Feature Selection and Extraction in Microsoft Azure

EXPLORING YOUR DATASET FOR FEATURE SELECTION AND EXTRACTION

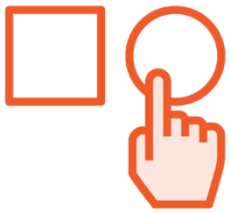
Xavier Morera
HELPING DEVELOPERS UNDERSTAND SEARCH & BIG DATA

@xmorera www.xaviermorera.com


ONE SIZE
DOES NOT
FIT ALL

Define the problem

Prepare the data



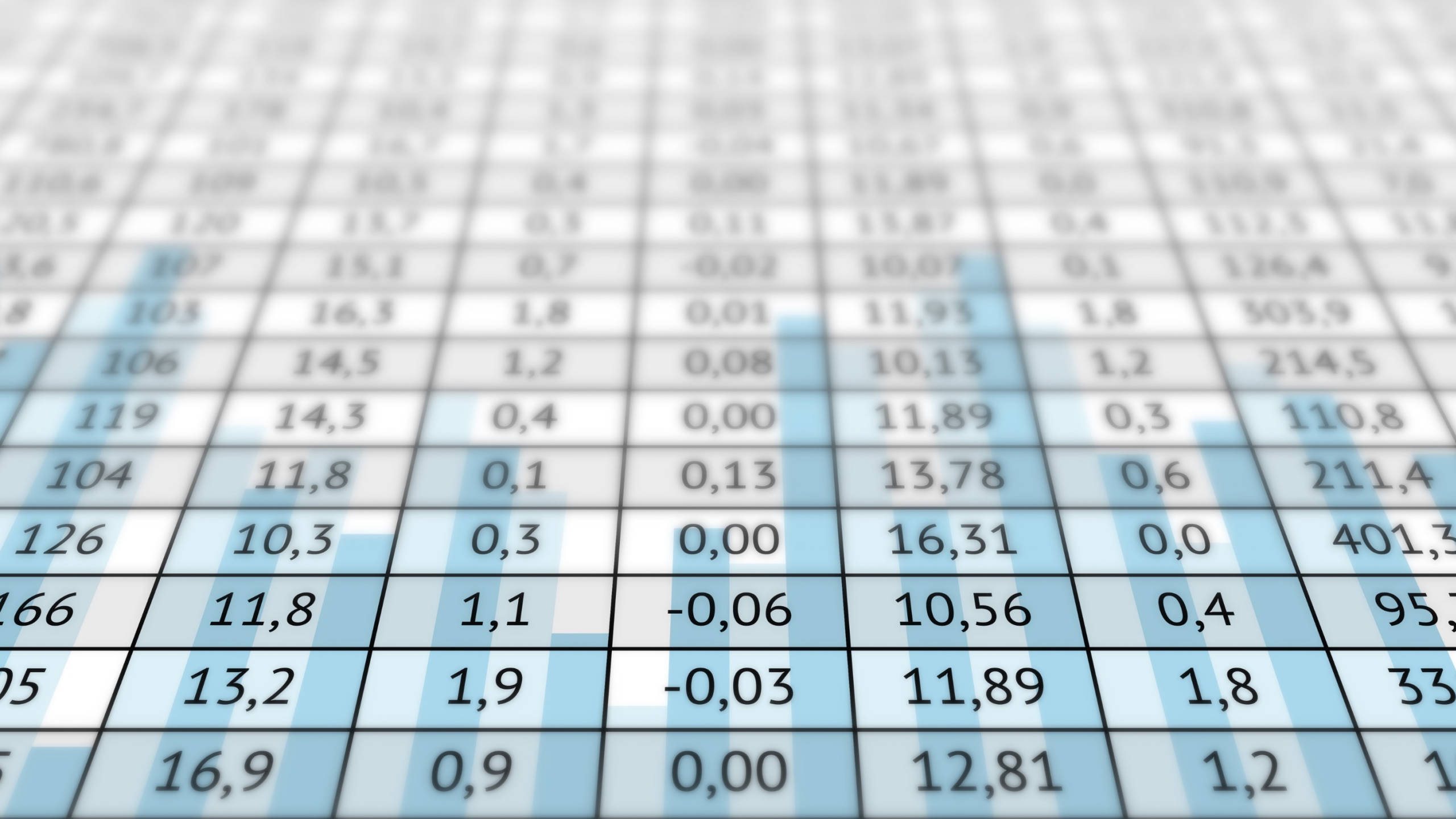
Pick the model

Train the model

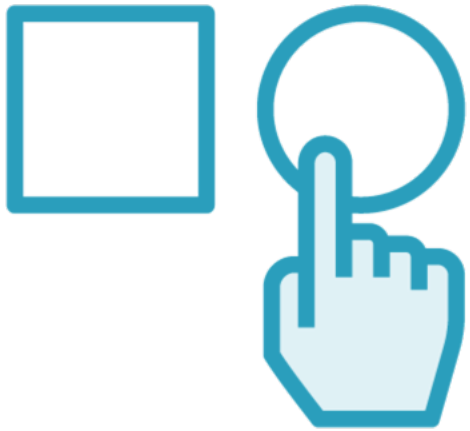
Test the model

Take the model
to production

Prepare the data


106 14,5 1,2 0,08 10,13 1,2 214,5
119 14,3 0,4 0,00 11,89 0,3 110,8
104 11,8 0,1 0,13 13,78 0,6 211,4
126 10,3 0,3 0,00 16,31 0,0
11,8 1,1 -0,06 10,56 0,4
13,2 1,9 -0,03 11,89 1,8
16,9 0,9 0,00 12,81 1,2

Features



Feature Selection

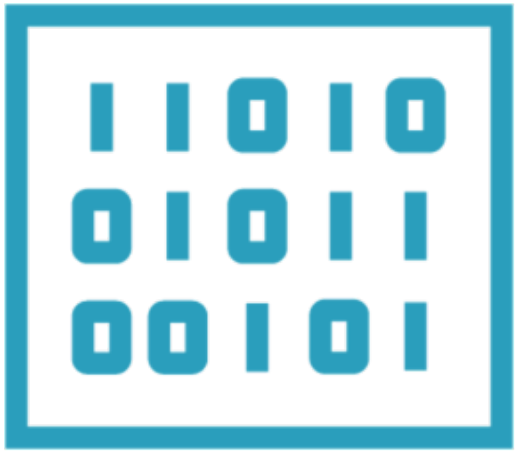
Feature Extraction

Feature Normalization

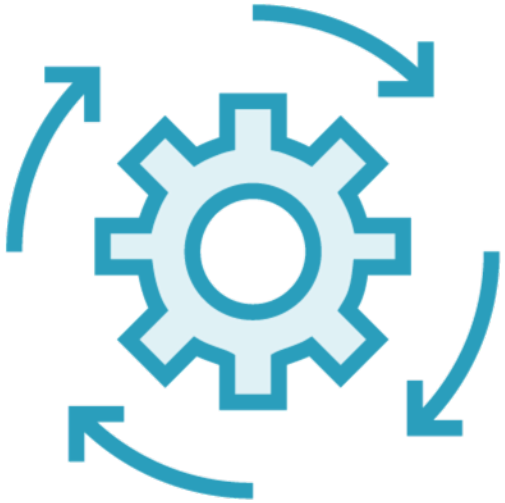
Feature

Individual measurable property or characteristic of a phenomenon being observed

Solving a Problem with Traditional Software



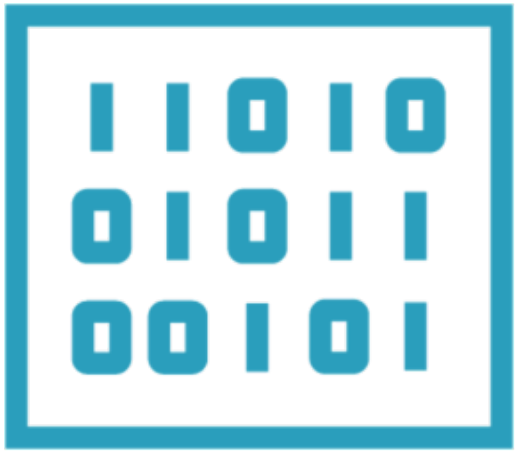
Data



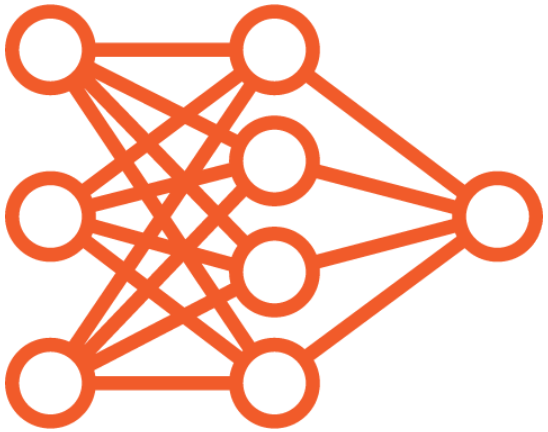
Apply a set of steps

Got a result

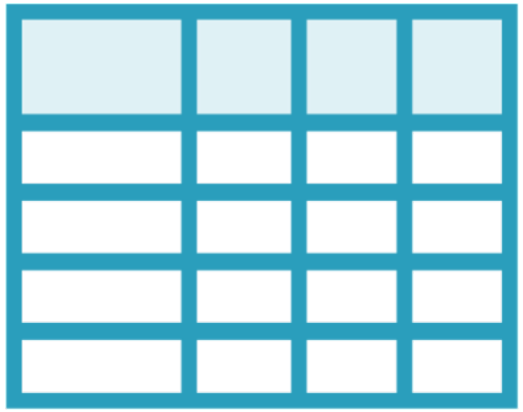
Solving a Problem with Machine Learning



Data

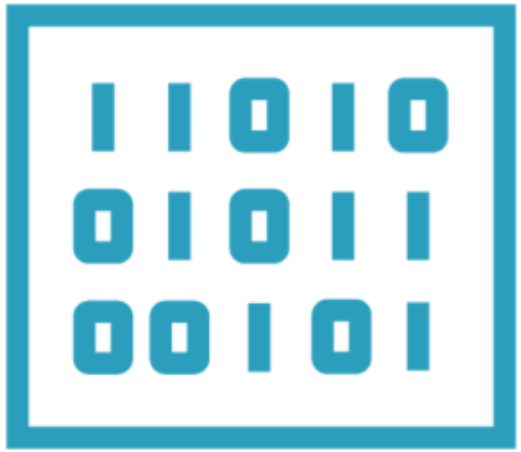


ML model



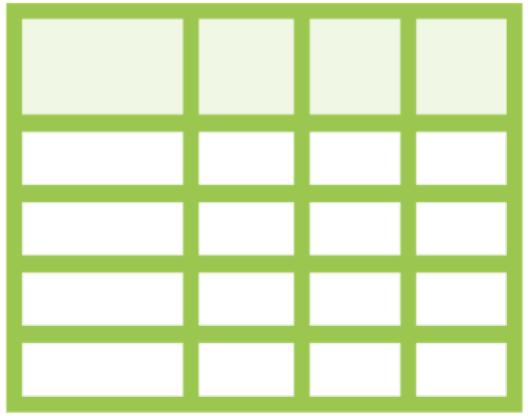
Expected results

Input Data



Features

Features



Text

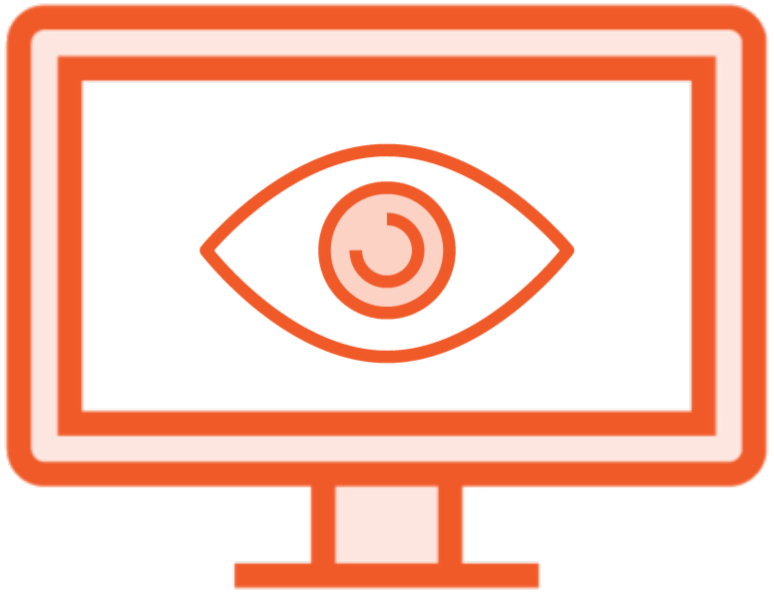
Images

Audio

Predict House Prices

Rooms	Area	Pool	Built	Price	Location
3	500	Yes	1980	200	SF
2	250	No	1999	100	Miami
3	300	Yes	1985	150	Seattle

Computer Vision



Feature extraction algorithm

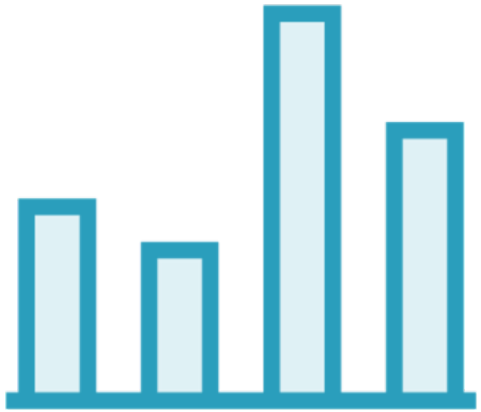
Not every feature can help improve the model

Important

Exploring and Identifying the Distribution of Your Data

Know Your Data

Statistics



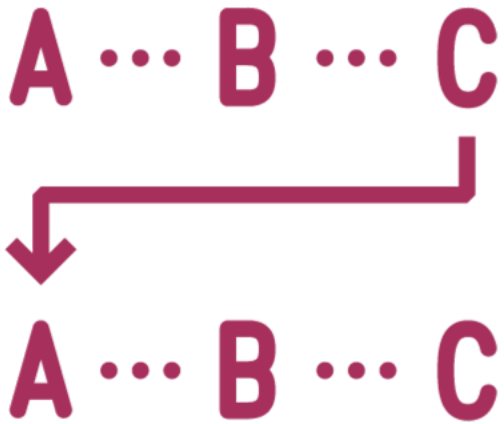
Histograms

Tendencies

What else?

Clean Your Data

Missing Values



Redundancy

Tidying Up

Other Steps

Categorize Your Problem

Input Data

Expected Output

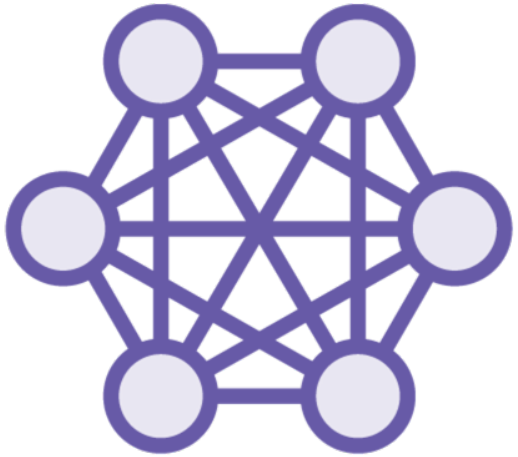
Answers

Quantity

Selecting the Right Model for Your Data

Regression

Classification



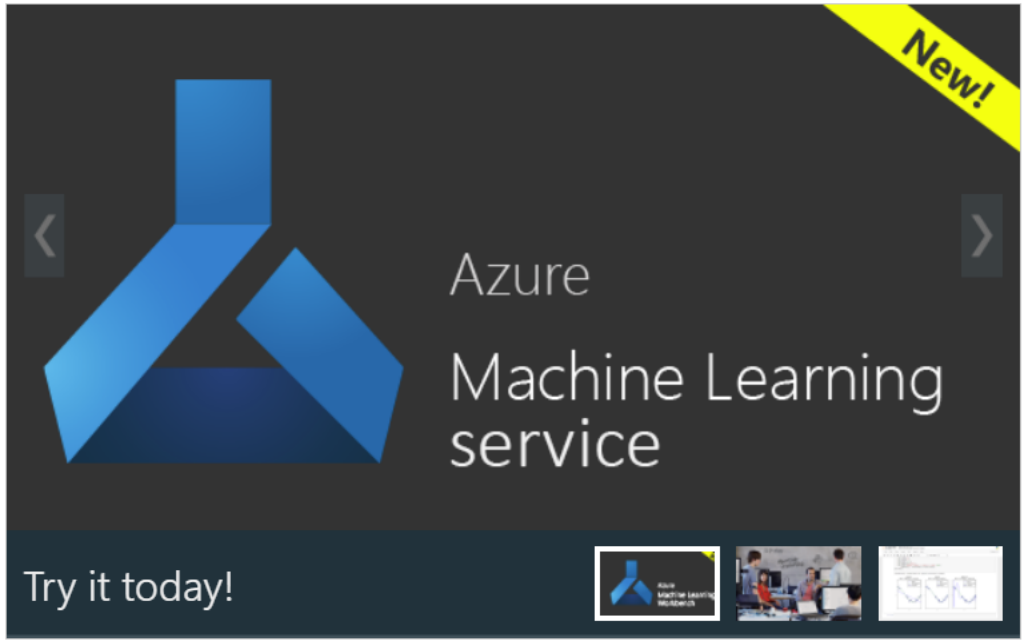
Clustering

Other Type of Model

Problem Solving

- [x] Define Problem
- [x] Develop a Plan
- [x] Implement Plan
- [x] Evaluate


T
U
V
W
XYZ
Q
R
S
O
P
Mc
N
M
K
L
I
J
H
F
G
D
E
C
B
A



Welcome to Azure Machine Learning Studio (classic)

Try it for free

No Azure subscription? No credit card? No problem! Choose anonymous Guest Access, or sign in with your work or school account, or a Microsoft account.

Sign In

Not an Azure ML Studio (classic) user?
Sign up here

Pricing & FAQ

Announcements NEW!

Azure Machine Learning Studio R Runtime Upgrade
Aired on October 31, 2018

The R language engine in the Execute R Script module of Azure Machine Learning Studio has added a new R runtime version -- Microsoft R Open (MRO) 3.4.4. MRO 3.4.4 is based on open-source CRAN R 3.4.4 and is therefore compatible with packages that works with that version of R.

Mining Campaign Funds
Aired on August 03, 2017

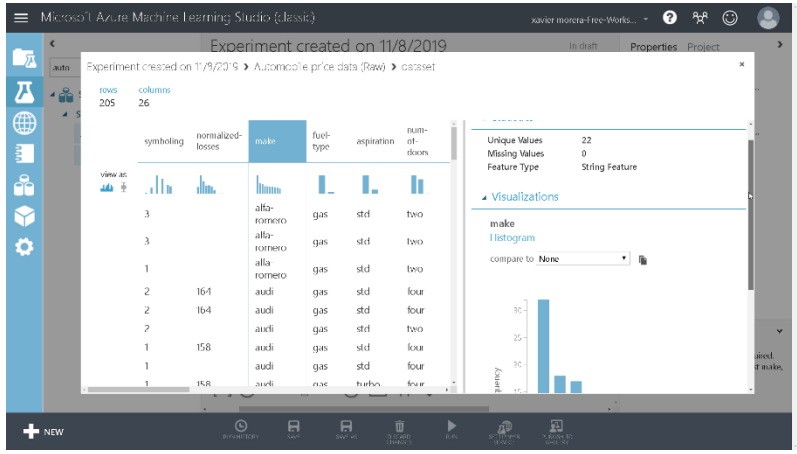
Play with 2016 Presidential Campaign finance data while learning how to prepare a large dataset for machine learning by processing and engineering features. This sample experiment works on a 2.5 GB dataset and will take about 20 minutes to run in its entirety.

> Learn More

Inside the Data Science VM
Aired on June 21, 2016

DSVM is a custom Azure Virtual Machine image that is published on the Azure marketplace and available on both Windows and Linux. It contains several popular data science and development tools both from Microsoft and from the open source community all pre-installed and pre-configured and ready to use. We will cover best practices that would show

Data Table



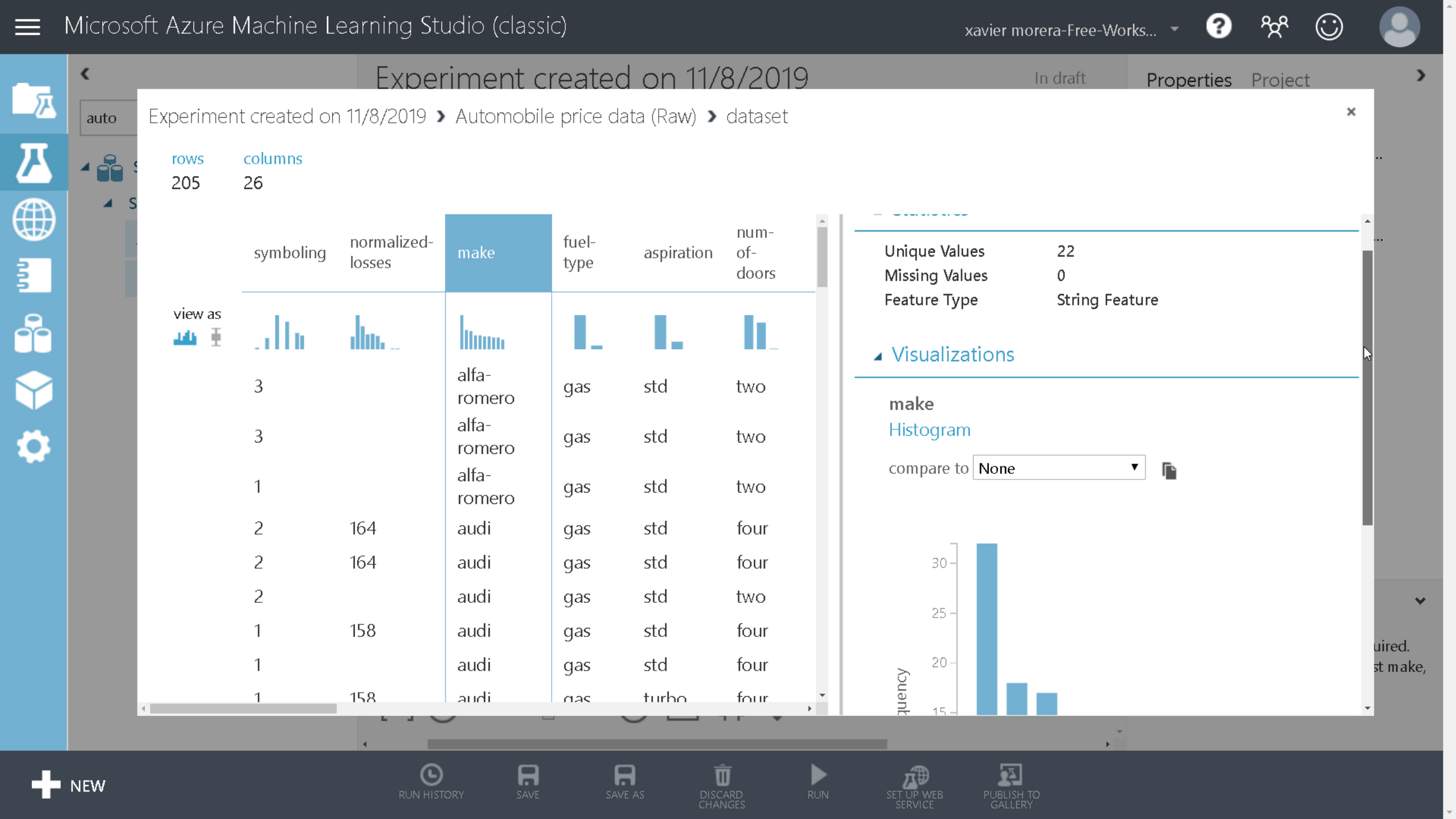
Internal representation

Many types of data

Multiple sources

Converted into Data Table

- And back


Microsoft Azure Machine Learning Studio (classic)
xavier morera-Free-Works...
Experiment created on 11/8/2019
In draft
Properties
Project
auto
Experiment created on 11/8/2019 › Automobile price data (Raw) › dataset
rows
205
columns
26
symboling
normalized-losses
make
fuel-type
aspiration
num-of-doors
view as
3
alfa-romero
gas
std
two
3
alfa-romero
gas
std
two
1
alfa-romero
gas
std
two
2
164
audi
gas
std
four
2
164
audi
gas
std
four
2
audi
gas
std
two
1
158
audi
gas
std
four
1
audi
gas
std
four
Unique Values
22
Missing Values
0
Feature Type
String Feature
Visualizations
make
Histogram
compare to
None
30
25
20
15
NEW
RUN HISTORY
SAVE
SAVE AS
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY

Data Table

Sources

Plain text (.txt)

Comma-separated values (CSV)

Tab-separated values (TSV)

Excel file

Azure table

Hive table

SQL database table

OData values

SVMLight data (.svmlight)

Attribute Relation File Format (ARFF)

Zip file (.zip)

R object or workspace file (.RData)

Data types Allowed

String

Integer

Double

Boolean

DateTime

TimeSpan

Demo

Dataset Exploration Demo

Takeaways

What is a feature?

Different types of features

- Text, image, or audio

Features

- Selection, extraction, & normalization

Data

- Other formats
- Data Table

Explore the data