Exercise 3: Training a Classification Model

The model you have built to predict daily sales is an example of a regression model. Classification is another kind of supervised learning in which instead of predicting a numeric value, the model is trained to predict the category or *class* of an observation. In this exercise, you will copy an existing training experiment from the Azure AI Gallery and run it to train a classification model that predicts whether or not a profit can be made on any given day.

Copy an Experiment from the Gallery to your Workspace

1. Navigate to https://gallery.cortanaintelligence.com/Experiment/Lemonade-Classification. This opens the **Lemonade Classification** example in the Azure AI Gallery

2. Click **Open in Studio** to copy the experiment and its dataset to your Azure Machine Learning Studio workspace. If you are prompted, sign in using your Microsoft account, and then select the region and workspace for your Azure Machine Learning Studio subscription.

3. When the experiment opens in Azure Machine Learning Studio, rearrange the panes and zoom level so you can see the modules it contains. The experiment data flow starts with a dataset named **Profitability.csv**.

4. Visualize the output of the **Profitability.csv** dataset, and note that it contains historic daily lemonade sales data and a column named **Profitable** in which a 1 indicates that a profit was made on that day, and a 0 indicates that the day was unprofitable

* 1. 5. Review the rest of the experiment, noting that it contains modules to perform the following tasks:
  2. • Create a new feature containing the normal log of **Rainfall**.
  3. • Scale the numeric features using Z-Score or MinMax normalization depending on the distribution of the numeric column data.
  4. • Mark **Day** as a categorical field.
  5. • Clear the **Date** and **Rainfall** features.
  6. • Split the dataset into two subsets for training (70%) and testing (30%).
  7. • Use the two-class logistic regression algorithm to train a classification model that predicts **Profitable** (in spite of being called “logistic *regression*”, this algorithm is used to predict classes rather than numeric values).
  8. • Score the trained model using the test data.
  9. • Evaluate the model based on the test results.

Run the Experiment and View the Results

1. Run the **Lemonade Classification** experiment and wait for it to complete.

2. When the experiment has finished running, view the output of the **Score Model** module, and note that it contains new fields named **Scored Labels** and **Scored Probabilities.** Compare some of the values in the **Scored Labels** field to the **Profitable** field. In most cases, the predicted value in the **Scored Labels** field should be the same as the **Profitable** field.

Compare the **Scored Labels** field to the **Scored Probabilities** field. The scored probability is the numeric value between 0 and 1 calculated by classification algorithm. When this value is closer to 0 than 1, the **Scored Labels** field is 0; and when its closer to 1 than to 0, the **Scored Labels** field is 1.

3. Visualize the output of the **Evaluate Model** module to open the **Evaluation results** window, and view the Received Operator Characteristic (ROC) chart. The larger the area under the curve in this chart, the better the model is performing. In this case, the line goes almost all the way up the left side before going across the top, resulting in an area under the curve that includes almost all of the chart.

4. In the **Evaluation results** window, scroll down to view the evaluation metrics, which includes the *confusion matrix* formed by true positive, false negative, false positive, and true negative predictions; the accuracy, recall, precision, and F1 score; the threshold, and the area under the curve (AUC). These results indicate that, based on the test data, the trained model does a good job of predicting whether or not a particular day will be profitable.

*Challenge*

Adjust the threshold by dragging the slider, and observe the effect on the model metrics.