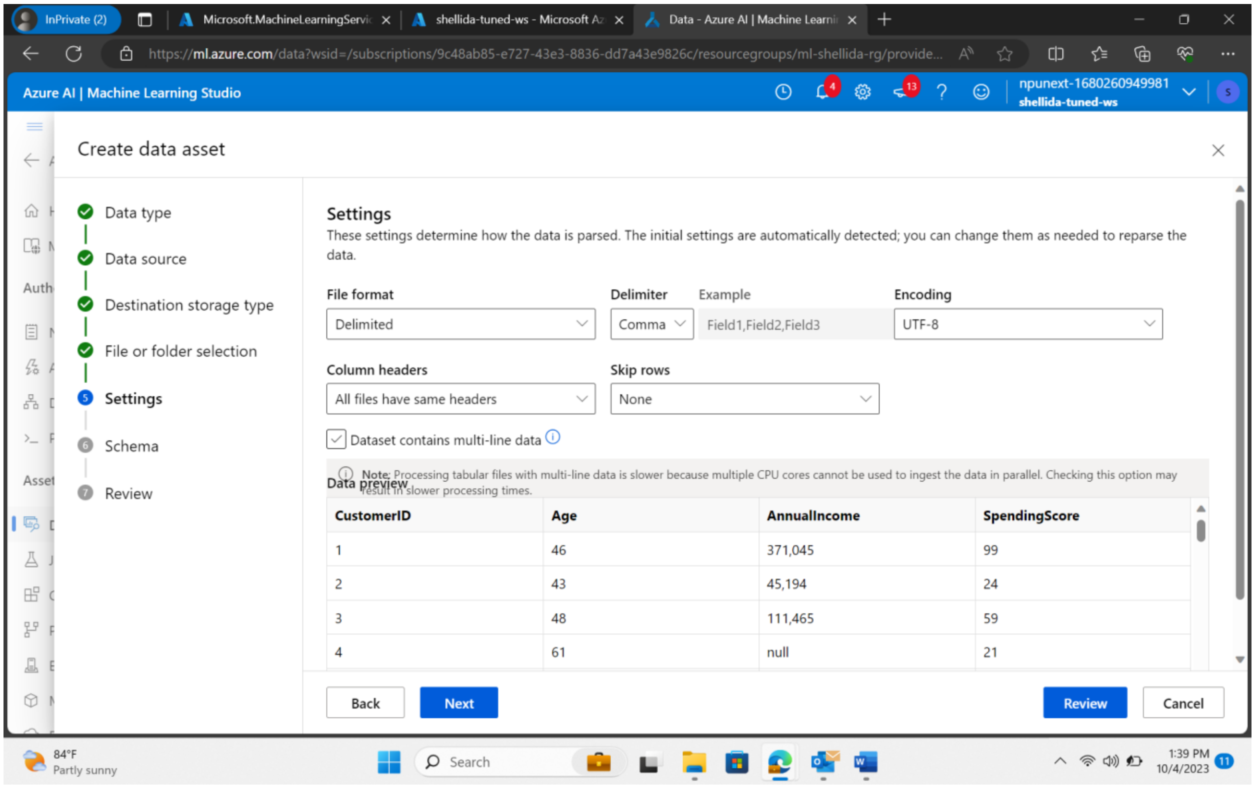
Azure ML Studio - Hands-On Assessment

-Musaddiq Shariff

Emp ID: 655023

1. Create Data set



InPrivate (2)
Microsoft.MachineLearningServic
shellida-tuned-ws - Microsoft Az
Data - Azure AI | Machine Learni
https://ml.azure.com/data?wsid=/subscriptions/9c48ab85-e727-43e3-8836-dd7a43e9826c/resourcegroups/ml-shellida-rg/provide...
Azure AI | Machine Learning Studio
npunext-1680260949981
shellida-tuned-ws

Create data asset

- Data type
- Data source
- Destination storage type
- File or folder selection
- 5 Settings
- 6 Schema
- 7 Review

Settings

These settings determine how the data is parsed. The initial settings are automatically detected; you can change them as needed to reparse the data.

File format: Delimited

Delimiter: Comma

Example: Field1,Field2,Field3

Encoding: UTF-8

Column headers: All files have same headers

Skip rows: None

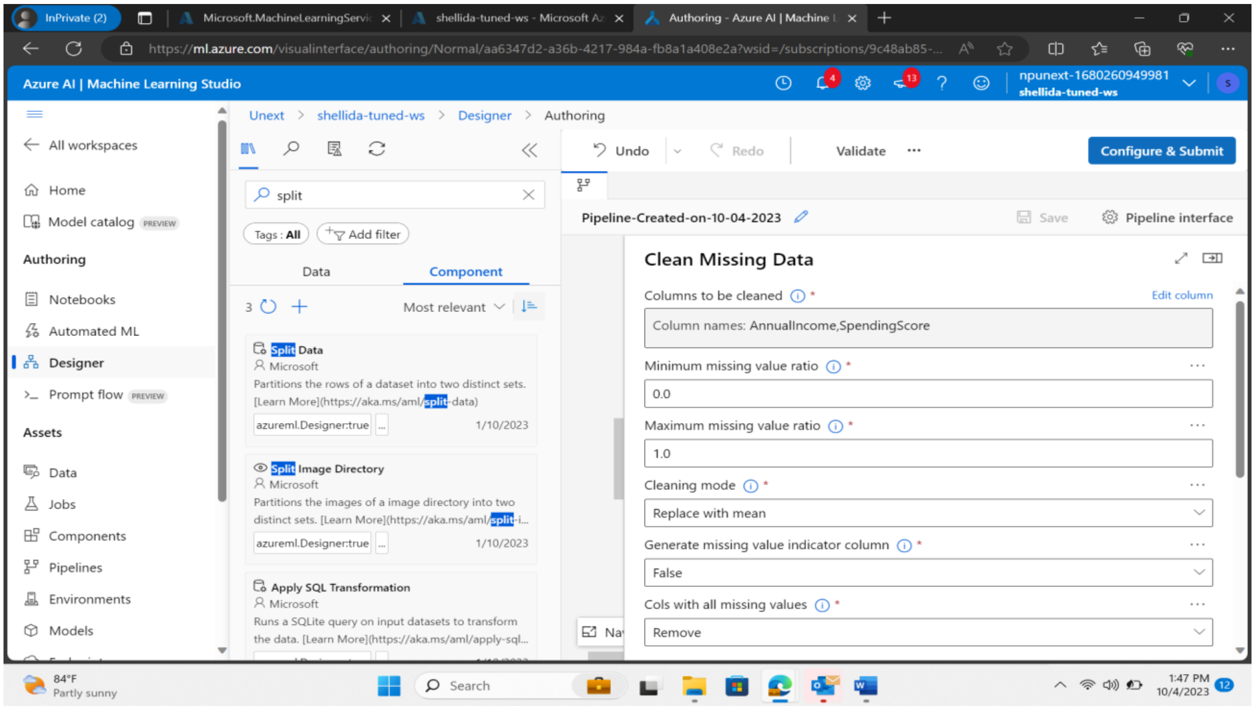
☑ Dataset contains multi-line data

Note: Processing tabular files with multi-line data is slower because multiple CPU cores cannot be used to ingest the data in parallel. Checking this option may result in slower processing times.

Data preview

CustomerID	Age	AnnualIncome	SpendingScore
1	46	371,045	99
2	43	45,194	24
3	48	111,465	59
4	61	null	21

2. Clean Missing Data



Split Data
Microsoft
Partitions the rows of a dataset into two distinct sets. [Learn More](https://aka.ms/aml/split-data)
azureml.Designer:true 1/10/2023

Split Image Directory
Microsoft
Partitions the images of a image directory into two distinct sets. [Learn More](https://aka.ms/aml/split-i...
azureml.Designer:true 1/10/2023

Apply SQL Transformation
Microsoft
Runs a SQLite query on input datasets to transform the data. [Learn More](https://aka.ms/aml/apply-sql...

Pipeline-Created-on-10-04-2023

Clean Missing Data

Columns to be cleaned *
Edit column
Column names: AnnualIncome,SpendingScore

Minimum missing value ratio *
0.0

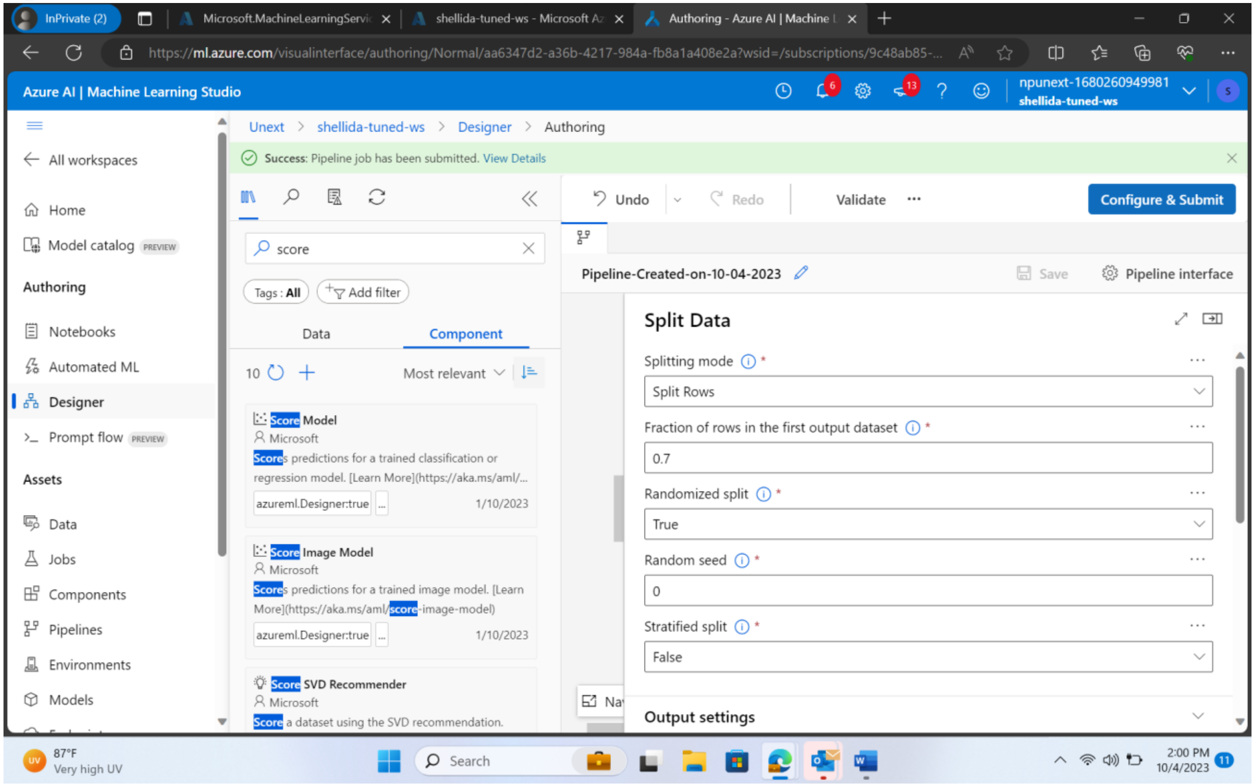
Maximum missing value ratio *
1.0

Cleaning mode *
Replace with mean

Generate missing value indicator column *
False

Cols with all missing values *
Remove

3. Split the dataset



Success: Pipeline job has been submitted. View Details

Score Model
Microsoft
Scores predictions for a trained classification or regression model. [Learn More](https://aka.ms/aml/...
azureml.Designer:true 1/10/2023

Score Image Model
Microsoft
Scores predictions for a trained image model. [Learn More](https://aka.ms/aml/score-image-model)
azureml.Designer:true 1/10/2023

Score SVD Recommender
Microsoft
Score a dataset using the SVD recommendation.

Pipeline-Created-on-10-04-2023

Split Data

Splitting mode *
Split Rows

Fraction of rows in the first output dataset *
0.7

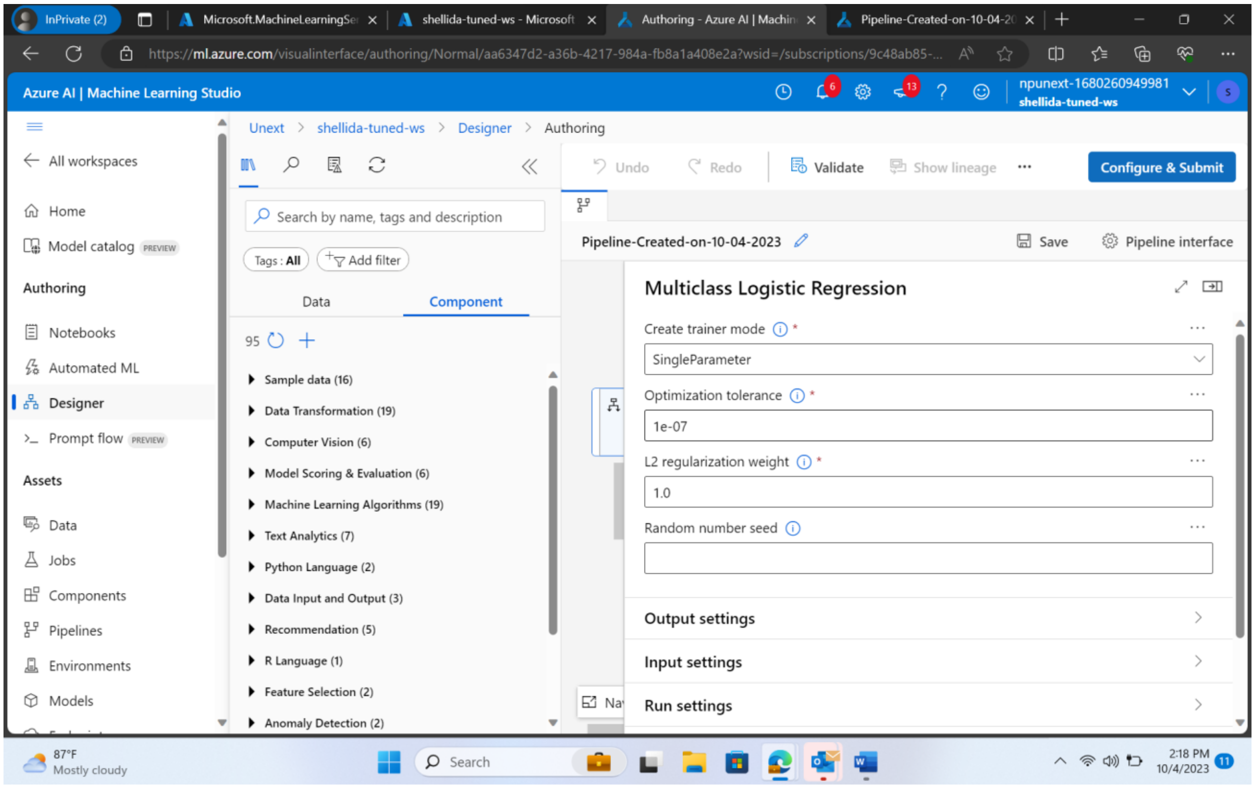
Randomized split *
True

Random seed *
0

Stratified split *
False

Output settings

4. Selecting appropriate model



Multiclass Logistic Regression

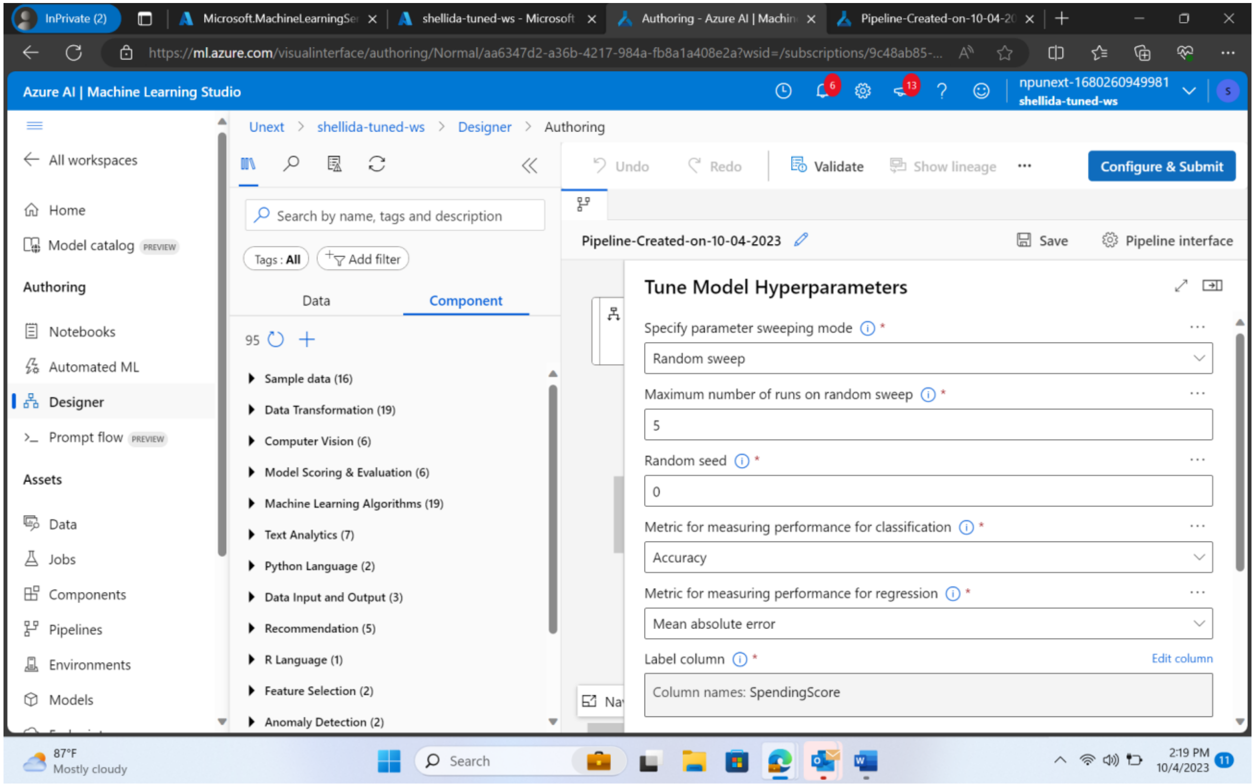
Create trainer mode *: SingleParameter

Optimization tolerance *: 1e-07

L2 regularization weight *: 1.0

Random number seed:

5. Tune model Hyperparameters



Tune Model Hyperparameters

Specify parameter sweeping mode *: Random sweep

Maximum number of runs on random sweep *: 5

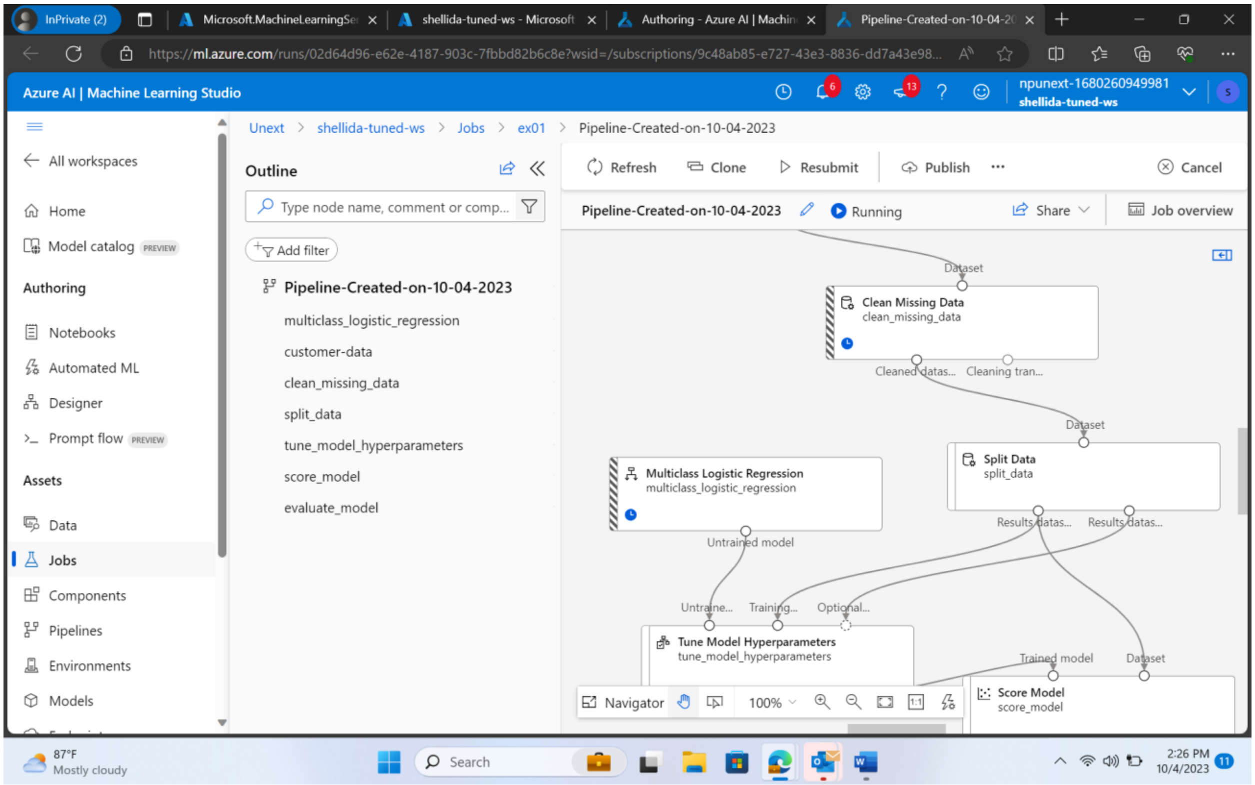
Random seed *: 0

Metric for measuring performance for classification *: Accuracy

Metric for measuring performance for regression *: Mean absolute error

Label column *: Column names: SpendingScore

6. Model Running



Assessment Questions:

1. What are the key steps involved in preparing the dataset for training a machine learning model using Azure Machine Learning? Briefly explain each step.

Ans: Steps involved in preparing dataset are:

1) Data collection: In this we use several types of sources like Gather relevant data from various sources, which may include databases, files, or external APIs. Azure Component: Utilize Azure Data Factory, Azure Databricks, or other services to ingest and collect data. We may also use Azure Data Factory, Azure Databricks, or other services to ingest and collect data.
2) Data Exploration: Analyze the dataset to understand its structure, identify missing values, outliers, and explore the distribution of features. For this we can use Use Azure Notebooks, Azure Databricks, or Jupyter notebooks in Azure Machine Learning to perform exploratory data analysis (EDA).
3) Data Cleaning: Handle missing values, outliers, and correct any inconsistencies in the dataset.t
4) Data Splitting: Divide the dataset into training and testing sets to assess the model's performance on unseen data.

2. Why is it important to split the dataset into training and testing sets when developing a machine learning model? How does this help in model evaluation?

Ans. It is necessary to split the dataset into training and testing sets because the model has to be trained with a new dataset and we refer here it as training dataset and again for the validation purposed we use the testing dataset to check the accuracy of the model because the data available in the testing dataset will be unique for the model means its new for the model. The best results are obtained by using this technique.

3. Describe a machine learning algorithm suitable for predicting customer purchasing behaviour in the given scenario. Explain why you chose this algorithm.

Ans. I chose Random Forest Algorithm

The reasons are:

1) Random Forest is an ensemble learning method, meaning it combines the predictions of multiple individual models (decision trees) to improve overall performance
2) Non-Linearity and Complex Relationships
3) Handling Categorical Variables:
4) Robust to Overfitting:
5) Handling Imbalanced Datasets

4. What is hyperparameter tuning, and why is it important in machine learning? Explain a technique used for hyperparameter tuning and its benefits.

Ans. In machine learning, a model's hyperparameters are configuration settings that are not learned from the data but are set prior to training. Hyperparameter tuning, also known as hyperparameter optimization, is the process of selecting the best set of hyperparameters for a machine learning model to improve its performance. The goal is to find the hyperparameters that result in the best generalization performance on unseen data.

Hyperparameters are configuration settings that are predefined before training and can significantly impact a model's ability to generalize to new data. Grid Search is a commonly employed technique for hyperparameter tuning, where a predefined grid of hyperparameter values is systematically searched. For each combination in the grid, the model is trained and evaluated, and the set of hyperparameters yielding the best performance is chosen. Grid Search offers a transparent and intuitive approach, ensuring a thorough exploration of the hyperparameter space. It is beneficial for improving model accuracy, preventing overfitting, and enhancing overall robustness, contributing to the reliability and effectiveness of machine learning models.