ML-8

Feature Engineering

Pandas profiling

Firstly, to do panda profiling we need to install the library of it.

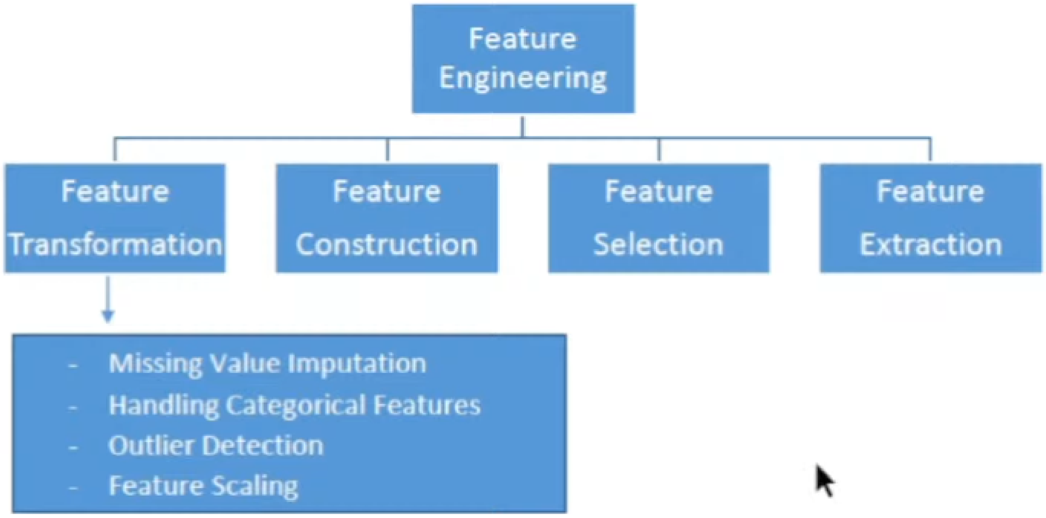
```
pip install pandas-profiling
```

after this , you have to import the profile report from this library like this

```
from pandas_profiling import ProfileReport
prof=ProfileReport(df)
prof.to_file(output_file="titanic.html")
```

Q) What is feature engineering?

Ans. Feature Engineering is the process that takes raw data and transforms it into features that can be used to create a predictive model using machine learning



Missing Value Imputation

Q) what are imputed values?

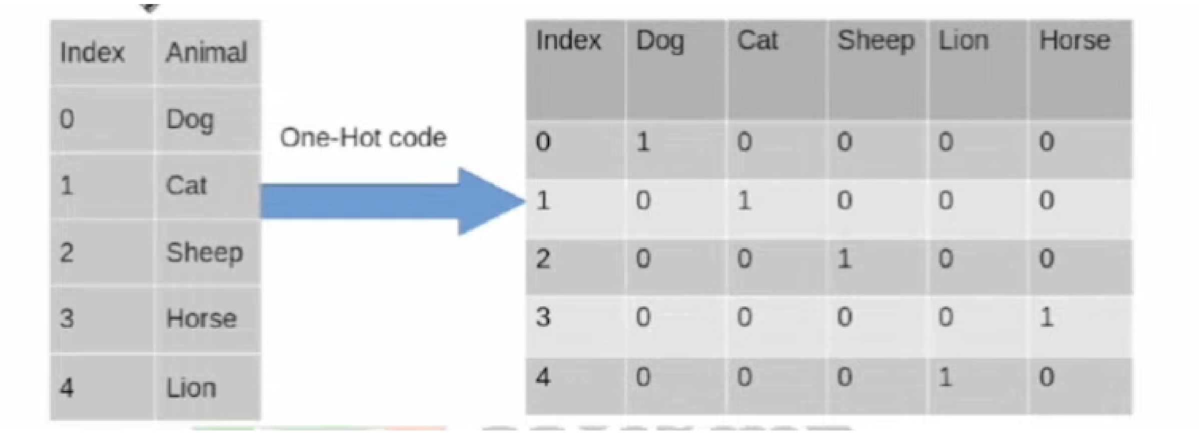
→ these values are also known as estimated imputation , is an assumed values which is given to an item when the actual value is not known . these are logical or implicit values for an item. he can either fill the values by doing some mathematical calculations eg. mean of the data or median.

Handling Categorical Features

Q) How do you handle categorical feature?

→ one of the most common ways to deal with the categorical data in ml is `one-hot encoding`.

in this , we convert the whole categorical data into numerical data by making a new binary feature for each category.



Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

One-Hot code

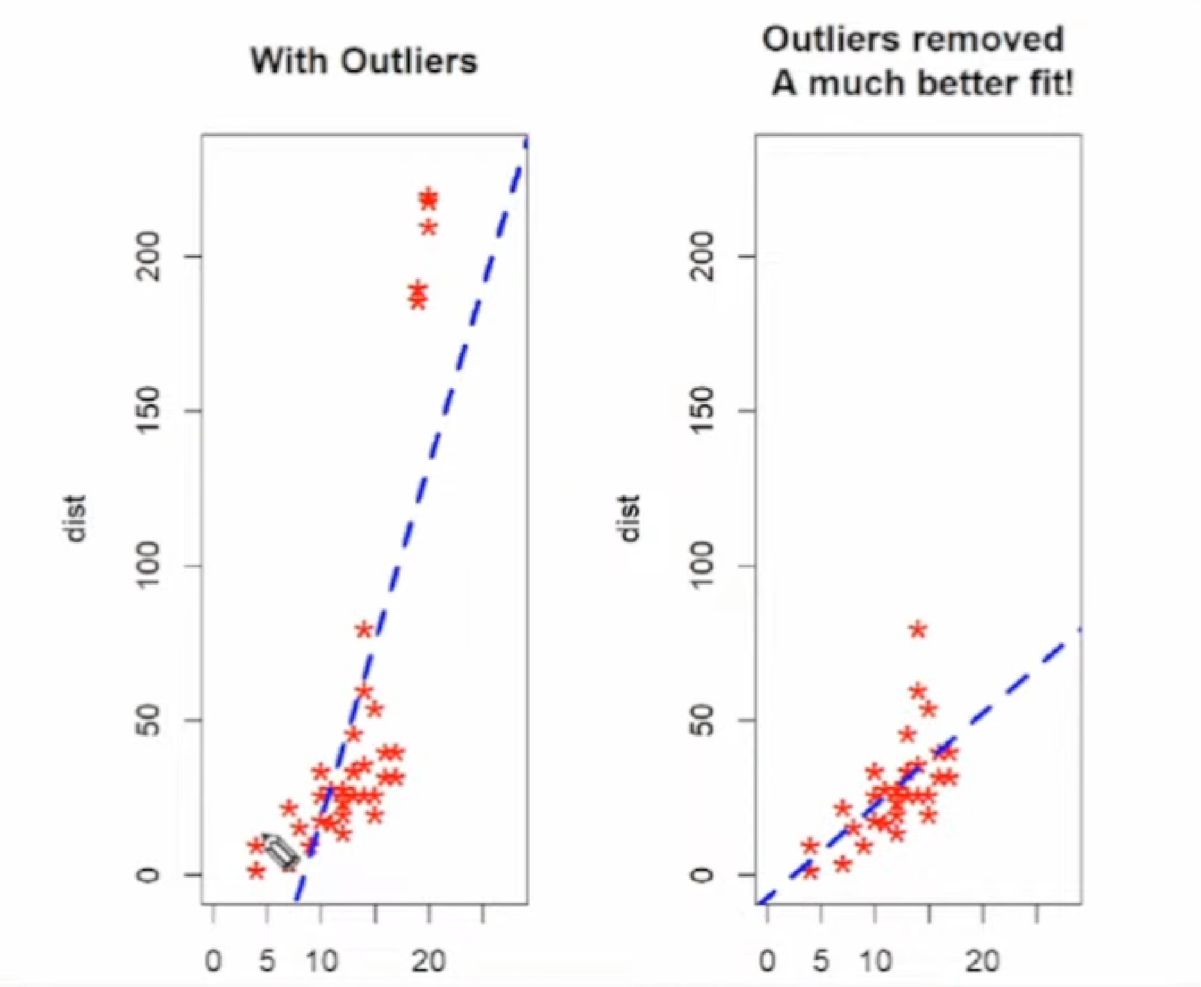
Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

Outlier Detection

process of detecting outliers , or the data point that’s far away from the avg. and depending upon what you are trying to accomplish ,potentially removing or resolving them from the analysis to prevent any potential skewing.

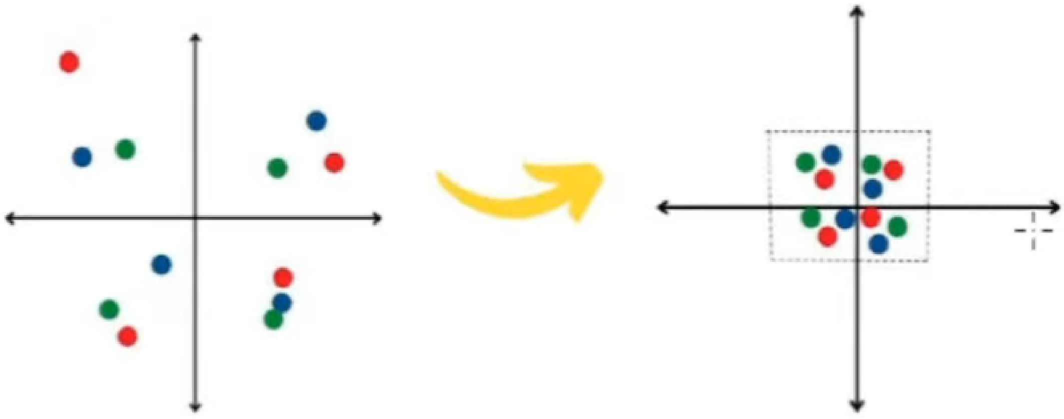
How do you find outliers in linear regression?

→

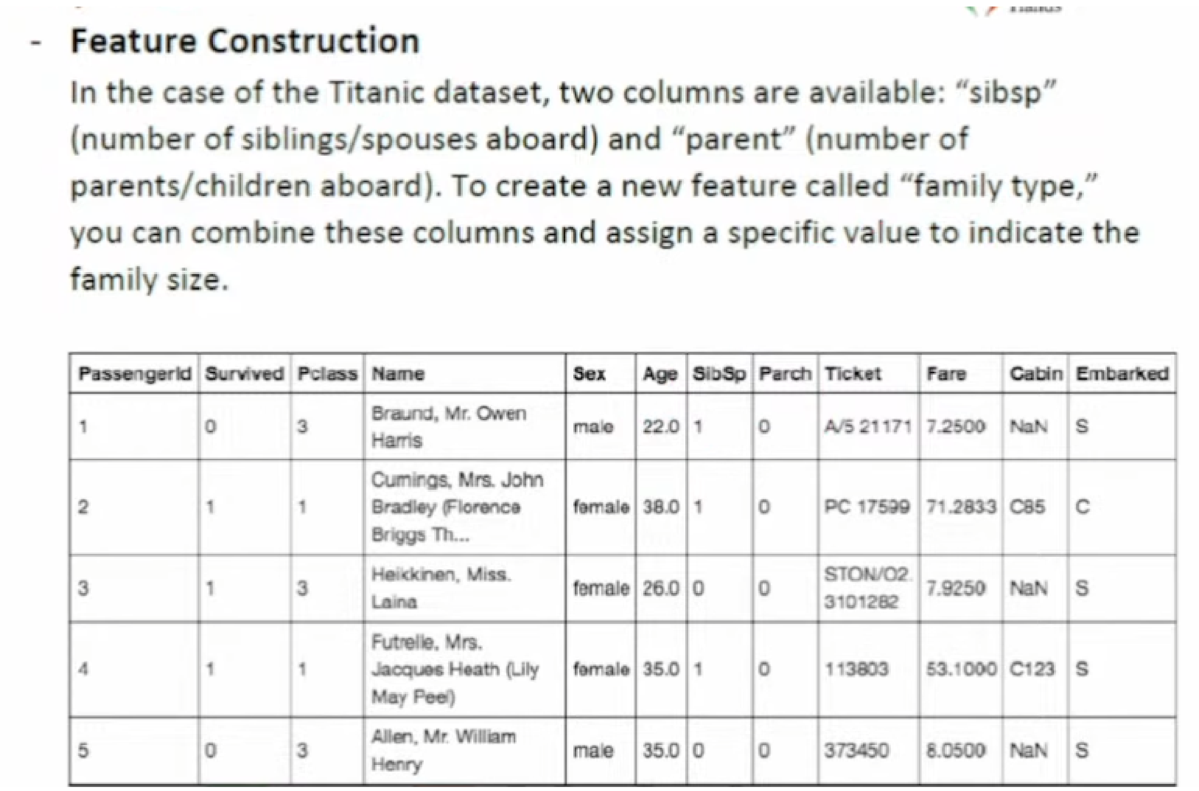


▼ Feature Scaling

it is the method used to normalize the range of the independent variables or features of data . In data processing, it is also known as `data normalization` and is generally done in the data pre-processing step.



▼ Feature Construction

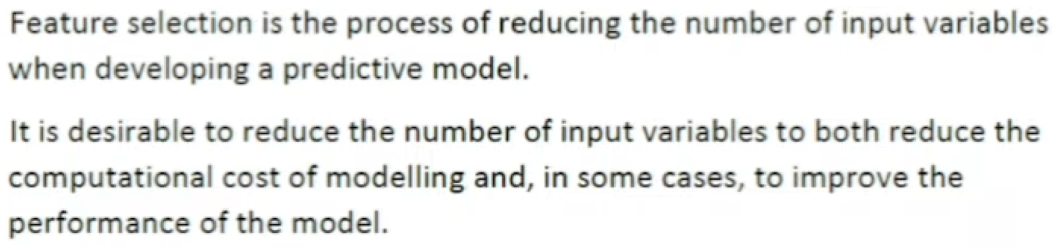


- **Feature Construction**

  In the case of the Titanic dataset, two columns are available: "sibsp" (number of siblings/spouses aboard) and "parent" (number of parents/children aboard). To create a new feature called "family type," you can combine these columns and assign a specific value to indicate the family size.

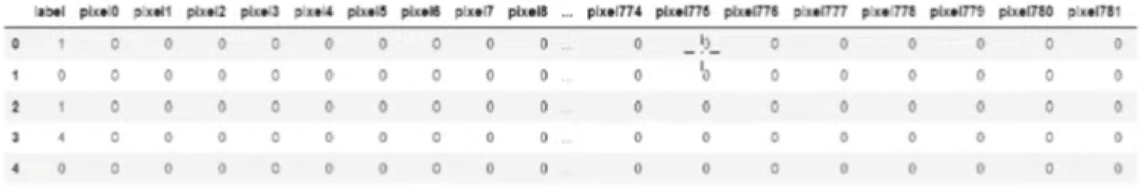
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

▼ Feature Selection

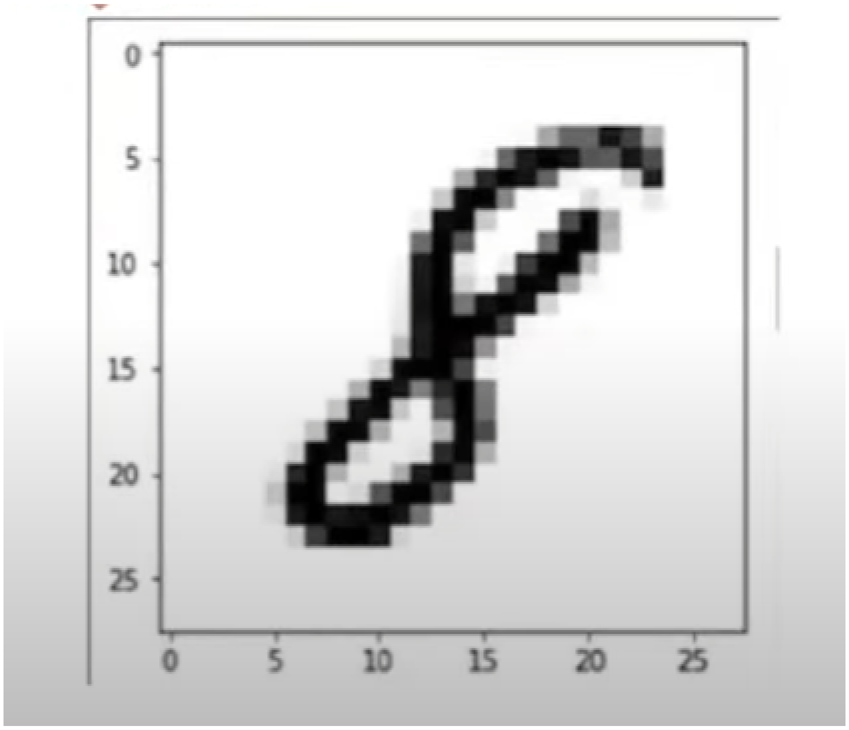


Feature selection is the process of reducing the number of input variables when developing a predictive model.

It is desirable to reduce the number of input variables to both reduce the computational cost of modelling and, in some cases, to improve the performance of the model.

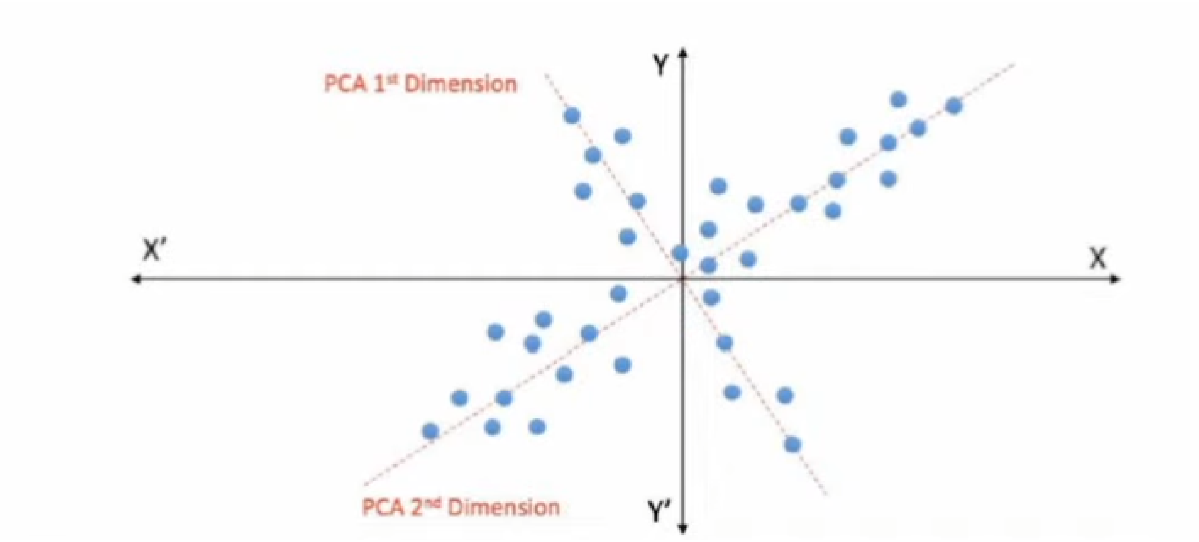


	label	pixel0	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	...	pixel774	pixel775	pixel776	pixel777	pixel778	pixel779	pixel780	pixel781
0	1	0	0	0	0	0	0	0	0	0	...	0	[illegible]	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
3	4	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0



▼ Feature Extraction

it is the process of transforming raw data into numerical features that can be processed while preserving the information in the original dataset . It yields better results than applying machine learning directly to the raw data.


PCA 1st Dimension
Y
X'
X
PCA 2nd Dimension
Y'