IS-507 - Data, Statistical Models, and Information - Fall 2024

Final Project Report
on
Statistical Modeling and Analysis of Bank's Telemarketing Campaigns

Group Members:
Hardik Lad (hlad2)
Praveena Acharya (ppa3)
Omkar Chalke(ochalke2)
Rujul Khatavkar(rujulsk2)
Radha Jawale(rjawale2)

1. OVERVIEW

The project uses a dataset from a Portuguese bank's telemarketing campaign aimed at selling term deposits. Term deposits are key fixed-interest products for the bank, and optimizing outreach efforts is crucial for maximizing client conversion. This project is used to predict which customers are most likely to subscribe, offering valuable insights for targeted marketing and potentially reducing telemarketing costs in future campaigns.

2. DATA DESCRIPTION

This project utilizes a publicly available dataset - https://archive.ics.uci.edu/dataset/222/bank+marketing. It provides detailed customer demographic, financial, and engagement information, making it a valuable resource for understanding customer behavior in direct marketing. Developed by Paulo Cortez (Univ. Minho) and Sérgio Moro (ISCTE-IUL) in 2012, this dataset was sourced from the UCI Machine Learning Repository *[S. Moro, P. Cortez, and P. Rita, A Data-Driven Approach to Predict the Success of Bank Telemarketing, Decision Support Systems, Elsevier, 62:22-31, June 2014]*.

3. DATA PREPROCESSING

The data preprocessing phase prepared the banking dataset for analyzing term deposit subscriptions. Test and training datasets were combined, and quality checks addressed empty strings in categorical variables by converting them to NA. Numerical variables like age and account balance were standardized using z-scores to ensure comparability, while categorical variables were converted to factors for proper modeling. The target variable, client_subscribed, was binarized (0/1) for classification. Finally, the dataset was split into training (70%) and testing (30%) sets using stratified sampling with a fixed seed for balanced representation. These steps ensured a solid foundation for reliable analysis.

4. LITERATURE REVIEW

1. *Predicting the Accuracy for Telemarketing Process in Banks Using Data Mining - Fawaz J. Alsolami1, Farrukh Saleem2 and Abdullah AL-Malaise AL-Ghamdi2* - found logistic regression achieved the highest accuracy (91.48%) in predicting telemarketing success, emphasizing preprocessing and a CRISP-DM-based approach.
2. *A data modeling approach for classification problems: Application to bank telemarketing prediction - Koumetio Tekouabou et al. (2019)* introduced a novel preprocessing method, enhancing multiple models, with neural networks excelling in some cases.
3. *Prediction of Term Deposit in Bank: Using Logistic Model (Enjing Jiang1, Zihao Wang2, Jiaying Zhao3)* - showed logistic regression outperformed decision trees in accuracy and robustness for term deposit prediction.

Our study incorporates advanced techniques like Random Forest and K-means clustering, enabling deeper insights through customer segmentation, feature analysis, and actionable strategies for modern marketing needs.

5. RESEARCH PROBLEM I - FACTORS SIGNIFICANTLY AFFECTING A CUSTOMER'S LIKELIHOOD OF SUBSCRIBING TO A TERM DEPOSIT

5.1. PROBLEM DESCRIPTION AND OBJECTIVE

This study analyzes key factors influencing customer subscription to term deposits through telemarketing campaigns using logistic regression. It examines demographics (age, job, education), financial indicators (account balance, loans), and campaign metrics (contact duration, previous attempts, past outcomes) to identify significant predictors.

5.2. MODEL FITTING

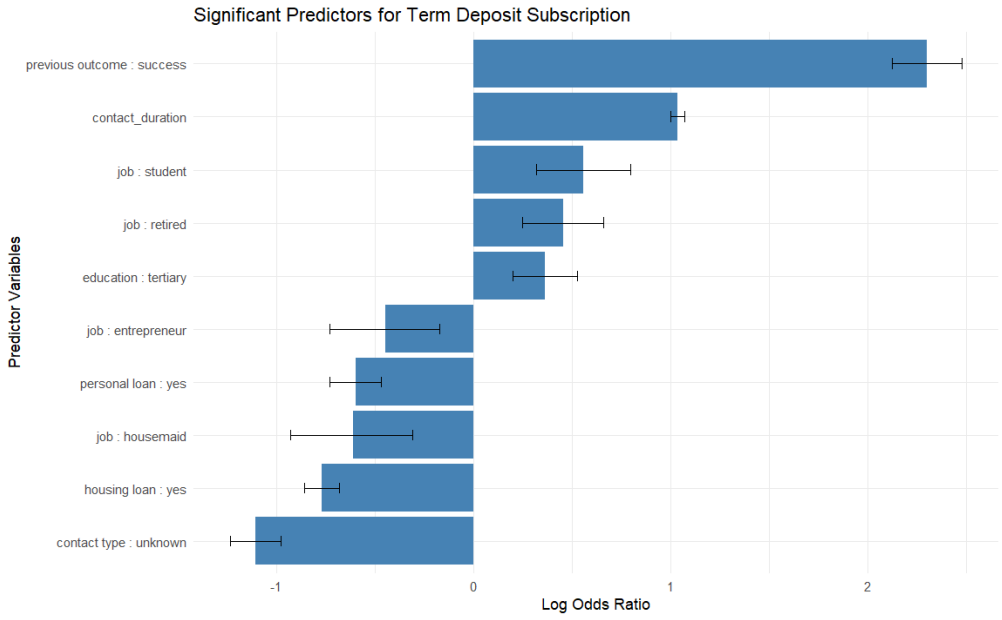
The glm() function was used to fit a logistic regression model predicting whether a client subscribed to a term deposit ("yes") or not ("no"). The dependent variable represents subscription status, while predictors include demographics (age, job, marital status, education), financial indicators (balance), and other attributes. Logistic regression, with family = binomial(), models binary outcomes and estimates how predictors influence subscription probability, offering interpretable and rigorous insights into key factors.

5.3. MODEL EVALUATION

The logistic regression model identifies key factors influencing term deposit subscriptions. Positive predictors, such as contact duration and past success, significantly increase subscription likelihood, while being a student or retired also raises the probability. Education, financial factors, and bank balance play important roles. The model shows strong performance, with deviance reduced from 25,201 to 17,534 and an AIC of 17,596, indicating a good fit.

ODDS RATIOS

The odds ratio analysis reveals key factors influencing term deposit subscriptions. Positive predictors include previous campaign success (OR = 9.96), longer contact duration (OR = 2.81), being a student (OR = 1.75) or retired (OR = 1.58), and having tertiary education (OR = 1.44). Negative predictors, such as housing loans (OR = 0.46), personal loans (OR = 0.55), and occupations like housemaids or entrepreneurs (OR = 0.54, 0.64), reduce subscription likelihood. Visuals show odds ratios and confidence intervals, highlighting significant predictors. Strategies should focus on targeting customers with successful past interactions, students, retirees, higher education, and those without loans while emphasizing longer contact durations and clear communication.

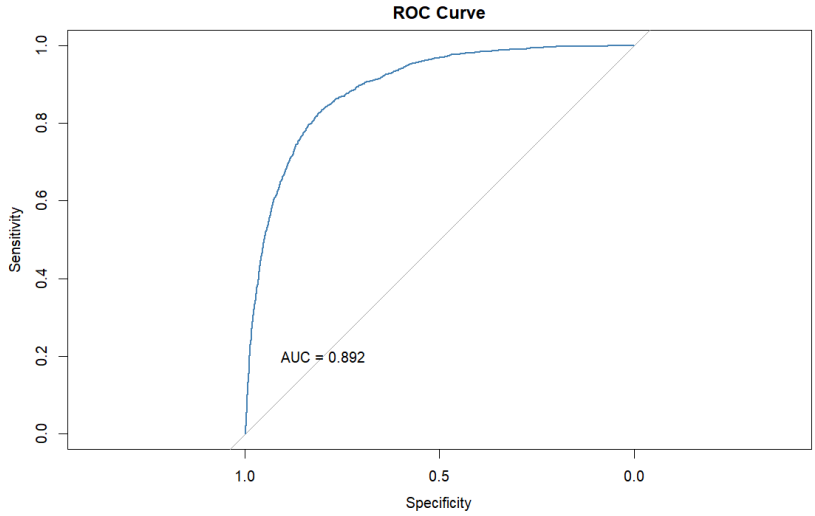


VIF (Variance Inflation Factor)

The VIF analysis shows acceptable multicollinearity levels, with all values below 5. Variables like job (VIF: 3.95), poutcome (VIF: 4.10), and pdays (VIF: 3.69) have moderate correlations, while default (VIF: 1.01), balance (VIF: 1.03), loan (VIF: 1.03), and campaign (VIF: 1.03) show low correlations. These low VIF values indicate model stability and reliability, with no need to exclude predictors due to multicollinearity, ensuring stable, interpretable predictions.

ROC AND AUC

The ROC curve with an AUC of 0.892 indicates strong predictive performance, effectively distinguishing subscribers from non-subscribers. The curve's shape highlights excellent class discrimination, ensuring accurate targeting and enhancing campaign effectiveness.



CONFUSION MATRIX

The model achieved an accuracy of 90.12%, though this is influenced by class imbalance, with 88.48% of cases being non-subscribers. Sensitivity was high at 97.77% for identifying non-subscribers, but specificity was low at 31.41% for detecting subscribers. Positive predictive value (91.63%) was strong for non-subscribers, while the negative predictive value (64.67%) was weaker for subscribers. This imbalance biases the model toward the majority class, as reflected in the moderate Kappa statistic of 0.3758. McNemar's test further highlighted significant differences in false positive and false negative rates ($p < 2.2e\text{-}16$). Addressing class imbalance is essential to improve predictions for the minority class.

6. RESEARCH PROBLEM II - IS THERE A STATISTICALLY SIGNIFICANT DIFFERENCE IN SUBSCRIPTION RATES ACROSS DEMOGRAPHIC GROUPS?

6.1 PROBLEM DESCRIPTION AND OBJECTIVE

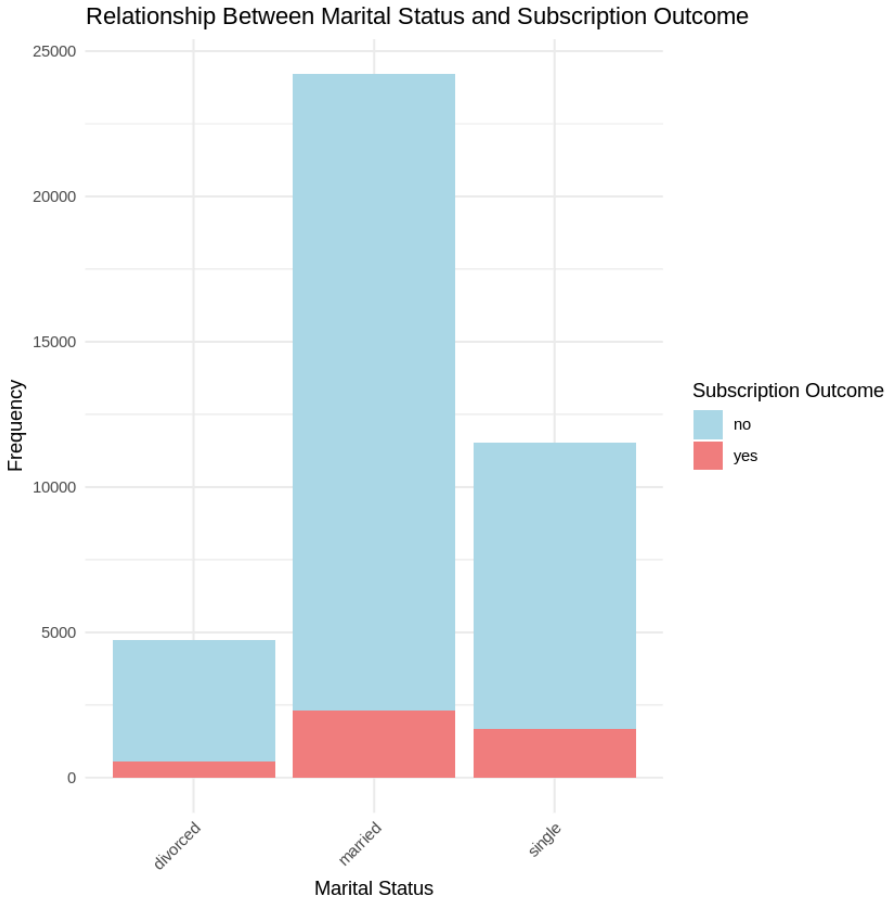
This study investigates whether there are statistically significant differences in subscription rates for a term deposit product across demographic groups. It focuses on three key demographic variables—marital status, education level, and job type—and examines their relationship with subscription outcomes. Using a Chi-Square Test for Independence, the objective is to identify any significant dependencies between these demographic factors and the likelihood of subscription.

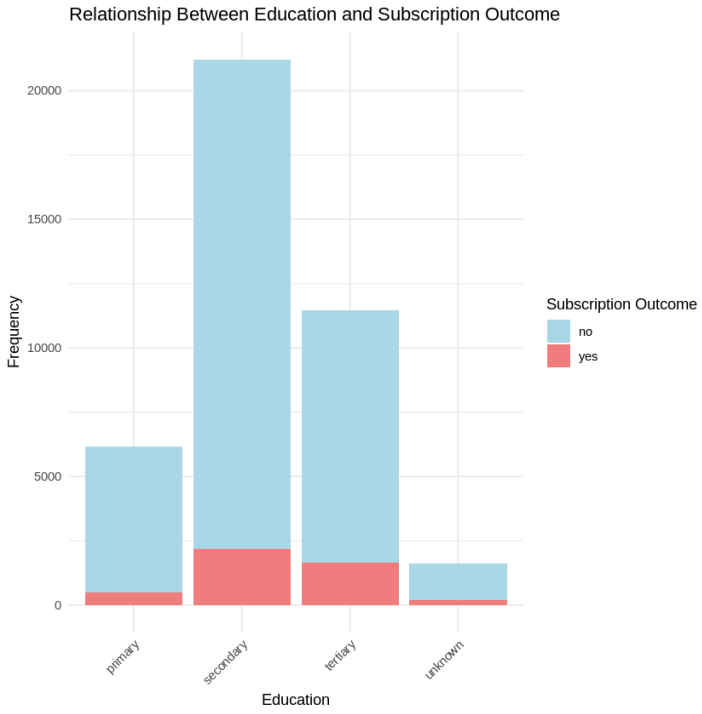
6.2 METHODOLOGY

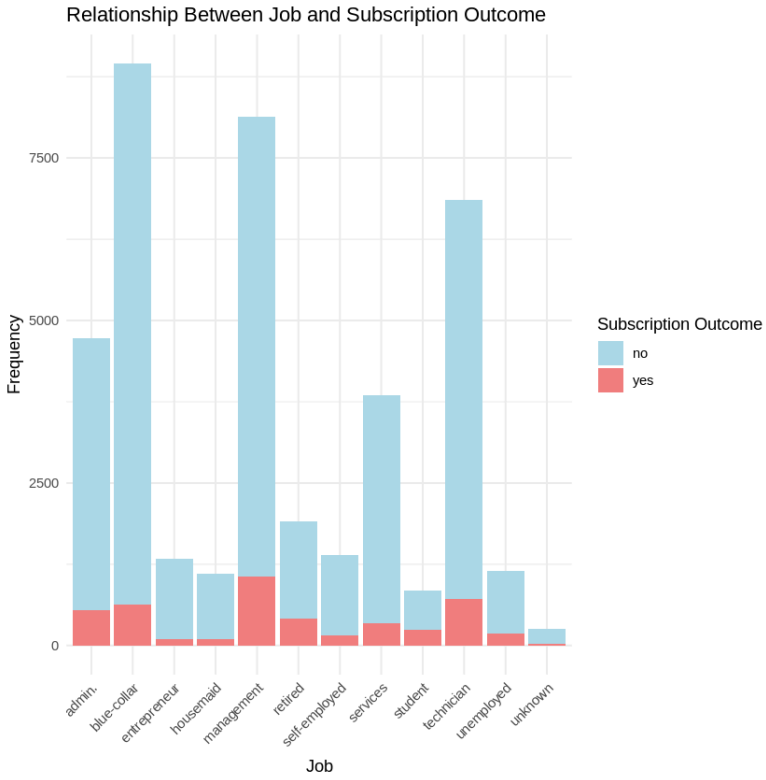
We employed the Chi-Square Test for Independence to analyze the relationship between demographic variables and subscription rates. This statistical method helps determine whether there is a significant association between categorical variables. The null hypothesis assumes independence between the variables, while the alternative hypothesis suggests a significant relationship.

6.3 RESULT AND ANALYSIS

The Chi-Square tests revealed statistically significant relationships between marital status, education level, job type, and subscription outcomes. For marital status ($X^2 = 197.41$, $p < 2.2e\text{-}16$), single individuals had higher subscription rates than married or divorced groups. Education level (X^2 = significant, $p < 0.001$) showed that tertiary-educated individuals were most likely to subscribe, while those with primary education had the lowest rates. Job type (X^2 = significant, $p < 0.001$) indicated that retired individuals and students were the most likely to subscribe, whereas blue-collar workers and technicians showed lower likelihoods. These findings highlight the importance of demographic factors in subscription behavior.







6.4 DISCUSSION

The analysis reveals significant demographic differences in subscription rates with important marketing implications. Singles exhibit higher subscription rates, requiring targeted campaigns, while married and

divorced individuals may need tailored approaches. Tertiary-educated individuals are more receptive, whereas simpler offers could engage those with primary education. Retired individuals and students show high subscription likelihood, suggesting focused campaigns, while messaging for blue-collar workers and technicians should emphasize long-term benefits. These findings highlight the critical role of demographics in predicting subscription likelihood, enabling financial institutions to design more effective, targeted strategies. Leveraging these insights can improve subscription rates and enhance customer engagement.

7. RESEARCH PROBLEM III - HOW DO PREVIOUS CAMPAIGN RESULTS IMPACT CURRENT SUBSCRIPTION RATES?

7.1 PROBLEM DESCRIPTION AND OBJECTIVE

This research examines how previous campaign results impact current subscription rates, crucial for optimizing marketing strategies. Using survival analysis, we aim to quantify the relationship between past interactions and subscription likelihood, providing insights to enhance success rates. Survival analysis is ideal for handling time-to-event data, censored observations, and assessing covariates' impact on subscription probability over time, offering intuitive outputs like survival probabilities and hazard ratios.

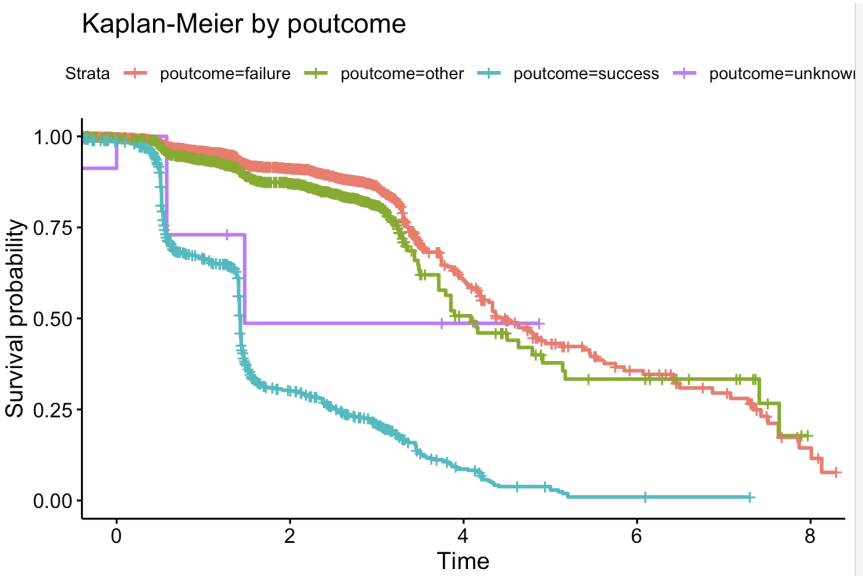
7.2 METHODOLOGY

The study analyzed subscription status using Kaplan-Meier analysis to visualize survival probabilities across poutcome categories and the Cox Proportional Hazards Model to assess the impact of covariates (pdays, previous, poutcome) on subscription likelihood. Key metrics included survival probability (non-subscription likelihood over time) and hazard ratios (relative subscription likelihood for different covariate levels).

7.3 RESULTS

Kaplan-Meier Analysis

The Kaplan-Meier analysis revealed significant variations in survival probabilities across different poutcome categories:

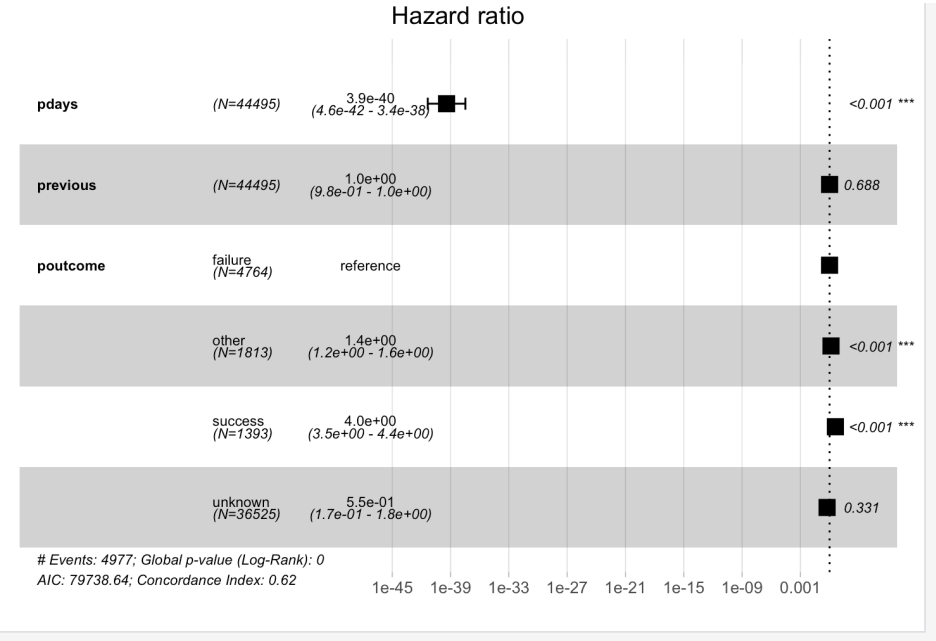


1. The 'success' category showed the steepest decline in survival probability, indicating the highest subscription rates.
2. 'Failure' and 'other' categories demonstrated moderate declines over time.
3. The 'unknown' category had the flattest curve, suggesting the lowest likelihood of subscription.

Cox Proportional Hazards Model

The Cox Proportional Hazards Model provided quantitative insights into the impact of various factors on subscription likelihood:

1. poutcome = success: Hazard ratio (HR) = 3.961 (highly significant). Clients with prior successful campaigns were nearly four times more likely to subscribe compared to the reference group (poutcome = failure).
2. poutcome = other: HR = 1.410, showing a moderately positive impact.
3. poutcome = failure: Insignificant impact, indicating prior failures have minimal influence on subscription likelihood.
4. pdays: HR = 3.937e-40, indicating that longer gaps between contacts significantly reduce the likelihood of subscription.



7.4 OBSERVATIONS

Clients with prior successful campaigns (poutcome = success) are significantly more likely to subscribe, emphasizing the importance of leveraging past successes. Moderate positive outcomes (poutcome = other) also contribute to future subscriptions, while prior failures have minimal impact, allowing opportunities for re-engagement. Timing is critical, as longer gaps between campaigns drastically reduce engagement likelihood, underscoring the need for consistent, timely follow-ups to maintain client interest. These findings highlight the value of prioritizing clients with successful histories and optimizing campaign timing to enhance subscription rates. Future research could explore demographic and campaign-specific factors or the long-term

impact of repeated successes and failures, further refining marketing strategies to boost engagement and effectiveness.

8. RESEARCH PROBLEM IV - WHAT IS THE OPTIMAL CONTACT STRATEGY IN TERMS OF FREQUENCY AND TIMING?

8.1 PROBLEM DESCRIPTION AND OVERVIEW

This study aims to identify how contact frequency and timing influence client conversion rates to optimize campaign effectiveness. Using a Random Forest model, patterns in customer interactions were analyzed to uncover key drivers of successful subscriptions.

8.2 METHODOLOGY

Key features were engineered, including a binary variable for prior contacts and a categorical variable for campaign intensity (Low, Medium, High, Very High). The dataset was split into training (70%) and testing (30%) subsets. The Random Forest model captured non-linear relationships between predictors like age, bank balance, campaign frequency, outcomes, contact month, type, and duration. Model performance was assessed with a confusion matrix and ROC curve, with the AUC summarizing sensitivity and specificity.

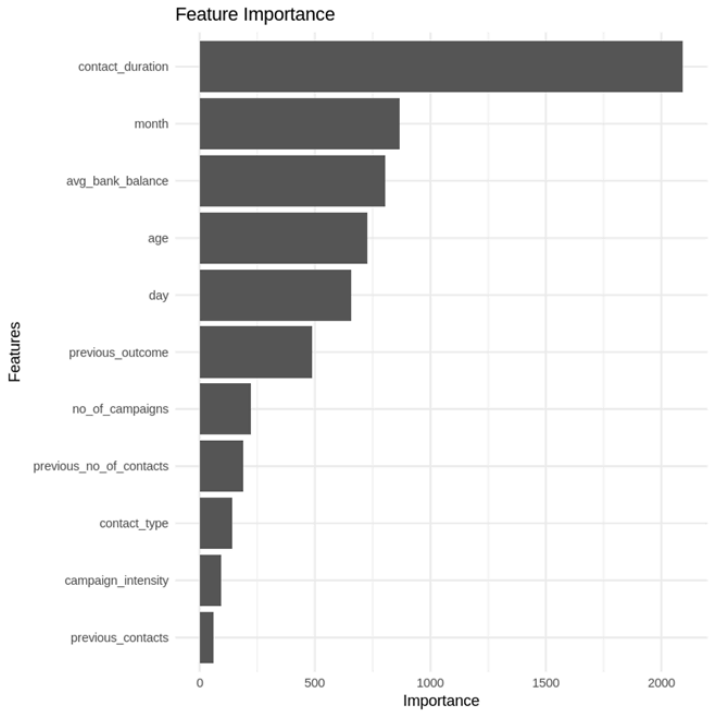
8.3 RESULTS

Model Performance Metrics

The model achieved 91% accuracy, with precision and recall rates of 0.87 and 0.83, respectively, and an F1-Score of 0.85. An AUC of 0.94 indicated strong predictive ability.

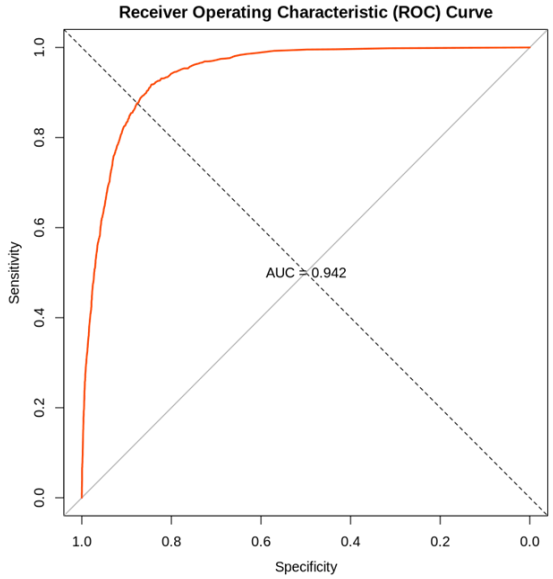
Visualization Explanation

To enhance understanding of the results, several visualizations were created:

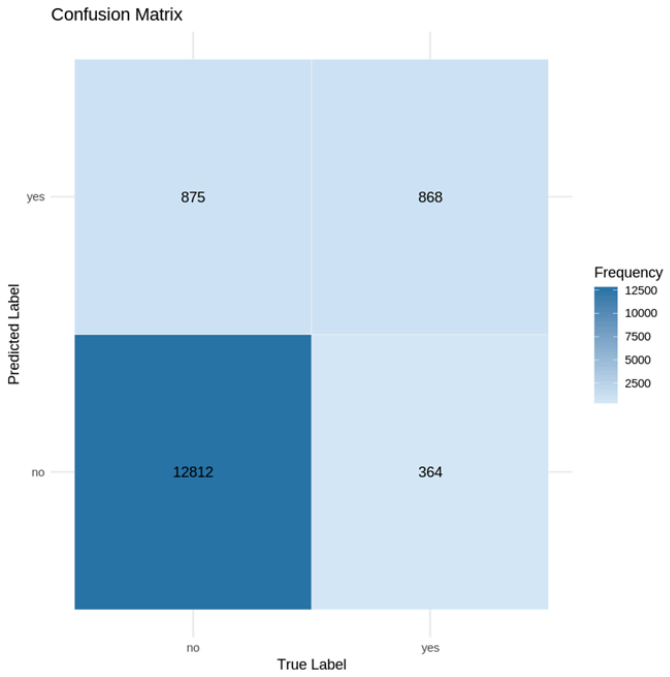


Feature Importance Plot:Highlighted contact duration, bank balance, and age as critical predictors.This visualization aids in identifying which aspects of the contact strategy are most influential.

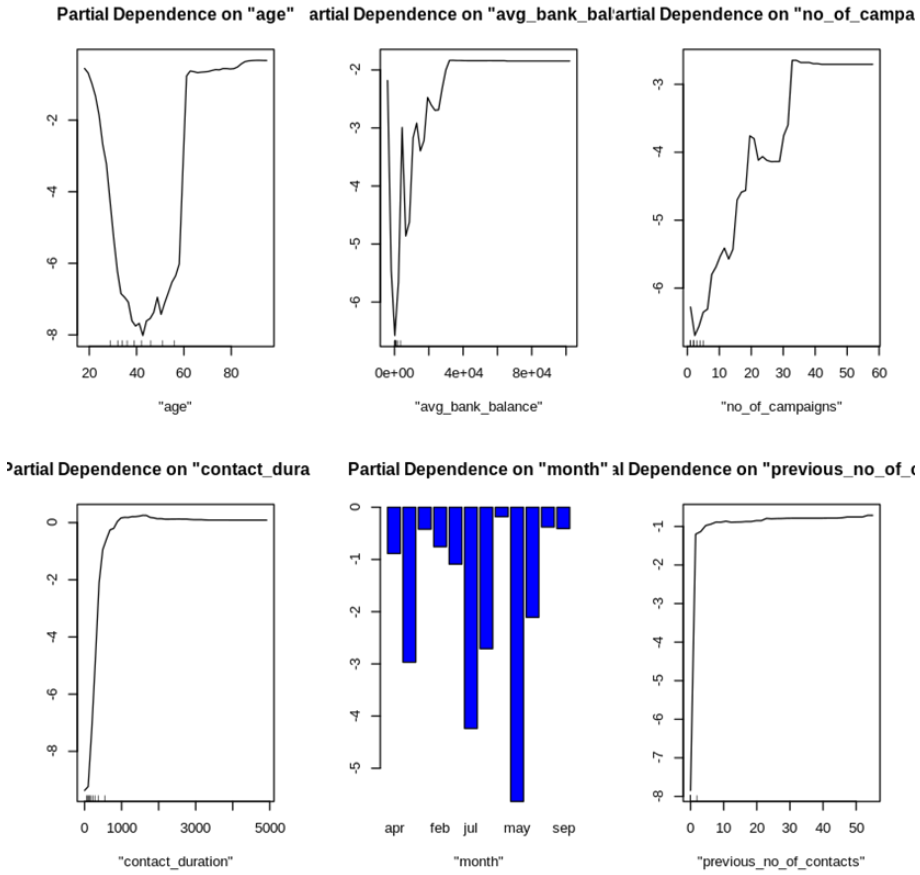
ROC Curve: The ROC curve illustrates how well the model distinguishes between subscribers and non-subscribers across various thresholds. A curve that approaches the top-left corner signifies better performance, while an AUC close to 1 indicates strong predictive ability.



Confusion Matrix: Displayed prediction accuracy with true positives/negatives along the diagonal.



Partial Dependence Plots: Demonstrated relationships like higher subscription likelihood for middle-aged clients and longer call durations.



Longer call durations and strategic timing during high-success months (e.g., March or August) significantly enhance subscription rates. Tailoring strategies based on age, bank balance, and campaign intensity further improves efficiency. By leveraging these data-driven insights, businesses can refine strategies, boost subscriptions, and optimize resource allocation, ultimately strengthening client relationships and increasing profitability in subscription-based services.

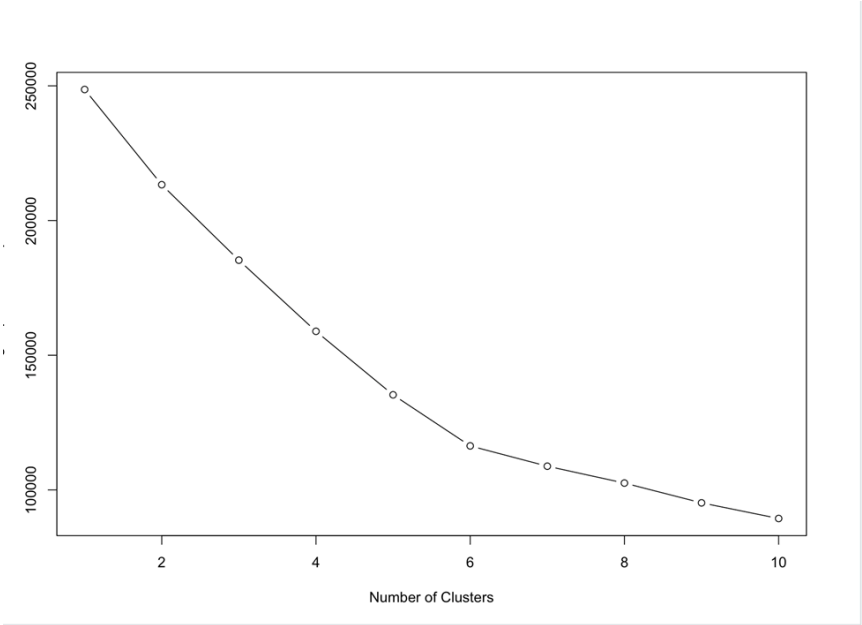
9. RESEARCH PROBLEM V - CUSTOMER SEGMENTATION BASED ON SUBSCRIPTION LIKELIHOOD

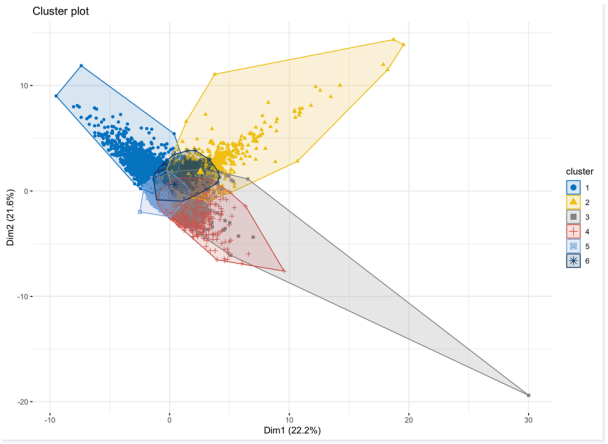
9.1 PROBLEM DESCRIPTION AND OBJECTIVE

Customer segmentation is vital for targeting and retaining customers effectively. In banking, identifying groups by subscription likelihood enhances marketing efficiency. This study uses K-means clustering, an unsupervised machine learning technique, to segment customers based on their likelihood to subscribe.

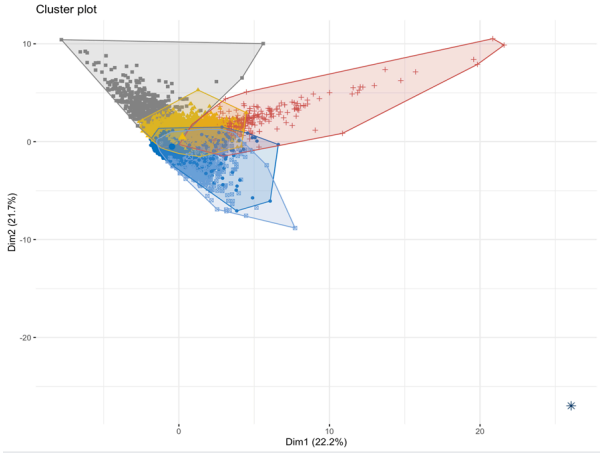
9.2 METHODOLOGY

The elbow method identified six clusters as optimal, balancing complexity and data fit. K-means, outperforming hierarchical clustering in silhouette scores and the Calinski-Harabasz index, was chosen for its scalability with large datasets.





K-means Cluster Plot



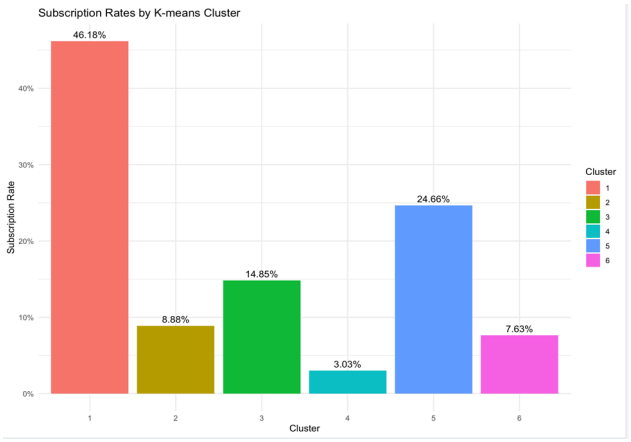
Hierarchical Cluster Plot

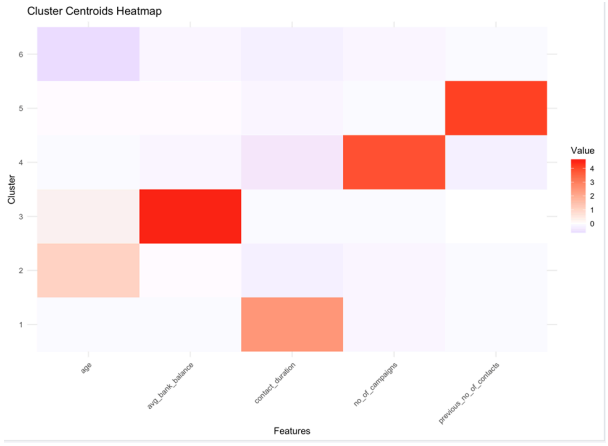
9.3 RESULTS

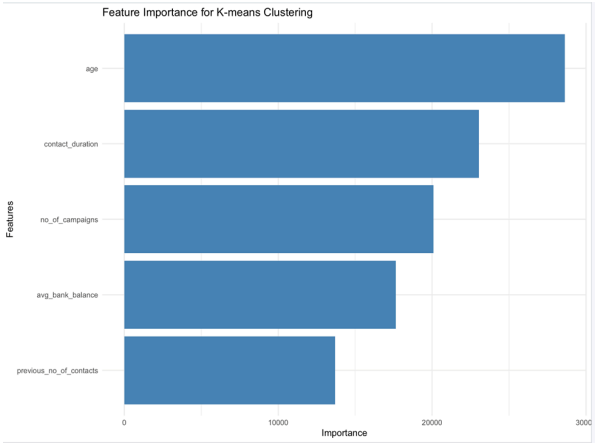
Customer segmentation identified six groups based on subscription likelihood:

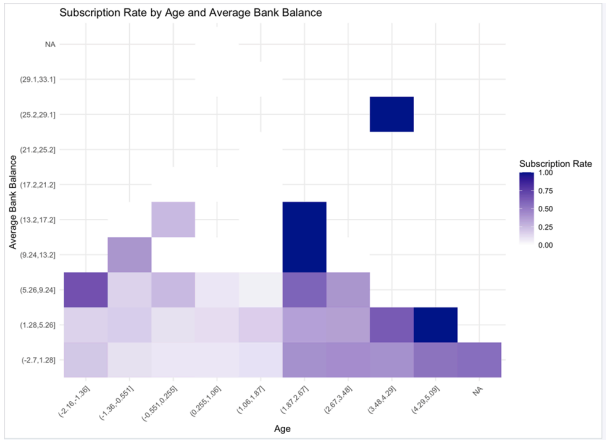
1. High Potential Subscribers (46.18%): High bank balances and longer contact durations.
2. Engaged Customers (24.66%): Strong engagement through frequent campaigns and prior contacts.
3. Moderate Prospects (14.85%): Balanced metrics for contact duration and bank balance.
4. Lower Potential Group (8.88%): Lower bank balances and moderate contact durations.
5. Low Engagement Segment (7.63%): Few contacts and minimal engagement.
6. Minimal Potential Group (3.03%): Lowest subscription rates despite varied attempts.

Age, contact duration, and campaign frequency were the most influential factors, emphasizing the need for strategies tailored to demographics, optimized contact duration, and appropriate campaign intensity.









9.4 CONCLUSION

The clustering analysis revealed distinct customer segments, offering actionable insights for targeted marketing. High subscription rates in Cluster 1 highlight the importance of focusing on customers with higher bank balances and engaging them in longer conversations. In contrast, low subscription rates in Cluster 4 suggest the need for alternative approaches or deprioritizing this segment. Age, a critical factor, underscores the need for tailored strategies across different demographics, while the role of contact duration emphasizes the value of quality interactions in driving subscriptions. This study validates the effectiveness of K-means clustering in segmenting customers based on subscription likelihood, enabling targeted strategies and efficient resource allocation. Future research could incorporate additional data points and advanced machine learning techniques to refine segmentation and better predict customer behavior.

10. SUMMARY AND RECOMMENDATIONS

The analysis identified key factors influencing term deposit subscriptions, including prior campaign success, contact duration, and demographics like age, education, and job type. Singles, retirees, students, and tertiary-educated individuals showed higher subscription rates, while blue-collar workers and those with primary education required tailored strategies. Timing was critical, with shorter gaps between campaigns and strategic targeting during high-success months boosting engagement. K-means clustering revealed actionable

customer segments, with high-potential groups prioritized for marketing. Despite the model's 90.12% accuracy, class imbalance limited specificity, underscoring the need for improvement.

Recommendations:

Focus marketing on high-potential segments using demographic insights.

1. Leverage successful campaign histories and ensure timely follow-ups.
2. Enhance engagement with targeted strategies based on age and education.
3. Address class imbalance using techniques like SMOTE or weighted models to improve prediction for minority classes.

11. FUTURE WORKS AND IMPROVEMENT

To improve model accuracy and balance, addressing class imbalance (88.48% non-subscribers) is crucial. Techniques like SMOTE, undersampling, and adjusting class weights can enhance sensitivity for minority classes. Exploring advanced algorithms like Random Forest or XGBoost and employing cross-validation, hyperparameter tuning, and regularization (Ridge/Lasso) can optimize performance. Incorporating additional variables such as income, spending habits, and geographic data, along with psychographic insights from surveys, can further refine marketing strategies. Employing customer segmentation and establishing a feedback loop for continuous model improvement will enhance predictions and provide a deeper understanding of subscription behavior for the financial sector.

12. TEAM CONTRIBUTION

The project showcased strong team collaboration, with tasks efficiently divided among all members. Preprocessing and modeling were tackled collaboratively, ensuring a unified approach and shared understanding of the data. Each of the five research questions was assigned to a team member, allowing for focused analysis and leveraging individual strengths. The final integration of results and report writing was conducted collectively, ensuring consistency and coherence in the overall presentation. This collaborative approach not only streamlined the workflow but also fostered effective communication and teamwork throughout the project.