Customer Segmentation Based on Marketing Response

Elaf Alshehri
Computer Science Department, College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia
443200821@student.ksu.edu.sa

Afnan Alsuliman
Computer Science Department, College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia
443200648@student.ksu.edu.sa

Asma Albahkaly
Computer Science Department, College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia
443200597@student.ksu.edu.sa

Lama Albediwy
Computer Science Department, College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia
443200503@student.ksu.edu.sa

Shaihanah Alhomidan
Computer Science Department, College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia
443203053@student.ksu.edu.sa

Sheikha Alhaqbani
Computer Science Department, College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia
443200534@student.ksu.edu.sa

***Abstract*— Customer segmentation plays a vital role in improving marketing campaign effectiveness by identifying groups of customers with similar behaviors. In this study, we aim to predict customer responses to term deposit marketing campaigns using the Bank Marketing Campaign Dataset provided by a Portuguese banking institution. We applied three classification algorithms, Random Forest, Logistic Regression, and Gaussian Naive Bayes, to classify whether a customer would subscribe to a term deposit. The dataset was preprocessed through one-hot encoding, normalization, and balancing using downsampling to address class imbalance. After evaluating the models based on accuracy, precision, recall, and F1-score, the Random Forest Classifier demonstrated the best performance. Our results suggest that ensemble methods are particularly effective for complex, real-world datasets like customer marketing data, offering robust and interpretable solutions to support decision-making in targeted marketing strategies.**

Keywords— ***Customer Segmentation, Classification, Marketing Campaigns, Random Forest, Logistic Regression, Naive Bayes, Machine Learning, Bank Marketing Dataset***

I. Introduction

Customer segmentation is a critical component in enhancing marketing strategies by allowing businesses to tailor their services and communications to specific groups of customers. Instead of adopting a one-size-fits-all approach, segmentation enables targeted marketing, ultimately improving customer satisfaction and increasing campaign success rates [1].

The financial sector, particularly banking, has increasingly utilized machine learning techniques to better understand customer behavior and predict their responses to marketing campaigns [2]. Classification algorithms such as Random Forest, Logistic Regression, and Naive Bayes have proven effective in categorizing customers based on historical interaction data [3]. These methods help predict the likelihood of a customer subscribing to a financial product, such as a term deposit, which allows banks to optimize marketing efforts and improve conversion rates.

Effective data preprocessing is a critical step in any machine learning project. Handling missing values, encoding categorical features, normalizing numerical attributes, and addressing class imbalance are essential for building accurate predictive models [4]. Without proper data preparation, models may produce biased or unreliable predictions, especially when dealing with real-world datasets where data inconsistencies are common.

In this study, we explore the application of different classification algorithms to the "Bank Marketing Campaign" dataset[1], which contains detailed customer and campaign-related information collected by a Portuguese banking institution. Our primary goal is to develop a predictive model capable of accurately identifying customers likely to subscribe to a term deposit based on their personal, financial, and interaction data.

The structure of the paper is as follows: Section II presents a brief review of related work; Section III describes the dataset and the preprocessing steps undertaken; Section IV discusses the classification algorithms employed; Section V presents the experimental results and discussion; and Section VI concludes the findings and suggests future directions.

II. Literature Review

Customer segmentation is a fundamental task in marketing to enhance the effectiveness of campaigns by grouping customers based on behavioral and demographic patterns. Various machine learning techniques have been employed to automate this process. Gomes and Meisen [5], conducted a comprehensive survey of customer segmentation methods, emphasizing the use of clustering algorithms such as K-Means, hierarchical clustering, and DBSCAN. They highlighted that effective preprocessing, including feature scaling, one-hot encoding of categorical variables, and dimensionality reduction, is crucial to improving model accuracy and interpretability. Moreover, the study stressed that understanding customer characteristics through unsupervised learning can significantly support targeted marketing strategies and resource allocation.

On the other hand, Khan et al. [6] explored predictive modeling techniques to anticipate customer responses to marketing initiatives. Their work implemented classification models such as Decision Trees, Random Forests, and Logistic Regression, demonstrating that data balancing techniques, such as down sampling or SMOTE, along with normalization, substantially boosting the classification performance. The study also pointed out the importance of evaluating models using precision, recall, and F1-score to ensure reliable decision-making. These findings provide strong guidance for our approach, where both data preprocessing and the choice of machine learning models are critical to achieving effective

[1] https://archive.ics.uci.edu/dataset/222/bank+marketing

and actionable customer segmentation. Customer segmentation plays a critical role in enhancing the efficiency of marketing strategies by categorizing customers based on shared characteristics.

Chaudhary et al. [7], implemented a hybrid machine learning framework combining clustering and classification algorithms to improve customer profiling. Specifically, the study utilized K-Means clustering to initially group customers into behavioral segments, followed by the application of a Random Forest classifier to predict future customer behavior based on these segments. Feature engineering techniques such as normalization, principal component analysis (PCA), and one-hot encoding were extensively employed to boost the model's performance. The study concluded that integrating unsupervised and supervised learning approaches led to a significant increase in marketing response rates by accurately identifying high-potential customer groups.

In contrast, Soni et al. [8] focused on developing a deep learning-based approach for customer segmentation using autoencoders. The authors demonstrated that using deep neural networks to automatically extract high-level features from complex customer data outperformed traditional machine learning models in terms of clustering quality and predictive accuracy. The study highlighted the importance of proper hyperparameter tuning, dropout regularization, and early stopping to avoid overfitting. Evaluation metrics such as Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) were used to assess the effectiveness of segmentation results. The findings emphasized that deep learning models, although computationally intensive, offer powerful tools for understanding customer behavior and enhancing targeted marketing efforts.

III. Material and Methods

A. Dataset Description

The dataset used in this study is the Bank Marketing Campaign Dataset, which was collected from a Portuguese banking institution. It contains information related to direct marketing campaigns aimed at promoting term deposit subscriptions. The dataset is widely used in classification and predictive modeling tasks due to its rich set of features and real-world relevance.

It consists of 45,211 records, with each record representing a different client. The dataset includes a variety of attributes that capture demographic, financial, and marketing information about the clients. Demographic features include age, job type, marital status, and education level. Financial attributes provide details such as account balance, housing loan status, and personal loan status. Additionally, the dataset contains marketing-related information such as the type of contact communication, the number of contacts performed during the campaign, the outcome of previous campaigns, and the duration of the last contact.

The target variable in the dataset is whether the client subscribed to a term deposit or not. It is a binary classification task with two possible outcomes: "yes" or "no." One important characteristic of the dataset is the imbalance in the target classes, where the number of clients who did not subscribe is significantly higher than those who did.

Some attributes require special attention during preprocessing. For example, the "duration" feature represents the last contact duration in seconds and has a strong influence on the target outcome, but it is not available before the contact is made, which may cause data leakage if not handled carefully. The "pdays" feature indicates the number of days since the last contact, with a value of -1 meaning the client was not previously contacted.

Overall, the dataset provides a comprehensive view of customers and their interactions with marketing campaigns. Its combination of categorical and numerical features makes it suitable for applying a variety of machine learning algorithms after proper preprocessing steps such as encoding, normalization, and balancing.

B. Correlation Analysis

We studied the relationships between the numeric features by calculating their correlations. Correlation helps us understand how two numbers change together. A high positive correlation means both numbers increase together, while a high negative correlation means when one number increases, the other decreases. Before analyzing the data, we removed any rows with missing values to keep the results accurate. The highest positive correlation was found between emp.var.rate (employment variation rate) and euribor3m (three-month interest rate), with a value of +0.969, showing a strong relationship influenced by economic factors. The highest negative correlation was between previous (number of previous contacts) and pdays (number of days since last contact), with a value of −0.590, meaning that clients who had a long gap since the last contact usually had fewer previous contacts. These findings help us better understand the dataset and choose the best features for building a good classification model.

C. Data Preprocessing

To prepare the dataset for classification, several preprocessing steps were applied to ensure compatibility with machine learning algorithms and to improve overall model performance.

Initially, categorical features were converted into numerical format using one-hot encoding, allowing the models to interpret them effectively. To prevent multicollinearity, one category from each encoded variable was removed.

Next, Min-Max normalization was applied to the numerical columns, including attributes like age, duration, and campaign, scaling all values to a range between 0 and 1. This ensured that no single feature disproportionately influenced the model due to differences in scale.

To handle the class imbalance in the target variable, a down sampling approach was applied. The majority class (customers who did not subscribe to a term deposit) was randomly reduced without replacement to match the number of samples in the minority class (customers who did subscribe). This helped create a balanced dataset without introducing duplicate entries.

Finally, the balanced dataset was shuffled, and feature attributes were separated from the target attribute. This processed data is fully ready to be used for training and evaluating the classification models.

D. Classification Algorithms

In this project, three different classification algorithms were selected to predict whether a customer would subscribe to a term deposit or not. The selected algorithms are Random Forest Classifier, Logistic Regression, and Gaussian Naive Bayes. Each algorithm was chosen based on its strengths and relevance to the nature of the dataset.

1) *Random Forest*

The Random Forest Classifier is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes for classification. It is known for its high accuracy, robustness to overfitting, and ability to handle datasets with a large number of features. Random Forest was chosen because it can capture complex relationships between attributes and is less sensitive to noise compared to individual decision trees.

2) *Logistic Regression*

Logistic Regression is a linear model commonly used for binary classification tasks. It estimates the probability that a given input belongs to a particular class. Logistic Regression was included because it is simple, efficient, and performs well when the relationship between the features and the target is approximately linear. Additionally, it provides clear insights into the influence of each feature on the prediction outcome.

3) *Naive Bayes*

Gaussian Naive Bayes is a probabilistic classifier based on Bayes' Theorem, assuming independence between the features. It is particularly effective when the independence assumption holds and is computationally efficient. Although this assumption is often not realistic for real-world datasets, Naive Bayes was chosen to provide a comparison with other models and to evaluate how simpler probabilistic models perform on this type of marketing data.

IV. Experiments

Before building the classification models, several preprocessing steps were applied to the dataset to ensure it was suitable for machine learning tasks. These steps included handling missing values by removing incomplete records, applying one-hot encoding to convert categorical variables into numerical format, and normalizing numerical features to bring them onto a similar scale. Furthermore, to address the class imbalance issue, downsampling was performed by reducing the majority class to match the size of the minority class, which helped create a balanced dataset.

After completing data cleaning, correlation study, and preprocessing steps, the dataset was ready for classification modeling. Three different classification algorithms were selected and tested to predict whether a customer would subscribe to a term deposit: Random Forest Classifier, Gaussian Naive Bayes, and Logistic Regression.

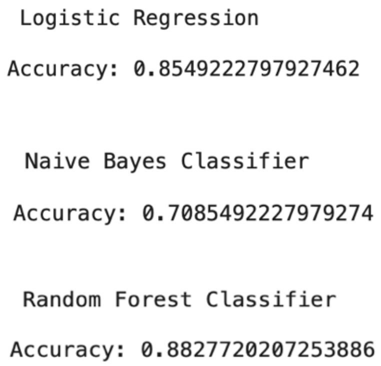
The preprocessed and balanced dataset was first split into training and testing subsets using an 80-20 ratio. Each classifier was then trained on the training set and evaluated using the testing set to ensure a consistent and fair comparison of their performances. All models were implemented using Scikit-learn libraries, with default hyperparameters unless otherwise specified. During training, models learned patterns from the feature set, while testing allowed the assessment of how well these models generalized to unseen data.

Throughout the experiments, the models were evaluated based on various performance metrics including accuracy, precision, recall, and F1-score. This allowed for a detailed comparison to determine which model provided the best predictive capability for customer subscription behavior. Random Forest Classifier, in particular, showed superior performance across most metrics, highlighting its robustness in handling complex datasets with mixed feature types.

V. Results and Discussion

The three models were evaluated based on accuracy score, confusion matrix, and detailed classification reports, including precision, recall, and F1-score for both classes.

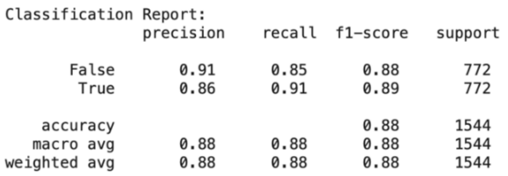
The Random Forest Classifier achieved the highest overall accuracy among the models. It demonstrated strong predictive performance, successfully classifying both customers who subscribed and those who didn’t subscribe to a term deposit. Logistic Regression followed closely, achieving good accuracy but slightly lower recall for the positive class (subscribers). Gaussian Naive Bayes showed the weakest performance, with lower overall accuracy and precision compared to the other two models. This lower performance can be attributed to its assumption of feature independence, which is not fully realistic for this dataset where many features are correlated.



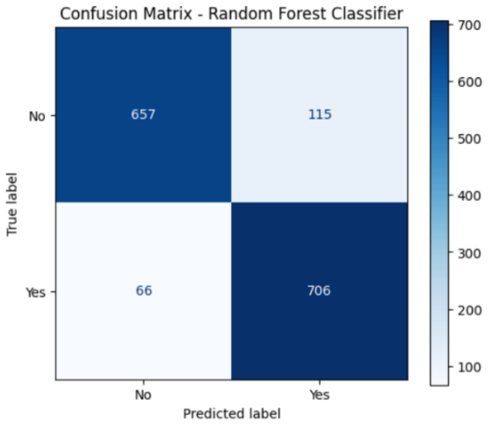
```
 Logistic Regression

Accuracy: 0.8549222797927462


 Naive Bayes Classifier

Accuracy: 0.7085492227979274


 Random Forest Classifier

Accuracy: 0.8827720207253886
```

The confusion matrix for Random Forest showed the highest number of correctly predicted samples and the fewest number of false positives and false negatives. Logistic Regression also performed well but had a slightly higher number of misclassifications. In contrast, Gaussian Naive Bayes tended to misclassify a larger number of subscribing customers, leading to lower recall and a higher false-negative rate.

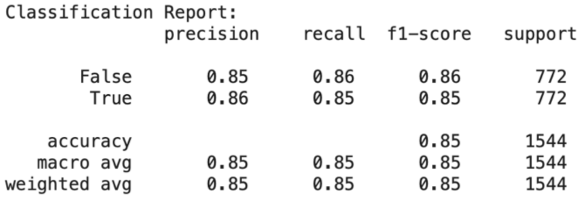
The results indicate that Random Forest is the most successful model for this dataset. Its high accuracy and balanced classification metrics suggest that it can capture complex interactions between the features without relying on strong assumptions. This is expected because Random Forest, as an ensemble method, builds multiple decision trees and averages their results, making it robust to noise and capable of modeling non-linear feature relationships.



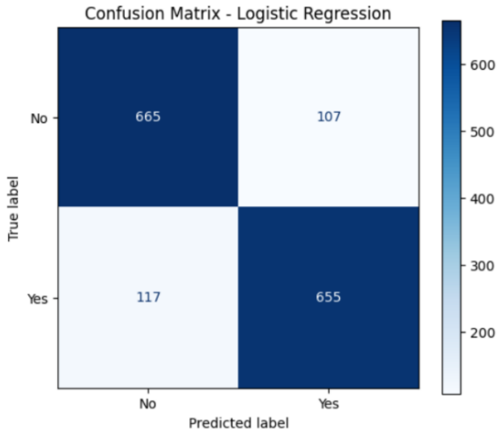
```
Classification Report:
              precision    recall  f1-score   support

       False       0.91      0.85      0.88       772
        True       0.86      0.91      0.89       772

    accuracy                           0.88      1544
   macro avg       0.88      0.88      0.88      1544
weighted avg       0.88      0.88      0.88      1544
```



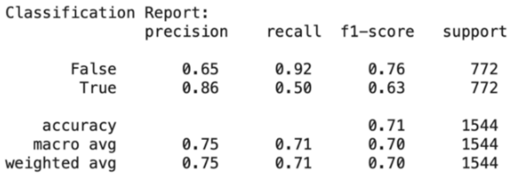
Logistic Regression, still performed well, highlighting that some linear relationships between the features and the target exist. However, it was slightly less effective in distinguishing between customers who subscribed and those who did not, mainly when the decision boundary was not strictly linear.



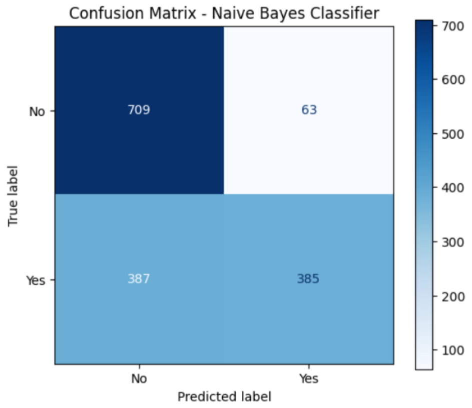
```
Classification Report:
              precision    recall  f1-score   support

       False       0.85      0.86      0.86       772
        True       0.86      0.85      0.85       772

    accuracy                           0.85      1544
   macro avg       0.85      0.85      0.85      1544
weighted avg       0.85      0.85      0.85      1544
```



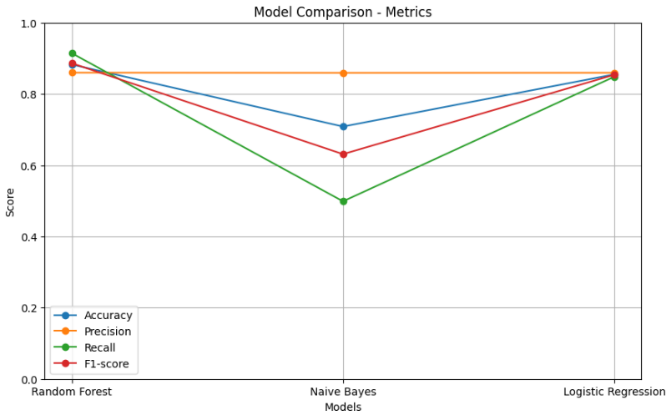
Another important factor that helped improve model performance was the balancing of classes through down sampling. Without balancing, models would have been biased toward predicting the majority class ("no subscription"), reducing their ability to detect true subscribers. Balancing the classes allowed models to treat both classes more equally, improving recall for the minority class.



```
Classification Report:
              precision    recall  f1-score   support

       False       0.65      0.92      0.76       772
        True       0.86      0.50      0.63       772

    accuracy                           0.71      1544
   macro avg       0.75      0.71      0.70      1544
weighted avg       0.75      0.71      0.70      1544
```



In conclusion, the Random Forest Classifier is the most appropriate algorithm for predicting customer subscription based on the tested models, offering high accuracy and robustness. It achieved the best balance across all performance metrics, including accuracy, precision, recall, and F1-score, making it the most reliable model for customer response prediction. Logistic Regression also demonstrated strong and consistent performance, suggesting it could serve as a simpler alternative when computational efficiency is needed. In contrast, Gaussian Naive Bayes performed significantly worse, particularly in recall, indicating its limited suitability for this type of marketing data. Further improvements could be explored in future work, such as hyperparameter tuning or trying more advanced ensemble methods like Gradient Boosting to enhance predictive performance.



VI. Conclusions

In this study, we explored the use of machine learning techniques to predict customer responses to a bank marketing campaign. By utilizing the Bank Marketing Campaign Dataset, we performed a series of preprocessing steps to clean and prepare the data, including handling missing values, encoding categorical variables, normalizing numerical features, and addressing class imbalance through downsampling. These steps ensured that the dataset was suitable for accurate and unbiased model training.

Three classification algorithms, Random Forest Classifier, Logistic Regression, and Gaussian Naive Bayes were implemented and evaluated. The experiments showed that the Random Forest Classifier achieved the highest overall performance, demonstrating strong predictive capability, robustness to noise, and effective handling of feature interactions. Logistic Regression also performed well, particularly benefiting from its simplicity and ability to model linear relationships between features. In contrast, Gaussian

Naive Bayes showed lower performance, likely due to its assumption of feature independence, which does not hold for the real-world complexities present in the dataset.

The results highlight the importance of selecting appropriate algorithms based on the characteristics of the dataset. Ensemble methods like Random Forest proved particularly effective in capturing the non-linear relationships and interactions between features, leading to better overall model performance. Additionally, careful data preprocessing, including balancing classes and normalization, played a critical role in improving model accuracy and fairness.

Overall, this study demonstrates that with the right combination of preprocessing techniques and classification models, it is possible to significantly enhance the ability to predict customer behavior in marketing contexts. Future work may focus on fine-tuning model hyperparameters, applying more advanced ensemble techniques such as Gradient Boosting, and exploring feature selection methods to further improve model efficiency and interpretability.

REFERENCES

[1] G. Wang, “Customer segmentation in the digital marketing using a Q-learning based differential evolution algorithm integrated with K-means clustering,” *PLoS One*, vol. 20, no. 2 February, Feb. 2025, doi: 10.1371/JOURNAL.PONE.0318519.

[2] Ibrahim Adedeji Adeniran, Angela Omozele Abhulimen, Anwuli Nkemchor Obiki-Osafiele, Olajide Soji Osundare, Edith Ebele Agu, and Christianah Pelumi Efunniyi, “Data-Driven approaches to improve customer experience in banking: Techniques and outcomes,” *International Journal of Management & Entrepreneurship Research*, vol. 6, no. 8, pp. 2797–2818, Aug. 2024, doi: 10.51594/IJMER.V6I8.1467.

[3] J. Brownlee, “Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning,” *Machine Learning Mastery*, pp. 1–463, 2020.

[4] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory To Algorithms*. 2015. [Online]. Available: https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

[5] M. A. Gomes, T. Meisen, M. A. Gomes, and T. Meisen, “A review on customer segmentation methods for personalized customer targeting in e-commerce use cases,” *Information Systems and e-Business Management 2023 21:3*, vol. 21, no. 3, pp. 527–570, Jun. 2023, doi: 10.1007/S10257-023-00640-4.

[6] M. A. R. Khan, M. S. Akter, and R. Islam, “Big Data Analytics And Predictive Analysis In Enhancing Customer Relationship Management (CRM): A Systematic Review Of Techniques And Tools,” *Non human journal*, vol. 1, no. 01, pp. 83–99, Nov. 2024, doi: 10.70008/JMLDEDS.V1I01.44.

[7] P. Chaudhary, V. Kalra, and S. Sharma, “A Hybrid Machine Learning Approach for Customer Segmentation Using RFM Analysis,” *Lecture Notes in Electrical Engineering*, vol. 836, pp. 87–100, 2022, doi: 10.1007/978-981-16-8542-2_7.

[8] J. Soni, N. Prabakar, and H. Upadhyay, “Deep Learning-Based Efficient Customer Segmentation for Online Retail Businesses,” pp. 147–164, 2023, doi: 10.1007/978-981-99-3970-1_9.