International Journal of Communication Networks and Information Security
ISSN: 2073-607X, 2076-0930
Volume 15 Issue 01 Year 2023

Study on the Influence of Knowledge-driven Technology on predicting consumer Repurchase Behaviour

Yajing Chen
School of Business and Economics, Universiti Putra Malaysia, Serdang 43300, Malaysia
gs59177@student.upm.edu.my
Yee Choy Leong
School of Business and Economics, Universiti Putra Malaysia, Serdang 43300, Malaysia
yee@upm.edu.my
Lee Shin Yiing
School of Business and Economics, Universiti Putra Malaysia, Serdang 43300, Malaysia
leeshin@upm.edu.my
Yunxia Xiao
College of Economic and Management, Chongqing Industry Polytechnic College, Chongqing, China
gs59177@student.upm.edu.my

Article History

Received: 05 March 2023
Revised: 21 April 2023
Accepted: 19 May 2023

Abstract

Consumer purchase behaviour has become a potential research area in business analytics, as exploring micro-level details would increase the business's profitability. In this prospect, many MNCs and other enterprises harness contemporary computing technologies like Big Data Analytics, Deep Learning, and Predictive Analytics to explore the latent knowledge in purchase patterns and customer behaviour. This work deploys a novel Multi-class Ada Boost (MAB) supported Convolutional Neural Network (CNN) to learn customer purchase behaviour by analysing the buying patterns and trends to predict the repurchases. The proposed model learns the trends sequentially as the CNN models are cascaded one after the other, thus preserving the contextual knowledge between the models. The proposed model is tested for its efficacy on Instacart Market Basket Analysis to predict whether the customer is repurchasing the same product. The performance of the proposed model is compared with other states of art Machine Learning algorithms like Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and XGBoost in terms of prediction accuracy, precision, and F1-score. In addition, synthetic noise is induced into the dataset at various levels to analyse the model's efficacy in handling noisy data. These results indicate that the model shows better results than its peers, thus making it more suitable to predict customer repurchase behaviour and pattern.

Keywords: Repurchase Behaviour, Multiclass AdaBoost-CNN, Transfer Learning, CNN Estimator

1. Introduction

The amount of information available is expanding rapidly with the significant data era. In just five years, the capacity of digital information expanded nine times in 2011. Global information volume will exceed 35 trillion gigabytes by 2020 [1]. Most information is derived from previous Internet user records, including web searches [2], clicks, and other actions. Because of ongoing technological progress, the business world is highly dynamic nowadays. The banking industry is in challenging and contentious situations due to these shifting conditions and a substantial financial market [3]. The economic growth of every country is primarily reliant on banks. Banks are undergoing rapid transformation due to the market's continual innovation and increased usage of technological services. Because of technological improvements, people may now learn about world events with a single touch [4].

The era of Industry 4.0 has begun with the advent of digitalisation. E-commerce systems will become more popular as technology and online culture advance [5]. The number of stores open for business shows a movement in trading patterns towards the online system. Each nation that wants to transition from a traditional retail system to an online marketplace must contend with the repercussions of digitalisation. Otherwise, it will lose out on revenue potential or potentially suffer a severe economic collapse. According to predictions, Indonesia will soon rank among the prominent participants in the digital media sector.

Additionally, three encouraging trends, including the rise in smartphone usage [6], the expansion of mobile marketing [7], and the expansion of e-commerce [8,9], lead to Indonesia being a significant user of digital media. Predicting client repurchase propensity/frequency in highly competitive consumer marketplaces has drawn significant academic interest [10]. Repurchase prediction has been critical in customer base analytics and targeted marketing [11]. Identifying clients from a significant clientele likely to purchase something within the following day, month, or quarter is a common data analytics task since it helps managers deploy sales personnel and perform marketing strategies more efficiently. This subject has expanded in economic importance with the introduction of many online platforms, which must comprehend client behaviours and forecast customer actions based on data records. Repurchase predictions using computer vision have lately gained prominence on the practical front [12],[13].

Nonetheless, marketing models are continuously being developed for this prediction [14],[15]. The willingness to make another purchase at the merchant where a consumer previously purchased is referred to as repurchase intent. Repurchase intent is critical since maintaining existing customers is significantly less expensive than obtaining new ones. Therefore, their regular purchasing habits boost corporate earnings [16],[17]. Keeping customers allows organisations to save money on gaining new customers, which increases profitability [18],[19],[20]. They are also more inclined to recommend the service to relatives and friends. This study suggests a technique that allows MulticlassAdaBoost to categorise enormous amounts of data using CNN. This article suggests an additional algorithm, MulticlassAdaBoost-CNN, that combines CNN's various advantages in analysing and finding trends in extensive data with MulticlassAdaBoost's functionality in handling enormous unbalance data in order to instil CNN's possibility in MulticlassAdaBoost-CNN for the task of having dealt with small datasets.

The contribution of the work is as follows:

Due to the incompleteness of the original data source used to collect the data, there are duplicated and missing values, Inconsistent attribute data formats, and a noise effect.

The core qualities that are essential to the research topic are recognised after quick data analysis and consumer purchase habits. The features are then linked and extracted in relation.

The proposed method MulticlassAdaBoost-CNN -overcomes this challenge and lowers it. The suggested MulticlassAdaBoost-CNN approach has a lower computational cost than standard MulticlassAdaBoost and CNN methods. This is accomplished through deep learning techniques' transfer learning characteristics.

The following is how the paper is organised. The second section introduces similar works. The suggested MulticlassAdaBoost-CNN concept is then explored in Section 3. Part 4 then discusses the experimental outcomes. The experimental section outlines the tests performed to assess the performance of the suggested approach. Section 5 finally presents a conclusion.

2. Related Works

Many researchers and scholars in the marketing and business analytics fields have put a lot of effort into predicting customers' repurchase behaviour, which eventually helps business firms to enhance their cross-selling campaigns. The standard techniques for repurchasing behaviour predictions include statistical and mathematical forecasting, Artificial Intelligence (AI) based methods, Machine Learning and Deep Learning.

Mathematical or statistical forecasting

Discounts, offers, and trade strategies are two unique kinds of merchandise exchange schemes that are widely practised. It delineates four decentralised models and uses the Stackelberg gaming procedure driven by the creator to settle on the best strategy. Consequently, a play-based benefitsharing agreement is intended to examine channel coordination. The analysis and approval of the best choices are done based on scientific and mathematical examinations [21]. This model is capable of doing short-term analysis, but its efficiency still needs to be tested in long-term trends; a comprehensive review of mathematical optimisation algorithms in predicting customer repurchase behaviour on the grounds of ten metaheuristics is done [22]. The estimates in this study incorporate the following optimisers: insect lion, dragonfly, grasshopper, Harris birds, moth-fire, multi-section, sine cosine, salp swarm, whale, and dim wolf. The review highlights that.

Moth Fire is better than its peers. A cosine-based prediction of repurchase behaviour is based on the characteristics of organisers, government officials, and administrators in organisations where money scandals are common [23]. The work considers three factors: frequency of purchase, redundancy, and client recurrence. The performance of the prediction is measured in terms of accuracy, Recall, and F1 score.

Machine Learning Algorithms

Web-based organisations use clustering techniques to learn the customers buying trends based on the information collected from the user activity. The K-means clustering algorithm is leveraged to predict whether the customer will buy the product [24]. Network intelligent recommendations are gaining momentum as they can easily bestow customer relations in web-based shopping platforms [25]. A Dynamic novel network intelligent hybrid recommendation algorithm is proposed in this work that can distinguish the operational periods. This model is integrated with TOPSIS and DEMATEL for assessing the customer preference index. The popular TOPSIS is integrated with a novel Intuitionistic Fuzzy Number for further evaluation. Though this method gives good prediction accuracy, it is computationally intensive. Data collected from Iran and Hungary were explored to predict the online purchase trend using a grocery app [26] by deploying Gaussian Mixture Model and a multi-layer perceptron. The data is segregated into three groups of users of the application. These models are straightforward, but their efficiency in handling overfitting has yet to be explored. Efficient feature engineering is the key to better prediction results. The Recency, Frequency, and Monetary aspects are excellent measures for improving the customer relationship by fostering repurchase behaviour. As the data is highly imbalanced, a SMOTE-based ENN is used to predict a customer's repurchase [27]. The salient feature of this work is the automatic hyperparameter tuning to give better accuracy than other compared state of art methods. This work delineates there is room for the deployment of better models.

Ensemble models

A novel ensemble model to predict the repurchasers is constructed using Tmall's dataset [28]. A SMOTE technique is used in this work to resolve the class imbalance issue. The classifiers used in this work are factorisation machine, LR, extreme gradient boosting, and a light version of gradient boosting machine. These algorithms are stacked into two layers to improve prediction accuracy. This is an effective method to predict the trends, but the model may need better generalisation. A deep ensemble model comprising individual-based learners like DeepCatboost, DeepGBM, and

DABiGRU is used to predict customer prediction behaviour [29]. These classifiers are combined using the vote-stacking method to give better performance.

A repurchase-and-return reputation method combines context-based information [30], which considers the repurchase and the product returns. A cost-sensitive dynamic learning classifier ensemble that can handle imbalanced data for predicting customer purchase behaviour is developed [31]. This model uses two types of dynamic ensemble: classifier selection and ensemble selection. This work also includes cost-sensitive selection criteria. The performance is validated on UCI accurate churn prediction dataset. Creating ensembles of complex algorithms is computationally expensive.

The detailed literature survey indicates that customer repurchase behaviour is an extensively researched area in business analytics. Various methods, from mathematical models to data-driven methods, are discussed here. However, only a few works exist in the literature that harnesses deep learning methods and their ensemble.

3. Proposed System

Optimal Consumer Buying Behaviors are the intentions and actions of both online and offline consumers before buying or repurchasing a product. This tedious process takes the help of logs of search engines, social media activities, or other influencing activities. Businesses must comprehend this complex process, as tapping this information would foster their sales and marketing initiatives. Figure 1 shows the various phases in the analysis of customer behaviour patterns.

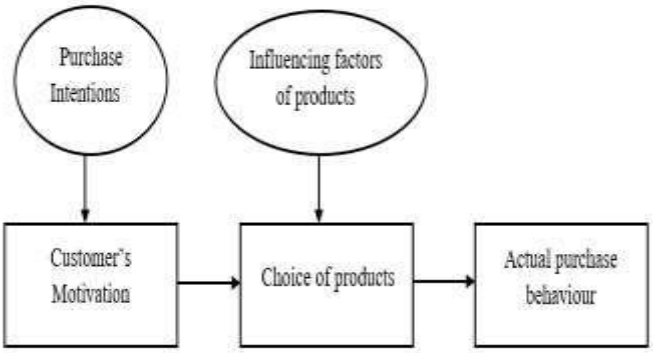

Figure 1. Phases in Analysing Customer Behaviour Prediction.

The predictions can be done by analysing multiple tangible and intangible factors like customer intentions, their influencers, available choice of products, etc. Figure 2 depicts the creation of the customer buying behaviour prediction model when the feature from the raw document is built to build the feature summary table. The characteristics of the training set are then extracted. The result of image retrieval is the selection of variables from the package that the proposed technique should just use. Finally, following feature selection, the training data is given into the learning process to produce a prediction model.

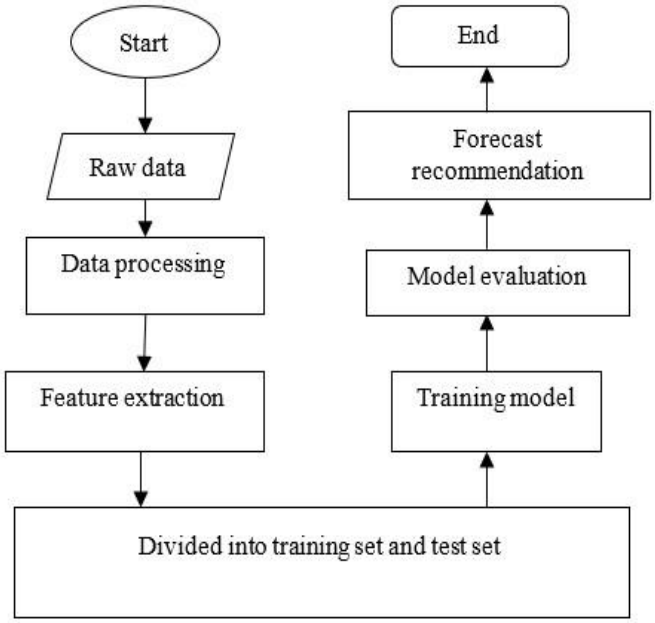

Figure 2. Consumer Buying Behaviour Prediction Process

E-Commerce Company's Return Policy: The Consumer Protection Law underwent revision by the government in 2014. This regulation formally established the "7-days without justification" return policy for internet purchases. According to this policy, Customers have the right to return things without providing a reason within seven days after delivery. On the one hand, a draconian return policy may protect customers' legal rights and interests. By giving an unfair return policy, renowned e-commerce platforms can attract users and create demand for goods. While making purchases online, consumers evaluate the level of service the seller offers and aspects relating to the product itself. The return procedure is a crucial measure.

Consumer Repurchase Behavior Analysis: Using Internet shopping platforms, consumers screen, gather, and add products to shopping carts in real time. These actions demonstrate desired good. The groups of individuals shopping are seniors, office professionals, and schoolchildren. Each has a different set of preferences. For instance, young individuals are enthusiastic about purchasing technology products, seniors enjoy purchasing everyday household goods or food items, and schoolchildren enjoy purchasing school supplies and snack foods. In light of this, certain and frequently purchase a particular category of items, such as daily fruits and vegetables, with a high repeat purchase rate. However, the repeat purchase rate is modest, and some customers only sporadically order electronics and other items. Some customers enjoy shopping, making various purchases, and placing plenty of orders. Yet, confident due to the intricacy of the product categories. This experiment aims to pick characteristic data through results analysis. The likelihood that customers will make another purchase is modelled and predicted using ensemble learning methods. Making forecasts using two categorisation systems is a workable solution. The supervised learning method used in this study employs extensive historical data to give training samples and goal values. Hence, when examining, product repurchase needs to be considered more. Knowing which products are frequently repurchased is crucial for a particular sort of consumer. The discovery of the underlying laws is quite significant.

Order Information Analysis: Key steps in prediction include locating relevant data and separating features from historical data. Although most of them are helpful for behaviour, a small number are mainly utilised for description and have little practical value. With feature extraction, appropriate fields be transformed into reliable features. Fields with low importance can simply be ignored. Reordering is a crucial feature. The accuracy of the forecast can be checked during training. When predicting a user's reconstructive behaviour, information like the time frame of the product purchase is crucial. Real-world consumer trends are typically predictable. Orders are concentrated on weekends and holidays compared to weekdays. The higher-quality items are frequently added to the

shopping basket first since they are more likely to be repeated purchases. A more thorough analysis of more relevant data is possible. To get the predicted result.

3.1 Pre-processing

The data set was made available on the Kaggle platform. Based on past performance, the competition's goal is to foretell which products customers will purchase again. Online downloads of Instacart's open-source historical sales data are available.

3.2 Feature Extraction

The original data is highly dimensional and has low-quality features, which increases the model's training duration. Appropriate data characteristics are crucial for increasing prediction accuracy and reducing modelling complexity.

3.3 Proposed MulticlassAdaBoost-CNN Algorithm

The proposed multi-class AdaBoost strategy is improved in this research to build an approach known as MulticlassAdaBoost-CNN. The multi-class ada boost is a popular ML algorithm with good prediction results. The CNN model is well known for its automatic feature selection from presented data. This work integrates the prowess of these algorithms, where both are the best of their kind, to achieve maximum prediction accuracy. In the typical acquisition, technique is employed. SAMME employs the actual value of the chance that an input sample belongs to several classes. R.

Assume the training dataset comprises the following elements: $(x_1, c_1), \ldots, (x_n, c_n)$, where x_i. The learned classifier can then determine what has yet to be seen. Each sample in the training data is given a weight, resulting in a data weight vector termed $D = \{d_i\}$, where $i = 1, 2, \ldots, n$.

The information loads are introduced by d_i=1/n. Then, M systems, for example, CNNs, are prepared successively. In the main cycle of the successive education method, the principal CNN loads are introduced arbitrarily and prepared for at least one age in light of the trouble. Feature selection is automatically made in the convolution layers of the networks. The primary quantity of assessors is prepared on all the preparation tests with a similar load of 1/n. There are no significant distinctions, for example, loads of various preparation tests for the principal CNN. In the wake of preparing, the result of the CNN is determined for preparing tests. MulticlassAdaBoost-CNN produces a K-layered yield vector for an information test as the CNN's result. The extended qualities are contained. Every part in the result vector addresses a genuine esteemed, evaluated expectation relating to a class. P(x_i)=[p_k (x_i)],k=1,...K is the result vector for an info test xi and it is given to the class with the highest probability when it is tested. The first CNN's output, $P^{m=1}(x_i) = [p^m{}_k{}^{=1}(x_i)]$, is utilised to update the data weights, $D = \{d_i\}$ by (1).

$$d^m{}_i{}^{+1} = d^m{}_i \exp\left(-\alpha_k \_\_\_{}^{-}{}_k{}^1 Y_i{}^T \log(P^m(x_i))\right), i = 1, \ldots . n \quad (1)$$

Where $d^m{}_i$ is the weight of the ith planning test used by the mth as a result of the ith providing the test, condition (1) is obtained from either the SAMME.R computation and is used in this work to refresh the example weights for a CNN. If the exponential of the mth CNN's result vector, Pm (x i), and the result name Y iT are linked, and their inside item has a high value, the dramatic capacity in (1) has a lesser esteem (in light of the negative sign). As a result of the low value of the outstanding capacity in (1), the weight of the preparation exam for the subsequent CNN, because the lasting result is near the mark preparation test has been prepared by the ongoing. The loads for all preparing tests for the ongoing CNN are refreshed, and afterward, they are standardised by separating them by the loads' all-out aggregate. The prepared CNN is saved, and the following CNN starts to learn. In the AdaBoost approach, a new plastic new classifier is haphazardly initialised and educated to be accompanied. The conventional AdaBoost doesn't fit for CNN because, for countless preparation tests, CNN brings about significant connections between's the ideal names, Y_i^T, and the actual results of the CNN. The worth of the remarkable component in (1) for the matching examples is like this diminished by the high relationships. Subsequently, a tiny subset of preliminary tests not prepared by the earlier CNN has essential qualities for the loads. The modest number of undeveloped

examples is rather than the significant number of CNN learning boundaries. The ensuing CNN, which depends on the conventional AdaBoost approach, is focused on a predetermined number of undeveloped information. The cutting-edge CNN is compelled to become overfitted on the short assortment of information since it was prepared entirely without any preparation on such few preparation tests. Likewise, preparing a CNN from starting costs a ton of processes.

It is recommended that the learning boundaries of the prepared CNN in the ongoing cycle be moved to the succeeding CNN so it masters utilising the moved boundaries instead of beginning the CNN's preparation from an irregular starting state. One entrance part of CNN is move realising, which helps the succeeding CNN keep up with the earlier information it got during the previous CNN's learning. The moved CNN doesn't need broad preparation over an enormous number of learning ages since it learns well about the whole dataset. The computational expense is diminished by moving the ongoing learning boundaries to the accompanying CNN. During the exchange stage, the new CNN goes through a similar cycle as the ancient one, which includes preparing the new CNN for one age, removing the prepared CNN yield vector from each preparing test, utilising the result to refresh the information loads, D={d_i}, lastly normalising the loads. For each CNN in AdaBoost, the cycle is rehashed.

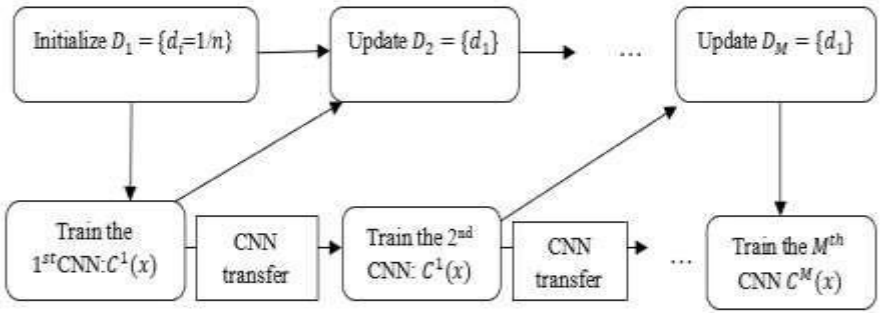

Figure 3. The Schematic Diagram of the Proposed MulticlassAdaBoostCNN Works Based on CNN Transfer Learning

Figure 1 depicts the MulitclassAdaBoost-suggested CNN's schematic diagram. The initial data

1 weight trains the first CNN once $D_1 = \{d_i = {}_n\}$. $D_2 =$ $\{d_i\}$, are then updated using the first CNN,

$C^1(x)$. Also, the second CNN receives the trained $C^1(x)$ model. This process is carried over repeatedly to train the M th CNN, $C^M(x)$.

3.3.1 Training a CNN with a Weighted Sample

NewMany convolutional layers, a pooling layer, and a fully linked layer are often stacked to create a CNN. The low-level features are collected in the bottom levels of a CNN's hierarchical structure, while the sophisticated features with more abstract information are extracted at the top layers. A CNN's bottom layers it to the subsequent layer in various feature maps. To translate an input into a feature map, CNN employs a set known as a kernel, W. Assume that the lt$\hbar$ layer contains a variety of feature maps. It is possible to determine y_i^l using (2).

$$y_{il} = \sum_j f(w_{il,j} * y_{il-1} + b_{li}) \quad (2)$$

bill addresses the predilection associated with the ith feature map within the lth tier, wherein w (i,j)l is the multi-layer portion used to decode the jth showcase map in the (l-1)th layer toward the ith highlights map in the denoted layer. The variable b is the bias value. The convolutional layer utilises a nonlinear enactment capability, like the sigmoid capability, f (.), or the corrected straight units (ReLU) capability. The convolutional administrator sign is shown by the '*'. After each convolutional layer, a top pooling layer is applied, passing the most extreme worth through a

neighbourhood window. The pooling layer diminishes the number of elements, bringing down the computational expense.

The past convolutional layers are associated close to completely associated hidden layers. The convolutional layers' gathered elements are levelled and handled to the completely associated layer.

$$F^i = f\,(W^l(F^{l-1})^T + b^l) \quad (3)$$

where b^l is the predisposition related to the last secret layer, W^l is the weight framework interfacing the last secret layer to the one preceding it, and F^(l)is the result of the last secret layer. A non-direct capability is f(.). Before being applied to the accompanying completely associated layer, the result of the last convolutional layer is smoothed to a vector.

The past layers are layered on top of a strategic relapse model to make an all-out yield. The relapse model's result is changed into the likelihood dissemination of the classes utilising the SoftMax capability, as demonstrated in (4).

$$Z = \text{softmax}\,(W^0(F^L)^T + b^0) \quad (4)$$

Where L seems to be the number of result neurons, which is equal to the total number of classes; Z is the organisation's result vector, with a component comparing to each class; W0 is the weight grid that connects the result layer to the result of the previous totally associated layer; FL is the result of the last wholly associated secret layer; and b0 is the inclination connected with both the result layer.

The back spread learning calculation is utilised to prepare the CNN. The learning calculation ascertains mistakes utilising cross-entropy. Each example in this study has a weight, di, and the example loads are presented in the mistake capability as shown in (5).

$$E_i = -\sum_{Lc=1} t_{ci}\log(z_{ic})d_i \quad (5)$$

E_i s the blunder related to the ith test, t_i^c is the ith test's comparing component of the mark vector, z_i^c is the ith test's comparing component of the result vector, and d_i is the example weight for the ith input test.

Table 1. Pseudo Code of the Proposed Muticlass AdaBoost-CNN

Table 1: Pseudo code of the proposed Multiclass AdaBoost-CNN.

Initialize the ith data sample weight with $d_i = \frac{1}{n}$ where $i = 1,2,\ldots,n$, and n is the total number of training samples, and initialize M, i.e. the total number of CNNs.

For m=1 to M:

1) *If* $m == 1$:

Train the first CNN, i.e. $C^{m=1}(x)$, on the training data using the initial sample weights, $D_{m=1} = \{d_i = 1/n\}$.

else:

Transfer the learning parameters of the previous CNN, $C^{m-1}(x)$, to the mth CNN, i.e. $C^m(x)$.

Train the mth CNN, i.e. $C^m(x)$ on the training data for one epoch using the sample weight vector $D_m = \{d_i\}$.

2) Obtain the output of the mth CNN, i.e. class probability estimates, for all the K classes:

$p_k^{(m)}(x)$ where $k = 1,2,\ldots,K$.

3) Update the data sample weight D_m based on $p_k^{(m)}(x)$ using (1).

4) Re-normalize the updated data sample weights, D_m.

5) Save the mth CNN, i.e. $C^m(x)$.

Table 1 presents the suggested MulticlassAdaBoost for CNN pseudocode in its entirety. The classifier for that iteration of the sequential learning approach is initially taught using training data and appropriate data weights, D=di, in each iteration of the technique. The data weights are changed

for the next group based on the outcome of the classification algorithm. For M weak classifiers, these two procedures are carried out successively.

3.3.2 *Testing with MulticlassAdaBoost-CNN*

The The M CNNs and M base classifiers were trained, and the resulting MulticlassAdaBoost-CNN is now prepared for testing, application (6),

$$C(x) = \underset{k}{\operatorname{argmax}} \sum_{m=1}^{M} h_k^m(x) \quad (6)$$

where h ( )ome is calculated by (7).

$$h_k^m(x)\ (K-1)\left(\log\left(p_k^m(x)\right) - \frac{1}{k}\sum_{k=1}^{k}\log\left(p_k^m(x)\right)\right)$$

After applying x as its input, $p^m{}_k(x)$ is the kth member of the output vector of the mth CNN. (7) was discovered using a multi-class AdaBoost and the Lagrange optimisation method on a constrained issue.

3.4 *Performance Measures*

Route The paper's findings fall into four categories: true positive (TP), false positive (FP), true negative (TN), and false negative. These categories are determined by combining the sample's correct category with the model's predicted category (FN). The greater the score on the central pixel and the smaller the number on the inter - and intra, conditional probability yields precision, recall, and F1score. To assess the training and test outcomes of the model for both the probability model of ecommerce consumers' purchase intention, the precision variable AUC values are all included. The following is the precise formula: Regarding accuracy, the ratio of the total is as mentioned in Equation 11. The precision, recall, F1 score and AUC are mentioned as Equations 8-12. TP

$$\text{Precision} = \frac{}{TP+FP\ TP} \quad (8)$$

$$\text{Recall} = \frac{}{TP+FN} \quad (9)$$

The F1-score is the total assessment recall and precision. "F1" scores range from 0 to 1, with 1 being the best possible result. The equation reads as follows:

$$F1 = \frac{2\times\text{Precision}\times\text{Recall}}{\text{precision}+\text{Recall}} \quad (10)$$

Accuracy is measured as the proportion of correctly identified samples to all samples. The standard cannot report the classifier's error categories or reflect the possible distribution of response value. However, it is simpler to comprehend. The equation reads as follows:

$$\text{Accuracy} = \frac{TP+TN}{(TP+FP)+(TN+FN)} \quad (11)$$

The binary classification model's evaluation measure, AUC (Area under the curve), shows that positive and negative samples were randomly chosen. The area under the ROC curve, which measures the probability that the positive sample's score will be larger than the negative sample, can be accurately provided by the classifier. The formula follows The more significant the area, the better the model effect.

$$AUC = \frac{\int_0^1 TPdFP}{(TP+FN)(TN+FP)} \quad (12)$$

4. Results and Discussion

The dataset comprises 400 customers buying information, including the unique ID, gender, age, and salary. The target variable to be predicted is the customer's decision to buy the product. The original data got refined in this experiment, and the customer information with return records was selected somewhere to examine the consumer's willingness to acquire the acquired items under the

return process. Thirty per cent of the filtered data are utilised as the test set, while the remaining seventy per cent are used as the training set. On the Kaggle data set "https://www.kaggle.com/competitions/instacart-market-basket-analysis/data," experiments were conducted.

Table 2. Basic Characteristics Table

Feature variable	Feature name	Remarks
User_id	User ID	Field of the prediction result set
Product_id	Product ID	Fields of the prediction result set
Order_number	Order ID	Medium importance
Add_to_cart_order	Item ID in order	Higher importance
Recordered	Repeat purchase logo	Fields of the training data set
Order-dow	What day of the week the order was placed	Generally important
Order-Hour_of_day	Time of order	Generally important
Days_since_prior_order	Number of days since the last order	Medium importance

Table 2 displays the chosen features. The features make up the majority of the 30 features that were chosen. Table 3 displays the chosen features.

Table 3. User Product Feature Table

Feature variable	Feature name
Avg_cart_priority	Average value of user product purchase priority
Times_ord	Number of purchases
Log_freq	Log value of user's historical purchase frequency
Days_to_last_order	The number of days since the most recent order
Order_streak	The number of other orders in the repurchase interval
Weekend	Weekend purchase logo
Hod_delta	Actual and average purchase time difference (hours)
Dow_delta	Actual and average purchase time difference (weeks)
Orders_since_last	The number of orders purchased since the last order

Table 4. Comparison of Prediction Results under Different Models

Model	**Accuracy**	**Precision**	**F1**
LR	0.7615	0.7176	0.7504
SVM	0.7705	0.7231	0.7435
RF	0.7910	0.7535	0.7846
XGBoost	0.8008	0.7532	0.7957
MulticlassAdaBoost-CNN	0.8493	0.7970	0.8324

In Table 4, the experimental findings are displayed. MulticlassAdaBoost-CNN is more accurate than the three LR, SVM, RF, and CNN algorithms. MulticlassAdaBoost-precision CNNs increased by 6%, 4%, and 3%, respectively. Correspondingly, there is a 5%, 3%, and 1% increase in the F1 of MulticlassAdaBoost-CNN, with CNN performing somewhat better. Table 5 displays the experimental findings of each comparison algorithm after adding noise data of 1%, 3%, and 5% to the initial data set.

Table 5. Comparison of Experimental Results after Adding Noise

Noise ratio(%)	**Algorithm**	**Accuracy**	**Index Precision**	**F1**
1	LR	0.7435 0.7602	0.7013 0.7112	0.7334 0.7472
	SVM	0.7843 0.7914	0.7535 0.7564	0.7655 0.7784
	RF	0.8352	0.7923	0.8014
	XGBoost			
	MulticlassAdaBoost-CNN			
3	LR	0.7403 0.7523	0.6832 0.6934	0.7161 0.7256
	SVM	0.7785 0.7801	0.7473 0.7421	0.7535 0.7678
	RF	0.8284	0.7839	0.7914
	XGBoost			
	MulticlassAdaBoost-CNN			
5	LR	0.7205 0.7392	0.6656 0.6745	0.7010 0.7062
	SVM	0.7536 0.7642	0.7311 0.7213	0.7445 0.7143
	RF	0.8157	0.7762	0.7825
	XGBoost			
	MulticlassAdaBoost-CNN			

This is due to the algorithm's introduction of a fuzzy method, which improves noise resistance. Nonetheless, there will undoubtedly be some noise in the data gathered from real-world production and life. XGBoost and MulticlassAdaBoost-CNN have predictive effects. Based on the two models, MulticlassAdaBoost-accuracy, CNN's precision, and F1 are improved. It is possible to deduce The outcomes in Table 5 display that the prediction performance of each method decreases with increasing noise. When the noise level is increased by 1%, the prediction accuracy of LR, SVM, RF, Support vector regression, and MulticlassAdaBoost-CNN are abridged by 2.3%, 1.3%, 1.2%, 2.4%, MulticlassAdaBoost-CNN are reduced by 1.9%, 2.5%, 4.6%, 3.8%, and 2.3%, respectively. LR, SVM, RF, XGBoost and MulticlassAdaBoost-CNN's precision is decreased by 5.3%, 4.5%, 7.1%, 5.8%, and 3.7%, respectively, when the noise is increased by 5%. The MulticlassAdaBoost-CNN method has the lowest reduction rate, as observed from the reduction rate. This demonstrates how much more noise-resistant the algorithm is.

Table 6. The Testing Accuracy Values

#Layers	Testing accuracy (%)	Layers
4	81.15	Fully-Droupout_fully_fully
5	70.81	Droupout_fully Droupout_fully__fully
6	90.91	Conv_Droupout_fully Droupout_fully__fully
7	92.05	Conv_Pooling_Droupout_fully Droupout_fully__fully
8	91.05	Conv_Pooling_Conv_Droupout_fully Droupout_fully__fully
9	90.27	Conv_Pooling_Conv_Pooling_Droupout_fully Droupout_fully__fully
10	89.99	Conv_Pooling_Conv_Pooling_Conv_Droupout_Droupout_fully Droupout_fully__fully

The results in Table 6 show greater testing accuracy values for networks with seven layers. The suggested MulticlassAdaBoost-CNN of 94.08% is higher than the primary network's of 92.05% with the optimal number of layers or 7 layers.

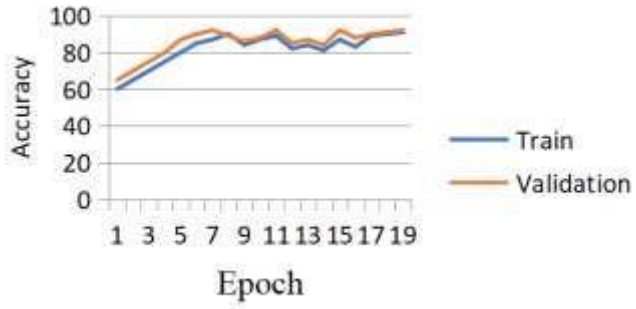

Figure 4. Training and Validation Accuracies of the Single CNN across Different Learning Epochs

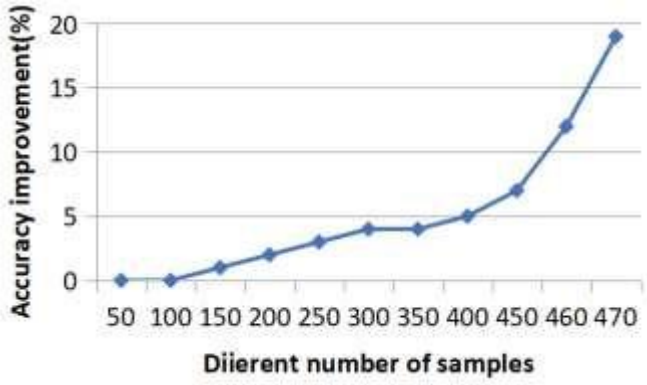

Figure 5. The Suggested Method's Testing Accuracy Improves as the Amount of Inequity, i.e. The discrepancy among the sample count in classes 2 and 3, inside the information

Figure 6 illustrates the increase in testing correctness when employing AdaBoost-CNN compared to a solitary CNN when the unbalanced dataset's sample count is varied.

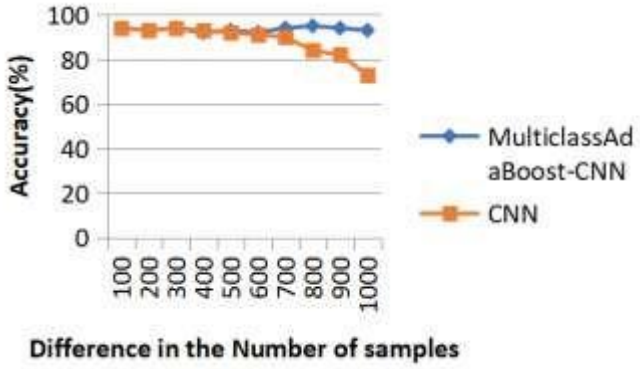

Figure 6. The Suggested AdaBoost-CNN and CNN Testing Accuracies for Various Levels of Imbalance, i.e. the alteration in the number of models in classes 2 and 3

The proposed approach improves exactness for higher imbalances in exercise data, as shown in Figure 7

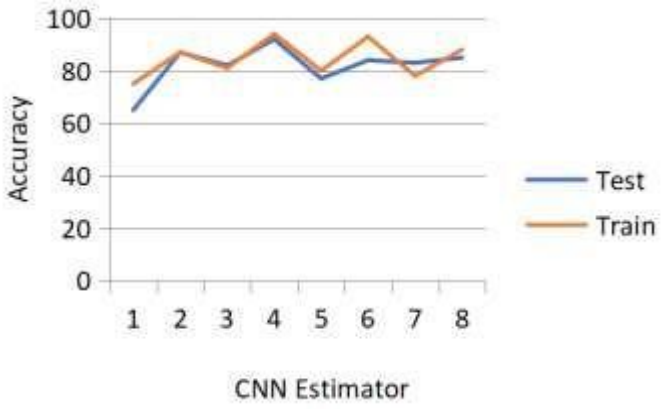

Figure 6. Accuracy of Different CNN Estimators in the Proposed MulticlassAdaBoost-CNN Algorithm

Figure 5 depicts multiple CNN estimation techniques in MulticlassAdaBoost-CNN. Several CNN estimators are displayed on the 'x-axis in the proposed MulticlassAdaBoost-CNN. MulticlassAdaBoost-CNN employs ten CNN estimators. Fig. 8 shows that MulticlassAdaBoost-CNN has the greatest. The testing accuracy score is 91.85%. Despite possessing an accuracy result of 94.91% and receiving instruction for ten learning epochs, a single CNN can only reach a testing accuracy of
90.97%. As a result, reducing the dimensionality of a CNN may be beneficial by minimising the impact of previously trained samples in succeeding epochs. MulticlassAdaBoost-CNN performs better overall than the single CNN and the CNN estimators that comprise its ingredients. Testing accuracy of 94.08% can be attained using MulticlassAdaBoost-CNN. MulticlassAdaBoost-enhancement CNNs can be attributed to their capacity to avoid overfitting. It is recognised that prevention can result in very modest losses on complicated, complex titles above the tables.

5. Conclusion

Customer repurchase behaviour is one of the main factors that indicate customer satisfaction level in business. This work deploys an ensemble of CNN integrated with MulticlassAdaBoost, which learns the features automatically from the presented dataset. This model sequentially learns

the features automatically through cascaded CNN, which learns the trends and patterns inherent in the purchase data. If the customer and product ID are matched multiple times, it implies a repurchase has been made. The experimental analysis is done to test the efficacy of the model in terms of prediction accuracy, precision and F1 score. In addition, external noise is induced in the dataset, and the model exhibits improved performance than other state-of-the-art ML models. As a future enhancement, a temporal analysis of the data can be done using time series-based prediction algorithms like LSTM and RNN. This will further enhance the prediction efficacy by analysing the data along a purchase timeline of the customer.

References

[1] D. Koehn, S. Lessmann M. Schaal, "Predicting Online Shopping Behaviour from Clickstream Data using Deep Learning," *Expert Systems with Applications (Journal)*, vol.150, 2020, pp. 113342.

[2] X.W. Chen, X. Lin, "Big Data Deep Learning: Challenges and Perspectives," *IEEE Access (Journal)*, vol. 2, No. 2, 2014, pp. 514-525.

[3] A.A. Alalwan, Y.K. Dwivedi, NP. Rana, "Factors influencing adoption of mobile banking by Jordanian bank customers: Extending UTAUT2 with trust," *International Journal of Information Management (Journal)*, vol.37, no. 3, 2017, pp.99-110.

[4] W.A. Alkhowaiter, "Digital payment and banking adoption research in Gulf countries: A systematic literature review," *International Journal of Information Management (Journal)*, vol. 53, 2020, pp. 102102.

[5] R. Mashur, B.I. Gunawan, B. I., M.F. Ashoer, M. Hidayat, HPKP Aditya, "Moving from traditional to society 5.0: Case study by online transportation business," *Journal of Distribution Science (Journal),* vol. 17, No. 9, 2019, pp. 93-102.

[6] K. Machmud, "The Smartphone Use in Indonesian Schools: The High School Students' Perspectives," *Journal of Arts and Humanities (Journal)*, vol. 7, no. 2, 2018, pp. 33-40.

[7] F.A.W. Wirawan, E. Oktivera, "Analysis on the implementation of digital marketing towards motorbike transport service case study: GO-JEK (online taxi motorbike)," Jakarta, Indonesia. In *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*, 2015, pp.1-6.

[8] A. Indahingwati, A. Launtu, H. Tamsah, A. Firman, AHPK Putra, A. Aswari, "How Digital Technology Driven Millennial Consumer Behaviour in Indonesia," *Journal of Distribution Science*, vol.17, no.8, 2019, pp. 25-34.

[9] S. Soebandhi, A. Wahid, I. Darmawanti, "Service quality and store atmosphere on customer satisfaction and repurchase intention," *BISMA (Bisnis Dan Manajemen)*, vol.13, no.1, 2020, pp.26-36.

[10] Y.C. Chou, H.H.C. Chuang, "A predictive investigation of first-time cus-tomer retention in online reservation services," *Service Business*, vol. 12, no. 4, 2018, pp. 685-699.

[11] A. Martínez, C. Schmuck, S.P. Jr, C. Pirker, M. Haltmeier, "A machine learning framework for customer purchase prediction in the non-contractual setting," *European Journal of Operational Research*, vol. 281, no. 3, 2020, pp. 588-596.

[12] C.G. Mena, A. De Caigny, K. Coussement, K.W. De Bock, S. Lessmann, "Churn prediction with sequential data and deep neural networks," *A Comparative Analysis arXiv preprint*, 2019, arXiv: 1909.11114.

[13] A. Tandon, A. Aakash, G.A. Aggarwal, "Impact of EWOM, website quality, and product satisfaction on customer satisfaction and repurchase intention: moderating role of shipping and handling," *International Journal of System Assurance Engineering and Management*, vol.11, 2020, pp.349-356.

[14] R. Dew, A. Ansari, "Bayesian nonparametric customer base analysis with model-based visualisations," *Marketing Science*, vol. 37, no. 2, 2018, pp.216-235.

[15] A. Gopalakrishnan, E.T. Bradlow, P.S. Fader, "A cross-cohort changepoint model for customer-base analysis," *Marketing Science*, vol. 36, no. 2, 2017, pp. 195-213.

[16] L. NGUYEN, TH NGUYEN, T.K. Phuong, "An Empirical Study of Customers' Satisfaction and Repurchase Intention on Online Shopping in Vietnam," *The Journal of Asian Finance, Economics and Business*, vol. 8, no. 1, 2021, pp. 971-983.

[17] S. Perumal, J. Ali, H. Shaarih, "Exploring nexus among sensory marketing and repurchase intention: Application of SOR Model," *Management Science Letters*, vol.11, no.5, 2021, pp.1527-1536.

[18] GK Amoako, LD. Caesar, R.K. Dzogbenuku, G.A. Bonsu, "Service recovery performance and repurchase intentions: the mediation effect of service quality at KFC," *Journal of Hospitality and Tourism Insights*, vol. 6, no. 1, 2023, pp.110-130.

[19] C. Khanijoh, C. Nuangjamnong, K. Dowpiset, "The impact of consumers' satisfaction and repurchase intention on E-commerce Platform: a case study of the top three E-commerce in Bangkok," In *AU Virtual International Conference Entrepreneurship and Sustainability in the Digital Era*, vol.1, no.1, 2020.

[20] E.C. Mendoza, "A study of online customers repurchase intention using the 4Rs of marketing framework," *International Review of Management and Marketing*, vol.11, no.2, 2021, pp.1.

[21] C. Mondal, BC. Giri, "Analysing strategies in a green e-commerce supply chain with return policy and exchange offer," *Computers & Industrial Engineering*, vol.171, 2022, pp.108492.

[22] R. Yazdani, M. J. Taghipourian, M. M. Pourpasha, S. S. Hosseini, "Attracting potential customers in E-commerce environments: a comparative study of metaheuristic algorithms," *Processes*, vol.10, no.2, 2022, pp.369.

[23] Z. Hu, X. Li, C. Wei, H. Zhou, "Examining collaborative filtering algorithms for clothing recommendation in e-commerce," *Textile Research Journal*, vol.89, no.14, 2019, pp.28212835.

[24] L. Rajput, S.N. Singh, "Customer Segmentation of E-commerce data using K-means Clustering Algorithm," In 2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), *IEEE*, 2023, pp.658-664.

[25] Y. Gao, H. Liang, B. Sun, "Dynamic network intelligent hybrid recommendation algorithm and its application in online shopping platform," *Journal of Intelligent & Fuzzy Systems*, vol.40, no.5, 2021, pp.9173-9185.

[26] A. Salamzadeh, P. Ebrahimi, M. Soleimani, M. Fekete-Farkas, "Grocery Apps and Consumer Purchase Behavior: Application of Gaussian Mixture Model and Multi-Layer Perceptron Algorithm," *Journal of Risk and Financial Management*, vol.15, no.10, 2022, pp.424.

[27] L. Yang, X. Niu, J. Wu, "RF-LighGBM: A probabilistic ensemble way to predict customer repurchase behaviour in community e-commerce," arXiv preprint, 2021, arXiv:2109.00724.

[28] M. Zhang, J. Lu, N. Ma, T.C.E. Cheng, G. Hua, "A Feature Engineering and Ensemble Learning Based Approach for Repeated Buyers Prediction," *INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL*, vol. 17, no. 6, 2022.

[29] H. Zhang, J. Dong, "Prediction of repeat customers on e-commerce platform based on blockchain," *Wireless communications and mobile computing*, pp. 1-15, 2020.

[30] Y. Liu, X. Zhou, H. Yu, "3R model: A post-purchase context-aware reputation model to mitigate unfair ratings in e-commerce," *Knowledge-Based Systems*, vol. 231, 2021, pp.107441.

[31] Xiao, Jin, Ling Xie, Changzheng He, and Xiaoyi Jiang, "Dynamic classifier ensemble model for customer classification with imbalanced class distribution," *Expert Systems with Applications*, vol.39, no.3, 2012, pp.3668-3675.