Journal of Applied Engineering and Technological Science
Vol 6(1) 2024: 48-59

DEVELOPMENT OF AN OPTIMIZED ENSEMBLE LEAST SQUARES MODEL FOR IDENTIFYING POTENTIAL DEPOSIT CUSTOMERS

Firman Aziz[1], Mutia Maulida[2]*, Jafar[3], Nurafni Shahnyb[4], Ampauleng[5], Norma Nasir[6]
Computer Science Study Program, Pancasakti University, Makassar, Indonesia[1]
Department of Information Technology, Universitas Lambung Mangkurat, Banjarmasin, Indonesia[2]
Language Education, Pancasakti University, Makassar, Indonesia[3]
Communication Science Study Program, Pancasakti University, Makassar, Indonesia[4]
Management Study Program, STIEM Bongaya, Makassar, Indonesia[5]
Mathematics Education, Universitas Negeri Makassar, Makassar, Indonesia[6]
firman.aziz@unpacti.ac.id[1], mutia.maulida@ulm.ac.id[2], jafarmahmud14@gmail.com[3]
nurafni.syahnyb@unpacti.ac.id[4], ampauleng@stiem-bongaya.ac.id[5], norma.nasir@unm.ac.id[6]

Received: 23 August 2024, Revised: 01 November 2024, Accepted: 02 November 2024
**Corresponding Author*

ABSTRACT
The banking sector faces significant challenges in effectively promoting its products and services. While direct marketing has proven to be a potent tool for customer acquisition, it often leads to customer dissatisfaction, thereby tarnishing the bank's reputation. Leveraging Business Intelligence (BI) technology offers a strategic advantage by enabling the classification and analysis of customer data, particularly for time deposit customers. This study presents the development and optimization of an Ensemble Least Squares (ELS) algorithm to enhance the classification of potential deposit customers. The proposed Ensemble Least Squares Support Vector Machine (ELS-SVM) algorithm demonstrated superior performance compared to traditional SVM and LS-SVM methods. Notably, the ELS-SVM achieved an average performance improvement of 10.04% over standard Support Vector Machine (SVM) techniques.
Keywords : *Business Intelligence, Bank Marketing, Classification, Potential Deposits Customers, Ensemble Least Square Support Vector Machine*

1. Introduction

Business Intelligence is currently very familiar among companies, which consists of strategies and techniques used to analyze business data to produce product-related information (Orjatsalo et al., 2024). Business Intelligence Technology provides a historical, current, and future view for businesses that includes reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics, and prescriptive analytics. Business Intelligence technology can handle large amounts of data to help identify, develop, and create new strategic business opportunities to facilitate data interpretation, identifying new opportunities, and implement effective strategies based on insights to provide a competitive business advantage and have long-term stability (Ajah et al., 2019).

The banking sector has started to apply Business Intelligence Technology by processing customer information, especially in the marketing sector. In general, the banking approach in introducing goods or services is by placing advertisements through television, radio, newspapers, etc, or specifically targeting customers or it is called direct bank marketing (Chan-Olmsted, 2019; Kapoor & Kapoor, 2021; Quayson et al., 2024; Rhay Vicerra et al., 2019). The application of direct bank marketing in introducing its products in the form of goods and services to customers via telephone, e-mail, and others is considered very effective in attracting customers. However, many customers feel annoyed, causing a negative view of the bank and also the costs and time spent by telemarketers (Hujić & Salihić, 2020; Kreituss et al., 2021). To overcome this problem, it is necessary to process data by classifying potential consumers based on widely available information (Chen et al., 2020; Dogra et al., 2022; Sen et al., 2020). In classification, there is a process of analyzing a set of data which results in a set of grouping rules based on the available data labels, which then becomes a classifier model used to obtain future information.

The studies examining the application of Business Intelligence Technology in the banking sector reveal several critical limitations. Firstly, many studies, such as (Zhuang & Yao, 2018), lack clear performance metrics for their data mining techniques, which hinders the assessment of their effectiveness and comparability with other methods. Additionally, the research by (Parlar & SK, 2017) highlights the use of feature selection methods like Information Gain and Chi-square without exploring their interactions or the potential impact of unconsidered features, potentially leading to suboptimal model performance. While (Ruangthong & S, 2015) successfully applied the SMOTE algorithm to address unbalanced datasets, the generalizability of their findings is limited, as their focus on specific algorithms like J48 may not apply across different contexts. Furthermore, the study by (Mutoi Siregar et al., 2020) acknowledges the need for more complex data collection to enhance predictive performance, underscoring the challenge of obtaining high-quality data that accurately represents the customer base.

The ANN approach in (Selma, 2020) achieved a high accuracy of 98.93%, but the unequal representation of subscribers versus non-subscribers raises concerns about potential bias towards the majority class, affecting practical applications. (Grzonka et al., 2016) pointed out the importance of randomness in bootstrap sampling for decision tree methodologies; however, their study may overlook the implications of sampling variations on the reliability of the models produced. Moreover, the use of a single data partition with unbalanced data in the studies by (Firman Aziz, 2020; Lawi, Aziz, et al., 2018; Lawi, Velayaty, et al., 2018) can lead to overfitting and less generalizable results, emphasizing the need for more robust validation methods. The sensitivity of single vector machine approaches to sample settings and parameters, as noted by (F Aziz et al., 2019), indicates a potential fragility in model performance, suggesting further tuning may be necessary for consistency. Additionally, some studies, like that of (Che et al., 2020), fail to identify influential attributes despite using advanced methods such as t-SNE and SVM, highlighting the importance of understanding which features significantly impact model predictions. Lastly, while some research aims for applicability across multiple industries, the unique characteristics of the banking sector necessitate further validation in diverse contexts. Overall, addressing these limitations is crucial for advancing research and enhancing practical applications in banking marketing.

The reliance on traditional classification methods without incorporating hybrid or ensemble techniques limits the potential to enhance predictive accuracy and robustness. Many studies inadequately address the integration of multiple algorithms to leverage each one's strengths, which can lead to models that may not perform optimally under varying data conditions. Additionally, the lack of exploration into the temporal dynamics of customer behavior and the impact of external factors on data trends presents another significant gap, as the banking environment is influenced by economic fluctuations and regulatory changes. Furthermore, while various studies highlight the importance of customer segmentation for targeted marketing, they often overlook the need for real-time data processing and adaptive modeling, which are essential for swiftly responding to market changes and customer needs. Addressing these limitations is crucial for advancing research and enhancing practical applications in banking marketing.

This study builds upon and addresses the limitations of previous research by proposing a hybrid method, specifically the Ensemble Least Square Support Vector Machine (ELSSVM), utilizing the Boosting algorithm, particularly Adaboost. A key advantage of the ensemble method is its ability to improve model accuracy by combining predictions from various base models. In this context, Adaboost serves to rectify the prediction errors generated by individual models, thus aiming to yield more accurate predictions compared to single approaches, including Support Vector Machine (SVM) and Least Square Support Vector Machine (LS-SVM), which are often sensitive to data variations. By adopting an ensemble approach, this research also seeks to reduce variance that can lead to overfitting, a common issue encountered when using a single model on large and complex datasets. This method enhances the model's resilience to individual errors, where if one model makes a mistake, other models within the ensemble can compensate for that error, improving overall performance.

Moreover, this study capitalizes on the strengths of each model by integrating them, creating a more holistic and adaptive approach while providing flexibility in selecting the most

suitable model for the challenges faced. Another important aspect is this study's focus on improving performance on imbalanced datasets, a significant challenge in many prior studies. By employing the Adaboost algorithm, the weights of models can be adjusted based on the difficulty of predicting specific classes, thereby enhancing performance on imbalanced data. The research also emphasizes the significance of identifying influential attributes using the Information Gain method, which aids in understanding which features contribute most to predictions, enhancing model interpretability and reducing data complexity by eliminating irrelevant features.

By analyzing the distribution of multiple data partitions, this research aims to evaluate the performance of the proposed method in more diverse contexts, providing a clearer picture of the model's capabilities in real-world situations. Overall, the hybrid approach suggested in this study is designed to tackle various limitations faced by earlier studies, with the aim of improving performance, robustness, and accuracy in relevant applications, particularly within the banking and direct marketing sectors.

2. Literature Review

2.1. Support Vector Machine

Support Vector Machine is used to solve the problem of pattern classification and is very suitable for use in data that can be linearly separated by maximizing the boundary plane (Hyperplane) to separate data into two classes in a feature space (Pisner & Learning, 2020; Shi, 2022). Figure 1 shows the possible hyperplane options for the best data sets and dividing fields with the maximum margins that are crossed between the two classes.

Support Vector Machine (SVM) is a powerful tool for pattern classification, particularly effective in scenarios where data can be linearly separated. By maximizing the hyperplane, SVM creates a boundary that effectively separates different classes within a feature space, making it suitable for various classification tasks, including customer classification in direct bank marketing. One of the key strengths of SVM is its effectiveness in high-dimensional spaces, allowing it to handle numerous features, such as demographic, behavioral, and transactional data. Additionally, SVM is robust to overfitting, as it aims to maximize the margin between classes, which leads to better generalization on unseen data. Its versatility is further enhanced by the kernel trick, enabling SVM to classify complex customer patterns effectively. However, SVM is not without its weaknesses. The computational complexity of training can be high, particularly with large datasets, potentially resulting in slower performance in real-time applications like immediate customer segmentation. Moreover, SVM is sensitive to noise and outliers in the training data, which can distort the hyperplane and negatively impact classification accuracy. While SVM primarily serves as a binary classifier, extensions such as One-vs-All can complicate its application in multi-class customer segmentation. In direct bank marketing, SVM can be utilized to classify customers based on their likelihood of responding to marketing campaigns, facilitating targeted marketing efforts that optimize resource allocation. It can also aid in churn prediction by identifying customers at risk of leaving the bank and help segment the customer base for personalized marketing. Furthermore, SVM can be applied in credit scoring to classify customers based on creditworthiness, assisting banks in making informed loan decisions.

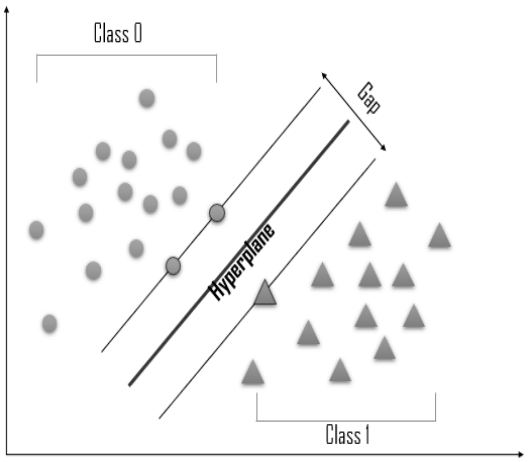

Fig. 1. Support Vector Machine Concept.

2.2. Least Square Support Vector Machine

Least Squares Support Vector Machine is an SVM development pioneered by Suykens and Vandewalle in 1999 (Heddam et al., 2022; Temeng et al., 2022). Least Squares Support Vector Machine produces lower error rates compared to Support Vector Machine because it performs an advanced classification by reprocessing incorrectly classified data using quadratic hyperplane. The shape of the Least Squares Support Vector Machine is shown with the following objective function:

$$\min \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n} {\xi_i}^2$$

Subject to

$$y_i\left(\varphi(\boldsymbol{x_i}).\boldsymbol{w^T} + b\right) = 1 - \xi_i, k = 1,2,\dots,N$$

Where ξ is the slack variable that determines the misclassification of the data sample. C>0 is the parameter that determines the amount of the penalty due to errors in data classification. k is the kernel function used.

Least Squares Support Vector Machine (LS-SVM) is a powerful extension of the traditional Support Vector Machine, offering notable advantages in the realm of customer classification, particularly within direct banking marketing. One of its primary strengths lies in its ability to produce lower error rates due to its unique approach of minimizing the least squares error rather than merely maximizing the margin. This characteristic allows LS-SVM to effectively reprocess incorrectly classified data, resulting in improved accuracy and robustness when handling complex datasets that are often encountered in banking environments. Furthermore, LS-SVM's reliance on a quadratic hyperplane enhances its flexibility in accommodating nonlinear relationships among features, making it particularly adept at capturing intricate patterns in customer behavior.

However, LS-SVM is not without its weaknesses. The method can be sensitive to outliers and noise within the training data, which can distort the hyperplane and negatively impact classification performance. Additionally, the computational complexity associated with LS-SVM increases with larger datasets, potentially leading to slower processing times in real-time applications such as direct marketing campaigns. This can pose a challenge when rapid decision-making is crucial for optimizing marketing strategies.

In practical applications, LS-SVM can significantly aid in identifying customer segments that are most likely to respond to marketing efforts, facilitating targeted outreach strategies. By classifying customers based on their predicted responses to promotional offers, banks can allocate resources more efficiently and personalize their marketing approaches, ultimately enhancing customer engagement and satisfaction. Moreover, LS-SVM can be employed in

churn prediction models, helping banks identify at-risk customers and implement retention strategies proactively.

2.3. Ensemble Least Square Support Vector Machine with AdaBoost Method

The Ensemble Least Squares Support Vector Machine (ELSSVM) combined with the AdaBoost method presents a powerful approach for customer classification in direct banking marketing. One of its key strengths is its ability to enhance classification accuracy through ensemble learning, which combines multiple weak learners to form a robust model. By leveraging AdaBoost, this method can focus on misclassified instances, adjusting their weights during training to improve the model's predictive power. This capability is particularly beneficial in a banking context, where customer data can be imbalanced and complex, as it allows the model to better identify and classify diverse customer segments based on their responsiveness to marketing initiatives.

However, ELSSVM is not without its limitations. The reliance on parameter tuning, such as determining the optimal gamma and sigma values, can be challenging and may require extensive experimentation. Additionally, while the AdaBoost method effectively addresses misclassifications, it may also amplify noise within the data if not managed carefully, potentially leading to overfitting. The computational complexity associated with ensemble methods can also pose challenges, especially when dealing with large datasets typical in banking, potentially impacting processing time and real-time applicability.

In practical applications, ELSSVM with the AdaBoost method can significantly improve targeted marketing efforts by accurately classifying customers based on their likelihood to respond to specific promotions. By effectively identifying high-potential customer segments, banks can optimize their marketing strategies, allocate resources more efficiently, and ultimately enhance customer satisfaction and loyalty. Moreover, this approach can facilitate better churn prediction models, allowing banks to proactively engage at-risk customers and improve retention rates.

Procedure Ensemble Least Square Support Vector Machine with AdaBoost Method

- Load Dataset
- Identify Label Attributes, class and amount of data.
- Determine the amount of training data and test data.
- Form a classification
  - Determining Gamma Value.
  - Determine the Value of Sigma.
  - Variable initialization, and initial weighting.
- Determine the number of iterations
  - Using the component learner algorithm (least squares support vector machine) to form a classification model.
  - Calculate the weight of misclassification
  - Update training data sample weights
  - Testing the model with test data
- Results from testing the model.

2.4. Performance Evaluation

Evaluation of the performance results of each classification is calculated based on 3 measurements, namely: Accuracy, Sensitivity, and Specificity based on the values of True Positive, False Negative, False Positive, and True Negative (Ibrahim et al., 2021; Tharwat, 2021).

True Positive (TP): The amount of data identified as Potential Customers and the results of their predictions of Potential Customers.

True Negative (TN): The amount of data identified as Non-Potential Customers and the results of their predictions Non-Potential Customers.

False Positive (FN): The amount of data identified as Non-Potential Customers but the results of the predictions are Potential Customers.

False Negative (FN): The amount of data data that is identified as Potential Customers but the results of the predictions are Non-Potential Customers.

Table 1 - Confuxion Matriks

	Actual: Potential Customers (Positive)	Actual: Non-Potential Customers (Negative)
Predicted: Potential Customers (Positive)	True Potential Customers (TP)	False Non-Potential Customers (FN)
Predicted: Non-Potential Customers (Negative)	False Potential Customers (FP)	True Non-Potential Customers (TN)

$$Acc = \frac{1}{N}(TP+TN)$$

3. Research Methods

3.1. Dataset

The data processed is bank direct marketing dataset data that can be accessed through the University of California at Irvine (UCI) Machine Learning Repository. The data came from a Portuguese bank, from May 2008 to June 2013. The choice of the Portuguese bank direct marketing dataset from the UCI Machine Learning Repository is highly relevant to the research objectives for several reasons. First, this dataset specifically focuses on bank marketing campaigns conducted over a significant period, from May 2008 to June 2013, allowing for a comprehensive analysis of customer behavior and responses to marketing efforts. This longitudinal aspect provides insights into changing customer preferences and the effectiveness of various marketing strategies over time, making it particularly valuable for understanding trends and patterns in bank marketing. Additionally, the dataset has been validated in previous studies, enhancing its credibility and reliability for predictive modeling. The validation by prior research suggests that the dataset is well-structured, with clear labels and relevant features necessary for effective classification of customers. This cleanliness is crucial for achieving accurate predictions regarding which customers are likely to subscribe to term deposits, a primary goal of the research.

Moreover, the Portuguese banking sector presents a unique context that may differ significantly from other regions or banks. Understanding customer behavior within this specific market can yield insights that are applicable to similar banking environments or inform marketing strategies tailored to particular demographic and economic conditions. By focusing on this dataset, the research can contribute to the broader understanding of how banks can effectively target and communicate with potential customers, thus enhancing marketing efficiency and customer satisfaction. The description of the dataset is shown in Table 2.

Table 2 - Description Dataset.

Attribute	Data type
Age	Numeric
Job	Categorical :Admin, Blue-collar, Enterpreneur, Housemaid, Management, Retired, Self-employed, Services, Student, Technician, Unemployed, Unknown
Martial	Categorical : Divorced, Married, Single, Unknown
Education	Categorical : Basic.4y, Basic.6y, Basic.9y, High.school, Illiterate, Professional.course, University.degree, Unknown
Default	Categorical : No, Yes, Unknown
Housing	Categorical : No, Yes, Unknown
Loan	Categorical : No, Yes, Unknown
Contact	Categorical : Celular, Telephone
Month	Categorical : Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec
Day_of_week	Categorical : Mon, Tue, Wed, Thu, Fri
Duration	Numeric
Campaign	Numeric
Pdays	Numeric
Previous	Numeric
Poutcome	Categorical : Failure, Nonexistent, Success
Y	Binary : Yes, No

3.2. Normalization Data

Normalization of the dataset is carried out using the SMOTE (Synthetic Minority Over-sampling Technique) method to address the class imbalance issue in customer data (Mujahid et al., 2024). In this context, there is a significant difference between the number of customers who subscribe to deposits (4,640) and those who do not (36,540), which can cause the classification model to lean towards predicting the majority class (customers who do not subscribe). By applying SMOTE, the normalization process involves creating synthetic samples from the minority class, namely potential deposit customers. SMOTE works by selecting customers from the minority class and then generating new data based on interpolation between existing customers. This method adds variability to the minority class without duplicating existing data, thereby creating a more balanced dataset. After normalization, the dataset is transformed to a total of 9,280, with the number of data points for both classes—potential deposit customers and non-potential customers—equalized to 4,640 each. In this way, the classification model developed will have a better chance of accurately predicting both classes, thereby enhancing the efficiency and reliability of the model in identifying customers who are likely to subscribe to deposits.

The division of the dataset into training and testing sets is a crucial step in the machine learning process, as it directly influences the model's performance and its ability to generalize. In this study, an 80:20 ratio is utilized for the training and testing sets, allowing 80% of the data to be used for training, which ensures the model can effectively learn from a diverse range of examples. This is particularly important for complex models like the Ensemble Least Squares Support Vector Machine (ELSSVM), which requires sufficient data to accurately capture underlying patterns. The remaining 20% serves as a robust testing set, providing an adequate sample for evaluating the model's performance and ensuring its ability to generalize to new data. A smaller testing set may not adequately represent the dataset's variability, potentially leading to misleading performance metrics. This 80:20 split aligns with standard practices in machine learning, offering a reliable evaluation of model performance and instilling confidence in its application in real-world scenarios, particularly in banking and direct marketing contexts.

To further enhance the model's robustness, cross-validation techniques will be employed to prevent overfitting and ensure better generalization of the machine learning model. By dividing the training dataset into multiple subsets (or folds) and training the model on these subsets while validating it on the remaining data, cross-validation provides a more reliable estimate of model performance. This methodological rigor allows for the identification of optimal hyperparameters and aids in assessing the model's ability to generalize to unseen data. This 80:20 split, coupled with cross-validation, aligns with standard practices in machine learning, offering a reliable evaluation of model performance and instilling confidence in its application in real-world scenarios, particularly in banking and direct marketing contexts.

4. Results and Discussions

The overall calculation results of the information gain for each attribute can be seen in Table 3. Accordingly, the highest information gain value is the ‘Duration’ attribute that means the most influential attribute on the data processed in determining the target class. The identification of the 'Duration' attribute as the most influential variable in determining the target class has significant implications for predictive modeling, particularly in the context of direct bank marketing. This finding suggests that the length of communication with potential deposit customers plays a crucial role in influencing their decision to subscribe to deposit products. From a theoretical perspective, this aligns with established marketing principles that emphasize the importance of engagement in consumer decision-making. For instance, a longer duration of communication may indicate more in-depth discussions regarding the benefits of the product, handling of customer inquiries, and relationship building, which can enhance the likelihood of conversion. According to the Elaboration Likelihood Model (Lin et al., 2019), when individuals are provided with more information and time to process it, they are likely to form more positive attitudes toward the marketed product. In this context, if bank representatives spend more time explaining the benefits of subscribing to a deposit, customers are more likely to feel informed and, therefore, more inclined to commit.

Moreover, literature on customer engagement highlights that longer interactions often lead to increased trust and satisfaction, both of which are critical components in customer decision-making (Busalim & Society, 2021). In the banking sector, where customers may have reservations about financial products, building trust through extended communication may be a key factor in converting potential customers into actual ones. Additionally, research on customer behavior indicates that interaction duration can serve as a proxy for the quality of the customer experience (Omoregie et al., 2019). Therefore, marketing strategies that prioritize longer and more meaningful interactions with customers may not only improve conversion rates but also enhance customer satisfaction and loyalty in the long run.

Table 3 - Information Gain Calculation Results

Attribute	Information Gain
Duration	0.1094127
Pdays	0.0444837
Poutcome	0.0438343
Month	0.0380966
Previous	0.0277329
Age	0.0184393
Contact	0.0168012
Job	0.0142231
Default	0.0083306
Campaign	0.0042851
Education	0.0034476
Marital	0.0020686
Day of week	0.0004645
Housing	0.0000997
Loan	0.0000193

The testing process utilizes MATLAB to implement the performance of the applied methods. The testing methodology involves using two types of test data: one with imbalanced categories and the other with balanced categories. The performance results for each method, based on both imbalanced and balanced data, are presented in the respective results. For the balanced dataset, 80% of the overall data is allocated for training, which corresponds to 7,424 data points, while the test data remains consistently at 9,280 data points. The purpose of this configuration is to assess the impact of the training data size on evaluating the performance of the resulting model.

The results of the total average obtained by the proposed ELS-SVM method using AdaBoost has the highest percentage compared to SVM and LS-SVM methods. This proves that the ELS-SVM using AdaBoost is better than other methods. The results obtained found that the selection of the kernel greatly affects the performance of each proposed method. Where in this research, the best kernel is linear kernel. The superior performance of the Ensemble Least Square Support Vector Machine (ELS-SVM) using AdaBoost compared to traditional Least Square Support Vector Machine (LS-SVM) and Support Vector Machine (SVM) can be attributed to several key factors. ELS-SVM benefits from an ensemble learning approach, which combines predictions from multiple base learners (LS-SVMs), enhancing robustness and reducing the likelihood of overfitting by leveraging the strengths of different models. Additionally, the AdaBoost algorithm assigns varying weights to training instances based on previous classification accuracy, enabling the model to focus more on difficult-to-classify instances in subsequent iterations. This capability improves accuracy, particularly in imbalanced datasets where certain instances may be harder to classify correctly. The choice of kernel function also plays a crucial role; in this study, the linear kernel was found to be the most effective, as it allows for simpler decision boundaries that are advantageous for linearly separable data. Moreover, ELS-SVM's handling of class imbalance is particularly adept, as it focuses more on the minority class during training through adaptive weighting, resulting in improved overall performance. Finally, the research indicated that the number of K-folds in cross-validation significantly affects model performance; while more folds generally enhance validation robustness, too many can lead to decreased performance due to reduced training data per fold.

Table 4 - The Results Accuracy Of Data Classification Are Imbalanced

Training Data: Test Data	Kernel Function	K-Fold	Accuracy (%)		
			SVM	LS-SVM	ELS SVM

Training Data: Test Data	Kernel Function	K-Fold	Accuracy (%) SVM	LS-SVM	ELS SVM
7424:1856	Linear	5	94.23	94.72	95.15
		10	92.94	94.59	94.99
		50	92.62	94.33	94.93
		100	93.59	94.72	94.51
	Gaussian	5	76.81	82.87	89.58
		10	76.81	82.87	89.58
		50	76.81	82.87	89.58
		100	76.81	82.87	89.58
	Polynomial	5	80.05	83.92	91.77
		10	80.03	73.11	87.74
		50	92.4	57.09	93.35
		100	75.74	55.14	83.06

Table 5 - The Results Accuracy of Data Classification Are Balanced

Training Data: Test Data	Kernel Function	K-Fold	Accuracy SVM	LS-SVM	ELS-SVM
7424:9280	Linear	5	97.97	97.47	97.47
		10	96.68	97.34	97.31
		50	96.36	97.08	97.25
		100	97.33	97.47	96.83
	Gaussian	5	80.55	85.62	91.9
		10	80.55	85.62	91.9
		50	80.55	85.62	91.9
		100	80.55	85.62	91.9
	Polynomial	5	83.79	86.67	94.09
		10	83.77	75.86	90.06
		50	96.14	59.84	95.67
		100	79.48	57.89	85.38

Our research findings provide a significant contribution to existing knowledge in the field of predicting customer behavior for deposit subscriptions in the context of bank marketing. Several key relationships and differentiating factors can be highlighted, namely our emphasis on the importance of the attribute 'Duration,' which is in line with previous studies such as (Mutoi Siregar et al., 2020; Parlar & SK, 2017; Zhuang & Yao, 2018), who also recognize the importance of certain attributes in predicting customer behavior. However, our study uniquely selected 'Duration' as the most influential attribute, providing a clear focus that differentiates it from other studies. Additionally, the alignment of our findings with existing literature reinforces the reliability of our results, suggesting that 'Duration' is a consistent predictor across different contexts.

The introduction of the ELS-SVM method using AdaBoost as a predictive model demonstrated innovation and effectiveness in dealing with imbalanced data sets. Although studies such as (Ruangthong & S, 2015; Selma, 2020) have addressed the imbalance problem through methods such as SMOTE and Artificial Neural Networks, our proposed approach stands out for its superior performance, especially with linear kernels, as revealed in the comparison with the SVM method and LS-SVM. This further highlights the uniqueness of our methodology and its potential practical application. Additionally, our study contributes to the understanding of the impact of the number of K-folds on model performance, a factor that has not been widely explored in previous research. Recognition of the influence of K-folds on the ELS-SVM method provides valuable insight into the robustness of the model, which differentiates our study from others. Overall, our research not only reinforces the importance of certain attributes in predicting customer behavior but also introduces an innovative method that outperforms existing approaches. The focus on 'Duration,' emphasis on handling imbalanced data sets, and exploration of K-folds contribute new insights to the field, making our research a valuable addition to the existing literature.

Furthermore, the LS-SVM method shows better performance compared to the ordinary SVM method for the case of imbalanced test data, based on the average classification results obtained. It is found that the performance of the SVM method is very good when it has a small amount of test data and decreases when the amount of test data is large. Meanwhile, the average

performance of LS-SVM and ELS-SVM is not affected when the number of test data is small or large. However, there is a decrease in performance when the amount of training data approaches the amount of test data. The LS-SVM and ELS-SVM methods succeeded in overcoming the shortcomings of the SVM method when the amount of test data was small, with significant improvements observed. The proposed ELS-SVM method has succeeded in improving the performance of the LS-SVM due to its mechanism of re-weighting classified data and re-training at each iteration to the specified iteration limit. Although ELS-SVM demonstrates strong performance, its computational time for model training can be higher than that of traditional SVM, which poses challenges in real-world applications where speed and efficiency are essential. Furthermore, increasing the number of iterations in this method does not necessarily guarantee improved performance of the component learners, highlighting the need for further development in future work. This interplay between our findings and those from previous studies underscores the ongoing relevance of exploring advanced machine learning techniques to enhance predictive accuracy in banking and marketing contexts.

The findings of this research have significant practical implications for bank marketing strategies, particularly in enhancing the effectiveness of marketing campaigns and making better-informed decisions regarding customer behavior. By identifying 'Duration' as the most influential attribute in predicting potential deposit subscribers, banks can focus on designing more targeted marketing strategies. For instance, they can segment customers based on the duration of their interactions or service usage, allowing them to target those most likely to subscribe to deposit products with tailored promotions and detailed information on the benefits of investing in deposits. Additionally, the application of the proposed ELS-SVM model, which has proven effective in addressing imbalanced datasets, can enhance the accuracy of future customer behavior predictions. This model can be integrated into customer relationship management (CRM) systems to analyze customer data in real-time and provide actionable recommendations. Furthermore, banks can develop better retention strategies by leveraging insights about influential attributes; for example, if customers exhibit long interaction durations but have not yet subscribed to deposits, banks can proactively approach these customers to convert them into deposit subscribers. To maximize these benefits, it is recommended that banks integrate various data sources to gain comprehensive insights into customer behavior, including transaction data, customer service interactions, and survey feedback. Training marketing staff on how to utilize data analysis tools and predictive modeling, including ELS-SVM, will enable them to design data-driven campaigns and tailor their approaches to customers more effectively. Additionally, banks should implement monitoring systems to continuously assess campaign effectiveness and adjust strategies based on data analysis results, ensuring responsiveness to changes in customer behavior and market demands. Investing in advanced analytics technology will also be crucial for enabling banks to identify customer behavior patterns and make more informed decisions. By applying these recommendations, banks can significantly enhance their marketing strategies, ultimately contributing to revenue growth from deposit products and increased customer satisfaction.

5. Conclusion

This paper has proposed an Ensemble Least Squares Support Vector Machine (ELS-SVM) method uses AdaBoost to classify potential deposit customers. The focus is to improve the performance of SVM that sensitively to the sample and parameter settings. Therefore, the paper has also investigated the LS-SVM to overcome the sensitivity, however, the LS-SVM has a problem with the randomness of the sample and thus the development is tackled by ensemble technique using AdaBoost method. The proposed ELS-SVM has succeeded in increasing the performance of the SVM and LSSVM methods with an average percentage performance of 10.04%. It can be concluded that the proposed ELS-SVM method successfully overcomes the weaknesses of the SVM and LS-SVM methods. The research has a limitation concerning the number of iterations. In future work, it would be interesting to explore the impact of accuracy fluctuations resulting from the iterations using the proposed Ensemble Least Square Support Vector Machine (ELS-SVM) method.

References

Ajah, I., computing, H. N.-B. data and cognitive, & 2019, undefined. (2019). Big data and business analytics: Trends, platforms, success factors and applications. *Mdpi.Com*. https://doi.org/10.3390/bdcc3020032

Aziz, F, Lawi, A., & Budiman, E. (2019). Increasing Accuracy of Ensemble Logistics Regression Classifier by Estimating the Newton Raphson Parameter in Credit Scoring. *Ieeexplore.Ieee.Org*. https://doi.org/10.1109/CAIPT.2017.8320700

Aziz, Firman. (2020). Klasifikasi Pelanggan Deposito Potensial menggunakan Ensembel Least Square Support Vector Machine. *Journal of System and Computer Engineering*, *1*(1), 1. http://journal.unpacti.ac.id/index.php/JSCE/article/view/80

Busalim, A., & Society, F. G. (2021). Customer engagement behaviour on social commerce platforms: An empirical study. *Elsevier*. https://www.sciencedirect.com/science/article/pii/S0160791X19307481

Chan-Olmsted, S. M. (2019). A Review of Artificial Intelligence Adoptions in the Media Industry. *JMM International Journal on Media Management*, *21*(3–4), 193–215. https://doi.org/10.1080/14241277.2019.1695619

Che, J., Zhao, S., & Li, Y. (2020). Bank telemarketing forecasting model based on t-SNE-SVM. *Scirp.Org*. https://www.scirp.org/journal/paperinformation.aspx?paperid=100260

Chen, R., Dewi, C., Huang, S., Data, R. C.-J. of B., & 2020, undefined. (2020). Selecting critical features for data classification based on machine learning methods. *Springer*, *7*(1). https://doi.org/10.1186/s40537-020-00327-4

Dogra, V., Verma, S., Chatterjee, P., Shafi, J., Choi, J., & Fazal Ijaz, M. (2022). A Complete Process of Text Classification System Using State-of-the-Art NLP Models. *Wiley Online Library*, *2022*. https://doi.org/10.1155/2022/1883698

Grzonka, D., Suchacka, G., in, B. B.-I. S., & 2016, undefined. (2016). Application of selected supervised classification methods to bank marketing campaign. *Yadda.Icm.Edu.Pl*, *5*(1), 36–48. https://yadda.icm.edu.pl/yadda/element/bwmeta1.element.desklight-3c731462-b2de-4c6d-9e43-95ccf418785f

Heddam, S., Ptak, M., Sojka, M., Kim, S., Malik, A., Kisi, O., & Zounemat-Kermani, M. (2022). Least square support vector machine-based variational mode decomposition: a new hybrid model for daily river water temperature modeling. *Environmental Science and Pollution Research*, *29*(47), 71555–71582. https://doi.org/10.1007/S11356-022-20953-0

Hujić, N., & Salihić, F. (2020). *Marketing in tourism-direct marketing as marketing communications technology*. https://www.ceeol.com/search/article-detail?id=871923

Ibrahim, D., Elshennawy, N., & And, A. S. (2021). Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. *Elsevier*. https://www.sciencedirect.com/science/article/pii/S0010482521001426

Kapoor, R., & Kapoor, K. (2021). The transition from traditional to digital marketing: a study of the evolution of e-marketing in the Indian hotel industry. *Worldwide Hospitality and Tourism Themes*, *13*(2), 199–213. https://doi.org/10.1108/WHATT-10-2020-0124/FULL/HTML

Kreituss, I., Vasiljeva, T., & Rokjane, B. (2021). Factors Important for Banks in Attracting and Retaining Customers. *Lecture Notes in Networks and Systems*, *195*, 691–702. https://doi.org/10.1007/978-3-030-68476-1_64

Lawi, A., Aziz, F., & Syarif, S. (2018). Ensemble GradientBoost for increasing classification accuracy of credit scoring. *Proceedings of the 2017 4th International Conference on Computer Applications and Information Processing Technology, CAIPT 2017*, *2018-January*, 1–4. https://doi.org/10.1109/CAIPT.2017.8320700

Lawi, A., Velayaty, A. A., & Zainuddin, Z. (2018). On Identifying Potential Direct Marketing Consumers using Adaptive Boosted Support Vector Machine. In K. Bali (Ed.), *Proceedings of the 2017 4th International Conference on Computer Applications and Information Processing Technology, CAIPT 2017* (pp. 1–4). https://doi.org/10.1109/CAIPT.2017.8320691

Lin, X., Featherman, M., Brooks, S. L., & Hajli, N. (2019). Exploring Gender Differences in

Online Consumer Purchase Decision Making: An Online Product Presentation Perspective. *Information Systems Frontiers*, *21*(5), 1187–1201. https://doi.org/10.1007/S10796-018-9831-1

Mujahid, M., Kına, E. R. O. L., Rustam, F., Villar, M. G., Alvarado, E. S., De La Torre Diez, I., & Ashraf, I. (2024). Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/S40537-024-00943-4

Mutoi Siregar, A., Faisal, S., Handayani, H. H., & Jalaludin, A. (2020). Classification Data for Direct Marketing using Deep Learning. *Scientific Jounal of PPI-UKM*, *7*(2). https://doi.org/10.27512/sjppi-ukm/se/a15052020

Omoregie, O. K., Addae, J. A., Coffie, S., Ampong, G. O. A., & Ofori, K. S. (2019). Factors influencing consumer loyalty: evidence from the Ghanaian retail banking industry. *International Journal of Bank Marketing*, *37*(3), 798–820. https://doi.org/10.1108/IJBM-04-2018-0099/FULL/HTML

Orjatsalo, J., Hussinki, H., & Stoklasa, J. (2024). Business analytics in managerial decision-making: top management perceptions. *Measuring Business Excellence*. https://doi.org/10.1108/MBE-09-2023-0130/FULL/HTML

Parlar, T., & SK, A. (2017). Using data mining techniques for detecting the important features of the bank direct marketing data. *International Journal of Economics and Financial Issues*, *7*(2), 692–696. https://dergipark.org.tr/en/pub/ijefi/issue/32035/354551?publisher=http-www-cag-edu-tr-ilhan-ozturk

Pisner, D., & Learning, D. S. (2020). Support vector machine. *Elsevier*. https://www.sciencedirect.com/science/article/pii/B9780128157398000067

Quayson, A., Issau, K., Gnankob, R. I., & Seidu, S. (2024). Marketing communications' dimensions and brand loyalty in the banking sector. *Revista de Gestao*, *31*(1), 115–132. https://doi.org/10.1108/REGE-10-2021-0191/FULL/HTML

Rhay Vicerra, R. P., James Loresco, P., Dadios, E. P., James Loresco, P. M., & Rhay PVicerra, R. (2019). Segmentation of lettuce plants using super pixels and thresholding methods in smart farm hydroponics setup. *Researchgate.Net*, *12*. https://www.researchgate.net/profile/Pocholo-Loresco-2/publication/334289031_Segmentation_of_Lettuce_Plants_Using_Super_Pixels_and_Thresholding_Methods_in_Smart_Farm_Hydroponics_Setup/links/5d2850e3a6fdcc2462d6b4d4/Segmentation-of-Lettuce-Plants-Using-Sup

Ruangthong, P., & S, J. (2015). Bank direct marketing analysis of asymmetric information based on machine learning. *Ieeexplore.Ieee.Org*. https://ieeexplore.ieee.org/abstract/document/7219777/

Selma, M. (2020). Predicting the success of bank telemarketing using Artificial Neural Network. *International Journal of Economics*. https://publications.waset.org/10010974/predicting-the-success-of-bank-telemarketing-using-artificial-neural-network

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised Classification Algorithms in Machine Learning: A Survey and Review. *Advances in Intelligent Systems and Computing*, *937*, 99–111. https://doi.org/10.1007/978-981-13-7403-6_11

Shi, Y. (2022). Support Vector Machine Classification. *Advances in Big Data Analytics*, 97–246. https://doi.org/10.1007/978-981-16-3607-3_3

Temeng, V. A., Arthur, C. K., & Ziggah, Y. Y. (2022). Suitability assessment of different vector machine regression techniques for blast-induced ground vibration prediction in Ghana. *Modeling Earth Systems and Environment*, *8*(1), 897–909. https://doi.org/10.1007/S40808-021-01129-0

Tharwat, A. (2021). Classification assessment methods. *Emerald.Com*. https://www.emerald.com/insight/content/doi/10.1016/j.aci.2018.08.003/full/html

Zhuang, Q., & Yao, Y. (2018). Application of data mining in term deposit marketing. *Iaeng.Org*. http://www.iaeng.org/publication/IMECS2018/IMECS2018_pp707-710.pdf