Revolutionizing Bank Marketing:

Machine Learning Approaches for Predicting Term Deposit Subscriptions

Student: Tran Phuong Anh, 11219258
Instructor: PhD. Le Thi Khuyen

May 19, 2025

Table of contents

Revolutionizing Bank Marketing:
Machine Learning Approaches for Predicting Term Deposit Subscriptions

Introduction →

Background **& Motivations**

Sector Overview

Banking is shifting toward digital. Marketing optimization and customer retention are now strategic priorities.

Core Business Challenge

Banks face challenges in identifying which customers are likely to subscribe to term deposits. Manual targeting leads to low conversion rates and inefficient use of resources.

Personal Motivation

Aspire to work in data analytics for banking. This project builds practical skills for real-world applications.

Research Questions

- How do data quality and preprocessing influence model performance in predicting term deposit subscriptions?

- Which machine learning models offer the best predictive results in bank marketing campaigns?

- What are the most impactful features for accurately predicting customer subscription?

- Which techniques and training strategies can enhance model performance and generalization?

- How can SHAP or LIME help explain and interpret model predictions?

Revolutionizing Bank Marketing:
Machine Learning Approaches for Predicting Term Deposit Subscriptions

Literature Review →

Research gaps

Limited model interpretability

Most studies focus on accuracy but overlook how and why predictions are made.

Lack of supplementary error analysis

Misclassified/Uncertain cases are rarely analyzed after prediction, limiting understanding of model weaknesses.

Insufficient evaluation metrics

Important metrics like sensitivity and balanced accuracy are often missing.

Note: *Due to time limitation, a more comprehensive summary of previous works is coverd in the thesis report.*

Material →

Material: Dataset Overview

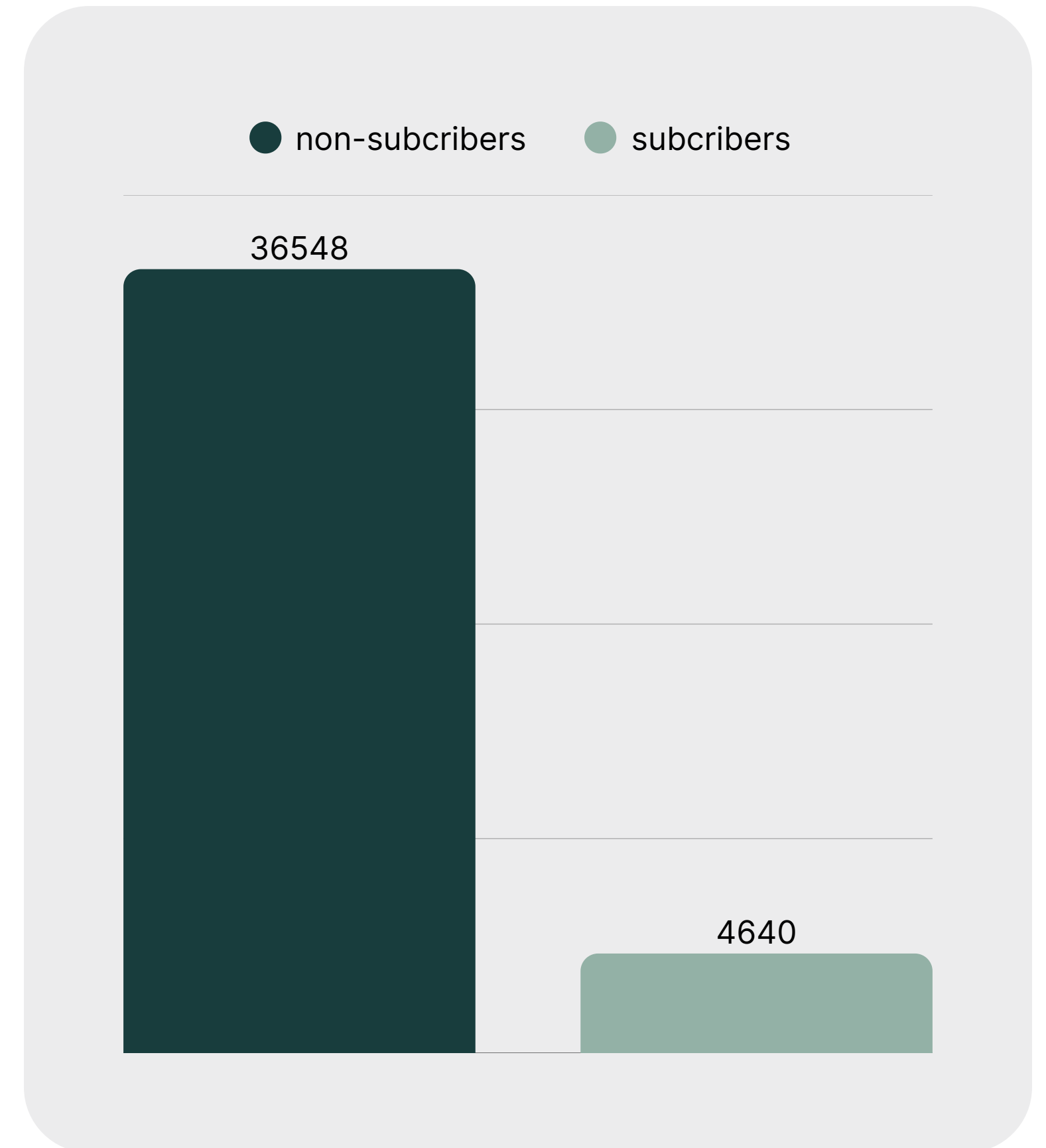**Source:** UCI Machine Learning Repository (Moro et al., 2014).
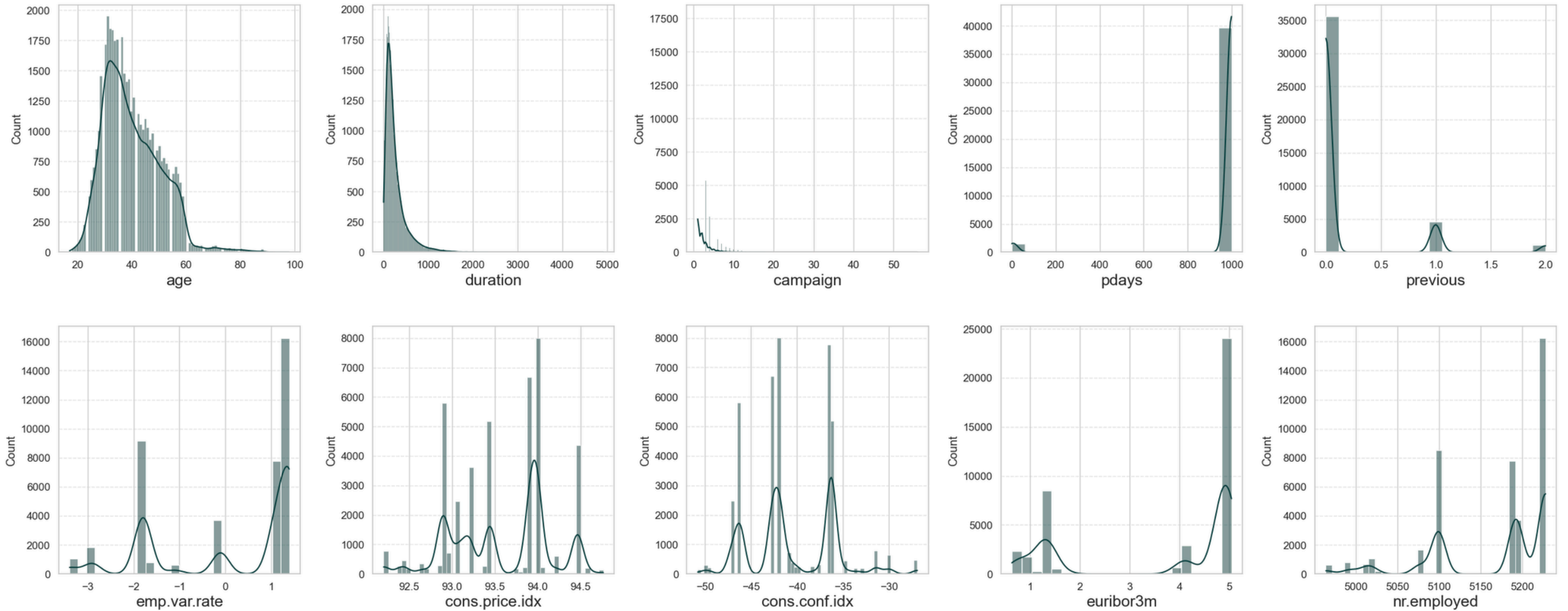**Size:** 41,188 observations × 21 variables.

Features:
- 20 input features (categorical + numerical).
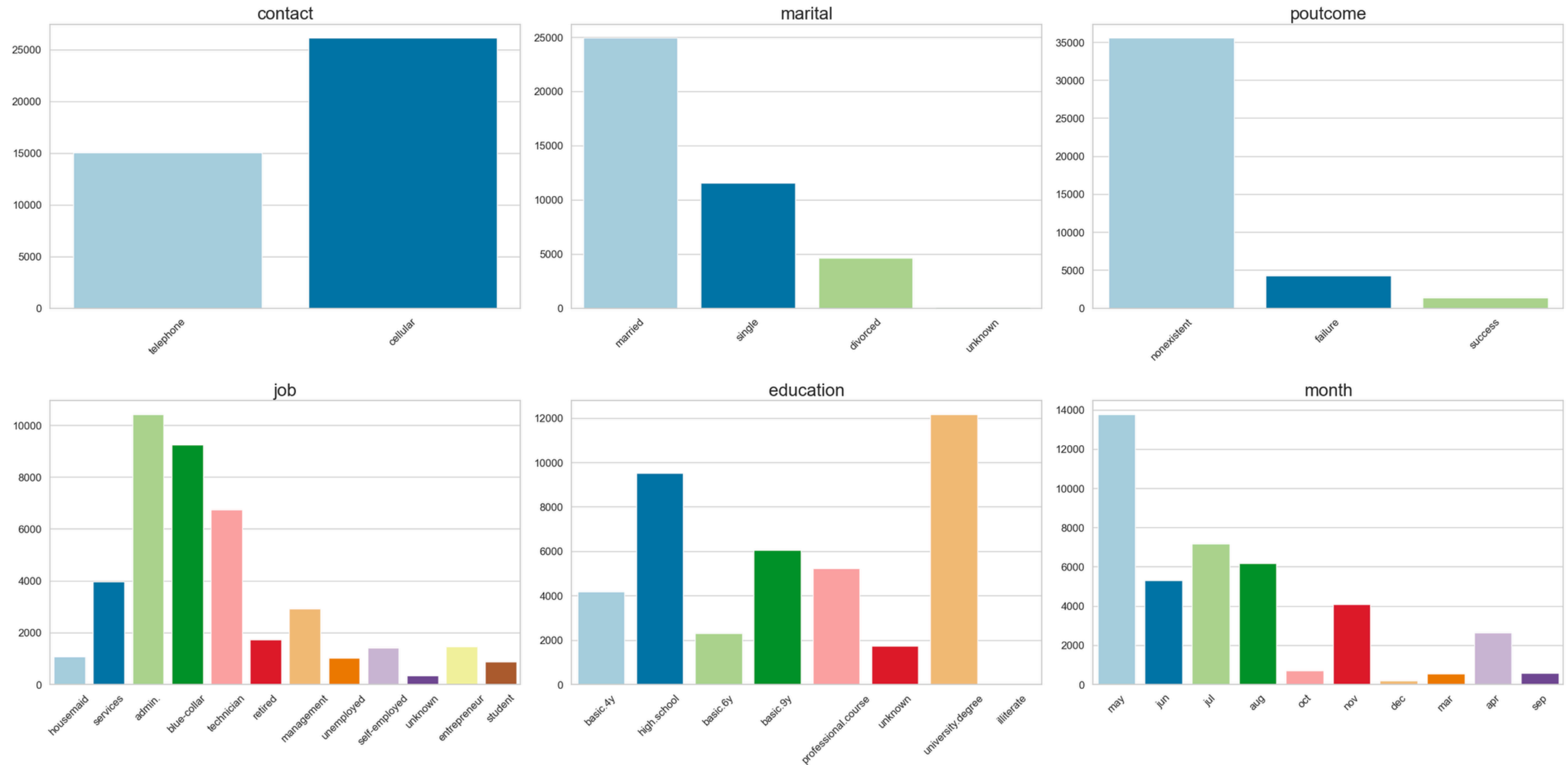- 1 binary target: Subscribed (yes/no).

Target imbalance:
- Subcribers/Yes: **11.27%**.
- Non-Subcribers/No: **88.73%**.

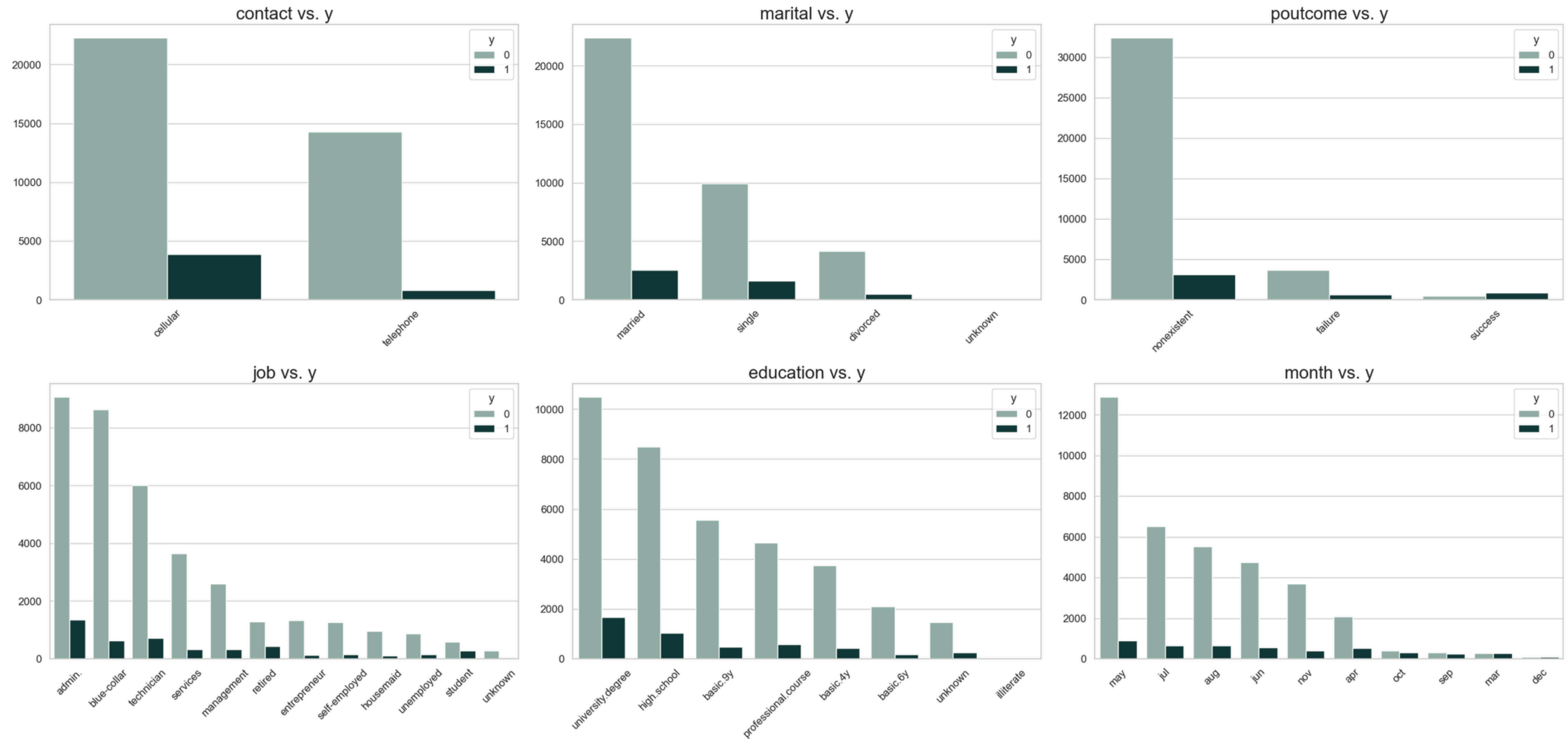● non-subcribers   ● subcribers

36548

4640

Material: Exploratory data analysis

*Univariate Analysis - Distribution of continuous variables.*

Univariate Analysis: Distribution of some categorical variables.

**Bivariate Analysis:** Distribution of some categorical by target.

Correlation matrix

The variable **duration** shows the strongest positive correlation with subscription behavior, while lower values of economic indicators such as **euribor3m, emp.var.rate** and **nr.employed** are associated with higher term deposit subscription rates.

Correlation Matrix of Numerical Variables
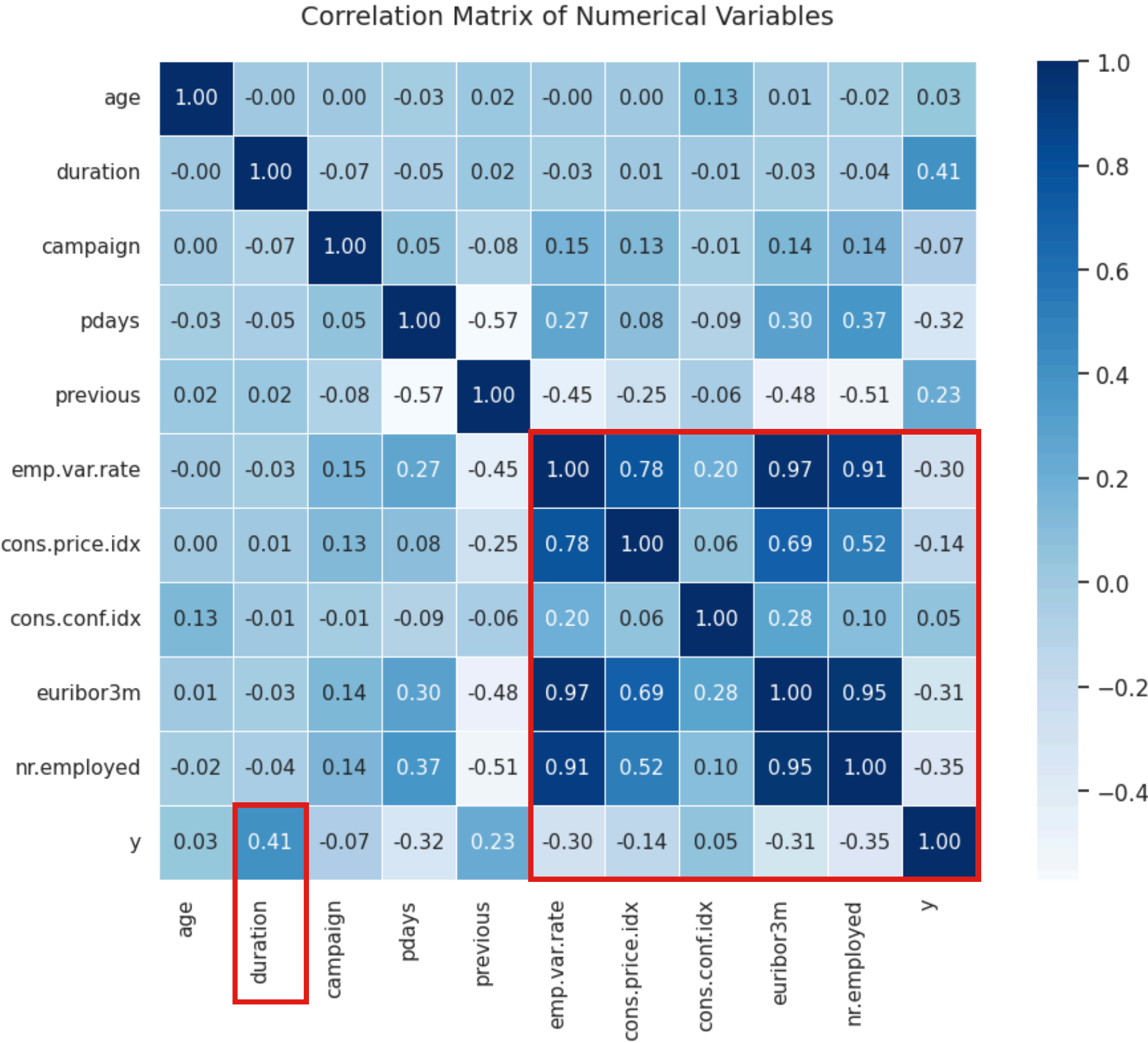
t-SNE visualization

Subscribed customers (darker points) distributed sparsely across multiple regions, indicating class imbalance and complex decision boundaries.

t-SNE Visualization of Bank Marketing Data

Sample Size: 3000 | Features: 61 | Run Time: 8.71s
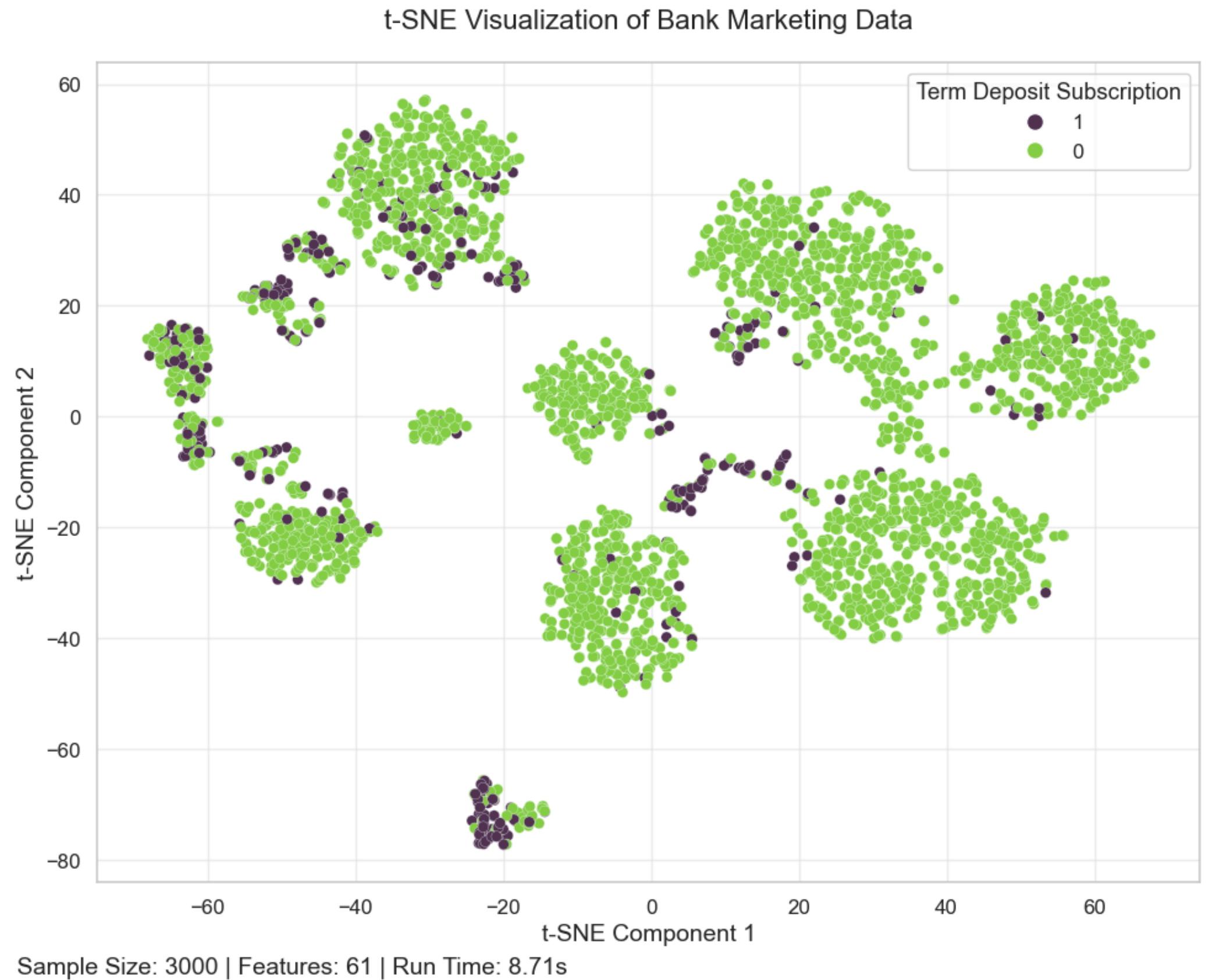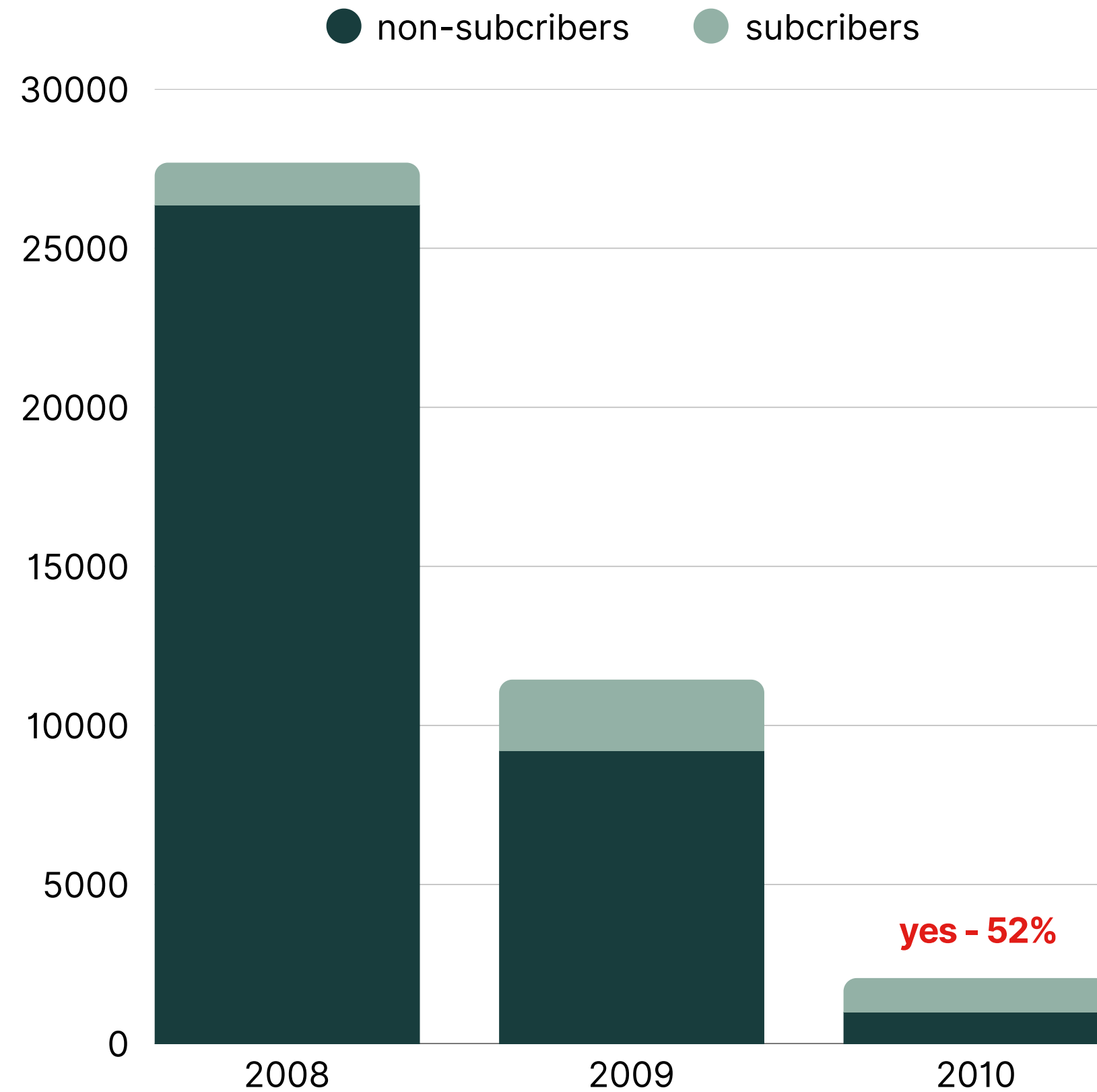
Temporal patterns

The 2010 campaign achieved the highest subscription rate (**52%**) despite having the fewest contacts, highlighting the impact of economic timing.

Material: Dataset Preprocessing

Data Cleaning	Data Transformation	Handling Imbalanced Data	Data Normalization

Duplicate values
- First Occurence Keep (FOK) strategy.

Outlier values
- IQR method.
- Winsorization.

Invalid value replacement

Feature encoding
- One-hot encoding.

Class imbalance
- Random oversampling.

Feature scaling
- Min-Max scaler.

Revolutionizing Bank Marketing:
Machine Learning Approaches for Predicting Term Deposit Subscriptions

Methodology →

Methodology: Models and Training Techniques

Baseline models:

1. Logistic Regression
2. Naive Bayes

Ensemble learning models:

1. Random Forest
2. Gradient Boosting
3. CatBoost
4. LightGBM

Training techniques:

1. Cross-validation: Stratified K-fold (k=5).
2. Hyperparameter tuning: RandomizedSearchCV.

Methodology: Evaluation Metrics & Explainability

- Confusion matrix, ROC-AUC.

- Balanced Accuracy, Sensitivity, Specificity.

- SHapley Additive exPlanations (SHAP)

- Local Interpretable Model-agnostic Explanations (LIME)

Revolutionizing Bank Marketing:
Machine Learning Approaches for Predicting Term Deposit Subscriptions
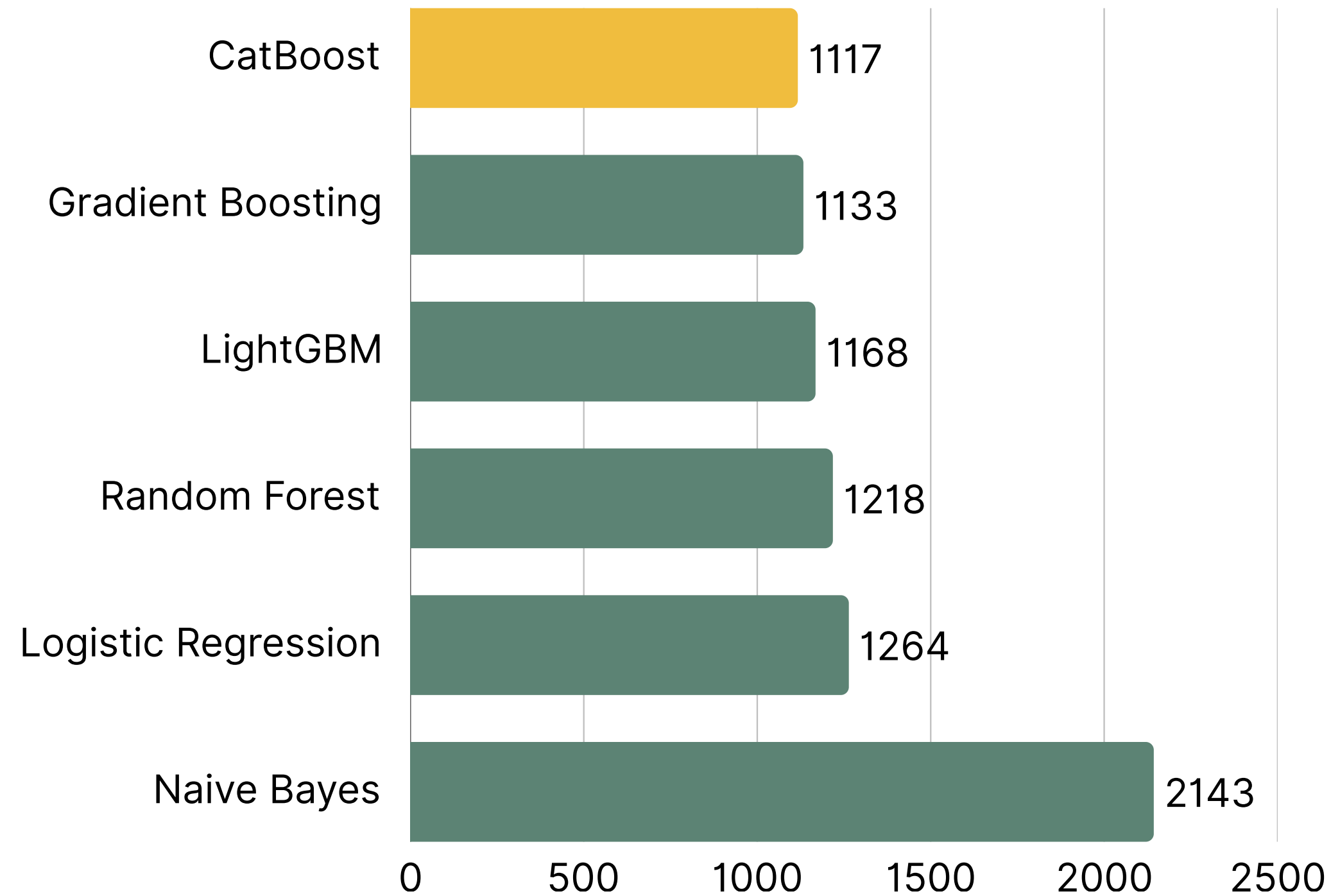
Results →

Model	Train score	Validation score	Test score
LightGBM	93.12%	**89.17%**	**89.06%**
CatBoost	95.14%	87.47%	88.10%
Gradient Boosting	89.68%	88.73%	87.99%
XGBoost	**96.02%**	86.21%	86.96%
Logistic Regression	88.16%	88.00%	86.66%
Random Forest	99.99%	75.44%	76.15%

Table 1. Model performance on training, validation and testing set.

Model	Balanced Acc	Sensitivity	Specificity	ROC-AUC
CatBoost	**89.01%**	92.38%	**85.64%**	94.69%
LightGBM	88.70%	**92.48%**	84.93%	**94.72%**
Gradient Boosting	88.67%	91.86%	85.48%	94.44%
Random Forest	88.00%	91.65%	84.35%	93.74%
Logistic Regression	86.84%	89.70%	83.98%	93.13%
Naive Bayes	75.92%	78.47%	73.38%	82.11%

**Table 2.** Model performance on testing set.

Results: Misclassified Cases

CatBoost achieves the lowest number of misclassified cases, demonstrating its strong ability to capture complex patterns and decision boundaries in customer behavior.

Results: Uncertainty Cases

CatBoost shows both a low number of uncertain cases and the lowest error rate within this group, highlighting its strong calibration and ability to handle ambiguous customer behavior effectively.
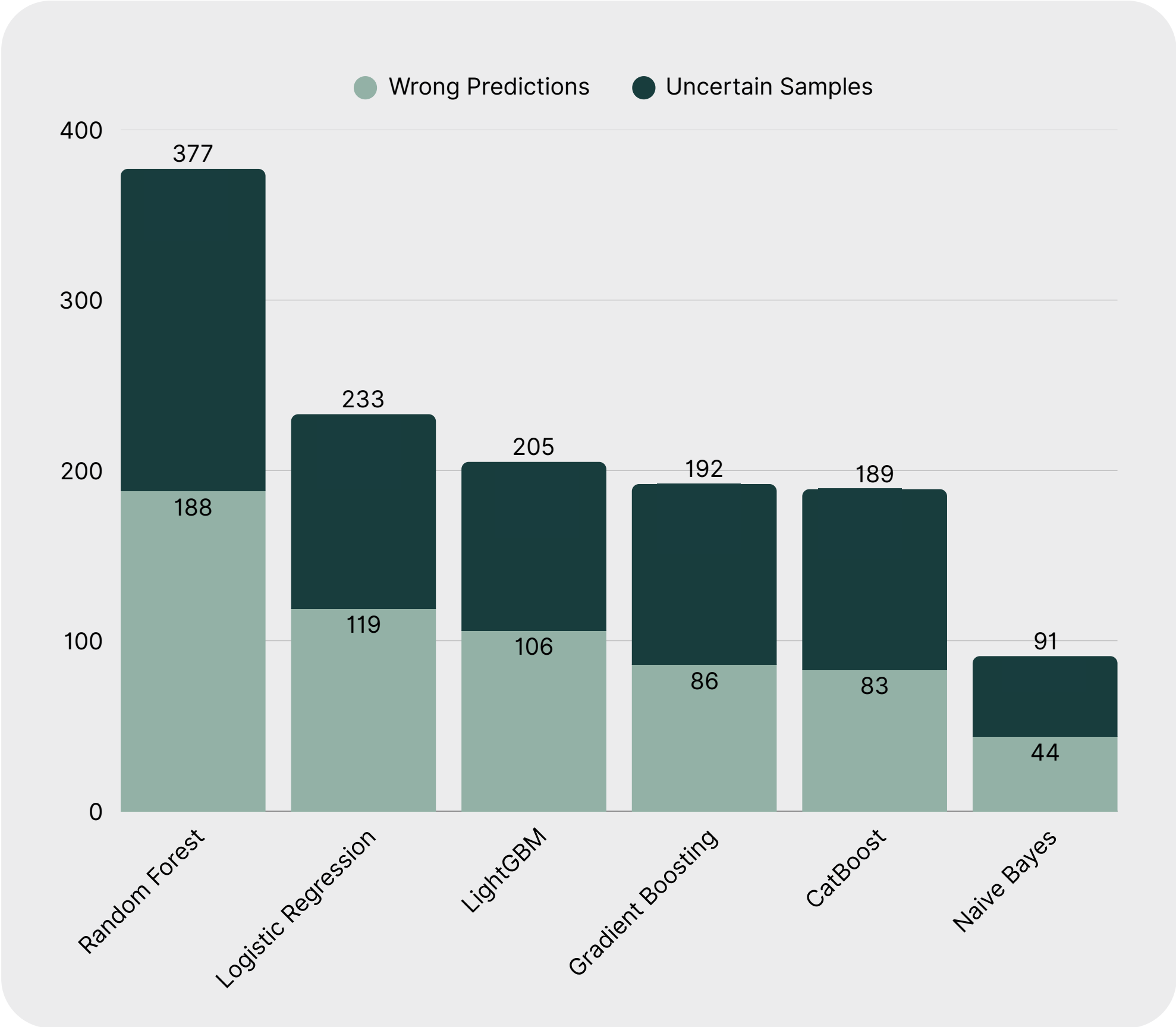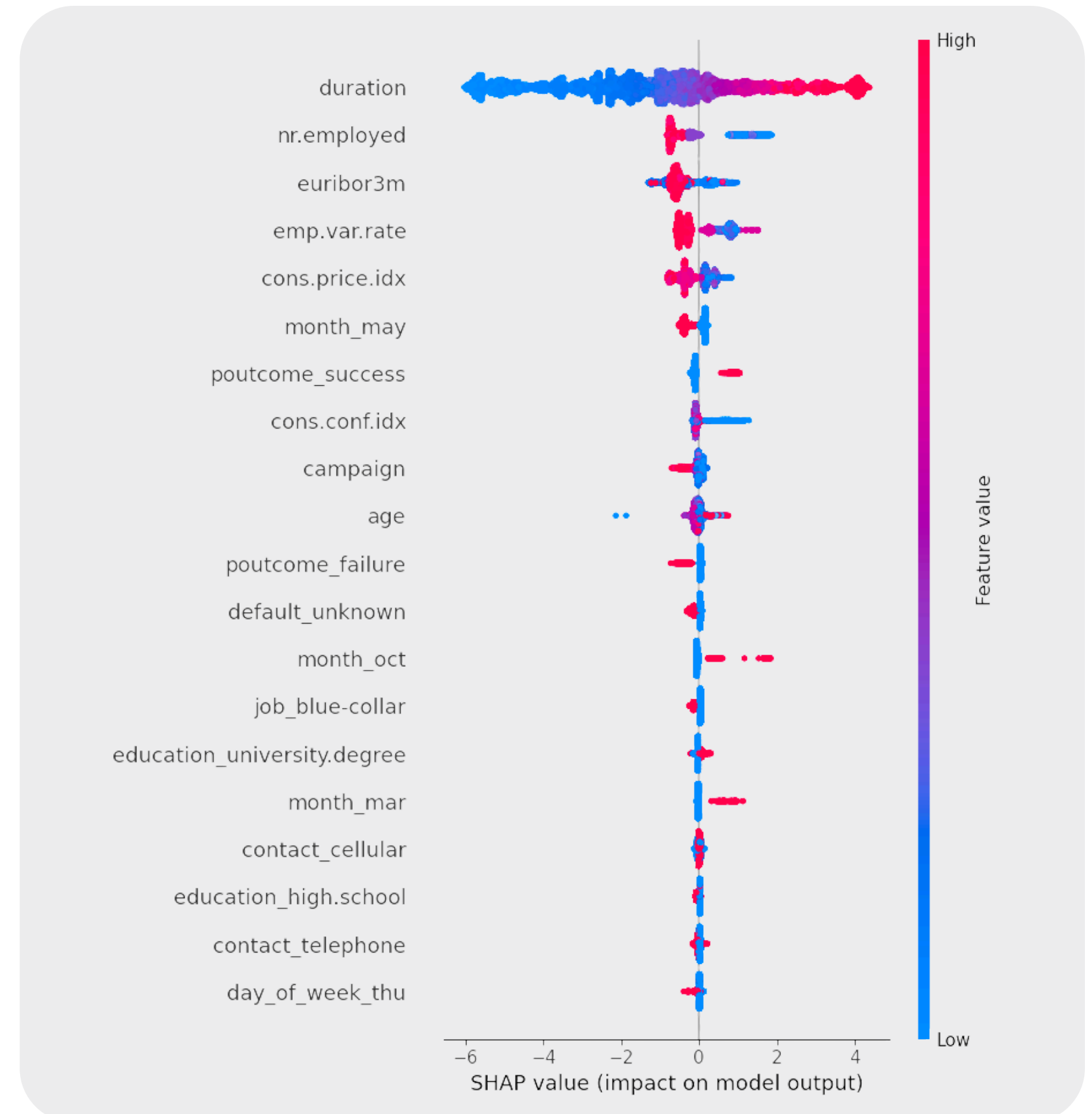
CatBoost

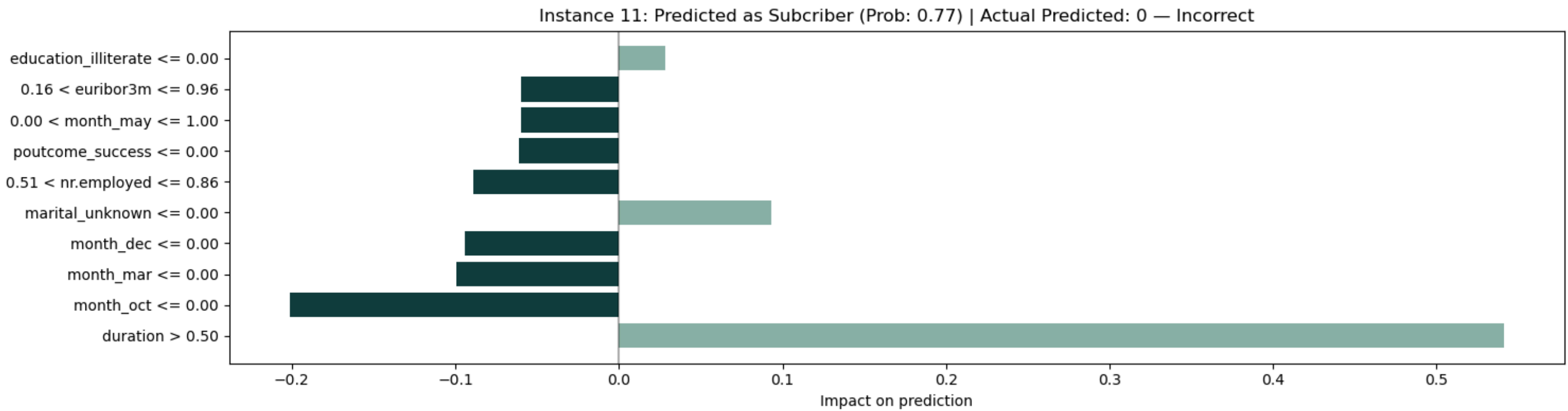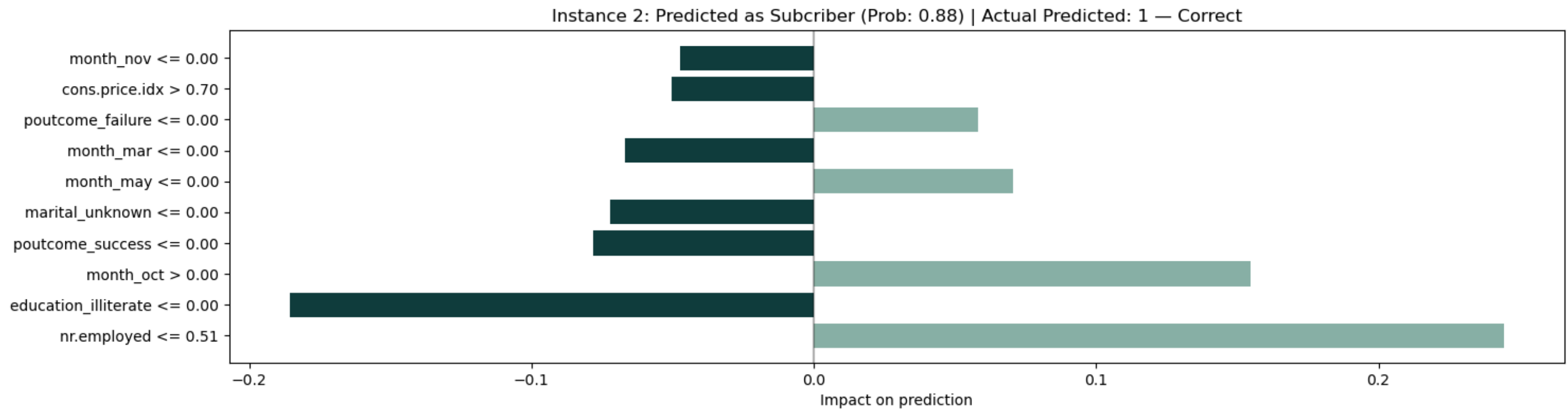| Balanced Accuracy	Sensitivity	Specificity
89%	92.4%	85.6%
ROC-AUC	**Misclassified rate**	**Correct rate**
94.7%	13.5%	56.1% *(in uncertain cases)*

Results: Top 15 Features importance
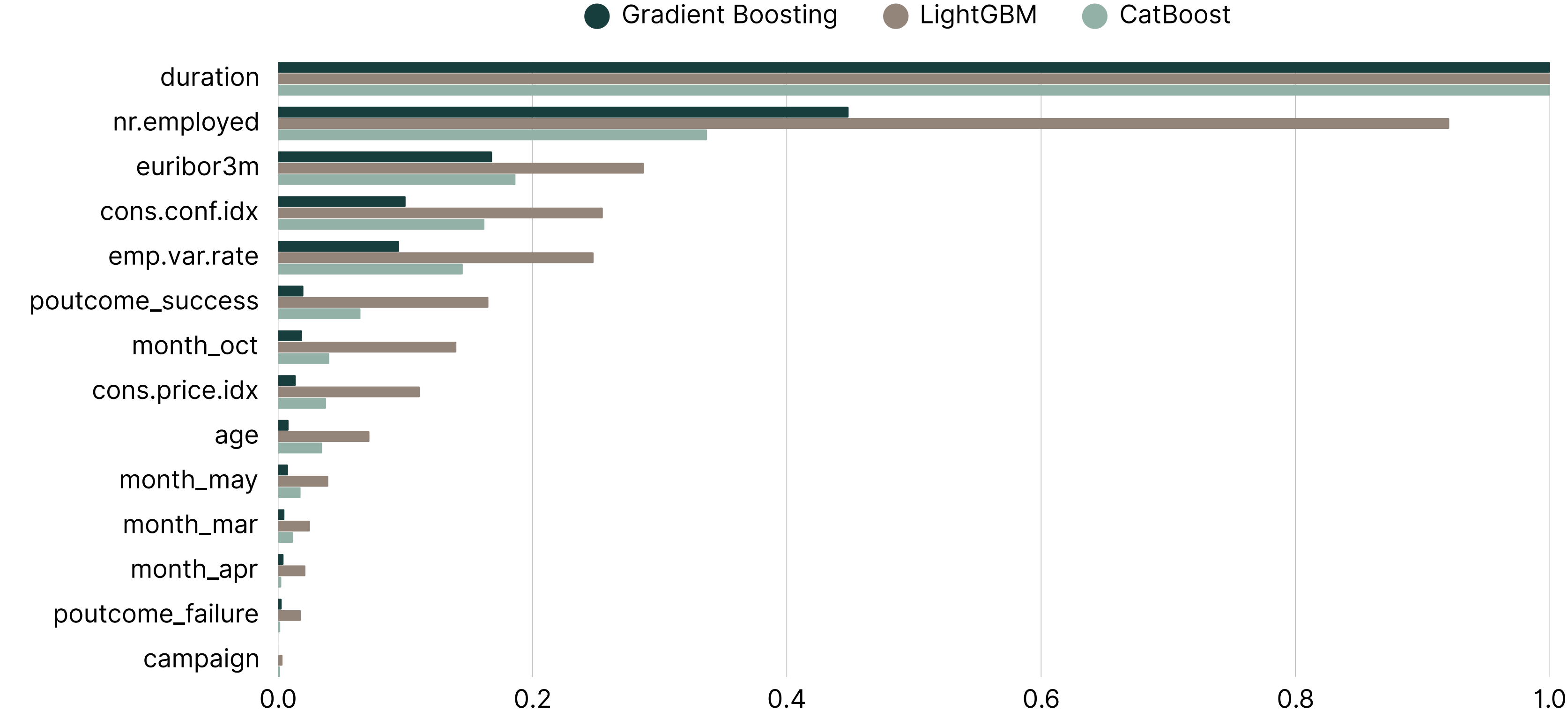
Results: SHAP analysis - CatBoost

Longer call durations and prior campaign success are strong positive predictors of term deposit subscription, while high economic indicators (e.g., **euribor3m**, **nr.employed**) tend to reduce the likelihood of subscription.

Instance 2: Predicted as Subcriber (Prob: 0.88) | Actual Predicted: 1 — Correct

Instance 11: Predicted as Subcriber (Prob: 0.77) | Actual Predicted: 0 — Incorrect

Results: LIME analysis - CatBoost

Conclusions →

Conclusions

Contribution of the thesis:
- A systematic comparison of ML models, combined with structured data preprocessing and model evaluation, and the integration of XAI to enhance model transparency and practical applicability in bank marketing.

Key findings:
- Ensemble models, particularly **CatBoost** (89%) and **LightGBM** (88.7%), significantly **outperform traditional methods**, especially when supported by proper hyperparameter tuning.
- CatBoost had fewer uncertain predictions and the lowest error rate in this group, showing better calibration and reliability than other models.
- SHAP and LIME revealed that **duration**, **poutcome**, and **macroeconomic** features play a key role in model decisions, supporting transparency and actionable business insights.

Research potential:
- This study opens up avenues for practical applications in bank marketing and the development of transparent, trustworthy AI systems in highly regulated industries such as finance.

Future Works

1. Enhance Deployment Feasibility

Evaluate training time, computational cost, and real-time responsiveness for practical implementation.

2. Incorporate Up-to-Date & Multi-Channel Data

Use real-time transactions or behavioral data to improve model relevance.

3. Advance Feature Engineering

Explore advanced strategies to enhance feature quality and representation for better model performance.

thank you for listening!

Questions and feedback are welcome.