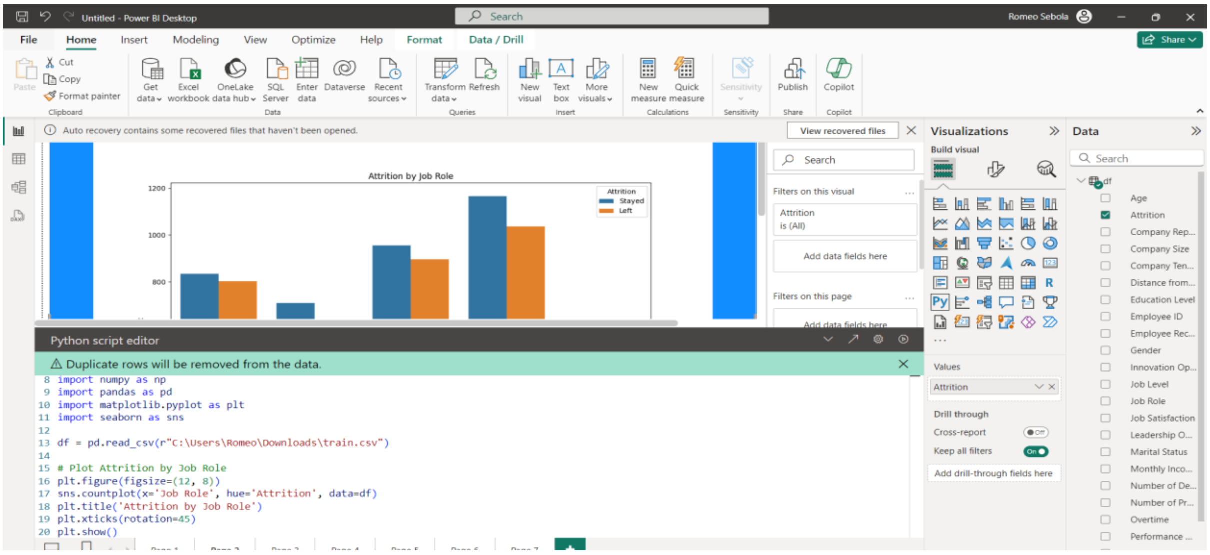
Predicting Employee Attrition Using Machine Learning Models

Introduction

In this report, I will use a machine learning model to predict employee attrition with the Synthetic Employee Attrition Dataset. This dataset has 74,498 samples, split into training and testing sets for model building and testing. Each sample includes a unique Employee ID and various features like demographics, job details, and personal circumstances that might affect attrition.

The main goal is to understand why employees leave and create models to identify employees at risk of leaving. This dataset is great for HR analytics, building machine learning models, and showing advanced data analysis methods. It gives a detailed and realistic view of what affects employee retention, making it useful for HR researchers and professionals.

EDA(PLOTS) IN POWERBI



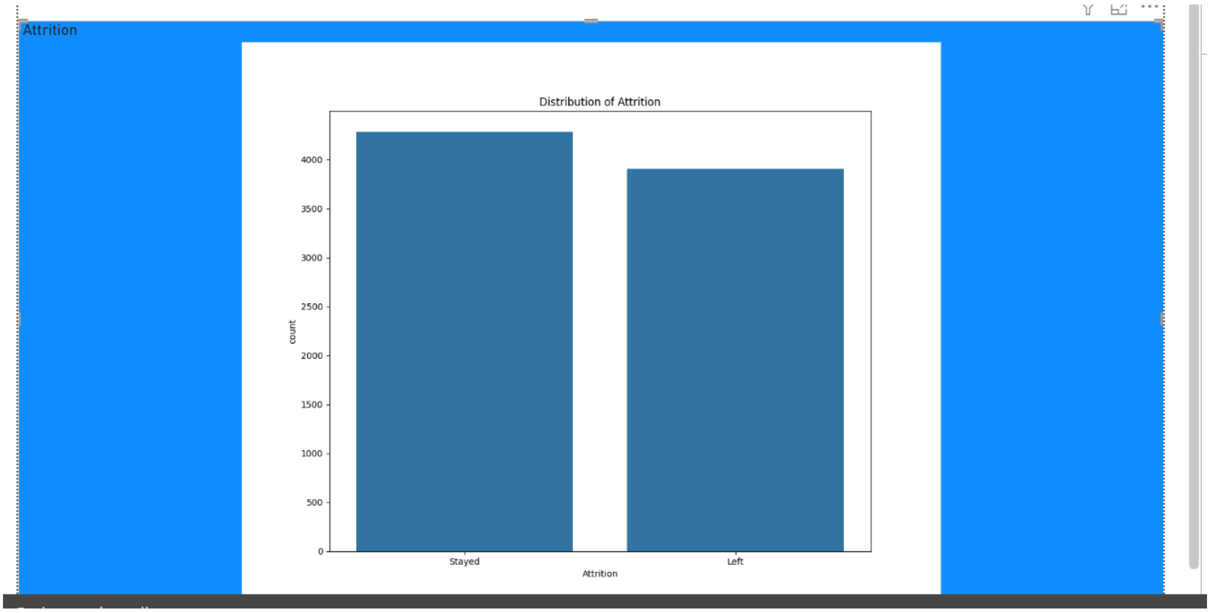
Auto recovery contains some recovered files that haven't been opened.

⚠ Duplicate rows will be removed from the data.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv(r"C:\Users\Romeo\Downloads\train.csv")

# Plot Attrition by Job Role
plt.figure(figsize=(12, 8))
sns.countplot(x='Job Role', hue='Attrition', data=df)
plt.title('Attrition by Job Role')
plt.xticks(rotation=45)
plt.show()
```

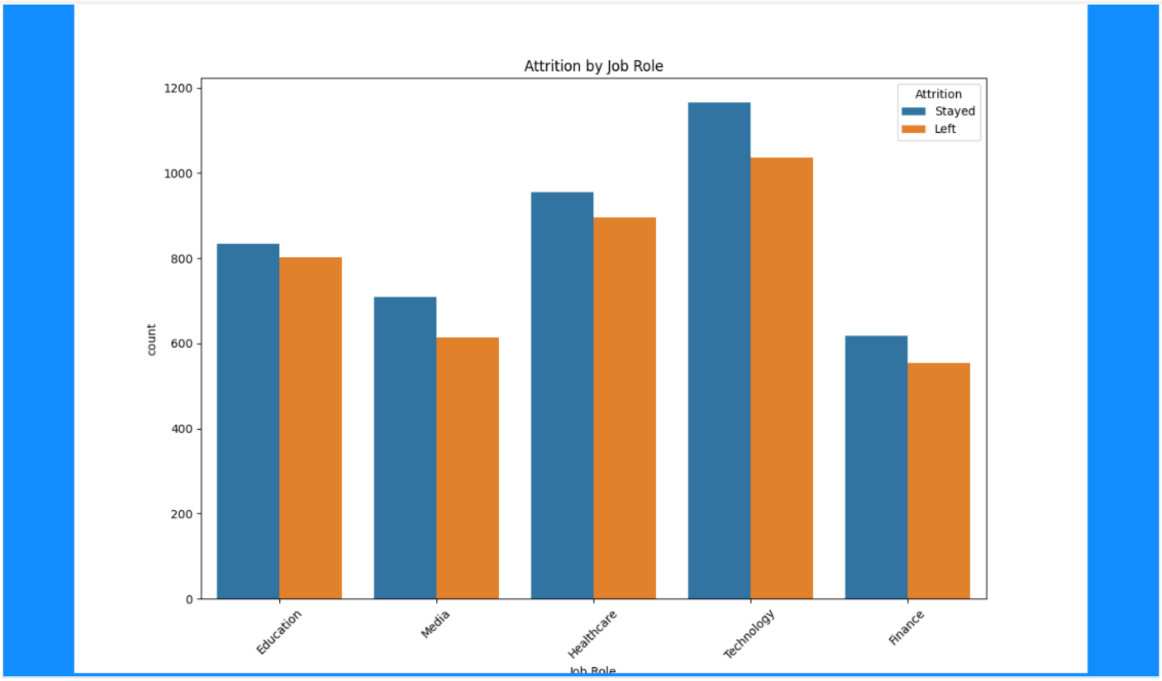
Distribution of attrition

The following graph illustrates the distribution of attrition within the dataset



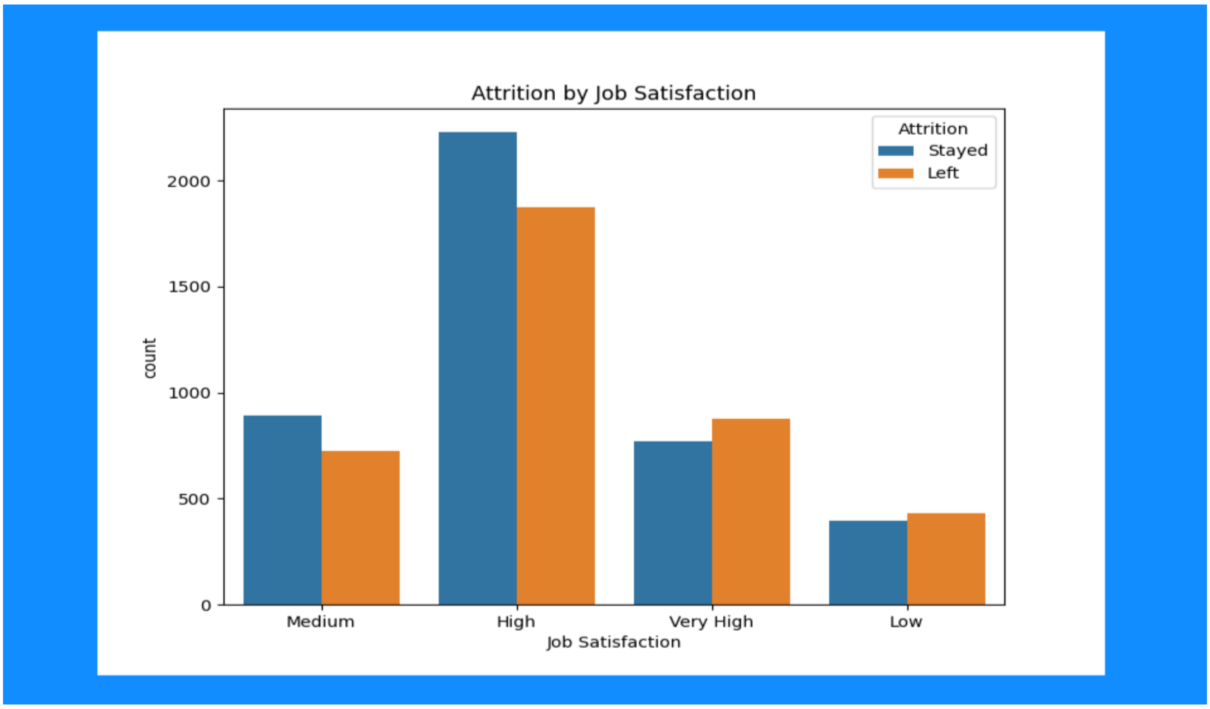
Attrition by Job Role

The following analysis explores the distribution of attrition across different job roles within the organization.



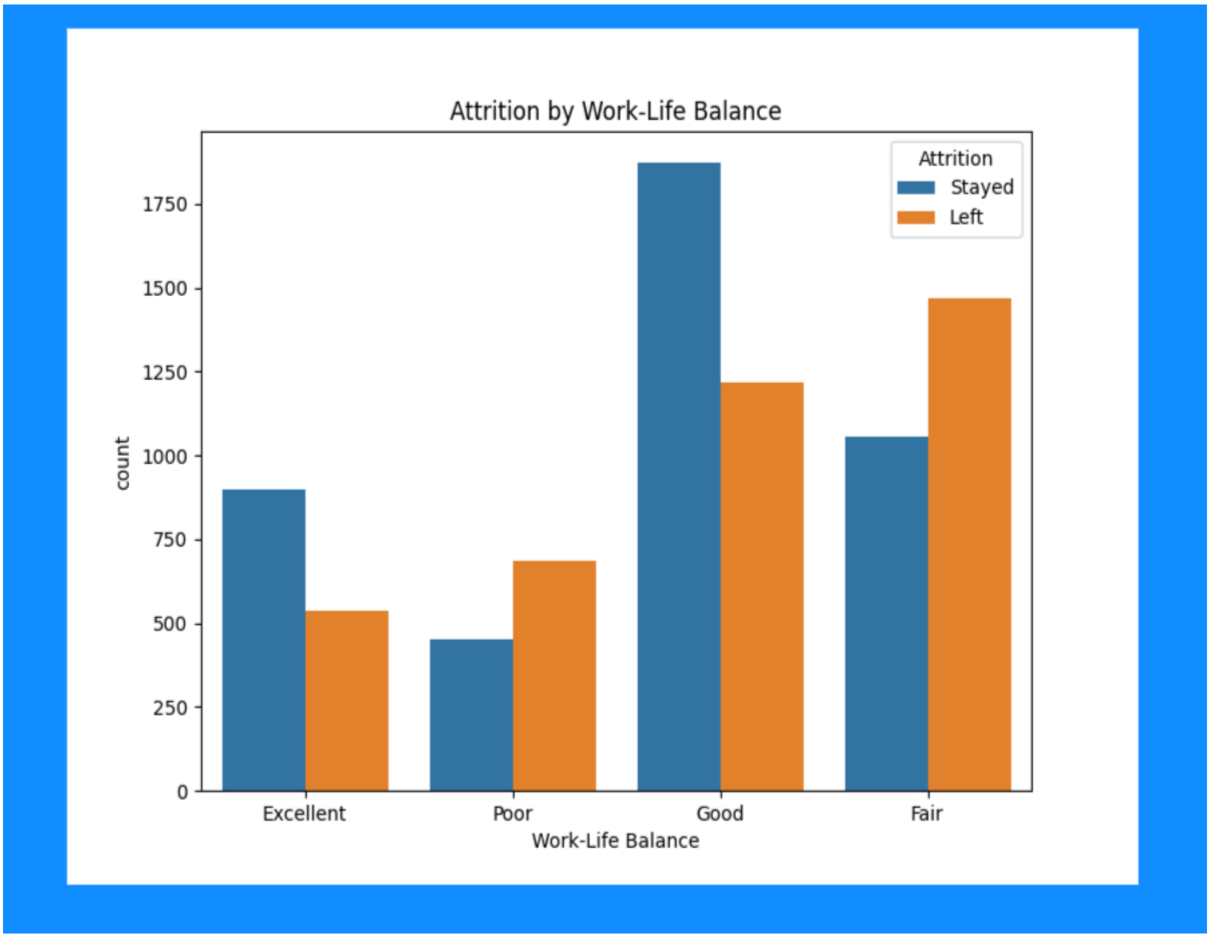
Attrition by job Satisfaction

The following analysis examines the relationship between job satisfaction levels and employee attrition within the organization.



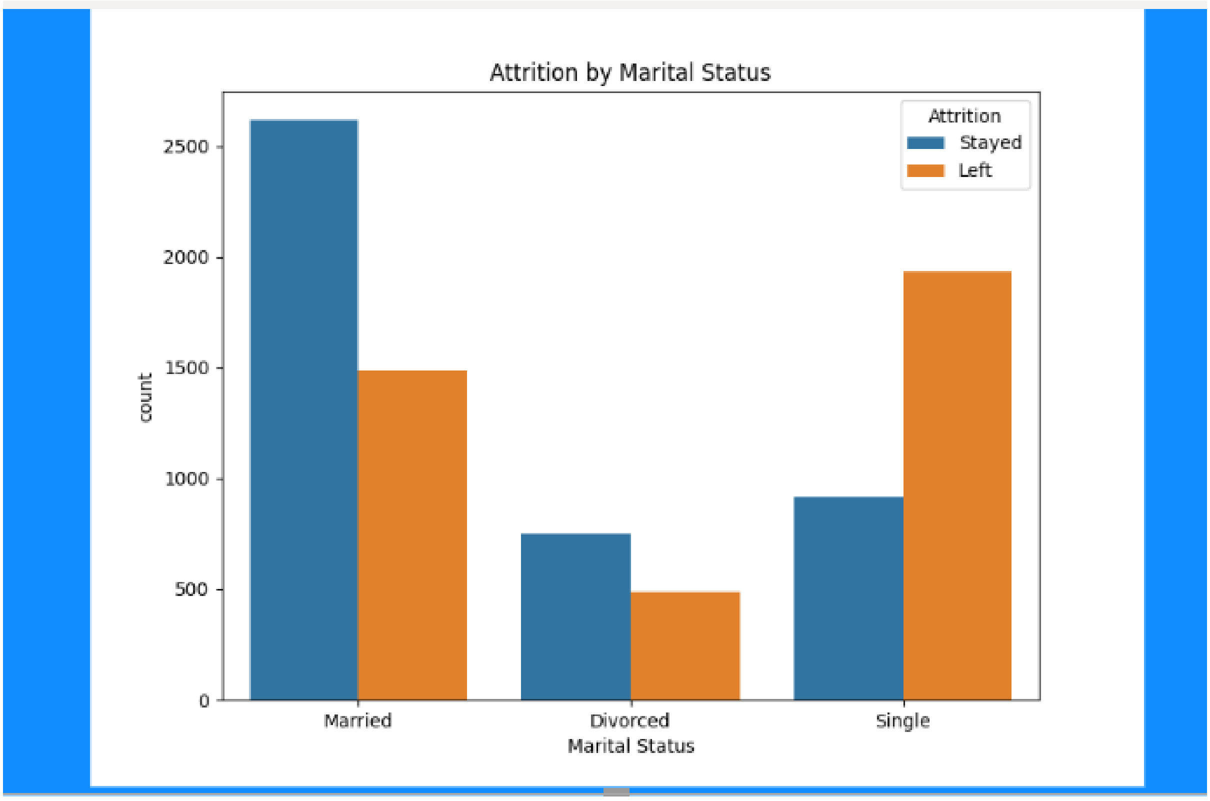
Attrition by Work-Life Balance

The following analysis investigates the relationship between work-life balance and employee attrition within the organization.



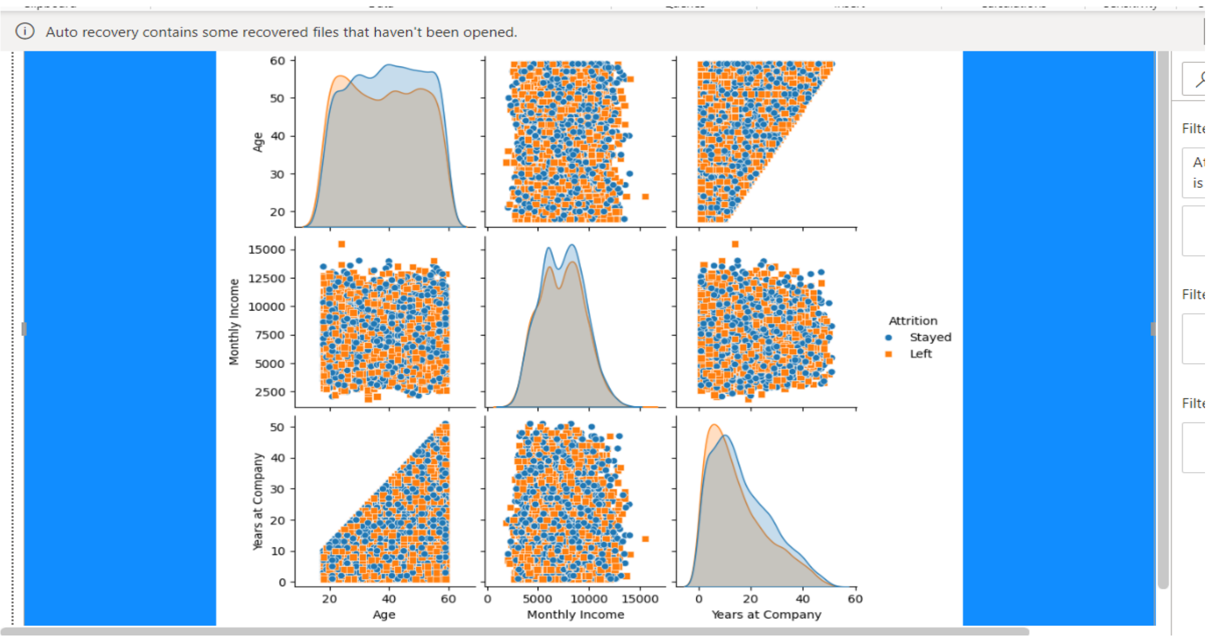
Attrition by Marital Status

The following analysis examines the relationship between marital status and employee attrition within the organization.



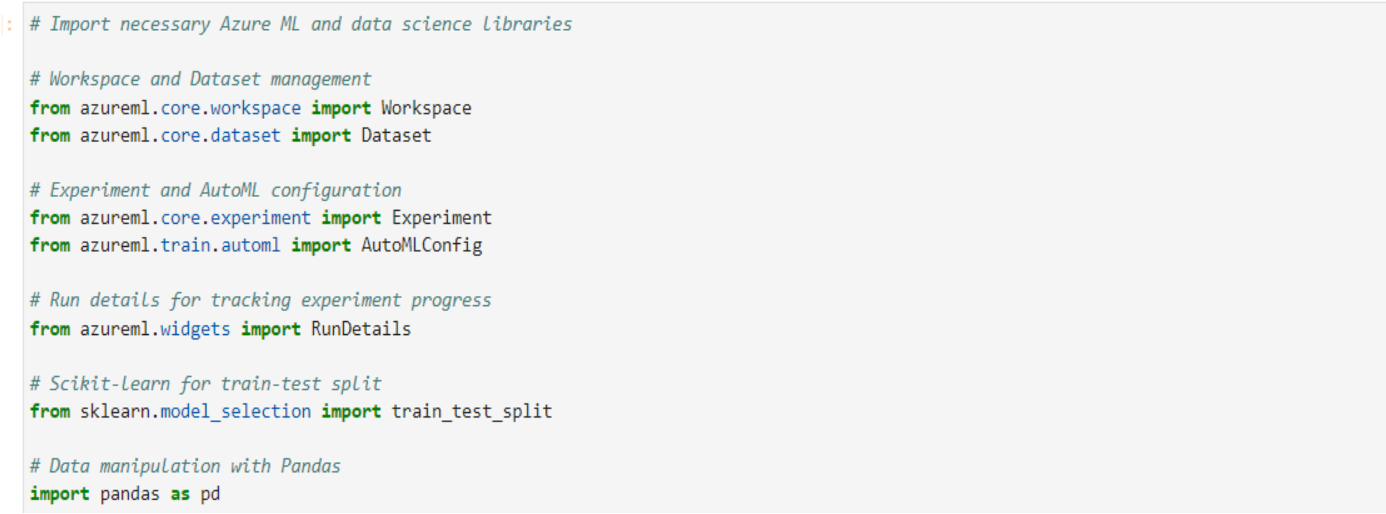
Pair plot for selected features

he following pair plot visualizes the relationships between selected features in the dataset.

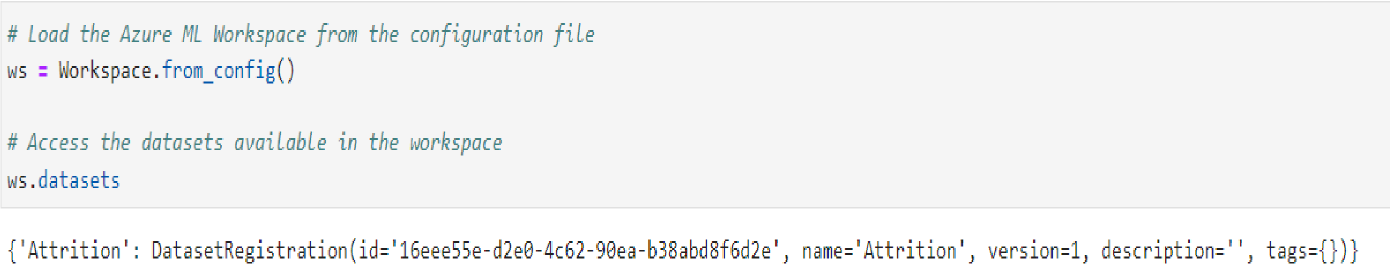


Auto recovery contains some recovered files that haven't been opened.

DATA PREP

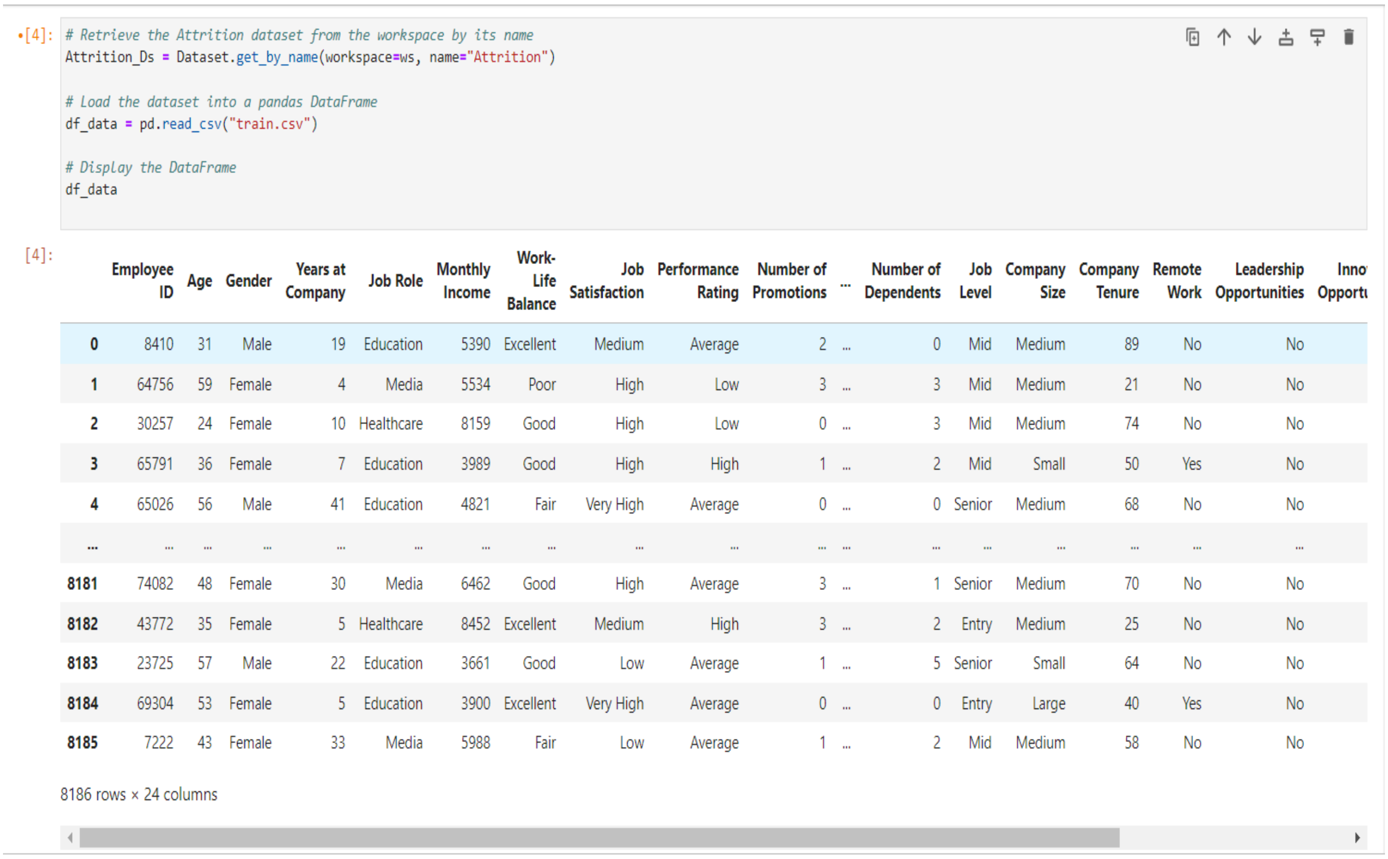


```
:  # Import necessary Azure ML and data science libraries

   # Workspace and Dataset management
   from azureml.core.workspace import Workspace
   from azureml.core.dataset import Dataset

   # Experiment and AutoML configuration
   from azureml.core.experiment import Experiment
   from azureml.train.automl import AutoMLConfig

   # Run details for tracking experiment progress
   from azureml.widgets import RunDetails

   # Scikit-learn for train-test split
   from sklearn.model_selection import train_test_split

   # Data manipulation with Pandas
   import pandas as pd
```



```
# Load the Azure ML Workspace from the configuration file
ws = Workspace.from_config()

# Access the datasets available in the workspace
ws.datasets
```

```
{'Attrition': DatasetRegistration(id='16eee55e-d2e0-4c62-90ea-b38abd8f6d2e', name='Attrition', version=1, description='', tags={})}
```

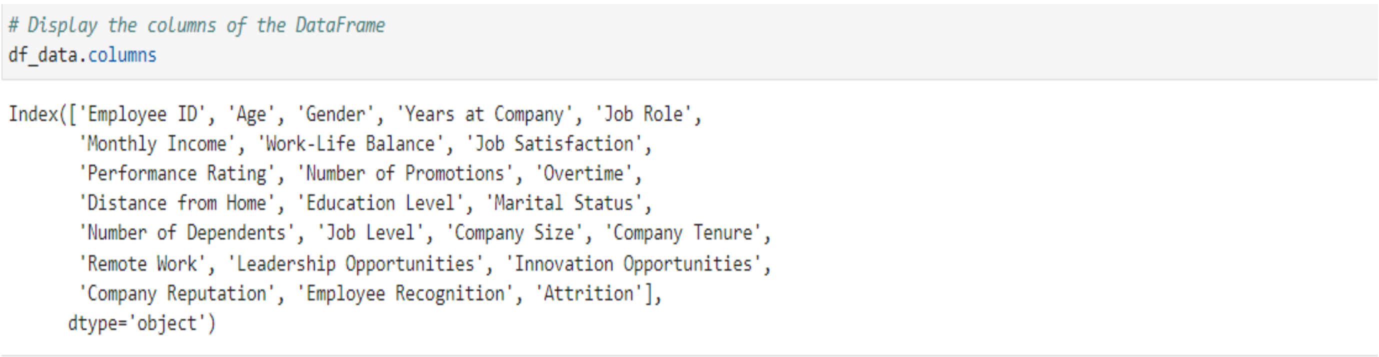


```
•[4]: # Retrieve the Attrition dataset from the workspace by its name
      Attrition_Ds = Dataset.get_by_name(workspace=ws, name="Attrition")

      # Load the dataset into a pandas DataFrame
      df_data = pd.read_csv("train.csv")

      # Display the DataFrame
      df_data
```

[4]:

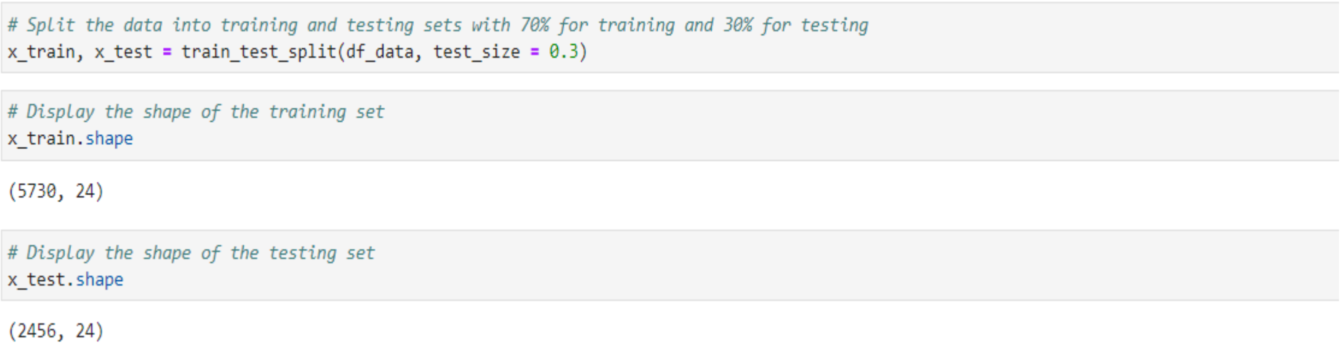
	Employee ID	Age	Gender	Years at Company	Job Role	Monthly Income	Work-Life Balance	Job Satisfaction	Performance Rating	Number of Promotions	...	Number of Dependents	Job Level	Company Size	Company Tenure	Remote Work	Leadership Opportunities	Inno Opportu
0	8410	31	Male	19	Education	5390	Excellent	Medium	Average	2	...	0	Mid	Medium	89	No	No	
1	64756	59	Female	4	Media	5534	Poor	High	Low	3	...	3	Mid	Medium	21	No	No	
2	30257	24	Female	10	Healthcare	8159	Good	High	Low	0	...	3	Mid	Medium	74	No	No	
3	65791	36	Female	7	Education	3989	Good	High	High	1	...	2	Mid	Small	50	Yes	No	
4	65026	56	Male	41	Education	4821	Fair	Very High	Average	0	...	0	Senior	Medium	68	No	No	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
8181	74082	48	Female	30	Media	6462	Good	High	Average	3	...	1	Senior	Medium	70	No	No	
8182	43772	35	Female	5	Healthcare	8452	Excellent	Medium	High	3	...	2	Entry	Medium	25	No	No	
8183	23725	57	Male	22	Education	3661	Good	Low	Average	1	...	5	Senior	Small	64	No	No	
8184	69304	53	Female	5	Education	3900	Excellent	Very High	Average	0	...	0	Entry	Large	40	Yes	No	
8185	7222	43	Female	33	Media	5988	Fair	Low	Average	1	...	2	Mid	Medium	58	No	No	

8186 rows × 24 columns



```
# Display the columns of the DataFrame
df_data.columns
```

```
Index(['Employee ID', 'Age', 'Gender', 'Years at Company', 'Job Role',
       'Monthly Income', 'Work-Life Balance', 'Job Satisfaction',
       'Performance Rating', 'Number of Promotions', 'Overtime',
       'Distance from Home', 'Education Level', 'Marital Status',
       'Number of Dependents', 'Job Level', 'Company Size', 'Company Tenure',
       'Remote Work', 'Leadership Opportunities', 'Innovation Opportunities',
       'Company Reputation', 'Employee Recognition', 'Attrition'],
      dtype='object')
```



```
# Split the data into training and testing sets with 70% for training and 30% for testing
x_train, x_test = train_test_split(df_data, test_size = 0.3)
```

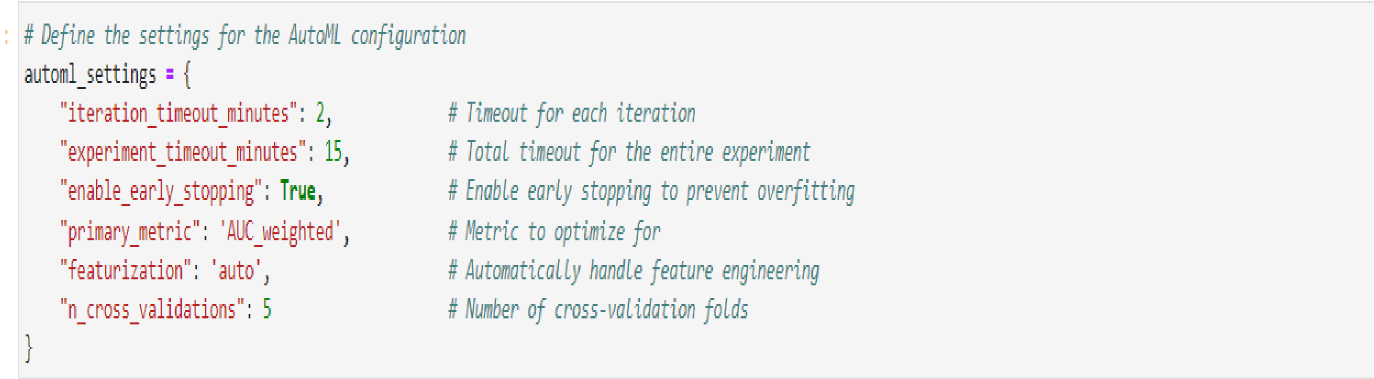
```
# Display the shape of the training set
x_train.shape
```

```
(5730, 24)
```

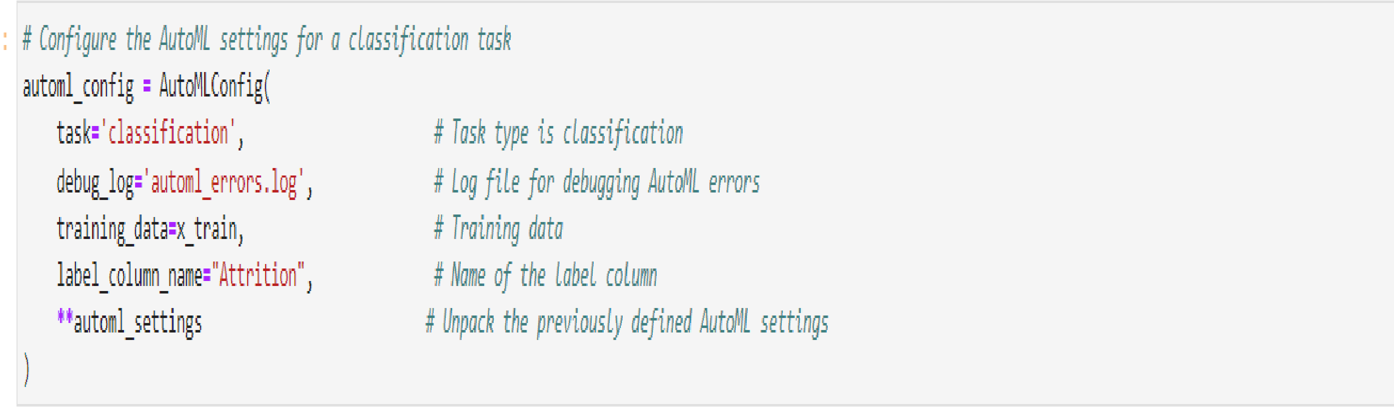
```
# Display the shape of the testing set
x_test.shape
```

```
(2456, 24)
```

2. AutoML Configuration

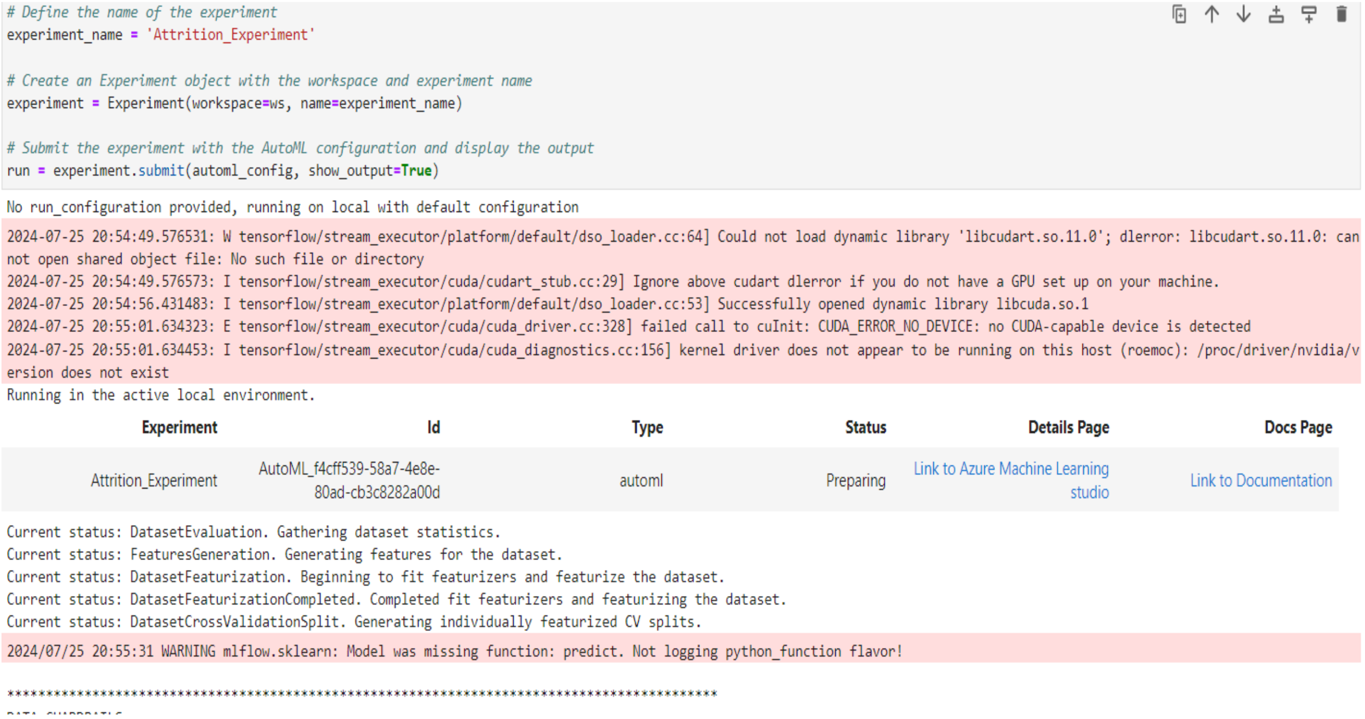


```
# Define the settings for the AutoML configuration
automl_settings = {
    "iteration_timeout_minutes": 2,            # Timeout for each iteration
    "experiment_timeout_minutes": 15,          # Total timeout for the entire experiment
    "enable_early_stopping": True,             # Enable early stopping to prevent overfitting
    "primary_metric": 'AUC_weighted',          # Metric to optimize for
    "featurization": 'auto',                   # Automatically handle feature engineering
    "n_cross_validations": 5                   # Number of cross-validation folds
}
```



```
# Configure the AutoML settings for a classification task
automl_config = AutoMLConfig(
    task='classification',                     # Task type is classification
    debug_log='automl_errors.log',             # Log file for debugging AutoML errors
    training_data=x_train,                     # Training data
    label_column_name="Attrition",             # Name of the label column
    **automl_settings                         # Unpack the previously defined AutoML settings
)
```

3. Experiment Management

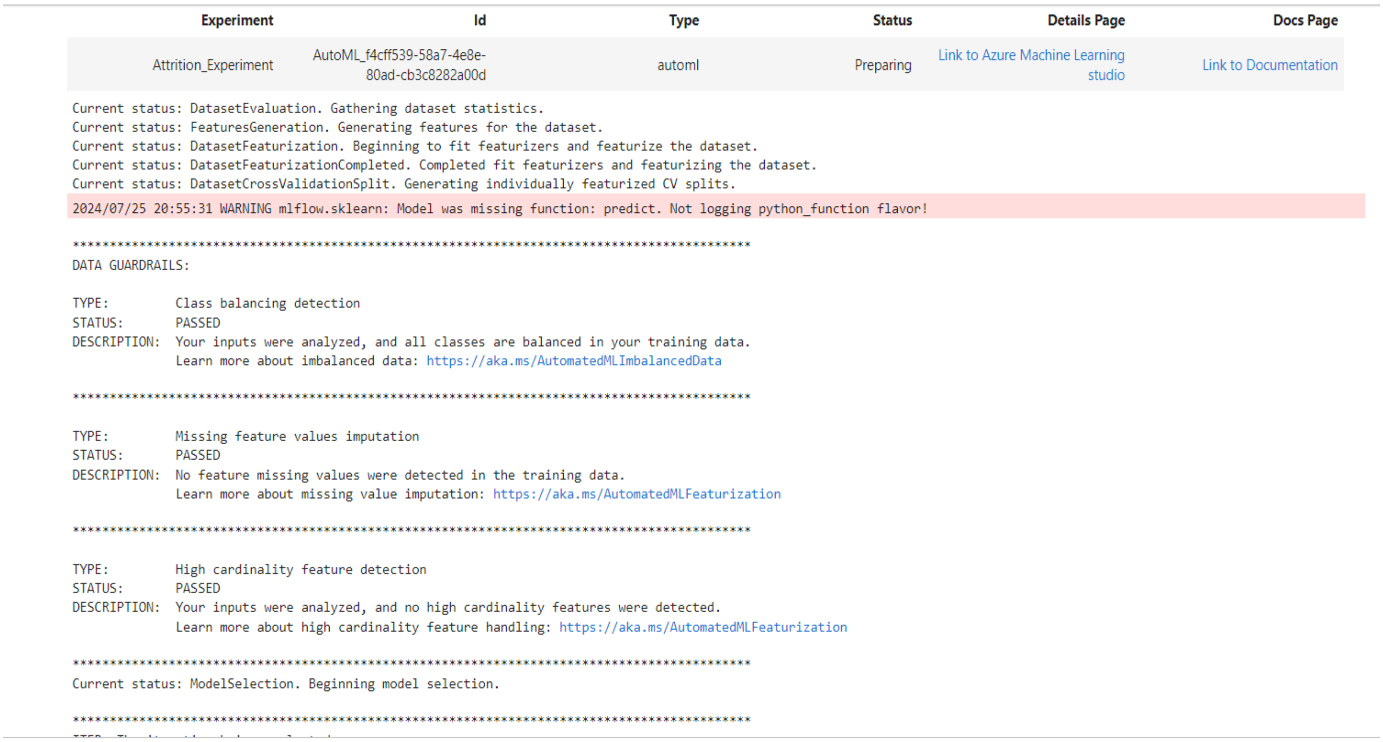


```
# Define the name of the experiment
experiment_name = 'Attrition_Experiment'

# Create an Experiment object with the workspace and experiment name
experiment = Experiment(workspace=ws, name=experiment_name)

# Submit the experiment with the AutoML configuration and display the output
run = experiment.submit(automl_config, show_output=True)
```

```
No run_configuration provided, running on local with default configuration
2024-07-25 20:54:49.576531: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dynamic library 'libcudart.so.11.0'; dlerror: libcudart.so.11.0: can
not open shared object file: No such file or directory
2024-07-25 20:54:49.576573: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] Ignore above cudart dlerror if you do not have a GPU set up on your machine.
2024-07-25 20:54:56.431483: I tensorflow/stream_executor/platform/default/dso_loader.cc:53] Successfully opened dynamic library libcuda.so.1
2024-07-25 20:55:01.634323: E tensorflow/stream_executor/cuda/cuda_driver.cc:328] failed call to cuInit: CUDA_ERROR_NO_DEVICE: no CUDA-capable device is detected
2024-07-25 20:55:01.634453: I tensorflow/stream_executor/cuda/cuda_diagnostics.cc:156] kernel driver does not appear to be running on this host (roemoc): /proc/driver/nvidia/v
ersion does not exist
Running in the active local environment.
```

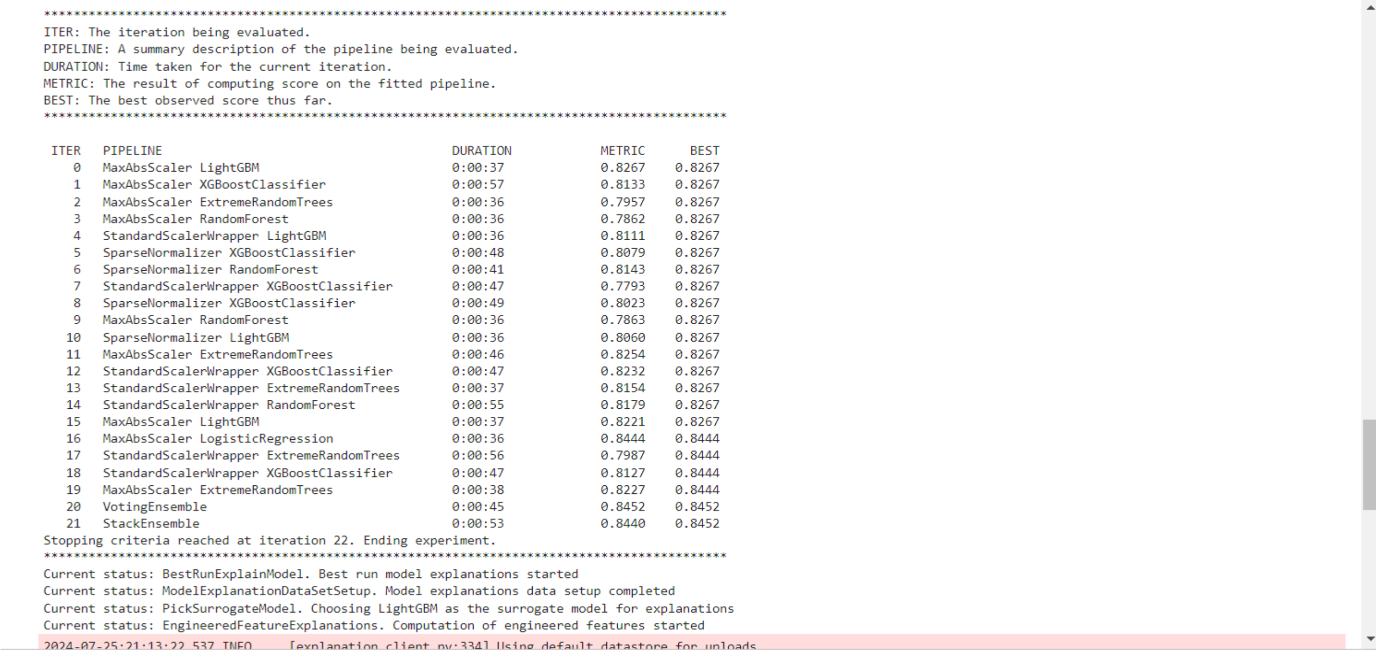
Experiment	Id	Type	Status	Details Page	Docs Page
Attrition_Experiment	AutoML_f4cff539-58a7-4e8e-80ad-cb3c8282a00d	automl	Preparing	Link to Azure Machine Learning studio	Link to Documentation

```
Current status: DatasetEvaluation. Gathering dataset statistics.
Current status: FeaturesGeneration. Generating features for the dataset.
Current status: DatasetFeaturization. Beginning to fit featurizers and featurize the dataset.
Current status: DatasetFeaturizationCompleted. Completed fit featurizers and featurizing the dataset.
Current status: DatasetCrossValidationSplit. Generating individually featurized CV splits.
2024/07/25 20:55:31 WARNING mlflow.sklearn: Model was missing function: predict. Not logging python_function flavor!

********************************************************************************************
```

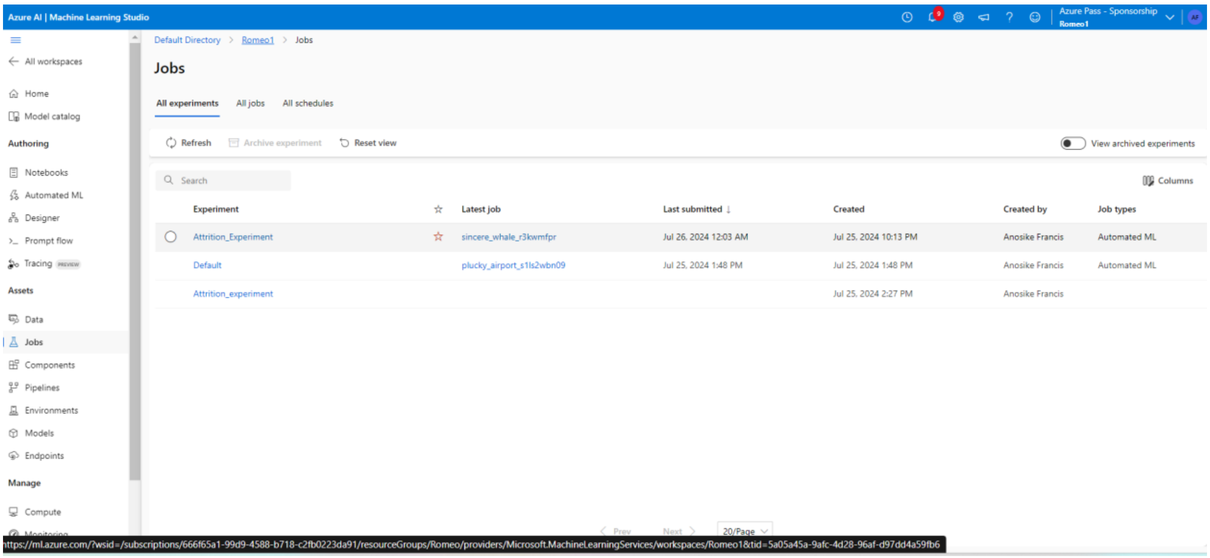


Experiment	Id	Type	Status	Details Page	Docs Page
Attrition_Experiment	AutoML_f4cff539-58a7-4e8e-80ad-cb3c8282a00d	automl	Preparing	Link to Azure Machine Learning studio	Link to Documentation

```
Current status: DatasetEvaluation. Gathering dataset statistics.
Current status: FeaturesGeneration. Generating features for the dataset.
Current status: DatasetFeaturization. Beginning to fit featurizers and featurize the dataset.
Current status: DatasetFeaturizationCompleted. Completed fit featurizers and featurizing the dataset.
Current status: DatasetCrossValidationSplit. Generating individually featurized CV splits.
2024/07/25 20:55:31 WARNING mlflow.sklearn: Model was missing function: predict. Not logging python_function flavor!

********************************************************************************************
DATA GUARDRAILS: 

TYPE:         Class balancing detection
STATUS:       PASSED
DESCRIPTION:  Your inputs were analyzed, and all classes are balanced in your training data.
              Learn more about imbalanced data: https://aka.ms/AutomatedMLImbalancedData

********************************************************************************************

TYPE:         Missing feature values imputation
STATUS:       PASSED
DESCRIPTION:  No feature missing values were detected in the training data.
              Learn more about missing value imputation: https://aka.ms/AutomatedMLFeaturization

********************************************************************************************

TYPE:         High cardinality feature detection
STATUS:       PASSED
DESCRIPTION:  Your inputs were analyzed, and no high cardinality features were detected.
              Learn more about high cardinality feature handling: https://aka.ms/AutomatedMLFeaturization

********************************************************************************************
Current status: ModelSelection. Beginning model selection.

********************************************************************************************
ITER: The iteration being evaluated.
```



```
********************************************************************************************
ITER: The iteration being evaluated.
PIPELINE: A summary description of the pipeline being evaluated.
DURATION: Time taken for the current iteration.
METRIC: The result of computing score on the fitted pipeline.
BEST: The best observed score thus far.
********************************************************************************************

 ITER   PIPELINE                                       DURATION            METRIC      BEST
    0   MaxAbsScaler LightGBM                          0:00:37             0.8267    0.8267
    1   MaxAbsScaler XGBoostClassifier                 0:00:57             0.8133    0.8267
    2   MaxAbsScaler ExtremeRandomTrees                0:00:36             0.7957    0.8267
    3   MaxAbsScaler RandomForest                      0:00:36             0.7862    0.8267
    4   StandardScalerWrapper LightGBM                 0:00:36             0.8111    0.8267
    5   SparseNormalizer XGBoostClassifier             0:00:48             0.8079    0.8267
    6   SparseNormalizer RandomForest                  0:00:41             0.8143    0.8267
    7   StandardScalerWrapper XGBoostClassifier        0:00:47             0.7793    0.8267
    8   SparseNormalizer XGBoostClassifier             0:00:49             0.8023    0.8267
    9   MaxAbsScaler RandomForest                      0:00:36             0.7863    0.8267
   10   SparseNormalizer LightGBM                      0:00:36             0.8060    0.8267
   11   MaxAbsScaler ExtremeRandomTrees                0:00:46             0.8254    0.8267
   12   StandardScalerWrapper XGBoostClassifier        0:00:47             0.8232    0.8267
   13   StandardScalerWrapper ExtremeRandomTrees       0:00:37             0.8154    0.8267
   14   StandardScalerWrapper RandomForest             0:00:55             0.8179    0.8267
   15   MaxAbsScaler LightGBM                          0:00:37             0.8221    0.8267
   16   MaxAbsScaler LogisticRegression                0:00:36             0.8444    0.8444
   17   StandardScalerWrapper ExtremeRandomTrees       0:00:56             0.7987    0.8444
   18   StandardScalerWrapper XGBoostClassifier        0:00:47             0.8127    0.8444
   19   MaxAbsScaler ExtremeRandomTrees                0:00:38             0.8227    0.8444
   20   VotingEnsemble                                 0:00:45             0.8452    0.8452
   21   StackEnsemble                                  0:00:53             0.8440    0.8452
Stopping criteria reached at iteration 22. Ending experiment.
********************************************************************************************
Current status: BestRunExplainModel. Best run model explanations started
Current status: ModelExplanationDataSetSetup. Model explanations data setup completed
Current status: PickSurrogateModel. Choosing LightGBM as the surrogate model for explanations
Current status: EngineeredFeatureExplanations. Computation of engineered features started
2024-07-25:21:13:22.537 INFO      [explanation_client.py:334] Using default datastore for uploads
```

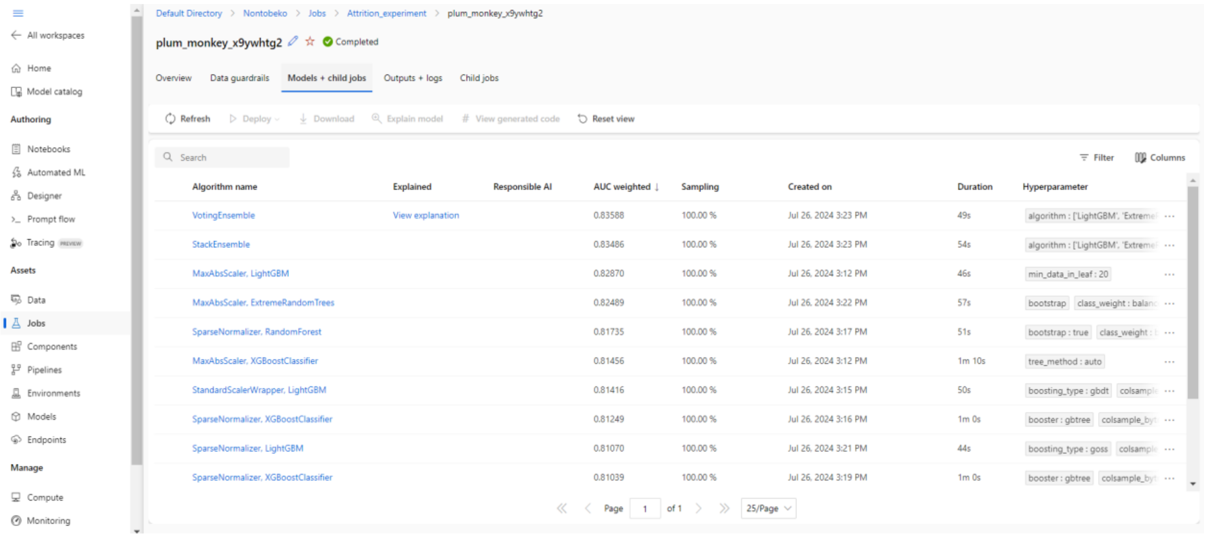


Jobs

	Experiment	☆	Latest job	Last submitted ↓	Created	Created by	Job types
○	Attrition_Experiment	☆	sincere_whale_r3kwmfpr	Jul 26, 2024 12:03 AM	Jul 25, 2024 10:13 PM	Anosike Francis	Automated ML
	Default		plucky_airport_s1ls2wbn09	Jul 25, 2024 1:48 PM	Jul 25, 2024 1:48 PM	Anosike Francis	Automated ML
	Attrition_experiment				Jul 25, 2024 2:27 PM	Anosike Francis	

ttps://ml.azure.com/?wsid=/subscriptions/666f65a1-99d9-4588-b718-c2fb0223da91/resourceGroups/Romeo/providers/Microsoft.MachineLearningServices/workspaces/Romeo1&tid=5a05a45a-9afc-4d28-96af-d97dd4a59fb6

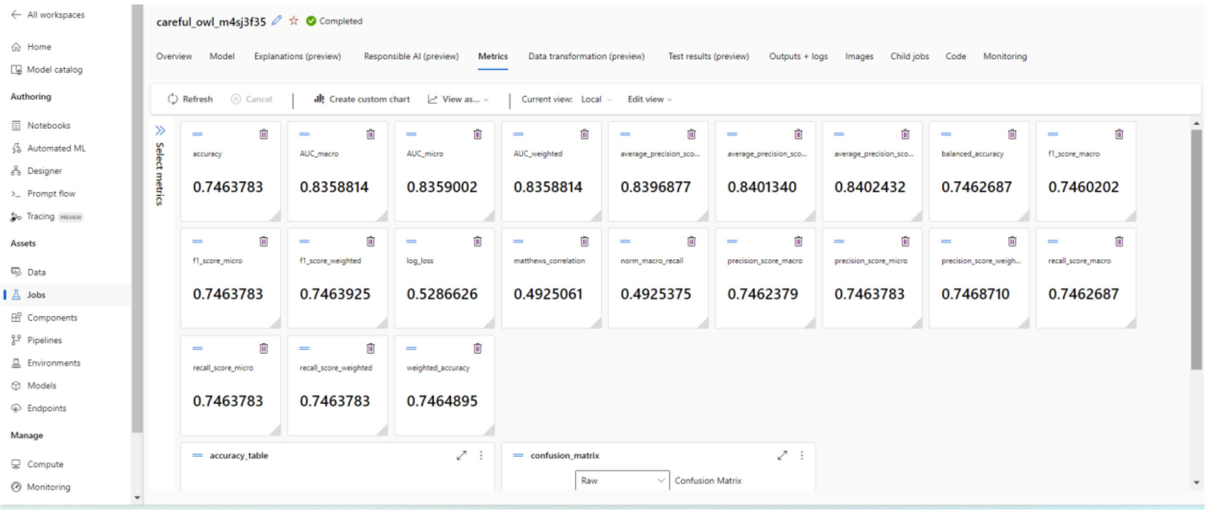
Choose best model based on high Accuracy



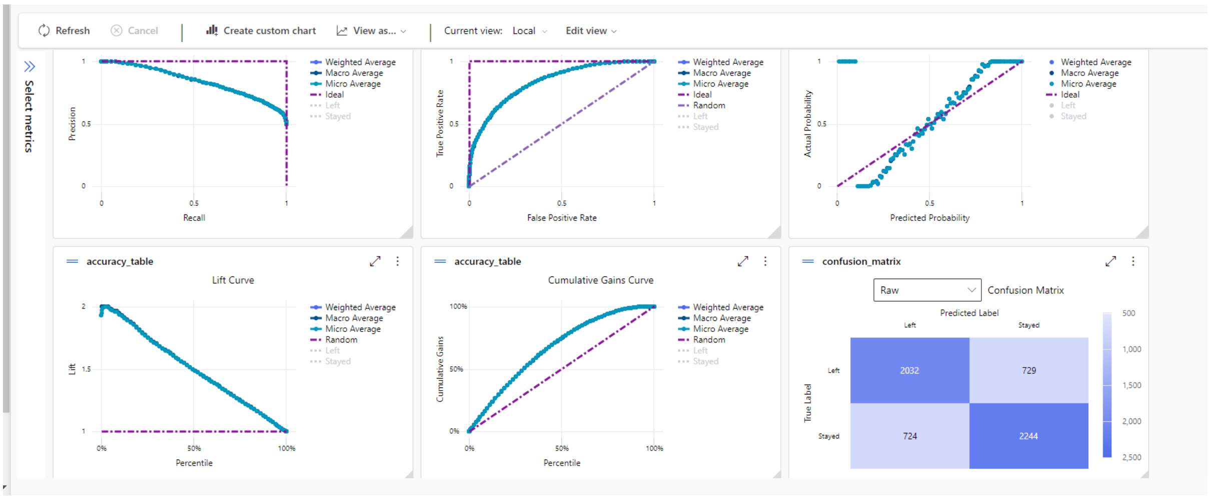
plum_monkey_x9ywhtg2 ☆ Completed

Algorithm name	Explained	Responsible AI	AUC weighted ↓	Sampling	Created on	Duration	Hyperparameter
VotingEnsemble	View explanation		0.83588	100.00 %	Jul 26, 2024 3:23 PM	49s	algorithm : ['LightGBM', 'Extreme
StackEnsemble			0.83486	100.00 %	Jul 26, 2024 3:23 PM	54s	algorithm : ['LightGBM', 'Extreme
MaxAbsScaler, LightGBM			0.82870	100.00 %	Jul 26, 2024 3:12 PM	46s	min_data_in_leaf : 20
MaxAbsScaler, ExtremeRandomTrees			0.82489	100.00 %	Jul 26, 2024 3:22 PM	57s	bootstrap class_weight : balanc
SparseNormalizer, RandomForest			0.81735	100.00 %	Jul 26, 2024 3:17 PM	51s	bootstrap : true class_weight :
MaxAbsScaler, XGBoostClassifier			0.81456	100.00 %	Jul 26, 2024 3:12 PM	1m 10s	tree_method : auto
StandardScalerWrapper, LightGBM			0.81416	100.00 %	Jul 26, 2024 3:15 PM	50s	boosting_type : gbdt colsample
SparseNormalizer, XGBoostClassifier			0.81249	100.00 %	Jul 26, 2024 3:16 PM	1m 0s	booster : gbtree colsample_byt
SparseNormalizer, LightGBM			0.81070	100.00 %	Jul 26, 2024 3:21 PM	44s	boosting_type : goss colsample
SparseNormalizer, XGBoostClassifier			0.81039	100.00 %	Jul 26, 2024 3:19 PM	1m 0s	booster : gbtree colsample_byt

Information about best model based on accuracy



accuracy	0.7463783
AUC_macro	0.8358814
AUC_micro	0.8359002
AUC_weighted	0.8358814
average_precision_sco...	0.8396877
average_precision_sco...	0.8401340
average_precision_sco...	0.8402432
balanced_accuracy	0.7462687
f1_score_macro	0.7460202
f1_score_micro	0.7463783
f1_score_weighted	0.7463925
log_loss	0.5286626
matthews_correlation	0.4925061
norm_macro_recall	0.4925375
precision_score_macro	0.7462379
precision_score_micro	0.7463783
precision_score_weigh...	0.7468710
recall_score_macro	0.7462687
recall_score_micro	0.7463783
recall_score_weighted	0.7463783
weighted_accuracy	0.7464895



Register the model

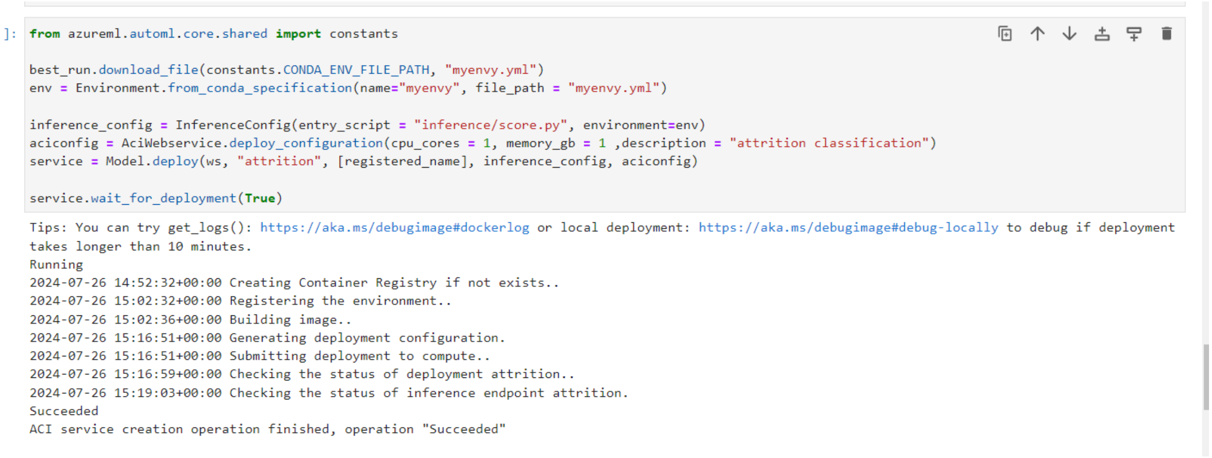
```
model_name = best_run.properties["model_name"]
```

```
registered_name = run.register_model(model_name = model_name, description = "AutoMl Attrition", tags = None)
```

```
from azureml.core.model import InferenceConfig
from azureml.core.webservice import AciWebservice, Webservice
from azureml.core.model import Model
from azureml.core.environment import Environment
```

```
#Download the scoring files
best_run.download_file("outputs/scoring_file_v_1_0_0.py", "inference/score.py")
```

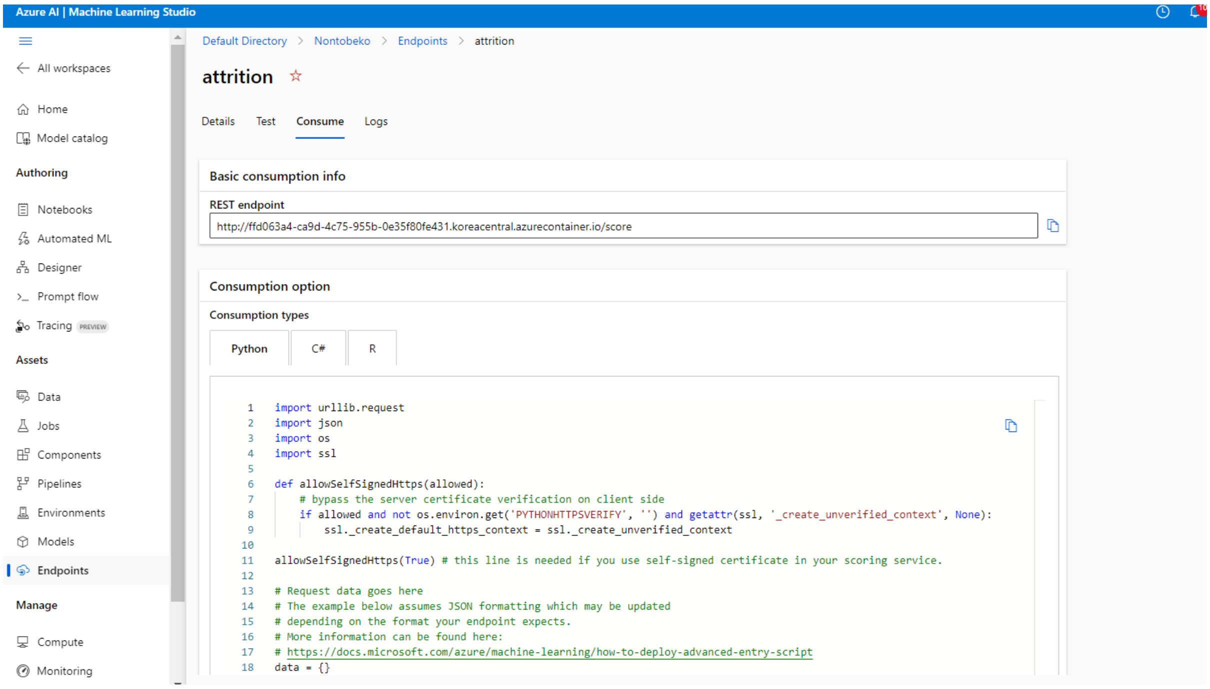
Start the deployment



```
]: from azureml.automl.core.shared import constants

best_run.download_file(constants.CONDA_ENV_FILE_PATH, "myenvy.yml")
env = Environment.from_conda_specification(name="myenvy", file_path = "myenvy.yml")

inference_config = InferenceConfig(entry_script = "inference/score.py", environment=env)
aciconfig = AciWebservice.deploy_configuration(cpu_cores = 1, memory_gb = 1 ,description = "attrition classification")
service = Model.deploy(ws, "attrition", [registered_name], inference_config, aciconfig)

service.wait_for_deployment(True)
```

```
Tips: You can try get_logs(): https://aka.ms/debugimage#dockerlog or local deployment: https://aka.ms/debugimage#debug-locally to debug if deployment
takes longer than 10 minutes.
Running
2024-07-26 14:52:32+00:00 Creating Container Registry if not exists..
2024-07-26 15:02:32+00:00 Registering the environment..
2024-07-26 15:02:36+00:00 Building image..
2024-07-26 15:16:51+00:00 Generating deployment configuration.
2024-07-26 15:16:51+00:00 Submitting deployment to compute..
2024-07-26 15:16:59+00:00 Checking the status of deployment attrition..
2024-07-26 15:19:03+00:00 Checking the status of inference endpoint attrition.
Succeeded
ACI service creation operation finished, operation "Succeeded"
```

Check endpoints and consume the model at test



attrition ☆

Basic consumption info

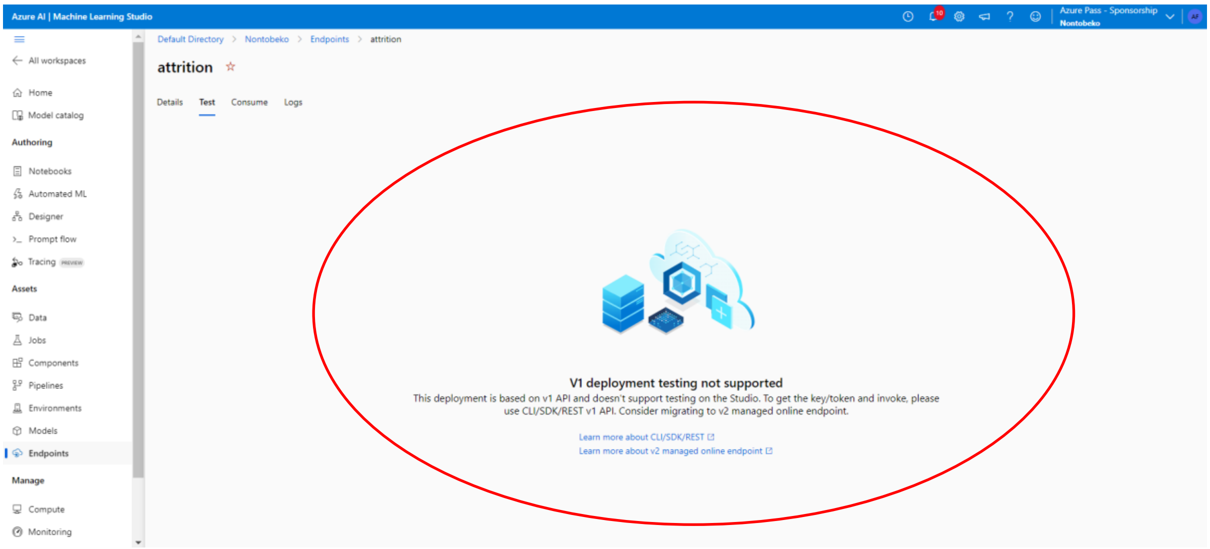
REST endpoint

http://ffd063a4-ca9d-4c75-955b-0e35f80fe431.koreacentral.azurecontainer.io/score

Consumption option

Consumption types

```
import urllib.request
import json
import os
import ssl

def allowSelfSignedHttps(allowed):
    # bypass the server certificate verification on client side
    if allowed and not os.environ.get('PYTHONHTTPSVERIFY', '') and getattr(ssl, '_create_unverified_context', None):
        ssl._create_default_https_context = ssl._create_unverified_context

allowSelfSignedHttps(True) # this line is needed if you use self-signed certificate in your scoring service.

# Request data goes here
# The example below assumes JSON formatting which may be updated
# depending on the format your endpoint expects.
# More information can be found here:
# https://docs.microsoft.com/azure/machine-learning/how-to-deploy-advanced-entry-script
data = {}
```



V1 deployment testing not supported

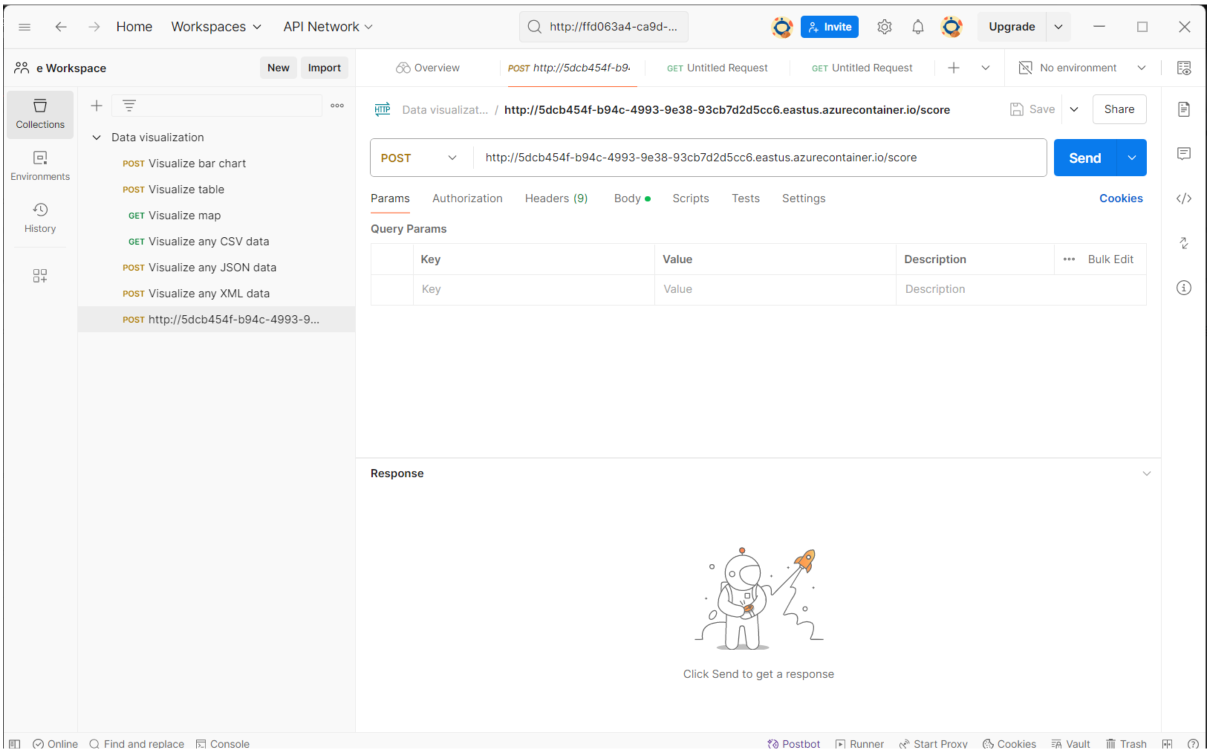
This deployment is based on v1 API and doesn't support testing on the Studio. To get the key/token and invoke, please use CLI/SDK/REST v1 API. Consider migrating to v2 managed online endpoint.

Learn more about CLI/SDK/REST

Learn more about v2 managed online endpoint

Use third party "POSTMAN" for consuming

http://ffd063a4-ca9d-4c75-955b-0e35f80fe431.koreacentral.azurecontainer.io/score



Data visualizat... / http://5dcb454f-b94c-4993-9e38-93cb7d2d5cc6.eastus.azurecontainer.io/score

POST http://5dcb454f-b94c-4993-9e38-93cb7d2d5cc6.eastus.azurecontainer.io/score

Query Params

Key	Value	Description
Key	Value	Description

Click Send to get a response