Feature Engineering

↳ Is the process of using domain knowledge to extract features from raw data.

Then these features can be utilized to improve performance of ML Algorithms.

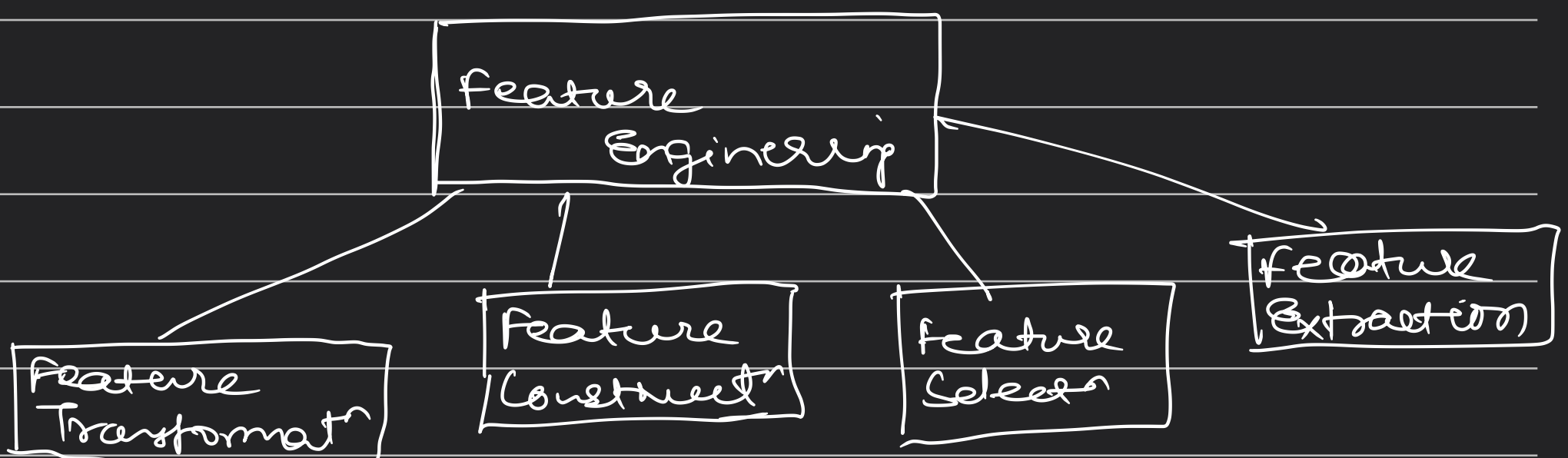
Its more of an ART ⇒ Feature Engineering

"Even a bad algorithm will perform extremly well if its features are fine grained picket"

"On the other hand a good algo will perform bad if it doesn't get good features"

ML Flow for Any Problem.

Flow Chart of feature Engineering



Feature Transformation

⇒ Handling Missing Values → This first step in feature Engineering

- Handling by adding values
- Handling by removing values.

⇒ Handling Categorical Value

- Your libraries only understand numerical data. So your categories must get converted into some numerical value.

→ Ex - One Hot Encoding.

⇒ Outlier Detection

Outliers are noisy data, the cause the

Algo to behave abnormaly when trained & give drastic result.

⇒ Feature Scaling

Few features will have dominance in value over others, this can impact our algo. For this we Scale features.

Ex: Min Max Scaling, Normalization etc.

Ex Age & Salary = 32, 100000 $

This difference in ⇐ Big Difference scaling of features effect the Algo.

Feature Construction

we create a complete new column or variable or feature based upon our intitution we got from domain knowledge of data.

⇒ Family Member column can be created using Sibling & ParArch column in Titanic Database.

↳ we can also create categorical data called Alone with 0 family Members & family Mar etc.

feature Selection

we have lot of input columns & we select based upon our domain knowledge

which columns / variables or features we are going to use in our Machine learning algorithm.

Now we start to operate on less features because we eliminated the unnecessary ones.

Feature Extraction

Programatically we create new features based upon the features provided.

Its bit different from feature construction

For Ex ⇒ Room, WashRoom, Price These are necessary data for Real State Predict but we can use Room + WashRoom = Carpet Area

This way new feature got created & we got rid of the other two.

In a way we reduce features & use the new created ones.

Used specially in High Dimensional Data.