Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

J Component report

Programme : B.Tech (CSE)

Course Title : Essentials of Data Analytics

Course Code : CSE3506

Slot : A1

Title: Customer Segmentation and Sales Prediction based on Web Activity

Team Members: Maddukuri Nivas | 19BCE1010

Lakshmi Sairam Kakarla | 19BCE1052

Kovvuri Uday Surya Deveswar Reddy | 19BCE1253

Faculty: Dr. Raghukiran Nadimpalli

Abstract 2

Introduction 3

Objective and goals of the Project 4

Problem Statement 4

Motivation 5

Challenges 5

Literature Survey 6

Requirements Specification 7

Hardware Requirements 7

Software Requirements 7

System Design 8

Implementation of System 8

Results and Discussion 8

Conclusion and Future Work 14

References 14

Appendix 15

Abstract

Online shopping applications are always trending these days. It is a space where people may go to the website, pick the products they want, fill out the relevant information, pay (or choose to pay later), and buy goods and services in a flash. There are many people who purchase the products on the websites but there is no proof to show that every person i.e. whoever is using the online shopping websites is intended to buy a product. They may be visited to compare the prices with other websites or they can be your competitors. The objective of our work is to predict whether the product will be purchased or not and to find key factors which play a role in making customers buy a product. Performing customer segmentation will optimize promoting techniques, maximize a customer's esteem for business, and make strides in client encounter and fulfillment.

Introduction

In the last few years, the business has increasingly moved into the electronic realm. The customer's role has a major impact on e-commerce websites. E-commerce websites allow people to buy products using the internet rather than face to face. In today's scenario, E-commerce websites make the relationship between people and shopping very strong, flexible, and viable. Customer service is the most salient factor for any kind of website. It was the place where the user engages with the website to buy a product. Customers of the websites utilize the services in various ways other than to buy a product. There are many benefits to using E-commerce websites such as it helps you save time and effort. With the ease of purchasing from the comfort of your own home, there is a wide choice of items accessible. Discounts and reduced rates are available. Find out everything there is to know about the product. We can compare and contrast various models and brands. By examining the above-mentioned benefits, they can also become threats to a dull E-commerce website. Customers can compare various products on the website with other websites. The quality of the website plays a major role in generating revenue in E-commerce websites.

The pandemic of Covid-19 has had a significant influence on the e-commerce business. During the epidemic, the industry saw both positive and negative effects. People are buying more products online since it is safer in terms of disease transmission. According to statistics, the e-commerce business is experiencing higher sales growth [1]. Online buying is developing at such a rapid pace that the global online shopping market is expected to reach about $4 trillion by 2020. In the United States alone, they predict 300 million internet shoppers by 2023. Because transactions related to purchase confirmation on E-commerce websites are often completed at the end of user sessions, certain information about current sessions can be watched and utilized to determine the likelihood of making a purchase [2]. The web click logs include a lot of potentially helpful information, and they can help you figure out how people's browsing habits relate to what they wish to buy [3]. The Web consumer is influenced by product involvement, Web skills, obstacles, and the usage of value-added search tools [4].

Consumers are increasingly using electronic channels to buy and sell their daily necessities. Recently, there has been an increase in customer interest in online buying. Consumer awareness has increased as a result of technological advancements and the growth of information technology, and individuals now prefer to buy online rather than visit a store to acquire products or goods [5]. In an online purchase, the decision-making process, which includes brand selection and price sensitivity, is unique. Several incentive elements, such as situational factors, product qualities, and previous e-purchasing experience, might influence consumers' views toward shopping online. Furthermore, accessible decision support tools can assist people in making sensible decisions in the face of overwhelming data [6].

Our work is based on customer segmentation and predicting the sales using the log activity of different websites. This type of analysis and clustering will be useful for e-commerce websites to improve the quality of websites such that they can follow some measures to increase their revenue. By considering the various factors like the average time customers are spending on product review pages to make a successful purchase of a product, we will find some key inferences. By using different packages which are available in R, we will analyze and we will perform customer segmentation. Performing customer segmentation will optimize promoting techniques, maximize a customer's esteem for business, and make strides in client encounter and fulfillment. For individuals who are about to start a new business or planning for a startup, these analytics and classifications would be helpful to convey how customers interact with the websites to buy a product.

Objective and goals of the project

This is a competitive world, and the population is rapidly expanding. The expansion of e-commerce websites was enormous as a result of these pandemic scenarios. Many businesses are launching their own e-commerce sites. Because of the rivalry, every organization must improve its website by using various technologies such as AI, NLP, and others. Not only that but each and every employee's log data must be tracked. So that the organization may do a thorough study of each website visitor. The individual who is purchasing should be given first consideration. We can forecast if the individual is willing to buy based on the data. So, the website can show more and more suggestions to the person who is willing to buy. This does not mean that the system should not recommend anything for those who are not willing to buy. But this is just a term that can be included while recommending.

The goal of this project is to use log data to forecast whether or not a user will buy a product. We may use Machine Learning classification algorithms to predict this, and the other goal is to do cluster analysis to determine the sales Key Performance Indicator (KPI). Finally, we determine the best cluster number by using the classification method with the highest accuracy and the elbow curve, which is basically the withiness sum of each cluster given a range of hyperparameters.

Problem Statement

The objective of our work is to predict whether the product will be purchased or not and to find key factors which play a role in making customers buy a product. Performing customer segmentation will optimize promoting techniques, maximize a customer's esteem for business, and make strides in client encounter and fulfillment. For individuals who are about to start a new business or planning for a startup, these analytics and classifications would be helpful to convey how customers interact with the websites to buy a product. The scope of our work is finding the ways to increase the revenue of e-commerce websites and finding the strategies that should follow to increase the users.

The dataset contains 12.3K records and 18 features including page value, exit rate, bounce rate, etc. It is possible to predict whether revenue will be generated or not. And using unsupervised learning, it is possible to do the customer segmentation and take action based on groups. To get the best accuracy, we need to do a few data preprocessing techniques like downsampling, Outliers elimination, etc.

Motivation

The main motivation of this project is to assist e-commerce websites to better serve users by using their online activity to predict the probability of revenue. The major application is to create firewalls and dedicated servers to subset user requests so as to provide an improved user experience to candidates with relatively more probability of purchase. Additionally, we can identify features and parameters and their weighted contributions to the sales made by users. This may be something as simple as the choice of the first webpage the user views while visiting a website.

We can implement real-time analytics like the average time user has to spend on a product page to make a purchase. If organizations are able to engage users for the required time, the probability of purchase will increase. The analytics also provides the SEO state of the organization's webpage. The average page weight of the pages in an organization, if it's lower than the required, organizations can explore techniques to improve this score.

Challenges

There are various challenges in this project to overcome, the major challenges are listed below.

1. Imbalance in Dataset is a major problem because we had to equip downsampling or oversampling techniques to remove bias in the models.
2. Quality and reliability of data is a major question we need to address. From the sourced logs of the server, the activity of the users is tracked. Additionally internally decided page scores and bounce rate are used, whose calculations are not made public, so the possibility of error or augmentation is to be considered.
3. With so much data available, it's challenging to sift through it all and find the information that's most important. Employees who are overworked may not thoroughly examine data or focus on the measurements that are simplest to obtain rather than those that genuinely provide value. Furthermore, if an employee must manually filter through data, real-time insights into what is actually happening may be impossible. Data that is no longer current can have a major detrimental impact on decision-making.
4. The next problem is analyzing data from various, disparate sources. Distinct systems often house different types of data. Employees may be unaware of this, resulting in inadequate or incorrect analyses. Manually integrating data takes effort and limits insights to what can be seen quickly.

5. Inaccurate data is the number one enemy of data analytics. The output will be unreliable if the input is poor. Manual data input errors are a major source of erroneous data. If the analysis is utilized to influence decisions, this can have serious negative repercussions. Another issue is asymmetrical data, which occurs when information in one system does not reflect changes in another, causing it to become obsolete.
6. Budget is another issue that risk managers face on a regular basis. Because risk is frequently a tiny department, getting permission for large acquisitions like an analytics system can be challenging.

Literature Review

Radhakrishnan, V., 2021 [1] Their study covers the effect of E-Commerce websites on the world and examines the issues and causes that affect online businesses in-depth, as well as learn about the state of e-commerce in various parts of the world., they interpret that more than half of web users aged 25 to 44 in the 17 nations surveyed said they've spent more time shopping online in the last few weeks, with men, in particular, saying they've increased the amount of time they spend on e-commerce activities. Suchacka G, Skolimowska-Kulig M, 2015 [2] in their study the authors formulate the task of predicting buying sessions in a Web store as a supervised classification problem with two target classes linked to whether a purchase transaction is completed in session or not, and a feature vector containing variables describing user sessions. The k-Nearest Neighbors (k-NN) classification is used in the described method.

Wang, Y.T. and Lee, A.J., 2011 [3], For e-commerce systems, the authors advocated using web browsing history mining to determine customer preferences. To model an individual's web surfing history, we provide a user browsing-history-hierarchical-presentation-graph. A new method called UPSAWBH (User Preference Similarity Calculation Algorithm Based on Web Browsing History) is proposed, which measures the level of preference similarity between users based on their web page click patterns, and an interesting web page detection algorithm is designed to extract users' preferences.

Koufaris, M., 2002 [4], look at how online shoppers' emotional and cognitive responses to their first visit to a Web-based store can influence their intention to return and their chance to make accidental purchases. The findings support the online shopper's dual identity as a shopper and a computer user, as both buying pleasure and perceived site utility highly influence the intention to return. Kasuma, J., Kanyan [5], their study's major goal is to look at the aspects that influence customers' online purchase intentions. A total of 200 people took part in the survey. Customers' online shopping intentions are influenced by aspects such as convenience, time savings, website/features, and security, according to the empirical findings. Wei, L., 2016 [6], Their research concluded that, due to a large number of product selections, a recommendation agent is required for clients to list their preferred options. It restricts products that customers do not regard to be their options and forces them to make successful product selections.

Requirements Specification

To function properly, hardware necessitates the use of the software. Your software may not run efficiently or at all if you don't have the right hardware. When making decisions, it's critical to think about both.

Hardware Requirements:

1) **64 - bit Platform:**

   The word 64-bit refers to a computer generation in which 64-bit processors are standard. A word size of 64 bits is used to describe particular types of computer architecture, buses, memory, and CPUs, as well as the software that runs on them.

2) **Free - Space:**

   On the hard disc, 250MB of free space is necessary for the setup installation. To store the projector to execute the project in its entirety, a minimum of 300MB is required.

3) **RAM:**

   RAM saves the data that your computer is currently working with so that it may be accessed quickly. For the project, the minimum required RAM is 1GB. But the recommended size would be a minimum of 2GB.

4) **Processor:**

   The processor, commonly known as the CPU, gives the computer the instructions and processing power it needs to function. For the project, Minimum-Intel Core i3 2.5G Hz, Recommended-Intel Core i5

Software Requirements:

1) **Rstudio:**

   R is a statistical computing and graphics programming language, and RStudio provides an integrated development environment for it. We may build up the platform by downloading it directly from the rstudio website. It may also be utilized from the cloud.

2) **Version:**

   To run this Rstudio, the windows version should be 7 or later. In Mac os, Rstudio can be used in the latest version 1.4. 1717 compatible with the M1 processor.

3) **Libraries:**

   There are certain libraries to be installed to run the project or to get better insights. They are ggplot2, dplyr, read_excel, cowplot, e1071, and respective packages for the classification and clustering models.

System Design

For this project, we aim to perform Predictions of sales and to cluster the customers to obtain KPI for sales. Clustering is also being used for Classification by setting the hyperparameter k(number of clusters) as 2. Apart from this, we can utilize the elbow curve, which is essentially the withiness sum of each cluster for a range of hyperparameters to conclude the optimal cluster number.

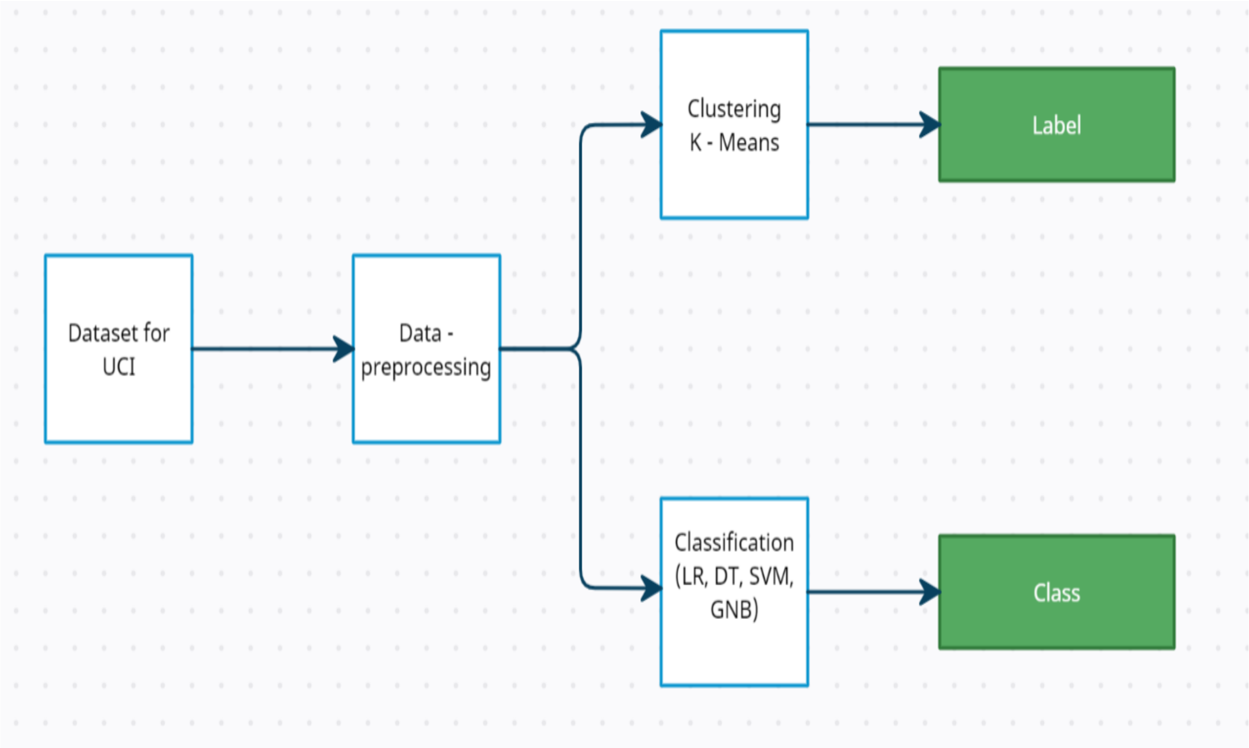


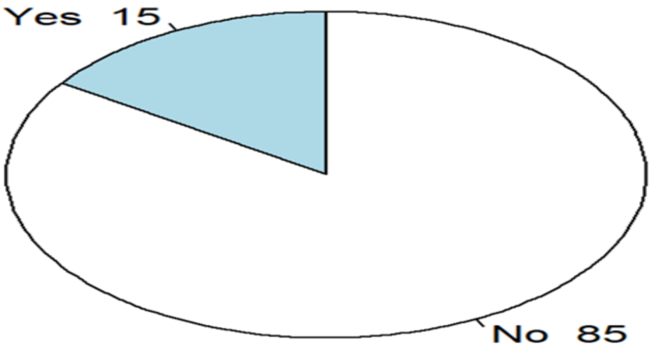
Figure-X: System Design

For classification of the users as revenue generating or non-revenue generating, we employ binary classification techniques such as Logistic Regression, Support Vector Machine, Decision Trees, Ensemble Methods such as Random Forest, and ensemble-based models containing SVM and logistic models. To extend the logistic regression we used boosting algorithms for better performance. This includes Xdgboost, and Adaboost, this can also be included in the ensemble models for experimenting for better accuracy of the classifiers.

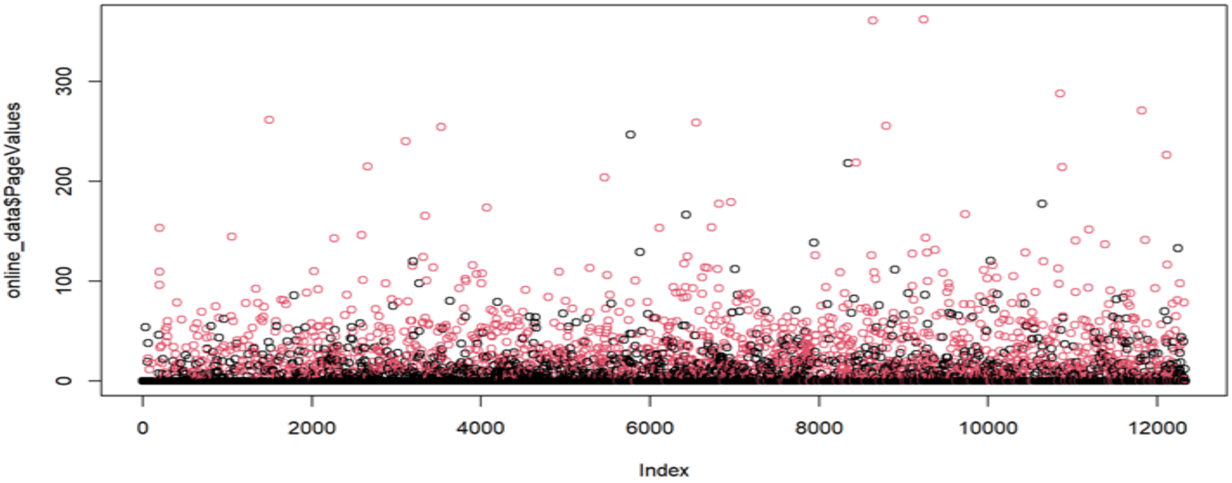
Results and Discussion

Feature Statistics:

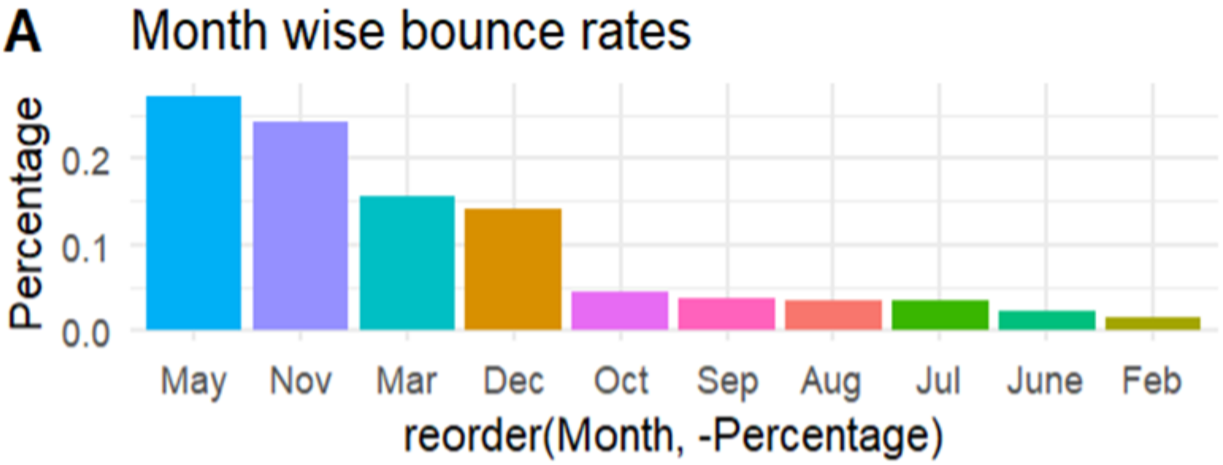
Feature statistics contain the relationship between various features and the target variable “Revenue”. It also contains the distribution and outlier’s plot.



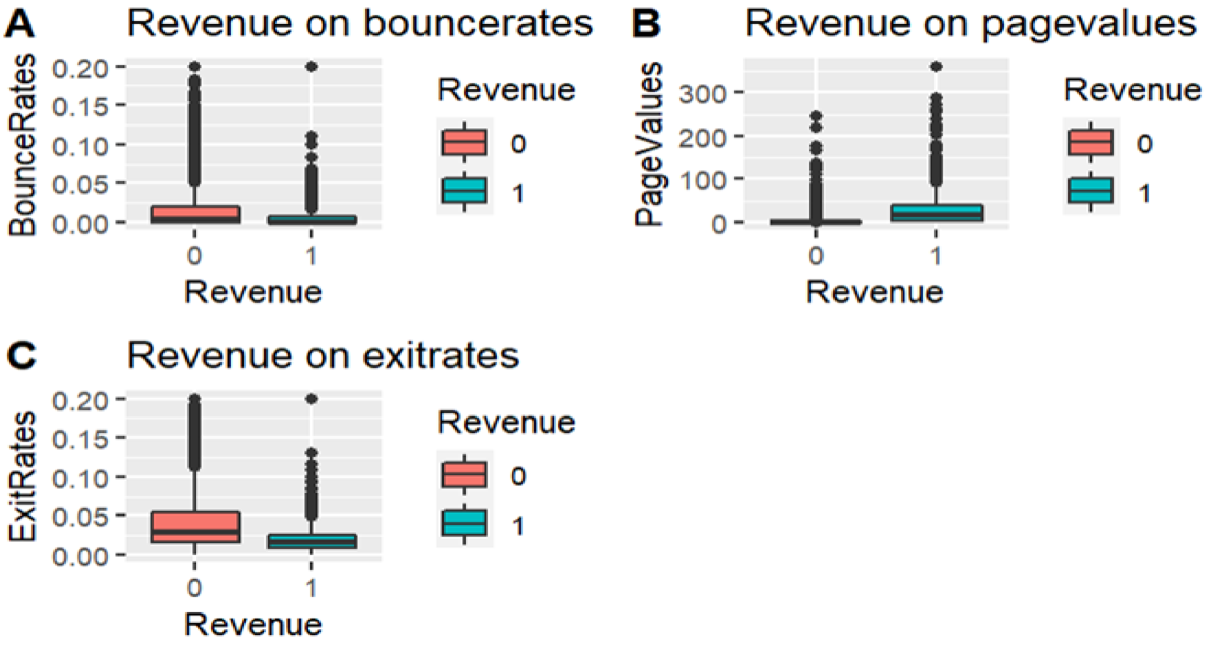
The above pie chart shows the proportion of observations. The plot shows that there are 15% of observations where revenue is generated, whereas 85% of records have been recorded that revenue was not generated. But 85% of the majority class may lead to bias. To overcome this problem we performed downsampling while performing training of data.



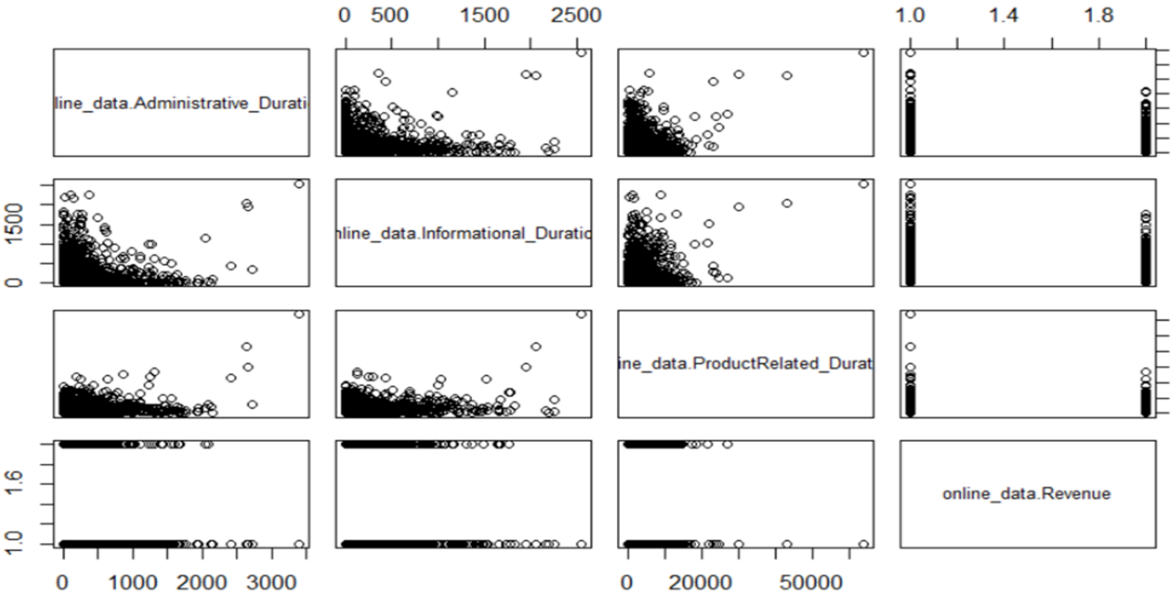
The above dot plot just represents the distribution of page values. Page value the average value for a page visited by a visitor before arriving at the goal page or completing an Ecommerce transaction (or both). Page values follow Skewed distribution, there are a lot of pages with very less page values, and very few pages with page values greater than 100.



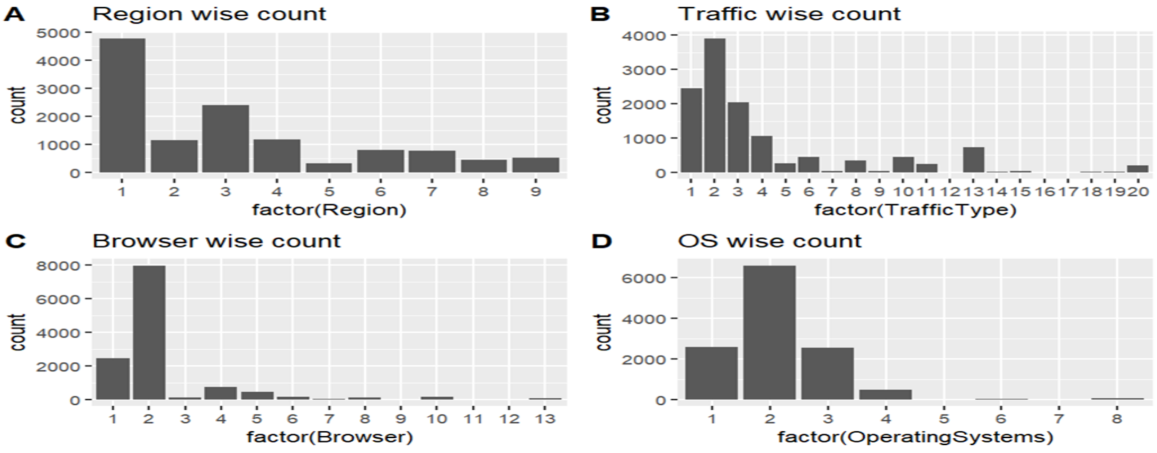
The above bar plot shows the comparison of bounce rates by month. Bounce rate is the percentage where a user visits the website and leaves the website by viewing only a single page. The bounce rates in May (27.3%) and November (24.3%) are very high. The percentage rates are increasing from June to November. There is a dwindle in February (1.50%). We can infer that the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests is more in May and November.



Boxplots are a way to see how well a data set's data is distributed. In this particular plot, we analyzed the revenue based on bounce rates, page values, and exit rates. There are many outliers in each attribute, which tells us that while predicting the data we have to either filter the data or, remove (or) change the data to maintain the accuracy of the predicted value.



The above pair plot is used to figure out which attributes are best for explaining a relationship between two variables or for forming the most separated groups. Drawing some simple lines or making a linear separation in our data-set also aids in the formation of some simple classification models. From the matrix, we can observe that as the informational duration increases, the Administrative Duration Decreases. This indicates a negative correlation similarly for Product-Related Duration and Informational Duration.

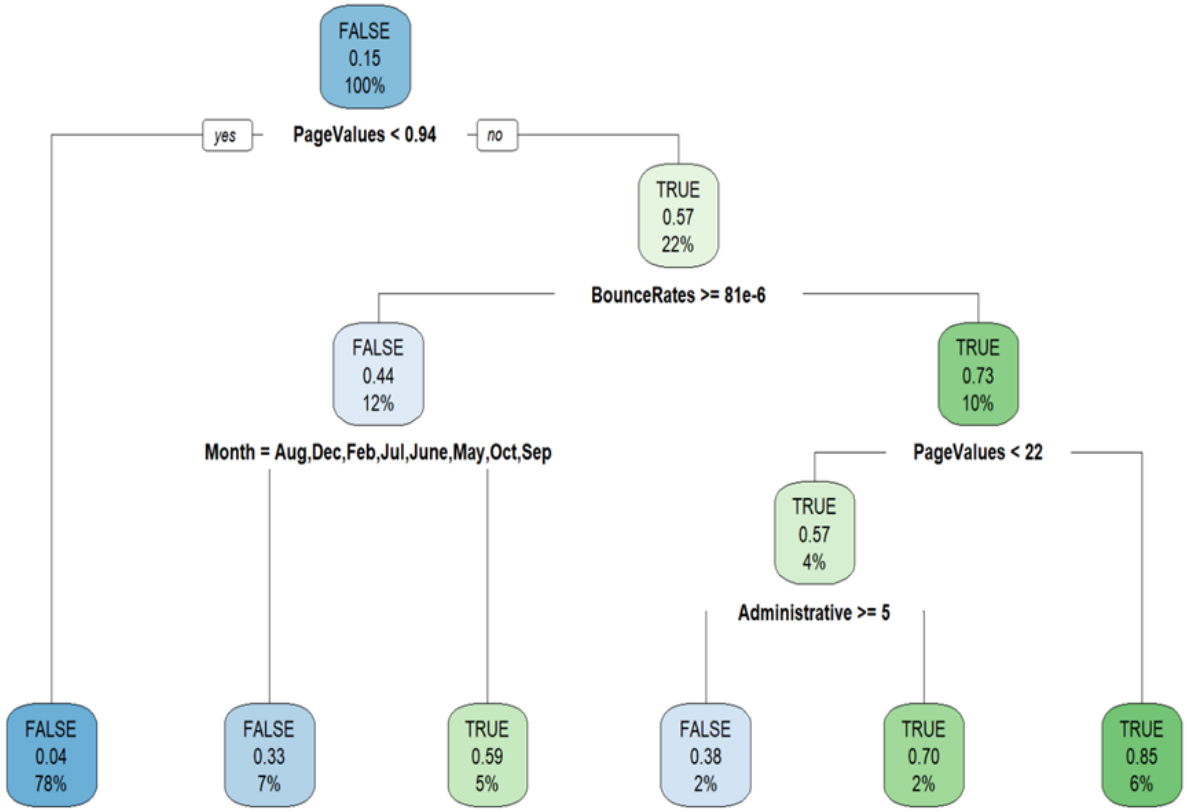


A count plot is a histogram across a category variable rather than a quantitative variable. It's used to employ bars to show the number of observations in each categorical division. From the above plot, we can infer that the 2nd type of browser, 2nd type of OS, 2nd type of website where it has high traffic had the highest number of distributions in the data.

Classification of Various Algorithms:

For predicting the values and accuracy the techniques used are decision tree classifier, support vector classification, Logistic Regression, and Naive Bayes classification.

A Decision Tree is a simple representation of classifying illustrations. It may be Administered through Machine Learning where the information is persistently part agreeing to a certain parameter. In this strategy, a set of preparing cases is broken down into smaller and smaller subsets whereas at the same time a related decision tree gets incrementally created. At the conclusion of the learning handle, a choice tree covering the preparing set is returned. The key thought is to utilize a choice tree to parcel the information space into cluster (or thick) districts and empty (or inadequate) regions.

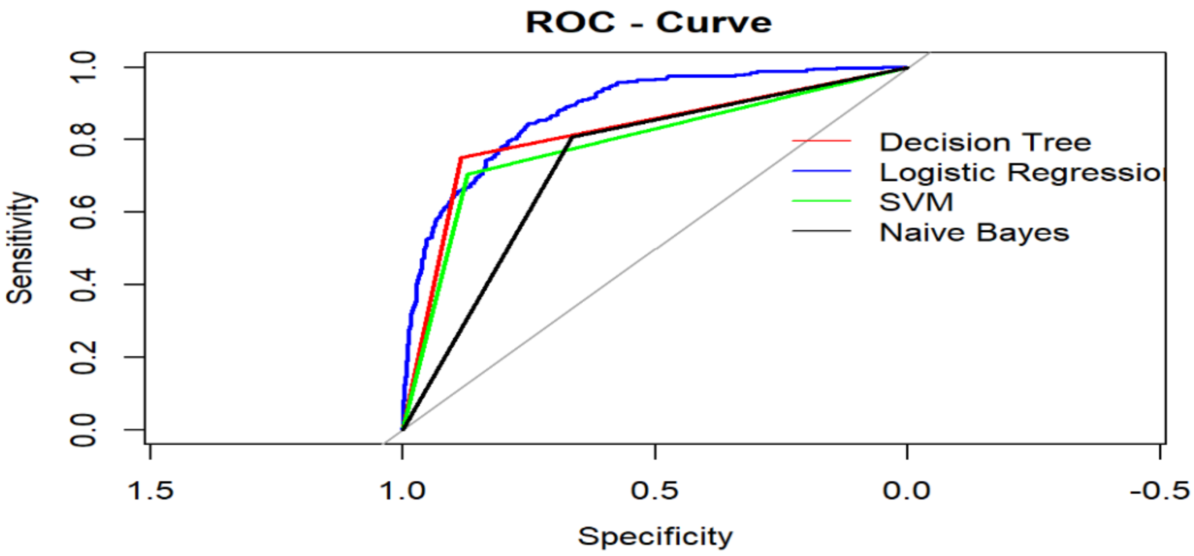


From the implementation of Decision Tree, we found out that PageValue is the most important parameter, so if a user is viewing a page with a page value lower than 0.94 there is a 78% chance of not revenue. If this page value is search engine-based, organizations can implement ways to increase this feature.

The Support Vector Machine, or SVM, is a popular Supervised Learning technique that may be used to solve both classification and regression issues. However, it is mostly utilized in Machine Learning for Classification difficulties. The SVM algorithm's purpose is to find the optimum line or decision boundary for categorizing n-dimensional space into classes so that additional data points can be readily placed in the correct category in the future. A hyperplane is a name for the optimal choice boundary.

The categorical dependent variable is predicted using logistic regression utilizing a set of independent variables. A categorical dependent variable's output is predicted using logistic regression. As a result, the result must be a discrete or categorical value. It can be Yes or No, 0 or 1, true or false, and so on, but instead of giving exact values like 0 and 1, it delivers probabilistic values that are somewhere between 0 and 1.

The Naive Bayes Classifier is a simple and effective classification method that aids in the development of fast machine learning models capable of making quick predictions. It's a probabilistic classifier, which means it makes predictions based on an object's probability.



Model	Accuracy	AUC	F1-Score
Decision Tree	87.3	0.831	0.901
Support Vector	91.7	0.919	0.943
Logistic Regression	84.1	0.755	0.555
Naive Bayes	66.4	0.824	0.762

The algorithms can be compared based on the F1 Score and AUC(Area Under Curve). F1 Score is the harmonic mean of Precision and Recall. We have observed that on a data split of 80%, the support vector has the highest testing accuracy of 91%. Based on Both AUC and F1 Scores, the support vector model has outperformed the other models. The data for the support vector and Decision Trees are transformed accordingly. The categorical data is labeled encoded. Similarly, for algorithms, we have eliminated some outliers. To uncover patterns and trends in the data set under investigation, we employed data sampling to select, alter, and analyze a representative selection of data points.

Conclusion

E-commerce websites have been facilitating online shopping in a better and enhanced security process. Users are utilizing the websites in various ways apart from buying a product such as comparing and contrasting the prices in distinct e-commerce websites. In our work, we analyzed the customer intentions using the data. Utilizing the average time that a user spends on a particular application, bounce rates and using other different types of attributes like the kind of operating system the user was using we predict whether the revenue is generated or not by different kinds of classification algorithms.

References

[1] Radhakrishnan, V., 2021. E-Commerce industry and its effect on the world today. International Research Journal on Advanced Science Hub, 3(2), pp.23-29.

[2] Suchacka G, Skolimowska-Kulig M, Potempa A (2015) A k-nearest neighbors method for classifying user sessions in e-commerce scenarios. J Telecommun Inf Technol 3:64

[3] Wang, Y.T. and Lee, A.J., 2011. Mining Web navigation patterns with a path traversal graph. Expert Systems with Applications, 38(6), pp.7112-7122.

[4]Koufaris, M., 2002. Applying the technology acceptance model and flow theory to online consumer behavior. Information systems research, 13(2), pp.205-223.

[5]Kasuma, J., Kanyan, A., Khairol, M., Sa'ait, N. and Paint, G., 2020. Factors influencing customers' intention for online shopping. International Journal of Modern Trends in Business Research, 3(11), pp.31-41.

[6]Wei, L., 2016. Decision-making Behaviors toward online shopping. International Journal of Marketing Studies, 8(3), pp.111-121.

Appendix

Down sampling Code:

```
ds_x<-sample(1:nrow(train[train$Revenue==0,]),1530)
train_p=train[train$Revenue==1,]
train_f=train[train$Revenue==0,][ds_x,]
train=rbind(train_f,train_p)
nrow(train[train$Revenue==0,])
nrow(train[train$Revenue==1,])
```

Decision Tree code:

```
train$VisitorType <- as.factor(train$VisitorType)
train$Month <- as.factor(train$Month)
mod <- rpart(Revenue~., data = train, method = 'class')
library(party)
```

```
modeldc <- ctree(Revenue ~
PageValues+Month+Administrative+ExitRates+VisitorType+OperatingSystems+TrafficType+B
ounceRates,train)
plot(modeldc)
```

Support Vector Machine (SVM):

```
library(e1071)
classifier = svm(formula = Revenue ~
Administrative+Administrative_Duration+Informational+Informational+Informational_Duration
+ProductRelated+ProductRelated_Duration+BounceRates+ExitRates+Weekend+VisitorType+Tr
afficType+Browser+Region+OperatingSystems+Month+SpecialDay+PageValues,data = train,
            type = 'C-classification',
            kernel = 'linear')
```

Logistic Regression Code:

```
logmodel <- glm(Revenue ~ ., data = train, family = "binomial")
logmodel1 <- glm(Revenue ~ Month+ PageValues + ExitRates + OperatingSystems + Region +
Administrative+BounceRates+SpecialDay, data = train, family = "binomial")
predictlogmo <- predict(logmodel1, test, type = "response")
logcm = table(test$Revenue, predictlogmo)
alogcm_Test <- sum(diag(logcm)) / sum(logcm)
print(paste('Accuracy for test', alogcm_Test))
```

Naive Bayes Code:

```
naivemodel <- naiveBayes(Revenue ~ .,data =train)
naive_pred <- predict(naivemodel, newdata = test)
summary(naivemodel)
naivecm = table(test$Revenue, naive_pred)
naivecm_Test <- sum(diag(naivecm)) / sum(naivecm)
print(paste('Accuracy for test', naivecm_Test))
```