Basic Terminologies

By - Aayush Shandilya

Sampling

A sample is an unbiased number of observations taken from a population.

Types of Sampling

1. Simple Random Sampling : Simple random sampling is ideal if every entity in the population is identical. Each member have equal opportunity.
2. Stratified Random Sampling : This type of sampling, also referred to as proportional random sampling or quota random sampling, divides the overall population into smaller groups. These are known as strata. People within the strata share similar characteristics

Feature Scaling

Normalization : Scales features from 0 to 1.Normalization is useful when there are no outliers as it cannot cope up with them. Eg: Image Pixels in CNN (Sklearn - MinMaxScaler)

Standardization : Transform feature s.t. Mean = 0 and Std. Deviation=1.Standardization can be helpful in cases where the data follows a Gaussian distribution. (Sklearn - StandardScaler)

Standardisation (Z-score Normalization)	Max-Min Normalization
$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$	$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$

When to Scale ?

1. KNN / K-Means - what if scale is not right ?
2. Linear Regression - convergence more quickly

When not to Scale ?

1. Decision Trees
2. Random Forest

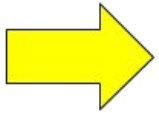
Encoding

Encoding used on categorical attributes/features.

A. Nominal Encoding : are variables that have no inherent order. They are simply categories that can be distinguished from each other.Eg : Gender - Male or Female
    a. One Hot Encoding
B. Ordinal Encoding :have an inherent order. They can be ranked from highest to lowest or vice versa. Eg: Qualification-10th,12th,BE,ME,PhD
    a. Label Encoding

Examples :

Color
Red
Red
Yellow
Green
Yellow



Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

One Hot Encoding

SAFETY-LEVEL (TEXT)	SAFETY-LEVEL (NUMERICAL)
None	0
Low	1
Medium	2
High	3
Very-High	4

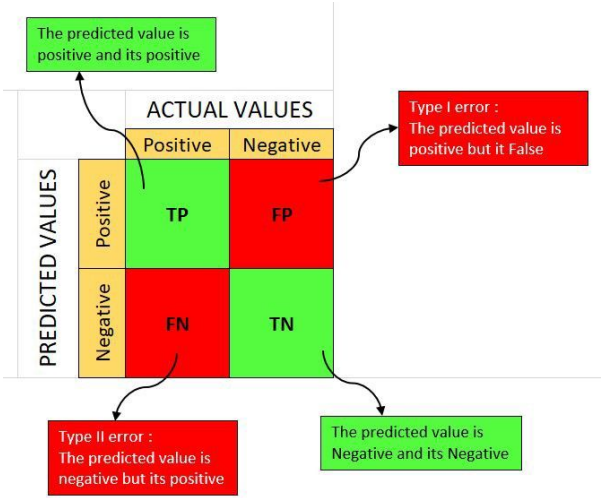
Label Encoding

Problem with one Hot Encoding:

1. Dummy Variable Trap
2. Curse of Dimensionality Eg: Pin Code

Performance Metrics

A. Confusion Matrix



B. Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

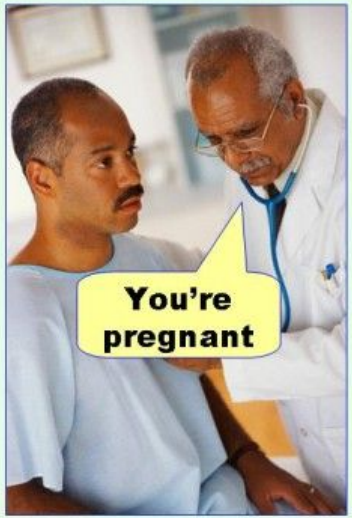
Problem :

T-900

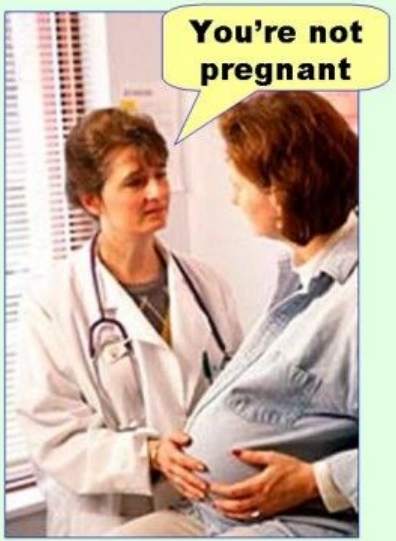
F-100 => Accuracy = 90%

False Positive and False Negative

Type I error
(false positive)



Type II error
(false negative)



Precision and Recall

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

TP = True positive

TN = True negative

FP = False positive

FN = False negative

When to use Precision ? -> What proportion of positive identifications was actually correct? FP Imp

- Eg: Spams [Domain Knowledge required]

When to use Recall ? -> What proportion of actual positives was identified correctly? FN Imp

- Eg: Cancer

What if we want both ? Eg: Stock crash [Company and Stockholders]

F-Beta

$$F_{beta} = \frac{(1+\beta^2)precision * recall}{\beta^2 * precision + recall}$$

Weighted harmonic mean of precision and recall

- If FP and FN both are Imp Beta=1 like stock crash [F-1 Score]
- If FP is more Imp than FN Beta=0.5 like Spam Mail
- If FP is less Imp than FN Beta=2 like Cancer

Find Beta for taking an umbrella for rain ?

Correlation Coefficients

Correlation deals with association between 2 or more variable measured by **correlation coefficient** summarizing direction and degree of correlation.

$$\textbf{correlation coefficient} \in [-1,1]$$

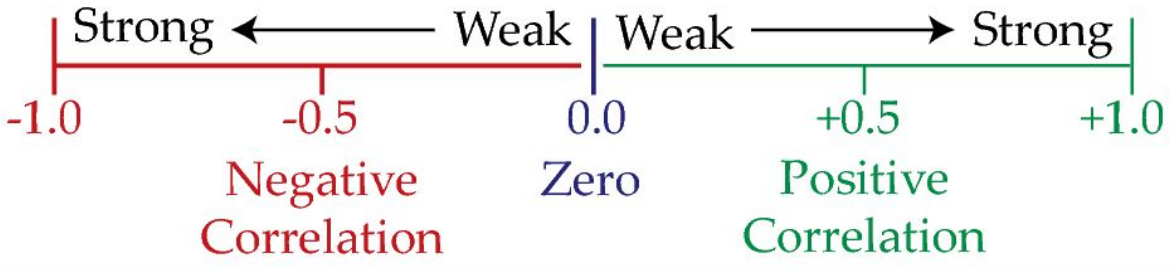
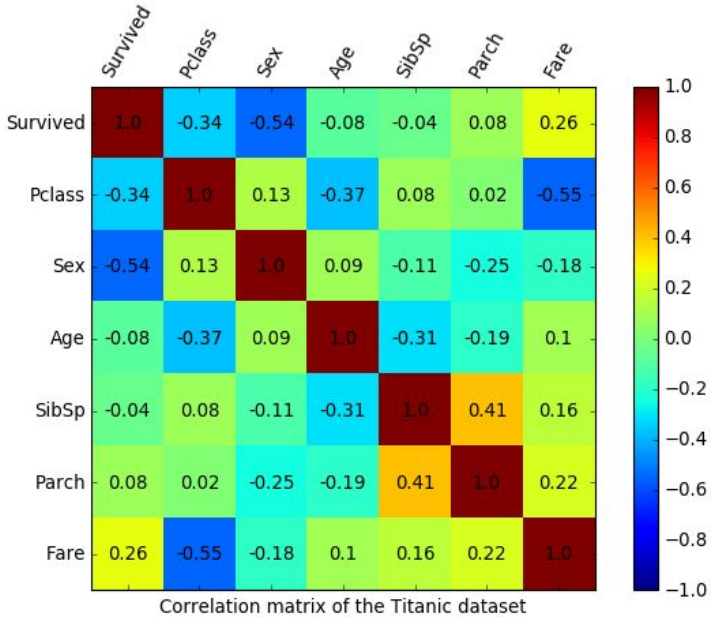


Fig. 2: The spectrum of the correlation coefficient (-1 to +1)

Correlation Matrices

Table which displays the correlation coefficients for different variables



Pclass: Passenger Class
Survival: (Yes/No)
Sex: Sex
age: Age
sibsp: Number of Siblings/Spouses Aboard
parch: Number of Parents/Children Aboard
fare: Passenger Fare

Questions … ?