BUAN/MIS 6356:004 Business Analytics with R – Group Project Team 2

Objective:

This dataset contains the information about online shopping done by the customers and our objective is to increase the profitability of the company by generating the insights of customers who did shop from the website and who did not and finding out the reasons.

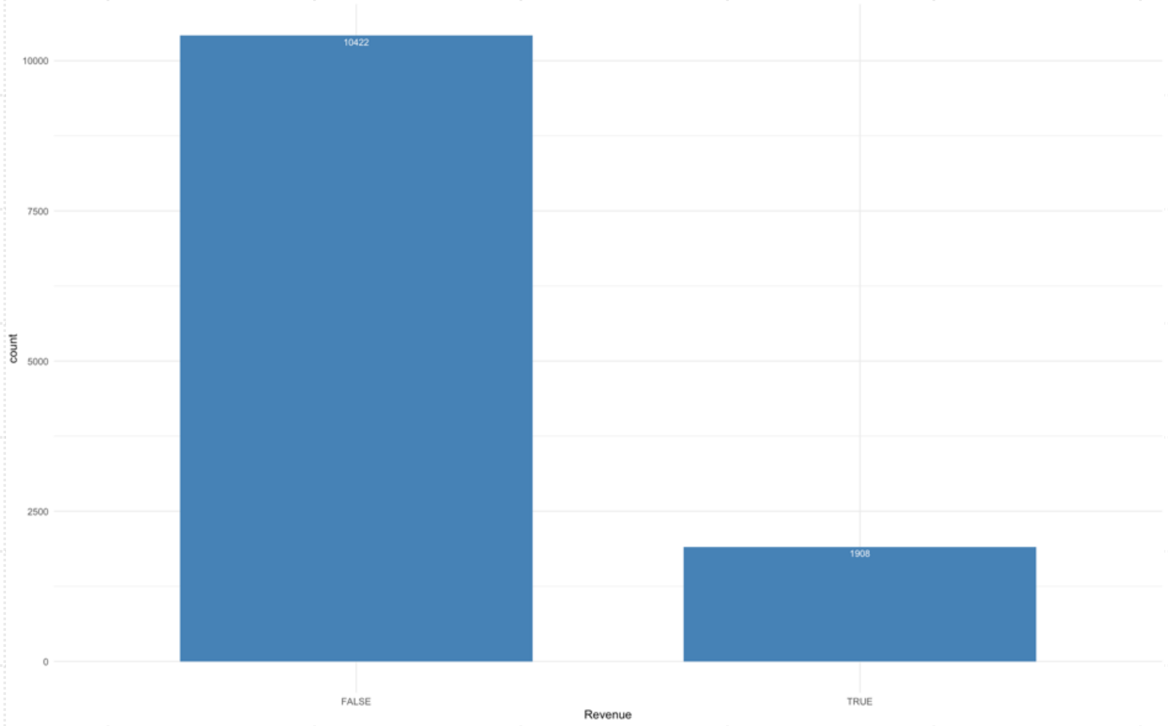
Insight Generation Points:

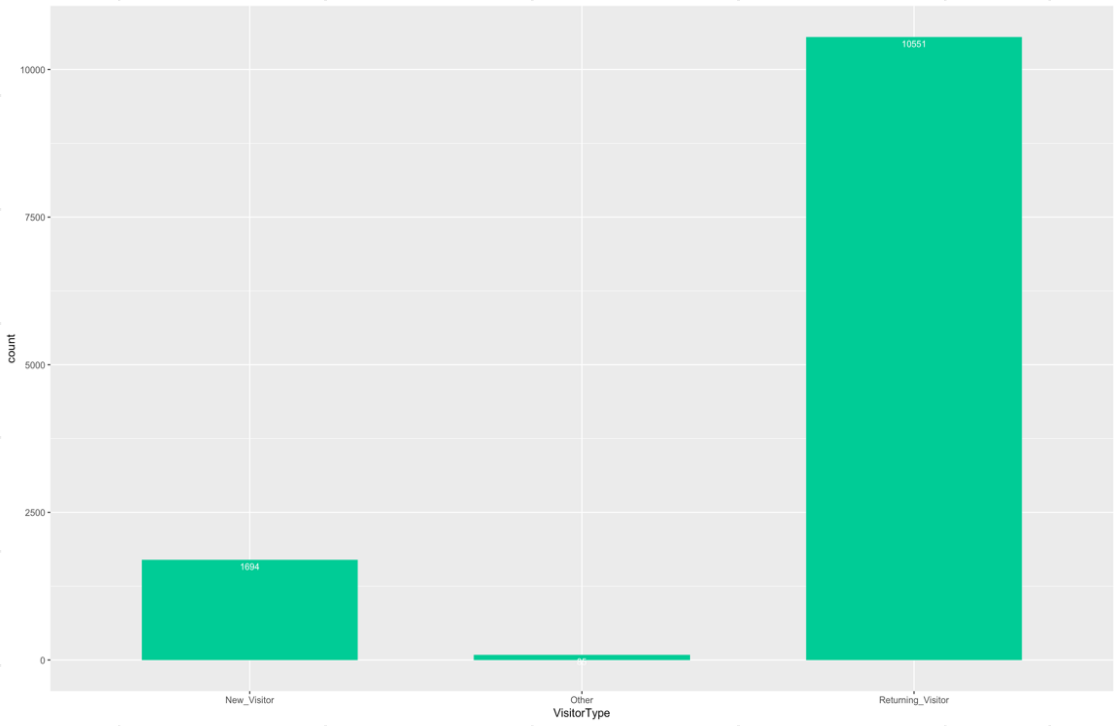
- Clustering of data of customers who did shop from the website and who did not
- Time spent by customers on website
- Association to find the relation between existing or new customer pattern along with their weekend, weekday, special day shopping trend
- Bounce rate Trend and month wise traffic insights

Attribute Information

- The dataset consists of 10 numerical and 8 categorical attributes.
- The 'Revenue' attribute can be used as the class label.
- "Product Related" and "Product Related Duration" columns contains number of different of pages visited by the customer in that session and total time spent in each of these pages.
- The "Bounce Rate", "Exit Rate" and "Page Value" columns represent the metrics measured by "Google Analytics" for each page in the e-commerce site.
- The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction.
- The "Special Day" column represents the site visiting time to a specific special day (e.g., Valentine's Day) during which customer is more likely to shop.

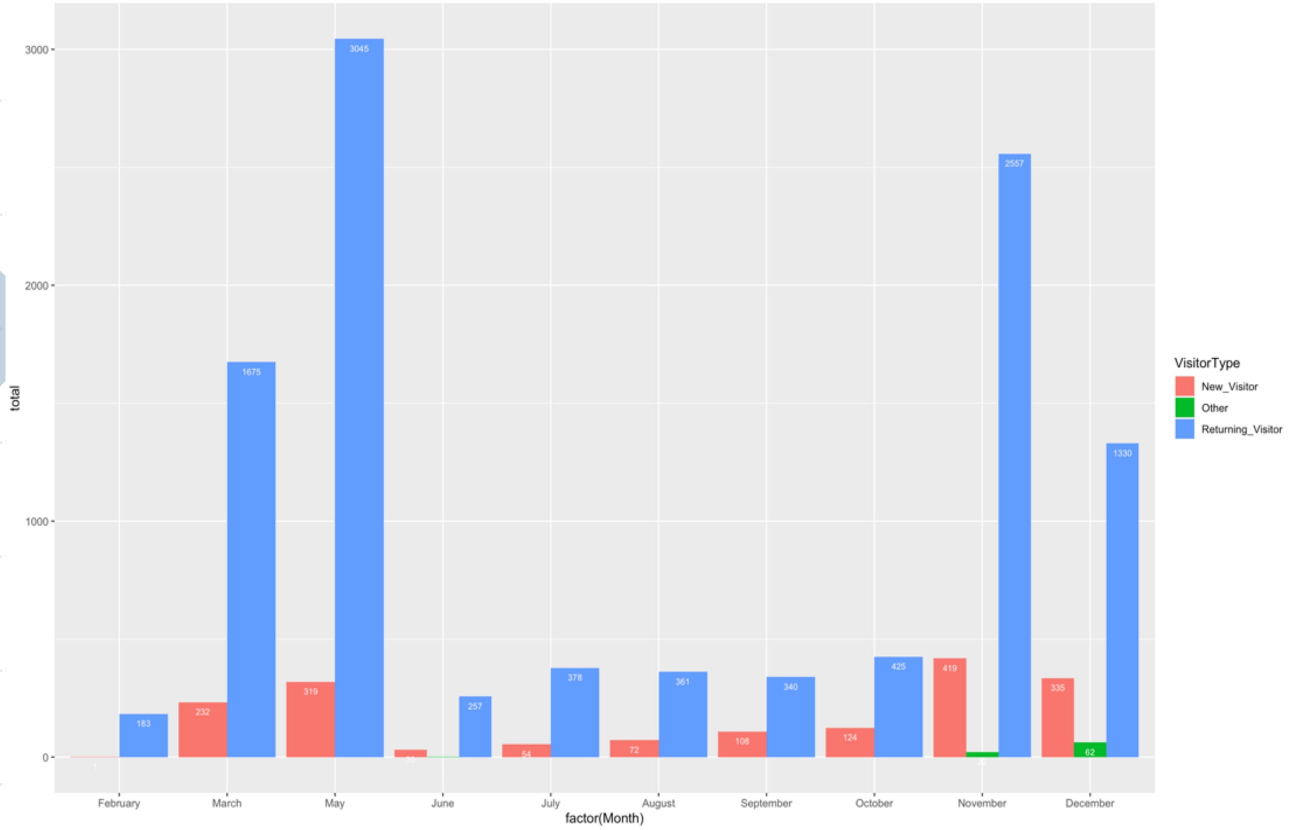
Exploratory Data Analysis

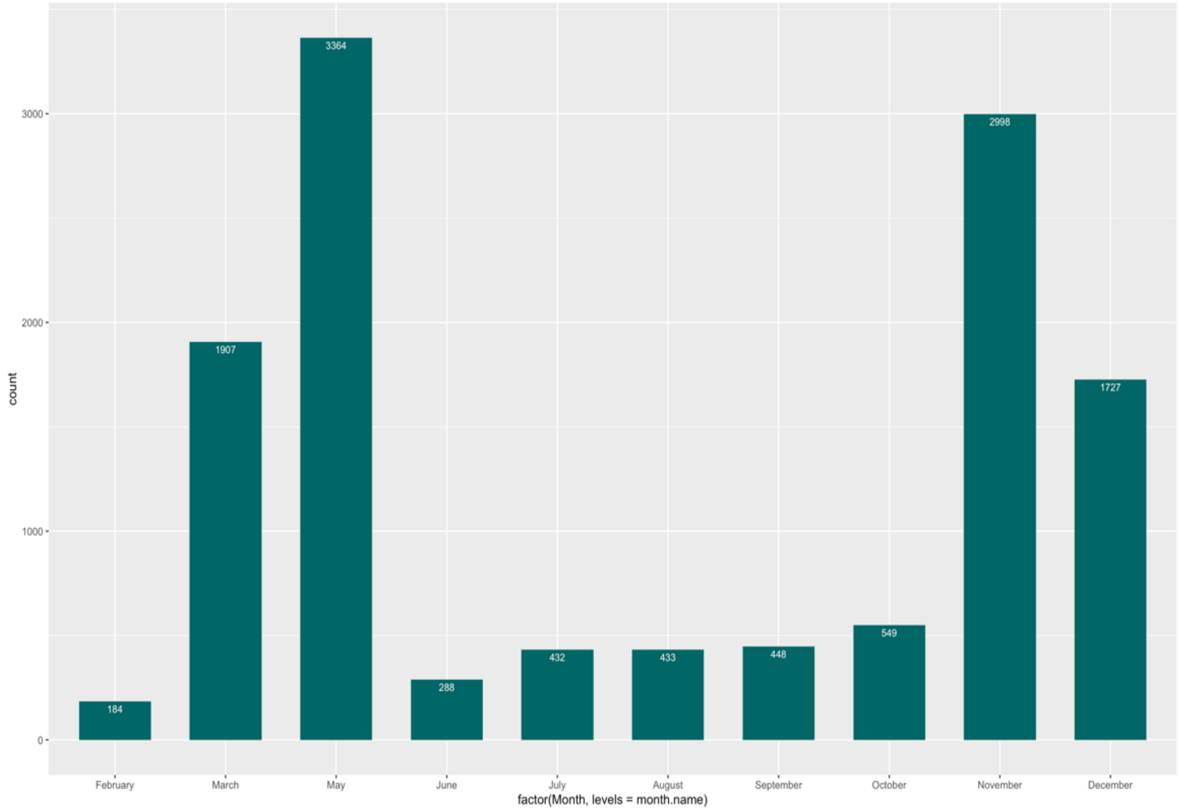




Revenue Wise Count:
Visitors: 10,422
Buyers: 1908

Visitor Wise Count:
New Visitors: 1694
Returning Customers / Visitors: 10551



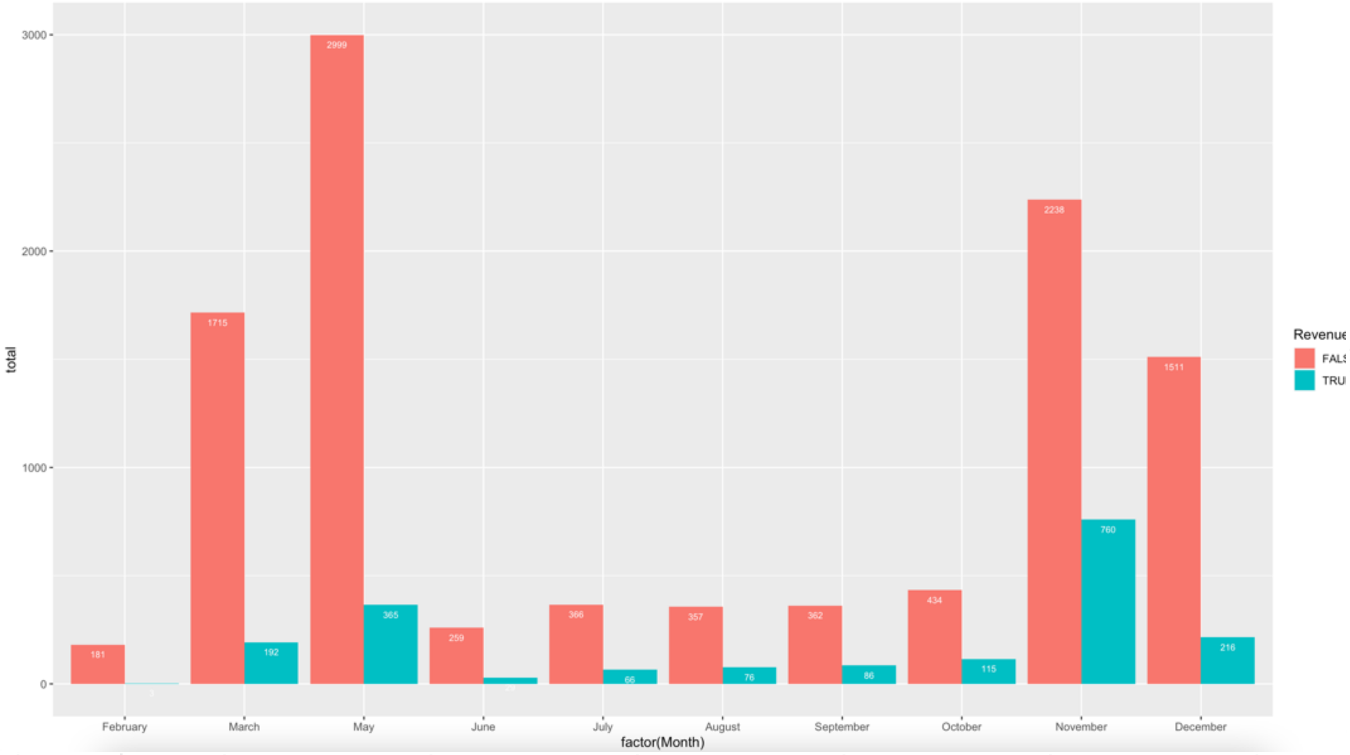


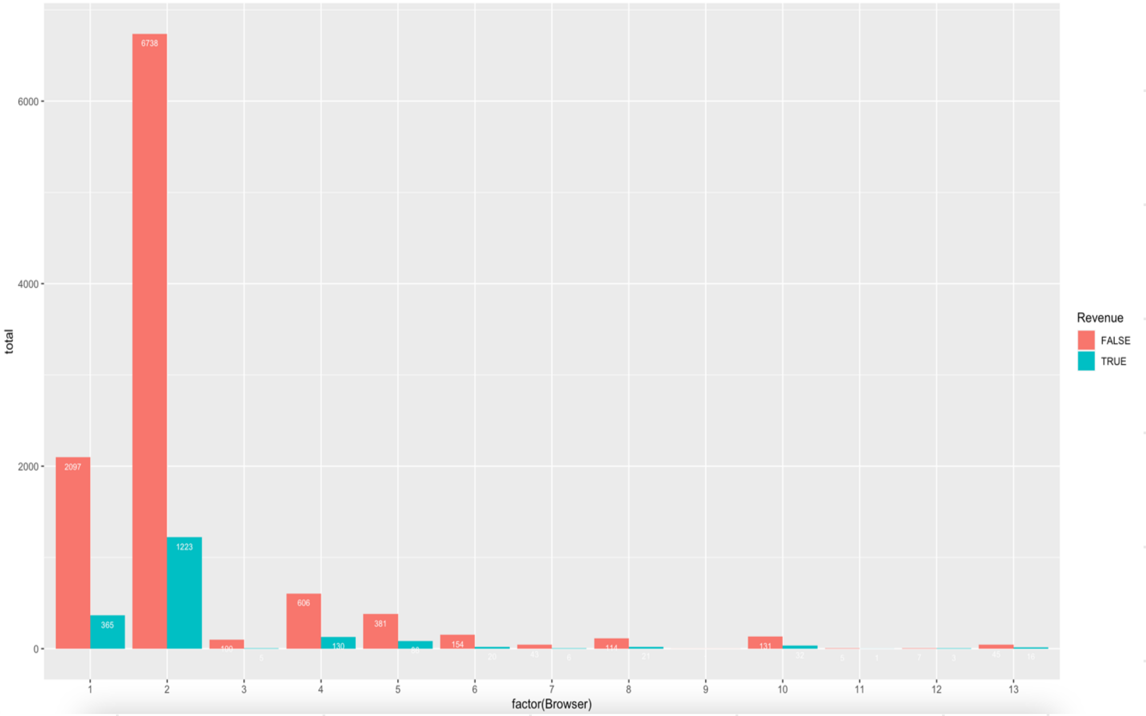
Month-Wise Visitors:

- Frequency of returning visitors greater than new visitors.
- Frequency of visitors is significant during March, May and November, December.

Month-Wise Count Plot:

- High Traffic during Summer months and Festive periods



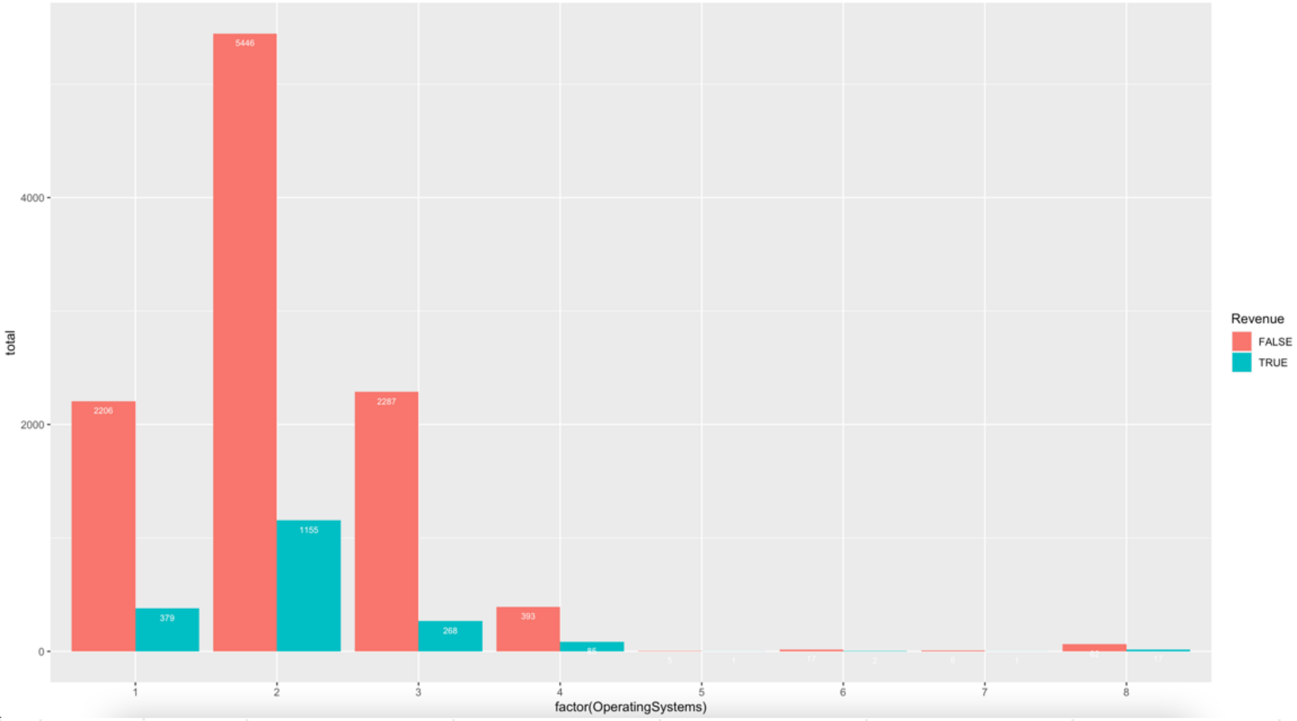


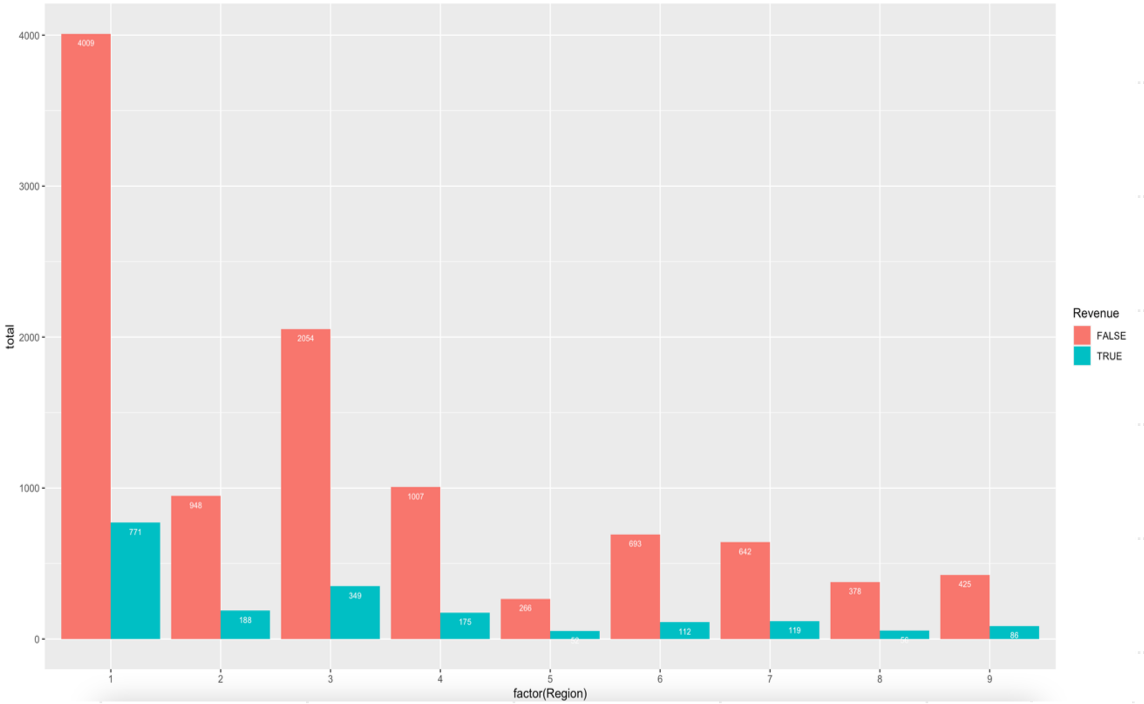
Month wise revenue count:

- Frequency of purchases in terms of the revenue generated.
- Orange Bars: People who visited the website but did not contribute to the revenue.

Browser wise revenue count:

- Plot is indicative of the fact that based upon the browser factor.



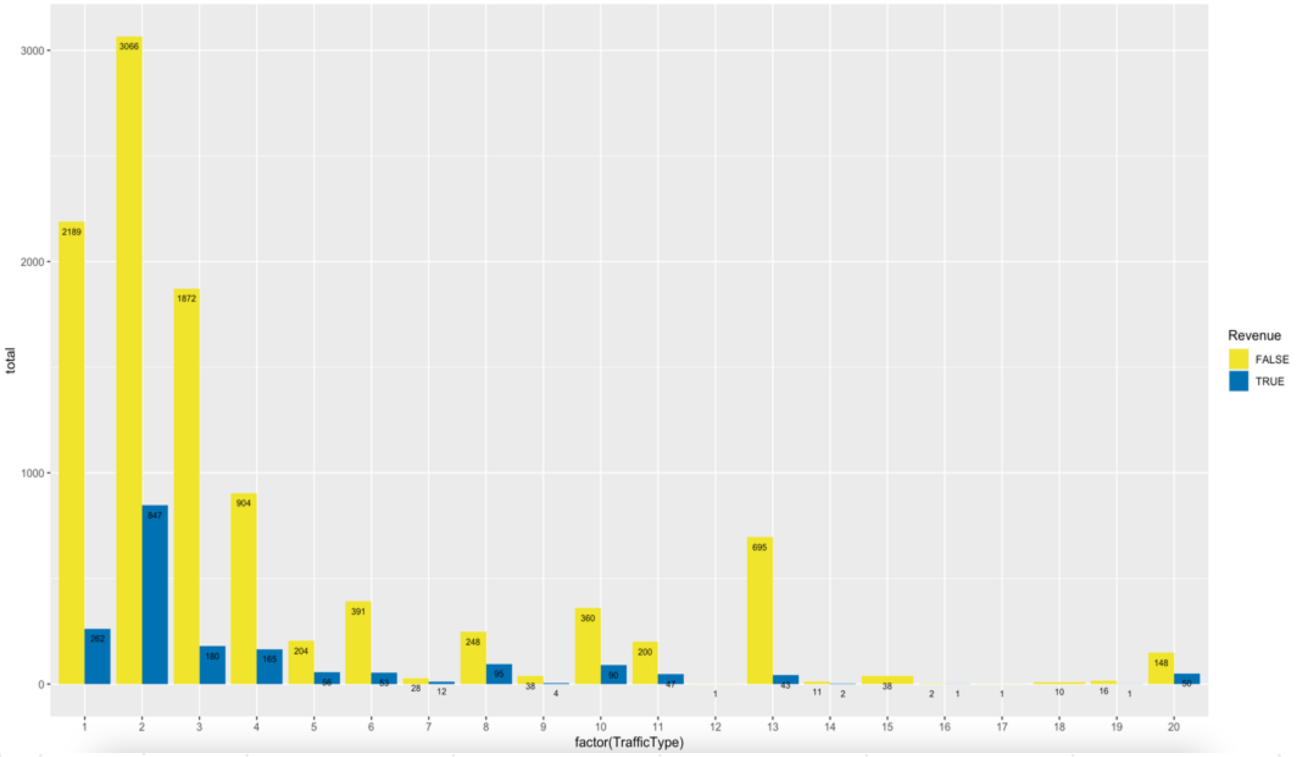


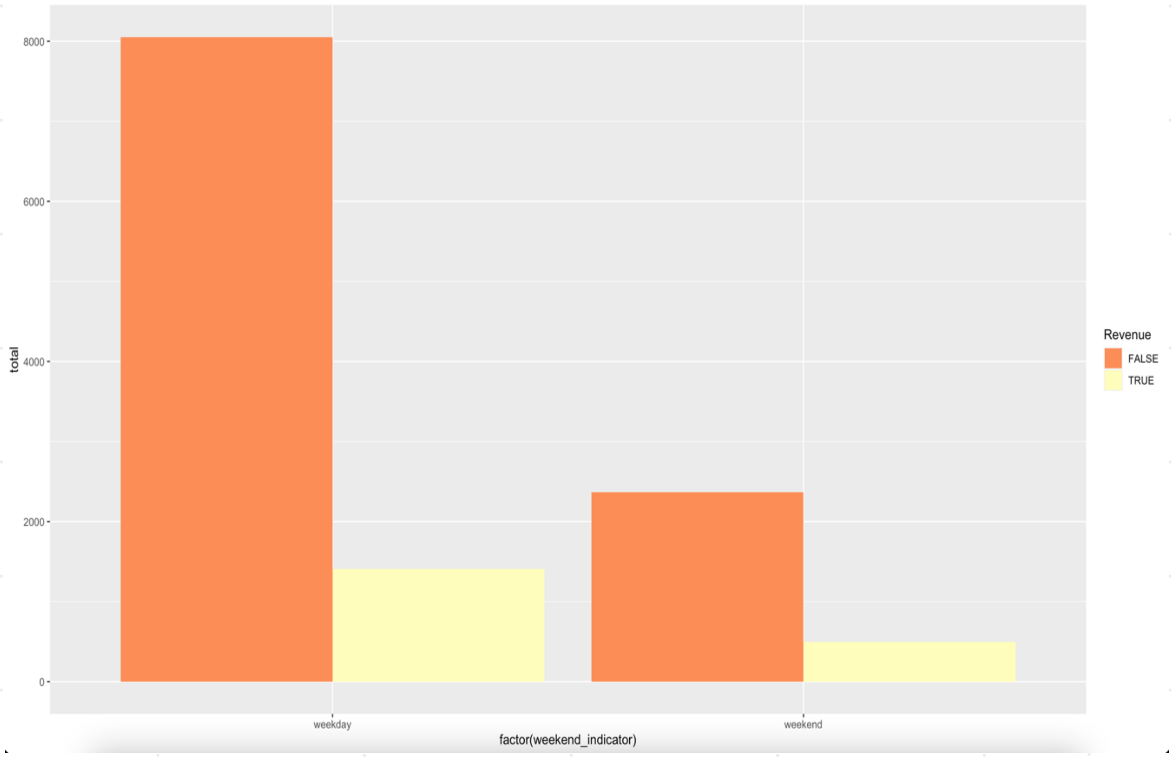
OS wise revenue count:

- Factor 2 has the highest revenue and highest traffic of visitors which did not contribute to the revenue.

Region wise revenue count:

- The variance of revenue in terms of region factor.





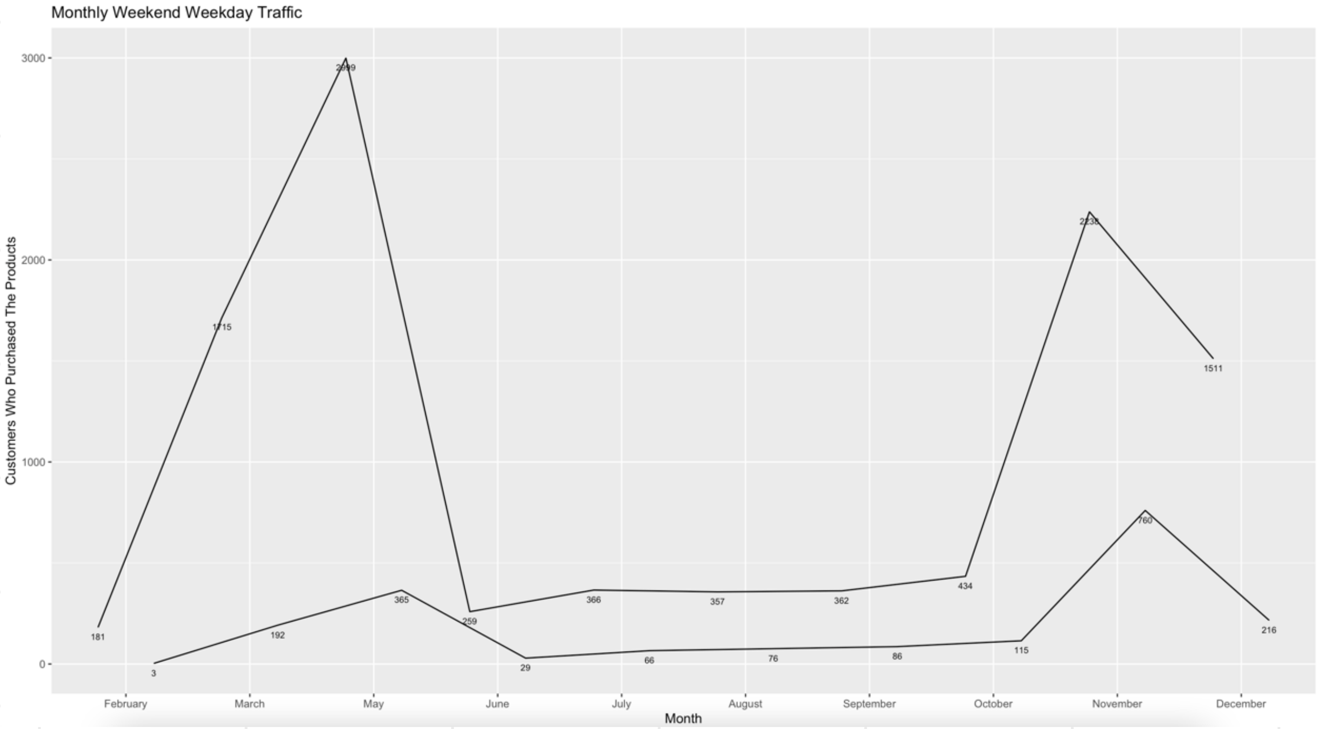
Traffic wise revenue count:

- Visualization of the revenue generated in terms of the traffic type factor.

Weekend-Weekday wise revenue count:

- During the weekday, higher revenue is generated as compared to over the weekend.

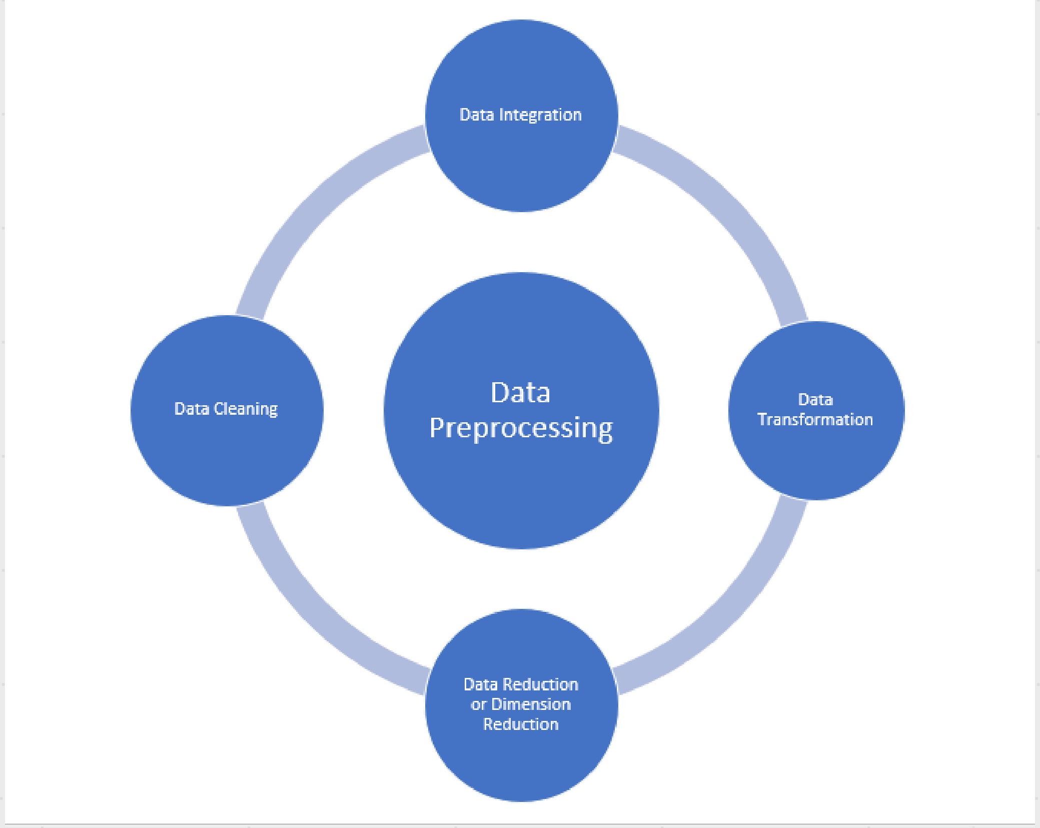
Monthly Weekend-Weekday Traffic



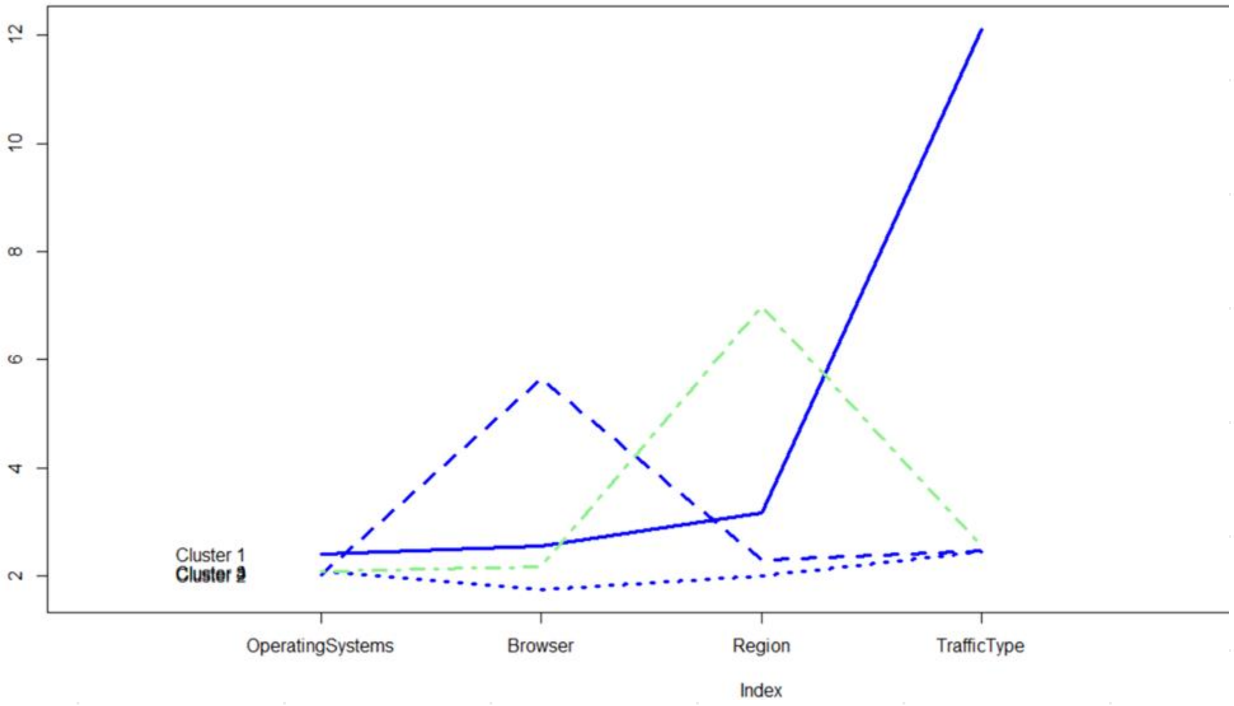
Seasonality Trend is observed for the weekend, weekday customer visits

Data Pre-processing steps

- Recoded variables ‘TrafficType’ and ‘Browser’ since there were too many values at factor level and to reduce the categories.
- Training And Validation split is kept as 65% and 35% respectively
- Data is scaled with ‘center’ , ‘scale’ technique to bring all the data points on the same scale



Clustering Analysis



Clusters:

- Operating Systems
- Browser
- Region Index
- Traffic Type.

01

Cluster 1: Remains between 2 to 4 for OS and Browser, but steeply rises for the Region and Traffic Type parameters.

02

Cluster 2: Remains between 2 and 4 for OS, but steeply rises for Browser, thereafter, decreasing again for Region and Traffic Type.

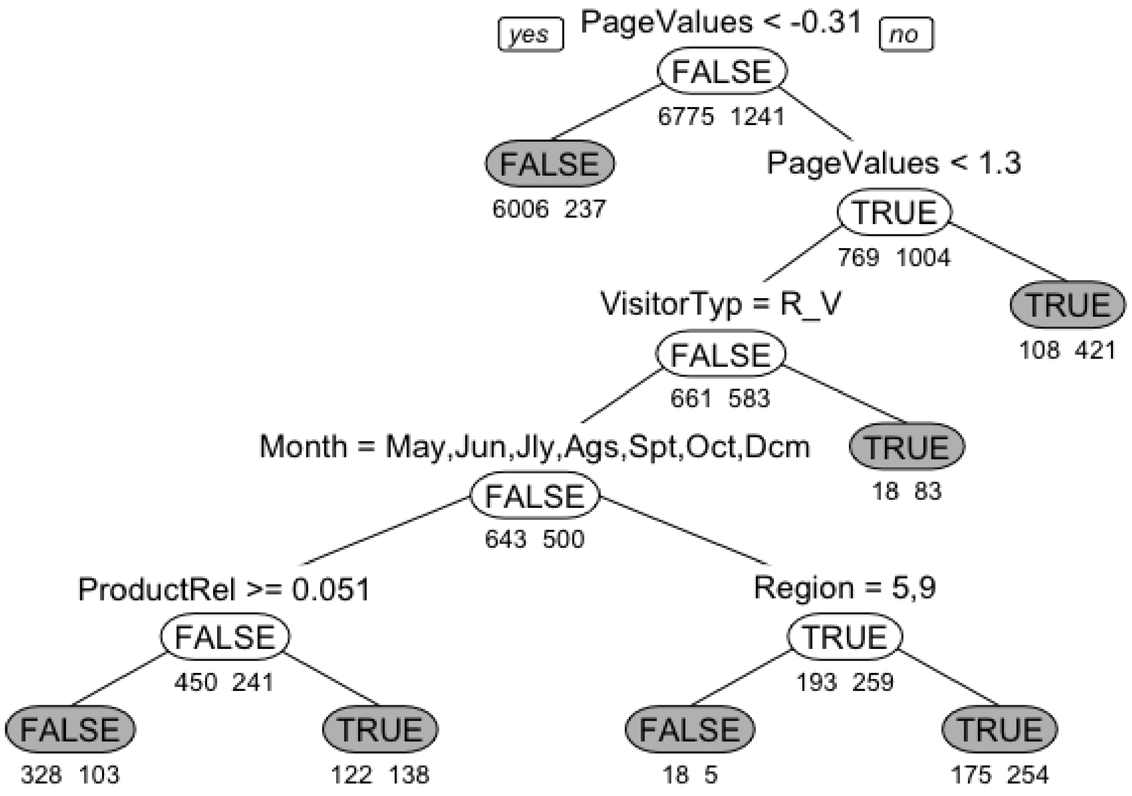
03

Cluster 3: Falls between 2 to 4 for OS but increases for browser and decreases again for Region and Traffic Type.

04

Cluster 4: Remains between 2 and 4 for all parameters except browser, for which it falls below 2 as well.

Decision Tree Model



Split Based On :

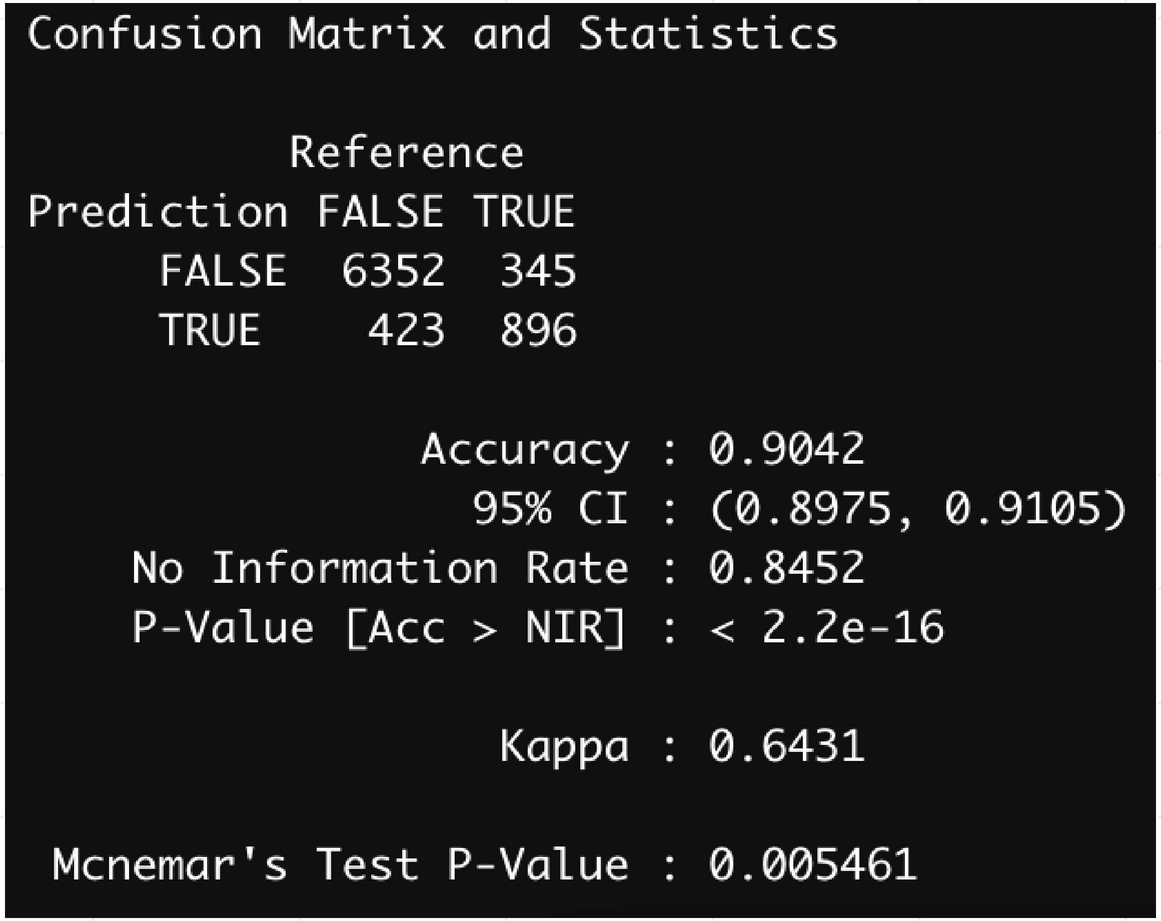
- Page Value
- Visitor Type
- Month
- ProductRel
- Region

Decision Tree Leaves: 7

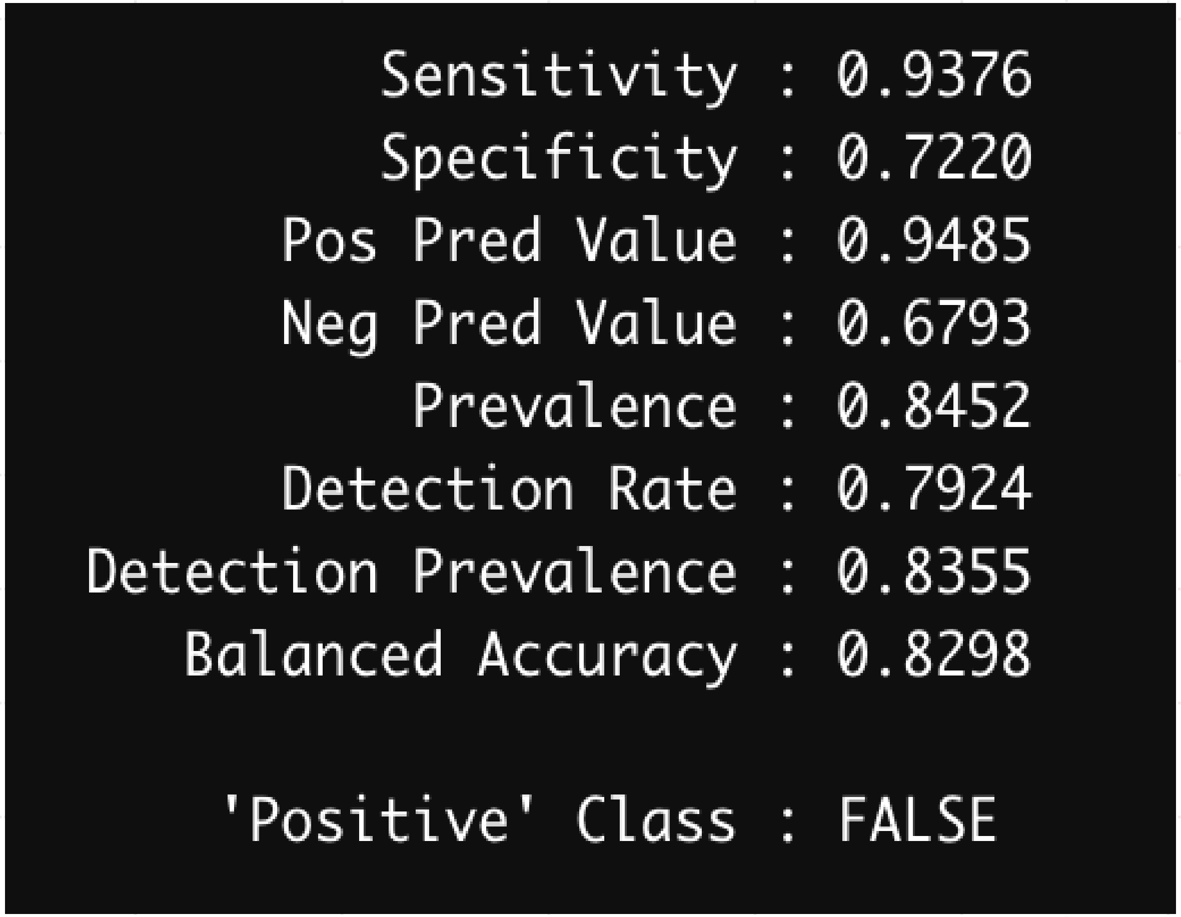
Shows if revenue is generated or not

Decision Tree Model

- **Confusion Matrix for Training Data – (Summary)**

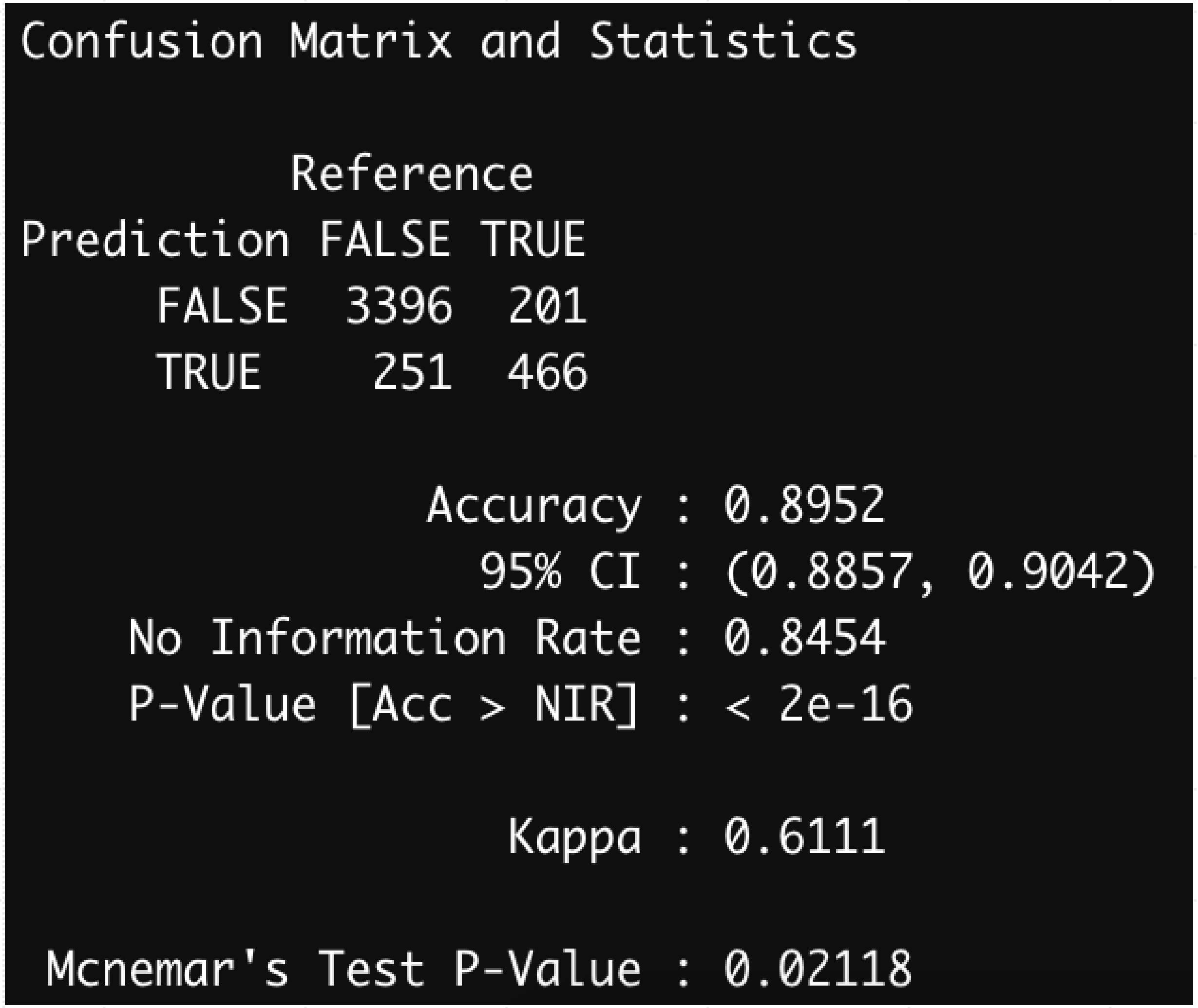


```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  6352  345
     TRUE    423  896

               Accuracy : 0.9042
                 95% CI : (0.8975, 0.9105)
    No Information Rate : 0.8452
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6431

 Mcnemar's Test P-Value : 0.005461
```

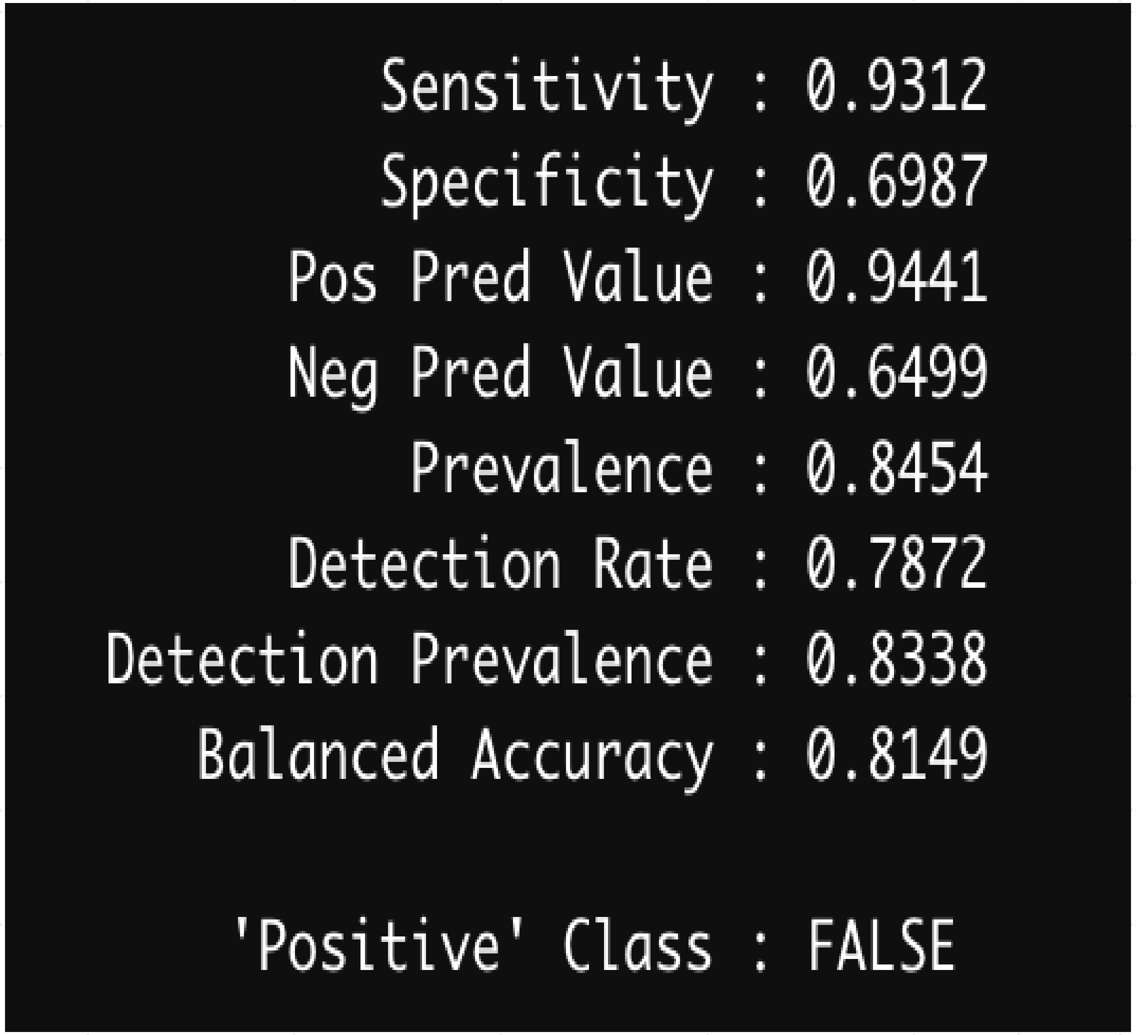


```
            Sensitivity : 0.9376
            Specificity : 0.7220
         Pos Pred Value : 0.9485
         Neg Pred Value : 0.6793
             Prevalence : 0.8452
         Detection Rate : 0.7924
   Detection Prevalence : 0.8355
      Balanced Accuracy : 0.8298

       'Positive' Class : FALSE
```

Confusion Matrix for Validation Set – (Summary)

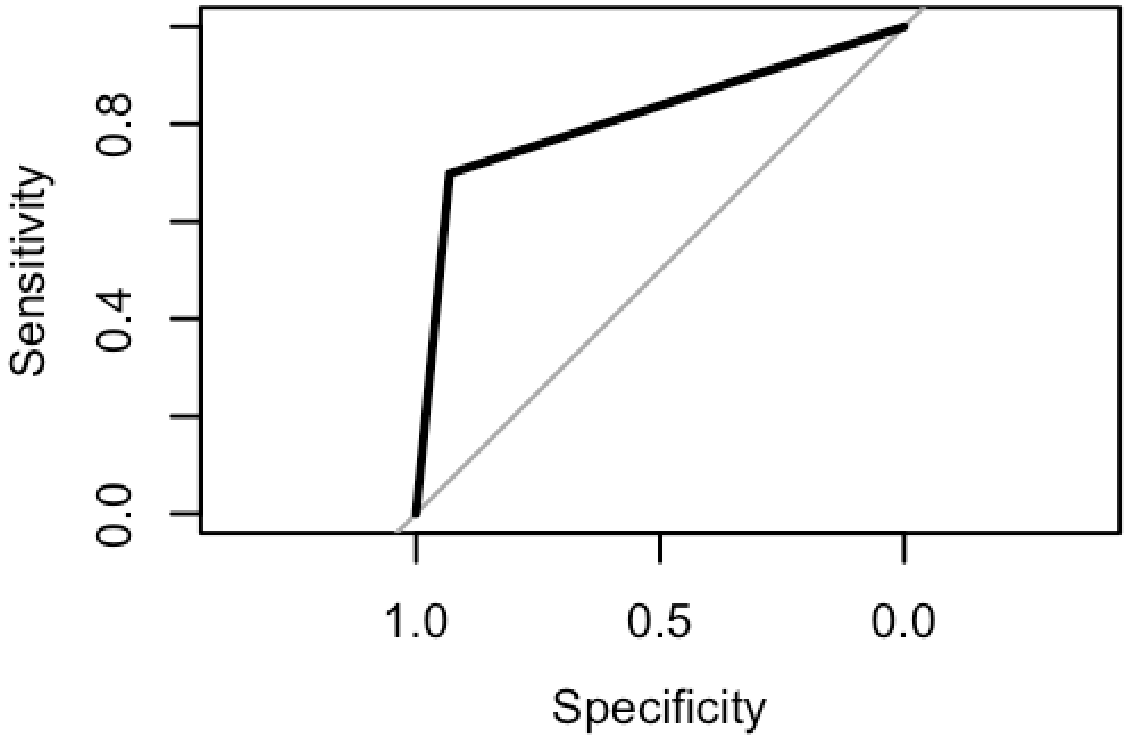


```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  3396  201
     TRUE    251  466

               Accuracy : 0.8952
                 95% CI : (0.8857, 0.9042)
    No Information Rate : 0.8454
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.6111

 Mcnemar's Test P-Value : 0.02118
```



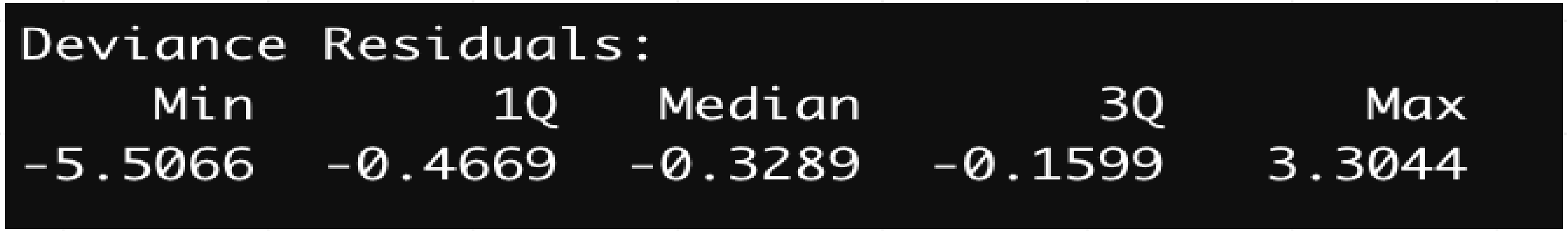
```
            Sensitivity : 0.9312
            Specificity : 0.6987
         Pos Pred Value : 0.9441
         Neg Pred Value : 0.6499
             Prevalence : 0.8454
         Detection Rate : 0.7872
   Detection Prevalence : 0.8338
      Balanced Accuracy : 0.8149

       'Positive' Class : FALSE
```

Actual and Predicted Records for Decision Tree – ROC

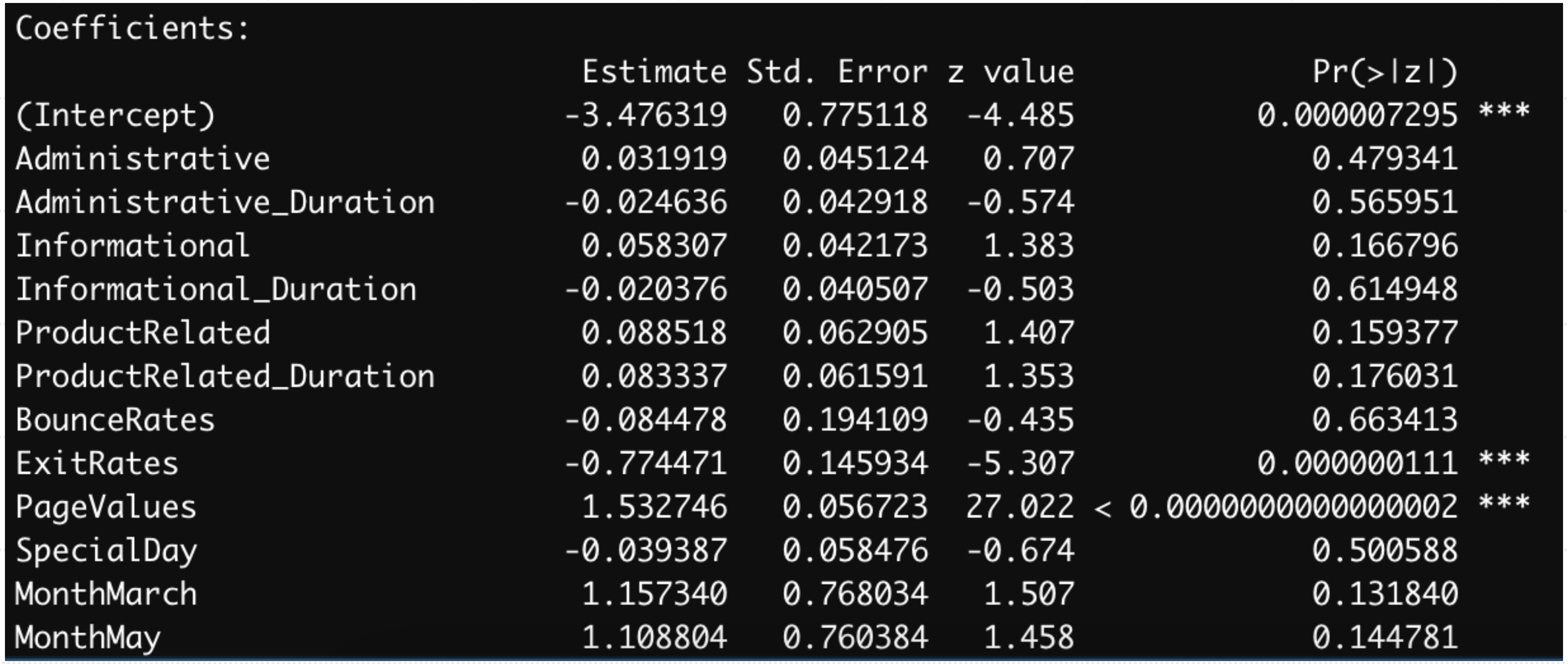


Area under the curve: 0.8149

Logistic Regression Model:
Summary for Logistic Regression Analysis

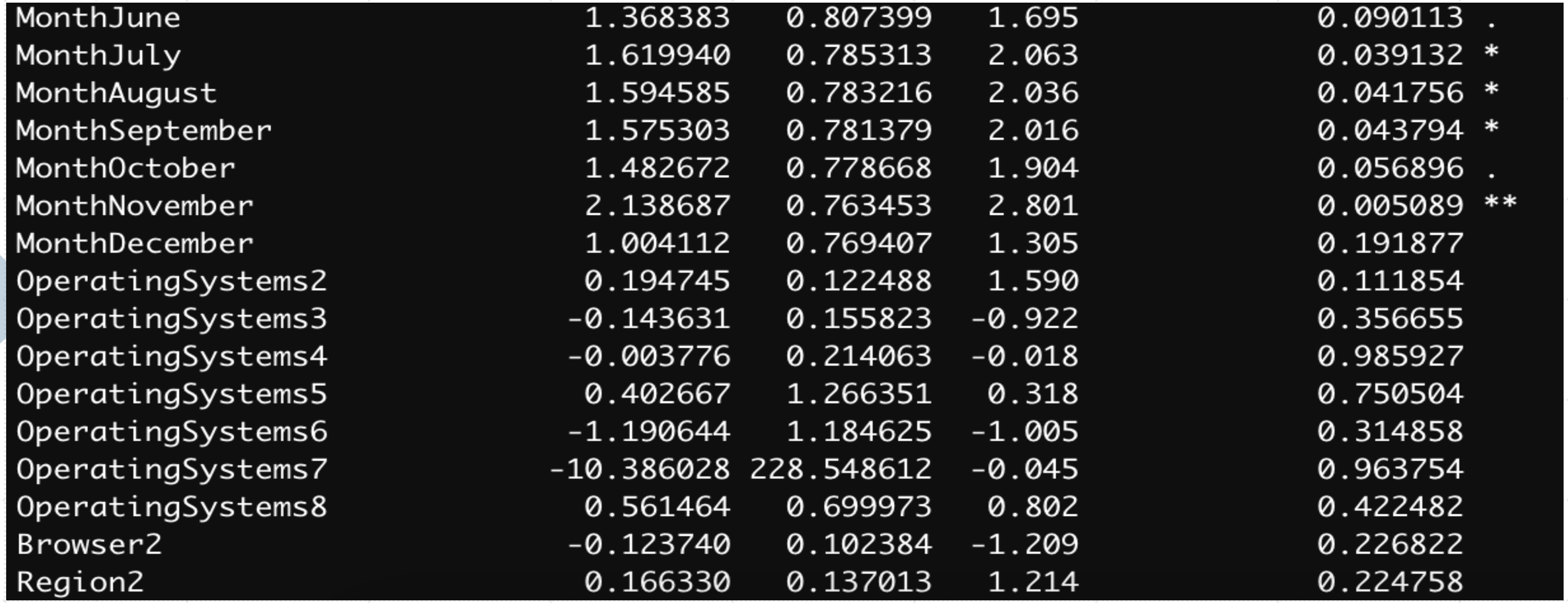


```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.5066  -0.4669  -0.3289  -0.1599   3.3044
```

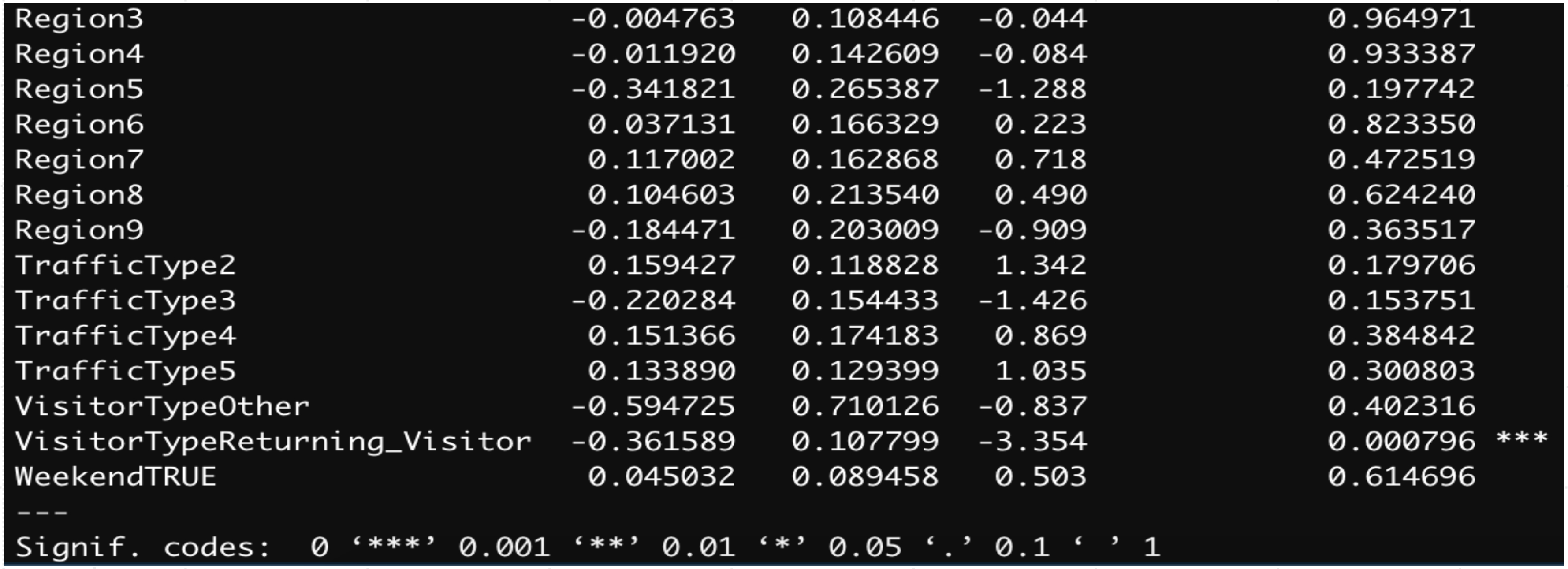


```
Coefficients:
                          Estimate Std. Error z value             Pr(>|z|)
(Intercept)              -3.476319   0.775118  -4.485          0.000007295 ***
Administrative            0.031919   0.045124   0.707             0.479341
Administrative_Duration  -0.024636   0.042918  -0.574             0.565951
Informational             0.058307   0.042173   1.383             0.166796
Informational_Duration   -0.020376   0.040507  -0.503             0.614948
ProductRelated            0.088518   0.062905   1.407             0.159377
ProductRelated_Duration   0.083337   0.061591   1.353             0.176031
BounceRates              -0.084478   0.194109  -0.435             0.663413
ExitRates                -0.774471   0.145934  -5.307          0.000000111 ***
PageValues                1.532746   0.056723  27.022 < 0.0000000000000002 ***
SpecialDay               -0.039387   0.058476  -0.674             0.500588
MonthMarch                1.157340   0.768034   1.507             0.131840
MonthMay                  1.108804   0.760384   1.458             0.144781
```

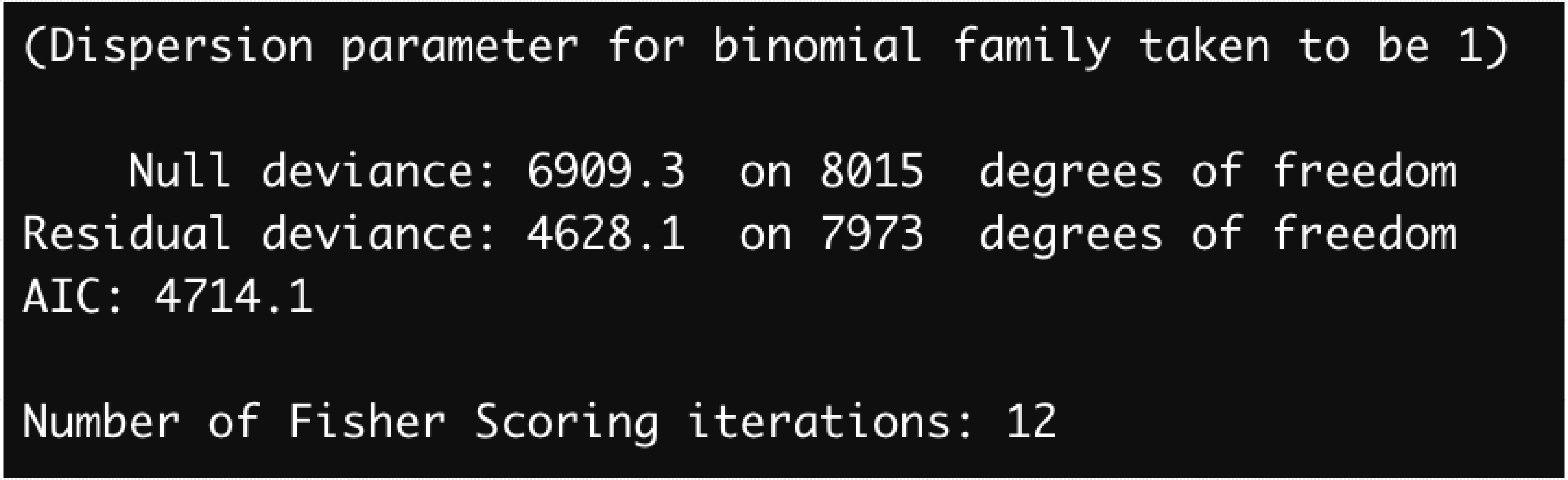
ExitRates, PageValues



```
MonthJune                      1.368383   0.807399   1.695             0.090113 .
MonthJuly                      1.619940   0.785313   2.063             0.039132 *
MonthAugust                    1.594585   0.783216   2.036             0.041756 *
MonthSeptember                 1.575303   0.781379   2.016             0.043794 *
MonthOctober                   1.482672   0.778668   1.904             0.056896 .
MonthNovember                  2.138687   0.763453   2.801             0.005089 **
MonthDecember                  1.004112   0.769407   1.305             0.191877
OperatingSystems2              0.194745   0.122488   1.590             0.111854
OperatingSystems3             -0.143631   0.155823  -0.922             0.356655
OperatingSystems4             -0.003776   0.214063  -0.018             0.985927
OperatingSystems5              0.402667   1.266351   0.318             0.750504
OperatingSystems6             -1.190644   1.184625  -1.005             0.314858
OperatingSystems7            -10.386028 228.548612  -0.045             0.963754
OperatingSystems8              0.561464   0.699973   0.802             0.422482
Browser2                      -0.123740   0.102384  -1.209             0.226822
Region2                        0.166330   0.137013   1.214             0.224758
```

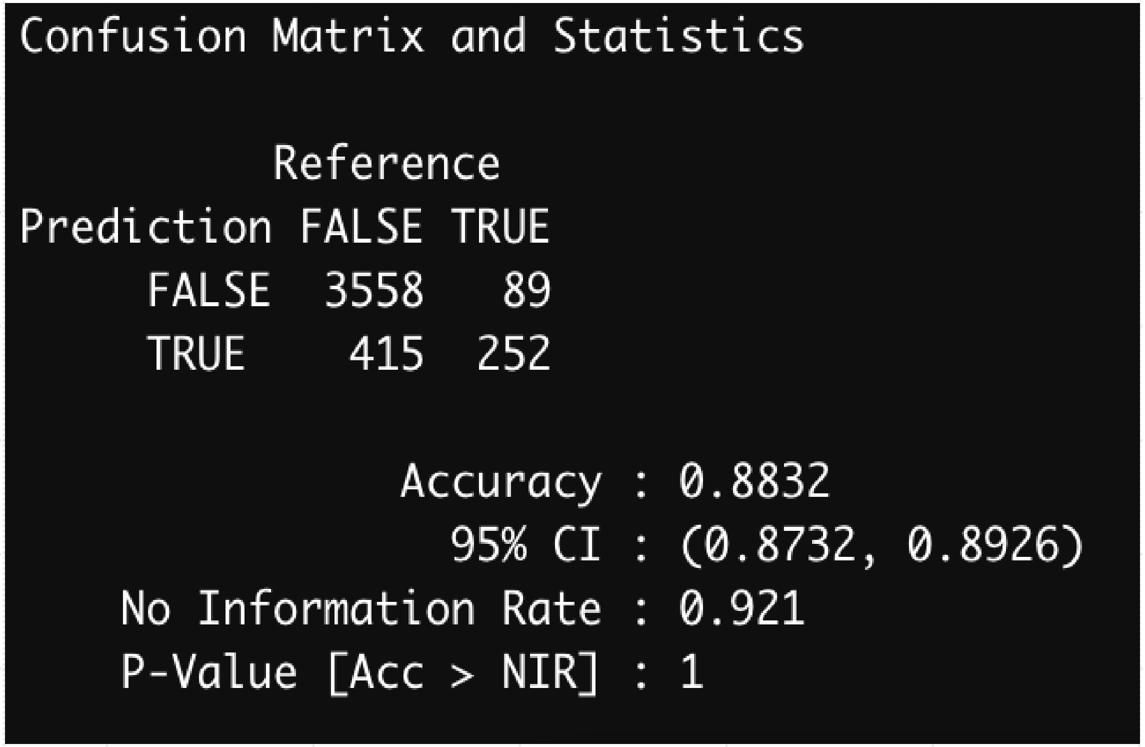


```
Region3                       -0.004763   0.108446  -0.044             0.964971
Region4                       -0.011920   0.142609  -0.084             0.933387
Region5                       -0.341821   0.265387  -1.288             0.197742
Region6                        0.037131   0.166329   0.223             0.823350
Region7                        0.117002   0.162868   0.718             0.472519
Region8                        0.104603   0.213540   0.490             0.624240
Region9                       -0.184471   0.203009  -0.909             0.363517
TrafficType2                   0.159427   0.118828   1.342             0.179706
TrafficType3                  -0.220284   0.154433  -1.426             0.153751
TrafficType4                   0.151366   0.174183   0.869             0.384842
TrafficType5                   0.133890   0.129399   1.035             0.300803
VisitorTypeOther              -0.594725   0.710126  -0.837             0.402316
VisitorTypeReturning_Visitor  -0.361589   0.107799  -3.354             0.000796 ***
WeekendTRUE                    0.045032   0.089458   0.503             0.614696
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

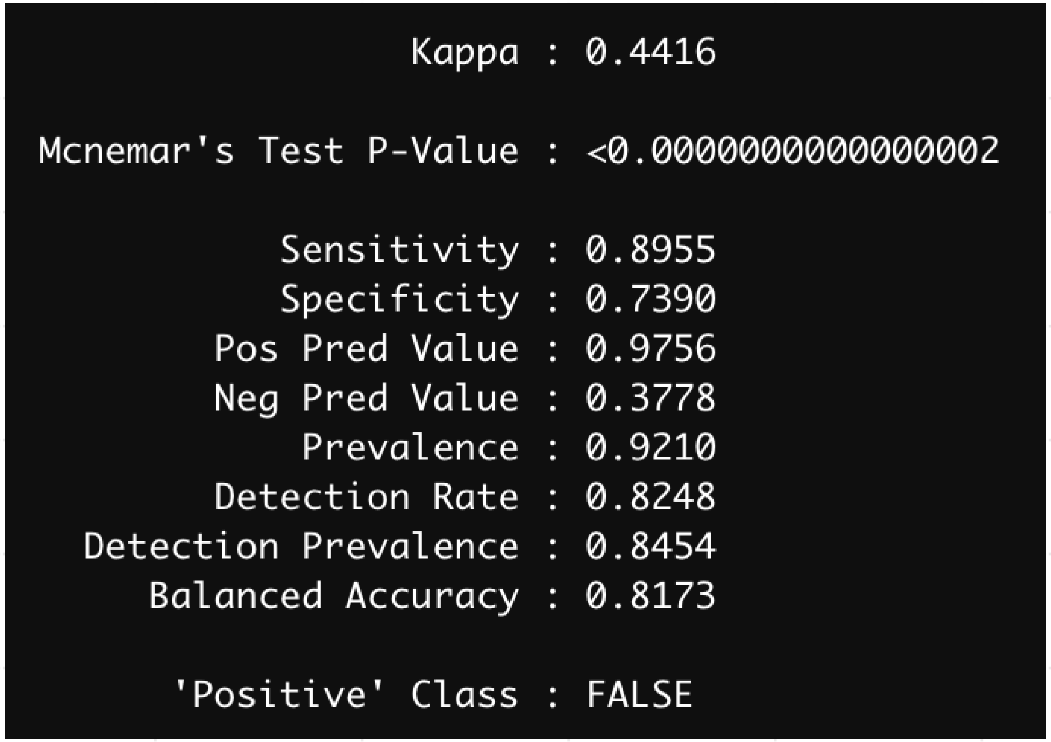


```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6909.3  on 8015  degrees of freedom
Residual deviance: 4628.1  on 7973  degrees of freedom
AIC: 4714.1

Number of Fisher Scoring iterations: 12
```

Statistics for Confusion Matrix

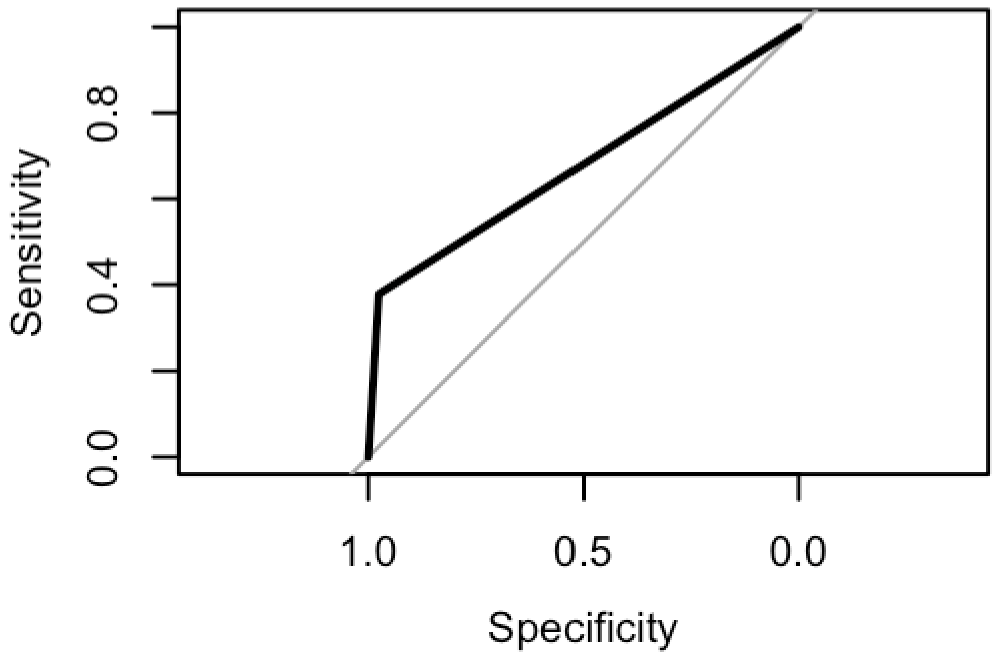


```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  3558   89
     TRUE    415  252

               Accuracy : 0.8832
                 95% CI : (0.8732, 0.8926)
    No Information Rate : 0.921
    P-Value [Acc > NIR] : 1
```



```
                 Kappa : 0.4416

Mcnemar's Test P-Value : <0.0000000000000002

           Sensitivity : 0.8955
           Specificity : 0.7390
        Pos Pred Value : 0.9756
        Neg Pred Value : 0.3778
            Prevalence : 0.9210
        Detection Rate : 0.8248
  Detection Prevalence : 0.8454
     Balanced Accuracy : 0.8173

      'Positive' Class : FALSE
```

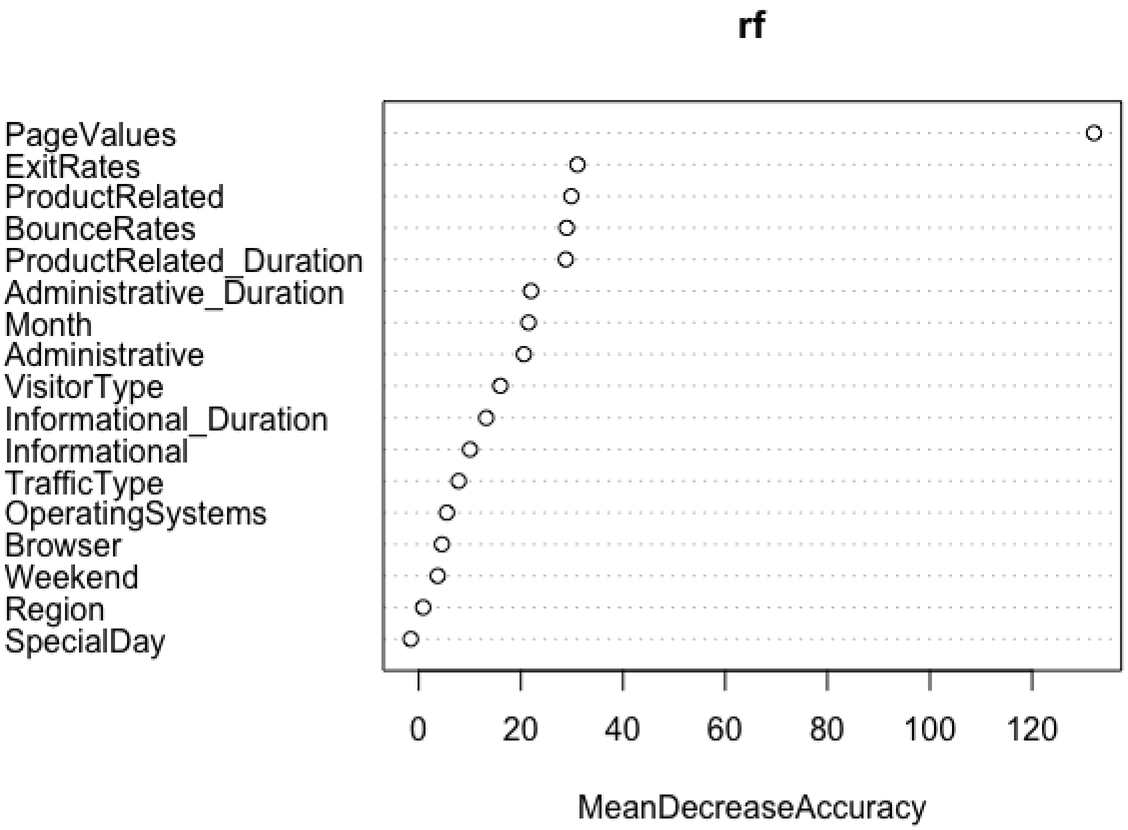
ROC Curve for Logistic Regression



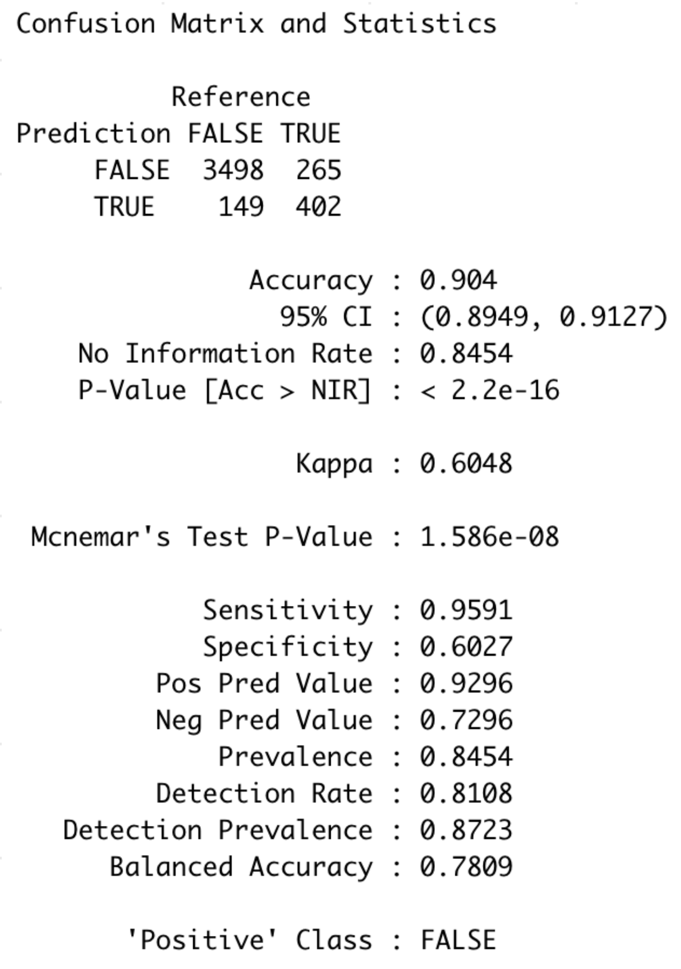
Area under the curve: 0.6767

Random Forest Model

- **Variable Importance Plot**



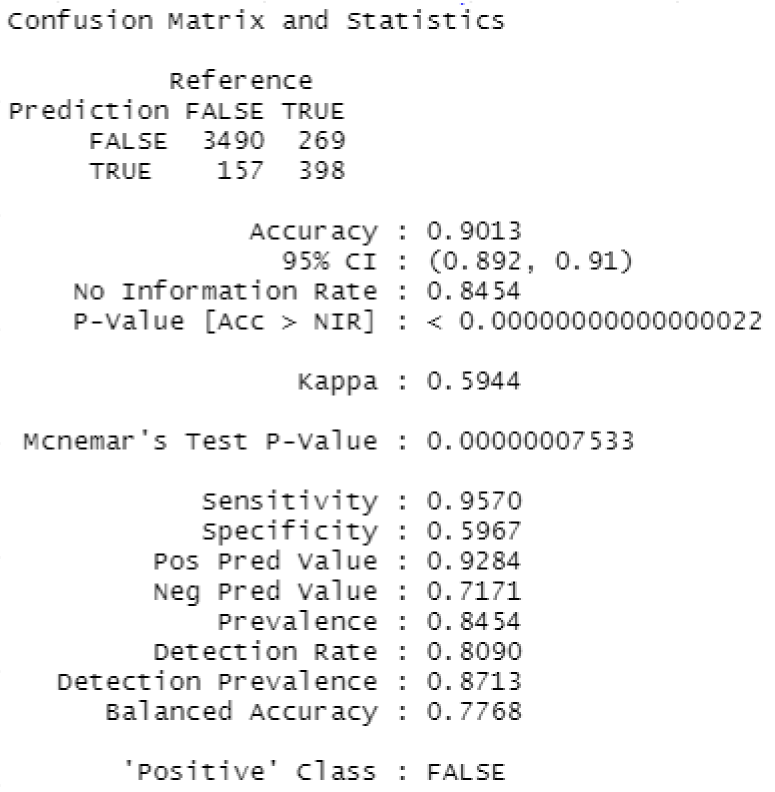
Confusion Matrix



```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  3498  265
     TRUE    149  402

               Accuracy : 0.904
                 95% CI : (0.8949, 0.9127)
    No Information Rate : 0.8454
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6048

 Mcnemar's Test P-Value : 1.586e-08

            Sensitivity : 0.9591
            Specificity : 0.6027
         Pos Pred Value : 0.9296
         Neg Pred Value : 0.7296
             Prevalence : 0.8454
         Detection Rate : 0.8108
   Detection Prevalence : 0.8723
      Balanced Accuracy : 0.7809

       'Positive' Class : FALSE
```

Model Performance Accuracy: 90.4%

Boosted Tree



```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  3490  269
     TRUE    157  398
                                               
               Accuracy : 0.9013               
                 95% CI : (0.892, 0.91)        
    No Information Rate : 0.8454               
    P-Value [Acc > NIR] : < 0.00000000000000022
                                               
                  Kappa : 0.5944               
                                               
 Mcnemar's Test P-Value : 0.00000007533        
                                               
            Sensitivity : 0.9570               
            Specificity : 0.5967               
         Pos Pred Value : 0.9284               
         Neg Pred Value : 0.7171               
             Prevalence : 0.8454               
         Detection Rate : 0.8090               
   Detection Prevalence : 0.8713               
      Balanced Accuracy : 0.7768               
                                               
       'Positive' Class : FALSE                
```

Model Performance Accuracy: 90.13%

Model Performances

- So far after evaluating decision tree, logistic regression, random forest, boosting performances, the random forest model had a better performance so far in terms of model accuracy of 90.4%, and for this reason, as we would use random forest in order to get the profitability of the online store/website. This profitability can be improved by stressing the following factors from the data by focusing on the UI of the website, exit rate, product-related information page and bounce rate compared to the other factors helping for better decision making.

Final Conclusion from The Data

Pointers To Improve Website Pages, Customer Experience:

- The significant importance of PageValue comprehends that the customers who will check out different products and their recommendations.
- Hence a good amount of improvement on recommendation engines and bundle packages would bring in more conversions for the website. This includes more products exploiting the long tail effect in e-commerce could drive more revenue.

Pointer For Better Conversion Rate:

- Minimalist and attractive UI Pages To retain more users on the website pages
- Being informative to the users about product information and their prices
- Bringing more users on the website through inorganic promotions, coupons and ads
- The bounce rate of a website can be reduced by implementing faster refresh rates and creating an attractive landing page which has highly good deals on products and offers exclusively for visitors
- Also creating personalized emails for existing members and introducing customer loyalty programs would help in bringing more retention.