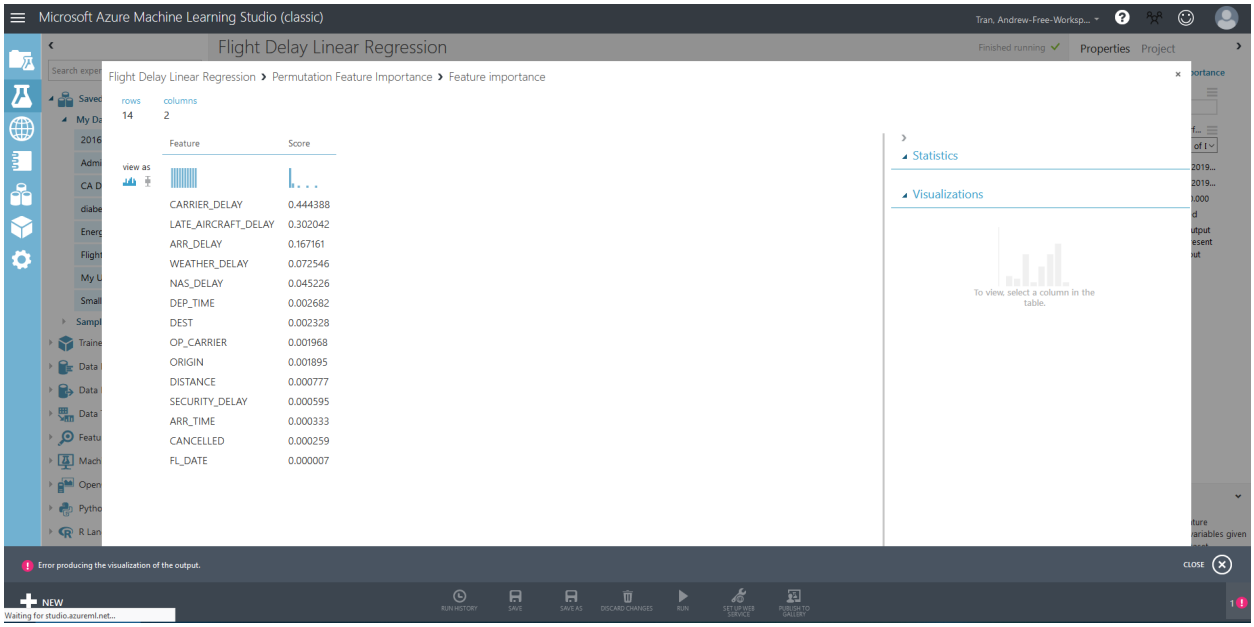
estimate departure delays. A lot of them are fairly close.



1. Select the **Evaluate Model** module's output and click **Visualize**. You should see what is shown below. Notice here, we have a **Root Mean Squared Error** of 11 and **Coefficient of Determination** of .93. What this means is that the model we have built is very accurate; (i.e.) Flight Delays can very accurately be predicted based on the other variables.



2. Click on the **Permutation Feature Importance** module's output and select **Visualize**. You should see something similar to what is shown below.



3. This is what we have been trying to find. The Permutation Feature Importance module tells us which features had the most impact, (i.e.) The most importance. We can use this to see which factors have the biggest role in causing flight delays. We can see that **CARRIER_DELAY**, **LATE_AIRCRAFT_DELAY**, and **ARR_DELAY** are in the top 3.

References

15. Dataset Source URL: https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018#2013.csv

16. GitHub URL: https://github.com/spatta983/CIS3200-Group4