HBRP PUBLICATION

Journal of Advance Research in Mobile Computing
Volume 3 Issue 1

Analysis of Consumer Purchase on Ecommerce

Sushama Khanvilkar[1], Maithili Salankar[2*], Ayesha Shetty[3], Vincia Marian[4]
[1]Professor, [2,3,4]Student
Computer Department, Xavier Institute of Engineering, Mumbai, Maharashtra, India

Corresponding Author
E-mail id:-*maithilissalankar@gmail.com*

ABSTRACT
The display thinks about points at dissecting and anticipating the human conduct of online shopping clients. This paper presents a scholarly audit of later inquiries about on client buy determining within the setting of e-commerce. This paper points to advantage websites that have begun on a little scale. The key commitment may be a conceptual investigation system that methodically maps this existing writing into three primary assignments, such as forecast of acquiring sessions, acquiring choices.

Keywords:-*Customer behaviour, feature selection, smote*

INTRODUCTION

Customer Behavior is an progressively developing range of consider. It may be a broader term that ponders the consumer's reasons for choosing the item that suits their needs or wants. Customer conduct is portrayed by the American Showcasing Affiliation as "Energetic Interaction".

The impact and discernment, activities, and the world in which human creatures share perspectives of their lives. Marketers have an pressing need to get it and foresee anything they can approximately clients in arrange to flourish in a competitive promoting environment.
The online shopping mode has made it simple and simple for the client to create the proper item choice at any time. Within the investigation, the viewpoint on the item measurements that cause clients to purchase online is talked about.

Presently with a quick rise in e-commerce, individuals don't have to go shopping anyplace for days. A person regularly spends hours and hours at domestic fairs checking out the things he or she isn't fascinated by buying for deals over the web. Such customer shopping inclinations astoundingly carry the online shopping industry to a distant more beneficial put.

CUSTOMER BEHAVIOR

- Buyer movement is the examination of people, gatherings or associations with respect to their strategy of selecting, securing, utilizing and sorting out items, organizations, gatherings or contemplations to meet necessities and the impacts on the customer and the common open of this process. While shopping, a client can act in a few diverse ways. Each sort of client can never be expected to be the same in state of mind, decision-making and mental terms. Individual, mental and social and social variables may impact a buyer. There are some other components that influence the shopping penchants of clients. Few of these are conceivable:
- Money related conditions of the Shopper
- Need of the client to buy
- Require for the Customer

There's too a concern of the client sharing

his/her personal data online together with these factors. A parcel of individuals discovers it helpless and unsafe. Given these behavioral challenges, millions and millions of people shop online each day. The reason of this paper is to look at the activities of people going by online shopping sites and investing their time there, analyzing different things. That will too take into consideration how numerous individuals are there and how numerous of them are their cash.

CUSTOMER BEHAVIOR INFLUENCING MARKET STRATEGIES

The trade is getting exceptionally competitive each day presently. Organizations exhaust a parcel of cash on arranging the foremost effective way to advertise their merchandise. This tried promoters to dismember changing shopper shopping conduct and to construct novel choices inside another, quickly making medium. Web-based research appears that the Internet is changing the way individuals utilize shopping stages to buy products.

EXISTING SYSTEMS

E-retailers perform questionnaires and consumer surveys to understand their purchasing habits, but they do not get truthful responses because e-retailers face difficulties in predicting potential customer buying patterns in online shopping. E-retailers, therefore, find it hard to attract customers. The customers often only view the product and then leave the web. It is difficult for e-retailers to analyze whether or not the consumer will buy the product and they face a challenge in building customer confidence and loyalty, and consumers often face a problem in choosing their desirable items.

ANALYSIS

We suggest a prediction system based on our dataset attributes to test human behaviour in online shopping according to decision-making styles. The dataset is classified using three algorithms: XGBoost, logistic regression, and Light GBM (Light Gradient Boosting Machine), which predict consumer behaviour and assist e-retailers in understanding their potential customers.

Dataset Information

Function vectors from 12,330 separate sessions are included in the dataset. The dataset was built in such a way that each session belongs to a different user over the course of a year to prevent any tendency to a particular campaign, special day, user profile, or period.

Attribute Information

The dataset has ten numerical attributes and eight categorical attributes. It is possible to use the class name 'Revenue' as an attribute. The following are names of the attributes and their purpose:

Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration - refer to the number of different types of pages that the user accessed during that session, as well as the total time spent on each of these page categories. These features' values are extracted from the URL details of the user's visited sites and are modified in real time when the user performs an action, such as switching from one page to another.

The metrics calculated by 'Google Analytics' for each page are represented by the e-commerce platform's Bounce Rate, Exit Rate and Page Value features.

Bounce Rate - refers to the percentage of visitors who arrive at the site from that page and then leave ("bounce") without making any additional requests to the analytics server during that session.

Exit Rate - The value of it is measured in the same way as it is for all page views on that page, with the percentage representing the session's last page view.

Page Value - It reflects the average value of a web page reached by a customer before an e-commerce transaction is completed.

Special Day - indicates that site visits closer to a particular special day (e.g., Father's Day, Valentine's Day) are more likely to result in a purchase. The importance of this attribute is calculated by taking into account e-commerce complexities such as the time between placing an order and receiving it. For Valentine's Day, this value is non-zero between February 2 and February 12, zero before and after that date unless it is close to another special day, and one on February 8.

The dataset also includes the operating system, browser, location, type of traffic, type of visitor returned or new, the weekend Boolean value, and the month of the year as the date of the visit.

Exploratory Data Analysis

Exploratory data analysis is a method of analyzing sets of information to summarize their main characteristics, often using visualization strategies. There are numerous complex representations that can be created with pandas, and there is usually no need to import additional libraries.

EDA is a data analysis phenomenon that is used to achieve a deeper understanding of data aspects such as:
-main features of data
-variables and relationships that hold between them
-determining which variables are essential to our problem

Conversion of Categorical to Numerical Values

With numerical variables, the majority of algorithms achieve better end results. The python library "sklearn" needs numerical array capabilities. To make it even easier to expect, we convert the expressed variable to a numeric value. For example, if we have two training variables, we can convert 'yes' to one and 'no' to zero. Variables with a finite set of label values are referred to as categorical data. The majority of gadget learning algorithms necessitate numerical input and output variables.

Feature Selection

The process of selecting a subset of suitable attributes to be used in machine learning to create the model is known as feature selection. Efficient feature selection eliminates redundant variables and keeps only the best subset of predictors in the model, which often have shorter training times. One of the options for feature selection that can be done for model design purposes is feature extraction techniques such as function extraction techniques such as PCA.

IMPLEMENTATION METHODOLOGY

Splitting Data

The dataset is divided into training and testing sets so that the models perform well and don't overfit. To build the model, the training set will be utilized, whereas the test set will be utilized to approve that it works well and can be generalized to new information. The appropriateness for the task of predicting which buyer would make a buy was assessed by the number of different classification models such as logistic regression, XGBoost and Light gbm.

Max Min Scaling

Min-max scaling is additionally referred to as min-max normalization, is the most

straightforward approach and comprises of rescaling the number of features to scale the extend to [0, 1] or [−1, 1]. It is dependent on the quality of the data to choose the target run. The formula is:

$m = (x - x_{min}) / (x_{max} - x_{min})$

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

Logistic regression is a regression analysis suitable to perform when the dependent variable is dichotomous (binary). Logistic regression is a statistical analysis, as are all regression analyses. It is used to display the characteristics of data and the relationships between one dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables.

Light Gradient Boosting

Light GBM is a high-performance distributed fast gradient boosting system based on decision tree algorithm, which can be used for ranking, classification and various other machine learning tasks. Since it is based on the decision tree algorithm, it divides the tree into leaves in the most suitable way, while other boosting algorithms divide the tree into depth or level instead of leaves. While expanding on a similar leaf in Light GBM, the leaf-wise algorithm would thus be able to decrease a bigger number of losses than the level-wise algorithm, bringing about much better exactness that can seldom be accomplished by any of the current boosting algorithm.

Advantages of Light GBM

1. Increased training speed and efficiency: Light GBM uses a histogram-based algorithm that stores values of continuous functions in discrete bins to speed up the training process.
2. Lower memory usage: By replacing discrete bins with continuous values, the memory usage is reduced.
3. Better than any other boosting algorithm in terms of accuracy: Compared with the level-wise approach, the tree generated by the leaf wise split approach is much more complicated, which is a key factor to obtain higher accuracy. However, it can frequently result in overfitting, which can be avoided by setting the parameter max depth.
4. Large data set support: Compared with XGBOOST, it can handle large data sets while greatly reducing training time.

XGBoost

XGBoost is an ensemble Machine Learning algorithm that uses a gradient boosting structure and is based on a decision tree. It's also known as Intense Gradient Boosting. It's a great blend of software and hardware optimization techniques that produce superior results in the shortest amount of time while using less computing resources.

The algorithm differentiates itself in the following ways:

1. To solve problems with regression, grouping, ranking, and user-defined prediction, a wide range of applications can be used.
2. Portability: Runs smoothly on Windows, Linux, and OS X.
3. Languages: All major C++, Python, R, Java, Scala, and Julia programming languages are supported.
4. Cloud Integration: Supports clusters on AWS, Azure, and Yarn, as well as Flink, Spark, and other habitats.

Popularity of XGBoost

1. It is simple to execute. In general, GB

is considered to be one of the best ML algorithms out-of-the-box, because even with limited tuning, it usually performs very well.
2. It is quick. Many speed enhancements are included in modern libraries like XGBoost, enabling a high-performing model to be trained in a short amount of time.
3. It's an ensemble learning algorithm that integrates the predictions of multiple base learners for each input/example to produce a single overall prediction. This makes it easier to learn more complex relationships between features and targets/labels in the training collection.
4. They learn in a logical sequence. With each iteration, it creates a new base learner to correct the errors of the previous learner sequence.
5. It has a lot of hyper-parameters for performance tuning, so it's a very capable of adapting learner.

Hyperparameter Tuning

The problem of selecting a collection of suitable hyperparameters for a learning algorithm is known as hyperparameter tuning. A hyperparameter is a variable whose value is used to control the learning process.

The model can have multiple hyperparameters, and finding the correct combination of parameters can be seen as a search problem. Two of Hyperparameter tuning's best methods are:

1. GridSearchCV
2. RandomizedSearchCV

SMOTE

Minority data is copied from the minority data population in the classical oversampling technique. Although the amount of data increases, it does not provide new information or changes to the machine learning model.

It stands for Synthetic Minority Over-sampling Technique. SMOTE works with the aid of using utilising a k-nearest neighbor set of rules to create artificial information. SMOTE first begin with the aid of using deciding on random information from the minority class, then k-nearest pals from the information are set. Synthetic information might then made among the random information and the randomly decided on k-nearest neighbor.

Confusion Matrix

A confusion matrix is a diagram that depicts the effects of forecasting on a classification problem. By counting values, the number of accurate and incorrect predictions is totaled and broken down by class. The confusion matrix shows how the classification model is confused when making predictions. It not only allows us to understand the types of errors made by the classifier, but also allows us to understand the number of errors made by the classifier.

Term Definition

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 measure the accuracy of predictions made by a classification algorithm. On a per-class basis, the analysis shows the precision, recall, and f1-score of the main classification metrics. The metrics are calculated using true and false positives, as well as true and false negatives. Positive and negative are generic names for the expected classes in this situation.

Classification Rate/Accuracy:
Accuracy = (TP + TN) / (TP + TN + FP + FN)

Recall:
Recall gives us an idea of how much it predicts yes and when it is really yes.
Recall = TP / (TP + FN)

Precision:
Precision tells us how much it is accurate when it predicts yes.
Precision = TP / (TP + FP)

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

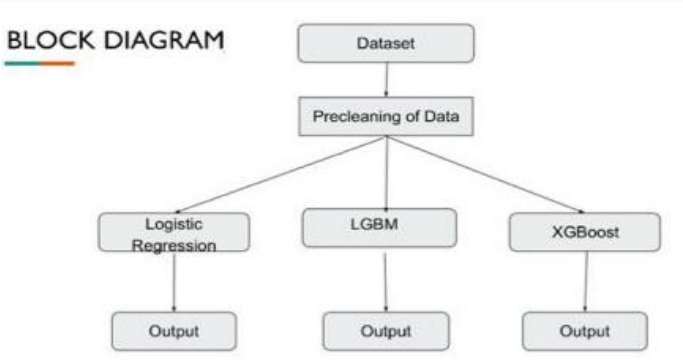

Fig.1:-*Model Flowchart*

IMPLEMENTATION RESULT
Logistic Regression
Without Smote

	precision	recall	f1-score	support
0	0.88	0.97	0.92	1022
1	0.71	0.38	0.50	211
accuracy			0.87	1233
macro avg	0.80	0.68	0.71	1233
weighted avg	0.85	0.87	0.85	1233

With Smote

	precision	recall	f1-score	support
0	0.88	0.97	0.92	1022
1	0.71	0.38	0.50	211
accuracy			0.87	1233
macro avg	0.80	0.68	0.71	1233
weighted avg	0.85	0.87	0.85	1233

LGBM
Without Smote

	precision	recall	f1-score	support
0	0.90	0.98	0.94	1022
1	0.81	0.50	0.62	211
accuracy			0.89	1233
macro avg	0.86	0.74	0.78	1233
weighted avg	0.89	0.89	0.88	1233

With Smote

	precision	recall	f1-score	support
0	0.96	0.90	0.93	1022
1	0.62	0.81	0.70	211
accuracy			0.88	1233
macro avg	0.79	0.85	0.81	1233
weighted avg	0.90	0.88	0.89	1233

XGBOOST
Without Smote

	precision	recall	f1-score	support
0	0.92	0.96	0.94	1022
1	0.76	0.59	0.66	211
accuracy			0.90	1233
macro avg	0.84	0.78	0.80	1233
weighted avg	0.89	0.90	0.89	1233

With Smote

	precision	recall	f1-score	support
0	0.93	0.96	0.94	1022
1	0.76	0.64	0.70	211
accuracy			0.90	1233
macro avg	0.84	0.80	0.82	1233
weighted avg	0.90	0.90	0.90	1233

Clearly from above results we can conclude that LGBM using SMOTE is most reliable and giving better accuracy.

CONCLUSION

As examined over the reason of this paper is to look at the activities of people going by online shopping locales and investing their time there, dissecting different things. The precision gotten might offer assistance the little scale websites run easily and offer assistance them make profit. The administration of these associations can take the desired activities based on the inner parts extricated from our examination.

ACKNOWLEDGEMENT

Prof. Sushama Khanvilkar, our teacher, who led us in working on this subject and assisted us in conducting extensive research, deserves our sincere and heartfelt gratitude. We appreciate her sharing her experience with us and helping us with the writing of this document.

REFERENCES

1. 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.
2. Jain, D.(2018). Analysis of consumer behavior towards online shopping. *International Conference on Contemporary Innovations in Library Information Science, Social Science & Technology for Virtual World*.
3. 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). IEEE.
4. Tjhin, V.(2016). Decision making in online purchase of movie ticket. *International Seminar on Application for Technology of Information and*

Communication.

5. Cervenka, P.(2016). Cognitive system in market analysis.
6. 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.
7. Karimi, S., Papamichail, K. N., & Holland, C. P. (2013). Purchase decision processes in the internet age. In *Decision Support Systems III-Impact of Decision Support Systems for Global Environments* (pp. 57-66). Springer, Cham.
8. Singh, A. K., & Sailo, M. (2013). Consumer behavior in online shopping: a study of Aizawl. *International Journal of Business & Management Research*, *1*(3), 45-49.
9. Zhang, A., Zheng, M., Jiang, N., & Zhang, J. (2013, November). Culture and consumers' decision-making styles: An experimental study in individual-level. In *2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering* (Vol. 2, pp. 444-449). IEEE.
10. Lee, J., Podlaseck, M., Schonberg, E., & Hoch, R. (2001). Visualization and analysis of clickstream data of online stores for understanding web merchandising. *Data mining and knowledge discovery*, *5*(1), 59-84.