Predicting Customer Response to Bank Campaigns with Artificial Intelligence Algorithms

Gulay Cicek[1], and Efe Genişoğlu[2]
[1,2]Department of Software Engineering, Faculty of Engineering Architecture
Istanbul Beykent University, Sariyer, Istanbul, Turkey
[1]gulaycicek@beykent.edu.tr, [2]2203013061@student.beykent.edu.tr

***Abstract*—**In the study, five different machine and deep learning algorithms (Random Forest, XGBoost, MLP, Naive Bayes, SVM) were compared on the UCI Bank Marketing dataset in order to increase the efficiency of telemarketing campaigns in the banking sector. The models were trained under two different feature selection scenarios, PCA and MLP, and with a unique data balancing (SMOTE) method based on customer segmentation (K-Means).

As a result of the experiments, the highest performance was obtained by the Random Forest model with an F1-Score of 0.6126. The most important result of the study is that the applied balancing method causes over-learning in all tested models.

Keywords: Bank Marketing, Customer Response Prediction, Machine Learning, Deep Learning, Feature Selection, Imbalanced Data, K-Means.

I. Introduction

Banks use various marketing strategies to reach new or existing customers with their products or services. In parallel with the advancement in technology, marketing techniques have also evolved over time. There are different marketing methods such as digital marketing, email marketing, and telemarketing [1]. Although digital marketing is preferred with the widespread use of social media, direct marketing methods are still critical in the banking sector.

Direct marketing is the process of determining the characteristics of potential customers of services and products and promoting these products to the determined customer base. In the 1990s, there was a significant increase in the number of businesses that preferred direct marketing in the international market [2]. The main purpose of this method is to establish low-cost, two-way and direct communication with individual customers. In order to use this marketing method effectively, it is necessary to reveal the characteristics of the current customer profile and to anticipate the demands of potential customers. Direct marketing advertisements help to establish one-on-one contact with current and potential customers in order to get fast and measurable reactions. The feedback data received from each potential customer that makes up the target audience is stored and this data is used to acquire new customers [3] [4]. One of the most commonly used direct marketing methods is telemarketing. It is cheaper in terms of both time and money compared to face-to-face communication. This method is very advantageous for both customers and the bank. Customers can contact a representative or an artificial intelligence assistant who can solve their problems even outside of business hours. In addition, the data obtained from the telemarketing method can be used effectively [5].For example, a bank creates a special low-interest loan campaign for customers who do not have an account in order to attract new customers. In order to introduce this campaign, they call each customer and tell them the interest rates. Then, each response from the people is recorded in the system. This data obtained is used to shape the audience to be marketed in the future and new campaign strategies. Ultimately, constantly calling and trying to market to a person who will never be convinced will be a burden on the company and prevent it from reaching other customers. In addition, the person who is constantly called may drift further away from the brand. Therefore, it is very important to predict the characteristics of the potential customer in advance.

It is very important for call center employees to only talk to the target audience because other customers who are experiencing problems will also call the call center. Otherwise, the effectiveness of marketing calls will decrease. Also, customers who call with a problem expect their problems to be resolved as soon as possible. The longer this process takes, the more customers can be lost.

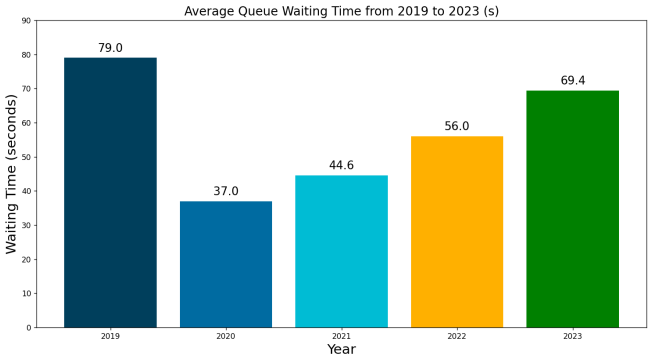

Fig. 1: Average waiting time in queue from 2019 to 2023 [6].

According to the study titled ”State of the Contact Center 2023-24” conducted by Ozonotel, the average waiting time for call centers between 2023-2024 was determined to be 69.4 seconds [6]. The time determined as 44.6 seconds in 2021 and 56 seconds in 2022 is increasing every year. Although there are various reasons for this, it is important to communicate only with the target audience in order to reduce these times.

Since each customer base has different habits, special marketing methods can be used for each. For example, while digital banking is common among young people, older customers may prefer different methods. 61% of young people under the age of 25 stated that they find telemarketing annoying [7]. In another study, individuals between the ages of 25-54 stated that they receive telemarketing calls frequently. It was found that this situation increases the tendency to respond negatively in the mentioned age group[8]. In a study conducted by Altuntaş and Boydak in 2022, Generations Z and Y prefer digital channels and a personalized advertising

experience, while they find traditional methods annoying [9]. In particular, Generation Z has shown little interest in traditional methods such as telemarketing, as proven by a study conducted by Ramazan Nacar and Ozan Karaarslan in 2024 [10]. Another important criterion is the customer's profession. While managers or white-collar employees respond more positively to campaigns, this rate is lower for blue-collar employees [11].This situation is directly related to the income level of the person. "Are they a homeowner?" "What is their current bank balance?" According to the study published by the OECD in 2022, people's financial moves are affected by their monthly income, education level, employment status and age [12]. Another factor that determines a person's financial situation is whether they have debt or not. It has been observed that individuals who already have a credit debt are more likely to respond negatively to marketing campaigns in order to avoid further debt [11] [13]. Similarly, individuals who are married are more likely to respond negatively to marketing campaigns to avoid further financial obligations [14]. As a result, studies in the literature have shown that married people and those with poor financial situations are less likely to respond to marketing campaigns. Young people, on the other hand, are generally more likely to respond negatively to telemarketing campaigns than older people because they find them annoying.

Although the financial situation of individuals is important, the method by which banks communicate with the target audience, how long the persuasion process takes, and whether the audience is correct are also determining factors. It has been observed that calls made from mobile phones have higher acceptance rates than calls made from landlines. In addition, the longer the talk time, the more likely the call will be to have a positive result [14] .

The aim and original contributions of the study are as follows:

- In order for bank campaigns to reach the best audience, it is aimed to predict the positive or negative response of customers using artificial intelligence methods.
- UCI Bank Marketing dataset was used [15]. This dataset of a Portuguese bank consists of 45,211 data. 88.3% of the customers answered no and 11.7% answered yes.
- In order to ensure originality, the K value of the K-Means method was determined by the Silhouette coefficient and customers were grouped according to this value. Then, each group was balanced with SMOTE.
- Random Forest (RF), Support Vector Machine (SVM) and XGBoost were used among machine learning methods. Multi-Layer Perceptron (MLPClassifier) was used among deep learning methods.
- Python is preferred for the implementation of machine learning, deep learning and other algorithms.

II. Literature Review

Many studies in the literature have used machine and deep learning methods to predict customer responses. Özge Cömert and Mesut Toğaçar used k-Nearest Neighbor (k-NN) and Bayesian algorithm to predict the success of telemarketing calls and automatically determined the hyperparameter values. The UCI dataset with 45,211 samples was used. However, no precautions were taken for the imbalance in this dataset. %94.68 accuracy, %62.96 sensitivity and %99.01 specificity values were obtained [16] .

Kevser Özdem and M. Ali Akcayol adopted the data mining method for the same data set in their study in 2021. The Apriori algorithm was used and no method was used to eliminate the imbalance in the data set. A lower rate was achieved with an accuracy rate of 87%. It is possible to achieve higher rates with SMOTE and different algorithms (SVM, RF, XGBoost, etc.) [17] .

Fahim Nasir and his colleagues used the UCI Bank Marketing dataset in their study in the same field, but unlike other studies, they used many balancing techniques (Random Oversampling (ROS), Random Undersampling (RUS), SMOTE, BorderlineSMOTE2, AdaSyn SMOTE). They examined these techniques separately with different combinations. After the examinations, the combination that gave the best result was the BorderlineSMOTE2 and XGBoost combination with an F1 Score of 0.87 [18].

In the study conducted by Muneeb Asif, UCI dataset was used again; Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF) and Artificial Neural Networks (ANN) algorithms were applied comparatively. By applying feature selection with Logistic Regression and LASSO methods, a separate simplified subset was created and the performance values were compared. While RF provided 90.63% accuracy in the full model, the result obtained with the subset was 90.56%. Successful results were obtained in both ways. The shortcoming of this study is that no steps were taken regarding data imbalance [14] .

Archit Verma used WEKA with Decision Tree, Naive Bayes, Multilayer Neural Network, Support Vector Machines, Logistic Regression and Random Forest algorithms in his study in 2019. While the unbalanced data set could not even exceed the 50% limit in F1 score; after SMOTE, Random Oversampling (ROS), Random Undersampling (RUS) methods were applied, this rate increased to 95.3%. The shortcomings of the study can be said to be limited to WEKA only (more advanced experiments could be done with Python, R etc.) and not using deep learning methods. [19].

Olatunji Apampa used classical methods in his study in 2016 via Orange 3.2, R-Studio, Excel and applied Manual Undersampling to balance the data set. While the highest AUC value was 0.678 when the data set was unbalanced, this value increased to 0.742 after the data set was balanced. This study was limited to classical methods and high values could not be obtained [20].

In the study conducted by Anamai Na-udom and his friends in 2022, they eliminated imbalances in the dataset with the random undersampling method. They compared J48 (Decision Tree), Random Forest, Random Tree, Naive Bayes algorithms in the WEKA environment. The most successful model was RF with 87.11%, followed by J48 with 87.10%. Although J48 is structurally less complex than RF, it achieved almost the same results [21].

Rutu S. Patel & Himanshu S. Mazumdar, in their 2018 study, created an Artificial Neural Network algorithm to predict customer responses. The main goal is to avoid over-learning and achieve high accuracy. 95% accuracy rate was achieved but different machine learning methods were not used [22].

In the study conducted by Alaa Abu-Srhan and his colleagues in 2019, the UCI dataset was examined with classical methods and the issue of visualization of the data was emphasized. A small part of the dataset with 4,521 samples was used. Since the dataset was unbalanced in its current state, this problem was resolved using SMOTE. The highest accuracy value was determined as 80.8% with Naive Bayes. The fact that the dataset used was small and that F1 scores were not specified can be considered as deficiencies [23].

In the study conducted by Shiueng-Bien Yang and Tai-Liang Chen in 2020, a new model called Uncertain Decision Tree (UDT) was developed. The model is supported by UGCA (Uncertain Genetic Clustering Algorithm) and thus the number of branches at each node is automatically optimized. Traditional methods C4.5,

Fuzzy Decision Tree (FDT), Genetic Fuzzy C4.5 (GCFDT) were used. UMA and UGCA were used to provide optimization. The UGCA model was the most successful model with an accuracy rate of 93.5%. The point that can be mentioned as a deficiency in this study is that the data set balancing method is not used [24] .

K. Wisaeng used a total of four methods in his study in 2013. Two of them are decision tree based J48-Graft and LAD Tree; the other two are machine learning based Radial Basis Function Network (RBFN) and Support Vector Machines (SVM). The highest accuracy rate belongs to SVM with 86.95%. This rate could have been increased if a balanced data set was selected or the balancing method was applied [25].

Amit Taneja used the Kaggle customer segmentation dataset in addition to the UCI dataset in his study in 2024. He applied Logistic Regression, Decision Tree, Random Forest and Support Vector Machines from classical machine learning methods, and the highest result among them was obtained with the Random Forest algorithm with 0.91. The fact that deep learning methods were not used and the sampling methods were not explained clearly can be considered as deficient [26].

In her study conducted in 2021, Mücella Özbay Karakuş examined the data set with a three-layer multilayer artificial neural network (MLNN). As a result, an accuracy rate of 94.3% was obtained. Although the data set was unbalanced and no separate precautions were taken in this regard, a very high rate was obtained. This rate could have been increased by using the SMOTE method. In addition, metrics such as F1 score and precision were not specified in the study, and the accuracy rate was limited [27].

In the study conducted by Youngkeun Choi and Jae Choi in 2022, variable effects were evaluated separately to predict customer responses. Some features in the data set were deemed necessary, while it was stated that some (education status, occupation, whether or not to own a house, marital status) would have a very low effect. Many studies in the literature have proven that such features positively or negatively affect the likelihood of people getting into more debt. Although it was said that the effect would be low, the study was conducted by taking these features into account and decision tree algorithms were applied. 78.4% accuracy rate was obtained. The shortcomings of the study can be listed as not taking precautions against the imbalance of the data set and not applying classical machine learning methods and being limited only to decision tree algorithms [28].

Chittem Leela Krishna and Poli Venkata Subba Reddy preferred the Deep Neural Network (DNN) model in their 2019 study to predict the response of customers to the campaign. The results obtained with the deep neural network model were compared with classical machine learning methods. The accuracy rate was 92.5% for the deep neural network, while it was 91.0% for k-NN. At this point, classical machine learning algorithms were slightly behind. The study was conducted with a small data set of 1,000 samples [29].

In the study conducted by Vira Bunga Tiara and his friends in 2024, they used Principal Component Analysis (PCA) and examined its effects, unlike others. In addition, ANN, Support Vector Machines, Decision Tree, Random Forest, k-NN algorithms were also applied. Kaggle Bank Marketing Dataset with 11,162 samples was used as the data set. The most successful model, ANN, achieved an accuracy of %80.51 before applying PCA, while this value increased to % 82.08 after applying PCA. The use of PCA increased the accuracy rate of all algorithms by approximately %2. The unbalanced data set can be balanced with the SMOTE method and higher accuracy rates can be achieved [30].

In the study conducted by Yasemin Gültepe and her team on the UCI dataset in 2019, it was seen that the ”duration” feature played a critical role. Naive Bayes and One-R algorithm were used on WEKA. In this study, the One-R algorithm achieved the highest accuracy rate with %89.39. Using only two algorithms, not including modern machine learning methods, and not being able to find a solution to the dataset imbalance can be mentioned as deficiencies. [31].

In the study conducted by Wei Jin and Yingying He in 2019, they used three different data mining models. Decision Tree (CART), Neural Network (NN) and Support Vector Machine (SVM) are the methods used to predict customer responses. A 10% section of the UCI Bank Marketing dataset was used in the study. This section of the dataset contains 4,119 samples and although it has an unbalanced structure, no solution was found for this issue. The most successful among the models was the Decision Tree model. An AUC value of 1.0 was obtained. Although it seems like a perfect result, this situation may indicate over-learning, it is unrealistic. In this direction, the SMOTE method can be recommended for over-learning control and data balancing as a future contribution [32].

Khor Kok-Chin and Ng Keng-Hoong tested the cost-sensitive learning method in their study in 2016. Naive Bayes (NB), C4.5 and Naive Bayes Tree (NBT) models were tested and compared at different error cost rates. UCI Bank Marketing dataset was used. The dataset imbalance problem was solved with SMOTE. As a result of all these, the best value was obtained with the combination of SMOTE and NBT. It was found that cost-sensitive learning was less effective on this dataset [33].

In the study conducted by Suraya Nurain Kalid, Kok-Chin Khor and Keng-Hoong Ng in 2014, they found that the combination of Naive Bayes Tree (NBT) and SMOTE models achieved the best results against imbalance. Data-level solutions were proposed to predict the ”yes” class with fewer samples. The NBT + SMOTE(%1000) combination achieved a ROC value of 0.987. Other metric results such as F1 score, accuracy, precision were not included in the study except for ROC. In addition, the issue that oversampling (SMOTE method) would cause overlearning was not evaluated [34].

In the study conducted by Jianguo Che and his team in 2020, they first proposed the t-SNE (t-distributed Stochastic Neighbor Embedding) and then the support vector machines (SVM) model as a solution to the complex structure in the UCI Bank Marketing data set. The accuracy rate achieved by these two models is %86.07. The points that can be mentioned as minuses; Only the random undersampling method was applied as a solution to the data imbalance, no oversampling method (e.g. SMOTE etc.) was applied. As a future contribution, steps can be taken for over-learning control and can also be supported by powerful models such as Random Forest and XGBoost [35].

In the study conducted by Md. Rashid Farooqi and Naiyar Iqbal in 2019, they applied J48 Decision Tree, SMO (SVM), Artificial Neural Network (ANN), Naive Bayes (NB) and k-Nearest Neighbor (kNN) models to predict the response of customers using the UCI Bank Marketing dataset. In the applications carried out on the WEKA platform, the model that gave the highest performance was J48 with % 91.2 accuracy and % 58 F1 score rate. No solution was sought for data imbalance. In the future, the imbalance can be eliminated with the SMOTE method and more comprehensive results can be obtained by testing the models with different combinations [36] .

In the study conducted by Md. Rashid Farooqi and Naiyar Iqbal in 2019, they applied J48 Decision Tree, SMO (SVM), Artificial Neural Network (ANN), Naive Bayes (NB) and k-Nearest Neighbor (kNN) models to predict the response of customers using the UCI Bank Marketing dataset. In the applications carried out on the WEKA platform, the model that gave the highest performance was J48 with %91.2 accuracy and %58 F1 score rate. No solution was sought for data imbalance. In the future, the imbalance can be eliminated with the SMOTE method and more comprehensive results can be obtained by testing the models with different combinations [37].

In the study conducted by Shamala Palaniappan and her team in 2017, they used the UCI Bank Marketing dataset to divide customers into profiles. Naive Bayes (NB), Decision Tree (DT) and Random Forest (RF) models were applied. The study was conducted on the RapidMiner platform. The most successful of the three models used was the Decision Tree (DT) algorithm with an accuracy rate of 90.68%. This dataset with 41,188 samples is unbalanced and no precautions were taken in this direction. It is possible to increase the accuracy rate with methods such as SMOTE etc. In addition, it was limited to only three basic models. More realistic results can be obtained with more powerful and diverse models [38].

In the research conducted by Chen and his team in 2020, Car Evaluation and Human Activity Recognition data series sets were also examined in addition to the UCI Bank Marketing data set. Random Forest (RF), Support Vector Machines (SVM), k-Nearest Neighbor (k-NN) and Linear Discriminant Analysis (LDA) models were used. varImp(), Boruta, RFE methods were used as feature selection methods. The most successful combination was RF + RF with a value of 0.9099. The data set is unbalanced and no solution was proposed for this issue [39].

In the study conducted by Safarkhani and Moro in 2021, J48 Decision Tree, Logistic Regression and Naive Bayes models were used to predict customer responses to campaigns on the UCI Bank Marketing dataset. The resampling method was used as a precaution against data imbalance and dimensionality reduction methods were applied to reduce complexity. The most successful model was J48 with an accuracy rate of 94.39%. In the future, powerful models such as XGBoost and Random Forest can be integrated, and even more reliable results can be obtained if deep learning methods are used [40].

In the study conducted by Yan and his team in 2025, they proposed a new algorithm called KM-DBSCAN to be tested on the UCI Bank Marketing dataset. This model divided customers into four different groups according to their characteristics such as age, education, and occupation. The highest positive response was received from the first group. Their average age was 42%, their current bank balance was high, and the rate of those who did not currently have a mortgage was 75%. The positive response rate of this group to the campaigns was 45%. It was also mentioned in the introduction that telemarketing is an annoying method for young people. An accuracy rate of 92.3% was achieved with this method [41].

In the study conducted by Elife Ozturk Kiyak and her team in 2023, they sought a solution to the low performance of the k-Nearest Neighbor (k-NN) model against complex data sets. In this direction, the HLKNN (High-Level k-Nearest Neighbors) model was applied. The HLKNN method is not limited to only the nearest neighbors of an example, but also aims to process the neighbors of these neighbors and obtain high results. 32 different data sets were used in this study. These data sets include UCI Bank Marketing's data set with 45,211 samples. While HLKNN achieved an accuracy rate of 88.72%, KNN achieved an accuracy rate of 88.52%. Although there was no serious difference for this data set, when all data sets were compared, HLKNN achieved an accuracy of 81.01% and KNN achieved an accuracy of 79.76%. The HLKNN model achieved better results than KNN in 26 out of 32 data sets. However, it was observed that the HLKNN model required much longer processing times than KNN. No solution was developed for the data set imbalance. These two situations can be mentioned as limitations.[42].

In the study conducted by Ioannis E. Livieris in 2019, a new algorithm called weight-limited feedback neural networks (WCRNN) was tested and compared with different weight limits. The aim of implementing this model is to eliminate the problems of the classical RNN model such as high memory usage and excessive weight growth tendency with weight limits (bounds) and to achieve higher performance. In this study, UCI Bank Marketing, German Credit and Banknote data sets were used. A small part of the UCI Bank Marketing data set with 4,119 samples was used. WCRNN2, with the weight limit determined as [-2, 2], was the model that gave the highest performance with % 42 accuracy and % 42 F1 score. No solution was found for the data set imbalance. In addition, a more comprehensive research can be carried out in the future by using models such as Random Forest (RF), Support Vector Machines (SVM) and XGBoost [43].

In his 2021 study, Ahmad Freij used Support Vector Machines (SVM) and Linear Regression (Linear Regression) models on the UCI Bank Marketing dataset to predict the positive or negative responses of customers to campaigns. The study was conducted on the RapidMiner platform. Three different features stand out that directly affect the results obtained. The first is the "age" feature; It has been observed that as age increases, the probability of receiving a "yes" answer increases slightly. As the number of searches for "campaign" increases, the probability of receiving a "no" answer increases. It has been observed that "previous", that is, someone who has previously participated in campaigns, is more likely to respond "yes" again. Data imbalance can be resolved with SMOTE [44].

The table below provides some information about the studies in the literature.

Article Authors (Year)	Dataset	Sample Numbers	Methods	Results	Deficiencies	Future Contributions
Kevser Özdem & M. Ali Akcayol (2021) [17]	UCI Bank Marketing	45,211 Samples (5289 "Yes", 39,922 "No")	Apriori algorithm	Accuracy:%87 Precision:%96 Sensitivity:%87	The data set is unbalanced and models such as RF, SVM are not used.	Classical models can be integrated and the imbalance can be solved with SMOTE.
Özge Cömert & Mesut Toğaçar (2023) [16]	UCI Bank Marketing	45,211 Samples (5,289 "Yes", 39,922 "No")	KNN, Bayes	Accuracy:%94.68 Specifity:%99.01 Sensitivity:%62.96	Data imbalance has not been resolved and is limited to KNN.	The number of models used should be increased and the imbalance should be resolved.
Fahim Nasir et al. (2024) [18]	UCI Bank Marketing	45,211 Samples (5,289 "Yes", 39,922 "No")	XGBoost, BorderlineSMOTE2	F1: %87 AUC: %94	Balancing was done only based on XGBoost and deep learning was not used.	The scope can be expanded by trying different combinations with different algorithms.
Muneeb Asif (2017) [14]	UCI Bank Marketing	45,211 Samples (5,289 "Yes", 39,922 "No")	DT, SVM, RF, ANN	Accuracy:%90.63 AUC: %93.7	Deep learning was not used. No comparison was made with other advanced algorithms.	Different algorithms should be tested and the number of samples should be increased
Archit Verma (2019) [19]	UCI Bank Marketing	45,211 Samples (5,289 "Yes", 39,922 "No")	DT, NB, SVM	F1: %95.3 Precision: %92.2 AUCPR: %99.5	Limited to WEKA and no deep learning.	Number of models and platforms used can be varied.
Olatunji Apampa (2016) [20]	UCI Bank Marketing	45,211 Samples (5,289 "Yes", 39,922 "No")	LR, DT, NB, RF	Accuracy:%100 Specifity:%90 Sensitivity:%70	Deep learning was not used. No comparison was made with other advanced algorithms.	Different algorithms should be tested and the number of samples should be increased
Anamai Naudom (2022) [21]	UCI Bank Marketing	36,548 samples (4,640 "Yes", 31,908 "No")	J48, NB, RF, RT	Accuracy:87.11% Precision:90.93% F1: 87.4%	Only undersampling is used. Deep learning is not used.	SMOTE and different deep learning models can be included.
Rutu S. Patel and Himanshu S. Mazumdar (2018) [22]	UCI Bank Marketing	41,188 samples	ANN	Accuracy:95.19% Specificity:92.31% Sensitivity:95.42%	Classical machine learning methods were not used and data imbalance was not resolved.	Machine learning algorithms can be integrated and imbalance can be resolved with SMOTE.
Alaa Abu-Srhan (2019) [23]	UCI Bank Marketing	4,521 samples (500 "Yes", 4,021 "No")	RF, SVM, NB, NN, KNN	Doğruluk:%87.27	The dataset is too small and unbalanced. Also, the number of performance metrics is limited.	The dataset can be selected as large and balanced. More metric results can be included.
Shiueng-Bien Yang and Tai-Liang Chen (2020) [24]	UCI Bank Marketing	41.188 örnek	KNN, SVM, RF	Accuracy:100% Specifity:90% Sensitivity:70%	Deep learning was not used. It was compared with different methods.	Different algorithms should be tested and the number of samples should be increased
K. Wisaeng (2013) [25]	UCI Bank Marketing	45,211 Samples (5,289 "Yes", 39,922 "No")	J48, SVM	Accuracy:86.95% Specifity:86.7% Sensitivity:87%	Dataset imbalance is not resolved and compared with different methods.	Dataset balancing methods can be used and deep learning models can be applied.

TABLE I: Literature Review Results

Article Authors (Year)	Dataset	Sample Numbers	Methods	Results	Deficiencies	Future Contributions
Amit Taneja (2024) [26]	UCI Bank Marketing	41.188 sample	LR, DT, RF, SVM	Accuracy:%91 Precision:%91 F1: %91	Hyperparameter optimization is not given in detail. Deep learning algorithms are not used.	Deep learning algorithms can be integrated.
Mücella Özbay Karakuş (2021) [27]	UCI Bank Marketing	500 Samples (ADHD: 300, Healthy: 200)	MLNN	Accuracy:94.3%	Only MLNN tested. Dataset imbalance is not resolved. Number of metrics is limited.	In the future, SMOTE can resolve the imbalance and incorporate classical machine learning algorithms.
Choi ve Jae Choi (2022) [28]	UCI Bank Marketing	45,211 Samples (5,289 "Yes", 39,922 "No")	DT	Accuracy:%78.4 Precision:%82.61	Only decision tree was applied and no solution was found for the imbalanced data set.	RF Models such as SVM can be used and imbalance can be solved with SMOTE.
Chittem Leela Krishna and Poli Venkata Subba Reddy (2019) [29]	UCI Bank Marketing	1.000 sample	DT, SVM, DNN, NB, KNN	Accuracy:92.5%	Dataset is too small and unbalanced.	A larger and more balanced dataset may be preferred.
Vira Bunga Tiara (2024) [30]	Kaggle - Bank Marketing Dataset	11.162 sample	ANN, KNN, SVM, RF, DT	Accuracy:82.08% Precision:88% Sensitivity: 76%	The dataset is unbalanced and hyperparameter optimization is only done for SVM.	Dataset balancing methods can be applied.
Yasemin Gültepe (2019) [31]	UCI Bank Marketing	45,211 Samples (5,289 "Yes", 39,922 "No")	NB, One-R	Accuracy: 89.39%	The number of algorithms used is small and the data set is unbalanced.	Data set balancing methods can be applied and the number of algorithms can be increased
Wei Jin and Yingying He (2019) [32]	UCI Bank Marketing	4.119 sample	DT, NN, SVM	Accuracy:100% Specificity:100% Sensitivity:100%	Values are too high, there may be over-learning.	Over-learning control can be done
Khor Kok-Chin and Ng Keng-Hoong (2016) [33]	UCI Bank Marketing	45,211 Samples (5,289 "Yes", 39,922 "No")	NB, C4.5	Accuracy:100% Specificity:90% Sensitivity:70%	Deep learning was not used. No comparison was made with other high-level algorithms.	The number of algorithms used should be increased
Suraya Nurain Kalid, Kok-Chin Khor ve Keng-Hoong Ng (2014) [34]	UCI Bank Marketing	45,211 Samples (5,289 "Yes", 39,922 "No")	KNN, NBT, C4.5	ROC: %98.7	Oversampling may cause overlearning, no precautions have been taken.	Overlearning must be controlled
Jianguo Che (2020) [35]	UCI Bank Marketing	500 Samples (ADHD: 300, Healthy: 200)	KNN, SVM, RF	Accuracy:100% Specificity:90% Sensitivity:70%	Deep learning was not used. No comparison with other advanced algorithms.	High-level algorithms and deep learning methods should be integrated
Smith (2020)[36]	Kaggle-Healthy and ADHD data	41,188 samples	t-SNE-SVM	Accuracy:86.07% Precision:83.64% Precision:89.38%	Only undersampling was used.	Oversampling methods can also be included

TABLE II: Literature Review Results

III. Method

A. Dataset

1) Data Source: In this study, the UCI Bank Marketing dataset, which belongs to a Portuguese bank, was used. [15].

2) Data Type: Two different data types, Categorical and Integer, were used to display the features in the data set.

3) Data Size: The dataset contains 45,211 samples. Of these samples, 5289 (11.7%) answer "Yes" and 39,922 (88.3%) answer "No".

B. Attribute Extraction

The feature extraction process was carried out with the One-Hot Encoding method, which was performed in the pre-processing phase. One-Hot Encoding is a method of converting categorical data into numerical columns to be used in machine learning models. For example, if a variable has k categories, One-Hot Encoding converts this variable into k 0/1 columns. It is a preferred method especially for "labeled" and "unordered" categorical variables. [45] With this technique, multi-class categorical features such as job, marital and education were transformed into a large number of new binary (i.e. 0/1) features that machine learning algorithms can process. While there were 16 features in the original dataset, this process created a new feature set with 48 features ready to be modeled.

C. Pre-processing

Some preprocessing steps were applied to bring the dataset into a format that machine-deep learning models can process and to optimize the model performance. First, it was determined that there was no missing data in the dataset. This was also stated on the dataset's page on UCI.

In order to process numerical data, binary (0/1) categorical columns such as default, housing, loan and the target variable y were recoded as 1 and 0, respectively. In order to prevent large-scale numbers from suppressing small-scale numbers during the model's learning process, all numerical attributes such as age and balance were standardized with StandardScaler.

The data set was divided into two, 80% training and 20% test data, preserving the class distributions in order to accurately measure the model generalization performance.

D. Attribute Selection

In this study, two different methods, namely Principal Component Analysis (PCA) and Multi-Layer Perceptron (MLP), a deep learning technique, were used to reduce the size of the 48-feature feature set obtained as a result of the feature extraction step and to compare the model performance on different feature sets.

As the first method, Principal Component Analysis (PCA), a dimensionality reduction technique, was applied. As a result of the analysis, it was determined that 23 principal components were sufficient to explain 95% of the total variance in the data set. Accordingly, the original feature space was reduced to a new dimension consisting of 23 synthetic features for both training and test data. With this method, it was aimed to manage possible relationships between features (multicollinearity) and to preserve a large part of the information with fewer features.

As the second method, a deep learning-based feature selection approach was adopted. For this purpose, a Multi-Layer Perceptron (MLP) model was trained and the contribution of each feature to the model's predictive power was measured with the permutation_importance technique. As a result of the analysis, a new feature set was created by selecting the first 25 features with the highest importance. With this method, it was determined that the most important feature was duration.

As a result of these two methods, in addition to the existing 48-feature data set, two new data sets with 23 PCA and 25 MLP features were prepared to be used in the model training and evaluation phase. This provides the opportunity to comparatively analyze the effect of different feature sets on the model performance.

E. Model Selection

1) Machine Learning Algorithms: Machine Learning (ML) is the process by which a computer program improves in performance with experience relative to a given performance metric. Its goal is to make the process of performing cognitive tasks, such as object recognition or natural language translation, fast and efficient [46].

- *Random Forest (RF):* The Random Forest algorithm is an ensemble learning technique that works by creating multiple decision trees (DTs). This method allows the trees to make independent predictions from each other and then concludes these predictions with a majority vote. Thus, more accurate and low error rate results are obtained in large data sets. Random Forest was included in this study because of the very high accuracy rates obtained in other studies in this field in the literature [47].
- *Support Vector Machines (SVM):* Support Vector Machines are a classification method that provides effective results, especially when working with two-class data. When data cannot be separated directly linearly, kernel functions come into play and the data is projected into a higher-dimensional space. Thanks to this transformation, non-linear boundaries can be created and more complex classifications can be made [48].
- *XGBoost:* XGBoost is an advanced machine learning method based on the Gradient Boosting algorithm. It usually uses weak learners such as decision trees, trains these models sequentially and combines them to create a strong prediction model. It shows high performance in terms of accuracy and speed even on large data sets. This model also has some features to prevent overfitting (a distinguishing feature from Gradient Boosting). It was included in the study due to its speed on large data sets and its effect against overfitting [49].
- *Naive Bayes:* Naive Bayes is a classification algorithm based on a probabilistic approach in principle and uses Bayes Theorem in classification processes. In this method, all features are considered independently of each other, so the algorithm works quickly. Due to the large size of the data set to be used in this study, the Naive Bayes model provides an advantage with its speed [50].

2) Deep Learning Algorithms: Deep learning is a machine learning concept based on artificial neural networks. This method works with multi-layered (deep) neural networks and can learn features on its own from raw data [46].

- *Multi-Layers Perceptron (MLP-Classifier):* Multilayer perceptron (MLP) is one of the examples of feedforward neural networks with multiple layers. It can model and classify complex datasets that are impossible to separate linearly. This model is important in terms of including deep learning algorithms in this study [51].

F. Model Training and Evaluation

1) Education Process: First, to create new segments on the dataset, it was tested how many different groups the data could be divided into most efficiently. To determine the number of segments, the K-Means algorithm was run for K values from 2 to 10, and the silhouette coefficient was calculated for each K value. Accordingly, the optimal K value that gave the highest silhouette score was determined as "2". After finding the optimal K value, the customers were divided into two different groups. Since the dataset was unbalanced, the imbalance problem of both groups was solved with the SMOTE method after the groups were separated.

2) Hyperparameter Tuning: In machine learning, instead of the parameters that the model learns from the data, the variables that the researcher sets are called hyperparameters. The performance of each machine learning model depends on its own hyperparameters, such as the number of trees. Determining the correct hyperparameter values is very important to optimize the model performance and directly affects the results to be obtained [52].

In this study, RandomizedSearchCV technique is used to find the hyperparameters that will give the best performance for each model. This method selects random combinations from a predefined parameter set, performs a predefined number of trials and determines the parameter set that obtains the highest Cross Validation score as the "best" as a result of the trials.

The hyperparameter space of the model is expressed by Λ, which is the Cartesian product of the value sets of the n parameters to be tuned:

$$\Lambda = \Lambda_1 \times \Lambda_2 \times \cdots \times \Lambda_n$$

For each hyperparameter setting $\lambda \in \Lambda$, the performance metric of the model obtained by cross-validation (e.g. F1-Score etc.) is represented by the function f(Λ). The RandomizedSearchCV method runs N number of trials, specified by n_iter, and follows these steps: N random configurations are sampled from a probability distribution defined on the Λ space:

$$\lambda^{(1)}, \lambda^{(2)}, \ldots, \lambda^{(N)} \sim p(\lambda)$$

Among these N trials, the best option Λ that maximizes the performance metric f(Λ) is selected:

$$\lambda^* = \underset{\lambda \in \lambda^{(1)}, \ldots, \lambda^{(N)}}{\operatorname{argmax}} f(\lambda)$$

This method was used to find the optimum hyperparameter for each model tested in the study. The best parameters found for each model as a result of optimization were used in the evaluation of the performance of the final model on training and test data, aiming to compare each algorithm with its highest potential.

3) Evaluation Metrics: Evaluation metrics are used to evaluate the performance of machine and deep learning. These metrics provide numerical data about the different features and success of the models [53]

The most important metric in model evaluation is accuracy. The calculation of this value is done with True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN) results. The ratio of the total of false positives and negatives (FP+FN) to the total number of samples (TP+TN+FP+FN) gives the accuracy value result [53].

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

Precision is a performance metric that measures how many of the examples a model predicts in the positive class are actually positive. In this study, Precision determines how many of the people the model predicted as "This customer will accept the campaign" actually accepted the campaign. A high precision result shows that the model's positive predictions are reliable and that marketing efforts are not wasted. On the contrary, if the precision result is low, it shows that a potential customer who will never accept the campaign offer is marked as a "potential customer" by the model. This negatively affects the operational efficiency of the relevant company and the budget allocated to marketing, it is important to obtain a high precision value to optimize this situation. [53].

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

Sensitivity is an evaluation metric that shows what proportion of actually positive examples were correctly identified as "positive" by the model. [53] This metric measures the model's tendency to "miss" the positive class. A high sensitivity result indicates that the model is very successful in finding customers who will accept to participate in the campaign.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

F1-Score is a measure that takes into account both Precision and Sensitivity metrics when evaluating the performance of a model. The harmonic mean of these two metrics is calculated and the result is obtained [53]. In this study, it is not enough to just not miss potential customers (high Sensitivity) or just not waste resources (high Precision). A balance must be established between these two metrics, and the F1-Score shows how well this balance is established. If a high F1-Score is obtained, it shows that the model makes both accurate predictions, i.e. reduces unnecessary searches, and can detect a large portion of potential customers, i.e. does not miss opportunities.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$$

AUC is a performance metric that measures the ability of a model to distinguish between positive and negative classes. The AUC value represents the area under the ROC curve and summarizes how well the model performs at different probability thresholds in a single value. For example, when a customer who is randomly selected from the dataset and who will answer yes to the campaign is placed side by side with a customer who will answer no, it gives the probability that the model will assign a higher "probability of saying yes" to the customer who will say yes.

Principal Component Analysis is a statistical dimensionality reduction technique used to reduce the number of features in large datasets [53]. The purpose of PCA is to transform the main features that may be related to each other into a new set of features called Principal Components that are not related to each other. The new components obtained are ranked according to their ratios explaining the total variance in the dataset; the first principal component explains the most variance, the second principal component explains the largest portion of the remaining variance, and so on.

In this study, PCA was used to represent a large portion (95%) of the information contained in the original dataset of 48 features with a smaller number (23) of features, both to speed up the model training process and to reduce the effects of problems that may arise from the relationships between features (e.g. multicollinearity).

IV. Experimental Results

The analyses in this study were performed on the UCI Bank Marketing dataset, which belongs to a Portuguese bank and is frequently used in the literature. Before proceeding to the model training process, K-Means clustering was performed to determine customer segments, and accordingly, customers were divided into two different segments. Then, the SMOTE technique was applied to each of these segments separately to eliminate the class imbalance problem. Then, Principal Component Analysis (PCA) and Multilayer Perceptron (MLP) feature selection methods were applied to compare the performances of the models on different feature sets.

A. Model Performance and Evaluation

1) Performance Metrics: The performances of machine and deep learning methods were measured with Accuracy, Precision, Sensitivity, Specificity, F1 Score and AUC metrics. PCA ON and PCA OFF states of the models are compared in Table-III and Table-IV. MLP ON and MLP OFF states are compared in Table-5 and Table-6.

TABLE III: Summary Performance Metrics of Models Based on PCA Effect

Model Name	PCA	Accuracy	F1-Score	AUC
Random Forest	PCA OFF	0.8936	0.6121	0.9288
Random Forest	PCA ON	0.8754	0.5623	0.9050
XGBoost	PCA OFF	0.8911	0.6020	0.9233
XGBoost	PCA ON	0.8725	0.5533	0.9006
MLP Classifier	PCA OFF	0.8717	0.5195	0.8814
MLP Classifier	PCA ON	0.8482	0.5379	0.8848
Naive Bayes	PCA OFF	0.8691	0.4807	0.7713
Naive Bayes	PCA ON	0.8478	0.4452	0.7706
SVM	PCA OFF	0.7646	0.2553	0.6139
SVM	PCA ON	0.8074	0.3243	0.6775

TABLE IV: Performance Results of Models Based on PCA Effect

Model Name	PCA	Precision	Sensitivity	Specificity
Random Forest	PCA OFF	0.5338	0.7174	0.9170
Random Forest	PCA ON	0.4773	0.6843	0.9007
XGBoost	PCA OFF	0.5258	0.7042	0.9158
XGBoost	PCA ON	0.4688	0.6749	0.8987
MLP Classifier	PCA OFF	0.4624	0.5926	0.9087
MLP Classifier	PCA ON	0.4177	0.7552	0.8605
Naive Bayes	PCA OFF	0.4484	0.5180	0.9156
Naive Bayes	PCA ON	0.3882	0.5217	0.8910
SVM	PCA OFF	0.2027	0.3450	0.8202
SVM	PCA ON	0.2750	0.3951	0.8620

In Random Forest and XGBoost models, PCA application negatively affected the performance. While F1-Score decreased from 0.6121 to 0.5623 in Random Forest model, it decreased from 0.6020 to 0.5533 in XGBoost model with a similar decrease. This situation shows that these models use the discriminative information in the main feature set more efficiently than the components formed by PCA.

MLP Classifier and Support Vector Machines (SVM) models were positively affected by PCA application. The highest increase was observed in SVM model, where F1-Score increased from 0.2553 to 0.3243. In MLP Classifier model, PCA slightly decreased the Precision result but increased the Sensitivity result from 0.5926 to 0.7552, thus increasing the F1-Score. The Naive Bayes model was not affected much by the PCA condition and only experienced a small decrease in the F1 Score. These results show that although PCA may improve some models, it is not efficient in all models.

TABLE V: Summary Performance Metrics of Models Based on the Effect of Feature Selection with MLP

Model Name	MLP Status	Accuracy	F1-Score	AUC
Random Forest	MLP OFF	0.8936	0.6121	0.9288
Random Forest	MLP ON	0.8870	0.6126	0.9259
XGBoost	MLP OFF	0.8911	0.6020	0.9233
XGBoost	MLP ON	0.8891	0.6102	0.9231
MLP Classifier	MLP OFF	0.8717	0.5195	0.8814
MLP Classifier	MLP ON	0.8645	0.5560	0.8974
Naive Bayes	MLP OFF	0.8691	0.4807	0.7713
Naive Bayes	MLP ON	0.8872	0.4780	0.8032
SVM	MLP OFF	0.7646	0.2553	0.6139
SVM	MLP ON	0.7363	0.3265	0.7157

TABLE VI: Diagnostic Performance Metrics of Models Based on the Effect of Feature Selection with MLP

Model Name	MLP Status	Precision	Sensitivity	Specificity
Random Forest	MLP OFF	0.5338	0.7174	0.9170
Random Forest	MLP ON	0.5114	0.7637	0.9033
XGBoost	MLP OFF	0.5258	0.7042	0.9158
XGBoost	MLP ON	0.5182	0.7420	0.9086
MLP Classifier	MLP OFF	0.4624	0.5926	0.9087
MLP Classifier	MLP ON	0.4509	0.7250	0.8830
Naive Bayes	MLP OFF	0.4484	0.5180	0.9156
Naive Bayes	MLP ON	0.5212	0.4414	0.9463
SVM	MLP OFF	0.2027	0.3450	0.8202
SVM	MLP ON	0.2328	0.5463	0.7614

The results presented in Table V and Table VI show that the second feature selection method of the study, MLP, has a positive effect on most of the models, unlike PCA.

Using 25 features selected by MLP method in Random Forest, SVM, XGBoost and MLP Classifier models increased F1-Score. Only in Naive Bayes model F1-Score decreased slightly, but Precision and AUC metrics increased. This can be interpreted as Naive Bayes makes more accurate but hesitant predictions.

2) Comparison of PCA and MLP Methods: When the effects of PCA and MLP models on performance results are compared, it is seen that MLP has a significant superiority. While Random Forest and XGBoost models obtained an F1-Score of approximately 0.61, this score decreased to 0.55 when PCA was applied. However, when MLP was applied, the value of 0.61 was preserved, and this result was achieved despite the number of features being reduced by 48% (25 out of 48 features were discarded).

As a result, my MLP method has a higher impact on the model metric results compared to PCA because it preserves the most informative original features and eliminates the less important ones that may create noise.

B. Error Analysis

Error analysis examines the type and amount of misclassifications (FP and FN) made by the model, revealing its strengths and weaknesses. This allows us to evaluate the risks (e.g. lost customers and wasted resources) and advantages of the model in real-life applications.

The following table shows the error analysis results obtained when the best performing method is MLP ON.

When the error types in Table VII are examined, two important errors stand out:

False Positive (FP): Marking a customer who will not accept the campaign as 'will accept'. This mistake means unnecessary searches and therefore a waste of resources.

TABLE VII: Error Analysis Results of Models in MLP ON Condition

Model Name	TP	TN	FP	FN
Random Forest	808	7213	772	250
XGBoost	785	7255	730	273
MLP Classifier	767	7051	934	291
Naive Bayes	467	7556	429	591
SVM	578	6080	1905	480

False Negative (FN): Marking a customer who will accept the campaign as 'will not accept'. This means a lost customer and is probably a more costly mistake.

Accordingly, it is seen that the Random Forest (FN: 250) and XGBoost (FN: 273) models, which give the highest F1-Score, have some of the lowest False Negative numbers. This shows that these models have a lower risk of missing marketing opportunities than others. On the other hand, SVM (FP: 1905), which has the highest FP value, stands out as the model that uses resources most inefficiently.

C. Evaluation of Training and Test Performance

It is not enough in practice for a machine learning model to achieve high results on training data. The main thing is that it exhibits consistent performance on test data that it has no relation to. Large differences between training and test performances indicate a serious problem called overfitting, which indicates that the model is memorizing the data instead of learning it. In this section, the overfitting tendencies of the developed models are analyzed by comparing the 5-fold Cross Validation (CV) results reflecting the generalization performance on the training data with the final test set results.

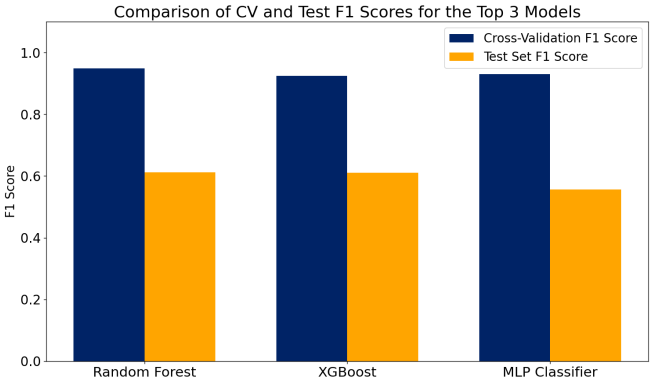

Fig. 2: Comparison of CV and Test F1 Scores for the Top 3 Models.

The graph clearly shows the difference between the 5-fold Cross Validation F1-Scores (dark blue bars) reflecting the performance of the models on the training data and their performance on the test set (orange bars).

This is the clearest visual evidence that even the most successful models overfit the synthetic training data generated by the applied K-Means and aggressive SMOTE balancing method. Even in the best models such as Random Forest and XGBoost, Cross Validation scores in the range of 0.93-0.95 drop to 0.61 on the test set, indicating a performance loss of over 30 points. This situation forms the basis for analyzing the effects of the data balancing method used on generalization in the "Discussion" section of the study and points to the potential risks of such synthetic data augmentation techniques.

D. Overfitting Case

Large differences between training and testing performances indicate a serious problem called over-learning, which indicates that the model memorizes the data instead of learning it [35]. In this section, the over-learning tendencies of the developed models are analyzed by comparing the 5-fold Cross-Validation (CV) results reflecting the generalization performance on the training data with the final test set results.

The data presented in Figure 2, which summarizes the results of the three most successful models, clearly demonstrates this over-learning situation. In all tested models, there is a significant and systematic drop between the Cross-Validation performance on the training data and the final performance on the never-before-seen test set. For example, even in the best models, Random Forest and XGBoost, the Cross-Validation F1-Scores, which were in the range of 0.93-0.95, dropped to 0.61 on the test set, indicating a performance loss of over 30 points.

It was determined that the main reason for this over-learning was due to the K-Means and SMOTE balancing method adopted as the original methodology of the study. Balancing the training set with synthetic data at a perfect ratio (50/50) caused the models to become experts on the distribution of this data. However, when faced with the test set, which was representative of the real world and was extremely unbalanced, these "memorized" rules failed to provide sufficient performance in generalization. The strongest evidence for this situation is that even the simplest model, Naive Bayes, shows a serious tendency for over-learning under this methodology.

After this problem was identified, various solutions were sought. First, the aim was to reduce the model complexity. "Aggressive pruning" was applied to the Random Forest model by setting hyperparameters such as max_depth (maximum tree depth) to more restrictive values. However, it was observed that this intervention did not provide a significant improvement in the test set performance and was insufficient to close the gap between CV and test scores. This result confirms that the problem is not only due to the complexity of the model, but also due to the mismatch between the training and test data distributions.

E. Time and Resource Usage

The time taken for training and testing processes directly affects the practicality, scalability and computational cost of the model in real-world applications. In this section, the training and testing times of all tested models under different feature selection methods are comparatively examined.

1) Training and Testing Times: Each model was tested in scenarios where two different feature selection models were applied and not applied, and the training and testing times measured in seconds were summarized in Table VIII. All tests were performed on a system with an Intel Core i7 processor, 16 GB RAM and an NVIDIA GeForce 3060 graphics card, using open source libraries such as the Python programming language Scikit-learn and XGBoost.

2) Effect of Feature Selection on Duration: When the results in Table VIII are examined, it is seen that feature selection methods have an effect on the training time depending on the models. It has been determined that PCA application slightly reduces the training time as expected in models such as NB and SVM, but significantly increases the time in tree-based RF and XGBoost models. This situation is thought to be due to the fact that the components formed by PCA make it difficult for decision trees to find the

TABLE VIII: Training and Testing Times of Models in Different Scenarios (seconds)

Model Name	Scenario	Training Time (s)	Testing Time (s)
Random Forest	PCA OFF	61.21	0.08
Random Forest	PCA ON	230.24	0.06
Random Forest	MLP ON	69.10	0.07
XGBoost	PCA OFF	109.44	0.05
XGBoost	PCA ON	392.02	0.08
XGBoost	MLP ON	2752.84	0.03
MLP Classifier	PCA OFF	491.46	0.01
MLP Classifier	PCA ON	469.35	0.03
MLP Classifier	MLP ON	514.43	0.01
Naive Bayes	PCA OFF	8.42	0.01
Naive Bayes	PCA ON	1.58	0.01
Naive Bayes	MLP ON	1.85	0.00
SVM	PCA OFF	731.65	0.17
SVM	PCA ON	690.95	0.97
SVM	MLP ON	506.18	0.19

most appropriate separation point due to the decision tree-based structure of both models and increase the computational cost.

On the other hand, MLP-based feature selection, despite reducing the number of features, has excessively extended the training time of the XGBoost model (approximately 46 minutes). This shows that the features selected by MLP create a very costly pair for the sequential learning mechanism of XGBoost. When we look at the other models, it is seen that it does not have a negative or positive effect in terms of time.

The fact that the test times of all models are quite low (all under one second) shows that the models can produce predictions very quickly after being trained. In terms of efficiency, RF, which is one of the models that gives the best performance, stands out in this sense as it can be trained much faster than XGBoost.

F. Generalization Ability of the Model and Cross-Validation Results

When evaluating the model success, it is measured not only by the results obtained on a specific test set, but also by how well it can adapt to different subsets of data that it has not seen before, i.e. its generalization ability. If the generalization capacity of the model is high, it is expected to show more reliable and consistent performance when faced with real-world data.

1) Cross-Validation Results: In this study, 5-fold Cross-Validation technique was used to measure the generalization performance of the models more reliably. In this method, the segmented and balanced training data is divided into 5 equal parts; the model is trained on 4 parts and tested on 1 part, and this process is repeated 5 times until each part is used as test data.

The average F1-Score and standard deviation values obtained as a result of 5-fold Cross-Validation in the preview selection method (mostly MLP ON) where each model showed the best performance are summarized in Table IX. A low standard deviation indicates that the model performs more consistently on different data subsets.

TABLE IX: 5-Fold Cross Validation F1-Score Results in Best-Scenario Models

Model Name	Mean F1-Score (CV)	Standard Deviation (CV)
Random Forest	0.9488	$\pm$0.0008
XGBoost	0.9252	$\pm$0.0023
MLP Classifier	0.9306	$\pm$0.0031
Naive Bayes	0.7242	$\pm$0.0073
SVM	0.6659	$\pm$0.0183

These results show that especially Random Forest, XGBoost and MLP Classifier models obtain low standard deviation values. This shows that the models have a high and stable generalization ability on the training data. This table has the data that forms the basis of over-learning, compared to the performance of the models on the test set, which was previously presented in Figure 2.

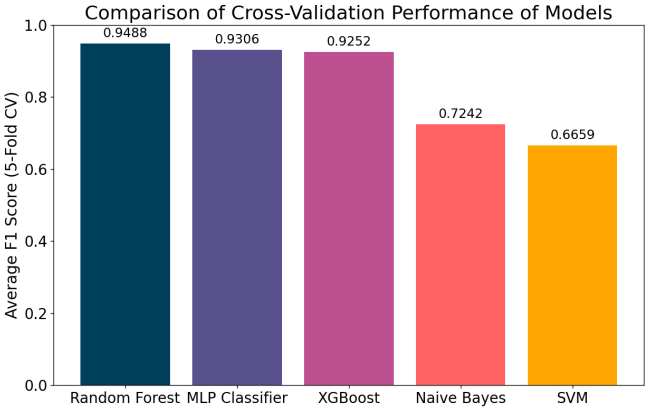

Fig. 3: Comparison of Cross-Validation Performance of Models.

The 5-fold Cross Validation results presented in Table IX are summarized visually in Graph 3. As can be seen from the graph, Random Forest, XGBoost and MLP Classifier models show very high and close generalization performance on the training data with average F1-Scores of 0.9488, 0.9252 and 0.9306 respectively. Naive Bayes and SVM are far behind these three models.

V. Discussion

The obtained results show that Tree Based Methods (Random Forest and XGBoost) are clearly superior to the other models for this classification problem. Both models have a significant difference over their closest competitors with a test set performance of 0.61 in the F1-Score metric. MLP Classifier, the Deep Learning representative of the study, is behind these two models but ranks third with a better performance than Naive Bayes and SVM. The lowest performance of SVM in this problem indicates that the decision boundaries in the dataset are too complex.

A. Effect of Feature Selection Methods

The two feature reduction methods compared in the study (PCA and MLP-based selection) have different effects on the models. The more traditionally accepted PCA method has reduced the performance of the most successful models, Random Forest and XGBoost. On the other hand, one of the original contributions of the study, deep learning-based feature selection MLP, has managed to maintain the performance of these models despite reducing the number of features by approximately %48. The main reason for this situation is that the MLP method preserves the most critical main features in terms of information and eliminates only those that may create noise, while PCA combines all the features and produces components that make it difficult to interpret the models.

B. Original Balancing Method and Systematic Over-Learning Problem:

One of the most important experiments of the study was an original data balancing approach based on segmentation. A strong aspect of this approach is that instead of randomly determining

the number of customer segments (K), the most suitable K value for the data structure (K=2 for all scenarios) is mathematically determined using the Silhouette coefficient. Despite this meticulous segmentation step, the subsequent SMOTE application, which perfectly balanced each cluster at a 50/50 ratio, triggered the over-learning problem, which is the biggest problem of the study.

A large difference was observed between the Cross-Validation scores and the Test Set scores of all tested models. It proves that the models memorized the artificial training data that deviates from the real-world data distribution. This situation shows how much the proposed original methodology (K determination with Silhouette) will affect the result in the following steps, no matter how robust it is. Therefore, it was concluded that the proposed original method, although conceptually innovative, causes the over-learning problem in its current form.

C. Efficiency and Time Comparison

Although RF and XGBoost are neck and neck in terms of performance, there is a significant difference between them in terms of efficiency. In their best case scenario, RF was trained in about 1 minute, while XGBoost took about 46 minutes to train. This shows that even though both models achieved good results, RF is a more efficient and useful method when the training time is considered. While the training times of SVM and MLP models are quite long, the fastest model was Naive Bayes, as expected.

D. Limitations of the Study and Future Recommendations

The biggest limitation of this study is the over-learning problem caused by the original balancing method. To overcome this problem, various methods such as reducing model complexity (tightening regularization parameters), balancing at the algorithm level (using class_weight), and reducing SMOTE intensity have been tried. Despite this, it has been observed that the difference between the performance on the training data and the test data continues. This situation has been determined that the root of the problem is caused by an element beyond simple hyperparameter settings.

Accordingly, more advanced methods can be suggested for future studies. In particular, more advanced techniques such as Borderline-SMOTE, which have been shown to yield successful results in the literature, can be used, or cost-sensitive learning algorithms that assign different costs to False Positive (FP) and False Negative (FN) errors can be used to solve this problem efficiently.

VI. Result

In this study, the problem of predicting customer responses with artificial intelligence models was addressed in order to increase the success rate of banks' telemarketing campaigns. Accordingly, the customers in the dataset were first optimized with the Silhouette coefficient and divided into two different segments using the K-Means algorithm, and both segments were balanced with the SMOTE technique. After this original data preparation process, different versions of the dataset were created by selecting features with PCA and MLP, and five different models, such as Random Forest, XGBoost, MLP, Naive Bayes and SVM, were tested on these sets.

As a result of the experiments, it was seen that the Random Forest model achieved the highest prediction success with an F1-Score of 0.6126 in the scenario where MLP was used. When the feature selection methods (PCA and MLP) were compared, it was determined that MLP gave better results than the PCA method in most of the models. The most important finding of the study is that the combination of K-Means + Silhouette + SMOTE proved to cause over-learning problem in all tested models. Although different methods were used to solve the problem, none of the experiments yielded results. This situation constitutes the biggest limitation of the study.

References

[1] Ruhan İri. Telefonla veya internet üzerinden sipariş verilen ürünlerin tüketicilere teslim edilmesinde oluşan müşteri memnuniyetinin önemine yönelik niğde ve yöresinde yapilan bir araştirma. *İktisadi ve İdari Yaklaşımlar Dergisi*, 1(1):1–14, 2019.

[2] Levent Gelibolu and Tufan Özsoy. Çağrı merkezlerinin satış amaçlı kullanılması: Doğrudan pazarlamanın bir unsuru olarak telepazarlama. *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 22(1):481–500, 2013.

[3] Muhsin Özgür Dolgun and Derya Ersel. Doğrudan pazarlama stratejilerinin belirlenmesinde veri madenciliği yöntemlerinin kullanımı. *İstatistikçiler Dergisi: İstatistik ve Aktüerya*, 7(1):1–13, 2014.

[4] Nurhan Babür Tosun. Doğrudan pazarlama reklamlarının etkisi. *Galatasaray Üniversitesi İletişim Dergisi*, (11):9–26, 2009.

[5] Muhammed Bilgehan Aytaç and Hasan Şakir Bilge. Tele pazarlama verilerinin birliktelik kurallariyla ve crisp-dm yöntemiyle analiz edilmesi. *Aksaray Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 5(2):25–40, 2013.

[6] Ozonetel. State of the contact center 2023-24, 2023. Erişim Tarihi: 3 Nisan 2025.

[7] Suresh P Machhar, Mr Prashant M Pilot, and Anand Vallabh Vidyanagar. "a study on consumer perception towards telemarketing: Special reference to female consumers in anand city". *Journal of Emerging Technologies and Innovative Research*, 5(11):490–497, 2018.

[8] Quirk's Editorial Staff. Rising refusal rates: The impact of telemarketing. *Quirk's Marketing Research Review*, 2023. Accessed: 2025-04-04.

[9] Elgiz Yılmaz Altuntaş and Ece Boydak. Kuşak teorisi ve marka gençleştirme stratejisi kapsaminda markalarin dijital reklam uygulamalari üzerinden değerlendirilmesi: Orkid, coca-cola, vichy ve exxen örnekleri. *Uluslararası Halkla İlişkiler ve Reklam Çalışmaları Dergisi*, 5(1):88–114, 2022.

[10] Ozan Karaarslan and Ramazan Nacar. Dijital pazarlama kanallarının kullanımının x, y, z kuşaklarına göre farklılaşması. *Turkish Journal of Marketing*, 9(4):97–113, 2024.

[11] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[12] OECD. Understanding financial consumer behaviour. https://www.oecd.org/finance/understanding-financial-consumer-behaviour.htm, 2022. Accessed: 2025-04-04.

[13] Laksana Vongchalerm. *Analysis of predicting the success of the banking telemarketing campaigns by using machine learning techniques*. PhD thesis, Dublin, National College of Ireland, 2022.

[14] Muneeb Asif. Predicting the success of bank telemarketing using various classification algorithms, 2018.

[15] S. Moro, P. Cortez, and P. Rita. Bank marketing dataset. https://archive.ics.uci.edu/ml/datasets/Bank+Marketing, 2014. UCI Machine Learning Repository.

[16] Özge Cömert and Mesut Toğaçar. Telepazarlama çağrılarının başarısını tahmin etmek üzere veriye dayalı bir yaklaşım. *Verimlilik Dergisi*, 57(4):735–746, 2023.

[17] Kevser Özdem and M Ali Akcayol. Müşteri davranış tahmini için bir model: Bankacılık sektörü için uygulama. *Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, 10(1):1–8, 2021.

[18] Fahim Nasir, Abdulghani Ali Ahmed, Mehmet Sabir Kiraz, Iryna Yevseyeva, and Mubarak Saif. Data-driven decision-making for bank target marketing using supervised learning classifiers on imbalanced big data. *Computers, Materials & Continua*, 81(1), 2024.

[19] Archit Verma. Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using weka. *International Research Journal of Engineering and Technology*, 5(13):54–60, 2019.

[20] Olatunji Apampa. Evaluation of classification and ensemble algorithms for bank customer marketing response prediction. *Journal of International Technology and Information Management*, 25(4):6, 2016.

[21] Waritpon Saengthongrattanachot, Anamai Na-udom, and Jaratsri Rungrattanaubol. A comparison of machine learning techniques for classification in bank marketing data. *Thai Journal of Mathematics*, pages 157–168, 2022.

[22] Rutu S Patel and Himanshu S Mazumdar. Prediction of bank investors using neural network in direct marketing. *International Journal of Engineering and Applied Sciences*, 5(2):257283, 2018.

[23] Alaa Abu-Srhan, Bara'a Alhammad, Rizik Al-Sayyed, et al. Visualization and analysis in bank direct marketing prediction. *International Journal of Advanced Computer Science and Applications*, 10(7), 2019.

[24] Shiueng-Bien Yang and Tai-Liang Chen. Uncertain decision tree for bank marketing classification. *Journal of Computational and Applied Mathematics*, 371:112710, 2020.

[25] K Wisaeng. A comparison of different classification techniques for bank direct marketing. *International Journal of Soft Computing and Engineering (IJSCE)*, 3(4):116–119, 2013.

[26] Amit Taneja. Customer behavior analysis in banking: Leveraging big data to enhance personalized services.

[27] Mücella Özbay Karakuş. A multi-layer neural network approach to predict the success of bank telemarketing. *Artificial Intelligence Theory And Applications*, 1(1):69–75, 2021.

[28] Youngkeun Choi and Jae Choi. How does machine learning predict the success of bank telemarketing? 2022.

[29] Chittem Leela Krishna and Poli Venkata Subba Reddy. An efficient deep neural network classifier in banking system. *International Journal of Research and Analytical Reviews*, 6(2):511–517, 2019.

[30] Vira Bunga Tiara, Amril Mutoi Siregar, Dwi Sulistya Kusumaningrum Kusumaningrum, and Tatang Rohana. Bank customer segmentation model using machine learning. *Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI*, 13(1):66–78, 2024.

[31] Yasemin Gultepe, Wisam Gwad, Yuosra Aljamel, and Yossf Ahmed. Mining a marketing campaigns data of bank. 2019.

[32] Wei Jin and Yingying He. Three data mining models to predict bank telemarketing. In *IOP Conference Series: Materials Science and Engineering*, volume 490, page 062075. IOP Publishing, 2019.

[33] Khor Kok-Chin and Ng Keng-Hoong. Evaluation of cost sensitive learning for imbalanced bank direct marketing data. *Indian Journal of Science and Technology*, 9:42, 2016.

[34] Suraya Nurain Kalid, Kok Chin Khor, Keng Hoong Ng, and Choo Yee Ting. Effective classification for unbalanced bank direct marketing data with over-sampling. In *Proceedings of the knowledge management international conference (KMICe), Langkawi, Kedah*, pages 12–15, 2014.

[35] Jianguo Che, Sai Zhao, Yongfan Li, and Kai Li. Bank telemarketing forecasting model based on t-sne-svm. *Journal of Service Science and Management*, 13(3):435–448, 2020.

[36] Rashid Farooqi and Naiyar Iqbal. Performance evaluation for competency of bank telemarketing prediction using data mining techniques. *International Journal of Recent Technology and Engineering*, 8(2):5666–5674, 2019.

[37] Israa M Hayder, Ghazwan Abdul Nabi Al Ali, and Hussain A Younis. Predicting reaction based on customer's transaction using machine learning approaches. *International Journal of Electrical and Computer Engineering*, 13(1):1086, 2023.

[38] Shamala Palaniappan, Aida Mustapha, Cik Feresa Mohd Foozy, and Rodziah Atan. Customer profiling using classification approach for bank telemarketing. *JOIV: International Journal on Informatics Visualization*, 1(4-2):214–217, 2017.

[39] Rung-Ching Chen, Christine Dewi, Su-Wen Huang, and Rezzy Eko Caraka. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7(1):52, 2020.

[40] Fereshteh Safarkhani and Sérgio Moro. Improving the accuracy of predicting bank depositor's behavior using a decision tree. *Applied Sciences*, 11(19):9016, 2021.

[41] Xiaohua Yan, Yufeng Li, Fuquan Nie, and Rui Li. Bank customer segmentation and marketing strategies based on improved dbscan algorithm. *Applied Sciences (2076-3417)*, 15(6), 2025.

[42] Elife Ozturk Kiyak, Bita Ghasemkhani, and Derya Birant. High-level k-nearest neighbors (hlknn): a supervised machine learning model for classification analysis. *Electronics*, 12(18):3828, 2023.

[43] Ioannis E Livieris. Forecasting economy-related data utilizing weight-constrained recurrent neural networks. *Algorithms*, 12(4):85, 2019.

[44] A. Freij. Classification models for bank marketing campaign: Towards smart bank marketing. *American Journal of Business and Operations Research*, pages 21–30, 2021.

[45] Alice Zheng and Amanda Casari. *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.", 2018.

[46] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic markets*, 31(3):685–695, 2021.

[47] Shiju Sathyadevan and Remya R. Nair. Comparative analysis of decision tree algorithms: Id3, c4.5 and random forest. In Lakhmi C. Jain, Himansu Sekhar Behera, Jyotsna Kumar Mandal, and Durga Prasad Mohapatra, editors, *Computational Intelligence in Data Mining - Volume 1*, pages 549–562, New Delhi, 2015. Springer India.

[48] Sonal Pathak, Suhail Javed Quraishi, Anupam Singh, Malikhan Singh, Kavita Arora, and Danish Ather. A comparative analysis of machine learning models: Svm, naïve bayes,

random forest, and lstm in predictive analytics. In *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pages 790–795, 2023.

[49] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM, August 2016.

[50] Ayşe Berna Altınel. Türkçe metinlerde makine öğrenmesi algoritmalarının duygu analizi problemi üzerindeki performansının kıyaslanması. *Avrupa Bilim ve Teknoloji Dergisi*, (28):1056–1061, 2021.

[51] Fionn Murtagh. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2(5):183–197, 1991.

[52] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The journal of machine learning research*, 13(1):281–305, 2012.

[53] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.