Machine Learning in Banking: Enhancing Marketing Campaigns Through Predictive Analytics

Divyansh Garg[1], Kunjal Joshi[2*], Janhvi Vanga[3] , Avinash Tandle[4]

[1,2,3,4] Mukesh Patel School of Technology Management and Engineering, SVKM's NMIMS, 40058 Mumbai, India

[1]divyansh.garg06@nmims.in,

[2*]kunjal.joshi48@nmims.in,[3]janhvi.vanga97@nmims.in, [4]avinash.tandle@nmims.in

Abstract. Machine Learning (ML) has revolutionized the banking sector by enabling precise customer segmentation, marketing, and fraud analysis. This study entails predicting customer churn from a credit card customer dataset, which is demographic, transactional, and account-based. Supervised Machine Learning algorithms were applied to classify customers according to their churn likelihood. Results highlight the strength of ensemble learning methods like XGBoost and Random Forest in improving classification results. Additionally, this study explores the broader impact of ML on the banking sector, including the optimization of marketing campaigns, enhanced fraud protection, and reduced operational costs. Spending on ML-based banking solutions is substantial, with major banks like JPMorgan Chase integrating AWS AI tools for massive data processing and Commonwealth Bank of Australia allocating nearly $1.1 billion to technology spending. Projections indicate that AI is expected to yield cost savings of up to $1 trillion by 2030 and an estimated profit increase of $340 billion by 2025. These findings provide financial institutions with datadriven insights to improve customer retention and efficiency

Keywords: Machine Learning (ML), Banking Industry, Customer Segmentation, Targeted Marketing, Fraud Detection.

1 Introduction

The banking sector is increasingly deploying ML to engage customers better and streamline operations. Through the study of enormous volumes of data, ML models aid banks in determining patterns, categorizing customers in terms of their behavior, and applying focused marketing [1]. This paper aims to determine customer churn by using a dataset that captures demographic characteristics (age, sex, income), transactional behavior (credit limit utilization, usage patterns), and engagement metrics (account activity, customer interactions) [2]. The intention is to categorize customers based on their potential to churn and facilitate pre-emptive retention efforts.

Machine learning-based predictive analytics enable banks to predict customer requirements, increasing loyalty and sales [3]. Ensemble learning techniques, in particular, have been helpful in identifying financial distress, which makes early intervention possible [4]. The Bank Marketing Dataset, which consists of 45,211

S. Sharma et al. (eds.), *Proceedings of the International Conference on Recent Advances in Intelligent and Sustainable Technologies (RAIST 2026)*, Atlantis Highlights in Intelligent Systems 17,
https://doi.org/10.2991/978-94-6239-707-1_13

records, has proved effective in predicting the success of telemarketing campaigns. Likewise, MLbased methods have been useful for classifying customers in order to forecast repeat buyers, which improves targeted marketing. Additionally, ML increases fraud detection by identifying anomalies in transaction data, protecting both financial institutions and consumers [5]. Ensemble learning methods have been shown to efficiently detect fraudulent activities and dramatically enhance security measures for banking systems. Artificial intelligence-powered automation has also brought cost savings, with estimates suggesting savings of as much as $1 trillion by 2030 [6]. Commonwealth Bank of Australia's AI solutions have resulted in a reduction of 40% in call center wait times, enhancing customer experience [7]. Rocket Mortgage has used AI to maximize call center performance, making operations more efficient.

These developments highlight the revolutionary function of ML in contemporary banking and render it a critical tool for customer satisfaction, marketing efficiency, fraud detection, and substantial reductions in operational costs. This research uses several ML algorithms, compares their performance, and examines their implications for financial operations and customer retention.

2 Literature Review

The use of machine learning in banking applications has gradually gained momentum, particularly for improving marketing campaigns and reaching out to customers. Different predictive models have been investigated to help banks identify potential customers more effectively and maximize outreach.

Moro et al. [8] investigated the use of data mining approaches to improve bank telemarketing campaigns. The authors concluded that machine learning models such as Decision Trees and Neural Networks were able to predict customer responses more accurately than traditional approaches. They highlighted that past interactions and customer demographics are critical factors in determining the success of marketing campaigns. Verbraken et al. [9] highlighted the problem of imbalanced datasets in banking applications, where very few customers respond positively to marketing campaigns. They noted that approaches such as Synthetic Minority Over-sampling Technique (SMOTE) improved prediction accuracy by balancing datasets. Similarly, Risselada et al. [10] compared different machine learning models for customer retention and concluded that ensemble models such as Random Forest performed better in identifying future customers than traditional regression models. A systematic review conducted by Fernández-Delgado et al. [11] compared different machine learning models and concluded that ensemble models, specifically Random Forest and XGBoost, performed better than other models in handling.Their study emphasized the need for choosing appropriate features and model parameter tuning for the best performance. Kumar and Ravi [12] investigated the influence of feature engineering on predictive performance in banking use cases. They established that categorical variable encoding, numerical feature standardization, and removing redundant data highly enhanced model efficiency. Their research corroborates the increasing trend of

using ensemble models in financial forecasting. Aside from structured data, Natural Language Processing (NLP) in banking has been the focus of recent studies. Syarif et al. [13] showed that sentiment analysis of customer feedback can assist banks in tailoring their marketing efforts, resulting in improved customer interaction and increased conversion rates.Though these studies demonstrate the possibilities of machine learning in banking, real-time execution and evolving customer behavior are issues yet to be resolved. Research in the future could focus on deep learning algorithms and reinforcement learning to design adaptive, targeted marketing campaigns. With further developments in machine learning, its place in banking is going to become even more central to optimizing customer interaction and enhancing business results.

3 Methodology

The work is based on a customer marketing dataset provided by a financial institution. It contains basic demographic information such as age, job, marital status, and education, along with financial details including loan and housing status. Campaign-related attributes like number of contacts, previous responses, and last interaction are also included. The target variable y shows whether a customer subscribed to the product, so the task is treated as binary classification.

3.1 Data Loading and Initial Exploration

The data was read into Python Using Pandas. Initial checks were done to look at column types, summary statistics, and missing values. Numerical and categorical columns were reviewed separately to get a general idea of their spread.

3.2 Exploratory Data Analysis (EDA)

EDA was carried out to study feature distributions and their relationship with the target variable. A correlation analysis was performed to identify multicollinearity among numerical features, and highly correlated variables were removed. Categorical variables were grouped where required to reduce sparsity. Outliers in numerical columns were detected using the Interquartile Range (IQR) method.

- Figure 1: Correlation Heatmap: The heatmap illustrates relationships among numerical variables and reveals strong correlations between selected economic features, indicating potential redundancy.
- Figure 2: Education after Simplification: After merging similar categories, the grouped education levels provide clearer trends in customer subscription.
- Figure 3: Job Title Distribution: Subscription responses across job types indicate differing success rates among occupational categories.

- Figure 4: Outlier Detection (Boxplot): Boxplots identify extreme values in numerical features such as age, campaign, and duration, which were handled during preprocessing

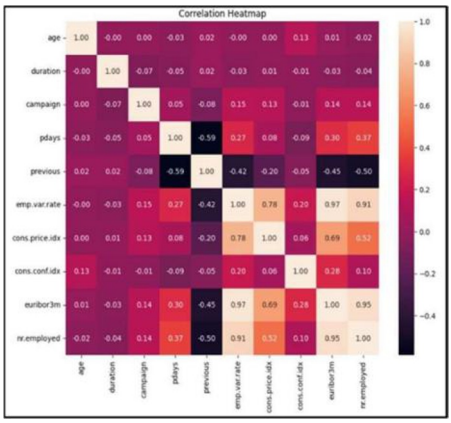

Fig. 1. Correlation Heatmap

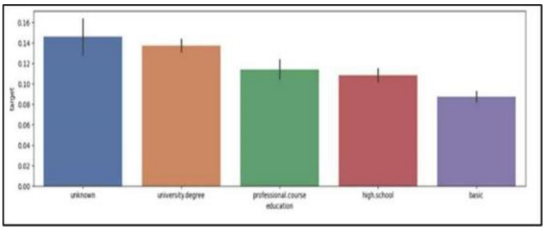

Fig. 2. Bar plot of Education after Simplification

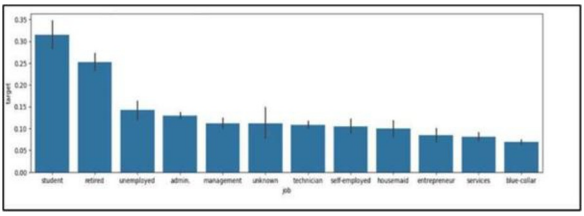

Fig. 3. Bar plot of Job title

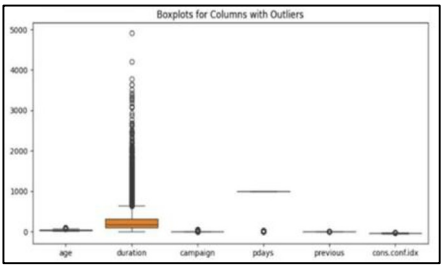

Fig. 4. Box and Whisker Plot for outlier detection

3.3 Data Cleaning and Preprocessing

The target variable was converted into binary form (yes = 1, no = 0). Similar categorical levels were merged to reduce sparsity, and unknown values were replaced with the most frequent category. Categorical features were encoded using label encoding for model compatibility. Highly correlated attributes were removed to avoid redundancy, and outliers were eliminated using the Interquartile Range (IQR) method. These preprocessing steps improved data consistency and overall model reliability.

3.4 Feature Scaling and Handling Class Imbalance

As class imbalance can occur in several financial marketing datasets, this research uses the Synthetic Minority Over-sampling Technique (SMOTE) to create synthetic samples of the minority class so models can learn patterns in the data effectively. Feature scaling is used through the StandardScaler, especially for such models as KNearest Neighbors (KNN) and Logistic Regression, which are scale-sensitive and will be affected by variations in scale. This converts numerical features to a standard distribution of mean zero and variance one, thus improving model convergence while training.

3.5 Model Selection, Training and Evaluation

Several machine learning models are trained and tested for predictive accuracy:

- K-Nearest Neighbors (KNN): A classifier based on distance that predicts labels based on the majority vote of k-nearest instances.
- Logistic Regression: A linear model that estimates the probability of customer subscription via the logistic function.
- Decision Tree Classifier: A tree-based model that splits data according to feature importance to predict customer classification.
- Random Forest Classifier: A type of ensemble learning that creates many decision trees and takes their predictions' average for better accuracy.
- XGBoost Classifier: A gradient boosting algorithm that improves decision tree performance by iterative learning.

Hyperparameter tuning is done on the n_neighbors (number of neighbors), weights (distance based or uniform), and algorithm (ball tree, KD tree, brute force) for KNN to select model fitting options to improve model performance.

4 Results and Discussion

This section contains the performance comparison of different machine learning models applied for predicting customer subscription propensity in banking campaigns. The comparison has been based on the confusion matrices, accuracy, precision, recall, and F1-score for all models. The optimal set of hyperparameters for each model has been calculated, allowing it to perform at its best.

Table 1. Modals Metrics Comparison

Modals	Mean Accuracy	Precision	Recall	Specificity	F1-Score
KNN	94.34%	90.75%	98.00%	89.96%	94.23%
Logistic Regression	87.56%	86.80%	88.69%	86.46%	87.74%
Decision Tree	95.55%	94.78%	96.45%	94.65%	95.61%
Random Forest	98.24%	95.06%	98.05%	94.92%	96.54%
XGBoost	98.85%	96.35%	98.19%	95.80%	97.26%
S.V.M	95.60%	91.90%	96.50%	92.40%	94.00%

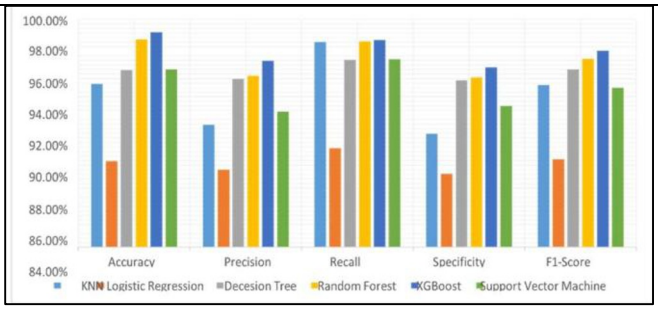

Fig. 5. Modals Metrices Comparison

The KNN model did well with an accuracy of 94.34 percent. The KNN model also had recall. However, the specificity of the KNN model was a little lower. This meant the KNN model had some positives. The Logistic Regression model had results with 87.56 percent accuracy. The Logistic Regression model was good at balancing precision and recall. The Logistic Regression model was not very good at capturing complex decision boundaries. The KNN model and the Logistic Regression model are different.

The KNN model is better at some things. The Logistic Regression model is better, at other things. The Decision Tree did a job of classifying things with a nice balance between being precise and remembering everything. However, the Decision Tree had a problem, with overfitting. The Random Forest was even better. Got it right most of the time with a score of 98.24 percent. The Random Forest reduced the chances of getting results each time and made more stable predictions.. The Random Forest took more time to think about it and used more com-
puter power. XGBoost delivered the best overall results, achieving the highest accuracy and a well-balanced precision–recall score, demonstrating strong generalization across classes. SVM also achieved competitive accuracy (95.60%) but performed slightly below the ensemble models.

Figure 6 to Figure 11 shows the confusion Matrix of the various models.

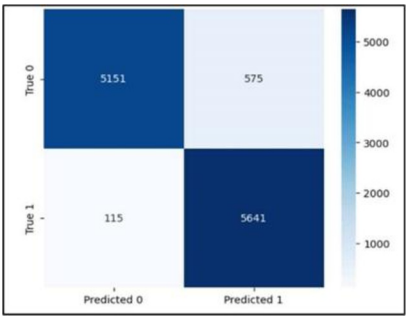

Fig. 6. Confusion Matrix of KNN

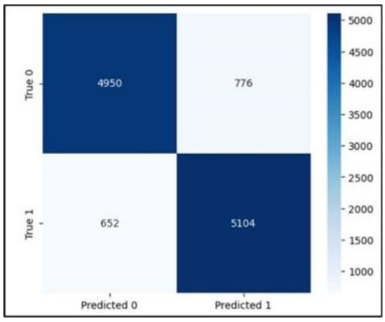

Fig. 7. Confusion Matrix of Logistic Regression

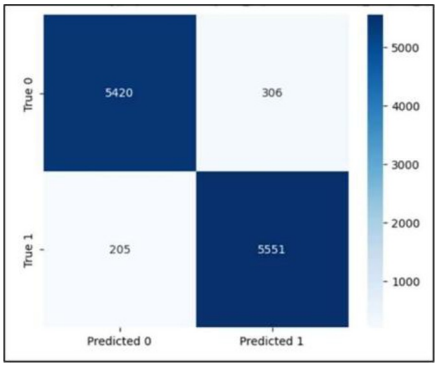

Fig. 8. Confusion Matrix of Decision Tree

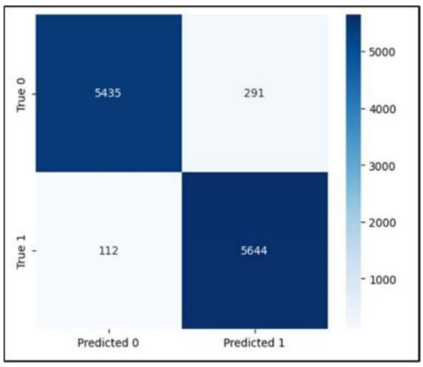

Fig. 9. Confusion Matrix of Random Forest

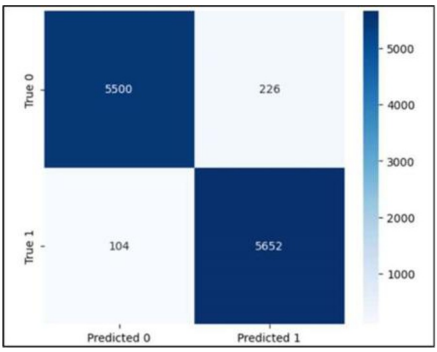

Fig. 10. Confusion Matrix of XGBoost

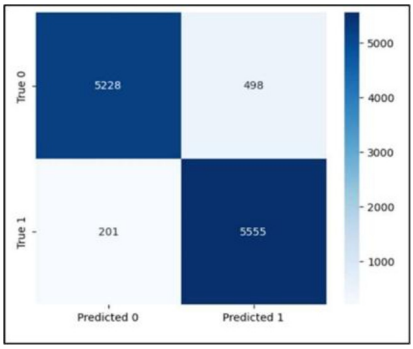

Fig. 11. Confusion Matrix of SVM

Out of all models, XGBoost was the highest scorer with the highest accuracy and F1score. Random Forest followed closely, then Decision Tree and SVM. KNN and Logistic Regression performed moderately, which suggests that the simpler models lack
the capability to deal with complicated patterns. These findings highlight the efficacy of ensemble methods in improving classification performance through minimization of overfitting and generalization.

5 Conclusions

This study investigated the use of machine learning methods to forecast customer subscription likelihood in banking campaigns based on a dataset with demographic, transactional, and account-related attributes. The dataset of-freed insightful information about customer behavior, such as age, income, type of job, credit usage, and previous marketing interactions. Analyzing these variables helps financial institutions understand customer interaction better and optimize marketing efforts to increase subscription rates. The research il-lustrated that machine learning models can adequately segment customers depending on their prospect of subscribing to a product offered by a bank, enabling them to manage their resources and orient their offerings appropriately to the appropriate customer profiles.

Among the models assessed, XGBoost was found to be the most effective classifier - achieving the highest accuracy and balance between recall and precision. Its ability to detect nuanced interactions between a variety of features made it a natural fit for financial contexts in which accuracy is critical. Random Forest, another ensemble model was also an effective model, providing high classification while having less chances for overfitting than the decision tree. Decision trees are easy to understand and can learn non-linear relationships but can easily overfit if not pruned. Logistic Regression, a very common approach for financial modeling performed reasonably but did not detect complicated relationships. K-Nearest Neighbors (KNN), was

computationally expensive in circumstances with full dimensionality data and therefore not appropriate for larger use cases in Banking.

Additionally, machine learning models, like those studied in this re-search, can help banks mitigate capital risks and better detect fraud or possible attempts at fraud by recognizing inconsistencies between customer transactions. The predictive capabilities of churn rate detection can allow banks to predict customer problems in advance, promote differentiated solutions, and ensure ongoing loyalty. The fact that the dataset is comprised of historic and behavioral features highlights its educational value for banks looking to improve marketing, loan authorization, credit risk management, and customer service

In the future, with the incorporation of real time transactional data, external economic environmental states, and customers' sentiment analysis, will further improve predictive modelling. Using advanced techniques like deep learning techniques, reinforcement learning techniques, and hybrid machine learning type AI methods could improve predictive accuracy and robustness in banking. In addition, using Explainable AI (XAI) frameworks would really improve compliance and transparency in line with financial regulations, and make machine learning-based marketing strategies more reliable and understandable to consumers. As the finance industry is increasingly applying AI, the role of predictive analytics will certainly extend in terms of customer service interactions, fraudulent avoidance, and developing bespoke banking products; paving the way for even cleverer, more customer-centric financial products.

References

1. Nguyen, V.H.: Customer churn prediction in the banking sector using machine learningbased classification models. ResearchGate (2023). Available at: https://www.researchgate.net/publication/368911804
2. Datrics: Bank churn prediction: Using ML to retain customers. Datrics (2023). Available at: https://www.datrics.ai/articles/bank-churn-prediction-using-ml-to-retain-customers
3. Predicting and preventing customer churn in retail banking using machine learning. Global Banking & Finance Review (2023). Available at: https://www.globalbankingandfinance.com/predicting-and-preventing-customer-churn-inretail-banking-using-machine-learning
4. AI in banking: Driving efficiency and innovation. FinTech Magazine (2025). Available at: https://fintechmagazine.com/articles/ai-in-banking-driving-efficiency-and-innovation
5. AI and the banking industry's $1 trillion opportunity. The Financial Brand (2023). Available at: https://thefinancialbrand.com/news/artificial-intelligence-banking/artificialintelligence-trends-banking-industry-72653
6. Eyers, J.: Commonwealth Bank bets big on artificial intelligence. The Australian (2025). Available at: https://www.theaustralian.com.au/business/financial-services/commonwealth-bank-unveils-ai-push-in-banking-arms-race
7. Santomassimo, M.: AI will affect nearly "every part" of Wells Fargo, CFO says. Barron's (2025). Available at: https://www.barrons.com/articles/wells-fargo-ai-banking-a6f84c15

8. Moro, S., Laureano, R., Cortez, P.: Using data mining for bank direct marketing: An application of the CRISP-DM methodology. Expert Systems with Applications 37(9), 12293–
12301 (2014). https://doi.org/10.1016/j.eswa.2014.02.034
9. Verbraken, R., Verbeke, W., Baesens, B.: A novel profit maximizing metric for measuring classification performance of customer churn prediction models. IEEE Transactions on Knowledge and Data Engineering 25(5), 961–973 (2013). https://doi.org/10.1109/TKDE.2012.64
10. Risselada, H., Verhoef, P., Bijmolt, B.: Staying power of churn prediction models. Journal of Interactive Marketing 24(3), 198–208 (2010). https://doi.org/10.1016/j.intmar.2010.07.005
11. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real-world classification problems? Journal of Machine Learning Research 15, 3133–3181 (2014). Available at: https://www.jmlr.org/papers/volume15/delgado14a/delgado14a.pdf
12. Kumar, M., Ravi, V.: A survey of the applications of text mining in financial domain. Knowledge-Based Systems 114, 128–147 (2016). https://doi.org/10.1016/j.knosys.2016.10.003
13. Syarif, I., Prugel-Bennett, A., Wills, G.: Data mining approaches for network intrusion detection. Information Security Journal: A Global Perspective 21(2), 84–94 (2012). https://doi.org/10.1080/19393555.2011.654304