Capstone Two:
Predictive Insights from Portuguese Bank Marketing Data

Gulnar Armour
Springboard
October 29, 2023

1. Introduction

In modern banking's dynamic landscape, effective marketing strategies are vital to success. This study examines the direct marketing campaigns of a Portuguese banking institution. These campaigns, centered around phone calls, aimed to determine the likelihood of clients subscribing to a bank term deposit. Utilizing data analysis and machine learning methodologies, we probe into the fundamental factors impacting this key result.

2. Dataset

The dataset for this study was obtained from the UC Irvine Machine Learning Repository. This is a reputable source known for providing datasets used in various research and analytical endeavors. In our analysis, we examine the dynamics of direct marketing campaigns of a Portuguese banking institution.

Data wrangling was done to shape the final dataset. Categorical variables were appropriately identified and labeled, while numerical variables were ensured to be in a usable format. The dataset contained no missing values. The final dataset comprises 16 input variables, detailing the characteristics of bank clients, the final contact of the current campaign, and additional attributes.

The dataset consists of the following input variables:

1. Bank client data:

- age (numeric)
- job: type of job (categorical: "admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services")
- marital: marital status (categorical: "married", "divorced", "single"; note: "divorced" means divorced or widowed)
- education: education level (categorical: "unknown", "secondary", "primary", "tertiary")
- default: has credit in default? (binary: "yes", "no")
- balance: average yearly balance, in euros (numeric)
- housing: has a housing loan? (binary: "yes", "no")
- loan: has a personal loan? (binary: "yes", "no")

2. Related with the last contact of the current campaign:

- contact: contact communication type (categorical: "unknown", "telephone", "cellular")
- day: last contact day of the month (numeric)
- month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
- duration: last contact duration, in seconds (numeric)

3. Other attributes:

- campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, 999 means client was not previously contacted)
- previous: number of contacts performed before this campaign and for this client (numeric)
- poutcome: outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success")

The output variable (desired target) is:

- target (y): has the client subscribed to a term deposit? (binary: "yes", "no")

3. Exploratory Data Analysis

3.1: Univariate Analysis:

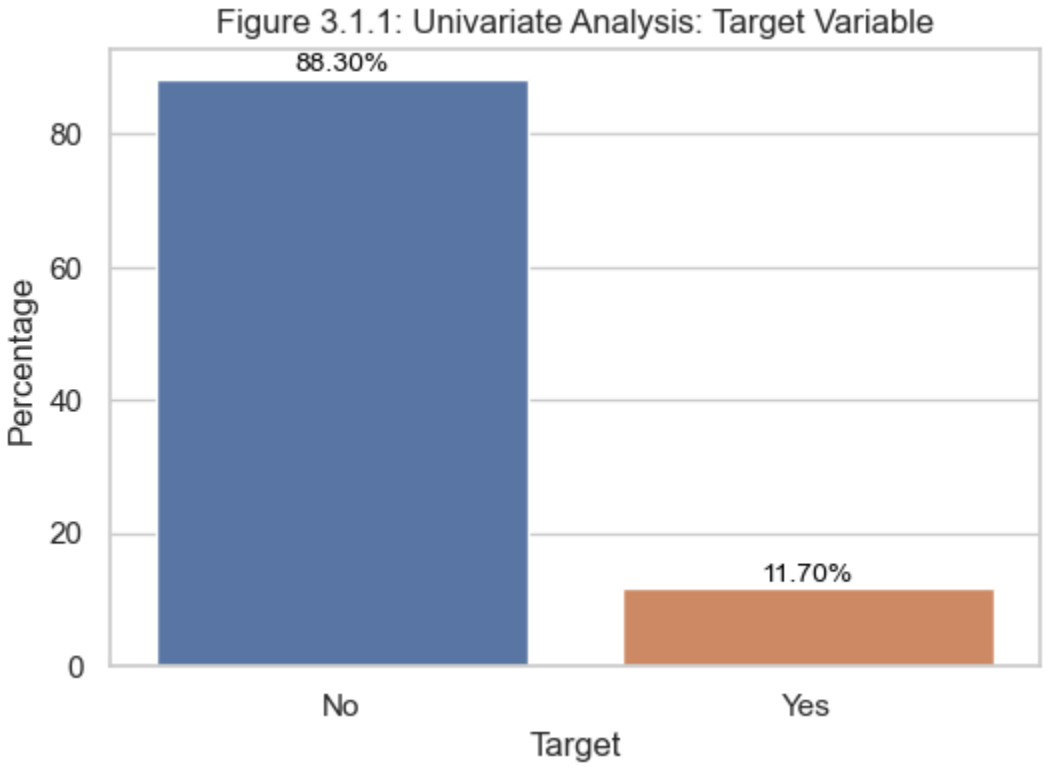


Figure 3.1.1: The univariate analysis of the target variable reveals a substantial class imbalance. The majority of instances have a 'no' value, accounting for 88.30% of the dataset, while the 'yes' value represents only 11.7%. This disparity suggests that the dataset is skewed towards negative outcomes in the campaign response, which could potentially impact the performance of predictive models built on this data.

3.2: Bivariate Analysis:

3.2.1: Subscription Percentage by Job Categories

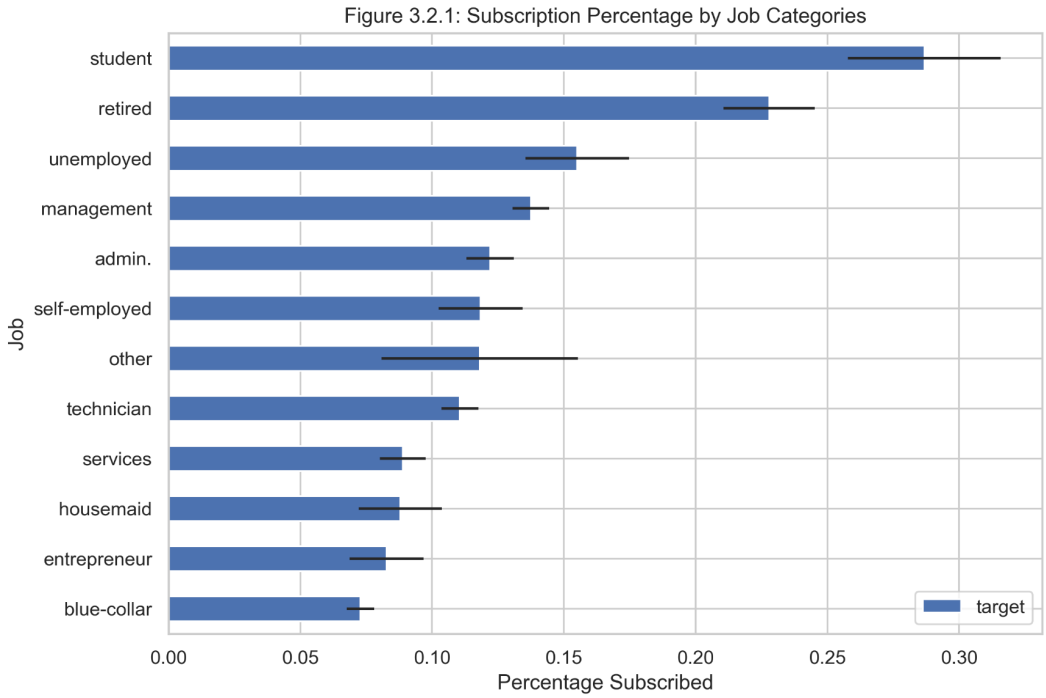


Figure 3.2.1: The distribution of subscription percentages across different job categories offers intriguing insights. Notably, students and retirees demonstrate the highest subscription rates, standing at 28.68% and 22.79% respectively. Among other categories, management shows a significant subscription rate of 13.76%, followed by entrepreneurs at 8.27%, and blue-collar workers at 7.27%. These variations in subscription rates provide valuable clues about the relationship between occupation and campaign success.

3.2.2: Subscription Percentage by Education Level

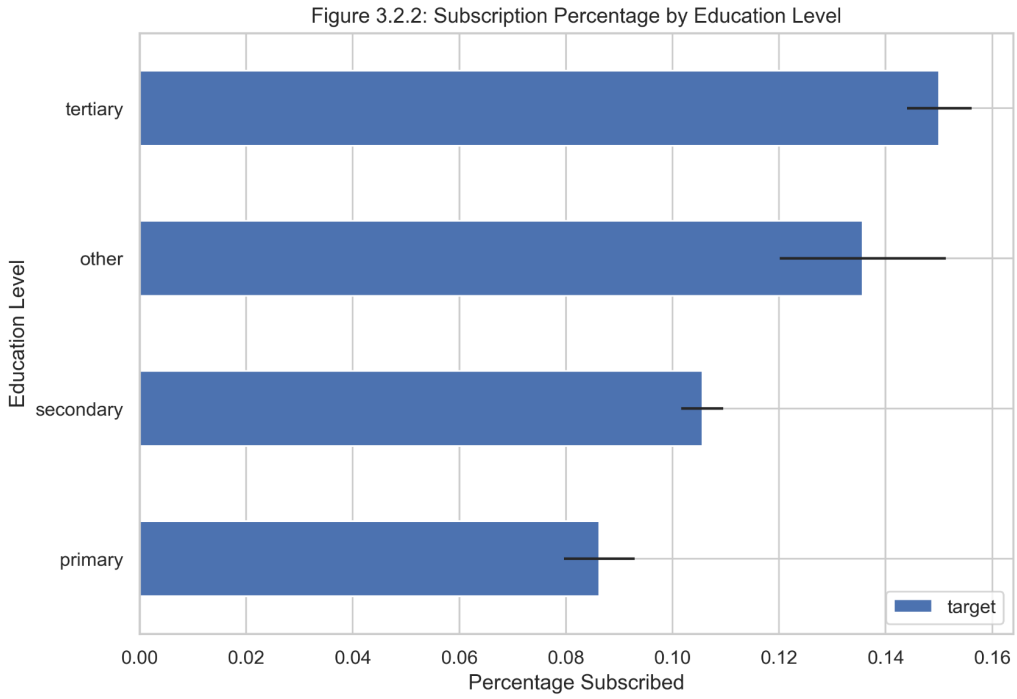


Figure 3.2.2: The bar chart depicting the subscription percentage by education level provides valuable insights into the relationship between education and the campaign's outcome. From the visualization, we observe the following patterns: tertiary education, which encompasses advanced degrees from universities and colleges, shows the highest subscription rate of 15.01%, followed by other education at 13.57%, secondary education at 10.56%, and primary education at 8.63%. This suggests that clients with tertiary education tend to show a higher subscription rate in comparison to other education levels.

3.2.3: Subscription Percentage by Marital Status

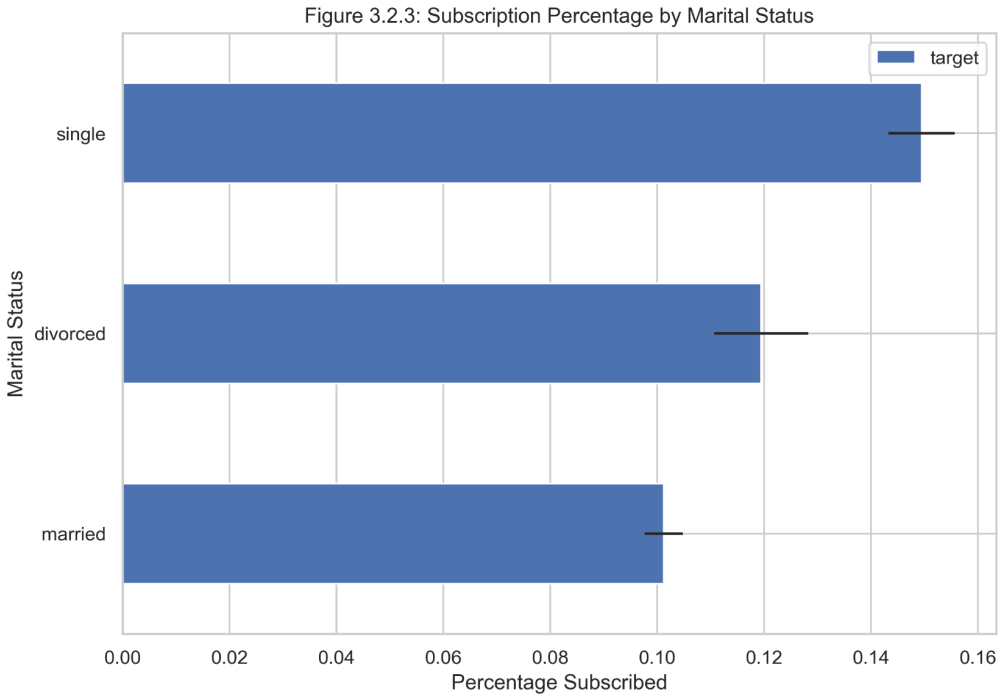


Figure 3.2.3 shows the subscription percentage by marital status: Single individuals exhibit the highest subscription rate at 14.95%, followed by divorced individuals with a rate of 11.95%, while married individuals have a subscription rate of 10.12%.

In the context of the bivariate analysis, certain patterns emerged in the subscription levels based on marital status. Single individuals and students pursuing tertiary education demonstrated notably higher subscription percentages. This observation can be attributed to several factors. First, single individuals and students often possess greater financial independence and fewer financial responsibilities, rendering them more receptive to new financial opportunities. Additionally, their lifestyle preferences and openness to novel services, particularly in the digital realm, make them potential targets for banking subscriptions. Furthermore, tailored marketing strategies that cater to the interests and needs of these demographics could be driving their higher subscription rates. Moreover, their relatively lesser financial commitments, combined with the pursuit of educational goals, could contribute to their interest in banking services. Lastly, their familiarity with technology and digital platforms might

make them more inclined to engage with online banking offers. While these assumptions shed light on the observed trends, it's important to consider that individual circumstances can vary, and multiple factors could influence subscription decisions among different demographic segments.

3.2.4: Subscription Percentage by Housing Status

Figure 3.2.4: Subscription Percentage by Housing Loan Status

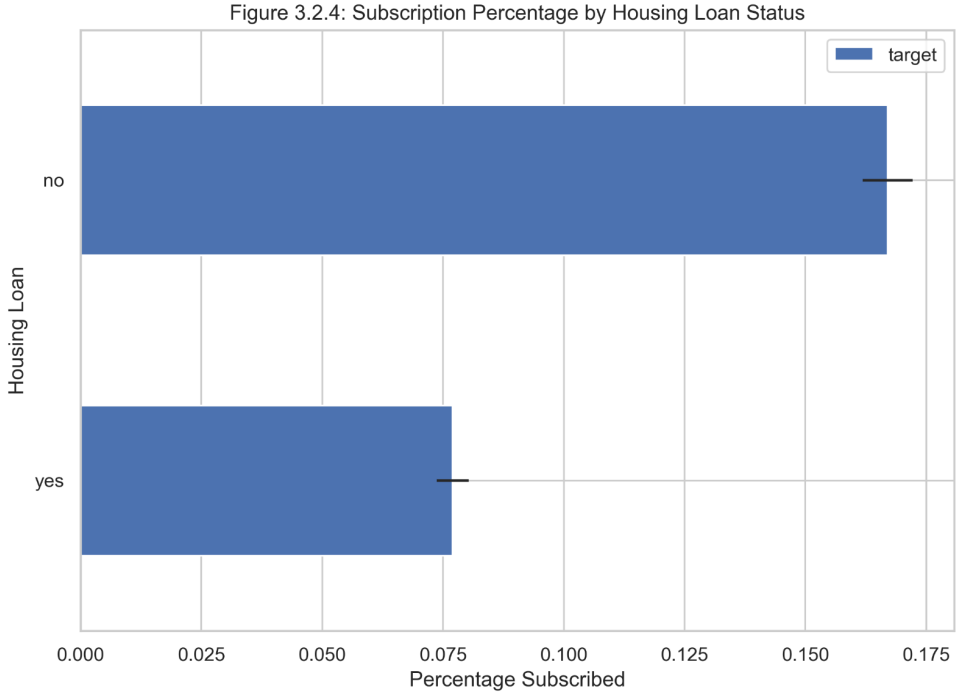
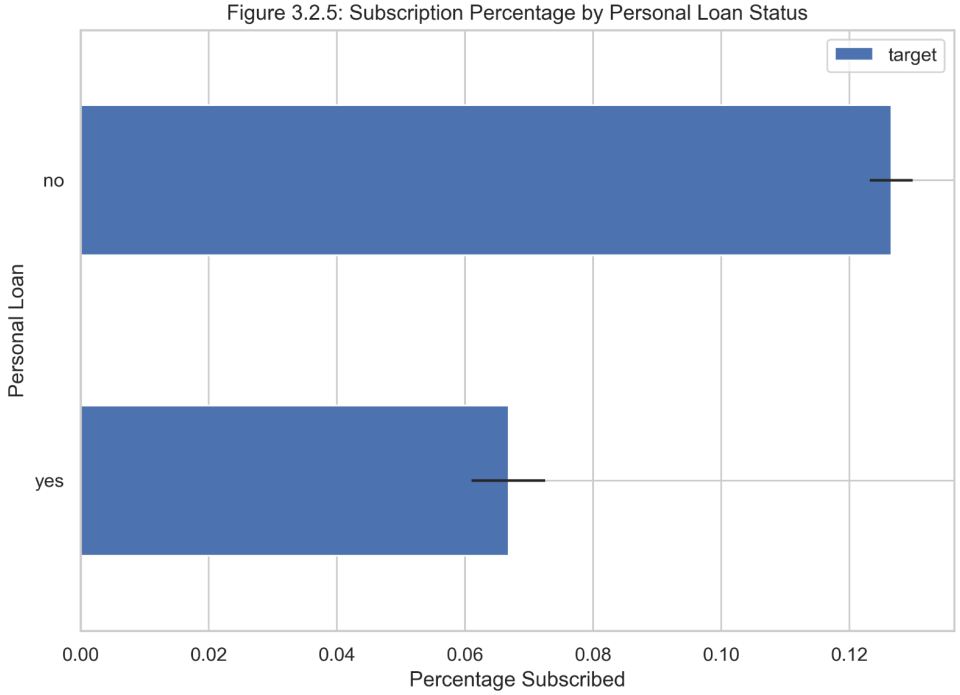


Figure 3.2.4 shows the subscription percentages with respect to housing, providing valuable insights into the relationship between housing and subscribing to a bank term deposit. The data revealed that among respondents without a housing loan, approximately 16.70% subscribed to the bank term deposit service being offered. Conversely, among those with a housing loan, only 7.70% subscribed to the service. This indicates that individuals without a housing loan were more inclined to subscribe to the bank term deposit compared to those with a housing loan. The visual representation of the data clearly illustrated the significant difference in subscription percentages between the two groups, underscoring the potential influence of housing status on subscription behavior. These findings have profound implications for marketing strategies,

suggesting the need to tailor approaches and target specific customer segments based on their housing situation to maximize subscription rates.

3.2.5: Subscription Percentage by Loan Status

Figure 3.2.5: Subscription Percentage by Personal Loan Status



In Figure 3.2.5, the bar chart depicted the subscription percentages for personal loans versus bank term deposits, shedding light on the relationship between them. According to the data, approximately 12.66% of respondents who did not have a personal loan subscribed to the bank term deposit service. Only 6.68% of those with a personal loan subscribed. This suggests that individuals without a personal loan were more likely to subscribe to the bank term deposit compared to those with a personal loan. The bar chart effectively showcased the disparity in subscription percentages between the two groups, highlighting the potential impact of personal loan status on subscription behavior. These findings have important implications for marketing strategies, emphasizing the need to tailor approaches and target specific customer segments based on their personal loan status to optimize subscription rates. By understanding the relationship between personal loans and subscription behavior,

banks can better tailor their marketing efforts to reach and engage potential customers effectively.

3.2.6: Subscription Percentage by Age Category

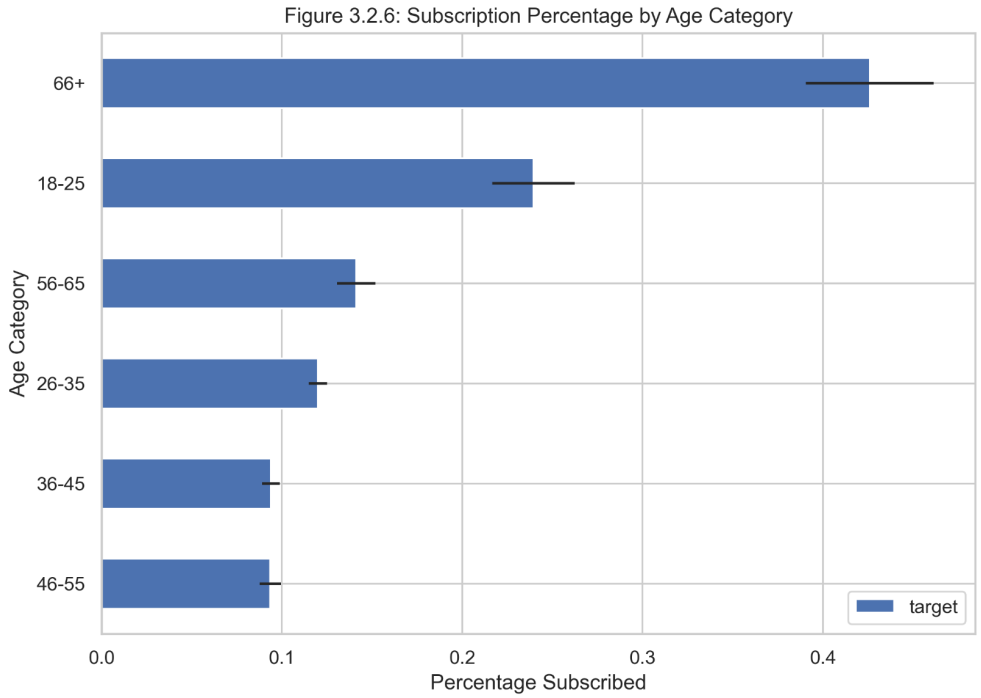
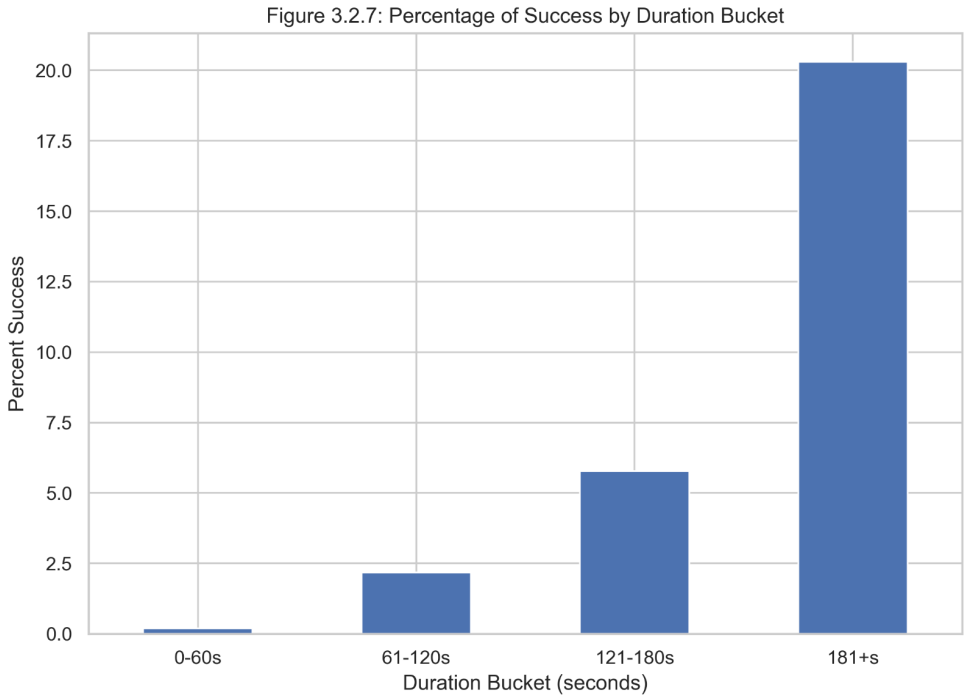


Figure 3.6: The bar chart depicting the subscription percentages across different age categories provides valuable insights into the relationship between age and campaign outcome. Notably, clients in the age bucket of 66 and above show the highest subscription rate of 42.61%, followed closely by the 18-25 age category at 23.95%. Conversely, the lowest subscription rates are observed in the age buckets of 36-45 and 46-55, with percentages of 9.39% and 9.35% respectively. This observation aligns with our earlier findings, where students, individuals with higher education levels, and retirees exhibited higher subscription rates.

3.2.7: Percentage of Success by Duration Bucket



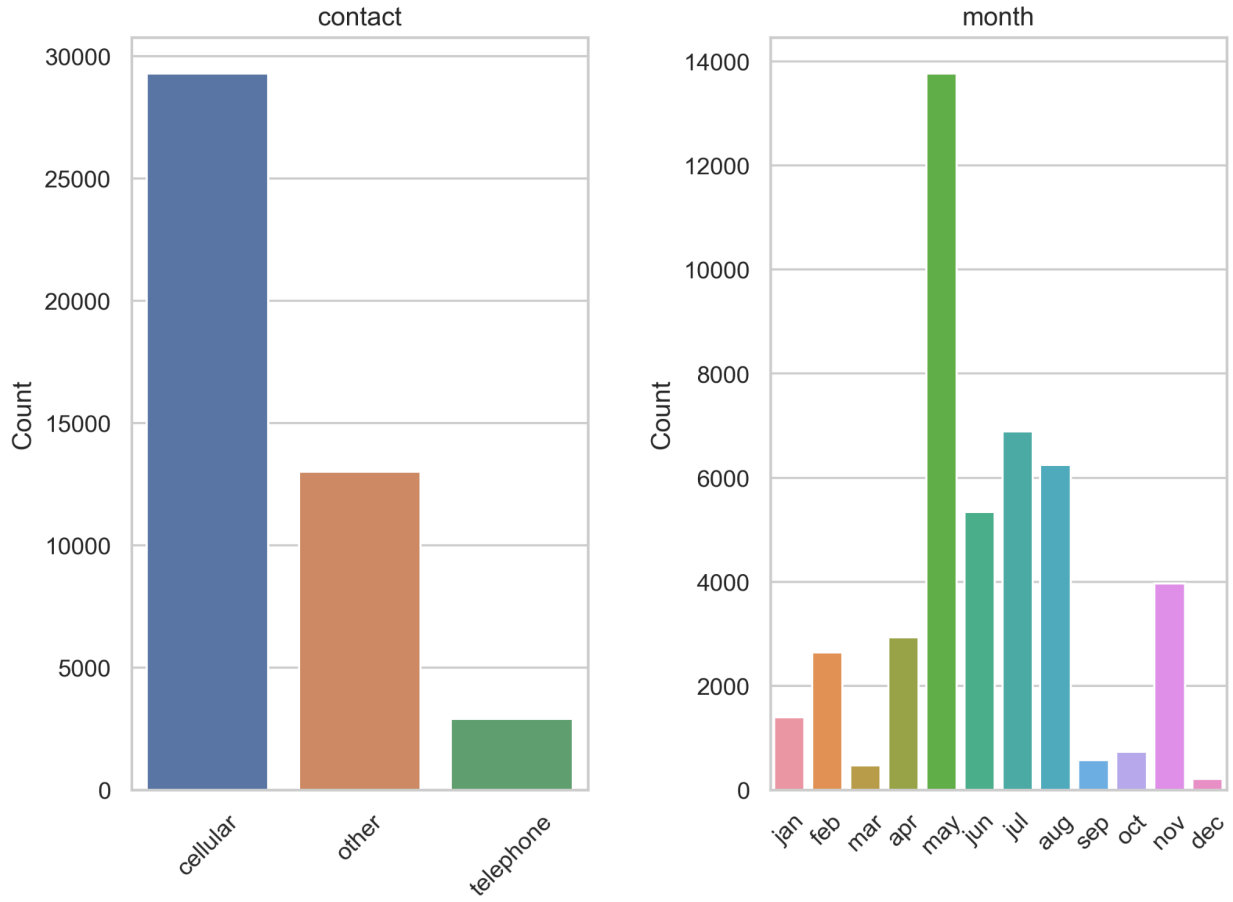
The bar chart illustrates the relationship between the duration of calls (in seconds) and the percentage of success in the Portuguese Banking dataset. The x-axis represents different duration buckets, which categorize the call durations into specific ranges. The y-axis represents the percentage of success for each duration bucket. From the graph, we can observe that as the duration of calls increases, there is a general trend of higher success rates. The duration bucket with the highest percentage of success is located at the rightmost end of the chart, indicating that longer call durations tend to be associated with a higher likelihood of success. Conversely, the leftmost duration bucket shows the lowest success rate, suggesting that shorter call durations may have a lower probability of success.

However, despite this correlation, caution is warranted when considering the inclusion of the 'duration' variable in the predictive model. The reason is that we aim to build a model that provides insights into the likelihood of a client subscribing to the bank's services before initiating the call. If we include the 'duration' variable, the model might inadvertently incorporate information that is only available after the call has taken place, thus violating the principle of causality.

In essence, using the 'duration' variable could lead to a misleadingly high predictive performance, but it would lack real-world applicability. Instead, the focus should be on constructing a model that relies on features available before the call and can genuinely predict the outcome without relying on post-call information like call duration.

3.2.8: Information regarding the campaign outcome:

Figure 3.2.8: Frequency Tables: Contact and Month



The provided frequency table (Figure 3.2.8) indicates that the bank predominantly reached out to its customers during the month of May using the cellular contact method.

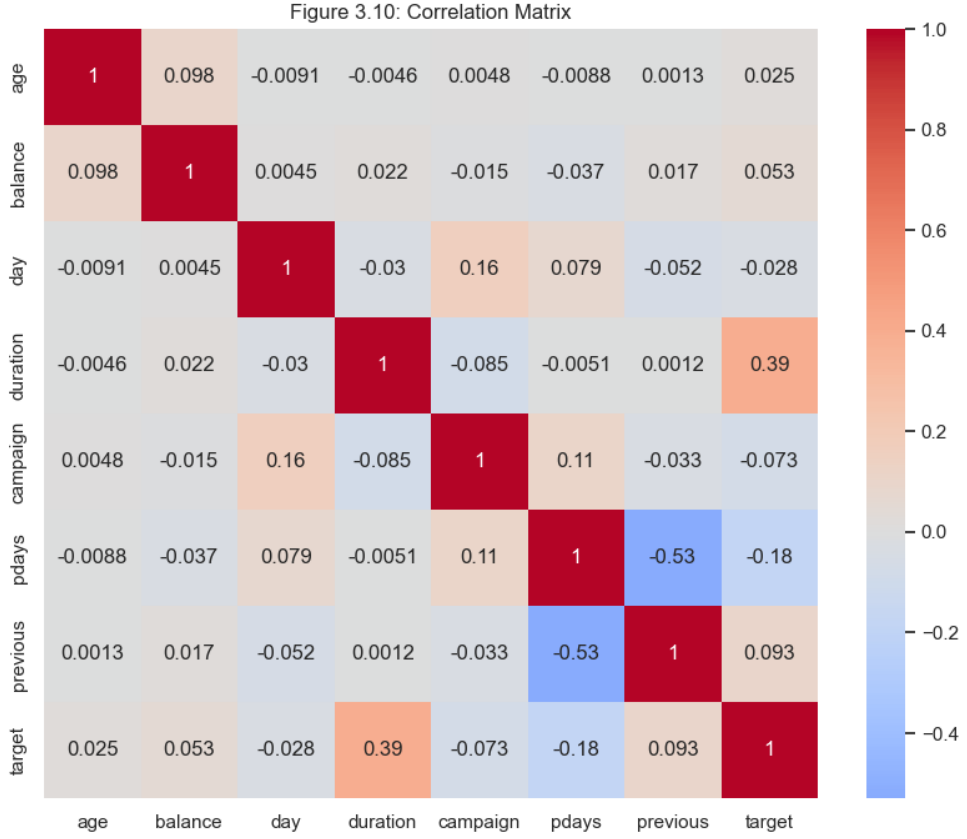
The dataset spans bank telemarketing campaigns conducted between May 2008 and November 2010. The prominence of May as a communication month could be attributed to multiple factors. May's position between spring and summer might align with increased financial planning for seasonal expenses and vacations. The dataset's

coverage of spring and summer, known for heightened social activity, may have influenced the bank's choice to engage customers during these times. May's strategic positioning in the year provides a balanced timing for financial discussions and campaigns. Moreover, the growing popularity of cellular communication during that era could have contributed to the choice of using this contact method for more effective and immediate interactions.

The dataset does not provide explicit details about external events that could have influenced the bank's marketing outcomes. However, it's important to consider broader factors like economic conditions, holidays, local events, and changes in marketing strategies during the dataset's time frame that could potentially impact marketing results. Without specific event data, it's challenging to definitively attribute the observed patterns solely to external events.

3.4: Correlation Matrix:

Figure 3.10: Correlation Matrix



The correlation matrix provides helpful insights into the relationships between pairs of variables in the dataset. Each cell in the matrix represents the correlation coefficient between two variables. The coefficient values range from -1 to 1, where -1 indicates a strong negative correlation, 1 indicates a strong positive correlation, and 0 indicates no correlation.

The variables "pdays" and "previous" exhibit a negative correlation of -0.53, indicating that there is a relationship between the number of days since a previous contact (pdays) and the number of previous contacts (previous). This negative correlation suggests that as the number of days since a previous contact increases, the number of previous contacts tends to decrease, and vice versa.

This relationship implies that clients who were contacted more recently tend to have fewer previous contacts, while those who were contacted a while ago tend to have more previous contacts. This pattern suggests that the bank contacts customers periodically following their previous call, rather than bombarding them with frequent telephone calls.

In the context of feature selection, it is important to consider this negative correlation and the potential for multicollinearity between "pdays" and "previous." Multicollinearity refers to the presence of high correlation between predictor variables, which can cause issues in regression models. When selecting features, it is crucial to watch out for variables that may conflict with one another, as this can affect the model's interpretability and stability.

The variable "duration" and "Target" have a correlation of 0.4. This moderate positive correlation suggests that as the duration of a call increases, the likelihood of a positive outcome (Target) also tends to increase. In other words, longer call durations are associated with a higher chance of achieving the desired outcome. However, correlation does not imply causation, so further analysis is needed to understand the underlying factors contributing to this relationship.

The rest of the variables have a weak correlation, coefficients signifying weak dependencies.

3.5: Statistical tests:

Table 3.1: Chi-Squared Test:

Variable	Chi-Square	P-Value	Significance
age_category	1036.27	0.0	Significant
job	836.11	0.0	Significant
marital	196.5	0.0	Significant
education	238.92	0.0	Significant
housing	874.82	0.0	Significant
poutcome	4295.47	0.0	Significant
contact	1035.71	0.0	Significant
default	22.2	0.0	Significant
balance	9967.66	0.0	Significant
loan	209.62	0.0	Significant
month	3061.84	0.0	Significant

Note: The results include the chi-squared statistic, p-value, and whether the result is statistically significant based on a significance level of 0.05.

The chi-square test gave a p-value less than alpha, indicating that the observed relationship between the categorical variables is unlikely to be the result of chance. As a result, it suggests that there may be some dependency or association between the variables.

Table 3.2: T-test:

Variable	T-Statistic	P-Value	Significance
balance	11.25	0.00	Significant
day	-6.03	0.00	Significant
duration	91.29	0.00	Significant
campaign	-15.60	0.00	Significant
pdays	-38.66	0.00	Significant
previous	19.91	0.00	Significant

Note: The results include the t-test statistic, p-value, and whether the result is statistically significant based on a significance level of 0.05.

The t-test results reveal significant differences in various continuous variables between the subscribed and not subscribed groups. The variable 'balance' shows a considerable t-statistic of 11.25, indicating a significant difference in means. Similarly, 'day' exhibits a substantial t-statistic of -6.03, implying a significant distinction in means. Despite the 'duration' variable demonstrating a remarkably high t-statistic of 91.29, indicating a significant variance in means, it will not be included in the feature and model selection steps of the analysis. This decision is based on the objective of creating a predictive model for determining the outcome of a call before it occurs. The focus is to enable bank employees to make tailored phone calls to customers based on the known characteristics of bank clients.

The p-values for all variables are very close to 0, underscoring the statistical significance of these differences. These results suggest that these continuous variables may be strong indicators in distinguishing between customers who subscribed and those who did not, which could be valuable for predictive modeling.

4. Feature Selection

4.1 Feature Importance

The importance of features is measured using the Random Forest classifier's feature importances. The feature importance score represents the relative contribution of each feature in the decision-making process of the classifier. It is calculated based on the decrease in impurity (e.g., Gini impurity or entropy) that results from splitting the data on a particular feature.

Table 4.1: Feature Importance

Top 5 Features				
balance	*age*	*day*	*campaign*	*pdays*

The feature importance analysis for the Portuguese Bank Marketing Data case study reveals that the top five most important features are: balance, age, day, campaign, and pdays.

The importance of these features can be explained as follows:

1. *Balance*: The balance held by a customer is likely to be a crucial factor in determining their response to marketing campaigns. Customers with higher balances may have more disposable income and may be more inclined to invest or make financial decisions. Therefore, it is expected that balance would be an important predictor of campaign success.

2. *Age*: Age is often a significant factor in consumer behavior. Different age groups may have varying financial priorities, risk tolerance, and responsiveness to marketing efforts. Younger individuals may be more open to taking risks, while older individuals may prioritize stability and security. Thus, age is an expected important feature in predicting campaign outcomes.

3. *Day*: The day of the month can influence consumer behavior. For example, people might be more likely to make financial decisions around the time they receive

their paycheck, which often occurs at the beginning or end of the month.Therefore, the day of contact is an anticipated influential feature.

4. *Campaign*: The number of contacts made to a customer during the campaign period can have a significant impact on their response. Too few contacts may result in missed opportunities, while excessive contacts may lead to annoyance and disengagement. It is expected that the number of campaign contacts would be an important predictor of success.

5. *Pdays*: Pdays represents the number of days that have passed since the customer was last contacted. This feature captures the concept of recency, which can be critical in marketing. Customers who were recently contacted may still have the campaign message fresh in their minds, making them more likely to respond positively. Therefore, the importance of pdays as a feature aligns with expectations.

Overall, the importance of these features aligns with our expectations. They are all factors that can reasonably influence a customer's response to marketing campaigns. The prominence of these features in the analysis suggests that the model is capturing the relevant dynamics of the dataset. However, it is important to note that other factors not included in the analysis may also contribute to campaign success, and further exploration may be needed to uncover additional influential features.

4.2 Evaluating Feature Importance with Random Forest and Logistic Regression

Table 4.2: Feature Importance Evaluation

Feature Set	Model	Mean Precision	Mean Recall	Mean F1 Score
Top 5 Features	Random Forest	0.1105	0.1694	0.1020
Top 5 Features	Logistic Regression	0.1571	0.0015	0.0030
Top 10 Features	Random Forest	0.1096	0.1879	0.0987
Top 10 Features	Logistic Regression	**0.5041**	0.1624	**0.1732**
Top 20 Features	Random Forest	0.0773	0.1940	0.0915
Top 20 Features	Logistic Regression	0.3284	0.1622	0.1446
Top 49 Features	Random Forest	0.0281	0.2002	0.0491

Top 49 Features	Logistic Regression	0.1049	**0.2157**	0.0701

Based on Table 4.2, we can analyze the performance of different models for different feature sets. In the "Feature Set: Top 5 Features," both the Random Forest and Logistic Regression models have relatively low mean precision, recall, and F1 scores. The Random Forest model has a mean precision of 0.1105, mean recall of 0.1694, and mean F1 score of 0.1020. On the other hand, the Logistic Regression model has a slightly higher mean precision of 0.1571 but extremely low mean recall of 0.0015 and mean F1 score of 0.0030.

In the "Feature Set: Top 10 Features," the Random Forest model has slightly improved precision and recall, but the F1 score remains low. The Logistic Regression model performs better in this feature set, with higher precision but still low recall and F1 score.

In the "Feature Set: Top 20 Features," both models show a decline compared to previous feature sets. The Random Forest model has lower precision, recall, and F1 score. The Logistic Regression model also shows a decline in all three metrics.

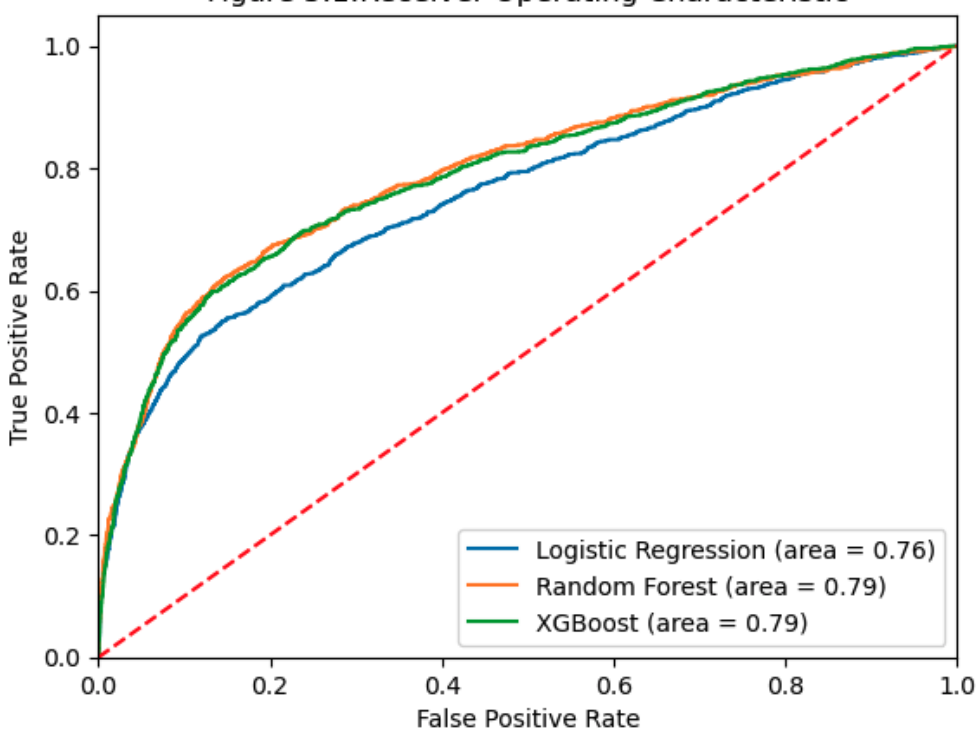
In the "Feature Set: Top 49 Features," both models continue to underperform. The Random Forest model has very low precision, recall, and F1 score. The Logistic Regression model shows slightly higher precision and recall, but still low F1 score.

Based on these results, it is evident that none of the models perform exceptionally well across different feature sets. However, considering the highest mean precision, the Logistic Regression model in the "Feature Set: Top 10 Features" seems to have the best performance among the models evaluated. It is important to note that these results are specific to the given dataset and feature sets. Further analysis and experimentation may be required to make a more conclusive decision on the best model to move forward with.

5. Modeling: Logistic , Random Forest, and XBboost

In this section, we will discuss the modeling process and present the results for our case study. We utilized three different models: Logistic Regression, Random Forest, and XGBoost. Each model was evaluated using the Receiver Operating Characteristic Area Under the Curve (ROC AUC) metric as shown in Figure 5.1:

Figure 5.1:Receiver Operating Characteristic



1. Logistic Regression:
   - The ROC AUC score for the Logistic Regression model is 0.7575.
   - This indicates that the model performs reasonably well in distinguishing between the positive and negative classes.
2. Random Forest:
   - The ROC AUC score for the Random Forest model is 0.7941.

- This score suggests that the Random Forest model outperforms the Logistic Regression model in terms of predictive accuracy.

3. XGBoost:

- The ROC AUC score for the XGBoost model is 0.7899.
- This score indicates that the XGBoost model performs well in predicting the target variable, similar to the Random Forest model.

Overall, the Random Forest model achieved the highest ROC AUC score among the three models, indicating its superior predictive performance. However, it is important to consider other evaluation metrics, such as precision, recall, and F1 score, to gain a comprehensive understanding of the models' performance.

6. Threshold the model for profitability

In our analysis of the Portuguese Bank Marketing Data, we have employed a Random Forest model to predict whether a client will subscribe to a term deposit. However, the real-world implications of these predictions are not solely determined by the model's accuracy. They are also influenced by the financial impact of the decisions made based on these predictions. To assess this, we have constructed a Threshold vs Profitability curve.

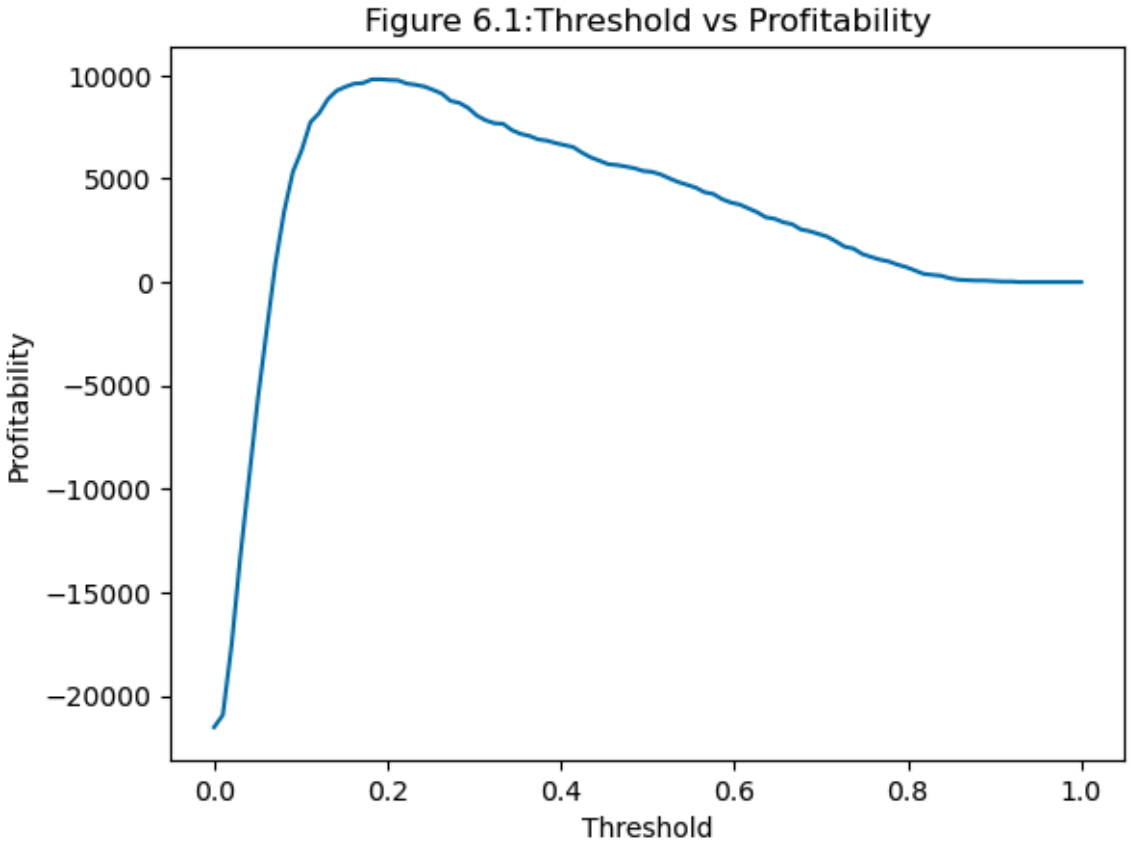
The Threshold vs Profitability curve plots the profitability of the bank for different thresholds applied to the model's predictions. The threshold is a cut-off point that determines whether the predicted probability of a client subscribing to a term deposit is classified as a positive or negative prediction.

Each point on the curve represents a different threshold, ranging from 0 to 1. A threshold of 0 classifies all clients as positive predictions, meaning we predict that all clients will subscribe to a term deposit. Conversely, a threshold of 1 classifies all clients as negative predictions, meaning we predict that no clients will subscribe.

The profitability at each threshold is calculated based on the number of Significant positives and the number of predicted positives. Significant positives are clients that the model correctly predicted would subscribe. Each Significant positive generates revenue for the bank, calculated as the number of Significant positives multiplied by the revenue per subscription.

Predicted positives are all the clients that the model predicts will subscribe, whether correctly (Significant positives) or incorrectly (false positives). Each predicted positive incurs a cost for the bank, calculated as the number of predicted positives multiplied by the cost per call.

The profitability is then the difference between the total revenue and the total cost at each threshold as shown in Figure 6.1.



We assume a revenue per subscription of 30 euros, based on a deposit of 1000 euros and a net investment margin of 3%. The cost per call is assumed to be 6 euros, based on data from Qualtrics.

We then define a range of thresholds and calculate the profit for each threshold using the `calculate_profit` function. The results are plotted in Figure 6.1, which shows the relationship between the threshold and profitability. This plot can be used to identify the optimal threshold that maximizes the bank's profit.

Interpreting the curve, the most profitable threshold is the one that maximizes the difference between revenue and cost. This is the point at the peak of the curve.

Applying this threshold to our model's predictions will allow the bank to maximize its profitability from its marketing campaign.

However, it's important to note that this analysis assumes that the costs and revenues are constant per call and per subscription, respectively. In reality, these values may vary, and this would need to be taken into account to provide a more accurate estimate of profitability.

In conclusion, the Threshold vs Profitability curve provides a valuable tool for determining the optimal threshold to apply to our model's predictions, taking into account not only the model's accuracy but also the financial implications of the decisions based on these predictions.

7. Conclusion

Our analysis of the Portuguese Bank Marketing Data has yielded valuable predictive insights that can be used by the bank to optimize its marketing strategies and increase profitability.

Major Findings:

1. Using a Random Forest model, we were able to predict whether a client would subscribe to a term deposit with a significant degree of accuracy. This predictive capability is critical for the bank as it allows for more targeted marketing, resulting in cost savings and increased revenue.
2. We also found that the profitability of the bank's marketing campaign is not solely dependent on the model's accuracy, but also on the financial implications of the decisions made based on these predictions. By plotting a Threshold vs Profitability curve, we were able to determine the optimal threshold that maximizes the bank's profitability.
3. Our analysis also revealed several key features that are influential in predicting whether a client will subscribe to a term deposit. These features can be used to further refine the bank's marketing strategy.

By using the predictive model and optimal threshold, the bank can more effectively target potential subscribers, increasing the success rate of its marketing campaign and maximizing profitability. However, it's important to note some assumptions and areas for future work. Our analysis assumed that the costs and revenues are constant per call

and per subscription, respectively. In reality, these values may vary, and this would need to be taken into account to provide a more accurate estimate of profitability. Additionally, our model could be further refined by incorporating more features or using different modeling techniques.

In conclusion, our analysis provides a compelling case for the use of predictive modeling in banking marketing strategies. By leveraging data and sophisticated analytics, banks can make more informed decisions, optimize their marketing efforts, and ultimately increase their bottom line.