CIS5560 Term Project Tutorial

Authors: Swarnim Jambhule, Tanvi Gawade, Sushant Burde

Instructor: [Jongwook Woo]

Date: 05/17/2019

Lab Tutorial

Swarnim jambhule

Tanvi Gawade

Sushant Burde

05/17/2019

Seattle library books data On Microsoft Azure Machine Learning

Objectives

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

- Get data manually
- Train data in the system
- Predicting total number of books available in a library using Decision Forest Regression and Boosted Decision Tree.
- Visualization
- **https://gallery.cortanaintelligence.com/Experiment/CIS-5560-Project-2**

Overview

In this lab, you will train and evaluate a classification model. Classification is one of the fundamental machine learning methods used in data science. Classification models enable you to predict classes or categories of a label value. Classification algorithms can be two-class methods, where there are two possible categories, or multi-class methods. Like regression, classification is a supervised machine learning technique, wherein models are trained from labeled cases.

Public libraries are one of society's great institutions. They provide an opportunity for anyone with an appetite to read, learn and socialize with their community. We specifically focus on the Floating Item types from the Seattle Public Library Inventory Data Collections. The Library uses floating collection management model for some item collections. Floating collections do not assign items to "owning" locations and instead allows those items to move around the library system based on where patrons are checking them out from and returning them to. Items usually stay at a branch until they get checked out and get returned to a different branch or are delivered to another branch to fulfill a hold request.

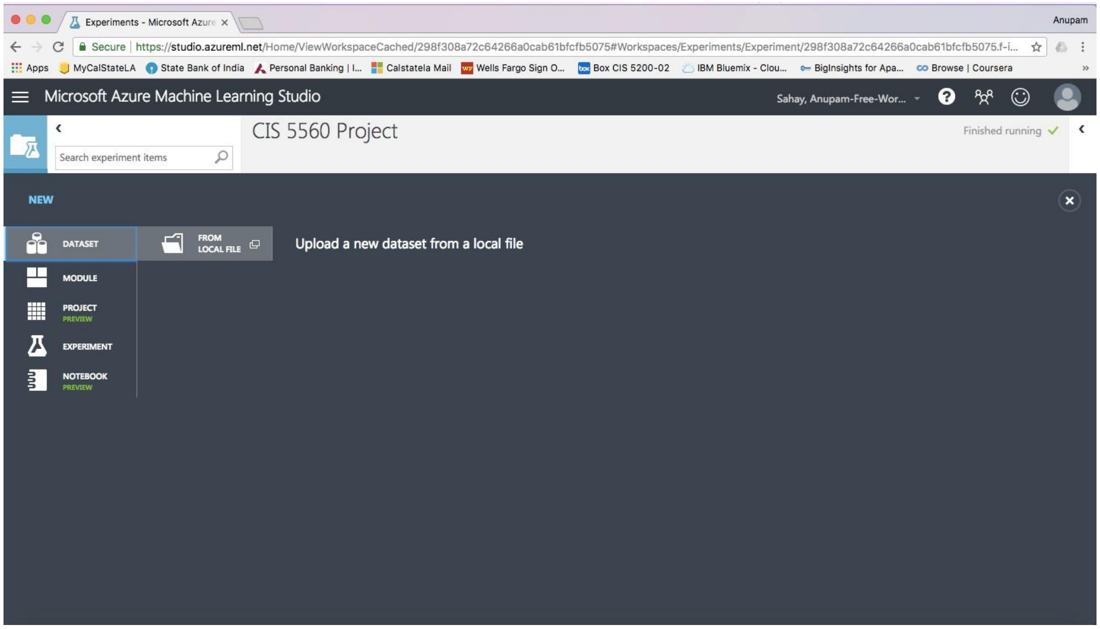
Libraries occasionally must rebalance the floating collection when too many of a specific item are returned to the same branch. Floating collections make books available for patrons more quickly, while reducing staff time and delivery vehicle expenses. Collections get refreshed continuously, meaning branch collections better reflect what their patrons are using. Furthermore, there's less wear-and-tear on materials, and centralized selectors don't need to make branch-by-branch decisions on who receives a copy. It is generally accepted that floating collections lead to an increase in checkouts. Based on generous list of data provided by our instructor, we have done some researches and exclusively decided which data we are using for this project. We have created a model to Classify whether the item collection is of a Floating type or Non-Floating type based on Feature columns the number of trips taken on a day

Platform Spec

- Microsoft Azure Machine Learning
- CPU Speed: 3.4GHz
- # of nodes: 1
- Total Memory Size: 10GB

Step 1: Upload the Data Set from the Local File

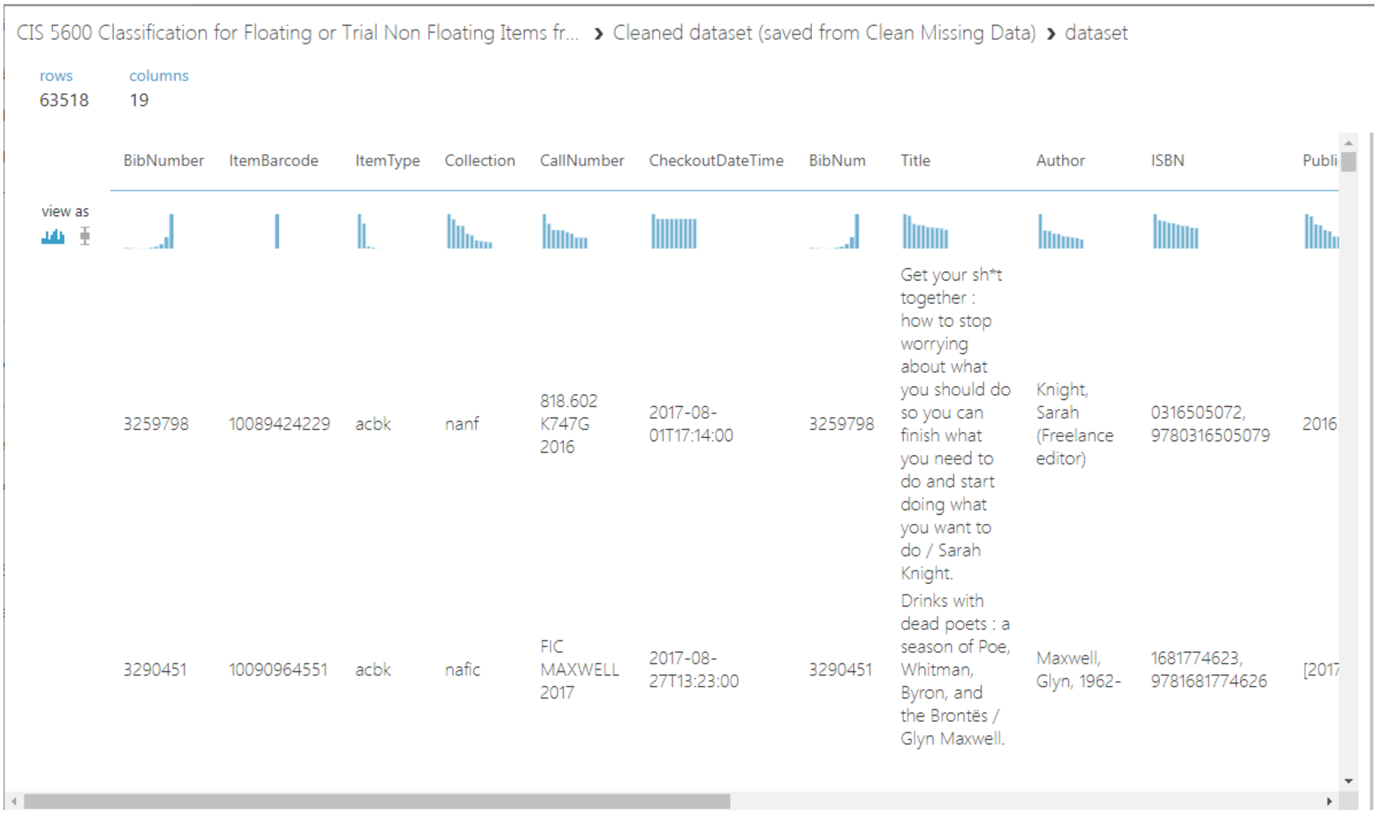
1. This step is to upload Library_collection_inventory.csv, checkouts_by_tittle.csv, intregredted_library_system.csv



- This dataset is available in the Kaggle website and was last updated 2 years ago
- All the dataset should be in format of Generic CSV file with a header(.csv)

Step 2: Visualization of the Dataset Loaded in Azure ML

This step is to verify if all the columns are present in the dataset from source.



CIS 5600 Classification for Floating or Trial Non Floating Items fr... › Cleaned dataset (saved from Clean Missing Data) › dataset

rows 63518 columns 19

BibNumber	ItemBarcode	ItemType	Collection	CallNumber	CheckoutDateTime	BibNum	Title	Author	ISBN	Publi
3259798	10089424229	acbk	nanf	818.602 K747G 2016	2017-08-01T17:14:00	3259798	Get your sh*t together : how to stop worrying about what you should do so you can finish what you need to do and start doing what you want to do / Sarah Knight.	Knight, Sarah (Freelance editor)	0316505072, 9780316505079	2016
3290451	10090964551	acbk	nafic	FIC MAXWELL 2017	2017-08-27T13:23:00	3290451	Drinks with dead poets : a season of Poe, Whitman, Byron, and the Brontës / Glyn Maxwell.	Maxwell, Glyn, 1962-	1681774623, 9781681774626	[2017

Step 3: Building a Classification Model

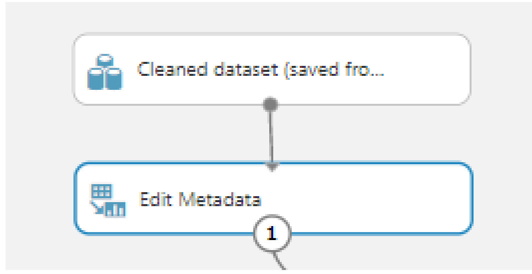
This step is to verify if all the columns are present in the dataset from source.

In this exercise, you will select all the string type data and select specific data column.We will select column, As the in classification, "itemType" is a code from the catalog record that describes the type of item. Some of the more common codes are: acbk (adult book), acdvd (adult DVD), jcbk (children's book), accd (adult CD) "Collection" is a collection code from the catalog record which describes the item. Here are some common examples: nanf (adult non-fiction), nafic(adult fiction), ncpic(children's picture book), nycomic (Young adult comic books).

1. Drag the **Cleansed dataset (Clean)** and onto the canvas.

2. Search for the **Edit Metadata** module and drag it onto the canvas. Connect the output

of the data set to the input (**Dataset1**) port of the **Edit Metadata** module. At this point your experiment should resemble the following:



3. Click the **Edit Metadata** module, and in the Properties pane, and set the parameter as follows:
   a) Launch column selector and select all String in column.
   b) Data type: Unchanged
   c) Categorical: Make categorical
   d) Fields: Unchanged

4. Click **Column selector**, and connect the output of Edit metadeata to **Coloum selector** in the Properties pane, and set the parameter as follows:
   1.1) Launch column: **Exclude** Itemtype (2)

5. Lunch **Split Data** and collect the output of edit metadata to **Split data.** In properties plane do the following.
   a) Splitting mode: Split Rows
   b) Fraction of rows in the first output dataset: 0.6
   c) Randomized split: Selected
   d) Randomized seed: 0
   e) Stratified split: False

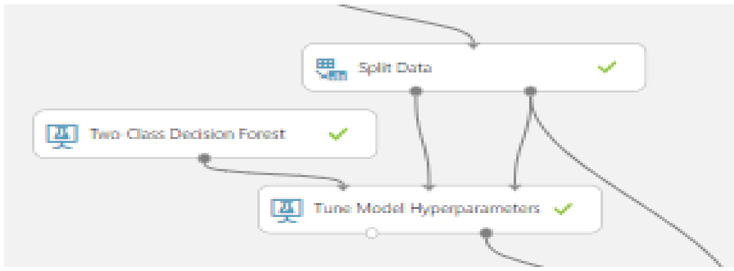
6. Select and drag **Tune Model Hyperparameter** to workspace, connect the output of Split data to Second input of **Tune Model Hyperparameter**, and connect the second output of split data to Third input of Tune Model Hyperparameter. In properties plane do the following.
   a) Specify parameter sweeping mode: Random sweep
   b) Maximum number of random sweep: 30
   c) Randomized seed: 4567
   d) Label column: Floating item
   e) Metric for measuring performance of classification: Recall
   f) Metric for measuring performance for regression: Mean absolute error.

(**part-1 Using Two-class Decision Forest**)

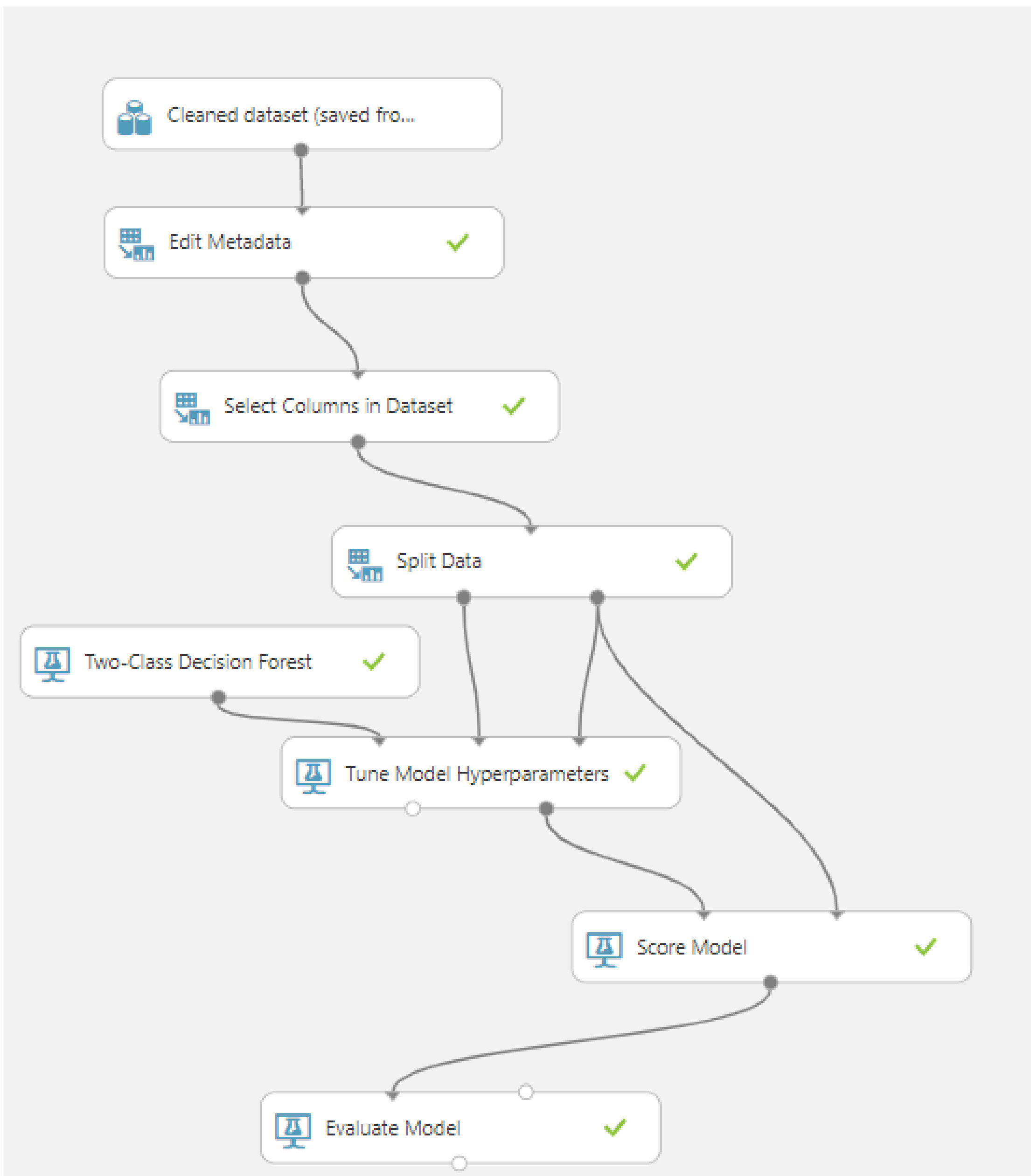
7. Select and drag **Two-class Decision Forest** into workspace and collect the output of **Two-class Decision Forest** to the first input of **Tune Model Hyperparameter.** In properties plane do the following.

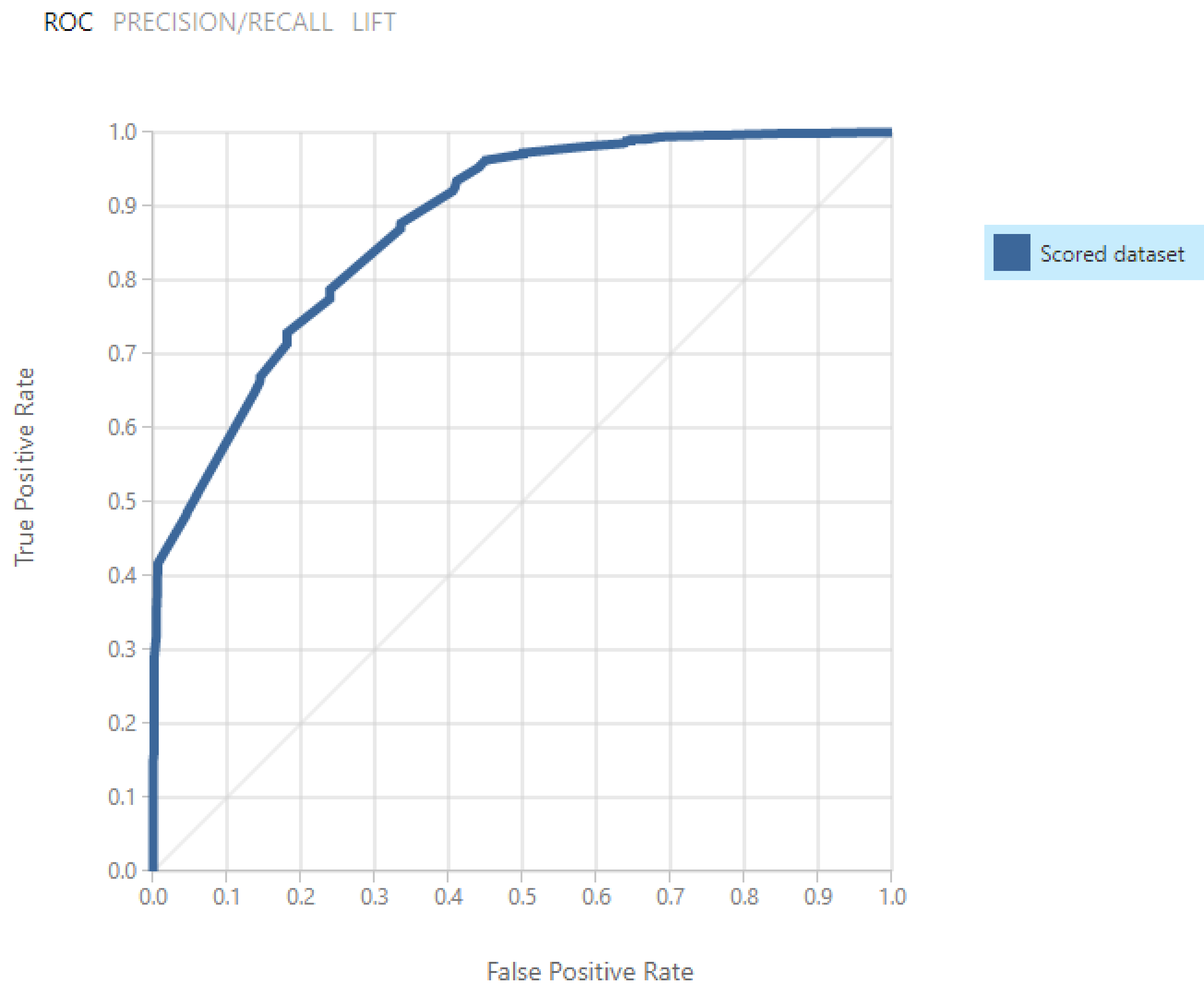
   a) Resampling method: Bagging
   b) Create trainer mode: Single parameter
   c) Number of decision trees: 40
   d) Number of random splits per node: 128

e) Maximum depth of decision tree: 32
f) Minimum number of samples per leaf node 4.
g) Allow unknow value for categorical features: selected.



a





Examine this ROC curve. Notice that the bold blue line is well above the diagonal grey line, indicating the model is performing significantly better than random guessing. The AUC is 0.874 (see below), which is significantly more than 0.5 obtained by random.

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
21939	0	0.864	0.864	0.5	0.874
False Positive	**True Negative**	**Recall**	**F1 Score**		
3463	5	1.000	0.927		

(**part-2 Using Two-class Decision Forest**)

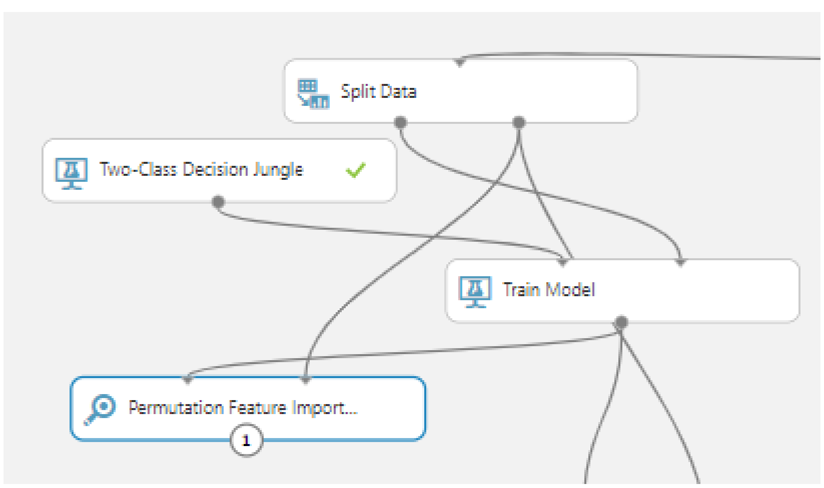
11. Lunch **Split Data** and collect the output of edit metadata to **Split data.** Connect the output from select column dataset In properties plane, do the following.

    a) Splitting mode: Split Rows
    b) Fraction of rows in the first output dataset: 0.6
    c) Randomized split: Selected
    d) Randomized seed: 0
    e) Stratified split: False

12 . Select and drag **Two-class Decision Jungle** into workspace and collect the output of **Two-class Decision Jungle** to the input of **Train model.** In properties plane do the following.

    a) Resampling method: Bagging
    b) Create trainer mode: Single parameter
    c) Number of decisions DAGS: 8
    d) Maximum width of decision DAGS : 28
    e) Maximum depth of decision DAGS: 32
    f) Number of optimization steps per desecion DAGS: 2048
    g) Allow unknow value for categorical features: selected.

13. Drag and drop **permutation feature import** and connect the **output of split data** to the right input of the **permutation feature import** and connect the output from train model to first input of **permutation feature import.**

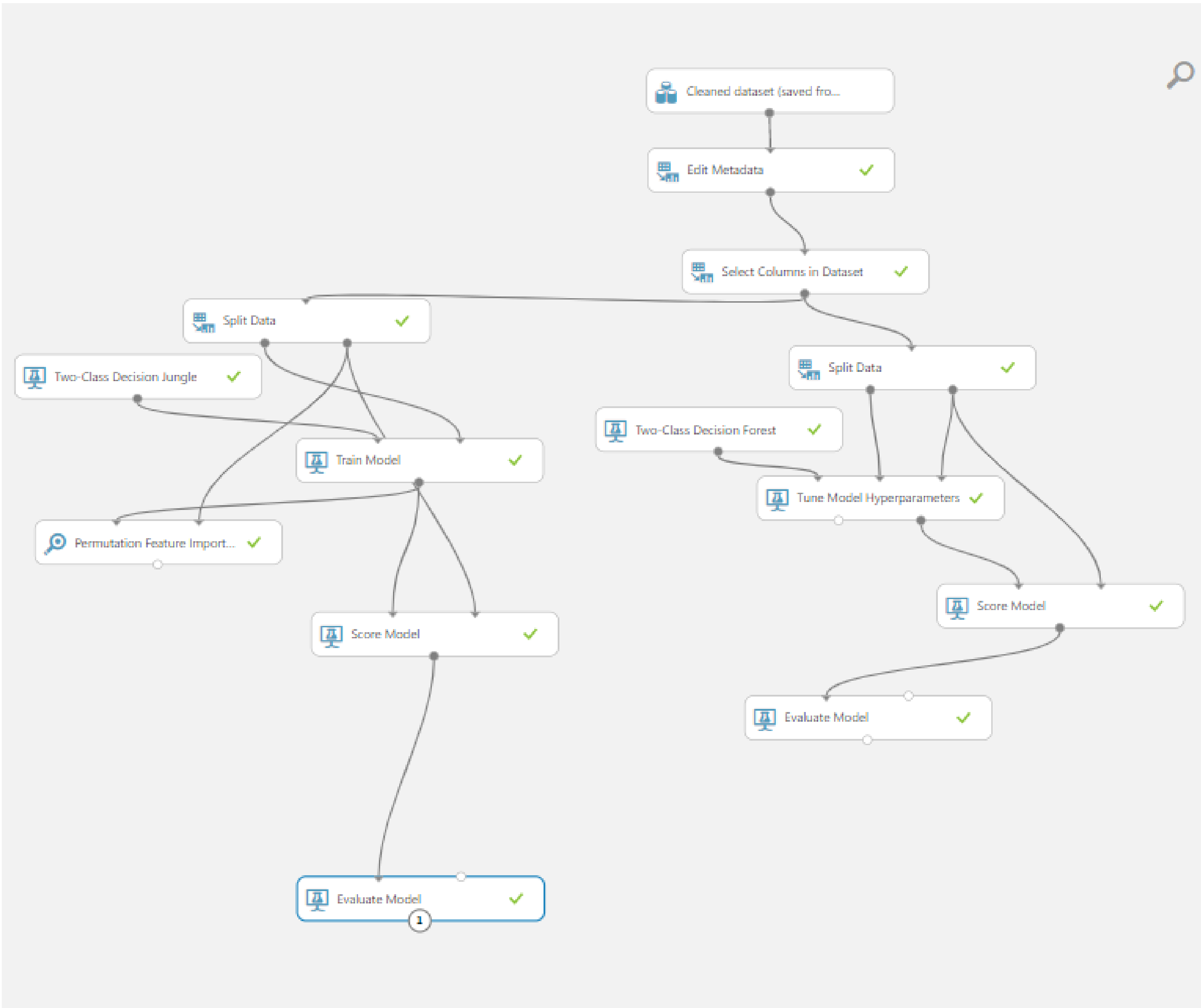


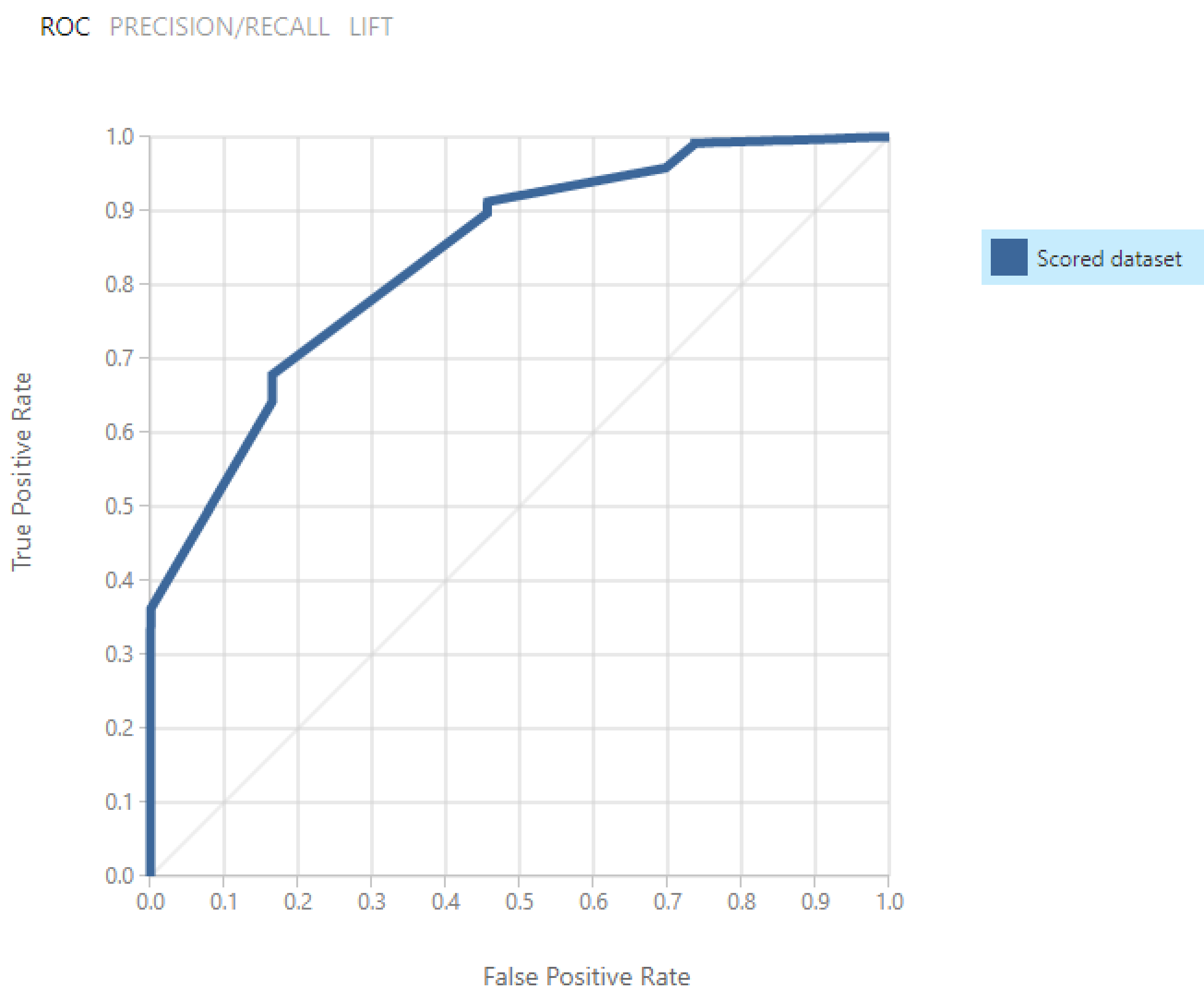
14. We will use Score model to generate predictions using a trained classification or regression model. select and drag **Score model** to workspace and connect second output of **Train model**

.

15. Now **Evaluate model** and connect the output of score model.

16. Save and run the experiment. When the experiment has finished running, visualize the output from the **Evaluate Model**.





True Positive	False Negative	Accuracy	Precision	Threshold	AUC
21938	1	0.864	0.864	0.5	0.838
False Positive	**True Negative**	**Recall**	**F1 Score**		
3458	10	1.000	0.927		

Examine this ROC curve. Notice that the bold blue line is well above the diagonal grey line, indicating the model is performing significantly better than random guessing. The AUC is 0.838 (see below), which is significantly more than 0.5 obtained by random guessing.

Summary

In this lab, you have constructed and evaluated a two class or binary classification model. Highlight from the results of this lab are:

- Visualization of the data set can help differentiate features which separate the cases from those that are unlikely to do so.
- Feature pruning and parameter sweeping can improve model performance
- Examining the classification behavior of features can highlight potential performance problems or provide guidance on improving a model.

Note: The experiment created in this lab is available in the Cortana Analytics library at https://gallery.azure.ai/Experiment/CIS-5600-Classification-for-Floating-or-Trial-Non-Floating-Items-from-the-Library-Inventory-Collections-or-Trial-Non-Floating-Items-from-the-Library-Inventory-Collections