Data Mining - EX4

Deadline: Friday, Aban 11, 1403 - November 01, 2024

Question 1:

You are analyzing a dataset of customer reviews for an e-commerce platform. The dataset includes a numerical variable `Review Length` (number of words in the review) and a categorical variable `Product Category`. Upon plotting, you notice that `Review Length` is heavily right-skewed, and the degree of skewness varies across different `Product Categories`.

- **a.** What steps would you take to explore and address the skewness in `Review Length`, and how might this impact your analysis?
- **b.** How can you use visualization techniques to compare the distributions of `Review Length` across different `Product Categories`?

Question 2:

In preparing data for a predictive model, you are given a dataset containing various customer attributes, including a categorical variable `Membership Tier` with levels: *Silver*, *Gold*, and *Platinum*. You plan to include this variable in your analysis but are concerned about how to appropriately encode it.

- **a.** Explain the different methods available for encoding categorical variables like `Membership Tier` and discuss the potential impact of each method on your analysis.
- **b.** Which encoding method would you choose if the `Membership Tier` has an ordinal relationship, and why?

Question 3:

You are analyzing a financial dataset containing variables such as `Annual Income`, `Credit Score`, `Loan Amount Requested`, and a binary target variable `Loan Default` (Yes or No). During your exploratory data analysis, you notice that `Annual Income` and `Credit Score` are highly correlated.

- **a.** What problems might arise from including both `Annual Income` and `Credit Score` as predictors in a predictive model, and what strategies can you employ during EDA to address these issues?

- **b.** Additionally, suppose you discover that the relationship between `Loan Amount Requested` and `Loan Default` varies significantly across different ranges of `Annual Income`. How would you investigate and visualize this mutual effect during EDA, and why is understanding this interaction important for your analysis?

---

practical exercise: Exploring Customer Churn Data with EDA

Objective

The objective of this exercise is to explore a customer churn dataset from a telecommunications company. By performing various EDA operations, you will learn how to understand data structure, clean data, analyze distributions, and draw insights that can help in understanding customer behavior and churn patterns.

You can use **this link** to access the data. Also upload your answer in E-learn as a PDF or ipynb or R Markdown file. Descriptions and codes must be complete and detailed.

Instructions

Step 1: Load and Inspect the Data

1. **Load the Data**: Read the CSV file into a DataFrame.
2. **Inspect the Structure**: Check the number of rows and columns, data types of each column, and the first few rows of the dataset.
3. **Questions**:
   - What types of variables are in the dataset (e.g., numerical, categorical)?
   - What are the column names, and what do you think they represent?

Step 2: Data Cleaning

1. **Check for Missing Values**: Identify any missing values in the dataset. Note if any column has a high percentage of missing values.
2. **Check for Duplicates**: Verify if there are duplicate rows, and remove them if necessary.
3. **Data Types**: Ensure each column has an appropriate data type (e.g., numerical for `Account Length`, categorical for `Int'l Plan`).
4. **Questions**:

- Are there any missing values or duplicates? If so, how would you handle them?
- Are there any columns where the data type needs adjustment?

Step 3: Descriptive Statistics

1. **Summarize Numerical Features**: Calculate basic descriptive statistics (mean, median, standard deviation) for numerical columns.
2. **Summarize Categorical Features**: Check the unique values and their frequencies for categorical columns.
3. **Questions**:
   - What do the mean and median tell you about each numerical feature?
   - How are the values distributed in categorical columns?

Step 4: Univariate Analysis

1. **Visualize Distributions**: Use histograms for numerical features (e.g., `Account Length`, `Day Mins`, `Intl Mins`) and bar plots for categorical features (e.g., `Int'l Plan`, `VMail Plan`).
2. **Analyze Churn Rate**: Plot the churn rate to understand the proportion of customers who have churned versus retained.
3. **Questions**:
   - What is the churn rate, and what does it tell you about the customer base?
   - Are there any features that show skewness or unusual patterns?

Step 5: Bivariate Analysis

1. **Service Plans vs. Churn**: Plot the churn rate against service plans (`Int'l Plan`, `VMail Plan`) to see if customers with certain plans have higher churn rates.
2. **Customer Service Calls vs. Churn**: Visualize the number of customer service calls segmented by churn status.
3. **Usage and Churn**: Compare distributions of high-usage metrics (e.g., `Day Mins`, `Intl Mins`) between churned and non-churned customers.
4. **Questions**:
   - Do customers with international or voicemail plans tend to churn more or less?
   - Is there a relationship between customer service call frequency and churn?
   - Do high-usage customers show higher churn rates?

Step 6: Scatter Plot Analysis

1. **Usage Features Correlation**: Create scatter plots to visualize the relationship between pairs of usage features such as:
   - `Day Mins` vs. `Day Charge`
   - `Intl Mins` vs. `Intl Charge`
2. **Churn Highlighting**: Color the points by churn status to observe if certain patterns emerge between these features for churned and non-churned customers.
3. **Questions**:
   - Is there a direct linear relationship between minutes used and charges? (This can help verify feature accuracy or identify outliers.)
   - Do certain usage patterns distinguish churned customers from non-churned customers?

Step 7: Draw Conclusions

1. **Summarize Insights**: Based on your findings, list the main factors that seem to influence customer churn.
2. **Interpret Trends**: Consider why certain features might impact churn. For example, why might frequent customer service calls correlate with churn?
