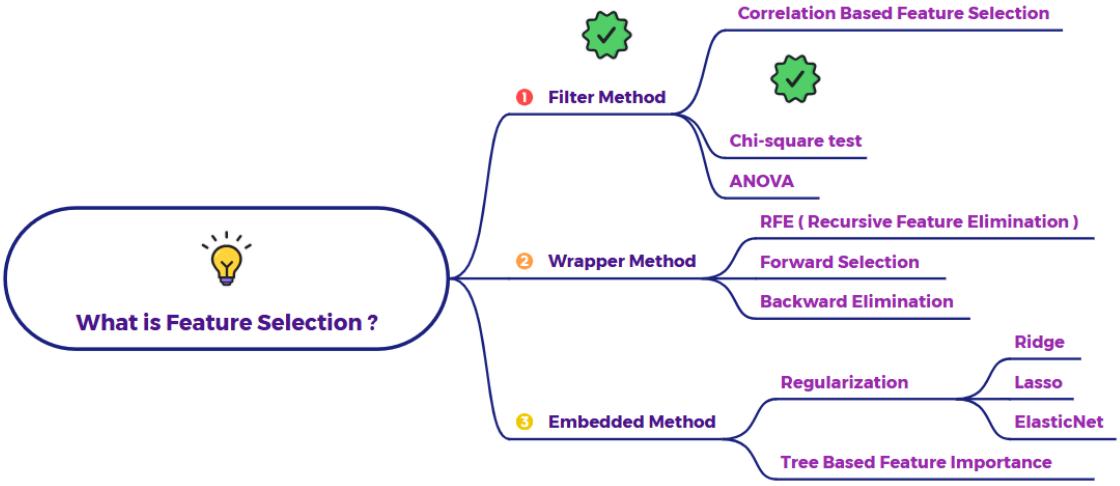
Explain Chi-square test-based feature selection



let's explore the **Filter Method - Chi-Square Test Based Feature Selection**. This method is particularly useful when you have a **categorical target variable,** and you want to select the categorical features that are most likely to be dependent on it.

Core Idea:

The Chi-Square test (χ^2) is a statistical test that measures the independence between two categorical variables. In the context of feature selection, we use it to determine if there's a statistically significant association between each categorical feature and the categorical target variable. If a feature is strongly associated with the target, it's likely to be a good predictor.

How it Works:

1. **Create Contingency Tables:** For each categorical feature, you create a contingency table (also known as a cross-tabulation) that shows the frequency distribution of the feature's categories across the categories of the target variable.
2. **Calculate the Chi-Square Statistic:** For each contingency table, the Chi-Square statistic is calculated. This statistic quantifies the difference between the observed frequencies in the table and the frequencies we would expect if the two variables were completely independent. The formula for the Chi-Square statistic is:

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where:

- O_{ij} is the observed frequency of the cell at the i-th row and j-th column.
- E_{ij} is the expected frequency of the cell at the i-th row and j-th column under the assumption of independence.
- R is the number of rows (categories in the feature).
- C is the number of columns (categories in the target variable).

3. **Calculate the p-value:** The Chi-Square statistic is then used to calculate a p-value based on the degrees of freedom (df = (R-1) * (C-1)). The p-value represents the probability of observing the data (or more extreme data) if the feature and the target were truly independent.
4. **Rank Features:** Features are ranked based on their Chi-Square statistic value (higher values indicate stronger association) or their p-value (lower values indicate stronger evidence against independence).
5. **Select Top Features:** You select the top-ranked features based on a chosen threshold for the Chi-Square statistic or the p-value. Typically, you would select features with a low p-value (below a significance level like 0.05), suggesting a statistically significant association with the target. Alternatively, you might select the features with the highest Chi-Square statistic values.

Example:

Let's say we want to predict whether a customer will **purchase a product (Yes/No)**. Our target variable is "Purchase (Yes/No)" (categorical). We have a categorical feature: **"Marketing Channel"** with categories: "Email", "Social Media", "Website", "Referral".

1. Create Contingency Table for "Marketing Channel" and "Purchase":

Marketing Channel	Purchase = Yes	Purchase = No	Total
Email	80	120	200
Social Media	50	150	200
Website	100	100	200
Referral	120	80	200
Total	**350**	**450**	**800**

2. Calculate Expected Frequencies:

Under the assumption of independence, the expected frequency for each cell is calculated as:

$$E_{ij} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

For example, the expected frequency for "Email" and "Purchase = Yes" is:
$E_{11} = \frac{200 \times 350}{800} = 87.5$

Calculating expected frequencies for all cells:

Marketing Channel	Purchase = Yes (Expected)	Purchase = No (Expected)	Total
Email	87.5	112.5	200
Social Media	87.5	112.5	200
Website	87.5	112.5	200
Referral	87.5	112.5	200
Total	**350**	**450**	**800**

3. Calculate the Chi-Square Statistic:

$$\chi^2 = \frac{(80-87.5)^2}{87.5} + \frac{(120-112.5)^2}{112.5} + \frac{(50-87.5)^2}{87.5} + \frac{(150-112.5)^2}{112.5} + \frac{(100-87.5)^2}{87.5} + \frac{(100-112.5)^2}{112.5} + \frac{(120-87.5)^2}{87.5} + \frac{(80-112.5)^2}{112.5}$$

$$\chi^2 \approx 0.643 + 0.494 + 16.071 + 12.321 + 1.786 + 1.375 + 12.321 + 9.494 \approx 54.505$$

4. Calculate the p-value:

The degrees of freedom (df) for this table are (4 - 1) * (2 - 1) = 3. We would then look up the p-value associated with a Chi-Square statistic of approximately 54.505 with 3 degrees of freedom. The p-value would be very small (much less than 0.05).

5. Interpret and Select:

A very small p-value indicates that there is strong evidence to reject the null hypothesis of independence between "Marketing Channel" and "Purchase". This suggests that the marketing channel used has a statistically significant association with whether a customer makes a

purchase. Therefore, "Marketing Channel" would be considered a relevant feature for predicting "Purchase" based on the Chi-Square test.

How to Handle Multiple Categorical Features:

If you have multiple categorical features, you would perform this Chi-Square test for each feature against the target variable independently. Then, you would rank the features based on their Chi-Square statistic (higher is better) or their p-value (lower is better) and select the top-k features or those that meet a certain significance level.

Limitations of Chi-Square Test for Feature Selection:

- **Only for Categorical Target:** The Chi-Square test is specifically designed for categorical target variables. If your target variable is numerical, you would need to use different filter methods (like correlation with numerical features or ANOVA if the features are categorical).
- **Sensitivity to Sample Size:** The Chi-Square statistic can be sensitive to large sample sizes, potentially leading to statistically significant results even for weak associations.
- **Assumptions:** The Chi-Square test has certain assumptions (e.g., expected frequencies should not be too small).
- **Only Evaluates Individual Feature Relevance:** Like other filter methods, it evaluates the relationship of each feature with the target in isolation and doesn't consider feature interactions.

In summary, Chi-Square Test based feature selection is a valuable filter method for identifying categorical features that have a statistically significant association with a categorical target variable. By analyzing the contingency tables and the resulting Chi-Square statistics and p-values, you can select the most relevant categorical predictors.