SUMMARY & INFERENCES

Likelihood Models Inferences:

As stated above, we have used two different techniques: One for dimensionality reduction and other for Feature selection. We used PCA for dimensionality reduction and WOE and IV for Feature selection.

Firstly we have applied PCA and based on that we have implemented different models:

Model scores after PCA implementation:

Decision Tree: We have got a testing accuracy of 77% with a precision of False as 73% and True as 70%.

The recall for False obtained is 68% and True is 75%. We got the F1 score of False as 71% and True as 72%.

The confusion matrix obtained

2101	976
776	2301

Naïve Bayes: We have got a testing accuracy of 70% with a precision of False as 68% and True as 72%.

The recall for False obtained is 75% and True is 65%. We got the F1 score of False as 72% and True as 69%.

The confusion matrix obtained

2307	770
1068	2009

SVM: We have got a testing accuracy of 87% and training accuracy of 90%, with a precision of False as 93% and True as 81%.

The recall for False obtained is 83% and True is 92%. We got the F1 score of False as 88% and True as 86%.

The confusion matrix obtained

2871	571
216	2506

Random Forest: We have got a testing accuracy of 80% and training accuracy of 99% with a precision of False as 90% and True as 71%. The recall for False obtained is 75% and True is 87%. We got the F1 score of False as 82% and True as 78%.

The confusion matrix obtained

2464	613
817	2260

Adaboost: We have got a testing accuracy of 79% and training accuracy of 76% with a precision of False as 79% and True as 75%. The recall for False obtained is 76% and True is 78%. We got the F1 score of False as 77% and True as 76%.

The confusion matrix obtained

2431	777
646	2300

Gradient Boosting: We have got a testing accuracy of 81% and training accuracy of 87% with a precision of False as 83% and True as 80%. The recall for False obtained is 81% and True is 83%. We got the F1 score of False as 82% and True as 81%.

The confusion matrix obtained

2567	614
510	2463

Model scores after WOE and IV implementation:

Decision Tree: We have got a testing accuracy of 82% with a precision of False as 92% and True as 90%.

The recall for False obtained is 90% and True is 92%. We got the F1 score of False as 91% and True as 91%.

The confusion matrix obtained

2779	298
251	2826

Naïve Bayes: We have got testing accuracy of 74% with a precision of False as 75% and True as 75%.

The recall for False obtained is 75% and True is 75%. We got the F1 score of False as 75% and True as 75%.

The confusion matrix obtained

2316	761
784	2293

SVM: We have got a testing accuracy of 87% and training accuracy of 90%, with a precision of False as 93% and True as 81%.

The recall for False obtained is 83% and True is 92%. We got the F1 score of False as 88% and True as 86%.

The confusion matrix obtained

2871	571
216	2506

Random Forest: We have got a testing accuracy of 89% and training accuracy of 86% with a precision of False as 93% and True as 94%. The recall for False obtained is 94% and True is 84%. We got the F1 score of False as 90% and True as 88%.

The confusion matrix obtained

2882	195
480	2597

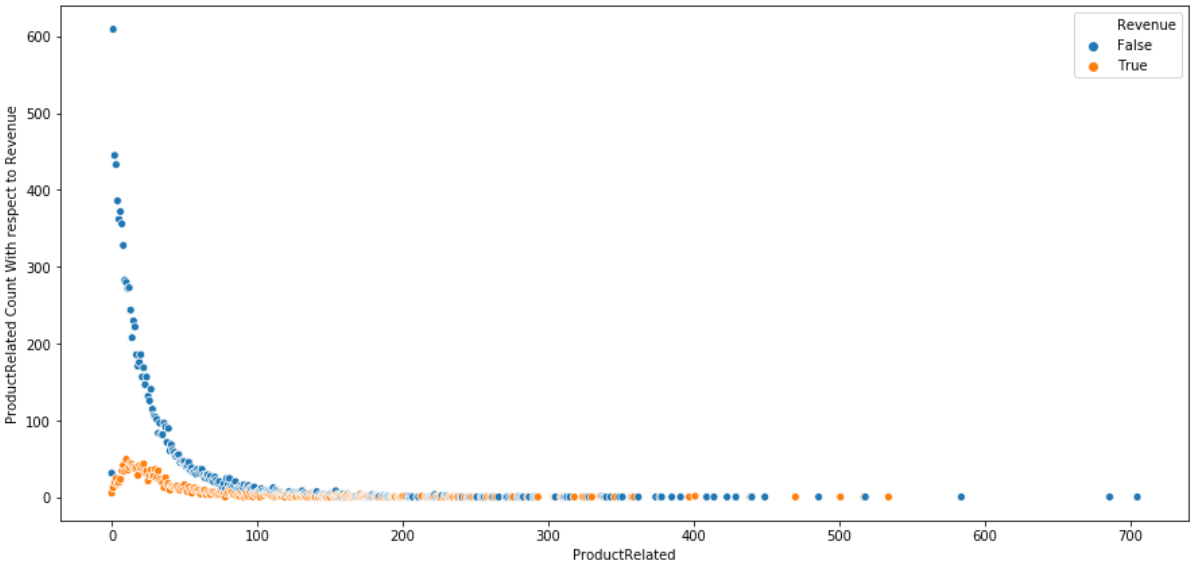
Gradient Boosting: We have got a testing accuracy of 91% and training accuracy of 87% with a precision of False as 92% and True as 91%. The recall for False obtained is 91% and True is 92%. We got the F1 score of False as 92% and True as 91%.

The confusion matrix obtained

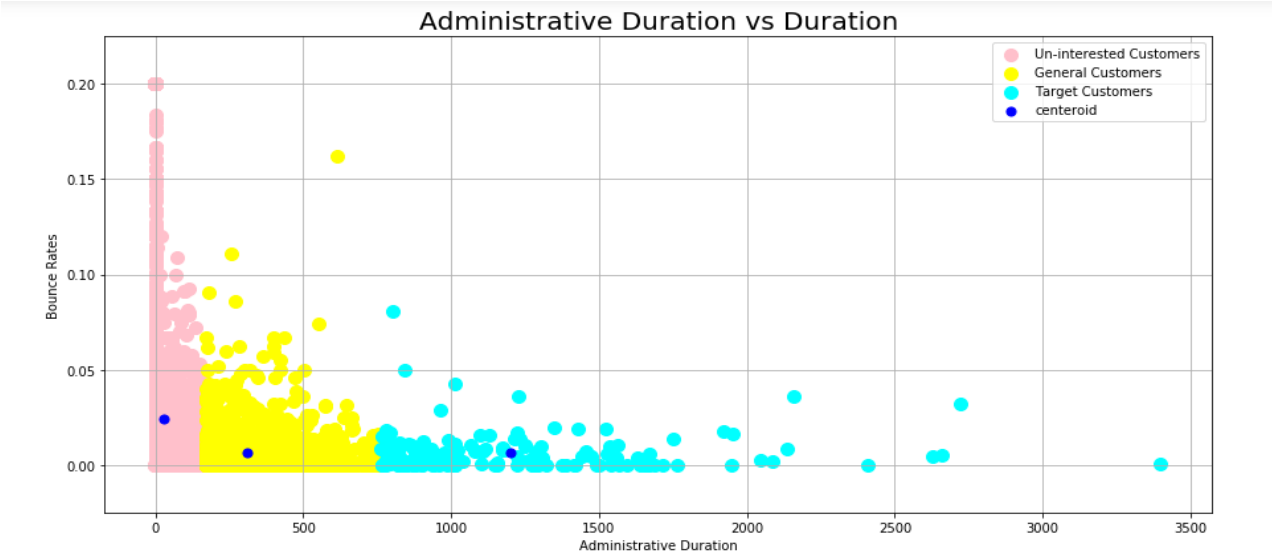
2814	263
257	2820

After applying both the techniques and obtaining accuracies of the models we can see that Gradient Boosting applied on WOE and IV gives better F1 score and confusion matrix than any other technique and model combination. Hence Gradient Boosting is so far the best model for prediction of Revenue.

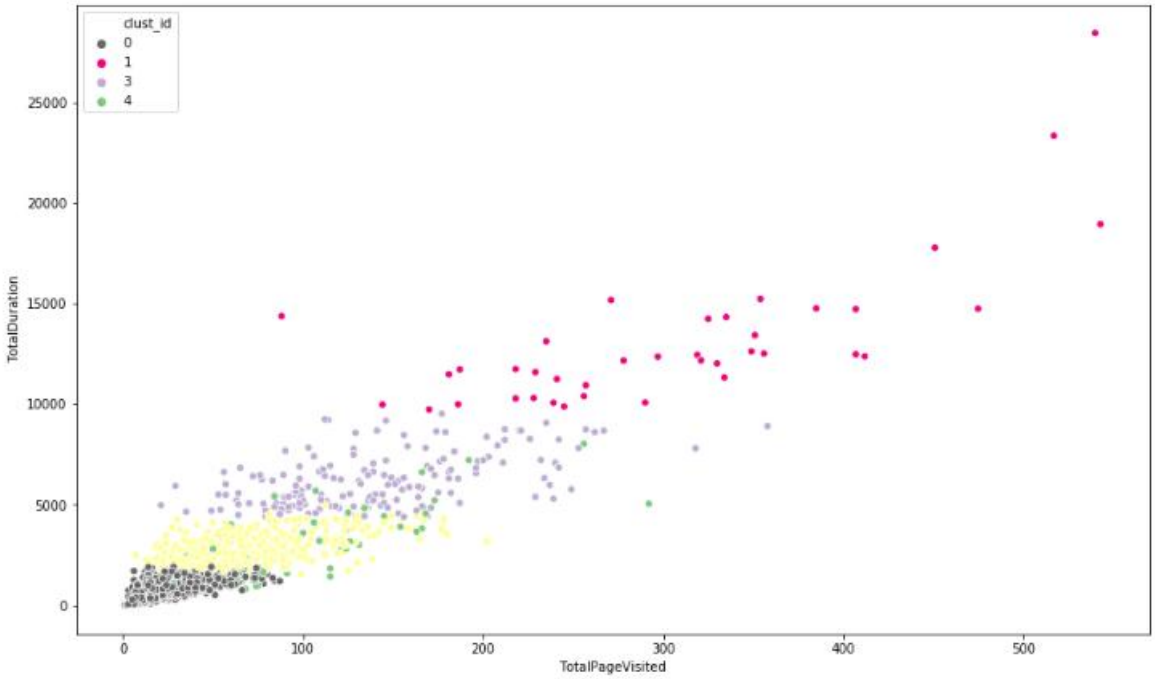
Patterns Inferences:



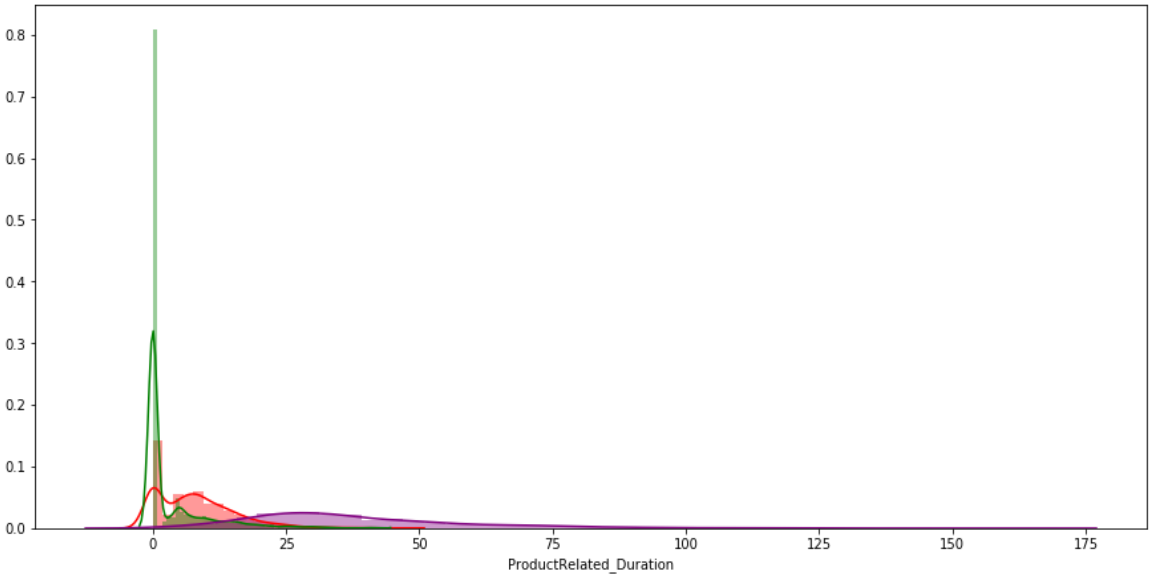
This scatterplot of ProductRelated and its duration with respect to revenue clearly show the distinction in their behavioral pattern. Most of the non-buyers tend to visit page between 0 to 100.This number of non-buyers is comparatively higher as compared buyers. As the page visits exceeds beyond 100 the count of number of buyers increases.

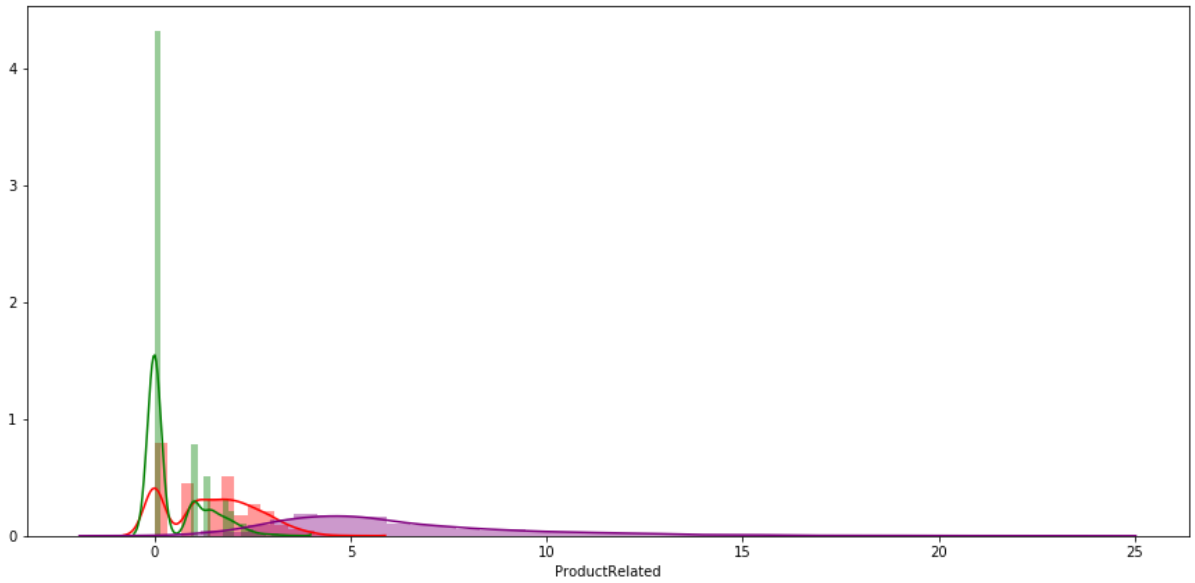


The scatterplot of Admin_Duration vs Bounce Rate reveal another user behavioral pattern. Many of the uninterested customer session had given less time in their session and they have corresponding high chances of bouncing i.e they tend to visit only single pages. But most of the buyers who had given more time have comparatively less Bounce Rates.

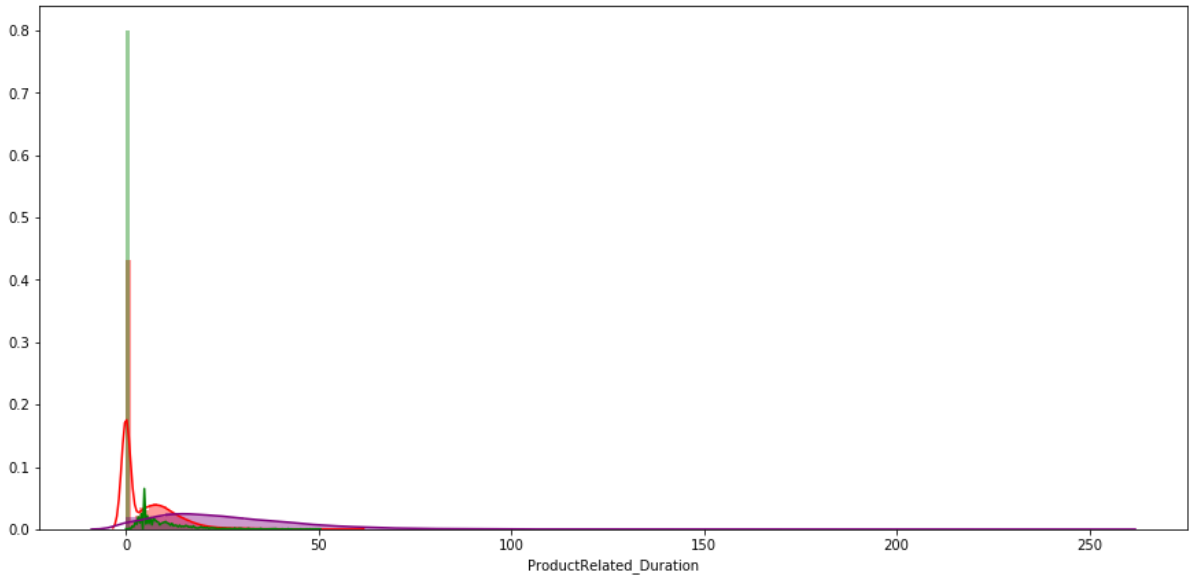


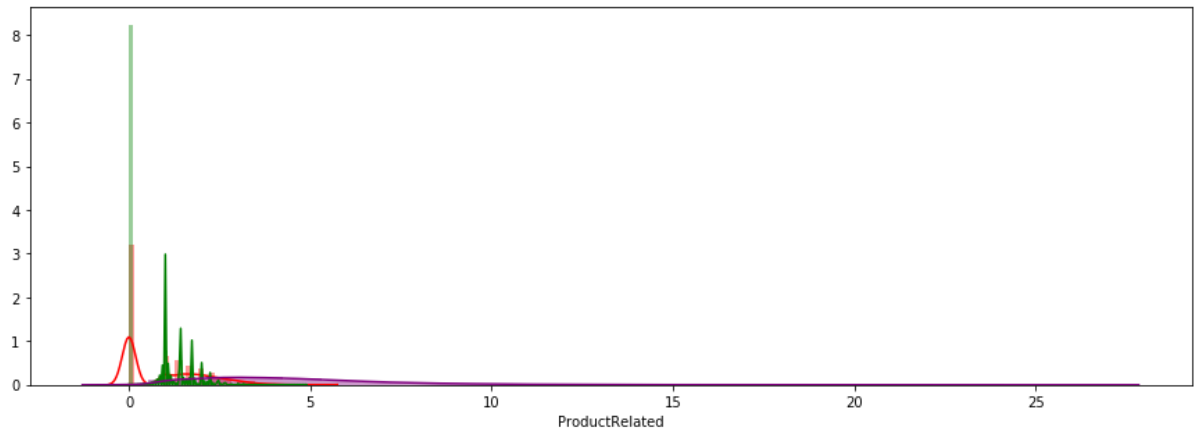
This scatterplot shows behavioral pattern on the basis of Time vs Frequency (Total Page Visited vs Total Duration) of all the buyers. Clusters were formed of user with similar behavioral pattern. The user in cluster 0 are specific and they are clear about which type of product they have to buy. Thus time given and corresponding Duration are less. User with behavioral pattern in the cluster 2 are highest in number. User in cluster 1 are tend to give more time and visits more pages but they are comparatively fewer in number who are not specific which kind of product they have to buy.





The above two graph represents the pattern of the page visits done by the Buyers. Administrative , Informational and ProductRelated pages are represented using red ,green and purple respectively .This pattern reveals that before going to ProductRelated pages most of the people are tend to give more time on Administrative than Informational which is further proceeded to ProductRelated which leads to generating Revenue.





The above two graphs reveals the behavioral pattern of non-buyers which clearly shows that most of the non-buyers are tend to visit more Informational pages than Administrative pages which further being proceeded to ProductRelated. Thus a session in which more time and visits is being given to Informational than Administrative then conversion are less likely to happen.