Online Shoppers Purchasing Intention

CS6220 - Data Mining Techniques

Rajath Kashyap
Mukund Wagh
Bishwarup Neogy

Team Members

- **Mukund**
  - Data Exploration
  - Interpretation
  - Understanding how user spends time on website, interacts with it and generates revenue.
- **Rajath**
  - Analysis
  - Data Preprocessing
  - Naive Bayes
  - Support Vector Machine
- **Bishwarup**
  - Logistic Regression
  - Random Forest Classifier
  - Neural Network
  - Comparison of Models.

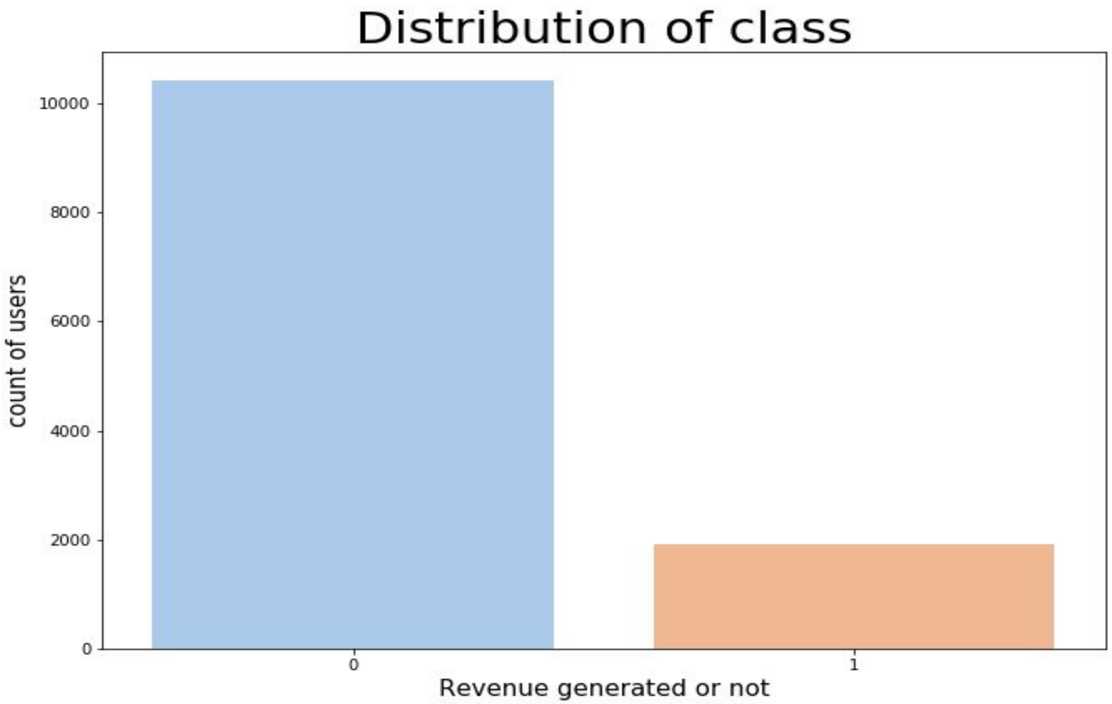
Data Set Description

1. **Administrative**
2. **Administrative Duration**
3. **Informational**
4. **Informational Duration**
5. **Product Related**
6. **Product Related Duration**
7. **Bounce Rate**
8. **Exit Rate**
9. **Page Value**
10. Special Day
11. Month
12. Operating System
13. Browser
14. Region
15. Traffic Type
16. **Visitor Type**
17. **Weekend**
18. Revenue ← Y

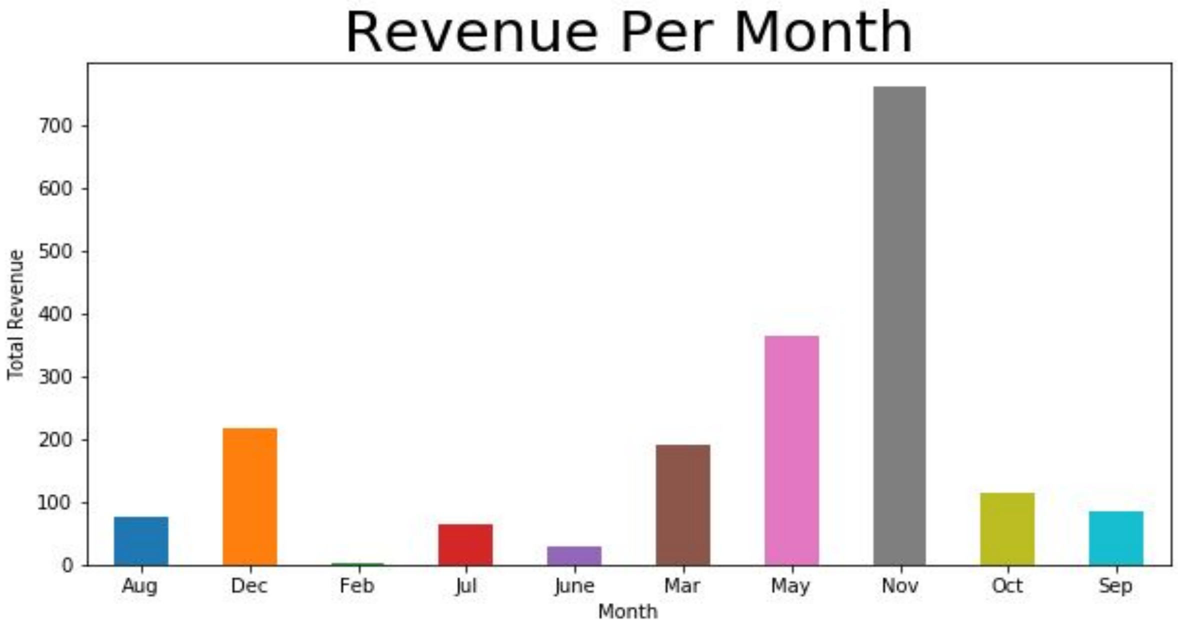
Data Exploration

Revenue Distribution

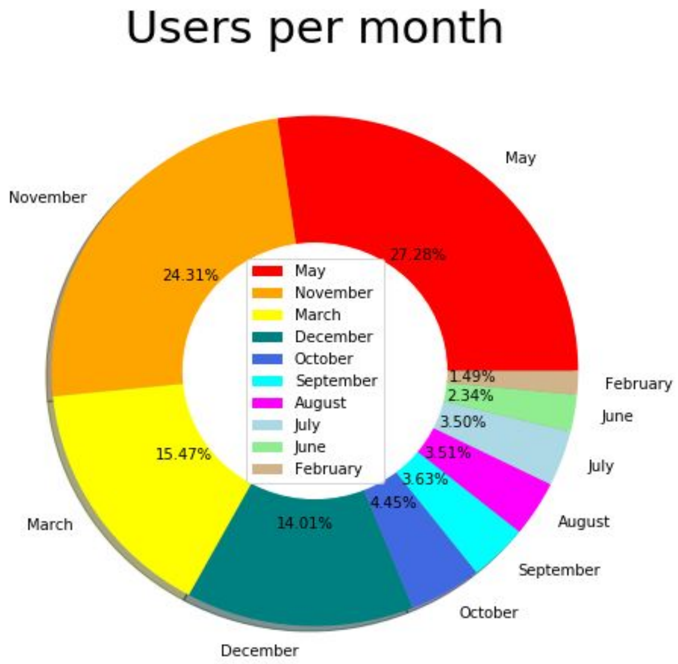
- Count of revenue distribution.
- Revenue distribution over months.
- User visits per month



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



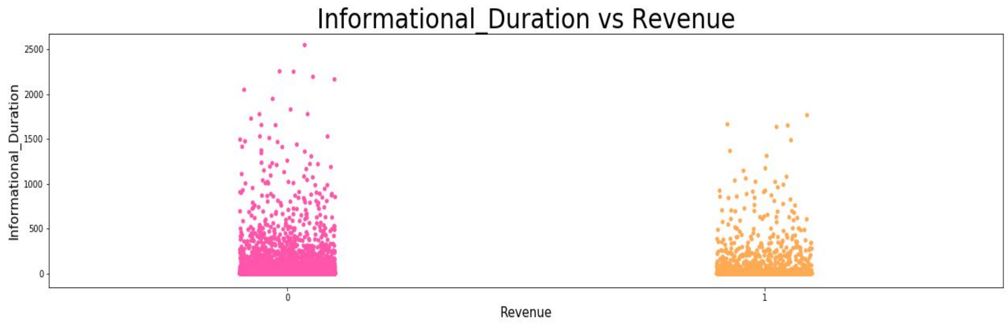
- More the number of visitors more is the sale.
- We have products that fulfill the needs of the customers.



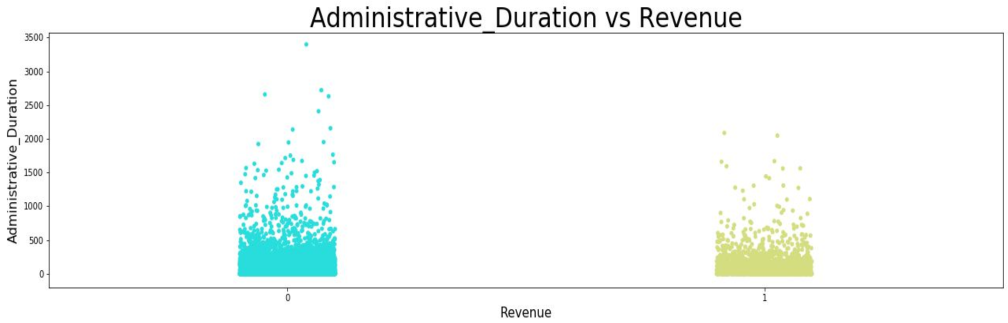
- More the number of visitors more is the sale.
- We have products that fulfill the needs of the customers.

Time Spent on different pages of the website

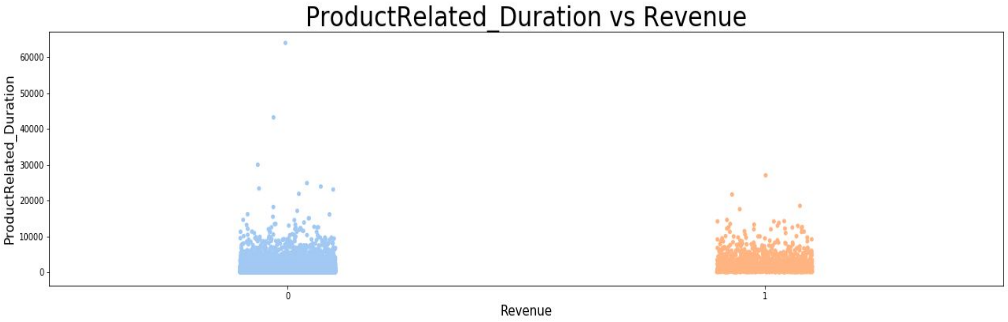
- Bivariate analysis of the time spent
- More time spent high probability that we may lose that potential customer.
- Scope of improvement on the website



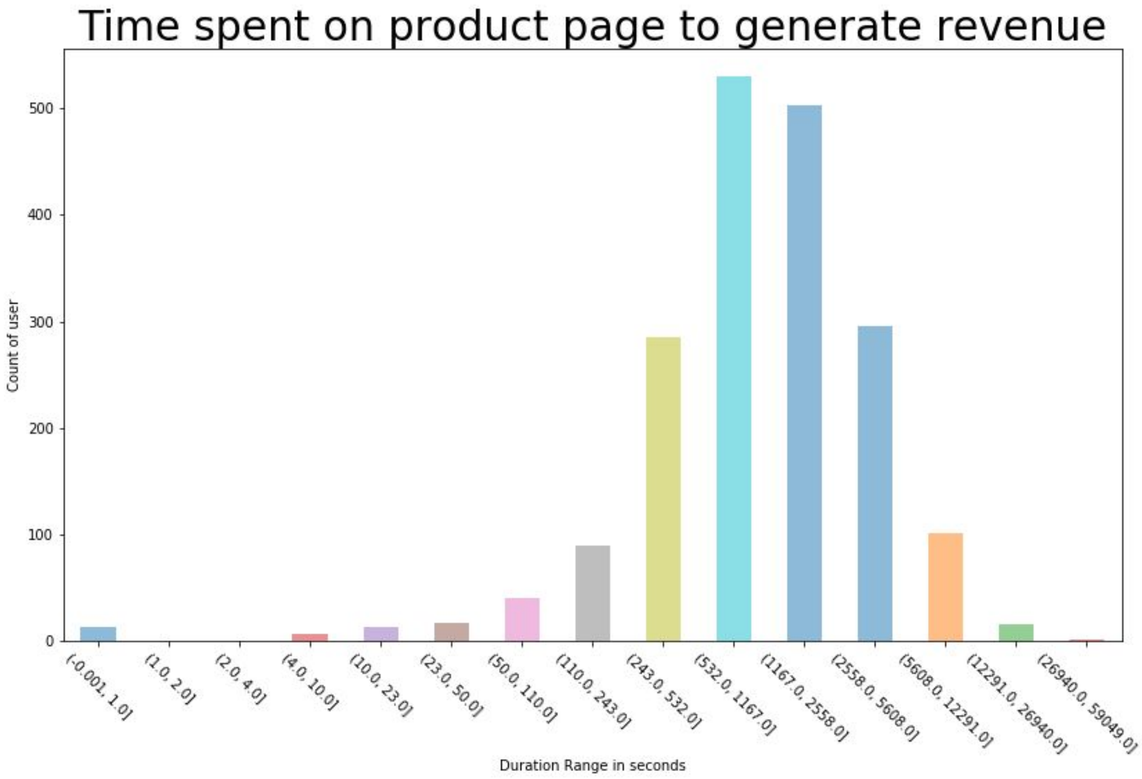
Informational_Duration vs Revenue
Informational_Duration
2500
2000
1500
1000
500
0
0
1
Revenue



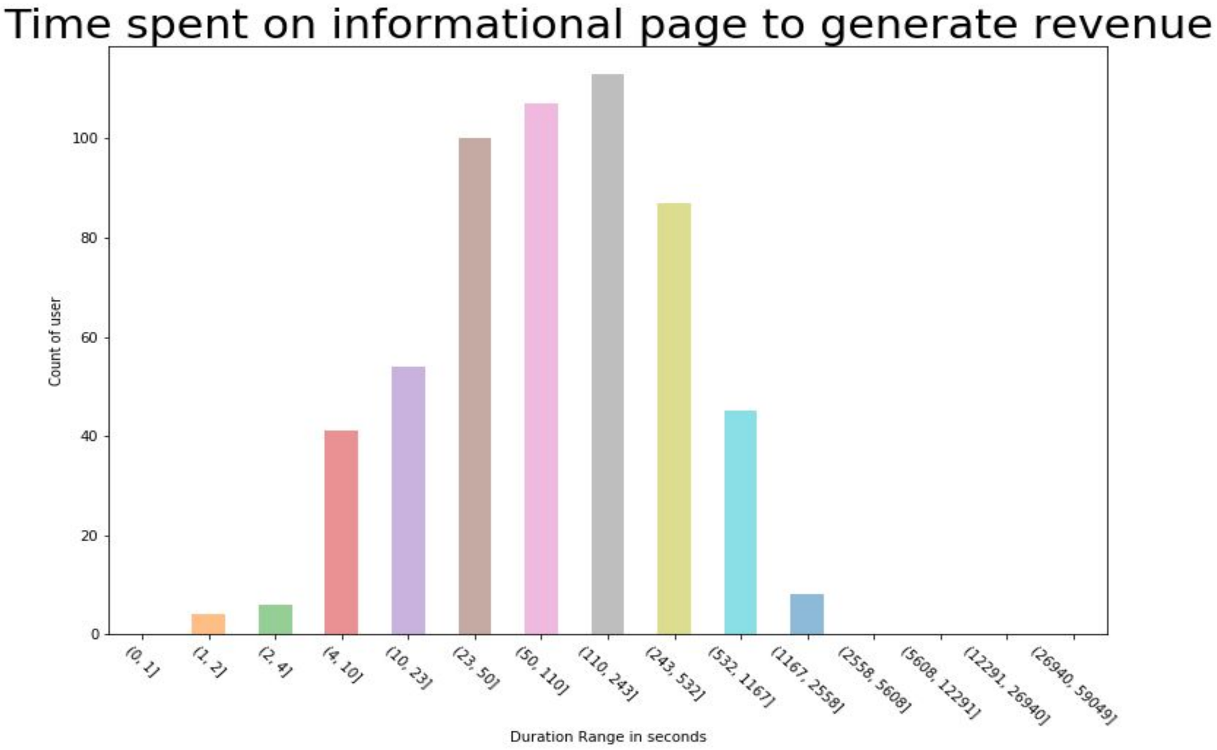
Administrative_Duration vs Revenue
Administrative_Duration
3500
3000
2500
2000
1500
1000
500
0
0
1
Revenue



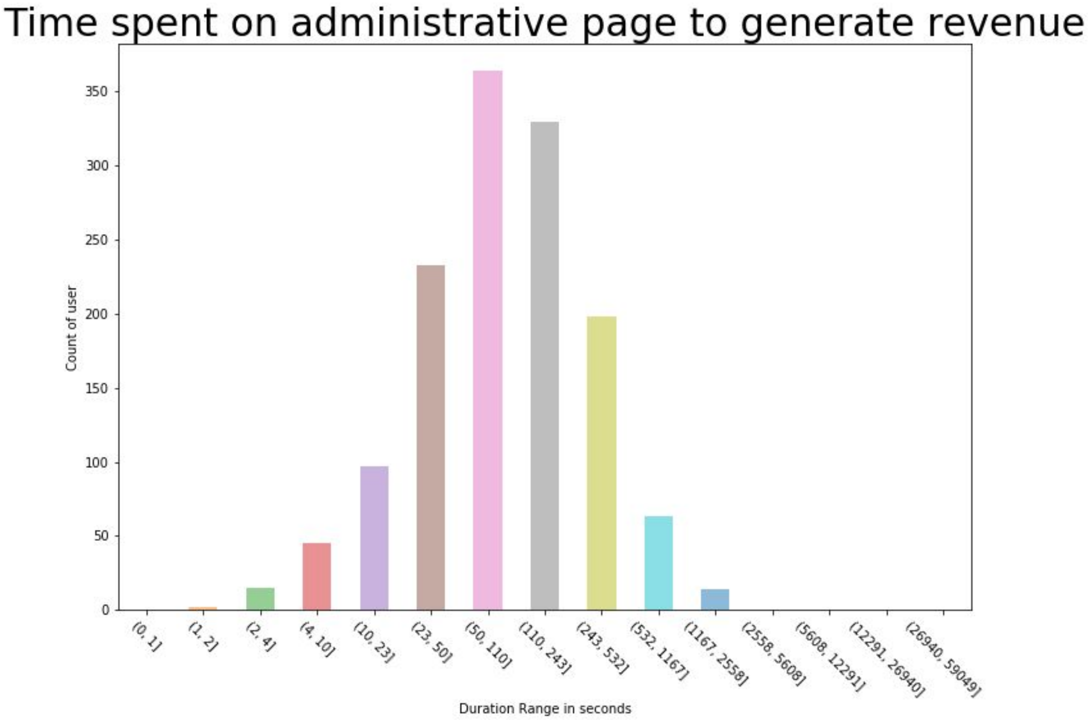
ProductRelated_Duration vs Revenue
ProductRelated_Duration
60000
50000
40000
30000
20000
10000
0
0
1
Revenue



- 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.



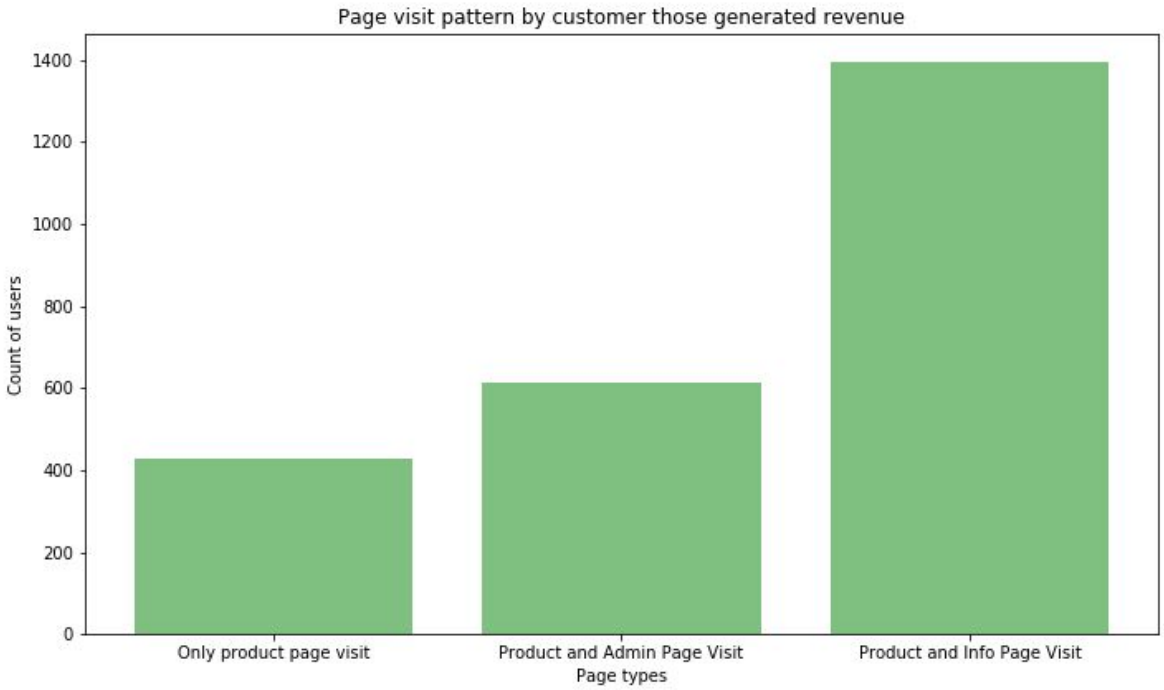
- Many users has to visit the info page to be sure of the product they are going to buy.



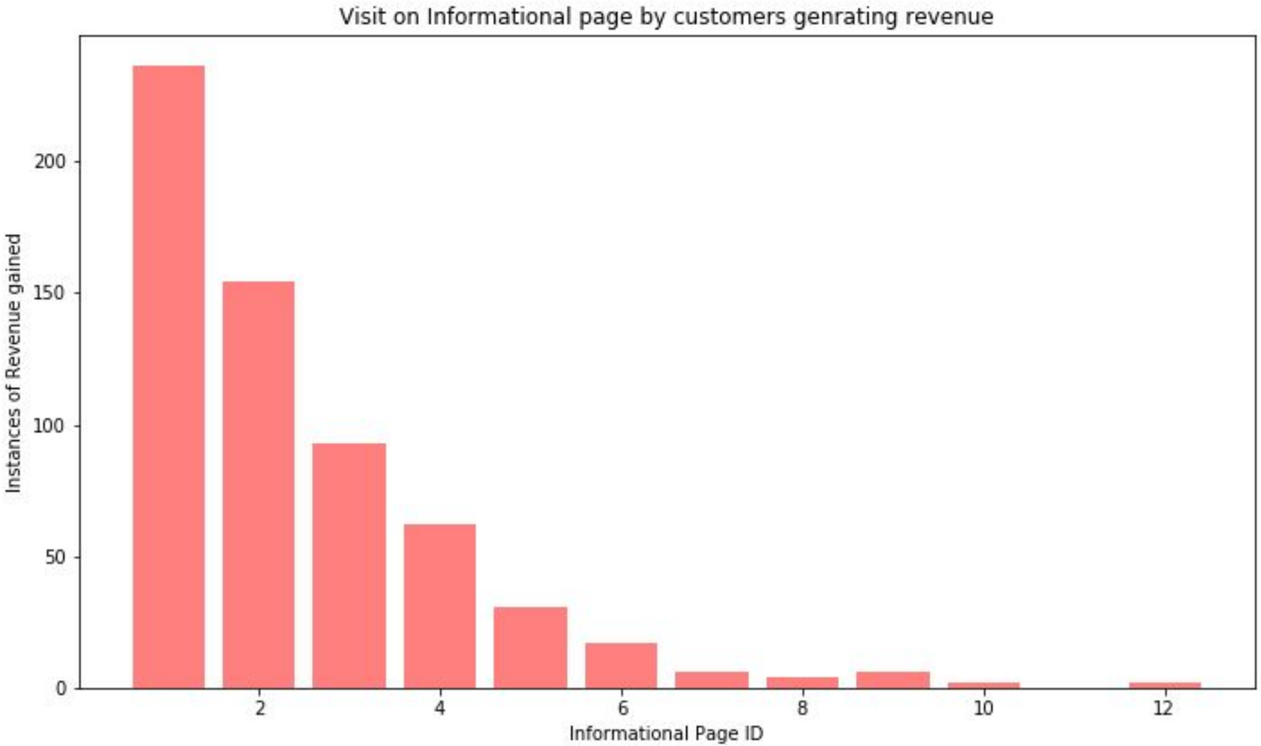
- We have around 2000 unique users who have given us the revenue, and we can see that more than 70% of the users have to visit the administrative page in order to buy the product, also around 50% customers have to spend more than a minute on the administrative pages.

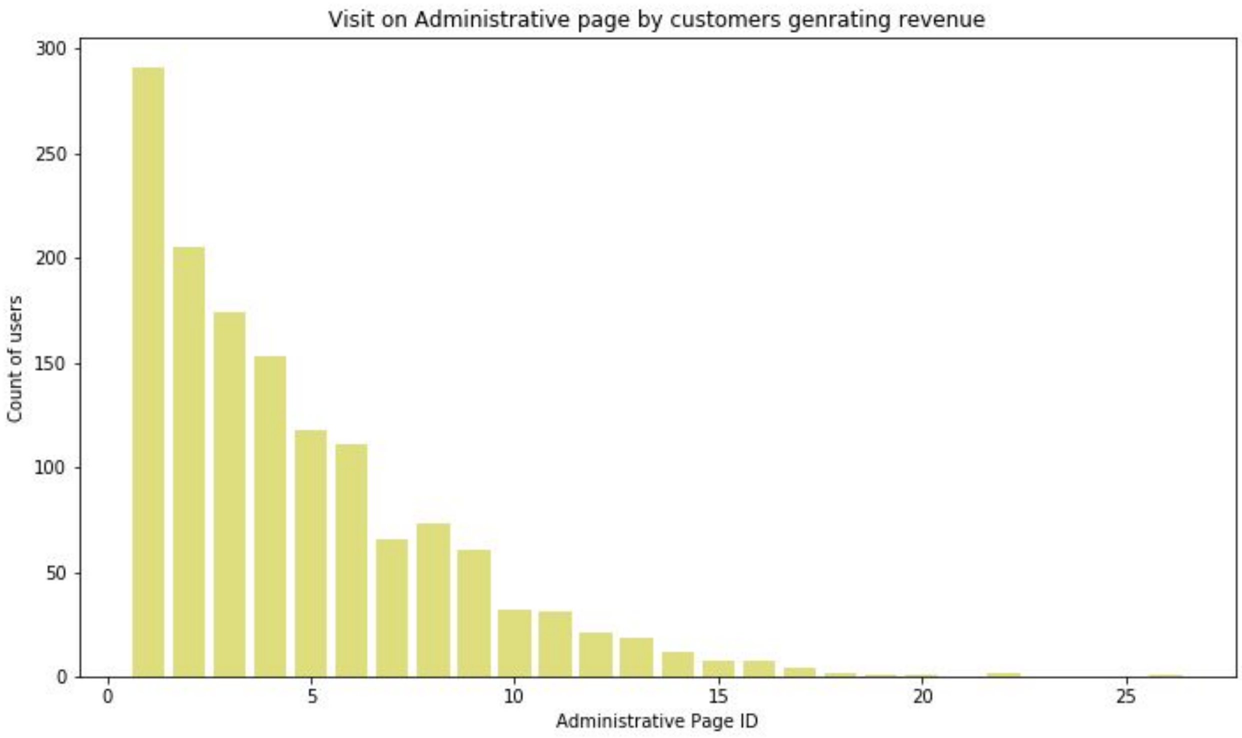
Understanding User Interaction with pages

- Will help in understanding the current system better.



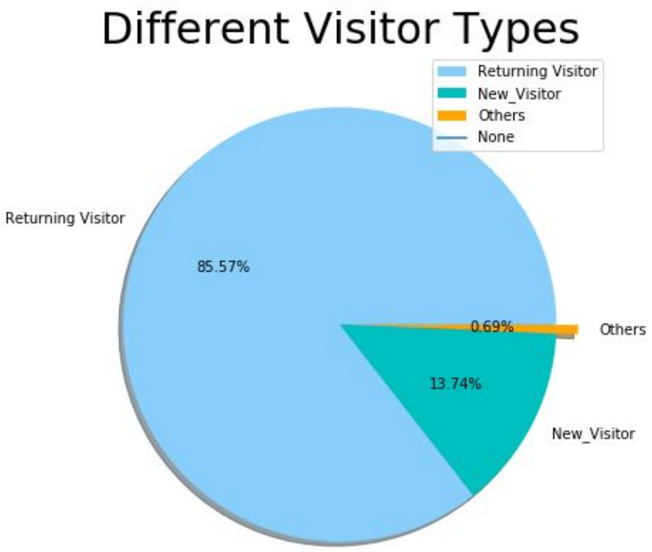
- We have data of users visiting different pages on website.
- Prioritizing the task to retain the potential customers.


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

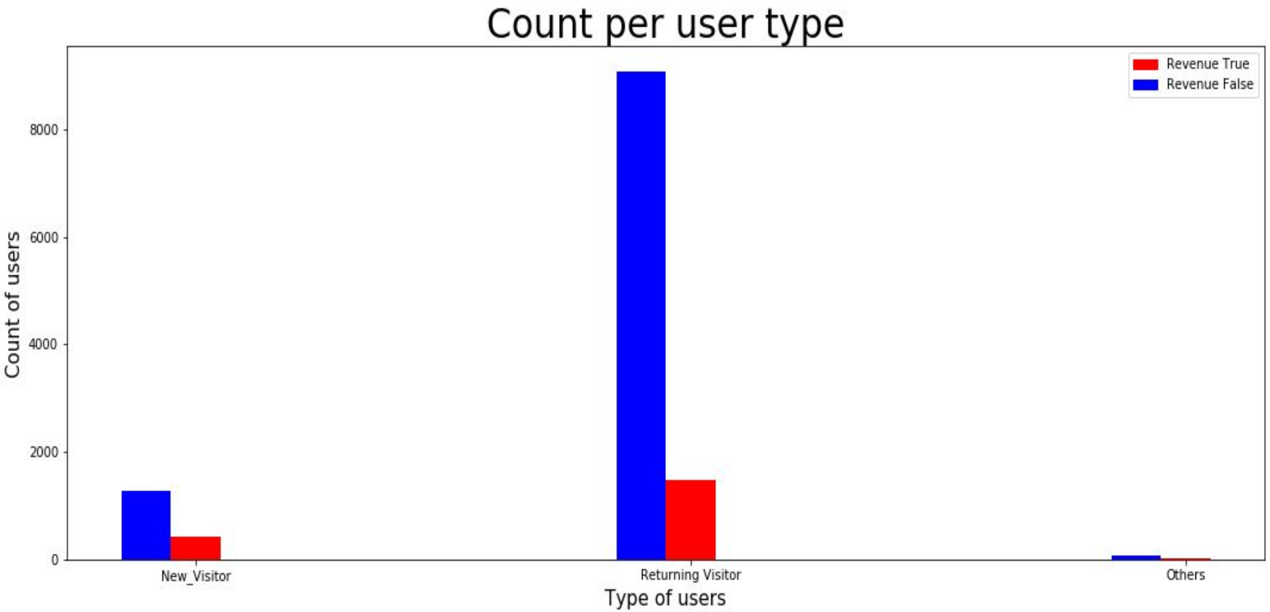

Visit on Administrative page by customers genrating revenue
Count of users
Administrative Page ID
0
50
100
150
200
250
300
0
5
10
15
20
25

Types of users

- Distribution of the user in each type of user.
- Intention to buy items.
- How to increase sales



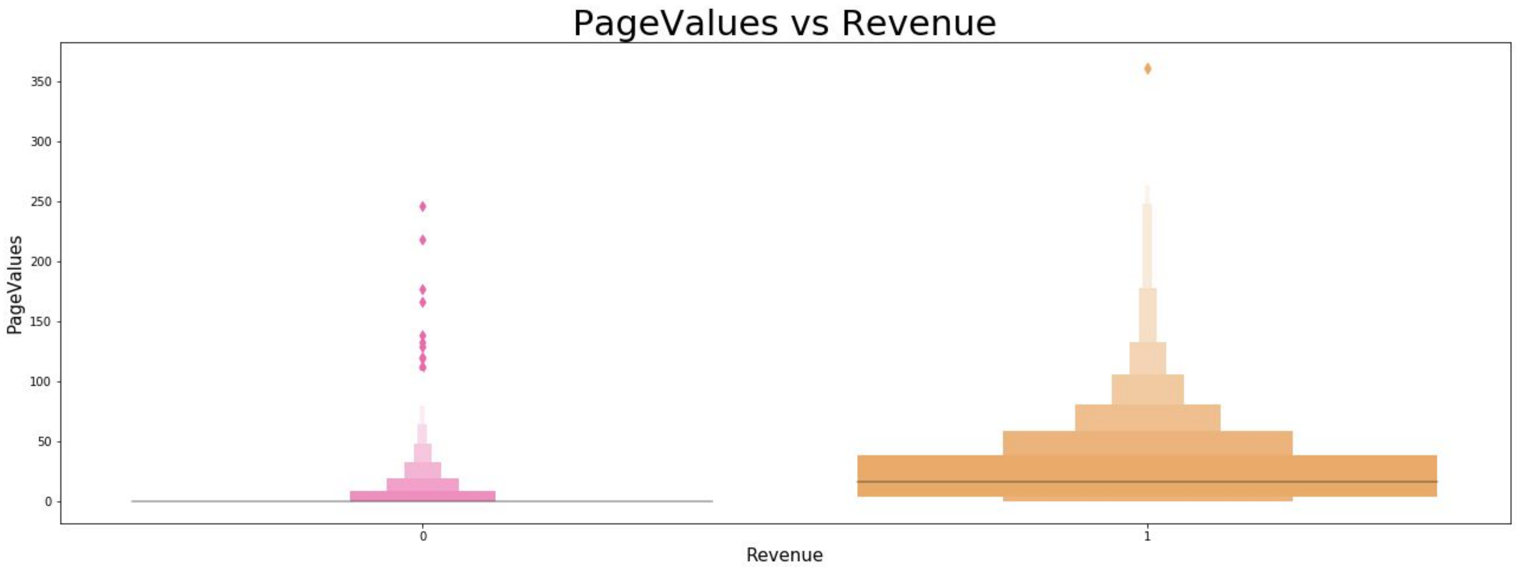
- We have 3 categories of users, the new users, returning users and others.



- 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

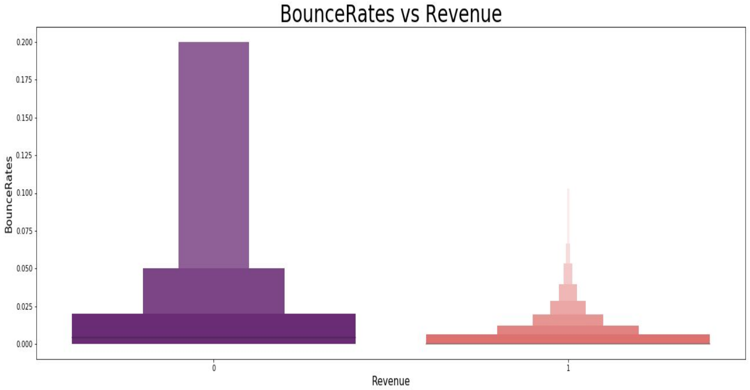


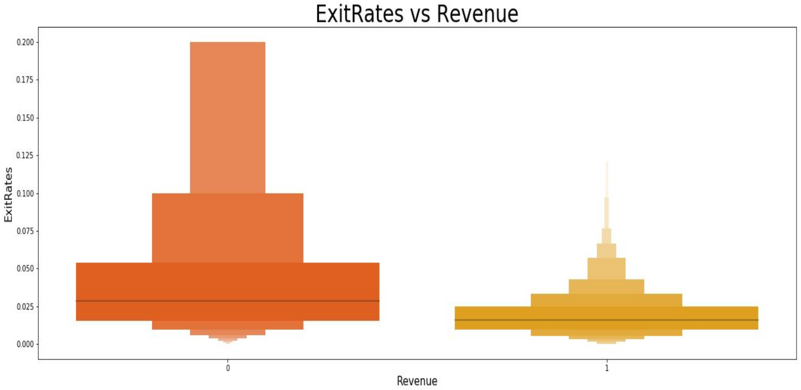
- Page Value is the average value for a page that a user visited before landing on the goal page or completing an Ecommerce transaction (or both). This value is intended to give you an idea of which page in your site contributed more to your site's revenue. If the page wasn't involved in an ecommerce transaction for your website in any way, then the Page Value for that page will be $0 since the page was never visited in a session where a transaction occurred.

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.

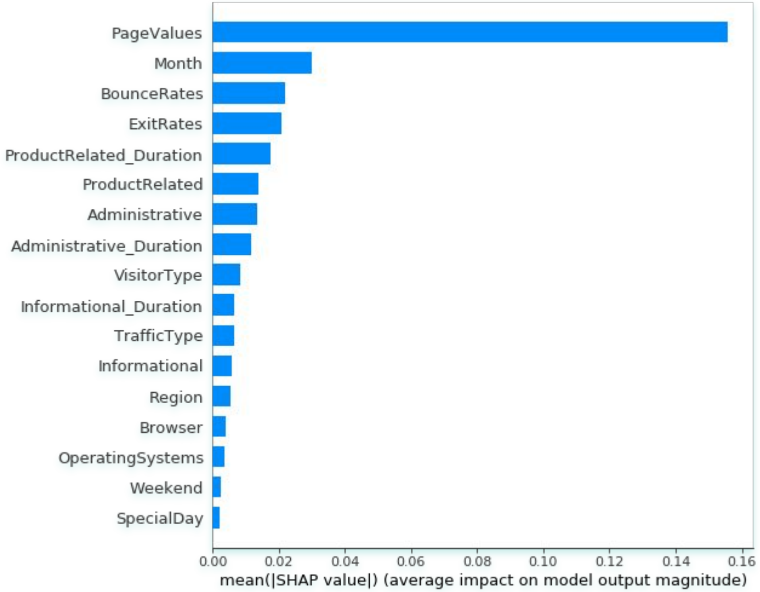


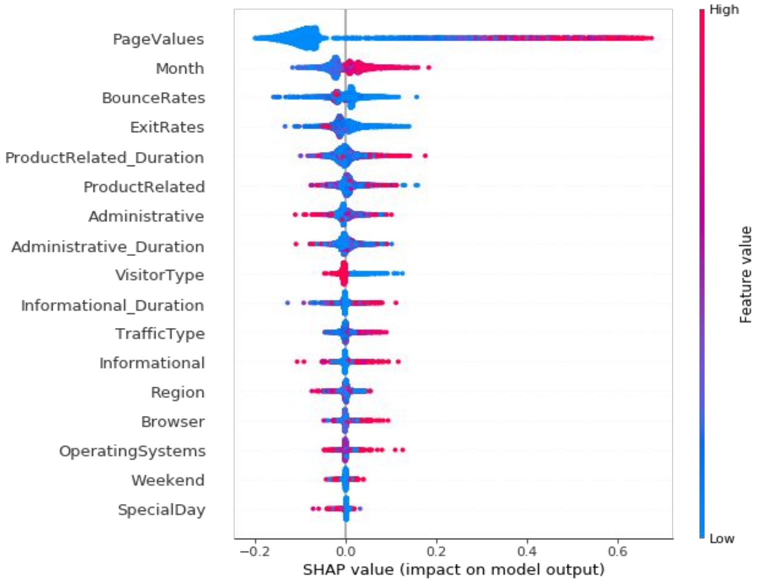


SHapley Additive exPlanation (SHAP)

- Explains the output of any machine learning model using Shapley values
- SHAP assigns a value to each feature for each prediction; the higher the value, the larger the feature attribution to the specific prediction
- Calculation of these values is simple but computationally expensive.
- To compute this, a model is trained with that feature present, and another model is trained with the feature withheld. Then, predictions from the two models are compared on the current input i.e. their difference is computed.
- Since the effect of withholding a feature depends on other features in the model, the preceding differences are computed for all possible subsets of features. The Shapley values are a weighted average of all possible differences and are used as feature attributions.

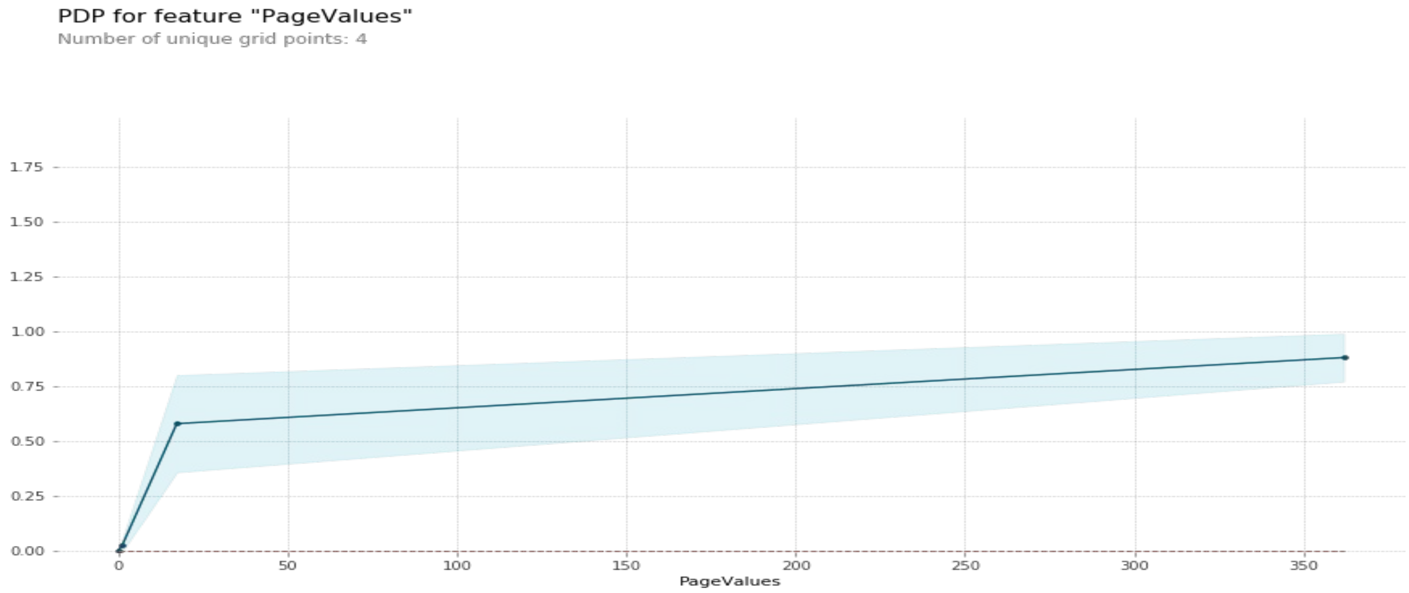
SHapley Additive exPlanation (SHAP) Analysis





Partial Dependence Plot

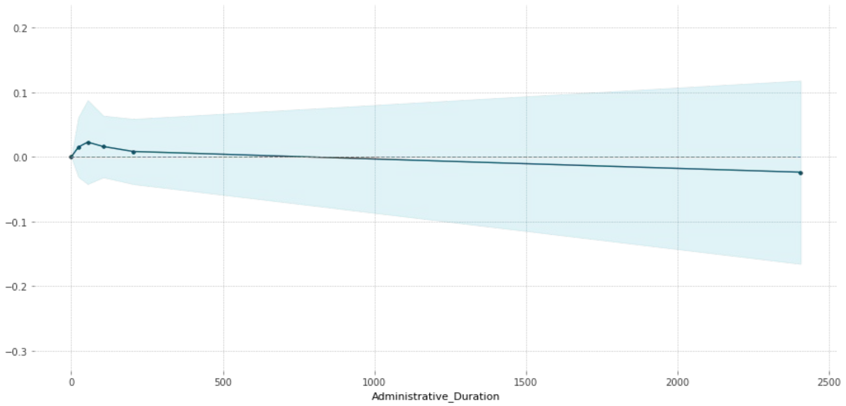
The partial dependence plot (PDP or PD plot) shows the marginal effect one or two features have on the predicted outcome of a machine learning model.The plot can show whether the relationship between the target and a feature is linear, monotonic or more complex



Partial Dependence Plot

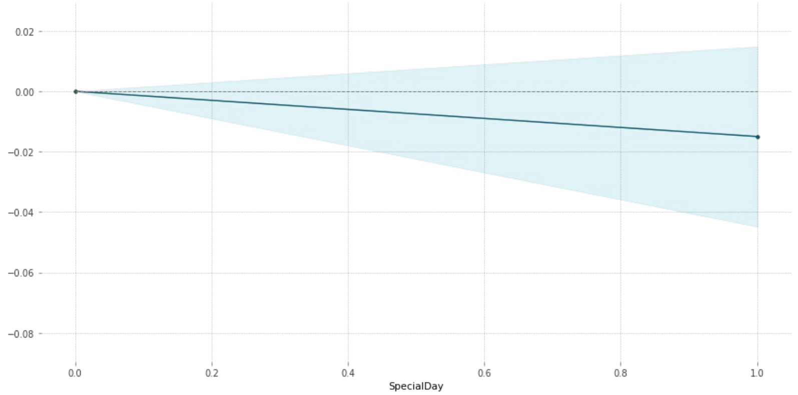
PDP for feature "Administrative_Duration"

Number of unique grid points: 6



PDP for feature "SpecialDay"

Number of unique grid points: 2



Data PreProcessing

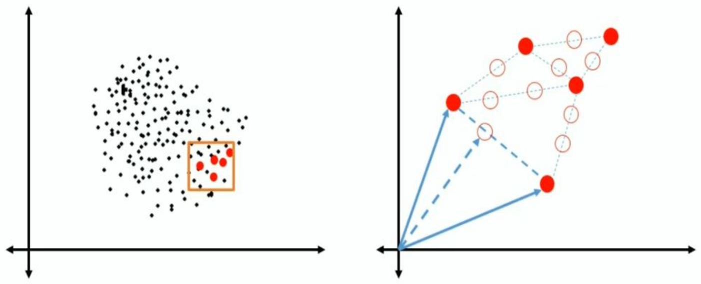
Standardizing the data

- Remove Null values
- Convert String values to numericals
  - Month - January to December
  - VisitorType - New User, Returning User and Other
- Convert boolean values to numericals
  - Weekend and Revenue

Synthetic Minority Oversampling Technique (SMOTE)

Challenges with imbalance Data

- produces biased predictions
- conventional model evaluation methods do not accurately measure model performance



How it works:

- creates synthetic samples from the minor class.
- calculates the k nearest neighbors and selects two or more similar instances (using a distance measure) and multiplies an instance by a random amount within the difference to the neighboring instances.

Data Modeling

Models

We used the given data set to train five different models and compared their performance.

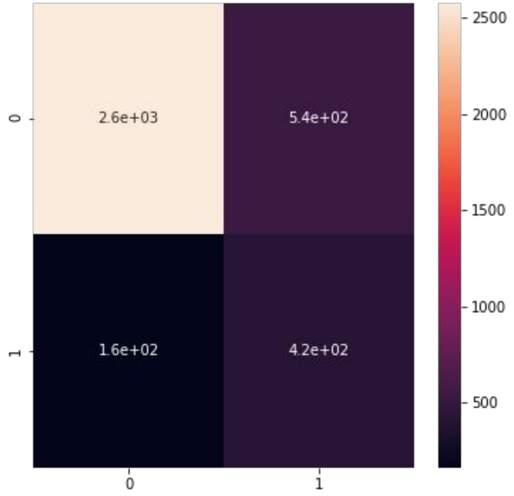
- **Naive Bayes**
- **Support Vector Machine**
- Logistic Regression
- Random Forest Classifier
- Neural Network

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



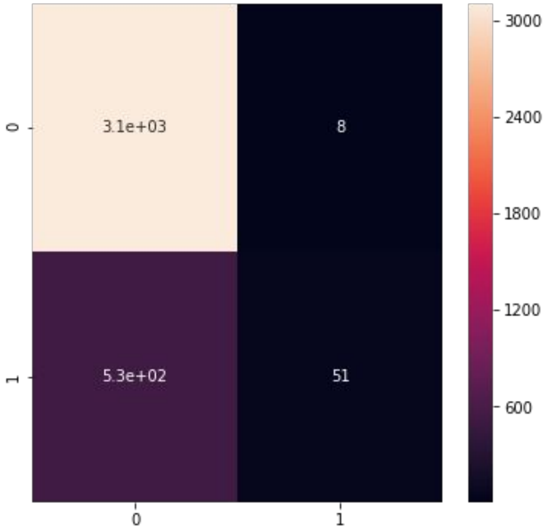
Support Vector Machine (SVM)

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:

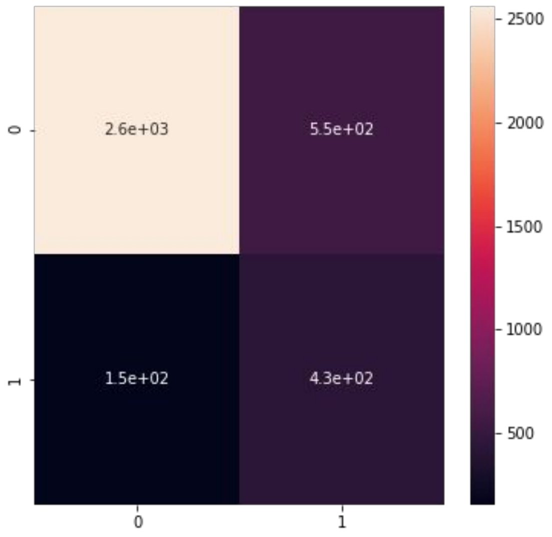


Support Vector Machine (SVM)

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:



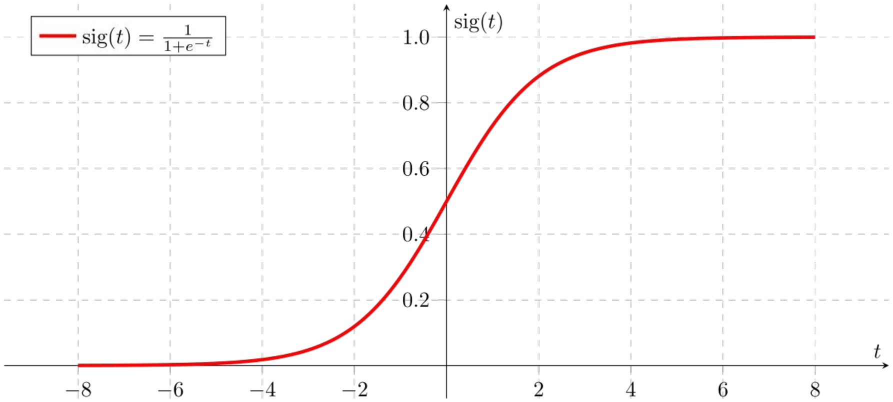
Models

We used the given data set to train five different models and compared their performance.

- Naive Bayes
- Support Vector Machine
- **Logistic Regression**
- **Random Forest Classifier**
- **Neural Network**

Logistic Regression

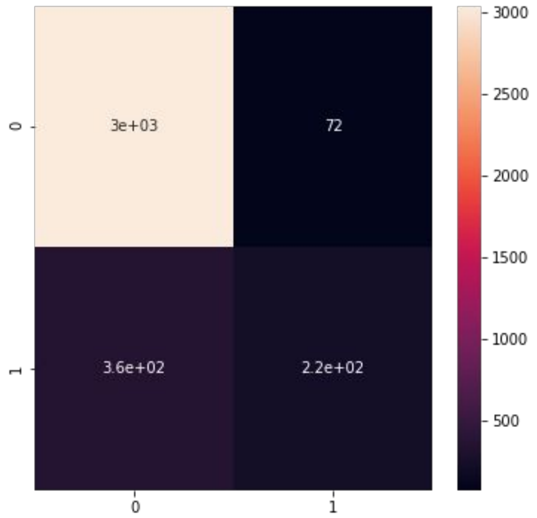
- An algorithm for Classification
- Used where response variable is categorical
    - Ex: Tumour - Malignant/Benign
- To find a relationship between:
    - **Features** and a **Particular Outcome**



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

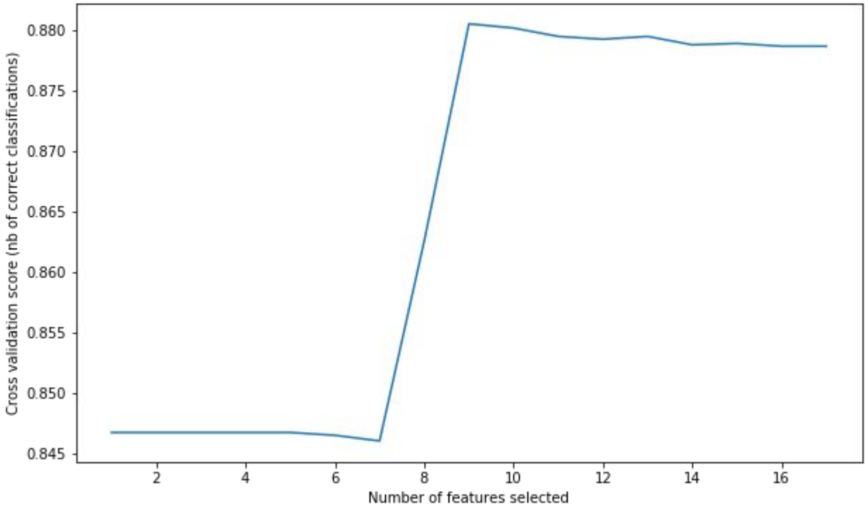


Logistic Regression

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

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

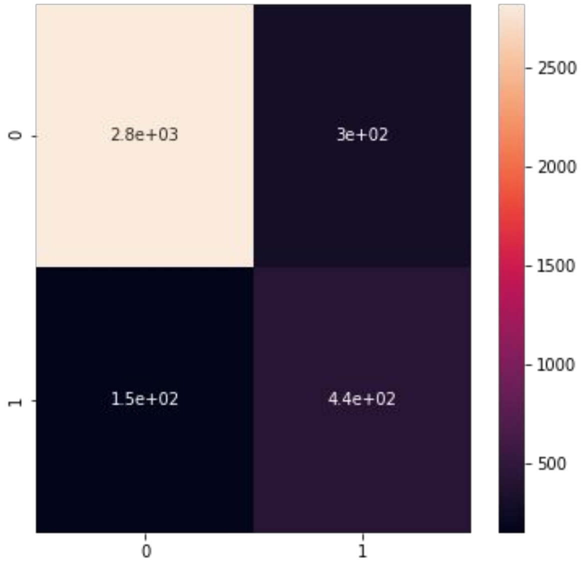
Selected features:
['Informational', 'BounceRates', 'ExitRates', 'PageValues', 'SpecialDay', 'Month', 'OperatingSystems', 'VisitorType', 'Weekend']



Logistic Regression

Classification Report : with SMOTE and RFE

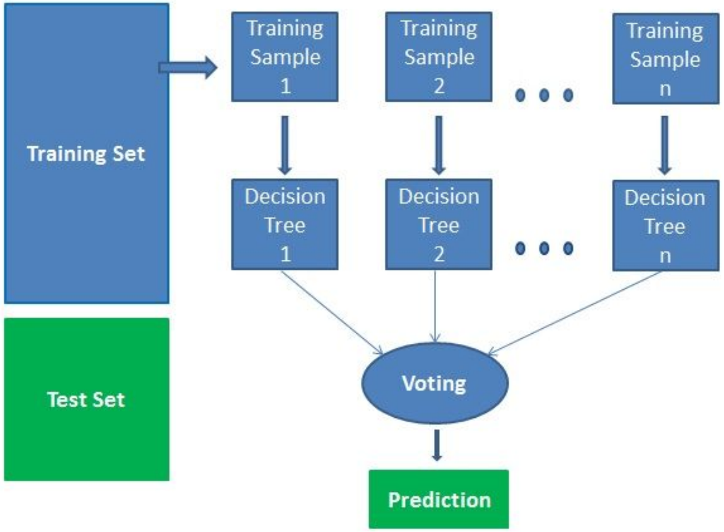
	Precision	Recall	f1-score	Support
0	0.95	0.91	0.93	3114
1	0.60	0.75	0.66	585
Accuracy			0.88	3699
Macro avg	0.77	0.83	0.79	3699
Weighted avg	0.89	0.88	0.89	3699



Random Forest Classifier

Many Trees ~ A Forest

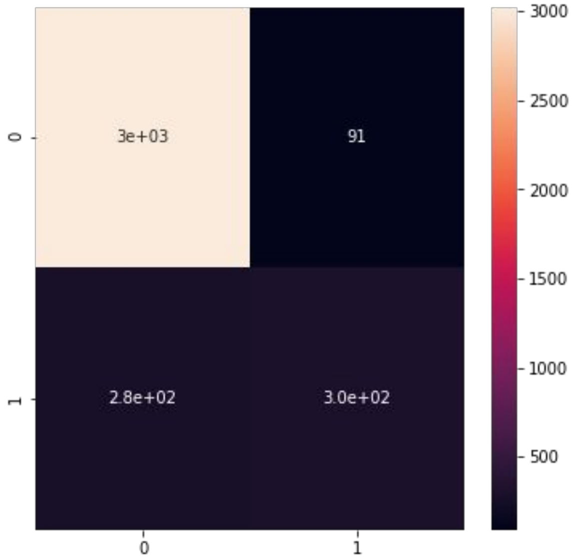
- Select **random samples** from dataset.
- Construct **decision tree** for each sample and get a prediction.
- Perform a **vote** for each predicted result.
- Select the prediction result with the **most votes** as the final prediction.



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

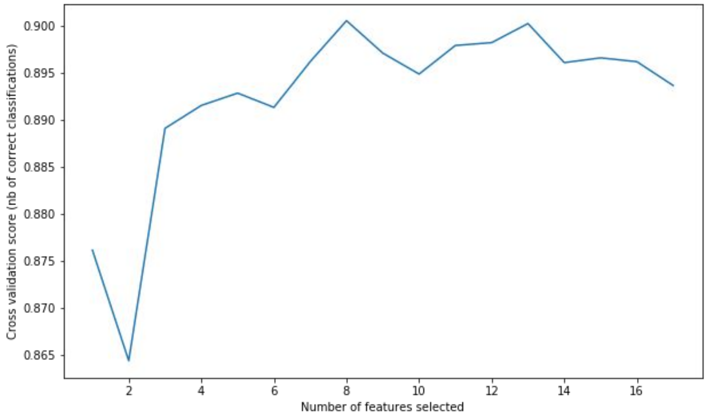


Random Forest Classifier

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']

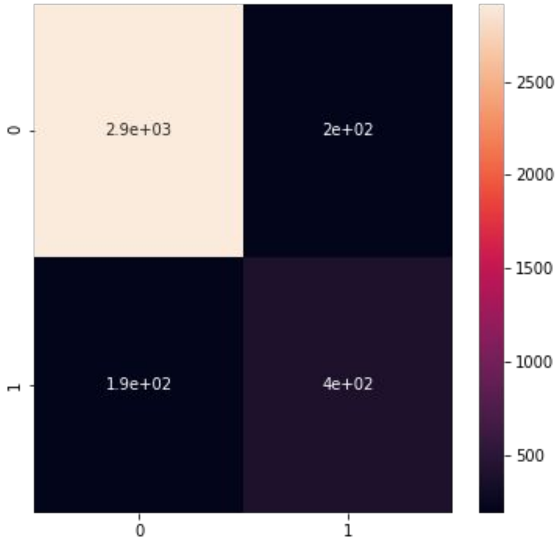


Random Forest Classifier

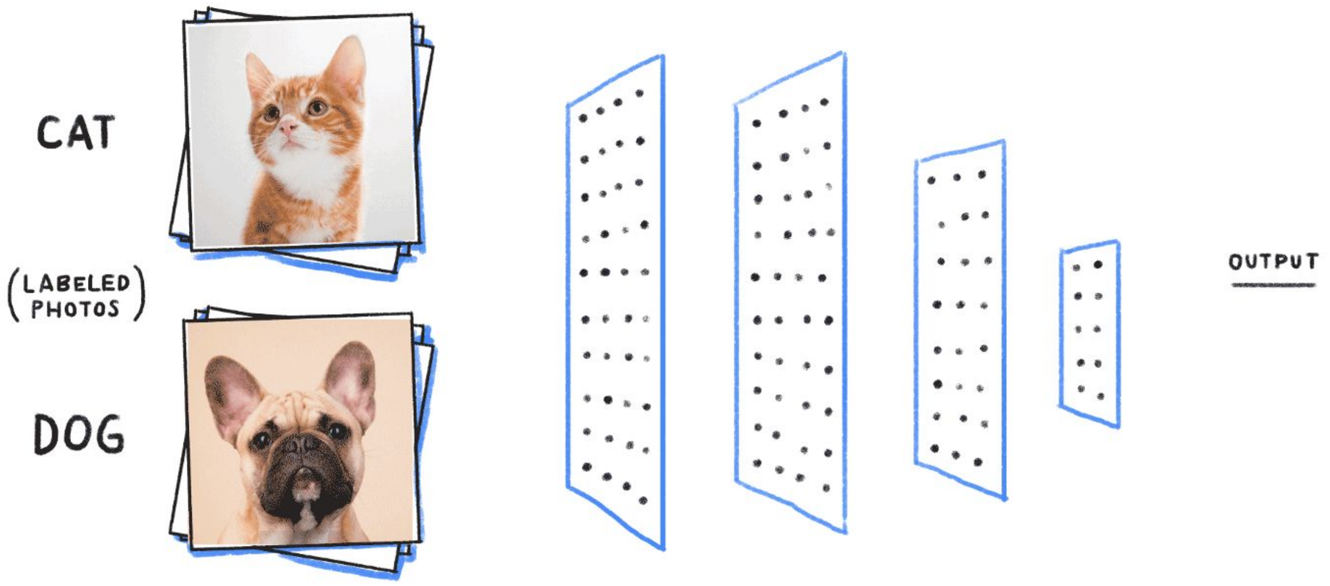
We again evaluated our classifier and obtained the following results:

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



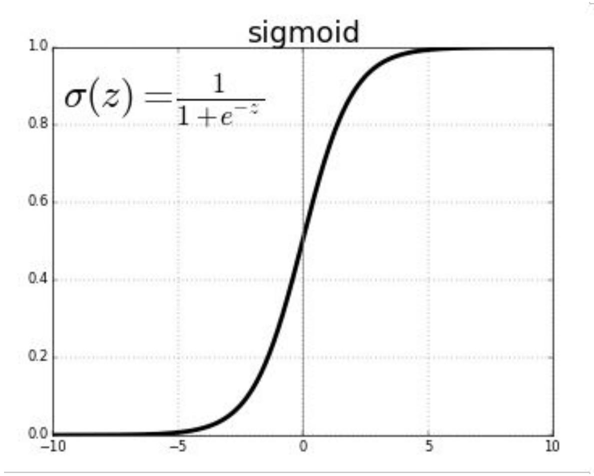
Neural Network

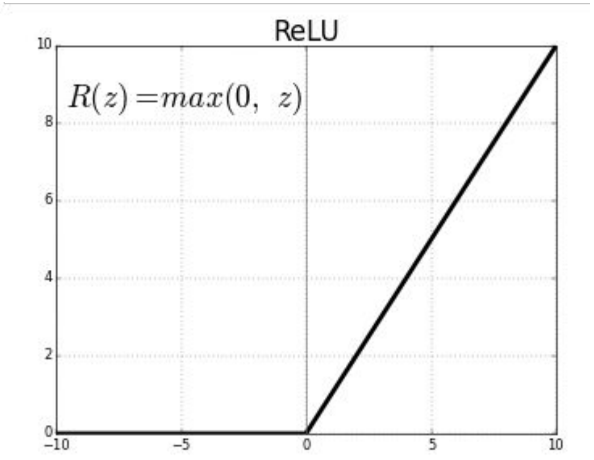
```
model = keras.Sequential([
    keras.layers.Dense(60, input_shape=(x_train.shape[1],), activation=tf.nn.relu),
    keras.layers.Dense(units=1, activation=tf.nn.sigmoid)
])
```

```
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

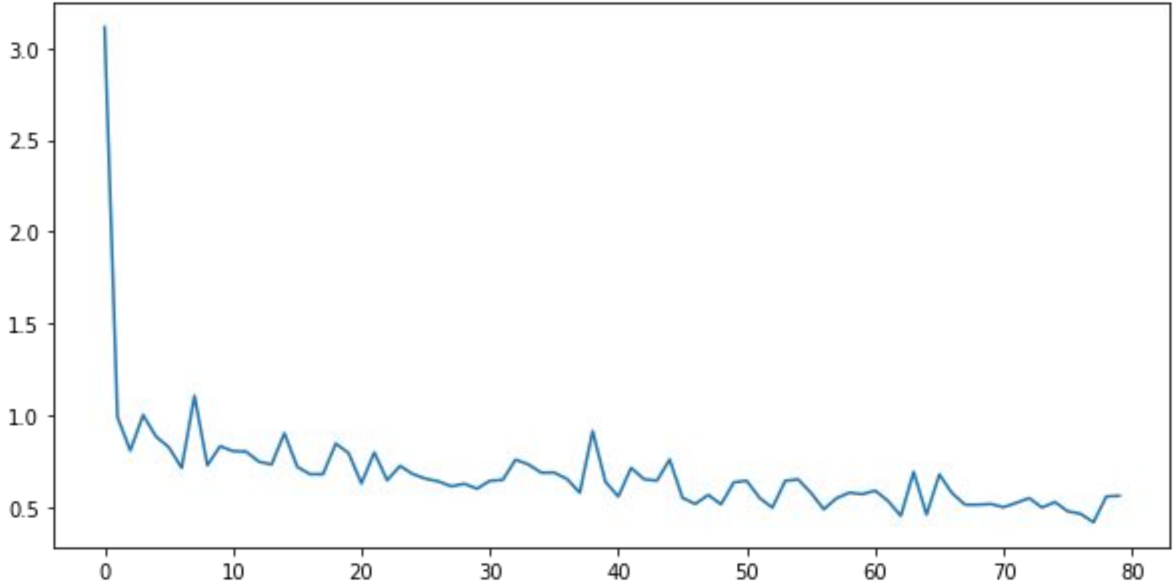
- Hidden layer activation : Rectified linear unit (ReLU)
  - Preferred because derivative is 1.
  - Sigmoid not suitable : Slows down gradient descent.
- Output Layer activation : Sigmoid function
  - Prefered because in a binary classifier we want the output to be between 0 and 1.





Neural Network

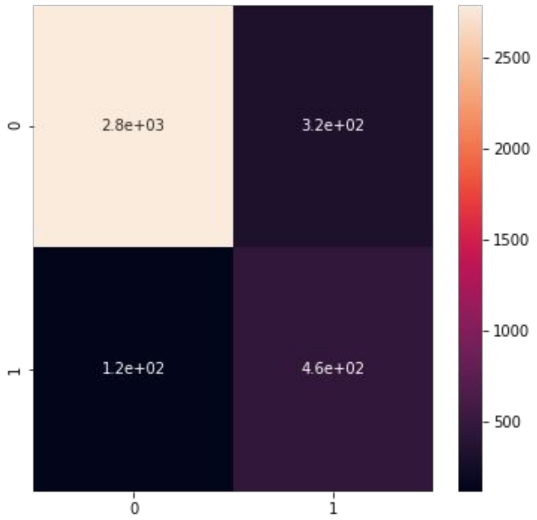
Training period : 80 Epochs



Neural Network

Classification Report : With SMOTE

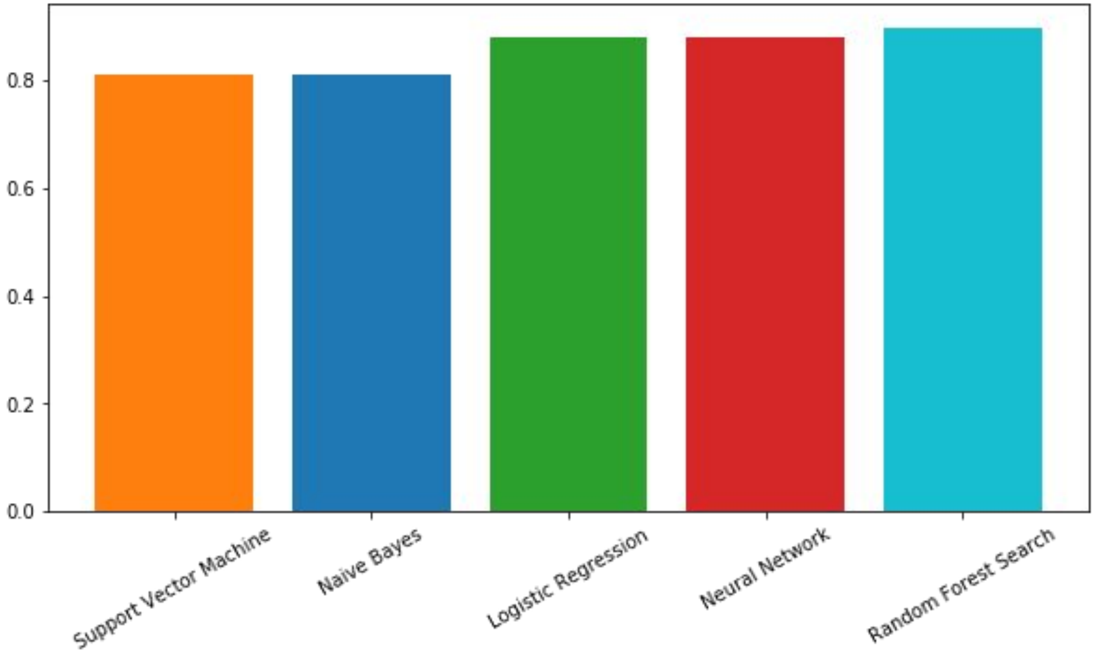
	Precision	Recall	f1-score	Support
0	0.96	0.90	0.93	3114
1	0.59	0.79	0.68	585
Accuracy			0.88	3699
Macro avg	0.77	0.85	0.80	3699
Weighted avg	0.90	0.88	0.89	3699



Training accuracy: 0.8807786
Testing accuracy: 0.88050824

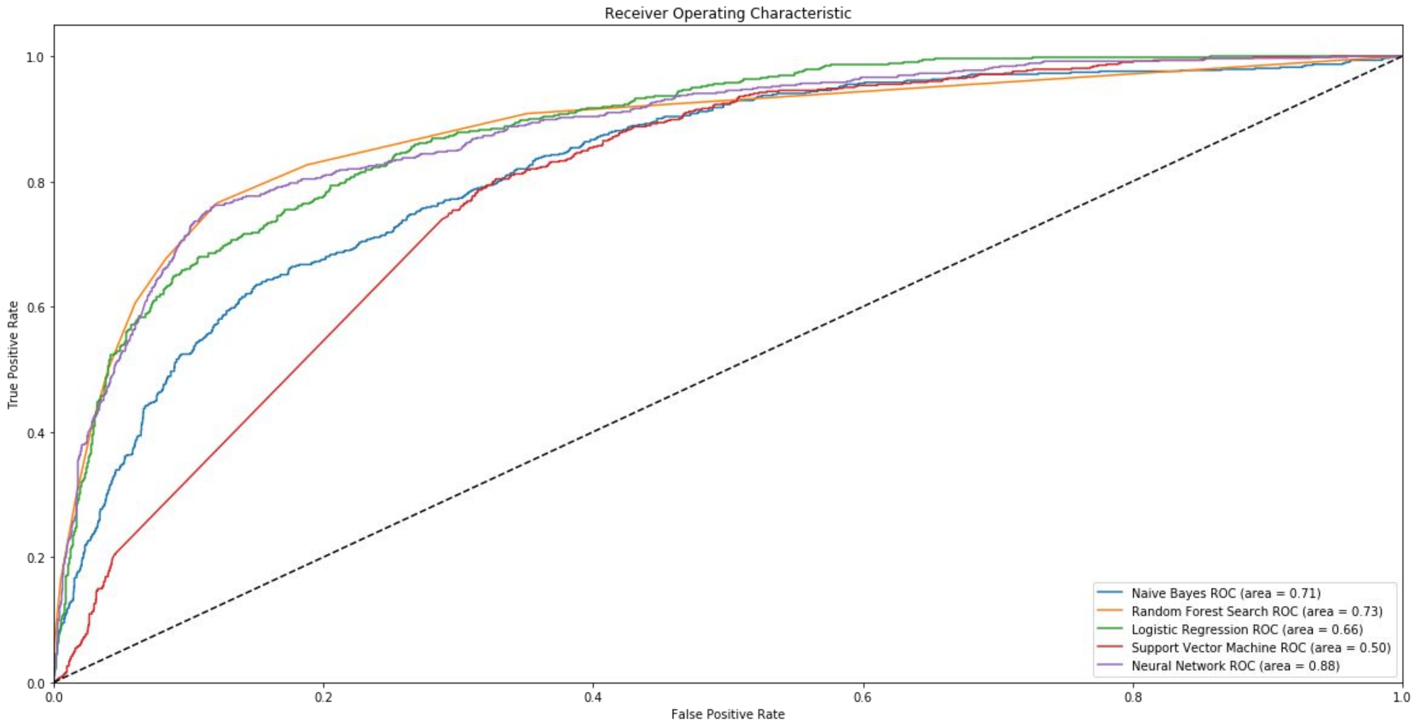
Comparison of Models

Accuracy:



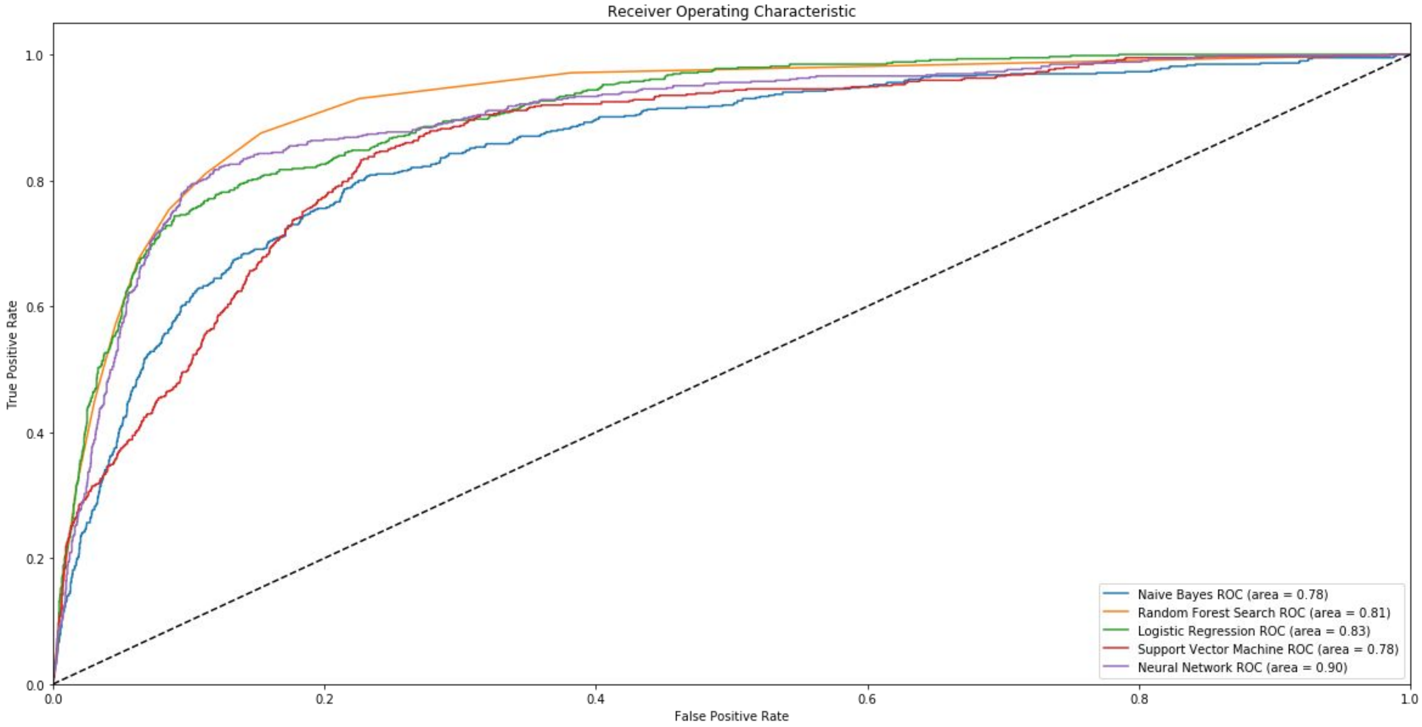
Comparison of Models

ROC Curves (Before optimization)



Comparison of Models

ROC Curves (After optimization)



Conclusion

Overall Results

- More the time spent on the website by the user, less is the probability of generating revenue.
- Users intend to buy more on weekdays.
- We have very few new visitors, will have to advertise more about the website to increase the sales.
- A large number of datasets are imbalanced.
    - Metrics like Accuracy is not always reliable.
    - Recall, ROC curves are better metrics.

Future Work

Purchasing Prediction is a basis for:

- Targeted online Ads.
- Recommendation Systems.
- Association between specific products on specific days.
- Try other techniques (Ex: under sampling, ADASYN) to tackle Imbalance in the dataset.

THANK YOU