

Core ML Pipeline

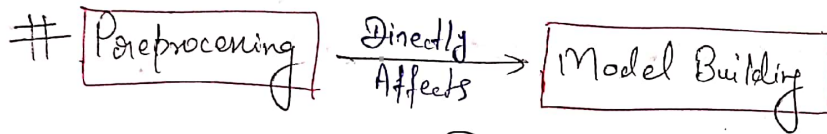
- ① Data Collection
- ② EDA (Analysis)
- ③ Preprocessing or FE
- ④ Model Building
- ⑤ Evaluation Matrix or Validation

EDA

- ① Profile Building
- ② Stats Based Analysis
- ③ Graph Based Analysis (using Python)

Preprocessing

- ① Missing Values
- ② Handle outliers
- ③ Scale
- ④ Transformation
- ⑤ Encoding
- ⑥ Handle Imbalanced data
- ⑦ Feature Selection
- ⑧ Dimension Reduction (PCA, LDA, tSNE)
- ⑨ Duplicate value / Duplicate column
- ⑩ Split / merge / Drop / Add

Different Ways of performing FE

- ① Handle Missing Values : \rightarrow
Different Techniques : \rightarrow

- ① Fill it with any Random Value
- ② Forward Filling / Backward Filling
- ③ Statistical Approach : \rightarrow Can fill with mean, median or mode.
- ④ "End of distribution" : \rightarrow Can fill the data with the help of end of distribution.

- ⑤ Drop that row
- ⑥ KNN - Imputer
- ⑦ Take M.L Algo which support missing values.
- ⑧ Create your own model and can predict missing value.

③ Transformation of Data

- ① Box-Cox transformation
- ② Power transformation
- ③ Log transformation
- ④ Square root transformation
- ⑤ Cube root transformation
- ⑥ Yeo Johnson transformation.

- ② Outlier : \rightarrow

Detection

- ① Z-Score
- ② IQR
- ③ Boxplot
- ④ Scatter Plot
- ⑤ Violine Plot etc

Handling

- ① Drop
- ② Fill with median
- ③ Replace with any value / Trimming that part

\rightarrow Since outliers affects 'mean' hence we can't choose 'mean' to replace the outliers. So we have chosen 'median'.

⑤ Encoding : \rightarrow

- ① One hot encoding
- ② Label encoding
- ③ Binary encoding
- ④ Target guided encoding.
- ⑤ Hash encoding.

④ Scaling

- ① Standardization
- ② Min-Max scale
- ③ Unit scaling

⑥ Imbalanced Data: → Inside the Particular column if ^{class} ratio is mismatching i.e called Imbalanced data. → Dynamic

Different technique: →

- (i) Collect more data
- (ii) Under sampling
- (iii) Over sampling.
- (iv) Cluster based over sampling.

* Dataset

Perform → EDA

Preprocessing

- (i) missing nos. → mean/mode/median
- (ii) outliers → Boop
- (iii) scaling → min-max
- (iv) Encoding → one hot encoding

↓
Model → 75%.

* Suppose after applying some technique we have done preprocessing and we have done model building & we get → 75% accuracy.

if Using some other technique if we do preprocessing and then after model building we get → 80% Accuracy

Hence ~~it~~ and it keeps increasing or decreasing.

* Hence we can say that preprocessing or PE is completely Dynamic.
It is a research area where ^{after} every few days new techniques keeps coming.

* Whether it is structured data or text data or image data we have to perform preprocessing/PE over the data.

Q. Then how can we build the best model?

By doing permutation and combination and it comes with practice.