Bank Marketing Case Study

Eva Beyebach, Pablo Chacon, & Danya Saed

I. Executive Summary

The case study aimed to develop Logistic Regression and Linear Discriminant Analysis (LDA) models to predict client subscription to term deposits from data of a Portuguese bank's marketing campaign. The objective was to determine which model provides superior predictive accuracy and to identify key influencing variables. This entailed processing a dataset with known issues, such as a significant number of "unknown" values, to ensure a robust analysis. A thorough literature review positioned our work within the context of existing research, emphasizing the established value of logistic regression in financial predictive analytics and the potential for LDA to offer insights into customer behavior patterns.

The results of the models showed that Logistic Regression with unknown variables included yielded the highest overall accuracy. However, it was the LDA model that demonstrated the highest specificity, suggesting better performance in predicting non-subscribers—a critical insight given the imbalanced nature of the data with a low number of term-deposit subscribers. Key variables such as contact method, month of contact, and previous campaign outcomes were identified as significant predictors. The study recommends prioritizing cellular contact over telephone, focusing on the month of March, and leveraging macroeconomic indicators to target customers more effectively. This analysis provides a foundation for strategic marketing and resource allocation to optimize the bank's campaign efforts.

II. The Problem

The objective of this case study was to develop and compare Logistic Regression and Linear Discriminant Analysis (LDA) models to predict client subscription (yes or no) to term deposits, based on data from direct marketing campaigns of a Portuguese financial institution, primarily conducted through phone calls. We aimed to construct and evaluate multiple Logistic and LDA models to determine which offers superior accuracy and predictive capability. Additionally, we intended to identify the variables most critical to the performance of these models.

Our goal was to pinpoint the optimal model and predictor variables for identifying potential customers likely to subscribe to a term deposit. Achieving this would allow us to more effectively allocate the firm's resources toward prospective customers, optimizing call efforts and targeting strategies. This report will provide a brief overview of the literature pertinent to our study, followed by a discussion of the methodology employed, the steps undertaken to process the dataset, and an analysis of our findings, all presented in a manner befitting an academic report.

III. Review of Related Literature

In their ResearchGate publication, "Prediction of Term Deposit in Bank Using Logistic Model," Jiang, Wang, and Zhao (2022) provide an in-depth analysis of how logistic regression can assist banks in predicting customer propensity to open a term deposit. This study represents a significant contribution to the field by refining the application of logistic regression—a method with a robust statistical foundation—for a more nuanced understanding of customer behavior within the banking sector. Building on the work of other researchers, such as Hou et al. (2022), Khan et al. (2022), and Dutta et al. (2020), who have explored a variety of algorithms from naive Bayes to decision trees, including advanced techniques like convolutional-GRU for similar predictive purposes, the team led by Jiang highlights the practical efficacy of logistic regression. Their research delves into the core of leveraging sophisticated data analytics to enhance bank marketing strategies, specifically targeting potential term deposit subscribers. By identifying the most effective ways to allocate banking resources to attract new clients, their findings significantly contribute to the broader dialogue on maximizing the use of predictive analytics in finance. This work underscores the importance of logistic regression in forecasting financial trends and shaping future banking strategies.

IV. Methodology

In this case study, we evaluate the outcomes of two logistic regression models and two Linear Discriminant Analysis (LDA) models. Initially, the data were cleansed to ensure compatibility with the models. A notable portion of the data contained the label "unknown," leading us to create a duplicate dataset, named "bank_unk," which included all observations labeled as "unknown." For the original dataset, referred to as "bank," we excluded all observations with "unknown" values. All categorical variables were converted into factor variables for this study. Additionally, the "duration" variable was omitted from all models, as its inclusion would not contribute to a realistic model, as outlined in the provided documentation.

The data were divided using a 70/30 ratio for the training and test sets, respectively. The first logistic regression model was developed using the dataset without "unknown" values, which comprised 3,089 observations and 22 variables. The training set consisted of 2,162 observations, while the testing set included 927 observations. The execution of the vif() function revealed multicollinearity issues, necessitating the removal of "nr.employed" and "euribor3m" from the model.

Subsequently, a new logistic regression model was formulated using the "bank_unk" dataset, which included all observations and contained 4,014 observations. Similar to the initial model, "euribor3m" and "nr.employed" were removed due to multicollinearity concerns.

Logistic regression models operate under certain assumptions to ensure the accuracy of their predictions. These assumptions encompass a linear relationship between the predictor variables and the log odds of the dependent variable, the absence of multicollinearity, no significant outliers among predictors, and the normal distribution and homoscedasticity of the data. Deviations from these assumptions could compromise the validity of the results.

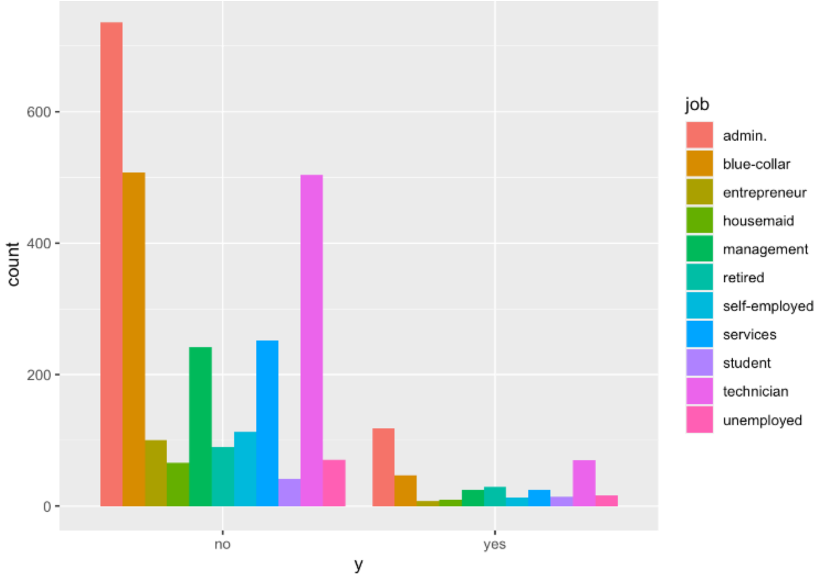
Linear Discriminant Analysis operates as a linear method used for classification and dimensionality reduction, frequently applied in feature extraction for pattern classification challenges. The foundational assumptions of this model include the normal distribution of the data and identical covariance matrices across all classes.

V. Data

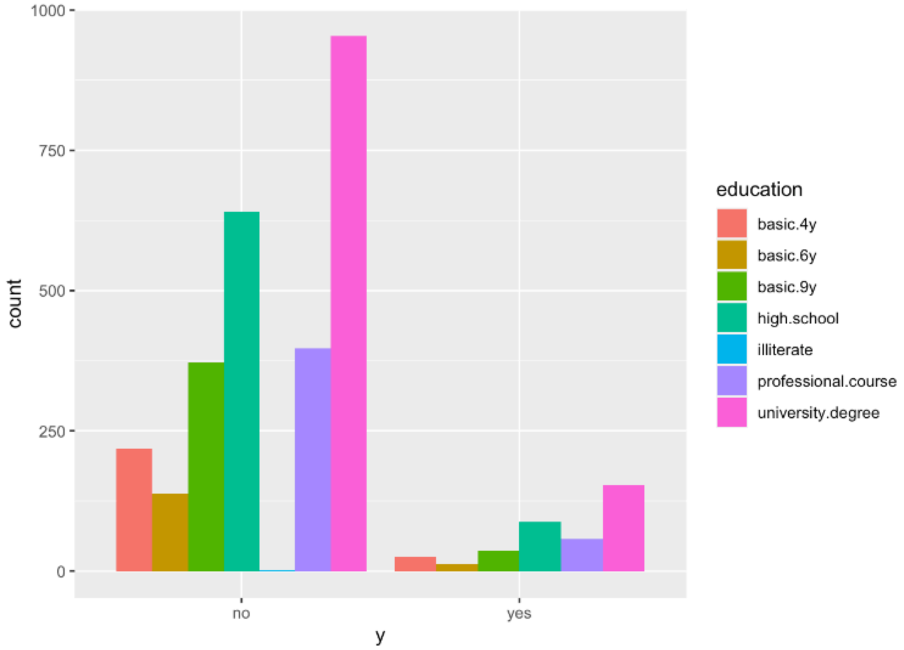
The `*bank*` dataset, comprising 4,119 observations from a Portuguese bank's marketing campaign, includes 21 variables that capture demographic, macroeconomic, and historical campaign performance data. Initial data cleaning revealed no missing values, streamlining the preprocessing phase. However, the dataset contained 1,230 "unknown" variables, potentially compromising data integrity and model results. To address this, a parallel dataset, bank_unk, was established to retain these "unknown" entries for comparative analysis, and a bank_num data frame was created to focus on the numeric variables, facilitating outlier detection and correlation analysis.

Outlier examination within the `*bank_num*` data frame indicated that only `*age*` and `*campaign*` variables required modification. Individuals over the age of 68 were capped at 68, and campaign contacts above 5 were standardized to 5. This adjustment aimed to mitigate the skewing effects of these outliers on the model's accuracy. The rest of the numerical variables, including `*cons.conf.indx*` and `*nr.employed*`, displayed no significant outliers warranting further adjustment.

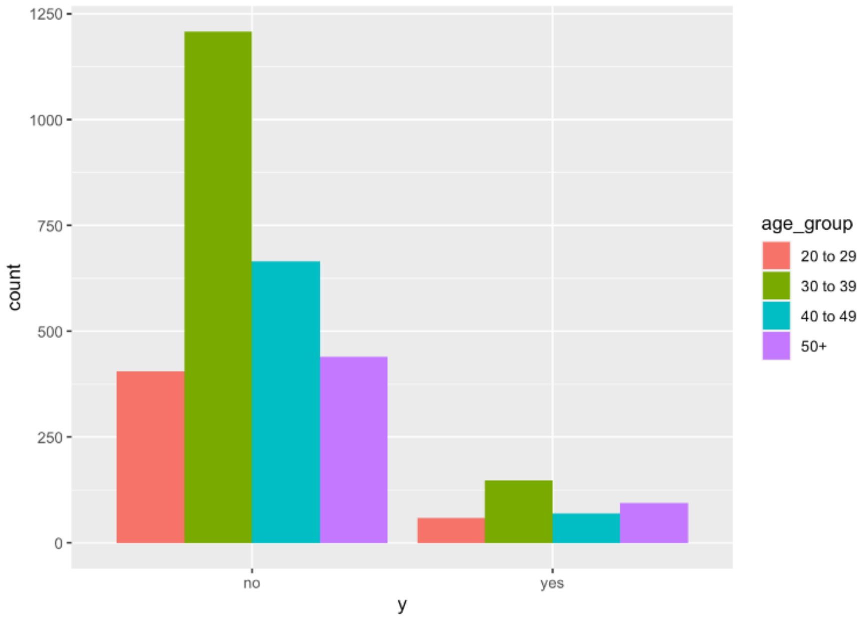
Visual analysis of the data highlighted a stark imbalance, with 88% of respondents not subscribing to a term deposit. This discrepancy could lead to adverse effects on model outcomes.



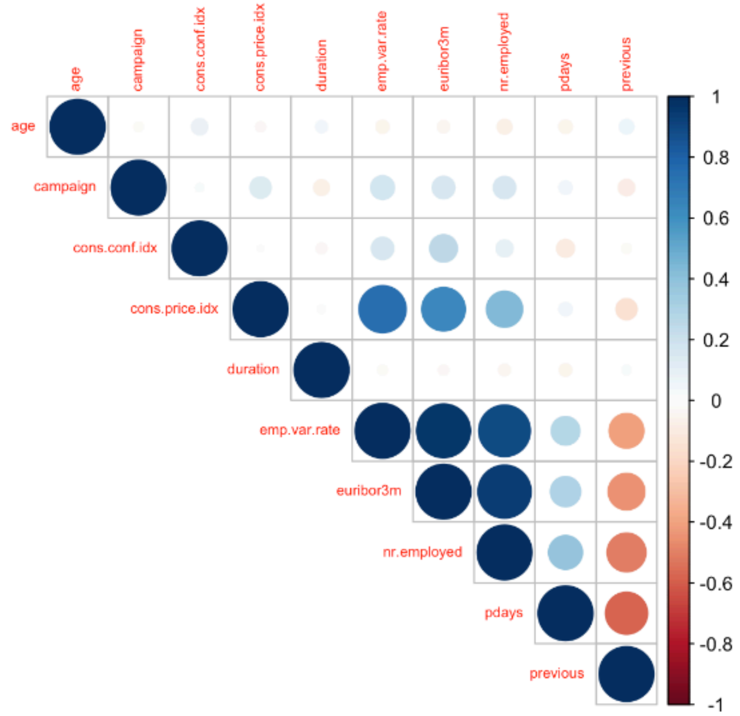
A closer look at demographics showed that individuals with administrative jobs, followed by blue-collar and technician roles, were the most common both in subscribing and not subscribing to term deposits.



Educational insights revealed that university graduates formed the largest group among both subscribers and non-subscribers, with the majority recorded as "no."



The majority of term deposit subscribers fall within the 30-39 age bracket, with the subsequent most substantial group being individuals aged 50 and above. Interestingly, the 40-49 age cohort, despite being the second-largest in the dataset, ranks third in terms of subscription rates. This discrepancy warrants further investigation to understand the underlying factors influencing these age-specific behaviors.



The correlation matrix indicates strong positive correlations between the euribor 3-month rate, employment variation rate, and the number of employees. A strong negative correlation is observed between the number of previous contacts and days since the last contact. These findings suggest multicollinearity concerns that could impact the predictive accuracy of models using this dataset.

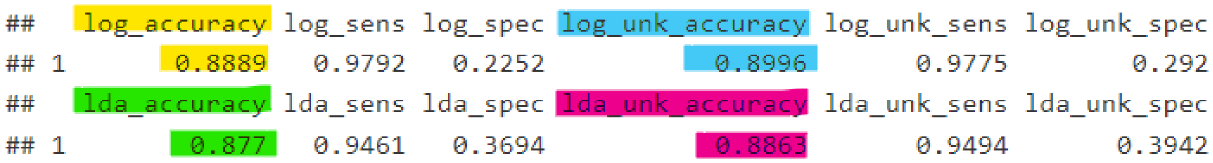
A logistic regression model was then employed to predict term deposit subscriptions, converting the dependent variable `*y*` into a binary numeric factor. Initial modeling efforts uncovered multicollinearity concerns, particularly with `*emp.var.rate*`, `*euribor3m*`, and `nr.*employed*`, necessitating the removal of `*nr.employed*` and subsequently `*euribor3m*` to refine the model. After adjustments, the model's predictive power was reassessed, showing an improvement in specificity when "unknown" variables were included. This finding suggests that "unknown" variables play a crucial role in model accuracy and should not be disregarded.

Finally, an LDA model was utilized to cluster term deposit subscribers by characteristics. The model, run on both the original and "unknown" variable-inclusive datasets, showed that while the LDA model had slightly lower overall accuracy and sensitivity than the logistic models, it offered the highest specificity. This indicates that the LDA model might be more adept at correctly predicting non-subscribers to term deposits, particularly when "unknown" variables were considered. The LDA's ability to cluster non-subscribers suggests a shared profile among this group, which could be leveraged for more targeted marketing strategies.

VI. Findings (Results)

Based on overall accuracy, the best-performing models are as follows:

- Logistic Regression with unknowns: 0.8996
- Logistic Regression without unknowns: 0.8889
- LDA model with unknowns: 0.8863
- LDA model without unknowns: 0.877

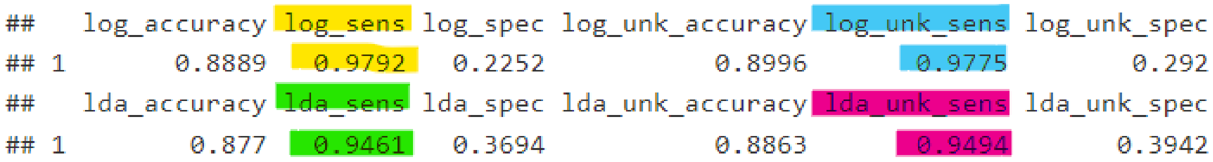


```
##   log_accuracy log_sens log_spec log_unk_accuracy log_unk_sens log_unk_spec
## 1       0.8889   0.9792   0.2252           0.8996       0.9775        0.292
##   lda_accuracy lda_sens lda_spec lda_unk_accuracy lda_unk_sens lda_unk_spec
## 1        0.877   0.9461   0.3694           0.8863       0.9494       0.3942
```

The Logistic Regression model with unknowns achieved the highest overall performance, marking a notable development. The preservation of accuracy despite the presence of unknown values was unexpected. This phenomenon could be attributed to the fact that the unknown values constitute approximately 925 variables, with a significant portion originating from the education variable. It is plausible that numerous instances of unknown education contributed to accurate predictions.

Comparing sensitivity across the models yields the following results:

- Logistic Regression without unknowns: 0.9792
- Logistic Regression with unknowns: 0.9775
- LDA model with unknowns: 0.9494
- LDA model without unknowns: 0.9461

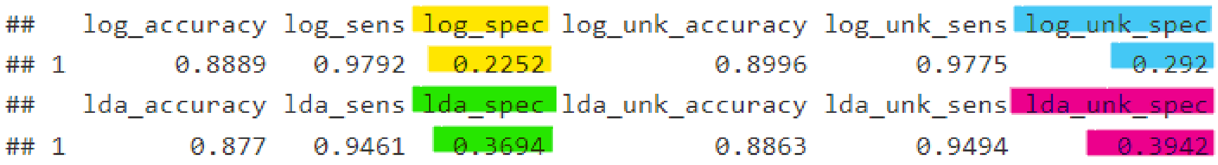


```
##   log_accuracy log_sens log_spec log_unk_accuracy log_unk_sens log_unk_spec
## 1       0.8889   0.9792   0.2252           0.8996       0.9775        0.292
##   lda_accuracy lda_sens lda_spec lda_unk_accuracy lda_unk_sens lda_unk_spec
## 1        0.877   0.9461   0.3694           0.8863       0.9494       0.3942
```

Remarkably, the Logistic Regression model without unknowns exhibited the highest sensitivity, registering at 0.9792. Despite being the second most accurate, its superior performance in correctly predicting term-deposit subscribers is surprising. The unknown variables might have facilitated accurate predictions of non-term deposit subscribers, which did not enhance sensitivity as anticipated. Although the sensitivity of the Logistic Regression models exceeded 0.97, a difference of 0.029 between these models and the LDA model in terms of sensitivity is noteworthy. Despite the LDA models' sensitivity hovering around 0.94, they cannot match the exceptional performance of the Logistic Regression models.

The specificity results are as follows:

- LDA model with unknowns: 0.3942
- LDA model without unknowns: 0.3694
- Logistic Regression with unknowns: 0.292
- Logistic Regression without unknowns: 0.2252



```
##   log_accuracy log_sens log_spec log_unk_accuracy log_unk_sens log_unk_spec
## 1       0.8889   0.9792   0.2252           0.8996       0.9775        0.292
##   lda_accuracy lda_sens lda_spec lda_unk_accuracy lda_unk_sens lda_unk_spec
## 1        0.877   0.9461   0.3694           0.8863       0.9494       0.3942
```

Surprisingly, the LDA models top the specificity rankings with 0.3942 and 0.3694, respectively, contrasting sharply with the 0.292 and 0.2252 scores of the Logistic Regression models. Given the significantly lower number of non-term deposits, it can be inferred that the LDA models were more adept at handling imbalanced data, especially considering the LDA model with unknowns' specificity being only 0.10 shy of 50%.

Best Predictors in Each Model.

Identifying the variables that significantly contributed to the models' accuracy is crucial. Examining both the Logistic Regression and LDA models without unknowns, since these served as the baseline models, reveals that variables marked with *** indicate statistical significance in the model, denoting them as the best predictors:

- ‘contacttelephone’: Clients contacted by telephone were among the best predictors.
- ‘monthmar’: Clients contacted in March were also top predictors.
- ‘poutcomesuccess’: Clients who had previously participated successfully in marketing campaigns were significant predictors, suggesting that targeting "success" in future campaigns could enhance model accuracy.
- ‘emp.var.rate’: The employment variation rate, as a macroeconomic condition, emerged as a strong predictor in the model. Further analysis is needed to ascertain the specific employment variation rate that yields the most accurate results, but this variable is crucial.
- ‘cons.price.idx’: The consumer price index is another vital macroeconomic variable. Depending on the CPI of a given month, inflation and term deposit rates might be higher, potentially making them more attractive to clients. This variable warrants further investigation.

VII. Conclusions and Recommendations

Upon analyzing the results derived from all four models, it is evident that:

- Each model exhibits exceptionally high accuracy, surpassing an 87% threshold.
- The sensitivity metric is similarly outstanding, exceeding 94% for every model.

It is apparent that neither accuracy nor sensitivity poses significant challenges. Instead, the critical concern lies with specificity, particularly the models' capacity to accurately predict long-term subscribers. Given the data's limited occurrences of affirmative responses, the specificity of these models raises concerns. Consequently, the LDA model, excluding unknowns, emerges as the superior choice. Although it ranks lower in both accuracy and sensitivity, its performance in terms of specificity makes it the most suitable option.

Regarding recommendations, our analysis suggests that, based on the logistic regression models, the mode of contact, specifically telephone communication, is a robust predictor for the target variable. The negative coefficient associated with telephone contact implies that it does not enhance subscription rates, advising a strategic shift towards cellular contact for client communication.

Furthermore, the month of March is identified as significantly influential, with a positive coefficient indicating a higher subscription rate during this period. However, the impact of other variables, such as the consumer price index ('cons.price.idx'), and the Euribor 3-month rate ('euribor3m'), cannot be overlooked. The significant positive correlation of 'outcome success' with the model underscores the importance of focusing on customers who have benefitted from previous campaigns. Given the significance of 'employment variation rate' ('emp.var.rate'), 'cons.price.idx', and 'consumer confidence index' ('cons.conf.idx'), it is advisable for the company to strategically leverage these indicators. Specifically, targeting customers when the 'emp.var.rate' is negative and both 'cons.price.idx' and 'cons.conf.idx' are positive could potentially enhance subscription rates.

This comprehensive analysis not only identifies the strengths and weaknesses of the existing models but also offers actionable insights for optimizing marketing strategies and improving customer engagement outcomes.