E-Commerce Customer Purchase Behavior Using Data Mining Prediction

Maryam Arif, Dr. Mubasher H. Malik, Dr. Hamid Ghous

Abstract— Customer Behavior is an important factor that helps the retailers to predict about the sale of the products and maintain cost efficiently, especially in e-commerce. Prediction and assessment of customer behavior has become one of the major issue for the researchers. Many of algorithms has been applied to predict the customer behavior on e-commerce platform. In this paper, we have analyzed past work and summarized it in one place to help the researchers in future. So, they can get all information and work on the limiting factors to enhance the accuracy of models. We have analyzed the previous work into three categories: (1) deep learning models (2) machine learning models (3) hybrid models. Limitations and future scope of previous literature has been investigated.

Index Terms— Customer behavior, Customer purchase intention, Data mining, Deep learning, Machine learning, E-commerce, Prediction of products

1 Introduction

Customer behavior towards the products is the most focus point of view for retailers as it will be helpful in their businesses. Customer behavior shows the choice and need of the customers. It suggests the customer purchasing intent that is useful for business in online and offline retailer markets. As the digital systems and internet are overtaking the world and our life, the trend of online shopping is increasing with the passage of time. In these virtual markets, the need of accurate prediction of customer intention has become more important.

E-retailers target the greater audience than a local supermarket that has more diverse preferences for buying. But to grasp the maximum business, they need to predict the accurate customer behavior about their choices and likings. If e-retailers can better predict the choices of the customers, they can save cost and can sell more trending products.

They can also estimate the production amount of the products according to user need that can help them in saving the storage space. Online questionnaires helped the retailers to study the customer behavior and their choices but it is not much helpful. To this problem, data mining has provided much better techniques and algorithms.

Data mining is a growing field of computer science nowadays. Interest of researchers has been shifted towards data mining due to its efficient results about predictions and sentimental analysis which are near to the reality.

2 Background

Customer behavior is the attitude or manner of a consumer towards the products in the market. Customers' behaviors show their interest in particular products. Customer behavior towards the product depends upon the pattern these products appear before them and highly impact their purchasing decision. Customer behavior understanding has great importance for the retailers as they need to produce products according to the customer choice and need.by studying the customer behavior, e-retailers can predict what influences customers for their shopping decisions. Understanding customer buying behavior is the key secret in reaching and fascinating your clients, and directing their intention to buy from you. Today, the world is changing and becoming more dependent on technology. In such hasty days, E-commerce and online shopping facilitates the life of each and every person. Now, there are many e-commerce platforms that attract the user's intention towards themselves and provide their desired products like amazon, eBay and AliBaba etc.

Amazon has become an international platform which captures the customers around the world. E-commerce has become the world's most growing business but Amazon.com gains most of the popularity among such platforms. Like [1] put forward the concept of the global e-commerce village that has been settled by Amazon.com. It has been owned by Amazon.com, Inc. This popular company has become one of the largest companies in America like Google, Apple, Microsoft and Facebook. It is well known for working on the variety of trending fields of computer sciences and technology like e-commerce, cloud computing, digital streaming and artificial intelligence. They have integrated the techniques of data mining in their e-commerce business. Their recommender system is the key secret of their success. Like [2] has explained how amazon.com product recommendation system is continuously becoming more and more predictive with respect to the choices of each and every person.

- *Maryam Arif is currently pursuing M. Phil degree program in computer science in Institute Southern Punjab, Multan, Pakistan. PH- +92 3358272977. E-mail: maryamarif.ma@gmail.com*
- *Dr. Mubasher H. Malik is currently working as assistant professor in the department of Computer Science at Institute of Southern Punjab, Multan, Pakistan. PH- +92 301 8630005. E-mail: mubasher@isp.edu.pk*
- *Dr. Hamid Ghoush is currently working as assistant professor in the department of Computer Science at Institute of Southern Punjab, Multan, Pakistan. PH- +92 315 6098599. E-mail: hamidghous@isp.edu.pk (He did his PHD from University of Technology Sydney.He got more than ten years of research experience from oversease and Pakistan.)*

On the other hand, Alibaba.com is rapidly gaining popularity in e-commerce. It has been recently founded by Jack Ma who focuses on integrating artificial intelligence in e-commerce business. [4] has explained the great achievements of Alibaba.com specially Taobao. As online shopping is increasing rapidly, the need to understand customer behavior on virtual platforms has become more important yet more difficult. To grasp more customers, retailers on online shopping platforms need to show their products to the targeted customers who need those products. It will help them to estimate the amount of production of products according to the customer demand. They can save their cost to invest in other products. They can also estimate the profit from the predicted sale of the products. So, to understand the customers' behaviors and their purchasing intention matters a lot. If a customer cannot find his desired products or random products appear in front of them, they do not buy such products.

Some users like to scroll a lot but check products just for time consumption. Such an attitude does not help retailers to promote businesses. In other cases, if different quality and cost products have been shown to the customers then it is more likely that customers ignore those products. So, to target active buyers, it is necessary to suggest the products according to the customer choice. As [4] have shown that how customer loyalty and satisfaction depend upon the customer purchasing intention and correct customer behavior analysis. By behavior analysis and purchase intent, their research has proved 56% more customer loyalty.

So, understanding the customer behavior and predict their purchasing intention has become a hot topic and many researchers provided different solutions for this problem. Nowadays, data mining, natural language processing, image processing and Big Data has become one of the most modern techniques to solve such problems. Many researchers have tried to predict the customer's purchase intention using data mining techniques. Like [5] verified this fact that data mining procedure answers our problems more quickly and accurately than any other process. [6] also provided the significance of data mining techniques in the better understanding of customer behavior in the banking sector. [7] has described the effectiveness of using data mining techniques in the business growth and for making marketing strategies. There are many algorithms in data mining for predicting the sales of e-commerce about a product. [8] has provided the list of efficient algorithms for forecasting the sales. It also helps in predicting the customer behavior on e-commerce platforms. [9] has investigated for mitigating the adverse effect of customer behavior.

Data mining has two forms that are machine learning and deep learning. Both of these share the same base. Like, both of these techniques use the existing information to draw the patterns that will be helpful for making decisions. In deep learning, patterns are extracted and decisions are made about the things or persons, like recognizing the sales pattern or any type of fraud. But in machine learning, extracted patterns are used in training the machine, like computers or developing an artificial intelligent system. Deep learning systems involve more human interactions while machine learning systems are mostly automated and do not depend upon human interaction.

3 Literature Review

As the data minin methods are capturing the attention of each researcher for solving the issues like prediction of customer behavior, we have analyzed the literature with such framework that either have used deep learning methods or machine learning methods. Some researchers also have used both of the algorithms that lie in hybrid model category. The structure of this research has been shown in Fig. 1.

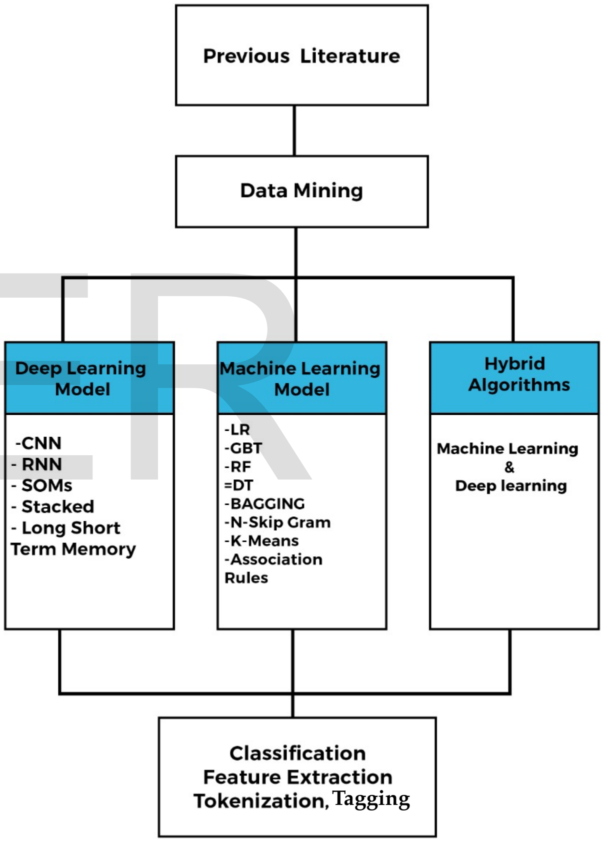

Fig. 1. Strucutre flow

This flow diagram has shown methods that are considered in this review. The detailed studied has been discussed in the following sections:

3.1 Deep Learning Methods

[10] have predicted the customers demand for a product by considering by marketing publicity and reviews about the products. They have used dataset of electric products from Amazon.com. To attaract the customer intention, count of reviews,

quality of reviews, charge-free deliveris has been investigated. They have applied a neural network to predict the factors influencing the customer demand of electronic products in an online environment. They have made a three-layer neural network. They have assigned 0 and 1 to starting weights and biases. To adjust the weight and They have used the Delta rule to adjust the connection weight and reduce the output errors. To train the algorithm, they have used a multilayer ceptron per training neural network. Then they have implemented a sigmoid function. They have concluded that online reviews and publicity of products attract customer's intentions more.

[11] have designed Deep Intent Prediction Network (DIPN) to anticipate the customers intent of purchase. They have designed two types of datasets: touch-interactive information by the Taobao app and browse-interactive behavior from Alibaba. To pre-process the data, they have to truncate the sample data and add labels. Their model has used an embedding lookup layer with a fully-connected layer. Then they have implemented a bifacel recurrent layer to remove dependencies. In the end, they have implemented multi-task learning. They have compared the results of their model with GBDT, RNN and DNN. their model provided 18.96% improvement for predicting the items for a customer purchasing behavior.

[12] have suggested a novel approach that can predict the customer demand for service-oriented products manufacturing. They have used the structural equation model for structuring the link between index of customer satisfaction and the influencing factors. Then they have implemented the least square support vector mechanism to reduce the adverse effect and predict the customer demand. To pre-process the data, they have applied normalization and then implemented KMO test and Bentley's sphere test for validating the prime factors of the dataset. They have compared their model with ANN, ARIMA and grey theory-based CDP models. By comparing with all models, the LSSVM model has shown less error than others by only having 0.263% of mean relative error.

[13] have analyzed the unstructured data for recognizing the purchase intention pattern of a customer on an e-commerce platform. They have collected the data from the voice recordings of customer interactions, and chat transcripts about cars collected from social media. First, they have created semantic annotation for the dataset. Then they have pre-processed the data by cleaning, POS tagging and filtering techniques. Then they have applied deep learning methods that are (1) CNN, (2) Recurrent Neural Networks (RNN), (3) Long Short-Term Memory (LSTM) and (4) the Gated Recurrent Units to attain desired results. They have predicted highly accurate customer behavior.

[14] has predicted the customers behavior and their purchase intention using deep learning methods and neural networks. Neural network has consisted of three layers: (1) an input layer, (2) hidden layer and (3) output layer which are connected with different network architecture. They applied RNN, LSTM and Word Embeddings. They have developed a system which has two hidden RNN layers by LSTM cells. They have used demonstrating "open research" data that is available on the internet. Their experimental results show the 99% correct prediction results that suggested that their model has been overfitting. Table 1 has provided the detailed investigation of previous frameworks based on deep learning methods.

3.2 Machine Learning Methods

[15] have predicted a new approach for proposing the demanded product for a customer. They have considered the different price of the same product sold by different sellers for proposing the demand of that product. They used datasets of e-commerce companies in Turkey mainly from n11. First dataset is preprocessed by filling the missing values, removing the attributes that have major missing values and removing the irrelevant attributes. The selling information of a particular product to same or different customers is also categorized and low outliers are also removed. They have used the Stacked generalization method and compared it with Random Forest (RF), Decision Tree (DT), Linear Regression (LR) and Gradient Boosted Trees (GBT). It is observed that the Stacked Generalization method provided the accuracy as much as other techniques provided.

[16] analyzed the forecasting of sales on e-commerce by appling market basket analysis with association rules. The data of supermarkets is collected from Vancouver Island University website. The dataset is prepared by converting purchased or not purchased data into numeric form then rearranged by NumericToNominal, minMetric and lowerBoundMindSupport parameters are also corrected. Visualization is implemented to better understand the dataset. They have used three data mining procedures to propose their study that are clustering, classification and association. They have tested Apriori and FP Growth algorithms for implementing Association rule algorithm but Apriori algorithm did not provide any satisfied result. So, FP Growth algorithm is best suited to their study by providing 21.06 Conviction and 1 (100%) confidence value.

[17] has investigated the repeating purchase behavior of customers on e-commerce platforms by a multisource information fusion and its impact on precision marketing by merchants. They have worked on the dataset collected from Tmall. Data is preprocessed by cleaning the past data of customers, removing erroneous and missing values to get a balanced dataset. Features are created by hidden layers of buyers, sellers and their interaction. This well-balanced sample dataset is used to run DABiGRU model. Then DeepCar boost and Deep GBM are used on the trained dataset to predict repeating behavior of buyers. These models are assembled with vote-stacking techniques. Proposed model provided the accuracy 91.28% and AUC 70.53%.

[18] predicted the non-linear purchasing behavior of customers using data mining methods. They considered market factors for remanufactured products in e-marketplaces. They used structured and unstructured variables of a real-world Amazon dataset to test this model.

TABLE 1
RESEARCH PAPERS USING DEEP LEARNING METHODS

Title	Authors	Preprocessing	Methods	Dataset	Results	Limitations/Future Direction
Predicting consumer product demands via Big Data: the roles of online promotional marketing and online reviews [10]	AYL Chong, E Ch'ng, MJ Liu, B Li	- cleaning the dataset - assigning 0 and 1 values	- Delta rule - Neural network	Electronic data set of Amazon.com	online reviews and promotional marketing strategies attract customer's intentions more	- large dataset size needed to examine - multiple categories of products must be investigated
Buying or Browsing? : Predicting Real-time Purchasing Intent using Attention-based Deep Network with Multiple Behavior [11]	Long Guo, Lifeng Hua, Rongfei Jia, Binqiang Zhao, Xiaobo Wang, Bin Cui	- truncate - tagging and labeling	- GBDT - RNN - DNN	Alibaba	18.96% improvement	- need to consider intra-view and inter-view attention practically for better performance of the model.
Customer demand prediction of service-oriented manufacturing incorporating customer satisfaction [12]	Jin Cao, Zhibin Jiang & Kangzhou Wang	-normalization - KMO test and Bentley's sphere test	-structural equation model - least square support vector mechanism - LSSVM model ANN, ARIMA	Online questionnaire	0.263% of mean relative error	- need to work on variety of datasets
Customer Perception Analysis Using Deep Learning and NLP [13]	Sridhar Ramaswamy, Natalie DeClerck	- tagging - filtering - text cleaning	- CNN - RNN - LSTM - the Gated Recurrent Units	Data from the voice recordings of customer interactions, and chat transcripts collected from social media	They have predicted highly accurate customer behavior	- more corpus is needed to develop
Predicting Process Behaviour using Deep Learning [14]	Joerg Evermann, Jana-Rebecca Rehse, Peter Fettke	-text cleaning	-RNN -LSTM -Word Embeddings	Demonstrating "open research" data	Shows prediction result 99% which proof the overfitting	- model is overfitting

For structured data, they preprocessed data for missing data values. Variables for more than 5% data missing, imputation procedures are applied while variables with less than 5% missing values are removed. Categorical variables are converted to 1 and 0. Outlier treatment is done by standard deviations and Univariate outliers is treated by box-and-whisker plot and winsorised. Unstructured textual data is preprocessed by tokenization, stopwords filtering and part-of-speech (POS) tagging.

Machine-learning and Data Mining (CRISP-DM) framework are implemented along with the CART model, M5 model tree and RF model of Regression tree model descriptions. Divide and conquer strategy is used in the Cart Model to identify a predictor and break-point value. So, the tree node of training space is splitted into a homogeneous subset by binary splitting. Cart Model results are improved by the M5 Model in which the LR model is implemented to foresee the values in every node. RF Model enhanced the performance by using a bagging ensemble method with multiple base classifiers. K-fold cross validation is used to test the results. Evaluation of this study shows that RF Model is best among other models used in this study by providing lowest values of MAE and RMSE and highest value of R2.

[19] have predicted the Best Seller Rank by using linear regression algorithms. They have collected hourly sales data from Amazon by using a web scrapper developed on python. They have used visualization of the correlations technique for observing the patterns through graphs. Then they have to classify the dataset into groups on the basis of a period of three days for analysing the patterns via linear regression method. They have obtained 69.46% value of R-squared. It provided a rough estimate about BSR.

[20] predicted the stock difference for some products on the basis of their sales available on different e-commerce platforms used in Turkey. They have collected data from the e-commerce site of Turkey that is Hepsiburada. They prepared their data set by running basic preprocessing techniques. To predict the stock difference, they have used Random forest technique with bagging algorithm and gradient boosting with boosting algorithms. CART (Classification and Regression Tree) algorithm is implemented for making decision trees. After running root mean square test, it is observed that price is the main element that determined the stock difference on these online platforms.

[21] have investigated the influence of items resemblance during the purchasing session and analyse their model by common recommendation method. First, they have collected session logs from RetailRocket dataset. At the pre-processing stage, sessions of less than three interacted items and items with short durations are removed. They have used the N-skip gram method of the Word2Vec model to train their model. Then they calculated the similarities using Word2Vec. In the third stage, they filtered out attributes and generated the features. On the gained attributes, they have solved the class imbalance problem using under sampling class imbalance and SMOTE methods. In the last stage, they have tested their model using three Ml Models: bagging, random forest method and decision trees. In experimentation, Bagging, when both similarities between fl, and the fm are included as features, they have gotten the 88.6% of F1 score while RF and DT has provided the F1 score of 81.8% and 87.6%.

[22] have studied about the methods to predict the customer behavior and future purchase pattern. They have applied machine learning algorithms of Association Rules, K-Means, Random Forest classifiers, and Survival Analysis. They have analysed their model by applying to Flowserve's customer purchase history. They have obtained a cleaned data set without missing values by using preprocessing methods. They have analysed the date of purchase, sell price, and item category for their study. Rule generation method has generated the set of items and their related rules like combinations of items. RFM analysis has consisted of clustering, classification, and association rules which consist of customers' marketing levels and provides future prediction. Logistic regression has applied to predict the relationship between one binary variable, in this case churn. Decision tree is applied on the resulting data set from Naive Bayes classifier, SVM and K-neighbor classifier. The predicted accuracy after optimization on the test data set is 84%.

[23] have used a data driven approach for predicting consumers' online purchasing preferences using their online lifestyle. They have built online lifestyles lexicon having 7 different dimensions with the help of text mining approach to learn consumer behavior. They have constructed lifestyle lexicon of customers online that has seven different characteristics by applying text mining approaches that considered consumers' textual behaviors. Their online lifestyles corpuses are efficiently filtered out using online review on Amazon. Then they applied collaborative filtering techniques to predict the recommended products for customers. They have applied these preprocessing techniques for obtaining clean datasets: tokenization, normalization, stop words removal, and stemming for text on lexicon and Amazon reviews. They have tested their model by experimenting it with RMSE, Tukey's HSD and analysis of variance (ANOVA). Their experimental results suggested that lifestyles of customers on online platform and all of its sub-dimensions can increase the predicting products power for customers purchase.

[24] have investigated the information diagnosticity for predicting the consumer behavior and their purchase intention. They have gathered the dataset from conducting a survey among the English Speaking Amazon.com customers that are recruited on Amazon's MTurk. They have constructed their model on some assumptions. To preprocess this survey data, they have converted the result of surveys in numerical form to test it. They have tested it with Root Mean Square values and bootstrapping. Their results suggested that online reviews about products and their combination play an important role towards purchasing behavior.

[25] have investigated the upcoming activities and their related products for targeting their customers to suggest these predicted products for their purchase. They have used the

Query intent method on product reviews of Amazon.com. They have targeted two entities: activity and audience. They have used the Wikimedia API to collect such data of activity and audience. Then they collected customer reviews for activity related products from Amazon. Then they preprocess the data into three stages: (1) punctuation deletion, converting all the word to lowercase, tokenization, applying stem unigrams; (2) pair up the intent aliases, annotating the review with the respective attributes; (3) combining the all pairs in a ASIN's reviews. on the training dataset, they implemented convolutional neural networks with mapping that followed by classification. To predict the search queries, they have implemented PySpark word2vec. Then they calculated the entropies for testing their model. Their experimental results show 95% accuracy for Activity datasets better than the Audience data whose accuracy is 90%.

[26] have analysed the classhing reviews on online platform with respect to their effects on customers behavior and the sales rank on Amazon.com. They have gathered the data of 6816 items from the shoe category of Amazon.com during Spring 2017. They have designed two hypotheses for their model that are: (1) effect of valence has higher impact on low and high range than on medium range (2) purchasing similarity is affected by the volume for medium and high range of valnce while low range is unaffected. First, they cleaned the dataset then classified the dataset in operationalize independent and control variables. Then they calculate the root mean square value and f score values of their model. Their second hypothesis provided good results with a value of 0.35 for root square and high valence.

[27] have combined reconstruction of phase space that is PSR technique and LSSVM to predict the future demand of for service-oriented products (SOM). First, they constructed the prediction sample space by the PSR to increase the time duration of the limited sample. After that they have improved the generalization and trained the LSSVM by the hybrid polynomial and kernel of radial basis function. At the end, they optimized the key factors of the LSSVM by implementing the particle swarm optimization algorithm. To test their model, they collected the dataset for the demand of air conditioner compressors in 2005 to 2012 for fortelling the demand in 2013 from an online platform in Shanghai, China. Then they first normalize the dataset in preprocessing. They have tested their model by running a comparison with ANN-based approach, grey theory-based approach, RE, the MAPE and RMSE. Their proposed model has showed the RMSE results of 0.1076 value and MAPE results of 0.572%.

[28] have designed a better system that can efficiently analyse the shopping behavior and predict the products to a customer. This system has tried to overcome shortcomings and challenges of previous shopping behavior prediction methods for an oonline platform. They have used linear model logistic regression and decision tree that are based on XGBoost models. But after optimizion, they have suggested the nonlinear model can for better results. First, they have combined the single model then implemented a fusion algorithm to link the foresee results. This approach has helped them to ignore the accuracy of the linear model for easy-fit and the DT model for over-fitting.

[29] have presented a system of decision support for allocating the retail price and revenue of high-end luxury retail products by predicting customer demand. They have collected sales data of 2.5 years of the retail stores from 45 distinct areas that provide sales purchase facilities of such products online. They have applied the algorithm based on regression tree or random forest-based machine learning techniqeu for anticipating the demands of customers on weekly based. They have considered prices of products, days of holidays, discounts, inventory and other influencing parameters for decision making. Then they have used interdependencies demand and price that are modified and then integrated into the model that has been built on integer linear method to allocate the best price. Branch-bound and branch-cut methods with root node have been implemented to anticipate the revenue from a product. Heuristic methods have also been used to further optimize their model. Their model has given the minimum 23.6% of MAPE.

[30] have investigated the probability statistics by user clicking behavior data for predicting the customer purchase intention on e-commerce platforms. To test their model, they have used the dataset from competition of Ali mobile in 2015. It has recorded the clicking behavior of users in huge amount. They have recorded each behavior for one click of a user to a specific product that also have client's ID, product ID, related category, intent of purchase type and time specifications. Behavior type tells whether a person is only browsing the product, adding it to cart or purchase it. Then they tagged the behavior of each user with an item. Then they classify the user behavior with purchase and non purchase items. Then they predicted the purchasing intention of the customer by the number of same users clicks on that specific item. They have used Bayes Model using classifiers of Bernoulli naive bayes, multinomial naive baye and poisson naive bayesx for predicting the customer intention. Then they apply F-measures to test their dataset. They have compared their model with User-based collaborative filtering that provided 0.099% precision with Bayes approach while their model provided 13.4% precision with one dataset and 85.9% precision with another dataset.

[31] have investigated the textual description of the products for predicting the customer purchase intention for the particular product. They have used 90,000+ product specification from Rakuten, that is Japanese famous online marketplace, to predict which keywords in product description attract more customer intention. They have suggested a novel approach using neural network architecture with the purpose to control for influencing parameters. They have extracted features from these specification textx and fit into a statistical model. To preprocess the data, they have segmented the data into tokens and selected such tokenz which are helpful for foretelling the high sales. Then they have divided the dataset into subsets then apply Odds Ratio (OR) to compare one copura with another. After

that they have implemented Mutual information to make correct classification decisions. Then they implement Lasso Regularization for selecting a variable to implement a linear regression model. In the end, they implemented a feature extraction method for prediction. Their experimental results suggested that correct use of keywords, appeals to authority, kind words and brief explanation of product attracts the most customers.

[32] have implemented both Supervised and unsupervised methods to predict the products for online customers depending on their behavior or search. They have classified the customers' behavior and check the relative behavior with existing customers' manners. Then they have used the supervised method to produce a new pattern related to the browsing history of customer. This pattern can also be used for other customers with the same behavior pattern. Their model generated the Ontology Language (OWL) file for each new customer that stored the relevant search data in history which have been utilized later for predicting the recommended products for that customer. Decision Support System has been implemented for predicting the products for that customer and for other customers with similar behavior patterns. To test their model, they have collected a large user dataset from ICS – Machine learning dataset. By experimental results, their model has shown 54.6% accuracy to predict the products for customers.

[33] have suggested such a support system that can efficiently suggest the products online according to customers choice. This system has included these modules: acquiring the information, transformation and integeration of knowledge in to the model. They have used the gaining knowledge module to collect fuzzy information from each review via sentimental analysis. Information transformation module has been implemented to transform gained fuzzy text into LINCs. Then they have implemented the integration module for obtaining the overall LINCs for each product to get the ranking of products. They have used the dataset of reviews that has been collected from Taobao.com under the skin-care category to test their model. They have preprocessed the dataset by cleaning text, eleminating stop words and segmentation of Chinese text. Their results showed that the predicted model can help to enhance the customer satisfaction by predicting accurate desirable products.

[34] have designed a customer behavior analysis system that classifies the high-value customers, investigates their purchasing behavior on an online platform and predicts their next purchasing. They have collected the data about opinions of customers' online shopping process for a travel agency (OTA). This system can datamine the competitors' prices, make segments for customers and analyse the future anticipations. Customer purchasing intent under the influence of competitors' price changes has also been analysed. They have first cleaned the dataset then run an algorithm to make segments of each customer profile. Then they have implemented the priori association rules for predicting the next destination and customized packages for each high-end value customer. Their results have shown that the LAX and NRT rules association provided 60% confidence in predictions.

[35] have suggested a novel approach by analysing the customer loyalty from predicting their purchase intention. They have used the RFM model to estimate customer loyalty. Then they classify the customers on their loyalty levels. Then they segment them on the basis of market variables. The groups of product are predicted for each market section by using cluster algorithms. Then they have associated the product bundle with each segment via Apriori algorithm. After that Classification models are implemented to associate product bundles to each customer. To test their approach, they have collected the data from online electronic transactions of the company. This dataset has 541910 records that are collected from 4340 customers during the time period of December 2014 to December 2015 from the Golestan province in Iran. To prepare the dataset, they have implemented data cleaning eliminating the missing and noisy values and normalization techniques. Their results showed higher SVM model results for this system than other algorithms.

[36] have designed a system to monitor human-website interaction for predicting the customer purchasing intention. They have collected data via tool from five major Polish online stores: (1) Merlin, (2) Agito, (3) Electro, (4) Empik and (5) Morele. Then they applied basic data cleaning techniques to preprocess the data. After that they have estimated behavior indicators and used them in Random forest algorithm and decision-tree classification to predict the customers purchase intention. Their experimental results have shown 52% accuracy of their model.

[37] have used data mining techniques to predict customers who came back for repurchase the product. They have broadened model, TAM, by including e-service quality, faith and leisuring of the customer to predict their intention. They have collected dataset from 360 PCHome. To preprocess the dataset, they clean it by removing the data of users who have incomplete information. Then they have implemented model of structural equation (SEM) and least squares of partial least (PLS) models. Their experimental results have predicted the 70% purchasing intention of a user.

[38] have analysed the purchasing decision and customer satisfaction through the effect of service quality and brand image. They have collected the data of Shopee customers by a questionnaire that is designed on the basis of a Likert scale of 1-5. First, they designed some hypotheses for their model. Then they have measured the Convergent validity and Discriminant validity of the model. After that they have estimated out Composite Reliability for their model by testing the trained data set after cleaning the data. They have tested their model by R-Square Model that has shown the 58% brand image influence on the purchasing behavior of a customer.

[39] have designed an explanatory model to predict the customer purchasing behavior by informational incentives and social influence affecting OSC consumer behavior. They have

collected the data from surveys of the customers that participated in the festival on Singles' Day in China. They have also conducted interviews for validating the dataset. To preprocess the dataset, the questionnaire with insufficient information filled has been excluded. Then SEM and PLS techniques has been implemented to analyse the data. They also used Harman's post hoc single-factor test to remove biasness from the dataset.

[40] have investigated the issues to anticipate the purchase behavior of online marketplace customers. They have collected the data from Alibaba. This dataset has consisted of 6 billion logs that are generated by 5 million Taobao users for almost 150 million products in one month. First, they have filtered out the original log and purchased data from a given dataset to preprocess the data. Then they have performed the feature extraction method to extract those factors by which customers purchasing behavior can be predicted. Then they have designed the sliding window which contains the log time on a product from the users. After that they have applied Gradient Boosting Decision Trees model combined with Logistic Regression model for predicting the customers intention. Their experimental results have shown the 8.66% F1 score.

[41] 2015 have developed recommendation algorithms to predict the products for a customer on the basis of their purchasing behavior. They have collected the dataset from Tianchi competition, held by Alibaba. First, they clean the dataset in pre-processing. Then they have designed their model on three modules: input module, recommendation algorithm module and output module. Implicit and explicit data mining rules have been implemented in the input module. Then they have used a collaborative filtering algorithm with sparse data in the recommendation algorithm module. Then they have implemented Top - N algorithm in the output module to gain the predicting items. Experimental results have shown that their model provided the precision value 0.031.

[42] have used data mining rules for predicting the customers preferences on the basis of users shopping behavior. They have used data from Alibaba's e-commerce platform. They have performed feature extraction to select the exact features for predicting the correct products from the user behavior. Then they have performed mining techniques on these features that are: user and product features, and users-products category features. Then they have implemented the deep forest method on their model and test it with other algorithms. Their experiment has shown the 9.51 F1 value for deep forest method.

[43] have used customer behavior log to predict the next purchasing behavior of the user. They have used feature engineering for user behavior log and then implemented supervised learning models. They have used dataset of competion that is held by AliBaba for mobile suggesting system in 2015. They have used Gradient Boosting Decision Trees and Random Forests for evaluating their model. Evaluated result has provided the 8.64% F1 score for their model.

[44] have combined the traditional model with behavior pattern extraction method to predict the suggested items for the customers on the basis of their behavior. They have collected the data desensitized online transactions that are taped by T-mall which is a branch of Alibaba. First, they have prepared the dataset by classifying the data into required groups with the time stamp. Then they have designed the model in three phases: behavior mining phase, filtering phase and suggestion phase. They have implemented the collaborative filtering technique and sequential mining in their model. Their results suggested that the classical collaborative filtering algorithm combined with traditional models have promising future in predicting the recommended products.

[45] have analysed the data mining approaches to predict the customer behavior for purchasing an item. They have designed a model CustOmer purchase pREdiction modeL (COREL) that has two phases: (1) building a customer product collection and (2) learning about customers preferences to make predictions. They have collected the data from a chinese e-commerce platform that is Jingdong.com. They have implemented logistic regression classification in their model that outperformed the Decision tree and k-nearest neighbor. They have also applied Collaborative filtering in their evaluation. Their experimental results have shown precision scores of 92.4% for LR, 90.6% for DT and 84.5 % for KNN.

[46] have proposed a network approach that is driven on data and analysis to predict one's selection sets for customers. They have used the association network to identify the communities related to products. For the customer heterogeneity, they have classified the customers into clusters based on their parameters and for each segment, they have calculated the frequency of the consideration. For predicting new suggesting sets, they have integrated the products links with customer sections. They have tested their model on two datasets: first they have generated the dataset of 10,000 customers using 100 different products and second the data of 2007 Vehicle Quality Survey from JDPA. First, they have implemented the segmentation method and k-means clustering then they have implemented association rules for prediction. Their evaluation has provided the average hit rate of 13.05% which is increased upto 26% after that.

[47] have investigated the influence of e-services influence using four phases: (1) description, (2) acceptance, (3) fulfillment and (4) remanufactured products demand. They have used data from the eBay platform collected in auction and fixed price. To pre-process the data, they removed outliers. Then controlled and relying variable are computed in each dataset. To evaluate the prediction results, they have applied OLS regression. Their experimental results have suggested e-service offerings play key factor in predicting the purchase intention of a customer. Table 2 has provided the detailed evaluation of previous work that have used machine learning methods.

TABLE 2
RESEARCH PAPERS USING MACHINE LEARNING METHODS

Title	Authors	Preprocessing	Methods	Dataset	Results	Future Work\ Limitations
Demand Prediction Using Machine Learning Methods and Stacked Generalization [15]	Resul Tugay, Sule Gunduz Oguducu	-Regression algorithms - data cleaning	-Stacked Generalization (SG) method - Random Forest - Decision tree method - Gradient Boosted Trees (GBT) - Linear Regression (LR)	- online e-commerce company, -n11 (www.n11.com)	- Root Mean Squared Error (RMSE)	The result of this model can be used for price optimization problems in future.
Market basket analysis with association rules [16]	Yuksel Akay, Unvan	- Clustering algorithm - Regression Algorithm	Apriori algorithm and FPGrwoth Algorithm	Sales data of supermarket of Vancouver Island University website	21.06 Conviction and 1 (100%) confidence value	It can be productive and profitable in case market basket analysis correctly translates the market operator.
Prediction of Repeat Customers on E-Commerce Platform Based on Blockchain [17]	Huibing Zhang and Junchao Dong	- subtime under sampling method - feature extraction	-DeepCatboost - DeepGBM - double attention BiGRU (DABiGRU) individual models using the vote-stacking method	Tmall dataset	accuracy= 91.28% and AUC=70.53%	Better ensembled learning fusion strategies can be used to improve the prediction performance of repeating purchase behavior of users.
Predicting customer demand for remanufactured products: A data-mining approach [18]	Truong Van Nguyen, Li Zhou, Alain Yee Loong Chong, Boying Li and Xiaodie Pu	- box-and-whisker plot - Tokenisation - stopwords filtering - part-of-speech (POS) tagging	-Machine-learning -Data Mining (CRISP-DM) framework - CART model, M5 model tree and RF model of Regression tree model descriptions	realworld Amazon dataset	Predictive power of RP demand provided in three forms that are strong, moderate and limited	more datasets can be aggregated for more beneficial results. ML Models have limited functionality in industry. So new methodologies can be developed that work more efficiently. data analytics in reverse logistics and CLSCs has limited research. So, this area needs more attention.
Best Seller Rank (BSR) to Sales: An empirical look at Amazon.com [19]	Hongrui Liu, Hongwei Liu, Amit Sharma	- visualization of the correlations	-Classification - Linear Regression	-Amazon hourly sales data	69.46% value of R-squared	This model can be improved by using machine learning methods like Long short-term memory (LSTM)

The Role of Machine Learning Algorithms in Determining Product Sales in E-commerce: A Case Study for Turkey [20]	Hilal Yıldız	- data cleaning - bootstrap	random forest gradient boosting	Sales record of E-commerce platform in Turkey Hepsiburada	Price plays important factor	different variable combinations can be tested for this model
Using Word2Vec Recommendation for Improved Purchase Prediction [21]	Ramazan Esmeli, Mohamed Bader-El-Den, Hassana Abdullahi	- removal of less than three interacted items and items with short durations	- N-skip-gram method of the Word2Vec model - SMOTE and Under-sampling class imbalance methods - Random Forest (RF), Bagging and Decision Trees (DT)	RetailRocket dataset	RF and DT with f and l value of 81.8% and 87.6 %	- predicted results are needed to integrated with RS
Identifying Customer Churn in After-market Operations using Machine Learning Algorithms [22]	Richard Farrow, William Trevino, Vitaly Briker, and Brent Allen	- clean the data and no missing values.	- Association Rules, K-Means, Random Forest classifiers, and Survival Analysis - Decision tree	- Flowserve's customer purchase history	Predicted accuracy is 84%.	- need to implement on larger scale - more meaning sub-sampling of customers is need
Lifestyles in Amazon: Evidence from online reviews enhanced recommender system [23]	Ling Chen, Zhang Tao, Hui Liu, Weiqing Li, Zichao Wang, Xiangen Hu, Weijun Wang	-tokenization, -normalization, -removal ofstop words, -stemming on the text	- RMSE - Tukey's HSD - analysis of variance (ANOVA)	Amazon Reviews collected by He and McAuley	Their experimental results suggested that lifestyles on online platform and all its sub-dimensions can increase the anticipating products power for customers purchase.	- it is needed to focus on the lifestyle of such consumer and products which have fewer reviews
"What you say, I buy!": Information Diagnosticity and the Impact of Electronic Word-of-Mouth (eWOM) Consumer Reviews on Purchase Intention [24]	Dr. Laura Gurney, Dr. John JD Eveland, Dr. Indira R. Guzman	- ignoring missing data surveys - converting surveys in numerical form on the basis of Lickert Scale	- Information Diagnosticity method - RMSE and bootstrapping	one-time cross-sectional survey among the English Speaking Amazon.com	online reviews about products and their combination play an important role towards purchasing behavior.	- need to test this model on product reviews

Subjective Search Intent Predictions using Customer Reviews [25]	Adrian Boteanu, Emily Dutile, Adam Kiezun, Shay Artzi	- punctuation removal, - converting text to lowercase, - tokenization - stemming unigrams - Matching and annotating respectiveentities; - combine the ASIN's reviews	- convolutional neural networks with mapping - classification - PySpark word2vec	- Wikimedia API - Amazon Product Reviews	Gain 95% accuracy for Activity datasets while 90% accuracy for Audience dataset	- large scale crowdsourced survey is needed - tri-grams as input can be helpful in increasing the accuracy
Investigating conflicting online review information: evidence from Amazon. Com [26]	Elika Kordrostami, Vahid Rahmani	- cleaning - classifying	- OLS regression	- reviews from amazon	0.35 value for root square and high valence	- only two categories have been used - more dependent variables are needed to consider
Customer demand prediction of service-oriented manufacturing using the least square support vector machine optimized by particle swarm optimization algorithm [27]	Jin Cao, Zhibin Jiang & Kangzhou Wang	-normalization	- PSO algorithm - PSR technique - swarm optimization algorithm - hybrid kernel	Online shopping data	RMSE is of 0.1076 value and MAPE results if of 0.572%	- it is needed to investigate the SOM with correlation
Machine learning-based e-commerce platform repurchase customer prediction mode [28]	Cheng-Ju Liu, Tien-Shou Huang, Ping-Tsan HoID, Jui-Chan Huang, ChingTang Hsieh	- data cleaning	- linear model logistic regression and decision tree based XGBoost model - fusion algorithm	Online customer reviews	Made improvement than single model	Machine learning models can not be easily over-fitted. So, there are robusts.
Demand Prediction and Price Optimization for Semi-Luxury Supermarket Segment [29]	T. Qu, J.H. Zhang, Felix T.S. Chan, R.S. Srivastava, M.K. Tiwari, Woo-Yong Park	- cleaning and normalization	- regression tree or random forest-based machine learning algorithm - root node analysis - heuristic methods	Online sales record	23.6% of MAPE	- more variables are needed to consider in this model

E-commerce Purchase Prediction Approach by User Behavior Data [30]	Ru Jia, Ru Li, Meiju Yu, Shanshan Wang	- tagging	- Bayes Model using classifiers of Bernoulli naive Bayes (BNB), multinomial naive Bayes (MNB) and Poisson naive Bayes (PNB)	Click data of Alibaba	13.4% to 85.9% precision	- need to test on larger scale - need to consider more dependent variable for stable results
Predicting Sales from the Language of Product Descriptions [31]	Reid Pryzant, Young-joo Chung, Dan Jurafsky	- tokenization	- Odds Ratio (OR) - Mutual information - Lasso Regularization - linear regression model -feature extraction	90,000+ product descriptions from Japanese e-commerce marketplace Rakuten	correct use of keywords, appeals to authority, polite language, and mentions of informative and seasonal language attracts the most customers	- need to focus on feature selector
Decision Support System for Online Product Recommendation Service based on Consumer Behavior [32]	Sajal Acharjee, Sheikh Abujar, Shusa Acharjee, Shahidul Islam	- data cleaning	- Decision Support System	user dataset from ICS – Machine learning dataset	54.6% accuracy	- better text filtering algorithm is needed
A Linguistic Intuitionistic Cloud Decision Support Model with Sentiment Analysis for Product Selection in E-commerce [33]	Ruxia Liang, Jianqiang Wang	- Text cleaning, Removal of stop words, Chinese text segmentation	- decision support model - information acquisition, information transformation, and integration model - sentiment analysis	online consumer reviews for five skin care products from Taobao.com	predicted model can help to enhance the customer satisfaction by predicting accurate desirable products	- weight of reviews is needed to consider - diverse data is needed to test this model
Customer online shopping experience data analytics: Integrated customer segmentation and customised services prediction model [34]	Eugene Wong, Yan Wei	- text cleaning - segmentation	- RFM - priori association rules	Review data from customers of an online travel agency (OTA)	60% prediction	- more customer demographic information and behavior information is needed
A novel model for product bundling and direct marketing in e-commerce based on market segmentation [35]	Arash Beheshtian-Ardakani, Mohammad Fathian and Mohammadreza Gholamian	- data cleaning - data normalization	- RFM mode - Apriori algorithm - segmentation - clustering - SVM model	Record of online electronic transactions of the company	Their model shows higher SVM model results for this system than other algorithms	- response model can be built - more efficient market segmentation and customer loyalty analysis are required

Modeling online user product interest for recommender systems and ergonomics studies [36]	Piotr Sulikowski, Tomasz Zdziebko, Dominik Turzyński	- data cleaning	- Random forest algorithm - decision-tree classification	five major Polish online stores: Merlin.pl, Agito.pl, Electro.pl, Empik.com, and Morele.net	52% accuracy	- more sensitive system can be develop by deeper analysis
Determinants of customer repurchase intention in online shopping [37]	Chao-Min Chiu, Chen-Chi Chang, Hsiang-Lan Cheng, Yu-Hui Fang	- removing the incomplete information	- Structural equation modelling - partial least squares	360 PCHome online shopping customer	70% purchasing intention	- model needs to be tested on multiple categorical dataset
Analysis of Service Quality and Brand Image on Customer Satisfaction Through Purchase Decisions as Intervening Variable [38]	Deviana Vierdwiyani, Afriapollo Syafarudin	- text cleaning	- Partial Least Square - R-Square Model	Shopee customers	- 58% influence of brand image on customers purchasing behavior	- more datasets with different categories are required
The impact of informational incentives and social influence on consumer behavior during Alibaba's online shopping carnival [39]	Xiaoyu Xu, Qi Li, Lifang Peng, Tzyh-Lih Hsia, Chih-Jung Huang, Jen-Her Wu	- removal of irrelevant data	- Structural equation modelling (SEM) - Partial least squares (PLS) -Harman’s post hoc single-factor test	surveys of the customers that participated in the festival on Singles’ Day in China	Results support their model to predict OSC consumer behavior	- dataset is not much reliable - influence of demographic information is needed to use to predict the behavior
Multi-classes Feature Engineering with Sliding Window for Purchase Prediction in Mobile Commerce [40]	Qiang Li, Maojie Gu, Keren Zhou and Xiaoming Sun	- filtering - feature extraction	- Gradient Boosting Decision Trees - Logistic Regression	Customers of Alibaba	8.66% F1 score	- geological information about the user is not included
Intelligent Mining on Purchase Information and Recommendation System for E-Commerce [41]	Weikang Xue, Bopin Xiao, Lin Mu	- cleaning text	- Implicit and explicit data mining - collaborative filtering algorithm	Purchasing data of AliBaba	Precision value is 0.031	- limited transaction data is used - demography is not sufficient

Research on a Prediction Model of Online Shopping Behavior Based on Deep Forest Algorithm [42]	Xin Hu, Yanfei Yang, Lanhua Chen, Siru Zhu	- feature extraction	- deep forest - deep neural network - Xgboost - random forest - support vector machine	data of Alibaba's e-commerce platform	9.51 F1 value for deep forest	- there are some uncertainties that are needed to resolve
A Method of Purchase Prediction Based on User Behavior Log [43]	Dancheng Li, Guangming Zhao, Zhi Wang, Wenjia Ma and Ying Liu	- text cleaning - feature selection	- feature engineering - supervised learning models - Gradient Boosting Decision Trees - Random Forests	AliBaba Mobile Recommendation Competition held in 2015	8.64% F1 score	- user segmentation is still not well efficient
A Behavior Mining Based Hybrid Recommender System [44]	Zhiyuan Fang, Lingqi Zhang, Kun Chen	-classification	- Collaborative filtering - Sequential pattern mining,	desensitized mobile transaction record provided by T-mall, Alibaba	classical collaborative filtering algorithms combined with traditional models have promising future in predicting the recommended products.	- more filtering variables are needed to consider
Predicting customer purchase behavior in the e-commerce context [45]	Jiangtao Qiu, Zhangxi Lin, Yinghong Li	- grouping	- logistic regression classification - decision tree - k-nearest neighbor -Collaborative filtering	Collected records from Jingdong.com	precision scores are 92.4%, 90.6% and 84.5 % for LR, DT and KNN	- brand preferences have not been increased
A Data-Driven Network Analysis Approach to Predicting Customer Choice Sets for Choice Modeling in Engineering Design [46]	Mingxian Wang, Wei Chen	- data cleaning	- Segmentation - K-means clustering -association rules	generated the dataset of 10,000 customers using 100 different products and other data of 2007 Vehicle Quality Survey from JDPA	the average hit rate for the predicted choice sets is 13.05% that increased upto 26% after that	- multiple forms and structure of choice model is needed to evaluate
The influence of e-services on customer online purchasing behavior toward remanufactured products [47]	Xun Xu, Shuo Zeng, Yuanjie He	- removing outliers	- OLS regression	Dataset from eBay	e-service offerings play an important role in predicting the purchase intention of a customer	- transaction phase is required to study - demographic analysis is needed to be done

3.3 Hybrid Method

[48] have used fusion of informationand ensemble algorithm for predicting the customers' behavior. They have used a forecasting dataset of the HI GUIDES, that have been collected for DataCastle competition. They have first preprocessed the data for cleaning the missing values, double values and unnecessary data from the original dataset. Then they have extracted the characteristics of users that are required for predicting the customer behavior. They have provided the stack model called SE by using fusion of information and ensemble algorithm. They implemented a stacking algorithm on extracted featured data. Then they have tested it with various algorithms. The stacking ensemble model has provided results which are 0.26% more accurate results than the results provided by bagging method with random forest model, 0.09% more accurate than the results of Catboost model and 1.77% more accurate than the results of logistic regression algorithm. It has shown a 98.40% F1 score.

[49] has reported a study about investigating the need of the products by customers by analysing the purchasing history of that customer and predicting discount offers on that products. They collected the dataset of 2017 from Instacart online grocery shopping site. They had designed Next Purchase Date predictor by using these four techniques: RNN and linear regression for extracting pattern for one pair of user and product and NN and extreme gradient boosting extract pattern for all user-products pair. To extract these patterns, a dataset was prepared by a featured extraction technique for creating required features. Absolute error of these models was calculated to test these. NN algorithm predicted NPD 31.8% of dataset while XGBoost predicted NPD 19.3% with error less than 24 hours. It is suggested that NN algorithm performed well than other models.

[50] predicted the next month's total sales on the basis of past sales data obtained from the total sales for every product implementing data mining. They had used sales data of Russian software firms 1C Company to train the dataset. They had implemented conventional data engineering procedures to solve their problem. First, they used a data cleaning method to correct the invalid and missing values by either removing those values or taking mean or median. Exploratory Data Analysis (EDA) was used for plotting graphs in a distributed manner that shows the sales of each month. To implement the feature engineering process, all datasets were merged and transformed into master dataset. Then required features are extracted. They have compared three methods for predicting sales of the coming month that are XGBoost, LSTM and ARIMA model. RMSE test was done for each model to test these models. XGBoost provided 0.87815 with RMSE, LSTM provided 0.92417 with RMSE, ARIMA provided 1.09266 with RMSE. To reduce bias and variance in output predictions of this model, more ensemble techniques can be used in future predictions.

[51] predicted the online behavior of customers using clickstream data with machine learning techniques. They had gathered sales records of an online store from May 20th to July 20th, 2018. They had prepared the dataset into 5 sessions by deleting page views generated by bots, all checkout page views and the last three page views of all sessions, pruning the large data and masking the data by zero-padding. Then features were extracted using GRU and LSTM classifiers. To predict the behavior of customers, they have implemented logistic regression, random forest, gradient boosting and multi-layer perceptron on the trained dataset. The ensemble model provided the AUC value of 0.9581 that showed RNN-based sequence classifiers predicted customer behavior more accurately than previous approaches.

[52] had provided the analysis for methodologies to predict buying sessions, purchase decisions, and customer intents. They had collected online time, logs, and clicking records for testing their model. They prepared the data by transforming it into transaction records. They implemented feature engineering techniques to obtain the desired features from the dataset. Recurrent neural networks, LSTMs, and Autoencoders techniques were used for the feature learning process. After obtaining the trained dataset, gradient boosting decision, logistic regression, multilayer perceptron, random forest and support vector machine were tested to predict the customer behavior. It was observed that each one method has its own benefits and disadvantages.

[53] has investigated the potential customer and predicted their purchasing behavior for a retail superstore using machine learning methods. They had used the sales record of an e-retail superstore. They implemented Supervised Machine Learning technique for modelling their study. First they implemented basic preprocessing techniques and labelled the data. The potential score of 0 and 1 obtained and dataset was processed through a classification method to extract the required information and patterns. To increase the accuracy, feature extraction was also implemented. Different tests had been implemented that show Multilayer Perceptron Classifier provided 99.41% accuracy for this methodology.

[54] has predicted the customer's purchasing decision using different machine learning methods including DT, MLP network, naïve bayes, RBF network, and SVM. They have used the data of 240 consumers in Vietnam 2018 which was collected by survey based on Likert scale. To preprocess, they have classified the data using three independent variables: Perceived price, Perceived quality and Consumers ethnocentrism. The results have shown the highest performance having 91.6667% correct anticiaptions.

[55] have investigated the predicting model for both anonymous and identified sessions of customers on an e-commerce platform. They also analysed the sessions when a customer purchase a product with the sessions when a customer do not purchase the product. They have considered a Eurpeon e-commerce site to gather the anonymized visit data of October 2019. To preprocess the data, they have collected sample data with a unique non-personal customer identifier, linking time with query and clicked page URL. They also recorded which

price customer decides for purchasing an item while surfing among different prices of the same product.

Then, they have selected the items of interest and removed all other products. They have used feature selection methods and Random Forest to evaluate the purchase intent of a user. They also used LR, KNN, SVM, neural classifier, and GBDT for evaluating the modified dataset. They have used points for each session for predicting the intent or purchase of a user. Their result has showed two predictors: one predictor provided 17.54% *F*1 score for purchase intent of anonymous session and second predictor provided *F*1 score of 96.20% for purchase intent of identified session.

[56] have designed a decision support tool to predict the sales data via analysing the customers purchasing combination. They have collected the customer preferences by purchasing record of electric toy cars from Tao Bao, an online platform belonging to Alibaba Group ™, to train a dataset of having three domains: customer with customer satisfaction index, functional with product specification and their combinations and physical with products components. To train the dataset, irrelevant and redundant specifications have been removed manually. Then Principal Components Analysis has been carried out to recognize the specifications with least variance. Then such entries with missing values have also reduced.

The dataset has converted into normalized form. After that Customers' satisfaction index (CSI) has been calculated. They have analysed the specific combinations by cluster analysis. They have used the K-means algorithm for this purpose. Then they have calculated Customer Satisfaction Index with every combination of products through regression techniques using Neural Network and Bayesian Regularization training algorithm. Their tool has shown good results about predicting the combination of products and customer behavior.

[57] have worked on sentimental analysis and data mining on social netwrok to design prediction models that will be valuable for retailers of the e-commerce platform and customers who want to search durable products. They have gathered the data of online customer's real-time search and review from Amazon.com only about the cameras. In this first dataset, they have removed irrelevant information about cameras to obtain an efficient dataset. In other dataset, they have collected the IDs who follow the brands, customer viewpoint for a specific product considering that brand's Twitter account and Amazon's reviews. Purchase behavior after extensive search of customers is also gathered from amazon for analysing in the model. A sentiment analysis is also performed for obtaining product unlikeness and attributes level unlikeness. They have applied the Vader rule-based model to calculate the percentage polarity in the range of positive, negative and neutral. they have linked the social network mining with sentimental analysis to obtain the predicting values for an item. Then they normalized the predicted data. On the predicted set, they performed Multiple Linear Regressions in case of linearity and GMDH Neural Network in case of nonlinearity so they can compare the accuracy of their model. Their experimental results have provided the high accuracy.

[58] have suggested a novel approach that includes two modules: (1) anticipating the visitor's intent and (2) identifying the abandonment of similar products. First, they have implemented pageview data to analyse the intention of customer. After feature extraction, they have performed RF, SVMs and MLP classifiers. To pre-process the data, they have used oversampling and feature selection. Their experiments have shown that their proposed system provided highly accurate results and F1 Score with MPL than with RF and SVM. Table 3 has explained the detailed investigation of researches using hybrid methods.

TABLE 3
RESEARCH PAPERS USING HYBRID METHODS

Title	Authors	Pre-processing	Methods	Dataset	Results	Limitations / Future Direction
SE-stacking: Improving user purchase behavior prediction by information fusion and ensemble learning [48]	Jing Xu, Jie Wang, Ye Tian, Jiangpeng Yan, Xiu LiID, Xin Gao	-clean the missing values, double values, and irrelevant values - Feature extraction	-ANFIS Model by RMSE -mean absolute deviation and mean absolute percentage error -Least square method and back-propagation gradient descent method	Dataset of previous e-order arrivals	78% accurate for ARIMA	-predictive methodologies or heuristics approaches can perform better

Using Machine Learning To Predict The Next Purchase Date For An Individual Retail Customer [49]	M. Droomer & J. Bekker	- Feature Extraction	- recurrent neural networks - linear regression - artificial neural networks - extreme gradient boosting	Instacart online grocery shopping	NPD accuracy = 31.8% with an error of less than one day and 16.8% with an error of between one and two days.	- enhancing the analysis to more datasets, - NPD predictor can work better at a retail chain using the NN algorithm, -finding ways to combine the sequence and non-sequence-based attempt
Predicting Future Sales of Retail Products using Machine Learning [50]	Alay Dilipbhai Shah, Devendra Swami, Subhrajeet K B Ray	- Data cleaning, - Exploratory Data Analysis (EDA) - Feature engineering	- eXtreme gradient boosting - Long short term memory -ARIMA model - Root mean squared error	Sales data of Russian software firms 1C Company	XGBoost test with RMSE = 0.87815, XGBoost test with RMSE = 0.92417, ARIMA test with RMSE = 1.09266	- ensemble techniques can be combined to combine the predictions for better results
Predicting online shopping behaviour from clickstream data using deep learning [51]	Dennis Koehn, Stefan Lessmann, Markus Schaal	- Deleted page views generated by bots - deleted all checkout page views - deleted the last three page views of all sessions. - pruning the large data - masking the data by zero-padding	- feature engineering - Logistic regression, - Random forest, - Gradient boosting - Multi-layer perceptron	Online sales records of a webstore.	AUC value is 0.9581	- tested only on limited level
Customer Purchase Behavior Prediction in E-commerce: Current Tasks, Applications and Methodologies [52]	Douglas Cirqueira, Markus Hofer, Dietmar Nedbal, Markus Helfert, Marija Bezbradica	-transformation - feature extraction	- Logistic regression - Random forest -Support vector machines - Gradient boosting decision - Multilayer perceptron	online time-period, shopping timelogs, and clicking records	Different algorithms provided different results	- multitask setting and time of purchase is needed to consider
A Machine Learning Approach to Identify Potential Customer Based on Purchase Behavior [53]	Adil Mahmud Choudhury and Kamruddin Nur	- data cleaning - labelling	- Classification - Feature engineering - Supervised machine learning technique	Sales record of an e-retail superstore Taradin	Multilayer Perceptron Classifier provided 99.41% accuracy	- buying frequency is needed to combine with product of interest

An approach based on machine learning techniques for forecasting Vietnamese consumers' purchase behaviour [54]	Quang Hung Do and Tran Van Trang	- data cleaning - data classification	- decision tree - multilayer perceptron - Naïve Bayes - radial basis function - support vector machine (SVM	data of 240 consumers in Vietnam 2018 which was collected by survey based on Likert scale	91.6667% correct prediction	- limited explanatory variables are considered
Analyzing and Predicting Purchase Intent in E-commerce: Anonymous vs. Identified Customers [55]	Mariya Hendriksen, Ernst Kuiper, Pim Nauts, Sebastian Schelter, Maarten de Rijke	- Collect data with selected indicators - removing irrelevant products	- Random Forest - Logistic regression - K-nearest neighbors - Support vector machines - Neural classifier - Gradient boosted - Decision tree	data of four weeks (28 days) of anonymized visits sampled from a European e-commerce platform in October 2019	*F*1 score is of 96.20% for purchase intent	- anonymous users are not considered in this study
Machine learning-based design features decision support tool via customers purchasing data analysis [56]	Jian Zhang, Xingpeng Chu, Alessandro Simeone and Peihua Gu	- cleaning the text	- Cluster analysis - regression techniques using Neural Network and Bayesian Regularization training algorithm	online-purchasing data of electric toy cars from Tao Bao, an online platform belonging to Alibaba Group ™	Their tool has shown good results about predicting the combination of products and customer behavior	- tool is needed to test with bigger datasets - accuracy of the algorithm is needed to enhance
Predicting the consumer's purchase intention of durable goods: An attribute-level analysis [57]	Sujoy Bag, Manoj Kumar Tiwari, Felix T.S. Chan	- cleaning the data set	- sentiment analysis - Vader rule-based model - Multiple Linear Regressions	- customer's real-time search and review from Amazon.com	Their experimental results showed the high accuracy of the predicted model	- needs to test on multiple products - utilitarianism is needed to study - more influential factors are needed to consider
Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks [58]	C. Okan Sakar, S. Olcay Polat, Mete Katircioglu, Yomi Kastro	- feature selection - oversampling	- feature extraction - random forest (RF), support vector machines (SVMs), and multilayer perceptron (MLP)	- data of online shoppers	High accuracy of 88.92%	- user based recommender system can be developed for better results

4 Discussion

Most of the researchers have adopted different procedures to predict the customer behavior but tested their algorithms on small scale. They have used limited number of categories of products for classification. In previous work, they have used limited dependent variables. So, more variables are needed to be investigated for classifying the features. They have not linked the purchasing time with the product categories as well. It is needed to consider. In this way, we can predict the exact sale of the products with particular time period. There is need of using more hybrid methods for achieving better results

In future, we can use feature selection method to extract more accurate features from the datasets. In previous work, some researchers have used questionnaires to know the point of view of customers but that method is not much reliable. So, we need to test such frameworks with other type of dataset that involves sale and combination of products.

5 Significance

As we have studied about the previous work of hardworking researches that has provided various algorithms for predicting good percentage of customer purchase intention, our study will help future researchers to study about the literature in one place. As we have provided the limitations of each study, it will help to focus on the limiting factors due to which results can not be achieved pretty good on complex level. This study will also help to design such system that helps e-retailers to predict more about customer's choice of products from multiple categories. By predicting customer purchase intention or choice, better suggesting system can be developed. It can also help e-retailers for better stock management by judging the need of customer. By good suggesting system, sales of an e-commerce platform can also be increased.

6 Conclusion

As the customer behavior prediction has become the important topic in e-commerce, many researchers have provided different algorithms for enhancing the accuracy of the prediction. It can help to make better suggestion for customers on online selling platforms like Amazon and Alibaba. To contribute in this respect, we have reviewed previous literature for analyzing the strengths and weaknesses of their work. We have structured the previous literature in the categories of machine learning methods, deep learning methods and hybrid methods. Most of them have predicted the customer behavior and intent of purchase highly accurate but hey have been tested on small scale with a single type of product or limited number of combinations. Some previous works has been concluded on the basis of questionnaires that are not much productive when system has been applied in real world. So, such a framework is needed that should be tested with multiple categories of products, their purchasing time and more dependent corpuses. In future, more feature selection methods like wrap-based, filter based, Pearson Correlation can be implemented for extracting the exact results from complicated datasets.

References

[1] Kai-Ingo Voigt, Oana Buliga, Kathrin Michl, "Creating the Global Shopping Mall: The Case of Amazon", Springer International Publishing Switzerland, pp. 67-77, 2016.

[2] Brent Smith, Greg Linden," Two Decades of Recommender Systems at Amazon.com,.2017.

[3] Young-Chan Kim, "Alibaba: Jack Ma's Unique Growth Strategy and the Future of Its Global Development in the Chinese Digital Business Industry", 2018.

[4] Usha Ramanathan, Nachiappan Subramanian, Wantao Yu & Rohini Vijaygopal, "Impact of customer loyalty and service operations on customer behaviour and firm performance: empirical evidence from UK retail sector", 2017.

[5] Chris Rygielski, Jyun-Cheng Wang, David C. Yen, "Data mining techniques for customer relationship management", 2002.

[6] Hassani, Hossein, "Banking with blockchain-ed big data",2018.

[7] Sergey Gutnik, "Application of Data Mining and Machine Learning Methods to Enhance the Effectiveness of Digital Marketing Strategies", 2021.

[8] Mingyang Zhang, Yixin Wang, and Zhiguo Wu, "Data Mining Algorithm for Demand Forecast Analysis on Flash Sales Platform", 2021

[9] W Reim, D Sjödin, V Parida, "Mitigating adverse customer behaviour for product-service system provision: An agency theory perspective", 2018.

[10] [AYL Chong, E Ch'ng, MJ Liu, B Li," Predicting consumer product demands via Big Data: the roles of online promotional marketing and online reviews", 2017.

[11] Long Guo, Lifeng Hua, Rongfei Jia, Binqiang Zhao, Xiaobo Wang, Bin Cui," Buying or Browsing? : Predicting Real-time Purchasing Intent using Attention-based Deep Network with Multiple Behavior", 2019.

[12] Jin Cao, Zhibin Jiang & Kangzhou Wang," Customer demand prediction of service-oriented manufacturing incorporating customer satisfaction", 2015.

[13] Sridhar Ramaswamy, Natalie DeClerck," Customer Perception Analysis Using Deep Learning and NLP", 2018.

[14] Joerg Evermann, Jana-Rebecca Rehse, Peter Fettke," Predicting Process Behaviour using Deep Learning", 2017.

[15] Resul Tugay, Sule Gunduz Oguducu, "Demand Prediction Using Machine Learning Methods and Stacked Generalization", 2020.

[16] Yuksel Akay, Unvan," Market basket analysis with association rules", 2020.

[17] Huibing Zhang and Junchao Dong," Prediction of Repeat Customers on E-Commerce Platform Based on Blockchain", 2020.

[18] Truong Van Nguyen, Li Zhou, Alain Yee Loong Chong, Boying Li and Xiaodie Pu, "Predicting customer demand for remanufactured products: A data-mining approach", 2020.

[19] Hongrui Liu, Hongwei Liu, Amit Sharma, “Best Seller Rank (BSR) to Sales: An empirical look at Amazon.com “, 2020.

[20] Hilal Yıldız, “The Role of Machine Learning Algorithms in Determining Product Sales in E-commerce: A Case Study for Turkey”, 2020.

[21] Ramazan Esmeli, Mohamed Bader-El-Den, Hassana Abdullahi,” Using Word2Vec Recommendation for Improved Purchase Prediction”, 2020.

[22] Richard Farrow, William Trevino, Vitaly Briker, and Brent Allen, “Identifying Customer Churn in After-market Operations using Machine Learning Algorithms”, 2020.

[23] Ling Chen, Zhang Tao, Hui Liu, Weiqing Li, Zichao Wang, Xiangen Hu, Weijun Wang, “Lifestyles in Amazon: Evidence from online reviews enhanced recommender system”, 2019.

[24] Dr. Laura Gurney, Dr. John JD Eveland, Dr. Indira R. Guzman, “‘What you say, I buy!’: Information Diagnosticity and the Impact of Electronic Word-of-Mouth (eWOM) Consumer Reviews on Purchase Intention”, 2019.

[25] Adrian Boteanu, Emily Dutile, Adam Kiezun, Shay Artzi, “Subjective Search Intent Predictions using Customer Reviews”, 2020.

[26] Elika Kordrostami, Vahid Rahmani, “Investigating conflicting online review information: evidence from Amazon. com”, 2020.

[27] Jin Cao, Zhibin Jiang & Kangzhou Wang, “Customer demand prediction of service-oriented manufacturing using the least square support vector machine optimized by particle swarm optimization algorithm”, 2016.

[28] Cheng-Ju Liu, Tien-Shou Huang, Ping-Tsan HoID, Jui-Chan Huang, ChingTang Hsieh, “Machine learning-based e-commerce platform repurchase customer prediction mode”, 2020.

[29] T. Qu, J.H. Zhang, Felix T.S. Chan, R.S. Srivastava, M.K. Tiwari, Woo-Yong Park, “Demand Prediction and Price Optimization for Semi-Luxury Supermarket Segment”, 2017.

[30] Ru Jia, Ru Li, Meiju Yu, Shanshan Wang, “E-commerce Purchase Prediction Approach by User Behavior Data”, 2017.

[31] Reid Pryzant, Young-joo Chung, Dan Jurafsky, “Predicting Sales from the Language of Product Descriptions”, 2017.

[32] Sajal Acharjee, Sheikh Abujar, Shusa Acharjee, Shahidul Islam,” Decision Support System for Online Product Recommendation Service based on Consumer Behavior”, 2017.

[33] Ruxia Liang, Jian-qiang Wang, “A Linguistic Intuitionistic Cloud Decision Support Model with Sentiment Analysis for Product Selection in E-commerce”, 2019.

[34] Eugene Wong, Yan Wei, “Customer online shopping experience data analytics: Integrated customer segmentation and customised services prediction model”, 2018.

[35] Arash Beheshtian-Ardakani, Mohammad Fathian and Mohammadreza Gholamian, “A novel model for product bundling and direct marketing in e-commerce based on market segmentation “, 2017.

[36] Piotr Sulikowski, Tomasz Zdziebko, Dominik Turzyński, “Modeling online user product interest for recommender systems and ergonomics studies”, 2017.

[37] Chao-Min Chiu, Chen-Chi Chang, Hsiang-Lan Cheng, Yu-Hui Fang, “Determinants of customer repurchase intention in online shopping”, 2008.

[38] Deviana Vierdwiyani, Afriapollo Syafarudin, “Analysis of Service Quality and Brand Image on Customer Satisfaction Through Purchase Decisions as Intervening Variable “, 2020.

[39] Xiaoyu Xu, Qi Li, Lifang Peng, Tzyh-Lih Hsia, Chih-Jung Huang, Jen-Her Wu, “The impact of informational incentives and social influence on consumer behavior during Alibaba's online shopping carnival”, 2017.

[40] Qiang Li, Maojie Gu, Keren Zhou and Xiaoming Sun, “Multi-classes Feature Engineering with Sliding Window for Purchase Prediction in Mobile Commerce”, 2015.

[41] Weikang Xue, Bopin Xiao, Lin Mu, “Intelligent Mining on Purchase Information and Recommendation System for E-Commerce”, 2015.

[42] Xin Hu, Yanfei Yang, Lanhua Chen, Siru Zhu, “Research on a Prediction Model of Online Shopping Behavior Based on Deep Forest Algorithm”, 2020.

[43] Dancheng Li, Guangming Zhao, Zhi Wang, Wenjia Ma and Ying Liu, “A Method of Purchase Prediction Based on User Behavior Log”, 2015.

[44] Zhiyuan Fang, Lingqi Zhang, Kun Chen, “A Behavior Mining Based Hybrid Recommender System”, 2016.

[45] Jiangtao Qiu, Zhangxi Lin, Yinghong Li, “Predicting customer purchase behavior in the e-commerce context”, 2015.

[46] Mingxian Wang, Wei Chen, “A Data-Driven Network Analysis Approach to Predicting Customer Choice Sets for Choice Modeling in Engineering Design”, 2015.

[47] Xun Xu, Shuo Zeng, Yuanjie He, “The influence of e-services on customer online purchasing behavior toward remanufactured products”, 2017.

[48] Jing Xu, Jie Wang, Ye Tian, Jiangpeng Yan, Xiu LiID, Xin Gao, “SE-stacking: Improving user purchase behavior prediction by information fusion and ensemble learning”, 2020.

[49] M. Droomer & J. Bekker, “Using Machine Learning to Predict the Next Purchase Date for An Individual Retail Customer”, 2020.

[50] Alay Dilipbhai Shah, Devendra Swami, Subhrajeet K B Ray, “Predicting Future Sales of Retail Products using Machine Learning”, 2020.

[51] Dennis Koehn, Stefan Lessmann, Markus Schaal, “Predicting online shopping behaviour from clickstream data using deep learning”, 2020.

[52] Douglas Cirqueira, Markus Hofer, Dietmar Nedbal, Markus Helfert, Marija Bezbradica, “Customer Purchase Behavior Prediction in E-commerce: Current Tasks, Applications and Methodologies”, 2020.

[53] Adil Mahmud Choudhury and Kamruddin Nur, “A Machine Learning Approach to Identify Potential Customer Based on Purchase Behavior”, 2020.

[54] Quang Hung Do and Tran Van Trang, “An approach based on machine learning techniques for forecasting Vietnamese consumers’ purchase behaviour”, 2020.

[55] Mariya Hendriksen, Ernst Kuiper, Pim Nauts, Sebastian Schelter, Maarten de Rijke, “Analyzing and Predicting Purchase Intent in E-commerce: Anonymous vs. Identified Customers”, 2020.

[56] Jian Zhang, Xingpeng Chu, Alessandro Simeone and Peihua Gu, “Machine learning-based design features decision support tool via customers purchasing data analysis”, 2020.

[57] Sujoy Bag, Manoj Kumar Tiwari, Felix T.S. Chan, “Predicting the consumer’s purchase intention of durable goods: An attribute-level analysis”, 2019.

[58] C. Okan Sakar, S. Olcay Polat, Mete Katircioglu, Yomi Kastro, “Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks”, 2018.