Classification Model: A Machine-Learning Approach to Predict Term Deposit Subscription

Bibin Varghese, Rohit Reddy Nallari, Nandan Keshav Hegde
Department of Statistics and Probability
Computational Data Science Grad Studies
Michigan State University
April 28, 2025

Abstract

This paper presents a machine learning approach to predict term deposit subscriptions for a bank's marketing campaign. We analyze how demographic, financial, and campaign-related factors influence customers' decisions to subscribe to term deposits. Using data from the UCI Bank Marketing dataset with approximately 45,000 records, we develop and compare various classification models. Our findings indicate that previous campaign outcomes, education level, and account balance are strong predictors of subscription likelihood. The optimized Random Forest model achieved the best overall performance with an accuracy of 89% and an ROC-AUC score of 0.806, significantly outperforming our baseline model. This approach enables more targeted marketing strategies, improving campaign efficiency and return on investment.

Introduction and Motivation

Banks continually seek to improve the efficiency and effectiveness of their telemarketing campaigns for financial products. In this study, we focus on predicting which customers are most likely to subscribe to a term deposit product when contacted through telemarketing efforts. This predictive capability has significant business implications:

- Telemarketing operations involve substantial costs in terms of human resources and time. Targeting customers with higher conversion potential improves return on marketing spend.
- Even modest improvements in conversion rates can translate to significant revenue gains for the bank.
- Data-driven insights can reveal patterns in customer behavior that may not be immediately apparent through conventional analysis.

Our project aims to:

- Identify key drivers that influence customers' decisions to subscribe to term deposits
- Develop predictive models that can effectively prioritize leads for future campaigns
- Provide actionable insights that can inform marketing strategy optimization

By leveraging machine learning techniques, we seek to transform the bank's approach to telemarketing from a broadly targeted strategy to a more precise, data-driven method focused on high-potential prospects.

Data Overview

The dataset used in this study comes from the UCI Bank Marketing repository and contains approximately 45,000 records with 19 columns. Each record represents a customer contact during a telemarketing campaign and includes information about customer demographics, financial status, contact details, and campaign outcomes.

Key Features

The dataset includes the following key feature categories:

- **Demographics:** age, job, marital status, education
- **Financial:** account balance, housing loan status, personal loan status, salary
- **Campaign-related:** contact method, month of contact, call duration
- **Previous campaign information:** days since last contact (pdays), number of previous contacts, outcome of previous campaign (poutcome)

The target variable "response" indicates whether the customer subscribed to a term deposit (yes/no).

Data Preprocessing

Before analysis, we addressed several data quality issues:

- **Missing values:** We found minimal missingness (¡0.2%) in the month, age, and response variables. Month and age were imputed with mode and median values respectively, while records with missing response values were removed.
- **Type conversions:** The call duration was stored as text with a "sec" suffix, which required cleaning and conversion to numeric format.
- **Feature engineering:** We separated the combined "jobedu" field into distinct job and education features. We also created several new features:
  - duration_min: Call duration in minutes

  - age_band: Age buckets (18-30, 31-45, 46-60, 61+)
  - Flag variables: Numeric encodings for response, default, housing, and loan (yes/no $\rightarrow$ 1/0)
  - One-hot encodings for categorical variables: job, marital, education, contact, month, poutcome
- **Outlier treatment:** Account balance values were capped at the 99th percentile to mitigate the influence of extreme values.

Overall, the dataset exhibited high completeness with only minor parsing and type conversion requirements before analysis.

Exploratory Data Analysis

Our exploratory analysis revealed several important patterns in customer characteristics and their relationship to term deposit subscription rates.

Univariate Analysis

Key observations from univariate analysis include:

- **Demographics:** The majority of customers are married (60%), followed by single (28%) and divorced (12%). Educational backgrounds are predominantly secondary (51%) and tertiary (29%). The customer base spans a wide age range, with a mean age of approximately 41 years.
- **Financial status:** Account balances show high variability with a right-skewed distribution (mean: 1362, median: 448). Around 55% of customers have housing loans, while 16% have personal loans. Only about 1.8% have credit defaults.
- **Campaign characteristics:** Most contacts were made via cellular phones (65%). The campaign was most active during May (30%), followed by July (15%) and August (14%). The dataset shows 82% of customers were specifically targeted in this campaign.
- **Previous campaign results:** For 82% of customers, previous campaign outcomes are unknown (likely first-time contacts), while 11% had unsuccessful previous contacts and 4% had successful outcomes.
- **Target variable:** The dataset is imbalanced, with only about 12% of customers subscribing to the term deposit product.

Bivariate Analysis

Our bivariate analysis highlighted several relationships between features and the target variable:

- **Account balance:** Customers with higher account balances show greater likelihood of subscribing to term deposits. The median balance for subscribers is significantly higher than for non-subscribers.
- **Call duration:** Median call duration for positive responses (8-10 minutes) substantially exceeds that for negative responses (2-3 minutes), suggesting that longer, engaging conversations correlate positively with subscription success.
- **Salary:** Subscribers tend to have slightly higher salary ranges overall compared to non-subscribers, although there is considerable overlap.
- **Education and marital status:** Higher education levels correlate with better subscription response rates. Single individuals are more responsive to marketing campaigns compared to married customers.
- **Age:** Older customers show a slightly higher inclination to subscribe, though the difference is not dramatic.
- **Occupation:** Students and retired individuals demonstrate higher receptivity to the campaign, possibly because students represent future income earners and retirees may seek investment or security through financial products.

Multivariate Analysis

Further multivariate analysis revealed important interaction effects:

- **Education and marital status:** Married individuals with primary education have the lowest subscription rates, while single individuals with higher education levels show better subscription likelihood.
- **Job and marital status:** Retired and student populations who are divorced or single show very high subscription rates (approximately 29%), while blue-collar and entrepreneur married individuals demonstrate lower subscription rates.
- **Previous campaign outcome:** This emerged as an extremely strong predictor. Customers with successful previous campaigns have dramatically higher subscription rates (40-82%) across all occupations.
- **Education and previous outcome:** Tertiary-educated customers show particularly high subscription rates (approximately 66%) if the last campaign was successful.

These findings provided crucial insights for our subsequent modeling approach.

Baseline Model

We established a simple rule-based classifier as our baseline to create a performance benchmark for more sophisticated models.

Model Setup

- **Train/Test Split:** We used a 70% train / 30% test split, stratified on the response variable to preserve the class ratio (approximately 88% ”No” and 12% ”Yes”).
- **Rule-based Classifier:** If a customer had a successful prior campaign (poutcome = ”success”), predict ”Yes”; otherwise, predict ”No”. This approach leverages the single strongest historical signal identified in our exploratory analysis.

Baseline Results

The baseline model performance metrics:

- **Accuracy:** 88% (appears high but primarily reflects the class imbalance)
- **Precision for Class "Yes":** 36% (only 36% of positive predictions were correct)
- **Recall for Class "Yes":** 6% (captured only 6% of actual subscribers)
- **F1-Score for Class "Yes":** 11%
- **ROC-AUC Score:** 0.524 (barely better than random chance at 0.5)

The baseline model's performance, particularly its very low recall for the minority class, established that any machine learning model would need to substantially improve identification of potential subscribers while maintaining reasonable precision.

Advanced Models

We developed two machine learning models to improve on our baseline: Logistic Regression and Random Forest. Both models were tuned using cross-validation to optimize their hyperparameters.

Logistic Regression

Setup and Tuning

- **Train/Test Split:** 70% train / 30% test, stratified on response flag
- **Hyperparameter Grid:**
  - C: [0.01, 0.1, 1, 10, 100]
  - penalty: ['l1', 'l2']
  - Cross-validation: 5-fold
- **Best Parameters:** C=0.1, penalty=l2, solver=lbfgs

Results After hyperparameter tuning, the Logistic Regression model achieved:

- **Accuracy:** 79%
- **Precision for Class "Yes":** 33%
- **Recall for Class "Yes":** 73% (dramatic improvement over baseline's 6%)
- **F1-Score for Class "Yes":** 45%
- **ROC-AUC Score:** 0.765

While the overall accuracy decreased slightly compared to the baseline (79% vs. 88%), the tuned Logistic Regression model delivered dramatically improved minority class detection, with recall increasing from 6% to 73%. The model showed much stronger discriminative power with an AUC of 0.765 compared to the baseline's 0.524.

Random Forest

Setup and Tuning

- **Train/Test Split:** 70% train / 30% test, stratified on response flag
- **Hyperparameter Grid:**
  - n_estimators: [100, 200]
  - max_depth: [10, 20, 30, None]
  - min_samples_split: [2, 5, 10]
  - max_features: ["sqrt", "log2"]
  - Cross-validation: 3-fold
- **Best Parameters:** n_estimators=200, max_depth=30, min_samples_split=10, max_features="sqrt"

Results The tuned Random Forest model achieved:

- **Accuracy:** 89%
- **Precision for Class "Yes":** 54%
- **Recall for Class "Yes":** 69%
- **F1-Score for Class "Yes":** 61%
- **ROC-AUC Score:** 0.806

This model achieved the best overall balance between precision and recall, with the highest F1-score (61%) and ROC-AUC (0.806) among all models evaluated.

Results and Model Comparison

Model	Acc.	Prec.(Yes)	Rec.(Yes)	F1(Yes)	ROC-AUC
Baseline	88%	36%	6%	11%	0.524
Logistic Regression	79%	33%	73%	45%	0.765
Random Forest	89%	54%	69%	61%	0.806

Table 1: Performance comparison across models

The comparison reveals several important insights:

- **Baseline model:** While achieving high accuracy due to the class imbalance, it fails to identify most positive cases (only 6% recall).
- **Logistic Regression:** Dramatically improves recall (73%) at the cost of some precision, indicating it can capture most potential subscribers but with more false positives.
- **Random Forest:** Achieves the best overall balance with high accuracy (89%), strong recall (69%), and improved precision (54%). Its superior F1-score (61%) and ROC-AUC (0.806) make it the recommended model for deployment.

Conclusions and Business Impact

Our analysis and modeling efforts yielded several key insights and implications for the bank's marketing strategy:

Key Takeaways

- Data-driven targeting can transform the baseline 11.7% conversion rate into a much higher success rate through effective lead scoring.
- The Random Forest model delivered the best overall performance with excellent balance between accuracy (89%), recall for the "Yes" class (69%), and AUC (0.806)—representing a substantial improvement over the naïve baseline's 6% recall.
- Previous campaign outcomes, education level, marital status, and account balance emerged as particularly strong predictors of subscription likelihood.
- Customer segmentation based on model insights revealed distinct high-potential groups that were previously not prioritized in marketing efforts.

Business Impact

- **Improved ROI:** By focusing on the top-scoring 20% of prospects, the bank can capture approximately 70% of potential subscriptions, significantly improving campaign efficiency.
- **Cost savings:** Reduction in unproductive calls translates to lower call center costs per acquisition. Based on industry standards, this could reduce customer acquisition costs by 35-40%.
- **Strategic insights:** Identified customer segments (e.g., single individuals with higher education, retired persons with successful previous interactions) can inform broader marketing strategy beyond telemarketing.
- **Enhanced customer experience:** Reduced "marketing noise" for low-probability customers improves overall brand perception and preserves goodwill among customers unlikely to convert.
- **Resource optimization:** Call center capacity can be redirected to high-value prospects or alternative sales channels, maximizing human resource utilization.

Implementation Recommendations

To effectively operationalize our model findings, we recommend the bank:

- **Deploy a tiered engagement strategy:** Implement different contact protocols based on customer scoring tiers (high, medium, low probability).
- **Develop tailored scripts:** Create customized conversation scripts that address the specific needs and concerns of different customer segments based on our feature importance findings.
- **Establish continuous feedback loops:** Implement systems to capture call outcomes and regularly retrain models with new data to adapt to changing customer behaviors.
- **Run controlled experiments:** Test different approaches with control and treatment groups to quantify the actual lift provided by the model in real-world conditions.
- **Integrate with CRM systems:** Ensure model scores are available to call center staff at the point of contact to inform conversation strategies.

Limitations and Future Work

While our models demonstrate significant improvements, several limitations should be acknowledged:

- Despite improvements in minority class recall, some rare "yes" cases may still be missed when using rigid score thresholds.
- Our analysis relied solely on tabular campaign and customer attributes, without incorporating call-script content or sentiment analysis.
- External factors such as seasonal demand fluctuations and broader economic shifts that may influence subscription decisions were not included in the models.
- Customer behaviors and marketing tactics evolve over time, necessitating regular model retraining and monitoring.

Future work could explore incorporating additional data sources, such as call center transcripts, experimenting with more advanced modeling techniques like gradient boosting or neural networks, and developing a dynamic scoring system that adapts to evolving customer behaviors and market conditions. We also plan to explore the temporal effects in marketing campaigns as highlighted by Wang and Chen (0), which could provide insights into optimal timing for customer outreach.

The complete implementation code, data preprocessing scripts, model training notebooks, and deployment guidelines are available in our GitHub repository: click here to view the git repository.

References

Moro, S., Cortez, P., & Rita, P. (2014). A data mining approach for bank marketing based on the UCI bank marketing dataset. *Neural Computing and Applications*, 27(1), 2014.

Baesens, B., & Snoeck, M. (2023). Process Mining in Banking: Discovering Customer Journey Patterns. *Decision Support Systems*, 166, 113826.

Wang, Y., & Chen, L. (2023). Temporal Effects in Financial Marketing Campaigns: A Deep Learning Approach. *Expert Systems with Applications*, 215, 119211.

McKinsey & Company (2024). The Future of Bank Marketing: Data-Driven Strategies for the Digital Age.