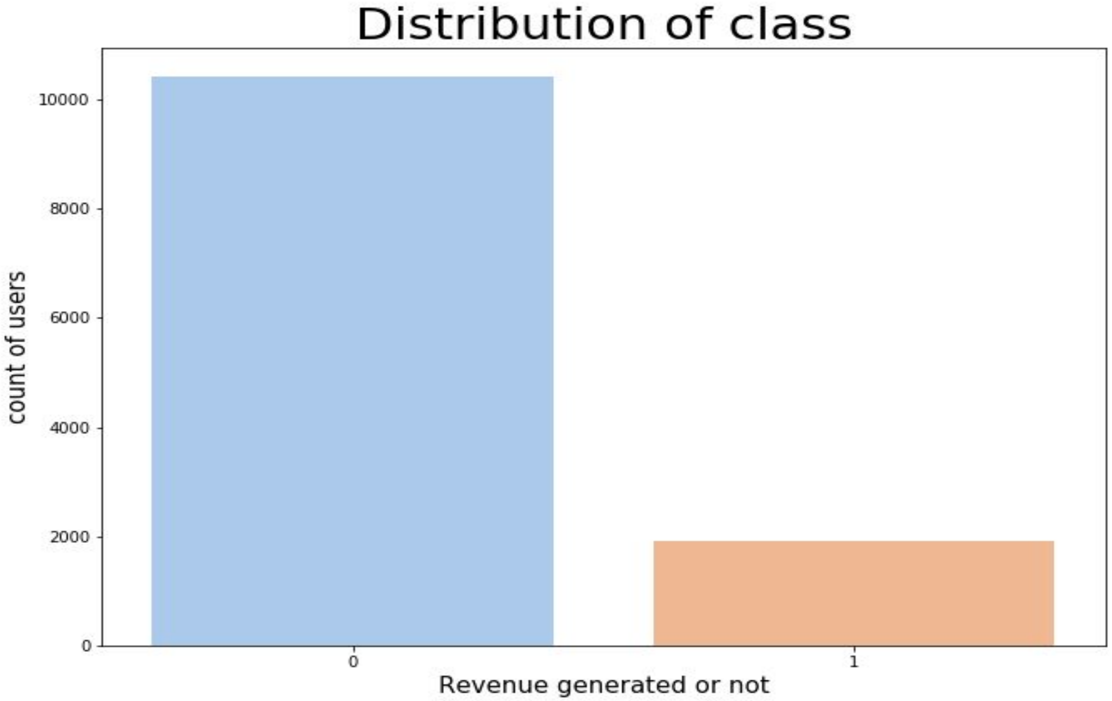
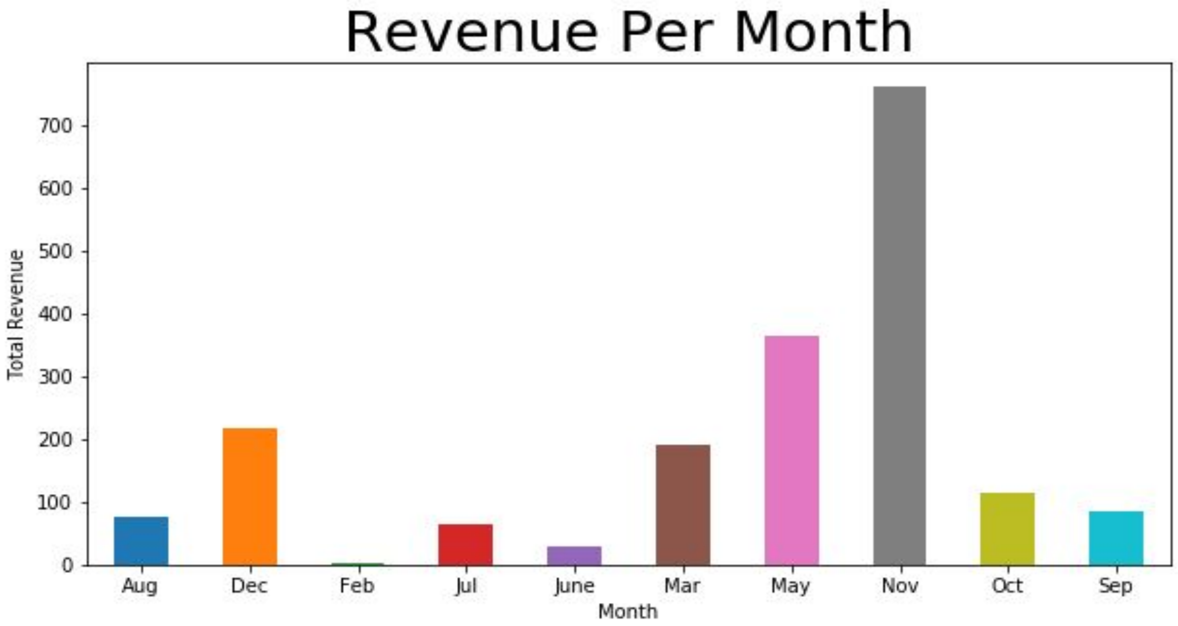
CS6220 - Data Mining Techniques

Online Shoppers Purchasing Intention - RESULTS

Rajath Kashyap
Mukund Wagh
Bishwarup Neogy

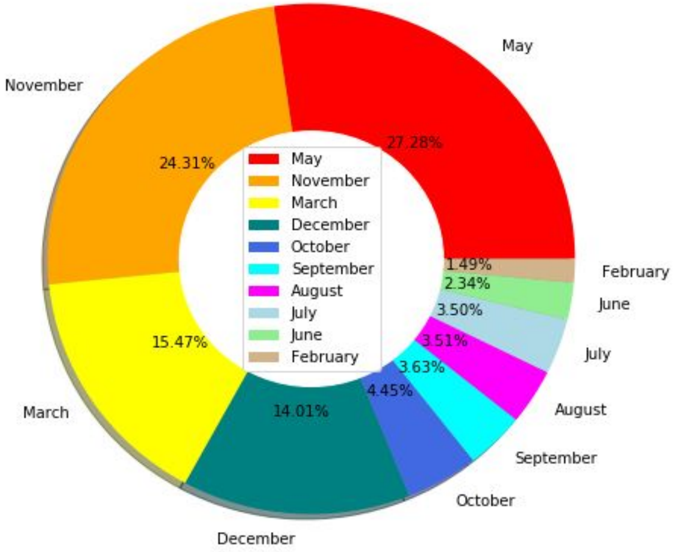


- Unbalanced dataset
- Get valuable insight from the available data.



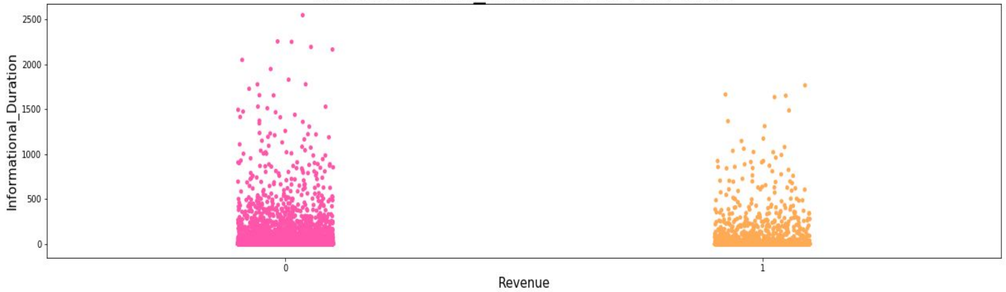
- More the number of visitors more is the sale.
- We have products that fulfill the needs of the customers.

Users per month

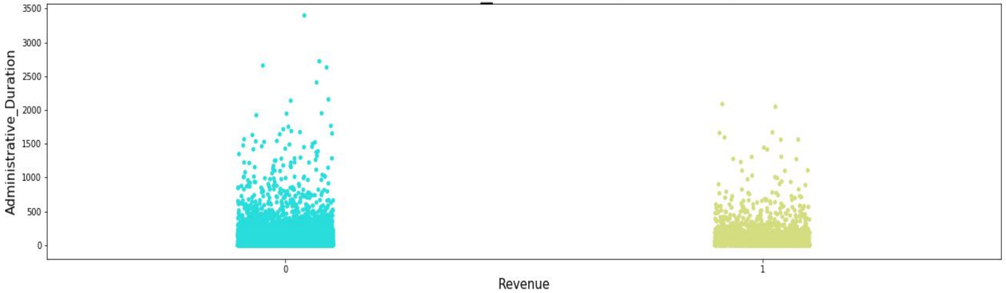


- More the number of visitors more is the sale.
- We have products that fulfill the needs of the customers.

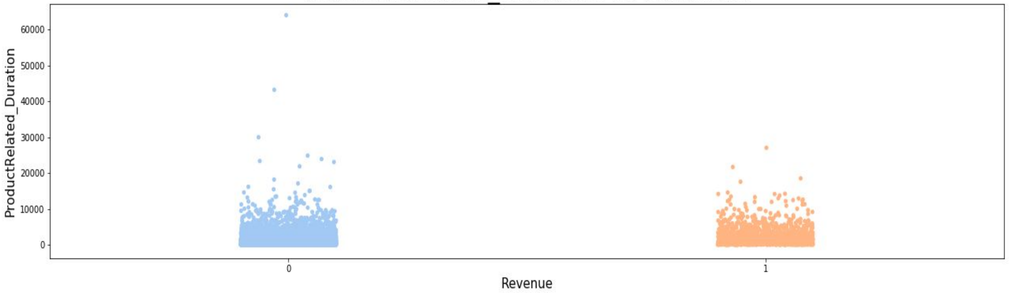
Informational_Duration vs Revenue

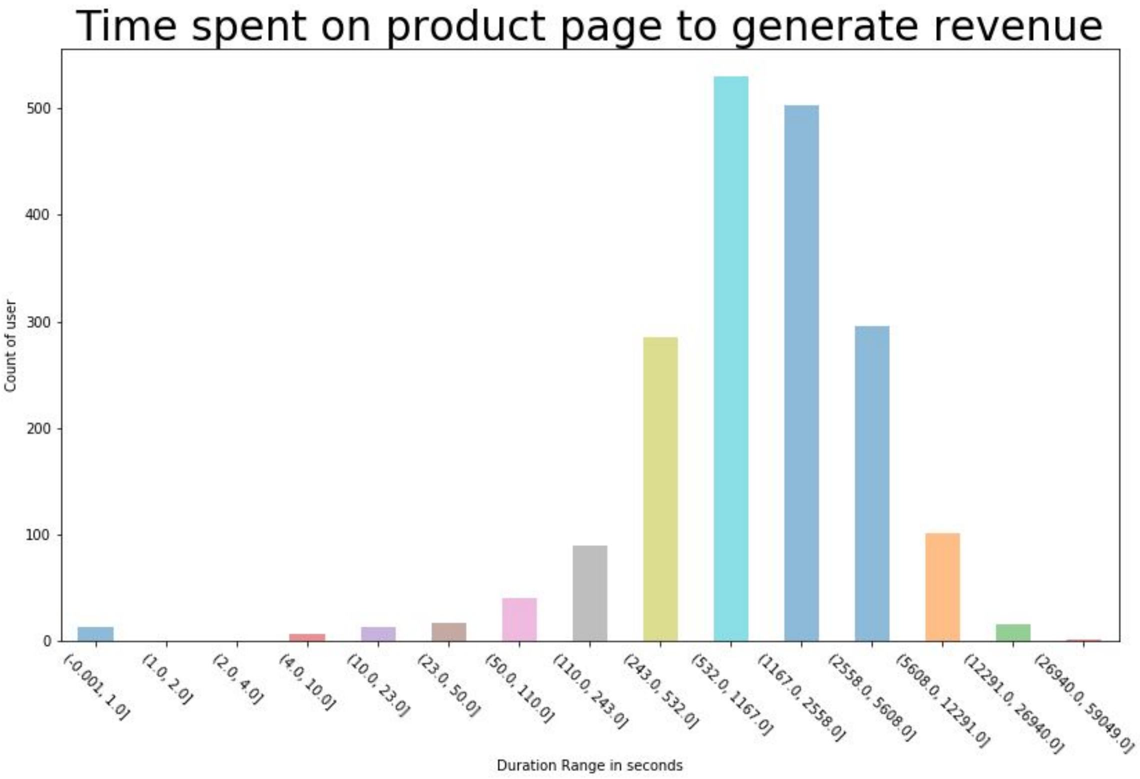


Administrative_Duration vs Revenue

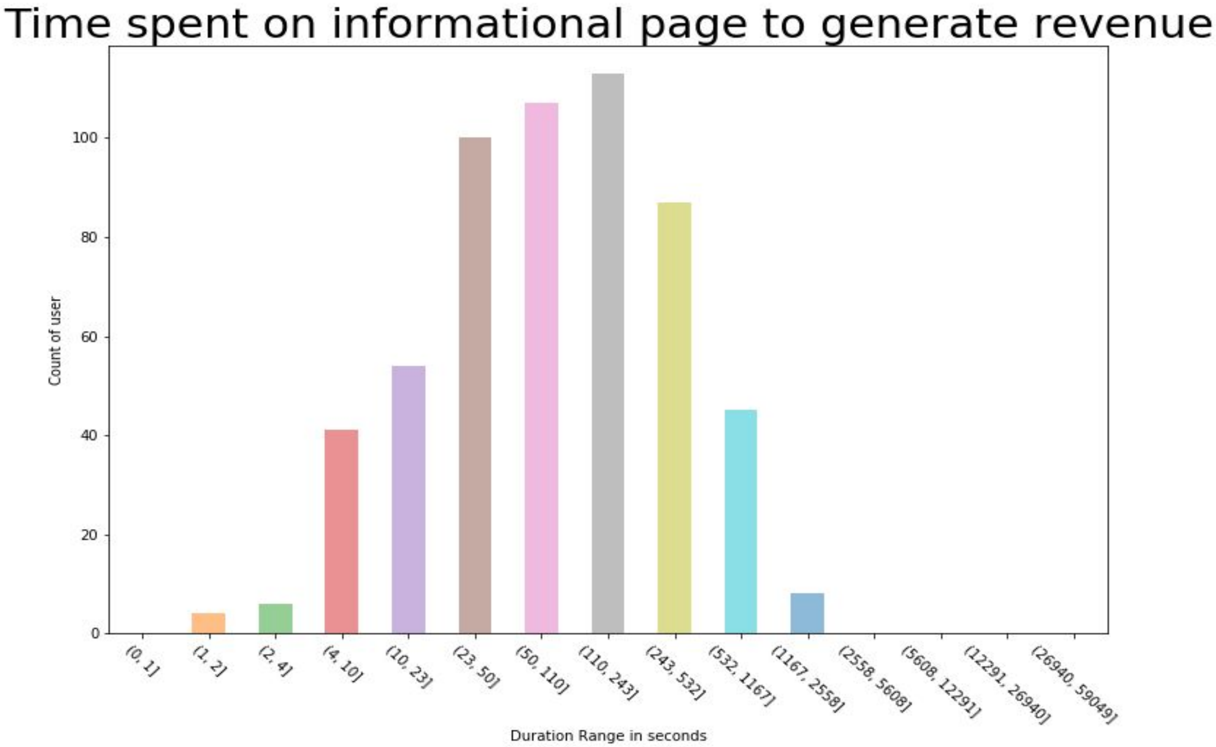


ProductRelated_Duration vs Revenue

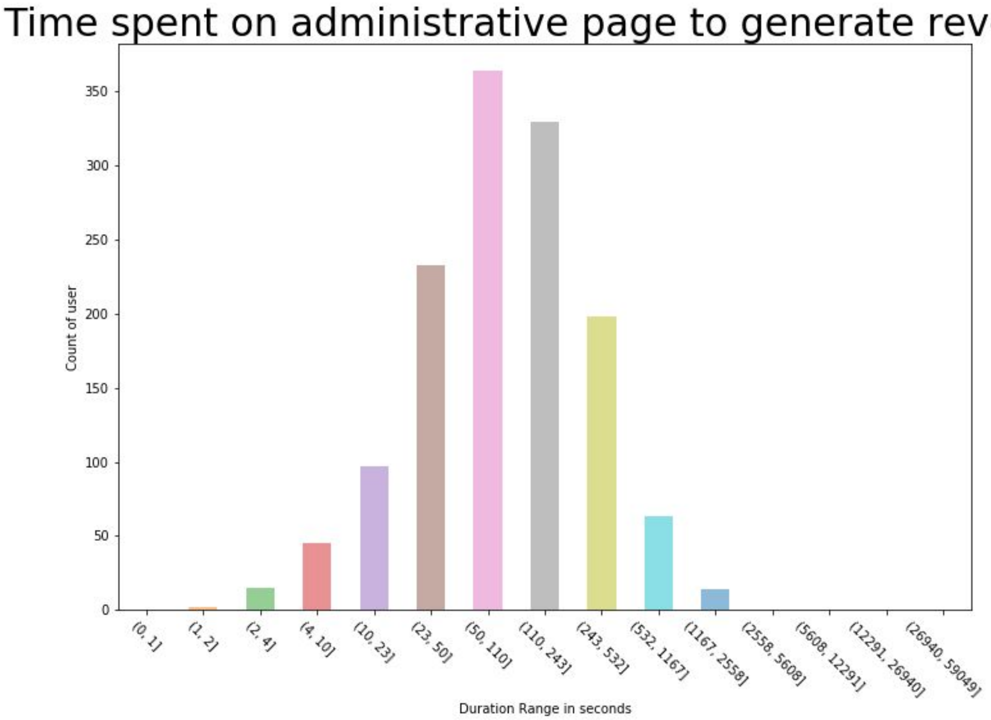




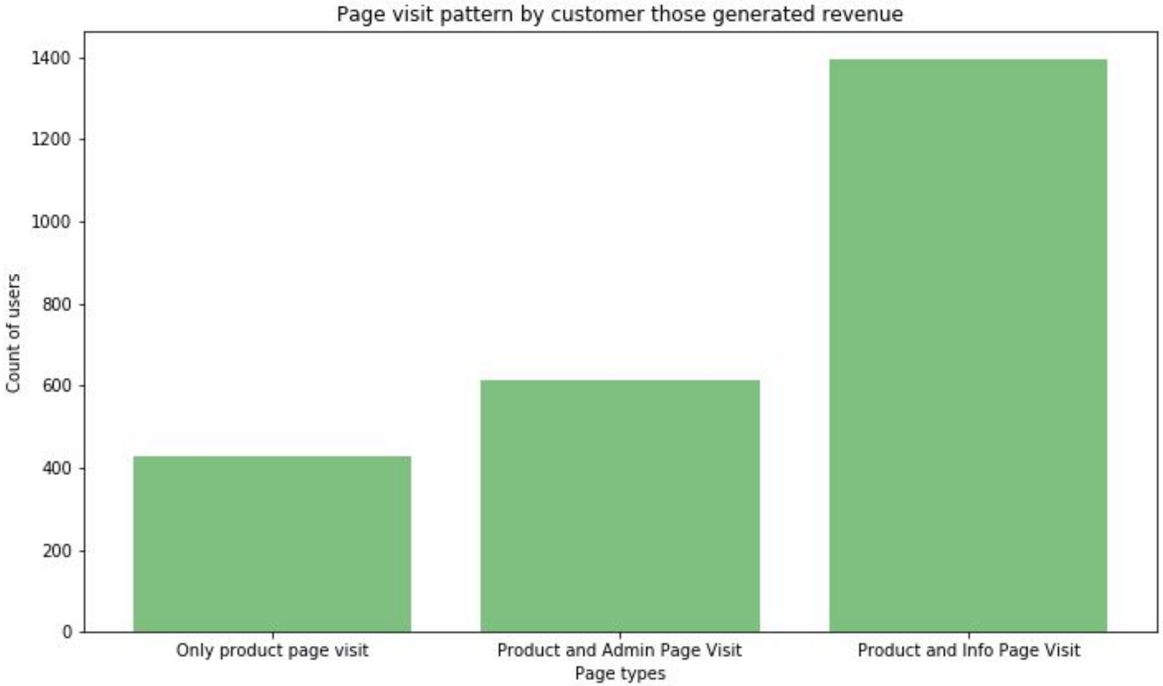
- We need to improve on the overall search engine of the website and cater them with the right product when the try to find one.



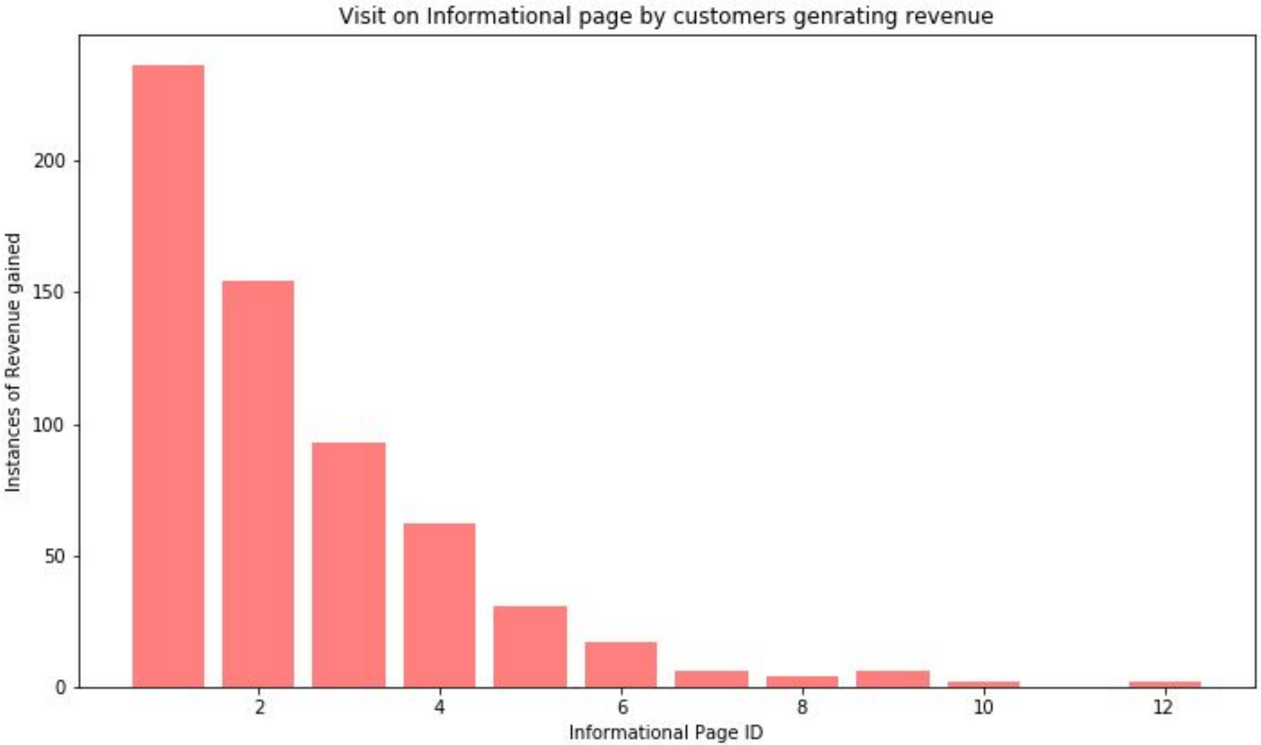
- Many users has to visit the info page to be sure of the product they are going to buy.

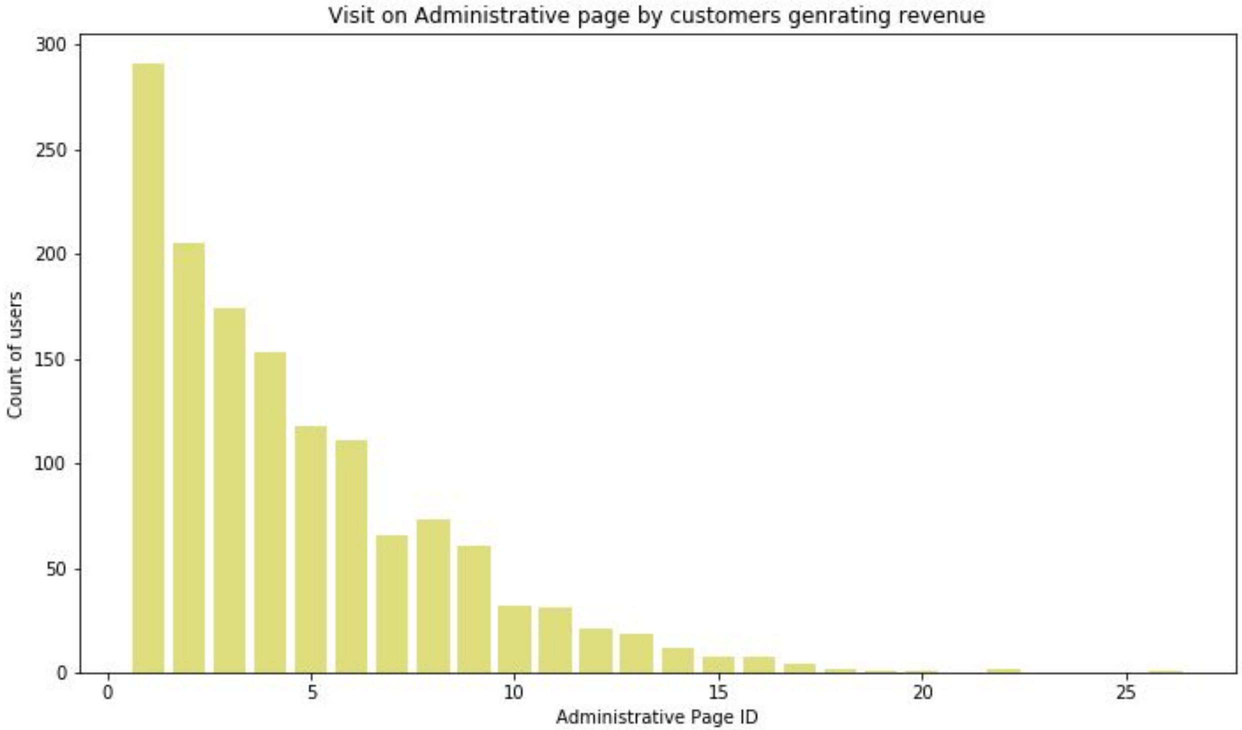

Time spent on administrative page to generate revenue

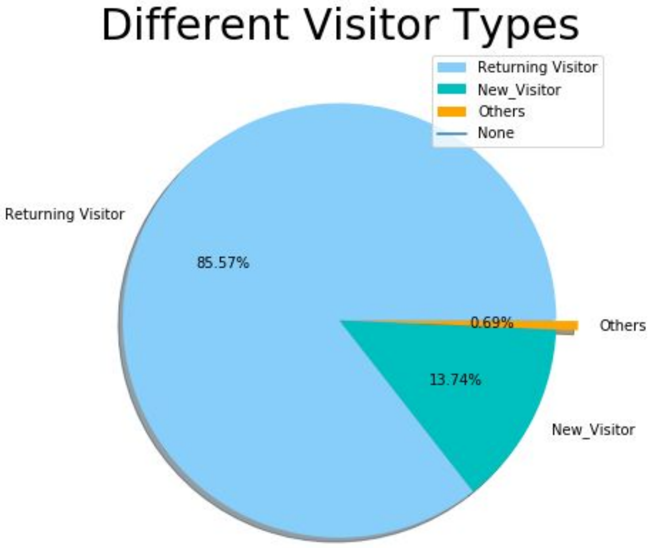
- We have around 2000 unique users who have given us the revenue, and we can see that more than 70% of the users have to visit the administrative page in order to buy the product, also around 50% customers have to spend more than a minute on the administrative pages.



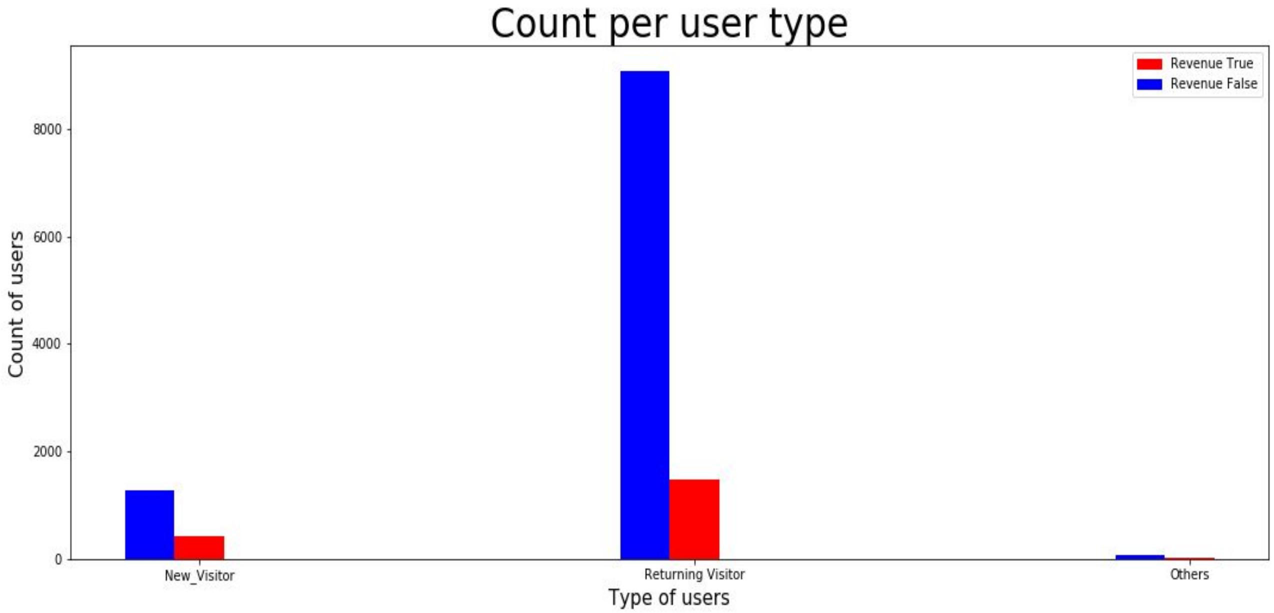
- We have data of users visiting different pages on website.
- Prioritizing the task to retain the potential customers.

Visit on Informational page by customers genrating revenue
Instances of Revenue gained
0
50
100
150
200
Informational Page ID
2
4
6
8
10
12

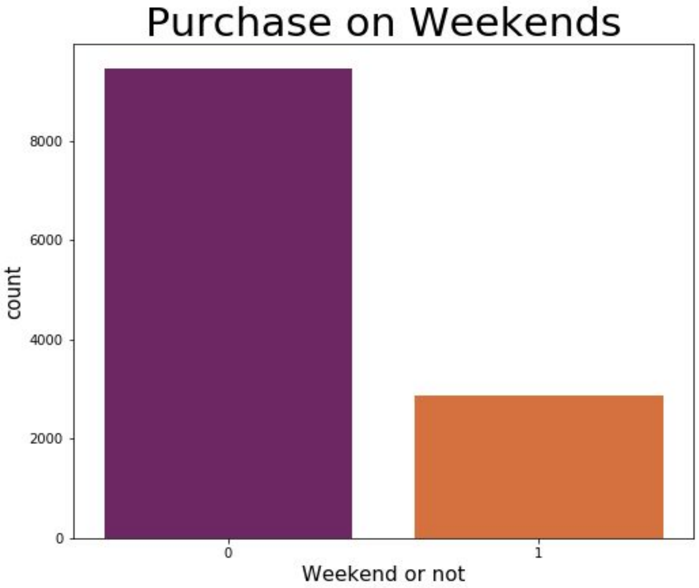

Visit on Administrative page by customers genrating revenue
Count of users
300
250
200
150
100
50
0
0
5
10
15
20
25
Administrative Page ID



- We have 3 categories of users, the new users, returning users and others.

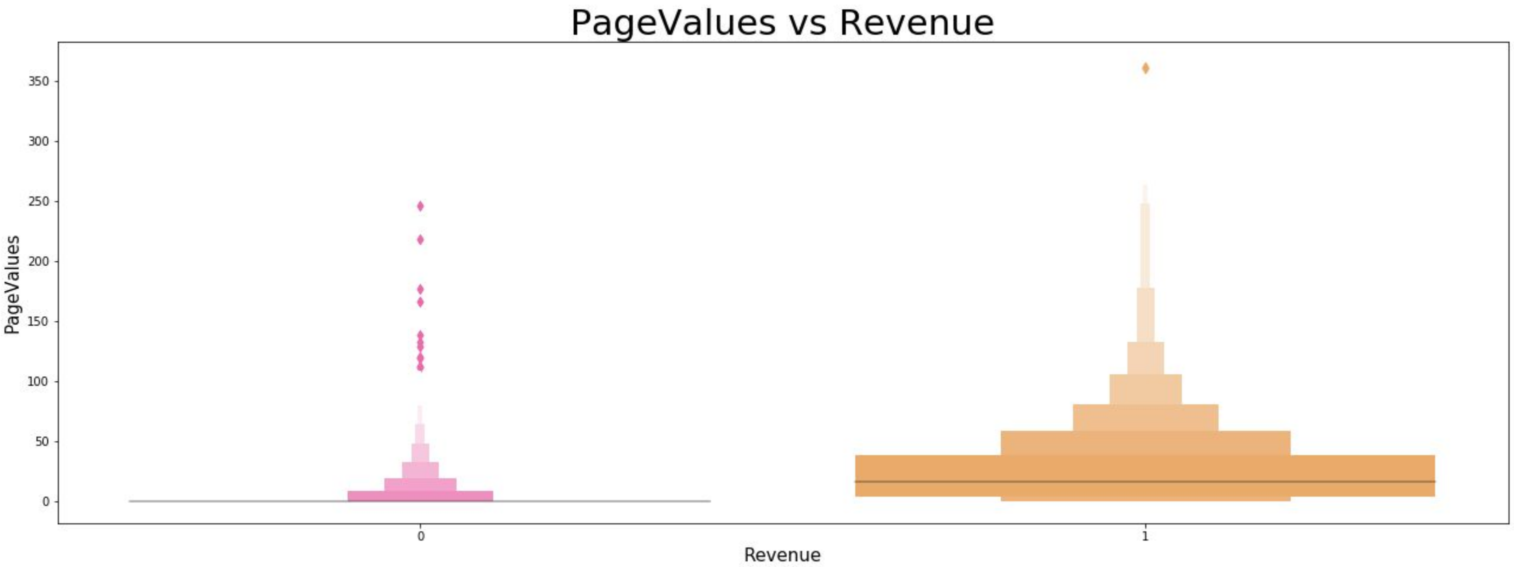


- Distribution of the user in each type of user.
- Returning users give us the most of the revenue.



- Most of the purchases are on the weekdays.
- We should come up with schemes and offers that will also attract customers on the weekends.

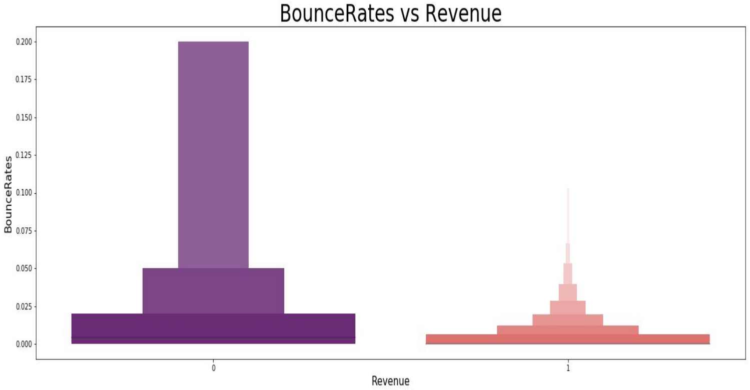
Bivariate Analysis: Page Value vs Revenue

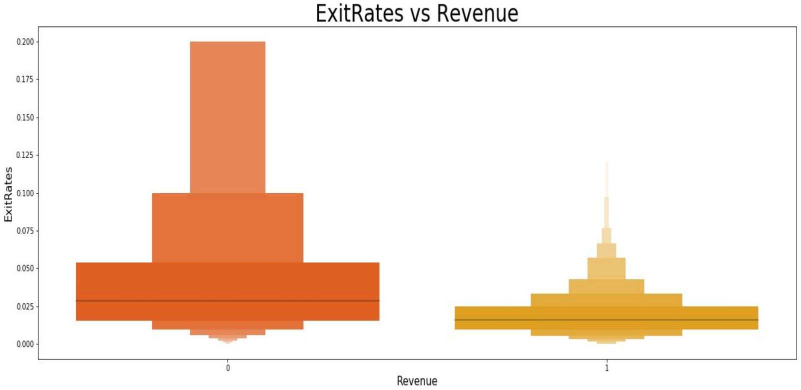


Bivariate Analysis: Bounce Rate and Exit Rate vs Revenue

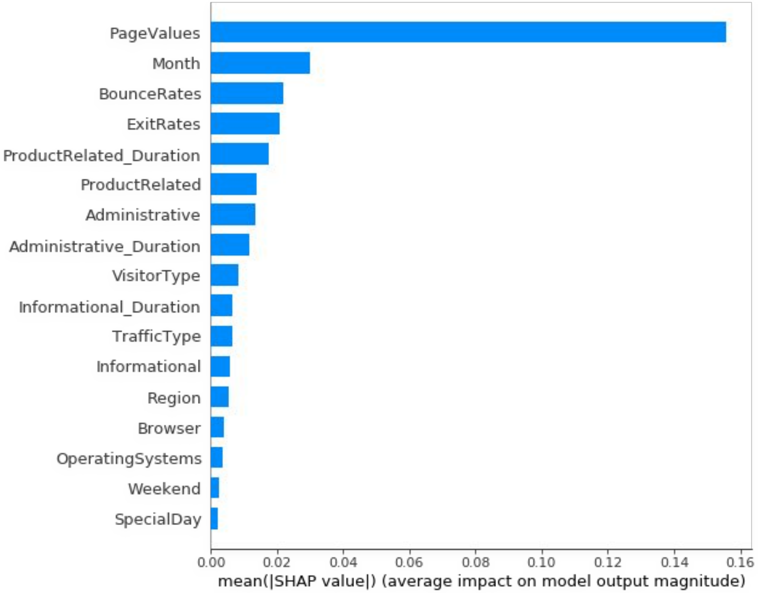
Bounce Rate : Avg time between a user opening a page on the site and exiting without triggering any other requests.

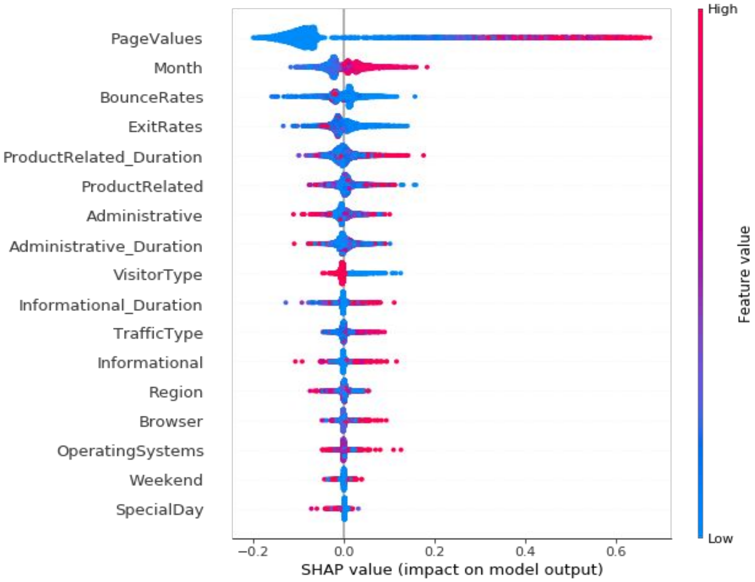
Exit Rate : Exit Rate is the percentage of users who exit the page and close out the session.





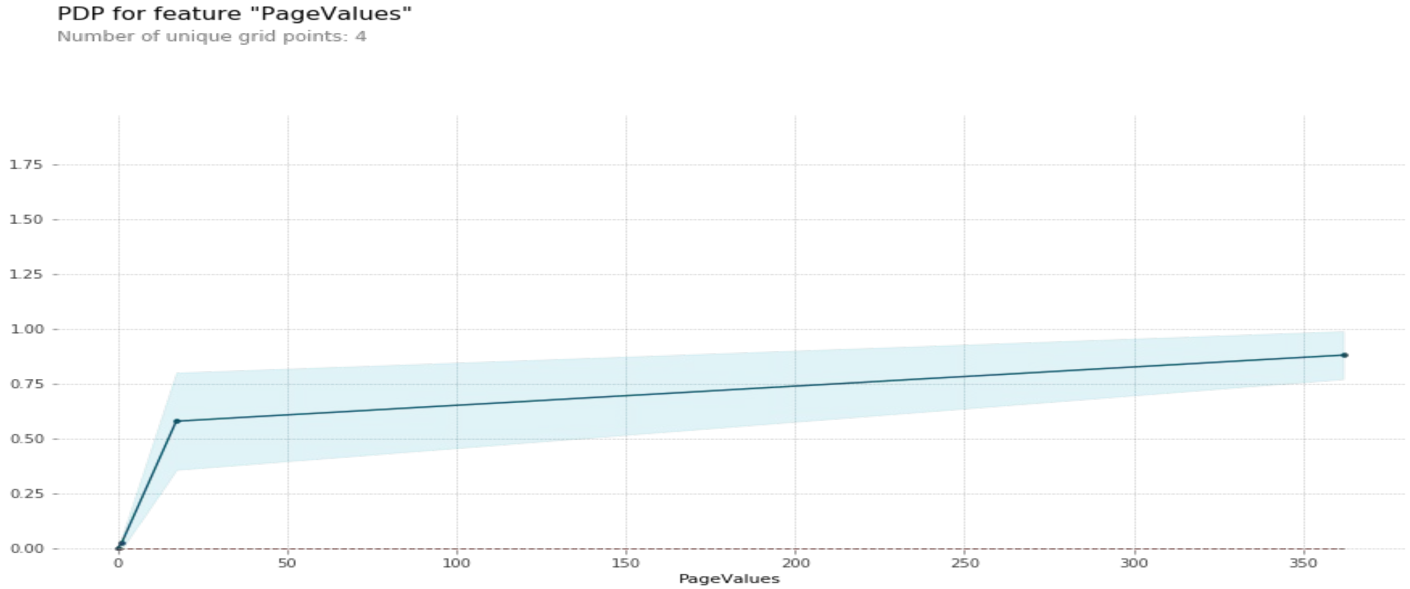
SHapley Additive exPlanation (SHAP) Analysis



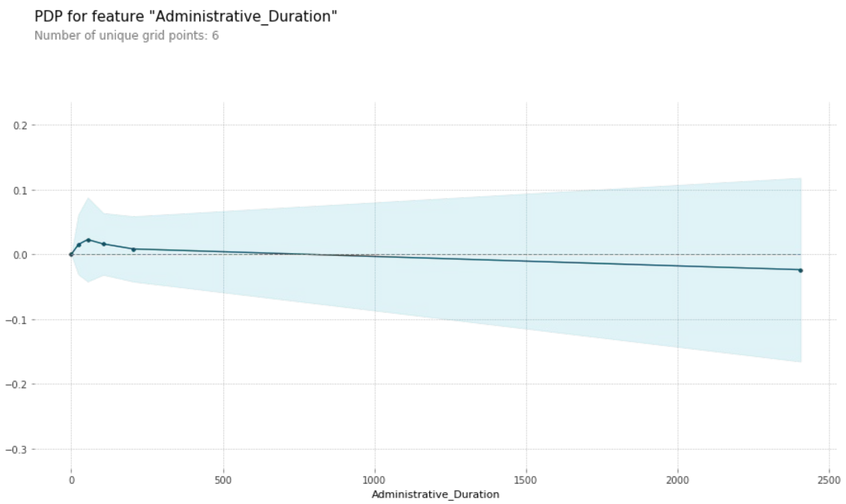


Partial Dependence Plot

The partial dependence plot (PDP or PD plot) shows the marginal effect one or two features have on the predicted outcome of a machine learning model.The plot can show whether the relationship between the target and a feature is linear, monotonic or more complex

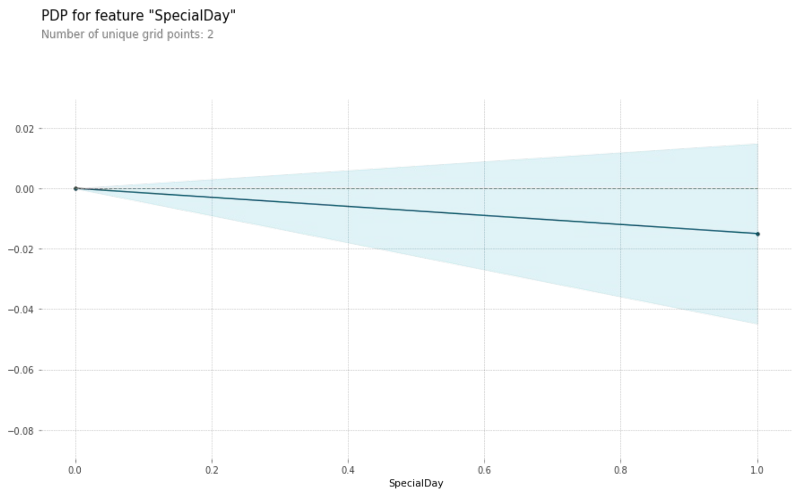


Partial Dependence Plot



PDP for feature "Administrative_Duration"

Number of unique grid points: 6



PDP for feature "SpecialDay"

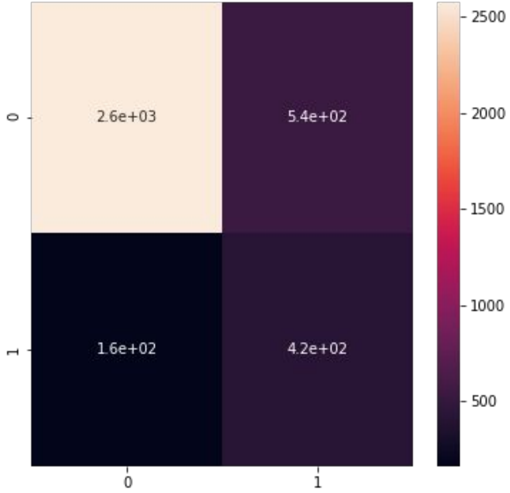
Number of unique grid points: 2

Naive Bayes

Naive Bayes Classifier is probabilistic classifier which uses Bayes' theorem with strong (naive) independence assumptions between the features

Classification Report :

	Precision	**Recall**	**f1-score**	**support**
0	0.94	0.83	0.88	3114
1	0.44	0.72	0.55	585
accuracy			0.81	3699
macro avg	0.69	0.78	0.71	3699
weighted avg	0.86	0.81	0.83	3699



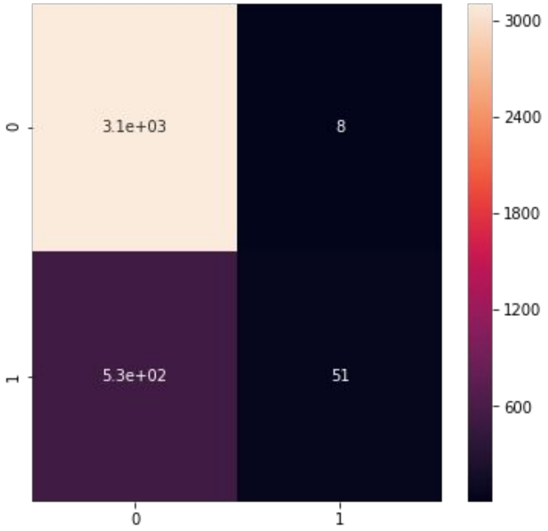
Support Vector Machine (SVM)

Supervised non-probabilistic binary classifier algorithm, when given labeled training data, outputs an optimal hyperplane which categorizes new examples.

Classification Report :

Without SMOTE	Precision	Recall	f1-score	support
0	0.85	1.00	0.92	3114
1	0.86	0.09	0.16	585
accuracy			0.85	3699
macro avg	0.86	0.54	0.54	3699
weighted avg	0.86	0.85	0.80	3699

Confusion Matrix:

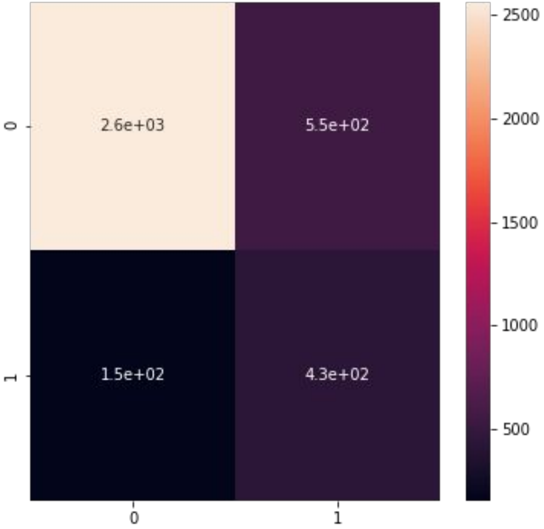


Support Vector Machine (SVM)

Classification Report :

With SMOTE	**Precision**	**Recall**	**f1-score**	**support**
0	0.94	0.82	0.88	3114
1	0.44	0.74	0.55	585
accuracy			0.81	3699
macro avg	0.69	0.78	0.71	3699
weighted avg	0.86	0.81	0.83	3699

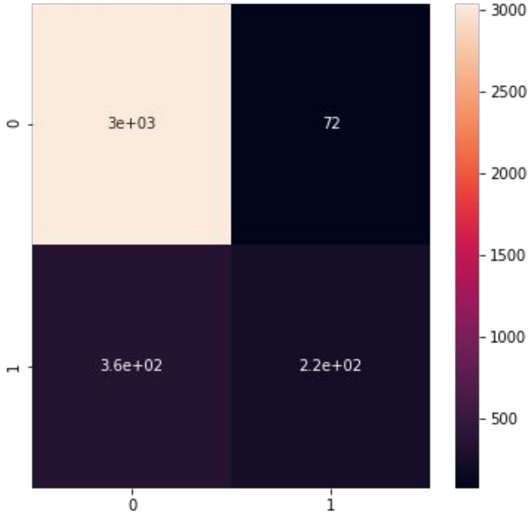
Confusion Matrix:



Logistic Regression

Classification Report :

	Precision	Recall	f1-score	support
0	0.89	0.98	0.93	3114
1	0.76	0.38	0.51	585
accuracy			0.88	3699
macro avg	0.82	0.68	0.72	3699
weighted avg	0.87	0.88	0.87	3699

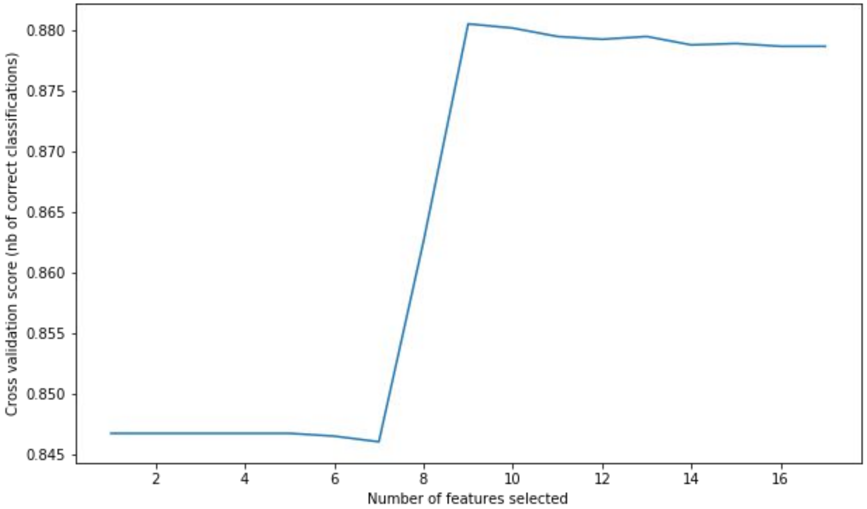


Logistic Regression

Next we reduced the dimensionality by using Recursive Feature Elimination (RFE).

With Logistic Regression as the model, RFE selected the following **9** Features :

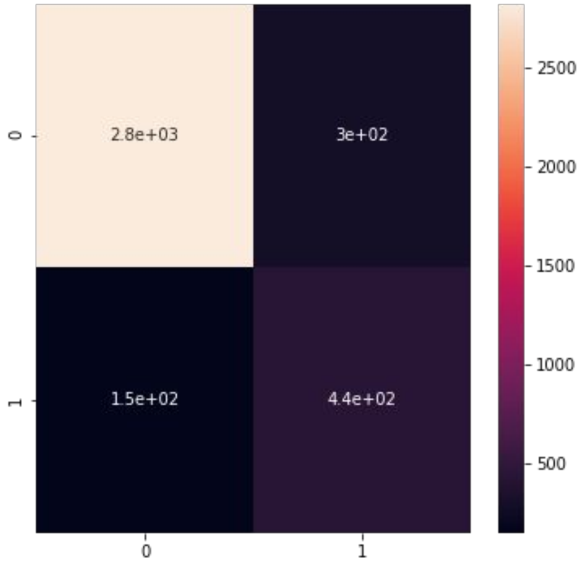
Selected features:
['Informational', 'BounceRates', 'ExitRates', 'PageValues', 'SpecialDay', 'Month', 'OperatingSystems', 'VisitorType', 'Weekend']



Logistic Regression

Classification Report : with SMOTE and RFE

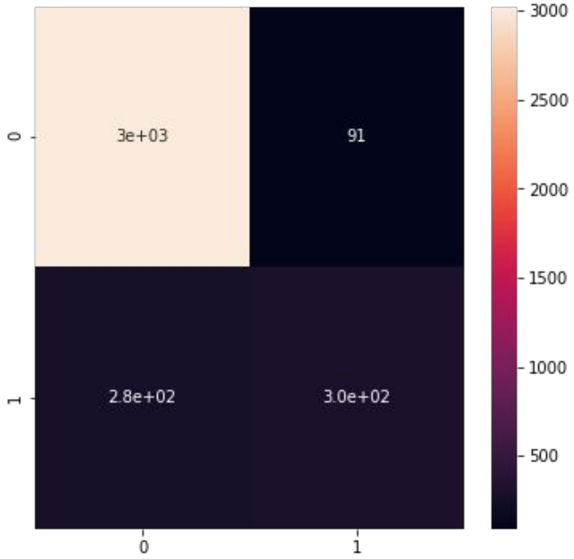
	Precision	Recall	f1-score	Support
0	0.95	0.91	0.93	3114
1	0.60	0.75	0.66	585
Accuracy			0.88	3699
Macro avg	0.77	0.83	0.79	3699
Weighted avg	0.89	0.88	0.89	3699



Random Forest Classifier

Classification Report :

	Precision	Recall	f1-score	Support
0	0.92	0.97	0.94	3114
1	0.77	0.52	0.62	585
Accuracy			0.90	3699
Macro avg	0.84	0.75	0.78	3699
Weighted avg	0.89	0.90	0.89	3699

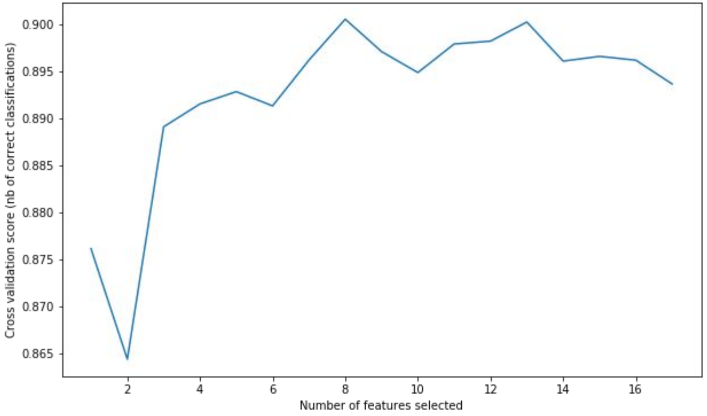


Random Forest Classifier

Next we reduced the dimensionality by using Recursive Feature Elimination (RFE).

With Random Forest Classifier as the model, RFE selected the following **12** Features :

Selected features:
['Administrative', 'Administrative_Duration', 'Informational_Duration', 'ProductRelated', 'ProductRelated_Duration', 'BounceRates', 'ExitRates', 'PageValues', 'Month', 'Browser', 'Region', 'TrafficType']

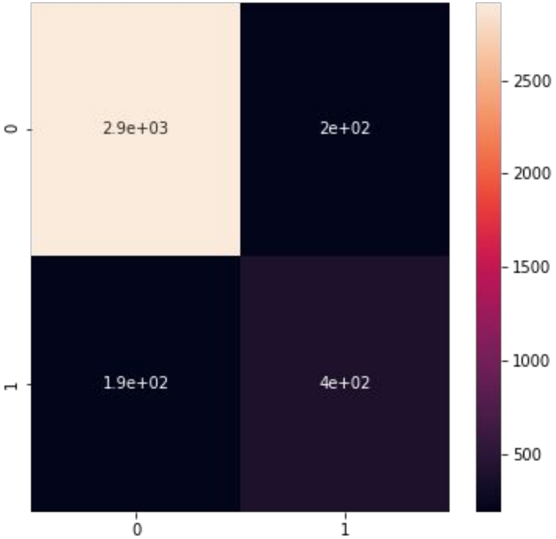


Random Forest Classifier

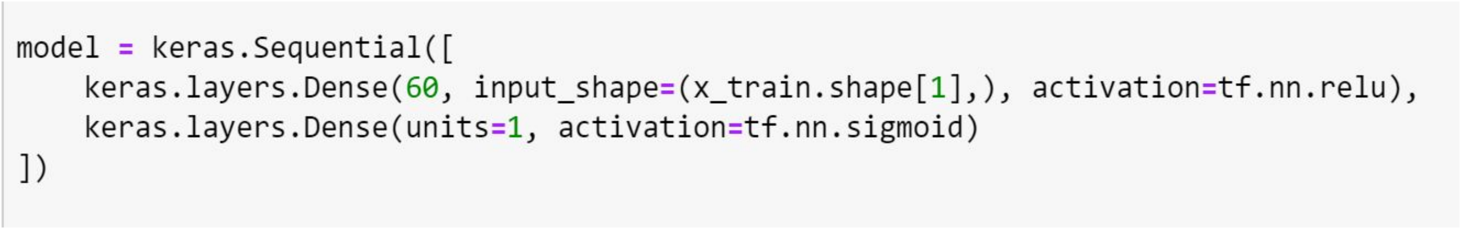
We again evaluated our classifier and obtained the following results:

Classification Report : With SMOTE and RFE

	Precision	**Recall**	**f1-score**	**Support**
0	0.94	0.94	0.94	3114
1	0.67	0.68	0.67	585
Accuracy			0.89	3699
Macro avg	0.80	0.81	0.80	3699
Weighted avg	0.90	0.90	0.90	3699



Neural Network

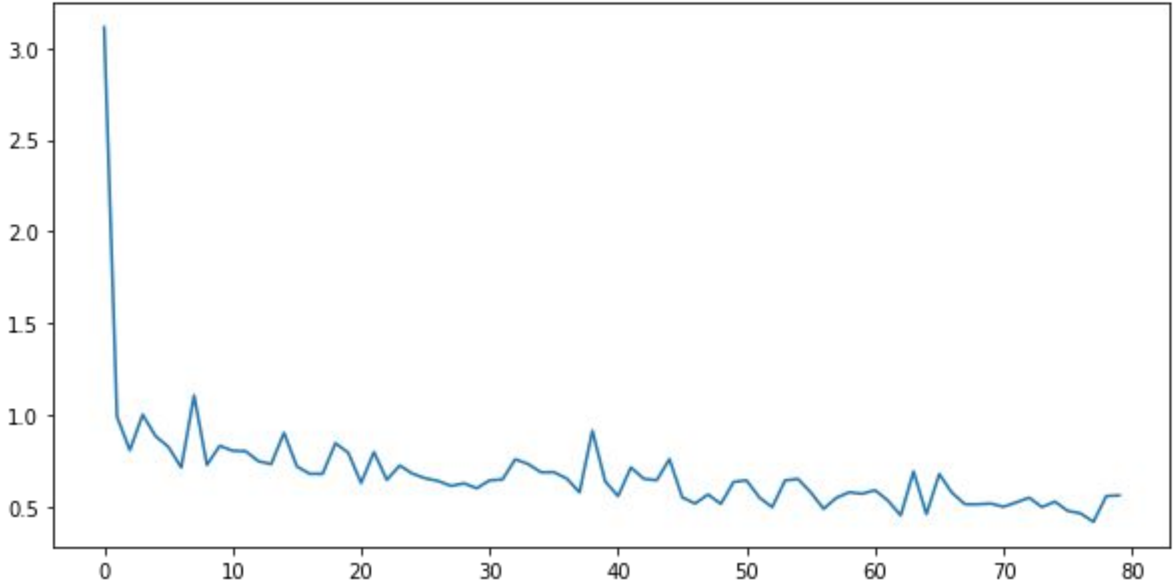


```
model = keras.Sequential([
    keras.layers.Dense(60, input_shape=(x_train.shape[1],), activation=tf.nn.relu),
    keras.layers.Dense(units=1, activation=tf.nn.sigmoid)
])
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 60)                1080
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 61
=================================================================
Total params: 1,141
Trainable params: 1,141
Non-trainable params: 0
```

Neural Network

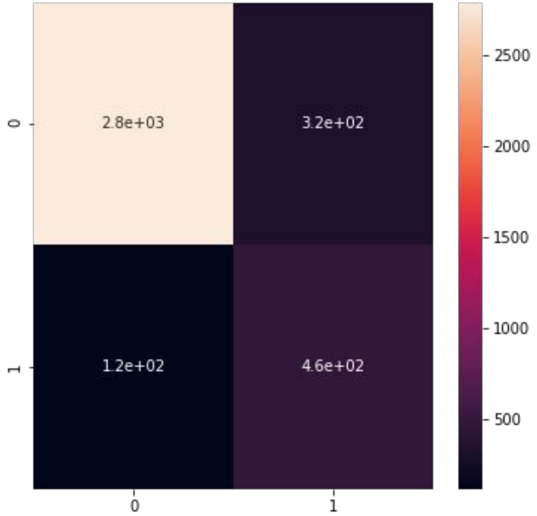
Training period : 80 Epochs



Neural Network

Classification Report : With SMOTE

	Precision	Recall	f1-score	Support
0	0.96	0.90	0.93	3114
1	0.59	0.79	0.68	585
Accuracy			0.88	3699
Macro avg	0.77	0.85	0.80	3699
Weighted avg	0.90	0.88	0.89	3699

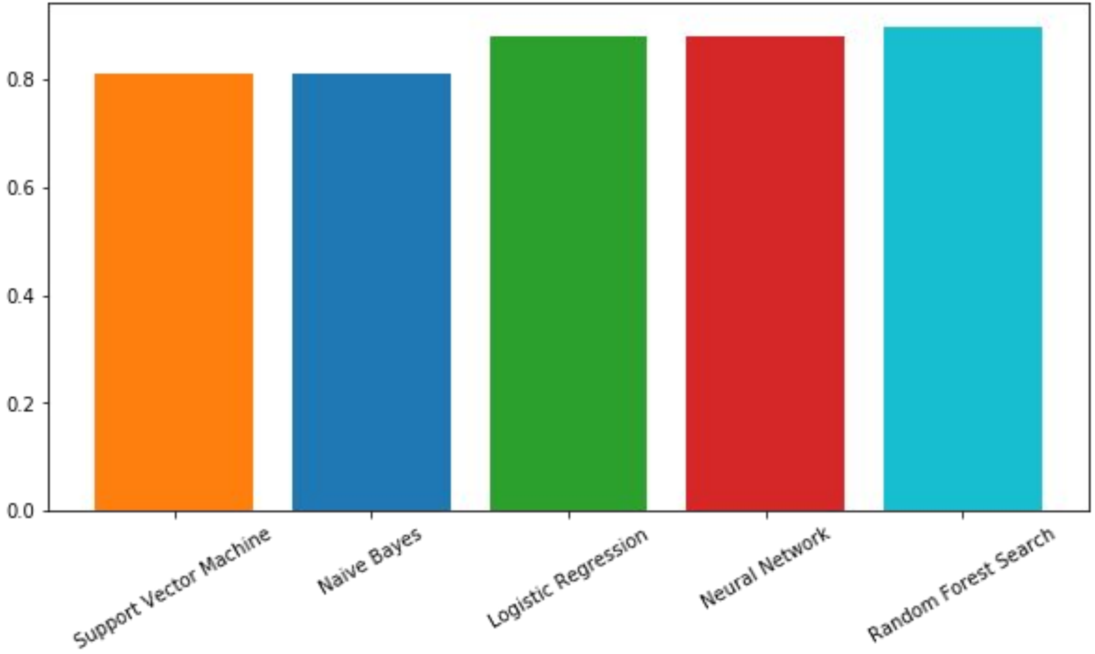


Training accuracy: 0.8807786
Testing accuracy: 0.88050824

Comparison of Models

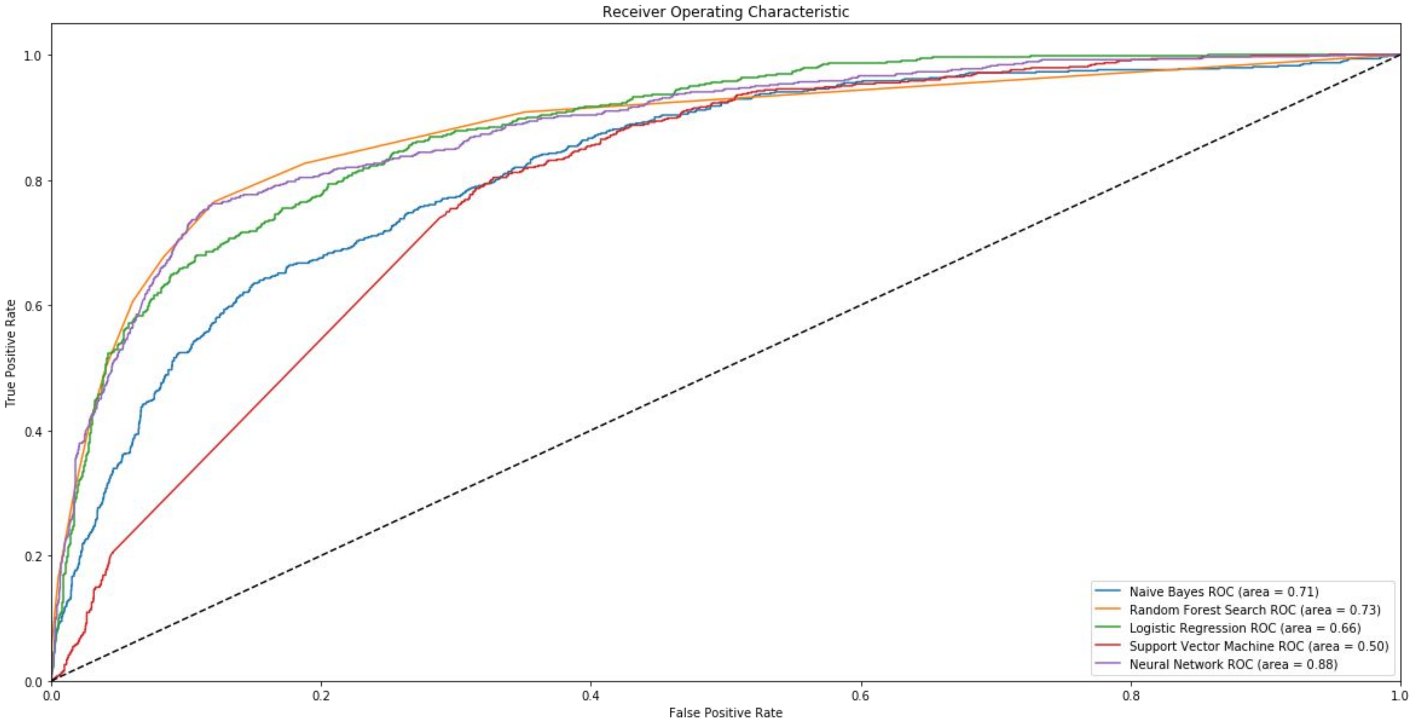
Accuracy

:



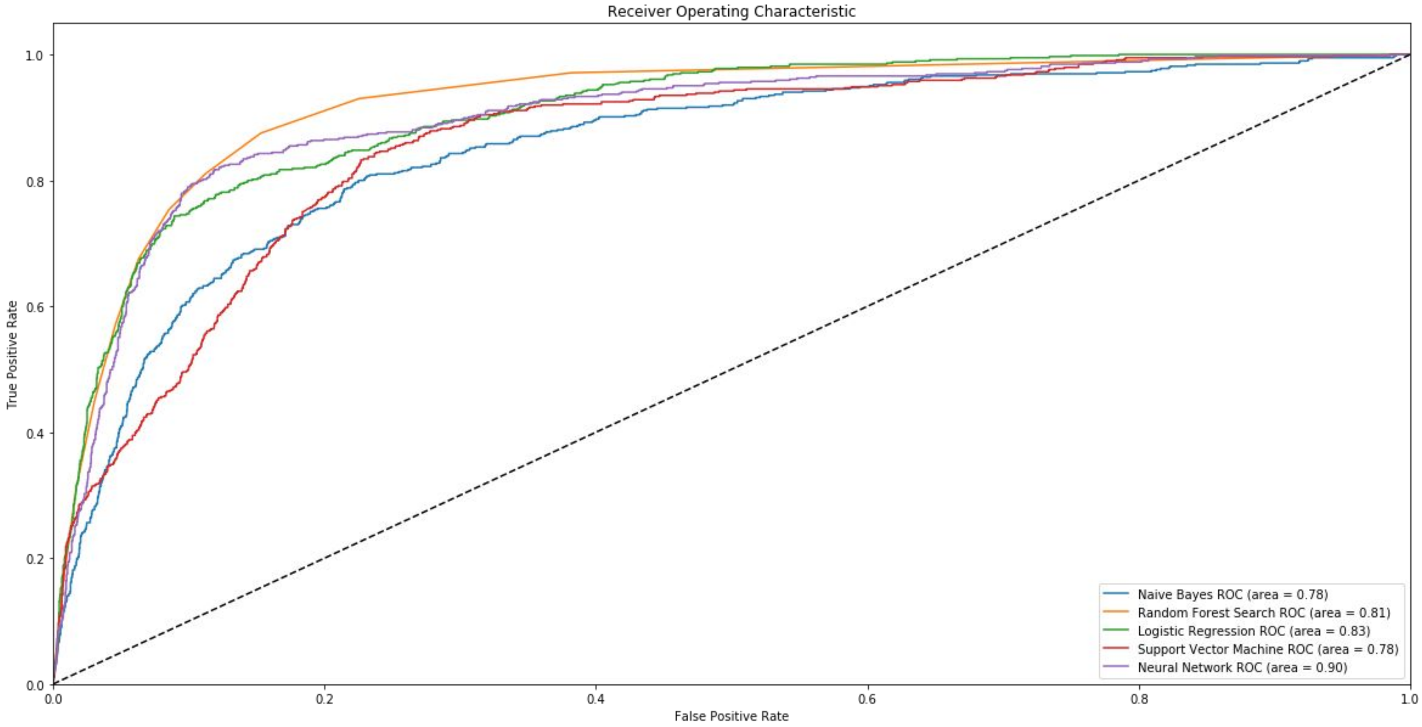
Comparison of Models

ROC Curves (Before optimization)



Comparison of Models

ROC Curves (After optimization)


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
Naive Bayes ROC (area = 0.78)
Random Forest Search ROC (area = 0.81)
Logistic Regression ROC (area = 0.83)
Support Vector Machine ROC (area = 0.78)
Neural Network ROC (area = 0.90)