PROJECT REPORT

Online Shopper's Intention

"submitted towards partial fulfilment of the criteria for award of PGPDSE by GLIM"

Submitted by Group 2:

Student Name	SIS ID
Ankita Shinde	PNKNQYSXVJ
Kiran Kendre	YI1KJHTCUA
Priyaranjan Gouda	4A50SBSO3E
Rutuja Gulhane	G36DP0Z4O8
Sanket Yeginwar	DB9EFH96V6

Batch: Group 2

Mentor: Animesh Tiwari

GREAT LAKES
INSTITUTE OF MANAGEMENT, CHENNAI

greatlearning
Learning for Life

Abstract & keyword

Abstract:

Consumers shopping activities on the internet turn out to be more important every year. Although the increase of e-commerce usage over the last few years has created potential in the market, most of the visitors still do not complete their online shopping process. This leads the online retailers the need for solutions to prevent the loss of their revenues. The aim of this study is to evaluate the actions taken by the visitors on ecommerce environment in real time and predicting the visitor's shopping intent. The extracted features from page view data kept track during the visit along with some session and user information are fed to machine learning classification methods to build a model. Of the 12,330 sessions in the dataset, 84.5% (10,422) were negative class samples that did not end with shopping, and the rest (1908) were positive class samples ending with shopping. Outliers treatment has been done by MICE (Multivariate Imputation via Chained Equations), central tendency method. We have also transformed data by square root method. We have also applied two base model: Decision Tree with accuracy score 85% and Logistic Regression with accuracy score 88%.

Keywords: Online Shopper's Intentions Analysis, Machine Learning, MICE (Multivariate Imputation via Chained Equations).

Acknowledgements

At the outset, we are indebted to our Mentor Mr. Animesh Tiwari for his time, valuable inputs and guidance. His experience, support and structured thought process guided us to be on the right track towards completion of this project.

We are fortunate to have Ms. Varsha Mali as our TA –Academic Delivery. Her in-depth knowledge coupled with her passion in delivering the subjects in a lucid manner has helped us a lot. We are thankful to her for her guidance towards entire coursework.

We also thank all the course faculty of the DSE program for providing us a strong foundation in various concepts of analytics & machine learning.

Ankita Shinde
Kiran Kendre
Priyaranjan Gouda
Rutuja Gulhane
Sanket Yeginwar

Date: 26-02-2016

Place: Pune

Certification of completion

I hereby certify that the project titled “Online Shopper’s Intentions” was undertaken and completed under my guidance and supervision by Ankita Shinde, Kiran Kendre, Priyaranjan Gouda, Rutuja Gulhane, Sanket Yeginwar, students of the September2019 batch of the Post Graduate Program in Data Science & Engineering, Pune.

Mr. Animesh Tiwari

Date: 26-02-2020

Table of Contents

Abbreviations used

Abbreviations	Expansions
LR	Logistic Regression
DT	Decision Tree
AUC	Area Under Curve
RF	Random Forest
LGBM	Light Gradient Boosting Method
Bag DT	Bagging Decision Tree
Boost DT	Boosting Decision Tree
FNR	False Negative Rate
FPR	False Positive Rate
SMOTE	Synthetic Minority Oversampling Technique
URL	Uniform Resource Locator

Executive summary

Background & need for study: The increase in e-commerce usage over the past few years has created potential in the market, but the fact that the conversion rates have not increased at the same rate leads to the need for solutions that present customized promotions to the online shoppers. In physical retailing, a salesperson can offer a range of customized alternatives to shoppers based on the experience he or she has gained over time. This experience has an important influence on the effective use of time, purchase conversion rates, and sales figures. Many e-commerce and information technology companies invest in early detection and behavioral prediction systems which imitate the behavior of a salesperson in virtual shopping environment. In parallel with these efforts, some academic studies addressing the problem from different perspectives using machine learning methods have been proposed. While some of these studies deal with categorization of visits based on the user's navigational patterns, others aim to predict the behavior of users in real time and take actions accordingly to improve the shopping cart abandonment and purchase conversion rates

Scope & Objectives: The objective of this project is to do a research and develop a methodology by building models for Online Consumer Commercial Intent Analysis. By analyzing the mouse movements, the link and button click information that the user has on the screen and the tracking data of the pages visited will be obtained and the actions taken as the result of these data will be determined. Acceptable actions will be used as labels during pattern definition with supervised learning algorithms. Thus, when any user receives actions that match the predefined pattern, they will be tagged with the obtained pattern function and the action to be taken instantaneously will be determined.

Approach & methodology: The data is extracted from google analytics web platform. After processing the dataset and cleaning the inconsistencies, the numerical and categorical features used in the shopper's intention prediction model is generated. Various Classification algorithms are used to predict online consumer commercial intent based on set of independent variables like traffic type, visitor type, duration on administration pages, informational pages and product pages along with technology used. The predictive models are also used to identify the variables that strongly influence the conversion using variable importance and probabilistic approaches. The models are evaluated using relevant model performance measures to arrive at the most robust models for prediction. Clustering algorithms are used to come up with emerging customer segments and relevant target marketing activities for each segment.

Key learnings: The session data obtained from the navigation path followed during the online visit convey important information about the online shopper's intention of the visitor, combining them with session information-based features that possess unique information about the purchasing interest improves the success rate of the system.

Recommendations & actionable insights: The high-level recommendations for the project are developed by predicting customers commercial intent on the website. These are then linked to the model findings to recommend actionable insights, which include providing offers and loyalty points for returning and to increase new customer visits by creating better landing page with high page value.

Chapter 1 - Project overview

The online marketing space is in constant shift as new technologies, services, and marketing tactics gain popularity and become the new standard. Online store owners are one of the many different segments affected by these constant evolutions. In order for these business owners to survive and thrive, they need to be able to make better decisions faster. This is where web analytics comes into play. The data thus made available provides ample scope for varied analytical use cases like customer segmentation & behavioral analysis

Need for study

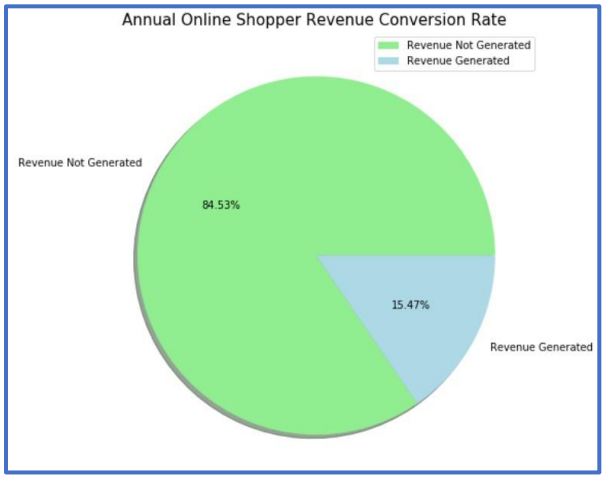


Figure 1.1 – Need for study

The increase in e-commerce usage over the past few years has created potential in the market, But the Sales conversion rates for Columbia Sportswear company have been very low.

The company have invested in early detection and behavioral prediction of users in real time and take actions accordingly to improve the shopping cart abandonment and purchase conversion rates. This study has an important influence on the effective use of time, purchase conversion rates, and sales figures

Current baseline & Business mission

The current yearly conversion rate of online shoppers is 15.47 % and business mission aims to increase the conversion rate.

Problem statement & project scope

To support the incremental revenue contribution from online shoppers, the business require insights related to prospect behavior & engagement with the websites at multiple stages. In the process, we intend to apply various predictive modeling techniques involving classification.

Data sources

In order to classify consumer on-site behavior, a training dataset is collected from online retailer site. This dataset is constructed by Google Analytics function for collecting statistical data about user online activities. The dataset consists of feature vectors belonging to 12330 sessions. The dataset was formed so that each session would belong to a different user in a one-year period to avoid any tendency to a specific campaign, special day, user profile, or period. Of the 12330 sessions in the dataset, 84.5 percentage (10422) were negative class samples that did not end with shopping, and the rest (1908) were positive class samples ending with shopping.

Dataset Description

During website session, browsing information about visited pages is collected and features are extracted
as follows

Feature Name	Feature Description	Type
Administrative	It is the admin page of website.	Categorical
Administrative Duration	The time spent by user on the Administrative Page.	Numerical
Informational	It is the information page of website.	Categorical
Informational Duration	The time spent by user on Information Page.	Numerical
Product Related	It is the product related page of website.	Categorical
Product Related Duration	The time spent by user on the Product Related Page	Numerical
Bounce Rate	It is a single-page session on your site.	Numerical
Exit Rate	For all page views to page, exit rate is percentage that were the last in session	Numerical
Page Value	It is the average value for a page that a user visited before landing on the goal page or completing an Ecommerce transaction (or both).	Numerical
Special Day	Closeness of the site visiting time to a special day	Categorical
Operating Systems	Operating system of the visitor	Categorical
Browser	Browser of the visitor	Categorical

Feature Name	Feature Description	Type
Region	Geographic region from which the session has been started by the visitor	Categorical
Traffic Type	Traffic source by which the visitor has arrived at the website (e.g. banner, SMS, direct)	Categorical
Visitor Type	Visitor type as “New Visitor”, “Returning Visitor” and "Other"	Categorical
Weekend	Boolean value indicating whether the date of the visit is weekend	Categorical
Revenue	Class label indicating whether the visit has been finalized with a transaction	Categorical

Table shows the numerical and categorical features. Among these features, "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor and total time spent in each of these page types in seconds. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another.

The "Operating Systems", " Browser", " Traffic Type" and “Visitor Type” features represent the metrics measured by "Google Analytics" for each page in the e-commerce site. “Weekend” and “Month” features are derived by looking date of visit.They give information about whether the date of visit is at the end of the week or not and the moth of the visit respectively. “Revenue” feature indicates that whether the visit results in transaction finalization.

Data preparation & clean up

The source dataset received has been prepared to ensure that the fields are cleaned up, the values are suitable for model building and the variable names are self-explanatory. The broad approach for data preparation can be outlined as:

Table 4 – Data pre-processing steps

Label Encoding	Outlier Treatment	Transformation
Categorical Variable Month, Operating System, Browser, Region, Traffic Type, Visitor type are label encoded. Weekend and Revenue feature is converted into binary value 0's and 1's	We have treated outliers by replacing outliers with NAN and treating NAN values by MICE(Multivariate Imputation via Chained Equations).	We have transformed the data by square root method.

Statistical tools & techniques

Various classification algorithms have been used to analyze customer purchase intention for conversion and to identify the extent to which each independent variable influence conversion. The independent variables can be broadly grouped as Visitor session information and visitor pageview information. The dependent variable is whether the customer will generate the revenue or not by his session navigation pattern.

The model building exercise has also considered cross validation and tuning techniques to ensure that the models built perform well when used for prediction.

The classification algorithms used for Commercial intent prediction include

- Logistic regression
- Decision Tree
- Random Forest
- Light Gradient Boosting
- KNN

Model performance measures used for evaluating models

The various models built, must be evaluated based on certain model performance measures to identify the most robust models. The choice of the right model performance measures is highly critical since the dataset is a highly imbalanced dataset and the conversion rate is 15.47%. Model accuracy alone may not be enough to evaluate a model. Hence the following model performance measures have been used to evaluate the models, based on the confusion matrix built for the predictions on the training and test datasets:

	Negative (Predicted)	**Positive (Predicted)**
Negative (Observed)	True Negative (TN)	False positive (FP)
Positive (Observed)	False negative (FN)	True positive (TP)

Accuracy

Accuracy is the number of correct predictions made by the model by the total number of records. The best accuracy is 100% indicating that all the predictions are correct.

Considering the response rate (conversion rate) of our dataset which is ~16%, accuracy is not a valid measure of model performance. Even if all the records are predicted as 0, the model will still have an accuracy of 84%. Hence other model performance measures need to be evaluated.

Sensitivity or recall

Sensitivity (Recall or True positive rate) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall or true positive rate (TPR).

For our dataset, it gives the ratio of actual customers who generated revenue by the total number of customers predicted who will generate the revenue.

Specificity

Specificity (true negative rate) is calculated as the number of correct negative predictions divided by the total number of negatives.

For our dataset, specificity gives the ratio of actual customers who will not generate revenue by the number of customers who are predicted who will not generate revenue.

Precision

Precision (Positive predictive value) is calculated as the number of correct positive predictions divided by the total number of positive predictions.

Precision tells us, what proportion of customers who generated revenue as customers actually generated revenue. If precision is low, it implies that the model has lot of false positives.

F1-Score

F1 is an overall measure of a model's accuracy that combines precision and recall A good F1 score means that you have low false positives and low false negatives, so you're correctly identifying real threats and you are not disturbed by false alarms. An F1 score is considered perfect when it's 1, while the model is a total failure when it's 0.

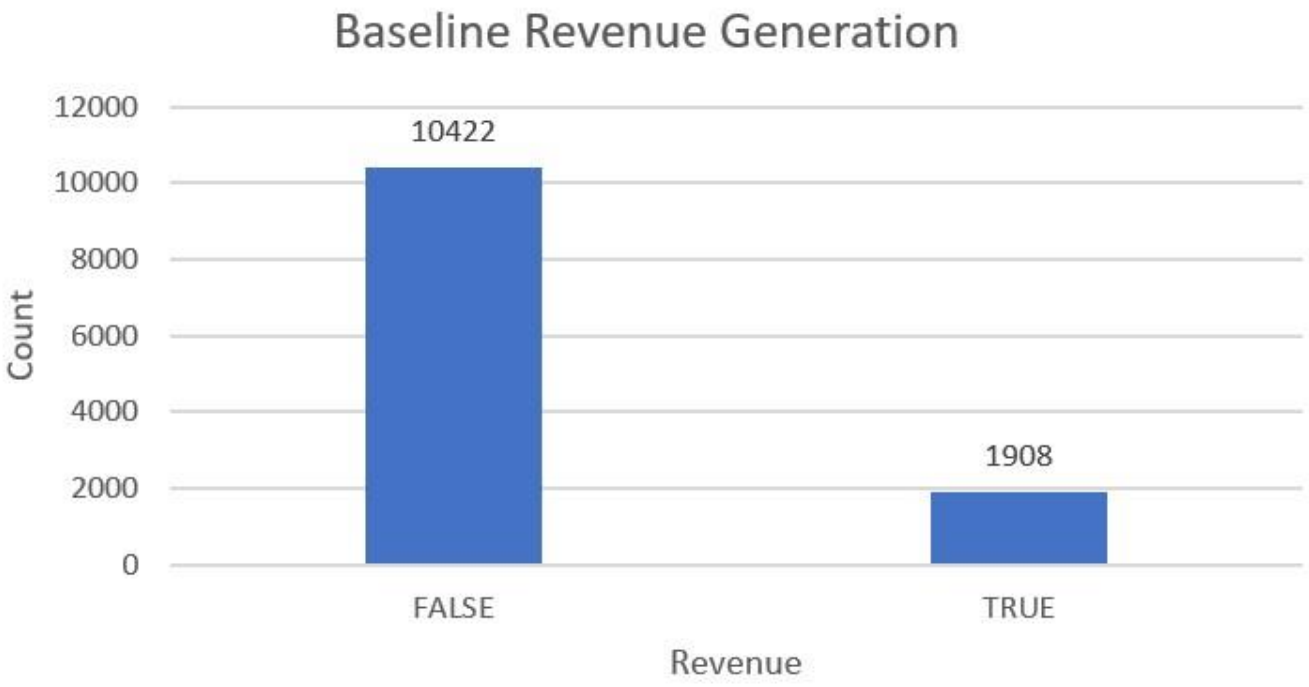
Chapter 2 - Exploratory data analysis

The purpose of exploratory data analysis is two-fold:

- ➢ to understand the data in terms of Visitor session information and visitor pageview information across various independent variables
- ➢ Get insights on various features.

Understand data distribution

Baseline conversion rate



The baseline conversion rate of visitors who generated Revenue vs overall visitors is = 1908/12330 = 15.47 %.

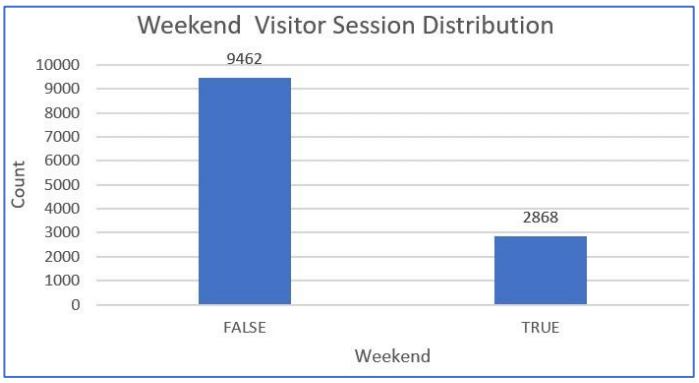
This indicates that the data set is an imbalanced dataset where the number of observations belonging to class 1 (True) is significantly lower than those belonging to class 0 (False)

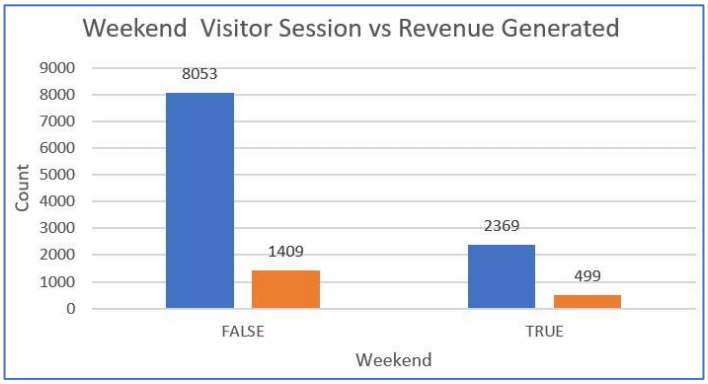
The conventional accuracy of the predictive models is not a relevant measure of model performance because machine learning algorithms are usually designed to improve accuracy by reducing the error. Thus, they do not take into account the class distribution / proportion or balance of classes.

Hence we will consider other model performance measures to evaluate a model, keeping in mind the class imbalance problem.

Insights for Feature Selection

Weekend Visitor session vs Revenue Generated



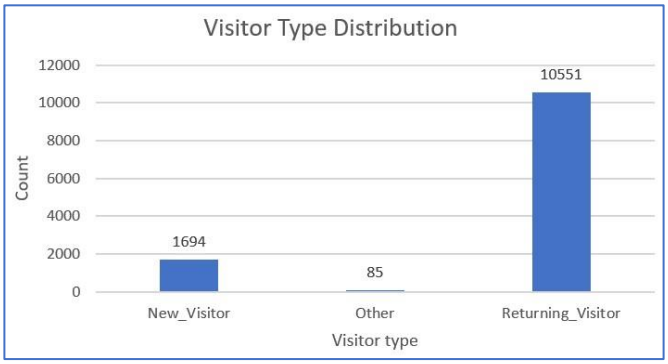


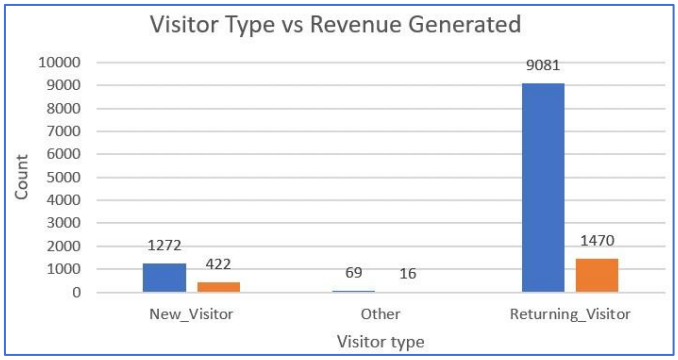
Revenue Conversion Rate

Revenue	FALSE	TRUE
Weekday	85.11%	14.89%
Weekend	82.60%	17.40%

- There is lot of visitor session is found during weekday rather than weekend. Which might be due to the reason that customer prefer to shop directly in stores during weekends rather than online
- Revenue conversion rate during weekend is slightly greater than weekday.

Visitor Type vs Revenue Generated



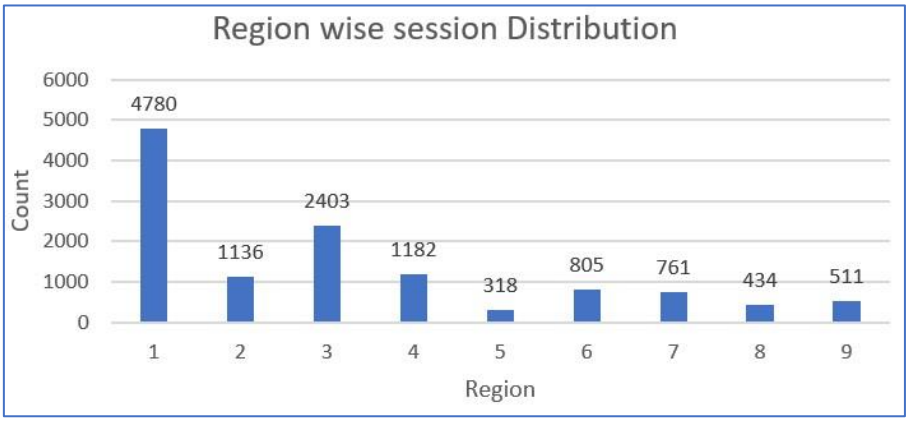


Visitor Type Conversion Rate

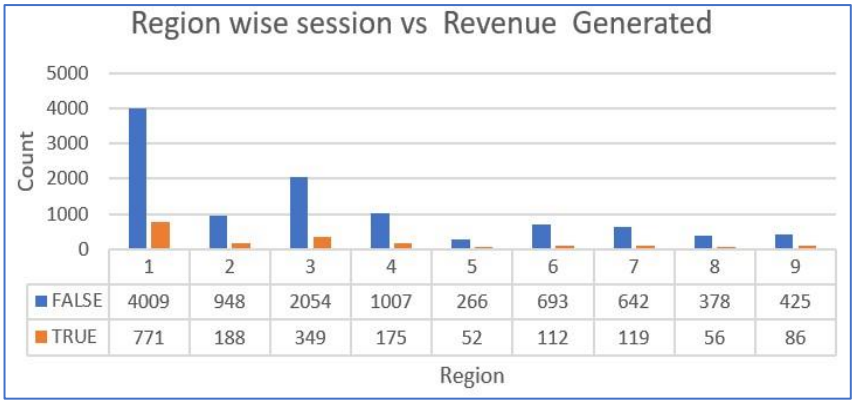
Visitor Type	FALSE	TRUE
New Visitor	75.09%	24.91%
Other	81.18%	18.82%
Returning Visitor	86.07%	13.93%

- Number of new visitors are very less when compared to returning customer to the website.
- Conversion Rate of new customer is nearly 10% greater than returning customer.
- More Efforts need to be made by digital marketing team to bring new visitors to the website.

Region wise visitor session vs Revenue Generated

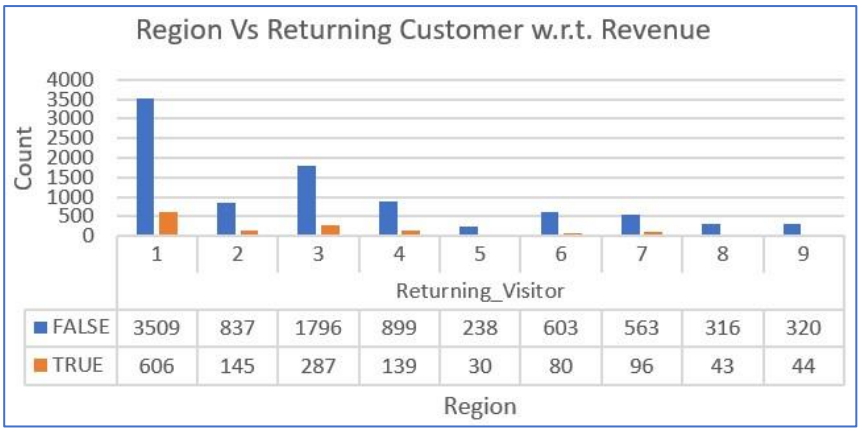


Region	% Visitors
Region 1	38.77%
Region 2	9.21%
Region 3	19.49%
Region 4	9.59%
Region 5	2.58%
Region 6	6.53%
Region 7	6.17%
Region 8	3.52%
Region 9	4.14%



Region Wise Conversion Rate

Region	FALSE	TRUE
Region 1	83.87%	16.13%
Region 2	83.45%	16.55%
Region 3	85.48%	14.52%
Region 4	85.19%	14.81%
Region 5	83.65%	16.35%
Region 6	86.09%	13.91%
Region 7	84.36%	15.64%
Region 8	87.10%	12.90%
Region 9	83.17%	16.83%

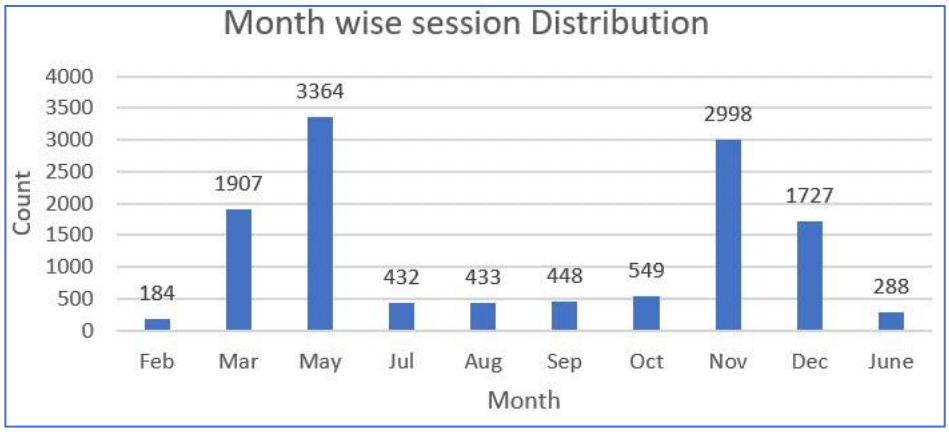


Returning customer conversion Rate

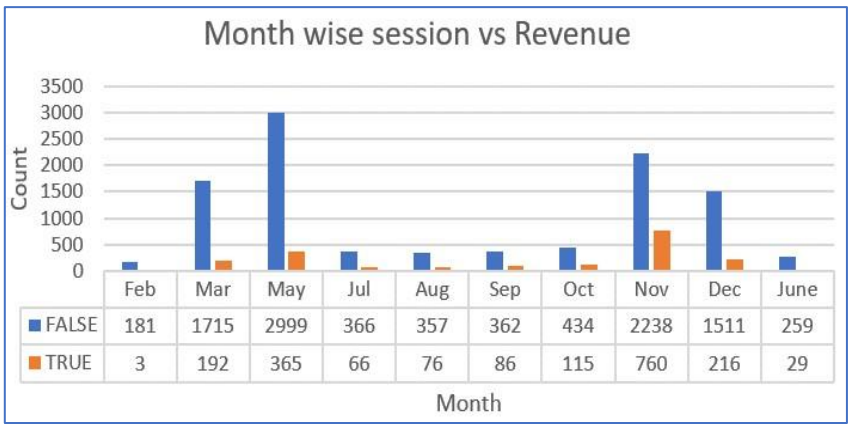
Region	FALSE	TRUE
Region 1	85.27%	14.73%
Region 2	85.23%	14.77%
Region 3	86.22%	13.78%
Region 4	86.61%	13.39%
Region 5	88.81%	11.19%
Region 6	88.29%	11.71%
Region 7	85.43%	14.57%
Region 8	88.02%	11.98%
Region 9	87.91%	12.09%

- ➢ More customer web session is found for Region 1.
- ➢ Even though customer web session of region 3 and region 4 is more region 2 its conversion rate is low.
- ➢ While considering Returning customer conversion rate by region wise region 3 and region 4 are low.
- ➢ More Efforts need to be made by digital marketing team to increase revenue conversion rate in region 3 and region 4.

Month wise visitor session vs Revenue Generated

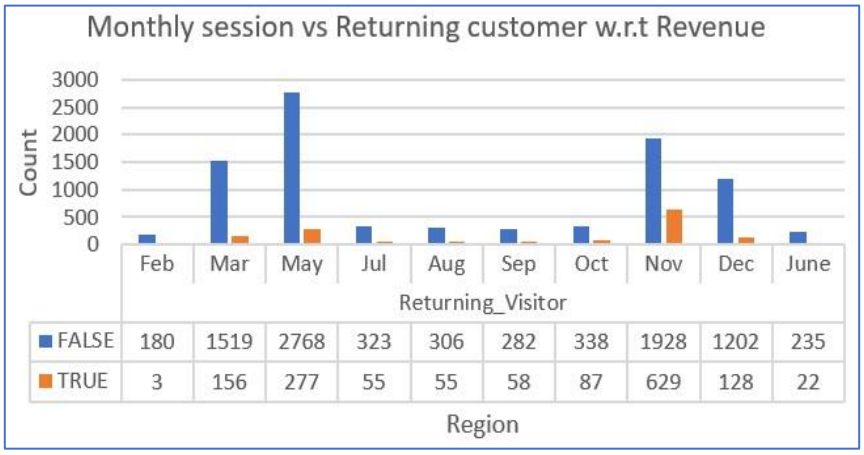


Month	% visitors
Feb	1.49%
Mar	15.47%
May	27.28%
June	2.34%
Jul	3.50%
Aug	3.51%
Sep	3.63%
Oct	4.45%
Nov	24.31%
Dec	14.01%



Month Wise Conversion Rate

Month	FALSE	TRUE
Feb	98.37%	1.63%
Mar	89.93%	10.07%
May	89.15%	10.85%
June	89.93%	10.07%
Jul	84.72%	15.28%
Aug	82.45%	17.55%
Sep	80.80%	19.20%
Oct	79.05%	20.95%
Nov	74.65%	25.35%
Dec	87.49%	12.51%

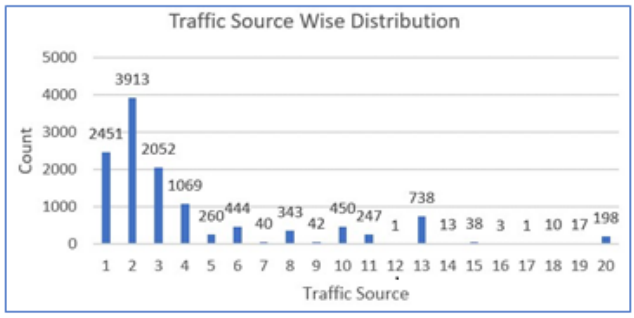


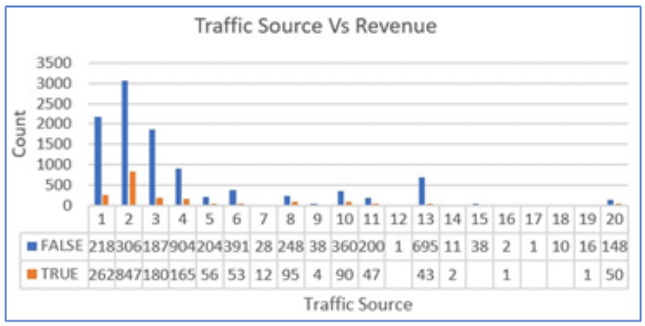
Returning customer Conversion Rate

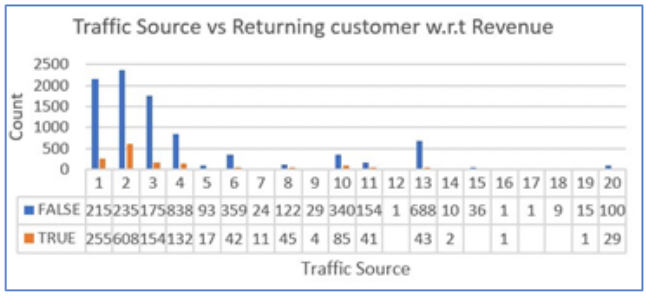
Month	FALSE	TRUE
Feb	98.36%	1.64%
Mar	90.69%	9.31%
May	90.90%	9.10%
June	91.44%	8.56%
Jul	85.45%	14.55%
Aug	84.76%	15.24%
Sep	82.94%	17.06%
Oct	79.53%	20.47%
Nov	75.40%	24.60%
Dec	90.38%	9.62%

- ➢ 81% of online user session is found in the month of March, May, November and December.
- ➢ Conversion rate of March, May and December is very low when compared to November Month.
- ➢ Even returning customer conversion rate is low on these three months.
- ➢ More offers can be given in these months to boost revenue generation.

Visitor Traffic Source vs Revenue Generate



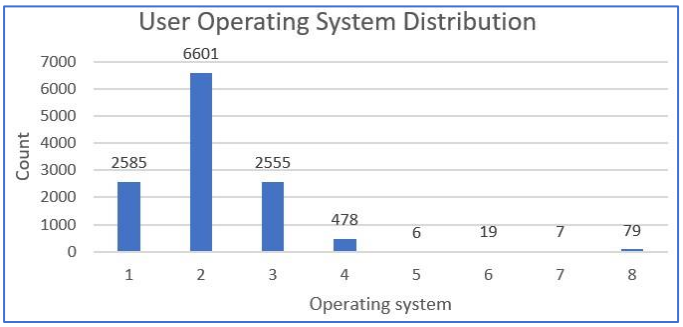


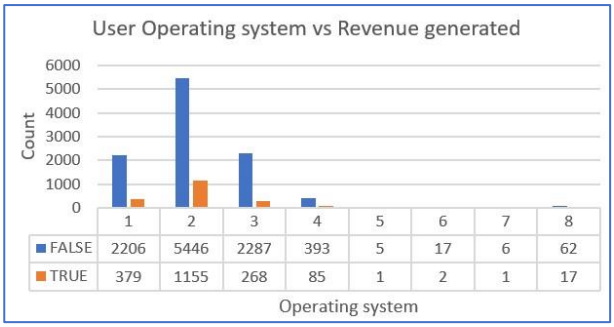


Source	Visitor in %	Traffic Conversion Rate		Returning Customer Conversion rate	
		FALSE in %	TRUE in %	FALSE in %	TRUE in %
Source 1	19.88	89.31	10.69	89.41	10.59
Source 2	31.74	78.35	21.65	79.5	20.5
Source 3	16.64	91.23	8.77	91.92	8.08
Source 4	8.67	84.57	15.43	86.39	13.61
Source 5	2.11	78.46	21.54	84.55	15.45
Source 6	3.6	88.06	11.94	89.53	10.47
Source 7	0.32	70	30	68.57	31.43
Source 8	2.78	72.3	27.7	73.05	26.95
Source 9	0.34	90.48	9.52	87.88	12.12
Source 10	3.65	80	20	80	20
Source 11	2	80.97	19.03	78.97	21.03
Source 12	0.01	100	0	100	0
Source 13	5.99	94.17	5.83	94.12	5.88
Source 14	0.11	84.62	15.38	83.33	16.67
Source 15	0.31	100	0	100	0
Source 16	0.02	66.67	33.33	50	50
Source 17	0.01	100	0	100	0
Source 18	0.08	100	0	100	0
Source 19	0.14	94.12	5.88	93.75	6.25
Source 20	1.61	74.75	25.25	77.52	22.48

- 68% of revenue are generated are three traffic sources 1,2 and 3.
- Revenue Conversion rate of source 1 and source 3 less when compared to Source 2.
- Returning customer conversion rate on traffic source 1 and source 3 are also relatively low when compared to Source 2. Less conversion rate of these source might be due to wrong landing page.

Visitor Operating system vs Revenue Generated

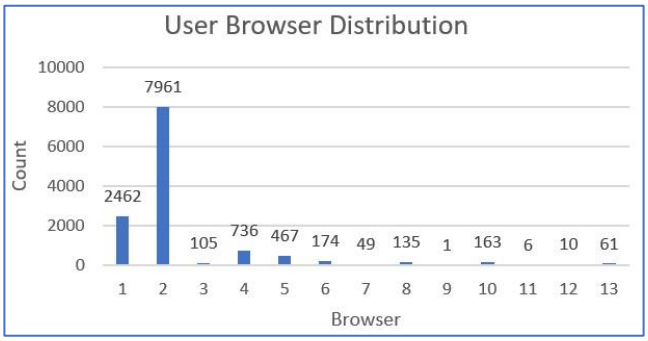


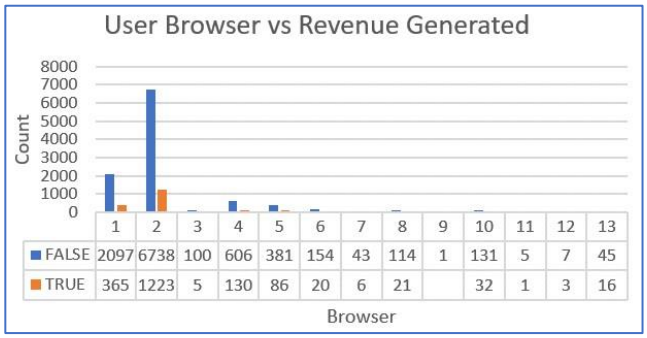


		OS Conversion Rate		Returning Customer Conversion Rate	
Operating System	% Visitor	FALSE	TRUE	FALSE	TRUE
OS 1	20.97%	85.34%	14.66%	87.40%	12.60%
OS 2	53.54%	82.50%	17.50%	84.12%	15.88%
OS 3	20.72%	89.51%	10.49%	89.89%	10.11%
OS 4	3.88%	82.22%	17.78%	83.86%	16.14%
OS 5	0.05%	83.33%	16.67%	83.33%	16.67%
OS 6	0.15%	89.47%	10.53%	94.12%	5.88%
OS 7	0.06%	85.71%	14.29%	83.33%	16.67%
OS 8	0.64%	78.48%	21.52%	100.00%	0.00%

- 95% Customer online session is done from three operating system OS1, OS2 and OS3.
- Conversion rate of OS3 is less when compared with OS1 and OS2.
- Similarly Returning customer conversion rate is less for OS1 and OS3 when compared with OS2.

Visitor Browser vs Revenue Generated



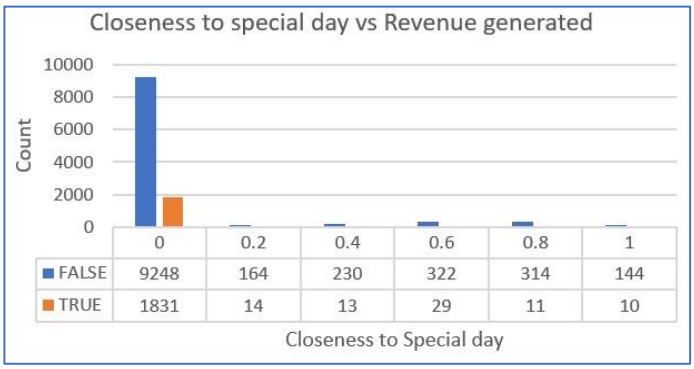


Browser Conversion

		Browser Conversion Rate		Returning Customer Conversion rate	
Browsers	Visitor	FALSE	TRUE	FALSE	TRUE
Browser 1	19.97%	85.17%	14.38%	87.28%	12.72%
Browser 2	64.57%	84.64%	15.36%	85.69%	14.31%
Browser 3	0.85%	95.24%	4.76%	95.24%	4.76%
Browser 4	5.97%	82.34%	17.66%	86.61%	13.39%
Browser 5	3.79%	81.58%	18.42%	84.37%	15.63%
Browser 6	1.41%	88.51%	11.49%	88.96%	11.04%
Browser 7	0.04%	87.76%	12.24%	85.37%	14.63%
Browser 8	1.09%	84.44%	15.56%	86.73%	13.27%
Browser 9	0.01%	100.00%	0.00%	100.00%	0.00%
Browser 10	1.32%	80.37%	19.63%	81.56%	18.44%
Browser 11	0.05%	83.33%	16.67%	83.33%	16.67%
Browser 12	0.08%	70.00%	30.00%	66.67%	33.33%
Browser 13	0.49%	73.33%	26.23%	87.50%	12.50%

- 85 % of user session occurs from using 2 browsers namely Browser 1 and Browser 2
- Both Browser conversion rate and returning conversion rate is low for these browser.

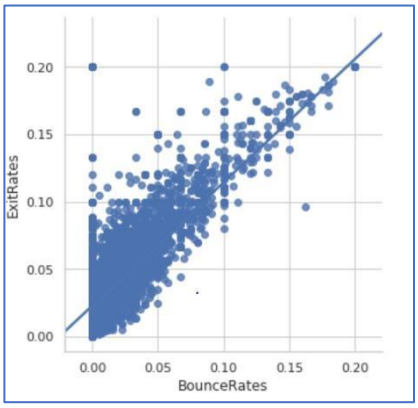
Special Day vs Revenue Generated

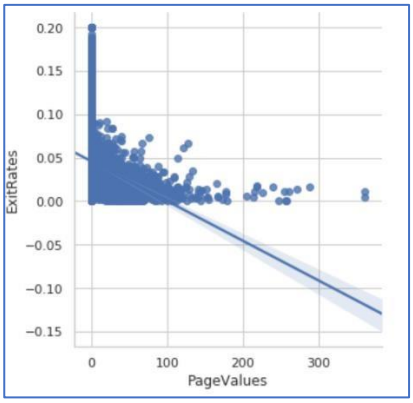


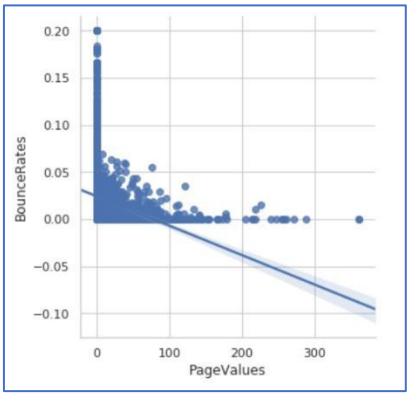
Special Day	% visitor
0	89.85%
0.2	1.44%
0.4	1.97%
0.6	2.85%
0.8	2.64%
1	1.25%

- ➢ 90% online session happen on non-special days.
- ➢ Since its sportswear company there is no affinity for special days to revenue generation.

Bounce Rate vs Exit Rate vs Page value

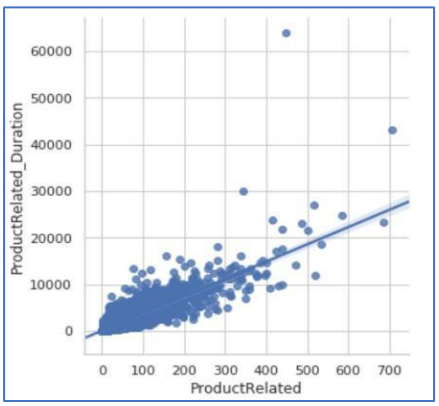


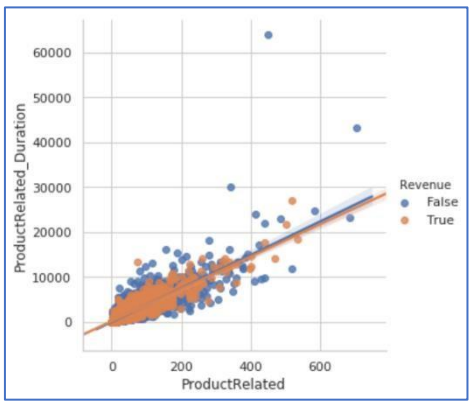




- ➢ Bounce Rate and Exit rate have positive correlation. With increase in Bounce rate the exit rate from the page increases.
- ➢ Page value and Exit rate are negatively correlated. With increase in page value the exit rate reduces.
- ➢ Page value and Bounce rate are negatively correlated. With increase in page value the bounce rate reduces.

Product related pageviews vs product page duration vs Revenue generated

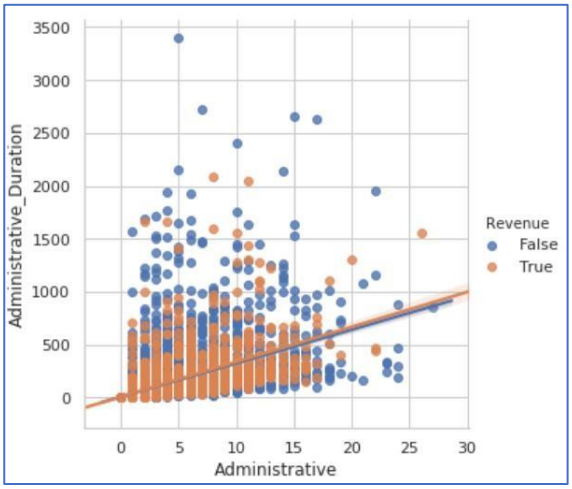




➢ Product related pageviews and product related pageview duration are positively correlated. With increase in number of products pageviews the product pageview duration also increase.

➢ Even tough customers spend more time on product pages they didn't make into revenue conversion.

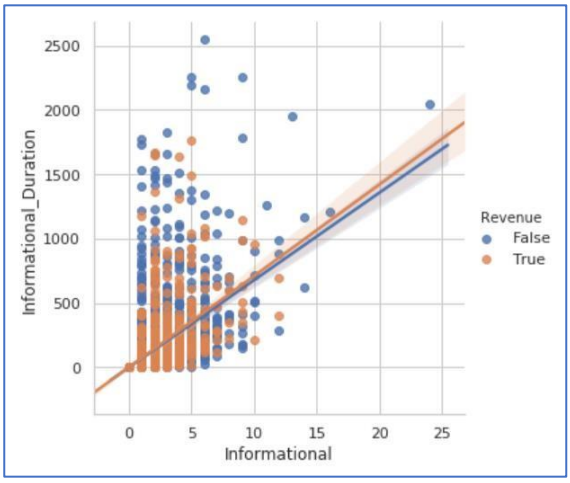
Administrative related pageviews vs Administrative duration vs Revenue generated



➢ Administrative related pageviews and Administrative related pageview duration are positively correlated. With increase in number of products pageviews the product pageview duration also increase.

➢ User who visited less number of administrative pages but tool more duration on those pages, this implies user might have problem in logging in pages

Information related pageviews vs Informational duration vs Revenue generated

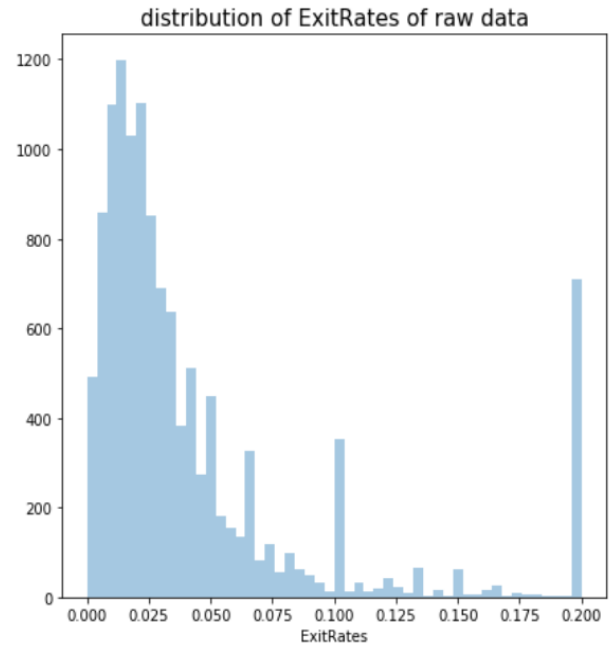


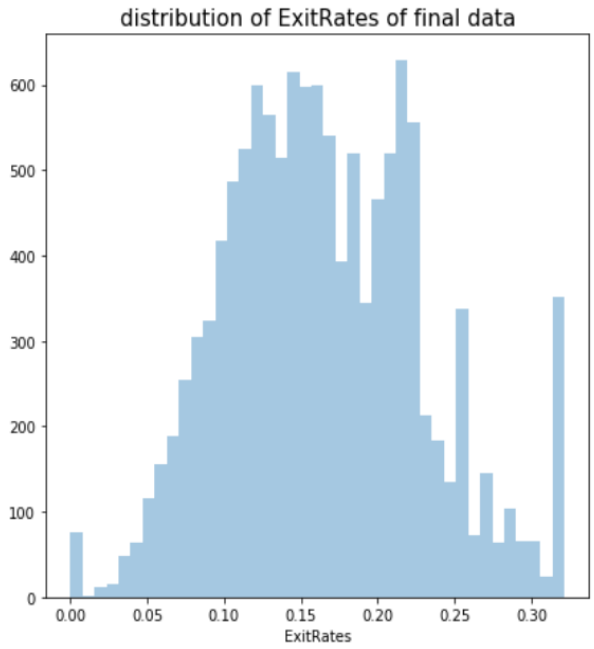
➢ Information related pageviews and information related pageview duration are positively correlated. With increase in number of products pageviews the product pageview duration also increase.

➢ Customer who have made online purchase visited lesser number of informational pages which implies informational pageview don't have much effect on revenue generation.

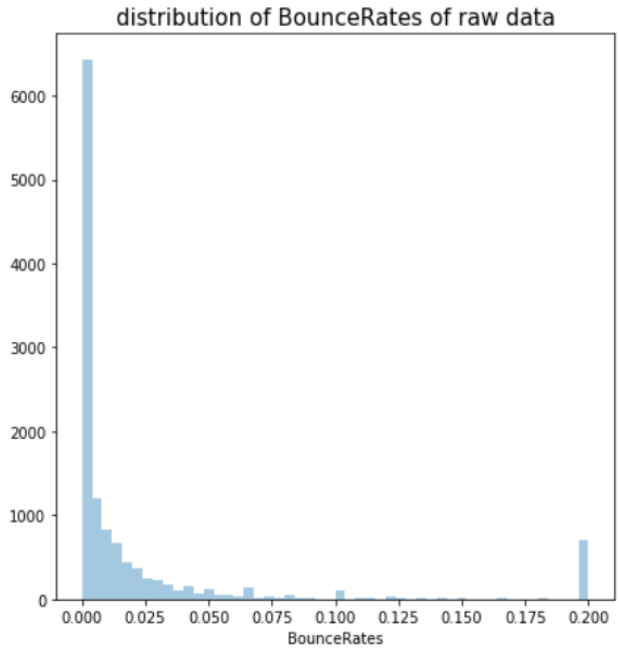
Distribution before and after Outlier Treatment

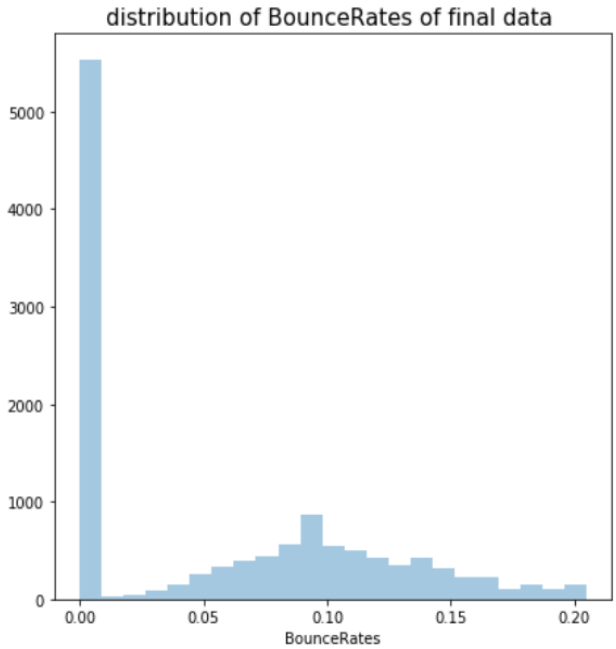
Exit rate:





Bounce Rate:





Chapter 3 - Feature Selection & Model Building

Features Statistical Inference

	Features	P-Value	Type	Test
0	Administrative_Duration	2.078261e-06	Continuous	Annova
1	ProductRelated_Duration	6.919878e-41	Continuous	Annova
2	BounceRates	8.087801e-09	Continuous	Annova
3	ExitRates	6.724954e-45	Continuous	Annova
4	Month	2.238786e-77	Categorical	Chi-Sqr
5	VisitorType	4.269904e-30	Categorical	Chi-Sqr
6	Weekend	1.266325e-03	Categorical	Chi-Sqr
7	TrafficType	1.652735e-67	Categorical	Chi-Sqr
8	Region	3.214250e-01	Categorical	Chi-Sqr
9	Browser	6.087543e-03	Categorical	Chi-Sqr
10	OperatingSystems	1.416094e-13	Categorical	Chi-Sqr
11	SpecialDay	3.543244e-19	Categorical	Chi-Sqr
12	PageValues	0.000000e+00	Categorical	Chi-Sqr
13	Informational	2.781744e-35	Categorical	Chi-Sqr
14	Administrative	2.270586e-77	Categorical	Chi-Sqr
15	ProductRelated	7.875455e-72	Categorical	Chi-Sqr

- All the features are important as p-value is less than 0.05, so they all are significant. We will use feature selection technique to get the important features

Feature Selection

Feature selection is the process of selecting a subset of relevant attributes to be used in making the model in machine learning. Effective feature selection eliminates redundant variables and keeps only the best subset of predictors in the model which also gives shorter training times. Besides this, it avoids the curse of dimensionality and enhance generalization by reducing overfitting.

In this project, feature selection techniques are applied to improve the classification performance and/or scalability of the system. Thus, we aim to investigate if better or similar classification performance can be achieved with a smaller number of features. An alternative of feature selection is the use a feature extraction technique such as Principal Component Analysis for dimensionality reduction. However, in this case, the features in the reduced space will be the linear combinations of 17 attributes, which brings the need of tracking all features during the visit and updating the feature vector after a new action is taken by the visitor. Therefore, it has been deemed appropriate to apply feature selection instead of feature extraction within the scope of this research. We have used backward elimination for feature selection.

Besides, considering the real-time usage of the proposed system, achieving better or similar classification performance with less number of features will improve the scalability of the system since less number of features will be kept track during the session.

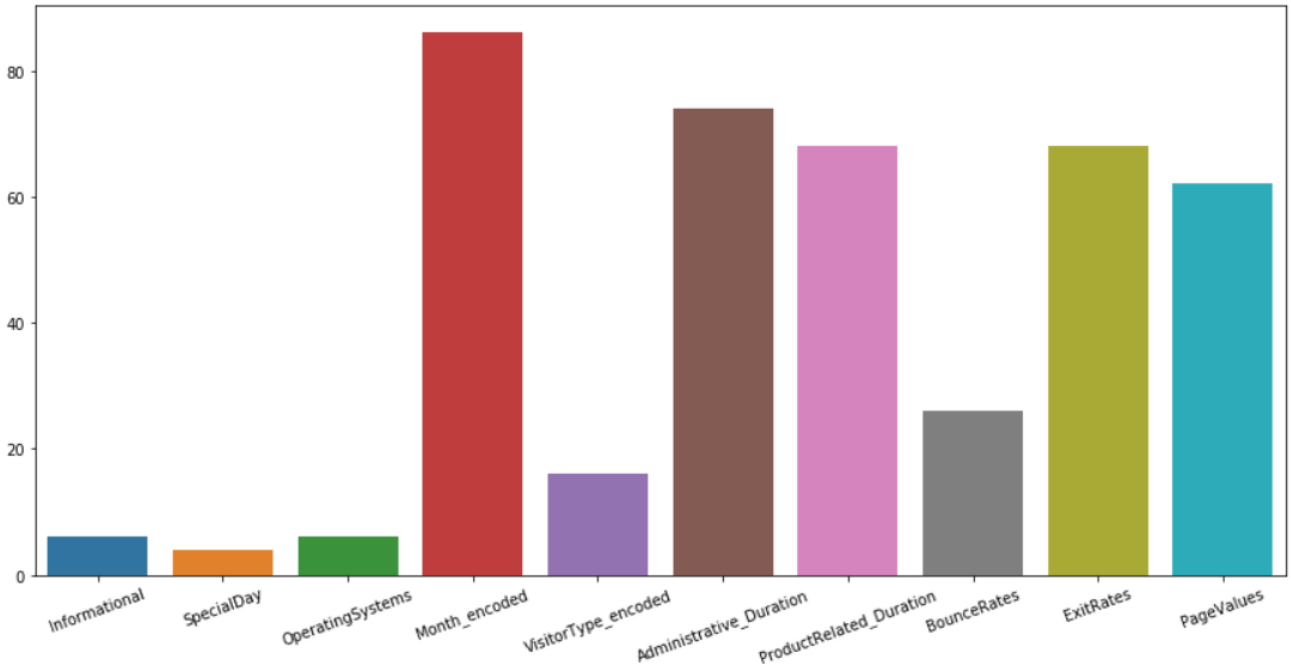
Model Building

Classification Results:

One of the purposes of this project is to get the analyses results of the measuring the user's intention to finalize the transaction and build a model for visitor behavior analysis. The dataset is fed to Logistic Regression, Decision tree, Random Forest and Light Gradient Boosting classifiers using fivefold cross validations. The Accuracy, Precision, Bias Error and Variance Error and F1-Score are presented for each classifier.

Algorithm	Train_accuracy	Test_accuracy	Test_Precision	Test_recall	F1_score
Logistic Regression (Raw data)	0.8814	0.8755	0.6736	0.3431	0.4547
Logistic Regression (Final data)	0.8888	0.8943	0.7450	0.4908	0.5917
Bagging Logistic Regression (Final data)	0.8891	0.8947	0.7421	0.4986	0.5965
Decision Tree (Raw data)	1.0000	0.8682	0.5634	0.5710	0.5672
Decision Tree (Final data)	1.0000	0.8705	0.5866	0.5772	0.5820
Bagging Decision Tree (Final data)	0.9894	0.8938	0.6993	0.5616	0.6229
K-nearest neighbours (Raw data)	0.8946	0.8527	0.5884	0.2493	0.3387
K-nearest neighbours (Final data)	0.8989	0.8914	0.7248	0.4908	0.5852
Random Forest (Raw data)	0.9998	0.8949	0.6979	0.5388	0.6081
Random Forest (Final data)	0.9998	0.9053	0.7403	0.6062	0.6666
Tuned Random Forest (Final data)	0.8974	0.8918	0.7067	0.5249	0.6024
Light Gradient Boosting Method (Final data)	0.9172	0.9029	0.7307	0.5984	0.6580

Variable Importance Plot for Final Model



Chapter 4 - Conclusions

In this project, we aimed to construct a real-time user behavior analysis system for online shopping environment. We used an online retailer data to perform the experiments. In order to predict the purchasing intention of the visitor, we used aggregated page view data that kept track during the visit along with some session and user information as input to machine learning algorithms. Data Cleaning (Outlier treatment) and feature selection pre-processing techniques are applied to improve the success rates and scalability of the algorithms. The best results are achieved with a Light Gradient Boosting algorithm. Our findings support the argument that the features extracted from session data during the visit convey important information for online shopper's intention prediction.

Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output. So, here we have used Backward Elimination method. Considering the real time usage of the proposed system, achieving better or similar classification performance with minimal subset of features is an important factor for the e-commerce companies since less number of features will be kept track during the session.

Chapter 5 – Recommendations and actionable Insights

➢ Conversion Rate of New visitors are high when compared to Returning customer. In order to bring new visitors to the website below actions needs to be taken.

- ❖ Discounting is not a long-term strategy but it can be highly effective in driving new customers to your store. Figure out your customer acquisition cost and from that, how much of a discount (on a limited amount of quantity/product) you can afford to offer in order to acquire new customers.
- ❖ Partnering with a non-competitive but audience-complementary partner can be a highly effective way of acquiring new customers. This can be something as simple as a traffic exchange – partnering with a highly-trafficked site in your customer's domain, putting up a banner to drive traffic to your shop, and paying the partner either a cut of the cart revenue or a flat fee for every customer acquired via the partner banner.
- ❖ Writing authoritative, interesting content in your online shop's contextual domain will pay huge dividends over the long term. Targeted content will help boost your site's SEO bringing in new customers organically, and will also encourage your existing visitors to share your content more. Every online shop should have blog content as part of its marketing strategy
- ❖ A super effective way to capture a whole new customer segment is to offer a whole new product or service! This doesn't even need to be complicated, it could simply be a repositioning, repackaging or even repricing of an existing product.
- ❖ The best and arguably most valuable method of customer acquisition is when existing customers *refer a friend*. When this method works really well, all the marketing is done by your existing customers meaning you can focus on running your online store instead of spending time bringing people to it. Referrals can happen organically via Word of Mouth marketing (focus on great products, great prices and excellent customer) but you can also implement a referral marketing program.

➢ Number of Returning customer to website is high but the conversion rate is low when compared to new customers. Retargeting is an effective way to generate a revenue.

- ❖ Target Individuals based on the searches they conduct on Web Brower.
- ❖ Target Individuals based on specific products viewed, actions taken and actions not taken (abandoning the cart)
- ❖ Target the customer based on the source they arrived to the website.
- ❖ Target customers who are interacting with email programs.
- ❖ Target customers who have visited a partner site that shares similar product.
- ❖ Target the customers who have interact with your distributed content (custom Facebook page, expandable ad unit.)
- ❖ Target individuals who consume similar content to your existing customers.

➢ Decrease the bounce rate of page and increase page value for more revenue generation.

Chapter 6 - References and Bibliography

- Exploration of shopping orientations and online purchase intention. European Journal of Marketing, 37(11/12), 2003.
- Yi Jin Lim, Abdullah Osman, Shahrul Nizam Salahuddin, Abdul Rahim Romle, Safizal Abdullah Factors Influencing Online Shopping Behavior: The Mediating Role of Purchase Intention.
- Referred www.towardsdatascience.com for MICE outlier treatment
- Arun Thamizhvanan, M.J. Xavier. Determinants of customers' online purchase intention: An empirical study in India.