Course Project
DeVry University
College of Engineering and Information Sciences

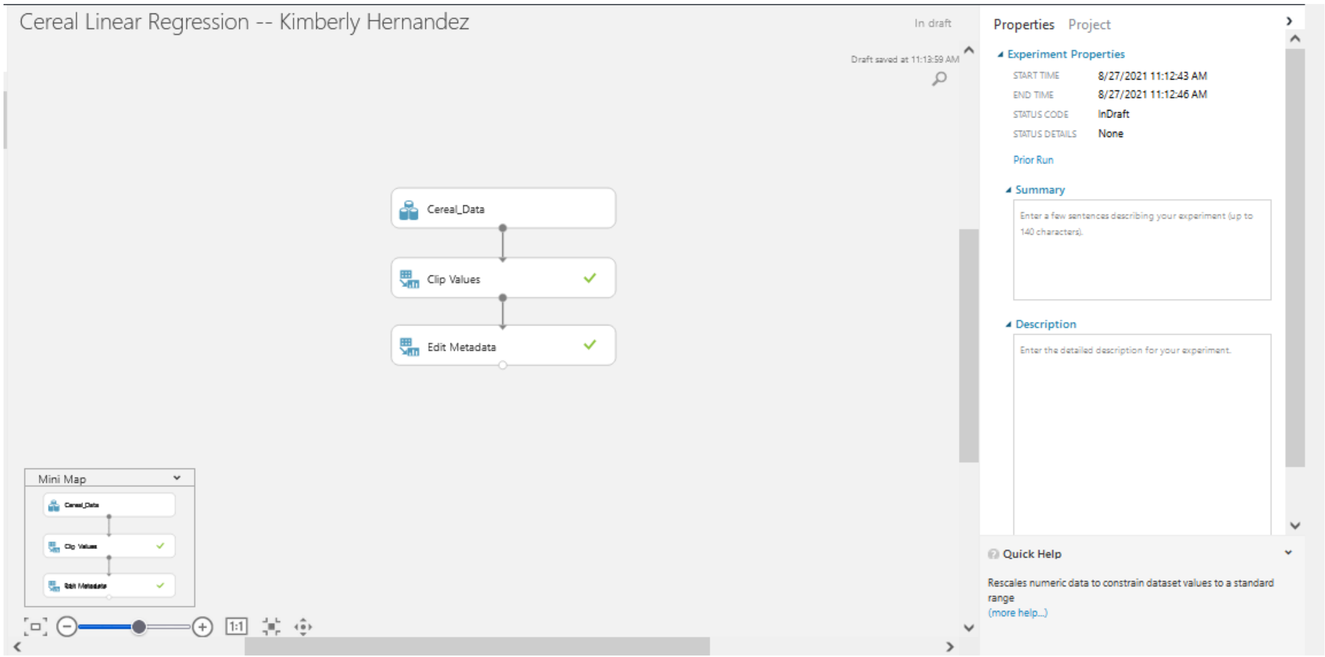
Course Number: CEIS312
Module 8 Report
Kimberly Hernandez

Introduction to the problem

A cereal data set from Kaggle.com is given in order to determine predictions of consumer report ratings on different cereal. This dataset will be added to the Azure Machine learning environment and linear regression will be performed on it. Several steps of the machine learning process will be taken to create a model accurate enough for future predictions of ratings on cereal. Some of the project's focus will be on data preparation and the iteration process. The model must reach at least a 0.7 Coefficient of determination or R² and to consider error and bias of the model the Relative Squared Error will be taken into consideration as well.

Uploading dataset

Cereal Dataset added.



START TIME 8/27/2021 11:12:43 AM
END TIME 8/27/2021 11:12:46 AM
STATUS CODE InDraft
STATUS DETAILS None

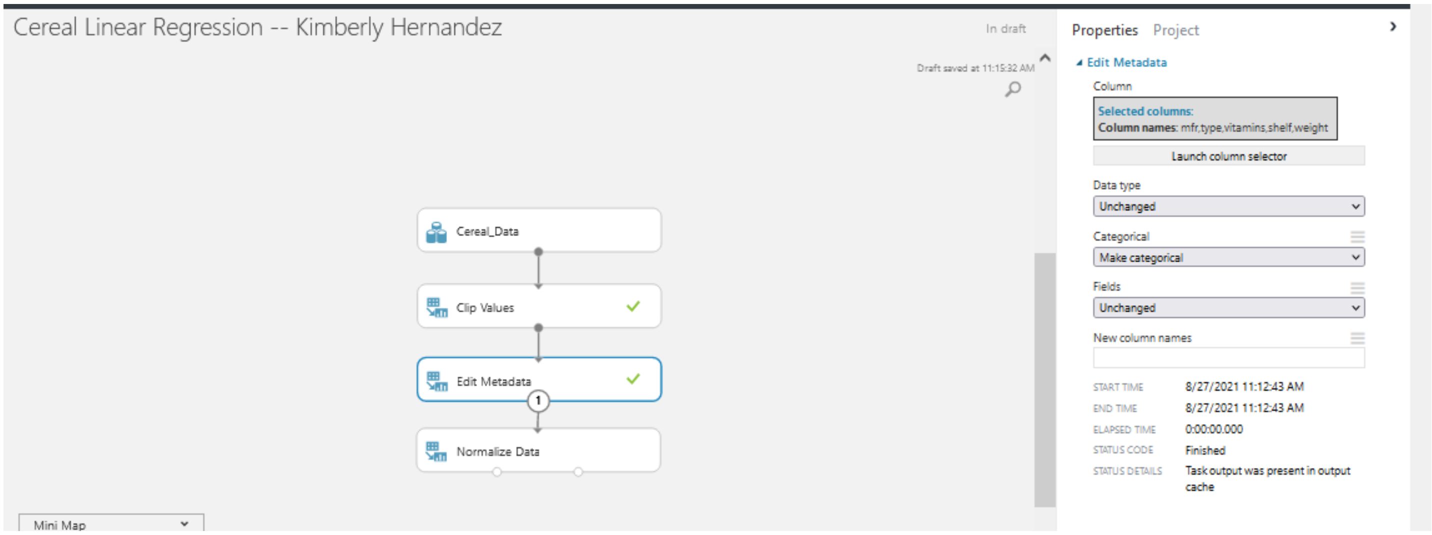
Enter a few sentences describing your experiment (up to 140 characters).

Enter the detailed description for your experiment.

Rescales numeric data to constrain dataset values to a standard range

Data preparation (normalization)

‘Edit Metadata’ module added to make category fields ‘Categorical’.



Edit Metadata

Column

Selected columns:
Column names: mfr,type,vitamins,shelf,weight

Launch column selector

Data type: Unchanged

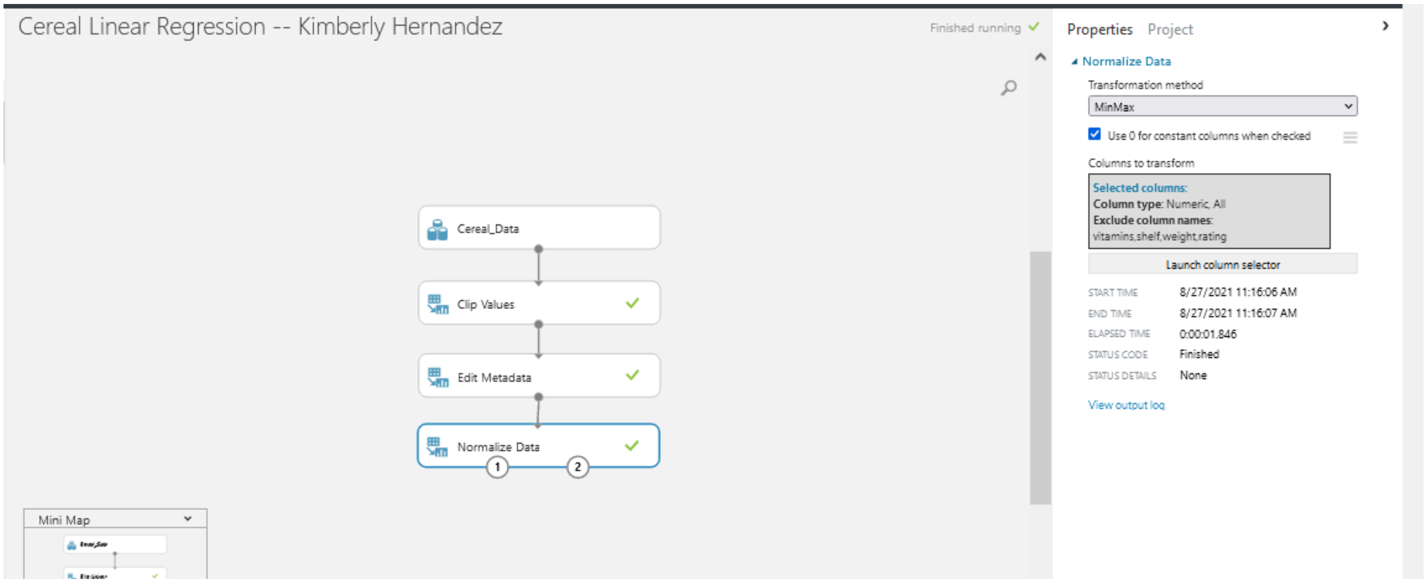
Categorical: Make categorical

Fields: Unchanged

New column names:

START TIME 8/27/2021 11:12:43 AM
END TIME 8/27/2021 11:12:43 AM
ELAPSED TIME 0:00:00.000
STATUS CODE Finished
STATUS DETAILS Task output was present in output cache

‘Normalized Data’ module added.



Normalize Data

Transformation method: MinMax

☑ Use 0 for constant columns when checked

Columns to transform

Selected columns:
Column type: Numeric, All
Exclude column names: vitamins,shelf,weight,rating

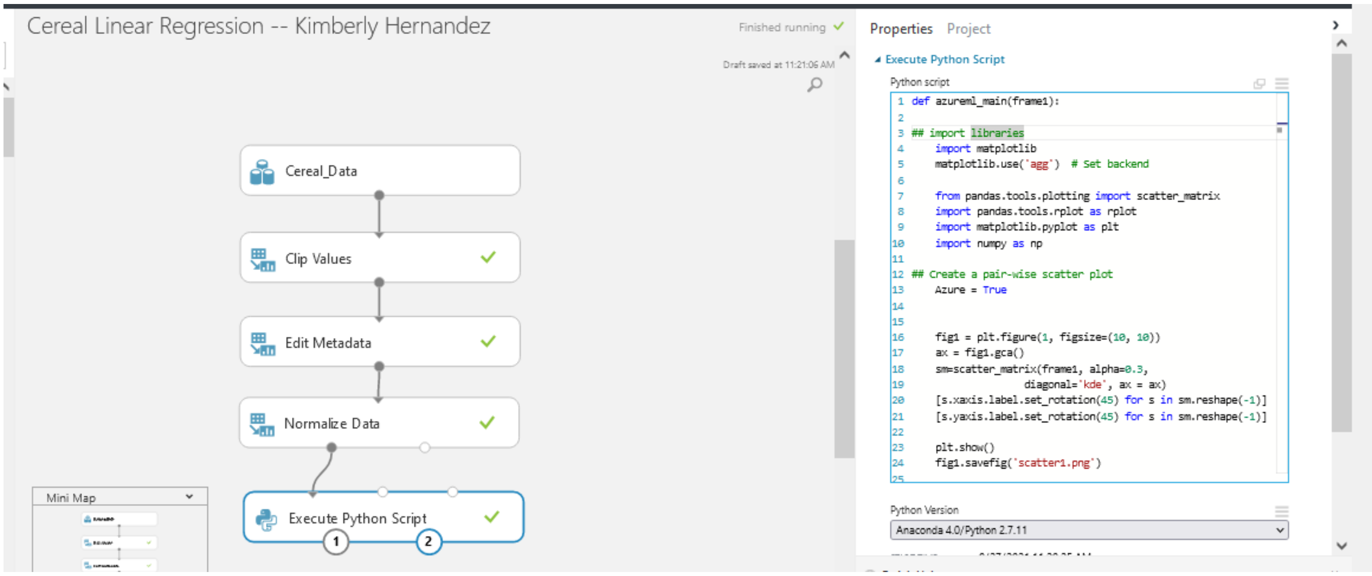
Launch column selector

START TIME 8/27/2021 11:16:06 AM
END TIME 8/27/2021 11:16:07 AM
ELAPSED TIME 0:00:01.846
STATUS CODE Finished
STATUS DETAILS None

View output log

Data Visualization (python script)

‘Execute Python Script’ module added.

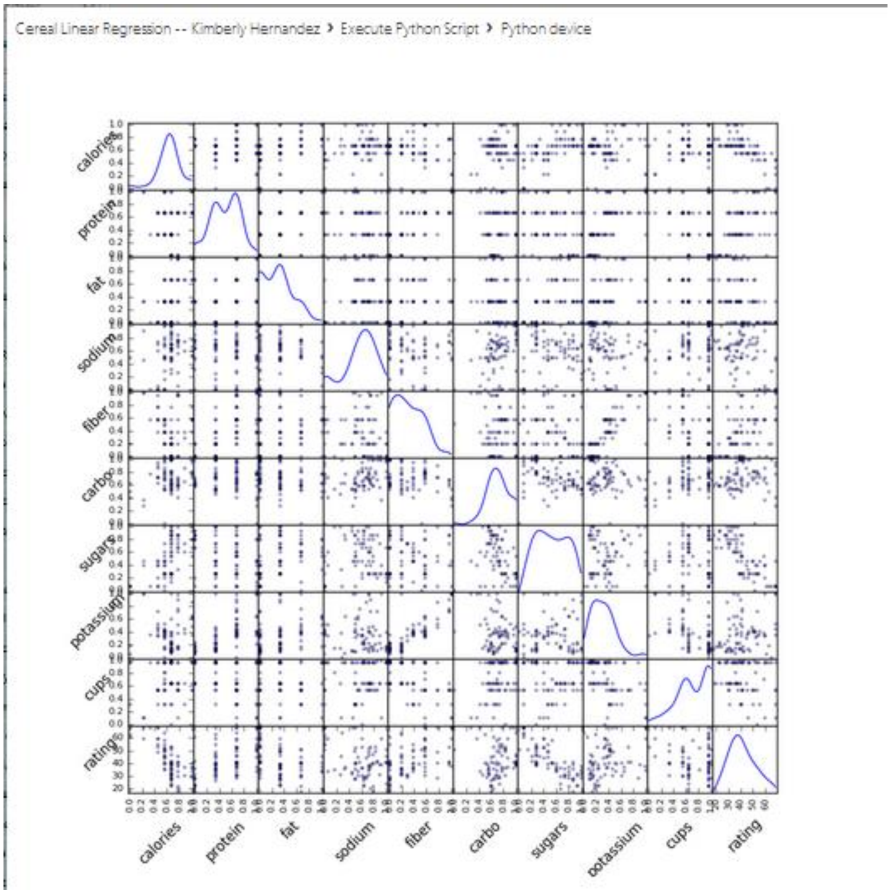


Python script

```python
def azureml_main(frame1):

## import libraries
    import matplotlib
    matplotlib.use('agg')  # Set backend

    from pandas.tools.plotting import scatter_matrix
    import pandas.tools.rplot as rplot
    import matplotlib.pyplot as plt
    import numpy as np

## Create a pair-wise scatter plot
    Azure = True


    fig1 = plt.figure(1, figsize=(10, 10))
    ax = fig1.gca()
    sm=scatter_matrix(frame1, alpha=0.3,
                  diagonal='kde', ax = ax)
    [s.xaxis.label.set_rotation(45) for s in sm.reshape(-1)]
    [s.yaxis.label.set_rotation(45) for s in sm.reshape(-1)]

    plt.show()
    fig1.savefig('scatter1.png')
```

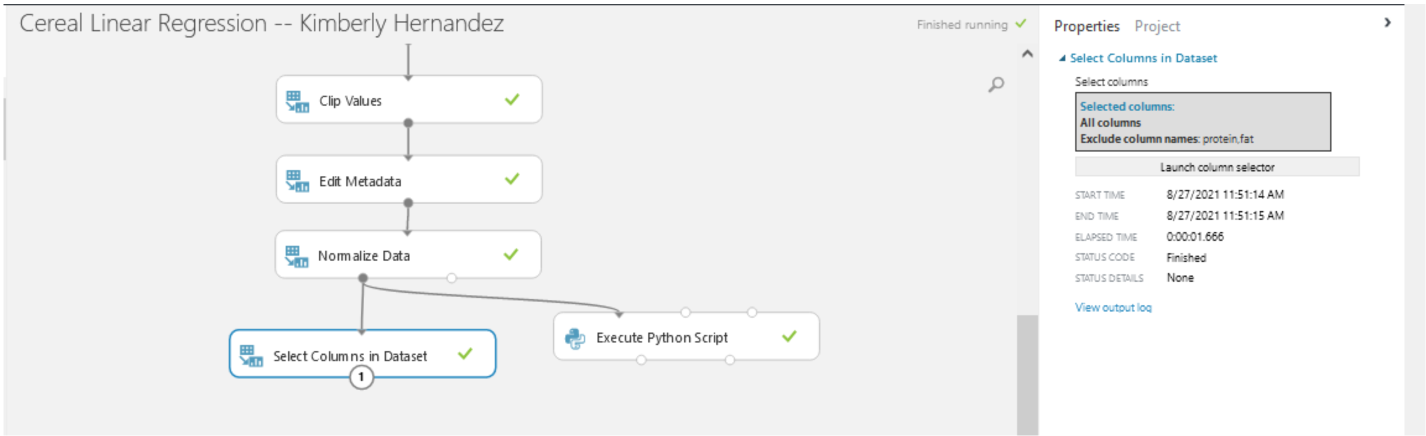
Python Version: Anaconda 4.0/Python 2.7.11

Code created Python charts.



Selecting features

Selected columns in dataset module added.
Fields chosen shown in properties panel.



Cereal Linear Regression -- Kimberly Hernandez

Finished running ✓

Properties Project

Select Columns in Dataset

Select columns

Selected columns:
All columns
Exclude column names: protein,fat

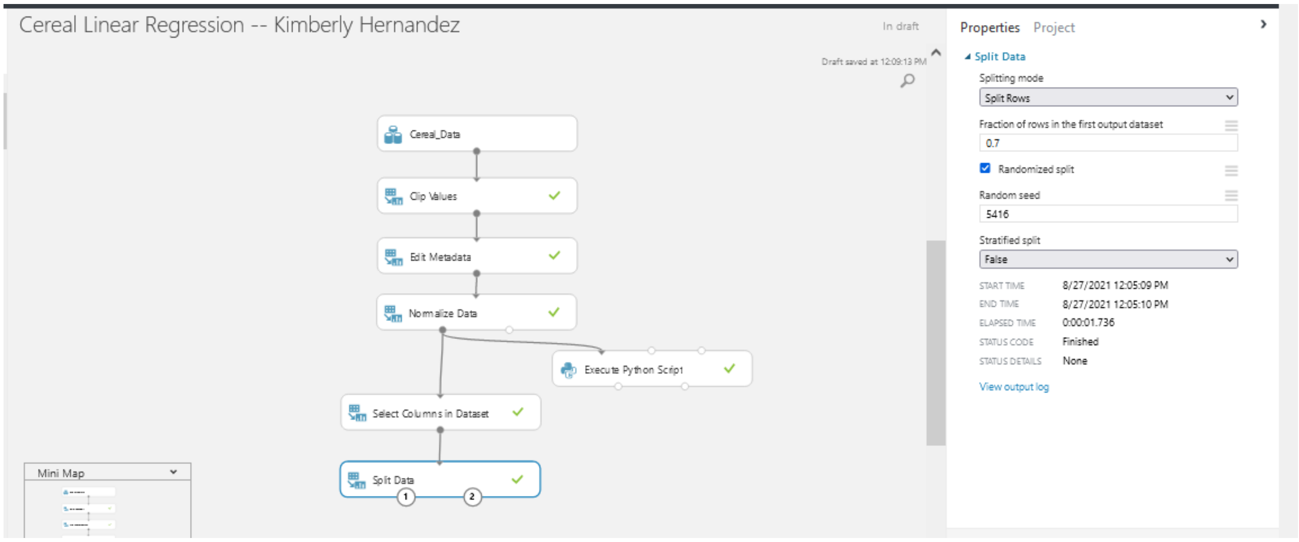
Launch column selector

START TIME	8/27/2021 11:51:14 AM
END TIME	8/27/2021 11:51:15 AM
ELAPSED TIME	0:00:01.666
STATUS CODE	Finished
STATUS DETAILS	None

View output log

Splitting data

Split module added.



Cereal Linear Regression -- Kimberly Hernandez

In draft

Draft saved at 12:09:13 PM

Properties Project

Split Data

Splitting mode: Split Rows

Fraction of rows in the first output dataset: 0.7

☑ Randomized split

Random seed: 5416

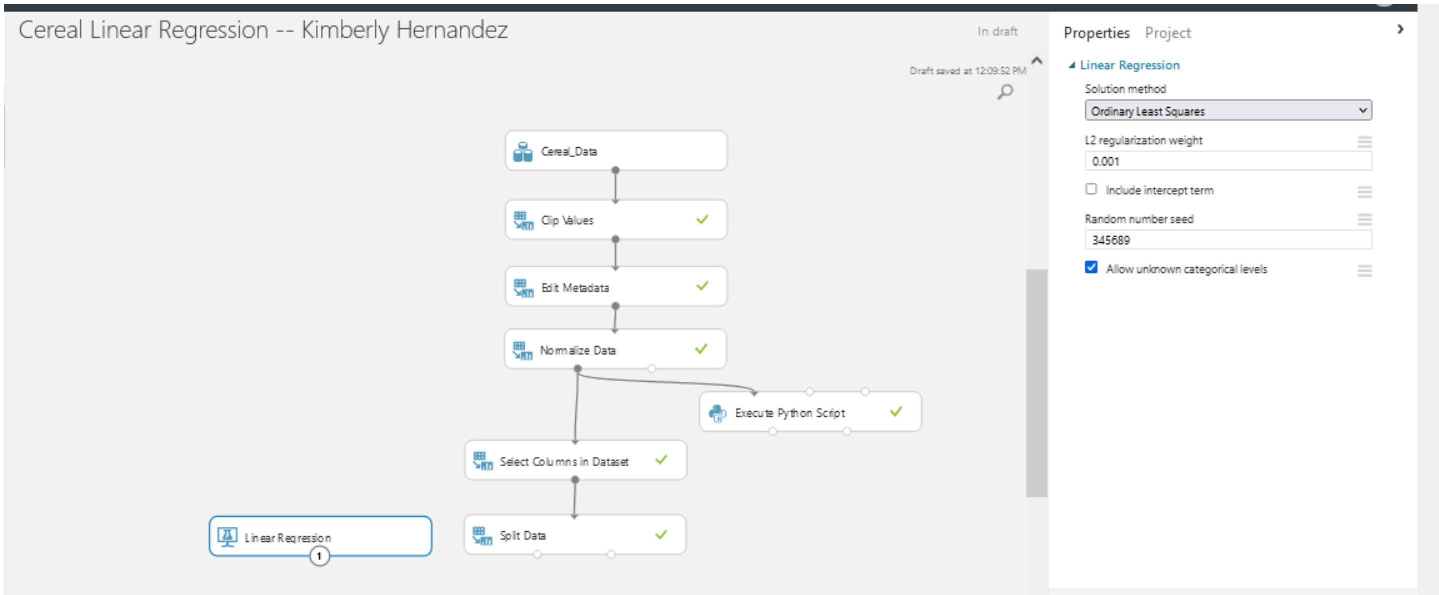
Stratified split: False

START TIME	8/27/2021 12:05:09 PM
END TIME	8/27/2021 12:05:10 PM
ELAPSED TIME	0:00:01.736
STATUS CODE	Finished
STATUS DETAILS	None

View output log

Linear regression model

Linear Regression Module added.



Linear Regression

Solution method: Ordinary Least Squares

L2 regularization weight: 0.001

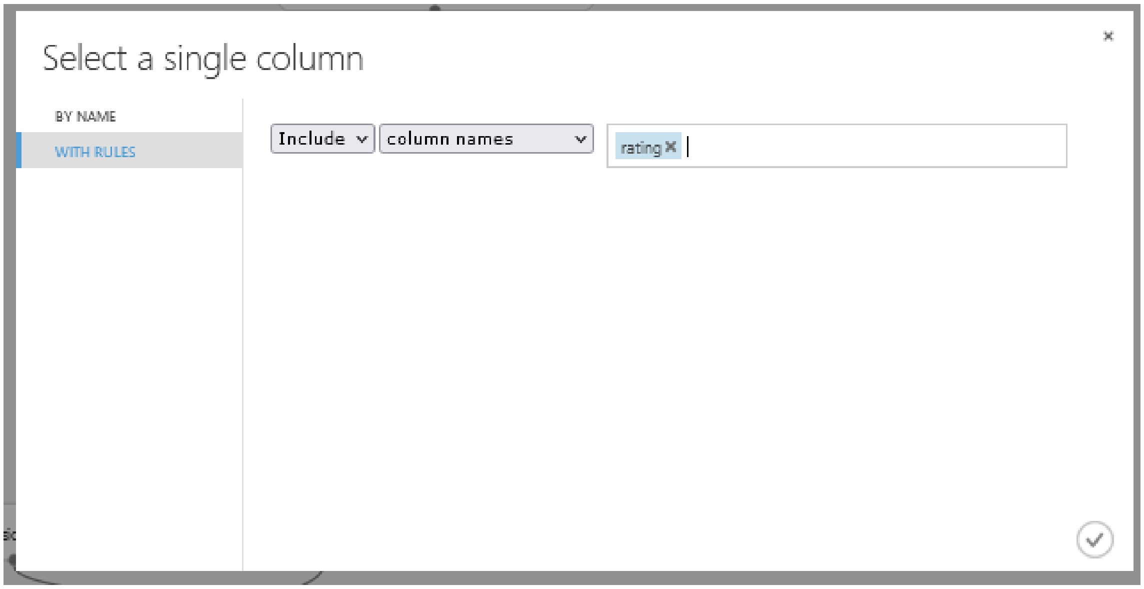
☐ Include intercept term

Random number seed: 345689

☑ Allow unknown categorical levels

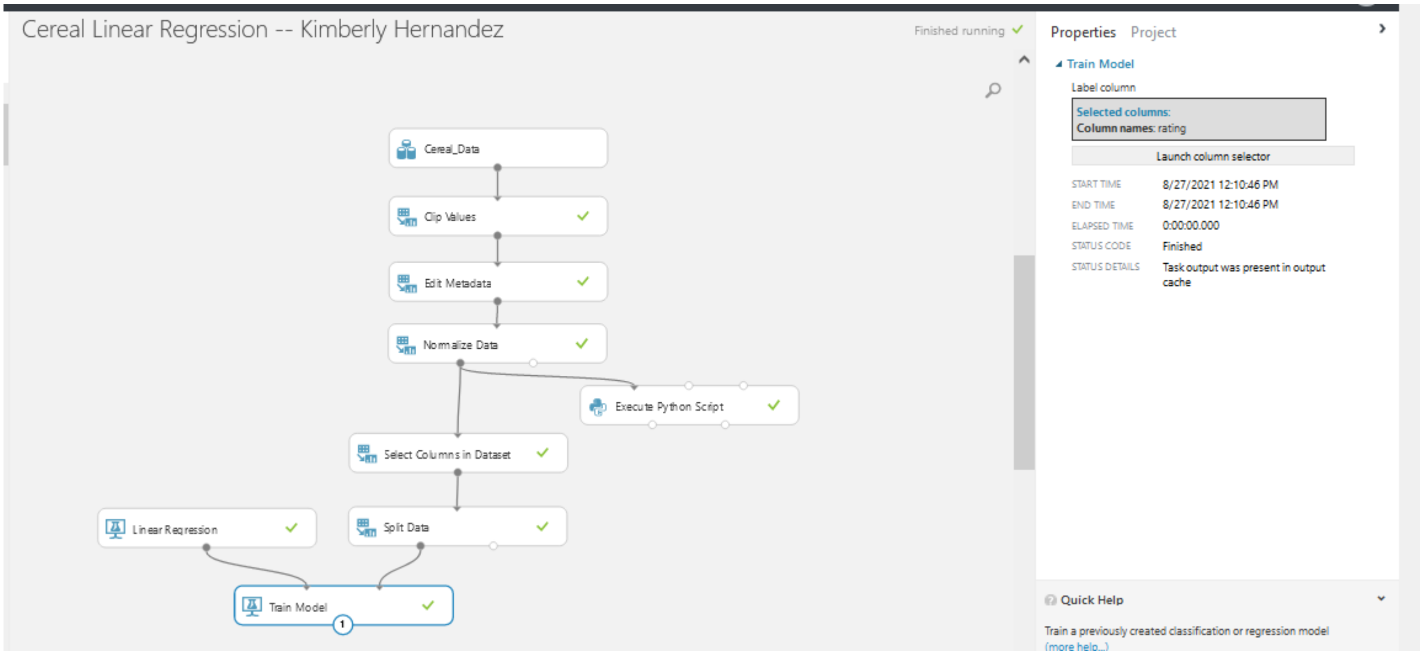
Training the model

‘Rating’ feature being selected under the ‘Train Model’ module.



Include | column names | rating ×

‘Train model’ module present.



Cereal Linear Regression -- Kimberly Hernandez

Finished running ✓

Properties Project

Train Model

Label column

Selected columns:
Column names: rating

Launch column selector

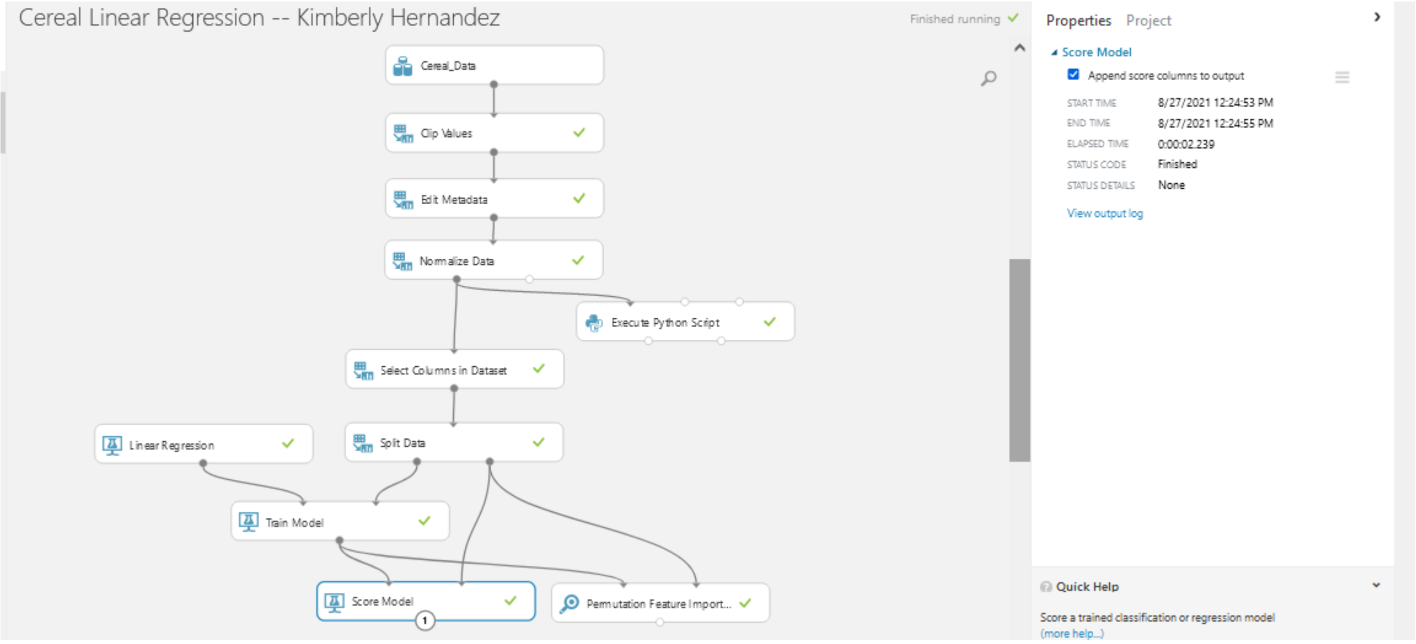
START TIME	8/27/2021 12:10:46 PM
END TIME	8/27/2021 12:10:46 PM
ELAPSED TIME	0:00:00.000
STATUS CODE	Finished
STATUS DETAILS	Task output was present in output cache

Quick Help

Train a previously created classification or regression model
(more help...)

Scoring the model – show scored labels

‘Score model’ module added.



Cereal Linear Regression -- Kimberly Hernandez

Finished running ✓

Properties Project

Score Model

☑ Append score columns to output

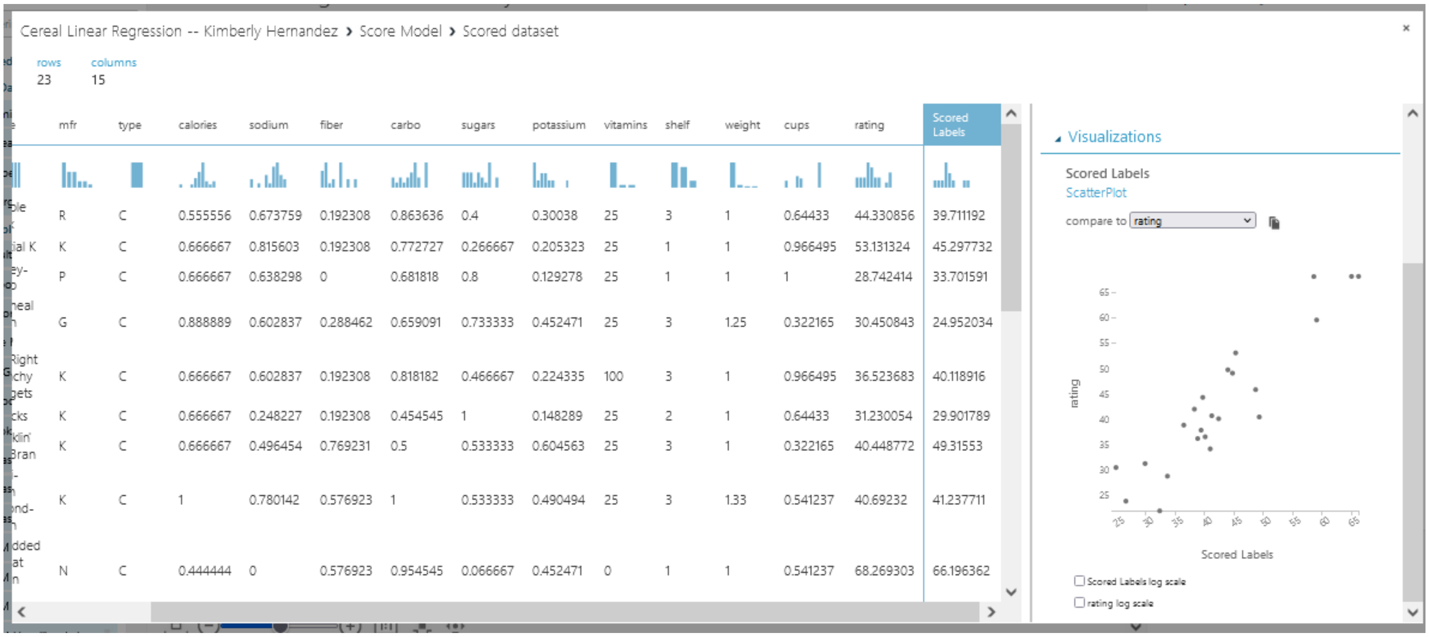
START TIME	8/27/2021 12:24:53 PM
END TIME	8/27/2021 12:24:55 PM
ELAPSED TIME	0:00:02.239
STATUS CODE	Finished
STATUS DETAILS	None

View output log

Quick Help

Score a trained classification or regression model
(more help...)

Scored labels visualization.



Cereal Linear Regression -- Kimberly Hernandez › Score Model › Scored dataset

rows 23 columns 15

e	mfr	type	calories	sodium	fiber	carbo	sugars	potassium	vitamins	shelf	weight	cups	rating	Scored Labels
ble	R	C	0.555556	0.673759	0.192308	0.863636	0.4	0.30038	25	3	1	0.64433	44.330856	39.711192
ial K	K	C	0.666667	0.815603	0.192308	0.772727	0.266667	0.205323	25	1	1	0.966495	53.131324	45.297732
ey- op	P	C	0.666667	0.638298	0	0.681818	0.8	0.129278	25	1	1	1	28.742414	33.701591
heal n	G	C	0.888889	0.602837	0.288462	0.659091	0.733333	0.452471	25	3	1.25	0.322165	30.450843	24.952034
Right chy gets	K	C	0.666667	0.602837	0.192308	0.818182	0.466667	0.224335	100	3	1	0.966495	36.523683	40.118916
cks	K	C	0.666667	0.248227	0.192308	0.454545	1	0.148289	25	2	1	0.64433	31.230054	29.901789
klin' Bran	K	C	0.666667	0.496454	0.769231	0.5	0.533333	0.604563	25	3	1	0.322165	40.448772	49.31553
- h nd- h	K	C	1	0.780142	0.576923	1	0.533333	0.490494	25	3	1.33	0.541237	40.69232	41.237711
dded at n	N	C	0.444444	0	0.576923	0.954545	0.066667	0.452471	0	1	1	0.541237	68.269303	66.196362

Visualizations

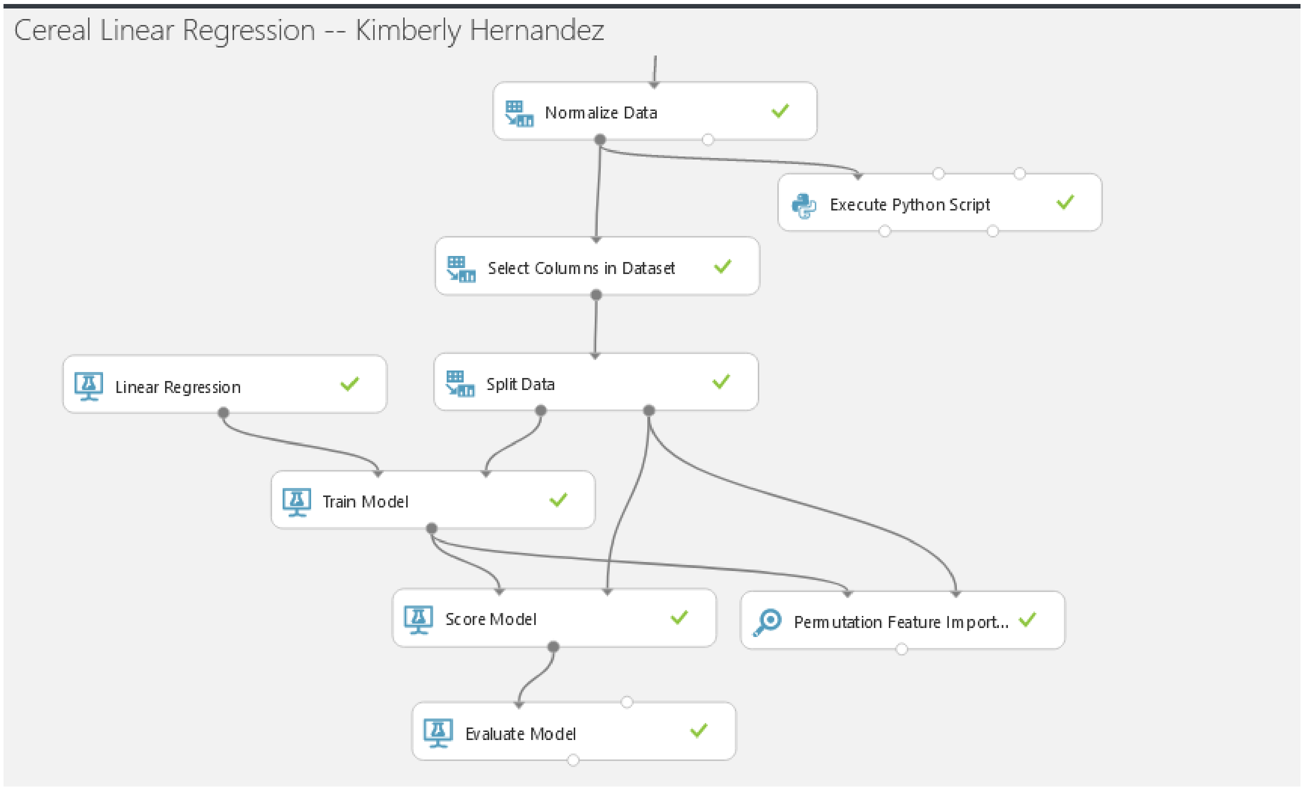
Scored Labels

ScatterPlot

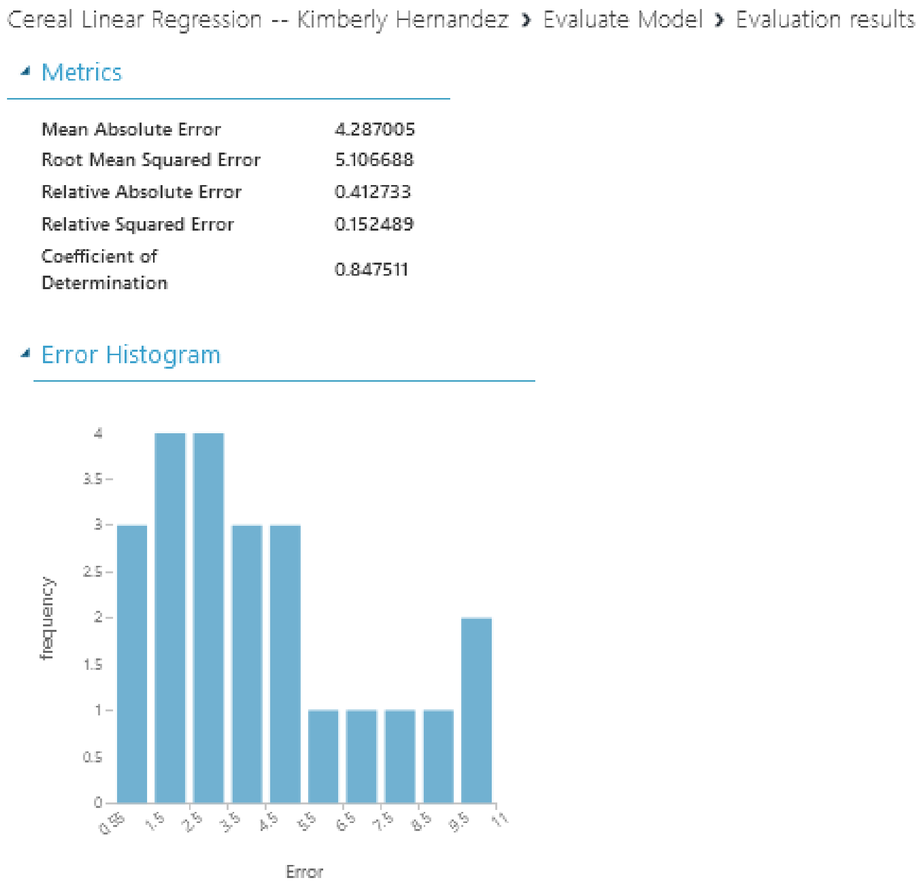
compare to rating

Evaluating the model

'Evaluate Model' module added.



All metrics and histograms.



Cereal Linear Regression -- Kimberly Hernandez › Evaluate Model › Evaluation results

Metrics

Mean Absolute Error	4.287005
Root Mean Squared Error	5.106688
Relative Absolute Error	0.412733
Relative Squared Error	0.152489
Coefficient of Determination	0.847511

Error Histogram

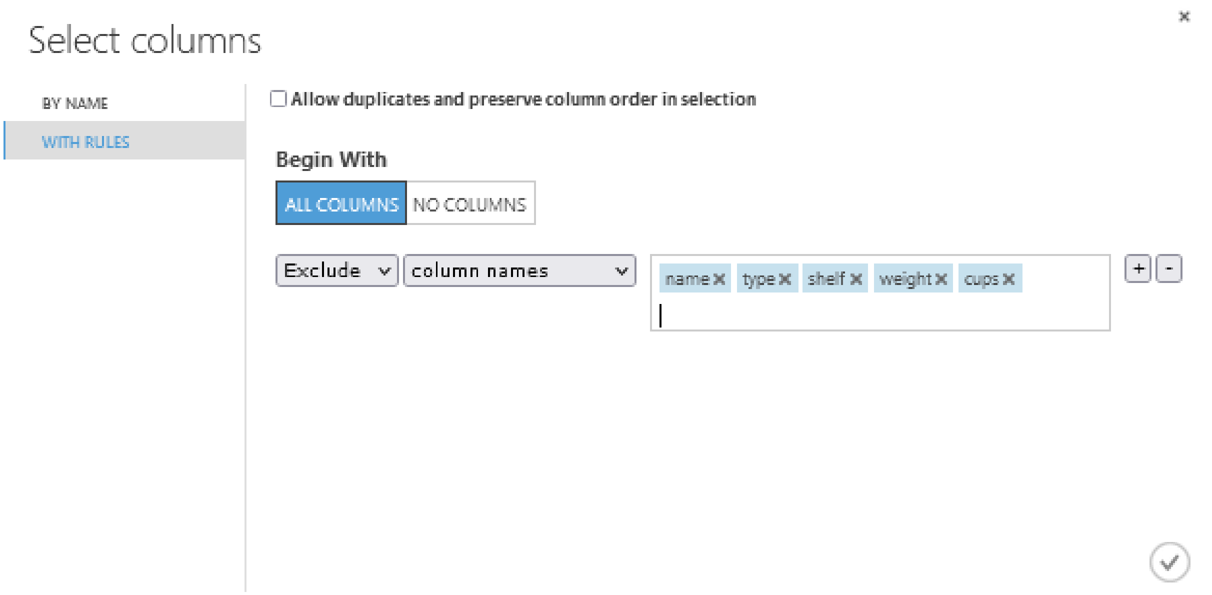
Explanation: This model excluded protein and fat because the scatter plot values from the python script showed scattered values in these two features. The current R² value is 0.847511 which is relatively close to the value of 1.0 which is considered a perfect model. Although there could be some improvement with the R² and Relative Squared Error, so an iteration process must take place.

Iteration process

Explanation: Since we want to consider error and bias of model some features must be put back in and some will be removed. Doing this will also take into consideration of the overall accuracy of model.

- ### Why were certain features excluded?

Explanation: At this point it has been determined through 'Permutation Feature Importance' that the model needs the name, type, shelf, weight, and cups to be excluded from the model. In addition to this process, protein and fat were added back in to evaluate where these values could bring the model's R² and Relative Squared Error. (Image below)



Select columns

BY NAME

WITH RULES

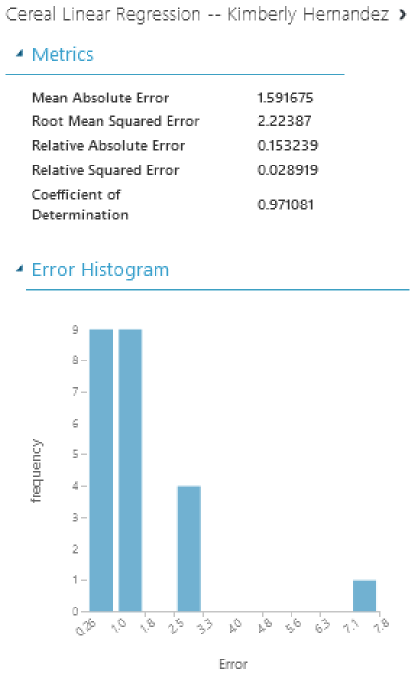
☐ Allow duplicates and preserve column order in selection

Begin With

ALL COLUMNS | NO COLUMNS

Exclude | column names | name ✖ type ✖ shelf ✖ weight ✖ cups ✖

- **New evaluation when those features were excluded.**



Cereal Linear Regression -- Kimberly Hernandez › Evaluate Model › Evaluation results

Metrics

Mean Absolute Error	1.591675
Root Mean Squared Error	2.22387
Relative Absolute Error	0.153239
Relative Squared Error	0.028919
Coefficient of Determination	0.971081

Error Histogram

Explanation: The model at this point includes protein & fat and excludes the ones listed above. Now the model has a R² of 0.971081 which is significantly closer to the value of 1.0 which would be considered a perfect model. There is also improvement of the Relative Squared Error which is at 0.028919 which is significantly closer to the value of 0.0 which would be considered a perfect model that has all model errors of 0. It seems that making these alterations to the model increasing its overall accuracy, bias, and error.

- **What features are most influential on the rating?**

  Explanation: When the model incorporated features such as Fiber, Sugars, mfr, carbo, vitamins, potassium, calories, and sodium the R² was at 0.847511. After adding protein and fat the model's R² significantly increased to 0.971081. It is safe to say after looking at the 'Permutation Feature Importance' module that the most influential features are sugars, calories, protein, carbo, and sodium.

Conclusion

The model reached an R² of 0.971081 and a Relative Squared Error(RSE) of 0.028919. It is determined that the model can predict cereal ratings due to the R² and RSE of the model reaching close to perfect parameters. It was important that the model undergo areas of importance such as the data preparation and iteration process. Without these two areas the model would not have reached an accuracy closer to 1.0.

Challenges Faced

An error occurred during normalization of model. Error was fixed when 'Vitamins', 'Shelf', and 'Weight' columns was excluded from the 'Normalized data' module.

Career skills obtained

Software Skills – These skills were obtained by using the Azure Machine Learning Studio.
Machine Learning Algorithms – Creating the model using linear regression.
Data Modeling and Evaluation – Finding patterns of the data to see which features make the model as accurate as possible. This also incorporates going through the evaluation and iteration process.
Problem Solving skills – Figuring out how to resolve an error during the process.
Communication skills – Creating a report on the dataset for others to understand the process of model development.