Diamond

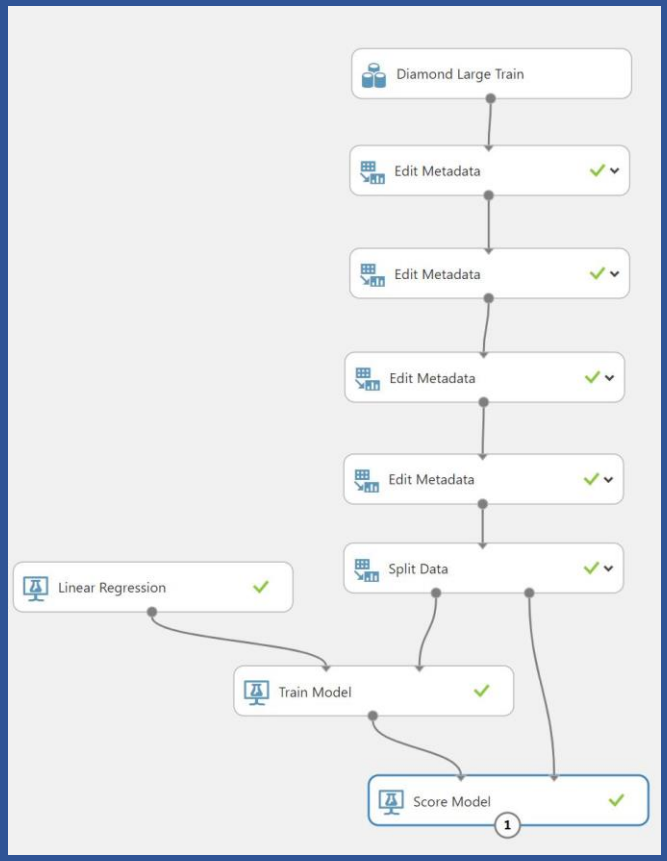
Regression experimental in Azure ML

What's in this session?

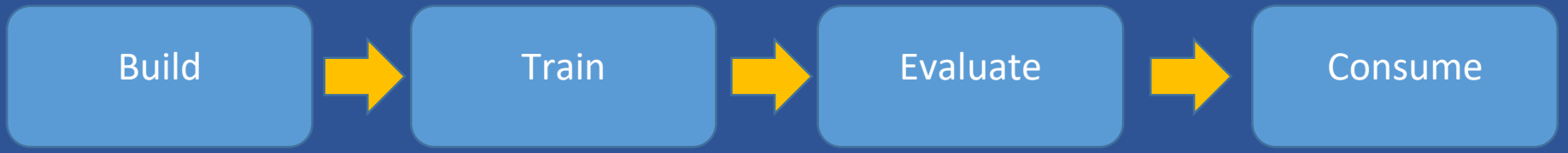
1. Objective: Predict price of a diamond
2. Prepare dataset
3. Understand dataset
4. Create ML model in Azure ML Studio
5. Score / Evaluate model

The finished model

https://raw.githubusercontent.com/laploy/ML.NET/master/Diamond%20Large/diamond%20large%20model.JPG



Work flow



Prepare dataset

1. Create folder D:\ml
2. Download dataset to D:\ml
3. diamond-Large-Train

https://raw.githubusercontent.com/laploy/ML.NET/master/Diamond%20Large/diamonds-Large-Train.csv

Understand dataset+

Look at the diamonds-Large-Test.csv

https://github.com/laploy/ML.NET/blob/master/Diamond%20Large/diamonds-Large-Test.csv

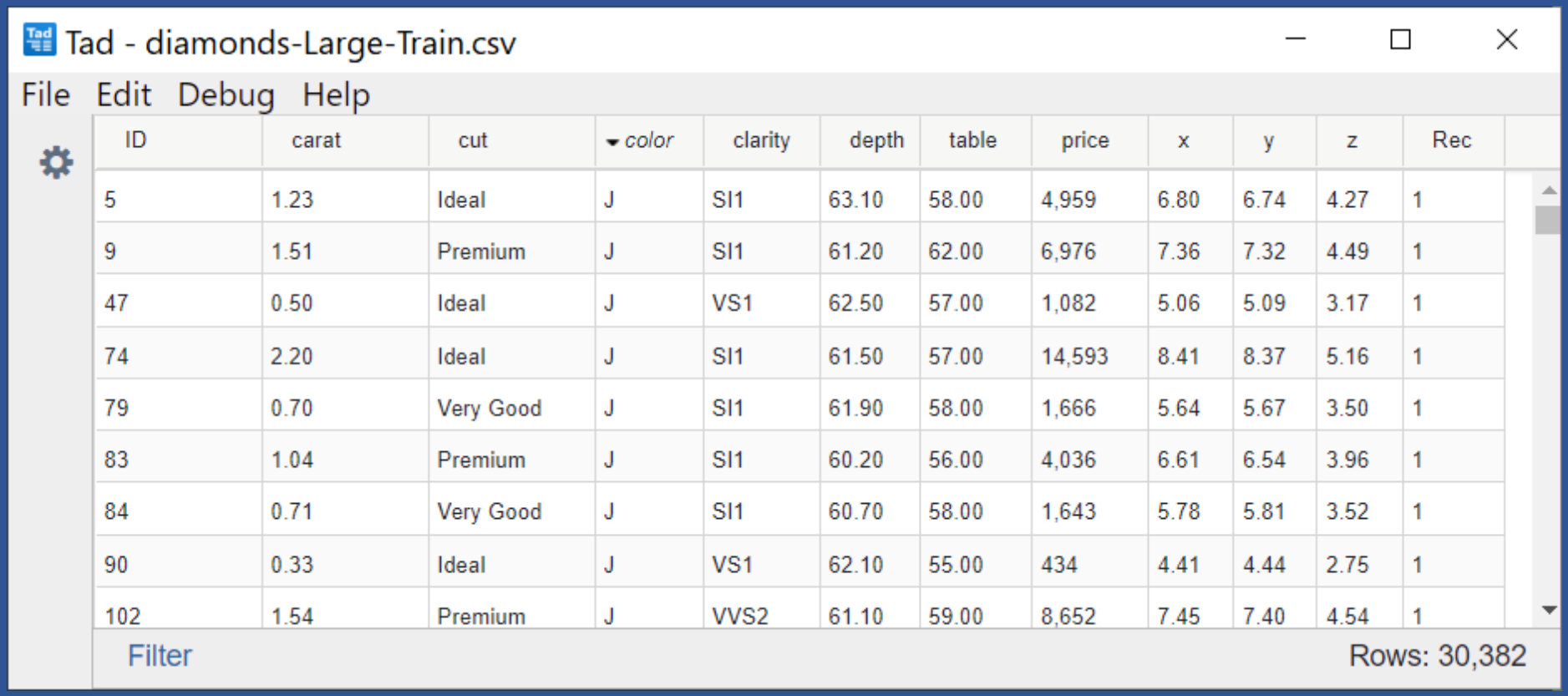
	A	B	C	D	E	F	G	H	I	J	K
1	ID	carat	cut	color	clarity	depth	table	price	x	y	z
2	34944	0.3	Ideal	E	VS2	61.5	56	658	4.29	4.33	2.65
3	34945	1.2	Very Gooc	J	VS2	62.5	57	4590	6.72	6.79	4.22
4	34946	1.01	Very Gooc	F	VVS1	62.9	57	10019	6.35	6.41	4.01
5	34947	0.5	Ideal	F	IF	62	55	2645	5.09	5.13	3.17
6	34948	0.28	Very Gooc	F	IF	62.2	55	612	4.23	4.26	2.64
7	34949	1.56	Ideal	D	SI1	62.2	58	10934	7.37	7.42	4.6
8	34950	2.05	Premium	J	VS1	60.1	58	15067	8.25	8.19	4.94
9	34951	1.03	Ideal	G	SI1	62.5	57	5337	6.4	6.49	4.03
10	34952	0.38	Very Gooc	I	VS2	61.7	56	680	4.65	4.68	2.88
11	34953	0.9	Very Gooc	G	SI1	62.2	58	4435	6.15	6.2	3.84

Download and Install Tad

Tad is a free (MIT Licensed) desktop application for viewing and analyzing tabular data.

https://www.tadviewer.com/

Open file: diamonds-Large-Train.csv



ID	carat	cut	color	clarity	depth	table	price	x	y	z	Rec
5	1.23	Ideal	J	SI1	63.10	58.00	4,959	6.80	6.74	4.27	1
9	1.51	Premium	J	SI1	61.20	62.00	6,976	7.36	7.32	4.49	1
47	0.50	Ideal	J	VS1	62.50	57.00	1,082	5.06	5.09	3.17	1
74	2.20	Ideal	J	SI1	61.50	57.00	14,593	8.41	8.37	5.16	1
79	0.70	Very Good	J	SI1	61.90	58.00	1,666	5.64	5.67	3.50	1
83	1.04	Premium	J	SI1	60.20	56.00	4,036	6.61	6.54	3.96	1
84	0.71	Very Good	J	SI1	60.70	58.00	1,643	5.78	5.81	3.52	1
90	0.33	Ideal	J	VS1	62.10	55.00	434	4.41	4.44	2.75	1
102	1.54	Premium	J	VVS2	61.10	59.00	8,652	7.45	7.40	4.54	1

Understand dataset

Information on 35000 round shape diamonds collected under 10 aspects;

- Price –The price of the diamond in US dollars ($326–$18,823)
- Carat- The weight of the diamond where 1carat=200mg.
- x- Length in mm (0–10.74)
- y- Width in mm (0–58.9)
- z- Depth in mm (0–31.8)
- Depth-Total depth percentage
- Table- Width of top of diamond relative to widest point (43–95)

Categorical Variables

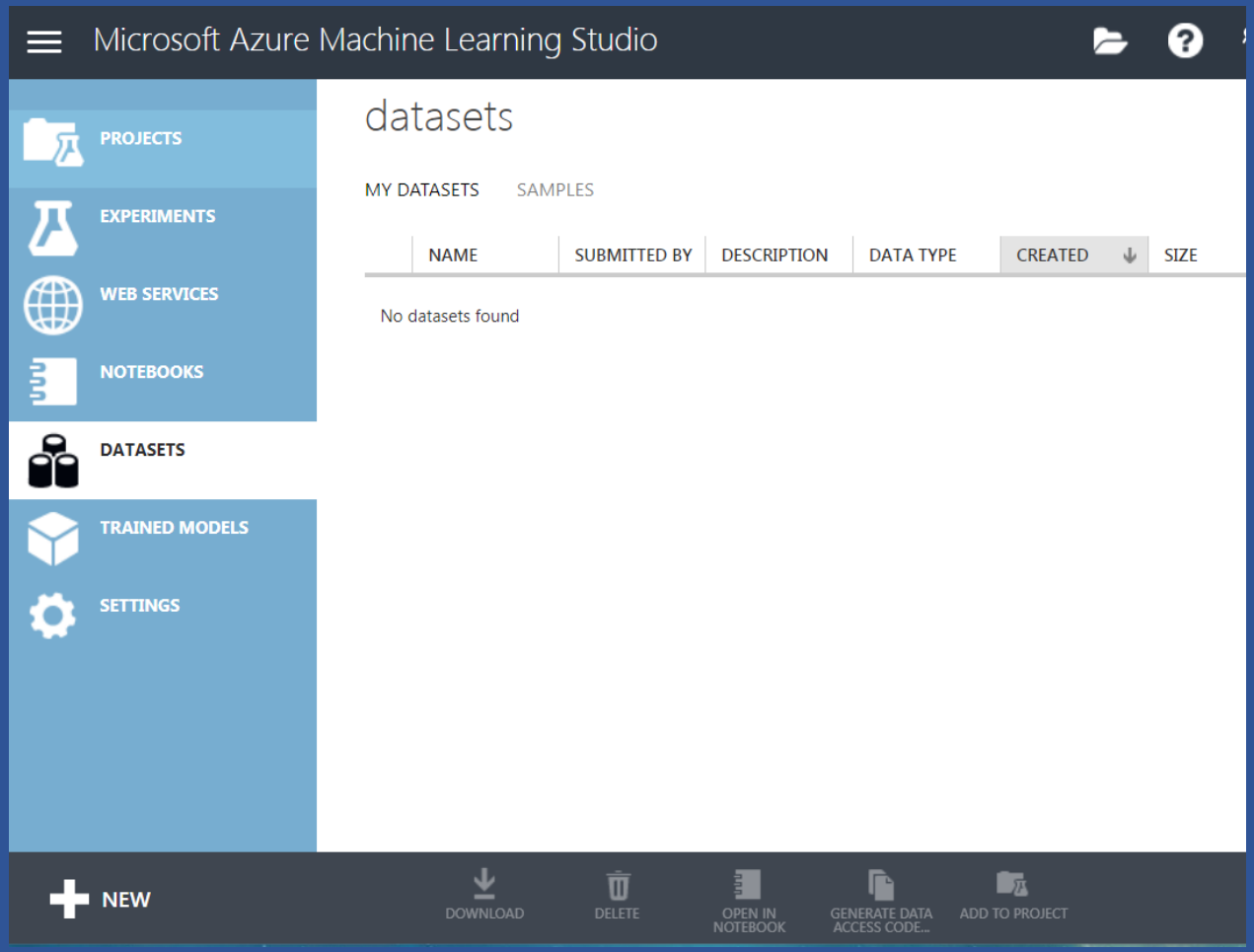
- Cut- The quality of the cut (Fair < Good< Very Good< Premium< Ideal)
- Color- Color of the diamond [D (best), E, F, G, H, I, J (worst)]
- Clarity- Measures how clear the diamond is (I1 (worst), SI1, SI2, VS1, VS2, VVS1, VVS2, IF (best))

Features = All column except ID

Label = Price

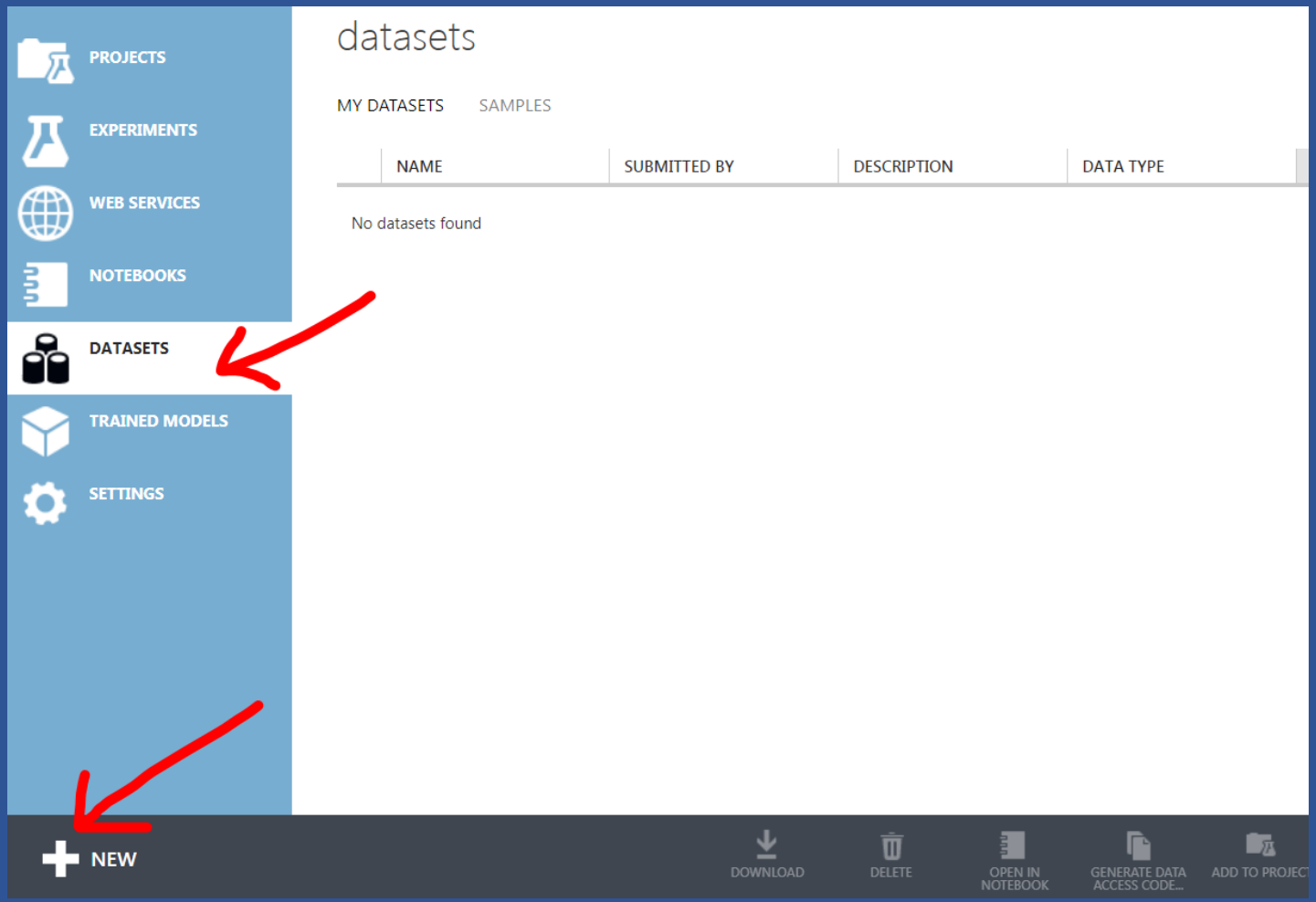
Create Data Set in Azure ML Studio

Open web browser / go to Azure ML Studio https://studio.azureml.net/



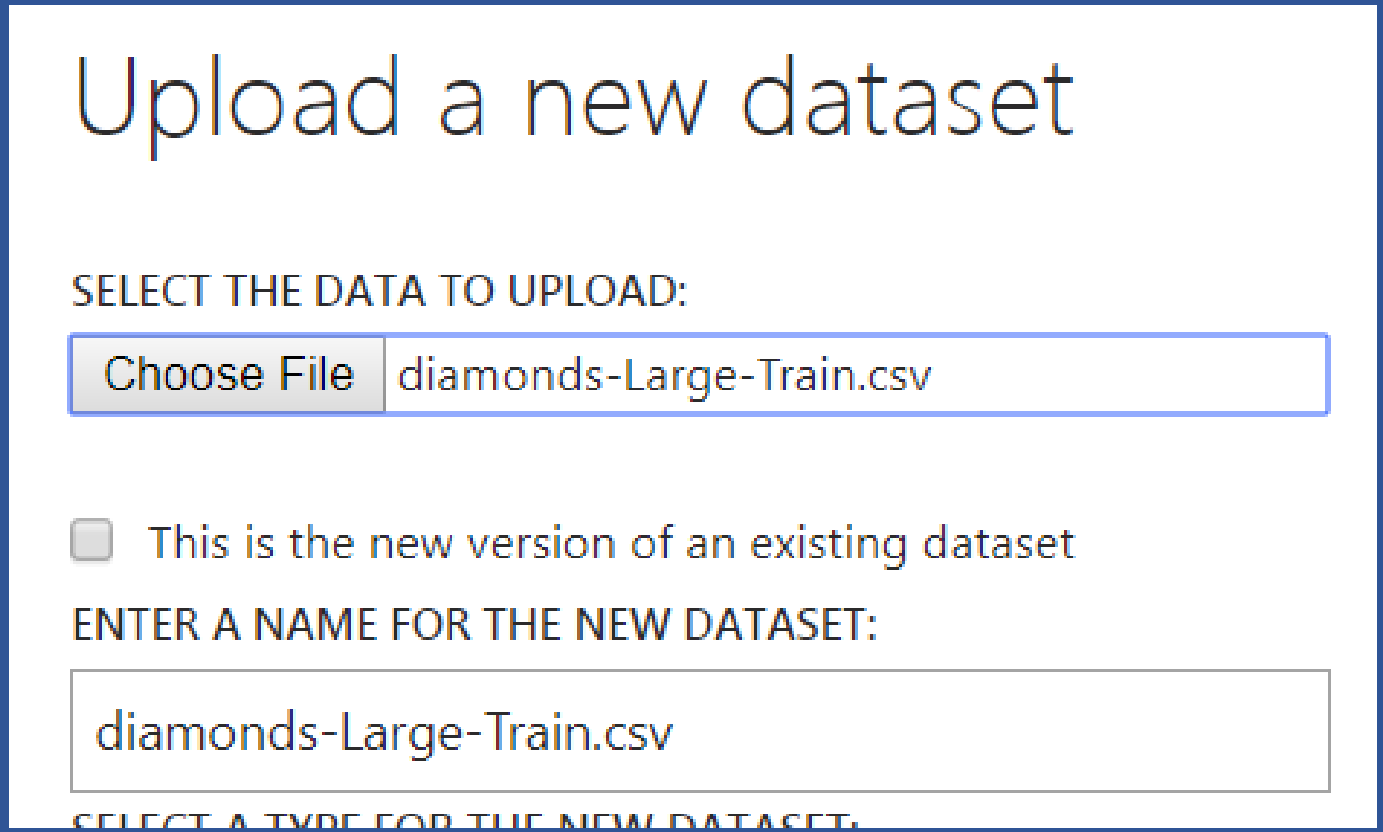
Create Data Set

Click DATASET / + NEW



Create Data Set

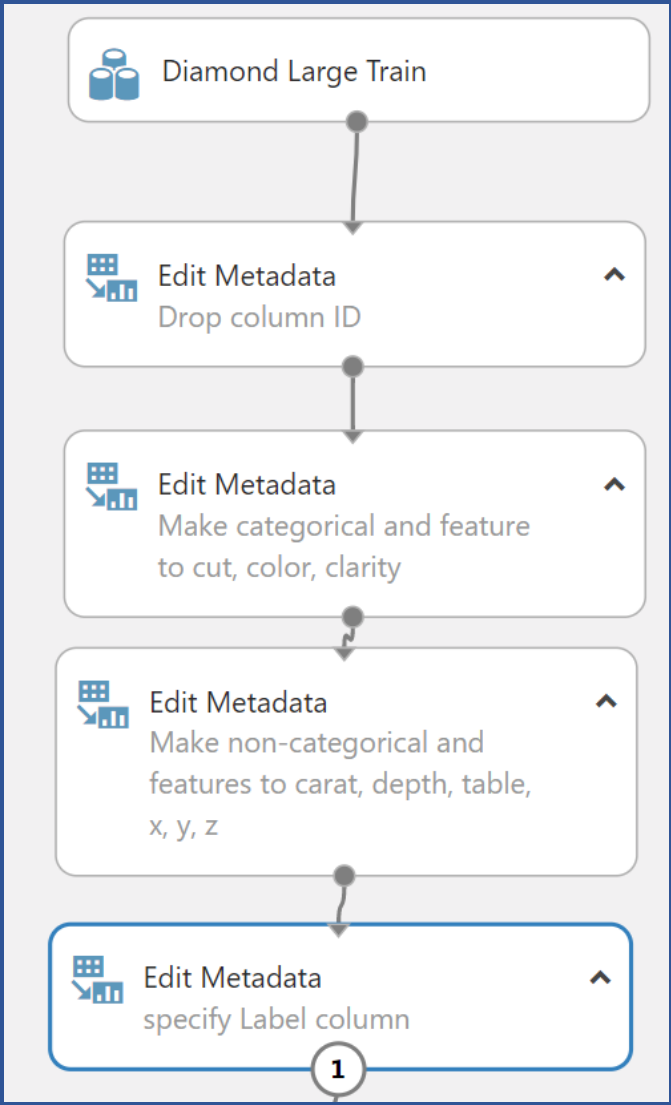
Choose file -> diamonds-Large-Train.csv



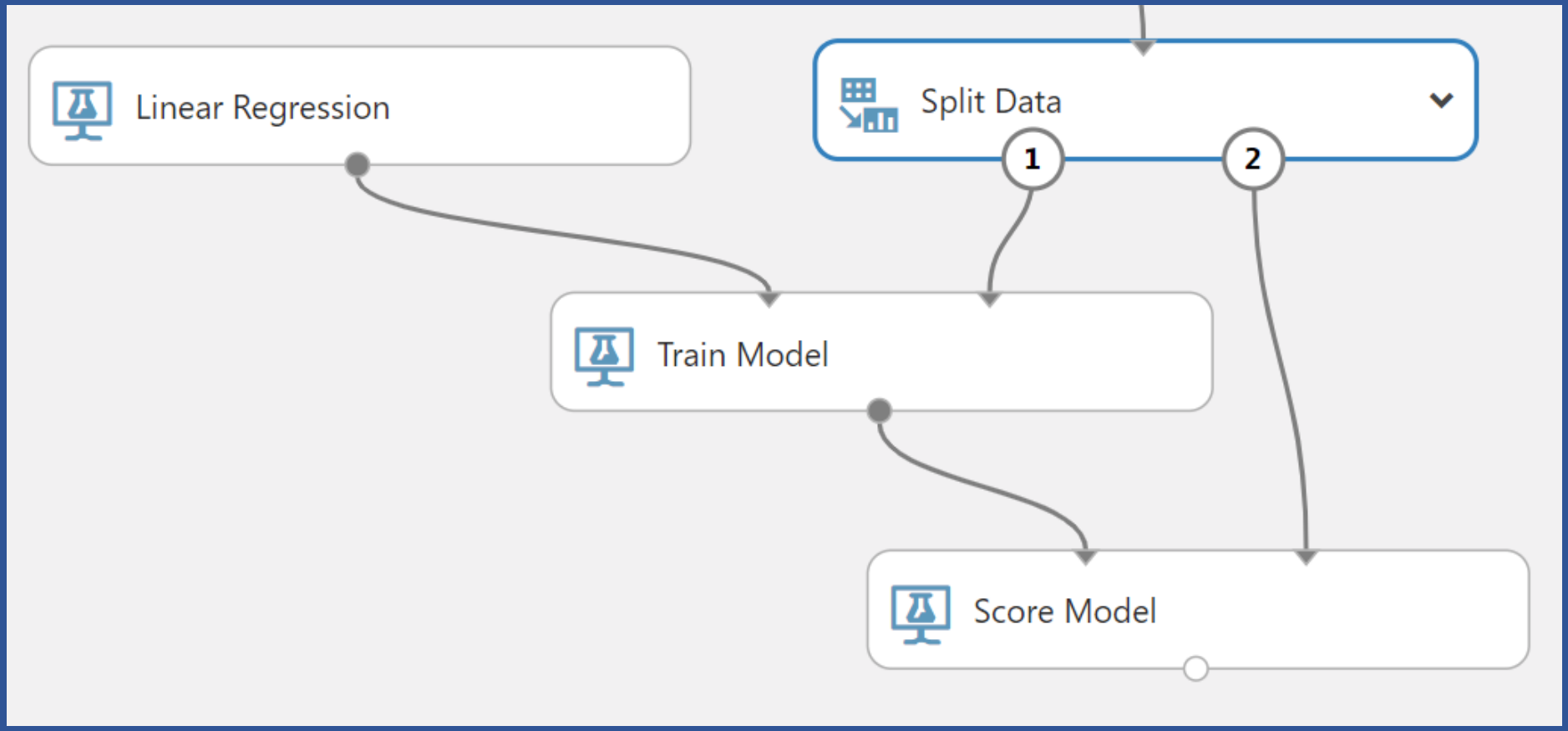
Create ML model

1. Import data
2. Drop column ID
3. Set column Cut, Color, Clarity to Categorical and Feature
4. Set column carat, depth, table, x, y, z to non Categorical and Feature
5. Set label = price
6. Split data 70% train 30% score
7. Train and score

Create ML model



Train and score



Next step

Create AutoML of Diamond prediction