E-Commerce Customer Churn Prediction for the Marketplace in Malaysia

Tang Hui Ying[1], Suraya Ya'acob[2]

Razak Faculty of Technology and Informatics Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Malaysia

tang.ying@graduate.utm.my[1], suraya.yaacob@utm.my[2]

Abstract

Customer churn, or the loss of customers, is a significant challenge for e-commerce businesses, as it leads to revenue loss and increased marketing costs. Most online marketplaces have insufficient awareness of the e-commerce customer churn rate that can be analysed before conducting customer retention activities or management. It is crucial for an online marketplace to handle customer churn as it is more high-value to manage existing customers than to have new customers on the platform. This study uses machine learning algorithms to predict e-commerce customer churn. This study has three (3) goals that must be met, which include identifying the attributes with high association to e-commerce customer churn for the online marketplace in Malaysia, constructing e-commerce customer churn prediction model with the important attributes by using the machine learning techniques, and evaluating and comparing the performance of the e-commerce customer churn prediction models using evaluation metrics. The machine learning techniques chosen based on the literature review are Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Logistics Regression (LR), K-Nearest Neighbours (KNN), and eXtreme Gradient Boosting (XGB). The dataset involved in the study was compiled from a leading e-commerce platform in Malaysia – online marketplace ABC, and it contains customer demographic information, purchase history etc. The dataset and model uses Python as the analytics tool and extended CRISP-DM with TDSP as the methodology. There are four (4) evaluation metrics to compare and evaluate the performance of the prediction models, including accuracy, precision, recall, and F1-score. The desired outcome shows that machine learning can predict Malaysian e-commerce customer churn. This study can help Malaysian e-commerce companies identify at-risk customers and retain them. Additionally, this study can inform e-commerce customer churn prediction research.

Keywords: e-commerce, marketplace, customer churn, prediction, machine learning

1. Introduction

The e-commerce marketplace, also known as online e-commerce marketing, is a venue or website where customers may access a wide range of product brands shown alongside those from other vendors, businesses, or people. The marketplace owner attracts customers and conducts transactions while third-party providers handle manufacturing and shipment.[1] Online marketplaces speed up production by enabling manufacturers to sell directly to consumers, avoiding wasteful stock-holding methods. Some of the popular online marketplaces in Malaysia include Lazada, Shopee, and 11street. These marketplaces provide a diverse selection of

---

* Corresponding author. suraya.yaacob@utm.my

items, including but not limited to fashion, electronics, and household items. In addition to that, the ever-increasing number of people who utilise mobile devices in the country has also helped drive the growth of online shopping. [2] Despite the growth, the online marketplace in Malaysia faces challenges such as competition from established brick-and-mortar retailers and the need to build trust among consumers. The industry also faces challenges related to logistics, delivery, and payment methods. To overcome these challenges, e-commerce companies in Malaysia are focusing on offering a superior customer experience and building strong customer relationships. [3] Generally, the online marketplace in Malaysia is projected to continue to expand in the future as more and more consumers embrace e-commerce as a way to shop and as the government continues to support its development.

Malaysia is an appealing Southeast Asian e-commerce market due to its varied economy and sophisticated digital technology infrastructure. In 2021, 80 percent of Malaysia's population (27.4 million) was active internet users, and mobile phone penetration was high (84.2 percent).[4] As of January 2021, Malaysia has 28 million social media users and 39.99 million mobile connections.[1] Additionally, e-commerce has seen significant growth in Malaysia in recent years. With online marketplaces are becoming a popular choice for consumers in Malaysia, especially during the pandemic. In 2020, the e-commerce industry in Malaysia was valued at over $2 billion, and it is projected to reach $3.6 billion by 2025. [5] The Malaysian government has also been supportive of the growth of e-commerce, as they see it as a way to drive economic growth and create new job opportunities.

Nevertheless, the high competition in the e-commerce industry has led to a significant challenge for online businesses - customer churn. Customer churn, also known as customer attrition or customer loss, may result in significant revenue loss and higher marketing expenditures for e-commerce enterprises.[6] In the case of an online marketplace, churn happens when a customer discontinues the use of a product or service. In Malaysia, the average customer churn rate for e-commerce businesses is around 20%. [4] This highlights the need for effective retention strategies to reduce customer churn. The practise of identifying consumers who are likely to discontinue using a product or service in the near future is known as churn prediction. To address this challenge, this research aims to use machine learning techniques to predict customer churn in the e-commerce marketplace in Malaysia. The study employs a combination of supervised learning algorithms, including decision tree (DT), random forest (RF), Naïve Bayes (NB), logistics regression (LR), K-Nearest Neighbours (KNN), and eXtreme Gradient Boosting (XGB) to predict customer churn. The dataset used consist of e-commerce customer data collected from a popular online marketplace in Malaysia.

2. Problem Background

Online marketplaces rely on a large customer base to drive revenue and growth. When customers leave, they take their purchasing power with them, and the marketplace loses out on future revenue. [7] Additionally, it can be costly to acquire new customers, so retaining existing customers is crucial to the success of an online marketplace. As a result, forecasting customer churn is critical for online marketplaces because it helps them to identify at-risk clients early and take efforts to keep them, such as delivering special incentives or specialised customer

support.[8] By identifying and retaining at-risk customers, online marketplaces can reduce the impact of churn on their revenue and growth.

In addition to customer retention, customer churn prediction can also help online marketplaces to identify the factors that influence customer behaviour, such as product prices, product quality, and customer service, which can help them improve their offerings and customer service. [9] In summary, customer churn prediction for an online marketplace is important because it enables the marketplace to identify and retain at-risk customers, reducing the impact of churn on their revenue and growth and improving their offerings and customer service.

Furthermore, the ability to predict customer churn in the e-commerce marketplace in Malaysia would be a valuable tool for online businesses. Predictive models may help firms identify at-risk clients and take proactive efforts to keep them.[8] However, predicting customer turnover in e-commerce is a difficult problem, and previous research has indicated that typical approaches, such as logistic regression and decision trees, may not be enough. [7] In light of these challenges, there is a need for research that employs machine learning techniques to predict customer churn in the e-commerce marketplace in Malaysia. The application of these techniques can provide a more accurate and reliable prediction of customer churn and inform the development of retention strategies for e-commerce businesses in Malaysia.

The gaps in research on e-commerce customer churn prediction include the lack of comprehensive data on customer behaviour, as the e-commerce industry is different across different countries and regions. [10] Cultural and geographic differences may affect customer behaviour and may not be captured by existing models; giving the quality and quantity of data available for customer churn prediction may not be adequate. [11] This can affect the accuracy of the prediction models. Besides limited access to customer data due to privacy concerns, researchers struggled to gain access to customer data that is needed to build effective prediction models. Moreover, the need for more advanced machine learning techniques to accurately predict customer churn or the complexity of the prediction models used may not be suitable for the e-commerce industry. It may not effectively capture the diverse and dynamic customer behaviour in this sector. Additionally, there is a need for more cross-industry studies to understand the generalizability of findings and to develop more robust models that can be applied across different e-commerce platforms and industries. [12] Predictive models may be biased towards certain customer segments, leading to inaccurate results and may not be integrated into the day-to-day business processes of e-commerce companies, making it difficult for companies to act on the results of the predictions, as well as may not be clear what metrics should be used to evaluate the success of customer churn prediction models.

The proposed model focuses online marketplace in Malaysia as there is no or minimal paper study on e-commerce customer churn in Malaysia; by understanding Malaysian consumer behaviour and focusing on seller-view, which is from SME, MSME. The effectiveness of the prediction model's performance are assessed using a range of metrics, including accuracy, precision, recall, and F1-score. The findings from this study offer valuable understanding regarding the elements that contribute to customer churn within the e-commerce market in Malaysia and inform the development of retention strategies for e-commerce businesses in the country.

3. Literature Reviews

Raeisi and Sajedi attempt to obtain an understanding of a data set offered by the most prominent online food ordering business in Tehran, Iran. Six data mining techniques, including k-NN, Naïve Bayes, Decision Tree, Random Forest, Gradient Boosted Tree, and Rule Induction, were utilised to forecast whether or not a user would discontinue using the service in view of the online properties and user behaviour. The findings demonstrate that Gradient Boosted Trees (gradient boosting employs two distinct models: A "weak" machine learning model, usually a decision tree) outperform other approaches with an accuracy of 86.90%. Besides, there is a study addressing the challenge of high-precision customer churn prediction using a single model. The combined model provides improved interpretability like the decision tree model and higher prediction accuracy like the neural network model. [13]

Besides, the study by Zhuang showed the accuracy of predictions is improved through consumer segmentation, and the XG-Boost algorithm is better than other algorithms, with an accuracy of 88%. Furthermore, Ait Daoud et al. proposed a customer attrition model with k-means approach, including the clustering step, in which the dependent variable is categorised as non-churn, partial churn and churn. As a result, the effectiveness of the decision tree ensemble model in detecting both partial churn and complete churn in comparison to other models. According to Gan's study, the XGBoost algorithm may successfully anticipate e-commerce customer loss and then give a decision-making reference for e-commerce firms' customer care strategy. [14]

In a study on a dataset from Kaggle for a leading e-commerce platform, the researcher uses the synthetic minority oversampling method (SMOTE) to equalize the intensity of the Churn column by creating clean sampling of the minority subclass (Churned) based on the nearest neighbours. The accuracy and kappa metrics of 93.5% and 0.75, respectively, indicate that Random Forest is the best-suited machine learning method [7]. Another study shows that RF performed with AUC = 0.7162 compared to DT and LR. The finding is that the chance of churning for Paradox whales seems predictable. More data and fine-tuning the concept of churn might help to enhance the model even more. [15] In the study by ÇAlli and Kasim, different machine learning algorithms were used to analyze ten features and determine the best algorithm for predicting customer churn. The random forest algorithm was found to be the most effective in predicting customer churn in the SaaS industry. [16]

On the other hand, Xia et al. developed a two-layer fusion model based on stacking integration to increase the accuracy of online shopping customer attrition prediction. The experimental findings reveal that a two-tier fusion model may significantly increase the accuracy of customer loss prediction for online shopping and that an upgraded data set can substantially improve the prediction impact of the two-tier model giving the best accuracy of 93.72% using logistic regression. Other than that, a study shows the established machine learning classifiers can predict learner churning behaviour with a high recall value of 0.824 and F1-score of 0.778 by using LR, which is better than RF and NB. This churn detection allows app developers to intervene in learner retention. [17]

Based on the literature review done, a summary of machine learning techniques used in the previous works is shown in Table 1 and Table 2. Besides, Figure 1 show the literature map on the related definitions and concepts on e-commerce customer churn prediction model.

Table 1. Comparison of the Related Works on the Attributes used in the e-Commerce Customer Churn Prediction Model

Attributes	Related Works on e-Commerce Customer Churn Prediction Model*										Total Count
	1	2	3	4	5	6	7	8	9	10	
Amount of consumption (price)	/		/	/	/	/	/	/	/		8
Number of days since last purchase	/	/	/		/	/	/	/	/		8
Frequency of purchasing	/		/		/	/	/	/	/		7
Enjoyment/ratings		/	/	/	/	/	/		/		7
Total amount of product	/	/	/		/	/		/			6
Number of time periods of first to last purchase	/		/		/		/	/			5
Customer group/categories	/							/		/	3
Payment method		/	/		/						3
Area/location		/	/							/	3
Gender		/	/	/							3
Number of visits/duration			/							/	2
Product category				/						/	2
Devices used			/		/						2
Discount/cashback			/	/							2
Age		/		/							2
Estimated salary		/		/							2
Credit balance		/									1
Marital status			/								1

Table 2. Comparison of the Related Works on the Machine Learning Techniques Used in the e-Commerce Customer Churn Prediction Model

Machine Learning Techniques	Related Works on e-Commerce Customer Churn Prediction Model*										Total Count
	1	2	3	4	5	6	7	8	9	10	
Logistic Regression		/	/	/		/	/	/		/	7
Random Forest	/			/	/		/	/		/	6
Decision Tree	/	/		/			/	/			5
XG-Boost Tree			/		/	/		/		/	5
K-NN	/			/		/				/	4
Naïve Bayes	/			/		/			/		4
Support Vector Machines		/	/							/	3
Gradient Boosted Trees	/									/	2

|									
BP Neural Network		/		/					2
ANN			/		/				2
Rule Induction	/								1
Multi-Layer Perceptron		/							1
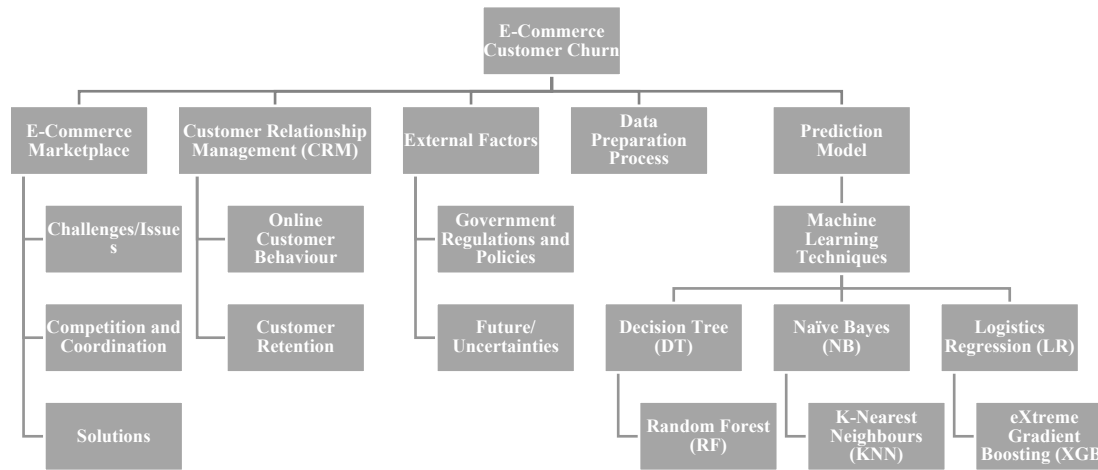
Figure 1. Literature Map

4. Research Methodology

This paper discussed the first research question which is "What are the important attributes that are highly associated to e-commerce customer churn for the online marketplace in Malaysia?" and relating to the question with the research objective of "To identify the attributes with high association to e-commerce customer churn for the online marketplace in Malaysia." together with selection of the machine learning techniques from the related papers. As show in section 2 on the literature reviews, the studies the related attributes, and modelling for e-commerce customer churn prediction were from Scopus, ScienceDirect, IEEE, Google Scholar etc. The initial screening of the related/relevant papers were based on titles and abstracts, followed by a full-text review. In addition, complete texts were collected via institutional access and assessed for relevance and quality. Furthermore, the data about the author, publication year, study technique, significant results, and limitations was extracted. The literature based on common themes and examined trends, controversies, and gaps were categorised. Each study's methodology, theoretical frameworks, and contribution were examined critically. A conceptual model was created by synthesising fundamental ideas and linkages. The output from this method resulted with the identification about related attributes, and modelling used in the e-commerce industry on customer churn and customer retention. As showed in Figure 1 on the literature map, the related keywords in the figure were taken into consideration in selecting the journals, papers etc. and were taken from the past five years which is from 2018 to 2022. The summary of the attributes and machine learning techniques showed in Table 1 and Table 2.

Figure 2 depicts the overall research design and procedural flowchart of the study. The flowchart explained the research's phases/stages, beginning with defining the

issue and case study. Data collecting from the online marketplace ABC was followed by development and experimentation, testing and validation, result analysis, assessment, and discussion, and documentation.
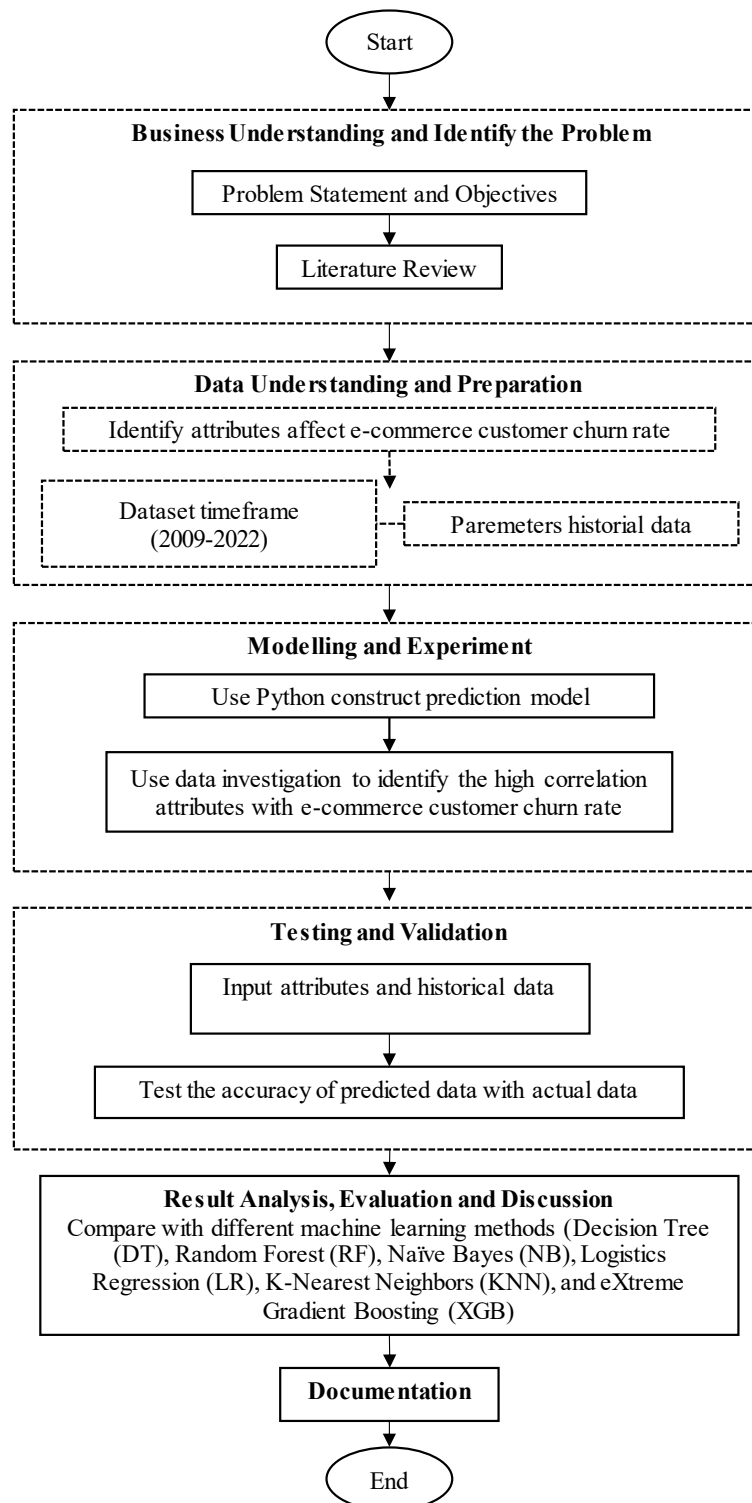
Figure 2. Overall Research Design and Procedure Flowchart

5. Conclusion and Expected Outcome

The paper mostly representing the initial stages to partially of the mid-stages of the prediction models for the e-commerce customer churn on seller view. However, this study constructs prediction models for e-commerce customer churn on seller view. Allowing online marketplace to retain the existing customers (sellers) who are more high-value than new customers. Actual sample data is gathered from the ABC online marketplace. This research also employs an improved data mining technique (CRISP-DM with TDSP), making it more user-friendly to the business, particularly in terms of visualisation.This may assist Malaysian online marketplaces in taking proactive actions to retain these clients and decrease churn rates. The output may also include a better knowledge of the primary elements that lead to customer turnover, which may be used to develop strategies for lowering churn and enhancing customer retention. A prediction model or algorithm that can continually monitor and forecast customer turnover in real time may also be included in the result.

On the other side, this initiative may help firms better understand Malaysian online buying behaviour and how to increase client retention.Aside from that, this study may assist companies in determining the most efficient machine learning approach for customer churn prediction, which may be useful for other online marketplace enterprises or e-commerce in Malaysia and other countries with comparable characteristics. This research might also give a comparison of the models' performance and accuracy by using different machine learning approaches. Finally, it might be a useful resource for academics and practitioners working in the fields of e-commerce, machine learning, and customer churn prediction.

Acknowledgments

References

[1]   D. M. Selvi and M. G. Md Johar, "Leveraging the Use of Online Marketplace for SMEs in Malaysia: A Review," *Journal of Management and Science,* vol. 20, 1, 2022.

[2]   Economic Planning Unit, Prime Minister's Department. (2021). *Malaysia Digital Economy Blueprint*.

[3]   Ts *et al.*, *E-Government and E-Commerce Issues in Malaysia*. 2021.

[4]   T. I. T. Administration. Malaysia - Country Commercial Guide [Online] Available: https://www.trade.gov/country-commercial-guides/malaysia-ecommerce

[5]   S. Y. Tham, "The Digital Free Trade Zone (DFTZ): Putting Malaysia's SMEs onto the Digital Silk Road ", no. 2018 No. 17.

[6]   N. Jha, D. Parekh, M. Mouhoub, and V. Makkar, "Customer Segmentation and Churn Prediction in Online Retail," 2020, pp. 328-334.

[7]   A. Alshamsi, "Customer Churn prediction in E-Commerce Sector " Professional Studies (MS) Master's Project, Graduate Programs & Research (Dubai), Rochester Institute of Technology, 2022. [Online]. Available: https://scholarworks.rit.edu/theses/11183

[8]   Y. Zhuang, "Research on E-commerce Customer Churn Prediction Based on Improved Value Model and XG-Boost Algorithm," 2018.

[9]   F. Guo, "Research on Customer Relationship Management in E-commerce Environment," presented at the 8th Annual International Conference on Geo-Spatial Knowledge and Intelligence 18-19 December 2020, Xi'an, Shaanxi, China, 2021.

[10]   R. Ait daoud, A. Amine, B. Belaid, and R. L. Rachid Lbibb, "Clustering Prediction Techniques in Defining and Predicting Customers Defection: The Case of E-Commerce Context," *International Journal of Electrical and Computer Engineering,* vol. 8, 08/01 2018, doi: 10.11591/ijece.v8i4.pp2367-2383.

[11]   N. A. Abdul Hamid, C. H. Cheun, N. H. Abdullah, M. F. Ahmad, and Y. Ngadiman, "Does Persuasive E-commerce Website Influence Users' Acceptance and Online Buying Behaviour? The Findings of the Largest E-commerce Website in Malaysia," in *ICT for a Better Life and a Better World*, (Lecture Notes in Information Systems and Organisation, 2019, ch. Chapter 17, pp. 263-279.

[12]   X. Du, B. Liu, and J. Zhang, "Application of Business Intelligence Based on Big Data in E-commerce Data Analysis," *Journal of Physics: Conference Series,* vol. 1395, p. 012011, 11/01 2019, doi: 10.1088/1742-6596/1395/1/012011.

[13]   X. Hu, Y. Yang, L. Chen, and S. Zhu, "Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network," *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA),* pp. 129-132, 2020.

[14]   L. Gan, "XGBoost-Based E-Commerce Customer Loss Prediction," *Computational Intelligence and Neuroscience,* vol. 2022, p. 1858300, 2022/07/31 2022, doi: 10.1155/2022/1858300.

[15]   V. Tryggvadottir, "Customer Churn Prediction for PC Games : Probability of churn predicted for big-spenders using supervised machine learning," Independent thesis Advanced level (degree of Master (Two Years)) Student thesis, TRITA-SCI-GRU, 2019:254, 2019. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-254198

[16]   L. ÇAlli and S. Kasim, "Using Machine Learning Algorithms to Analyze Customer Churn in the Software as a Service (SaaS) Industry," *Academic Platform Journal of Engineering and Smart Systems,* vol. 10, no. 3, pp. 115-123, 2022, doi: 10.21541/apjess.1139862.

[17]   D. Singh, R. Pathan, G. Banerjee, and R. Rajendran, *From Hello to Bye-Bye: Churn Prediction in English Language Learning App*. 2021.