National College of Ireland

Impact of Direct Marketing Strategies on Consumer Behavior in the Banking Sector

MSc Research Project
MSCAIBUS

Shilpa Pilla
Student ID: X23154713

School of Computing
National College of Ireland

Supervisor: Anderson Simiscuka

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Shilpa Pilla
Student ID:	X23154713
Programme:	MSCAIBUS
Year:	2023-2024
Module:	MSc Research Project
Supervisor:	Anderson Simiscuka
Submission Due Date:	12-08-2024
Project Title:	Impact of Direct Marketing Strategies on Consumer Behavior in the Banking Sector
Word Count:	5867
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	p.shilpa
Date:	12-08-2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	*yes*
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	*yes*
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	*yes*

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Impact of Direct Marketing Strategies on Consumer Behavior in the Banking Sector

Shilpa Pilla
MSCAIBUS

Abstract

Digitalization has radically changed the entire business of banking, increased competition, and increased consumer expectations. This paper discusses the efficiency of some popular direct marketing strategies for the banking industry using machine learning algorithms like logistic regression, decision trees, random forest, and gradient boosting. Some of the major objectives of this paper were to check the performance of these models on customer response prediction, examine demographic factors which may make a difference, and analyze the most important features contributing towards marketing campaign success. Conducted on detailed data from Kaggle, extensive data preprocessing, feature scaling, model training, and model evaluation formed part of the research. The findings revealed that Gradient Boosting and Random Forest models achieved the highest overall performance, with accuracies around 90.2 % and ROC AUC scores above 92 %, making them highly effective in distinguishing between positive and negative customer responses. The study also highlighted the role of feature engineering and demographic factors in determining marketing outcomes. Despite the fact that such models have been successful, there are still some areas to improve, such as enhancing recall rates and class imbalance. This study sets a path toward individualized, efficient, and successful campaigns that aim at better customer engagement and retention in the highly competitive banking industry of today.

1 Introduction

The rise of digitalization has completely transformed how the banking sector operates, causing a dent in the way banks communicate with their customers. Indeed, with increasing competition and consumer expectations, banks are now focusing on direct digital marketing practices to reach out to prospective customers more successfully, targeting customer acquisition and retention in the long run. This paper examines trends in direct marketing in the banking industry and how they translate into actual consumer behavior, the decision-making process, and loyalty.

Direct marketing is direct communication with the customersrequests by phone, email, or physical mail (Nayak and Siddiqui; 2024). It offers personalization in direct interaction between the bank and its customers. Messages sent through this technique are more likely appropriate to customer needs and preferences than traditional broad-scale advertising. This personalized method enhances not only the chance of conversing to a sale but also the strengthening of a customer's relationship with loyalty and long-term bonding

Figure 1: Direct Marketing

The digital age has made direct marketing more relevant, and more so, the data now available to banks on behavioral information and preference of customers. It thus becomes more accessible to identify and locate loci, which are then targeted with relevant tailor-made promotions and offers. For instance, a bank shall conduct analytics on customer data and identify an overdrawn customer to whom it shall target with offers on credit products that more effectively serve his financial needs. On the contrary, the effectiveness of direct marketing campaigns could significantly vary with the demographic characteristics of the receptive individuals, including their age, occupation, marital status, or even educational level. The younger customers, who usually have a positive attitude toward technology, could be more open to digital marketing activities. Older customers will be biased toward traditional skills, including direct mail or personal phone calls. The subtleties are something the banks need to appreciate as they proceed with such strategies in marketing, ensuring they have just the right way of interfacing with each set of their customer base. Moreover, the regulatory environment within the banking environment and finance sectors specifies challenges and constraints under which direct marketing may actually be practiced. For banks, related legal and ethical considerations bind their marketing practices in ensuring consumer protection and data privacy. This is important, as any wrong practice could cause significant reputational damage and heavy penalties.

This in brief, sets the stage for an in-depth investigation into the strategic application of direct marketing in the banking sector. It is this line of inquiry, analyzing the effects of the different strategies on consumer behavior and decision-making, that the work will seek to bring into relief actionable insights that would help the bank optimize its marketing endeavors. This study will explore how demographic factors interact with marketing effectiveness to provide insight into how banks use direct marketing to engage customers better, convert more new customers, and build lasting brand loyalty in a competitive digital environment.

1.1 Research Objectives and Questions

The primary objectives of this study are:

1. To analyze how different direct marketing strategies, influence customer decisions in banking.

2. To evaluate the role of demographic factors such as age, job, marital status, and education in the effectiveness of these marketing strategies.

3. To identify key attributes that enhance the success rates of direct marketing campaigns in the banking industry.

4. To utilize machine learning models to predict customer responses to direct marketing campaigns and improve the accuracy of targeting strategies

The research questions guiding this study are:

1. How can machine learning algorithms such as logistic regression, decision trees, random forest, and gradient boosting be applied to analyze the impact of direct marketing strategies on consumer behavior in the banking sector

2. Which demographic factors contribute to the effectiveness of these marketing strategies

1.2 Contribution to Scientific Literature

This work contributes to the available literature with a detailed examination of direct marketing by banks and an emphasis on demographic factors and machine learning models. In this way, this study offers new insights into maximizing marketing campaigns for customer engagement and loyalty. The paper considers in greater detail the regulatory and ethical challenges of direct marketing and makes some practical recommendations to enable banks to handle these complexities.

1.3 Structure of the report

The report is structured to provide insight into the impact that direct marketing strategies have on the behavior of consumers in the banking sector. This introduction part briefly describes the topic, research objectives, and questions and summarizes the study's contribution, which added value to the scientific literature. The literature review focuses on carefully considering existing studies about direct marketing strategies, demographic factors, and machine learning applications for the banking industry. The methodology section will present an overview of the research design, data collection tools, and techniques for analysis used in the study. The data analyses section, followed by descriptive and inferential statistics, and applications of machine learning models present their results. At last, the Discussion section will interpret the findings concerning prior literature and discusses practical implications for banking practice. Finally, there is the Conclusion section, wherein one summarizes the main findings, points out the study's limitations, and provides recommendations for future research. The results would then be clearly and logically presented in a very structured approach to research and its results.

2 Related Work

2.1 Direct Marketing Strategies in the Banking Sector

The literature review of academic sources on direct marketing strategies within the banking sector reveals several approaches and their outcomes. For instance, (Mero and Taiminen; 2016) focus on the role of personalization in marketing communications and further support that those banks using direct marketing for purposes connected with the introduction of new products and personalized advice in financial issues realized higher engagement rates. They, however, refer to the limited ability for integration of data sources as one of the primary reasons that makes it not easy to compile a customer

view. To that respect, (Nayak and Siddiqui; 2024) discuss traditional marketing shifting towards digital direct marketing due to increased reach and precision as some of the enormous strengths. For them, though with significant benefits, digital direct marketing has the huge problem of excluding those demographics who are not so tech-savvy, which may be a significant limitation. (Sahni et al.; 2018) confirm the effectiveness of personalized email marketing due to higher open rates and click-through rates for the individualized campaign. Their results remain encouraging, but they raise the complex issue of data privacy within banking. De Bruyn and Lilien contrast that targeted online advertising is more effective in stimulating. Their study, however, also makes mention of the significant cost and hassle involved in conducting a behavioral targeting advertising campaign. (Verhoef et al.; 2010) talk about customer engagement through direct marketing. Value-added campaigns, according to them, through direct marketing, possess a strong influence on creating customer loyalty. The positive aspect of their research is that it covers a detailed examination of customer data; however, at the same time, it also accepts that obtaining updated information about customers is a highly challenging task.(Kumar and Shah; 2004), talk about loyalty programs based on direct marketing strategies and illustrate that the rate of customer retention can be increased to a considerable extent. But they also indicate that over time, constant innovation is needed to make such programs work correctly.

2.2 Machine Learning Applications in Direct Marketing

Machine learning applications in direct marketing are becoming popular because many studies have underscored their potential to improve marketing effectiveness. (Cui et al.; 2006) illustrate how Bayesian networks, coupled with evolutionary programming, can predict customer response to marketing campaigns. Their results indicate a significant improvement in targeting accuracy using the suggested methodologies; however, the complexity of implementing these machine learning models is equally revealed. (Tekouabou et al.; 2022) use a case to explain how a European bank uses the independent variables to run ML algorithms to optimize direct marketing efforts, achieving an added acquisition of customers by 20 %. While these results may be auspicious, the study highlights the high cost and expertise required to deploy such ML solutions. (Dwivedi et al.; 2021) connect with the broader application of digital and social media marketing; they comment that ML models allow marketers to collect and analyze consumer data more holistically. At the same time, they alert against possible ethical concerns due to data privacy and the transparent data usage policy. (Ellahi et al.; 2024) confirm the role of omni channel strategies boosted by ML but mention a high investment required in technology and training. (Gefen and Straub; 2000), discuss the role of trust in the adoption of e-commerce and direct marketing strategies and tout that perceived ease of use affects customer engagement. This article is nearly strident about the psychological sides of marketing; however, it even addresses the issue of how the measurement of trust comes with many pitfalls. (Acquisti et al.; 2020) present challenges in the preservation of privacy in this digital age. One of the tensions they bring out is between personalized marketing and consumer privacy. Their findings indicate that ethical concerns are pretty relevant for implementing how ML-driven marketing strategies should be conducted.

2.3 Case Studies on ML-Enhanced Marketing

As many case studies prove, the effectiveness of applying machine learning to direct marketing is the one that, for example, in the case of Bank of America in 2017, the employment of ML algorithms in customer data analysis increased the relevancy of their email marketing campaigns, which led to the increase of response rates by 30 %. For example, (Crespo and Govindarajan; 2018) report showed how a European bank used machine-learning algorithms to fine-tune its direct marketing. This intelligence improved customer acquisition by 20 % and cross-selling differentiation by 25%.

2.4 Summary of Findings

The literature reviewed describes the vast potential of direct marketing strategies in the banking sector, more so when aided by machine learning applications. This literature points to personalized and targeted marketing campaigns that bring about better customer engagement, loyalty, and acquisition. However, a few limitations persist, like challenges concerning data integration, high implementation costs, and ethical concerns about data privacy. Arguably, a litany of existing solutions leaves a lacuna in effectively tackling issues, hence the necessity for more research. The present study will, therefore, close these gaps through an intricate understanding of how different direct marketing strategies and demographic factors interact with each other and finally come up with the sophisticated use of ML models in optimizing marketing efforts with ethical guarantees.

3 Methodology

This research applies a mixed-method approach to combine both quantitative analysis in parallel with machine learning techniques for the impacts of direct marketing strategies in their banking sectors. The research uses a publicly available data set from Kaggle. The data is very detailed, and it contains the actual information pertaining to clients and banks with whom the clients have had some prior interaction from direct marketing campaigns.

3.1 Data Collection

This study was conducted by making use of the "Banking Dataset Marketing Targets" dataset from Kaggle. The dataset is licensed through CC0: Public Domain. This dataset consists of features: age, job, marital status, education, default history, housing loan status, contact type, month of contact, day of the week, details of the campaign, outcomes resulting during earlier campaigns, and outcomes pertaining to subscription.

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

Figure 2: Dataset Sample

The dataset is publicly available on Kaggle under the license CC0: Public Domain. Thus, there are no restrictions for its use. The sources of data are online collection methods and phone conversation tracking that ensure a record of all interactions with the customers. This dataset is already used in many predictive modeling tasks; it's quite effective to learn about the effectiveness of direct marketing in the banking sector. It is an instrument with richness in its features and a perfect data collection methodology that can make this research great.

3.2 Raw Data Analysis

3.2.1 The Dataset

Some of the features contained in the dataset used in this research include age, job, marital status, education, default history, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome, and target variable y. The dataset contains 45,211 entries; each entry is an instance that corresponds to a client who has had interaction with the bank in terms of direct marketing campaigns. As it were, this dataset is detailed and well-structured with both numerical and categorical data types.

3.3 Data Cleaning

3.3.1 Handling Missing Values:

The dataset was checked for missing values. As shown by the result of checking missing values, the dataset does not have missing values. This ensures that the dataset is complete and ready for further pre-processing without imputation or handling missing data.

3.4 Data Transformation

3.4.1 Handling Outliers:

As shown in the Fig. 5, there are a number of features that have extreme outliers. Some of these features are age, balance, duration, campaign, pdays, and previous. Since outliers can have an effect on the performance not only in the way machine learning models are fitted but also in their performance, thus it is very important to manage them to have an accurate prediction.

Outliers in the dataset were handled using the interquartile range method.

3.5 Exploratory Data Analysis

Exploratory Data Analysis (EDA) involves analyzing the main characteristics of the dataset, often using visual methods. Here, the study explores the distributions of various features in the dataset, their relationships with each other, and how they relate to the target variable.

3.5.1 Distribution of numerical Features

These different distributions of numerical features convey several insights into this dataset: the distribution of age is very much skewed toward the young population, with a significant peak around age 30. It indicates that there may be a large proportion of

```
# The structure of our datasets
train_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        45211 non-null  int64
 1   job        45211 non-null  object
 2   marital    45211 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
dtypes: int64(7), object(10)
memory usage: 5.9+ MB
```

Figure 3: Dataset Structure

```
# Checking missing values
missing_values_train = train_df.isnull().sum()
missing_values_test = test_df.isnull().sum()

missing_values_train, missing_values_test
```

```
(age          0
 job          0
 marital      0
 education    0
 default      0
 balance      0
 housing      0
 loan         0
 contact      0
 day          0
 month        0
 duration     0
 campaign     0
 pdays        0
 previous     0
 poutcome     0
 y            0
 dtype: int64,
```

Figure 4: Missing Values

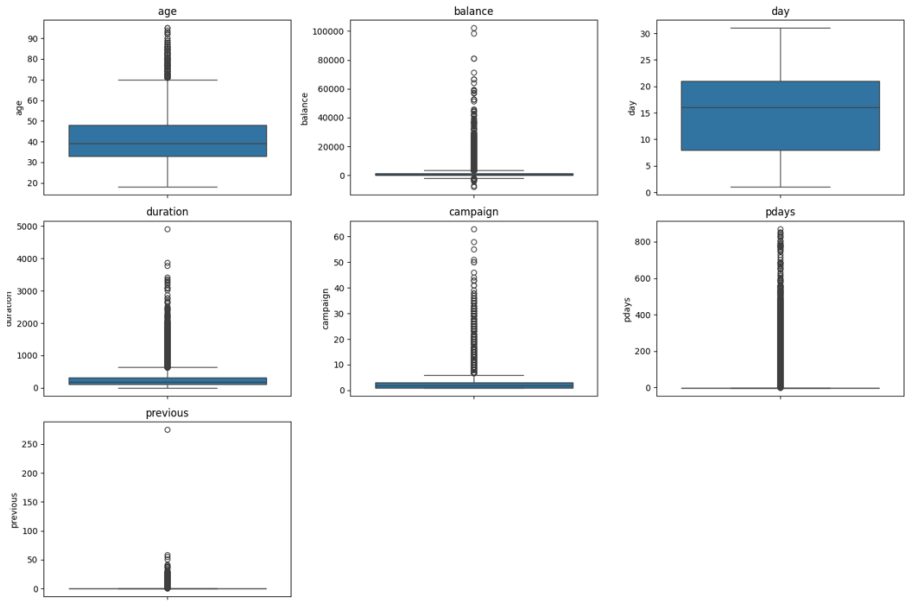

Figure 5: Outliers

clients within this age group and therefore most likely one of the key demographics to target with marketing strategies. The balance feature indicates a right-skewed distribution, with most clients having a balance close to zero and very few having large balances, showing it has a small number of rich clients. The day feature, representing the day of the month when the client was last contacted, seems uniformly distributed, suggesting that marketing campaigns were spread evenly throughout the month and no preference toward particular days of the month was found.

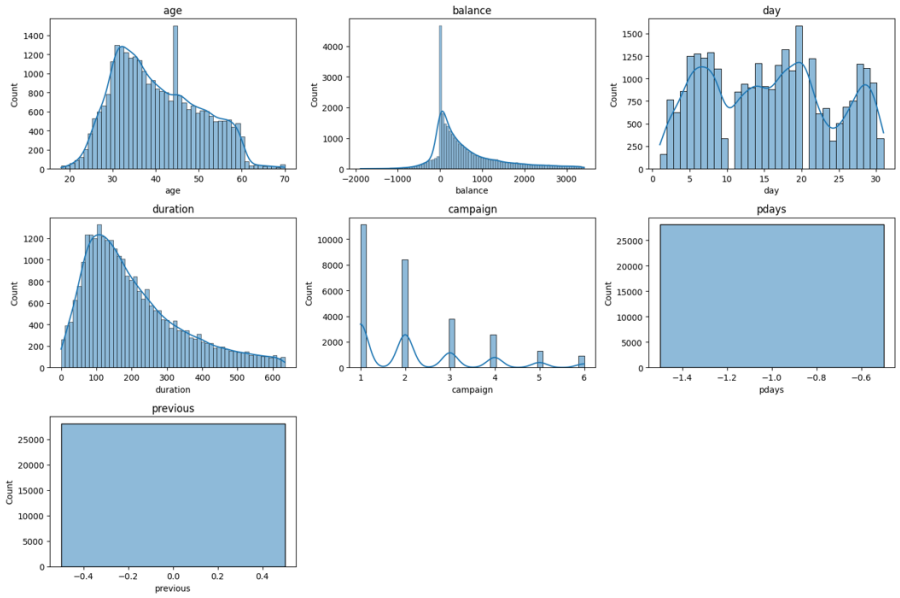

Figure 6: Feature Distributions

The duration feature, which is the last-contact duration in seconds, is right-skewed and indicates most of its contacts to last less than 200 seconds. However, there are some cases where the contacts had taken a much longer time. This would seem to suggest the higher values of this feature will be meaningful, and maybe a lot of talk times would result in reached goals. The feature campaign, where the representation is numerical: the number of contacts executed during this campaign, is severely right-skewed in that the majority of clients were contacted either once or twice. The pdays feature, representing the number of days since the client was last contacted from a previous campaign, shows a strange distribution with a concentrated number of observations in the -1 range, meaning that many of the clients have not been contacted before. Finally, as for the number of contacts performed before this campaign, it has a lot of density at zero, showing that for many clients, the current campaign is the first time they were ever contacted. Again, these distributions speak volume on the variation in engagement for the clients and suggest different marketing strategies needed to be harnessed upon different segment of client base.

3.5.2 Distribution of categorical features

The categorical feature distribution includes demographic information and characteristics of clients in the dataset. This job distribution indicates that most clients fall into

the blue-collar/management category, with fewer clients being, for example, students or housemaids. This might mean targeted marketing based on people's various occupations. Marital status shows a greater number of married clients than single or divorced, a factor that probably points to a large fraction of clients with other financial priorities and stability.

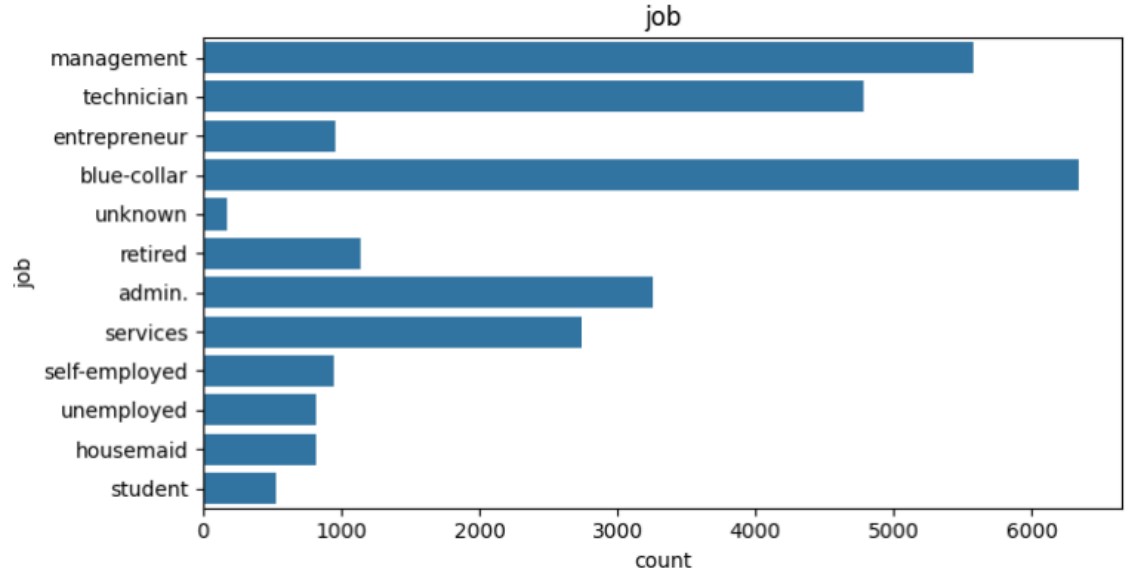

Figure 7: Job Features

Education levels reveal a large number of clients with secondary and tertiary education, with a smaller number of clients having primary education. The majority of clients do not have credit defaults, which indicates a generally financially responsible client base. Housing and loan statuses show that many clients have housing loans, but fewer have personal loans, which could influence their response to loan-related marketing campaigns.

The contact feature shows that most clients were reached via cellular phones, with fewer contacted via telephone or unknown methods. The month feature indicates that the highest number of contacts were made in May, suggesting a potential seasonal trend in marketing efforts. Finally, the outcome feature, which shows the outcome of the previous marketing campaign, has a large number of unknowns, which could be further investigated to understand past campaign performance. The target variable y shows that a majority of clients did not subscribe to the term deposit, indicating a need for more effective targeting and personalized marketing strategies.

The correlation heatmap provides the relationships between numerical features in the dataset. The values range from -1 to 1, indicating the strength and direction of the correlation. Most features show weak correlations with each other, as evidenced by the values close to zero. For instance, age and balance have a slight positive correlation, suggesting that older clients may have slightly higher balances. The strongest positive correlation is observed between day and campaign, indicating that the number of contacts is somewhat related to the day of the month. Negative correlations, such as duration and campaign, imply that longer call durations are slightly associated with fewer contact attempts. Overall, the heatmap indicates that the numerical features in the dataset are relatively independent of each other, which is useful information for model building as it

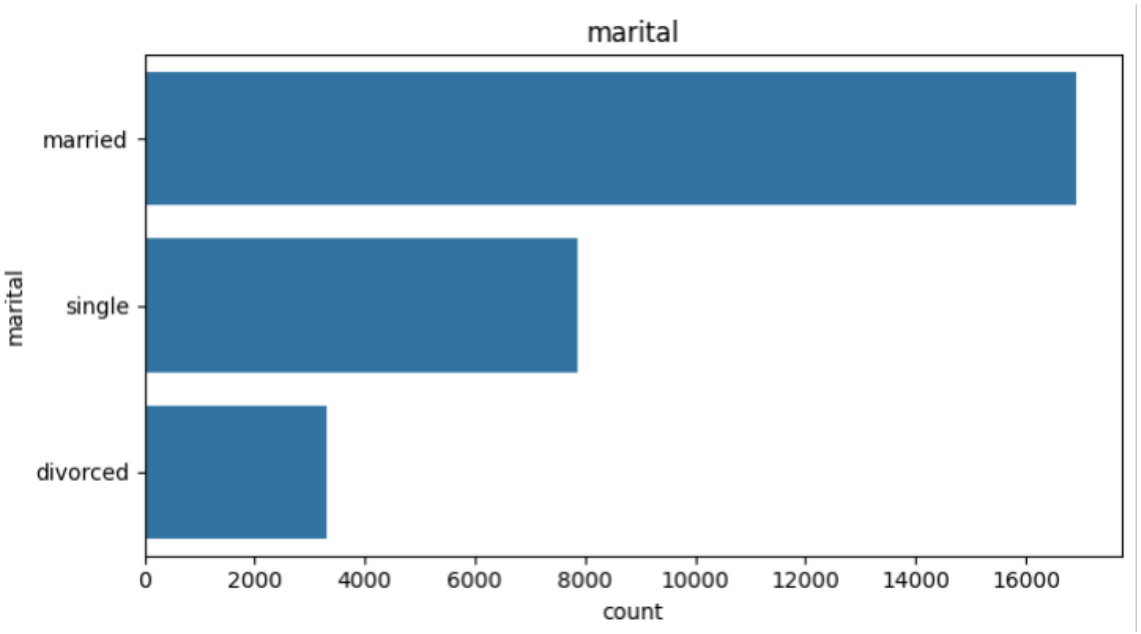

Figure 8: Marital Feature

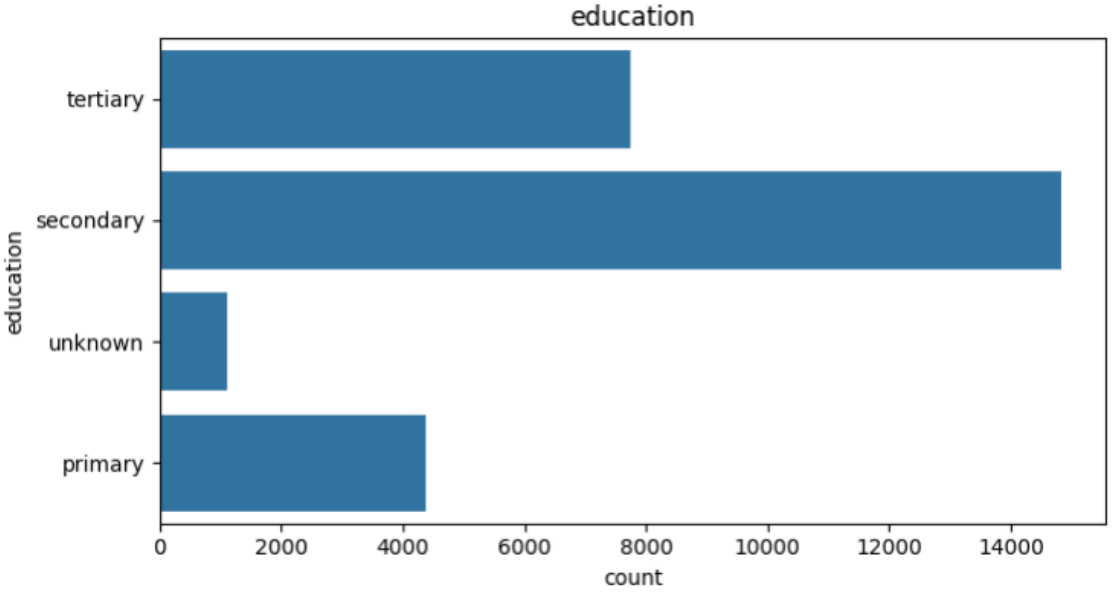

Figure 9: Education Features

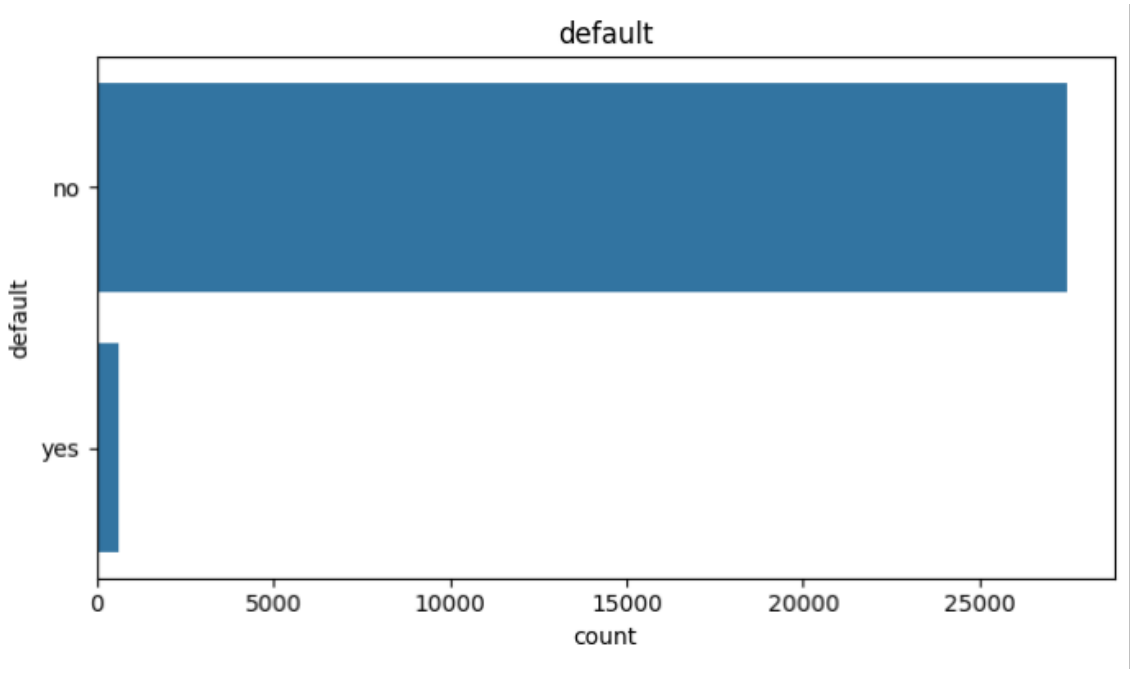

Figure 10: Credit Default Feature

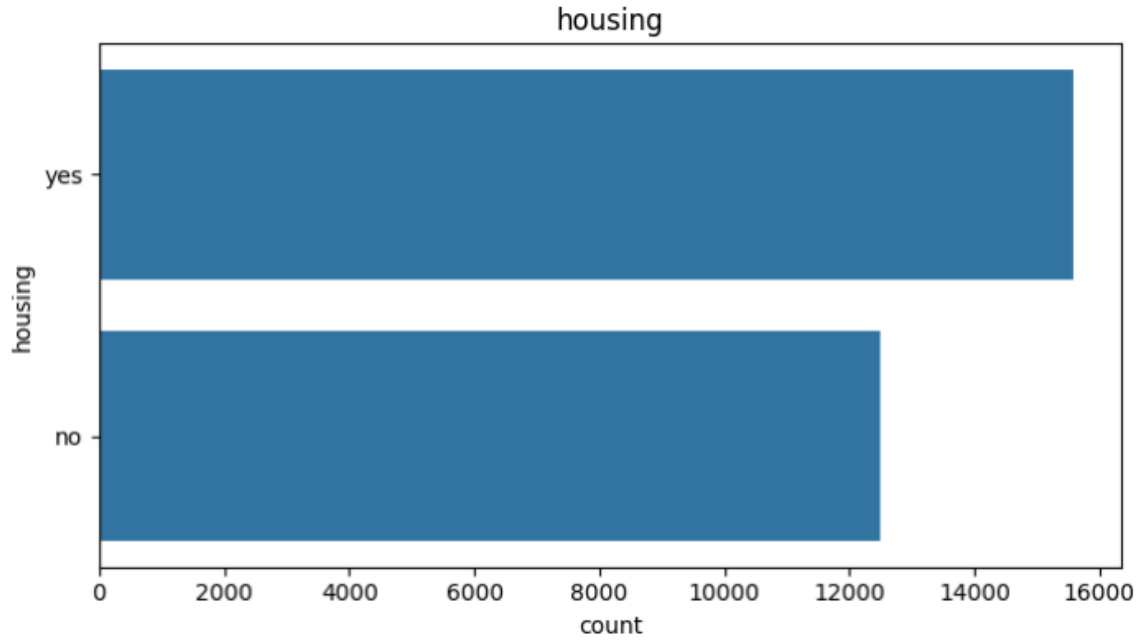

Figure 11: Housing Feature

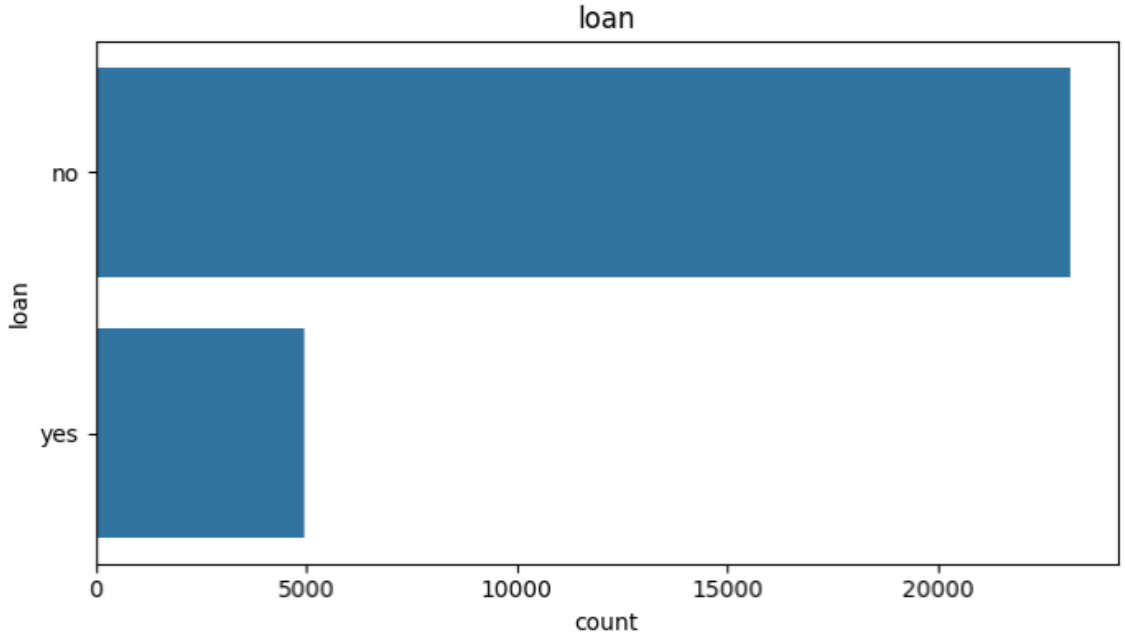

Figure 12: Loan Feature

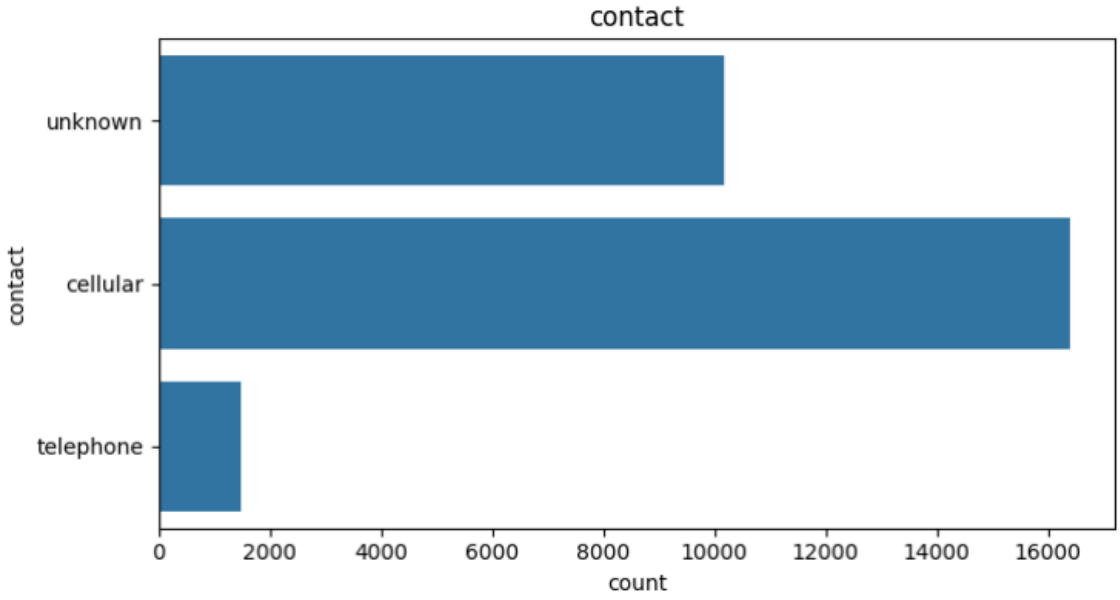

Figure 13: Contact Feature

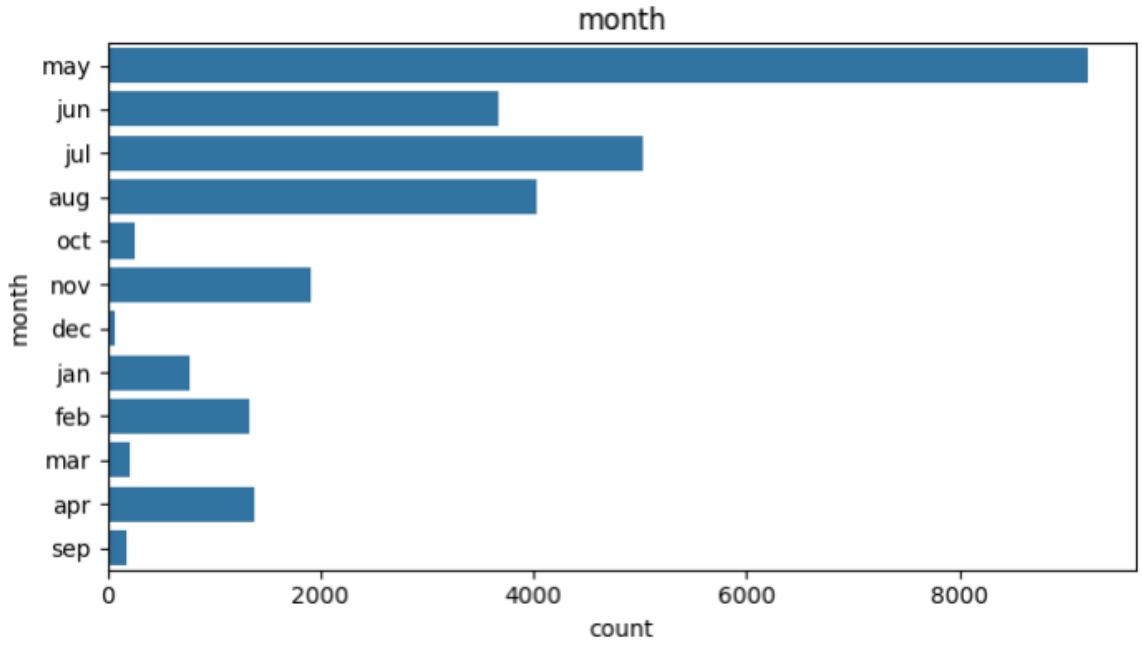

Figure 14: Month Feature

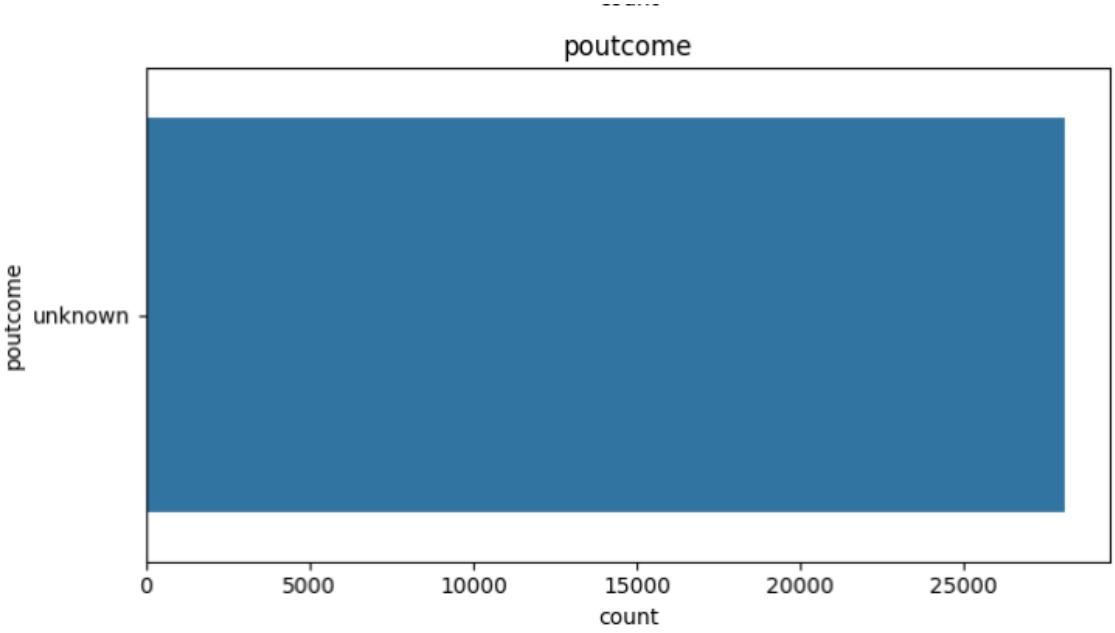

Figure 15: P-Outcome Feature

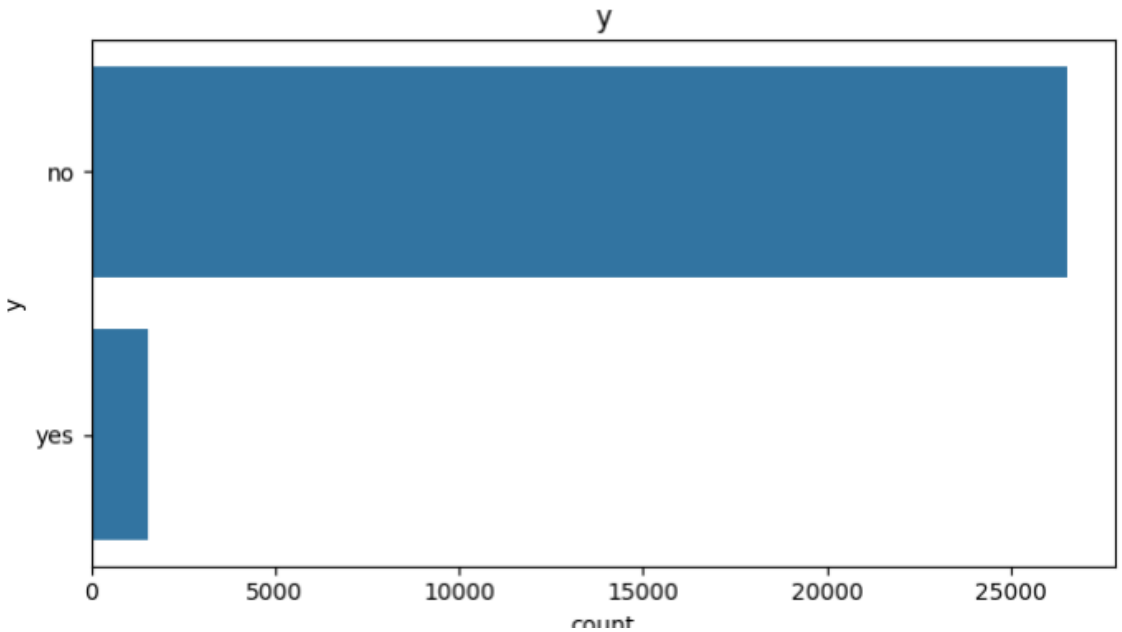

Figure 16: Target Variable

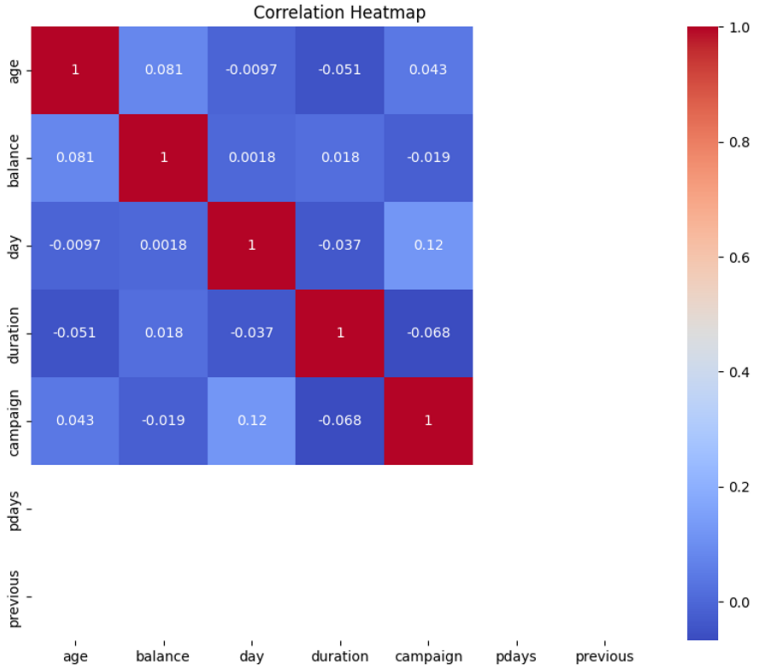

Figure 17: Heatmap

reduces concerns about multicollinearity.

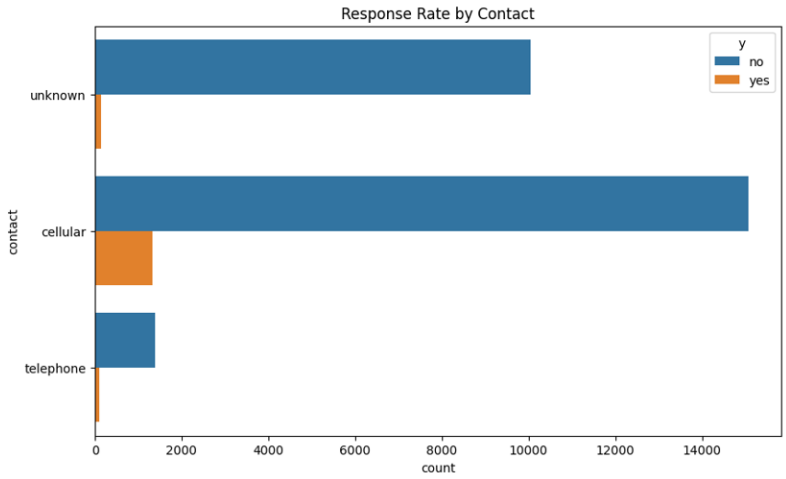

Figure 18: Response Rete by Contact

The bar graph shows that the majority of clients were contacted via cellular phones, followed by an unknown method and telephone. Among these contact methods, cellular contact has the highest number of positive responses (yes) for subscribing to the term deposit, indicating it is the most effective communication channel. The unknown contact method, despite having a significant number of total contacts, has a very low positive response rate. Telephone contact has the least number of contacts and responses overall. This descriptive statistic suggests that cellular contact is the most promising method for direct marketing campaigns in the banking sector, significantly outperforming other methods in terms of client engagement and response rate.

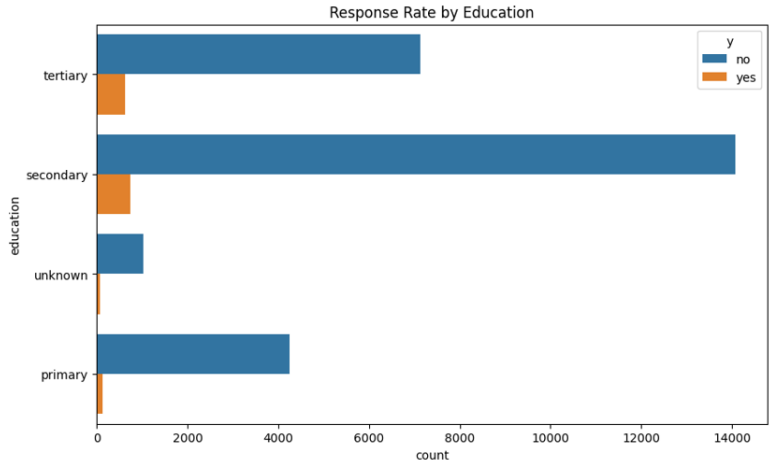

Figure 19: Response Rate by education

The bar plot shows the response rate to the marketing campaign categorized by the clients' education levels (primary, secondary, tertiary, and unknown). Clients with secondary education constitute the largest group, followed by those with tertiary education. The response rate for clients with tertiary education is notably higher compared to

those with primary or secondary education, indicating that higher educational attainment might correlate with a greater likelihood of subscribing to the term deposit. The lowest response rate is in the primary education group, so it can be that marketing strategies were more oriented towards clients with higher levels of education and should be adjusted to attract more clients with lower educational levels. The 'unknown' category has the smallest number of contacts and responses; hence, this is just incomplete data. This is a simple descriptive statistic telling about the high impact of educational level upon the effectiveness of campaigns: better-educated clients respond more.

4 Design Specification

In carrying out this research, sophisticated blending of techniques, architecture, and frameworks was used to unravel how direct marketing strategies impacted the banking sector.

4.1 Techniques/Architecture/Framework

4.1.1 Data Processing Pipeline:

In this work, a seamless data processing pipeline was developed to ensure that the dataset is meticulously prepared for rigorous analysis. This involves a series of steps in logical order: from data loading to pre-processing, feature engineering, model training, and evaluation. All phases are very important in ensuring the integrity of the data and accuracy of models.

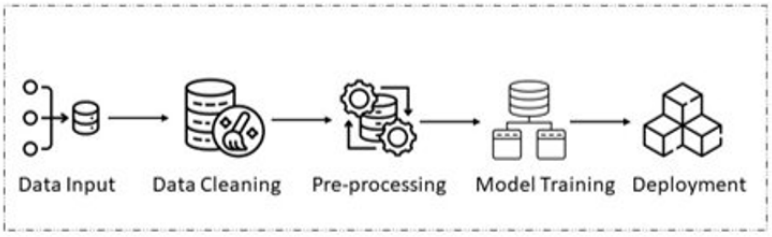

Figure 20: Data Processing Pipeline

4.1.2 Machine Learning Framework

The backbone of the machine learning efforts is built on Scikit-learn, a powerful and flexible Python library. Scikit-learn's comprehensive suite of tools for data pre-processing, model training, and evaluation makes it an ideal choice for this research. The important components within the machine learning framework are preprocessing tools, machine learning algorithms, and evaluation metrics. Preprocessing tools include handling of missing values, encoding categorical variables, and scaling features in a single line of code. The machine learning algorithms adopted range from logistic regression to decision trees, random forest, and gradient boosters. For the proper evaluation of a model, a set of metrics is used featuring accuracy, precision, recall, F1-score, and ROC-AUC.

Table 1: Metrics

Accuracy	Proportion of correctly predicted instances
Precision	Proportion of true positive predictions among all positive predictions
Recall	Proportion of true positive predictions among all actual positive instances
F1-Score	Harmonizes precision and recall into a single metric
ROC-AUC	Evaluates the models discrimination ability between classes

4.1.3 Evaluation Framework

In order to allow a holistic assessment of the models, a multiple-metrics-based evaluation framework was designed. This would provide an in-depth understanding of the performance of a model and the areas that potentially require improvement.

4.1.4 Logistic Regression:

Logistic regression is used to predict binary outcomes from modeling based on the probability of a categorical dependent variable, giving considerations to one or more predictor variables. Predicted probabilities output would thus be binary predictions computed based on some thresholdusually 0.5.

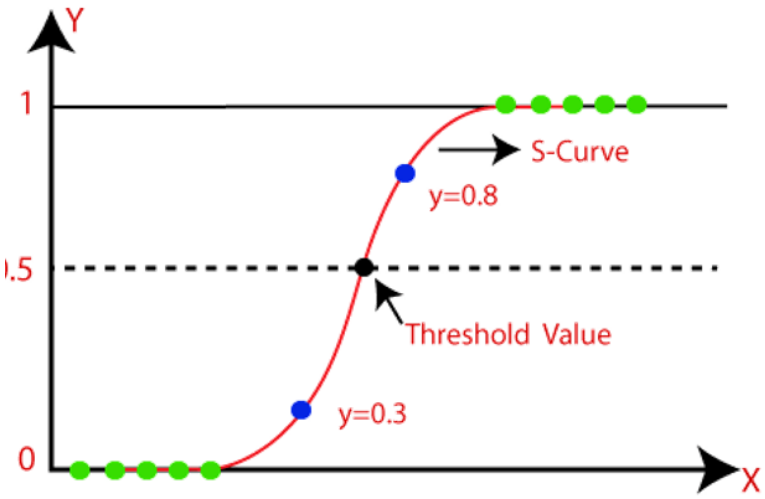

Figure 21: Logistic Regression

This study will use this model to predict whether a customer will respond to a direct marketing campaign based on their demographic and interaction data.

4.1.5 Decision Trees:

Decision trees imitate the decision-making process and the probable consequences involved. The tree structure is such that nodes of the tree represent decisions; the branches represent the outcomes, and leaf nodes give the final decisions. All this makes it an interpretable model.

Will be utilized to pinpoint significant factors influencing customer responses to marketing campaigns and make data-driven predictions.

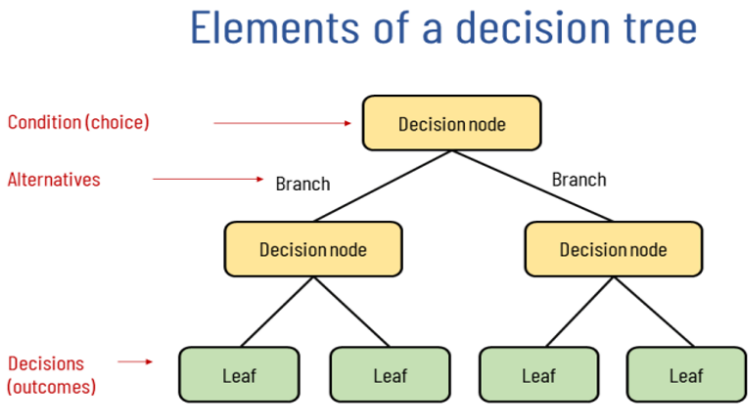

Figure 22: Decision Tree

4.1.6 Random Forest:

Random forest is an ensemble learning method whereby several decision trees are created and their primitives combined to improve accuracy and strength. In this case, each tree is created based on a random subset of data and features, thereby improving strength by minimizing overfitting.

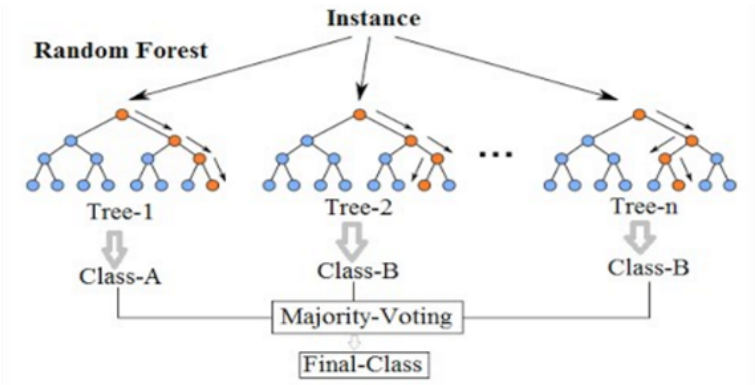

Figure 23: Random Forest

Provides more reliable and precise predictions for customer responses by leveraging the collective intelligence of multiple decision trees.

4.1.7 Gradient Boosting:

Gradient Boosting refines predictions through the use of a sequential ensemble learning technique where every new model looks at the mistakes made by the previous models. This iterative process, which gives much importance to the reduction of the loss function, has driven very accurate predictions.

Applied to boost prediction accuracy for customer responses by iteratively enhancing model performance through multiple stages.

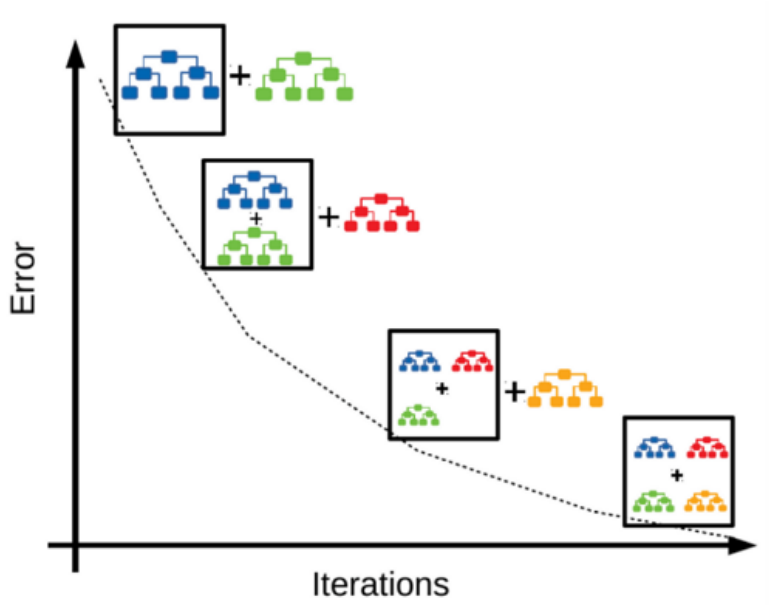

Figure 24: Gradient Boosting

5 Implementation

5.1 Trains Test Split

Data splitting is an important task in modeling; the dataset will be divided into a training subset and a test subset. Further in this example, independent variables (X) are derived by removing the dependent variable (y) by dropping the column 'y'. The dataset is then split into X-train, X-test, y-train, and y-test using an 80-20 ratio, where 80 % of the data is used for training the model and 20 % for evaluating its performance. A random state of 42 is set to ensure the reproducibility of the results. This method helps in assessing the model's generalizability to new, unseen data.

```
# Prepare the data
X = train_df.drop('y', axis=1)
y = train_df['y']

# Split the data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Figure 25: Test Train Split

5.2 Feauture Scalling

Feature scaling is an important preprocessing step in machine learning, particularly with algorithms sensitive to the scale of the input data. Now, here it is applied to standardize features using Standard Scaler on dataset columns. It removes the mean and scales to unit variance; that is, a distribution with a mean of zero and a standard deviation of one is returned.

```
# Scale the features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Figure 26: Scaling

5.3 Training Models

In the model training stage, different algorithms of machine learning have been utilized in predicting this target variable. Four models were selected: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. All of these models were trained on the processed and scaled data. It involves passing the training data to the models so that they learn the relationship between the features and the target variable. All models trained successfully; this was indicated by the output messages which were confirming that each model was trained. This step is extremely critical because it's getting the models ready to make predictions on new, unseen data by understanding patterns from the training data.

```
# Train models
models = {
    'Logistic Regression': LogisticRegression(),
    'Decision Tree': DecisionTreeClassifier(),
    'Random Forest': RandomForestClassifier(),
    'Gradient Boosting': GradientBoostingClassifier()
}

for name, model in models.items():
    model.fit(X_train_scaled, y_train)
    print(f"{name} trained.")
```

```
Logistic Regression trained.
Decision Tree trained.
Random Forest trained.
Gradient Boosting trained.
```

Figure 27: Training Models

6 Evaluation

To assess the effectiveness of each model, the evaluation process involved comparing the predicted results with the actual outcomes from the test dataset. This is a step that involves again several key metrics. One of them is accuracy, which refers to the proportion of total correct predictions made by the model. It's a simple measure that tells overall abilitywhat the model is capable of in terms of prediction. Other important metrics are precision and recall. The former is the number of correct positive predictions out of all positive predictions formed, while the latter is a measure of a model's ability to identify all actual positive cases. This set of metrics is applied to infer information about specificity and sensitivity, respectively. F1-score provides a balance between precision and recall. It

proves useful in cases when the class distribution may not be even. It gives one single measure of accuracy for the model, considering both false positives and false negatives. Another important metric in evaluation is the ROC AUC. It estimates the ability of the model to differentiate between positive and negative cases, thus giving in one number the summary of model performance at various threshold settings. The higher the ROC AUC, the better the performing model. The further assessment was carried out through a confusion matrix that gives the visual representation of the counts of true positives, true negatives, false positives, and false negatives. This matrix is key to the identification of specific types of errors the model tends to make, thus providing more granular information about its strengths and weaknesses. Each model was subjected to a very rigorous evaluation process so that the performance of each could be determined, leading to the selection of the best model according to its ability to make accurate predictions on the test data according to various metrics.

6.1 Logistics Regression

The Logistic Regression model provided baseline performance of 88.80 % in accuracy and 87.00 % in ROC AUC. The precision for this was 59.85 %, with the recall remaining significantly low at 21.72 %, which means a large number of actual positive cases were missed by this model. Again, there is an imbalance reflected in an F1-score of 31.88 %, which shows how much more conservative the model is to make positive predictions. While it offers simplicity and interpretability, the predictive power was limited compared with complex models for logistic regression, making this method less suitable for application at hand.

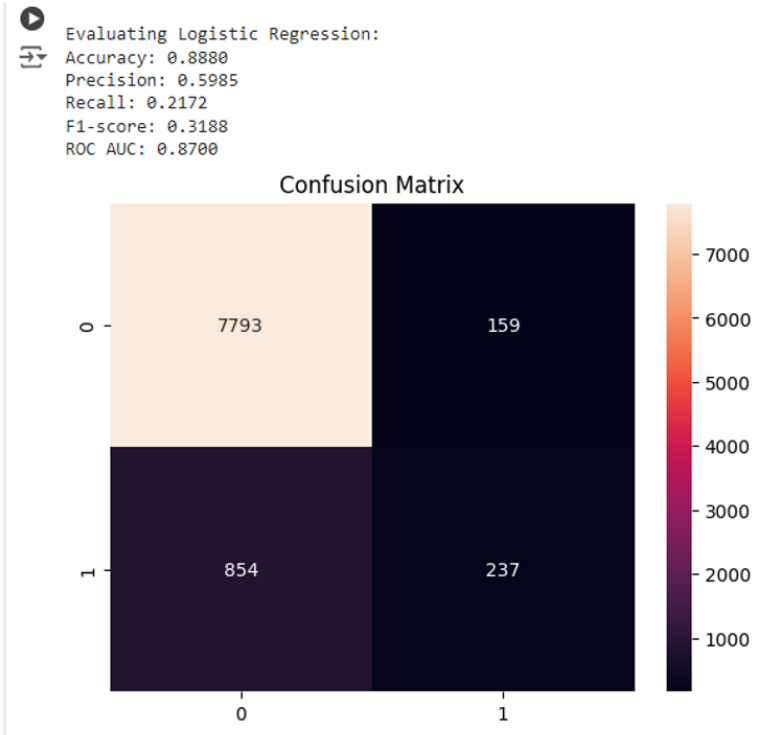

Figure 28: Logistic Regression Model Evaluation

6.2 Decision Tree

This model had an accuracy of 87.16 % and a ROC AUC of 70.12 %, which is lower than that of the Gradient Boosting and Random Forest models. The precision was 46.85 %, with a recall equating to 47.66 %. It shows the model is balanced but less ranked with its ability to detect or correctly predict positive incidences. This balance was reflected in an F1-score of 47.25 %, indicating a lower overall performance as compared to the ensemble methods. Although aggressive compared to the others, especially the ensemble methods, the Decision Tree model is nevertheless useful in interpreting the importance of different features.

6.3 Random Forest

The Random Forest model also exhibited strong performance, with an accuracy of 90.21 % and an ROC AUC of 92.33 %, slightly outperforming Gradient Boosting in terms of ROC AUC. Its precision was 64.63 %, similar to Gradient Boosting, indicating a high rate of correct positive predictions. The recall was slightly higher at 41.70 %, suggesting better identification of actual positive cases. The F1-score of 50.70 % shows a balanced performance between precision and recall. Random Forest's ensemble nature contributed to its robustness and reliability in predictions, making it another strong candidate for the final model.

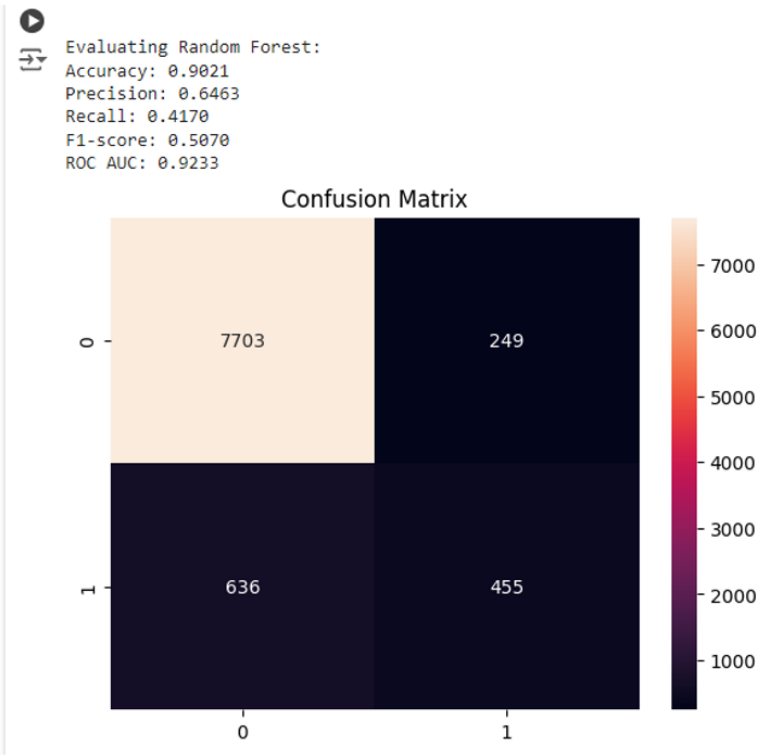

Figure 29: Random Forest Model Evaluation

6.4 Gradient Boosting

Gradient Boosting did well with an accuracy of 90.19 % and a really nice ROC AUC of 92.03 %, thus indicating the existence of strong classifiers for separating positive from negative cases. Model precision was 65.22 %, hence indicating a good proportion of correct positive predictions out of all positive predictions assigned by the model. The recall, however, was relatively lower at 40.05 %, indicating that as good as it was in predicting positive cases, it missed a substantial number of actual positives; this is reflected in its F1-score of 49.63 %. Overall, Gradient Boosting returned quite an impressive display of the ability to make accurate predictions, making this a reliable choice for application to the problem at hand.

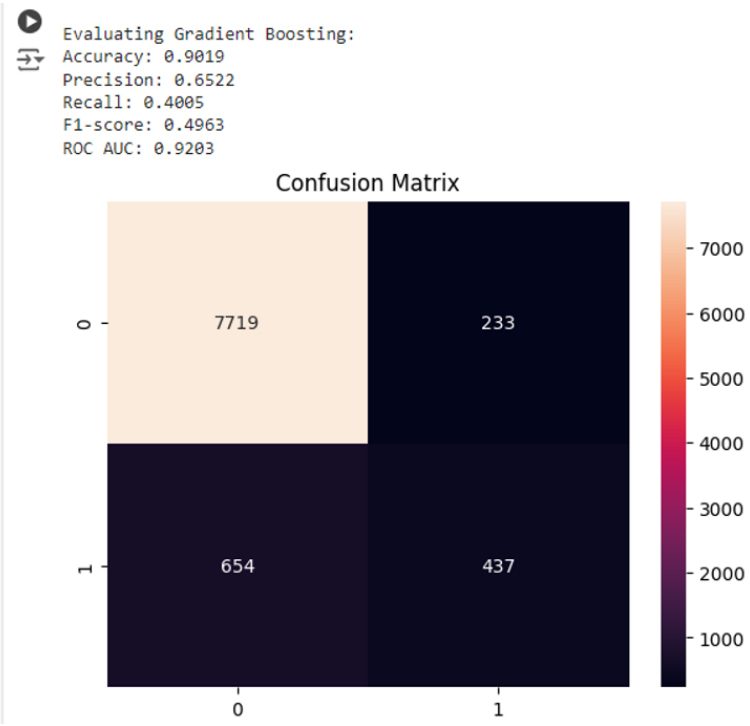

Figure 30: Gradient Boosting Model Evaluation

6.5 Discussion

Comparing these models shows the trade-offs among different performance metrics. Gradient Boosting and Random Forest are the most resilient models for this application, driven by their high accuracy and ROC AUC scores. Their lower recall values indicate that further tuning or probably the introduction of additional data features may be necessary to positively identify more cases. Later on, decision trees provide poor performance overall but offer interpretability that can be very useful in guiding further feature engineering and model refinement. Logistic regression gives a good baseline and insight into the linear relationships within data. For practical applications in the banking sector, the choice of the model would depend on which factors are most important in the marketing strategy to be implemented. Gradient Boosting or Random Forest if maximum accuracy in positive predictions is the most important objective. On the other hand, it is in Decision

Trees that one gets interpretability and insight into decision pathways if those things are critical. For a fast, interpretable, relatively good starting point, Logistic Regression still has its place.

7 Conclusion and Future Work

This paper set out to study the application of some machine learning algorithmsthe logistic regression, decision trees, random forest, and gradient boostingin evaluating how these direct marketing strategies impact customer behaviors in banking sectors. Some of the key objectives include establishing the effectiveness of these models with respect to prediction, explaining their role in demographic factors, and identifying key features that drive the success rates of campaigns. Through a comprehensive mixed-method approach combining quantitative analysis and machine learning techniques, the study successfully addressed the research question and met the stated objectives. This research uses a granular, public dataset obtained from Kaggle, entailing extensive data pre-processing, feature scaling, model training, and finally, the evaluation phase. Among the major findings, it was evident that Gradient Boosting and Random Forest models were ranked as the top two overall best-performing models, with an accuracy of approximately 90.2 % and above 92 % in terms of their respective ROC AUC scores. Models did an exemplary job of differentiating between positive versus negative responses and, therefore, were applicable in predicting customer behavior towards direct marketing campaigns. The Decision Tree model, despite its lower overall performance, provided valuable insights into feature importance and decision-making processes. Logistic Regression served as a robust baseline, offering simplicity and interoperability, though its predictive power was limited compared to more complex models. This research further has very huge implications for the banking sector in applying actionable insights to optimize their direct marketing strategies. Graduate Boosting and Random Forest performed quite well, indicating that advanced machine learning could bring very high accuracy to targeting strategies, hence improving customer engagement and success rates of campaigns. Some limitations were also found in the study, such as lower recall values into the best models, still not allowing the identifying of all the potential positive responders. The findings of the current study allow for the following future research and commercialization routes. To increase model recall, future works might aim to enhance ensemble recalls through ensemble stacking, hyperparameter tuning, or the addition of more informative features. This will help recognize more actual positive cases in order to increase the efficacy of these models. Addressing data imbalance is another important area. In general, having a more balanced training dataset can improve model performance, especially for the identification of instances of the minority class. One interesting technique for this work is Synthetic Minority Over-sampling Technique (SMOTE), an oversampling method that balancing a dataset to obtain more reliable and fair predictions. Real-time data integration and model updating can further strengthen the practical applicability of models under dynamic marketing environments. This will allow for more timely and better predictions based on how customers behave in the most recent interaction, hence making marketing strategies relevant and productive. Further ahead have to be the ethical and regulatory considerations in future research. Not only for consumer trust, but also far from legal implications, it is very important to make sure there is conformance to data privacy regulations addressing the possible biases within the models. This becomes doubly important in the banking sector due to

the sensitive customer information involved. Already, several very wide possibilities open up for the commercialization of the models developed in this research work. The banks and financial institutions can embed these developed models within their CRM system to further fine-tune marketing efforts for optimum customer acquisition and retention. Such user-friendly software solutions, based on these models, would also be very useful to the marketing professionals across banks to make data-driven decisions and enhance the effectiveness of their campaigns.

References

Acquisti, A., Brandimarte, L. and Loewenstein, G. (2020). Secrets and likes: The drive for privacy and the difficulty of achieving it in the digital age, *Journal of Consumer Psychology* **30**(4): 736–758.

Crespo, I. and Govindarajan, A. (2018). The analytics-enabled collections model. Accessed: Aug 2024.
URL: *https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/the-analytics-enabled-collections-model*

Cui, G., Wong, M. L. and Lui, H.-K. (2006). Machine learning for direct marketing response models: Bayesian networks with evolutionary programming, *Management Science* **52**: 597–612.

Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., Jain, V., Karjaluoto, H., Kefi, H., Krishen, A. S., Kumar, V., Rahman, M. M., Raman, R., Rauschnabel, P. A., Rowley, J., Salo, J., Tran, G. A. and Wang, Y. (2021). Setting the future of digital and social media marketing research: Perspectives and research propositions, *International Journal of Information Management* **59**: 102168.

Ellahi, A. Q. U. A., Rehman, H. M., Hossain, M. B., Ills, C. B. and Tanweer, A. (2024). The impact of omnichannel integration towards customer interest in alternatives: Retailer uncertainty and web rooming in retailing, *Cogent Business & Management* **11**(1): 2316931.

Gefen, D. and Straub, D. (2000). The relative importance of perceived ease of use in is adoption: A study of e-commerce adoption, *J. AIS* **1**: 0–.

Kumar, V. and Shah, D. (2004). Building and sustaining profitable customer loyalty for the 21st century, *Journal of Retailing* **80**(4): 317–329.

Mero, J. and Taiminen, H. (2016). Harnessing marketing automation for b2b content marketing, *Industrial Marketing Management* **54**: 164–175.

Nayak, P. and Siddiqui, I. N. (2024). Study on significance of direct marketing in modern era, *IRE Journals* .

Sahni, N., Wheeler, S. and Chintagunta, P. (2018). Personalization in email marketing: The role of noninformative advertising content, *Marketing Science* **37**.

Tekouabou, S. C. K., Gherghina, . C., Toulni, H., Neves Mata, P., Mata, M. N. and Martins, J. M. (2022). A machine learning framework towards bank telemarketing prediction, *Journal of Risk and Financial Management* **15**(6).

Verhoef, P. C., Reinartz, W. J. and Krafft, M. (2010). Customer engagement as a new perspective in customer management, *Journal of Service Research* **13**(3): 247–252.