EasyChair Preprint

№ 10460

Characterising Online Purchasing Behaviour

Ganesh Gurram, Nikhil Reddy Kodumuru and Mukesh Tripathi

June 28, 2023

Characterising Online Purchasing Behaviour

G Ganesh
Department of Information Technology
Vasavi College of Engineering
Hyderabad, India
ganeshgurram000@gmail.com

K Nikhil Reddy
Department of Information Technology
Vasavi College of Engineering
Hyderabad, India
nikhilreddy024@gmail.com

Dr Mukesh Tripathi
Department of Information Technology)
Vasavi College of Engineering
Hyderabad, India
mukeshtripathi016@gmail.com

Abstract—This paper suggests a system that classifies customers based on their purchase behaviour and examines their patterns using supervised and unsupervised learning techniques. The technology analyses consumer behaviour at the session and user journey levels, anticipating consumer behaviour with excellent accuracy and recall rates. Five main client clusters with different behaviour patterns were identified by the study: "new customers," "active shoppers," "returning decisive shoppers," "comparison shoppers," and "window shoppers." These patterns can offer insightful information for developing personalised marketing tactics, enhancing the entire consumer experience, boosting client retention, and promoting business expansion. The study emphasises the value of client segmentation and its possible commercial advantages.

I. INTRODUCTION

The need for e-Commerce platforms has become widespread due to safety measures and recent trends, which demand a seamless shopping experience for virtually all products. As the pandemic has caused a shift in digital shopping experiences and social distancing, it is crucial to have scalable analytical frameworks that automatically analyse virtual shopping experiences. These frameworks can assist in shopping personalization, inventory management, and marketing for both customers and manufacturers. One of the primary reasons for inconsistencies in virtual shopping processes across products and platforms is the variation in product cost, delivery wait times, and platform usability. In this study, we aim to analyse customers' online shopping history using a consistent framework that can scale across different products and platforms. By doing so, we can identify patterns and trends over time, which may be indicative of specific shopping behaviours. The proposed system is designed to predict customer conversion from browsing to purchasing at both the session level and user journey level.

The objective is to establish a standardized workflow that includes feature engineering, feature selection, and predictive modelling for user-product interactions, specifically user journeys. Moreover, the system can use sequence modelling to predict when a purchase event is likely to occur during a session. Lastly, we aim to classify customer purchasing behaviour into five unique categories based on their actions of product viewing, carting, and purchasing. This approach can provide valuable insights to guide business and marketing intelligence.

Upon logging into a shopping website, customers typically accept cookies to initiate the session. This session is identified by a unique combination of session-ID and client-ID, which can be utilized to record information related to product-level browsing, additions to cart, removals from cart, and completed purchases. By consolidating the session level data, we can create user journeys and assess the likelihood of a sale per client for each product-type interaction. This approach was inspired by the methodology presented in [1].

In a previous study [5], the authors utilized timestamps of clicks within a clicking stream during a session to develop a model that predicted buying patterns. Bidirectional LSTM models were employed for this task, and the results showed that the accuracy from click stream sequences and LSTM models was similar to that obtained using engineered features and classification models. In the present work, we expand upon this analysis of predictive models at both session-level and user-journey level for purchase events. Specifically, we investigate the sensitivity of these models to online shopping portal and product-level features.

In addition to assessing sessions, previous research [6] highlights the importance of predicting repeat customers and their likelihood of completing their orders. This finding serves as a motivation for our analysis of user-journey level interactions. Specifically, we examine both session-level and user-journey level features to categorize user interaction clusters. Understanding these clusters can enable more accurate predictive modelling for purchasing events per cluster, which can ultimately improve our ability to gauge customer-specific demands.

II.PROBLEM STATEMENT – OVERVIEW

The objective of this project is to identify patterns and trends in shopping behaviour over time to predict customer

conversion from browsing to purchasing. Previous studies have analysed product demand and price variations at session-levels, but our proposed system takes it a step further by analysing user-journey level interactions to predict repeat customers and their likelihood of returning to finish their orders. By analysing both session-level and user-journey level features, we aim to categorize user interaction clusters, which can then be used to improve predictive modelling for purchasing events specific to each cluster. Ultimately, this will allow us to accurately gauge customer-specific demands and improve customer conversion rates.

III. LITERATURE SURVEY

Online purchasing behaviour analysis has been an active area of research in recent years. Machine learning has been employed to predict purchasing behaviour in real-time using online shopping data. In [1], Roychowdhury et al. proposed an Online Purchasing-behaviour Analysis using Machine learning (OPAM) system that utilizes machine learning techniques to analyse online shopping data. The system can analyse customer behaviour and provide personalized recommendations. In [2], Sakar et al. proposed a real-time prediction model that uses multilayer perceptron and LSTM recurrent neural networks to predict the purchasing intentions of online shoppers.

In [3] the authors used web usage mining techniques to analyse customer behaviour and made recommendations to improve the user interface of the website. In [4] E-commerce behaviour data from a multi-category store were made available as an open dataset for research purposes.

Neural modelling has also been used to understand buying behaviour in e-commerce. In [5], Wu et al. proposed a neural model that can predict purchasing behaviour based on clicking patterns. Charan Somboon and Viyanon compared different machine learning models to predict repeat buyers in a Kaggle acquired value shopper case study [6].

In addition, TPOT, a tree-based pipeline optimization tool, has been developed to automate the process of machine learning pipeline optimization [7].

Machine learning techniques have also been employed for churn prediction in telecom, where the objective is to identify customers who are likely to leave a service provider. In [8], Idris et al. used random forest and PSO-based data balancing along with various feature selection strategies to predict customer churn. Internal clustering validation measures have been used to evaluate clustering models in machine learning [9]. Dynamic clustering of histogram data based on adaptive squared Wasserstein distances has also been proposed [10].

IV. PROPOSED WORK

Our proposed system provides two key contributions. Firstly, we examine the sensitivity of online shopping portal and product-level features for session-level and user-journey level classification, respectively, in the context of purchase prediction. Secondly, we utilize unsupervised learning techniques to identify distinct user-behaviour clusters/categories.

A. FLOW DIAGRAM

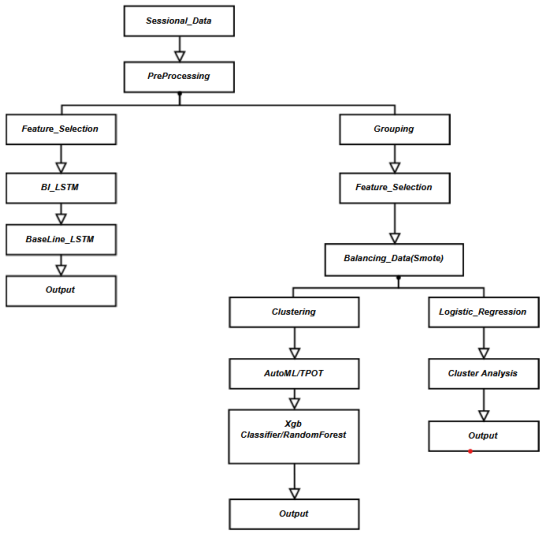

Fig. 1. Flow Diagram

Classification of User Session

We begin by utilizing a dataset with columns including 'event_time', 'event_type' (one of [view, cart, remove_from_cart, purchase]), 'product_id', 'category ID', 'category_code', 'brand', 'price', 'user_id', and 'User session ID'. This data is preprocessed and transformed to include new columns for user journey classification, such as 'NumOfEventsInJourney', 'NumSessions', 'interactionTime', 'maxPrice', 'minPrice', 'NumCart', 'NumView', 'InsessionCart', 'InsessionView', 'Weekend', 'Fr', 'Mon', 'Sat', 'Sun', 'Thu', 'Tue', 'Wed', 2019, 'Oct', 'Afternoon', 'Dawn', 'EarlyMorning', 'Evening', 'Morning', 'Night', 'Noon', and 'Purchase'.

Feature Ranking

we subject the user-session data to feature ranking. Upon exploratory data analysis and feature ranking using Random Forest, we observe little to no variance in the distribution of purchase journeys versus non-purchase journeys when measuring against date-time attributes. Our analysis highlights that features such as theTotalEventsInSession,interactionTime,NumTimesCartedInSession,NumTimesViewedInSession,maxPrice,minPrice,AvgAmtCartedInSession, AvgAmtViewedInSession,NumCategoriesCartedInSession,NumCategoriesViewedInSession,Purchase hold significantly higher weightage than features like date, time, or month of purchase. Thus, based on this analysis, we select the top-ranked 11 features with significant weights ranging from total interaction time, number of events, total carting and viewing time, to the maximum and minimum price range. Each feature is then scaled within the range of [0,1].

Balancing The Data

To address the issue of heavily imbalanced data, we employed a technique, SMOTE. SMOTE stands for Synthetic Minority Over-sampling Technique, which is used to balance the class distribution by under sampling the minority class.

We then Performed analysis using RNN (Recurrent Neural Network) algorithms Baseline LSTM and Bidirectional LSTM. (LSTM stands for Long-Short-Term-Memory).

Baseline LSTM and Bidirectional LSTM is implemented using keras.

Classification of User Journeys

We begin by utilizing a dataset with columns including 'event_time', 'event_type' (one of [view, cart, remove_from_cart, purchase]), 'product_id', 'category ID', 'category_code', 'brand', 'price', 'user_id', and 'User session ID'. This data is preprocessed and transformed to include new columns for user journey classification, such as 'NumOfEventsInJourney', 'NumSessions', 'interactionTime', 'maxPrice', 'minPrice', 'NumCart', 'NumView', 'InsessionCart', 'InsessionView', 'Weekend', 'Fr', 'Mon', 'Sat', 'Sun', 'Thu', 'Tue', 'Wed', 2019, 'Oct', 'Afternoon', 'Dawn', 'EarlyMorning', 'Evening', 'Morning', 'Night', 'Noon', and 'Purchase'.

Feature Ranking

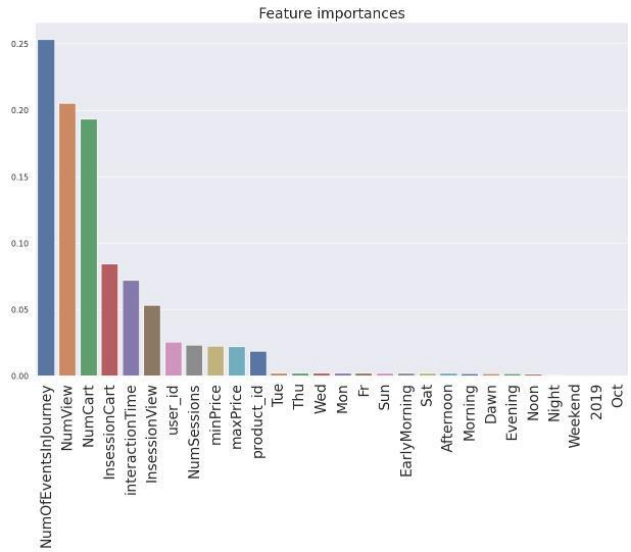

Fig 2. Feature Ranking Using Random Forest

Like the session-level data, we subject the user-journey level data to feature ranking. Upon exploratory data analysis and feature ranking, we observe little to no variance in the distribution of purchase journeys versus non-purchase journeys when measuring against date-time attributes. Our analysis highlights that feature such as the number of events in the user journey, total interaction time, number of sessions, number of carts, views, and removals hold significantly higher weightage than features like date, time, or month of purchase. Thus, based on this analysis, we select the top-ranked 11 features with significant weights ranging from total interaction time, number of events, total carting and viewing time, to the maximum and minimum price range. Each feature is then scaled within the range of [0,1].

Balancing The Data

To address the issue of heavily imbalanced data, we employed two techniques, SMOTE and PCA. SMOTE stands for Synthetic Minority Over-sampling Technique, which is used to balance the class distribution by oversampling the minority class. PCA, on the other hand, helps reduce the dimensionality of the data by identifying the principal components that explain most of the variance in the data. These techniques make training a model easier and help prevent bias towards one class. By balancing the data, the model will no longer favor the majority class simply because it contains more data.

we then performed data modeling and analysis using logistic regression.

Clustering of user journey data

Unsupervised clustering is conducted on the user-journey data using k-means clustering. The Elbow method, implemented through the Yellow brick Library in Python, is utilized to determine the optimal number of user-journey clusters (Q) that minimizes overall distortion score per cluster. The analysis identifies K = 5 as the optimal number of clusters.

Classification of user journey using AUTOML-TPOT

After clustering the data, we included a new column called "cluster" that corresponds to the cluster ID for each user journey row. This way, we can classify customers more effectively and the data is now more suitable for purchase prediction. We utilized stratified sampling to extract a sample of data from the user journey data, and then performed AUTOML-TPOT to determine the best set of hyperparameters and models that provide high accuracy, recall, and precision.

V. DATASET

Multi-Category Store Dataset

The multi-category dataset comprises behavior data of a large online store for seven months, starting from October 2019 to April 2020, and the total size of the dataset is about 9 GB. The dataset contains 42,448,764 rows and nine attributes. In this study, we focus on specific smartphones as our category code and extracted 309,934 rows related to these products. Each row in the dataset represents an event associated with users and products, and these events create a many-to-many relationship between products and users. The columns in the dataset are event_type, product_id, category_id, category_code, brand, price, user_id, and user_session.

VI. REQUIREMENTS

A. SOFTWARE REQUIREMENTS

Kaggle notebook
Operating system – Windows/Linux

B. HARDWARE REQUIREMENTS

Hard Drive – 16 GB
RAM – 12GB

VII. RESULT

We ran three sets of experiments to gauge the effectiveness of the three parallel modules that make up the proposed system. In the first group, we classified and predicted journeys and sessions that led to purchases in comparison to those that did not, using an ideal set of designed traits. The second set entailed applying an LSTM model to project session-level buying behaviour. The model offers a likelihood score that indicates whether a sale is likely to occur during the following session. To discover customer insights that can assist guide marketing and customer retention efforts, we lastly performed unsupervised clustering of user journey-level data.

User Session:

To determine the most effective model for two datasets comprising session-level data, we employed LSTM models with various layer and neuron topologies. There are 9.22% of sessions that result in purchases. To forecast the results of a sequence input, LSTM models are frequently utilised. We contrasted the performance of Bi-LSTM Feature models to "baseline" LSTM models to assess the significance of session-level features. The Bi-LSTM feature models contain information about product pricing and brand, whereas the baseline models consider the order of events in a session and the amount of time spent on each event. A sequence of events with a maximum of 100 occurrences, categorised as view, cart, remove from cart, or purchase, serves as the input for both models. The result is the likelihood of a purchase event occurring. The results for both types of models for the multi-category dataset are presented in a comparison table.

	Baseline LSTM	**Bi LSTM**
Best Optimizer	SGD	Adam
Accuracy	0.778	0.7792
Precision	0.389	0.6072
Recall	0.5	0.7792

Table 1. User Session Classification Performance

The study shows that the Bi-LSTM model, which includes additional features, performs better than the baseline LSTM models on a dataset with multiple categories. The higher accuracy and recall of the Bi-LSTM model are particularly relevant to the goal of ensuring that there is enough stock and inventory at the session level to satisfy customer demand consistently.

User Journey:

The user journey data we have is imbalanced, so we used binary classification with a 70/30 split to create models. We trained a Logistic Regression model on the unbalanced dataset, then performed sample balancing using class weighting and SMOTE, a technique that oversamples the minority class. The results of the classification performance on the smartphone dataset are shown in a table.

	Unbalanced	**Balanced**
Accuracy	0.9659	0.7589
Precision	0.7562	0.7503
Recall	0.2996	0.7763

Table 2. User Journey Classification Performance using Logistic Regression

We found that balancing the data significantly improved precision and recall for the classification. As the number of purchase samples is low in both datasets, we aimed to maximize recall to ensure that purchase events are not missed. This is important for predicting quantities for inventory and stocking, and to make sure customers have access to the products they want. False positive events lead to overstocking, which does not harm the customer shopping experience significantly.

Our main objective is to determine how accurately we can predict whether a user will make a purchase during a browsing session by analyzing various factors and the behavior of similar users. To achieve this, we will utilize sequence models to predict the probability of a purchase occurring in the next session. We will also employ classification models at a user-journey level using the TPOT library.

These are the pipelines using TPOT

→RandomForestClassifier(CombineDFs(CombineDFs(CombineDFs(input_matrix, input_matrix), input_matrix), input_matrix), bootstrap=False, criterion=entropy, max_depth=None, max_features=log2, min_samples_leaf=1, min_samples_split=2, n_estimators=400)

Accuracy	0.9856
Precision	0.9267
Recall	0.8924

Table 3. Using Random Forest Classifier

The model is a pipeline consisting of three estimators: FastICA, MLPClassifier and XGBClassifier. The hyperparameters for each of these estimators are as follows:

FastICA:
tol=0.4
MLPClassifier:
alpha=0.0, learning_rate_init=0.5
XGBClassifier:
learning_rate=0.01, max_depth=6, min_child_weight=20, n_estimators=100, n_jobs=1,subsample=1.0,verbosity=0

Accuracy	0.9973
Precision	0.9854
Recall	0.9514

Table 4. User Journey Classification Performance using Best Pipeline

Clusters:

Finally, we will examine clusters of user journeys to gain insights into customer behaviour, which can be used to inform marketing and customer retention strategies.

Cluster	0		1		2		3		4	
Feature	Purchase	No-Purchase	Purchase	No-Purchase	Purchase	No-Purchase	Purchase	No-Purchase	Purchase	No-Purchase
NumOfEventsInJourney	5.99	1.56	13.73	4.63	21.51	6.8	12.12	4.25	15.71	5.18
NumSessions	1.56	1.08	4.52	2.84	7.5	4.24	3.93	2.57	5.33	3.19
InteractionTime	14590.29	3638.27	800264.68	799326.23	2055509	2027766.24	352939.39	354794.8	1345301	1340321.18
maxPrice	434.75	504.25	509.51	527.83	594.67	593.98	453.7	490.68	540.28	554.24
minPrice	434.31	504.13	498.6	515.33	572.5	572.44	447.75	483.27	525.06	537.77
NumCart	1.21	0.02	2.34	0.11	3.48	0.15	2.03	0.11	2.65	0.12
NumView	3.6	1.53	9.28	4.51	14.48	6.64	8.28	4.13	10.61	5.05
InsessionCart	1.4	0.15	1.89	0.31	2.11	0.31	1.75	0.32	1.94	0.31
InsessionView	5.83	9.21	9.05	13.04	9.91	13.65	9.01	12.98	9.24	13.31
Cluster Definition	New Shoppers		Quick/Active Shoppers		Returning Decisive Shoppers		Comparision Shoppers		Window Shoppers	

Fig 3. User Journey based cluster definitions

Cluster 1: New Inquisitive Shoppers
This cluster represents customers who have recently joined and are actively exploring the store's offerings. They are open to trying new things and are likely to experiment with different products. However, they may not be very decisive about their purchases yet.
Cluster 2: Quick/Active Shoppers
This cluster has relatively high values for most features, particularly in terms of the number of events in a customer's journey and the number of sessions they have. These customers are likely to be more decisive and engaged.
Cluster 3: Returning Decisive Shoppers
This cluster represents customers who are loyal to the store and have made multiple purchases in the past. They have a clear idea of what they want and are generally decisive about their purchases. They are likely to be regular shoppers and may have specific preferences.
Cluster 4: Comparison Shoppers
This cluster represents customers who are decisive about their purchases but may not be as loyal to the store. They are likely to do research before making a purchase and consider factors such as price, quality, and practicality. They may switch between different stores based on their needs.
Cluster 5: Window Shoppers
This cluster has relatively low values for most features, apart from in-session view events, indicating that these customers are browsing the site but not engaging in many other actions.

VIII. CONCLUSION

The focus of this study is on developing an analytical framework that can perform predictive modelling on session-level and user-product interaction journey levels using automated modules for feature engineering and feature selection. What sets this work apart is its emphasis on analysing user-journeys to predict purchase events and categorize customers based on their purchasing patterns. This approach is novel and can provide valuable insights for businesses looking to improve their marketing and sales strategies.

Our experiments led to three significant conclusions. Firstly, predicting purchase events at the session level has lower accuracy (75-80%) and precision (58-65%) compared to user-journey-based classification, which had an accuracy and precision of 97-99%. This result is not surprising since user journeys provide repeated interaction information, which better represents the intent to purchase than any session. Therefore, predicting purchase events at the user-journey level is more accurate and useful for stocking and inventory purposes.
In our study, we used clustering techniques to analyse user-journey data and identify patterns in purchasing behaviour. We found five distinct clusters for both data sets and analyzed feature trends to distinguish between purchase and no-purchase events. While the majority of user-journeys (>90%) belonged to clusters representing "New Shoppers" with a higher tendency to research rather than purchase, there were other minority clusters that demonstrated varying degrees of engagement and purchasing intent. This analysis can help businesses make targeted recommendations and follow-ups, such as marketing nudges, coupons, and offers, to improve personalized online shopping behaviour. In the future, we plan to extend our framework to larger and more diverse datasets and use sequence modelling at the cluster level.

REFERENCES

[1] S. Roychowdhury, E. Alareqi and W. Li, "OPAM: Online Purchasing-behavior Analysis using Machine learning," 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 2021, pp. 1-8, doi: 10.1109/IJCNN52387.2021.9533658.

[2] C. O. Sakar, S. O. Polat, M. Katircioglu, and Y. Kastro, “Real-time
prediction of online shoppers’ purchasing intention using multilayer
perceptron and lstm recurrent neural networks,” Neural Computing and
Applications, vol. 31, no. 10, pp. 6893–6908, 2019.

[3] C. J. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M. J. del

Jesus, and S. Garc´ıa, “Web usage mining to improve the design of an ecommerce website: Orolivesur. com,” Expert Systems with Applications,
vol. 39, no. 12, pp. 11 243–11 249, 2012.

[4] M. Kechinov:, “Ecommerce behavior data from multi category store,”
https://www.kaggle.com/mkechinov/ecommerce-behavior-data-frommulti-category-store, 2019.

[5] Z. Wu, B. H. Tan, R. Duan, Y. Liu, and R. S. Mong Goh, “Neural
modeling of buying behaviour for e-commerce from clicking patterns,”
in Proceedings of the 2015 International ACM Recommender Systems
Challenge, 2015, pp. 1–4.

[6] T. Charanasomboon and W. Viyanon, “A comparative study of repeat
buyer prediction: Kaggle acquired value shopper case study,” in Proceedings of the 2019 2nd International Conference on Information Science
and Systems, 2019, pp. 306–310.

[7] R. S. Olson and J. H. Moore, “Tpot: A tree-based pipeline optimization
tool for automating machine learning,” in Workshop on automatic
machine learning. PMLR, 2016, pp. 66–74.

[8] A. Idris, M. Rizwan, and A. Khan, “Churn prediction in telecom using
random forest and pso based data balancing in combination with various feature selection strategies,” Computers & Electrical Engineering,
vol. 38, no. 6, pp. 1808–1819, 2012.

[9] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, “Understanding of internal.
clustering validation measures,” in 2010 IEEE International Conference
on Data Mining. IEEE, 2010, pp. 911–916.

[10] A. Irpino, R. Verde, and F. d. A. De Carvalho, “Dynamic clustering of
histogram data based on adaptive squared wasserstein distances,” Expert.
Systems with Applications, vol. 41, no. 7, pp. 3351–3366, 2014.