Step 1: Exploratory Data Analysis (EDA)

University: Tek-Up
Group: ING-4-SDIA

Tutor: Haythem Ghazouani

1 Overview

Exploratory Data Analysis (EDA) is the process of analyzing datasets to summarize their main characteristics, often with visual methods. In this step, we will explore the **Bank Customer Churn** dataset.

2 Dataset Structure

The dataset contains information about bank customers and whether they left the bank (`Exited`). Key columns include:

- `CreditScore`, `Age`, `Tenure`, `Balance`.
- `NumOfProducts`, `HasCrCard`, `IsActiveMember`.
- `Exited` (Target variable).

3 Essential Python Tools

For this module, we use:

- **Pandas:** For data manipulation and summary stats.
- **Seaborn/Matplotlib:** For static visualizations.
- **Plotly (Optional):** For interactive dashboards.

4 Implementation

The following code (available in `code/data`$_{e}xploration.py)performsabasicEDA$:

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df = pd.read_csv("data/bank_churn.csv")

# 1. Distribution of the Target Variable
```

```
sns.countplot(x='Exited', data=df)
plt.title('Churn Count')
plt.show()

# 2. Correlation Analysis
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.show()
```

5 Key Insights to Look For

During your analysis, ask yourself:

1. Is the dataset balanced? (Ratio of 0s to 1s in Exited).
2. Which features are most correlated with churn?
3. Are there outliers in the Balance or EstimatedSalary?

6 Exercise

Create a histogram for the Age column, grouped by the target variable Exited. What can you conclude about the relationship between age and customer churn?