Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Optimizing E-Commerce in Indonesia: Ensemble Learning for Predicting Potential Buyers

Faiz Nur Fitrah Insani, Denny
faiznfi@gmail.com, denny@cs.ui.ac.id
Faculty of Computer Science, Universitas Indonesia

Article Information

Submitted : 12 Jan 2024
Reviewed: 20 Jan 2024
Accepted : 15 Feb 2024

Keywords

Prediction, Xgboost, LightGBM, Random Forest, Soft Voting.

Abstract

In the competitive Indonesian e-commerce sector, data-driven decision-making is crucial for success. This study addresses the challenge faced by a leading e-commerce company, where despite a 134% increase in promotional expenses, active user transactions remained low. Focusing on predicting potential buyers to optimize promotional spending, the research evaluates various ensemble learning methods, including Random Forest, XGBoost, and LightGBM algorithms. Through extensive testing, all three models demonstrated high precision in identifying potential buyers. Remarkably, XGBoost achieved an exceptional precision score of 89.5%. Further enhancement through a soft voting strategy combining XGBoost and LightGBM resulted in the highest precision rate of 89.8%, suggesting a promising approach for targeted marketing and improved promotional strategies in the e-commerce industry.

A. Introduction

The recent advancement of Artificial Intelligence technology has rendered data as an asset that is essential for any company or organization. Mainly in the e-commerce sector in Indonesia, which is currently engaged in fierce competition to gain numerous transactions and expand its customer base. The success in these endeavors significantly relies on how data is processed into valuable information, serving as a powerful tool for companies to excel in various aspects. One notable application is in predicting potential buyers. By leveraging potential users, e-commerce businesses can more easily identify the types and variations of customers accessing their applications. Furthermore, they can implement more targeted marketing strategies, thereby gaining a competitive edge [1]. The primary objective of predicting potential buyers is to minimize marketing costs and enhance company revenue through customer transactions. This is closely tied to promotional strategies such as discounts, cashback, and free shipping, which serve as the ecommerce industry's cutting-edge weapons for acquiring a large customer base [2].

One of the largest e-commerce companies in Indonesia is facing challenges related to the high cost of promotions to users without a corresponding increase in the revenue from active customers in its application. Last August 2023, the promotional expenses increased by 134%, but when correlated with the number of users conducting transactions, it remained stable at 2% of the total active users that month. This is attributed to the failure to address potential buyers, which leads to ineffective distribution of promotions and coupons such as cashback, discounts, and free shipping. Ineffective targeting of promotions has driven up costs significantly. Therefore, there is a need for a method to predict potential buyers, enabling the precise distribution of promotions using customer segmentation that has the potential to conduct transactions.

To predict potential buyers, classification machine learning can be employed by leveraging the analysis of customer behavior data [3]. The objective of analyzing customer behavior data is to identify patterns, trends, and habits performed by customers, which can be used to enhance user experience, optimize marketing strategies, prevent fraud, and improve security [4]. By utilizing this data, the behavior of customers accessing the application can be understood, including their transaction patterns. Subsequently, a model can be created to automatically detect and categorize customers who are inclined to make purchases and those who are not based on their behavior. However, the classification methods using machine learning are diverse; hence, this research aims to find a classification method that can identify users with a tendency to buy or potential buyers. These users can then be classified into a new customer segment, making promotional efforts more precise and targeted. Motivated by these challenges, this paper addresses the following research question: "What is the most precise classification method in predicting potential buyers for e-commerce?" By answering this question, this research aims to identify the most precise classification method, providing recommendations to predict potential buyers in e-commerce companies.

The prediction of potential buyers has been a major focus in the field of customer segmentation, with various studies exploring different methodologies to achieve this goal. For example, paper [5] investigated the effectiveness of several classification algorithms, including Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and K-Nearest Neighbors (KNN). Interestingly, Random Forest emerged as the best-performing algorithm, achieving an impressive accuracy of 89.71% in predicting potential buyers. The paper utilized the Online Shoppers Purchasing Intention dataset from the UCI Repository [6], which provides 17 supporting features relevant to consumer purchasing behavior and a main feature indicating the likelihood of a purchase. These findings highlight the potential of machine learning algorithms in identifying potential buyers, offering valuable insights for businesses to optimize their marketing strategies and improve customer targeting.

In a similar context within the realm of e-commerce, paper [7] explains the methodology for predicting user behavior related to online transactions. By employing the same dataset, this study adopts the CatBoost algorithm, one of the methods of ensemble learning. The evaluation results, based on accuracy, reached 88.51%. Similarly, paper [8] follows the same approach and utilizes the same dataset as the previous research. The distinction lies in this paper's utilization of another ensemble learning algorithm, namely AdaBoost, which yields an accuracy with an ROC area of 91%.

Although addressing different scenarios, paper [9] share a common goal, which is to predict transactions in Online-to-Offline (O2O) settings This study employs the Extreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LightGBM), and Random Forest algorithms. The outcome seeks the highest precision, which is at 89.5% after combining XGBoost and LightGBM. Paper [10] also used the same approach with [9] and achieved an AUC score of 69.5%, exceeding the performance of individual XGBoost and LightGBM models for predicting loyal customers.

Based on previous research and literature reviews, this paper employs ensemble learning algorithms, specifically Random Forest [5], XGBoost [9], and LightGBM [9],[10]. Using ensemble learning algorithms can yield better evaluation results compared to individual algorithms [5],[9].

The evaluation metrics used in papers [5] and [7] prioritize classifying the correct values, with accuracy as the primary measure. However, in the author's research, data imbalance renders accuracy less reliable. Similarly, while recall aims to identify all potential buyers, it can lead to costly promotions for many non-buyers in e-commerce. Therefore, precision emerges as the most appropriate evaluation method, accurately prioritizing the identification of true potential buyers for cost-effective promotion.

B. Literature Review

This section discusses the theoretical basis used in the research, which consists of knowledge sharing and organizational culture.

1. Behavior Analytics

Behavior analytics is a widely used technology in online businesses, particularly in e-commerce. The purpose of employing behavior analytics is to enhance user experience in online transactions, optimize marketing strategies, prevent fraud, and improve security [4]. One of the key features is tracking every user activity, such as buttons clicked, pages viewed, and the duration of user activity on the website. This way, every user's activity on the website can be traced from the moment they enter until they exit the page.

Behavior analytics involves collecting information about user actions and interactions on a website, app, or other digital platform. However, they differ in their scope and purpose. Behavioral analytics goes beyond simply collecting data and involves analyzing it to understand user motivations, needs, and preferences. Machine learning algorithms can generate accurate visual representations of consumer behavior, enabling us to study the dataset in greater detail. This allows for deeper understanding and informed conclusions about overall consumer behavior within the e-commerce platform [11].

2. Classification Model

Classification models in machine learning function as automated sorting mechanisms, assigning data points to specific categories based on their inherent characteristics. In supervised learning, these models are trained using pre-labeled data, where each data point has a predetermined category assigned by experts [12]. The training process aims to extract generalizable patterns from this labeled data, enabling the model to accurately categorize new, unlabeled data points into their corresponding classes.

Two primary objectives guide the development of classification models. Firstly, high performance is crucial, ensuring the model accurately predicts the appropriate category for new data points based on their features. Secondly, interpretability is of utmost importance, providing insight into the model's decision-making process and the relationships between input features and output classifications. This transparency fosters trust in the model and allows for further refinement and improvement.

3. Ensemble Learning

This paper leverages ensemble learning, a method that strengthens predictive performance by training and combining multiple models, for its classification model [13]. Unlike other techniques that learn a single model from data, ensemble learning builds multiple models and combines their predictions to form a final, more robust prediction [14]. Ensemble learning encompasses several popular methods such as bagging and boosting.

Bagging, also known as Bootstrap Aggregating, works by training multiple independent models on slightly different versions of the original training data. These versions are created by randomly sampling with replacement, a technique called bootstrapping. By combining the predictions of these individual models, bagging aims to achieve a more accurate and stable overall performance.

Boosting, on the other hand, takes a sequential approach. It builds a series of models one at a time, with each new model focusing on correcting the errors of the previous one. This is achieved by adjusting the weights assigned to data points based on how well they were classified in the previous model. By focusing on the more challenging data points, boosting iteratively builds a strong predictive model by combining the strengths of these individual "weak learners" into a single, powerful ensemble [13].

4. Random Forest

Random Forest is a one of machine learning algorithm used for classification, regression, and any other tasks involving data processing. This algorithm is an ensemble learning method that combines multiple decision trees to produce predictions that are more accurate and stable than using a single decision tree alone [15].

The Random Forest model is a tree-based ensemble algorithm, meaning it calculates the average prediction from a multitude of individual decision trees. Each tree is built on a unique sample drawn with replacement from the original data set, a process known as bagging or bootstrap aggregating. This technique effectively reduces overfitting, a common problem in individual decision trees. While interpretation of individual decision trees is straightforward, their combination within a Random Forest sacrifices this interpretability for the benefit of significantly improved predictive performance [16]. Notably, Random Forest provides a more accurate estimation of the error rate compared to a single decision tree. This accuracy has been mathematically proven to consistently increase as the number of trees grows [15].

5. Gradient Boosting Machine

Gradient Boosting Machine is a type of ensemble machine learning algorithm used to create a robust predictive model. The basic idea is to build a series of small decision trees, known as weak decision trees, and combine them into a stronger model. The process begins by creating the first tree to predict the outcome. Then, examine where the first model makes errors, and we build the second tree to correct those errors. This process continues, creating additional trees to rectify the remaining errors [13]. The outcome is a combination of decision trees working together to create accurate predictions. Although each decision tree is weak individually, their combination produces a robust model that can be used to predict values of interest. XGBoost and LightGBM are popular implementations of this Gradient Boosting Machine concept. XGBoost (eXtreme Gradient Boosting) and LightGBM are two widely used open-source libraries for implementing Gradient Boosting Machine (GBM) in machine learning. XGBoost is known for its high scalability, algorithm optimization, and success in data science competitions [17]. In the other hands, LightGBM employs a histogram-based algorithm for better training time efficiency and memory consumption, while also leveraging parallel learning optimization for faster execution speed [18].

C. Research Method

1. Research Design

The research design was guided by the CRISP-DM (Cross Industry Standard Process for Data Mining) framework, encompassing the following stages, business understanding, data understanding, data preparation, modeling, and evaluation [19]. This framework provides a structured and efficient approach to the research process, ensuring thorough data analysis and robust model development.

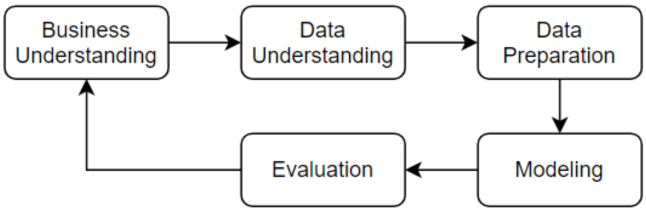

Figure1. Research Design

2. Data Understanding

In this study, the data utilized belongs to ecommerce company behavior analytics data. This dataset comprises information on the behavior of active users engaging within the application, including those using Android and iOS operating systems, as well as web browsers.

The sampling process utilizes the stratified sampling method, ensuring that the samples are randomly chosen while remaining proportional to the target feature. As this study aims to identify potential buyers, the target feature is the transaction, indicating whether the user made a purchase (1) or not (0). The data is directly extracted from the data warehouse, but due to limitations and company policies, only a sample is taken, approximately 10% of the total data for October, which is 7,644,652, with 18 unprocessed features.

Table 1. Feature of Dataset

No	Feature	Description
1	day_name	Day of the week when the user accesses the application
2	visitor_type	Returning user or new user
3	day_type	Weekday or weekend
4	special_day	Special days, such as 10.10
5	onscreen_duration	Duration of user access to the application, in seconds
6	total_session	Number of sessions a user accesses the application
7	client_id	User ID associated with their device(s) (a user can have more than one device)
8	browser	Name of the browser used by the user
9	channel_grouping	Name of the marketing channel that brings the user to the application
10	total_interaction	Number of interactions within the application
11	total_impression	Number of impressions within the application
12	product_clicked	Number of products clicked by user
13	product_view	Number of products impressed by user
14	promo_clicked	Number of promotions clicked
15	promo_viewed	Number of promotions viewed
16	add_to_cart	Whether the user has ever added a product to the cart
17	checkout	Whether the user has ever completed the checkout process (selected a courier)
18	transaction (target)	Whether the user has ever made a transaction

To determine the feature in the dataset, we amalgamate various references as benchmarks, such as visitor_type, onscreen_duration, browser, and channel_grouping, referring to the Online Shoppers Purchasing Intention Dataset [6]. Some other features include total impressions and total interaction [20], action types like add_to_cart, checkout and as well as special_day [10]. The remaining are additional features by author that might have an impact on the transaction feature. The complete list can be observed in table 1.

3. Data Preparation

Data preparation or preprocessing constitutes a crucial stage in machine learning workflows, enabling the identification of anomalies within datasets [21]. This process facilitates data cleaning, deduplication, transformation, and normalization, ultimately ensuring that the data is prepared for machine learning algorithms to learn from.

The initial step involved addressing missing values in the dataset. However, prior to this, the client_id feature was removed as it represented user identification and lacked significant impact on the target feature. Subsequent analysis revealed missing values in both the visitor_type and onscreen_duration features. The type of visitor identified as user type within the application – a value of "1" indicates a new user, while a null value signifies a returning user. Similarly, onscreen_duration represents the time spent within the application, with a null value signifying a user who immediately exited without engaging with the platform (bounce rate). To address these missing values, all instances were replaced with the number "0", indicating a returning user for visitor_type and 0 seconds for onscreen_duration.

Next stage is identifying if there are users with the same interactions within the application that lead to duplication data. This can affect the model's performance as it is deemed insignificant. The number of duplications is 1,277,633 rows and need to perform deduplication using the drop duplicates function from the pandas library. Thus, the final clean data count is 6,367,019.

After the deduplication process, data transformation was implemented to modify values and achieve normalization. This was necessary as the dataset contained numerical features with varying measurements, such as "onscreen_duration" measured in seconds and "product_viewed"/"product_clicked" measured in counts. To ensure comparable ranges across features, min-max normalization was employed using the sklearn library. This technique transformed the original data ranges to a consistent 0-1 scale, all numerical features were scaled to the same range, facilitating a more accurate comparison and analysis. This is particularly important when dealing with machine learning algorithms, as they may be sensitive to the scale of the input data.

Following data transformation and normalization, one-hot encoding was employed to convert categorical features into binary vectors. This approach represents each possible category value as a separate dimension in the vector, where 1 denotes the presence and 0 its absence [22]. In this study, all categorical features, such as platform, browser, day_name, day_type, special_day, and channel grouping, were transformed into individual classes represented by binary values (0 or 1). This process utilized the get_dummies() function from the pandas library, resulting in a final set of 65 features ready for training.

Due to the limited data available for the transaction feature class, comprising only 13% (Table 2), undersampling of the majority class was performed. This targeted the reduction of the majority class, users active in the application but not making transactions. The author conducted an experiment, exploring the impact of class imbalance on model performance through four different scenarios. The first scenario used raw data with an extreme imbalance of 87% majority class and 13% minority class. Subsequent scenarios used progressively more balanced ratios: 75:25, 60:40, and finally, 50:50. Interestingly, the 50:50 ratio yielded the highest precision score, indicating that addressing class imbalance is crucial for achieving optimal performance. Consequently, the final distribution consisted of 767,962 users conducting transactions and 767,962 users not engaging in transactions.

Table 2. Imbalance Data Handling

Transaction	Count before (%)	Count after (%)
false	5,123,878 (87%)	767,962 (50%)
true	767,962 (13%)	767,962 (50%)

4. Modeling

Ensemble learning algorithms, specifically Random Forest, LightGBM, and XGBoost, will be employed to generate models from the cleaned and balanced data. The implementations will utilize the sklearn, xgboost, and lgbm libraries.

This phase utilizes cross-validation to guarantee a robust model. This technique repeatedly divides the data into k equal folds, where each fold is used as the validation set once while the remaining k-1 folds serve for training. This process is iterated k times (in this case, k=5) to prevent overfitting and generate a more reliable estimate of the model's generalizability [23].

Hyperparameter tuning was also performed to identify the most effective model configuration. This process involved two distinct phases. First, the model was trained and evaluated using its default set of parameters, establishing a baseline performance measure. Subsequently, a grid search strategy was employed to explore a comprehensive range of hyperparameter combinations. This technique systematically assesses model performance across various parameter settings, enabling the discovery of optimal values. Leveraging grid search, along with insights from relevant articles and previous research [9],[10] optimal hyperparameters were identified to enhance model performance.

5. Evaluation

Since the business objective is to increase transactions per user while maintaining cost-effectiveness, prioritizing precision as the primary model evaluation metric is critical. This prioritization stems from the need to minimize False Negatives (FN), which represent missed opportunities to engage potential transacting customers. By focusing on precision, the model can accurately identify true positive (TP) users who are likely to transact, ensuring a more targeted and cost-effective promotional strategy compared to targeting all users. Precision, which measures the proportion of predicted positives that are truly positive, helps achieve this goal by precisely assessing the model's ability to identify potential buyers.

D. Result and Discussion

The results and experiments from implementing the classification algorithms were presented based on two experimental scenarios. The first scenario involved using the algorithms with their default hyperparameter settings, while the second scenario employed grid search to identify the optimal hyperparameter configurations and subsequently applied them to the algorithms. A comparative analysis of the precision scores and a comprehensive interpretation of the results were then conducted. This approach was implemented in a Python environment with version 3.9.17, utilizing the scikit-learn library version 1.3.0 and the Spyder interface within Anaconda.

Table 3. Result of cross validation on Classification Model

Evaluation	RF	LightGBM	XGBoost
Accuracy	0.931	0.932	0.933
Precision	0.889	0.888	0.892
Recall	0.985	0.988	0.984
F1 Score	0.935	0.935	0.936
AUC Score	0.931	0.932	0.933

The first experiment, which utilized the default hyperparameter settings, revealed that XGBoost outperformed the other algorithms in terms of accuracy, precision, F1-score, and AUC score. LightGBM achieved the highest recall score. These results were averaged over multiple iterations. Random Forest had the lowest performance of all the models. In conclusion, XGBoost emerged as the most precise algorithm among the tested models in the first experiment, achieving a precision score of 89.2%. LightGBM followed closely with a score of 88.9%, while Random Forest delivered the lowest precision at 88.8%.

Table 4. Grid Search Hyperparameter Tuning input and result

Random Forest	XGBoost	LightGBM
'max_depth': [5,50]	'learning_rate': [0.1, 0.2],	'learning_rate': [0.01, 0.1, 0.3],
'max_features': [2,3]	'n_estimators': [100, 200],	'n_estimators': [100, 200],
'min_samples_leaf': [3, 10],	'max_depth': [25, 50],	'max_depth': [10,50],
'min_samples_split': [5, 10],	'subsample': [0.3, 0.5],	'min_child_samples': [10, 20 ,30],
'n_estimators': [100, 500]	'colsample_bytree': [0.5, 1.0]	'colsample_bytree': [0.5, 1.0]

The second experiment involved hyperparameter tuning to optimize the performance of the classification models. Utilizing the grid search method, various parameters were configured and iteratively tested to identify the optimal configuration for each model. The specific parameters employed are presented in Table 4. Specifically, XGB utilized subsample to determine the fraction of data samples used in each iteration, while LightGBM employed min_child_samples to control the minimum number of samples required in a leaf node. The results of hyperparameter tuning are highlighted in red.

Table 5. Hyperparameter Tuning Evaluation Result

	RF	LightGBM	XGBoost
Precision before	0.889	0.888	0.892
Precision after	0.889	0.890	0.895
Recall before	0.985	0.988	0.984
Recall after	0.985	0.985	0.985

The algorithm yielded a model with improved precision compared to the baseline, as shown in Table 5. XGBoost achieved the highest individual precision of 89.5%. The average recall across all algorithms was 98.5%. Encouraged by these results, the authors plan a third experiment to explore the potential for further performance gains. This experiment involves combining each algorithm combination (RF with LightGBM, RF with XGBoost, and XGBoost with LightGBM) using a soft voting classifier for enhanced accuracy, as suggested in [24].

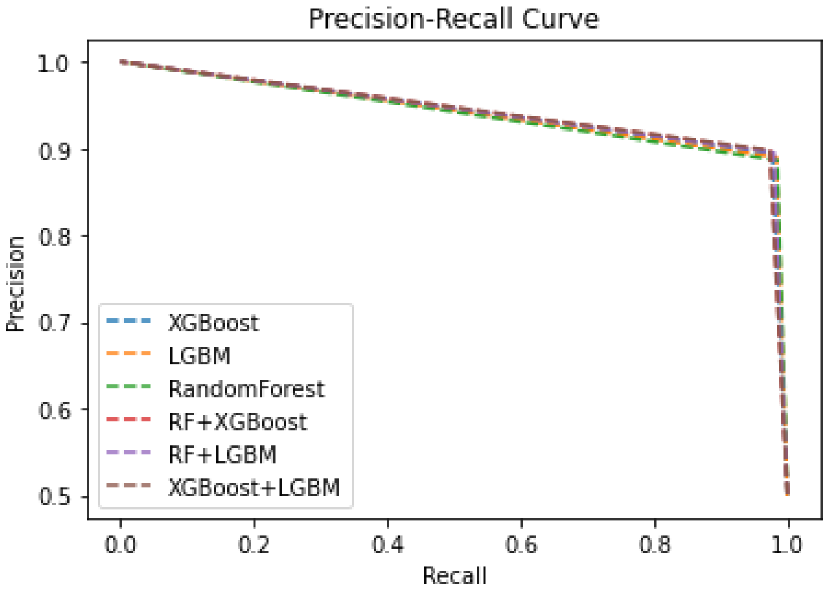

Figure 2. Precision-Recall Curve

A precision-recall curve was created to visualize the performance of each model, especially the combined models. The results showed that all three models had improved precision, with RF + LightGBM achieving 89.5%, RF + XGBoost achieving 89.7%, and the highest precision of 89.8% achieved by the combination of XGBoost and LightGBM. This is shown in Figure 2, where XGBoost and LightGBM have the highest line on the y-axis, which represents precision

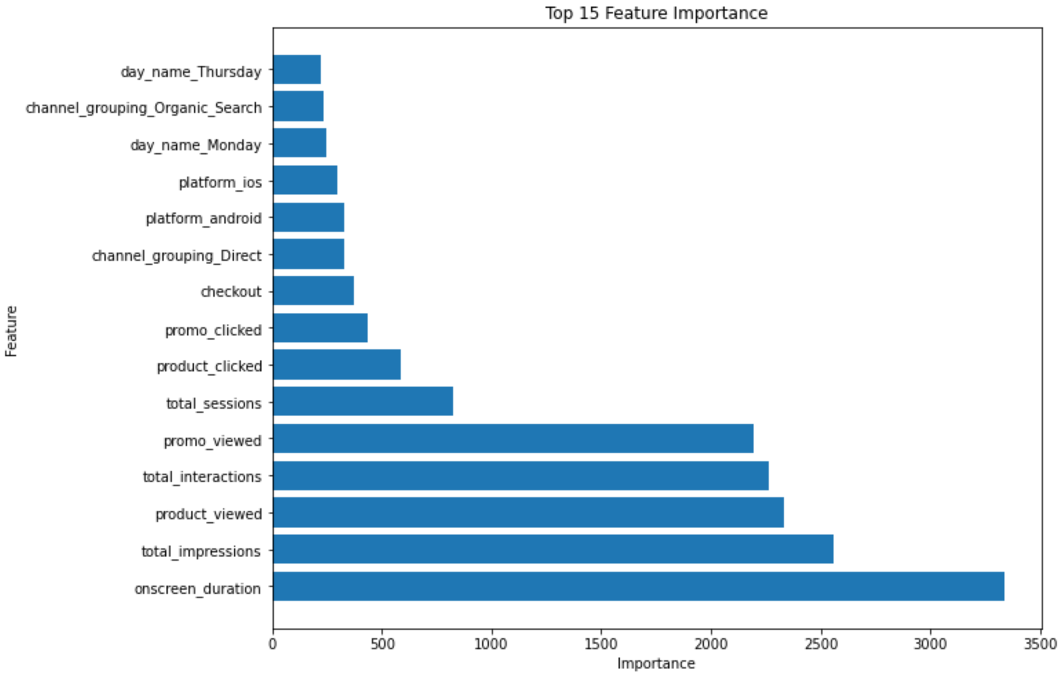

Figure 3. Feature Importance

Following this, an examination of feature importance was conducted to identify the features most influential on the target variable, in this case, the transaction. Figure 3 presents the top 15 most influential features. According to the combined XGBoost and LightGBM models, "onscreen_duration" emerges as the most influential feature, indicating that longer user engagement with the e-commerce application significantly increases the likelihood of a purchase. This supports the researcher's hypothesis that extended app usage correlates with increased purchase potential. The second most influential feature is "total_impression," suggesting that increased exposure to items (widgets, videos, tickers, etc.) within the e-commerce platform also incentivizes purchases.

E. Conclusion

This research employed ensemble learning for classification, specifically leveraging the Random Forest, XGBoost, and LightGBM algorithms. Through extensive testing and experimentation, all three models achieved high precision scores, with XGBoost reaching the peak at 89.5%. A subsequent soft voting approach combined the individual models to further improve overall precision. This resulted in the XGBoost and LightGBM combination achieving the highest precision of 89.8%.

The limitation of this research lies in the constraints imposed by company policies, which prohibit the acquisition of whole data, resulting in only a limited sample being used. For future works, the addition of more data is suggested to enhance the models' performance in predicting potential buyers. The consideration of alternative cases and scenarios is also recommended, given the continuous advancement of technology. It is anticipated that more sophisticated and reliable classification algorithms will emerge over the years, contributing to improved binary classification predictions.

F. Acknowledgment

This paper owes its completion to the invaluable support and resources provided by the University of Indonesia, whose contributions have been fundamental in facilitating our research. Additionally, we extend our heartfelt gratitude to a prominent e-commerce company in Indonesia, whose cooperation and insights have been pivotal, yet due to confidentiality agreements, their name cannot be disclosed. Their willingness to participate and provide essential data has been a cornerstone in the success of this study. We deeply appreciate the guidance, expertise, and resources these organizations have generously offered, making this research not only possible but also more impactful.

G. References

[1] B. Shen, "E-commerce Customer Segmentation via Unsupervised Machine Learning," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jan. 2021. doi: 10.1145/3448734.3450775.

[2] X. Cheng, S. Deng, X. Jiang, and Y. Li, "Optimal promotion strategies of online marketplaces," *Eur J Oper Res*, vol. 306, no. 3, pp. 1264–1278, May 2023, doi: 10.1016/j.ejor.2022.08.020.

[3] A. Martínez, C. Schmuck, S. Pereverzyev, C. Pirker, and M. Haltmeier, "A machine learning framework for customer purchase prediction in the non-contractual setting," *Eur J Oper Res*, vol. 281, no. 3, pp. 588–596, Mar. 2020, doi: 10.1016/j.ejor.2018.04.034.

[4] S. Steensen, R. Ferrer-Conill, and C. Peters, "(Against a) Theory of Audience Engagement with News," *Journal Stud*, pp. 1662–1680, 2020, doi: 10.1080/1461670X.2020.1788414.

[5] S. Sharma and H. K. Soni, "Discernment of potential buyers based on purchasing behaviour via machine learning techniques," in *Proceedings of 2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering, ICADEE 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020. doi: 10.1109/ICADEE51157.2020.9368935.

[6] C. Sakar and Y. Kastro, "Online Shoppers Purchasing Intention Dataset." 2018.

[7] X. Dou, "Online Purchase Behavior Prediction and Analysis Using Ensemble Learning," *IEEE 5th International Conference on Cloud Computing and Big Data Analytics*, 2020.

[8] D. Algawiaz, G. Dobbie, and S. Alam, "Predicting a User's Purchase Intention Using AdaBoost," *IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, 2019.

[9] L. Yang, J. Wu, X. Niu, and L. Shi, "Towards purchase prediction: A voting-based method leveraging transactional information," in *2022 5th International Conference on Data Science and Information Technology, DSIT 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/DSIT55514.2022.9943898.

[10] B. Zhao, A. Takasu, R. Yahyapour, and X. Fu, "Loyal consumers or one-time deal hunters: Repeat buyer prediction for E-commerce," in *IEEE International Conference on Data Mining Workshops, ICDMW*, IEEE Computer Society, Nov. 2019, pp. 1080–1087. doi: 10.1109/ICDMW.2019.00158.

[11] V. Shrirame, J. Sabade, H. Soneta, and M. Vijayalakshmi, "Consumer Behavior Analytics using Machine Learning Algorithms," in *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2020, pp. 1–6.

[12] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *SN Appl Sci*, vol. 3, no. 2, Feb. 2021, doi: 10.1007/s42452-021-04148-9.

[13] O. Sagi and L. Rokach, "Ensemble learning: A survey," *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 8, no. 4, p. e1249, 2018.

[14] J. Alzubi, A. Nayyar, and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Nov. 2018. doi: 10.1088/1742-6596/1142/1/012012.

[15] L. Breiman, "Random Forests," 2001.

[16] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata Journal*, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.

[17] T. Chen and T. He, "xgboost: eXtreme Gradient Boosting," 2023.

[18] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[19] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 2000, pp. 29–39.

[20] S. Patil, K. Raut, P. Palsodkar, T. Singh, Y. Dubey, and R. Umate, "Click Prediction Learning for Effective Advertising," in *2022 International Conference on Emerging Trends in Engineering and Medical Sciences, ICETEMS 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 283–288. doi: 10.1109/ICETEMS56252.2022.10093291.

[21] J. J. Davis and A. J. Clark, "Data Preprocessing for Anomaly Based Network Intrusion Detection: A Review," *Comput. Secur.*, vol. 30, no. 6–7, pp. 353–375, Sep. 2011, doi: 10.1016/j.cose.2011.05.008.

[22] L. Yu, R. Zhou, R. Chen, and K. K. Lai, "Missing data preprocessing in credit classification: One-hot encoding or imputation?," *Emerging Markets Finance and Trade*, vol. 58, no. 2, pp. 472–482, 2022.

[23] T. T. Wong and P. Y. Yeh, "Reliable Accuracy Estimates from k-Fold Cross Validation," *IEEE Trans Knowl Data Eng*, vol. 32, no. 8, pp. 1586–1594, Aug. 2020, doi: 10.1109/TKDE.2019.2912815.

[24] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, 2021.