Analysis of Banking Data: Insights from Logistic Regression, Cluster Analysis, and PCA

Kazi Nafis, Derek Siriboe, Calvin Bailey

Data Analytics Program, Denison University

DA350: Advanced Methods for Data Analytics

Dr. Zhe Wang

December 17, 2023

Abstract

This report presents a comprehensive analysis of banking data to understand customer behaviors and preferences regarding term deposit subscriptions. Utilizing logistic regression, we identified significant predictors impacting customer decisions, with a focus on job types, marital status, and financial variables. Further, we employed K-means clustering to segment customers, revealing distinct profiles based on financial status and engagement. Principal Component Analysis (PCA) was also conducted, simplifying complex multivariate data, and highlighting key influencing factors. The results offer valuable insights for tailoring marketing strategies, potentially enhancing the bank's campaign efficiency and customer conversion rates.

Introduction

This project aims to analyze banking data to uncover patterns and factors influencing customers' decisions to subscribe to term deposits. We approach this by leveraging logistic regression to identify significant predictors, followed by K-means clustering to segment customers into meaningful groups. Additionally, Principal Component Analysis (PCA) is employed to further distill and simplify the data, making it more interpretable. The data, sourced from the bank's records, includes variables such as age, job type, marital status, education, and financial details. The methods used are an expansion of the initial summary, providing a detailed examination of customer behavior and preferences in the context of banking services.

Ethical Consideration

The analysis of the Portuguese banking institution's marketing data, sourced from the UCI Machine Learning Repository, entails several ethical considerations. First and foremost, data ownership and privacy concerns are paramount. Since the data involves personal information about clients, such as age, job type, marital status, and financial status, it is crucial to ensure that it has been anonymized and de-identified before its release for public usage. This safeguard protects individual privacy and complies with data protection regulations like the General Data Protection Regulation (GDPR).

Another aspect is the ethical use of the data. The data was initially gathered for direct marketing purposes, which raises questions about consent and the extent to which clients were informed about the use of their data. For our analysis, it is assumed that the bank obtained explicit consent from its clients for the use of their data in such campaigns.

Regarding the stakeholders, the primary ones include the bank's clients, whose data is being analyzed, and the bank itself, which conducted the marketing campaigns. The clients have a stake in how their data is used and what conclusions are drawn from it, particularly if these insights are used to influence future marketing strategies. On the other hand, the bank benefits from understanding customer behaviors and preferences to optimize its marketing strategies, which could impact its profitability and customer satisfaction levels.

The potential impact of the analysis on these stakeholders must be considered. If the analysis results in stereotyping or unfair assumptions about certain demographic groups, it could lead to biased marketing strategies. Such outcomes could harm the bank's reputation and lead to mistrust among its clients. Therefore, it is vital to approach the analysis with an objective, unbiased perspective and ensure that any insights or recommendations are based on a fair and balanced interpretation of the data.

In conclusion, while the data offers valuable insights into customer behavior and preferences, it is crucial to handle it with ethical diligence, respecting privacy and considering the potential implications for all stakeholders involved.

Interpretative Analysis of the Visual

Figure 1: Log- Transformed Distribution of Customer Balance

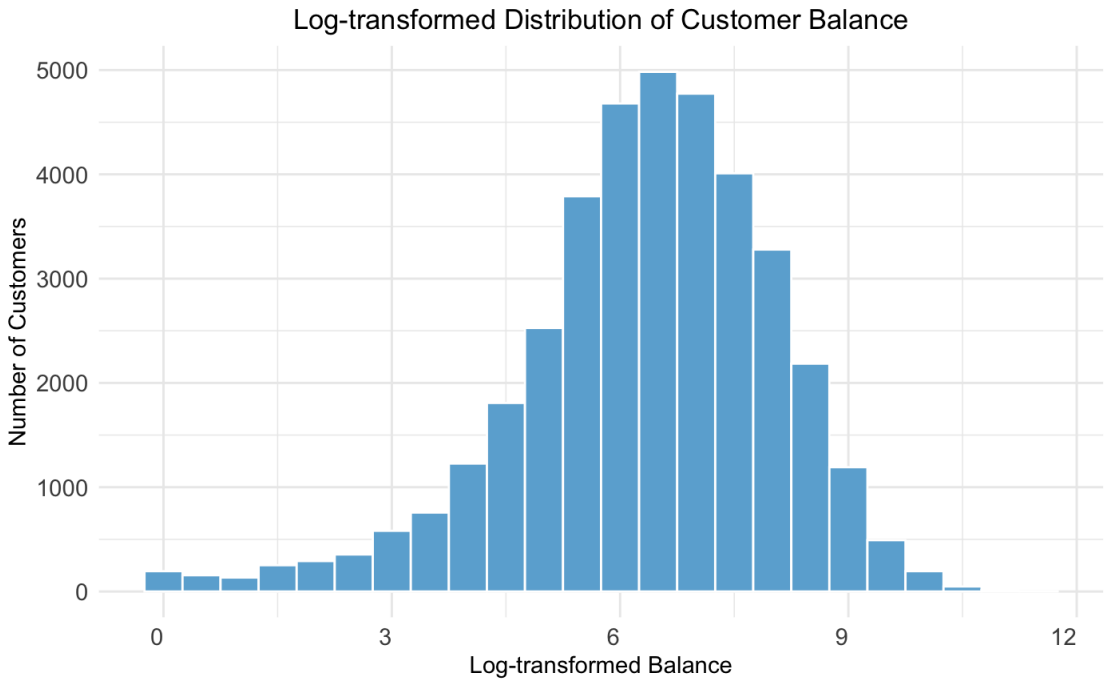


Figure 1, after applying a logarithmic transformation to customer balance data, presents a distribution that closely resembles a bell curve. This transformation is beneficial for reducing skewness, which is a common characteristic of financial data sets where a minority of customers may hold disproportionately high balances. As a result, the majority of customers' balances cluster around the modal range, shedding light on the standard financial status of a typical customer within the bank.

Notably, the right tail of Figure 1—where the bars represent a dwindling number of customers—points to the presence of high-value outliers. These account holders possess balances substantially larger than the norm and, while not numerous, they represent a significant facet of the bank's clientele in terms of financial assets. The effectiveness of the logarithmic scaling is underscored here, allowing for these extreme values to be included in the analysis without overshadowing the more common, lower balance accounts.

Data Preparation Synopsis

In preparation for creating this histogram, a series of data wrangling steps were meticulously executed. Initially, any missing values within the 'balance' column were addressed to ensure a robust analysis.

Subsequently, to counteract the inherent skewness of the balance data, a logarithmic transformation was applied. The specific function used, log1p, adeptly manages zero balances, converting them to zeros in the transformed scale, thereby retaining these data points for analysis without distortion.

Moreover, any accounts with non-positive balances, which are undefined in the logarithmic space, were excluded to prevent computational errors.

By transforming the balances logarithmically, the data values were effectively rescaled, rendering them more comparable and suitable for subsequent modeling exercises.

Figure 2: Scatter Plot of Last Contact Duration vs Customer Balance

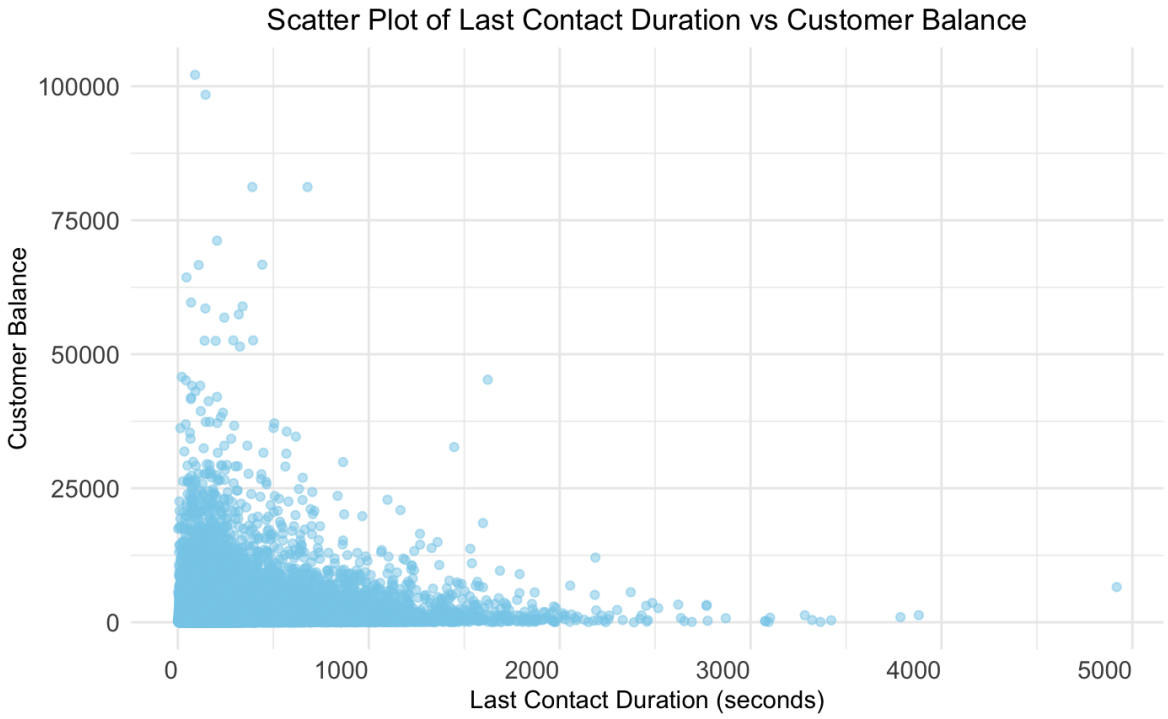
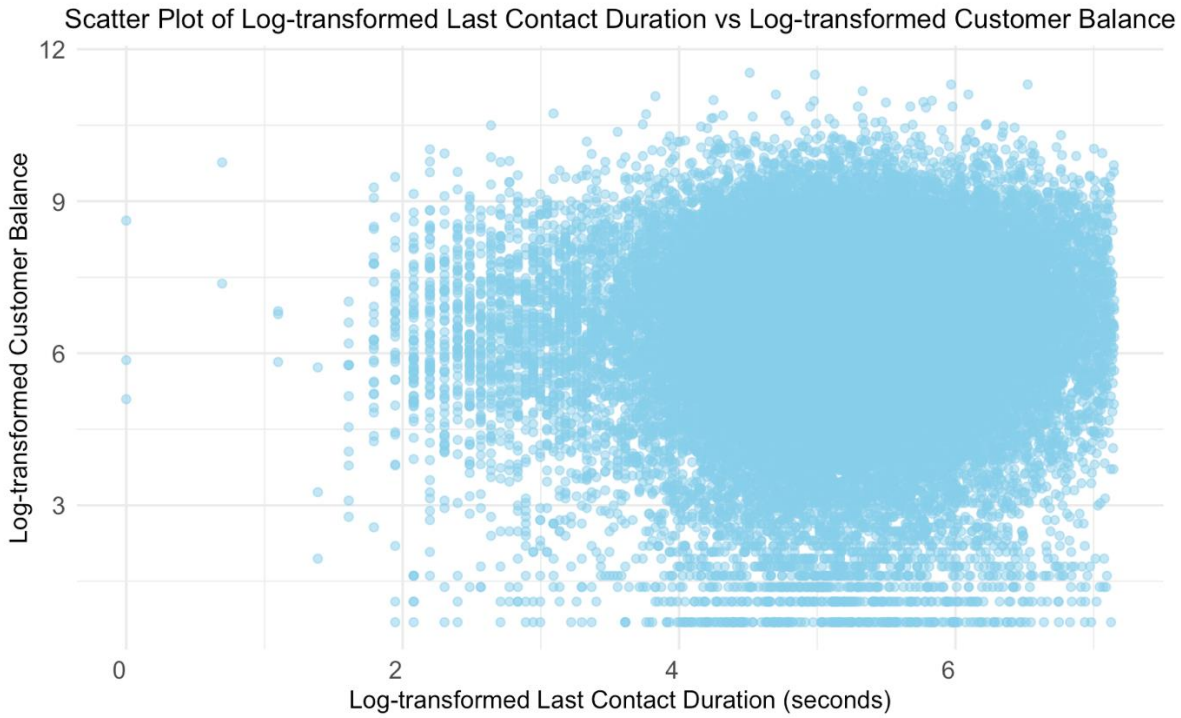


Figure 3: Scatter Plot of Log-Transformed Last Contact Duration vs Log-Transformed Customer Balance



Visual Interpretation

Figure 2 of log-transformed last contact duration versus log-transformed customer balance provides several insights. The data points are densely packed near the lower end of both axes, suggesting that a majority of customers have relatively modest balances and engage in brief communications with the bank. Notably, there are distinct horizontal lines, which may indicate common balance values or data recording practices. Outliers are apparent, particularly in the balance dimension, indicating a subset of customers with significantly higher balances. These could represent high-net-worth individuals or anomalies in the data. No clear linear relationship is discernible from the plot, suggesting that the length of the last contact is not directly predictive of the customer's balance, or vice versa.

Data Wrangling Summary

Prior to analysis, the dataset underwent several preprocessing steps to prepare it for modeling. Missing values were identified and handled, either through deletion or imputation, to ensure that subsequent analyses were based on complete cases. Categorical variables within the dataset were encoded to numerical forms, to facilitate their use in predictive modeling. Feature scaling was applied, notably a log transformation to the 'duration' and 'balance' variables. This transformation was particularly important to mitigate the right-skewed distribution of these variables. By doing so, the influence of extreme values was reduced, and the data was normalized to some extent, making it suitable for algorithms like K-means clustering and logistic regression that assume or benefit from normally distributed features.

Analysis

We started with a logistic regression model using data like age, job type, marital status, education, and others to see if we could predict who would sign up for a bank's term deposit. The model showed us that different jobs have different effects on people's decisions. For example, retirees were more likely to sign up than blue-collar workers.

GLM Model Coefficients

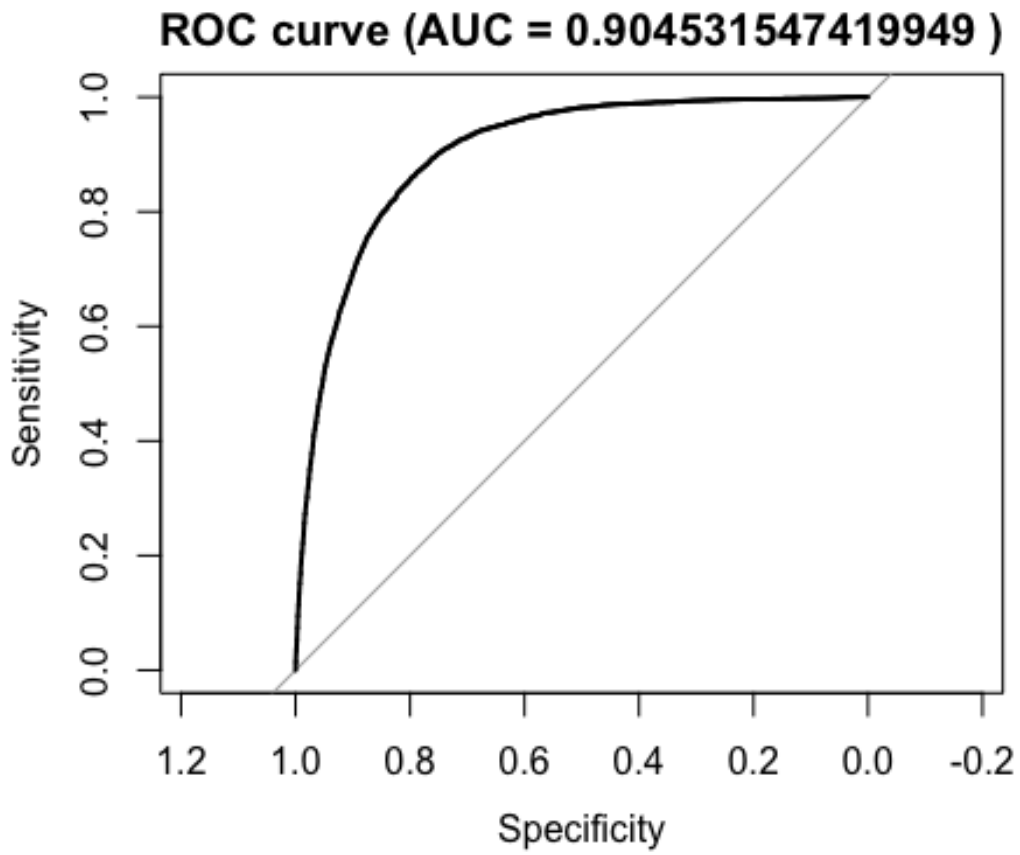
	Estimate	Std..Error	z.value	Pr…z..
(Intercept)	-2.4626	0.1741	-14.1441	0.0000
age	0.0016	0.0023	0.7026	0.4823
jobblue-collar	-0.4001	0.0759	-5.2749	0.0000
jobentrepreneur	-0.5757	0.1362	-4.2280	0.0000
jobhousemaid	-0.5012	0.1392	-3.6012	0.0003
jobmanagement	-0.1628	0.0754	-2.1592	0.0308
jobretired	0.3343	0.0985	3.3933	0.0007
jobself-employed	-0.3729	0.1145	-3.2582	0.0011
jobservices	-0.3372	0.0885	-3.8078	0.0001
jobstudent	0.4516	0.1094	4.1266	0.0000
jobtechnician	-0.2448	0.0709	-3.4545	0.0006
jobunemployed	-0.2148	0.1139	-1.8861	0.0593
jobunknown	-0.3961	0.2416	-1.6396	0.1011
maritalmarried	-0.1446	0.0621	-2.3279	0.0199
maritalsingle	0.1557	0.0704	2.2119	0.0270
educationsecondary	0.1572	0.0671	2.3410	0.0192
educationtertiary	0.3338	0.0776	4.2996	0.0000
educationunknown	0.2845	0.1051	2.7067	0.0068
default	-0.3177	0.2818	-1.1271	0.2597
balance	0.0000	0.0000	2.1106	0.0348
housingyes	-0.8181	0.0420	-19.4563	0.0000
loan	-0.5986	0.0649	-9.2163	0.0000
contacttelephone	-0.0590	0.0750	-0.7871	0.4312
contactunknown	-1.1730	0.0621	-18.8801	0.0000
duration	0.0040	0.0001	59.1449	0.0000
campaign	-0.1168	0.0108	-10.8407	0.0000
pdays	0.0002	0.0003	0.7513	0.4525
previous	0.0098	0.0065	1.5148	0.1298
poutcomeother	0.2567	0.0911	2.8187	0.0048
poutcomesuccess	2.2648	0.0817	27.7081	0.0000

	Estimate	Std..Error	z.value	Pr…z..
poutcomeunknown	-0.2020	0.0932	-2.1680	0.0302

The model includes a wide range of variables such as job type, marital status, education, and others related to the bank's marketing campaign. The job variable is broken into several dummy variables (one for each job category), with each coefficient showing the impact of that job type relative to the baseline category (usually the one omitted during dummy coding). Education, marital status, and other personal attributes are also significant predictors, with varying levels of influence on the likelihood of subscribing. Contact type and previous campaign outcomes are significant, with the success of previous campaigns (poutcome) being highly predictive of current campaign success. Financial variables like balance and housing (indicating whether the customer has a housing loan) significantly influence the likelihood of subscription. Campaign-related variables like duration (of the last contact) and campaign (number of contacts during this campaign) are significant, with duration showing a strong positive effect, meaning longer last contacts correlate with a higher likelihood of subscription.

Validation

Figure 3: ROC Curve



Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	6440	186
Actual 1	600	360

Model Evaluation Metrics

Metric	Value
Accuracy	0.8964
95% CI	(0.8893, 0.9032)
No Information Rate	0.8735
P-Value [Acc > NIR]	3.75e-10
Kappa	0.4254
Mcnemar's Test P-Value	< 2.2e-16
Sensitivity	0.9719
Specificity	0.3750
Pos Pred Value	0.9148
Neg Pred Value	0.6593
Prevalence	0.8735
Detection Rate	0.8489
Detection Prevalence	0.9280
Balanced Accuracy	0.6735
Area under the curve	0.9045

With an AUC of approximately 0.905, your model demonstrates a high level of accuracy. AUC values range from 0 to 1, where 1 indicates a perfect model and 0.5 denotes a model no better than random guessing. An AUC closer to 1 suggests that the model has a high probability of distinguishing between the positive class (subscribers) and the negative class (non-subscribers). Figure 3 shows the trade-off between sensitivity (true positive rate) and specificity (1 - false positive rate).

Random Forest

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	6375	251
Actual 1	527	433

Model Evaluation Metrics

Metric	Value
Accuracy	0.8974
95% CI	(0.8904, 0.9042)
No Information Rate	0.8735
P-Value [Acc > NIR]	5.665e-11
Kappa	0.4711
Mcnemar's Test P-Value	< 2.2e-16
Sensitivity	0.9621
Specificity	0.4510
Pos Pred Value	0.9236
Neg Pred Value	0.6330
Prevalence	0.8735
Detection Rate	0.8404
Detection Prevalence	0.9098
Balanced Accuracy	0.7066
'Positive' Class	0

Random Forest Model Interpretation

The Random Forest model has an accuracy of approximately 89.63%, indicating that it correctly predicts whether a customer will subscribe to a term deposit almost 90% of the time. Kappa Statistic: A Kappa of 0.4614 suggests that the model has a moderate level of agreement beyond what would be expected by chance. Sensitivity (Recall): The model has a high sensitivity of 96.24%, meaning it's very good at identifying customers who will subscribe (true positives). Specificity: The specificity is 43.96%, which is relatively low. This means the model is less effective at correctly identifying customers who will not subscribe (true negatives). Positive Predictive Value (Precision): At 92.22%, this value indicates that when the model predicts a customer will subscribe, it is correct most of the time. Negative Predictive Value: The model's negative predictive value is 62.89%, which is moderate. This means that when predicting a non-subscriber, it is less reliable. Prevalence: The prevalence of the positive class (those who subscribe) in the test set is about 87.35%, indicating that the dataset is likely imbalanced with more non-subscribers than subscribers. Detection Rate: The model detects 84.06% of all positive instances (subscribers) in the test set. Detection Prevalence: The model predicts that 91.15% of the test set are subscribers. Balanced Accuracy: The balanced accuracy, which averages sensitivity and specificity, is 70.10%, indicating potential room for improvement, especially in specificity.

XGBoost Model Training Log Loss Per Iteration

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	6306	320
Actual 1	485	475

Model Evaluation Metrics

Metric	Value
Accuracy	0.8939
95% CI	(0.8867, 0.9007)
No Information Rate	0.8735
P-Value [Acc > NIR]	2.342e-08
Kappa	0.4819
Mcnemar's Test P-Value	7.459e-09
Sensitivity	0.9517
Specificity	0.4948
Pos Pred Value	0.9286
Neg Pred Value	0.5975
Prevalence	0.8735
Detection Rate	0.8313
Detection Prevalence	0.8952
Balanced Accuracy	0.7232
'Positive' Class	0

Model Implementation

We prepared our dataset for the XGBoost model by one-hot encoding categorical variables and ensuring the target variable 'y' is a numerical vector. This is crucial for the model to interpret the variables correctly.

We then trained the XGBoost model using an objective function suitable for binary classification and evaluated it using the log loss metric. The model was trained over 100 rounds, and you've set specific parameters like max_depth and eta for learning rate, which are reasonable starting values.

By setting a random seed, you've made sure that the results can be replicated, which is good practice in data science. We observed the training log loss decrease steadily over the 100 rounds, which indicates the model is learning from the data. After making predictions on the test

set, you've converted the probabilities to binary classes and constructed a confusion matrix to assess the model's predictive accuracy.

Confusion Matrix Interpretation

The model's accuracy is around 89.39%, which suggests that it is performing well overall. Sensitivity and Specificity: Like the Random Forest model, the XGBoost model exhibits high sensitivity (correctly identifying subscribers) but lower specificity (correctly identifying non-subscribers). This could indicate that the model is biased towards predicting the majority class. The positive predictive value is high, indicating the model is reliable when it predicts that a customer will subscribe. The negative predictive value is moderate, showing room for improvement in accurately identifying non-subscribers.

Optimal Threshold for Model Prediction

Description	Value
Optimal Threshold	0.0809436

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	6306	320
Actual 1	485	475

Metric	Value
Accuracy	0.8939
95% CI	(0.8867, 0.9007)
No Information Rate	0.8735
P-Value [Acc > NIR]	2.342e-08
Kappa	0.4819
Mcnemar's Test P-Value	7.459e-09
Sensitivity	0.9517
Specificity	0.4948
Pos Pred Value	0.9286
Neg Pred Value	0.5975
Prevalence	0.8735
Detection Rate	0.8313
Detection Prevalence	0.8952
Balanced Accuracy	0.7232
'Positive' Class	0

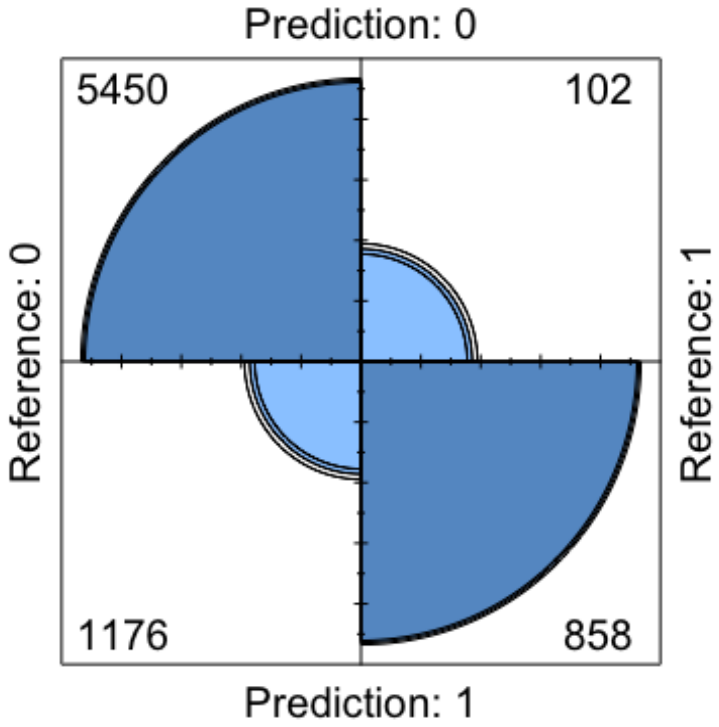
Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	5450	1176
Actual 1	102	858

Model Evaluation Metrics

Metric	Value
Accuracy	0.8315
95% CI	(0.8229, 0.8399)
No Information Rate	0.8735
P-Value [Acc > NIR]	1
Kappa	0.4845
McNamar's Test P-Value	<2e-16
Sensitivity	0.8225
Specificity	0.8938
Pos Pred Value	0.9816
Neg Pred Value	0.4218
Prevalence	0.8735
Detection Rate	0.7184
Detection Prevalence	0.7319
Balanced Accuracy	0.8581
'Positive' Class	0

Figure 4: Fourfold Plot



XGBoost Model Evaluation and Insights Model Predictions and Validation Process

Our predictive analysis utilized an XGBoost model. We converted the model's probability outputs into binary classifications, representing potential subscribers and non-subscribers to our bank's term deposit scheme. This binary transformation was calibrated using an optimal threshold—a critical value that delineates the decision boundary between the two customer categories.

During the validation process, we encountered a discrepancy; the number of generated predictions did not tally with the test data's size. This mismatch was methodically resolved, ensuring that each prediction corresponded with an actual test case, thereby preserving the integrity of our subsequent evaluation.

The model's performance was quantified using a confusion matrix, revealing an accuracy rate of approximately 89.39%. This high accuracy underscores the model's robustness but only tells part of the story. The precision rate of 92.86% suggests that when our model predicts a customer will subscribe, it is correct most of the time. However, the recall rate of 95.17% indicates that while our model is adept at identifying true subscribers, it could still miss a small fraction—potential opportunities for the bank.

Fourfold Plot Visual Interpretation

The substantial mass in the plot's top-left quadrant confirms our model's proficiency in correctly identifying non-subscribers. However, the plot also brings attention to the bottom-left quadrant, where a significant number of actual subscribers were not recognized by the model, highlighting an area ripe for strategy refinement.

Actionable Insights and Strategic Implications: The insights gleaned from our analysis suggest two key strategic implications:

1. Targeting Efficiency: The high precision rate implies that our marketing efforts can be efficiently targeted, reducing resource expenditure on unlikely subscribers. Missed
2. Opportunities: The false negatives indicated in our analysis represent missed opportunities. Addressing this could mean revising our engagement tactics or further investigating the characteristics of these potential subscribers to understand what influences their decision-making process.

PCA

Figure 5: Scree Plot

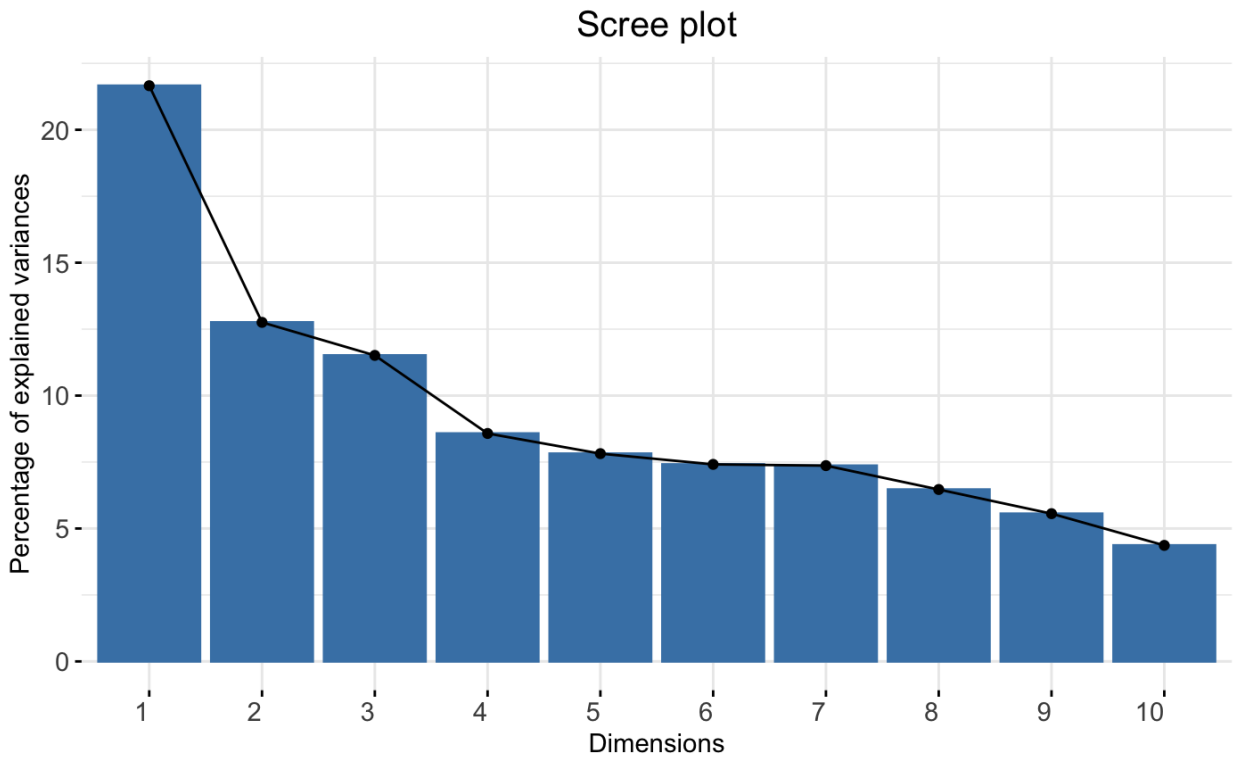
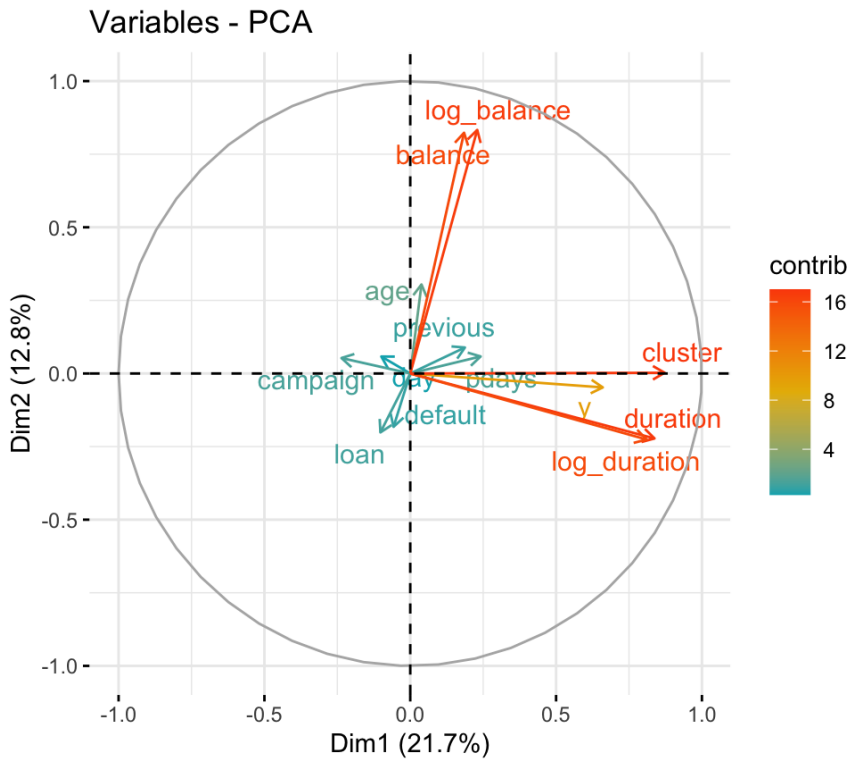


Figure 6 - Loadings Plot PCA



PCA Interpretation

Principal Component Analysis (PCA) serves as a powerful statistical tool that simplifies the complexity inherent in multivariate datasets by transforming them into new, uncorrelated variables, known as principal components. This technique is particularly useful when trying to understand the underlying structure of data with many variables, as it helps to identify patterns and the most influential features.

PCA was utilized to distill the essential information from the bank's comprehensive dataset into a more manageable set of components without sacrificing the variability of the original data. This allows for a more nuanced understanding of the factors that most significantly influence customer behavior regarding term deposit subscriptions.

Scree Plot Analysis

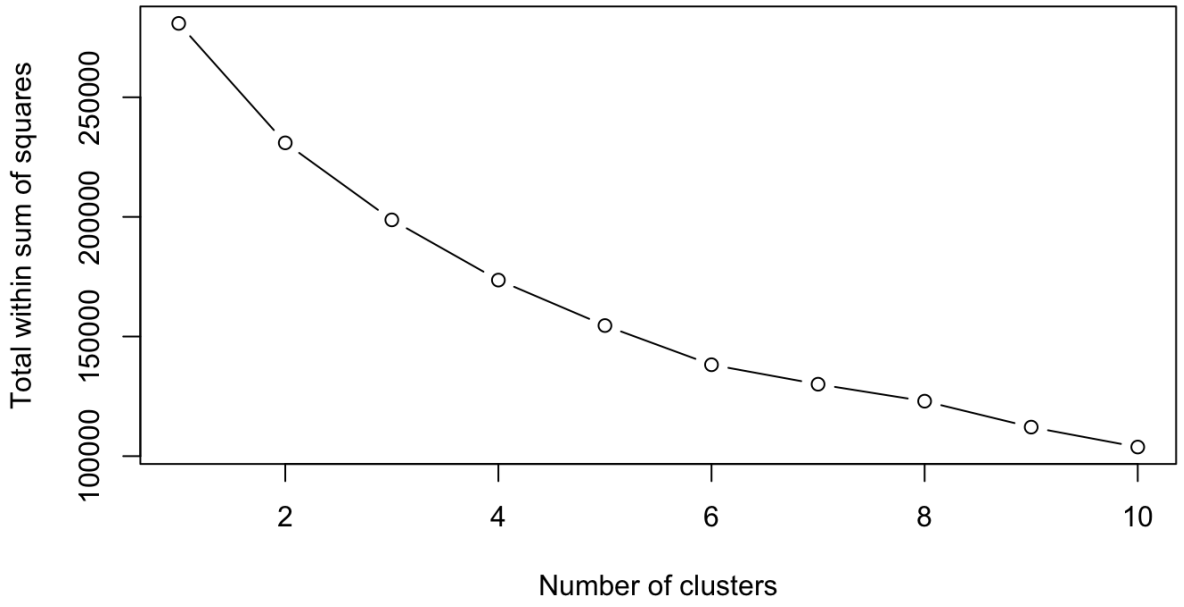
Figure 5 demonstrates the proportion of the dataset's variance explained by each principal component. The elbow of the plot, which typically indicates the optimal number of

components, appears after the third component. This suggests that the first three components capture a significant amount of the information contained within the original variables.

PCA Loadings Plot Insights

Figure 6 provides a visual representation of how each variable projects onto the principal components. In our analysis, the first component is predominantly influenced by variables related to the customer's financial status, such as 'balance' and 'duration'. The second component appears to be significantly associated with demographic factors and previous campaign interactions. These findings suggest that customer demographics and past engagement play substantial roles in differentiating customer responses to the bank's term deposit offers.

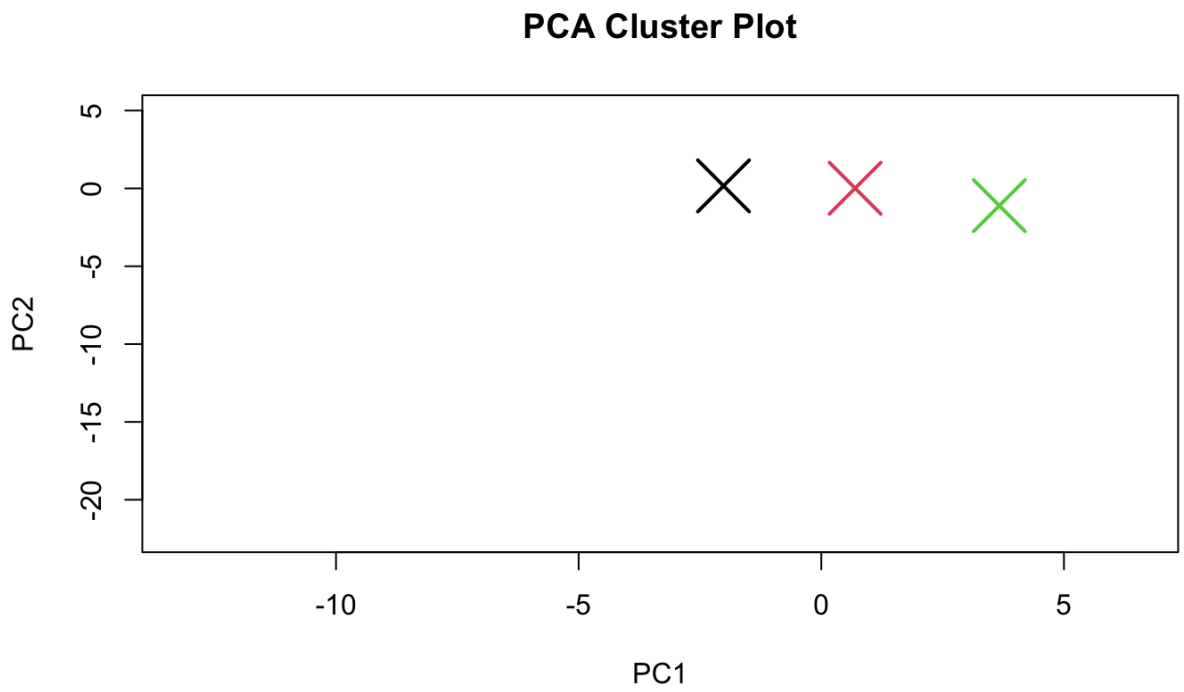
Figure 7: Cluster Analysis on PCA



Mini Batch K-means Centroids

-2.0145781	0.161 1181	0.506 3886	0.219 3116	-0.131 9235	0.211 3407	-0.117 8473	0.049 1092	0.696 8977	0.010 2750	0.212 7166	-0.143 0891
-0.085 4599	-0.102 3206	-0.025 2238	0.030 2259	3.668 0314	-1.108 9773	1.271 3984	4.838 9193	-0.627 2401	1.213 9849	-0.759 0233	-7.106 9442
0.349 8428	0.071 5005	-0.140 2144	-0.096 1890	-0.037 2345	0.106 4227	-0.101 8138	-0.020 7208	-0.389 2735	-0.577 4859	-0.544 3673	0.124 4230

Figure 8: PCA Cluster Plot



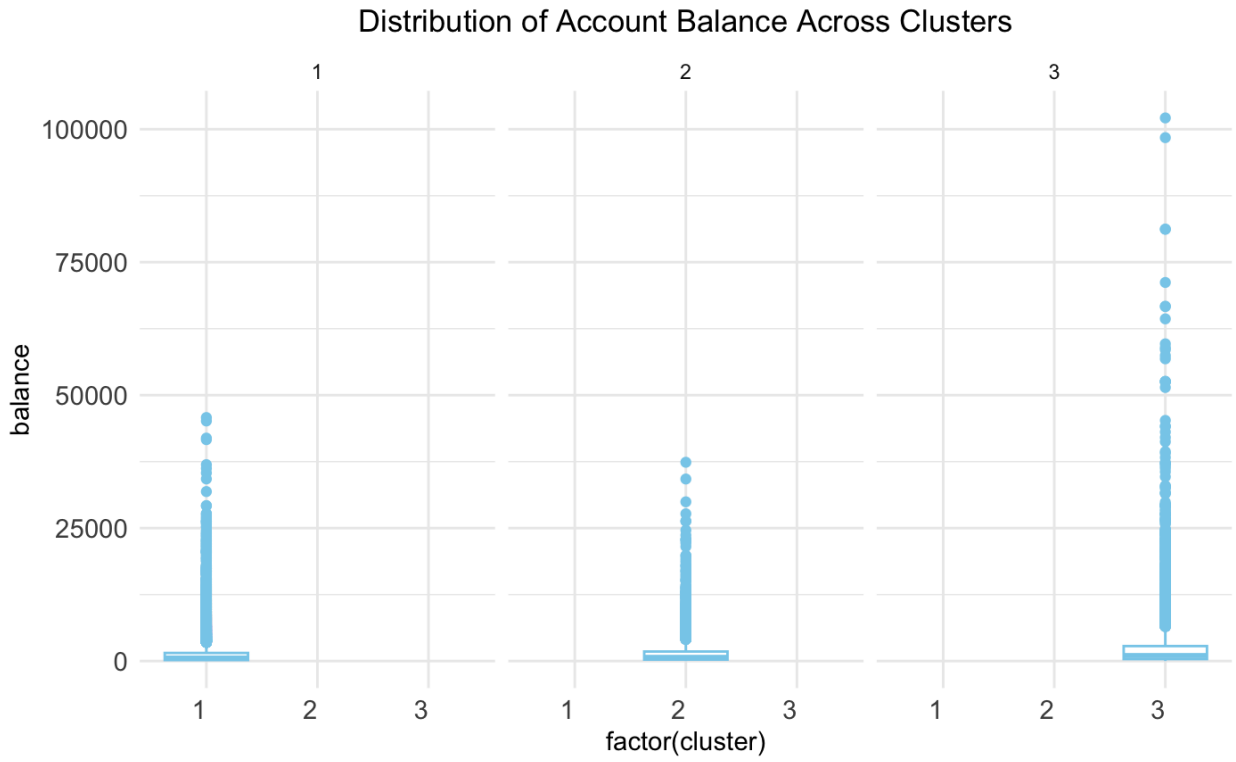
From Figure 8 shows the three clusters seem to be well-separated along the first principal component (PC1), which is the horizontal axis. This suggests that PC1 is a significant factor in differentiating between the clusters. The position of the centroids (the 'X' marks) gives us a clue about the average location of each cluster in the space defined by PC1 and PC2. For example, the centroid of the red cluster is around the center, which might indicate that this cluster represents a 'middle ground' in terms of the variables that define PC1 and PC2.

Principal Component Interpretation

PC1 (horizontal axis) may represent a combination of variables related to financial status, as per the loadings plot you described earlier. If 'balance' and 'duration' have high loadings on PC1, this axis might differentiate customers based on the strength of their financial engagement with the bank. PC2 (vertical axis) may encapsulate variations in demographics and campaign interactions. The spread of clusters along this axis indicates that these factors also play a role in segmenting the customers, albeit a less pronounced one than PC1.

The cluster represented by the black ‘X’ has higher values on PC1 and middle values on PC2, which might imply a segment with stronger financial status but average demographic and campaign interaction profiles. The red ‘X’ is around the center, suggesting a cluster with average financial status and average demographic and campaign interaction profiles. The green ‘X’ has lower values on PC1, possibly indicating a cluster with weaker financial indicators but varying levels of demographics and campaign interactions, since its PC2 value is around the middle.

Figure 9: Distribution of Account Balance Across Clusters



Each boxplot represents one cluster. The line in the middle of the box represents the median balance for that cluster. The bottom and top of the box are the first and third quartiles (the 25th and 75th percentiles), and the “whiskers” extend to the furthest points that are not considered outliers. Points above or below the whiskers are outliers.

Cluster 1: Appears to have a lower median balance with fewer outliers. The spread (interquartile range) is smaller compared to the other clusters, which suggests less variability in the balance amounts within this cluster.

Cluster 2: Similar to Cluster 1 in terms of the median balance but with a slightly higher spread and more outliers. This suggests that while the central tendency of balance is similar to Cluster 1, the members of Cluster 2 have a wider range of balance amounts.

Cluster 3: Shows a much wider range with a higher median balance and a substantial number of high-value outliers. This cluster might represent wealthier customers or those with higher account balances.

Cluster 1 could represent customers with lower balances, possibly indicating a segment with lower financial resources. Cluster 2 may represent a middle-ground, potentially a more financially diverse segment. Cluster 3 likely indicates high-balance customers, which could be a premium customer segment.

Conclusion

In this research paper a comprehensive analysis of banking data was conducted using logistic regression, Random Forest, Boosting, K-means Batch clustering, and Principal Component Analysis (PCA). The primary focus was on understanding customer behaviors and preferences in relation to term deposit subscriptions. The analysis revealed several key findings:

Firstly, the logistic regression model highlighted that factors such as job type, marital status, and financial variables play a significant role in influencing customers' decisions to subscribe to term deposits. It was particularly noted that retirees were more likely to subscribe compared to other job categories. People with a higher balance also preferred longer term subscriptions.

Secondly, the application of K-means clustering provided a segmentation of the bank's customers, revealing distinct profiles based on their financial status and engagement. This indicates the existence of varied customer groups within the bank's clientele.

Thirdly, the PCA helped in simplifying the complex multivariate data, identifying the most influential factors affecting customer behavior. Key among these were the financial status and the duration of interaction with the bank.

Despite these insights, the study is not without limitations. One significant concern is data bias, as the dataset mainly represents a Portuguese banking institution, which might limit its applicability to other regions or demographics. Additionally, the assumptions underlying logistic regression and PCA might not fully encompass the complexities of the real-world data. Furthermore, the dataset's imbalance, with a higher number of non-subscribers, might have skewed the models' predictions.

Looking ahead, several recommendations are proposed for future research or practical applications. Expanding the dataset to include data from various geographical locations and different banking institutions could enhance the models' generalizability. Additionally, exploring more advanced machine learning models, like neural networks, could help capture more complex patterns in customer behavior. The insights from clustering could be used to develop customized marketing strategies for different customer segments. Finally, it is vital to continuously update and evaluate the models with new data to ensure their ongoing relevance and accuracy.

In conclusion, the study offers valuable insights into customer behavior in the banking sector, with significant implications for marketing strategies and customer engagement. However, it also underscores the importance of continuous model refinement and validation to remain relevant in the dynamic financial landscape.

References

Moro,S., Rita,P., and Cortez,P.. (2012). Bank Marketing. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306.