PROJECT REPORT

Online Shopper's Purchasing Intention

Abstract

One of the most popular activities on the Web is shopping. It has much allure in it — you can shop at your leisure, anytime, and in your PJs. Literally anyone can have their pages built to display their specific goods and services and create a market for buyers to purchase their products.

History of ecommerce dates back to the invention of the very old notion of "sell and buy,'' electricity, cables, computers, modems, and the Internet. E-commerce became possible in 1991 when the Internet was opened to commercial use. Since that date thousands of businesses have taken up residence at web sites.

Even though E-commerce industry is experiencing perennial growth since its inception, one of the crucial problems is that most of the visitors still do not complete their online shopping process. This leads to loss of revenues for the online retailer's. This study is done in order to provide a solution for the above mentioned problem by evaluating the actions taken by the visitors on E-commerce environment in real time and predicting the visitor's shopping intent.

The information provided by the visits of users is fed to machine learning classification algorithms to build a predicting model. In the process of refining the model and making it better to provide more insightful results, oversampling and feature selection pre-processing steps are employed.

Table of Contents

Chapter 1 - Introduction

Because of the numerous advantages and benefits, more and more people say they prefer online shopping over conventional shopping these days. The buyer's decision-making process has changed dramatically in recent years. Buyers are conducting extensive research online before ever speaking to a salesperson. Buyers are also making more direct purchases online and via their smartphone, never stepping foot into traditional brick-and-mortar locations. The internet makes doing business much easier and faster. The data thus made available provides ample scope for varied analytical use cases like customer segmentation and behavioural analysis. An e-commerce system provides real-time data and analytics about products and customers. You can see how people interact with the site, what products interest them, what they left in their cart and how much the average purchase was. Valuable metrics that allow you to make adjustments to meet your customer's needs.

Background

Online shopping is one of the most popular online activities worldwide with global e-retail sales reaching 3.5 trillion U.S. dollars in 2019. It is no surprise that India is the fastest growing online commerce market. Globally, in 2019, ecommerce sales will account for 14.1% of all retail sales worldwide. This figure is expected to reach 22% in 2023. With such a growth opportunity, every business is making strides to capture a share of the market.

Need for Study

In spite of improving underlying factors like a young demographics, rising internet penetration, and relative better economic performance the e-commerce sector has not been able to solve some of the major issues that plague the market like delivering a personal experience, high exit rates and shopping cart abandonment and maintaining customer loyalty. Most of them lack the necessary insight into customer behaviour and buying patterns, data which can help them thrive in the current e-commerce environment. These factors ultimately cause a low acquisition rate, the industry standard is only 1 to 2% or the cost of acquisition rises drastically.

Current Situation and Mission

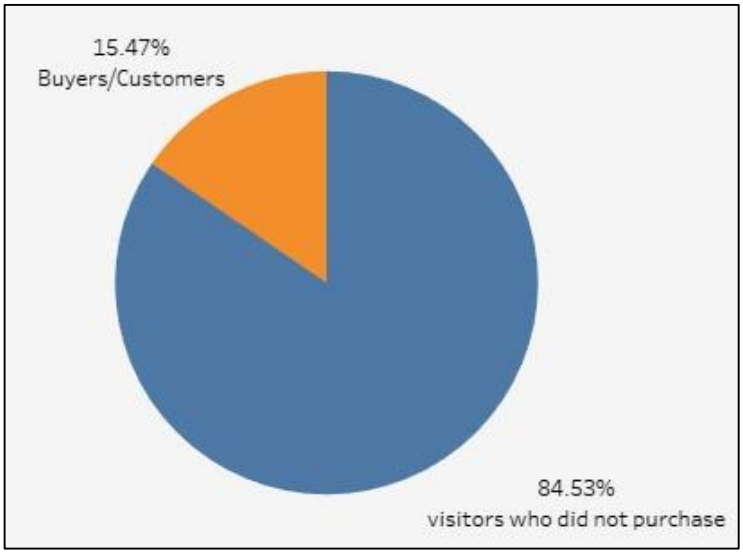


Figure 1 – Conversion Rate

The current yearly conversion rate of online at this e - commerce website is 15.47 % and business mission aims to increase the conversion rate to 30%.

Problem Statement

To recolonize potential customers with high purchase intent so that the e-commerce can nurture the leads and/or use remarketing tools to convert them improving the conversion rate. To provide these deliverables, the project would analyse data of user session logged in by visitors on the e-commerce website.

Data Source

To classify consumer on-site behaviour, a training dataset is collected from an e-commerce website. This dataset is consisting of a combination of both user data collected by the e-commerce and Google Analytics metrics showing statistical data about the pages. 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 to avoid any tendency to a specific campaign, special day, user profile, or period.

Data Description

The dataset consists of 10 numerical and 8 categorical attributes.

Table 1 - Numerical features used in the user behaviour analysis model

Name of the Feature	Description	Range	Mean	Standard Deviation
Administrative	Number of Admirative Pages (Login Page, Logout, Register, Account, forgot password etc.) visited by the user.	0 – 27	2.31	3.32
Administrative Duration	Total amount of time (in seconds) spent by the visitor on account management related pages.	0 – 3398	80.81	176.7

Information	Number of pages visited by the visitor about website, communication and address information of the shopping site.	0 – 24	0.503	1.26
Informational Duration	Total amount of time (in seconds) spent by the visitor on informational pages.	0 – 2549	34.47	140.64
Product Related	Number of pages visited by visitors about product related pages like cart, search, topic etc.	0 – 705	31.73	44.45
Product Related Duration	Total amount of time (in seconds) spent by the visitor on product related pages.	0 – 63973	1194.74	1912.25
Bounce Rate	The percentage of visitors to a particular website who navigate away from the site after viewing only one page.	0 – 0.2	0.022	0.04
Exit Rate	The percentage of people who left a site from that page.	0 – 0.2	0.043	0.05
Page Value	The percentage of visitors to a particular website who navigate away from the site after viewing only one page.	0 – 361	5.88	18.55
Special Day	The closeness of the site visiting time to a specific special day (e.g. Mother's Day, Valentine's Day)	0 – 1	0.061	0.19

Table 2 - Categorical Features used in the User Behaviour Analysis Model

Name of the feature	**Description**	**Number of categorical values**
Operating System	Operating system of the visitor.	8
Browser	Browser of the visitor.	13
Region	Geographic region from which the session has been started by the visitor.	9
Traffic Type	Traffic source by which the visitor has arrived at the website (e.g. banner, SMS, direct).	20

Visitor Type	Visitor type as “New Visitor”, “Returning Visitor” and “Other”.	3
Weekend	Boolean value indicating whether the date of the visit is weekend.	2
Month	Month value of the visit date.	12
Revenue (Target Variable)	Class label indicating whether the visit generate revenue or not,	2

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.

Understanding Data Distribution

Baseline Conversion Rate

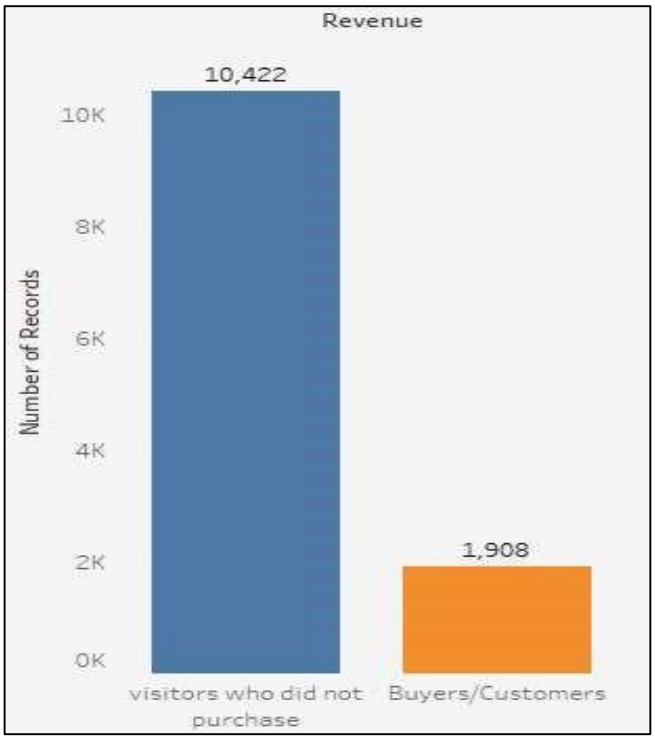


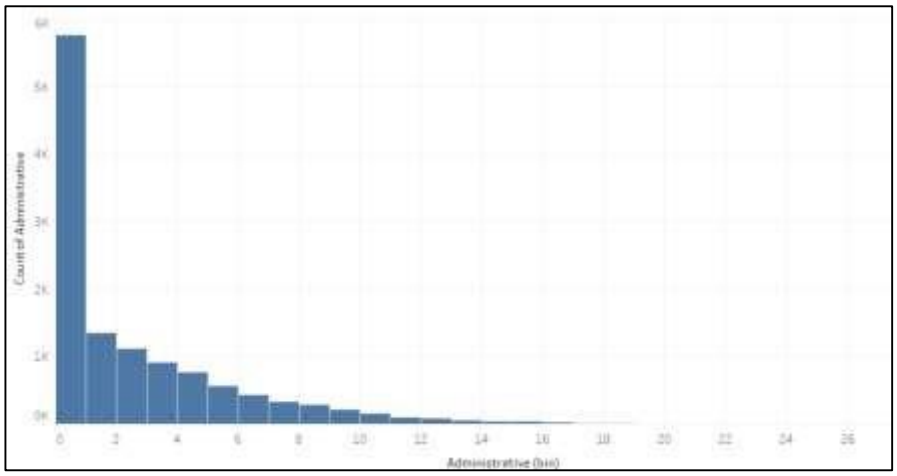
Figure 2 – count of Revenue feature

The conversion rate for the e-commerce is 1908 / 12330 = 15.47%

The model performance will suffer due to this **imbalance in the data** as most machine learning algorithms and works best when the number of instances of each classes are roughly equal. When the number of instances of one class far exceeds the other, problems arise.

Histogram for Administrative, Informational and Page Related visits

Figure 3 – Distribution of Administrative Visits on the page



As observed from figure 3,4 and 5 the distribution of the number of Administrative, Informational and Page Related visits are highly skewed with large number of zeros. The zeros represent that the type of visit by the visitor is other than that particular visit type.

We can observe similar trend for the duration of visit.

Figure 4 – *Distribution of Informational Visits on the page*

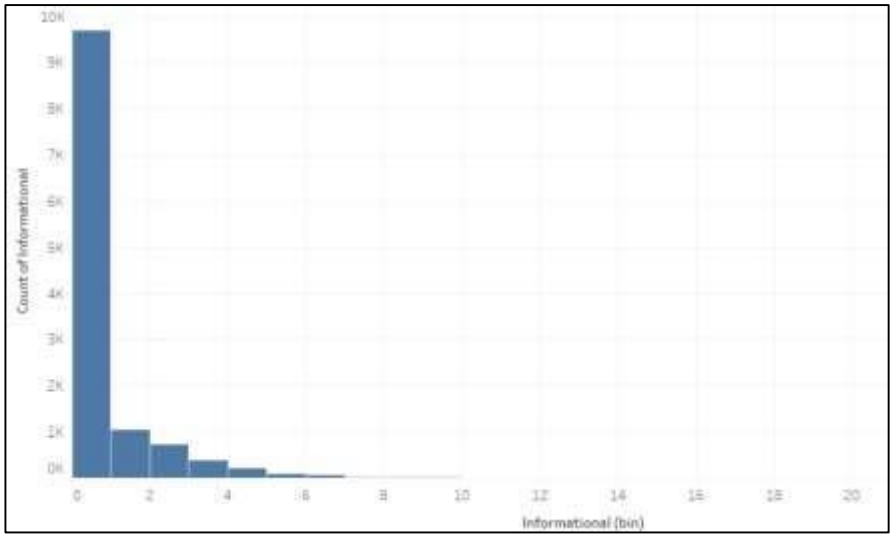
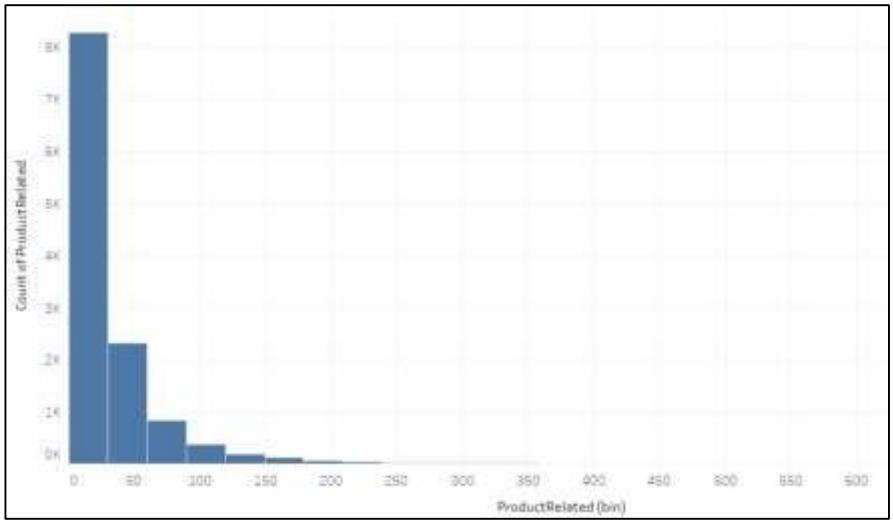


Figure 5 – *Distribution of Product Related Visits on the page*



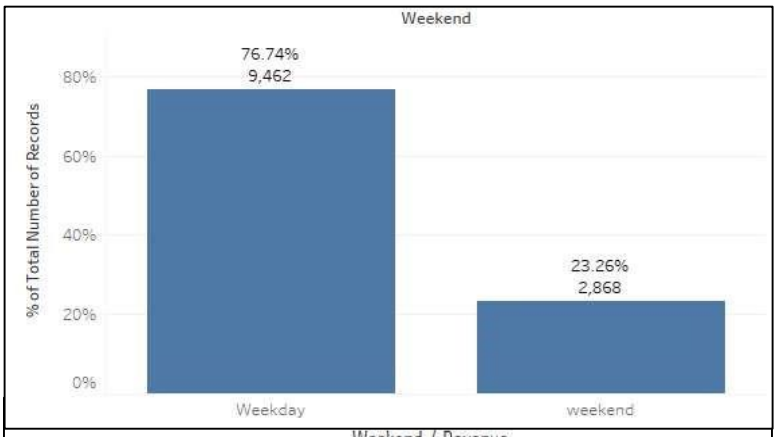
Effect of Weekend on customers

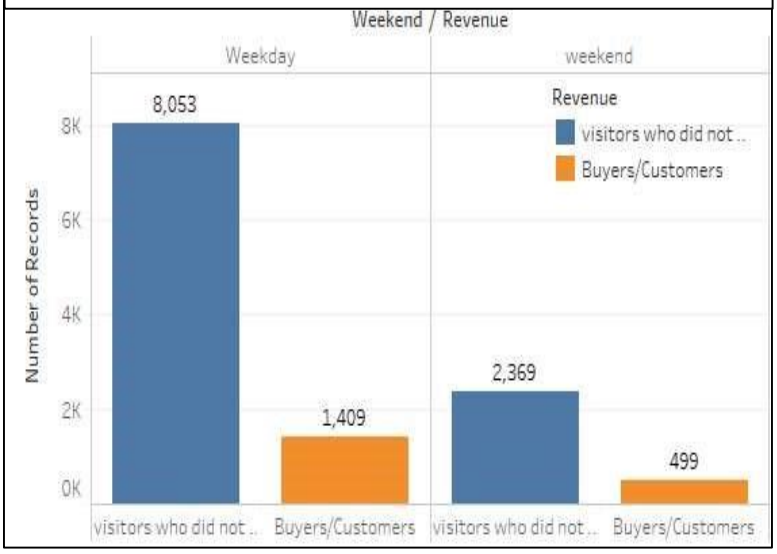
Table 3 - Weekend Visitor session vs. Revenue Generated

Revenue	Weekday	Weekend
True	85.11%	82.60%
False	14.89%	17.40%

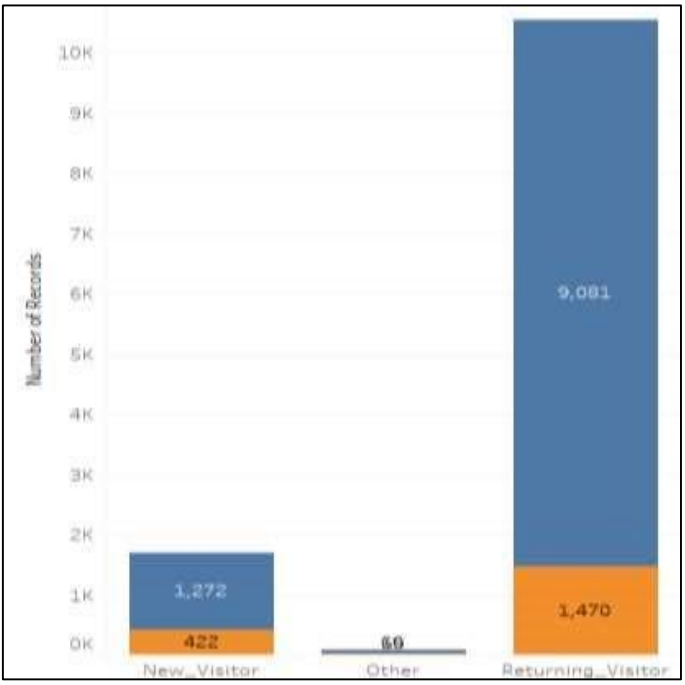
There is a high number of visitors during the weekday rather than the weekend. This may be due to the reason that customers like to shop directly in shops rather than online on weekends.

There is not a significant difference in the conversion rates between weekdays and weekends.





Visitor Type and Revenue



10%.

customers.

Visitor Type	Revenue True	False
New	24.91%	75.09%
Other	18.82%	81.18%
Returning	13.93%	86.07%

While the returning visitors are significantly more than the new visitors the conversion rate of the returning customers is less by

The company should make more efforts towards returning

Special Day

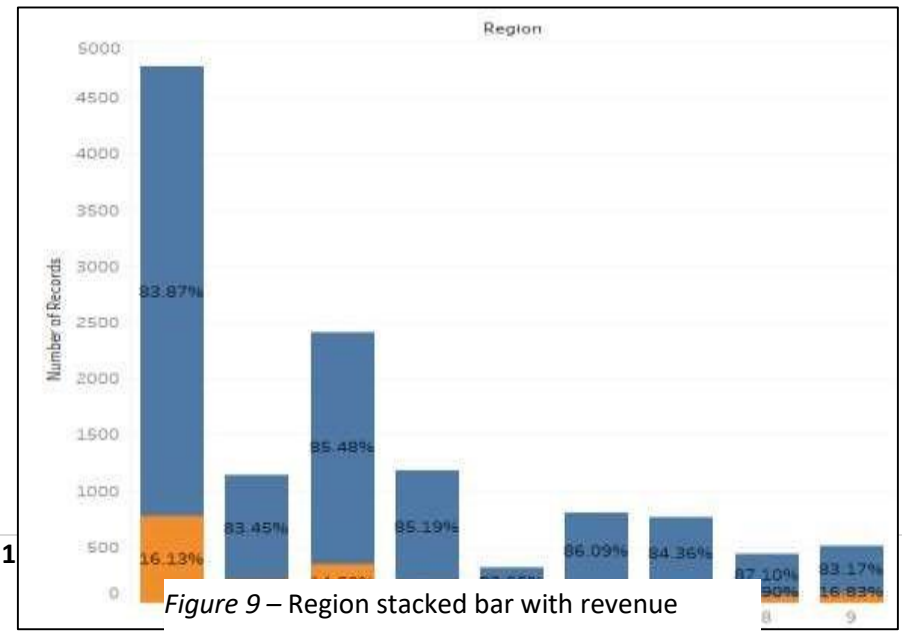
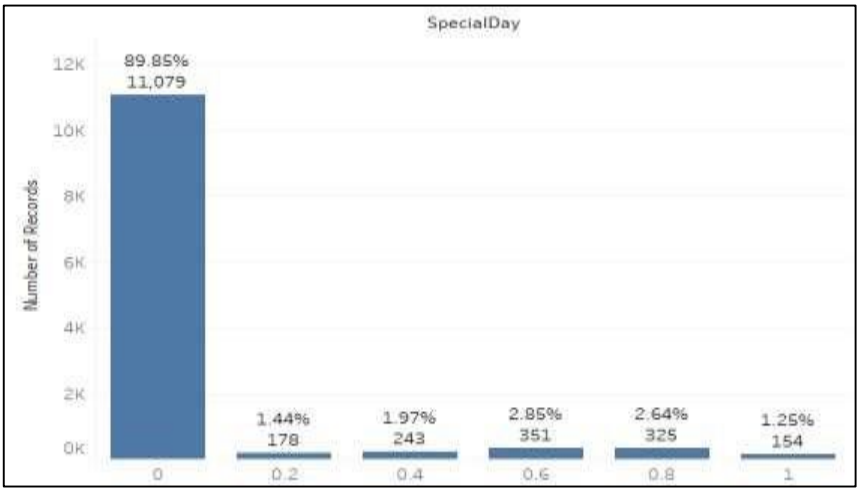


Figure 9 – Region stacked bar with revenue



Special Day does not seem to affect revenue.
People are visiting the website more and more with increasing closeness to the special day but lesser on the special day itself.
The e -commerce brings in major of its visits away from special days suggesting they do not specialize in gifting items

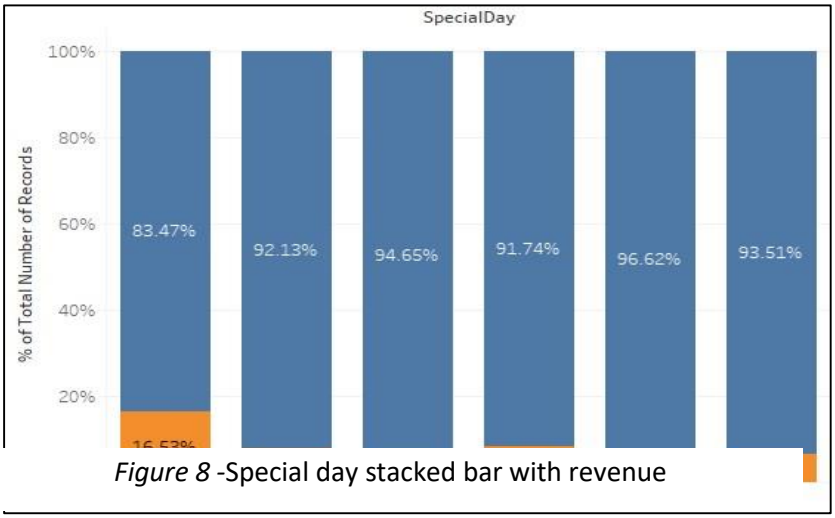


Figure 8 -Special day stacked bar with revenue

Region –

Most visitors come from region 1.
Though traffic from region 2 is lesser than that of region 3 and 4, the conversion rate for region 2 is higher. Hence, the business should try to increase the traffic for region 2.
They should also try to improve the conversion rate for region 3, 4, 6 and 8 as their conversion rate is below average.

Analysis of revenue by traffic type

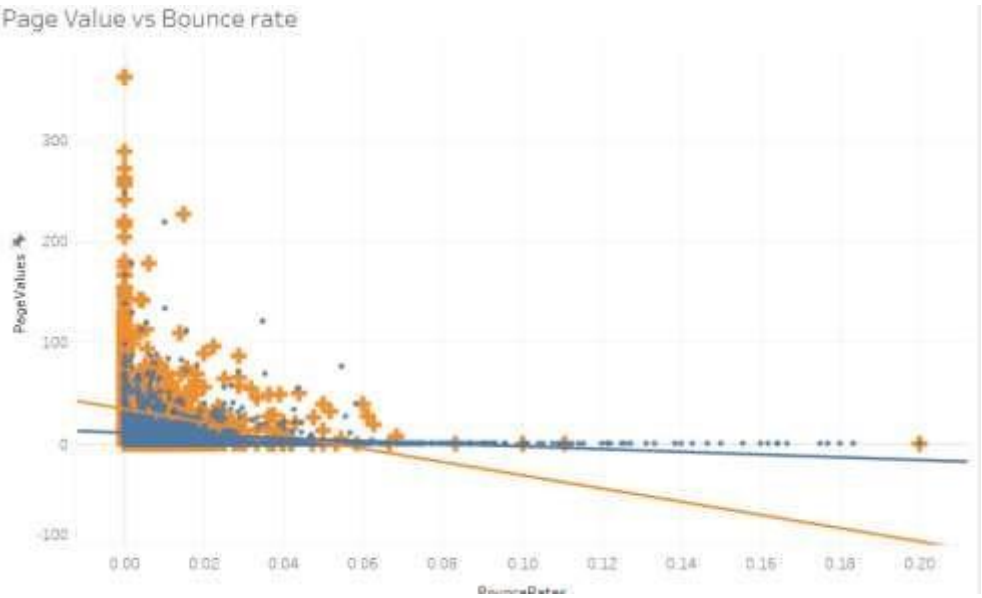
Traffic type

TrafficType		Revenue		Returning customers	
		visitors who did not pur..	Buyers/Cus..	visitors who did not pur..	Buyers/Cus..
1	19.88%	89.310%	10.690%	89.406%	10.594%
2	31.74%	78.354%	21.646%	79.501%	20.499%
3	16.64%	91.228%	8.772%	91.916%	8.084%
4	8.67%	84.565%	15.435%	86.392%	13.608%
5	2.11%	78.462%	21.538%	84.545%	15.455%
6	3.60%	88.063%	11.937%	89.526%	10.474%
7	0.32%	70.000%	30.000%	68.571%	31.429%
8	2.78%	72.303%	27.697%	73.054%	26.946%
9	0.34%	90.476%	9.524%	87.879%	12.121%
10	3.65%	80.000%	20.000%	80.000%	20.000%
11	2.00%	80.972%	19.028%	78.974%	21.026%
12	0.01%	100.000%		100.000%	
13	5.99%	94.173%	5.827%	94.118%	5.882%
14	0.11%	84.615%	15.385%	83.333%	16.667%
15	0.31%	100.000%		100.000%	
16	0.02%	66.667%	33.333%	50.000%	50.000%
17	0.01%	100.000%		100.000%	
18	0.08%	100.000%		100.000%	
19	0.14%	94.118%	5.882%	93.750%	6.250%
20	1.61%	74.747%	25.253%	77.519%	22.481%

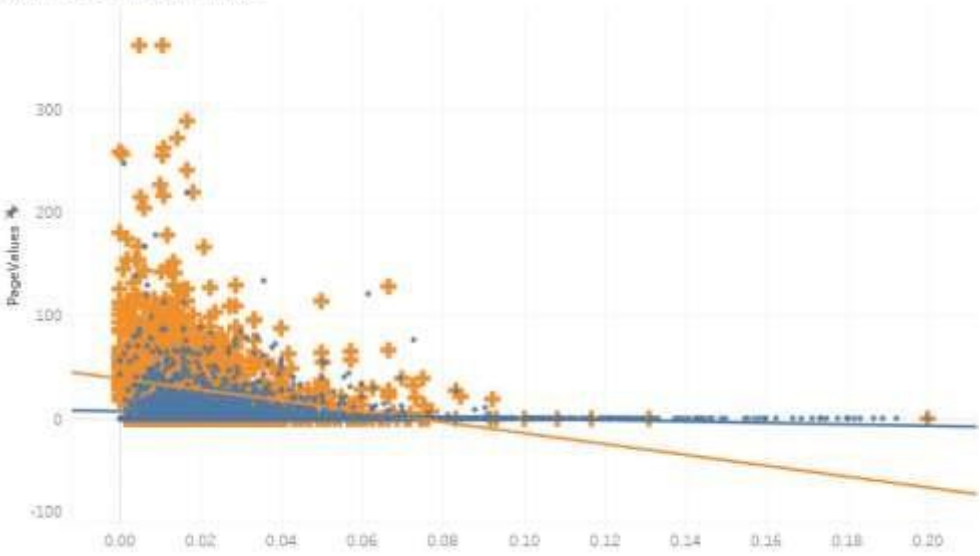
Inference:

- 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 sources might be due to wrong landing page.

Page Value vs Bounce rate



Exit rate vs Page Values



Inferences:

Exit rate and bounce rate have a direct relationship, while page value has an inverse relationship with both the exit rate and bounce rate.

For the sessions in which led to a purchase, the pages had a lower bounce rate and or higher page value.

Chapter 3 – Data Pre-Processing for Model building purposes

Treating Categorical Variables

For categorical variables to be considered in analysis we have to first convert them into its analogous numerical values which is comprehensible by the machine learning model. This was achieved by employing encoding techniques on categorical attributes.

1. SUPERVISED RATIO

Supervised Ratio was used on categorical variables except 'Weekend' and 'Revenue' to convert categorical variable into numerical values so that it can be used as inputs for machine learning models.

In the supervised Ratio Algorithm, the numerical value is a function of a number of records with the categorical value in question and how they break down between positive and negative class attribute values as follows.

Vi = pi/ti where

Vi = numerical value for the ith value of some categorical attribute

Pi = number of records with positive class value for the categorical attribute value in question ti

= total number of records with the categorical attribute value in question.

We performed will do Supervised ratio encoding for all other categorical values since OHE results in drastic increase of number of columns.

For eg this is the result of SR conversion for month column

Month	true	false	SR
Aug	76	357	0.175520
Dec	203	1462	0.121922
Feb	3	181	0.016304
Jul	66	366	0.152778
June	29	258	0.101045
Mar	192	1715	0.100682
May	365	2999	0.108502
Nov	757	2219	0.254368
Oct	115	434	0.209472
Sep	86	362	0.191964

Outlier Treatment:

Outliers were identified by using IQR method as well as by plotting boxplots for each attribute.

Imbalance Treatment:

As observed during the EDA process, the revenue column also the target variable is highly imbalanced.

Before proceeding with the model building, SMOTE was employed to overcome this issue.

SMOTE stands for Synthetic Minority Oversampling Technique. This is a statistical technique for increasing the number of cases in your dataset in a balanced way. The module works by generating new instances from existing minority cases that you supply as input.

The entire above mentioned can be summarized in the following table:

Encoding	Outlier Treatment	Standardization	Oversampling
Categorical Variable Month, Operating System, Browser, Region, Traffic Type, Visitor type are converted into numerical values using Supervised Ratio	Box plot were drawn for Independent features against Target variable and outlier were detected.	Standard Scalar function from Scikit learn library since the numerical variable are of different scale in order to obtain better performance.	Since our Dataset is highly imbalanced, we used SMOTE oversampling technique in order to tackle class imbalance situation.
Weekend and Revenue feature is converted into binary value 0's and 1's	Since the outliers are legitimate, we have decided to retain them in data		

Removing Junk Data

In the dataset there are rows with Admin Duration, Inform duration, Product duration are all 0.Which is logically not possible since in a session you must be spending some time on any of the types of pages. Hence all the row where Admin Duration, Inform Duration and Product duration equalled to Zero were removed

Chapter 3 - 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 enhances generalization by reducing overfitting.
One of the alternatives of feature selection which can be performed for model building purpose is feature extraction technique such as feature extraction techniques like PCA (Principal Component Analysis). 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.

In this project, feature selection techniques are applied to improve the classification performance and/or scalability of the system. So we tried applying different feature selection techniques

1. KBest features using Chi^2 values

One common feature selection method that is used with text data is the Chi-Square feature selection. The $\chi 2 \chi 2$ test is used in statistics to test the independence of two events. More specifically in feature selection we use it to test whether the occurrence of a specific term and the occurrence of a specific class are independent. More formally, given a document DD, we estimate the following quantity for each term and rank them by their score:

$$\chi^2(D,t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

Where

- NN is the observed frequency in and EE the expected frequency
- E_t takes the value 1 if the document contains term tt and 0 otherwise
- E_c takes the value 1 if the document is in class cc and 0 otherwise

For each feature (term), a corresponding high $\chi 2 \chi 2$ score indicates that the null hypothesis H0H0 of independence (meaning the document class has no influence over the term's frequency) should be rejected and the occurrence of the term and class are dependent.

2. Variance Inflation Factor (VIF)

Multicollinearity is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy. Issues with multicollinearity are that the standard errors of the affected coefficients tend to be large, the data redundancy in the explanatory variables may result in model overfitting.

Predictive uncertainty caused by multicollinearity thus poses a challenge for predictive environmental niche or species distribution modelling, especially when used to predict the distribution of a species under novel conditions.

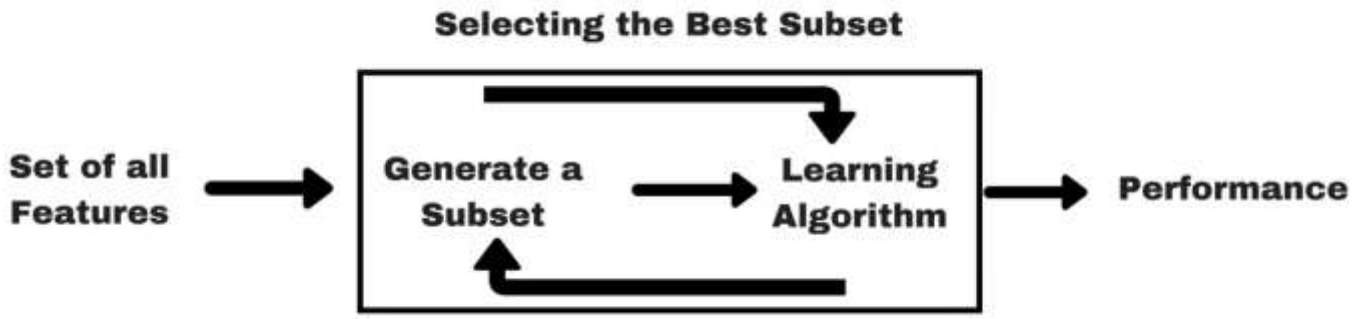
One way to detect multicollinearity is the variance inflation factor analysis. The VIF is widely used as a measure of the degree of multi-collinearity of the i^{th} independent variable with the other independent variables in a regression model. If we have explanatory variables X_1, X_2, X_3, … X_i, the VIF for an explanatory variable X_1 can be calculated by running an ordinary least square regression that has X_1 as a function of all the other explanatory variables X_2 … X_i. The VIF is than computed following Equation,

$$\mathrm{VIF} = \frac{1}{1 - R^2}$$

where R^2 is the coefficient of determination of the regression equation. This can be repeated for each of the explanatory variables. The size of VIF gives the magnitude of the multicollinearity. The square root of the VIF shows how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other predictor variables in the model.

3. Wrapper method

In wrapper methods, we try to use a subset of features and train a model using them. Based on the inferences that we draw from the previous model, we decide to add or remove features from your subset. The problem is essentially reduced to a search problem.



4. Backward Elimination

In backward elimination, we start with all the features and remove the least significant feature in each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.

4.1 Recursive Feature Elimination

It is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination.

In the following Table we can see the features selected in each Feature selection technique.

S.NO	KBest features Using chi2 values	VIF	Backward Elimination	RFE
1	Page Values	Page Values	Page Values	Page Values
2	Exit Rates	Exit Rates	Exit Rates	Exit Rates
3	Product Related	Product Related	Product Related	Product Related
4	Product Related Duration	Product Related Duration	Product Related Duration	Product Related Duration
5	Bounce Rates	Bounce Rates	Bounce Rates	Bounce Rates
6	Administrative	Administrative	Administrative	Administrative
7	Returning_visitor	Returning_visitor	Returning_visitor	Returning_visitor
8	Other_visitor	Other_visitor	Other_visitor	Other_visitor
9	Informational	Informational	Informational	Informational
10	Administrative Duration	Administrative Duration	Administrative Duration	Administrative Duration
11	Special Day	Special Day	Special Day	Special Day
12	Month	Month	Month	Month
13	Informational Duration	Informational Duration	Informational Duration	Informational Duration
14	Weekend	Weekend	Weekend	Weekend
15	Browser	Browser	Browser	Browser
16	Operating Systems	Operating Systems	Operating Systems	Operating Systems
17	Region	Region	Region	Region
18	Traffic Type	Traffic Type	Traffic Type	Traffic Type

From the above process we can see that different combination features were selected.

Hence we decided to build models with all the combinations.

Chapter 4 – Model Building

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.

ROC chart & Area under the curve (AUC)

ROC chart is a plot of 1-specificity in the X axis and sensitivity in the Y axis. Area under the ROC curve is a measure of model performance. The AUC of a random classifier is 50% and that of a perfect classifier is 100%. For practical situations, an AUC of over 70% is desirable.

Level of significance

For all the hypothesis tests in the project, the level of significance is assumed as 5% unless specified otherwise.

Algorithms used for building models

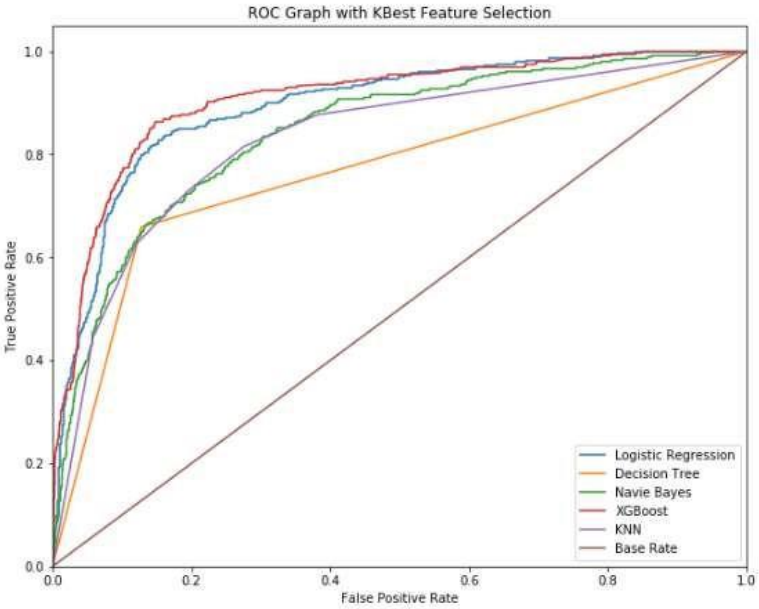
- **Logistic Regression**
- **KNN**
- **Naïve Bayes**
- **Decision Tress**
- **XGBoost**

Results obtained when no feature selection was done:

	Model	Accuracy	F1Score
0	Logistic Regression	0.882860	0.527778
1	Decission Tree	0.846684	0.520216
2	KNN	0.880706	0.528109
3	Navie Bayes	0.840224	0.551391

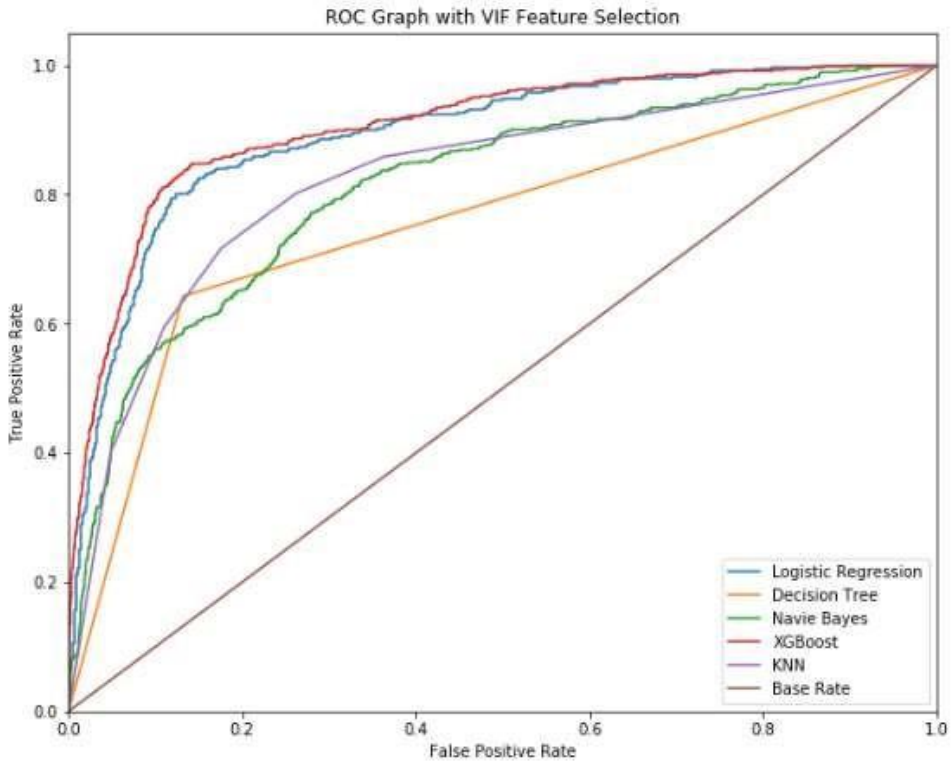
Results obtained when KBest features using chi^2 Values were selected:

	Model	Accuracy	F1Score	Sensitivity	Specificity	FPR	FNR
0	Logistic Regression_KB	0.874246	0.642157	0.948498	0.572052	0.427948	0.051502
1	Decissiom Tree_KB	0.839363	0.558580	0.933515	0.484600	0.515400	0.066485
2	KNN_KB	0.793282	0.520958	0.942193	0.405280	0.594720	0.057807
3	Navie Bayes_KB	0.772610	0.503759	0.944307	0.379603	0.620397	0.055693
4	XGBoost_KB	0.872524	0.655012	0.957739	0.562000	0.438000	0.042261



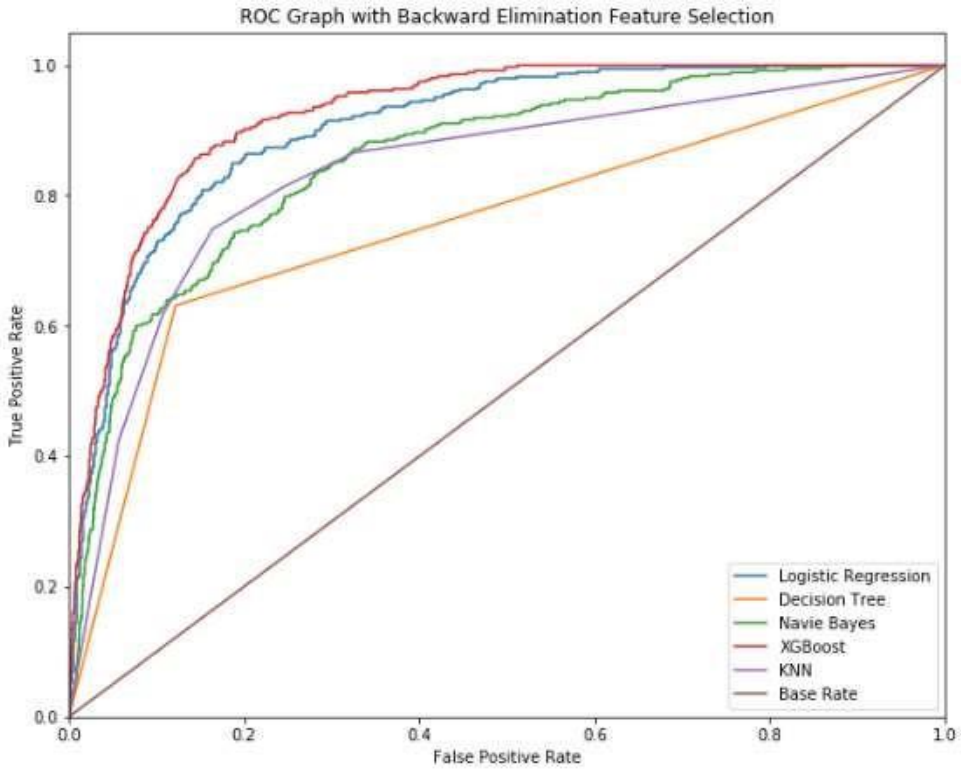
Results obtained when features using VIF were selected:

	Model	Accuracy	F1Score	Sensitivity	Specificity	FPR	FNR
0	Logistic Regression_VIF	0.875108	0.653938	0.944823	0.593931	0.406069	0.055177
1	Decission Tree_VIF	0.831754	0.553354	0.926170	0.485944	0.514056	0.073830
2	KNN_VIF	0.806776	0.546190	0.937622	0.441176	0.558824	0.062378
3	Navie Bayes_VIF	0.750502	0.479329	0.930643	0.362319	0.637681	0.069357
4	XGBoost_VIF	0.880850	0.685368	0.958593	0.599469	0.400531	0.041407



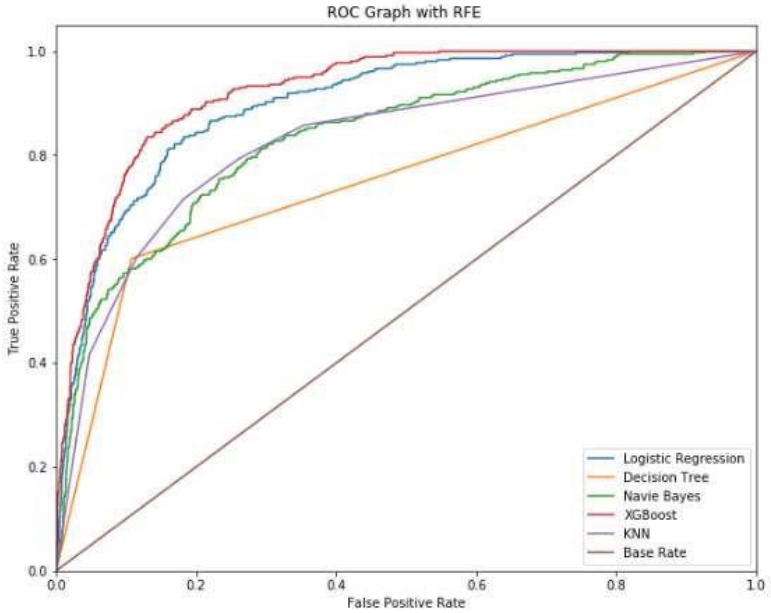
Results obtained when features using Backward Elimination were selected:

	Model	Accuracy	F1Score	Sensitivity	Specificity	FPR	FNR
0	Logistic Regression_BE	0.850560	0.615725	0.954980	0.510092	0.489908	0.045020
1	Decission Tree_BE	0.840224	0.549210	0.928918	0.486022	0.513978	0.071082
2	KNN_BE	0.822567	0.565401	0.948037	0.454237	0.545763	0.051963
3	Navie Bayes_BE	0.828596	0.542529	0.932597	0.460938	0.539062	0.067403
4	XGBoost_BE	0.878553	0.662679	0.956074	0.579498	0.420502	0.043926



Results obtained when features using Recursive Feature Elimination were selected:

	Model	Accuracy	F1Score	Sensitivity	Specificity
0	Logistic Regression_RFE	0.840224	0.601504	0.955403	0.488656
1	Decission Tree_RFE	0.847545	0.548469	0.924578	0.504695
2	KNN_RFE	0.802326	0.527291	0.940316	0.417618
3	Navie Bayes_RFE	0.742463	0.482699	0.948163	0.349624
4	XGBoost_RFE	0.876830	0.660333	0.956474	0.574380



Results

	Model	Accuracy	F1Score	Sensitivity	Specificity
0	Logistic Regression_KB	0.874246	0.642157	0.948498	0.572052
1	Decissiom Tree_KB	0.839363	0.558580	0.933515	0.484600
2	KNN_KB	0.793282	0.520958	0.942193	0.405280
3	Navie Bayes_KB	0.772610	0.503759	0.944307	0.379603
4	XGBoost_KB	0.872524	0.655012	0.957739	0.562000
0	Logistic Regression_VIF	0.875108	0.653938	0.944823	0.593931
1	Decission Tree_VIF	0.831754	0.553354	0.926170	0.485944
2	KNN_VIF	0.806776	0.546190	0.937622	0.441176
3	Navie Bayes_VIF	0.750502	0.479329	0.930643	0.362319
4	XGBoost_VIF	0.880850	0.685368	0.958593	0.599469
0	Logistic Regression_BE	0.850560	0.615725	0.954980	0.510092
1	Decission Tree_BE	0.840224	0.549210	0.928918	0.486022
2	KNN_BE	0.822567	0.565401	0.948037	0.454237
3	Navie Bayes_BE	0.828596	0.542529	0.932597	0.460938
4	XGBoost_BE	0.878553	0.662679	0.956074	0.579498
0	Logistic Regression_RFE	0.840224	0.601504	0.955403	0.488656
1	Decission Tree_RFE	0.847545	0.548469	0.924578	0.504695
2	KNN_RFE	0.802326	0.527291	0.940316	0.417618
3	Navie Bayes_RFE	0.742463	0.482699	0.948163	0.349624
4	XGBoost_RFE	0.876830	0.660333	0.956474	0.574380

We can see from the above table that the highest accuracy score and F1- score was achieved in XGBoost model which had the features selected from the VIF process.

Chapter 5- Conclusion

In this project, we aim to construct a real-time user behaviour analysis system for online shopping environment. We use an online retailer data to perform the experiments. In order to predict the purchasing intention of the visitor, we use aggregated page view data kept track during the visit along with some session and user information as input to machine learning algorithms. Oversampling and feature selection pre-processing techniques are applied to improve the success rates and scalability of the algorithms. The best results are achieved with XGBoost algorithm. Our findings support the argument that the features extracted from clickstream data during the visit convey important information for online purchasing intention prediction.

The findings show that choosing a minimal subset of combination of clickstream data aggregated statistics and session information such as the date and geographic region results in a more accurate and scalable system. 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.