Ozkan Emre Ozdemir 07/07/2016

Deriving Knowledge from Data at Scale

Assignment # 1

Build an experiment in Azure Machine Learning using Decision Trees or Logistic Regression.

1. What is the percentage of correct classification results (using all attributes)?
2. What is the percentage of correct classification results (using a subset of the attributes)?
3. What is the AUC of your model?
4. What is your best AUC that you can achieve?
5. Which are the minimum number of attributes? Why?

The experiment is conducted using the following options in Azure Machine Learning:

- **ReadWhiteWine.csv** data is selected
- **"Select Colum in Dataset"** option is used to define the attributes that are used in the experiment
- **"Edit Metadata"** option is used to prepare and build the model, and "R/W" column is selected
- **"Split Data"** option is used to split the data as 70% to train the model and 30 % is to test
- **"Two Class Boosted Decision Three"** option is selected and default values are used
- **"Train Data"** option is selected to and "R/W" column is selected. Model is strained using the 70% of the data
- **"Score Model"** option is selected to take the information from the testing data, run it through the model, and compare the predictions.
- **"Evaluate Model**" option is selected to evaluate the experiment predictions

The nodalization of the wine quality experiment is shown in Figure 1. The experiment is run multiple times by changing the selected column dataset based on the given tasks and results are presented for each task in the next section.

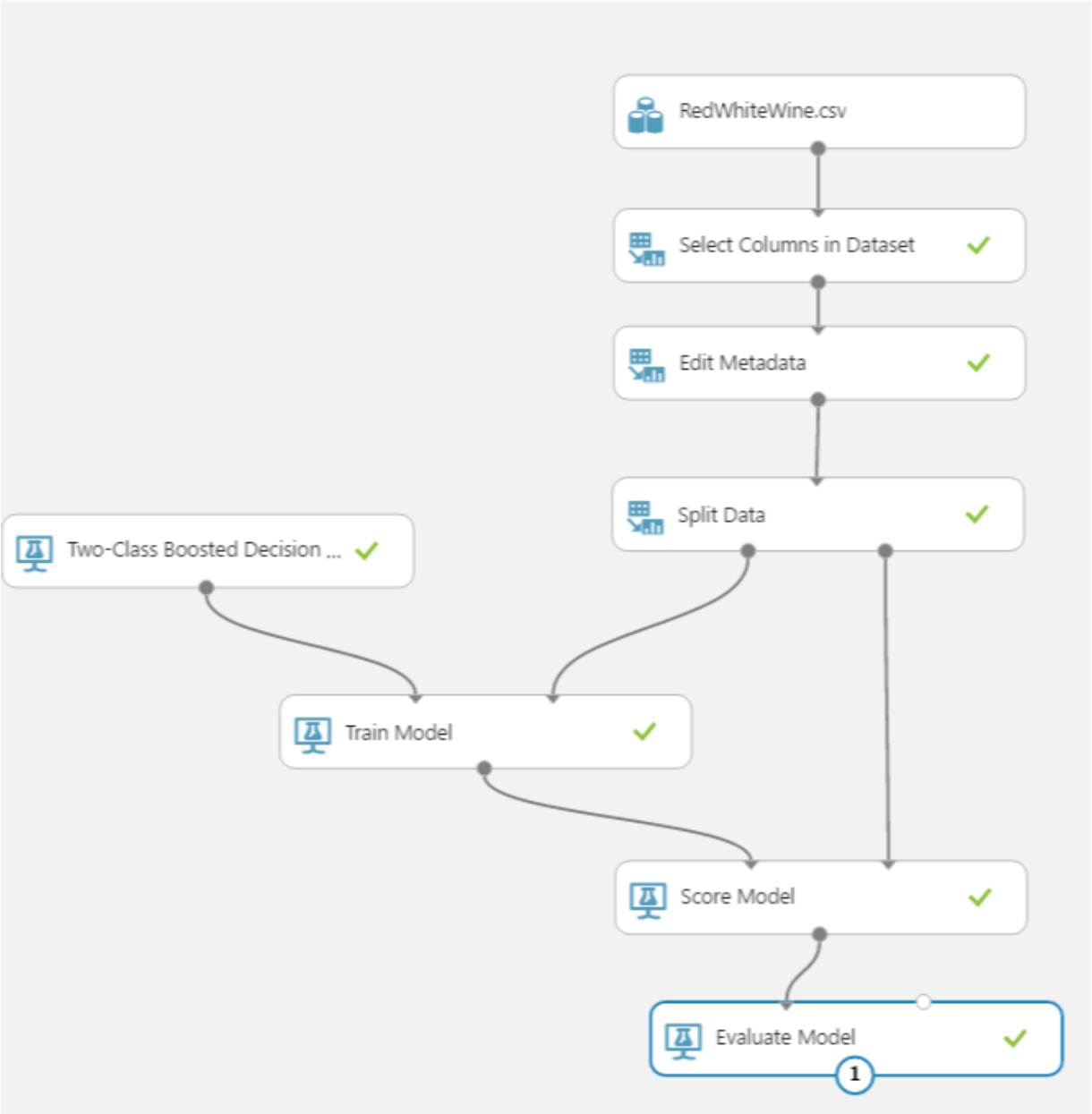
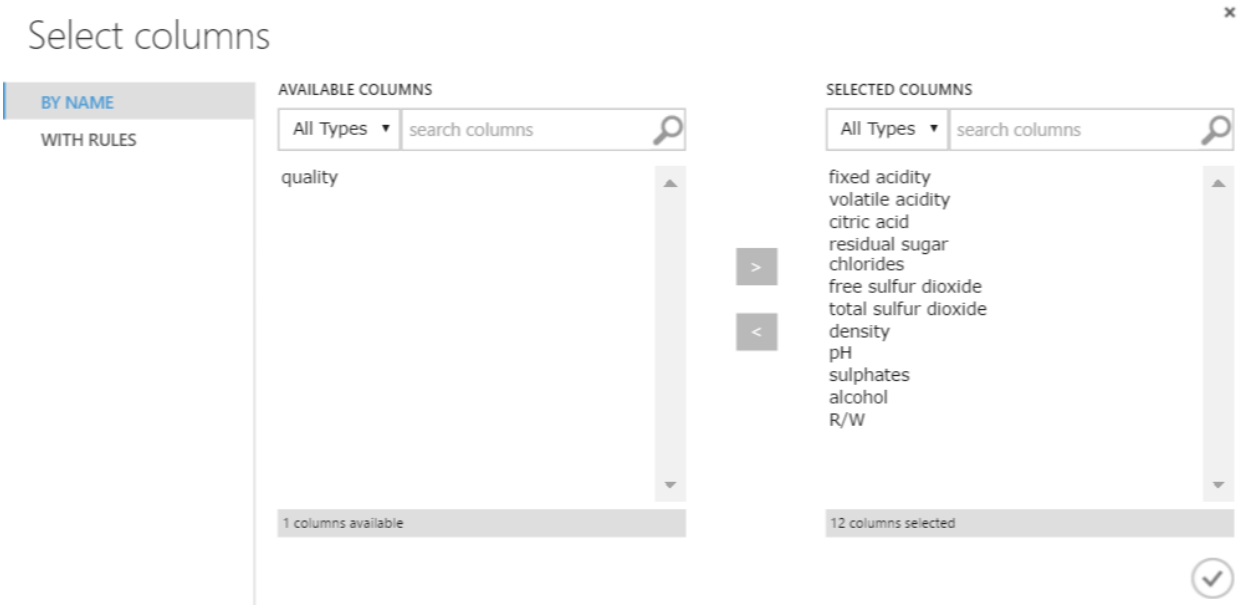
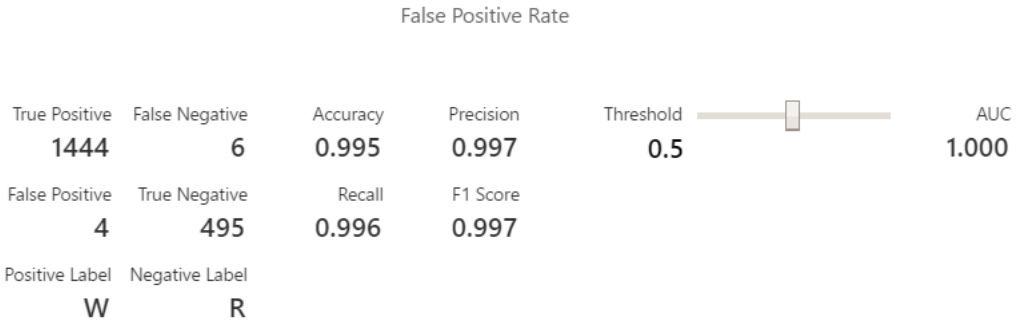


Figure 1: Wine Quality Model Experiment Nodalization

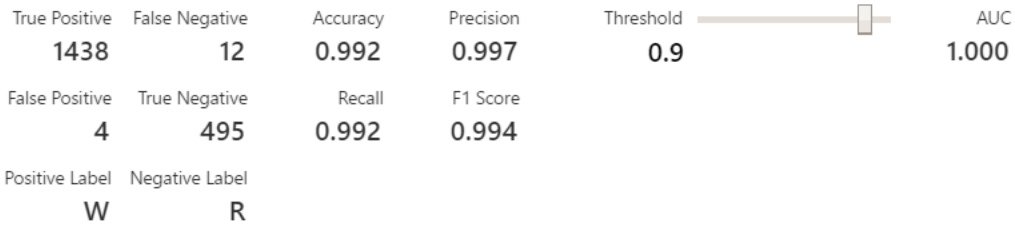
Task 1: What is the percentage of correct classification results (using all attributes)?



- Default AUC = 0.5 Results

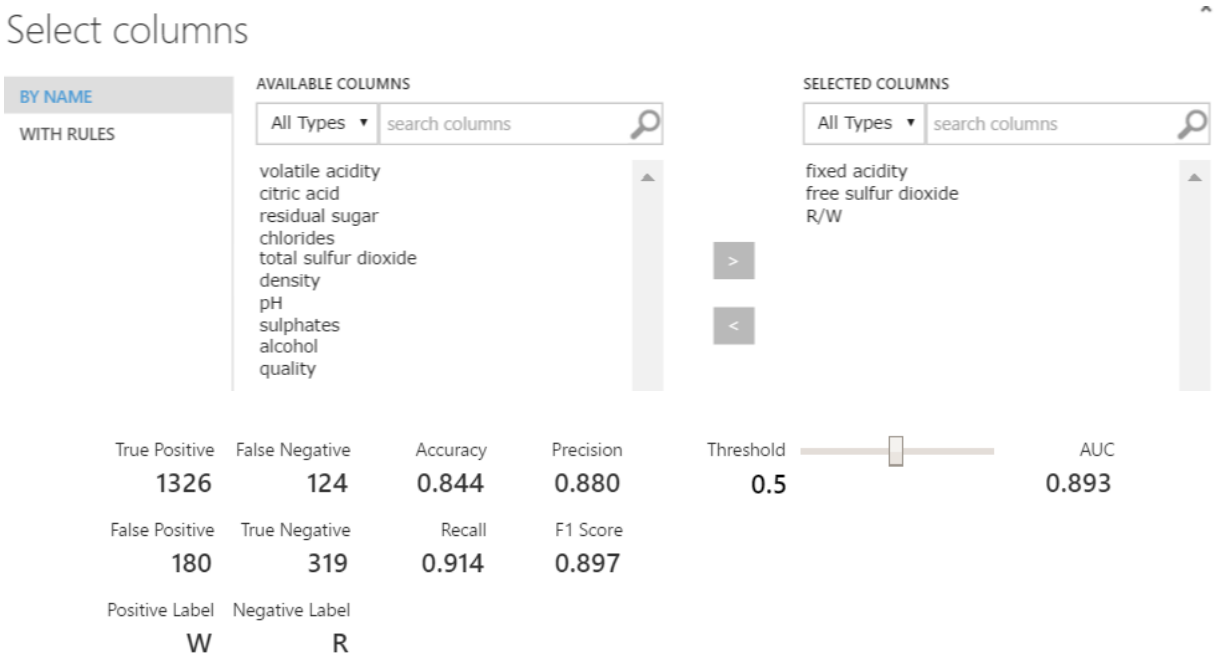


- Modified AUC = 0.9 Results

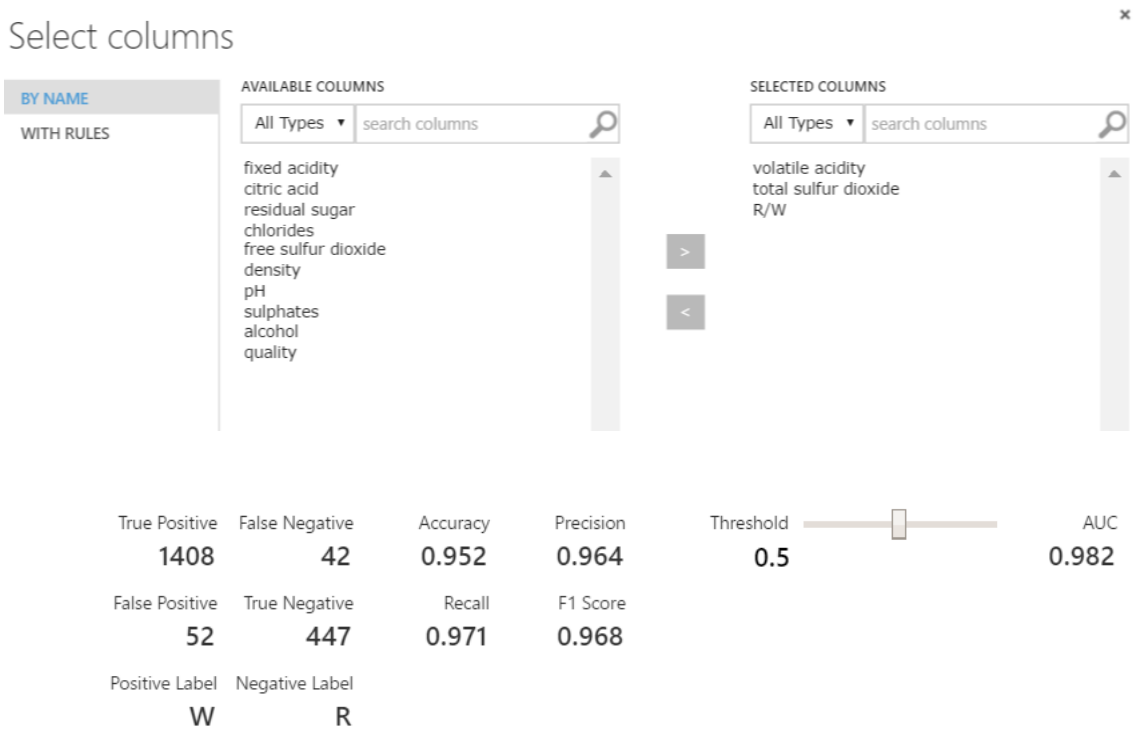


Task 2: What is the percentage of correct classification results (using a subset of the attributes)?

- **Fixed Acidity - Free Sulphur Dioxide**



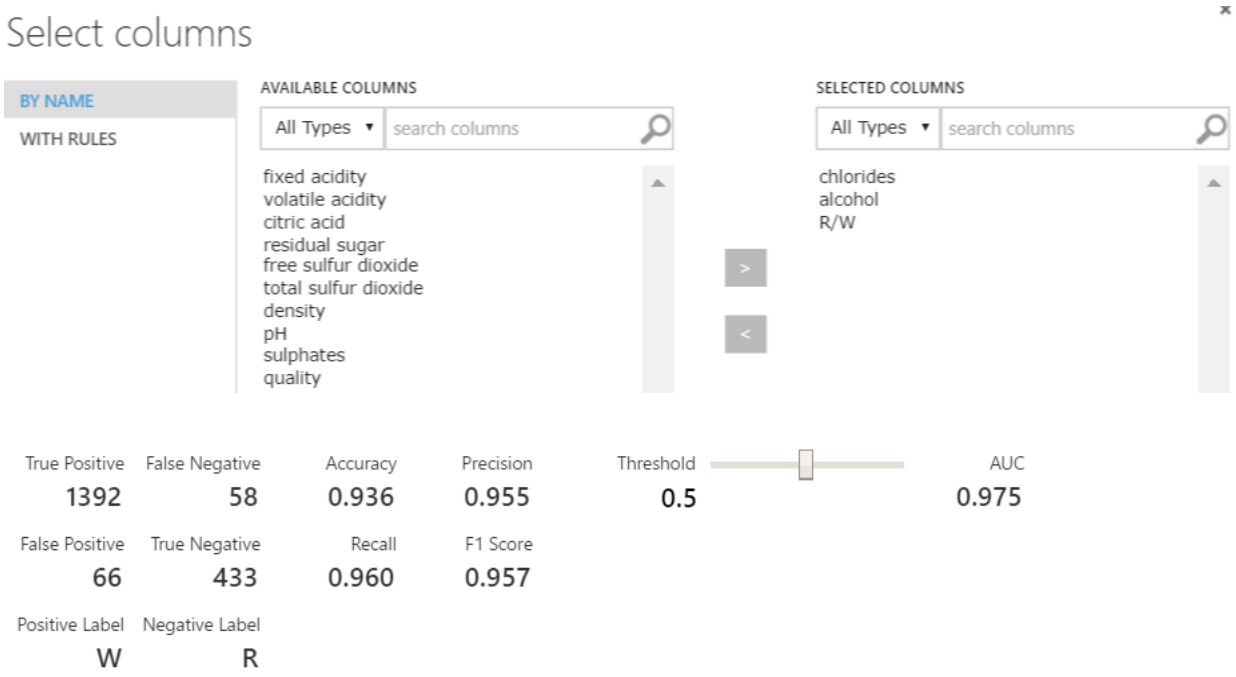
- **Volatile Acidity - Total Sulphur Dioxide**



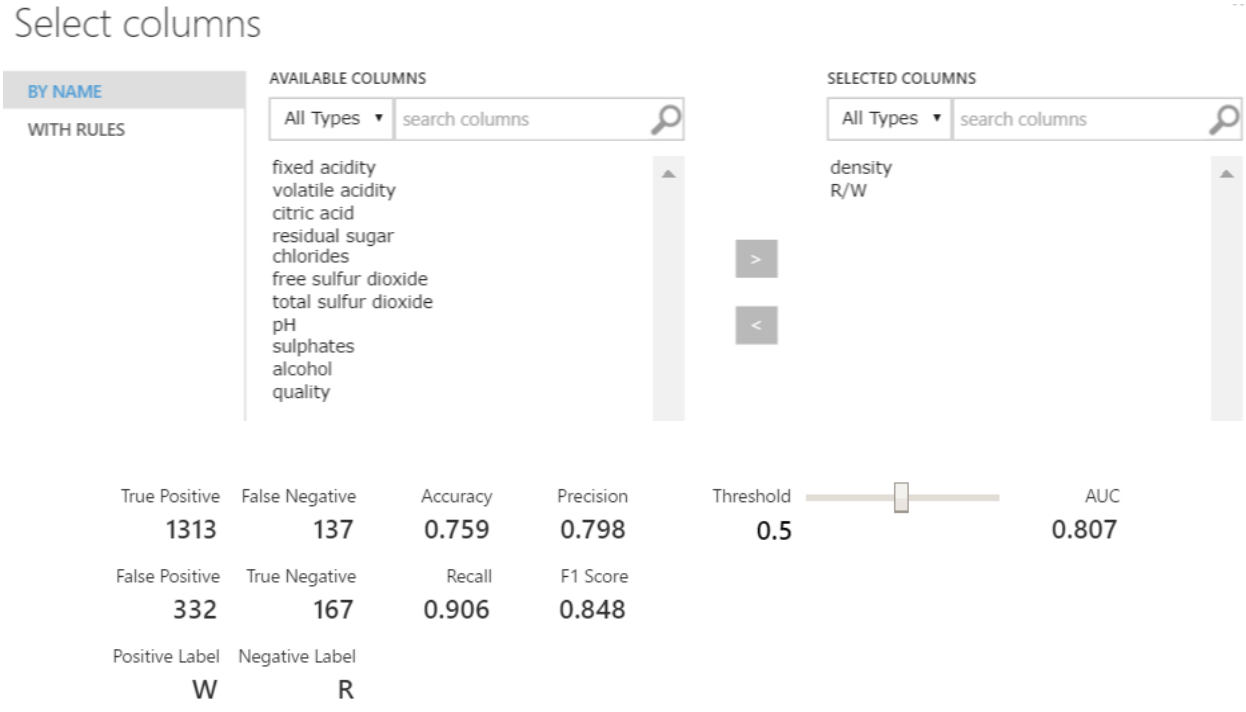
- **Citric Acid - Sulphates**

- **Residual Sugar - pH**

- **Chlorides - Alcohol**


Select columns
BY NAME
WITH RULES
AVAILABLE COLUMNS
All Types
search columns
fixed acidity
volatile acidity
citric acid
residual sugar
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
quality
SELECTED COLUMNS
All Types
search columns
chlorides
alcohol
R/W
True Positive 1392
False Negative 58
Accuracy 0.936
Precision 0.955
Threshold 0.5
AUC 0.975
False Positive 66
True Negative 433
Recall 0.960
F1 Score 0.957
Positive Label W
Negative Label R

- **Density**


Select columns
BY NAME
WITH RULES
AVAILABLE COLUMNS
All Types
search columns
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
pH
sulphates
alcohol
quality
SELECTED COLUMNS
All Types
search columns
density
R/W
True Positive 1313
False Negative 137
Accuracy 0.759
Precision 0.798
Threshold 0.5
AUC 0.807
False Positive 332
True Negative 167
Recall 0.906
F1 Score 0.848
Positive Label W
Negative Label R

Results can be summarized as follows:

Table 1 : : Wine Quality Model Experiment Results

Attributes	Accuracy [%]	AUC
All Attributes	99.5	1
Fixed Acidity - Free Sulphur Dioxide	84.4	0.893
Volatile Acidity - Total Sulphur Dioxide	95.2	0.982
Citric Acid - Sulphates	85.8	0.904
Residual Sugar - Ph	84.2	0.899
Chlorides - Alcohol	93.6	0.975
Density	75.9	0.807

- What is the percentage of correct classification results (using all attributes)?
  Answer: The accuracy is 99.5 % when the default 0.5 threshold is selected where the AUC is 1
- What is the percentage of correct classification results (using a subset of the attributes)?
  Answer: Tabulated above (Table 1)
- What is the AUC of your model?
  Answer: When the threshold is kept as default 0.5 and provided high accuracy 99.5 %. The threshold is tuned and increased to 0.9 and the accuracy was obtained as 99.2 %
- What is your best AUC that you can achieve?
  Answer: The highest AUC was obtained as 1 when all the attributes were selected
- Which are the minimum number of attributes? Why?
  Answer: When only the Volatile Acidity - Total Sulphur Dioxide attributes was selected; a highest accuracy (95.2%) and AUC values (0.987) were obtained compared to the other subsets of attributes cases (Table 1).