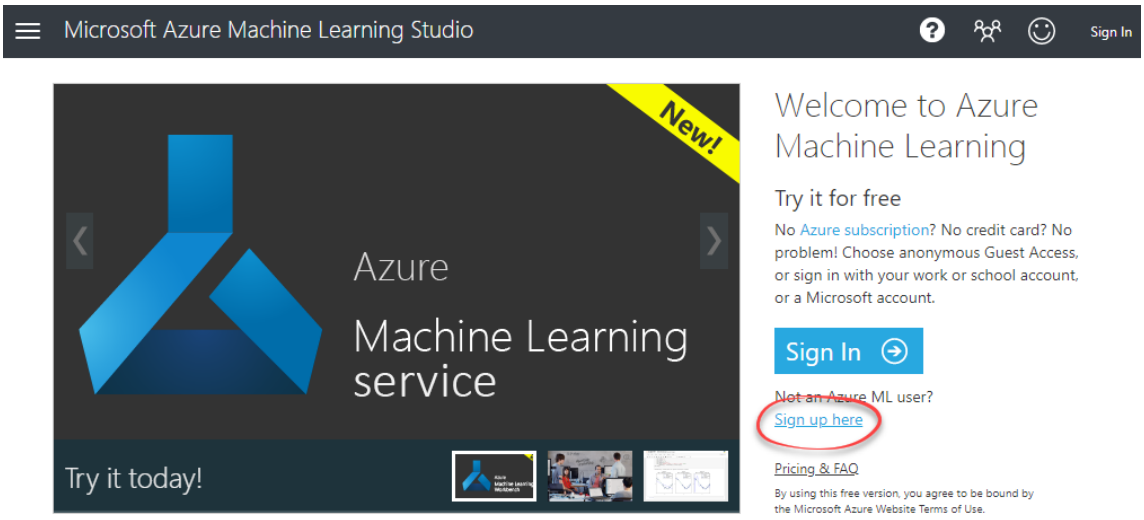
Future & Aspiring Data Professionals Workshop

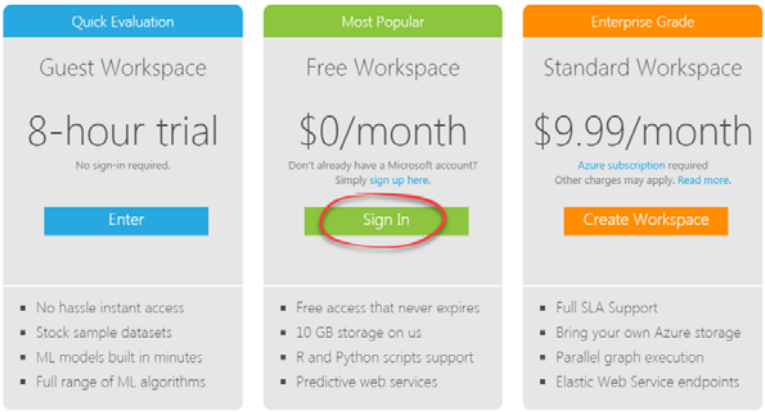
Registration

This section will start you down the path of registering for free workshop space in Azure Machine Learning.

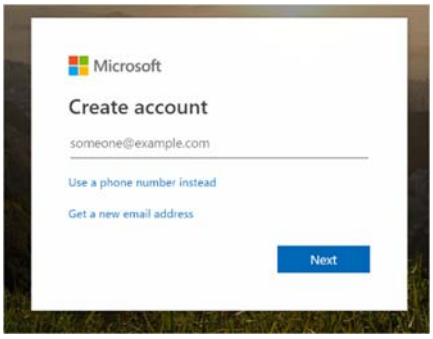
1. Go to https://studio.azureml.net/
2. Click **Sign up here**



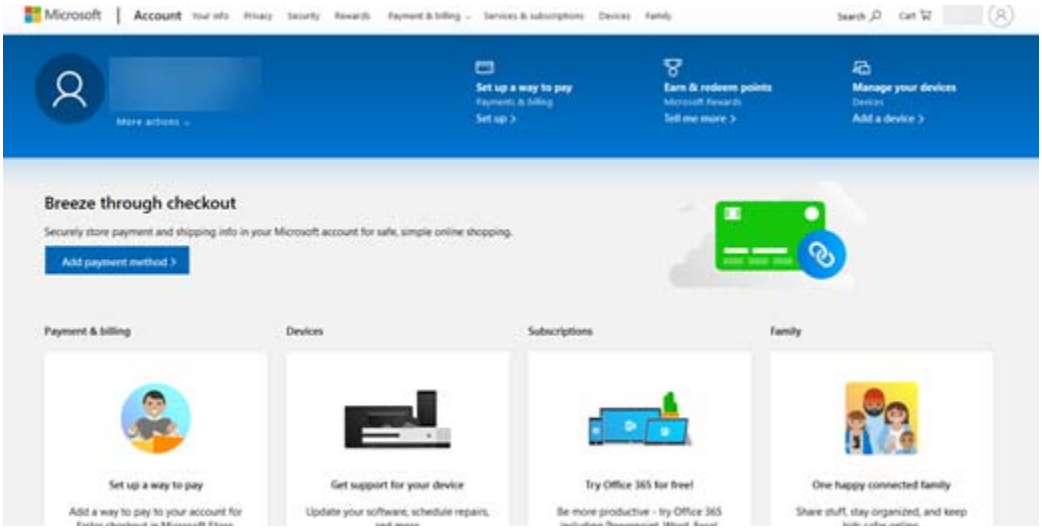
3. Select “Free Workspace” option as shown below and then click **Sign In**



4. Create an account if you don’t have a Microsoft account already. You may use an existing email address (@gmail.com, etc.) or Get a new email address from Microsoft (@outlook.com).



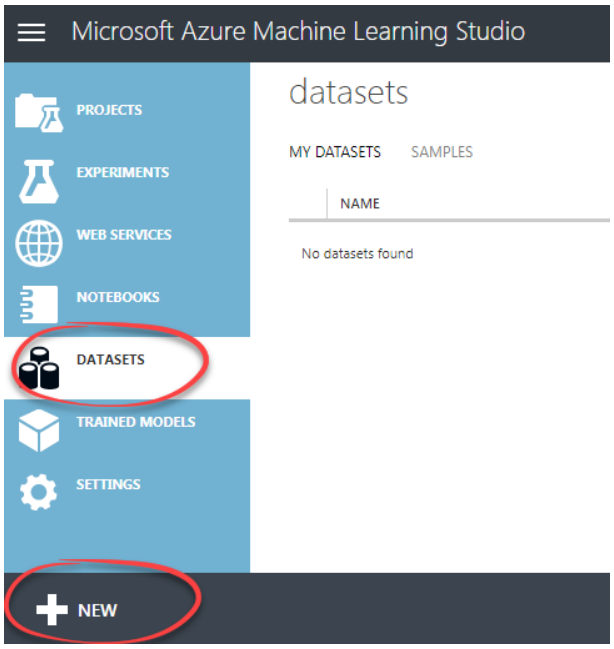
Follow the prompts given and once you get to this page you have successfully made an account.



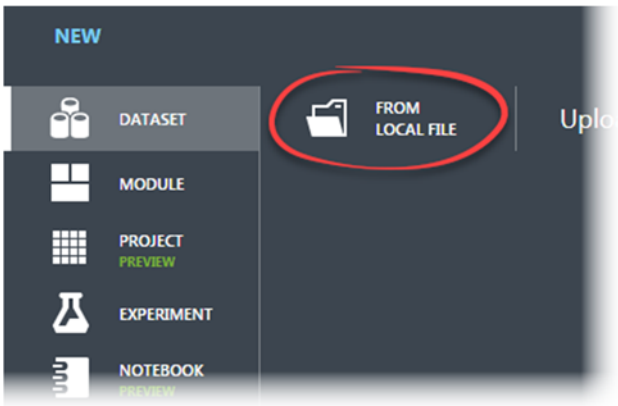
Data upload and creating an experiment

This section will walk you through uploading datafiles to your free workspace. As you increase your level of support, then you will be allowed to use your own databases stored in Azure, but this demonstration is meant to introduce you to algorithms. Most of the datasets for this demonstration came from https://www.kaggle.com/snehal1409/movielens.

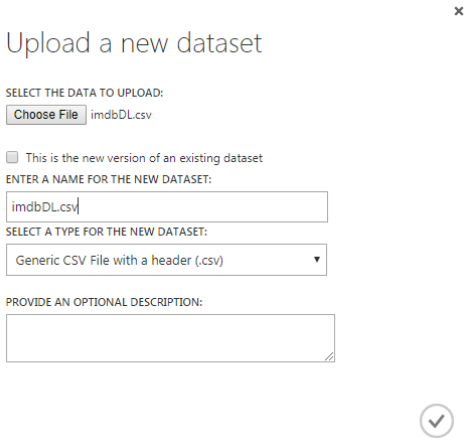
1. Sign into https://studio.azureml.net/
2. Select **DATASETS** on the left side of the workspace.
3. Click on **NEW** on the lower left corner.



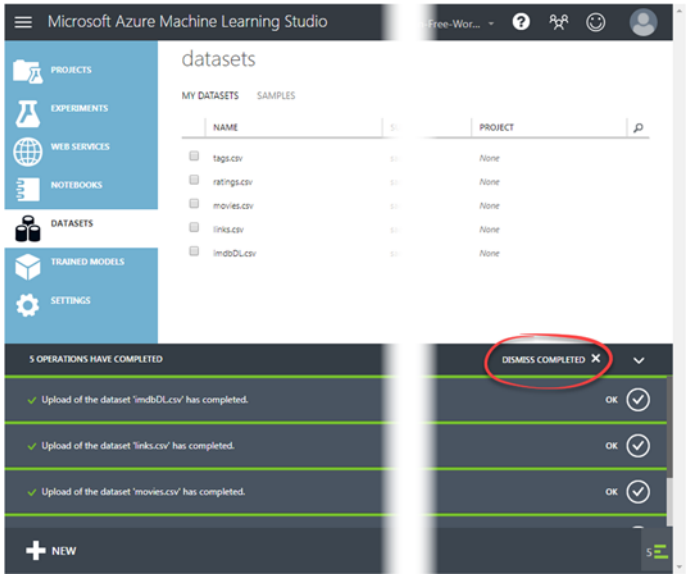
4. Click **FROM LOCAL FILE**.



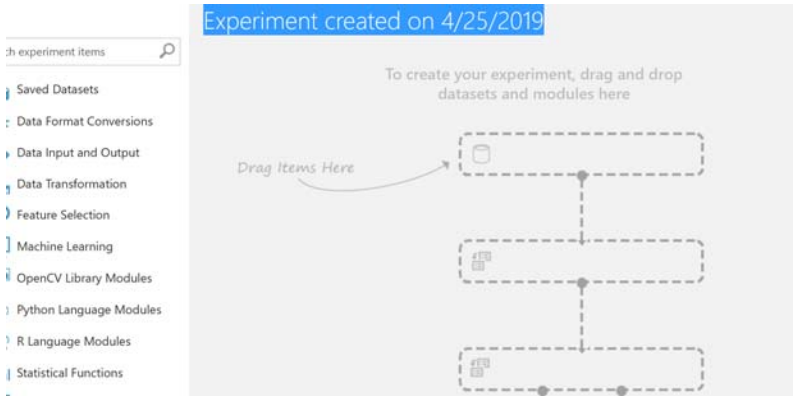
5. Click **BROWSE** and follow the instructor's directions to find the folder that contains the files. You will add each file individually. You may use the default settings with each file and complete the addition by clicking the check mark.



6. Repeat the above steps until you have added all five files as shown below and then click **DISMISS COMPLETED**.



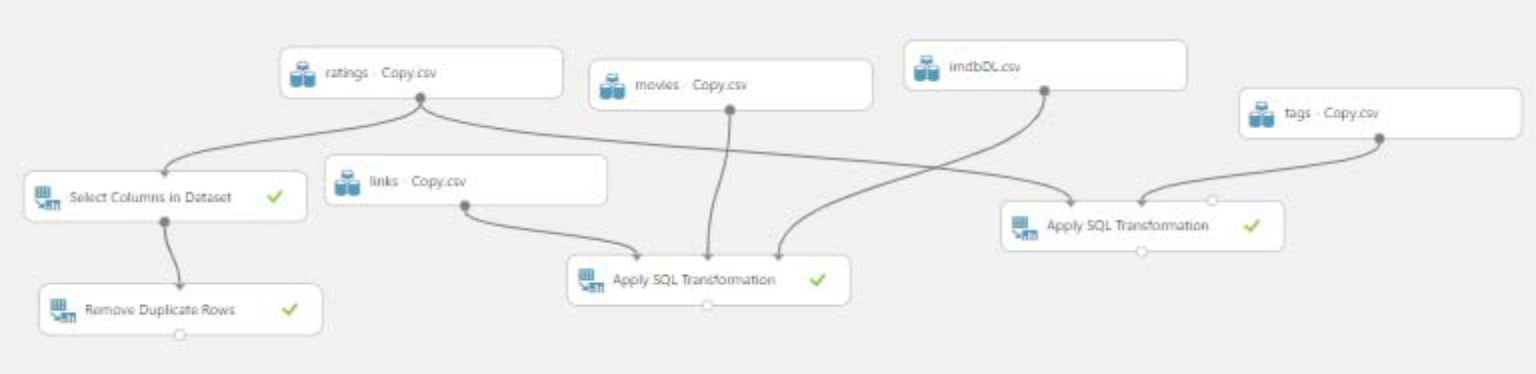
7. Click on **EXPERIMENTS** in the upper left corner and then select **NEW** in the lower left corner.
8. Click **Blank Experiment**. Your screen should appear like the figure below. You may rename the experiment.



Future & Aspiring Data Professionals Workshop

Adding data – Cleaning data- Creating Datasets (User, Item, And Rating)

This section is designed to introduce you to some of the concepts of data integration. The files that we are working with may be viewed as representing different databases. While this is a primitive demonstration of data integration, it is important to know the concepts. For the matchbox recommender, you will use ratings, information about the user, and information about the item (movie). By the end of this section your model will appear as follows.



1. Go to **Saved Datasets**, click on **My Datasets**, then drag in *imdbDL.csv*, *links.csv*, *movies.csv*, *ratings.csv*, and *tags.csv*.
2. From the Data Transformation group, in the Manipulation group, drag a **Select Columns in Dataset** and connect the *ratings.csv* output to the input.
3. Click on *Select Columns in Dataset* and click **Launch column selector** and select the userId, movieId, and rating columns.
4. Next, drag an **Apply SQL Transformation** and connect *links.csv* to the first input, *movies.csv* into the second input, and *imdbDL.csv* into third input.
5. Click on **Apply SQL Transformation** and click in the *SQL Query Script* area and copy and paste this block of SQL into the blank space.

```
SELECT *
FROM t1 JOIN t2 ON t1.movieid = t2.movieid
JOIN t3 ON t1.imdbid = t3.tid;
```

6. Right-click on the Apply SQL Transformation block and click **Run selected**.
7. Click the output connector and select **Visualize** to visualize the transformation. You should see a table. The row count should be 5,449 and the column count 51.

rows 5449 columns 51

	movieId	imdbId	tmdbId	movieId (2)	title	genres	tid
view as							
1	114709	862	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	114709
2	113497	8844	2	Jumanji (1995)	Adventure|Children|Fantasy	113497
3	113228	15602	3	Grumpier Old Men (1995)	Comedy|Romance	113228
4	114885	31357	4	Waiting to Exhale (1995)	Comedy|Drama|Romance	114885
5	113041	11862	5	Father of the Bride Part II (1995)	Comedy	113041
6	113277	949	6	Heat (1995)	Action|Crime|Thriller	113277
7	114319	11860	7	Sabrina	Comedy|Romance	114319

8. Different join types return different data. With the `LEFT OUTER JOIN` you will return all the data from the table on the left regardless of if there is a match on the right. Copy and paste this SQL code into the Apply SQL transformation module.

```
SELECT *
FROM t1 LEFT OUTER JOIN t2 ON t1.movieid = t2.movieid
LEFT OUTER JOIN t3 ON t1.imdbid = t3.tid;
```

9. Run the experiment again and visualize the output from the Apply SQL Transformation again. The row count should now be 9,125.

rows 9125 columns 51

	movieId	imdbId	tmdbId	movieId (2)	title	genres	tid	title (2
view as								
1	114709	862	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	114709	Toy St (1995)
2	113497	8844	2	Jumanji (1995)	Adventure|Children|Fantasy	113497	Juman
3	113228	15602	3	Grumpier Old Men (1995)	Comedy|Romance	113228	Der d Frühli
4	114885	31357	4	Waiting to Exhale (1995)	Comedy|Drama|Romance	114885	Waitin Exhale Warte Right
5	113041	11862	5	Father of the Bride Part II (1995)	Comedy	113041	Ein Ge des H Vater 2 (19
6	113277	949	6	Heat (1995)	Action|Crime|Thriller	113277	Heat

10. Drag another **Apply SQL Transformation** and connect the *ratings.csv* to the first input and the *tags.csv* to the second input.
11. Click on **Apply SQL Transformation** and open the SQL Query Script and copy and paste this block of SQL to replace the suggestion.

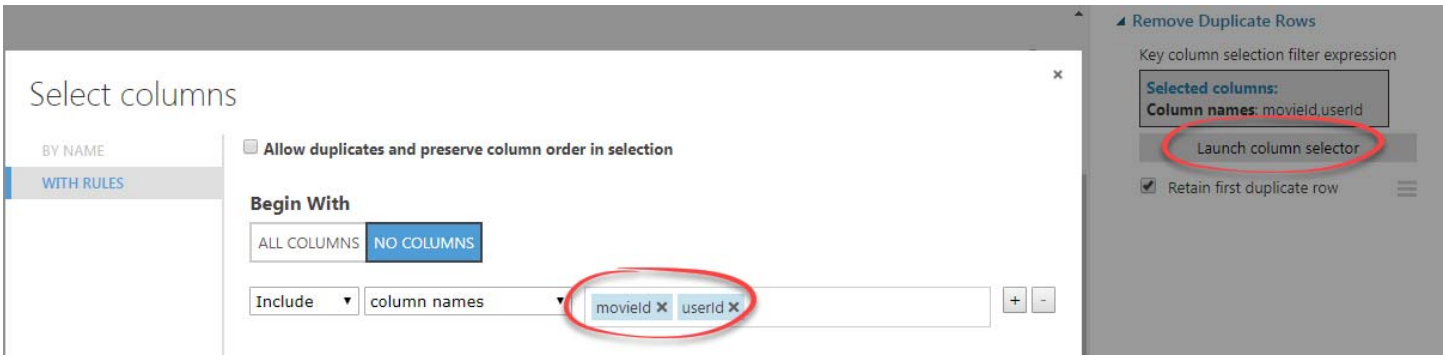
```
SELECT t1.userid, AVG(rating) as ARating, CASE WHEN t2.movieid IS NULL
THEN 0 ELSE 1 END AS tag
FROM t1 LEFT OUTER JOIN t2 ON t1.userid=t2.userid
GROUP BY t1.userid;
```

rows 671 columns 3

view as

	userId	ARating	tag
1	1	2.521739	1
2	2	3.577778	0
3	3	3.744048	0
4	4	4.525	0
5	5	3.926573	0
6	6	3.492308	1
7	7	3.948529	0
8	8	3.927313	0
9	9	3.813333	1
10	10	3.77027	0
11	11	4.140351	0
12	12	2.686747	0
13	13	3.528409	0

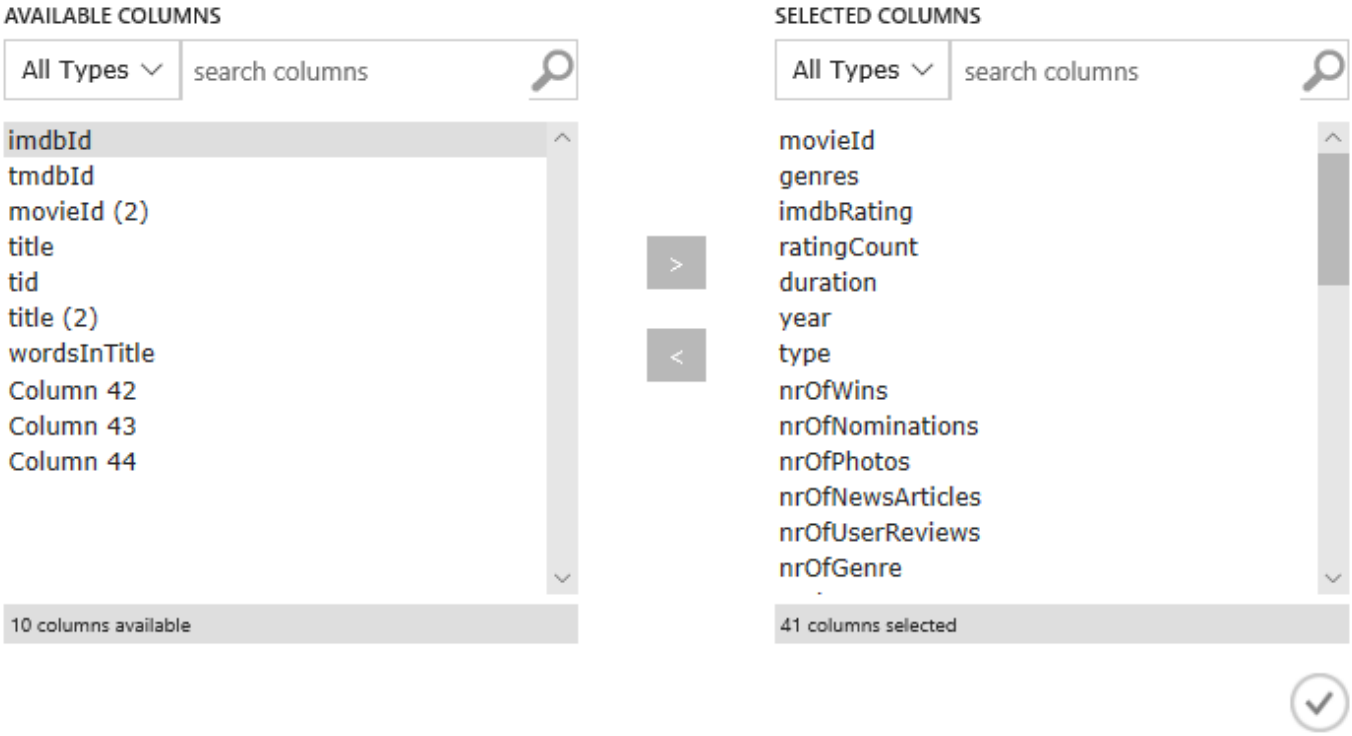
12. Find and add a **Remove Duplicate Rows** module to the *Select Columns in Dataset*. Click **Launch column selector**. Select the *movieId* and *userId* to include by column name. Remove duplicates ensures that there are no duplicate ratings for a user of a particular movie. Once you have done this, run the experiment.



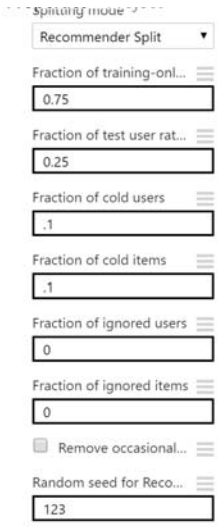
Recommender Model (Split – Scoring)

This section splits the data so that we may have a dataset to train our machine learning with as well as one to evaluate the machine learning. Azure machine learning includes many different models. In this demonstration we are using the matchbox recommender.

1. Search for the **Select Columns in Dataset** module on the left-hand side of the screen.
2. Drag a **Select Columns in Dataset** module and connect the output of the *Apply SQL Transformation* off of the 'links,' 'movies,' and 'imdbId' datasets.
3. Click the *Select Columns in Dataset* and launch the column selector. Click **BY NAME** and select the desired columns as shown below. Exclude the left columns and keep the right columns (41 columns).

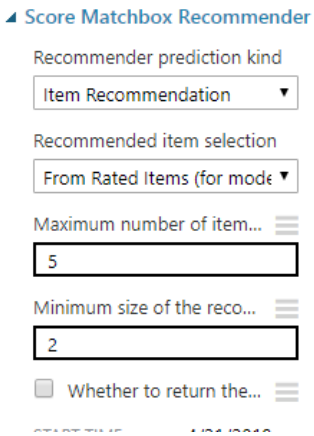


4. Search for the **Split Data** module on the left-hand side of the screen and drag it into to the model, connecting the output of the Remove Duplicate Rows module.
5. Click the Split Data module and set the following parameters for the split module:

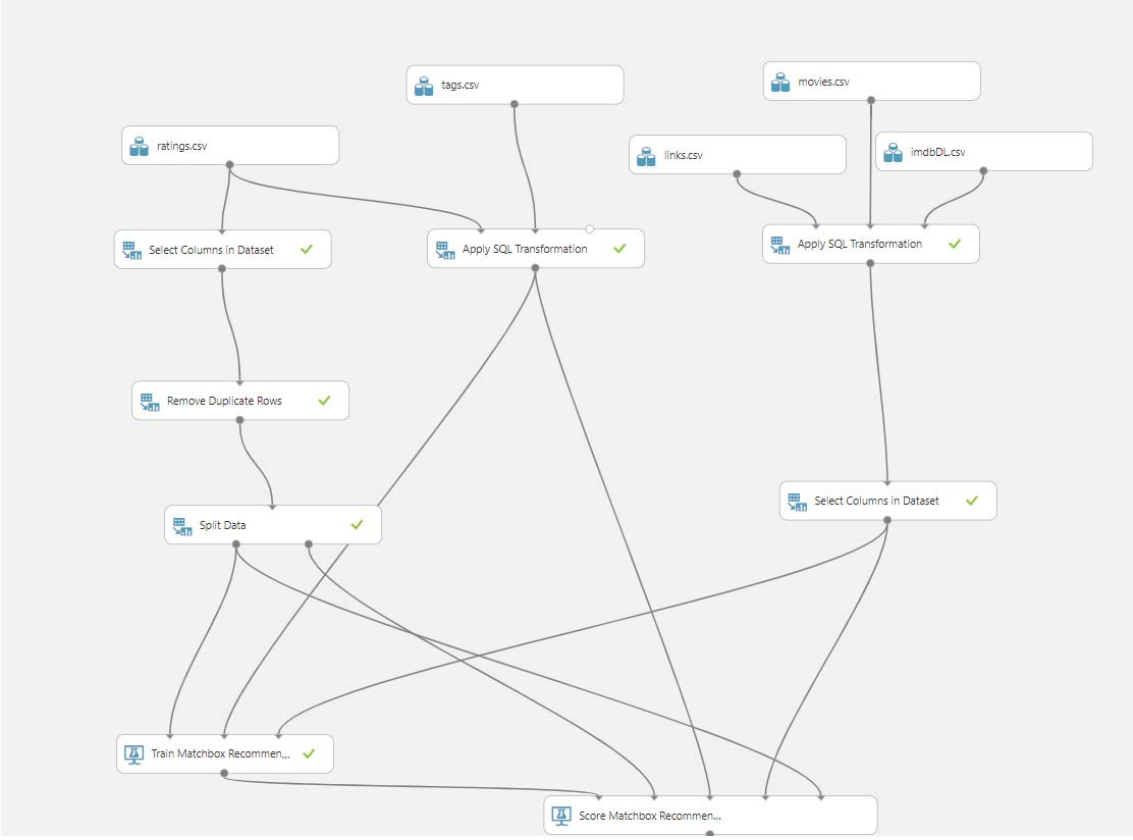


6. Add the **Train Matchbox Recommender** module and connect as follows:
    - Left side of the Split Data module to the left side of the Train Matchbox Recommender module
    - Apply SQL Transformation off the 'tags' dataset to the center
    - Select Columns in Dataset off of the 'links', 'movies', and 'imdbDL' datasets to the right side
7. Click the Train Matchbox Recommender module and input the following properties:

8. Search for and add the **Score Matchbox Recommender** module to the model. Connect the following to the Score Matchbox Recommender module in the following manner:
    - Train Matchbox Recommender to the outer left side
    - Right side of the Split module to the inner left side
    - Apply SQL Transformation off the 'tags' dataset to the center
    - Select Columns in Dataset off the 'links', 'movies', and 'imdbDL' datasets to the inner right side
    - Left side of the Split Data to the outer right side of the Score Matchbox Recommender module
9. Click the Score Matchbox Recommender module and select the following properties:



10. At this point your model should look something like this:



Evaluate recommender (Scoring and Evaluate recommender)

Machine learning is an iterative process. This puts an emphasis in knowing how to evaluate each model. This section will add the evaluate model and demonstrate making a change to the model to see if it improves the performance.

1. Search for the **Evaluate Recommender** module and drag and drop it into the model. Attach the right output from the Split Data and attach it to the left side input of the Evaluate Recommender. Next, attach the output of the Score Matchbox Recommender and attach it to the right input of the Evaluate Recommender.

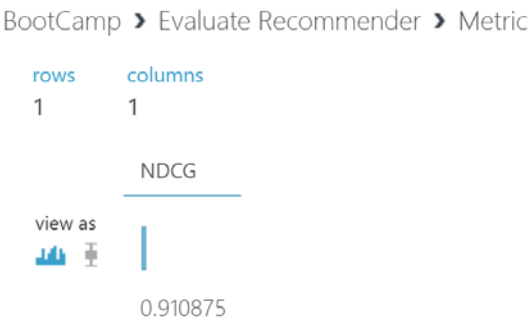
2. Run the program.
3. Click to visualize the results of the scored model. You will see 5 recommended items for each user. These are movies that the user has scored so that the we may evaluate the model.

rows 641 columns 6

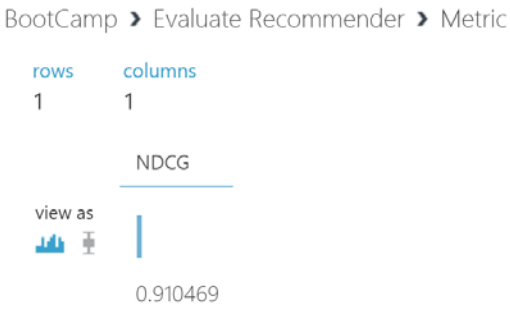
view as

User	Item 1	Item 2	Item 3	Item 4	Item 5
73	1293	111	3424	4103	36
475	7502	1221	85342	296	106920
199	111	104879	80489	541	59810
544	1147	7502	1217	106920	527
176	260	7153	1196	1198	5989
442	2571	1200	1240	260	589
30	1197	912	527	913	1945

4. Click to visualize the results of the evaluate model. This will show us the NDCG. This is a measure of the accuracy of ratings and the value will range from 0 – 1 with a value that is closer to 1 being more accurate. This value uses the ratings given to movies in the test dataset to evaluate whether the model provided movies that the user would like.



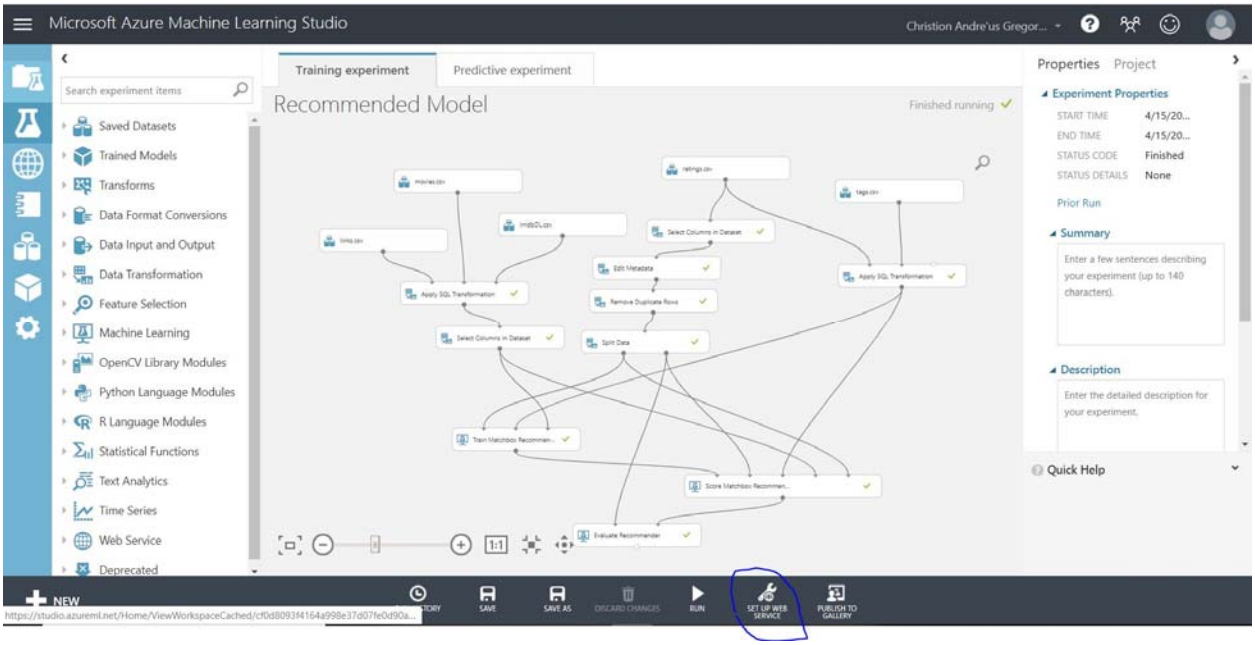
5. In an iterative process, we should change the model and evaluate whether the model improves. One thing that we can change is add the column “Words in the Title”. We can then rerun the model and see if it improved the results. If it does not improve the model, then remove it.



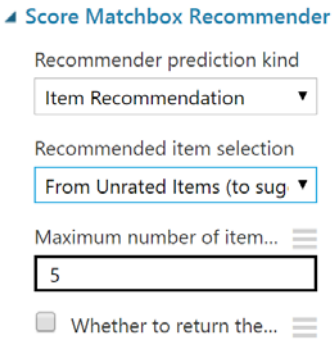
Deploying the Model

This section shows you how to deploy the model to an API that allows you to reference the model through other applications. We will also download an Excel spreadsheet and demonstrate using the spreadsheet to show the recommended movies by each user.

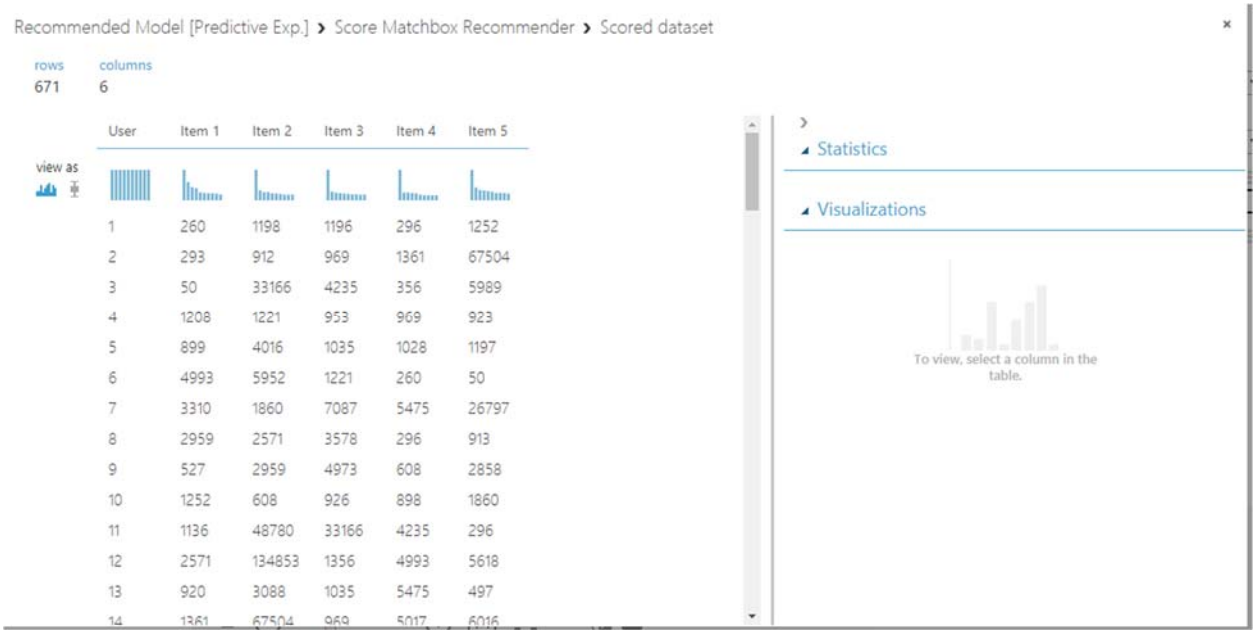
1. With the Movie Recommendations experiment open, click the **SET UP WEB SERVICE** icon at the bottom of the Azure ML Studio page and click **Predictive Web Service [Recommended]**. A new Predictive Experiment tab will be automatically created.



2. Select the Score Matchbox Recommender module and change the recommended item selection property from Rated Items (for model evaluation) to From Unrated Items (All items would suggest new or previously watched).

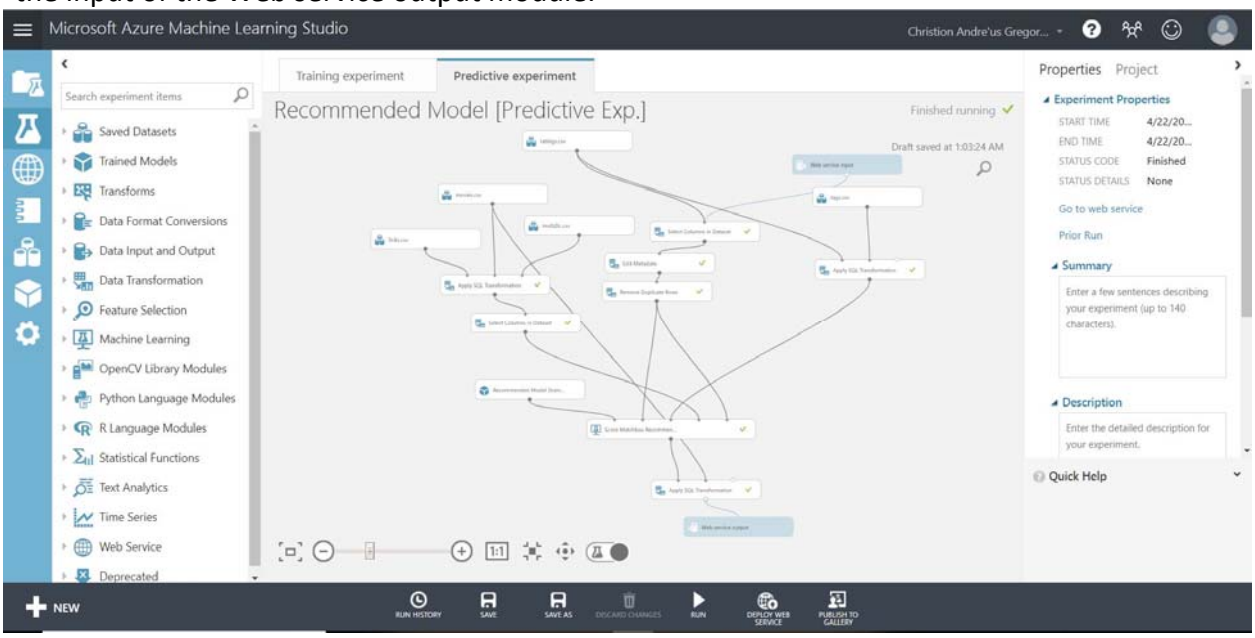


3. Run the predictive experiment (this can take a while – over 10 minutes in some cases)
4. When the experiment has finished running, visualize the output from the Score Matchbox Recommender module, verifying that it shows three recommendations for each user. However, the recommendations are movie IDs, and the web service will be more useful if it returns movie titles.



	User	Item 1	Item 2	Item 3	Item 4	Item 5
1		260	1198	1196	296	1252
2		293	912	969	1361	67504
3		50	33166	4235	356	5989
4		1208	1221	953	969	923
5		899	4016	1035	1028	1197
6		4993	5952	1221	260	50
7		3310	1860	7087	5475	26797
8		2959	2571	3578	296	913
9		527	2959	4973	608	2858
10		1252	608	926	898	1860
11		1136	48780	33166	4235	296
12		2571	134853	1356	4993	5618
13		920	3088	1035	5475	497
14		1361	67504	969	5017	6016

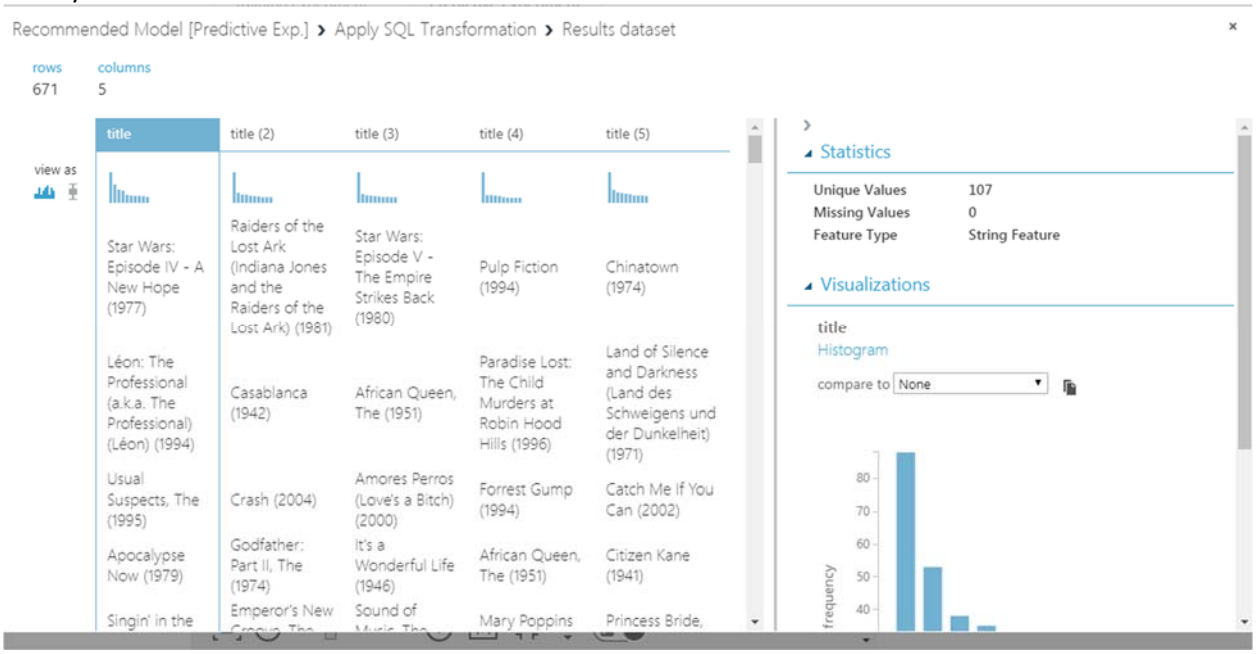
5. Add an Apply SQL Transformation module to the experiment and drag the output from the Score Matchbox Recommender to its Table1 (left-most) input and drag the output from the IMDB Sample dataset to its Table2 (middle) input. Then drag the output of the Apply SQL Transformation module to the input of the Web service output module.



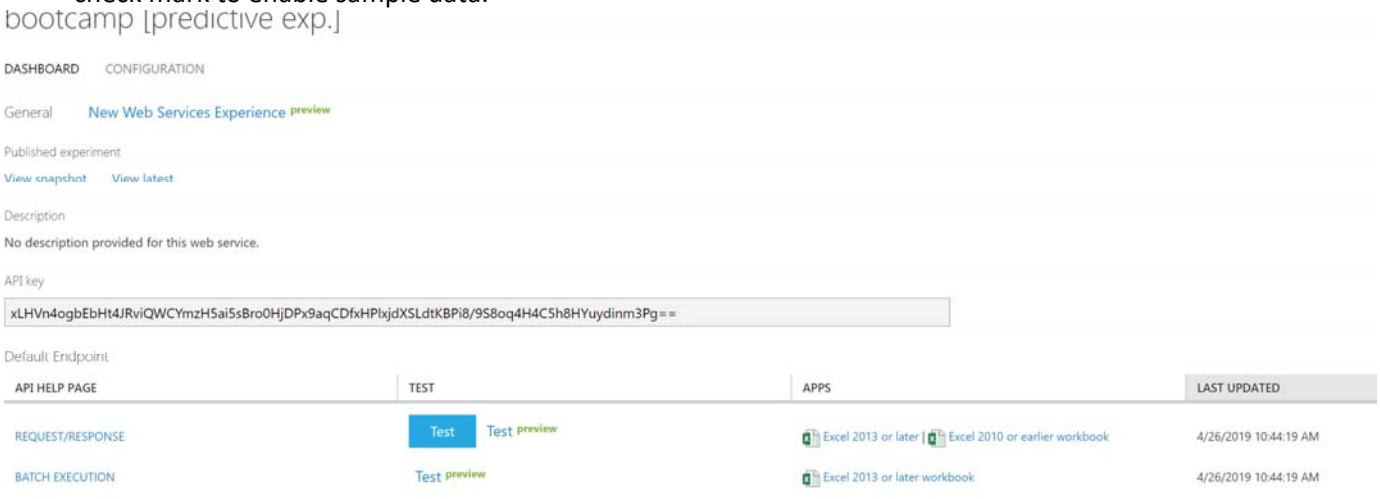
6. Select the Apply SQL Transformation, and replace its default SQL script with the following code:

```
SELECT r1.[title], r2.[title], r3.[title] , r4.[title], r5.[title]
FROM t1 JOIN t2 AS r1 ON t1.[Item 1] = r1.[movieId]
JOIN t2 AS r2 ON t1.[Item 2] = r2.[movieId]
JOIN t2 AS r3 ON t1.[Item 3] = r3.[movieId]
JOIN t2 AS r4 ON t1.[Item 4] = r4.[movieId]
JOIN t2 AS r5 ON t1.[Item 5] = r5.[movieId];
```

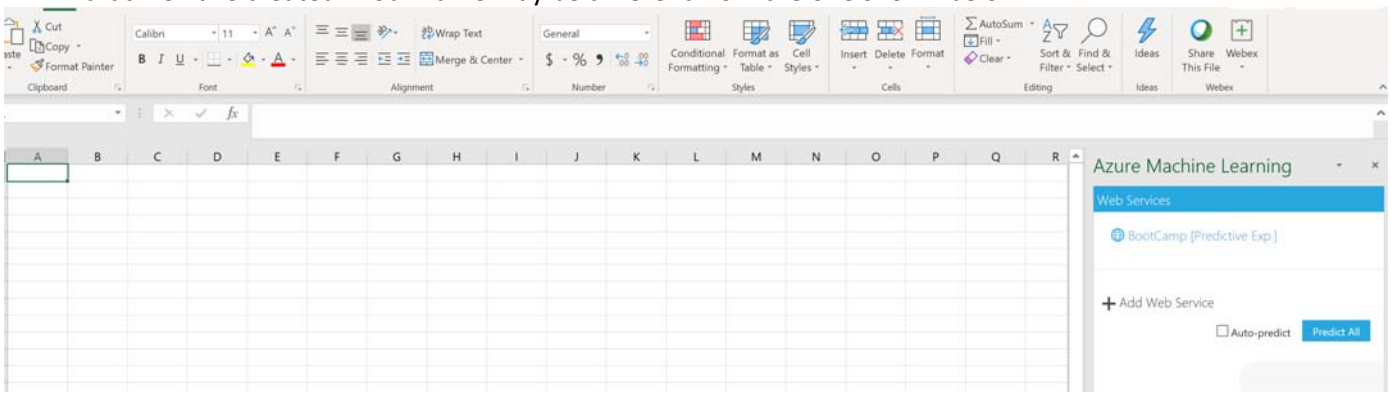
7. Save and run the experiment again. Then visualize the output of the Apply SQL Transformation module and verify that the recommended movie titles are returned.



8. In the Movie Recommendations [Predictive Exp.] experiment, click the Deploy Web Service icon at the bottom of the Azure ML Studio window.
9. Click on Excel 2013 or later link to download the spreadsheet you will be working with. You should click the check mark to enable sample data.



10. Once you have downloaded and opened the spreadsheet enable it for editing and select the predictive model that we have created. Your name may be different from the one shown below.



11. Now we will select to use sample data and set the Input to be the same cells within the spreadsheet. You should also set the userids to be 1,2,3,4,5 and the output to be cell E1. You can then click to predict, and the output will appear as below.