Determination of Likelihood of Purchase

In Online Shopping

Submitted towards partial fulfilment of the criteria for the award of PGP-DSE by GREAT LAKES INSTITUTE OF MANAGEMENT

Group – 7

Candidate Name	Roll Number
Ambar Ghosh	DSEFTBJAN19003
Ankita Deore	DSEFTBJAN19006
Niveditha Ramayanam	DSEFTBJAN19015
Prahlad Agnihotri	DSEFTBJAN19017
Pratik Nag	DSEFTBJAN19026

Project Mentor – Mr P.V.Subramanian

greatlearning

CERTIFICATE OF ORIGINALITY

This is to certify that the project titled ‘Determination of Likelihood of purchase in online shopping’ is an original work of the students namely Ambar Ghosh, Ankita Deore, Niveditha Ramayanam, Prahlad Agnihotri and Pratik Nag; and the final report is being submitted towards the fulfilment of the criteria for the award of PGP-DSE by GLIM.

Table of Content:

Executive Summary:

We ‘Bangalore Group7’ are working on Project - ‘Determination of Likelihood of Purchase in Online Shopping.’

It has often been observed that there are a lot people visiting numerous online shopping websites to browse for their needs but the conversion rate into purchases is low.

The main purpose of this project is to study the customer’s behavioural patterns for a getting a insight about the features that influence the sales.

The dataset given to us contains 12330 rows and 18 columns/features.

Introduction:

Nowadays, online shopping has become a very common scenario. There are many shopping websites offering a wide range of goods for people. However, sometimes it is observed that there are many users who browse through various websites but their probability of purchasing a product is very low.

There are quite a number of factors influencing the likelihood of purchasing a product from any shopping website.

Need of the study:

- For every customer, there is some behavioural pattern observed while visiting a specific shopping website that needs to be taken into consideration.

- There are divisions having a substantial amount of customers who often browse through certain category of products still their conversion rate into purchase is very low based upon certain parameters that needs to be focused upon.

- The probability that customers will make a transaction ensuring the purchase plays an important role in understanding the influence of sales.

Scope & Objectives:

Scope:

The scope of the project is to get insights from the given data with respect to the features, by understanding the features effectively; we will visualize the importance of each feature. By doing so we will understand how each feature will influence the purchase of the product.

Objective:

Predict how likely it is for a customer to make a purchase by building classification models.

Evaluate the model performance measures and choose the optimum model

Data-analysis:

1) Preparation of Dataset:

The dataset provided for this capstone contains 12330 rows and 18 columns.

Features	Description
Administrative	Number of pages visited by the user for user account management related activities
Administrative_Duration	Time spent on Admin pages by the user
Informational	Number of pages visited by the user about the website
Informational_Duration	Time spent on Informational pages by the user
ProductRelated	Number of product related pages visited by the user
ProductRelated_Duration	Time spent on Product related pages by the user
BounceRates	Average bounce rate of the pages visited by the user
ExitRates	Average exit rate of the pages visited by the user
PageValues	Average page value of the pages visited by the user

SpecialDay	Closeness of the visiting day to a special event like Mother's Day or festivals like Christmas
Month	Month of the visit from Jan to Dec
OperatingSystems	Operating System of the visitor
Browser	Browser of the visitor
Region	Geographic region from which the session has been started by the visitor
TrafficType	Traffic source through which user has entered the website
VisitorType	Visitor type as New visitor, Returning user or Others
Weekend	If the user visited on a weekend or not
Revenue	If the user visit resulted with a transaction

1. Formatting the dataset:

Formatting the data is converting the given data to machine-readable format. This means all the data points in each features should be of same data type. As our data was rectangular readable data, this step was not necessary.

2. Data Cleaning:

After formatting the data to machine-readable format, we have to clean the data to get better understanding of the data and to make it ready for the model building.

Basic Data Cleaning involves:

1. Handling-missing values.
2. Handling Outliers.
3. Data formatting.

1) Handling-missing values: – Our data did not had any missing values so this step was not necessary. However, if it had any we would have replaced the data in categorical feature with mode. In addition, data in continuous feature with median/mean depending on the data.

2) Handling Outliers: - Our data had too many outliers. Outliers were present in almost all the features. This will affect the predictive models like logistic regression, K-nn, etc.

   So we tried to handle these outliers by applying some transformations like Log transform, square transform, cube transform for left skewed features and square root, cube root, Log transform for right skewed features. However, even after trying all these transformations we still had lot of outliers so we decided to go for models, which can handle outliers like Decision tree – Predictive model, Random Forest – Ensemble technique

(bagging), Gradient Boosting – Ensemble Technique (boosting). However, we have tried all the models but as expected, the results were comparably bad.

3) Data formatting.

   Cleaning the data inside each features. As our data was in proper readable format this step was not necessary.

   Only step, which was necessary, was outlier handling. Because our data was in proper cleaned readable rectangular structure. However, data had too many outliers so we tried many possible ways to handle these outliers but still we could not. Therefore, we decided to go with models, which can handle outliers.

2) EDA for the given dataset:

   EDA is a general approach to exploring datasets by means of simple summary statistics and graphic visualizations in order to gain a deeper understanding of the data.

   We will start by exploring the Basic Descriptive statistics.

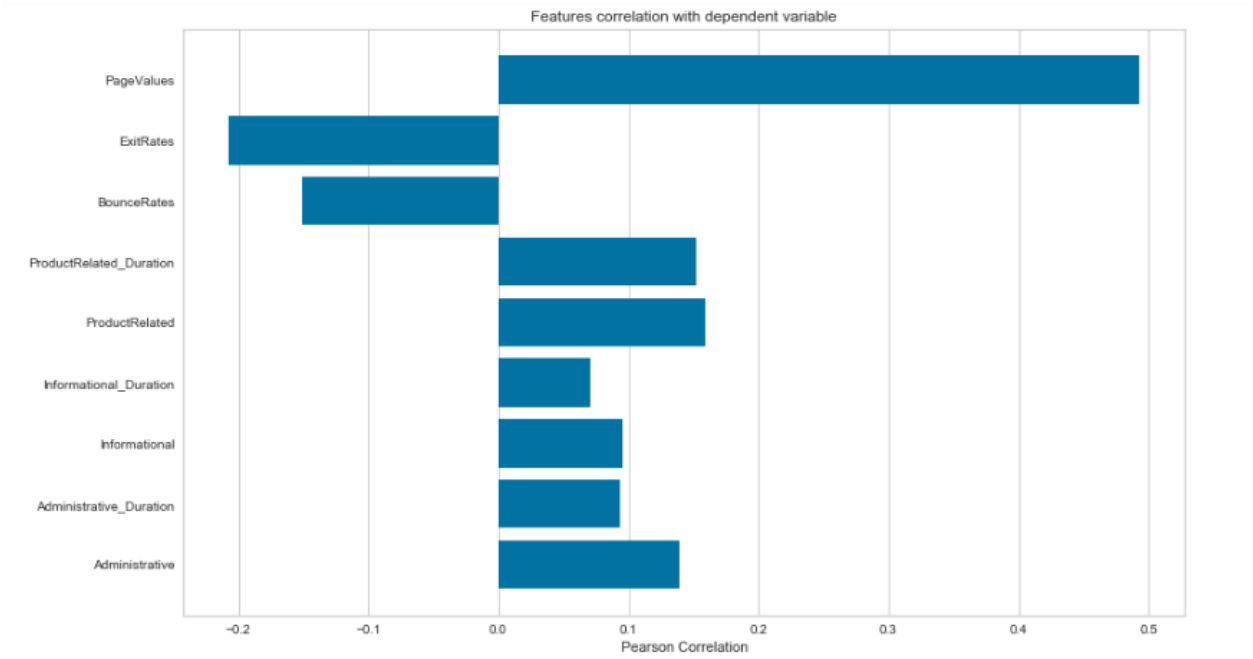
Descriptive Statistics:

	count	mean	std	min	25%	50%	75%	max
Administrative	12330.0	-1.375191e-15	1.000041	-0.696993	-0.696993	-0.395938	0.507228	7.431499
Administrative_Duration	12330.0	2.074316e-15	1.000041	-0.457191	-0.457191	-0.414764	0.070360	18.769559
Informational	12330.0	6.987391e-15	1.000041	-0.396478	-0.396478	-0.396478	-0.396478	18.499599
Informational_Duration	12330.0	1.765777e-16	1.000041	-0.244931	-0.244931	-0.244931	-0.244931	17.868683
ProductRelated	12330.0	-2.849753e-16	1.000041	-0.713488	-0.556092	-0.308755	0.140949	15.138577
ProductRelated_Duration	12330.0	1.021684e-15	1.000041	-0.624348	-0.528121	-0.311357	0.140788	32.806777
BounceRates	12330.0	1.333384e-15	1.000041	-0.457683	-0.457683	-0.393490	-0.110935	3.667189
ExitRates	12330.0	-2.622846e-16	1.000041	-0.886371	-0.592393	-0.368691	0.142551	3.229316
PageValues	12330.0	-4.953810e-15	1.000041	-0.317178	-0.317178	-0.317178	-0.317178	19.166337

Insights:

1. It is evident from the descriptive statistics that all the variables are highly skewed.
2. The standard deviation for all the variables are same

Pearson's Correlation



Insights:

1. As we expect, the attributes like PageValue, Product related duration, product related information, informational, information related duration, administrative, administrative duration are directly proportional to the target class

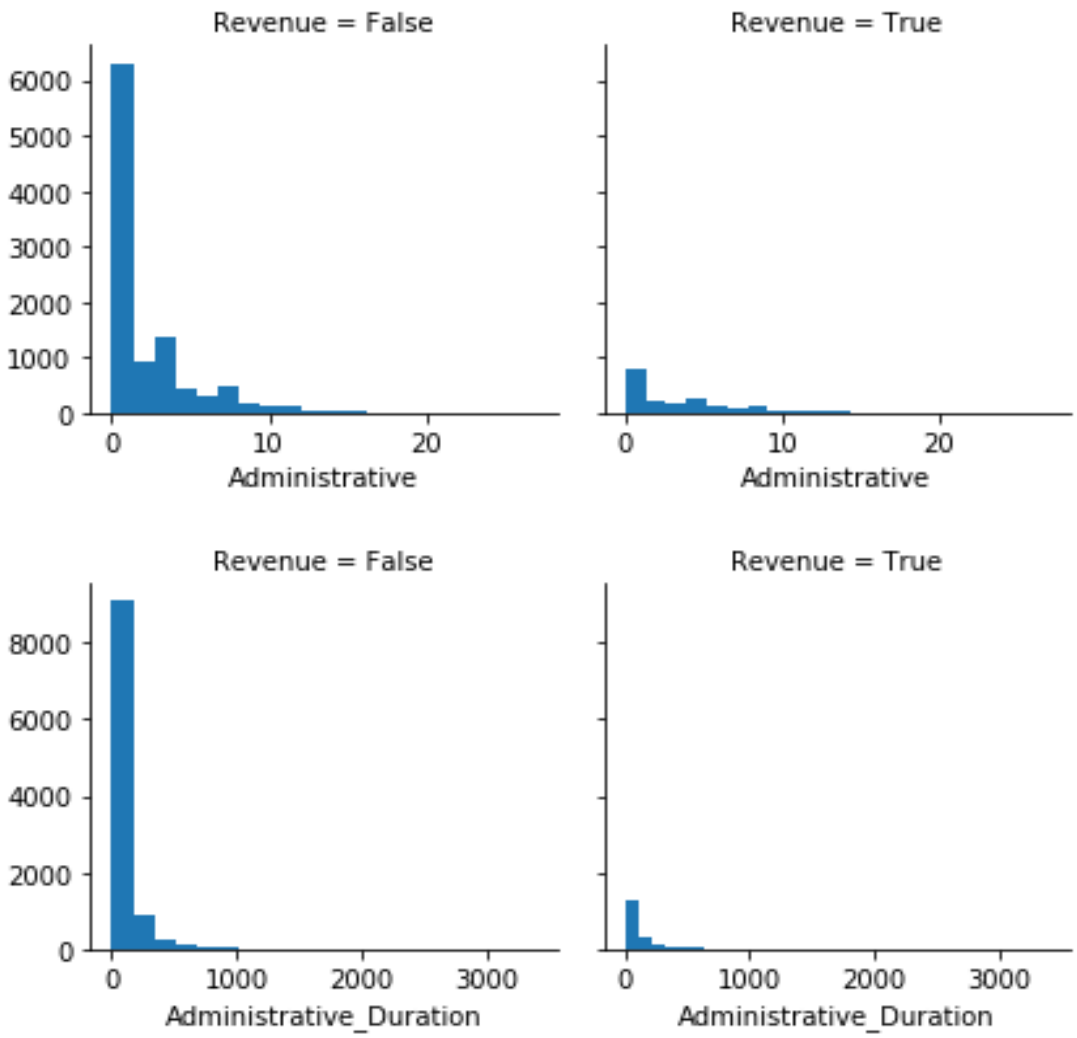
Classes like BounceRate and ExitRate are inversely proportional to the target class.

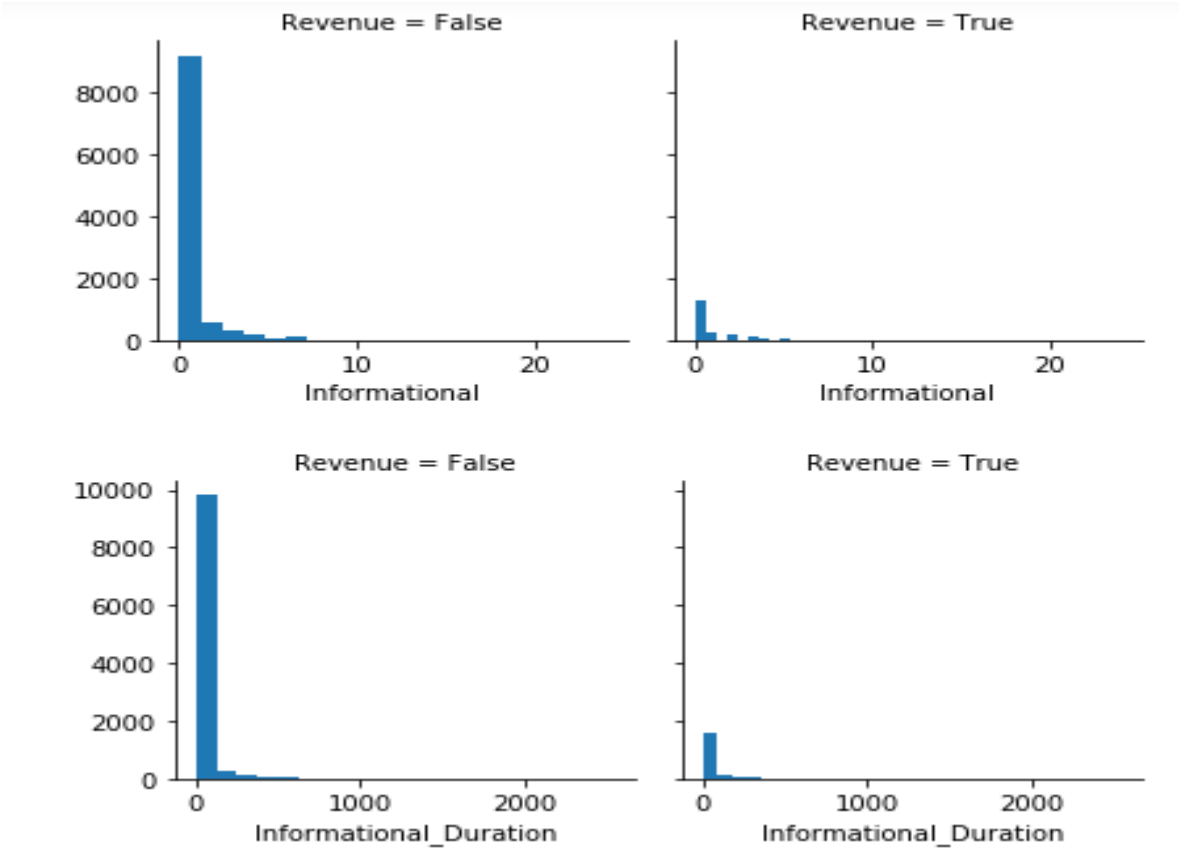
What we need to know:

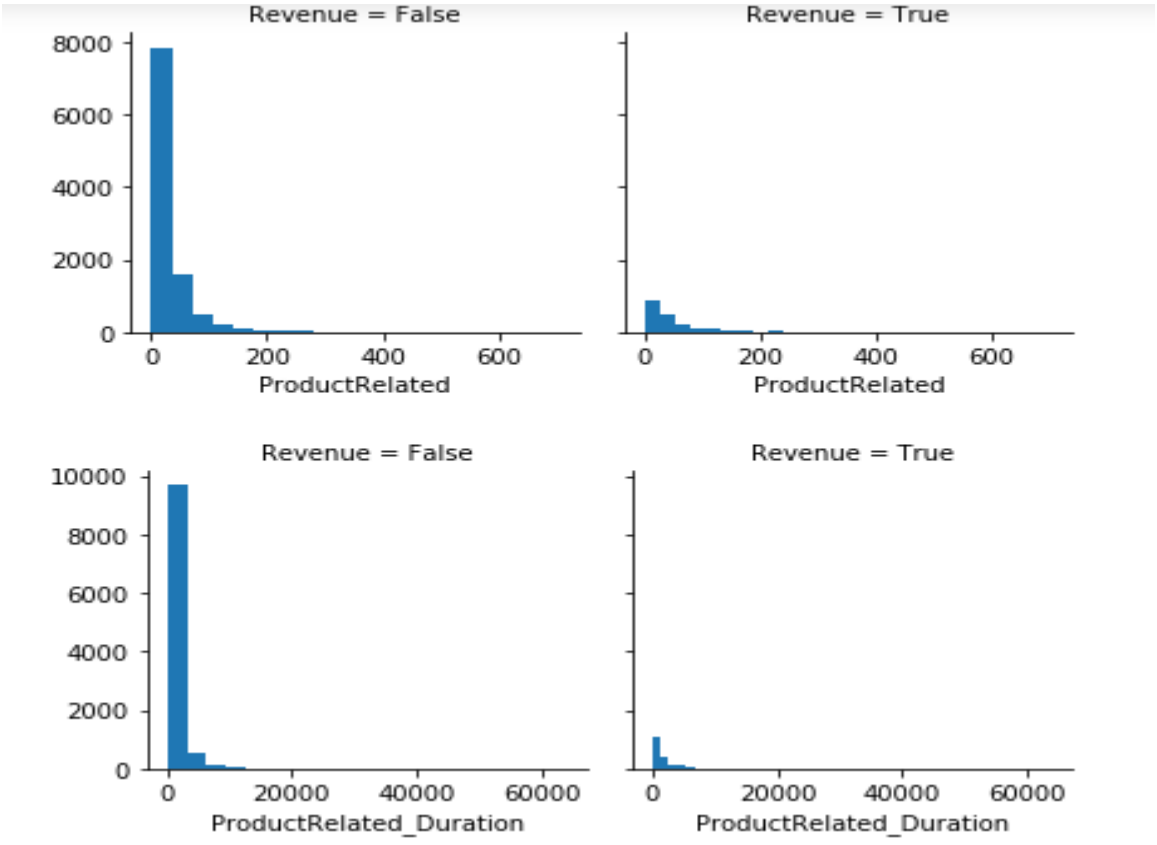
The numerical columns taken for EDA are:

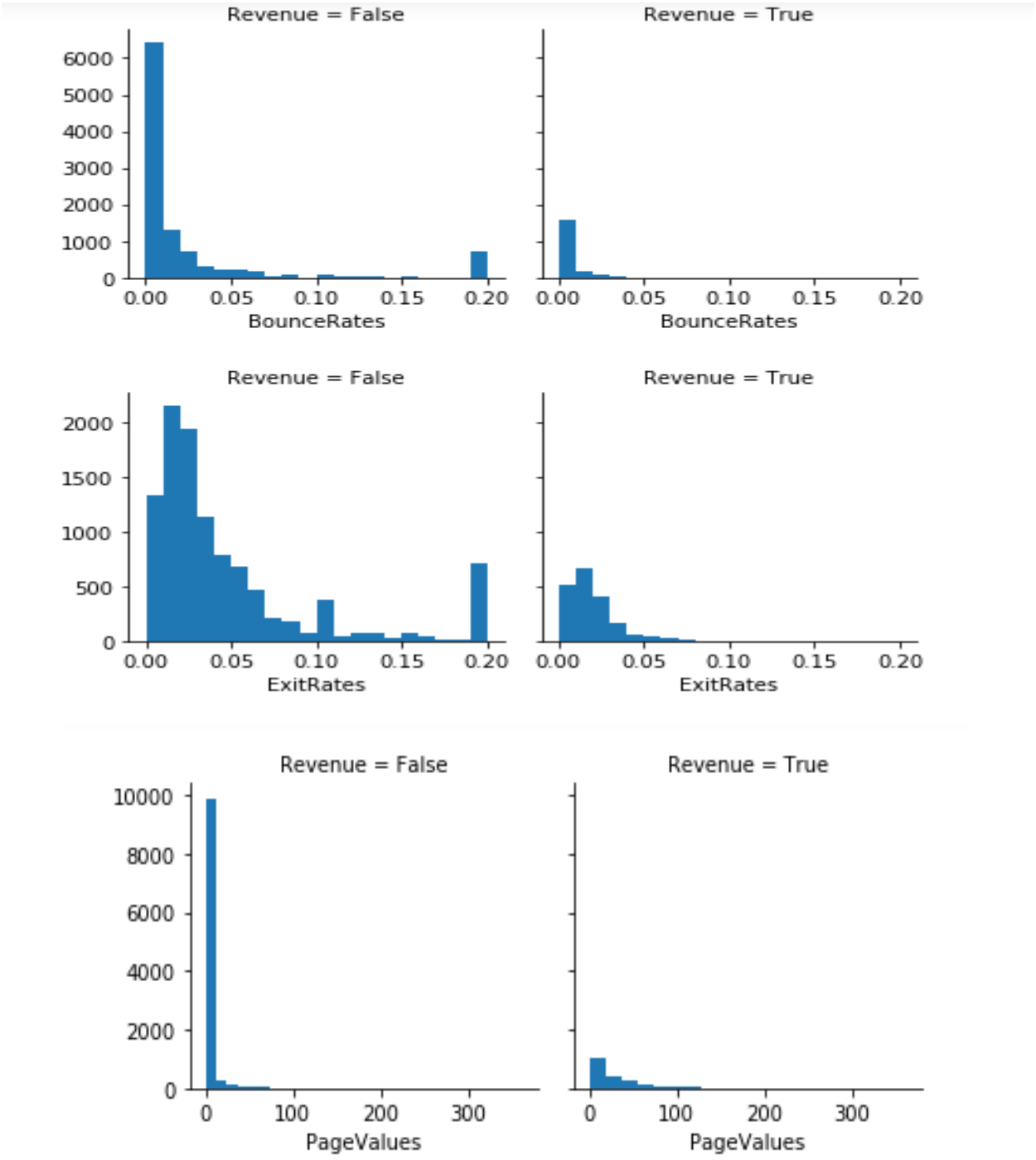
1. Administrative
2. Administrative_duration
3. Informational
4. Informational_duration
5. ProductRelated
6. ProductRelated_Duration
7. BounceRates
8. ExitRates
9. PageValues

Distribution of each Continuous attribute with respect to the Target Variable




Revenue = False
Revenue = True
8000
6000
4000
2000
0
0
10
20
Informational
Revenue = False
Revenue = True
10000
8000
6000
4000
2000
0
0
1000
2000
Informational_Duration


Revenue = False
Revenue = True
8000
6000
4000
2000
0
0
200
400
600
ProductRelated
Revenue = False
Revenue = True
10000
8000
6000
4000
2000
0
0
20000
40000
60000
ProductRelated_Duration


Revenue = False
Revenue = True
6000
5000
4000
3000
2000
1000
0
0.00
0.05
0.10
0.15
0.20
BounceRates
Revenue = False
Revenue = True
2000
1500
1000
500
0
ExitRates
Revenue = False
Revenue = True
10000
8000
6000
4000
2000
0
100
200
300
PageValues

Insights:

1. The data looks highly skewed. In addition, most of them are right skewed.
2. Therefore, we applied log transformations to those columns but the data still had lot of outliers. Therefore, we have decided to go for models, which can handle outliers.

Skewness and Kurtosis of the fields:

```
skewness of  Administrative  is: 0.57
kurtosis of  Administrative  is: -1.03
-----------------------------------------------
skewness of  Administrative_Duration  is: 1.61
kurtosis of  Administrative_Duration  is: 2.57
-----------------------------------------------
skewness of  Informational  is: 1.95
kurtosis of  Informational  is: 2.73
-----------------------------------------------
skewness of  Informational_Duration  is: 3.61
kurtosis of  Informational_Duration  is: 14.28
-----------------------------------------------
skewness of  ProductRelated  is: 0.84
kurtosis of  ProductRelated  is: 0.48
-----------------------------------------------
skewness of  ProductRelated_Duration  is: 1.01
kurtosis of  ProductRelated_Duration  is: 0.82
-----------------------------------------------
skewness of  BounceRates  is: 1.94
kurtosis of  BounceRates  is: 3.02
-----------------------------------------------
skewness of  ExitRates  is: 0.44
kurtosis of  ExitRates  is: -0.14
-----------------------------------------------
skewness of  PageValues  is: 2.6
kurtosis of  PageValues  is: 6.6
-----------------------------------------------
```

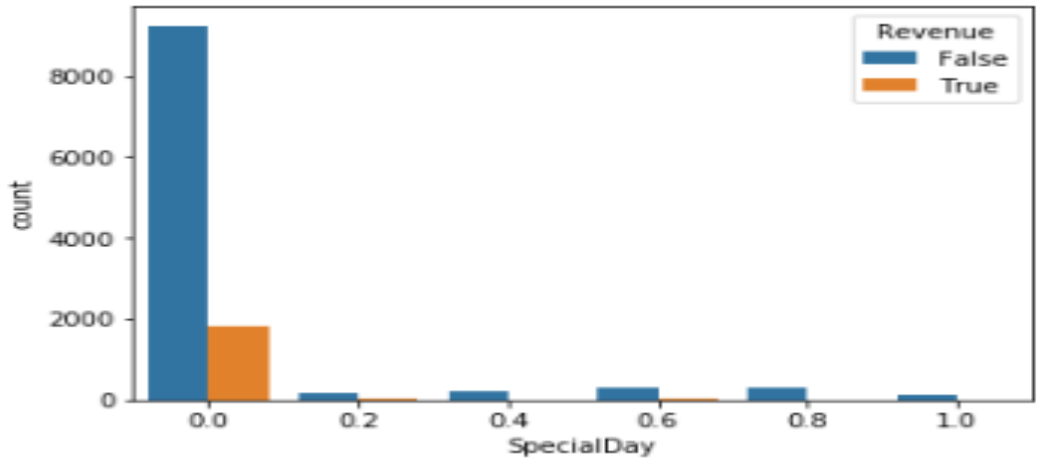
The data looks highly skewed. So, performed log transformations.

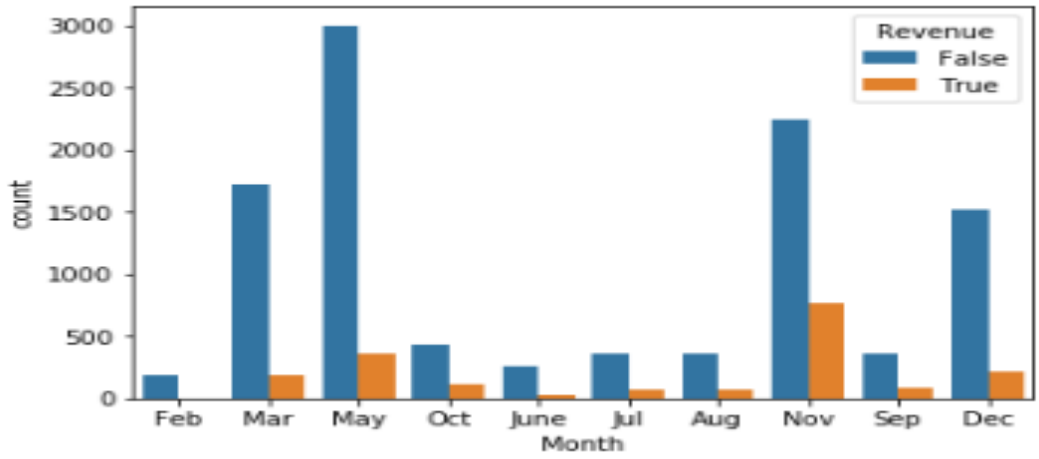
Outlier check: Quantile view

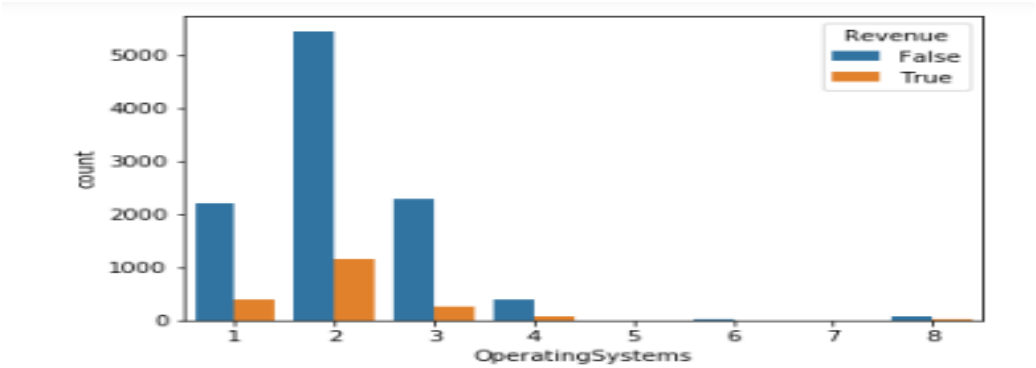
The categorical columns taken for EDA are:

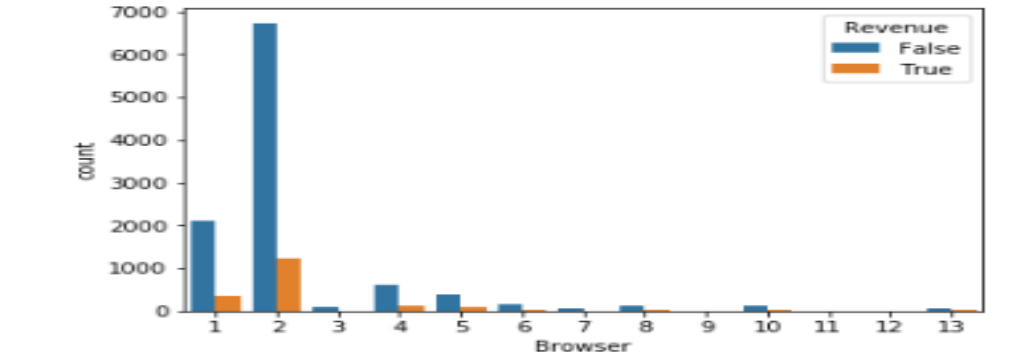
1. SpecialDay
2. Month
3. Operating Systems
4. Browser
5. Region
6. TrafficType
7. VisitorType
8. Weekend
9. Revenue

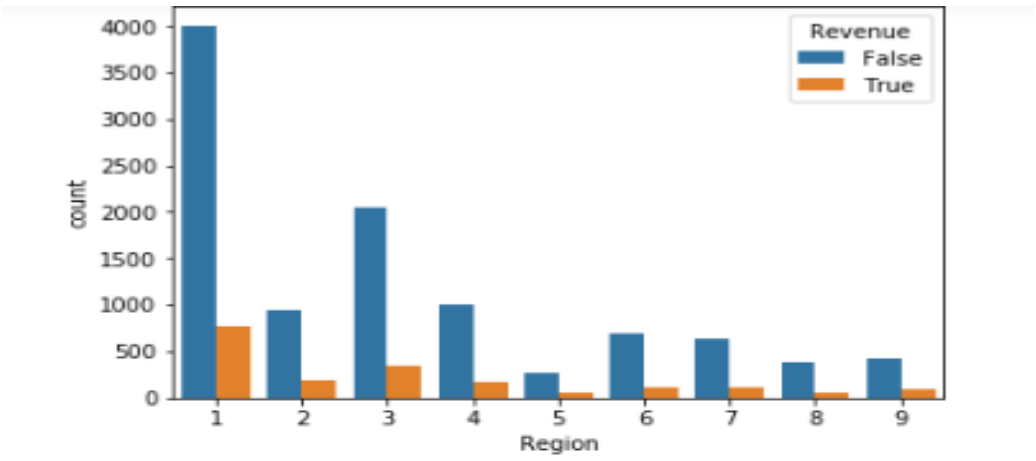
Categorical values EDA:

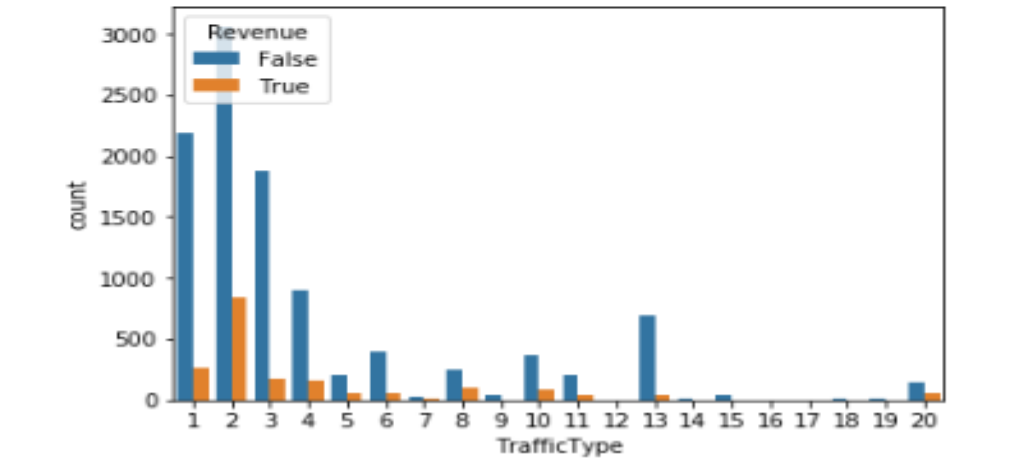


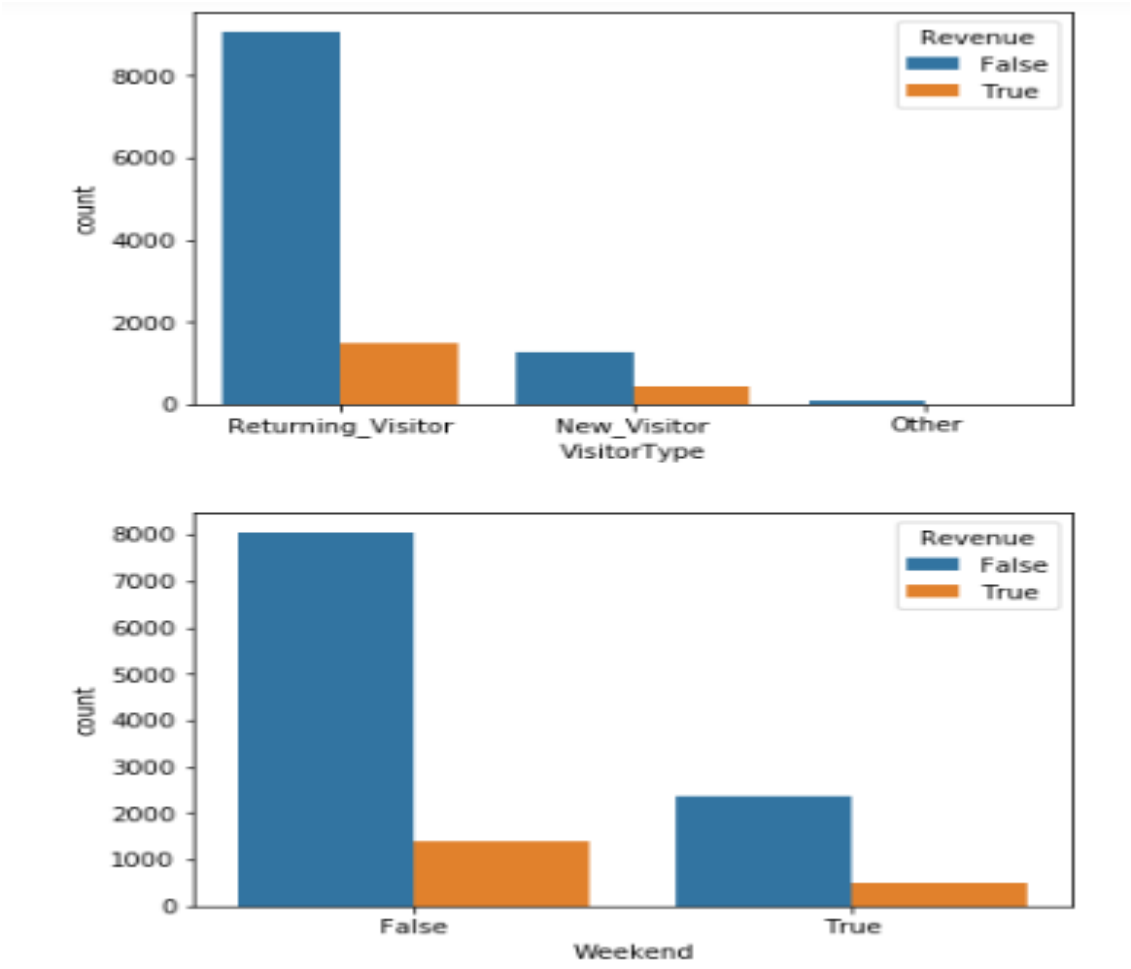



Revenue
False
True
count
OperatingSystems


Revenue
False
True
count
Browser


Revenue
False
True
count
Region

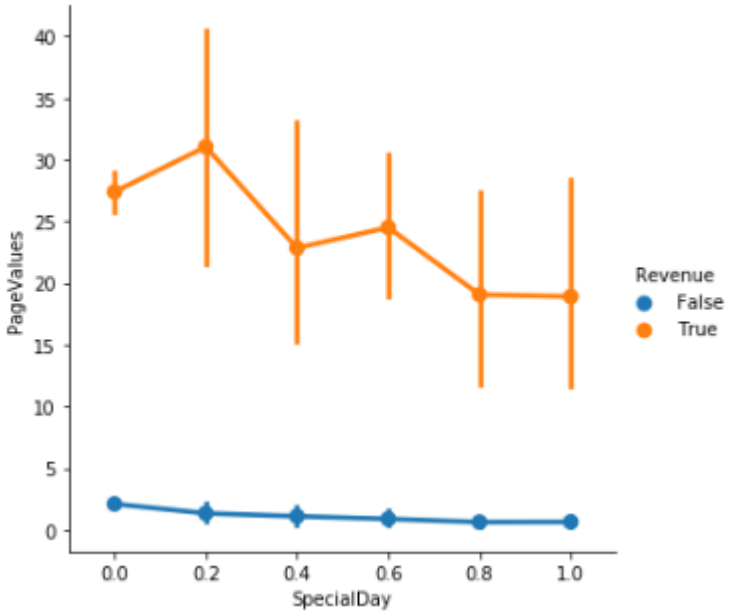

Revenue
False
True
count
TrafficType



Insights:

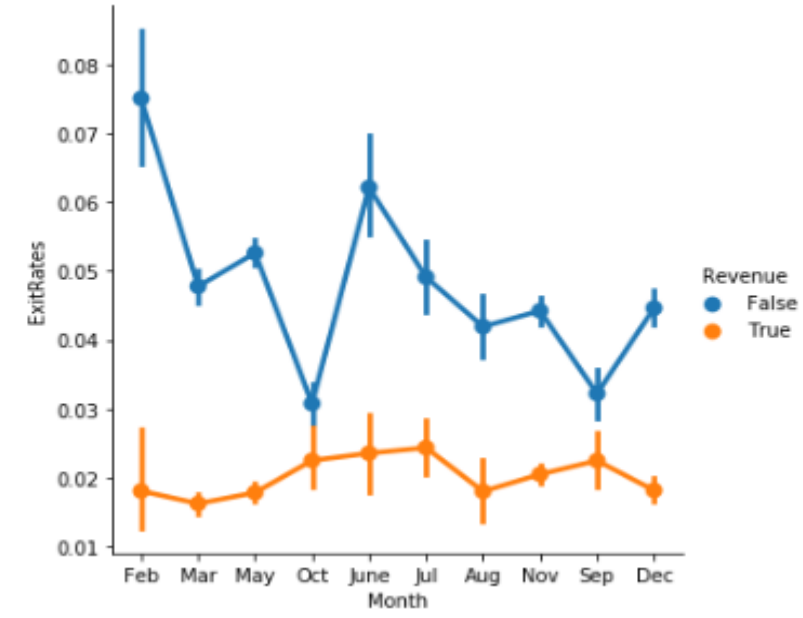
1. Interestingly the revenue is achieved on weekdays is more than on a weekend
2. The inflow of new visitors is very less. The Business needs to focus on getting more customers to the website.
3. Operating System 2 is somehow bringing more revenue.

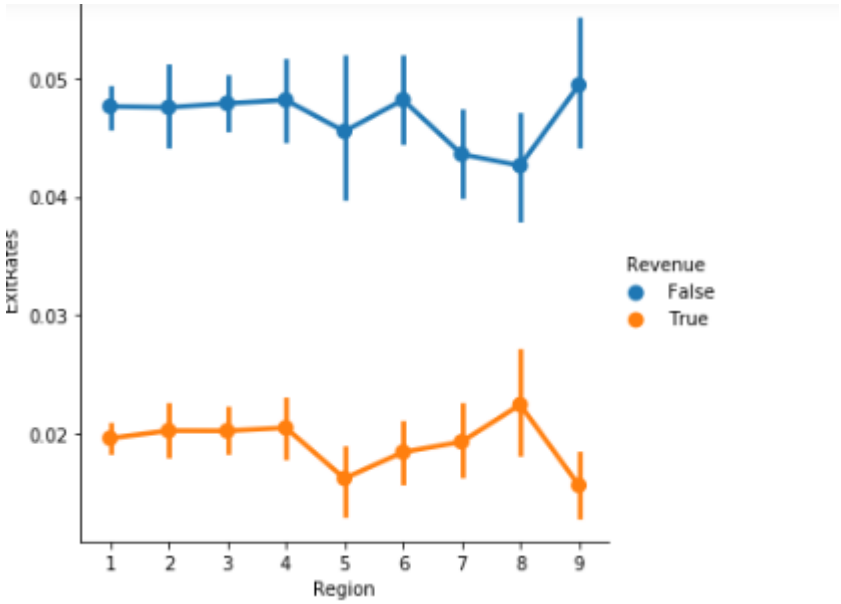
We did a factor plot and count plot on these columns for getting an idea.



Insight:

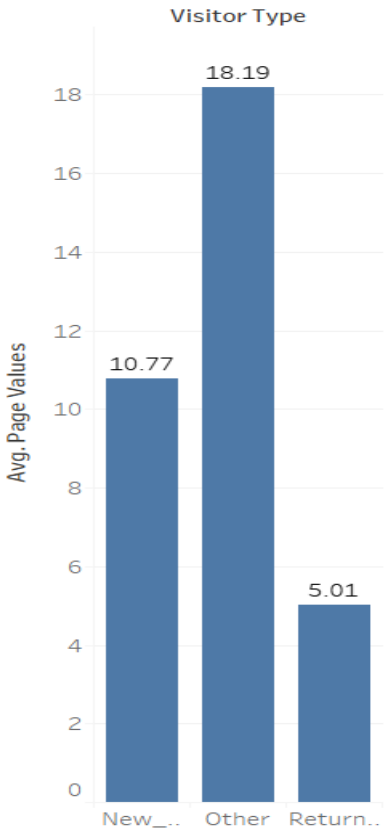
From above figure, it is clear that on special days, the page values are high and there is more scope to achieve revenue if some discounts are offered on the products using recommendation systems.





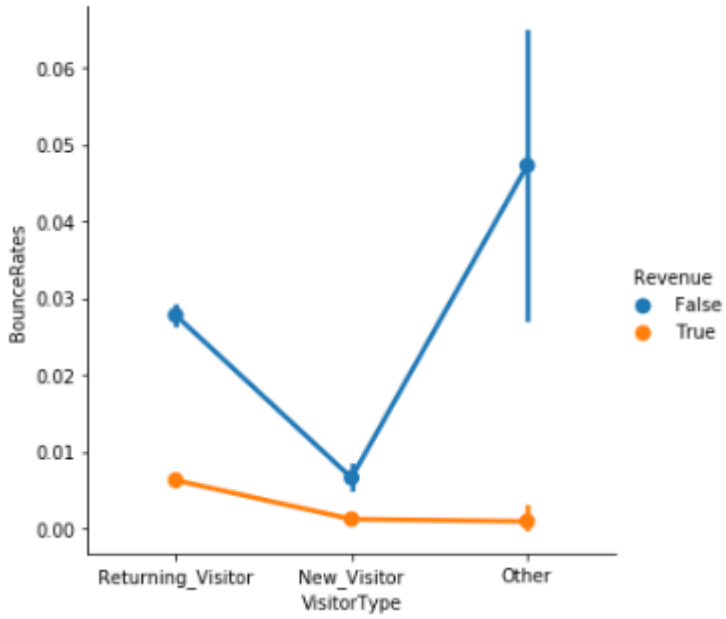
Insight:

From the above plot, The Exit rates of the customers are high no matter what month or what region it is. Which means that the business is in loss. It has to reduce the exit rates of the customers to achieve more revenue.



If we have a look at the above obtained bar graph, it is clear that avg. page values for the 'other' category of visitors are high while that of the returning visitors is least. What is surprising is that Avg. page value and Revenue are directly proportional but the above graph is showing otherwise.

One more insight from cater plot:



Insight:

It is clear from the above figure that the Bounce rate is high in 'Other' Visitors and in most of the Returning Visitors. There is scope to improve visual attractions of the app by offers or by other advertisement mechanisms.

Methods and Outputs:

We have used two methods to cross validate the model performance.

1. 70/30(80/20) split model
2. K-fold cross validation(k=3,10)

1) 70/30(80/20) split method

We have mainly used this model to get classification report of the model. Because sometimes model accuracy will not tell the complete story about the model, especially when data is heavily imbalanced like our data predicting the revenue as zero (not buying) will be very high which will influence the model accuracy even though predicting precision/recall of the buyers is very low. Therefore, to check the classification report we have used this technique. In this method we have used stratify to reduce the further imbalance of the data.

Split the data:

The data is split into 70:30 split.

The shape of dataset looks like below after the split:

Shape of x_train : (8631, 26)

Shape of y_train : (8631,)

Shape of x_test : (3699, 26)

Shape of y_test : (3699,)

2) Cross Validation:

   Performed K-fold cross validation for 10 folds. We have even used this method for all the models to check model performance and stability of the model with respect to different splits.

Feature selection: - To reduce the variance error and to reduce the dimensionality we have opted for feature selection using random forest.

We have considered all the features with feature importance of 0.01 and above.

Important features are -

Administrative

Administrative_Duration

Informational

Informational_Duration

ProductRelated

ProductRelated_Duration

BounceRates

ExitRates

PageValues

OperatingSystems

Browser

Region

TrafficType

Weekend

Month_Nov

VisitorType_Returning_Visitor

Basic Predictive models applied:

1. Logistic Regression
   i) Threshold = 0.5
   ii) Threshold = 0.3
2. K-nn model
3. Decision Tree

Ensemble models applied:

1. Random forest model
2. Gradient Boosting

Note : - From the business point of view getting better recall for '1' and at the same time maintaining F1 score and other values like precision and recall is important.

1) Logistic Regression: - We have applied logistic regression model to the data using default 0.5 threshold. But we got very bad results for recall of '1'

	Precision	recall	f1-score	support
0	0.90	0.97	0.94	2116
1	0.68	0.39	0.48	350
avg / total	0.87	0.89	0.87	2466

Then we have changed the threshold value to 0.3. We got very good results compared to previous results. Still these were bad. As we have talked before these are getting affected by outliers and multi collinearity.

	Precision	recall	f1-score	support
0	0.92	0.96	0.94	3127
1	0.72	0.56	0.63	572
micro avg	0.90	0.90	0.90	3699
macro avg	0.82	0.76	0.79	3699
weighted avg	0.89	0.90	0.89	3699

2) K-nn model: - We have applied K-nn model but due to too many outliers, even this model is performing badly. As we know K-nn model works on the distance, so we scaled the data then applied K-nn model. However, there are too many outliers, which affects the model. The highest recall we got for 1 is 51%.

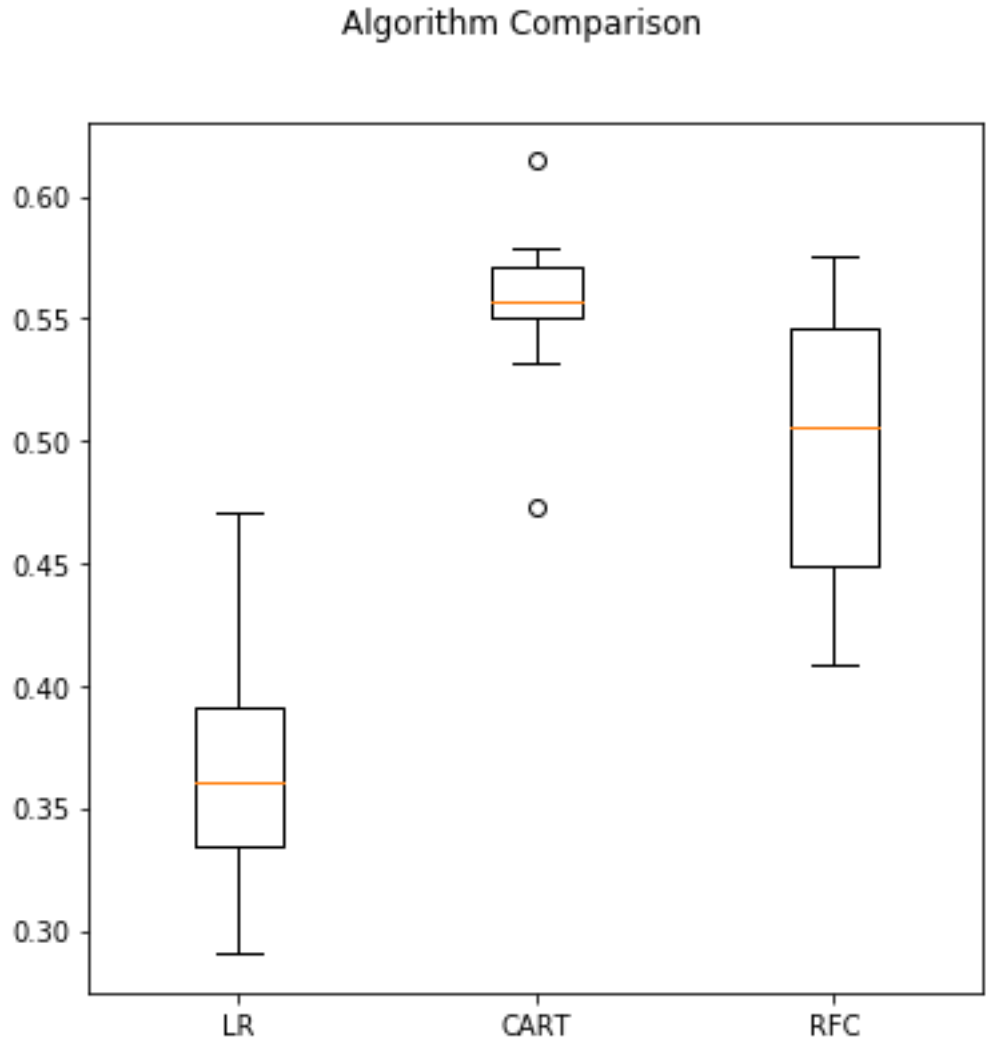
3) Decision Tree Model (CART): - Decision tree model can handle outliers but still outliers will affect it as it divides continuous variables on mean instead of median. However, we got much better results for decision tree than other predictive models.

   Basic model: -

	precision	recall	f1-score	support
0	0.92	0.97	0.95	3127
1	0.78	0.59	0.65	572
micro avg	0.91	0.91	0.91	3699
macro avg	0.85	0.77	0.80	3699
weighted avg	0.90	0.91	0.90	3699

After applying hyper-parameters, we have achieved higher recall for '1' (62%).

Base Model Check:



Insight:

1. From the above base model check with respect to the recall score, the decision tree algorithm was performing better than all other model. Therefore, we used hyper-parameters to get further better results.
2. Therefore, the idea is to apply ensemble techniques on these two models to improvise the model further.

Ensemble models applied:

1. Random Forest modelling: - Random forest model can handle outliers, multi co-linearity. As this model, works on bootstrapping technique over-fitting of the model will not happen. Basic Random forest model was providing a recall of 62% and after hyper-parameter tuning we got a recall of around 68% and we further tried to decrease the probability threshold of the model to 0.3(30%). For this model, we got a recall of 75% and during the modelling we made sure that the accuracy and the F1 score remains same. (Both around 90%)

Basic model classification report:

	Precision	recall	f1-score	support
0	0.92	0.97	0.95	3127
1	0.75	0.62	0.65	572
micro avg	0.91	0.91	0.91	3699
macro avg	0.85	0.77	0.80	3699
weighted avg	0.90	0.91	0.90	3699

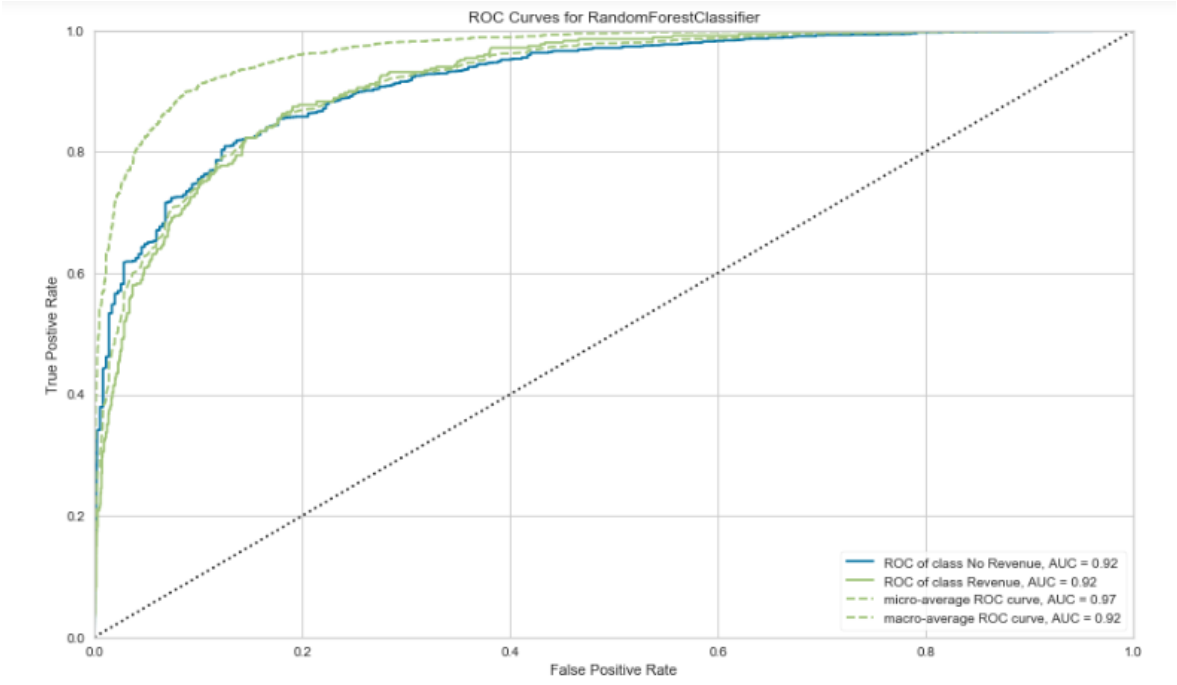
Model with P (0.3) threshold classification report:

	Precision	recall	f1-score	support
0	0.95	0.91	0.93	3127
1	0.61	0.75	0.67	572
micro avg	0.89	0.89	0.89	3699
macro avg	0.78	0.83	0.80	3699
weighted avg	0.90	0.89	0.89	3699

2. Gradient Boosting: - we have applied gradient boosting (type of boosting technique). Even though we expected it to work better than the random forest for the basic model (as it corrects the error in every next model). It was only a bit better than the random forest model. We did not go for the hype-parameter tuning of this model, as grid search was taking lot of time. This model works sequentially so it will take lot of time.

Graphical representation of ROC-AUC of the models.

ROC-AUC:

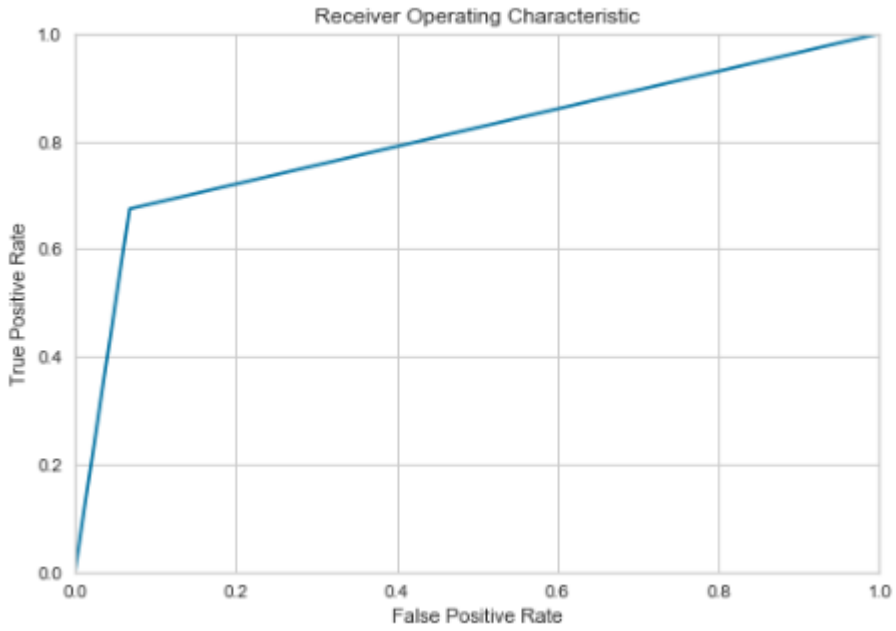


Insight:

1. The AUC metric of the curve is 0.92 with the following hyper parameters.(criterion='gini',max_depth=15,max_features=12,min_samples_leaf=12,min_samples_split=7)

2. There is a 92% chance that your model will classify the positive and negative classes correctly. However, all we need is to improve the recall of class 1 than that of the accuracy or precision of class 1.

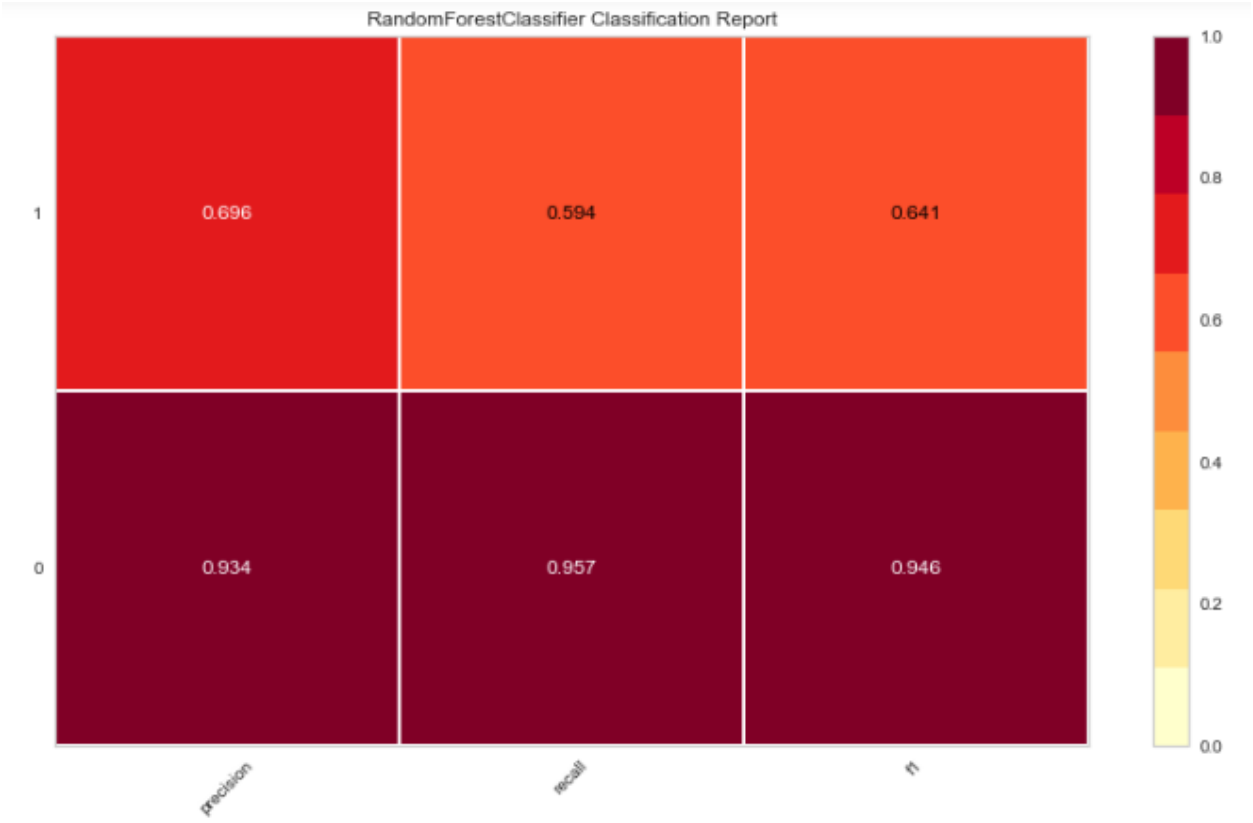
ROC-AUC after threshold change to 0.3:



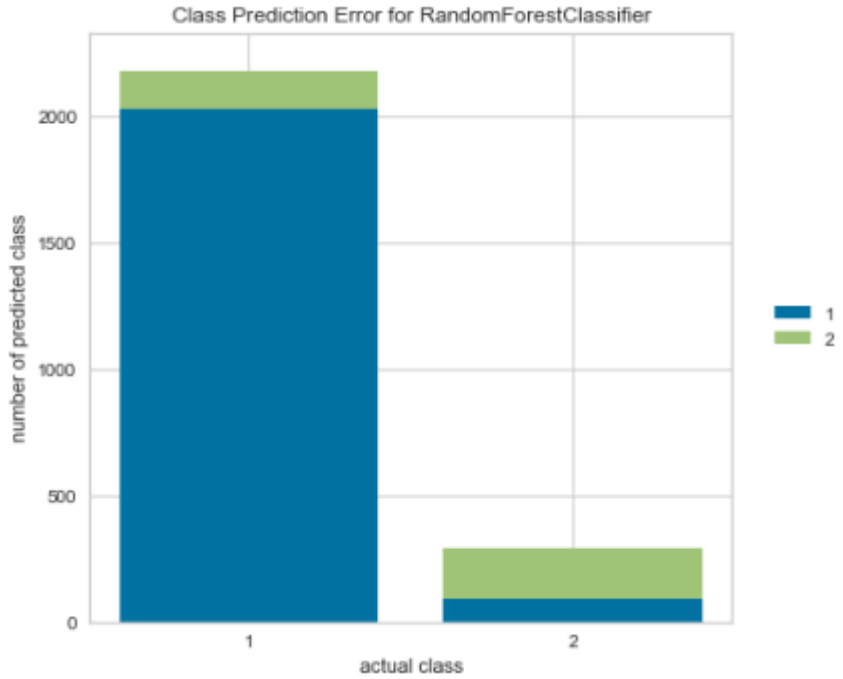
Insights:

1. The model's AUC after modifying the threshold to 0.3 is 0.89
2. There is an 89% chance that your model will classify the positive and negative classes correctly.
3. However, the recall achieved with this model is better than the prior model. Recall for this model is 75% with almost same accuracy and f1-score

Classification Report Before Threshold settings:



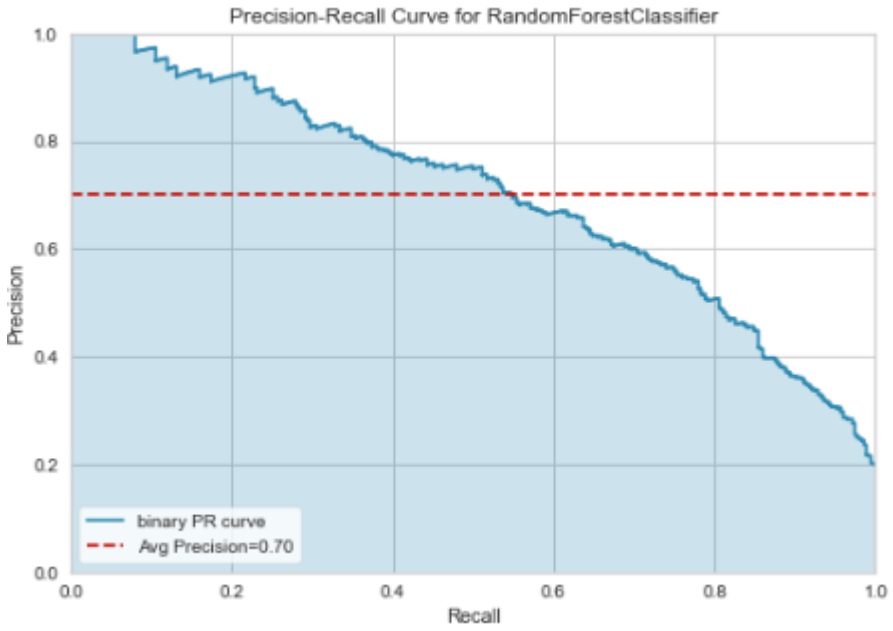
Class Prediction Error graph:



Insight:

1. The Class Prediction Error is good with respect to the target class, which means that the recall of the target class has improved when compared to the previous model.

The Precision Recall Trade-off:



Insight:

1. The Recall achieved is around 70% after the threshold setting

Conclusion:

The purpose of this project is to examine consumer behavior pattern toward online shopping and the possibility that the customer will buy the product. After analyzing the data, we have given insights and the pattern of customer behavior towards online shopping. In addition, we have used predictive models to predict the class (generates revenue or not). In addition, we got an accuracy of 91%, f1 score of 90%.

References:

- https://www.researchgate.net/publication/324694092_EFFECTS_OF_ONLINE_SHOPPING_ON_CONSUMER_BUYING_BEHAVIOUR

- Chang, M., K., Cheung, W., Lai, V., S. (2005), 'Literature derived reference models for the adoption of online shopping', Information & Subject:capstone interim report Version 2.docxManagement, Vol.42(4), pp.543-559.

- E-commerce Facts (2012), 'Eurostat: 43 percent of Europeans make purchases online', [online] Available at: http://www.ecommercefacts.com/research/2012/05/eurostat-eu-online-shoppi/ [accessed on 10 December 2016]