[**Machine Learning Techniques for Predicting the Success of Bank Telemarketing**](https://www.sciencedirect.com/science/article/pii/S016792361400061X)

Ani P. Petkova

Sofia University „St. Kliment Ohridski”,

Faculty of Economics and Business Administration

Sofia, Bulgaria

****

**1. Introduction**

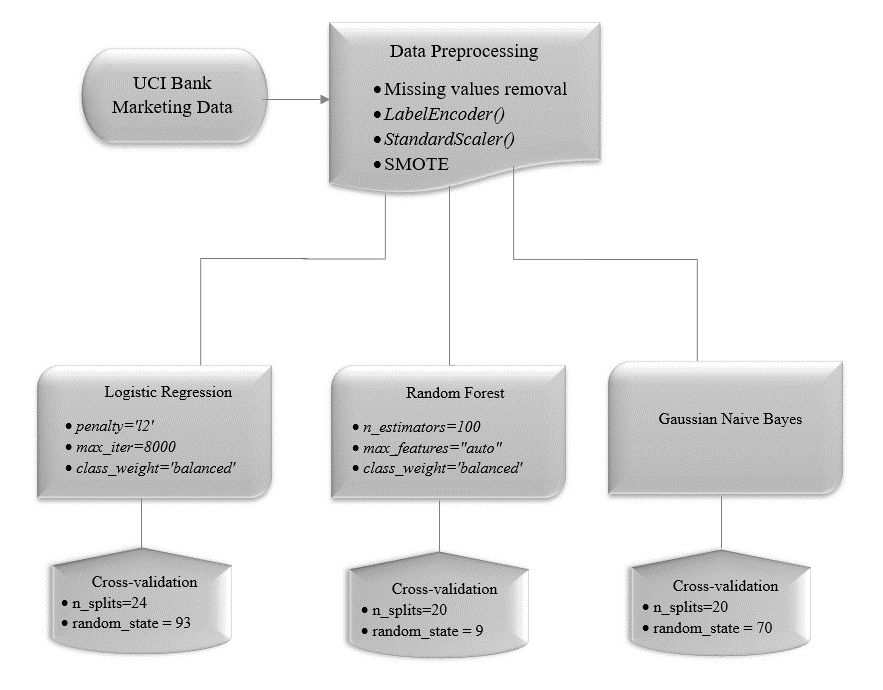
Machine Learning (ML), part of Artificial Intelligence, includes a range of techniques that help computers make informed decisions. [1] ML can be divided to two main types based on different learning approaches: supervised and unsupervised machine learning. Supervised learning is mostly used for the purpose of classification tasks. [2] These classification tasks can be used by companies to help with decision-making processes in order to achieve their goals. To enhance their business and reach specific goals, companies use marketing selling campaigns. A preferred marketing campaign is telemarketing owing to its remote nature. [3] Due to internal competition and different financial crises, financial institutions, such as large retail banks, are sometimes compelled to increase their long-term deposits. Hence, they rely on direct marketing campaigns to secure clients subscribing to a deposit. [4]

Several studies on bank marketing have been conducted in recent years. Notably, Feng et al.(2022) [5] employed a dynamic ensemble selection method, META-DES-AAP, to predict the success of bank telemarketing. This novel model considers both the accuracy of the model and AP maximization. META-DES-AAP model reaches accuracy **89.39%**, AUC **89.44%**, sensitivity **92.62%** and specificity **86.27%**. In previous work, T. F. Bahari et al.(2015) [6] proposed an efficient CRM-data mining framework for the prediction of customer behavior in the domain of banking applications. Two classification models were assessed: Multilayer Perception Neural Network (MLPNN) and Naive Bayes (NB). MLPNN classifier model demonstrated higher accuracy at **88.63%** compared to NB, which achieved **87.97%.**  The rest of the statistical measures namely AUC, sensitivity and specificity for MLPNN are as follows: 84.7%, 40.9% and 94.85%. Similarly, for NB – 85.8%, 47.2%, 93.28%. The model proposed by Feng et al. achieves higher level of accuracy and shows better overall performance when compared to the models introduced by T. F. Bahari et al.

In this paper, a comparison of three ML models (Logistic Regression, Naïve Bayes, and Random Forest) is conducted in order to predict bank telemarketing success. Techniques such as SMOTE and K-fold are utilized to address class imbalance and assess model performance**, aiming to attain higher AUC while maintaining balance between specificity and sensitivity.**

**2. Methodology**

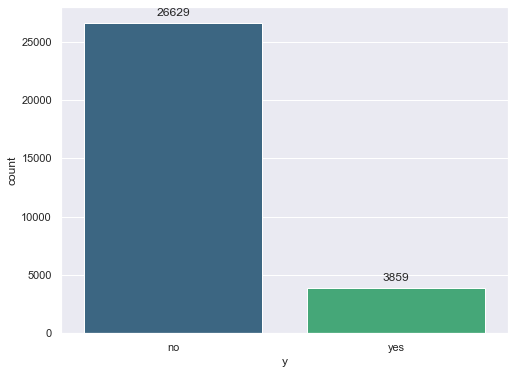
**Figure 1.** Proposed framework

****

Source: Created by the author

For the purpose of the analysis, we utilized a dataset comprising real data collected from a Portuguese retail bank over the period from May 2008 to November 2010, encompassing a total of 41,188 instances and 20 features1 (available at the UCI Machine Learning Repository [7]). There are several missing values in some categorical attributes, all coded with the "unknown" label. Therefore, all observations containing the word "unknown" were removed. Subsequently, a check for any remaining null values in the dataset was conducted, resulting in none found. The dataset is then left with 30,488 instances where only 3,859 records (12.67%) are corresponding to successes, while the remainder is classified as unsuccessful. To mitigate the imbalanced data issue (*Fig.2*) arising from substantial differences between the two samples, resampling techniques were implemented on the dataset. [8]

1­See *Appendix 6.1* for a detailed description of all variables

**Figure 2.** Count plot of target variable

Source: Created by the author

In addition to the imbalance, the data exhibits heterogeneity, compounded by the presence of non-numerical data. Addressing these challenges necessitates a more diverse preprocessing approach.[9] The *LabelEncoder()* function is used for transforming categorical labels into numeric format, making the dataset much easier to work with. Standardization is an extra step taken to tackle this issue. Utilizing *StandardScaler()* to standardize the data helps in centering the data and removing any bias due to varying scales.

Two techniques commonly used to address imbalanced data are oversampling and undersampling. The preferred method is oversampling because, with the undersampling approach, a considerable number of class samples are often removed, resulting in information loss. In contrast, oversampling guarantees the preservation of information integrity. The **Synthetic Minority Oversampling Technique** (SMOTE) stands out as a widely recognized and effective oversampling method for tackling imbalanced learning.[10] SMOTE was implemented from the imbalanced-learn library, with a random seed set to ensure reproducibility *(random\_state=42)*. It synthetically generates instances of the minority class, thereby balancing the distribution of target classes.

In seeking an effective predictive model, three different classifiers were utilized: **Logistic Regression, Random Forest** and **Gaussian Naive Bayes**. The purpose was to evaluate their performance on the classification task.

Logistic Regression, a linear model known for its interpretability, was subjected to a K-Fold cross-validation with 24 folds. The dataset was shuffled with a random state of 93. Logistic Regression was configured with L2 regularization, a maximum iteration limit of 8000, and balanced class weights to address potential imbalances.

The Random Forest ensemble model, chosen for its ability to handle complex relationships, was constructed with a forest of 100 decision trees. The 'auto' setting for maximum features and class weights were set to “balanced”. K-Fold cross-validation with 20 folds and a random state of 9 were applied.

The Gaussian Naive Bayes classifier, known for its simplicity and computational efficiency, was employed.[11] Similarly to the previous approaches, K-Fold cross-validation with 20 folds was utilized to assess its generalization capabilities. The dataset was randomly shuffled to ensure unbiased splits, and the random state was set to 70 to maintain reproducibility.

**3. Results**

Results from all classification models are summarized in *Table 1*. For the purpose of evaluating the models, four measure were used – **AUC, accuracy, sensitivity** and **specificity**.

The **Receiver Operating Characteristic** (**ROC) curve** for binary classification visually represents the trade-off between correctly identifying positive instances (true positive rate) and incorrectly classifying negative instances (false positive rate). The curve is constructed by adjusting the classification threshold across a range of values. In the case of random classification, the ROC curve follows a straight line from the origin to (1, 1). Any deviation above this line indicates an improvement over random performance. **The Area Under the Curve (AUC)** quantifies this improvement, reflecting the quality of the classification in terms of ranking. The higher the AUC, the better the model's ability to distinguish between positive and negative instances. [12]

Classification accuracy, is calculated by summing the True Positives (TP) and True Negatives (TN) and then dividing by the total number of cases (N). It signifies the proportion of correctly classified cases:

*\*Accuracy*=

Sensitivity is determined by the ratio of TP to the sum of TP and False Negatives (FN). It represents the rate of correctly classified positive instances or the True Positive Rate:

*\*Sensitivity =*

Specificity is TN divided by the sum of TN and False Positives (FP). It indicates the rate of correctly classified negative instances or the True Negative Rate:

*\*Specificity =*

\*Equations are as specified in [6].

These metrics offer insights into the model's ability to accurately classify both positive and negative instances.

**Table 1.** Summary of results from current paper

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Classification Accuracy(%)** | **Sensitivity(%)** | **Specificity(%)** | **AUC(%)** |
| **LR** | 85.9 | **93.47** | 84.98 | 94.69 |
| **RF** | **93.11** | 58.53 | **97.27** | **100** |
| **NB** | 83.13 | 62.5 | 86.11 | 85.46 |
|  |  |  |  |  |

Source: Created by the author

At first glance, the RF model could be considered the best-performing model, exhibiting the highest percentages for accuracy, specificity, and AUC – 93.11%, 97.27%, and 100%, respectively. It's worth noting, however, that considering our goal of maximizing AUC values while maintaining a balance between sensitivity and specificity, the RF model performs a lot lower in terms of sensitivity. When dealing with imbalanced datasets, machine learning algorithms often excel in accurately predicting the majority class but struggle to provide accurate predictions for the minority class. Therefore, we select the classification model that gives the best accuracy by increasing sensitivity and specificity. [13]

The LR model demonstrates good overall accuracy and achieves balance between sensitivity and Specificity. Its Area Under the Curve (AUC) value, a measure of its discriminative ability, is also notably high at 94.69%. In this sense, LR best meets out criteria. The Naïve Bayes (NB) classifier exhibits the least favorable performance among the three models. With an accuracy of 83.13%, it falls behind both the Logistic Regression (LR) and Random Forest (RF) models. The AUC value of 85.46% further underscores that the NB classifier has the lowest discriminative power among the models, signaling its comparatively weaker performance in capturing patterns within the data.

ROC curves for each model are attached to the Appendix section for reference.

**4. Discussion**

For a predictive model to be efficient, it should not only achieve high accuracy but also aim for balance between sensitivity and specificity with high AUC values. Low sensitivity indicates that a model is not effectively identifying positive instances (e.g., true positives) among the total positive cases. In other words, the model is missing or misclassifying a significant portion of the actual positive cases.

As seen in  *Table 2,* the worst performance for the sensitivity measure is achieved with ***MLPNN*** model from T. F. Bahari et al.(2015)[6]. The model demonstrates relatively high accuracy, however the MLPNN’s AUC value is the lowest among all models. Similarly, our RF model has a sensitivity of 58.53% which is considered low compared to the specificity 97.27% for the same model. RF scores best results for three of the measures – accuracy, Specificity and AUC, but fails to keep a sensitivity-specificity balance. Another model chosen for comparison is from Feng et al.(2022)[5], who proposed a novel model termed ***META-DES-APP***. While their model manages to keep a sensitivity-specificity balance with good overall accuracy 89.39%, it performs worse than LR and RF in terms of AUC. On the other hand, LR reaches an AUC score of 94.69% and performs best for sensitivity 93.47% while still maintaining balance with specificity. An AUC of 95% indicates a high level of accuracy in predicting the true positive rate while maintaining a low false positive rate.[14]

**Table 2.** Comparison of results from current paper with results from previous relevant papers dealing with direct bank telemarketing prediction using ML techniques

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Classification Accuracy(%)** | **Sensitivity(%)** | **Specificity(%)** | **AUC(%)** |
| **LR** | 85.9 | **93.47** | 84.98 | 94.69 |
| **RF** | **93.11** | 58.53 | **97.27** | **100** |
| **META-DES-AAP1** | 89.39 | 92.62 | 86.27 | 89.44 |
| **MLPNN2** | 88.63 | 40.9 | 94.85 | 84.7 |

1Feng et al.(2022)[5] 2T. F. Bahari et al.(2015)[6]

Source: Created by the author

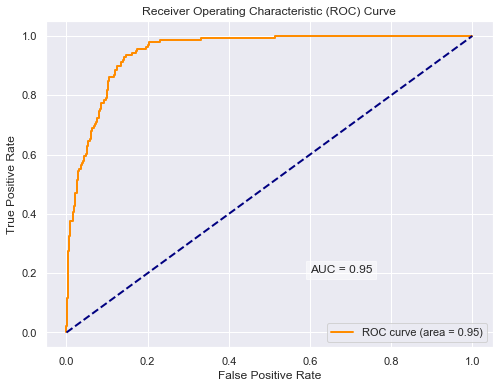
**5. Conclusion**

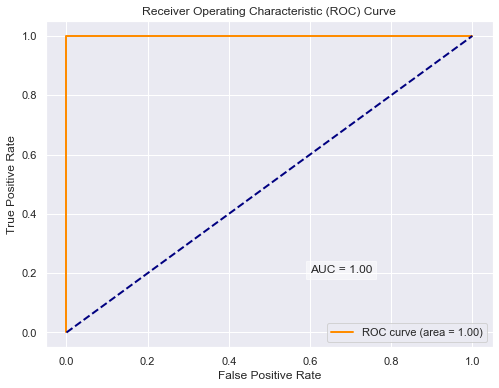
For this study, we emphasize the significance of achieving a balance between sensitivity and specificity, particularly when dealing with imbalanced datasets in the context of predicting bank telemarketing success for long-term deposits. Our choice of a Logistic Regression model is guided by its optimization for both AUC value and sensitivity-specificity balance. The higher AUC value indicates the model's superior ability to distinguish between positive and negative classes, while the careful consideration of sensitivity and specificity ensures a well-rounded performance. This approach is crucial in providing reliable predictions for the success of bank telemarketing campaigns, where imbalances in the dataset could otherwise lead to biased results.

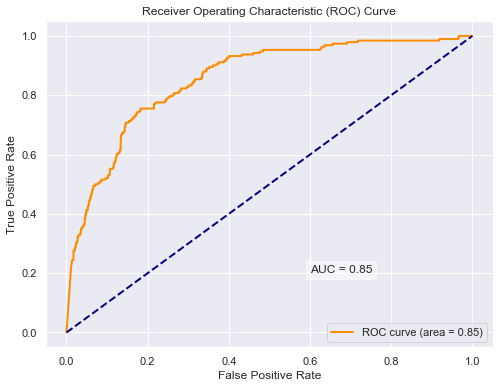
**6. Appendix**

|  |  |  |  |
| --- | --- | --- | --- |
| Number | Name of variable | Type | Description |
| 1 | age | Numeric | Age of the individual |
| 2 | job | Categorical | Type of job |
| 3 | marital | Categorical | Marital status |
| 4 | education | Categorical | Level of education |
| 5 | default | Categorical | Credit in default |
| 6 | housing | Categorical | Housing loan |
| 7 | loan | Categorical | Personal loan |
| 8 | contact | Categorical | Communication type for last contact |
| 9 | month | Categorical | Last contact month of the year |
| 10 | day\_of\_week | Categorical | Last contact day of the week |
| 11 | duration | Numeric | Last contact duration in seconds |
| 12 | campaign | Numeric | Number of contacts during this campaign |
| 13 | pdays | Numeric | Days since the client was last contacted |
| 14 | previous | Numeric | Number of contacts before this campaign |
| 15 | poutcome | Categorical | Outcome of the previous marketing campaign |
| 16 | emp.var.rate | Numeric | Employment variation rate |
| 17 | cons.price.idx | Numeric | Consumer price index |
| 18 | cons.conf.idx | Numeric | Consumer confidence index |
| 19 | euribor3m | Numeric | Euribor 3-month rate |
| 20 | nr.employed | Numeric | Number of employees |
| 21 | y | Binary: “yes”, “no” | has the client subscribed a term deposit? |

**6.2 List of Variables**

**6.2 Logistic Regression ROC Curve**

**6.3 Random Forest ROC Curve**

**6.4 Naive Bayes ROC Curve**

**Acknowledgments**: All materials presented in the Appendix are created by author.

**References**

[1] Kourou K., Exarchos T. P., Exarchos K. P., Karamouzis M. V., Fotiadis D. I. (2015). Machine

learning applications in cancer prognosis and prediction, *Computational and Structural Biotechnology Journal,* Volume 13, Pages 8-17.

[2]Sun Z., Yuan Y., Dong X., Liu Z., Cai K., Cheng W., Wu J., Qiao Z., Chen A. (2023).

Supervised machine learning: A new method to predict the outcomes following exercise intervention in children with autism spectrum disorder, *International Journal of Clinical and Health Psychology,* Volume 23, Issue 4.

[3] Moro S., Cortez P., Rita P., (2014). A data-driven approach to predict the success of bank

Telemarketing. *Decision Support Systems*, Volume 62, Pages 22-31.

[4] Moro S. & Cortez P. & Laureano R. (2011). Using Data Mining for Bank Direct Marketing:

An Application of the CRISP-DM Methodology. *Proceedings of the European Simulation and Modelling Conference.*

[5] Feng Y., Yin Y., Wang D., Dhamotharan L., (2022). A dynamic ensemble selection method

for bank telemarketing sales prediction, *Journal of Business Research*, Volume 139, Pages 368-382.

[6] Bahari T. F., Elayidom M. S.,(2015). An Efficient CRM-Data Mining Framework for the

Prediction of Customer Behaviour, *Procedia Computer Science*, Volume 46, Pages 725-731.

[7] Moro,S., Rita,P., and Cortez,P.. (2012). Bank Marketing*. UCI Machine Learning*

*Repository*. https://doi.org/10.24432/C5K306.(accessed on 16 January 2024).

[8] Safarkhani, F.; Moro, S., (2021). Improving the Accuracy of Predicting Bank Depositor’s

Behavior Using a Decision Tree. *Appl. Sci*., 11, 9016. <https://doi.org/10.3390/> app11199016.

[9] Tékouabou, K. S. C., ¸Gherghina S. C., Toulni H., Mata P. N., Mata M. N., and Martins J.

M.(2022). A Machine Learning Framework towards Bank Telemarketing Prediction. *Journal of Risk and Financial Management* 15: 269.

[10] Sun P., Wang Z., Jia L., Xu Z.,(2024). SMOTE-kTLNN: A hybrid re-sampling method

based on SMOTE and a two-layer nearest neighbor classifier, *Expert Systems with Applications*, Volume 238, Part A.

[11] Blanquero R., Carrizosa E., Ramírez-Cobo P., Sillero-Denamiel M. R., (2021). Variable

selection for Naïve Bayes classification, *Computers & Operations Research*, Volume 135.

[12] Cortes, C., Mohri, M., (2003). AUC optimization versus error rate minimization, *Advances*

*in Neural Information Processing Systems*, Vancouver, Canada, 16.

[13] Garde A., Voss A., Caminal P., Benito S., Giraldo B. F., (2013). SVM-based feature

selection to optimize sensitivity–specificity balance applied to weaning, *Computers in Biology and Medicine*, Volume 43, Issue 5.

[14] Bowers, A.J. and Zhou, X., (2019). Receiver operating characteristic (ROC) area under the

curve (AUC): a diagnostic measure for evaluating the accuracy of predictors of education outcomes. *Journal of Education for Students Placed at Risk* (JESPAR), 24(1), pp.20-46.