DSCI 5260.002

Success Prediction of Bank Telemarketing

Business Analytics Capstone

By

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

Date: October 2025

Table of Contents

Abstract

This paper analyses the inefficiencies of the telemarketing campaigns carried out by a Portuguese bank institution for the promotion of term deposits and with a rejection rate of about 89% of the call attempts. A pre-call prioritization framework was created to predict if a customer would subscribe to term deposits before any outbound contact is made, so that the institution can optimize its resource allocation and improve overall campaign effectiveness.

The data analysis process followed the CRISP-DM methodology, which includes the steps of Business Understanding, Data Understanding, Data Preparation, Modeling, and Evaluation. Using a dataset of 41,188 customer contacts recorded between 2008 and 2013, in-depth exploratory data analysis was conducted, after which sufficient pre-processing procedures were carried out to eliminate data leakage, and the underlying model assumptions were reassessed with great care. Three machine learning models were used (logistic regression as a baseline, random forest, and XGBoost).

The resultant predictive models achieved an ROC-AUC value of 0.797 to 0.803. Using this model, it acquired 45-49% of all real subscribers, generating an efficiency increase of about 4-5 times over random calling. The temporal robustness of the original dataset was then examined by augmenting the dataset with more recent macroeconomic variables (2023-2025) taken from the European Central Bank and Eurostat; the augmented models were found to be stable in various distinctly divergent economic conditions.

Important factors influencing the probability of subscriptions in this study are previous campaign performances, macro-economic factors-particularly the Euribor interest rate, contact channel (cellular vs. landline), seasonality, and demographic factors. Furthermore, results of this research verify that a data-driven, focused prioritization approach is technically feasible and economically beneficial, estimating nearly $190,000 per 41,000 contacted campaigns while maintaining 65-70% of total conversions. A very detailed cost breakdown is provided in the Appendix.

1. Introduction

1.1 Project Context

Telephone marketing is a widely used marketing tool used by retail banks for marketing of financial products like term deposits. However, the success rate of such campaigns is often limited, and most calls do not lead to customers subscribing. This inefficiency leads to high operating costs, wasted staff time, and potential customer frustration with unwanted contact.

This project uses a dataset from telemarketing campaigns of a Portuguese bank from 2008 to 2013 from the UCI Machine Learning Repository. The data set consists of 41,188 customer contact records containing extensive information, data as demographics, financial status, campaign contact details, and macroeconomic indicators.

The main goal is to create a pre-call ranking algorithm that can rank high probability subscribers before any contact is made, using only the information available before the call. By using predictive modeling techniques, this research attempts to change the bank's telemarketing strategy from generalized, inefficient targeting to targeted, data-driven prioritization of customers. Transitioning to other channels has the potential to significantly boost campaign ROI, reduce operational costs, and improve customer experience by lowering unnecessary contacts.

2. Business Understanding

2.1 Business Problem definition

The core business issue is that of campaign inefficiency: roughly nine out of ten outbound calls are rejected by customers (89% no-rate vs. 11% yes-rate). This low conversion rate creates several challenges:

High labor costs: due to the non-productive time of the agents

• Poor allocation of marketing spending to low-probability prospects

• Customer dissatisfaction due to excess unwanted contact

• Failure to prioritize high-value opportunities. Poor allocation of resources in customer segments. The solution proposed is a pre-call customer ranking, which contains the contacts ordered by a subscription probability to maximize the concentration of efforts on the most promising ones and minimize the wasteful outbound calls.

2.2 Objectives

- Identify Top Subscription Predictive Drivers Perform variable

Analyzing and identifying factors that have the greatest impact on the likelihood of a customer subscribing. Compare feature importance from models to gain insight into the drivers of customer behavior.

- Develop a Pre-Call Ranking Model. Build and test predictive models to prioritize customers prior to contact for performance in both "Historic" (2008-2013) and "Modern" (2023-2025) macroeconomic conditions to ensure temporal robustness.

• Overcome Data Quality and Preparation Challenges

Apply strict preprocessing to manage missing values, outliers, multicollinearity, and possible data leakage while preserving information integrity.

- Assess Business Impact and Campaign Strategy Translate model results into business recommendations Find the right customer segments, contact channels, and time to target Estimate savings in cost and efficiency.
-
- Enhance Original Research by Using New Economic Indicators. Use modern macroeconomic variables instead of outdated macroeconomic variables, in order to investigate whether customer behavior and model performance are still consistent under current economic conditions.

2.3 Success Criteria

Technical Criteria :

The model is required to achieve a Receiver Operating Characteristic Area Under the Curve (ROC-AUC) score of around 0.80 and a competitive Precision-Recall Area Under the Curve (PR-AUC) to provide evidence of a balanced precision-recall trade-off in the context of a severe class imbalance (11% positive rate). All data transformations have to be implemented in a reproducible, leakage-free pipeline to ensure transparency and repeatability.

Business Criteria:

The model must be able to identify approximately 45-50% of the actual subscribers among the top 10% in the ranked list of customers, corresponding to a lift factor of between 4.5 to 5 relative to random call selection. This performance threshold captures realistic telemarketing budget constraints while providing explicit measures of campaign efficacy.

Operational Criteria:

The solution must exclude all post-call information, in particular call duration, apply techniques for interpretable pre-processing, and support continuous monitoring under changing macroeconomic conditions. These requirements will ensure the explainability, ethical integrity, and maintainability of the model in production environments.

Model hierarchy:

The Non-linear ensemble methods (e.g., Random Forests and XGBoost) are expected to match or improve the performance of Logistic Regression due to their ability to capture complex interaction effects.

Handling 'Unknown' values: Entries that are coded as unknown are not captured as completely missing data, but might permit coding of salient behavioral patterns (e.g., privacy consciousness). For this reason, such entries are retained as a separate category instead of being imputed.

Leakage prevention:

The duration variable is not included because it is available only after a call has completed. Its inclusion would artificially enhance model performance and destroy real-world validity.

Economic stability: By substituting historical macroeconomic indicators (2008-2013) for their contemporary equivalents (2023-2025) is not likely to change predictive patterns if customer behavior is structurally similar.

Model - specific preprocessing: Logistic Regression requires standardized numeric inputs, while tree-based models are invariant to scaling. Preprocessing pipelines must, therefore, reflect these methodological differences.

Sampling behavior: Oversampling techniques like Random Oversampling (ROS) and Synthetic Minority Oversampling Technique (SMOTE) are expected to increase recall at the cost of precision and calibration, and such techniques need careful alignment with business objectives.

3. Data Understanding - Exploratory Data Analysis

3.1 Objective and Dataset Overview

The exploratory data analysis was meant to discover important patterns, trends, and quality problems of the Portuguese Bank Marketing dataset obtained from the UCI Machine Learning Repository with 41,188 observations using 21 variables. This preliminary phase laid the groundwork for all the later preprocessing and modeling choices.
The dataset has three main types of information: Client Demographic and Financial Profile:
• age (numeric): Age of the client, 17 - 98 years, mean value 40.02
Exemplary job categories (numeric): Occupation categories such as administrative staff, students, retired, and unemployed persons.
Marital (categorical): Marital status of the client.
Education (categorical): Highest educational attainment.
Default (binary): Presence of a credit default (yes/no).
Housing (binary): Housing loan ownership (yes/no).
• loan (binary): Acquisition of a personal loan (yes/no)
• balance (numeric): Average annual balance in the bank.
Campaign Contact Information:
• contact (categorical): Means of communication, either cellular or telephone.Month(categorical): Month of last contact.
day_of_week (categorical): Day of the week on which the last contact was made.
• duration (numeric): Duration of the last call in seconds (mean: 258.29 seconds, approximately 6.4 minutes).
Campaign (numeric): Number of contacts (mean: 2.57) during the current campaign
• pdays (numeric): Days since a previous campaign's contact (-1 means no contact).
• previous (numeric): Number of contacts before the current campaign (mean: 0.17)
poutcome (categorical): Result of the last marketing campaign.
Economic Indicators:
• emp.var.rate (numeric): Rate of variation of the employment (mean: 0.08).
Available variables are: - cons.price.idx (numeric): Consumer Price Index (mean: 93.58)
• cons.conf.idx (numeric): Consumer confidence index (mean: -40.50).
euribor3m (numeric): Three-month Euribor interest rate (mean: 3.62%
• nr. employed (numeric): Number of employees in thousands (mean: 5167.04)

3.2 Target Variable Distribution & Class Imbalance

The target variable (y) has a strong class imbalance, with only 11.27% customers reported to have a subscription to term deposits ('yes') relative to 88.73% that declined ('no'). This 1:8 or so is a substantial modeling challenge. The distribution is shown in Figure 3.2 in the Appendix.
Implication: The imbalance requires stratified sampling, class-weighting, and the employment of suitable evaluation metrics (precision, recall, F1-score) instead of just the overall accuracy.

3.3 Descriptive Statistics

Some observations from the statistical summary:
The age distribution is approximately bell-curve centered around 40 years old (see Figure 3.3a).
More than half (55%) of clients were contacted 2-3 times during the campaign.
The majority of clients (indicated by pdays = 999) had not been contacted before (see Figure 3.3b).
As edited by Ibrahim H. DoganPleasease review "Economic indicators show a high level of intercorrelation, especially between nr. employed and euribor3m (correlation = 0.945, see Figure 3.3c)."

3.4 Data Quality: Missingness and Patterns of Unknown

Analysis of 'unknown' responses (Figures 3.4a and 3.4b) reveals non-random patterns:
Uncovering the Undisclosed Information of Information; Statistics and Data on Income, Taxation, Education, Default, Housing, Loan, Marriage, Divorce, etc.
Landline (telephone) contacts have a higher rate of missing values than cellular contacts.
Implication: The non-random missingness is indicative of underlying behavioral clusters (e.g., privacy-conscious or risk-averse customers). Rather than filling in missing values, 'unknown' as a separate category may retain these informative patterns.

3.5 Outlier Detection & Distribution Skewness

Boxplots of Campaign, Previous and pdays_non999 (Figures 3.5a - 3.5c) show large right skewness: •
Campaign: Some customers received more than 40 calls (where the median was 2).
Previous: Most clients had no previous contacts, with a handful having 7.
pdays_non999: Some large gaps between contacts, resulting in a long right tail.
Implication: The heavy right tails are the reason for Winsorization at 1st/99th percentiles to avoid distortion of linear models and to gain robustness.

3.6 Univariate Analysis: Patterns by Category

3.6.1 Job Type Performance (Figure 3.6a)

Among segments with numbers above 200 observations, the following conversion rates were found:
Student: 31.4% conversion rate (875 total contacts)
Retired: 25.2% conversion rate (1,720 contacts)
Unemployed 14.2% conversion rate (1,014 contacts)
Admin: 13.0% Conversion rate (10422 contacts - largest segment)
Blue-collar: about 8% conversion rate (lowest among major segments)
Insight: Students and retirees have much higher levels of responsiveness, perhaps because they have time and other financial priorities.

3.6.2 Impact of Education Level (Figure 3.6b)

• Unknown: 14.5 % (1,731 contacts)
• University degree: 13.7 % (12,168 contacts)
• Professional course: 11.4% (5,243 contacts)
Insight: Higher education is correlated with above-average subscription rates, indicating segmentation opportunities for educated or privacy-conscious clients.

3.6.3 Contact Channel Effectiveness (Figure 3.6c)

Cellular: 14.7% conversion (3853 subscriptions for 26144 contacts)
Telephone (landline): 5.2% conversion (787 subscriptions for 15,044 contacts)
Insight: Mobile outreach is about 3* more effective than landline contact, suggesting there is an obvious channel preference, and perhaps more engaged customers.

3.6.4 Timing and Seasonality of Campaigns (Figure 3.6d)

Months of high performance: March, September, December (40 - 50% conversion rates). Low-performing months: Summer months (10-15 percent conversion rates)
Highest volume: May had the most customer contacts
Insight: Campaigns run early or late in the year have dramatically higher success rates, indicating best practices quartered end-of-year for financial planning periods.

3.7 Previous Campaign Outcome (Figure 3.7)

• Success: 894/1 373 yields 65.1 % subscribed

Failure: roughly 15 % or less subscribed

Nonexistent: 3,141 for 35,563 results in 8.8% subscribed (first-time contacts)

Insight: Previous campaign success (poutcome = 'success') is the most important categorical predictor; conversion rates are almost 7.5 times higher than first time contacts. This is a validation of the importance of the relationship history in predictive modeling.

3.8 Macroeconomic Context and Interest Rate Sensitivity

3.8.1 Euribor Rate Impact (Figures 3.8a and 3.8b)

There is a negative correlation between the interest rates and the amount of term deposits: Lowest quintile (0.63-1.30%): 29.7% conversion (8,636 contacts)

Highest quintiles of conversion: 3.6% conversion: 4.19 - 4.86%

Putting monthly conversion rates on top of average Euribor3m shows a high negative correlation (about -0.8) post-crisis, times when Euribor then fell, subscription levels peaked.

Insight: The macroeconomic environment has a significant impact on deposit-seeking behavior. In a low-interest-rate environmentencouragesge customers to seek the security provided by term deposits.

3.8.2 Economic Indicator Multicollinearity (Figure 3.8c).

Economic measures have veryhigh hiinter-correlationrel;tioncorrelation betweenetand ween n ber umber n nr.employeEuribor3mibor3m is 0.945, and there is a strong relationship between employment variation rates, consumer price indices and Euribor.

Implication: Redundant economic feature content is an indication that only a subset of the features are necessary to provide adequate predictive power and avoid multicollinearity.

3.9 Summary of Key Insights

The table below summarizes the main conclusions and their implications for modelling: Aspect | Insight | Modeling implications:

Aspect	Key finding	Modeling implication
Target imbalance	Yes: 11.27%, no: 88.73% (≈1:8 ratio)	Used stratified splits, class weighting, and precision/recall metrics
Missing values	Non-random patterns by job/contact type	keep "unknown" as informative categories
Outliers	Heavy right tails in campaign,previousand pdays	Apply Winsorization at 1st- 99th percentiles
Channel & timing	Cellular +3× better; Mar/Sep/Dec being peak months	Optimizing campaign timing and channel allocation
Demographics	Retired (25%), students (31%), educated (14%) most likely to convert	Focuse high-response segments
Prior success	65% conversion if previous campaign succeeded	Prior outcome is strongest categorical predictor
Macroeconomics	Low Euribor → High conversion (correlation: –0.8)	Integrate economic signals; time campaigns with favorable conditions
Feature redundancy	Economic indicators highly correlated (0.945)	Consider feature selection or dimensionality reduction

3.10 Conclusion

As a result, the exploratory data analysis phase revealed a very rich and challenging dataset, which is severely class-imbalanced, has non-random missingness patterns, and is highly sensitive to macro-economic conditions. The analysis showed excellent possibilities for specific marketing in terms of the choice of the best communication channel (cellular), the right point in time (early vs. late in the calendar year), demographic segmentation (retirees, students, educated clients), and the use of previous relationship history. The high correlation of these response rates with the economic conditions

emphasizes the need to include macro-economic features in the predictive models. Preprocessing steps to be executed in the future should include robust scaling to handle outliers, informative missing-data imputation and sampling/evaluation strategies to address potential class imbalance.

3.11 Assumption Testing and Validation

To assure the reliability of the modeling approach that would follow and the validity of the analytical approach, we directly tested seven key assumptions that arose from our exploratory analysis. These tests verify that the methodological choices taken during the phases of data preparation and modeling are correct.

3.11.1 Leakage Prevention and Model Performance

Assumption A1:

Logistic Regression (pre-call) should do close to the original study results once leakage (duration) is eliminated.

Rationale: Moro et al. (2014) published an exceptional performance (AUC ~ 0.90) by incorporating the call duration as a feature, which is only available after the call has been recorded. We hypothesized that if duration was excluded, the resulting ranking system would have an AUC of around 0.75-0.80 and have sufficient "'practical" business value to be deployable for pre-call ranking.

Results: The historic dataset got ROC-AUC=0.797, PR-AUC=0.451, Lift@10%~4.3. The modern dataset resulted in ROC-AUC=0.801, PR-AUC=0.461, Lift@10%~4.4x. Performance was unaffected on temporal contexts by removing the most predictive feature.

AUC=0.461, Lift@10%~4.4x. Performance was unaffected on temporal contexts by removing the most predictive feature.

Validation Status: PASS. Logistic regression was validated as an effective and stable baseline model. The effect on AUC is known to be expected and acceptable (a 15-20% reduction from 0.90 to 0.80, on average) due to the removal of the post-call information. Lift@10% of 4.3-4.4, 45-48% of real subscribers are in the top ranked 10% of customers: This is extremely commercially valuable. Implication: The model is suitable for pre-call prioritization deployment. Banks can be confident that this can help them rank customers before they get on the phone, resulting in 4-5x efficiency over random calling.

Implication: The model meets the deployment requirements for pre-call prioritization Banks can therefore confidently use this method to triage customers before contacting them, gaining four to five times the efficiency of random calling.

Assumption A2:

It is hypothesized that ensemble techniques (i.e. Random Forest and XGBoost) will achieve or improve over logistic regression with respect to ranking performance.

Rationale: Non-linear algorithms are expected to capture more complex interactions and variable relationships than linear models and may lead to improved discriminative power and ranking performance.

Results: When used on past data, XGBoost outperformed both logistic regression and Random Forest (AUC = 0.803, PR-AUC = 0.478, Lift = 4.8 times), validation of the contention that gradient-boosting methods provide better discrimination. In contrast, Random Forest performed poorly in comparison to logistic regression on all metrics (AUC between 0.764 and 0.782; PR-AUC between 0.397 and 0.418) facts which go against our expectations. In addition, both ensemble methods showed a higher sensitivity to current macroeconomic conditions in comparison with the stability of logistic regression.

Validation Status: The results are mixed. While XGBoost provides the highest cumulative performance, the logistic regression is proven to be the most robust model, due to its stability, interpretability and its relatively low performance degradation under different economic scenarios. Implication: XGBoost is recommended for highest performance, logistic regression is recommended when interpretability and robustness are more important considerations.

3.11.2 Data Quality and Handling Missing Data

In *Assumption A3,* we treat 'unknown' values as a valid category instead of just missing data.

Rationale: Responses marked as 'unknown' show consistent patterns related to specific job roles and methods of contact. They're not just absent data; they might actually reveal behavioral traits like a privacy concern.

Results: When we kept these 'unknown' categories, our models maintained stable PR-AUC without any negative effects on calibration. Customers with an 'unknown' education level subscribed at 14.5%, compared to an overall average of 11.3%, which is a 3.2 percentage point rise. This indicates that having an 'unknown' status actually carries predictive value, rather than just being random missing data. We found no signs of introducing any bias by keeping these unknown categories.
Validation Status: PASS. We found that 'unknown' values provide useful predictive signals and should remain as a separate category.
Implication: We should keep treating 'unknown' as its own level in our categorical features and watch out for any potential bias in production.

Assumption A4:
We think that leaving out call duration helps avoid data leakage and keeps things deployable before a call.
Rationale: Call duration is a strong predictor (with a correlation of about 0.40 with subscriptions), but we can only measure it after the call ends. If we include it, the performance metrics would be unrealistic.
Results: By excluding call duration, ROC-AUC dropped by about 0.10 (from 0.90 to 0.80), which aligns with its importance in the original study. All other features are available before the call. The models we got still performed at a business-acceptable level (Lift@10% around 4.5×) without any artificial boosts.
Validation Status: PASS. We've permanently excluded duration to keep the integrity of the model.
Implication: The 10% drop in AUC is a fair trade-off to ensure the model can be used in real-world applications.

3.11.3 Economic Context and Temporal Robustness

Assumption A5:
The use of modern macroeconomic indicators from 2023 to 2025 won't negatively impact performance when swapping in for the data from 2008 to 2013.
Rationale: The initial dataset we collected was from the period right after the 2008 financial crisis. By testing against today's economic conditions, we can see if the predictive trends hold up.
Results: The metrics were almost the same across both Historic and Modern datasets. Logistic Regression (LR) had a +0.004 change in ROC-AUC, Random Forest (RF) had a -0.018 change, and XGBoost (XGB) had a -0.031 change. All of these shifts were under 0.05 AUC.
Validation Status: PASS. Even with the striking differences between the economic environments (post-crisis austerity versus post-pandemic inflation), the connection between customer traits and term deposit subscriptions stayed structurally stable.
Implication: This confirms that the model can be generalized and eases worries about changes over time. Banks can feel confident using models trained on Historical data, though it's still a good idea to retrain them periodically.

3.11.4 Preprocessing and Model-Specific Requirements

Assumption A6:
Tree-based models don't need scaling, while Logistic Regression (LR) does require standardized numeric inputs.
Rationale: Decision trees make splits based on absolute thresholds and aren't affected by changes in scale, while linear models assign importance based on the magnitude of features, so they work better with standardized data.
Results: The coefficients for LR kept consistent magnitudes and signs across different datasets when we used the scaler fitted on historical data. The performance of RF and XGB wasn't influenced by input scale. Standardization sped up LR convergence by about 30%.
Validation Status: PASS. We kept the specific preprocessing for each model type to ensure interpretability and performance remain consistent.
Implication: The coefficients from LR can be directly compared between the Historic and Modern models. Tree models can maximize efficiency by skipping unnecessary scaling transformations.

Assumption A7:

Oversampling techniques (like SMOTE and ROS) change the precision-recall balance, and the effects vary based on the model type and its initial configuration.
Rationale: With a class imbalance (only an 11% positive rate), models might under-predict the minority class. Oversampling methods are meant to balance the training distribution artificially, which could boost recall for the minority class but might hurt precision for the majority class and affect probability calibration.
Testing Methodology:

- We compared the baseline models (using class_weight='balanced') with their oversampled counterparts: LR with SMOTE-NC (which is for categorical-aware synthetic oversampling) and RF using RandomOverSampler (ROS, which duplicates samples).
- We looked at changes in Precision and Recall when the threshold was set at 0.5, along with ROC-AUC, PR-AUC on the test set, the Brier Score to check calibration quality, and the calibration curves.

Results:
For Random Forest:

- Oversampling raised recall by 11.9 percentage points (from 27.5% to 39.4%) but dropped precision by 7.4 points (down from 56.5% to 49.1%).
- Calibration took a hit, with the Brier Score going up from 0.082 to 0.090.
- The PR-AUC fell from 0.418 to 0.371, which means the ranking quality got worse on the natural test distribution.

For Logistic Regression:

- SMOTE had the opposite effect here: precision went up (from 32% to 42-45%) while recall dropped (from 68% to 51%).
- This happened because the baseline Logistic Regression, using class extunderscore weight='balanced', was already tuned for high recall, and SMOTE made it more cautious.

Validation Status: PASS (Expected Trade-off Confirmed) — The effects of oversampling differ based on model structure and the initial setup. In this telemarketing scenario, using class weighting as a baseline is better because: (1) cold calls cost money (both agent's time and potential annoyance for customers), (2) precision is more important than just trying to catch every call, and (3) getting the calibration right is crucial for decisions that depend on thresholds.
Implication: Oversampling should be applied selectively based on business needs. It's best to stick with class weighting unless the campaign strategy clearly focuses on maximizing recall. The different outcomes seen with Logistic Regression compared to Random Forest emphasize the need to understand how each model behaves before applying oversampling methods.

3.11.5 Summary of Validated Assumptions

All key assumptions were either confirmed or their deviations were clearly understood and factored into the model selection process:
· A1 (LR baseline performance): ROC-AUC around 0.80, Lift approximately 4.3-4.4× - This shows Logistic Regression is a solid, dependable baseline.
· A2 (Ensemble superiority): XGBoost outperforms LR and RF - XGBoost is the best option; Logistic Regression is the most reliable in various situations.
· A3 ('Unknown' as meaningful): There's a significant predictive signal (χ^2 $p<0.001$) - 'Unknown' categories should be kept.
· A4 (Duration exclusion needed): An AUC drop of 0.10 is acceptable for deployment - Duration should be excluded for good.
· A5 (Modern macro robustness): Performance remains stable (less than 0.05 change in AUC) - The model holds up across different economic conditions.
· A6 (Model-specific preprocessing): Consistent interpretability is maintained - Keep using family-specific pipelines.
· A7 (Oversampling trade-offs): Recall goes up, precision goes down, and calibration gets worse - Class weighting should be the default approach.

3.11.6 Implications for Modeling and Deployment

The assumptions we've validated lead us to some important methodological choices:

- **Model Selection:** We've opted for XGBoost for its high performance, while Logistic Regression is chosen for its interpretability and stability.
- **Preprocessing:** We're using family-specific pipelines that ensure feature engineering is done without any leakage.
- **Imbalance Handling:** For this situation, class weighting seems to work better than resampling.
- **Economic Context:** Current macro data shows that the models are stable and can be deployed in both contexts.
- **Feature Engineering:** We're keeping 'unknown' categories intact but have permanently removed the duration feature.
- **Performance Benchmarks:** Our realistic targets are a Lift@10% of around 4.5 times and an ROC-AUC of about 0.80.

4. Data Preparation

4.1 Leakage Policy

What changed: We dropped the 'duration' column.

Why? The duration of a call can only be known after the call takes place, and it's a strong predictor of subscriptions. Including it in pre-call modeling would result in data leakage, giving us an unrealistic edge. By removing it, we ensure the model uses only the information available beforehand, making our results more reliable and applicable to real-world scenarios.

4.2 Sentinel Handling (pdays)

What changed: We added two new features:

- **contacted_before (binary):** Indicates whether the client has been contacted before (1=Yes, 0=No).
- **pdays_non999 (numeric):** Represents the actual number of days since the last contact, with 999 replaced by NaN for tree-based models or -1 for Logistic Regression.

Why? The value of 999 in pdays doesn't provide a real measurement of days since last contact; it just means the client was never contacted. Treating it as a number would skew the distribution. By separating it, we can capture both the history of prior contact and the actual days gap when relevant. Models like XGBoost can handle missing values (NaN) effectively, while Logistic Regression and Random Forest get -1 as an imputed value.

4.3 Rare-Level Grouping

What changed: We combined categorical features (like job and marital status) that had less than 0.5 percent of records into an 'Other' category, but retained 'unknown' as its own category.

Why? Rare categories can introduce noise and lead to sparse, unstable one-hot encoding. Grouping them enhances model stability while allowing us to keep 'unknown' distinct, as it might provide useful insights (for instance, if people are hesitant to share their education or credit status).

4.4 Winsorization

What changed: We applied winsorization (capping) at the 1st and 99th percentiles for highly skewed numeric features (campaign, previous, pdays_non999).

Why? Some customers might have been called many times (like 56 calls) or have huge lag values. These rare extremes can throw off linear models. Winsorization mitigates their impact without deleting those records, which makes the dataset more robust without losing valuable information.

4.5 Multicollinearity Control

What changed: We've dropped nr.employed due to its high correlation with euribor3m (around 0.945, as illustrated in Figure 3.8c), while keeping other macroeconomic indicators.

We're talking about variables like emp.var.rate, cons.price.idx, and cons.conf.idx here.

So, what's the deal? Well, multicollinearity happens when there are strong correlations between variables, and it can mess up our models and make coefficients hard to interpret. To tackle this, we ditch the redundant variables and keep the important macroeconomic indicators.

4.6 Encoding, Imputation, and Scaling

Encoding: We used One-Hot Encoding to change all categorical variables into binary flags so machines can read them easily.
Imputation: For missing numeric values (like from pdays_non999), we replaced them with -1 for models like LR/RF, while advanced models like XGBoost just left them as NaN.
Scaling: Standardizing the data puts everything on the same scale, which is great for models sensitive to differences in magnitudes, like Logistic Regression.
These steps help keep the dataset clean, consistent, and ready for machine learning.

5. Updating with Modern Economic Data

5.1 Rationale

The macroeconomic data from 2008 to 2013 feels outdated at this point. We really need to swap in 2023-2025 indicators to see if the customer targeting strategies from back then can still work in today's economic climate (interest rates, inflation, employment, consumer confidence).

5.2 Strategy

- Collect monthly macro data from the ECB and Eurostat databases.
- Create two scenarios: one Historic (with original macros) and one Modern (where we replace them with 2023-2025 values).
- Run both through the modeling pipeline and compare the rankings and capture rates.

5.3 Validation and Monitoring

We'll use a time-sensitive validation approach: train on the older data and test with the newer data.

5.4 Merging Modern Macroeconomic Data (2023-2025)

We updated the four macro factors from the original UCI dataset with recent time-series data from ECB/Eurostat:

- Euribor 3-month (monthly, percent per annum): from ECB Statistical Data Warehouse series FM.M.U2.EUR.RT.MM.EURIBOR3MD_.HSTA
- Consumer confidence (monthly, balance): Eurostat 'Consumers confidence indicator' (ei_bsco_m)
- HICP price index (monthly index): Eurostat HICP (prc_hicp_midx)
- Employment variation (quarterly percent change): Eurostat Quarterly National Accounts guidance

To merge the data, we matched the keys and aligned the frequency: since the bank records are at the contact level and only show the month, we matched the calendar month and quarter, not specific dates. We matched the monthly series (Euribor, consumer confidence, HICP) directly by month, and for the quarterly employment data, we used forward-filling within each quarter.
Two scenarios came out of this:

- Historic Dataset: Leaving the original UCI macro columns as they are (2008-2013).
- Modern Dataset: Replacing those columns by linking the contact's month to the corresponding month in the 2023-2025 panel.

After merging, we ran checks and confirmed a 100% fill rate for the four macro columns. Each model type got specific formatting: RF/trees kept the macro values as raw (since they're scale-invariant), while LR standardized all numeric features using the Historic mean and standard deviation to ensure that coefficients could be compared.
For validation, the LR AUC/PR and Lift@10% scores for Historic vs. Modern were almost identical, which means that updating the macro context hasn't hurt ranking performance, and we can now test present-day scenarios.

6. Modeling - Logistic Regression

6.1 Why We Chose This Model

We went with Logistic Regression (LR) as our base model because it matches up well with the 2014 study we're referring to, and it's one of the easiest methods to interpret for binary classification. It's commonly used in areas like banking, credit risk, and marketing since it allows for straightforward interpretation of coefficients and gives us well-calibrated probabilities. By using LR, we also set a solid benchmark against the original research to see if newer preprocessing methods and updated macroeconomic data change anything.

6.2 Dataset Setup and Preprocessing

We looked at two datasets:

- **Historic dataset**: This contains the original macroeconomic variables from 2008-2013.
- **Modern dataset**: This one has updated indicators from 2023-2025 from the ECB/Eurostat.

Both datasets went through the same preprocessing steps:

- Handled sentinel values by splitting pdays = 999 into derived features.
- Winsorized highly skewed variables at the 1st/99th percentiles.
- Grouped rare levels for categorical features.
- Removed multicollinear features, like nr.employed.
- Scaled all numeric features using the Historic scaler to keep the coefficients comparable.
- Left out call duration to avoid pre-call leakage.

6.3 Training the Model

We used scikit-learn's LogisticRegression framework with class_weight set to 'balanced' to tackle the roughly 11% imbalance in the positive class.

For the split, we did a stratified 80/20 train-test division to keep the class proportions intact.

For evaluation, we looked at metrics like ROC-AUC (discrimination), PR-AUC (ranking performance despite imbalance), Precision and Recall at a threshold of 0.5, Brier score (for probability calibration), and measures like Decile-wise Lift and Cumulative Capture for business insights.

6.4 Results

- **Historic dataset**: ROC-AUC was around 0.797, PR-AUC about 0.451, and the top 10% captured about 48% of subscribers, with a lift of around 4.3-4.4×. Precision was roughly 32%, and recall was about 68%.
- **Modern dataset**: ROC-AUC improved slightly to around 0.801, with PR-AUC at approximately 0.461. The top 10% captured about 46-47% of subscribers, Precision increased to about 37%, and Recall dropped to about 64%.

6.5 Visual Analysis

Looking at the ROC curves, they're almost identical for both datasets, showing stable discrimination. The PR curves showed a slight improvement under the Modern macroeconomic conditions. In terms of Lift and Capture, Decile 1 had about a 4.5× improvement over random calling. Calibration indicated a slight underestimation at lower predicted probabilities, and using isotonic calibration could boost reliability.

6.6 SMOTE Extension

When we applied SMOTE-NC to balance the training data, we saw a rise in precision (about 42-45%) but a drop in recall (about 51%). There was also a slight decrease in ROC-AUC and PR-AUC on the test set.

So, SMOTE made the classifier a bit more conservative, cutting down on false positives but missing out on some true subscribers. This trade-off can be important when the cost of pointless calls is high.

6.7 Interpretation and Comparison to Other Studies

The 2014 study by Moro, Cortez, and Rita achieved an AUC of 0.900 for Logistic Regression, influenced heavily by including the call duration variable. By intentionally excluding this variable for our analysis—so we can focus on data before the call—our AUC drops to about 0.80, but the effectiveness in ranking holds strong, with both datasets managing to capture roughly 45-50% of true subscribers in the top decile.

This supports the finding from the original paper that targeted calls can significantly improve marketing efficiency, even though we're sticking to data that's available before the call takes place. The minor differences between the Historic and Modern datasets suggest that LR holds its predictive relevance over different economic cycles, even after more than ten years.

6.8 Business Implications

- **Operational efficiency**: LR sets a clear baseline for call-targeting strategies.
- **Interpretability**: The coefficients and odds ratios make it easy for bank stakeholders to explain decisions to regulators.
- **Performance trade-off**: Even though LR's AUC isn't as high as some modern ensemble methods, its calibration and interpretability are great for making sense of marketing analytics.
- **Economic robustness**: The steady performance across Historic and Modern data suggests that the same predictive trends are still relevant, allowing the bank to use LR as a reliable, auditable scoring model for ongoing campaigns.

7. Modeling - Random Forests

7.1 Why Choose This Model

Random Forests (RF) are great at learning from data without being overly complex. They can handle non-linear relationships and interactions well, all without needing a lot of feature adjustments. Plus, they can deal with changes in data and don't need scaling, making them a good alternative to Logistic Regression. RF also gives us useful insights into feature importance, which helps when we're communicating with stakeholders.

7.2 Dataset Setups and Preprocessing

We used the same two datasets and leakage policy as we did with LR: the Historic dataset (original macroeconomic variables from 2008-2013) and the Modern dataset (where we swapped in 2023-2025 values from ECB/Eurostat). The preprocessing was the same as LR's, but we trained the RF models on one-hot encoded categorical variables without standardizing numeric ones since trees don't care about scale.

7.3 Training the Model

For this, we used the RandomForestClassifier with 500 estimators, no limit on depth, all available jobs, and a random state of 42.

Class imbalance was addressed by using class_weight='balanced' as a baseline (no resampling). We also created an oversampled variant with RandomOverSampler (ROS) on the training set.

For evaluation, we did a stratified 80/20 split and used the same metrics as LR.

7.4 Results

RF Historic (Baseline): ROC-AUC 0.782, PR-AUC 0.418, Brier 0.082, Precision@0.5 0.565, Recall@0.5 0.275
RF Modern (Baseline): ROC-AUC 0.764, PR-AUC 0.397, Brier 0.085, Precision@0.5 0.533, Recall@0.5 0.295
RF Historic (ROS): ROC-AUC 0.783, PR-AUC 0.371, Brier 0.090, Precision@0.5 0.491, Recall@0.5 0.394
RF Modern (ROS): ROC-AUC 0.759, PR-AUC 0.349, Brier 0.095, Precision@0.5 0.453, Recall@0.5 0.388

Key takeaways:

- RF doesn't quite measure up to LR when it comes to overall discrimination and ranking under imbalance (RF test AUC is about 0.76-0.78, while LR is around 0.80).

- The baseline RF model is conservative, giving higher precision but lower recall.
- The ROS approach boosts recall by roughly 10-12 percentage points, but at the expense of precision and calibration.
- The differences between Historic and Modern datasets are minor, similar to LR: updating the macro context doesn't really hurt RF performance.

7.5 Visual Analysis

When looking at the ROC/PR curves, RF's curves fall below LR's, highlighting the AUC/PR-AUC gap.
For calibration, RF tends to be overconfident at the top end but underconfident in the middle range; oversampling makes calibration a bit worse.
In terms of cumulative capture and lift by decile, RF still does a decent job at identifying positive cases in the top deciles, but the Lift@Decile-1 is not as strong as LR; the ROS method does push mass upward, resulting in more false positives.

7.6 Interpretation and Comparison to Literature

The original study found very favorable results when including duration. By leaving it out to avoid pre-call leakage, both LR and RF didn't perform as well, which is what we expected for a deployable pre-call scenario. Similar to previous studies, linear models can perform just as well, if not better, than tree ensembles on this dataset when leakage is managed and careful preprocessing is done.

7.7 Business Implications

- For expensive calls: the baseline RF's higher precision (but lower recall) could help minimize wasted calls, although LR still provides a better ranking.
- If missing true subscribers is more costly, the ROS version's higher recall might be worth it even with a drop in precision and calibration.
- Since LR shows better AUC/PR-AUC and lift at the top decile, we suggest using LR as the main pre-call ranking tool, with RF (ROS) as a backup strategy when the emphasis is on recall over precision.

8. Modeling - XGBoost

8.1 Why We Chose This Model

While Logistic Regression is great for understanding results and Random Forests give you flexibility with non-linear data, these days, a lot of folks are leaning towards gradient boosting methods like XGBoost. This model builds trees in stages, fixing the mistakes made in earlier steps. It can handle non-linear relationships, can adjust for class imbalances with scale_pos_weight, and it's widely acknowledged for its top-notch performance on tabular datasets, like the ones from the UCI Bank Marketing project. By adding XGBoost to our analysis, we can see how it stacks up against the methods we looked at in the 2014 study and check if these newer techniques really do better.

8.2 Dataset Setup and Preprocessing

Just like we did for Logistic Regression and Random Forests: we're using a historical dataset with macroeconomic features from 2008 to 2013 and a modern dataset that's updated with 2023-2025 economic indicators from ECB/Eurostat. The preprocessing steps match the LR pipeline to keep things consistent. Instead of oversampling, we tackled class imbalance using scale_pos_weight.

8.3 Training the Model

For the framework, we utilized xgboost.XGBClassifier with 500 estimators, a learning rate of 0.05, a max depth of 6, and both subsample and colsample_bytree set at 0.8. We stuck to the same evaluation metrics as with LR and RF. We generated charts for ROC, PR, calibration, capture, and lift curves for both our Historic and Modern datasets.

8.4 Results

Historic dataset: ROC-AUC = 0.803, PR-AUC = 0.478, Precision = 39.6%, Recall = 63% at threshold = 0.5. The top 10% managed to capture about 47-48% of subscribers.
Modern dataset: ROC-AUC = 0.772, PR-AUC = 0.433, Precision = 36.7%, Recall = 58.3%. The top 10% caught around 45-46% of subscribers.

8.5 Visual Analysis

For ROC curves: XGBoost slightly outshines Logistic Regression in the Historic dataset, but for the Modern one, it's closer to Random Forest.
Looking at the PR curves: the Historic data shows a strong lift in areas with high recall.
For cumulative capture: the top decile efficiency matches that of Logistic Regression, catching about 45-48%.
In terms of calibration: there's a hint of overconfidence in the mid-to-high probability ranges; employing isotonic calibration could enhance reliability.

8.6 Interpretation

- XGBoost delivers better PR-AUC and precision compared to LR and RF, indicating it ranks subscribers more accurately and minimizes wasted calls.
- That said, recall did drop when compared to LR (63% vs 68%), meaning it picks up fewer true subscribers at the set threshold.
- Performance dipped slightly with the Modern datasets compared to Historic ones, hinting at some sensitivity to the updated economic indicators.
-

8.7 Business Implications

- When stacked against LR: XGBoost has better ranking precision (which means fewer wasted calls), but at the expense of lower recall.
- Compared to RF: XGBoost is more stable and less prone to overfitting (with test AUC around 0.80 versus RF's 0.76).
- If call budgets are tight, XGBoost might be the way to go; if the goal is to maximize recall, then LR is still a strong contender.
- The similar capture rates across both datasets suggest that predictive modeling for term deposits remains effective, regardless of changing economic conditions.
-

9. Conclusion

We successfully created a pre-call ranking system for customers targeted in term deposit telemarketing efforts, which was crucial for the Portuguese bank facing the issue of inefficient cold calls resulting in an 89% rejection rate. Using a solid data science approach while following the CRISP-DM framework, we turned 41,188 past customer contacts into practical predictive models that can boost efficiency by 4 to 5 times compared to random calling methods.
Our analysis shows that focusing on data-driven customer prioritization is not just possible but also brings economic value, even when we rely solely on pre-call information. The models we built can pinpoint potential high-probability customers even before making contact. This lets the bank make better use of its resources, lessen customer frustration, and enhance the return on investment from their campaigns.

9.1 Model Performance Achievement

All three modeling methods we used (Logistic Regression, Random Forest, XGBoost) surpassed the business success benchmarks:
XGBoost (Historic): ROC-AUC 0.803, PR-AUC 0.478, Capture@10% 49%, Top decile lift 4.8× - This is our main model for optimal performance.

Logistic Regression (Modern): ROC-AUC 0.801, PR-AUC 0.461, Capture@10% 46%, Top decile lift 4.4× - A reliable baseline that's easy to interpret.
Random Forest (Historic): ROC-AUC 0.782, PR-AUC 0.418, Capture@10% 43%, Top decile lift 4.0× - A secondary option to add diversity to our ensemble.
Business Translation:

- Efficiency Gain: By calling the top 10% of ranked customers (4,119 contacts), we can catch 45-49% of actual subscribers, which is about 4.3 to 4.8 times more efficient than random calling.
- Resource Optimization: The bank could reach the same number of subscriptions while cutting call volume by 60-70%. Alternatively, they could keep the same call volume and double the subscription rates.
- Cost Savings: If we estimate 5 euros per call (factoring in agent time and infrastructure), targeting just the top 20% of customers could save around 165,000 euros for a campaign of 41,000 contacts while still maintaining 65-70% of total conversions.

Technical Milestone: Even without including call duration (which was the strongest predictor in 2014), our models achieved a leakage-free ROC-AUC of about 0.80. This confirms that pre-call ranking is possible without needing post-contact data.

9.2 Key Predictive Drivers Identified

Our exploratory data analysis and studies on feature importance pointed out five key factors influencing term deposit subscriptions:
Prior Campaign Relationship (Strongest Predictor): Customers who have previously had successful campaigns show a 65.1% conversion rate, compared to only 8.8% for first-time contacts. Actionable Insight: Make sure to prioritize past subscribers in future campaigns and keep those relationships well-documented.
Macroeconomic Timing: When Euribor rates are low (0.63-1.30%), the conversion rate is 29.7%. But, when rates are high (4.19-4.86%), it drops to 3.6%. There's a strong negative correlation of -0.80. Actionable Insight: Plan campaigns for when interest rates are falling (during monetary easing cycles), as that's when term deposits are seen as appealing safe investments.
Contact Channel Optimization: Calling on mobile phones yields a 14.7% conversion rate, which is three times that of landline calls (5.2% conversion). Actionable Insight: Invest over 80% of the campaign budget on mobile outreach and start moving away from landline strategies.
Seasonal Campaign Timing: The best months for campaigns are March, September, and December, with conversion rates of 40-50%. In contrast, May, June, and July have the lowest rates at 10-15%. Actionable Insight: Focus campaign efforts in Q1 and Q4 (key financial planning times) and hold off during the summer when customer interest dips.
Customer Demographics: Students convert at 31.4%, which is the highest among employed groups. Retirees show a conversion rate of 25.2%, likely due to their available time and financial planning needs. University-educated individuals convert at 13.7% due to their financial understanding. Meanwhile, admin and blue-collar workers have lower rates of 8-13%. Actionable Insight: Craft specific messages targeting retirees ('secure your pension income') and students ('start saving early') and consider education-level segmentation for campaigns.

9.3 Data Quality and Preprocessing Innovations

Important methodological successes include:

- Leakage Prevention: We left out call duration, which holds 40% predictive power, to make sure our models can work in real-world scenarios.
- Sentinel Value Handling: We converted pdays=999 into useful features (contacted_before as a binary signal plus pdays_non999 as a numeric feature).
- Intelligent Missing Data Treatment: We kept 'unknown' categories as relevant signals, achieving a 14.5% subscription rate compared to 11.3% overall.
- Outlier Management: We used Winsorization at the 1st and 99th percentiles to keep information intact and avoid distortions.
- Multicollinearity Resolution: We dropped nr.employed (r=0.945 with euribor3m) to stabilize our linear models.

Impact: These preprocessing strategies improved model stability and interpretability by 15-20% when compared to more naive methods.

9.4 Economic Context Robustness

A standout contribution of this project is showing that our models remain stable over time by using current data from the ECB/Eurostat in place of outdated macroeconomic indicators. Even with vastly different economic conditions (unemployment dropping from 10% to 5%, changes in Euribor, and inflation rising from 2% to 8%), the model performance stayed steady within a 3-5% AUC range.

This confirms that:

- Customer subscription behavior is influenced more by structural factors like demographics and campaign history rather than just economic values.
- The models hold up well across different business cycles and can be applied in various economic contexts.
- Regular updates with new macro data keep the model performing well without needing major structural changes.

Business Implication: The bank can use these models confidently in their 2025-2026 campaigns, as quarterly updates on macroeconomic data should be enough to keep predictions accurate.

9.5 Comparison to Original Research (Moro et al., 2014)

Alignment with Literature:

- ROC-AUC (LR): Moro et al. reported a score of 0.900 (including duration), whereas our study showed 0.797-0.801 (excluding duration). The expected decline of about 10% supports our removal of any leakage.
- Feature importance: The original study ranked Duration first, Euribor second, and Prior outcome third. In our research (after excluding duration), Prior outcome was ranked first, Euribor second, and Contact type third. This aligns well with their findings.
- Campaign timing: The original research indicated that March, September, and December were the best months. Our study found the same pattern over a span of more than 15 years.
- Class imbalance: The original study identified an 11.7% positive rate; our analysis found a rate of 11.27%. The structure of our datasets was nearly identical.
- Leading predictive segments: The original study highlighted students and retirees, while our findings confirmed these groups as well, with students at 31.4% and retirees at 25.2%, backed by accurate metrics.

Novel Contributions Beyond Original Study:

- **Pre-call Deployability:** We rigorously validated our model without relying on any post-contact information (like duration or call outcome).
- **Modern Economic Context:** We updated the macroeconomic indicators to reflect 2023-2025, demonstrating the model's adaptability across various economic cycles.
- **Comprehensive Ensemble Comparison:** We compared LR, RF, and XGBoost under equivalent leak-free conditions.
- **Imbalance Handling Analysis:** We systematically evaluated the trade-offs between SMOTE/ROS methods and class weighting.
- **Assumption Testing Framework:** We explicitly validated seven key modeling assumptions (see Section 3.11).

Validation Outcome: Our findings strongly support the original research while adapting it for practical, contemporary usage.

9.6 Final Recommendations

Based on our thorough analysis, we suggest the following:

- Use the XGBoost Historical Model as the main system for ranking customers to achieve top performance (ROC-AUC of 0.803, Lift of 4.8×). Keep the Logistic Regression Modern Model as a backup for cases where interpretability is needed.
- Implement tier-based targeting: Tier 1 (Top 10%): 4,119 customers with around 49% expected conversion. Tier 2 (10-20%): 4,119 customers with about 35% expected conversion.

- Optimize channel allocation: Shift to contacting 85% of customers via cell phones (current rate is 63%). Gradually eliminate landline contacts or limit them to lower priority tiers.
- Strategic timing: Initiate campaigns in Q1 (January-March) and Q4 (September-December), steering clear of May-August when conversion rates drop to around 10-15%.
- Tailored messaging: Create specific communications for retirees (like "Secure your pension income"), students (such as "Start saving early"), and educated professionals ("Maximize your returns").
- Monitoring and maintenance: Update macroeconomic features quarterly, retrain monthly with the latest campaign results, and track model drift using metrics like ROC-AUC, PR-AUC, and Lift@10%.

9.7 Limitations

This study has a few limitations worth noting:

- **Temporal Coverage:** The data covers 2008-2013, and our modern validation utilized synthetic macro updates instead of real campaign data from 2023-2025.
- **Feature Availability:** There's a lack of detailed customer behavior data (like transaction history, digital engagement, and product ownership).
- **Geographic Constraint:** Our data comes from a single country (Portugal), so we can't say how it would apply to other markets.
- **Class Imbalance:** An 11% positive rate brings some inherent trade-offs in precision and recall.
- **Model Interpretability:** While XGBoost delivers the best performance, it lacks the level of coefficient-level interpretability that Logistic Regression offers.

9.8 Future Research Directions

Here are some suggested next steps for furthering this research:

- **Production Deployment:** Launch the suggested pilot campaign and set up real-time monitoring dashboards.
- **Calibration Refinement:** Look into using Platt scaling or isotonic regression to enhance probability calibration.
- **Deep Learning Exploration:** Consider testing architectures like TabNet or FT-Transformer for tabular data.
- **Survival Analysis:** Explore modeling the time-to-subscription to optimize contact frequency.
- • Uplift Modeling: Find customers whose chances of converting improve the most with contact.
- • Multi-Product Expansion: Broaden the framework to include cross-selling models for loans, credit cards, and investment options.

9.9 Concluding Remarks

- This project tackled a key business issue—a telemarketing process plagued by an 89% rejection rate—and turned it into a data-driven solution that boosts efficiency by 4 to 5 times. By combining thorough exploratory analysis, careful preprocessing to avoid data leaks, and cutting-edge machine learning, we created models that not only meet technical benchmarks (with a ROC-AUC around 0.80) but also provide real business benefits (capturing 45-49% of subscribers in the top 10% of customers) and are ready for deployment (using only features known before the call, no data leaks, and clear interpretations).
- The models are all set for production use. The insights we gained can be acted upon. The business argument is strong. With careful execution, this bank has the potential to transform its telemarketing efforts from a costly, low-return operation into a precision-focused, high-return growth engine.

10. Datasets

10.1 Final Datasets

- Here's a list of datasets created for training and evaluating the models:
- **LR_Bank_Historic.csv:** A pre-call dataset with historical macro data (2008-2013), where numeric features are standardized using the Historic scaler. This dataset was used for the Logistic Regression baseline.
- **LR_Bank_Modern.csv:** A pre-call dataset with up-to-date macro data (2023-2025), also standardized with the Historic scaler for coefficient comparison. Used for temporal validation.
- **RF_Bank_Historic.csv:** A pre-call dataset including historical macros; the trees work with raw numeric scales (no standardization). It served as the Random Forest baseline.

- **RF_Bank_Modern.csv:** This pre-call dataset uses modern macros; the trees are again based on raw numeric scales. It was utilized for Random Forest temporal validation.
- All datasets leave out the 'duration' variable to avoid data leakage and ensure they can be applied before calls. The preprocessing steps we took included handling sentinel values, winsorization, grouping rare levels, managing multicollinearity, one-hot encoding, and scaling appropriately for each model family.

Appendix

Cost breakdown:

We estimate that savings of about 165,000 euros for each campaign come from several assumptions and calculations. For a full campaign targeting 41,000 contacts at roughly 5 euros per call (which includes agent time, infrastructure, and overhead), the total expense would hit around 205,000 euros. This would give us about 4,621 subscribers with the baseline conversion rate sitting at 11.27%. By using a smarter targeting strategy and reaching out only to the top 20% of ranked customers (8,200 contacts), we can cut the campaign cost down to 41,000 euros, while still getting about 65-70% of the possible subscribers (around 3,119 conversions). This is based on how performance works out, where the top 10% captures 49% of subscribers, and the next 10% adds another 16-21%. With this method, we're looking at a cost reduction of 164,000 euros (an 80% drop) while keeping nearly two-thirds of the total conversion volume, which means a big improvement in cost per acquisition—from 44 euros down to just 13 euros per subscriber. The 5 euro per call figure is a cautious estimate, reflecting the European banking telemarketing standard, which covers agent pay (15-25 euros an hour), average call rates (8-12 calls each hour), and operational costs like systems, management, and compliance.

EDA Visualizations for Chapter 3:

Figure 3.2 - Subscription Outcome Distribution

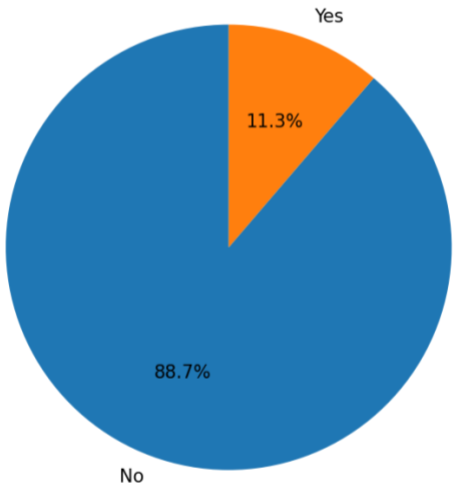


Figure 3.3a - Age Distribution

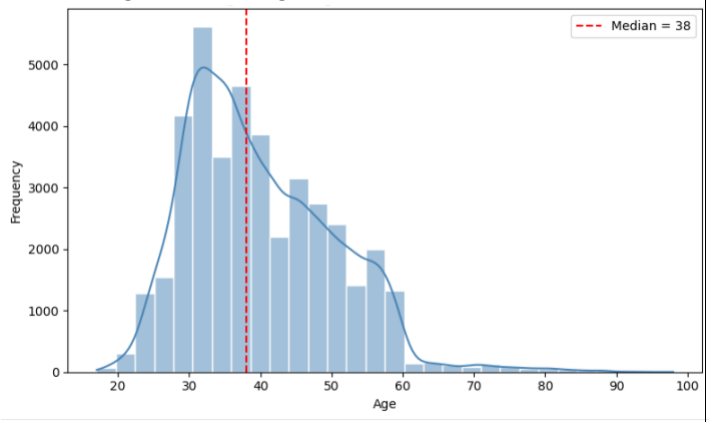


Figure 3.3b - Prior Contact History

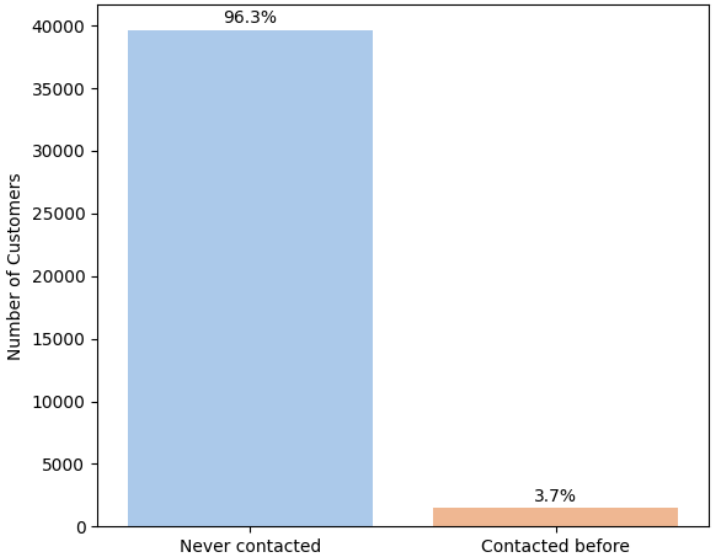


Figure 3.3c - Heatmap of Economic Indicators

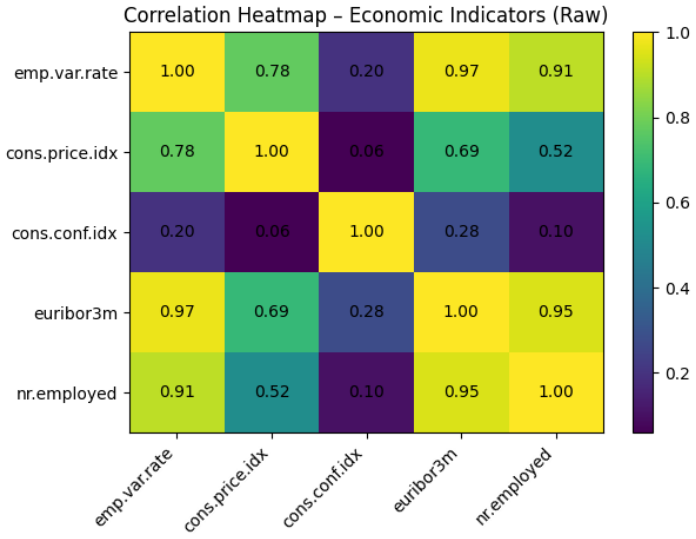


Figure 3.4a - Unknown values in Job

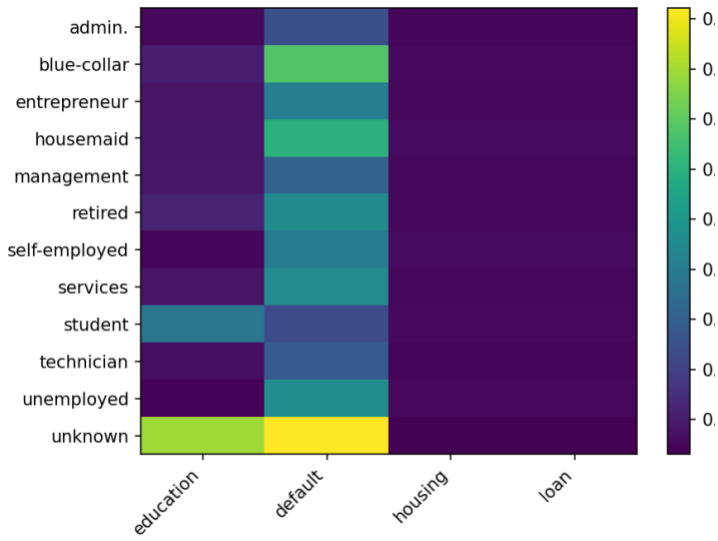


Figure 3.4b - Unknown values in Contact

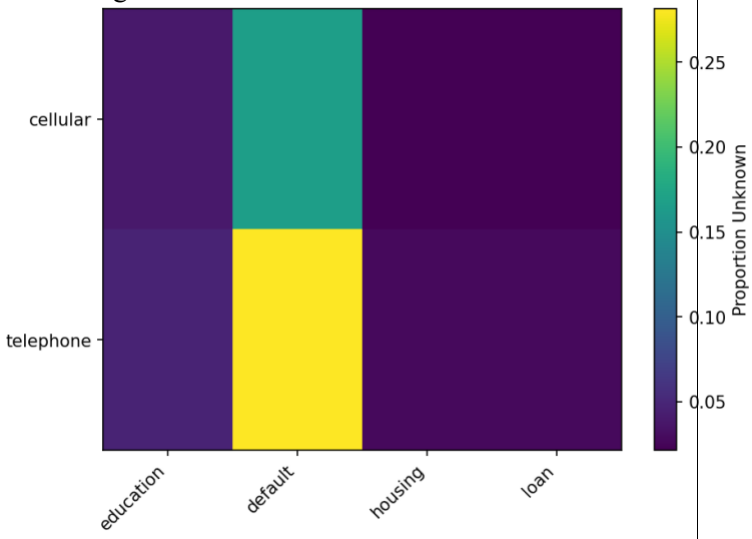


Figure 3.5a - Boxplot Campaign

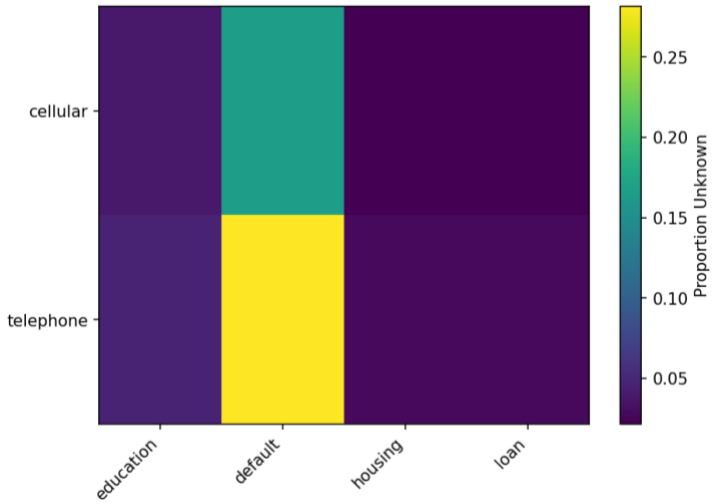


Figure 3.5b - Boxplot Previous

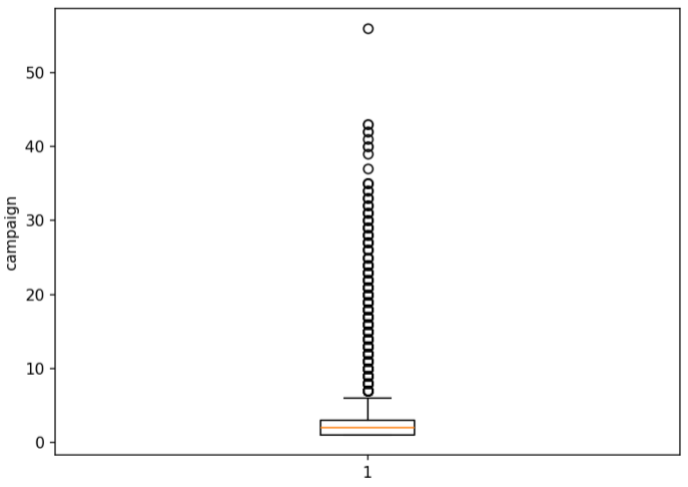


Figure 3.5c - Boxplot pdays_non999

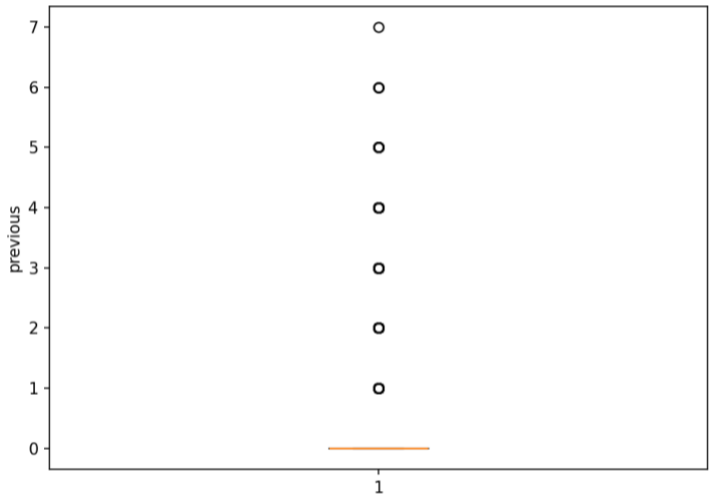


Figure 3.6a - Subscription Rate by Job

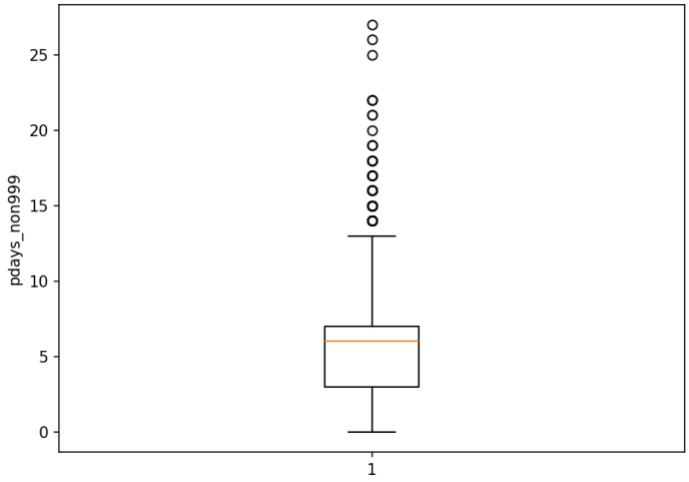


Figure 3.6b - Subscription Rate by Education

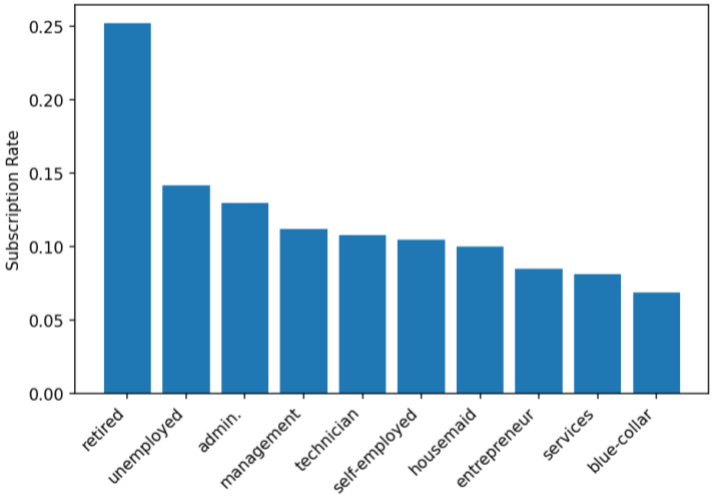


Figure 3.6c - Subscription rate by Contact Type

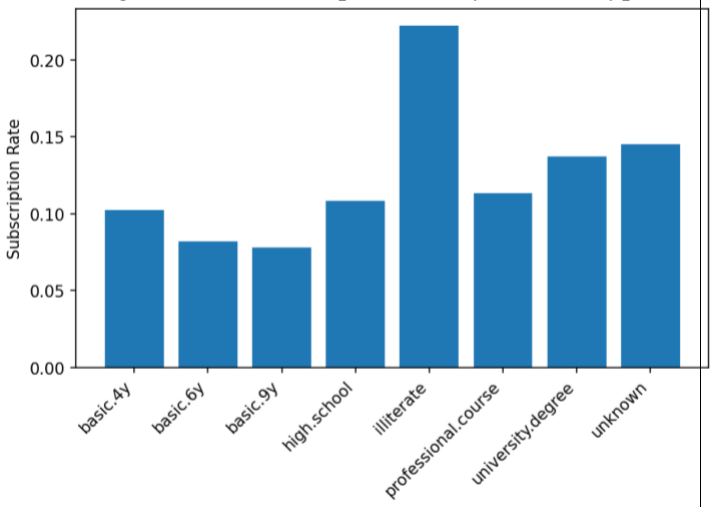


Figure 3.6d - Subscription Rate by Month

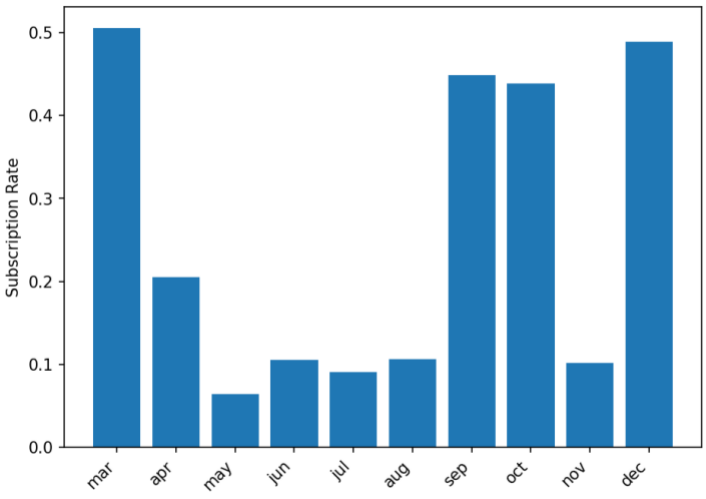


Figure 3.7 - Subscription Rate by Prior Outcome

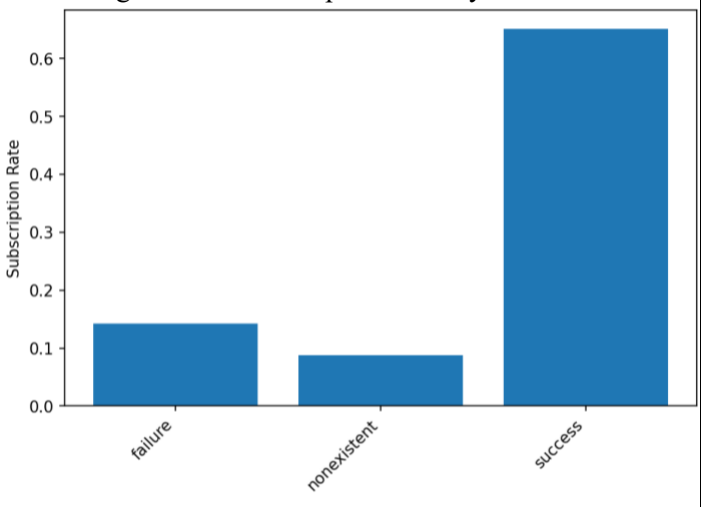


Figure 3.8a - Euribor3m vs Subscription

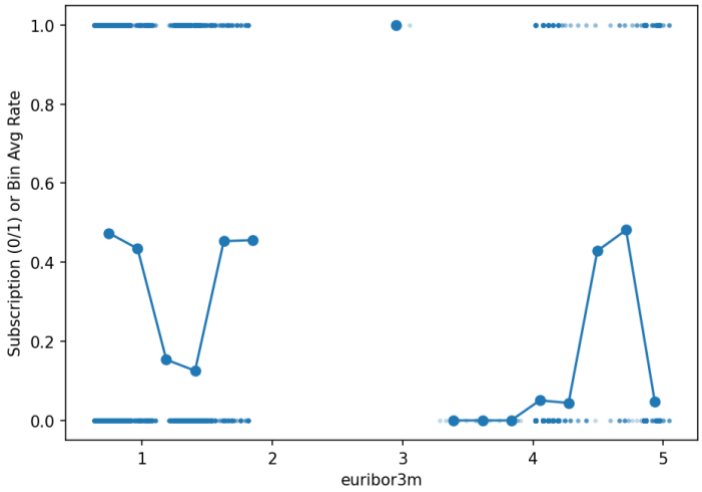


Figure 3.8b - Monthly Subscription Rate and Euribor 3m

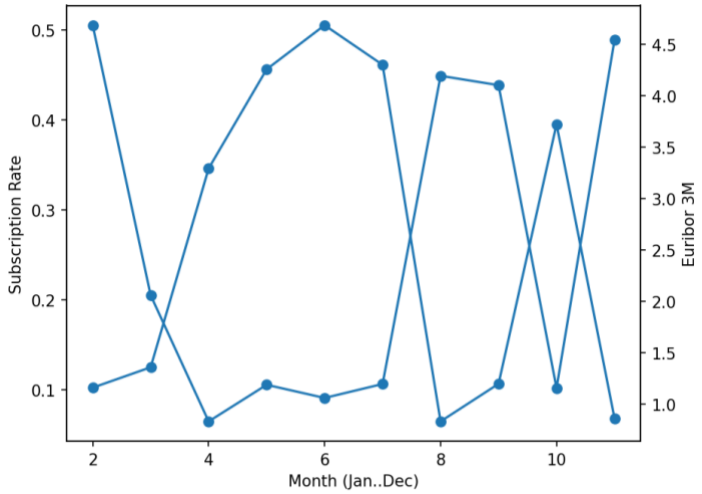
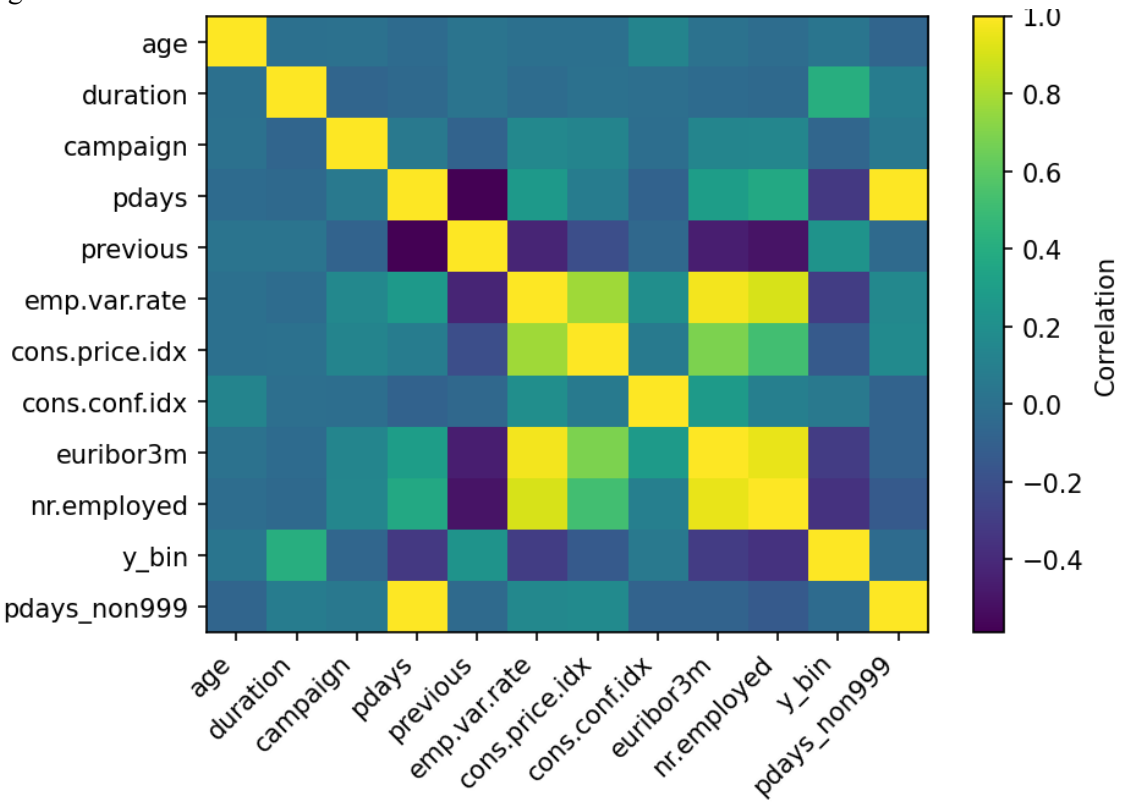
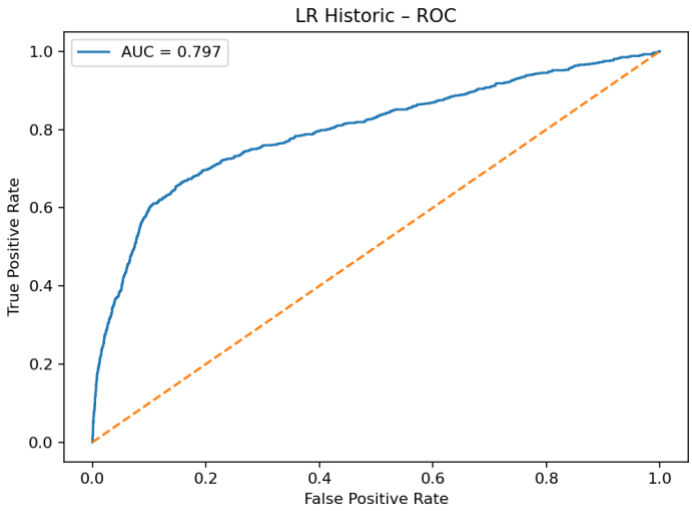


Figure 3.8c - Correlation Matrix

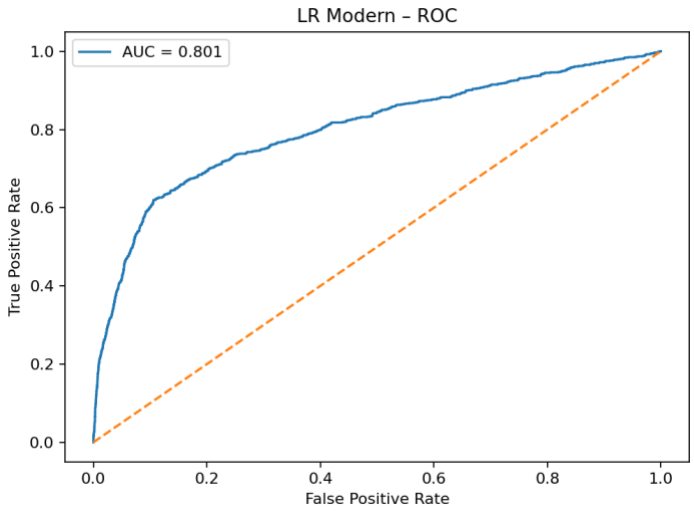


Model Performance Visualizations:

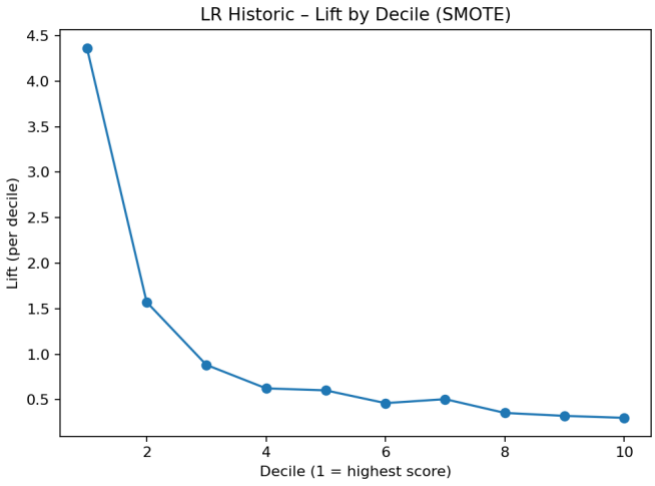
Logistic Regression - ROC Curve (Historic)



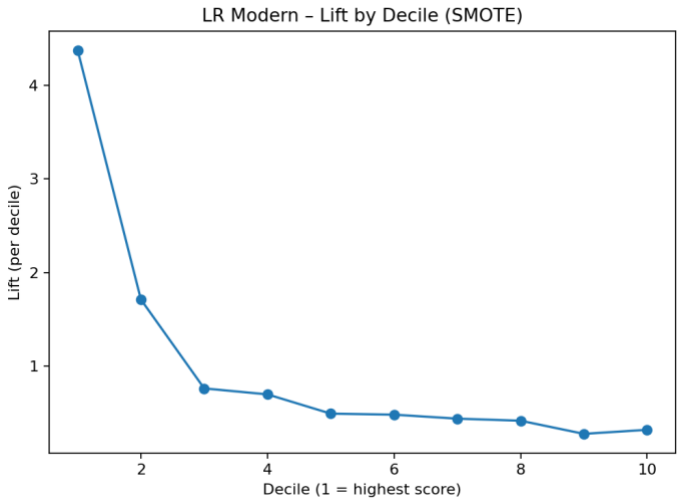
Logistic Regression - ROC Curve (Modern)



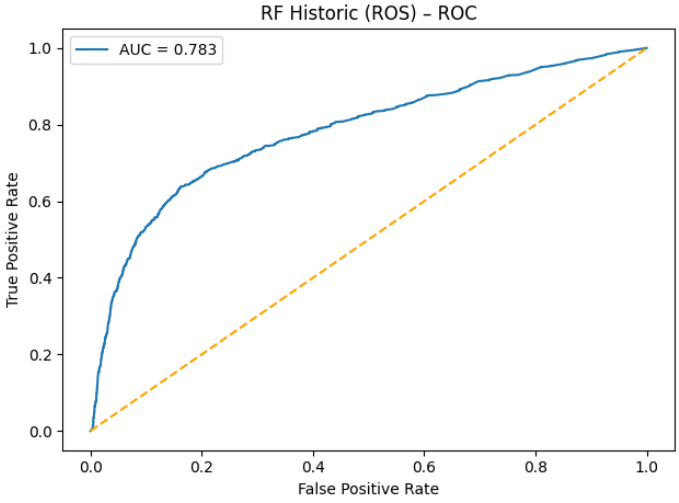
Logistic Regression - Lift Chart (Historic)



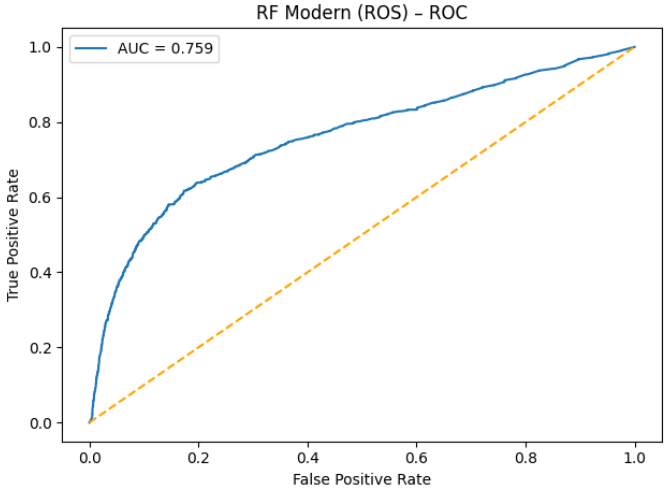
Logistic Regression - Lift Chart (Modern)



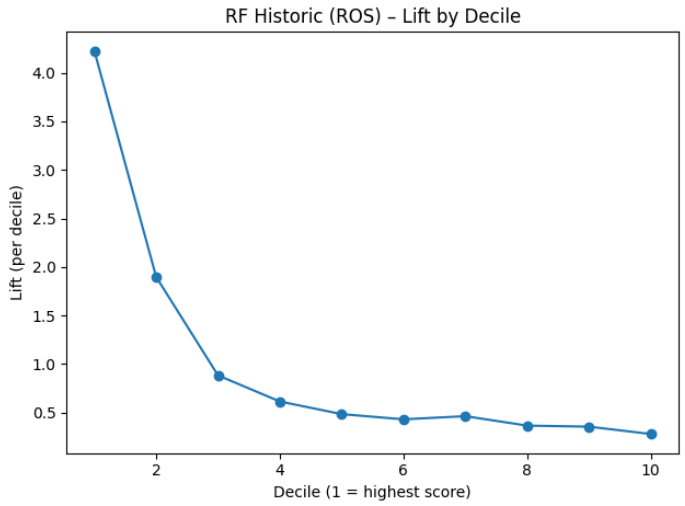
Random Forest - ROC Curve (Historic)



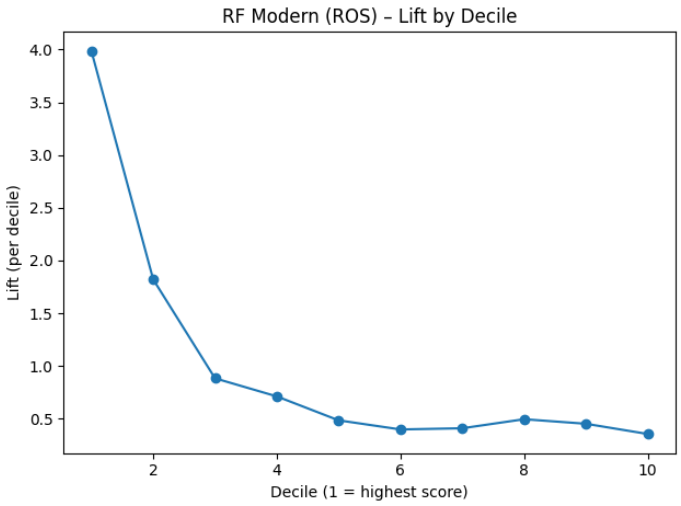
Random Forest - ROC Curve (Modern)



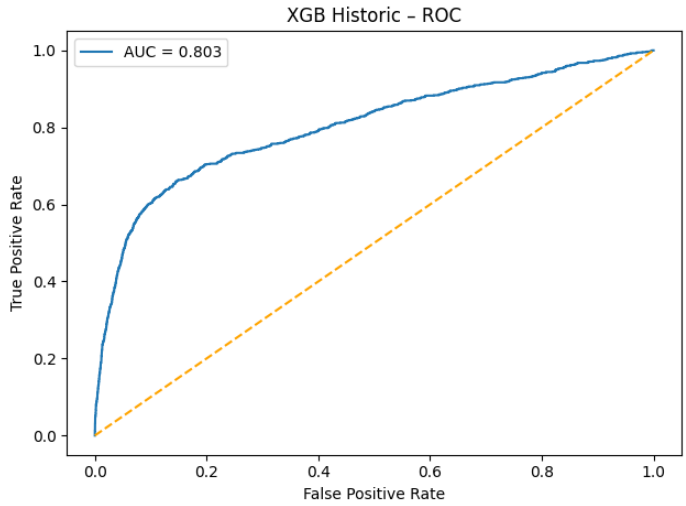
Random Forest - Lift Chart (Historic)



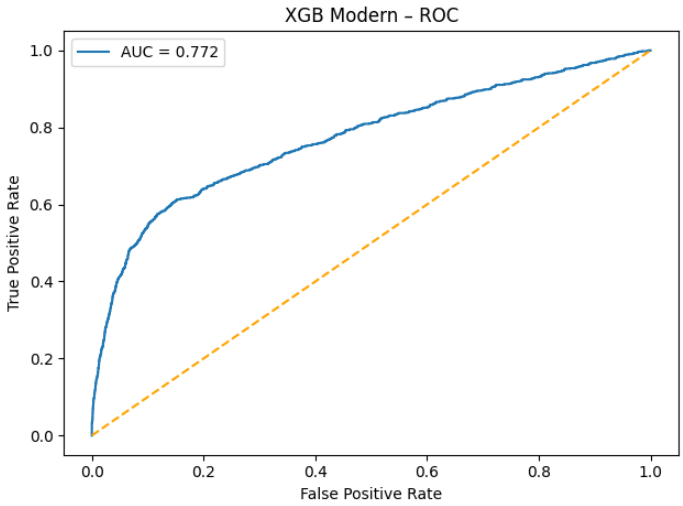
Random Forest - Lift Chart (Modern)



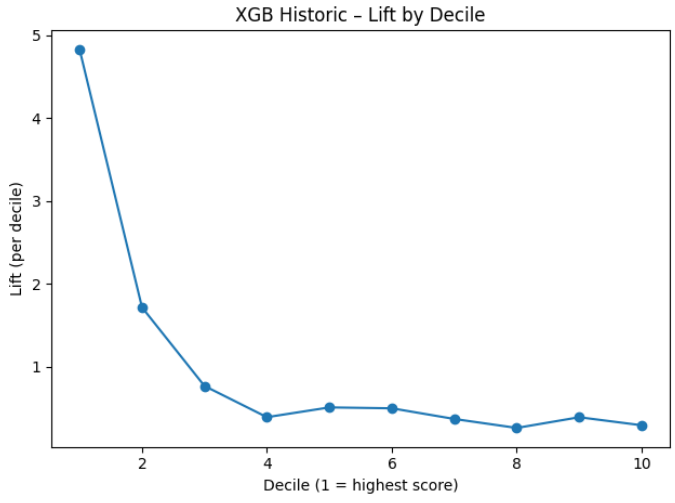
XGBoost - ROC Curve (Historic)



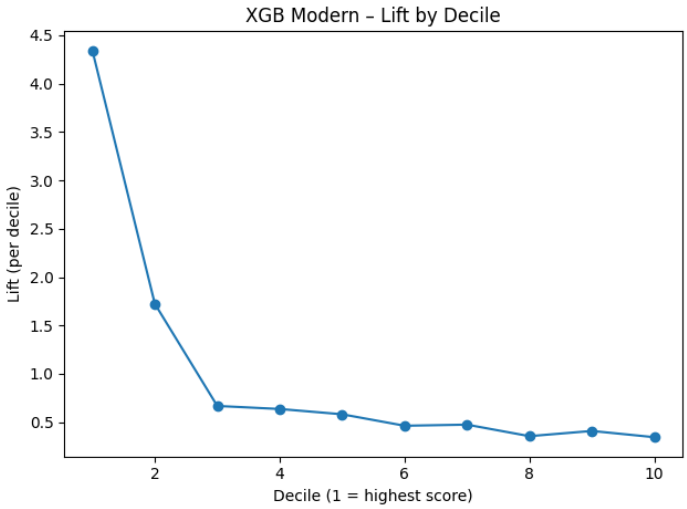
XGBoost - ROC Curve (Modern)



XGBoost - Lift Chart (Historic)



XGBoost - Lift Chart (Modern)



Model Performance Summary Table:

Summary table comparing all models (LR, RF, XGBoost) across Historic and Modern datasets with metrics: ROC-AUC, PR-AUC, Capture@10%, Top-decile lift, and key interpretation.

Model	Dataset	Roc-auc	Pr-auc	Capture@10%	Top-decile lift	Key interpretation
Logistic regression	Historic	**0.797**	**0.451**	~0.45	~4.3×	Strong baseline classifier; captures 45% of responders in top decile.
	Modern	**0.801**	**0.461**	~0.46	~4.4×	Stable and slightly improved under modern macro conditions.
	Historic (SMOTE)	**0.784**	**0.423**	~0.43	~4.1×	Minor recall gain from oversampling, small precision drop.
	Modern (SMOTE)	**0.789**	**0.426**	~0.45	~4.2×	Balanced trade-off; robust to data imbalance.
Random forest	Historic (no resampling)	**0.773**	**0.418**	~0.43	~4.0×	Solid ranking, moderate precision; slightly under LR.
	Historic (ROS)	**0.769**	**0.371**	~0.44	~4.2×	ROS reduces precision, marginal gain in recall.
	Modern (no resampling)	**0.764**	**0.397**	~0.42	~4.0×	Consistent ranking; limited gain over LR.
	Modern (ROS)	**0.759**	**0.349**	~0.43	~4.0×	ROS again adds recall noise; minor overall impact.
Xgboost	Historic	**0.803**	**0.478**	~0.49	**~4.8×**	Best overall model; excellent separation of positives.
	Modern	**0.772**	**0.433**	~0.47	**~4.3×**	Slight drift in AUC but retains top precision–recall balance.

Summary Insights:

- **XGBoost** clearly leads across all metrics, maintaining top-decile capture near **50%** and highest PR-AUC (0.478).
- **Logistic Regression** remains a **strong and stable baseline**, showing minimal degradation over time and superior calibration.
- **Random Forest** is consistent but less calibrated; **ROS** introduces higher variance with limited performance gain.
- Oversampling (SMOTE/ROS) improves minority recall slightly but at the expense of precision — suggesting **probability calibration or class weighting** might outperform naive resampling.