



Workshop on Machine learning – Day1

ML development for classification
Problems

Date: 20th Aug 2021

dubrangala@gmail.com

Day 1 Agenda

- Introduction
- General Steps to Build a ML Model
- Classification Model Use Case
- Exploratory Data Analysis
- General Feature engineering techniques
- Practical Demo using Python



Dayananda Ubrangala

Statistician and Data Science Specialist

Publications



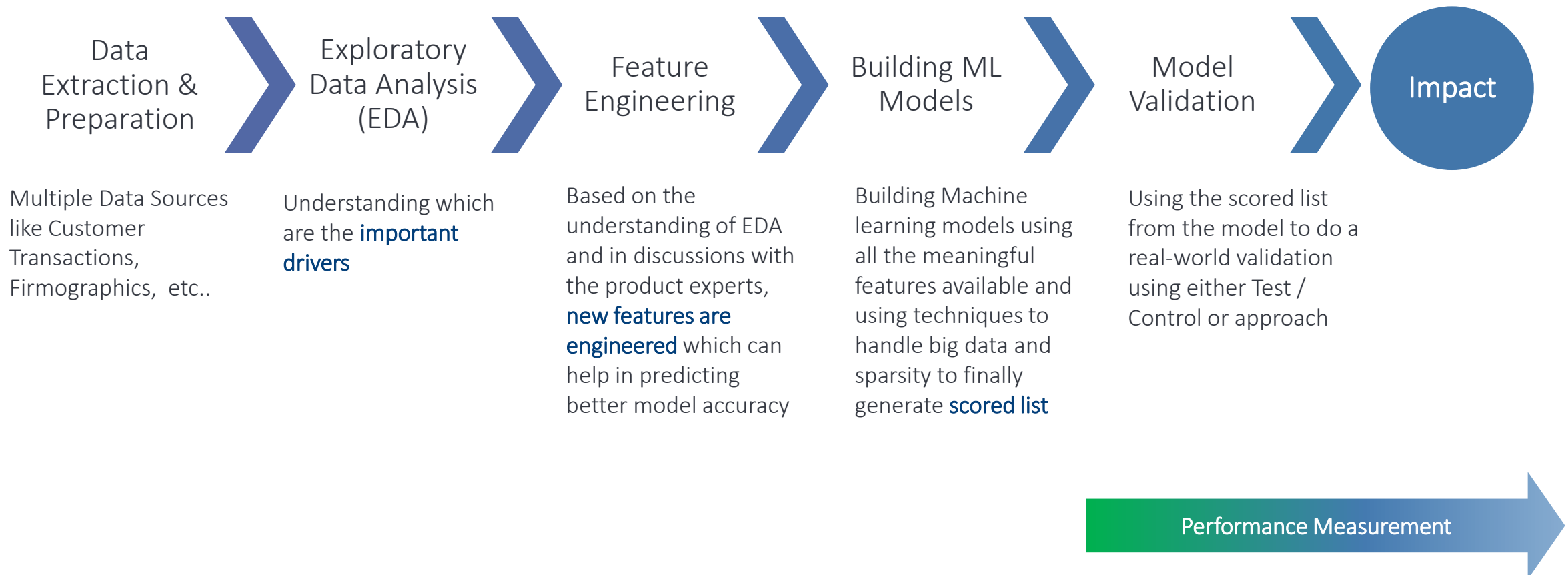


General Steps to Build a Model

A view of data mining process

Bucket	Steps	Industry success	Competition Success
Translate Business problem into Data Mining Problem	Define Business problem	Very high	Not applicable
	Determine entity level to model at	Very high	Not applicable
	Determine the target variable	Very high	Not applicable
Plan for generalization	Determine cross-validation techniques	Critical	Critical
	Find the right validation dataset	Critical	Critical
Feature engineering	Engineer features talking to stakeholders with business knowledge	High	Not applicable
	Engineer features from the dataset	Very high	Critical
Modeling	Hyper-parameter tuning	Medium	Critical
	Ensembles of models	Medium	Critical
	Last mile optimizations	Medium	Critical
Implementation	Use by the business resulting in business impact	Critical	Not applicable

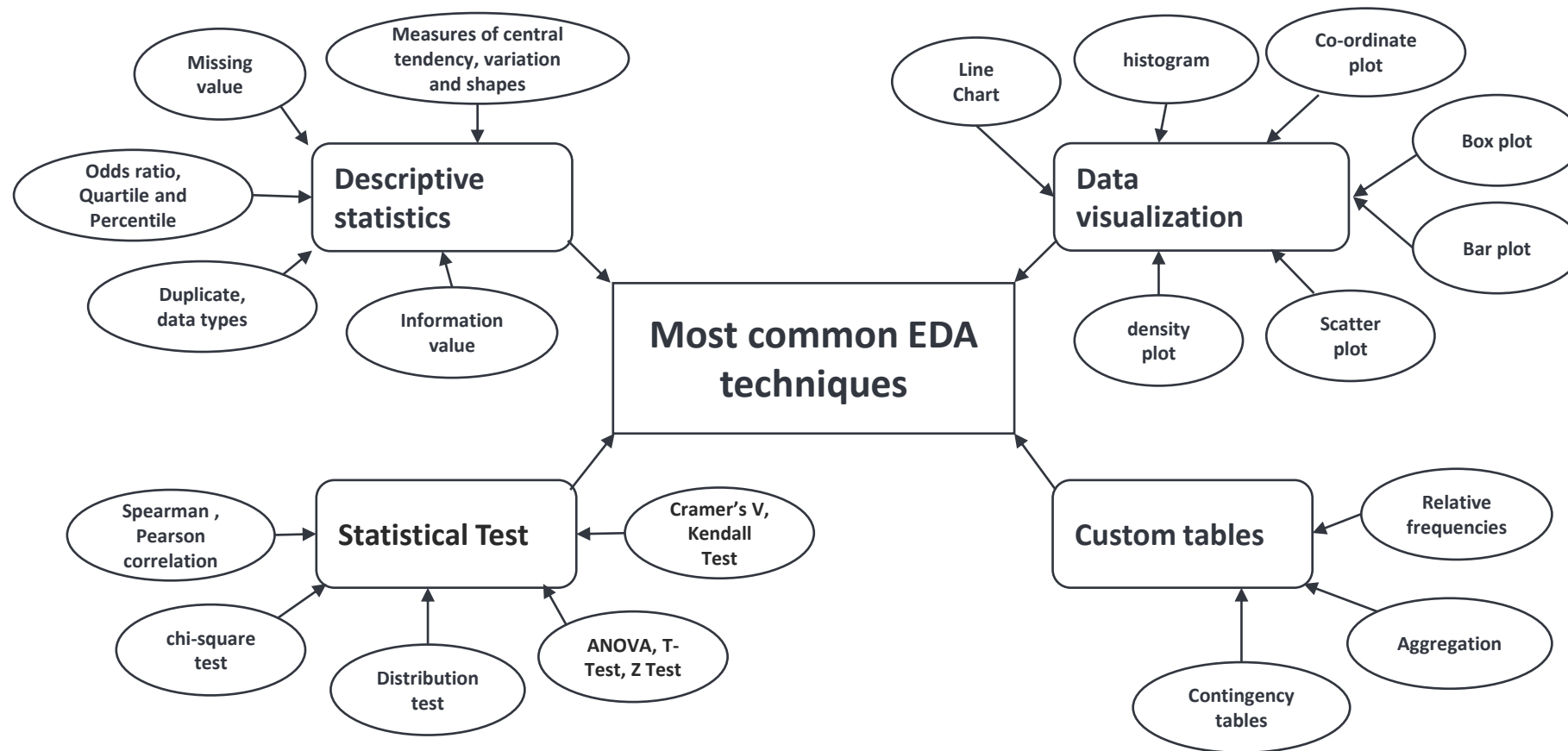
Stages in Building Machine Learning Classification Model





Exploratory Data Analysis (EDA)

Typical EDA used in machine learning



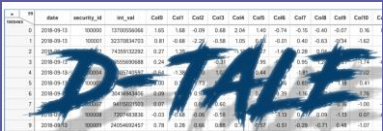
Top Libraries that can perform EDA in one line of python code



Pandas profiling is an open-source python library that automates the EDA process and creates a detailed report. Pandas Profiling can be used easily for large datasets as it is blazingly fast and creates reports in a few seconds



Sweetviz is an open-source python auto-visualization library that generates a report, exploring the data with the help of high-density plots. It not only automates the EDA but is also used for comparing datasets and drawing inferences from it.



D-Tale is an open-source python auto-visualization library. It is one of the best auto data-visualization libraries. D-Tale helps you to get a detailed EDA of the data. It also has a feature of code export, for every plot or analysis in the report.



Autoviz is an open-source python auto visualization library that mainly focuses on visualizing the relationship of the data by generating different types of plot.




Feature Engineering Techniques (FE)

Types of Data Cleaning and Feature Engineering

A

Missing variable imputation

X1	X2	X3
12	A	1
NA	B	0
22	NA	NA
34	A	1




X1	X2	X3
12	A	1
23	B	0
22	A	1
34	A	1

D

one hot encoding for categorical features

X1	X2	X3
12	A	1
23	B	1
22	C	0
34	A	1



X2_A	X2_B	X2_C
1	0	0
0	1	0
0	0	1
1	0	0

F

Bulk interactions for numerical features

$$X1_X3_multiple = X1 * X3$$

X1	X2	X3
68	A	2
55	B	4
1	C	6
62	A	3




X1_X3_multiple
136
220
6
186

B

Removing duplicates

X1	X1	X2
1	1	0
2	2	1
1	1	0
2	2	1




X1	X2
1	0
2	1

E

Outlier flag and imputation

X1	X2	X3
68	A	2
55	B	4
1	C	6
62	A	3



X1_cap	X1_imp
0	68
0	55
1	61.6
0	62

G

Frequent Transformer

X1	X2	X3
68	A	2
55	B	4
1	C	6
62	A	3




X2_Freq	X2_Prop
2	0.50
1	0.25
1	0.25
2	0.50

C

Cleaning column names

X1	X2	X?3
1	A	0
2	B	1
1	A	0
2	A	1



X1	X2	X3
1	A	0
2	B	1
1	A	0
2	A	1

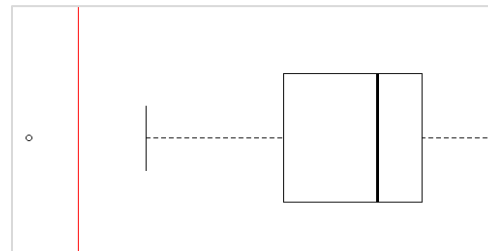
H

Date Variable Transformer

X1	X2	X3
1/11/2019	A	2
12/1/2020	B	4
23/6/2018	C	6
18/9/2017	A	3



X1_month	X1_year	X1_quarter
11	2019	4
1	2020	1
6	2018	2
9	2017	3





Classification Models Use Cases

HR Analytics: Job Change of Data Scientists

Objective

Predict the probability of a candidate looking for a new job

Data Source

Open-source Competition Data set

Evaluation

Area under the curve score

Details

This dataset designed to understand the factors that lead a person will work for the company(leaving current job) ,and the goal of this task is building model(s) that uses the current credentials, demographics, experience to predict the probability of a candidate looking for a new job or will work for the company.

Data Note

- The dataset is imbalanced so it might affect your result if you don't handle it
- Most features are categorical (Nominal, Ordinal, Binary), some with high cardinality so encoding methods and techniques will help to boost models performance
- Missing imputation strategy might affect the results so it can be a part of your pipeline as well.

Practical Demo using Python Packages

EDA and ML Model (Random Forest and XGBoost)





Q&A