International Journal of Scientific Research & Engineering Trends
Volume 11, Issue 5, Sep-Oct-2025, ISSN (Online): 2395-566X

Survey on Customer Behavior Data Analysis for Product Purchasing

Keerti Pal[1], Prof. Jayshree Boaddh[2], Prof. Rahul Patidar[3]
[1]MTech Scholar Vaishnavi Institutes of Technology and Science, Bhopal.
[2]HOD, CSE Vaishnavi Institutes of Technology and Science, Bhopal
[3]Asst.Prof., CSE Vaishnavi Institutes of Technology and Science, Bhopal

Abstract- — Product Sales Dataset is a comprehensive collection of sales data for a wide range of products available on the E-commerce e-commerce platform. This kind of dataset provides invaluable insights into customer behavior, product performance, and market trends, making it an essential resource for data analysis, market research, and business strategy development. This dataset is indispensable for market research, allowing businesses to discern market trends, consumer preferences, and competitive landscapes. This paper presents a comprehensive approach to customer behavior analysis and predictive modelling within the context of supermarket retail. This paper finds techniques that extract patterns in shopping data for the learning and prediction of user preference. This work list different proposed models with techniques. Paper has list various evaluation parameters of user purchase prediction models.

Keywords: Shopping Store, Product Recommendation, Feature Optimization.

INTRODUCTION

In the contemporary landscape of retail, particularly within the dynamic realm of supermarket operations, the quest to understand and predict customer behaviour standcus as a cornerstone of strategic decision-making. Amidst the proliferation of data sources and the advent of sophisticated analytical methodologies, retailers are compelled to adopt comprehensive data mining approaches to derive actionable insights that drive business success. This paper embarks on a journey through the intricate domain of customer behaviour analysis and predictive modelling within the context of supermarket retail, elucidating the pivotal role of data-driven strategies in optimizing operational efficiencies and enhancing customer satisfaction.

However, the retail sector is a complex and ever-evolving market that heavily relies on customer behavior [1]. Studying consumer behavior is a complex and challenging task that requires a deep understanding of the factors that influence customer decision-making. Customers are influenced by a variety of factors, including cultural, social, economic, and personal factors, and their behavior is often difficult to predict. Analyzing and understanding consumer behavior allows retailers to stay ahead of their competition, respond better to market changes, and remain competitive in the ever-changing retail sector [2]. Moreover, customer behavior is constantly evolving, making it difficult to predict future trends. To effectively analyze consumer behavior, businesses need to leverage the power of big data analysis and use sophisticated analytical techniques to identify trends and patterns in customer behavior [3].

This provides opportunities for research based on real data collected by businesses which may reveal actual customer behaviour. Contrary to data collected through surveys which are subject to more or less bias and distortion as customers (as respondents) may not reply truthfully (Bolarinwa 2015), transaction data collected by businesses do not suffer from such limitations, as it captures customers' decisions and behaviour.

Transaction data refer to events (sequences of events) whose main purpose is to provide information about a transaction and can be analysed to better understand the activities of the enterprise (Hand 2018). Transaction data are very high-frequency data where all transactions are recorded, and these data contain valuable information about the time between transactions as well as underlying information about the type of trades (Hautsch & Pohlmeier 2001).

II. RELATED WORK

R. A. Moral et al. in [6], propose a novel Bayesian hierarchical joint model that is able to characterise customer profiles based on how many events take place within different television watching journeys, and how long it takes between events. The model drastically reduces the dimensionality of the data from thousands of observations per customer to 11 customer-level parameter estimates and random effects.

Shakeel, P. M. et. al. in [7] proposes the adaptive hybridized intelligent computational model (AHICM) as an effective tool to analyze consumer behavior for business development using consumer data from different sources. Consumer data include perception, affect, cognition beliefs, market research, and business strategy to inform choices and preferences. The AHICM was applied to online advertising and consumer buying behavior.

Wen (2023) [8], analyzed customer churn in banks using logistic regression and decision tree models, with the latter proving more precise and unbiased. Key predictors of churn included age, salary, and product usage, with certain customer profiles showing a higher tendency to leave. They recommend that banks refine their systems to retain customers at high risk of churning (Wen 2023).

Xiao and Piao (2022) et. al. in [9], explored the significant data era's onset and how mobile edge computing's transformative potential has led to a surge in consumer data production. They highlighted the critical need to analyze this vast data for valuable insights. The study focused on understanding customer groups' consumption patterns to develop effective marketing strategies, particularly in the telecom industry. It pointed out the growing use of association rules for analyzing customer behavior due to the complex data and algorithms involved. The paper noted that while association rules and mobile computing are potent for identifying patterns in large data sets, they have limits, such as the inability to infer causality (Xiao and Piao 2022).

Stuti et al. (2022) in [10], presented a detailed method to extract insights from customer purchasing data to improve product recommendations. Their two-step approach identified product correlations from user transactions and then formulated utility-based association rules to map purchasing trends. These rules were integrated into a recommender system to suggest new products. The study also evaluated the accuracy of these rules against those derived from Frequent Item Set Mining and Improved Utility-Based Mining on an e-commerce platform (Stuti et al. 2022).

Wang, W. et. al. in [11], In this paper, a prediction model based on XGBoost is proposed to predict user purchase behavior. Firstly, a user value model (LDTD) utilizing multi-feature fusion is proposed to differentiate between user types based on the available user account data. The multi-feature behavior fusion is carried out to generate the user tag feature according to user behavior patterns. Next, the XGBoost feature importance model is employed to analyze multi-dimensional features and identify the model with the most significant weight value as the key feature for constructing the model. This feature, together with other user features, is then used for prediction via the XGBoost model.

III. PREDICTION METHODS

Regression Methods Regression techniques are widely used in purchase prediction models to analyze relationships between dependent (e.g., purchase decision) and independent variables (e.g., user demographics, product ratings, social signals). Both linear and non-linear models can be applied, but linear regression often proves more effective when working with numerical and structured data like click-through rates, product views, and purchase history [12]. However, for tasks like sentiment analysis, where the relationship between input and output is complex and non-linear, non-linear regression models or advanced machine learning techniques (e.g., deep learning) may offer better performance [13]. For instance, studies have utilized regression-based sentiment models to analyze early and recent user posts across multiple platforms, using this data to predict future purchases or preferences [14].

Naive Bayes Classifier The Naive Bayes classifier is a probabilistic model based on Bayes' theorem. It calculates the posterior probability of a class (e.g., purchase or no purchase) based on prior probabilities and the observed features. Despite its assumption of conditional independence among features—which may not always hold in real-world data—it has shown promising results in various user behavior prediction scenarios [15]. When predicting discrete outcomes like "will purchase" or "will not purchase," the model can be directly applied. For continuous outcomes, preprocessing or discretization of the prediction variable is required [16].

K-Nearest Neighbor (KNN) Classifier The KNN algorithm is a non-parametric, instance-based learning method used to classify users by comparing them with similar users (neighbors). It predicts a user's purchase behavior based on the behavior of the k most similar users, using distance metrics such as Euclidean or Manhattan distance over multidimensional feature vectors (e.g., browsing time, past purchases, clicks) [17]. This method is intuitive and effective when the dataset is well-structured and not too sparse, making it useful for purchase behavior clustering.

Artificial Neural Networks (ANN) Artificial Neural Networks (ANN) simulate the human brain's functioning and have shown significant success in modeling complex, non-linear relationships in user behavior data [18]. An ANN comprises input, hidden, and output layers. The input layer receives user

and product data, the hidden layer processes patterns, and the output layer predicts the purchase outcome.

Neural networks are particularly useful for handling high-dimensional data. Additionally, Self-Organizing Maps (SOM)—a form of ANN—can be used for dimensionality reduction, helping to extract important features before performing classification or prediction [19].

Decision Trees: Decision trees are transparent, rule-based models that predict outcomes by traversing from a root node to leaf nodes based on user attributes. Unlike neural networks, which are often considered "black box" models, decision trees are interpretable and can handle missing data and categorical variables effectively.

Model-Based Prediction Model-based prediction involves constructing mathematical representations of user behavior before applying predictive algorithms [20]. These models aim to simulate decision-making processes, user-product interactions, and network influences. Although building such models requires deep understanding of user dynamics and social contexts, they offer insights into the mechanisms behind purchasing behavior. Some studies have proposed frameworks to model user activities on social platforms, but challenges remain due to the complexity and variability of human behavior online [21].

SOCIAL NETWORKS AND PURCHASE BEHAVIORS

Social networks have a significant impact on consumer purchasing decisions, and this influence has been widely studied. One notable study by Aral et al. (2009) highlights the concept of peer pressure in online environments. Their findings indicate that if a user's friends adopt a particular product or service, the likelihood that the user will also make the purchase increases substantially. This social contagion effect demonstrates how behavior can spread through social ties [22].

Further evidence is presented in the work of Zhang et al. (2012), who investigated the Taobao e-commerce platform—China's largest online shopping network. Their research revealed a strong information passing mechanism: buyers tend to purchase from sellers who have previously sold to their friends, even when other sellers offer lower prices or have better ratings. This suggests that social influence can outweigh objective product attributes during decision-making [23].

Online social networks serve various purposes, and the nature of interactions within them also differs. For example, Facebook enables users to engage in a range of activities such as liking, commenting, tagging, or posting on another user's wall. These actions can be treated as features representing user behavior. However, not all interactions result in the formation of new links within the social network graph. For instance, sending a message or a friend request does not immediately create a graph edge. A formal connection is only established once a friend request is accepted, thereby generating a new edge between the two nodes in the social graph [24].

A key challenge in behavior analysis is feature extraction. One effective method is to construct a behavioral dataset by logging user-generated events and their frequencies. Let us consider a social graph with Nodes = {U1, U2, U3, ..., Un} and Links = {L1, L2, L3, ..., Ln}. Each user's behavior can be captured through event-frequency metrics (e.g., number of likes, comments, or messages sent), which provide the foundation for deeper behavioral modeling.

The area of social media recommendation is still emerging, with most work focusing on recommending content (e.g., URLs, posts, videos) or new friends. Unlike traditional recommender systems that rely mainly on user-item interactions, social media recommender systems utilize the user's social connections to enhance the accuracy of suggestions. Systems that incorporate social ties often outperform those based solely on user preferences or collaborative filtering [25].

IV. EVALUATION PARAMETER

To assess the performance of a customer purchase prediction model, several key evaluation metrics are commonly used, including accuracy, precision, recall, and F1-score. These metrics help determine how effectively the model predicts whether a customer will make a purchase.

Each metric is calculated based on the values derived from the confusion matrix, which includes true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). By substituting the model's results into the respective formulas for these metrics, we can quantitatively evaluate its predictive accuracy and reliability.

For instance:

- True Positive (TP): The model correctly predicts that a customer will make a purchase, and the customer indeed makes a purchase.
- False Positive (FP): The model predicts that a customer will make a purchase, but the customer does not.

- True Negative (TN): The model correctly predicts that a customer will not make a purchase, and the customer does not.
- False Negative (FN): The model predicts that a customer will not make a purchase, but the customer actually does.

Accurate evaluation using these parameters ensures the robustness of the customer purchase prediction model and supports its deployment in real-world retail and e-commerce applications.

V. CONCLUSION

Mining World Wide Web has necessitated the users to make use of automated tools to locate desired information resources and to follow and assess their usage pattern. Web item pre- has been widely used to reduce the user confusion problem. Proposed model is mostly used for prediction in social network because of its high accuracy. It is a powerful method for arranging users' choices with there social relation into clusters according to their similarity. This survey helps to develops a techniques for overcoming the issues of web item prediction. However, research of the web item prediction is just at its beginning and much deeper understanding needs to be gained.

REFERENCES

1. Chong Tat Chua, Hady W. Lauw, and Ee-Peng Lim. "Generative Models for Item Adoptions Using Social Correlation". IEEE Transactions On Knowledge And Data Engineering, Vol. 25, NO. 9, SEPTEMBER 2013.
2. Liben-Nowell, David, and Kleinberg, Jon. (2007). The Link Prediction Problem for Social Networks . Journal of the American Society for Information Science and Technology, 58(7):1019-1031.
3. J. Herlocker, J. Konstan, L. Terveen, and J. Riedl. Evaluating collaborative filtering recommender systems. ACM Trans. on Information Systems, 22(1):5–53, 2004.
4. G. Szabo and B. a. Huberman, "Predicting the popularity of online content," Communications of the ACM, vol. 53, no. 8, p. 80, Aug. 2010.
5. S. Kak, Y. Chen, L. Wang, "Data Mining Using Surface and Deep Agents Based on Neural Networks," in Proceedings of the Sixteenth Annual Americas' Conference on Information Systems, 2010.
6. R. A. Moral et al., "Profiling Television Watching Behavior Using Bayesian Hierarchical Joint Models for Time-to-Event and Count Data," in IEEE Access, vol. 10, pp. 113018-113027, 2022.
7. Shakeel, P. M., & Baskar, S. (2020). Automatic Human Emotion Classification in Web Document Using Fuzzy Inference System (FIS): Human Emotion Classification. International Journal of Technology and Human Interaction, 16(1), 94–104.
8. Wen, Z. 2023. Feature analysis and model comparison of logistic regression and decision tree for customer churn prediction. Journal of Communication and Computer 20: 1073.
9. Xiao, B., and G. Piao. 2022. Analysis of infuencing factors and enterprise strategy of online consumer behavior decision based on association rules and mobile computing. Wireless Communications and Mobile Computing 2022: Article ID 6849017.
10. Stuti, S., K. Gupta, N. Srivastava, and A. Verma. 2022. A novel approach of product recommendation using utility-based association rules. International Journal of Information Retrieval Research (IJIRR) 12 (1): 1–19.
11. Wang, W.; Xiong, W.; Wang, J.; Tao, L.; Li, S.; Yi, Y.; Zou, X.; Li, C. A User Purchase Behavior Prediction Method Based on XGBoost. Electronics 2023, 12, 2047.
12. Jindal, N., & Liu, B. (2007). Review spam detection. Proceedings of the 16th International Conference on World Wide Web, 1189–1190.
13. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.
14. Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. Management Science, 57(8), 1485–1509.
15. McCallum, A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. AAAI-98 Workshop on Learning for Text Categorization, 752(1), 41–48.
16. Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29(2), 103–130.
17. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27.
18. Zhang, G. P. (2000). Neural networks for classification: A survey. IEEE Transactions on Systems, Man, and Cybernetics, 30(4), 451–462.
19. Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), 1464–1480.
20. Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.
21. Lerman, K., & Ghosh, R. (2010). Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, 90–97.

22. Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proceedings of the National Academy of Sciences, 106(51), 21544–21549.
23. Zhang, J., Ackerman, M. S., & Adamic, L. A. (2012). Expertise networks in online communities: structure and algorithms. Proceedings of the 16th ACM Conference on Computer Supported Cooperative Work, 71–80.
24. Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. N., & Zhao, B. Y. (2009). User interactions in social networks and their implications. Proceedings of the 4th ACM European Conference on Computer Systems, 205–218.