ONLINE SHOPPERS INTENTION

UMA MAHESWARI

07.01.2022

ABSTRACT

The purpose of this study is to examine the influence of online shopping experience on perception of specific types of risks associated with online shopping and how each type of risk perceptions influences online purchase intentions. A conceptual model was proposed to illustrate the relationships between online shopping experience and perceptions of product, financial, and privacy risks associated with online shopping, and how both experience and risk perceptions impact online purchase intentions. The results indicate that online shopping experience is a strong positive predictor of online shoppers' purchase intentions for the two product categories (i.e. non-digital and digital products) examined. Online shopping experience is negatively related to perceptions of product and financial risks associated with online shopping regardless of product category; but only reduce privacy risk associated with shopping non-digital products online. Interestingly, although both product and financial risks are negatively related to online purchase intentions for non-digital and marginally for digital product, privacy risk perception is not related to online shopping intentions for either of the product categories. Findings are discussed with theoretical and managerial implications.

Keywords: Online Purchase Intentions; Perceived Risk; Experience; Product Category

1.Problem statement

People often spend a lot of time browsing through online shopping websites, but the conversion rate into purchases is low. Determine the likelihood of purchase based on the given features in the dataset. The dataset consists of feature vectors belonging to 12,330 online sessions. The purpose of this project is to identify user behaviour patterns to effectively understand features that influence the sales.

Objective

The purpose of this project is building machine learning models to identify user behaviour patterns and determine the likelihood of purchase based on the given features for online purchase consumer data.

Dataset

online_shoppers_intention.csv

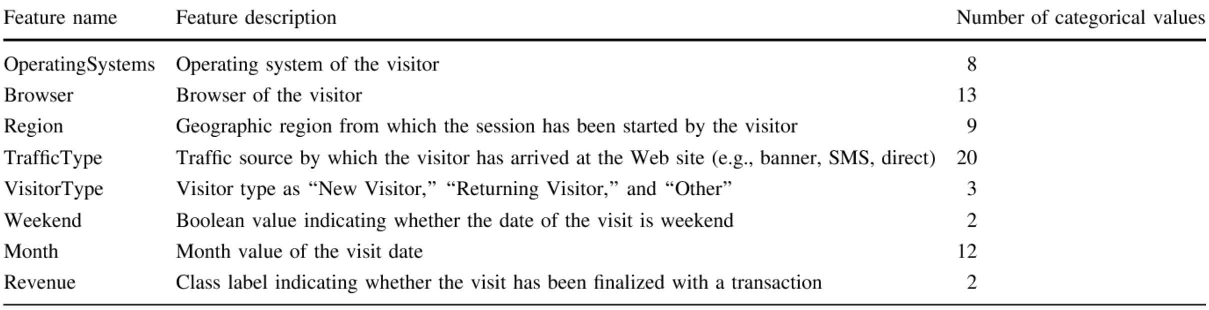
- The dataset consists of feature vectors belonging to 12,330 online sessions.
- We use 'Revenue' as our target variable and it's in a Boolean Format.
- Out of which
    - Numerical Features – 14
    - Categorical Features - 2
    - Boolean Features - 2

Feature Analysis

Numerical Features

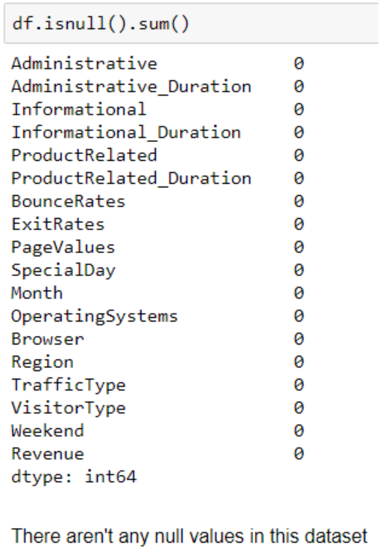
Feature name	Feature description	Min. value	Max. value	SD
Administrative	Number of pages visited by the visitor about account management	0	27	3.32
Administrative duration	Total amount of time (in seconds) spent by the visitor on account management related pages	0	3398	176.70
Informational	Number of pages visited by the visitor about Web site, 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	63,973	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	0.19

Categorical Features



Feature name	Feature description	Number of categorical values
OperatingSystems	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
TrafficType	Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)	20
VisitorType	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

Data Clean up



```
df.isnull().sum()
```

```
Administrative             0
Administrative_Duration    0
Informational              0
Informational_Duration     0
ProductRelated             0
ProductRelated_Duration    0
BounceRates                0
ExitRates                  0
PageValues                 0
SpecialDay                 0
Month                      0
OperatingSystems           0
Browser                    0
Region                     0
TrafficType                0
VisitorType                0
Weekend                    0
Revenue                    0
dtype: int64
```

There aren't any null values in this dataset

2. Market/Customer/Business Need Assessment

We do not need to allocate resources to improve user experience for people with high bounce and exit rates. Only few people in the lower spectrum end up making a purchase. Our target customers are the ones who visit the product from 10 to 400 times. The outliers generally do not end up purchasing. The covid-19 pandemic and the lockdown has negatively affected small businesses like shops and vegetable vendors. They have been forced to shut down their shops and stalls early, which has resulted in significant and rapid decrease in sales. Also, the customer buying preferences have been significantly changed due to the pandemic. Therefore, by using this technique, we aim to provide small businesses with useful insights from the available data and ways to generate more revenue.

3. Target Specifications and Characterization

The proposed system/service will provide the shopkeepers and vendors with some techniques so that their sales boost up and they no longer have to go through an economic crisis. It will suggest them to group certain items together, based on the analysis performed by the algorithm, so that the customer buys these grouped items together. Also, applying certain discount strategies on such grouped items will also increase the sales as required. As far as the local vegetable vendors are concerned, the analysis aims at suggesting to them the frequently bought veggies in a certain area, which they can eventually consider for cultivation and hence, increase their sales as well.

4. External Search (information sources/references)

The sources I have used as reference for analyzing the need of such a system for local businesses and how E-commerce giants have been using the technique to boost up online sales, They are

- Understanding Customer Behaviour
- How E-commerce sites benefit from online shoppers Analysis
- Increasing purchase and Improving ROI
- A study on Understanding Changing Trends of Customer Behaviour

5. Bench marking alternate products

E-commerce giants like Amazon,Flipkart have been using affinity analysis to perform Online Shoppers Analysis, which identifies purchasing habits of customers and uses this information to cross-sell and up-sell relevant items. But this technique would also be beneficial when applied to the small businesses since most of the daily needs and other essentials are still being bought from these shopkeepers and vendors.

6. Application Regulations (government and environmental)

- Data protection and privacy regulations (Customers)
- Govt Regulations for small businesses
- Employment Laws
- Antitrust Regulations
- Regulations against false advertising

7. Applicable Constraints (need for space, budget, expertise)

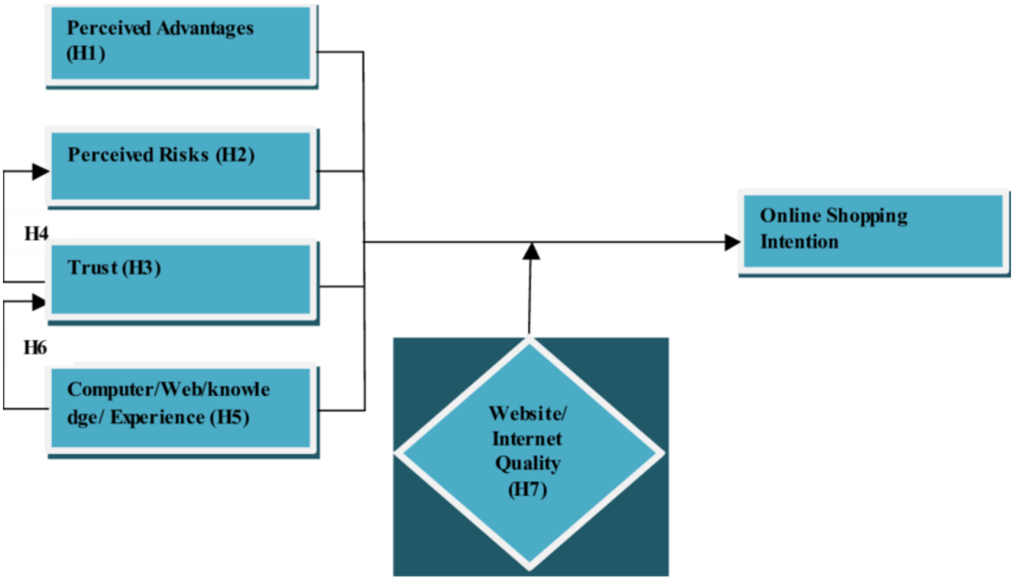
- Data Collection from shopkeepers and vendors
- Continuous data collection and maintenance
- Lack of technical knowledge for the user(vendors)
- Taking care of rarely bought products
- Convincing the shopkeepers to implement the system in their shops.

8.Business Opportunity

Since the above technique has only been used by large companies, this can be extended for small businesses, not only shopkeepers or vendors, but also food businesses and takeaways. Therefore, there is a fair chance of this service being a great business opportunity. Every small business that depends on sales can and would want to opt for using this service in order to always know what their customers want. The emergence of every small business is thus a fairly great business opportunity for the service provided by us

9. Concept Generation

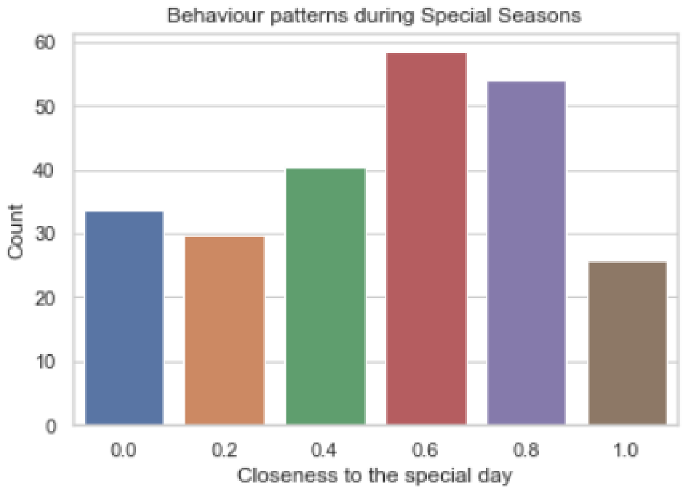
Purpose Deposite the increasing utilization of webpages for the purposes of information seeking, customers' concerns havig become a crucial impediment for online shopping. The purpose of this paper is to examine the influence of the effectiveness of web assurance seals services (WASS) and customers' concerns on customer's willingness to purchase.


Perceived Advantages (H1)
Perceived Risks (H2)
H4
Trust (H3)
H6
Computer/Web/knowledge/ Experience (H5)
Online Shopping Intention
Website/ Internet Quality (H7)

10.Concept Development

- Represents the special day(eg :Christmas , Halloween , thanksgiving)

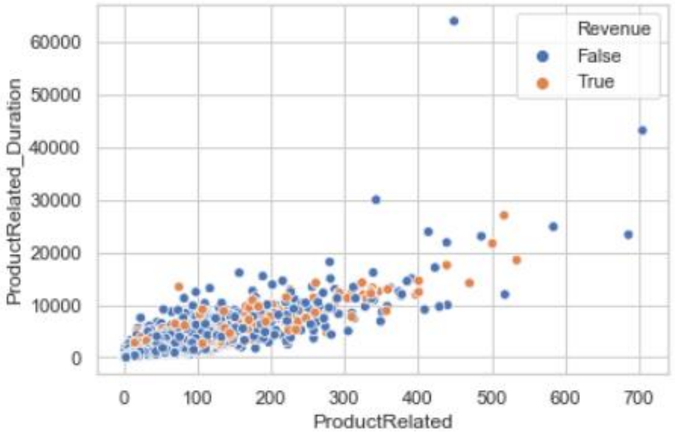
- Represents all the 11 months preceding the special day. (eg : nov 1 to sep 30 on the case of Halloween)
- 0.2 to 1.0 – each of the decimal bucket represents 6 days from oct 1 to oct 30 with 0.2 being toward the week on the start of the month and 0.8 being the week before the special day.



- We find that there is a general uptrend in buying pattern as the special day approaches and peaks during the 2 weeks preceding the festival and drops down on the day of the festival itself.
- It might be a great idea to stock up on extra inventory and offer special discounts and gift cards during these couple of weeks before the holiday season to boost sales even further.

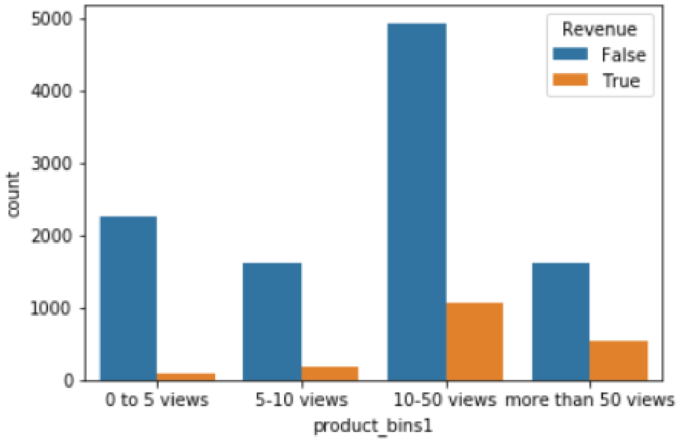
11. Product Details (How does it work)

- Graph 1 is a plot of Prouct related pages viewed vs amount of time spent.
- We find the conversion points(orange) to be distributed throughout.



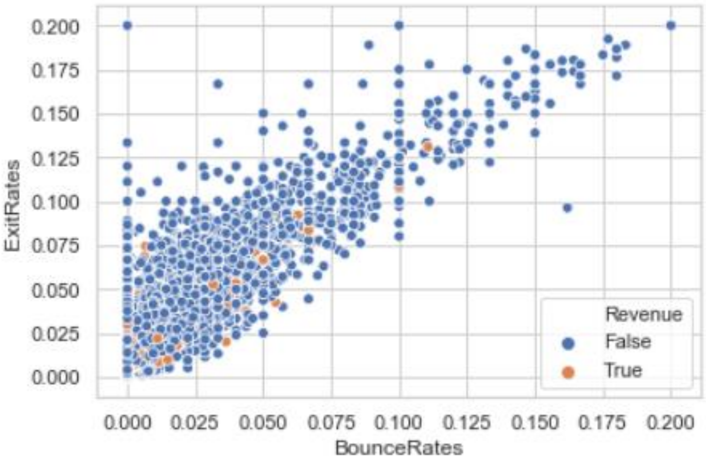
Graph 1

- Graph 2 is a bin-wise analysis and we find conversion percentage to be highest for people who visit more than 50 times followed by people who visit 10-50times.



Graph 2

- Graph 3 shows bounce rates and exit rates w.r.t Revenue.
- We can see the vast majority of people who bounce or exit the website don’t make a purchase.



Graph 3

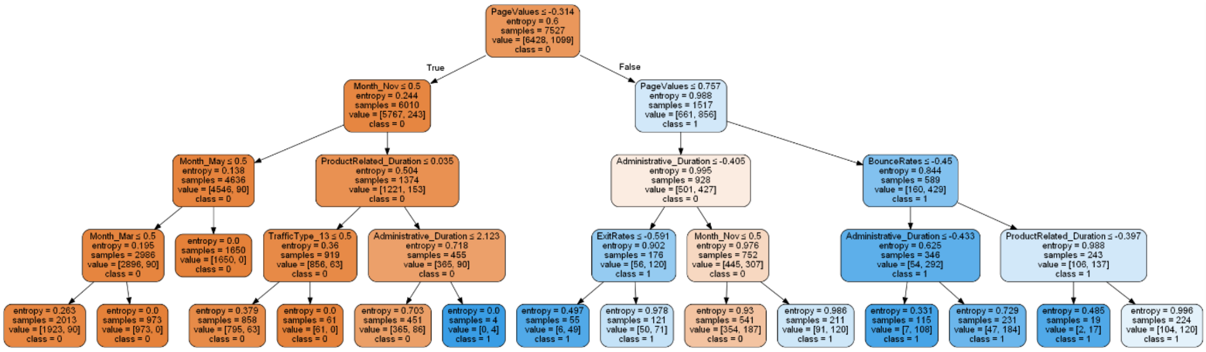
Business Insight:

1. Our target customers are the ones who visit the product from 10 to 400 times. The outliers generally do not end up purchasing.
2. We do not need to allocate resources to improve user experience for people with high bounce and exit rates. Only few people in the lower spectrum end up making a purchase.

12.Final Product Prototype(abstract) with Schematic Diagram

1. Summary of problem statement, data and findings Every good abstract describes succinctly what was intended at the outset, and summarizes findings and implications.
2. Overview of the final process Briefly describe your problem solving methodology. Include information about the salient features of your data, data pre-processing steps, the algorithms you used, and how you combined techniques.
3. Step-by-step walk through of the solution Describe the steps you took to solve the problem. What did you find at each stage, and how did it inform the next steps? Build up to the final solution.

4. Model evaluation Describe the final model (or ensemble) in detail. What was the objective, what parameters were prominent, and how did you evaluate the success of your models(s)? A convincing explanation of the robustness of your solution will go a long way to supporting your solution.

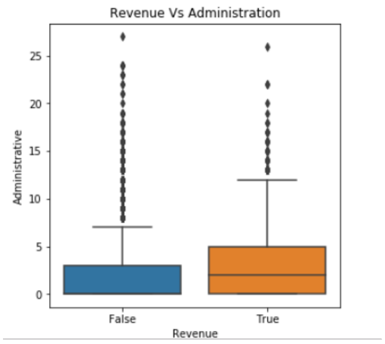

PageValues ≤ -0.314 entropy = 0.6 samples = 7527 value = [6428, 1099] class = 0
True
False
Month_Nov ≤ 0.5 entropy = 0.244 samples = 6010 value = [5767, 243] class = 0
PageValues ≤ 0.757 entropy = 0.988 samples = 1517 value = [661, 856] class = 1
Month_May ≤ 0.5 entropy = 0.138 samples = 4636 value = [4546, 90] class = 0
ProductRelated_Duration ≤ 0.035 entropy = 0.504 samples = 1374 value = [1221, 153] class = 0
Administrative_Duration ≤ -0.405 entropy = 0.995 samples = 928 value = [501, 427] class = 0
BounceRates ≤ -0.45 entropy = 0.844 samples = 589 value = [160, 429] class = 1
Month_Mar ≤ 0.5 entropy = 0.195 samples = 2986 value = [2896, 90] class = 0
entropy = 0.0 samples = 1650 value = [1650, 0] class = 0
TrafficType_13 ≤ 0.5 entropy = 0.36 samples = 919 value = [856, 63] class = 0
Administrative_Duration ≤ 2.123 entropy = 0.718 samples = 455 value = [365, 90] class = 0
ExitRates ≤ -0.591 entropy = 0.902 samples = 176 value = [56, 120] class = 1
Month_Nov ≤ 0.5 entropy = 0.976 samples = 752 value = [445, 307] class = 0
Administrative_Duration ≤ -0.433 entropy = 0.625 samples = 346 value = [54, 292] class = 1
ProductRelated_Duration ≤ -0.397 entropy = 0.988 samples = 243 value = [106, 137] class = 1
entropy = 0.263 samples = 2013 value = [1923, 90] class = 0
entropy = 0.0 samples = 973 value = [973, 0] class = 0
entropy = 0.379 samples = 858 value = [795, 63] class = 0
entropy = 0.0 samples = 61 value = [61, 0] class = 0
entropy = 0.703 samples = 451 value = [365, 86] class = 0
entropy = 0.0 samples = 4 value = [0, 4] class = 1
entropy = 0.497 samples = 55 value = [6, 49] class = 1
entropy = 0.978 samples = 121 value = [50, 71] class = 1
entropy = 0.93 samples = 541 value = [354, 187] class = 0
entropy = 0.986 samples = 211 value = [91, 120] class = 1
entropy = 0.331 samples = 115 value = [7, 108] class = 1
entropy = 0.729 samples = 231 value = [47, 184] class = 1
entropy = 0.485 samples = 19 value = [2, 17] class = 1
entropy = 0.996 samples = 224 value = [104, 120] class = 1

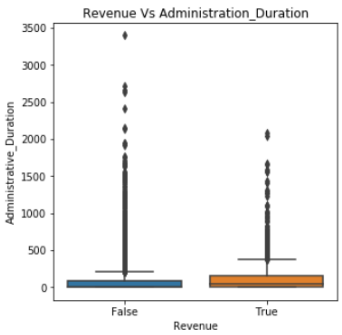
13.1 Some Basic Visualizations on Real World or Augmented Data

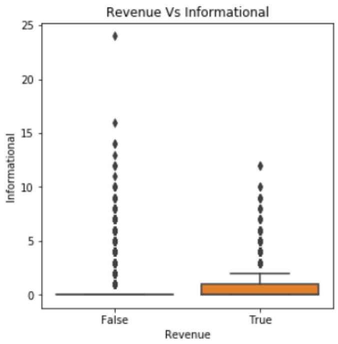
Detecting Outlier

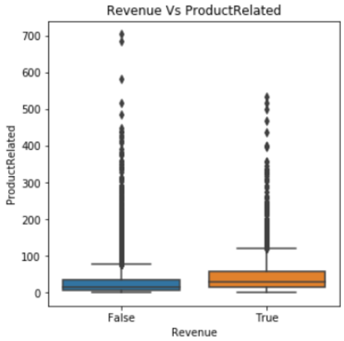
Boxplots

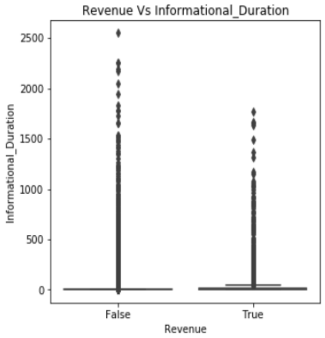
- Outliers were detected and analysed using the Outlier Box-plots.
- From the Outliers box-plot we can infer that the data consists of many outliers for the target variable, Revenue.
- Hence, we concluded that these outliers are legitimate outliers and we decided to retain them in the data.

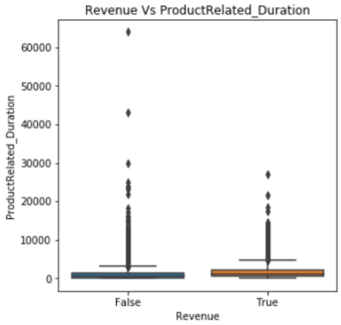

Revenue Vs Administration
Administrative
0
5
10
15
20
25
False
True
Revenue

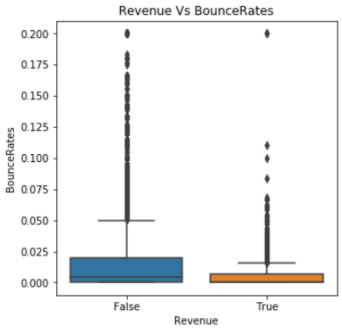

Revenue Vs Administration_Duration
Administrative_Duration
0
500
1000
1500
2000
2500
3000
3500
False
True
Revenue


Revenue Vs Informational
Informational
0
5
10
15
20
25
False
True
Revenue


Revenue Vs ProductRelated
ProductRelated
0
100
200
300
400
500
600
700
False
True
Revenue

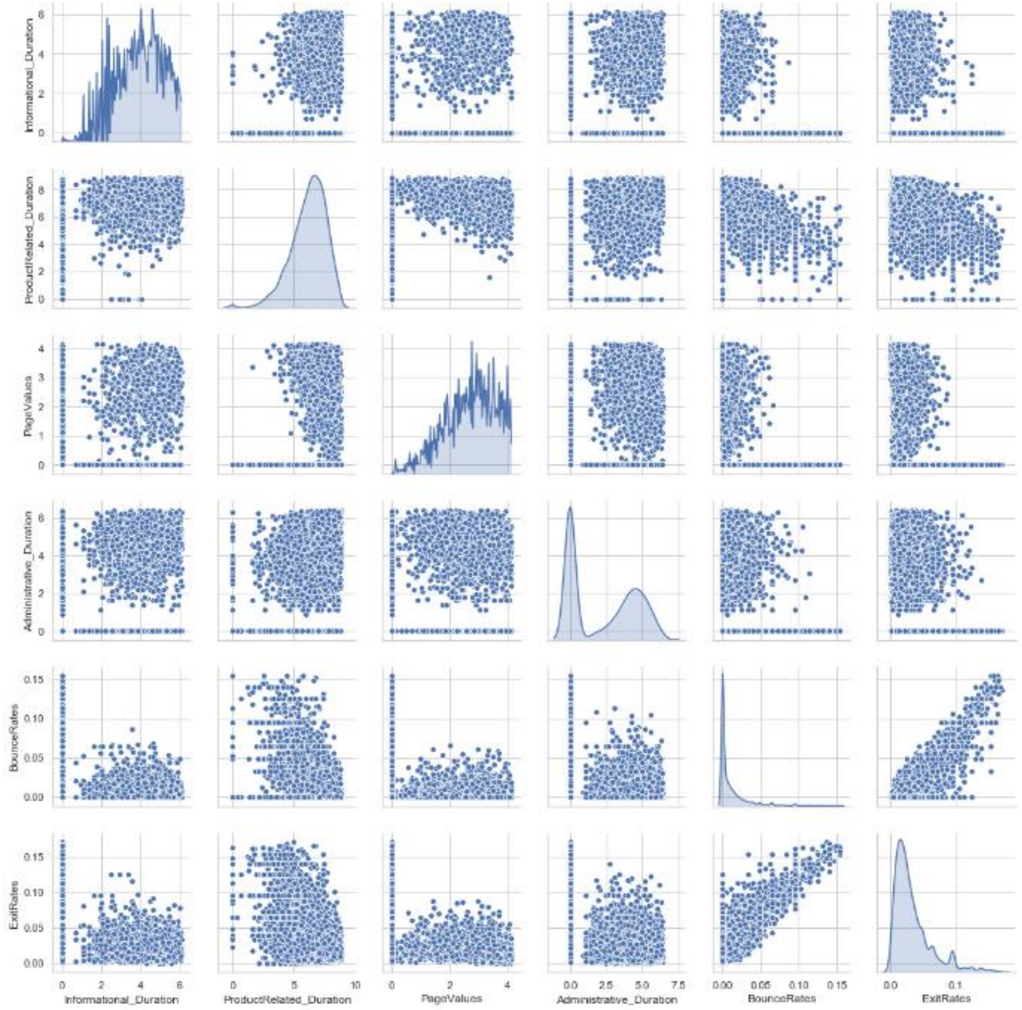

Revenue Vs Informational_Duration
Informational_Duration
0
500
1000
1500
2000
2500
False
True
Revenue


Revenue Vs ProductRelated_Duration
ProductRelated_Duration
0
10000
20000
30000
40000
50000
60000
False
True
Revenue


Revenue Vs BounceRates
BounceRates
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
False
True
Revenue

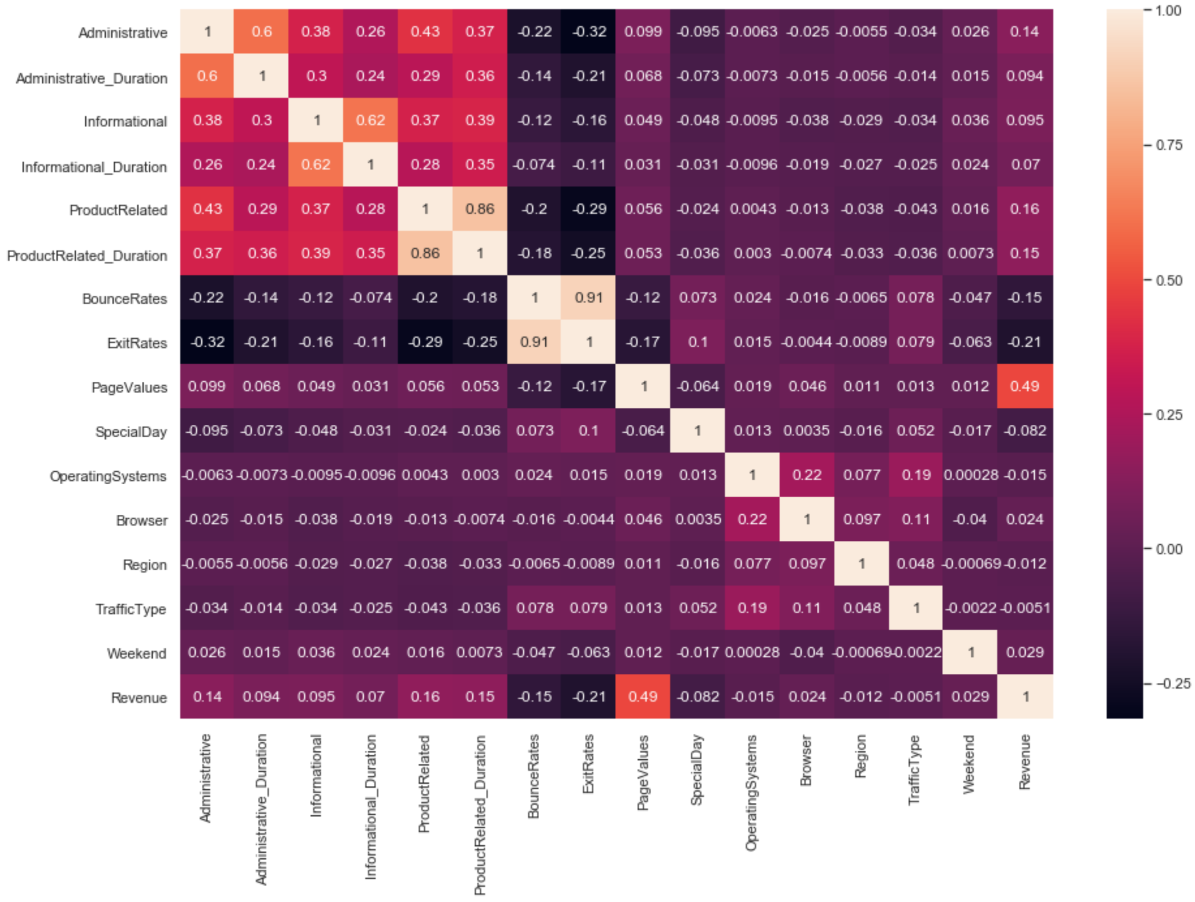
Pair plot

After the removed the Outliers form the data



CORRELATION TABLE

- Among the entire feature "Pagevalues" have high correlation with the Target variable.
- Most of the columns have positive correlation with the target variable. Only few have negative correlation. Among that "ExitRates" have high correlation.
- Most of the features have multicollinearity. Top 3 very strongly correlated Features are
    - 'BounceRates' & 'ExitRates'   = 0 .91
    - 'ProductRelated' & 'ProductRelated_Duration' = 0.86
    - 'Administrative' & 'Administrative_Duration' = 0.61



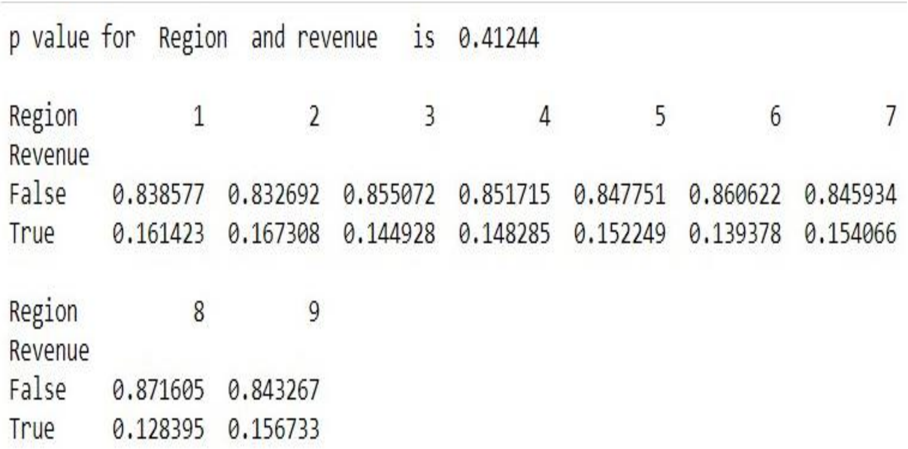
13.2 Exploratory Data Analysis

Chi-Square

- p value is less than 0.05, null hypothesis is rejected.
- p value for Region and revenue is 0.41244
- We can drop the column Region as there is no statistical relationship between Region and Revenue.

Null hypothesis

- Online buying behaviour doesn’t vary with the visitor type.
- Online buying behaviour doesn’t vary with the page values.
- Online buying behaviour doesn’t vary with the special day.
- Online buying behaviour doesn’t vary with the month
- Online buying behaviour doesn’t vary with weekday or weekend.
- Online buying behaviour doesn’t vary with the browser type.
- Online buying behaviour doesn’t vary with the Region.
- Online buying behaviour doesn’t vary with Administrative, Product or Information Related Pages.



```
p value for  Region  and revenue   is  0.41244

Region          1         2         3         4         5         6         7
Revenue
False    0.838577  0.832692  0.855072  0.851715  0.847751  0.860622  0.845934
True     0.161423  0.167308  0.144928  0.148285  0.152249  0.139378  0.154066

Region          8         9
Revenue
False    0.871605  0.843267
True     0.128395  0.156733
```

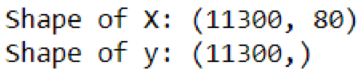
SMOTE MODEL for class imbalance

Our target variable 'Revenue' is highly unbalanced (ie: around 80% who don't buy the product and 20 % who buy the product) and hence we need to adopt some measures to balance the dataset in order to predict the model with a high precision score.

A simple way to generate synthetic samples is to randomly sample the attributes from instances in the minority class.

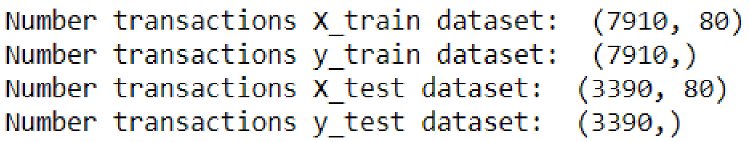
As its name suggests, SMOTE is an oversampling method. It works by creating synthetic samples from the minor class instead of creating copies. The algorithm selects two or more similar instances (using a distance measure) and perturbing an instance one attribute at a time by a random amount within the difference to the neighbouring instances.

Shape of the X and Y variables



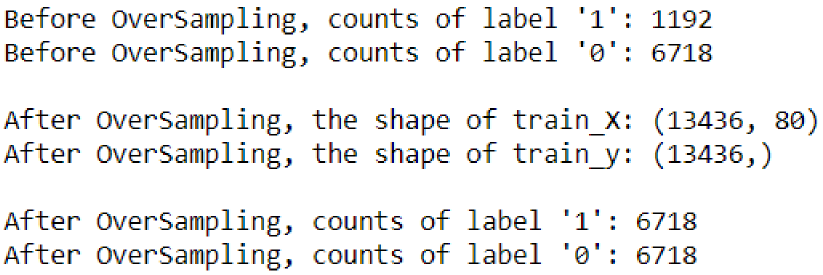
```
Shape of X: (11300, 80)
Shape of y: (11300,)
```

Shape of Train and Test



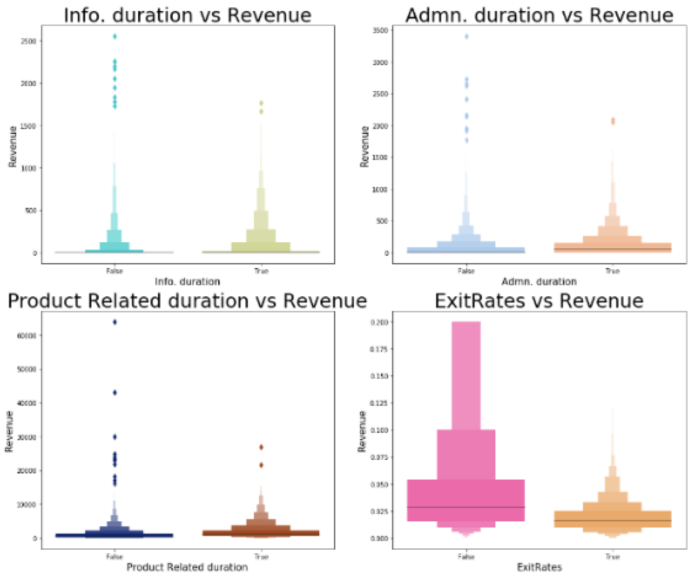
```
Number transactions X_train dataset:  (7910, 80)
Number transactions y_train dataset:  (7910,)
Number transactions X_test dataset:  (3390, 80)
Number transactions y_test dataset:  (3390,)
```

Smote Model values before and after Oversampling

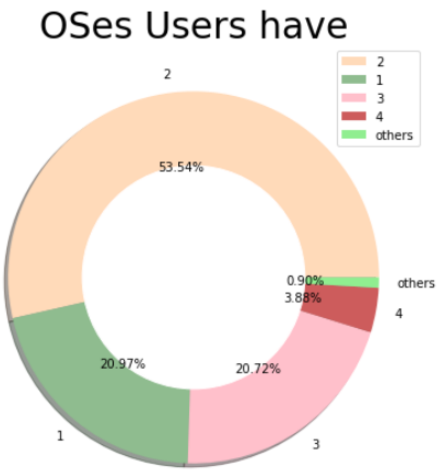


```
Before OverSampling, counts of label '1': 1192
Before OverSampling, counts of label '0': 6718

After OverSampling, the shape of train_X: (13436, 80)
After OverSampling, the shape of train_y: (13436,)

After OverSampling, counts of label '1': 6718
After OverSampling, counts of label '0': 6718
```

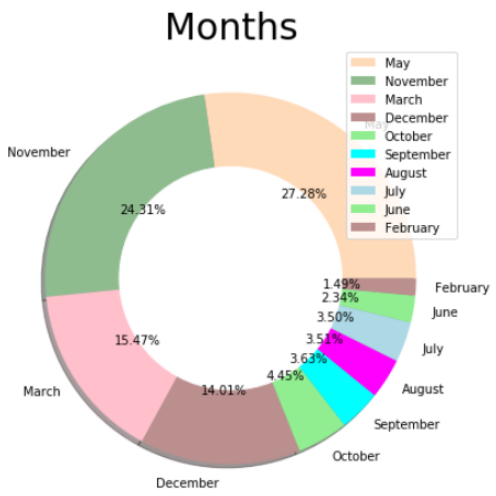
Below Graph shows the relation between Informational_Duration and Revenue, Administrative_Duration and Revenue, ProductRelated_Duration and Revenue, ExitRates and Revenue



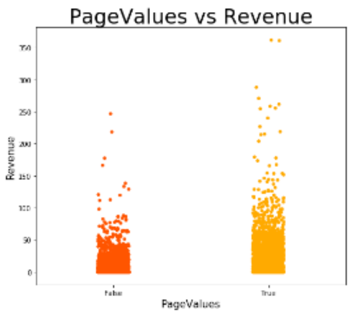
Below Graph shows us the distribution of revenue with respect to the type of OS used by the user based on their purchase and Special Days purchase based on month

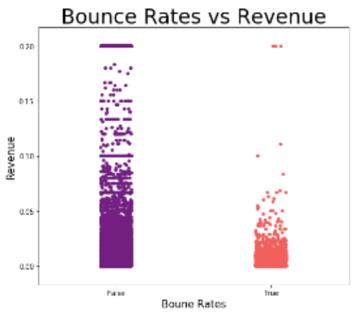


Below Graph shows the spread of revenue through the different months of the year. The revenue is higher just before the holiday seasons. i.e. In May and November.

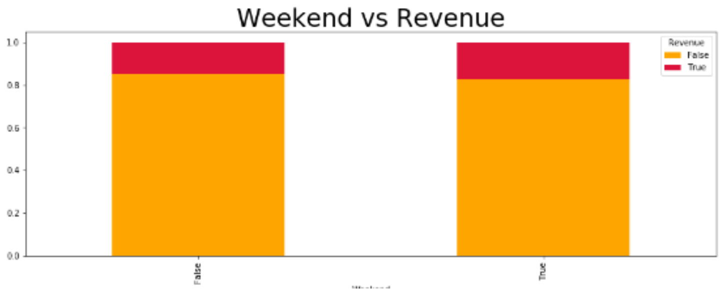


Below Graph illustrates the relation between these columns PageValues and Revenue, Bounce Rates and Revenue.

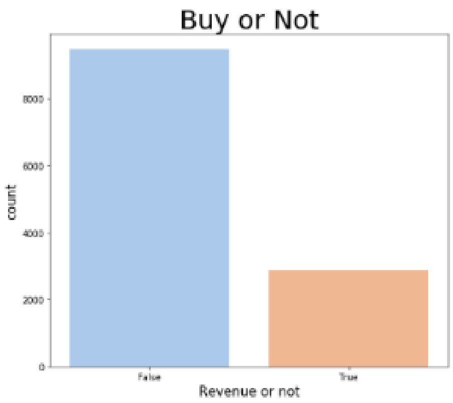


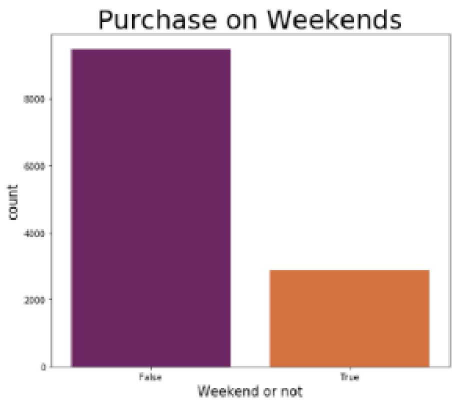


Below Graph illustrates the relation between Weekend and Revenue

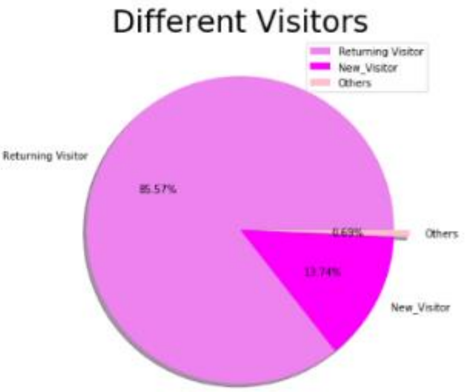


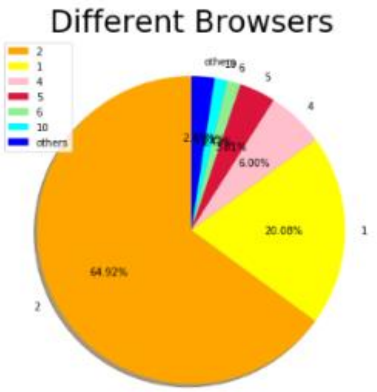
Below diagram shows the whether the customers purchase on Weekends or not.





Below diagram shows us the different types of visitors and Different types of Browsers which are used to buy the products.

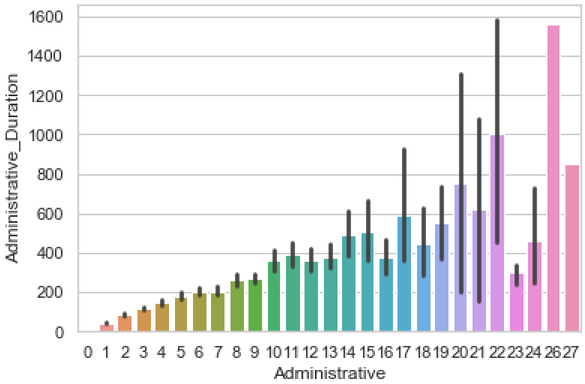




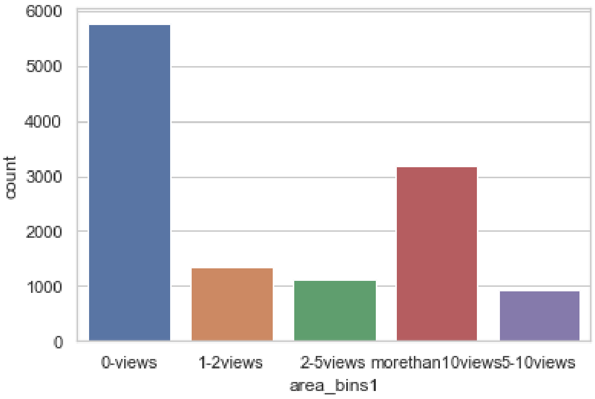
Converting into bins

Administration Related

Administrative vs Administrative_Duration (Graph 1)



Count vs Area_bins (Graph 2)

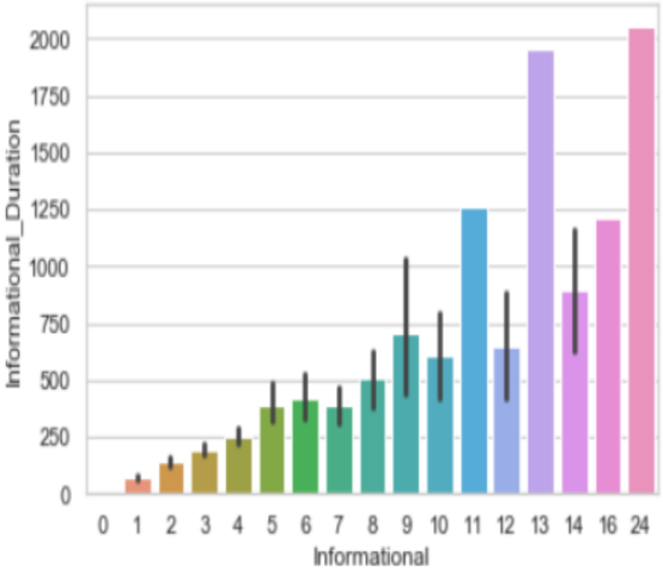


Administration Related Issues

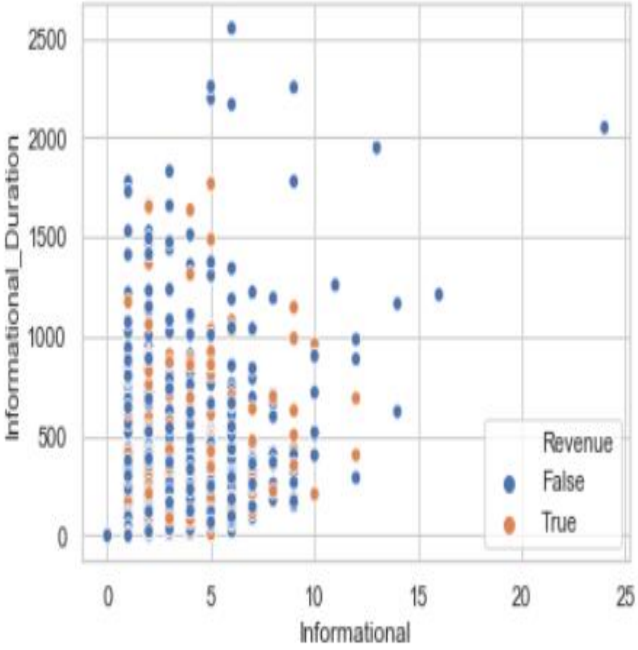
- Graph 1 shows the distribution for count of admin related issues vs duration spent on them.
- We can see that there is a general Uptrend.
- In Graph 2 we have created bins for admin related count and we can see that people who didn’t have any admin issue is highest.
- However, there are lot of people who visited admin related content more than 10 times also.
- This tells us that we might need to look into our tech support and IT content and improve them.

Informational Related

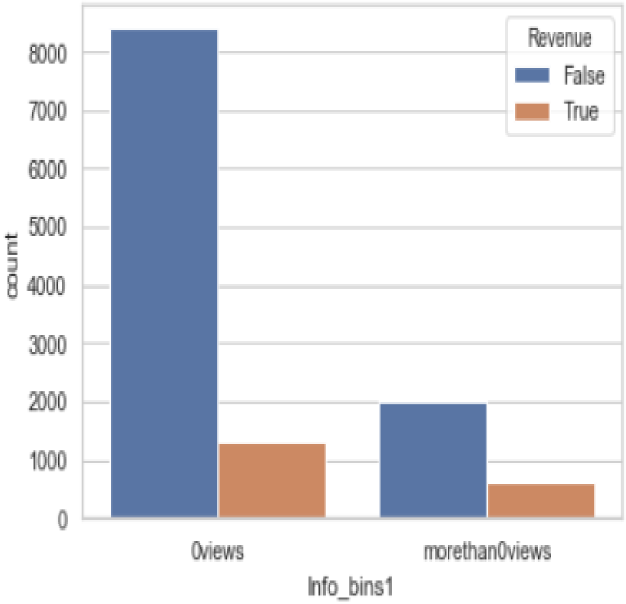
Informational_Duration vs Informational (Graph 3)



Informational_Duration vs Informational (Graph 3)



Revenue vs Info_Bins1(Graph 5)

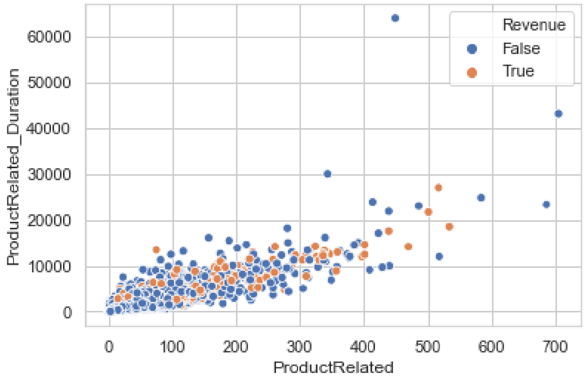


Information Related Content

- Graph 3 shows the count of info related websites visited by the user vs duration spent on it.
- We see that this is also on a general uptrend.
- Graph 4 shows people who research too much don't end up buying.
- Ideal region – people who visit below 5 info related content and spend below 1000secs.
- Graph 5 show that conversion percentage is higher for people who have visited info related content (i.e.: people who have done their research)

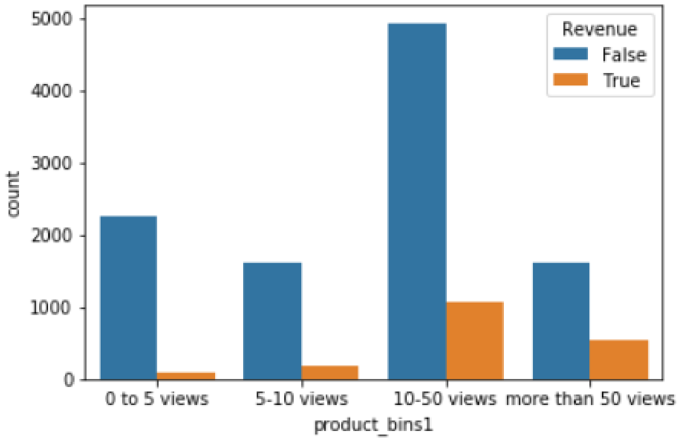
Product Related

ProductRelated_Duration vs Product_Related



Graph 6

Revenue vs Product_bins



Graph 7

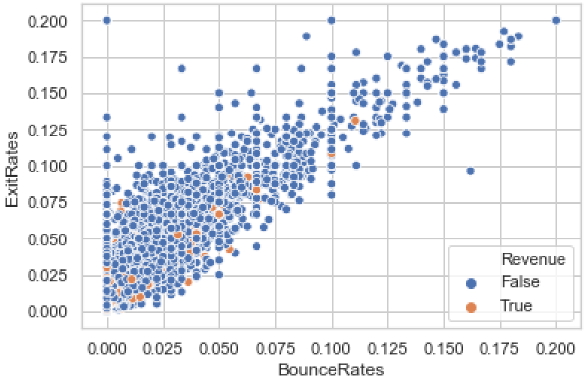
Product Related Analysis

- Graph 6 is a plot of Product related pages viewed vs amount of time spent.
- We find the conversion points(orange) to be distributed throughout.
- Graph 7 is a bin-wise analysis and we find conversion percentage to be highest for people who visit more than 50 times followed by people who visit 10-50times.

Business Insight

Our target customers are the ones who visit the product from 10 to 400 times. The outliers generally do not end up purchasing.

ExitRates vs BounceRates



Bounce Rates & Exit rates Analysis

- Graph 3 shows bounce rates and exit rates w.r.t Revenue.
- We can see the vast majority of people who bounce or exit the website don’t make a purchase.

Business Insight

We do not need to allocate resources to improve user experience for people with high bounce and exit rates. Only few people in the lower spectrum end up making a purchase.

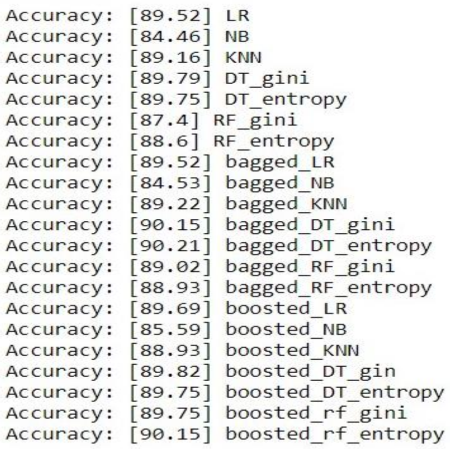
14.3 ML Modelling

We have used 21 models to know the Accuracy and AUC for the given data. In that models, there is 7 base models, 14 models with bagging and boosting each. Here I have included snips for them.

- Base Models:
  - Linear Regression
  - Naive Bayes
  - KNN
  - Decision Tree – GINI
  - Decision Tree – Entropy
  - Random Forest – GINI

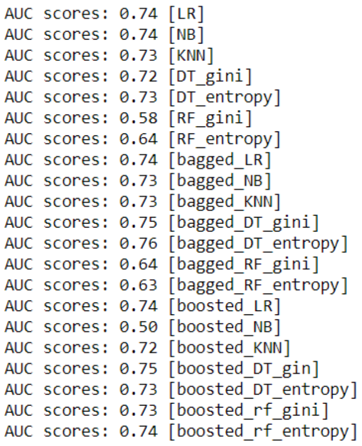
- Random Forest – Entropy

Accuracy for the models



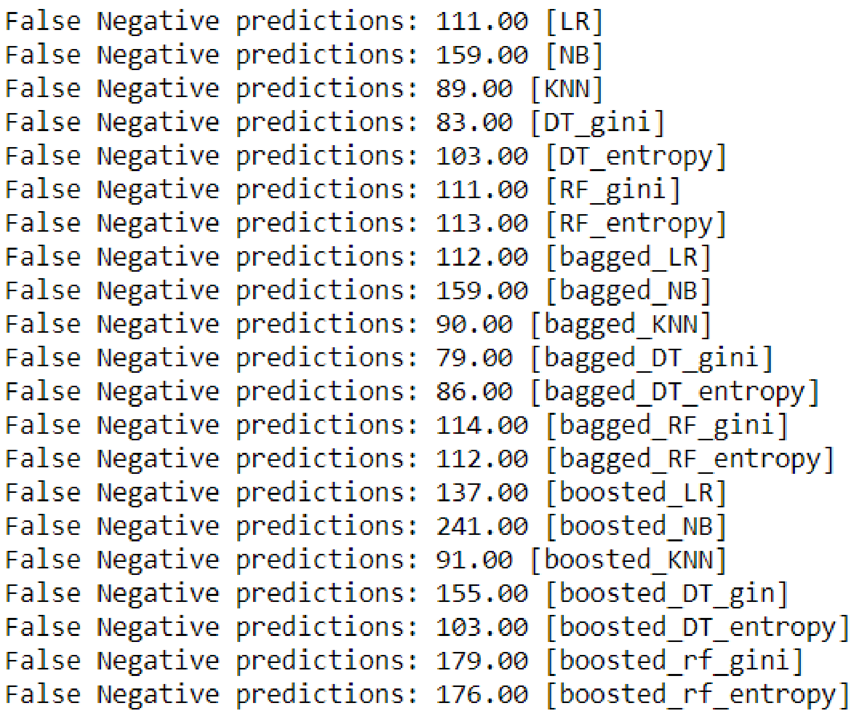
```
Accuracy: [89.52] LR
Accuracy: [84.46] NB
Accuracy: [89.16] KNN
Accuracy: [89.79] DT_gini
Accuracy: [89.75] DT_entropy
Accuracy: [87.4] RF_gini
Accuracy: [88.6] RF_entropy
Accuracy: [89.52] bagged_LR
Accuracy: [84.53] bagged_NB
Accuracy: [89.22] bagged_KNN
Accuracy: [90.15] bagged_DT_gini
Accuracy: [90.21] bagged_DT_entropy
Accuracy: [89.02] bagged_RF_gini
Accuracy: [88.93] bagged_RF_entropy
Accuracy: [89.69] boosted_LR
Accuracy: [85.59] boosted_NB
Accuracy: [88.93] boosted_KNN
Accuracy: [89.82] boosted_DT_gin
Accuracy: [89.75] boosted_DT_entropy
Accuracy: [89.75] boosted_rf_gini
Accuracy: [90.15] boosted_rf_entropy
```

AUC for the models



```
AUC scores: 0.74 [LR]
AUC scores: 0.74 [NB]
AUC scores: 0.73 [KNN]
AUC scores: 0.72 [DT_gini]
AUC scores: 0.73 [DT_entropy]
AUC scores: 0.58 [RF_gini]
AUC scores: 0.64 [RF_entropy]
AUC scores: 0.74 [bagged_LR]
AUC scores: 0.73 [bagged_NB]
AUC scores: 0.73 [bagged_KNN]
AUC scores: 0.75 [bagged_DT_gini]
AUC scores: 0.76 [bagged_DT_entropy]
AUC scores: 0.64 [bagged_RF_gini]
AUC scores: 0.63 [bagged_RF_entropy]
AUC scores: 0.74 [boosted_LR]
AUC scores: 0.50 [boosted_NB]
AUC scores: 0.72 [boosted_KNN]
AUC scores: 0.75 [boosted_DT_gin]
AUC scores: 0.73 [boosted_DT_entropy]
AUC scores: 0.73 [boosted_rf_gini]
AUC scores: 0.74 [boosted_rf_entropy]
```

False Negative predictions



```
False Negative predictions: 111.00 [LR]
False Negative predictions: 159.00 [NB]
False Negative predictions: 89.00 [KNN]
False Negative predictions: 83.00 [DT_gini]
False Negative predictions: 103.00 [DT_entropy]
False Negative predictions: 111.00 [RF_gini]
False Negative predictions: 113.00 [RF_entropy]
False Negative predictions: 112.00 [bagged_LR]
False Negative predictions: 159.00 [bagged_NB]
False Negative predictions: 90.00 [bagged_KNN]
False Negative predictions: 79.00 [bagged_DT_gini]
False Negative predictions: 86.00 [bagged_DT_entropy]
False Negative predictions: 114.00 [bagged_RF_gini]
False Negative predictions: 112.00 [bagged_RF_entropy]
False Negative predictions: 137.00 [boosted_LR]
False Negative predictions: 241.00 [boosted_NB]
False Negative predictions: 91.00 [boosted_KNN]
False Negative predictions: 155.00 [boosted_DT_gin]
False Negative predictions: 103.00 [boosted_DT_entropy]
False Negative predictions: 179.00 [boosted_rf_gini]
False Negative predictions: 176.00 [boosted_rf_entropy]
```

Stacking Classifier

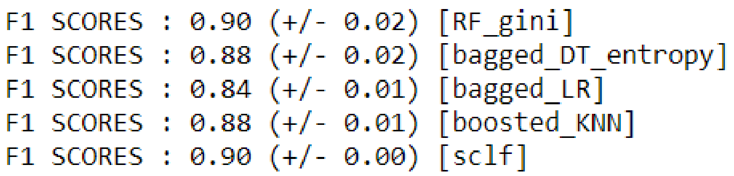
We finally build a stacked model to try and improve our F1 and recall scores by using a stacked model.

The point of stacking is to explore a space of different models for the same problem. The idea is that you can attack a learning problem with different types of models which are capable to learn some part of the problem, but not the whole space of the problem.

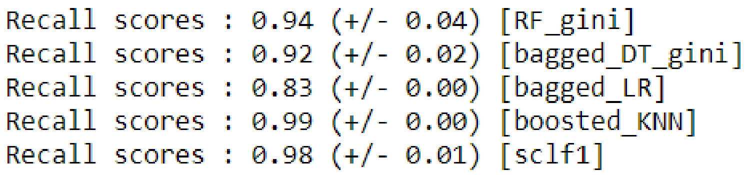
So, you can build multiple different learners and you use them to build an intermediate prediction, one prediction for each learned model. Then you add a new model which learns from the intermediate predictions the same target.

This final model is said to be stacked on the top of the others, hence the name. Thus, you might improve your overall performance, and often you end up with a model which is better than any individual intermediate model.

Below diagrams shows the F1 Scores and Recall Scores for the models



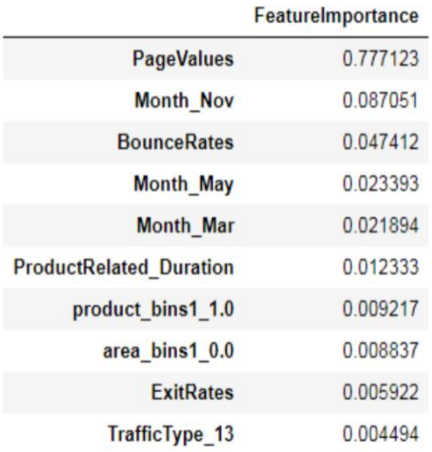
```
F1 SCORES : 0.90 (+/- 0.02) [RF_gini]
F1 SCORES : 0.88 (+/- 0.02) [bagged_DT_entropy]
F1 SCORES : 0.84 (+/- 0.01) [bagged_LR]
F1 SCORES : 0.88 (+/- 0.01) [boosted_KNN]
F1 SCORES : 0.90 (+/- 0.00) [sclf]
```



```
Recall scores : 0.94 (+/- 0.04) [RF_gini]
Recall scores : 0.92 (+/- 0.02) [bagged_DT_gini]
Recall scores : 0.83 (+/- 0.00) [bagged_LR]
Recall scores : 0.99 (+/- 0.00) [boosted_KNN]
Recall scores : 0.98 (+/- 0.01) [sclf1]
```

Feature Importance

The following snip shows some of the most important features that had a greater impact on the model and its performance.

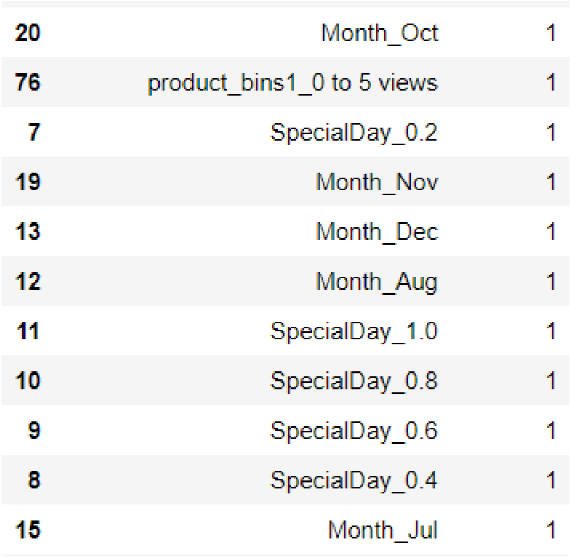


	FeatureImportance
PageValues	0.777123
Month_Nov	0.087051
BounceRates	0.047412
Month_May	0.023393
Month_Mar	0.021894
ProductRelated_Duration	0.012333
product_bins1_1.0	0.009217
area_bins1_0.0	0.008837
ExitRates	0.005922
TrafficType_13	0.004494

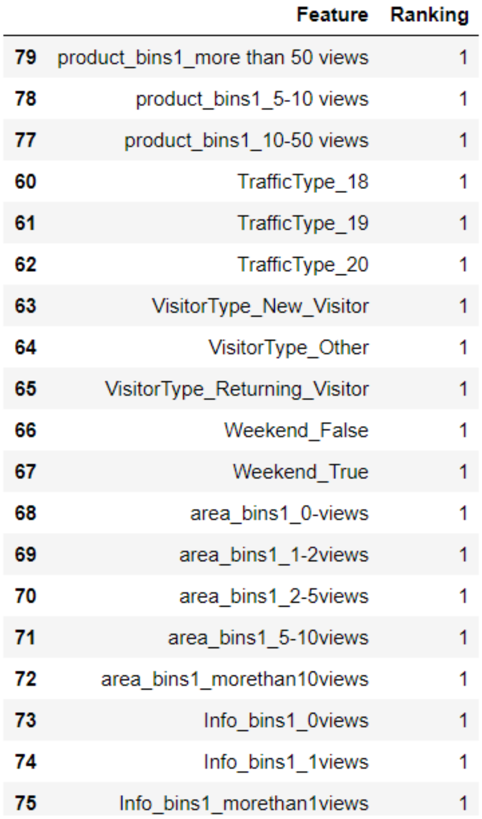
We find that PageValues is the most important feature that influences our model followed by the Month of purchase.

Bounce rates /exit rates and the duration for which a person browsed a product are also contributing features that affect our model.

Recursive Feature Selection



20	Month_Oct	1
76	product_bins1_0 to 5 views	1
7	SpecialDay_0.2	1
19	Month_Nov	1
13	Month_Dec	1
12	Month_Aug	1
11	SpecialDay_1.0	1
10	SpecialDay_0.8	1
9	SpecialDay_0.6	1
8	SpecialDay_0.4	1
15	Month_Jul	1



	Feature	Ranking
79	product_bins1_more than 50 views	1
78	product_bins1_5-10 views	1
77	product_bins1_10-50 views	1
60	TrafficType_18	1
61	TrafficType_19	1
62	TrafficType_20	1
63	VisitorType_New_Visitor	1
64	VisitorType_Other	1
65	VisitorType_Returning_Visitor	1
66	Weekend_False	1
67	Weekend_True	1
68	area_bins1_0-views	1
69	area_bins1_1-2views	1
70	area_bins1_2-5views	1
71	area_bins1_5-10views	1
72	area_bins1_morethan10views	1
73	Info_bins1_0views	1
74	Info_bins1_1views	1
75	Info_bins1_morethan1views	1

Python Code file

Capstoneproj.ipynb

13.4Github link to the code implementation

https://github.com/UmaMaheswariBalasubramanian/Project-1

15. Conclusion

Insights

Page Value vs Revenue

- Since Page Value is the attribute that has the highest influence on our model, once we know the page value of different pages, we can extensively promote those pages with low traffic volume and high page value to boost your conversion rate.
- **Page value = (Total page value + Transaction revenue) / Total unique pageviews**

Month Vs Revenue

- The footfalls are highest in May followed by November.
- However, conversion percentage seems to be low in May, this might be a good time to give blockbuster discounts and EOS sale to boost conversion rate.
- The sale seems to be good in Nov, it’s a good idea to stock up extra inventory and also offer good discounts in Nov.
- There are very few footfalls and almost zero purchase in Feb, we might need to look into that as well.

Weekday Vs Weekend Revenue

- The number of footfalls seem to be higher on weekdays compared to weekends.
- However, the conversion rate seems to be slightly better on weekends.
- A good way to attract more people to the website and improve conversion rate would be to send notifications to users regarding weekend discounts right from Thursday or Friday.

Visitor Type Vs Revenue

- Here we see that footfalls for returning visitor is way higher compared to new visitors, however the conversion rate of new visitors is higher compared to returning visitor.

- To honour returning visitors we could introduce loyalty points and bonuses which they could redeem if they are a returning user.
- Since the no of new visitors is very low, we might need to look into our digital marketing and SEO options to increase footfalls.