

Machine Learning (ML) Workshop

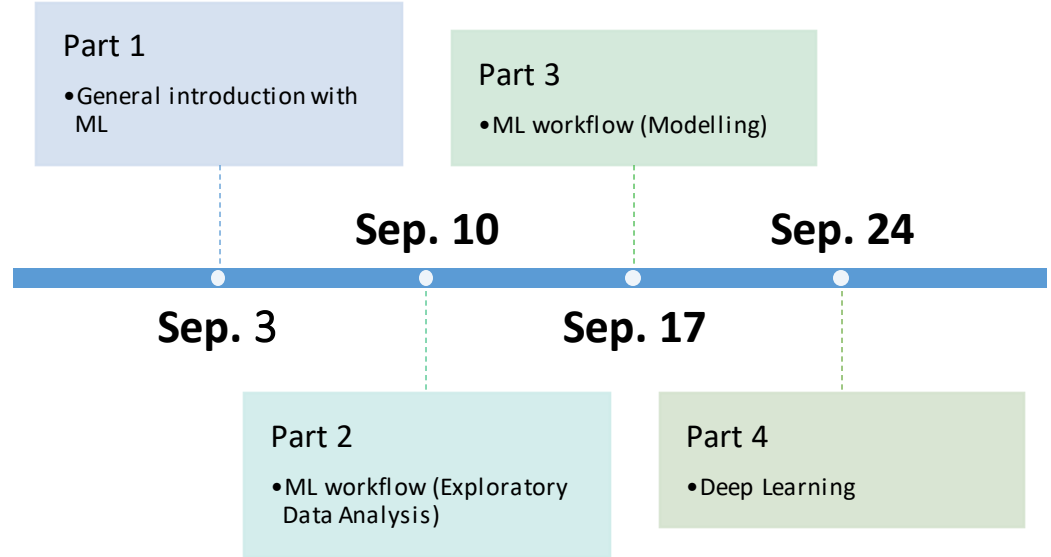


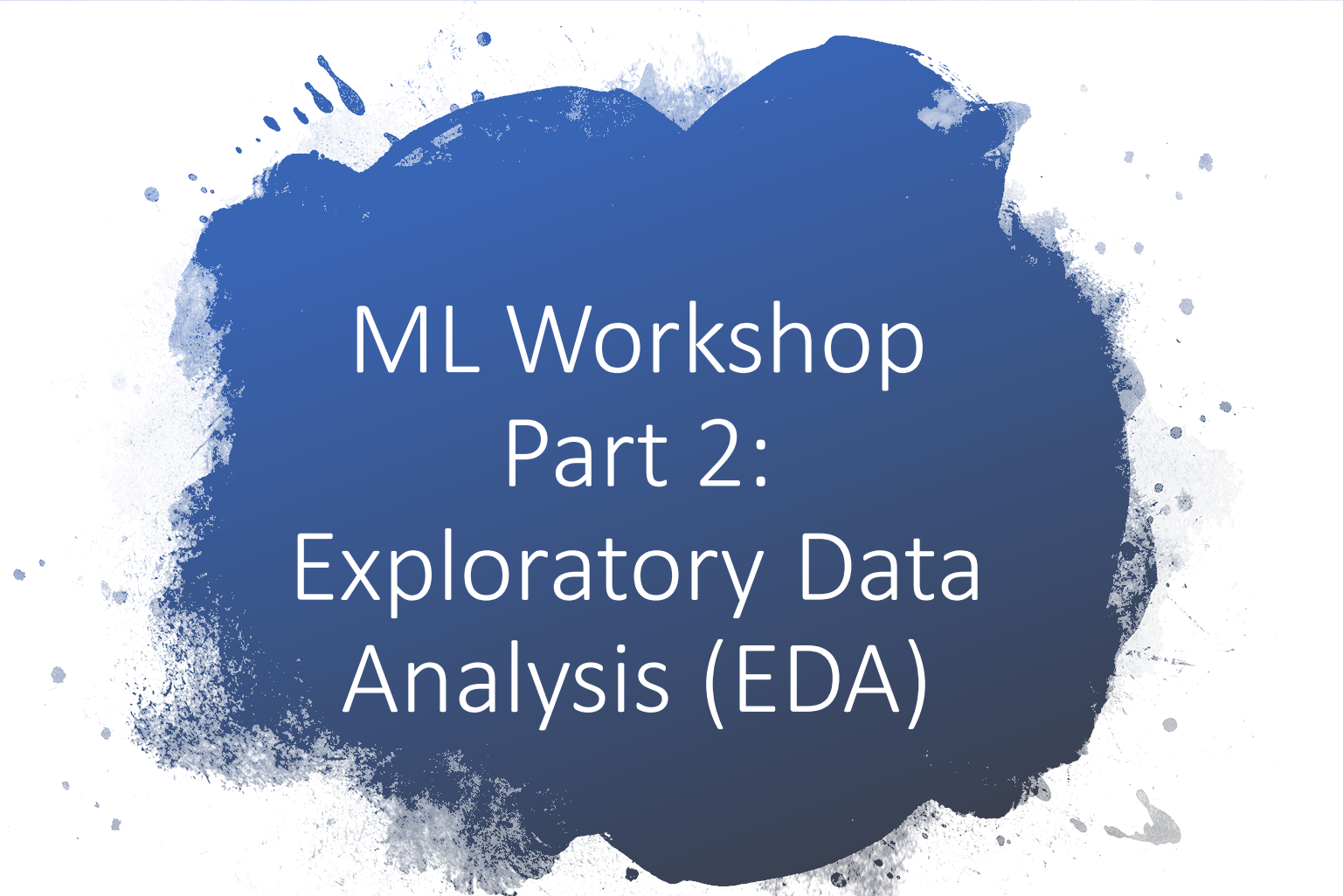
Dr Sara Soltaninejad
Fall 2021

Who am I?

- Sara Soltaninejad
- PhD in MRC, Computing Science Department, UofA, 2016-2020
 - PhD Thesis: Intelligent Parkinson's Disease Classification and Progress Monitoring Using Non-Invasive Techniques
 - Supervisors:
 - Dr Anup Basu
 - Dr Irene Cheng
- ML Developer AltaML, 2020-Now

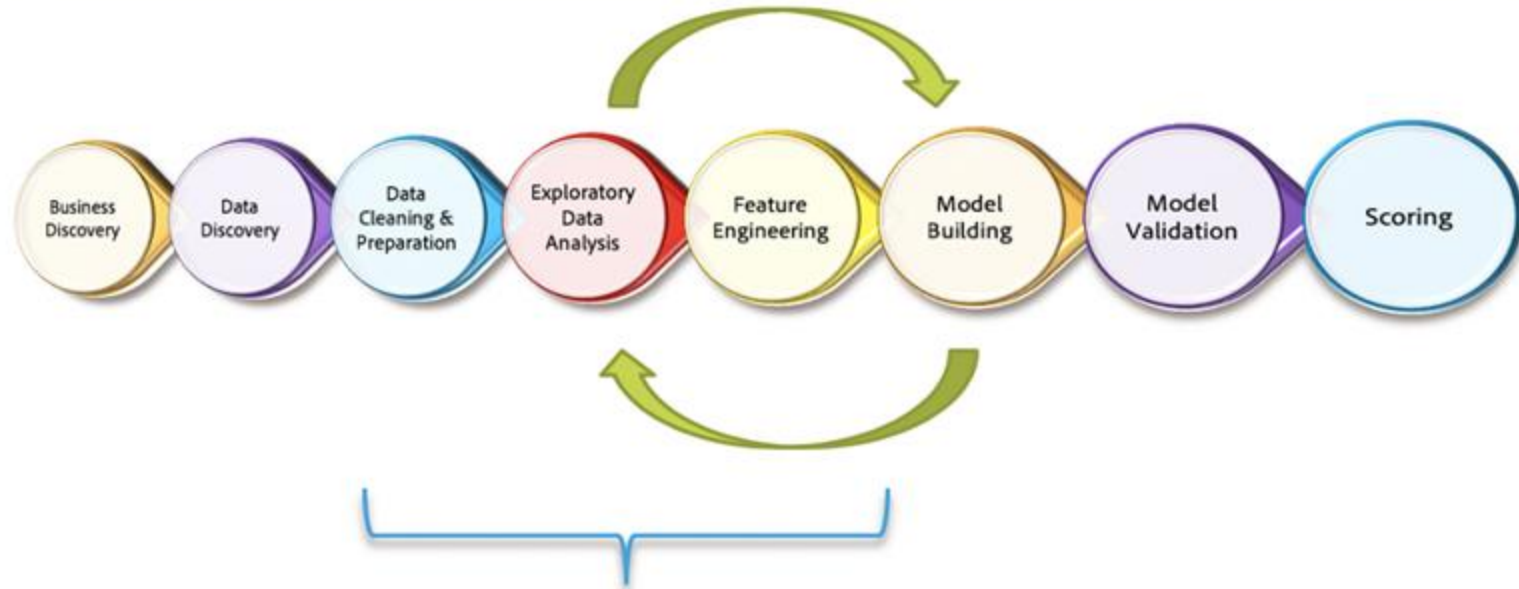
ML Workshop Outline





ML Workshop Part 2: Exploratory Data Analysis (EDA)

Where does EDA fit into a data science project



Why we need exploratory data analysis

How data is capable
of solving important
business problems is
essential and key
skill of a data
scientist



Find relationships of
different attributes



Accurate predictions
through right data

EDA is the best way
to gain knowledge
about any data for
any given business
problem



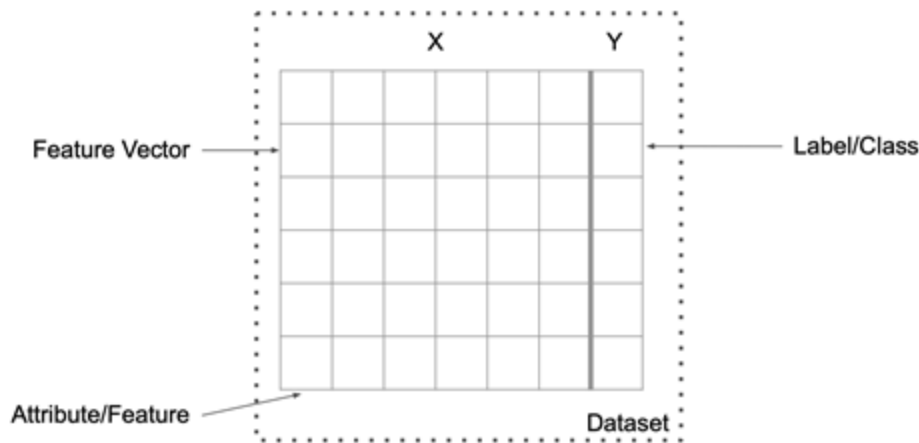
Stable predictive models
by right choice of
attributes



Valuable business
insights and data driven
marketing strategies

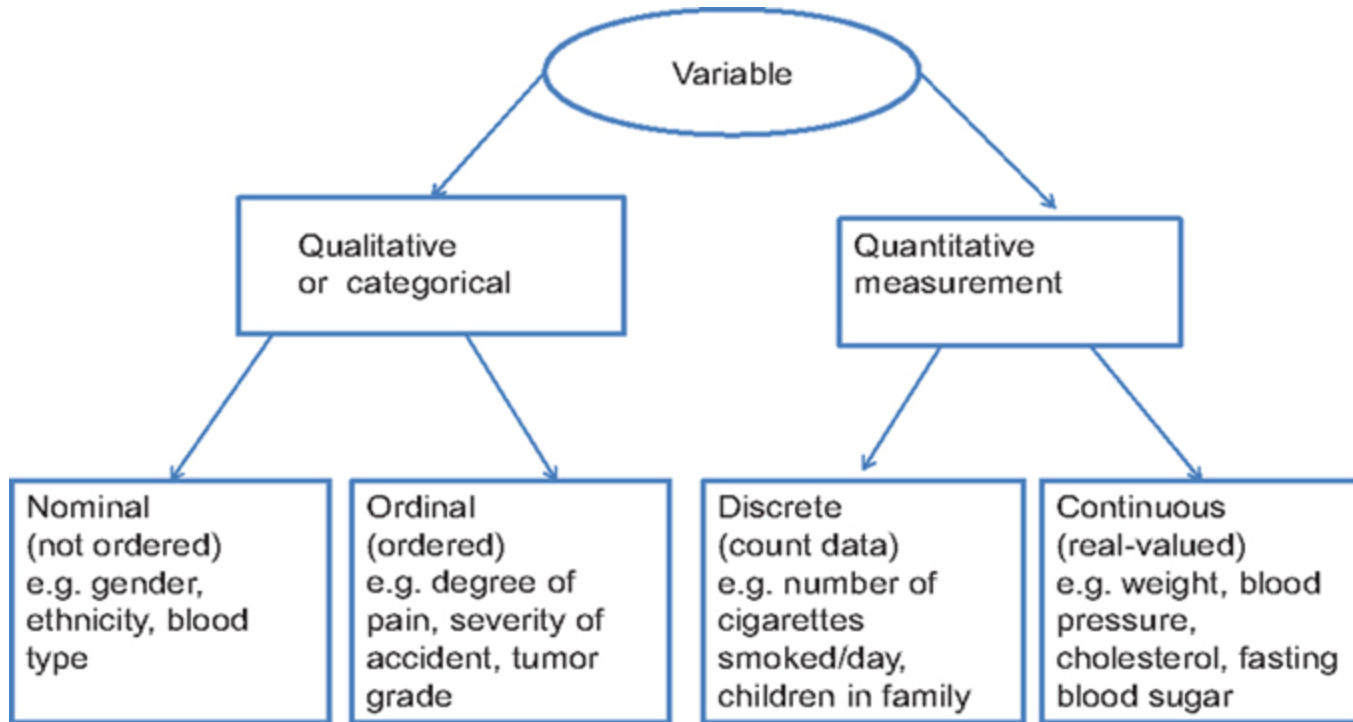
ML Input: Data Frame Terminology

- Feature/Attribute: A single variable (binary, nominal, numerical)
- Instance/Feature vector: One entity described by features
- Label/Class/Target Variable: An extra information that categorizes/classifies a given instance
- Dataset: Collection instances



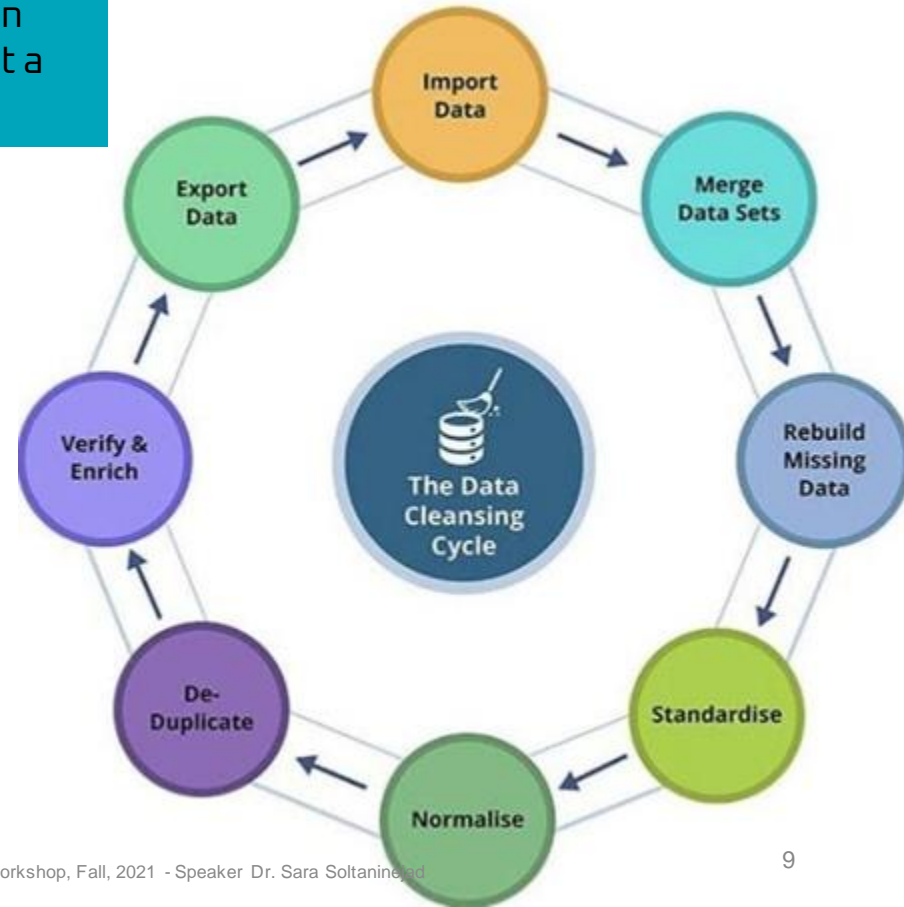
	state	color	food	age	height	score
Jane	NY	blue	Steak	30	165	4.6
Niko	TX	green	Lamb	2	70	8.3
Aaron	FL	red	Mango	12	120	9.0
Penelope	AL	white	Apple	4	80	3.3
Dean	AK	gray	Cheese	32	180	1.8
Christina	TX	black	Melon	33	172	9.5
Cornelia	TX	red	Beans	69	150	2.2

Data Frame Variable Data Types

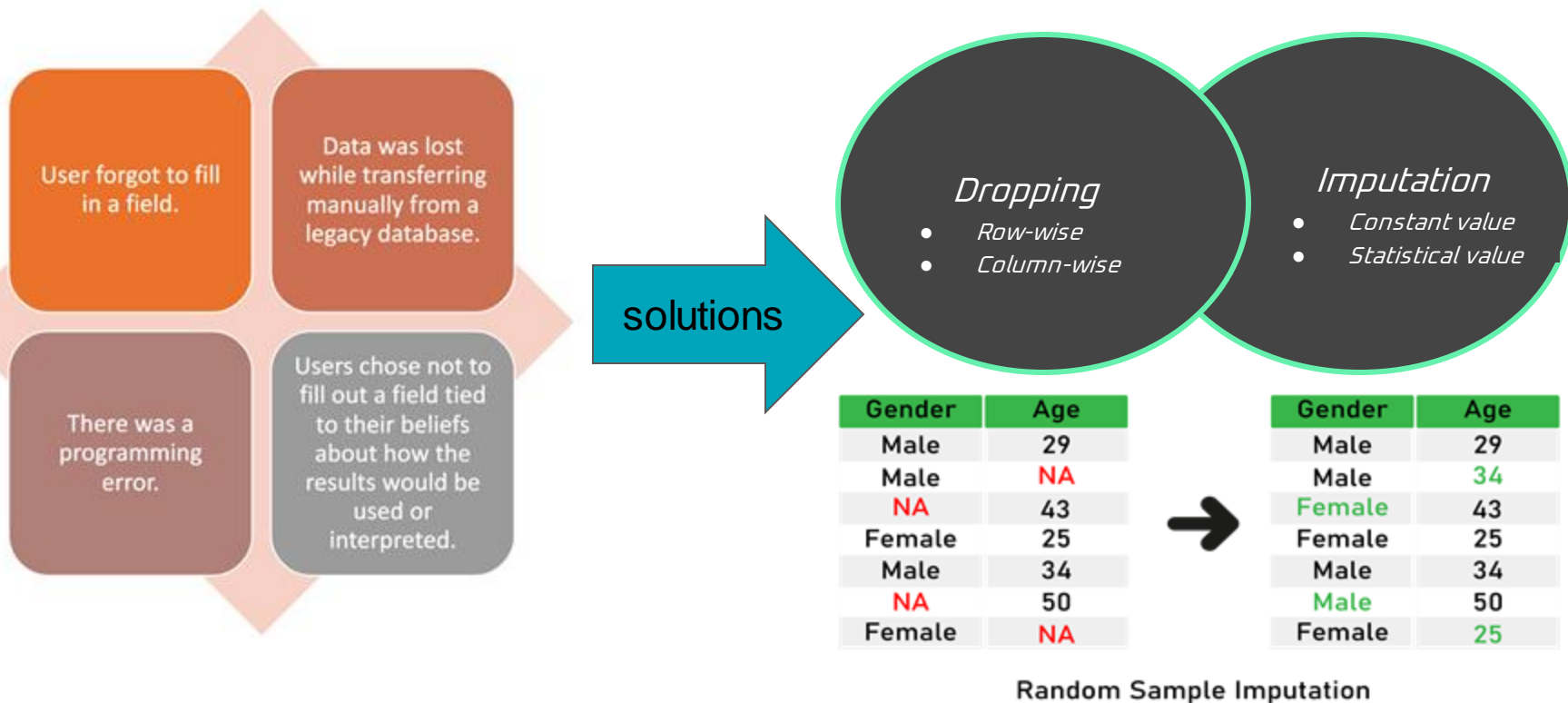


EDA helps us understand various facets of our data. In this step, we analyze different attributes of data, uncover interesting insights, and even visualize data on different dimensions to get a better understanding.

Data Preprocessing



Preprocessing - Missing Values



Preprocessing - outliers

In statistics, an outlier is an observation point that is distant from other observations. Possible reasons for outliers are recording errors, unusual sampling and laboratory procedures or conditions.

Outlier
Detection

Data Visualization

- *Box plot*
- *Scatter plot*

Math Analysis

- *Z score*
- *Quantile Analysis*

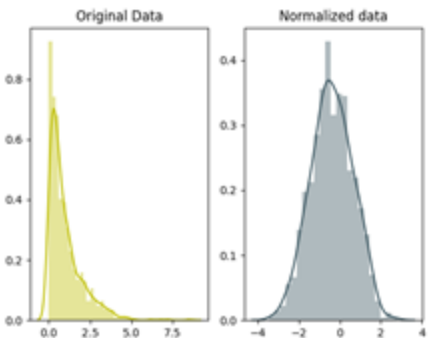
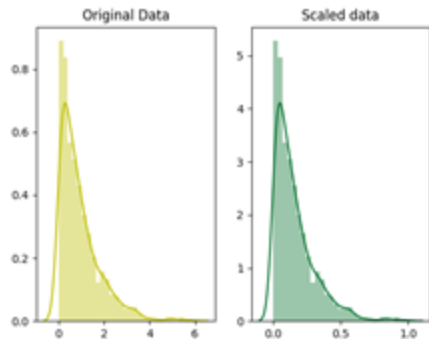
Outlier
Removal

Dropping

Imputation

Preprocessing - Scaling

Data may contain attributes with a mixture of scales for various quantities such as dollar, kilogram, and sales volume.



$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Scaling for
numerical variable

$$x_{scaled} = \frac{x - mean}{sd}$$

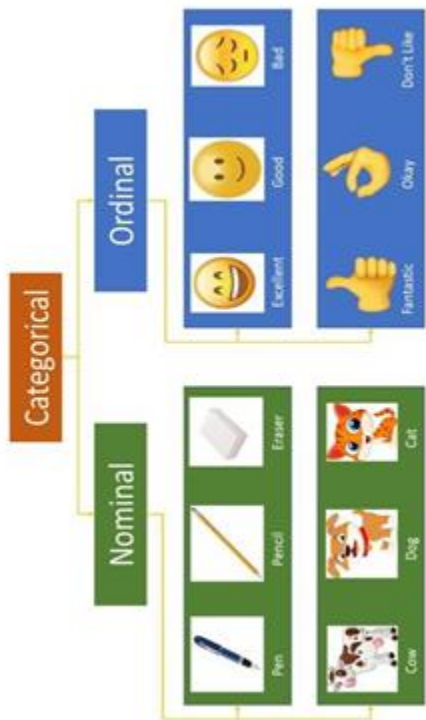
Normalization

Sensitive to outliers

Standardization

Values are not bounded.

Preprocessing - Scaling



Scaling for
categorical
variable

Label Encoding

- *Dataset label*
- *Ordinal Variable*

Careful
about
variable
order!!

*One Hot Encoding
Nominal Variables*

Breakfast
Every day
Never
Rarely
Most days
Never



Breakfast
3
0
1
2
0

color
red
green
blue
red



color_red	color_blue	color_green
1	0	0
0	0	1
0	1	0
1	0	0

Feature Creation

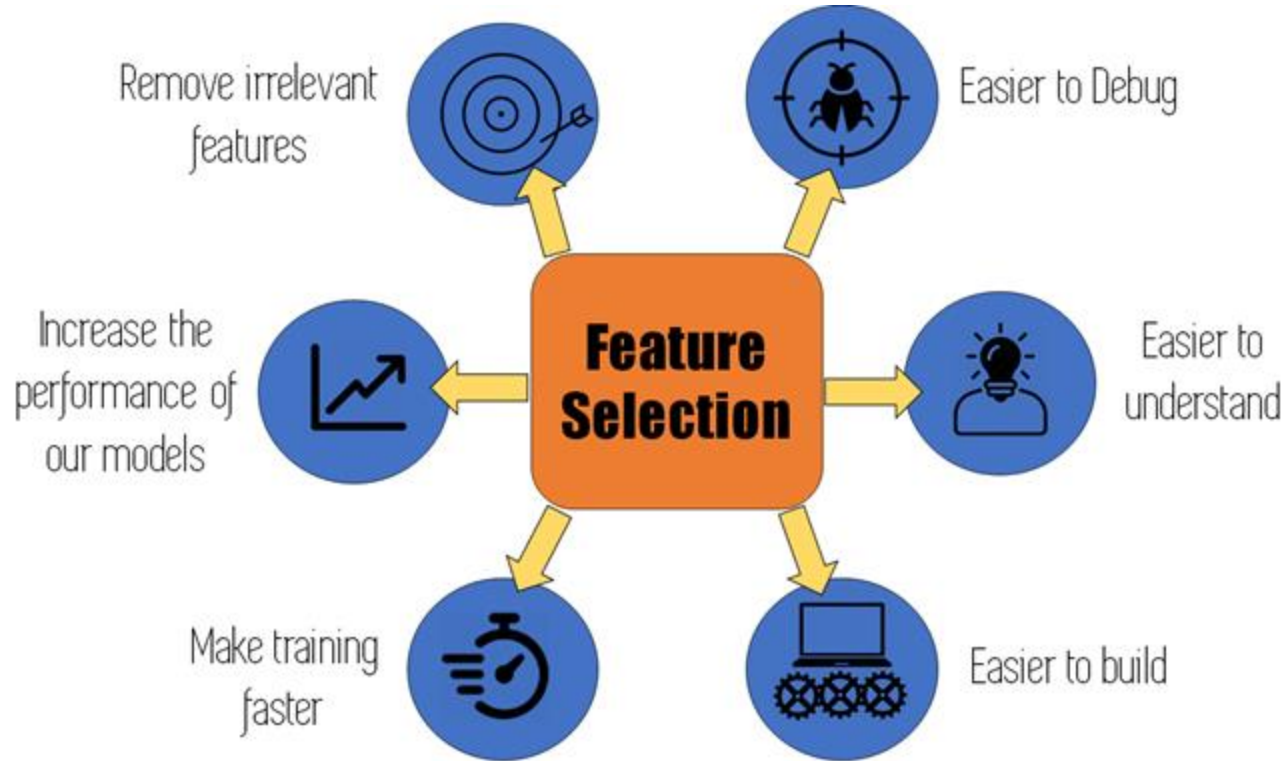
Feature engineering is the process of using domain knowledge to transform raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

New feature

client_id	joined	income	credit_score	join_month	log_income
46109	2002-04-16	172677	527	4	12.059178
49545	2007-11-14	104564	770	11	11.557555
41480	2013-03-11	122607	585	3	11.716739
46180	2001-11-06	43851	562	11	10.688553
25707	2006-10-06	211422	621	10	12.261611



Feature Selection



Feature Selection

Filter Methods



Wrapper Methods





Before We
Move
Forward..

This tutorial is working on Data Frame.

It assumes Python version 2.7 or 3.6+

There are 5 key libraries that you will need to install

scipy

numpy

matplotlib

pandas

Sklearn

Others:
Notebook/GoogleColab

Check the libraries version

Import pandas

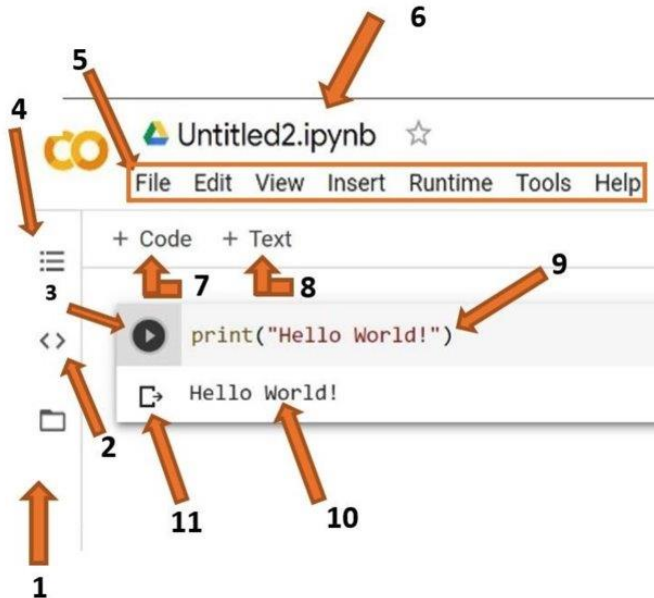
pandas.__version__

Google Colab

- Google Colab is a free Jupyter Notebook environment hosted by Google.
- It has all the features of Jupyterlab and more.
- It is a great platform used by data scientists and machine learning programmers because it takes away the hassle of having to do installations on your own machine.
- Colab has many data science libraries pre-installed and allows you to save your files on Google Drive.

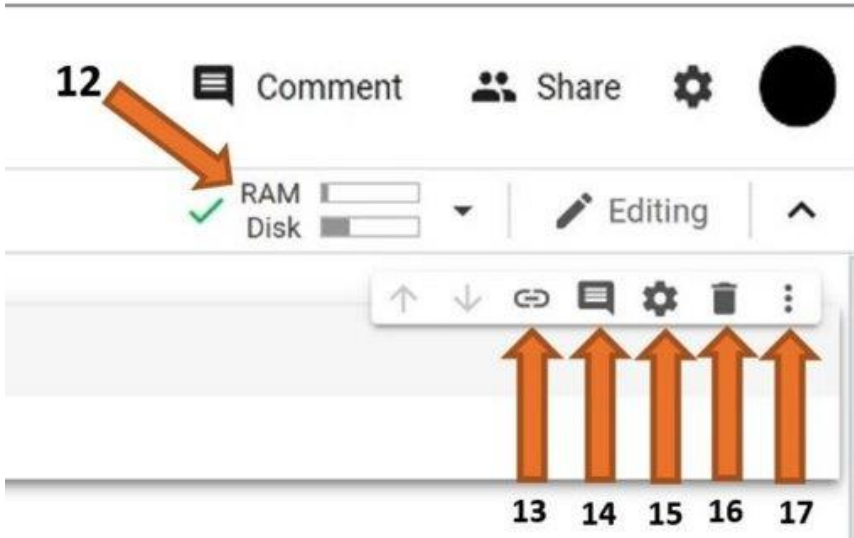


Google Colab



1. **Files:** Here you will be able to upload datasets and other files from both your computer and Google Drive
2. **Code Snippets:** Here you will be able to find prewritten snippets of code for different functionalities like adding new libraries or referencing one cell from another.
3. **Run Cell:** This is the run button. Clicking this will run any code that is inserted in the cell beside it. You can use the shortcut shift+enter to run the current cell and exit to a new one.
4. **Table of Contents:** Here you will be able to create and traverse different sections inside of your notebook. Sections allow you to organize your code and improve readability.
5. **Menu Bar:** Like in any other application, this menu bar can be used to manipulate the entire file or add new files. Look over the different tabs and familiarize yourself with the different options. In particular, make sure you know how to upload or open a notebook and download the notebook (all of these options are under "File").
6. **File Name:** This is the name of your file. You can click on it to change the name. Do not edit the extension (.ipynb) while editing the file name as this might make your file unopenable.
7. **Insert Code Cell:** This button will add a code cell below the cell you currently have selected.
8. **Insert Text Cell:** This button will add a text cell below the cell you currently have selected.
9. **Cell:** This is the cell. This is where you can write your code or add text depending on the type of cell it is.
10. **Output:** This is the output of your code, including any errors, will be shown.
11. **Clear Output:** This button will remove the output.

Google Colab



- 12. **Ram and Disk:** All of the code you write will run on Google's computer, and you will only see the output. This means that even if you have a slow computer, running big chunks of code will not be an issue. Google only allots a certain amount of Ram and Disk space for each *user*, so be mindful of that as you work on larger projects.
- 13. **Link to Cell:** This button will create a URL that will link to the cell you have selected.
- 14. **Comment:** This button will allow you to create a comment on the selected cell. Note that this will be a comment on (about) the cell and not a comment in the cell.
- 15. **Settings:** This button will allow you to change the Theme of the notebook, font type, and size, indentation width, etc.
- 16. **Delete Cell:** This button will delete the selected cell.
- 17. **More Options:** Contains options to cut and copy a cell as well as the option to add form and hide code.

Some General Functions

- Load the dataset, preview it and check general properties

Method	Application	Comment
<code>df = pd.read_csv("nba_all_elo.csv")</code>	Loading the data	
<code>df.head()</code> <code>df.tail()</code>	outputs the first/last five rows of your DataFrame	we could also pass a number as well: <code>movies_df.head(10)</code> would output the top/down ten rows.
<code>pd.set_option("display.max_rows", 100)</code> <code>pd.set_option("display.max_columns", 100)</code>	configure Pandas to display	
<code>df.info()</code>	display all columns and their data types	
<code>df.describe()</code>	basic descriptive statistics for all numeric columns	<code>df.describe(include=np.object)</code> some descriptive statistics
<code>df.shape</code>	fast and useful attribute which outputs just a tuple of (rows, columns)	
<code>df.columns</code>	List of columns in the dataframe	



Q&A

contact me

soltanin@ualberta.ca
Sara@altaml.com

Wants to know more about AltaML
<https://www.altaml.com>



AltaML
Applied AI Lab

*Never
Stop
Learning*



Thank
You