Feature Engineering and Selection

1. After EDA (Analysis & Visualization) of data we move to next step feature engineering.

If EDA is not done well (i.e if you didn't do your understanding of data well) you can't select it's good features to make good AI/ML models.

Eg:- Car buying
then features → min & max speed.

∴ In ML features are things that can tip your decisions and that of your model.

2. Model → An algo trained to take in input and give out prediction output.

features (of car) → Model → Prediction. (Buying car yes (or) no).

3. Features engineering → Act of converting raw observations into derived features.

Features
- → Raw → Prediction directly on (one variable) (buying car only depends on price) only.
- → Derived.
  - → feature from two (or) more variables.

different kind of feature eng is applied for diff datatypes (numerical, date time, categorical).

(4) Variables
- → Numeric
  - → Continuous
  - → Discrete.
- → Categorical
  - → Nominal
  - → Ordinal (order).
  - → Date/time.

(5) Creating features from numerical columns.

a) Using raw features as is.

b) binning

c) Binarization (creating a feature for 0 & 1)

d) Log transformation.

(6) For categorical variables. [cannot be directly used, ∴ raw features as is can't be used].

a) label Encoding

b) One hot encoding

c) Target Encoding

→ Giving numerial value/ labeling categorical data [Eg:- Male → 1, Female → 0.]

But this might lead to bias, thats when one hot encoding is used.

(7). Mixed variable (Num & Cate) dataset can also be present.
mix

(8) Handling outliers.

Outlier detection with standard deviation.

Either drop outlier (or) cap outliers.

Page No.
Date

9) Normalization

a) → Min max scalar. } Scaling of data in distance based algo.
b) → Standard scalar. }

c) Feature Scaling.

X_train Y_train
X_test Y_test.

Standardization in Normalisation

Feature Selection.

Feature Selection.

- Filter method [Similar to data processing]
- Wrapper method (takes long time) exhaustive.
- Embedded method. (L1, L2, tree)

If corelation very high i.e columns similar then drop.

Constant
Quasi constant
& Duplicate colums } to be dropped

---

delay column create → clear date - when project bill. delivered.

where clear bill = null → test set.

linear regression
decision tree. | Hitmap & filter } for Leggings.

Preparation for Quiz (Preprocessing, EDA and Feature Engineering)

(1) ~~Preprocess of data~~

Constant Column/features removal → nunique().
Quasi constant removal → Variance threshold

1.2 Target Variable → Variable for which you want to get a deeper understanding of.
→ ~~Depends on your business & goal.~~
→ ~~Very imp in case of supervised learning.~~

1.3 → ~~Removing duplicate columns~~ → ~~If same kind of feature again & again, the model might get biased.~~

For small dataset. → drop_duplicates in pandas after Transpose.

→ But if rows > 50,000 & columns > 100 then error will be thrown bcz of recursion stack will be full. Seperate func to be written.

1.4 → Date / Time in dataset

→ Special type of categorical variable. Gives lot of data.

Can give info in form of → Year, month, day, Quarter, Semester, week(or) weekend, day of the week.

```
data['Date of Birth'] = pd.to_datetime(data['Date of Birth'])
```

then derive conclusions.

```
data['month'] = data['Date of Birth'].dt.month
```

table name

etc.

* isin for sat, sun.

difference → (datetime.datetime.today() - data['DOB']).head().

Page No.
Date

1.5 Missing Values → 3 data mechanism for missing

- completely random. (MCAR)
- almost ~~random~~. (MAR)
- systematic loss of data. (MNAR)

for identification

→ Including completely random data doesn't bias the model.

→ We can proceed with ~~almost~~ random also if we can manage it.

→ Not at random are containing relevant info, will create impact.

MNAR

★ → data.isnull().mean() → for finding null percentage.
data.isnull().sum() → quantifies num of missing values.

★ → Generally groupby & np.where() is used in this process.

MCAR → Ignore, won't make a diff. → data[data.columnname.isnull()]

MAR → data.emp_title.unique()[0:20]

1.6 Mean & Median Imputation.

→ means replacing the data which is missing