Milestone 1: Exploratory Data Analysis & Preprocessing

Customer Churn Analysis

1. Introduction

This report presents a systematic exploratory data analysis (EDA) and preprocessing workflow applied to the Churn dataset collected from two raw files (churn-bigml-80.csv and churn-bigml-20.csv).

The goal of this milestone is to ensure data integrity, identify structural patterns, handle inconsistencies, and prepare a clean dataset suitable for downstream modeling tasks.

2. Data Acquisition

Two raw datasets were loaded and concatenated vertically:

Data sources:

churn-bigml-80.csv

churn-bigml-20.csv

Operation: pd.concat([df1, df2], axis=0)

Result: A unified dataset containing all observations across both files.

This ensured full data coverage before conducting any preprocessing or exploratory examination.

3. Exploratory Data Analysis

3.1 Structural Assessment

Key dataset diagnostics included:

Shape: The combined dataset's dimensions were inspected to determine the total number of rows and columns.

Missing values: A column-wise assessment confirmed the absence of null entries prior to preprocessing.

Duplicate entries: Duplicate rows were quantified to evaluate potential redundancy.

Data types: Using data.info(), numerical and categorical variables were identified.

This preliminary examination provided a foundational understanding of dataset quality and structure.

3.2 Feature Classification

Numerical features: Extracted using select_dtypes(include=[np.number]).

Categorical features: Extracted using select_dtypes(exclude=[np.number]).

This classification later informed the choice of preprocessing techniques for each feature type.

3.3 Distribution and Behavioral Patterns

Multiple visualization techniques were used to analyze customer behavior and churn patterns:

Univariate plots: Histograms and boxplots for numerical variables.

Categorical distributions: Countplots and pie charts for:

International plan

Voice mail plan

Area code

Customer service calls

Churn imbalance check: Class distribution revealed clear class imbalance.

Multivariate exploration:

Pair plots to inspect variable interactions.

Churn-conditioned histograms.

Scatterplots for aggregated national charges vs. minutes.

These visual analyses highlighted behavioral distinctions between churners and non-churners and provided insights into feature relevance.

3.4 State-Level Churn Analysis

Top states contributing to churn and non-churn populations were identified:

Extracted via value_counts().head(5)

Visualized using comparative pie charts.

This offered a geographic perspective on customer churn tendencies.

4. Data Cleaning & Preprocessing

4.1 Encoding Categorical Variables

Binary categorical variables were transformed using direct mapping:

International plan: "Yes"→1, "No"→0

Voice mail plan: "Yes"→1, "No"→0

Churn: True→1, False→0

This ensured consistency and model-readiness.

4.2 Standardization of U.S. States

State abbreviations were converted into full state names using the us Python library:

Custom function state_to_full() applied across the State column

"DC" explicitly mapped to "District of Columbia"

This standardization eliminated categorical fragmentation and improved downstream interpretability.

4.3 Outlier Detection and Removal

Outliers were processed using the Interquartile Range (IQR) method:

Applied to selected continuous variables:

["Account length","Total day minutes","Total eve minutes","Total night minutes","Total eve calls","Total night calls"]

Computation:

IQR = Q75 − Q25

Outlier bounds:

min = Q25 − 1.5 * IQR

max = Q75 + 1.5 * IQR

Handling strategy:

Values outside bounds were replaced with NaN

All rows containing outliers were removed using dropna()

This approach ensured a robust dataset free from extreme anomalies without applying capping or scaling.

4.4 Data Type Normalization

Following outlier removal:

All float64 columns were converted to int64 where appropriate.

The State column was cast to a categorical type.

This enforced schema consistency and optimized memory efficiency.

4.5 Feature Engineering

Two aggregate variables were created:

Total national minutes

= day minutes + evening minutes + night minutes

Total national charge

= day charge + evening charge + night charge

These engineered features provided holistic insight into customer usage patterns.

4.6 Saving the Processed Dataset

The final cleaned dataset was exported to:

data/processed/churn_cleaned.csv

This file serves as the official input for subsequent modeling tasks.

5. Correlation Analysis

A correlation heatmap was constructed based on numerical features:

Identified strongest predictors of churn.

Highlighted variables with negligible predictive contribution.

Provided early indications of potential multicollinearity.

This analytical step guided expectations for subsequent feature selection and modeling.

6. Conclusion

This milestone successfully:

Integrated raw datasets into a unified structure.

Conducted a detailed exploratory analysis of churn behavior.

Performed cleaning, encoding, normalization, and outlier removal.

Engineered relevant aggregated features.

Saved a high-quality processed dataset ready for advanced analysis.

These steps establish a rigorous foundation for statistical modeling and machine learning pipelines in future