Lab - Building Rapid Prototypes

Summary

In this lab you will:

- Run a Watson Studio AutoAI Experiment
- Generate a Python notebook from the AutoAI

Refer to the demo videos from this lesson for a step-by-step demonstration of how to complete the lab.

Instructions

1. Create a new project in Watson Studio

    a. Select Create an empty project
    b. Name your project AutoAI_exp
    c. Select your Cloud Object Storage from the Select storage dropdown
    d. Click Create to finish creating the empty project

2. Add a data source to project

    a. Click the 0001 icon on the menu bar at the top right
    b. Click on browse to add a local .csv file
    c. Upload the provided file banknote_auth.csv

3. Explore the data

    a. You should see banknote_auth.csv listed as a Data asset on the project page. Select it to open a preview and profile of the data.
    b. Review the description of the data source and the individual attributes here: http://archive.ics.uci.edu/ml/datasets/banknote+authentication
    c. Optional: In order to complete this step, add the free Watson Knowledge Catalog as a service to your cloud account. Then, click Profile (next to Preview) in the menu bar of Watson Studio and select Create profile
    d. Optional: A data profile will be generated. You may need to refresh the page in your browser to show the profile. Take a moment to explore the provided statistics. How might these statistics inform your choice of a prediction algorithm for this problem?

4. Add an AutoAI Experiment

    a. Return to the project menu by clicking the name of the project in the top left menu bar
    b. Click Add to Project +
    c. Select AutoAI Experiment
    d. Give the experiment a name
    e. Under Associate Services, select Associate a Machine Learning service instance
    f. Click on WatsonMachineLearning from the drop down menu in the new tab and Select
    g. In the original tab, select the Reload button
    h. Click Create

5. Add a data source to the AutoAI Experiment

    a. Click Select from project
    b. Select banknote_auth.csv and click Select asset
    c. Once the columns have been parsed, click on the column class to select it as a prediction column
    d. Click Run experiment
    e. Watch the animation and the changing Pipeline leaderboard as the experiment runs
    f. When the experiment finishes running (should take slightly over 2min), notice the two algorithms that have been selected as top performing. Are these algorithms that you would have tried on your own?

6. Save the top-ranked pipeline as a Notebook

    a. Rollover the mouse on the top-ranked pipeline to show the Save as button. Click Notebook to save that pipeline as a Notebook.
    b. Click Create on the New notebook screen.
        i. You may need to enable popups in your web browser in order for the notebook to open.
    c. In the rest of this course, we'll work our way through the notebook examining how each step of the Data Science workflow is represented in the notebook.
        i. Review the notebook and run each cell.
        ii. In particular, we'll focus on the Compose Pipeline cell.
        iii. If you're not familiar with `scikit-learn`'s `Pipeline` class, review the documentation here: https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html
        iv. Documentation for IBM's autoai_libs can be found here: https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/autoai-lib-python.html
    d. Under 8. Fit pipeline, predict on Holdout set, calculate score, perform cross-validation, notice the output indicating the pipeline's accuracy on the holdout set and for cross-validation. These match the results in the Pipeline leaderboard of the UI. Note that there is some rounding to the values so 0.999… appears as 1 in the UI.

7. Optional: Run an experiment for a regression problem

    a. Return to the project page and add a new data source parkinsons_updrs.csv
    b. Read about the data source here: http://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring. Note that the target column in this case is the total_UPDRS column
    c. Repeat the above steps to run a new experiment for this data set
        i. Before running the experiment, select Experiment Settings. Under Data Source settings and Select columns to include, deselect the following columns:
            1. subject# - id number for subject
            2. motor_updrs – subtotal for total_updrs target label
        ii. Click Save settings and Run experiment