National College of Ireland

Predictive Analysis of Online Shoppers Intention Using Extreme Gradient Boosting Algorithm

MSc in Data Analytics (MSCDA-B)
Domain Application of Predictive Analytics – CA2

Sai Chethan Singu
Student ID: x18181937

School of Computing
National College of Ireland

Lecturer: Vikas Sahni

National College of Ireland

Project Submission Sheet – 2020/2021

School of Computing

Student Name:	Sai Chethan Singu		
Student ID:	x18181937		
Programme:	MSCDA	**Year:**	2020
Module:	Domain Applications of Predictive Analytics		
Lecturer:	Vikas Sahni		
Submission Due Date:	23-08-2020		
Project Title:	Predictive Analysis of Online Shoppers Intention		
Word Count:	3058		

Ihereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Sai chethan singu

Date: 23-08-2020

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1.Please attach a completed copy of this sheet to each project (including multiple copies).
2.**You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predictive Analysis of Online Shoppers Intention Using Extreme Gradient Boosting Algorithm

Sai Chethan Singu
MSc. Data Analytics
(School Of Computing)
National College of Ireland
Dublin, Ireland
x18181937@student.ncirl.ie

***Abstract*— Growth in technology and changing lifestyles were important reasons for the growing interest in online shopping. Now it has become new dimension in business sector. With the increase in online shoppers, created a potential market across world for e-commerce. The accessibility, time saving, variety of products, return policies were the main factors behind the influence of online shopping by consumers. Internet shopping has gained lot of attention nowadays due to modernization. This new era of technology has completely replaced the traditional way of buying the goods. Understanding patterns and changes in the behavior of customers has become crucial factors for the e-commerce websites for enhancing customers experience and marketing. At the same time, demand has been increased for sellers to determine the behavioral changes and patterns of online shoppers. So that they can predict their customers intention for purchase. This emphasizes the importance of predicting the customers intention to yield the desired revenue. In order to study the aspects, that influence the intention of online purchaser's machine learning models have been used to build the prediction model. The results showed that model was performed better with Extreme Gradient boosting, which is one of the optimized Gradient boosting algorithms. XGBoost algorithm has been applied on the data which is pre-processed and achieved accuracy of 90.34%. This model was used to analyze the different factors which are responsible for generating revenue from the customers. It also determines the visitors having intention of purchasing but instead quit the website. This enhances the firms or companies to capture the attention of potential customers in the market.**

Keywords—Online Shopping, Modernization, Machine Learning, Gradient Boosting, Extreme Gradient Boosting.

I. INTRODUCTION

Internet usage has been increasing daily which eventually making things easier for everyone. It impacts customers lifestyle and their behaviour. With reduced search effort, great deal of prices, convenience were the main reasons for driving the attention of people. Huge number of business transactions have been made every year across the globe. Now many people have begun investigating the factors that are affecting online shoppers to do online shopping so that by improving marketing strategies and customer's experience they can raise sales and revenue in return. Some of the important concerns for companies is the conversion rate and maintaining customers in long run. Hence the trend of analysing online shopper's behaviour and changes has become one of the emerging research field in retail industry. By analyzing the past data using data mining and machine learning algorithms, companies can predict the intention of shoppers whether they will purchase or abandon the website from buying goods or services. Now companies are focusing on potential customers who are more likely to generate high revenue for the company. By machine learning algorithms, companies can segregate the customers into potential customers by taking many factors into consideration. Thus, paying attention to existing and potential customers in order to retain them in the long run. Other important thing for the companies to have is customer satisfaction as physical interaction will not take place between online shoppers and sellers.

Nowadays companies use different strategies to maintain customer satisfaction so that they can remain competitive in the market. A research work proposed the prediction model of online shoppers by using data like user and session information, page view info when customers visit website [1]. Based on that information model predicts whether the person purchase from that respective website or quit the website from purchasing. The factors that affects the online shoppers to do shopping were analysed by using this prediction model. Another recent study states that to background of their education, occupation, income and age were the main contributing factors that influences the purchaser's intention. Different age group people with different occupations showed wide range of interests when doing online shopping [2]. Other research work proposed

that purchase intention was mainly influenced by the perception of their family, friends and media. In this study it was proved that the intention itself was the notable predictor for shopping online. By using hypotheses testing, correlation between intention and behaviour of online shopping seems to have strong relationship. And also the influence of family and friends [3].

The analysis of online shoppers intention was done by taking important factors into consideration that includes session info, traffic, page value, exit rate and region. By using those data elements, whether revenue from a person can be generated or not. This will eventually help all the companies to target on the potential customers that means who are more likely to purchase from online during their visit. Elements like page value, special day and traffic were the deciding factors for the generation of revenue. This will let the companies to plan effective campaigns of advertisements and proper budget allocations can be made. The prediction analysis of shoppers intention was performed by using Extreme gradient boosting otherwise known as XGBoost. It is the one of ensemble technique based on sequential method. Due to its high performance as well as accuracy, XGBoost algorithm was used when compared to other algorithms.

II. RELATED RESEARCH AND APPLICABLE TECHNIQUES

A prediction model for predicting the purchase intention of online products and services via internet was proposed in paper [1]. The features that are extracted from the model are fed to multilayer perceptron (MLP) and random forest. In order to boost the efficiency, oversampling was done. The findings indicate that, MLP with back propagation generates substantially high accuracy and F1 score when compared to random forest.

Social media, nowadays has became one of the key influential factor in terms of purchasing intention. Most of the people are relying on recommendations and opinions based on social media flatform. In this research work author proposed that decision making related online shopping depends on social media influence than the information quality. With a sample of around 500 consumers from Facebook indicated that social media has great impact on shopping intention [4].

As online shopping gained much popularity in these days, it has become a new dimension in business sector. This has made the sellers to analyze the intentions and patterns of those who do online shopping. By using different classification algorithms and gradient boosting, prediction model was built where compared to classification algorithms like Decision tree, Random forest and SVM were proved to show less accuracy than gradient boosting. As gradient boosting algorithm is one of ensemble algorithm, it outperforms the other algorithms by giving highest accuracy of 90.34% [5].

Different kinds of promotions and flexible return policies were proved to be the important factors in terms of revenue generation for online shopping. In this study, author finds the relationship between shoppers repurchase intention and the effect of return policies and promotions by using full connected LSTM. Proposed methods for extracting customer's browsing features with respect to different channels across sales promotions. In this proposal, deep learning framework have been used along with sigmoid layer. [6].

People purchase intention depends mainly on the promotions or offers as per availability. People normally tend to compare the prices with the promotion channels. In this paper, author proposed a deep learning model using long short term networks to find the relationship between promotion channels and customers. With the previous history, demographic details and other elements, model was able to predict their purchase intention. Using full connected deep neural networks outperforms the other machine learning algorithms with respect to accuracy and other metrics [7].

E-commerce companies must maintain a good relationship with the customers in order to maintain them in long run. This research work examines the factors that favors the shopping experience. As there will be no interaction with people by the companies, it is more important to know their needs and concerns than any other. This study reveals that product pictures or visuals contribute more towards the intention of online shoppers [8].

Recommender systems were launched long ago that allows consumers to choose the right products or services from different alternatives. The basic idea behind this is to aid consumers new products or items rather than manual search this is the fundamental principle of recommendation system. In this paper [9] applications and challenges of recommendation system was analyzed and further study was conducted on predicting the shoppers intention where it was proved that Random Forest algorithm performs with best accuracy and ROC value of around 93% when compared to other algorithms.

With the hit of commercial websites and their products on internet, online shopping has become a part of daily life for many people around the world. With the ease of convenient shopping and enhanced options in online websites also contributes to raising demand. A decision-based support

system was proposed in order to predict the purchase intention of consumers during their browsing sessions. Extreme gradient boosting which is one of Ensemble methods was used to classify the purchase intention of consumers. With the features of content entropy, model was able to gain around 41% and 34% of recall and F1 scores respectively. It was showed that, feature engineering techniques like outlier handling, imputation and one hot encoding played an important role in building the model.

Online shopping in India and across many parts of the world has shown a tremendous growth in recent 10 years. An attempt has been made to predict the behavior of online shoppers. With the help of collected data based on survey, Random Forest based prediction model has been built. At first, missing values in the data have been replaced followed with encoding of data into binary format. Random forest algorithm has performed well in terms accuracy and ROC curve.

III. THE TECHNIQUE EMPLOYEED

The insights of literature survey presented that ensemble techniques are most apt methods to be followed for the data scenario in this work. Of various ensemble techniques available XGBoost is opted as primary modeling technique in this work.

XGBoost

Extreme Gradient Boosting (or) XGBoost is a variant of gradient tree boosting with regularization involved. XGBoost acts as a scalable system for learning tree ensembles, it improves the model performance by devising better regularization objective. Xgboost is a sparse aware algorithm which supports parallelization and cache optimization. Xgboost learns a self-contained derivation of gradient boosting algorithm which preliminary calculus.

Boosting creates collection of predictors. Here, learners will learned sequentially with early learners fitting simple models to the data and then data is analyzed for errors. Consecutive trees (random sample) are fit and at every step, goal is to improvise the accuracy from the prior trees. When the input is misclassified by hypothesis, its weight will be increased so that next hypothesis is more likely to classify it correctly. This process converts weak learners into better performing model. XGBoost is exactly a tool motivated by the principle of boosted trees. Importantly, it emphasis on consideration in terms of systems optimization and principles in machine learning. The goal of XGboost is to push the extreme of the computation limits to provide a scalable, portable and accurate results.

Features of XGboost:

Regularization: XGBoost penalize the complex models with the help of L1 and L2 regularization which helps in prevents overfitting of model.

Handling sparse data: XGBoost implements sparsity-aware split finding algorithm in order to handle multiple types of sparsity patterns in data.

Weighted quantile sketch: XGBoost also has distributed weight quantile sketch algorithm to manage the weighted data effectively.

Block structure for parallel learning: XGBoost can use multiple cores on CPU for fast computing. It is possible due to block structure in which the data will be sorted and stored in terms of memory units called blocks. This enables data layout to be reused by the subsequent iterations, rather than computing it once again.

Cache awareness: For XGBoost, no continuous memory access is necessary to fetch the gradient statistics in terms of row index. XGBoost has been designed in such a way to allow optimal use of the hardware. This can be done by allocating the internal buffers present in each thread, where it is possible to store the gradient statistics.

Out-of-core computing: This feature will optimize the accessible disk space and also maximize the usage to handle the huge datasets which doesn't fit into memory

XGBoost with extreme computing capability and unique set of features make it most suitable algorithm for this work.

IV. METHODOLOGY

This section work describes the steps involved in developing the model. The dataset is obtained as 'csv' file from UCI Machine Learning Repository and read into our frame-work using 'pandas' library.

A. Exploratory Data Analysis:

The Exploratory Data Analysis (EDA) is implemented on our dataset providing us insights that data is of dimensions (12330 x 18) and there are null values present in various independent variables like {'Administrative', 'Administrative_Duration', 'Informational', Informational _Duration' , 'Product Related', 'Product Related_Duration'} all these null values are filled with median values of respective variables , as median values are not much effected by outliers. The categorical variables of {'Month', 'Visitor Type', 'Weekend', 'Revenue'} which are of object and Boolean data types were label encoded. A correlation map in figure.1 is drawn between variables to observe presence of any significant relation between independent and dependent variables and also for multi-collinearity.

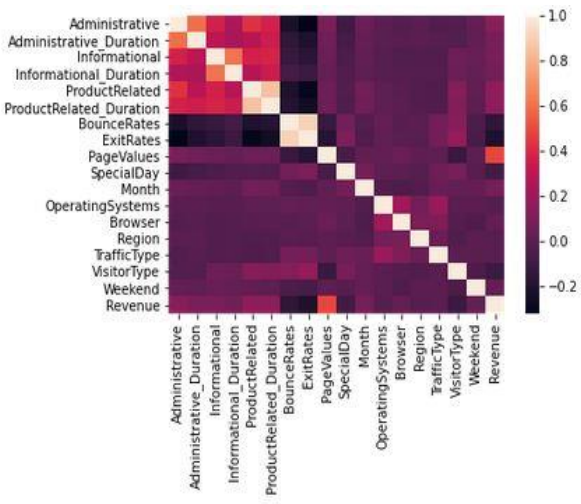


Figure.1: Correlation Plot

Based on insights obtained from correlation heatmap, the variables {'ProductRelated_Duration', 'BounceRates'} were dropped from consideration as they are highly correlated with each other. A count plot is drawn in figure.2 to see the count on nature of visitors visiting the website.

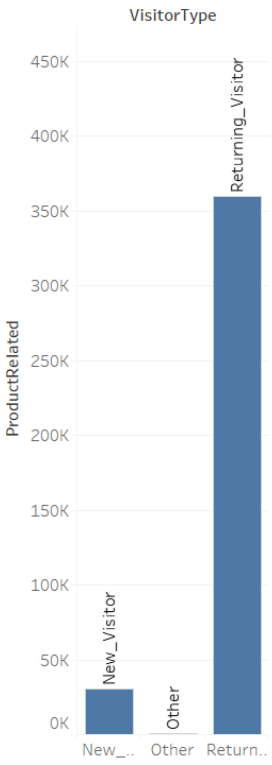


Figure.2: Types of visitors

From figure.2, returning visitors are more in number when compared to new visitors and other visitors. This shows that number of people who visited the website earlier, they have the high chance of visiting website again than people who are visiting for the first time. The count plot for customers vs month in figure.3 backs the insights drawn from pie chart above presenting that more customers visited during the months of May and November.

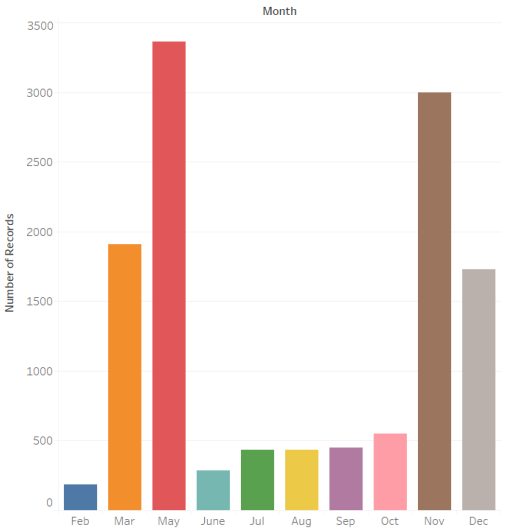


Figure.3: Customer's shopping in different months

An effort is made to understand the Revenue variable, as part of these efforts 'Revenue' variation with respect to 'Month', 'Visitor Type', 'Weekend', 'Special Days' were presented as plots.

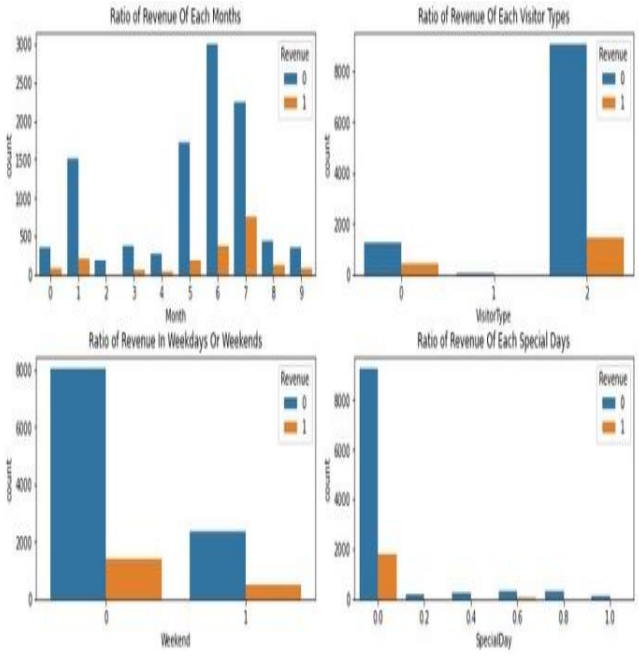


Figure.4: Generation of Revenue in different categories

B. **Data Preparation:**

The data obtained after null values filling and dropping based on correlation is made split into train and test for classifier building. The entire dataset is divided in terms train and test

subsets with ratios of 65% and 35% respectively. This subset division is done with 'sklearn model selction' framework and is very important to evaluate the performance of the model.

C. Findings:

The XGBoost classifier from sklearn framework is supplied with train data we prepared. XGBoost will get trained on this data which enables it to learn the pattern inherently present in data. The trained XGBoost classifier is used to make predictions on test data and the predictions are validated against the ground truth. The XGBoost classifier modeled in this work presented an accuracy of 90.28% on test subset. Confusion matrix and classification report of XGBoost classifier are obtained in figure.5.

```
[[6485  289]
 [ 490  750]]
              precision    recall  f1-score   support

           0       0.93      0.96      0.94      6774
           1       0.72      0.60      0.66      1240

    accuracy                           0.90      8014
   macro avg       0.83      0.78      0.80      8014
weighted avg       0.90      0.90      0.90      8014
```

Figure.5: Classification Report

The Receiver Operating Characteristics (ROC) curve is a plot between true positive and false positive rates respectively. The model is more accurate if the curve is closer towards left border in the graph shown in figure.6.

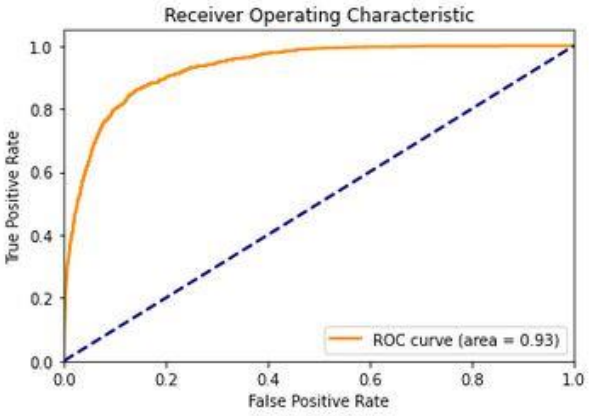
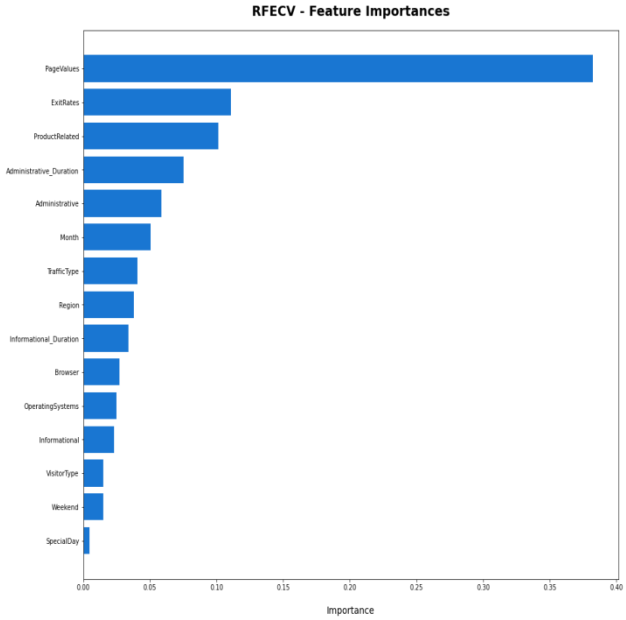


Figure.6: ROC Curve

D. Feature Importance:

Finding the optimal set of features plays a vital role in classifiers performance. Combination of Random-Forest algorithm and Recursive Feature Elimination With Cross Validation (RFECV) is used to find the feature importance. According to ranking of important features in the dataset, Page value, Exit rate, Product related are the most influenced features in terms of predicting shoppers intention.



V. Conclusion

The main objective behind this study was to analyze the important factors which contributes in influencing the online shopping. And by using suitable machine learning algorithm, which is Extreme gradient Boosting, a prediction model has been built to predict the purchase intention of online shoppers. This study provided important insights based on which the revenue generation from online websites can be predicted. By using XGBoost algorithm, 90% of accuracy was achieved by the model. Recursive feature elimination technique, enabled to find the rankings of the features in a hierarchy. It was showed that Page values, Product related and Exit rates are among the important features based on which revenue generation from online shoppers can be predicted. Although good accuracy was achieved, it can further be increased with the help of deep neural networks. Finally, by using XGBoost for the prediction of shoppers intention could make e-commerce easier, more profitable and convenient for all stakeholders.

References

[1] Sakar, Polat, S. Katircioglu, M. Kastro :" Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks". Neural Comput. Appl. 2018.

[2] Malik G. & Guptha A. "An Empirical Study on Behavioral Intent of Consumers in Online Shopping. Business Perspectives and Research". 2013.

[3] Lim. Y, Osman A, S. Shahrul, Salahuddin S, Romle A." Factors Influencing Online Shopping Behavior: The Mediating Role of Purchase Intention" Procedia Economics and Finance, 2016.

[4] Jen Ruei Fu, I. Wei Lu, Jessica H.F. Chen, Cheng Kiang: "Investigating consumers' online social shopping intention: An information processing perspective", International Journal of Information Management, 2020.

[5] M. Kabir, F. Ashraf and R. Ajwad, "Analysis of Different Predicting Model for Online Shoppers' Purchase Intention from Empirical Data," 2019 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2019.

[6] Xiaolin Lia, Zhuanga Y, Benjiang L, Guoqing C "A multi-stage hidden Markov model of customer repurchase motivation in online shopping",2019.

[7] Zhang, T, Ling C & Chen, Y. "Customer Purchase Intent Prediction Under Online Multi-Channel Promotion: A Feature-Combined Deep Learning Framework".2019.

[8] C. Changchit, S.J Douthit and B. Hoffmeyer, "Online shopping: what factors are important to shopping?," Journal of Electronic Commerce, vol. 10,2004.

[9] Sharma R., Kaur, B, Rani S, and Gupta D, "Recommender system: Towards classification of human intentions in e-shopping using machine learning", Journal of Computational and Theoretical Nanoscience, 2019.

[10] B. Liu and B. Zheng, "A scalable purchase intention prediction system using extreme gradient boosting machines with browsing content entropy," 2018 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, 2018.

[11] Joshi R, Saravanan, P and Gupte, R, "A random forest approach for predicting online buying behavior of Indian customers". Theoretical Economics Letters,2018.