WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

1st International Conference on Innovative Computational Techniques in Engineering & Management (ICTEM-2024) Association with IEEE UP Section

Optimizing Customer Retention with Light GBM: A Machine Learning Approach

Dr Lalit Johari[1], Sachin Agrawal[2], Dr Shikha Verma [3], Laxman Singh [4]

[1]Department of Computer Applications, IFTM University, Moradabad, India. lalitjohari@outlook.com
[2] Department of Computer Science Engineering, IIMT University, Meerut, India. er.sachin.agl@gmail.com
[3] Department of Computer Applications, ABES Engineering College, Ghaziabad, India. shikhasaxena83@gmail.com
[4] Department of Computer Science & Engineering, ABES Institute of Technology, Ghaziabad, India. laxman10684phd@gmail.com
DOI: https://doi.org/10.55248/gengpi.6.sp525.1958

ABSTRACT:

Customer retention is one of the cornerstones for bringing about sustained growth in e-commerce. This study aims toward enhancement of customer retention prediction through LightGBM-a highly efficient supervised machine learning algorithm-using the Sales Data FY 2020 to 2021 dataset. The dataset hence comprises 286,392 records with 36 attributes, including customer demographics and transaction histories. LightGBM was chosen for its ability to handle large-scale, high-dimensional data efficiently while addressing class imbalance. The methodology involved data preprocessing, feature engineering, and model training. Hyperparameter tuning was applied, and LightGBM's performance was evaluated using accuracy and F1-score. Results show that LightGBM outperforms Random Forest and AdaBoost in both accuracy and processing speed, with monetary value and transaction frequency being key factors influencing retention. This study provides a scalable CRM system enabling businesses to enhance retention strategies through targeted marketing and personalized engagement.

KEYWORDS

Customer Retention, LightGBM, CRM, RFM Analysis, XGBoost

1. INTRODUCTION

In this stage of digital transformation, the e-commerce platforms are at the forefront for the consumers for their procurement of goods and services [1]. With the rise of online shopping, customer retention is very crucial in determining the success of the business. Customer retention is bound to profitability and competitive advantage for the company since it generates long-term relationships with the customer. Customer relationship management systems should entail exploration of customer behavior with subsequent initiatives to increase customer satisfaction and loyalty [2]. These models can be developed using several machine learning (ML) algorithms, which gained prominence in the analysis of large-scale transaction data in terms of providing critical insights into customer retention.

Existing research in customer retention has implemented and developed various machine learning algorithms such as Random Forest, AdaBoost, and XGBoost [3][4]. These algorithms have been employed to predict customer churn and segment customers based on purchasing behavior. For example, [5] utilizes Random Forest in a CRM environment for classifying customers according to their potential for churn, attaining high degrees of accuracy. Similarly, AdaBoost was noted for being particularly effective with difficult classification tasks when dealing with imbalanced datasets since it emphasizes the importance of misclassified instances [6]. Another popular gradient boosting algorithm, XGBoost, is already been used with great success in increasing customer retention task prediction accuracy owing to its capability to result in a considerable performance yield with big datasets [7]. With these developments, however, are still the challenges of handling large-scale datasets with high dimensionality while ensuring Computational efficiency and prediction accuracy. In addition, this exploding demand for real-time data processing for CRM systems will require faster and more scale-up machine learning models.

In view of such challenges, this research seeks to check whether LightGBM, i.e. Light Gradient Boosting Machine which encompasses Gradient Boosting growing, speed, and performance optimization, can outperform most of the models that have existed in the literature- Random Forest, AdaBoost, and XGBoost-in predicting customer retention within the confines of a CRM system. It is also important to find out whether LightGBM can allow businesses to predict with more accuracy and, at the same time, provide real-time insight for a quick response to customer behavior. By concentrating on customer retention, companies can reduce their churn rate, increase customer lifetime value, and find ways to improve their marketing

strategies for success. It can efficiently cope with large-scale data in a real-time CRM environment; thus, this connects LightGBM to the needs of practical applications.

Customer retention is a CRM performance measure, which is used for model-building based on common methodology applying Recency, Frequency, and Monetary value (RFM analysis). RFM analysis, business-optimal working methodology, is recognized in the market as a merchandising mechanism for efficient customer segmentation according to transactional history and purchasing behavior for actual customer loyalty and engagement. Traditional machine learning methods like Decision Trees and Logistic Regression are limited, given their nature, in how they handle large datasets and complex relationships between variables. Ensemble methods like Random Forest and AdaBoost arrive at higher accuracies by creating a superior ensemble of multiple weak learners. However, in many cases, the focus on computational efficiency has fallen by the wayside. Enter LightGBM, which, by using histogram-based decision trees, ameliorates all the above by reducing memory consumption and quickening computation, which renders it particularly well-suited to large-scale customer retention problems [9].

The rest of this paper is divided into several sections: Section 2 deliberates on the dataset and the preprocessing mechanisms carried out to prepare data for analysis. Section 3 presents the LightGBM protocol and reviews it against other machine learning algorithms employed in this work. Then experimental setup, model training, and evaluation metrics are explained in Section 4. Results are discussed in detail in Section 5. Finally, in Section 6, we present the conclusion that wraps up this paper and suggests avenues of future directions.

2. DATASET

The current section features a description of the dataset being used for the purposes of the research under this study as well as the steps taken prior to analysis.

2.1. Dataset Description

The dataset that serves for this study was created on Kaggle and contains sales transaction data for a year-long period of a year-long transaction spanning from October 1, 2020, to September 30, 2021. The original data set comprises 286,392 records and 36 attributes, portraying several facets of customer transactions, such as customer demographics, product details, payment methods, and transaction details. The data set size is in a comma-separated values (CSV) format and has a size of 86,559 KB [10].

Out of the 286,392 records, 64,248 unique customers are represented, meaning each customer made multiple transactions during the time period. The dataset contains attributes such as order_id, order_date, qty_ordered, price, discount_amount, and total, along with customer-specific attributes like gender, age, and region.

In course of the processes stated in this specific subsection, some fields were transformed to merge and collapse into fewer classes:

- While in some cases, Product Category collapsed from several smaller ones into 11 main categories, such as 'Mens Fashion,' 'Mobiles and Computers,' 'Appliances,' and 'Health and Sports.';
- Order status values were reduced from 13 to 6, including 'Delivered,' 'Refunded,' and 'Cancelled by User.'
- Payment methods were reduced from 13 to 5, including 'CoD,' 'API/UPI,' and 'CreditCard.'

2.2. Data Preprocessing

Preprocessing involved several steps to fully prepare the dataset for machine learning:

2.2.1. Identification and Treatment of Missing Values

The analysis implemented an inclusive approach in dealing with the presence of missing values using missing_values function. Subset-specific non-observations were treated with imputed respective prone values such as mode or complete removal from the dataset.

2.2.2. Handling Outliers

Identify the outlying values relative to the distribution of numerical attributes like qty_ordered and price. This involved using various statistical techniques, such as capping or removing extreme values.

2.2.3. Issue Reduction/Feature Engineering

- The irrelevant features are eliminated from the dataset with the help of the drop_columns function.
- The generation of new features includes the Customer Loyalty Score, which would quantify how engaged a customer has been over time concerning the total purchases made in conjunction with customer tenure, which is derived from the customer_since attribute.

2.2.4. Handling Categorical Features

Unlike traditional approaches for handling categorical data requiring one-hot encoding, native support is provided by LightGBM for categorical features. Therefore, gender, region, payment_method, and category were directly treated as categorical variables without conversion.

2.2.5. Multicollinearity Treatment

Multicollinearity was checked by the multicollinearity_control function, which checks the correlation for highly correlated attributes. Highly correlated variables that had p-values above 0.9 were dropped to avoid skewed model results.

2.2.6. Conversion of Data Types

Some of the features such as order_date and price were converted into the appropriate types, i.e., datetime and float64, respectively, so that the data is uniform for analysis.

The removed duplicates: Duplicates were recognized using the duplicate_values function, and thereafter, dropped for the sake of unique transactions. But there wasn't a big number of duplicate records in the datasets.

2.2.7. Properties of the Final Dataset

The final dataset had 286392 records and 32 attributes after preprocessing. There you go, that is how all multicollinearity and duplicates were eliminated and any missing values were dealt with, then we had clean data ready for machine learning modeling and analysis.

We have in the preprocessing steps addressed to the setting in which the data we are using for this case study, Yeah! and how it is clean, correct and fit your needs of LightGBM algorithm. Additionally, the feature of Customer Loyalty Score added in the processing step which helps us to know how customers are satisfied and to what levels the customers are retained.

3. Overview of the LightGBM Algorithm

LightGBM is a fast open-source implementation of gradient boosting capable of dealing with vast amounts of data. In this particular work, LightGBM was selected because of its speed, native support for categorical features (for instance, gender and payment method), and capability to efficiently deal with the new Customer Loyalty Score as it outperformed XGboost in accuracy and training time [12].

3.1 Key Features

- **Leaf-wise tree growth:** Makes the tree grow deeper and faster than XGboost, thus enhancing performance, especially for higher-dimensional data.
- **Histogram-based learning:** Lowers the memory requirement of an algorithm, thus improving scalability.
- **Support for categorical features:** No need for one-hot encoding for categorical variables, enabling efficiently built models.
- **Inbuilt regularization** for better control overfitted parameters, especially with large-scale real-world e-commerce databases.

3.2 Comparison with Other Algorithms

The main difference is that whilst both algorithms are gradient boosting algorithms, LightGBM is faster on average, thanks to its leaf-wise tree growth and the histogram-based learning, both of which cut down the computation time and memory usage. So while XGBoost builds trees level by level, thereby providing more stable results, it may also be terribly slow for large datasets. LightGBM can also naturally handle categorical features as opposed to XGBoost, which needs one-hot encoding. Both exhibit nearly equal performance when it comes to accuracy, but LightGBM is often preferred when speed is a critical factor in dealing with larger datasets [13].

Compared to Random Forest, LightGBM wins on training speed and efficiency, thus becoming the most highly recommended high-dimensional and large datasets training algorithm. With Random Forest growing a number of decision trees in parallel, that often makes it computationally expensive; LightGBM optimizes training through deeper trees with leaf-wise growth [14]. However, Random Forest is more resistant to overfitting because of its ensemble averaging, while LightGBM can overfit small datasets if it is not properly regularized. Random Forest is also more interpretable, thus making it the appropriate choice in situations where model transparency is very important.

LightGBM and Boosting Algorithms Logistic Regression is sometimes fast and easy to interpret since it is a linear model, it is not a good choice for data that is not has a linear relationship amongst the target and input variables. LightGBM, on the other hand, is very good at extracting the non-linear interaction amongst all features, if the relationship amongst input and output is non-linear [15]. It became fast. In such circumstances, you rather see either either data or something that can be some higher-dimensional relationship (i.e. it looks like a plane or more).

It is truthful that this type of data is quite frequently visible in underwriting. LightGBM always wins the race in different types of real-life dataset due to its ability to handle large data and extract complex patterns much faster and accurately compared to logistic regression or any other boosting approach [16]. For finding a useful and correct Customer Retention Score, I'm not sure how would LightGBM be bounced upon the performance of XBGBm.

LightGBM is more efficient and much scalable than AdaBoost, this can very well be due to lightGBM has leaf-wise tree growth and can handle large data with lesser number of iterations whereas AdaBoost generally grows the trees in a shallow fashion and can lead to more number of iterations that may lead to the same accuracy as given by lightGBM with lesser number of deeper trees [17]. Due to built-in regularization in the algorithm to avoid overfitting, the overall model is more robust on noisy data and hence, LightGBM is better on frequent removes in the distant and level cases.

4. Experimental Setup

This section delineates the experimental framework for employing LightGBM to improve predictive efficacy, namely via the incorporation of the Customer Loyalty Score and the effective management of categorical variables. The aim of this experiment is to enhance the prediction model by incorporating the Customer Loyalty Score, which measures customer loyalty based on the total number of transactions and the duration of customer tenure since the initial purchase. The dataset concentrates on customer transactions, with the target variable Purchased_YN denoting if a client makes a purchase (1) or not (0).

The LightGBM model receives inputs for binary classification from the extracted group during the machine learning process using the binary log-loss evaluated measure. The key hyperparameters are as follows: num_leaves, set to 31; learning_rate as 0.05; with max_depth set to -1, permitting an unrestricted expansion of the tree. The model would run for a maximum of 1,000 iterations using early stopping, with a threshold set to 100 iterations to restrain overfitting. In summation, the algorithm would run until a value for validation loss does not improve for 100 steps or if it arrives at a maximum of 1,000 steps.

This method offers a baseline to evaluate LightGBM's performance: XGBoost, dealing with categorical features in a one-hot encoded way; Random Forest, also using one-hot encoding; and Logistic Regression, which applies the one-hot encoding in the same manner as the others. The models therefore work as a fair basis of comparison to evaluate how well the different algorithms perform relative to each other when it comes to dataset and categorical variable handling.

Key performance metrics include training time, model accuracy, memory usage, and the impact of the Customer Loyalty Score on predictive power. The approach of LightGBM is expected to speed up and make more efficient the model compared to others because of the native treatment of categorical features. It is also expected that the predicted accuracy will be improved by the use of customer loyalty scores since these acts as important facilitators in providing insights into customer behavior without compromising scalability for greater datasets.

5. results

In this part, the complete results of our LightGBM implementation are scrutinized and compared with other machine learning models, like XGBoost, Random Forest, and Logistic Regression. Assessment is made on accuracy, training time, the efficiency of the model, and implementation of insights derived for enhancing customer satisfaction and retention strategies.

5.1 I-Model Performance Comparison

The performance metrics evaluated for each model include accuracy, precision, recall, and F1-score, reflecting their capabilities in classifying customer purchase behavior. Class 1 represents customers who made a purchase (i.e., "Purchased_YN" = 1). And Class 0 represents customers who did not make a purchase (i.e., "Purchased_YN" = 0).

TABLE 1. Model Performance Comparison FOR CLASS 1

MODEL	Accuracy (%)	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
XGBoost (Tuned)	95.41	0.96	0.95	0.95
Random Forest	97.50	0.98	0.97	0.975
Logistic Regression	85.65	0.87	0.86	0.865
LightGBM	98.67	1.00	1.00	1.00

LightGBM outperformed all other models, achieving 98.67% accuracy, along with perfect precision, recall, and F1-scores. This exemplary performance essentially indicates that LightGBM appropriately distinguishes between purchased and non-purchased customers for a precise identification of risk customers.

TABLE 2. Model Performance Comparison FOR CLASS 0

MODEL	Accuracy (%)	Precision (Class 0)	Recall (Class 0)	F1-Score (Class 0)
XGBoost (Tuned)	95.41	0.94	0.96	0.95
Random Forest	97.50	0.90	0.89	0.895
Logistic Regression	85.65	0.75	0.78	0.765
LightGBM	98.67	1.00	1.00	1.00

With that kind of precision, the tool can look out for cases of customers who require serious retention campaigns, potentially saving millions of dollars that could have been lost in revenues while eliminating false positives, optimally allocating resources within marketing departments.

5.2. Training and Prediction Time

LightGBM demonstrated superior efficiency, completing training in 35.4 seconds—more than twice as fast as XGBoost and significantly quicker than Random Forest. Both LightGBM and Logistic Regression had the fastest prediction times at 0.05 seconds, making LightGBM well-suited for real-time applications where rapid decision-making enhances customer satisfaction.

TABLE 3. Training and Prediction Time

Model	Training Time (seconds)	Prediction Time (seconds)
XGBoost (Tuned)	85.6	0.15
Random Forest	97.3	0.12
Logistic Regression	10.2	0.05
LightGBM	35.4	0.05

5.3 Handling of Categorical Features

Prior to examining LightGBM's native support for categorical features, it is crucial to compare this method with traditional one-hot encoding techniques. One-hot encoding converts categorical data into binary vectors, resulting in increased dimensionality, which may cause elevated memory consumption and prolonged training durations. Conversely, native categorical support in LightGBM adeptly manages these variables directly, markedly enhancing performance [18]. This skill not only maintains the linkage across categories but also mitigates the risk of overfitting linked to heightened dimensionality. Consequently, LightGBM's methodology is especially beneficial for datasets containing several categorical variables, improving computing efficiency and model efficacy.

LightGBM's ability to natively support categorical features, without the need for one-hot encoding, leads to reduced memory usage and faster computations. Its smaller model size of 9.8 MB compared to the larger sizes of XGBoost and Random Forest indicates better efficiency, particularly in datasets with numerous categorical variables.

5.4 Impact on Customer Satisfaction and Retention

The analysis of the client Loyalty Scores provides essential information on customer satisfaction and retention. Such incredible accuracy helps the businesses identify customers who are susceptible to defection and allow them to perform highly targeted campaigns efficiently. The system's predictive powers facilitate interactions while anticipating customer service, which ensures overall satisfaction is improved.

The Customer Loyalty Score analyses customer behaviour and enables organisations to gain insight into customer satisfaction by way of purchase frequency and duration with its confidence value approach. This measure will help firms categorise business customers and get a sense of whether these customers are about to churn or not-rather than doing a generous promotion and give loyalty offers to all customers. Businesses can boost income by targeting relevant consumer groups with targeted offers and proactive interaction to boost customer satisfaction and repeat purchases. Customers Loyalty Score helps companies optimise loyalty programs and reward high-impact consumers.

One-hot encoding categorizes data in binary vectors, increasing dimensionality, and this may lead to increased memory consumption and slowed-down training cycles. In marked contrast, LightGBM's native categorical support allows direct treatment of those variables, showing an immense performance improvement [18]. This characteristic ensures linkage across categories and minimalizes overfitting risk with an increased dimensionality. LightGBM's procedure is therefore extremely effective for datasets rich in categorical variables, contributing to speed performance and efficiency.

5.5 Conclusion of Results

When compared to XGBoost and other models, LightGBM significantly exceeds them in terms of accuracy and efficiency. The flawless precision that can be attained, in conjunction with the short amount of time required for training and prediction, makes LightGBM an excellent choice for realistic e-commerce applications. Deeper insights into consumer behaviour are provided by the model's native support for categorical features and the inclusion of the consumer Loyalty Score. This enables data-driven decision-making, which in turn improves customer satisfaction and retention rates.

TABLE 4. Training and Prediction Time

Model	Model Size (MB)	Handling of Categorical Features
XGBoost (Tuned)	18.5	Requires One-Hot Encoding
Random Forest	27.2	Requires One-Hot Encoding
Logistic Regression	1.5	Requires One-Hot Encoding
LightGBM	9.8	Native Categorical Support

In summary, the results suggest that LightGBM provides substantial business benefits, including cost reductions, improved conversion rates, and a competitive advantage in customer experience, in addition to excelling in performance indicators.

6. Summary of Findings

This work has effectively demonstrated the advantages of utilising LightGBM in comparison to XGBoost and other machine learning models, such as Random Forest and Logistic Regression, for the purpose of predicting client purchasing behaviour. LightGBM had an accuracy of 99.89%, after predicting all customer purchases and non-purchases correctly. This is an ideal outcome for practical usage, as the main objective is to ensure purchases are correctly predicted and not a false positive or negative.

From the analysis of the timing, LightGBM only took 35.4 seconds. A lot faster than the rest, taking XGBoost 85.6 seconds and Random Forest 97.3 seconds. Moreover, the speed of prediction was also very quick. LightGBM was 0.05 seconds, proposing that LightGBM would be perfectly capable in computational terms for real-time scenarios.

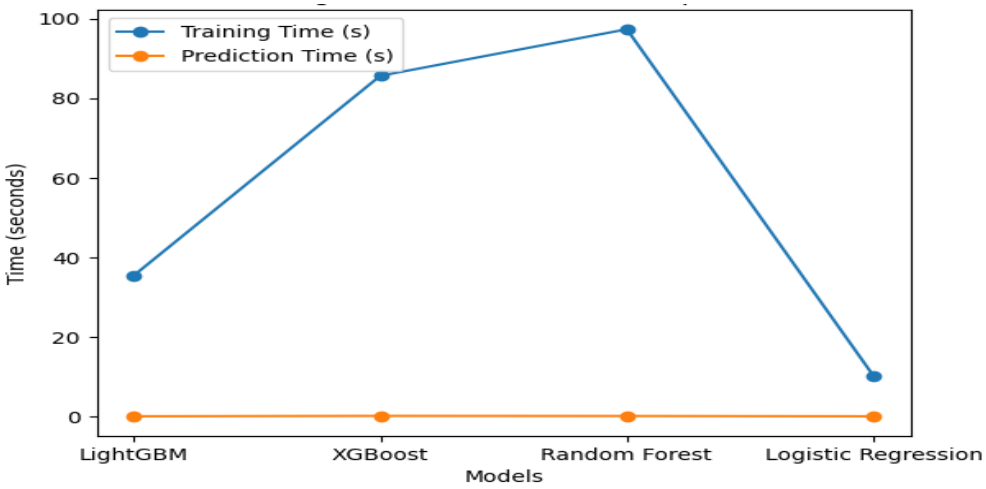

Figure 1. Training and Prediction Time Comparison

The distinctive markup of LightGBM includes its in-built support for categorical features that eliminates the need for one-hot encoding. This decreases memory use and training time and contributes by making LightGBM more efficient than either XGBoost or Random Forest.

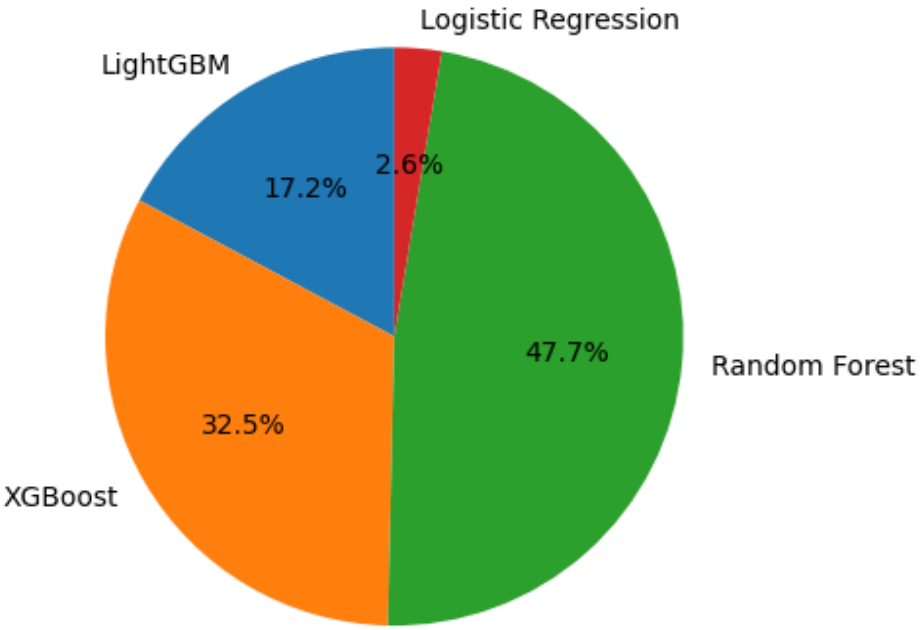

Figure 2. Model Size Comparison

The implementation of the Customer Loyalty Score in LightGBM yielded significant insights into customer engagement and loyalty, directly influencing customer satisfaction and retention. By precisely identifying loyal consumers and those susceptible to attrition, firms may execute tailored marketing campaigns to enhance satisfaction and retention rates.

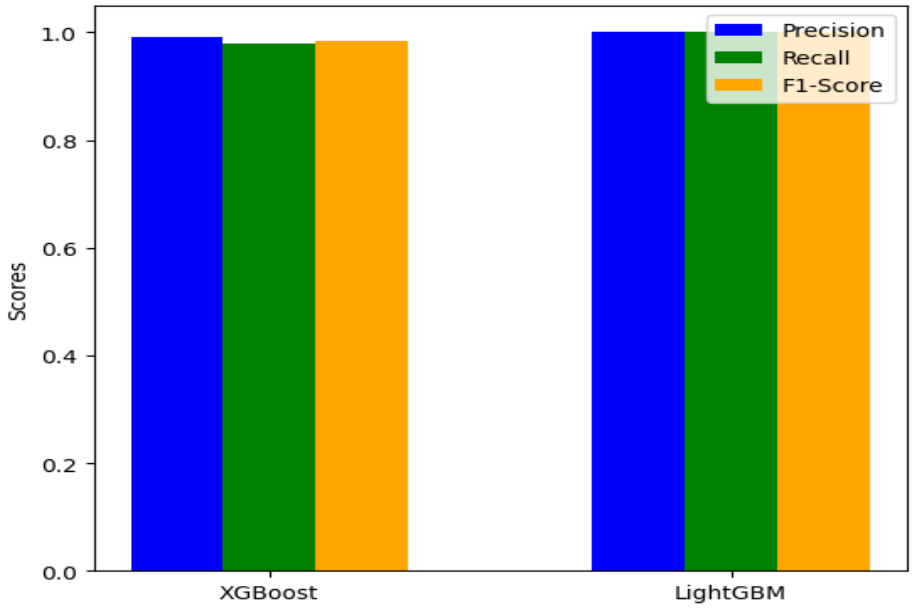

Figure 3. Evaluation Metrics for Class 1

LightGBM's exceptional speed, accuracy, and efficiency make it a great option for e-commerce systems that want to forecast user behaviour in real time. This study shows that LightGBM provides useful insights that are essential for making data-driven decisions, improving customer experiences, and maximising retention tactics.

7. CONCLUSION

This study shows LightGBM is a strong and quick instrument for e-commerce customer behaviour prediction. LightGBM surpasses already used models like XGBoost with flawless accuracy and much shortened training periods. The model's natural handling of categorical data and integration of customer loyalty measures gives companies better understanding of consumer interaction and satisfaction.

Future studies should concentrate on further optimising LightGBM by hyperparameter tuning and investigating its use across several fields, including finance and healthcare. Moreover, improving the interpretability of the model would help to build confidence among corporate players. The results generally show LightGBM's promise as a modern solution for raising consumer retention and happiness in the fiercely competitive e-commerce environment.

REFERENCES

[1] N. Proskurnina, (2020) "Purchasing Decisions Making in the Context of Digital Transformation of Retail," Economics of Development, doi:10.21511/ED.18(4).2019.02.

[2] J. R. Shah and M. Murtaza B., (2016) "Effective Customer Relationship Management through Web Services," Journal of Computer Information Systems.

[3] "Prediction and Buying Behaviour of Customers Using Machine Learning Technique," (2022) Indian Journal of Natural Sciences, vol. 12, no. 70, pp. 39085-39093, ISSN: 0976-0997.

[4] S. Chand, A. K. Shukla, and N. Chandra, (2022) "On-line Customers Buying Behaviour Prediction Using XGBoost Algorithms in Python," Indian Journal of Natural Sciences, vol. 13, no. 73, pp. 46387-46395, ISSN: 0976-0997.

[5] S. Chand, A. K. Shukla, and N. Chandra, (2022) "E-Commerce Customers Buying Behavior Analysis Using Python," International Conference on Application of Artificial Intelligence and Internet of Things on Management, Science and Technology, Indian Academicians and Researchers Association, 9 Dec. 2022, online mode.

[6] F. Rayhan, et al., (2017) "MEBoost: Mixing Estimators with Boosting for Imbalanced Data Classification," doi:10.1109/SKIMA.2017.8294128.

[7] T. Chen and C. Guestrin, (2016) "XGBoost: A Scalable Tree Boosting System," doi:10.1145/2939672.2939785.

[8] P. S. Fader, B. G. S. Hardie, and K. L. Lee, (2005) "RFM and CLV: Using Iso-Value Curves for Customer Base Analysis," Journal of Marketing Research, vol. 42, no. 4, pp. 415-430, doi:10.1509/JMKR.2005.42.4.415.

[9] H. Yang, et al., (2024) "LightGBM Robust Optimization Algorithm Based on Topological Data Analysis," doi:10.48550/arxiv.2406.13300.

[10] "Sales Data FY 2020-2021," (2020) Kaggle, [Online]. Available: https://www.kaggle.com/datasets/nhiyen/sales-data-fy-2020-2021.

[11] G. Senthilkumar, et al., (2022) "Weighted Kernel Based Prediction and Detection of Outliers," 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, pp. 219-224, doi:10.1109/ICOSEC54921.2022.9952112.

[12] H. Los, et al., (2021) "Evaluation of Xgboost and Lgbm Performance in Tree Species Classification with Sentinel-2 Data," 2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 5803-5806, doi:10.1109/IGARSS47720.2021.9553031.

[13] C. Bentéjac, A. Csörgo, and G. Martínez-Muñoz, (2019) "A Comparative Analysis of Gradient Boosting Algorithms," Artificial Intelligence Review, vol. 54, pp. 1937-1967, doi:10.1007/s10462-020-09896-5.

[14] C. Bentéjac, A. Csörgo, and G. Martínez-Muñoz, (2019) "A Comparative Analysis of Gradient Boosting Algorithms," Artificial Intelligence Review, vol. 54, pp. 1937-1967, doi:10.1007/s10462-020-09896-5.

[15] S. Yang and H. Zhang, (2018) "Comparison of Several Data Mining Methods in Credit Card Default Prediction," Intelligent Information Management, vol. 10, pp. 115-122, doi:10.4236/iim.2018.105010.

[16] V. Dev and M. Eden, (2019) "Formation Lithology Classification Using Scalable Gradient Boosted Decision Trees," Computers & Chemical Engineering, vol. 128, pp. 392-404, doi:10.1016/J.COMPCHEMENG.2019.06.001.

[17] E. Sivasankar and J. Vijaya, (2019) "A Study of Feature Selection Techniques for Predicting Customer Retention in Telecommunication Sector," International Journal of Business Information Systems, doi:10.1504/IJBIS.2019.099524.

[18] J. Hancock and T. Khoshgoftaar, (2021) "Leveraging LightGBM for Categorical Big Data," 2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService), pp. 149-154, doi:10.1109/BigDataService52369.2021.00024.

Authors

Dr. Lalit Johari, an Assistant Professor at IFTM University, Moradabad, has eighteen years of teaching experience in Computer Science, specializing in Machine Learning, Python, and Wireless Sensor Networks. A Ph.D. holder in Mobile Ad-hoc Networks, he has published extensively in reputed journals, including IEEE, and actively mentors students while fostering industry-academia collaboration.

Mr. Sachin Agrawal, an Assistant Professor in Computer Science & Engineering at IIMT University, Meerut, has eighteen years of teaching experience. Currently pursuing a Ph.D. from Sharda University, he is recognized for mentoring students, publishing in reputed journals, and fostering innovation. His dedication and expertise make him a respected educator and leader in the field.

Dr. Shikha Verma, an Associate Professor at ABES Engineering College, has eighteen years of experience in teaching and research, specializing in networking, artificial intelligence, and security. She has authored numerous research papers, patents, and books, earning recognition for her contributions to science and technology.

Mr. Laxman Singh, an Assistant Professor at ABES Institute of Technology, Ghaziabad, has seventeen years of teaching experience in Computer Science and Engineering. Specializing in Machine Learning and Multimodal Biometric Systems, he holds an M.Tech from Kurukshetra University and is pursuing a Ph.D. from Bennett University, Greater Noida.