



Regression

Citizen Analytics – An Initiative by Data Science Team

START ►

© 2020 Petroliaam Nasional Berhad (PETRONAS)

All rights reserved. No part of this document may be reproduced in any form possible, stored in a retrieval system, transmitted and/or disseminated in any form or by any means (digital, mechanical, hard copy, recording or otherwise) without the permission of the copyright owner.

Learning Objectives

By the end of this module, you will be able to:



01

Define regression and how it works

02

Describe common regression algorithms

03

Learn how to perform regression in Azure ML Studio

Content

01.	Linear Regression	04
	a. What is Linear Regression?	
	b. Simple regression and fitting functions	
	c. Regression in Azure ML	
02.	Decision tree regression	12
	a. What is a decision tree?	
	b. From discrete model to continuous prediction	
	c. Decision Tree regression in Azure ML	
03.	Summary	16
04.	References	18

Linear Regression

What is Linear Regression?

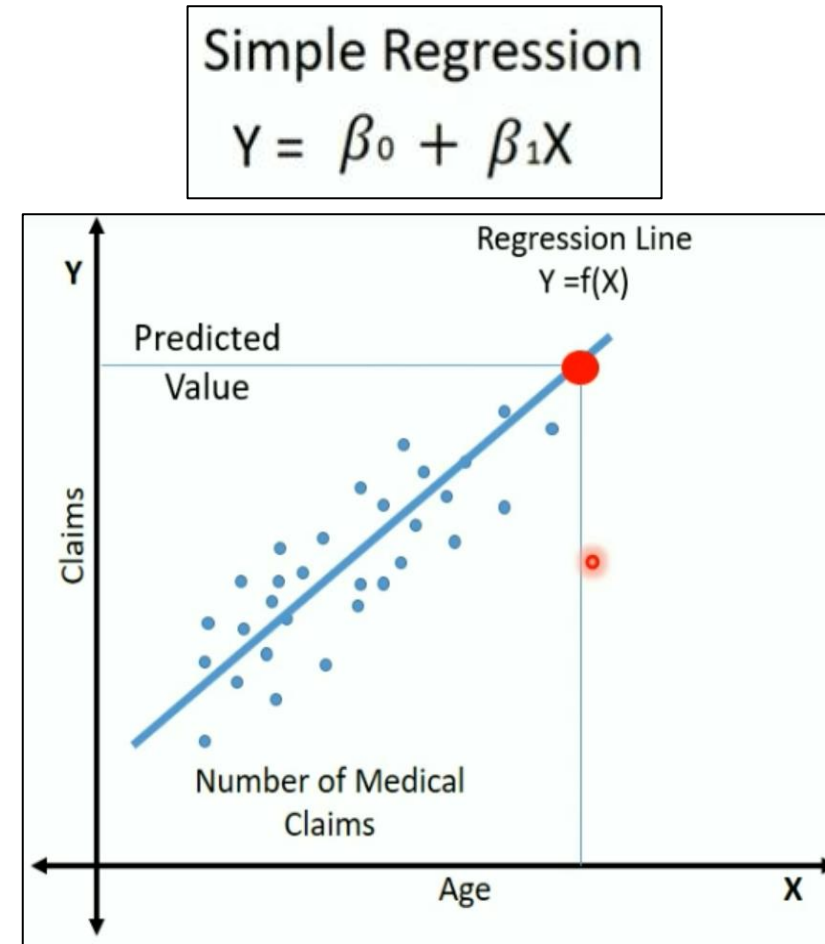
- Use Linear Regression when you want to understand something which is continually varying.
- Linear regression attempts to fit lines to your data, by minimizing an error function.
- Determine the relationship between the dependent variable (response) and one or more independent variables - predictors.
- Predictors and response variables are continuous variables.

Examples of continuous variables:

- Temperature of a boiler
- Flow rates
- Numbers of people expected
- Total spend

Simple Regression

- $Y = mX + C$
- m is gradient, C is y-intercept
- The line is fit to best match the data, based on error function.
- Given a new X , new estimate of Y can be made after parameters fitted.



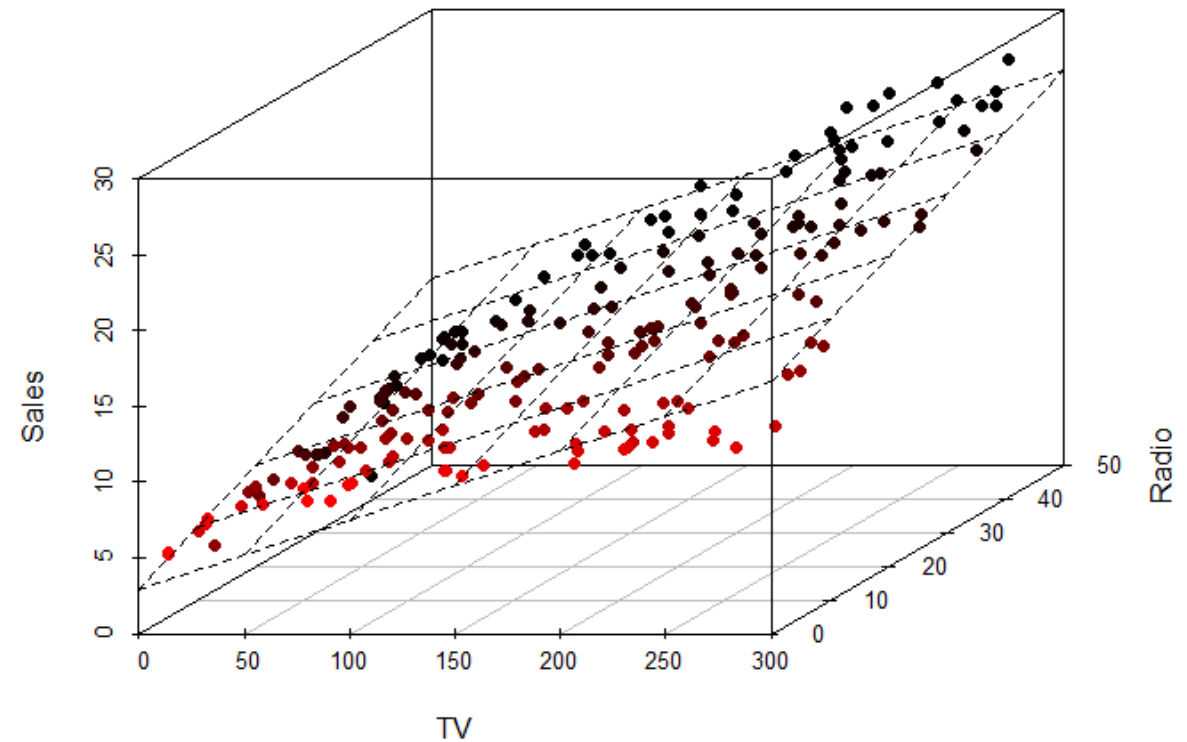
Source: Simple Linear Regression. <https://www.tech-quantum.com/classification-logistic-regression/>

Multivariate Linear Regression

- $Y = mX + C$ but in higher dimensions.
- Gradient and y-intercept parameters for every predictor.
- Hyperplane fit through the data in higher dimensional space.
- Given a new set of X_n , the new estimate of Y can be made after parameters fitted.

multivariate linear regression.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n$$

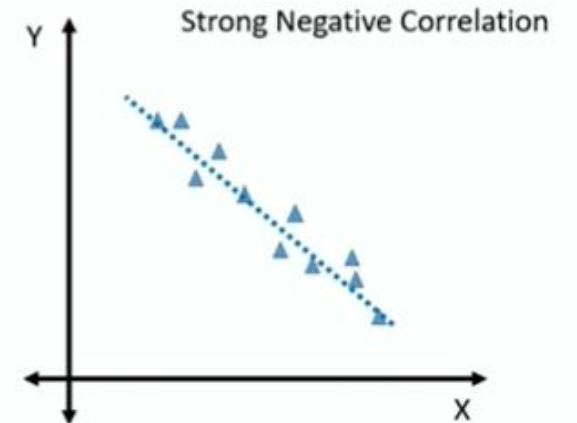
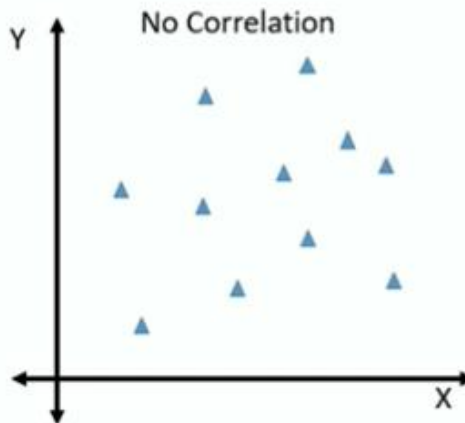
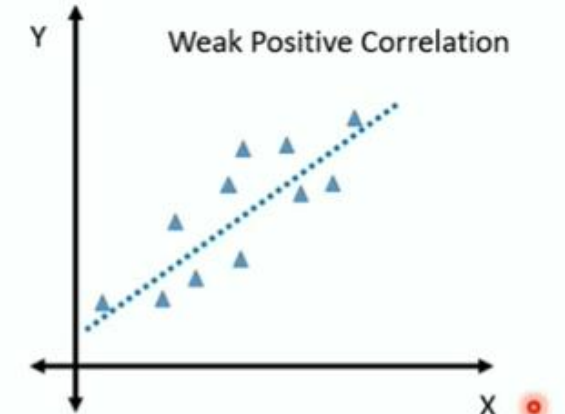
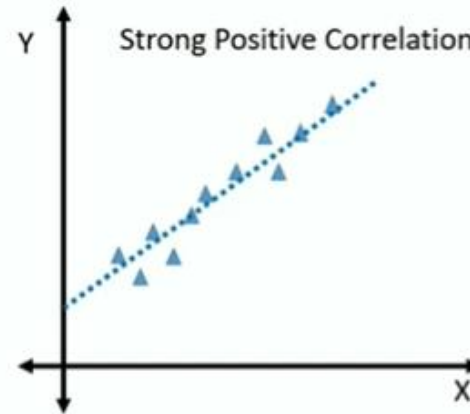


Source: *Multivariate Linear Regression*.

<https://stackoverflow.com/questions/26431800/plot-linear-model-in-3d-with-matplotlib>

Fitting functions

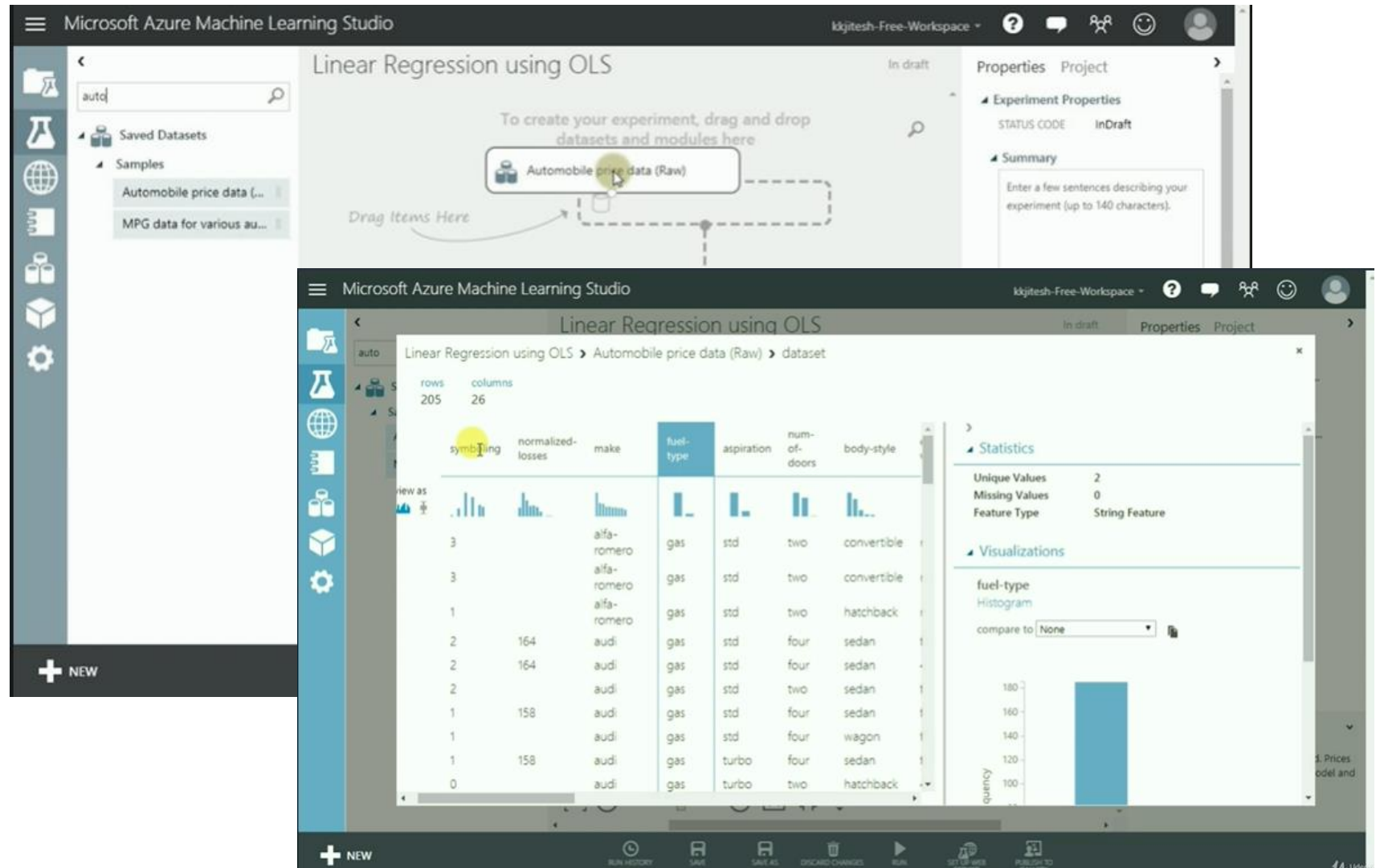
- What is the best line through your data?
- Controlled by fitting functions:
 - Ordinary least square in Y
 - $\sum_{\text{data}} (Y_{\text{data}} - Y_{\text{line}})^2$
 - Mean Absolute Error in Y
 - $\sum_{\text{data}} \frac{1}{n} |Y_{\text{data}} - Y_{\text{line}}|$
 - Root Mean Sq Error in Y
 - $\sqrt{\frac{1}{n} \sum_{\text{data}} (Y_{\text{data}} - Y_{\text{line}})^2}$



Source: Correlation and Linear Regression. https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Correlation-Regression/BS704_Correlation-Regression_print.html

Linear Regression in Azure ML

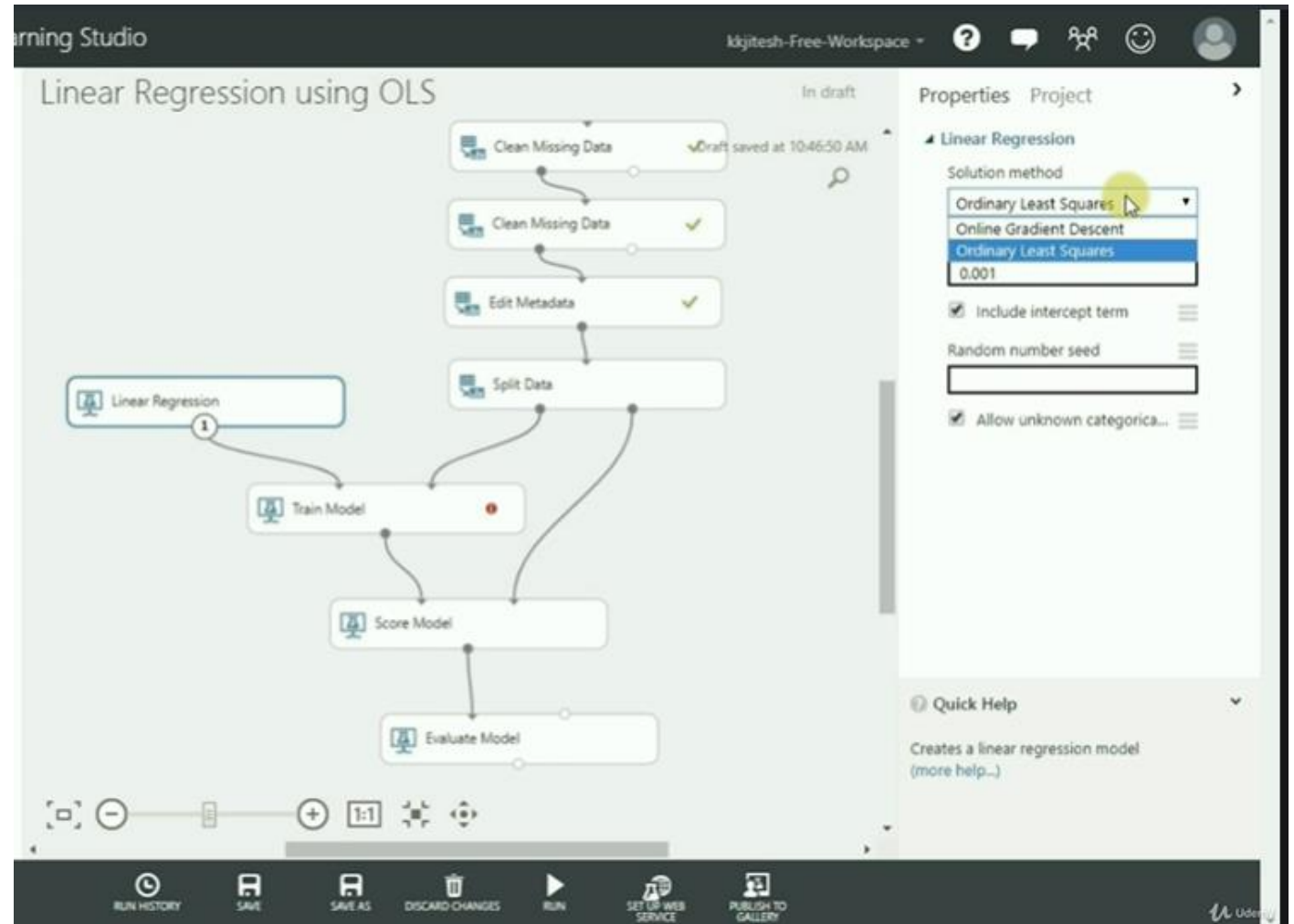
1. Load the data.
2. Right click on the data module and select Visualize.



(Cont...)

Linear Regression in Azure ML

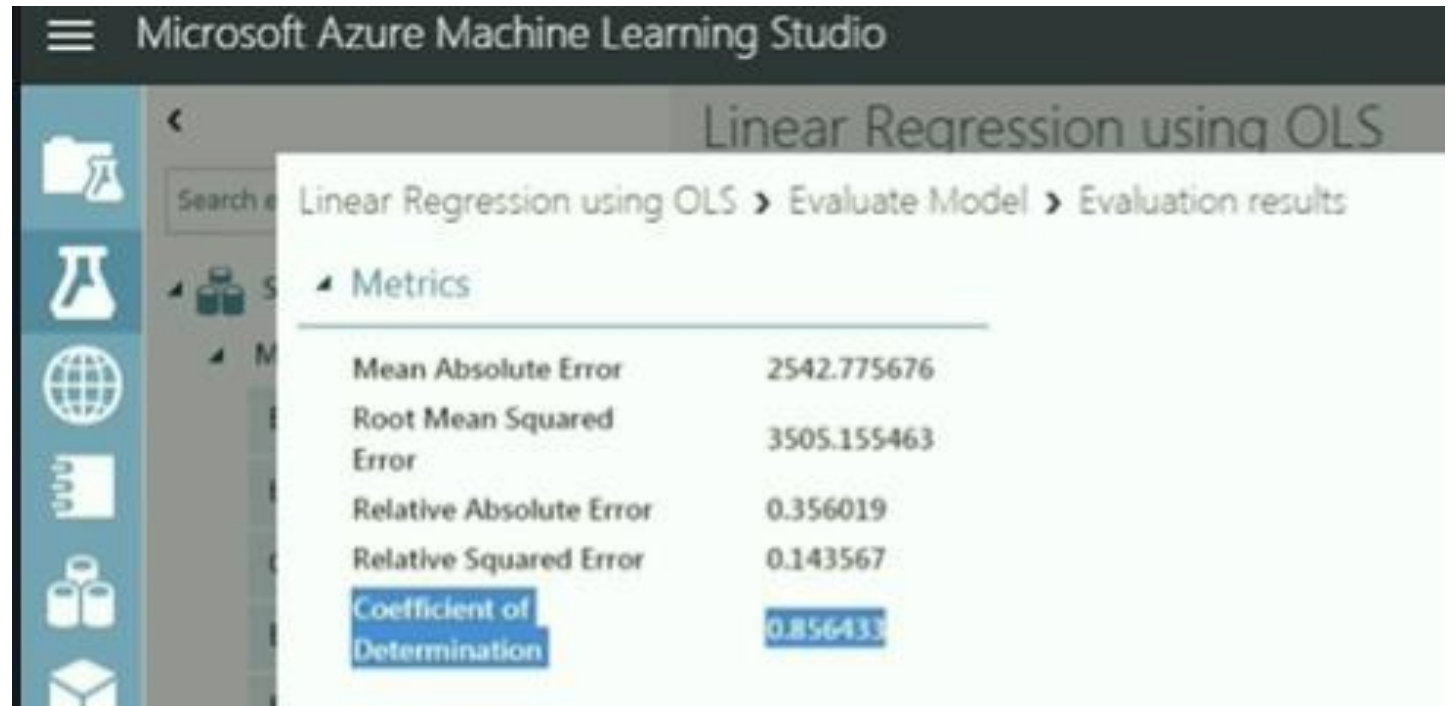
3. Add data cleansing steps and split data into test and training parts.
4. Add Linear Regression model, choose Ordinary Least Squares for error function to minimize over data.
5. Add Train Module block to fit the parameters based on training data.
6. Score Module and Evaluate Model blocks allow evaluation of the quality of fit of data, and performance against the test dataset.



(Cont...)

Linear Regression in Azure ML

7. Check the quality of fit metrics to evaluate the quality of the model.



Microsoft Azure Machine Learning Studio

Linear Regression using OLS

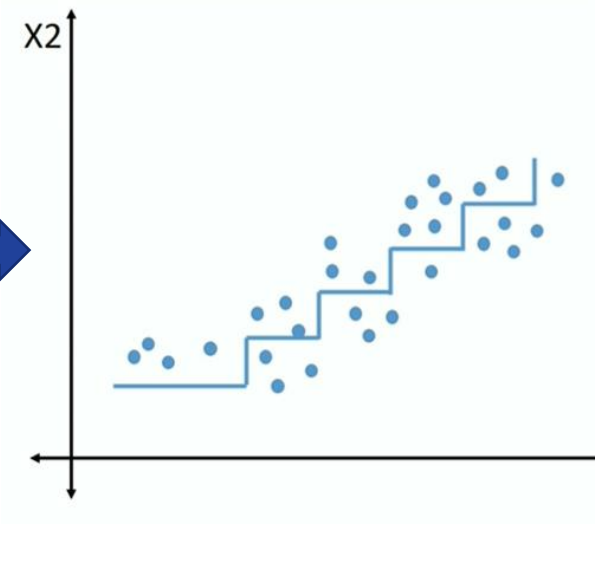
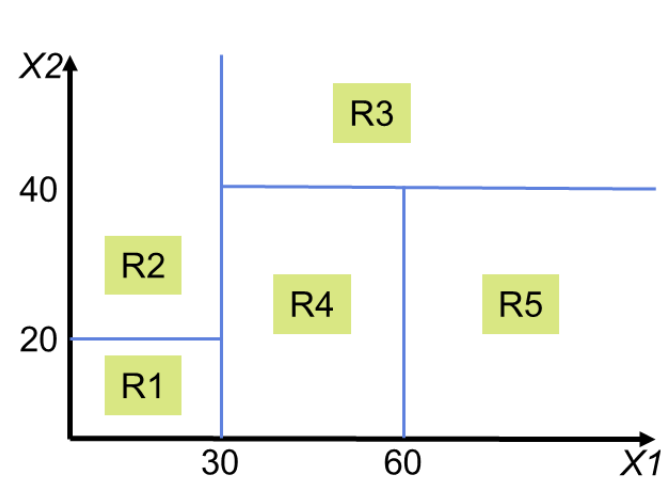
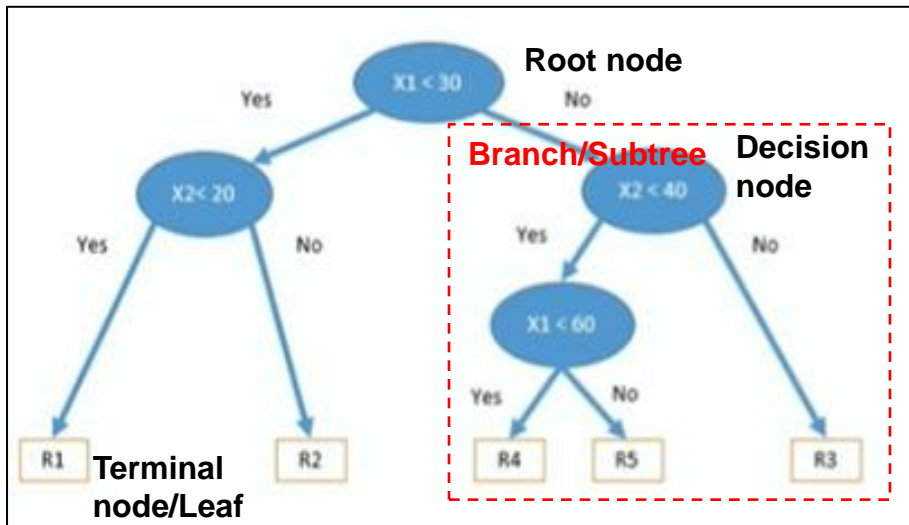
Linear Regression using OLS > Evaluate Model > Evaluation results

Metrics	
Mean Absolute Error	2542.775676
Root Mean Squared Error	3505.155463
Relative Absolute Error	0.356019
Relative Squared Error	0.143567
Coefficient of Determination	0.856433

Decision Tree Regression

What is Decision Tree Regression?

- Allows a decision tree approach to be taken with predictors.
- Dependent variable (response) still continuous variable.
- Categorical (only takes certain values) data can be used as a predictor.



To Fit Y:

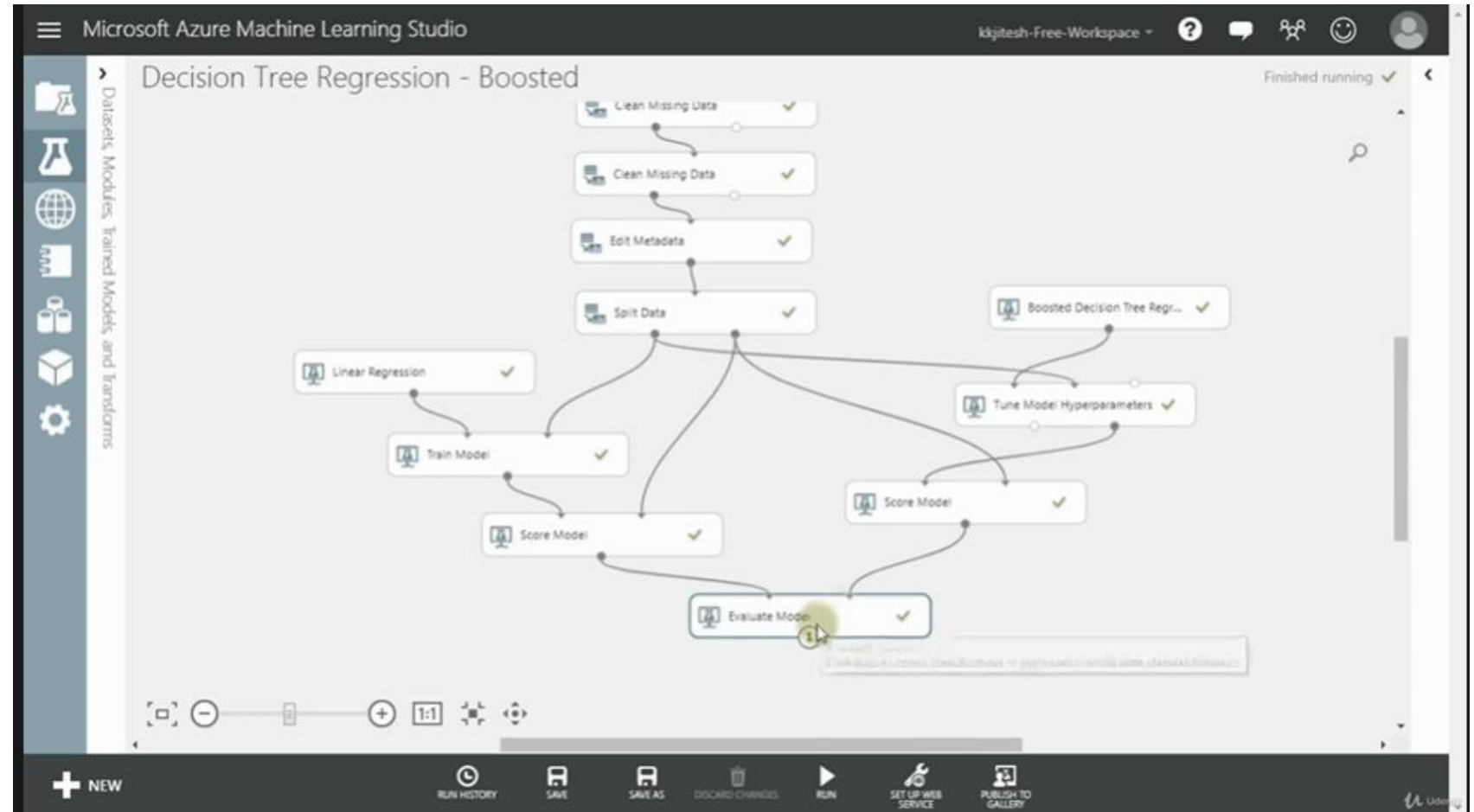
- Output of each region is mean of region.
- Every leaf may have regression line for points contained within that region.

Source: The parts of Power BI. <https://docs.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview>

Decision Tree Regression in AzureML

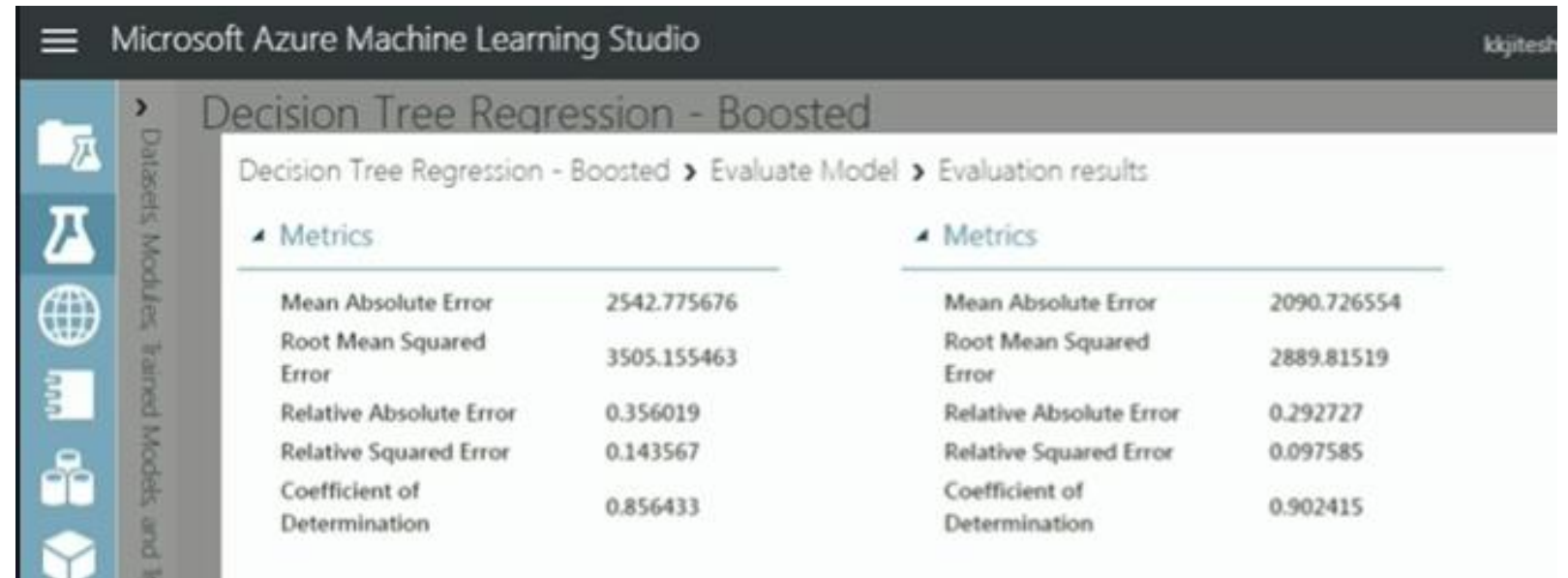
1. Can put Decision Tree Model in same canvas as Linear Regression for comparison.
2. Add Boosted Decision Tree module block.
3. Add Tune model hyperparameters for training model (different from Linear Regression).
4. Add score model block and connect into evaluate model block to compare 2 approaches.

(Cont...)



Decision Tree Regression in AzureML

5. Check quality of fit metrics to ensure model results look good.
6. By running two approaches in the same canvas you can compare 2 models against each other with the Evaluate model block.



The screenshot displays the Microsoft Azure Machine Learning Studio interface. The main title is "Decision Tree Regression - Boosted". Below it, the breadcrumb navigation shows "Decision Tree Regression - Boosted > Evaluate Model > Evaluation results". The left sidebar contains icons for Datasets, Modules, Trained Models, and Jobs. The main content area shows two side-by-side tables of metrics for comparison.

Metrics	
Mean Absolute Error	2542.775676
Root Mean Squared Error	3505.155463
Relative Absolute Error	0.356019
Relative Squared Error	0.143567
Coefficient of Determination	0.856433

Metrics	
Mean Absolute Error	2090.726554
Root Mean Squared Error	2889.81519
Relative Absolute Error	0.292727
Relative Squared Error	0.097585
Coefficient of Determination	0.902415

Summary

Summary

1

What is regression?

- Regression is a supervised machine learning method to predict a continuously varying output. (e.g. Temperature of a boiler, flow rates, number of people expected, total spend)

2

Different regression algorithms available in Azure ML Studio

- Linear Regression and Decision Tree Regression etc.

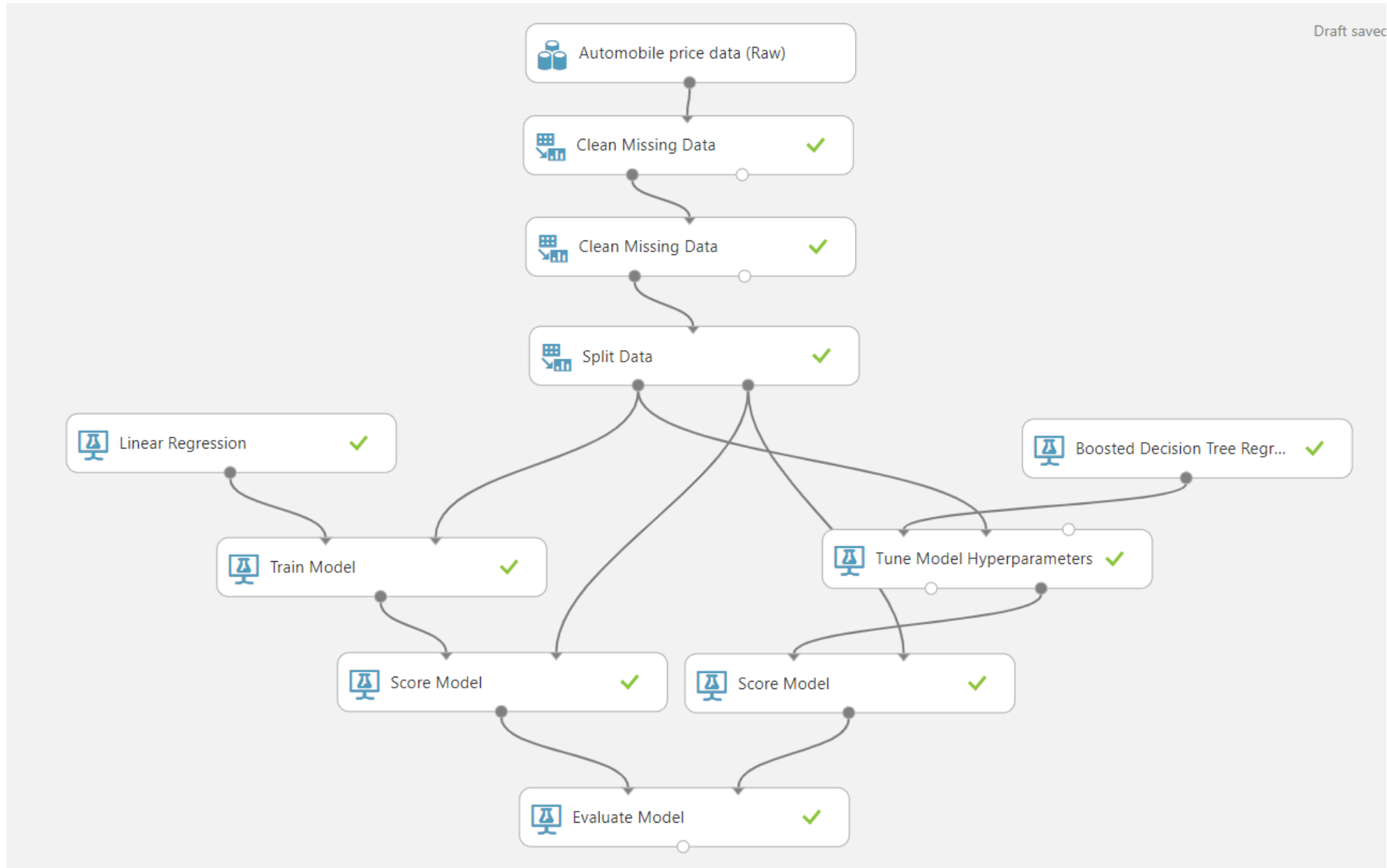
3

Other facts on regression

- Linear regression can be used to extrapolate (predict outside range used for training) but be careful!
- Regression models can fit the data in different ways, controlled by the error function.

Exercise

Recreate linear regression and boosted decision tree flow



Clean missing data details

Clean Missing Data

Columns to be cleaned

Selected columns:

Column names:

normalized-losses,wheel-base,length,width,height,cur weight,engine-size,bore,stroke,compression ratio,horsepower,peak-rpm,city-mpg,highway-mpg,price

Launch column selector

Minimum missing value ra...

0

Maximum missing value r...

1

Cleaning mode

Replace with mean

Cols with all missing values

Remove

Columns to be cleaned

Selected columns:

Column names:

make,fuel-type,aspiration,num-of-doors,body-style,drive-wheels,engine-location,engine-type,num-of-cylinders,fuel-system,symboling

Launch column selector

Minimum missing value ra...

0

Maximum missing value r...

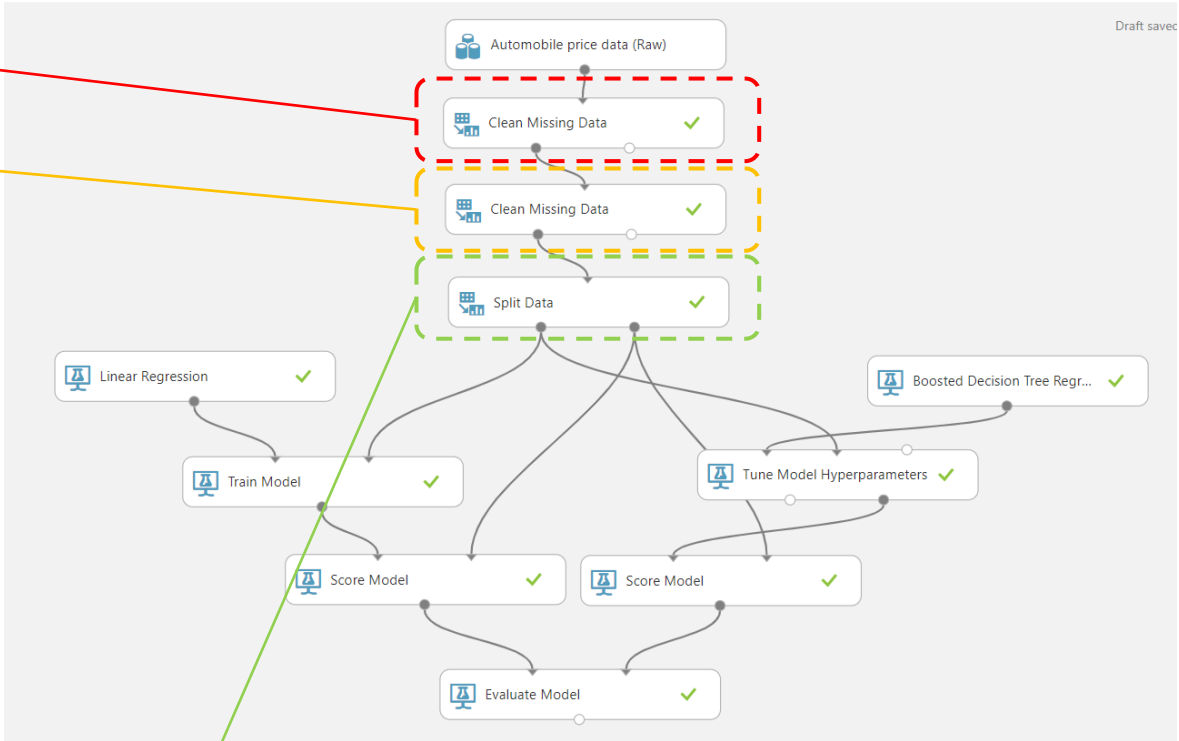
1

Cleaning mode

Replace with mode

Cols with all missing values

Remove



Split ratio = 0.8

Regression details – Target variable for training is PRICE

Linear Regression

Solution method
Online Gradient Descent

Create trainer mode
Parameter Range

Learning rate
☐ Use Range Builder
0.025, 0.05, 0.1, 0.2

Number of training epochs
☐ Use Range Builder
1, 10, 100

L2 regularization weight
☐ Use Range Builder
0.001, 0.01, 0.1

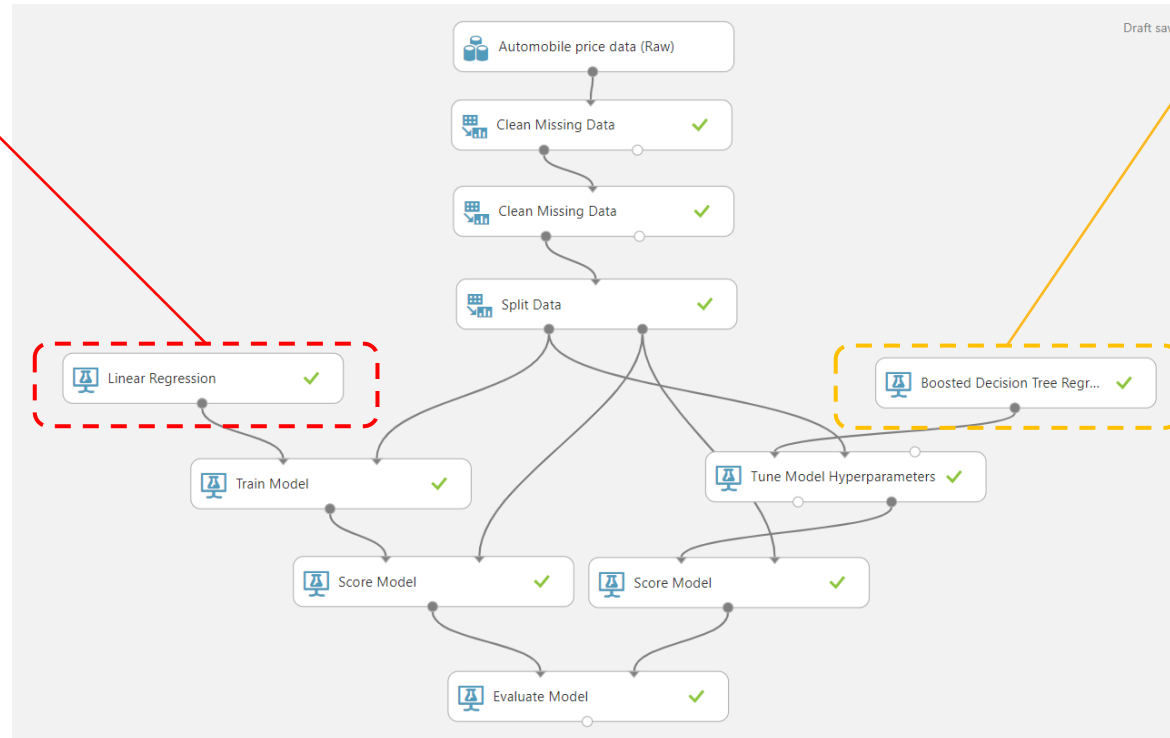
☒ Normalize features

☒ Average final hypothe...

☒ Decrease learning rate

Random number seed
123

☒ Allow unknown categ...



Try both

Boosted Decision Tree Regression

Create trainer mode
Single Parameter

Maximum number of leaves
20

Minimum number of samples
10

Learning rate
0.2

Total number of trees constructed
100

Random number seed

☒ Allow unknown categories

START TIME 9/24/2020 ...

END TIME 9/24/2020 ...

ELAPSED TIME 0:00:03.191

STATUS CODE Finished

STATUS DETAILS None

[View output log](#)

Boosted Decision Tree Regression

Create trainer mode
Parameter Range

Maximum number of leaves
☐ Use Range Builder
2, 8, 32, 128

Minimum number of samples
☐ Use Range Builder
1, 10, 50

Learning rate
☐ Use Range Builder
0.025, 0.05, 0.1, 0.2, 0.4

Number of trees constructed
☐ Use Range Builder
20, 100, 500

Random number seed

☒ Allow unknown categories

Target variable for training is
PRICE

References

References

Regression in Azure ML

<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/machine-learning-initialize-model-regression>

Linear Regression

<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/linear-regression>

Boosted Decision Tree Regression

<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/boosted-decision-tree-regression>

Decision Forest Regression

<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/decision-forest-regression>

Thank you for your passion!

