Predicting and Understanding Customer Subscription Decisions in Banking Campaigns

Team 10: Tim Su, Anita Lin, Charmaine Lin, Sherry Zhu, Jessica Yang

University of Southern California
DSO 585: Data Driven Consulting
December 2025

1. Executive Summary

Retail banks rely heavily on outbound direct marketing to promote term deposits, but these campaigns face tight operational constraints: call center capacity is limited, outreach is expensive, and contacting low-propensity customers can reduce customer satisfaction. This project develops a predictive targeting approach that uses historical customer and campaign data to estimate term deposit subscription likelihood, rank customers, and support more efficient allocation of outreach resources.

Using the UCI Bank Marketing dataset (45,211 observations; May 2008–Nov 2010) with an imbalanced outcome (11.7% “yes”), multiple supervised learning classifiers were evaluated under an 80/20 stratified train–test split and cross-validation. Models included Logistic Regression (baseline), Lasso Logistic Regression (feature selection), Random Forest, XGBoost, LightGBM, and a Neural Network. Overall, gradient boosting models delivered the strongest discrimination performance (LightGBM ROC-AUC = 0.806; XGBoost ROC-AUC = 0.805), while Random Forest achieved the most balanced precision–recall trade-off (F1 = 0.48; precision = 0.43). Logistic Regression and Lasso provided interpretable baselines (ROC-AUC ≈ 0.772–0.773) but produced many false positives at default thresholds (precision = 0.27). The Neural Network achieved high accuracy and precision (accuracy = 0.89; precision = 0.61) but low recall (0.24), indicating it is overly conservative without threshold tuning.

To translate model outputs into operational decisions, SHAP explanations from the LightGBM model were used to identify practical drivers of predicted subscription. Prior campaign success and higher account balance consistently increase predicted propensity, while repeated contact attempts during the same campaign (high campaign) are associated with lower predicted conversion, suggesting diminishing returns and supporting the use of contact caps or lower-cost follow-ups. The “contact_unknown” indicator is highly influential and negative, likely reflecting data quality and incomplete channel logging, highlighting an opportunity to improve CRM capture for more reliable targeting. Month effects appear in the model but are treated as contextual signals rather than fixed seasonality due to sensitivity to changing economic conditions and campaign design.

Overall, the findings support a deployable workflow in which customers are scored and ranked by predicted probability, outreach is prioritized within capacity constraints, and contact intensity rules are used to reduce wasted calls and customer fatigue. Future enhancements include threshold optimization aligned to campaign capacity and objectives, monitoring for performance drift, and incorporating additional economic indicators to strengthen robustness.

2. Introduction & Business Problem

2.1 Business Context

Direct marketing is still one of the most common acquisition strategies in retail banking, especially for term deposit products. These campaigns often rely on outbound calls, where agents contact existing or prospective customers to promote enrollment. Because term deposits are trust-based products, effective outreach requires good timing and relevance—meaning banks need to be strategic about *who* they contact.

However, banks face real operational constraints: call center capacity is limited, outbound campaigns are expensive, and contacting uninterested customers can harm customer experience. As a result, the central challenge is not whether marketing exists, but how to allocate outreach resources efficiently.

This is where predictive modeling becomes essential: instead of contacting customers broadly, banks can use historical customer and campaign data to estimate subscription likelihood, rank customers, and prioritize outreach toward those with the highest probability of converting.

2.2 Problem Statement

Traditional campaign targeting often depends on heuristic segmentation and broad contact lists, which can be costly and inefficient. While banks collect rich customer and campaign data, they frequently lack a systematic way to translate that data into actionable prioritization decisions.

To address inefficiencies in outreach and improve campaign ROI, we define the core problem as follows:

“We aim to help banks optimize term deposit campaigns by using predictive modeling to prioritize which customers to contact and to support more efficient, targeted outreach.”

This scope focuses on building models that:

- Estimate subscription likelihood using customer demographics, financial attributes, and campaign-related features
- Rank customers to support targeting decisions under limited operational capacity
- Help reduce wasted contacts and improve conversion efficiency while being mindful of customer experience

3. Data Source & Description

3.1 Data Source

The data used in this project comes from the UCI Bank Marketing Dataset, a publicly available dataset containing detailed records from a Portuguese bank’s direct marketing campaigns. These campaigns were conducted between May 2008 and November 2010, during which the bank repeatedly contacted customers to promote subscription to a term deposit product.

The dataset consists of 45, 211 records and 17 variables as in Table 1. Each row represents an individual customer contacted during a campaign, capturing demographic attributes, financial standing, and prior engagement history at the time of contact. The target variable is y, indicating whether the customer subscribed to the term deposit (yes/no).

Table 1. Variable Descriptions

Variable	Description
Age	Customer age
Job	Job type / employment category
Marital	Marital status
Education	Education level (e.g., basic, high school, university)
Default	Whether the customer has credit in default (yes/no)
Balance	Average annual account balance
Housing	Whether the customer has a housing loan (yes/no)
Loan	Whether the customer has a personal loan (yes/no)
Contact	Contact communication type (e.g., cellular, telephone)
Day	Day of the month of the last contact
Month	Month of the last contact
Duration	Duration of the last contact (seconds)
Campaign	Number of contacts made with the customer during the current campaign
Pdays	Days since the customer was last contacted in a previous campaign
Previous	Number of contacts made with the customer before the current campaign
Poutcome	Outcome of the previous marketing campaign
Target	Whether the customer subscribed to a term deposit (yes/no)

* Modeling note: *duration* is not known before the call happens and is strongly tied to the outcome (if duration = 0 then y = “no”). For a predictive model used before outreach, this should be excluded.

3.2 Exploratory Data Analysis (EDA)

To understand the structure of the dataset and identify early patterns in customer behavior, we conducted a detailed exploratory analysis. The dataset includes 7 numerical variables (Age, Balance, Day, Duration,

Campaign, Pdays, Previous) and 10 categorical variables (Job, Marital, Education, Default, Housing, Loan, Contact, Month, Poutcome, Target).

Figure 1. Distribution of Target Variable

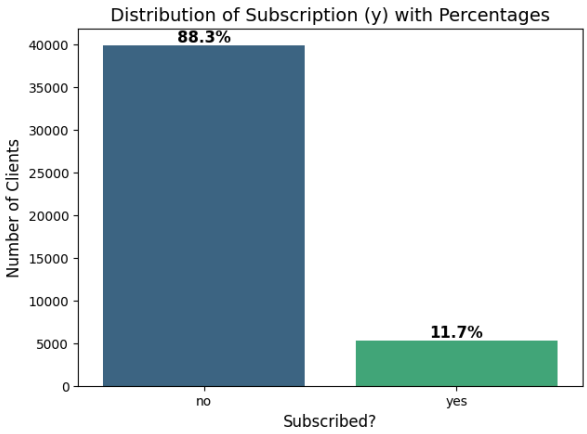


Figure 1 shows that the target variable y is highly imbalanced. Approximately 88.3% of clients did not subscribe to a term deposit ("no"), while only 11.7% subscribed ("yes"). This indicates that positive outcomes are relatively rare in the dataset. As a result, overall accuracy alone may be misleading, and model evaluation should emphasize imbalance-aware metrics (e.g., ROC-AUC, precision/recall, F1-score, and confusion matrix), along with thoughtful threshold selection to align with campaign objectives.

Figure 2. Correlation Heatmap of Numeric Features

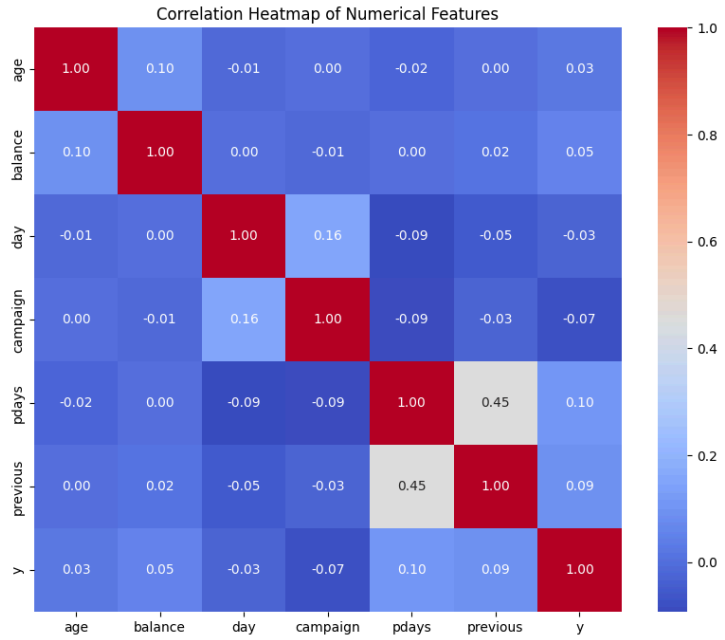


Figure 2 suggests limited multicollinearity among most numerical variables, with pairwise correlations generally close to zero. The main exception is a moderate positive correlation between pdays and previous ($r \approx 0.45$), which is expected because both variables reflect prior contact history and campaign engagement.

A few other relationships are weak but noticeable, such as day and campaign ($r \approx 0.16$) and age and balance ($r \approx 0.10$), indicating only minor linear association.

With respect to the target y, linear correlations are small across all numerical features (e.g., pdays ≈ 0.10, previous ≈ 0.09, balance ≈ 0.05, campaign ≈ −0.07). This implies that no single numeric variable is strongly predictive in a purely linear sense.

4. Data Preprocessing

4.1 Handling Missing Values

Based on our initial exploration, the dataset contains no true missing values; however, several categorical fields include a substantial number of entries labeled “unknown,” indicating incomplete or unspecified information rather than blank records. The largest concentrations are poutcome (81.75%) and contact (28.80%), followed by education (4.11%) and job (0.64%). To preserve sample size and avoid potential bias from deletion, we kept all rows and treated “unknown” as a valid category during preprocessing.

4.2 Variable Distribution and Outlier Checks

We observed that age, balance, and pdays exhibit skewness. We did not apply transformations (e.g., log) because tree-based models can naturally handle skewed distributions, and linear models benefit primarily from standardization.

Outliers in balance were intentionally retained, as high-balance customers may represent important business cases rather than data errors. Removing these observations could reduce the model’s relevance for high-value targeting scenarios.

4.3 Encoding Categorical Variables

Because many predictive models require numerical inputs, categorical variables were converted into numeric form as follows:

- One-hot encoding for nominal variables such as job, marital, education, contact, month, and poutcome.
- Binary encoding for yes/no variables such as housing, loan, and default, which were mapped to 1/0.

These transformations ensure compatibility with models including logistic regression and tree-based methods, while preserving interpretability of categorical effects.

4.4 Feature Standardized

For models sensitive to feature scale, especially logistic regression and other regularized linear methods, we standardized continuous variables (e.g., age, balance, pdays, and other numeric indicators when included).

Tree-based models (e.g., Random Forest, XGBoost) were trained on unscaled features, consistent with best practices since these methods are generally scale-invariant.

4.5 Train/Test Split

We split the dataset into training and test sets using an 80/20 split. To ensure the class distribution of the imbalanced target variable was preserved in both sets, we applied stratified sampling.

5. Machine Learning Modeling

We apply supervised learning models to predict whether a customer will subscribe to a term deposit and compare their performance on a held out test set. Specifically, we evaluate classification algorithms: Logistic Regression, Random Forest, XGBoost, LightGBM, and Neural Networks. These methods span a range of model complexity, from interpretable linear decision boundaries to more flexible non linear approaches, allowing us to assess predictive performance and practical trade offs.

5.1 Logistic Regression

Logistic regression was selected as the baseline model because it is a well established method for binary classification, produces calibrated probability estimates, and offers high interpretability through its coefficients. This makes it a strong reference point for evaluating whether more complex models provide meaningful performance improvements.

L1 regularization (Lasso) was then applied to logistic regression to reduce overfitting and improve model parsimony by shrinking less informative coefficients to exactly zero. This feature selection property is particularly useful in the presence of many one hot encoded categorical variables, as it helps identify the most predictive variables while simplifying the final model.

5.2 Random Forest

Random Forest is well suited for this task because it can capture non linear relationships and feature interactions, performs well on data with mixed variable types, and does not require feature scaling. Despite these advantages, Random Forest has several limitations. It is generally **l**ess interpretable than logistic regression or a single decision tree, and common feature importance measures can be biased toward variables with more potential split points.

5.3 XGBoost

XGBoost can also capture non linear relationships and complex feature interactions, handle mixed feature types (after encoding), and is generally robust to skewed numerical distributions without requiring feature scaling. However, it is sensitive to hyperparameter choices and can overfit if not properly tuned, especially with many boosting rounds or deep trees.

5.4 LightGBM

Compared with traditional gradient boosting implementations, LightGBM is designed for efficiency and scalability through optimized tree growing and histogram based split finding, which often results in faster training and strong performance on large datasets. Like XGBoost, it is also sensitive to hyperparameter settings (e.g., learning rate, number of leaves, max depth), and improper tuning can lead to overfitting. In addition, when many categorical variables are one hot encoded, the feature space can become high dimensional, which may increase training complexity and make careful regularization more important.

5.5 Neural Networks

Neural Networks are flexible non linear models that learn complex relationships between features and the subscription outcome through multiple layers of weighted transformations, making them capable of capturing interactions and non linear patterns that may be missed by simpler models. However, neural networks are generally less interpretable than logistic regression, and can be sensitive to architecture and training hyperparameters

6. Model Results

6.1 Baseline Logistic Regression

The dataset was split into training and test sets using an 80/20 stratified split to preserve the class distribution of the target variable. Continuous variables (age, balance, day, campaign, pdays, previous) were standardized using z-score normalization, with the scaler fit on the training set and applied to the test set to prevent data leakage. A baseline logistic regression classifier was then trained with class_weight = “balanced” to account for class imbalance.

6.2 Lasso Logistic Regression (Feature selection)

To reduce model complexity and perform embedded feature selection, an L1-regularized logistic regression (Lasso) was trained using -fold cross-validation. The inverse regularization strength C was selected by maximizing ROC-AUC on the cross-validation folds. After fitting the final model, features with non-zero coefficients were retained as selected predictors, while coefficients shrunk to zero were treated as dropped features.

6.3 Tree-Based Model (Hyperparameter Tuning)

For tree-based models (Random Forest, XGBoost, and LightGBM), hyperparameter tuning was conducted using 3-fold stratified cross-validation to improve generalization and reduce overfitting. Key parameters such as tree depth, number of trees/boosting rounds, learning rate, and sampling-related settings were tuned using ROC-AUC as the primary selection metric. The best-performing configuration from cross-validation was then refitted on the full training set and evaluated on the held-out test set for final reporting.

6.4 5 Fold Cross Validation

5-fold cross validation was conducted during model development to provide a more reliable estimate of out of sample performance and to reduce sensitivity to any single train test split, thereby strengthening confidence in model generalization.

Table 2. Model Performance Summary

Model	ROC-AUC	Accuracy	Precision (pos)	Recall (pos)	F1 (pos)
Logistic Regression	0.772	0.76	0.27	0.62	0.37
Lasso Logistic	0.773	0.75	0.27	0.64	0.38

Regression					
Random Forest	0.802	0.86	0.43	0.55	0.48
XGBoost	0.805	0.82	0.35	0.65	0.45
LightGBM	0.806	0.82	0.35	0.65	0.45
Neural Networks	0.788	0.89	0.61	0.24	0.24

Figure 3. ROC Curves

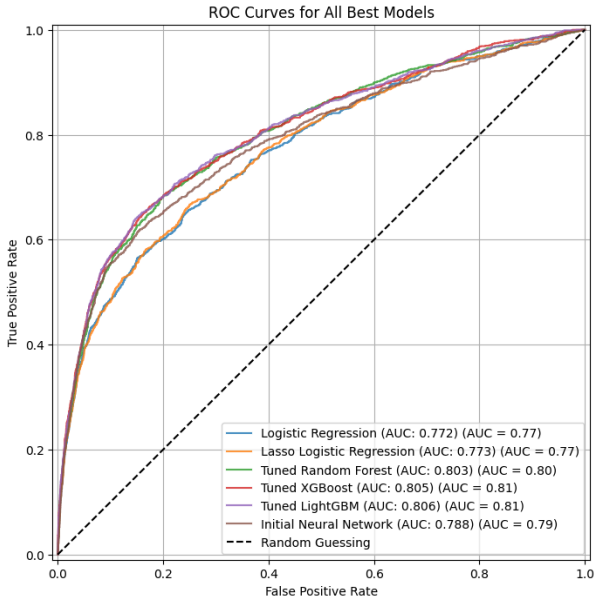


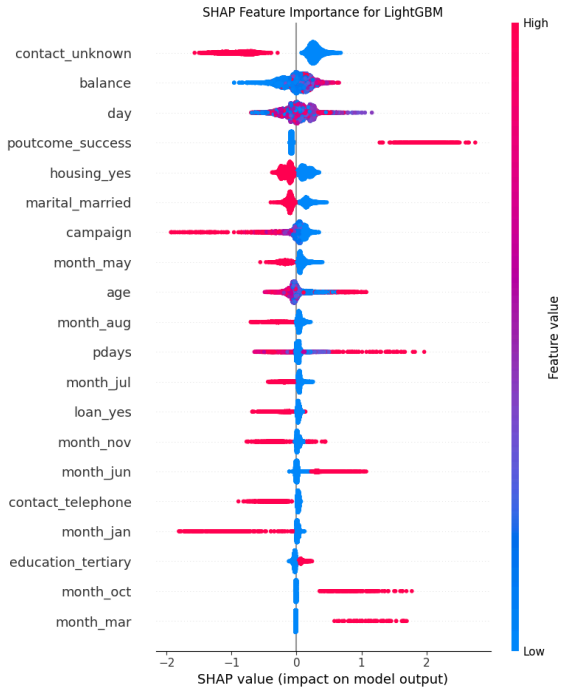
Table 2 shows that model choice materially changes the precision–recall trade-off, which matters more than accuracy in this highly imbalanced setting. Logistic Regression and Lasso perform similarly in ranking ability (ROC-AUC ≈ 0.772–0.773) and achieve relatively high recall (0.62–0.64), but their low positive precision (0.27) indicates many predicted subscribers would be false positives, reducing outreach efficiency. Tree-based methods provide the most balanced performance: Random Forest improves ROC-AUC to 0.802 and delivers the best F1-score (0.48), driven by the highest precision among non-neural models (0.43), making it attractive when minimizing wasted contacts is important. Gradient boosting models (XGBoost and LightGBM) achieve the strongest overall discrimination (ROC-AUC ≈ 0.805–0.806) and the highest recall (0.65), making them well-suited for campaigns focused on capturing more potential subscribers, even if precision is moderate (0.35). The neural network displays the highest accuracy (0.89) and very high precision (0.61), but extremely low recall (0.24), meaning it identifies only a small fraction of actual subscribers; this suggests the default threshold yields an overly conservative classifier that may be better used only when outreach capacity is very limited or after threshold tuning.

6. Business Implementations

The SHAP analysis highlights which customer and campaign features most strongly influence the LightGBM model's subscription predictions. These insights can be operationalized to improve campaign

efficiency by prioritizing outreach toward higher-probability customers and reducing wasted contacts. contact_unknown is the top feature and strongly reduces predicted subscription. From a business standpoint, this likely reflects missing channel logging or inconsistent campaign recording rather than customer preference.

Figure 4. SHAP Feature Importance for LightGBM



6.1 Targeting and Prioritization Strategy

Use the model's predicted probability as a ranking score and contact customers in descending order until the outreach capacity (agents, call hours, budget) is reached.

- High-priority segment: customers with strong positive drivers such as poutcome_success (previous campaign success) and higher balance, since these features consistently push predictions toward subscription.
- Lower-priority segment: customers with features that push predictions downward, such as housing_yes, and loan_yes, unless there is sufficient capacity or a strategic reason to include them.

6.2 Re-contact Policy and Contact Intensity Control

The SHAP results show that higher values of campaign (more contacts during the current campaign) tend to decrease predicted subscription likelihood, suggesting diminishing returns for repeated contacts.

- Set a soft cap on the number of contact attempts per customer during a campaign (e.g., stop after a defined number of unsuccessful attempts).
- For customers already contacted multiple times (high campaign), shift strategy from repeated calls to lower-cost follow-ups (e.g., email/SMS) if available.

6.3 Timing Effects

Month indicators show up as meaningful drivers in the SHAP results. For example, customers contacted in March and October tend to receive positive SHAP contributions, while January, May, and August more often push predictions downward. However, these patterns should not be treated as fixed seasonality. Month effects can also reflect changing economic conditions, product offers, or campaign execution differences across time. For this reason, month is best used as a contextual feature in the model, while operational decisions should rely primarily on the predicted score and be monitored over time for drift.

7. Conclusion

This project demonstrates how predictive modeling can meaningfully improve the efficiency of direct marketing campaigns in retail banking. Using historical campaign and customer data, multiple supervised learning models were developed and compared to estimate the probability that a customer will subscribe to a term deposit.

Across models, tree-based methods such as XGBoost and LightGBM achieved the strongest predictive performance, outperforming baseline logistic regression in ROC-AUC while maintaining reasonable precision–recall trade-offs under class imbalance. At the same time, logistic regression and Lasso provided valuable interpretability and feature selection benefits, serving as strong benchmarks and offering insight into key subscription drivers. The use of stratified train/test splits, cross-validation, and hyperparameter tuning strengthened confidence in model generalization and robustness.

From a business perspective, the results highlight the importance of prioritization rather than broad outreach. Instead of contacting customers uniformly, banks can use predicted probabilities to rank customers and allocate limited call center capacity toward those most likely to convert. Feature insights from SHAP analysis further support practical strategies such as emphasizing customers with prior campaign success, managing contact frequency to avoid diminishing returns, and monitoring timing effects carefully rather than relying on fixed seasonal assumptions.

Overall, this study shows that data-driven targeting can increase conversion efficiency, reduce unnecessary outreach, and improve campaign ROI while maintaining customer experience. Future work could incorporate additional economic indicators, dynamic threshold optimization based on campaign capacity, and real-time deployment considerations to further enhance operational impact.