Bank Marketing Dataset: EDA, Regression, Classification, Clustering & Pattern Mining

University of the Cumberlands

Advanced Big Data and Data Mining

Instructor: Dr. Satish Penmatsa

Group Members :

Suresh Ghimire

Sagar Bhetwal

Nirajan Acharya

Umesh Dhakal

Enjal Chauhan

Date: 15 February, 2026

Presentation Link :

https://cumber-my.sharepoint.com/:v:/g/personal/sghimire38288_ucumberlands_edu/IQB1BVQMvYzYRpO1aTxZtL9_AfiIvO3bNbkV5mlvI5MfZrI?nav=eyJyZWZlcnJhbEluZm8iOnsicmVmZXJyYWxBcHAiOiJTdHJlYW1XZWJBcHAiLCJyZWZlcnJhbFZpZXciOiJTaGFyZURpYWxvZy1MaW5rIiwicmVmZXJyYWxBcHBQbGF0Zm9ybSI6IldlYiIsInJlZmVycmFsTW9kZSI6InZpZXcifX0%3D&e=s7rWwT

Roadmap

EDA & Cleaning (Deliverable 1)

Regression Modeling (Deliverable 2)

Classification, Clustering, Association Rules (Deliverable 3)

Key findings + implications

Dataset Overview

- Dataset: Bank Marketing (rows ≈ 45,211)
- Target used in EDA & classification: y (subscription: 0/1)
- Mixed features: numeric + categorical (job, education, loan, etc.)
- Common “unknown” coded values; pdays has special meaning

Data Validation & Cleaning Decisions

Validation checks: shape, schema, duplicates, missing values

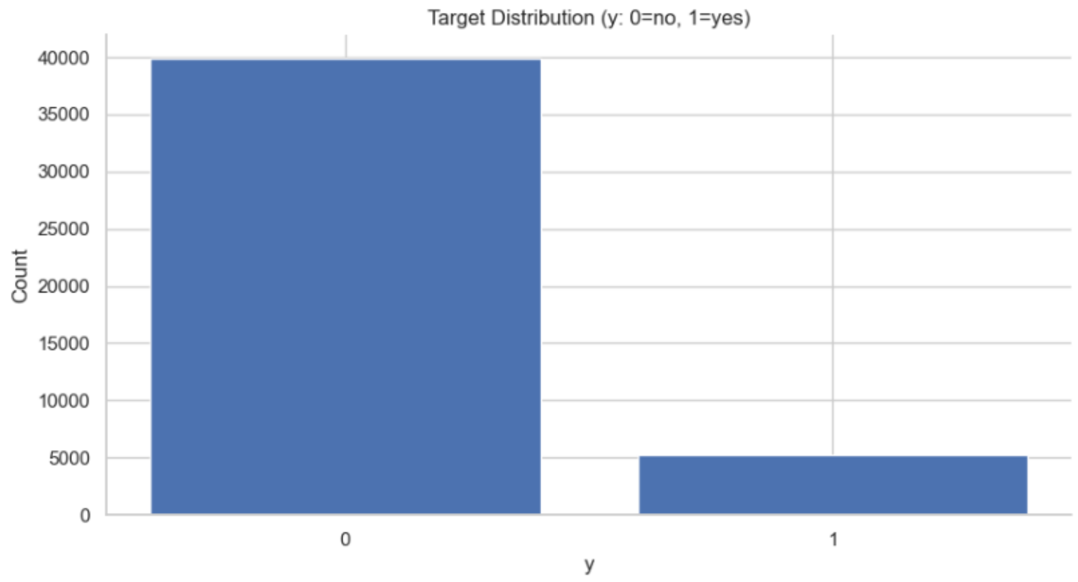
Cleaning actions from notebook:

- Standardize column names
- Map target y: {no->0, yes->1}
- Convert pdays = -1 to missing + create prev_contacted indicator
- Add *_unknown flags for categorical columns containing “unknown”

Target Balance (Class Imbalance)

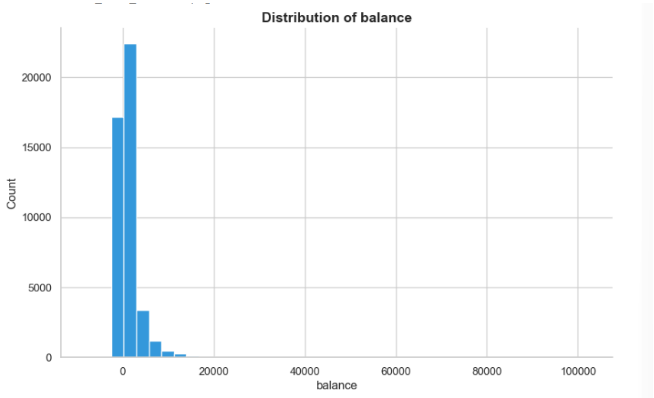
- What to show:
    - Short explanation: subscription “yes” is minority

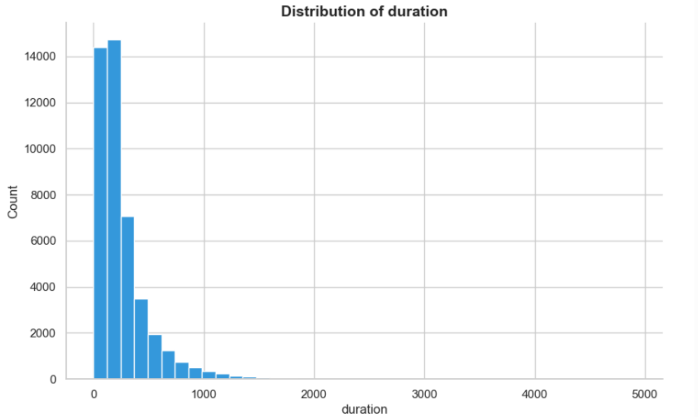
y	count	percent
0	39922	88.3
1	5289	11.7

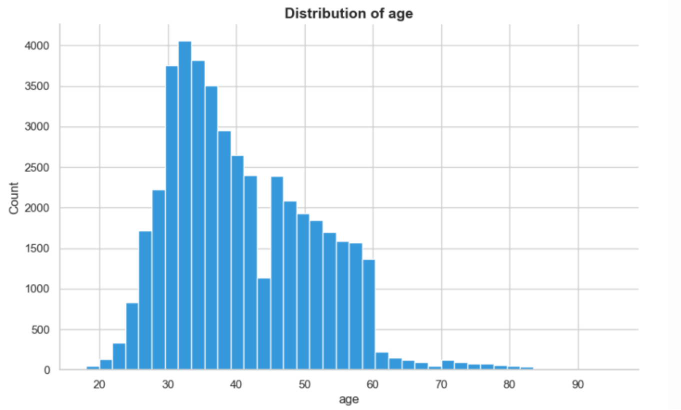


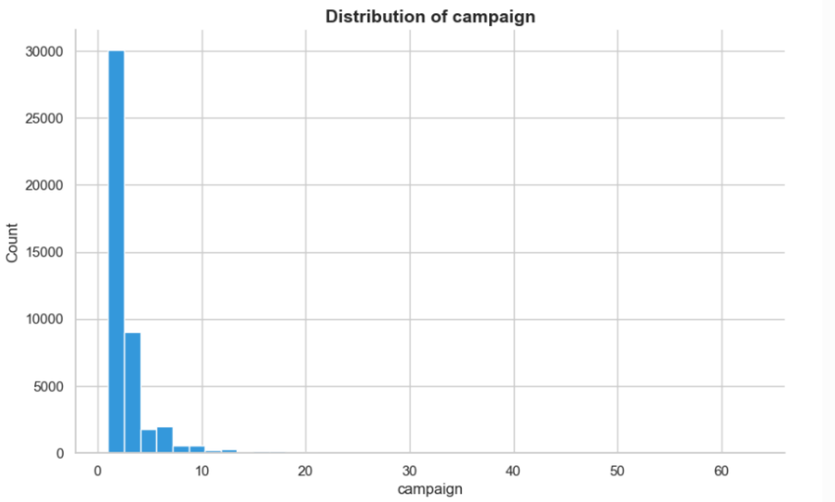
Numeric Feature Distributions

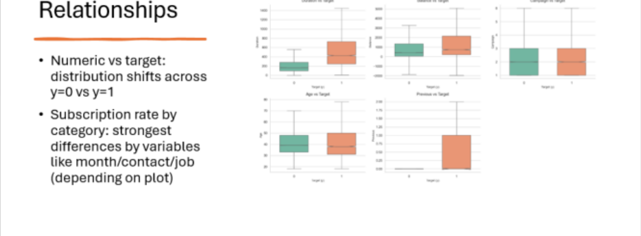
- Brief callouts: skew and long tails (balance, duration, campaign)











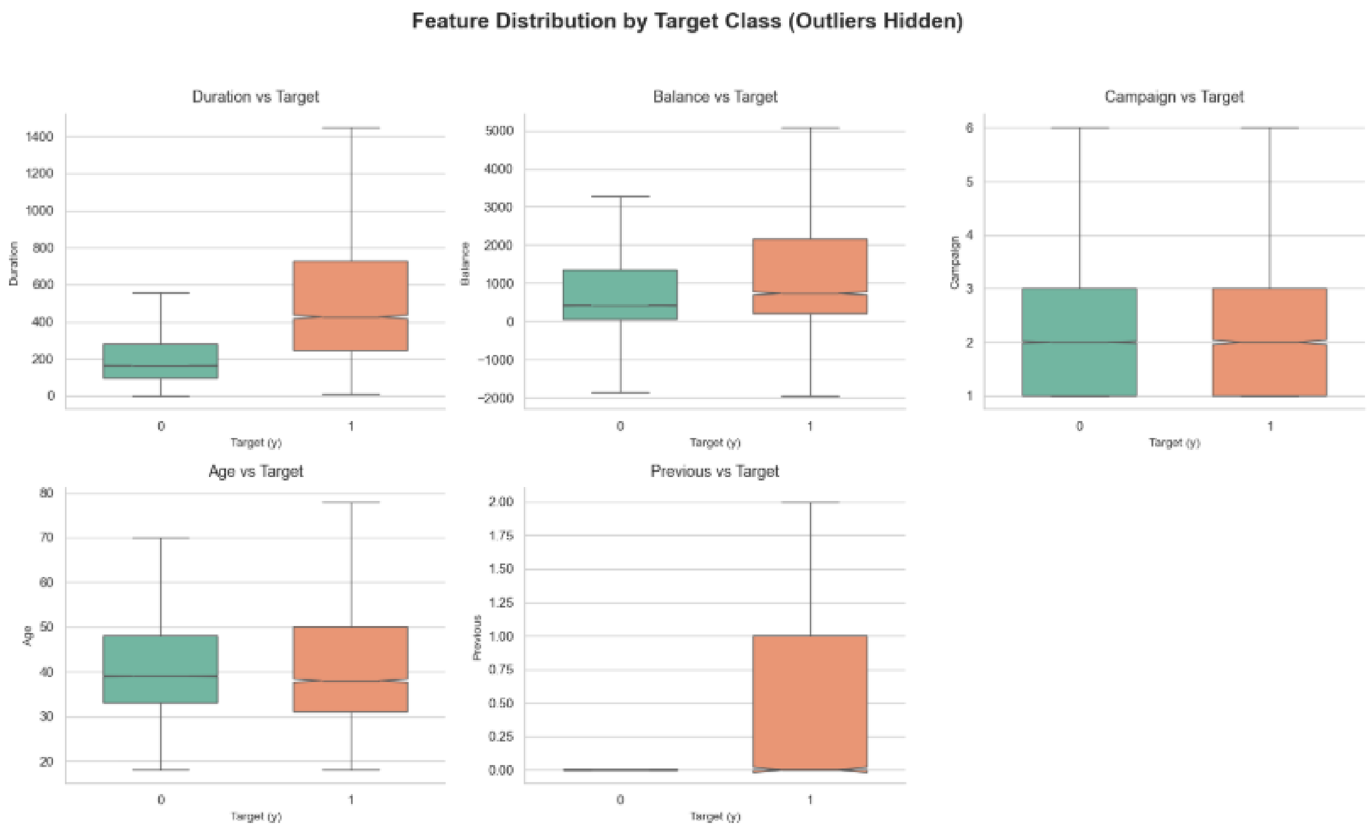
Outliers + Categorical Distributions

Outliers observed in balance, duration, campaign, and previous

Categorical breakdowns show dominant categories and “unknown” presence

Feature - Target Relationships

- Numeric vs target: distribution shifts across y=0 vs y=1
- Subscription rate by category: strongest differences by variables like month/contact/job (depending on plot)



Regression Modeling Setup

Regression target in notebook: **age**

Feature engineering used:

- age_squared (nonlinear relationship)
- pdays_was_missing, pdays_filled
- campaign_per_previous

Preprocessing pipeline:

- Numeric: impute median + scale
- Categorical: impute most frequent + one-hot encode

Models compared:

- Linear Regression, Ridge, Lasso

Regression Results + Visualization

- Key result: Ridge / Linear Regression / Lasso are extremely close
- Reported metrics from notebook (test):
  - $R^2 \approx 0.981$
  - RMSE ≈ 1.46

Classification Models

- Models evaluated:
  - Logistic Regression (baseline)
  - Decision Tree (baseline)
  - Linear SVM (baseline + tuned)

  - Why F1 matters: class imbalance (yes ≈ 11.7%)
  - Best tuned parameter:
    - SVM best C = 0.5 (from GridSearch)

Clustering: Selecting K + Visualizing Clusters

- Clustering method: MiniBatch K-Means on preprocessed features (no target column)
  - Elbow (inertia) + silhouette
  - Best k reported: 2
- K selection:
- 2D visualization: TruncatedSVD projection

Cluster Profiling (What the clusters mean)

- Cluster size summary
  - Cluster 0: 36,954
  - Cluster 1: 8,257
- Numeric averages table (age, balance, duration, campaign, previous, pdays)
- Top categorical breakdowns (example shown in notebook: job)

Association Rule Mining (Patterns among successful subscriptions)

- Association Rule Mining (Patterns among successful subscriptions)
- Method: Apriori + association rules
- Constraints used:
  - min_support = 0.05
  - max_len = 3
  - rules filtered to confidence ≥ **0.60**

Key Findings, Challenges, Next Steps

Key findings

- Imbalance drives metric choice (F1/ROC-AUC)
- SVM tuned performed best overall in F1 (per notebook table)
- Clustering suggests two major customer segments
- Association rules reveal recurring profile combinations in subscribers

Challenges

- Handling “unknown” + pdays sentinel
- Avoiding misleading accuracy
- Keeping plots readable (outliers)

Next steps:

- Try tree ensembles (Random Forest / Gradient Boosting)
- Threshold tuning for business objectives
- Deeper cluster interpretation with campaign strategy