Project

Aashesh Vontivillu Nareshchand

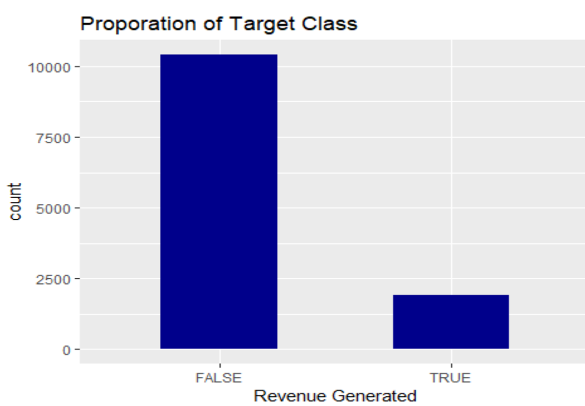
November 18, 2019

Online Shopper Purchasing Intention

Online shopping has made our life easy with purchasing the items done in minutes. But it is not everytime that we end up purchasing the item. Or in other words, we can say that there is no guaratanee that customer has the intention to purchase whenever he visits an ecommere website. The goal of this project is analyse the factors that help in determining the visitor purchasing intent and the predict if customer has purchasing intent or not given a new set of test attributes that has various Information related to customer behavior in online shopping websites. The outcome of the project can recommend the employers in targeting customers and help the employers in improvising the marketing strategies.

About the Dataset

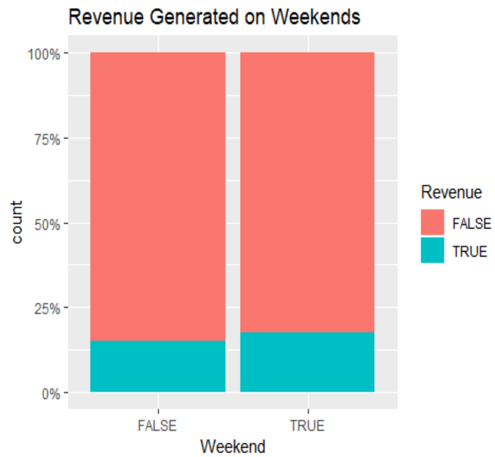
The dataset consists of feature vectors belonging to 12,330 sessions. Each session would belong to a different user in a 1-year period and any tendency to a specific campaign, special day, user profile, or period is avoided. The dataset consists of 10 numerical and 8 categorical attributes. The Revenue or Purchasing Intention attribute is used as the class label



Revenue True: 1909 observations, Revenue False: 12331 observations

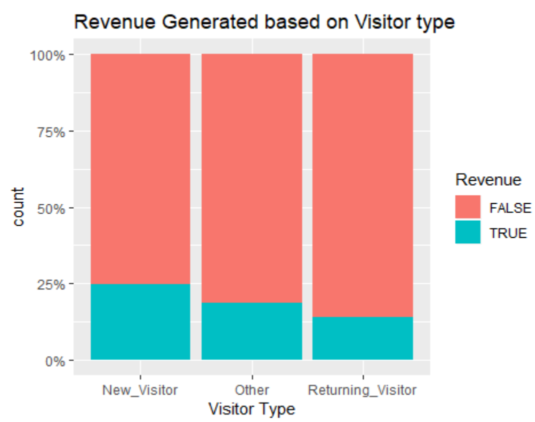
Proportion of the target class is 85: 15, which means the dataset is highly imbalanced.

Revenue or Purchasing Intention based on Weekends

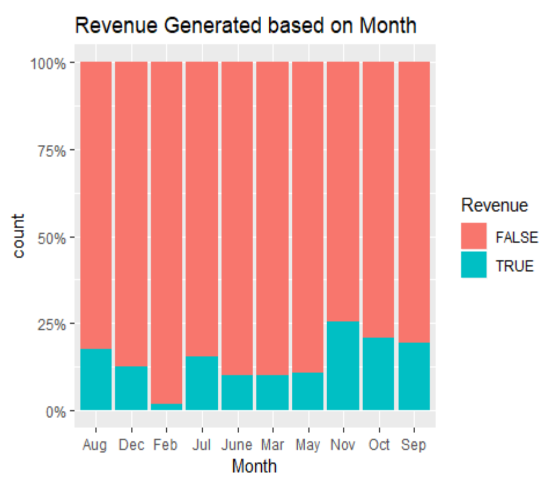


The revenue generated on weekends is slightly higher than the non-weekends. Out of all the user sessions, 17.4 % of the users ended up purchasing on weekends and 14.9% of the users ended up not purchasing

Revenue or Purchasing Intention based on Visitor Type & Month

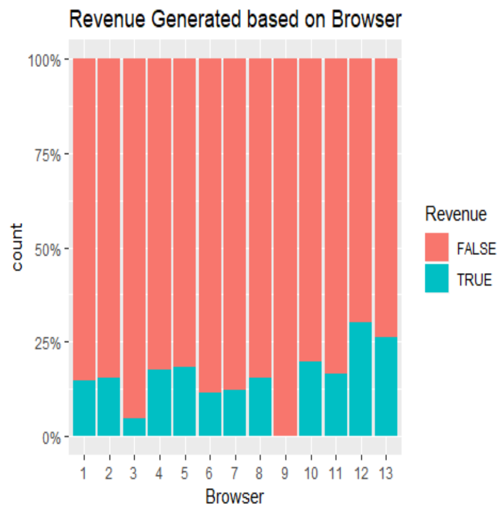


'New_Visitor' has more intention to purchase than the 'Returning' and 'Other visitors'. So, users are are creating an account and visiting the website for the sole reason to purchase the item.



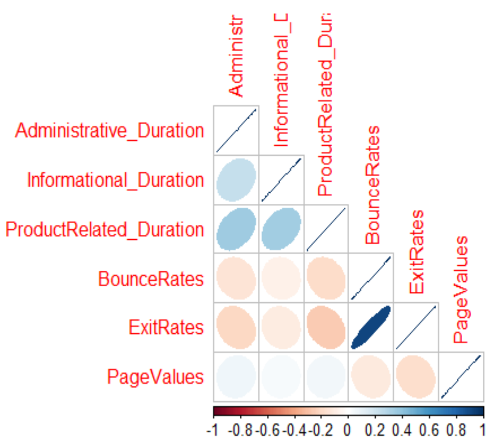
User intention's to purchase is the highest across November. And this is due to the week of Thanksgiving, where people tend to purchase more whenever they visit the shopping website. It is interesting to see users has the least intention to purchase in February, although there is valentine's day in February. Looks like a lot of single people out there.

<u>**Revenue or Purchasing Intention based on Browser**</u>



Users using the browsers 12 and 13 has more revenue compared to the other browsers used.

<u>**Correlation Plot**</u>



Bounce Rate and Exit Rate are highly correlated with 0.91. All the others feature have low to medium amount of correlation with each other.

<u>**Outliers Detection**</u>

Feature Administrative_Duration has 9% of its observations to be outliers

```
## [1] 1167 number of observations in the dataset
```

Feature ProductRelated_Duration has 7% of its observations to be outliers.

```
## [1] 960 number of observations in the dataset
```

Feature Informational_Duration has 19% of its observations to be outliers.

```
## [1] 2404 number of observations in the dataset
```

```
## Revenue Class or the Target Variable before Sampling
## FALSE  TRUE
##  7286  1336

## Revenue Class or the Target Variable after SMOTE Sampling
## FALSE  TRUE
##  7481  6680
```

__Metric to be considered for model evaluation - Recall__

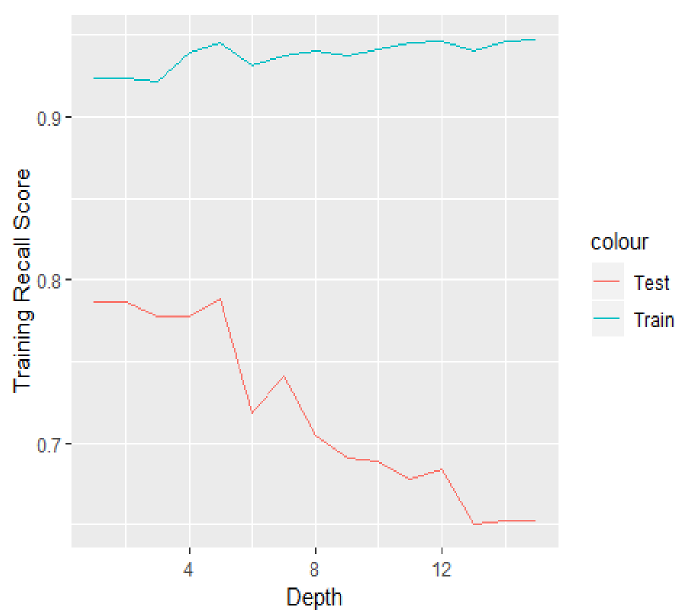
Why is Recall Score used here?

It is metric that determines how well the classifier was able to predict a specific target class. For this dataset, our class of interest is to determine and find the users having the intention to purchase (Revenue Feature is True). We do not worry much if user who is no intention to purhcase is classified as interested users as its misclassification cost is very low. The goal of the models is to know how well the model can generalise and predict the interested users (Revenue Feature is True)

Decision Tree

Implementing the Decision Tree Algorithm with various depths on training upsampled data and test data

Selecting the Optimal Depth:



As we see from the plot above the training recall score is reaching close to 1 with increasing depths. This is due to the below reasons

1) Decision Trees are very prone to overfitting as its depth increases.
2) Training data used in SMOTE sampled data.

The testing recall score reached its maximum at depth 5 and can be chosen as its optimal parameter

Note: *Decision Trees overfit on training data even with the use of cross validation with increasing depths.*

Implementation of Decision Tree using the Optimal Depth = 5:

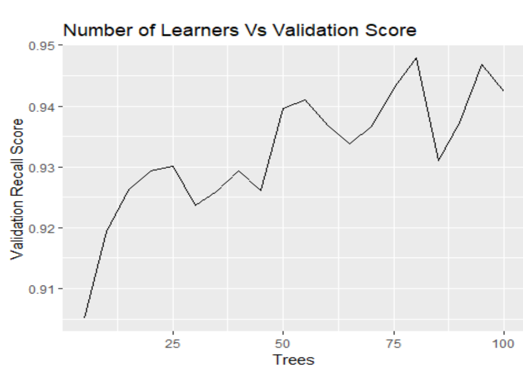
```
## Confusion Matrix

##        Predicted
## Actual  FALSE TRUE
##   FALSE  2771  351
##   TRUE    121  451

##
## Test recall score is 0.7884615
##
## Feature Importances

##            PageValues        Informational       Administrative
##           4575.425142          2300.679754          2263.219318
##             ExitRates       ProductRelated ProductRelated_class
##           1728.454116          1652.555562          1440.327478
##                 Month  Informational_class          BounceRates
##             68.469721            67.300469            28.472570
##           TrafficType              Browser     OperatingSystems
##             13.879644            12.263216             9.289387
##                Region          VisitorType Administrative_class
##              4.799135             2.565859             1.538462
```

Decision Tree can also be used as a feature selection process as the nodes in the trees are split based on the information and entropy provided by each of the features

RandomForest

Hyperparameter tuning of the RandomForest Algorithm using cross validation to find the optimal number of learners

Selecting the number of learners with cross validation:



Random Forest seems to be performing really well on the training data giving us a recall score of 0.95 when the number of trees used is greater than 70.

Unlike decision trees, random forests are not prone to overfitting as we are setting the depth of each tree at its optimal and is constant. Here, for training the random forest, maxnodes of 5 was used as the decision tree gave an optimal depth at 5. Random forest uses the maximum votin of the trees to perform its classification. And as the number of learners increases, we will not have any bias in the classification and each of the trees will contribute to the voting and hence this avoids the overfitting issue.

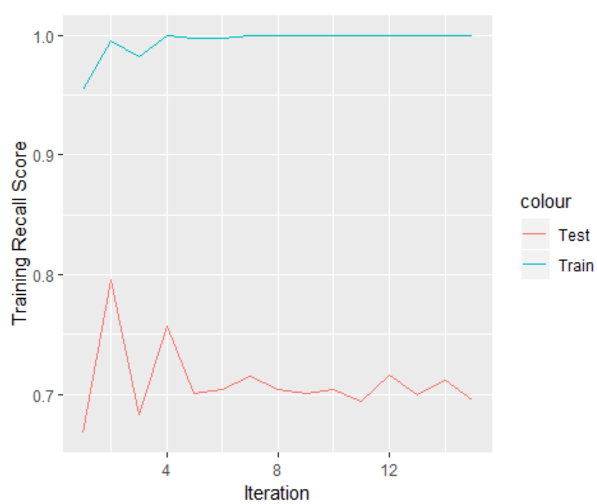
Implementation of Random Forest using the Optimal Number of trees =80:

```
## Confusion Matrix

##         Predicted
## Actual   FALSE  TRUE
##    FALSE  2568   554
##    TRUE    113   459

##
## Test recall score is 0.8024476
```

With Random Forest, there is a slight increase in the test recall score by close to 2% compared to the Decision Tree Model

AdaBoosting

Selecting the Optimal number of estimators of learners:



Optimal number of learners can be chosen as 2. The model is clearly overfitting to the training data. This is because on each iteration adaboosing gives more weight to the misclassified classes and duplicate them in order to learn better on the misclassified results.

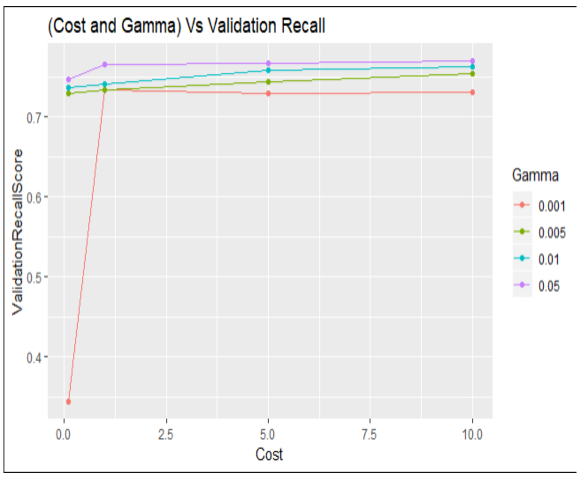
Implementing AdaBoost with the Optimal number of Learners =2:

```
## Confusion Matrix

##         Predicted
## Actual   FALSE TRUE
##    FALSE  2559  563
##    TRUE    117  455

##
## Test recall score is 0.7954545
```

Boosting the Decision Trees is increasing the test recall score by 1%.

Support Vector Machine

Hyperparameter tuning using cross validation to select the best Cost and Gamma



The plot shows the result of validation recall scores with training upsampled data. The optimal parameters are Cost: 1 and Gamma: 0.05

Implementing the SVM Radial Model with Optimal Cost and Gamma:

```
## Confusion Matrix

        Predicted
## Actual   FALSE TRUE
##    FALSE  2937  185
##    TRUE    194  378

##
## Test recall score is 0.6608392
```

Recall Score for SVM Radial seems to be very poor. The model is not generalising the data well and also data points are not distributed across the center.

Naive Bayes:

Close to 80% of the Page values are zeros. We create an indicator page value feature to check if the page value is zero or non-zero.

This will give us an idea on how important the page value is on interested customers

```
## Predicted Probability for PageValues_class

##        PageValues_class
## Y        Not Zero      Zero
##   FALSE 0.1160273 0.8839727
##   TRUE  0.6411677 0.3588323

## If the customer visited a zero page value, there is a probability of 89% t
hat the customer will not purchase anything.

## If the customer visited a non zero page value, there is a probability of 6
2% that the customer will purchase.

## Confusion Matrix

       Predicted
## Actual  FALSE TRUE
##   FALSE  1455 1667
##   TRUE     50  522

##
## Test recall score is 0.9125874
```

Naive Bayes is giving the best recall score of 0.90 when compared to the other models.

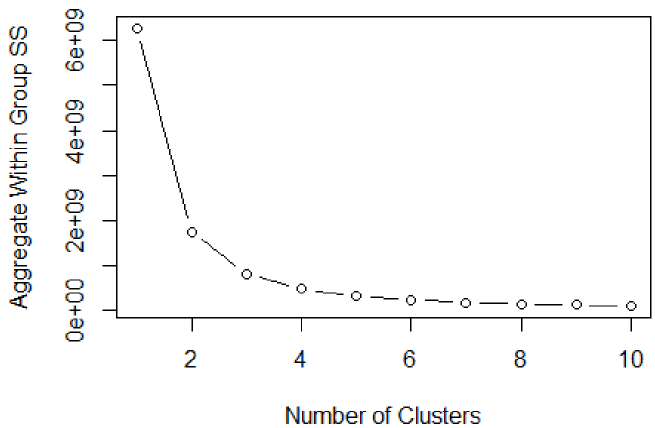
So, is Naive Bayes the best model for this dataset? Although, we had the best recall score, precision of this model seems to be too less as seen from the confusion matrix. Out of the 2200, that were predicted TRUE, only 516 of it were correct. We can find out the best model only we know other factors like the misclassification cost, marketing cost for each customer and profits generated by the customer if he ends up purchasing.

Clustering

Removing the outliers from features ProductRelated_Duration and Administrative_Duration as k-means is very sensitive to outliers.

Filtered the dataset with features Administrative_Duration','ProductRelated_Duration','BounceRates' to perform the clustering.

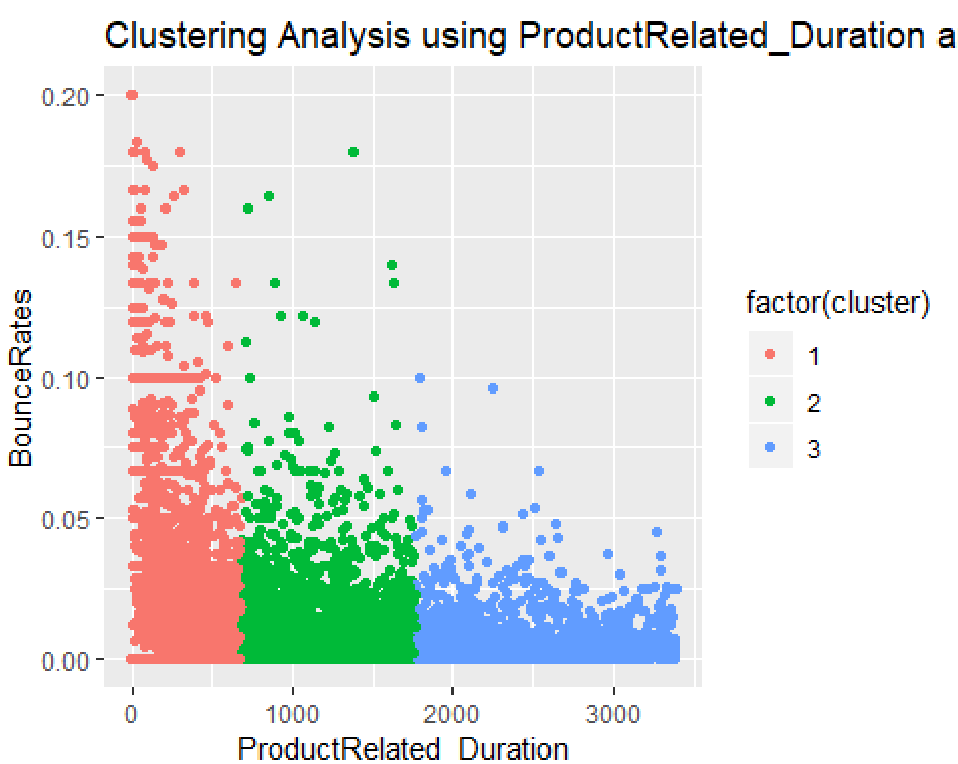
Elbow Plot to find the number of clusters



From the elbow plot, we can choose 3 as the number of clusters as the Within Sum of Squares seems to have reached its minimum saturation.

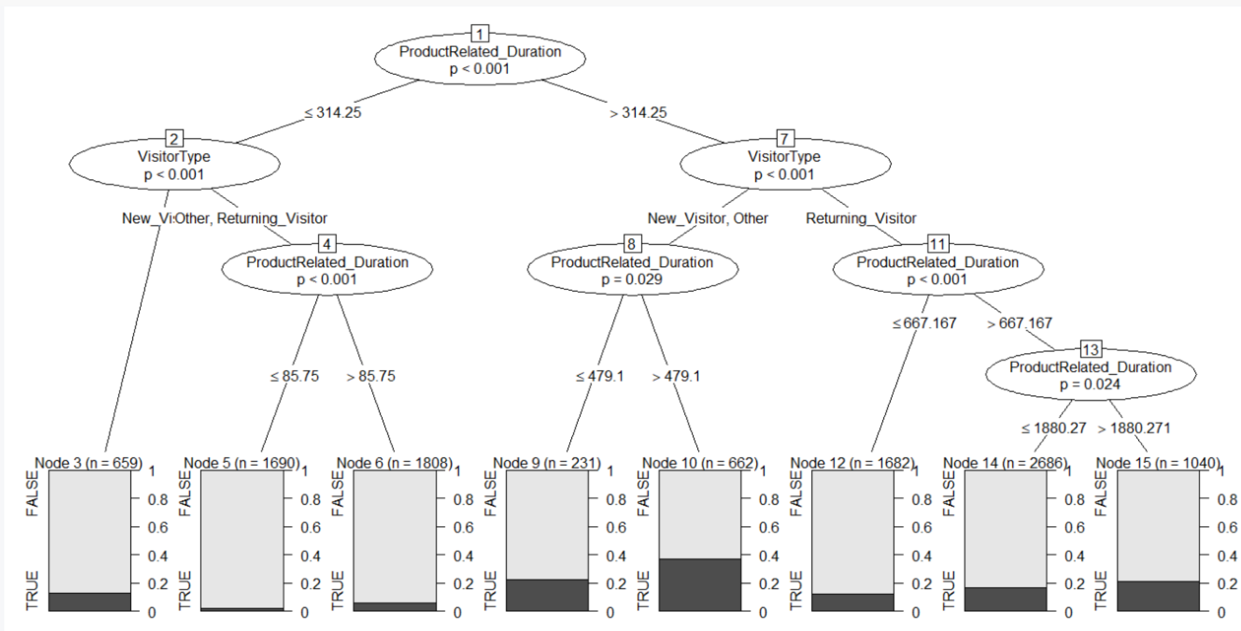
Clustering:

```
## Number of records in each cluster
##    1    2    3
## 6366 2850 1242
```



```
## Cluster Centres

##   Administrative_Duration ProductRelated_Duration BounceRates
## 1                25.92242                243.8954 0.034993808
## 2                48.77771               1131.1517 0.009498654
## 3                61.48693               2408.7304 0.007778928
```

Ctree Decision Tree



Interpretations: If the user is a new visitor and the product related duration is greater than 429, then there is an 40% chance that the user will purchase. ### Purchasing intention is the lowest when the user is a returning visitor and spends less than 85 on product related page

Comparisons of the Algorithms Implemented

Algorithm	DecisionTree	RandomForest	Adaboosting	SVM	NaiveBayes
Recall	0.788	0.802	0.795	0.660	0.912

Conclusion

1) Based on Recall Scores, Naive bayes algorithm seems to be performing the best. But we cannot conclude it as the best model for this data but we do not the other factos like misclassification cost,marketing cost, average profit when the users purchases an item

2) More marketing to be performed on New Visitors, users who spent more time on the Product Related pages and users who visit the pages with page value greater than zero.
3) SMOTE sampling the dataset helped the algorithms to perform better on this data
4) Model Performance can be improved when more data with the minority class is available or more relatable features are present in the dataset.