MANTECH Publications
Journal of Research in Computer Science and Engineering
Volume 7 Issue 3
ISSN- 2457-0818 (Online)

Analysis of Customer Behavior on Ecommerce Websites

Shiv Shankar Prasad[1], Mahesh Rai[2], Roshani Kumari[3]
Assistant Professor[1], Students[2, 3]
Department of Computer Science Engineering
Sethu Institute of Technology, Kariapatti
Corresponding Author's E-mail:- *maheshrai95@gmail.com[2]*

Abstract

The current study attempts to analyse and anticipate online buyers' human behaviour. The goal of this research is to give a review of recent work on consumer purchase forecasting in the context of e-commerce. This article is intended to assist in the establishment of small-scale websites. The primary contribution is a conceptual study framework that methodically maps the available knowledge into two key goals: purchasing session prediction and purchase decisions.

Keywords: *Clustering, feature selection, Customer behaviour.*

INTRODUCTION

Consumer behaviour is a fast growing area of study. It is a broad term for investigating the consumer's motives for picking a product that matches their desires or objectives. Consumer behaviour is referred to as "dynamic interaction" by the American Marketing Association. Influence and perception, behaviours, and the environment in which individuals share aspects of their lives: in order to compete in a competitive marketing climate, marketers must learn and foresee everything they can about their clients. Customers may make the greatest product selection at any time thanks to the ease of online shopping. The report includes a viewpoint on the product qualities that motivate individuals to buy online.

Because of the growing popularity of e-commerce, people no longer need to go shopping for days. An individual would frequently spend hours at home browsing things for sale on the internet that he or she is not interested in acquiring. Such

Page 161-170 © MANTECH PUBLICATIONS 2022. All Rights Reserved

purchase habits rocket the internet shopping sector to a considerably more profitable position.

Customer Behaviour

Buyer activity is the investigation of people, gatherings, or organisations into their method of selecting, acquiring, utilising, and organising things, administrations, gatherings, or thoughts to fit requirements, as well as the consequences of this process on the shopper and the general public.

When shopping, a client might act in a variety of ways. No two clients are same in terms of attitude, decision-making, and intellectual capacity. Personal, psychological, social, and cultural aspects can all impact a buyer's decision. A number of other factors impact customer purchase behaviour. Only a few of these are practical:

- Financial conditions of the Consumer
- Desire of the customer to purchase
- Need for the Consumer

Along with these concerns, there is also concern that the client would disclose critical information online. Many people believe it is delicate and dangerous. Despite these behavioural barriers, millions of consumers make internet purchases every day. The purpose of this study is to investigate and analyse the actions of people who visit and spend time on online shopping sites. This will also take into account the number of people there and how many of them are carrying cash.

Customer Behaviour Influencing Market Strategies

Every day, the firm improves its competitiveness. Companies invest a lot of money in developing the most successful method for selling their products. This tasked advertising with analysing altering client purchase behaviours and developing new opportunities within another rapidly expanding media. According to web-based surveys, the Internet is changing the way people utilise shopping platforms to make purchases.

Existing Systems

E-retailers undertake questionnaires and consumer surveys to better understand their clients' purchasing habits, but they do not obtain honest findings since they are unable to foresee likely client spending patterns in online shopping. As a result, e-commerce retailers are struggling to attract customers. Customers usually browse the products before leaving the website. E-retailers struggle to forecast whether or not

Page 161-170 © MANTECH PUBLICATIONS 2022. All Rights Reserved

a consumer would purchase a product, and thus fight to build client confidence and loyalty, while consumers typically struggle to choose desired things.

ANALYSIS

Based on the qualities of our dataset, we provide a prediction method for analysing human behaviour in online purchasing based on decision-making styles. Three algorithms are used to classify the dataset: XGBoost, logistic regression, and Light GBM (Light Gradient Boosting Machine), which predict consumer behaviour and help e-retailers identify their potential customers.

Dataset Information

The dataset includes function vectors from 12,330 sessions. To avoid any inclination to a specific campaign, special day, user profile, or timeframe, the dataset was created so that each session in a 1-year cycle belongs to a distinct user. We are utilising a dataset obtained online; however, adjustments must be made.

Attribute Information

The dataset has ten number characteristics and eight categories attributes. The property 'Revenue' can be used as the class name. "Administrative," "Administrative Duration," "Informational," "Informational Duration," "Product Related," and "Product Related Duration" show the number of various types of pages seen by the visitor during that session, as well as the total time spent on each of these page categories. The values of these characteristics are derived from the URL information of the sites visited by the user and are adjusted in real time when the user performs an action, such as moving from one page to another.

The 'Bounce Rate,' 'Exit Rate,' and 'Page Value' elements on the e-commerce platform represent the numbers computed by 'Google Analytics' for each page. The "Bounce Rate" function on a web page refers to the proportion of users that reach the site from that page and then leave ("bounce") within that session without prompting any more queries from the analytics server. The value of the "Exit Rate" function for a certain web page is determined, with the percentage being the final page visit in the session.

The "Page Value" field represents the average value of a web page visited by a user prior to the e-commerce transaction. was finished The "Special Day" function displays that the site visit time is closer to a certain special day (for example, Mother's Day or Valentine's Day), when

Page 161-170 © MANTECH PUBLICATIONS 2022. All Rights Reserved

sessions are more likely to end with a purchase. The relevance of this feature is assessed by taking into consideration the complications of e-commerce, such as the delay between the order date and the delivery date.

For example, between February 2 and February 12, this value has a non-zero value, a zero value before and after that day unless it is identical to another special day, and a maximum value of 1 on February 8. The dataset also includes the operating system, the browser, the location, the kind of traffic, the type of visitor (return or new), the weekend, and the month of the year as the date of the visit.

Feature Selection

The process of picking a subset of acceptable characteristics to be used in machine learning to form the model is known as feature selection. Efficient feature selection eliminates redundant variables and keeps only the best subset of predictors in the model, resulting in reduced training periods. Feature extraction techniques, such as function extraction techniques such as PCA, are examples of feature selection strategies that may be used for model creation.

Clustering

In order to identify diverse patterns among online consumers, we employ the elbow methodology to determine the total number of ideal clusters. The elbow technique in cluster analysis is a heuristic way of evaluating and assessing accuracy aimed at assisting in identifying the appropriate number of clusters in a dataset. The elbow approach computes an average score for all clusters for each k value after running a cluster of k-means on the dataset for a number of k-values (say 1–10). Design determines the distortion ranking, which is the number of square distances from each point to its allocated centre. When these overall measures for each model are shown, the optimal value for k can be visually estimated. If the line chart resembles an arm, the optimal value of k is "elbow" (the inflection point on the curve). If there is a significant inflection point, the "arm" may be up or down, but it is a clear signal that the underlying model suits better at that moment. This approach employs the K-means clustering algorithm. K-means is a basic unsupervised machine learning technique that divides data into k clusters. The approach is relatively naïve in that the user must choose k in advance, allocating all members to k clusters even if it is not the proper k for the dataset.

Page 161-170 © MANTECH PUBLICATIONS 2022. All Rights Reserved

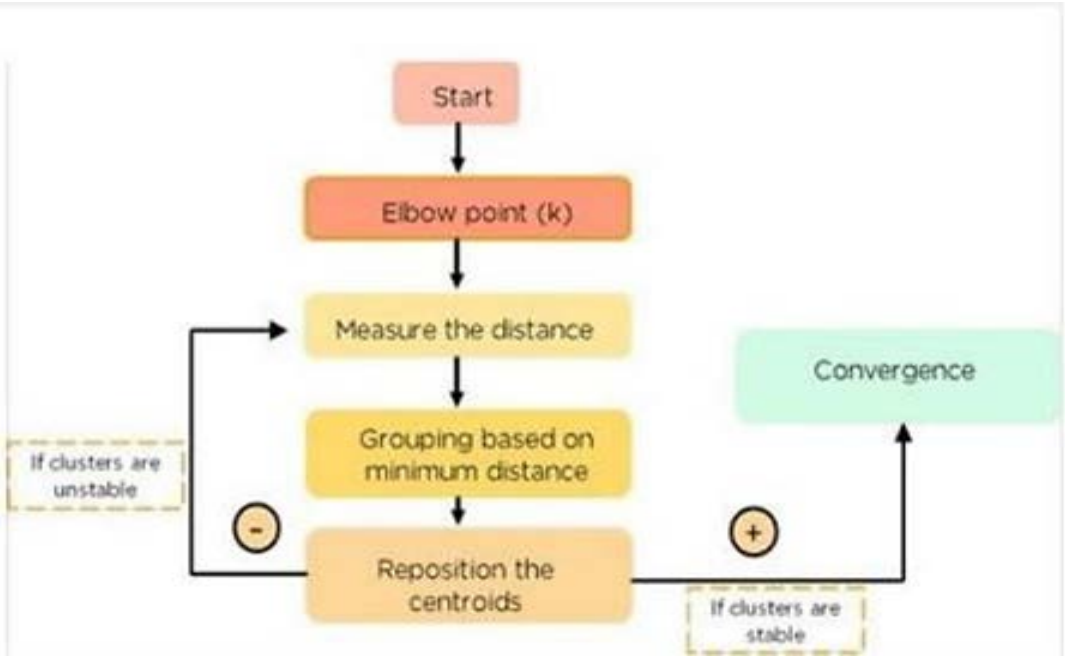


Fig.1:-Clustering Process

IMPLEMENTATION METHODOLOGY

Splitting Data

To ensure that the models function properly and do not overfit, the dataset is divided into training and testing sets. The training set will be used to build the model, while the test set will be used to check that it performs properly and can be adapted to new knowledge. Several classification models, like logistic regression, XGBoost, and Light GBM, were looked at to see if they could be used to predict which consumer will buy something.

Max Min Scaling

Min-max scaling is also referred to as min-max normalization, is the easiest approach and consists of rescaling the number of features to scale the range to [0, 1] or [−1, 1]. It is dependent on the quality of the data to select the target range. The formula, in this case, is:

m = (x -xmin) / (xmax -xmin) Where,

m is our new value

x is the original cell value

xmin is the minimum value of the column

xmax is the maximum value of the column

Using this formula, we will see that the values of each

Column will now be between zero and one.

Logistic Regression

When the dependent variable is binary, logistic regression is the best regression methodology to use (binary). Logistic regression is a predictive analysis, as are other regression studies. Logistic regression is used to categorise data and

Page 161-170 © MANTECH PUBLICATIONS 2022. All Rights Reserved

demonstrate the link between one dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables.

It is often difficult to evaluate logistic regressions; the Intellects Statistics technique swiftly assists you in doing so. The result is then translated into plain English to do the analysis. The dependent variable in nature should be dichotomous (e.g., presence vs. absent).

Light Gradient Boosting

Light GBM is a fast, distributed, high-performance gradient boosting system based on the decision tree technique that is used for ranking, classification, and many other machine learning applications. Because it is based on decision tree algorithms, it splits the tree leaf-wise with the greatest fit, unlike other boosting methods that split the tree depth-wise or level-wise rather than leaf-wise.

When rising on the same leaf in light GBM, the leaf-wise approach may therefore minimise more losses than the level-wise technique, resulting in considerably superior accuracy that can seldom be attained by any of the existing boosting algorithms.

Advantages of Light GBM

1. Increased training speed and efficiency: Light GBM employs a histogram-based technique that buckets continuous function values into discrete bins to speed up the training process.

2. Reduced memory use: Replaces discrete bins with continuous values to save memory usage.

3. Better accuracy than any other boosting algorithm: By using a leaf-wise split technique rather than a level-wise approach, it builds considerably more complex trees, which is the fundamental component in obtaining higher accuracy. It can, however, frequently result in overfitting, which can be avoided by setting the option to max depth.

4. Big Dataset Compatibility: When compared to XGBOOST, it can perform similarly well with huge datasets while requiring significantly less training time.

XGBoost

XGBoost is an ensemble machine learning technique that employs a framework for gradient boosting and is based on a

Page 161-170 © MANTECH PUBLICATIONS 2022. All Rights Reserved

decision tree. It is also known as "intense gradient boosting." It is the best way to combine software and hardware optimization methods to get the best results in the least amount of time and with the least amount of computing resources.

The algorithm differentiates itself in the following ways:

1. A wide variety of applications can be used to overcome problems with regression, classification, ranking, and user-defined prediction.

2. Portability: Runs on Windows, Linux, and OS X smoothly.

3. Languages: Supports all major C++, Python, R, Java, Scala, and Julia programming languages.

4. Cloud Integration: Supports AWS, Azure, and Yarn clusters and works well with Flink, Spark, and other ecosystems.

Why is it most popular?

1. It is straightforward to carry out. In general, GB is regarded as one of the top ML algorithms out-of-the-box, because it typically performs quite well even with minimal customization.

2. It is quick. Modern libraries, like XGBoost, include considerable speed improvements, allowing a high-performing model to be trained in a short period of time.

3. It is an ensemble learning method that aggregates the predictions of numerous base learners for each input or example to get one overall prediction. This promotes the learning of more dynamic correlations between characteristics and targets or labels in the training set.

4. They learn in a systematic manner. With each iteration, a new base learner is added to rectify the mistakes of the preceding learner series.

5. It has various hyperparameters for performance adjustment, making it a very adaptable learner.

Hyper parameter Tuning

The difficulty of selecting a set of ideal hyper parameters for a learning algorithm is known as "hyper parameter tuning." A hyper parameter is a parameter whose value is used to control the learning process. Models can include several hyper parameters, and finding the optimal combination of parameters can be considered a search issue. Two of the most

Page 161-170 © MANTECH PUBLICATIONS 2022. All Rights Reserved

effective strategies for hyper parameter adjustment are:

1. GridSearchCV
2. RandomizedSearchCV

Confusion Matrix

A confusion matrix is a description of the effects of forecasting on a classification problem. By counting values, the number of correct and incorrect predictions is summarised and broken down by class. This is the key to the confusion matrix. The confusion matrix reveals the ways in which the classification model is confused when it makes predictions. It gives us information about not only the errors made by a classifier, but also the types of errors made.

Definition of the Terms

Positive (P): Observation is positive (for example: is an apple).

Negative (N): Observation is not positive (for example: is not an apple).

True Positive (TP): Observation is positive, and is predicted to be positive.

False Negative (FN): Observation is positive, but is predicted negative.

True Negative (TN): Observation is negative, and is predicted to be negative.

False Positive (FP): Observation is negative, but is predicted positive.

Classification Report

A classification algorithm is used to calculate the accuracy of predictions made by the algorithm. The study shows the accuracy, recall, and f1-score of the important classification metrics for each class. The metrics are generated by combining true and false positives and true and false negatives. Positive and negative are general labels for the projected groupings in this situation.

Classification Rate/Accuracy

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Recall

Recall gives us an idea about when it'sactually yes, how often does it predict yes. Recall = TP / (TP + FN)

Precision

Precision tells us about when it predicts yes, how often is it correct.

Precision = TP / (TP + FP)

E-Measure

F-measure=(2*Recall*Precision)/ (Recall+Precision)

Page 161-170 © MANTECH PUBLICATIONS 2022. All Rights Reserved

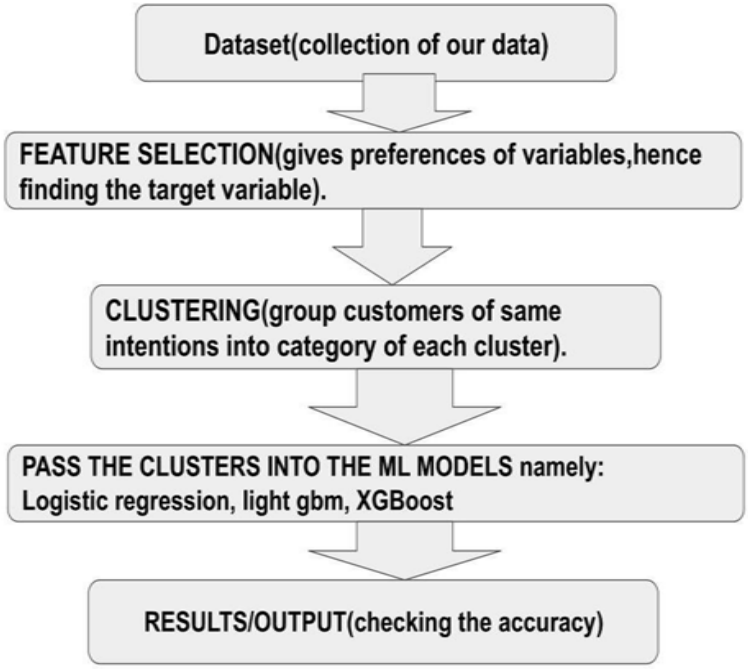


Fig.2:-Block diagram of Model

CONCLUSION

As previously said, the purpose of this study is to examine the actions of people who visit and spend time on online shopping sites. The accuracy gained will help small-scale websites function smoothly and financially. The management of these companies may take the required actions based on the inside knowledge obtained during our inquiry.

REFERENCES

1. Sun, T., Wang, M., & Liang, Z. (2017, December). Predictive modeling of potential customers based on the customers clickstream data: A field study. In 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 2221-2225).

2. Hoch, R. (2001). Visualization and analysis of clickstream data of online stores for understanding web merchandising. Data mining and knowledge discovery, 5(1-2), 59-84.

3. Tjhin V. Decision making in online purchase of movie ticket. International Seminar on Application for Technology of Information and Communication. 2016.

Page 161-170 © MANTECH PUBLICATIONS 2022. All Rights Reserved

4. Cervenka, P., Cognitive system in market analysis. October 2016

5. Tanksale, D., Neelam, N., &Venkatachalam, R. (2014). Consumer decision making styles of young adult consumers in India. Procedia-Social and Behavioral Sciences, 133, 211- 218.

6. Puspitasari, N. B., WP, S. N., Amyhorsea, D. N., &Susanty, A. (2018). Consumer's Buying Decision- Making Process in E-Commerce. In E3S Web of Conferences (Vol. 31,p. 11003). EDP Sciences.Lee, J., Podlaseck, M., Schonberg, E.

Page 161-170 © MANTECH PUBLICATIONS 2022. All Rights Reserved