Predicting Term Deposit Subscription Using Logistic Regression: A GLM-Based Analysis on Bank Marketing Data

Abstract

This project applies logistic regression—a widely used Generalized Linear Model (GLM)—to predict customer subscription behavior using bank marketing data. The analysis focuses on both predictive accuracy and interpretability, with attention to class imbalance challenges. Key client attributes such as past campaign outcomes, timing of contact, and communication methods were identified as strong predictors. The model provides practical insights for improving marketing strategies, enabling more targeted and effective client outreach.

1. Introduction

In modern banking, marketing strategies are shifting from traditional broad outreach approaches toward data-driven, targeted campaigns. Among various financial products, term deposits are particularly valuable—not only do they offer customers a stable return, but they also help banks retain clients and manage liquidity more effectively. However, given the vast pool of potential customers, banks face crucial questions: which clients are most likely to subscribe to a term deposit? How can conversion rates be improved? And which customers should be prioritized for marketing contact? These challenges directly impact both operational costs and customer satisfaction.

While conventional marketing decisions often rely on intuition or rule-of-thumb segmentation, the availability of detailed historical data has made statistical modeling and machine learning the mainstream in marketing optimization. Banks typically collect a wide array of customer information, including age, job type, marital status, loan history, and interaction patterns. These variables hold latent behavioral signals that can be uncovered through modeling.

This project uses the Bank Marketing dataset from the UCI Machine Learning Repository, which contains data from a Portuguese bank's telemarketing campaigns conducted between 2008 and 2010. The goal of these campaigns was to promote subscription to term deposit products. The dataset includes over 45,000 customer records with demographic and interaction-related features, along with a binary outcome variable indicating whether each customer subscribed. Notably, only about 12% of the customers subscribed, making this a highly imbalanced binary classification problem and presenting a modeling challenge.
Accordingly, the project goals are:

- To explore relationships between client attributes and subscription behavior using descriptive statistics and visualizations;

- Construct and refine a logistic regression model, a widely used form of Generalized Linear Models (GLM), to predict term deposit subscription;

- To evaluate model performance using metrics such as confusion matrices, ROC curves, and AUC scores;

- Analyze variable importance to understand which features most influence customer decisions and offer practical recommendations for marketing strategy.

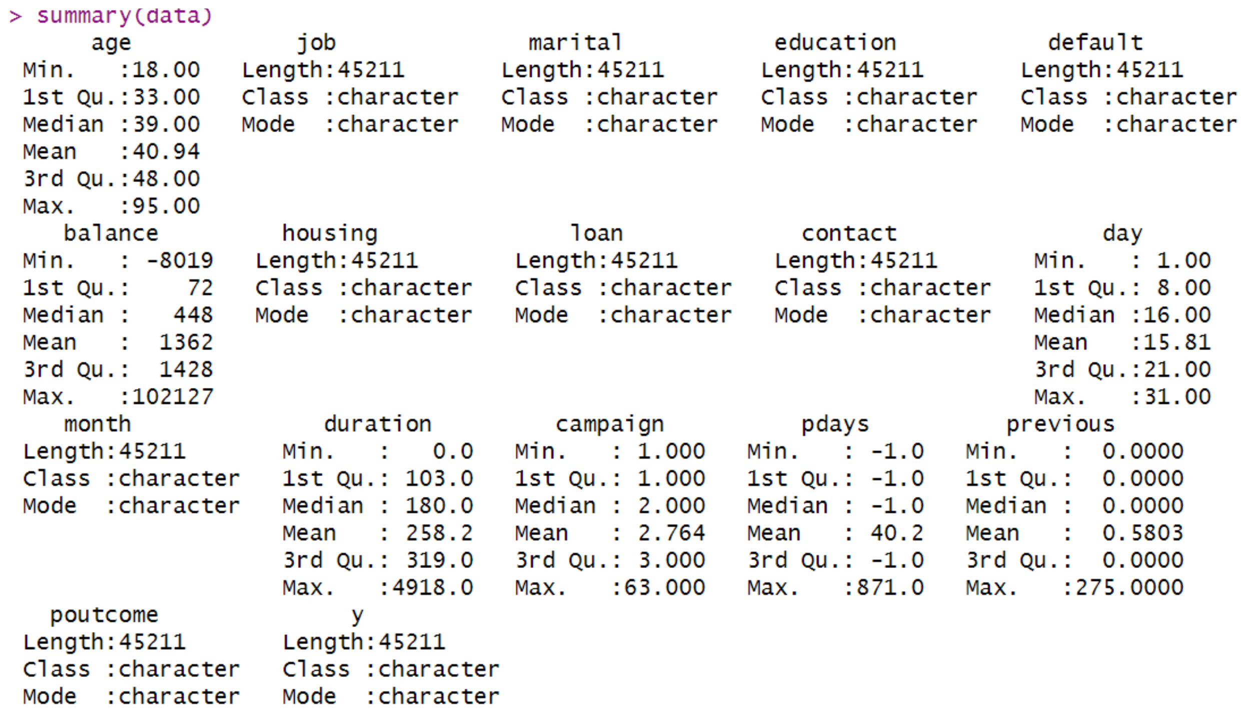
This work builds upon prior research in bank marketing analytics. For example, Moro et al. (2011) applied the CRISP-DM methodology and machine learning techniques to similar data, highlighting the tradeoff between predictive power and interpretability. Theoretical foundations for logistic regression are provided by Hosmer et al. (2013) and Menard (2002), while He and Garcia (2009) emphasize the importance of metrics such as recall and AUC when dealing with class imbalance.

Accordingly, the primary objective of this project is to build a logistic regression model that predicts whether a client will subscribe to a term deposit based on their demographic and behavioral attributes. In addition to predictive accuracy, the project aims to interpret the most influential factors behind subscription decisions and assess how class imbalance affects model performance. Ultimately, the goal is to generate data-driven insights that support more effective and targeted bank marketing strategies.

This report integrates predictive modeling, variable importance analysis, and performance evaluation under class imbalance, offering both methodological rigor and practical relevance for real-world applications.

2. Descriptive Statistics and Data Visualization

Before modeling, I conducted a systematic descriptive analysis of the Bank Marketing dataset to gain a comprehensive understanding of its structure and variables, as shown in figure 1. The dataset includes 45,211 records and 17 variables, covering demographic information (e.g., age, job, marital status), financial status (e.g., account balance, loan information), and historical marketing data (e.g., number of contacts, outcomes of previous campaigns).



```
> summary(data)
      age            job              marital           education           default         
 Min.   :18.00   Length:45211       Length:45211       Length:45211       Length:45211      
 1st Qu.:33.00   Class :character   Class :character   Class :character   Class :character  
 Median :39.00   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :40.94                                                                              
 3rd Qu.:48.00                                                                              
 Max.   :95.00                                                                              
    balance         housing              loan             contact               day       
 Min.   : -8019   Length:45211       Length:45211       Length:45211       Min.   : 1.00  
 1st Qu.:    72   Class :character   Class :character   Class :character   1st Qu.: 8.00  
 Median :   448   Mode  :character   Mode  :character   Mode  :character   Median :16.00  
 Mean   :  1362                                                            Mean   :15.81  
 3rd Qu.:  1428                                                            3rd Qu.:21.00  
 Max.   :102127                                                            Max.   :31.00  
    month              duration         campaign          pdays           previous       
 Length:45211       Min.   :   0.0   Min.   : 1.000   Min.   : -1.0   Min.   :  0.0000  
 Class :character   1st Qu.: 103.0   1st Qu.: 1.000   1st Qu.: -1.0   1st Qu.:  0.0000  
 Mode  :character   Median : 180.0   Median : 2.000   Median : -1.0   Median :  0.0000  
                    Mean   : 258.2   Mean   : 2.764   Mean   : 40.2   Mean   :  0.5803  
                    3rd Qu.: 319.0   3rd Qu.: 3.000   3rd Qu.: -1.0   3rd Qu.:  0.0000  
                    Max.   :4918.0   Max.   :63.000   Max.   :871.0   Max.   :275.0000  
   poutcome              y            
 Length:45211       Length:45211      
 Class :character   Class :character  
 Mode  :character   Mode  :character  
```

Figure 1. Descriptive Statistics

Among the numerical variables, age ranges from 18 to 95, with a mean of 40.94 and a median of 39, indicating most clients fall in the young to middle-aged category. Account balance (balance) varies dramatically from -8019 to 102,127, revealing a highly skewed distribution with a few high-net-worth clients. Contact-related variables such as campaign (current campaign contact count), pdays (days since last contact), and previous (previous contact count) also exhibit strong skewness, which may affect model stability and require careful handling.

The response variable "y" indicates whether a customer subscribed ("yes" or "no"). Count statistics show 5,289 "yes" responses versus 39,922 "no" responses—indicating a subscription rate of only 12%. This kind of imbalanced binary classification problem can significantly influence model performance, requiring special attention during training and evaluation.

To visualize the data structure, I created several charts, summarized as follows:

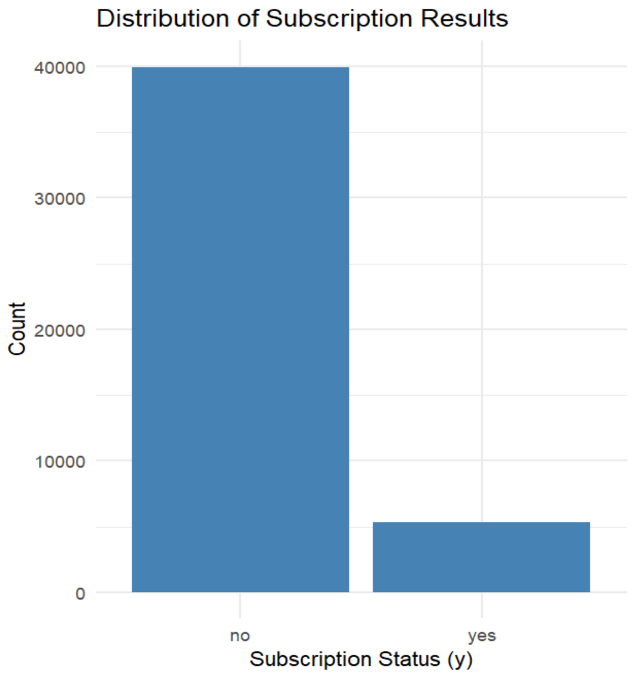


Figure 2. Distribution of Subscription Results

Figure 2 clearly shows that most clients did not subscribe, with only a small portion opting in, highlighting the severe class imbalance. In such cases, relying solely on accuracy is misleading—robust metrics like F1 score, recall, and AUC are emphasized instead.

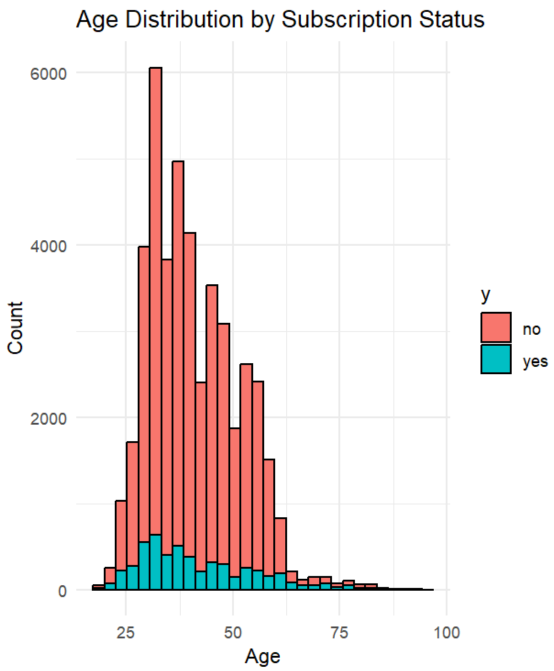


Figure 3. Age Distribution by Subscription Status

Figure 3 demonstrates that subscription behavior differs significantly by age. Most subscribers are aged 30–60, with the 30–40 group being the most active. Virtually no subscriptions occur among clients under 30 or over 60, indicating a potential nonlinear relationship between age and subscription likelihood. Techniques such as binning, polynomial terms, or splines may help model this effect.

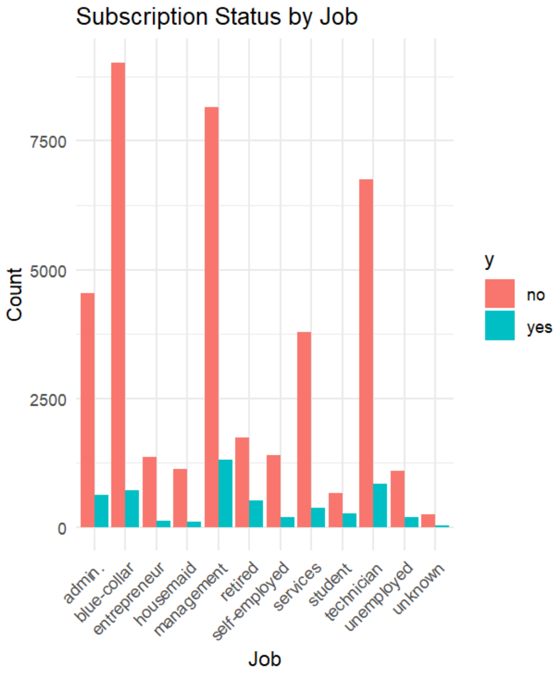


Figure 4. Subscription Status by Job

Subscription likelihood varies by occupation, as shown in figure 4. Higher “yes” ratios appear among “students” and “retired” individuals, while “blue-collar” and “services” workers show lower interest. Though jobs like “management” and “technician” are common, their subscription rates are still modest. This suggests job is a powerful predictor and may interact with age or education in meaningful ways.

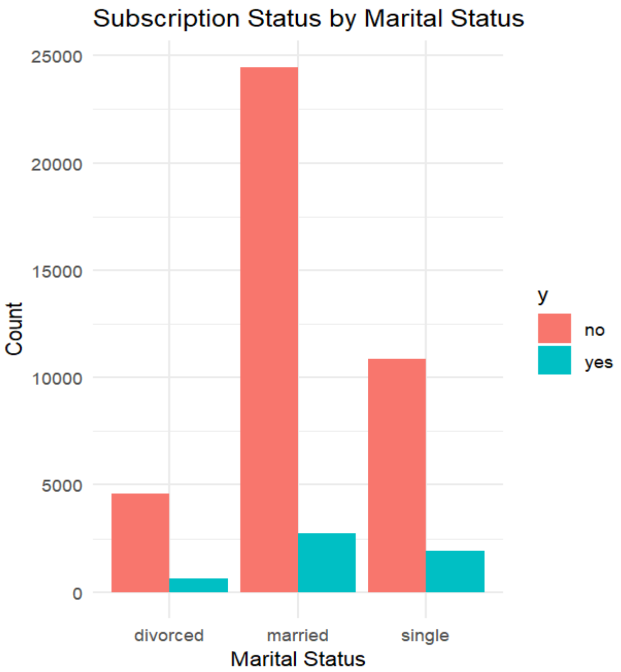


Figure 5. Subscription Status by Marital Status

As shown in figure 5, Although “married” clients form the largest group, their subscription rates are lower than those of “single” and “divorced” individuals, particularly the former. This might reflect autonomy in decision-making or differing risk preferences. Marital status is therefore a valuable predictor.

Together, these descriptive insights and visualizations not only reveal key data distributions but also suggest meaningful relationships among variables. The charts are well-structured with clear labels, reinforcing the statistical insights and guiding variable selection and modeling strategy.

3. Method

This project applies a binary logistic regression model, a widely used form of Generalized Linear Model (GLM), to estimate the probability that a client subscribes to a term deposit. GLMs generalize linear regression by allowing the response variable to follow non-normal distributions and link it to predictors through a transformation function. For binary outcomes, logistic regression employs the logit link function, which maps predicted values to probabilities between 0 and 1.

Mathematically, the model can be expressed as:

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1}$$

where p is the probability that a client subscribes (i.e., y=1), β_0 is the intercept, and β_1, β_2 ,…, β_{p-1} are the coefficients for the predictor variables x_1 , x_2 ,…, x_{p-1}. This formulation not only ensures that the predicted probabilities are constrained between 0 and 1, but also allows us to interpret each β_i as the change in the log-odds of subscription per unit change in x_i. Exponentiating these coefficients gives the odds ratios, which provide an intuitive measure of each predictor's impact.

During preprocessing, I confirmed that the dataset contained no missing values, so no imputation or deletion was necessary. The binary response variable y was encoded such that "yes" = 1 and "no" = 0. All categorical variables (e.g., job, marital, education, default, housing, loan, contact, month, poutcome) were converted into factors to enable proper model interpretation.

The dataset was randomly split into training (80%) and testing (20%) sets to ensure generalizability. A full model including all variables was first fitted, followed by stepwise selection based on the Akaike Information Criterion (AIC). This bidirectional process aims to reduce overfitting and enhance interpretability by iteratively adding or removing variables based on their contribution to model fit.

The final selected model retained key predictors such as job, marital status, education, account balance, housing and loan status, contact method, contact day and month, call duration, number of contacts in the campaign, and prior campaign outcome. Variables like age were excluded due to low predictive power or multicollinearity.

This process significantly simplified the model—reducing the total number of variables from 16 to around 12—while improving model quality (AIC reduced from 26,136 to 17,292). The result is a more parsimonious, interpretable, and stable model suited for real-world applications.

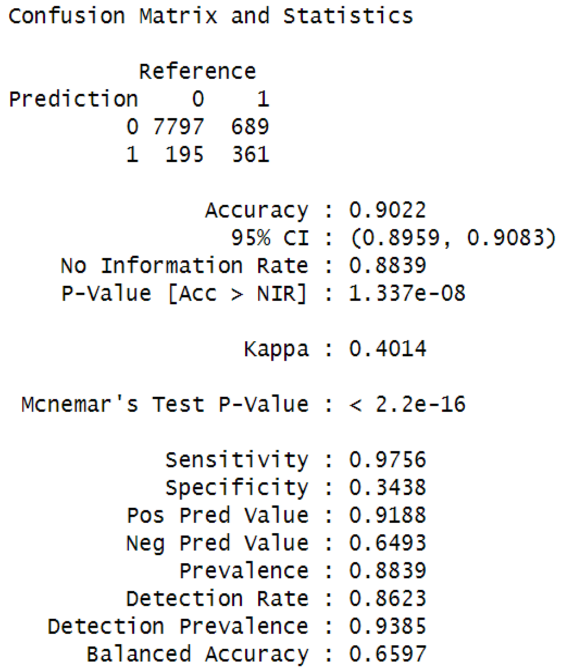
Given the pronounced class imbalance (as only about 12% subscribed), I also implemented a **class-weighted logistic regression** to boost minority class influence. However, while it slightly improved precision in some thresholds, it consistently lowered recall and F1 scores—indicating weaker ability to detect actual subscribers. Therefore, the final model uses a standard (unweighted) logistic regression, but tunes the classification threshold (e.g., lowering from 0.5 to 0.2) to improve recall without sacrificing too much precision.

Finally, I evaluated model performance on the test set using confusion matrices, accuracy, precision, recall, F1 score, and balanced accuracy. ROC curves and AUC were also computed to assess overall discrimination ability. To aid interpretation, I extracted model coefficients and transformed them into odds ratios, revealing both the direction and strength of influence for each predictor.

4. Results

Following model development, I evaluated the unweighted logistic regression model on the test set using both metrics and visualizations.

As shown in figure 6, the confusion matrix showed strong performance: out of 9,000 test cases, 7,797 “non-subscribers” were correctly classified, with 689 false positives. For “subscribers,” 361 were correctly predicted, with 195 misclassified. This yields an overall accuracy of 90.22%, strong performance given the imbalance.



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7797  689
         1  195  361

               Accuracy : 0.9022
                 95% CI : (0.8959, 0.9083)
    No Information Rate : 0.8839
    P-Value [Acc > NIR] : 1.337e-08

                  Kappa : 0.4014

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9756
            Specificity : 0.3438
         Pos Pred Value : 0.9188
         Neg Pred Value : 0.6493
             Prevalence : 0.8839
         Detection Rate : 0.8623
   Detection Prevalence : 0.9385
      Balanced Accuracy : 0.6597
```

Figure 6. Confusion Matrix and Statistics

Sensitivity (97.56%) indicates the model's high effectiveness in identifying non-subscribers, while the relatively low specificity (34.38%) reflects room for improvement in detecting subscribers. However, due to the imbalance, accuracy alone can be misleading. Thus, I also examined recall and F1 score to assess the model’s ability to capture the minority class. At the default threshold of 0.5, the model achieved a precision of 91.88%, a recall of 34.38%, and an F1 score of approximately 50%, indicating a moderate ability to identify subscribing clients. Lowering the threshold (e.g., to 0.2) improved F1 performance, supporting threshold tuning as a practical strategy for handling imbalanced data and enhancing the detection of true subscribers.

I also experimented with a class-weighted logistic regression model to address the class imbalance, with the results presented in the Appendix. While it slightly improved precision at certain thresholds, it consistently reduced recall and F1 score across the board. For example, at a threshold of 0.5, the weighted model yielded a precision of 74.0%, but recall dropped to only 3.5%, resulting in an F1 score below 7%. Given this limited ability to detect actual subscribers, I retained the unweighted model and focused on threshold tuning as a more effective strategy.

The ROC curve, which plots sensitivity against (1 – specificity), shows a clear arc toward the top-left corner. The AUC of 0.904 confirms excellent discriminative power—meaning the model correctly distinguishes a subscriber from a non-subscriber in over 90% of random pairwise cases.

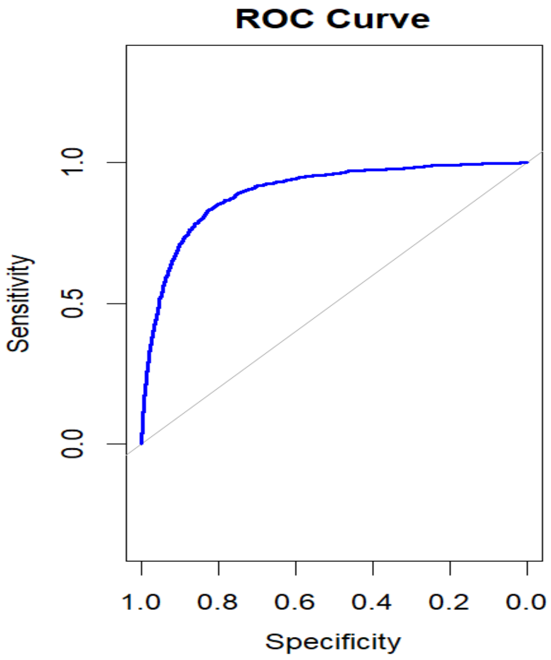


Figure 7. ROC Cureve

Figure 8 presents the variable importance analysis, based on model coefficients and their corresponding odds ratios (OR). The results highlight several key predictors:

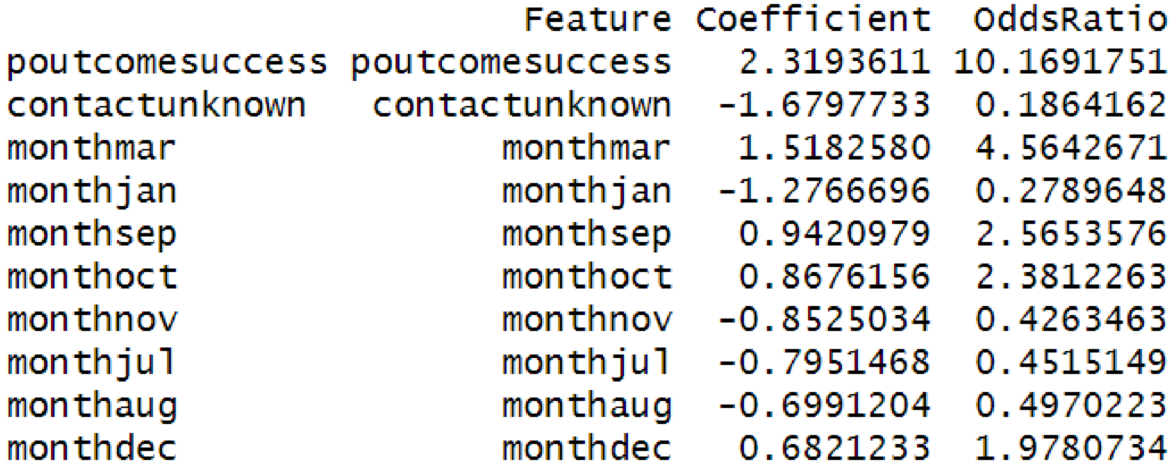
- The most influential feature is poutcomesuccess—the client's response to past marketing efforts—with a coefficient of 2.32 and an odds ratio of 10.17, indicating that these clients are over ten times more likely to subscribe again.

- The contact month also plays a crucial role. For example, March (OR = 4.56) is strongly associated with higher subscription likelihood, while January (OR = 0.28) and November (OR = 0.43) are linked to much lower engagement. This underscores the importance of campaign timing.

- Clients with unknown contact type are significantly less likely to subscribe (OR = 0.19), possibly due to unreachable channels or lower engagement—making this a key area for customer data enrichment.

These findings offer clear implications for marketing strategy: banks should prioritize clients who responded positively in the past, schedule outreach during high-response months such as March and September, and address gaps in contact information to improve conversion.

These results not only validate the predictive power of the logistic regression model but also provide actionable insights into client behavior, which I further explore in the following discussion section.



```
                         Feature Coefficient  OddsRatio
poutcomesuccess poutcomesuccess   2.3193611 10.1691751
contactunknown   contactunknown  -1.6797733  0.1864162
monthmar               monthmar   1.5182580  4.5642671
monthjan               monthjan  -1.2766696  0.2789648
monthsep               monthsep   0.9420979  2.5653576
monthoct               monthoct   0.8676156  2.3812263
monthnov               monthnov  -0.8525034  0.4263463
monthjul               monthjul  -0.7951468  0.4515149
monthaug               monthaug  -0.6991204  0.4970223
monthdec               monthdec   0.6821233  1.9780734
```

Figure 8. Variable Importance Analysis

5. Discussion

This project sets out to predict customer subscription to term deposits using logistic regression and to uncover the behavioral factors that drive conversion. Through systematic data preprocessing, statistical modeling, and comprehensive evaluation, I successfully built a well-performing model and extracted key insights relevant to marketing strategy.

The final model demonstrated robust overall performance: over 90% accuracy and an AUC of 0.904, indicating strong discriminative power. It performed particularly well in identifying non-subscribers, and—with threshold tuning—achieved a fair balance between precision and recall for the minority (subscriber) class. These metrics support the model's utility as a tool for identifying high-potential clients and guiding marketing outreach.

Variable analysis revealed several important predictors. Most notably, “poutcome_success” had an odds ratio of 10.17, underscoring that clients who responded positively to prior campaigns are far more likely to convert again. The negative effect of “contact_unknown” illustrates the value of complete contact data. Timing, captured by the month of contact, also played a significant role—suggesting potential seasonality in customer responsiveness. Additionally, client attributes such as job type, education level, and loan status also influenced outcomes, offering further opportunities for segmentation.

In answering the core question — "Who is more likely to subscribe?"—the model emphasizes that marketing history, communication channels, and timing are key behavioral indicators. Banks can leverage these findings to improve targeting, avoid low-response periods, and enhance contact databases for better campaign performance.

While the project yielded meaningful insights, it also faced limitations. The class imbalance made it difficult to optimize for the minority class. Although class-weighted logistic regression was tested, it underperformed in terms of recall and F1 score. In practice, this means high-value potential clients may be under-identified. Future research could incorporate advanced models (e.g., Random Forest, XGBoost), sampling techniques (e.g., SMOTE), or nonlinear modeling (e.g., spline regression, GAMs) to capture richer relationships.

Overall, this project demonstrates the practical value of statistical modeling in financial marketing. Logistic regression not only enables reliable prediction, but also delivers interpretable, actionable insights. The approach equips banks with a foundation for smarter, data-driven decision-making—and sets the stage for future exploration into more sophisticated modeling techniques.

References:

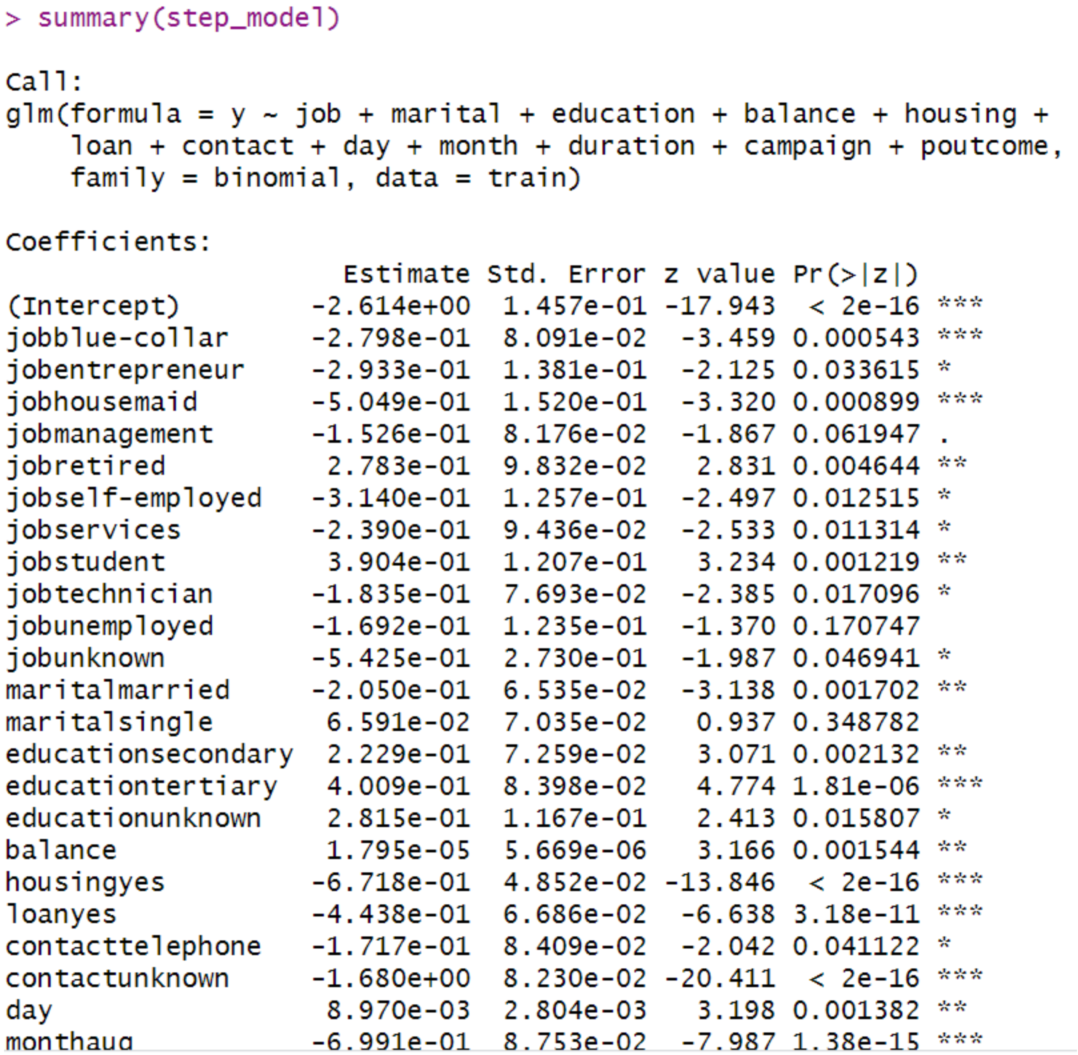
He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263-1284.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Menard, S. (2001). *Applied logistic regression analysis*. SAGE publications.

Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology.

Appendix:

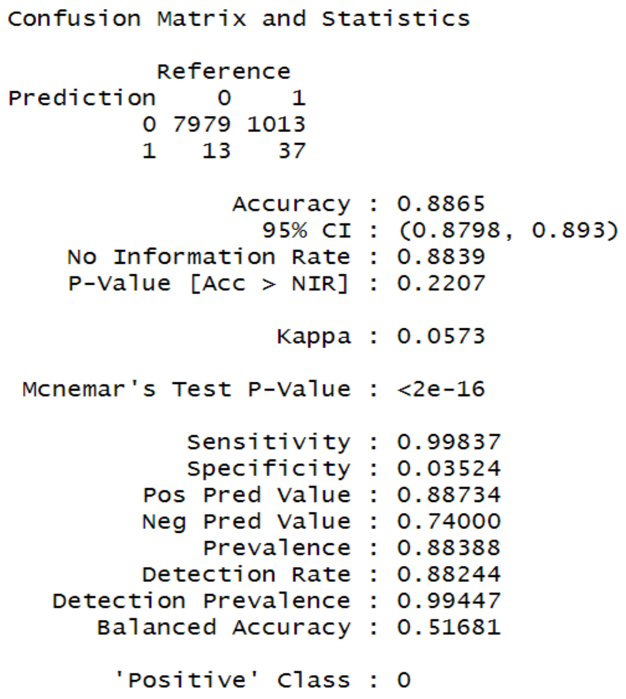


```
> summary(step_model)

Call:
glm(formula = y ~ job + marital + education + balance + housing + 
    loan + contact + day + month + duration + campaign + poutcome, 
    family = binomial, data = train)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -2.614e+00  1.457e-01 -17.943  < 2e-16 ***
jobblue-collar     -2.798e-01  8.091e-02  -3.459 0.000543 ***
jobentrepreneur    -2.933e-01  1.381e-01  -2.125 0.033615 *  
jobhousemaid       -5.049e-01  1.520e-01  -3.320 0.000899 ***
jobmanagement      -1.526e-01  8.176e-02  -1.867 0.061947 .  
jobretired          2.783e-01  9.832e-02   2.831 0.004644 ** 
jobself-employed   -3.140e-01  1.257e-01  -2.497 0.012515 *  
jobservices        -2.390e-01  9.436e-02  -2.533 0.011314 *  
jobstudent          3.904e-01  1.207e-01   3.234 0.001219 ** 
jobtechnician      -1.835e-01  7.693e-02  -2.385 0.017096 *  
jobunemployed      -1.692e-01  1.235e-01  -1.370 0.170747    
jobunknown         -5.425e-01  2.730e-01  -1.987 0.046941 *  
maritalmarried     -2.050e-01  6.535e-02  -3.138 0.001702 ** 
maritalsingle       6.591e-02  7.035e-02   0.937 0.348782    
educationsecondary  2.229e-01  7.259e-02   3.071 0.002132 ** 
educationtertiary   4.009e-01  8.398e-02   4.774 1.81e-06 ***
educationunknown    2.815e-01  1.167e-01   2.413 0.015807 *  
balance             1.795e-05  5.669e-06   3.166 0.001544 ** 
housingyes         -6.718e-01  4.852e-02 -13.846  < 2e-16 ***
loanyes            -4.438e-01  6.686e-02  -6.638 3.18e-11 ***
contacttelephone   -1.717e-01  8.409e-02  -2.042 0.041122 *  
contactunknown     -1.680e+00  8.230e-02 -20.411  < 2e-16 ***
day                 8.970e-03  2.804e-03   3.198 0.001382 ** 
monthaug           -6.991e-01  8.753e-02  -7.987 1.38e-15 ***
```

Figure A1. Selected Variables from Stepwise Logistic Regression (AIC-Based)



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7979 1013
         1   13   37

               Accuracy : 0.8865
                 95% CI : (0.8798, 0.893)
    No Information Rate : 0.8839
    P-Value [Acc > NIR] : 0.2207

                  Kappa : 0.0573

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99837
            Specificity : 0.03524
         Pos Pred Value : 0.88734
         Neg Pred Value : 0.74000
             Prevalence : 0.88388
         Detection Rate : 0.88244
   Detection Prevalence : 0.99447
      Balanced Accuracy : 0.51681

       'Positive' Class : 0
```

Figure A2. Confusion Matrix and Evaluation Metrics – Weighted Logistic Regression

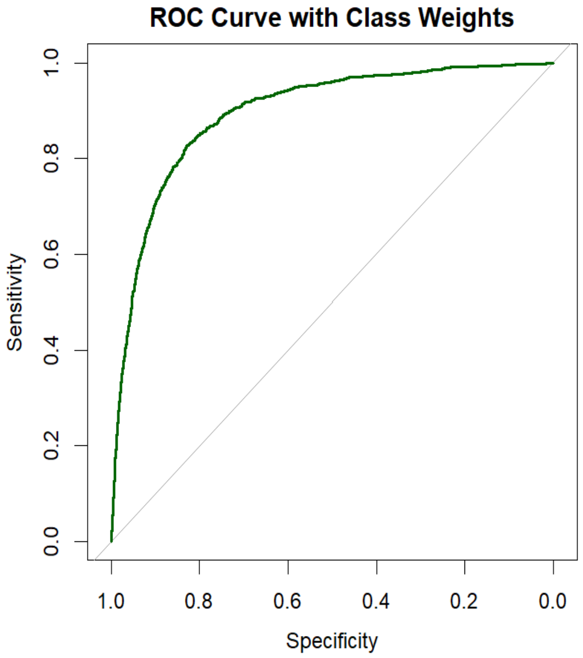


Figure A3. ROC Curve – Weighted Logistic Regression

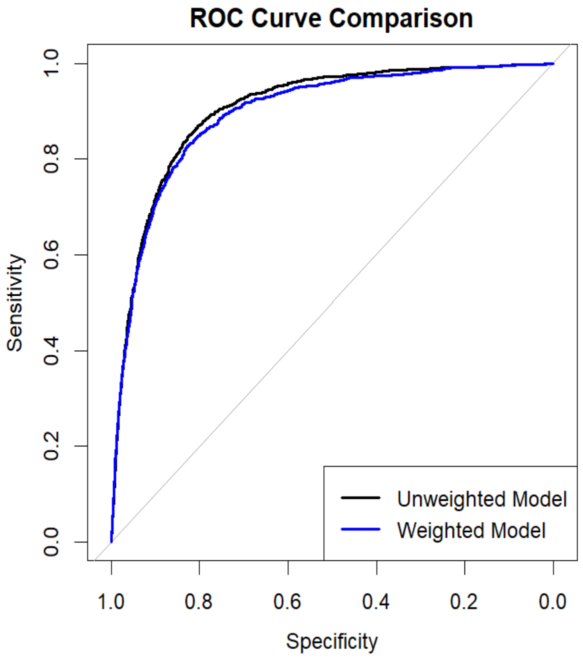
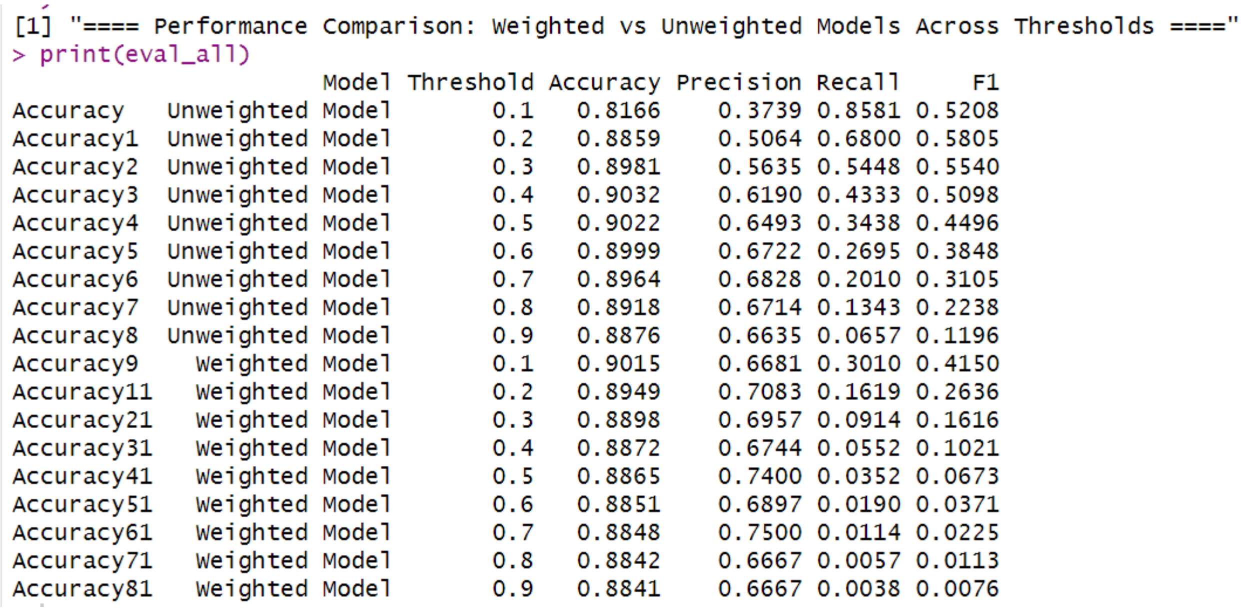


Figure A4. ROC Curve Comparison



```
[1] "==== Performance Comparison: Weighted vs Unweighted Models Across Thresholds ===="
> print(eval_all)
                      Model Threshold Accuracy Precision Recall     F1
Accuracy   Unweighted Model       0.1   0.8166    0.3739 0.8581 0.5208
Accuracy1  Unweighted Model       0.2   0.8859    0.5064 0.6800 0.5805
Accuracy2  Unweighted Model       0.3   0.8981    0.5635 0.5448 0.5540
Accuracy3  Unweighted Model       0.4   0.9032    0.6190 0.4333 0.5098
Accuracy4  Unweighted Model       0.5   0.9022    0.6493 0.3438 0.4496
Accuracy5  Unweighted Model       0.6   0.8999    0.6722 0.2695 0.3848
Accuracy6  Unweighted Model       0.7   0.8964    0.6828 0.2010 0.3105
Accuracy7  Unweighted Model       0.8   0.8918    0.6714 0.1343 0.2238
Accuracy8  Unweighted Model       0.9   0.8876    0.6635 0.0657 0.1196
Accuracy9    Weighted Model       0.1   0.9015    0.6681 0.3010 0.4150
Accuracy11   Weighted Model       0.2   0.8949    0.7083 0.1619 0.2636
Accuracy21   Weighted Model       0.3   0.8898    0.6957 0.0914 0.1616
Accuracy31   Weighted Model       0.4   0.8872    0.6744 0.0552 0.1021
Accuracy41   Weighted Model       0.5   0.8865    0.7400 0.0352 0.0673
Accuracy51   Weighted Model       0.6   0.8851    0.6897 0.0190 0.0371
Accuracy61   Weighted Model       0.7   0.8848    0.7500 0.0114 0.0225
Accuracy71   Weighted Model       0.8   0.8842    0.6667 0.0057 0.0113
Accuracy81   Weighted Model       0.9   0.8841    0.6667 0.0038 0.0076
```

Figure A5. Performance Comparison