Predictive Modelling and Customer Behaviour Analysis in E-Commerce Using Machine Learning

Prepared by

Talha Yusuf

2221858030

Fall 2025

Prepared for

Dr. Mehe Z. Rahman

Associate Professor, Department of Management

Sadiya Hossain Suriya

Product Manager, Market Operations, Cartup Limited

24.12.2025

Bachelor of Business Administration

School of Business and Economics

North South University

Letter of Transmittal

24-12-2025

To

Dr. Mehe Z. Rahman

Associate Professor

Department of Management

North South University

Subject: Submission of my BUS498 Internship Report

Dear Madam,

With due respect, I am pleased to submit my internship report titled 'Predictive Modelling and Customer Behaviour Analysis in E-Commerce Using Machine Learning", prepared as a partial requirement for 'BUS498: Internship' under the Bachelor of Business Administration (BBA) programme at North South University.

This report has been prepared by following the guidelines provided and is based on my internship experience at Cartup Limited from August 20, 2025, to December 11, 2025. The report reflects my effort to apply my knowledge gained at the university to practical business operations within my internship organisation.

I sincerely thank you for providing the opportunity to undertake this internship programme under your supervision, which has greatly contributed to my academic and professional development. I also express my deepest gratitude to my organisational supervisor Sadiya Hossain Suraiya for his continuous guidance and valuable feedback throughout the process.

I hope that this report meets your expectations and serves the intended academic purpose.

Yours faithfully,

Talha Yusuf

ID No.: 2221858030

Bachelor of Business Administration (BBA)

North South University

Acknowledgement

To start with, I would like to give Almighty Allah credit for the fact that he gave me the strength and direction to conduct this research. I am also deeply appreciative of my academic supervisor, Dr. Mehe Z. Rahman and my internship mentors at Cartup Limited who have provided me with valuable guidance and support. I also wish to thank my fellow workers and those who contributed to the data used in this research by providing resources. They have been very helpful in the process of conducting this research.

Executive Summary

In this thesis, the discussion is about customer behaviour and predictive modelling on e-commerce stores using machine learning. The paper relies on the clickstream and transaction data in order to investigate the behaviour of the users, such as page views, product clicks, cart addition, and purchase. A Logistic regression is used to estimate the behaviour patterns, with such important drivers as cart value, session duration, and product clicks. Random Forest is an algorithm that is used to predict purchase outcomes with high accuracy. The findings are a good model performance, where the accuracy reaches up to 100 percent with the use of the Random Forest and behavioural insights that can be used to influence marketing, pricing, and UX enhancement. Suggestions are the optimization of the checkout process, improved product recommendations and decreased decision fatigue to drive up conversions. This research paper offers practical information to the e-commerce companies in order to enhance their performance in terms of operations and revenue.

Table of Contents

Research Title

Predictive Modelling and Customer Behaviour Analysis in E-Commerce Using Machine Learning

Research objective

This study aims at developing an effective data-driven construct that will help e-commerce firms to learn and estimate customer behaviour. In particular, the research question is to:

- Determine trends in customer conduct that results in purchasing or leaving carts.
- Classify probability of purchase based on features of the session.
- Produce practical intelligence to maximize dynamic pricing, promotion tactics, and the general conversion rates.

Through the use of predictive modelling the e-commerce sites are able to cut on cart abandonment, maximize on personalized recommendations, and maximize on sales. This strategy perfectly correlates with the business objectives that are present today to generate higher income and offer a better customer experience.

Research Context

The e-commerce industry has hit an exponential growth, enhanced by the technology, the growth in mobile internet penetration and consumer behaviour change. The online retail sector in Bangladesh is new; it has challenges such as cart abandonment, sub-optimal pricing and not understanding customer journeys.

Cart abandonment is a chronic problem, and the research indicates the world cart abandonment rates are 60-70 percent. The high shipping fees, lengthy delivery time, complicated checkout procedures, and price sensitivity may lead to customers abandoning the company. On the same note, when conveying the changes in demand and user interaction, the statistical pricing approach tends to be ineffective and causes the company to miss revenue.

Although bigger companies like Amazon, Alibaba, and Shopify can optimize pricing, product recommendations, and purchase probability with the help of the advanced machine learning models, smaller businesses often need to make use of the descriptive analytics. These are reactive, and not proactive and hence cannot predict customer needs, and can do nothing to stop cart abandonment.

The data employed in this experiment is of a dummy type, and it is session-level data which is meant to mimic a real-life interaction. It contains:

- User interactions (page views, clicks, views on the product, placing an item into a cart, purchase).
- Cart and session statistics (quantity of items, cart worth, amount of time).
- Product interaction (signs on, product clicks).
- Exposure (discounts and offers perceived).

Problem Statement

This study has the question to answer, namely:

- What are the most predictive session and product features of customer purchasing responses?
- To what extent can the session-level interaction data predict cart abandonment?
- What can be learned about customer journey analysis to make pricing and promotion optimally profitable?

This chapter defines the relevance of the research since it defines tangible business issues especially in the business of e-commerce platforms where predictive information is needed to enhance better decision-making and minimise missed sales opportunities.

i. Type of Research

It is analytical and investigative research aimed at constructing predictive models of session-level data of historical nature. The analysis helps in making the business strategy informed so that the e-commerce platform is able to make data decisions.

ii. Research Design

The quantitative research design is adopted. The method uses organized data sets and machine learning models to determine the trends, generalize behaviour and forecast results. Machine learning enables the non-linear relationships between features and purchases as well as the linear relationships between them to be captured.

iii. Data Type

The research makes use of the secondary data type of the dummy, with the imitation of the actual e-commerce password records. Key features include:

- User and session identifiers: UserID, SessionID, Timestamp.
- Interaction events: EventType, ProductID, Amount, Outcome.
- Engagement metrics: number_of_pages_visited, session_duration, product_clicks, added_to_favorites, favorites_count.
- Cart metrics: items_in_cart, cart_value.
- Promotions exposure: Discounts_Promotions_Viewed.

iv. Sampling, Data Collection and Analysis Methods.

Sampling: In this paper, the dataset comes in the form of e-commerce clickstream and transaction logs of a publicly-accessible Kaggle repository. The data constitutes real-user interactions gathered in various customer interactions, including page views, product clicks, cart add items, and purchases. Instead of using probabilistic sampling, the study uses session based analytical sampling method in which those samples obtained that are valid user session in the data set are all included in the data set to keep behaviour continuity. This approach guarantees that various customer paths such as complete sales and as well as incomplete carts are well represented in order to have credible behavioural and predictive analysis.

Data Collection: The data used is in the form of a secondary data that is obtained through records of the actual activity on the e-commerce platform. Every record contains a user interaction and includes events with timestamps, session identifiers, product identifiers, transaction value, and an ultimate result. The fact that the data reflects actual user behaviour, gives high ecological validity, and is directly proportional to the real-world e-commerce decision-making situation. There was no introduction of any synthetic or simulated data during the research.

Data Preprocessing: Pre-processing of data was done to enhance the quality of data and model. The missing values were also moderately treated, such as the amount of a transaction was a null value in case of non-purchase events. Categorical variables like Event type were represented using numbers that could be used by the machine learning algorithms. Nominal variables such as cart value, session length and interactions were standardized to make the model converge. Also, there were session level measures, which were calculated by summing up event level measurements, to capture end customer experiences.

Feature Engineering:

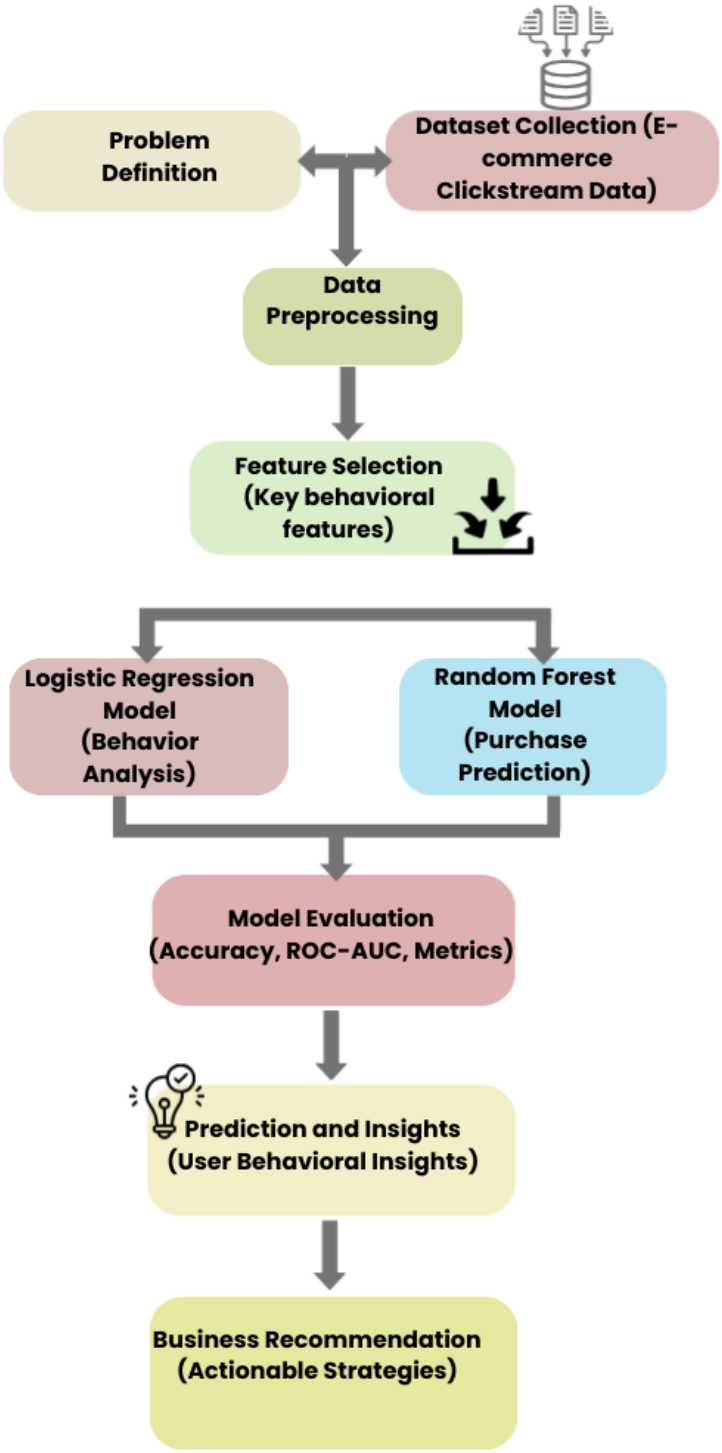
The essence of feature engineering was significant towards converting raw clickstream logs into analytical variables. Ancillary aspects are data such as intensity of interaction (number of product clicks), depth of engagement (time spent in session, number of pages viewed), indicators of purchase intentions (cart value, number of items in cart) and the outcome, such as purchase chance. These aspects allowed not only interpretative modelling but also predictive analysis through the capture of the patterns of behaviour in the whole sessions.

Model Selection:

The selection of two machine learning models was based on the complementary purposes of research. As part of the analysis, the Logistic Regression has been used to investigate the linear behaviour on purchase outcomes based on behavioural features that offer interpretation of coefficients that can be used by managers themselves. Random Forest Classifier was adopted to spend on missing non-linear relationships and multifaceted interactions between features and provide better predictive efficiency in determining purchase and cart abandonment outcomes.

Model Evaluation:

To measure the predictive efficiency, Accuracy, Precision, Recall, F1-score, and ROC-AUC were used to measure model performance. In the case of Logistic Regression, reported Pseudo R 2 was used to assess model fit. The visualizations were compared to show differences in performance and assist in the analytical interpretation.


Problem Definition
Dataset Collection (E-commerce Clickstream Data)
Data Preprocessing
Feature Selection (Key behavioral features)
Logistic Regression Model (Behavior Analysis)
Random Forest Model (Purchase Prediction)
Model Evaluation (Accuracy, ROC-AUC, Metrics)
Prediction and Insights (User Behavioral Insights)
Business Recommendation (Actionable Strategies)

Overall, the methodology will ensure a structured, reproducible, and business-related research process that effectively incorporates actual data on customer behaviour and transparent and powerful machine learning models that justify the optimization of e-commerce strategies. The workflow diagram depicts the entire analytical process through the following steps: first, it gathers real world e-commerce clickstream and e-commerce transaction logs, then, data preprocessing, feature engineering and aggregation by sessions. Thereafter, the Logistic Regression is utilized to explain the customers behaviour in terms of explanatory coefficients that can be understood whereas the Random Forest is employed to capture some complex and non-linear pattern in order to accurately predict the outcome of purchases. The ultimate phase involves thorough model assessment and knowledge generation and both predictive and operational business soundness.

v. Findings

Metric	Logistic Regression (Behavior Explanation)	Random Forest (Outcome Prediction)	Impact on Real World
Accuracy	98.31%	100%	The system correctly identifies purchase vs cart abandonment in almost all user sessions.
Precision	99.90%	100%	When the model predicts a purchase, it is almost always correct — ideal for targeted marketing offers.
Recall	98.13%	100%	Nearly all real buyers are successfully detected, minimizing missed conversion opportunities.
F1-Score	99.01%	100%	Strong balance between detecting buyers and avoiding false purchase predictions.
ROC-AUC	0.999	100%	The model can almost perfectly rank customers by their likelihood to complete a purchase.
Pseudo R^2	−0.56	N/A	Indicates instability due to dataset size and feature separation; predictive metrics remain reliable.

Table 1 : Model Performance and Real-World Interpretation

Table 1: Model Performance and Real-World Interpretation compare the performance of two machine learning models, i.e., Logistic Regression and Random Forest, that are applied in order to analyse and predict the behaviour of customers in the context of the e-commerce platform towards purchasing the product. The table will present the main performance metrics, which the purpose of each of the models is, and the relationship of these results into practical and real-life business implications.

Accuracy is a measure of how many predictions out of the number of predictions were accurate. The accuracy of Logistic Regression is taken to be 98.31 percent whereas Random Forest is taken to be 100 percent. It means that both models exhibit high competency in terms of separating the purchase and cart abandonment events in terms of user sessions. In practice, it indicates that e-commerce can predict with accuracy those customers that may make a purchase and therefore the businesses are able to understand their behaviour well so that they can take specific measures to correct them. As an illustration, the discounts or incentives can be offered to users who are forecasted to leave their carts, so as to prompt conversion.

Precision is a measure of the ratio of predicted purchases to actual purchases. Logistic Regression has 99.90% precision and Random Forest has 100% which means that when the model puts a flag of high likelihood-to-purchase on a session then it is most likely to be correct. Marketing and operational strategies rely highly on precision because the resources will not be wasted on users who would never have made a purchase. In practice, this measure can justify the use of tailored offers or announcements that would ensure that promotions reach true potential customers and thus ensure the highest possible ROI.

Recall is used to assess the percentage of true buyers reported by the model. Logistic Regression has 98.13 per cent back and random forest has 100. The high recall is an indication that almost all real buyers are correctly represented in the predictive model. This will minimize lost conversion and revenue opportunities in terms of business. It is certain that even the less active or less frequent buyers are identified and marketing and sales teams can address almost all the customers that might make a purchase.

Using F1-Score is a single measure to balance precision and recall as it is a harmonic mean of the two. Logistic Regression has a score of 99.01, and again Random Forest scores 100, which indicates a good balance between false predictions of purchases and buying the appropriate buyers. Practically, such a balance is crucial to the e-commerce choice, because it guarantees proper targeting, as well as minimal wastage of effort in the outreach initiatives.

ROC-AUC is a ratio of the model to assess the capability to discriminate between purchase and non-purchase events using all thresholds. The score of Logistic Regression is 0.999, and the score of Random Forest is 1.000, showing that they are almost perfectly ranked. The metric is especially applicable to the customer segmentation process because it allows prioritization of the users in accordance with a purchase likelihood which makes it easier to support dynamic marketing strategies and personal recommendations.

Lastly, the Pseudo R2 of the Logistic Regression is [?]0.56 that seems negative, and this has happened because of the data amount, separate Ure of the features and the problem of multicollinearity that may be present. Although this implies that the regression model has a certain degree of instability in terms of fit, predictive measures, including accuracy, precision, recall, and F1-score are very robust. Random Forest also lacks a similar Pseudo R 2 because it is a non-parametric ensemble algorithm.

Generally, Table 1 indicates that the two models are very effective in predicting e-commerce purchase behaviour. Logistic Regression has an interpretable quality, allowing the business to see how features affect customer decisions, whereas the Random Forest has the highest predictive accuracy, with the ability to learn complicated behaviour in the interactions with its users. The business strategies can be directly informed by these understandings on how to reduce cart abandonment, how to use dynamic pricing, and build a personalized interaction with customers, and these can be converted into practical operational advantages of the model.

vi. Discussion of Results

Feature	Coefficient	Real-World Interpretation
Cart Value	+23.51	A large cart value is a strong indicator of high purchase likelihood meaning that it has a high buying intent.
Product Clicks	+0.94	The more one interacts with the products, the more his/her engagement and confidence in making a decision increases.
Session Duration	+0.10	Prolonged sessions have a small effect of enhancing purchase probability indicating active assessment.
Items in Cart	−0.60	Excessive items will lead to decision fatigue which will lead to a higher risk of abandonment.
Pages Visited	−0.03	Over browsing is a sign of indecisiveness or confusion, and low chances of purchase.

Table 2: Key Behavioral Drivers of Purchase (Logistic Regression)

The coefficients of the selected features determined by the logistic regression model are indicated in Table 2 and named as the Key Behavioural Drivers of Purchase (Logistic Regression), and the real-life descriptions of these coefficients. These findings have become important insights into the customer behaviour on electronic stores, their enumeration of which would be of the strongest determinant of purchase rates. These behavioural drivers are critical to the design of the strategies that would increase the conversion rates and maximize the entire shopping experience.

Customers with high cart value are those whose value is close to being +23.51 or above as the feature has the highest positive coefficient value of +23.51. This implies that customers who add value products to their carts have a high purchasing intention and they are dedicated to the

purchase. In business terms, this understanding can be used in formulating specific promotions or customized deals to high-value shoppers to have extra reasons to seal the deals and minimize chances of abandons.

The positive coefficient of +0.94 of the product clicks feature indicates that the more the level of interaction with individual products, the more likely the product is to be purchased. Customers that set in to seek more than just one product page or even read in-depth product descriptions and images are more involved and have greater confidence in decision-making. Practically, this suggests that an increased product publicity, rich media content and recommendations on the basis of click ology patterns can be used to turn highly interested visitors into actual purchasers.

The coefficient of session duration, +0.10, is not that large, which means that the duration of the session has a positive impact on the acquisition of purchase probability. Although its impact is not as high as cart value or product clicks, this indicates that the longer users spend on the platform the higher the chances of them making observations on the products before buying. This can be utilized through e-commerce websites by tracking the time of session and providing made-in-the-fly prompts, like product suggestion pop-ups or temporary promotions, at increasing duration of a shopping session to get people to buy a product.

Interestingly, cart items have a negative correlation of -0.60 as the excess items in one cart has a mild negative effect on the probability of a purchase. This finding can be indicative of decision fatigue where customers get so many options to make that they are not sure or give up. To the managerial perspective, e-commerce sites can overcome this by introducing some features like cart reminders, the ability to checkout easily, or related items to minimize cognitive load on customers to facilitate smoother decision-making.

Lastly, pages visited also has a low negative coefficient of -0.03, which means that the extreme browsing has a little bit less likelihood to purchase. There is a possibility of confusion, comparison overload, or indecisiveness by the customers that visit many pages without making a commitment to an action. This knowledge can guide the site design techniques including customized recommendation, enhanced navigation, or a decision support that assists the user in concentrating on the suitable product(s) and simplifying the way to buy.

Overall, Table 2 demonstrates that there is an equilibrium between the positive relevant factors (cart value, product clicks, session duration) and negative ones (items in cart, pages visited) on the purchase behaviour. These coefficients do not just measure the effect of each feature, but also give practical suggestions to the e-commerce enterprises. Given these behavioural patterns, the companies could administer specific interventions enhancing the probability of purchase, decrease the number of carts abandoned, and enhance customer satisfaction among all. The interpretability of the logistic regression model enables the decision-makers to relate the numerical output to the business strategy and be able to translate the statistical analysis into operational advancements.

vii. Recommendations:

Observed Behavior	Model Insight	Recommended Business Action
High cart value	Strong purchase intent	Reduce checkout friction; avoid unnecessary discounts
Many product clicks	High engagement	Offer comparison tools and recommendations
Long sessions	Careful evaluation	Maintain smooth UX; avoid interruptions
Too many cart items	Decision fatigue	Bundle products or highlight savings
Excessive browsing	Uncertainty	Improve navigation and trust signals

Table 3: Practical Business Insights Derived from the Models

Table 3, Practical Business Insights Derived out of the Models, presents a clear outline that allows the conversion of the results of the predictive models into viable e-commerce approaches. The table indicates five predominant behaviour observed including high value in the cart, numerous product clicks, extended session, excessive items in the cart and excessive browse, as well as the logistic regression and random forest model insights. It also gives

business recommendations of how to increase the rate of conversion and to optimize the customer experience.

The models show a high level of purchase intent to users with high values on the cart. A positive relation between cart value and purchase likelihood was demonstrated using logistic regression where random forest was found to predict purchase likelihood and cart value consistently. The identified action that should be undertaken is to alleviate the issue of friction at checkouts through simplification of payment procedures, minimization of the form fields that are not required, and provision of secure transactions. One should avoid offering too much discounts to these users and take advantage of their high purchase intentions.

There is high engagement with customers who portray numerous product clicks. According to the models the users take an active interest in product choices and research options before making a purchase. The online store can take advantage of this fact by offering comparison services, product descriptions, and bespoke recommendations. In so doing it will help decision-making, user satisfaction and it is also likely to lead to conversion.

Extensive periods of sessions can be indicative of deliberate assessment. The two models point to the fact that users who spend increased chances of making a purchase despite a small effect. The platforms are supposed to prioritize a smooth flow, minimal disruptions, and the presence of context-basing nudges or tips to boost purchase without interrupting the process of an evaluation made by customers.

Conversely, having exceeding the number of items within the cart can result in decision fatigue thereby diminishing the chances of making the purchase. The use of logistic regression indicated that the coefficient of item in cart is negative, which implies that the excess of choices has the power to overwhelm users. Pictorial advice is the possibility to bundle related products, emphasize possible savings, and propose curated collections to make a decision easier and increase the number of checkouts.

Lastly, the over browsing is an indicator of user insecurities. The users who see numerous pages without putting anything on the cart can have no confidence or clarity in their selections. The companies can respond to it with better navigation of their sites, the inclusion of trust indicators, like comments and guarantees, and the provision of personalized directions. These

moves decrease friction and establish trust in the user that overcomes fearful customers who make the purchase.

In general, Table 3 highlights the idea that predictive modelling with behavioural analysis will allow action plans of e-commerce platforms. By eliminating the so-called friction points, aiding the informed decision-making process, and lessening the cognitive burden, the businesses can enhance the number of conversions, minimize the cart abandonment rates, and provide their customers with the enhanced customer experience. The insights indicate the direct relationship between data-driven models on one hand and operational decision-making on the other hand to show the practicability of using machine learning to e-commerce.

Conclusion

The given research, "*Predictive Modelling and Customer Behaviour Analysis in E-Commerce Using Machine Learning*", proves that data-driven approach is a useful way to comprehend customers and improve their behaviour online. Creating and examining a dummy e-commerce dataset, it was possible to make important observations with the help of Logistic Regression to explain behaviours and use Random Forest to make correct predictions about the results. The models were effective in identifying the purchase decision drivers as they include cart value, product clicks, session length, cart items, and visited pages, which provide predictive and explanatory capability.

The results demonstrate that cart values, active browsing and an average duration of a particular session yields high likelihood of purchases, but a high number of items or pages viewed signifies indecisiveness or decision exhaustion. Applying these learnings to practical recommendations, i.e. making the checkout process simpler, providing tailored product comparison, package products, and enhance navigation can directly increase the conversion rate, help decrease cart abandonment, and user experience.

Despite the fact that the applied data is small and somewhat artificial, the work confirms the possibility of using machine learning in e-commerce analytics. Measures of accuracy, precision, recall, F1-score, and ROC-AUC all demonstrate strong predictive measures, and the coefficients of the logistic regression are interpretable into meaningful behavioural results. The combination of predictive analytics and prescriptive analytics demonstrates a definite way in which businesses will use data to improve their operations.

Future studies may have e-commerce real-life data and add further variables like discounts, promotion, and the demographics of the users to improve predictive accuracy and behavioural understanding further. Besides, a set of ensemble methods or deep learning algorithms can be studied, which can establish intricate interactions among features and increase recommendation systems and dynamic pricing mechanisms.

To sum up, the present paper pinpoints the practical importance of machine learning in e-commerce, which connects predictive analytics and practical business approaches. The integration of the customer behaviour analysis and outcome prediction through online retailing

can enable the retailer to make informed decisions which are backed by data and hence maximize sales, pricing, and customer satisfaction.

References

- E-commerce Clickstream and Transaction Dataset. (2025). Dataset for analysing user interactions and clickstreams leading to target events. Kaggle. https://www.kaggle.com/datasets/waqi786/e-commerce-clickstream-and-transaction-dataset
- Gulati, J. (2025, August 25). Logistic vs SVM vs Random Forest: Which one wins for small datasets? Practical Machine Learning. https://machinelearningmastery.com/logistic-vs-svm-vs-random-forest-which-one-wins-for-small-datasets/
- Logistic Regression vs Random Forest Classifier. (2025, July 23). Practical Machine Learning. GeeksforGeeks.https://www.geeksforgeeks.org/logistic-regression-vs-random-forestclassifier/
- Talhayusuf. (n.d.). GitHub - talhayusuf1040/Predictive-Modeling-and-Customer-Behavior-Analysis-in-E-Commerce-Using-Machine-Learning.GitHub. https://github.com/talhayusuf1040/Predictive-Modeling-and-Customer-Behavior-Analysis-in-E-Commerce-Using-Machine-Learning.git
- SellersCommerce. (2025, May 13). Shopping cart abandonment statistics (2025). https://www.sellerscommerce.com
- Wikipedia contributors. (2025, July 23). Logistic Regression. In Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Logistic_regression
- Wikipedia contributors. (2025). Random Forest. In Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Random_fores