Feature Engineering

Linear Relationship among numerical variables/features

Day	ABC Returns (%)	XYZ Returns (%)
1	1.1	3
2	1.7	4.2
3	2.1	4.9
4	1.4	4.1
5	0.2	2.5

Daily returns for two stocks using the closing prices

Does Stocks ABC and XYZ are related or not?

How much strong these two stocks are related?

Covariance

Covariance measures how two variables move together. It measures whether the two move in the same direction (a positive covariance) or in opposite directions (a negative covariance).

Covariance

$$\text{Covariance} = \frac{\sum (\text{Return}_{ABC} - \text{Average}_{ABC}) * (\text{Return}_{XYZ} - \text{Average}_{XYZ})}{(\text{Sample Size})}$$

Interpretation

In the example there is a positive covariance, so the two stocks tend to move together. When one has a high return, the other tends to have a high return as well.

If the result was negative, then the two stocks would tend to have opposite returns; when one had a positive return, the other would have a negative return.

Correlation

Covariance can tell how the stocks move together, but to determine the strength of the relationship, we need correlation.

Quantifying the relationship

- *Correlation* describes the strength of the *linear* association between two variables.
- It takes values between -1 (perfect negative) and +1 (perfect positive).
- A value of 0 indicates no linear association.

Correlation

Since $Cov(X, Y)$ depends on the magnitude of X and Y we would prefer to have a measure of association that is not affected by changes in the scales of the variables.

The most common measure of *linear* association is correlation which is defined as

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

$$-1 < \rho(X, Y) < 1$$

Where the magnitude of the correlation measures the strength of the *linear* association and the sign determines if it is a positive or negative relationship.

Linear Relationship among factor variables/features

Correlation Analysis (Categorical Data)

- X^2 (chi-square) test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related

- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count

Chi-Square Calculation: An Example

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- Compute p-value from χ^2 value.

If p-value < 0.05 then the variables are not independent otherwise we conclude that the variables are dependent.

Common Tasks we do in Feature Engineering

1. Remove the variables that contains NA's more than some threshold

2. Remove the near-zero variance features

3. Handle the feature which are collinear i.e., whose correlation is higher

    a) Remove the features such that all pair-wise correlations are higher than some given threshold.

    b) Build the new features with linear combination of existing features sothat they are independent.