KHATULISTIWA SMART : JOURNAL OF ARTIFICIAL INTELLIGENCE
e-ISSN : 3089-9710
p-ISSN : XXXX-XXXX

Khatulistiwa : Journal of Artificial Intelligence
Vol. 3 No. 1 (2026) : 1 - 17
Available online at : https://ejournal.narasikhatulistiwa.org/index.php/kjarti

Predicting Potential Deposit Customers Using CatBoost with Feature Selection and Hyperparameter Optimization

Jiawei Han [1], Zainal Arifin [2]
[1] University of Illinois Urbana Champaign, Amerika Serikat,
[2] Universitas Nurul Jadid, Indonesia
hanj@illinois.edu [1], zainal@unuja.ac.id [2]

Doi
Submitted: 21 May 2026 | Revision: 20 Jun 2026 | Accepted: 25 Jun 2026 | Published: 30 Jun 2026

Abstract The banking industry requires increasingly targeted marketing strategies to identify potential customers for deposit products. This study aims to build a predictive model for identifying potential deposit customers using the CatBoost algorithm combined with feature selection based on feature importance and hyperparameter tuning using Hyperopt. This study employs an experimental method using the Bank Marketing Dataset from Kaggle, a secondary dataset consisting of 45,211 data points and 17 attributes, including a target variable indicating the customer's decision to subscribe to a deposit product or not. The research stages included Exploratory Data Analysis (EDA), preprocessing, data cleaning, class balancing using SMOTE-NC, feature selection, hyperparameter tuning, modeling, and model evaluation. The EDA results revealed outliers in the duration and previous features, as well as class imbalance in the target variable. After preprocessing, the model was built using 10 selected features: duration, month, contact, day, outcome, balance, housing, pdays, age, and job. Evaluation results using a confusion matrix showed that the CatBoost model achieved an accuracy of 92.8%, a sensitivity of 91.0%, and a specificity of 94.8%. These findings demonstrate that the combination of CatBoost, feature selection, and hyperparameter tuning is capable of producing an accurate and efficient predictive model for handling both numerical and categorical data. This model can help banks develop more effective, cost-efficient, measurable, and data-driven deposit marketing strategies, while also supporting business decision-making that is more adaptive to customer behavior.

Keywords *CatBoost, Bank Marketing, Feature Selection, Hyperparameter Tuning, Classification*

1. INTRODUCTION

In the modern era, the banking industry faces increasingly complex challenges, particularly regarding the intensifying competition to attract customers (Fejza Ademi et al., 2022). This fierce competition drives banking institutions to proactively design and implement new strategies to maintain and expand their market share. One strategy that can be implemented is to improve the ability to accurately and comprehensively predict potential customers (Sagar, 2023). This predictive capability is a crucial aspect for banks in developing more effective, targeted, and efficient marketing strategies, particularly when promoting various financial products or instruments to segments of the population that are more likely to respond to offers.

One of the most widely promoted banking products is the time deposit. Under Law No. 10 of 1998, a time deposit is defined as a type of savings account from which funds can only be withdrawn after a specified period, in accordance with an agreement between the customer and the bank (Republic of Indonesia, 1998). By being able to identify potential deposit customers, banks can more accurately pinpoint segments of the population with the financial capacity, investment tendencies, and interest in term deposit products. This enables banks to develop deposit promotion strategies that are more relevant, targeted, and aligned with the characteristics of their target market.

Machine learning approaches are increasingly being used in the process of predicting potential deposit customers. This is due to machine learning's ability to learn patterns, analyze large amounts of data, and generate predictions that are more adaptive than those of conventional approaches. The effectiveness of this approach has been demonstrated in various previous research studies (Fajri et al., 2022; Hudawi et al., 2021; Putri et al., 2023; Rafrastara et al., 2024; Tholib et al., 2023). In the banking context, machine learning can help process demographic data, transaction histories, and customer responses to marketing campaigns to generate more accurate predictive models.

A number of previous studies have applied machine learning to predict potential deposit customers. The study (Puteri et al. 2021) used correlation-based feature selection and Multilayer Perceptron Neural Networks (MLPNN) classification. The study successfully identified several key attributes, such as duration, previous, contact, cons.price.idx, month, cons.conf.idx, age, job, marital, and housing, with a prediction accuracy of 80.5%. Meanwhile, a study (Zaki et al. 2024) tested several classification models, namely the SGD Classifier, k-NN, and Random Forest, with the best results obtained by Random Forest, which achieved an accuracy of 87.5%. Another study conducted by (Riyyasy et al. 2023) applied several machine learning classification algorithms, such as Random Forest, XGBoost, SVC, and Logistic Regression. The results of this study showed that Random Forest and XGBoost were the best models for predicting potential deposit customers, with an accuracy of 91.7%.

Nevertheless, previous studies still have limitations, particularly regarding the handling of categorical data. In some machine learning algorithms, categorical data generally must undergo a transformation process, such as one-hot encoding. This process can increase data dimensionality, especially when the dataset contains many categorical features (Cerda & Varoquaux, 2020). Increased dimensionality not only adds to the computational load but also has the potential to affect the efficiency and performance of

classification models. Therefore, an approach is needed that can handle categorical data more efficiently without excessively increasing data complexity.

To address this issue, this study proposes the use of the CatBoost algorithm to predict potential deposit customers. CatBoost is a gradient-boosting-based machine learning algorithm that is specifically designed to handle categorical data efficiently without requiring one-hot encoding, unlike many other algorithms (Hancock & Khoshgoftaar, 2020). With this capability, CatBoost is expected to reduce computational complexity and data dimensionality, particularly in datasets with many categorical features. Additionally, this study combines feature selection techniques based on feature importance to select the most relevant features for predicting potential deposit customers, as well as hyperparameter tuning using Hyperopt to optimize the CatBoost model parameters for optimal performance.

Through this approach, this study aims to develop a CatBoost-based predictive model for identifying potential deposit customers, combined with feature selection and hyperparameter tuning. The resulting model is expected not only to improve prediction accuracy but also to reduce computational complexity. Practically, the results of this study are expected to assist the banking industry in developing more targeted, efficient, and data-driven deposit marketing strategies. Additionally, this study is also expected to contribute to the development of machine learning applications in predicting customer behavior within the banking sector.

2. RESEARCH METHODS

This study employs an experimental method to implement the CatBoost algorithm in combination with feature selection and hyperparameter tuning techniques. The experiments were conducted using Google Colaboratory as the computational environment, supported by various Python libraries required for data analysis, preprocessing, model building, and machine learning performance evaluation.

The research process began with uploading the dataset used as the research data source. Following that, data characteristics were analyzed through Exploratory Data Analysis (EDA) to understand distribution patterns, the presence of outliers, and class imbalance in the dataset. Subsequently, the data underwent a preprocessing stage consisting of data cleaning and class balancing. The processed data is then used in the feature selection stage to select features relevant to the prediction target. Once the selected features are obtained, hyperparameter tuning is performed to find the optimal parameter combination for the CatBoost algorithm. The final stage involves building the CatBoost model using the selected features and optimal parameters, followed by model evaluation using a confusion matrix. The research workflow is illustrated in Figure 1.

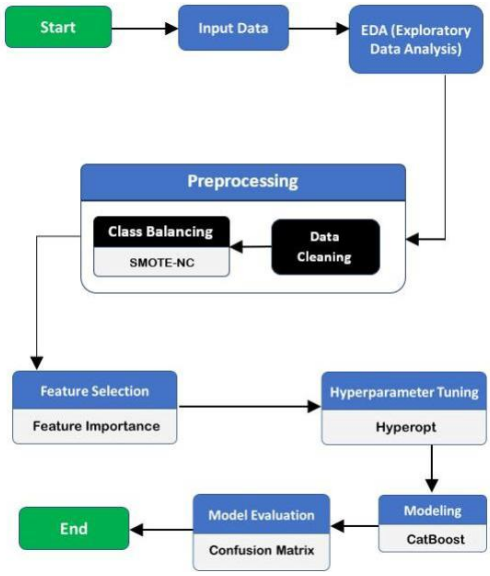


Figure 1. Research stages

2.1. Data Entry

The dataset used in this study is a secondary dataset obtained from the website www.kaggle.com. The dataset was uploaded by Hariharanpavan in 2022 under the title "Bank Marketing Dataset." The data comes from a direct marketing campaign conducted over the phone by a banking institution in Portugal between May 2008 and November 2010.

The dataset consists of 45,211 data rows with 16 independent attributes covering customer demographic information, interaction history with the bank, and marketing campaign details. These attributes include age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, and outcome. The target variable used is the binary label y, which indicates whether the customer made a deposit (category: yes) or did not make a deposit (category: no). This data was obtained through a combination of bank transaction records and telemarketing call results. Details of the dataset attributes can be seen in Table 1.

Tabel 1. Deskripsi Dataset

No	Attribute	Description
1	Age	The customer's age in numerical form
2	Job	Customer occupation, such as administrative staff, unknown, unemployed, management, housekeeper, entrepreneur, student, blue-collar worker, self-employed, retired, technician, and service worker
3	Marital	Marital status, including married, divorced, and single
4	Education	Level of education, consisting of unknown, secondary, primary, and tertiary
5	Default	Customer's history of delinquent loans, consisting of "yes" and "no"
6	Balance	Customer account balance in euros
7	Housing	Homeownership status, consisting of "yes" and "no"
8	Loan	Status of other loans, consisting of "yes" and "no"
9	Contact	The methods used to contact customers include unknown, telephone, and cell phone

10	Day	The last day customers were contacted as part of the marketing campaign
11	Month	The last month the customer was contacted, such as Jan, Feb, Mar, through Dec
12	Duration	Duration of contact with customers during promotions, in seconds
13	Campaign	The number of promotional messages received by customers during the campaign
14	Pdays	The number of days since the customer was last contacted before this campaign was launched; a value of -1 indicates that the customer has never been contacted before
15	Previous	The number of contacts made prior to this campaign
16	Poutcome	The results of previous marketing campaigns include unknown, other, failure, and success
17	Y	Whether or not a customer has a time deposit account, with options of "yes" and "no"

2.2. EDA (Exploratory Data Analysis)

Exploratory Data Analysis (EDA) is a process for understanding the characteristics and patterns present in data through visual and descriptive analysis, thereby enabling the determination of appropriate data handling steps to achieve optimal results. Through EDA, researchers can obtain information regarding data distribution patterns, the presence of outliers, and the characteristics of both numerical and categorical variables. EDA is performed prior to algorithm selection and implementation, whether on structured or random data. In this study, EDA was performed through data visualization using Box Plots and Count Plots. Box Plots were used to display the minimum value, first quartile, median, third quartile, and maximum value, as well as to identify potential outliers in numerical variables. Meanwhile, Count Plots were used to display the frequency distribution of categorical variables, including the class distribution of the target variable.

2.3. Preprocessing

Preprocessing plays a crucial role in improving data quality, reliability, and consistency, thereby supporting more optimal machine learning model performance. Through the process of removing outliers, handling missing values, and checking for duplicate data, preprocessing can help improve model accuracy and make it easier for algorithms to read, use, and interpret the dataset. In this study, two preprocessing methods were used: data cleaning and class balancing. Both methods were applied based on the results of the EDA conducted previously.

a. a. Data Cleaning

   Data cleaning is a process performed to address missing values, noise, and inconsistencies in the data. This process aims to improve the quality and integrity of the data used in model training, so that the resulting model is expected to have better performance and prediction accuracy. At this stage, checks are performed for the presence of missing values, duplicate data, and extreme values or outliers based on information obtained from the EDA process.

b. Class Balancing

   Class balancing is performed to balance the class distribution in the dataset. This step is crucial to ensure that each class is proportionally represented, so

that the model is not biased toward the majority class. In this study, the SMOTE-NC oversampling technique was applied to balance the target class. SMOTE-NC is an oversampling technique that can generate new synthetic samples from the minority class by considering both numerical and categorical features. This technique was chosen because the dataset used contains many categorical features. With the application of SMOTE-NC, the model is expected to be able to learn patterns from the minority class more effectively without neglecting the characteristics of the categorical data.

2.4. Feature Selection

Feature selection is the process of selecting the features most relevant to the target class. This process is performed to reduce computational costs, minimize the memory requirements of the predictive model, and improve model generalization by ensuring that only features with significant contributions are used. In this study, feature selection was performed using the feature importance function available in CatBoost. This function is used to evaluate the importance of each feature relative to the target after the model has been trained using the dataset. The formula for calculating the feature importance value is shown in Equation (1).

$$\sum_{tree,leafs_F} \begin{matrix} feature\_importance_F = \\ (v_1 - avr)^2 . c_1 \\ + (v_2 - avr)^2 . c_2 \end{matrix} \quad (1)$$

Description:

a. avr displays the average value of the formula based on the weights of the objects on the left and right leaves.
b. c1 and c2 represent the total weights of the objects in the left and right leaves, respectively. If the weights are not specified in the dataset, these values are equal to the number of objects in each leaf.
c. v1 and v2 represent the formula values in the left and right leaves, respectively.

2.5. Hyperparameter Tuning

Hyperparameter tuning is a technique used to adjust the parameters in machine learning or deep learning algorithms with the aim of significantly improving model performance. This study uses Hyperopt to perform the hyperparameter tuning process. Hyperopt is a Python library developed to optimize hyperparameters in machine learning algorithms. The Tree-structured Parzen Estimator (TPE) method is used by Hyperopt in the process of searching for the best hyperparameters. In this method, the probability of hyperparameters is defined using two density functions, namely l(x) and g(x), as shown in Equation (2).

$$p(x|y) \begin{cases} l(x) & if\ y < y^* \\ g(x) & if\ y \geq y^* \end{cases} \quad (2)$$

Description:

a. l(x) is the density estimated using the observation x(i) with a loss value f(x(i)) that is lower than y*.
b. g(x) is a density function constructed using other observations.
c. The TPE algorithm selects y* as the γ-quantile of the observed y values, such that $p(y < y^*) = \gamma$. The hyperparameter configuration that yields the best value is then evaluated against the objective function to obtain the most optimal combination of parameters.

2.6. Modeling

Modeling is the stage of applying algorithms to discover, identify, and form patterns found in the research data. In this study, the algorithm used is CatBoost. CatBoost is an algorithm developed from the Gradient Boosting Decision Tree (GBDT) by the Russian technology company Yandex in 2017. The primary objective of this algorithm is to combine several weak learners through a loss function optimization process to produce an optimal model. The mechanism of the CatBoost algorithm is illustrated in Figure 2.

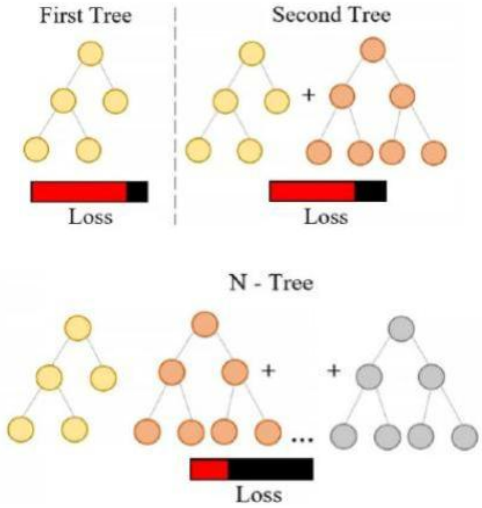


Figure 2. The CatBoost Mechanism

CatBoost has a unique ability to automatically handle categorical data using a statistical approach. By optimizing input parameters, CatBoost is able to reduce the risk of overfitting without requiring special processing of categorical features. Additionally, CatBoost uses random permutations on categorical data and calculates the average label, so it does not rely on binary conversion processes such as one-hot encoding.

2.7. Model Evaluation

Model evaluation is the process of measuring the performance of a built model. In this study, a confusion matrix was used to evaluate the performance of the CatBoost model. A confusion matrix is a table that shows the number of test data points classified correctly and incorrectly, and can therefore be used to assess the effectiveness of a classification system. A confusion matrix has four main terms, namely:

a. True Positive (TP) refers to the number of data points that actually fall into the category of potential customers and are predicted to be potential customers.

b. True Negative (TN), which is the number of data points that actually fall into the non-prospective customer category and are predicted to be non-prospective customers
c. False Positive (FP): the number of data points that actually fall into the category of non-potential customers but are predicted to be potential customers.
d. False Negative (FN): the number of data points that actually fall into the category of potential customers but are predicted to be non-potential customers.

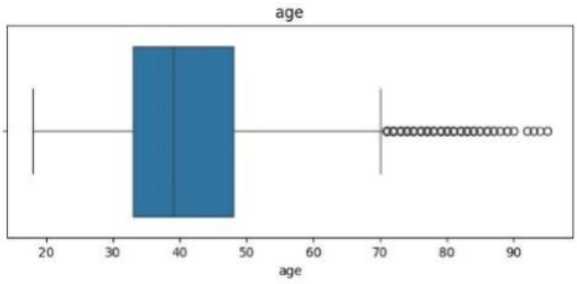
Based on these four components, a confusion matrix is used to calculate accuracy, sensitivity, and specificity. Accuracy indicates the proportion of correct predictions relative to the entire test dataset. Sensitivity indicates the model's ability to identify positive cases, while specificity indicates the model's ability to identify negative cases.

3. RESULTS

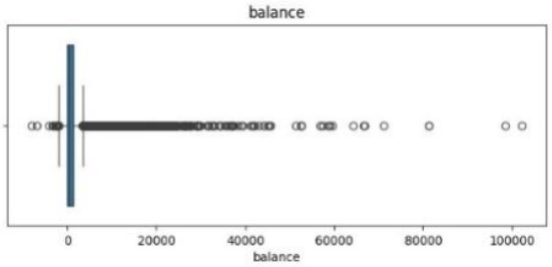
3.1. EDA (Exploratory Data Analysis)

Exploratory Data Analysis was conducted by visualizing the dataset using box plots and count plots. Box plots were used to identify outliers in numerical features, while count plots were used to determine the class distribution of the target variable. This step provided an initial overview of data quality, distribution patterns, and potential issues that needed to be addressed before building the model.

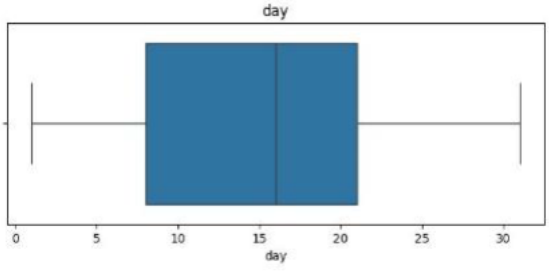
a. Box Plot



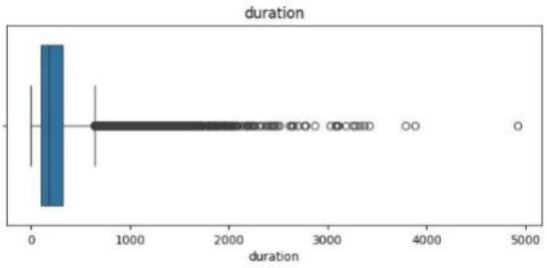
a. age feature



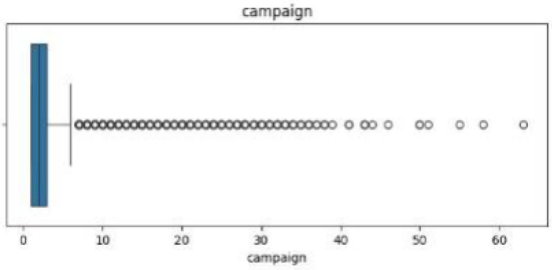
b. balance feature



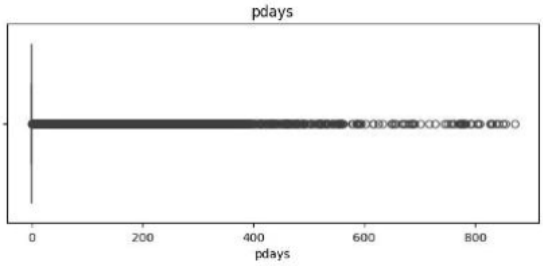
c. day feature



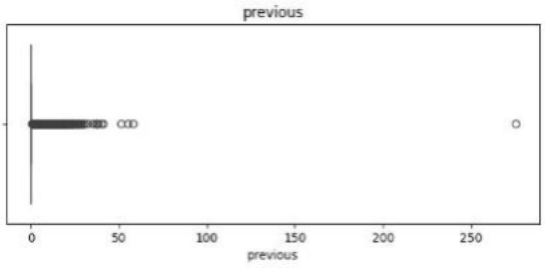
d. duration feature



e. campaign feature



f. pdays feature



g. Previous feature

Figure 3. Box Plot Visualization of Numeric Features

Based on the box plot visualization in Figure 3, several important insights were identified. For the age feature, there are some customer records with ages approaching 100 years. This indicates the presence of elderly customers in the dataset; therefore, this information can still be considered as long as it remains within reasonable limits. For the duration feature, negative values in seconds were found. These values are invalid because communication duration cannot be negative, so they need to be addressed during the data cleaning stage. Additionally, for the "previous" feature, extremely high values were found, approaching 300. This value is considered unrealistic in the context of the number of previous contacts with

the customer, so it has the potential to be an outlier that could affect the model's learning process.

b. Count Plot

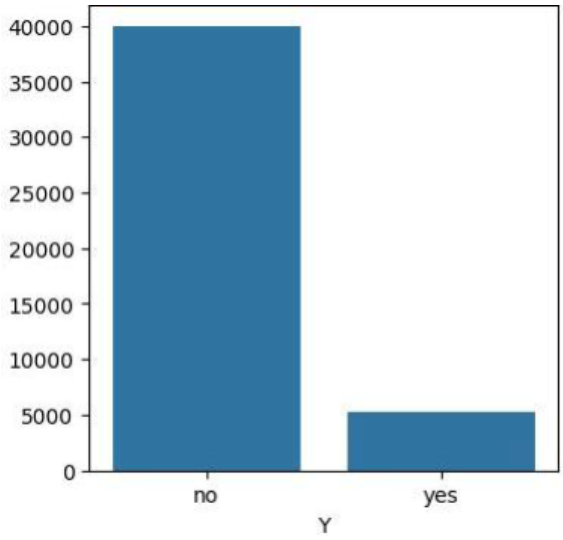


Figure 4. Distribution of the "Y" Feature Class

Based on the count plot visualization in Figure 4, there appears to be an imbalance in the class distribution of the target variable. The number of customers who did not make a deposit is significantly higher than the number of customers who did. This class imbalance can cause the model to favor the majority class and perform less optimally in predicting the minority class. Therefore, it is necessary to apply oversampling techniques to balance the class distribution before the modeling process begins.

3.2. Preprocessing

a. Data Cleaning

The data cleaning stage focuses on handling outliers or extreme values that could potentially interfere with the model training process. Based on the analysis results, outliers were found in the duration and previous features. Table 2 shows the number of data points that were removed because they contained outliers in those two features.

Table 2. Data Deleted During the Data Cleaning Phase

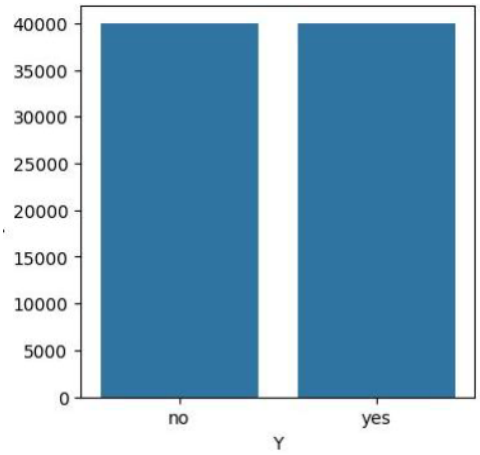
No	Reason	amount of data
1	Negative values in the duration field	3
2	The previous value is greater than 100	1

A total of 3 data rows were removed because they contained negative values in the duration feature, while 1 data row was removed because it had a previous value greater than 100. These values were considered invalid outliers in the context of the bank's marketing campaign. After the outlier removal process was completed, the remaining data consisted of 45,207 rows out of the initial total of 45,211 rows. These outliers may be caused by input errors, special conditions, or data anomalies. At this

stage, no issues related to missing values or duplicate data were found, so data cleaning focused only on handling outliers in the duration and previous features.

b. Class Balancing

Before class balancing was performed, the data distribution between the "yes" and "no" classes was still unbalanced, as shown in Figure 4. The number of samples in the "no" class was much larger than in the "yes" class. This situation can cause the model to produce predictions that are biased toward the majority class. To address this issue, the SMOTE-NC oversampling technique was used. This technique was chosen because it can handle datasets with a combination of numerical and categorical features. The results of the class balancing are shown in Figure 5.



Gambar 5. "Y" Feature Class After Oversampling

After applying the oversampling technique using SMOTE-NC, the number of samples in the minority class—the "yes" class—increased significantly, resulting in a more balanced distribution between the two classes. With a balanced class distribution, the model is expected to learn patterns from both classes more proportionally and achieve better predictive performance.

3.3. Feature Selection

Feature selection was performed using the feature importance function in CatBoost. Based on this process, 10 features with the highest importance scores relative to the target were identified. These features are considered the most relevant and influential in predicting whether a customer will make a deposit or not. The results of the feature selection are shown in Table 3.

Table 1. Number of Datasets

No	Fitur	Importance
1	Duration	23.034733423174092
2	Month	13.002639478005829
3	Contact	9.552350988843171
4	Day	8.613479074516862
5	Poutcome	8.010717868596675
6	Balance	6.790500599865537

7	Housing	6.014267236653641
8	Pdays	4.891514503880521
9	Age	4.166553246346829
10	Job	3.9649717785453573

Based on Table 3, the duration feature has the highest importance score, making it the most influential feature in the prediction process. Furthermore, the month, contact, day, and outcome features also contribute significantly to the prediction results. By using these 10 selected features, the model training process becomes more efficient because less relevant features are no longer used in the modeling process.

3.4. Hyperparameter Tuning

The hyperparameter tuning process was performed using the Hyperopt library by defining a search space for each hyperparameter to be optimized. The search space serves to limit the range of parameter values tested by the tuning algorithm in its search for the best combination. The hyperparameter search space used is shown in Figure 6.

```
# Mendefinisikan ruang pencarian hyperparameter
space = {
    'max_depth': hp.quniform('max_depth', 6,10, q=1),
    'l2_leaf_reg': hp.uniform('l2_leaf_reg', 0.1, 10),
    'learning_rate': hp.uniform('learning_rate', 0.01, 0.5),
    'subsample': hp.uniform('subsample', 0.5, 1.0)
}
```

Figure 6. Hyperparameter Search Space

Based on Figure 6, the parameters used in the tuning process include max_depth, l2_leaf_reg, learning_rate, and subsample. The max_depth parameter specifies the maximum depth of each tree built, with a range of values from 6 to 10. The l2_leaf_reg parameter specifies the L2 regularization level on each tree's leaf nodes, with tested values ranging from 0.1 to 10. The learning_rate parameter specifies the model's learning rate, with a range of values from 0.01 to 0.5. Meanwhile, the subsample parameter indicates the percentage of the dataset used in building each tree, with a range of values between 0.5 and 1.0. After the search space was determined, the tuning process was performed to find the best combination of hyperparameters based on the logloss evaluation metric. After 20 tuning iterations, the best combination of hyperparameter values was obtained, as shown in Table 4.

Table 4. Best CatBoost Hyperparameters

No	Hyperparameter	Best Value
1	max_depth	9.0
2	l2_leaf_reg	2.1509964758438764
3	learning_rate	0.15324979289799676
4	subsample	0.6258962666318186

This optimal combination of hyperparameters was used in the development of the CatBoost model. With this configuration, the model is expected to deliver optimal

classification performance in predicting whether customers will choose to open a deposit account.

3.5. Modeling

After completing the preprocessing, feature selection, and hyperparameter tuning stages, the next step is to build a classification model using the CatBoost algorithm. Before building the model, the 10 previously selected features were analyzed to identify categorical features. The identified categorical features are shown in Table 5.

Table 5. Identified Categorical Features

No	Fitur	Tipe Data
1	Month	Object
2	Contact	Object
3	Poutcome	Object
4	Housing	Object
5	Job	Object

These categorical features were then used as the value for the `cat_features` parameter in the CatBoost model. This parameter informs the model that the features are categorical, allowing CatBoost to process them directly without requiring one-hot encoding. Next, the dataset was split into training and test data with a 75:25 ratio. The training data consists of 59,877 rows and is used to train the model, while the test data consists of 19,959 rows and is used to evaluate the model's performance. After that, the CatBoost model was built using the 10 selected features and the optimal hyperparameter combination obtained in the previous stage.

3.6. Model Evaluation

The model was evaluated to assess the performance of CatBoost when combined with feature selection and hyperparameter tuning. The evaluation was conducted using a confusion matrix, as this method can show the number of data points correctly and incorrectly classified by the model. The results of the CatBoost model's confusion matrix in predicting the likelihood of customers opening a time deposit account are shown in Figure 7.

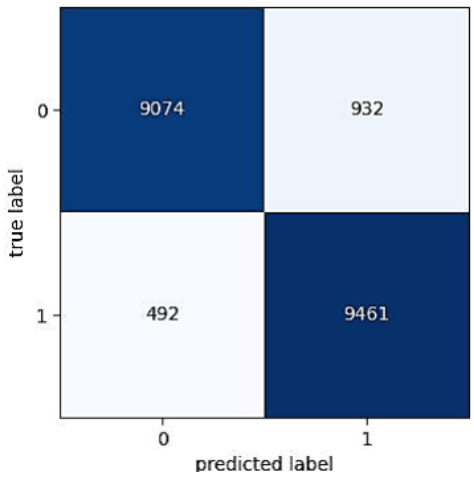


Figure 6. Hyperparameter Search Space

Based on the confusion matrix in Figure 7, the True Positive (TP) value is 9,461. This value indicates that the model successfully and accurately predicted 9,461 customers who actually opened a time deposit account. This figure demonstrates that the model is effective at identifying potential customers, thereby helping the bank design more targeted marketing strategies. The False Positive (FP) value of 492 indicates that there were 492 customers predicted to subscribe to a deposit account, but in reality did not. Although this can lead to less effective marketing costs, the number of FPs is relatively small compared to the TP. Thus, the model can still be considered quite efficient in reducing potential waste of promotional costs on customers who are less interested.

The False Negative (FN) value of 932 indicates that there are 932 customers who actually hold time deposits but were predicted by the model not to do so. This suggests that potential marketing opportunities may have been missed. Therefore, the bank needs to evaluate the characteristics of customers in the FN group so that marketing strategies can be further improved to reach potential customers who have not yet been optimally identified. Meanwhile, the True Negative (TN) value of 9,074 indicates that the model successfully and correctly predicted 9,074 customers who indeed do not subscribe to deposits. This result demonstrates the model's ability to identify customers who are less likely to respond to deposit offers. With this information, the bank can allocate marketing resources more efficiently and avoid promotions that are not well-targeted.

Based on the results of the confusion matrix, the accuracy was 92.8%, the sensitivity was 91.0%, and the specificity was 94.8%. These evaluation results indicate that the use of feature selection and hyperparameter tuning can reduce the complexity of the dataset while achieving high classification performance in the CatBoost model. Thus, this combination of methods is effective for predicting potential deposit customers.

In practical terms, the findings of this study can benefit banks in managing their deposit product marketing strategies. With a model accuracy of 92.8%, banks can segment customers more intelligently, target promotions at customers who are more likely to open deposit accounts, and reduce ineffective marketing costs. Additionally, this model can help bank staff gain preliminary insights into customers' interest trends regarding deposits, enabling faster and more personalized service. It can also support banks in designing deposit products that better align with customer needs. Customers predicted to have high potential can be offered special deals, relevant product information, or a more personalized marketing approach. Thus, this predictive model is not only useful for improving the effectiveness of marketing campaigns but can also strengthen the long-term relationship between the bank and its customers.

In addition, this model can support more data-driven business decision-making, such as estimating potential inflows from deposit products and planning the use of those funds more efficiently. However, this model should still be viewed as a decision-making tool, not as the sole basis for business decisions. Therefore, regular data updates and model evaluations are necessary to ensure the model remains relevant to changes in customer behavior and the dynamics of the banking market.

4. CONCLUSIONS

Based on the results of the study, it can be concluded that the CatBoost algorithm can be effectively used to predict potential deposit customers. The main advantage of CatBoost lies in its ability to handle categorical features directly without requiring one-hot encoding, thereby reducing data complexity and improving modeling efficiency. This is important because the bank marketing dataset used in this study contains a combination of numerical and categorical features. The use of feature selection based on feature importance successfully identified the 10 features most relevant to the prediction: duration, month, contact, day, outcome, balance, housing, pdays, age, and job. This feature selection helps reduce data dimensionality and focuses the model on the variables that contribute most significantly to customers' decisions to open a deposit account. In addition, the application of hyperparameter tuning using Hyperopt was able to produce the best combination of parameters, thereby optimizing the model's performance. The evaluation results show that the CatBoost model, combined with feature selection and hyperparameter tuning, achieved an accuracy of 92.8%, a sensitivity of 91.0%, and a specificity of 94.8%. These results indicate that the model has a strong ability to identify both potential customers and those unlikely to make deposits. Thus, the approach used in this study can assist the banking industry in developing more targeted, efficient, and data-driven deposit marketing strategies. Practically, this model can be utilized by banks to enhance the effectiveness of marketing campaigns, reduce costs associated with misdirected promotions, and strengthen personalized services for customers. However, the use of the model must be accompanied by periodic evaluations to ensure prediction accuracy remains maintained as customer behavior and market conditions evolve.

5. REFERENCES

Alexandropoulos, S.-A., Kotsiantis, S., & Vrahatis, M. (2019). Data preprocessing in predictive data mining. Knowledge Engineering Review, 34, e1.

Bach, J., & Böhm, K. (2024). Alternative feature selection with user control. International Journal of Data Science and Analytics, 1–23.

Bergstra, J., Yamins, D., & Cox, D. D. (2013). Hyperopt: A Python library for optimizing the hyperparameters of machine learning algorithms. SciPy, 13, 20.

Cerda, P., & Varoquaux, G. (2020). Encoding high-cardinality string categorical variables. IEEE Transactions on Knowledge and Data Engineering, PP, 1.

Cholissodin, I., & Soebroto, A. A. (2021). AI, machine learning & deep learning: Teori & implementasi.

Fajri, F. N., Tholib, A., & Yuliana, W. (2022). Application of machine learning algorithm for determining elective courses in informatics study program. Jurnal Teknik Informatika dan Sistem Informasi, 8(3), 485–496.

Fejza Ademi, V., Avdullahi, A., Tmava, Q., & Durguti, E. (2022). Analysis of the banking sector competition in Kosovo. Studia Universitatis "Vasile Goldis" Arad – Economics Series, 32, 2022–2054.

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. Journal of Big Data, 7(1), 94.

Hasanah, M. A., Soim, S., & Handayani, A. S. (2021). Implementasi CRISP-DM model menggunakan metode decision tree dengan algoritma CART untuk prediksi curah hujan berpotensi banjir. Journal of Applied Informatics and Computing, 5(2), 103–108.

Hudawi, A., Octavia, N., Elfandiono, A., Setiawan, A. B., Ghafur, A. A., & Susanto, A. E. (2021). Klasifikasi pemahaman santri dalam pembelajaran Kitab Kuning menggunakan algoritma C4.5 pohon keputusan di Pondok Pesantren Nurul Jadid. TRILOGI: Jurnal Ilmu Teknologi, Kesehatan, dan Humaniora, 2(3), 266–269.

Ibrahim, A. A., Ridwan, R. L., Muhammed, M. M., Abdulaziz, R. O., & Saheed, G. A. (2020). Comparison of the CatBoost classifier with other machine learning methods. International Journal of Advanced Computer Science and Applications, 11(11).

Jiménez-Cordero, A., & Maldonado, S. (2021). Automatic feature scaling and selection for support vector machine classification with functional data. Applied Intelligence, 51(1), 161–184.

Kumar, P. S., Kumari, A., Mohapatra, S., Naik, B., Nayak, J., & Mishra, M. (2021). CatBoost ensemble approach for diabetes risk prediction at early stages. In 2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON) (pp. 1–6).

Minarno, A. E., Mandiri, M. H. C., & Alfarizy, M. R. (2021). Klasifikasi COVID-19 menggunakan filter Gabor dan CNN dengan hyperparameter tuning. ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika, 9(3), 493.

Mukherjee, M., & Khushi, M. (2021). SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features. Applied System Innovation, 4(1), 18.

Nogales, R. E., & Benalcázar, M. E. (2023). Analysis and evaluation of feature selection and feature extraction methods. International Journal of Computational Intelligence Systems, 16(1), 153.

Purbolaksono, M. D., Pratama, D. T. B., & Hamzah, F. (2023). Perbandingan Gini index dan Chi Square pada sentimen analisis ulasan film menggunakan Support Vector Machine classifier. JEPIN: Jurnal Edukasi dan Penelitian Informatika, 9(3), 528–534.

Puteri, A. N., Arizal, A., & Achmad, A. D. (2021). Feature selection correlation-based pada prediksi nasabah bank telemarketing untuk deposito. MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 20(2), 335–342.

Putri, V. V., Tholib, A., & Novia, C. (2023). Deteksi Kaggle bot account menggunakan deep neural networks. NJCA: Nusantara Journal of Computers and Its Applications, 8(1), 13–21.

Rafrastara, F. A., Supriyanto, C., Amiral, A., Amalia, S. R., Al Fahreza, M. D., & Ahmed, F. (2024). Performance comparison of k-nearest neighbor algorithm with various k values and distance metrics for malware detection. Jurnal Media Informatika Budidarma, 8(1), 450–458.

Riyyasy, M. R. A., Aghniya, W. N., & Tantyoko, H. (2023). Penerapan algoritma machine learning untuk memprediksi term deposit nasabah perbankan. LEDGER: Journal of

Informatics and Information Technology, 2(3), 145–156.

Saber, M., et al. (2022). Examining LightGBM and CatBoost models for wadi flash flood susceptibility prediction. Geocarto International, 37(25), 7462–7487.

Sagar, K. M. (2023). Customer deposit forecasting: Optimizing deposit predictions through CRM and machine learning integration. International Journal of Scientific Research in Engineering and Management, 7, 1–11.

Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory data analysis using Python. International Journal of Innovative Technology and Exploring Engineering, 8(12), 4727–4735.

Shudiq, W. J., As, A. H., & Rahman, M. F. (2020). Penentuan metode terbaik dalam menentukan jenis pohon pisang menurut tekstur daun menggunakan metode K-NN dan SVM. Jurnal Teknologi dan Manajemen Informatika, 6(2), 128–136.

Tholib, A., Agusmawati, N. K., & Khoiriyah, F. (2023). Prediksi harga emas menggunakan metode LSTM dan GRU. Jurnal Informatika dan Teknik Elektro Terapan, 11(3).

Wang, D., & Qian, H. (2023). CatBoost-based automatic classification study of river network. ISPRS International Journal of Geo-Information, 12(10).

Widiari, N. P. A., Suarjaya, I., & Githa, D. P. (2020). Teknik data cleaning menggunakan Snowflake untuk studi kasus objek pariwisata di Bali. Jurnal Ilmiah Merpati: Menara Penelitian Akademika Teknologi Informasi, 8(2), 137.

Xiao, W., Wang, C., Liu, J., Gao, M., & Wu, J. (2023). Optimizing faulting prediction for rigid pavements using a hybrid SHAP-TPE-CatBoost model. Applied Sciences, 13, 12862.

Zaki, A. M., Khodadadi, N., Lim, W. H., & Towfek, S. K. (2024). Predictive analytics and machine learning in direct marketing for anticipating bank term deposit subscriptions. American Journal of Business and Operations Research, 11(1), 79–88.