PROJECT REPORT

PROPENSITY OF BUYING BY CUSTOMERS IN E-COMMERCE INDUSTRY

Submitted towards partial fulfillment of the criteria

for award of PGPBA by GLIEMR

<u>Submitted By:</u>

Akash Kapoor (Roll No 1503)

Anjali Gupta (Roll No 1506)

Course & Batch: **PGPBA APR 2014**
Mentor: **Ms. Sarita Digumarti**

Great Lakes Institute of Management

GREAT LAKES
GREAT LAKES IEMR, GURGAON
Global Mindset. Indian Roots.

815, Udyog Vihar, Phase V,
Gurgaon, India 122016
Tel: +91-124-4934000,
Fax: +91-124-4934001
www.greatlakes.edu.in/gurgaon

Acknowledgements

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. We would like to extend our sincere thanks to all of them. We are highly indebted to Ms Sarita Digumarti for her guidance and constant supervision as well as for providing necessary information regarding the project & also for the support in completing the project. Her experience and support guided us to make the research process look simple. She provided valuable inputs at every step of the project.

We would also like to express our gratitude towards Mr. Sanjay Sethi, Ms. Radhika Ghai, Mr. Nitin Agarwal & Analytics team of ShopClues.com for their kind co-operation which help us in completion of this project.

Last but not the least we wish to thank Prof. Bappaditya Mukhopadhyay, our course Director, for constant supervision, guidance and for being a source of inspiration in helping us to work on this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: March 15, 2015 Akash Kapoor

Place: Gurgaon Anjali Gupta

Certificate of Completion

I hereby certify that the project titled “Propensity of buying by customers in E-Commerce Industry” was undertaken and completed under my supervision by Akash Kapoor and Anjali Gupta, both students of second batch of Postgraduate Program in Business Analytics (PGPBAAPR2014).

Date: March 15, 2015 (Sarita Digumarti)

Place: Gurgaon Mentor

TABLE OF CONTENTS

List of Tables

List of Figures

Executive Summary

ShopClues.com is an online retail website, enabling small and medium-size merchants to transact online, positions itself as a marketplace for the smaller cities buyers and sellers. ShopClues is the first e-commerce website in India that operated on the managed marketplace model. They wanted to find out the likelihood of the users buying the products. The purpose of this study is to predict and in doing so hopefully provide a better understanding of individual-level online buying behavior. This will improve the conversion rate and in turn will decrease the acquisition cost of the customer.

The duration of data extracted was 10th Nov'2014 to 9th Dec'2014. There were several reasons to select this period. One of the major reasons was that the company has moved to Adobe Omniture from Google Analytics for their web analytics needs and the data captured in the new tool was starting from 10th Nov onwards. The data in Google analytics was at macro level and hence was not appropriate for the study. Secondly, the study started in December 2014 and hence the duration was maximum possible. Thirdly any seasonality in the data was avoided.

The research takes following categories to study the propensity – Demographics of the customer/prospect, Behavior of the customer/ prospect on website, Acquisition Channels by which customer has visited website and Outcome. Each of the categories is further drilled down to variable level present in the data base. Orders in outcome category are taken as response variable. Demography comprises of identification of the users registered on the website. Acquisition has browser information, domain and campaign information which brought the customer to website. Behavior comprises of 10 variables which includes how much time the customer spends on the site, how many pages he visits, what day he comes to the website and so on.

The study assesses the impact of all the variables by creating deciles and dummies on orders. Finally, it validates the model on another sample data. A logistic regression approach was chosen to model the relation between predictor variable and other category variables. Results transpire that 72% of the orders are placed by 50% of the customers. ShopClues should invest their marketing dollars on these 50% customers for better revenue and ROI.

1.0 Introduction

1.1 Title & Objective of the Study

"Propensity of buying by customer in Indian E-Commerce Industry" is the title of this study. The objective of this study is to predict–and in so doing hopefully provide a better understanding of–individual-level online buying behavior. This will improve the conversion rate and in turn will decrease the acquisition cost of the customer.

1.2 Need of the Study

Since 2012, e-commerce in India has grown the fastest in the Asia-Pacific region. With over 250 million internet users and growth being shown at an extremely fast rate, e-commerce in India is estimated to reach $20 billion in 2015, which is approx. eight times ($2.5 billion) the amount it was worth in 2009. Driven by strong adoption of cheaper smart phones and affordable data plans, the number of only mobile Internet users in India is expected to reach 213 million by June'2015. Overall users are expected to reach 357 million and will become number one in the world.

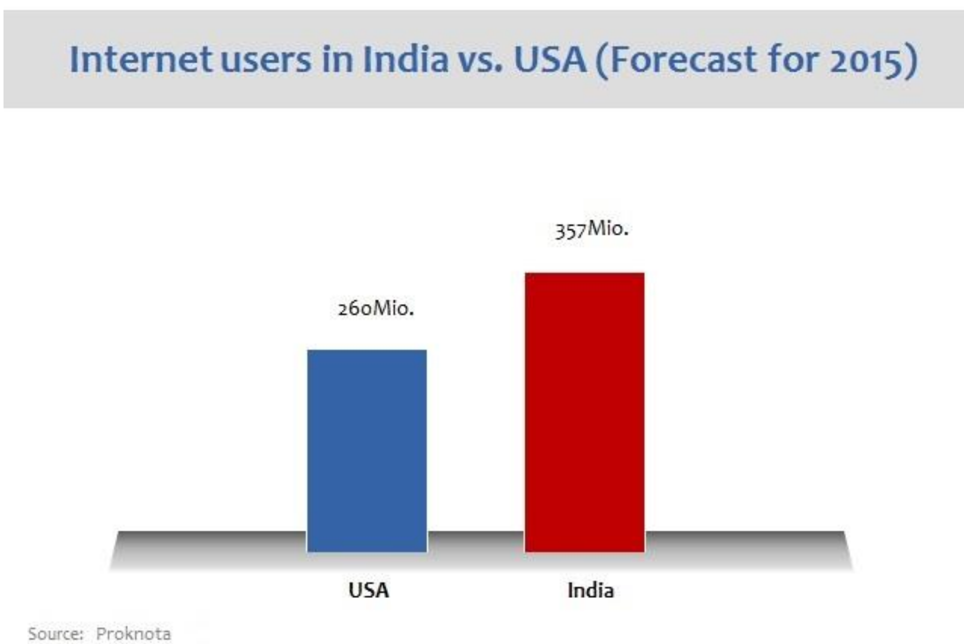

Figure-1.1 Internet Users

The trend expected for Indian Retail E commerce market i.e. B2C market shows continuous growth, with the rate of growth slowing down as penetration increases, but still estimated to be a healthy 25% in 2018.

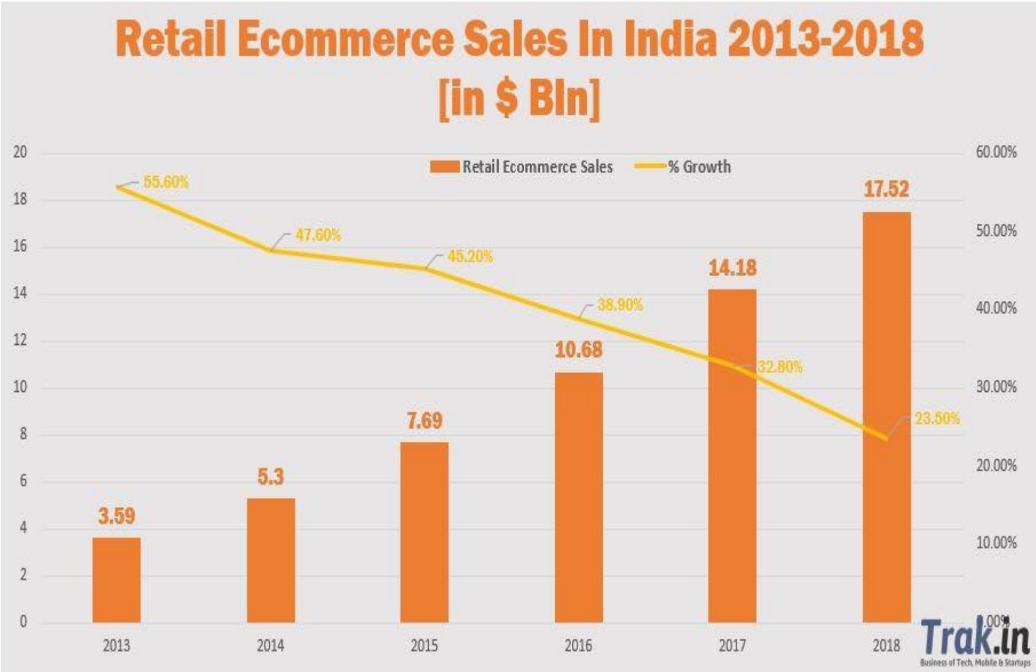

Figure-1.2 Ecommerce Sales

Key drivers for such growth in Indian e-commerce are:

- Increasing broadband Internet (growing at 20% MoM) and 3G penetration.
- Rising standards of living and a burgeoning, upwardly mobile middle class with high disposable incomes.
- Availability of much wider product range (including long tail and Direct Imports) compared to what is available at brick and mortar retailers.
- Busy lifestyles, urban traffic congestion and lack of time for offline shopping.
- Lower prices compared to brick and mortar retail driven by disintermediation and reduced inventory and real estate costs.
- Increased usage of online classified sites, with more consumers buying and selling second-hand goods.

- Evolution of the online marketplace model with sites like Jabong.com, Flipkart.com, Snapdeal.com and ShopClues.com.
- The government is also focusing on promoting e-commerce, innovation and entrepreneurship.

The above statistics makes a compelling case for companies to invest in web analytics to study how to increase their market share in the fast growing market. In fact understanding of the customer is a first-priority for any E-commerce company. E-commerce companies are depending on knowledge management systems for growth, customer acquisition and retention and to manage variable costs. A bird's eye view on conversion rate of major online ecommerce stores like amazon.com, jabong.com, Flipkart.com or ShopClues.com ranges from 0.78 to 3 %.

1.3 Company under Study

ShopClues.com is an online retail website, headquartered in Gurgaon, India. The company was founded in the Silicon Valley, USA in the year 2011 by an alumnus of Washington University and renowned Wall Street internet analyst Sandeep Aggarwal and eBay's former Global Product Head, Sanjay Sethi.

ShopClues is the first e-commerce website in India that operated on the managed marketplace model. This company enables small and medium-size merchants to transact online, positions itself as a marketplace for the smaller cities buyers and sellers. Unlike the top three marketplaces -- Amazon, Flipkart and Snapdeal -- they sells lesser known or unbranded items online, while the other top three focus more on branded stuff. They deals in more than 2 million products and guarantee authenticity of products, warranty and even ensure lowest price. They even have a record of catering to more than 42 million online visitors. The company has even more than 350 registered employees across the country. They offer wide range of branded products to consumers from every corner of the country.

Company initiate free delivery services at the doorsteps of customers through reputed courier services to nearly 10,000 cities across the country. They even initiate a 30-day return guarantee if unable to meet customer satisfaction. The leader marketplace for more than 1,00,000 small and local businesses seeking to reach the mass consumer in India's tier 2 and 3 cities. They lists products sold by about 100,000 merchants and intends to ramp up this number up to 300,000 by the end of this year.

1.4 Data Sources

ShopClues has two sources of data – one is the online data residing on the cloud and the other one is their own database. The online data captures all the variables listed above but the data is available post Oct'2014, as the company has implemented Omniture in Oct'2014 only. Google Analytics provides data on macro level and not User ID level.

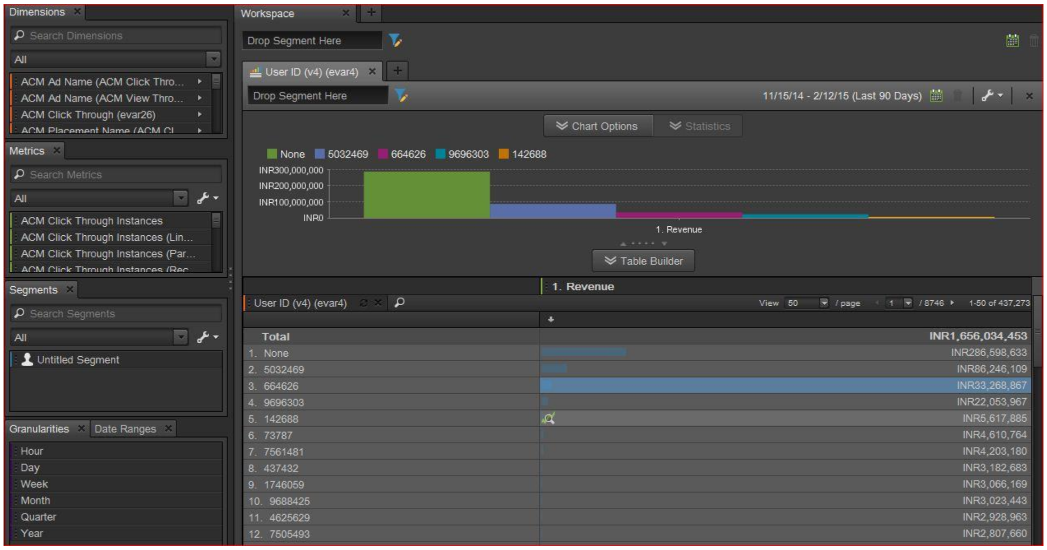

Figure-1.3 Ominture

1.5 Tools & Techniques

We have conducted our analysis in R and SAS environment. Data visualization is done with tableau and excel. Data extraction and organizing was the most challenging and time consuming part of the entire research. Most of the variables taken are dimensions in itself and constant metrics are taken against each dimension. Each dimension file is downloaded and cleaned in a similar format. Data extraction took long time due to limitation of software. Following steps were executed after the data was extracted from Omniture-

- Data Integration in R
- Data Profiling in SAS
- Dependent Variable Analysis
- Data Treatment in SAS
    - Missing values
    - Outliers
    - Redundant Variables
- Derived Variable Creation in SAS
- Bi-variate Analysis in SAS

- Fine Classing in SAS
    - Dummy Variable Creation for categorical variables
    - Bins for continuous variables
    -
- Coarse Classing in SAS
- Multi Co linearity in SAS, by using VIF
- Sampling- Test and Train Data
- Logistic Regression Modeling in SAS
- Statistical Analysis
- Lift Charts for order predictions
- Validation

1.6 Limitations

Predicting propensity of buying in E-Commerce has certain limitations. These includes-

- Online buying probabilities are usually low which can lead to a lack of predictive and explanatory power from models. 2% is the conversion rate in this industry, in current scenario.
- It is difficult to effectively account for what Web users do, and to what they are exposed while browsing a site.
- Online stores reach a diverse user population across many competitive environments, models of online buying must account for the corresponding user heterogeneity.
- Organizing Data is the biggest challenge in building such models-
    - Data is very noisy.
    - 70% of the transactions have no User ID's, as they log in as guests.
    - Demographic variables are not giving the correct information.
    - Data is having lot of multiple entries for the same variable at the same time.
    - De-duping of customers and accounts, as transactional systems usually do not provide safeguards to stop the generation of duplicate customer records.

2.0 Literature Review

Ying (2006) in his study "Essay on modeling consumer behavior in online shopping environments" examined online purchase behavior across multiple shopping sessions. Shopping cart abandonment is the bane of many e-commerce websites. He investigated abandoned shopping carts in an online grocery shopping setting. Specifically, he developed a joint model for the cart, order, and purchase quantity decisions. The interdependence between the three decisions is captured by the correlations between the error terms. Empirical analysis shows that not all abandoned shopping carts result in lost sales. Customers routinely pick up abandoned carts and complete the final orders. Among the factors that propel customers to continue with aborted shopping are the time of shopping, time elapsed since the previous visit, the number of items left in the abandoned cart, and promotion intensity. The study offers marketers important managerial implications on how to mitigate the shopping cart abandonment problem.

Another study “Know Your Buyer: A predictive approach to understand online buyers’ behavior” by Sandeep Pal attempted to study the online buying behavior of buyers and concluded that clicks, session duration, previous session, purchase session, clicks rate per session etc. were the factors influencing the buying behavior.

Different variables like visitor engagement, behavior, demographics, and acquisition channels can lead toward high propensity of a prospect’s willingness to buy. This paper provides insights on how the online data can support customized targeting, resulting in incremental increases in e-commerce revenues by advanced predictive modeling on visitors’ behavior.

Most of the previous researches have been focused on just the customer behavior. Although, this study also takes customer behavior into consideration, in addition it also considers other factors. The broad category of factors is mentioned below.

1. Demography
2. Acquisition
3. Behavior
4. Outcome

3.0 Data Description & Preparation

3.1 Identification of variables

After observing the website and understanding all the metrics and dimensions that were captured in Omniture, following variables were extracted –

Demographics - In this category, User ID was captured. It would have been good to have gender, age, education, income and address of the users but unfortunately, the company database did not have good data for these variables and hence User ID was the only relevant category. Company assigns user-id to the users who register on the company website. The customers who buy and do not register are not captured in the study. Location variable was also giving information as per geo-location at the time of order (which varies, if person is travelling) and not the permanent location. Thus it is also dropped.

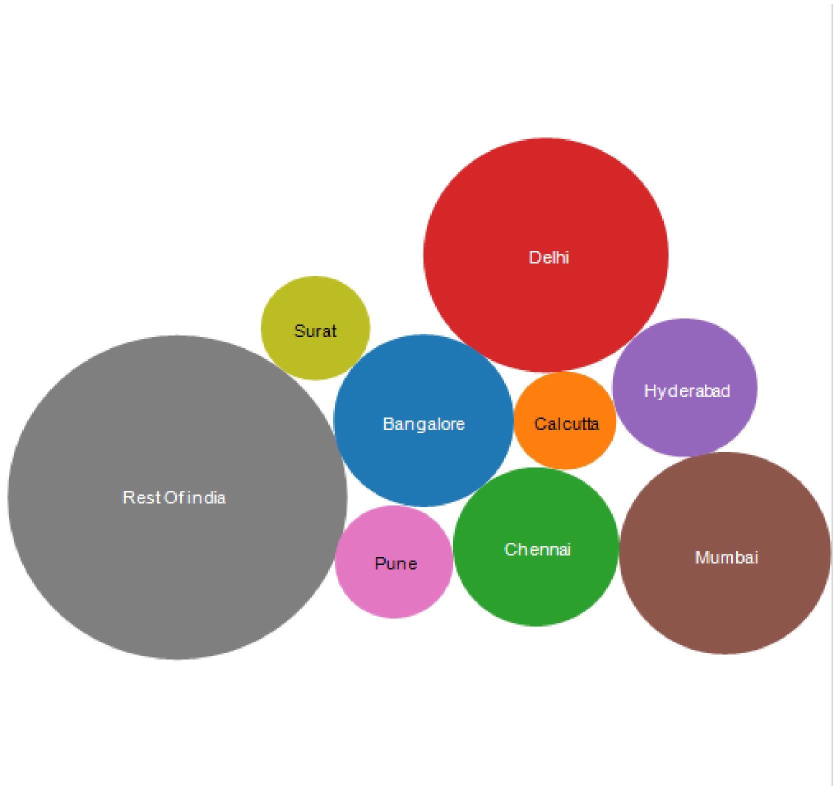

Figure-3.1 Geo Location

Insights- Company is focusing on Tier II and Tier III cities but 66%of the revenue is generated from Metros & Tier I cities.

Acquisition - How the customer is acquired- by which domain did he/she came through, what browser did he/she use, which campaign brought him to the website. After extraction it was observed that 99% of the User ID's are coming through different versions of Google Chrome, thus this variable is dropped. There were not many campaigns during this period and this variable is also having very few observations.

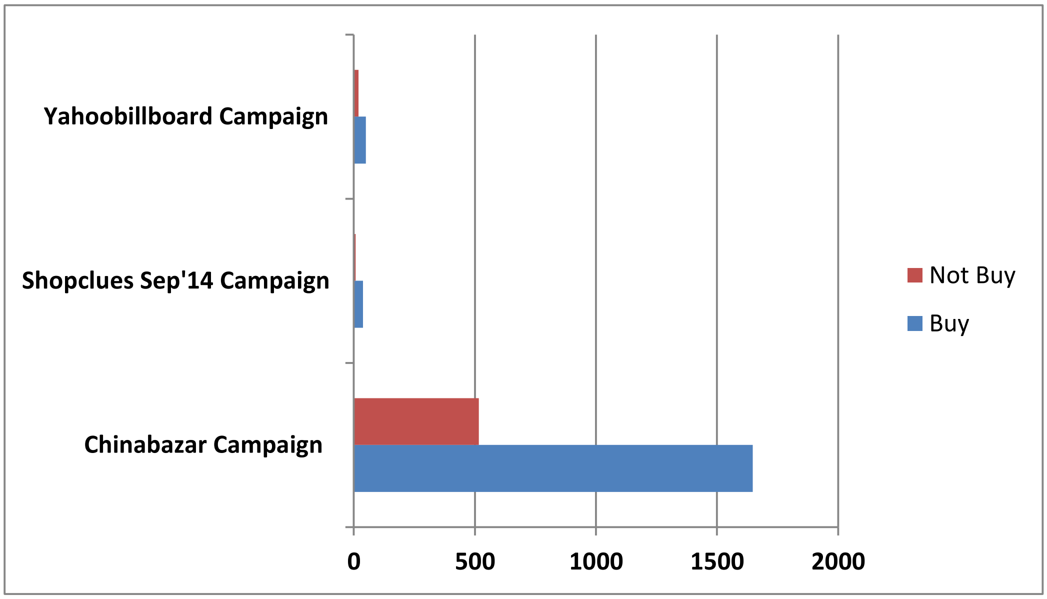

Figure-3.2 Campaigns

Behavior - This is very significant category for the study as the behavior on website combined with other categories provides a much better understanding of buying by customers relative to just studying the customer behavior. This category comprises of following variables-

- Entry hour of the day- This is giving information for every minute for each day. Thus it was dropped as it is having 25,92,000 minutes details and in each minute, there are number of User ID's.
- Time spent on website – It captures the total time spent by the user in the duration of the study which is from 10th Nov '2014 to 09th Dec'2014.

Insights- Figure 3.3 shows the time spent by users on the website. For product view, there is a decreasing trend as the time increases. Maximum orders took place when time spent is between 1 to 3 minutes. It decreases as the time spent on website increases.

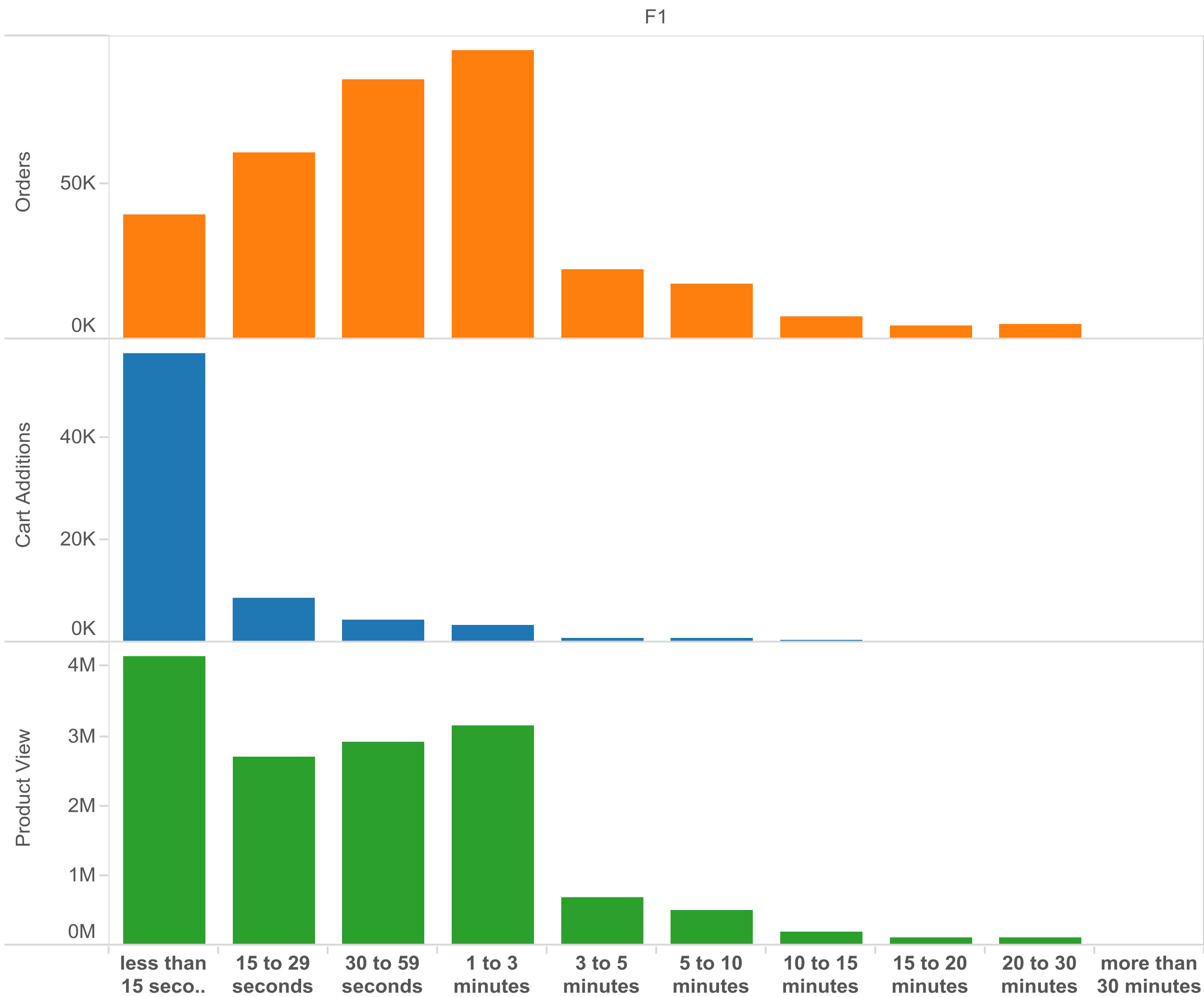

Figure-3.3 Time spent on website

- Visits - Total number of visits in the duration of our study.
- Entry day of the week - This variable provides the day of the week when user visited the e - retail market, whether it is Monday, Tuesday, Wednesday and so on.

Insights- Figure 3.4 shows the product views, cart additions, orders and revenue generated on different days of the week. As Sunday is a holiday and people have leisure time, maximum product views, cart additions and orders took place on this day. Monday is the second best day for orders, as the cart additions done on previous day are converted into orders.

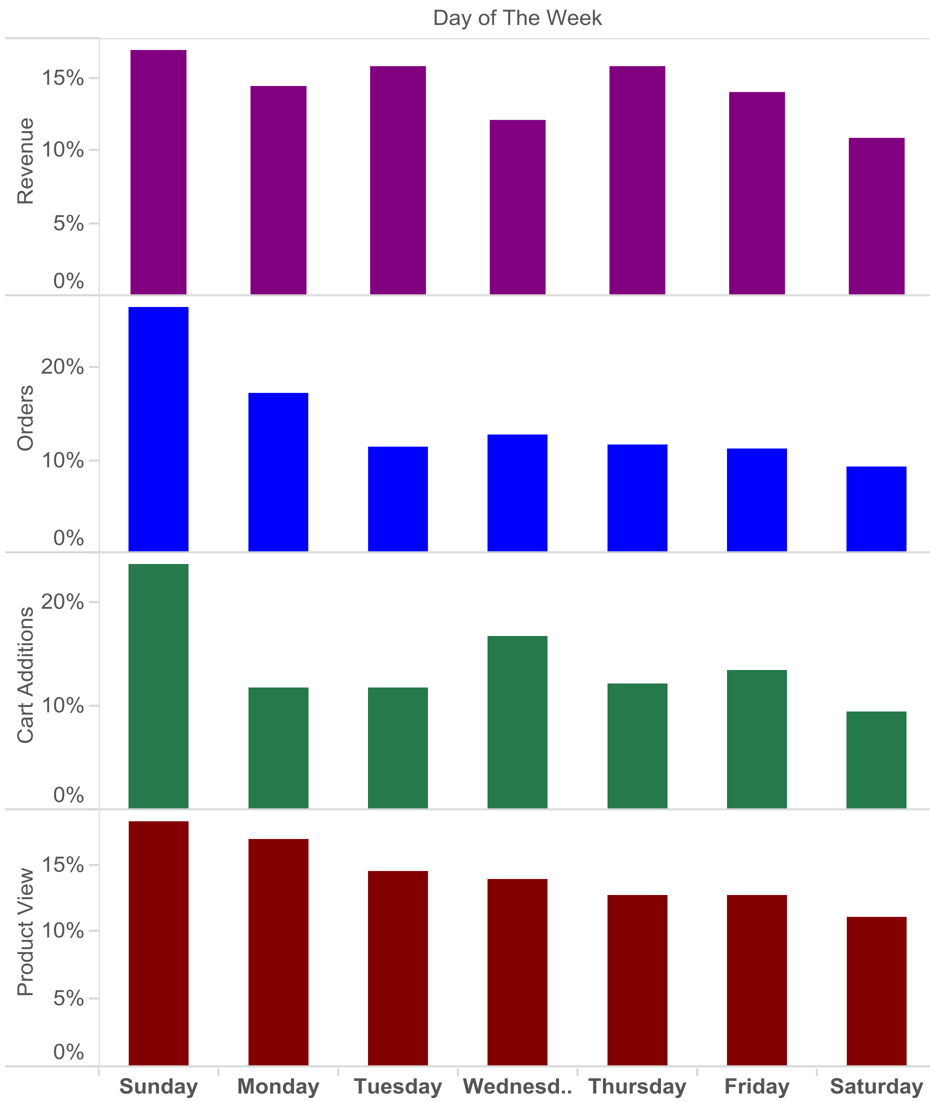

Figure-3.4 Entry day of the Week

- Page Views- Total pages viewed by the Users during the time of study.
- Return frequency - The length of time that passes between visits from returning visitors.

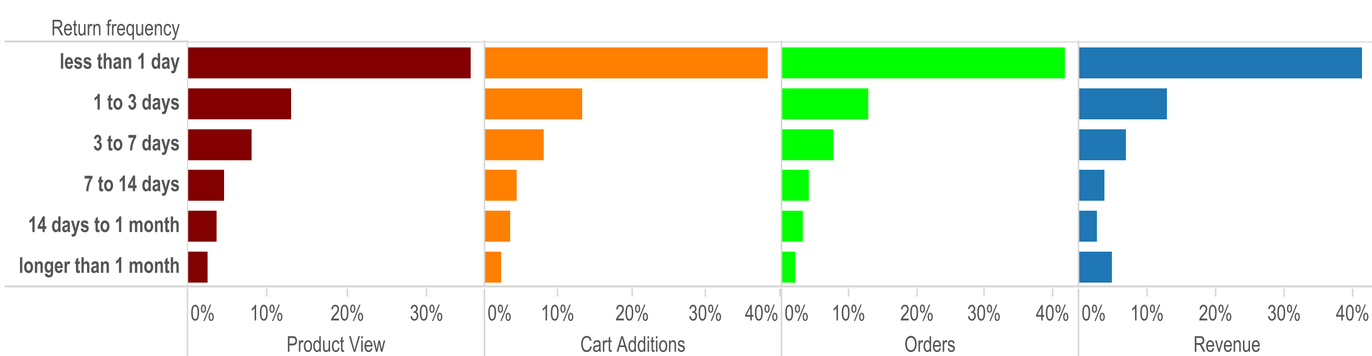

Figure-3.5 Recency

Return frequency	Product View	Cart Additions	Orders	Revenue
less than 1 day	35.5%	38.4%	41.7%	41.2%
1 to 3 days	12.9%	13.3%	12.8%	13.0%
3 to 7 days	8.0%	8.0%	7.6%	6.8%
7 to 14 days	4.5%	4.4%	4.1%	3.8%
14 days to 1 month	3.7%	3.4%	3.1%	2.7%
longer than 1 month	2.4%	2.2%	2.0%	4.8%

Table -3.1 Recency

Insights- The users visiting again within a day, orders more than other users. As time increases, the conversion % decreases. This can be attributed to – someone who sees a product and wants to buy it, would generally come sooner to buy it, and may be after comparing with other websites.

- Loyalty of Customer – There are three type of customers
    a. New Customer – This variable is described as 1 visit and 1 purchase by the user.
    b. Return Customer – This is described as more than 1 visit and 2 purchases by the user.
    c. Loyal Customer - More than 1 visit and 3+ purchases by the user is captured in this variable.

Insights – New customers account for maximum orders but return and loyal customers places orders directly.

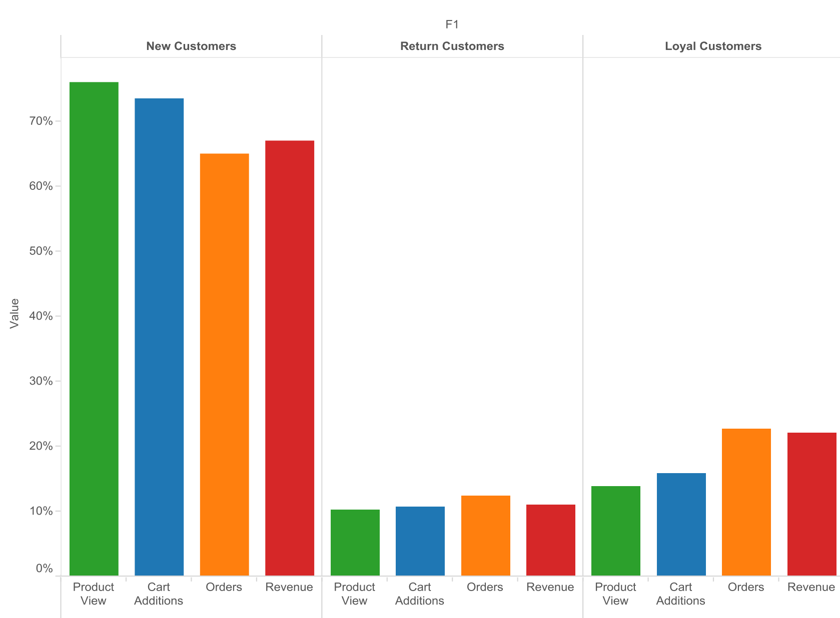

Figure -3.6 Loyalty

- Entry Page - This was the entry page by the user. Was it home page or fashion page or some other page.

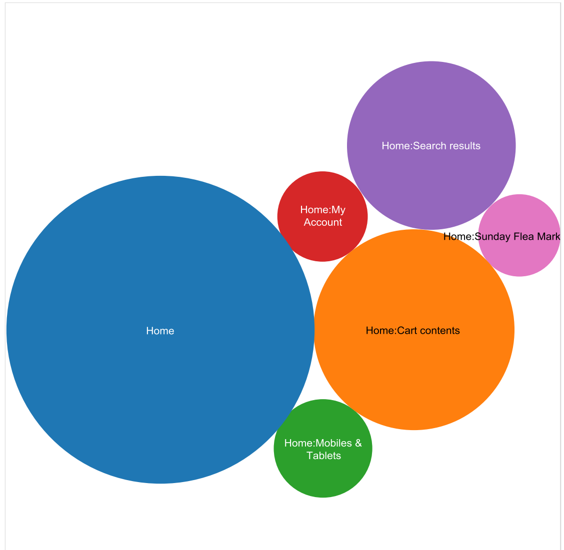

Figure -3.7 Entry Page

Insights - Above graph shows around 26% of the page views are home page which is way ahead of rest of the page views, it makes sense as most of the users logging in would come to home page. Next is cart contents page followed by search results page.

- Days before first purchase - How many days passed after the customer registered and before the first product was purchased. The study takes maximum of 6 days before first purchase, beyond that all the days are taken into one category.
- Days since last visit - How many days passed since the user last visited is captured in this variable.
- Product Views – This variable captures the number of days a user views the products.
- Cart Addition - Additions made to cart, on per day basis. The ones which did not convert into purchase on the same day are captured.

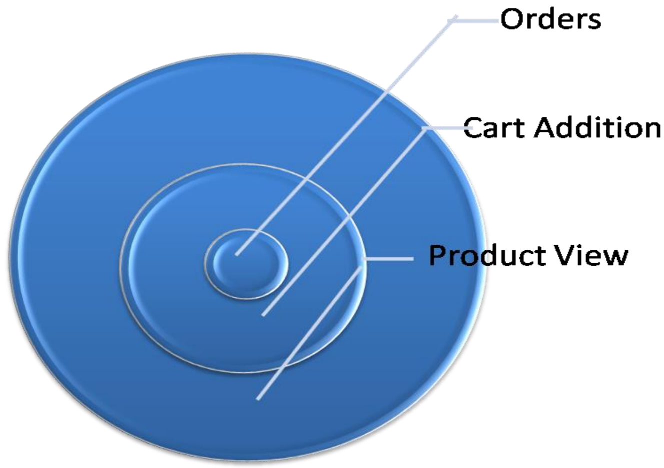

Figure-3.8 Three variables in order of their spread

Outcome - Orders is our target or response variable. This variable captures the number of times a user orders during one month on daily basis. For instance, a user orders on 10^{th} Nov, 15^{th} Nov and then on 20^{th} Nov, this variable for the above user would have value as 3 regardless of how many orders were placed. Order is the number of times the purchase event is set on daily basis.

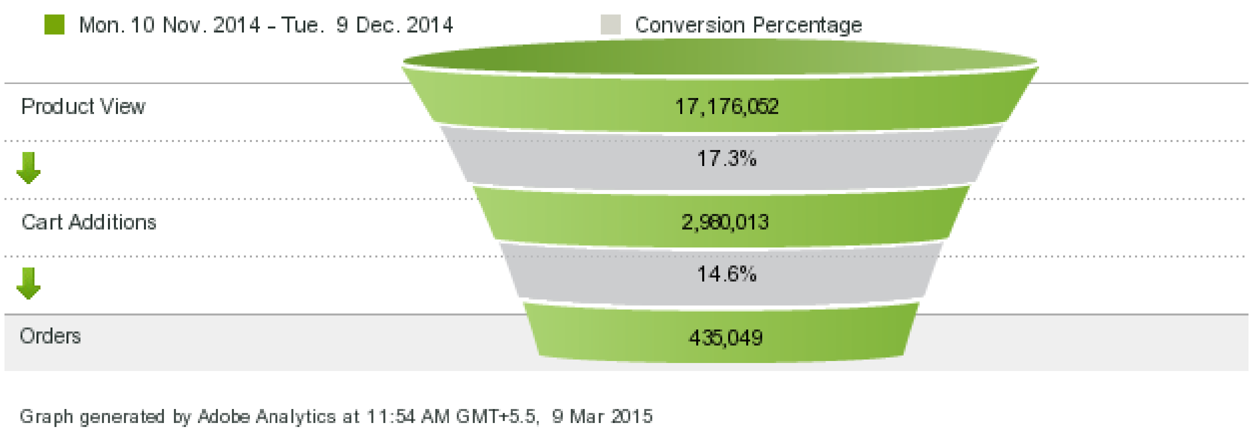

Figure-3.9 Conversion Funnel

Insights - Figure 3.9 shows the conversion funnel from product view to cart additions to Orders. ShopClues does better than the industry average of 2%. Its conversion percentage is 2.53%.

Conversions	10 Nov. 2014 - 9 Dec. 2014
Product View to Orders	2.53%
Cart Additions to Orders	14.60%
Order Averages	**10 Nov. 2014 - 9 Dec. 2014**
Average Orders per Product View	0.02
Average Orders per Cart Addition	0.14

Table -3.2 Conversions

Insights - Table 3.2 shows the conversions from Product views to Orders which is above the industry average for an ecommerce company. The percentage of users converting the products from cart to orders is 14.6% which means around 85% of users do not checkout. The above percentages are shown in numbers in the Order averages column.

3.2 Other Useful information

Revenue - Figure 3.10 shows the revenue for the duration of study and prior 4 weeks. The bump on Oct 20, 2014 is due to diwali sale. Other than that we see the sales pretty much the same. The total revenue generated during the period of study was Rs 467,772,052.

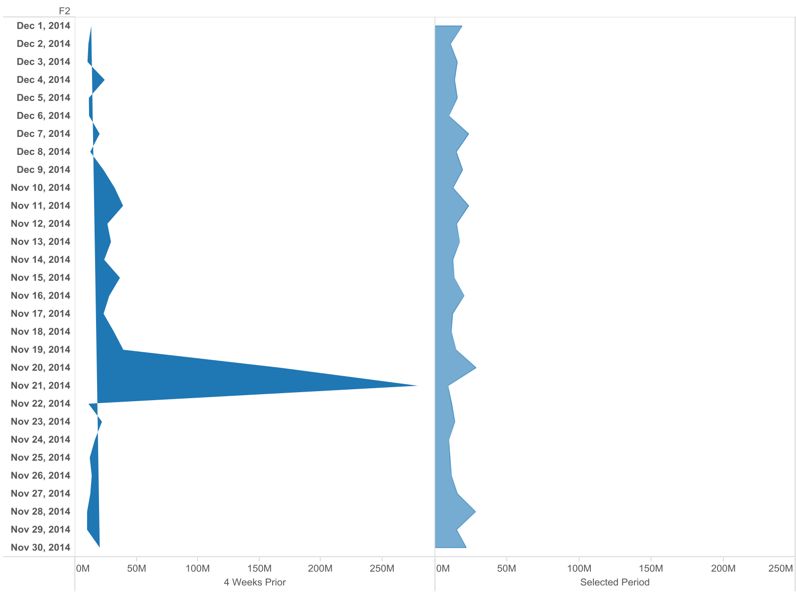

Figure 3.10 Revenue

Monthly Unique Visitors **-** Monthly unique visitors for the duration of the study and preceding four weeks are shown in the graph above. For some reason at the beginning of the period, maximum number of unique visitors is seen, it decreases gradually and at the end of the period it starts increasing again.

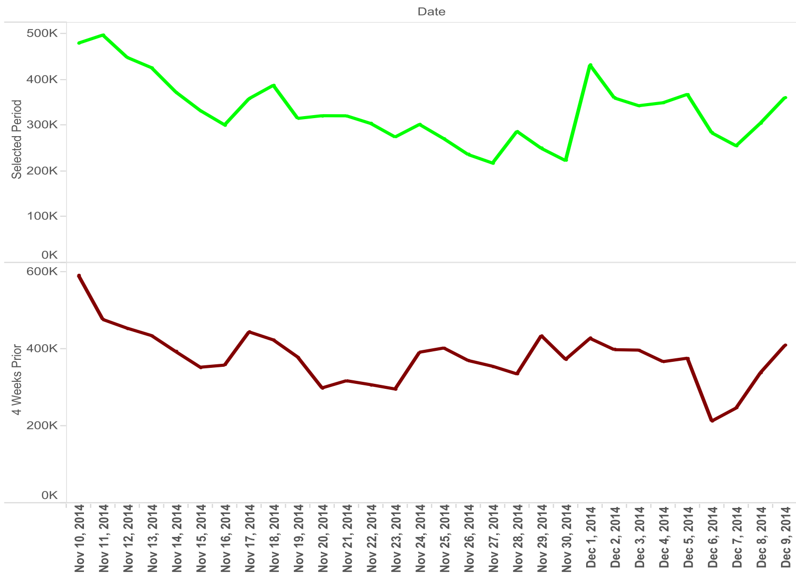

Figure 3.11 Monthly Unique Visitors

New Vs Repeat Customers - The graph below shows that revenue generated by repeat customers is more than new customers. The revenue generated by repeat customers makes more than 70% of the total revenue whereas new customers contribute around 30%. The visits do not show much difference, both the categories contribute around 50% of the visits.

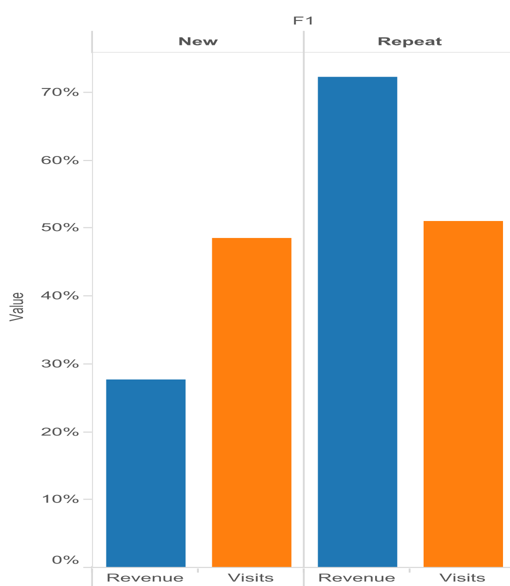

Figure -3.12 New Vs Repeat Customers

Merchant Category – Among top six categories, fashion has the maximum orders but Jewelry and watches is generating maximum revenue as average cost per order is high.

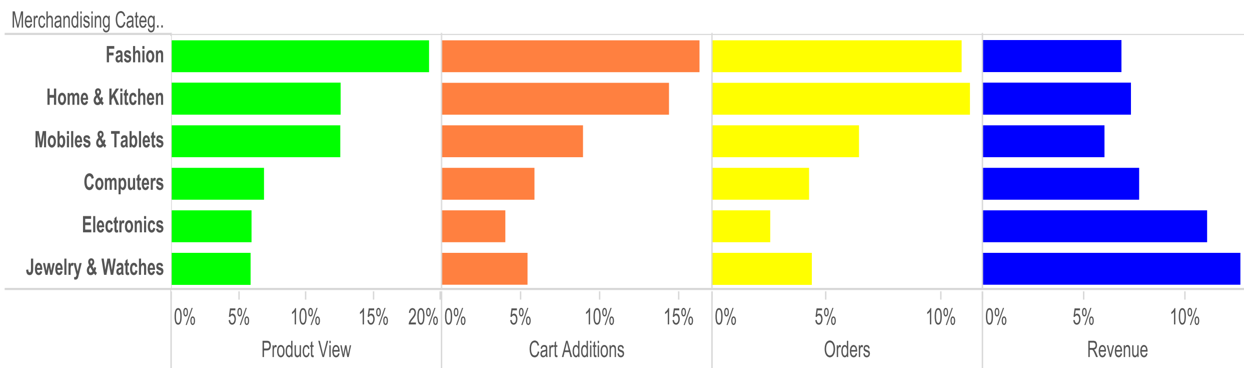

Figure -3.13 Merchant Categories

3.3 Data Dictionary

Explanations for the 36 original variables are presented in Table 3.3.

Placement (Col no)	Variable	Type	Description
1	User_ID	Num	Unique ID of Registered user
2	Product_View	Num	Occurs when the product detail page is viewed on daily basis
3	Cart_Additions	Num	Additions in cart per day basis
4	Orders	Num	Order is the number of times the purchase event is set on daily basis.
5	Ave_Time_on site	Num	Average time spent on website during the month
6	Ave_pages_per_visit	Num	Derived variable (page views/no of visits)
7	Mon	Num	Entry day of the week
8	Tues	Num	Entry day of the week
9	Wed	Num	Entry day of the week
10	Thurs	Num	Entry day of the week
11	Fri	Num	Entry day of the week
12	Sat	Num	Entry day of the week
13	Sun	Num	Entry day of the week
14	New_cust	Num	1 visit and 1 purchase
15	Return_Cust	Num	More than 1 visit and 2 purchases

Placement (Col no)	Variable	Type	Description
16	Loyal_Cust	Num	More than 1 visit and 3+ purchases
17	RF_lessthan1day	Num	the length of time that passes between visits from returning visitors, and the number of visits that fall into each time length category
18	RF_1to3days	Num	
19	RF_3to7days	Num	
20	RF_7to14days	Num	
21	RF_morethan14days	Num	
22	Days_Before_First_Purchase	Num	no of days between the first time customers visit and when they finally make a purchase.
23	Days_Since_last_Visit	Num	Determines the number of days since a user last visited
24	Home	Num	Entry Page
25	Sunday_Flea_Market	Num	Entry Page
26	ShopClues_Offers	Num	Entry Page
27	Search_Results	Num	Entry Page
28	Sunday_Flea_Market_Sale	Num	Entry Page
29	Super_Saver_Bazar	Num	Entry Page
30	Fashion_Bollywood_Store	Num	Merchant Category
31	My_Account	Num	Entry Page
32	Cart_contents	Num	Entry Page
33	Black_Friday	Num	Entry Page
34	Click_Through_Chinabazar	Num	Campaign (ACM Click thru)
35	Click_Through_Shopclues_sep14	Num	Campaign (ACM Click thru)
36	Click_Through_Yahoobillboard	Num	Campaign (ACM Click thru)

Table -3.3 Data Dictionary

Each of the variables is assigned a best (primary) category.

3.4 Quality Concerns

One concern regarding data quality comes from the high percentage of duplicate values in variables like location. Geo-location is captured, which varies if the person is travelling continuously. Most of the demographic variable fields are empty or have messy data. Date of Birth is having values like year 2020 etc. Time spent in a day on website (given in seconds) is more than 24 hrs. Most of the transactions take place as Guest User. After plummeting all such variables and observations, the data was merged. There were 36 variables with 2,60,725 observations (User ID).

3.5 Data Preparation

Variables Transformation

- Product Views, Cart additions & Orders are transformed on basis of day. Thus maximum of these 3 variables can be 30 only (study is for 30 days).
- Average pages per visit is derived from total no of pages seen during the 30 days span, divided by total number of visits.
- Average time spent onsite is calculated by total time divided by total number of visits.
- Dummy variables are created for days before 1st purchase and days since last visit.

Missing values and Outliers

- **Average Time spent on site** - 499 observations had missing values, which is approx 2% of the entire data. This was replaced by the median which is 762 seconds. There were 5 outliers on both the extremes. The values on the higher end are replaced by 4215 sec (99th percentile value) and the values on the lower end are replaced by 55 sec.

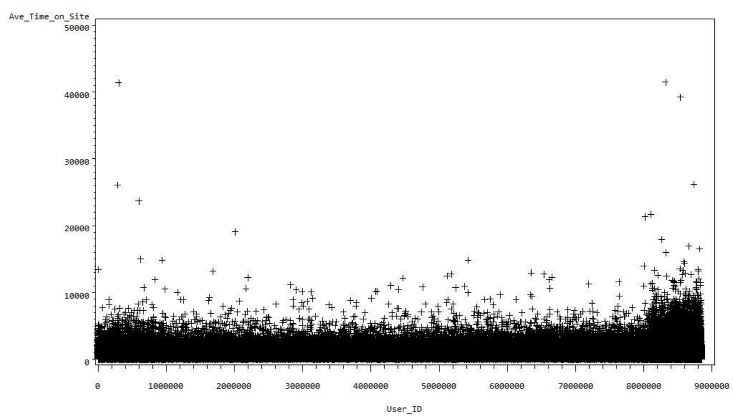

Figure -3.14 Average time on site

- **Days before 1st purchase**- The observations where users have made their first purchase after visiting the site for 7 or more than 7 days are shown as missing. These were taken as greater than 7 and were converted to dummies which will be discussed later in the section.

- **Days Since last visit**- Similar to the observations in previous variable. This variable too had missing values which were actually those users who had visited after 6 or more than 6 days. These were also taken care off while creating dummies.
- **Average pages per visit**- There were five extreme values on both the ends. They were replaced by the cut off value on both the ends. It was 2 on the lower end and values higher than 66 were replaced by 66.

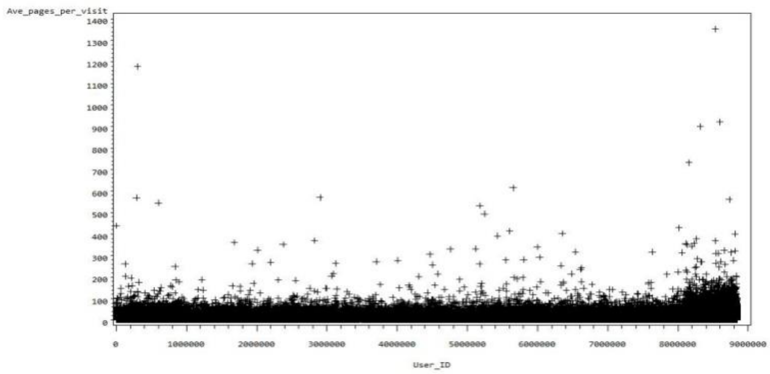

Figure -3.15 Average pages per visit

Bivariate Analysis

There are only two continuous variables – Average time on site and average pages per visit. The SAS visualization shows that both are directly propotional to each other.

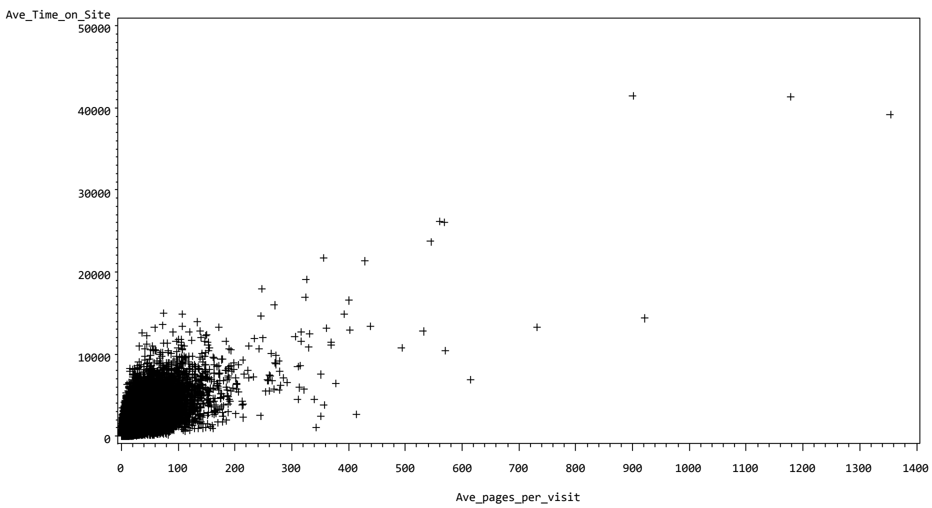

Figure -3.16 Bivariate Analysis

Fine Classing

Maximum variables are discrete and only two of them are continuous i.e. Avg time per visit, average pages per visit. Whether it is discrete or continuous, binning is better to analyze the data as it reduces the spread. In order to do that, deciles were created for the following variables - Avg time per visit, average pages per visit, product views, cart additions. While creating deciles, the rule was that no bin should overlap and each bin should have atleast 5% of the poulation. For product views, 6 bins are created, cart additions have 4 bins, average pages per visit and average time per visit, each has 10 bins.

These bins were further converted into dummies. The variable with “n” bins are converted to “n-1” dummies. Total no of variables are 68 independent variables and one dependent variable.

Coarse Classing

On checking the collinearity, no collinearity was observed between the independent variables. Some of the variables were found with high VIF. The ones with highest VIF (VIF>3) were dropped based on lower information value (IV). Three variables are dropped during this procedure and left with 65 independent variables.

4.0 Modeling

The modeling set was split into two parts -70% modeling data and 30% validation data. Logistic regression was run on modeling data.Using the logistic procedure from SAS and finding out the two variables with maximum **'p'** values. Then drop the one having low **Information Value**. Also it was observed that 'same day visit' and 'RF_less than 1day' are having multicollinearity. Thus 'RF_less than 1day' is dropped. Repeated the logistic procedure for 11 times and dropped the variables on the basis of high '**p**' values and low Information value.. The resulting significant predictors, their p-values and the estimated signs for numeric predictors are shown in Table 4.1.

Parameter	Estimate	P-Value	Parameter	Estimate	P-Value
Intercept	2.6836	<.0001	samedayvisit	0.0944	<.0001
Mon	0.3433	<.0001	Av_time_bin1	-1.0744	<.0001
Tues	0.2289	<.0001	Av_time_bin2	-0.4953	<.0001
Wed	0.2155	<.0001	Av_time_bin3	-0.4179	<.0001
Thurs	0.2364	<.0001	Av_time_bin4	-0.3399	<.0001
Fri	0.268	<.0001	Av_time_bin5	-0.3367	<.0001
Sat	0.2209	<.0001	Av_time_bin6	-0.2663	<.0001
Sun	0.5737	<.0001	Av_time_bin7	-0.163	<.0001
New_Cust	0.5787	<.0001	Av_time_bin8	-0.1297	<.0001
Return_Cust	0.7076	<.0001	Av_time_bin9	-0.1012	<.0001
Loyal_Cust	0.6926	<.0001	Av_pgvisit_bin1	-0.68	<.0001
RF_1to3days	-0.079	<.0001	Av_pgvisit_bin3	-0.1551	<.0001
RF_7to14days	0.0585	0.0016	Av_pgvisit_bin4	-0.1786	<.0001
RF_morethan14days	0.105	<.0001	Av_pgvisit_bin5	-0.1529	<.0001
Home	-0.148	<.0001	Av_pgvisit_bin6	-0.1102	<.0001
Sunday_Flea_Market	0.1787	<.0001	Av_pgvisit_bin7	-0.0804	0.0004
ShopClues_Offers	-0.1763	<.0001	Av_pgvisit_bin8	-0.0463	0.0189
Search_results	0.1856	<.0001	pro_view_bin2	-3.6538	<.0001
Sunday_Flea_Market_S	0.2409	<.0001	pro_view_bin3	-3.4619	<.0001
Super_Saver_Bazar	0.3067	<.0001	pro_view_bin4	-3.3776	<.0001
Fashion_Bollywood_St	-0.3422	<.0001	pro_view_bin5	-3.303	<.0001
My_Account	-0.2783	<.0001	cart_add_bin1	-0.3858	<.0001
Cart_contents	0.4641	<.0001	cart_add_bin3	0.3052	<.0001

Table-4.1 Logistic Output

The output of the data is shown in section-1 of Appendix. Table– 1 shows the name of the data set, response variable and number of response levels. Number of observations are shown in table-2. They are 182563. Table-3 shows that 99977 users had ordered during one month of study and 82586 users did not order.

Table-5 tests that at least one of the beta coefficient is not zero.This condition is satisfied aas the p- value <0.0001. Table-6 shows all the significant variables and the beta coefficents for them. Table-7 shows the odds ratio or the coefficients for variables in multiplicative model. We will interpret them in the results section.Table-8 shows another important statistics,the concordant and discordant pairs. It shows a significant,81.7% concordant pairs.The c value of 81.8 implies that the model is around 27% better than the random model which had 55% users with orders.

4.1 Results

Odds Ratios

In this section we will interpret the odds ratio for variables. This would provide much better interpretation than the additive model or the log odds model. The variable Mon or the entry day of the week as Monday has odds ratio of 1.41 which implies that a user logging in on Monday increases the odds ratio by 41%. If the entry day is Tuesday the odds ratio is increased by 25% as the odds ratio is 1.257. Similarly for other days, the odds ratio can be interpreted. Maximum increase in odds ratio is on Sunday, it’s a 77% which also makes sense as most people have holiday and they can order during the leisure time (also proved from data visualization).

Entry day	Odds Ratio
Mon	1.41
Tues	1.257
Wed	1.24
Thurs	1.267
Fri	1.307
Sat	1.247
Sun	1.775

Table 4.2 Entry Day

Next set of variables is New customer, Returning customer and Loyal customer. The value of odds- ratio for these variables is 1.78, 2.02 and 1.99 respectively. This implies that a new customer has 78% increased odds ratio of buying a product, its 102% increase in case of returning customer and 99% in case of loyal customer.

Loyalty	Odds ratio
New_Cust	1.784
Return_Cust	2.029
Loyal_Cust	1.999

Table 4.3 Customers

Return frequency has three significant variables return frequency 1 to 3 days, 7 to 14 days and more than 14 days. The odds ratio for these variables is 0.92, 1.06 and 1.11 respectively. This means a user returning between 1 to 3 days decreases the odds ratio of buying a product by 8%. In case of user returning between 7 to 14 days the odds ratio increases by 6 % whereas the increase is 11% if the return frequency is more than 14 days. Same day visit has almost 10% increased odds ratio.

Recency	Odds ratio
RF_1to3days	0.924
RF_7to14days	1.06
RF_morethan14days	1.111

Table 4.4 Return Frequency

Entry page is the next set of variables to be analyzed. The users entry at Home page, ShopClues_offers page, Fashion_Bollywood_St and My_Account page have decreased odds ratio by 13.8%, 14%, 29% and 24% respectively. In contrast Sunday_Flea_Market, Search_results, Sunday_Flea_Market_S, Super_Saver_Bazar and cart_contents show an increase in odds ratio by 19.6%, 20%, 27%, 36% and 59% respectively.

Entry Page	Odds ratio
Home	0.862
Sunday_Flea_Market	1.196
ShopClues_Offers	0.838
Search_results	1.204
Sunday_Flea_Market_S	1.272
Super_Saver_Bazar	1.359
Fashion_Bollywood_St	0.71
My_Account	0.757
Cart_contents	1.591

Table 4.5 Entry Page

Average time spent on the website has a trend, as the time spent on the website increases, the odds ratio of buying a product increases. Although, overall the time spent on the website shows a decrease in odds ratio but the decrease reduces as the time increases. See Table 4.5

Avg time spent bins	Odds Ratio
2- 228	0.341
229-362	0.609
363-492	0.658
493-624	0.712
625-761	0.714
762-914	0.766
915-1102	0.85
1103-1366	0.878
1367-1866	0.904

Table 4.6 Avg Time

The 2-228 range bin has a decreased odds ratio by around 66% which means a user spending on an average 2-228 minutes on a website in a month has decreased odds ratio of buying a product by 66%. The decrease odds ratio reduces as the user spends more time on the website, to the extent that the decrease in odds ratio goes to 10% for the users spending 1367 to 1866 minutes ina month on the website.

Average pages per visit are shown is table 4.6 The odds ratio for a user visiting 1 to 5 pages is decreased by around 49%. The odds ratio increases as the user visits more pages on the website. A user visiting 16 to 19 pages has decreased odds ratio of 4.5% which means a user visiting on an average 16 to 19 pages reduces the chances of buying by 4.5%.

Avg pages per visit	Odds Ratio
greater than 1 but less than 5	0.507
greater than 6 but less than 8	0.856
8 or more than 8 but less than 10	0.836
10 to 11	0.858
12 to 13	0.896
14 to 15	0.923
16 to 19	0.955

Table 4.7 Avg Pages

Product-view bins are shown in Table 4.7. The users viewing products one day in a month have a decreased odds ratio of 97.4% whereas user with product views on 4 to 5 days decreases odds ratio by 96.3%

Product Views	Odds Ratio
1	0.026
2	0.031
3	0.034
4 to 5	0.037

Table 4.8 Product Views

Cart Addition bins are shown in Table-4. The users not adding products in cart addition have decreased odds ratio by 32%, in contrast the users adding products on 2 days in a month have increased odds ratio of 35.7% which means a user adding products in cart on 2 days ina month and not ordering the same day has increased chances of buying a product by 35.7%

Cart Addition	Odds Ratio
0	0.68
2	1.357

Table 4.9 Cart Addition

Lorenz Curve

This is a graphical representation of the cumulative distribution function of the empirical probability distribution.

Decile	Observations	Events Predicted	Events Expected	Lift	Cumulative% of Expected	Cumulative% of Predicted
1	18256	16911	9997	1.69161	10%	16.91%
2	18256	17282	9997	1.728721	20%	34.20%
3	18256	15095	9997	1.509955	30%	49.30%
4	18256	12344	9997	1.234772	40%	61.65%
5	18256	10146	9997	1.014906	50%	71.79%
6	18256	8430	9997	0.843254	60%	80.23%
7	18256	7126	9997	0.712815	70%	87.35%
8	18257	5878	9998	0.587945	80%	93.23%
9	18257	4690	9998	0.469116	90%	97.92%
10	18257	2075	9998	0.207551	100%	100.00%

Table – 4.10 Deciles

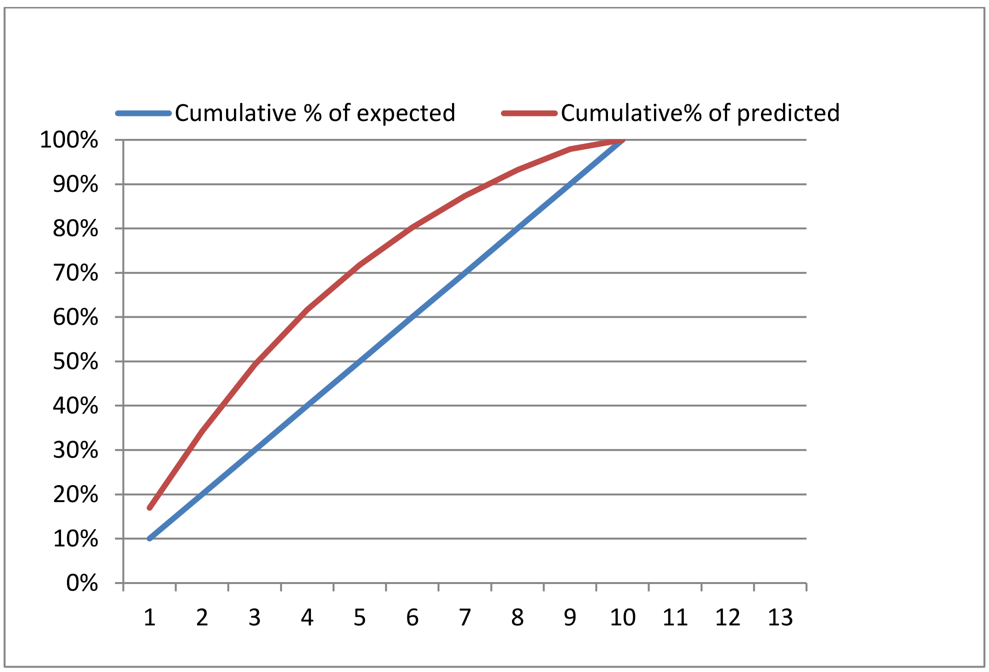

Figure 4.1 Lorenz Curve (Test)

Figure 4.1 shows the lift for the predicted model and the straight line shows the cumulative percentage of expected events. By the fifth decile the model shows a lift of around 22% which impiles that compared to random model, the new model will capture around 72% of cutomers buying the product in first 50% of the population.

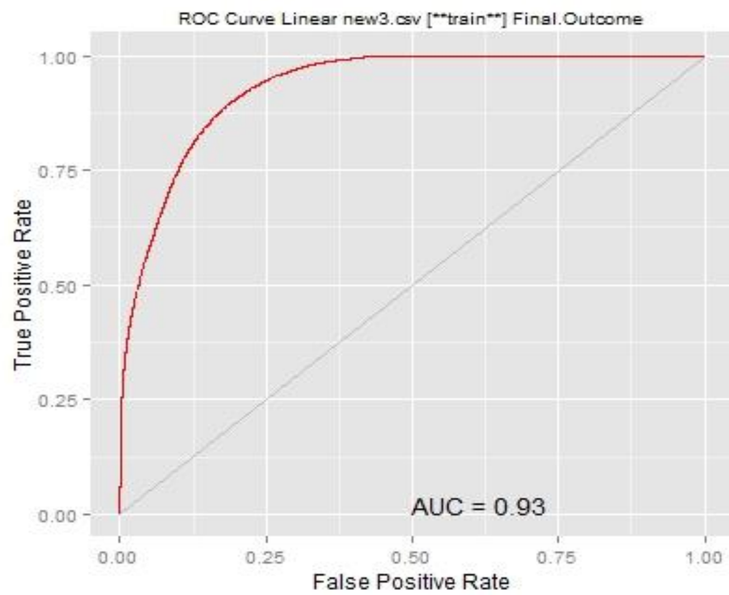

Figure 4.2 ROC Curve (Test)

The ROC curve (Figure-4.2)shows high true positivity which again shows that it's a good model, with an Area Under Curve of 0 .93 Area under curve measures the discrimination, that is the ability of the model to correctly classify people who will buy and who will not buy.

4.2 Validation

Same model is run on validation data. Number of observations in this dataset are 78162, out of which 54.47% are having positive outcome. The resulting significant predictors, their p-values and the estimated signs for numeric predictors are shown in Table 4.11.

Parameter	Estimate	P-Value	Parameter	Estimate	P-Value
Intercept	2.7184	<.0001	samedayvisit	0.1322	<.0001
Mon	0.3708	<.0001	Av_time_bin1	-1.1208	<.0001
Tues	0.2105	<.0001	Av_time_bin2	-0.5482	<.0001
Wed	0.2289	<.0001	Av_time_bin3	-0.4441	<.0001
Thurs	0.27	<.0001	Av_time_bin4	-0.3844	<.0001
Fri	0.2727	<.0001	Av_time_bin5	-0.3781	<.0001
Sat	0.2436	<.0001	Av_time_bin6	-0.2959	<.0001
Sun	0.588	<.0001	Av_time_bin7	-0.2528	<.0001
New_Cust	0.6244	<.0001	Av_time_bin8	-0.1589	<.0001
Return_Cust	0.7399	<.0001	Av_time_bin9	-0.0994	0.0059
Loyal_Cust	0.6982	<.0001	Av_pgvisit_bin1	-0.6963	<.0001

RF_1to3days	-0.0856	<.0001	Av_pgvisit_bin3	-0.0927	0.0155
RF_7to14days	0.0785	0.0052	Av_pgvisit_bin4	-0.1675	<.0001
RF_morethan14days	0.1347	<.0001	Av_pgvisit_bin5	-0.1605	<.0001
Home	-0.2018	<.0001	Av_pgvisit_bin6	-0.1025	0.0015
Sunday_Flea_Market	0.2224	<.0001	Av_pgvisit_bin7	-0.0981	0.0045
ShopClues_Offers	-0.1908	<.0001	Av_pgvisit_bin8	-0.0882	0.0034
Search_results	0.1733	0.0002	pro_view_bin2	-3.7119	<.0001
Sunday_Flea_Market_S	0.2599	<.0001	pro_view_bin3	-3.496	<.0001
Super_Saver_Bazar	0.2343	<.0001	pro_view_bin4	-3.4973	<.0001
Fashion_Bollywood_St	-0.3729	<.0001	pro_view_bin5	-3.4017	<.0001
My_Account	-0.2236	<.0001	cart_add_bin1	-0.3966	<.0001
Cart_contents	0.5051	<.0001	cart_add_bin3	0.1961	<.0001

Table – 4.11 Logistic Output on Validation

The ROC curve and the Lorenz curve (Figure- 4.3 and Figure -4.4) were run on the model with validation data. The results were pretty much the same as the test data. By the fifth decile 72% of the customers buying the products are captured. This shows that the model is robust.

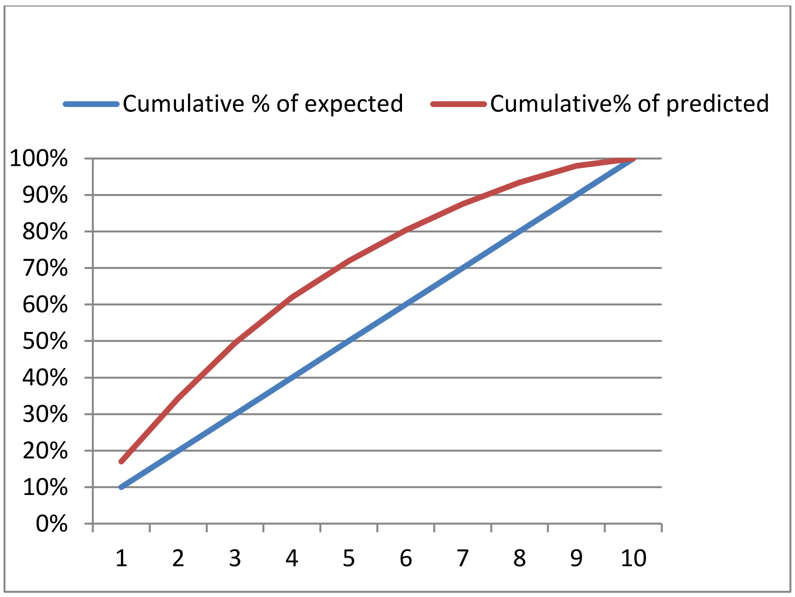

Figure 4.3 Lorenz Curve Validation

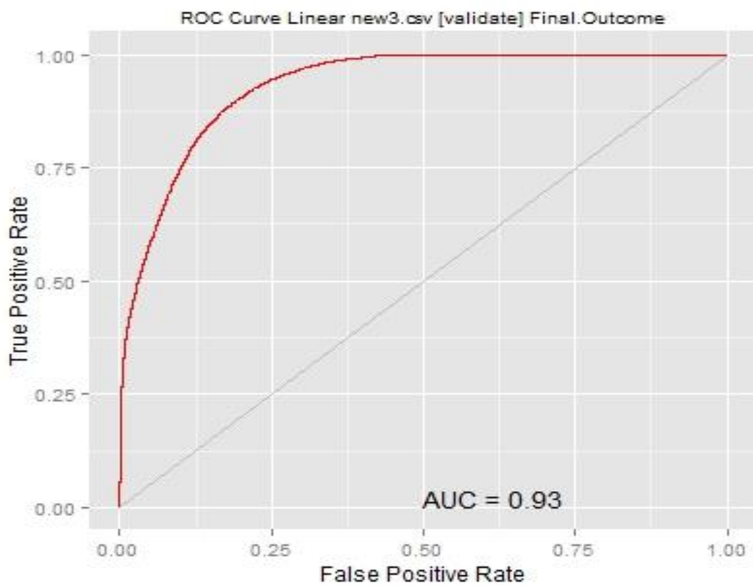

Figure 4.4 ROC Curve Validation

5.0 Conclusion

The logistic models discovered sets of variables bearing statistically significant impacts on the likelihood of buying. Based on the findings, it is seen that 72% of buying is done by 50% of the customers. ShopClues should invest their marketing dollars on these customers to up-sell and cross-sell their products. This will enhance their revenue and also provide better ROI for marketing dollars.

Appendix

Table-A1 Model Information

Model Information	
Data Set	FIRSTLIB.NEW_DEVLOP
Response Variable	outcome
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Table-A2 Observations

Number of Observations Read	78162
Number of Observations Used	78162

Table-A3 Response Profile

Response Profile		
Ordered Value	**outcome**	**Total Frequency**
1	1	99977
2	0	82586

Table-A4 Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	251428.87	190906.09
SC	251438.99	191371.37
-2 Log L	251426.87	190814.09

Table-A5 Likelihood Ratio

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	60612.784	45	<.0001
Score	49571.4191	45	<.0001
Wald	33461.7795	45	<.0001

Table-A6 Likelihood Estimates

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.6836	0.0405	4388.9536	<.0001
Mon	1	0.3433	0.0174	388.5051	<.0001
Tues	1	0.2289	0.0177	168.1737	<.0001
Wed	1	0.2155	0.0183	138.3044	<.0001
Thurs	1	0.2364	0.0182	168.223	<.0001
Fri	1	0.268	0.0183	215.5873	<.0001
Sat	1	0.2209	0.0187	138.9729	<.0001
Sun	1	0.5737	0.0185	958.2264	<.0001
New_Cust	1	0.5787	0.0189	936.5616	<.0001
Return_Cust	1	0.7076	0.0159	1991.9477	<.0001
Loyal_Cust	1	0.6926	0.0172	1630.7187	<.0001
RF_1to3days	1	-0.079	0.0144	30.1349	<.0001
RF_7to14days	1	0.0585	0.0185	9.9864	0.0016
RF_morethan14days	1	0.105	0.0172	37.2136	<.0001
Home	1	-0.148	0.0131	128.4338	<.0001
Sunday_Flea_Market	1	0.1787	0.0219	66.4453	<.0001
ShopClues_Offers	1	-0.1763	0.0275	41.1401	<.0001
Search_results	1	0.1856	0.0299	38.417	<.0001
Sunday_Flea_Market_S	1	0.2409	0.0326	54.5222	<.0001
Super_Saver_Bazar	1	0.3067	0.0347	78.0408	<.0001
Fashion_Bollywood_St	1	-0.3422	0.0577	35.1675	<.0001
My_Account	1	-0.2783	0.0219	161.3206	<.0001
Cart_contents	1	0.4641	0.0344	181.9532	<.0001
samedayvisit	1	0.0944	0.0132	51.1175	<.0001
Av_time_bin1	1	-1.0744	0.0285	1419.2496	<.0001

Av_time_bin2	1	-0.4953	0.0254	380.6719	<.0001
Av_time_bin3	1	-0.4179	0.025	280.4625	<.0001
Av_time_bin4	1	-0.3399	0.0248	187.3523	<.0001
Av_time_bin5	1	-0.3367	0.0251	180.0539	<.0001
Av_time_bin6	1	-0.2663	0.025	113.2103	<.0001
Av_time_bin7	1	-0.163	0.025	42.5708	<.0001
Av_time_bin8	1	-0.1297	0.0247	27.5508	<.0001
Av_time_bin9	1	-0.1012	0.0238	18.0872	<.0001
Av_pgvisit_bin1	1	-0.68	0.027	633.644	<.0001
Av_pgvisit_bin3	1	-0.1551	0.0251	38.2617	<.0001
Av_pgvisit_bin4	1	-0.1786	0.0193	85.402	<.0001
Av_pgvisit_bin5	1	-0.1529	0.0198	59.5708	<.0001
Av_pgvisit_bin6	1	-0.1102	0.021	27.6119	<.0001
Av_pgvisit_bin7	1	-0.0804	0.0225	12.7336	0.0004
Av_pgvisit_bin8	1	-0.0463	0.0197	5.5124	0.0189
pro_view_bin2	1	-3.6538	0.0293	15602.2745	<.0001
pro_view_bin3	1	-3.4619	0.0339	10402.0936	<.0001
pro_view_bin4	1	-3.3776	0.0417	6567.7011	<.0001
pro_view_bin5	1	-3.303	0.0499	4382.0771	<.0001
cart_add_bin1	1	-0.3858	0.0127	917.0377	<.0001
cart_add_bin3	1	0.3052	0.0235	168.8015	<.0001

Table-A7 Odds Ratio Estimates

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald-Confidence Limits	
Mon	1.41	1.362	1.459
Tues	1.257	1.214	1.301
Wed	1.24	1.197	1.286
Thurs	1.267	1.222	1.313
Fri	1.307	1.261	1.355
Sat	1.247	1.202	1.294
Sun	1.775	1.712	1.841
New_Cust	1.784	1.719	1.851
Return_Cust	2.029	1.967	2.093
Loyal_Cust	1.999	1.933	2.067
RF_1to3days	0.924	0.898	0.95
RF_7to14days	1.06	1.022	1.099
RF_morethan14days	1.111	1.074	1.149

Home	0.862	0.841	0.885
Sunday_Flea_Market	1.196	1.145	1.248
ShopClues_Offers	0.838	0.794	0.885
Search_results	1.204	1.135	1.277
Sunday_Flea_Market_S	1.272	1.194	1.356
Super_Saver_Bazar	1.359	1.27	1.455
Fashion_Bollywood_St	0.71	0.634	0.795
My_Account	0.757	0.725	0.79
Cart_contents	1.591	1.487	1.702
samedayvisit	1.099	1.071	1.128
Av_time_bin1	0.341	0.323	0.361
Av_time_bin2	0.609	0.58	0.64
Av_time_bin3	0.658	0.627	0.691
Av_time_bin4	0.712	0.678	0.747
Av_time_bin5	0.714	0.68	0.75
Av_time_bin6	0.766	0.73	0.805
Av_time_bin7	0.85	0.809	0.892
Av_time_bin8	0.878	0.837	0.922
Av_time_bin9	0.904	0.863	0.947
Av_pgvisit_bin1	0.507	0.48	0.534
Av_pgvisit_bin3	0.856	0.815	0.899
Av_pgvisit_bin4	0.836	0.805	0.869
Av_pgvisit_bin5	0.858	0.826	0.892
Av_pgvisit_bin6	0.896	0.86	0.933
Av_pgvisit_bin7	0.923	0.883	0.964
Av_pgvisit_bin8	0.955	0.919	0.992
pro_view_bin2	0.026	0.024	0.027
pro_view_bin3	0.031	0.029	0.034
pro_view_bin4	0.034	0.031	0.037
pro_view_bin5	0.037	0.033	0.041
cart_add_bin1	0.68	0.663	0.697
cart_add_bin3	1.357	1.296	1.421

Table-A8 Predicted Probability

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	81.7	Somers' D	0.637
Percent Discordant	18.1	Gamma	0.638
Percent Tied	0.2	Tau-a	0.315
Pairs	8256700522	c	0.818

Table-A9 Range for Bins Created

Bins created	Ave_Time_on site (sec)
Av_time_bin1	2- 228
Av_time_bin2	229-362
Av_time_bin3	363-492
Av_time_bin4	493-624
Av_time_bin5	625-761
Av_time_bin6	762-914
Av_time_bin7	915-1102
Av_time_bin8	1103-1366
Av_time_bin9	1367-1866
Bins created	**Ave_pages per visit**
Av_pgvisit_bin1	greater than 1 but less than 5
Av_pgvisit_bin3	greater than 6 but less than 8
Av_pgvisit_bin4	8 or more than 8 but less than 10
Av_pgvisit_bin5	10 to 11
Av_pgvisit_bin6	12 to 13
Av_pgvisit_bin7	14 to 15
Av_pgvisit_bin8	16 to 19
Bins created	**Product Views(days)**
pro_view_bin2	1
pro_view_bin3	2
pro_view_bin4	3
pro_view_bin5	4 to 5
Bins created	**Cart Addition (days)**
cart_add_bin1	0
cart_add_bin2	1
cart_add_bin3	2

List of References

1. *"An analysis of Factors Affecting on Online Shopping Behavior of Consumers"* **by Mohammad Hossein Moshref Javadi, Hossein Rezaei Dolatabadi, Mojtaba Nourbakhsh, Amir Poursaeedi & Ahmad Reza Asadollahi - Department of Management, University of Isfahan, Isfahan, Iran.**
2. *"Know Your Buyer: A predictive approach to understand online buyers' behavior"* **by Sandip Pal,Happiest Minds, Analytics Practice**

Books

1. *Web Analytics 2.0* **by Avinash Kaushik**
2. *SiteCatalyst user guide* – **Adobe Online Marketing Suite**