

Feature Engineering and Selection

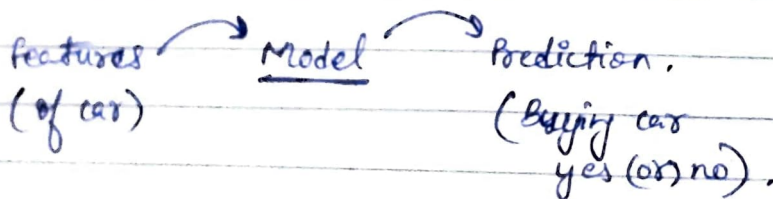
- ① After EDA (Analysis & Visualization) of data we move to next step feature engineering.

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.

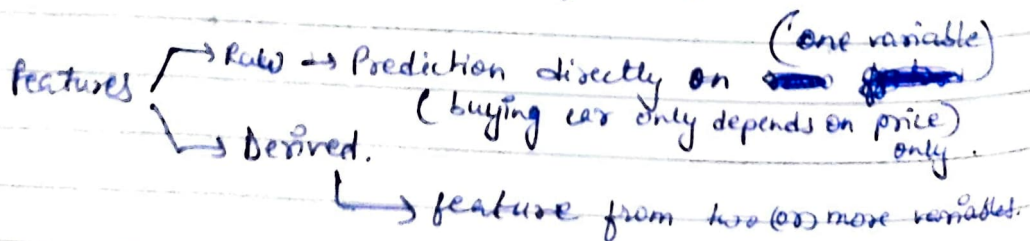
Eg:- Car buying
then features \rightarrow min & max speed.

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

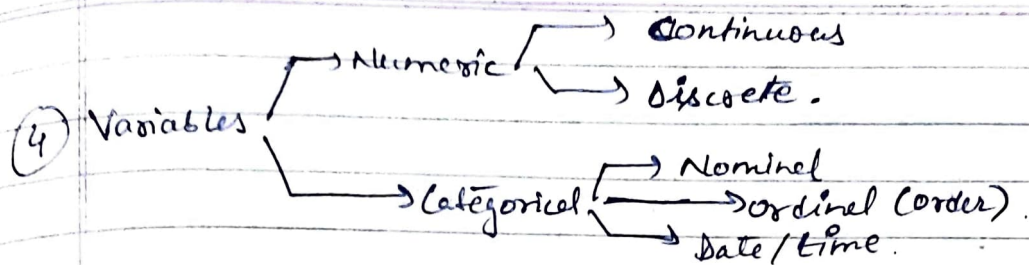
- ② Model \rightarrow An algo trained to take in input and give out prediction output.



- ③ Features engineering \rightarrow Act of converting raw observations into derived features.



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



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

But this might lead to bias, that's when one hot encoding is used.

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

⑧ Handling outliers.

Outliers detection with standard deviation.

Either drop outlier OR cap outliers.

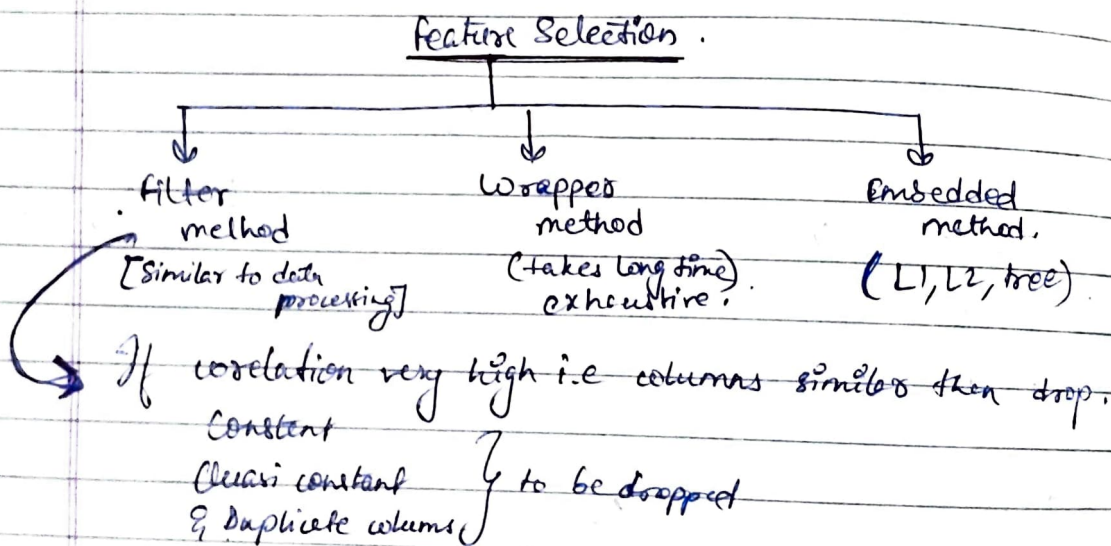
9) Normalization

a) \rightarrow min max scalar. } Scaling of data in distance based algo.
 b) \rightarrow Standard scalar.

c) Feature Scaling. X_train Y_train
X_test Y_test

10) Standardization in Normalization

Feature Selection



delay column create \rightarrow clear date - when project delivered.

where clear bill = null \rightarrow test set.

Linear regression
decision tree.

Hitmap for loggins
& filter

Preparation for Quiz (Preprocessing, EDA and Feature Engineering)

(i) Preprocess of data

Constant column/features removal \rightarrow unique().

Occur constant removal \rightarrow Variance threshold

1.2 Target Variable \rightarrow Variable for which you want to get a deeper understanding of.

\rightarrow Depends on your business & goal.

\rightarrow Very imp in case of supervised learning.

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

For small dataset \rightarrow drop_duplicates in pandas after Transpose.

But if rows $> 50,000$ & columns > 100 then error will be thrown by recursion stack will be full. Separate func to be written.

1.4 Date/Time in dataset

\rightarrow Special type of categorical variable. Gives lot of data.

Can give info in form of \rightarrow 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 \rightarrow

\times isin for sat, sun.

etc.

difference \rightarrow `(datetime.datetime.today() - data['DOB']).head()`

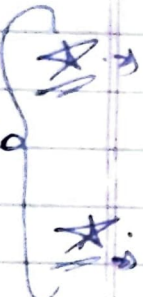
①.5 Missing Values

- 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 all random are containing relevant info, will create impact.

MNAR



`data.isnull().mean()`
`data.isnull().sum()`

- ↳ for finding null percentage.
- ↳ quantifies num of missing values.

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

MCAR

→ Ignore, won't make a diff. → `data[data.columns[0].isnull()]`

MAR

`data.emp_title.unique()[0:20]`

①.6 Mean & Median Imputation

→ means replacing the data ~~with~~ which is missing