National College of Ireland

Project Submission Sheet – 2021/2022

Student Name:	AMRUTA VENKATESH, KAPIL LALWANI, RAJ SHRIKANT SONAWANE, ROHAN SANJAY KORE, VIPLAV VIJAY GADEWAR		
Student ID:	X21168580, X21123292, X21155054, X19214413, X21164274		
Programme:	MSCDAD_JAN22A_I	**Year:**	2022
Module:	DOMAIN APPLICATION OF PREDICTIVE ANALYTICS		
Lecturer:	VIKAS SAHANI		
Submission Due Date:	11-08-2022		
Project Title:	PREDICTION OF CUSTOMER'S INTENTION TO PURCHASE		
Word Count:	2367		

I hereby 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 references 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:	AMRUTA VENKATESH, KAPIL LALWANI, RAJ SHRIKANT SONAWANE, ROHAN SANJAY KORE, VIPLAV VIJAY GADEWAR
Date:	11-08-2022

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **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. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

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

Prediction of Customer's Intention to Purchase

1st Amruta Venkatesh
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x21168580@student.ncirl.ie

2nd Kapil Lalwani
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x21123292@student.ncirl.ie

3rd Raj Shrikant Sonawane
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x21155054@student.ncirl.ie

4th Rohan Sanjay Kore
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x19214413@student.ncirl.ie

5th Viplav Vijay Gadewar
MSc. Data Analytics
National College of Ireland
Dublin, Ireland
x21164274@student.ncirl.ie

***Abstract*—Everyday millions of users do shopping using the websites around the world. Conventional stores spend a large amount of money to go online. As a result of utilising the websites' session data, a lot of information is generated from the users. This data can be of very good value if analysed properly it can generate great value to business. To improve and upgrade the predictive capacity of the customers in purchasing behaviours on websites platforms, a new method has been used in this paper. This study introduced the fundamental standards of the XGBClassifier, investigated the historical data of an web sessions, pre-processed the original data and constructed an Online Shoppers Purchasing Intention model based on the XGBClassifier. With the help of this classification merchants who wanna sell things online can add a good business value.The characteristics of the significance of the results were analysed using visual representations.The outcomes demonstrated that by utilizing the XGBClassifier to anticipate the purchasing intentions of the customers, can improve the performance and a better prediction effort.**

***Index Terms*—web sessions analysis; purchasing behaviour prediction; XGBClassifier**

I. Introduction

The e-vending sector has grown significantly in recent years. Online shoppers have a wide range of choices. When a buyer needs to purchase a product, it only takes a few clicks for them to compare the goods on other websites. Therefore, businesses must make every effort to turn potential clients into customers in order to succeed in this cutthroat economy. To do this, we may enhance our business tactics by predicting clients' purchasing intentions.

In the competitive marketplace, the e-commerce industry has seen both benefits and challenges. The value and significance of the web session data that is hidden may be assessed through information mining, which also helps to increase the consumers' desire to buy. In order to increase the exchange volume of web platforms and further the development of web platforms, this article provides a powerful technique for the prediction of users' purchasing patterns. Python programming and the machine learning XGBClassifier technique was used to finish the investigation. Because it can read a big dataset and forecast results with the highest degree of accuracy, the XGBClassifier model is regarded as one of the greatest machine learning mitigation strategies.

The objective of this study is to :

- Predicting a customer's buying intention using Web session dataset.
- Thorough examination of historical data and forecasting of future patterns that will aid in creating positive income.

This work incorporates calculations for XGBClassifier to clasiify client buying patterns on web sessions. The features of consumers' desire to buy are investigated and analyzed in this article utilizing some Python tools. Extensible learning frameworks may be used by the XGBClassifier to benefit from huge datasets and produce models since they can accurately capture the circumstances of large datasets. Targeted planning, intelligent choices, and prompt marketing decision-making are used to provide the desired results, which include increased client attraction, significant financial gains, and the anticipation of favorable outcomes.

II. Hypothesis

- **Null Hypothesis:** The amount of money made depends on how many people come on a weekend day.
- **Alternate Hypothesis:** The weekend that customers arrive and the revenue generated are unrelated.

III. Related Works

Online shopping gives a platform for retailers to reach out maximum target customers. Today the buyers have many options for purchasing a product online. Therefore, the sellers online should consider various factors to increase their online conversion rates. Various studies have been carried to determine the buyer's intention. We aim to research the elements that make precise and scalable purchase intention prediction for virtual retail environments feasible.

In this study [1], the author employed RF, SVM, and MLP to forecast website abonnement probability and users' propensity to shop. When applied to clickstream data and other crucial criteria for prediction that we have considered for our investigation, MLP has the greatest accuracy. To increase the ratio of our abandonment to buy conversion rate, the modules work together to assist us separate the possible positive users who are likely to abandon in the forecast horizon.

The author of this study [2] has looked at what causes a client to leave his shopping cart empty at the transaction stage. According to the study, client dissatisfaction with the transaction is what causes most people to leave. The consumer qualifies as a prospective income producer because they have already completed their shopping. Therefore, increasing the conversion of these clients should be the goal. The author has looked at several reasons that might cause a consumer to leave before paying. The conclusion is that neglecting customer expectations throughout the checkout process may be just as harmful as ignoring customer demands early in the customer purchase process.

The author of this work [3] has provided a machine learning-based online recommendation system. The author divided users based on their activity and targeted users who could potentially generate any potential revenue using a combination of techniques including tracking online activity using business process modeling, collecting statical data using Google Analytics, and classification algorithms.

The author of this research [4] used analytics to examine and enhance the website OrOliveSur.com's favorable online discourse. Clustering, association rules, and subgroup finding are the techniques employed. With regard to the approaches employed, various knowledge may be analyzed in this online usage mining study. The study also demonstrates how important it is for administrators to monitor visitors to reference websites.

IV. Selection of Techniques Employed

The purpose of this domain application is to make a prediction about whether customers will generate revenue or not. Hence, the target variable is of dichotomous type, the applicable techniques considered were Logistics Regression, Support Vector Machine, Random Forest Classification, K-means clustering, and lastly XGBClassifier. In every online shopping business, customers either visit the website to gain knowledge on the prices of the product, or to compare prices from other websites. Some customers may find a product suitable to be purchased while some may not and leave the website without buying any product. As shown in figure 1, the data-set gathered from the public library, 20 percent of the customers who visited the website have generated revenue for the company by making some purchases. While the rest of 80 percent have not. So, the data is highly imbalanced and may not provide the best accuracy in the prediction of customers purchasing intention.

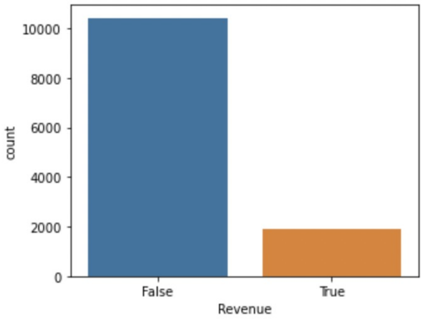


Fig. 1. Revenue Generated: True vs False

While the above-mentioned models work well only when the predicting variable is balanced or closely balanced with an ample amount of data provided. Supervised Machine Learning models such as SVM, Logistics Regression, and Random Forest may produce poor results by being biased towards a class that has accounts large enough which is the customer not making any purchase and not generating any revenue to the company.

As defined, the domain is a website that wants to study customers' intentions without being biased towards the majority class. To deal with class imbalance, we have chosen the XGBClassifier model with a combination of performance-enhancing techniques such as Hyper-parameters tunning. By setting the parameters to desired values we can predict the positive outcomes which account for only 20% of total revenue generation results.

The selected algorithm has proven to provide acceptable results, which have been evaluated in the Evaluation section with appropriate metrics.

V. Implementation of the technique

A. Methodology

In order to better serve the demands of the study, a few small adjustments to the CRISP-DM technique were made in order to implement the predictive analysis. The CRISP-DM technique, as opposed to the KDD methodology, which largely focuses on the project's technical components, displays and makes use of the Business Understanding phase, which is how this was accomplished. Below figure 2 depicts the methodology.

B. Data Understanding

The dataset obtained through UCI machine learning repository [6] consists of feature vectors from 12,330 web sessions with 18 attributes. The class label is created using the 'Revenue' attribute. To eliminate any inclination to a particular campaign, special day, user profile, or time frame, the dataset has each session belonging to a distinct user over the course of a year.

C. Data Handling and Pre-processing

Dataset is loaded to data frame using python in Jupiter Notebook. For Pre-processing, Data is checked for null values and null rows are removed from dataset. Data set is further split into 80% for training purpose and 20% for testing purpose.

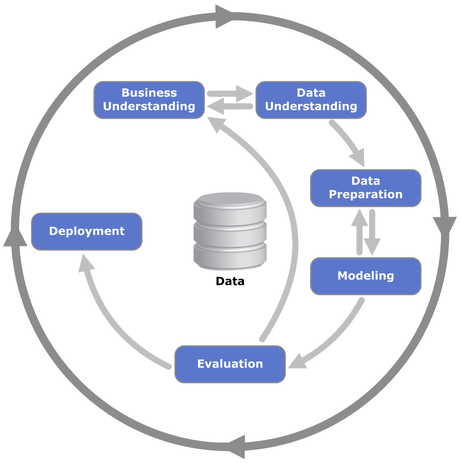
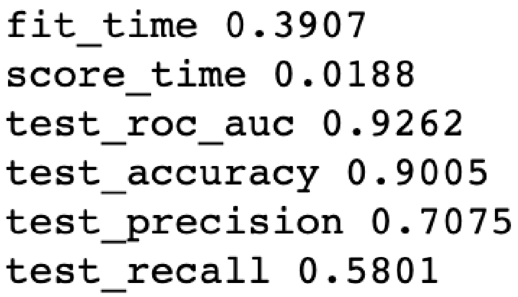


Fig. 2. CRISP-DM Process Diagram

D. Predictive Model

After examining the relevant work that has been done in the field, it was discovered that the XGBClassifier was the most well-known predictive model utilized in order to get the best accuracy for the predictions and a benchmark accuracy for the predictive analysis. The dataset was separated into training and testing sets, and then the models were trained on the training set before being applied to predict the location of these independent variables. The test accuracy for the model was 90.05% as shown from below results figure 3.



```
fit_time 0.3907
score_time 0.0188
test_roc_auc 0.9262
test_accuracy 0.9005
test_precision 0.7075
test_recall 0.5801
```

Fig. 3. XGBClassifier Result

VI. QUANTITATIVE AND QUALITATIVE INTERPRETATION

A. Exploratory Data Analysis

The arrangement of the properties was modified based on their categories and numeric foundation. The categorical qualities were transformed into organized factor variables and numerically encoded for modeling purposes. The dataset's numeric variables were standardized for clustering algorithms and scaled for classification. During the training session, eighty percent of the data was used, while the remaining twenty percent was used to validate our model. No values are missing from the dataset.

Relationship between Bounce Rates and Exit Rates: A high bounce rate may suggest problems with customer satisfaction [1] due to one or more factors, such as an unpleasant user interface, a sluggish throughput, or other technical concerns. A high departure rate may be indicative of under performing sectors in sales funnels, indicating opportunities for improvement, as if consumers are departing, then no one is buying. According to Big-Commerce [2], an appropriate bounce rate is from 30 to 55 percent. Our data reveal that the 7 bounce rates are widely dispersed below 10% as shown in figure 4. According to upward business [3], a bounce rate of less than 5% is grounds for caution. Therefore, more research is required on these facts. Assuming there is in fact no mistake, we might search for methods to improve bounce rates and departure rates to guarantee sales preservation and client retention.

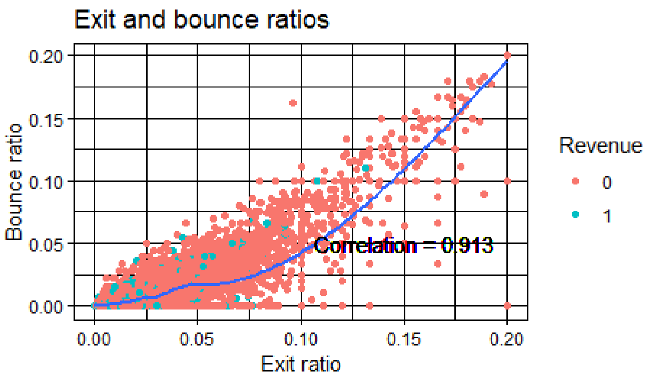


Fig. 4. Exit and Bounce Rates

Correlation: As dipicted in figure 5 majority of the numerical qualities seem to display a considerable degree of positive skewness, whereas others demonstrate a negligible amount of negative skewness.

Revenue based on loyalty and weekend: As figure 6 illustrates that the majority of clients, regardless of whether they generate money or not, are repeat customers, indicating that the company has dealt well with customer retention. However, it is evident that conversion rates need improvement. It is usual for businesses to prioritize one transformation and neglect the other. While customer retention is an indicator of brand value, a lack of new client acquisition might drastically influence business results. The majority of customers entered and made purchases over the week. This might be exploited further by attempting to increase weekend client watching and purchases.

Figure 7 and figure 8 demonstrate the seasonal rise in revenue. The overall pattern is declining after the months of February,

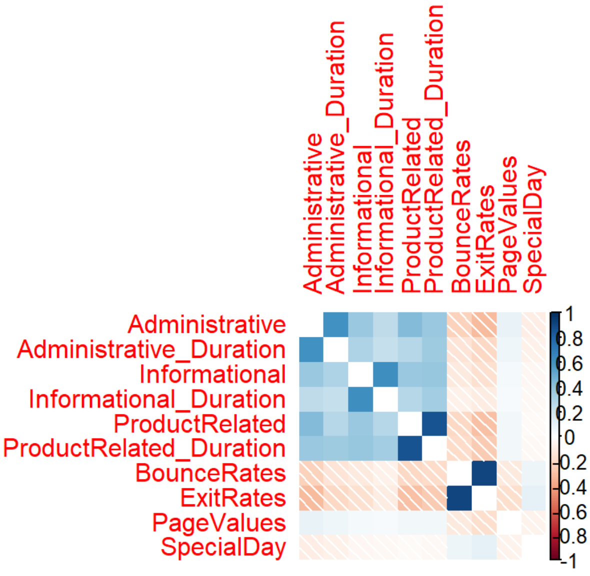


Fig. 5. Segment Correlation

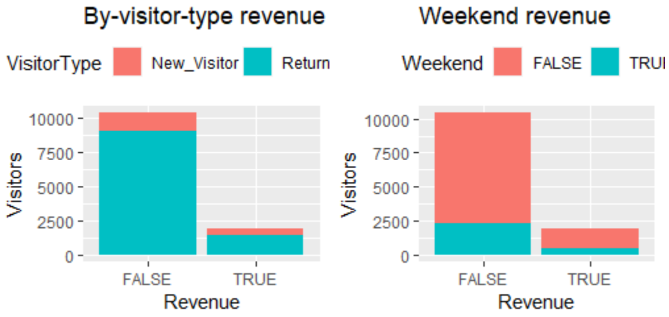


Fig. 6. Visitor-type Revenue and Weekend Revenue

March, and May when client involvement appears to be at its highest. In addition, the tendency seems to plateau from June to October, after which there appears to be an increase in interaction since the holiday on account of Black Friday arrives. When demand looks to be strong, there appears to be a great deal of interaction, but conversion rates are substantially lower since the majority of these sales are fueled by recurrent consumers. While this implies the existence of strong customer loyalty, more attention must be paid to conversion, since the plots above indicate that a large number of consumers examine your items but do not complete a purchase.

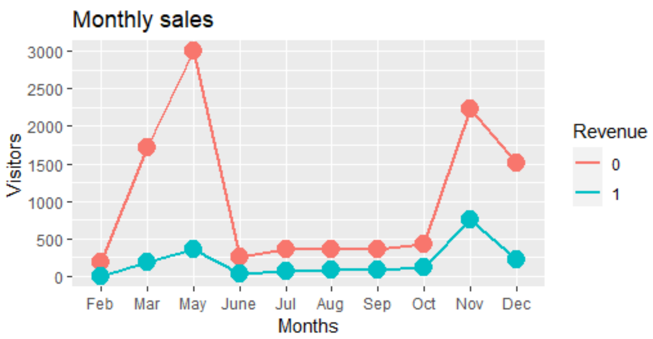


Fig. 7. Monthly Sales Trend

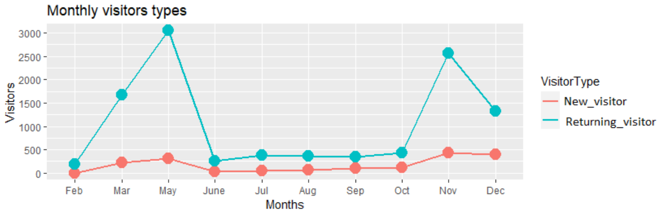


Fig. 8. Monthly Visitor-type Trend

VII. The Business Value Qualitative Interpretation of the Findings

1) According to the interpretation given above, a high bounce rate may be one of the reasons why 80% of people that visit a website don't generate any money for it. High product prices, bad user interference, or a delay in the reaction are all potential causes. By considering the market competition and setting rates appropriately, you may draw more people's attention and convert them into customers. Additionally, having a UI that is appealing and easy to use encourages consumers to stay on the website longer.
2) Recurring clients make a significant contribution to producing money, according to another report. In order to better understand what else the organization may do to enhance its services, it is advised that it conduct feedback or survey procedures.
3) The general pattern begins to change after the months of February, March, and May, when customer involvement appears to be at its peak. The website should develop various methods, such as summer deals or "Back to School" offers, to increase income generating all year round and increase client engagement.

References

[1] J. Yang, R. Sarathy and J. K. Lee, "The effect of product review balance and volume on online Shoppers' risk perception and purchase intention," Decision Support Systems, vol. 89, pp. 66-76, 2016

[2] D. V. d. Poel and W. Buckinx, "Predicting online-purchasing behaviour," European Journal of Operational Research, vol. 166, no. 2005, pp. 557-575, 2005.

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

[4] C. O. Sakar, S. O. Polat, M. Katircioglu and Y. Kastro, ”Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks,” Neural Computing and Applications, vol. 31, pp. 6893-6908, 2019.

[5] Rajamma, R.K., Paswan, A.K. and Hossain, M.M., 2009. Why do shoppers abandon shopping cart? Perceived waiting time, risk, and transaction inconvenience. Journal of Product & Brand Management.

[6] https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset