How Chi-Squared Test of Independence Informs Feature Engineering for Logistic Regression

Overview

Feature engineering is a critical step in building a robust logistic regression model, as it involves selecting, creating, or transforming features that improve model performance. When working with categorical data, the **chi-squared test of independence** can inform feature engineering by identifying statistically significant relationships between categorical variables and the target variable, indicating which variables are likely to be valuable predictors.

Chi-Squared Test of Independence: Purpose and Application

The chi-squared test of independence assesses whether there is a **statistical association between two categorical variables**. It provides a p-value to test the null hypothesis that the variables are independent, with low p-values (typically < 0.05) suggesting a significant association. In feature engineering, these insights help us focus on features that provide meaningful information about the target variable.

Steps to Use Chi-Squared Test Results in Feature Engineering

1. **Identify Candidate Variables**:
    - Use chi-squared tests to analyze categorical variables (both potential predictors and the target variable) in the dataset.
    - Create a contingency table for each predictor vs. target pair, then apply the chi-squared test to determine associations.
2. **Interpret the Chi-Squared Results**:
    - A **significant p-value** (e.g., $p < 0.05$) suggests an association, indicating that the predictor variable may have a relationship with the target and could be informative for the logistic regression model.
    - A **non-significant p-value** (e.g., $p > 0.05$) suggests independence, indicating that the predictor may not contribute much to the model.
3. **Guide Feature Selection**:
    - **Select significant variables** as candidate predictors: Include only the variables that show significant association with the target, focusing the model on predictors with statistical relevance.

   - **Filter out non-significant variables**: This simplifies the model, reduces dimensionality, and improves interpretability by excluding less relevant variables.

   - For instance, if the chi-squared test reveals that **Customer Segment** and **Region** are significantly associated with the likelihood of a purchase (target variable), these are strong candidates for inclusion.

4. **Engineer Interaction Terms**:

   - For variables that show a significant association with the target and with each other, consider creating **interaction terms**. Interaction terms capture joint effects of predictor pairs, which can provide additional predictive power in logistic regression.

   - Example: If **Segment** and **Order Month** are both associated with the target, creating an interaction term (Segment * Order Month) might enhance the model by capturing combined effects on purchase likelihood.

5. **Determine Variable Encoding Based on Association**:

   - For significant categorical variables, use appropriate encoding (e.g., **one-hot encoding** or **binary encoding**) to ensure that logistic regression can interpret these features.

   - Chi-squared informs encoding by highlighting which categories within each variable are associated with the target. In cases where there are many categories, it may be useful to consolidate or group categories with low or similar association levels.

Example Workflow

Suppose we aim to predict **purchase probability** (Yes/No) based on customer demographics and transaction details:

1. Conduct chi-squared tests between the target variable (Purchase) and predictors like **Segment**, **Region**, **Ship Mode**, and **Order Month**.

2. Interpret results:

   - If **Segment** and **Region** have significant p-values, include them as predictors.

   - If **Ship Mode** shows no association (high p-value), consider excluding it.

3. Engineer features based on interactions, e.g., **Segment * Order Month**, if both variables are significantly associated with the target.

Benefits of Using Chi-Squared Test Results in Feature Engineering

- **Improved Model Accuracy**: By selecting only the relevant features, we reduce noise and make the model more accurate.
- **Reduced Dimensionality**: Excluding irrelevant features improves model efficiency and interpretability.
- **Enhanced Interpretability**: Each feature in the final model has a statistically justified relationship with the target variable, making model results easier to explain.

Limitations

While chi-squared tests are valuable, they do not measure the **strength** or **direction** of relationships as correlation measures do with continuous data. Therefore, additional analysis may be required to fully understand relationships and optimize feature engineering.