07/02/22 medium.com.

Data Preprocessing & EDA (Exploratory Data Analysis)
(cleaning)

1) Data Preprocessing also called data cleaning.

Step 1
2) Constant features

→ whole column has only one value.
With one constant value only these are not needed, they are generally dropped.

Quasi-constant feature

→ threshold defined.

→ A column which has 99% or 80 values are same and only one, two outlier.
i.e almost a constant feature. These are also generally dropped.

Step 2
3) Data filtering

i) → Drop columns which have very high rows of null..

df. isna(). mean() % 100 → (gives percentage of null in columns).

ii) Dropping duplicate values.

iii) dropna fn used to drop NON values.
This is used to drop rows.

iv) Dropping columns that have data, but it is not relevant to you.

v) If null values not dropped they are generally replaced by a string (or) number. Bcz null is not understood by data model later. [fillna used].

4) Date Time Conversion

→ HRC mainly handles time-series data So date-time conversion is very imp.

→ format = '%Y-%m-%dT%H:%M:%SZ' format

Syntax → df 1["ColumnName"].dt.year
└→ has data. └→ date └→ week, day, dayofweek etc also available.

5) Data Splitting → Missed bcz of some work. but basic is, → Train set data ← Data Set → Test set data (80:20 or 70:30).

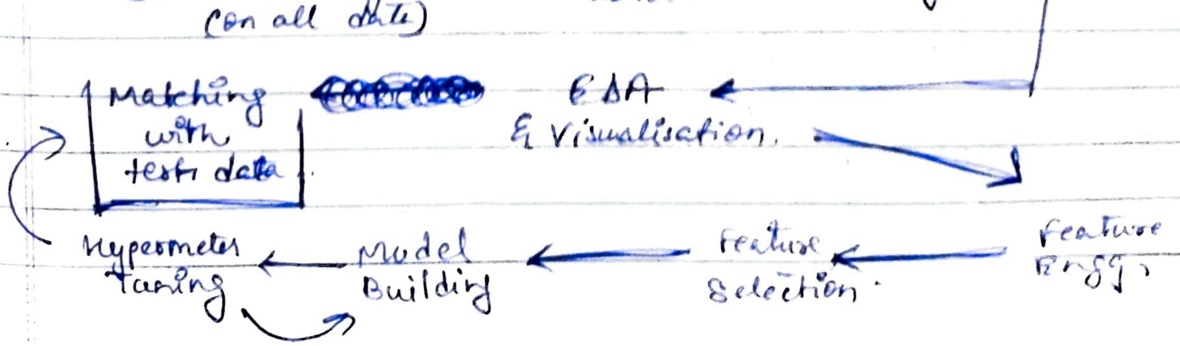
Statistics

6) → Continuous Variables → A seq of possible values which are infinite and uncountable

Categorical Variables →

7) → Distribution → Uniform, exponential, normal} →

Binomial → based on a histogram
↳ for a categorial variable



Page No.
Date

Data Analysis Maths Overview

EDA [We will only cover numerical analysis].

→

1.) Univariant & Multivariant Analysis.

→ Mean, Median or mode.
↳ not generally used.

→ Variance, Standard deviation.

→ Skewness and Kurtosis.

} Univariant. (With only one columns).

Studying multiple variables at once → multivariant.

→ Correlation.

→ Covariance.

→ Principal Component Analysis (PCA).
↳ Not used in Fintech.

★ In fintech generally data is not modified. Just analysis is done.

2.) Distribution & IQR (Inter Quartile Range).

→ Population, Sample, Distributions —

→ Probability · concepts discussed here.

→ Skewness of distribution → Left, Right discussed.
↳ Measured by mean, median & mode.
(Measures of central tendency)

Median → Positional Avg

Mean → Mathematical Avg
(AM, GM, HM)

3) Outlier Detection → Disrupt data & conclusions, generally taken out from data.

Detection generally done by boxplot.

IQR → Inter Quartile Range. (Outside the IQR → Outlier).

4. EDA (Expolatory Data Analysis).

① ⇒ Converting numbers to visual [Matplotlib → Private, Seeborn → Client & Plotly ↳ very high level.]

② ⇒ Barplot, Scatterplot, Boxplot } diff kinds of plots

③ ⇒ import matplotlib as plt.
↳ Like pandas is pd & numpy is np.

④ ⇒ Other plot → histogram, piechart, etc.
Distplot, Violin plot, Colon density plot.

Final Project

Clear date column in dataset. → when bill will be cleared. of each of 50,000 invoices.

We will predict clearing date.

PDP.