Step 1: Advanced Exploratory Data Analysis (EDA)

Tutor: Haythem Ghazouani

1 Overview

Exploratory Data Analysis (EDA) is the process of using summary statistics and visualizations to understand the structure of your data. This tutorial goes beyond basic averages to uncover hidden interactions between features.

Implementation Note

The source code for this EDA is provided in `code/data_exploration.py`.

2 Phase 1: Structure & Quality Check

Before visualizing, you must validate the data's integrity.

- **Imbalance Check:** Does the target variable have a huge disparity? (e.g., in Churn, if 95% stay, a model predicting "Stayed" only will have 95% accuracy but 0 value).
- **Missingness:** Are there N/A values that require imputation?
- **Data Leakage:** Remove IDs or timestamps that uniquely identify rows but provide no predictive power.

3 Phase 2: Deep Univariate Analysis

Focus on the spread of single variables.

- **Skewness:** Use histograms with Kernel Density Estimation (KDE). If data is skewed, you may need a log-transform later.
- **Outlier Detection:** Use Boxplots. Outliers can skew your model's perception of "normal" behavior.

4 Phase 3: Advanced Bivariate Analysis

How does a feature act when conditioned on the target?

Step Categorical vs. Target

Compare percentages using stacked bar charts or `countplot` with `hue`. This reveals if certain groups (e.g., Gender or Country) are significantly more likely to churn.

Step Numerical vs. Target

Use Violin plots or Boxplots. Does the distribution of "Balance" shift significantly for customers who left? If so, Balance is a strong predictor.

5 Phase 4: Multivariate Analysis

- **Pairplots:** Visualize relationships between multiple pairs of variables at once. This helps identify clusters or non-linear trends.
- **Correlation Heatmap:** Identify multicollinearity. If two features correlate at 0.9+, you might only need one of them.

6 Code Implementation

```
import seaborn as sns
import matplotlib.pyplot as plt

# 1. Categorical Impact
sns.countplot(x='Geography', hue='Exited', data=df)
plt.title('Churn by Country')
plt.show()

# 2. Outlier Analysis
sns.boxplot(y=df['Balance'], x=df['Exited'])
plt.title('Balance Distribution by Churn Status')
plt.show()

# 3. Feature Interaction
sns.pairplot(df[['Age', 'Balance', 'CreditScore', 'Exited']], hue=
    'Exited')
plt.show()
```

Statistical Note

Correlation does not imply causation. A high correlation between "Age" and "Churn" in the Bank dataset might be due to life stages rather than the bank's services themselves.

7 Self-Guided Tasks

1. Calculate the median age of customers who exited versus those who stayed.
2. Find which `NumOfProducts` has the highest churn rate.
3. Identify if "Active Membership" mitigates the impact of a low "Credit Score".

Data Reference

The primary dataset used here is `data/bank_churn.csv`. All visualization outputs (PNGs) are saved directly into the `data/` folder for inclusion in your reports.