Chi – Square Test for Feature Selection

Purpose: Select Features which are most relevant to the Response variable

- The Independent and Dependent variables are categorical

The Formula for Chi Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

c = degrees of freedom

O = observed value(s)

E = expected value(s)

Hypothesis Testing

H_0: Two variables are Independent (Not Related to each other) → Feature is Not Selected

Ha: Two variables are Not Independent (Related to each other) → Feature is Selected

Example from Dataset

df['gender']: Male and Female
df['bichoice']: 0 or 1 (Electric Vehicle or Gasoline Vehicle)

Contingency Table

Buy Choice\ Gender	Electric Vehicle	Gasoline Vehicle	Total
Male	1724	1210	2934
Female	1469	1399	2868
Total	3193	2609	5802

Expected Value Calculation:
P(A and B) = P(A) * P(B)

Required Probability Calculation:
P(Male), P(Female), P(EV), P(GV) = 0.56, 0.494, 0.55, 0.44
P(Male and EV), P(Male and GV)
P(Female and EV), P(Female and GV)

Chi-Squared Value = 10.45

Accept or Reject Null Hypothesis

- 95% Confidence Interval → alpha = 0.05

K-Mode Clustering Algorithm

*Modified version of K-means Clustering Algorithm, designed for Categorical variables

Notations:

$$d_1(X,Y) = \sum_{i=1}^{n} \delta(x_i, y_i)$$

$$\delta(x_i, y_i) = \begin{cases} 0, & x_i = y_i \\ 1, & x_i \neq y_i \end{cases}$$

X and Y: Two Categorical Object
m: Number of categorical attributes

Initial K Observation: Define the number of K (cluster) to be defined

Dissimilarity Measure: Total Mismatches between the corresponding attribute values of two objects

- 0: When two categorical values are equal (Dissimilarity Score = 0)
- 1: When two categorical values are unequal (Dissimilarity Score = 1)

Comparing Each Sample to K-th Cluster -> Sample will be assign to the Cluster having lowest Dissimilar Score

Dissimilarity Score Table: Yellow Indicating Chosen Cluster for each sample

	Cluster 1	Cluster 2	Cluster 3
Sample 1	0	1	2
Sample 2	3	1	2
Sample 3	3	3	0
Sample 4	2	1	2

Choosing New K using Mode (Instead of Mean for K-means)

- Defined by the Frequency of categorical feature (ith), the most frequent feature within the cluster will be selected as a new standard for cluster

The Iteration Stops when there is no more change in the Update of Cluster

Optimal K Choice – Elbow Method

Within Cluster Difference

- K: Number of cluster
- m: Number of Observations in each cluster
- c: Centroid of the cluster
- d1: Dissimilarity measure

$$WCD = \sum_{j=1}^{k} \sum_{i=1}^{m} d_1(x_i, y_c)$$

Dissimilarity between the centroid of the K and the Sample (X) is measured for range of k (0-nth K)

*Elbow indicated the sudden decrease in the Dissimilarity measure (slope changes)

Monte Carlo Simulation

Purpose: Estimate the possible outcomes of uncertain events

Method that express the Law of Large numbers

- When sampling is done continuously from a population, the mean of the samples extracted will have the sample mean as the population mean

Detailed Explanation:
Probability of an event is unknown, therefore repeated occurrence of an event over time for many numbers can identify the Probability of an uncertain event → The resulting PDF is identical to the population PDF

Box Cox Transformation

Purpose: Transforming Non-Gaussian distribution into Gaussian-form Distribution

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

Exponent Lambda = Range of -5 to 5

- All the Lambda values are considered → Selection of Optimal Lambda (The one that provides best Normal-Shaped curve)

Common Box-Cox Transformations

Lambda value (λ)	Transformed data (Y')
-3	$Y^{-3} = 1/Y^3$
-2	$Y^{-2} = 1/Y^2$
-1	$Y^{-1} = 1/Y^1$
-0.5	$Y^{-0.5} = 1/(\sqrt{(Y)})$
0	log(Y)**
0.5	$Y^{0.5} = \sqrt{(Y)}$
1	$Y^1 = Y$
2	Y^2
3	Y^3