International Journal of Advance and Innovative Research
Volume 9, Issue 4 October - December 2022
ISSN 2394 - 7780

E-COMMERCE CUSTOMERS BUYING BEHAVIOR ANALYSIS USING PYTHON

[1]Subhash Chand and [2]Dr. A.K. Shukla and [3]Naresh Chandra
[1,3]Research Scholar and [2]Associate Professor, IFTM University, Moradabad

ABSTRACT

Any business have at least two components i.e. sellers and buyers. Sellers are the entity who are selling the products, services, etc. and buyers are the consumers who consume the products, services, etc. The buying behavior of the buyers are very dynamic in nature and very difficult to predict. Buyers may be classified as offline buyers and online buyers. Online buyers purchase the products/services using e-commerce websites and payment is also made using online mode. Know your customer (KYC) methods is used by the companies for digitally and accurately record of the on-line journey of each customer. Ecommerce companies use KYC to serve to customers in more efficiently and accurately. Statistical methods are also applied on this digitally recorded data to understand the customer behavior and change the business accordingly. Customer behavior is very broad in nature, which includes buying behavior, searching behavior, browsing behavior, payment behavior, writing feedback behavior, product liking and disliking behavior, etc.. Customer buying behavior is highly personal activity and depends on many factors/situation related to customer like gender, financial background, marital status, mental level, having kids or not, locality, age, etc.. Therefore, it is very difficult to predict the customer buying behavior generally. A dataset related to online buying is used to identify and analyze the customer behavior using the python.

Keywords: Sellers, Buyers, Behavior, E-commerce, KYC, Python.

INTRODUCTION

Prediction of customer buying behavior is a challenging task. Customer buying behavior depends on many factors like customer nature, financial status, age, gender, marital status, having kids or not, number of kids, discount available, festival session, favorite day, favorite month, party occasion, spouse birthday, kids birthday, etc.. According to their convenience customers set some day and some months as favorite day and favorite months and he/she try to buy more on these days or months. A study on customer buying behavior is carried out on a secondary data. This dataset contains more than one lakh thirty thousand records and more than thirty attributes. Initially many attributes contain null values which lateron treated by removing and fill by relevant values by using some methods available in python. Finally a dataset containing no null values is prepared and used for analysis of customer buying behavior.

A study on gender, marital status, number of children, customer level, favorite day, favorite month is carried out from different angles. The graphs plotted between these attributes shows valuable information's which definitely helps a company to improve their business and fulfill the customer's needs also. It is observed that customers having more children buy more, females buy more than males, gender wise customers buy more on their favorite day and months, etc.Customers having more children are an asset to the firm selling products on-line. This type of customers wants to avail more discounts on the products to save money for future.

DATA USED

A secondary dataset contains more than 1, 30,000 records and more than 30 attributes is taken from github.com website. This dataset is analyzed from different angles by using python. The attribute names are self exploratory i.e. ID is the customers unique number, GENDER is the customers gender, MARITAL_STATUS is the marital status of the customer i.e. single or married, LT used at the beginning of attributes name stands for Last Transaction and FT used at the beginning of attributes name stands for First Transaction, TOT used at the beginning of attributes name stands for Total, CUST used at the beginning of attributes name stands for Customers, AVG used at the beginning of attributes name stands for Average, FAV used at the beginning of attributes name stands for Favorite.A complete record of this dataset is given as follows:

Attribute Names	Values
ID	364259
GENDER	FEMALE
MARITAL_STATUS	MARRIED
NO_OF_CHILDREN	0
LT_OCCASION	0

LT_OFFER_APPLIED	1
LT_QUANTITY	2
LT_AMOUNT	752260
FT_OCCASION	0
FT_QUANTITY	2
FT_OFFER_APPLIED	1
FT_AMOUNT	313995
TOT_TRANSACTIONS_OVERALL	4
CUST_LEVEL_OVERALL	Medium
TOT_TRANSACTION_AMOUNT_OVERALL	1218512
TOT_QUANTITY_OVERALL	6
TOT_DISCOUNT_OVERALL	49028
AVG_TRANSACTION_AMOUNT_OVERALL	304628
AVG_QUANTITY_OVERALL	1
AVG_DISCOUNT_OVERALL	12257
TOT_BIRTHDAY_PRODUCT	0
PREMIUM_PRODUCT	6
TOTAL_ANNIVERSARY_PRODUCT	2
TOTAL_OCASSION_PRODUCT	1.0
FAV_DAY	7
FAV_MONTH	7
AGE	56.0
CITY	DEHRADUN
STATE	UTTARAKHAND
RECENCY	430

This dataset is related to a festival session and contains more attributes. Therefore, it can be used to analyzed from different angles to extract useful information's about customer's behavior. During this study the data is analyzed mainly with respect to gender, marital status, and number of child's, total transaction amount, favorite day and months of the customers. Later on, data can be analyzed w.r.t. age, product types, recency, city and state of the customers.

METHOD

The first step to analyze any dataset is the treatment of null values available in dataset. The different method are applied to handle null values like numeric values may be replaced by mean, mode or median value of the column/attribute of the dataset. This method is not applicable to the categorical values, therefore categorical null values may be replaced among the set of most frequent values appeared in particular column randomly. Initially this dataset contains many null values, which is treated by deleting some unwanted attributes containing null values and by replacing other column's null values by frequently appeared values. The following is the details of attributes containing null values. The percentage of null values is also calculated by using following code in python.

```
column_having_missing_value = df.columns [df.isna ().any ()].tolist ()

Total_missing_count = []

For i in column_having_missing_value:

total_missing_count.append (sum (pd.isnull (df [str (i)])))
```

International Journal of Advance and Innovative Research
Volume 9, Issue 4 October - December 2022
ISSN 2394 - 7780

missing_dict = {'Missing_value_column_name':column_having_missing_value,'Total_missing_Values': total_missing_count}

Print (missing_dict,'\n\n')

Missing_value_DF = pd.DataFrame (missing_dict)

Missing_value_DF ["Percentage of Missing Value"] = (Missing_value_DF ['Total_missing_Values'] / df.shape [0]) * 100

Print (Missing_value_DF.sort_values (by= ['Percentage of Missing Value'], ascending =False))

Missing_Value_Column_Name	**Total_missing_Values**	**Percentage ofMissing Value**
CHILD3GENDER	127267	96.11%
CHILD2GENDER	109981	83.05%
CHILD1GENDER	97444	73.59%
SPOUSEAGE	77588	58.59%
SPOUSE_DOB_DAYS_LEFT	77588	58.59%
ANNIVERSARY_DAYS_LEFT	69502	52.48%
TOTAL_OCASSION_PRODUCT	62877	47.48%
MARITAL_STATUS	43786	33.06%
STATE	34160	25.79%
AGE	28372	21.42%
DOB_DAYS_LEFT	28372	21.42%
CITY	26952	20.35%
GENDER	21887	16.52%

After null values treatment, bar graph is plotted on the whole data frame(df) values. The bottom of the graph shows name of attributes and top of the graph shows total number of records available in that attributes. All the attributes containing the same available record number i.e., 132414, which indicates that no null values are present in any attributes of the data frame (df).

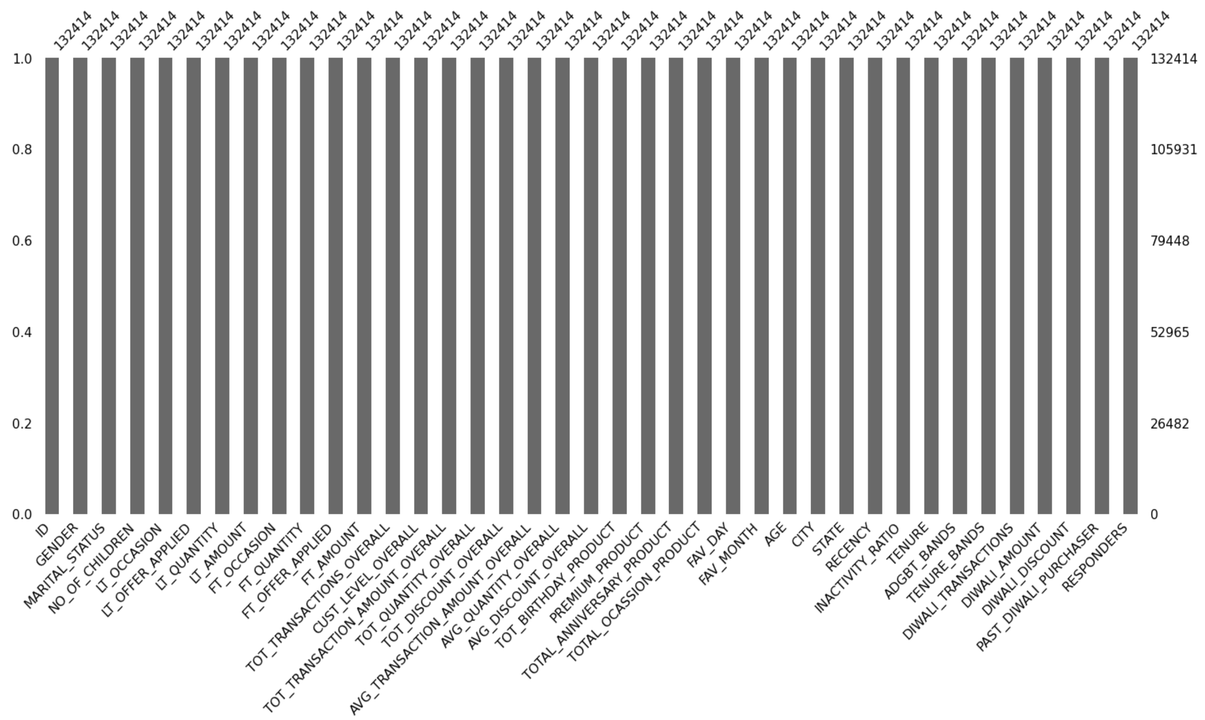

Fig. 1

 ISSN 2394 - 7780

In the dataset favorite day and months are represented by number's i.e. 1,2,3,4,5,6,7,8,9,10,11,12.These numerical values are treated as continuous values by the software during analysis,therefore before analysis these numerical values must be converted to categorical values i.e. day 1 convert to "SUNDAY", day 2 to "MONDAY",day 3 to "TUESDAY", day 4 to "WEDNESDAY", day 5 to "THURSDAY", day 6 to "FRIDAY", day 7 to "SATURDAY. Similarly, month 1 convert to 'JANUARY', month 2 to 'FEBRUARY', month 3 to 'MARCH', month 4 to 'APRIL', month 5 to 'MAY', month 6 to 'JUNE', month 7 to 'JULY', month 8 to 'AUGUST', month 9 to 'SEPTEMBER', month 10 to 'OCTOBER', month 11 to:'NOVEMBER', month 12 to 'DECEMBER'.

The attributes 'Lt_Occasion', 'Lt_Offer_Applied', 'Ft_Occasion', 'Ft_Offer_Applied', 'Past_Diwali_Purchaser' contains the values 0 and 1. These values are also converted to categorical values as 0 replaced by 'N' and 1 replaced by 'Y'.After null values treatment and other operations of preprocessing has been completed on the dataset the dataset is ready for analysis.

A graph between marital status, number of children's and overall total transactional amount is plotted (fig. 2). The fig indicates that married customers having three child's purchase more than two children and one child. customers having one child or no child behaves the same in terms of overall total transaction amount. The unmarried customers having the total transaction amount at par with the customers having two children.

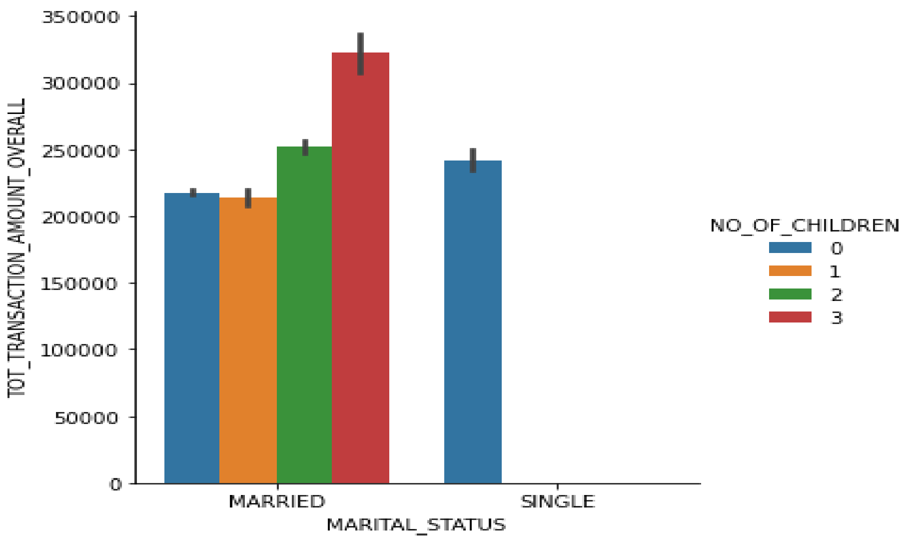

Fig. 2

The next graph plotted between gender of the customer, number of children of customers and overall total transaction value (fig. 3). The graph shows that male or female having three children's expend more amount purchase although females having three children's expend more amount on purchase than males having three children's. As a general observation females do more purchase and expend more amount on purchase than males. The graph also indicates that customers having three children, purchase more and spend more money than others.

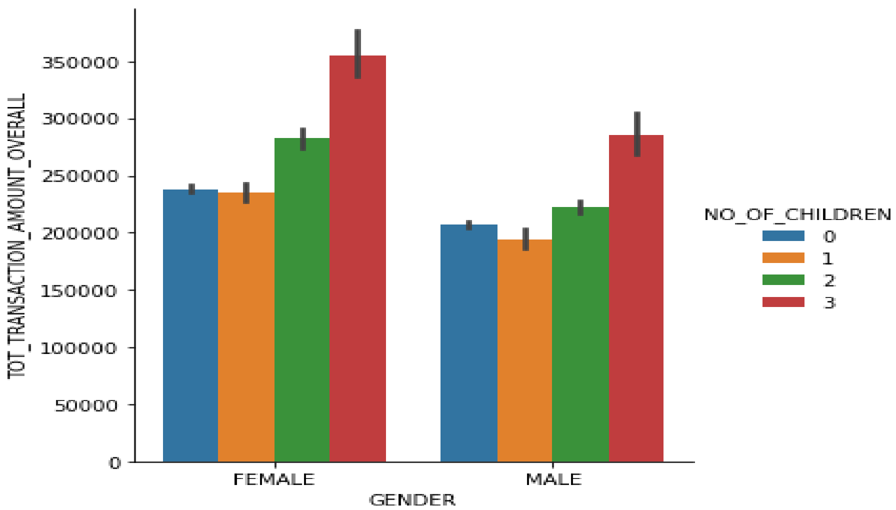

Fig. 3

The next graph plotted between customer level, number of children of customers and overall total transaction value (fig. 4).The customers having high customer level purchase more and expend more money on online buying, customers having low customer level purchase less and spend less money on online buying. Customers having medium customers level are lies within the range of low customer level and high customer levels. A graph between customer level, number of children and overall transaction value is plotted. The graph indicates that customers having two or three children expend more within their category than the customers having zero or one child. The purchase amount difference between high and medium level customers are four time than medium level customers which is a big difference and needs focus on it.

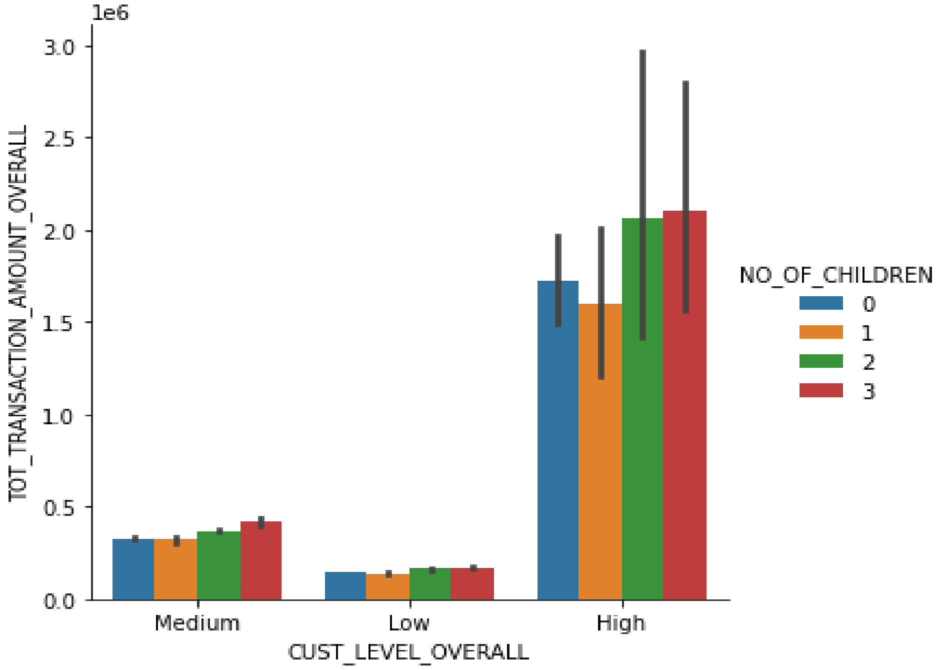

Fig. 4

Figure 5 which is the graph between overall customer level and number of children's customer have, shows that when customers move from no children to one child the customers buying level slightly goes down and when customers move from one child to two child than customers buying level slightly goes up and when customers move two children to three child the customers buying level significantly goes up an more purchase done by the customer. The number of children's have significant impact on customers buying level and buying behavior.

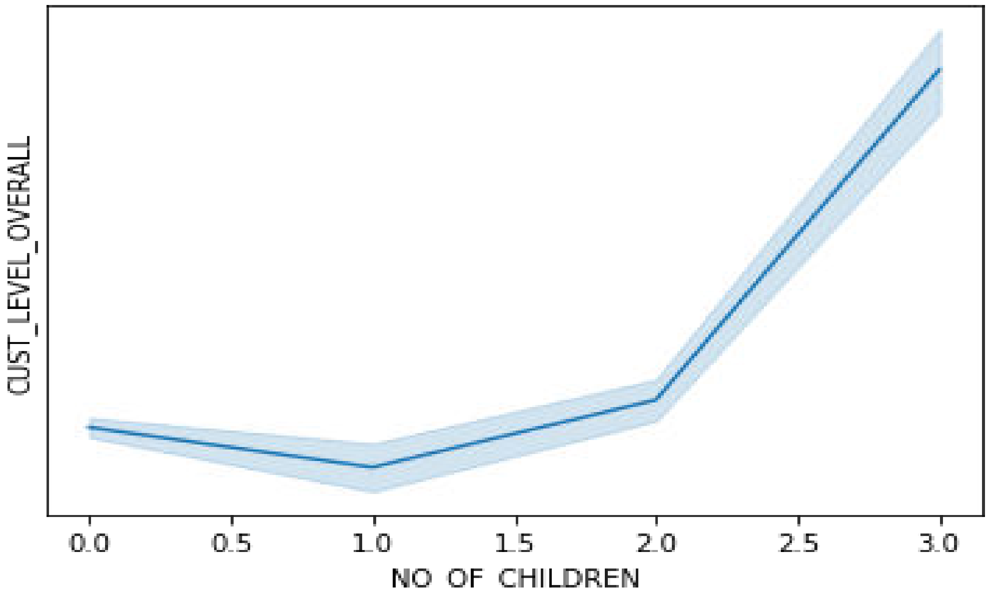

Fig. 5

Another graph plotted between number of children and number of items purchased by the customers (fig. 6). The graph is also indicating that customers having more child's purchase more items than the customers having less numbers of children.

 ISSN 2394 - 7780

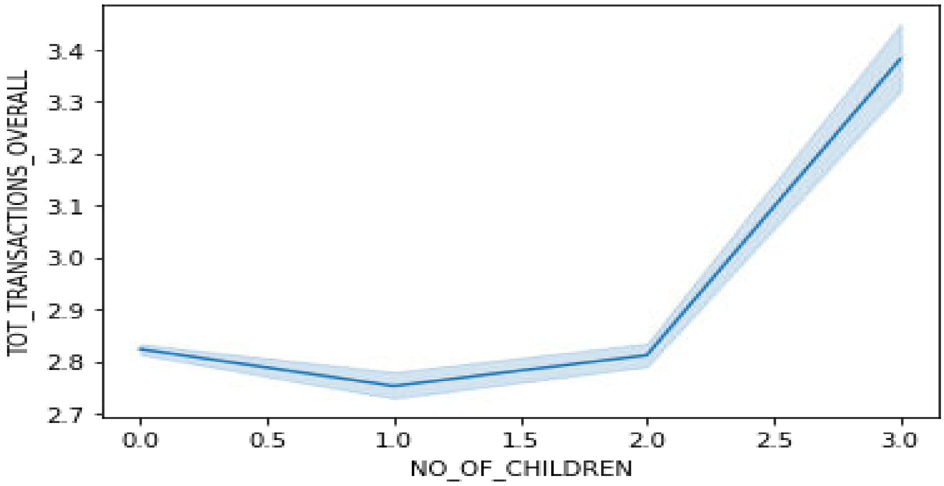

Fig. 6

Similarly, a graph between overall total discount availed by the customer and the number of children is plotted (fig. 7). This graph shows that customers having more children wants availed more discount. This is an important factor to firms who are selling products of services. Offering more discount will improve the business and Return on Investment (RoI) will also be improved.

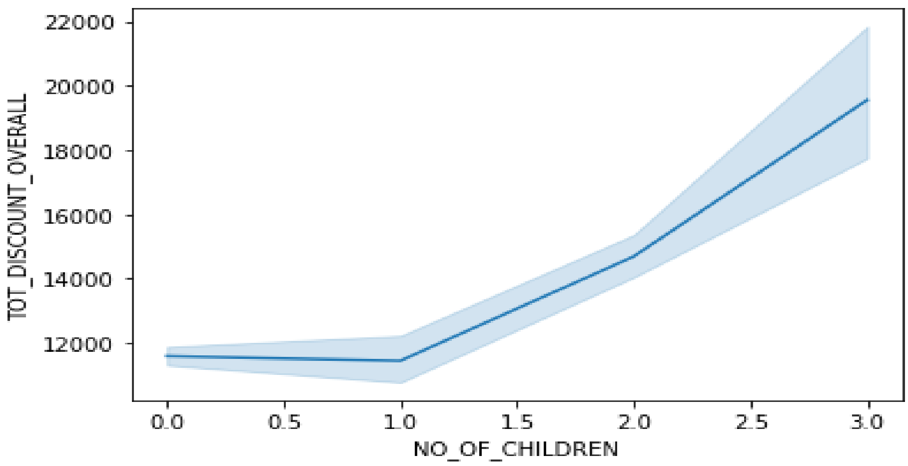

Fig. 7

Further a study is made on the favorite day, favorite month and gender of the customer's. A graph between favorite day, gender of customers and overall total transaction amount is plotted (fig. 8). The graph shows that female customers have only one favorite day on which they buy more and that day is Thursday. On other days female customers buy less and Tuesday is the less preferred day by female customers. On the other hand male customers have two favorite days on which they buy more and these days are Thursday and Friday. On other day male customers buy less and Wednesday & Monday are the less preferred day by the male customers.

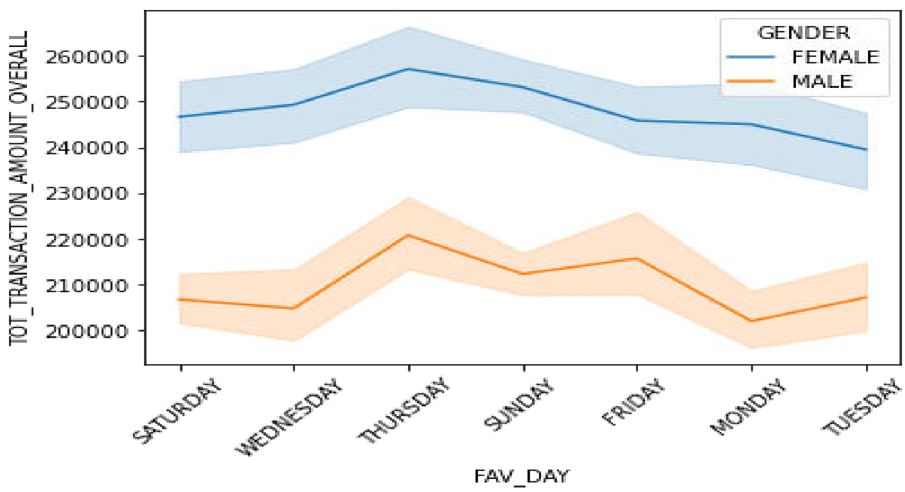

Fig. 8

ISSN 2394 - 7780

Similarly, a graph between favorite months, gender of the customer and overall average transaction amount is plotted (fig. 9). The graph shows that October is the favorite month of all the customers irrespective of gender of customers. October is also a festival months, therefore customers (both male & female) buy more during this month. The March is the less preferred month of both male and female customers. During March month the salaried customers pay all taxes from their salaries, therefore this may be the reason that both male and female customers buy less during the month of March.

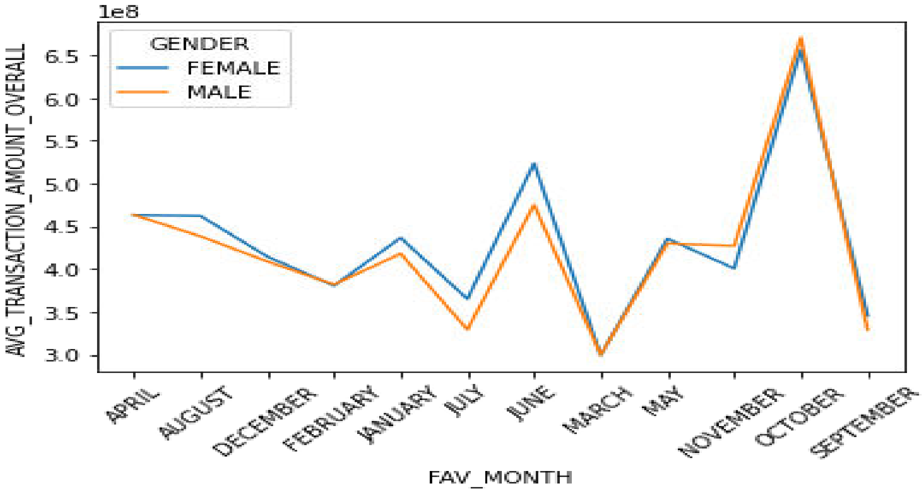

Fig. 9

CONCLUSION

A dataset taken from GitHub, related to customers on-line buying is analyzed from different angles like effect of customer gender on total transaction amount, effect of customers marital status on total transaction amount, effect of number of children on total transaction amount, relation between number of children and discount amount, relation between number of children and customer level, gender wise favorite day of the customers on which customers buy more, gender wise favorite months of the customer on which customers buy more.

It is found that personal attributes of the customers like gender, marital status, number of children, etc. have significant effect on customers buying behaviors. Female customer buys more than male customers. Female customers having no child or one child buy less than the female customers having two or three children. Female customers having three children buy more and have high customer level. Similarly, male customers having three children buy more than the male customers having no child or one child or two children. Tuesday is the favorite day of female customers and Tuesday & Friday are the favorite day of the male customers. October is the favorite month for both male and female customers and March is the month when both male and female customers buy less. Irrespective of customers gender, customers having more children wants to avail more discount to save the money.

REFERENCES

[1] Subhash Chand, et al (2022), “On-line Customers Buying Behaviour Prediction using XGBoost Algorithms in Python”, Indian Journal of Natural Sciences, Vol.13, Issue 73.

[2] Naresh Chandra, et al (2022), “Statistical Hypothesis Testing for Determining the Relationship of Categorical Variables using Python Code”, Indian Journal of Natural Sciences, Vol.12, Issue 70.

[3] Subhash Chand, et al (2022), “Prediction and Buying Behaviour of Customers using Machine Learning Technique”, Indian Journal of Natural Sciences, Vol.12, Issue 70.

[4] Naresh Chandra, et al (2022), “Customers’ Segmentation using RFM Model with k-means”, Indian Journal of Natural Sciences, Vol.13, Issue 73.

[5] Sharmeen Binti Syazwan Lai, et al (2021), “Comparing the Performance of AdaBoost, XGBoost, and Logistic Regression for Imbalanced Data”, Mathematics and Statistics, vol 9, no. 3, DOI: 10.13189/ms.2021.090320.

[6] JIKHAN JEONG, (2021), “3123301 Identifying Consumer Preferences From User-Generated Content on Amazon.Com by Leveraging Machine Learning “, IEEE Access Digital Object Identifier 10.1109/ACCESS.2021.

[7] ELENA CUEVAS-MOLANO, et al,(2021), “Factors Increasing Consumer Engagement of Branded Content in Instagram”, Digital Object Identifier 10.1109/ACCESS.2021.3121186 1.

[8] Syeda Sarah Azmi, et al (2020), “An Overview of Boosting Decision Tree Algorithms utilizing AdaBoost and XGBoost Boosting strategies”, International Research Journal of Engineering and Technology, vol 7 issue 5.

[9] LUCIANO FERREIRA DA SILVA, et al (2020), “Relationship Between Marketing and Project Management Success Through Cognitive Process Lens”, IEEE Access, Digital Object Identifier 10.1109/ACCESS.2020.3023897.

[10] MOHAMMAD ALOJAIL, et al (2020), “A Novel Technique for Behavioral Analytics Using Ensemble Learning Algorithms in E-Commerce”, IEEE Access, Digital Object Identifier 10.1109/ACCESS.2020.3016419

[11] Javier fabra,et al (2020), Log-Based Session Profiling and Online Behavioral Prediction in E_Commerce Websites, Digital Object Identifier 10.1109/ACCESS.2020.3024649.

[12] Dehua Kong1, et al (2019), Research on Product Recommendation Based on Web Space-Time Customer Behavior Trajectory, International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI).

[13] Chen ling, et al (2019), Customer Purchase Intent Prediction Under Online Multi-Channel Promotion: A Feature-Combined Deep Learning Framework, Digital Object Identifier 10.1109/ ACCESS.2019.2935121.

[14] Kokwaitham, et al (2019), Perceived Risk Factors Affecting Consumers’ Online Shopping Behaviour, Journal of Asian Finance, Economics and Business Vol 6 No 4 pp. 249-260.

[15] Yu jiang, et al (2019), “A Pedestrian Detection Method Based on Genetic Algorithm for Optimize XGBoost Training Parameters”, Digital Object Identifier 10.1109/ACCESS.2019.2936454, vol 7.

[16] Oly Mishra, et al (2018), Sources of Need Recognition in Online Consumers Decision Making Process, IJEMR – April 2018 - Vol 8 Issue 04 - Online - ISSN 2249–2585.

[17] Sergio hernández, et al (2017), Analysis of Users' Behavior in Structured e-Commerce Websites, Digital Object Identifier 10.1109/ACCESS.2017.2707600.

[18] Popli Ankita, et al (2015), Factors of Perceived Risk Affecting Online Purchase Decisions of Consumers, Pacific Business Review International Volume 8, Issue 2.

[19] Liu Nengbao, et al (2014), Handling Class Imbalance in Customer Behavior Prediction, IEEE, 978-1-4799-5158-1/14.

[20] Gupta Mona, et al (2014), Characterizing Comparison Shopping Behavior: A Case Study, IEEE 115 ICDE Workshops, 978-1-4799-3481-2/14.