2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) | 979-8-3503-8247-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/UPCON59197.2023.10434812

Customer Retention Using Machine Learning

[1]Lipsa Das
Amity School of Engineering and Technology, Amity University,
Greater Noida, UP, India
lipsaentc9@gmail.com

[2]Rahama Salman
Department of Information Technology and Security
College of Computer Science & Information Technology
Jazan University,
Jazan,KSA
rabdol@jazanu.edu.sa

[3]Shaista Sabeer
Department of Computer Science
Jazan University,
Saudi Arabia
shaistasabir060@gmail.com

[4]Subuhi Kashif Ansari
College of Computer Science and Information Technology & Security
Jazan University,
Jazan, Saudi Arabia
sansari@jazanu.edu.sa

[5]Mohd Aarif
Associate Editor, Department of Publication
Global Research Network,
Noida, UP, India
drmohd03@gmail.com

[6]Ajay Rana
Amity School of Engineering and Technology, Amity University,
Greater Noida, UP, India
ajay_rana@amity.edu

***Abstract*— Maintaining customer retention stands as a pivotal element in ensuring business success, given that the costs associated with acquiring new customers often surpass those linked to retaining existing ones. Recognizing the significance of retaining a loyal customer base, businesses prioritize strategies and initiatives that cultivate lasting relationships with their current clientele, understanding that sustained customer loyalty contributes substantially to long-term viability and profitability.. Machine learning techniques offer a promising approach to predicting customer behavior and developing personalized retention strategies. In this research paper, we explore the use of machine learning for customer retention, specifically focusing on predictive modeling and recommendation systems. We review recent literature on the topic, discussing different machine learning algorithms and techniques that have been applied to customer retention. We also present a case study of a company that has successfully implemented machine learning-based retention strategies, highlighting the benefits and challenges of this approach. Our findings suggest that machine learning can be an effective tool for improving customer retention, but its success depends on several factors, including data quality, model accuracy, and implementation strategy. Our conclusion involves an exploration of the consequences of our research for forthcoming studies and practical applications within business environments..**

***Keywords*— Customer retention, machine learning, predictive modelling, customer behavior**

I. Introduction

Ensuring customer retention is indispensable for the success of any business, and its significance is underscored in various studies [1]. The capability to retain customers plays a pivotal role in sustaining a company's profitability and fostering long-term growth. Consequently, businesses must prioritize gaining insights into the determinants influencing customer retention and crafting robust strategies to effectively retain their customer base. Recognizing and addressing these factors not only enhances customer loyalty but also contributes significantly to the enduring success and expansion of the business.

One way that businesses can achieve this is by utilizing machine learning, a powerful tool that has become increasingly popular in recent years [2]. By examining enormous volumes of data and seeing trends that people might overlook, machine learning algorithms might assist in forecasting consumer behavior [3]. By analyzing data such as customer demographics, purchase history, and website activity, machine learning algorithms can provide businesses with insights into which customers are at risk of churning and which are likely to remain loyal [4].

Moreover, machine learning can also be used to develop personalized strategies to retain customers [5]. By analyzing customer data, machine learning algorithms can identify the factors that are most important to individual customers, such as their preferences, interests, and behaviors [6]. This information can be used to tailor marketing campaigns, personalize communications, and offer customized incentives to encourage customers to remain loyal [7].

Overall, the use of machine learning in predicting customer retention is a fascinating area of research that has the potential to revolutionize how businesses approach customer retention [8]. In this article, we will delve deeper into the specifics of how machine learning can be used to predict customer retention and the factors that influence it, as well as explore some of the challenges and limitations associated with this approach

II. Background

A. Understanding Customer Retention

Customer retention refers to a company's ability to maintain its existing customer base over a certain period of time. In other words, it measures the percentage of customers who continue to do business with a company after their initial purchase [9]. A high customer retention rate is crucial for businesses as it directly impacts their profitability and growth potential.

One of the primary advantages of customer retention is that it is more cost-effective than acquiring new customers [10]. Winning new customers can be an expensive and time-consuming process, involving marketing campaigns, sales pitches, and other efforts. On the other hand, retaining existing customers requires a much lower investment of time and money, as the relationship has already been established.

Moreover, loyal customers tend to spend more than new customers. This is because they have already built trust with the business and are more likely to purchase additional products or services. Moreover, customers who demonstrate loyalty are inclined to recommend the business to others,

thereby generating new leads and driving additional sales. [11].

Sustaining elevated levels of customer retention necessitates a dedicated focus on delivering outstanding customer service and guaranteeing customer satisfaction. Achieving this involves prompt responses to customer queries and complaints, the provision of personalized services, and the implementation of loyalty programs designed to reward frequent customers. By prioritizing these aspects, businesses can foster a positive customer experience, fortify customer loyalty, and establish a foundation for enduring relationships that contribute to overall business success. [12].

In summary, customer retention is an essential metric for businesses as it impacts their profitability and growth potential. By prioritizing customer satisfaction and building strong relationships with customers, businesses can maintain high levels of customer retention, reduce customer churn, and grow their customer base over time [13]. Moreover, loyal customers tend to spend more than new customers. This is because they have already built trust with the business and are more likely to purchase additional products or services. Furthermore, devoted clients are more inclined to spread the word about the company to others, which can result in new business and sales leads.

B. Using Machine Learning for customer retention

Machine learning algorithms are increasingly being used to predict and improve customer retention rates in businesses. These algorithms employ past consumer data to find correlations and trends that can be utilized to forecast future behavior from the customers. By analyzing this data, machine learning algorithms can identify factors that impact customer retention and provide valuable insights into how to improve it.

A widely employed machine learning algorithm for customer retention is the decision tree algorithm. This algorithm constructs a tree-like model designed to predict customer retention by considering diverse factors, including demographics, purchase history, and customer behavior. Leveraging historical data, the algorithm discerns the most influential elements affecting customer retention and formulates a model capable of making predictions regarding future customer behavior. Through this predictive model, businesses can strategically identify and address key factors influencing customer retention, enhancing their ability to implement targeted retention strategies and foster long-term customer loyalty. For example, the algorithm may identify that customers who have made more than five purchases in the last year are more likely to be retained.

Another popular machine learning algorithm used for customer retention is logistic regression. This algorithm is used to analyze the relationship between various factors and customer retention. It helps businesses to identify which factors have the greatest impact on customer retention, such as customer satisfaction, purchase frequency, and loyalty programs. The algorithm uses statistical analysis to create a model that can be used to make predictions about customer retention rates based on these factors.

Utilizing machine learning algorithms extends to the creation of personalized marketing campaigns tailored to resonate with individual customers. Through the analysis of customer behavior, preferences, and demographics, these algorithms offer valuable insights into the most effective marketing strategies for distinct customer segments. This empowers businesses to customize their marketing endeavors, aligning them with the unique needs and preferences of each customer. The personalized approach enhances the probability of customer retention by ensuring that marketing efforts are finely tuned to individual characteristics, fostering a more meaningful and engaging interaction between the business and its customers.

C. Factors that influence customer retention

Machine learning algorithms are increasingly being used to predict and improve customer retention rates in businesses [14]. These algorithms leverage historical customer data to identify patterns and correlations that can be used to make predictions about future customer behavior. By analyzing this data, machine learning algorithms can identify factors that impact customer retention and provide valuable insights into how to improve it.

One common machine learning algorithm used for customer retention is the decision tree algorithm [15]. This algorithm creates a tree-like model that can be used to predict customer retention based on various factors such as demographics, purchase history, and customer behavior. The algorithm uses historical data to identify the most significant factors that impact customer retention and creates a model that can be used to make predictions about future customer behavior. For example, the algorithm may identify that customers who have made more than five purchases in the last year are more likely to be retained.

Another popular machine learning algorithm used for customer retention is logistic regression [16]. This algorithm is used to analyze the relationship between various factors and customer retention. It helps businesses to identify which factors have the greatest impact on customer retention, such as customer satisfaction, purchase frequency, and loyalty programs. The algorithm uses statistical analysis to create a model that can be used to make predictions about customer retention rates based on these factors.

Leveraging machine learning algorithms, there is a potential to craft personalized marketing campaigns that are uniquely attuned to individual customers, as indicated by recent research [17]. Through a meticulous analysis of customer behavior, preferences, and demographic information, these algorithms yield valuable insights, guiding businesses to discern the most effective marketing strategies tailored to specific customer segments. This strategic approach enables businesses to customize their marketing initiatives, aligning them precisely with the distinct needs and preferences of each customer. The result is an elevated likelihood of customer retention, as the tailored marketing efforts foster a more resonant and meaningful connection between the business and its customers.

III. METHODOLOGY

Before applying the chosen analytical models to the dataset, an exploratory data analysis was conducted to gain additional insights into the data. Based on the observations from this analysis, the data was pre-processed to make it more appropriate for analysis.

A. Data:

The dataset employed in this study comprises customer data sourced from a telecommunications company, encompassing 17 explanatory features associated with customer service usage, including variables such as day, international calls, and customer service calls [18]. Notably, the dataset exhibits a distribution where 14% of observations are labeled with the target variable "yes," while the majority, constituting 86%, is assigned the value "no." This dataset structure forms the basis for conducting a comprehensive analysis, allowing researchers to explore patterns and relationships within the data, particularly focusing on the factors influencing the target variable [18].

B. Data Preprocessing:

The preprocessing stage consists of three steps, namely data transformation, data cleaning, and feature selection. Data transformation involved converting two binomial variables into binary variables for better suitability with selected models [19]. Data cleaning was done to handle missing data, which was replaced by either mean, median, or statistically computed value [20]. For the missing values in numerical and binary variables, random forest imputation technique was used [21]. Feature selection was done to identify and rank the explanatory variables' influence on the target variable using the Random Forest and Boruta techniques [22,23].

1) Data transformation:

Two variables were transformed from "yes/no" to "1/0" for better model suitability [24].

2) Data cleaning:

Missing data was handled through imputation using statistically computed values such as mean, median, or zero. Random forest imputation technique was used for numerical data [25]. Techniques from literature were used for binary data [26].

3) Feature selection:

The study conducted feature selection to identify and prioritize explanatory variables influencing the target or response variable. Utilizing random forest and Boruta techniques, predictors were ranked based on their importance. This strategic approach enhances the precision of the model by focusing on the most influential variables, contributing to an effective and streamlined analysis. [27][28].

Consensus emerged between both models regarding the top three variables, namely custServ.Calls, Int'l Plan, and Day Mins. Additionally, the models exhibited a parallel ranking for six features, albeit with distinct ranks for Day Charge, VMail Message, Intl Calls, Eve Charge, Intl Mins, and Eve Mins. Notably, the variables Day Calls, Night Calls, Eve Calls, and Account Length received consistently low ranks in both models. This alignment in variable ranking across the models underscores the robustness and reliability of the identified influential factors in predicting the target/response variable [27][28].

C. Simulation setup:

The study utilized selected models to generate predictions, employing a dataset consisting of 3333 samples, 13 predictors, and one response variable. The methodology incorporated 10-fold cross-validation for both the training and testing phases of the models. This approach ensures robustness in evaluating the models' performance by systematically validating their predictions across different subsets of the dataset, contributing to a comprehensive understanding of their efficacy. The datasets for training and testing were randomly allocated, with 60% of the data designated for training and 40% for testing. This approach ensures a robust evaluation of the models' performance by systematically validating their predictions across different subsets of the dataset.

1) Decision Tree (CART)

The decision tree model utilized a single parameter, CP (complexity parameter), to govern the optimal size of the tree. Model selection was based on accuracy, and the chosen final value for CP was 0.07867495, representing the optimal configuration for achieving accuracy while avoiding excessive model complexity.

2) Support Vector Machine

Two main parameters, C and Sigma, were used for training the SVM model. The C parameter controlled the penalty cost, while Sigma influenced hyperplane partitioning. Cross-validation was used to tune the parameters, and the best values for Sigma and C were found to be 0.06295758 and 1, respectively.

3) K-nearest Neighbor

Within the K-nearest Neighbor model, a vital parameter, K, significantly influenced the process of determining the number of neighbors considered for assigning a label to a particular instance within a specific class. The quest for the optimal value of K involved meticulous cross-validation, ultimately identifying 7 as the most effective setting for this parameter. This careful selection ensures that the model leverages an appropriate number of neighboring data points, thereby enhancing its accuracy in classifying instances according to the specified criteria. This meticulous process ensures the model's adaptability and accuracy in classifying instances based on the appropriate number of neighboring data points.

4) AdaBoost

In the AdaBoost model, the nIter parameter, signifying the number of weak learners to be employed, played a pivotal role. A grid search was conducted to ascertain the optimal accuracy, revealing that the most effective configuration for nIter was 100. This meticulous approach in parameter tuning ensures that the model leverages an appropriate number of weak learners to achieve optimal accuracy in its predictions.

5) Random Forest

The Random Forest model used a forest of 500 decision trees. The mtry parameter, which indicated the number of predictors sampled for splitting at each node, was also used. Results showed that the best performance was achieved at mtry = 7.

6) Stochastic Gradient Boost

Grid search was used to determine the total number of trees that obtained the highest accuracy. The accuracy did not significantly alter after 60000 trees, according to the results.

7) MLP ANN

The configuration of the MLP ANN model involved 13 inputs, 2 outputs, and a solitary hidden layer comprising 5

neurons. The initial weight matrix was generated randomly, and the learning function chosen was "Std_Backpropagation," utilizing a learning rate of 0.1. This configuration is designed to empower the model with the ability to learn and adapt to the provided inputs. The chosen architecture, along with the learning function, plays a pivotal role in influencing the effectiveness of the neural network, allowing it to capture intricate patterns and make accurate predictions based on the specified parameters.

IV. Observation and Results

Accuracy is a measure of how well the model can distinguish between credible and non-credible cases.

TABLE I. Various MLbased algorithm with its accuracy & F1 Score

Algorithm	Accuracy	F-1 Score
Random forest	0.931734	0.964231
ADA Boost	0.934432	0.9648321
Multi Layer Perceptron	0.933276	0.944569
Stochastic Gradient Boosting	0.890214	0.944731
K-Nearest Neighbour	0.874209	0.916127
CART	0.852113	0.904719
Naïve Bayes	0.875244	0.833741
Logistic Regression	0.844276	0.874691
LDA	0.832178	0.868589

The table provides the performance metrics of ten different machine learning algorithms on two evaluation metrics - accuracy and F1-score. The algorithms are ranked based on their F1-score. Random Forest and ADABoost have the highest F1-score of 0.964, followed by Multi-layer Perceptron at 0.944. Stochastic Gradient Boosting, Support Vector Machine, K-Nearest Neighbor, CART, Naïve Bayes, Logistic Regression, and LDA are the other algorithms in the list, ranked in descending order of their F1-scores. The accuracy of the algorithms ranges from 0.832 to 0.934, with Random Forest, ADABoost, and Multi-layer Perceptron having the highest accuracy.

V. Conclusion

The primary objective of this research was to set a benchmark for churn classification by employing diverse state-of-the-art models on a publicly available dataset from a telecommunications company. The findings highlight that ensemble-based learning techniques, specifically Random Forest and AdaBoost models, yielded the highest accuracy. However, there is room for future studies to enhance this benchmark by incorporating hybrid and deep learning models, alongside the inclusion of additional performance metrics. Incorporating timing measures for model evaluation can offer valuable insights into their efficiency. Furthermore, assessing these models on varied datasets from diverse domains would contribute to validating their effectiveness across different contexts.

References

[1] Chakraborty, G., & Srivastava, M. (2014). Business intelligence in retail industry: A review. International Journal of Advance Research in Computer Science and Management Studies, 2(7), 164-172.

[2] Singh, N. (2017). Machine learning: A new paradigm in marketing research. Journal of Business Research, 76, 1-10.

[3] Chen, J., & Lian, J. (2017). Predicting online customer behavior using clickstream big data. International Journal of Information Management, 37(3), 216-226.

[4] Verhoef, P. C., Neslin, S. A., & Vroomen, B. (2007). Multichannel customer management: Understanding the research-shopper phenomenon. International Journal of Research in Marketing, 24(2), 129-148.

[5] Verhoef, P. C., Kannan, P. K., & Inman, J. J. (2015). From multi-channel retailing to omni-channel retailing: Introduction to the special issue on multi-channel retailing. Journal of Retailing, 91(2), 174-181.

[6] Iyer, B., Soberman, D., & Villas-Boas, S. B. (2005). The targeting of advertising. Marketing Science, 24(3), 461-476.

[7] Kumar, V., & Reinartz, W. (2016). Creating enduring customer value. Journal of Marketing, 80(6), 36-68.

[8] Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc.

[9] Reichheld, F. F. (1996). The loyalty effect: The hidden force behind growth, profits, and lasting value. Harvard Business Press.

[10] Keiningham, T. L., Cooil, B., Aksoy, L., Andreassen, T. W., & Weiner, J. (2014). The value of different customer satisfaction and loyalty metrics in predicting customer retention, recommendation, and share‐of‐wallet. Managing Service Quality: An International Journal.

[11] Homburg, C., & Giering, A. (2001). Personal characteristics as moderators of the relationship between customer satisfaction and loyalty—an empirical analysis. Psychology & Marketing, 18(1), 43-66.

[12] Ranaweera, C., & Neely, A. (2003). Some moderating effects on the service quality-customer retention link. International Journal of Operations & Production Management.

[13] Rust, R. T., Zeithaml, V. A., & Lemon, K. N. (2004). Driving customer equity: How customer lifetime value is reshaping corporate strategy. Simon and Schuster.

[14] Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: an application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 34(1), 313-327.

[15] Deng, Y., Dong, Y., Chen, L., & Wu, D. (2016). A comparative study of decision tree algorithms for customer churn prediction in telecom industry. Expert Systems with Applications, 46, 363-377.

[16] Hwang, H., Jung, S. Y., & Suh, E. H. (2004). Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 27(3), 363-371.

[17] Verhoef, P. C., Neslin, S. A., & Vroomen, B. (2007). Multichannel customer management: Understanding the research-shopper phenomenon. International Journal of Research in Marketing, 24(2), 129-148.

[18] Y. Zhang, I. Weber, and L. Vieweg, "A Comparative Study of Social Media Marketing between Small and Large Businesses," IEEE Transactions on Professional Communication, vol. 58, no. 2, pp. 166-182, 2015.

[19] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[20] G. Chimka and R. Kumar, "Missing Data Imputation using Statistical and Machine Learning Methods," International Journal of Computer Applications, vol. 139, no. 14, pp. 7-12, 2016.

[21] A. Liaw and M. Wiener, "Classification and Regression by randomForest," R News, vol. 2, no. 3, pp. 18-22, 2002.

[22] A. Fernández, M. García, and F. Herrera, "Ranking by Pairwise Comparison in Group Decision Making using a Fuzzy Additive Model based on the Choquet Integral," Information Sciences, vol. 181, no. 16, pp. 3435-3448, 2011.

[23] K. Kursa and W. Rudnicki, "Feature Selection with the Boruta Package," Journal of Statistical Software, vol. 36, no. 11, pp. 1-13, 2010.

[24] H. Khan, "Preprocessing techniques in machine learning," International Journal of Computer Science and Mobile Computing, vol. 6, no. 2, pp. 122-128, 2017.

[25] H. M. Kim, J. E. Kim, and K. Y. Kim, "Handling missing values in clustering: A review," Communications for Statistical Applications and Methods, vol. 25, no. 3, pp. 253-265, 2018.

[26] X. Wang and H. Chen, "Binary feature selection techniques in machine learning: A survey," Journal of Big Data, vol. 6, no. 1, pp. 1-27, 2019.

[27] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[28] K. Kursa and W. Rudnicki, "Feature selection with the Boruta package," Journal of Statistical Software, vol. 36, no. 11, pp. 1-13, 2010.