Predicting the Consumer's Product Purchase Intention Using Regression Analysis at Attribute Level

K Radha[#1], V Karthik [*2], K Manish[#3]

[#1]*Asst Professor,CSE,GITAM University,Rudraram,Telangana,India*

***Abstract* —** *Recently, Retail 4.0 is higher demand for accurate prediction of consumer's purchase intention. In this regard, an attribute level decision support prediction model has been created for providing an influential online shopping platform to the customers. In order to build the prediction model, brand's social reviews' polarity are calculated from social network mining and sentiment analysis, respectively. Afterward, an appropriate regression analysis and required instances have been found for each attribute to predict the appropriate product stats. One of the key findings, the camera attributes: sensor, display, and image stabilization make the customer attention at the end of the search. The outcomes of this analysis can be profitable to online retailers and prepare an efficient platform for the customers to obtain the desired goods. Finally, the sensitivity analysis has also been done to test the robustness of the applied model.*

Keywords — *attribute level decision support prediction model, regression analysis, social network mining, sentiment analysis, e-commerce retailers*

I. INTRODUCTION

Recently, Retail 4.0 is progressively demanding the accurate prediction of consumer's purchase intention. In this regard, an attribute level decision support prediction model has been developed for providing an influential e-commerce platform to the customers. In order to build the prediction model, brands' social perception score and reviews' polarity are computed from social network mining and sentiment analysis, respectively. Afterward, an appropriate regression analysis and suitable instances have been identified for each attribute to predict the appropriate product attributes. One of the key findings, the camera attributes: sensor, display, and image stabilization pursue the customer attention at the end of the search. The outcomes of this analysis can be beneficial to e-commerce retailers and prepare an efficient search platform for the customers to obtain the desired durable goods in an adorable form. Finally, the sensitivity analysis has also been performed to test the robustness of the proposed model. Online shopping tendency is meritoriously boosting after the advent of bricks-and-mortar retailers. In the year of 2016, e-retailers have generated the estimated revenue of 1.9 trillion U.S. dollars from 1.61 billion customers globally. Amazon, the leading international e-retail company, has more than 310 million active customer accounts who bought near 136 billion U.S. dollars' goods in 2016 (Statista, 2017). In the first month of demonetization, the growth of digital payment in the world third purchasing power parity country (India) was escalated 271% and simultaneously the cash on delivery was dropped about 30-40% (Chronicle, 2017). Furthermore, out of the total e-commerce market, consumers approximately purchase 34% of durable goods. Thus, an analysis of online consumer's buying behaviour of durable goods is a vital aspect in e-commerce market to represent the online shopping in an eloquent way.

Consumers conduct a deep search before the purchase of required goods. As per consumer searches camera on an average of 14 times before the purchase. Initially, they search the product based on their needs to gain experience and thorough screening of reviews is exercised by them before a confirmation of purchase. Later, a goal-oriented customer goes for the deeper search to extract the deep level information and read reviews to make a final decision. Tracking the melody of consumers' online purchasing behavior and storing it in a structural form is a stimulating work. ComScore is a US based leading company that does the job and stores it as comScore Panel Data. In the recent years, an enormous consumers online review data is becoming a valuable research area for exploring the influential factors in the e-commerces domain. It is found that they have increased the revenue of 2.7 billion dollars (Spool, 2009) after setting the question "Was this review helpful to you?" on each customers' review records. Investigating the effect of consumers review data for influencing the purchase intention is a important work for providing the prominent e-commerce platform to the customers. Brightlocal.com has found out that 84% of individuals follow the online reviews before purchase. Online reviews impact on consumers' confidence for purchasing a product as well as to provide the real-life experiences, whereas product

company benefits from the product feedback to improve the quality of goods.

It is essential to explore the consumer's behavior on attribute level for desired goods. Importantly, the question "which search pattern are responsible for influencing the consumer's choice?" has become an exclusive question of recent time in e-commerce market. In the favor of raised question, different influential factors have been differentiated by many researchers in the past for purchasing an online product . Social network enhances the consumers' perception on brand name and also helps the customer to recognize new branded products . Furthermore, brand names impact the consumer's mind for selecting and the able to pay for an individual product . Likewise, consumer online reviews influence the customer decision for buying goods and assists the product company for forecasting the product sales .

An in-depth observation has been made by using linear regression analysis for seeing the deep level consumer's search and selection patterns. To address the query, researchers investigated the consumers search patterns and detected the influence of online reviews for purchasing goods. However, the combined effect of deep level consumer's search and screening reviews patterns on purchasing products has not yet been done. Further, social perception score is an important aspect in this recent times which should be incorporated in the prediction model. Our analysis considers the SPS and encounters its usefulness, and further investigates the joint effect of consumer online search and screening reviews on the purchase decision. According to our best knowledge, another limitation in the existing literature is that no one performed the regression analysis based on the linear and non-linear property of the attributes to deal with consumer buying behavior. However, it is observed that their diverse searching interests on distinct attributes vary with both the linear and nonlinear patterns in the collected data set. This issue has been addressed by simultaneously considering linear and nonlinear regression analysis based on the consumers' searching nature for the specific attributes.

The present research develops a deep level product selection strategies: (i) search to choose, (ii) view reviews , (iii) search attributes with looking product overall reviews to choose, and (iv) search attributes with screening corresponding attribute's reviews to choose. Further, predicted attributes values are searched on database to recommend the relevant products those customers desired to purchase. In the previous study, a number of researchers have considered the influence of consumer's review on purchasing products whereas they have overlooked the reviews which do not affect the customers' intention towards purchasing a product.This research is important for investigating which attributes are significant for changing the consumer's mindset towards purchasing any product. Furthermore, the total number of consumer's online search has been normalized to ten equal deciles.

Rest of the paper is organized as follows. In section 2, background and related work are discussed in detail. The case description is shared in section 3. Preprocessing of the data with detailed data description is provided in Section 4. Further, section 5 elucidates the proposed research methodology. The results of Proposed Algorithm and managerial insights are shown in section 6. Finally, the paper is concluded with limitations and future extensions in section 7.

II. RELATED WORK

A number of studies have been performed for analyzing the insights of online consumers buying behavior. However, only a few of them have found the customers buying behavior for durable goods. Still, an attribute-level prediction model with the integration of consumer search pattern, social perceptual score, and online reviews has not been addressed in the existing literature. By the critical examination of the background and related work, it is divided into four subsections, (i) Influential Factors of e-commerce, (ii) Brand and Social Network, (iii) Impact of Online Consumers Review, and (iv) Predict the Consumers Purchase.It seems that the research growth in the related domain is continuously booming markedly. The collection process of the records is done by a simple technique. The keywords of each classification are searched on Scopus database with the constraint 'Article' in document type to collect the number of research articles published in Scopus indexed journals. The aim of this finding is just to show the importance of this research domain.

A. Influential Factors of Online Shopping

A number of key factors exist within the literature that influences the shoppers toward a web purchase. Primarily, Consumers get influenced for buying a particular branded product from three main online information sources, eWOM , manufacturer/retailer, and neutral/third party. The gender variations and product sort additionally impact of buying digital and non-digital product . As per Malc et al. the price fairness isn't solely influenced the shoppers to shop for a product however additionally unfold a negative perception regarding the vendor. Some online resources like video blogs changed the consumers' mindset on the physical and social attractiveness of luxury brand perceptions and attitude homophily on

para-social interaction (PSI) have observed that the compromise result for getting consumer goods is strong than the fast-moving goods (FMCG).

B. Impact of Online Consumers Review

In the recent era, on-line review systems square measure creating the biases on social influence and merchandise choice. In order to reduce the biases, have differentiated and investigated the retailer promoted reviews and self-motivated reviews for the same product. explored the results of on-line shoppers review, goods type, and the perceptions in the decision-making process on consumers buying intention. They derived that the negative reviews impact more on purchase decision than the positive reviews, and suggested the retailing management for delivering a quick response to negative comments. Initially, shoppers get influenced by the standard of product info. Then, things rating and overall standing build the client call towards the acquisition. According to, previous sales data and consumer reviews are helpful for forecasting the product sales, which was verified by integrating the model of sentiment analysis and Bass/Norton model. have been observed the same findings by summarizing the reviews based on the different feature.

Consumers perceived risk, usefulness, structural assurance, effectiveness, and so on from the product reviews and get influenced to buy goods individually or in the group . Further, have used consumer reviews to build the mining perceptual map which provides a practical vision for smartphone companies to take suitable marketing decision. In the gift study, a number of machine learning based algorithm has been applied on product reviews to classify the reviews into multiple feature vectors . For example, have proposed a logistic regression-based prediction model to classify the reviews into two sets, high and low trustworthiness. In order to investigate the consumers' screening reviews activities. introduced Associate in Nursing eye-tracking technique and finds that the majority of the shoppers concentrate on attributes level reviews. Besides, process and extracting the data from the large quantity of on-line

client review knowledge modify students to explore the analysis space of information management systems, huge knowledge analytics, and natural language processing.

C. Predict the Consumers Purchase Intention and Recommendations

In the recent digital atmosphere, e-business firms wide used the merchandises recommendation and on-line advertising for raising the product sales. Google and Rival Microsoft spent nearly 350 and 746 million US dollar respectively in the advertising of its products and services. A novel active learning approach has been developed by [30] to improve the prediction accuracy of recommendation system and search advertising. People typically request on-line product with/without any specific target.have presented that e-commerce detects two types of search information such as uncertain and goal-oriented. A study has been conducted on 1261 Dutch automotive homeowners by bunch the pre-purchase info search and modeling the structural equation in characteristic the sequence of search. The search data of 109 participants, those who have purchased at least single water bottled, has been collected and evaluated the influence in product choice. It is also observed that when participants visit the environmental, economic, or health-related websites where emphasized the benefits for bottled water are more likely to buy. Moreover,surveyed the online information search behavior for purchasing mobile phone and laptop on 643 participants from London and Birmingham. They have predicted the influential factor by using multiple linear regression analysis. Similarly, Poel & Buckinx (2005) proposed a prediction model using Logit regression to find the customer purchase intention in their next visit to the retailer website. Initially, they determine ninety two input variables and classified them into four classes like general clickstream, customer demographics, detailed clickstream, and historical purchase behavior and finally choose 9 variables by using forward and backward variable selection techniques.categorise the physical product into sturdy, nondurable and industrial goods and then analyze the correlation between consumers' web search and purchase behavior. The result shows that predicting the acquisition of durables is critical from the search traffic. However, they exempted the direction of that search path deals with the buyer towards purchase. A number of machine learning based prediction models have been developed with the help of several textual features like polarity, entropy, subjectivity, and reading ease those mitigates the Matthew effect and detects the helpfulness of customer reviews.

In the past analysis, the mutual effect of online search and screening reviews data is hardly seen due to the complexity of tracking and linking that information. This analysis carries out the combined influence of on-line search and screening reviews for predicting attributes level.

consumers' purchase intention. The waves of looking out and screening consumers' review on attribute level have additionally been incorporated during this analysis individually to find the prestigious factors in client purchase call. In addition, predicting the brand name is difficult from the history of searched brands keyword. Therefore, in this study, the brand's name has been converted into a numerical form of a social perceptual score for the brand's eco-friendly and

luxuries nature. This social sensory activity score helps the shoppers to select up personalised branded product and solves the starvation downside of recent branded or less fashionable product. The aim of this exploration is to help the buyer to settle on their desired product quickly and to assist the e-retailing company for taking an appropriate social control call. An attribute level prediction model has been planned to realize the key objectives of this study.

III. CASE DESCRIPTION

In this paper, a period of time shoppers 'online search and selection state of affairs are captured for predicting their purchase intention of durables. Amazon has been selected as an e-commerce platform for collecting the consumers' real-time search and reviews data. Furthermore, the camera has chosen as durable goods for attribute level analysis. Generally, Amazon predicts the consumers' purchase intention and recommends the product in many ways that on their web site. Recommendation system having desire to make the application simple and user-friendly in the e-commerce domain and the research on this area is still highly active. Amazon uses item-based collaborative filtering in their recommendation systems, which has been performing well since 2003. However, existing recommendation systems have few limitations where some of the drawbacks are especially for durable goods. Consumers pay an extended amount of your time in looking out the product and screening reviews before getting any durables. Therefore, capturing the consumers' semipermanent search Associate in Nursingd screening reviews information and so storing it in an analytical kind for predicting their purchase intention could be a difficult task. An attribute-level prediction model has been developed with both linear and non-linear regression analysis for predicting the consumers' purchase intention. Moreover, the sentiment analysis has applied to convert the reviews into polarity and, then incorporate the reviews' polarity with the search patterns. Besides, new items or less popular branded products suffer from cold start problem which causes starvation. The planned model is used the social network mining to work out the brand's social perception score and to pursue the starvation drawback.

IV. PROBLEM STATEMENT

To predict the amount of purchases in Retail shop using Machine Learning algorithm called MULTIPLE LINEAR REGRESSION

DATA SET:
The given data set consists of the following parameters:
A - User_ID
B - Product_ID
C - Gender
D - Age
E - Occupation
F - City_Category
G - Stay_In_Current_City_Years
H - Marital_Status
I - Product_Category_1
J - Product_Category_2
K - Product_Category_3
L – Purchase

A. OBJECTIVE:

To get a better understanding and chalking out a plan of action for solution of the client, we have adapted the view point of looking at product categories and for further deep understanding of the problem, we have also considered gender age of the customer and reasoned out the various factors of choice of the products and they purchase , and our primary objective of this case study was to look up the factors which were dampening the sale of products and correlate them to product categories and draft out an outcome report to client regarding the various accepts of a product purchases.

B. PREPROCESSING OF THE DATA:

Preprocessing of the data actually involves the following steps:

C. GETTING THE DATASET

We can get the data set from the database or
we can get the data from client.

D. IMPORTING THE LIBRARIES

We have to import the libraries as per the requirement of the algorithm.

We have to import the libraries as per the requirement of the algorithm.

IMPORTING THE LIBRARIES

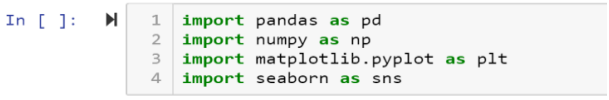

Figure1 : Importing Libraries

E. IMPORTING THE DATA-SET:

Pandas in python provide an interesting method read_csv(). The read_csv function reads the entire dataset from a comma separated values file and we can assign it to a DataFrame to which all the

operations can be performed. It helps us to access each and every row as well as columns and each and every value can be access using the dataframe. Any missing value or NaN value have to be cleaned.

READING THE DATA-SET

```
[3]: 1 df=pd.read_csv('datafiles/train.csv')
     2 df.head()
```

Out[3]:

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status
0	1000001	P00069042	F	0-17	10	A	2	0
1	1000001	P00248942	F	0-17	10	A	2	0
2	1000001	P00087842	F	0-17	10	A	2	0
3	1000001	P00085442	F	0-17	10	A	2	0
4	1000002	P00285442	M	55+	16	C	4+	0

Fig2 : Reading the dataset

F. HANDLING MISSING VALUES:

Missing values can be handled in many ways using some inbuilt methods:
(a)dropna()
(b)fillna()
(c)interpolate()
(d)mean imputation and median imputation

(a)dropna():

dropna() is a function which drops all the rows and columns which are having the missing values(i.e. NaN)

```
In [48]: 1 data
```

	day	temperature	windspeed	event
0	2017-01-01	32.0	6.0	Rain
1	2017-01-02	NaN	7.0	Sunny
2	2017-01-03	28.0	NaN	Snow
3	2017-01-04	NaN	7.0	0
4	2017-01-05	32.0	NaN	Rain
5	2017-01-06	31.0	2.0	Sunny
6	2017-01-06	34.0	5.0	0

Fig 2 : data before using dropna()

```
In [52]: 1 data.dropna()
```

Out[52]:

	day	temperature	windspeed	event
0	2017-01-01	32.0	6.0	Rain
5	2017-01-06	31.0	2.0	Sunny
6	2017-01-06	34.0	5.0	0

Fig 3 : data after using dropna()

- dropna() function has a parameter called how which works as follows
  - if how = 'all' is passed then it drops the rows where all the columns of the particular row are missing
- if how = 'any' is passed then it drops the rows where all the columns of the particular row are missing

(b)fillna():

fillna() is a function which replaces all the missing values using different ways.

```
In [45]: 1 data.fillna(0)
```

Out[45]:

	day	temperature	windspeed	event
0	2017-01-01	32.0	6.0	Rain
1	2017-01-02	0.0	7.0	Sunny
2	2017-01-03	28.0	0.0	Snow
3	2017-01-04	0.0	7.0	0
4	2017-01-05	32.0	0.0	Rain
5	2017-01-06	31.0	2.0	Sunny
6	2017-01-06	34.0	5.0	0

Fig 4 : functioning of fillna(0)

- fillna() also have parameters called method and axis
- if we use method = 'ffill' where ffill is a method called forward fill, which carry forwards the previous row's value

• if we use method = 'bfill' where bfill is a method called backward fill, which carry backward the next row's value

• if we use method = 'ffill' , axis = 'columns' then it carry forwards the previous column's value

• if we use method = 'bfill' , axis = 'columns' then it carry backward the next column's value

(c)interpolate():

• • interpolate() is a function which comes up with a guess value based on the other values in the dataset and fills those guess values in the place of missing values

```
In [51]: 1 data.interpolate()
```

Out[51]:

	day	temperature	windspeed	event
0	2017-01-01	32.0	6.0	Rain
1	2017-01-02	30.0	7.0	Sunny
2	2017-01-03	28.0	7.0	Snow
3	2017-01-04	30.0	7.0	0
4	2017-01-05	32.0	4.5	Rain
5	2017-01-06	31.0	2.0	Sunny
6	2017-01-06	34.0	5.0	0

Fig 5: functioning of interpolate()

(d)mean and median imputation

• mean and median imputation can be performed by using fillna().

• mean imputation calculates the mean for the entire column and replaces the missing values in that column with the calculated mean.

• median imputation calculates the median for the entire column and replaces the missing values in that column with the calculated median.

Mean Imputation

```
In [77]: data.fillna(data.mean())
```

Out[77]:

	day	temperature	windspeed	event
0	2017-01-01	32.0	6.0	Rain
1	2017-01-02	31.4	7.0	Sunny
2	2017-01-03	28.0	5.4	Snow
3	2017-01-04	31.4	7.0	0
4	2017-01-05	32.0	5.4	Rain
5	2017-01-06	31.0	2.0	Sunny
6	2017-01-06	34.0	5.0	0

Fig 6: mean imputation

Median Imputation

```
In [78]: data.fillna(data.median())
```

Out[78]:

	day	temperature	windspeed	event
0	2017-01-01	32.0	6.0	Rain
1	2017-01-02	31.4	7.0	Sunny
2	2017-01-03	28.0	5.4	Snow
3	2017-01-04	31.4	7.0	0
4	2017-01-05	32.0	5.4	Rain
5	2017-01-06	31.0	2.0	Sunny
6	2017-01-06	34.0	5.0	0

Fig 7 : median imputation

```
In [88]: price.head()
```

Out[88]:

	town	area	price
0	monroe township	2600	550000
1	monroe township	3000	565000
2	monroe township	3200	610000
3	monroe township	3600	680000
4	monroe township	4000	725000

Fig 8 : categorical data

```
In [89]: dummy_set = pd.get_dummies(price.town)
         dummy_set
```

Out[89]:

	monroe township	robinsville	west windsor
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0
5	0	0	1
6	0	0	1
7	0	0	1
8	0	0	1
9	0	1	0
10	0	1	0
11	0	1	0
12	0	1	0

Fig 9 : dummy set for the above data

```
In [90]: merged = pd.concat([price, dummy_set], axis = 1)
In [91]: merged
```

Out[91]:

	town	area	price	monroe township	robinsville	west windsor
0	monroe township	2600	550000	1	0	0
1	monroe township	3000	565000	1	0	0
2	monroe township	3200	610000	1	0	0
3	monroe township	3600	680000	1	0	0
4	monroe township	4000	725000	1	0	0
5	west windsor	2600	585000	0	0	1
6	west windsor	2800	615000	0	0	1
7	west windsor	3300	650000	0	0	1
8	west windsor	3600	710000	0	0	1
9	robinsville	2600	575000	0	1	0
10	robinsville	2900	600000	0	1	0
11	robinsville	3100	620000	0	1	0
12	robinsville	3600	695000	0	1	0

Fig 10 : adding the dummy set to dataframe

- Getting dummies using label encoder from scikit learn package

We have a method called label encoder in scikit learn package .we need to import the label encoder method from scikitlearn package and after that we have to fit and transform the data frame to make the categorical data into dummies.

If we use this method to get dummies then in place of categorical data we get the numerical values (0,1,2….)

Dummy Variables using LabelEncoder

```
In [95]: from sklearn.preprocessing import LabelEncoder
         le = LabelEncoder()
```

Fig 11 : importing the method

```
dfle = price
dfle.town = le.fit_transform(price.town)
dfle
```

t[96]:

	town	area	price
0	0	2600	550000
1	0	3000	565000
2	0	3200	610000
3	0	3600	680000
4	0	4000	725000
5	2	2600	585000
6	2	2800	615000
7	2	3300	650000
8	2	3600	710000
9	1	2600	575000
10	1	2900	600000
11	1	3100	620000
12	1	3600	695000

Fig12: handling categorical data

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size= 0.2, random_state = 0)
```

Fig 13 : importing train_test_split

Using OLS: OLS is Ordinary Least Squares .This method is available in formula.api package in statsmodel library. So to use this method we have to import statsmodel.formula.api

- In this method we need to first mention the output column names followed by "~" followed by input column names, and this entire thing should be mentioned in double quotes.

- Multiple input and output columns can be mentioned using concat(+) symbol.

• This method gives the summary of the model.

• From this summary we have to observe the R-squared value, R-squared value gives the accuracy mentioned by the client(i.e. if accuracy is 80% then R-squared value must be greater than 0.8)

```
import statsmodels.formula.api as smf
a = smf.ols("AT ~ Waist", data = df).fit()
a.summary()
```

OLS Regression Results

Dep. Variable:	AT	R-squared:	0.670
Model:	OLS	Adj. R-squared:	0.667
Method:	Least Squares	F-statistic:	217.3
Date:	Sat, 08 Jun 2019	Prob (F-statistic):	1.62e-27
Time:	14:29:40	Log-Likelihood:	-534.99
No. Observations:	109	AIC:	1074.
Df Residuals:	107	BIC:	1079.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-215.9815	21.796	-9.909	0.000	-259.190	-172.773
Waist	3.4589	0.235	14.740	0.000	2.994	3.924

Fig 14 : ols and its functioning

• If the model doesn't meet the accuracy we can perform mathematical operations on input and output columns, so that it meets the accuracy

```
sqrta = smf.ols("np.sqrt(AT) ~ Waist", data=df).fit()
sqrta.summary()
```

OLS Regression Results

Dep. Variable:	np.sqrt(AT)	R-squared:	0.710
Model:	OLS	Adj. R-squared:	0.707
Method:	Least Squares	F-statistic:	261.5
Date:	Sat, 08 Jun 2019	Prob (F-statistic):	1.69e-30
Time:	14:34:03	Log-Likelihood:	-202.91
No. Observations:	109	AIC:	409.8
Df Residuals:	107	BIC:	415.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6.8984	1.036	-6.660	0.000	-8.952	-4.845
Waist	0.1803	0.011	16.170	0.000	0.158	0.202

Fig 15 : mathematical operation on input or output

V. EVALUATION OF MULTIPLE LINEAR REGRESSION

• **MULTIPLE LINEAREGRESSION**:
A multiple linear regression model allows us to capture the relationship between multiple feature columns and the target column. Here's

• what the formula looks like:
y^=a0+a1x1+a2x2+...+anxn

• • Importing the required libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Fig 16. Importing the libraries

• *Reading the Data-Set*

```
df=pd.read_csv('datafiles/train.csv')
df.head()
```

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Cat
0	1000001	P00069042	F	0-17	10	A	2	0	
1	1000001	P00248942	F	0-17	10	A	2	0	
2	1000001	P00087842	F	0-17	10	A	2	0	
3	1000001	P00085442	F	0-17	10	A	2	0	
4	1000002	P00285442	M	55+	16	C	4+	0	

Fig 17 : Reading the Data-Set

Handling the missing values

o There is a method called isnull() which gives the number of missing values in each and every column.

o Using fillna() method each and every missing value is replaced by 0.

```
df.isnull().sum()
```

```
User_ID                          0
Product_ID                       0
Gender                           0
Age                              0
Occupation                       0
City_Category                    0
Stay_In_Current_City_Years       0
Marital_Status                   0
Product_Category_1               0
Product_Category_2          173638
Product_Category_3          383247
Purchase                         0
dtype: int64
```

Fig 18 : Before handling the missing the values

```
df1=df.iloc[:,:].apply(lambda x:x.fillna(0))
df1.isnull().sum()
```

```
User_ID                     0
Product_ID                  0
Gender                      0
Age                         0
Occupation                  0
City_Category               0
Stay_In_Current_City_Years  0
Marital_Status              0
Product_Category_1          0
Product_Category_2          0
Product_Category_3          0
Purchase                    0
dtype: int64
```

Fig 19 : After handling the missing values

• Dealing with categorical data
Using label encoder from preprocessing package which is present in scikit learn, we can get dummies in place of categorical data

• Once we get dummies we need to fit and transform that to our dataframe

```
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
df1.Gender=le.fit_transform(df1.Gender)
df1.head()
```

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Cat
0	1000001	P00069042	0	0-17	10	A	2	0	
1	1000001	P00248942	0	0-17	10	A	2	0	
2	1000001	P00087842	0	0-17	10	A	2	0	
3	1000001	P00085442	0	0-17	10	A	2	0	
4	1000002	P00285442	1	55+	16	C	4+	0	

Fig 20 : Getting dummies for Gender column

```
df1.Age=le.fit_transform(df1.Age)
df1.head()
```

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Cat
0	1000001	P00069042	0	0	10	A	2	0	
1	1000001	P00248942	0	0	10	A	2	0	
2	1000001	P00087842	0	0	10	A	2	0	
3	1000001	P00085442	0	0	10	A	2	0	
4	1000002	P00285442	1	6	16	C	4+	0	

Fig 21 : Getting dummies for Age column

```
df1.City_Category=le.fit_transform(df1.City_Category)
df1.head()
```

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Cat
0	1000001	P00069042	0	0	10	0	2	0	
1	1000001	P00248942	0	0	10	0	2	0	
2	1000001	P00087842	0	0	10	0	2	0	
3	1000001	P00085442	0	0	10	0	2	0	
4	1000002	P00285442	1	6	16	2	4+	0	

Fig 22 : Getting dummies for City_Category column

- Removing the columns which are not required using drop method

```
df1=df1.drop(['Stay_In_Current_City_Years','User_ID','Product_ID'],axis=1)
df1.head()
```

	Gender	Age	Occupation	City_Category	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3
0	0	0	10	0	0	3	0.0	0.0
1	0	0	10	0	0	1	6.0	14.0
2	0	0	10	0	0	12	0.0	0.0
3	0	0	10	0	0	12	14.0	0.0
4	1	6	16	2	0	8	0.0	0.0

Fig 23 : dropping the columns which are not required

- To get the correlation we have a method called corr() which gives the correlation between each and every column

```
df1.corr()
```

	Gender	Age	Occupation	City_Category	Marital_Status	Product_Category_1	Product_Categor
Gender	1.000000	-0.004262	0.117291	-0.004515	-0.011603	-0.045594	-0.000
Age	-0.004262	1.000000	0.091463	0.123079	0.311738	0.061197	0.018
Occupation	0.117291	0.091463	1.000000	0.034479	0.024280	-0.007618	0.006
City_Category	-0.004515	0.123079	0.034479	1.000000	0.039790	-0.014364	0.016
Marital_Status	-0.011603	0.311738	0.024280	0.039790	1.000000	0.019888	0.001
Product_Category_1	-0.045594	0.061197	-0.007618	-0.014364	0.019888	1.000000	-0.067
Product_Category_2	-0.000954	0.018770	0.006712	0.016003	0.001099	-0.067877	1.000
Product_Category_3	0.036146	-0.007422	0.012269	0.035525	-0.004629	-0.385534	0.094
Purchase	0.060346	0.015839	0.020833	0.061914	-0.000463	-0.343703	0.052

Fig 24 : correlation

Building the model(using the statsmodel library):

- We need to import the formula.api package from statsmodel library
- By importing this package we can use ols(ordinary least squares) to get the summary which gives R-squared value.
- As we have to predict the amount of purchases we have to take Purchase column as output and the remaining columns as input
- These output and input columns are separated by "~" symbol and the multiple inputs are taken using concat(+) symbol.
- This entire thing should be in quotations.

```
import statsmodels.formula.api as smf
import numpy as np
df2=smf.ols("Purchase ~ Gender + Age + Occupation + City_Category+Marital_Status+Product_Categ
df2.summary()
```

OLS Regression Results

Dep. Variable:	Purchase	R-squared:	0.170
Model:	OLS	Adj. R-squared:	0.170
Method:	Least Squares	F-statistic:	2436.
Date:	Fri, 14 Jun 2019	Prob (F-statistic):	0.00
Time:	15:14:25	Log-Likelihood:	-1.6448e+06
No. Observations:	166821	AIC:	3.290e+06
Df Residuals:	166806	BIC:	3.290e+06
Df Model:	14		
Covariance Type:	nonrobust		

Fig 25 : Getting R-squared value using ols

- Applying mathematical function to improve R-squared value.

```
exp=smf.ols("np.log(Purchase) ~ Gender + Age + Occupation + City_Category+Marital_Status+Produ
exp.summary()
```

OLS Regression Results

Dep. Variable:	np.log(Purchase)	R-squared:	0.219
Model:	OLS	Adj. R-squared:	0.219
Method:	Least Squares	F-statistic:	3348.
Date:	Fri, 14 Jun 2019	Prob (F-statistic):	0.00
Time:	15:24:31	Log-Likelihood:	-1.3006e+05
No. Observations:	166821	AIC:	2.601e+05
Df Residuals:	166806	BIC:	2.603e+05
Df Model:	14		
Covariance Type:	nonrobust		

Fig 26 : Improved R-squared value

A. Building the model (using splitting):

- First we have to retrieve the input and output sets from the given dataset

```
X=df.iloc[:,0:8]
X.head()
```

	Gender	Age	Occupation	City_Category	Marital_Status	Product_Category_1	Product_Category_2	Product_Category_3
0	0	0	10	0	0	3	0.0	0.0
1	0	0	10	0	0	1	6.0	14.0
2	0	0	10	0	0	12	0.0	0.0
3	0	0	10	0	0	12	14.0	0.0
4	1	6	16	2	0	8	0.0	0.0

Fig 27 : Retrieving the input columns

```
y=df.iloc[:,8]
y.head()
```

```
0     8370
1    15200
2     1422
3     1057
4     7969
Name: Purchase, dtype: int64
```

Fig 28 : Retrieving the output columns

- Import the train_test_split from model_selection package from scikitlearn library

• Then assigning the output to four different variables, before assigning we have to mention the train size or test size as a parameter to train_test_split. Then this method will split according to the size and assigns it to four variables.

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_state=0)
print(X_train.shape)
print(y_train.shape)
print(X_test.shape)
print(y_test.shape)
```

```
(440054, 8)
(440054,)
(110014, 8)
(110014,)
```

Fig 29 : Splitting the data

Import linear regression method which is available in linear_model package from scikit learn library

```
from sklearn.linear_model import LinearRegression
reg=LinearRegression()
reg.fit(X_train,y_train)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

Fig 30 : Importing linear regression

Once the model is built we need to check for accuracy

This can be done using predict method which is used to predict the output for input test set, and compare the predicted output with original output test set.

```
pred=reg.predict(X_test)
pred
```

```
array([10481.08806468,  6353.84743449, 11087.34680048, ...,
       12617.06615915, 11490.83221345, 10154.2145273 ])
```

Fig 31 : Predicting the output

```
np.mean(pred)
```

```
9261.993616616272
```

```
np.mean(y_test)
```

```
9269.135110076899
```

Fig 32 : Comparing with original output

VI. Limitations and Future Extensions

Nowadays, the consumers' internet usage strategy has been upgraded; some degree of changeability might be seen from the previous patterns which should think over in the model. In the future, proposed attribute level prediction model can be divided into four types. (i) First of all in this study a single category of durable goods is considered, further proposed model can be tested on multiple durable goods those consumers searched in various website.(ii) A generic Vader rule-based sentiment dictionary has been used in this analysis to find the reviews polarity. The sentiment analysis can be extended by proposing a personalized sentiment dictionary for each category products. Further, utilitarianism performs a similar role as that of sentiment analysis which can be explored in the future research. (iii) Basically, in this prediction model, a pilot study has been evolved by only considering the three months' consumers searched and screening reviews data.

In future, big data analysis can be involved for data collection, sentiment analysis, social network mining, and regression analysis separately or jointly to do the analysis in more sophisticated way. Specifically, Apache Spark, an open-source cluster-computing framework, can be performed for extracting, cleaning, storing and integrating the huge amount of user generated contents, and implementing the prediction model followed by sentiment analysis, social network mining, and regression analysis. (iv) In addition, some more influential factors like sales, discount, offers, deals, seasonality, etc. can also be incorporated to perform the model more efficiently and predict adequately.

VII. CONCLUSION

It is concluded after performing thorough Exploratory Data analysis which include Stats models which are computed to get accuracy and also Heat maps which are computed to get a clear understanding of the data set (which parameter has most abundant effect on the study case) and its come to point of getting the solution for the problem statement being , that the retail shopkeeper should strategically plan on the marketing in such a way that he could offer complementary products on the purchase of that particular product and if in possibility of bearing expenses offer more quantity of the product at the same price.

REFERENCES

[1] Chen, J., Teng, L., Yu, Y., & Yu, X. (2016). The effect of online information sources on purchase intentions between consumers with high and low susceptibility to informational influence. Journal of Business Research, 69(2), 467–475.

[2] Godey, B., Manthiou, A., Pederzoli, D., Rokka, J., Aiello, G., Donvito, R., & Singh, R. (2016). Social media marketing efforts of luxury brands: Influence on brand equity and consumer behavior. Journal of Business Research, 69(12), 5833–5841.

[3] Lim, C. H., Kim, K., & Cheong, Y. (2016). Factors affecting sportswear buying behavior: A comparative analysis of luxury sportswear. Journal of Business Research, 69(12), 5793–5800.

[4] Lacroix, C., & Jolibert, A. (2017). Mediational role of perceived personal legacy value between consumer agentic generativity and attitudes/buying intentions toward luxury brands. Journal of Business Research.

[5] Banerjee, S., Bhattacharyya, S., & Bose, I. (2017). Whose online reviews to trust? Understanding reviewer trustworthiness and its impact on business. Decision Support Systems, 96, 17–26.

[6] Fan, Z.-P., Che, Y.-J., & Chen, Z.-Y. (2017). Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis. Journal of Business Research, 74, 90–100.

[7] Bronnenberg, B. J., Kim, J. B., & Mela, C. F. (2016). Zooming in on choice: How do consumers search for cameras online? Marketing Science, 35(5), 693–712.

[8] Culotta, A., & Cutler, J. (2016). Mining brand perceptions from twitter social networks. Marketing Science, 35(3), 343–362.

[9] Baker, A. M., Donthu, N., & Kumar, V. (2016). Investigating how word-of-mouth conversations about brands influence purchase and retransmission intentions. Journal of Marketing Research, 53(2), 225–239.

[10] Otterbring, T., Ringler, C., Sirianni, N. J., & Gustafsson, A. (2017). The Abercrombie & Fitch Effect: The Impact of Physical Dominance on Male Customers' Status-Signaling Consumption. Journal of Marketing Research.

[11] Pascual-Miguel, F. J., Agudo-Peregrina, Á. F., & Chaparro-Peláez, J. (2015). Influences of gender and product type on online purchasing. Journal of Business Research, 68(7), 1550–1556.

[12] Malc, D., Mumel, D., & Pisnik, A. (2016). Exploring price fairness perceptions and their influence on consumer behavior. Journal of Business Research, 69(9), 3693–3697.

[13] Lichters, M., Müller, H., Sarstedt, M., & Vogt, B. (2016). How durable are compromise effects? Journal of Business Research, 69(10), 4056–4064.

[14] Askalidis, G., Kim, S. J., & Malthouse, E. C. (2017). Understanding and overcoming biases in online review systems. Decision Support Systems, 97, 23–30.

[15] Hsu, C.-L., Yu, L.-C., & Chang, K.-C. (2017). Exploring the effects of online customer reviews, regulatory focus, and product type on purchase intention: Perceived justice as a moderator. Computers in Human Behavior, 69, 335–346.

[16] Maslowska, E., Malthouse, E. C., & Viswanathan, V. (2017). Do customer reviews drive purchase decisions? The moderating roles of review exposure and price. Decision Support Systems, 98, 1–9.

[17] Filieri, R. (2015). What makes online reviews helpful? A diagnosticity-adoption framework to explain informational and normative influences in e-WOM. Journal of Business Research, 68(6), 1261–1270.

[18] Fan, Z.-P., Che, Y.-J., & Chen, Z.-Y. (2017). Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis. Journal of Business Research, 74, 90–100.

[19] Kangale, A., Kumar, S. K., Naeem, M. A., Williams, M., & Tiwari, M. K. (2016). Mining consumer reviews to generate ratings of different product attributes while producing feature-based review-summary. International Journal of Systems Science, 47(13), 3272–3286.

[20] Elwalda, A., Lü, K., & Ali, M. (2016). Perceived derived attributes of online customer reviews. Computers in Human Behavior, 56, 306–319.

[21] Shi, X., & Liao, Z. (2017). Online consumer review and group-buying participation: The mediating effects of consumer beliefs. Telematics and Informatics, 34(5), 605–617.

[22] Zhao, K., Stylianou, A. C., & Zheng, Y. (2017). Sources and impacts of social influence from online anonymous user reviews. Information & Management.

[23] Lee, J. E., & Watkins, B. (2016). YouTube vloggers' influence on consumer luxury brand perceptions and intentions. Journal of Business Research, 69(12), 5753–5760.

[24] Liu, M., Pan, W., Liu, M., Chen, Y., Peng, X., & Ming, Z. (2017). Mixed similarity learning for recommendation with implicit feedback. Knowledge-Based Systems, 119, 178–185.

[25] Luan, J., Yao, Z., Zhao, F., & Liu, H. (2016). Search product and experience product online reviews: An eye-tracking study on consumers' review search behavior. Computers in Human Behavior, 65, 420–430.

[26] Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. Journal of Business Research, 69(2), 897–904.

[27] Khan, Z., & Vorley, T. (2017). Big data text analytics: an enabler of knowledge management. Journal of Knowledge Management, 21(1), 18–34.

[28] Pauleen, D. J., & Wang, W. Y. C. (2017). Does big data mean big knowledge? KM perspectives on big data and analytics. Journal of Knowledge Management, 21(1), 1–6.

[29] Sullivan, L. (2017). Google, Facebook, Microsoft Spent Hundreds Of Millions To Advertise In 2016. Retrieved from https://www.mediapost.com/publications/article/302460/bing-google-facebook-spent-hundreds-of-millions.html

[30] Deodhar, M., Ghosh, J., Saar-Tsechansky, M., & Keshari, V. (2017). Active Learning with Multiple Localized Regression Models. INFORMS Journal on Computing, 29(3), 503–522.

[31] Ozkara, B. Y., Ozmen, M., & Kim, J. W. (2016). Exploring the relationship between information satisfaction and flow in the context of consumers' online search. Computers in Human Behavior, 63, 844–859.

[32] Roscoe, R. D., Grebitus, C., O'Brian, J., Johnson, A. C., & Kula, I. (2016). Online information search and decision making: Effects of web search stance. Computers in Human Behavior, 56, 103–118.

[33] Dutta, C. B., & Das, D. K. (2017). What drives consumers' online information search behavior? Evidence from England. Journal of Retailing and Consumer Services, 35, 36–45.

[34] Jun, S.-P., & Park, D.-H. (2016). Consumer information search behavior and purchasing decisions: Empirical evidence from Korea. Technological Forecasting and Social Change, 107, 97–111.

[35] Ngo-Ye, T. L., Sinha, A. P., & Sen, A. (2017). Predicting the helpfulness of online reviews using a scripts-enriched text regression model. Expert Systems with Applications, 71, 98–110.

[36] Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., & Roy, P. K. (2017). Predicting the "helpfulness" of online consumer reviews. Journal of Business Research, 70, 346–355.

[37] Mohammadiani, R. P., Mohammadi, S., & Malik, Z. (2017). Understanding the relationship strengths in users' activities, review helpfulness and influence. Computers in Human Behavior, 75, 117–129.

[38] Qazi, A., Syed, K. B. S., Raj, R. G., Cambria, E., Tahir, M., & Alghazzawi, D. (2016). A concept-level approach to the analysis of online review helpfulness. Computers in Human Behavior, 58, 75–81.

[39] Wei, J., He, J., Chen, K., Zhou, Y., & Tang, Z. (2017). Collaborative filtering and deep learning based recommendation system for cold start items. Expert Systems with Applications, 69, 29–39.

[40] Smith, B., & Linden, G. (2017). Two Decades of Recommender Systems at Amazon. com. IEEE Internet Computing, 21(3), 12–18.

[41] Sujoy Bag, Manoj Kumar Tiwari ,"Predicting the consumer's purchase intention of durable goods: An attribute-level analysis", Journal of Business Research · January 2019.