Proceeding
International Conference on Economics, Business and Information Technology
E-ISSN 3025-4892
Vol. 6 (2025), Page 616-622

Smart Campaign for Smart Business Using Machine Learning on Improving Product Marketing Strategy

Difari Afreyna Fauziah[1*], Agung Muliawan[2], Iqbal Sabilirrasyad[3], Bima Wahyu Maulana[4]

[1] Sistem dan Teknologi Informasi, Institut Teknologi dan Sains Mandala, Indonesia
[2] Manajemen Informatika, Politeknik Negeri Jember, Indonesia
[3] Rekayasa Perangkat Lunak, Institut Teknologi dan Sains Mandala, Indonesia
[4] Teknik Komputer, Politeknik Negeri Jember, Indonesia
Corresponding Author: difariafreyna@itsm.ac.id
Received: 10-08-2025 Accepted: 13-08-2025 Published: 30-09-2025

Abstract

Improving the effectiveness of banking marketing campaigns is a major challenge in retaining and attracting new customers. This research aims to predict the success of direct marketing campaigns using the Random Forest algorithm on the UCI Bank Marketing dataset. The dataset includes various demographic variables and historical customer interactions. Preprocessing was done through min-max normalization and division of training and test data with a ratio of 80:20. Model evaluation results show that Random Forest is able to classify uninterested customers with high accuracy (true negative = 794), but has a weakness in detecting interested customers (true positive = 9), indicating a class imbalance in the data. The overall accuracy of the model reached 87%, with a precision of 64% and a recall of 8.7%. Feature importance analysis showed that the variables balance, age, and day were the most influential factors in customer decisions. Overall, the Random Forest algorithm successfully uncovered important patterns in the data that are relevant for more targeted marketing decisions. Nonetheless, improving the model's performance towards minority classes needs to be done through the approach of handling data imbalance. This research contributes to the utilization of machine learning in supporting data-driven marketing strategies in the banking sector.

Keywords: Random Forest; technopreneur; smart campaign

1. Introduction

The development of digital technology in the last two decades has driven significant transformation in the global business world, including in Indonesia. Rapid digitization not only impacts the production and distribution process, but also drastically changes the way companies market their products. Consumers are now increasingly connected, have extensive access to information, and exhibit more dynamic and unpredictable behavior. In this modern business ecosystem, companies can no longer rely on traditional marketing strategies that are generic and one-way. Instead, a more personalized, data-driven approach oriented towards a deep understanding of customer needs and preferences is a must.

This condition creates an urgency for the application of marketing intelligence, which is an approach that utilizes information technology and data analysis to collect, process, and interpret market information as a basis for formulating marketing strategies. Marketing intelligence allows companies to detect changes in market trends, assess the effectiveness of marketing campaigns, and conduct more precise segmentation and targeting. For start-ups and technopreneurs, the ability to execute adaptive and intelligent marketing strategies is a key factor in surviving and growing amidst intense competition and high market volatility.

This research takes a case study of the Bank Marketing dataset published by the UCI Machine Learning Repository. This dataset represents data on the results of telemarketing campaigns from banking institutions to promote time deposit products. Through the Random Forest algorithm approach, this research aims to identify the most influential factors in determining the success of the campaign and evaluate the performance of the prediction model used.

The ultimate goal of this research is to offer an analytical model that can be used by digital businesses in designing smart campaigns, which are marketing intelligence-based campaigns that

are not only technically effective, but also strategic in facing competition in the digital economy era. By integrating data science into the marketing decision-making process, technopreneurs are expected to improve targeting accuracy and marketing budget efficiency.

2. Methods

This research uses an exploratory quantitative approach with the support of data mining techniques to analyze the effectiveness of data-based product marketing campaigns. The stages of the methodology used include several steps shown in this figure:

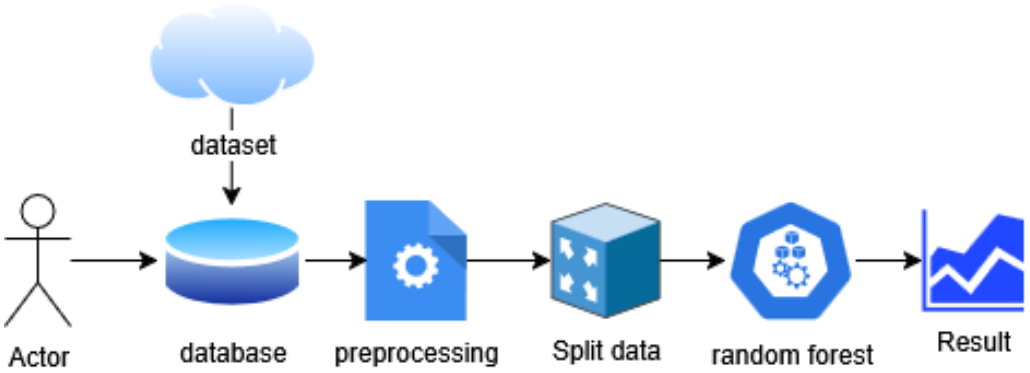

Figure 1. Methods

A. Data Collection

The data source used in this research comes from the Bank Marketing open dataset provided by the UCI Machine Learning Repository. This dataset contains historical data on the results of marketing campaigns for time deposit products from a banking institution in Europe. The dataset includes 17 independent variables as well as 1 target variable which is the campaign response (subscribe or not) with a total of 4567 rows described in the following table:

Table 1. Variabel Dataset

Variabel Name	Type Data	Description
age	Integer	Customer's age
Job	Categorical	Customer's occupation type (e.g. admin., technician, services, management, unemployed, etc.)
marital	Categorical	Marital status (e.g. married, single, divorced)
education	Categorical	Education level (e.g. primary, secondary, tertiary, unknown)
default	Binary	Does the customer have any previous bad debts?
balance	Integer	Annual average balance in the account (in euros)
housing	Binary	Does the customer have a housing loan?
loan	Binary	Does the customer have a personal loan?
contact	Categorical	Type of contact communication (cellular, telephone)
day_of_week	Date	Day of the month when last contacted
Month	Categorical	Month when last contacted (jan, feb, mar, etc.)
duration	Numeric	Duration of last call (in seconds)
campaign	Numeric	Number of contacts made during this campaign to the same customer
pdays	Numeric	Number of days since the customer was last contacted in a previous campaign
previous	Numeric	Number of contacts made before this campaign

poutcome	Categorical	Results of previous marketing campaigns (success, failure, nonexistent)
y	Categorical	Does the customer subscribe to time deposits?

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
2	30	unemploye	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
3	33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
4	35	manageme	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
5	30	manageme	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
6	59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no
7	35	manageme	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no
8	36	self-emplo	married	tertiary	no	307	yes	no	cellular	14	may	341	1	330	2	other	no
9	39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2	-1	0	unknown	no
10	41	entrepren	married	tertiary	no	221	yes	no	unknown	14	may	57	2	-1	0	unknown	no
11	43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	147	2	failure	no
12	39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	-1	0	unknown	no
13	43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	-1	0	unknown	no
14	36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	-1	0	unknown	no
15	20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	-1	0	unknown	yes
16	31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	241	1	failure	no

Figure 2. Sampel Dataset

B. Preprocesing

Data pre-processing is done to ensure the quality of the data used in the analysis is in optimal condition (M et al., 2025). The first stage starts with checking for missing values, but the Bank Marketing dataset from UCI is known to contain no missing values (Hermansyah et al., 2024). Next, categorical variables were mapped into numerical form using label encoding and one-hot encoding techniques on features such as job, marital, education, contact, month, and poutcome, to enable machine learning algorithms to process the data effectively. The duration variable was omitted from the model training process as the value is only available after the interaction with the customer is complete and may lead to data leakage. The pdays feature that has a dominant value of 999 is considered for the transformation of the value into the categories “has been contacted” and “has not been contacted” to improve the interpretability of the model (Rustam et al., 2020).

The next stage is normalizing numeric features such as age, balance, and campaign using the Min-Max Scaling method in order to have a uniform scale and avoid the dominance of certain features in the classification process. The following is the formula for the Min-Max Scaling method (Afrianto et al., 2024):

$$Xscaled \frac{X - X\,min}{X\,\max - X\,min}$$

Description:
X = original value
X min = minimum value of the feature (column)
X max = maximum value of the feature (column)
X scaled = normalized value in the range [0, 1]

C. Split Data

After the pre-processing process is complete, the next step in this research is to perform data splitting to separate the training data (training set) and test data (test set) (Sabilirrasyad et al., 2024). The purpose of this division is to test the generalization ability of the model built on data that has never been seen before. The data is divided with a ratio of 80:20, where 80% is used for model training and the remaining 20% is used for testing. This process is done randomly but still maintains the distribution of the proportion of target classes using stratified sampling techniques to avoid bias towards the majority class (Malik et al., 2024). Thus, the model can be trained with sufficiently representative data and tested with unbiased data, so that the results of model performance evaluation become more objective and reliable (Ahuja et al., 2025).

D. Random Forest

In this study, the Random Forest algorithm was selected as a classification method to predict the success of product marketing campaigns based on historical data from the Bank Marketing dataset. Random Forest is one of the bagging-based ensemble learning techniques, which forms a number of decision trees from a subset of data taken randomly with a bootstrap technique (Aini et al., 2024). Each tree provides a prediction result, and the final result is determined based on the majority vote of all trees. The main advantage of Random Forest lies in its ability to handle high-dimensional data,

reduce the risk of overfitting, and provide important information about feature importance (Devella et al., 2020). In the context of this research, Random Forest is not only used to build predictive models, but also to identify the key variables that have the most influence on customer decisions to accept or reject banking product offers (Devella et al., 2020). Random Forest does not have one explicit formula like regression, because it is non-parametric and based on voting from many trees (Erdiansyah et al., 2022).

E. Confusion Matrix (Result)

To evaluate the performance of the Random Forest classification model, the confusion matrix is used as a tool to analyze the prediction performance of the target class. Confusion matrix presents classification results in the form of a two-dimensional table that compares between model predictions and actual labels (Fauziah et al., 2024). In this case, the target class consists of two categories, namely yes (customer accepts the product offer) and no (customer rejects). The four main components in the confusion matrix are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). TP indicates the number of cases where the model successfully predicted yes correctly, while TN is for no predictions that were also correct. Conversely, FP indicates the number of incorrect yes predictions, and FN for incorrect no predictions. Based on the resulting confusion matrix, evaluation metrics such as accuracy, precision, recall, and F1-score are obtained that reflect the model's ability to classify data correctly and in a balanced manner. This evaluation is important to ensure that the model not only has high accuracy, but is also able to avoid risky misclassifications in a marketing context. The following is the confusion matrix formula (Badriyah et al., 2024):

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

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - score = 2\,x \frac{Presisi\ x\ Recall}{Presisi + Recall}$$

Description:
TP = True Positive: Predicted "Yes" and actually "Yes"
TN = True Negative: Predicted "No" and actually "No"
FP = False Positive: Predicted "Yes" but actually "No" (Type I Error)
FN = False Negative: Predicted "No" but actually "Yes" (Type II Error)

3. Results and Discussion

The classification model built using the Random Forest algorithm is evaluated using a confusion matrix to measure performance in predicting the success of marketing campaigns. Based on the confusion matrix generated in the following figure:

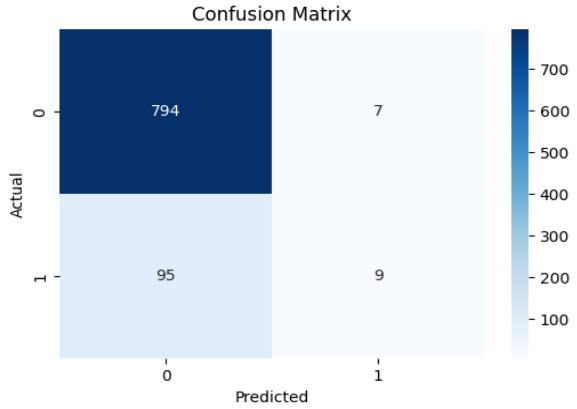

Figure 3. Confusion matrix

```
Confusion Matrix:
 [[794   7]
 [ 95   9]]

Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.99      0.94       801
           1       0.56      0.09      0.15       104

    accuracy                           0.89       905
   macro avg       0.73      0.54      0.54       905
weighted avg       0.86      0.89      0.85       905

Akurasi: 0.887292817679558
```

Figure 4. Accuracy Random Forest

Although the model produced high accuracy (88.73%), the recall and F1-score values were very low, especially in detecting the positive class (customers who responded to the campaign). This indicates that the model is more likely to classify most of the data into the negative class, i.e. customers who did not respond. This is common in imbalanced datasets, where there are far fewer customers who respond to campaigns than those who do not. In a business context, this is a serious concern as models tend to miss potential opportunities from customers who are actually interested (Muliawan et al., 2022).

Furthermore, the following visualization displays the 10 most important features in the Random Forest model that contribute to the prediction of marketing campaign success. Each feature's importance is calculated based on how much it is used to split the data in the decision tree within the Random Forest ensemble.

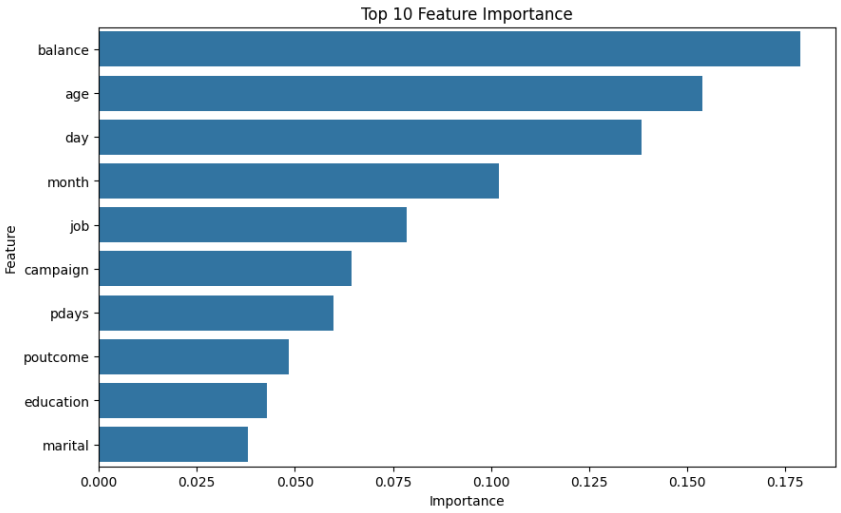

Figure 5. Top Fiture

The most significant variable is balance, which is the average balance of customers. The higher the balance, the more likely the customer is to be financially stable and therefore more likely to respond to marketing campaigns. Furthermore, age is also an important indicator because preferences and responses to financial products often differ between age groups. For example, younger customers tend to be more interested in long-term investment products, while older ones may be more conservative.

Day and month indicate the time when the customer was contacted. This time pattern is important because the success of the campaign can be affected by seasonal factors or the best time to reach out to customers, such as towards the end of the month or in certain months that coincide with annual bonuses. Meanwhile, job indicates the socio-economic background of the customer, which can influence interest in certain products.

The variables campaign (number of contacts during the campaign) and pdays (number of days since the last contact in the previous campaign) provide information on the intensity and frequency of communication by the bank. Too much contact can have a negative impact, but the right approach can increase success. Poutcome, which describes the outcome of previous marketing campaigns (e.g., success or failure), is also decisive as it reflects trends in customer behavior over time.

The last two variables are education and marital. Education level affects customers' understanding of banking products, while marital status can relate to different financial needs, such as family insurance or children's education savings. Overall, these ten variables show that a combination of demographic, economic, and historical behavioral factors play a significant role in determining the success of a bank's marketing campaign. Insights from this analysis can help banks strategize more targeted campaigns.

4. Conclusion

This study aims to predict the success of direct marketing campaigns to bank customers using the Random Forest algorithm based on data from the UCI Bank Marketing Dataset. Based on the model evaluation results through confusion matrix, the Random Forest algorithm produces high accuracy in classifying customers who are not interested in the bank's product offer (794 true negatives), but is less optimal in predicting customers who are actually interested (only 9 true positives out of 104 actual cases). This indicates a class imbalance problem in the data, where the majority of labels are negative (not interested), so the model tends to predict towards the majority.

However, the results of the feature importance analysis show that the model is able to identify the variables that are most influential in customer decisions. The ten most important variables include balance, age, day, month, job, campaign, pdays, poutcome, education, and marital. These variables reflect demographic, economic, and historical customer interactions with the bank, which collectively provide important insights to strategize a more targeted campaign. Thus, this study concludes that while the classification performance of minority classes still needs improvement, the Random Forest algorithm still makes an important contribution in uncovering key patterns that influence the success of bank marketing campaigns. The implications of these results can be used as a basis for improving customer segmentation strategies and optimizing the time and intensity of contact with customers. In the future, improved model performance can be achieved by handling data imbalance, such as using oversampling, undersampling, or specialized algorithms for imbalanced data.

References

Afrianto, E., Wiranto, F., Muliawan, A., & Muhdar, M. (2024). DATA MINING ANALYST FOR CLASSIFYING PLANT GROWTH DATA USING THE NAIVE BAYES METHOD. *PROCEEDING INTERNATIONAL CONFERENCE ON ECONOMICS, BUSINESS AND INFORMATION TECHNOLOGY (ICEBIT)*, *5*, 233–239. https://doi.org/10.31967/prmandala.v5i0.1190

Ahuja, J., Sharma, V., Gupta, R., Kapoor, P., & Arora, S. (2025). Decoding Machine Learning Algorithms for DR Detection. *2024 International Conference on IT Innovation and Knowledge Discovery (ITIKD)*, 1–7. https://doi.org/10.1109/ITIKD63574.2025.11004715

Aini, E. D. N., Khasanah, R. A., Ristyawan, A., & Diniati, E. (2024). Penggunaan Data Mining untuk Prediksi tingkat Obesitas di Meksiko Menggunakan Metode Random Forest. *Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi)*, *8*(3), 1256–1265.

Badriyah, J., Ramadhani, N., Muliawan, A., Ummah, K. R., & Amrullah, A. (2024). Penerapan Dimensi Reduksi Pada Machine Learning Dalam Klasifikasi Kanker Payudara Berdasarkan Parameter Medis. *Jurnal RESTIKOM : Riset Teknik Informatika Dan Komputer*, *6*(3), Article 3. https://doi.org/10.52005/restikom.v6i3.379

Devella, S., Yohannes, Y., & Rahmawati, F. (2020). Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT. *JATISI (Jurnal Teknik Informatika Dan Sistem Informasi)*, *7*, 310–320. https://doi.org/10.35957/jatisi.v7i2.289

Erdiansyah, U., Lubis, A. I., & Erwansyah, K. (2022). Komparasi Metode K-Nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *6*(1), Article 1. https://doi.org/10.30865/mib.v6i1.3373

Fauziah, D. A., Muliawan, A., & Dimyati, M. (2024). IMPLEMENTATION OF MACHINE LEARNING ON EMPLOYEE ATTRITION BASED ON PERFORMANCE PARAMETERS USING PARTICLE SWARM OPTIMIZATION AND ENSEMBLE CLASSIFER METHODS. *Jurnal Teknik Informatika (Jutif)*, *5*(6), Article 6. https://doi.org/10.52436/1.jutif.2024.5.6.3442

Hermansyah, M., Prasetyo, N. A., Muliawan, A., Firdausi, M. F., & Wiranto, F. (2024). Classification of Student Readiness for Educational Unit Exams: Decision Tree Approach C4.5 Based on Try Out Scores at MTs Nahdlatul Arifin. *LOREM: Computational Engineering and Computer Information Systems*, *1*(1), Article 1.

M, K., M, J. R., & K, P. (2025). Metaheuristic Feature Selection for Diabetes Prediction with P-G-S Approach. *Procedia Computer Science*, *252*, 165–171. https://doi.org/10.1016/j.procs.2024.12.018

Malik, I., Iqbal, A., Gu, Y. H., & Al-antari, M. A. (2024). Deep Learning for Alzheimer's Disease Prediction: A Comprehensive Review. *Diagnostics*, *14*(12), Article 12. https://doi.org/10.3390/diagnostics14121281

Muliawan, A., Badriyah, T., & Syarif, I. (2022). Membangun Sistem Rekomendasi Hotel dengan Content Based Filtering Menggunakan K-Nearest Neighbor dan Haversine Formula. *Technomedia Journal*, *7*, 231–247. https://doi.org/10.33050/tmj.v7i2.1893

Rustam, R., Rahmatullah, S., Supriyato, S., & Wahyuni, S. (2020). PENERAPAN DATA MINING UNTUK PREDIKSI PENJUALAN PRODUK TRIPLEK PADA PT PUNCAK MENARA HIJAU MAS. *Jurnal Informasi dan Komputer*, *8*(2), Article 2. https://doi.org/10.35959/jik.v8i2.186

Sabilirrasyad, I., Hermansyah, M., Prasetyo, N. A., Muliawan, A., & Wahid, A. (2024). Unveiling X/Twitter's Sentiment Landscape: A Python Crawler That Maps Opinion Using Advanced Search. *LOREM: Computational Engineering and Computer Information Systems*, *1*(1), Article 1.