Volume: 13 Issue: 06 Pages: 5169-5182(2022) ISSN: 0975-0290

Comparative Analysis of Different Machine Learning Algorithms to Predict Online Shoppers' Behaviour

Veena Parihar
Research Scholar, Department of Computer Science and Engineering, Career Point University, Kota
Veena2parihar@gmail.com
Surendra Yadav
Professor, Department of Computer Science and Engineering, Vivekananda Global University, Jaipur
syadav66@gmail.com

ABSTRACT

The trend of online-shopping has gradually increased and this trend is growing with a fast pace in the present scenario. As the trend of online-shopping is growing day by day, the prediction of consumer purchasing behavior and choices is becoming as a topic of curiosity for the researchers and business-organizations. It is very challenging to predict buying behaviour of clients in advance. The discovery of consumer purchase patterns in advance can be proven useful for increasing the growth of businesses and generation of revenue. This proposed research work is an effort to develop a framework that presents some useful insights and predicts consumers' shopping behaviour by applying effective machine learning techniques.The present research work studies and analyses the various aspects and dimensions of online shopping which may impact the experience of purchasing by examining the considered data-set. Further, the thorough study of different machine-learning classification algorithms was performed to be applied for developing a new and better model for analyzing the online purchase data. Some chosen algorithms were applied on the selected data-set and performance evaluation was done using the performance metrics. The algorithm that performed well in terms of accuracy and other factors were chosen for developing the new model.

Keywords – **Online shopping, customer prediction, future preferences, machine learning, e-commerce, recommendation system, classification**

Date of Submission: Apr 27, 2022 Date of Acceptance: May 25, 2022

1. INTRODUCTION

The evolution of internet and e-commerce has totally transformed not only the consumers' way of purchasing but also the marketing strategies of the businesses or brands for attracting and retaining the customers by customizing the communications for specific customers [1].This growth of such e-shopping platforms has also created challenges for the businesses or marketing professionals to run their businesses properly and successfully. The future of e-commerce is totally dependent on technology and professionals who are able to develop such personalized purchasing experience for the coming consumers to increase the purchase revenue [2]. With the benefits of this new trend, there are so many new challenges arise which have to be faced for implementing this new kind of framework as it is completely dependent on technology. There are many factors which may affect the business of e-commerce such as internet, proper UI, offering personalized choices, appropriate presentation of the product and many more [3,4].

The motivation of opting this particular research area and topic is that it is an emerging topic of interest in present time as huge number of consumers are opting for online shopping and also the duration of COVID has directed the uninterested consumers also towards e-commerce and this trend will continue in coming future also [5]. Thus, prediction of consumers' purchase pattern is crucial for running the any kind of online businesses. A marketing-firm which is capable to make predictions about its clients' buying and browsing behaviour will be able to add and retain more clients, thus, increased revenue [6]. The present research work is an effort to develop such improved model which can effectively predict the purchase behaviour by analysing click stream data of user sessions by the application of various Machine Learning (ML) techniques [7]. A dataset of users' session's data has been considered for analysis and feature importance is also considered to be applied to ML algorithms as it will perform better [8]. The most appropriate features have been fed to the training model for the improved accuracy and performance. The results of the experimental frame work states that the classic-gradient-boosting algorithm performed the best among all the classification algorithms. The key contributions of the work are summed up as follows.

1. Firstly, a dataset has been collected characterizing the users' different sessions data. Pre-processing and restructuring of data have been done.

Volume: 13 Issue: 06 Pages: 5169-5182(2022) ISSN: 0975-0290

2. Data analysis has been performed for generating various insights from the existing data for developing a better system.
3. A variety of ML classification algorithms have been considered studied and applied. The performance evaluation of all the techniques has been performed.
4. The best performed classification algorithm has been considered for designing the new improved framework for predicting the consumer behaviour.

The organization of the paper is in following manner. The first section presents introduction about the area of research work. The second section provides an overview about the trends of online shopping. The third section states the benefits of predicting consumer buying behaviour. The fourth section reviews the literature of the related work. The fifth section is all about the research methodology and proposed work. The sixth section presents the performance analysis of all the techniques whereas the paper ends with the last section that is conclusion.

2. CONSUMER-BEHAVIOUR ESTIMATION

Consumer-behaviour is a factor that involves the track record of whole journey of purchasing any product that how an item was reached, selected, purchased and what other items were viewed etc. [9]. Various factors are there that may affect this behaviour as shown in the fig 1 below.

Fig. 1 Various aspects of consumer behaviour

To figure out the choices and interests of consumers, there is a need of combining the operation-data and experience-data together which are called O-data and X-data [10]. The O-data includes information about finance, sales or HR whereas the X-data may involve the information about consumer satisfaction score and net-promoter score. This combination of this O-data and X-data may provide the businesses a thorough understanding about the growth, revenue generation and consumers' purchase pattern for future insights [11]. By predicting behavioural patterns of consumers in advance, one will be able to figure out the points such as reduction in consumer churn, identification of high valued consumers, encouragement of loyal clients, reaching to the demand properly, less expenditure on market campaigns and improving the consumer experience etc. [12].Thus, the businesses are following some tactics for winning more number of customers, increasing their revenue and performance which are as follows.

2.1. Decrease the consumer churn

Consumer churn can be described as the number of consumers who leave or stop purchasing from a company in particular time duration. It can have significant impact on the growth and revenue of a company [13]. Generally, the strategies are made for targeting new clients but it is far more costly to acquire new clients rather than retaining the old ones. Thus, by analyzing the X-data and O-data combinedly, the possible reasons of churn can be identified for the growth of business [14].

2.2. Improve the consumer retention

Consumer-retention is an important aspect for growing rate of business. By examining O and X data properly, it can be identified which consumers will stay and who is about to left. Thus, by planning some actions and strategies accordingly may help in retaining the likely to churn consumers. This reduces the requirement of targeting new customers as well as increases the profit [15].

2.3. Improved consumer-satisfaction

Having an understanding and idea about the consumers' demands, interests and satisfactory levels are essential for the successful execution of a business. It may be proven useful in creating more revenue and recommending new consumers for purchase [16]. So, the increment in consumers' satisfaction level can greatly increase the consumer loyalty and purchase rate.

2.4. Improved consumer involvement

By collecting the users' experience data through the surveys, feedbacks or social-media, the metrics of consumer involvement can be acquired such as what are motivations that attract a consumer towards your business [17]. After the exploration of such motivational data, the mapping of such information can be done for achieving proposed business outcomes like sales, customer satisfaction scores and NPS etc. [18].

2.5. Triangulation of Customer-Experience (CX) data

Data-triangulation can be defined as the validation of the data from multiple resources. Such approach may provide the more accurate data for the analysis that results into more accurate and precise insights [19]. Triangulation of data mainly depends on the process of collecting, examining, and actioning on the CX data for predicting the behavioural patterns before a transaction occurs. As an example, the NPS data is a measure of how much the customers are satisfied with a company's products. It can be useful in developing regression models and analyzing

Volume: 13 Issue: 06 Pages: 5169-5182(2022) ISSN: 0975-0290

the clients particularly on their own terms and conditions [20].

2.6. Utilization of product feedbacks

Listening to consumer feedbacks can be a significant factor in improving the overall engagement. The feedbacks are the mirror image of what a consumer wants and what motivation can retain him to your products [21]. Actions taken in time according to the feedbacks are useful in enhancing the user-experience as well as creating good products and services according to the user needs. By applying this tactic, there is a reduction in the problem generated because of a bad product, thus, increased focus on the revenue generation ideas [22].

2.7. Tracking of brand-value

The task of analyzing and examining the consumer demands is an ongoing and never-ending process as there are very dynamic. For creating a suitable CX-program, it is very important to track the brand health. Tracking the trend of your products provides a clear picture of where you stand in this fast-paced economic environment also helps in providing the insights about the future trends before their occurrence. This will also be useful for identifying the effects of consumer experience on existing brand [23, 24].

2.8. Apply the ML techniques

The application of ML techniques is tremendously useful for making the predictions about the consumers' choices, interests and requirements [25]. These techniques turn the huge amount of the data into significant predictions. It enables the brands and businesses towards creative and critical thinking through the data analysis. With the automation of this prediction process, one will become capable of taking quick operational and marketing actions to resolve the problems before they become public [26].

3. RELATED WORK

The prediction of online buying behaviour has studied by various researchers of this field [27]. The current researchers have developed various frameworks to inspect purchasing predictions by utilizing the data of previous users' sessions [28]. A web-site for a e-commerce business is an important aspect for increasing sales and marketing the products among customers. The studies shows that a web-site of any e-business may increase the popularity and revenue and build the trust in the people [29].Experimental findings presented that integrating the behavioural attributes of users to the ML techniques increases the performance of ML algorithms to predict the buying pattern.

Zeng et al. (2019) have proposed a prediction model of that analyses the purchase behaviour of China people during a festive season. A finding says that if a user is interested towards a product, then he/she spend more time to investigate that particular product [30]. Mokryn et al. (2019), analysed the Logistic-Regression (LR), Decision-Tree (DT) classification and bagging to examine the impact of the attributes such as time, trendiness of the product etc. on the performance of predicting purchase pattern and it was discovered that bagging method performed the best among all [31]. Wu et al. (2015) presented a model of predicting buying behaviour that aims to identify click-stream patterns instead of session attributes. The findings of experiment proven that by using features of click-stream patterns, the accuracy of prediction can be enhanced [32].

Bo Wang et al. (2018), presented a customized recommendation frame work on the basis of user feedbacks generated implicitly. The input to the framework is buying behaviour, their comparisons and product sequences. This whole data can be accessed through the session logs of the users [33].

Andres Ferraro et al. (2018), developed a new method for improving the recommendations on the basis of a metric-of-choice by combining various different ML algorithms for individual users based on their performance. The proposed technique is able to forecast any probable fault, that a system may produce for every user on the basis of their previous behaviour. Thus, the new method suggested was a regression-model considering various metrics predicting performance of the system based on the parameters that predict system performance based on a variety of parameters that describe the previous activities of the users [34].Reis et al. (2018) aims at providing insights about the whole digital-transformation and future directions for the research. By considering almost 300+ articles, this paper tends to present a thorough review of related literature [35].

According to Çelik et al. (2019), with the growing use of social-networking sites, the business organizations have created the reach to customers by delivering their services and products to the social- media platforms. Every person has different choices of products or services. Gender is an important aspect to figure out the interests. The social-media audience can be targeted on the basis of gender to increase the revenue by creating offers and discounts on particular products. The prime objective of this study is to estimate the gender of social-media commentators through the application of ML algorithms mostly by analyzing the names[36].

Kai Wang et al., 2019 proposed a customized product recommender system based on clustering. The naïve k-Nearest Neighbour (KNN) method has its own limitations of selecting near-by data-points. For developing this product recommender framework, the Recurrent Neural Network (RNN) and attention methods were merged. This research work was proven to be successful to solve the issues of diversity and scalability [37].

Kim et al. (2020) proposed a new system for predicting buying behaviour in real-time by analyzing the users' pattern of visiting physical stores. They settled up some cameras and sensors and applied the object-detection techniques to record the purchase actions. It was a very

Volume: 13 Issue: 06 Pages: 5169-5182(2022) ISSN: 0975-0290

challenging and expensive task. The models of making prediction about purchase are effective, configured easily and well-integrated with the existing system [38].

Esmeli et al. (2020) proposed and developed a novel-framework to predict the consumers' purchasing intention at the early phase. The framework is capable of personalizing the content and providing offers and discounts accordingly. They analysed the purchase behaviours of the users by discovering the hidden patterns. These patterns are useful to create customized marketing schemes for individual customers, thus, improving the sales ultimately [39].

4. PROPOSED METHODOLOGY AND WORK

The trend of online shopping is evolving day by day rapidly. It is an important aspect to predict consumer purchase behaviour in advance for automating the marketing strategies for more revenue generation and targeting the customers [40]. The prime objective of this research work is to propose such a framework that predicts the shoppers' intention by applying the best ML algorithm in terms of performance and accuracy.

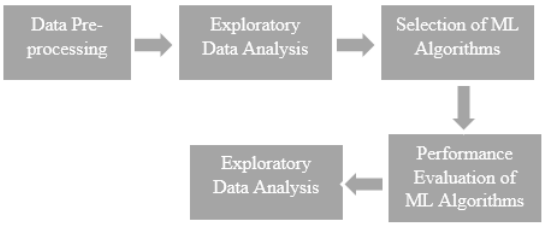

Fig. 2 Flowchart of proposed methodology

4.1. Dataset Description

The data-set used here is Online-Shoppers'-Intention opted from UCI repository that consists of 18 features which includes both, categorical and numerical features, collected from around 12330 sessions belonging to different users over a period of one year [41].Almost 85% samples are negative and rest are the positive samples which indicates a transaction was completed. The proposed work was carried out in two phases- first is analysis of data-set and selection of classification-algorithm. In the first phase, various factors affecting the consumer purchase, were analysed and in the second phase, different ML algorithms such as Support Vector machine (SVM), Ada-Boost, Naïve Bayes (NB), Random Forest (RF), Gradient Boost (GB), KNN and LR classifier [42].

4.2. Exploratory Analysis

This section tries to analyse various aspects that may create an impact on the consumers' purchase behaviour when handled properly [43]. Thus, by examining those features useful suggestions and recommendations can be made for the growth of e-businesses. The analysis of the significant factors is as follows.

4.2.1. correlation among various factors

By observing the following plot in fig3, it can be seen that there is a very less correlation the different attributes of the considered data-set. Very few attributes are there which are highly correlated i.e., the value of correlation is greater than or equal to 0.7 such as Exit-Rate & Bounce-Rate and Product-Related and Product-Related-Duration. The other attributes are less correlated i.e., the value of correlation lies between 0.3 to 0.7

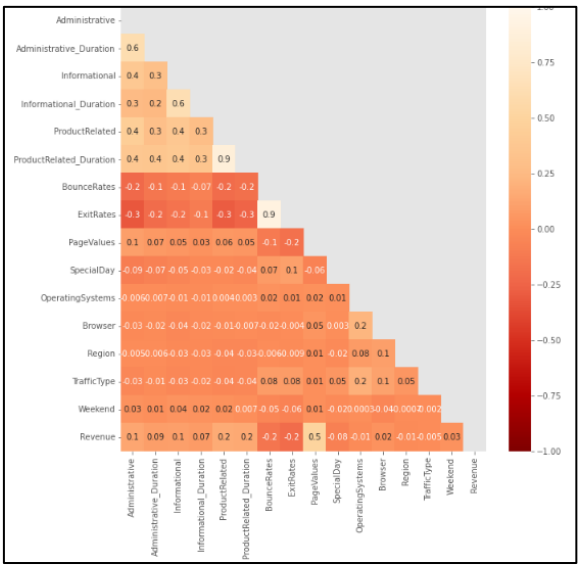

Fig.3 Relation between different features

4.2.2. Analysis of web-pages

The analysis of the various attributes related to web-pages was performed and the following graphs were plotted against the revenue attribute.

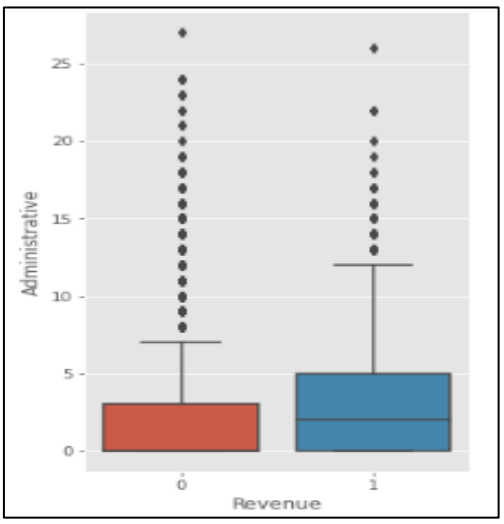

Fig. 4 Plot between administrative and revenue

Volume: 13 Issue: 06 Pages: 5169-5182(2022) ISSN: 0975-0290

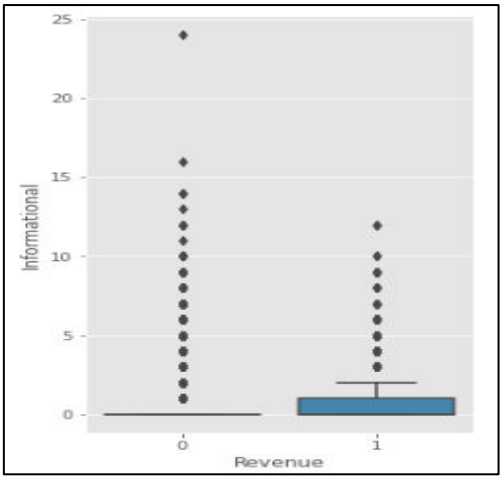

Fig. 5 Plot between informational and revenue

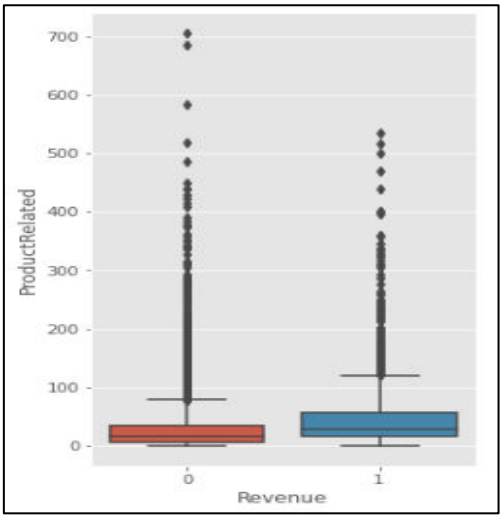

Fig. 6 Plot between Product Related and revenue

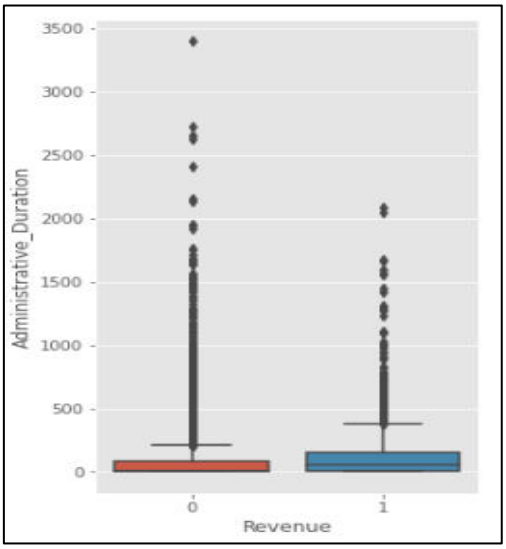

Fig. 7 Plot between Administrative Duration and revenue

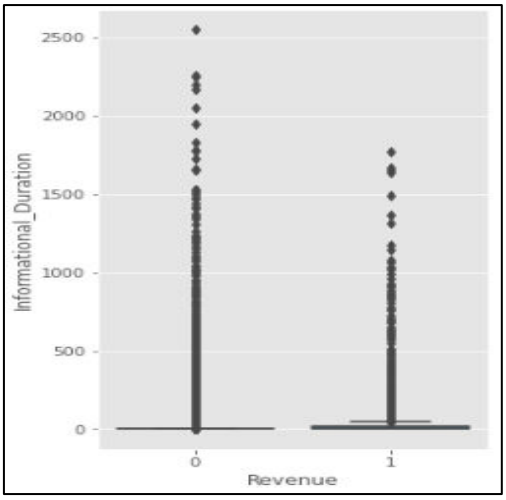

Fig. 8 Plot between Informational Duration and revenue

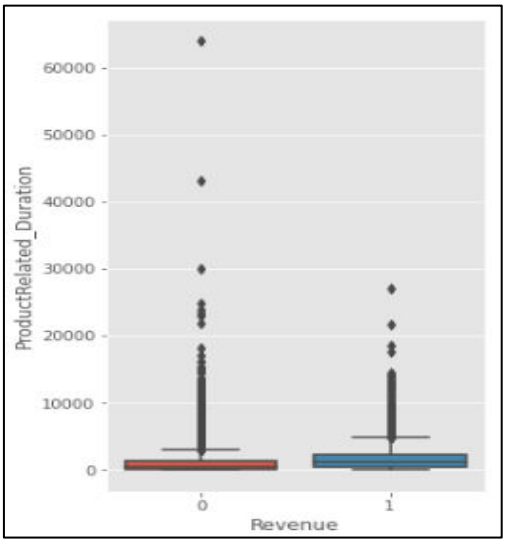

Fig. 9 Plot between Product Related Duration and revenue

By examining the above plotted box-plots in fig. 4 to fig. 9, it can be observed that generally the viewers visit a lesser number of pages and for less time duration if they are not intended to purchase something. The time spent on the Product Related pages is higher than the time spent on the Administrative or informational web-pages. The very first three attributes show a skewed distribution towards the revenue feature.

4.2.3. Analysis of Page-Metrics

The all three google-analytics metrics related to bounce-rate, exit-rate and page-value have been visualized in this section.

Volume: 13 Issue: 06 Pages: 5169-5182(2022) ISSN: 0975-0290

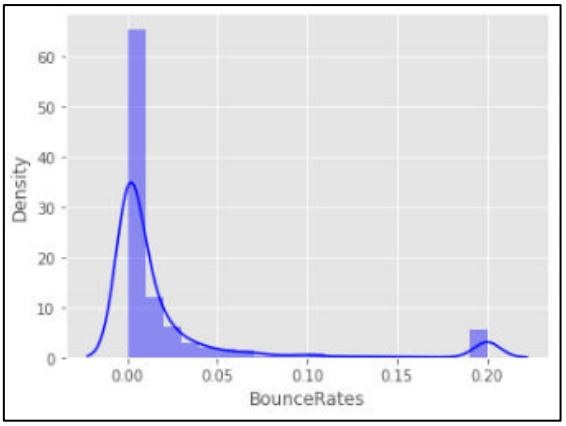

Fig. 10 Distribution of bounce-rate

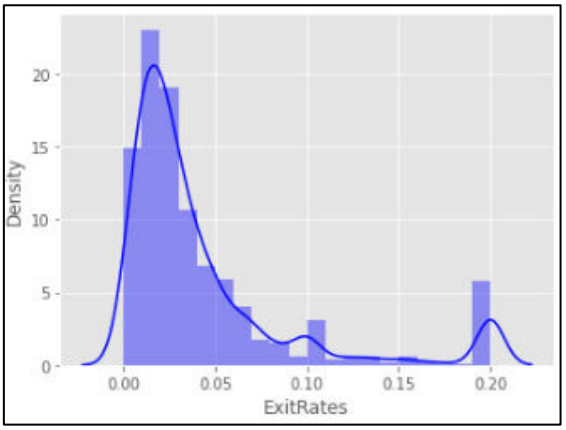

Fig. 11 Distribution of exit-rate

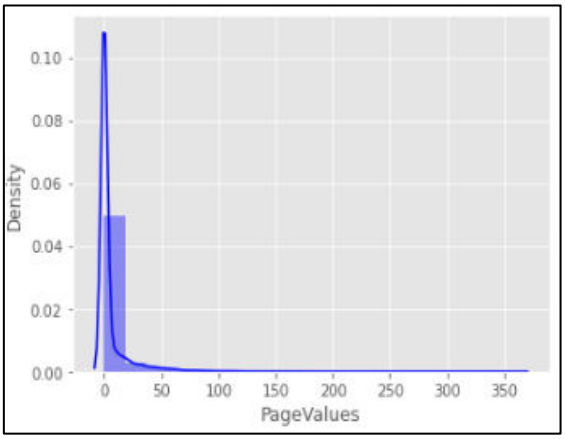

Fig. 12 Distribution of page-values

By observing the above three plots in fig. 10 to fig. 12, it can be pointed out that the bounce-rate and page-value attributes are not having a normal-distribution. All the three attributes are skewed towards right direction and have more outliers. The average value of exit and bounce rates is subsequently low for most of the data-points which is a good indication that customers are visiting and engaging themselves with the web-site. Exit-rates has higher values than the bounce-rate attribute which is good.

4.2.4. Visitor-Analysis Based on Various Factors

In this section, the analysis of visitors has been presented on the basis of various aspects such as type of operating system or browser used, from which region the users are belonging to and the type of the traffic. By discovering the insights through this data, appropriate strategies may be designed for improving the existing system.

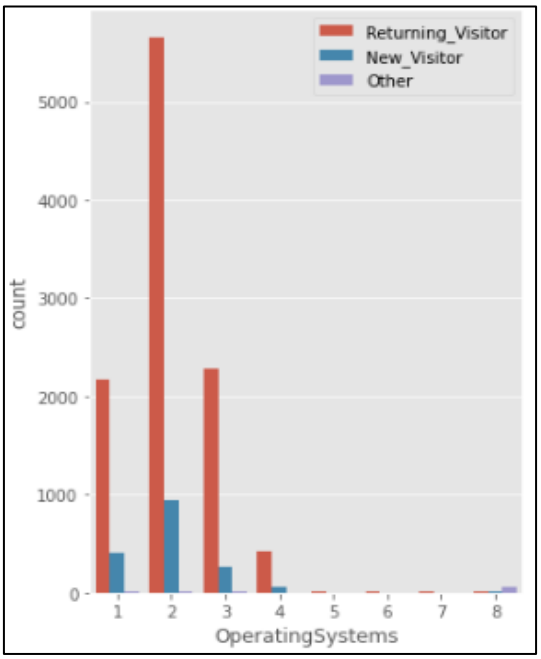

Fig. 13 Visitors analysis based on operating system

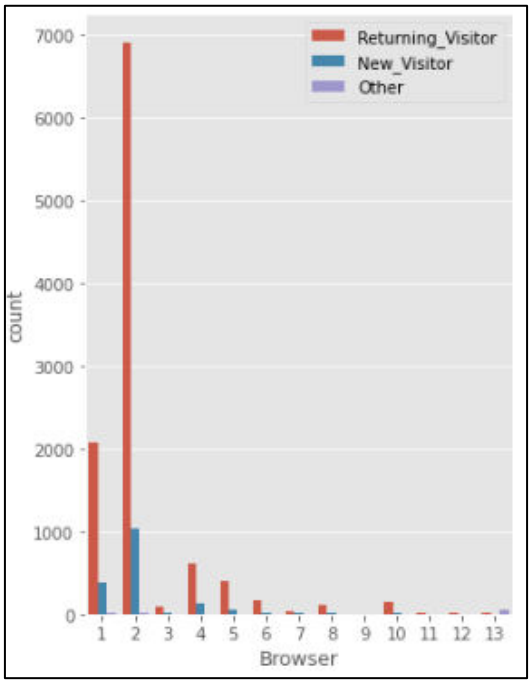

Fig. 14 Visitors analysis based on browser

Volume: 13 Issue: 06 Pages: 5169-5182(2022) ISSN: 0975-0290

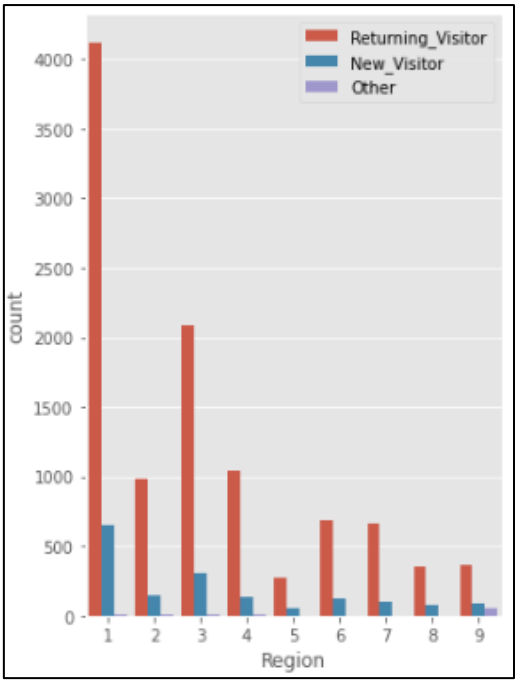

Fig. 15 Visitors analysis based on region

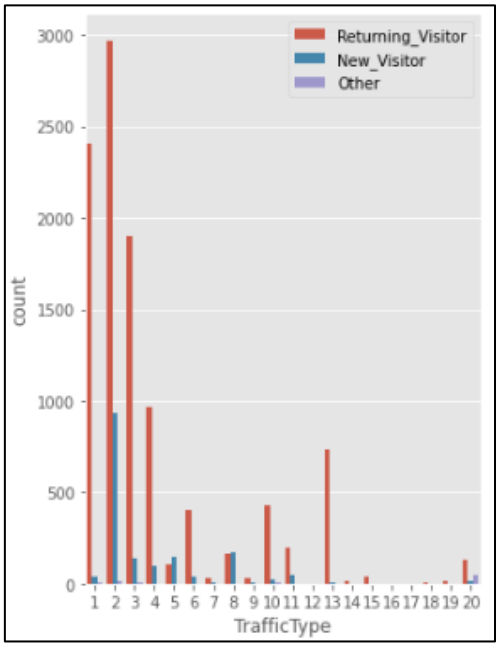

Fig. 16 Visitors analysis based on traffic-type

By examining the above fig. 13 to fig. 16, it can be observed that the operating system of category 2 is having around 7000 samples in the dataset i.e., it is mostly used by the users and the second most popular category is category 1 which have around 2000 samples. It indicates that there is need to design such UI which is compatible with the all categories of operating systems. The similar situation is with the categories of browsers where category 2 is dominant followed by category 1 and rest of the browsers are used very rarely. The plot related to region entails that customer from the region of category 1 are mostly involved followed by category 2. A smaller number of customers are involved related to the other categories of the region. It means there should be appropriate marketing strategies to be applied to these regions for improving the customer involvement. According to the traffic analysis plot, the sources of traffic are very diverse and may be improved by improving ads and SEO optimization.

4.2.5. Analysis of Visit-Date

This section analyses the revenue data based on the month, weekend, week days and special days. By observing the following first plot, it can be concluded that in May and March month, there is significantly huge number of visits although the purchase rate is low comparatively. No visits are there in the month of January and April. A larger number of transactions have been completed at the yearend i.e., during November and December.

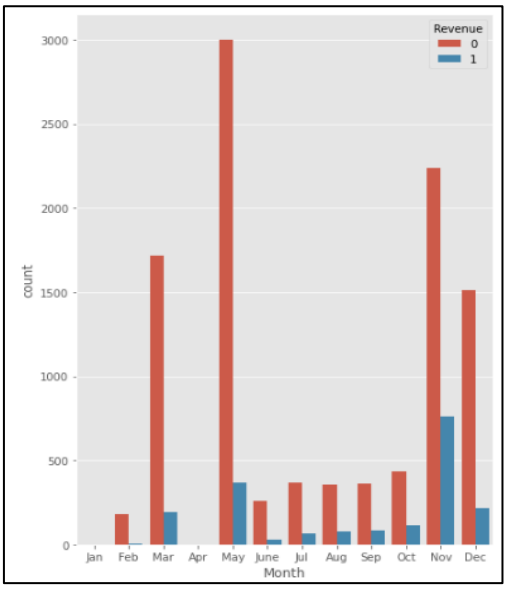

Fig. 17 Month based revenue analysis

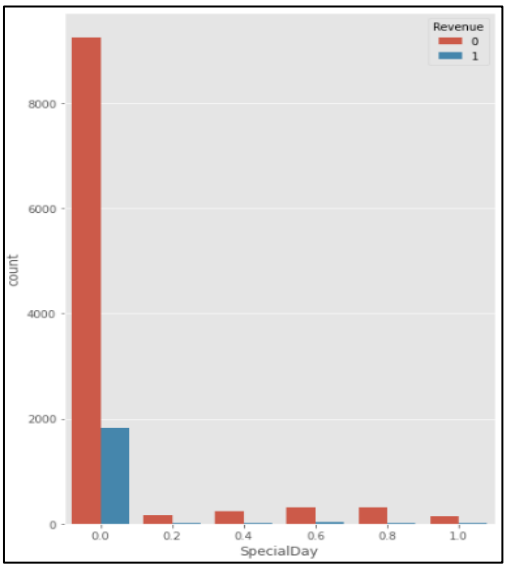

Fig. 18 Revenue analysis based on special-days

Volume: 13 Issue: 06 Pages: 5169-5182(2022) ISSN: 0975-0290

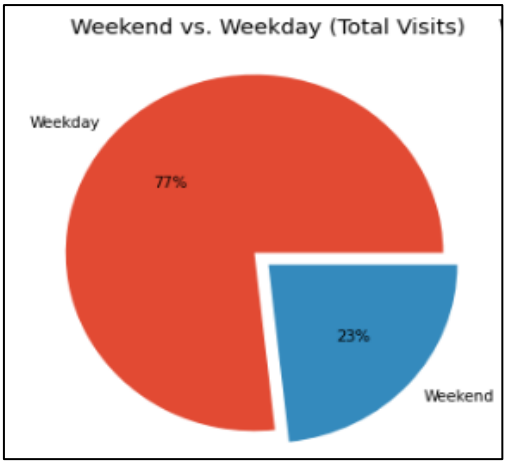

Fig. 19 Visitors analysis on weekend vs week-days

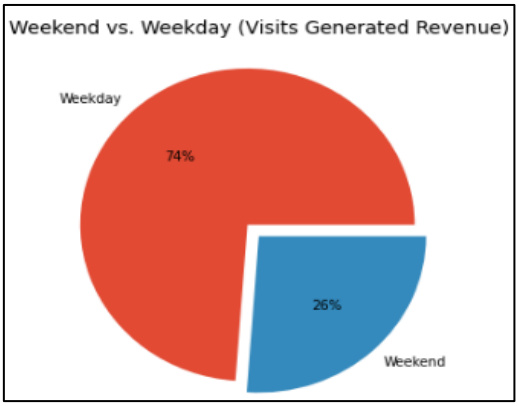

Fig. 20 Revenue generation on weekend vs week-days

By observing fig. 17 to fig. 20, the assumption can be made that if the visit date is near to a special-day, it will more likely to be ended with a transaction. Most number of the transactions occurred on the special-days. The weekend graph indicates that there is a slight increment in the transactions on weekends in comparison to the weekdays.

4.3. Selection of Classification Model

There are various ML algorithms that can be applied for classification as well as regression tasks. The proposed problem is related to classification i.e., revenue generated or not. There are various classification techniques related to ML which are going to be applied in this research work which are as follows.

4.3.1. Naive Bayes (NB) Classification

Some features of the data-set have normal distribution whereas the others are not normally distributed [44]. That is why the gaussian NB classification techniques has been applied in this work. The performance shown by the NB classifier is as follows.

```
Gaussian Naive Bayes Performance:
---------------------------------
Accuracy        :  0.775013520822066
F1 Score        :  0.4914425427872861
Precision       :  0.3941176470588235
Recall          :  0.6525974025974026
Confusion Matrix:
  [[1232  309]
 [ 107  201]]
```

Fig. 21 performance metric-NB classifier

The results are not up to the mark and less desirable asgaussian-distribution may not fit to all the attributes of the data-set.

4.3.2. KNN classifier

The scaled dataset version was used to be fed into KNN classifier.It is a type of non parametric algorithm [45]. The following performance was observed through KNN classifier.

```
K-Nearest Neighbor Initial Performance:
---------------------------------------
Accuracy        :  0.8734451054624122
F1 Score        :  0.5021276595744681
Precision       :  0.7283950617283951
Recall          :  0.38311688311688313
Confusion Matrix:
  [[1497   44]
 [ 190  118]]
```

Fig. 22 performance metric-KNN classifier

When compared to NB classifier, KNN classifier performs well in terms of f1-score and accuracy but there is a decrease in recall value. In next step, hyper-parameter tuning was applied to KNN for improving the performance through grid-search.

KNN Tuning

The following parameters were tuned to be fed into KNN classifier such as leaf-size, number of neighbours, distance-metric (p) and weights. Leaf-size may have impact on the memory and speed parameters.

```
K-Nearest Neighbor Tuned Performance:
-------------------------------------
Accuracy        :  0.8734451054624122
F1 Score        :  0.5063291139240506
Precision       :  0.7228915662650602
Recall          :  0.38961038961038963
Confusion Matrix:
  [[1495   46]
 [ 188  120]]
```

Fig. 23 Performance metric-tuned KNN classifier

Volume: 13 Issue: 06 Pages: 5169-5182(2022) ISSN: 0975-0290

Thus, by seeing above fig. 23, it is observed that there is a slight improvement in performance of KNN after the application of hyper-parameter tuning.

4.3.3. SVM Classifier

The dataset used for SVM was after scaling. SVM proposes a higher accuracy for classification task. It is able to handle the non-linear data through the kernel trick. This classifier represents the separation of data-points through a hyper-plane with a significant amount of margin [46]. The task of SVM is to find an optimum hyper-plane that can classify the new data-points effectively.

```
SVM Initial Performance:
------------------------
Accuracy         :  0.8875067604110329
F1 Score         :  0.5856573705179283
Precision        :  0.7577319587628866
Recall           :  0.4772727272727273
Confusion Matrix:
  [[1494   47]
 [ 161  147]]
```

Fig. 24 Performance metric-SVM classifier

With comparison to the KNN classifier, SVM offers significantly better performance in terms of all the metrics. In this direction, next the SVM is applied with tuning of hyper-parameters using grid-search.

SVM-Tuning

In the process of SVM-tuning, the parameters tuned were kernel, regularization and gamma.

Kernel - The task of kernel is to converts the existing data-set into a required format. Various functions are there to perform this task such as linear functions, polynomial functions and radial functions. These transformations of data-set can offer the classifiers with improved accuracy and performance [47].

Regularization- This process is used for avoiding misclassification of training data. It represented by alphabet 'C'. The lower values of C results into a hyper-plane of larger margin whereas the higher values of C will result into a small area of hyper-plane.

Gamma: The lower values of gamma consider the far away data-points whereas the higher values will consider only the nearby data-points which may cause over-fitting.

```
SVM Tuned Performance:
----------------------
Accuracy         :  0.8891292590589508
F1 Score         :  0.6003898635477583
Precision        :  0.751219512195122
Recall           :  0.5
Confusion Matrix:
  [[1490   51]
 [ 154  154]]
```

Fig. 25 Performance metric-tuned SVM classifier

By tuning the above explained parameters, there is an increase can be seen in the performance of SVM i.e., accuracy and f1-score values are significantly improved as shown in fig. 25.

4.3.4. LR Analysis

LR technique is a kind of statistical-analysis that makes estimation about the association between the parameters such as dependent and independent variables by implementing logistics function [48].

```
Logistic Regression initial Performance:
----------------------------------------
Accuracy         :  0.8777717685235262
F1 Score         :  0.5232067510548523
Precision        :  0.7469879518072289
Recall           :  0.4025974025974026
Confusion Matrix:
  [[1499   42]
 [ 184  124]]
```

Fig. 26 Performance metric- LR

The simple LR classifier resulted into less f1-score value and accuracy in comparison with the SVM classifier. The next version of LR was implemented with hyper tuning of parameters.

LR Tuning

The LR classifier is implemented with tuning of hyper-parameters which resulted into the following metrics.

```
Logistic Regression Tuned Performance:
--------------------------------------
Accuracy         :  0.879394267171444
F1 Score         :  0.5285412262156448
Precision        :  0.7575757575757576
Recall           :  0.40584415584415584
Confusion Matrix:
  [[1501   40]
 [ 183  125]]
```

Fig. 27 Performance metric- tuned LR

The above metrics clearly states that with the use of hyper parameter tuning, there is a slight improvement in the value of f1-score and accuracy of LR classifier but the performance of SVM is better still.

4.3.5. RF Classification

The RF classifier works on a group of different DT classifiers. It considers the various samples of DT's and calculates the average value of the sub samples for increasing the prediction accuracy and controlling the over fitting [49]. The size of the subsamples is handled with the attribute "max-samples" otherwise all the samples of data-set would be used to create every tree.

Volume: 13 Issue: 06 Pages: 5169-5182(2022) ISSN: 0975-0290

```
Random Forest initial Performance:
----------------------------------
Accuracy        :  0.8945375878853434
F1 Score        :  0.6228239845261122
Precision       :  0.7703349282296651
Recall          :  0.5227272727272727
Confusion Matrix:
  [[1493   48]
 [ 147  161]]
```

Fig. 28 Performance metric- RF classifier

The above fig. 27 represents the performance of RF algorithm. The RF classifier with default values of parameter offers higher f1-score and accuracy than all the other algorithms implemented in this work. Next, The RF classifier is implemented with tuning of hyper-parameters.

RF Tuning

In RF tuning, various parameters were used for tuning purpose such as n-estimator, max-features, max-depth, min-sample-split and min-sample-leaf. N-estimator represents the number of trees in RF. Max-feature represents the maximum number of features that are used to split a node. Max-depth is used to define the maximum number of levels in each DT. Min-sample-split denotes the number of data-points in a tree node before it is splitted and min-sample-leaf defines the minimum number of data-points that can be allotted to a leaf-node of tree [50].

```
Random Forest Tuned Performance:
--------------------------------
Accuracy        :  0.9015684153596538
F1 Score        :  0.6617100371747212
Precision       :  0.7739130434782608
Recall          :  0.577922077922078
Confusion Matrix:
  [[1489   52]
 [ 130  178]]
```

Fig. 29 Performance metric- tuned RF classifier

The tuned version of RF algorithm has presented the best values of precision, accuracy, f1-score and recall till now.

4.3.6. GB Classifier

GB is ML technique used for regression as well as classification. It is a type of binary classifier. It is a robust and effective technique to generate prediction models. It generates a predictive model as a collaboration of various weak models [51].The key advantage is that it is capable of handling mixed kind of data naturally.

```
Gradient Boost initial Performance:
-----------------------------------
Accuracy        :  0.9053542455381287
F1 Score        :  0.6891651865008881
Precision       :  0.7607843137254902
Recall          :  0.6298701298701299
Confusion Matrix:
  [[1480   61]
 [ 114  194]]
```

Fig. 30 Performance metric- GB classifier

The simple GB classifier performs better than the RF classifier as shown in fig. 30. The next version of GB classifier was generated by applying hyper-parameter tuning.

GB Tuning

In this, various numbers of hyper-parameters were tuned such as n-estimators, loss, rate of learning, sub-sample, max-feature, max-depth, min-sample-split and min-sample-leaf [52]. Because of the huge number of parameters for testing, the random search has been used for this.

```
Gradient Boost Tuned Performance:
---------------------------------
Accuracy        :  0.8950784207679827
F1 Score        :  0.6407407407407407
Precision       :  0.7456896551724138
Recall          :  0.5616883116883117
Confusion Matrix:
  [[1482   59]
 [ 135  173]]
```

Fig. 31 Performance metric- tuned GB classifier

By examining the above metrics in fig. 31, it can be stated that the default GB classifier is slightly efficient than the tuned one.

4.3.7. AdaBoost Algorithm

Ada-Boost algorithm is a type of ensemble method used as boosting process. In this, weights for every instance are re-assigned which are classified incorrectly for finding the best model [53].

```
AdaBoost initial Performance:
-----------------------------
Accuracy        :  0.8848025959978366
F1 Score        :  0.6148282097649187
Precision       :  0.6938775510204082
Recall          :  0.551948051948052
Confusion Matrix:
  [[1466   75]
 [ 138  170]]
```

Fig. 32 Performance metric- Ada-Boost

Volume: 13 Issue: 06 Pages: 5169-5182(2022) ISSN: 0975-0290

It can be noted from the above metrics that the performance of Ada-Boost is lower than the GB classifier. Tuned version is also tried for Ada-Boost by adjusting the parameters such as number of estimators and learning rate.

```
AdaBoost Tuned Performance:
---------------------------
Accuracy         :  0.8885884261763115
F1 Score         :  0.624087591240876
Precision        :  0.7125
Recall           :  0.5551948051948052
Confusion Matrix:
  [[1472   69]
 [ 137  171]]
```

Fig. 33 Performance metric- tuned Ada-Boost

The tuned version of Ada-Boost has a performance improvement in comparison to the previous version, especially in terms of f1-score and precision as per fig. 33.Still, it is not higher than the GB classifier.

5. COMPARATIVE ANALYSIS OF MODELS

Seven different classification algorithms have been studied and implemented in this work. The performance evaluation and comparison of all the algorithms is as follows.

Classifier	Accuracy	F1-Score	Precision	Recall
NB	0.775	0.491	0.394	0.652
KNN	0.873	0.506	0.723	0.39
SVM	0.889	0.6	0.751	0.5
LR	0.879	0.529	0.758	0.406
RF	0.902	0.662	0.774	0.578
GB	0.905	0.689	0.761	0.63
AdaBoost	0.889	0.624	0.713	0.555

Fig. 34Performance comparison of different classifiers

ROC stands for receiver operating characteristics curve. It is a significant metrics for evaluating the performance of binary classification techniques. It is basically a probability plot which tells how much a prediction model is accurate in distinguishing between the classes [54].An ROC curve is plotted for visualizing the performances of all the seven classifiers for selecting best among them for generating the prediction model.

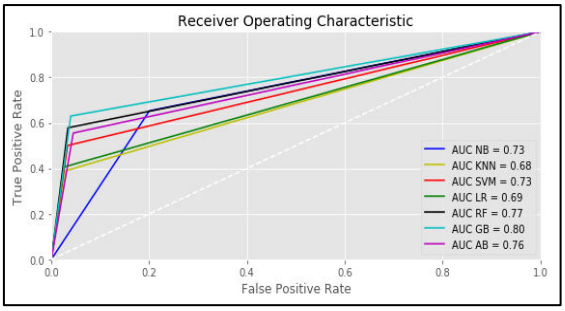

Fig. 35 ROC curve for performance evaluation

It can be clearly stated from the above ROC curve in fig. 35 that among all the classification algorithms, the GB classifier is having highest performance in terms of all the measures such as accuracy, precision and f1-score. Thus, it is the algorithm which will be chosen for the further work.

6. CONCLUSION

In this era of internet, e-business and online shopping, it is a key requirement to make predictions about the users' shopping behaviour and the other collected information for better execution of these kind of businesses and customer retention. The main aim of the work done is to propose such a framework which can present insights generated out of the analysis of data and to implement and compare various ML classification techniques to choose the best technique among all for developing a model of best performance. In this direction, seven distinct algorithms have been selected and implemented for making a comparison. For enhancing the performance of classifiers, the hyper-parameter tuning was used.

Thus, by analyzing the performance metrics of all the classifiers, it has been concluded that the best performing algorithm is GB classifier followed by RF, SVM and others. The best performing algorithm have accuracy around 91% and precision around 76%. So, the GB classifier is an efficient classifier for predicting that a customer will make a purchase or not. This prediction of shopping behaviours of customers is useful for emphasizing the interested customers for making them generate a revenue in future.

There is much scope of research in future also in this particular field. As, it can be seen that the considered data-set is skewed and unbalanced in nature. Thus, different methods may be used in future for handling better and reducing the impact of the un-balanced data for achieving a much better performance.

REFERENCES

[1] Goularas D and Kamis S, (2019) "Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data," 2019 International Conference on Deep Learning and Machine Learning in Emerging

Volume: 13 Issue: 06 Pages: 5169-5182(2022) ISSN: 0975-0290

Applications (Deep-ML), 2019, pp. 12-17, doi: 10.1109/Deep-ML.2019.00011.

[2] Ståhl, N., Falkman, G., Karlsson, A., Mathiason, G., & Boström, J. (2019). Deep reinforcement learning for Multiparameter optimization in de novo drug design. doi:10.26434/chemrxiv.7990910.v2

[3] Li Y.F, Guo L.Z and Zhou Z.H, (2021) "Towards Safe Weakly Supervised Learning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 334-346, 1 Jan. 2021, doi: 10.1109/TPAMI.2019.2922396.

[4] Parihar V., Yadav S. (2021), "Comparison Estimation of Effective Consumer Future Preferences with the Application of AI", in Vivekananda Journal of Research October, 2021, Vol. 10, Special Issue, Pg No. 133-145 ISSN 2319-8702(Print) ISSN 2456-7574(Online)

[5] WANG N. (2021), "Research on the influence of the cross-border e-commerce development of small and medium-sized enterprises in Dongguan in the post-epidemic era," 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT), 2021, pp. 176-180, doi: 10.1109/ECIT52743.2021.00047.

[6] Madiudia I., Porplytsya N., Zelenetska K. and Shevchuk R. (2021), "Modeling Dynamics of Traffic for the E-Commerce Website in the Process of its Search Engine Optimization During the COVID-19 Pandemic," 2021 11th International Conference on Advanced Computer Information Technologies (ACIT), 2021, pp. 61-64, doi: 10.1109/ACIT52158.2021.9548565.

[7] Surjandy, Cassandra C., Meyliana, Eni Y., Marcela Y. and Clarissa S. (2021), "Analysis of Product Trust, Product Rating and Seller Trust in e-Commerce on Purchase Intention during the COVID-19 Pandemic," 2021 International Conference on Information Management and Technology (ICIMTech), 2021, pp. 522-525, doi: 10.1109/ICIMTech53080.2021.9534964.

[8] Sahu P. K., & Gupta R. (2021). Frequent sequential traversal pattern mining for next web page prediction. International Journal of Advanced Networking and Applications, 13(03), 4983-4987. https://doi.org/10.35444/ijana.2021.13306

[9] Kim R. Y. (2020), "The Impact of COVID-19 on Consumers: Preparing for Digital Sales," in IEEE Engineering Management Review, vol. 48, no. 3, pp. 212-218, 1 thirdquarter,Sept. 2020, doi: 10.1109/EMR.2020.2990115.

[10] Wang C. -N., Nguyen N. -A. -T., Dang T. -T. and Hsu H. -P. (2021), "Evaluating Sustainable Last-Mile Delivery (LMD) in B2C E-Commerce Using Two-Stage Fuzzy MCDM Approach: A Case Study From Vietnam," in IEEE Access, vol. 9, pp. 146050-146067, 2021, doi: 10.1109/ACCESS.2021.3121607.

[11] Shehata N. S., Nasr M., Fangary L. E., & EL hamid L. A. (2021). Algorithms of deep Learning: Convolutional neural network role with colon cancer disease. International Journal of Advanced Networking and Applications, 13(01), 4827-4832. https://doi.org/10.35444/ijana.2021.13104

[12] Tufail H., Ashraf M. U., Alsubhi K. and Aljahdali H. M. (2022), "The Effect of Fake Reviews on e-Commerce During and After Covid-19 Pandemic: SKL-Based Fake Reviews Detection," in IEEE Access, vol. 10, pp. 25555-25564, 2022, doi: 10.1109/ACCESS.2022.3152806.

[13] Li H. and Peng T. (2020), "How Does Heterogeneous Consumer Behavior Affect Pricing Strategies of Retailers?," in IEEE Access, vol. 8, pp. 165018-165033, 2020, doi: 10.1109/ACCESS.2020.3022491.

[14] Quan J., Wang X. and Quan Y. (2019), "Effects of Consumers' Strategic Behavior and Psychological Satisfaction on the Retailer's Pricing and Inventory Decisions," in IEEE Access, vol. 7, pp. 178779-178787, 2019, doi: 10.1109/ACCESS.2019.2958685.

[15] Yang Z., Xiong G., Cao Z., Li Y. and Huang L. (2019), "A Decision Method for Online Purchases Considering Dynamic Information Preference Based on Sentiment Orientation Classification and Discrete DIFWA Operators," in IEEE Access, vol. 7, pp. 77008-77026, 2019, doi: 10.1109/ACCESS.2019.2921403.

[16] Chen H., Yan Q., Xie M., Zhang D. and Chen Y. (2019), "The Sequence Effect of Supplementary Online Comments in Book Sales," in IEEE Access, vol. 7, pp. 155650-155658, 2019, doi: 10.1109/ACCESS.2019.2948190.

[17] Lim J., Grover V. and Purvis R. L. (2012), "The Consumer Choice of E-Channels as a Purchasing Avenue: An Empirical Investigation of the Communicative Aspects of Information Quality," in IEEE Transactions on Engineering Management, vol. 59, no. 3, pp. 348-363, Aug. 2012, doi: 10.1109/TEM.2011.2164802.

[18] Wang X., Fan Z. -P. and Liu H. (2020), "How Can Sellers React to Consumers' Anticipated Regret in an Online Markdown Policy?," in IEEE Access, vol. 8, pp. 224911-224921, 2020, doi: 10.1109/ACCESS.2020.3041002.

[19] Alsheikh S. S., Shaalan K. and Meziane F. (2019), "Exploring the Effects of Consumers' Trust: A Predictive Model for Satisfying Buyers' Expectations Based on Sellers' Behavior in the Marketplace," in IEEE Access, vol. 7, pp. 73357-73372, 2019, doi: 10.1109/ACCESS.2019.2917999.

[20] Jeong J. (2021), "Identifying Consumer Preferences From User-Generated Content on Amazon.Com by Leveraging Machine Learning," in

IEEE Access, vol. 9, pp. 147357-147396, 2021, doi: 10.1109/ACCESS.2021.3123301.

[21] Zhao H. -H., Luo X. -C., Ma R. and Lu X. (2021), "An Extended Regularized K-Means Clustering Approach for High-Dimensional Customer Segmentation With Correlated Variables," in IEEE Access, vol. 9, pp. 48405-48412, 2021, doi: 10.1109/ACCESS.2021.3067499.

[22] Imaniya T. and Agus A. A. (2019), "The Influence of Consumer's Decision Making Process to Online Purchase Behavior Analysis," 2019 2nd International Conference of Computer and Informatics Engineering (IC2IE), 2019, pp. 116-121, doi: 10.1109/IC2IE47452.2019.8940898.

[23] Goyal N., Singh Y., Shukla R. and Srivastava S. (2021), "A Study on Online Shopping Behaviour for Apparel," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2021, pp. 1986-1991, doi: 10.1109/ICAC3N53548.2021.9725504.

[24] Rajvanshi A. (2021), "Influence of Demographics on Female Consumer's Attitude, Intention and Shopping Behaviour Towards Online Apparel Purchase in Delhi & NCR," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, pp. 1-7, doi: 10.1109/ICRITO51393.2021.9596548.

[25] Chaitanya C. and Gupta D. (2017), "Factors influencing customer satisfaction with usage of shopping apps in India," 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017, pp. 1483-1486, doi: 10.1109/RTEICT.2017.8256844.

[26] Zhao T., Hu M., Rahimi R. and King I. (2017), "It's about time! Modeling customer behaviors as the secretary problem in daily deal websites," 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 3670-3679, doi: 10.1109/IJCNN.2017.7966318.

[27] Singh S. P., Kumar A., Yadav N. and Awasthi R. (2018), "Data Mining: Consumer Behavior Analysis," 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2018, pp. 1917-1921, doi: 10.1109/RTEICT42901.2018.9012300.

[28] Morales-Rodríguez, F. M., Martínez-Ramón, J. P., Méndez, I., & Ruiz-Esteban, C. (2021). Stress, coping, and resilience before and after COVID-19: A predictive model based on artificial intelligence in the University environment. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.647964

[29] Seckler, M., Heinz, S., Forde, S., Tuch, A. N., & Opwis, K. (2015). Trust and distrust on the web: User experiences and website characteristics. *Computers in Human Behavior*, *45*, 39-50. https://doi.org/10.1016/j.chb.2014.11.064

[30] Zeng, M., Cao, H., Chen, M., & Li, Y. (2018). User behaviour modeling, recommendations, and purchase prediction during shopping festivals. Electronic Markets, 29(2), 263-274. https://doi.org/10.1007/s12525-018-0311-8

[31] Mokryn, O., Bogina, V., Kuflik, T. (2019). Will this session end with a purchase? Inferring current purchase intent of anonymous visitors. Electronic Commerce Research and Applications, 34, 100 836. https://doi.org/10.1016/j.elerap.2019.100836

[32] Wu, Z., Tan, B.H., Duan, R., Liu, Y., Mong Goh, R.S. (2015).Neural modeling of buying behaviour for E-commerce from clicking patterns. In Proceedings of the international ACM recommender systems challenge 2015 (p. 12). https://doi.org/10.1145/2813448.2813521.

[33] Wang, B.; Ye, F.; Xu, J. A Personalized Recommendation Algorithm Based on the User's Implicit Feedback in E-Commerce. Future Internet 2018, 10, 117.

[34] Ferraro, A.; Bogdanov, D.; Choi, K.; Serra, X. Using offline metrics and user behavior analysis to combine multiple systems for music recommendation. In Proceedings of the RecSys'18, REVEAL Workshop, Vancouver, BC, Canada, 2–7 October 2018; pp. 1–6.

[35] Reis, J., Amorim, M., Melão, N., & Matos, P. (2018). Digital transformation: A literature review and guidelines for future research. Advances in Intelligent Systems and Computing, 411-421. https://doi.org/10.1007/978-3-319-77703-0_41

[36] Çelik, Ö., & Aslan, A. F. (2019). Gender prediction from social media comments with artificial intelligence. *Sakarya University Journal of Science*, 1256-1264. https://doi.org/10.16984/saufenbilder.559452

[37] Wang, K.; Zhang, T.; Xue, T.; Lu, Y.; Na, S.G. E-Commerce Personalized Recommendation Analysis by Deeply-learned Clustering. J. Vis. Commun. Image Represent. 2020, 71, 1–7. [CrossRef]

[38] Kim, D.H., Lee, S., Jeon, J., Song, B.C. (2020). Real-time purchase behavior recognition system based on deep learning based object detection and tracking for an unmanned product cabinet. Expert Systems with Applications, 143, 113063. https://doi.org/10.1016/j.eswa.2019.113063.

[39] Esmeli, R., Bader-El-Den, M., & Abdullahi, H. (2020). Towards early purchase intention prediction in online session based retailing systems. Electronic Markets, 31(3), 697-715. https://doi.org/10.1007/s12525-020-00448-x

Volume: 13 Issue: 06 Pages: 5169-5182(2022) ISSN: 0975-0290

[40] N. T. HA, T. L. H. NGUYEN, T. V. PHAM, and T. H. T. NGUYEN (2021), "Factors Influencing Online Shopping Intention: An Empirical Study in Vietnam," The Journal of Asian Finance, Economics and Business, vol. 8, no. 3, pp. 1257–1266, Mar. 2021.

[41] UCI machine learning repository: Online shoppers purchasing intention dataset data set. (n.d.). https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

[42] Nasr M., Karam A., Atef M., Boles K., Samir K., & Raouf M. (2020). Natural language processing: Text categorization and classifications. International Journal of Advanced Networking and Applications, 12(02), 4542-4548. https://doi.org/10.35444/ijana.2020.12201

[43] Wu C. et al. (2019), "Exploratory Analysis for Big Social Data Using Deep Network," in IEEE Access, vol. 7, pp. 21446-21453, 2019, doi: 10.1109/ACCESS.2019.2898238.

[44] Setyaningsih E. R and Listiowarni I, (2021) "Categorization of Exam Questions based on Bloom Taxonomy using Naïve Bayes and Laplace Smoothing," 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT), 2021, pp. 330-333, doi: 10.1109/EIConCIT50028.2021.9431862.

[45] Sapkota N, Alsadoon A, Prasad P.W.C, Elchouemi A and Singh A.K, (2019) "Data Summarization Using Clustering and Classification: Spectral Clustering Combined with k-Means Using NFPH," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 146-151, doi: 10.1109/COMITCon.2019.8862218.

[46] Li Y.F, Guo L.Z and Zhou Z.H, (2021) "Towards Safe Weakly Supervised Learning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 334-346, 1 Jan. 2021, doi: 10.1109/TPAMI.2019.2922396.

[47] Rojas-Domínguez A., Padierna L. C., Carpio Valadez J. M., Puga-Soberanes H. J. and Fraire H. J. (2018), "Optimal Hyper-Parameter Tuning of SVM Classifiers With Application to Medical Diagnosis," in IEEE Access, vol. 6, pp. 7164-7176, 2018, doi: 10.1109/ACCESS.2017.2779794.

[48] Wang D, Gao Y and Tian Z, (2017) "One-Variable Linear Regression Mathematical Model of Color Reading and Material Concentration Identification," 2017 International Conference on Smart City and Systems Engineering (ICSCSE), 2017, pp. 119-122, doi: 10.1109/ICSCSE.2017.37.

[49] Yang F, (2019) "An Extended Idea about Decision Trees," 2019 International Conference on Computational Science and Computational Intelligence (CSCI), 2019, pp. 349-354, doi: 10.1109/CSCI49370.2019.00068.

[50] Kelkar K. M. and Bakal J. W. (2020), "Hyper Parameter Tuning of Random Forest Algorithm for Affective Learning System," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020, pp. 1192-1195, doi: 10.1109/ICSSIT48917.2020.9214213.

[51] Dutta J., Kim Y. W. and Dominic D. (2020), "Comparison of Gradient Boosting and Extreme Boosting Ensemble Methods for Webpage Classification," 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2020, pp. 77-82, doi: 10.1109/ICRCICN50933.2020.9296176.

[52] Thosar K., Tiwari P., Jyothula R. and Ambawade D. (2021), "Effective Malware Detection using Gradient Boosting and Convolutional Neural Network," 2021 IEEE Bombay Section Signature Conference (IBSSC), 2021, pp. 1-4, doi: 10.1109/IBSSC53889.2021.9673266.

[53] Yong Z., Jianyang L., Hui L. and Xuehui G. (2018), "Fatigue driving detection with modified ada-boost and fuzzy algorithm," 2018 Chinese Control And Decision Conference (CCDC), 2018, pp. 5971-5974, doi: 10.1109/CCDC.2018.8408177.

[54] Hassan A., & B S L. (2022). A study of emerging image processing and machine learning methodologies for classification of plant leaf disease. International Journal of Advanced Networking and Applications, 13(04), 5057-5062. https://doi.org/10.35444/ijana.2022.13407

Biography and Photograph

Veena Parihar, **Research Scholar** at Career Point University, Kota, Rajasthan. Enthusiast about teaching, learning and acquiring new knowledge. Having almost 5 years of experience of teaching and mentoring engineering graduates. Pursued B.tech. and M.Tech. from computer science and engineering stream. In depth knowledge of various programming languages and web frameworks. Thoroughly worked with python and java programming. Areas of interest are artificial intelligence, machine learning, personality identification and classification, data science and data security.

Dr. Surendra Yadav, Professor at Vivekananda Global University, Jaipur, Rajasthan. Having almost 15+ years of experience of teaching and guiding engineering students and PhD scholars. Pursued B.tech. and M.Tech. from computer science and engineering stream. Areas of interest are computer networks, machine learning, IoT and data mining.