CLIENT SUBSCRIPTION PREDICTION REPORT

By Kelvin Awuku-Boateng

1. Introduction

As part of the data analytics initiative, this project aims to support the business in identifying clients most likely to subscribe to a term deposit product. Leveraging historical banking data, we build a predictive model that determines whether a client will subscribe (target variable y: "yes" or "no"). The project encompasses the full data science workflow from data exploration and preprocessing, to feature engineering, modeling, and insight generation. These insights will help guide strategic decisions for more effective marketing campaigns.

2. Exploratory Data Analysis (EDA)

2.1 Dataset Overview

Training Samples: 45,211

Test Samples: 4,521

Features: 17 attributes (including both numerical and categorical)

2.2 Data Quality Checks

Missing Values: None detected.

Outliers: Boxplots revealed outliers in age, balance, and duration. These were retained as they may carry signal.

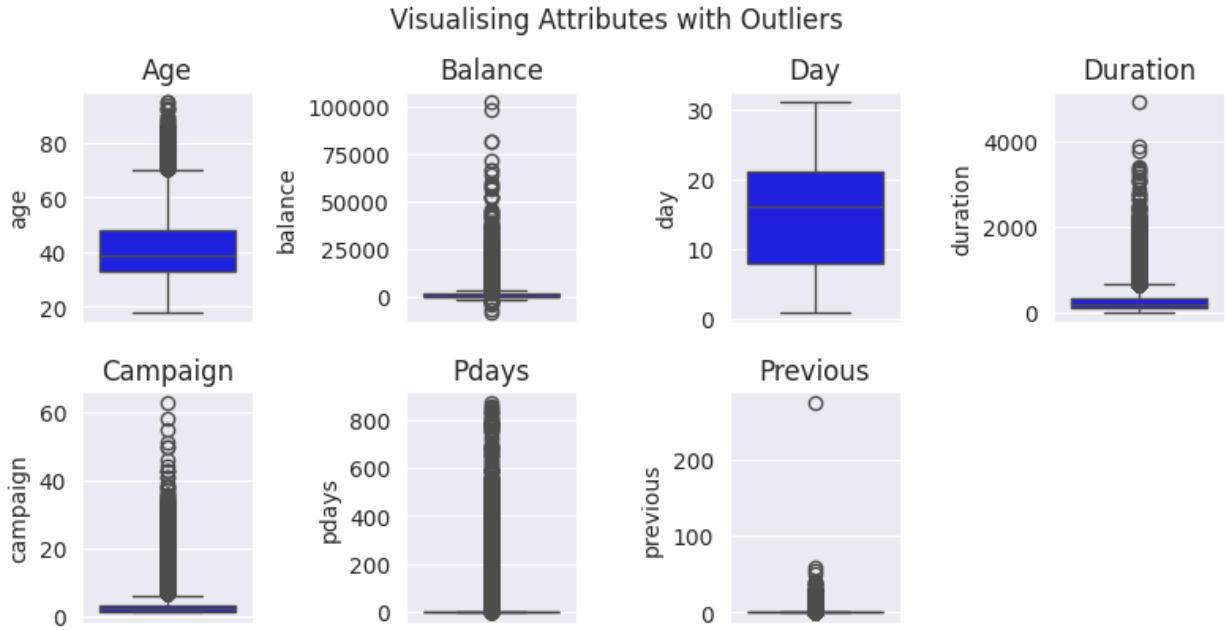

Figure 2.1: Boxplot Analysis on the training dataset.

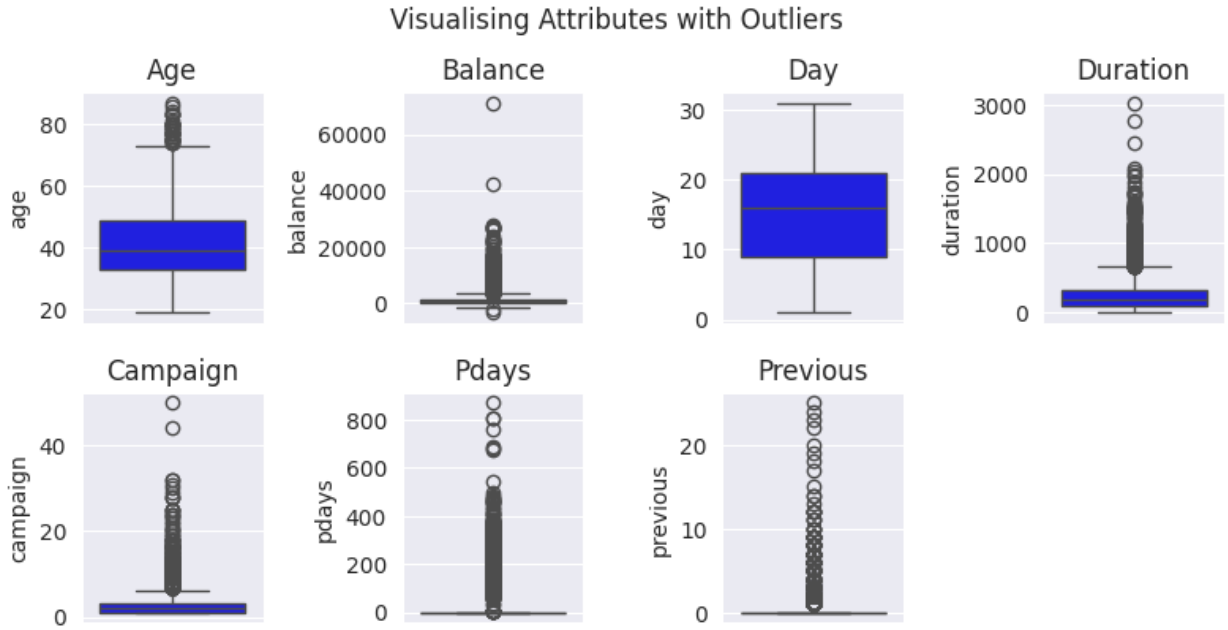

Figure 2.2: Boxplot Analysis on the test dataset.

Distributions: Histograms showed skewness in variables like balance and duration.

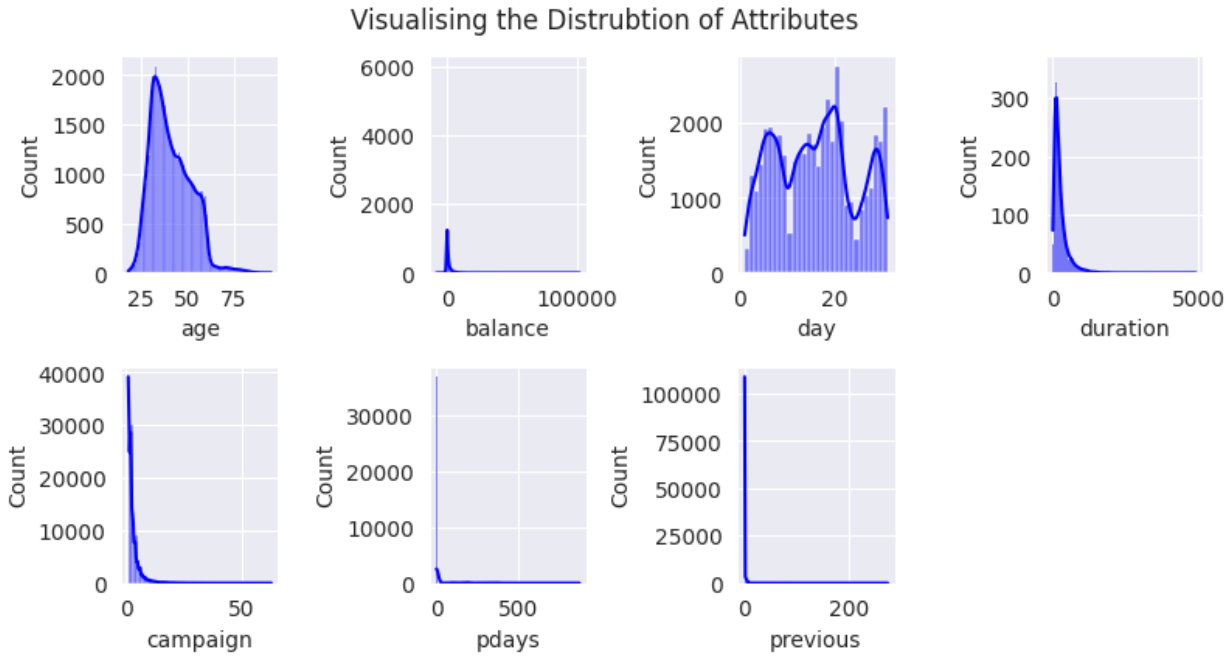

Figure 2.3: Distribution Pattern of Attributes on the training dataset.

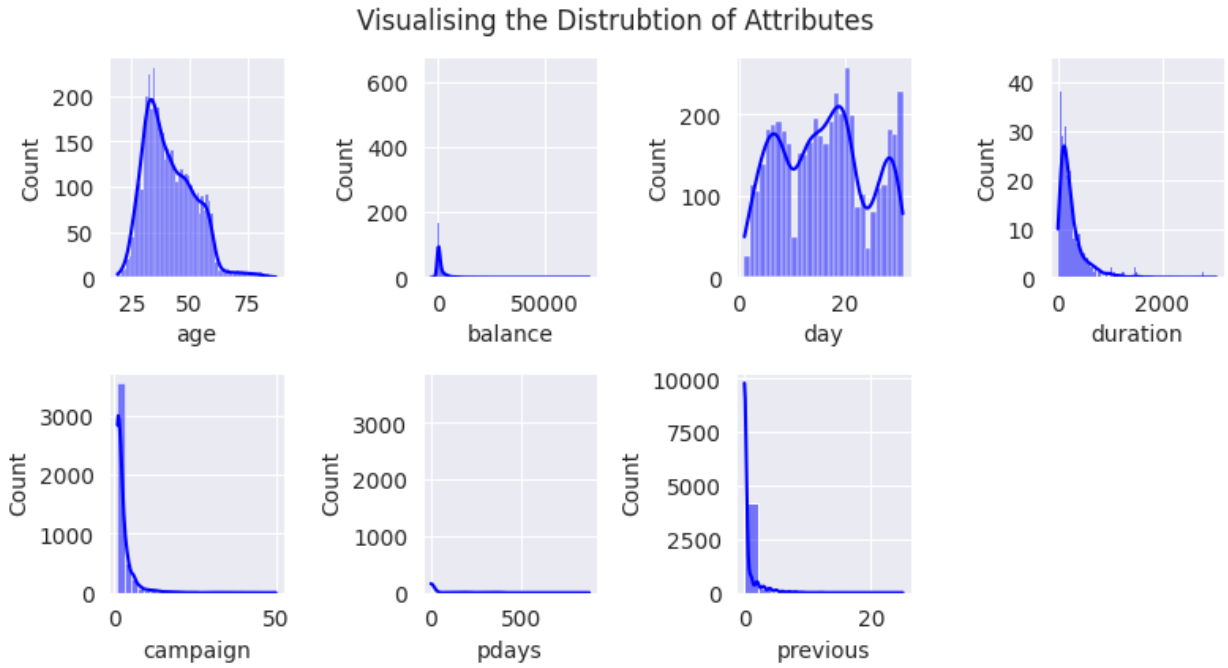

Figure 2.4: Distribution Pattern of Attributes on the test dataset.

Correlations: Positive association observed between duration, poutcome, and y. Correlation heatmaps and pairplots helped highlight impactful variables.

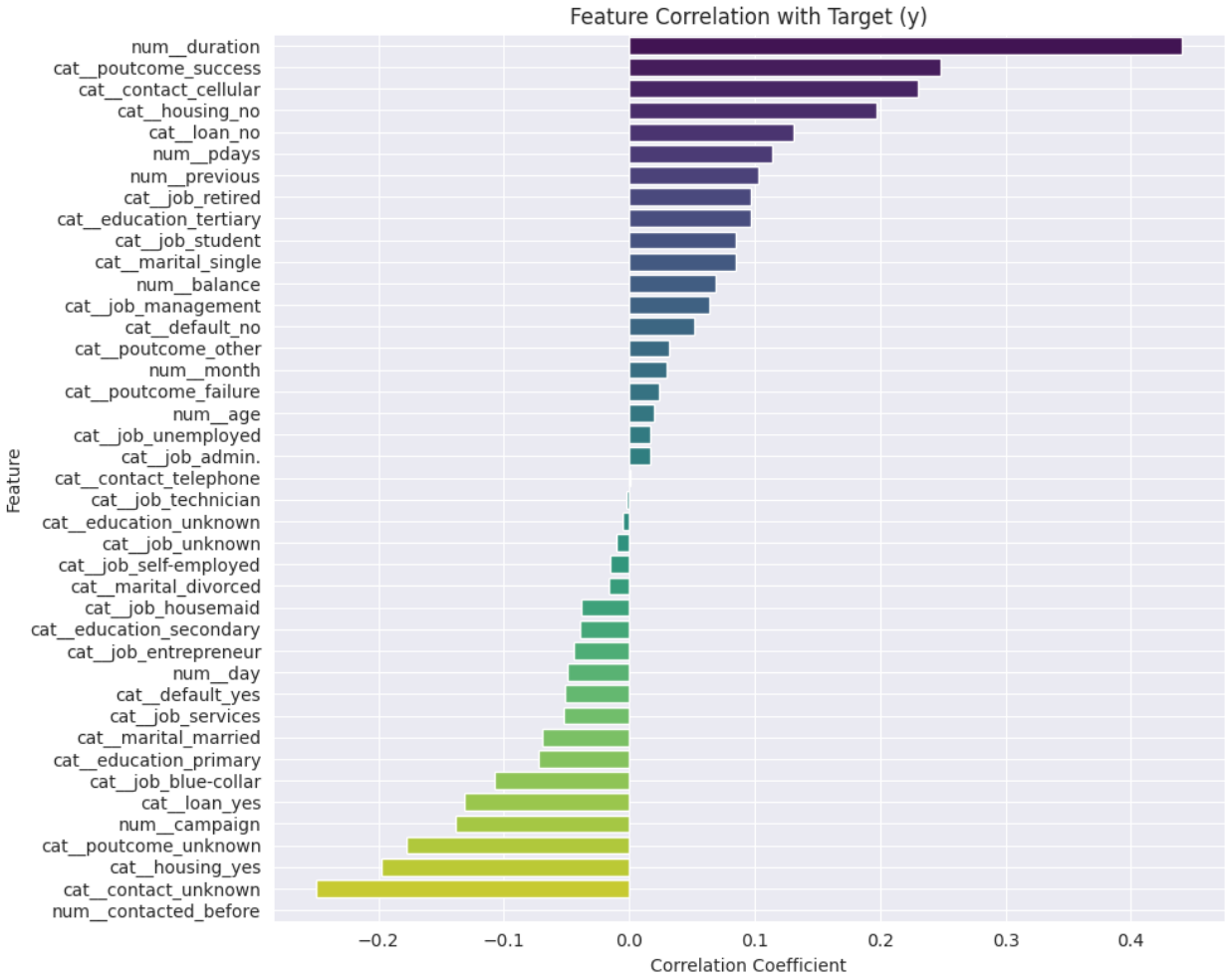

Figure 2.5: Feature Correlation with Target class.

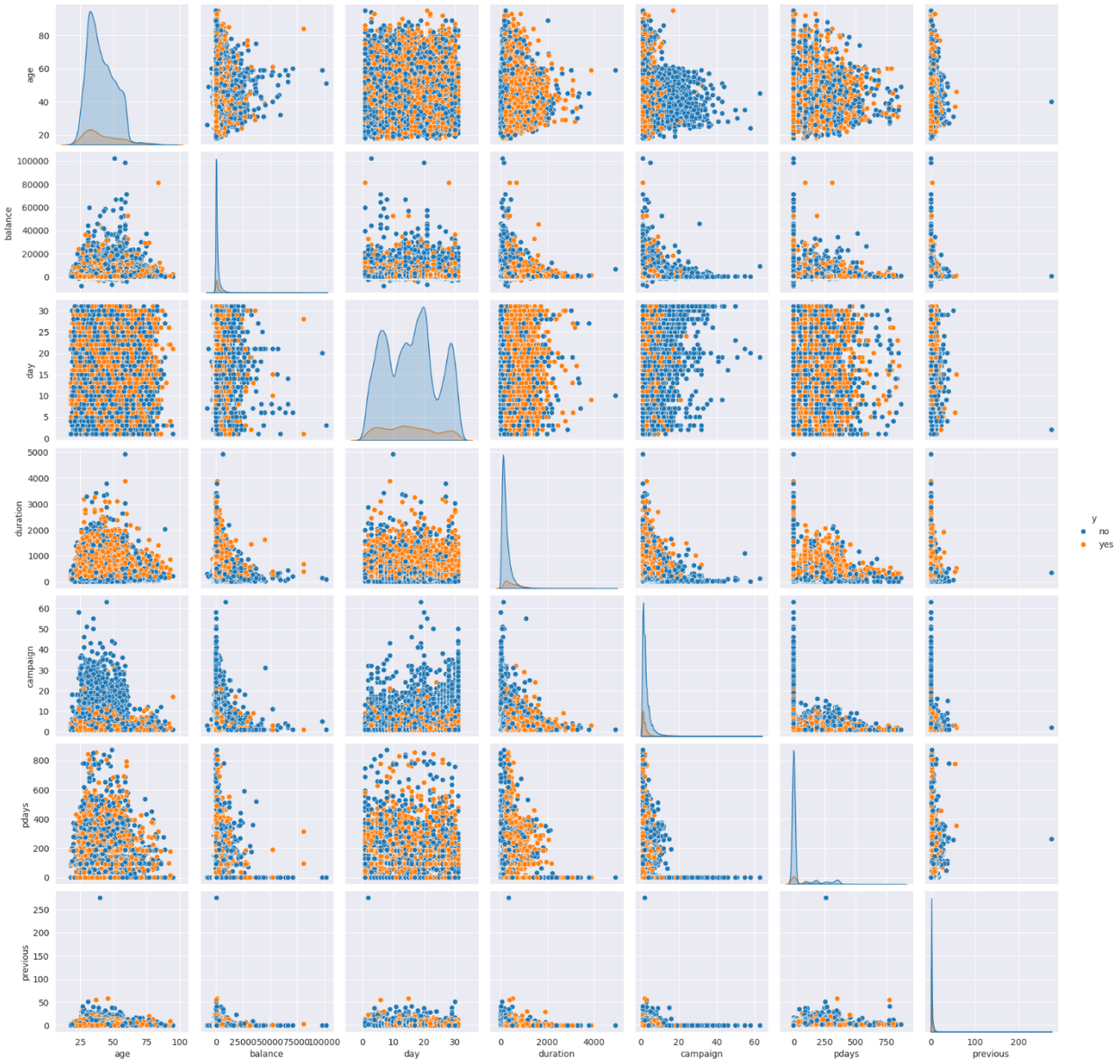

Figure 2.6: Feature Distribution & Correlation with Pairplot.

3. Feature Engineering

3.1 Preprocessing Steps

Normalization: Min-Max Scaler applied to numerical features for consistent scaling.

$$\text{Min-Max Scaler} = \frac{x_i - min(x)}{max(x) - min(x)}$$

Categorical Encoding: One-Hot Encoding applied to categorical variables (e.g., job, marital, education).

Target Transformation: Target variable y encoded to binary (yes = 1, no = 0).

Ordinal Mapping: month mapped to numerical order to retain sequence.

Feature Creation: New binary feature contacted_before derived from the pdays attribute to indicate prior contact.

4. Handling Class Imbalance

Given a significant imbalance between subscribed and non-subscribed classes, we applied ADASYN (Adaptive Synthetic Sampling) using strategic sampling of 80% to increase minority class representation.

Original: Majority class = ~88%, Minority class = ~12%

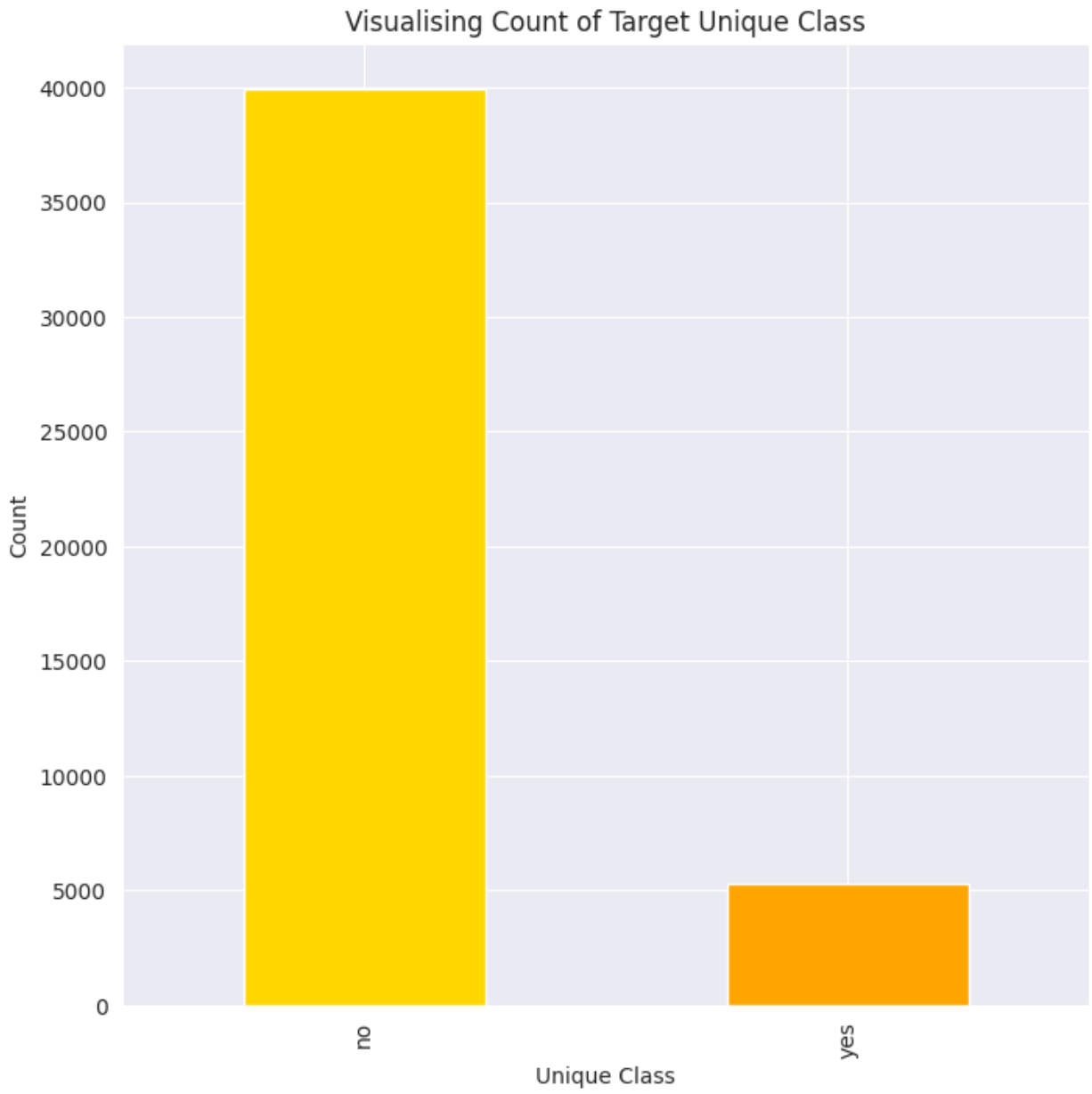

Figure 4.1: Visualizing Imbalanced Dataset.

Post-Resampling:

Total entries = 72,224

No: 40,000

Yes: 32,302

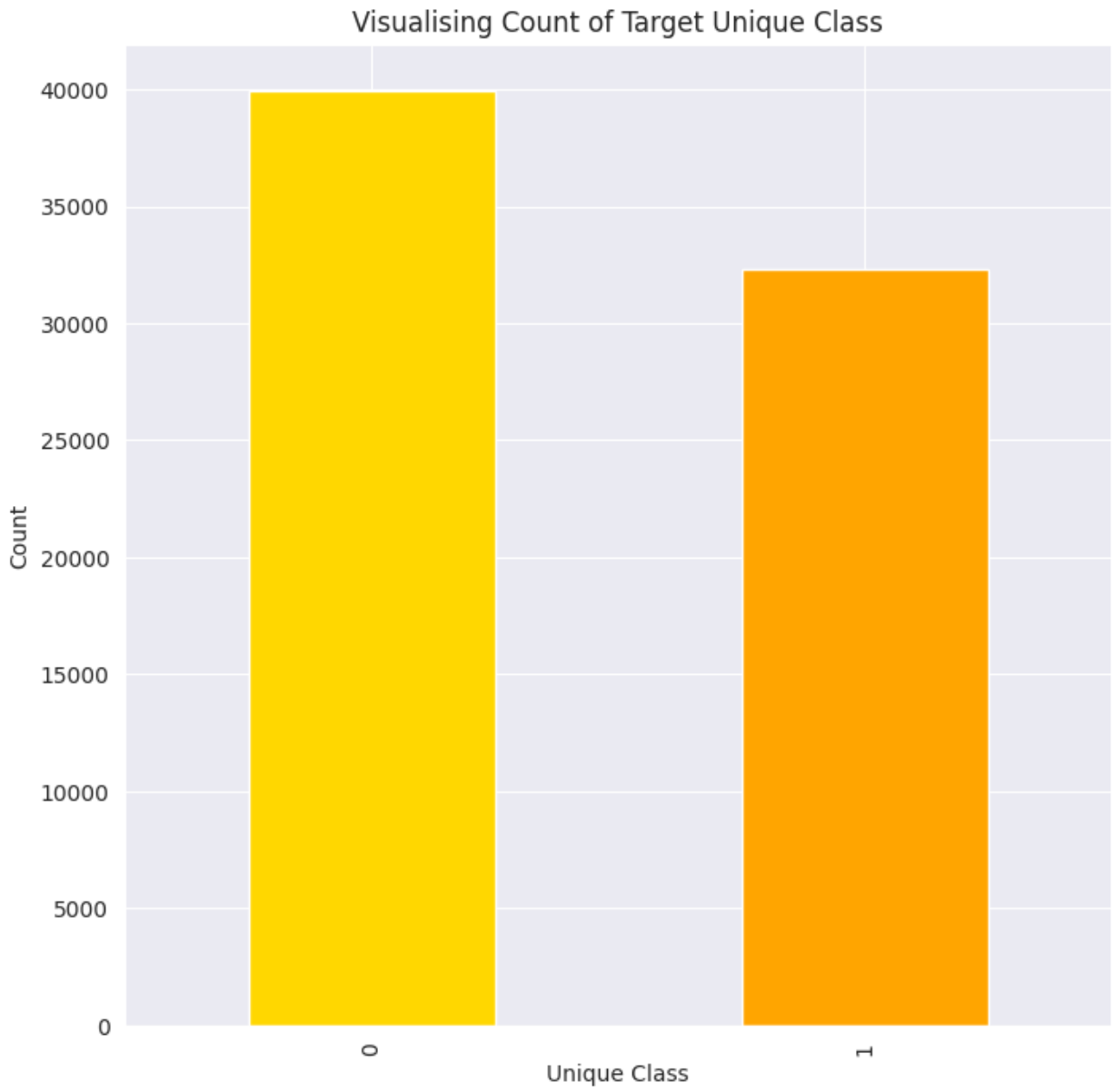

Figure 4.2: Visualizing Dataset after Adaptive Synthetic Sampling.

This step ensured more balanced learning and improved model performance, particularly for recall and F1-score.

5. Model Development

5.1 Models Trained

Based on research and initial experiments, the following models were developed:

- Logistic Regression – Baseline model offering interpretability.
- Decision Tree – Easy to explain and capable of handling mixed data types.
- Random Forest – Robust ensemble model with superior accuracy and resistance to overfitting.
- XGBoost – Gradient boosting model known for high performance on structured data.

5.2 Model Evaluation Metrics

- Accuracy – Overall correctness.
- Precision – Correct positive predictions out of all predicted positives.
- Recall – Ability to find all actual positives.
- F1 Score – Harmonic mean of precision and recall.
- ROC-AUC – Area under the Receiver Operating Characteristic curve.

6. Results and Evaluation

Table 6.1: Model Evaluation.

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Random Forest	89.45%	53.07%	72.93%	61.44%	93.05%
XGBoost	72.77%	28.53%	90.59%	43.40%	87.86%
Logistic Reg.	72.17%	28.02%	90.21%	42.77%	88.37%
Decision Tree	77.53%	29.18%	66.60%	40.58%	72.77%

7. Key Insights and Recommendations

7.1 Findings

- Best Overall Model: Random Forest, due to its high F1 and ROC-AUC scores.
- Highest Recall: XGBoost, ideal for minimizing false negatives in marketing.
- Class Imbalance had a major effect on precision, justifying the use of ADASYN.

7.2 Feature Impact

Key predictive features included:

- duration (length of last contact)
- poutcome (outcome of previous campaigns)
- contacted_before (prior engagement)
- month and campaign also showed moderate importance.

7.3 Client Behavior Patterns

- Clients more likely to subscribe:
- Had longer contact durations
- Were contacted in recent months
- Had previously responded positively to campaigns

8. Actionable Recommendations

- Target high-likelihood clients with longer calls and personalized messaging.
- Leverage prior engagement history—clients contacted earlier are more receptive.
- Focus outreach during successful months and tailor messaging by campaign effectiveness.

9. Next Steps

- Model Optimization: Tune hyperparameters for Random Forest and XGBoost.
- Feature Selection: Apply recursive feature elimination (RFE) to simplify the model.

- Deployment: Integrate the chosen model into the bank's CRM system for campaign execution.

10. Deliverables

- Report: Summary of data exploration, modeling approach, results, and insights
- Code: Full implementation in Predictive_Model.ipynb (Colab environment)
- Supporting Documents: Included research rationale and performance logs

References

1. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
2. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. IEEE International Joint Conference on Neural Networks.
3. Pedregosa et al., (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.