Specialization - 02

Week 02 → "Feature Engineering"

- Mapping Raw data into feature
- Mapping numeric value & categorical value
- Empirical Knowledge of data.

→ processing step

Data cleansing / Feature tuning / Representation transform / Feature extraction / Feature construction

→ Feature Engineering techniques

* Numerical Range { · Scaling
· Normalizing (Min-max) $N \sim [0, 1]$
· standardizing (z-score $\Rightarrow x_{std} = \frac{x - M}{\sigma}$)
$x_{std} \sim (0, \sigma)$

* Group { · Bucketizing
· Bag of words

* Dimensionality Reduction { · PCA
· t-SNE
· UMAP

Tools to use
↳ Tensorflow Embedding projector

Ⓐ Apache Beam run on top of other frameworks like (Spark or Flink) on google cloud data Flow.

Tools
↳ Tensorflow Transform
Apache Beam →

Ⓐ Data Affected by → Seasonality, Trend, drift

* Feature selection → [supervised] → { – filter → [-correlation, -univariate feature selection
– wrapper → [- Forward selection, - Backward elimination, - Recursive feature elimination
– Embedded → [-L1 & L2 regularization, -Feature importance

→ Correlation
- Pearson's correlation
- Kendall Tau Rank
- Spearman's Rank
- F-Test
- Chi-squared test

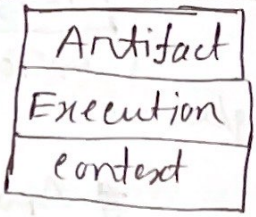
week-03 (Data journey)

→ Data Versioning
- Version control of dataset (Keep track of dataset change over the period)
- Tools = DVC, Git-LFS.

→ Environment version → Docker, Terraform

→ ML Metadata ⇒ Helps to track the data in order to change anything.

ML Metadata is a library for recording and retrieving metadata associated with ML production pipeling among other applications.



Data storage.

Week-04

→ Advance Data Labeling
- Semi-supervised learning
- Active learning
- Weak supervision with Snorkel.