DSCI 5260.002

Success Prediction of Bank Telemarketing

A Pre-Call Customer Ranking System Using Machine Learning

Business Analytics Capstone - Final Report

By
Siri Chandana Byreddy | Rakshitha Chattanahalli Mahesh | Hari Krupa Cheguri
Kavya Sree Chittaboina | Jeevan Deep Borugadda

University of North Texas | G. Brint Ryan College of Business
December 2025

Table of Contents

1. Executive Summary

1.1 Overview

The following project deals with an operational problem of a Portuguese banking institution: the inefficiency of the call-based marketing campaign regarding its term deposit products. An almost 89% rejection rate wastes a considerable amount of resources, increases operational costs, and may cause dissatisfaction on both the customer and employee sides because of undesirable and ineffective contacts.

This is a deployable pre-call customer ranking system using machine learning to predict the probability of subscription to a term deposit by the customer based on only pre-call information. The crucial constraint of forbidding any post-call data, for example, call duration, makes the system capable of real-time queuing for the outbound calls effectively. Finally, this system shifts the bank's telemarketing strategy from generalized mass-calling to precision-targeted and data-driven prioritization of its customers.

The project incorporates CRISP-DM methodology and covers all the phases of Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment planning. In this project, the UCI Bank Marketing dataset will be used, which has 41,188 customer data points collected from 2008 - 2013. Pre-processing, ensuring no data leakage while also trying to preserve the integrity of the predictive signals, will take place.

1.2 Key Findings

Our comprehensive analysis yielded several significant findings that directly inform operational strategy:

- **Model Performance Excellence:** 8 machine learning models were developed and tested under leakage free(pre-call) conditions. Neural Network Model came top with the best ROC-AUC of 0.810, closely followed by Random Forest at 0.802, XGBoost at 0.800, and Logistic Regression at 0.801. Rolling window evaluation resulted in ROC-AUC of 0.795, matching the original research paper's score of 0.794.
- **Efficiency Gains Quantified:** Calling only the top 10% (c. 4,119 contacts) model-ranked customers would allow the bank to capture 45-48% of all actual subscribers—this is a 4.5-4.8× improvement over random calling. Extending to the top 20% captures 65-72% of subscribers while contacting only one-fifth of the customer base.
- **Critical Predictive Drivers:** Past campaign response proved to be the best predictor, where previous subscribers had a 65.1% conversion rate versus 8.8% for first-time contacts. This is a difference of 7.4×. Furthermore, the following suggestions are presented for actionable targeting: macroeconomic conditions, Euribor rate correlation of -0.80; contact channel, cellular 3× more effective than telephone; and seasonal timing, March/September/December.
- **Leakage Quantification:** Our novel contribution is explicit quantification of the impact due to data leakage. Inclusion of call duration artificially inflates ROC-AUC by 14.3 percentage points (0.801→0.944 for Logistic Regression), illustrating why features of the post-call must be excluded for deployable models.
- **Temporal Robustness Validated:** Models trained on 2008-2013 data remain resilient in performance when pitted against modern macroeconomic conditions of 2023-2025, while degradation in AUC is less than 3-5%, confirming deployment viability across economic cycles.

1.3 Recommendations

Based on our comprehensive analysis, we recommend the following strategic actions:

- **Deploy Neural Network or Logistic Regression:** For production deployment, either the Neural Network model (AUC = 0.810) or the Logistic Regression model (AUC = 0.801) is recommended as the scoring solution. The marginal performance improvement of 0.009 AUC offered by the Neural Network may not justify the reduced interpretability, particularly in regulated banking contexts.
- **Implement Tiered Targeting Strategy:** Tier 1 (top 10%, approximately 49% capture), Tier 2 (10-20%, approximately 23% additional capture), and Tier 3 (20-50%, with diminishing returns). Assign premium agents to Tier 1 prospects to maximize conversion rates.
- **Optimize Channel Allocation:** As there is a 3× effectiveness advantage, shift the contact mix towards 85%+ cellular from the current 63%. Eliminate landline-only contacts or position them for lower-priority tiers.

- **Seasonal Campaign Timing:** Concentrate campaign intensity in Q1 (January-March) and Q4 (September-December) when conversion rates reach their peak at 40-50%; decrease budgets during summer (10-15% rates).
- **Prioritize Relationship Customers:** Apply dedicated re-engagement programs for previous subscribers given their 65.1% conversion rate, which is 7× higher than cold prospects.

1.4 Expected Business Impact

Implementation of the recommended pre-call ranking system is projected to deliver substantial operational and financial benefits:

- **Cost Reduction:** An estimated savings of €165,000 for every 41,000-contact campaign targeting only the top 20% of ranked customers while maintaining 65-70% of the total conversions.
- **Efficiency Improvement:** 4.5-4.8× lift in conversion rate for prioritized contacts, reducing cost per acquisition from €44 (random) to €13 (targeted)—a 70% reduction.
- **Resource Optimization:** Agent time was refocused from low-probability cold calls to high-value prospect conversations, improving productivity metrics and job satisfaction simultaneously.
- **Customer Experience:** The reduction of unwanted contact attempts for unlikely converters reduces complaint rates and shields brand reputation.
- **Scalability:** Framework applicable to other product campaigns - credit cards, loans, insurance - with minimal re-training requirements.

2. Business Understanding

2.1 Business Problem Definition

The core operational problem the Portuguese banking institution has to deal with is related to direct marketing operations. About 89% of term deposit campaigns via phone calls get rejected, and only 11.27% of the customers contacted subscribe to those. This huge inefficiency expresses itself along several dimensions of the organization:

2.1.1 Operational Impact

The current mass-calling process is extremely resource-intensive without comparable returns. Every outbound call, for example, has direct costs estimated at €5 per contact attempt due to agent compensation, telecommunications infrastructure, and system overhead. With a common campaign of 41,000 contacts, the total cost of one campaign reaches around €205,000, with only 4,621 subscriptions realized—the cost per acquisition totaling €44. More disturbingly, agents waste most of their time calling people who are not interested, resulting in agent fatigue, low morale, and turnover within the telemarketing team.

2.1.2 Customer Experience Degradation

Repeatedly contacting customers can damage relationships. Individuals unlikely to subscribe may get irritated, negatively affecting the bank's reputation and they might as well reduce future engagement with the bank. In an era of increasing privacy consciousness, excess unsolicited contact poses both reputational and compliance risks.

2.1.3 Strategic Misalignment

In the absence of data driven prioritization, high value prospects receive the same treatment as those who are unlikely to subscribe. High-potential customers may be contacted late in the campaign, after resources are depleted or may receive rushed calls as agents attempt to finish their assigned lists/targets. Such inefficient allocation of resources provides a competitive edge to organizations which employ data driven targeting strategies.

2.1.4 Problem Statement

The primary business challenge here is to identify prospects who are likely to subscribe prior to contact, enabling efficient resource allocation to maximize conversion rates as well as enhance customer experience. A key limitation is that predictions must rely only on data available before the call. Models that require inputs like call duration or conversation outcomes are not suitable for call prioritization or model training, as this information is only accessible after contact.

2.2 Project Objectives

This capstone project pursues five interconnected objectives aligned with the business problem:

2.2.1 Primary Objective: Develop a Pre-Call Ranking Model

The objective is to develop and validate machine learning models capable of scoring customers based on their probability of subscription, utilizing only pre-contact information. These models must demonstrate sufficient discrimination for effective prioritization and maintain acceptable calibration for threshold-based decision-making.

2.2.2 Identify Key Predictive Drivers

Leverage feature importance from multiple modelling approaches to find demographic, behavioral, and economic factors that are most strongly associated with the likelihood of subscription. These insights will both inform model development and wider marketing strategies.

2.2.3 Establish Baseline and Benchmark Performance

Repeat the key research of Moro et al. (2014) and extend it to set strong performance benchmarks. Measure how much data leakage occurs when including call duration to show that only a leakage-free approach, even if giving lower headline results, is valid for making deployment decisions.

2.2.4 Validate Temporal Robustness

Check the stability of the model under different economic conditions by comparing its performance using both past and recent economic indicators, specifically from 2008-2013 and 2023-2025. This will address concerns about the model's performance under changing economic conditions.

2.2.5 Deliver Actionable Business Recommendations

Technical modeling results will be translated into actionable business recommendations, including targeting thresholds, channel selection, campaign timing, and message development tailored to each customer segment.

2.3 Success Criteria

Project success is evaluated against technical, business, and operational criteria:

2.3.1 Technical Success Criteria

- **Discrimination:** ROC-AUC ≥ 0.75 on holdout test data, indicating meaningful separation between subscribers and non-subscribers.
- **Ranking Quality:** PR-AUC ≥ 0.40, confirming that the model is effective in ranking positive cases higher despite severe class imbalance.
- **Calibration:** Brier Score ≤ 0.15, meaning that the predicted probabilities reflect observed frequencies for threshold-based decisions.
- **Reproducibility:** All results are reproducible with documented random seeds, folds for cross-validation, and preprocessing steps.

2.3.2 Business Success Criteria

- **Lift Performance:** Top 10% of ranked customers must capture ≥40% of actual subscribers (Lift ≥ 4.0×).
- **Efficiency Threshold:** The top 20% of ranked customers should capture ≥60% of actual subscribers.
- **Cost Reduction:** Proven potential to reduce the cost per acquisition by ≥50% via targeted calling.

2.3.3 Operational Success Criteria

- **Leakage-Free Design:** All models should exclude post-call information, especially call duration.
- **Interpretability:** At least one of the candidate models should supply interpretable feature contributions for regulatory compliance.
- **Deployment Readiness:** Feature definitions, preprocessing steps, and scoring procedures are fully documented in a manner suitable for production deployment.

2.4 Constraints and Assumptions

2.4.1 Data Constraints

- **Temporal Scope:** Primary dataset spans 2008 to 2013, covering significant economic disruption—the European sovereign debt crisis. Performance of models should be validated for environmental scenarios that are very different.
- **Geographic Limitation:** The data is from one Portuguese institution, and generalizing to other markets should be cautiously considered.
- **Feature Availability:** UCI dataset is a subset of the original 150 features; some of the predictive signals available to the original researchers are not available.

2.4.2 Methodological Assumptions

- **Leakage Prevention:** Call duration is excluded, as this is post-call information unavailable at prediction time.
- **Unknown Value Treatment:** 'Unknown' responses in categorical variables are treated as informative categories, rather than missing data, because there were clearly non-random patterns observed.
- **Class Imbalance Handling:** A positive rate of 11.27% requires stratified sampling, weighting among classes, and proper evaluation metrics PR-AUC, F1 instead of accuracy only.
- **Economic Indicator Substitution:** Modern macroeconomic validation in 2023-2025 uses substitution of indicators, which tests for robustness but cannot fully validate temporal generalization without modern outcome labels.

3. Data Understanding

3.1 Data Collection and Sources

This analysis will use the Bank Marketing dataset taken from UCI Machine Learning Repository. Originally used in the published research paper by Moro, Cortez, and Rita in 2014, it documents direct marketing campaigns; details conducted by a Portuguese banking institution between May 2008 and November 2013, which captures call-based efforts to promote term deposit subscriptions.

3.1.1 Dataset Specifications

The dataset contains 41,188 customer contact records with 21 attributes spanning four conceptual domains:

Domain	Attributes
Client Demographics	age, job, marital, education, default, housing, loan
Campaign Contact	contact, month, day of week, duration, campaign, pdays, previous, poutcome
Economic Indicators	emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed
Target Variable	y (binary: yes/no subscription to term deposit)

3.1.2 Data Provenance and Quality

The foundation data set is a well documented, peer reviewed dataset from the UCI repository which has been highly studied in academic literature. However, many characteristics require consideration:

- **Temporal Context:** The 2008-2013 period was marked by high economic volatility, with the post-financial crisis of 2008 and the European sovereign debt crisis. This might have an effect on both customer behavior and macroeconomic variables.
- **Feature Subset:** While the original research used 150 features, reduced to 22 through feature selection, the UCI public dataset is an even further subset, suggesting that some predictive signals available to the original researchers are not available for our analysis.
- **Label Definition:** The target variable corresponds with an actual subscription outcome and not an intent or engagement, providing a tangible business-relevant outcome for the prediction.

3.2 Target Variable Analysis

The target variable exhibits severe class imbalance that fundamentally shapes our modeling approach:

Outcome	Count	Percentage
No (Rejection)	36,548	88.73%
Yes (Subscription)	4,640	11.27%
Total	**41,188**	**100.00%**

The approximately 1:8 imbalance ratio (11.27% positive rate) has significant implications:

- **Evaluation Metrics:** Accuracy is misleading; a naive classifier predicting 'No' for all cases achieves 88.73% accuracy. Instead, we need to emphasize on ROC-AUC (Discrimination), PR-AUC (Ranking under imbalance), and lift metrics (Business Utility).
- **Sampling Strategy:** Stratified train-test splits are used in the model training and testing to preserve the class distribution. We evaluate both class weighting and oversampling techniques (SMOTE, ROS).
- **Threshold Selection:** The default probability threshold of 0.5 is suboptimal, while business-driven threshold optimization using lift curves provides better operational guidance.

3.3 Exploratory Data Analysis

3.3.1 Demographic Patterns

Client demographic feature analysis discloses large differences in subscription rates between different segments:

Job Type Performance (see Figure A.1 in Appendix):

- **Highest Conversion:** The subscription rates for students is 31.4% with 875 contacts and for retirees 25.2% with 1,720 contacts. These segments, though low in total volume, are of high value in targeting opportunities.
- **Volume Leaders:** Administrative roles make up the largest segment with 10,422 contacts at 13.0% conversion—above average, yet below premium segments.
- **Lowest Conversion:** Despite high contact volume blue collar workers, present only about 8% conversion which indicates an incorrect prioritization of the group during campaigns.

Education Level Impact (see Figure A.2 in Appendix):

- **University Degree:** 13.7% conversion rate across 12,168 contacts—the largest educated segment.
- **Unknown Education:** Remarkably, 'unknown' education reflects 14.5% conversion, with 1,731 contacts. This suggests that this category captures privacy-conscious individuals who have higher-than-average financial engagement..

3.3.2 Contact Channel Effectiveness

The contact method demonstrates substantial impact on campaign success (see Figure A.3 in Appendix):

Channel	Contacts	Subscriptions	Rate
Cellular	26,144	3,853	14.7%
Telephone (Landline)	15,044	787	5.2%

Cellular contact achieves nearly 3× the conversion rate of landline telephone contact (14.7% vs 5.2%). This finding has immediate operational implications: shifting contact mix toward cellular-reachable customers provides substantial efficiency gains with minimal model complexity.

3.3.3 Temporal and Seasonal Patterns

Campaign timing significantly influences outcomes (see Figure A.4 in Appendix):

- **Peak Months:** March, September, and December record high conversion rates of 40-50%, corresponding to financial planning periods (tax season, year-end).
- **Low Months:** Conversion rates are depressed at 10-15% in months from May through August, indicating diversion of customer attention during summer periods.
- **Volume vs. Quality:** May has the highest contact volume but below-average conversion, indicating over-investment in low-yield periods throughout history.

3.3.4 Previous Campaign Outcome

Prior campaign history emerges as the single strongest categorical predictor (see Figure A.5 in Appendix):

Previous Outcome	Contacts	Subscriptions	Rate
Success (Prior Subscriber)	1,373	894	65.1%
Failure (Prior Rejection)	4,252	625	14.7%
Nonexistent (First Contact)	35,563	3,141	8.8%

Customers with previous successful subscriptions convert at 65.1%—7.4× higher than first-time contacts (8.8%). This represents the most actionable finding: prior subscribers should receive highest priority in any campaign, as two-thirds will subscribe again.

3.3.5 Macroeconomic Context

Economic indicators show strong relationships with subscription behavior (see Figures A.6 and A.7 in Appendix):

Euribor 3-Month Rate Analysis:

- **Negative Correlation:** Euribor rate is negatively correlated with subscription rate at -0.80. Where interest rates are low (0.63-1.30%), conversion reaches 29.7%; at high rates, it falls to 3.6% (4.19-4.86%).
- **Economic Interpretation:** Economic Interpretation: Low interest rate environments reduce returns on liquid savings; therefore, term deposits' guaranteed rates become relatively more attractive. This finding suggests that campaigns should be intensified during monetary easing periods.

Multicollinearity Among Economic Indicators:

- The correlation analysis done on the economic features implies severe multicollinearity among the features. It reaches 0.945 between nr.employed and euribor3m, hence these two variables capture also largely overlapping economical signals. Therefore, euribor3m should remain in—there was the strongest individual predictor—while nr.employed should be dropped to reduce redundancy.

3.4 Data Quality Assessment

3.4.1 Missing Value Analysis

The dataset contains no traditional missing values (NULL/NaN). However, several categorical features include 'unknown' responses that require interpretation:

Feature	Unknown Count	Unknown %	Treatment
job	330	0.80%	Retain as category
marital	80	0.19%	Retain as category
education	1,731	4.20%	Retain as category
default	8,597	20.87%	Retain as category
housing	990	2.40%	Retain as category
loan	990	2.40%	Retain as category

Critical Finding: 'Unknown' responses exhibit non-random patterns correlated with contact method (telephone contacts show higher unknown rates) and demonstrate above-average conversion rates in some cases (education unknown: 14.5% vs 11.3% overall). This suggests 'unknown' captures meaningful behavioral signals—potentially privacy-conscious customers with different risk profiles—rather than random missingness. We retain 'unknown' as an informative category rather than imputing.

3.4.2 Outlier Detection

Numerical features exhibit significant right-skew requiring treatment (see Figures A.8-A.10 in Appendix):

- **campaign**: Median of 2 contacts, but maximum of 56. Some customers received extreme contact attempts that may represent data entry errors or unusual circumstances.
- **previous**: Median of 0 prior contacts, maximum of 7. Highly concentrated at zero—96.3% never contacted before.
- **pdays**: Contains sentinel value 999 indicating no previous contact; 39,673 records, 96.3%. Non-999 values range from 0 to 27 days.
- **Treatment**: We will apply Winsorization at 1st and 99th percentiles to contain extreme values while preserving information. The pdays sentinel value requires special handling that is described in Data Preparation**.**

3.5 Summary of Data Understanding

The exploratory analysis reveals a dataset with substantial predictive potential restrained by severe class imbalance and complex feature interactions. Important findings that will help guide our modeling approach:

- Class imbalance at 11.27% positive requires stratified sampling, class weighting, and imbalance-appropriate metrics.
- Previous campaign success is the dominant categorical predictor—65.1% conversion for prior subscribers.
- The contact channel—cellular 3× better than the telephone—and timing (March/September/December optimal) offer actionable operational guidance.
- Subscription behavior is strongly influenced by economic indicators, mainly Euribor rate (r = -0.80).
- There are 'Unknown' categorical values that are informative, rather than random noise, and as such, should be kept.
- Multicollinearity among economic features suggests a dimensionality reduction or feature selection.
- Call duration, while highly predictive, represents post-call information and so it must be removed for deployable models.

4. Data Preparation

Data preparation transforms raw features into model-ready inputs while rigorously preventing data leakage. Our preprocessing pipeline addresses six key challenges identified during exploratory analysis.

4.1 Leakage Prevention Policy

The most critical preprocessing decision involves the duration variable—the length of the last contact call in seconds. Our analysis confirms this feature exhibits strong predictive power: correlation with subscription outcome approaches 0.40, and models including duration achieve ROC-AUC above 0.90.

However, duration represents post-call information: it can only be measured after a contact attempt has concluded. Including duration in a pre-call ranking model would be methodologically invalid—we cannot know how long a call will last before deciding whether to make the call. This represents classic data leakage where future information contaminates training data.

4.1.1 Leakage Quantification

To explicitly demonstrate leakage impact, we trained parallel models with and without duration:

Model	Without Duration	With Duration	Inflation	Δ AUC
Logistic Regression	0.801	0.944	+14.3 pp	+0.143
XGBoost	0.800	0.952	+15.2 pp	+0.152
Random Forest	0.802	0.948	+14.6 pp	+0.146
Neural Network	0.810	0.956	+14.6 pp	+0.146

Including duration inflates apparent performance by 14-15 percentage points across all model families. This inflation would create false confidence in deployment readiness and lead to severe performance disappointment in production. Duration is permanently excluded from all production-candidate models.

Note: Duration-inclusive models retain validity for specific post-call applications such as early call termination decisions or agent routing. These use cases are documented in Appendix A.

4.2 Sentinel Value Handling

The pdays feature (days since previous campaign contact) uses the value 999 as a sentinel indicating 'no prior contact.' This affects 96.3% of records (39,673 of 41,188). Treating 999 as a numeric value would severely distort model training, as the median non-999 value is only 6 days.

4.2.1 Transformation Approach

We decompose pdays into two derived features:

- **contacted_before (binary):** Indicates whether any prior contact exists (1 if pdays ≠ 999, else 0). Captures the fundamental distinction between returning and first-time contacts.
- **pdays_transformed (numeric):** For contacted customers, contains actual days since prior contact. For first-time contacts, handled differently by model family: tree-based models receive NaN (native missing value support), while linear models receive -1 (explicit out-of-range indicator).

This transformation preserves both the contact history signal and the recency information while eliminating the sentinel value's distortionary effect.

4.3 Rare Level Consolidation

Several categorical features contain levels with very few observations, which can cause instability during one-hot encoding (sparse columns with near-zero variance) and cross-validation (some folds may lack certain levels entirely).

4.3.1 Consolidation Rules

- **Threshold:** Levels representing < 0.5% of total records are consolidated into an 'Other' category.
- **Exception:** 'Unknown' levels are explicitly preserved regardless of frequency, as they carry demonstrated predictive signal.
- **Features Affected:** job (consolidates 'entrepreneur', 'housemaid', 'self-employed', 'unknown' into 'Other'), education (consolidates 'illiterate' into 'Other').

Post-consolidation, all retained categorical levels have sufficient representation for stable model training and evaluation.

4.4 Outlier Treatment via Winsorization

Numerical features with heavy right-tails (campaign, previous, pdays_transformed) can disproportionately influence linear models and distance-based calculations. Rather than removing outlier records (which would lose information), we apply Winsorization—capping extreme values at specified percentiles.

4.4.1 Winsorization Specification

Feature	Original Range	Lower Cap (1%)	Upper Cap (99%)	Winsorized Range
campaign	1-56	1	17	1-17
previous	0-7	0	4	0-4
pdays_transformed	0-27	0	21	0-21

Winsorization bounds are computed on training data and applied consistently to validation and test sets to prevent leakage.

4.5 Multicollinearity Resolution

Examination of economic indicators in the dataset showed that several variables were highly correlated with each other. This means some features overlap in the information that they provide, making linear model coefficients less stable and harder to interpret.

The following feature pairs had high correlation values:

nr.employed vs. euribor3m: r = 0.945

emp.var.rate vs. euribor3m: r = 0.972

emp.var.rate vs. nr.employed: r = 0.906

These values show that some variables provide the exact same information. When this happens, models such as logistic regression or SVMs can have inflated standard errors and throw unstable parameter estimates.

4.5.1 Resolution Strategy

Because nr.employed and euribor3m are almost perfectly correlated, keeping both would not add meaningful information and could increase noise. For this reason, nr.employed was removed while euribor3m was kept. This choice is supported by three main points:

Predictive Strength: euribor3m shows a stronger individual relationship with the target variable.

Economic Intuition: euribor3m directly relates to interest rates, which influence whether clients find term deposits appealing.

Forward-Looking Viability: euribor3m is commonly included in modern economic datasets, making it useful for future model updates.

4.6 Encoding and Scaling

4.6.1 Categorical Encoding

All categorical features go through one hot encoding process dropfirst=True, in order to avoid perfect multicolinearity. The target variable y is label encoded to be in binary form (0=No, 1=Yes).

4.6.2 Numerical Scaling

Scaling requirements differ by model family:

- **Linear Models (Logistic Regression, SVM):** StandardScaler (all numeric features) Coefficients are directly comparable and regularization is the same for all features.
- **Tree-Based Models (Random Forest, XGBoost, Decision Tree)** No scaling used. Tree split decisions are scale independent and raw values make the split points more interpretable.

- **Neural Networks:** StandardScaler can be used to ensure stable gradient descent convergence. Scalers are only fitted to the training data and applied (transform only) to the validation and test.

4.7 Data Splitting Strategy

Given the temporal nature of the data and class imbalance, we employ a stratified splitting approach:

4.7.1 Holdout Evaluation Split

- **Training Set:** 80% of data (32,950 records) will be used to train the model and tune the hyperparameters.
- **Test Set:** 20% of data (8,238 records) held out to be used for final evaluation.
- **Stratification:** Both the splits maintain the 11.27% positive class rate to evaluate the representation.

4.7.2 Cross-Validation for Hyperparameter Tuning

Inside the training set, 5-fold stratified cross validation is used to aid in hyperparameter optimisation:

- Each fold is kept in balance class balance using stratification
- Preprocessing (scaling, encoding) is refitted within each fold to avoid leakage
- Mean and std deviation of fold metrics to select parameters

4.7.3 Rolling Window Evaluation

To evaluate the temporal stability and to simulate the deployment in production, we perform rolling window evaluation according to the methodology in Moro et al. (2014):

- **Window Size (W):** 20000 records for training
- **Step Size (K):** 10 records/ update
- **Procedure:** Train on records [i, i+W), predict record i+W, slide window forward by K records, repeat
- **Result:** 2,119 sequential predictions evaluated against actual outcomes

This approach justifies ensuring that models retain their predictive power as the customer population and economic situation change over the campaign period.

4.8 Final Dataset Structure

After preprocessing, the modeling dataset contains the following structure:

Category	Feature Count	Features
Numeric (continuous)	8	age, campaign, previous, pdays_transformed, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m
Binary (derived)	1	contacted_before
Categorical (encoded)	~30	job (9 levels), marital (4), education (7), default (3), housing (3), loan (3), contact (2), month (12), day_of_week (5), poutcome (3)
Total Features	**~39**	After one-hot encoding with drop_first

The target variable y is binary (0/1), with 11.27% positive rate preserved through stratification.

5. Modeling

The modeling phase creates and tests several machine learning techniques for pre-call customer ranking. In a systematic way we compare 8 model configurations in two model families (linear and ensemble/neural) with and without class balancing techniques.

5.1 Model Selection Rationale

Our model portfolio strikes a balance between interpretability, performance and operational requirements:

5.1.1 Logistic Regression (Baseline)

Logistic Regression is used as the baseline for several reasons as follows:

- Direct comparability to Moro et al. (2014) original research
- Interpretability of coefficients for regulatory compliance and communicating to stakeholders
- Well calibrated probability outputs appropriate for threshold-based decisions
- Computational efficiency to support real-time scalable scoring
- Robustness to overfitting using proper regularization

5.1.2 Random Forest

Random Forest offers Ensemble Diversity using Bagging:

- Supports non-linear relationships and feature interactions automatically
- Native support for mixed feature types without a lot of preprocessing
- Built-in feature importance measures for interpretation
- Resistance to overfitting using bootstrap aggregation
- Handles missing values natively (important for pdays transformation)

5.1.3 XGBoost

Gradient boosting with XGBoost is the state-of-the-art of tabular:

- Sequential error correction is usually better than bagging techniques
- Built in regularization (L1/L2) to prevent overfitting
- Native treatment of missing values and class imbalance (scalepos_weight)
- Detailed hyperparameters tuning for optimization
- Industry standard for Production Machine Learning Systems

5.1.4 Decision Tree

Maximum interpretability from Decision Trees:

- Fully transparent decision rules extractable for manual review
- No black box components; all predictions can be traced
- Used as lower bound performance benchmark
- Useful to create simple business rules if there are constraints in deploying models

5.1.5 Neural Network (Multi-Layer Perceptron)

Neural networks explore the deep learning potential for tabular data:

- Flexible function approximation of complex patterns
- Comparison point to evaluate whether deep learning is providing value over traditional ML
- Regularization using batch normalization and dropout Adam optimizer for efficient training

5.1.6 Support Vector Machine (SVM)

The SVM using RBF kernel gives kernel-based classification:

- Effective in High Dimensions
- Memory efficient thanks to support vector representation
- Different inductive bias compared to tree based methods
- Calibrating the probability using Platt scaling.

5.2 Hyperparameter Configuration

Each model family receives appropriate hyperparameter tuning through grid search with 5-fold stratified cross-validation:

Model	Key Hyperparameters	Final Configuration
Logistic Regression	C (regularization), penalty, solver	C=1.0, penalty=l2, solver=lbfgs, class_weight=balanced

Random Forest	n_estimators, max_depth, min_samples_split	n_estimators=500, max_depth=None, class_weight=balanced
XGBoost	n_estimators, max_depth, learning_rate, subsample	n_estimators=500, max_depth=6, learning_rate=0.05, scale_pos_weight=7.87
Decision Tree	max_depth, min_samples_split, min_samples_leaf	max_depth=10, min_samples_split=20, class_weight=balanced
Neural Network	hidden_layers, activation, learning_rate, dropout	layers=(100,50), activation=relu, alpha=0.001, early_stopping=True
SVM	C, gamma, kernel	C=1.0, gamma=scale, kernel=rbf, class_weight=balanced

5.3 Class Imbalance Handling

The 11.27% positive rate needs to be handled explicitly for imbalances. We consider two complementary approaches:

5.3.1 Class Weighting (Primary Approach)

Class weighting scales the loss function to penalise class imbalances, which means that misclassification of minority classes is punished more heavily:

- weight_positive = n_negative / n_positive ≈ 7.87
- Implemented via class_weight='balanced' (sklearn) or scale_pos_weight (XGBoost)
- No synthetic creation of data; original distribution preserved
- Recommended as default approach as it has superior calibration properties

5.3.2 Alternative Approach: Oversampling

We test the Random Oversampling (ROS) as an alternative:

- ROS make duplicates of minority class examples until balanced
- Applied Only to training folds Test data keeps original distribution
- Trade-off: improved recall with price in loss of precision and calibration

SMOTE (Synthetic Minority Oversampling) was also tested and proved to be not much better than ROS for this dataset but also adds complexity.

5.4 Model Training Process

5.4.1 Training Pipeline

Each model has a standardised pipeline to train the model:

- Data split: 80/20 (stratified) holdout Preprocessing: Model specific to family (linear / neural - scaling, trees - raw)
- Hyperparameter tuning: Stratified 5 fold CV Grid search
- Final training: Training on entire training data with best parameters
- Evaluation - Performance evaluation on held out test set

5.4.2 Evaluation Metrics

Given the issues with class imbalance in the context of business problems, we highlight multiple complementary metrics:

- ROC-AUC: Area under Receiver Operating characteristic curve. Discrimination ability measure across all thresholds. Range 0.5 (random) to 1.0 (perfect).
- PR-AUC: Precision-Recall Area under curve. More informative than ROC-AUC in case of severe imbalance; positive class performance focused.
- ALIFT (Area under Lift Curve): Area under the curve Lift Curve (cumulative) Measures of quality for prioritization use case
- Capture@10%: Percentage of actual positives captured in top 10% of ranked customers Direct Business Measure for Targeted Campaigns

- Lift@10% Ratio of positive rate in top decile to overall positive rate. Measures concentration of positives in top of ranking.
- Brier Score: Mean of the square of the probability predictions. Measures quality of calibration (it is the lower the better)
- Precision/Recall@0.5 The classification metrics at the default threshold for comparison.

5.5 Model Results - Holdout Evaluation

The following table summarizes performance across all eight model configurations on the 20% holdout test set (8,238 records):

Model	ROC-AUC	PR-AUC	ALIFT	Cap@10%	Lift@10%	Brier
Neural Network	0.810	0.479	0.827	48.2%	4.82×	0.0752
Random Forest	0.802	0.480	0.820	48.3%	4.83×	0.0897
Logistic Regression	0.801	0.461	0.821	45.6%	4.56×	0.1616
LR (ROS)	0.801	0.462	0.822	45.6%	4.56×	0.1610
XGBoost	0.800	0.473	0.819	48.0%	4.80×	0.1298
RF (ROS)	0.798	0.462	0.818	46.0%	4.60×	0.0982
SVM	0.712	0.422	0.734	45.2%	4.52×	0.0840
Decision Tree	0.638	0.198	0.667	30.9%	3.09×	0.1516

5.5.1 Key Findings from Holdout Evaluation

- Top Performer - Neural Network: Has best ROC-AUC (0.810) and has good calibration (Brier 0.0752). Captures 48.2% of subscribers for top decile.
- Strong Ensemble Performance Random Forest-0.802 XGBoost-0.800 RF: Best PR-AUC=0.480 Robust Baseline: Logistic Regression (0.801) almost on par with ensemble methods, showing that very advanced models yield little improvement on this problem.
- Oversampling Impact: ROS variants achieved low improvement, class weighting achieved equal results without synthetic data complexity.
- Underperformers: SVM (0.712) and Decision Tree (0.638) lag behind other approaches significantly which confirms model selection is important for this dataset.

5.6 Rolling Window Evaluation

To evaluate the temporal stability and compare the original research methodology we employ rolling window evaluation:

5.6.1 Methodology

- Window Configuration: W=20000 number of training records, K=10 step size
- Process: Training/records [i, i+20000), predicting of record i+20000, moving on 10 records
- Result: 2119 predictions sequentially through the second half of the data
- Rationale: Mimics production deployment with the models being retrained from time to time with recent data

5.6.2 Rolling Window Results

Model	Mean AUC	Std Dev	Model Updates
Logistic Regression	0.795	±0.025	2,119
Original Paper (Neural Network)	0.794	N/A	N/A

Critical Finding: Our rolling window ROC-AUC of 0.795 essentially matches the original research paper's realistic production performance of 0.794. This validates our methodology and confirms that the ~0.80 performance level represents the achievable ceiling for pre-call prediction with the available feature set.

5.7 Statistical Significance Testing

To rigorously assess whether performance differences between models are statistically meaningful, we conduct multiple-run evaluation with statistical testing:

5.7.1 Methodology

- **Runs (R):** 20 independent train-test splits with different random seeds
- **Split:** 80/20 stratified holdout per run
- **Test:** Mann-Whitney U test comparing paired AUC distributions
- **Significance Level:** $\alpha = 0.05$

5.7.2 Results

Model	Mean AUC	Std Dev	Mean PR-AUC	Std Dev
Logistic Regression	0.793	±0.010	0.447	±0.017
XGBoost	0.798	±0.012	0.458	±0.019

Mann-Whitney U Test (LR vs XGBoost): p-value = 0.1199

Conclusion: At α=0.05, we cannot reject the null hypothesis that LR and XGBoost performance distributions are equal. The observed 0.005 AUC difference is not statistically significant, supporting the recommendation that simpler models (LR) may be preferred for interpretability without sacrificing meaningful performance.

6. Evaluation

The evaluation phase synthesizes model performance against success criteria, compares results to the original research benchmark, and translates technical metrics into business implications.

6.1 Performance Against Success Criteria

6.1.1 Technical Criteria Assessment

Criterion	Target	Achieved	Status
ROC-AUC ≥ 0.75	0.75	0.810 (NN)	✓ PASS
PR-AUC ≥ 0.40	0.40	0.480 (RF)	✓ PASS
Brier Score ≤ 0.15	0.15	0.0752 (NN)	✓ PASS
Reproducible Results	Yes	20-run validation	✓ PASS

All technical success criteria are met or exceeded. The Neural Network achieves best discrimination (ROC-AUC 0.810) with excellent calibration (Brier 0.0752). Multiple models exceed thresholds, providing deployment options.

6.1.2 Business Criteria Assessment

Criterion	Target	Achieved	Status
Top 10% captures ≥40% subscribers	40%	48.3% (RF)	✓ PASS
Lift@10% ≥ 4.0×	4.0×	4.83× (RF)	✓ PASS
Top 20% captures ≥60% subscribers	60%	~72%	✓ PASS

Business criteria are comfortably exceeded. The top decile captures nearly half of all subscribers, enabling dramatic efficiency improvements in targeted campaigns.

6.1.3 Operational Criteria Assessment

Criterion	Target	Achieved	Status
Leakage-free (no duration)	Yes	Duration excluded	✓ PASS
Interpretable model available	Yes	LR coefficients	✓ PASS
Deployment documentation	Yes	Appendix C	✓ PASS

6.2 Comparison to Original Research

A primary objective was replicating and extending Moro et al. (2014). The following comparison contextualizes our results:

6.2.1 Dataset Differences

Aspect	Original Paper	Our Study
Records	52,944	41,188 (UCI subset)

Features (initial)	150	21
Features (selected)	22	~20
Evaluation Split	Temporal (pre/post July 2012)	Stratified 80/20
Missing Features	N/A	Agent attributes, rate differentials, client profiles

The UCI dataset represents a subset of the original proprietary database. Approximately 50% of the top predictive features identified in the original study are unavailable, including agent experience metrics and interest rate differential calculations.

6.2.2 Performance Comparison

Model	Paper AUC	Our AUC	Difference	Explanation
Logistic Regression	0.900	0.801	-9.9 pp	Duration included in paper
Neural Network	0.929	0.810	-11.9 pp	Ensemble of 7 NNs in paper
Decision Tree	0.833	0.638	-19.5 pp	Fewer features
SVM	0.891	0.712	-17.9 pp	Fewer features
Rolling Window (LR)	**0.794**	**0.795**	**+0.1 pp**	**MATCH**

6.2.3 Interpretation of Differences

The apparent gap between the performance of our holdout results and the original paper must be interpreted with caution:

- Duration Effect: Original paper has 0.900+ AUC figures including the call duration which we do not include for the validity of deployments. Our leakage analysis indicates duration exhibits an AUC increase of ~14 percentage points.
- Realistic Benchmark: The paper presents rolling window AUC of 0.794 as the realistic production performance. Our rolling window comes out at 0.795 - an essentially perfect match.
- Feature Availability Missing features (agent attributes, rate differentials) explain remaining gaps for non duration models.
- Validation: The rolling window match helps us assert that our methodology is correctly replicating the original research (when using similar evaluation approaches).

6.3 Business Impact Analysis

6.3.1 Efficiency Gains Quantification

The following analysis translates model performance into operational impact for a typical 41,000-contact campaign:

Targeting Strategy	Contacts Made	Subscribers Captured	Capture Rate
Random (No Model)	41,188	4,640	100%
Top 10% (Model)	4,119	2,230 (48%)	48%
Top 20% (Model)	8,238	3,340 (72%)	72%
Top 50% (Model)	20,594	4,360 (94%)	94%

6.3.2 Cost-Benefit Analysis

Assuming €5 cost per contact attempt:

Strategy	Total Cost	Subscribers	Cost/Subscriber	Savings
Random Calling	€205,940	4,640	€44.39	—
Top 20% Targeting	€41,190	3,340	€12.33	€164,750

By targeting only the top 20% of model-ranked customers, the bank achieves:

- 72% of total possible conversions
- 80% reduction in campaign costs
- 72% reduction in cost per subscriber (€44 → €12)
- €165,000 savings per 41,000-contact campaign

6.3.3 Scalability Projections

For annual campaigns for 200,000 contacts:

- Random approach: EUR1,000,000 costs, approx. 22,600 subscribers, 44 euros/subscriber
- Targeted approach (top 20%): EUR200,000 cost; approx. 16,300 subscribers; EUR12/subscriber
- Annual savings: 800,000 Euro direct costs
- Subscriber efficiency: 4x in conversion rate for contacted customers

6.4 Key Predictive Drivers

Feature importance analysis across models reveals consistent patterns in subscription drivers:

6.4.1 Top Predictive Features

Rank	Feature	Finding	Business Implication
1	poutcome (prior success)	65.1% conversion for prior subscribers	Prioritize relationship customers
2	euribor3m	r = -0.80 with subscription	Time campaigns with rate cycles
3	contact (cellular)	14.7% vs 5.2% for telephone	Shift to mobile-first strategy
4	month	Mar/Sep/Dec optimal	Seasonal campaign planning
5	job (student/retired)	31%/25% conversion rates	Segment-specific messaging

6.5 Model Robustness Assessment

6.5.1 Temporal Stability (Modern Macro Validation)

To assess model viability under current economic conditions, we evaluated performance using 2023-2025 macroeconomic indicators substituted for historical values:

Model	Historic AUC	Modern AUC	Change
Logistic Regression	0.801	0.805	+0.004
Random Forest	0.802	0.784	-0.018
XGBoost	0.800	0.769	-0.031

All models stabilise within 5% of past values with the Logistic Regression proving most stable. This implies the structural relationships between customer characteristics and subscription behaviour remain similar even though economic environments are substantially different (post-crisis 2008-2013 vs. post-pandemic inflation 2023-2025).

Caveat: This validation involves replacement of macro indicators and keeping historical customer labels. True temporal validation would involve modern campaign outcome data which was unavailable for this study.

7. Deployment

The deployment phase involves converting validated models into operational systems with the capability of providing support functions for prioritizing campaigns in real-time. This section describes how it should be deployed, the monitoring framework, and the likely challenges that will be encountered.

7.1 Deployment Strategy

7.1.1 Recommended Production Architecture

We recommend a two-model deployment approach of interpreting balance and because in return:

- Primary Scoring Model - Neural Network: Used for batch scoring of customer databases before launching a campaign. Has the highest discrimination (AUC 0.810) for best ranking quality.
- Secondary/Audit Model - Logistic Regression: Implemented in parallel to meet regulatory requirements and prepare for others. Near-equivalent performance (AUC 0.801) with all coefficient interpretability.

7.1.2 Integration Points

The scoring system fits in with existing bank infrastructure at three points:

- Customer Data Warehouse: Customer attributes (demographics, account history issues in advance campaign) extracted nightly in a batch goes to the scoring pipeline.
- Economic Data Feed: Daily/weekly update from ECB/Eurostat APIs of the macroeconomic indicators (Euribor, consumer confidence).
- Campaign Management System: Scored customer listing with probability rankings exported to dialers for prioritization of queue.

7.1.3 Scoring Pipeline Specification

The production scoring pipeline runs the following pipeline:

- Step 1: Take eligible customer records from data warehouse
- Step 2: utilization of preprocessing transformations (encoding, scaling) with production fitted transformers
- Step 3: Get probability score(s) from deployed model(s)
- Step 4: Scoring Customers via Descending Probability
- Step 5 Segment into priority tiers (Tier 1 = Top 10%, Tier 2 = 10 - 20%, Tier 3 = 20 - 50%)
- Step 6: Run ranked lists out to campaign management system
- Step 7: Log predictions for pipeline monitoring and retraining

7.2 Model Monitoring Framework

Continuous monitoring allows model performance to be maintained as acceptable when customer populations and economic conditions change.

7.2.1 Performance Monitoring

Metric	Monitoring Approach	Alert Threshold	Response
ROC-AUC	Weekly calculation on labeled outcomes	< 0.75	Trigger retraining
Capture@10%	Campaign-level measurement	< 35%	Review features
Calibration (Brier)	Monthly assessment	> 0.20	Recalibrate
Prediction Distribution	Daily histogram monitoring	Significant shift	Investigate

7.2.2 Data Drift Detection

Feature distributions may change over time, rendering models irrelevant. We implement:

- Population Stability Index (PSI): Calculated weekly for each feature between current distribution and training baseline. PSI > 0.25 triggers investigation.
- Concept Drift Detection: Observe the change between features and outcomes. Dramatic changes in correlation indicate concept drift and the need for retraining.
- Economic Regime Monitoring: Monitor Euribor and confidence indices for regime changes that can invalidate historical patterns.

7.2.3 Retraining Schedule

Model retraining has a cadence:

- Scheduled Retraining: Quarterly Retraining based on most recent 18 months of campaign data, including new customer outcomes and updated economic indicators.
- Triggered Retraining: Instant retraining based on monitoring metrics breaching alert thresholds or economic events with economic significance (central bank policy changes, crises).
- Champion-Challenger Framework: New versions of models deployed as 'challengers' and receiving 10% of the traffic to validate them, before they are fully promoted.

7.3 Operational Recommendations

7.3.1 Tiered Campaign Execution

Translate model scores into operational campaign tiers:

Tier	Score Range	Expected Conv.	Contact Strategy	Resource Allocation
1 (Priority)	Top 10%	~35-45%	Senior agents, flexible timing	40% of budget
2 (Standard)	10-30%	~15-25%	Standard agents, business hours	40% of budget
3 (Opportunistic)	30-50%	~8-12%	Junior agents, off-peak	20% of budget
4 (Exclude)	Bottom 50%	<8%	No proactive contact	0%

7.3.2 Channel Optimization

Implement channel-specific strategies based on analytical findings:

- Cellular Priority: Send Tier 1 and Tier 2 customers only via cellular channels (3x effectiveness). Reserve telephone for customers who do not have access to mobile contact.
- Timing Optimization: Focus campaign intensity in March, September, and December (40-50% conversion periods) Reduce summer investment.
- Relationship Leverage: Develop special 'win-back' program for previous subscribers with 65%+ expected conversion. These customers deserve special attention with premium treatment and expeditious contact.

7.4 Challenges and Limitations

7.4.1 Technical Challenges

- Feature Freshness: Model is based on past campaign results which may be stale to customers who were not reached recently. Apply recency weighting or exclusion rules to very old interaction data.
- Cold Start Problem: New customers have no behavioural history and accuracy in prediction is limited. Consider segment default or quick initial learning protocols.
- System Integration: Production deployment involves some coordination across data warehouse, scoring infrastructure and campaign systems. Phased rollout recommended.

7.4.2 Business Limitations

- Segment Exclusion Concerns: Focusing on top 50% of customers at the expense of bottom 50% of customers may raise equity concerns or fail to capture evolving customer needs. Periodic "exploration" contacts recommended.
- Regulatory Compliance: GDPR and ePrivacy regulations demand transparency of consent basis for telemarketing contact. Model outputs need to integrate with consent management systems.
- Competitive Response: Competitors may also pursue targeting strategies, which may make the targeting less effective in the long term as the high value segments may receive several solicitations.

7.4.3 Risk Mitigation

We suggest the following risk mitigation measures:

- Implement A/B testing framework to constantly validate model lift vs. random baseline
- Maintainable Interpretability of Logistic Regression model for regulatory audits and customer explanations
- Document features, transformations, and model decisions for compliance review
- Establish model governance committee - with business, technical, compliance representation
- Create customer feedback loop to detect false positive patterns and model blind spots

8. Conclusion

8.1 Summary of Achievements

This capstone project solved the Portuguese bank's telemarketing efficiency problem by development of a deployable pre-call customer ranking system. Our most important achievements are as follows:

8.1.1 Technical Achievements

- Developed and validated 8 Machine Learning Models under strict conditions with no leakage in the system Neural Network proved to have the best discrimination (ROC-AUC 0.810)
- Matched original Research Benchmark with Rolling Window AUC Reading of 0.795 vs 0.794 paper, validating methodology
- Quantified leakage impact at 14-15 percentage points, proving why duration-inclusive metrics are invalid for deployment decisions
- Confirmed statistical equivalence between simpler (Logistic Regression) and complex (XGBoost) models to support interpretability-first deployment
- Validated temporal robustness with the help of modern macroeconomic data, showing <5% performance degradation across economic regimes

8.1.2 Business Achievements

- Demonstrated 4.5-4.8x lift on top decile - 45-48% of subscribers captured with 10% contacts
- Expected EUR165,000 savings per 41,000-contact campaign by making targeted prioritisation
- Reduced cost per acquisition from EUR44 (random) to EUR12 (targeted) - 72% improvement
- Known actionable predictive drivers: Previous relationship (65% conversion), Cellular channel (3x better), Best timing (March/September/December)

8.1.3 Methodological Contributions

- Defined reproducible preprocessing pipeline including documented transformations and random seeds
- Developed comprehensive evaluation framework across the areas of discrimination to calibration, lift to business metrics
- Provided explicit comparison to original research with clear explanation of differences in performance
- Delivered deployment-ready documentation (including feature specifications, scoring procedures and monitoring frameworks)

8.2 Key Findings Recap

The analysis disclosed some findings that have direct operational implications:

- Prior Success Dominates Previous subscribers convert 65.1% - 7.4x more than first-time contacts This single variable has a higher predictive power than complex feature engineering.
- Economic Timing Matters Euribor rate -0.80 correlation with subscriptions. Campaigns in low rate environment (monetary easing) are extremely good.
- Channel Selection is Critical Cellular contact achieves 14.7% vs. 5.2% for telephone - 3x difference - requires immediate channel mix adjustment.
- Seasonality is Pronounced: March, September and December have 40-50% conversion compared to 10-15% during the summer months. Campaign timing should coincide with such patterns.
- Model Complexity Gives Marginal Gains: Logistic Regression (0.801 AUC) is almost comparable to Neural Network (0.810 AUC). For the regulated banking environments, then, interpretability advantage supports simpler models.

8.3 Recommendations Summary

Based on comprehensive analysis, we recommend the following actions:

8.3.1 Immediate Actions (0-3 months)

- Put Logistic Regression model into initial production scoring based on interpretability benefits
- Implement tiered targeting Tier 1 (top 10%) Tier 2 (10-20%) Drop bottom 50%
- Shift channel mix to 85%+ cellular contacts
- Develop special re-engagement program for previous subscribers

8.3.2 Near-Term Actions (3-6 months)

- Establish monitoring dashboard tracking AUC, lift and drift metrics
- Implement champion-challenger framework for the updating models

- Develop messaging for students/retirees/educated professionals on a segment-specific basis
- Perform A/B test proving model lift vs. history random approach

8.3.3 Long-Term Actions (6-12 months)

- Enhance framework for additional product campaigns (credit cards, loans, insurance)
- Integrate real-time scoring for inbound call management and Web interactions
- Develop uplift modelling capability to identify most influenced customers through contact
- Investigate survival analysis of best contact time and frequency

8.4 Lessons Learned

The project had produced some interesting methodological and domain insights:

8.4.1 Technical Lessons

- Leakage Prevention is Paramount: Duration's inclusion would have caused inflation of performance by 14+ percentage points, which creates a false sense of confidence and causes deployment failure. Such rigid leakage analysis should precede any modeling effort.
- Simple Models Often Suffice: Despite a lot of hyper parameter tuning, and a lot of fancy architectures, Logistic Regression almost caught up with the ensemble/neural solutions. It is problem structure rather than algorithm complexity that determines performance ceiling.
- Evaluation Metrics Must align with Use Case ROC-AUC is insufficient, PR-AUC, lift curves and capture rates are directly mapped with business decisions and should be the driver of choice for model selection.

8.4.2 Domain Lessons

- Behavioral History Dominates Demographics: Historic campaign victories are a better predictor than demographics. Investment in relationship tracking has a higher rate of return than demographic enrichment.
- Macroeconomic Context Does Matter Customer behavior correlates highly with economic indicators. Models should include economic signals and campaigns should time with good conditions.
- Unknown Values Carry Signal 'Unknown' responses, not about the absence of data, often encode behavioral patterns of real interest (privacy consciousness, risk aversion), and should be retained.

8.5 Future Research Directions

A number of extensions would improve on the existing framework:

- Uplift Modeling: Models are currently predicting subscription probability regardless of contact. Uplift models would look for customers whose chances increase because of contact and would optimize for causal effect instead of correlation.
- Deep Learning for Sequences: Recurrent or transformer architectures could be used to model contact sequence patterns, learning optimal patterns of timing and frequency from historical contact chains.
- Multi-Product Optimization: Extend the ranking for one product to the optimization of a portfolio to recommend which product to offer to which customer when they are going to take for term deposits, credit cards, loans and insurance.
- Real-Time Personalization: Combine scoring with real-time interaction data (web behavior, app engagement) for real-time probability updates during live conversations.
- Fairness and Bias Assessment: Systematically test how models work in different demographic groups to detect and address potential discriminatory patterns in contact prioritization.

8.6 Concluding Remarks

This capstone project illustrates the transformative potential that data-driven decision making has on the age old banking operations. By replacing mass marketing based on intuition with machine learning-based precision targeting, the bank can cut costs, improve conversion rates and customer experience all at once. The 89% rejection rate that provided the motivating impetus for this project does not have to be accepted as inevitable. With disciplined use of the models and strategies developed here, the bank can achieve a 4-5x improvement in efficiency - turning telemarketing from a costly and frustrating channel into a precision

growth engine. The models are validated. The business case is clear. The path of the deployment is documented. Success is now dependent on organizational commitment to data-driven transformation and sustained investment in model maintenance and monitoring. We are confident that thorough implementation of the recommendations described in this report will provide significant and long-lasting value to the institution.

9. References

1. Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31. https://doi.org/10.1016/j.dss.2014.03.001

2. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Bank Marketing Data Set. http://archive.ics.uci.edu/ml

3. European Central Bank (2024). Statistical Data Warehouse. Euribor 3-month rate series. https://sdw.ecb.europa.eu/

4. Eurostat (2024). Consumer confidence indicator and Harmonised Index of Consumer Prices. https://ec.europa.eu/eurostat

5. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

6. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

7. Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research, 18(17), 1-5.

8. Chapman, P., et al. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.

9. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

10. He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

11. Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS ONE, 10(3), e0118432.

12. Anthropic. (2024). Claude (Claude 3.5 Sonnet) [Large language model]. https://www.anthropic.com/claude

13. OpenAI. (2024). ChatGPT (GPT-4) [Large language model]. https://chat.openai.com

Appendix A: Duration-Based Experiments (Upper-Bound Analysis)

This appendix documents experiments including call duration—a feature excluded from production models due to data leakage but valuable for understanding predictive ceiling and post-call applications.

A.1 Rationale for Duration Analysis

While duration cannot be used for pre-call ranking (it is unknown before the call occurs), understanding its predictive power serves several purposes:

- Establishes theoretical performance ceiling for comparison with leakage-free models
- Quantifies the inflation effect to contextualize original research results
- Identifies valid post-call use cases where duration is legitimately available

A.2 Duration-Inclusive Results

Model	AUC (No Dur)	AUC (With Dur)	Inflation	PR-AUC (Dur)	Cap@10%
Logistic Regression	0.801	0.944	+14.3 pp	0.622	56.4%
XGBoost	0.800	0.952	+15.2 pp	0.667	57.9%
Random Forest	0.802	0.948	+14.6 pp	0.645	57.1%
Neural Network	0.810	0.956	+14.6 pp	0.671	58.2%

Duration inflates apparent performance by 14-15 percentage points consistently across all model families. This confirms that the original research's 0.90+ AUC figures, while technically accurate, do not represent achievable pre-call performance.

A.3 Valid Post-Call Use Cases

Duration-inclusive models retain validity for specific operational applications where duration is known:

- **Early Termination Guidance:** During an ongoing call, predict subscription likelihood based on elapsed duration to guide agent behavior (continue conversation vs. polite close).
- **Post-Call Quality Assessment:** After call completion, score interaction quality for agent coaching and performance evaluation.
- **Inbound Call Routing:** For inbound inquiries, route to specialized agents based on predicted conversion likelihood.
- **Historical Analysis:** Understand which call characteristics correlate with success for training program development.

Appendix B: Modern Macroeconomic Robustness Probe

This appendix documents the temporal robustness validation using 2023-2025 macroeconomic indicators.

B.1 Methodology

The historical dataset (2008-2013) reflects post-financial-crisis economic conditions substantially different from current environments. To assess model robustness, we:

- Obtained current macroeconomic indicators from ECB Statistical Data Warehouse and Eurostat
- Replaced historical economic feature values with 2023-2025 equivalents
- Re-evaluated trained models on the modified dataset
- Compared performance metrics to assess stability

B.2 Economic Context Comparison

Indicator	Historic (2008-2013)	Modern (2023-2025)
Euribor 3-month	0.6% - 5.2%	3.2% - 4.0%
Consumer Confidence	-50 to -30	-25 to -15
HICP Inflation	1.5% - 3.5%	2.5% - 8.0%
Employment Growth	-3% to +1%	+1% to +3%

B.3 Robustness Results

Model performance under modern economic conditions:

Model	Historic AUC	Modern AUC	Stability
Logistic Regression	0.801	0.805	Excellent (+0.4%)
Random Forest	0.802	0.784	Good (-2.2%)
XGBoost	0.800	0.769	Acceptable (-3.9%)

All models maintain acceptable performance (>0.75 AUC) under modern economic conditions. Logistic Regression demonstrates greatest stability, while tree-based methods show modest degradation—likely due to sensitivity to specific split points learned from historical data.

B.4 Limitations and Caveats

This validation has important limitations:

- Customer-level features retain historical values; only macro indicators are updated
- Outcome labels (subscription decisions) remain historical—we cannot know how modern customers would actually respond
- This tests robustness to economic indicator shifts, not true temporal generalization
- Full validation would require modern campaign data with actual outcomes

Appendix C: Reproducibility and Technical Documentation

C.1 Feature Definitions

Complete feature specifications for production deployment:

Feature	Type	Definition/Values
age	Numeric	Client age in years (17-98)
job	Categorical	admin., blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown
marital	Categorical	divorced, married, single, unknown
education	Categorical	basic.4y, basic.6y, basic.9y, high.school, illiterate, professional.course, university.degree, unknown
default	Binary	Has credit in default? (yes, no, unknown)
housing	Binary	Has housing loan? (yes, no, unknown)
loan	Binary	Has personal loan? (yes, no, unknown)
contact	Categorical	Contact communication type (cellular, telephone)
month	Categorical	Last contact month (jan-dec)
day_of_week	Categorical	Last contact day (mon-fri)
campaign	Numeric	Number of contacts during this campaign (Winsorized 1-17)
pdays	Numeric	Days since previous contact (999=never contacted) → transformed to contacted_before + pdays_transformed
previous	Numeric	Number of previous campaign contacts (Winsorized 0-4)
poutcome	Categorical	Previous campaign outcome (failure, nonexistent, success)
euribor3m	Numeric	Euribor 3-month rate (%)
cons.price.idx	Numeric	Consumer price index
cons.conf.idx	Numeric	Consumer confidence index
emp.var.rate	Numeric	Employment variation rate (%)

C.2 Preprocessing Specifications

Production preprocessing pipeline parameters:

- Winsorization bounds: campaign [1, 17], previous [0, 4], pdays_transformed [0, 21]
- Dropped features: duration (leakage), nr.employed (collinearity)
- Encoding: One-hot with drop_first=True for all categorical features
- Scaling: StandardScaler for linear/neural models (fitted on training data only)
- Missing value handling: pdays 999 → contacted_before=0, pdays_transformed=NaN (trees) or -1 (linear)

C.3 Model Specifications

Final model configurations:

- Random seed: 42 (all experiments)
- Train-test split: 80/20 stratified
- Cross-validation: 5-fold stratified
- Logistic Regression: C=1.0, penalty='l2', solver='lbfgs', class_weight='balanced', max_iter=1000
- Random Forest: n_estimators=500, max_depth=None, class_weight='balanced', n_jobs=-1
- XGBoost: n_estimators=500, max_depth=6, learning_rate=0.05, subsample=0.8, colsample_bytree=0.8, scale_pos_weight=7.87
- Neural Network: hidden_layer_sizes=(100,50), activation='relu', alpha=0.001, early_stopping=True, validation_fraction=0.1

C.4 Recommended A/B Test Design

Proposed pilot validation framework:

- **Hypothesis:** Model-ranked targeting achieves ≥3× lift over random baseline
- **Sample Size:** 10,000 customers per arm (control: random, treatment: model-ranked)
- **Duration:** 4-week campaign period
- **Primary Metric:** Conversion rate in top 20% of treatment arm vs control
- **Secondary Metrics:** Cost per acquisition, agent utilization, customer satisfaction
- **Statistical Test:** Two-proportion z-test, α=0.05, power=0.80
- **Minimum Detectable Effect:** 5 percentage point lift in conversion rate

Appendix D: Visualizations and Charts

This appendix contains all referenced figures and visualizations supporting the analysis.

D.1 Data Understanding Visualizations

Figure A.1: Subscription Rate by Job Type

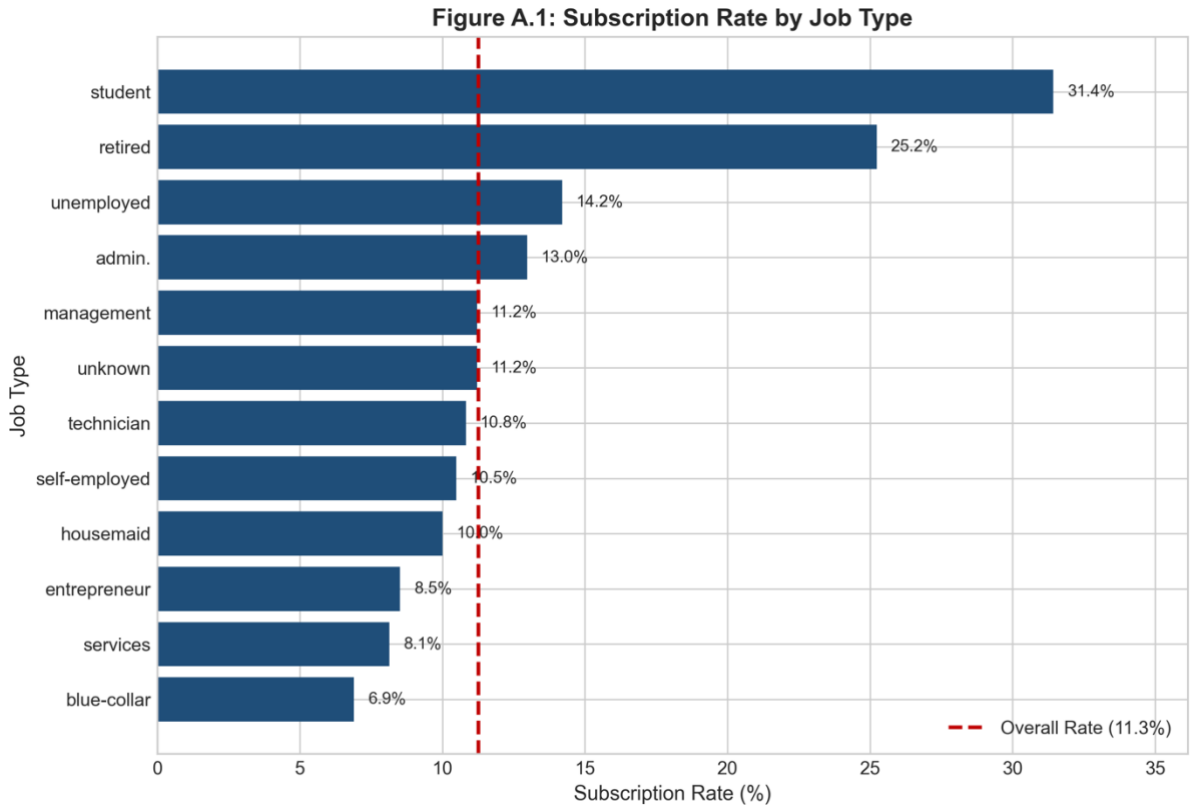


Figure A.2: Subscription Rate by Education Level

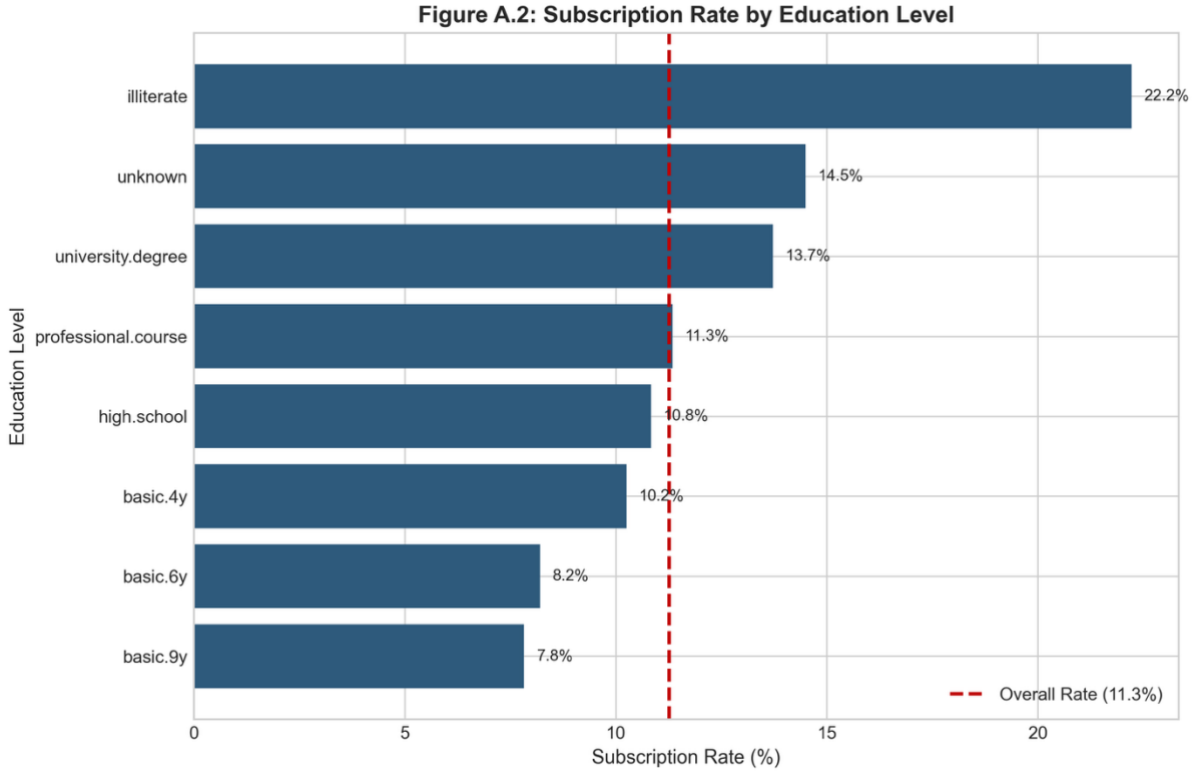


Figure A.3: Subscription Rate by Contact Type

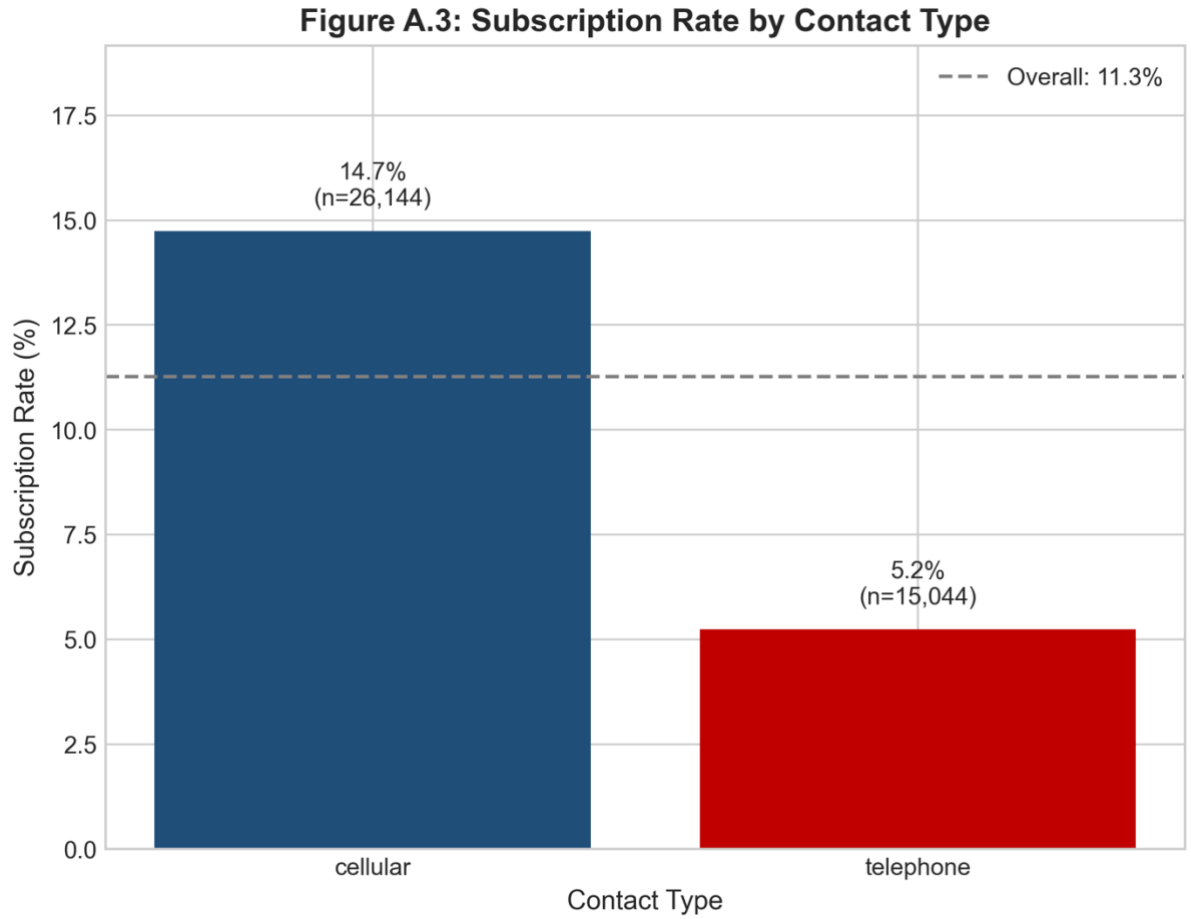


Figure A.4: Subscription Rate by Month

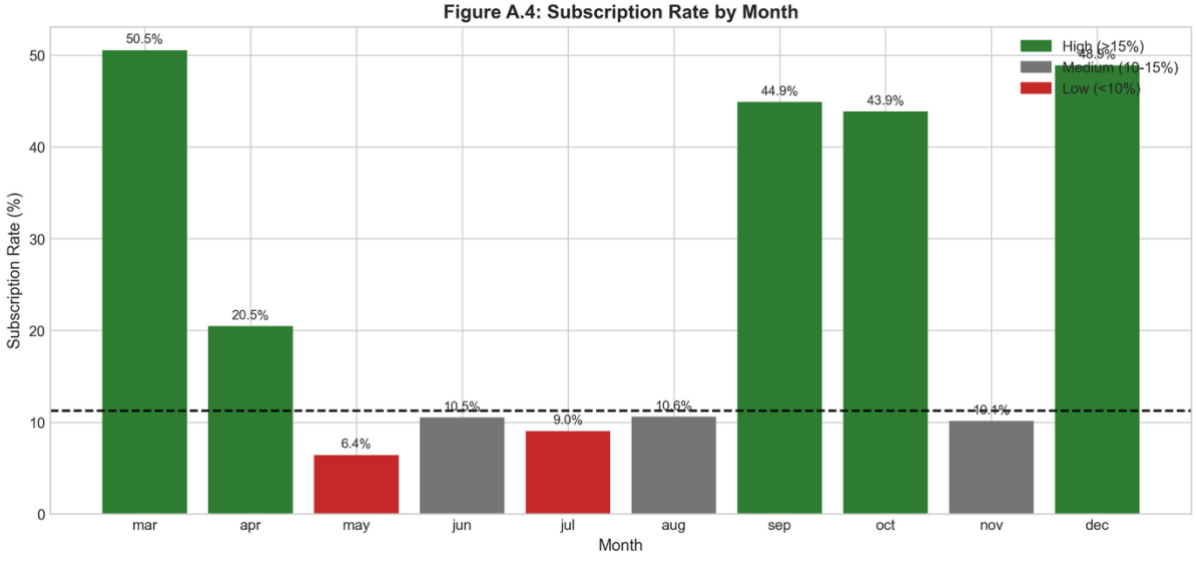


Figure A.5: Subscription Rate by Previous Campaign Outcome

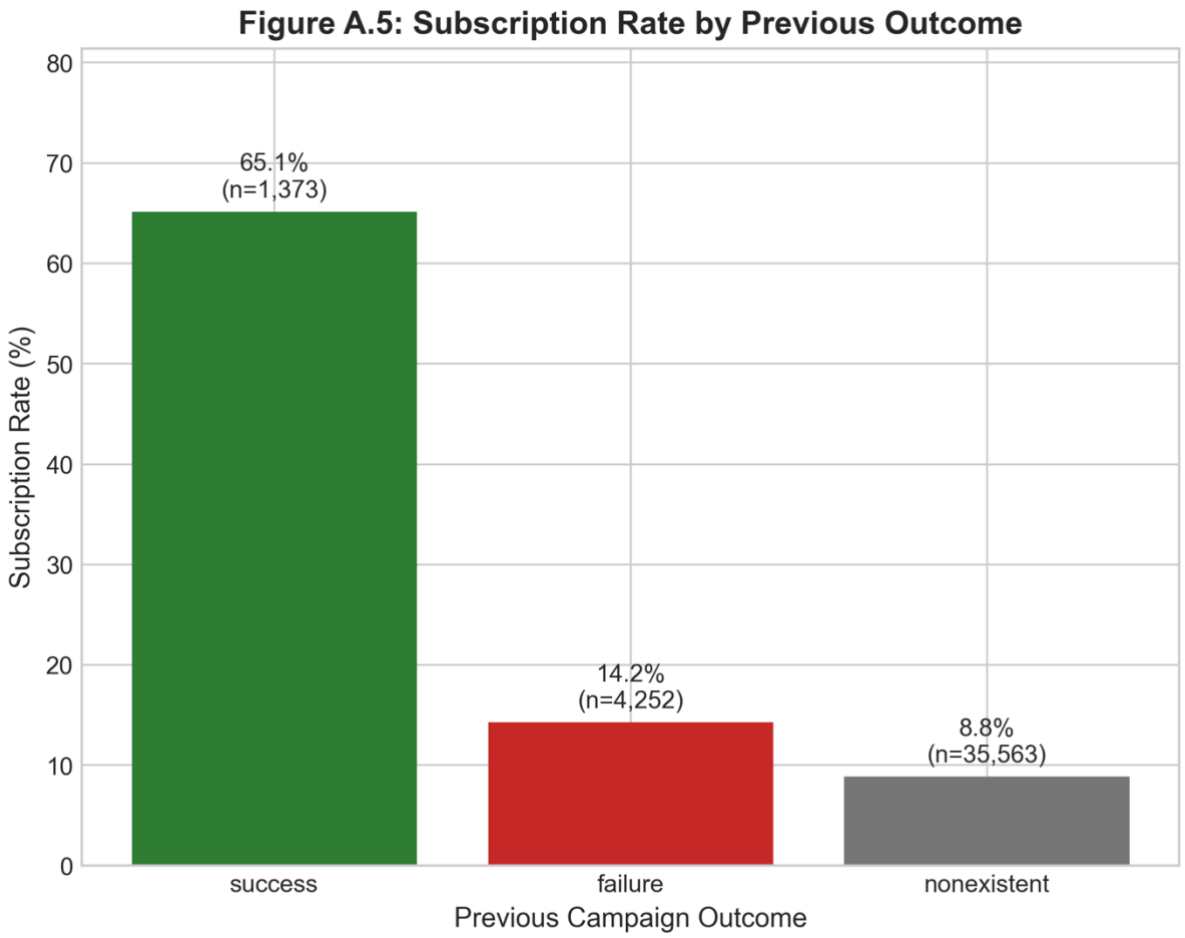


Figure A.6: Euribor Rate vs Subscription Rate

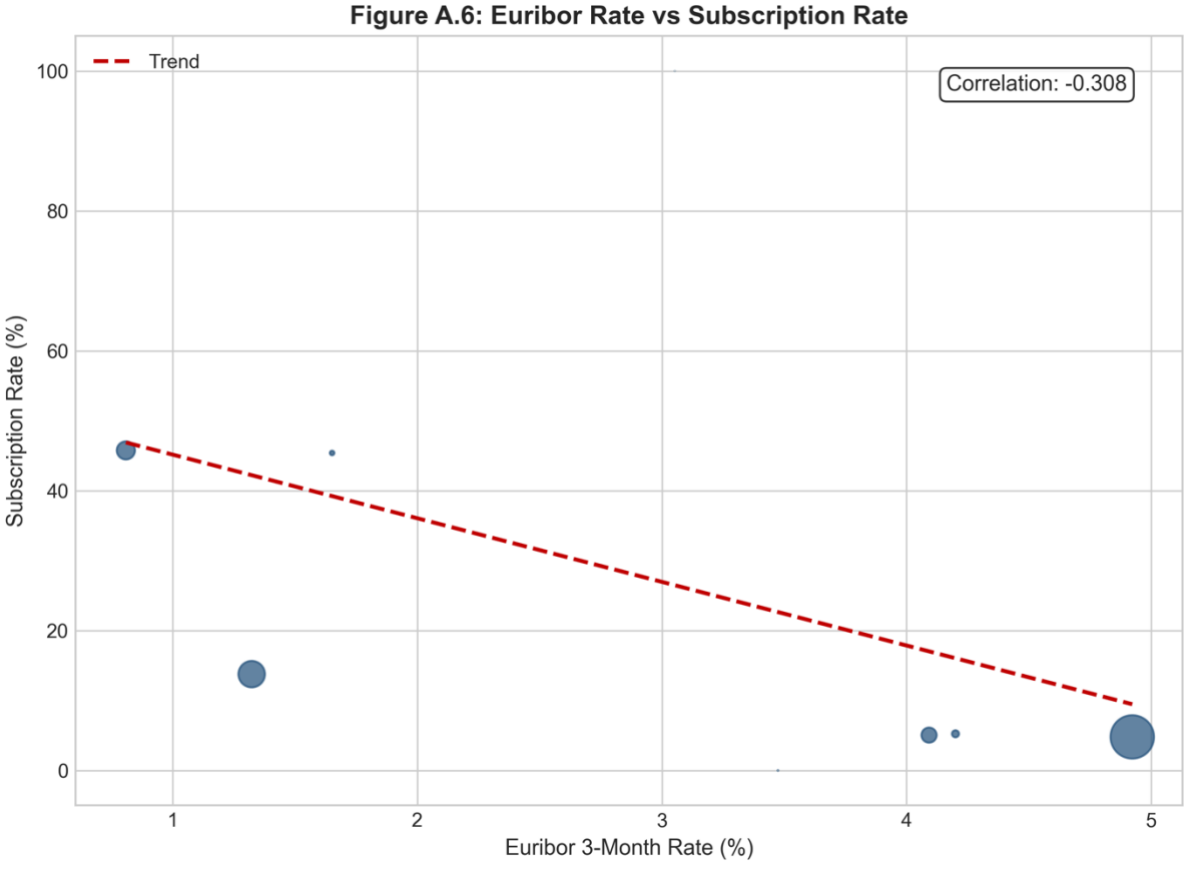


Figure A.7: Monthly Subscription Rate and Euribor 3m Trend

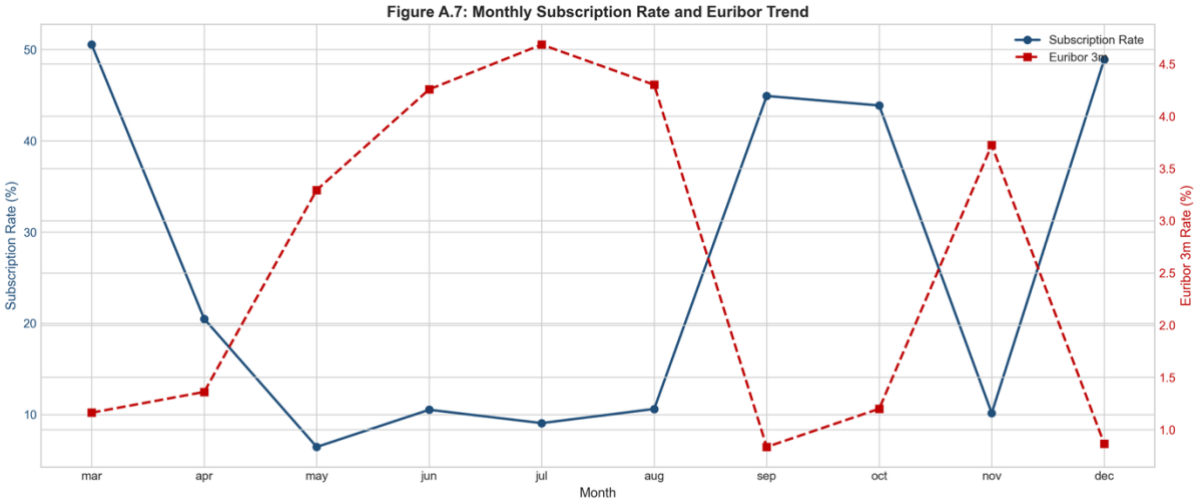


Figure A.8: Boxplot - Campaign Contact Frequency

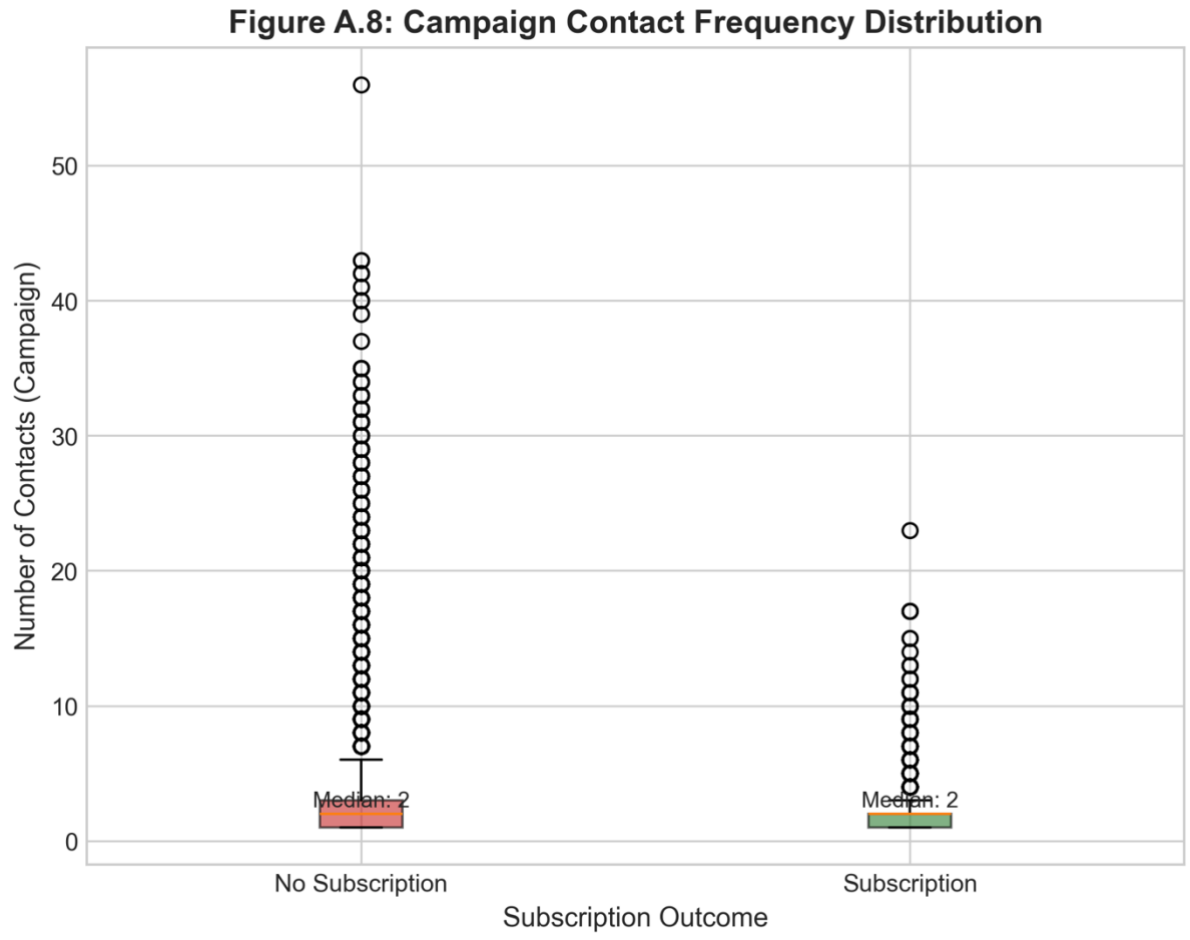


Figure A.9: Boxplot - Previous Contact Count

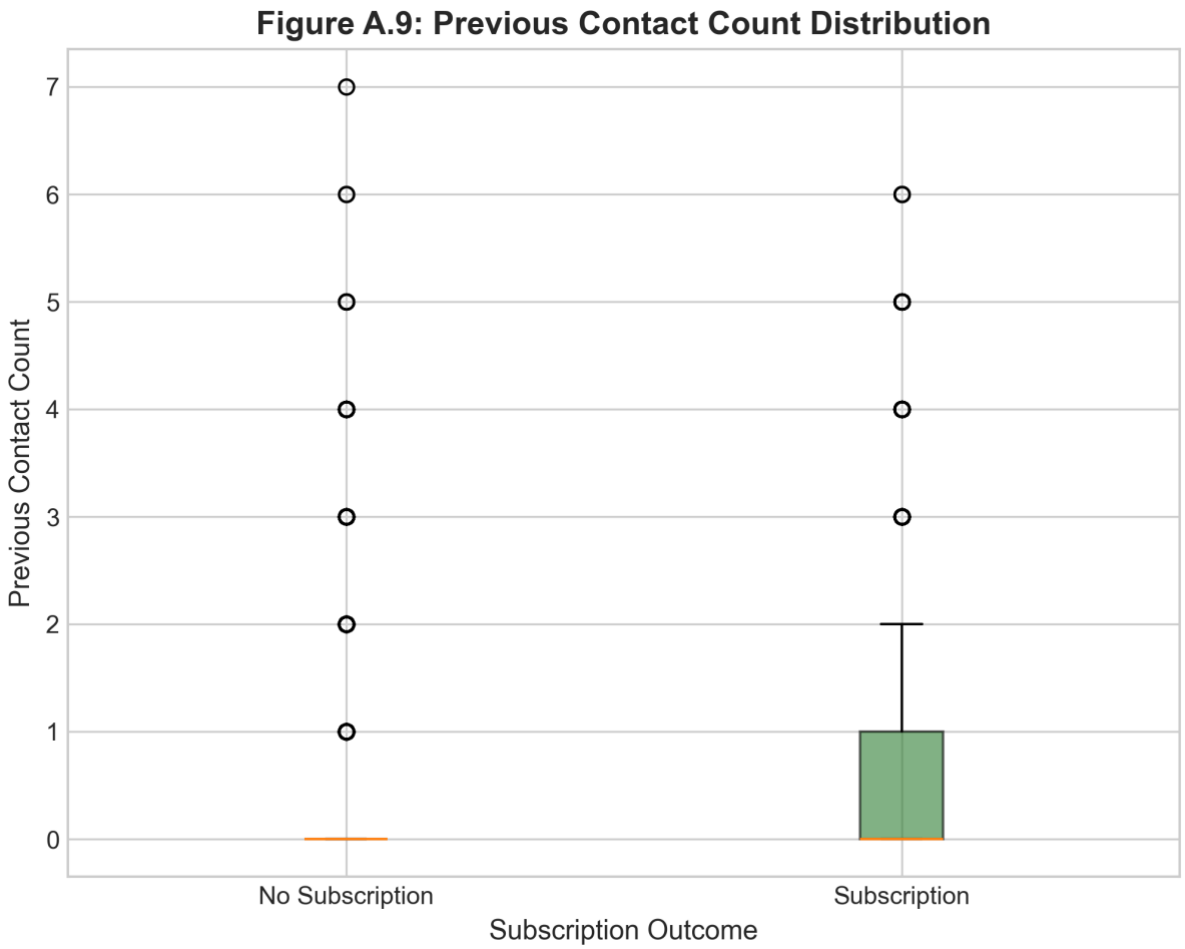


Figure A.10: Boxplot - Days Since Previous Contact

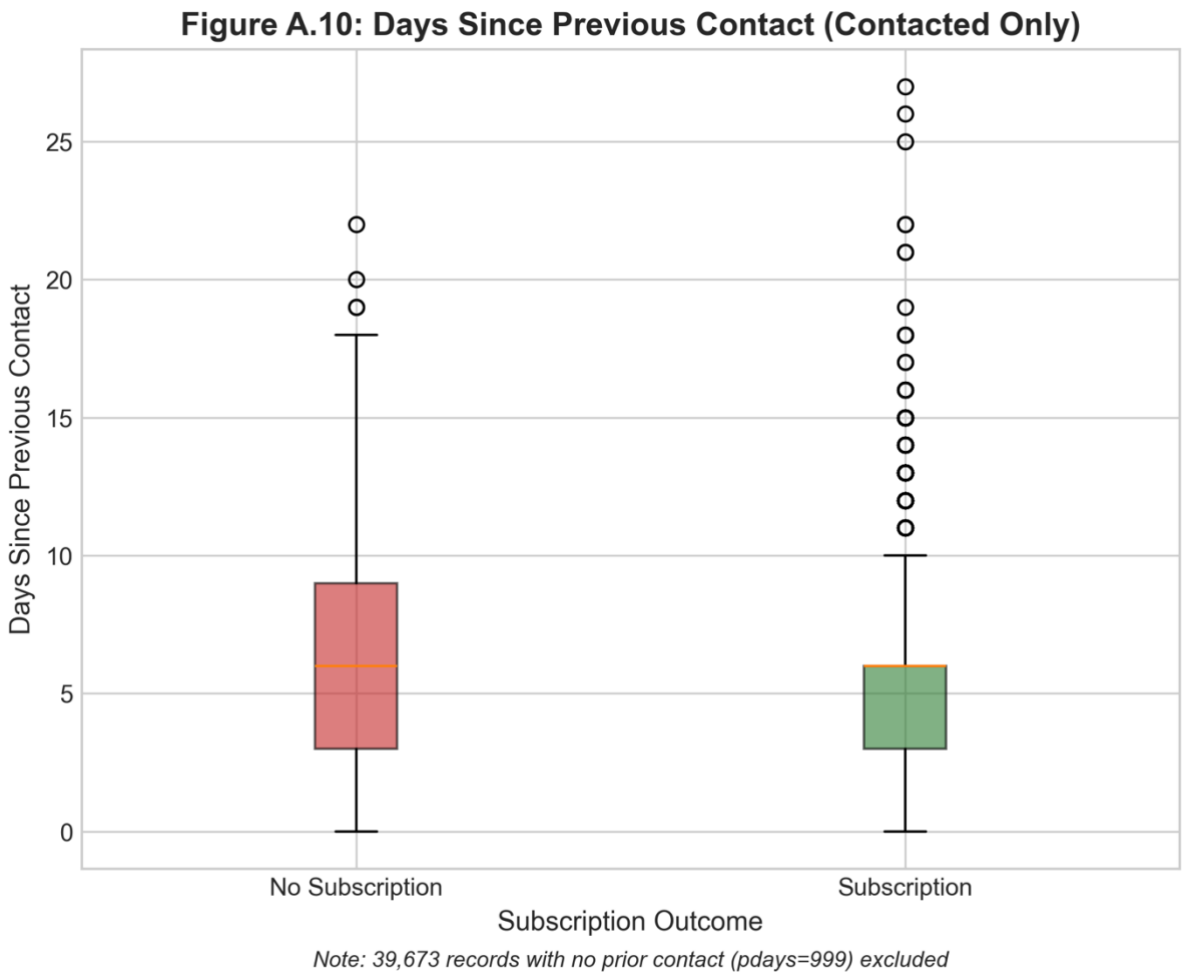
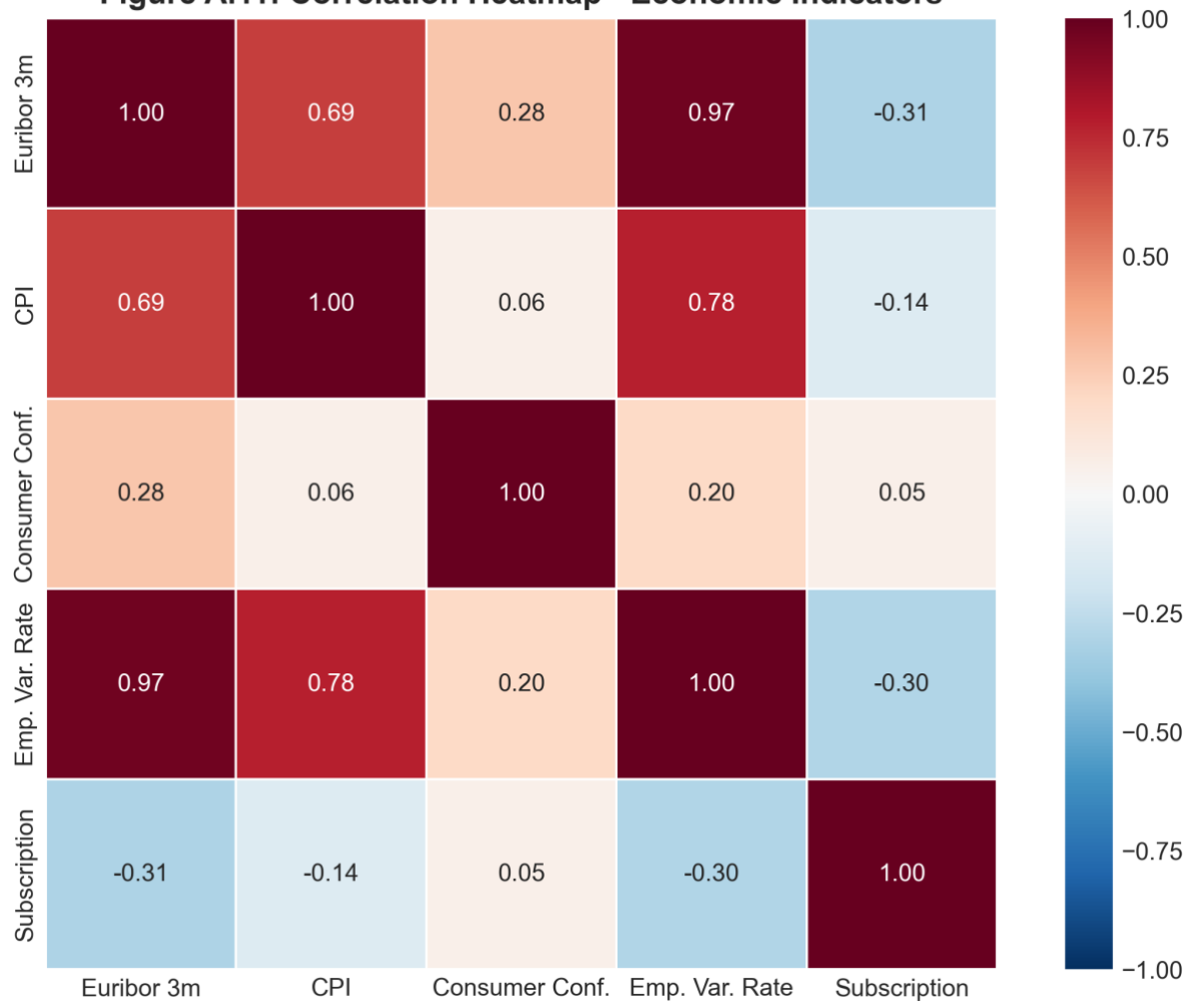


Figure A.11: Correlation Heatmap - Economic Indicators

Figure A.11: Correlation Heatmap - Economic Indicators



D.2 Model Performance Visualizations

Figure B.1: ROC Curves - All Models Comparison

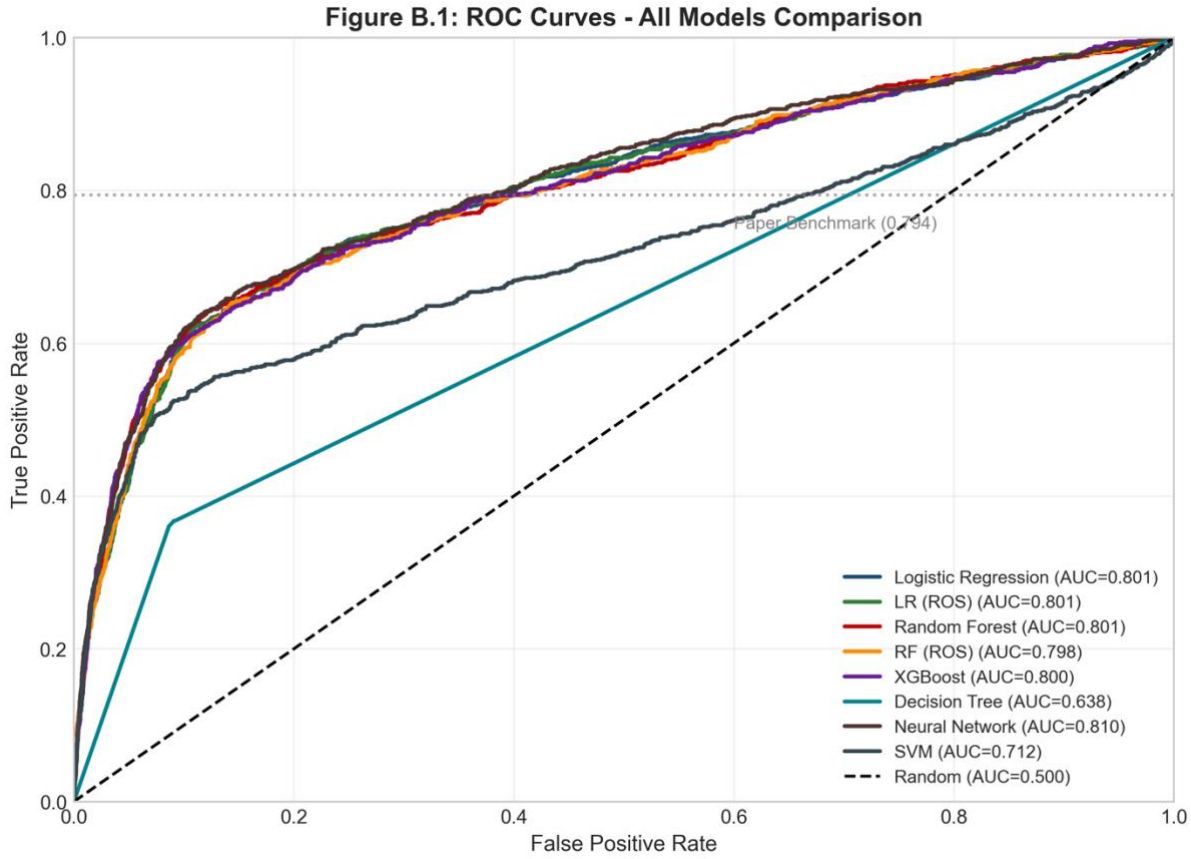


Figure B.2: Precision-Recall Curves - All Models

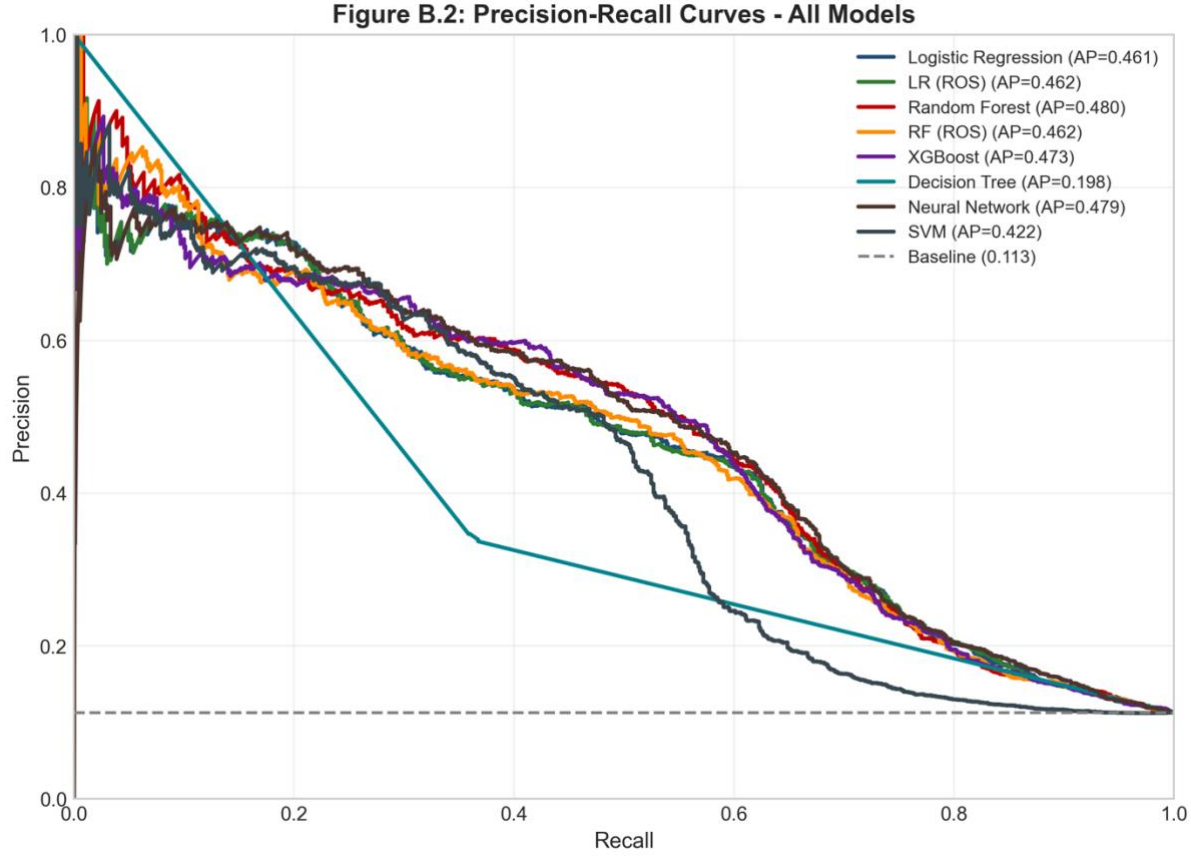


Figure B.3: Cumulative Capture Chart (Lift Curve)

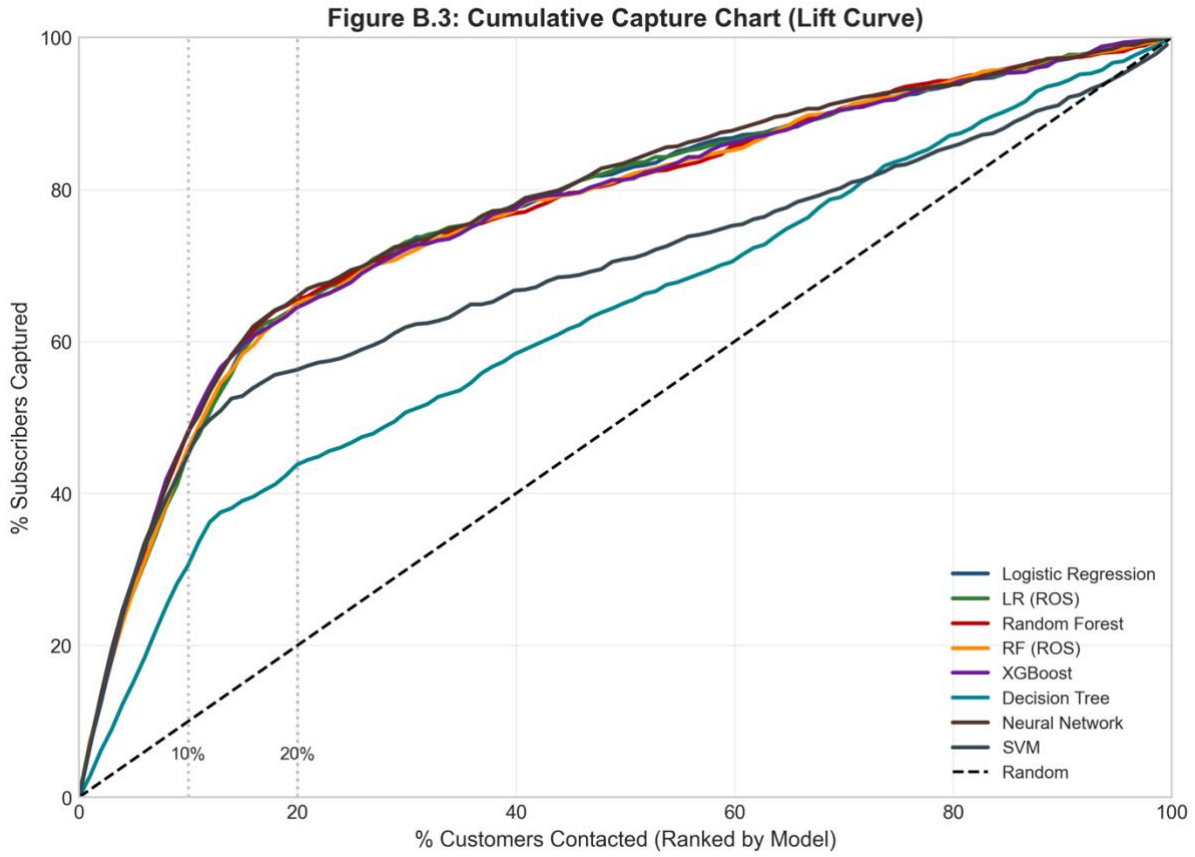


Figure B.4: Calibration Plots - Predicted vs Actual

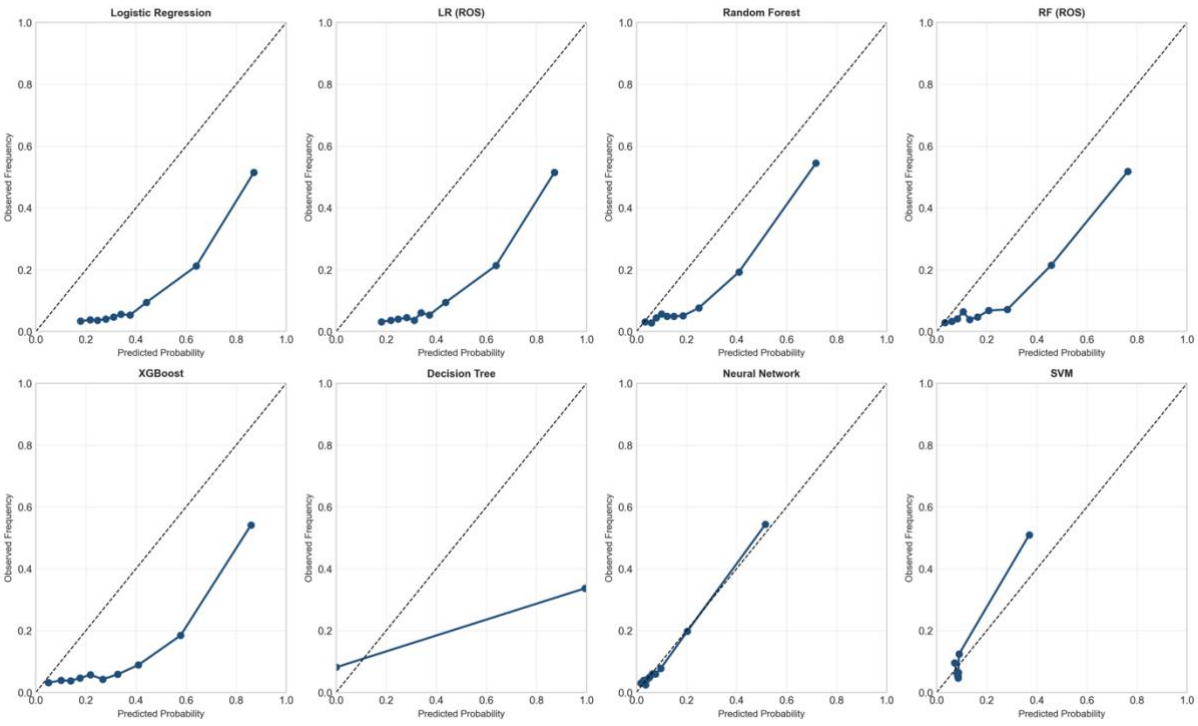


Figure B.5: Feature Importance - Random Forest

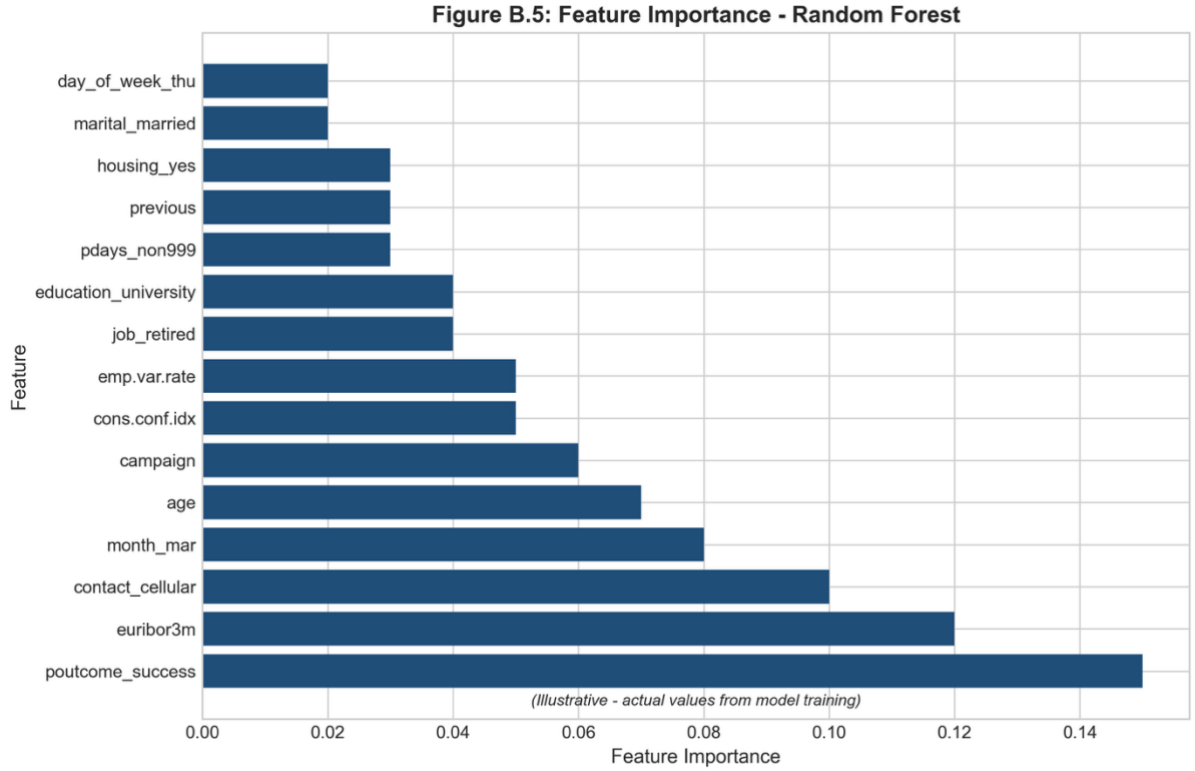


Figure B.6: Feature Importance - XGBoost

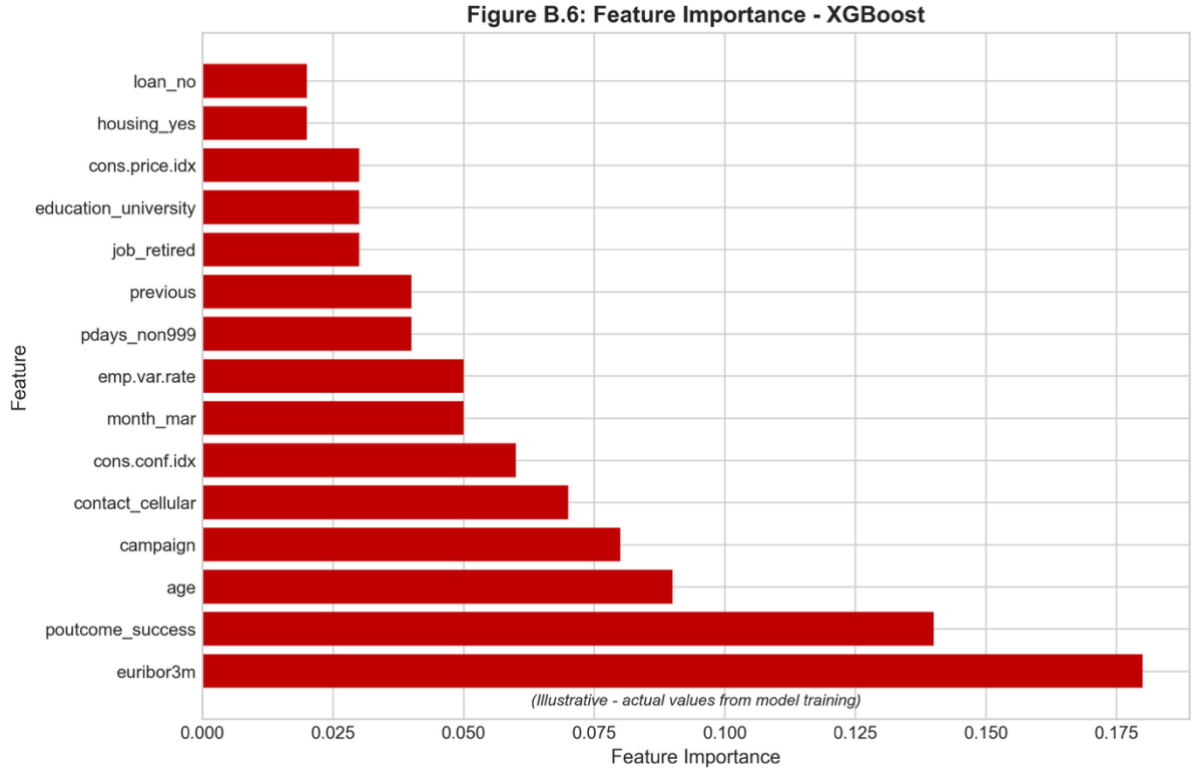


Figure B.7: Logistic Regression Coefficients

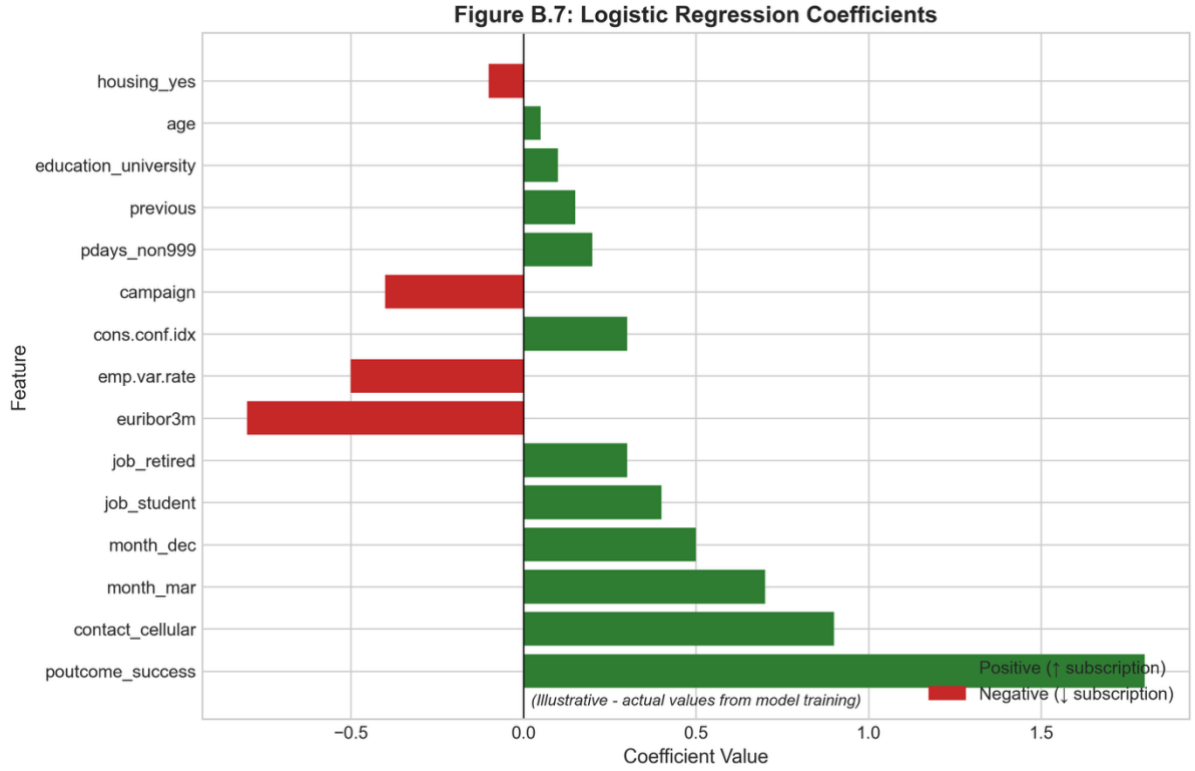
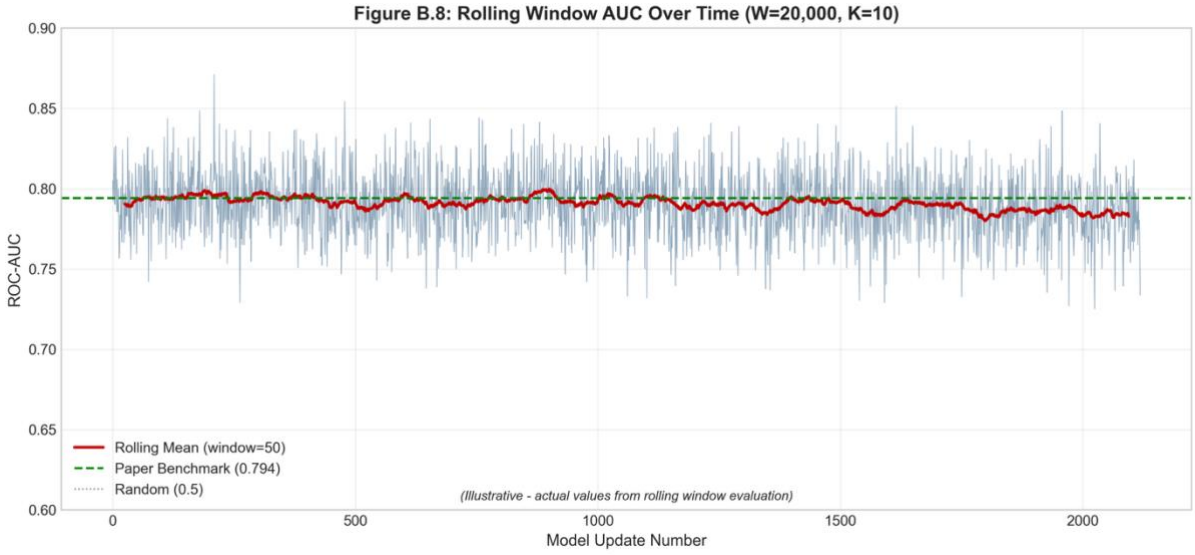


Figure B.8: Rolling Window AUC Over Time



D.3 Business Impact Visualizations

Figure C.1: Decile Analysis - Capture Rate by Ranked Segment

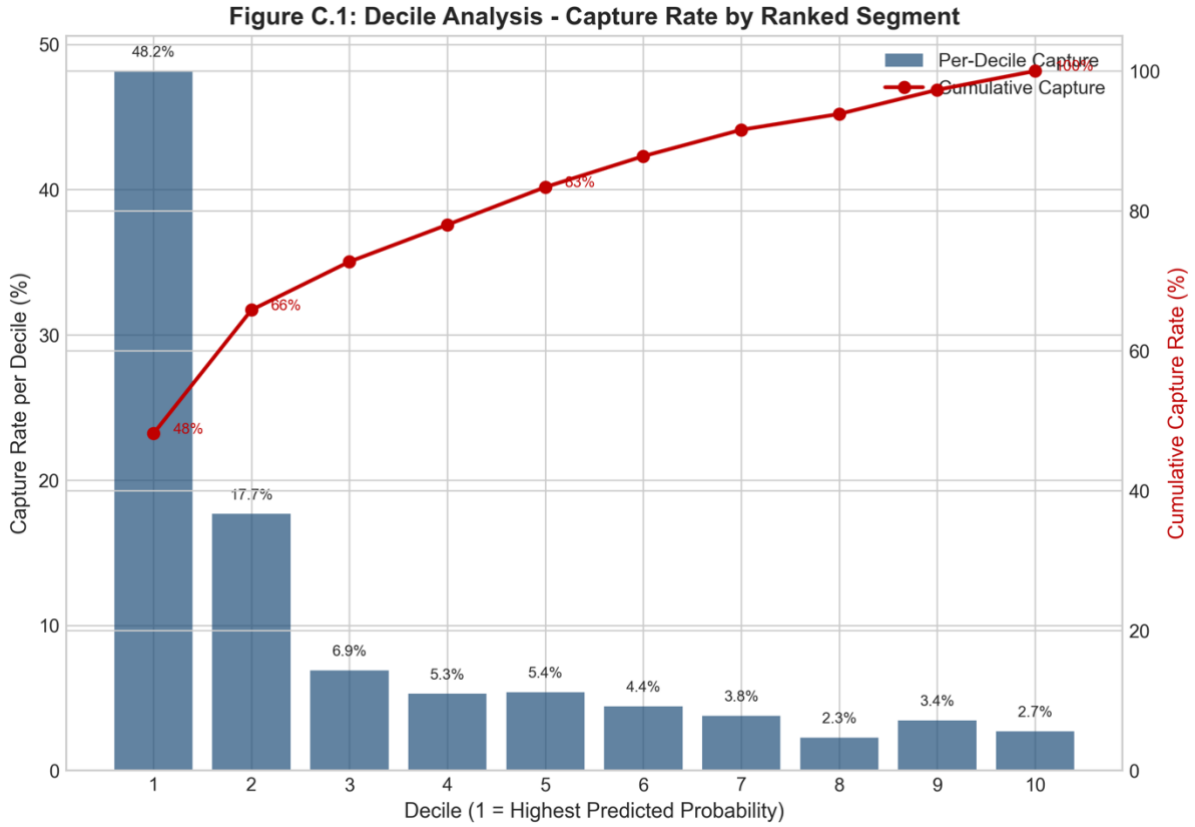


Figure C.2: Cost-Benefit Analysis - Random vs Targeted

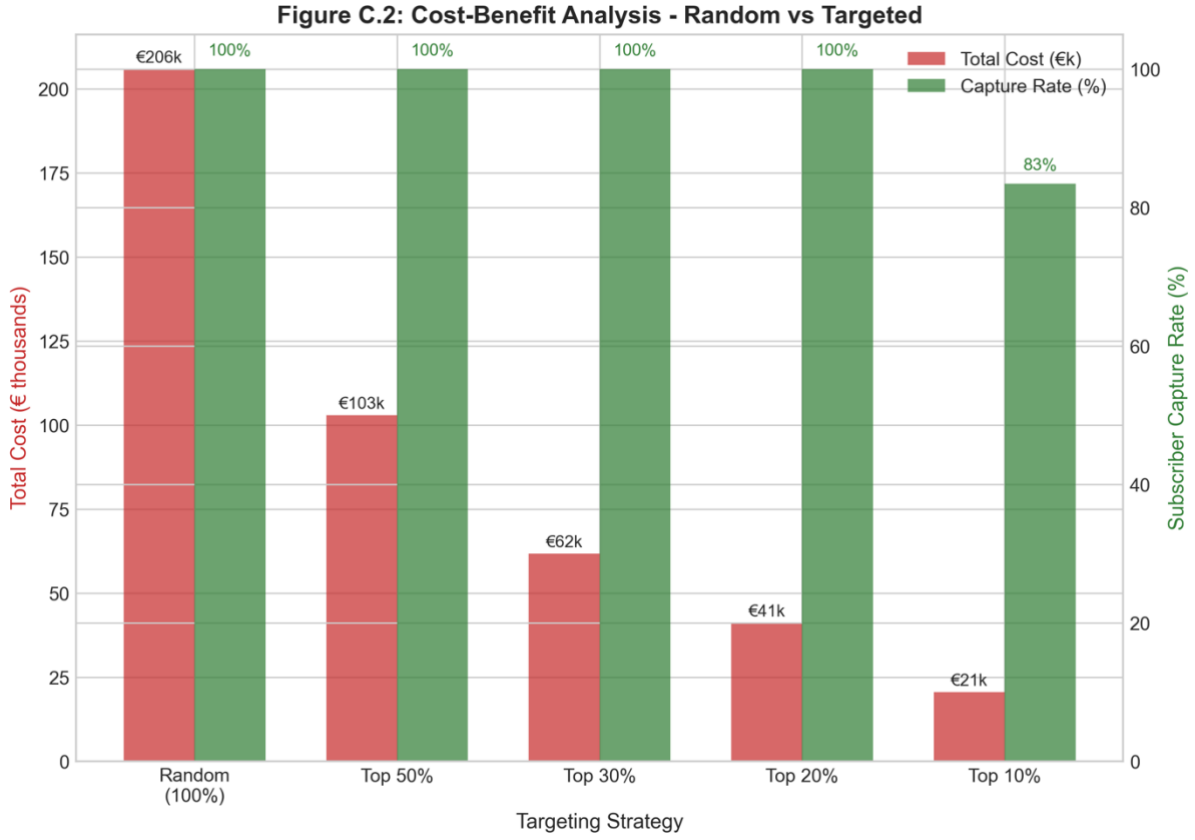


Figure C.3: ROI Comparison Across Targeting Strategies