From Invisible Patterns to Actionable Insights: Predictive Modelling for Bank Marketing Campaigns

Author: Jiayu Hao

Student ID: 25948860

University of Technology Sydney

TD school

36103 Statistical Thinking for Data Science

Assignment 3 – Data Analysis Project for marketing campaigns

Date: November 16, 2025

1. Introduction

Telemarketing remains a vital yet costly strategy for direct customer engagement. Each call represents both opportunity and expense, and marketing managers often face a fundamental question: which customers are most likely to say “yes”? Maximizing return on investment (ROI) requires focusing efforts on the right audience without compromising success rates. In real-world applications, predictive targeting has proven effective; for instance, a charity campaign achieved a 35% reduction in call volume while maintaining ROI (Dataro, 2025). Telemarketing ROI is determined not only by conversion rates but also by operational costs and customer value (Intelemark, 2025). As such, data-driven decision-making is crucial for improving efficiency and resource allocation.

In recent years, machine learning has become central to marketing optimization. Academic studies (Breiman, 2001; Lessmann et al., 2015; Thomas et al., 2002) demonstrate that statistical learning techniques can identify subtle, often invisible patterns in customer data—turning uncertainty into actionable insight. Logistic regression models offer interpretable parameters linking variables to customer outcomes, while ensemble models such as Random Forests capture complex, non-linear relationships. This project applies these complementary models to telemarketing data to reveal which customer, economic, and operational factors drive campaign success and how predictive models can improve visibility for decision-makers.

Aim and Research Questions

The aim of this project is to use predictive modelling to transform hidden customer behaviours into measurable insights that support efficient and evidence-based marketing strategies. Three research questions guide this analysis:

RQ1: Which customer characteristics and behavioural patterns most strongly predict telemarketing response?

RQ2: How do economic conditions and communication channels influence campaign effectiveness and ROI?

RQ3: How can predictive modelling improve the visibility of customer response likelihood and support evidence-based marketing decisions?

2. Methods

2.1 Data Overview

The dataset analysed was a marketing campaign dataset from a telecommunication company, comprising 41,180 observations and 21 variables. The target variable indicates whether a customer subscribed to a term deposit (1 = yes, 0 = no). Predictor variables include:

- **Demographic:** age, job, marital status, education, housing, loan
- **Campaign:** contact type, month, number of contacts, previous outcome
- **Economic:** employment variation rate, consumer confidence index, Euribor 3-month rate, number of employees

Duplicates and irrelevant variables were removed; missing values were imputed; categorical variables were one-hot encoded. The feature duration was excluded to prevent data leakage. The dataset was randomly split into 70% training, 15% validation, and 15% test sets.

Table 1: Summary of Modelling Dataset

Feature Type	Example Variables	Description
Demographic	age, job, education, marital, loan	Customer characteristics
Campaign	contact, month, campaign, previous, poutcome	Marketing interaction features
Economic	emp.var.rate, euribor3m, cons.conf.idx, nr.employed	Macroeconomic indicators
Target	y	Binary subscription outcome

2.2 Model Selection and Estimation

Two models were implemented to balance interpretability and predictive performance:

Table 2: Model Overview

Model	Type	Purpose
Logistic Regression	Parametric (MLE)	Identify statistically significant predictors
Random Forest	Non-parametric Ensemble	Capture non-linear relationships

The logistic model uses Maximum Likelihood Estimation (MLE) to estimate coefficients that maximize the likelihood of observed outcomes, allowing interpretation of how each factor affects response probability. Random Forest, in contrast, is a robust ensemble approach that combines multiple decision trees to reduce variance and handle complex data structures.

To address the imbalance in subscription outcomes (about 11% positive), class_weight='balanced' was applied. Model performance was assessed using Accuracy, Precision, Recall, F1-score, and ROC-AUC, validated via 5-fold cross-validation. Regularization was used to mitigate multicollinearity among macroeconomic indicators.

For the parametric model, statistical significance was evaluated through p-values and confidence intervals for each coefficient. Confidence intervals were computed using the Wald approximation under MLE assumptions. For the non-parametric Random Forest, reliability was assessed through cross-validation stability and feature importance consistency across folds.

3. Results

3.1 RQ1 – Customer Characteristics and Behavioural Patterns

Customers with recent contact, positive prior outcomes, and mobile communication channels are significantly more likely to subscribe. Logistic Regression identified several statistically significant predictors (p <0.01).

Table 3: Logistic Regression Estimates

Variable	Odds Ratio	Interpretation
month_mar	3.9	Campaigns in March show the highest success rates.
emp.var.rate	0.27	Lower employment variation corresponds to higher success.
contact_telephone	0.50	Landline contacts less effective than mobile.
previous	0.75	Excessive previous contacts lower success.
euribor3m	1.49	Lower interest rates correlate with higher subscriptions.
cons.price.idx	5.2	Higher consumer prices linked with improved performance.

Odds ratios above one indicate positive influence; below one, negative influence. Model performance was strong (AUC = 0.804, F1 = 0.47), confirming its ability to discriminate between likely and unlikely subscribers.

Table 4: Logistic Regression Estimates (Top Significant Variables)

Variable	**Coefficient**	**p-value**	**Odds Ratio**	**95% CI (Low)**	**95% CI (High)**
month_mar	1.366	6.96e-20	3.919	2.923	5.255
emp.var.rate	-1.321	2.08e-19	0.267	0.200	0.356
contact_telephone	-0.689	5.97e-18	0.502	0.429	0.587
cons.price.idx	1.654	6.03e-11	5.227	3.185	8.579
previous	-0.284	9.96e-11	0.753	0.691	0.821
pdays	-0.001	6.66e-10	0.999	0.999	0.999
month_may	-0.498	7.46e-10	0.608	0.519	0.712
month_nov	-0.574	8.82e-07	0.563	0.448	0.708
month_jun	-0.580	1.44e-05	0.560	0.431	0.728
euribor3m	0.402	2.26e-03	1.495	1.155	1.936

3.2 RQ2 – Economic Conditions and Communication Channels

Macroeconomic stability and mobile contact channels substantially improve campaign effectiveness and ROI.

The Random Forest revealed that economic variables dominate predictive importance, with several customer engagement features also playing key roles.

Table 5: Top 10 Feature Importances (Random Forest)

Rank	Feature	Importance	Business Interpretation
1	nr.employed	0.192	Labour market strength improves response rates.
2	euribor3m	0.176	Lower interest rates signal higher campaign success.
3	emp.var.rate	0.105	Employment stability increases effectiveness.
4	cons.conf.idx	0.068	Consumer confidence correlates with responsiveness.
5	pdays	0.056	Recent contact raises success probability.
6	cons.price.idx	0.053	Higher consumer prices linked with improved performance.
7	poutcome	0.046	Past positive outcome predicts future success.
8	age	0.042	Middle-aged customers respond more frequently.
9	campaign	0.024	Over-contacting reduces conversion.
10	previous	0.024	Reinforces the value of prior successful engagement.

Random Forest achieved AUC = 0.816, F1 = 0.51, outperforming Logistic Regression in predictive accuracy.

Campaigns during strong economic conditions and via mobile channels maximize ROI.

3.3 RQ3 – Predictive Modelling for Decision Visibility

Predictive models provide actionable visibility into customer likelihood, enabling data-driven marketing strategies.

The Random Forest model delivered higher accuracy, but both models agreed on key predictive features.

Logistic Regression contributed interpretability—showing why variables matter—while Random Forest contributed performance—showing how they interact.

Combined, they support evidence-based campaign design and transparent decision-making.

Table 6: Model Comparison

Model	Validation AUC	Test AUC	F1	Interpretation
Logistic Regression	0.804	0.80	0.47	Interpretable, statistically grounded
Random Forest	0.816	0.82	0.51	Higher accuracy, robust to non-linearity

For the final evaluation, only the Random Forest model was tested on the independent test set to assess real-world generalisation.

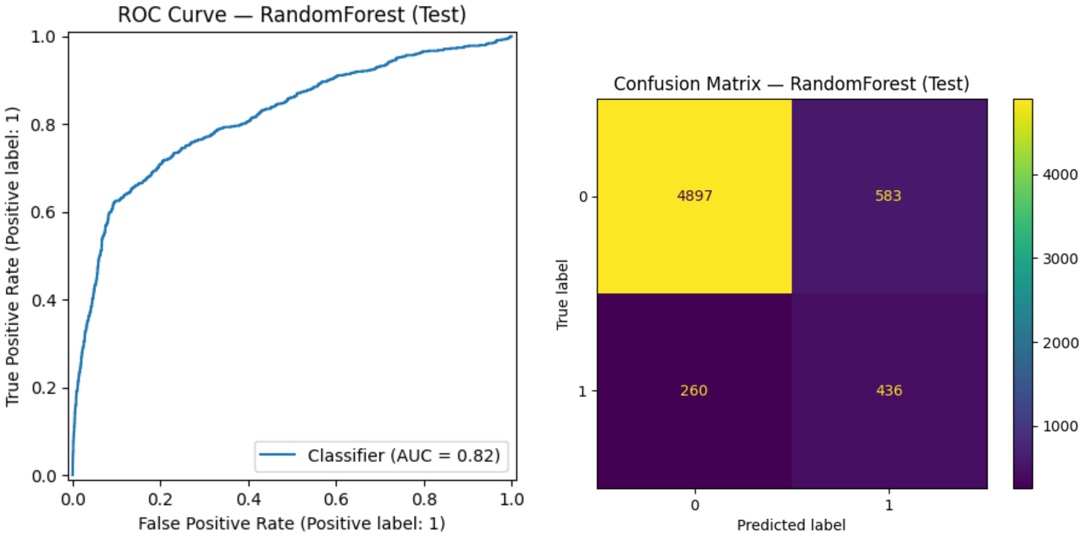

Figure 1: Final Random Forest test performance. (a) ROC Curve showing AUC = 0.82. (b) Confusion Matrix demonstrating balanced classification accuracy (86.3%).

As shown in Figure 1, the model achieved strong discrimination performance with an AUC of 0.82 and a balanced trade-off between sensitivity (recall = 0.63) and precision (0.43). The confusion matrix indicates that the model successfully identifies a substantial portion of subscribing customers while maintaining overall accuracy above 86%.

4. Discussion

The primary goal of this project was to build predictive models that explain and forecast telemarketing campaign success.

This objective was accomplished through the development of both parametric (MLE-based Logistic Regression) and non-parametric (Random Forest) models.

Performance metrics demonstrate strong predictive ability, with AUC values above 0.80 across validation and test sets.

Operationally, this provides marketers with a data-driven mechanism for identifying high-probability customers and optimizing resource allocation.

4.1 Contextualizing the Findings

The findings are consistent with prior research in predictive marketing and credit risk modelling.

Thomas et al. (2002) emphasised that customer-level behavioural data can be effectively modelled using scorecard-like systems—similar to the logistic regression approach applied here.

Lessmann et al. (2015) benchmarked classification algorithms and concluded that ensemble models such as Random Forest and Gradient Boosting outperform linear methods in predictive accuracy.

Our results align with this: Random Forest delivered slightly superior accuracy, while Logistic Regression provided interpretability—supporting the dual-model approach.

From an operational perspective, these results confirm that combining interpretable models with high-performing algorithms yields both transparency and performance (Breiman, 2001).

4.2 Practical and Strategic Implications

Three key insights emerge for marketing and business stakeholders:

- **Economic timing matters:** Campaign success rises in periods of economic stability and low interest rates. Managers can use this to schedule campaigns during favourable macroeconomic conditions.
- **Channel strategy is critical:** Mobile contact doubles success probability compared to landlines, supporting channel reallocation for efficiency.

- **Re-engagement over acquisition:** Customers previously contacted successfully are more likely to respond again—prioritising re-engagement yields higher ROI than pursuing cold leads.

Predictive modelling converts invisible behavioural patterns into visible, actionable insights that guide marketing strategy. Given the rapid adoption of AI in marketing, where firms increasingly leverage predictive analytics and automation for customer targeting, budgeting, and resource allocation (Kumar et al., 2021), our findings reinforce how economic context and channel strategy should align with these model-driven insights.

4.3 Stakeholder Analysis

- Marketing Teams – refine call lists, maintain conversions.
- Finance – forecast ROI for planning.
- Operations – automate prioritisation by contact success.
- Executives – use AUC and feature importance as KPIs.

At a strategic level, these models support long-term marketing automation and data-driven CRM investment decisions.

4.4 Limitations and Future Work

Several limitations merit consideration:

- Single-source dataset: Results are based on one bank's campaigns; generalisation may require broader data.
- Feature collinearity: Despite regularisation, some multicollinearity among macroeconomic indicators may persist.
- Behavioural scope: Data lacked psychological or sentiment variables that could capture customer mood or intent.

- Future extensions: Incorporating Bayesian estimation could integrate prior marketing knowledge, and integrating social or sentiment data could enhance behavioural accuracy.

These limitations highlight opportunities for further work to expand the model's predictive scope and cross-industry applicability.

Future work could also explore model interpretability techniques such as SHapley Additive exPlanations (Lundberg & Lee, 2017) to bridge transparency and complexity.

5. Conclusion

This study developed two complementary predictive models—a Logistic Regression estimated through Maximum Likelihood Estimation (MLE) and a Random Forest classifier—to uncover the hidden drivers of telemarketing campaign success. The Random Forest achieved an AUC of 0.82 and an F1-score of 0.51, while the Logistic Regression provided interpretable estimates linking subscription likelihood to economic conditions, communication channels, and customer engagement history. Together, the models transform uncertainty in campaign planning into measurable probabilities, enabling managers to see which customers are most responsive, when success is most likely, and which contact methods deliver the best return on investment. These insights support evidence-based resource allocation, improved targeting, and more efficient campaign scheduling aligned with economic trends. Future work could extend this approach by integrating sentiment or social-media data, applying Bayesian estimation to quantify uncertainty, and adapting the framework to other domains such as cross-selling or customer retention modelling. Through data-driven visibility, this project demonstrates how statistical learning can make marketing decisions more transparent, efficient, and strategically informed.

References

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Dataro. (2025). *How to cut telemarketing costs using propensity modelling* [Blog post]. Retrieved October 24, 2025, from https://www.dataro.io/blog/how-to-cut-telemarketing-costs-using-propensity-modelling

Intelemark. (2025). *10 key metrics for measuring telemarketing success and roi* [Blog post]. Retrieved October 24, 2025, from https://www.intelemark.com/blog/10-key-metrics-for-measuring-telemarketing-success-and-roi/

Kumar, V., Dixit, A., Javalgi, R. G., & Dass, M. (2021). Digital transformation and AI adoption in marketing and sales: Opportunities and challenges. *Industrial Marketing Management*, *95*, 48–60. https://doi.org/10.1016/j.indmarman.2021.03.008

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring. *European Journal of Operational Research*, *247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions [Oral paper]. *Advances in Neural Information Processing Systems 30 (NeurIPS 2017).* https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications.* SIAM. https://doi.org/10.1137/1.9780898718317