KS Knowledge Engineering and Data Science (KEDS)
Vol 7, No 1, April 2024, pp. 58–70

pISSN 2597-4602
eISSN 2597-4637

Optimising the Fashion E-Commerce Journey: A Data-Driven Approach to Customer Retention

Hasna Luthfiana Fadhila [a,1], Vynska Amalia Permadi [a,2,*], Sylvert Prian Tahalea [a,b,3]

[a] *Department of Informatics, Universitas Pembangunan Nasional Veteran Yogyakarta*
Jl. Padjajaran Jl. Ring Road Utara No.104, Sleman 55283, Indonesia
[b] *Institute of Informatics, University of Szeged, Szeged*
Szeged, Dugonics tér 13, 6720, Hungary
[1] *hasnaluthfiana@gmail.com;* [2] *vynspermadi@upnyk.ac.id*;* [3] *sylvert@inf.u-szeged.hu*
**corresponding author*

ARTICLE INFO

Article history:
Received 26 June 2024
Revised 08 July 2024
Accepted 05 August 2024
Published online 24 August 2024

Keywords:
Churn
Customer Retention
CRISP-DM
Data Driven
Fashion e-commerce

ABSTRACT

A fashion e-commerce company offers a wide range of products from domestic and international brands that are popular with young people. However, there has been an increase in non-organically acquired customers, many of whom do not return to make repeat purchases. This has led to a higher customer churn rate, with a significant proportion of non-organically sourced customers failing to become repeat purchasers. Consequently, a churn analysis and prediction model were developed to address this issue. This paper employs the Recency, Frequency, and Monetary (RFM) framework for churn analysis and prediction. The framework is underpinned by three key dimensions: last purchase recency, purchase frequency, and total transaction value. Seven machine learning algorithms were evaluated to identify the optimal approach. Following a comparative analysis of these models, Random Forest emerged as the superior algorithm, demonstrating an accuracy of 0.99, precision of 0.97, recall of 0.99, ROC AUC of 0.98, and F1-score of 0.97. Consequently, this model will be utilized for churn prediction. Based on the analysis and modelling, several recommendations are offered to enhance customer retention for the fashion e-commerce platform. In addition to predicting churn, this paper provides insights into potential refinements to the churn prediction model, such as real-time monitoring, personalized customer experiences, analysis of customer feedback, and lifetime value analysis.

I. Introduction

The fashion e-commerce market in Indonesia has experienced substantial growth over the past five years, driven by increased internet penetration, smartphone adoption, and a burgeoning middle class with higher disposable incomes. Several studies were conducted in academic literature focusing on the themes of data-driven marketing and e-commerce performance; however, there is a lack of research on the specific applications of this approach to the fashion sector [1]. Moreover, the research conducted so far mainly belongs to the earliest phase of the customer lifecycle, ignoring the importance of understanding how the customer is engaged and stimulated to perform specific actions in the following lifecycle stages: usage, purchases, loyalty, and reactivation. Vezzetti [2] and Silva [3] highlight the potential of big data analytics in the fashion industry, particularly in understanding customer tastes and trends and enhancing the customer experience. However, they do not specifically address using predictive analytics to improve the e-commerce experience. Cirqueira [4] provides a conceptual framework for predicting customer purchase behaviour in e-commerce, which could be applied to the fashion sector. However, there is a gap in the literature in combining data collection with predictive analytics to provide specific suggestions for improving the customer e-commerce experience in the fashion sector.

The role of data-driven recommendation in fashion is crucial, particularly in the context of e-commerce and the ever-changing desires of fashion lovers [5]. These systems are used to address complex problems in fashion e-commerce. Examples include social fashion-aware recommendations,

https://doi.org/10.17977/um018v7i12024p58-70

product recommendations, and size and fit recommendations [6]. However, there is a need for further research in this area, such as leveraging other data-driven approaches to bring recommendations to business owners. A range of studies have explored churn prediction in the fashion industry. Hakim [7] used machine learning techniques, achieving a 72.472% accuracy rate using the RandomForestClassifier model. Khodabandehlou [8] used Recency, Frequency, and Monetary (RFM) analysis with the framework of customer churn through six stages and achieved a 97.92% accuracy rate. Khan [9] presented a framework for early churn detection, achieving an 89.4% accuracy rate in a mobile phone network. Tamaddoni [10] compared churn prediction techniques, finding that boosting and logistic regression were practical in different contexts. These studies collectively highlight the potential of machine learning and data analysis in predicting customer churn in the fashion industry.

The potential application of machine learning to churn prediction within the fashion industry offers a promising avenue for gaining valuable insights to inform business strategy. It has been employed to predict or forecast churn [11][12][13][14][15], with specific algorithms demonstrating auspicious results [16][17][18]. While support vector machines (SVMs) were once a popular choice for churn prediction [19], it is not the best-performing algorithm by far [14][20][21]. Logistic regression has been successfully applied to churn prediction, with performance improvements achieved through optimization [22][23]. However, decision trees and random forests have consistently yielded superior predictive results, even without extensive parameter tuning [24][25][26][27]. Furthermore, the incorporation of ensemble methods such as AdaBoost [28][29][30], and XGBoost [31][32] has often led to significant performance enhancements for decision trees and random forests. Extra trees have also emerged as a high-performing algorithm for churn prediction, comparable to modified random forests and decision trees [13][33][34][35].

This article presents the viewpoint of utilizing a data-driven approach to predict customer churn in the fashion industry. The usage of several algorithms, such as Logistic Regression, k-NN, Random Forest, ADA Boost, XGBoost, Decision Tree, and Extra Tree, demonstrates the potential for accurate churn prediction. The findings of this research offer valuable insights for business owners to implement strategies to predict churn, such as enhancing the user experience in several ways.

II. Methods

Machine learning modelling requires thorough data preparation and feature engineering as essential preliminary steps. The following datasets will be employed to develop a robust model: The Customer Dataset includes detailed information on 10,000 registered customers, each characterized by 15 distinct data points. The Product Dataset encompasses 44,424 entries, where each product is described by 10 data attributes, offering a comprehensive overview of the available items on the platform. The Transaction Dataset contains an extensive collection of 852,584 transactions, each accompanied by 14 data points, which provide critical insights into customer purchasing behaviours and prevailing market trends. Additionally, the Clickstream Dataset records a vast amount of user activity on the platform, with 12,833,602 entries, each comprising 6 data points, offering valuable information on user behaviour and interaction patterns.

The Customer segmentation analysis used the RFM framework to identify the dependent variable (churn). This technique analyses customer behaviours based on three key dimensions: recency of the last purchase, purchase frequency, and total transaction value for each customer [36]. Following the RFM analysis, customers are categorized into eleven segments based on their assigned RFM scores. The segmentation labels used in this model are provided in Table 1.

This research employs a Data Science analysis model shown in Figure 1 based on the CRISP-DM framework [37]. The first stage, business understanding, focuses on establishing the context of the problem. As detailed in the previous section, four datasets of the fashion e-commerce platform will be utilized for machine learning model development in predicting customer churn. This stage delves into understanding the scope of the business and the datasets involved. The process involves an in-depth dataset analysis, paying close attention to their fields or columns.

The Customer Dataset comprises detailed information about registered customers and 15 data fields. This dataset includes a unique customer identifier, an integer, and 12 text-based fields such as names, emails, genders, and locations. Additionally, two numerical fields capture birthdays and join

dates, representing floats. These fields provide a comprehensive profile of each customer, which is vital for understanding customer behaviour and predicting churn.

Table 1. RFM segment categorisation

Segment	Description
Champions	Customers who have made the most recent, frequent, and high-value transactions.
Loyal	Customers who have made recent and frequent transactions but have lower transaction values.
Potential Loyalists	Customers who have made recent and high-value transactions but with lower transaction frequency.
New Customers	Customers who have made recent transactions but with lower transaction frequency and value.
Promising	Customers who have made recent transactions with increasing frequency and value.
Need Attention	Customers who have made recent transactions with decreasing frequency and value.
About to Sleep	Customers who have not made transactions recently but have a history of high-value transactions.
Cannot Lose Them	Customers who have not made transactions recently but have a history of high-value and frequent transactions.
At Risk	Customers who have not made transactions recently but have a history of low-value and frequent transactions.
Hibernating Customers	Customers who have not made transactions recently and have a history of low-value transactions.

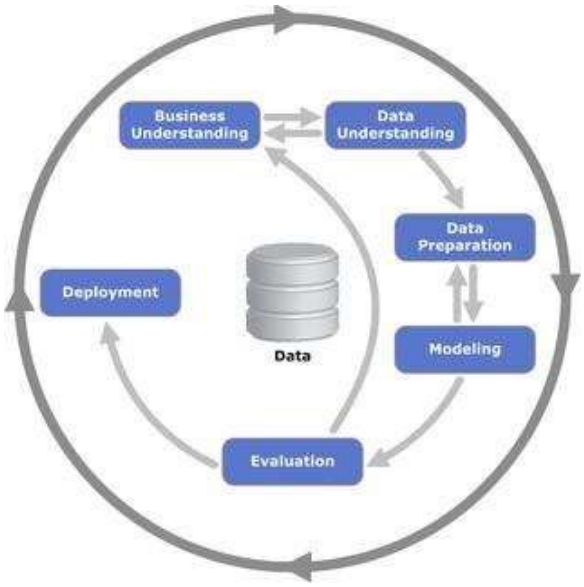

Fig. 1. CRISP-DM framework

The Product Dataset offers detailed insights into the products available on the platform, featuring 10 data fields. This dataset includes a unique product identifier, an integer, and eight text-based fields describing various product aspects, including names, categories, brands, descriptions, and shipping details. Additionally, a single numerical field represents product prices stored as floats. These details are essential for analyzing product performance and their influence on customer retention.

The Transaction Dataset captures detailed transaction data from the platform, consisting of 14 fields. It likely includes unique identifiers for transactions, customers, products, and payment methods, all represented as integers. The dataset also contains timestamps for transaction dates, stored as strings or objects, and text-based data describing shipping methods and event types. Numerical fields in this dataset capture transaction amounts and product discounts, represented as floats. This dataset is crucial for understanding purchasing behaviours and identifying patterns that may lead to customer churn.

Lastly, the Clickstream Dataset records customer activity on the platform, comprising six text-based fields stored as strings or objects. This dataset provides valuable insights into user interactions on the platform, which are critical for understanding how online customer behaviour correlates with the likelihood of churning. Each of these datasets plays a significant role in building a comprehensive model to predict customer churn, which is the goal of this research.

Further data exploration was conducted through visualization techniques. Pie charts, histograms, bar plots, scatter plots, tree map diagrams, and kernel density plots were employed to uncover insights from the data. Next, during customer data preparation, time data in the "birthdate" and "first_join_date" fields (originally in string format) was converted to the DateTime data type for improved analysis. A new field named "age" was also created based on the birthdate information.

The product dataset required handling missing values. Several fields, including "baseColour," "season," "year," "usage," and "productDisplayName," contained null values. Due to the large number of affected rows, deletion was deemed impractical. Instead, missing values in these fields will be addressed later by replacing them with the most frequent value (mode) within each field. Similar to the customer data, the transaction dataset had time data in string format. The "created_at" field, indicating transaction time, was converted to the DateTime data type.

Furthermore, two metadata fields, "product_metadata" and "event_metadata" in the transaction and clickstream datasets, respectively, were identified. These fields will be restructured to extract valuable information for further analysis. After data preparation has been carried out on each dataset, the four datasets above are merged to create new features, which will later be selected as churn features for machine learning models.

Following data preparation, we evaluated several machine learning algorithms to predict customer churn in the fashion e-commerce platform. These algorithms include Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, AdaBoost (Adaptive Boosting), XGBoost, Decision Tree, and Extra Trees (Extremely Randomised Trees). This modeling step aimed to select the optimal model. Following the training and testing of these seven models, we used various evaluation metrics, such as accuracy, precision, recall, ROC (Receiver Operating Characteristic)/ AUC (Area Under the Curve), and F1 Score, to determine the model that best predicts customer churn.

III. Results and Discussion

In business data understanding, a data exploration phase was initiated to understand better customer behavior and potential churn factors in the Fashion E-Commerce platform. This analysis revealed several key insights from each dataset.

An analysis of the customer dataset visualization in Figure 2 to Figure 4 reveals that the Fashion E-commerce platform predominantly caters to a young female demographic, with the highest concentration of customers residing in Jakarta. Interestingly, a significant surge in new customers coincided with the peak of the COVID-19 Delta variant in July 2021. This observation suggests a potential shift towards online shopping during lockdown periods, which may have influenced customer spending habits.

While the product data showcased in Figure 5 to Figure 6 primarily focuses on apparel, it is interesting that the selection leans more heavily towards men's clothing despite the customer base comprising a higher proportion of female customers.

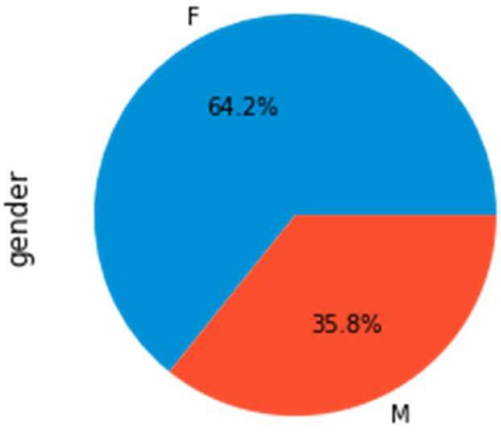

Fig. 2. Customer dataset by gender distribution

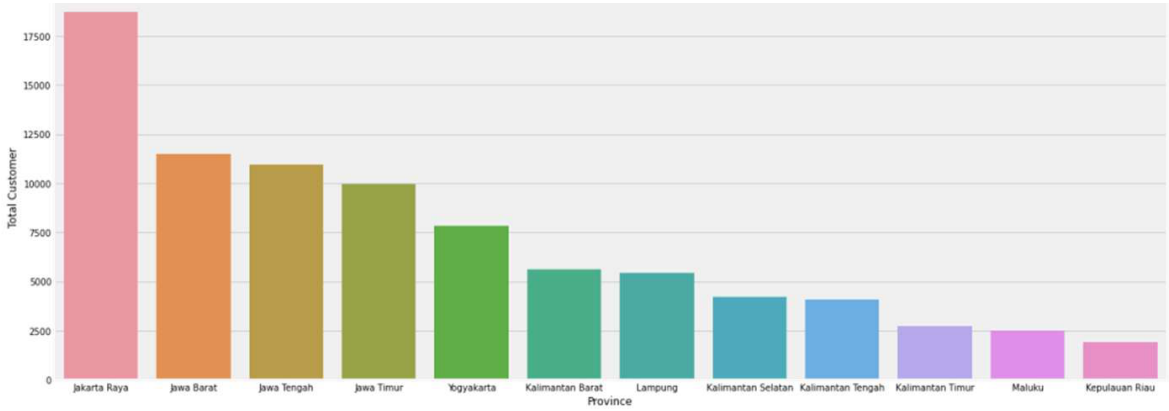

Fig. 3. Customer dataset by province

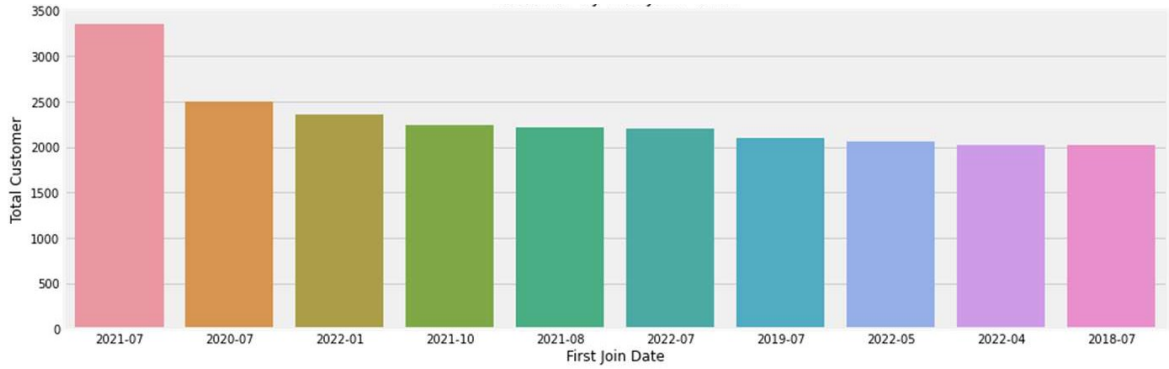

Fig. 4. Customer dataset by first join date

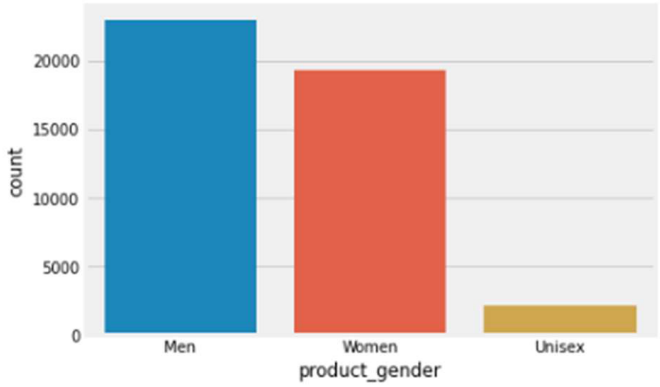

Fig. 5. Product dataset: product categories by gender

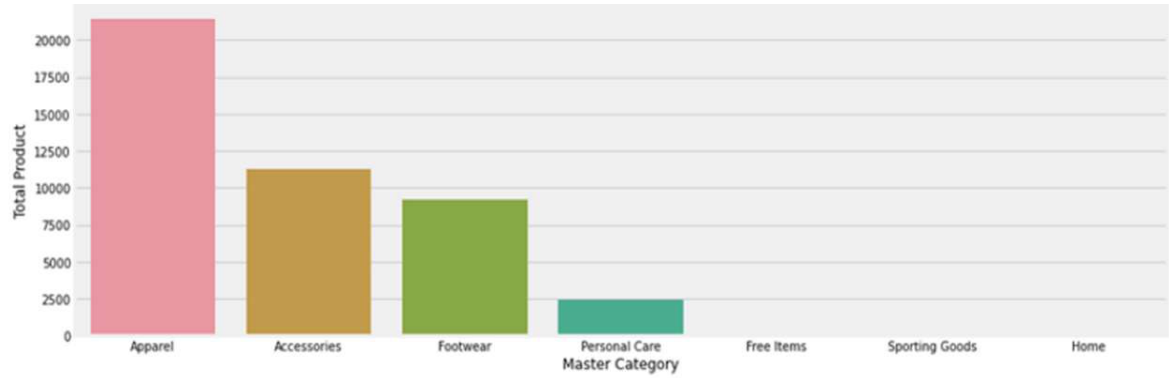

Fig. 6. Product categories by master category

Data visualization of the transaction dataset (Figure 7 to Figure 10) revealed that the "AZ2022" promo code was the most popular. At the same time, credit cards were the preferred payment method, followed by Gopay and OVO. Transaction volumes were unsurprisingly highest in Jakarta Raya and West Java. Additionally, the data indicated a late-night peak in transactions (10 PM until midnight) compared to mornings and afternoons.

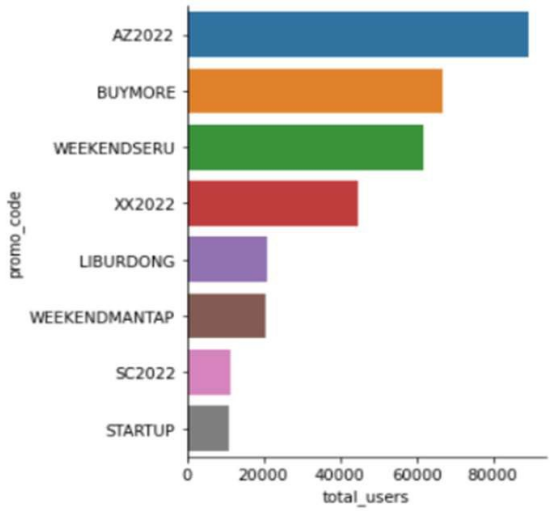

Fig. 7. Transactions by voucher code used

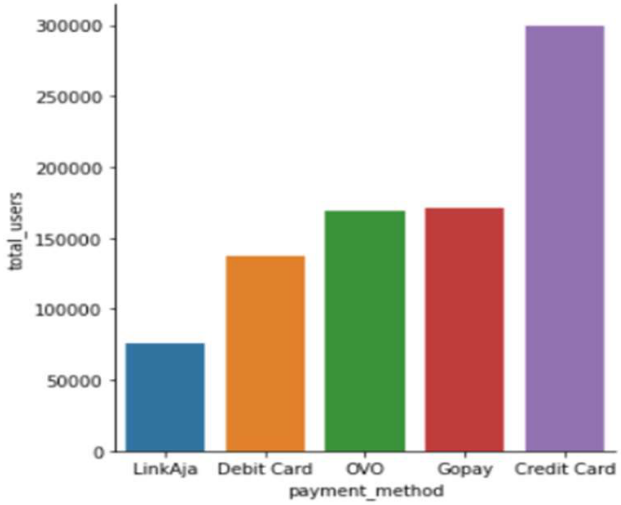

Fig. 8. Transactions by payment methods

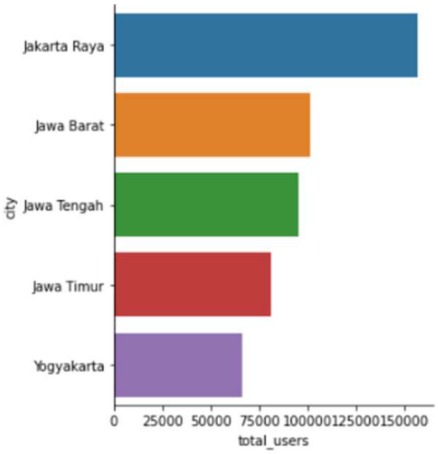

Fig. 9. Transactions by regions

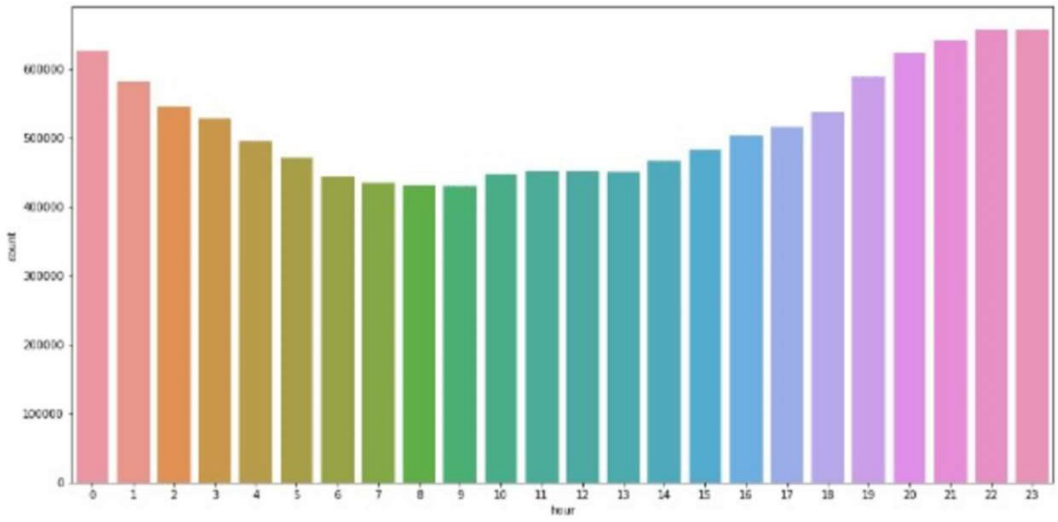

Fig. 10. Transactions by hours

The clickstream data reveals a strong preference for mobile access (Figure 11 to Figure 13), with nearly 90% of users accessing the platform via mobile devices. Clicking is the most common user activity, followed by adding promos. Interestingly, user activity peaks on Sundays and at 8 PM, with the lowest activity at 7 AM.

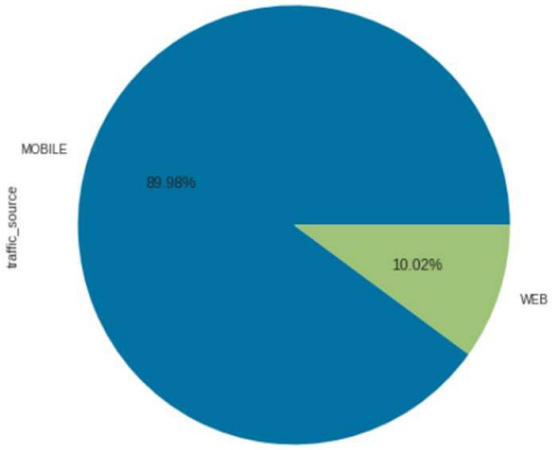

Fig. 11. Clickstream by users' devices

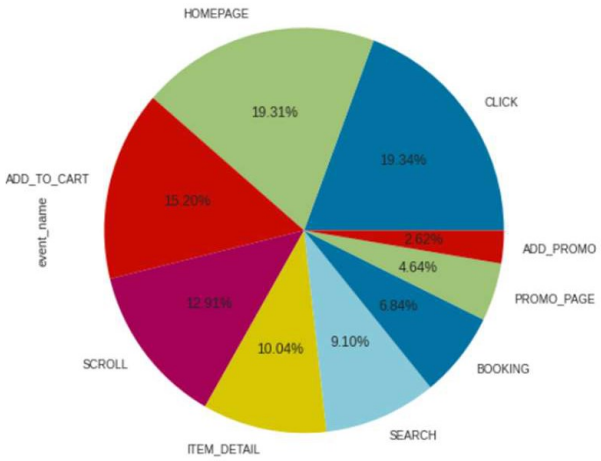

Fig. 12. Clickstream by click activity

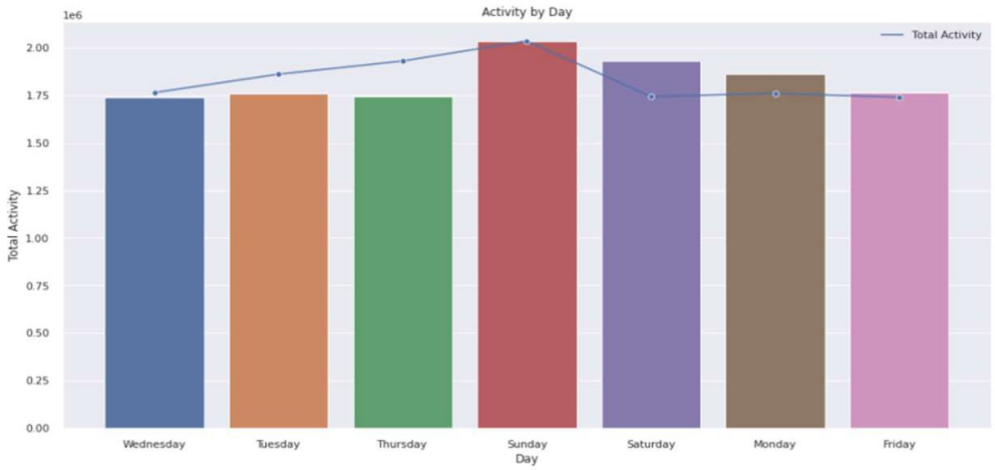

Fig. 13. Clickstream by day

The next phase is data preparation. The first step in this phase involved merging the four available datasets. Following the merger, RFM segmentation was applied to identify churned and non-churned customers. Recency represents the time elapsed since the last transaction, frequency refers to the number of transactions per customer, and monetary value reflects the total transaction amount. These three variables were then integrated into the merged dataset. Finally, based on the RFM variables, each variable was assigned a score, resulting in a combined RFM score for each customer, as illustrated in Table 2.

Table 2. RFM segmentation result

Customer_id	recency	frequency	monetary	recency_ score	frequency_ score	monetary_ score	rfm_ score	segment
3	33	682	299044279	4	5	5	455	Champions
8	77	202	63812004	3	4	3	343	Need Attention
9	64	59	33507505	4	3	3	433	Potential Loyalist
11	146	14	2765462	3	1	1	311	New Customers
15	207	57	23704990	2	3	3	233	Hibernating customers

Customer segments were then assigned churn labels based on their RFM characteristics to facilitate further analysis. Four segments – "Cannot Lose Them," "At Risk," "Hibernating Customers," and "Lost Customers" are designated as the "Churn" class, indicating a higher likelihood of customer churn. The remaining seven segments – "Champions," "Loyal," "Potential Loyalists," "New Customers," "Promising," "Need Attention," and "About to Sleep" classified as "Not Churn," representing customers with a lower churn risk. After integrating the "churn" classification field into the primary dataset, we conducted a correlation test to assess the relationships between other variables and the churn variable. The newly added field unearthed valuable insights. Customers who churn quickly tend to be more discount-driven (Figure 14), while older customers (over 50) might churn due to decreased spending (Figure 15).

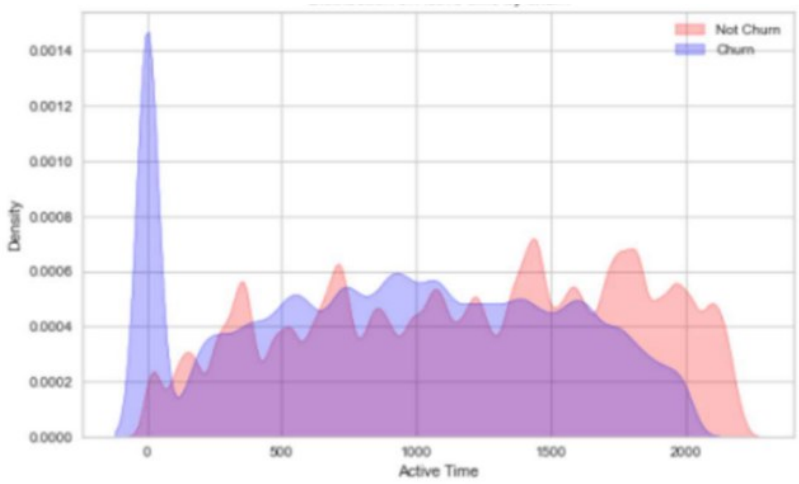

Fig. 14. Distribution of active time by churn

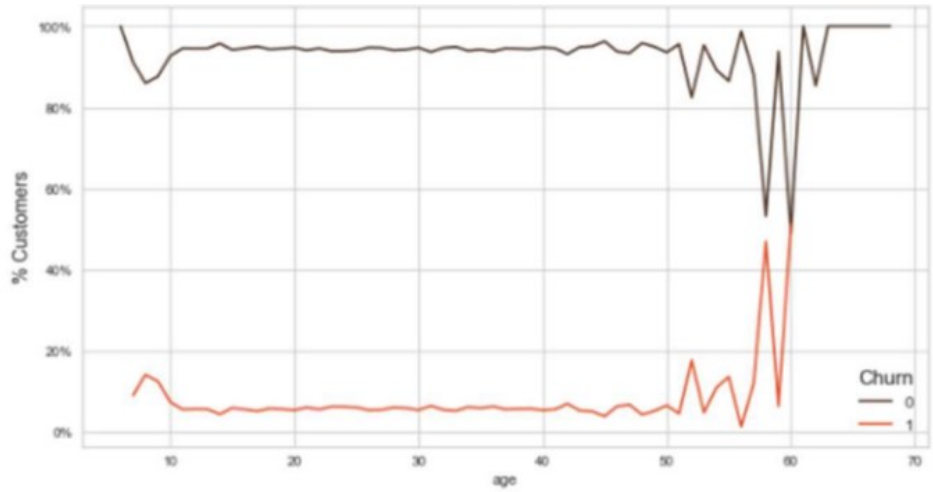

Fig. 15. Distribution of active time by churn

The correlation test results (Figure 16) suggest that individual features may not exert solid and independent effects on churn. The colour intensity within the given colour scheme corresponds to the strength of the correlation. Positive correlations indicate that as one variable increases, churn tends to increase as well. Conversely, negative correlations suggest that as one variable increases, churn tends to decrease. While the correlation matrix provides an initial overview of potential relationships between variables, further analysis and domain expertise are necessary to draw definitive conclusions.

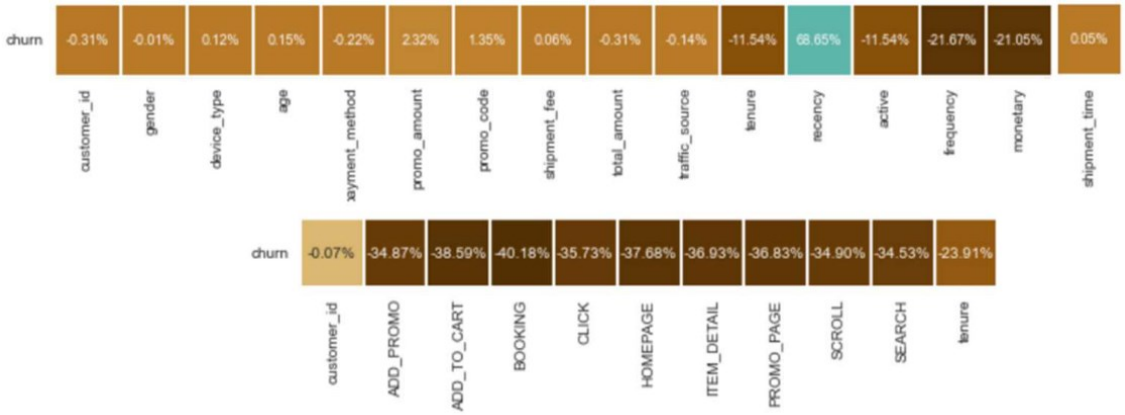

Fig. 16. Correlation test result

A classical assumption test examining Variance Inflation Factor (VIF) values was conducted to address potential multicollinearity and ensure model robustness. Multicollinearity, arising from high correlations between features, can negatively impact model performance by introducing redundancy and instability. By analyzing VIF values (Table 3), features exhibiting high multicollinearity can be identified and potentially removed, thereby refining feature selection for churn prediction. An analysis of VIF values reveals potential multicollinearity issues among the features "age," "active," "shipment_time," "ADD_TO_CART," "BOOKING," and "tenure," as their VIF scores exceed the threshold of 5. These elevated values indicate strong correlations with other variables within the dataset, potentially compromising model stability and interpretability. Conversely, "payment_method," "promo_code," "promo_amount," and "shipment_fee" exhibit VIF values ranging from 1 to 3.9. While this average level of multicollinearity is generally considered acceptable, it warrants further investigation. Nevertheless, in subsequent modelling, all ten features will be employed as independent variables to predict customer churn.

Table 3. VIF test result

Index	Feature	VIF
0	age	7.244217
1	payment_method	3.93064
2	promo_code	2.739201
3	promo_amount	2.883215
4	shipment_fee	1.824232
5	active	5.410465
6	shipment_time	4.584368
7	ADD_TO_CART	8.426836
8	BOOKING	9.032999
9	tenure	5.321233

In the modelling phase, the chosen features are combined into a single variable, typically denoted as "X." The churn label is then assigned its variable, often named "y" or "churn", for clarity. The data is split into two sets for model training: a training set (typically around 70%) used to build the model, and a testing set used to evaluate the model's performance on unseen data. Then, seven machine learning algorithms with detailed parameters are employed for model training, as given in Table 4. These algorithms are all common choices for classification tasks, which include churn prediction. They are chosen because they represent a variety of approaches and have a good track record of performance on telecommunication and banking industry datasets [38][39][40]. However, in this research, we only evaluate a single parameter option and do not perform hyperparameter tuning or cross-validation.

Table 4. Machine learning models and parameters

Algorithm	Parameters
Logistic Regression	Default
k-NN (k-Nearest Neighbor)	n_neighbors=10, metric='euclidean'
Random Forest	n_jobs=-1
ADA Boost (Adaptive Boosting)	Default
XG Boost	Default
Decision Tree	max_depth=4
Extra Tree (Extremely Randomised Trees)	Default

The results of the seven previous models were compared, and the best model was selected for churn prediction. After evaluating seven machine learning models for churn prediction, according to Table 5, the Random Forest model was likely chosen as the best model because it has the highest values for all the listed metrics (Accuracy, Precision, Recall, ROC AUC, F1-Score) compared to the other models.

Regarding the results, it is difficult to definitively determine the Random Forest model's performance concerning overfitting and generalizability since we only evaluated the model on a single dataset and did not perform parameter tuning. However, random forests are generally less prone to overfitting than models like decision trees [41][42]. This is because they average the predictions from many trees, which helps reduce the model's variance. Further research should be conducted using techniques like cross-validation to assess generalizability more effectively. This would involve

splitting the data into training and testing sets, training the model on the training data, and evaluating its performance on the unseen testing data.

Table 5. Model evaluation result

Algorithm	Accuracy	Precision	Recall	ROC AUC	F1-Score
Logistic Regression	0.96	0.96	0.74	0.74	0.81
k-NN (k-Nearest Neighbor)	0.92	0.81	0.55	0.54	0.56
Random Forest	0.99	0.97	0.99	0.98	0.97
ADA Boost (Adaptive Boosting)	0.98	0.93	0.94	0.94	0.94
XG Boost	0.99	0.95	0.97	0.97	0.96
Decision Tree	0.99	0.95	0.97	0.97	0.96
Extra Tree (Extremely Randomised Trees)	0.99	0.96	0.97	0.97	0.97

Compared to other research on churn analysis in the fashion industry, Granov [43] focuses on customer segmentation and classification (loyal, churner, returner) using K-prototypes clustering and logistic regression. While it does not directly use Random Forest, it provides a benchmark for churn prediction accuracy (0.98). This research employs logistic regression, but the Random Forest model outperforms it with higher performance metrics across all categories provided. Similarly, Karaarslan [44] utilizes machine learning to analyze customer churn in a fashion retail company. Their study explores various algorithms, including Gradient Boosting, which outperforms other models in their case. While this research employs logistic regression, the Random Forest model again surpasses it with superior performance metrics across all provided categories. From these comparisons, the Random Forest model demonstrates promising results. However, variations in datasets, model architectures (including parameterization), and evaluation metrics employed across studies hinder a more definitive comparison. Crucial factors influencing model performance include data size, class imbalance, and the problem domain.

Furthermore, while the Random Forest model demonstrated substantial predictive accuracy for customer churn based on the available historical data, it is essential to acknowledge a potential limitation: the model's performance with new users with limited historical data. New users typically exhibit fewer data points, impacting the model's ability to assess churn risk accurately. The Random Forest algorithm relies on numerous decision trees, each constructed based on various data attributes. When historical data is scarce, the model's predictive power diminishes as it lacks sufficient information to make reliable inferences. Consequently, the model might generate less accurate churn predictions for new users, potentially leading to suboptimal retention strategies

IV. Conclusions

The customer churn analysis on the Fashion E-commerce platform revealed a distinct customer profile, predominantly young women aged 20-29 residing in Jakarta. A notable surge in new customers coincided with the peak of the COVID-19 Delta variant in July 2021, suggesting a shift towards online shopping driven by lockdowns. Despite a majority female customer base, there was a noticeable preference for purchasing men's apparel. Regarding transactions, the promo code "AZ2022" was the most popular, with credit cards being the leading payment method, followed by Gopay and OVO. Transactions were most frequent late at night (10 PM to midnight) in Jakarta and West Java, with nearly 90% of users accessing the platform via mobile devices. Clicking was the most frequent activity, followed by adding promos, with user activity peaking on Sunday nights at 8 PM and dipping to its lowest at 7 AM.

This research makes two significant contributions. First, it provides deep insights into the demographics and behavioural patterns of the customer base. Second, it identifies ten key features that significantly impact churn: age, preferred payment methods, promo code usage, and purchase history. These insights were crucial in selecting the optimal churn prediction model, with the Random Forest algorithm emerging as the top performer in accuracy. The implications of these findings suggest that Fashion E-commerce can leverage the Random Forest model to design targeted strategies for reducing churn, such as offering competitive promotions for young female customers or enhancing the mobile user interface for a more user-friendly experience.

However, the study's limitations include static data, which may not fully capture dynamic changes in customer behaviour. For future research, it is recommended that the model be continuously updated with fresh data, more advanced machine learning techniques like deep learning be explored, and real-time monitoring of customer behaviour be implemented to enable early detection of churn risks and quicker interventions. By focusing on these future directions, Fashion E-commerce can enhance its churn prediction model, refine retention strategies, and achieve sustainable growth.

Declarations

Author contribution

All authors contributed equally as the main contributor of this paper. All authors read and approved the final paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

The authors declare no known conflict of financial interest or personal relationships that could have appeared to influence the work reported in this paper.

Additional information

Reprints and permission information are available at http://journal2.um.ac.id/index.php/keds.

References

[1] C. Stahl, N. Stein, and C. M. Flath, "Analytics applications in fashion supply chain management—A review of literature and practice," IEEE Trans. Eng. Manag., vol. 70, no. 4, pp. 1258–1282, 2021.

[2] E. Vezzetti, M. Alemanni, C. Balbo, and A. L. Guerra, "Big data analysis techniques for supporting product lifecycle management in the fashion industries," in Business Models and ICT Technologies for the Fashion Supply Chain: Proceedings of IT4Fashion 2017 and IT4Fashion 2018 7, Springer, 2019, pp. 25–34.

[3] E. S. Silva, H. Hassani, and D. Ø. Madsen, "Big Data in fashion: transforming the retail sector," J. Bus. Strategy, vol. 41, no. 4, pp. 21–27, 2020.

[4] D. Cirqueira, M. Hofer, D. Nedbal, M. Helfert, and M. Bezbradica, "Customer purchase behavior prediction in e-commerce: A conceptual framework and research agenda," in International workshop on new frontiers in mining complex patterns, Springer, 2019, pp. 119–136.

[5] S. Jaradat, N. Dokoohaki, H. J. C. Pampín, and R. Shirvany, "Workshop on recommender systems in fashion (fashionXrecsys2019)," in Proceedings of the 13th ACM Conference on Recommender Systems, 2019, pp. 552–553.

[6] S. Jaradat, N. Dokoohaki, H. J. C. Pampin, and R. Shirvany, "Third Workshop on Recommender Systems in Fashion-fashionXrecsys2021," in 15th ACM Conference on Recommender Systems (RECSYS), SEP 27-OCT 01, 2021, Amsterdam, NETHERLANDS, Association for Computing Machinery (ACM), 2021, pp. 810–812.

[7] M. A. Hakim and T. Terttiaavini, "Predictive Buyer Behavior Model as Customer Retention Optimization Strategy in E-commerce," INSYST J. Intell. Syst. Comput., vol. 6, no. 1, pp. 32–38, 2024.

[8] S. Khodabandehlou and R. M. Zivari, "Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior," J. Syst. Inf. Technol., vol. 19, no. 1/2, pp. 65–93, 2017.

[9] M. R. Khan, J. Manoj, A. Singh, and J. Blumenstock, "Behavioral modeling for churn prediction: Early indicators and accurate predictors of custom defection and loyalty," in 2015 IEEE International Congress on Big Data, IEEE, 2015, pp. 677–680.

[10] A. Tamaddoni, S. Stakhovych, and M. Ewing, "Comparing churn prediction techniques and assessing their performance: a contingent perspective," J. Serv. Res., vol. 19, no. 2, pp. 123–141, 2016.

[11] K. B. Agbemadon, R. Couturier, and D. Laiymani, "Churn detection using machine learning in the retail industry," in 2022 2nd International Conference on Computer, Control and Robotics (ICCCR), IEEE, 2022, pp. 172–178.

[12] Prithi Madhavan and K. Tamizharasi, "Churn Detection Using Machine Learning in the Retail Industry," Int. J. Eng. Technol. Manag. Sci., vol. 7, no. 1, pp. 344–349, 2023.

[13] S. R. Labhsetwar, "Predictive analysis of customer churn in telecom industry using supervised learning," ICTACT J. Soft Comput., vol. 10, no. 2, pp. 2054–2060, 2020.

[14] S. De, P. Prabu, and J. Paulose, "Application of machine learning in customer churn prediction," in 2021 Innovations in Power and Advanced Computing Technologies (i-PACT), IEEE, 2021, pp. 1–7.

[15] S. De, P. P, and J. Paulose, "Effective ML Techniques to Predict Customer Churn," in 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE, Sep. 2021, pp. 895–902.

[16] A. A. Chang et al., "Fashion trend forecasting using machine learning techniques: a review," Data Sci. Intell. Syst. Proc. 5th Comput. Methods Syst. Softw. 2021, Vol. 2, pp. 34–44, 2021.

[17] S. N. Günesen, N. Şen, N. Yıldırım, and T. Kaya, "Customer churn prediction in FMCG sector using machine learning applications," in IFIP International Workshop on Artificial Intelligence for Knowledge Management, Springer, 2021, pp. 82–103.
[18] B. Prabadevi, R. Shalini, and B. R. Kavitha, "Customer churning analysis using machine learning algorithms," Int. J. Intell. Networks, vol. 4, pp. 145–154, 2023.
[19] J. Shobana, C. Gangadhar, R. K. Arora, P. N. Renjith, J. Bamini, and Y. D. Chincholkar, "E-commerce customer churn prevention using machine learning-based business intelligence strategy," Meas. Sensors, vol. 27, p. 100728, 2023.
[20] P. R. Srivastava and P. Eachempati, "Intelligent employee retention system for attrition rate analysis and churn prediction: An ensemble machine learning and multi-criteria decision-making approach," J. Glob. Inf. Manag., vol. 29, no. 6, pp. 1–29, 2021.
[21] P. Nagaraj, V. Muneeswaran, A. Dharanidharan, M. Aakash, K. Balananthanan, and C. Rajkumar, "E-Commerce Customer Churn Prediction Scheme Based on Customer Behaviour Using Machine Learning," in 2023 International Conference on Computer Communication and Informatics (ICCCI), IEEE, 2023, pp. 1–6.
[22] R. Sharma, "Customer Churn Analysis in Telecom Industry using Logistics Regression in Machine Learning with Kaplan–Meier and Cox Proportional Hazards Model," INTERANTIONAL J. Sci. Res. Eng. Manag., vol. 08, no. 04, pp. 1–5, Apr. 2024.
[23] F. A. Mohamed and A. K. Al-Khalifa, "A review of machine learning methods for predicting churn in the telecom sector," in 2023 International Conference On Cyber Management And Engineering (CyMaEn), IEEE, 2023, pp. 164–170.
[24] S. Zhao, "Customer Churn Prediction Based on the Decision Tree and Random Forest Model," BCP Bus. Manag., vol. 44, pp. 339–344, Apr. 2023.
[25] H. W. Chow, Z. J. Lim, and S. Alam, "Data-driven Runway Occupancy Time Prediction using Decision Trees," in 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), IEEE, 2021, pp. 1–9.
[26] U. Pujianto, M. I. Akbar, N. T. Lassela, and D. Sutaji, "The Effect of Resampling on Classifier Performance: an Empirical Study.," Knowl. Eng. Data Sci., vol. 5, no. 1, pp. 87–100, 2022.
[27] J. Rohith and P. U. Priyadarsini, "Classification And Prediction Of Chronic Kidney Disease Using Novel Decision Tree Algorithm By Comparing Random Forest For Obtaining Better Accuracy," Cardiometry, no. 25, pp. 1800–1807, 2022.
[28] N. S. F. Putri, A. P. Wibawa, H. A. Rosyid, A. B. P. Utama, and W. Uriu, "Performance of Ensemble Classification for Agricultural and Biological Science Journals with Scopus Index," Knowl. Eng. Data Sci., vol. 5, no. 2, p. 137, Dec. 2022.
[29] T. Asra, A. Setiadi, M. Safudin, E. W. Lestari, N. Hardi, and D. P. Alamsyah, "Implementation of AdaBoost Algorithm in Prediction of Chronic Kidney Disease," in 2021 7th International Conference on Engineering, Applied Sciences and Technology (ICEAST), IEEE, 2021, pp. 264–268.
[30] P. Jeyaprakaash and K. Sashirekha, "Accuracy measure of customer churn prediction in telecom industry using Adaboost over Decision Tree algorithm," J. Pharm. Negat. Results, pp. 1495–1503, 2022.
[31] I. Hanif, "Implementing extreme gradient boosting (xgboost) classifier to improve customer churn prediction." 2020.
[32] E. al Abhinav S. Thorat, "Optimizing Churn Identification in Telecommunications Using Natural Language Processing and XG Boost Machine Learning Paradigm," Int. J. Recent Innov. Trends Comput. Commun., vol. 11, no. 9, pp. 4226–4232, Nov. 2023.
[33] P. G. Priyanka and S. A. Rahaman, "Customer Churn Prediction in Telecom Industry Using Regression Algorithms," Int. J. Innov. Res. Comput. Sci. Technol., vol. 10, no. 3, pp. 54–57, 2022.
[34] A. Dijendra and J. Sisodia, "Telecomm Churn Prediction Using Fundamental Classifiers To Identify Cumulative Probability," in 2021 International Conference on Communication information and Computing Technology (ICCICT), IEEE, 2021, pp. 1–5.
[35] F. E. Usman-Hamza et al., "Intelligent decision forest models for customer churn prediction," Appl. Sci., vol. 12, no. 16, p. 8270, 2022.
[36] C. Rungruang, P. Riyapan, A. Intarasit, K. Chuarkham, and J. Muangprathub, "RFM model customer segmentation based on hierarchical approach using FCA," Expert Syst. Appl., vol. 237, p. 121449, 2024.
[37] C. Shearer, "The CRISP-DM model: the new blueprint for data mining," J. data Warehous., vol. 5, no. 4, pp. 13–22, 2000.
[38] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," Computing, vol. 104, no. 2, pp. 271–294, 2022.
[39] H. Jain, A. Khunteta, and S. Srivastava, "Telecom churn prediction and used techniques, datasets and performance measures: a review," Telecommun. Syst., vol. 76, pp. 613–630, 2021.
[40] K. G. M. Karvana, S. Yazid, A. Syalim, and P. Mursanto, "Customer churn analysis and prediction using data mining models in banking industry," in 2019 international workshop on big data and information security (IWBIS), IEEE, 2019, pp. 33–38.
[41] L. Breiman, "Random forests Mach Learn 45 (1): 5–32." ed, 2001.
[42] S. Hussain and G. C. Hazarika, "Educational Data Mining Model Using Rattle," Int. J. Adv. Comput. Sci. Appl., vol. 5, no. 6, 2014.
[43] A. Granov, "Customer loyalty, return and churn prediction through machine learning methods: for a Swedish fashion and e-commerce company." 2021.
[44] H. Karaarslan, M. Baştuğ, C. G. Şen, and E. E. Işik, "A Comparative Study on Customer Churn Analysis Using Machine Learning and Data Enrichment Techniques," J. Soft Comput. Decis. Anal., vol. 2, no. 1, pp. 225–235, Jun. 2024.