**INTERIM REPORT**

MACHINE LEARNING MODEL FOR ONLINE SHOPPERS PURCHASING INTENTIONS

***(Dataset: Online shoppers purchasing intentions Data Set UCI Repository)***

Problem Statement: Developing a Machine Learning model aimed at

Predicting the online shoppers purchase intentions

**Mentor Submitted by**

Mr. Naveen Mooventh

Gowtham

Yaswanth Praveen

Samuel Mathews

**Preamble**

****

**What About Website Analytics?**

Monitoring the traffic of your eCommerce website is something that every successful website does so that they can see how well their site is performing. This means that eCommerce business owners should be noting how many visitors their website is getting, how long they are staying on the site, and where exactly they are losing business.

**Data For eCommerce**

In the last two years, this is a term that has become synonymous with eCommerce analytics. It entails sifting through massive amounts of data (Think hundreds of terabytes) in order to reach a scientific, statistic-based solution to a problem. It’s a burgeoning field, with a continuously growing scope. Big data can be invaluable for eCommerce business owners, for a variety of reasons:

##### **1. Behavioral Targeting:**

This remains the single most important factor upon which the success or failure of your eCommerce business rests. Big Data helps you understand your customers better – The products they like, why they leave their carts without paying, the times of the day and the months of the year that see the most customers, and a variety of other salient business decisions. You can plan your marketing campaigns around this – A practice known as behavioral targeting, or behavioral marketing – to maximize conversions. Examples include sending personalized emails with discount coupons to customers on their birthdays, targeted pop-ups informing customers of seasonal offers and so on.

**2. eCommerce Personalization:**

Customers like to see discounts, recommended products, and most importantly, they love a personal touch. This is impossible without data: What they tend to buy, how much time they spend browsing through products, how much they spend per transaction, how often they visit eCommerce sites, how often they abandon their cart and a myriad other bit of information.

##### **3. Customer Satisfaction:**

It’s been found that close to 68% of visitors tend to leave websites because they aren’t happy with the customer service being offered. Using Big Data, you can coordinate and synchronize multiple communication channels –Emails, phone calls and live chat, if applicable. Big Data is crucial to understanding recurring issues and fixing them immediately.

##### **4. Dynamic Pricing:**

With enough data, customers can be gauged for their spending patterns and on-site behavior. Once personas are attributed to customers, prices can be varied to entice them back to the site, were they to leave. Dynamic discounts are a powerful toolbutare only successful when executed with enough data. This has been one of the most effective customer retention strategies to date.

##### **5. End-To-End Analytics**

You can use Big Data to understand not just your customers, but how your entire business is doing: Stocking, shipping, inventory and sales. This gives you the sort of bird’s eye view that you can use to forecast future cash flows, Gross Merchandise Volumes, customer acquisition rates and other parameters crucial to the success of eCommerce businesses.

**Data Monography:**

The dataset consists of feature vectors belonging to 12,330 sessions.  
The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 12330 | **Area:** | Business |
| **Attribute Characteristics:** | Integer, Real | **Number of Attributes:** | 18 | **Date Donated** | 2018-08-31 |
| **Associated Tasks:** | Classification, Clustering | **Missing Values?** | N/A | **Number of Web Hits:** | 78453 |

Dataset link:

<https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset>

**Attributes’ Information:**

#### **Dataset 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 in that session and total time spent in each of these page categories.The Bounce Rate, Exit Rate and Page Value features represent the metrics measured by "Google Analytics" for each page in the e-commerce site.

#### **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.

#### **Exit Rate-**

Features for a specific web page are calculated as for all pageviews to the page, the percentage that were the last in the session.

#### **Page Value-**

Feature represents the average value for a web page that a user visited before completing an e-commerce transaction.

#### **Special Day-**

Feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day).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 Valentine'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.

The dataset also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.

**METHODOLOGY:**

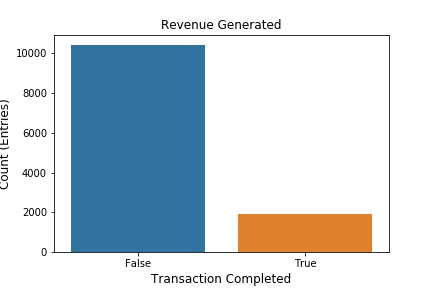
I Exploratory Data Analysis

II Modelling Technique

**I** -**Exploratory Data Analysis:-**

* To understand the data in terms of Visitor session information and visitor pageview information across various independent variables
* Get insights on various features.
* The data set was loaded in a pandas DataFrame object and a thorough investigation was done to unearth vital information about the data.

**Revenue Generated:**

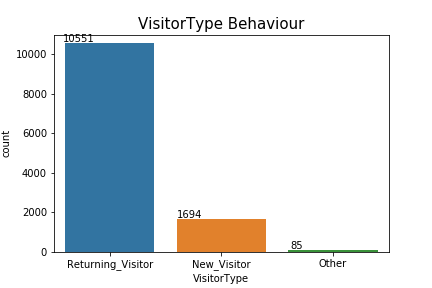
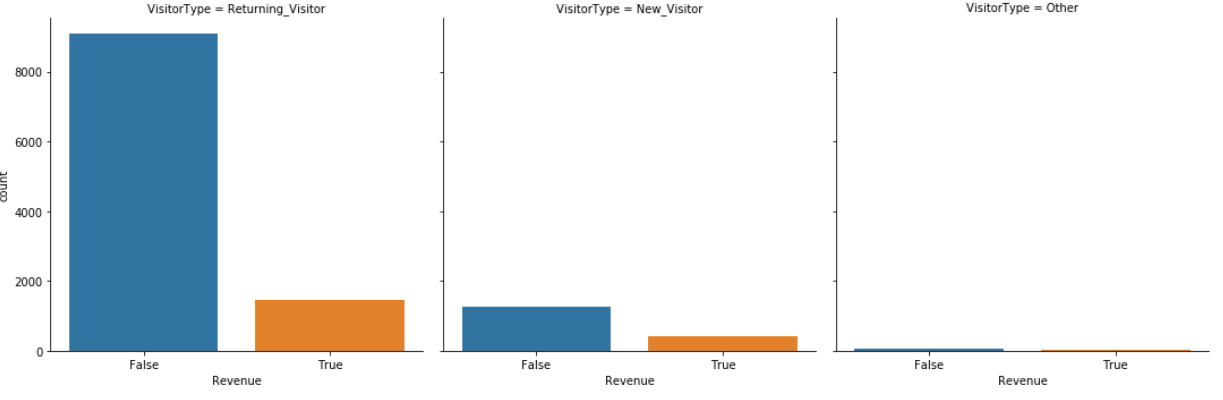


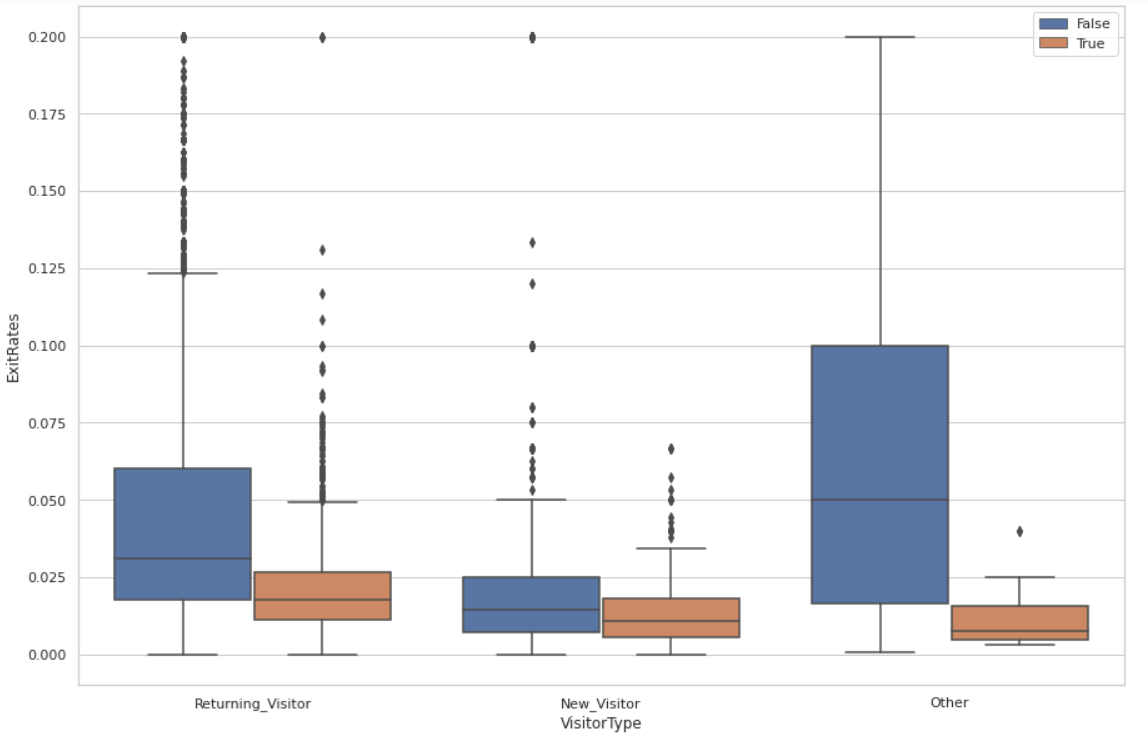
**Inference**-

False 10422 84%

True 1908 15%

* We need to focus on the 15% to generate more revenue.
* Based on the distribution, it's an imbalance dataset.

**Visitors Behavior :**

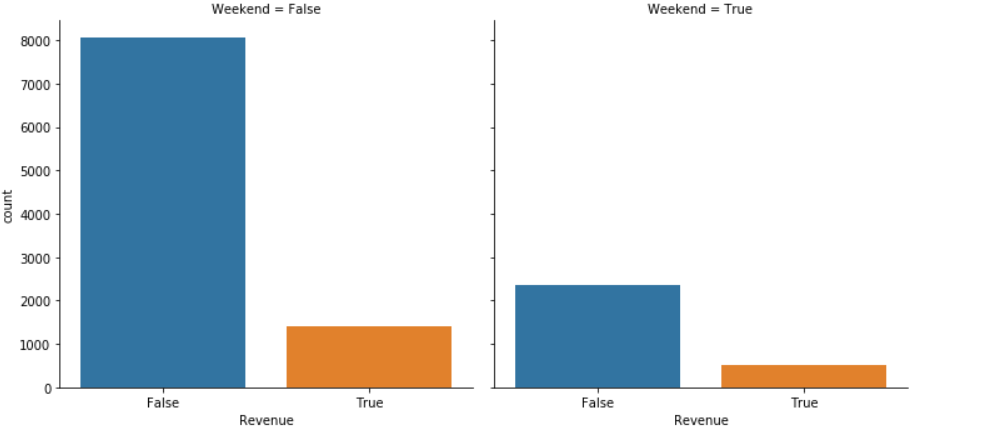
****

|  |  |  |
| --- | --- | --- |
| Revenue | True | False |
| Returning visitor | 86.07% | 13.93% |
| New visitor | 75.09% | 24.91% |

**Inference**-

* If you see that 85% percentage of returning visitors on this website. so this the reason why revenue is not generated. We need to focus on the new visitor also with ads/digital marketing
* You can observe that the conversion rate is also high for the new-Visitor, so there is a high possibility.
* Average exit rate for the new-visitor is comparatively less than the returning customer(it may be due to the distribution,let’s see)

**Weekend Revenue:**

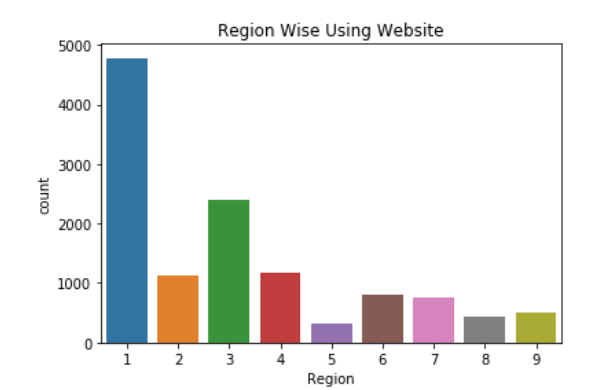


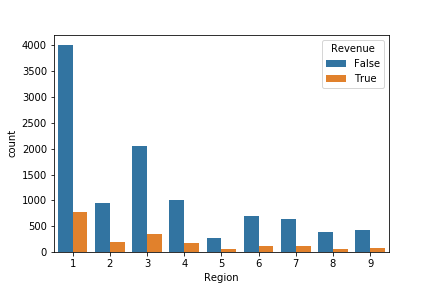
**Inference**-

* There are a lot of visitor sessions 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 the weekend is slightly greater than weekday.
* 76% of customers are visiting the website on weekdays.

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

**Region Wise:**



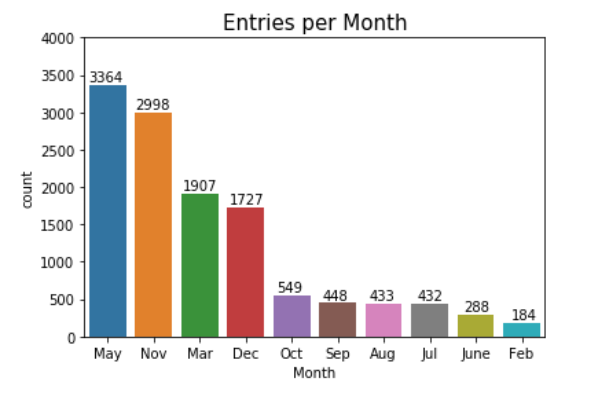
****

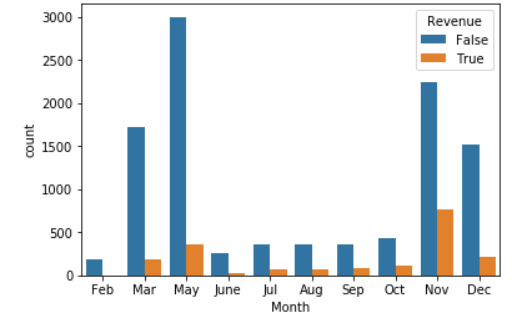
**Inference**-

* The customers visiting the online website are most of them region 1 and 3- contributing around more than 50% of the overall region.
* Region 4 has more conversion rate so let's take them into consideration.

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

**Month-Wise**:





**Inference**-

* Jan & april month are missing.
* 80% of customers are visiting the website in the month of may,nov,mar,dec.They should mainly concentrate on these months.
* Website visitors may be high in May,but actual Sales conversion took place in the month of November. We need to find a successful conversion rate Month wise for further understand .

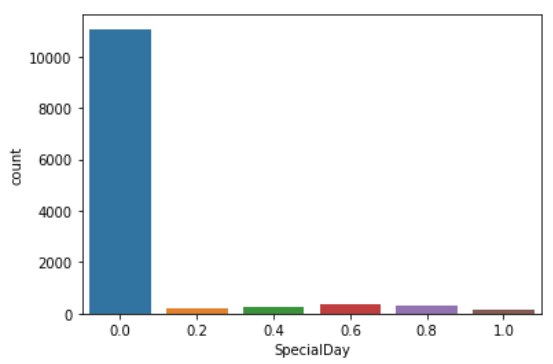
**Month wise visitor session(Revenue Generated )**

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

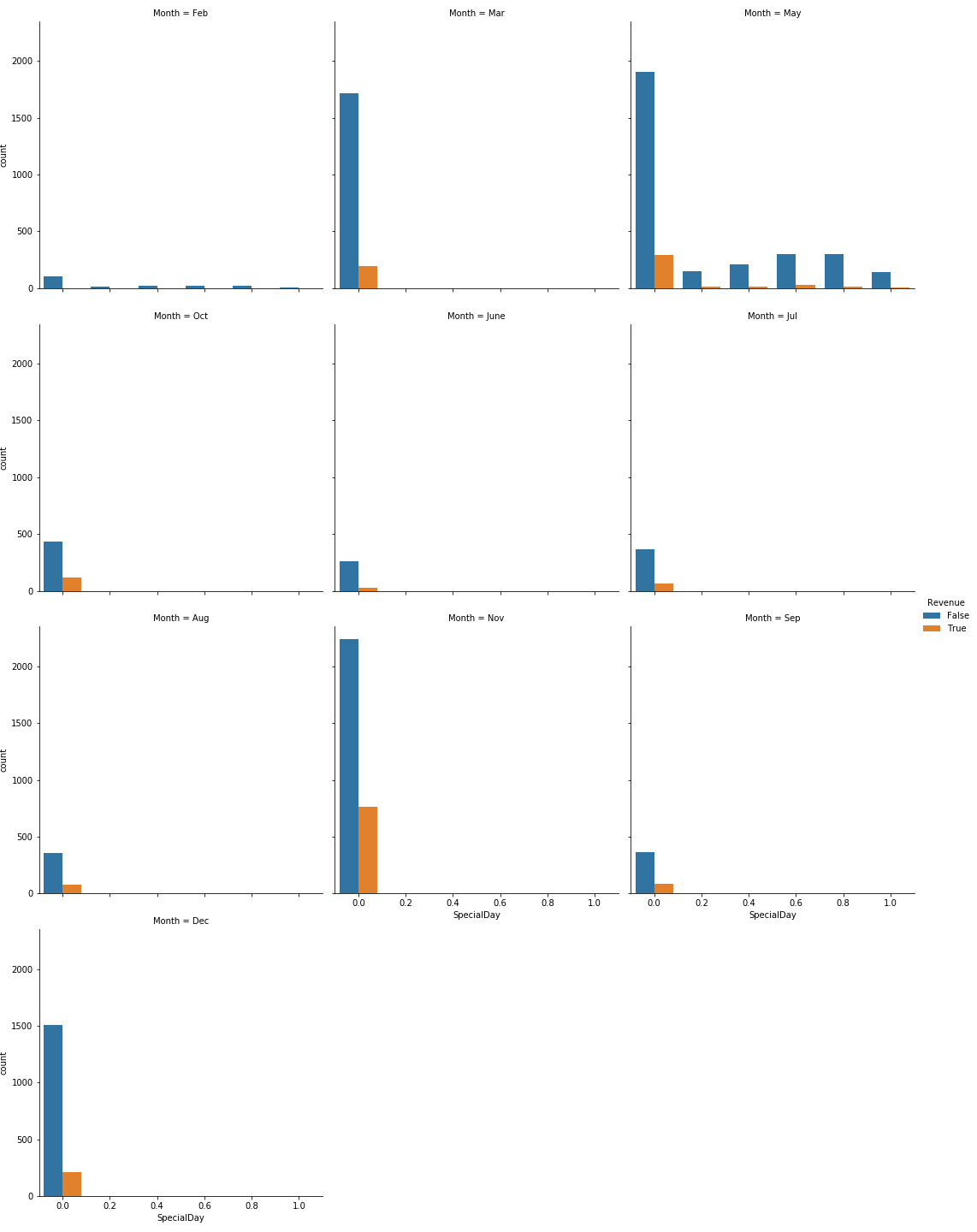
**Month wise percentage of visitor**

|  |  |
| --- | --- |
| **Month** | % of visitor |
| **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% |

**Special Day:**



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



**Inference-**

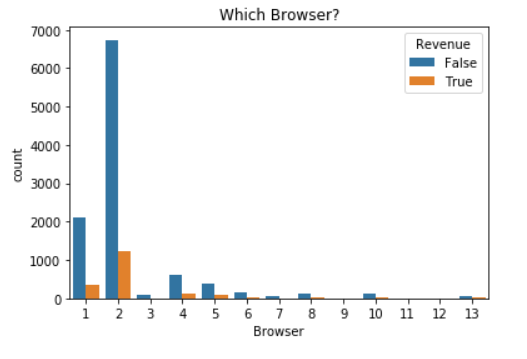
-90% online session happen on non-special days they buy 10 days before the function

- We need to look on November,December special day

- See why the people are missing so many functions in July month.

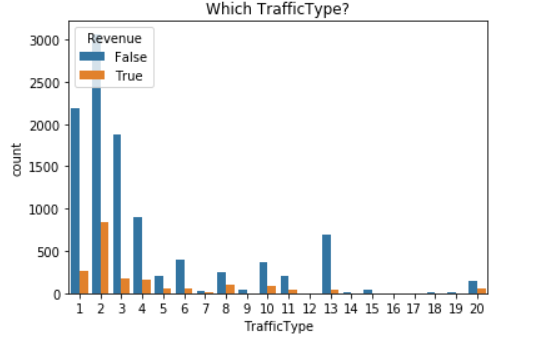
**\**

**Browser & Traffic type & OS type:**

****

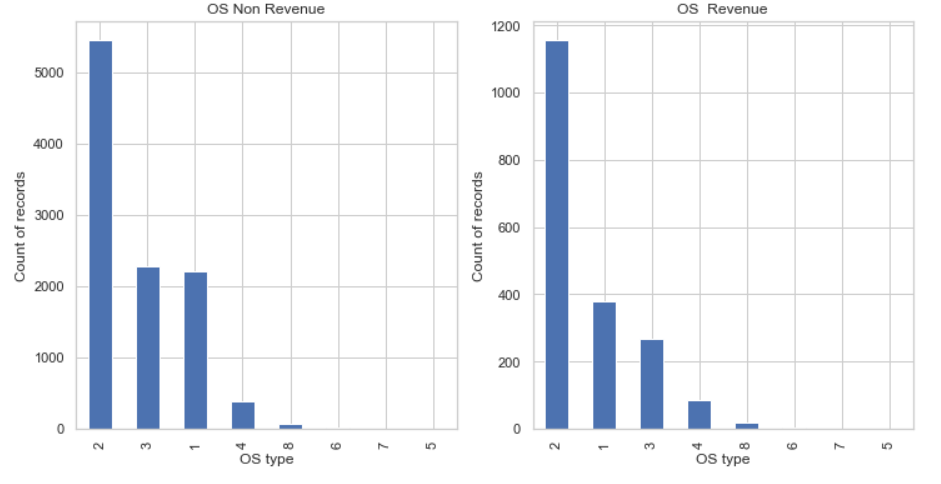
**Inference :**

Browser choice is even more polarizing than the Operating System. Here we see that a large majority of users use browser 2, with a smaller number of users using browser 1. All other browsers represent a small subsection of online users. We will not use this as it does not contribute much to our model.



**Inference:**

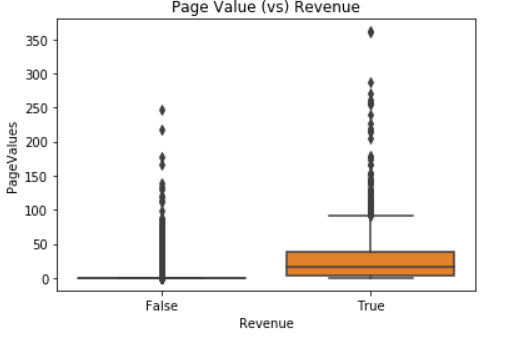
Traffic Type aids website owners in gauging traffic sources and can assist with determining where they should invest in advertisement. General certain ads and relate to the source page.



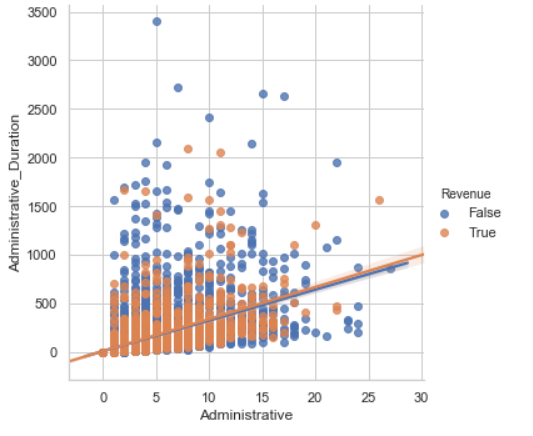
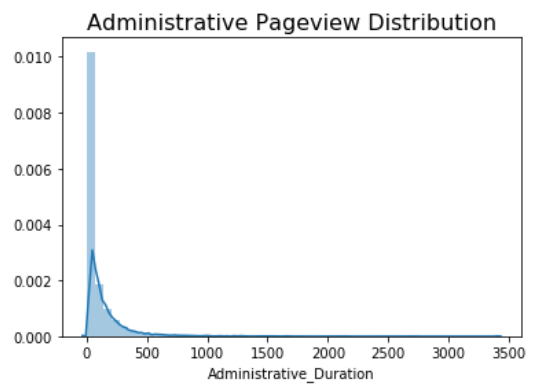
**Inference:**

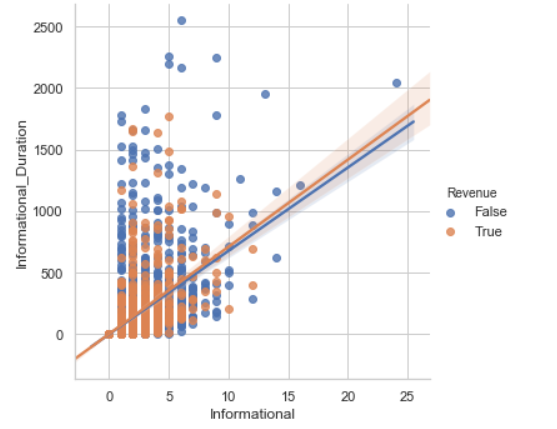
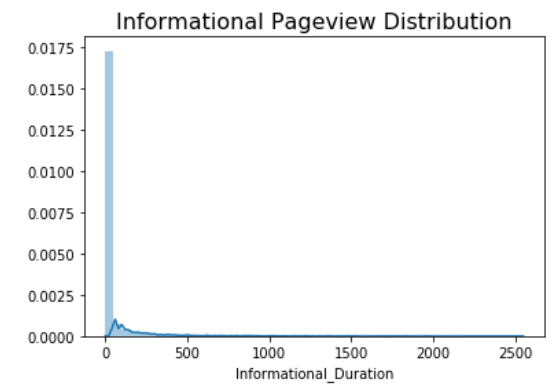
* Even though OS 2 concentrates almost the same among the revenue generated.
* OS 1 creates 5% more revenue generation compared to the OS type 3.

**Page View:**

****

**Inference-** When the page value increases the revenue is generated.Let’s explore



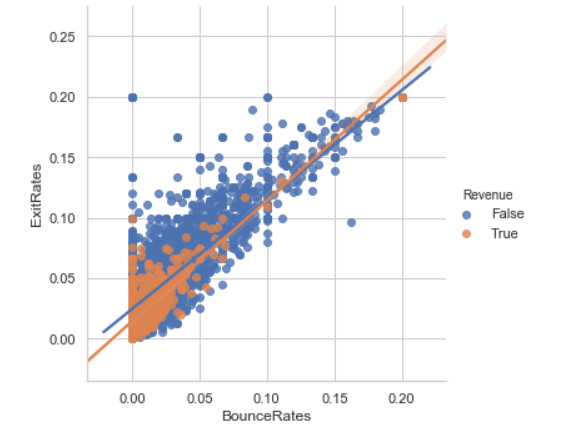


**Inference**:

* Administrative,Informational and their duration have been left skewed.
* Spend less time in the particular region , we need to find the reason to improvise.
* Administrative, check if there are any login problems
* Informational, even if they spend so much time they end up without buying.
* customers who spent a longer administrative duration in a website are very less likely to bounce from the website that is navigating away from the website just after navigating one page of that website.

**Bounce Rate & Exit Rates:**

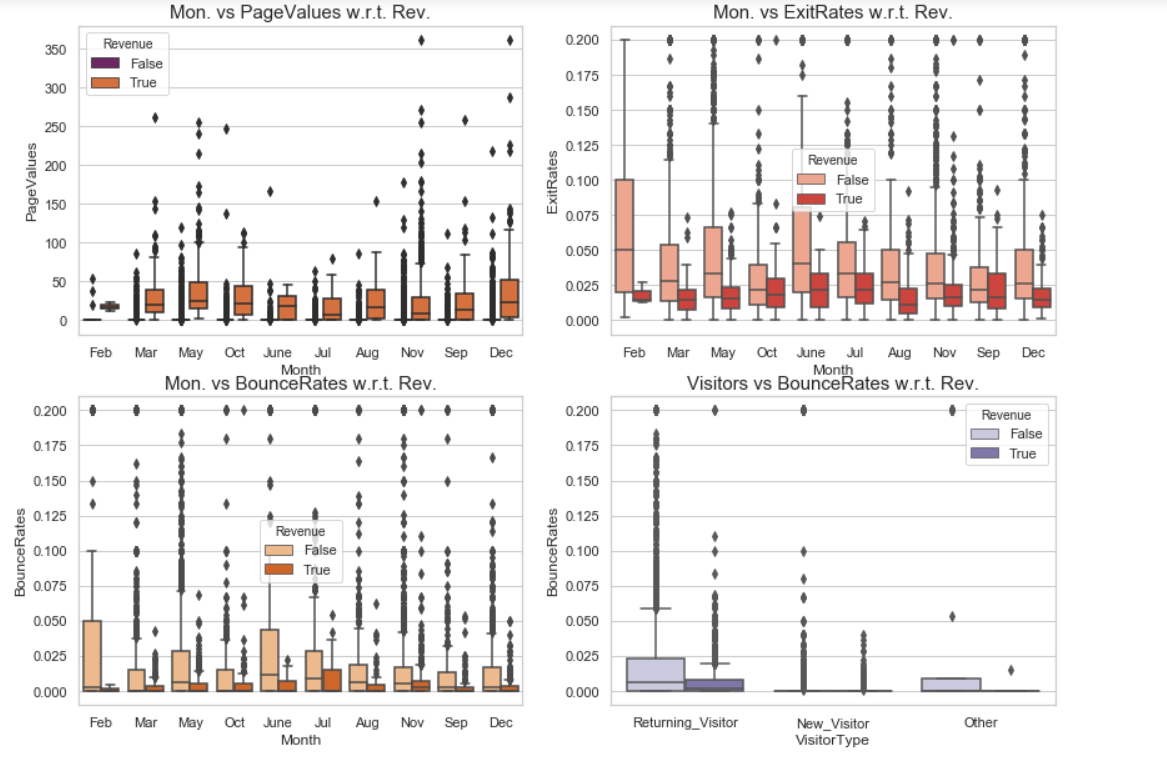
Bounce rate is the overall percentage of a single engagement session whereas exit rate is the percentage of exits from a page.



**Inference:**

* A high bounce rate could indicate issues with user satisfaction owing to one or many reasons such as unfriendly user interface of the website, extremely slow throughput or other technical matters.
* A high exit rate could be a sign of lower performing sectors in funnels, showing areas open to optimization as if customers are leaving then at the end of the day no one is buying.
* Here both of them exhibiting a positive relation- we need to try to control both to gain new customers and old customers to sustain.

**MonthWise Analysis:**

****

**Inference:**

* Bounce rates among the New\_Visitor is acceptable but why Returning\_visitor are bouncing. Need to optimise the reason.
* Page Value has a direct relation among the revenue generation and its wide spread across May and Dec- let’s find that and expand more or improvise February by providing ads.
* **Feb-** Bounce rate and Exit rate range is comparatively high, attract them.

**Base Model:**

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.

**Target feature --- > Revenue**

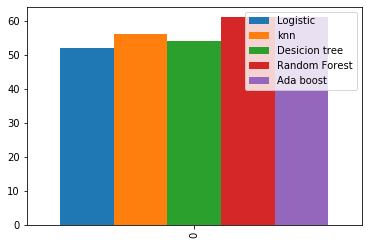
It’s a classification model so we try to draw some conclusions from the input values given for the training(independent features) . Here we have a binary classification with two possible outcomes(Revenue Generated- Yes/No)

Performance Metrics:

It's an imbalance classification problem, the performance metrics planned to estimate the performance are f1\_score,Precision,Recall,Roc\_curve, Bias and Variance error..

The dataset is fed to Logistic Regression, Decision tree, Random Forest and Gradient Boosting,KNN(K-Nearest neighbors).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy(%) | | Precision(%) | Recall(%) | f1score(%) | Bias\_Error | Variance\_Error |
|  | Train | Test |  |  |  |  |  |
| Logistic | 88 | 88 | 78 | 39 | 52 | 0.491418 | 0.000148 |
| KNN | 91 | 88 | 66 | 49 | 56 | 0.413108 | 0.000131 |
| Decision Tree | 100 | 85 | 55 | 54 | 55 | 0.452311 | 0.000175 |
| Random Forest | 100 | 89 | 76 | 51 | 61 | 0.373984 | 0.000000 |
| Ada Boost Classifier | 89 | 89 | 68 | 55 | 61 | 0.401740 | 0.000031 |

****

**Feature Selection & Modelling:**

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 our dataset we are having 12330 rows and 18 columns. If we build the model without feature selection the model will produce low bias and high variance. We have done feature selection for 18 columns based on statistical methods they are

\*Chi square test

\*Anova

**Chi Square Test**

Chi Square test is works when the independent and dependent variable is categorical. It is a non-parametric test deals with degree of freedom and alpha. Frequency should be greater than 5. It is a right skewed method. In this dataset we are having 8 columns as categorical and we are going to check each columns whether they are dependent on target column or not.

**Features which are dependent on target column(Revenue) based on p value from chi square test**

Weekend, Special day, Visitor type, Browser, Month, Traffic type

**Features which are independent on target column(Revenue) based on p value from chi square test.**

Region

**Anova Test**

Anova test is works when the independent variable is numerical and dependent variable is categorical. It is a two tailed test. In this dataset we are having 9 columns as numeric and we are going to check numerical columns whether they are dependent on target column or not.

**Features which are dependent on target column(Revenue) based on p value from Anova test**

Product related,Product related duration,Page values,Exit rate, adminstrative, adminstrative duration,Bounce rate,informational,informational duration.

**Wrapper Method**

A wrapper method needs one machine learning algorithm and uses its performance as evaluation criteria.Feed the features to the selected Machine Learning algorithm and based on the model performance you add/remove the features.It is an iterative and computationally expensive process but it is more accurate than the filter method.The wrapper methods are

\*Forward selection

\*Backward selection

\*Recursive feature selection

**Backward selection**

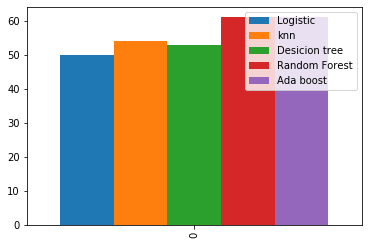
In backward elimination, we start with all the features and removes the least significant feature at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.But in this project we have mainly used backward selection method.

**Features which are dependent on target column(Revenue) based on p value from Backward elimination test**

ProductRelated, ProductRelated\_Duration,BounceRates, ExitRates, PageValues,Month\_Dec, 'Month\_Feb', 'Month\_Mar', 'Month\_May', 'Month\_Nov', 'OperatingSystems\_2', 'Browser\_12', 'TrafficType\_10', 'TrafficType\_11', 'TrafficType\_13', 'TrafficType\_2', 'TrafficType\_20', 'TrafficType\_5', 'TrafficType\_8', 'VisitorType\_Other', 'VisitorType\_Returning\_Visitor'.

**Results obtained with Feature selection:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Train Accuracy | Test Accuracy | Recall | F1 Score | Bias Error | Variance Error |
| Logistic Regression | 88 | 88 | 38 | 50 | 0.48 | 0.00032 |
| Random Forest | 99 | 90 | 52 | 61 | 0.35 | 0.00003 |
| Ada Boost | 99 | 85 | 56 | 61 | 0.38 | 0.00017 |
| Decision Tree | 89 | 89 | 54 | 53 | 0.44 | 0.00009 |
| Knn | 91 | 88 | 49 | 54 | 0.39 | 0.00003 |

****

**Inference**

Both after and before feature selection the is not much vairation in the performance in terms of accuracy,recall and f1 score.

**Modeling**

**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

**Results on class imbalanced dataset:**

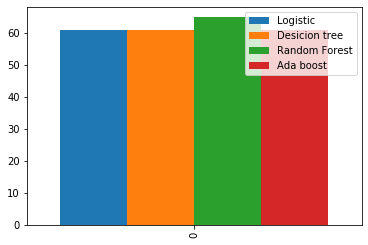
Tables below show the results obtained with Logistic Regression, Decision Tree, Random forest, Light gradient boosting, bagging classifier (Logistic regression & Decision Tree) and boosting classifier ( Logistic regression & Decision Tree) respectively. The results show that Ada Boosting ( AdaBoost ) 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. To test on this imbalance dataset we have done a technique on this project called smote.

**Smote**

There are a number of methods available to oversample a dataset used in a typical classification problem (using a classification algorithm to classify a set of images, given a labelled training set of images). The most common technique is known as SMOTE: Synthetic Minority Over-sampling Technique.

**Results obtained with Smote:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Train Accuracy | Test Accuracy | Recall | F1 Score | Bias\_Error | Variance  Error |
| Logistic Regression | 82 | 86 | 75 | 61 | 0.188918 | 0.00017 |
| Decision Tree | 100 | 85 | 65 | 56 | 0.089644 | 0.00006 |
| Random Forest | 100 | 89 | 68 | 65 | 0.061715 | 0.00000 |
| Ada Boost | 92 | 87 | 66 | 60 | 0.079090 | 0.000002 |



**Hyper Parameter tuning and feature Selection:**

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

**Hyper Parameter tuning**

**Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm.**The methods that are used most commonly in hyper parameter tuning are

\*Grid Search

\*Random Search

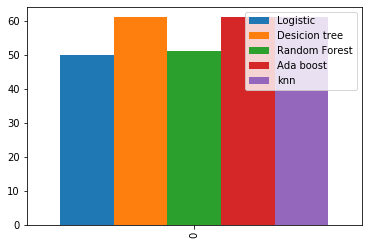
**Grid Search**

Grid search is a traditional way to perform hyperparameter optimization. It works by searching exhaustively through a specified subset of hyperparameters.Using sklearn’s GridSearchCV, we first define our grid of parameters to search over and then run the grid search.

The benefit of grid search is that it is guaranteed to find the optimal combination of parameters supplied. The drawback is that can be a time consuming process.

**Results obtained with Hyper parameter tuning(Grid Search) and feature Selection:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Train Accuracy | Test Accuracy | Recall | F1 Score |
| **Logistic Regression** | **88** | **88** | **37** | **50** |
| **Random Forest** | **89** | **89** | **37** | **51** |
| **Decision Tree** | **90** | **89** | **52** | **61** |
| **Ada Boost** | **91** | **89** | **52** | **61** |
| **Knn** | **90** | **87** | **43** | **51** |

****

**Inference**

As you see that in above result the decision tree and ada boost performs well in grid search

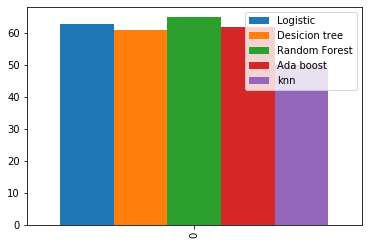
based on f1 score.The decision tree and ada boost has better f1 score when compared with logistic and random forest tree.

**Randomized Search**

Random search differs from grid search mainly in that it searches the specified subset of hyperparameters randomly instead of exhaustively. The major benefit being decreased processing time.Let’s give random search a try with sklearn’s RandomizedSearchCV. Very similar to grid search above, we define the hyperparameters to search over before running the search.It is very fast when compared with grid search.While this is an important step in modeling, it is by no means the only way to improve performance.

**Results obtained with Hyper parameter tuning and feature Selection:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Algorithm | Train Accuracy | Test Accuracy | Recall | F1 Score | Bias  Error | Variance  Error |
| **Logistic Regression** | **89** | **90** | **57** | **63** | **0.490607** | **0.000141** |
| **Random Forest** | **89** | **90** | **59** | **65** | **0.358793** | **0.000174** |
| **Decision Tree** | **90** | **89** | **53** | **61** | **0.356150** | **0.000193** |
| **Ada Boost** | **90** | **88** | **55** | **62** | **0.345290** | **0.00020** |
| **Knn** | **90** | **87** | **40** | **50** | **0.488720** | **0.000141** |

****

**Inference**

As you see that in above result the random forest tree and Ada Boost classifier performs well in randomized search based on f1 score.The random forest and logistic has better f1 score when compared with KNN and decision tree.Random forest is best when we consider f1 score.

**Conclusion and Recommendation**

* Problem statement of our project is to construct a real time user behaviour analysis of online shoppers to predict their intention and revenue generation.
* As per our observed data,we have an imbalance target where we have lesser revenue generated (only 16%).
* We have analysed and visualized various features to derive an inference among the importance of each. Aggregated the page value among the administration,information and product related and other chosen features as an input to the machine learning algorithms.
* Even though the model performed at its minimal 50% in correctly identifying the target customers. Proper feature selection and hyper parameter tuning have improved the performance rate to 65% f1\_score metric ( avoiding the false chosen).
* Findings show that choosing a proper combination of features with data aggregated statistics and session information including the location would increase the performance in classifying the customers based on their intentions.

**Reccomendation:**

**Recommendations can be more relevant with the comparison of the distribution and the conversion rate among the different features.**

- Returning customers are more but their conversion rate is lesser.There are few suggestions to improve their intention towards purchase such as:

* Providing coupons by identifying the loyal and returning customers with more attraction.
* Identify the page where the exit rate is more and try to build the particular page more informative and value-added.
* Maintaining the database of the returning customers with regular updates on the special days as a promotion of the products to generate more revenue.
* Providing small add-on products to the customer as a marketing strategy with minimal/discounted prices.

- Other categories the new customers even though they are less in count their conversion rate is more. It's a good start, now it's time to take certain action to retain them with possible actions such as

* First order with discounts and referral codes to generate more customers into the website.
* 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.
* New customers need to make sure that our website is user friendly with lesser ads and pop-ups.
* Administrate page and information page should be quick to help and connect for help to make sure that they don't get irritated.
* Returns and refunds with a certain time limit can be afforded with a valid reasons to provide trust to the shopping and company.

- Bounce rate/Exit rate is another major thing to focus as both are very high. Websites should be made more user friendly with attractive images and discounts to attract. Exit rate on the other hand can be controlled with the help of contacting them regarding the issues that might help them via email or message (making sure it's not more annoying).

- Special days are high in certain months yet many prefer not to shop within a few days of celebration due to the delivery dates. We can provide fast delivery on special days attracting more customers.

- Regions with high conversion rates can be updated on regular basis with the update on offers and marketing of the new products.

**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.