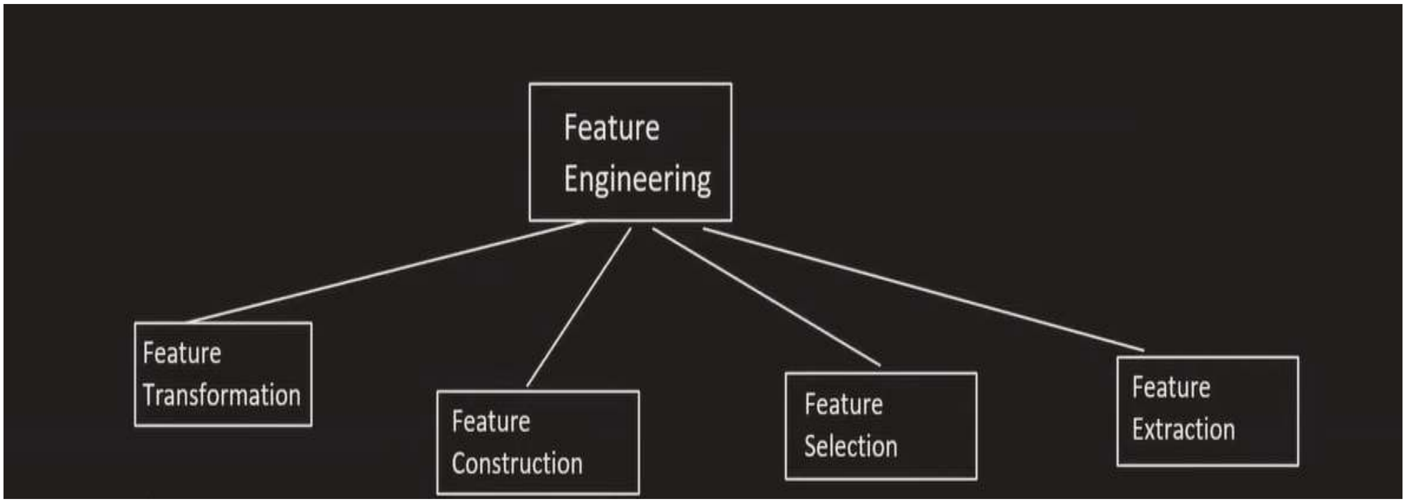
Feature Engineering

06 July 2023 19:50

Feature Engineering is the process of converting the raw data into the data which is comparatively more clean, useful and informative in terms of feeding to Our Machine Learning Models.

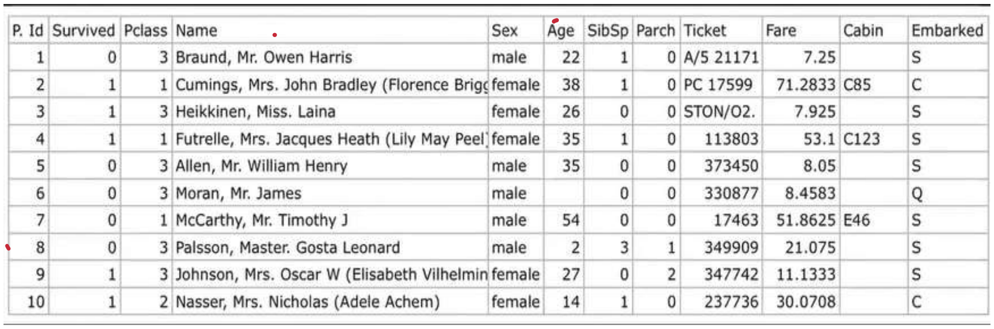
This generally involves following steps:



1. **Feature Transformation** refers to the process of converting and modifying the existing features to enhance their usefulness.
   a. Missing Value Removal/Imputation
   b. Handling Categorical Features like One-Hot Encoding
   c. Outlier Detection
   d. Feature Scaling

2. **Feature Construction** refers to creating new features from existing ones to capture additional information for better performance of the model.

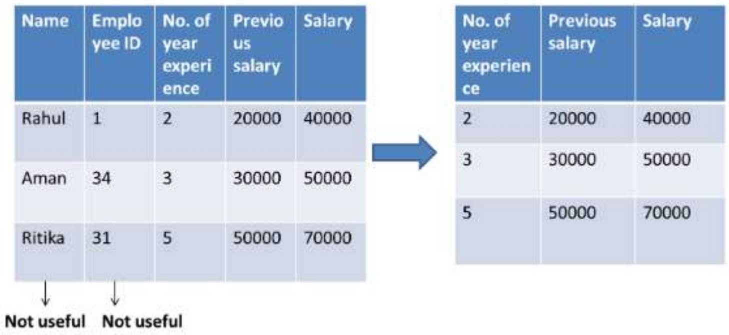
   For Example creating one additional feature "num_family_members" by combining SiblingSpouse and ParentsChildren features.



P. Id	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Brigg	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmin	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C

SibSp	**ParCh**	**num_family_members**	**Family_size**
0	0	0	**Alone** - 0
3	1	4	**Medium** - [3-4]
0	2	2	**Small** [1-2]
1	0	1	**Small**
1	4	5	**Large** [>4]

3. **Feature Selection** refers to the process of identifying and selecting the most relevant and informative features only from the available set of features and not using the remaining features for training the model. This improves the overall performance of the model.

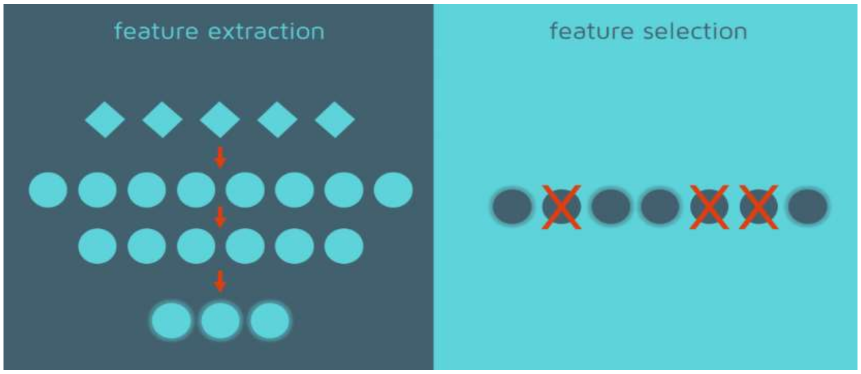


Name	Employee ID	No. of year experience	Previous salary	Salary
Rahul	1	2	20000	40000
Aman	34	3	30000	50000
Ritika	31	5	50000	70000

Not useful Not useful

No. of year experience	Previous salary	Salary
2	20000	40000
3	30000	50000
5	50000	70000

4. **Feature Extraction** refers to **programmatically extracting completely new features** from a given set of features **OR** finding a smaller set of **new features** by combining the existing features, containing basically the same or additional information. - This is completely different from Feature Construction.



Id	AvgBfast_Cal	AvgLunCal	AvgDin_Cal	AvgExer_Time	Wt.loss/week?

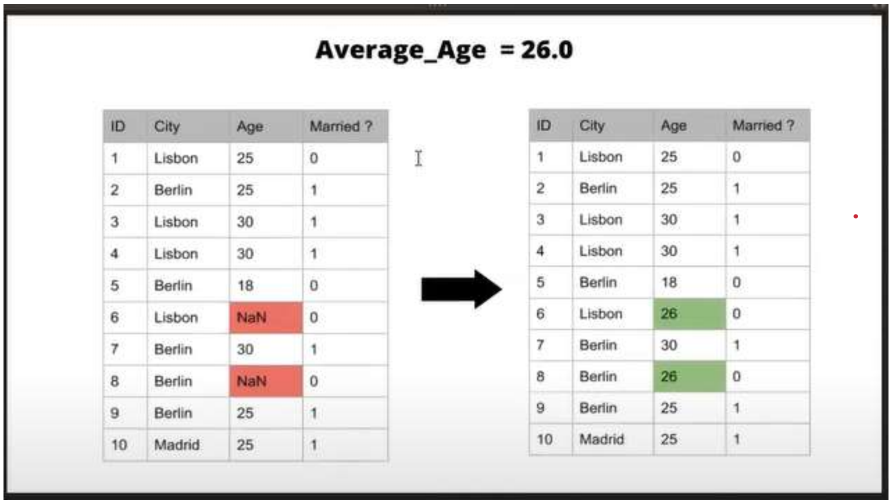
AvgCal_intake	AvgCal_burnt	Wt.loss/week?

----

1. **Feature Transformation** refers to the process of converting and modifying the existing features to enhance their usefulness.

a. Missing Value Removal/Imputation: - Today's Agenda

As **Sci-kit Learn** needs **data with no null values** to train the model, So we would either **Remove** the missing values(**if <5%**) or **Replace** the missing/null values with the **Mean, Median, Mode (Most Frequent Category)** etc.



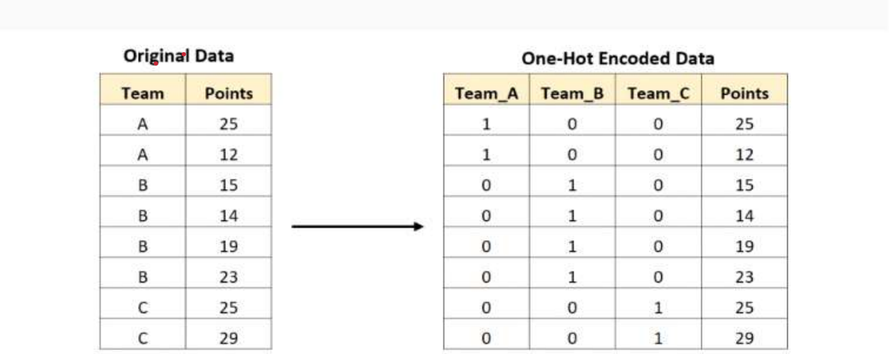
Average_Age = 26.0

ID	City	Age	Married ?
1	Lisbon	25	0
2	Berlin	25	1
3	Lisbon	30	1
4	Lisbon	30	1
5	Berlin	18	0
6	Lisbon	NaN	0
7	Berlin	30	1
8	Berlin	NaN	0
9	Berlin	25	1
10	Madrid	25	1

ID	City	Age	Married ?
1	Lisbon	25	0
2	Berlin	25	1
3	Lisbon	30	1
4	Lisbon	30	1
5	Berlin	18	0
6	Lisbon	26	0
7	Berlin	30	1
8	Berlin	26	0
9	Berlin	25	1
10	Madrid	25	1

b. Handling Categorical Features like One-Hot Encoding:

As Sci-kit Learn works with numerical data only, that is why we have to transform/convert the Categorical Features in Numerical Features.



Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

One-Hot Encoded Data

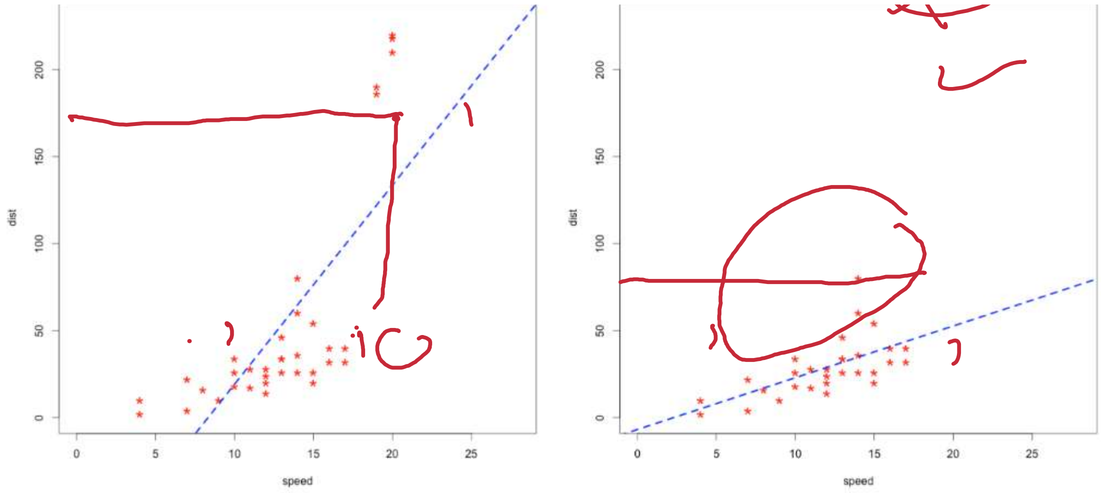
Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

c. Outlier Detection:

Outliers refer to the data points that exist outside the expected range. Or we can say those data points which lie at an abnormal distance from all other values.

Removing outliers is must to ensure better accuracy and performance of the model.

We'll cover the Outlier Detection Techniques later.

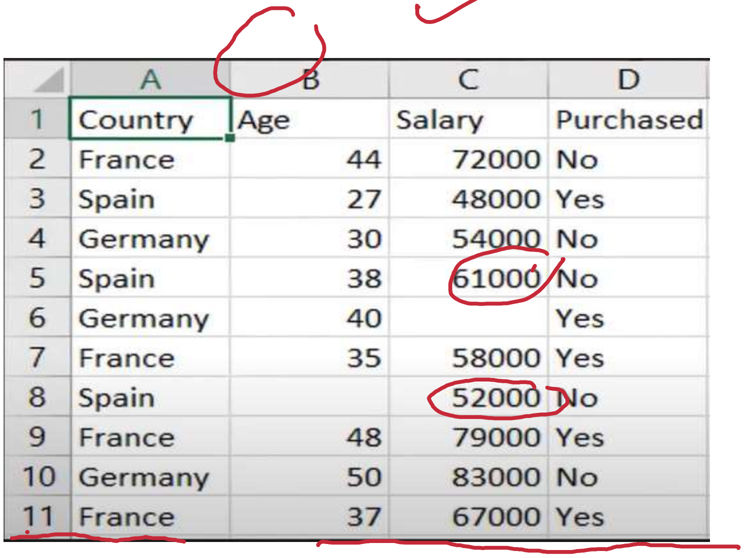


d. **Feature Scaling:** like Normalization and Standardization

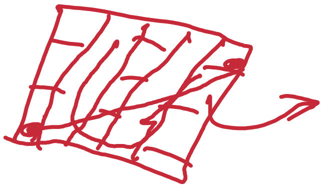
Feature scaling is a data preprocessing technique used to transform the values of features or variables in a dataset to a similar scale.

The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.

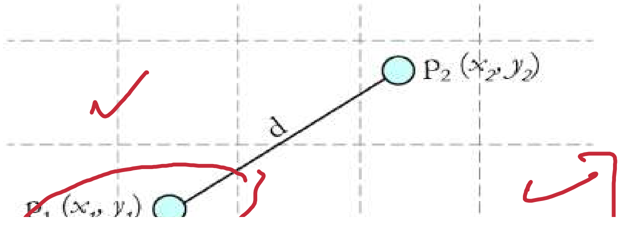
Without scaling features, the algorithm may be biased towards the feature with values higher in magnitude.



	A	B	C	D
1	Country	Age	Salary	Purchased
2	France	44	72000	No
3	Spain	27	48000	Yes
4	Germany	30	54000	No
5	Spain	38	61000	No
6	Germany	40		Yes
7	France	35	58000	Yes
8	Spain		52000	No
9	France	48	79000	Yes
10	Germany	50	83000	No
11	France	37	67000	Yes



Euclidean Distance formula i.e. the shortest distance between the 2 points which is used in kNN Algorithm.



Euclidean distance (d) = $\sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$