Predicting Online Shoppers' Purchase Intention Using Machine Learning Techniques

Akshay Kumar Valappil Thodi, [redacted]

School of Computing, Engineering & Intelligent Systems, Ulster University, Londonderry Campus

Abstract: Online Soppers' Purchase Intention dataset has been used to build classification models to predict whether a website visitor will end up making a purchase or not. Out of three classification model built, random forest performed the best with an accuracy of 86%. Variable-wise observations were made which were used to suggest relevant modifications to the website and the insights generated from the models were actionable enough to cause improvement in e-commerce business process.

Introduction

There is rarely any commodity that people can't purchase online and there is rarely any e-commerce company that can't target people with their promotional activities. If there is internet access, if you have got a history of surfing the internet, you are a potential revenue generator for these companies.

Different companies use different methods for this purpose. Big giants like Amazone gather and use customer data and behaviour to build advance recommendation systems and frequent item sets, while, small - time entities use tools like email marketing to keep reminding people about their existence. A lot of academicians also have embarked on the journey of introducing new methods and technologies in this area. R.R. Yager introduced fuzzy intelligent agents to autonomously determine what ads to be shown in a website dynamically [1] while MT Ballestar and colleagues tried to predict the e-commerce customer quality using machine learning [2].

Going on an e-commerce website and spending time researching different items doesn't necessarily mean purchase and revenue. This poses a challenge to the companies whether such segments of potential customers should be targeted by their marketing activities. This decision involves two steps. First, predicting whether they will eventually end up making a purchase or not, if yes, target them and if not, take necessary actions to convert them. The intention of this study is to build a system which would do this job of predicting the purchase intention of an e-commerce surfer.

Dataset Description

The dataset used for this study is 'Online Shoppers' purchasing intention' from UCI machine learning repository [3]. It contains data for 12330 online sessions on 18 different attributes. The target attribute here is "Revenue" which is binary with TRUE being the customer ended up making a purchase and vice versa. Out of 12330, 10422(84.5%) of response variable is 'False' which makes the data highly imbalanced.

Out of the rest of the variables, 9 are numerical and 8 are categorical. The numerical variables include information regarding different types of web pages surfed in the session, the time spent during those sessions and the exit and bounce rates of those pages. The categorical variables include information on the month, types of operating systems and browsers used by the customer, whether the day is a special day or a weekend or not and the type of visitor. Check the appendix A for the initial summary of the dataset.

Methodology

There are three major steps in the analysis of the data. First is loading and understanding the data, second is pre-processing and the third is building the model. First part is explained clearly in the previous section and Appendix A. Steps involved in pre-processing is listed and explained further in this section.

Duplicate entries

Duplicate entries were identified and removed by the function *distinct(df)*. By taking the dimensions of the data before and after this step shows that there were 125 duplicate entries in the dataset.

Outlier Detection

Box plots were used to get an idea of outlier presence in the data. From the plots it is seen that there were a lot of outliers and made it difficult to make a conclusion. So, as a further step, the function *identify_outliers()* was used to count the extreme outliers in each variable. There were still too many such values in each attribute which is inconclusive (Check Appendix B). So, as a solution, it was decided to use machine learning algorithms which is outlier agnostic to build the prediction models.

Multicollinearity

This is caused by multiple independent variables being correlated to each other. This can be easily identified by a correlation plot. From the plot below it is clear that variable pairs (product related, product related duration) and (Exit rate, bounce rate) have relatively significant correlation. This was proved after calculating the correlation and it was decided to remove on variable from each pair, namely, *product related* and *Bounce rate.*

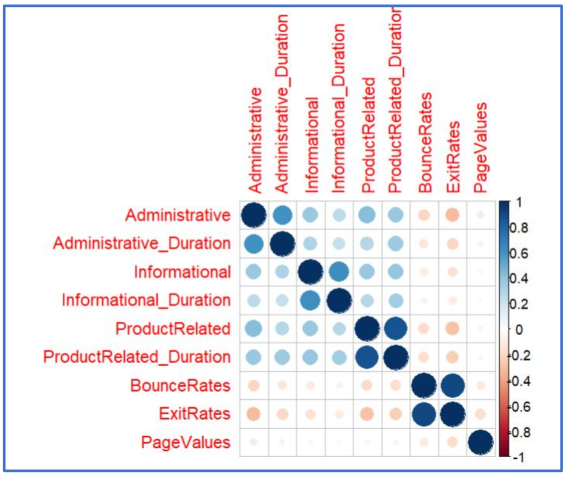


Figure 1: Correlation plot

Data Balancing

An imbalanced data gives a biased result in the model. There are multiple methods of data balancing. In this study 'SMOTE' has been used which is a data oversampling method which synthesizes dummy data in the minority class in a random manner [4]. The SMOTE-NC function that has been used here is specifically designed to SMOTE a data that has both nominal and continuous attributes in it. The before and after SMOTE distribution of the target variable 'revenue" is given below.

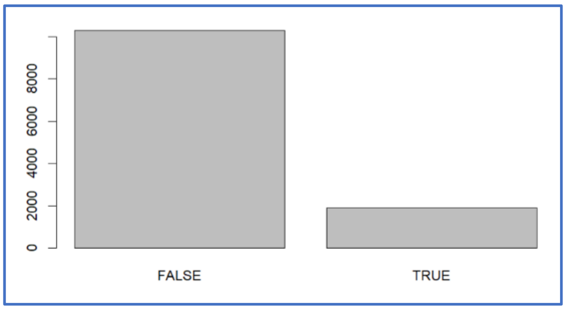


Figure 2: Class distribution before SMOTE

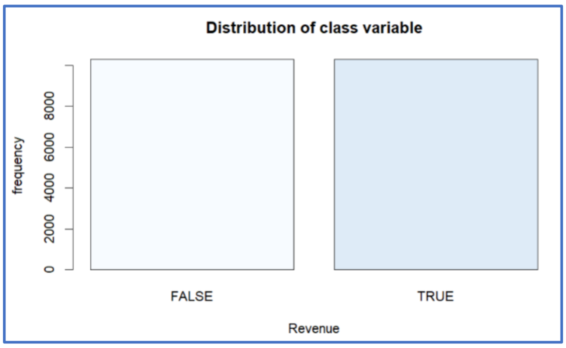


Figure 3: After SMOTE

Scaling

Different variables are in different scales which affects the calculations in the models' algorithm. To solve this, the numerical variables in the data has been scaled using the *scale(df)* function. The summary after scaling is given in Appendix C.

Classification Tree

Trees are a non-parametric classification algorithm which uses a tree like structure with root nodes, internal nodes and leaf nodes in visualizing the model [5]. On the pre-processed data, the build CT model showed an accuracy of 73%. The resulting tree is given below.

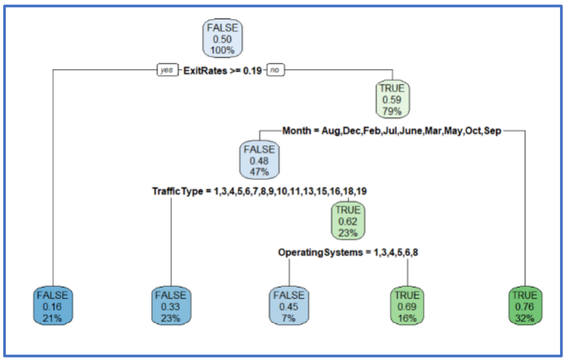


Figure 4: Classification Tree

Support Vector Machine

This is another supervised classification algorithm which is tolerant towards outliers. This model classifies data using maximum margin hyperplane [6]. The constructed model had an accuracy of 76% which is better than that of classification tree. For summary result of the model check Appendix D.

Random forest

Random forest is an ensemble classifier which consists of a collection of trees casting votes to find the most popular class for the new data point. This method is expected to show better results than CT since this involves multiple trees. As expected, the built model showed an accuracy of 86%. Find the results and confusion matrix of the mode in Appendix E.

Results

The figure in Appendix F summarises the performance of the models built. It is clear from the summary and the table below that random forest classifier outperforms other models built and it is obvious why. Random forest is an ensemble method which takes the classification decision by considering the suggestions from a number of classification trees.

Table 1: Performance summary

	CT	SVM	RF
Accuracy	73.70%	76.20%	86.10%
Precision	72.50%	77.50%	86.00%
Recall	76.10%	73.80%	86.10%
F1-Score	74.30%	75.60%	86.10%

Discussion & Conclusion

Following insights are actionable in the business point of view.

Webpages with exit rate > 0.019 should be worked on for improvement since it doesn't generate any revenue.

Spend less time and resources on the traffic types 2,7,12,17 & 20 in the month of November.

Duration of a website visit is irrelevant in revenue generation.

Confusion Matrix		**Actual**	
		+	-
Prediction	+	True Positive (TP)	False Positive (FP)
	-	False negative (FN)	True Negatives (TN)

False positives are those which has been classified as revenue generator but actually didn't. This means, they showed the same characteristics of a revenue generator, but for some reason they decided not to in that session. This should be the group for targeted marketing activities. Ince there is a higher chance of sales conversion in this group.

Recommendations

1) Get web page-wise exit and bounce rate data to improve the page quality in terms of call to action, hyperlinks to better and/or similar products, beauty features etc.
2) Gather demographic data on the customers for better modelling

3) Gather traffic type data to understand where the traffic is coming from so that marketing actions can be customized and focused depending on the types

Reference

[1] R. R.yager, "Targeted E-commerce," Institute of Electrical and Electronics Engineers, vol. 15, no. 6, pp. 42-45, 2000.

[2] P. G.-C. &. J. S. María Teresa Ballestar, "Predicting customer quality in e-commerce social networks: a machine learning approach," Review of Managerial Science, vol. 13, pp. 589-603, 2019.

[3] U. I. M. Repository, "Online Shoppers Purchasing Intention Dataset," UC Irvine ML Repository, 30 August 2018. [Online]. Available: https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset. [Accessed November 2023].

[4] N. V. C. Alberto Fern´andez, "SMOTE for Learning from Imbalanced Data: Progress andChallenges, Marking the 15-year Anniversary," Journal of Artificial Intelligence Research , vol. 61, pp. 863-905, 2018.

[5] A. A. Freitas, "Comprehensible Classification Models – a position paper," ACM SIGKDD Explorations Newsletter, vol. 15, no. 1, pp. 1-10, 2014.

[6] W. S. Noble, "What is a support vector machine?," Nature biotechnology, vol. 24, pp. 1565-1567, 2006.

Appendix

A. Summary of the dataset

```
[1] 12330    18
'data.frame':   12330 obs. of  18 variables:
 $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
 $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
 $ ProductRelated         : int  1 2 1 2 10 19 1 0 2 3 ...
 $ ProductRelated_Duration: num  0 64 0 2.67 627.5 ...
 $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
 $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
 $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
 $ Month                  : Factor w/ 10 levels "Aug","Dec","Feb",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ OperatingSystems       : int  1 2 4 3 3 2 2 1 2 2 ...
 $ Browser                : int  1 2 1 2 3 2 4 2 2 4 ...
 $ Region                 : int  1 1 9 2 1 1 3 1 2 1 ...
 $ TrafficType            : int  1 2 3 4 4 3 3 5 3 2 ...
 $ VisitorType            : Factor w/ 3 levels "New_Visitor",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Weekend                : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
 $ Revenue                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

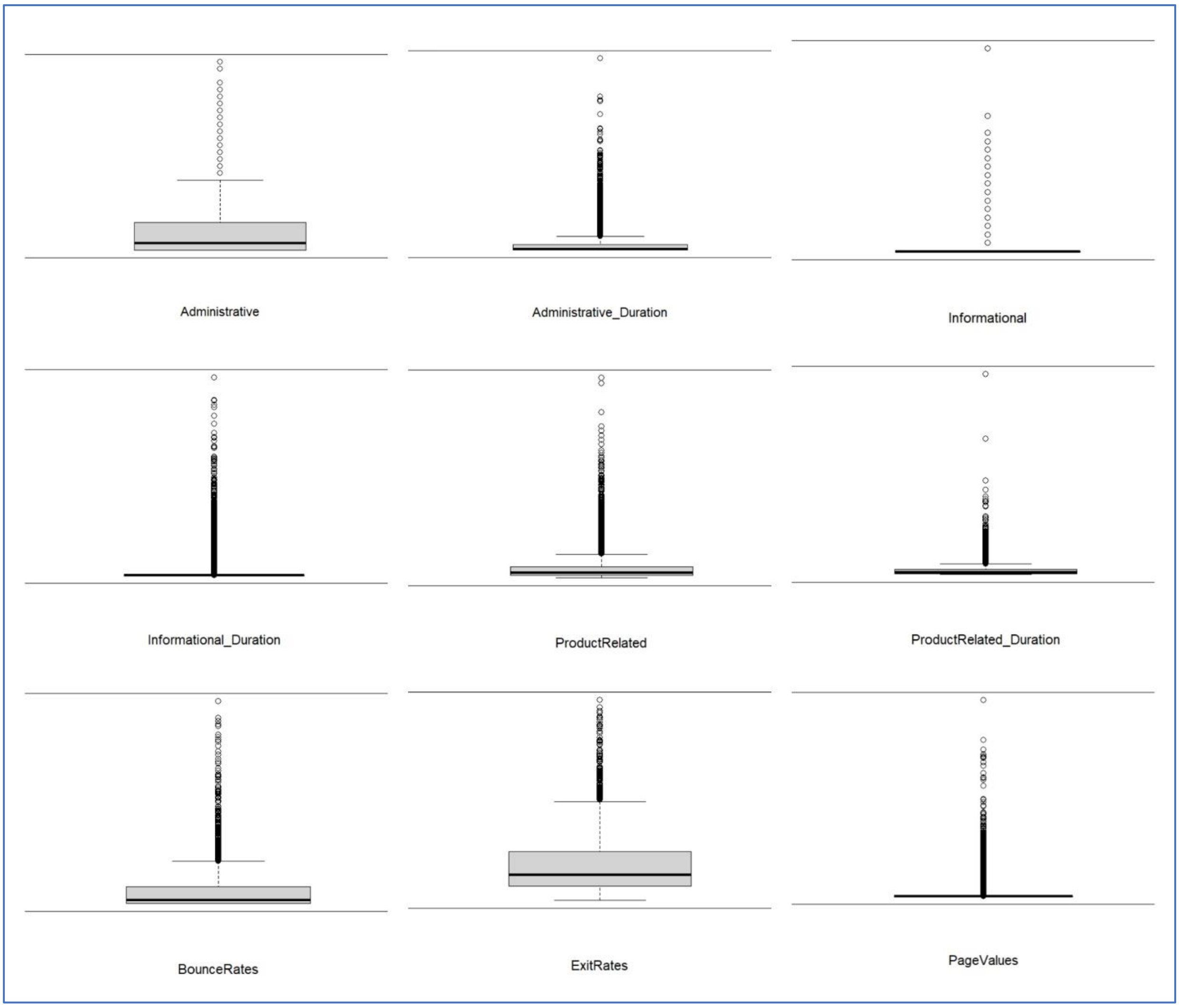
```
 Administrative   Administrative_Duration Informational      Informational_Duration
 Min.   : 0.000   Min.   :    0.00        Min.   : 0.0000    Min.   :    0.00
 1st Qu.: 0.000   1st Qu.:    0.00        1st Qu.: 0.0000    1st Qu.:    0.00
 Median : 1.000   Median :    7.50        Median : 0.0000    Median :    0.00
 Mean   : 2.315   Mean   :   80.82        Mean   : 0.5036    Mean   :   34.47
 3rd Qu.: 4.000   3rd Qu.:   93.26        3rd Qu.: 0.0000    3rd Qu.:    0.00
 Max.   :27.000   Max.   :3398.75         Max.   :24.0000    Max.   :2549.38

 ProductRelated   ProductRelated_Duration  BounceRates          ExitRates
 Min.   :  0.00   Min.   :    0.0          Min.   :0.000000   Min.   :0.00000
 1st Qu.:  7.00   1st Qu.:  184.1          1st Qu.:0.000000   1st Qu.:0.01429
 Median : 18.00   Median :  598.9          Median :0.003112   Median :0.02516
 Mean   : 31.73   Mean   : 1194.8          Mean   :0.022191   Mean   :0.04307
 3rd Qu.: 38.00   3rd Qu.: 1464.2          3rd Qu.:0.016813   3rd Qu.:0.05000
 Max.   :705.00   Max.   :63973.5          Max.   :0.200000   Max.   :0.20000

   PageValues          SpecialDay          Month        OperatingSystems     Browser
 Min.   :  0.000   Min.   :0.00000   May      :3364   Min.   :1.000    Min.   : 1.000
 1st Qu.:  0.000   1st Qu.:0.00000   Nov      :2998   1st Qu.:2.000    1st Qu.: 2.000
 Median :  0.000   Median :0.00000   Mar      :1907   Median :2.000    Median : 2.000
 Mean   :  5.889   Mean   :0.06143   Dec      :1727   Mean   :2.124    Mean   : 2.357
 3rd Qu.:  0.000   3rd Qu.:0.00000   Oct      : 549   3rd Qu.:3.000    3rd Qu.: 2.000
 Max.   :361.764   Max.   :1.00000   Sep      : 448   Max.   :8.000    Max.   :13.000
                                     (Other):1337
     Region         TrafficType              VisitorType     Weekend          Revenue
 Min.   :1.000   Min.   : 1.00   New_Visitor      : 1694   Mode :logical   Mode :logical
 1st Qu.:1.000   1st Qu.: 2.00   Other            :   85   FALSE:9462      FALSE:10422
 Median :3.000   Median : 2.00   Returning_Visitor:10551   TRUE :2868      TRUE :1908
 Mean   :3.147   Mean   : 4.07
 3rd Qu.:4.000   3rd Qu.: 4.00
 Max.   :9.000   Max.   :20.00
```

B. Outlier results

Boxplots



List of extreme outliers in each attribute

```
"No. of outliers in"
"Administrative"
51
"No. of outliers in"
"Administrative_Duration"
540
"No. of outliers in"
"Informational"
2631
"No. of outliers in"
"Informational_Duration"
2405
"No. of outliers in"
"ProductRelated"
446
"No. of outliers in"
"ProductRelated_Duration"
398
"No. of outliers in"
"BounceRates"
924
"No. of outliers in"
"ExitRates"
666
"No. of outliers in"
"PageValues"
2730
```

C. Summary after scaling

```
 Administrative.V1  Administrative_Duration.V1  Informational.V1   Informational_Duration.V1
Min.    :-0.814608  Min.    :-0.530245          Min.    :-0.460093  Min.    :-0.283980
1st Qu.:-0.814608   1st Qu.:-0.530245           1st Qu.:-0.460093   1st Qu.:-0.283980
Median :-0.366052   Median :-0.369864           Median :-0.460093   Median :-0.283980
Mean   : 0.000000   Mean   : 0.000000           Mean   : 0.000000   Mean   : 0.000000
3rd Qu.: 0.410765   3rd Qu.: 0.110504           3rd Qu.: 0.088239   3rd Qu.:-0.257117
Max.   : 7.174024   Max.   :18.631435           Max.   :17.845091   Max.   :17.310563

ProductRelated_Duration.V1    ExitRates.V1        PageValues.V1      SpecialDay
Min.    :-0.703392          Min.    :-0.854843  Min.    :-0.543071  0  :19334
1st Qu.:-0.555823           1st Qu.:-0.527104   1st Qu.:-0.543071   0.2:  178
Median :-0.315042           Median :-0.317015   Median :-0.527962   0.4:  243
Mean   : 0.000000           Mean   : 0.000000   Mean   : 0.000000   0.6:  361
3rd Qu.: 0.154367           3rd Qu.: 0.062826   3rd Qu.: 0.187555   0.8:  324
Max.   :29.697496           Max.   : 4.388980   Max.   :13.636028   1  :  154

    Month      OperatingSystems    Browser          Region        TrafficType
Nov    :7588   2      :13262    2      :15061   1      :9903   2      :9661
May    :5028   1      : 3686    1      : 3438   3      :3923   1      :3511
Dec    :2413   3      : 3026    4      :  871   2      :1588   3      :2496
Mar    :2392   4      :  513    5      :  512   4      :1556   4      :1448
Oct    : 900   8      :   75    6      :  175   7      :1037   13     : 807
Sep    : 655   6      :   19    10     :  169   6      :1030   10     : 591
(Other):1618   (Other):   13    (Other):  368   (Other):1557   (Other):2080
          VisitorType      Weekend        Revenue
New_Visitor      : 2911   FALSE:16805   FALSE:10297
Other            :   81   TRUE : 3789   TRUE :10297
Returning_Visitor:17602
```

D. Summary of SVM Model

```
Call:
svm(formula = train$Revenue ~ ., data = train, kernel = "linear", cost = 0.1,
    scale = FALSE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  0.1

Number of Support Vectors:  9321

 ( 4664 4657 )


Number of Classes:  2

Levels:
 FALSE TRUE



Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  1520  440
     TRUE    539 1619
```

E. Summary of RF Model

```
                Length Class  Mode
call                4  -none- call
type                1  -none- character
predicted       16476  factor numeric
err.rate         1500  -none- numeric
confusion           6  -none- numeric
votes           32952  matrix numeric
oob.times       16476  -none- numeric
classes             2  -none- character
importance         13  -none- numeric
importanceSD        0  -none- NULL
localImportance     0  -none- NULL
proximity           0  -none- NULL
ntree               1  -none- numeric
mtry                1  -none- numeric
forest             14  -none- list
y               16476  factor numeric
test                0  -none- NULL
inbag               0  -none- NULL
terms               3  terms  call
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  1774  287
     TRUE    285 1772
```

F. Summary of Model Performance