Microsoft

MLADS

MACHINE LEARNING, AI,
AND DATA SCIENCE CONFERENCE

November 14- 18, 2022

1

Microsoft

Data Engineering And the Data Science Process

Shep Sheppard
shep.sheppard@microsoft.com
https://www.linkedin.com/in/sqlshep/
@sqlshep

2

Microsoft

Introduction to Data Engineering

Shep Sheppard
Customer Engineer
Fast Track ISV

3

Agenda

- The Data Science Process
- Intro to data engineering
- Scalers
- Statistics

4

4

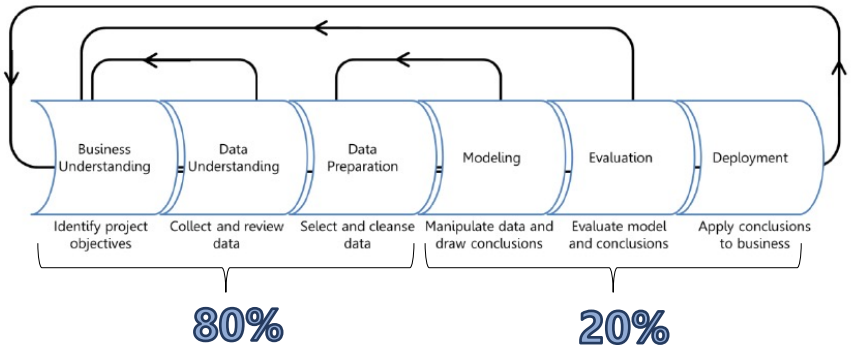
Session goals

- Gain an understanding of the Data Science Process
- Understand the need for data engineering
- Understand a few basic statistical methods

5

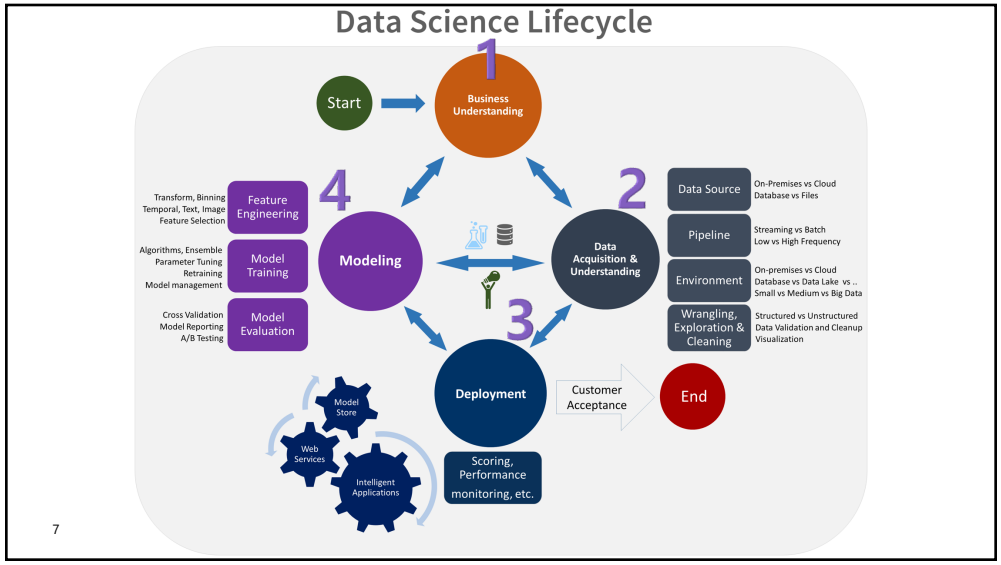
5

The Data Science Process – CRISP-DM Cross Industry Standard for Data Mining



6

6



7

Start with Data

- Where did the data come from?
- Is this all the data or a sample of the data?
- Is the sample representative of the population?
- Is the data representative of the Business Question?
- How much engineering is required?

8

8

Why data engineering?

- Where is the data coming from?
- How many sources of data do you have?
- Is the data denormalized?
- Will it be streaming data?
- Is the data time sensitive, does a prediction need to be made now?
- Online predictions vs batch predictions?
- Data cannot be passed into a model as is in most cases
- String data must be converted to numeric data
- Categorical data must be one hot encoded
- Image data may be processed as a numpy array

9

9

The EPA Notebook

The business problem?

Can we predicate the MPG of a vehicle with the data submitted to the EPA?

05-Linear Regression.ipynb

10

10

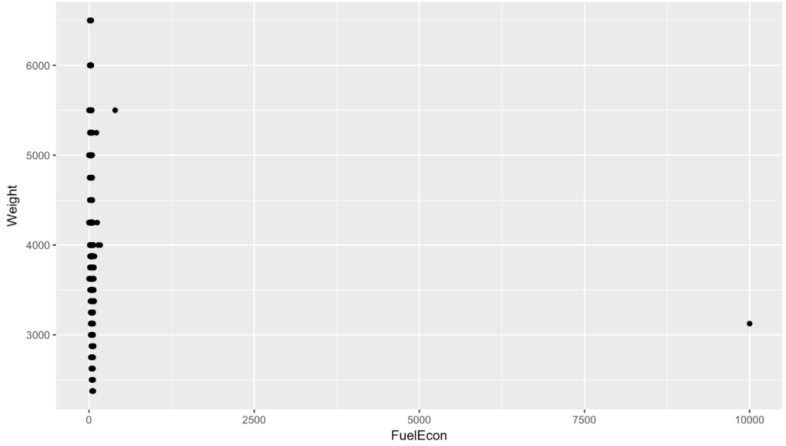
The data

.Veh.Make	Model	Vehicle.Type	HorsePower	Cylinders	Tested.Transmission.Type.Code	Tested.Transmission.Type	Gears	Drive.System.Code	Weight	AxleRatio
ston Martin	Rapide S	Car	552	12	SA	Semi-Automatic	8	R	4750	2.73
ston Martin	Vanquish	Car	568	12	SA	Semi-Automatic	8	R	4500	2.73
BENTLEY	Continental GT	Car	616	12	SA	Semi-Automatic	8	F	6000	2.85
BENTLEY	Continental GT	Car	616	12	SA	Semi-Automatic	8	F	6000	2.85
BMW	230i Convertible	Car	248	4	SA	Semi-Automatic	8	R	4000	2.81
BMW	230i Coupe	Car	248	4	M	Manual	6	R	3625	3.91
BMW	230i Coupe	Car	248	4	SA	Semi-Automatic	8	R	3625	2.81
BMW	230i xDrive Convertible	Car	248	4	SA	Semi-Automatic	8	R	4000	2.81
BMW	230i xDrive Coupe	Car	248	4	SA	Semi-Automatic	8	R	3750	2.81
BMW	320i	Both	181	4	A	Automatic	8	R	3625	3.20

11

11

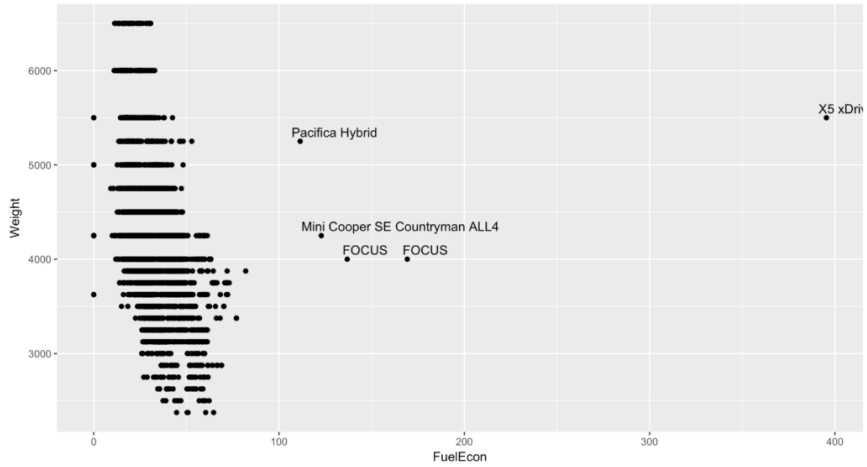
Look for bad data/outliers?



12

12

More outliers?



13

13

Is there data that should be removed?

Represented.Test.Veh.Make	Model	Vehicle.Type	HorsePower	Cylinders	Tested.Transmission.Type.Code	Tested.Transmission.Type
Aston Martin	Rapide S	Car	552	12	SA	Semi-Automatic
Aston Martin	Vanquish	Car	568	12	SA	Semi-Automatic
BENTLEY	Continental GT	Car	616	12	SA	Semi-Automatic
BENTLEY	Continental GT	Car	616	12	SA	Semi-Automatic
BMW	230i Convertible	Car	248	4	SA	Semi-Automatic
BMW	230i Coupe	Car	248	4	M	Manual
BMW	230i Coupe	Car	248	4	SA	Semi-Automatic
BMW	230i xDrive Convertible	Car	248	4	SA	Semi-Automatic
BMW	230i xDrive Coupe	Car	248	4	SA	Semi-Automatic
BMW	320i	Both	181	4	A	Automatic

14

14

What to do with Categorical data?

Represented.Test.Veh.Make	Model	Vehicle.Type	HorsePower	Cylinders	Tested.Transmission.Type.Code	Tested.Transmission.Type
Aston Martin	Rapide S	Car	552	12	SA	Semi-Automatic
Aston Martin	Vanquish	Car	568	12	SA	Semi-Automatic
BENTLEY	Continental GT	Car	616	12	SA	Semi-Automatic
BENTLEY	Continental GT	Car	616	12	SA	Semi-Automatic
BMW	230i Convertible	Car	248	4	SA	Semi-Automatic
BMW	230i Coupe	Car	248	4	M	Manual
BMW	230i Coupe	Car	248	4	SA	Semi-Automatic
BMW	230i xDrive Convertible	Car	248	4	SA	Semi-Automatic
BMW	230i xDrive Coupe	Car	248	4	SA	Semi-Automatic
BMW	320i	Both	181	4	A	Automatic

15

15

One hot-encode categorical data

· What was a datapoing in a column becomes a column

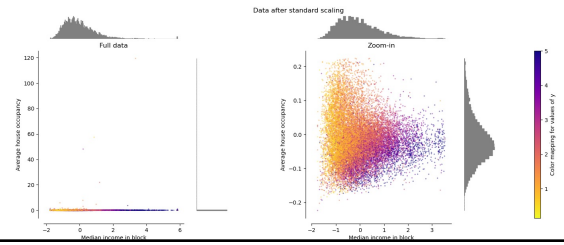
Tested_Transmission_Type_Code_A	Tested_Transmission_Type_Code_AM	Tested_Transmission_Type_Code_AMS	Tested_Transmission_Type_Code_CVT
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
...	...	...	...
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

16

16

Scalers?

- In general, learning algorithms benefit from standardization of the data set
- Standardization of datasets is a common requirement for many machine learning estimators
- Models might behave badly if the individual features do not more or less look like standard normally distributed data



17

17

sklearn.preprocessing.MaxAbsScaler

- Scale each feature by its maximum absolute value
- This estimator scales and translates each feature individually such that the maximal absolute value of each feature in the training set will be 1.0.
- It does not shift/center the data, and thus does not destroy any sparsity

```
from sklearn.preprocessing import MaxAbsScaler

X = [[ 10., -10.,  200.],
     [ 30.,  40.,  50.],
     [ 0.,  1., -1.]]

transformer = MaxAbsScaler().fit(X)
transformer

transformer.transform(X)
```

```
array([[ 0.33333333, -0.25      ,  1.        ],
       [ 1.        ,  1.        ,  0.25      ],
       [ 0.        ,  0.025     , -0.005     ]])
```

18

18

sklearn.preprocessing.StandardScaler

- Standardize features by removing the mean and scaling to unit variance.

```
from sklearn.preprocessing import StandardScaler
data = [[ 10., -10.,  200.],
     [ 30.,  40.,  50.],
     [ 0.,  1., -1.]]
scaler = StandardScaler()
print(scaler.fit(data))

print(scaler.mean_)

print(scaler.transform(data))
```

```
StandardScaler(copy=True, with_mean=True, with_std=True)
[13.33333333 10.33333333 83.        ]
[[-0.26726124 -0.94781764  1.37144955]
 [ 1.33630621  1.38288148 -0.3868191 ]
 [-1.06904497 -0.43506384 -0.98463045]]
```

19

sklearn.preprocessing.normalize

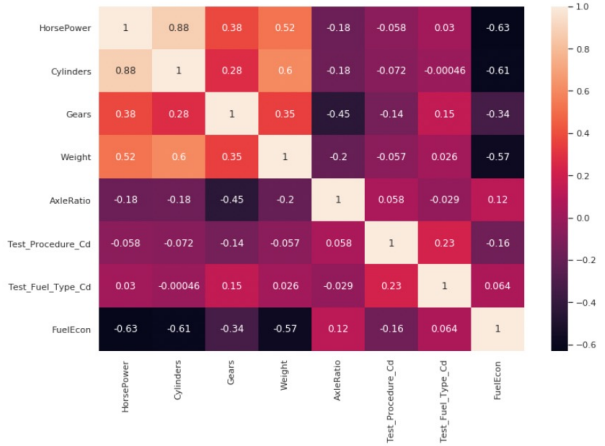
- **Normalization** is the process of **scaling individual samples to have unit norm**

```
from sklearn.preprocessing import normalize

X = [[ 10., -10.,  200.],
     [ 30.,  40.,  50.],
     [ 0.,  1., -1.]]
X_normalized = normalize(X, norm='l2')

X_normalized
```

```
array([[ 0.04987547, -0.04987547,  0.99750934],
       [ 0.42426407,  0.56568542,  0.70710678],
       [ 0.        ,  0.70710678, -0.70710678]])
```

20

Perform statistical summaries of the data, Correlations



21

21

Univariate Statistics

	RowNumber	HorsePower	Cylinders	Gears	Weight	AxleRatio	Test.Procedure.Cd	Test.Fuel.Type.Cd	FuelEcon
count	1034.000000	1034.000000	1034.000000	1034.000000	1034.000000	1034.000000	1034.000000	1034.000000	1034.000000
mean	521.361702	291.824952	5.431335	6.509671	4191.852031	3.411064	24.993230	56.366538	28.216538
std	300.241933	144.294932	1.905214	1.992824	787.821434	0.586484	22.078601	11.633650	9.496233
min	1.000000	72.000000	3.000000	1.000000	2375.000000	1.560000	2.000000	19.000000	9.200000
25%	262.250000	181.000000	4.000000	6.000000	3625.000000	3.070000	11.000000	61.000000	21.525000
50%	520.500000	271.500000	4.000000	7.000000	4000.000000	3.360000	21.000000	61.000000	26.800000
75%	781.750000	355.000000	6.000000	8.000000	4750.000000	3.700000	31.000000	61.000000	33.400000
max	1040.000000	1500.000000	16.000000	10.000000	6500.000000	5.440000	95.000000	61.000000	71.600000

22

22

Other data issues

- Do you need to deal with dates? Time Series
- Do you have location data, lat/long zip code / FIPS code
- Image data has its own techniques and processes
- Down sampling the image?
- Convert it to black and white image?
- Don't underestimate how long the data engineering may take
- Will the prediction need to be correlated to a datapoint later

23

23

Thank you for attending the MLADS Conference and helping to build a strong community

To find recordings, presentations, and other resources from the event, go to: http://aka.ms/november2022mlads

More information about the AI & ML Connected Community: http://aka.ms/aiml-cc

We want your feedback!

We read each evaluation and incorporate your comments into future MLADS events.
Both sessions and event evaluations will be available within the MyMLADS event portal during the conference dates—to access them, navigate to the **Evaluations tab**.

We request that session evaluations are submitted following each day of programming.

24

24


Q&A
This slide is required—do not delete—please read the notes for this slide, then delete this text box
25

25


Microsoft
© Copyright Microsoft Corporation. All rights reserved.

26