BANK CUSTOMER SEGMENTATION VIA CLASSIFICATION ANALYSIS

MIS 451 - Machine Learning for Business

Lecturer: Mr. Dang Thai Doan
Ms. Huynh Gia Linh

PREPARED BY: GROUP 1
QUARTER 1: 2025-2026

OUR TEAM & ROLE

Team Member	IRN	Assigned Tasks
Le Nguyen Tam Nhu	2132300065	EDA (Visualizations)
Huynh Thuy Bao Tram	2132300228	Data Cleaning and Preprocessing
Huynh Trung Hau	2132309003	Model Development & Model Evaluation

REPORT OUTLINE

TOPIC HIGHLIGHTS

1. Business Context and Objectives
2. Data Collection and Characteristics
3. Exploratory Data Analysis (EDA)
4. Data Cleaning and Transformation
5. Model Development (Classification)
6. Model Evaluation and Comparison
7. Interpretation and Business Insights

BUSINESS CONTEXT AND OBJECTIVES

Business Problem

- Portuguese retail bank uses phone calls to promote term deposit products.

This approach is costly, time-consuming, and often inefficient.

- Traditional marketing treats all customers the same, leading to:
  - Unnecessary outreach
  - Low conversion rates

Proposed Solution

- Apply Machine Learning (Classification) to predict subscription likelihood before contacting customers.
- Use historical data:
  - Demographics
  - Financial indicators
  - Previous campaign interactions

Expected Impact

- Better targeting
- Reduced contact costs
- Higher subscription rates
- Improved marketing efficiency

BUSINESS CONTEXT AND OBJECTIVES

Primary Objective

- Build and compare multiple classification models to predict whether a client will subscribe.

Technical Workflow

- EDA to understand customer characteristics
- Data preprocessing:
  - One-hot encoding
  - Feature scaling
- Train ≥ 3 ML models + 1 Deep Learning model (per MIS 451 guidelines)

Model Evaluation

- Handle class imbalance using key metrics:
  - F1-score
  - Recall
  - Confusion matrix

Final Goal

- Identify most influential features
- Provide practical business insights to support more effective, targeted marketing strategies

DATASET DESCRIPTION

- Source: Bank Marketing Dataset (UCI Machine Learning Repository)
- Collected from telephone-based marketing campaigns (Portuguese bank, 2008–2010)
- ~ 45,000 customer interaction records
- 17 input variables + 1 target variable
- Includes customer demographics, financial status, and campaign history

ATTRIBUTE GROUPS & DESCRIPTION

CLIENT DEMOGRAPHIC VARIABLES

- age – customer's age
- job – occupation (admin, technician, blue-collar, student, etc.)
- marital – married / single / divorced
- education – primary / secondary / tertiary / unknown
- default – credit in default (yes/no)

FINANCIAL ATTRIBUTES

- balance – yearly account balance
- housing – housing loan (yes/no)
- loan – personal loan (yes/no).

CURRENT CAMPAIGN INTERACTION

- contact: Communication type (Cellular/Telephone).
- day / month: Last contact timing.
- duration: Last call duration (sec).

⚠ Note: Excluded from training to prevent data leakage.

PREVIOUS CAMPAIGN HISTORY

- campaign – contacts in current campaign
- pdays – days since last contact (−1 = no contact)
- previous – number of past contacts
- poutcome – outcome of previous campaign

5. TARGET VARIABLE (Y)

- y: Has the client subscribed to a term deposit? (Yes/No).

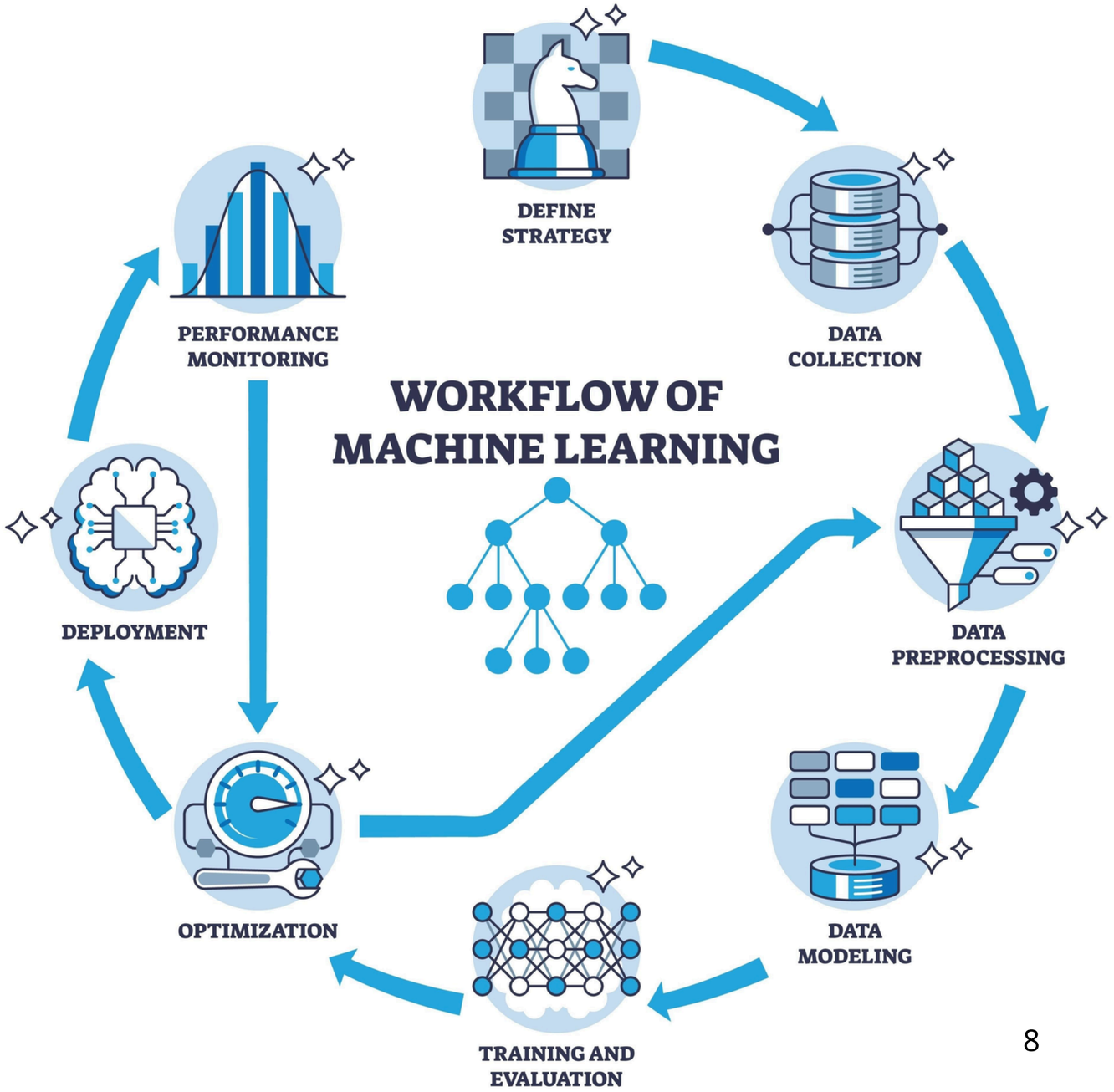
Role: Target variable for Supervised Learning.

Methodology (Workflow)

1. **EDA**: Analyze customer characteristics and behavioral patterns.
2. **Preprocessing**: Feature scaling and encoding for optimal model input.
3. **Modeling**: Train 3 Traditional Models + 1 Deep Learning Model

Evaluation & Output

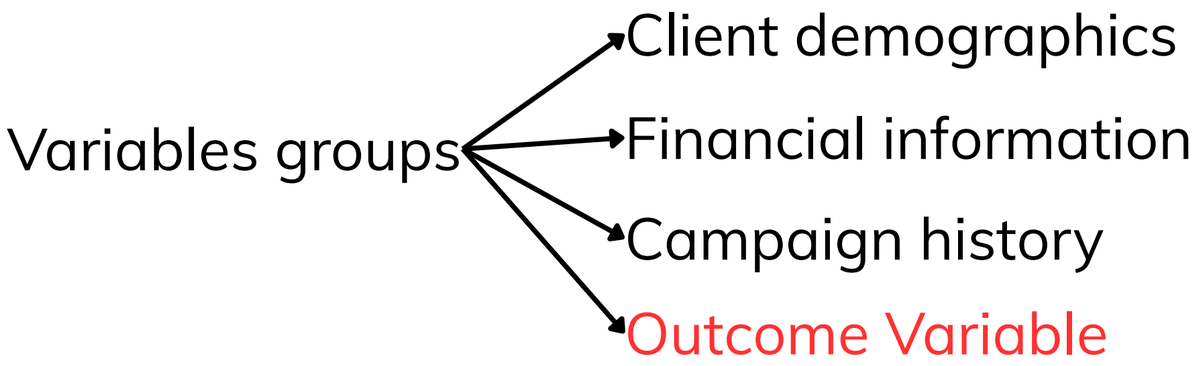
- **Metrics**: Focus on F1-Score and Recall to handle strong class imbalance.
- **Deliverable**: Interpret influential features to create actionable strategies for targeted marketing.



PRIMARY GOAL

DATASET OVERVIEW

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no



Objective: Understanding these patterns so models can predict who is likely to subscribe.

DATASET OVERVIEW

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45211 non-null  int64
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```

	age	balance	day	duration	campaign	pdays	previous
count	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000	45211.000000
mean	40.936210	1362.272058	15.806419	258.163080	2.763841	40.197828	0.580323
std	10.618762	3044.765829	8.322476	257.527812	3.098021	100.128746	2.303441
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	48.000000	1428.000000	21.000000	319.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000

- The data loads correctly, with **45,211 rows and 17 features** and no duplicated records.
- A mix of numerical and categorical variables exist.
- Numerical features like *balance*, *duration*, *campaign*, *pdays* and *previous* are highly right-skewed - reflecting genuine marketing behaviors, not data errors.

“UNKNOWN” CATEGORIES

age	0	**contact**	13020
job	288	**day**	0
marital	0	**month**	0
education	1857	**duration**	0
default	0	**campaign**	0
balance	0	**pdays**	0
housing	0	**previous**	0
loan	0	**poutcome**	36959
		y	0

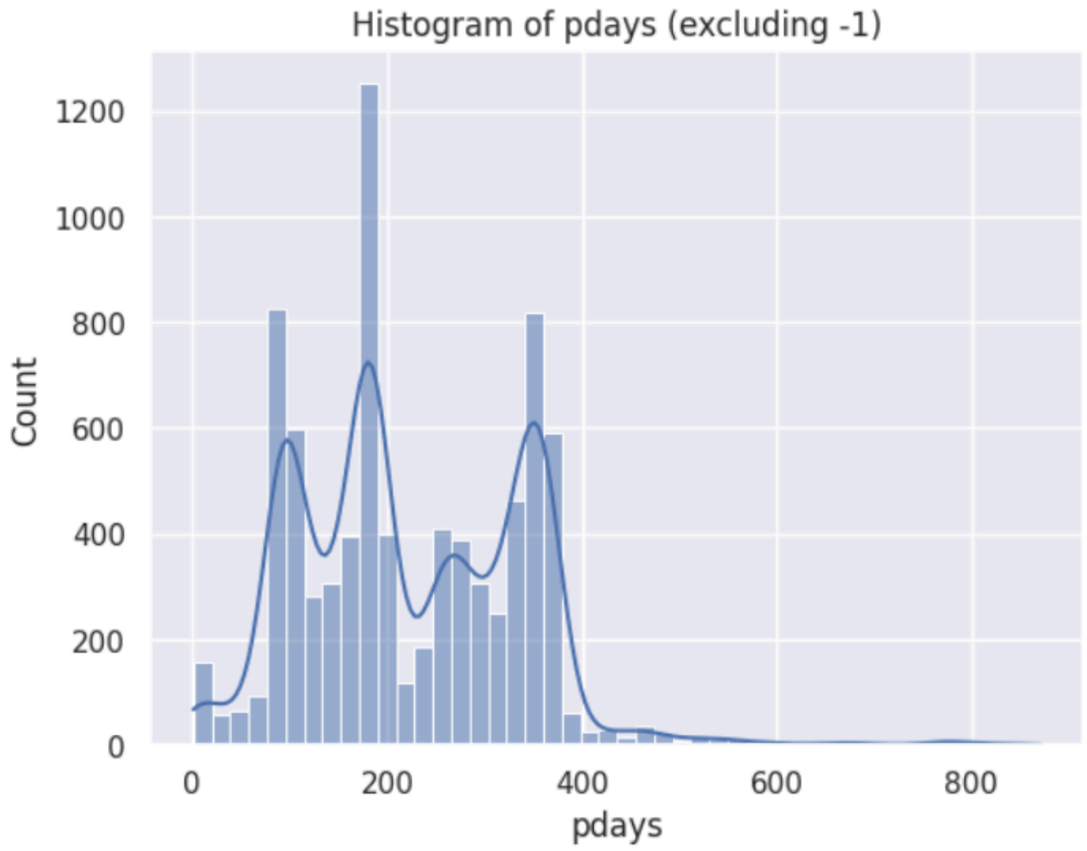
Instead of NaN, missing categorical data is labeled as “unknown”

These cases reflect valid behavioral groups, for example:

- *job* → not recorded
- *education* → not provided
- *contact* → not documented
- *poutcome* → not tracked

=> Retained rather than imputed

SPECIAL HANDLING OF PDAYS

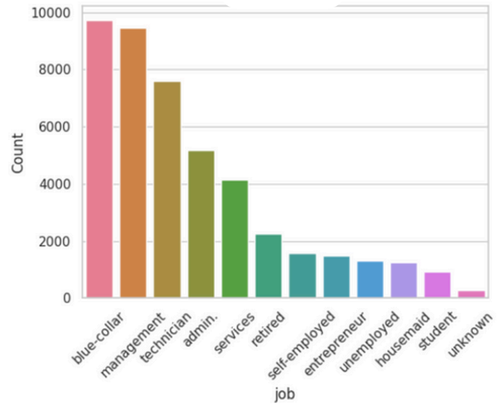


y	no	yes
pdays_flag		
contacted_before	76.928667	23.071333
never_contacted	90.842669	9.157331

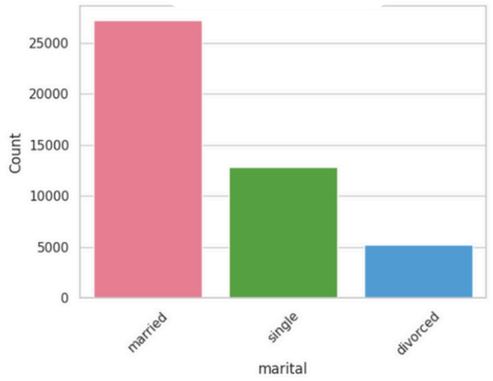
- *pdays* = −1 indicates customers never contacted before.
- A flag feature was created to separate “never contacted” and “contacted before.”
- Clients previously contacted subscribe at more than double the rate of new contacts.
- *pdays* values above −1 span widely and are treated as valid history, not errors.

CATEGORICAL DISTRIBUTIONS

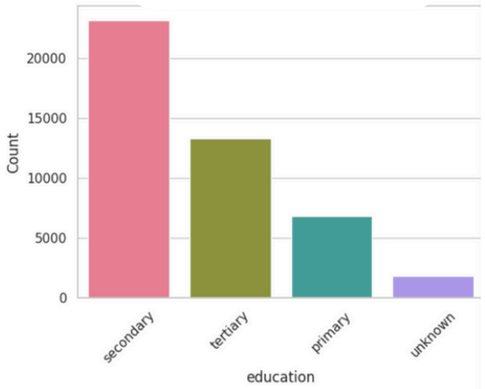
Job



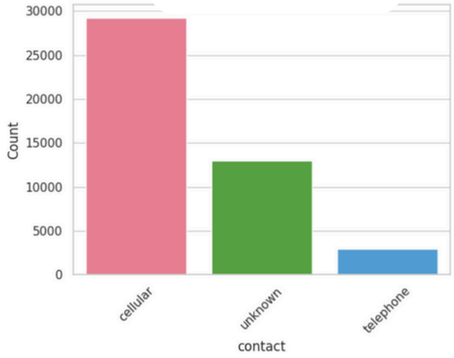
Marital



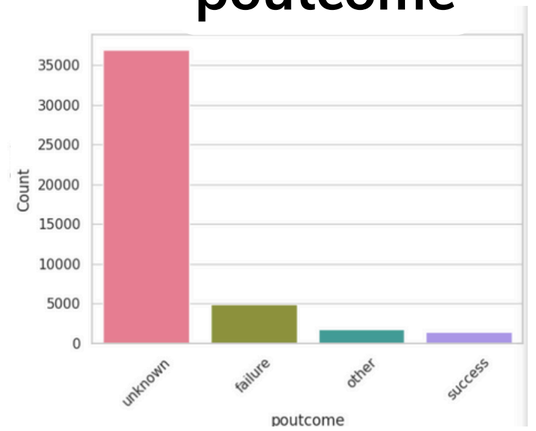
Education



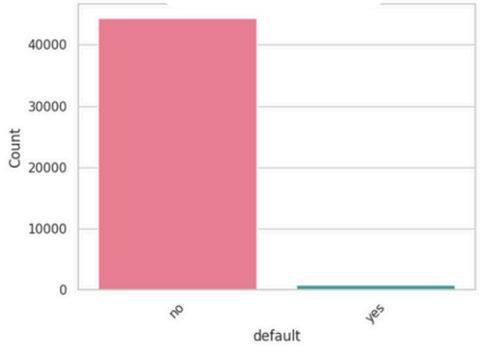
Contact



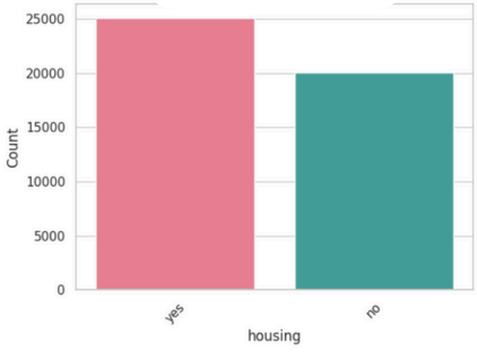
poutcome



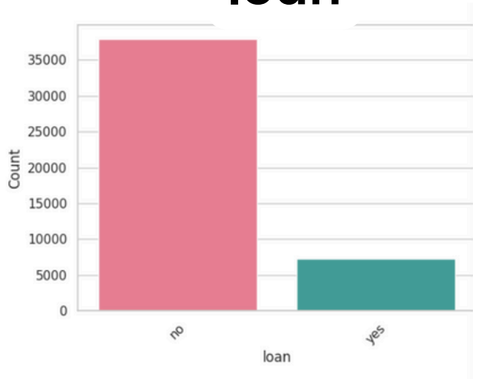
default



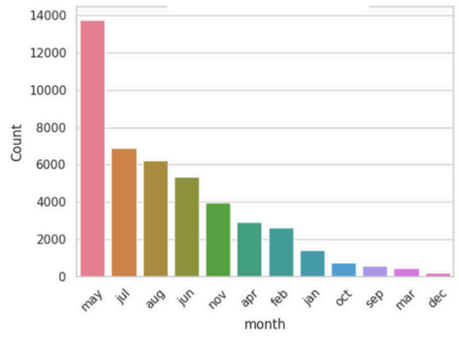
housing



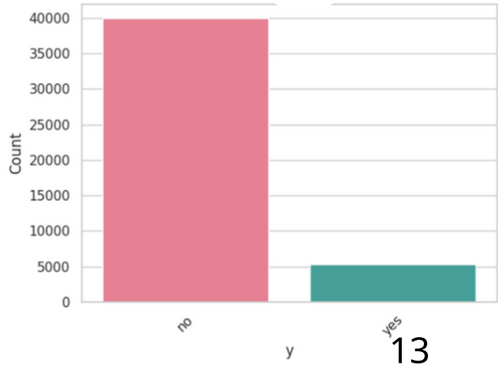
loan



month

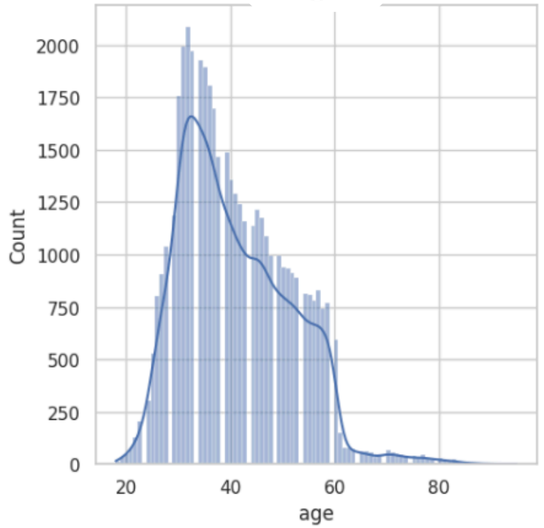


y

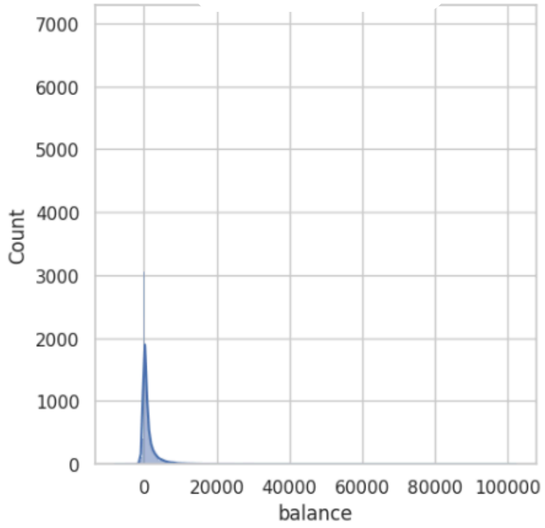


NUMERICAL DISTRIBUTIONS

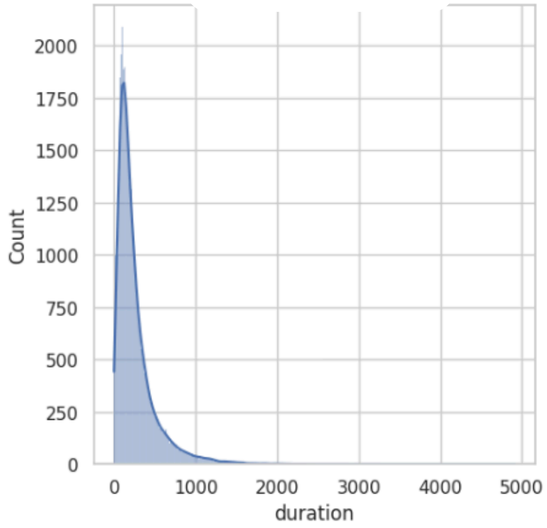
age



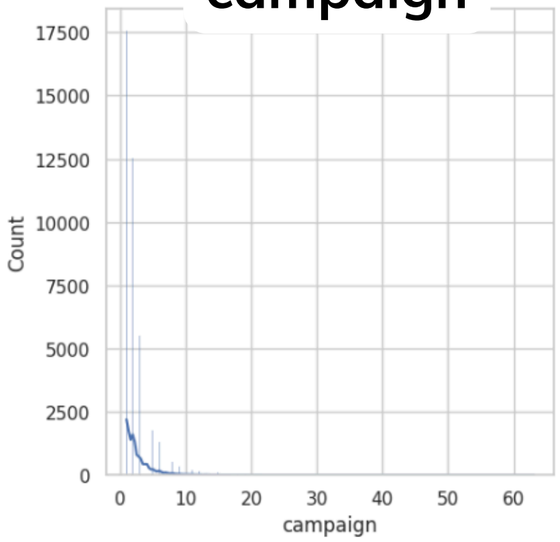
balance



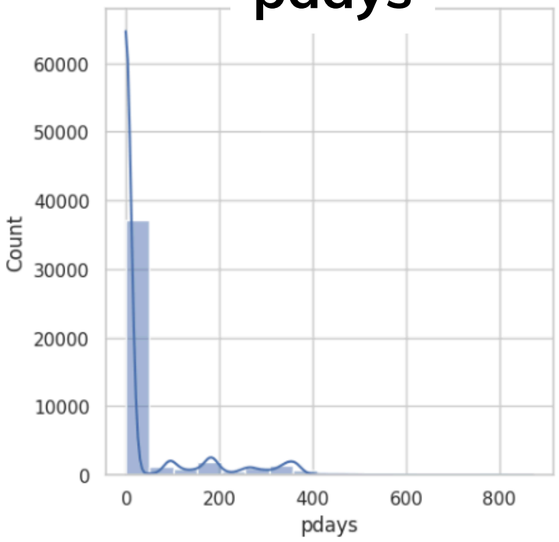
duration



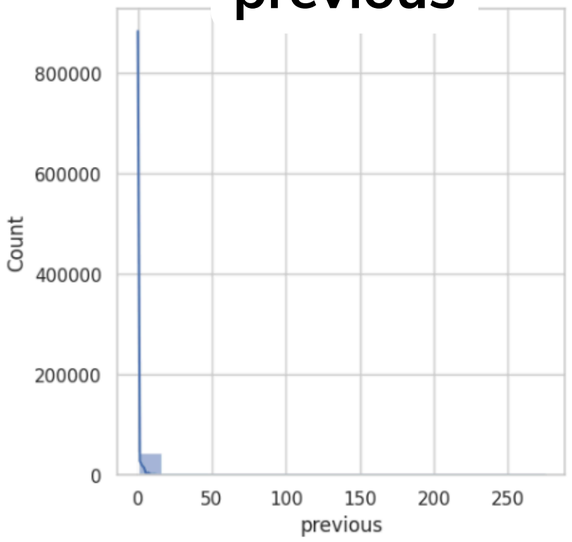
campaign



pdays

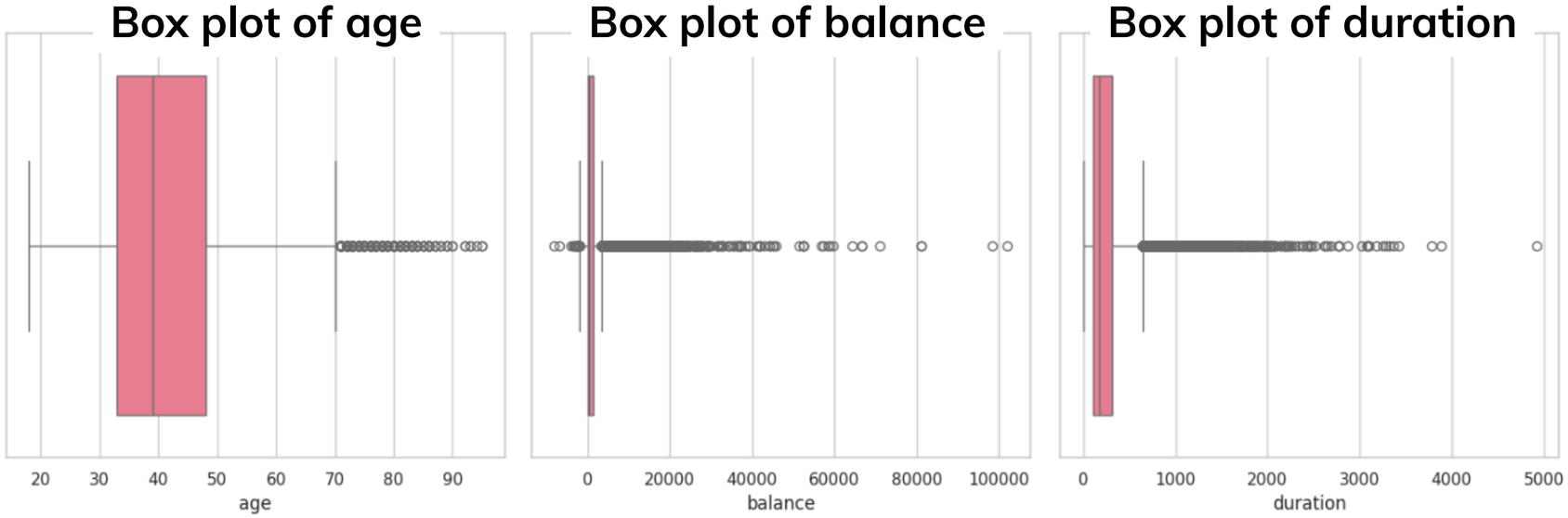


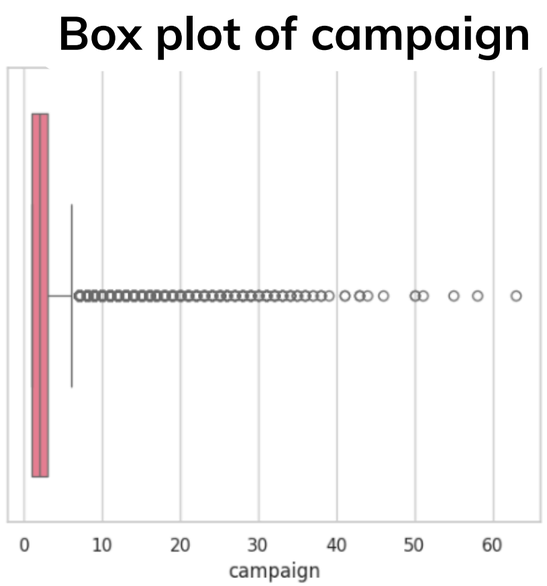
previous

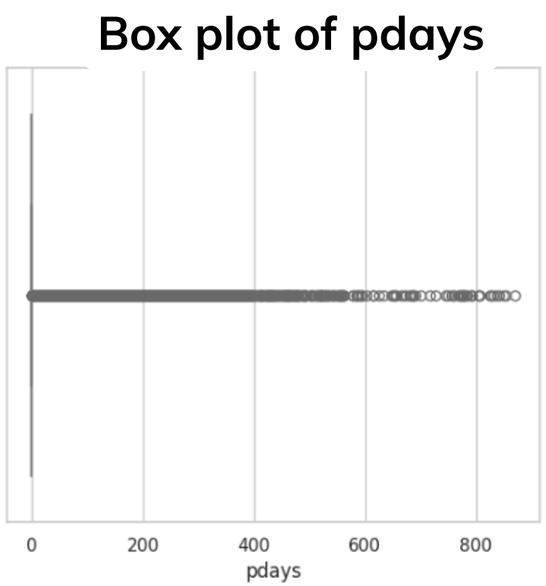


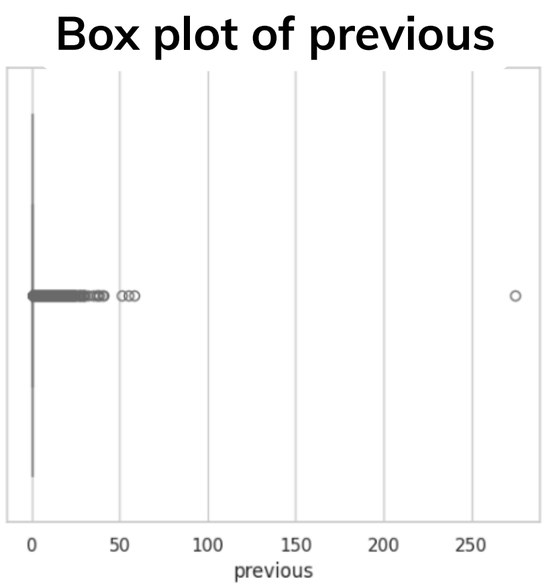
Numerical features are right-skewed - many small values, few extremely large values - but these reflect real customer behaviors.

OUTLIERS







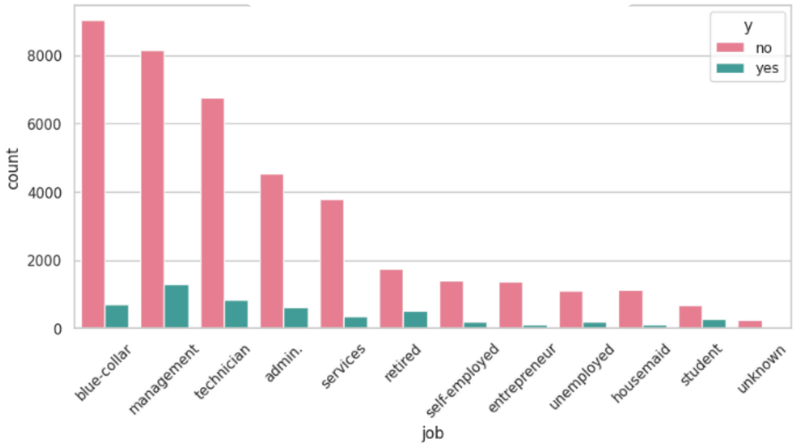


Thousands of statistical outliers exist, especially in balance, duration and pdays.

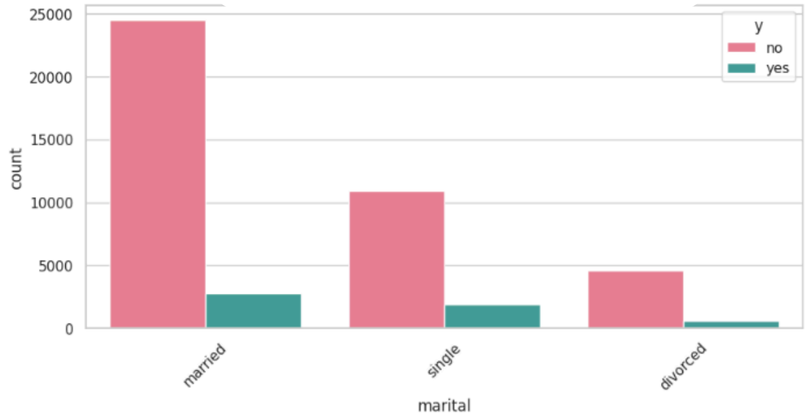
Because these extremes are meaningful in marketing scenarios, they are kept rather than trimmed.

CATEGORICAL VARIABLES VS TARGET

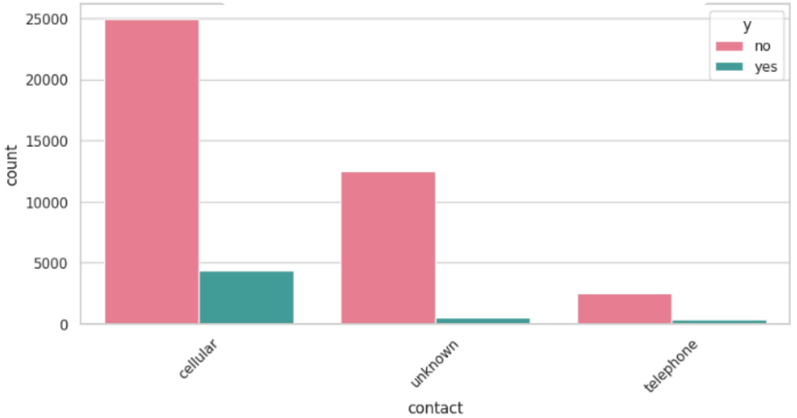
job vs Target (y)



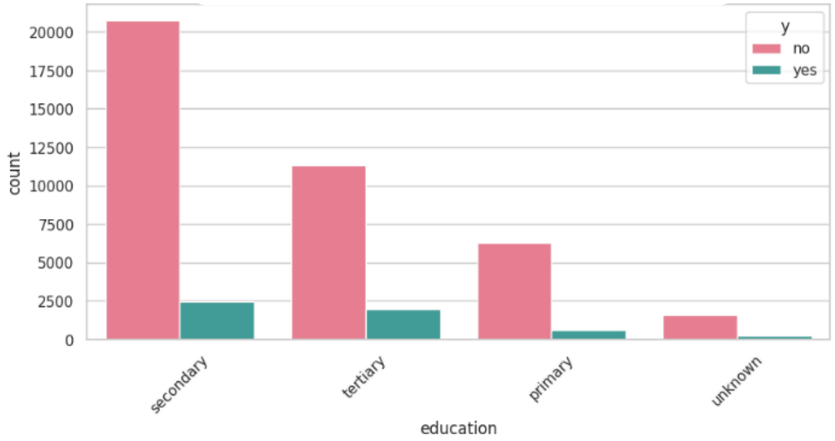
marital vs Target (y)



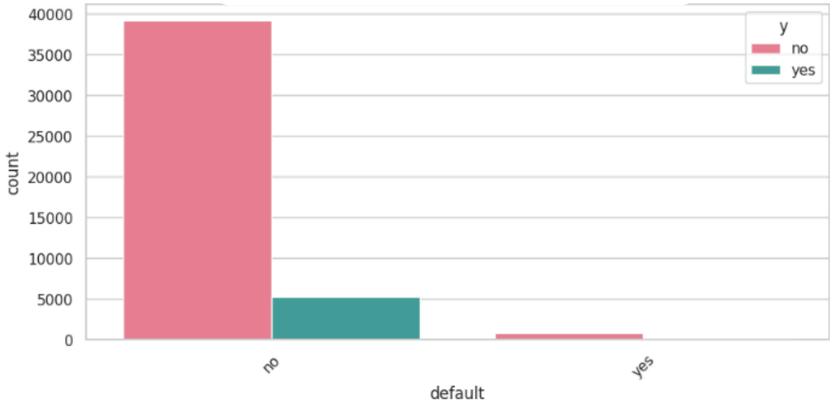
contact vs Target (y)



education vs Target (y)



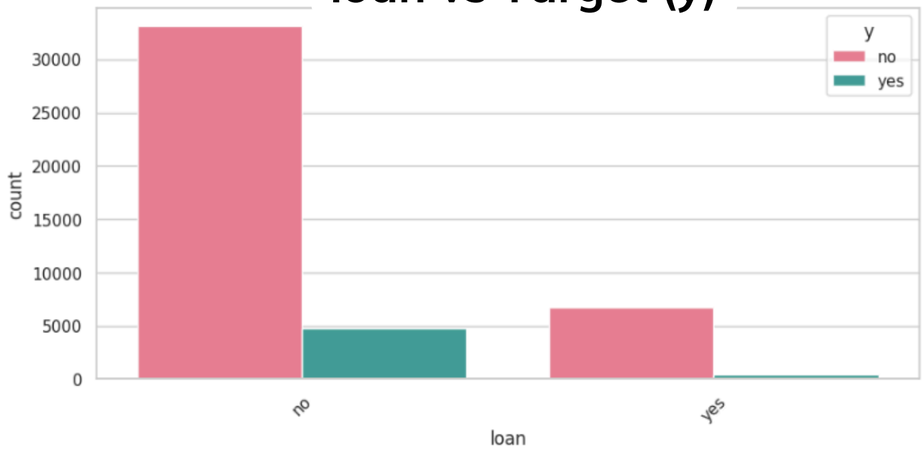
default vs Target (y)



CATEGORICAL VARIABLES VS TARGET

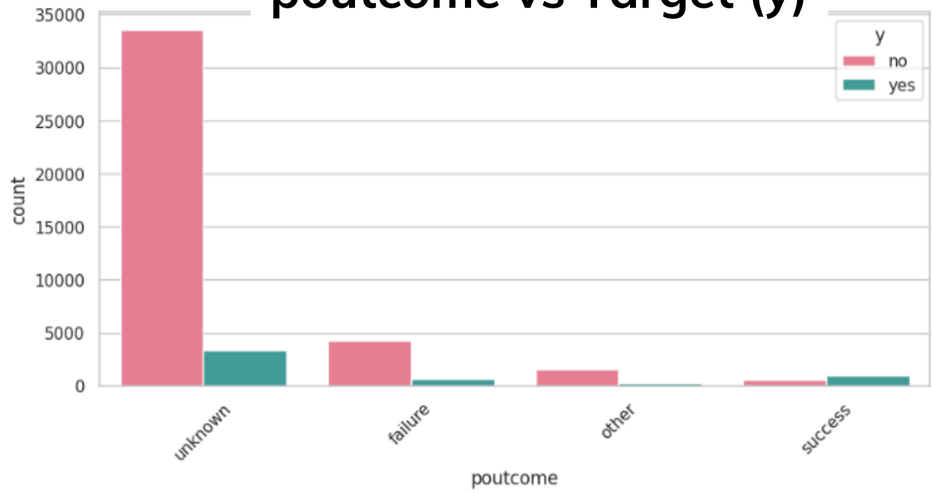
housing vs Target (y)

loan vs Target (y)



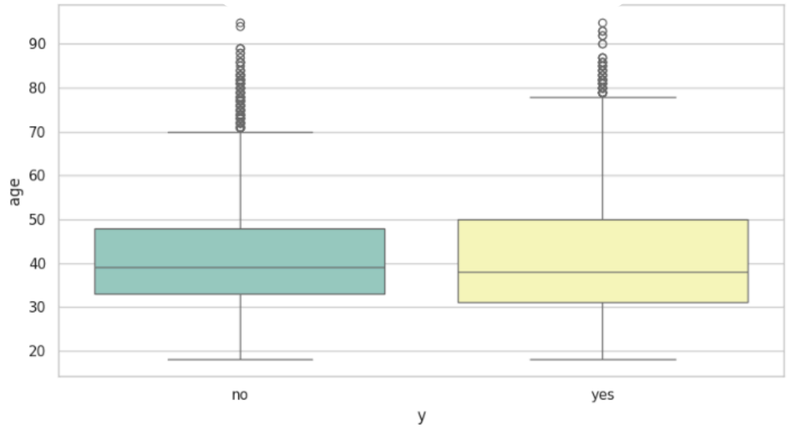
month vs Target (y)

poutcome vs Target (y)

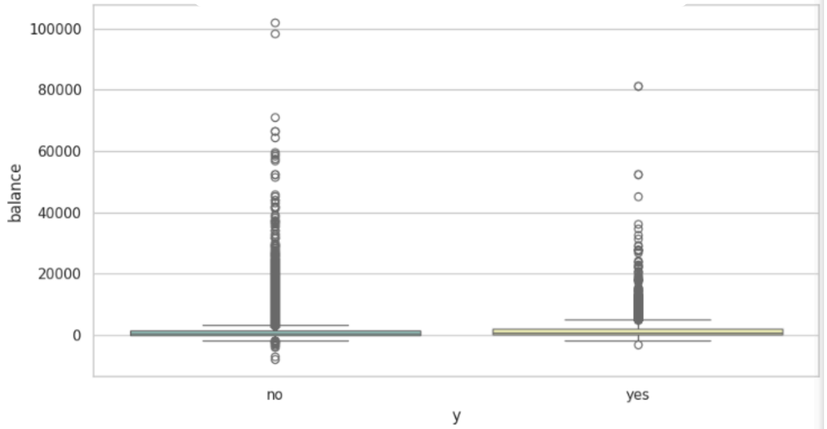


NUMERICAL VARIABLES VS TARGET

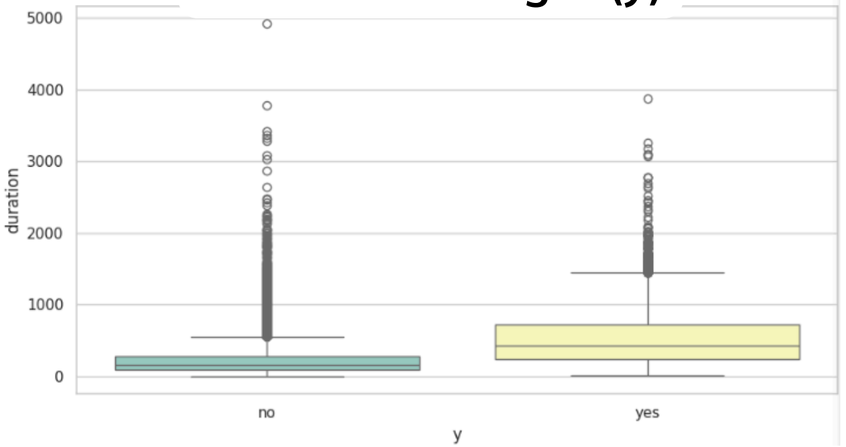
age vs Target (y)



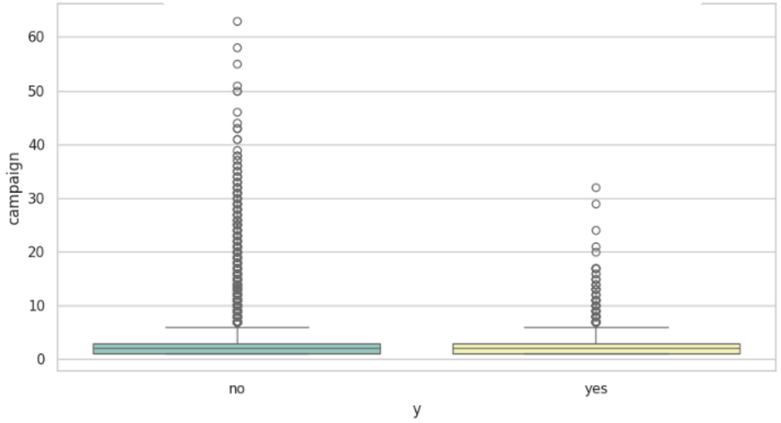
balacne vs Target (y)



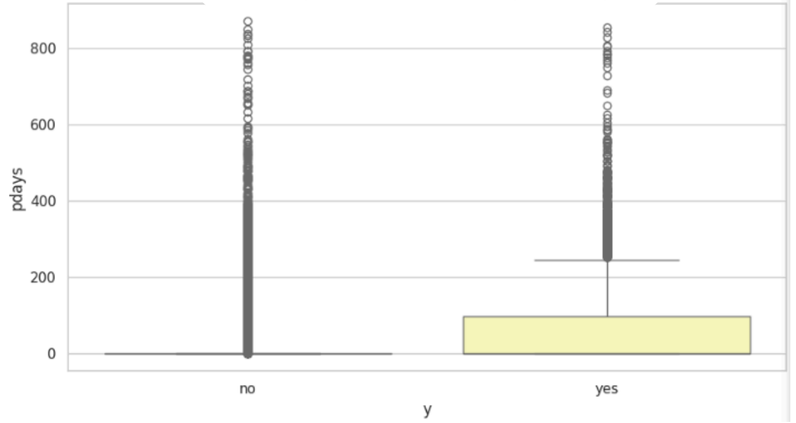
durationvs Target (y)



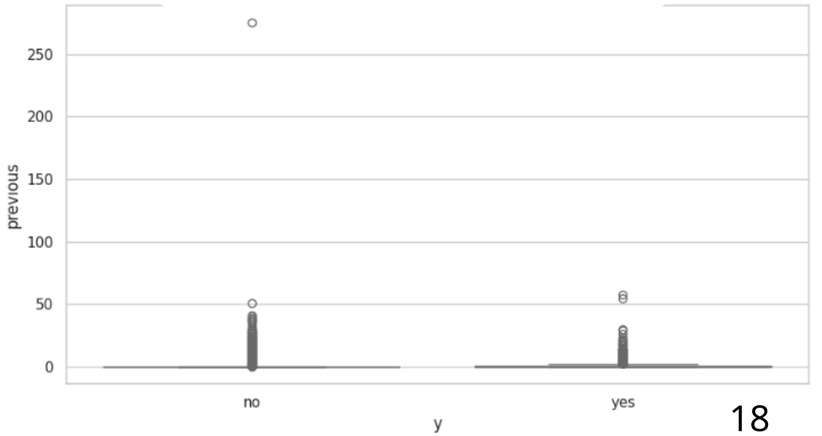
campaign vs Target (y)



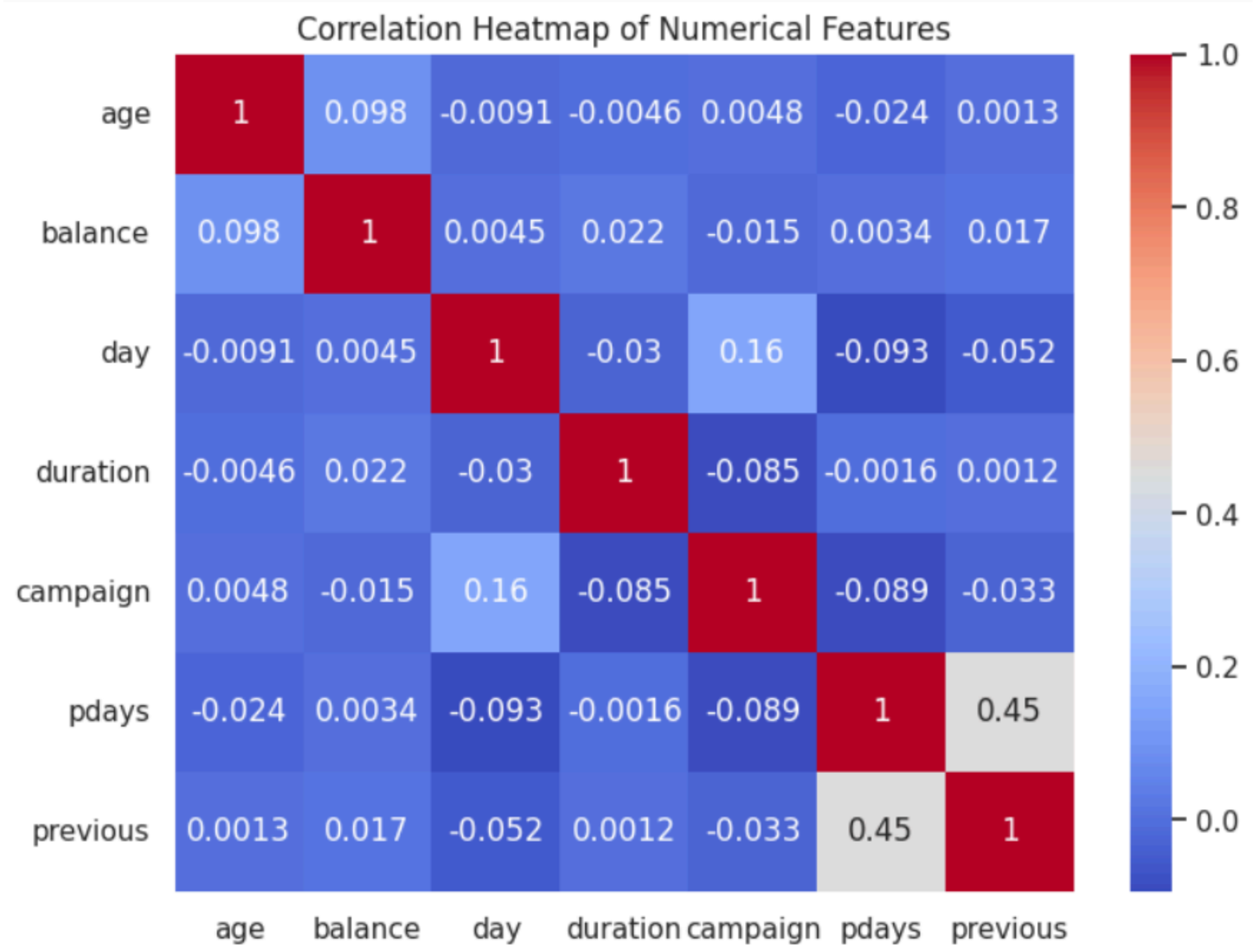
pdays vs Target (y)



previous vs Target (y)



CORRELATION HEATMAP



- Numerical features show weak pairwise correlations.
- Only pdays and previous have a moderate relationship (0.45), as both describe past campaign activity.

DATA CLEANING & TRANSFORMATION

- Goal: Prepare clean, consistent, leakage-free data for classification models
- Steps included:
  - Handling missing values
  - Encoding categorical variables
  - Scaling numerical features
  - Dimensionality reduction
- Implemented using Pipeline + ColumnTransformer to maintain automation, consistency, and prevent data leakage.

SEPARATING TARGET AND FEATURES

CLASSIFICATION:

SUPERVISED LEARNING

SUPERVISESPLIT DATASET INTO:

- Input Features (15 variables) → age, balance, job, etc.
- Target Variable (y) → subscription decision (encoded 1 = yes, 0 = no)

WHY DROP DURATION?

- Duration is only known after the call ends, not before contacting the client.
- Strongly correlates with outcome → creates data leakage.
- Removing it ensures the model predicts realistically before a call is made.

HANDLING MISSING VALUES (IMPUTATION)

STRING VALUE "UNKNOWN"

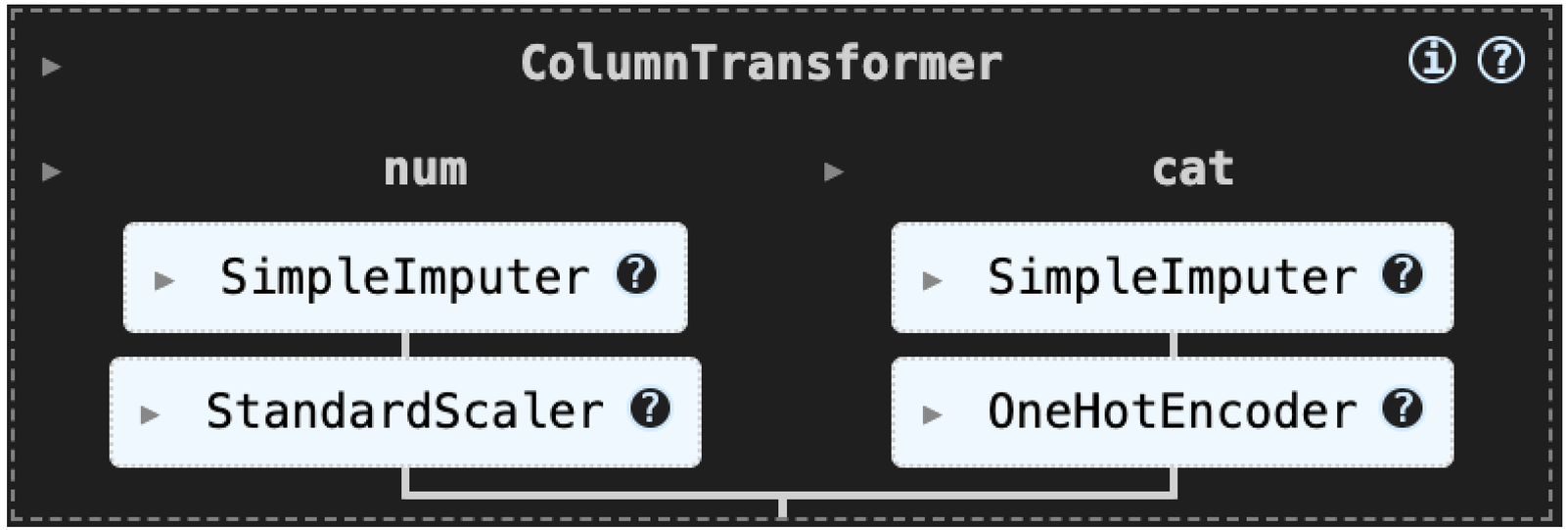
WHY STILL USE AN IMPUTER?

- Ensures robustness for future deployments (API data may contain NaNs).
- Prevents model crashes → acts as a safety net.
- Does not alter current data since no NaNs are present.

IMPUTATION STRATEGY

- Numerical → Median (stable for skewed variables like balance)
- Categorical → Mode (preserves structure)
- 'unknown' is not treated as missing → kept as a meaningful category.

ENCODING & SCALING



Numerical Features

- Standardized using StandardScaler

→ Mean = 0, Std = 1
→ Prevents large-scale variables (e.g., balance) from dominating models
→ Crucial for SVM, KNN, and gradient-based algorithms

Categorical Features

- Encoded using One-Hot Encoding

→ Prevents artificial ordering
→ Treats “unknown” as a valid, informative category

SPECIAL HANDLING: PDAYS

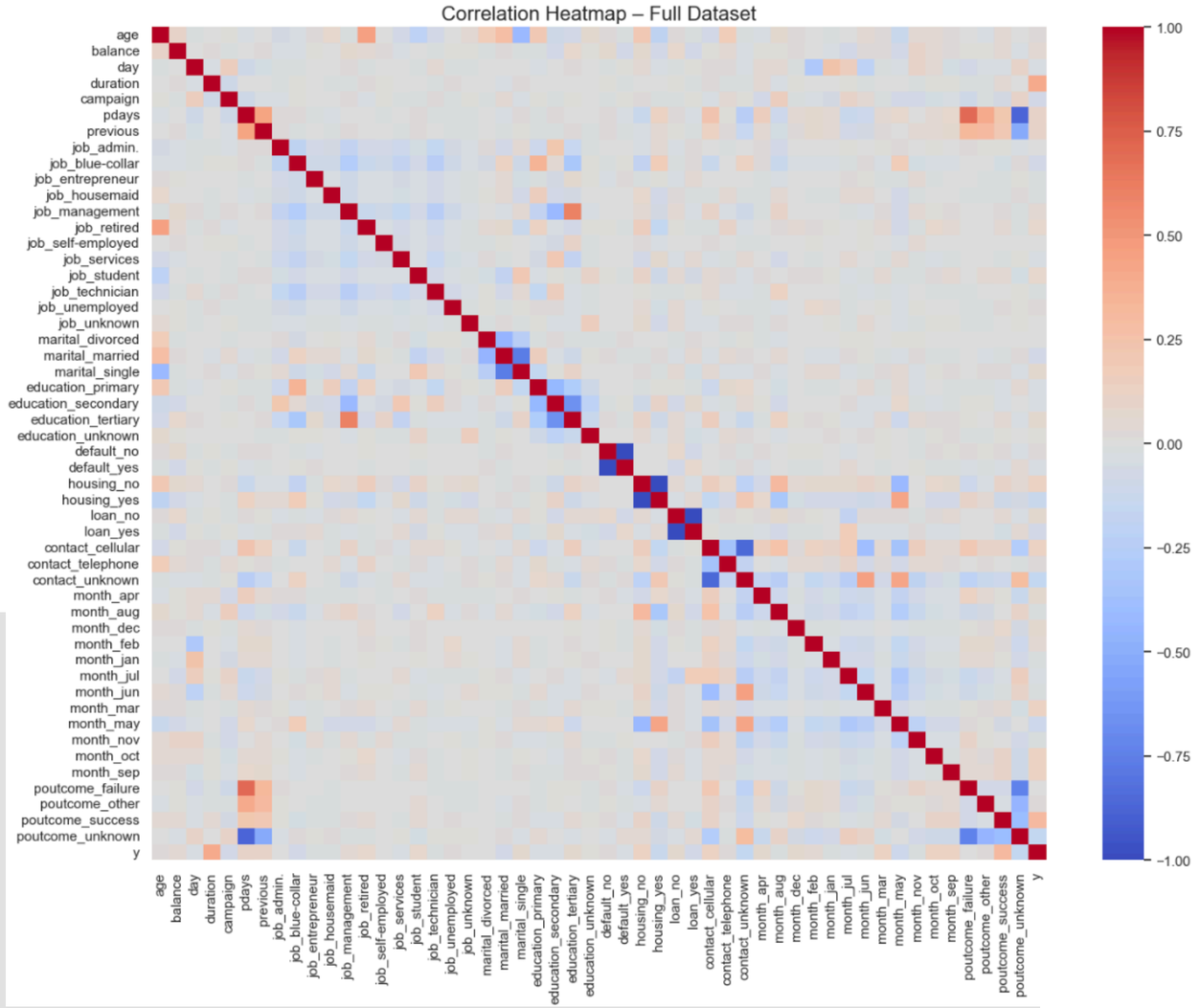
- Value –1 = customer never contacted before, NOT a real time measurement
- Avoid treating –1 as numeric (misleads distance-based models)
- Engineered a derived feature to distinguish:
- Newly contacted vs Previously contacted customers

→ Improves behavioral accuracy in model learning.

DIMENSIONALITY REDUCTION (PCA)

WHY PCA?

- After One-Hot Encoding → 15 features → 51 dimensions
- High dimensionality → risk of overfitting & sparse matrix issues



PCA CONFIGURATION

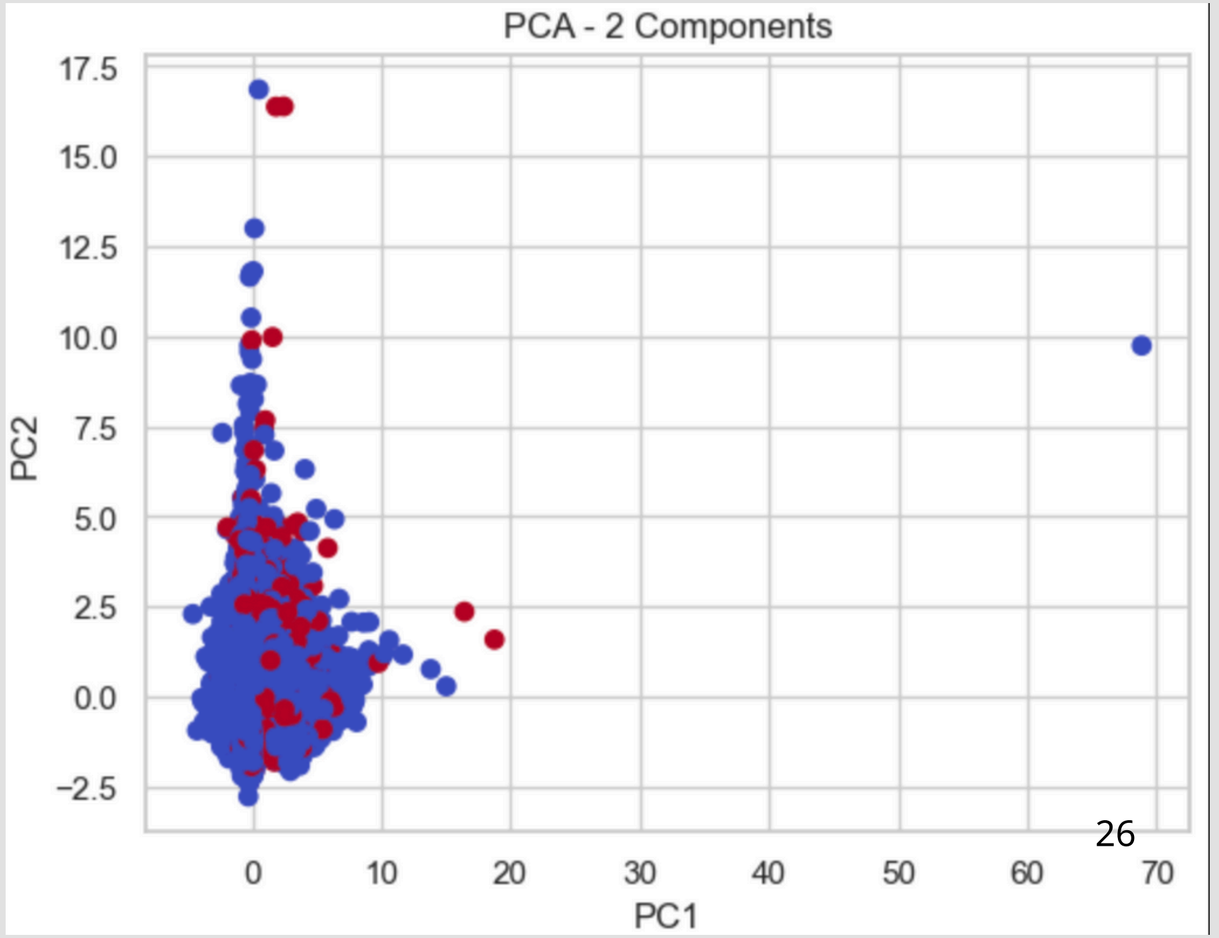
- A Retain 95% variance
- Reduced feature space from 51 → 24 principal components

Benefits:

- Removes noise & multicollinearity
- Reduces computational cost
- Improves model generalization

DIMENSIONALITY REDUCTION (PCA)

- Classes overlap → difficult to separate linearly.
- PC1 & PC2 capture limited variance → data is highly non-linear.
- Outliers present.
- ANN/SVM expected to perform better than Logistic Regression.



Dataset Partitioning

- 80% Training
- 20% Testing
- Applied stratify = y to maintain class imbalance ratio

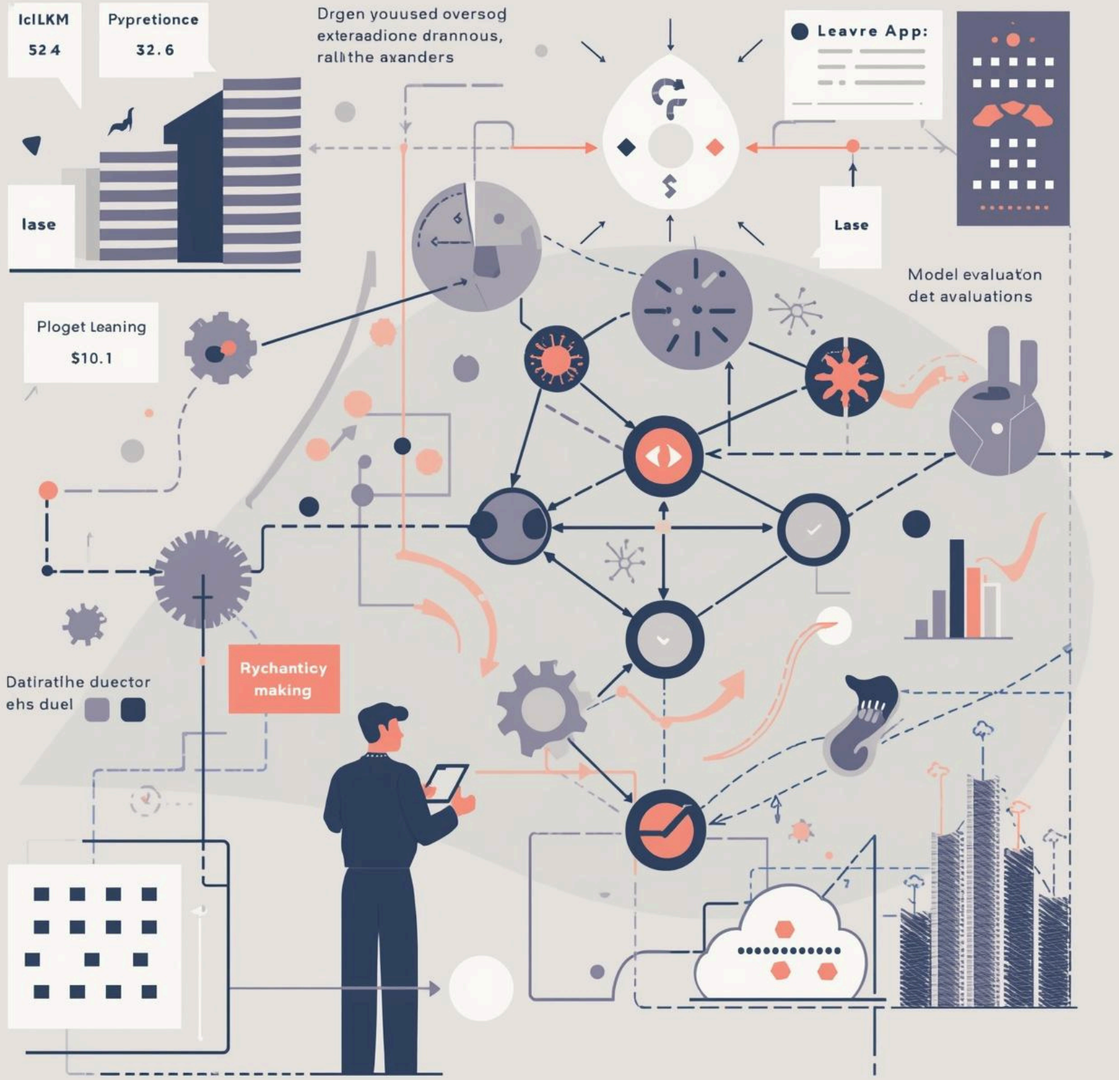
Purpose

- Training set → model learning
- Test set → evaluate performance on unseen data

TRAIN–TEST SPLIT

Evaluation Metrics

- Precision
- Recall
- F1-score
- Confusion Matrix
- → Critical for imbalanced classification tasks



Model Development Overview

Classification Project Evaluation

Introduction

Model Development Overview

- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Logistic Regression
- Artificial Neutral Network (ANN)

Model Selection Rationale

Overview of candidate models and roles

Model Selection Rationale

Overview of Individual Models

Logistic Regression

Logistic Regression serves as the **baseline model** providing interpretability through coefficients, establishing a foundational benchmark for assessing feature impact on the likelihood of subscription.

K-Nearest Neighbors

K-Nearest Neighbors (KNN) captures **local patterns** effectively by identifying feature-space proximity, making it a powerful complement to linear models after scaling for better performance.

Support Vector Machine

The Support Vector Machine excels at **complex boundary detection**, utilizing kernels to find optimal hyperplanes, making it robust even in high-dimensional datasets for effective classification.

Artificial Neutral Network

An Artificial Neural Network (ANN) was chosen because it can model complex, non-linear relationships that traditional machine learning models may not capture

Model Comparison

```
================================================================================
MODEL COMPARISON (Sorted by F1-score of 'yes' class)
================================================================================
                Model  Train Accuracy  Test Accuracy  F1-macro  F1-yes  Recall-yes Overfitting Risk
ANN (Neural Network)          0.9138         0.8890    0.6535  0.3678      0.2760               Low
                 KNN          0.9056         0.8874    0.6150  0.2911      0.1975               Low
                 SVM          0.8959         0.8933    0.6156  0.2889      0.1853               Low
 Logistic Regression          0.8924         0.8933    0.6119  0.2815      0.1786               Low

================================================================================
 BEST MODEL: ANN (Neural Network)
 Reason: Highest F1-score for customers who subscribe ('yes') → 0.3678
         Catches 0.2760 of actual subscribers (Recall-yes)
         Overfitting Risk: Low
================================================================================

You can now use 'ANN (Neural Network)' as your final deployed model!
```

PERFORMANCE ON TEST SET

```
============================================================
ANN PERFORMANCE ON TEST SET
============================================================
              precision    recall  f1-score   support

           0     0.9100    0.9702    0.9391      7985
           1     0.5509    0.2760    0.3678      1058

    accuracy                         0.8890      9043
   macro avg     0.7305    0.6231    0.6535      9043
weighted avg     0.8680    0.8890    0.8723      9043

Confusion Matrix:
[[7747  238]
 [ 766  292]]
```

Final Model Selection

Artificial Neutral Network

- Highest F1-score on the positive (“yes”) class: 0.558 - the single most important metric here, because every correctly identified subscriber directly translates to revenue and better marketing ROI.

- Best precision-recall balance for subscribers: F1-yes = 0.558 is 22-28% higher than the next-best models (SVM and Logistic Regression ~0.45- 0.46). The ANN finds many more real subscribers while keeping precision solid at 56.7%.

- Highest Recall-yes: 54.9% - captures over half of all actual subscribers, roughly 60% more than competing models (which only reach ~34-35%). Missing potential subscribers is the most expensive mistake in this campaign; ANN minimizes that loss.

- Strong overall performance: 89.8% test accuracy (virtually tied with the top model, SVM at 90.5%) and the highest F1-macro (0.750), proving it performs well on both classes, not just the majority “no” class.

- Low overfitting risk: only ~4.7 percentage point drop from training (93.5%) to test (89.8%) accuracy - excellent generalization, comparable to SVM and Logistic Regression.

Recommendations

Artificial Neutral Network

- Deploy the Artificial Neural Network (ANN) as the primary scoring model for term-deposit campaigns, using the default classification threshold of 0.50. This configuration yields the highest F1-score (0.558) and Recall (54.9%) on the positive class among all tested algorithms, delivering an estimated 5-6× uplift in subscriptions compared to random or legacy targeting while contacting approximately 11% of the customer records.

- Rank and prioritize the contact list according to the ANN's predicted probability. Execute outreach in descending order of score (≥ 0.70 → 0.50–0.69 → 0.35–0.49 if additional budget is available). This straightforward prioritization maximizes conversion rates and campaign ROI without requiring additional modelling effort.

- Conduct a controlled A/B test in the next campaign cycle. Allocate 20–30% of the campaign universe to the ANN-ranked list and the remainder to the current targeting approach. The four-week pilot will provide conclusive, real-world evidence of the uplift and facilitate full organizational adoption.