AZURE MACHINE LEARNING ASSIGNMENT – 1

NISHA - 2211566 - AI&ML

Use the feature engineering process for building a regression model using bike rental demand data. In machine learning predictions, effective feature engineering will lead to a more accurate model. The Bike Rental dataset is based on real data from the capital Bikeshare company, which operates a bike rental network in Washington DC in the United States. The dataset contains 17,379 rows and 17 columns, each row representing the number of bike rentals within a specific hour of a day in the years 2011 or 2012. Weather conditions (such as temperature, humidity, and wind speed) were included in this raw feature set, and the dates were categorized as holiday v/s weekdays etc.

The field to predict is “cnt” which contains a count value ranging from 1 to 977, representing the number of bike rentals within a specific hour. Our main goal is to construct effective features in the training data, so we build two models using the same algorithm, but with two different datasets. Using the Split Data Module in the visual designer, we split the input data such a way that the training data contains records for the year 2011, and the testing data, records for 2012. Both datasets have the same raw data at the origin, but we added different additional features to each training set.

- Set A = weather + holiday + weekday + weekend features for the predicted day
- Set B = number of bikes that were rented in each of the previous 12 hours

We are building two training datasets by combining the feature set as follows:

- Training set 1 : feature set A only.
- Training set 2: feature sets A+B

Steps….

1. Download bike-rental-hour.csv and create a blob object and copy URL.
2. Create a new data asset and name it **BikeRentalHourly** and with Type as File and choose **From web files** and give the copied URL from previous step as **Web URL.**
3. Open Pipeline Authoring Editor and Setup Compute Target.
4. Drag and drop on the canvas, the available BikeRentalHourly dataset under the Data.
5. Drag and drop the **Edit Metadata module** (Component/Data transformation category), connect the module to the dataset, and double click on the **Edit MetaData** to open Edit Metadata details

6. Click on the Edit column and the season and weathersit columns.
7. Drag and drop the **Select Columns in Dataset** (Data transformation category), connect the module to the Edit Metadata module, and double click on the **Select Columns in the Dataset.**
8. Click Edit Columns, Configure the Select Columns in Dataset module as follows:
   Include: All columns
   Select +
   Exclude Columns names : instant, dteday, casual, registered.
   Select Save

9. Drag and drop the Execute Python Script module (Python Language Category) and connect it with the Select Columns in Dataset. (Make sure the connector is connected to the very first input of the Execute Python Script module)

   Use Python script to append a new set of features to the dataset: number of bikes that were rented in each of the previous 12 hours. Feature set B captures very recent demand for the bikes. This will be the B set in the described feature engineering approach. The script adds 12 new columns to the dataset containing the number of bikes that were rented in each of the previous 12 hours.

10. Use the **Split Data** module (Data Transformation module) and connect its input with output from the **Select Columns in Dataset** module. Use the following configurations:

    Splitting mode : Relative Expression
    Relational Expression: \"y" == 0

11. Select the **Split Data** module block and use the menu buttons to Copy and Paste it on the Canvas. Connect the second one to the output of the Python Script execution step, which is the featured B set.
12. Drag and drop **Select columns in Dataset module** (Data transformation category), create four identical modules to exclude the yr column from all the outputs.
13. Drag and drop Boosted Decision Tree Regression module (Machine Learning Algorithms, Regression category)
14. Drag and drop **Train model** module (Model training category) and enter the cnt column in the Label column field.
15. Link the Boosted Decision Tree Regression module as the first input and the training dataset as the second input to the Train Model module.
16. Use two Score Model modules (Model Scoring and Evaluation category) and link it to the two Score Model modules.
17. Drag and drop the Evaluate Model module (Model Scoring and evaluation category) and link it to the two Score Model modules.

18. Select Submit to open the Setup pipeline run editor. In the Setup pipeline run editor, select Experiment, Create new and provide a New experiment name.

Results : -

Observe the values for the Mean_Absolute_Error, RMSE and R- Squared and other metrics. The first values corresponds to the model trained on feature set A. The second values corresponds to the model trained on the feature sets A + B.

Using simple feature engineering to derive new features from the existing data set allows the model to better understand the dynamics of the data and hence, produce a better prediction.

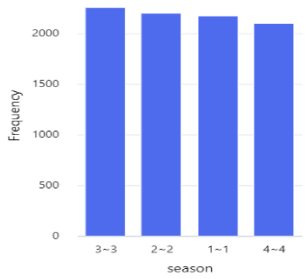
SCORED DATASETS- Exploring the Data sets.

season

Statistics

Mean	-
Median	-
Min	-
Max	-
Standard deviation	-
Unique values	4
Missing values	0
Feature type	Categorical Feature

Visualizations

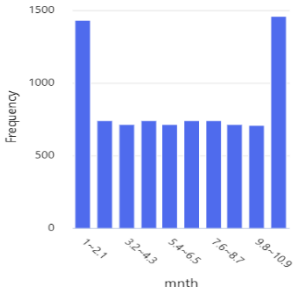


mnth

Statistics

Mean	6.5019
Median	7
Min	1
Max	12
Standard deviation	3.4491
Unique values	12
Missing values	0
Feature type	Numeric Feature

Visualizations

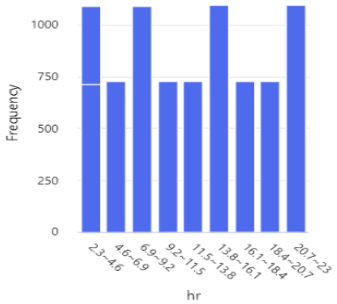


hr

Statistics

Mean	11.5202
Median	12
Min	0
Max	23
Standard deviation	6.9212
Unique values	24
Missing values	0
Feature type	Numeric Feature

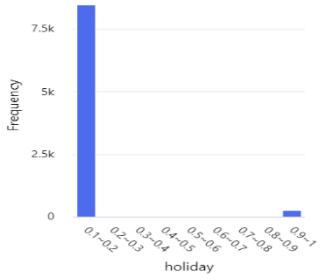
Visualizations



holiday

Statistics

Mean	0.0299
Median	0
Min	0
Max	1
Standard deviation	0.1703
Unique values	2
Missing values	0
Feature type	Numeric Feature

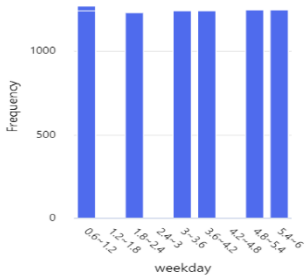


weekday

Statistics

Mean	2.9947
Median	3
Min	0
Max	6
Standard deviation	2.0053
Unique values	7
Missing values	0
Feature type	Numeric Feature

Visualizations

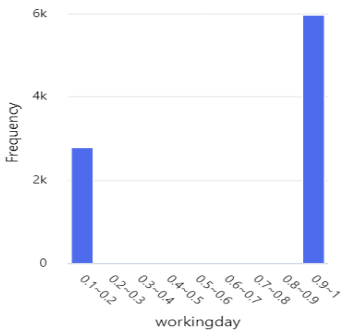


workingday

Statistics

Mean	0.6817
Median	1
Min	0
Max	1
Standard deviation	0.4658
Unique values	2
Missing values	0
Feature type	Numeric Feature

Visualizations

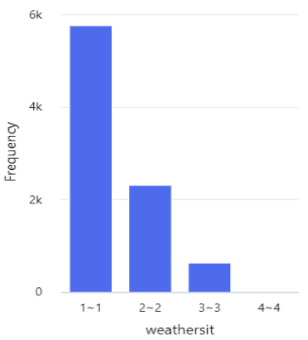


weathersit

Statistics

Mean	-
Median	-
Min	-
Max	-
Standard deviation	-
Unique values	4
Missing values	0
Feature type	Categorical Feature

Visualizations

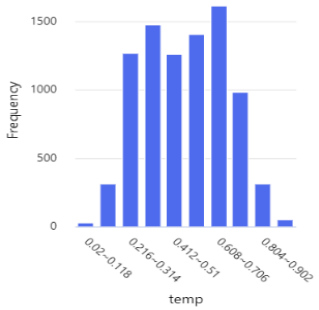


temp

Statistics

Mean	0.5048
Median	0.52
Min	0.02
Max	1
Standard deviation	0.1868
Unique values	50
Missing values	0
Feature type	Numeric Feature

Visualizations

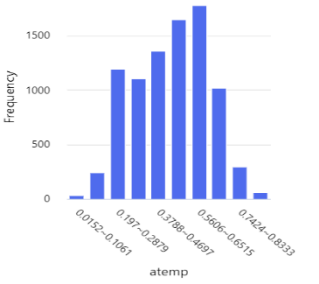


atemp

Statistics

Mean	0.4825
Median	0.4848
Min	0.0152
Max	0.9242
Standard deviation	0.1666
Unique values	61
Missing values	0
Feature type	Numeric Feature

Visualizations

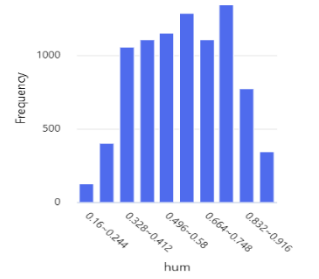


hum

Statistics

Mean	0.6112
Median	0.61
Min	0.16
Max	1
Standard deviation	0.1882
Unique values	81
Missing values	0
Feature type	Numeric Feature

Visualizations

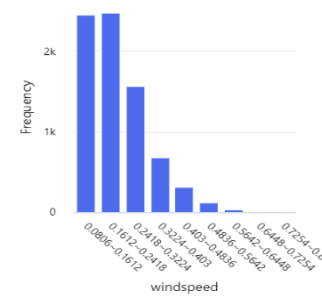


windspeed

Statistics

Mean	0.189
Median	0.1642
Min	0
Max	0.806
Standard deviation	0.1215
Unique values	26
Missing values	0
Feature type	Numeric Feature

Visualizations

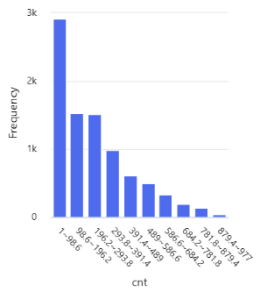


cnt

Statistics

Mean	234.6664
Median	191
Min	1
Max	977
Standard deviation	208.9109
Unique values	866
Missing values	0
Feature type	Numeric Label

Visualizations

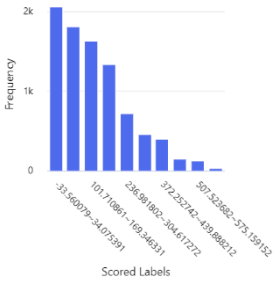


Scored Labels

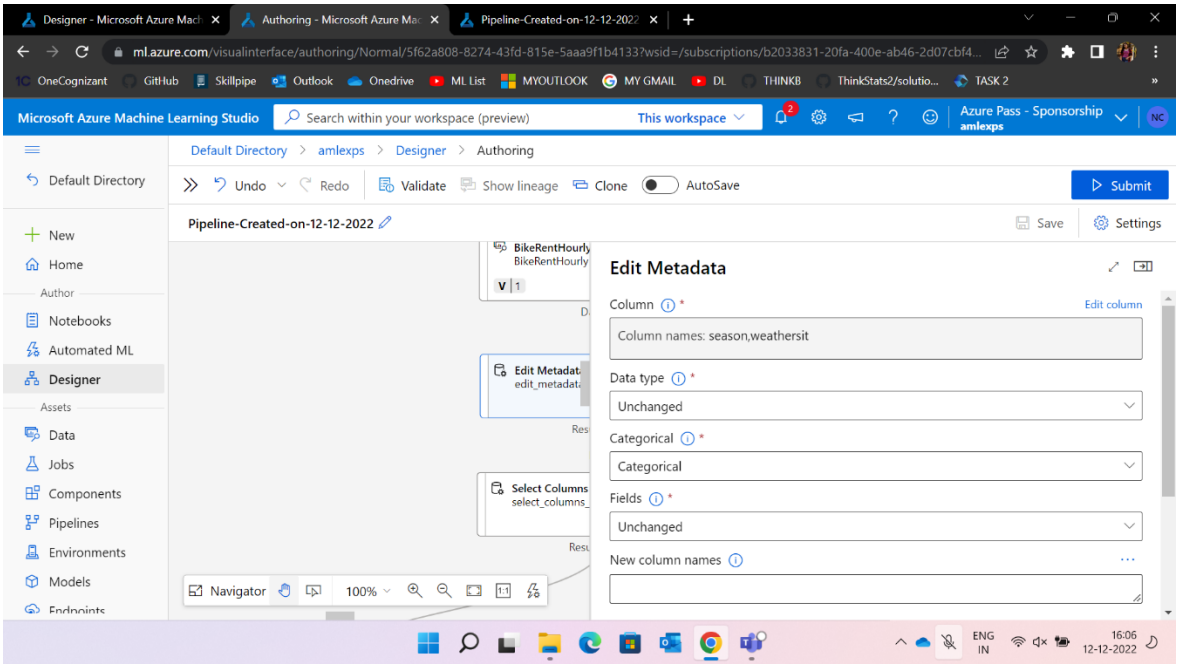
Statistics

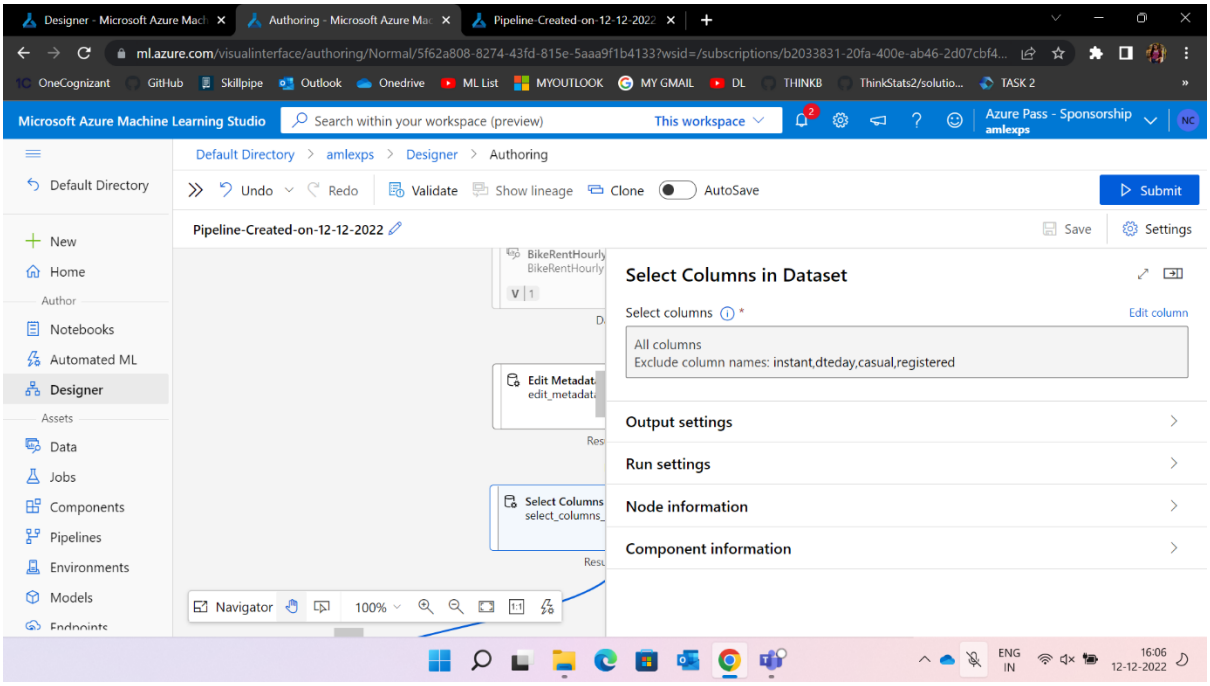
Mean	150.1674
Median	121.5931
Min	-33.5601
Max	642.7946
Standard deviation	131.5421
Unique values	8641
Missing values	0
Feature type	Numeric Score

Visualizations

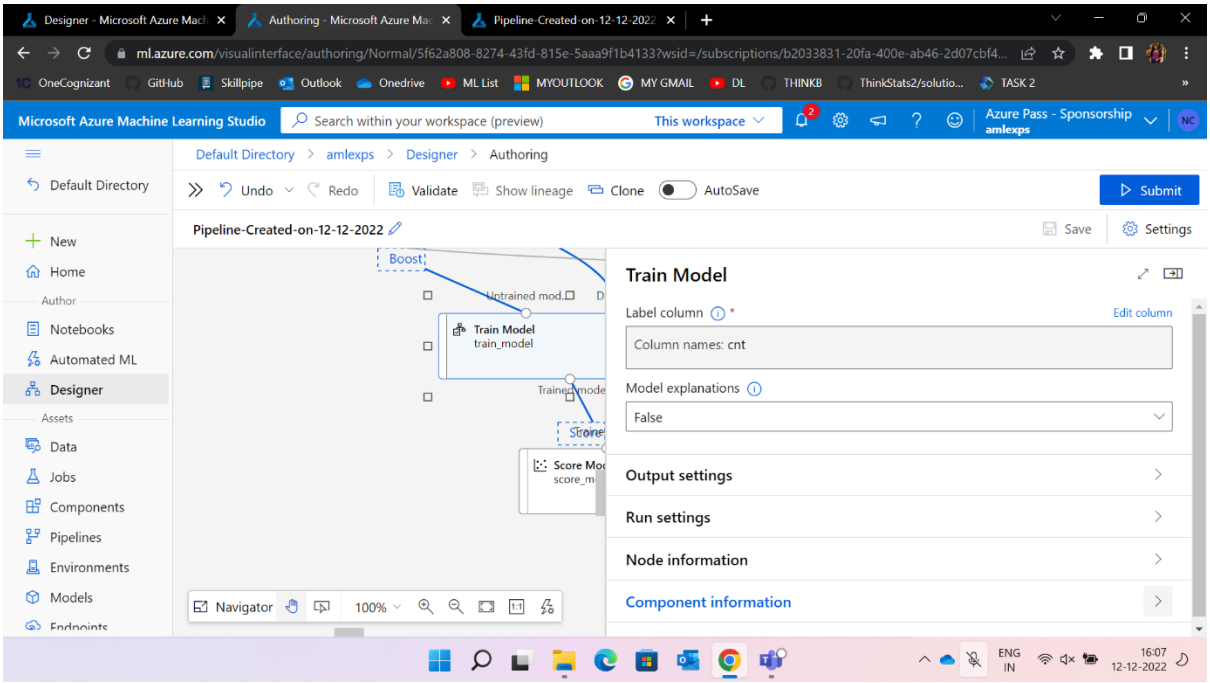


EDIT METADATA

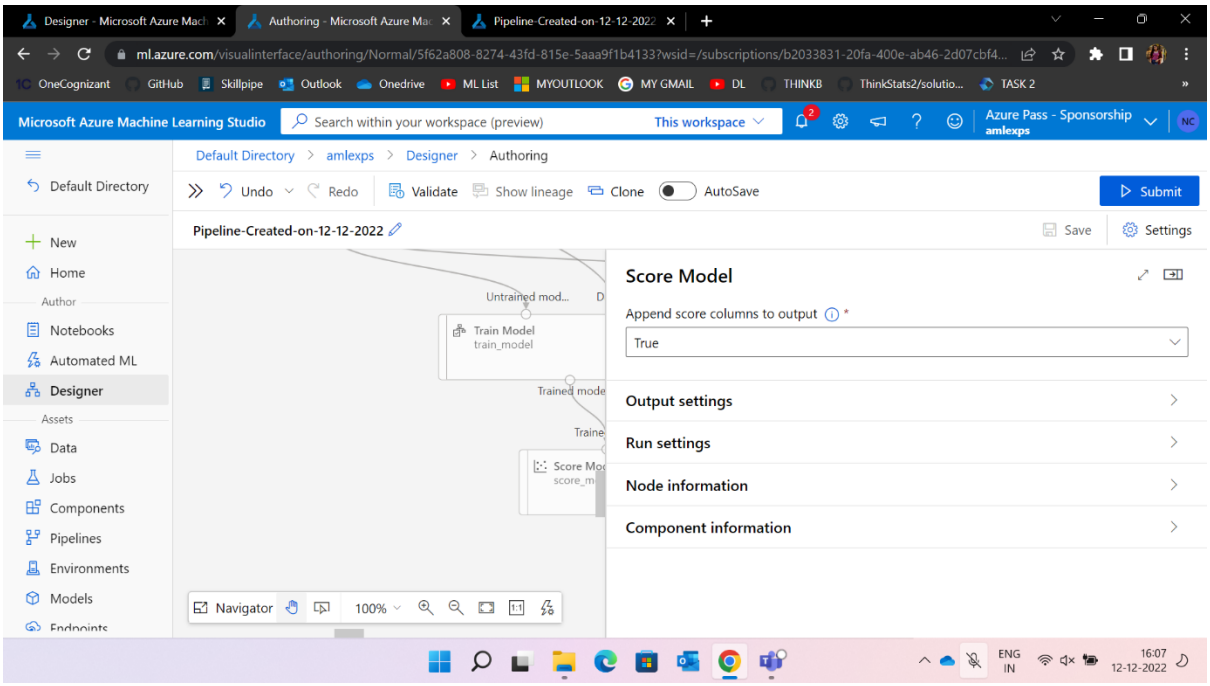



Microsoft Azure Machine Learning Studio
Search within your workspace (preview)
This workspace
Azure Pass - Sponsorship
amlexps
Default Directory
amlexps
Designer
Authoring
Undo
Redo
Validate
Show lineage
Clone
AutoSave
Submit
Pipeline-Created-on-12-12-2022
Save
Settings
New
Home
Author
Notebooks
Automated ML
Designer
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Select Columns in Dataset
Select columns
Edit column
All columns
Exclude column names: instant,dteday,casual,registered
Output settings
Run settings
Node information
Component information
Navigator
100%
ENG
IN
16:06
12-12-2022

TRAIN MODEL


Microsoft Azure Machine Learning Studio
Search within your workspace (preview)
This workspace
Azure Pass - Sponsorship
amlexps
Default Directory
amlexps
Designer
Authoring
Undo
Redo
Validate
Show lineage
Clone
AutoSave
Submit
Pipeline-Created-on-12-12-2022
Save
Settings
New
Home
Author
Notebooks
Automated ML
Designer
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Train Model
Label column
Edit column
Column names: cnt
Model explanations
False
Output settings
Run settings
Node information
Component information
Navigator
100%
ENG
IN
16:07
12-12-2022

SCORE MODEL



EVALUATION RESULT –

Coefficient_of_Determination

Statistics

Mean	0.7325
Median	0.7325
Min	0.6428
Max	0.8222
Standard deviation	0.1268
Unique values	2
Missing values	0
Feature type	Numeric Feature

Relative_Absolute_Error

Statistics

Mean	0.4261
Median	0.4261
Min	0.3173
Max	0.5349
Standard deviation	0.1539
Unique values	2
Missing values	0
Feature type	Numeric Feature

Relative_Squared_Error

Statistics

Mean	0.2675
Median	0.2675
Min	0.1778
Max	0.3572
Standard deviation	0.1268
Unique values	2
Missing values	0
Feature type	Numeric Feature

Root_Mean_Squared_Error

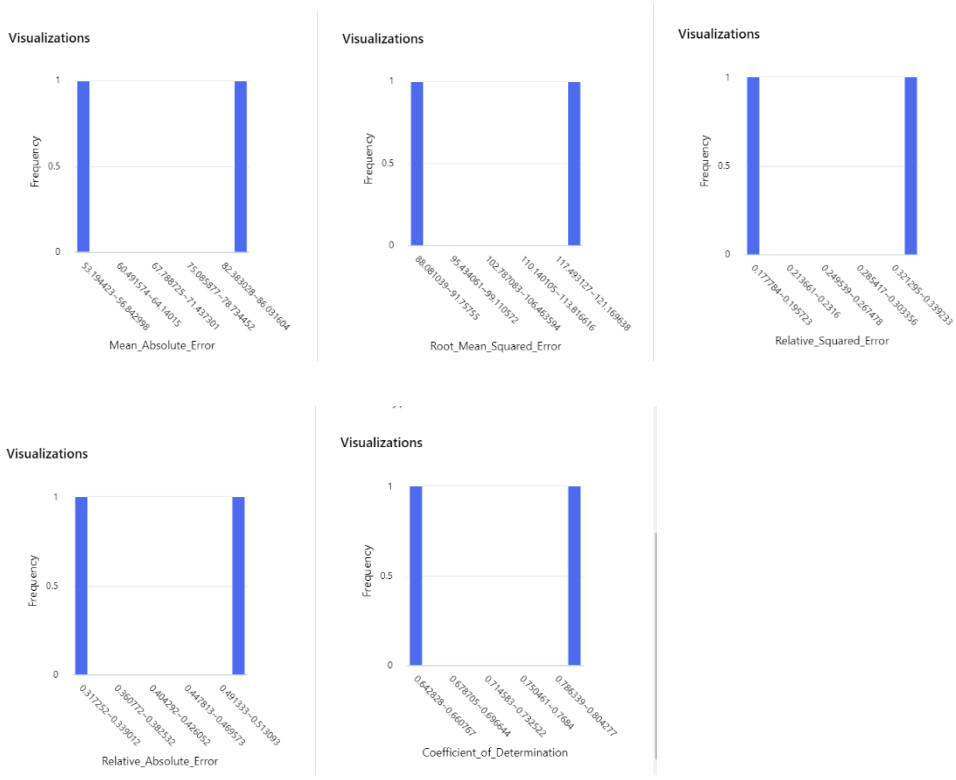
Statistics

Mean	106.4636
Median	106.4636
Min	88.081
Max	124.8461
Standard deviation	25.9969
Unique values	2
Missing values	0
Feature type	Numeric Feature

Mean_Absolute_Error

Statistics

Mean	71.4373
Median	71.4373
Min	53.1944
Max	89.6802
Standard deviation	25.7993
Unique values	2
Missing values	0
Feature type	Numeric Feature



RUNNED MODEL :- Successfully runned!