# TOC Report

1. Data Set Fields:

My dataset which is "Employees Promotion.csv" has 13 fields (12 features and 1 target). The fields are as follows: EmployeeID, Department, Region_Employment, Education Level, Gender, Recruitment Channel, NO_Trainings_LstYear, Age, previous_year_rating, Service Length, Awards, Avg_Training_Score, and Is Promoted. Is Promoted is the target variable and the rest of the field are features. The data types of the fields are as follows:
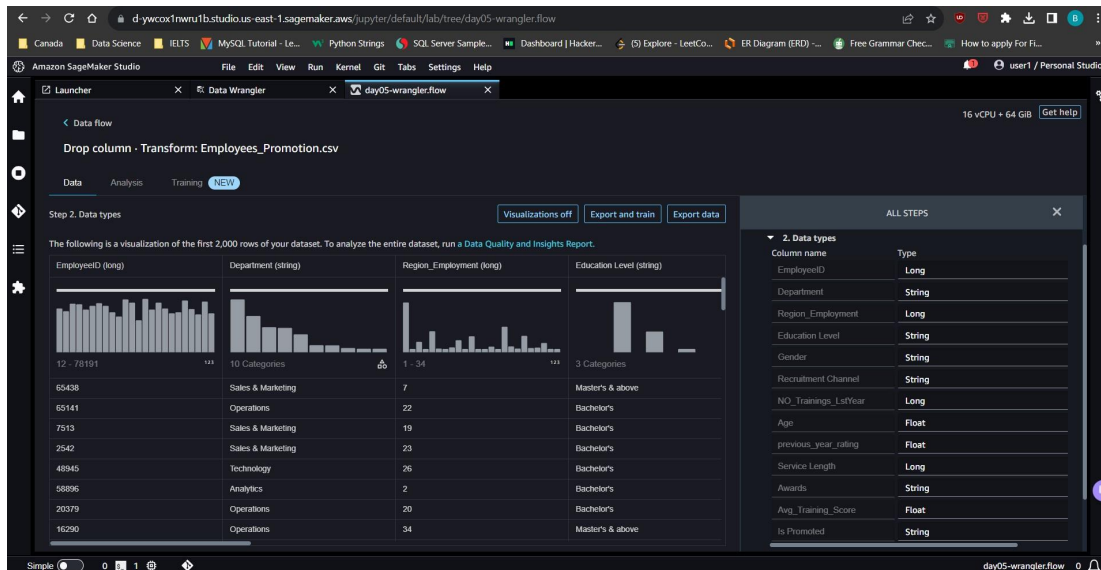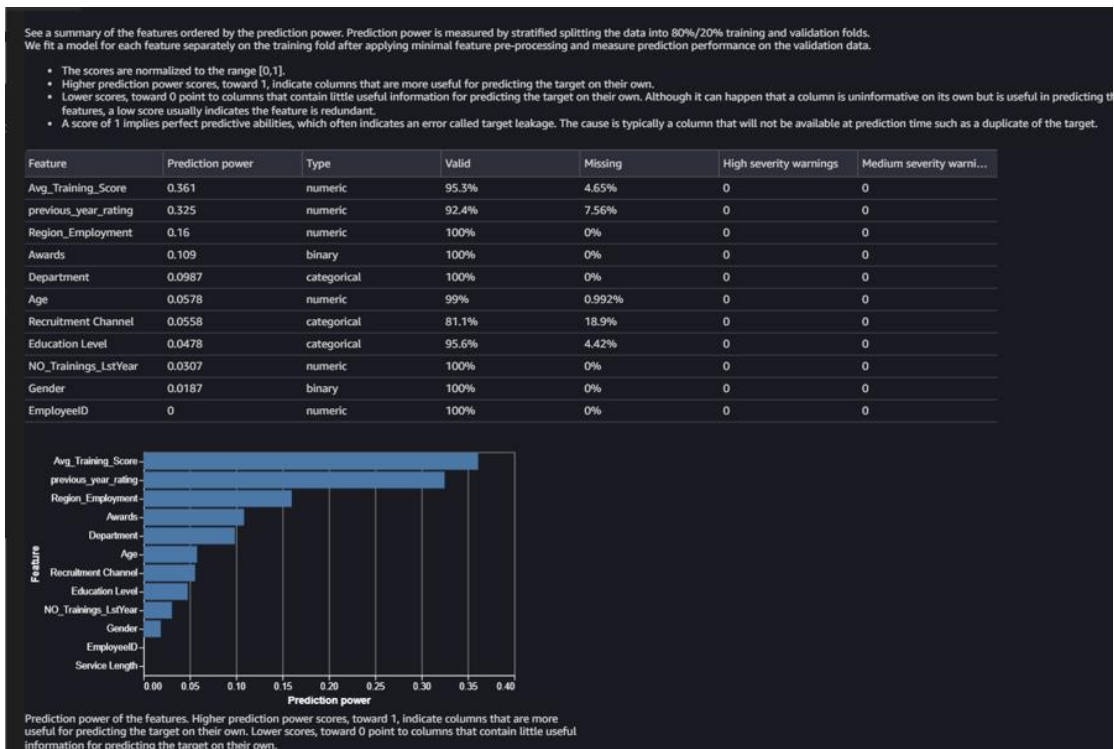

The aim is to train the model and predict whether the employees gets promoted or not based on the above mentioned fields.

2. Visualization:

After generating Data Quality Report, we got to know that we have 50,000 rows of data and there were 3.5% missing values in the dataset.

Prediction power of the features. Higher prediction power scores, toward 1, indicate columns that are more useful for predicting the target on their own. Lower scores, toward 0 point to columns that contain little useful information for predicting the target on their own.

Above feature shows the prediction power of the features on target variable, and also we can see percentage of missing values of each column. The prediction power of the features, Higher prediction power scores, toward 1, indicate columns that are more useful for predicting the target on their own. Lower, scores, toward 0 point to columns that contain little useful information for predicting the target on their own. So during transformation step, we will drop columns Service Length, EmployeeID, Gender, and Recruitment Channel. And we will fill the column which has missing values.

3. Transformations:
Dropping Columns: We drop columns Service Length, EmployeeID, Gender, and Recruitment Channel by using drop transform method.

Filling missing values: We fill missing values of column Age, previous_year_rating, and Avg_Training_Score by using Impute transform method and using Approximate Median Strategy, since these were numerical columns. Also the column Education_level is of categorical type, we use impute method to fill the missing values.
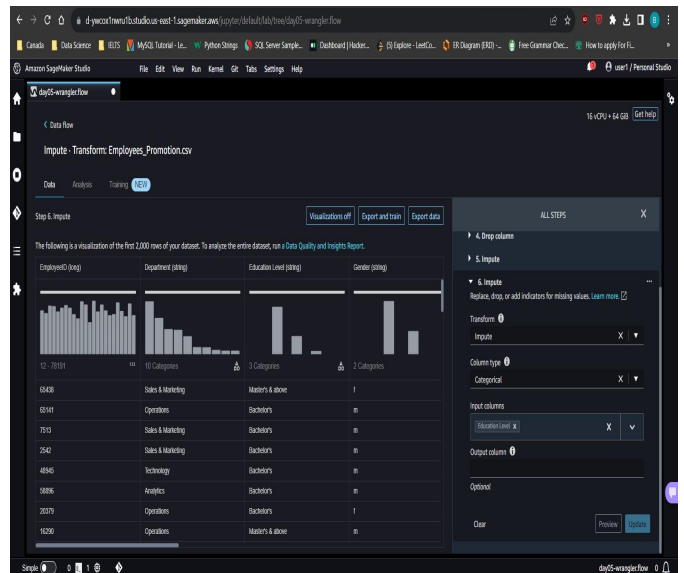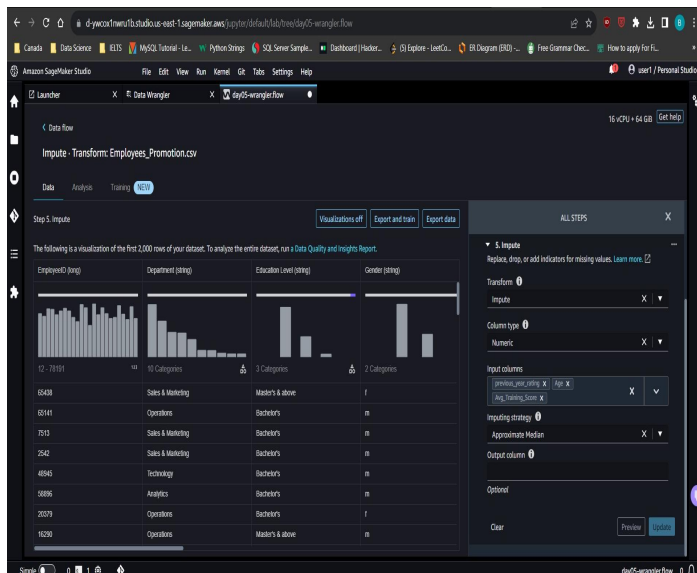
We then again perform Data Quality Report, to see for any missing values or redundant columns. Then we specify the destination of the bucket and create job to store it in the bucket.