Predictive Modelling of Customer Response to Marketing Campaigns

MIGLENA PAVLOVA, University of Twente, The Netherlands
SUPERVISOR: DR. MOHAMMED ELHAJJ, University of Twente, The Netherlands

In today's data-driven marketing landscape, accurately predicting customer responses to marketing campaigns is critical for optimizing engagement and return on investment (ROI). This study utilizes a Decision Tree (DT) model to identify key factors influencing customer behaviour. Initially, the model achieved a high accuracy of 87.3% but struggled with precision and recall due to class imbalance. By applying a resampling technique, the model's performance improved significantly, with a recall increase from 44% to 83.1% and an F1-score improvement from 49% to 74.2%. Key influential features identified include how recently a customer made a purchase, the number of days they have been a customer, and the number of previous campaigns they responded to. The study highlights the DT model's interpretability, making it a practical tool for marketing professionals to improve campaign effectiveness and customer targeting.

1 Introduction

In today's highly competitive and data-driven business environment, understanding customer behaviour is crucial for effective marketing strategies [5]. Predicting how customers will respond to marketing campaigns not only improves the effectiveness of these efforts but also significantly increases ROI [12]. As companies gather vast amounts of data from various customer touchpoints, the challenge is transforming this data into actionable insights for targeted and personalized marketing. Predictive modelling, particularly using DT models, offers a promising solution by using historical data to forecast future customer behaviour [21]. Marketing strategies can be divided into mass marketing and direct marketing. Mass marketing uses widespread media platforms like television, and radio, to reach existing and potential customers. In contrast, direct marketing focuses on contacting specific clients directly, often proving more cost-effective and resource-efficient. Understanding the effectiveness of these strategies requires a deep understanding of customer behaviour. A paper by Raorane and Kulkarni [15] suggests that studying consumer psychology, mindset, behaviour, and motivation allows companies to refine their marketing strategies. Therefore, collecting and analyzing customer data is essential for businesses. Customer Relationship Management (CRM) allows for the automatic collection of this data, which includes demographics, purchase history, and interactions with the company. This field revolves around identifying, establishing, and sustaining long-term relationships with clients. Utilizing CRM data is crucial for informed marketing decisions [17]. Traditionally, customer behaviour prediction relied on managers' intuition and experience, with decisions based on general trends rather than analytical support. However, the rise of Machine Learning (ML) has transformed predictive analytics, leading to the development of more advanced models. Tree-based ML classifiers, such as DT and Random Forest (RF) models, are known for their high accuracy and interpretability. DT models are particularly favored for their ease of understanding and visualization [21], as they create a tree-like structure of decisions based on input features. RFs, on the other hand, are an ensemble method that improves the predictive power of DT by aggregating the results of multiple trees, improving generalization, and reducing overfitting, meaning preventing the model to become too finely tuned to the training data[11]. The question of whether RF outperforms DT in predicting customer marketing responses is multi-dimensional. However, an advantage of both models is their capability to assess feature importance. This analysis can identify the most influential factors in predicting customer responses to promotions and marketing campaigns. By determining which customer attributes have the greatest impact, businesses can tailor their strategies and allocate resources more effectively. While the RF method is more robust against noisy data compared to just using a single DT[10, 11], when the dataset is relatively small, and the interpretability of the model is crucial, DT is the better choice. They are easier to interpret than RF, due to their representation of simple decision rules, making it easier to understand how each feature contributes to the model's predictions[14]. Despite the potential of predictive modelling, businesses often face significant difficulties in accurately predicting customer responses to marketing campaigns. The complexity of customer behaviour, influenced by many factors such as demographics and past interactions, makes it difficult to develop reliable models. Traditional approaches tend to overlook these complexities, leading to generalized and less effective marketing strategies [17]. This study seeks to address this gap by focusing on the interpretability and explainability of the predictive model, by utilizing the DT algorithm [21]. The primary objective is to identify and understand the most influential demographic factors, such as age, income, marital status, and education level, as well as examining the impact of past interactions with the company, including previous purchases and engagement with earlier campaigns. The research aims to achieve this through the following questions:

RQ1 What are the challenges and limitations presented in the literature regarding predicting customer marketing responses?
RQ2 How effective is the DT model at predicting customer response to marketing campaigns?

RQ3 What are the key factors influencing customer response to marketing campaigns as identified by the DT model?
- Which demographic factors are most influential in predicting customer response to marketing campaigns according to the DT model?
- How do past interactions with the company affect future responses according to the DT model?

TScIT 41, July 8, 2024, Enschede, The Netherlands

The rest of the paper is organized as follows: Section 2 reviews related work, discussing existing literature and the performance of DTs in predictive analytics. The methodology and practical implementation are detailed in Section 3, while Section 4 presents the research findings. Section 5 discusses the results and their implications, and Section 6 concludes with a summary of key findings, limitations, and directions for future research.

2 Related Work

This section reviews key studies that investigate the application of various predictive models in direct marketing, highlighting their methodologies and results. A paper by K. Wisaeng [20] compares different classification techniques in bank direct marketing, using a UCI repository data set with 16 attributes and 45,211 instances. The study examines two decision tree methods, J48-graft and LAD tree, and two machine learning approaches, Radial Basis Function Network (RBFN) and Support Vector Machine (SVM). The results indicate that among the algorithms tested, the SVM outperformed others, achieving the highest accuracy of 86.95%. In contrast, the RBFN showed the least effective performance with an accuracy of 74.34% [20]. Research by Sérgio Moro et al. [18] applied Logistic Regression (LR), Neural Networks (NN), DT, and SVM on a dataset sourced from a Portuguese bank, including 22 selected features. Their study highlighted the performance of the NN in predicting customer behaviour. To optimize marketing strategies the study provided practical insights, revealing that targeting the top half of customers classified as more likely to respond positively could lead to successful outcomes in 79% of cases. Suggesting that a selective approach to customer engagement can potentially reduce costs while maximizing campaign efficiency [18]. Another paper by Sérgio Moro [19] applied different data mining algorithms such as Naive Bayes (NB), DT, and SVM. The findings indicated that SVM has the highest prediction performance, with NB and DT following. The call duration was found to be the most significant feature, followed by the month of contact. DTs have emerged as a fundamental tool in predictive analytics for marketing. By offering a transparent and interpretive model, DTs provide marketers with valuable insights into the influence of different customer attributes on marketing outcomes [14, 21]. Many studies underscore the potential of DT models as a powerful tool for businesses seeking to optimize their marketing strategies and maximize customer engagement. The study conducted by authors in [22] demonstrated the effectiveness of DT models in forecasting customer responses to direct marketing. The researchers utilized DT models to analyze historical data from various marketing campaigns, to predict future customer behaviour. The DT models were trained on a range of features, including demographic information and past interactions with the company. Among the customers who were predicted not to respond to direct marketing, the model's accuracy was 87.23%. This means that in 87.23% of cases, the customers who were predicted not to respond indeed did not respond [22]. On the other hand, among the customers who were predicted to respond to direct marketing, the model's accuracy was 66.34%. This indicates that in 66.34% of cases, the customers who were predicted to respond did indeed respond [22]. Another study conducted on customer churn analysis for live stream e-commerce platforms used DT, Naive Bayes, and K-nearest neighbour algorithms to classify customers into churners and non-churners groups. The DT algorithm outperformed the other models with an accuracy of 93.6%. A similar research by Usman-Hamza et al. [3] highlighted the effectiveness of tree-based classifiers in customer churn prediction, outperforming other forms of classifiers in most cases. The RF ensemble arguably increases the generalization accuracy of Decision Tree-based classifiers without trading away accuracy on training data[7]. As per Chaubey et al. [5], this suggestion translates into the problem of customer purchasing behaviour prediction. Their paper suggests that when comparing the accuracy of models for churn prediction, RF has been found to perform better than the DT model, suggesting its potential to improve accuracy in specific predictive tasks. However, a study by Apampa [1] examines to what extent the use of RF ensemble improves the performance of the DT classification algorithm for the bank customer marketing response prediction. In this study it was concluded that the use of RF ensemble does not improve or improve the performance of the DT algorithm, suggesting that RF might not consistently improve DT's performance, particularly in contexts such as predicting bank customer responses to marketing. Additionally, interpreting the resulting RF model remains a challenging task, as even machine learning experts struggle to precisely analyze and uncover the detailed predictive structure [11]. Making the DT algorithm the most appropriate when interpretability is favored. Previous research has primarily focused on applying various ML techniques and comparing their efficiency. However, there has been a notable gap regarding the treatment of complexity issues. Decision-makers with limited technical backgrounds often struggle to grasp the complex relationships between attributes in traditional ML models. Therefore, this study aims to address this gap by applying a straightforward DT model that is easy to interpret.

3 Proposed Solution

This research follows a six-stage methodology that is designed to be straightforward and interpretable for individuals with a moderate understanding of data mining. The whole procedure is shown in Fig. 1:

3.1 Hardware and Software Configuration

The hardware and software configuration for this research ensures the reproducibility of the experiment. In Table 1 are listed the specific components and tools used. To ensure transparency and ac-

Table 1. Hardware and Software Configurations

	Component	Configuration
Hardware	Processor	Intel Core i7-10510U
	RAM	16 GB
	Storage	952 GB
	OS	Windows 11 Pro
Software	Language	Python
	Libraries	pandas, seaborn, matplotlib, scikit-learn
	Environment	Jupyter Notebook

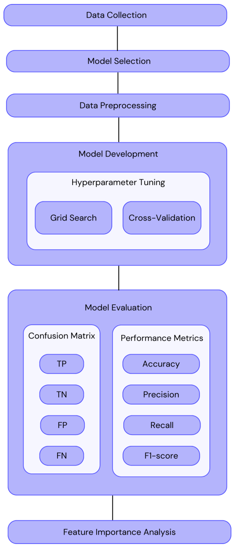

Fig. 1. Sketch of the proposed solution

cessibility in the research process, the source code for this research is made publicly available in a GitHub repository[1].

3.2 Data Collection

The dataset used in this study was obtained from the online platform Kaggel and it belongs to the Brazilian food ordering platform iFood [9]. As presented in Table 2 it includes various demographic data, such as age, income, marital status, and education level. As well as customer interaction data, such as previous purchases and previous marketing responses. The total number of instances is 2206. The dataset consists of 39 attributes, with the target variable 'Response' being a binary indicator. This target variable has two classes, "yes," indicating that the customer responded positively to a marketing campaign, and "no," indicating that the customer responded negatively. Notably, the dataset contains no categorical data. All attributes are either numerical or binary indicators. This structure eliminates the need for encoding categorical variables. However, the target class is imbalanced, highlighting the need for resampling techniques or adjusting class weights to address this issue.

Table 2. Data Dictionary

Demographic	Income	Kidhome	Age
	Teenhome	Customer_Days	marital_Together
	marital_Single	marital_Divorced	marital_Widow
	education_PhD	education_Master	education_Graduation
	education_Basic	education_2n Cycle	
Customer Interaction	MntWines	MntFruits	MntGoldProds
	MntMeatProducts	MntFishProducts	MntSweetProducts
	NumStorePurchases	NumCatalogPurchases	NumWebVisitsMonth
	NumDealsPurchases	NumWebPurchases	Recency
	Z_CostContact	Z_Revenue	MntTotal
	MntRegularProds	Complain	Response
	AcceptedCmp1	AcceptedCmp2	AcceptedCmp3
	AcceptedCmp4	AcceptedCmp5	AcceptedCmpOverall

3.3 Model Selection

DT is a supervised ML method, aiming to establish a relationship between input features and the target variable for accurate predictions [21]. Structurally, decision trees resemble a tree where each node signifies a decision based on an attribute, each branch corresponds to an outcome of that decision, and each leaf node represents a target class label. The classification process involves tracing a path from the root node, the primary attribute, to a leaf node [21]. This intuitive method uses an "if-else" logic, making it straightforward to understand and interpret [14, 21]. This is especially useful in marketing, where decisions are often made by individuals with limited technical knowledge, making decision trees an appropriate choice.

3.4 Data Preprocessing

It is observed that the dataset is significantly imbalanced, with a considerably higher number of negative responses ("no") compared to positive responses ("yes"). This class imbalance poses a notable challenge because the model tends to predict the majority class more frequently. While this may lead to high overall accuracy, it results in poor identification of the minority class, which is crucial for the campaign's success [6]. To address the issue of class imbalance, a technique called resampling is implemented. Resampling involves adjusting the dataset to balance the class distribution, ensuring that the model has an equal representation of both classes during training. This can be achieved through various methods such as oversampling the minority class or undersampling the majority class [6]. In this study, the undersampling technique is applied. This approach involves decreasing the number of instances in the majority class (negative responses) to match the number of instances in the minority class (positive responses), resulting in a more balanced dataset that allows the model to learn the characteristics of both classes more effectively. In addition to resampling, another effective approach that is used is adjusting the class weights [6]. By assigning higher weights to the minority class, the model further improves its sensitivity towards positive responses [13].

3.5 Model Development

In the next part of the research, the DT model is developed using a structured and methodical approach. Initially, the dataset is prepared, by partitioning the features into predictors (X) and the target variable (y). This method ensures that the model learns to predict the target variable based on the features [4]. The predictors consist of everything except the 'Response' column, which serves as the target variable. The dataset is divided into training and testing sets with an 80-20 ratio, meaning 80% of the data is used to train the model, and the remaining 20% is used to test it. This partitioning allows for the evaluation of the model's performance on unseen data, which simulates real-world scenarios where the model will encounter new data. This way the model generalizes well and is not overfitted to the training data [4]. Additionally, a random state of 42 is specified to guarantee reproducibility of the results, ensuring that the random processes involved in data splitting will produce the same results every time the code is run.

[1] https://github.com/megi2002/Predictive-Modelling-of-Customer-Response-to-Marketing-Campaigns

3.5.1 Hyperparameter-tuning After resampling, a grid search method, combined with cross-validation, is applied to explore different combinations of hyperparameters. One of the key ones is the 'criterion,' which determines the function used to measure the quality of a split. The options for the 'criterion' parameter include Gini impurity and entropy [4]. Gini impurity is defined in Equation 1:

$$G = 1 - \sum_{i=1}^{n} p_i^2 \tag{1}$$

Where p_i^2 represents the proportion of instances belonging to class i in the dataset. Gini impurity measures the probability of incorrectly classifying a randomly chosen element. An impurity of 0 indicates that all elements in a node belong to a single class, representing perfect purity. In practical terms, a lower Gini impurity means that the DT is better at creating homogeneous groups of customers, which can lead to more accurate predictions.[4]. Entropy is defined in Equation 2:

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{2}$$

It measures the amount of disorder within a set of classes. When the entropy is 0, it means there is no disorder, and all customers within a node share the same classification. Higher entropy values indicate greater disorder and less purity. The criterion of entropy often leads to more balanced splits compared to Gini impurity, as it creates splits that increase the information gain, making it a preferred choice when the goal is to achieve higher accuracy and a more informative model [4]. Another important hyperparameter is the 'splitter'. The 'splitter' can be set to 'best' or 'random.' The 'best' option selects the optimal split among all features, aiming to maximize information gain or minimize Gini impurity. On the other hand, the 'random' option selects a random feature and then finds the best split within that feature. Parameter 'best' might result in a more accurate but computationally intensive model, whereas 'random' can lead to faster training times and increased generalization [4]. The 'max_depth' parameter controls the maximum depth of the tree. It ranges from no limit, allowing the tree to expand until all leaves are pure, to a specified maximum depth, such as 5, 10, 15, or 20. A shallower tree generalizes better on unseen data, whereas a deeper tree can capture more details from the training data but risks overfitting [4]. The 'min_samples_split' parameter specifies the minimum number of samples required to split an internal node. It ranges from 2 to 15. A higher value prevents the model from learning too much from the noise in the training data, thus improving its generalization capability [4]. Finally, the 'min_samples_leaf' parameter indicates the minimum number of samples required to be at a leaf node. It ranges from 1 to 6. A higher value can lead to a more generalized model, whereas a lower value might allow the tree to capture more patterns [4]. By conducting an exhaustive grid search across these parameters, the model is evaluated through cross-validation for each combination. Meaning the model is trained and evaluated on different subsets of the training data to ensure that the hyperparameters are not overfitted to a particular subset. The cross-validation divides the training data into five parts, training the model on four parts and validating it on the fifth, rotating this process to cover all combinations [16]. The best combination of hyperparameters is identified based on the average performance across these folds [16]. The best estimator from the grid search is then selected as the final model (best_clf) for further evaluation.

3.6 Model Evaluation

Evaluating the performance of the predictive model is crucial in understanding how well it generalizes to new, unseen data. In this research, several key metrics are utilized to assess the effectiveness of the DT model in predicting customer responses to marketing campaigns. These metrics include accuracy, precision, recall, F1 score, and the confusion matrix

3.6.1 Confusion Matrix To gain a comprehensive understanding of a model's effectiveness in imbalanced scenarios, the use of a confusion matrix is essential. It summarizes the prediction results, showing the count of correct and incorrect predictions broken down by each class. The matrix is structured in Table 3: True Positives (TP) refer to the number of instances where the model correctly predicts a customer will respond positively to a campaign, aligning with actual positive responses. True Negatives (TN) denote cases where the model accurately identifies customers who will not respond, matching the actual negative responses. False Positives (FP), often termed "false alarms," occur when the model incorrectly predicts a positive response from customers who, in reality, do not respond to the campaign. Conversely, False Negatives (FN) happen when the model fails to predict a positive response from customers who indeed respond [8].

Table 3. Confusion Matrix

Predicted \Actual	Positive (+)	Negative (-)
Positive (+)	TP	FP
Negative (-)	FN	TN

3.6.2 Accuracy Accuracy is a measure of the overall correctness of the model [8], representing the proportion of correctly predicted instances out of the total instances, as shown in Equation 3.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

For this model, the accuracy indicates how well it can correctly classify both positive and negative responses. In the case of the imbalanced dataset in this study, high accuracy can be achieved by simply predicting the majority class most of the time. However, this high accuracy is deceptive because the model fails to identify the customers who actually respond, making it ineffective for practical purposes. The limitations of accuracy in the context of imbalanced datasets highlight the importance of alternative metrics such as precision, recall, and the F1 score [6].

3.6.3 Precision Also known as positive predictive value. As defined in Equation 4 precision measures the accuracy of positive predictions [8].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

In this study, precision indicates the proportion of customers who are predicted to respond positively and indeed did respond positively.

3.6.4 Recall As shown in Equation 5, recall measures the ability of the model to identify all actual positive instances [8].

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

In this study, recall indicates the proportion of actual positive responses that were correctly predicted by the model.

3.6.5 F1-Score The F1-score is the harmonic mean of precision and recall, providing a single metric that balances the two. It is particularly useful when there is an uneven class distribution [8]. The formula for the F1-score is shown in Equation 6:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

The F1-score ranges from 0 to 1, where 1 indicates perfect precision and recall, and 0 indicates the worst possible performance. This metric is beneficial when seeking a balance between precision and recall, especially in the presence of class imbalance [8].

3.7 Feature Importance Extraction

In this step, feature importance scores are extracted from the trained DT model and the top 10 features are identified and visualized. Feature importance is a metric that indicates the significance of each input variable in contributing to the prediction accuracy of the DT classifier.

3.8 Decision Rules Generation

In this step, the decision tree rules are generated in the form of *if.. else* statements. They allow for easy interpretation of the decision-making process, where one can understand how a particular prediction is made. The clarity of the DT rules enables stakeholders, who may not have a deep technical background, not only to pinpoint these influential factors accurately but also to utilize them effectively.

4 Results

In this section of the study, the best hyperparameters resulting from the grid search combined with cross-validation are presented before and after resampling is applied. Additionally, a comparative analysis of the model evaluation results before and after resampling is conducted. The analysis focuses on the confusion matrix and various performance metrics, including accuracy, precision, recall, and F1-score, to evaluate the model's effectiveness. Furthermore, the results include feature importance scores and the generated decision rules, which are extracted from the decision tree classification model after resampling. This approach is taken because the resampled dataset provides a more balanced and accurate representation of the underlying patterns, leading to more reliable and interpretable decision rules and feature importance scores.

4.1 Best Hyperparameters

4.1.1 Before Resampling The grid search combined with cross-validation identified the optimal hyperparameters to be the ones presented in Table 4 before resampling was applied.

Table 4. Parameter values before resampling

Parameter	Value
criterion	entropy
max_depth	5
min_samples_leaf	2
min_samples_split	2
splitter	random

These hyperparameters reflect a conservative approach to handling the significant class imbalance in the dataset. The criterion of entropy helps in maximizing information gain at each split. By limiting the maximum depth to 5, the model avoids overfitting to the majority class of negative responses, which dominates the dataset. The parameters for minimum samples per leaf and split ensure that each node has enough data to make reliable decisions, thus reducing the likelihood of splits based on noise or anomalies. The use of a random splitter adds an element of randomness to the decision-making process, which helps prevent the model from becoming overly complex and biased towards the majority class during training.

4.1.2 After Resampling: Following the application of undersampling to balance the class distribution, the grid search with cross-validation identified a different set of optimal hyperparameters, presented in Table 5.

Table 5. Parameter values after resampling

Parameter	Value
criterion	entropy
max_depth	None
min_samples_leaf	2
min_samples_split	2
splitter	best

The shift in hyperparameters post-undersampling indicates a significant change in the model's complexity and its approach to decision-making. With the maximum depth set to None, the model is allowed to grow without constraints until all leaves are pure or until they contain fewer samples than the minimum samples split threshold. This unrestricted growth enables the model to capture more detailed patterns in the balanced dataset. The switch to the best splitter means the model now selects the optimal split at each node, based on the entropy criterion, to maximize information gain, leading to more precise and effective splits that better separate the classes.

4.2 Confusion Matrix

4.2.1 Before Resampling The confusion matrix before resampling is presented in Table. 6 and it reveals that the model correctly identifies 27 true positives and 357 true negatives, while there were 21 false positives and 36 false negatives. This indicates that the model was successful in predicting customers who would respond positively to marketing campaigns in 27 instances and correctly identifying customers who would not respond in 357 instances. The high number of true negatives compared to true positives is attributed to the imbalance in the target class 'Response'. The model is

Table 6. Confusion matrix before resampling

Predicted \Actual	Positive (+)	Negative (-)
Positive (+)	TP = 27	FP = 21
Negative (-)	FN = 36	TN = 357

exposed to more instances of non-response during training, which makes it better at identifying non-responders (true negatives) but limits its capacity to detect responders (true positives). The model also produced 21 false positives (FP) representing instances where the model incorrectly predicted a positive response from customers who did not respond. Conversely, the 36 false negatives (FN) produced, indicate cases where the model failed to predict a positive response from customers who did respond positively. This means that the model occasionally mistakes non-responders for responders, potentially leading to unnecessary marketing efforts toward those unlikely to engage. More critically, the higher number of false negatives signifies that the model misses many potential customers who would have responded positively, ultimately resulting in missed opportunities for engagement.

4.2.2 After Resampling: The confusion matrix after resampling is presented in Table. 7 and it reveals that the model correctly identifies 49 true positives and 51 true negatives, while there were 24 false positives and 10 false negatives. Post-resampling, the model's ability to correctly identify positive responses improves significantly, evidenced by the increase in true positives from 27 to 49. This improvement is primarily due to the undersampling technique, which balances the class distribution by reducing the number of majority class instances, thereby allowing the model to learn more effectively from the minority class. However, this adjustment also leads to a slight increase in false positives (from 21 to 24) and a decrease in true negatives (from 357 to 51), as the model now encounters fewer non-responders during training. This trade-off is typical when addressing class imbalance; while the model becomes better at identifying the minority class, it may lose some accuracy in predicting the majority class. Despite this, the drop in false negatives from 36 to 10 is significant, indicating a more balanced and effective model that is better equipped to predict both responders and non-responders.

Table 7. Confusion matrix after resampling

Predicted \Actual	Positive (+)	Negative (-)
Positive (+)	TP = 49	FP = 24
Negative (-)	FN = 10	TN = 51

The breakdown made in the confusion matrix is crucial for calculating the performance metrics.

4.3 Model Evaluation

4.3.1 Before Resampling The performance of the model before resampling is presented in Fig. 2

Despite the high accuracy of 87.3%, the precision, recall, and F1-score are relatively low. Accuracy alone can be misleading in cases of imbalanced datasets, where one class significantly outweighs the other. Here, the high accuracy mainly reflects the model's ability to correctly identify non-responders, but it does not adequately capture the performance in predicting the responders.

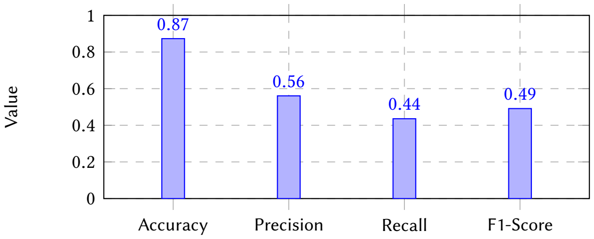

Fig. 2. Evaluation metrics before resampling

The precision, which is calculated to be 56%, measures the proportion of true positive predictions among all positive predictions. This means that out of all the instances, that the model predicted as responders, only 56% were actually correct. The recall, calculated to be 44%, measures the proportion of actual positive instances that were correctly identified by the model. This means that the model only identified 44% of the actual responders correctly. The low F1-score reflects the overall inefficiency of the model in handling the imbalanced dataset, as it struggles to achieve a good trade-off between precision and recall. While the model appears to perform well based on accuracy alone, the low precision, recall, and F1-score reveal its limitations in predicting the minority class effectively.

4.3.2 After Resampling: The performance of the model after resampling is presented in Fig. 3 Post-resampling, the model's performance improved significantly. The accuracy dropped to 74.6%, which is expected as the model now faces a more balanced dataset, making predictions more challenging. However, this decrease in accuracy is not necessarily a negative outcome. The balanced dataset has allowed for improvements in other critical metrics. The precision increased to 67.1%, indicating that the model is now better at correctly identifying true responders, reducing the number of false positives where non-responders are incorrectly predicted as responders. The recall increased to 83.1%, demonstrating a substantial improvement in capturing most of the true positive cases, thereby reducing the number of false negatives where actual responders are missed. Finally, the F1-score improved to 74.2%, providing a balanced measure of the model's precision and recall. The significant improvement in the evaluation metrics indicates that the model is now well-suited to identify both responders and non-responders accurately, making it more effective for practical applications in marketing campaigns.

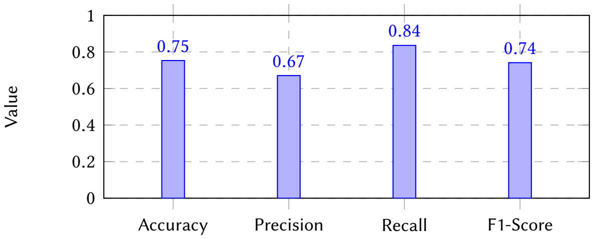

Fig. 3. Evaluation metrics after resampling

4.4 Feature Importance Scores

The top 10 most influential features are presented in Fig. 4. Demographic factors such as age and income are reported to play a crucial role in customer behaviour. Past customer interactions with the company, indicated by variables like Recency (days since last purchase), Customer_Days (days since customer registration), and AcceptedCmpOverall (number of accepted campaigns), are significantly influential to customer response. Additionally, product-specific purchases such as MntGoldProds (spending on gold products) and MntMeatProducts (spending on meat products), along with purchase channels including NumCatalogPurchases (number of catalog purchases), NumStorePurchases (number of store purchases), and NumWebPurchases (number of web purchases), influence the model's prediction of customer response to direct marketing. The visualization of the most influential features and the detailed decision tree, including all decision rules, is available in the Jupyter Notebook environment within the GitHub repository[2].

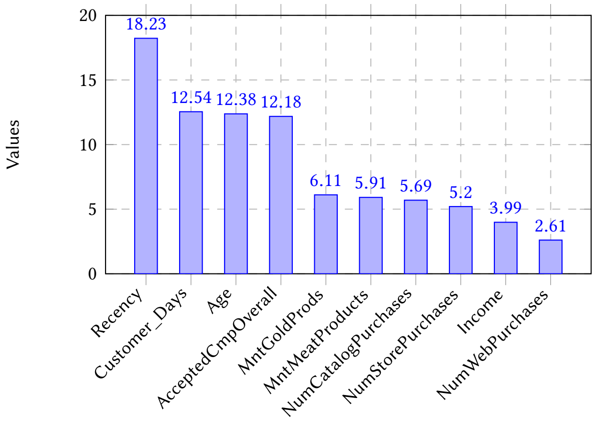

Fig. 4. Feature Importance Scores

5 Discussion

In this section, the results presented in Section 4 are interpreted and their implications for marketing strategies are discussed.

5.1 Results Interpretation

The findings, before resampling, indicate that, although the model had a high accuracy of 87.3%, it struggled to effectively predict the customers that responded positively to marketing campaigns. This is reflected in the relatively low precision (56%), recall (44%), and F1-score (49%), as well as in the confusion matrix that showed a significant number of false negatives (36) and a moderate number of false positives (21). The high accuracy was primarily due to the model's ability to correctly identify non-responders. However, this high accuracy is misleading in the context of the objective of this research, where the performance on the minority class of positive responders is more critical. This imbalance necessitates the use of techniques to improve the model's sensitivity to the minority class.

After applying resampling, and adjusting the class weights the findings demonstrate a significant improvement in the model's ability to predict positive responses. The confusion matrix post-resampling shows a more balanced performance, with 49 true positives and 51 true negatives. Although the overall accuracy decreased to 74.6%, this drop is expected and acceptable given the context of a more balanced dataset. The model's precision increased to 67.1%, indicating a higher proportion of correctly identified positive responders among all predicted positives. The recall improved dramatically to 83.1%, meaning the model is now much better at identifying actual responders, reducing the number of false negatives to 10. The F1-score also increased to 74.2%, providing a balanced measure of the model's precision and recall. These improved results post-resampling mean that the model is now better suited to address the research questions related to predictive modelling in marketing campaigns. The improved precision and recall imply that marketing efforts can be more accurately directed toward potential responders, maximizing the effectiveness of the campaigns and reducing unnecessary marketing expenses. The findings highlight the importance of balancing the dataset to improve model performance, ensuring that both responders and non-responders are effectively identified. Overall, the resampling approach has led to a more robust predictive model, capable of providing actionable insights for marketing strategies. By focusing on the key influential features and understanding the dynamics of customer behaviour, businesses can optimize their marketing efforts to achieve better engagement and conversion rates.

5.2 Implications for Marketing Strategies

In particular, the feature importance analysis in Fig. 4 highlights several key factors influencing customer responses to marketing campaigns. Demographic factors such as age and income play significant roles. Age suggests that certain age groups are more likely to respond to marketing efforts. Income also impacts response rates, indicating that customers with higher income levels might engage more with marketing offers. Past interactions with the company are also really important in shaping the model's predictive power. Recency is the most influential feature suggesting that marketing efforts should focus on customers who have interacted with the company recently, as they are more likely to respond positively to new campaigns. Similarly, the duration of the customer's relationship with the company, measured by Customer_Days, indicates that long-term customers, who have developed loyalty, are more receptive to marketing initiatives. The acceptance of previous campaigns (AcceptedCmpOverall) reflects customers' historical engagement with marketing efforts, suggesting that those who have positively responded in the past are more likely to do so in the future. Additionally, specific product categories, such as MntGoldProds and MntMeatProducts, influence customer responses, indicating preferences for certain products. Understanding these preferences allows for more effective product-specific promotions. The results in this study align with the findings of previous studies, such as those by Apampa [1] and Choi et al. [22], which also highlighted the importance of demographic and past interaction data. However, our study found that Recency and Customer_Days were more influential than previously reported, possibly due to the specific characteristics of

[2] https://github.com/megi2002/Predictive-Modelling-of-Customer-Response-to-Marketing-Campaigns

our dataset and the context of the marketing campaigns analyzed. Furthermore, the model is interpretable, providing clear and understandable decision rules. This interpretability is a significant advantage in the context of marketing campaigns. For example, one of the key decision rules, visualized in the GitHub [3], indicates that if a customer has accepted half of the previous campaigns (AcceptedCmpOverall 0.50), the model then considers their recency of interaction (Recency 42.50). If the customer has interacted with the company in the past 42 days, the model further refines its decision based on the number of catalog purchases (NumCatalogPurchases 0.50). Such rules are straightforward and easily comprehensible for marketing professionals, enabling them to understand the logic behind the model's predictions and make informed decisions based on these insights. This clarity builds trust in the model's recommendations. Marketing teams can confidently use the model to target customers, knowing that the predictions are based on logical and understandable criteria. This transparency is crucial for the practical application of the predictive models. Moreover, the interpretability ensures that the model can be easily updated and adjusted as new data becomes available. As marketing campaigns evolve and customer behaviours change, the decision rules can be re-evaluated and refined.

6 Conclusion

In conclusion, this study demonstrates the effectiveness of using DT models for predicting customer responses to marketing campaigns. By addressing the challenges of class imbalance through resampling and adjusting class weights, the model's ability to accurately predict positive responses improved significantly. This study not only identifies key demographic and interaction factors influencing customer behaviour but also provides a transparent and interpretable model, which is crucial for practical applications in marketing strategies. This study aims to answer three primary research questions. The first question regarding the challenges and limitations presented in the literature was addressed as the "Related Work" part of this study in Section 2, highlighting the complexities of customer behaviour and the limitations of traditional predictive models. The second question on the effectiveness of the DT model in predicting customer response to marketing campaigns is explored through the comparative analysis of model evaluation metrics before and after resampling, as presented in Sections 4.2 and 4.3, and interpreted in Section 5.1. Finally, the key factors influencing customer response are identified through feature importance analysis and decision rules extraction, presented in Section 4.4 and discussed in Section 5.2. Despite the significant improvements achieved, there are several limitations to this study. The dataset, while comprehensive, is limited to a specific context and may not generalize to other industries or geographical regions. Additionally, the use of undersampling, while effective in balancing the classes, reduces the overall dataset size, potentially excluding valuable information from the majority class. Future research should explore the integration of ensemble methods to improve model performance. Studies have shown that ensemble methods, such as RF, can provide significant improvements in handling imbalanced datasets and improving prediction accuracy [2].

References

[1] Olatunji Apampa. 2016. Evaluation of Classification and Ensemble Algorithms for Bank Customer Marketing Response Prediction. (English). *Journal of International Technology and Information Management* 25, 4 (2016). https://doi.org/scholarworks.lib.csusb.edu/jitim/vol25/iss4/6.

[2] Leo Breiman. 2001. Random Forests. *Machine Learning* 45 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[3] Salahdeen K. Nasiru Luiz Fernando Capretz Hammed A. Mojeed Shakirat A. Salihu Abimbola G. Akintola Modinat A. Mabayoje Fatima E. Usman-Hamza, Abdullateef O. Balogun and Joseph B. Awotunde. 2024. Empirical analysis of tree-based classification models for customer churn prediction. *Scientific African* (2024).

[4] Johannes Fürnkranz. 2011. Decision Tree. In *Encyclopedia of Machine Learning*, Claude Sammut and Geoffrey I. Webb (Eds.). Springer, Boston, MA, 263–267. https://doi.org/10.1007/978-0-387-30164-8_204

[5] Dhananjay Bisen Gyanendra Chaubey, Prathamesh Rajendra Gavhane and Siddhartha Kumar Arjaria. 2023. Customer purchasing behavior prediction using machine learning classifcation techniques. *Journal of Ambient Intelligence and Humanized Computing* 14 (2023). https://doi.org/10.1007/s12652-022-03837-6

[6] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data: open challenges and future directions. *IEEE Transactions on Knowledge and Data Engineering* 21, 9 (2009), 1263–1284.

[7] Tin Kam Ho. 1995. Random Decision Forests. In *In Proceedings of 3rd international conference on document analysis and recognition*. IEEE. https://doi.org/10.1109/ICDAR.1995.598994

[8] Mohammed Hossin and Mohd Nazri Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process* 5, 2 (2015), 1–11.

[9] iFood. 2024. iFood DF. https://www.kaggle.com/datasets/diniwilliams/ifood-df.

[10] Miron B. Kursa and Witold R. Rudnicki. 2011. The All Relevant Feature Selection using Random Forest. (2011). https://doi.org/10.48550/arXiv.1106.5112

[11] Gilles Louppe. 2014. *Understanding Random Forests: From Theory to Practice*. Ph. D. Dissertation. University of Liège.

[12] Yige Yuan Maggie Wenjing Liu, Qichao Zhu and Sihan Wu. 2023. The Impact of Predictive Analytics and AI on Digital Marketing Strategy and ROI. *The Palgrave Handbook of Interactive Marketing* (2023). https://doi.org/10.1007/978-3-031-42455-7_31

[13] M M Mehta and S B Talbar. 2017. Class imbalance problem in data mining: review. *International Journal of Computer Applications* 169, 6 (2017), 15–18.

[14] Yao-Yuan Yang Michal Moshkovitz and Kamalika Chaudhuri. 2021. Connecting Interpretability and Robustness in Decision Trees through Separation. (2021). https://doi.org/10.48550/arXiv.2102.07048

[15] Abhijit Raorane and R.V.Kulkarni. 2011. Data Mining Techniques: A Source for Consumer Behavior Analysis. *International Journal of Database Management Systems* (2011). https://doi.org/10.5121/ijdms.2011.3304

[16] Payam Refaeilzadeh, Lei Tang, and Huan Liu. 2009. Cross-Validation. In *Encyclopedia of Database Systems*, Ling Liu and M. Tamer Özsu (Eds.). Springer, Boston, MA, 532–538. https://doi.org/10.1007/978-0-387-30164-8_190

[17] Werner J. Reinartz and V. Kumar. 2002. The mismanagement of customer loyalty. *Harvard Business Review* 80, 7 (2002), 86–94.

[18] Paulo Cortez Sérgio Moro and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. *International Journal of Soft Computing and Engineering (IJSCE)* (2014).

[19] Raul M. S. Laureano Sérgio Moro and Paulo Cortez. 2011. *Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology*. Technical Report. Universidade do Minho.

[20] K. Wisaeng. 2013. A Comparison of Different Classification Techniques for Bank Direct Marketing. *International Journal of Soft Computing and Engineering (IJSCE)* (2013).

[21] Yan yan Song and Ying Lu. 2015. Decision tree methods: Applications for classification and prediction. *Shanghai Arch Psychiatry* 27 (2015). https://doi.org/10.11919/j.issn.1002-0829.215044

[22] Sangmyung Youngkeun Choi and Jae W. Choi. 2023. Assessing the Predictive Performance of Machine Learning in Direct Marketing Response. *International Journal of E-Business Research* 19 (2023). https://doi.org/10.4018/IJEBR.3214581

[3] https://github.com/megi2002/Predictive-Modelling-of-Customer-Response-to-Marketing-Campaigns