**Azure ML - Hands-on Assessment**

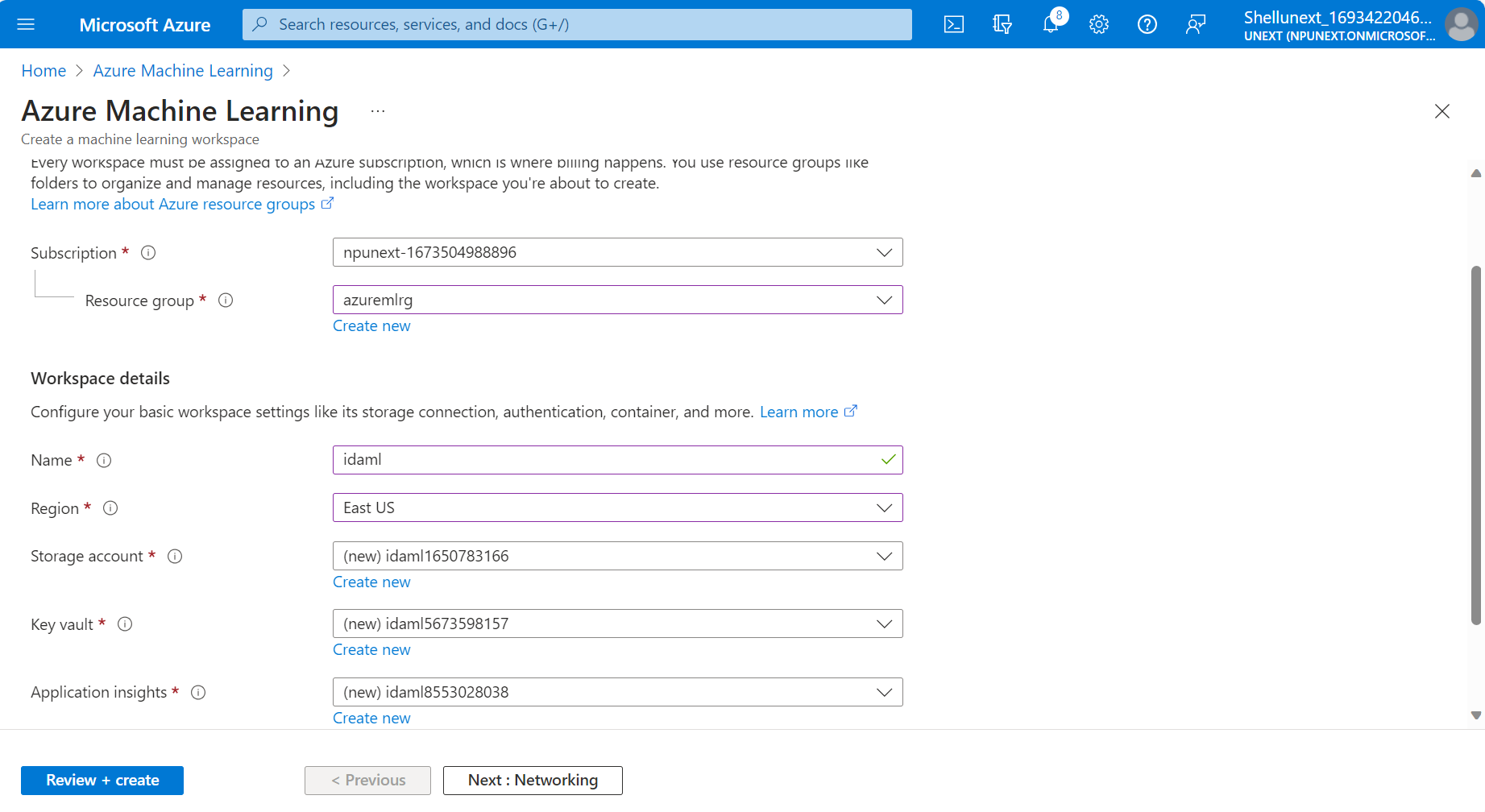
*By Yuva Sahith Varma Sangaraju*

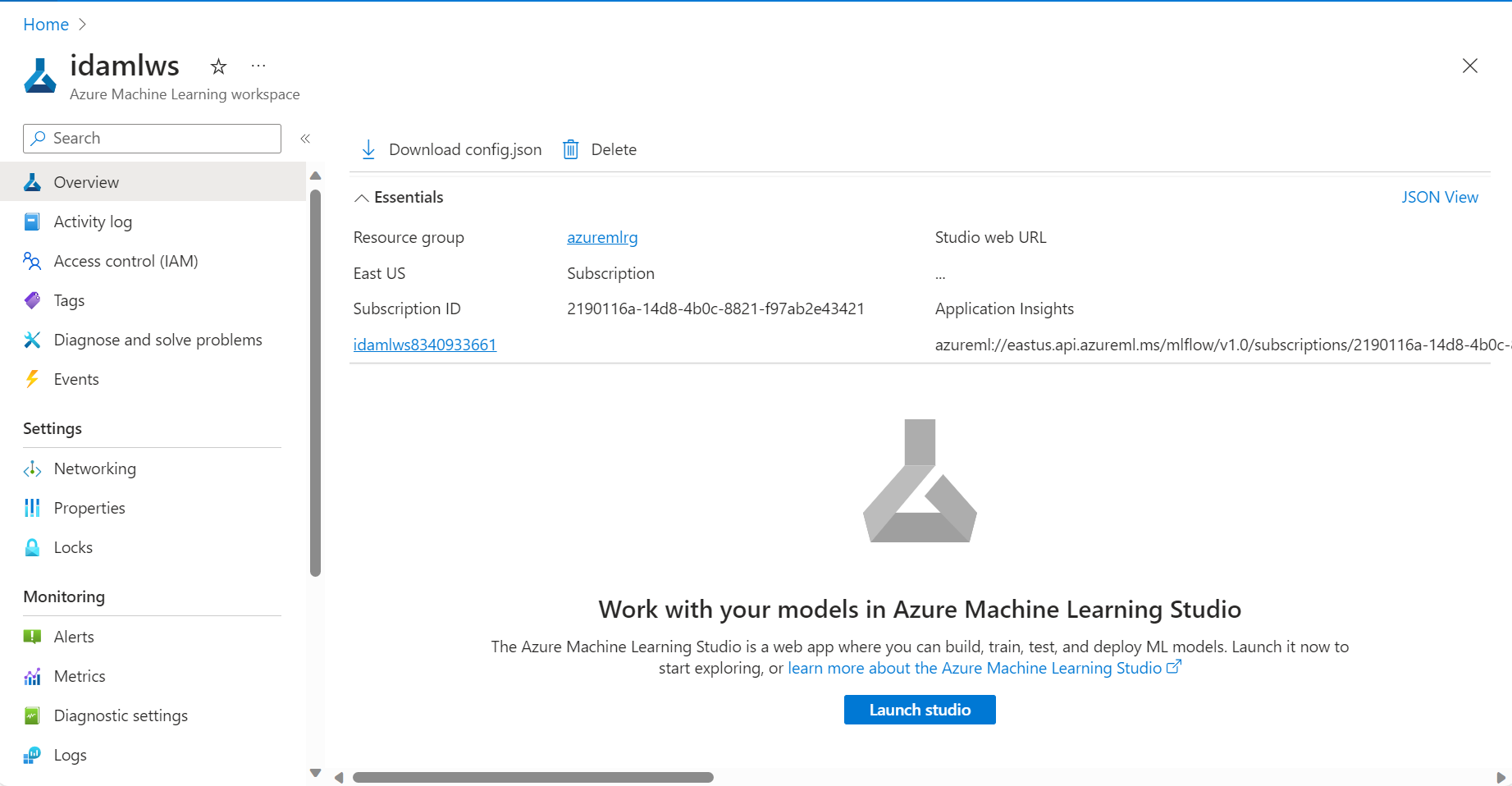
# Scenario:

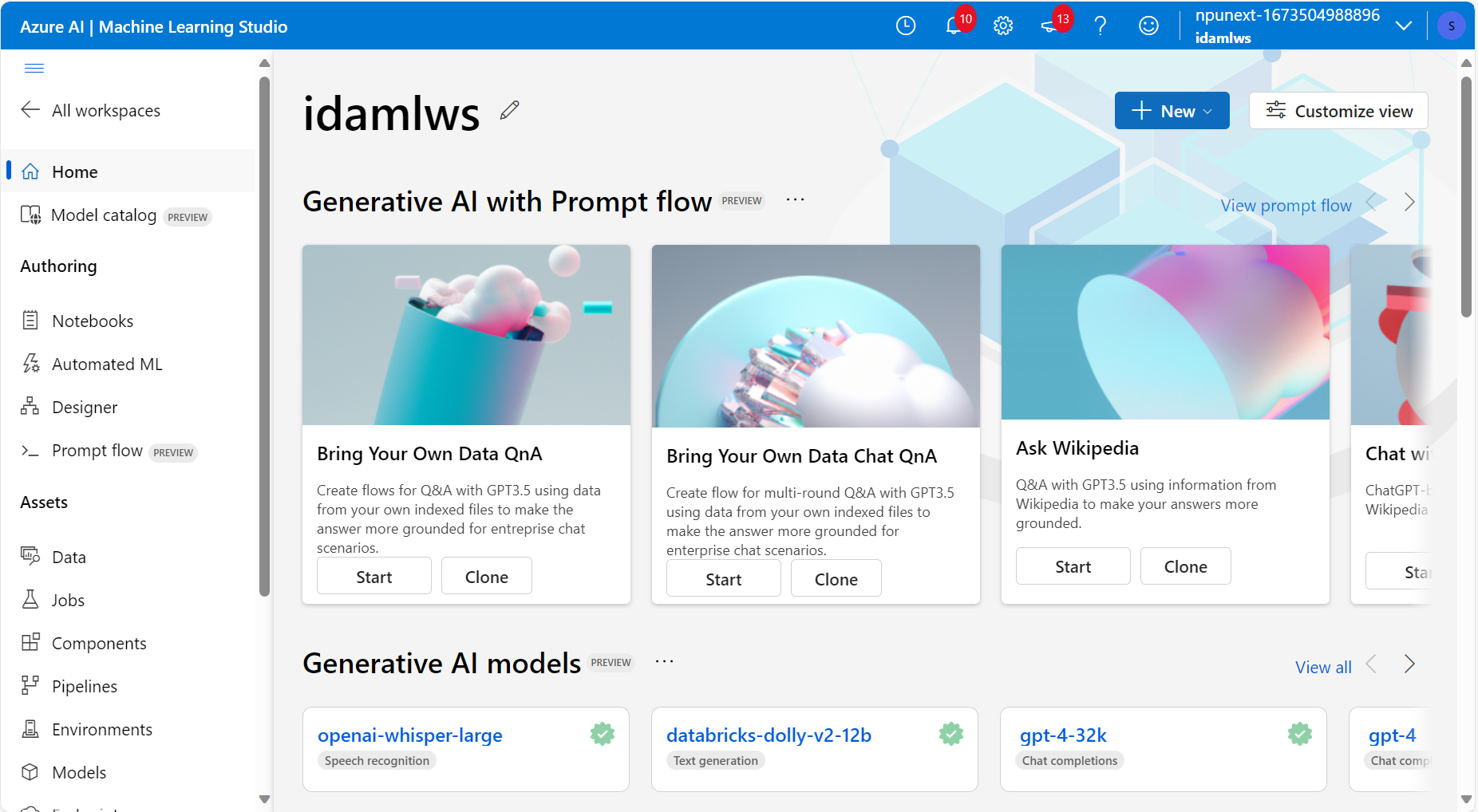
You are a data scientist at a retail company, and you have been tasked with developing a machine learning model to predict customer purchasing behaviour based on historical data. The company wants to personalize marketing strategies to target specific customer segments more effectively. You will use Azure Machine Learning to build and train the machine learning model.

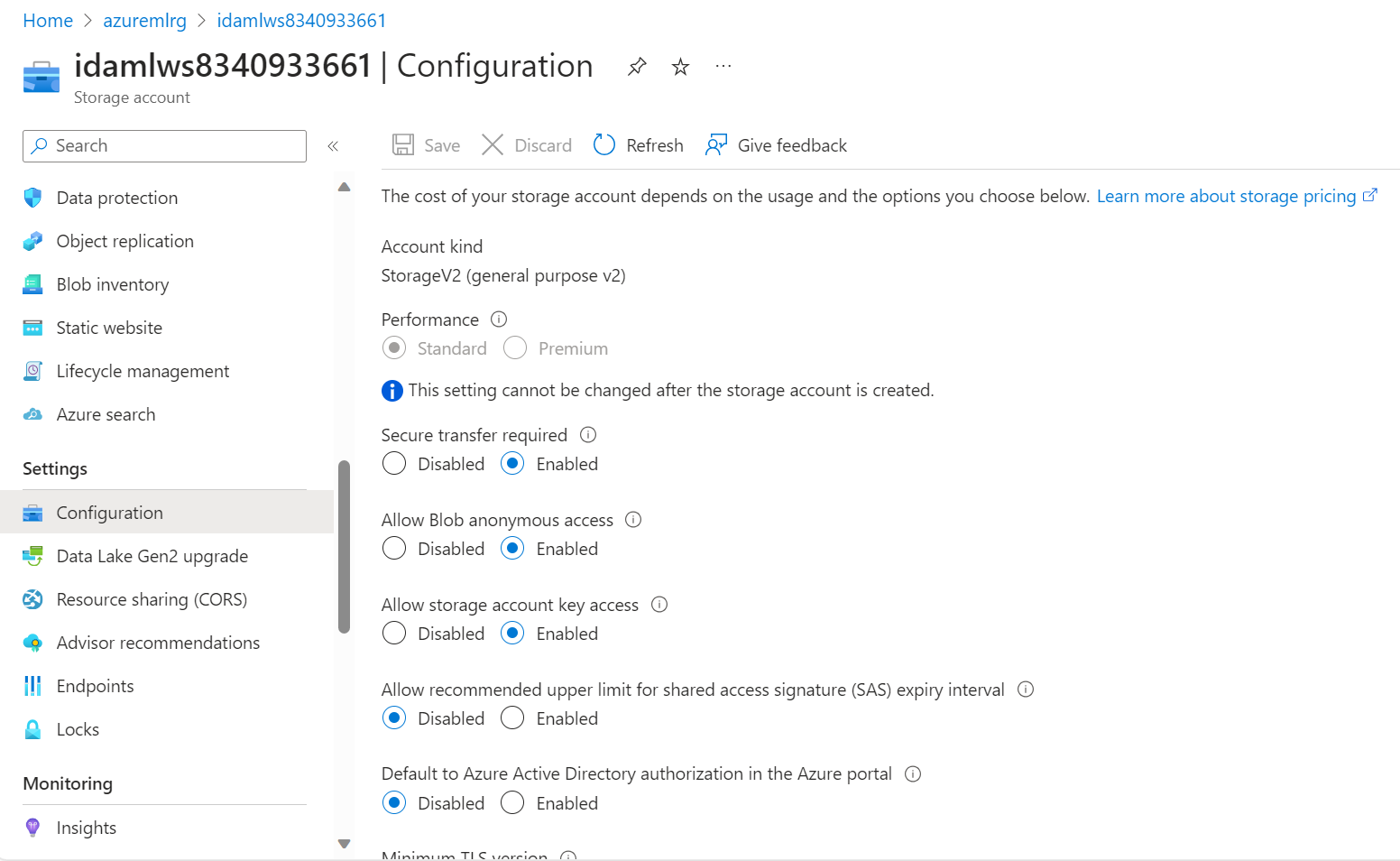
# Tasks:

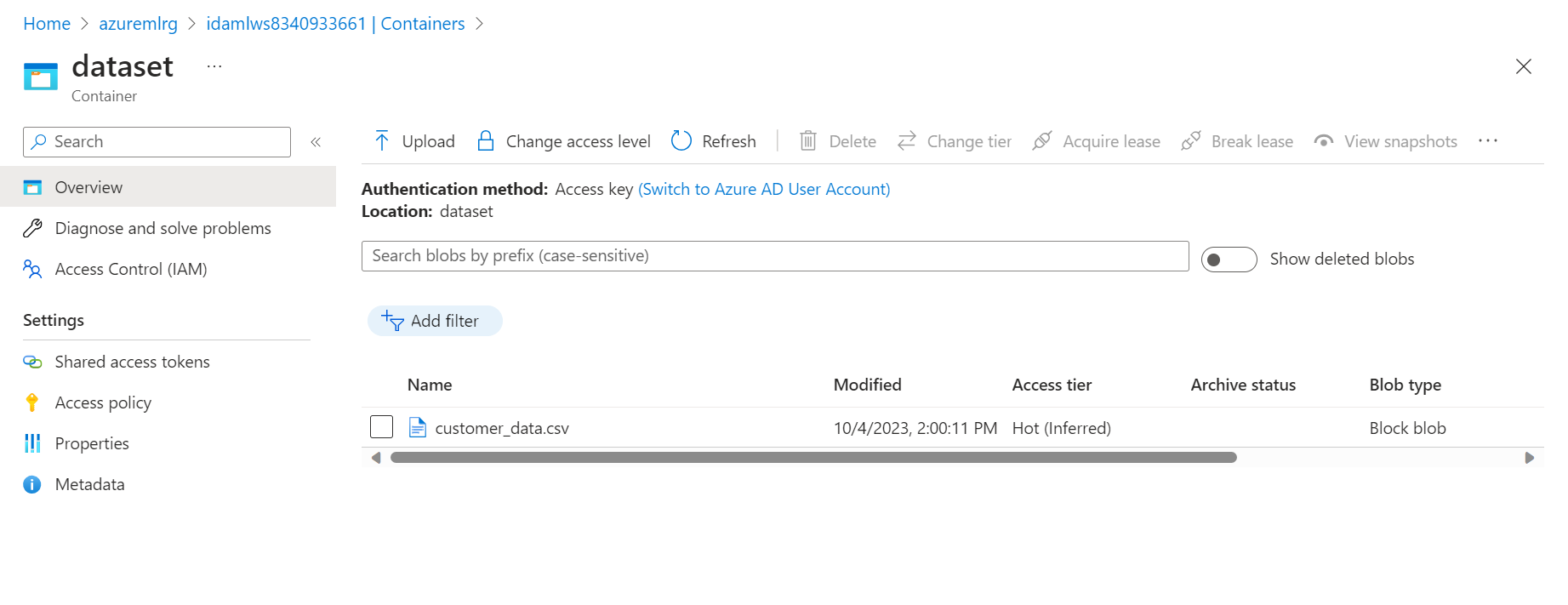
1. Data Preparation: a. Access the provided dataset in Azure Blob Storage containing historical customer data. b. Load the dataset and preprocess it to prepare it for training. Perform necessary data cleaning, feature selection, and feature engineering.

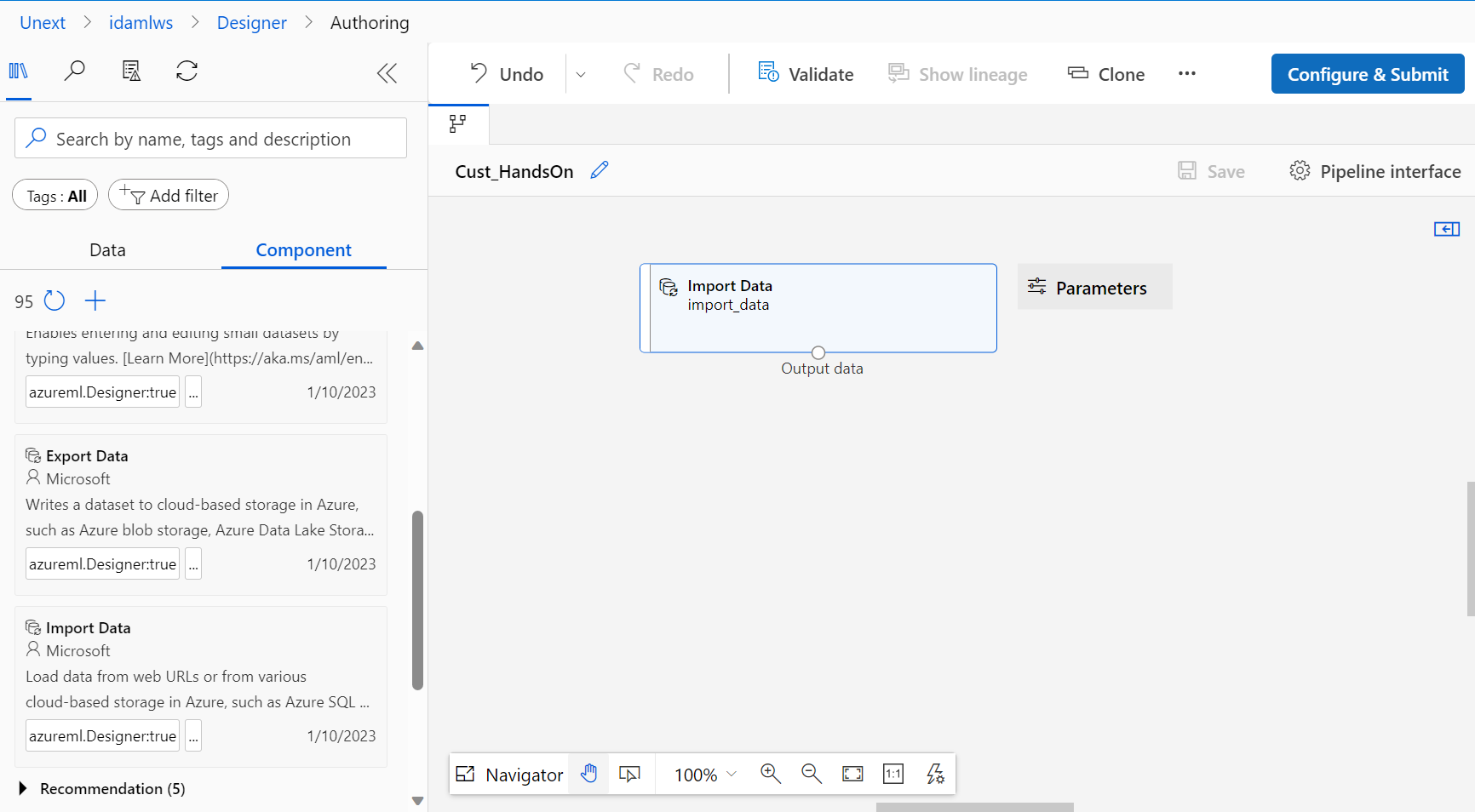


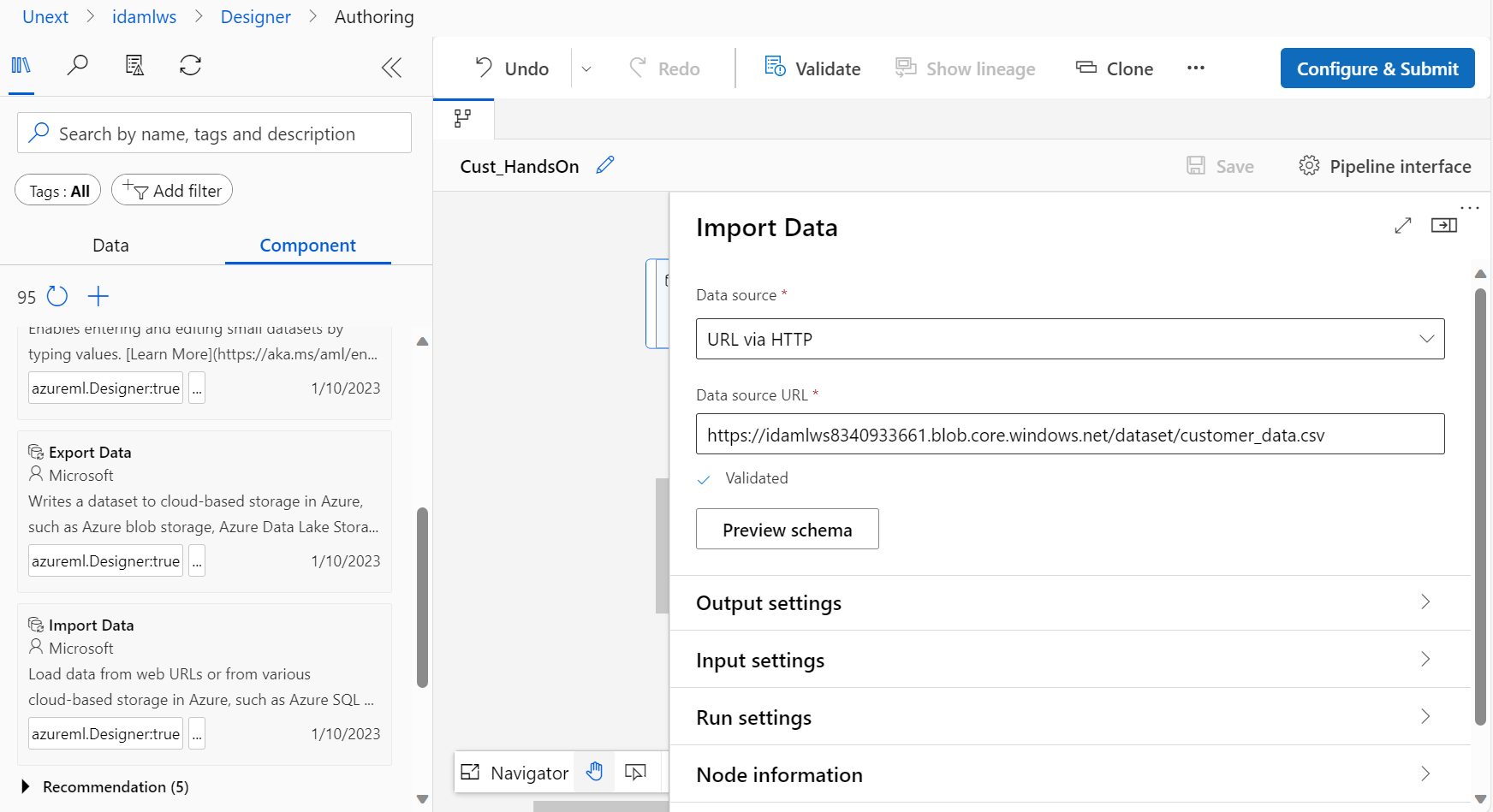


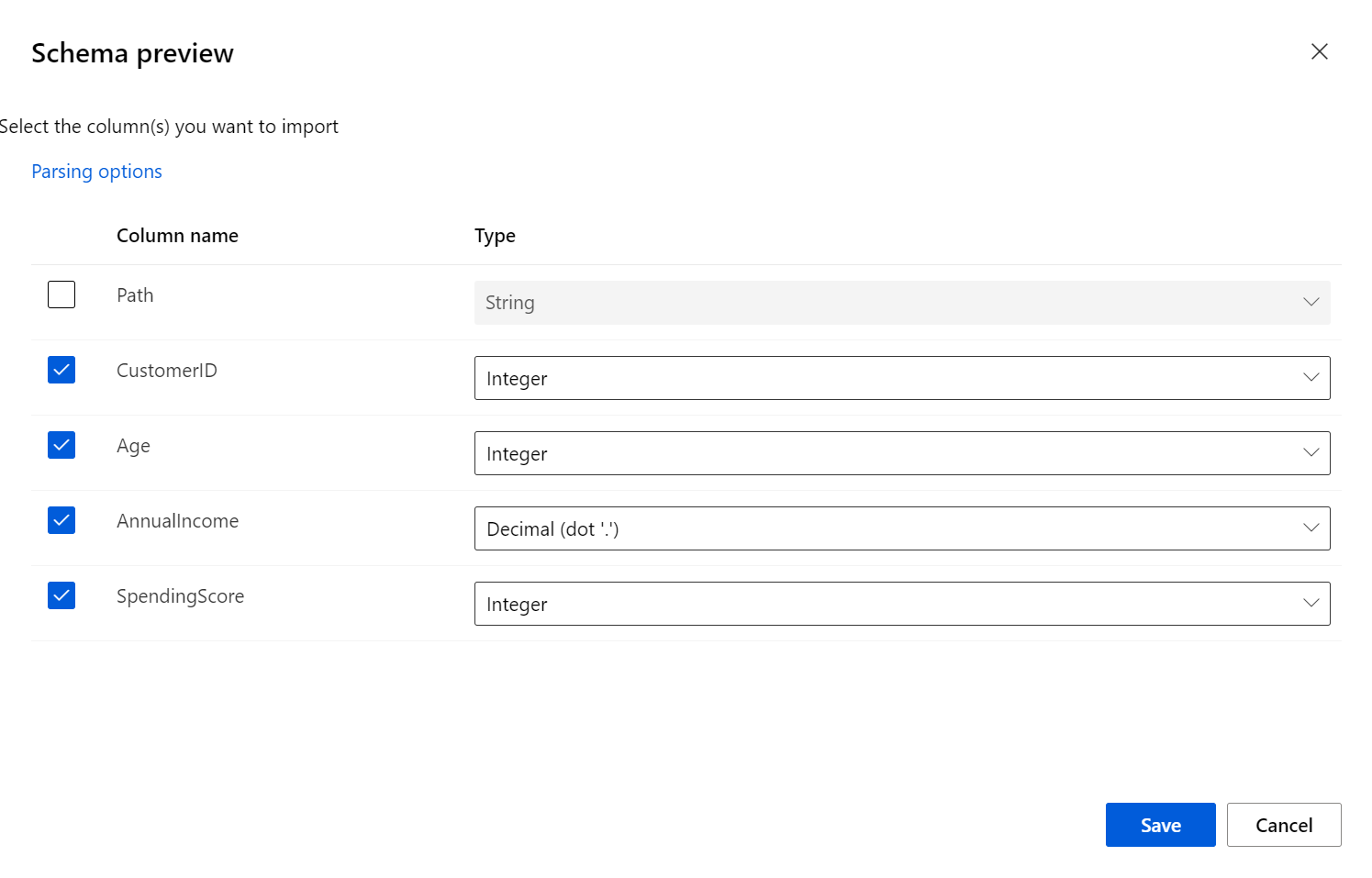


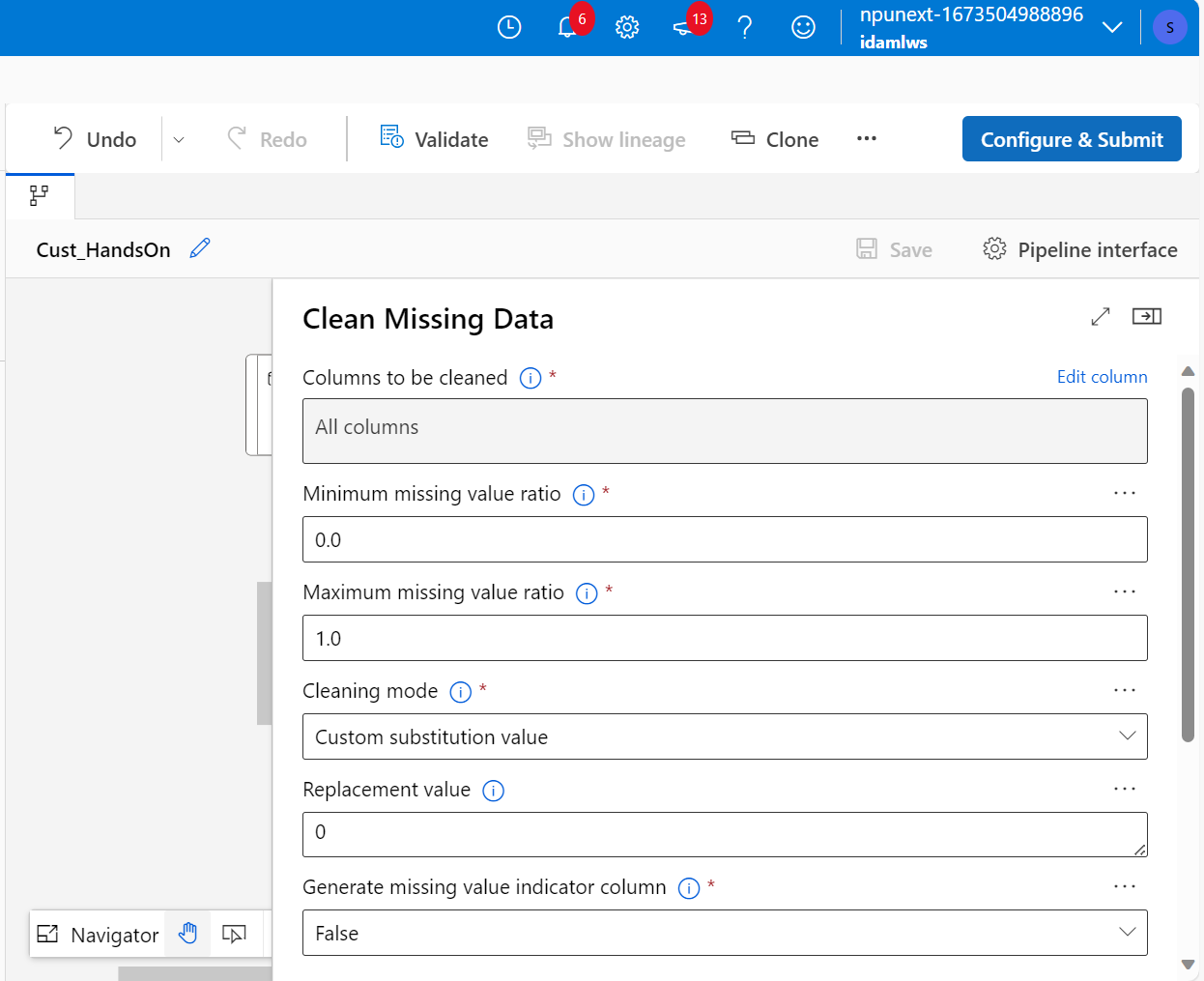


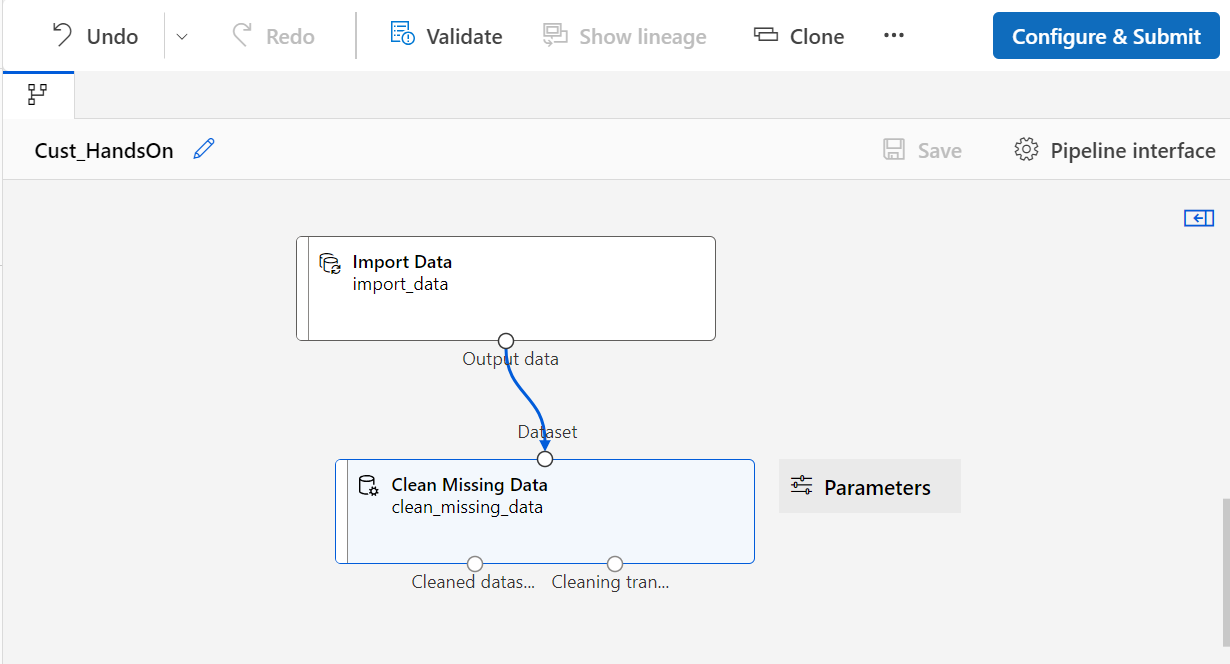




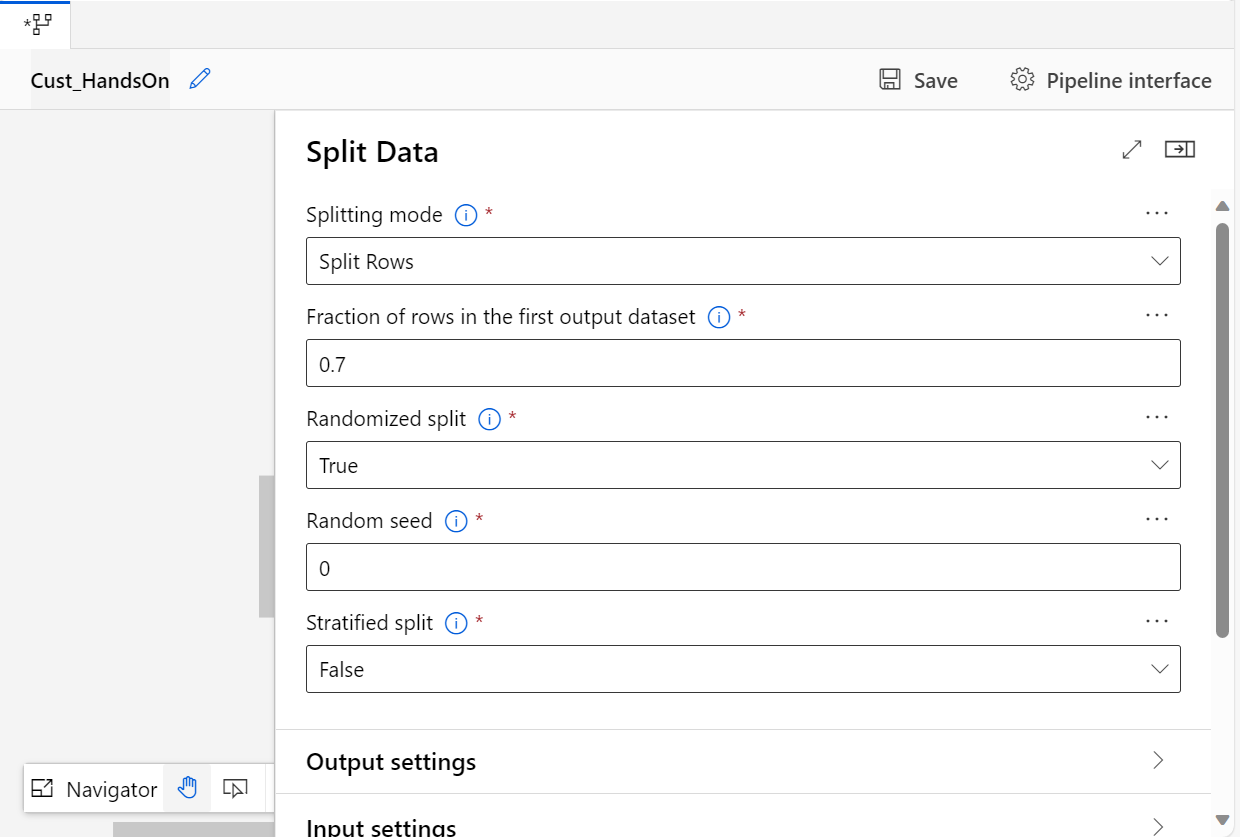




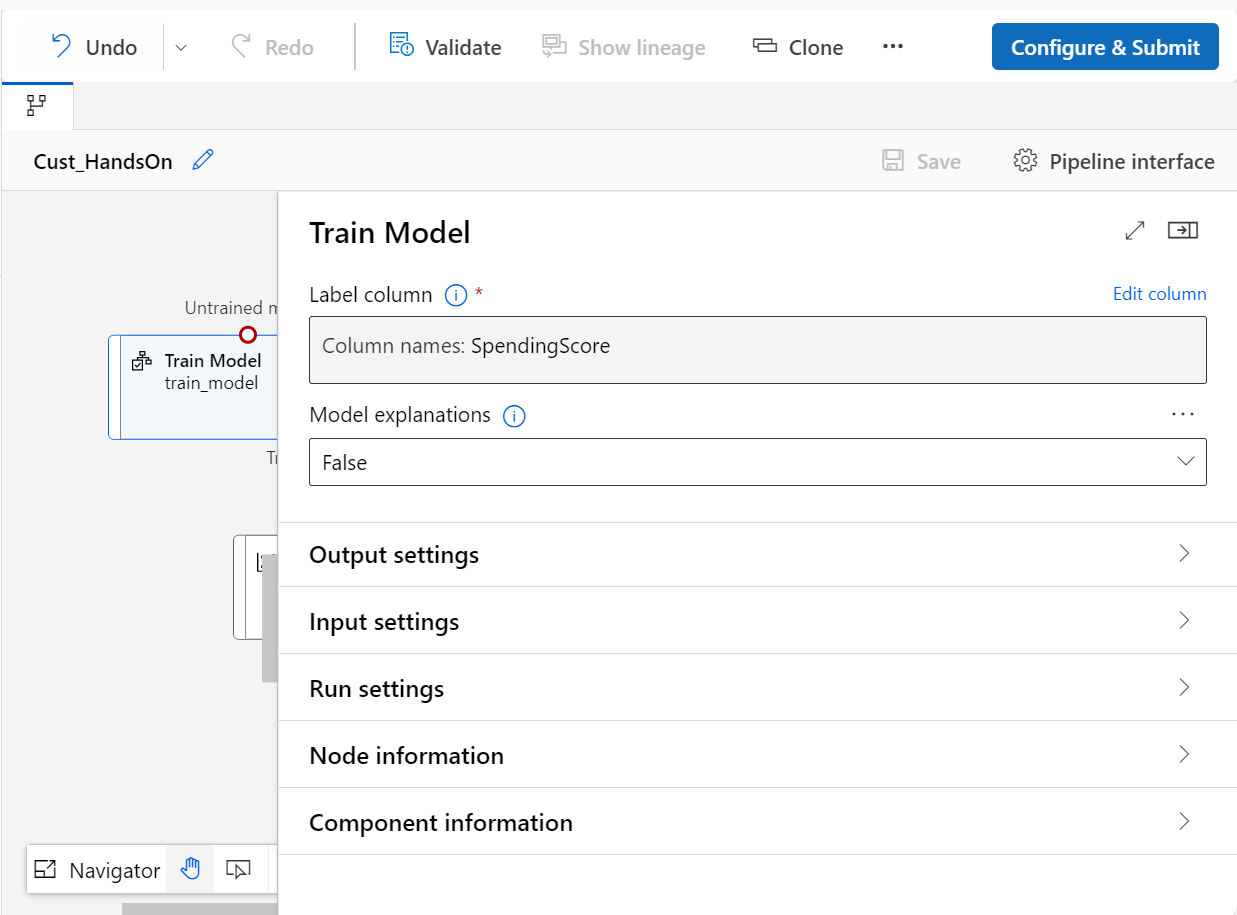


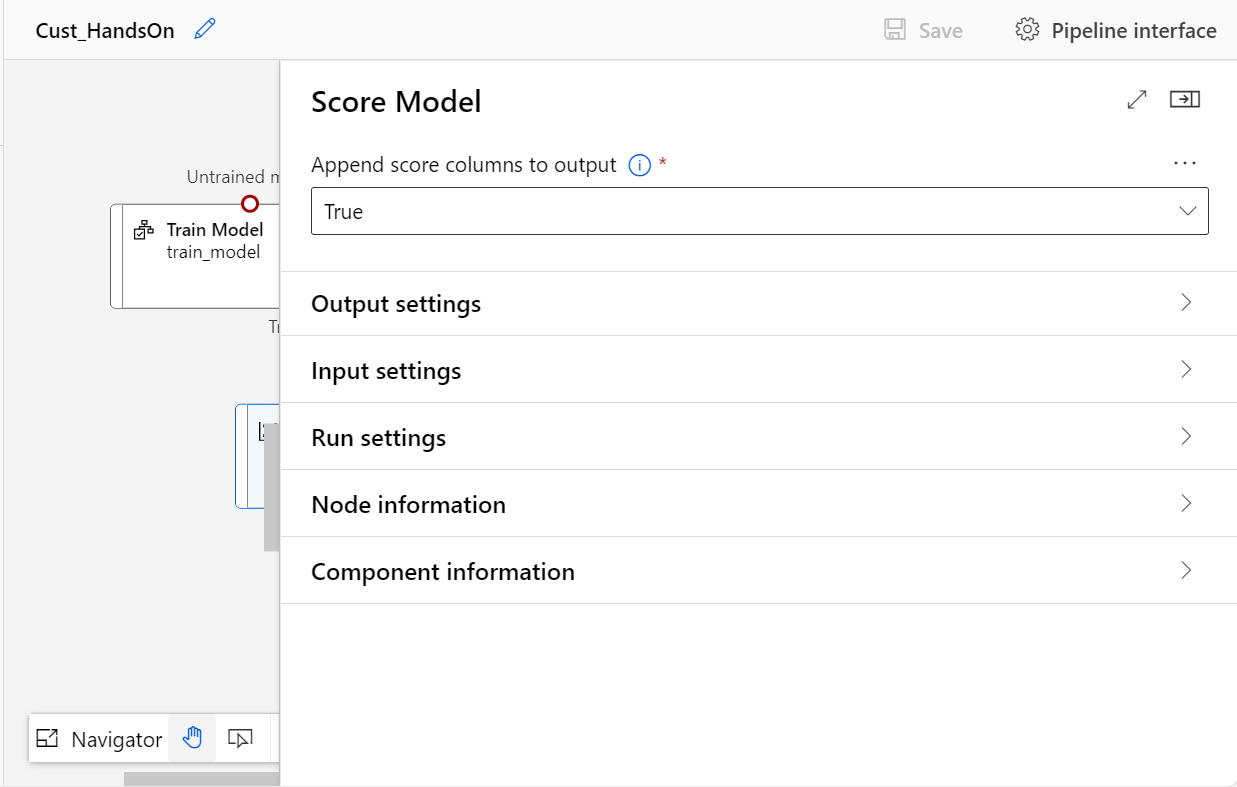


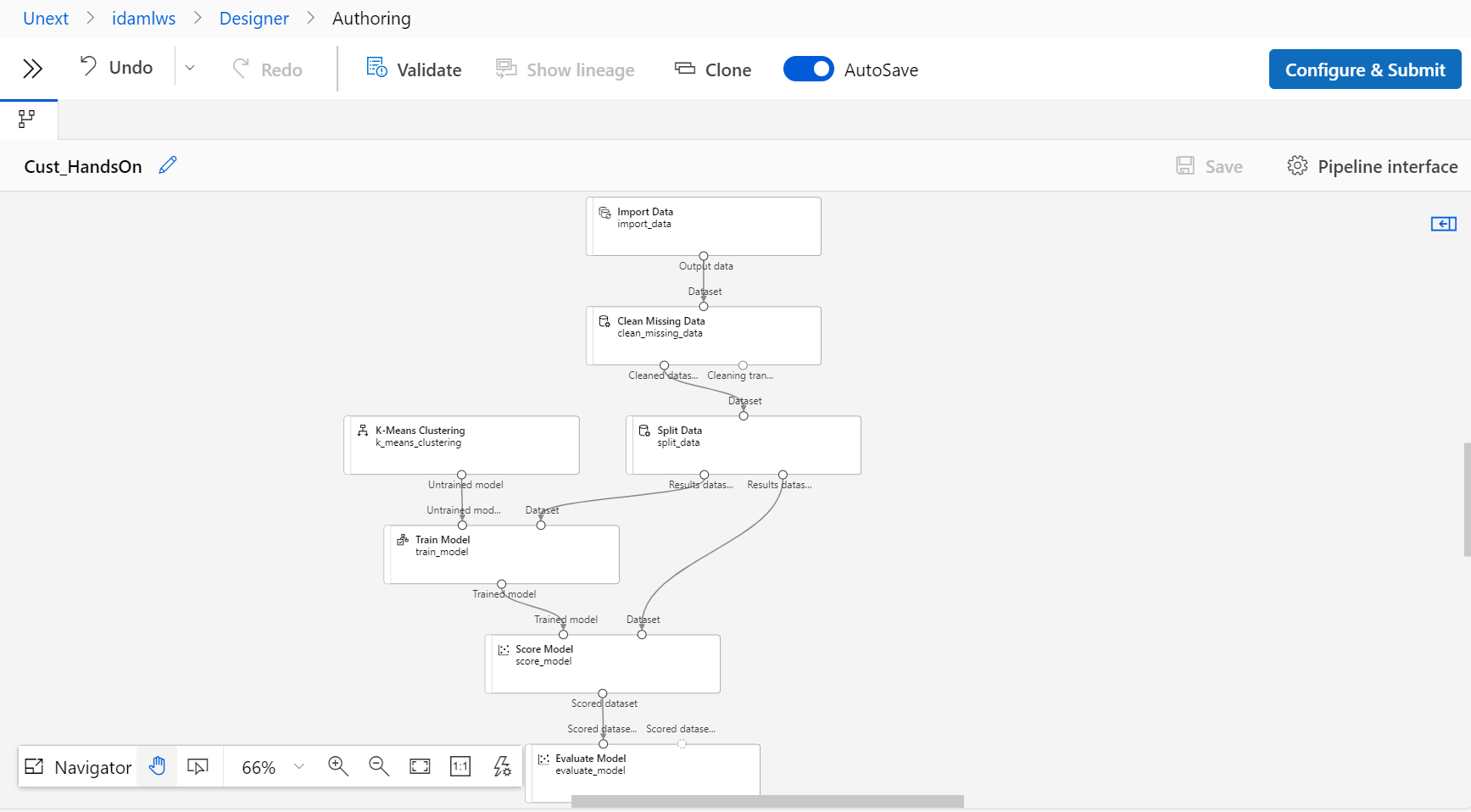
1. Model Development: a. Choose an appropriate machine learning algorithm for predicting customer purchasing behaviour. b. Split the dataset into training and testing sets. c. Train the machine learning model using the training dataset. d. Evaluate the model's performance using appropriate evaluation metrics.

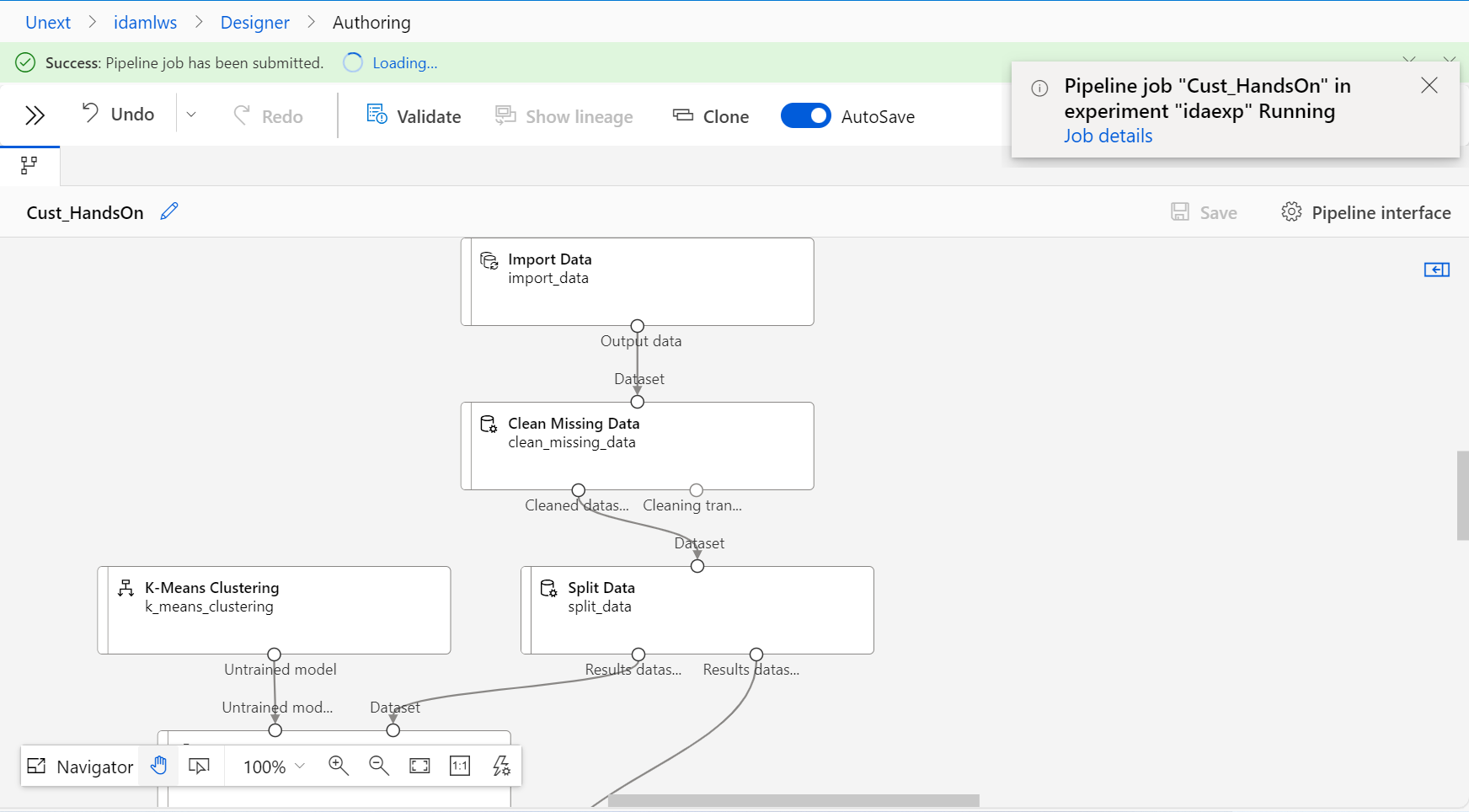


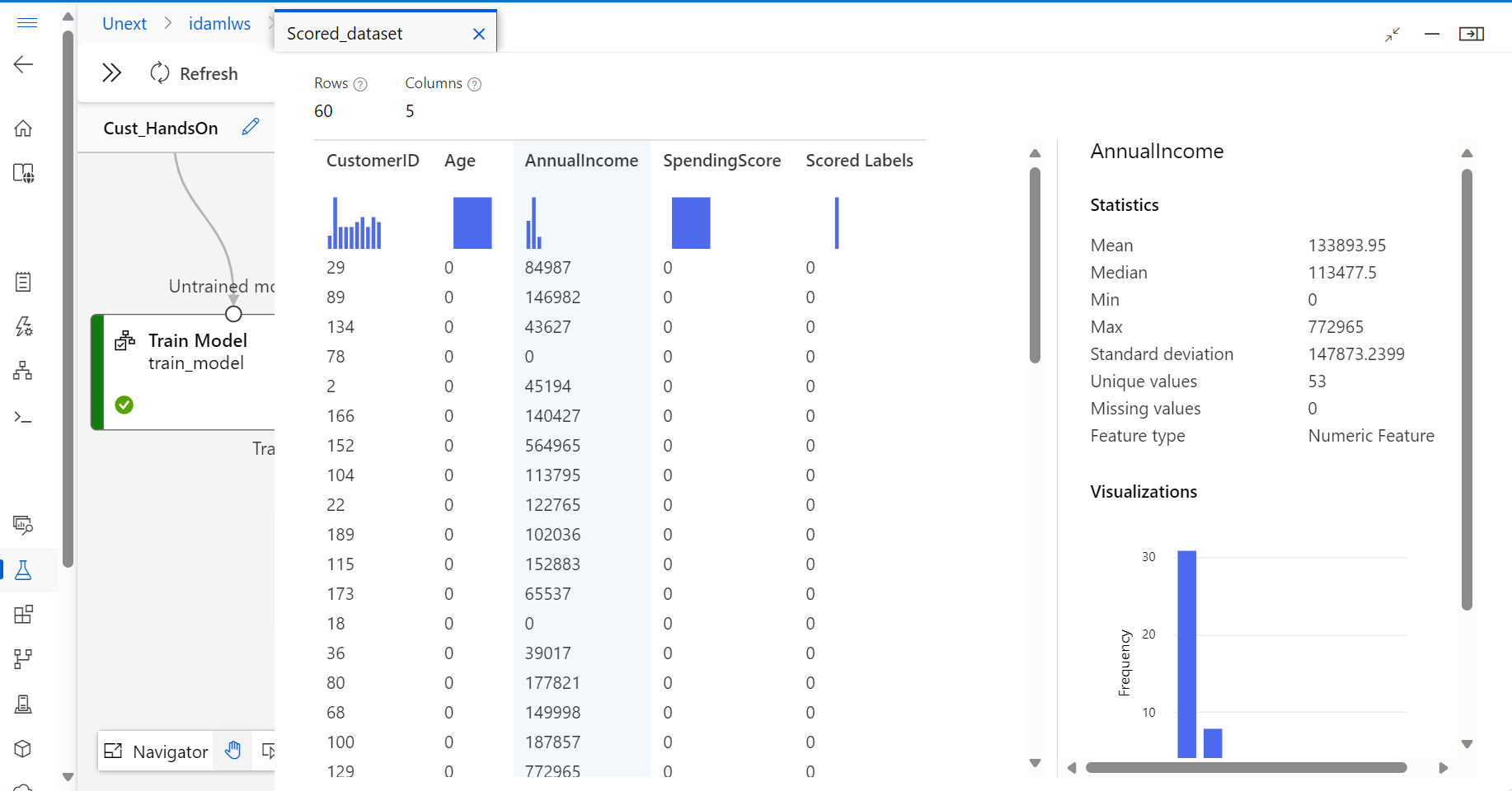
Selecting k-means clustering



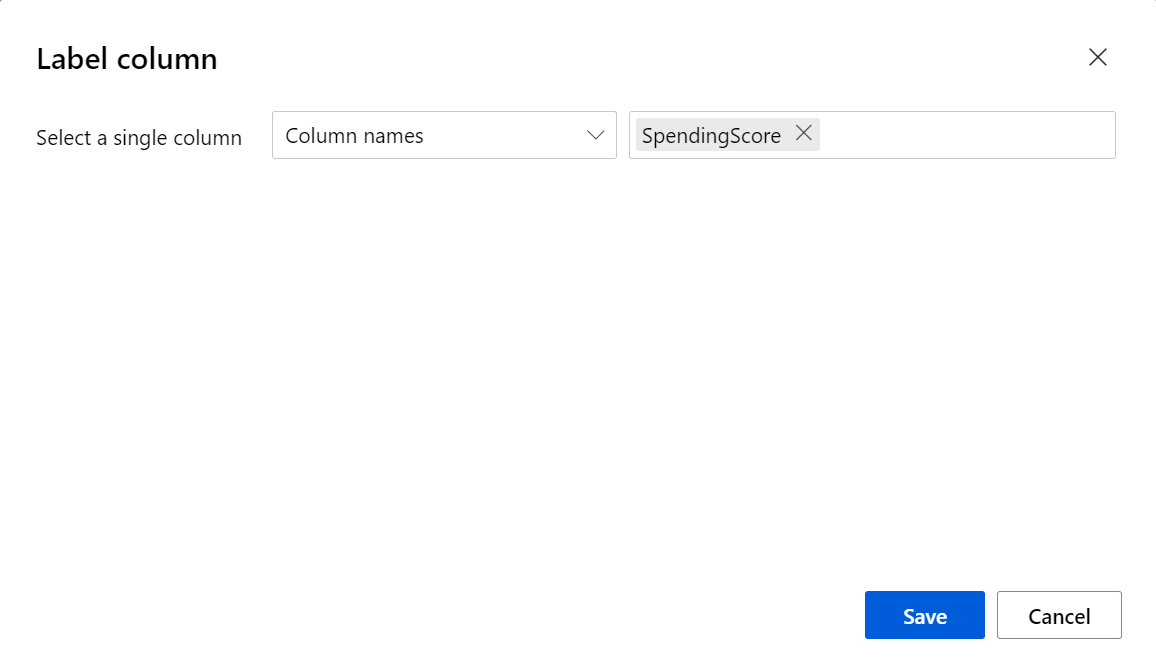


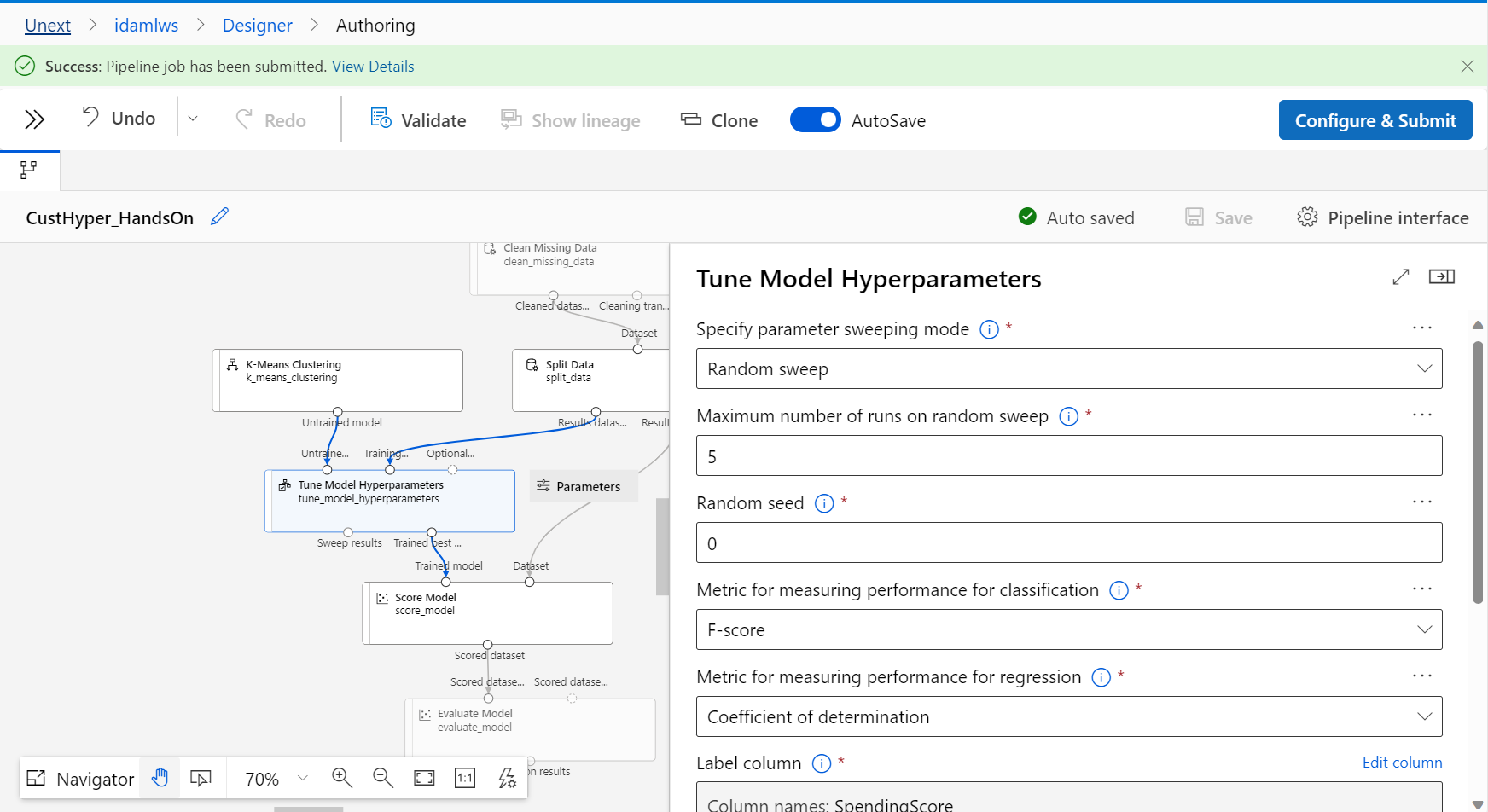




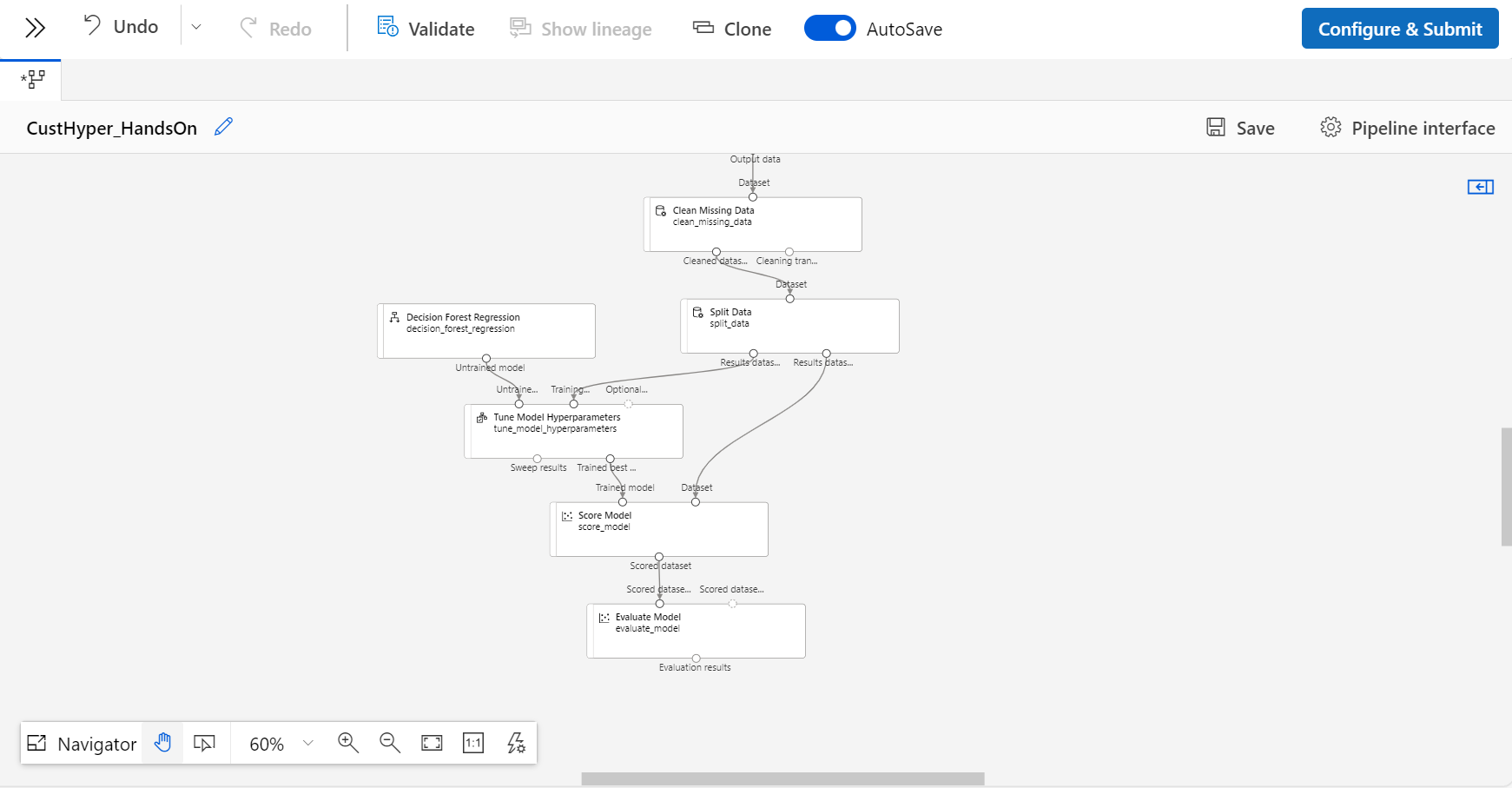


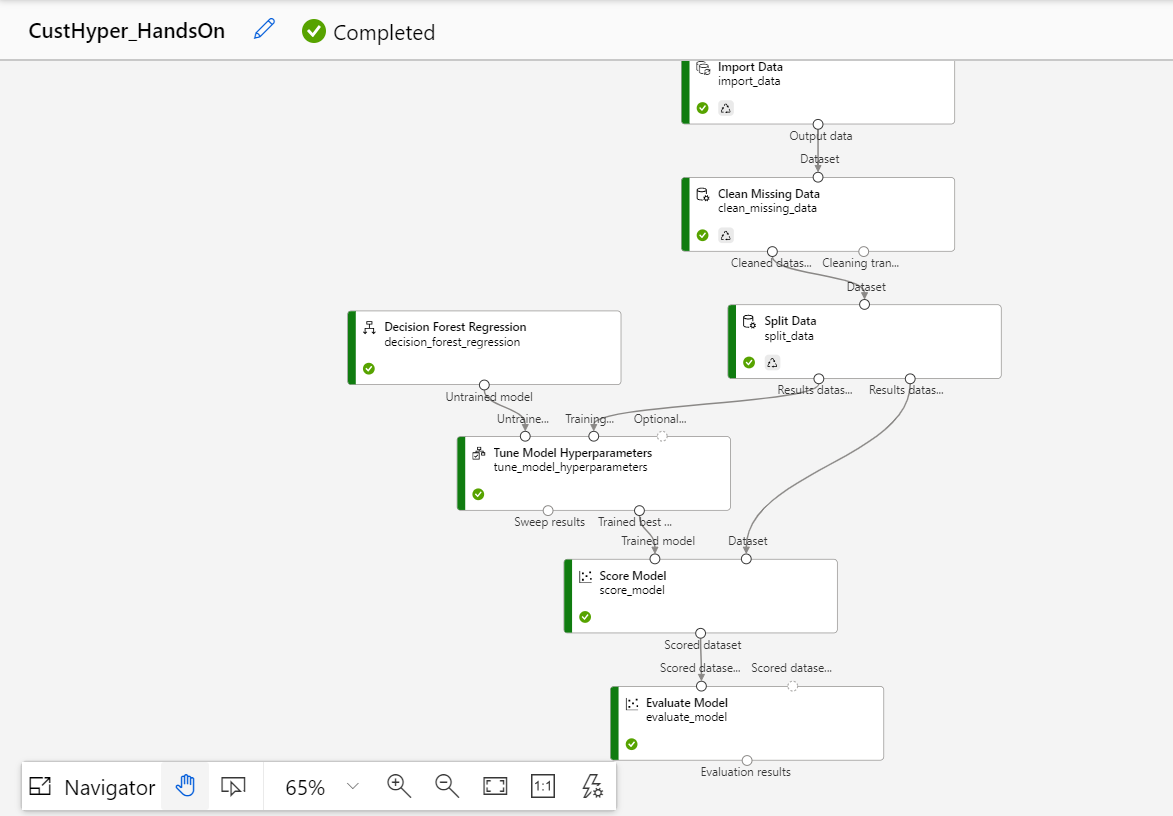
1. Hyperparameter Tuning: a. Optimize the model's performance by tuning hyperparameters. Use techniques such as grid search or random search to find the best set of hyperparameters.

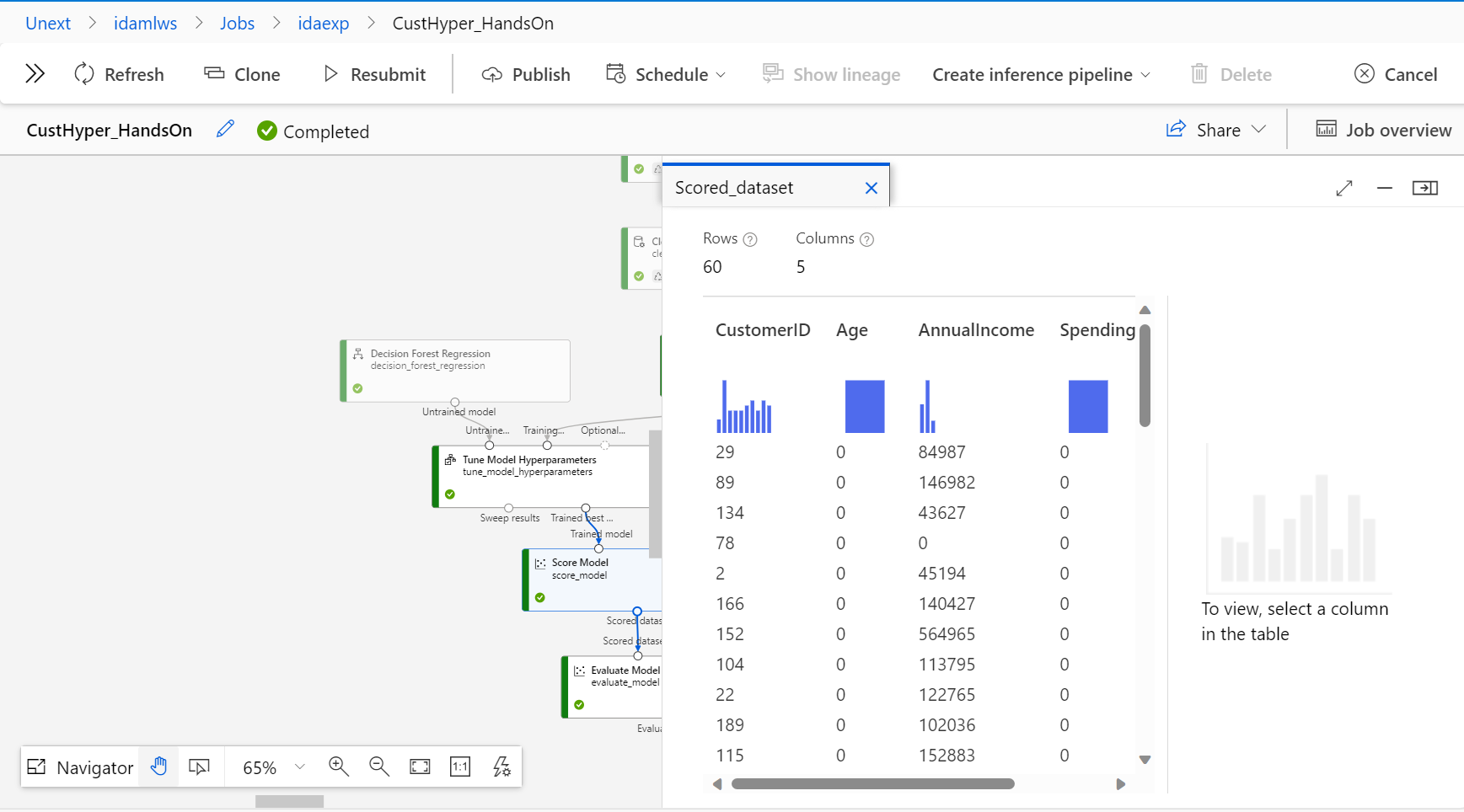


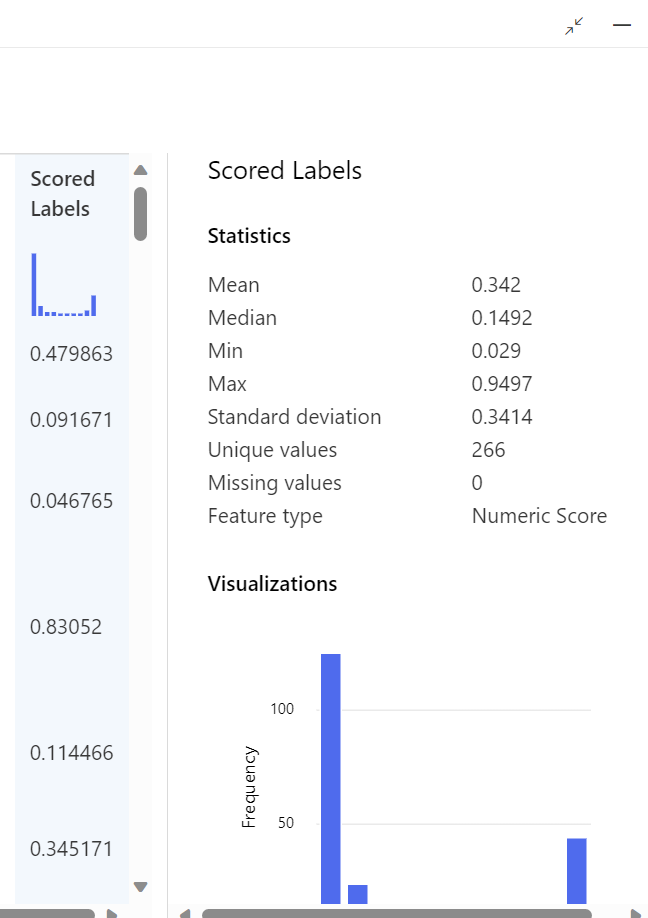


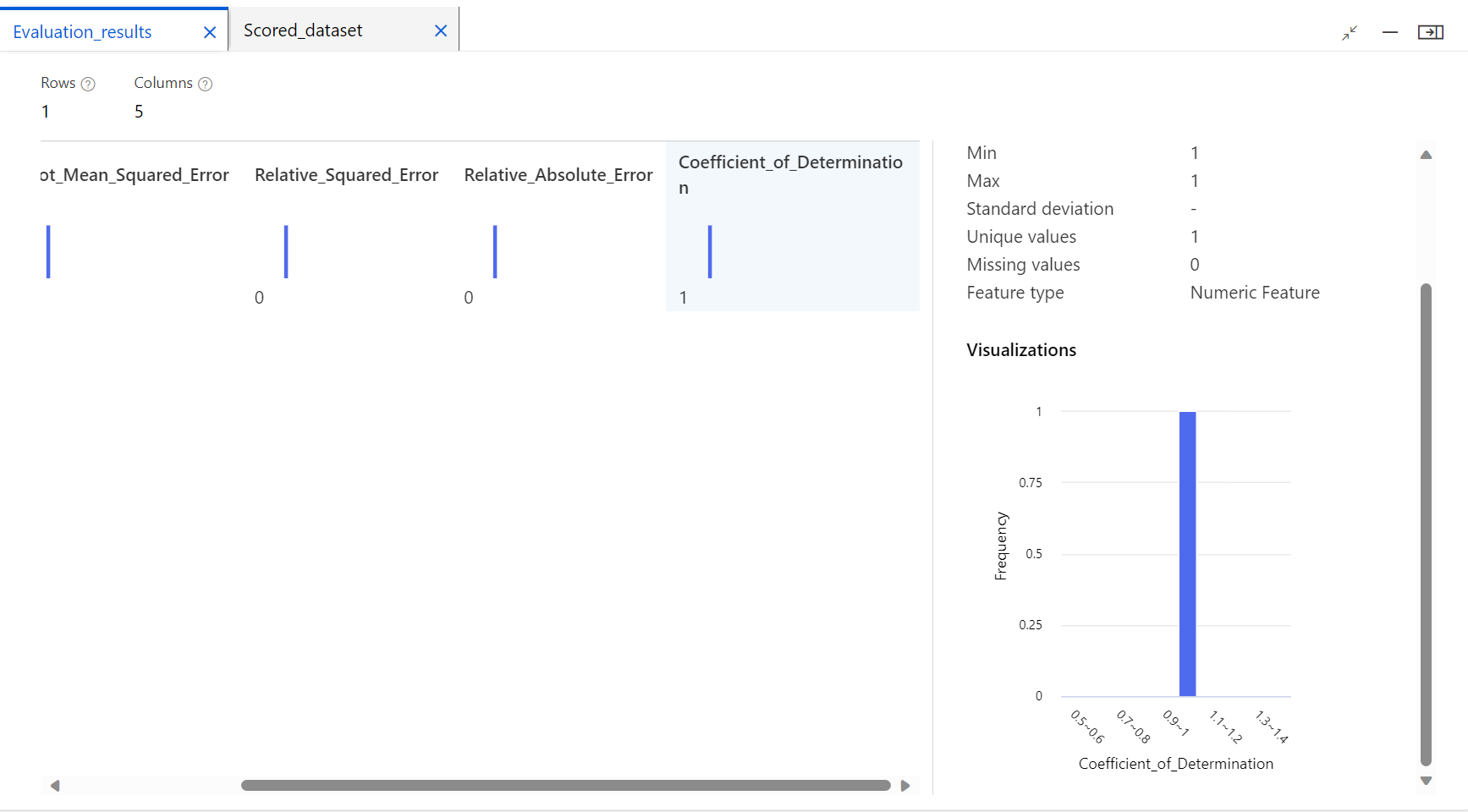
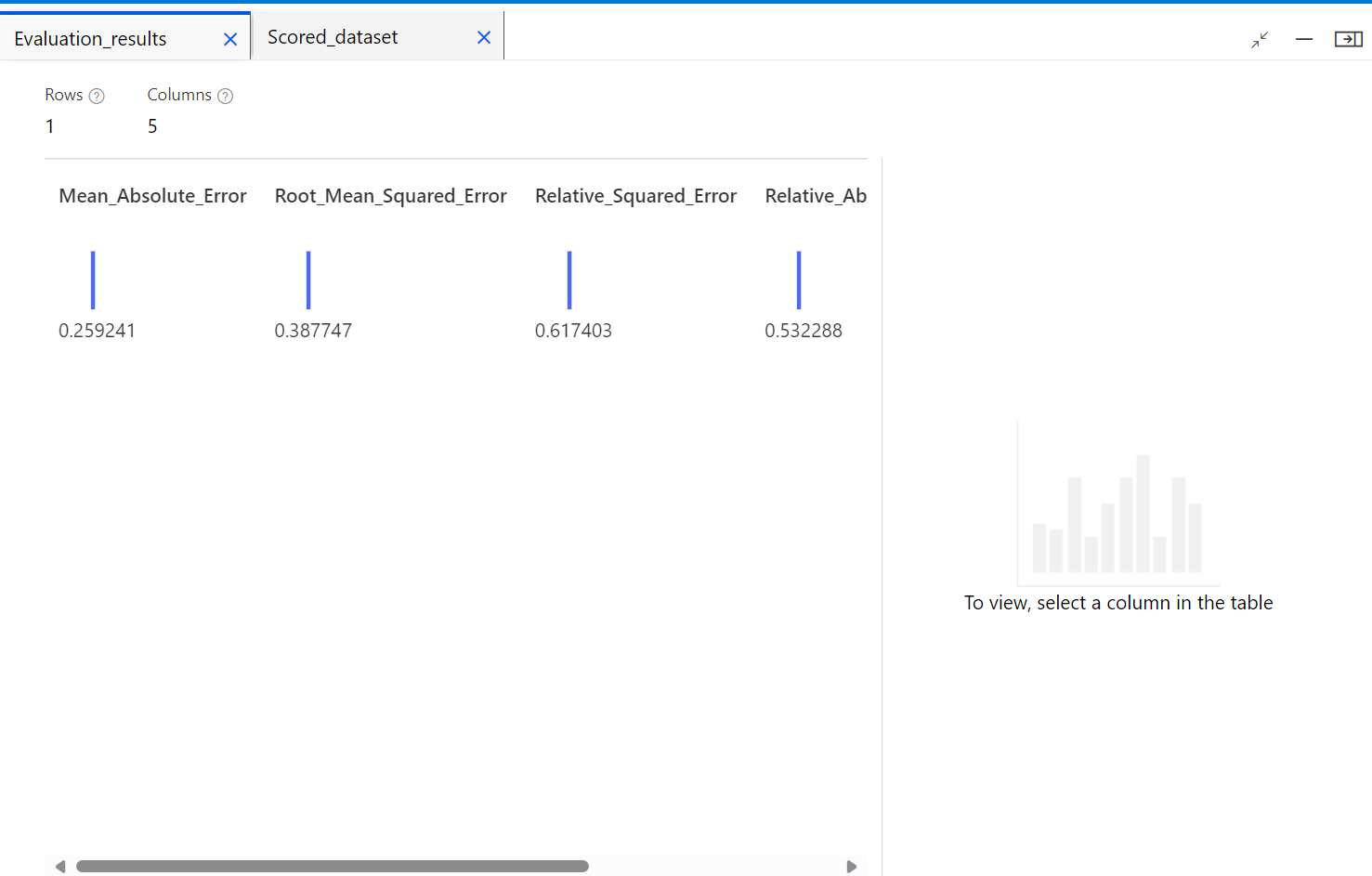
Selecting Decision Forest Regression

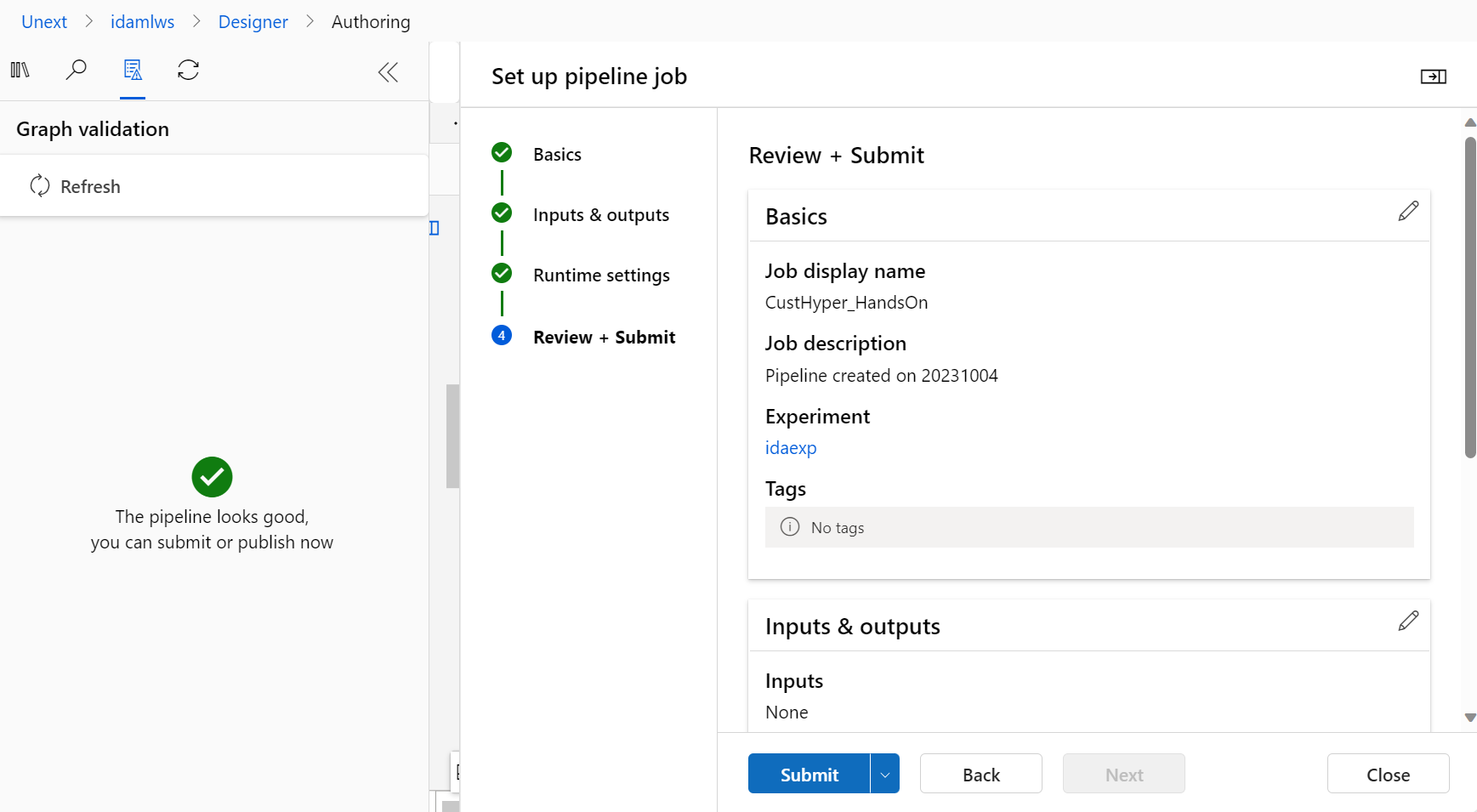












# Assessment Questions:

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.

1. **Data Collection and Ingestion:**
   * This initial step involves gathering data from various sources, such as databases or files.
   * Data is ingested into Azure Machine Learning workspace or storage, making it accessible for further processing.
2. **Data Exploration and Analysis:**
   * Data exploration involves understanding the dataset's characteristics, including its size, structure, and content.
3. **Data Cleaning:**
   * Data cleaning focuses on handling missing data, duplicate records, and outliers.
   * Techniques such as imputation, removal of duplicates, and anomaly detection are used to clean the data.
4. **Data Transformation:**
   * Data often needs to be transformed into a suitable format for machine learning. This includes encoding categorical variables, scaling numeric features, and creating new features (feature engineering).
   * Feature selection may also be performed to choose the most relevant features for model training.
5. **Data Splitting:**
   * The dataset is split into multiple subsets, typically into training, validation, and test sets.
   * The training set is used to train the model, the validation set helps tune hyperparameters, and the test set is used for model evaluation.
6. **Data Validation:**
   * Data validation ensures that the prepared dataset is consistent and meets the desired format.
   * It involves checking data types, ensuring no data leakage, and validating that transformations are applied correctly.
7. **Data Uploading to Azure Machine Learning:**
   * The prepared dataset is uploaded to Azure Machine Learning workspace or storage to make it accessible for model training.

These steps collectively ensure that the dataset is well-prepared and ready for training a machine learning model within Azure Machine Learning.

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?

Splitting the dataset into training and testing sets is a fundamental practice in machine learning for several important reasons, and it plays a critical role in model development and evaluation. Here's why it's essential:

1. **Model Assessment and Generalization:**
   * The primary goal of developing a machine learning model is to create a model that generalizes well to unseen data. In other words, the model should perform well on new, real-world data it hasn't encountered during training.
2. **Preventing Overfitting:**
   * Overfitting occurs when a model learns to memorize the training data rather than learning the underlying patterns in the data.
3. **Hyperparameter Tuning:**
   * During model development, you often need to tune hyperparameters, such as learning rates or regularization strengths, to achieve optimal model performance.
4. **Evaluating Model Metrics:**
   * Machine learning models are evaluated using various metrics like accuracy, precision, recall, F1-score, etc., depending on the problem type (classification, regression, etc.).
   * The testing set allows you to compute these metrics to quantify the model's performance and compare it to your project's goals and requirements.

Splitting the dataset into training and testing sets is crucial for evaluating a machine learning model's performance and its ability to generalize to new, unseen data.

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

In the scenario of predicting customer purchasing, a suitable machine learning algorithm is the **Decision Forest Regression**. Here's why Decision Forest is a good choice:

1. **Ensemble Learning:**
   * Decision Forest is an ensemble learning technique that combines multiple decision trees to make predictions. Each tree is trained on a different subset of the data, and the final prediction is an aggregation of the individual tree predictions.
   * Ensemble methods are known for their ability to improve prediction accuracy and generalization by reducing overfitting.
2. **Highly Flexible and Versatile:**
   * Decision Forest can handle both classification and regression tasks, making it suitable for a wide range of purchasing behavior prediction scenarios, such as binary classification (e.g., predicting whether a customer will make a purchase) or regression (e.g., predicting the amount spent by a customer).
3. **Feature Importance:**
   * Decision Forest provides a measure of feature importance, which is valuable in understanding which customer attributes or behaviors contribute the most to purchase decisions.
   * This feature can help businesses identify critical factors affecting purchasing behavior and make data-driven decisions.

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

**Hyperparameter tuning** is the process of finding the best set of hyperparameters for a machine learning model. Hyperparameters are configuration settings that are not learned from the data but are set prior to training a model. They control aspects of the learning process and model behavior. Examples of hyperparameters include learning rates, the number of hidden layers in a neural network, the depth of a decision tree, and regularization strengths.

Hyperparameter tuning is important in machine learning for several reasons:

1. **Improving Model Performance:** The choice of hyperparameters can significantly impact a model's performance. Selecting appropriate hyperparameters can lead to better predictive accuracy and generalization to unseen data.
2. **Avoiding Overfitting:** Proper hyperparameter settings can help prevent overfitting, where a model learns to memorize the training data rather than learning meaningful patterns. Overfit models perform poorly on new data.

One common technique for hyperparameter tuning is **Grid Search**:

**Grid Search:**

* Grid Search is a systematic method for hyperparameter tuning that involves specifying a range of values for each hyperparameter of interest.

Benefit of Grid Search:

1. **Exhaustive Search:** Grid Search explores all possible combinations within the specified hyperparameter ranges, ensuring that no promising set of hyperparameters is missed.

Dataset: <https://github.com/manojkumarsingh77/Shell2023/blob/main/AssessmentData/customer_data.csv>