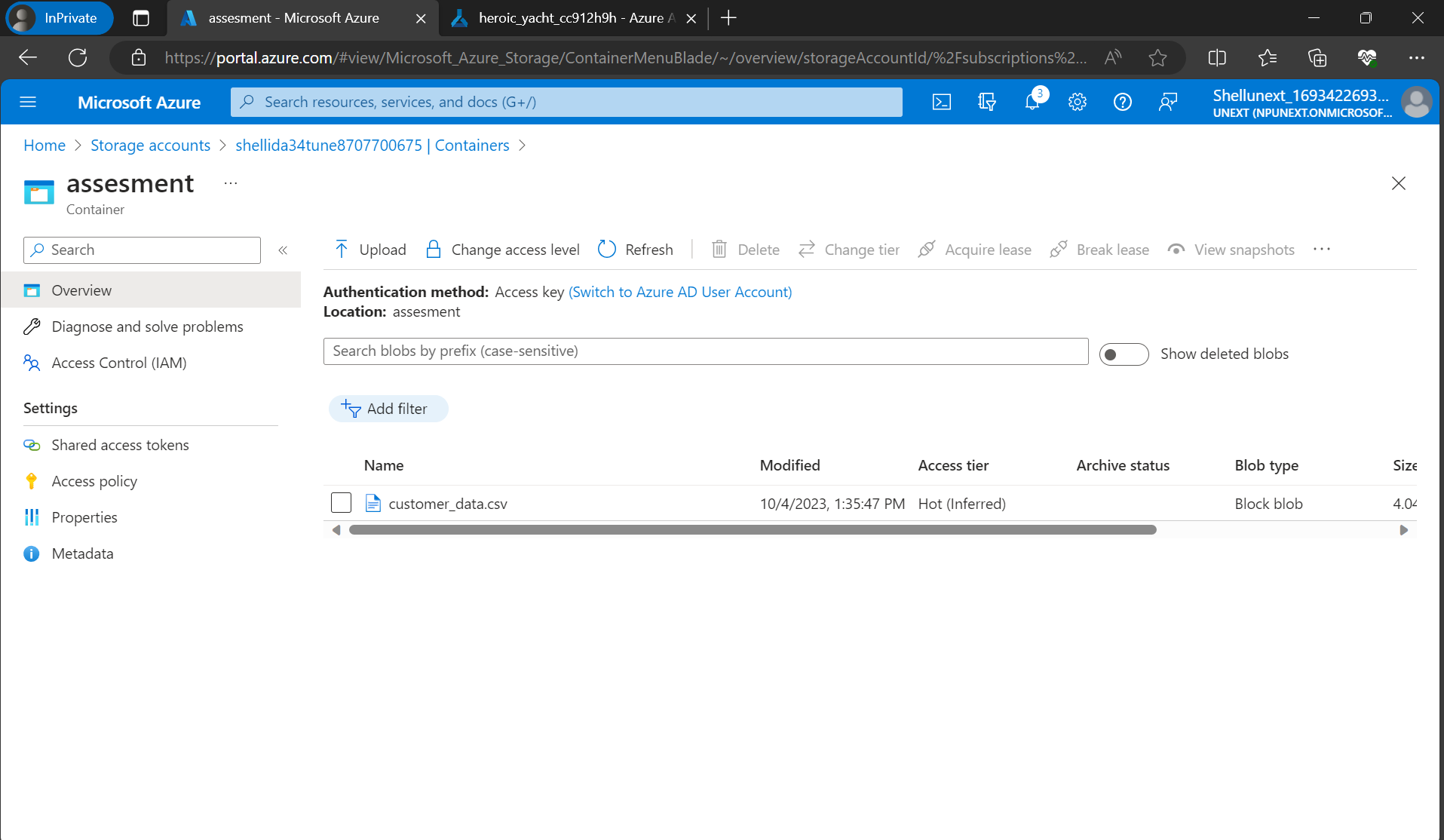
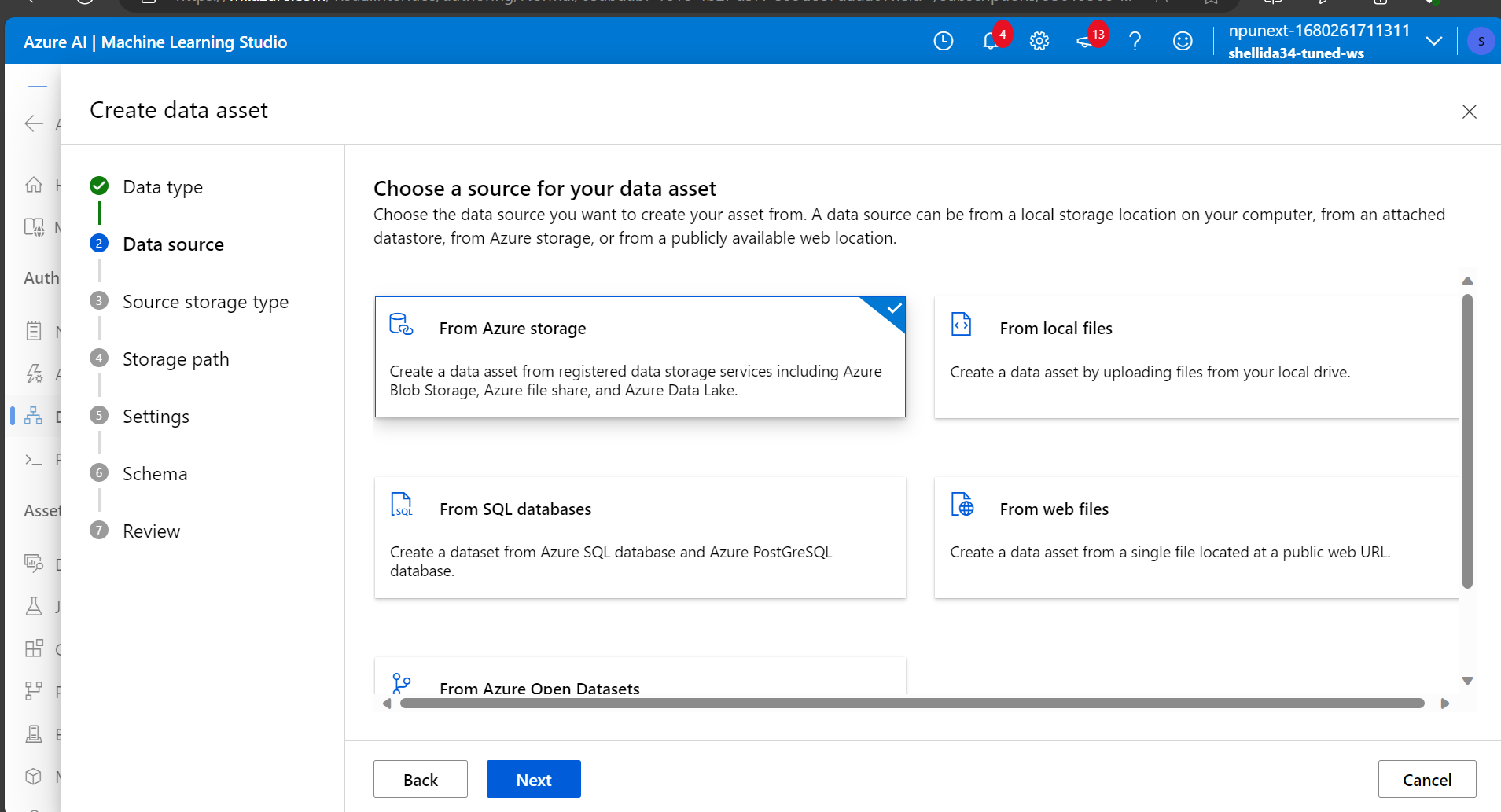
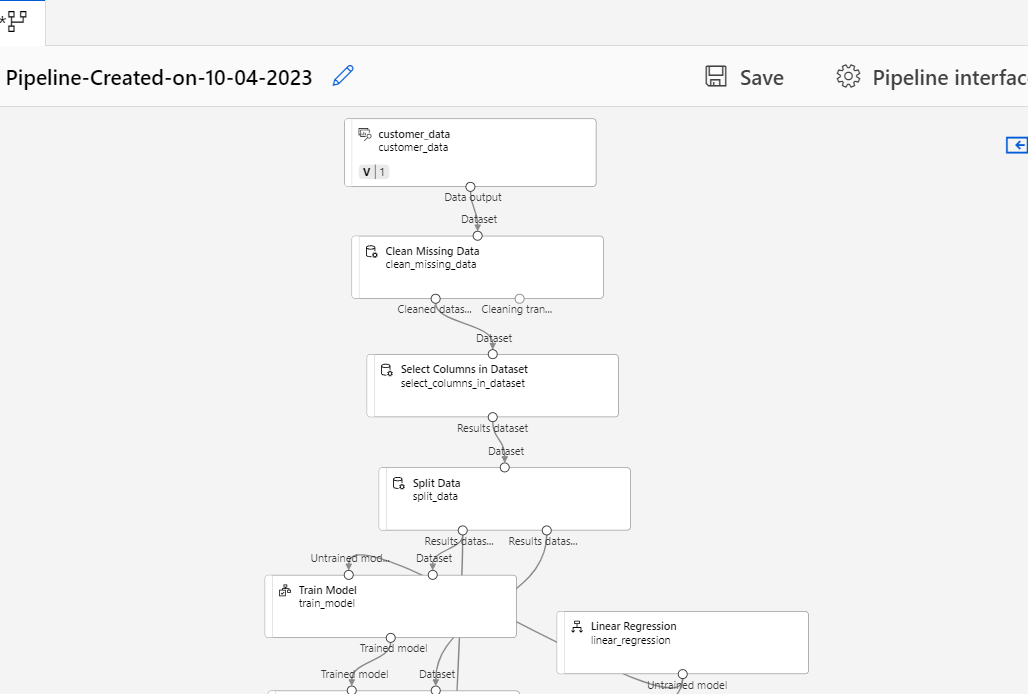
Uploading dataset in the blob storage

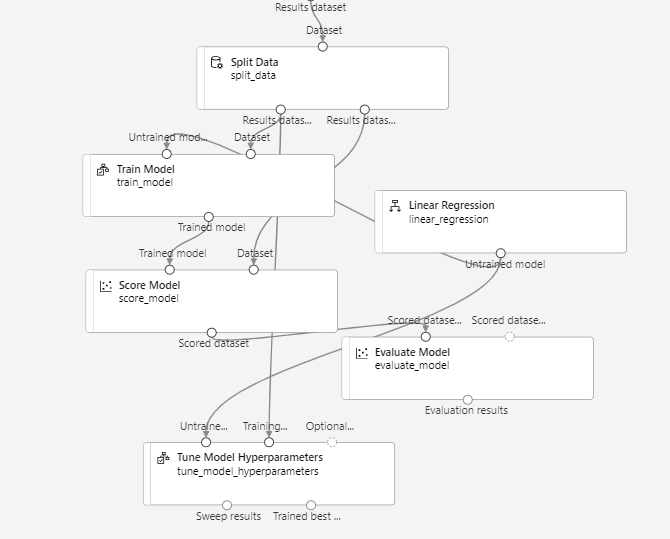


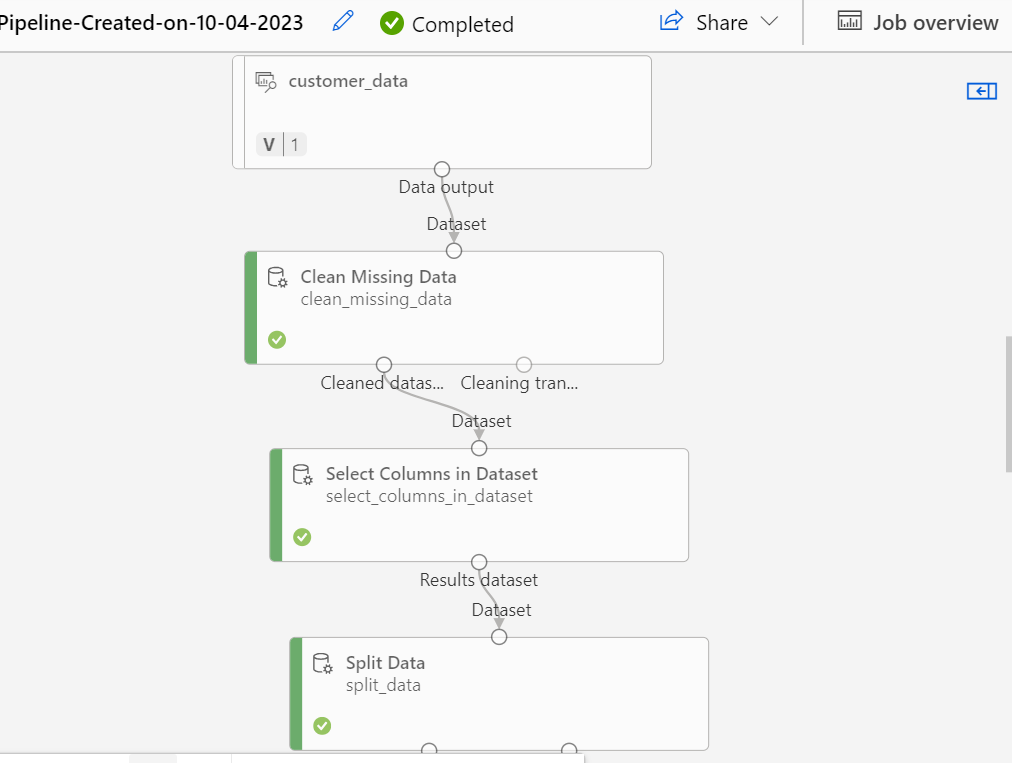
Accessing the dataset from the blob storage

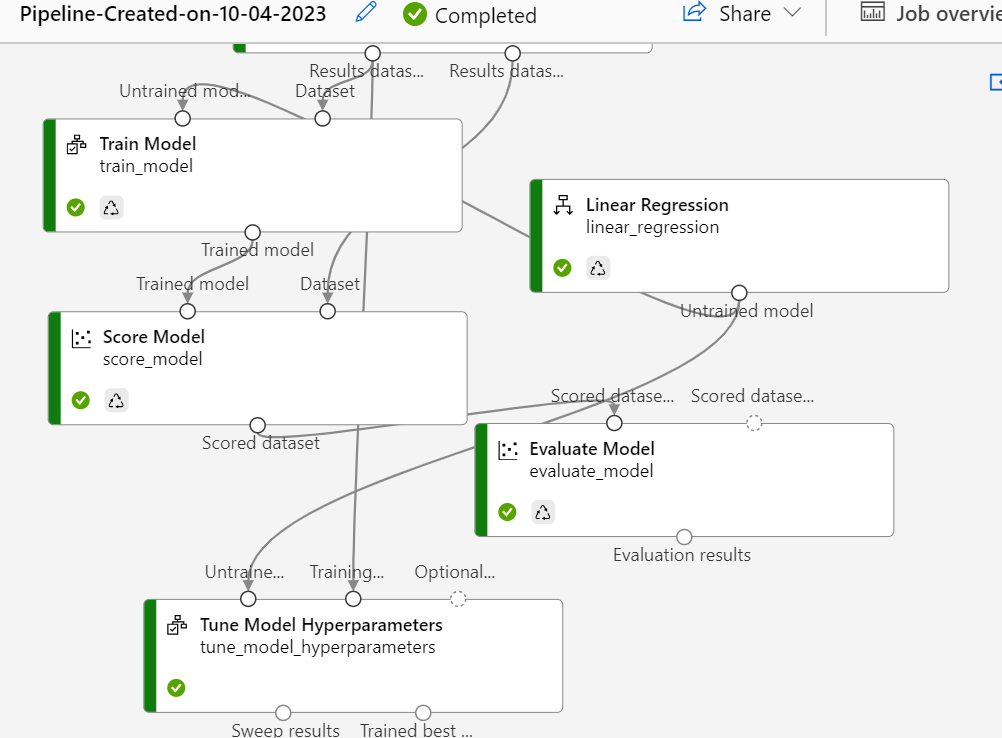


Pipeline:

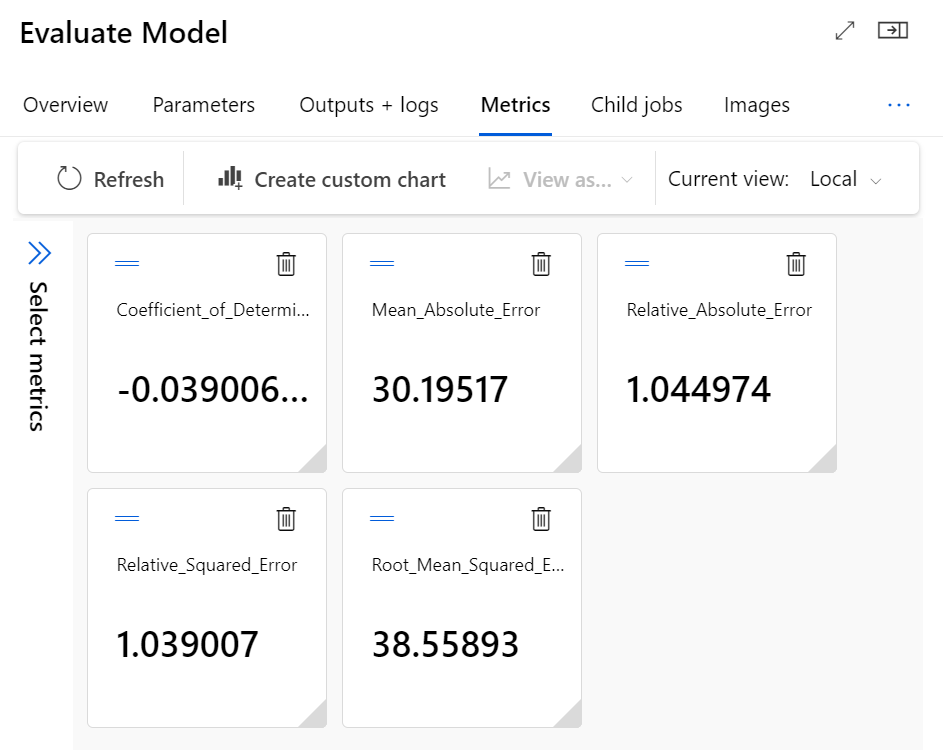








Evaluation metrics:



**ASSESSMENT QUESTION**

1. The key steps involved for training a dataset includes:
2. Handling the null values: It is the most important job prior to train the ML model. It involves replacing the null values with mean, median or mode depending on the situation. Sometime we delete the row where the field in empty. If the data of the column is having null values generally more than 60% we delete the entire column
3. Scaling the data: If the data doesn’t follow the normal curve, we have the normalize or standardize the data.
4. Selection of columns: we are required to select the particular columns which are necessary for the model training or else the model will be overfitted. Columns like ID shouldn’t be selected for the model training
5. Train-test split: the entire cleaned dataset splits into train and test data
6. The dataset splits into train and test because to train and test the model respectively. Generally we give more data to train dataset than that of test data. Once the model is trained, the data from the test dataset is used to evaluate the model. Then the predicted values and actual values are compared so that we can see how the model performs
7. It is basically a regression problem to determine the spending score. The dataset is relatively small hence I choosed linear regression for it. We can select random forest regressor too.
8. Hyperparameters are the parameters that we select manually like seed, batch size, train-test split ratio etc. As it is selected manually and randomly, our chosen value may not the help the model to perform better. As there is a range of values we can choose so to optimize it, we use hyperparameter tuning. One of the hyperparameter tuning is grid search