Enhancing Bank Marketing Strategies: The Impact of Feature Reduction Techniques on Machine Learning Model Performance

Zeynep Özer[1]
[1]*Bandırma Onyedi Eylül University, Dept. of management Information Systems, Bandırma, Balıkesir, Turkey, zozer@bandirma.edu.tr, ORCID: 0000-0001-8654-0902*

This research investigates the application of machine learning models in the banking sector, specifically focusing on the classification of bank marketing datasets. We explore the use of feature reduction techniques, including Principal Component Analysis (PCA) and SelectKBest, to enhance the performance of various models such as Multi-Layer Neural Networks (MLNN), K-Nearest Neighbors (KNN), Random Forest(RF), and Linear Discriminant Analysis (LDA). Our study reveals the pivotal role of these reduction techniques in addressing the challenges posed by high-dimensional data and imbalanced distributions in customer behavior prediction. Through comparative analysis, we demonstrate how PCA and SelectKBest, through F-Value and Chi-Squared methods, influence model accuracy and efficiency, providing insights into the effectiveness of these models in strategic bank marketing efforts.

Keywords: *Bank Marketing, Machine Learning, Feature Selection*

1. Introduction

Marketing strategies are evolving, especially in the banking sector where direct marketing plays a crucial role. Centralized contact centers have streamlined the operational management of these campaigns, with telemarketing emerging as a key channel for customer interaction [1,2]. This remote approach, which includes both inbound and outbound contacts, leverages technology to focus on customer value maximization. Decision support systems (DSSs) and Business Intelligence (BI), encompassing data mining (DM) techniques, are instrumental in this evolution. They aid in identifying potential clients and extracting valuable insights from data, with classification being a primary DM task for categorizing client interactions as successful or unsuccessful in terms of product subscription [3,4].

The critical role of data-driven decision-making in responding to challenging business environments is increasingly recognized, especially in the banking industry [1]. Identifying profitable customers is key to sustaining competitive advantage and maintaining long-term relationships. However, there are untapped opportunities in bank marketing, and one major challenge is improving database marketing efficiency. A significant issue in business analytics within this sector is the imbalanced data distribution of potential target customers, which complicates knowledge extraction from bank marketing data. Effective handling of this imbalance is necessary, as common approaches often lead to either processing overhead or information loss.

Machine learning models (MLMs) are widely used in marketing for customer behaviour prediction, which, as in many fields [5-10], can identify complex, non-linear relationships in data such as bank telemarketing datasets [11,12]. MLMs are adept at generalizing models for unseen inputs and can handle diverse data distributions. Adapting MLMs to be cost-sensitive can significantly improve their performance, especially in dealing with imbalanced data distributions, which is common in bank telemarketing. Unlike other methods that may compromise data quality through re-sampling, cost-sensitive MLMs maintain the integrity of the original dataset, making them suitable for predicting customer responses in bank telemarketing [13].

Feature reduction in machine learning is a crucial process that involves reducing the number of input variables used to build predictive models. It aims to simplify models, improve their performance, and minimize overfitting. Techniques like PCA and SelectKBest are used to identify the most relevant features. PCA reduces dimensionality by transforming features into a new set of variables, the principal components, which capture the most variance in the data. SelectKBest, on the other hand, selects features based on statistical tests, like the ANOVA(Analysis of Variance) F-Value and chi-

square test, to keep those with the highest significance. Effective feature reduction enhances model accuracy and efficiency, particularly in high-dimensional data scenarios.
In this study, we explore the application of machine learning models to the classification of the Bank Marketing dataset, a task central to enhancing strategic marketing efforts in the banking sector. Our focus lies in assessing the efficacy of feature reduction algorithms, specifically PCA and SelectKBest, the latter applied through F-Value and Chi-Squared methods, in refining model performance. We delve into the intricacies and outcomes of deploying Support Vector Machine (SVM), MLNN, KNN, RF, and LDA in this context. The evaluation aims to shed light on how these reduction techniques impact the accuracy and efficiency of predictive models, providing valuable insights into the dynamics of feature selection in bank marketing analytics.

2. Material and Methods

2.1 Principal Component Analysis (PCA)

PCA is a statistical technique used for dimensionality reduction in data analysis. It transforms a set of possibly correlated variables into a smaller number of uncorrelated variables called principal components. These components are identified in such a way that the first principal component accounts for the largest possible variance in the data, and each succeeding component, in turn, has the highest variance possible under the constraint that it is orthogonal to the preceding components. PCA is often used to simplify data, reduce noise, and identify underlying patterns. The method is particularly valuable when dealing with high-dimensional data, allowing for the visualization and interpretation of complex datasets by projecting them onto a lower-dimensional space while retaining most of the data's original variance.

2.2 SelectKBest Algorithm

The SelectKBest algorithm in conjunction with ANOVA F-Value and Chi-Squared tests is a feature selection technique used in machine learning. ANOVA F-Value is employed for linear models, particularly useful for regression tasks, where it evaluates the individual contribution of each feature by comparing the variance between different groups and within groups. The Chi-Squared test, on the other hand, is used for categorical data in classification tasks to assess the independence of two variables, thus determining the usefulness of a categorical feature in predicting a category outcome. SelectKBest then ranks these features based on their test scores and selects the 'K' highest scoring features, where 'K' is a user-defined parameter. This method is instrumental in enhancing model performance by reducing dimensionality, improving accuracy, and expediting training processes. The ANOVA F-Value is calculated using the equations:

$$F = \frac{Between\ Group\ Variance}{Within\ Group\ Variance}$$

The Chi-Squared test for categorical data is given by:

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency and E_i is the expected frequency under the null hypothesis of independence.

2.3 Bank Marketing Dataset

The bank marketing dataset comprises a total of 45,211 entries, spanning across 17 different columns. Among these, seven are numerical in nature, including attributes such as 'age', 'balance', 'day_of_week', 'duration', 'campaign', 'pdays', and 'previous'. The dataset also contains 10 categorical columns, with the 'y' column specifically indicating the class information, which is crucial for predictive modelling.

A closer look at the statistical summary of the numerical columns reveals a diverse range of data. For instance, the 'age' of individuals varies significantly, starting from 18 and going up to 95 years, with an average age of around 40.94 years. The 'balance' column, indicating the account balance, shows a substantial spread from -8,019 to 102,127, with an average balance of 1,362.27. This suggests a wide disparity in the financial status of the individuals in the dataset. The 'day_of_week' column, likely representing the day of the month when contact was made, spans from 1 to 31. The

'duration' of calls ranges vastly, from as short as 0 seconds to as long as 4,918 seconds, with an average duration of about 258.16 seconds. This indicates varied levels of engagement from the clients. The 'campaign' column, showing the number of contacts made during this campaign for a client, ranges from 1 to 63, hinting at diverse marketing efforts. The 'pdays' column, representing the number of days since the client was last contacted, varies from -1 (indicating no previous contact) to 871. Lastly, the 'previous' column, denoting the number of contacts made before this campaign, ranges from 0 to 275.

An important aspect of this dataset is the class distribution in the 'y' column, which is a binary classification of 'yes' or 'no', reflecting the outcome of the marketing campaign. The data shows a pronounced class imbalance, with a significantly higher number of 'no' responses (39,922 entries) compared to 'yes' responses (5,289 entries). This skewness in class distribution is a crucial factor to consider in any predictive modelling or analysis derived from this dataset.

Overall, the dataset offers a rich variety of information that can be leveraged for insightful analysis, particularly in understanding the factors that influence the success of a bank marketing campaign. However, the class imbalance and the wide range in some numerical values suggest that careful preprocessing and analysis techniques would be required to glean meaningful conclusions.

2.4 Evolution Metrics

In the context of classification models, the performance is typically assessed using a suite of metrics, each offering unique insights. Accuracy is a broad measure, calculating the proportion of true results (both positives and negatives) in the total dataset. Sensitivity, also known as recall, focuses on the model's ability to correctly identify actual positives, while specificity measures its success in correctly pinpointing actual negatives. Precision takes into account the accuracy of positive predictions. F1-score, harmonizing precision and sensitivity, provides a balanced view, especially important in situations where there is an uneven class distribution. These metrics, collectively, offer a comprehensive understanding of a model's strengths and weaknesses in classification tasks.

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

Here, TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

$$Sensitivity\ (Sens)\ = \frac{TP}{TP\ +\ FN}$$

$$Specifity\ (Spec)\ = \frac{TN}{TN\ +\ FP}$$

$$Precision\ (Prec)\ = \frac{TP}{TP\ +\ FP}$$

$$F1\ = 2\ x\ \frac{Prec\ x\ Sens}{Prec\ +\ Sens}$$

3. Proposed Model

In the developed model, an initial preprocessing step was applied to the dataset. During this phase, the 'age' variable, which is continuous, was categorized into three groups: young ($18 \leq age \leq 35$), middle-aged ($36 \leq age \leq 55$), and old ($age \geq 56$). All instances of NaN values were replaced with 'unknown'. Subsequently, all data points were transformed into numerical format and underwent normalization. Five distinct machine learning models were then trained using this normalized dataset. Additionally, the classification process involved applying feature reduction to the normalized data using techniques such as PCA and SelectKBest.

For the MLNN model, a structure of two hidden layers was employed, each comprising 100 neurons. The 'tanh' function was utilized as the activation function, and the Adam optimizer was chosen for optimization. In the KNN model, the

number of neighbours was set to 5, and the Minkowski metric was used for distance measurement. In the SVM model, the Radial Basis Function kernel was selected.

4. Results and Discussion

Table 1 provides a baseline comparison of the models using the full feature set. RF shows the highest Acc, Sens and F1 score, indicating its superior performance in correctly classifying instances, particularly in identifying true positives. The MLNN Acc value is quite close to this. SVM, while slightly lower in overall Accuracy, demonstrates the highest Specificity, suggesting its effectiveness in true negative classification.

Table 1. *Machine learning models classified with original data*

	Acc	Sens	Spec	Prec	F1
KNN	89.09	26.95	97.25	56.26	36.45
MLNN	90.32	34.29	97.69	**66.06**	45.14
SVM	89.58	21.24	**98.56**	65.98	32.13
RF	**90.5**	**40.86**	97.02	64.32	**49.97**
LDA	88.92	31.43	96.47	53.92	39.71

Table 2 shows the results of reduced features with PCA. PCA for feature reduction, there is a noticeable decline in Sensitivity across all models, particularly in MLNN, RF and LDA. This suggests that PCA, while reducing dimensionality, may have also omitted features crucial for detecting true positives. The decrease in Accuracy and F1 Scores across models indicates a general reduction in performance due to the loss of informative features.

Table 2. *Classified machine learning models using reduced features with PCA*

	Acc	Sens	Spec	Prec	F1
KNN	88.34	**18.86**	97.47	49.5	**27.31**
MLNN	89.16	13.05	99.16	**67.16**	21.85
SVM	**89.2**	15.33	98.9	64.66	24.79
RF	89.11	13.62	99.02	64.71	22.5
LDA	88.39	0	**100**	0	0

Table 3 shows the results of reduced features with FANOVO. Using FANOVO for feature reduction, there is an improvement in Sensitivity for KNN and MLNN compared to PCA, suggesting that FANOVO retains more relevant information for positive instance classification. However, the Precision and F1 Scores vary, indicating that while FANOVO may preserve certain critical features, it does not consistently enhance overall model performance

Table 3. *Classified machine learning models using reduced features with SelectKBest /* FANOVO

	Acc	Sens	Spec	Prec	F1
KNN	89.19	37.71	95.95	55	44.75
MLNN	**90.19**	**38.76**	96.95	62.52	**47.85**
SVM	89.55	21.71	**98.46**	64.96	32.55
RF	90.07	27.81	98.25	**67.59**	39.41
LDA	89.03	32	96.52	54.72	40.38

Feature reduction using the Chi-squared technique yields mixed results. Table 4 shows the Chi-squared results. The MLNN model shows improved Sensitivity and F1 Score compared to PCA and FANOVO, hinting at Chi-squared 's effectiveness in retaining features important for classification. The consistent low Sensitivity in SVM across all feature reduction methods points to its potential limitations in handling reduced feature sets for identifying true positives.

Table 4. *Classified machine learning models using reduced features with SelectKBest /* Chi-squared

	Acc	Sens	Spec	Prec	F1
KNN	89.17	35.9	96.17	55.2	43.51
MLNN	**90.14**	**46**	95.93	59.78	**51.99**
SVM	89.67	21.71	**98.6**	67.06	32.81
RF	90.12	27.71	98.32	**68.47**	39.46
LDA	89.02	32.19	96.48	54.6	40.5

The comparison illustrates the impact of feature selection and reduction techniques on the performance of machine learning models. MLNN generally maintains higher Accuracy and F1 Scores across different methods, indicating its robustness and adaptability to feature space variations. The fluctuating Sensitivity scores emphasize the need for careful feature selection, especially in applications where accurately identifying true positives is critical. Figure 1 shows the Acc and F1 scores of the MLNN model with Chi-squared for different feature numbers between and 3-14. This analysis highlights the significance of choosing appropriate feature reduction techniques in machine learning. It underscores the necessity of understanding the interplay between feature selection methods and model performance, particularly in domains where the balance between precision and recall is crucial for effective decision-making.

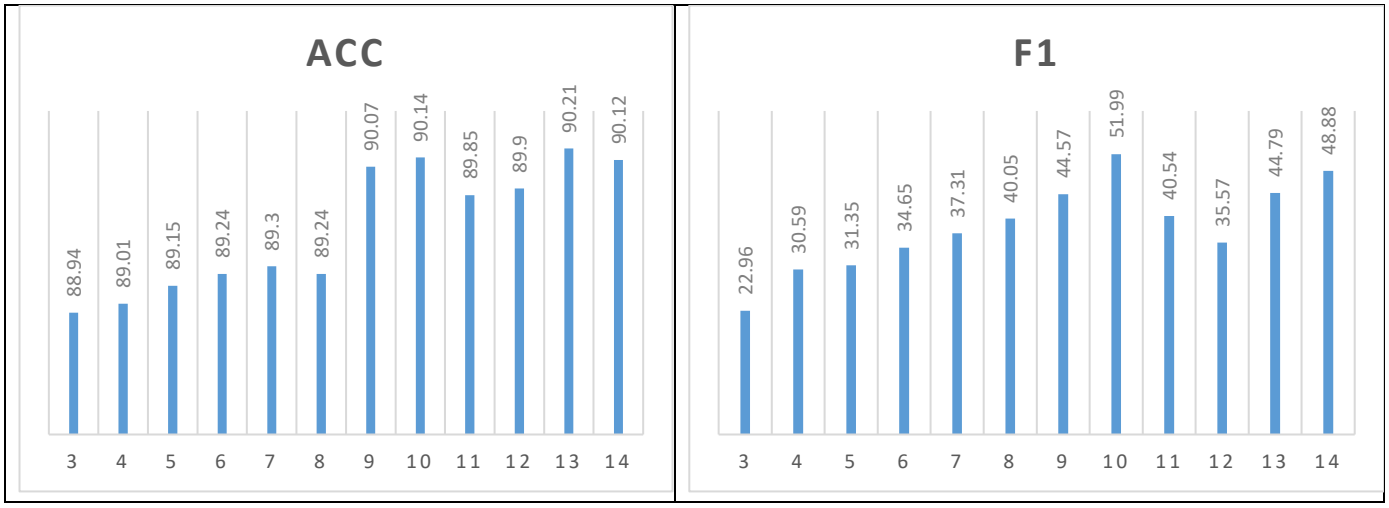

Figure 1. *Acc and F1 scores of the MLNN model with Chi2*

5. Conclusion

The study concludes that feature reduction plays a vital role in optimizing machine learning models for bank marketing. We found that RF exhibited superior performance in terms of Accuracy, Sensitivity, and F1 Score when using the full feature set, while the MLNN model showed considerable efficacy as well. The application of PCA resulted in a notable decrease in model sensitivity, suggesting potential loss of critical information for positive classification. In contrast, the use of FANOVO improved sensitivity in certain models, indicating its effectiveness in retaining relevant features. The Chi-squared method presented mixed results, with MLNN showing improved performance, but SVM consistently exhibited low sensitivity across all reduction methods. These findings underscore the importance of selecting appropriate feature reduction techniques to enhance the predictive power of machine learning models in bank marketing, aiding in the identification of profitable customer segments and the development of targeted marketing strategies.

REFERENCES

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

[2] Kotler, Philip, and Kevin Lane Keller. Marketing Management: Philip Kotler, Kevin Lane Keller. Pearson, 2012.

[3] Rust, Roland T., Christine Moorman, and Gaurav Bhalla. "Rethinking marketing." Harvard business review 88.1/2 (2010): 94-101.

[4] Nobibon, Fabrice Talla, Roel Leus, and Frits CR Spieksma. "Optimization models for targeted offers in direct marketing: Exact and heuristic algorithms." European Journal of Operational Research 210.3 (2011): 670-683.

[5] Seyman M. N., Taşpınar N., (2013), "Channel Estimation Based on Neural Network in Space Time Block Coded MIMO-OFDM System, Digital Signal Processing, Vol.23, No.1, pp. 275-280.

[6] Seyman M. N., Taşpınar N., (2013), "Radial Basis Function Neural Networks for Channel Estimation in MIMO-OFDM Systems", Arabian Journal for Science and Engineering, Vol.38, No. 8, pp. 2173-2178.

[7] Seyman M. N., (2023), "Convolutional Fuzzy Neural Network Based Symbol Detection in MIMO NOMA Systems", Journal of Electrical Engineering, Vol. 74, No. 1, pp. 60-64.

[8] Ozer, Ilyas, Zeynep Ozer, and Oguz Findik. "Noise robust sound event classification with convolutional neural network." Neurocomputing 272 (2018): 505-512.

[9] Ozer, Ilyas, Zeynep Ozer, and Oguz Findik. "Lanczos kernel based spectrogram image features for sound classification." Procedia computer science 111 (2017): 137-144.

[10] Bardak F. K., Seyman M. N., Temurtaş F., (2022), " EEG Based Emotion Prediction with Neural Network Models", Tehnički Glasnik, Vol. 16, No. 4, pp. 497-502.

[11] Ghochani, Mahmood, et al. "Simulation of customer behavior using artificial neural network techniques." International Journal of Information, Business and Management 5.2 (2013): 59.

[12] Kim, YongSeog, et al. "Customer targeting: A neural network approach guided by genetic algorithms." Management Science 51.2 (2005): 264-276.

[13] Ghatasheh, Nazeeh, et al. "Business analytics in telemarketing: Cost-sensitive analysis of bank campaigns using artificial neural networks." Applied Sciences 10.7 (2020): 2581.

[14] Moro,S., Rita,P., and Cortez,P.. (2012). Bank Marketing. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306.