International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 12 (2020), pp. 5105-5112
© International Research Publication House. http://www.irphouse.com

Investigating Factors that Affect Purchase Intention of Visitors of E-commerce Websites Using a High Scoring Random Forest Algorithm

Martha Teiko Teye[1], **Yaw Marfo Missah**[2]

[1] Graduate Student, College of Science, Kwame Nkrumah University of Science and Technology, Accra Rd, Kumasi, Ghana.

[2] PhD., Lecturer, College of Science, Kwame Nkrumah University of Science and Technology, Accra Rd, Kumasi, Ghana.

ORCIDs: 0000-0002-2370-4700 (Martha Teiko Teye), 0000-0002-2926-681X (Yaw Marfo Missah)

Abstract

Making sales to generate revenue is the prime aim of any business, including e-commerce. Research shows that one major barrier hindering the growth and expansion of e-commerce is the inability to effectively determine shoppers from browsers at e-commerce websites. This is attributed to the varying patterns in the shopping process of shoppers. Although there are predictive models based on e.g. naïve Bayes, multilayer perceptron and decision trees for predicting shopper purchase intention, predictive models based on large datasets with diverse features are better, as found in the study reported in this article. This research combines the classical random forest algorithm with hyperparameter tuning and the adaptive synthetic oversampling technique to predict the purchase intentions of shoppers at e-commerce websites whiles highlighting the importance of location in purchase decisions. The study uses a large dataset from the **University of California Irvine** Machine Learning Repository to build an effective machine learning model. The accuracy of the improved model is competitive to those of decision tree, support vector machine, multi-layer perceptron and classical random forest.

Keywords: E-commerce, purchase intention, random forest, consumer.

I. INTRODUCTION

How extreme would users go in deciding between buying from an e-commerce company (aka online shopping) or a physical shop? Recently, technology has been a part of most aspects of life and businesses are not an exception. The main aims and motivations of e-commerce businesses are to harness the advancement in technology, especially the Internet, to reduce expenses and make shopping more comfortable for consumers to thrive the business [1]. E-commerce businesses are faced with the problem of not being able to reliably predict the intention of visitors to their websites. He, Lu and Zhou [2] found that one major barrier hindering the growth of e-commerce is the inability to effectively determine the intention to purchase.

Unlike in the traditional brick and mortar setting, where shop attendants can have physical interactions to know reasons why visitors did not purchase particular goods, the e-commerce presence leaves businesses no idea why transactions never took place. According to [2,3], the nature of e-commerce websites with regards to navigation, content and simplicity of use determines how useful these websites appear to customers. The rate of usefulness plays a vital role in convincing customers in deciding whether or not to proceed with a transaction. Owing to different marketing strategies that have been adopted by the e-commerce industry, advertisements reach thousands of customers and hence websites are able to drive a lot of traffic. On the marketing point of view, this initiative is successful since numerous customers are attracted [4]. However, from the business perspective, unless the website visitors make actual purchases or transactions, their goal is not yet achieved.

The introduction of data science techniques (including machine learning) have been embraced in most decision making processes. Several machine learning (ML) models have been developed to forecast the shopping intention of consumers within the e-commerce space. The ML algorithms used in these models include support vector machines (SVM) [5], random forest and decision tree [6]. However, it is realized that most of the dataset used is biased towards false revenue which was a reason for accuracy levels falling between 82% and 89% for all the models used [7].

This paper uses a large dataset from University of California Irvin (UCI) machine learning online repository [8], preprocesses the data using one-hot encoding [9], builds an effective ML model using feature importance technique in the sklearn's random forest classifier to select the top 15 features. Final training is performed and compares the accuracy of the random forest model with that of decision tree, SVM, multi-layer perceptron and other ensemble algorithms as done in [5,6,7,10].

The primary objective of this research is to develop a high scoring random forest algorithm for e-commerce businesses to predict shoppers' purchase intention. The objective is achieved through the three related sub-objectives:

1. Identify existing weaknesses in the random forest algorithm for predicting purchase intention [7,10].
2. Train a random forest model using same data as in Objective 1 to provide better decisions and compare the results with that of random forest algorithm, SVMs, multi-layer perceptron and naïve bayes algorithms.
3. Identify basic factors that influence actual purchase intention based on the model.

II. LITERATURE REVIEW

Online shopping intention became a crucial concern to e-commerce business owners in 2000 when the Internet became a major infrastructure. There are a number of proposals addressing how e-commerce businesses can track customers' logs and use them to reliably predict the likelihood of purchase. Different studies have introduced diverse ideologies and perspectives in techniques for predicting decisions of shoppers based on patterns from shopping behaviour.

II.I Features of E-commerce that Affect Purchase Decision

The theory of buyer behaviour [11] shows a repetitive nature of shopping pattern taking into consideration the exogenous variables through the traditional brick and mortar mode of selling. Although behavioural patterns might be similar in both traditional and e-commerce businesses, other factors which include trust and security relating to the Internet also affect the rate of purchase in e-commerce businesses [12]. The theory concluded that factors such as the price of items, unavailability of needed products, pressure, financial capabilities of customers and societal traits are the major inhibitory behaviors of online shoppers which could affect their decisions in the presence of e-commerce-based challenges.

First, focus is placed on the physical design features of e-commerce websites and their interactivity with consumers, an aspect of human-computer interaction. This design features covers usability, information quality, platform quality, service quality and playfulness [4]. Whether customers visit a website just to "window shop" or to look for something specific, their stay must be made simple and effective [13]. It could be either suggesting similar items, use of corrective searches other than an error message due to e.g. wrongful spellings or by categorizing products. Owing to this, it is not prudent to use primitive data to predict current purchase trends but could be a way of providing comparative analysis between both mediums. In [14], Gupta applied neural networks to predict the likelihood of purchase by a visitor taking into consideration the prices. Gupta's algorithm was improved by Kukar-Kinney and Angeline [15] by extending it to online shopping which was not focused on only pricing but also other factors such as the loading time, ease of use and design which is now of key concern to shoppers. Next, empirical studies on how the design features and business models can be intertwined to create an experience that would lead to purchase decisions is considered. How well do consumers trust and have confidence in doing business at these websites?

II.II Factors Influencing Consumer Purchase

Gupta [16] focused on the dynamic pricing of goods and services by enabling e-commerce supply vendors to provide optimum prices for goods based on the current competition. The model used logistic regression to predict the optimal price of goods and services with four clusters of eight main features which had a significance level between 0.01 and 0.05. Using K-means clustering algorithm, the coefficient of the variation was calculated as 83% that indicates a large area of the data being used. Pricing of items is a deciding factor in all business models. Higher online prices compared to the physical market could reduce the purchases from e-commerce sites [14]. These researches outlined the factors that affect the decision of purchase which later also highlighted benefit and risk perspective and behavioural pattern trend as the theory of reasoned action which could lead to the abortion of online transactions [16].

The mental and societal/environmental behaviors, such as impatience and frustration, with which people go through the shopping and check-out process of e-commerce sites were identified. These researches outlined the factors that affect the decision to purchase which later also highlighted benefit and risk perspective and behavioral pattern trend as the theory of reasoned action which could lead to the discontinue of online transactions [16].

A theoretical approach to investigate the theory of buyer behavior with respect to online shopping was conducted to compare the relationship (similarities and differences) in characteristics of the traditional way of shopping and e-commerce. Individuals that visited e-commerce websites with a positive attitude were more likely to make purchases in the end. However, this is not a standalone factor; the positive attitude could be gotten from the user's experience at the website. It implies that the web revolution used as a measure of sentiment analysis to provide customization and personalization in the look of e-commerce sites improves the needs of customers [17]. The five indicators that influence the purchase intention of e-commerce businesses are simplicity in its usage, the usefulness of the e-commerce website, vendor competence, recommendations by friends and other third-party services and perceived vendor attitudes [2].

Aside these analytical and mathematical researches done, neural networks and machine learning have begun making predictions based on similar data and factors. The decision tree classifier was demonstrated to be "an effective data mining technique that can provide accurate prediction and determine the marketing effectiveness in most businesses of today" [18]. This book carried out an investigation to identify the best marketing activities business should take into consideration for effective marketing plans. The loss functions calculated were the basis to confirm the accuracy of the decision tree classifier.

II.III Modern ML Models

Sakar et al. [7] identified under sampling of the dataset as an effect on previous models used. Oversampling and under sampling are very common in most machine learning algorithms and this leads to higher accuracies but with high rates of false positives and true negatives, respectively. Therefore, to deal with oversampling of the dataset, the paper introduced more negative decision data to the original dataset. Also, due to a large number of features in the dataset used, a feature selection method was used to reduce the original features captured in the data. This was tested on a SVM, decision tree and multi-layer perceptron classifiers. The output revealed that, although SVM and decision tree had fair accuracies, the multi-layer perceptron had a good balance of accuracy and F1-score than both the SVM and decision tree.

Access to a large dataset, which is not readily available, led to a predictive model on fewer data with different features for the prediction of purchasing behaviour of online consumers [6]. The paper suggested a single feature with high accuracy for determining the purchase intention of consumers. A model using the decision tree algorithm with a k-fold cross-validation was used to determine the train and test datasets after selecting a feature on open data using Fisher score. Accuracies were maintained at acceptable values of over 80% with a high dependency on the page value of a particular purchase session. It was found that advertising on multiple pages (e-commerce sites) increased purchases.

Christian [5] researched into two top-scoring comparison tools, WEKA and scikit-learn, to test the efficiency of popular ML algorithms which include SVM, gradient descent and naïve bayesian. The test compared values of F1-score, kappa statistics, absolute mean error and the accuracies of each ML algorithm. The F1-score is a blend of both the precision and recall values of the algorithm. Precision focuses on maximizing the number of true positive predictions out of all predictions that are positive [19]. Recall focuses on maximizing the number of true positive predictions out of all predicted results (i.e. true positives plus false negatives). Based on the investigation conducted, the random forest algorithm appeared to be the most suitable algorithm for classifying online shoppers' intentions. Algorithms that use information gain or entropy criteria were considered to provide optimal results in determining the purchase intention of e-commerce shoppers [7].

Baati and Mohsil [10] considered customers' decision to purchase in their initial visit to a website and introduced the synthetic minority oversampling technique (SMOTE) [20] as a way of dealing with the class imbalance. SMOTE has been identified as one of the best ways of oversampling data. However, after considering the weekend feature which is the top-scoring feature with the dataset, the research ignored the other top six features as a reason of wanting to determine the purchase intention right when a user visits a website. This approach could lead to misleading data and contribute to low prediction accuracies. Although the accuracy using the random forest algorithm was high (88.78), the sensitivity of the prediction remained at 0.62. This clearly showed that the model's ability to identify the true positives was not the best.

Looking at the above ML models, it could be identified that even though oversampling improved the performance of the models experimented on, the difference in the improvement was still low and accuracies were still not up to 90%. It was realized that after oversampling the dataset, the choice of features selected to train the random forest model also affected the results obtained. Most of the researches conducted focused on about 8-12 relevant features, which omits some other important features in a dataset that play a role in determining consumers' intent to purchase. Hence, using the ADASYN oversampling technique and paying much attention to features from the datasets used in previous researches, this paper provides a comparatively high predictive random forest model as described in Section 3.

III. MATERIALS AND METHODOLOGY

This paper reports a classifier model built with two random forest algorithms to predict the purchase intentions of consumers who visit e-commerce websites. The model is built using Python and its libraries numpy, pandas, sklearn and matplotlib. Data on online shoppers' activities derived from Google Analytics was downloaded from the UCI machine learning repository by Sakar et al. [8] for the model building. To facilitate the learning of the algorithm, we started with a large amount of data and then extracted important features that are relevant to predicting shoppers' intention. Then, the random forest algorithm was used to predict the purchase intention of consumers.

III.I Dataset

The UCI dataset contains 12,330 rows of data from users of a popular e-commerce site, Amazon. The dataset contains their visits to the Amazon website and the navigations made through the site to the decision at the end of the visit. The full dataset was randomly split into a 70/30 per cent train to test ratio respectively. The dataset contained 18 features which were gathered from each user visiting the page. An overview of the structure of the raw dataset obtained is shown in Table 1 with detailed explanation to the headers in Table 2. Out of the 18 initial features, feature selection using the random forest classifier (feature importance model) was performed to reduce the features to be used by the algorithm to 15. It was observed that the purchasing intention is imbalanced which is common for online consumers to not make a purchase of any product during most visits to their shopping sites.

III.II Data Preprocessing

Data preprocessing which was the most time-consuming aspect of the model building is also one of the important factors for achieving reliable results. Unprocessed data usually would result in misleading and inaccurate predictions or false results. Some columns in the dataset such as the month, visitor type and weekend, contained categorical data while the rest had numerical data. To create uniformity in the dataset, categorical features in the dataset as shown in Table 1, were converted to numerical data using hot encoding [22]. Also, since the data had some missing and null values, preprocessing was necessary to ensure the dataset was credible to achieve a reliable model. For such entries, the most occurring or frequently occurring values (modal) of the data was calculated to fill the null-valued columns. This method was used as an alternative to deleting all empty entries to avoid losing data especially the zero-labelled revenue rows. The decision variable in the dataset is made up of two values (i.e. binary classification); either made purchase (True) or did not make a purchase (False). The purchase/non - purchase rate is approximately 85% (10,422)/ 15% (1,908). To cater for the class imbalance, adaptive synthetic (ADASYN) [21] sampling method was used on the training data which contained 8631 rows with 7305 zero revenues and 1326 one revenues as shown in Figure 1 and Figure 2, respectively.

Table 1. Sample of Raw Dataset

Ad	Ad_D	If	If_D	PR	PR_D	BR	ER	PV	SD	M	OS	BT	Reg	TT	VT[1]	Wk[2]	Rev
0	0	0	0	1	0	0.2	0.2	0	0	Feb	1	1	1	1	RV	F	F
0	0	0	0	2	64	0	0.1	0	0	Feb	2	2	1	2	RV	F	F
0	-1	0	-1	1	-1	0.2	0.2	0	0	Feb	4	1	9	3	RV	F	F
0	0	0	0	8	126.25	0	0.05	0	0	Dec	2	2	3	2	NV	T	0
0	0	0	0	2	2.66667	0.05	0.14	0	0	Feb	3	2	2	4	RV	F	F
0	0	0	0	10	627.5	0.02	0.05	0	0	Feb	3	3	1	4	RV	T	F
0	0	0	0	19	154.217	0.015789	0.024561	0	0	Feb	2	2	1	3	RV	F	F
0	-1	0	-1	1	-1	0.2	0.2	0	0.4	Feb	2	4	3	3	RV	F	F
4	1005.27	0	0	35	655.6584	0	0.005128	0	0	Nov	2	10	7	2	NV	T	0

Table 2. Data Variables and their description

Variables	**Description**
Administrative (Ad), Informational (If), Product Related (PR)	Pages browsed by consumers. Determining which of these pages they navigated.
Administrative Duration (Ad_D), Informational Duration (In_D), Product Related Duration (PR_D)	Total period of time a consumer spends on a particular page.
Bounce Rate (BR)	Average bounce rate value of the pages browsed by the visitor
Exit Rate (ER)	Average exit rate value of the pages visited by the visitor
Page Value (PV)	Average page value of the pages visited by the visitor
Special Day (SD)	Specifies whether the date visited is a special occasion (holiday, discount day) or not.
Month (M)	Month of visit
Region (Reg)	Location of consumer
Browser Type (BT)	Specific browser the consumer used to access the site
Operating System (OS)	Operating system of the consumer’s device.
Traffic Type (TT)	The specific URL which redirected the customer to the e-commerce page. (direct URL, ads from other pages, etc.)
Visitor Type (VT)	Specifies whether the consumer is a new or returning customer.
Revenue (Rev)	Whether or not a user made a purchase represented by a Boolean.
Weekend (Wk)	A Boolean variable that tells whether the website was visited on a weekend or a weekday

[1] Values of VT (Vistor Type) are VT = Returning_Visitors and NV = New_Visitors

[2] Wk (Weekend) values are F = False and T = True

III.III Feature Selection

Some selected columns from the data are used to create the machine learning model since there are some features which are irrelevant to the model or do not have an impact on the revenue introduced in this research. The random forest classifier filter-based feature selection method which is focused on the feature importance other than the statistical coefficient was implemented. Sklearn's SelectKBest was used to find the top scoring features in the processed dataset. The initial dataset had 18 features (as explained in Table 2) which were increased to 57 after preprocessing using one-hot encoding to affect features such as the Operating System, Month, Browser, Region and VisitorType. Feature selection then reduced the number of columns from the 57 in the dataset to 15 relevant features. This included Informational, Informational_Duration, ProductRelated, ProductRelated_ Duration, Administrative, Administrative_Duration, Boune Rates, ExitRates, PageValues, SpecialDay, Month (December), TrafficType, VisitorType, Region, and Weekend. The final selected features were also based on the assumptions made in the Section 1 and the classification method used (i.e. random forest) in this research.

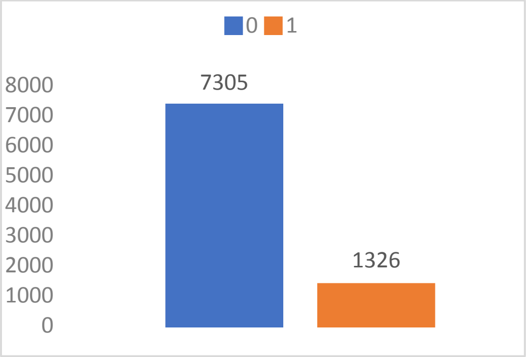


Figure 1. Revenue count of raw dataset

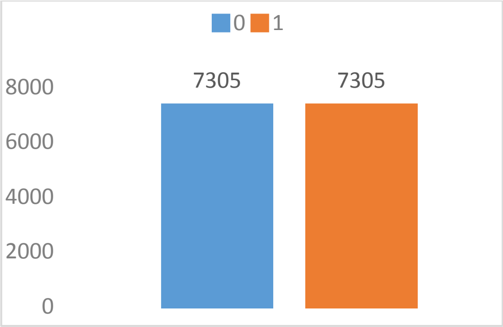


Figure 2. Revenue count after oversampling

III.IV Handling of Outliers

Outliers that are caused by experimental errors are filtered out of the dataset during data preprocessing prior to model building. Most of the outliers noticed in this data were due to either very high or very low values in some of the columns. A sample scatter plot using the Page Values as shown in Figures 3 and 4 was generated to visualize the data and identify some of the outliers. In order to ensure a reliable model, two approaches were used to handle any outliers in the dataset. The first approach identified and deleted all rows in the dataset containing outliers. In doing so, it was realized that the model gave an accuracy of 91.45%. Therefore, a second approach was used by calculating an anomaly score and then used this value as a totally new feature to the model. This helped in keeping data points which were outliers but resulted in a purchase.

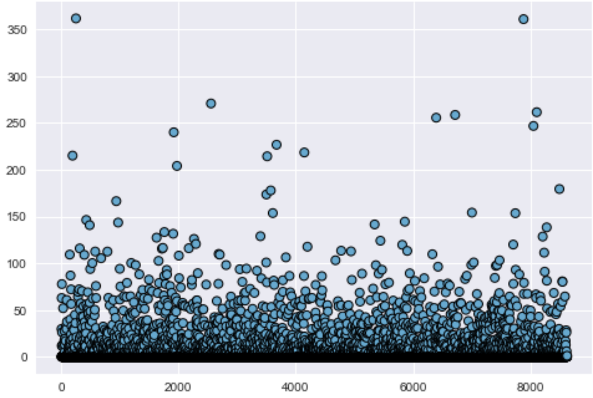


Figure 3. Unblanced train dataset

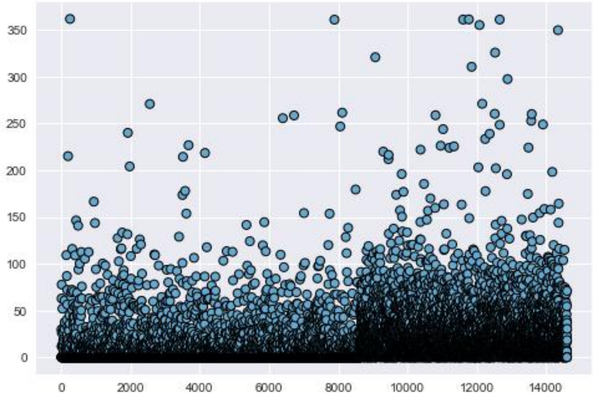


Figure 4. ADASYN oversampling- train dataset

III.V ML Model – Random Forest Classifier

Since the main aim of this study is to determine whether or not a purchase is made by an individual visiting an e-commerce website, a two-class classification algorithm is employed. This work used random forest classifier because it is a ML algorithm which have not been highly considered in previous works of predicting customer intention. Having a random forest classifier also has other desirable features, such as minimizing the risk of overfitted model as its decision is based on the decisions of multiple classifiers. Training time is less, it runs efficiently on large datasets, and estimates missing data values. The gini index criteria was used with this random forest classifier to calculate the node impurity within each tree in the decision-making process.

$$\text{Gini} = 1- \sum_{j} P_{sj}^{4} \qquad (1)$$

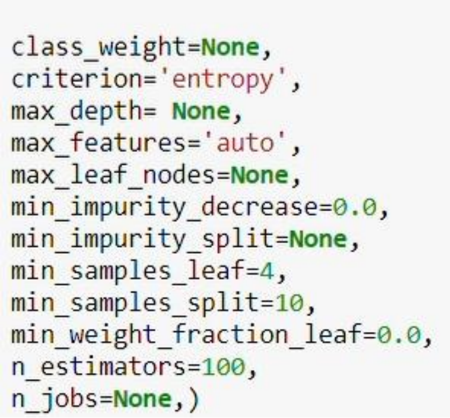
In building the machine learning model, optimization of the model through varying the parameters allows visualizing the classifier performance. By using a wide range of values, such as the maximum depth, minimum samples leaf, minimum samples split and the number of estimators the model performance also changes, either improves or worsens.

IV. RESULTS AND DISCUSSIONS

From the dataset, it was easily identified that a class imbalance towards non-purchase was present. This was so because, by observation, most people visit e-commerce sites to window shop and also check out the prices of goods and services. However, the factors that play a role in determining the purchase intention of customers go beyond the aesthetical design of the user interface, such as the circumstances surrounding their visits to these e-commerce sites. This section analyzes the results obtained to conclude on the factors that could influence the decision to purchase via e-commerce sites.

IV.I Hyper Parameter Tuning

To obtain optimal but effective accuracies, the random forest parameters were tweaked until an optimal result was obtained (Table 3). After a series of testing, the best values in terms of accuracy, true positive rate (TPR), true negative rate (TNR) and F1-score were obtained through the random forests parameter specification as shown in Figure 5. Initially, the criterion with "gini" (1) selection was considered but its accuracy values falls between 86% and 91% after several tuning of the classifier with both oversampled and original dataset. With the results from the gini criteria another random forest classifier using the entropy criteria was performed which boosted the accuracy of the model to 95.12%.



```
class_weight=None,
criterion='entropy',
max_depth= None,
max_features='auto',
max_leaf_nodes=None,
min_impurity_decrease=0.0,
min_impurity_split=None,
min_samples_leaf=4,
min_samples_split=10,
min_weight_fraction_leaf=0.0,
n_estimators=100,
n_jobs=None,)
```

Figure 5. Final parameter values for random forest classifier

IV.II Results

After calculating the class-wise accuracy, it can be inferred that the prediction was biased towards the 0 revenues. This amounts to a 10,422 positive decision and 1908 negative decisions. This meant that the model might be able to predict that a purchase would be made very easily than predicting users who might not make purchases. This led to oversampling of the training data using ADASYN [21] to cater for biases. Initially, oversampling alone was not enough to improve the accuracy. Therefore good feature selection was performed. Without oversampling, prediction using the random forests classifier resulted in an accuracy of 89.38% which comparatively was similar to results obtained in Table 3 and Table 4. Even with such an accuracy which can be considered good, the ability of the model to predict instances of non-purchasing was low as compared to the rate and precision it gave to predicting instances where purchases are made.

After oversampling to achieve an equal representation of both negative and positive class samples, the prediction on the training dataset had the accuracy of 95.12% while the prediction on the test data achieved the accuracy of 95.16%. This accuracy, precision and recall values are an improvement from the previous researches due to effective data preprocessing and tuning of model parameters. Table 3 shows an average value of the summary of results obtained from the random forest model using the training dataset compared to results by researches described in section 2.0 shown in Table 4 and Table 5.We realized that, although the decision tree algorithm used in both Tables 4 and 5 had relatively low accuracies compared to the support vector machine and multilayer perceptron models, the decision tree tended to produce high scoring predictive values in determining the purchasing intention of e-commerce consumers provided good features were present.

Table 3. Results from Random Forest Algorithm used in this paper

Classifier	Feature Splitting Criterion	Accuracy (%)	TPR	TNR	F1-Score
Random forest with oversampling	Entropy	95.12	0.83	0.89	0.79
Random forest without oversampling	Entropy	89.38	0.76	0.82	0.81

Table 4. Results obtained by Saker et al. [7]

Model	Classifier/ kernel	Accuracy (%)	TPR	TNR	F1-Score
Support Vector Machine	Linear	84.26	0.75	0.93	0.82
Decision Tree	Random Forest	82.29	0.74	0.90	0.81
Decision Tree	C4.5	82.34	0.79	0.85	0.82
Support Vector Machine	RBF	84.88	0.75	0.94	0.82
Multilayer Perceptron	10 hidden layers	87.94	0.84	0.92	0.86

Table 5. Results Obtained by Baati and Mohsil [8]

Model	Accuracy	TPR	TNR	F1-Score
Naïve Bayes	86.66	0.05	0.95	0.07
C4.5	86.59	0.55	0.92	0.56
Random Forest	88.78	0.62	0.91	0.60

Considering only the accuracy of the model can produce misleading results. Therefore, there was the need to consider the precision, recall and F1-score values of the model. From our results, a good balance of precision, recall and F1-score were also obtained and verified using the equations (2), (3) and (4) respectively.

Precision: $\frac{TP}{TP+FP}$ TP/TP+FP (2)

Recall: TP/(TP+FN)$\frac{TP}{TP+FN}$ (3)

F1: (2*Precision*Recall)/(Precision + Recall) (4)

IV.III Findings

The output from previous researches revealed that, although SVM and decision tree had fair accuracies, the multi-layer perceptron had a good balance of accuracy and F1-score as shown in Table 4. The high results were obtained from effective weight backtracking with backpropagation. However, all the ML models used had accuracies between 82% - 88%. None of these models had a score greater than 90% and this was largely attributed to the type of feature selection as well the balance in dataset used.

The initial claim of having a good feature selected data was confirmed from Table 3. Special day and region although considered to have less impact on the decision, helped improved the model. The different split data features which were tested using the “gini index” and “entropy” criteria values gave different results which were then combined to produce the final output. Although using one-hot encoding for preprocessing might not be necessary in most cases, the best practice of good feature selection and splitting aids in achieving high accuracies while reducing the probability of overfitting.

Table 6. Confusion Matrix

n = 3699	Predicted 0	Predicted 1	
Actual 0	3152 (TP)	118 (FP)	**3270**
Actual 1	129 (FN)	300 (TN)	**429**
	3281	**418**	

Validating the accuracy obtained by the random forest algorithm, a confusion matrix was generated from the 30% test dataset which comprises of 3699 rows as shown in Table 6. The model was able to predict 3152 zero revenues out of the 3270. Further calculations of the precision (2), recall (3) and F1 (4), after plotting the confusion matrix proved initial accuracy expressed as (5) was valid. Finally, it was observed that although pricing and design features play a role in determining the purchase decision of consumers, the nature of the site, special shopping days, and the time spent navigating various pages of the website, and how well the business is promoted also affects their decision to purchase.

V. CONCLUSION

In this paper, the random forest algorithm was applied on an online dataset of shoppers to perform the experiments with much emphasis on the data preprocessing. After analyzing the raw dataset and identified biases towards the outcomes of a positive decision, oversampling was used to create an equal distribution of data. The random forest algorithm’s parameters were tweaked until a suitable accuracy, precision and recall values were obtained.

The results provide a better prediction as compared to other existing models and classifiers used. With an accuracy of 90% and above for the random forest model, the findings show that data preprocessing and feature selection is of utmost importance in the process of building an ML model. Through this, it was observed that although geographic location and special days have not been in consideration of most researches, it plays a vital role in determining whether or not a purchase decision would be made. Different groups of people react to different interfaces in their own way of which e-commerce is not an exception. The nature of the dataset plays a vital role in determining how the model behaves and hence affecting the decisions of the model.

Most often than not, customers lookout for more interactive ways of stimulating engagement, that is, to interact with marketers and engage with peer communities. The crucial part of maintaining e-commerce is building an impressive presence. The web revolution used as a measure of sentiment analysis improves the needs of customers in addition to the factors described in Section 2.3. This emotive web would enable e-commerce businesses to interact and have personalized real-time experience with both employees and customers. This would help in creating a more personalized experience for customers.

Due to the imbalance within the revenue decisions, further improvements such as exploring other oversampling or under sampling techniques to the dataset could be made to cater for the imbalances. Further research could also take into consideration the optimization of the models to be able to determine an efficient output in terms of compilation and results.

REFERNCESS

[1] G.D. Gary, K. Munib, Z. Shaoming, The effects of e-commerce drivers on export marketing strategy, J. Int. Mark. 15 (2007) 30–57. https://doi.org/10.1509/jimk.15.2.30.

[2] D. He, Y. Lu, D. Zhou, Empirical Study of Consumers' Purchase Intentions in C2C Electronic Commerce, Tsinghua Sci. Technol. 13 (2008) 287–292. https://doi.org/10.1016/S1007-0214(08)70046-4.

[3] Y. Lee, K.A. Kozar, Investigating the Effect of Website Quality on E-business Success: An Analytic Hierachy Process (AHP) Approach, Decis. Support Syst. 42 (2006) 1383–1401. https://doi.org/0.1016/j.dss.2005.11.005.

[4] Z. Huang, M. Benyoucef, From e-commerce to social commerce: A close look at design features, Electron. Commer. Res. Appl. 12 (2013) 246–259. https://doi.org/10.1016/j.elerap.2012.12.003.

[5] Y. Christian, Comparison of Machine Learning Algorithms Using WEKA and Sci-Kit Learn in Classifying Online Shopper Intention, J. Informatics Telecommun. Eng. 3 (2019) 58. https://doi.org/10.31289/jite.v3i1.2599.

[6] İ. Topal, Estimation of Online Purchasing Intention Using Decision Tree, J. Manag. Econ. Res. 17 (2019) 269–280. https://doi.org/10.11611/yead.542249.

[7] C.O. Sakar, S.O. Polat, M. Katircioglu, Y. Kastro, Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks, Neural Comput. Appl. 31 (2019) 6893–6908. https://doi.org/10.1007/s00521-018-3523-0.

[8] C.O. Sakar, S.O. Polat, M. Katircioglu, Y. Kastro, Online Shoppers Purchase Intention Dataset, UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets (accessed on February 13, 2020).

[9] B. Jason, Why One-Hot Encode Data in Machine Learning, Machine Learning Mastery. https://machinelearningmastery/why-one-hot-encode-data-in-machine-learning (accessed on July 18, 2020)

[10] K. Baati, M. Mohsil, Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest, Artif. Intell. Appl. Innov. AIAI 2020. IFIP Adv. Inf. Commun. Technol. 583 (2020) 43–51. https://doi.org/10.1007/978-3-030-49161-1.

[11] G.H. Haines, J.A. Howard, J.N. Sheth, The Theory of Buyer Behavior., J. Am. Stat. Assoc. 65 (1970) 1406. https://doi.org/10.2307/2284311.

[12] R.C. Marchany, J.G. Tront, E-commerce security issues, in: Proc. 35th Annu. Hawaii Int. Conf. Syst. Sci., IEEE Comput. Soc, 2002: pp. 2500–2508. https://doi.org/10.1109/HICSS.2002.994190.

[13] M.G. Helander, H.M. Khalid, Modeling the customer in electronic commerce, Appl. Ergon. 31 (2000) 609–619. https://doi.org/10.1016/S0003-6870(00)00035-1.

[14] R. Gupta, C. Pathak, A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing, Procedia Comput. Sci. 36 (2014) 599–605. https://doi.org/10.1016/j.procs.2014.09.060.

[15] M. Kukar-Kinney, A.G. Close, The determinants of consumers' online shopping cart abandonment, J. Acad. Mark. Sci. 38 (2010) 240–250. https://doi.org/10.1007/s11747-009-0141-5.

[16] J. Cho, Likelihood to abort an online transaction: influences from cognitive evaluations, attitudes, and behavioral variables, Inf. Manag. 41 (2004) 827–838. https://doi.org/10.1016/j.im.2003.08.013.

[17] F. Almeida, J. D. Santos, J. A. Monteiro, E-Commerce Business Models in the Context of Web 3.0 Paradigm, Int. J. Adv. Inf. Technol. 3 (2013) 1–12. https://doi.org/10.5121/ijait.2013.3601.

[18] N. Gordini, V. Veglio, Customer Relationship Management and Data Mining, in: P. Hershey (Ed.), Mark. Consum. Behav., P. Vasant, IGI Global, 2015: pp. 789–828. https://doi.org/10.4018/978-1-4666-7357-1.ch036.

[19] D.L. Olson, D. Delen, Performance Evaluation for Predictive Modeling, in: Adv. Data Min. Tech., 1st ed., Springer, 2008: pp. 137–147. https://doi.org/10.1007/978-3-540-76917-0.

[20] N. V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, J. Artif. Intell. Res. 16 (2002) 321–357. https://doi.org/10.1002/eap.2043.

[21] H. He, Y. Bai, E.A. Garcia, S. Li, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, IEEE Int. Jt. Conf. Neural Networks (IEEE World Congr. Comput. Intell. (2008) 1322–1328.

[22] M.A. Dharmasiri, Preprocessing data for Predicting Online Shoppers Purchasing Intention, Anal. Vidhya. (2019). https://medium.com/analytics-vidhya/preprocessing-data-for-predicting-online-shoppers-purchasing-intention-ml-ba78186b7e85 (accessed February 27, 2020).

[23] L. Breiman, J. H Friedman, R. A Olshen, C. J Stone,. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. (1984). ISBN 978-0-412-04841-8.

[24] D. M. W. Powers, Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation, Mach. Learn. Technol., (2008).