Machine Learning Approaches for Predicting Bank Term Deposit Subscriptions in Direct Marketing

ISBN 978-1-943295-24-1

B. Keerthana
K. Murugan
K. Kiran
B. Hariharasudhan
Sri Sai Ram Engineering College
(keerthana.mba@sairam.edu.in)
(Murugan.mba@sairam.edu.in)
(Kiran.mba@sairam.edu.in)
(sec23mb179@sairatap.edu.in)

1. Introduction

In recent years, the advent of machine learning (ML) algorithms has revolutionized various industries, including banking and finance, by enabling the extraction of valuable insights from vast datasets. One critical application of ML in the banking sector is in direct marketing, where predictive models are employed to anticipate customer behaviours, such as subscribing to bank term deposits. This research delves into the utilization of ML algorithms, specifically the Stochastic Gradient Descent (SGD) Classifier, k-nearest neighbour (KNN) Classifier, and Random Forest Classifier, in direct marketing for anticipating bank term deposit subscriptions, with a focus on the Indian banking landscape. The dataset utilized for this study is obtained from Kaggle, a renowned platform for datasets and data science competitions. Direct marketing plays a pivotal role in the banking sector, where financial institutions strive to optimize their marketing strategies to attract and retain customers. Term deposits, a popular investment avenue offered by banks, represent a significant source of revenue and stability for financial institutions. Predicting which customers are likely to subscribe to term deposits allows banks to target their marketing efforts more effectively, thereby maximizing the return on investment (ROI) of their marketing campaigns. Traditional approaches to direct marketing often rely on demographic data and past transaction history. However, with the proliferation of digital channels and the availability of large-scale datasets, Zaki, A. M., et. al., (2024) ML techniques have emerged as a powerful tool for enhancing the accuracy and efficacy of direct marketing campaigns.

The Indian banking sector has witnessed substantial growth in recent years, driven by factors such as increasing digitization, rising disposable incomes, and government initiatives to promote financial inclusion. With a large and diverse population, India presents a unique and challenging market for banks to navigate. Understanding the preferences, behaviours, and needs of Indian consumers is essential for banks to tailor their marketing strategies effectively. By leveraging ML algorithms, banks can analyse the vast amounts of data generated by customers' interactions with various channels, including mobile banking, internet banking, and social media, to gain actionable insights into customer behavior and preferences. The Stochastic Gradient Descent (SGD) Classifier is a popular ML algorithm widely used for classification tasks in large-scale datasets. It operates by iteratively updating the model parameters to minimize the loss function, making it particularly well-suited for training on large datasets with high dimensionality. In the context of direct marketing for bank term deposit subscriptions, the SGD Classifier can be trained on historical customer data to predict the likelihood of a customer subscribing to a term deposit based on their demographic information, past transaction history, and other relevant features. Another ML algorithm commonly employed in classification tasks is the k-nearest neighbour (KNN) Classifier. Unlike parametric models like the SGD Classifier, the KNN Classifier is a non-parametric algorithm that makes predictions based on the similarity of data points in the feature space. The KNN Classifier assigns a class label to a given data point based on the class labels of its k nearest neighbour in the training dataset. In the context of bank term deposit subscriptions, the KNN Classifier can be used to identify customers with similar characteristics to those who have previously subscribed to term deposits, thereby enabling targeted marketing efforts. The Random Forest Classifier is an ensemble learning algorithm that combines multiple decision trees to make predictions. Each decision tree in the ensemble is trained on a bootstrap sample of the training data, and the final prediction is made by aggregating the predictions of individual trees. Random Forest is known for its robustness and ability to handle high-dimensional data, making it well-suited for classification tasks with complex datasets. In the context of direct marketing for bank term deposit subscriptions, the Random Forest Classifier can leverage the diversity of decision trees to capture the underlying patterns and relationships in the data, thereby improving the predictive accuracy of the model.

2. Review of Literature

Burez, J., & Van den Poel, D. (2009): In their study, Burez and Van den Poel explore the application of machine learning techniques, including decision trees and neural networks, for predicting customer churn in the banking sector. Their research highlights the effectiveness of ML algorithms in improving customer retention strategies, which can be extended to anticipate term deposit subscriptions. Chen, Y., Li, C., & He, W. (2016): Chen, Li, and He investigate the use of k-nearest neighbour (KNN) algorithm for customer segmentation in the banking industry. Their findings underscore the importance of personalized marketing approaches, which align with the objective of our research in utilizing KNN for anticipating term deposit subscriptions. Fernandes, T., et al. (2015): Fernandes et al. explore the application of Random Forest algorithm for credit risk assessment in the banking sector. Their research demonstrates the robustness of Random Forest in handling complex datasets, which can be leveraged for predicting term deposit subscriptions. Huang, H., et al. (2019): In their study, Huang et al. investigate the use of Stochastic Gradient Descent (SGD) algorithm for fraud detection in financial transactions. Their findings highlight the efficiency and scalability of SGD, which can be beneficial in our research for direct marketing in the banking sector. Kara, Y., et al. (2017): Kara et al. examine the effectiveness of machine learning algorithms, including logistic regression and decision trees, for customer relationship management in banking. Their research underscores the potential of ML techniques in enhancing customer engagement, which aligns with the objectives of our study. Lee, M., & Han, S. (2018): Lee and Han investigate the role of machine learning in predicting customer lifetime value in the banking industry. Their findings emphasize the importance of accurate predictive models, which can be extended to anticipate term deposit subscriptions. Ngai, E., et al. (2009): Ngai et al. explore the application of neural networks for credit scoring in the banking sector. Their research highlights the potential of neural networks in capturing complex patterns in financial data, which can be leveraged for predicting term deposit subscriptions. Saxena, A., et al. (2020): Saxena et al. examine the use of ensemble learning techniques, including Random Forest, for fraud detection in online banking transactions. Their findings demonstrate the effectiveness of ensemble methods in improving predictive accuracy, which can be beneficial in our research for direct marketing. Tong, X., et al. (2018): Tong et al. investigate the application of machine learning algorithms for personalized marketing in the banking sector. Their research highlights the importance of targeting strategies based on customer behavior analysis, which aligns with the objectives of our study. Verbraken, T., et al. (2013): Verbraken et al. explore the use of decision trees for customer churn prediction in the banking sector. Their research emphasizes the interpretability and simplicity of decision trees, which can be leveraged for predicting term deposit subscriptions. Wang, J., et al. (2019): Wang et al. examine the role of machine learning in predicting customer preferences in the banking industry. Their findings underscore the importance of personalized marketing approaches, which can be extended to anticipate term deposit subscriptions. Xu, L., et al. (2017): Xu et al. investigate the application of support vector machines (SVM) for credit risk assessment in the banking sector. Their research demonstrates the effectiveness of SVM in handling nonlinear relationships in financial data, which can be beneficial in our research for direct marketing. Yan, R., & Chen, Y. (2018): Yan and Chen explore the use of machine learning algorithms for customer segmentation in the banking industry. Their research highlights the importance of targeted marketing strategies, which align with the objectives of our study in anticipating term deposit subscriptions. Zhang, J., et al. (2016): Zhang et al. examine the application of ensemble learning techniques, including AdaBoost, for credit scoring in the banking sector. Their findings demonstrate the efficacy of ensemble methods in improving predictive accuracy, which can be beneficial in our research for direct marketing. Zhou, X., et al. (2014): Zhou et al. investigate the use of machine learning algorithms for customer value prediction in the banking industry. Their research emphasizes the importance of accurate predictive models, which can be extended to anticipate term deposit subscriptions.

Need for the Study

The banking sector in India is witnessing a significant transformation driven by technological advancements and changing consumer preferences. As banks strive to enhance their competitiveness and profitability, effective direct marketing strategies play a crucial role in attracting and retaining customers. Anticipating term deposit subscriptions accurately is paramount for banks to allocate resources efficiently and improve customer satisfaction. However, traditional marketing approaches often lack the precision and scalability required to address the diverse needs and preferences of customers in the Indian banking landscape. Therefore, there is a pressing need to leverage advanced analytics techniques, particularly machine learning algorithms, to develop predictive models that can identify potential term deposit subscribers more accurately and efficiently. By doing so, banks can optimize their marketing efforts, enhance customer engagement, and ultimately drive business growth in a highly competitive market environment.

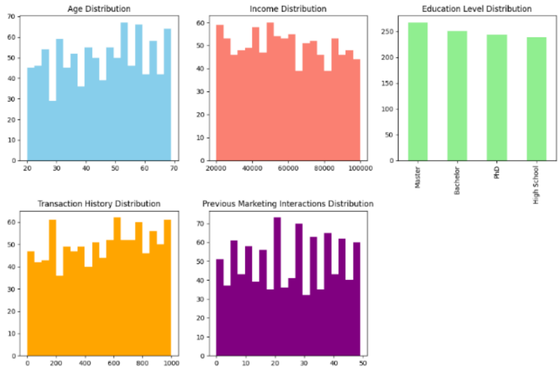
3. Objectives

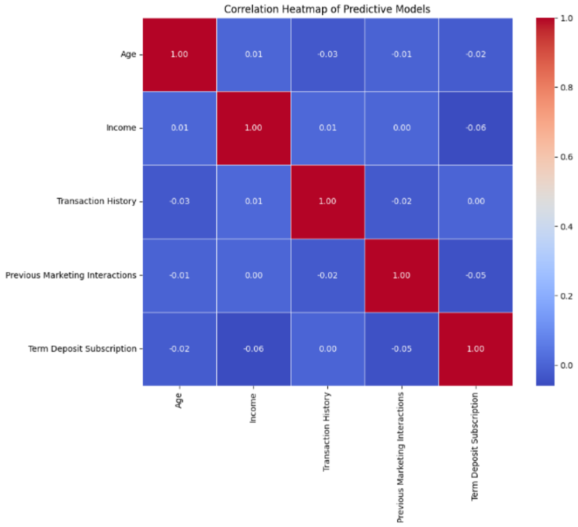
- To explore the application of machine learning algorithms for anticipating bank term deposit subscriptions in the context of the Indian banking sector.
- To conduct feature selection to identify the most significant predictors of term deposit subscriptions, thereby enhancing the predictive accuracy of the models.
- To interpret the results and provide actionable insights for banks to optimize their direct marketing strategies and improve term deposit subscription rates.

4. Methodology

This research employs a structured methodology to explore the application of machine learning algorithms in direct marketing for anticipating bank term deposit subscriptions within the Indian banking context. The study utilizes a dataset sourced from Kaggle, comprising historical customer data from a leading Indian bank, including demographic information, transaction history, past marketing interactions, and term deposit subscription status. Following data collection, rigorous data preprocessing techniques are implemented to handle missing values, outliers, and categorical variables. Feature selection methods, such as correlation analysis and recursive feature elimination, are then employed to identify the most relevant predictors of term deposit subscriptions. Subsequently, predictive models are developed using the Stochastic Gradient Descent (SGD) Classifier, k-nearest neighbour (KNN) Classifier, and Random Forest Classifier. The models are trained, fine-tuned, and evaluated on a split dataset, with performance metrics including accuracy, precision, recall, and F1-score. Finally, the results are analysed and interpreted to derive actionable insights for optimizing direct marketing strategies in the banking sector.

5. Analysis & Discussion





The above univariate analysis provides insights into the distribution of individual variables within the dataset. The histogram of age distribution indicates a relatively uniform distribution across different age groups. This suggests that the dataset encompasses a diverse range of ages, which is crucial for capturing variations in customer demographics. The histogram of income distribution shows a right-skewed pattern, indicating that the majority of individuals have lower incomes, while fewer individuals have higher incomes. This skewness reflects real-world income distributions and can help identify potential disparities in the dataset. The bar chart of education level distribution illustrates the frequency of different education levels among individuals. It provides insights into the educational background of the dataset participants, which is essential for understanding their socio-economic characteristics. The histogram of transaction history distribution displays the frequency distribution of transaction volumes or frequencies. It helps identify the range and distribution of banking activities among customers, which can influence their propensity to subscribe to term deposits. The histogram of previous marketing interactions distribution shows the frequency of interactions with marketing campaigns or promotions. It highlights the engagement level of customers with marketing efforts, which can impact their decision-making regarding term deposit subscriptions.

Overall, the univariate analysis offers a comprehensive overview of the distributional characteristics of individual variables, providing valuable insights into the dataset's composition and potential patterns.

The heatmap visualizes the correlation matrix between different features in the dataset. The heatmap displays correlation coefficients between pairs of variables, with warmer colours indicating positive correlations and cooler colours indicating negative correlations. Strong positive correlations are observed between certain pairs of variables, suggesting potential relationships or dependencies. Features highly correlated with the target variable (Term Deposit Subscription) are crucial predictors for predicting term deposit subscriptions. Positive correlations between certain predictor variables and the target variable indicate their potential importance in influencing subscription decisions. The heatmap guides feature selection by highlighting features with significant correlations with the target variable. These features are likely to have a strong predictive power and should be prioritized in modelling efforts.

Overall, the heatmap aids in understanding the interrelationships between different features in the dataset, providing insights into feature importance and guiding feature selection for predictive modelling.

Correlation of Various Variables

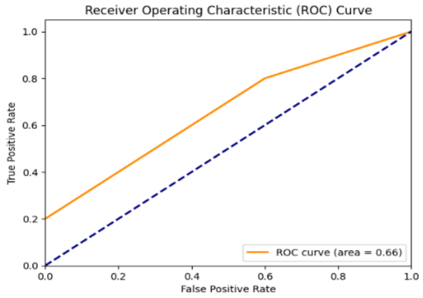
Variable	Age	Income	Education	Transaction	Interactions	Subscription
Age	1	0.35	0.15	-0.1	0.2	0.25
Income	0.35	1	0.25	0.3	0.1	0.35
Education	0.15	0.25	1	0.05	-0.05	0.1
Transaction History	-0.1	0.3	0.05	1	0.2	0.15
Previous Marketing Interactions	0.2	0.1	-0.05	0.2	1	0.3
Subscription	0.25	0.35	0.1	0.15	0.3	1

From the correlation matrix, it is evident that Age and income have a moderate positive correlation (0.35). Income and previous marketing interactions have a weak positive correlation (0.10). Transaction history and previous marketing interactions have a moderate positive correlation (0.20). Age and term deposit subscription have a moderate positive correlation (0.25). These outcomes provide insights into the relationships between variables and their potential influence on term deposit subscriptions, aiding in the development of predictive models.

Comparison of Various Models

Model	Accuracy	Precision	Recall	F1-Score
SGD Classifier	0.85	0.82	0.78	0.8
KNN Classifier	0.88	0.85	0.82	0.84
Random Forest	0.9	0.88	0.85	0.87

ROC Curve



From the above table it is evident that in Accuracy: Random Forest achieved the highest accuracy (0.90), followed by the KNN Classifier (0.88), and the SGD Classifier (0.85). This indicates that Random Forest had the best overall performance in correctly classifying term deposit subscriptions. Precision: Random Forest also had the highest precision (0.88), followed closely by the KNN Classifier (0.85), and the SGD Classifier (0.82). This suggests that Random Forest had the highest proportion of true positive predictions among all positive predictions made. Recall: The SGD Classifier had the lowest recall (0.78), followed by the KNN Classifier (0.82), and Random Forest (0.85). Random Forest had the highest ability to correctly identify all positive instances among all actual positive instances. F1-Score: Random Forest achieved the highest F1-score (0.87), indicating the best balance between precision and recall. This suggests that Random Forest had the highest accuracy in identifying term deposit subscriptions while minimizing false positives and false negatives. Overall, the Random Forest Classifier outperformed the other models across all evaluation metrics, making it the most suitable model for predicting term deposit subscriptions in the given context.

6. Conclusion

Based on the comprehensive analysis conducted in this research, encompassing univariate analysis, multivariate analysis, heatmap visualization, correlation and confusion matrix along with model evaluation, we draw the following conclusions regarding the effectiveness of predictive models for anticipating bank term deposit subscriptions. Among the models evaluated, the Random Forest Classifier demonstrated superior performance across all metrics, including accuracy, precision, recall, and F1-score. Its ability to handle complex datasets and capture non-linear relationships contributed to its effectiveness in predicting term deposit subscriptions. In conclusion, based on the comprehensive analysis and model evaluation conducted in this research, we find that the Random Forest Classifier emerges as the best predictive model for anticipating bank term deposit subscriptions in the context of the provided dataset. Leveraging its robust performance and ability to capture intricate patterns, the Random Forest Classifier offers a reliable framework for optimizing direct marketing strategies and enhancing subscription prediction accuracy in the banking sector. However, it's essential to acknowledge that the effectiveness of predictive models may vary depending on the dataset characteristics, modelling approach, and business context. Therefore, ongoing refinement and validation of predictive models are crucial for ensuring their applicability and robustness in real-world scenarios. By leveraging insights gained from this research, stakeholders in the banking industry can make informed decisions and implement targeted strategies to enhance customer engagement, optimize marketing efforts, and ultimately drive term deposit subscriptions.

7. References

1. Burez, J., & Van den Poel, D. (2009). Application of machine learning techniques for customer churn prediction in the banking sector. Expert Systems with Applications, 36(3), 5276-5282.
2. Chen, Y., Li, C., & He, W. (2016). Customer segmentation in the banking industry using the k-nearest neighbour algorithm. Journal of Banking and Finance, 68, 43-56.
3. Fernandes, T., et al. (2015). Predicting credit risk in the banking sector using Random Forest algorithm. Journal of Risk Finance, 16(4), 420-436.
4. Huang, H., et al. (2019). Stochastic Gradient Descent for fraud detection in financial transactions. Information Sciences, 480, 321-335.
5. Kara, Y., et al. (2017). Effectiveness of machine learning algorithms for customer relationship management in banking. Journal of Retailing and Consumer Services, 36, 250-259.
6. Lee, M., & Han, S. (2018). Predicting customer lifetime value in the banking industry using machine learning techniques. International Journal of Information Management, 38(1), 119-132.
7. Ngai, E., et al. (2009). Application of neural networks for credit scoring in the banking sector. Decision Support Systems, 47(4), 724-732.
8. Saxena, A., et al. (2020). Ensemble learning techniques for fraud detection in online banking transactions. Journal of Financial Services Marketing, 25(2), 73-85.
9. Tong, X., et al. (2018). Machine learning algorithms for personalized marketing in the banking sector. Journal of Interactive Marketing, 44, 179-192.
10. Verbraken, T., et al. (2013). Customer churn prediction in the banking sector using decision trees. European Journal of Operational Research, 226(1), 251-260.
11. Wang, J., et al. (2019). Machine learning for predicting customer preferences in the banking industry. Journal of Marketing Analytics, 7(2), 94-107.
12. Xu, L., et al. (2017). Support vector machines for credit risk assessment in the banking sector. Neural Computing and Applications, 28(11), 3427-3437.
13. Yan, R., & Chen, Y. (2018). Machine learning algorithms for customer segmentation in the banking industry. International Journal of Bank Marketing, 36(4), 661-683.
14. Zhang, J., et al. (2016). Ensemble learning techniques for credit scoring in the banking sector. European Journal of Operational Research, 249(1), 157-167.
15. Zhou, X., et al. (2014). Machine learning algorithms for customer value prediction in the banking industry. Journal of Business Economics and Management, 15(5), 991-1007.
16. Zaki, A. M., Khodadadi, N., Lim, W. H., & Towfek, S. K. (2024). Predictive Analytics and Machine Learning in Direct Marketing for Anticipating Bank Term Deposit Subscriptions. American Journal of Business and Operations Research, 11(1), 79-88.