Predicting Subscription to Bank Long-Term Deposits

Modelling direct marketing success for long-term deposits using Binomial Logistic Regression

---

Thao Le

Table of Content

I. Problem Description

1.1 Business Objective

Atlantic Savings Bank, a retail institution, is looking to refine its direct marketing strategy. Currently, their campaigns, primarily conducted via phone calls to promote long-term deposits, are proving to be both costly and resource-intensive. The challenge lies in the fact that clients often require multiple contacts before subscribing, and not all outreach efforts result in a successful conversion. Atlantic Savings Bank is seeking a more efficient approach to allocate its marketing resources, ultimately aiming to boost the success rate of these crucial campaigns.

In today's highly competitive capital market, investors have a wide array of financial instruments for saving, leveraging, and investing. However, many younger customers are grappling with financial hurdles such as stagnant wages and escalating housing costs. This has led to a lack of confidence in traditional savings methods, pushing them towards higher-risk assets like cryptocurrency and NFTs in pursuit of quicker returns.

This shift presents a significant challenge for retail banks trying to attract and retain customers, especially given that their long-term deposit interest rates are relatively low compared to other financial products. To address this, banks must first understand what makes customers likely to deposit their money. Then, they need to implement targeted marketing campaigns, potentially by offering competitive interest rates and additional services like insurance and capital management.

1.2 Data Mining Objective

The main objective of this project is to maximize the subscription rate from direct marketing efforts while reducing marketing costs and optimizing target strategy by developing a predictive model that accurately identifies which customers are most likely to subscribe to long-term deposits.

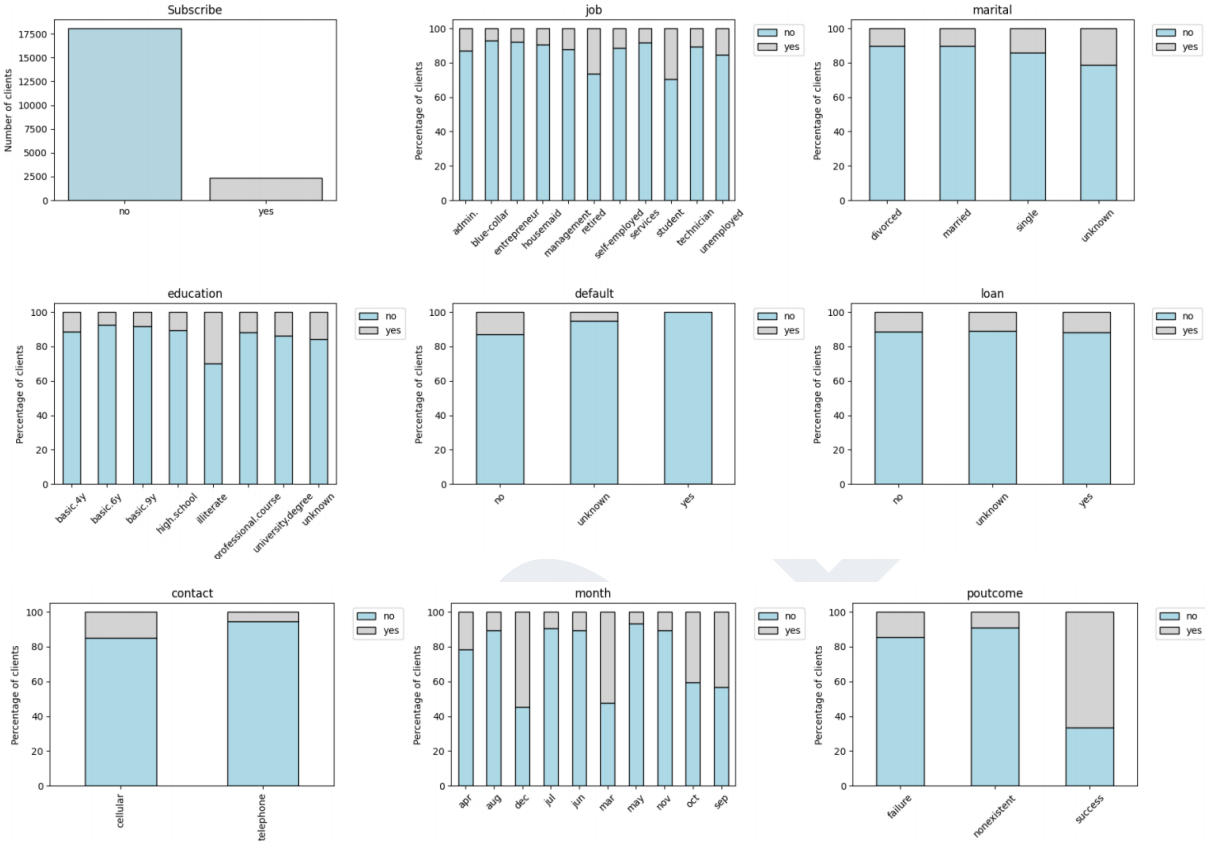
Our dataset consists of information from 20,420 bank clients, organized into 18 different variables. These variables include personal details like age, job type, marital status, and education level, along with financial indicators such as whether a client has defaulted on credit or taken out a personal loan. It also has data on marketing efforts, including how clients are last contacted (e.g., phone call, email) and when, plus details on past campaign outcomes and the total number of contacts made. Finally, the dataset incorporates broader economic factors like unemployment rates, inflation (consumer price index), consumer confidence, and interest rates.

II. Research Methodology and Results

2.1 Exploratory Analysis

In this dataset, we have 'Subscribe' as response variable which indicates whether the customer subscribes to the bank's long-term deposit. According to the following bar charts, it can be seen that dataset predominantly composed of clients who did not subscribe to the bank's offering (i.e., 18,082 out of 20,420 with 'No' answer), highlighting a significant class imbalance in the target variable, 'Subscribe'. Demographic variables like 'Job', 'Marital' status, and 'Education' show varying proportional subscription rates, suggesting their influence on client behaviours. Notably, 'single' individuals and those with

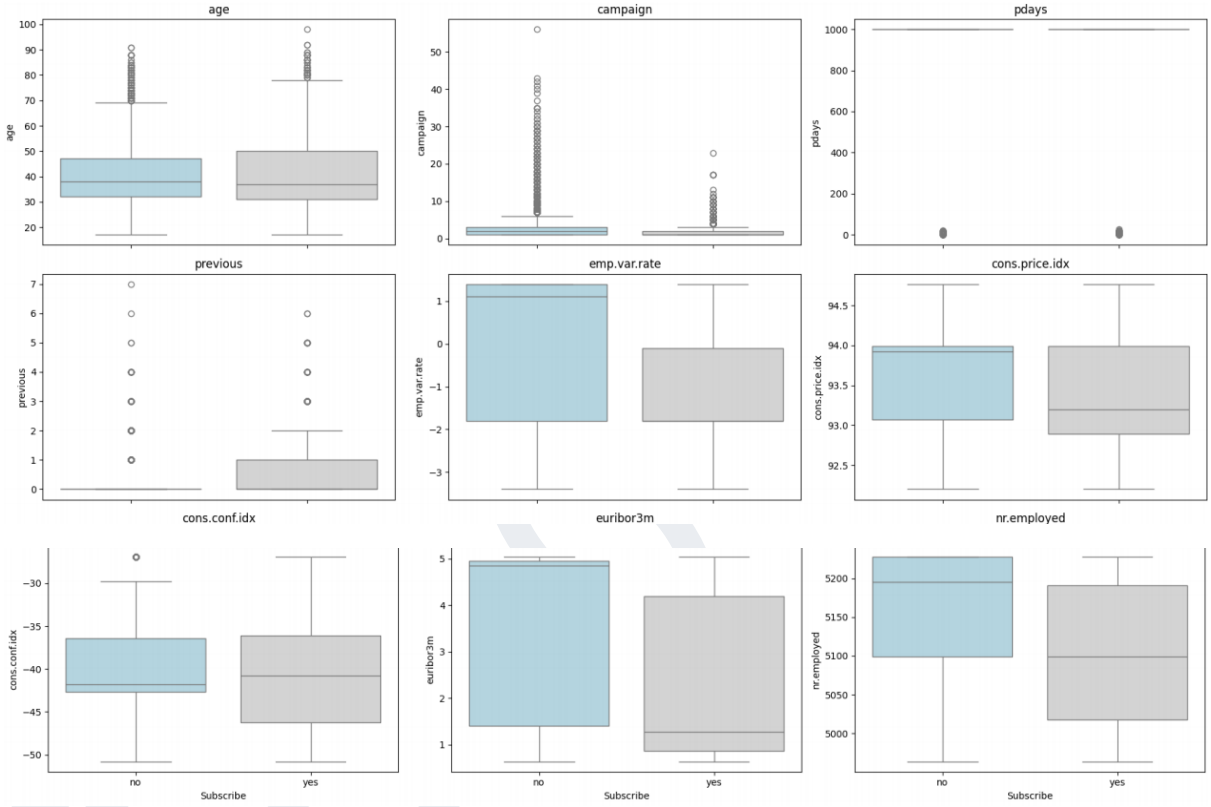
‘university.degree’ or ‘illiterate’ education appear somewhat more likely to subscribe. ‘student’ and ‘retired’ seem to have a relatively higher percentage of ‘yes’ subscriptions compared to their overall client count, although their absolute numbers might be small. Marital status seems to influence subscription propensity, with ‘single’ individuals potentially being more open to subscribing compared to ‘married’ or ‘divorced’ clients.



To gain a comprehensive understanding of the relationships between our target variable, ‘Subscribe’ and other numerical variables, we employ box plots. These visualizations effectively display the distribution of key numerical features - namely ‘age’,’campaign’ (number of contacts during the current campaign), ‘pdays’ (days since last contact from a previous campaign), and ‘previous’ (total previous contacts) - as well as variables reflecting the broader social and economic context. By splitting these distributions based on whether a client subscribed (‘yes’) or not (no’), we can discern potential differences and identify influential factors.

While the median ‘age’ is similar for both groups, ‘yes’ subscribers show a slightly older median age and a broader age distribution, including a higher proportion of older individuals, suggesting that older demographics might be more receptive to the bank’s offerings. However, the most striking differentiators emerge from the social and economic indicators. Clients who subscribed did so during periods characterized by lower employment variation rates (emp.var.rate), indicating more stable or slightly declining employment uncertainty. They also subscribed during times of lower consumer price index (cons.price.idx), implying reduced inflation, and crucially, during periods of higher consumer confidence index (cons.conf.idx), reflecting a more optimistic economic outlook. Furthermore, a highly significant relationship is observed with the Euribor 3-month rate

(euribor3m); subscribers are associated with periods of substantially lower interest rates. Lastly, ‘yes’ subscribers are also linked to periods with a slightly lower number of employees (nr.employed) in the economy, which could signify economic adjustments that make financial products more appealing. These comprehensive insights underscore the profound impact of the macroeconomic environment on client subscription behaviours, indicating that a favourable economic climate, particularly with lower interest rates and higher consumer confidence, plays a crucial role in successful marketing campaigns.



2.2 Missing values and outliers handling

The dataset contains no missing values but does contain outliers in its numeric variables. From the box plots, outliers are particularly noticeable in variables such as ‘age’ where both ‘no’ and ‘yes’ subscribers show extreme older ages, and critically in ‘campaign’ and ‘previous’ contacts, where some clients (especially non-subscribers) were contacted an exceptionally high number of times without subscribing. ‘Previous’ contacts stand out with the highest percentage of outliers at 13.73%, indicating a significant portion of clients with more than zero prior contacts. ‘Campaign’ also has a notable 5.93% of outliers, suggesting a number of clients were contacted excessively. ‘Pdays’ which is number of days that passed by after the client was last contacted from a previous campaign also exhibits outliers. Crucially, ‘emp.var.rate’, ‘cons.price.idx’, ‘euribor3m’, and ‘nr.employed’ show no outliers according to the IQR method where data points falling below Q1−1.5×IQR or above Q3+1.5×IQR are flagged as outliers. To deal with these outliers, a multi-faceted approach is recommended: extreme values in ‘age’ should be investigated for data entry errors or retained if genuine demographic segments; for ‘campaign’ and ‘previous’ contacts, transformation (e.g., logarithmic) or capping (winsorization) can reduce their

disproportionate influence, or alternatively, utilizing tree-based models which are inherently robust to outliers.

In this project, we use winsorization to deal with outliers in above-mentioned variables. Winsorization is a statistical method that handles outliers by capping them at a specified percentile. Instead of removing extreme values, it replaces them with the value of a specific percentile. For example, if we winsorize at the 5th and 95th percentiles, any value below the 5th percentile is replaced with the 5th percentile's value, and any value above the 95th percentile is replaced with the 95th percentile's value. For 'age' and 'cons.conf.idx', with their relatively low outlier percentages (around 1%), winsorization can effectively cap the extreme values without significantly altering the main distribution. This preserves more data than outright removal and can make models more robust.

2.3 Variable identification

In building our predictive model, the variable 'subscribe,' indicating whether a client opts for the long-term deposit, serves as our dependent variable or response. Our initial data exploration, visually represented through the bar plots and box plots provided earlier, highlights significant proportional differences in subscription rates ('Yes' vs. 'No') across several key features. Past campaign results, education, age, and job type are particularly anticipated to shape how customers choose long-term deposits. For example, older clients often prefer safe options like these deposits, while students or those with less education might also choose them for good interest, as they have fewer other investment choices. Additionally, customers who had a good experience in a past campaign are more likely to resubscribe, trusting the bank's service and understanding the potential benefits. Specifically, among categorical variables, we observed notable variations in 'poutcome' (previous campaign outcome), 'month' of contact, 'job' type, 'age' and 'education' level. Similarly, continuous variables like 'emp.var.rate' (employment variation rate), 'cons.conf.idx' (consumer confidence index), and 'nr.employed' (number of employees) also demonstrated distinct distributions by subscription status, suggesting their strong influence on a client's decision to subscribe.

2.4 Research Methodology

We choose binomial logistic regression method (logit model) to model choice probabilities.

The customers have only two choices between 1 (subscribe) and 0 (not subscribe) of long-term deposit (the choices are discrete dependent variables having only two values and distributed under Bernoulli distribution). The logit regression is less sensitive to outliers than linear regression. The linear regression to model the observed choices (1 and 0) might not be useful in capturing the relationship between IVs and DV and predicting DV. In addition, binomial logistic regression should have results similar to the probit model but it is easier to interpret and the probabilities are easier to be computed. More advanced methods such as decision trees and support vector machines can be applied. However, those methods will not be covered in our report due to their complexity.

In order to apply logit model, we need some following assumptions:

- The probability Yi = 1 depends on a continuous latent variable

$$Y_i = 0 \quad \text{if} \quad Y_i^* \leq 0$$

$$Y_i = 1 \quad \text{if} \quad Y_i^* \geq 0$$

- Yi∗ is a continuous variable with the following linear function

$$Y_i^* = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_k x_{i,k} + \epsilon_i$$

2.5 Regression models

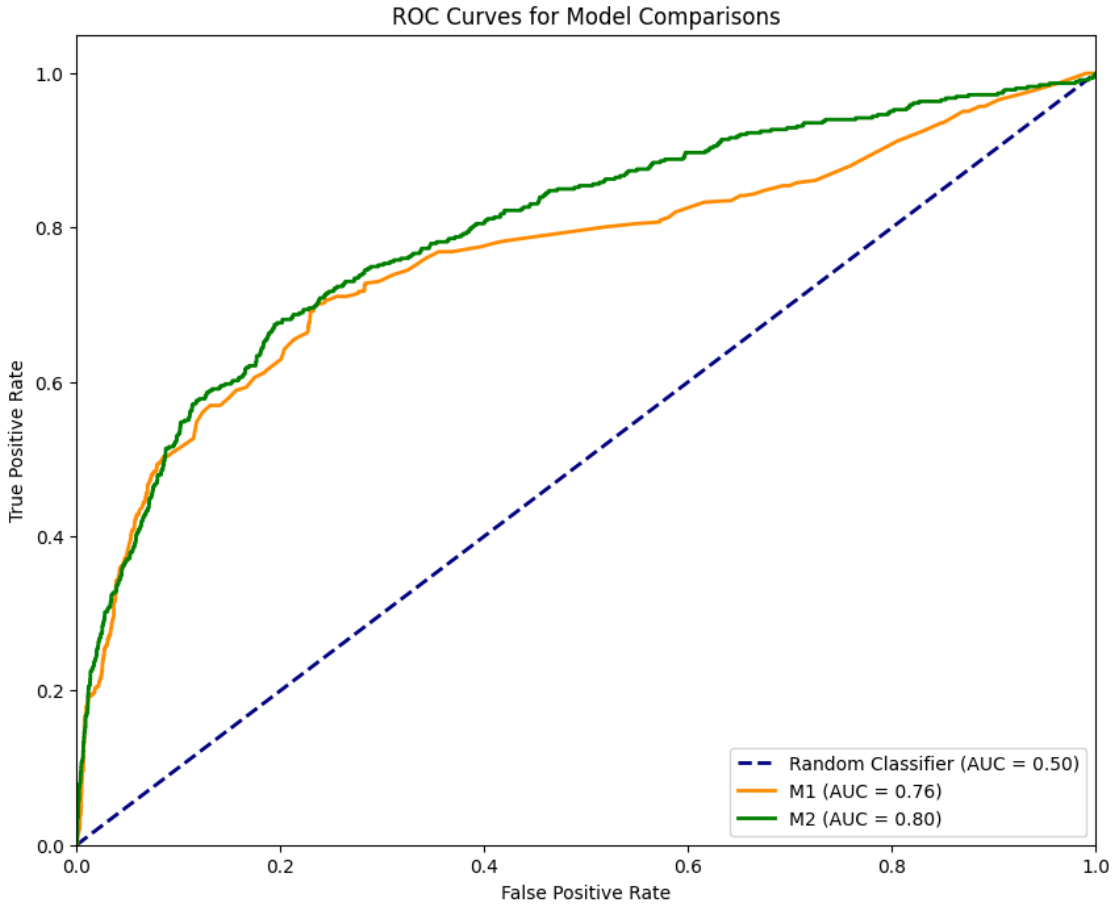
To enhance subscription rates through direct marketing, while simultaneously cutting costs and refining our targeting, we will construct two distinct models. Model M1 (Baseline Model) will analyse the correlation between client subscription behaviour and previous campaign outcomes, alongside key socio-economic indicators such as the employment variation rate, consumer price index, consumer confidence index, Euribor 3-month rate, and the number of employees. Model M2 (Alternative Model) will then explore the significance of contact month, job type, client age, and education level in influencing subscription decisions.

Before modelling, we split dataset into training and testing set with ratio 4:1. As our dataset is imbalanced with 'No' answer, there's a risk that a purely random split might assign all (or disproportionately many) samples of the minority class to either the training or the test set, leading to biased model training or unrealistic performance evaluation. Therefore, it is crucial to ensure that the proportion of samples for each class is roughly the same in both the training set and the test set as it is in the complete input dataset.

Model	Log likelihood	p-value	Pseudo R^2	AIC	BIC	Hit Rate
M1	-4751.8	0.000	0.1824	9517.6204	9571.5283	0.8976
M2	-4601.1	0.000	0.2083	9270.1972	9532.0355	0.8989

Analyzing the performance of our two models, M1 and M2, we can see that model M2 consistently outperforms M1 across multiple key metrics. Both models are statistically significant, evidenced by their p-values < 0.05, confirming their ability to explain subscription behaviour beyond random chance. However, M2's higher log-likelihood of -4601.1 compared to M1's -4751.8 indicates that M2 provides a better fit to the data. This is further corroborated by M2's higher Pseudo R^2 (0.2083 vs. 0.1824). Furthermore, M2's lower AIC (9270.1972) and BIC (9532.0355) values, relative to M1, signal that it achieves a better balance between model fit and model simplicity. While both models demonstrate very high hit rates (around 0.89), this metric should be interpreted cautiously given the significant class imbalance in the 'subscribe' variable, necessitating a closer examination of precision, recall, and F1-score for the minority 'Yes' class to fully assess their predictive capabilities for actual subscriptions. The ROC (Receiver Operating Characteristic) curve graph below compares the performance of two models, M1 and M2, against a random classifier in predicting client subscriptions. The dashed blue line represents a random classifier, which performs no better than chance. Its Area Under the Curve (AUC) is 0.50, meaning it has a 50% chance of distinguishing between positive and negative classes, serving as a baseline for comparison. Overall, both M1 and M2 are significantly better than

a random guess. However, model M2 stands out as the superior performer, exhibiting a greater capacity to accurately differentiate between clients who will subscribe and those who will not. The higher AUC of M2 suggests it is more effective in ranking potential subscribers from non-subscribers.



Based on log-likelihood, Pseudo R^2, AIC, BIC, and AUC, Model M2 generally performs better than Model M1 in explaining and fitting the subscription data. Therefore, we will re-estimate M2 with entire dataset. In order to understand the model and interpret the effect sizes, we calculate the exponential of coefficients.

In binomial logistic regression model, odds = $\exp(\beta_0 + \beta_1 x_i)$, can be understood as the likelihood of happening ($Y_i = 1$) versus not happening ($Y_i = 0$). For example, odds = 2 means that the likelihood of happening is 2 times larger than the likelihood of not happening. When x changes by 1 unit, odds change by a factor of $\exp(\beta_1)$ - the odds ratio (new odds/old odds). As such, for one unit change in x, the odds change by (odds ratio - 1)%.

$$Change\ in\ odd = (exp(\beta_1) - 1) * 100$$

The model results indicate that certain variables significantly influence a client's likelihood of subscribing to a long-term deposit. Positive impacts are strongly associated with clients who are illiterate (in the education category), those with a successful outcome from a previous campaign (poutcome_success), and being contacted in March or July. Conversely, factors that significantly decrease subscription odds include higher emp.var.rate, higher nr.employed, being in a blue-collar job, and being contacted during May, November,

October, September, or June. Interestingly, variables such as age and cons.conf.idx do not demonstrate a statistically significant relationship with subscription.

Variable	coef	std err	z	P>|z|	exp(coef)-1
age	-0.0009	0.003	-0.311	0.756	-0.0009
emp.var.rate	-0.6799	0.111	-6.131	0	-0.4939
cons.conf.idx	-0.0067	0.008	-0.886	0.376	-0.0067
nr.employed	-0.0091	0.001	-9.218	0	-0.0091
job_blue-collar	-0.2296	0.095	-2.405	0.016	-0.2048
job_entrepreneur	-0.1721	0.155	-1.112	0.266	-0.1579
job_housemaid	-0.2654	0.183	-1.452	0.147	-0.233
job_management	0.0319	0.101	0.315	0.753	0.0324
job_retired	0.2362	0.132	1.788	0.074	0.2664
job_self-employed	0.0153	0.138	0.111	0.911	0.0154
job_services	-0.1386	0.105	-1.322	0.186	-0.1293
job_student	0.1867	0.139	1.344	0.179	0.2052
job_technician	-0.0176	0.088	-0.199	0.842	-0.0174
job_unemployed	0.099	0.154	0.642	0.521	0.1042
education_basic.6y	-0.0838	0.147	-0.569	0.57	-0.0804
education_basic.9y	-0.0653	0.113	-0.576	0.565	-0.0632
education_high.school	-0.101	0.111	-0.912	0.362	-0.096
education_illiterate	1.053	0.761	2.173	0.03	1.866
education_professional.course	-0.0559	0.122	-0.456	0.648	-0.0544
education_university.degree	-0.0192	0.11	-0.174	0.862	-0.019
education_unknown	0.0769	0.146	0.528	0.598	0.0799
month_aug	0.2486	0.126	1.98	0.048	0.2822
month_dec	0.3832	0.256	1.499	0.134	0.4668
month_jul	0.3011	0.108	2.779	0.005	0.3512
month_jun	-0.2364	0.116	-2.047	0.041	-0.2098
month_mar	1.0227	0.148	6.922	0	1.7808
month_may	-0.6883	0.089	-7.719	0	-0.4972
month_nov	-0.4635	0.131	-3.53	0	-0.3703
month_oct	-0.3944	0.162	-2.43	0.015	-0.3263
month_sep	-0.3893	0.168	-2.311	0.021	-0.3235
poutcome_nonexistent	0.3955	0.076	5.185	0	0.4851
poutcome_success	1.2281	0.111	16.517	0	2.415

III. Conclusions and Recommendations

Our logistic regression model predicts the probability of a customer subscribing to a long-term deposit based on various characteristics, essentially allowing us to identify potential customers from their historical data. The model quantifies the impact of each characteristic on this probability through its estimated coefficient (and the derived exp(coef)-1 value, representing the percentage change in odds).

This insight empowers the bank to tailor future campaigns. For instance, being contacted in March or July significantly boosts subscription odds (178% and 35% increases, respectively), making these prime months for outreach. Furthermore, a successful outcome

from a previous campaign (poutcome_success) is an extremely powerful predictor, increasing the odds of subscription by over 241%, while clients with illiterate education also show a remarkably higher propensity (over 186% increase in odds).

Conversely, some factors significantly decrease subscription odds. A higher emp.var.rate reduces odds by nearly 50%, and being a blue-collar worker decreases odds by 20%. Higher employement variation rate can be understood as an indicator of economic uncertainty or instability in the job market. During periods of fluctuating employment, individuals may become more cautious with their finances, prioritizing liquidity and avoiding long-term financial commitments due to concerns about their job security or overall economic outlook. Contact during May, November, October, September, and June also leads to substantial decreases in subscription probability (ranging from 21% to nearly 50% reduction in odds). This might be linked to seasonal patterns in consumer behavior and banking activity. For example, May might be associated with spring spending (e.g., home improvements), and the latter months (September, October, November) often precede the holiday season, when consumer spending shifts towards gifts and other seasonal expenses, potentially diverting attention and funds away from long-term financial products like deposits. Variables like age and cons.conf.idx do not show a statistically significant relationship with subscription in this model.

By applying a probability cutoff (e.g., 50%), customers can be classified into "potential" and "non-potential" groups. This segmentation allows the bank to optimize resource allocation, investing more heavily (staff time, incentives) in potential customers, thereby minimizing acquisition costs while maximizing new long-term deposits.

However, our current method of selecting this cutoff prioritizes overall accuracy, which may sacrifice the identification of some true potential customers. Reducing the cutoff to increase true positives would also increase false positives, leading to higher acquisition costs. Addressing this, the management team faces the critical question of the optimal cutoff, requiring further analysis of the trade-off between the cost of missing a potential customer and the cost of a false positive acquisition, potentially necessitating more data. Additionally, a more advanced consideration for the management team might involve using clustering algorithms to group potential customers by their distinct characteristics, enabling tailored incentive offerings.