Data Mining
HW 4

Omid Naeej Nejad
610301189

November 7, 2024

1 Question 1

- **a. Steps to Explore and Address Skewness in Review Length:**

  1. **Descriptive Statistics:** Start by calculating descriptive statistics (mean, median, mode, standard deviation, and skewness) for Review Length across different Product Categories.
  2. **Visual Inspection:** Create ***histograms or box plots*** to visually inspect the distribution of Review Length for each Product Category in the data set. Visualization helps you to identify patterns and the degree of skewness.
  3. **Transformation:** If the skewness is significant, consider applying transformations to normalize the data and reduce skewness:
     - Log Transformation
     - Square Root Transformation
  4. **Impact on Analysis:** Addressing skewness is crucial as it can affect statistical analyses (e.g., t-tests, ANOVA) that assume normality. Skewed data can lead to inaccurate results, so normalizing the data can improve the reliability of your findings.

- **b. Visualization Techniques to Compare Distributions of Review Length across Product Categories**

1. ***Box Plots:*** Create box plots for Review Length segmented by Product Category. Box plots effectively show the median, quartiles, and potential outliers, making it easy to compare distributions across categories.
2. ***Overlaid Histograms & Normalized Histograms:*** Plot histograms of Review Length and mark different Product Categories by different colors.

2 Question 2

- **a. Methods for Encoding Categorical Variables**

  When preparing categorical variables for predictive modeling, there are several encoding methods available, each with its advantages and disadvantages. Here are some common methods:

  1. ***Label Encoding:*** Each category is assigned a unique integer value. For example, Silver = 0, Gold = 1, Platinum = 2.
     This method is simple and preserves the order of categories if they are ordinal.
  2. ***One-Hot Encoding:*** Each category is converted into a binary indicator (0 or 1). For example, for Membership Tier, you would create three new columns: Silver (0 or 1), Gold (0 or 1), and Platinum (0 or 1). Each row would have a 1 in the column corresponding to its category and 0 elsewhere.

     This method avoids introducing any ordinal relationships and works well with non-ordinal categorical variables. However, it can lead to a high-dimensional feature space if there are many categories.
  3. ***Binary Encoding:***
     For example: $Silver = 01$, $Gold = 10$, $Platinum = 00$

     This method reduces dimensionality compared to one-hot encoding.

- **b.** If the Membership Tier has an ordinal relationship (i.e., $Silver < Gold < Platinum$), **Label Encoding** would be the most appropriate choice.

Why Choose Ordinal Encoding for Ordinal Data?

- **Preserves Order:** It retains the rank order of the categories, which is crucial for ordinal data.
- **Dimensionality Maintenance:** Unlike one-hot encoding, it does not increase the dataset dimension.
- **Algorithm Compatibility:** Many algorithms, like decision trees, can benefit from the preserved order in the data.

3 Question 3

- ***a.***

Problems from Including Both Annual Income and Credit Score as Predictors When two (or more) predictors are highly correlated, some learning algorithms e.g. Linear Regression couldn't learn and some others couldn't explorate and recognize patterns and trends perfect. Also it can lead to unreliable estimates of the coefficients, making it difficult to determine the individual effect of each predictor on the target variable.

So:
- At best, this will overemphasize one data component.
- At worst, it will cause the model to become unstable and deliver unreliable results.

Strategies to Address Correlated Features During EDA

1. **Correlation Matrix:** Generate a correlation matrix to quantify the correlation between predictors. This helps identify which variables are highly correlated.
2. **Scatter Matrix Plot** If observed points occurred in a line, that numeric features would be correlated.
3. **Principal Component Analysis (PCA):** Using PCA to transform correlated predictors into a smaller set of uncorrelated components, retaining most of the variance in the data.
4. **Project Domain Knowledge:** to decide which variable is more relevant or reliable for your predictive model and consider dropping one of them.

- **b. Investigating and Visualizing the Interaction Between Loan Amount Requested and Loan Default Across Different Ranges of Annual Income.**

  How you can investigate and visualize the interaction between Loan Amount Requested and Loan Default:

  Investigative Steps:

  1. **Segmenting Data**
     Divide Annual Income into categories (e.g., low, medium, high) based on quantiles or business logic. This allows you to analyze how Loan Default rates change within these segments.
  2. **Group Analysis**
     Calculate the default rate for each segment of Annual Income at different levels of Loan Amount Requested. This can be done using aggregation functions (e.g., mean, count) to summarize defaults.

  Visualization Techniques:

  - **Clustered Bar Plots**
    Create clustered bar plots where each group represents a segment of Annual Income, and within each group, bars represent different ranges of Loan Amount Requested. This visually illustrates how default rates vary across income levels for different loan amounts.
  - **Box Plots**
    Use box plots to show the distribution of Loan Amount Requested for each category of Loan Default (Yes/No) within segments of Annual Income. This can highlight differences in loan amounts associated with defaults across income groups.

  Importance of Understanding This Interaction:

  - **Risk Assessment & Determining Strategies:**
    Understanding how Loan Amount Requested interacts with Annual Income helps in assessing risk more accurately for different borrower profiles.
  - **Model Improvement:**

Recognizing interactions can lead to more complex models that capture these relationships better, potentially improving predictive performance.