**Customer Purchase Behavior(Analysis by Vamsi Vivek Teja)**

**Business Problem Context:**

We are provided with information about 12,000+ online shopping sessions of users.

We are trying to:

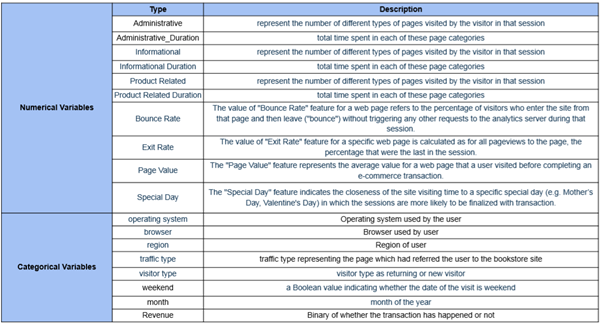
·        Predict if customers will end up generating revenue or not at the end of their session, based on their online behaviour.

·        Identify critical factors that would determine if a customer makes a purchase or not, and establish a relationship between how the changes in these factors impact the desired outcome.

**The Dataset:**

The [dataset](http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset) contains session level information of 12,330 user online sessions for a particular year.

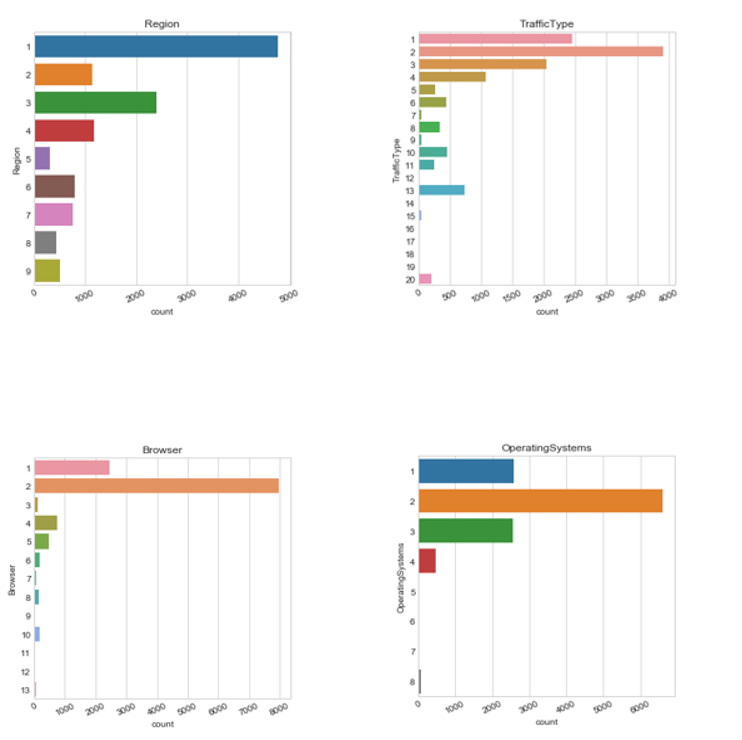
A summary of the variables in the data:



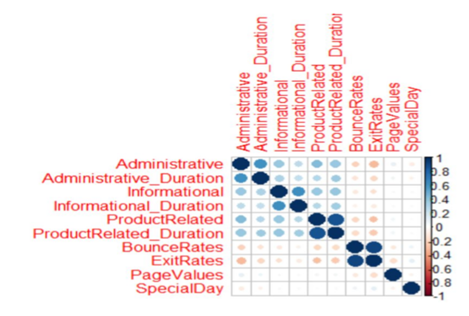
There are 10 numerical variables and 8 categorical variables. Our dependent binary variable of interest is ‘Revenue’.

**Data Analysis and Modelling:**

***Data cleaning and Feature Engineering:***

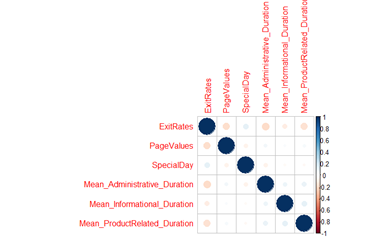


It is seen that there are categories with very few observations for the above variables. We group categories with low frequency and put them as ‘Others’.



We see there is a relatively high correlation between the number of pages visited and the time spent for the different page types (Administrative, Informational, ProductRelated).

To circumvent this, we combine each of these two columns and create three new columns for the mean duration spent on these types of pages in a session. These 3 variables will be used in our model:



The multicollinearity problem has been taken care of this way.

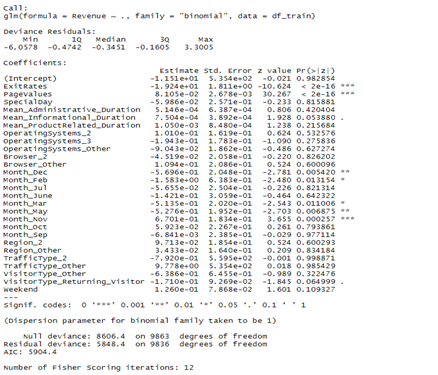
**Modelling:**

***Train-test split:***

We perform an 80-20 train-test split on the data.

***Base Logit Model:***

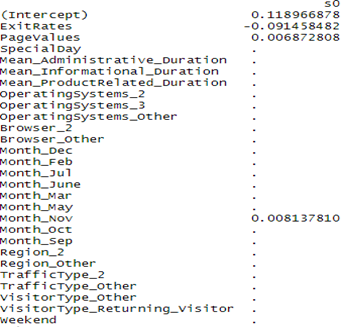
We create a logistic regression model including all the independent variables in it and revenue as the response variable.



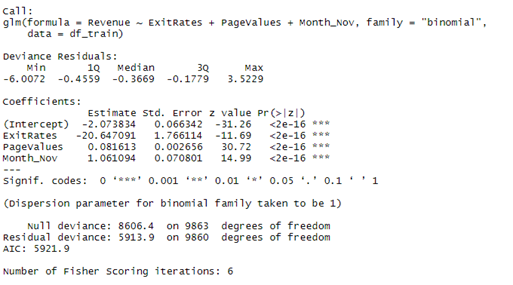
We obtain an accuracy of 0.89 and AUC = 0.68

***Lasso Regression for feature selection:***

Using Lasso Regression, we are able to reduce the number of significant variables to 3:

******

***Final logit model:***

******

Logistic model with three features. We then optimise the threshold and obtain a value of p=0.275.

We obtain an accuracy of 0.9 and AUC=0.77

**Model Interpretation:**

****

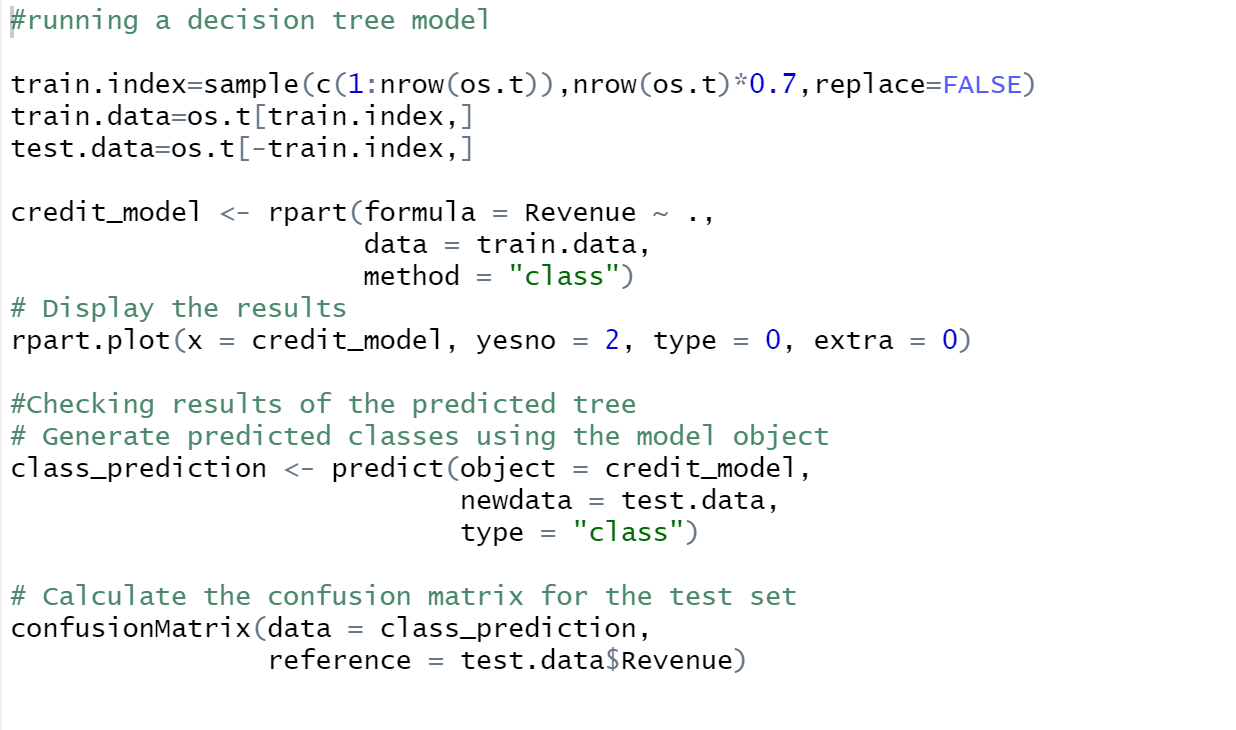
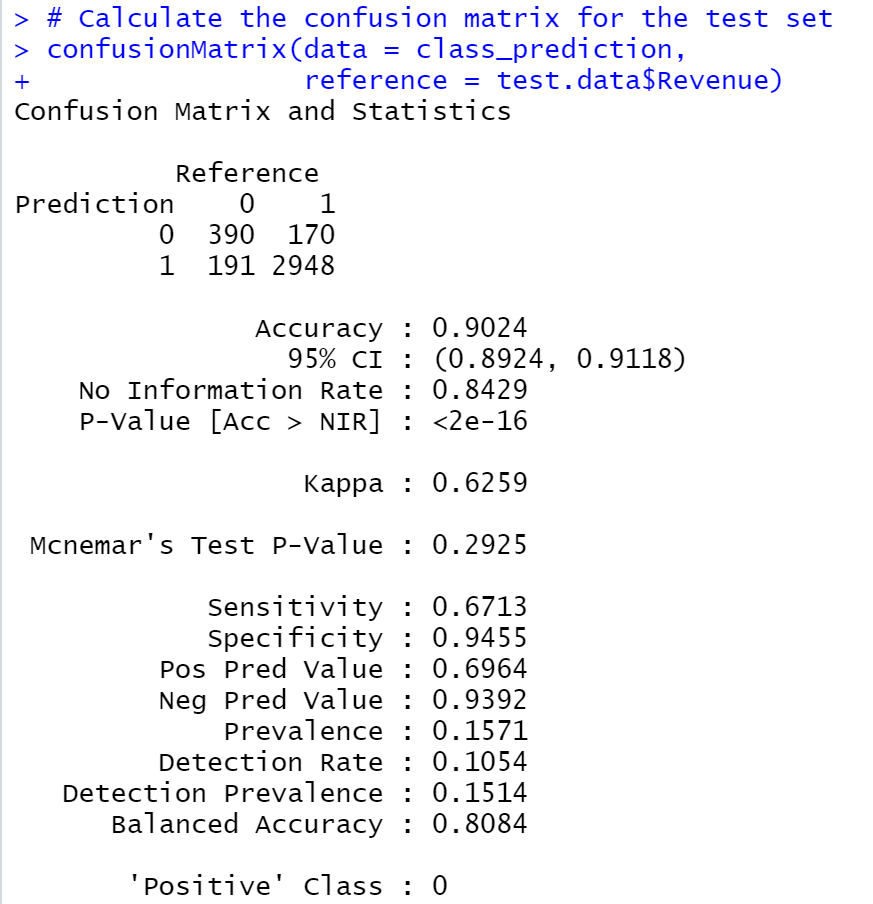
It is seen that Exit Rate, Page Values and Month\_Nov are the three factors that have a huge impact on a customer making an online purchase.

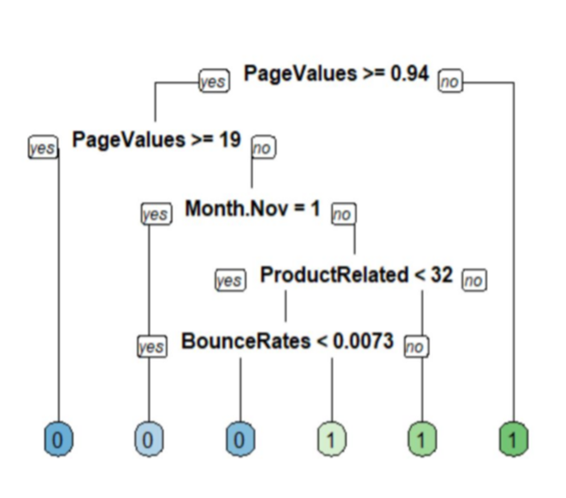
·        November has significantly more sales compared to other months. This might be because of Black Friday online sales, where a lot of customers will be looking to spend.

·        Odds of a purchase increases by 8.5% with every unit increase in page value. The business should look to increase and promote Page Value by SEO and other techniques.

·        Exit rate has a significant negative impact on revenue. Reducing the Exit rate by customer engagement and better UI can increase the chances of making a transaction.

***Decision Tree Classification:***

We build a decision tree model and obtain a classification accuracy of 90.3% on test data. Page value is the most important variable that impacts the outcome. Other variables in the model are Month\_Nov, ProductRelated and BounceRates. 



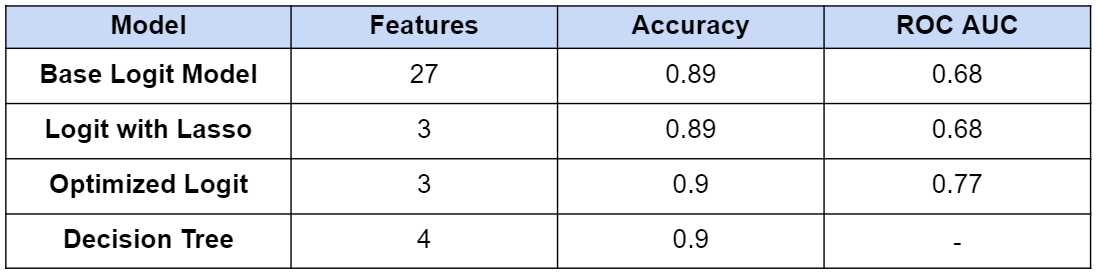
Interpretation:

·        If Page Value is lower than 0.94, then we predict that the customer will purchase.

·        If Page Value is between 0.94 and 19 and the month is November, then the transaction will not happen. For other months, it depends on Product Related and Bounce Rate values

·        If Page Value is greater than 19, then there is no purchase done.

**Summary:**

****

·        Our initial logistic model with 27 features was able to generate an accuracy of 89% on test data.

·        Lasso is then used to reduce the number of features to 3

·        We then use optimization to find the best threshold value, and obtain p=0.275. Using this in the model, we obtain an accuracy of 0.90 and AUC of 0.77.

·        The decision tree model increases the accuracy further and we obtain a score of 90.4%