MA 678 Final project

Telemarketing Strategies for Term Deposit Subscriptions: Analysis of Customer Response Patterns in a Portuguese Bank

Master of Statistical Practice

2023-2024

Authors:

Yuyang Sun

Course Coordinator:

Fotios Kokkotos

13th December 2023

Contents

1 Abstract

The marketing campaigns for bank institutions are lacking in effectiveness for customers over time. Consequently, economic pressures and intense competition have led many bank institutions to narrow their target markets and focus on more promising customer segments. However, identifying and selecting the target individuals remains a significant challenge. Thus, this report focuses on an explanatory data analysis of the success of a direct telemarketing campaign run by a Portuguese bank to promote subscription to term deposit accounts. The analysis applies logistic regression, and multilevel modeling from simple to complex in determining the key factors driving term deposit subscription using real-world data from a Portuguese bank (Moro et al., 2011).

The methods and insights provided by building multiple models and comparing their performance. Some analysis are conducted in the conclusion

keywords: Explanatory Data Analysis; Telemarketing Campaign; Logistic Regression; Multilevel Modeling

2 Introduction

Marketing strategies in banking institutions typically oscillate between mass campaigns targeted at the public and direct marketing at targeted contacts. Traditional mass campaigns are expensive and have been less effective in attracting customers to subscribe to their products, whereas those specific targets are more likely to be receptive to specific products or services, thereby enhancing campaign efficiency (Ou et al., 2003).

Nevertheless, direct marketing is not without its own set of challenges. As highlighted by Page and Luding (2003), there's a risk for direct marketing strategy due to perceived intrusiveness and privacy violations as the existence of global competition and internal pressure. European banks are compelled to augment their financial assets, especially during the financial crisis. A common strategy adopted in response to this pressure is the offering of long-term deposit plans with higher interest rates, which could potentially attract customers. Telemarketing thus emerges as a cost-effective method for conducting direct campaigns. Telemarketing are usually via phones, by providing a direct, one-on-one conversation. Thus, talking through phones can significantly simplify the process of introducing the range of services bank offered, and enabling customers grasp the key aspects of the products or services being presented.

This report focuses on an explanatory data analysis a direct telemarketing campaign conducted by a Portuguese bank from 2008 to 2010. The analysis leverges logistic regression and multilevel modeling techniques, encompassing various degrees of pooling methods. It will include the examination of null models and variable selection processes. The primary goal is to identify the most influential factors contributing to the successful subscription of term deposits. Additionally, the analysis will compare different models to determine which one demonstrates the highest performance in predicting successful outcomes. The dataset of this report comes from a realistic a Portuguese retail bank, as documented by (Moro et al. 2011).

3 Methodology

3.1 Preliminary Setup

3.1.1 Dataset description and modification

The dataset used in this analysis gives a general view of various client characteristics and their interactions with a Portuguese retail bank's direct marketing campaigns. The dataset, meticulously detailed by Moro et al. (2011), comprises a range of variables, each offering unique insights into client profiles and behaviors. The following provides an overview of each predictor and how it's related to the dataset.

- **Age (age):** This integer variable represents the age of the client. Age is a demographic factor that can influence a client's financial decisions.
- **Job (job):** A categorical variable indicating the client's occupation.
- **Marital Status (marital):** This categorical variable describes the marital status of the client.
- **Education (education):** Representing the education level of the client, this categorical variable includes several educational backgrounds.
- **Credit Default (default):** A binary variable indicating whether the client has credit in default.
- **Balance (balance):** Expressed in euros.
- **Housing Loan (housing):** A binary variable that indicates whether the client has a housing loan.
- **Personal Loan (loan):** Similar to housing, this is a binary variable that denotes whether the client has any personal loan.
- **Contact Communication Type (contact):** This categorical variable describes the method of communication used in the last contact (e.g., 'cellular', 'telephone'). Missing values in 'contact' indicate that some clients' contact methods were not recorded.

- **Last Contact Duration (duration):** An integer variable measuring the duration of the last contact in seconds. It's a crucial variable but should be used cautiously, as it might lead to target leakage if used for prediction.
- **Last Contact Day and Month ('day', 'month'):** Date variables indicate when the last contact with the client was made.
- **Campaign Contacts (campaign):** It counts the number of contacts made during the current campaign for the client, including the last contact.
- **Days Since Last Campaign Contact (pdays):** An integer indicating the number of days that passed since the client was last contacted from a previous campaign. Contains values equal to -1, signifying that the client wasn't contacted previously.
- **Previous Contacts (previous):** Reflecting the number of contacts made before the current campaign for a client.
- **Previous Campaign Outcome (poutcome):** A categorical variable showing the result of the previous marketing campaign.
- **Target Variable - Term Deposit Subscription (y):** The binary target variable indicates whether the client subscribed to a term deposit, serving as the primary focus of this analysis. This variable is the response variable of interest.

The dataset is initially presented in a sequential order but lacks an explicit 'year' attribute. To remedy this, we have introduced a 'year' attribute into the dataset. This addition is pivotal for enabling the consideration of years as a hierarchical structure within the modeling process. Prior to the construction of the model, it is of utmost importance to thoroughly examine the interrelationships among the predictor variables. A deep understanding of these relationships lays a solid foundation for building robust models and facilitates insightful subsequent analyses.

In the initial step of our data analysis, we conducted a thorough check for missing values (NA) within the dataset. Thus, we utilize the entire dataset.

3.1.2 Predictor Analysis

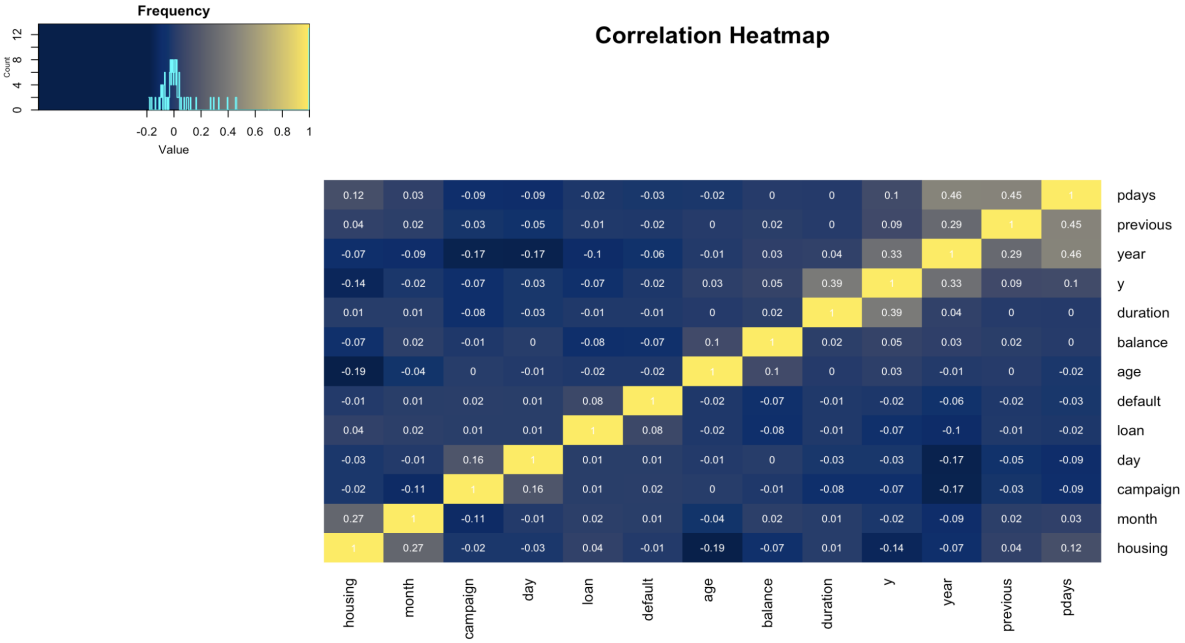

Figure 1: Correlation Heatmap for all Numeric Variables

Figure 1 indicates the correlation heatmap for all numeric variables in the dataset. Disregarding the yellow colors, which represent a variable's correlation with itself, the grey color indicates a positive correlation. It's observed that the response variable 'y' has a relatively positive relationship with both 'year' and 'duration'. On the other hand, darker blue indicates a negative correlation.'y' exhibits a negative correlation with 'housing', but clearly not a strong one. Additionally, it's worth noting that 'pdays', 'previous', and 'years' are also correlated.

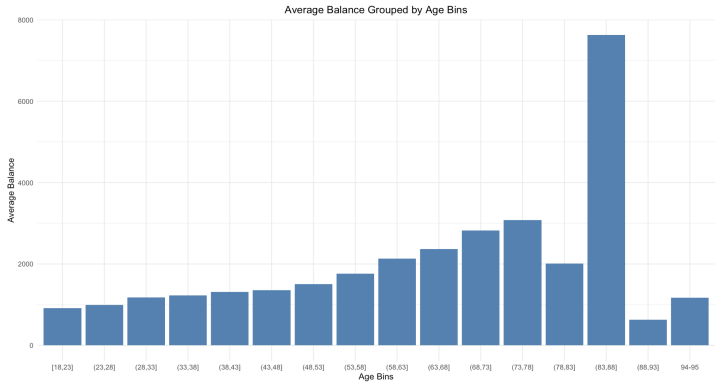

Figure 2: Bar Plot for Average Balance Grouped by Age Bins

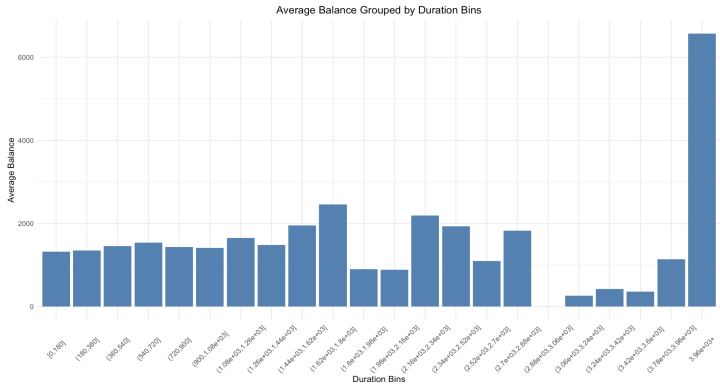

Figure 3: Bar Plot for Average Balance Grouped by Duration Bins

Figure 2 displays the average balance in bank accounts across different age groups. It suggests that younger and middle-aged groups have relatively similar average balances, with amounts gradually increasing with age. Notably, there is a significant increase in the average balance for individuals in the 88-93 age group, which could be due to factors such as the accumulation of retirement savings or decreased spending.

Figure 3 shows the average balance with different durations, likely representing how wealthy people tend to spend more time listening to bank products. The balance is relatively stable across many duration bins but shows notable peaks. For calls that last for a very long time, people tend to be much richer and generally, they have more interest in those bank products like high-interest rate term deposits.

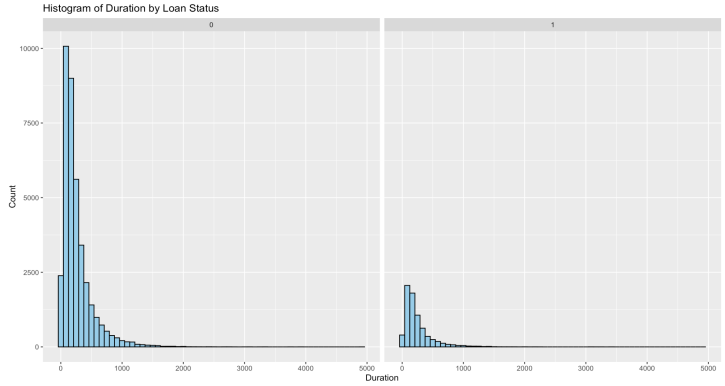

Figure 4: Dual Histogram of Duration by Loan Status

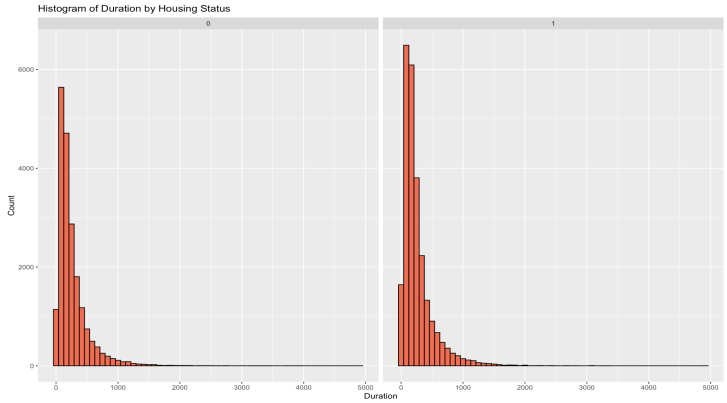

Figure 5: Dual Histogram of Duration by Housing Status

Figure 4 shows the connections between the duration distribution and loan status. The two groups are separated in this plot, with '0' indicating customers without loans and '1' indicating those with loans. Both distributions are right-skewed, but the '1' category exhibits a smaller count as not many people have a personal loan. It turned out that having a personal loan does not affect the overall trends of durations on the phone.

Figure 5 has a similar setup, the only difference is the housing status. Both distributions are right-skewed, indicating that shorter durations are more common than longer ones. It turned out that having a housing loan does not affect the overall trends of durations on the phone.

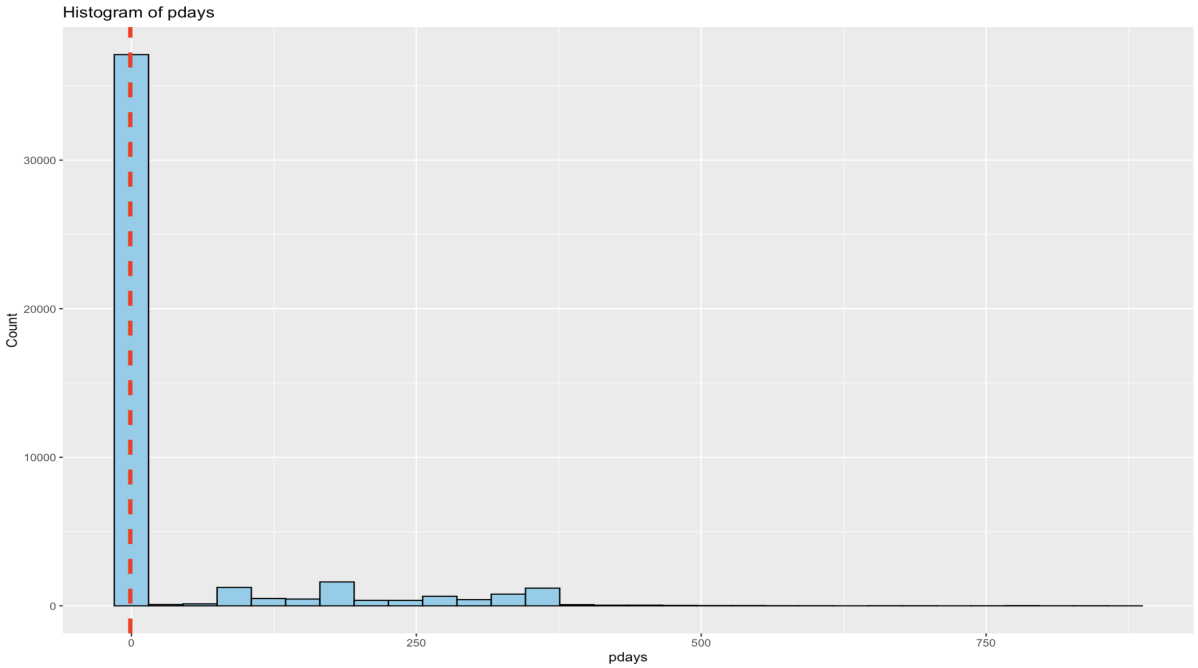

Figure 6: Bar plots for pday

This plot for pday indicates that most of the clients do not have any contact with the bank before this campaign (shown with the red dashed line). This indicates that the bank institution is currently trying some approaches to attract new clients. For the rest of the bars, nearly all of the bars are below 500 days, which approximately within in 1-year range, indicating those contacts are not far from the current campaign.

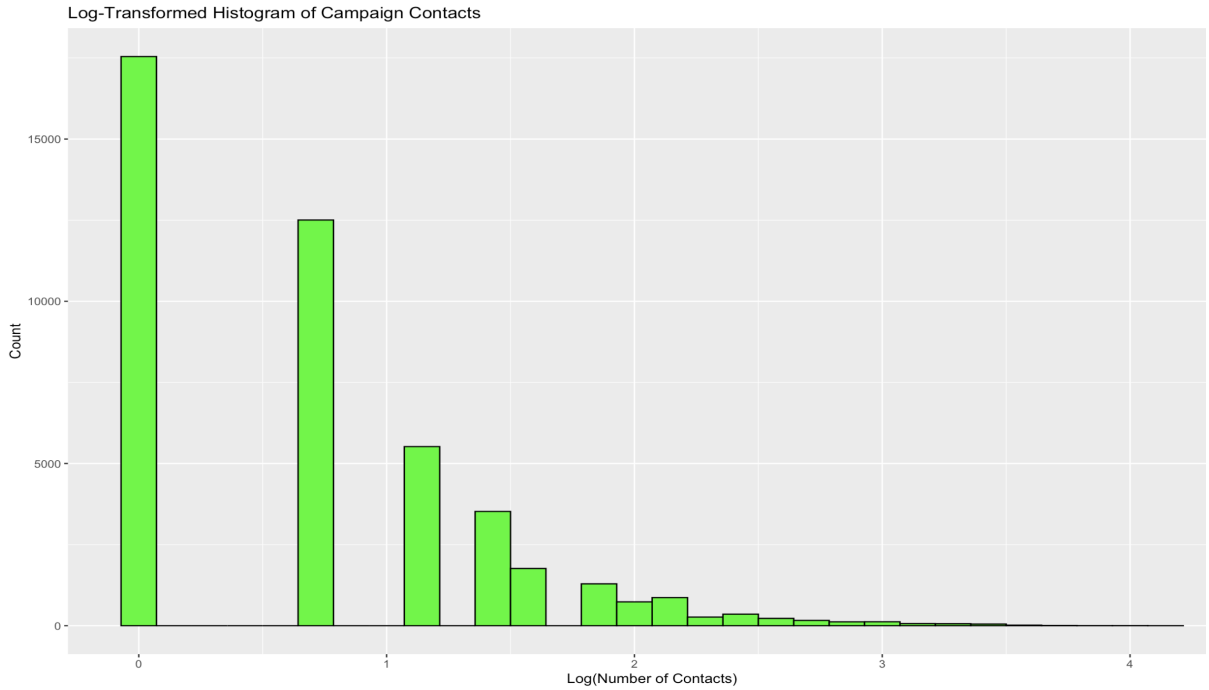

Figure 7: Bar plot for campaign

This histogram illustrates the distribution of campaign contacts after applying a log transformation as taking log reduces skewness. The distribution shows that lower values are much more common, as indicated by the taller bars for lower log(contact) values, suggesting that most campaigns had a relatively small number of contacts, and that could be a major challenge that bank institutions face.

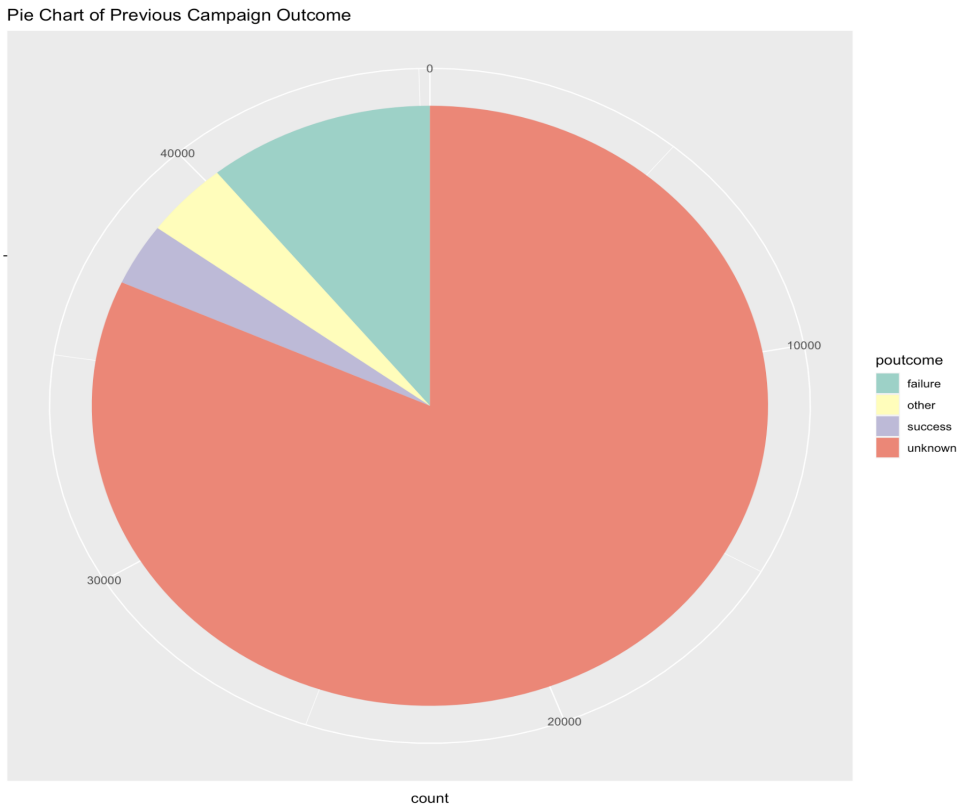

Figure 8: Piechar for poutcome

The pie chart shows the distribution of outcomes from a previous campaign. The largest section, marked as 'unknown', suggests that for many cases the outcome of the previous campaign was not recorded or is unknown.

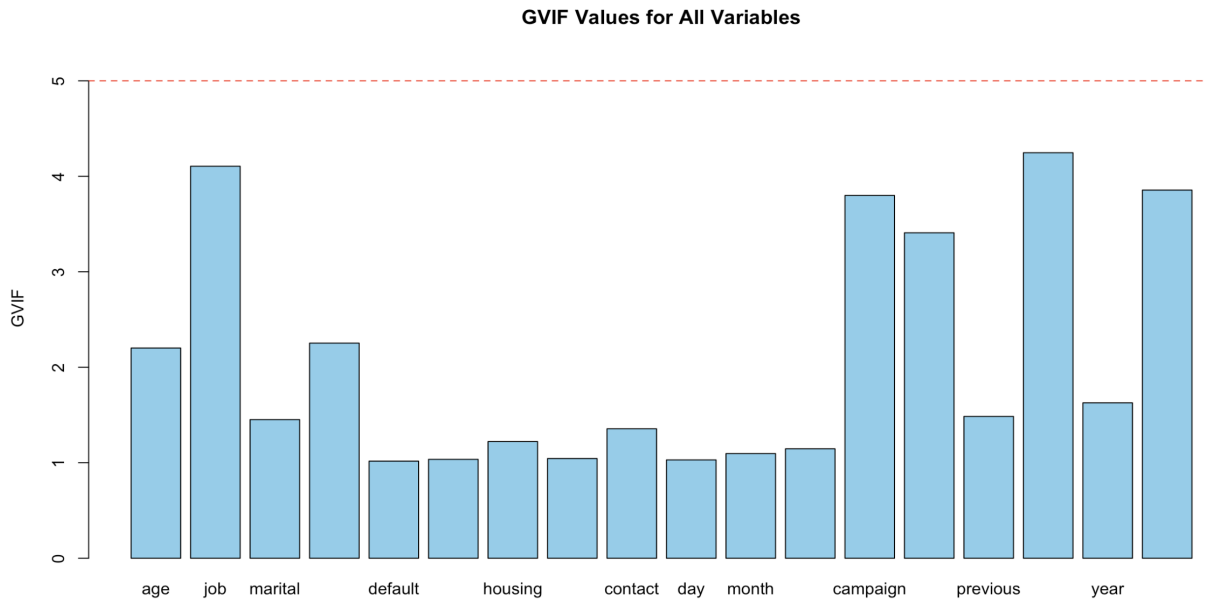

Figure 9: Bar Plot for Multicollinearity Check

The Multicollinearity checking plot displays the Generalized Variance Inflation Factor (GVIF) for different predictor variables in a statistical model. Notice that the red dashed line represents a threshold for concerning levels of multicollinearity, and all VIF values of the predictors in the dataset are below the threshold. Thus, there is no multicollinearity.

Given that no multicollinearity is observed among the predictors, there is no necessity to remove any of them. Consequently, we can fit our models while retaining all predictors. Our next step involves applying a logistic regression model, incorporating various levels of pooling. It's important to note that our response variable is binary, represented as 0 or 1, corresponding to 'fail' and 'success', respectively.

3.2 Model Description

3.2.1 Null model

Other than multilevel models, the null model serves as a baseline for all the models in this project. It tells how the results for a simplistic or minimalistic approach to the problem would be if we against which the performance of more complex models can be compared. In this project, we simply generate:

```
null_model <- glm(y ~ 1, data = train_set, family = binomial())
```

The ~ 1 part specifies the formula for the null model. y represents the success of the subscription of term deposit, and indicates that we are modeling y based on other variables. However, in this specific formula, 1 on the right side of represents a constant term (intercept) without any predictor variables.

3.2.2 Logistic model and Variable Selection

Logistic regression is a statistical modeling technique commonly used for binary classification tasks. In logistic regression, we model the probability of a binary outcome (in our case, the success of the subscription of term deposit) as a function of one or more predictor variables. The model estimates a set of parameters that describe the relationship between the predictors and the probability of the y. This is achieved by using the logistic function to transform a linear combination of the predictor variables into a probability scale, ensuring that the predicted probabilities are bounded between 0 and 1. The model parameters ($\beta_0, \beta_1, \ldots, \beta_p$) are learned from the data through methods such as maximum likelihood estimation, and the model can be used to make predictions and infer the influence of predictors on the binary outcome.

In math, the logistic model can be represented as follows:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p)}}$$

Where:

- $P(Y = 1|X)$ is the probability of the binary outcome Y being 1 given the predictor variables X.

- $\beta_0, \beta_1, \ldots, \beta_p$ are the model parameters to be estimated. - $X_1, X_2, \ldots, X_p$ are the predictor variables.

- e is the base of the natural logarithm, which is a constant.

In the process of refining our logistic regression model for the bank dataset, we employed a systematic approach known as backward selection to enhance the model's effectiveness. This method is particularly useful when the initial model with all predictors includes some with non-significant p-values, indicating that they may not contribute meaningfully to the model.

Backward selection begins by fitting the logistic regression model with all available predictors. We then indicated predictors by their p-values. The key strategy involves iteratively removing the predictor with the highest p-value. As the highest p-value is considered insignificant. This step is followed by refitting the model with the remaining predictors.

In each iteration, we re-evaluate and re-fit the model, ensuring that the exclusion of a predictor does not compromise the model's overall explanatory power. This process is repeated, continually eliminating the least significant predictor until all remaining predictors in the model demonstrate statistically significant contributions, as evidenced by p-values below a threshold, in our project, we use a threshold equal to 0.05.

For our bank dataset, this variable selection procedure entailed six iterations. Each iteration refined the model by discarding the least impactful predictor, culminating in a significantly enhanced logistic model. The final iteration yielded a model where all predictors had p-values below the chosen significance threshold, indicating that each remaining predictor is statistically significant and contributes meaningfully to the model's predictive capacity.

To be more specific, here is the r-code indicating the processes of variable selection as follow:

Action	**Details**
Removed: age	p-value: 0.72390832120019
Removed: job	p-value: 0.705746554127651
Removed: previous	p-value: 0.65466983870508
Removed: default	p-value: 0.546380456636024
Removed: poutcome	p-value: 0.327293615317516
Removed: marital	p-value: 0.061638671811334
Final model summary: glm(formula = y ~ education + balance + housing + loan + contact + day + month + duration + campaign + pdays + year, family = binomial(), data = train_set)	

Table 1: Backward Selection Process and Final Model

3.2.3 Probit Function

The probit function is similar to logistic regression. In a probit model, the probability of an event occurring is linked to the predictors through the cumulative distribution function (CDF) of the standard normal distribution. Mathematically, it can be expressed as follow:

$$P(Y = 1) = \Phi(\beta_0 + \beta X)$$

where Φ is the CDF of the standard normal distribution, β_0 is the intercept, β is the vector of coefficients, and X is the vector of predictors.

3.2.4 No Pooling

While building the multilevel model, we have to ensure that the dataset has a hierarchical structure. In the banking institution, the year attribute could be considered a key factor for building the models.

In a no-pooling model, each year factor (indexed by n) has a separate chance-of-success parameter $\theta_n \in [0, 1]$, and these θ_n values are assumed to be independent of each other.
To parameterize this model, we consider each year as a separate factor for the "no pooling" approach and make comparisons between different years. With this parameterization, the likelihood function describes the number of successes y_n by utilizing the log-odds of success α_n in the following manner:

$$p(y_n|\alpha_n) = \text{Binomial}(y_n|K_n, \text{logit}^{-1}(\alpha_n))$$

- y_n – the number of successes of term deposit subscription for year n.
- K_n – the number of clients for year n.
- $\text{logit}^{-1}(\alpha_n)$ – The inverse logit function, which converts α_n to the probability scale(θ_n).

Suppose the y_n values are independent conditional on θ, the likelihood for the entire dataset is:

$$p(y|\alpha) = \prod_{n=1}^{N} \text{Binomial}(y_n|K_n, \text{logit}^{-1}(\alpha_n))$$

This illustrates how data is generated in a no-pooling model, where the probability of success for each client is individually modeled using the log-odds parameter α_n.

3.2.5 Complete Pooling

The complete pooling model represents the chance of success for all term deposits regardless of years.

First, we have to ensure that all success/failure of subscription of term deposits are independent, the probability distribution for each year's successful subscription y_n is:

$$p(y_n|\theta) = \text{Binomial}(y_n|K_n, \theta).$$

Assuming each player is independent leads to the complete data likelihood

$$p(y|\theta) = \prod_{n=1}^{N} \text{Binomial}(y_n|K_n, \theta).$$

Log-odds α, which are defined by the logit transform as

$$\alpha = \text{logit}(\theta) = \log\left(\frac{\theta}{1-\theta}\right).$$

3.2.6 Partial Pooling

It's the combination of complete pooling and partial pooling and balances the use of individual and group information in data analysis. In this case, we leverage information from different years to make predictions, allowing each year to influence the results while still considering the patterns across all years.

$$y_{ij} = \beta_0 + \beta X_{ij} + b_i + \epsilon_{ij} \tag{1}$$

Where:

- y_{ij} is the response for the j-th observation in the i-th group.
- β_0 is the overall intercept across all groups.
- β represents the fixed effects coefficients, reflecting shared effects across all groups.
- X_{ij} is the matrix of predictors for the j-th observation in the i-th group.

- b_i is the random effect for the i-th group, capturing group-specific deviations.
- ϵ_{ij} is the residual error term, assumed to be normally distributed.

In this model, β captures the overall trends across groups, while b_i allows each group to have its unique adjustments, leading to the concept of partial pooling.

3.3 Train Set and Test Set

In the context of this project, we aim to partition the dataset into distinct sets: one for model training and another for testing. Specifically, we have opted for an 80-20 split, allocating 80 percent of the data for training the model and reserving 20 percent for rigorous testing and evaluation.

4 Results

4.1 Model Results

4.1.1 Null Model

Coefficients	**Estimate**	**Std. Error**	**z value**	**Pr(> |z|)**
(Intercept)	-2.02056	0.01636	-123.5	$< 2 \times 10^{-16}$ ***

The output of the null model suggests that while we are predicting the probability of a successful subscription to a term deposit, denoted as y. The intercept's coefficient, at -2.02056, is significantly different from zero with a p-value less than 2e-16, indicating a meaningful baseline log odds of success. This negative intercept suggests a lower likelihood, less than 50 percent, of successfully subscribing to a term deposit under the model's simplistic conditions.

4.1.2 Logistic Model

Variable	Estimate	Std. Error	z value	Pr(> |z|)
Intercept	-3219	72.66	-44.298	
dducationsecondary	0.3019	0.06722	4.491	7.08e-06
educationtertiary	0.5856	0.06953	8.422	¡ $< 2 \times 10^{-16}$
educationunknown	0.2999	0.1120	2.679	0.007385
balance	1.539e-05	5.635e-06	2.730	0.006327
housing	-0.8107	0.04315	-18.787	¡ $< 2 \times 10^{-16}$
loan	-0.3950	0.06651	-5.939	2.87e-09

This model shows that several factors significantly impact the likelihood of a client subscribing to a term deposit (y). Notably, higher education levels, such as secondary and tertiary, increase the probability, as indicated by positive coefficients. The average balance (balance) and contact duration (duration) also positively influence the subscription likelihood. In contrast, having a housing loan (housing) and a personal loan (loan) are negatively associated with subscription.

4.1.3 Probit Model

The intercept, also with a substantial negative value, is similar to logistic regression. Education also emerges as a significant predictor, with higher levels of education, such as secondary and tertiary, positively correlating with the likelihood of subscription. This suggests that clients with higher educational backgrounds are more inclined to subscribe to term deposits.

Variable	Estimate	Std. Error	z value	Pr(> $\|z\|$)
Intercept	-1.661	0.04518	-36.755	¡ $< 2 \times 10^{-16}$
educationsecondary	0.1597	0.03288	4.856	1.20e-06
educationtertiary	0.3164	0.03422	9.245	¡ $< 2 \times 10^{-16}$
educationunknown	0.2497	0.05567	4.485	7.30e-06
balance	9.855e-06	2.794e-06	3.527	0.00042
housing	-0.5348	0.02174	-24.600	¡ $< 2 \times 10^{-16}$
loan	-0.3790	0.03249	-11.666	¡ $< 2 \times 10^{-16}$

Table 2: Selected Predictors from the Probit Model

The variable balance shows a positive relationship with subscription likelihood, but its impact is relatively smaller. This indicates that clients with higher balances are slightly more likely to subscribe.

On the contrary, having loans, both housing (housing) and personal (loan), are associated with a decreased likelihood of subscription. This negative relationship might reflect the financial constraints or risk aversion among clients with outstanding loans.

4.1.4 No Pooling Model

Variable	Estimate	Std. Error	z value	Pr(> $\|z\|$)
Intercept	-4.780	0.1763	-27.112	¡ $< 2 \times 10^{-16}$
educationsecondary	0.2027	0.1159	1.749	0.0803
educationtertiary	0.1511	0.1262	1.197	0.2312
educationunknown	-0.1823	0.2348	-0.777	0.4374
balance	2.405e-06	1.232e-05	0.195	0.8453
housing	-0.1972	0.08554	-2.305	0.0212
loan	-0.2041	0.1031	-1.980	0.0477
contacttelephone	0.4599	0.1670	2.754	0.0059
contactunknown	-0.5382	0.1001	-5.379	7.5e-08
duration	0.005136	0.0001098	46.759	¡ $< 2 \times 10^{-16}$

Table 3: Selected Predictors from the Model and Their Relationship to Target Variable from No Pooling Year 2008

For this model with the year equal to 2008, the intercept is smaller than those logistic regressions. The coefficients for education levels, such as secondary (0.2027) and tertiary (0.1511), though positive, are not statistically significant at conventional levels, indicating a marginal influence on the likelihood of subscribing to term deposits, that is different from those logistic regression models.

The variables housing and loan, with negative coefficients of -0.1972 and -0.2041 respectively, and significant p-values, suggest that clients with housing loans or personal loans are

less likely to subscribe to term deposits. Notably, contact methods such as contacttelephone (0.4599) and contactunknown (-0.5382) show significant effects, indicating the importance of communication mode in the success of the marketing campaign.

The duration of the last contact, with a positive coefficient (0.005136) and a very significant p-value, influences the subscription likelihood while the duration on the phone is long.

Variable	Estimate	Std. Error	z value	Pr(> |z|)
Intercept	-2.580	0.1287	-20.037	¡ 2×10^{-16}
educationsecondary	0.2080	0.09291	2.239	0.025154
educationtertiary	0.6828	0.09519	7.173	7.34e-13
educationunknown	0.3305	0.1549	2.134	0.032834
housing	-1.246	0.06070	-20.525	¡ 2e-16
loan	-0.5125	0.1017	-5.038	4.71e-07
contacttelephone	-0.4648	0.1134	-4.100	4.13e-05
contactunknown	3.629	1.180	3.076	0.002098
duration	0.003640	0.0001105	32.954	¡ 2×10^{-16}

Table 4: Selected Predictors from the 2009 Model and Their Relationship to Target Variable

While year 2009, indicates significant relationships between several predictors and the likelihood of a client subscribing to a term deposit with a relatively larger intercept.

Educational background again plays a significant role, with educationtertiary (0.6828) showing a strong positive influence, and educationsecondary (0.2080) also contributing positively. Other factors such as contact and duration also are quite similar to the year 2008.

Variable	Estimate	Std. Error	z value	Pr(> |z|)
Intercept	-0.8780	0.2391	-3.672	0.000240
educationsecondary	0.5847	0.1734	3.373	0.000744
educationtertiary	0.5575	0.1757	3.174	0.001505
educationunknown	0.5832	0.2434	2.396	0.016576
housing	0.01627	0.1234	0.132	0.895138
loan	-0.4003	0.2036	-1.966	0.049269
contacttelephone	-0.5170	0.1634	-3.164	0.001556
contactunknown	-1.641	0.2390	-6.863	6.72e-12
duration	0.003838	0.0003238	11.852	¡ 2×10^{-16}
campaign	-0.1247	0.03545	-3.517	0.000436

Table 5: Selected Predictors from the 2010 Model and Their Relationship to Target Variable

The situation is a little bit different from the previous years. While education, contact play important roles, campaign shows a negative value of (-0.1247), suggesting that repeated contacts might not always be effective.

4.1.5 Complete Pooling Model

Variable	Estimate	Std. Error	z value	Pr(> |z|)
Intercept	-2.912	0.08646	-33.686	¡ 2×10^{-16}
educationsecondary	0.2968	0.06320	4.696	2.66e-06
educationtertiary	0.5920	0.06529	9.067	¡ 2×10^{-16}
educationunknown	0.4755	0.1044	4.554	5.26e-06
balance	1.521e-05	5.013e-06	3.034	0.00241
housing	-1.042	0.04158	-25.060	¡ 2×10^{-16}
loan	-0.7491	0.06417	-11.674	¡ 2×10^{-16}
contacttelephone	-0.08937	0.07629	-1.172	0.24138
contactunknown	-1.390	0.06481	-21.444	¡ 2×10^{-16}
day	-0.007012	0.002317	-3.026	0.00247
month	0.03582	0.006147	5.827	5.65e-09
duration	0.004051	0.0000689	58.797	¡ 2×10^{-16}
campaign	-0.1243	0.01120	-11.095	¡ 2×10^{-16}
pdays	0.002799	0.0001642	17.049	¡ 2×10^{-16}

The negative intercept (-2.912), which is similar to no-pooling methods. Similarly, education, and duration play a positive role, whereas factors like contact, housing, and loans play a negative role.

4.1.6 Partial Pooling Model

Variable	Estimate	Std. Error	z value	Pr(> |z|)
Intercept	-1.68950	0.67477	-2.504	0.012287
educationsecondary	0.30178	0.06706	4.500	6.79e-06
educationtertiary	0.58536	0.06941	8.433	¡ 2×10^{-16}
educationunknown	0.30001	0.11165	2.687	0.007210
balance	0.04691	0.01719	2.729	0.006354
housing	-0.80992	0.04433	-18.270	¡ 2×10^{-16}
loan	-0.39590	0.06657	-5.947	2.72e-09
contacttelephone	-0.26975	0.08266	-3.264	0.001100
contactunknown	-0.52218	0.07325	-7.128	1.02e-12
day	0.05037	0.01990	2.531	0.011374
month	0.07900	0.01919	4.116	3.85e-05
duration	1.12359	0.01887	59.544	¡ 2×10^{-16}
campaign	-0.12841	0.03306	-3.884	0.000103
pdays	-0.04077	0.01857	-2.196	0.028112

The results are a little different. Day (0.05037) and month (0.07900) are positively correlated with subscriptions, indicating potential seasonal or time-related trends in client subscriptions.

The duration of the last contact (duration at 1.12359) is a strong predictor, emphasizing the effectiveness of longer, quality interactions in convincing clients are potentially increase the success of subscriptions.

4.2 Model Performance

Model	AIC	BIC	ROC_AUC	Sensitivity	Specificity	RMSE
Partial Pooling	17234.174	17361.613	0.5121	0.0031	1.0000	0.9382
Complete Pooling	19355.757	19474.700	0.8649	0.9792	0.2066	0.2862
Probit Model	19349.417	19468.361	0.8654	0.9823	0.1754	0.2857
Logistic Model	17209.710	17337.149	0.9022	0.9711	0.3242	0.2746
Null Model	26118.131	26126.627	0.5000	1.0000	0.0000	0.3210
No Pooling (Year 2008)	5549.771	5661.861	0.8367	0.9884	0.1147	0.3084
No Pooling (Year 2009)	8716.710	8820.083	0.6304	0.6768	0.3109	0.4929
No Pooling (Year 2010)	2534.007	2613.149	0.8335	0.7391	0.7668	0.4271

Table 6: Comparison of Model Performance Metrics

To compare model performances, we first look into AIC and BIC values. Lower values of AIC and BIC are indicative of better model fits. In our analysis, the ”No Pooling (Year 2010)” model consistently exhibited the lowest AIC and BIC scores among the models considered, suggesting its superiority in terms of model fit and complexity.

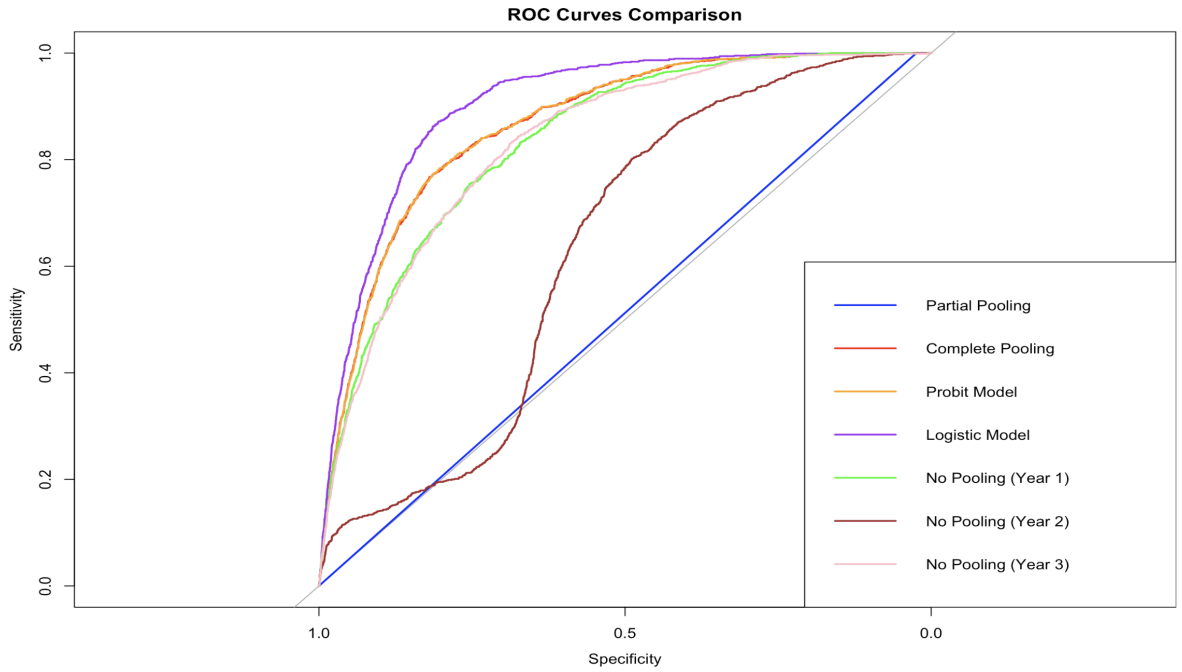

Figure 10: ROC AUC Curves for Models

The ROC AUC curve: ranging from 0 to 1, measures the models' ability to discriminate between positive and negative cases. Higher values indicate better discrimination. The Logistic Model shows the highest ROC AUC value (0.9022).

From this graph, we can also notice that partial pooling might seem somewhat unrealistic or overly idealized. In contrast, the performance metrics of other models make more sense. Especially, for logistic models, plus it has a low value of RMSE. As a result, it may be considered the preferred choice for predicting term deposit subscriptions within the scope of our analysis.

5 Conclusion

In summation, factors such as education level, contact method, and contact duration could influence customer behavior in banking institutions. Specifically, in higher levels of education, such as secondary and tertiary, people tend to subscribe to the term deposit. Also, the duration of contact plays an important role as more duration via telephones can lead to a potential increase in the likelihood of subscription. Among the models evaluated, the logistic model is the most effective with lower AIC/BIC and RMSE values and high ROC area compared with other models. This insight provides a rough framework for banking institutions to refine their marketing strategies and optimize customer engagement efforts.

To address this in future studies, we plan to implement the Synthetic Minority Over-sampling Technique (SMOTE). This approach will help balance our dataset, thereby enhancing the robustness and generalizability of our models.

Additionally, we are considering incorporating Random Forest models in our subsequent analyses. Random forest models average and summarize the behaviors of all models, which gives concrete conclusions other than just comparing models. Moreover, we aim to delve into more sophisticated methodologies, such as deep learning algorithms and advanced machine learning techniques. These approaches are renowned for their ability to model complex, non-linear relationships and interactions within data, which can help bank sectors with direct telemarketing campaigns with more complicated situations.

6 Bibliography

Fawcett, T. (2005). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

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

Page, C., Luding, Y. (2003). Bank managers' direct marketing dilemmas–customers' attitudes and purchase intention. International Journal of Bank Marketing, 21(3), 147-163.

Ou, C., Liu, C., Huang, J., Zhong, N. (2003). On data mining for direct marketing. In Proceedings of the 9th RSFDGrC conference (Vol. 2639, pp. 491–498).