International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 01 | Jan 2022 www.irjet.net p-ISSN: 2395-0072

A Survey on Customer Analytics Techniques for the Retail Industry

Anuj Kinge[1], Yash Oswal[2] , Hrithik P. B.[3], Nilima Kulkarni[4]

[1]*Student, Dept. of Computer Science and Engineering, MIT School of Engineering, Maharashtra, India*
[2]*Student, Dept. of Computer Science and Engineering, MIT School of Engineering, Maharashtra, India*
[3]*Student, Dept. of Computer Science and Engineering, MIT School of Engineering, Maharashtra, India*
[4]*Professor, Dept. of Computer Science and Engineering, MIT School of Engineering, Maharashtra, India*

Abstract - *As the retail industry grows increasingly competitive, the ability to streamline company operations while meeting customer expectations has never been more crucial. As a result, controlling and channelling data to strive towards customer satisfaction while generating healthy revenues is critical to surviving and prospering. It is widely recognized that a better understanding of consumers may improve both customer happiness and business success. Understanding the client purchase journey allows retailers to employ effective marketing and managerial resources, optimise the trip, and give distinct experiences to different customer segments, assuring customer loyalty and generating sales. Several methods for collecting consumer insights have been established, and a few of these pieces of literature are covered in this study. These methods include conventional approaches using classification, segmentation, and association rule generation approaches for customer churn prediction, customer segmentation, and Market Basket Analysis, respectively.*

Key Words: **Customer Churn, Customer Segmentation, Market Basket Analysis, Retail Analytics, K means, Apriori, Eclat.**

1. INTRODUCTION

The way customers study and buy items has evolved, marketing managers are now challenged to enhance business efficiency across channels while justifying marketing investment. Consumers are actively using many channels at various phases of their decision-making and purchasing cycles. Customers visit and experience many platforms during the buying path, making marketing management crucial for businesses today. In order to give insights on how to manage an extended marketing mix that comprises many media and channels, including social media, research is required. Marketers must understand how customers behave to influence their purchasing journey and change their marketing strategy accordingly.

Analytics is the most crucial factor in this process since it drives retail organisations to understand consumer decision policies. It generates insights such as how a corporation would be able to boost profits at the product level, and it also gives insights into what the consumer is like, or why the customer would want to buy a given product, through the many activities that it does. The analytics also assist organisations in identifying which things a client is most likely to purchase together, which promotions and offers would work best for particular products, and customised recommendations for each unique customer. These insights are entirely focused on the client, but there are also those that are wholly focused on the firm. Analytics may provide insights into how much expenditure a firm will have to undertake, store-specific product mix, appropriate pricing to attract more consumers, efficient stock tactics, and many other things. A crucial part of the business is having a thorough grasp of the demands of the clients so that holistic perspectives of their patterns can be analysed. When customers are satisfied with the service or products, they become more loyal [1].

The approaches listed below are some of the most often utilised in retail analytics:

- Customer Churn
- Customer Segmentation
- Market Basket Analysis

1.1 Customer Churn

Customer churn occurs when a customer decides to discontinue using a company's products. When a customer leaves, there are typically early warning indicators that may be identified through churn research. Customers do not instantly stop purchasing from the company, but rather gradually defect - that is, they gradually go to a competitor [2]. Production cost analysis and advertising play critical roles in a company's success. As a result, it becomes even more critical as the organisation expands. During a company's initial growth period, it is quite simple to recruit consumers, and the number of customers that are prone to leaving is likewise relatively low. However, as the company expands, finding new customers becomes more challenging. In the meantime, the client turnover rate rises.

The fundamental goal of a firm should be to identify consumers who are likely to leave a particular bank dataset. The qualities of the data we have for a certain period must be appropriately recognized. It is usually preferable for any firm or organisation to keep existing clients rather than strive to acquire new ones. After identifying these consumers, the firm can offer services based on the reasons, uses, or criteria that they appreciate the most. A variety of classification algorithms such as Support Vector Machine, Artificial Neural Networks, Decision Tree, Random Forest, Logistic Regression,

XGBoost, CatBoost, AdaBoost can be used to fulfil the criterion of anticipating whether or not the customer will leave the organisation.

1.2 Customer Segmentation

Customer segmentation is the technique of categorising a firm's customers into groups based on their analogy. The main idea of customer segmentation is to determine how to engage with customers in each category to maximise each customer's value to the business. Segmentation is not a one-time effort but rather a continual and iterative process of knowledge discovery from massive amounts of raw and unorganised data [3]. Customer segmentation enables marketers to reach out to each customer in the most efficient way possible. Customer segmentation takes advantage of the large amount of data available on clients marketers to accurately identify separate groups of customers based on demographic, behavioural, and other criteria.

Because the marketer's aim is generally to maximise the profit from each client, it is vital to understand how each given marketing action will affect the consumer ahead of time. Clients must be grouped or segmented based on their CLV. In order to fulfil the purpose of customer segmentation, unsupervised machine learning could be utilised, such as K means clustering, Fuzzy K means clustering, DBSCAN, etc.

1.3 Market Basket Analysis

Market basket analysis is a process that seeks for links between entities and things that often occur together, such as the products in a shopper's cart. Finding frequent itemsets in a transaction dataset and generating association rules is one of the most often used data mining techniques [4]. In the retail industry, market basket analysis explores the relationship between items by taking into account the co-occurrence of purchases in previous transactions. Association analysis, like market basket analysis, is a generalisation of applications that is now widely used in clickstream analysis, cross-selling recommendation engines, and information security. Association analysis is an unsupervised data science approach that does not need a target variable to be predicted. Instead, the system examines each transaction including a number of items (products) and identifies valuable relationship patterns. The difficulty in association analysis is to distinguish between a relevant observation and unethical rules. The Apriori, Eclat, and Frequent Pattern Growth algorithms provide effective methods for extracting these rules.

2. LITERATURE SURVEY

An in-depth study was conducted to learn about the most prevalent methods of customer attrition. The following tables list some of the selected literature papers.

Table -1.1: Literature paper-1 [5]

Year	2017
Title	Comparison of Deep Learning Algorithms to Predict Customer Churn within a Local Retail Industry
Aim	The objective of this paper is to predict customer churn for a local supermarket by considering transactional data features and point of sales systems.
Methodologies, Algorithms, Techniques	The implementation of Convolutional Neural Network (CNN) and Restricted Boltzmann (RBM)algorithms is done in this study that can be used to predict customer churn. For this purpose, a supermarket's data was considered for research, and Kimball methodology was applied, followed by the Extract, Transform and Load (ETL) process. Parameter Selection and a couple of churn models are created for two different algorithms (CNN and RBM).
Result/ Accuracy	For CNN, Accuracy: 63%(Model-1) , 74%(Model-2) Sensitivity: 59%(Model-1) , 66%(Model-2) For RBM, Accuracy: 77%(Model-1) , 83%(Model-2) Sensitivity: 67%(Model-1), 74%(Model-2) The RBM produced the best results when it came to categorising churners as compared to CNN.
Conclusion	This article explored the significance of accessible data of the customer's transactions that supermarkets retain and the optimal factors for predicting customer churn and, as a result, predicting turnover.

Table -1.2: Literature paper-2 [6]

Year	2021
Title	Deep learning for customer churn prediction in e-commerce decision support.

Aim	The aim of the work is to develop a model of deep learning that predicts client attrition by using the complete history of each customer's transactions.
Methodologies, Algorithms, Techniques	The work was done using actual data from the e-commerce industry, with a significant percentage of buyers being one-time consumers. Model tuning involved using two base ANN topologies. The research's approach was very well structured as it performed a 10-fold split over the dataset, used fully connected dense layers, RNN as the first hidden layer, chose a particular number of neurons by observing accuracy and F1 scores, and optimised the network using binary cross-entropy as its loss function. It used both cases of the model with and without dropout for comparing the training results. Leaky ReLU activation function was used for dying ReLU problems along with the standard rectified linear unit activation function.
Result/ Accuracy	Sixteen models were compared based on precision, recall, accuracy, activation function, and AROC. The variance in most models was analogous. Accuracy: 74% Precision: 78% Recall: 68%
Conclusion	Regular consumers (those who make more than five purchases) account for barely 2% of the total population. Thus, the dataset becomes unbalanced, and it becomes a very challenging task to create a good churn model. Rather than using random selection to create the training and test datasets, it could be useful to devise a mechanism to ensure all of a consumer's transactions can just occur for one dataset.

Table -1.3: Literature paper-3 [7]

Year	2018
Title	The Research of Online Shopping Customer Churn Prediction Based on Integrated Learning
Aim	This research paper plans to identify lost customers and categorise them into different types by applying integrated learning theory.
Methodologies, Algorithms, Techniques	RFM(Recency-Frequency-Monetary) theory is used to categorise the various values of lost consumers. ANN and SVM form the basis for applying the integrated learning theory. The Schmittlien Morrison and Colombo (SMC) model is used to predict customers' future activity to solve non-contract customer churn. The approach of using a combined model can provide better prediction results.
Result/ Accuracy	For ANN, Correct rate: 82.64% Error rate: 17.36% For SVM, Correct rate: 77.36% Error rate: 22.64% Correct rate of combined prediction: 93.01% Error rate of combined prediction: 6.99%
Conclusion	According to the findings of customer churn prediction, enhancing the accuracy of customer churn prediction is required to minimise the lost customers forecast for the current customers. The empirical results reveal that the combined forecasting model improves the hit rate, coverage rate, accuracy rate, and lift degree.

Table -1.4: Literature paper-4 [8]

Year	2020
Title	An Empirical Study on Customer Segmentation by Purchase Behaviours Using an RFM Model and K-Means Algorithm.
Aim	To undertake customer segmentation and value analysis to develop various approaches to enhance customer satisfaction.
Methodologies, Algorithms, Techniques	This proposed work uses online sales data, where an RFM (recency, frequency, and monetary) model, and the K-means clustering method is used to perform consumer segmentation and value analysis. Customers are divided into four groups depending on their purchasing habits. The process follows these steps: 1. Data preprocessing and

	preparation 2. Normalisation of RFM model indices Index Weight analysis 3. Clustering of customers using K-means.
Result/ Accuracy	No. of active customers- Added 529 more customers Total purchase Volume - went up by 279% Total consumption amount - went up by 101.97%
Conclusion	The efficiency of the analytical approach described in this article is demonstrated by improving the company's primary performance metrics. Learning how to put algorithms into a Customer relationship management service to support managers in making choices is a good way to increase performance.

Table -1.5: Literature paper-5 [9]

Year	2013
Title	Customer Segmentation Using Clustering and Data Mining Techniques
Aim	To develop a real-time and online system for a specific supermarket to forecast sales by using clustering techniques like k-means and SPSS tools.
Methodologies, Algorithms, Techniques	This study paper comprehensively analyzes the k-means clustering approach, and the SPSS Tool used to construct an online system for predicting sales in multiple yearly seasonal cycles for a particular supermarket. The model compares actual day-to-day sales figures to forecasted statistics. Different variables were considered for market segmentation. To assess the clusters' stability ANOVA study was also performed
Result/ Accuracy	K-means algorithm used - k = 4 groups / clusters n = 2138 customers Variables = 15 Highest Cluster MS = 9.650 (for Var-7) Highest Error MS = 1.880 (for Var-14) Highest F-Statistic = 24.704 (for Var-7) Highest P-value = 0.567 (for Var-9)
Conclusion	The analysis provided further examples of applying the cluster approach for market segmentation and forecasting. The computing-based system created automatically gave results to managers so that they could make rapid and informed decisions. Cluster branding and other cluster traits indicating a specific group of people were also subjected to simulation testing.

Table -1.6: Literature paper-6 [10]

Year	2021
Title	Clustering Approaches to Offer Business Insights
Aim	To study the relevance of Customer Segmentation as a core CRM capability (Customer Relationship Management) for fragmenting clients, utilising bunching procedures.
Methodologies, Algorithms, Techniques	This examination uses three approaches for Customer Segmentation: 1. K-means - Centre Based clustering Algorithm, 2. Hierarchical - Connectivity Based clustering algorithm 3. DBSCAN - Density-Based clustering algorithm
Result/ Accuracy	K-Means clustering performs better for a large number of perceptions, but hierarchical clustering can cope with fewer information foci.
Conclusion	This study presents three distinct ways that enable the choice of selecting a solution in specific situations. K-Means, DBSCAN, and Hierarchical clustering all have certain types of drawbacks that render them unsuitable when used solely.

Table -1.7: Literature paper-7 [11]

Year	2012
Title	Association Rule - Extracting Knowledge Using Market Basket Analysis
Aim	To study a large quantity of data to exploit customer behaviour and make the best decision possible, giving a business a competitive edge over its competitors.

Methodologies, Algorithms, Techniques	Market Basket Analysis from Association Rule mining may be performed on the customer transactions data at retail stores. 1. Proposed a system based on data mining for the analysis using the association rules and its two measures: Rule support and confidence. 2. Categorical data from transaction records are fed into the analysis, and the result is a set of association rules derived directly from the data.
Result/ Accuracy	Total transactions = 50 Association rules threshold values: Support: 20% Confidence: 60% 3, 10, 15, 17, 19, 20, and 23 have higher values than threshold values of support and confidence.
Conclusion	The presented research clearly demonstrates that data mining technologies may be utilised to optimise patterns connected with the dynamic behaviour of consumer transactions while purchasing a specific product.

Table -1.8: Literature paper-8 [12]

Year	2012
Title	Market Basket Analysis with Map/Reduce of Cloud Computing
Aim	This paper aims to apply the Market Basket Analysis technique, which is used to sort datasets and convert them to pairs of (key, value) that can be used with Map/Reduce.
Methodologies, Algorithms, Techniques	The following are the paper's key contributions: Proposing a Map/Reduce algorithm implemented in parallel computing. Instead of utilising HBase DB, input/output files are handled on HDFS. Data is composed in the form of a list structured with (key, value) pair values, and the amount of values per key is accumulated using a reducer.
Result/ Accuracy	Some nodes boost performance linearly for some transaction data sets; however, this has a constraint. For 6.7M transaction: Nodes = 20 Best execution time = 2868s For 13M transaction: Nodes = 20 Best execution time = 2911s For 26M transaction: Nodes = 20 Best execution time = 5671s
Conclusion	The results show that as additional nodes are added, the code using Map/Reduce increases performance, but there is a bottleneck at some point that precludes performance improvement. The bottleneck in Map/Reduce is revealed to be the processes of spreading, collecting, and lowering data.

Table -1.9: Literature paper-9 [13]

Year	2013
Title	Market Basket Analysis of Sports Store using Association Rules
Aim	To maximise the marketing and sale of sports equipment through the use of Market Basket Analysis and the FP Growth Algorithm.
Methodologies, Algorithms, Techniques	In this work, the Frequent Pattern Growth approach (for frequent itemset mining) is used for successful mining of recurring patterns in large databases since it offers several advantages over other strategies. Proposed a data mining-based approach for analysing association rules and their measures: rule body, rule head, support, confidence, and lift.
Result/ Accuracy	Association Rules: Minimum confidence = 1 was considered, Minimum Support of 0.527 and lift = 1.897 were observed for the most frequent items.
Conclusion	The Frequent Pattern Growth method used in this study shows pragmatic mining of frequent patterns in massive databases since it has various benefits over other techniques.

Table -1.10: Literature paper-10 [14]

Year	2016
Title	Market Basket Analysis using FP Growth and Apriori Algorithm: A case study of Mumbai Retail Store
Aim	The primary goal of this research work is to explore how different items in a grocery store variety interact with one another and how to use marketing actions to exploit these relationships.
Methodologies, Algorithms, Techniques	This study introduces a technique known as market basket transactions, which may be used to find intriguing correlations in large datasets. A binary representation format is employed when using association rule mining to detect the frequency of item sets. Frequent Pattern (FP) Growth and Apriori algorithms are used for association rules. On top of that, RapidMiner Platform is used.
Result/ Accuracy	In comparison to Apriori, FP growth is rather modest for a high number of transactions, according to the analysis results. To make frequent itemsets using Rapid Miner, you'll need more time. The Apriori technique in R programming generates many rules in milliseconds.
Conclusion	During the analysis, it was revealed that FP growth takes longer than Apriori for a high number of transactions. To make frequent itemsets using Rapid Miner, more time is required.

Table -1.11: Literature paper-11 [15]

Year	2021
Title	Comparative Analysis of Apriori and ECLAT Algorithm for Frequent Itemset Data Mining
Aim	This study aims to demonstrate that ECLAT outperforms Apriori when it comes to experimental and execution time in frequent set mining.
Methodologies, Algorithms, Techniques	This research employs the Cross-Industry Standard Process for Data Mining technique. In this research, the Apriori algorithm's purpose is to identify the most frequent itemset by considering the minimum support value and the minimum confidence value. In contrast, the ECLAT method uses the itemset patterns to do so.
Result/ Accuracy	Results of manual analysis of:- ECLAT algorithm: Taken - 9 rules Execution time - 0.2s Apriori algorithm: Taken - 11 rules Execution time - 0.4s
Conclusion	According to the research findings, the execution time required to run the 1846 items in the program demonstrates that the ECLAT method is 0.2 seconds faster than the Apriori technique which takes 0.4 seconds. In terms of scalability, the ECLAT method outperforms the Apriori algorithm. For huge datasets, the Apriori method is insufficient.

3. CONCLUSION

This literature survey paper examines several techniques to acquire consumer insights. We examined contemporary studies utilising machine learning and data mining approaches, including classification, segmentation, and market basket analysis. Retailers can gain a lot from an analytics-driven strategy that helps them understand how their consumers use their goods, as well as how to anticipate major risks like customer attrition - information that they can act on. These advancements have the potential to impact and accelerate the adoption of these approaches in the retail business.

REFERENCES

[1] S. Neslin, S. Gupta, W. Kamakura, L. Junxiang, and C. H. Mason, "Defection detection: Measuring and understanding the predictive accuracy of customer churn models," Journal of Marketing Research, vol. 43, pp. 204-211, 2006

[2] W. Buckinx and D. V. den Poel, "Customer base analysis: Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting," European Journal of Operational Research, pp. 252-268, 2005.

[3] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pp. 881-892, 2002.

[4] Wu X, Kumar V, Quilan JR., Ghosh J, Yang Q, Motoda H. Top 10 Algorithms in Data Mining. Springer-Verlay London Limited 2007:14:1-37

[5] Dingli, A., Marmara, V. and Fournier, N.S., 2017. Comparison of deep learning algorithms to predict customer churn within a local retail industry. International journal of machine learning and computing, 7(5), pp.128-132.

[6] Pondel, M., Wuczyński, M., Gryncewicz, W., Łysik, Ł., Hernes, M., Rot, A. and Kozina, A., 2021, July. Deep learning for customer churn prediction in e-commerce decision support. In Business Information Systems (pp. 3-12).

[7] Xia, G. and He, Q., 2018, March. The Research of online shopping customer churn prediction based on integrated learning. In Proceedings of the 2018 International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2018), Qingdao, China (pp. 30-31).

[8] Wu, J., Shi, L., Lin, W.P., Tsai, S.B., Li, Y., Yang, L. and Xu, G., 2020. An empirical study on customer segmentation by purchase behaviours using a RFM model and K-means algorithm. Mathematical Problems in Engineering, 2020.

[9] Kashwan, K.R. and Velu, C.M., 2013. Customer segmentation using clustering and data mining techniques. International Journal of Computer Theory and Engineering, 5(6), p.856.

[10] Sharma, B.T.S.S., Ahmad, B.T.S.K. and Singh, B.T.S.V., 2021. Clustering Approaches to Offer Business Insights.

[11] Raorane, A.A., Kulkarni, R.V. and Jitkar, B.D., 2012. Association rule–extracting knowledge using market basket analysis. Research Journal of Recent Sciences ISSN, 2277, p.2502.

[12] Woo, J. and Xu, Y., 2011. Market basket analysis algorithm with map/reduce of cloud computing. In Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA) (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[13] Kaur, H. and Singh, K., 2013. Market basket analysis of sports store using association rules. International Journal of Recent Trends in Electrical & Electronics Engg, 3(1), pp.81-85.

[14] Venkatachari, K. and Chandrasekaran, I.D., 2016. Market basket analysis using fp growth and apriori algorithm: a case study of mumbai retail store. BVIMSR's Journal of Management Research, 8(1), p.56.

[15] Krishnan, M.S., Nair, A.S. and Sebastian, J., 2022. Comparative Analysis of Apriori and ECLAT Algorithm for Frequent Itemset Data Mining. In Ubiquitous Intelligent Systems (pp. 489-497). Springer, Singapore.

BIOGRAPHIES

Anuj Kinge is from Pune City, Maharashtra, India. He is currently pursuing B.Tech from MIT-ADT University, School of Engineering, Pune, India. He has won the Best Paper Award at the Springer conference (CIIR 21). He has worked on projects such as Automatic Timetable Generator, Mini-Alexa, Chatbots. He is interested in Web Development, Data Analysis, Machine Learning and Deep Learning.

Yash Oswal is a computer science and engineering undergraduate student at MIT ADT University. Machine learning, deep learning, and algorithmic approaches are some of his areas of interest. He had worked on projects such as customer churn for retail banking organisations and People Safety Android app for Campus Students.

Hrithik P. B. is a student pursuing B.tech at MIT ADT University, Pune, India. He has worked on projects based on IOT and machine learning. He has contributed to the ICAR project and is also keenly interested in topics like Data Analysis, Deep Learning and Natural language processing.

Dr. Nilima Kulkarni is currently working as Associate Professor at MIT-ADT University, School of Engineering, Pune, India. She has completed BE (CSE), ME (CSE), from SRTMU Nanded, Maharashtra, and PhD from Amrita Vishwa Vidyapeetham, Bangalore India. Image processing, medical image processing, computer vision, artificial intelligence, eye tracking, machine learning, and deep learning are all areas in which she is interested in conducting research.