Customer Churn Prediction

Task 1: Data Gathering, Exploratory Data Analysis, and Data Preparation Report

1. Introduction
The objective of this task is to prepare a high-quality dataset for building a customer churn prediction model. This involves gathering relevant data sources, performing exploratory data analysis (EDA), and applying appropriate data cleaning and preprocessing techniques to ensure reliability and predictive performance.

2. Data Sources and Selection Rationale
The analysis combines multiple customer-level datasets joined using a unique customer identifier (CustomerID). Each dataset captures a different dimension of customer behavior and experience.

Customer Demographics: Age, Gender, Marital Status, Income Level. These variables provide contextual background and may indirectly influence churn behavior.

Transaction History: Transaction count, amount spent, and product category. Transaction behavior is a strong indicator of engagement and customer value.

Customer Service Interactions: Interaction type, service usage, and resolution status. High interaction frequency and unresolved issues are commonly associated with dissatisfaction and churn.

Online Activity: Login frequency and last login date. Lower engagement often precedes customer churn.

Churn Status: The target variable indicating whether a customer has churned (1) or not (0).

3. Exploratory Data Analysis (EDA)
The consolidated dataset contains 6,812 customers and 17 features. The observed churn rate is approximately 20%, indicating moderate class imbalance. This reflects a realistic business scenario and necessitates the use of evaluation metrics beyond accuracy.

Missing values were primarily found in behavioral and service-related variables and were determined to represent lack of activity rather than data quality issues. These missing values were therefore retained and treated appropriately during preprocessing.

Univariate analysis revealed right-skewed transaction amounts and varying engagement levels. Bivariate analysis showed that behavioral features such as login frequency, service usage, and transaction activity exhibited stronger separation between churned and retained customers compared to demographic variables.

4. Data Cleaning and Preprocessing
Customer identifiers were excluded from modeling features. Numeric missing values were imputed using the median, while categorical missing values were imputed using the most

frequent category. Outliers were capped using an interquartile range approach. Numeric features were standardized using z-score normalization, and categorical variables were encoded using one-hot encoding.

5. Final Prepared Dataset

The final dataset is fully cleaned, encoded, and scaled, making it suitable for training machine learning classification models. It includes all relevant predictors along with the churn target variable and is ready for model development.

6. Conclusion

This task established a strong foundation for customer churn modeling. Behavioral and service-related features emerged as the most influential predictors, while demographic variables provided supportive context. The resulting dataset is well-prepared for subsequent model training and evaluation.