**E-commerce Churn Prediction by using machine learning**

Student ID: 22089776

Student name: Amulya Panithi

GitHub Link: https://github.com/Amulya07534776064/Project-Work

Table of Contents

[Chapter 1: Introduction 3](#_Toc181266494)

[1.1: Background of the study 3](#_Toc181266495)

[1.2: Aim and objectives and research questions 3](#_Toc181266496)

[1.3: Rationale of the study 3](#_Toc181266497)

[1.4: Scope of the research 3](#_Toc181266498)

[1.5: Structure of the research 3](#_Toc181266499)

[Chapter 2: Literature review 4](#_Toc181266500)

[2.1: Introduction 4](#_Toc181266501)

[2.2: Customer churn in ecommerce: Reasons and consequences 4](#_Toc181266502)

[2.3: Machine learning and customer churn prediction 5](#_Toc181266503)

[2.4: Summary 7](#_Toc181266504)

[References 8](#_Toc181266505)

# Chapter 1: Introduction

1.1: Background of the study:

The data suggest that ecommerce sector’s penetration into the UK market will continue to rise. By 2029, the market share of ecommerce sector will be increased by 15.20 percent (Statista, 2029). Despite of such growth, the concern of customer churn is high for the ecommerce companies. The market size is growing but the developing the loyal customer base is yet high challenge for the ecommerce players. The online sales are growing and ecommerce market generated the revenue of whopping 2089 billion pounds in the year 2021 (SwiftERM, 2022). Therefore, it become clear that the growth opportunities are high but the problem of customer churn can take a toll on the profit margins and sales of ecommerce companies.

This particular issue needs to be addressed and for the same purpose, the technology-based solutions are introduced. The machine learning is the most prominent technology which has the potential to change the way companies understand its customers and design strategies to retain them. The strategic approach in customer relationship is that the companies must invest in high retention rather than acquisition (Akula, 2024). Artificial intelligence can enable the brands to become customer centric and understanding their behaviour. Machine learning can provide better solutions to do customer profiling and gaining insights from huge chunk of data. Overall, the identified problem was customer churn and identified solution was machine learning; so, the current research has focused on developing machine learning-based system to predict customer churn in ecommerce sector.

1.2: Aim and objectives and research questions:

*Aim*:

The project aimed at using machine learning to predict customer churn in ecommerce sector.

*Objectives*:

* To identify the factors causing customer churn in ecommerce
* To understand the negative consequences of customer churn for ecommerce
* To perform EDA in customer churn dataset in ecommerce
* To develop machine learning model to predict customer churn
* To test and evaluate the performance of applied algorithms

*Research question*:

* What factors cause customer churn in ecommerce sector?
* How machine learning help in predicting customer churn in ecommerce sector?

1.3: Rationale of the study:

The need for the customer churn prediction system was realised due to the high rate of customer leaving in ecommerce sector. It causes several negative consequences to the companies. The pressure on profits can be observed. The rationale of the study was to understand the reasons behind the customer churn and then putting such reasons as major data points for the model training. After this study, the future researches can focus on ideal datasets or variables which must be utilised to predict the customer churn specifically for ecommerce industry. Apart from this, the rationale of the study also deals with improving the managerial implications with respect to strengthening the customer relationship management. The machine learning model can not only inform about the possibilities of churn but the reasons can also be categorised. On the basis of such information, the managers can frame better strategies to keep customers satisfied and retention rate can become robust.

1.4: Scope of the research:

The scope of the current research revolved around the inclusion of multiple machine learning algorithm. This becomes important factor because each algorithms serve different purpose and have different characteristics. The readers can get to know about the pros and cons of algorithms and accordingly, the implementation in future tasks can be determined. Further, the scope of the study also includes the discussion around customer behaviour. It can help in understanding the customer mindset and their satisfaction or dissatisfaction parameters can be recorded.

1.5: Structure of the research:

The structure was made on the standard process which included the chapters like introduction, literature review, research method, data analysis and conclusion and recommendation chapter. The aims and objectives were clearly defined in introduction chapter and it also provided information about the rationale and scope of the research. The accumulation of past studies was necessary to get insights on the topic; so, the literature review chapter was documented. It has sub-themes which aligned with the core objectives. Afterwards, the third chapter of research was methodology and it has justification for the research type, the reasons to select any algorithm. The experimentation steps like data collection, data split, data pre-processing, feature selection, model training and even evaluation all were described in this chapter. The experimentation results were presented in fourth chapter and in the same chapter, the discussion was also made. The final chapter was the conclusion and recommendation which also has future scope and brief reflection of the author’s learning.

# Chapter 2: Literature review

2.1: Introduction:

The literature review chapter was prepared to get insights about the discussion made in previous works regarding the customer churn and application of machine learning. The scope of the current literature review included the understanding of reasons behind churn and consequences of churn and most importantly, the algorithms used in past works were also presented.

2.2: Customer churn in ecommerce: Reasons and consequences:

Low customer churn rate is essential for the companies especially to maintain profit margins. It was found in the work of Shobana et al. (2023) that the customers shift to other brands when industry competition is high. The customers tend to look for new offers and deals which put excessive pressure on the promotional campaigns of ecommerce firms. If the brand is not able to meet the customer expectations, then the customer churn can be observed. Due to high competition, the customers regularly get lucrative offers which affect their loyalty levels. Therefore, the customer churn must be prevented because it can help in reducing the customer acquisition cost. When brands have retained customers, then its money and power and other resources get utilised to achieve other goals rather than acquiring new customers.

Robust customer relationship is one way to prevent customer churn. The work done by Matuszelanski and Kopczewska (2022) revealed that the loyal customers show less intent of leaving the brand or shifting to another telecom players. However, in the absence of loyalty factors, the customer behaviour become highly uncertain. Their assessment of value buying or quality services can be changed on the regular basis and then meeting their demands become difficult task. It was found that the if the customer remain dissatisfied, then also the customer churn rate remain high. The study stressed on the fact that customer churn prediction tools and technologies like machine learning must be deployed to control the churning rate.

The poor handling of customer churn indicates towards the poor implementation of marketing strategies by the companies. The work done by Xiahou and Harada (2022) stressed that when churn is not controlled by appropriate measures then the shopping behaviour of existing customers gets changed and they become more critical of brand’s positioning and its products or services. The study confirmed that the major negative consequence of high customer churn is that the profit margins of the companies get declined. The author considered customers as major asset or resources to generate revenue. So, it is necessary that the customers desires should be understood and steps to prevent customer churn must be taken. As per the author, the major challenge for ecommerce firms is that they do not have direct relationship or interaction with the customers; so, it is hard to find that when churn can happen and why exactly churn ratio increases. Though ecommerce firms have ample amount of data about the customer shopping patterns so with the help of advanced techniques like machine learning the prediction about the customer churn can be made.

2.3: Machine learning and customer churn prediction:

Through early prediction of customer churn, the robust retention strategies can be framed. Prabadevi, Shalini and Kavitha (2023) used multiple algorithms to predict the customer churn. In this paper, the data was extracted from Kaggle which has 7044 alternatives and 21 attributes like phone service, tenure, multiple lines, internet service, online backup, online security, device protection, tech support, streaming TV, streaming movies and contract. The analysis of dataset variables suggested that it helped in understanding the customer behaviour in different ways and the goal of the paper, to predict the customer churn, was achieved. The author applied the algorithms like stochastic gradient boosting, random forest, logistic regression and KNN methods. Out of these algorithms, the best performing model was stochastic gradient boosting method which has the accuracy of 83.9 percent. As per the author, this model performed well in churn prediction because it can handle bias and it uses the delta rule in training process.

Khattak et al. (2023) considered customer churn as major financial loss for the brands. This paper utilised the deep learning models in predicting the customer churn. The dataset was picked from the telecom sector and it has 7033 rows and 20 different features. The variables in the dataset were online backup, monthly charges, total charges, paperless billing, payment method, streaming TV, streaming movies and tech support. These factors were helpful in gaining insight about the shopping patterns of the customers and accordingly the customer churn was predicted and this data was helpful in achieving the goal. The author applied the hybrid model where bi-directional LSTM model was applied along with CNN model. It was found that the proposed model achieved the accuracy of 81 percent in predicting the customer churn. This study picked data from telecom sector and size of the data was 7033. The data split was done in the ratio of 80 and 20 percent which is conventional ratio of training and testing phases. The model performed well because it has layers like convolutional, pooling and flatten layer. The accuracy of BI LSTM was high because the training process include backward and forward processing of given inputs.

One of the advantages with machine learning is that it has variety of algorithms dealing with different types of problems. The supervised method, unsupervised models and neural networks along with ensemble techniques all have different characteristics. Neural networks fall into the category of deep learning methods. Kim and Lee (2022) used machine learning to predict customer churn in influencer commerce in Korea. The dataset was picked for the period of 2018 to 2020 and models like decision tree was applied to predict the customer churn. The dataset has the variables like number of purchases, number of refunds, average payment, average refund and the number of influencers was 510 and number of customers was 100213. The goals were achieved from this dataset because it has clearly explained about the refund and payment history of the customers which become the basis of predicting future churn rates. It was the supervised learning algorithm. On the basis of experimentation, it was found that the model achieved the algorithms of 90 percent and predicted the churn with higher efficiency.

Olaniyi et al. (2020) focused study on the customer churn Nigerian banking sector. The author took the data of 200 customers and it has 5 major attributes. The size of the dataset was small in this paper and the applied algorithms were the support vector machine and KNN. The clustering method was applied on the data with the help of KNN model and then with the help of support vector machine model, the prediction process was done. Support vector machine model has different kernels and in the current study, the author has focused on application of polynomial kernels. In this paper, the model evaluation was done on the basis of accuracy levels, sensitivity, specificity and precision scores. On the basis of evaluation measures, it was found that the accuracy score for the Support vector machine model was 97 percent. It gives an edge to banking sector to understand the mindset of customers and timely actions can be taken to prevent any type of customer churn. Despite of small dataset, the algorithm achieved higher accuracy because support vector machine provides better results when dataset is limited. However, in this paper, the number of attributes could have been increased because currently, it has only 5 attributes.

Nhu et al. (2022) mentioned that one of the strategic goals of companies include customer retention. The study was focused on telecom sector and here support vector machine model applied to understand the customer churn. The dataset has variety of features so to understand the importance of the features, the author has focused on dimension reduction techniques. The dataset was available in depth because it has the variables like total night calls, total international calls, customer service calls, account length, area code, voice mail plan, total day minutes and so on. This dataset was also useful in achieving the goal of predicting customer churn because it has variables to assess the satisfaction level and engagement level of the customers. These methods were sequential forward selection and sequential backward selection. The dataset has imbalanced features as well and to deal with this problem SMOTE method was incorporated within the data pre-processing stage. The final results of the paper suggested that the model accuracy was 98.9 percent and its f1 score was 99 percent. these results were gained on different kernel functions of the support vector machine model. These kernels were polynomial, linear and radio basis. It was found that selected model showed highly reliable results on the radio-basis kernel function. For future research, the author set an agenda that the more advanced hyperparameter tuning techniques can be applied.

Sudarshan and Ganesh (2022) considered the customer churn issue in telecommunication sector and mentioned that constantly changing customer demands become the reasons behind customer churn. The author used deep learning algorithms to predict the customer churn. The dataset has 19 variables and 4250 training samples. The data has information about the telecommunication plan, voice plan, target variable, international plan and telecommunication service provider and so on. This data has multiple variables which was enough to understand the customer shopping patterns. The 19 attributes help in assessing the mindset or the relationship of customers with brand in different ways and strong churn prediction can be done. The applied algorithm was Swich recurrent neural network (S-RNN). Once the reasons behind the customer churn are known, then network utilisation theory was used to retain the customers. In this paper, the techniques like data normalisation were used and for feature selection, butterfly optimization program was applied. It was combined with Brownian movement. The experimentation results revealed that the proposed model has the accuracy of 95.99 percent. Its accuracy was compared with other models like convolutional neural network and artificial neural network. It was found that proposed model outperformed other deep learning techniques.

2.4: Summary:

The outcome of the above literature review provided the clear insight that customer churn is not an ideal scenario for the companies. The sectors like ecommerce have intense competition due to which, the customers have higher alternatives to make purchases. It became one of the major reasons behind frequent customer churn. It was also found that low customer loyalty and high customer dissatisfaction could lead towards the customer churn. When it comes to the negative impact, then it was found that the cost of customer acquisition increases and company’s potential to make higher sales gets affected poorly. The literature review analysis informed about the fact that previously, several studies have used machine learning algorithms to predict the customer churn. The major reason behind the same was that algorithms can work on non-linear data, large and complex data and can made churn predictions on different types of data like texts, images, ratings and so on.

References:

Akula, V. (2024). Council Post: The Proactive Edge: Leveraging AI For Customer Retention Success. *Forbes*. [online] 12 Aug. Available at: https://www.forbes.com/councils/forbestechcouncil/2024/05/02/the-proactive-edge-leveraging-ai-for-customer-retention-success/.

Khattak, A., Mehak, Z., Ahmad, H., Asghar, M.U., Asghar, M.Z. and Khan, A. (2023). Customer churn prediction using composite deep learning technique. *Scientific Reports*, [online] 13(1), p.17294. doi:https://doi.org/10.1038/s41598-023-44396-w.

Kim, S. and Lee, H. (2022). Customer Churn Prediction in Influencer Commerce: An Application of Decision Trees. *Procedia Computer Science*, 199, pp.1332–1339. doi:https://doi.org/10.1016/j.procs.2022.01.169.

Matuszelański, K. and Kopczewska, K. (2022). Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach. *Journal of Theoretical and Applied Electronic Commerce Research*, 17(1), pp.165–198. doi:https://doi.org/10.3390/jtaer17010009.

Nhu, N., Ly, T.V. and Son, D.V.T. (2022). Churn prediction in telecommunication industry using kernel Support Vector Machines. *PLOS ONE*, 17(5), p.e0267935. doi:https://doi.org/10.1371/journal.pone.0267935.

Olaniyi, A.S., Olaolu, A.M., Jimada- Ojuolape, B. and Kayode, S.Y. (2020). Customer Churn Prediction in Banking Industry Using K-Means and Support Vector Machine Algorithms. *International Journal of Multidisciplinary Sciences and Advanced Technology*, [online] 1(1), pp.48–54. Available at: https://www.researchgate.net/profile/Micheal-Arowolo/publication/346910130\_Customer\_Churn\_Prediction\_in\_Banking\_Industry\_Using\_K-Means\_and\_Support\_Vector\_Machine\_Algorithms/links/5fd1f34da6fdcc697bf30d22/Customer-Churn-Prediction-in-Banking-Industry-Using-K-Means-and-Support-Vector-Machine-Algorithms.pdf.

Prabadevi, B., Shalini, R. and Kavitha, B.R. (2023). Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks*, [online] 4. doi:https://doi.org/10.1016/j.ijin.2023.05.005.

Shobana, J., Gangadhar, Ch., Arora, R.K., Renjith, P.N., Bamini, J. and Chincholkar, Y. devidas (2023). E-commerce customer churn prevention using machine learning-based business intelligence strategy. *Measurement: Sensors*, 27, p.100728. doi:https://doi.org/10.1016/j.measen.2023.100728.

Statista (2023). *Digital buyer penetration in the United Kingdom 2017-2025*. [online] Statista. Available at: https://www.statista.com/forecasts/891311/digital-buyer-penetration-in-the-united-kingdom.

Sudharsan, R. and Ganesh, E.N. (2022). A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy. *Connection Science*, 34(1), pp.1855–1876. doi:https://doi.org/10.1080/09540091.2022.2083584.

SwiftERM (2022). *The UK Ecommerce Market Insights — SwiftERM Predictive Analytics For Ecommerce*. [online] swifterm.com. Available at: https://swifterm.com/the-uk-ecommerce-market-insights/ [Accessed 31 Oct. 2024].

Xiahou, X. and Harada, Y. (2022). B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. *Journal of Theoretical and Applied Electronic Commerce Research*, 17(2), pp.458–475. doi:https://doi.org/10.3390/jtaer17020024.