# Introduction

**Central question:** *Which combination of client traits, prior-campaign outcomes, and phone-call characteristics most reliably signals that a customer will subscribe to a term deposit?*

The Portuguese “Bank Marketing” dataset (OpenML ID 1461) records the outcome of 45,211 telephone contacts made between 2008 and 2010. Each row captures the client’s demographics (age, job, marital status, education), relationship attributes (balance, loans), details of the most recent call (month, day, duration), and a history snapshot (number of calls this campaign, days since last contact, previous-campaign outcome). After removing rows with missing job or education, dropping the non-informative contact channel, and de-duplicating, 43,193 observations remained. The target variable, y, is binary (1 = subscribed, 0 = not), but the data is imbalanced, with only about 12% being positive.

**Why does it matter?** Telephone campaigns are expensive. Dialling fewer, better-chosen prospects could significantly lift profit, thereby enhancing campaign efficiency and profitability. Targeting fewer but higher-quality prospects allows banks to allocate resources strategically and improve overall client engagement.

# Exploratory Data Analysis

## 1. Correlation Landscape

A heat-map of the seven numeric predictors plus *y* tells a surprisingly simple story: call duration dominates. At $r \approx 0.40$ it carries four times more linear signal than any other numeric variable. Visualising duration in box-plots confirms the jump: the median length of successful calls is roughly twice that of unsuccessful ones, and the upper quartile stretches well beyond two minutes. This is intuitive—longer conversations imply interest—but the strength of the relationship suggests preserving duration in raw form and engineering log or bucketed versions to capture its nonlinear climb.

Recency variables offer mild support: *pdays* (days since previous contact) and *previous* (number of earlier calls) correlate around 0.10. Interestingly, *campaign* (the number of calls in the current effort) correlates negatively after five attempts, hinting at diminishing returns rather than simple persistence paying off. Age, day-of-month, and balance show negligible linear association; their influence, if any, is likely non-linear or conditional on other factors.

## 2. Single-feature subscription rates

### *2.1 Previous-campaign outcome*

The horizontal bar chart of subscription rates across *poutcome* levels is unequivocal: prior success matters. Clients who previously accepted a product convert at approximately 65%, a six-fold increase over the "unknown" baseline (<10%). Even the ambiguous *other* outcome triples the baseline. This magnitude reinforces *poutcome* as the categorical analogue of duration—a single variable that almost pre-decides the outcome.

### *2.2 Seasonality*

Conversion is not constant through the calendar. March stands out (>50%), followed by September, December, and October (40–45%). May—the traditional spring push—actually performs worse (<10%), counter to any naïve assumption that "big marketing months" equal high success. Whether this reflects macro-economic noise or campaign fatigue, month must be encoded as ordered categorical (and perhaps cyclic if periodicity is suspected).

### *2.3 Job group*

Labour segments reveal clear tiers. Students top the list (≈32%), likely because small balances and flexible schedules make a term deposit appealing. Retirees follow (≈23%), plausibly attracted by capital-preservation. Manual and entrepreneurial roles cluster near the bottom (≤10%), suggesting that resources or risk preferences matter significantly.

### *2.4 Education & Marital*

Education maps monotonically onto probability—tertiary graduates reach approximately 16%, whereas primary-educated clients languish below 10%. Marital status shows a mild but notable tilt: singles (~15%) edge out married (10%) and divorced (12%), reflecting life-stage influences.

## 3. Cross-category interactions

### *3.1 Job × Education*

A heat-map of subscription rates by the 11 job classes and three education tiers reveals consistent lift from education within every occupation. Education seems to amplify inherent job-level propensity: retired-tertiary (28%) still beats retired-primary (22%), but the absolute gap is small compared with the spread across jobs. Conversely, self-employed primary-educated clients form the nadir (~4%). Thus, separate dummy variables for education and job type are essential, as combining them would dilute critical information.

### *3.2 Job × Poutcome*

Every row in the heat-map's "success" column exceeds 50%, while the "failure" column rarely exceeds 20%. Crucially, the lift is not uniform—managerial success hits 68%, blue-collar only 57%. Capturing this heterogeneity suggests adding interaction terms explicitly or relying on tree-based models for automatic feature exploitation.

### *3.3 Outcome × Month*

Seasonality amplifies outcome effects: March success climbs to 77%, whereas May failure collapses to 8%. Unexpectedly, even "unknown" outcomes experience a boost from September to December (~40%), perhaps reflecting end-of-year promotions. These patterns justify composite features (e.g., *poutcome_success* × *month_Mar_Sep*).

### 4. Call-intensity dynamics (Figure 9)

Plotting average success versus the number of calls (capped at 20) illustrates diminishing returns: the first call delivers 14%, the second 11%, and by the tenth call, conversion is <5%. Calls beyond 15 yield anecdotal and volatile results; therefore, campaigns are capped at 10, with a binary “persistently pursued” flag.

### 5. Binned numeric trends *(Figure 10)*

For top numeric features, dividing ranges into 50 equal-width bins and plotting average *y* uncovers insightful patterns. *Duration* exhibits an S-curve: conversion shoots up quickly and plateaus around 1,500 seconds. Very recent *pdays*contacts (0–100 days) hover near 25%, while extremely long gaps yield unreliable extremes due to scant data. The *previous* feature peaks when a client has been approached 5–15 times in earlier campaigns, declining beyond 30. The *campaign* feature trends monotonically downward, reinforcing earlier caps.

### 6. Key Insights

1. **Quality over quantity.** Long, substantive conversations are worth more than repeated short attempts. Duration dwarfs all other numeric predictors.
2. **History outweighs demographics.** A single prior success multiplies the chance of a new sale five-fold; failures suppress it, and the effect is season-sensitive.
3. **Propensity clusters by life-stage.** Students and retirees—especially those with tertiary education—are the easiest wins, whereas blue-collar and self-employed primary-educated clients are least receptive.
4. **Persistence has limits.** Additional calls after the second sharply erode returns. Campaign and previous counts must be capped, and class-weighted losses applied during modelling to discourage over-dialling.

These observations are broadly intuitive—engaged, informed, or demographically aligned clients respond better—but quantify the magnitude of each factor and highlight an unexpected seasonal spike in March.

# Models & Methods: Predictive Modelling

I trained and evaluated 5 classification models—Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbours (KNN), and Multi-layer Perceptron (MLP) to predict whether customers would subscribe to a term deposit. Each model was trained and tested on the same cleaned and processed dataset containing 43,193 records and 19 features. The preprocessing steps included scaling numerical variables, encoding categorical variables, and adding engineered features based on the exploratory analysis results, such as capturing prior campaign interactions and seasonal effects.

Because only about 12% of customers subscribe, models were compared primarily using the ROC-AUC metric, as it effectively measures performance despite class imbalance. I also reviewed precision, recall, accuracy, and the confusion matrix counts. In this context, false negatives (missing a potential subscriber) are costlier than false positives(unnecessary additional calls), making recall particularly important.
Below are the results, interpretation, and conclusions for each model.

---

**1. Logistic Regression**

The logistic regression model achieved a ROC-AUC of 0.909, with good recall (0.821) but relatively low precision (0.405). The confusion matrix showed 1,236 correctly identified subscribers (true positives), but missed 270 potential subscribers (false negatives) and made 1,817 unnecessary calls (false positives).
The strongest predictors identified by logistic regression included call duration, successful previous campaign outcomes, and contacts in March. Negative factors were unsuccessful previous campaigns, January calls, and certain blue-collar jobs. Logistic regression provided clear, easily interpretable insights into feature impacts.

---

**2. Decision Tree**

The decision tree achieved a ROC-AUC of 0.900, slightly lower than logistic regression, but notably improved recall to 0.862, meaning fewer missed subscribers (208 false negatives). However, this came with lower precision (0.341), resulting in many additional unnecessary calls (2,514 false positives).
The tree relied heavily on call duration as its primary decision factor, followed by successful previous campaigns and the customer's housing status. While the decision tree's logic is straightforward and easy to explain to stakeholders, its heavy reliance on a few factors makes it less robust.

**3. Random Forest**

The random forest model improved significantly, achieving a ROC-AUC of 0.927, with balanced recall (0.723) and precision (0.523). It correctly identified 1,089 subscribers while reducing false positives (992), though it increased missed subscribers to 417.

Feature importance was more evenly distributed, with duration still the most influential, followed by customer age, account balance, day of the month, and past campaign success. Random forest thus improved reliability and generalisation, trading off some recall for fewer false positives.

**4. K-Nearest Neighbours (KNN)**

KNN underperformed compared to other models, with a ROC-AUC of 0.882—the lowest among tested methods. Precision was high (0.725), but recall was very poor (0.221), leading to 1,173 missed subscribers. KNN struggled due to difficulty in handling the large number of categorical features. It is not recommended for practical deployment in this case.

**5. Multi-layer Perceptron (MLP)**

The neural network (MLP) showed strong results, achieving a ROC-AUC of 0.926, similar to the random forest. Precision and recall were balanced at around 0.59 and 0.60, respectively, resulting in fewer missed subscribers (596) than random forest but more false positives (637). The MLP offered strong predictive capability but at the cost of interpretability and increased complexity.

## Cross-Model Comparison and Deeper Insights

Across all models, call duration consistently emerged as the most powerful predictor. Its significance across model types reinforces the practical recommendation to focus resources on cultivating more extended, meaningful customer interactions.

Random forest and MLP stood out due to their ability to capture intricate, nonlinear relationships, outperforming logistic regression and decision tree models, which were more constrained by linear assumptions or simplistic decision logic. Despite MLP's complexity, its superior recall and balanced confusion matrix position it as a powerful tool if interpretability is less critical.

The decision tree model was notable for exceptionally high recall, ideal for campaigns prioritising subscriber acquisition at all costs. However, this method would incur substantial unnecessary follow-up costs due to many false positives.

Logistic regression, while easier to interpret, sacrificed precision to achieve recall comparable to more sophisticated models. Its insights were invaluable for clearly communicating to stakeholders why certain clients were targeted.

KNN was consistently inferior in our testing, underscoring the importance of model selection tailored to dataset characteristics—particularly handling categorical features and scalability.

# Conclusion

A consistent and structured comparison across five distinct models has illuminated important methodological insights and practical recommendations for enhancing bank marketing efficiency.

The random forest model emerged as the most balanced and effective choice for deployment. It achieved the highest ROC-AUC, effectively balanced recall and precision, and showed robustness through diversified feature importance. Its ability to handle complex interactions without substantial interpretative sacrifice makes it an ideal candidate for practical implementation.

However, if the business case prioritises recall heavily—maximising subscriber capture even at the expense of increased call volume—the decision tree provides superior recall, albeit with trade-offs.

The MLP neural network offers excellent predictive performance but is more challenging to interpret. Its effectiveness depends on stakeholder willingness to accept complexity for improved results.

Overall, the comprehensive feature engineering informed by detailed exploratory analysis significantly contributed to model performance. Critical insights, such as the diminishing returns of repeated calls, the heightened impact of previous campaign outcomes, and nuanced seasonal variations, allowed targeted and effective predictor transformations.

Thus, strategically deploying the random forest model, guided by careful attention to customer call duration, historical interaction outcomes, and demographic sensitivities, offers significant efficiency gains. Ultimately, leveraging these analytical insights can materially enhance customer targeting precision, reduce campaign costs, and substantially improve deposit acquisition performance.