Improved Ensemble-Based Online Machine Learning Approach for Telemarketing Success Prediction Using Hyperparameter-Tuned XGBoost

SK. Himambasha[1], P. Sai Ram[2]
Assistant Professor [1], PG Scholar [2]
Department of Master of Computer Applications
QIS College of Engineering & Technology (Autonomous), Ongole
Prakasam Dist., AP

Abstract:

Accurately forecasting potential customers, reducing resource waste, and improving efficiency may all be necessary for telemarketing to be productive. To find clients, we use machine learning models such as Random Forest, XGBoost, AdaBoost, and Gradient Boosting. F1 score, recall, accuracy, and precision are used to assess these algorithms. To address class imbalance in telemarketing datasets, oversampling techniques like SMOTE, ADASYN, and Borderline SMOTE generate synthetic data for underrepresented classes. Model performance is improved by MinMax scaling data normalization and Mutual Information feature selection. Training and testing are conducted using telemarketing data from the Portuguese Bank Marketing Data Set. Modifying the XGBoost hyperparameters raises the model's accuracy to more than 98% for both the top 10 features and the complete set. To promote user involvement, the system has a real-time prediction interface built on Flask. Ensemble-based online machine learning, a successful telemarketing optimization technique, enables models to adjust to shifting customer preferences.

Index terms - *Ensemble Learning; Online Machine Learning; Telemarketing; Class Imbalance; Synthetic Oversampling; SMOTE; ADASYN; Borderline SMOTE; Mutual Information; Feature Selection; Data Normalization; XGBoost; Random Forest; Gradient Boosting; Real-Time Prediction; Flask Interface; Bank Marketing Data*

INTRODUCTION

Direct marketing is a successful marketing strategy that reaches current and future consumers and persuades them to make a purchase by using advertising and a response channel. Cutting-edge technology enable businesses to save and examine campaign data to provide distinctive insights on consumer purchasing trends, brand preferences, pricing, and other factors. They can then use this information to develop focused marketing campaigns that could boost sales. Telemarketing is a popular direct marketing strategy because businesses can easily contact a broad client base and use business intelligence (BI) tools to evaluate data thanks to advancements in information technology and telecommunication. Telemarketing is effectively used by banks, insurance firms, financial institutions, and telecom corporations. According to a survey, 80% of consumers were irritated by telemarketing and thought it would breach their privacy. Other investigations have confirmed these findings. This highlights the necessity of a targeted and astute marketing strategy that targets a certain consumer category based on their financial, social, and personal information pertinent to the product or products.

LITERATURE SURVEY

1 A two-step system for direct bank telemarketing outcome classification:

A two-step method is offered to enhance the forecasting of telemarketing results and assist the marketing management team in efficiently handling client connections in the banking sector. To create preliminary predictions, a number of neural networks are initially trained using various types of data. A single neural network combines all of the initial predictions to create a final prediction in the second stage. Each neural network in the ensemble system has its initial weights optimized via particle swarm optimization. The provided two-step method outperforms all of its separate components, according to empirical findings. Furthermore, the two-step method performs better than a baseline method that trains a single neural network using all kinds of marketing data. The suggested two-step approach is a neural network ensemble model that is quick and simple to deploy, robust to noisy and nonlinear data, and appropriate for big and diverse marketing datasets.

2 Machine Learning for Prediction of Business Company Failure in Hospitality Sector:

It's unclear how company success and failure relate to one another. A significant amount of research has tried to create and evaluate company failure models since Altman's groundbreaking work. Artificial intelligence has showed tremendous promise in recent years, and several studies have proven that machine learning approaches outperform more conventional statistical techniques.

By creating a dataset of financial indicators of businesses for which it is known whether they were successful or not, we address the prediction of business success or failure in a supervised learning framework. 34 financial metrics that describe 1642 enterprises make up our sample.

Based on this information, we developed a number of models to examine how well they predicted fresh, unknown data from other businesses. We assess each model's predictive capacity using the standard classification metrics—precision, recall, f-score, and accuracy—that are employed by the machine learning research community.

3 A decision-making framework for precision marketing:

Through high-efficiency marketing, precision marketing helps businesses boost their revenues while providing individualized customer care. In this study, a novel data-mining-based decision-making paradigm for precision marking is presented. In order to accurately predict monthly supply quantity, this study first presents a trend model; next, it employs an RFM (Recency, Frequency, and Monetary) model to select attributes to cluster customers into different groups; third, it uses Pareto values and CHAID decision trees to identify important attribute values to distinguish different customer groups; and finally, it develops various supply strategies targeting each customer group. The suggested precision-making framework aims to assist managers in recognizing the potential traits of various client categories and proposing suitable precision marketing tactics that can significantly lower inventory for each customer type. In order to demonstrate how to use the suggested framework, real-world data from a Chinese corporation was gathered and utilized in a case study. This case study shows how effective our suggested framework for making decisions is and how well it can provide businesses a precise marketing plan.

4 An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour:

In today's sophisticated business environment, CRM-data mining frameworks manage ties between enterprises and customers and create intimate customer relationships. In recent years, data mining has been more popular in a variety of CRM applications, and one significant data mining approach that is helpful in the industry is the categorization model. In order to improve the decision-making processes for keeping valuable consumers, the model is used to forecast client behavior. This research proposes an effective CRM-data mining framework and examines two classification models, Naïve Bayes and Neural Networks, to demonstrate the superior accuracy of Neural Networks.

5 Information and reformation in KM systems: big data and strategic decision-making:

Purpose

This study provides a theoretical foundation for how knowledge management (KM) systems may help strategic choices incorporate large data. From private to public, for-profit to non-profit, advanced analytics are becoming essential for strategic decision-making. It is uncertain how big data and advanced analytics may influence and, if required, change KM system design and execution, despite the rising importance of gathering, sharing, and implementing people's knowledge in businesses.

Design/methodology/approach

To fill this gap, a mixed strategy was used. Company KM and data analysis systems were evaluated, along with a literature assessment.

Findings

Four data-based judgments and ground principles help KM systems handle huge data and sophisticated analytics.

Practice implications

A realistic approach that accounts for varied data-based judgments is proposed in the study. Restructuring KM systems to include big data and sophisticated analytics into strategic decision-making is suggested.

Originality/value

This is the first advanced analytics-based data-based decision-making typology.

METHODOLOGY

i) Proposed Work:

The proposed method increases the accuracy and efficiency of telemarketing campaigns by utilizing ensemble-based online machine learning algorithms as Random Forest, AdaBoost, Gradient Boosting, and XGBoost. These models are selected because they handle complex and nonlinear data interactions. Synthetic oversampling techniques such as SMOTE, ADASYN, and Borderline SMOTE are used to address the class imbalance in the telemarketing dataset. In order to improve generalization, these methods balance training data. The model is guaranteed to concentrate on the most crucial consumer characteristics when Mutual Information is used for feature selection. Data is normalized for model consistency using MinMax scaling. The Portuguese Bank Marketing Data Set, a well-known standard for telemarketing, is used for training and testing.

In order to enhance model performance and prediction accuracy, XGBoost is hyperparameter tuned in the extended system. The model does well on the top ten most important features as well as the entire feature set. A Flask-based web interface for real-time forecasting allows marketing teams to make quick, informed decisions. This interface is appropriate for real campaigns since it makes it simple to submit test data and provides quick response. It is a robust and scalable solution for contemporary telemarketing optimization because of its adaptable architecture, which permits future changes like adding datasets, customer behavior elements, and algorithms.

ii) System Architecture:

To make telemarketing prediction easier, the system design makes advantage of an end-to-end machine learning pipeline. The first step is data preprocessing, which involves cleaning, standardizing, and balancing the raw Portuguese Bank Marketing Data Set telemarketing dataset using SMOTE, ADASYN, and Borderline SMOTE oversampling. The most crucial characteristics are determined via mutual information. Random Forest, Gradient Boosting, AdaBoost, and an improved XGBoost model with hyperparameters customized for prediction performance are among the ensemble machine learning models that receive data. A Flask-based web interface allows users to upload new data and receive real-time predictions after the models have been trained. This architecture increases the effectiveness of telemarketing campaigns and their response to shifting customer behavior by offering ongoing learning and model upgrades.

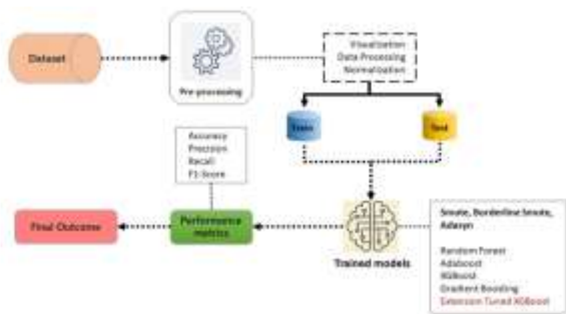

Fig.1. Proposed Architecture

iii) MODULES:

- Data loading: This module will be used to import the dataset.
- Visualization: One method of emphasizing relevance is to plot the selected qualities against their Mutual Information scores. Confusion matrices and ROC curves may also be used to display model performance metrics like recall, accuracy, and precision. Visualizing these KPIs facilitates comprehension of the effectiveness of the telemarketing prediction algorithms.
- Data Processing: A number of essential steps are involved in data processing to ensure high-quality input for machine learning models. The telemarketing dataset must first be loaded and examined for anomalies. characteristics are chosen and normalized using MinMax scaling following the Mutual Information approach's identification of relevant characteristics. Oversampling techniques like SMOTE, ADASYN, and Borderline SMOTE are used to address class imbalance and enhance model training and accuracy by generating synthetic data for underrepresented classes.
- Normalization: Techniques like MinMax scaling are used to standardize feature values within a specific range, often between 0 and 1. This process increases the overall accuracy of telemarketing forecasts by ensuring that each feature contributes equally to the model's performance, particularly in algorithms that are sensitive to feature sizes.
- Dividing data into train and test: This module will be used to divide data into train and test.
- Model creation: Gradient Boosting, Extension Tuned XGBoost, Adasyn Random Forest, Adaboost, Smote, and Borderline Smote are some of the model construction tools. The performance evaluation metrics for each algorithm are calculated.
- Admin login: The administrator can log in using this module.
- Drug Side Effect Prediction: Users can upload test results to this module.

- Prediction: The ultimate forecast was displayed.

iv) **ALGORITHMS:**

- Random Forest: To predict the likelihood that a customer will make a purchase, Random Forest builds many decision trees and combines their output. This ensemble approach increases accuracy and robustness by decreasing overfitting. The model's ability to identify patterns in both the majority and minority classes is improved by balancing classes utilizing methods such Border SMOTE and ADASYN, which improves prediction performance.
- XGBoost: XGBoost is utilized due to its effectiveness and efficiency in handling large datasets. Gradient boosting is used to consistently minimize errors, leading to a high degree of accuracy when identifying possible customers. The method employs a number of techniques, including Border SMOTE and ADASYN, to address class imbalance. This ensures accurate forecasts and allows it to perform well even with limited data for minority groups.
- AdaBoost is used to enhance the performance of weak classifiers by combining their outputs into a strong learner. Our system improves the overall accuracy of identifying potential clients by focusing on examples that were misclassified. By using techniques like Border SMOTE and ADASYN, AdaBoost addresses class imbalance and ensures that minority class samples are effectively learned—a crucial element of effective telemarketing.
- Gradient Boosting: Gradient boosting builds a prediction model by adding decision trees one after the other, each of which fixes the mistakes of the preceding tree. This method improves the accuracy and flexibility of the model. By balancing class distributions, oversampling techniques like SMOTE and ADASYN enable the model to efficiently gather data from both majority and minority classes for improved customer targeting.
- Tuned XGBoost: Tuned XGBoost enhances the predictive power and performance of the model by optimizing hyperparameters. This approach ensures that the most relevant attributes are prioritized when combined with techniques like Top 10 Border SMOTE and Top 10 ADASYN, increasing the accuracy and efficacy of predicting potential customers and their purchasing habits.

EXPERIMENTAL RESULTS

The system architecture uses an end-to-end machine learning pipeline to optimize the telemarketing prediction process. Data preparation is the first step. The Portuguese Bank Marketing Data Set's raw telemarketing dataset is cleaned, standardized using MinMax scaling, and balanced using oversampling methods including SMOTE, ADASYN, and

ISSN: 3049-0952
www.jsetms.com

Borderline SMOTE. The Mutual Information approach is then used for feature selection in order to extract the most pertinent properties. Random Forest, Gradient Boosting, AdaBoost, and an enhanced XGBoost model—which is adjusted for hyperparameters to improve prediction performance—are among the ensemble machine learning models that get the processed data. After the models are trained, real-time engagement is made possible with a Flask-based web interface that lets users submit fresh data and get predictions right away. This architecture ensures adaptation to shifting consumer behavior and boosts the efficacy of telemarketing campaigns by supporting ongoing learning and model changes.

Accuracy: A test's accuracy is determined by its capacity to distinguish between healthy and ill cases. To gauge the accuracy of the test, find the percentage of examined instances that had true positives and true negatives. According to the computations:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

Precision: Precision is the number of affirmative cases or the classification's accuracy rate. The following formula is applied to assess accuracy:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: A model's ability to recognise every instance of a pertinent machine learning class is measured by its recall. The ratio of accurately predicted positive observations to the total number of positives indicates how well a model can identify class instances.

$$Recall = \frac{TP}{(FN + TP)}$$

mAP: Mean Average Precision is one ranking quality metric (MAP). It considers the number of relevant recommendations and their position on the list. MAP at K is calculated as the arithmetic mean of the Average Precision (AP) at K for each user or query.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$
$$AP_k = \text{ the AP of class } k$$
$$n = \text{ the number of classes}$$

F1-Score: An accurate machine learning model is indicated by a high F1 score. combining precision and recall to increase model correctness. The accuracy statistic quantifies the frequency with which a model correctly predicts a dataset.

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\,cision)}{(Recall + \Pr e\,cision)}$$

CONCLUSION

In conclusion, the recommended strategy significantly increases telemarketing success by utilizing state-of-the-art machine learning algorithms to predict potential customers' propensity to make purchases. By combining techniques like SMOTE, ADASYN, and Borderline SMOTE with ensemble approaches like Random Forest, XGBoost, AdaBoost, and Gradient Boosting, the system successfully corrects class imbalance while optimally detecting interested customers. Applying the Mutual Information technique to feature selection ensures that the most relevant features are prioritized, improving model performance. In several settings, such as All Features SMOTE and Top 10 Features Borderline SMOTE, the optimized XGBoost model attains above 98% accuracy. The Portuguese Bank Marketing Data Set offers a solid foundation for testing and training, and the outcomes are superb. The use of MinMax scaling for data normalization and hyperparameter modification further enhances model accuracy and ensures that the predictive capabilities adapt to shifting customer preferences. The inclusion of a Flask-based interface, which enables efficient user interaction and dynamic adjustments based on user behavior, makes real-time predictions possible. All things considered, this comprehensive strategy provides a robust and effective means of optimizing telemarketing strategies, which improves resource efficiency and increases marketing campaign ROI, ultimately revolutionizing telemarketing practices in the industry.

FUTURE SCOPE

The future scope of this system will include integrating more data sources, such customer feedback and social media interactions, to increase prediction accuracy. Examining complex algorithms, such deep learning techniques, might improve client preference determination performance. Additionally, using real-time data for immediate insights and personalization can enhance marketing strategies. Examining automated customer segmentation based on preferences and behavior might improve targeting efforts. Last but not least, enhancing the Flask interface's visualization features might provide deeper insights and let marketers to make educated decisions more quickly and effectively, boosting the effectiveness of telemarketing campaigns.

REFERENCES

[1] N. Ghatasheh, I. Altaharwa, and K. Aldebei, Modeling the telemarketing process using genetic algorithms and extreme boosting: Feature selection and cost-sensitive analytical approach, IEEE Access, vol. 11, pp. 67806–67824, 2023.

[2] S. C. K. Tékouabou, Ş. C. Gherghina, H. Toulni, P. N. Mata, M. N. Mata, and J. M. Martins, A machine learning framework towards bank telemarketing prediction, J. Risk Financ. Manag., vol. 15, no. 6, p. 269, 2022.

[3] Y. Feng, Y. Yin, D. Wang, and L. Dhamotharan, A dynamic ensemble selection method for bank telemarketing sales prediction, J. Bus. Res., vol. 139, pp. 368–382, 2022.

[4] S. E. Saeed, M. Hammad, and A. Alqaddoumi, Predicting customer's subscription response to bank telemarketing campaign based on machine learning algorithms, in Proc. 2022 Int. Conf. Decision Aid Sciences and Applications (DASA), Chiangrai, Thailand, 2022, pp. 1474–1478.

[5] A. Vitorio and G. Marques, Impact of imbalanced data on bank telemarketing calls outcome forecasting using machine learning, in Proc. 2021 Int. Conf. Data Analytics for Business and Industry (ICDABI), Sakheer, Bahrain, 2021, pp. 380–384.

[6] S. Lahmiri, A two-step system for direct bank telemarketing outcome classification, Int. J. Intell. Syst. Account. Finance Manag., vol. 24, no. 1, pp. 49–55, 2017.

[7] J. H. Brito, J. M. Pereira, A. F. D. Silva, M. J. Angélico, A. Abreu, and S. Teixeira, Machine learning for prediction of business company failure in hospitality sector, in Advances in Tourism, Technology and Smart Systems, Á. Rocha, A. Abreu, J. V. D. Carvalho, D. Liberato, E. A. González,

and P. Liberato, eds. Singapore: Springer, 2019, pp. 307–317.

[8] Z. You, Y. W. Si, D. Zhang, X. Zeng, S. C. H. Leung, and T. Li, A decision-making framework for precision marketing, Expert Syst. Appl. Int. J., vol. 42, no. 7, pp. 3357–3367, 2015.

[9] T. F. Bahari and M. S. Elayidom, An efficient CRM-data mining framework for the prediction of customer behaviour, Procedia Comput. Sci., vol. 46, pp. 725–731, 2015.

[10] A. Intezari and S. Gressel, Information and reformation in KM systems: Big data and strategic decision-making, J. Knowl. Manag., vol. 21, no. 1, pp. 71–91, 2017.

Author Profiles

SK. Himambasha is an Assistant Professor in the Department of Master of Computer Applications at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. He earned his Master of Computer Applications (MCA) from Anna University, Chennai. With a strong research background, He has authored and co-authored research papers published in reputed peer-reviewed journals. His research interests include Machine Learning, Artificial Intelligence, Cloud Computing, and Programming Languages. He is committed to advancing research and fostering innovation while mentoring students to excel in both academic and professional pursuits.

P. Sai Ram is an MCA Student in the Department of Computer Application at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. He has Completed Degree in MPCS from Sri Harshini Degree College Ongole, Prakasam district.