



# Lending Club Loan



**Abstract:**

**Lending Club**, a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return. Each **borrower** fills out a comprehensive application, providing their past financial history, the reason for the loan, and more. Lending Club evaluates each borrower’s credit score using past historical data (and their own data science process!) and assigns an interest rate to the borrower. Approved loans are listed on the Lending Club website, where qualified investors can browse recently approved loans, the borrower’s credit score, the purpose for the loan, and other information from the application. Once an investor decides to fund a loan, the borrower then makes monthly payments back to Lending Club. Lending Club redistributes these payments to investors. This means that investors don’t have to wait until the full amount is paid off to start to see returns. If a loan is fully paid off on time, the investors make a return which corresponds to the interest rate the borrower had to pay in addition to the requested amount. Many loans aren’t completely paid off on time, however, and some borrowers [default](https://www.lendingclub.com/public/collections-process.action) on the loan. That’s the problem we’ll be trying to address as we clean some data from Lending Club for machine learning.

**Keywords**

: Lending Club Loan Analysis, Machine Learning, Tracking Data, Oversampling



# Table of Contents

[Chapter 1 - Project overview](#_bookmark0) [08](#_bookmark0)

[Need for study](#_bookmark0) [08](#_bookmark0)

[Current baseline & business mission](#_bookmark0) [08](#_bookmark0)

[Problem statement & project scope](#_bookmark1) [08](#_bookmark1)

[Data sources](#_bookmark1) [09](#_bookmark1)

Dataset Description [09](#_bookmark1)

[Data preparation & cleanup](#_bookmark2) [11](#_bookmark2)

[Statistical tools & techniques](#_bookmark2) [11](#_bookmark2)

[Model performance measures used for evaluating models 12](#_TOC_250001)

[Chapter 2 - Exploratory data analysis](#_bookmark3) [14](#_bookmark3)

[Understand data distribution](#_bookmark3) [14](#_bookmark3)

Insights into feature selection 15

[Chapter 3 - Feature Selection](#_bookmark4) & Model Building [22](#_bookmark4)

[Classification](#_bookmark4) Results [22](#_bookmark4)

Results on class imbalance dataset 22

[Results obtained with oversampling 23](#_TOC_250000)

Results obtained by oversampling and feature selection 23

Insights obtained from Odds Ratio Exploration 25

Variable importance plot for Tree based algorithm 25

Chapter 4 - Conclusion 27

Chapter 5 – Recommendations 28

Chapter 6 – References 29

[Chapter 7 - Appendix](#_bookmark5) [30](#_bookmark5)

Abbreviations used

|  |  |
| --- | --- |
| **Abbreviation** | **Expansion** |
| LR | Logistic Regression |
| DT | Decision Tree |
| AUC | Area Under the 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 |
| mRMR | Minimum Redundancy and maximum relevance |
| 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 behavioural prediction systems which imitate the behaviour 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 behaviour 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 analysing 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 purchasing 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 clickstream data obtained from the navigation path followed during the online visit convey important information about the purchasing 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 customers and to increase new customer visits by creating better landing page with high page value.



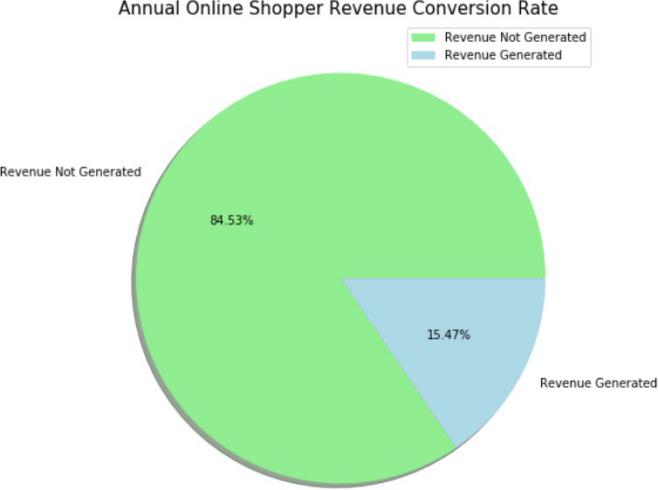
# Chapter 1 - Project overview

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 & behavioural analysis

The Columbia Sportswear Company is an American company that manufactures and distributes outerwear, sportswear, and footwear, as well as headgear, camping equipment, ski apparel, and outerwear accessories. It was founded in 1938 by Paul Lamfrom. The company is headquartered in Cedar Mill, an unincorporated area in Washington County, Oregon, in the Portland metropolitan area near Beaverton. Columbia Sportswear distributes its products in more than 72 countries and 13,000 retailers. Columbia also operates its own chain of retail stores, including its flagship store located in downtown Portland, Oregon

## 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 behavioural 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

**Figure 1.1 – Need for study**

## Current baseline & Business mission

The current yearly conversion rate of online shoppers in Columbia Sportswear Company website is 15.47 % and business mission aims to increase the conversion rate by 10 %

## Problem statement & project scope

To support the incremental revenue contribution from online shoppers, the business would require pointers about customer behaviour & focus areas. To provide these deliverables, the project would analyze data pertaining to online user session of **Columbia Sportswear Company for a period of 1 year**.

## Data sources

In order to classify consumer on-site behaviour, 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

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature Name** | **Feature Description** | **Min Value** | **Max Value** | **Std Dev.** |
| Administrative | Number of pages visited by the visitor about account management | 0 | 27 | 3.32 |
| Administrative Duration | Total amount of time (in seconds) spent by the visitor on account management related pages | 0 | 3398 | 176.7 |
| Informational | Number of pages visited by the visitor about website, communication and address information of the shopping site | 0 | 24 | 1.26 |
| Informational Duration | Total amount of time (in seconds) spent by the visitor on informational pages | 0 | 2549 | 140.64 |
| Product Related | Number of pages visited by visitor about product related pages | 0 | 705 | 44.45 |
| Product Related Duration | Total amount of time (in seconds) spent by the visitor on product related pages | 0 | 63973 | 1912.25 |
| Bounce Rate | Average bounce rate value of the pages visited by the visitor | 0 | 0.2 | 0.04 |
| Exit Rate | Average exit rate value of the pages visited by the visitor | 0 | 0.2 | 0.05 |
| Page Value | Average page value of the pages visited by the visitor | 0 | 361 | 18.55 |
| Special Day | Closeness of the site visiting time to a special day | 0 | 1 | 0.19 |

***Table 1*** shows the numerical features along with their statistical parameters. 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. ***Table 2*** shows this URL information of the pages visited by the user by their category types.

Table 2 – Page Types Derived By URL

|  |  |
| --- | --- |
| **Page Type** | **URL** |
| Administrative | /?login |
| Administrative | /?logout |
| Administrative | /LoginRegister |
| Administrative | /login' |
| Administrative | /passwordrecovery |
| Administrative | /?ref |
| Administrative | /?refer |

|  |  |
| --- | --- |
| Administrative | /?returnurl |
| Administrative | /customer |
| Administrative | /emailwishlist |
| Administrative | /omnicards |
| Product Related | / |
| Product Related | /c |
| Product Related | /urun, |
| Product Related | /search |
| Product Related | /cart |
| Informational | /Topic |
| Informational | /t-popup |
| Informational | /t |
| Informational | /contactus |
| Informational | /Catalog |
| Informational | /stores |

The "Bounce Rate", "Exit Rate" and "Page Value" features shown in Table 4 represent the metrics measured by "Google Analytics" for each page in the e-commerce site. These values can be stored in the system for all web pages of the e-commerce site in the developed system and updated automatically at regular intervals. The value of "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session. The value of “Exit Rate” feature for a specific web page is calculated as for all page views to the page, the percentage that were the last in the session. The “Page Value” feature represents the average value for a web page that a user visited before completing an E- commerce transaction. The “Special Day” feature indicates the closeness of the site visiting time to the special days (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina’s day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8.

Table 3 – Categorical Features used in the User Behaviour Analysis Model

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Feature Description** | **Number of**  **Categorical Values** |
| Operating Systems | 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 | Class label indicating whether the visit has been  finalized with a transaction | 2 |

***Table 3*** shows the categorical features along with their categorical values. The "Operating Systems", " Browser", " Traffic Type" and “Visitor Type” features shown in ***Table 3*** 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** | **Standardization** | **Oversampling** |
| Categorical Variable Month, Operating System, Browser, Region, Traffic Type, Visitor type are converted into Dummy variable | Box plot is drawn for Independent features against Target variable and outlier had been 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. |
| 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 |

## 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

## 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.



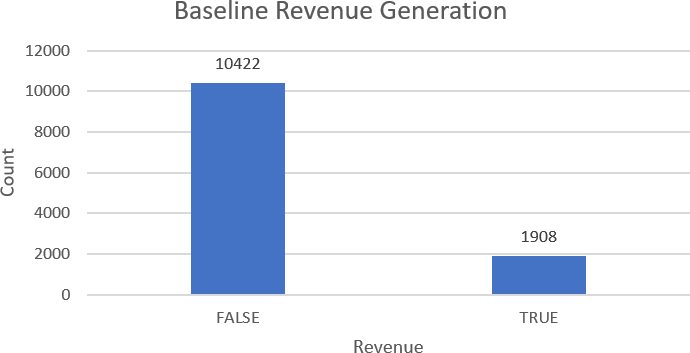
# Chapter 2 - Exploratory data analysis

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 Online visitor’s vs overall visitors is = 1908/12330 = 15.47 %.**

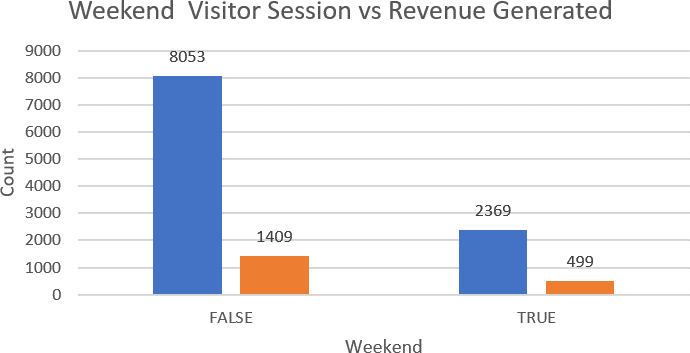
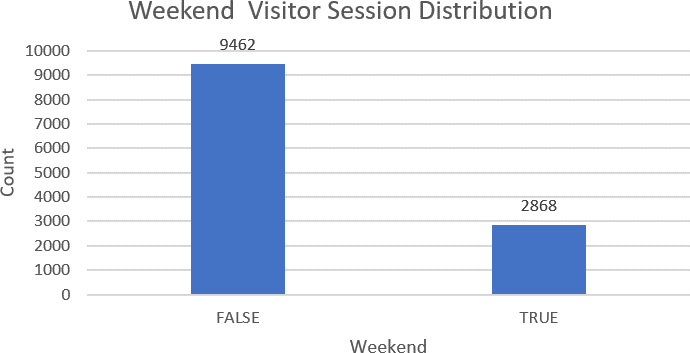
This indicates that the data set is an imbalanced dataset where the number of observations belonging to class 1 (True) is significantly lower than those belonging to class 0 (False)

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

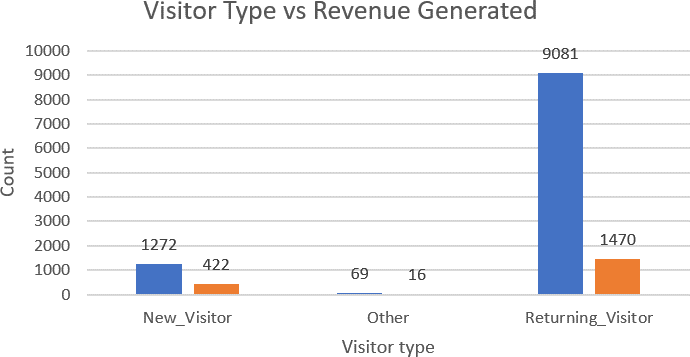
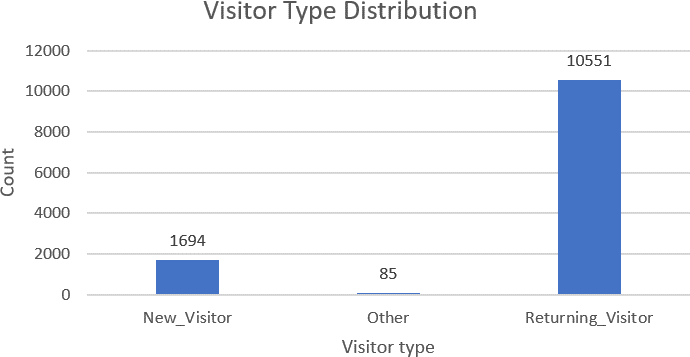


Revenue Conversion Rate

|  |  |  |
| --- | --- | --- |
| **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

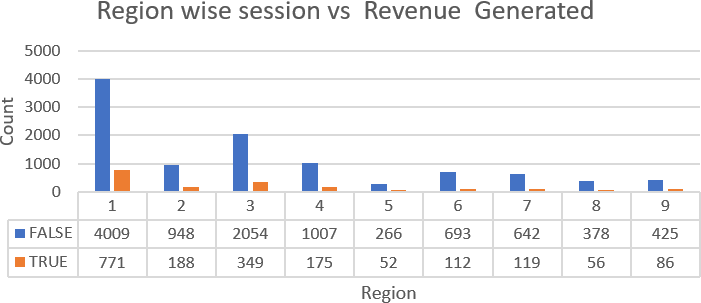
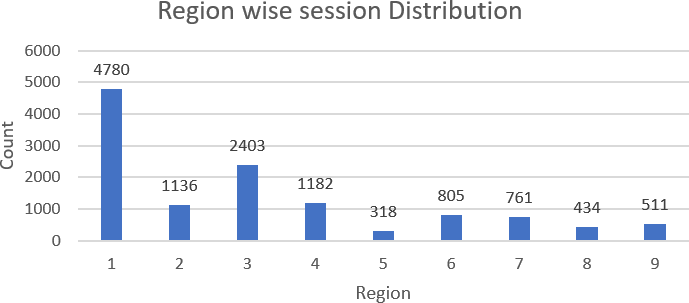
|  |  |  |
| --- | --- | --- |
| **Visitor Type** | **FALSE** | **TRUE** |
| New Visitor | 75.09% | 24.91% |
| Other | 81.18% | 18.82% |
| Returning Visitor | 86.07% | 13.93% |

* + 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

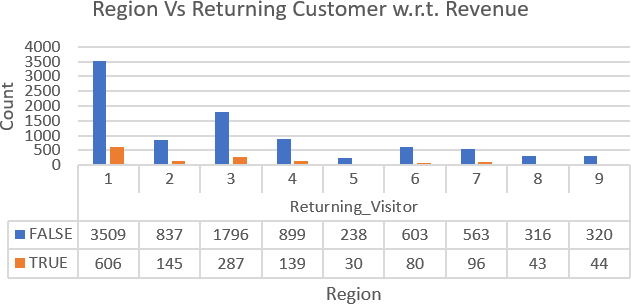
|  |  |
| --- | --- |
| **Region** | **% Visitors** |
| Region 1 | 38.77% |
| Region 2 | 9.21% |
| Region 3 | 19.49% |
| Region 4 | 9.59% |
| Region 5 | 2.58% |
| Region 6 | 6.53% |
| Region 7 | 6.17% |
| Region 8 | 3.52% |
| Region 9 | 4.14% |

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



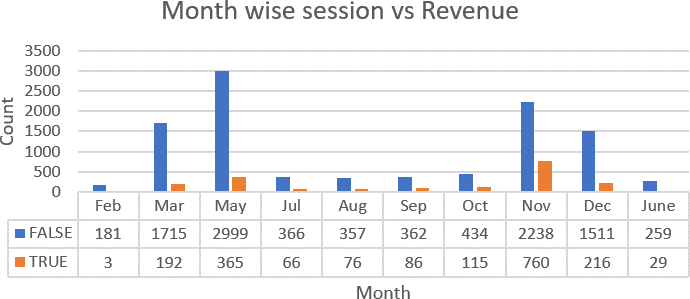
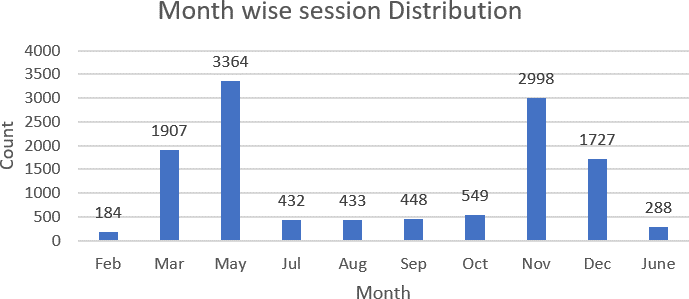
|  |  |  |
| --- | --- | --- |
| **Region** | **FALSE** | **TRUE** |
| Region 1 | 85.27% | 14.73% |
| Region 2 | 85.23% | 14.77% |
| Region 3 | 86.22% | 13.78% |
| Region 4 | 86.61% | 13.39% |
| Region 5 | 88.81% | 11.19% |
| Region 6 | 88.29% | 11.71% |
| Region 7 | 85.43% | 14.57% |
| Region 8 | 88.02% | 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

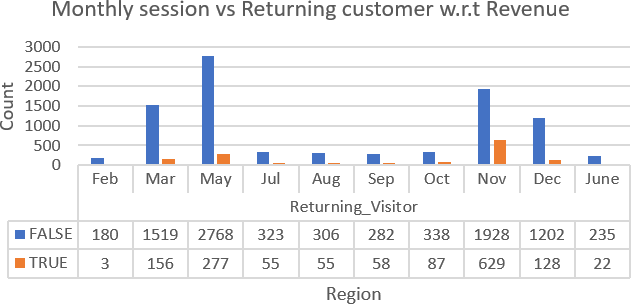
|  |  |
| --- | --- |
| **Month** | **% visitors** |
| Feb | 1.49% |
| Mar | 15.47% |
| May | 27.28% |
| June | 2.34% |
| Jul | 3.50% |
| Aug | 3.51% |
| Sep | 3.63% |
| Oct | 4.45% |
| Nov | 24.31% |
| Dec | 14.01% |

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% |

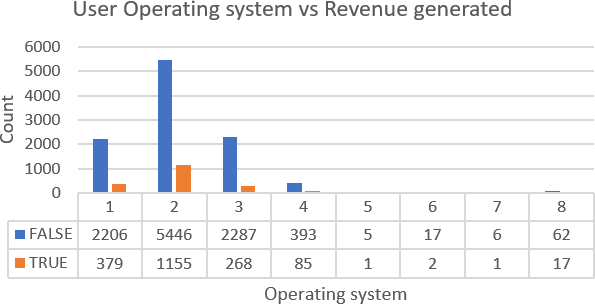
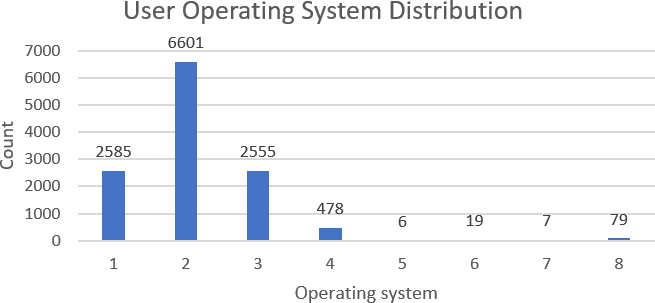
* + 81% of online user session is found in the month of March, May, November and December.
  + Conversion rate of March, May and December is very low when compared to November Month.
  + Even returning customer conversion rate is low on these three months.
  + More offers can be given in these months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

### Visitor Traffic Source vs Revenue Generated

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | | **Traffic Conversion Rate** | | **Returning Customer Conversion Rate** | |
| **Traffic**  **Source** | **%**  **Visitor** | **FALSE** | **TRUE** | **FALSE** | **TRUE** |
| Source 1 | 19.88% | 89.31% | 10.69% | 89.41% | 10.59% |
| Source 2 | 31.74% | 78.35% | 21.65% | 79.50% | 20.50% |
| 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 | 3.60% | 88.06% | 11.94% | 89.53% | 10.47% |
|  | |
| Source 7 | 0.32% | 70.00% | 30.00% | 68.57% | 31.43% |
| Source 8 | 2.78% | 72.30% | 27.70% | 73.05% | 26.95% |
| Source 9 | 0.34% | 90.48% | 9.52% | 87.88% | 12.12% |
| Source 10 | 3.65% | 80.00% | 20.00% | 80.00% | 20.00% |
| Source 11 | 2.00% | 80.97% | 19.03% | 78.97% | 21.03% |
| Source 12 | 0.01% | 100.00% | 0.00% | 100.00% | 0.00% |
| 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% | 100.00% | 0.00% | 100.00% | 0.00% |
|  | |
| Source 16 | 0.02% | 66.67% | 33.33% | 50.00% | 50.00% |
|  | |
| Source 17 | 0.01% | 100.00% | 0.00% | 100.00% | 0.00% |
| Source 18 | 0.08% | 100.00% | 0.00% | 100.00% | 0.00% |
| 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.

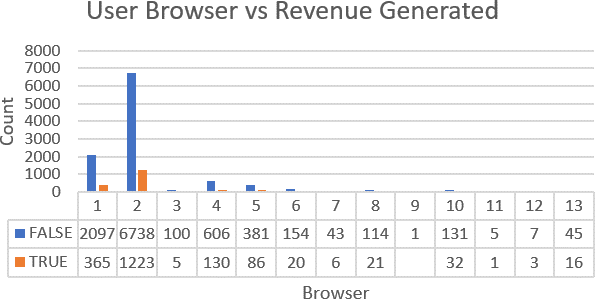
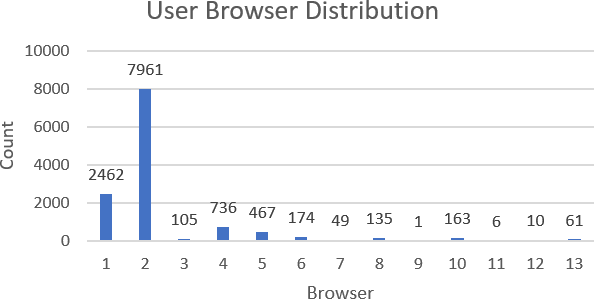
### Visitor Operating system vs Revenue Generated



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **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.

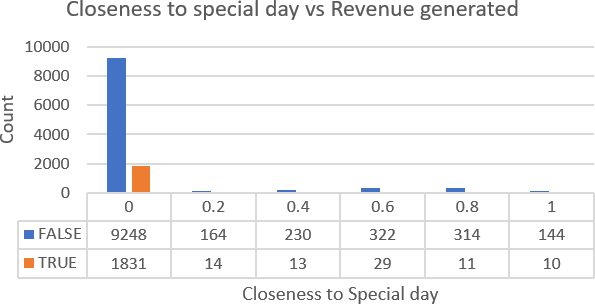
### Visitor Browser vs Revenue Generated



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | **Browser Conversion**  **Rate** | | **Returning Customer**  **Conversion Rate** | |
| **Browser Type** | **% Visitor** | **FALSE** | **TRUE** | **FALSE** | **TRUE** |
| Browser 1 | 19.97% | 85.17% | 14.83% | 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 | 0.40% | 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% | 73.77% | 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.

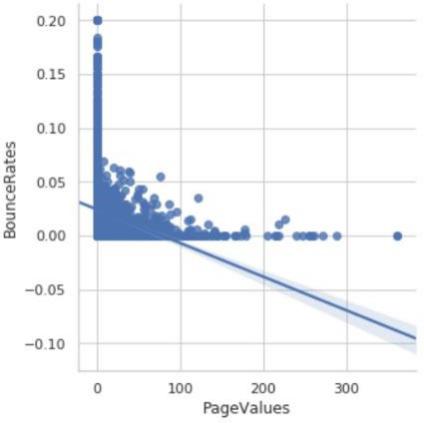
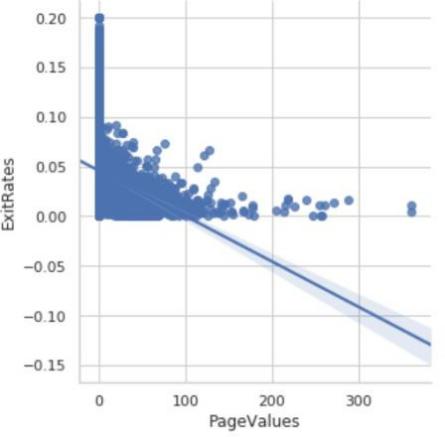
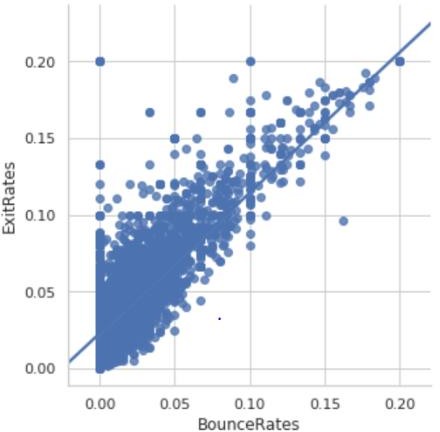
### Special Day vs Revenue Generated



|  |  |
| --- | --- |
| **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% |

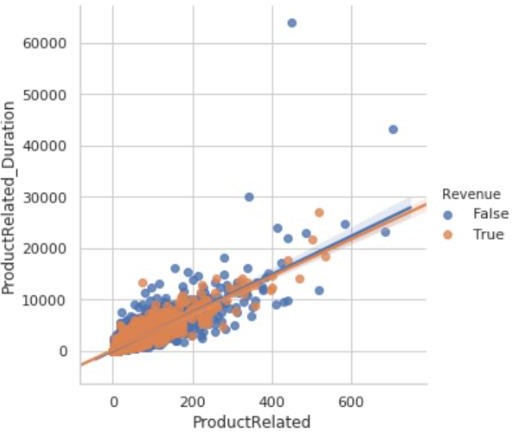
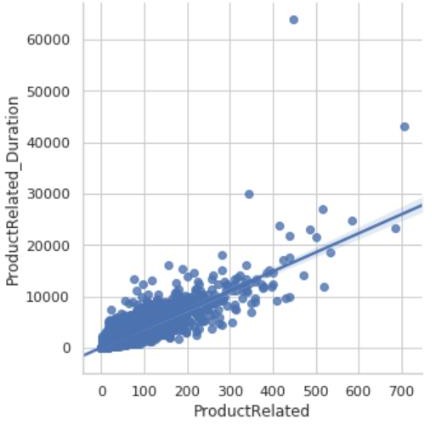
* + 90% online session happen on non-special days.
  + Since its sportswear company there is no affinity for special days to revenue generation.

### 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.

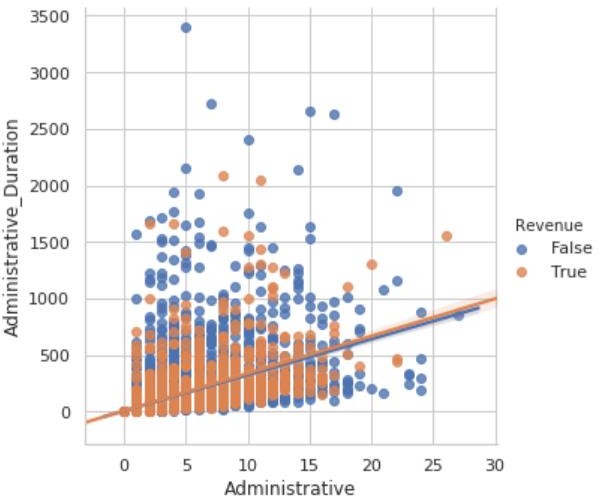
### Product related pageviews vs product page duration vs Revenue generated



* + 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

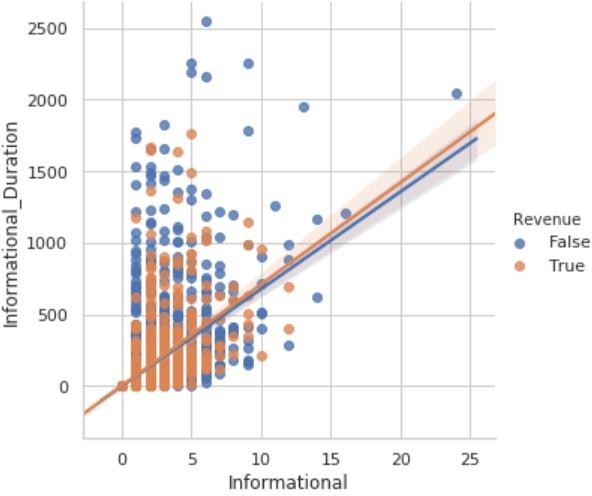
conversion.

### 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.

# Chapter 3 - Feature Selection & Model Building

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. For feature ranking, instead of wrapper algorithms that require a learning algorithm to be used and consequently can result in reduced feature sets specific to that classifier, filter-based algorithms are tested in which no classification algorithm is used. Correlation Attribute Evaluation, Information Gain Attribute Evaluation and Minimum Redundancy Maximum Relevance Filters were used in our experiments. In mRMR algorithm, the aim is to maximize the relevance between the selected set of features and class variable while avoiding the redundancy among the selected features. Thus, maximum classification accuracy is aimed to be obtained with minimal subset of features.

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.

## 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 behaviour 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.

## Results on class imbalanced dataset:

Tables below show the results obtained with Logistic Regression, Decision Tree, Random forest, Light gradient boosting, bagging classifier (Logistic regression & Decision Tree) and boosting classifier ( Logistic regression & Decision Tree) respectively. The results show that Light Gradient Boosting ( LGBM) gives the highest accuracy rate on test set. However, a class imbalance problem arises since the number of negative class instances in the data set is much higher than that of the positive class instances, and the imbalanced success rates on positive (TPR) and negative (TNR) samples show that the classifiers tend to label the test samples as the majority class. This class imbalance problem is a natural situation for the problem since most of the e-commerce visits do not end with shopping.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Train\_accuracy** | **Test\_accuracy** | **Test\_Precision** | **Bias Error** | **Variance Error** | **F1-score** |
| LGBM | 0.96675 | 0.89862 | 0.74713 | 0.10138 | (+/- 0.00008) | 0.78 |
| Bag\_DT | 0.88356 | 0.89754 | 0.67974 | 0.10246 | (+/- 0.00009) | 0.78 |
| RF | 0.99096 | 0.89186 | 0.67647 | 0.10814 | (+/- 0.00020) | 0.76 |
| DT | 1.00000 | 0.86348 | 0.53578 | 0.13652 | (+/- 0.00024) | 0.73 |
| Boost\_DT | 0.86919 | 0.86267 | 0.53358 | 0.13733 | (+/- 0.00021) | 0.72 |
| Bag\_LR | 0.88356 | 0.88781 | 0.74713 | 0.11219 | (+/- 0.00021) | 0.71 |
| LR | 0.88310 | 0.88754 | 0.74615 | 0.11246 | (+/- 0.00021) | 0.71 |
| Boost\_LR | 0.86919 | 0.87402 | 0.77465 | 0.12598 | (+/- 0.00014) | 0.63 |

## Results obtained with oversampling:

The results presented in this section show that the classifiers tend to minimize their errors on majority class samples, which leads to an imbalance between the accuracy rates of the positive and negative classes. However, in a real-time user behaviour analysis model, correctly identifying directed buying visits, which are represented with positive class in our dataset, is as important as identifying negative class samples. Therefore, a balanced classifier is needed to increase the conversion rates in an e-commerce website. To deal with class imbalance problem, we use oversampling method, in which a uniform distribution over the classes is aimed to be achieved by adding more of the minority (positive class in our dataset) class instances. Since this dataset is created by selecting multiple instances of the minority class more than once, first oversampling the dataset and then dividing it into training and test sets may lead to biased results due to the possibility that the same minority class instance may be used both for training and test. For this reason, in our study, 30 percentage of the data set consisting of 12330 samples is first left out for testing and the oversampling method is applied to the remaining 70 percentage of the samples.

The results obtained on the balanced dataset are shown in the table below. Since the number of samples belonging to positive and negative classes is equalized with oversampling, both accuracy and F1-score metrics can be used to evaluate the results.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model with**  **oversampling** | **Train\_accuracy** | **Test\_accuracy** | **Test\_Precision** | **Bias**  **Error** | **Variance Error** | **F1-score** |
| LGBM | 0.97080 | 0.93540 | 0.93318 | 0.06460 | (+/- 0.00001) | 0.94 |
| RF | 0.99589 | 0.92581 | 0.91991 | 0.07419 | (+/- 0.00000) | 0.93 |
| Bag\_DT | 0.83242 | 0.93045 | 0.91250 | 0.06956 | (+/- 0.00001) | 0.93 |
| DT | 1.00000 | 0.90006 | 0.89263 | 0.09994 | (+/- 0.00001) | 0.90 |
| Boost\_DT | 0.79513 | 0.89575 | 0.88669 | 0.10425 | (+/- 0.00001) | 0.90 |
| LR | 0.83557 | 0.83163 | 0.87142 | 0.16837 | (+/- 0.00006) | 0.83 |
| Bag\_LR | 0.83242 | 0.82667 | 0.86640 | 0.17333 | (+/- 0.00006) | 0.83 |
| Boost\_LR | 0.79513 | 0.78574 | 0.80343 | 0.21426 | (+/- 0.00008) | 0.79 |

## Results obtained with oversampling and feature Selection:

The LGBM algorithm, which achieved the highest accuracy and F1-score, has been chosen to identify the directed buying visits. In this section, we apply feature selection to further improve the classification performance of LGBM classifier. 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.

|  |  |  |  |
| --- | --- | --- | --- |
| **Ranking** | **Correlation** | **Info Gain** | **mRMR** |
| 1 | Page Values | Page Values | Administrative Duration |
| 2 | Exit Rates | Exit Rates | Product Related |
| 3 | Product Related | Product Related Duration | Informational |
| 4 | Product Related Duration | Bounce Rates | Exit Rates |
| 5 | Bounce Rates | Product Related | Special Day |
| 6 | Administrative | Administrative Duration | Browser |
| 7 | Visitor Type | Month | Informational Duration |
| 8 | Informational | Administrative | Page Values |
| 9 | Administrative Duration | Traffic Type | Product Related Duration |
| 10 | Special Day | Visitor Type | Region |
| 11 | Month | Informational Duration | Weekend |
| 12 | Informational Duration | Special Day | Visitor Type |
| 13 | Weekend | Operating Systems | Month |
| 14 | Browser | Weekend | Bounce Rates |
| 15 | Operating Systems | Informational | Traffic Type |
| 16 | Region | Browser | Operating Systems |
| 17 | Traffic Type | Region | Administrative |

Table above shows the feature rankings obtained with the filters used in this study. The results showed that the "Page Value" feature of Google Analytics tracking tool is selected in the first place by Correlation and information gain filters, it a carries discriminative information about the intent of the visitor. Considering that the “Page Value” feature represents the page that a user visited before completing an e-commerce transaction, it can be seen as a natural measure of visitor’s transaction finalization intention. In our system, this feature is represented as the average of the "Page Value" values of pages visited by the visitor during the session and is updated when the visitor moves to another page. As seen in table above, the other two Google Analytics features, “Exit Rate” and “Bounce Rate”, are also highly correlated with the class variable and take place near the top in correlation and information gain filter rankings. However, since "Bounce Rate" is also highly correlated with "Exit Rate" and so contains redundant information, mRMR, which suggests incrementally selecting the maximally relevant variables while avoiding the redundant ones, chooses it in the 14th order. Similarly, although the "Product Related" and "Product Related Duration" attributes are closely related to the class variable, they have been ranked in 2nd and 9th orders by mRMR because of their high correlation with “Page Value” feature which has already been chosen by the algorithm. It is seen that correlation and information gain methods give similar rankings since both algorithms ignore the relations among the selected variables, whereas the mRMR method gives a quite different ranking compared to these methods.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm used**  **with Filter** | **Train\_accuracy** | **Test\_accuracy** | **Test\_Precision** | **Bias Error** | **Number of**  **Features** |
| LGBM ( Correlation  Feature Selection ) | 0.99986 | 0.93540 | 0.92681 | 0.06460 | **13** |
| LGBM ( Information  Gain Feature Selection ) | 0.99993 | 0.93396 | 0.92040 | 0.06604 | **9** |
| LGBM ( mRMR Feature  Selection ) | 0.99986 | 0.93284 | 0.92210 | 0.06716 | **13** |

The top 10 features selected by the filters are incrementally fed to the best algorithm using the oversampled dataset. The highest accuracy obtained by each method and the number of input variables in the corresponding best model are shown in the table above The highest accuracy (93.40 percentage) and F1-score (0.94) are obtained using the feature subset containing the top 9 features of the information gain ranking. It is also seen that the correlation and mRMR algorithms use more features than Information gain algorithm in their best models. Since information gain filter performs significantly higher accuracy with less number of features than correlation and mRMR, Light Gradient Boosting Method with top 9 features selected by information gain is determined as the final model considering its better performance as well as scalability of the real-time storage and update of the feature vector periodically during the session.

## Insights inferred through odds-Ratio using Logistic Regression:

Using the coefficients of the X variables in the new model, the odds ratio and probability have been calculated as follows:

Odds ratio = exp(coef(LR model)) Probability = Odds/(1+Odds)

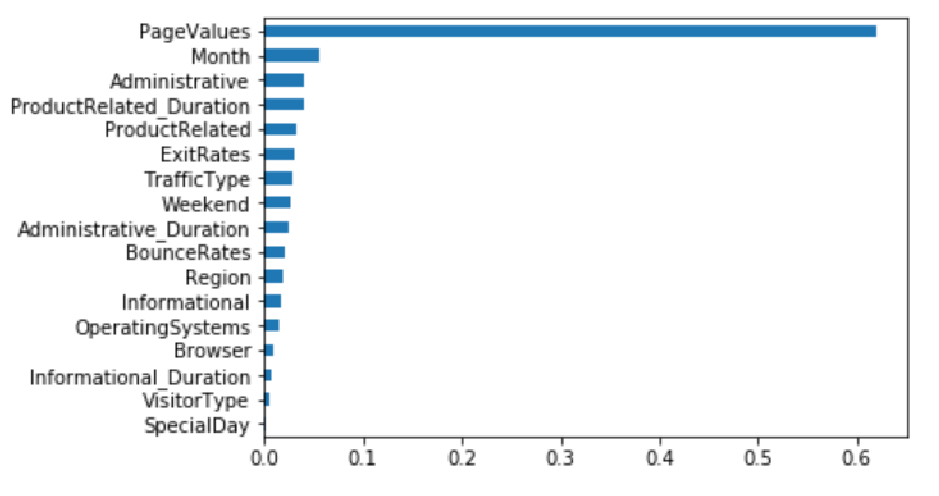
The Odds ratio and probability of some of the variables, based on their practical significance have been listed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor Variable** | **Co-eff** | **Odds Ratio** | **Probability** | **Correlation** |
| Administrative | 0.032539 | 1.033074 | 50.81% | 14% |
| Administrative\_Duration | -0.075243 | 0.927518 | 48.12% | 9% |
| Informational | 0.060954 | 1.06285 | 51.52% | 10% |
| Informational\_Duration | -0.01965 | 0.980542 | 49.51% | 7% |
| ProductRelated | 0.362581 | 1.437034 | 58.97% | 16% |
| ProductRelated\_Duration | 0.047996 | 1.049167 | 51.20% | 15% |
| BounceRates | -0.345979 | 0.707527 | 41.44% | -15% |
| ExitRates | -0.535751 | 0.585229 | 36.92% | -21% |
| PageValues | 2.521888 | 12.452082 | 92.57% | 49% |
| SpecialDay | -0.304253 | 0.737674 | 42.45% | -8% |
| Month | 0.116084 | 1.12309 | 52.90% | 8% |
| OperatingSystems | -0.110851 | 0.895072 | 47.23% | -1% |
| Browser | 0.000542 | 1.000542 | 50.01% | 2% |
| Region | -0.023389 | 0.976882 | 49.42% | -1% |
| TrafficType | 0.001356 | 1.001357 | 50.03% | -1% |
| VisitorType | 0.10415 | 0.880856 | 46.83% | -10% |
| Weekend | -0.126861 | 1.109766 | 52.60% | 3% |

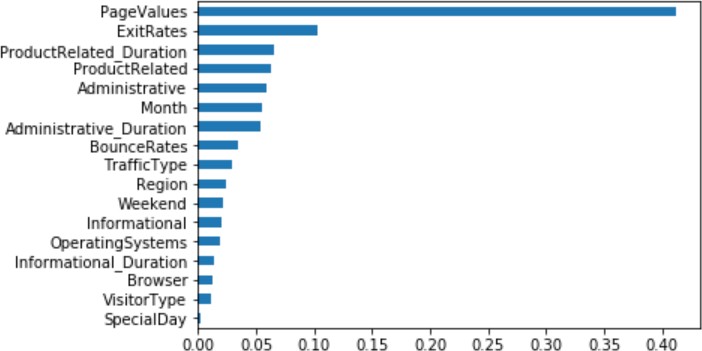
* + If the page value of E-commerce website is increased there is a 92% probability of Revenue generation to happen.
  + If current average bounce rate continues in future there is only there will be only 41 % chance for revenue generation.
  + By making better informational and product pages there will be 50-60% probability revenue will be generated.
  + Unless new customer had been brought into website there will be 47% of Probability of Revenue generation to happen.

## Variable Importance Plot for Tree Based Algorithms

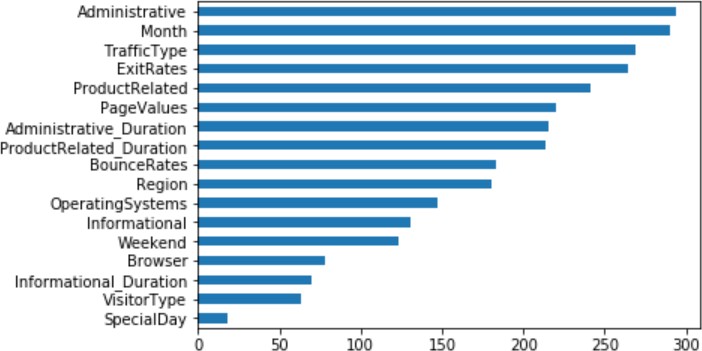
Plot for Decision Tree Model:



Plot for Random Forest Model:



Plot for Light Gradient Boost Model:



# Chapter 4 - Conclusions



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 a Light Gradient Boosting algorithm calculated using resilient backpropagation with weight backtracking. Our findings support the argument that the features extracted from clickstream data during the visit convey important information for online purchasing intention prediction. The features that represent aggregated statistics of the clickstream data obtained during the visit are ranked near the top by the filter feature ranking algorithms.

However, these metrics are also highly correlated with each other. On the other hand, although the session information-based features are less correlated with purchasing intention of the visitor, they contain unique information different from click stream based features. Therefore, we apply a feature ranking method called information gain Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output. 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.



# 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 a 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



* S. Olcay, PolatMete Katircioglu ,Yomi Kastro. Real-time prediction of online shoppers’

purchasing intention using multilayer perceptron and LSTM recurrent neural networks

* 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.
* Arun Thamizhvanan, M.J. Xavier. Determinants of customers' online purchase intention: an empirical study in India.

|  |
| --- |
| Chapter 18 - Appendix |
| Detailed data dictionary |

|  |  |
| --- | --- |
| **Feature Name** | **Feature Description** |
| Custom Variable | Visitor information |
| Page | Relative link of pages visited by the visitor. |
| Page Views | Page view value of the pages visited by the visitor |
| Unique Page Views | Unique page view value of the pages visited by the visitor |
| Average Time on Page | Average time of the pages visited by the visitor |
| Entrances | Value of entrances of pages visited by the visitor |
| Bounce Rate | Average bounce rate value of the pages visited by the visitor |
| Exit Rate | Average exit rate value of the pages visited by the visitor |
| Page Value | Average page value of the pages visited by the visitor |
| Date | Date of visit |
| Search Term | Search keyword which used by the visitor |
| Exit Screen | Last page of the of the visitor before session ends |
| City | City of the visitor |
| Browser Type | Browser type of the visitor |
| Operating System | Operating system of the visitor |
| Traffic Type | Traffic type by which the visitor has arrived at the website (e.g. banner, SMS, direct) |
| Visitor Type | Visitor type as “New Visitor”, “Returning Visitor” and “Other” |
| Source | Source by which the visitor has arrived at the website (e.g.Email, Google, Facebook) |
| Sessions | Session value of the pages visited by the visitor. |
| New Sessions Rate | Average session rate value of pages visited by the visitor |
| New Users | Numeric value of Visitor Type |
| Session Average Page | Average session time of the pages visited by the visitor |
| Session Duration | Session time of the pages visited by the visitor |
| Transactions | Number of transactions made by the visitor |
| Revenue | Total amount purchased by the visitor |
| Ecommerce Conversion Rate | Percentage of visits resulted in a transaction by visitor |