

# Data Mining in Online Shoppers Behaviour

## Table of Content

|                                                                                                                                                                                              |    |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Table of Content .....                                                                                                                                                                       | 2  |
| Table of Figures .....                                                                                                                                                                       | 3  |
| 1 Introduction .....                                                                                                                                                                         | 5  |
| 2 Data Set Description .....                                                                                                                                                                 | 5  |
| 2.1 The Origins of the Data .....                                                                                                                                                            | 5  |
| 2.2 About the Dataset .....                                                                                                                                                                  | 5  |
| 3 Data Pre-processing .....                                                                                                                                                                  | 10 |
| 4 Data Visualisation .....                                                                                                                                                                   | 11 |
| 5 Data Analysis .....                                                                                                                                                                        | 21 |
| 5.1 Dataset Split .....                                                                                                                                                                      | 21 |
| 5.2 Question 1: Can data mining techniques be used to automatically predict online shoppers' intention? .....                                                                                | 21 |
| 5.2.1 J4.8 Classifier .....                                                                                                                                                                  | 21 |
| 5.2.2 Other Classifiers .....                                                                                                                                                                | 26 |
| 5.3 Question 2: Is there any co-occurrence between browsing behaviour (pattern) and the transaction ending in a purchase? .....                                                              | 27 |
| 5.4 Question 3: Is it possible to identify the pages that are winners (i.e. those that contribute to purchases) and pages that are losers (i.e. those that do not encourage purchase)? ..... | 29 |
| 6 Reflection .....                                                                                                                                                                           | 30 |
| 7 Conclusion .....                                                                                                                                                                           | 30 |
| References .....                                                                                                                                                                             | 30 |

## Table of Figures

|                                                                                                                                                                                                                                                                        |    |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figure 1: Showing the number of session colour-code into those that ended in a purchase and those that did not. ....                                                                                                                                                   | 11 |
| Figure 2: Showing the percentages of session colour-code into those that ended in a purchase and those that did not. ....                                                                                                                                              | 12 |
| Figure 3: Showing the average time spent on Administrative Pages. Average of Administrative_Duration for each VisitorType broken down by Revenue. Colour shows details about Revenue. The data is filtered on Administrative, values greater than or equal to 1.....   | 13 |
| Figure 4: Showing the average time spent on Informative Pages. Average of Informational_Duration for each VisitorType broken down by Revenue. Colour shows details about Revenue. The data is filtered on Informational, values greater than or equal to 1.....        | 14 |
| Figure 5: Showing the average time spent on Product-Related Pages. Average of ProductRelated_Duration for each VisitorType broken down by Revenue. Colour shows details about Revenue. The data is filtered on ProductRelated, values greater than or equal to 1. .... | 15 |
| Figure 6: Showing the Number of Transactions across the Various Special Day Values, i.e. 0, 0.2, 0.4, 0.6, 0.8 and 1. Count of Administrative for each SpecialDay broken down by Revenue. Colour shows details about Revenue. ....                                     | 16 |
| Figure 7: Showing the Number of Transactions across the Various Weekend Values. ....                                                                                                                                                                                   | 17 |
| Figure 8: Showing the Number of Transactions across Region. ....                                                                                                                                                                                                       | 17 |
| Figure 9: Showing the Number of Transactions across Region. ....                                                                                                                                                                                                       | 18 |
| Figure 10: Visualization of all the attributes, in Weka.....                                                                                                                                                                                                           | 18 |
| Figure 11: Visualization of Administrative, Administrative_Duration, ProductRelated, ProductRelated_Duration, PageValues, SpecialDay, Browser, Region, Weekend and Revenue Attributes, in Weka .....                                                                   | 19 |
| Figure 12: Visualization of Informational, Informational_Duration, BounceRates, ExitRates, Month, OperatingSystems, TrafficType, and VisitorType Attributes, in Weka .....                                                                                             | 20 |
| Figure 13: Weka J4.8 Classifier excluding the Bounce Rate, Exit Rate and Page Value Attributes.....                                                                                                                                                                    | 22 |
| Figure 14: Weka J4.8 Classifier using only Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration Attributes with the Revenue Class Label.....                                                   | 23 |
| Figure 15: Weka J4.8 Classifier using all the attributes, without excluding any.....                                                                                                                                                                                   | 23 |
| Figure 16: Test Result for model generated excluding the Bounce Rate, Exit Rate and Page Value Attributes .....                                                                                                                                                        | 24 |
| Figure 17: Test Result using the model generated using only the attributes of Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration.....                                                        | 25 |
| Figure 18: Test Result using the model generated using all the attributes .....                                                                                                                                                                                        | 25 |
| Figure 19: Weka RandomTree Classifier Results.....                                                                                                                                                                                                                     | 28 |
| Figure 20: Weka RandomTree Classifier Result using Cross-Validation 10 folds.....                                                                                                                                                                                      | 28 |
| Figure 21: Prediction Result on Test Set using the two models from RandomTree .....                                                                                                                                                                                    | 29 |
| Table 1: Showing the Numerical attributes of the dataset.....                                                                                                                                                                                                          | 6  |
| Table 2: Minimum, Maximum, Mean and Standard Deviation Value of the Numerical Attributes .....                                                                                                                                                                         | 7  |
| Table 3: Description of the SpecialDay Class Attribute.....                                                                                                                                                                                                            | 7  |
| Table 4: Description of the Month Class Attribute .....                                                                                                                                                                                                                | 7  |

|                                                                   |    |
|-------------------------------------------------------------------|----|
| Table 5: Description of the OperatingSystems Class Attribute..... | 8  |
| Table 6: Description of the Browser Class Attribute.....          | 8  |
| Table 7: Description of the Region Class Attribute .....          | 8  |
| Table 8: Description of the TrafficType Class Attribute .....     | 9  |
| Table 9: Description of the VisitorType Class Attribute .....     | 9  |
| Table 10: Description of the Weekend Class Attribute.....         | 9  |
| Table 11: Description of the Revenue Class Attribute .....        | 10 |
| Table 12: Weka JRip Classifier Results .....                      | 26 |
| Table 13: Weka Multilayer Perceptron Classifier Results .....     | 26 |
| Table 14: Weka Naive Bayes Classifier Results .....               | 26 |
| Table 15: Weka J4.8 Classifier Results .....                      | 27 |

## 1 Introduction

Online Shopping is prevalent today. I believe that more could be done in the area of analysing the browsing pattern of online shoppers. Such analysis could help in predicting potential buyers so that more enticing content could be suggested to them, just the way a sales person might do in a departmental store. Such analysis could also help us identify pages that are not too productive in conversion. Such analysis could help identify those who are simply window-shopping to prevent them using all online resources. Hence, my interest in this Dataset.

## 2 Data Set Description

### 2.1 The Origins of the Data

The dataset was obtained from the UCI Machine Learning Repository [1]. Each record of the dataset is a collection of feature vectors recorded during a browsing session of an online shopper on a website. The website identity is not known because its identity was not disclosed, presumably to maintain the confidentiality of the shoppers as well as integrity of the website. More so, it protects the business so that its competitors do not use the information here as an edge. I attempted to contact by email the person who originally uploaded the dataset. I have not got any response, to this and other questions, which I shall disclose during this report.

The dataset was uploaded to UCI by C. Okan Sakar and Yomi Kastro. According to the UCI website, C. Okan Sakar was of the Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Bahcesehir University, Besiktas, Istanbul, Turkey. Yomi Kastro was of Inveon Information Technologies Consultancy and Trade, Istanbul, Turkey [1].

Sakar and Kastro together with S. Olcay Polat and Mete Katircioglu published a paper based on their analysis of the dataset. The paper was published on the 9th of May, 2018. The paper was titled: *Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks* [2]. It was published in the *Neural Computing and Applications* Journal. Neural Computing & Applications is an international journal which publishes original research and other information in the field of practical applications of neural computing and related techniques such as genetic algorithms, fuzzy logic and neuro-fuzzy systems [3].

In their paper, they were proposing a real-time online shopper behaviour analysis system consisting of two modules which simultaneously predicts the visitor's shopping intent and Web site abandonment likelihood. They claim that their findings support the feasibility of accurate and scalable purchasing intention prediction for virtual shopping environment using clickstream and session information data [4].

### 2.2 About the Dataset

This dataset contains 12,330 sessions. 84.5% (10,422) were negative class samples that did not end with shopping, and the rest (1,908) were positive class samples ending with shopping [1].

According to the original uploader, the dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period [4].

The dataset consists of 9 numerical and 9 categorical attributes. The Numerical attributes are:

Table 1: Showing the Numerical attributes of the dataset

| Label                    | Format | Explanations                                                                           |
|--------------------------|--------|----------------------------------------------------------------------------------------|
| Administrative           | Number | The number of times a user visited an administrative page                              |
| Administrative Duration  | Float  | The number of seconds spent on the administrative page                                 |
| Informational            | Number | The number of times a user visited an information page                                 |
| Informational Duration   | Float  | The number of seconds spent on the information page                                    |
| Product Related          | Number | The number of times a user visited a product related page                              |
| Product Related Duration | Float  | The number of seconds spent on product related pages                                   |
| Bounce Rate              | Float  | Percentage of users who visited a page without visiting another page in a session [5]. |
| Exit Rate                | Float  | Percentage of users who left the site from that page [6].                              |
| Page Value               | Float  | The Value that a page has contribute to your site's revenue [7].                       |

By Administrative Page is meant Pages about account management. By Information Page is meant Pages about the Web site, communication and address information of the shopping site [1].

The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page [1].

Bounce rate is a measure of retention. The thinking being that an effective website will engage visitors deeper into the website. Encouraging visitors to continue with their visit. Bounce rate is the percentage of people who landed on a page and immediately left. Bounces are always one-page sessions. A Bounce occurs when a user opens a single page on your site and then exits without visiting another page. The Bounce rate of a page is such single-page sessions on the page divided by all sessions [5].

Exit rate is the percentage of people who left your site from that page. Exits may have viewed more than one page in a session [6]. Exit rate of some web pages is not necessarily a bad thing. For example, if the exit rate of the page informing users of their successful transaction is high, this is not necessarily a bad thing as many users are expected to exit after this page, except those who want to continue shopping.

The "Page Value" is a metric that tells you if the page has contributed to a conversion or not. It is the average value for a page that a user visited before landing on the goal page or completing an ecommerce transaction (or both). This value is intended to give you an idea of which page in your site contributed more to your site's revenue [7].

The 9 categorical attributes are: Special Day, Operating system, Browser, Region, Traffic type, Visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, Month of the year, as well as the 'Revenue' attribute indicating whether the session ended in a purchase or not.

The Special Day attribute indicates Proximity of the Session to a special day such as Christmas, Valentine's Day, etc. For example, for Valentina's day which is February 8, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8. [1]

Below is the descriptive statistics of the values of the various attributes. The first table is for the numerical attributes, while other tables are each for the various categorical attributes.

*Table 2: Minimum, Maximum, Mean and Standard Deviation Value of the Numerical Attributes*

| Attribute                | Minimum | Maximum   | Mean     | Std Dev  |
|--------------------------|---------|-----------|----------|----------|
| Administrative           | 0       | 27        | 2.315    | 3.322    |
| Administrative Duration  | 0       | 3398.75   | 80.819   | 176.779  |
| Informational            | 0       | 24        | 0.504    | 1.27     |
| Informational Duration   | 0       | 2549.375  | 34.472   | 140.749  |
| Product Related          | 0       | 705       | 31.731   | 44.476   |
| Product Related Duration | 0       | 63973.522 | 1194.746 | 1913.669 |
| Bounce Rate              | 0       | 0.2       | 0.022    | 0.048    |
| Exit Rate                | 0       | 0.2       | 0.043    | 0.049    |
| Page Value               | 0       | 361.764   | 5.889    | 18.568   |

*Table 3: Description of the SpecialDay categorical Attribute*

| Values                       | Count |
|------------------------------|-------|
| 0                            | 11079 |
| 0.2                          | 178   |
| 0.4                          | 243   |
| 0.6                          | 351   |
| 0.8                          | 325   |
| 1                            | 154   |
| Missing: 0%      Distinct: 6 |       |

*Table 4: Description of the Month categorical Attribute*

| Values                        | Count |
|-------------------------------|-------|
| Jan                           | 0     |
| Feb                           | 184   |
| Mar                           | 1907  |
| Apr                           | 0     |
| May                           | 3364  |
| June                          | 288   |
| Jul                           | 432   |
| Aug                           | 433   |
| Sep                           | 448   |
| Oct                           | 549   |
| Nov                           | 2998  |
| Dec                           | 1727  |
| Missing: 0%      Distinct: 10 |       |

Table 5: Description of the OperatingSystems categorical Attribute

| Values                       | Count |
|------------------------------|-------|
| 1                            | 2585  |
| 2                            | 6601  |
| 3                            | 2555  |
| 4                            | 478   |
| 5                            | 6     |
| 6                            | 19    |
| 7                            | 7     |
| 8                            | 79    |
| Missing: 0%      Distinct: 8 |       |

Unfortunately, I don't know what the various labels 1, 2, 3, etc stand for. Probably, 1 for Windows, 2 for Macintosh, etc. I do not know. I tried reaching the uploader of the dataset, like I said before, but to no avail. However, I don't think this would affect my analysis and findings. I would elaborate on this later.

Table 6: Description of the Browser categorical Attribute

| Values                        | Count |
|-------------------------------|-------|
| 1                             | 2462  |
| 2                             | 7961  |
| 3                             | 105   |
| 4                             | 736   |
| 5                             | 467   |
| 6                             | 174   |
| 7                             | 49    |
| 8                             | 135   |
| 9                             | 1     |
| 10                            | 163   |
| 11                            | 6     |
| 12                            | 10    |
| 13                            | 61    |
| Missing: 0%      Distinct: 13 |       |

Again, I don't know what the various labels stand for.

Table 7: Description of the Region categorical Attribute

| Values | Count |
|--------|-------|
| 1      | 4780  |
| 2      | 1136  |
| 3      | 2403  |
| 4      | 1182  |
| 5      | 318   |

|                              |     |
|------------------------------|-----|
| 6                            | 805 |
| 7                            | 761 |
| 8                            | 434 |
| 9                            | 511 |
| Missing: 0%      Distinct: 9 |     |

Again, I don't know what the various labels stand for.

*Table 8: Description of the TrafficType categorical Attribute*

| Values                        | Count |
|-------------------------------|-------|
| 1                             | 2451  |
| 2                             | 3913  |
| 3                             | 2052  |
| 4                             | 1069  |
| 5                             | 260   |
| 6                             | 444   |
| 7                             | 40    |
| 8                             | 343   |
| 9                             | 42    |
| 10                            | 450   |
| 11                            | 247   |
| 12                            | 1     |
| 13                            | 738   |
| 14                            | 13    |
| 15                            | 38    |
| 16                            | 3     |
| 17                            | 1     |
| 18                            | 10    |
| 19                            | 17    |
| 20                            | 198   |
| Missing: 0%      Distinct: 20 |       |

Again, I don't know what the various labels stand for.

*Table 9: Description of the VisitorType categorical Attribute*

| Values                       | Count |
|------------------------------|-------|
| Returning_Visitor            | 10551 |
| New_Visitor                  | 1694  |
| Other                        | 85    |
| Missing: 0%      Distinct: 3 |       |

*Table 10: Description of the Weekend categorical Attribute*

| Values                       | Count |
|------------------------------|-------|
| False                        | 9462  |
| True                         | 2868  |
| Missing: 0%      Distinct: 2 |       |

Table 11: Description of the Revenue Class Attribute

| Values                       | Count |
|------------------------------|-------|
| False                        | 10422 |
| True                         | 1908  |
| Missing: 0%      Distinct: 2 |       |

### 3 Data Pre-processing

None of the records is empty, as is obvious from the descriptive statistics above. I did not delete any attributes. But I might ignore some of them during some of the analysis, especially when I do not think such attributes could contribute to that particular analysis.

I did not add any attribute. I changed the data type of the Special Day attribute. Its datatype was numeric from the original dataset. I changed it to a categorical datatype. I did that because I believe the numeric datatype was not appropriate. It would be inappropriate to be talking of the mean value of the Special Day attribute since each of the values in the attribute is simply a label to indicate its nearness to a special day. Instead of 0.2, 0.8 etc, we could as well have used such label as Close, Closer or On the Day, etc. So, even if 0.2 and 0.4 were used, they are not floats as in numbers, but more like an indication of nearness or otherwise.

Below is the copy of my ARFF header

```
@relation online_shoppers_intention

@attribute Administrative numeric
@attribute Administrative_Duration numeric
@attribute Informational numeric
@attribute Informational_Duration numeric
@attribute ProductRelated numeric
@attribute ProductRelated_Duration numeric
@attribute BounceRates numeric
@attribute ExitRates numeric
@attribute PageValues numeric
@attribute SpecialDay {0,0.2,0.4,0.6,0.8,1}
@attribute Month {Jan,Feb,Mar,Apr,May,June,Jul,Aug,Sep,Oct,Nov,Dec}
@attribute OperatingSystems {1,2,3,4,5,6,7,8}
@attribute Browser {1,2,3,4,5,6,7,8,9,10,11,12,13}
@attribute Region {1,2,3,4,5,6,7,8,9}
@attribute TrafficType {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20}
@attribute VisitorType {Returning_Visitor,New_Visitor,Other}
@attribute Weekend {FALSE,TRUE}
@attribute Revenue {FALSE,TRUE}
```

## 4 Data Visualisation

I carried out some data visualization in Tableau to see if there are any interesting patterns. Below are some of the data visualization done in Tableau.

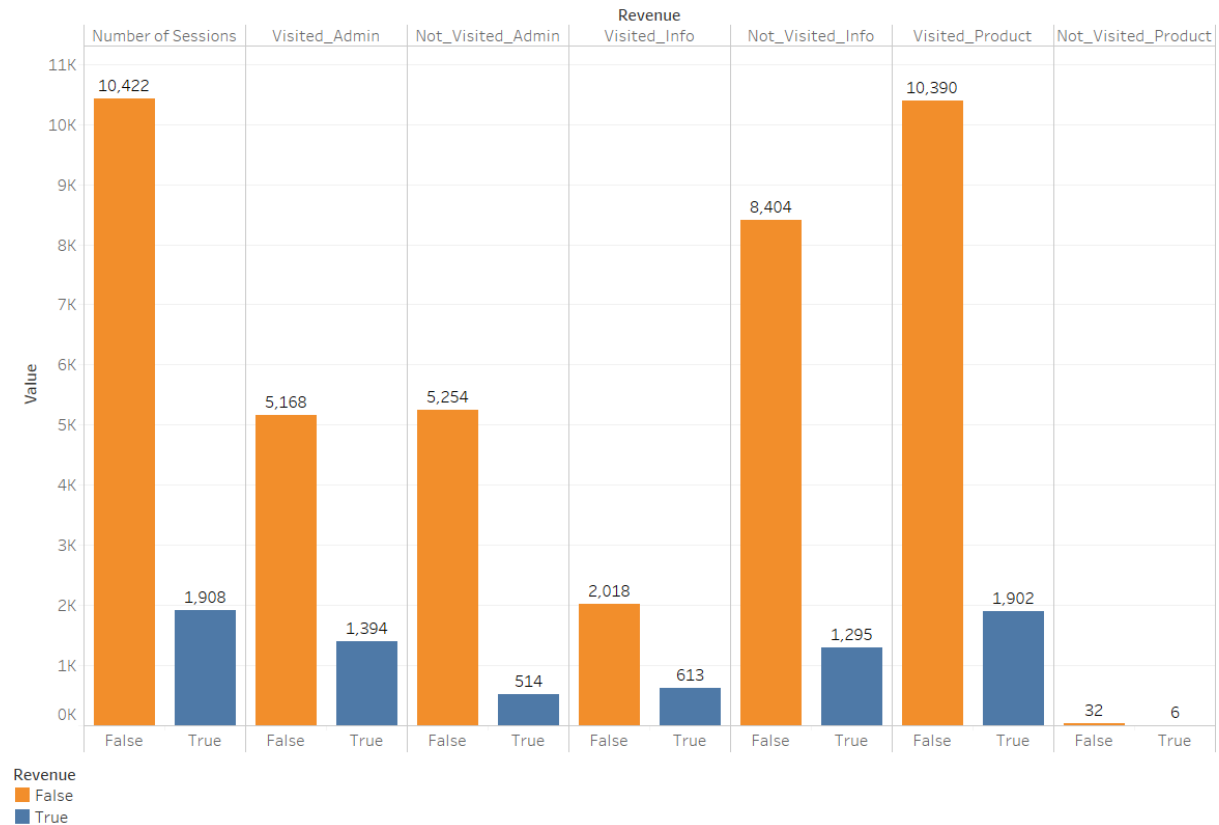


Figure 1: Showing the number of session colour-code into those that ended in a purchase and those that did not.

The above figure shows that 10,422 records did not end in a purchase, whereas 1,908 ended in a purchase. Of these 10,422, 5,168 visited administrative pages, 5,254 did not. 2,018 visited informative pages, 8,404 did not. 10,390 visited Product-Related pages, 32 did not. Of the 1,908 purchase records, 1,394 visited administrative pages, 514 did not. 613 visited informative pages, 1,295 did not. 1,902 visited Product-Related pages, 6 did not. The percentages are therefore as follows:

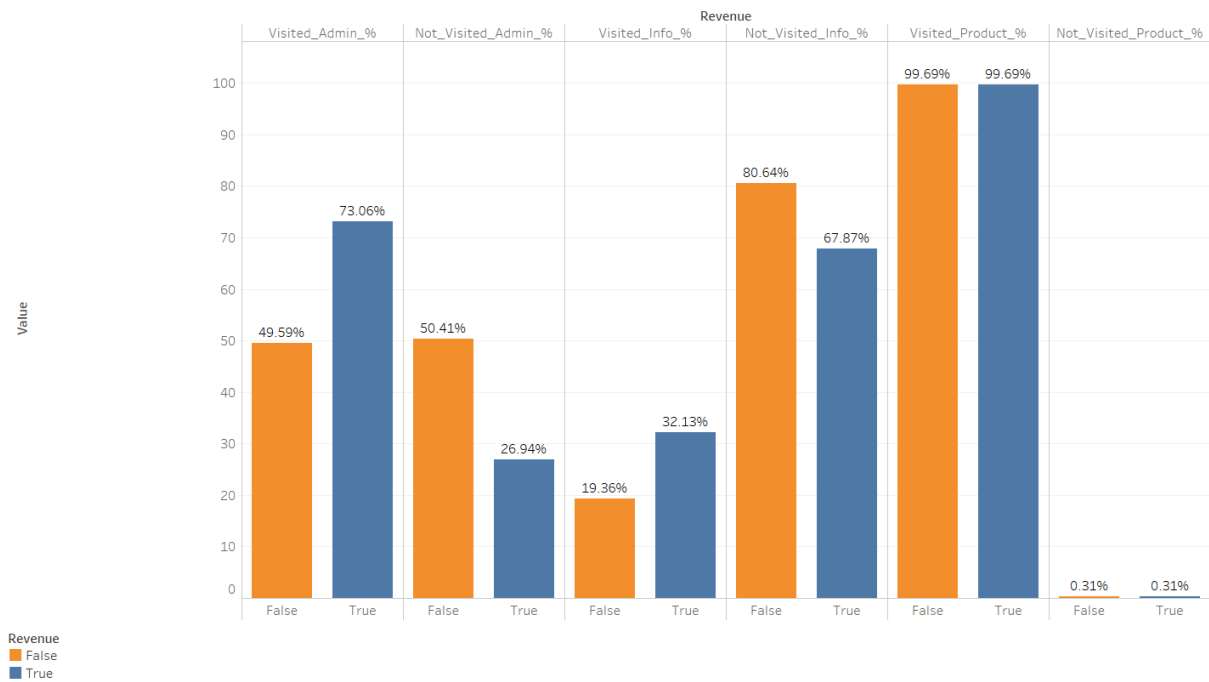


Figure 2: Showing the percentages of session colour-code into those that ended in a purchase and those that did not.

From these percentages, something interesting props up. There is a larger percentage of Successful Transaction visiting Administrative Pages (73.06%). Also, more percentage of Successful Transaction visited Informative Pages. For me, this begs the question: from browsing pattern of visitors on the website, is it possible to predict which would end in a transaction. Even if the prediction is not 100% accurate, is there any possibility? To what extent or percentage of accuracy is this prediction possible?

On this question of co-occurrence between browsing behaviour (pattern) and the transaction ending in a purchase, I carried out the following visualization in Tableau.

### Average Time spent on Administrative Page

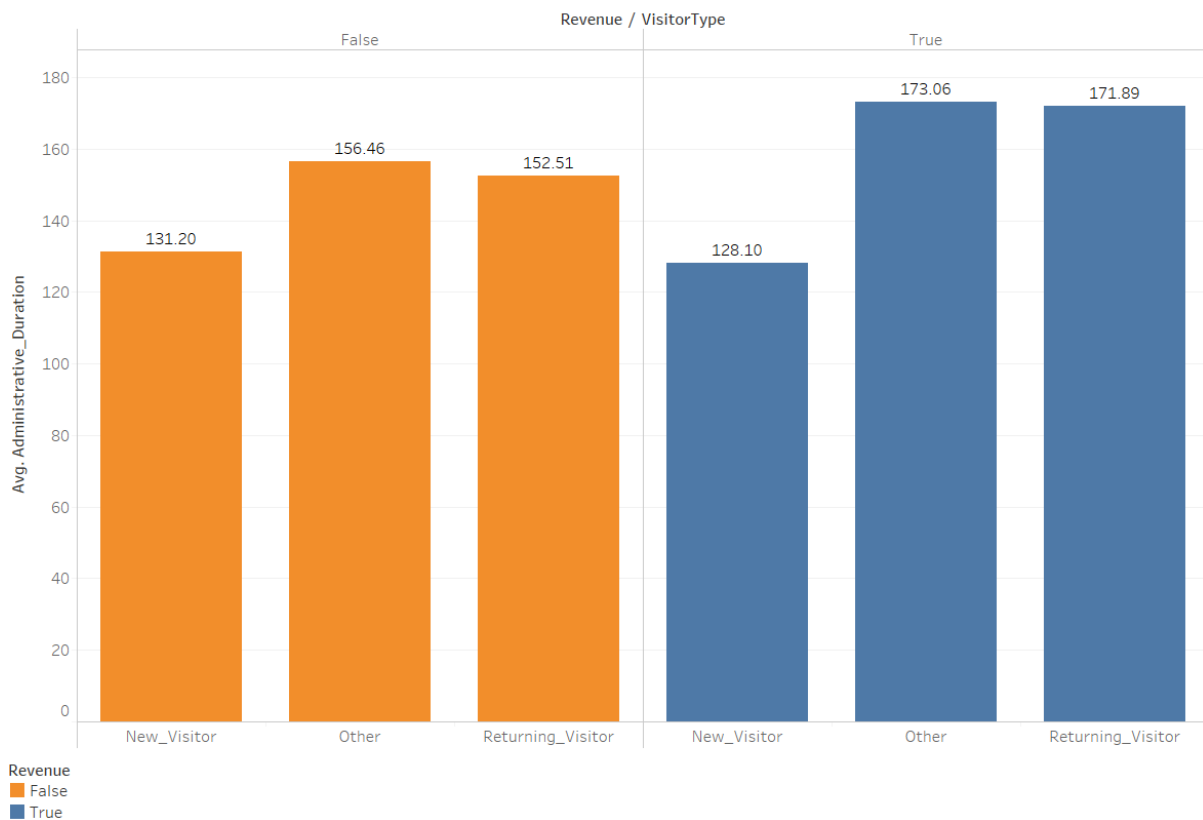


Figure 3: Showing the average time spent on Administrative Pages. Average of Administrative\_Duration for each VisitorType broken down by Revenue. Colour shows details about Revenue. The data is filtered on Administrative, values greater than or equal to 1.

The above shows the average time in seconds spent on Administrative Pages, colour-coded according to successful or unsuccessful purchase and divided into the various Visitor Types: Returning Visitor, New Visitor and Other. This shows that successful purchases tend to spend slightly more time browsing administrative pages, though the difference is very slim. To achieve a true average, I filtered out transactions that did not visit the administrative page at all.

## Average Time spent on Informational Page

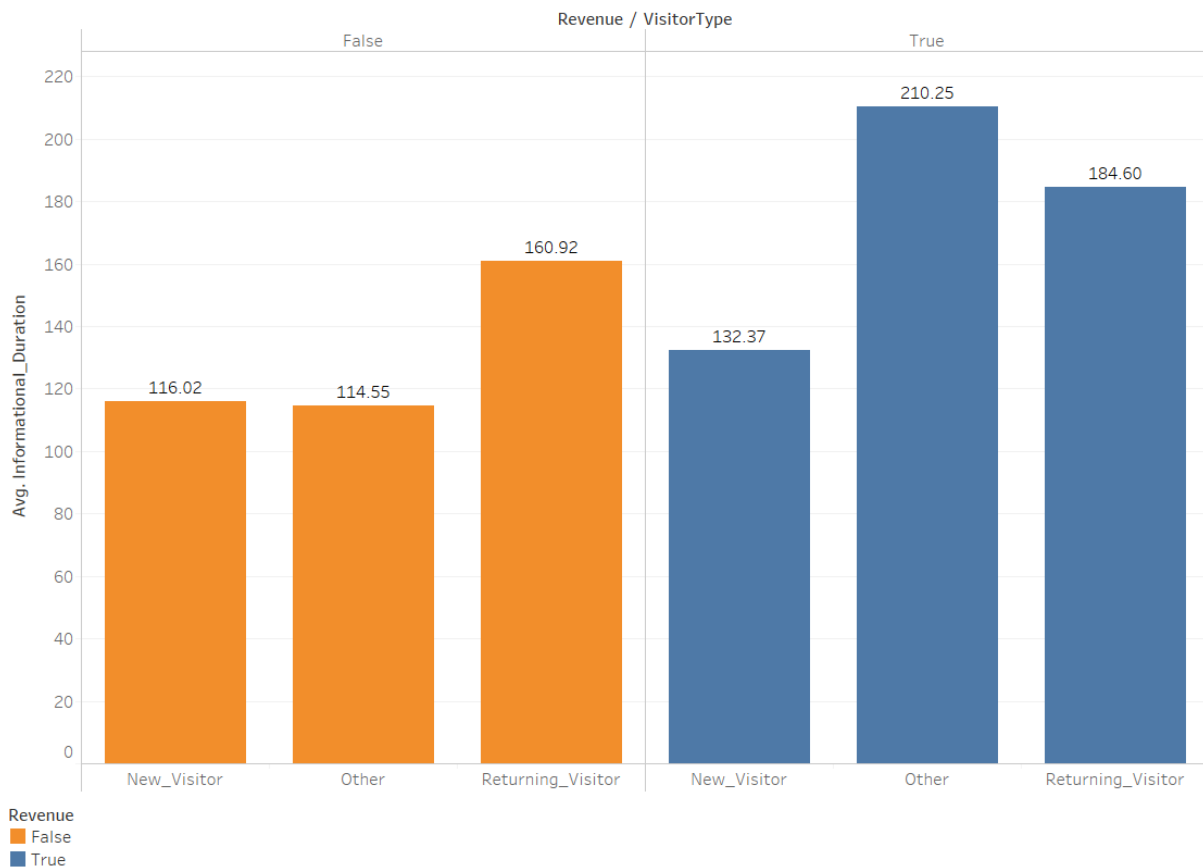


Figure 4: Showing the average time spent on Informative Pages. Average of Informational\_Duration for each VisitorType broken down by Revenue. Colour shows details about Revenue. The data is filtered on Informational, values greater than or equal to 1.

This shows that successful purchases spent more time browsing informative pages. Again, to achieve a true average, I filtered out transactions that did not visit the informative pages at all.

## Average Time spent on Product Related Page

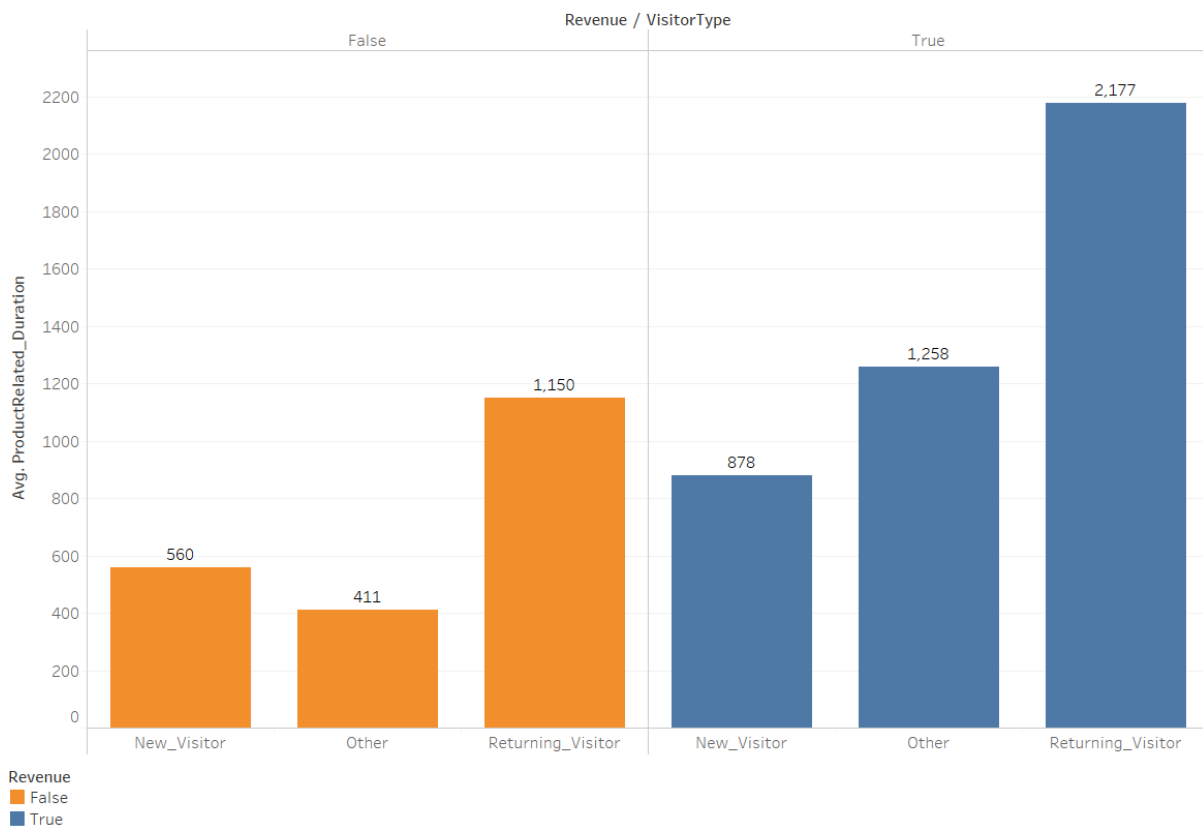


Figure 5: Showing the average time spent on Product-Related Pages. Average of ProductRelated\_Duration for each VisitorType broken down by Revenue. Colour shows details about Revenue. The data is filtered on ProductRelated, values greater than or equal to 1.

This shows that successful purchases spent significantly more time browsing Product-Related pages. This makes me more confident to explore the possibility of co-occurrence between browsing behaviour (pattern) and the transaction ending in a purchase.

There is also the problem of the meaning of the label used for these categorical attributes: operating system, browser, region, and traffic type. I attempted contacting the uploader of the dataset for the meaning of the various labels used, 1, 2, 3, etc. He has not replied. Nonetheless, I don't think these attributes would affect my analysis, because the uploader did say that the dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. [4] So, I imagine that the operating system, browser, region and traffic type are not critical determining factors. Does an operating system determine a user's browsing pattern, or does the browser determine a user's browsing pattern? This would be something interesting to investigate, but I would not do so in this project. I would simply just use them as they are in the dataset, without making an major assumptions.

Sheet 1 (6): Number of Transactions across the Various Special Day Values

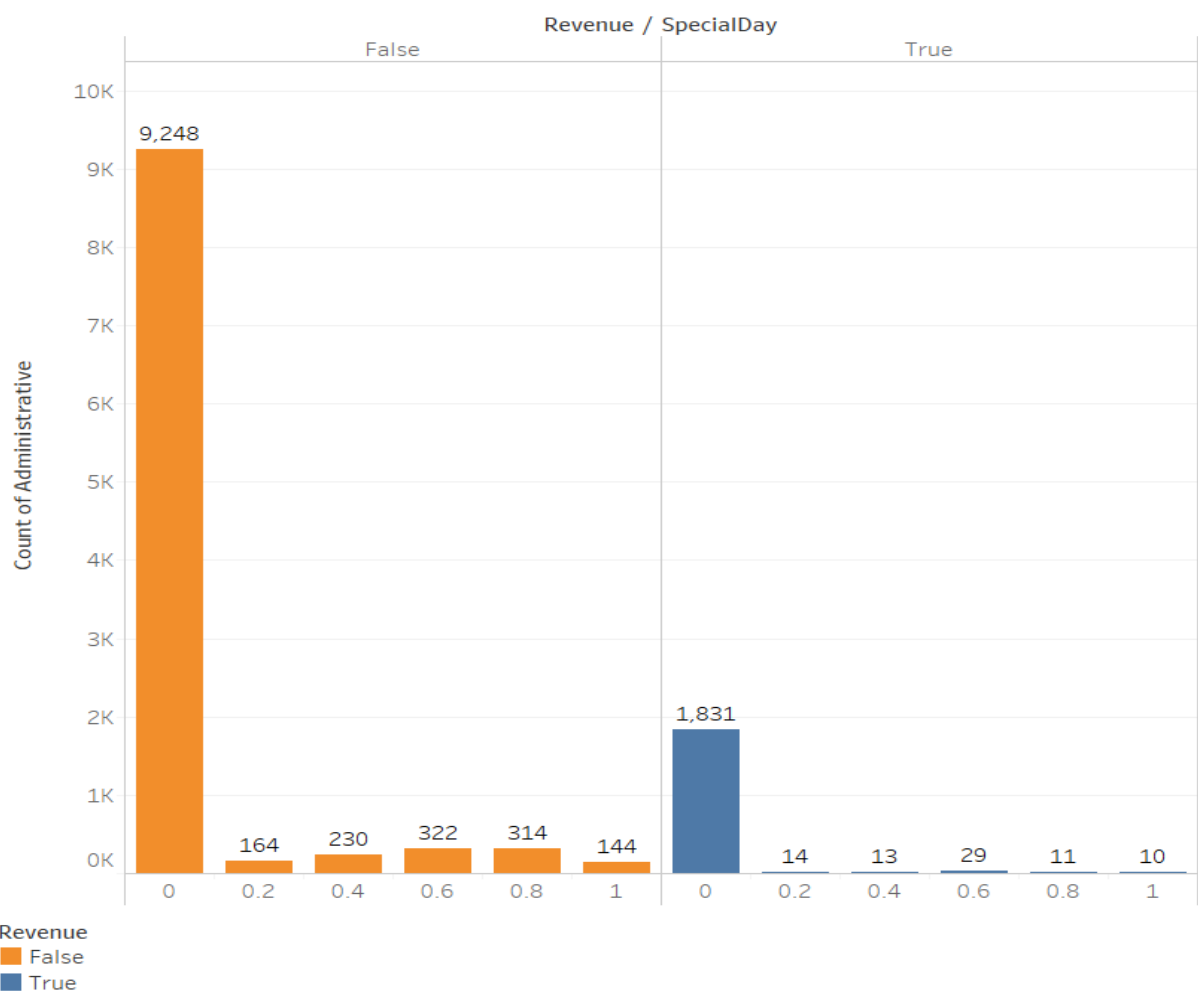


Figure 6: Showing the Number of Transactions across the Various Special Day Values, i.e. 0, 0.2, 0.4, 0.6, 0.8 and 1. Count of Administrative for each SpecialDay broken down by Revenue. Colour shows details about Revenue.

From the above figure, it is obvious that Special Day is not a determining factor. Majority of the transactions were actually outside proximity to any special day.

Sheet 1 (7): Number of Transactions across the Various Weekend Values (False/True)

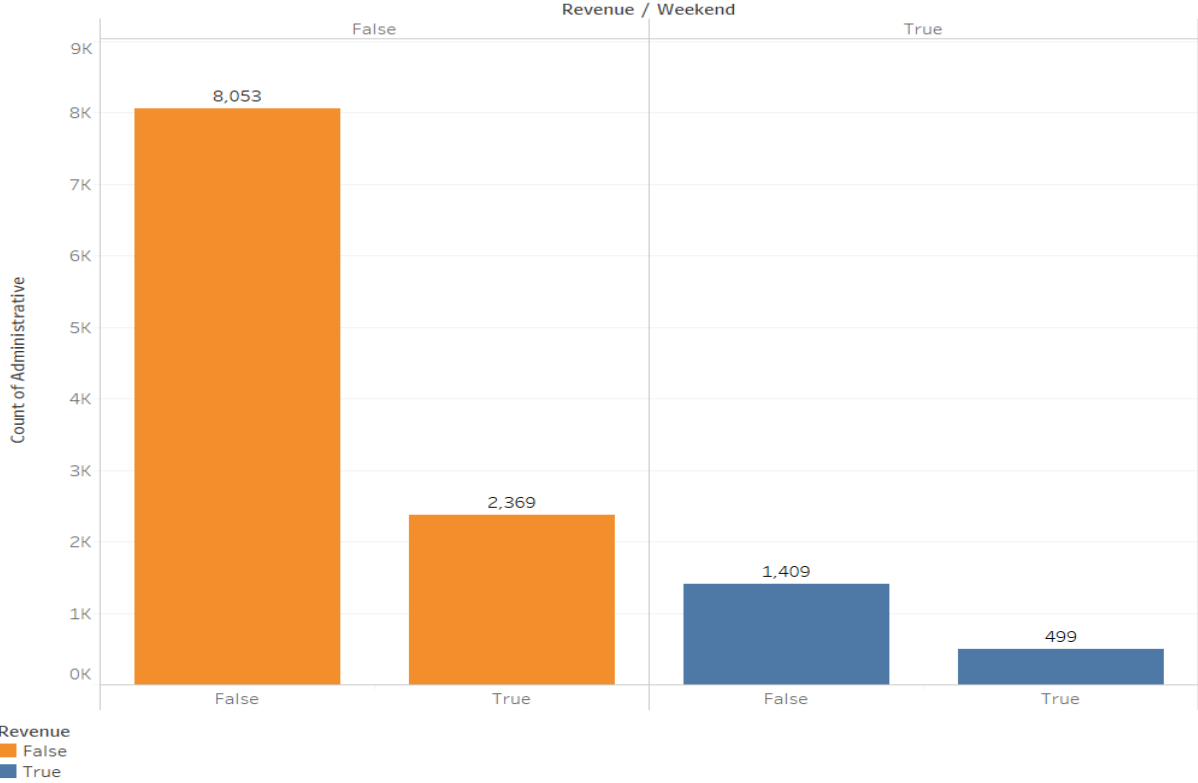


Figure 7: Showing the Number of Transactions across the Various Weekend Values.

Sheet 1 (8): Number of Transactions across Region

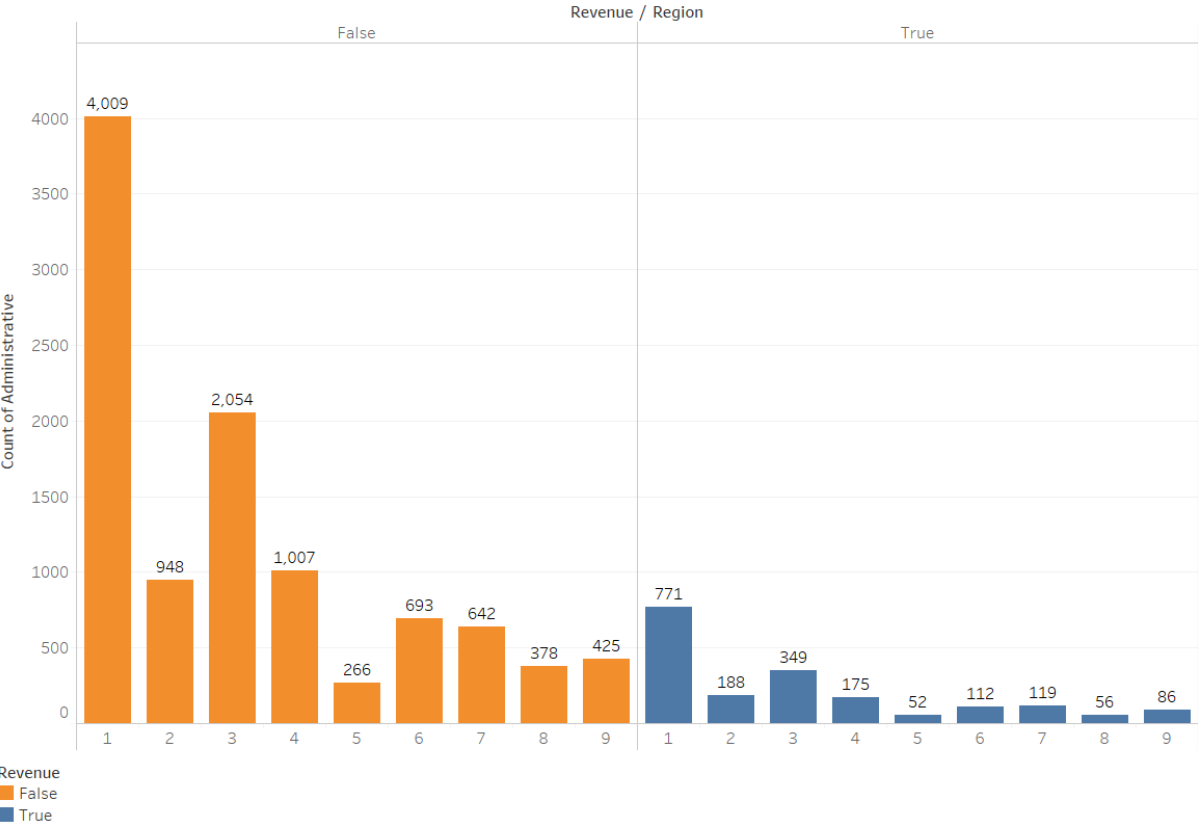


Figure 8: Showing the Number of Transactions across Region.

Sheet 1 (9): Number of Transactions across Traffic Type

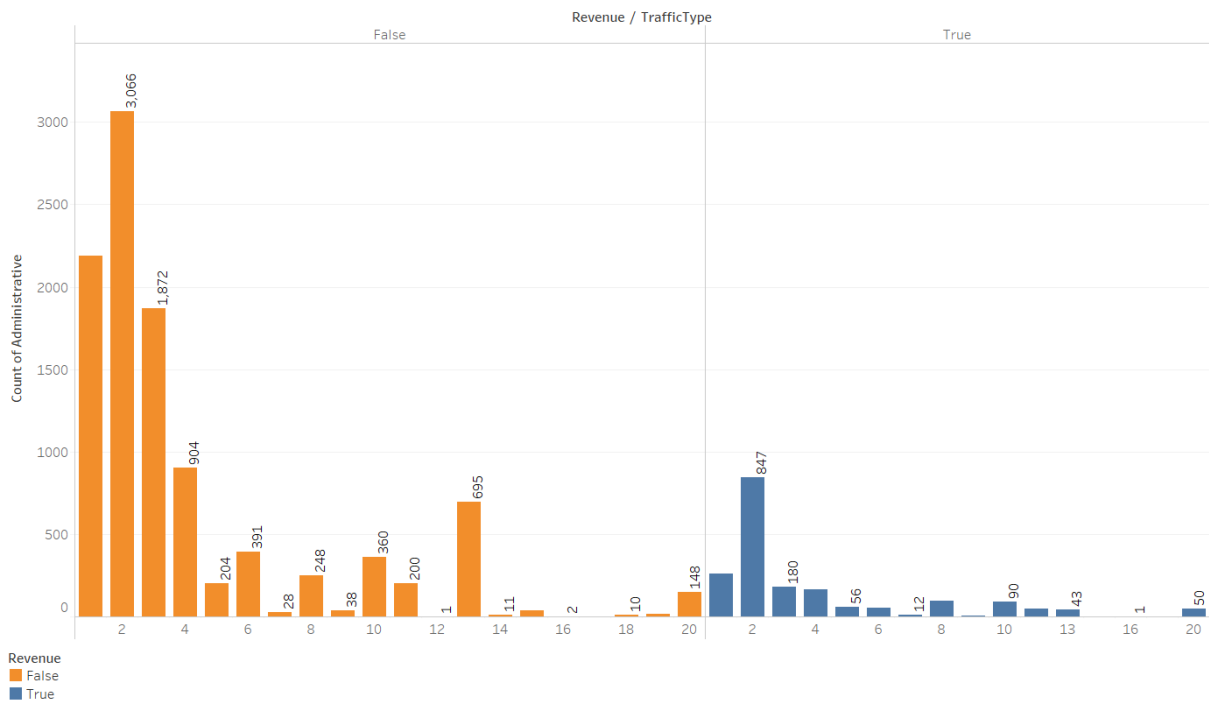


Figure 9: Showing the Number of Transactions across Region.

The following visualization is from Weka. It is a visualization of all the attributes.

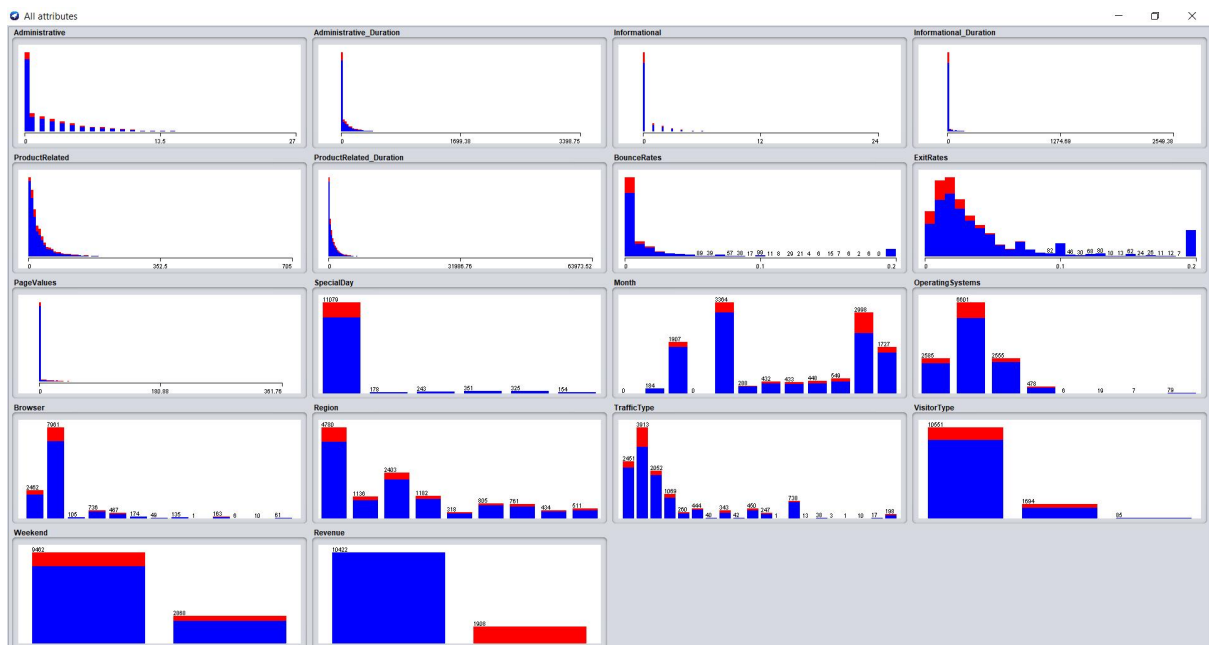


Figure 10: Visualization of all the attributes, in Weka

The above diagram is too small, so I would break it into two so they are clearer.

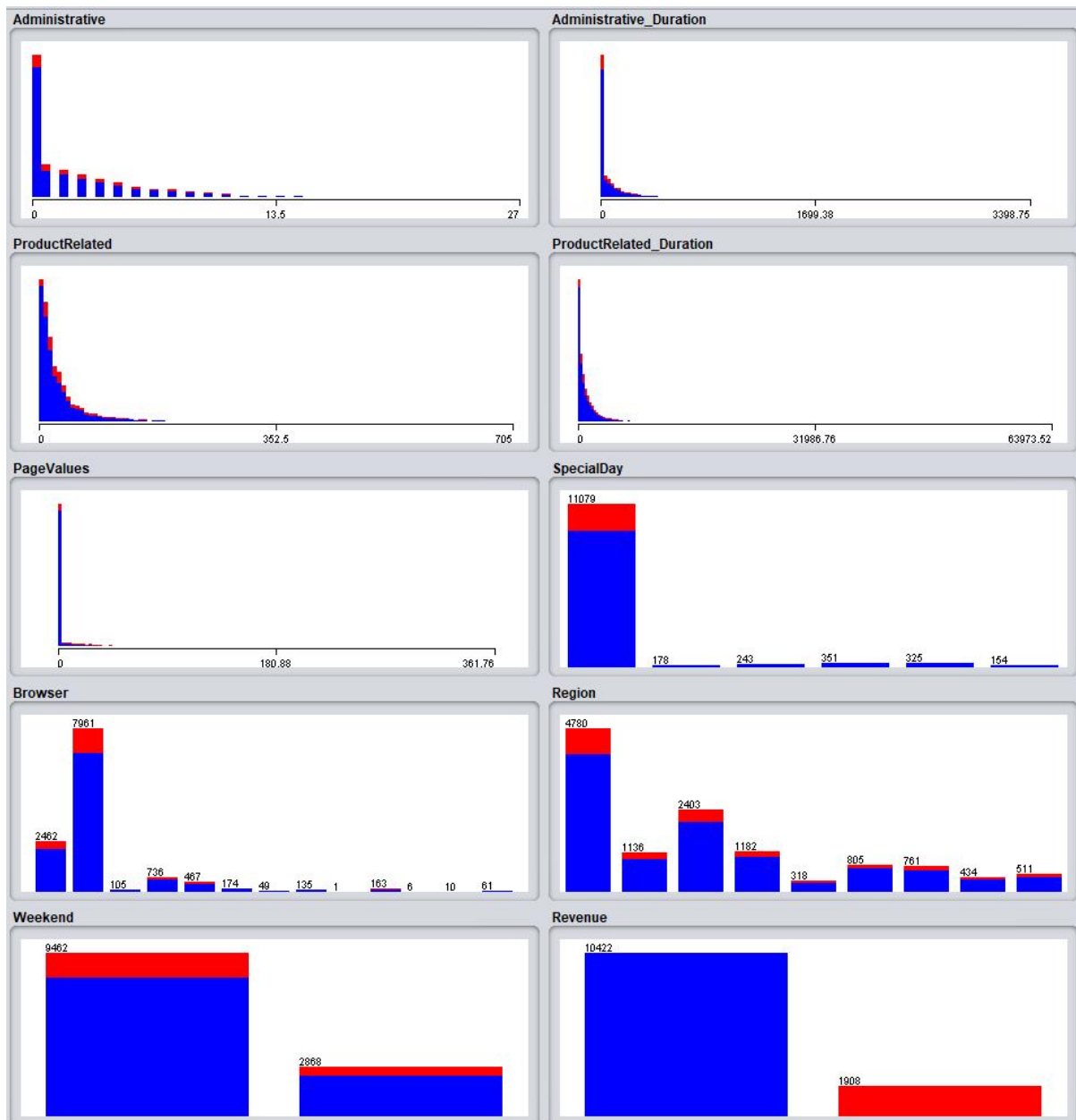


Figure 11: Visualization of Administrative, Administrative\_Duration, ProductRelated, ProductRelated\_Duration, PageValues, SpecialDay, Browser, Region, Weekend and Revenue Attributes, in Weka

The above visualization is explained below:

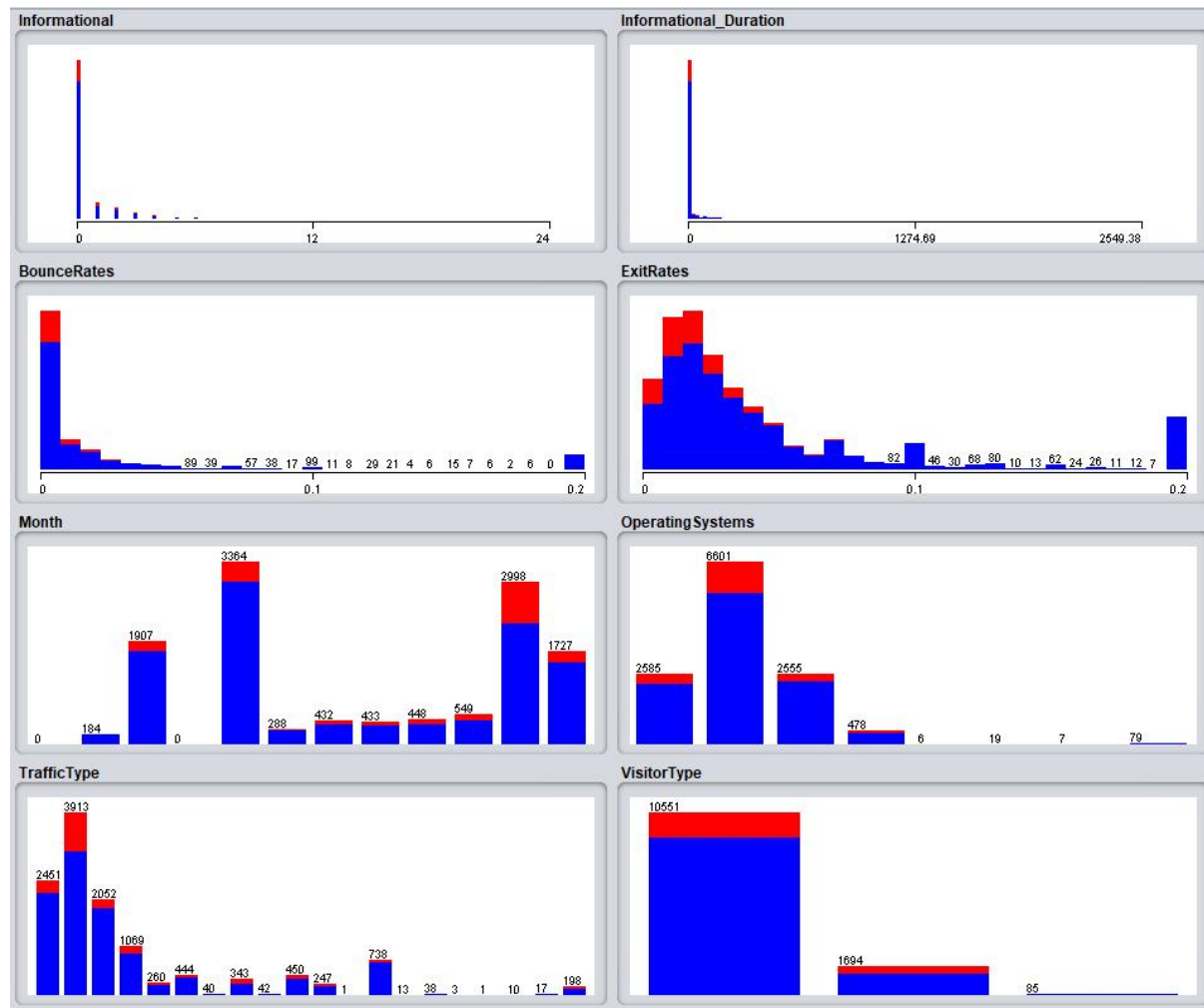
The number of times that visitors visited the Administrative Pages in a session range from 0 to 27 times. The diagram shows that many did not visit the Administrative Pages at all, as evident in the spike at 0. The number of seconds spent on the Administrative Pages in a session range from 0 to 3398.75 seconds. Certainly, most visitors did not visit the administrative pages at all as evident in the huge spike at 0.

The number of times that visitors visited Product-Related pages in a session range from 0 to 705 times. The number of seconds that visitors spent during a session on Product-Related Pages range from 0 to 63973.52 seconds.

PageValues shows the average Page Values of pages visited during a session. It ranges from 0 to 361.76. SpecialDay visualization shows that 11079 sessions occurred outside proximity to any Special

Day. The rest are explained on Table 3 above. Browser visualization shows the type of browsers used by the visitors. Browser 1 is the highest, followed by Browser 0. I do not know what the labels 0, 1, etc stand for. This attribute is not really integral for my research questions.

Region shows the various region of our visitors. Again, I do not know what the various labels 1, 2, etc stands for. Weekend shows how many visitors visited outside weekends (9462) and those that visited during the weekends (2868). Revenue shows that 10422 session did not end in purchase while 1908 ended in purchase.



TrafficType shows the various traffic type that visited the site represented as 1, 2, etc. I do not know what the labels 1, 2, etc stand for. VisitorType shows the various types of visitors that visited the site. It is broken down as Returning Visitors, New Visitors and Others. Returning Visitors is the highest.

## 5 Data Analysis

This project aims to address the following three research questions:

1. Can data mining techniques be used to automatically predict online shoppers' intention?
2. Is there any co-occurrence between browsing behaviour (pattern) and the transaction ending in a purchase?
3. Is it possible to identify the pages that are winners (i.e. those that contribute to purchases) and pages that are losers (i.e. those that do not encourage purchase)?

### 5.1 Dataset Split

Before I carried out any Data Analysis, I divided the dataset into a training set and a test set. I first randomized the dataset by using Weka *Randomize* filter. Next, I used the Weka *RemovePercentage* filter to split to 70% for the training set and used the *invertSelection* parameter to get the remaining 30% for the test set.

The training set has 8631 instances while the test set has 3699 instances. Subsequently, the training set would be used for training the model and the test set for the testing and validation.

### 5.2 Question 1: Can data mining techniques be used to automatically predict online shoppers' intention?

In this question, we want to investigate if it is possible to predict whether a session would end in a purchase or not. Since the revenue attribute is a categorical (discrete) variable, this is a classification problem [8]. Thus, we shall attempt several classification algorithms to see if any can help us predict the revenue attribute correctly [9] [10].

#### 5.2.1 J4.8 Classifier

When I ran J4.8 Classifier on Weka, excluding the Bounce Rate, Exit Rate and Page Value Attributes, I got the following result:

```

=== Summary ===

Correctly Classified Instances      7458           86.4095 %
Incorrectly Classified Instances    1173           13.5905 %
Kappa statistic                     0.2396
Mean absolute error                 0.2224
Root mean squared error             0.3334
Relative absolute error              84.9152 %
Root relative squared error         92.1592 %
Total Number of Instances          8631

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                -----  -----  -
                0.991    0.827    0.867      0.991    0.925      0.324    0.709    0.909    FALSE
                0.173    0.009    0.775      0.173    0.283      0.324    0.709    0.367    TRUE
Weighted Avg.   0.864    0.701    0.853      0.864    0.825      0.324    0.709    0.825

=== Confusion Matrix ===

  a    b  <-- classified as
7227  67 |    a = FALSE
1106 231 |    b = TRUE

```

Figure 13: Weka J4.8 Classifier excluding the Bounce Rate, Exit Rate and Page Value Attributes

From the summary in the figure 13 above, though the overall Precision is 0.85 and the overall Recall is 0.86, it is misleading. Why? Because the Recall for Sessions that end in Purchase (True Class) is only 0.17, while the Recall for Sessions that did not end in Purchase (False Class) is 0.99. This shows that this algorithm does better with predicting false class than true class. Probably, it is possible this is the case because there is more false than true classes in the dataset.

Note that recall is about how complete a result is, i.e. it is a measure of completeness [11]. The precision is, however, good for both false and true class. The precision for false case is 0.87 and the precision for true case is 0.78.

I ran the J4.8 Classifier again using only the useful attributes of Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration. The result is as follows:

```

=== Summary ===

Correctly Classified Instances      7294      84.5093 %
Incorrectly Classified Instances    1337      15.4907 %
Kappa statistic                     0
Mean absolute error                 0.2618
Root mean squared error             0.3618
Relative absolute error              99.9789 %
Root relative squared error         100 %
Total Number of Instances          8631

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                1.000    1.000    0.845     1.000    0.916     ?        0.500    0.845    FALSE
                0.000    0.000    ?         0.000    ?         ?        0.500    0.155    TRUE
Weighted Avg.   0.845    0.845    ?         0.845    ?         ?        0.500    0.738

=== Confusion Matrix ===

  a    b  <-- classified as
7294   0 |    a = FALSE
1337   0 |    b = TRUE

```

Figure 14: Weka J4.8 Classifier using only Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration Attributes with the Revenue Class Label.

As is obvious from the above figure, this time around the classifier identifies everything as false. This is disturbing given the fact that that attributes used (Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration) are the key determiners of a user's browsing behaviour.

Using all the attributes, the result was as follows:

```

=== Summary ===

Correctly Classified Instances      8001      92.7007 %
Incorrectly Classified Instances     630       7.2993 %
Kappa statistic                     0.7054
Mean absolute error                 0.1185
Root mean squared error             0.2434
Relative absolute error              45.2618 %
Root relative squared error         67.284 %
Total Number of Instances          8631

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0.969    0.301    0.946     0.969    0.957     0.708    0.889    0.963    FALSE
                0.699    0.031    0.804     0.699    0.748     0.708    0.889    0.749    TRUE
Weighted Avg.   0.927    0.260    0.924     0.927    0.925     0.708    0.889    0.930

=== Confusion Matrix ===

  a    b  <-- classified as
7067  227 |    a = FALSE
 403  934 |    b = TRUE

```

Figure 15: Weka J4.8 Classifier using all the attributes, without excluding any

This time around using all the attributes, we get the best results. The overall Precision is 0.92 and the overall Recall is 0.93. More so, the Recall for Sessions that end in Purchase (True Class) is 0.70 while

its Precision is 0.80. For Sessions that did not end in Purchase (False Class), the Recall is 0.97 and its Precision is 0.95. As good as these results are, it is still disturbing because why should such peripheral attribute such as Operating System, among others, contribute to giving a good result, but when they are taken away, the results are poorer. Could this be a case of overfitting or does this show that all the attributes are critical factors in predicting online shoppers' intentions?

I saved the models from each of the training above and ran them on the test set and got the following results:

For the first model, excluding the Bounce Rate, Exit Rate and Page Value Attributes:

```

3685  1:FALSE  1:FALSE  0.829
3686  1:FALSE  1:FALSE  0.951
3687  1:FALSE  1:FALSE  0.951
3688  2:TRUE   1:FALSE  + 0.829
3689  1:FALSE  1:FALSE  0.951
3690  1:FALSE  1:FALSE  0.829
3691  1:FALSE  1:FALSE  0.829
3692  1:FALSE  1:FALSE  0.829
3693  1:FALSE  1:FALSE  0.829
3694  1:FALSE  1:FALSE  0.829
3695  1:FALSE  1:FALSE  0.951
3696  1:FALSE  1:FALSE  0.951
3697  1:FALSE  1:FALSE  0.829
3698  1:FALSE  1:FALSE  0.951
3699  2:TRUE   1:FALSE  + 0.829

=== Summary ===

Correctly Classified Instances      3120           84.3471 %
Incorrectly Classified Instances    579           15.6529 %
Kappa statistic                    0.122
Mean absolute error                 0.2411
Root mean squared error             0.3601
Total Number of Instances          3699

```

Figure 16: Test Result for model generated excluding the Bounce Rate, Exit Rate and Page Value Attributes

I got an 84.3% accuracy.

For the second model, using only the useful attributes of Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration:

```

3685  1:FALSE  1:FALSE  0.845
3686  1:FALSE  1:FALSE  0.845
3687  1:FALSE  1:FALSE  0.845
3688  2:TRUE   1:FALSE  +  0.845
3689  1:FALSE  1:FALSE  0.845
3690  1:FALSE  1:FALSE  0.845
3691  1:FALSE  1:FALSE  0.845
3692  1:FALSE  1:FALSE  0.845
3693  1:FALSE  1:FALSE  0.845
3694  1:FALSE  1:FALSE  0.845
3695  1:FALSE  1:FALSE  0.845
3696  1:FALSE  1:FALSE  0.845
3697  1:FALSE  1:FALSE  0.845
3698  1:FALSE  1:FALSE  0.845
3699  2:TRUE   1:FALSE  +  0.845

=== Summary ===

Correctly Classified Instances      3128           84.5634 %
Incorrectly Classified Instances    571           15.4366 %
Kappa statistic                     0
Mean absolute error                 0.2614
Root mean squared error             0.3613
Total Number of Instances          3699

```

Figure 17: Test Result using the model generated using only the attributes of Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration

I got an accuracy of 84.6%, slightly higher than the previous.

And finally, for the model generated using all the attributes:

```

3688  2:TRUE   2:TRUE   0.83
3689  1:FALSE  1:FALSE  0.964
3690  1:FALSE  1:FALSE  0.964
3691  1:FALSE  1:FALSE  0.964
3692  1:FALSE  1:FALSE  0.964
3693  1:FALSE  1:FALSE  0.964
3694  1:FALSE  1:FALSE  0.964
3695  1:FALSE  1:FALSE  0.964
3696  1:FALSE  1:FALSE  0.964
3697  1:FALSE  2:TRUE   +  0.595
3698  1:FALSE  1:FALSE  0.964
3699  2:TRUE   1:FALSE  +  0.889

=== Summary ===

Correctly Classified Instances      3298           89.1592 %
Incorrectly Classified Instances    401           10.8408 %
Kappa statistic                     0.5531
Mean absolute error                 0.1452
Root mean squared error             0.2932
Total Number of Instances          3699

```

Figure 18: Test Result using the model generated using all the attributes

I got an accuracy of 89.2%, higher than the two previous ones.

From these results, we could say that data mining techniques can be used to automatically predict online shoppers' intention. More so, the prediction is higher when all the attributes including operating system, browsers, region etc are included. However, before we can say this conclusively more trials with new data are required.

### 5.2.2 Other Classifiers

I ran other classifiers and got similar results. The following tables shows the various results of other classifiers: JRIP, Multilayer Perceptron and Naïve Bayes. I included the J4.8 Classifier results for comparison purpose.

In the tables below, whenever I say Only the six attributes, I am referring to the attributes of Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration.

Table 12: Weka JRip Classifier Results

| JRIP                                            |               |           |        |                                         |
|-------------------------------------------------|---------------|-----------|--------|-----------------------------------------|
|                                                 | Class         | Precision | Recall | Summary                                 |
| Using All Attributes                            | FALSE         | 0.944     | 0.954  | 91.31% Correctly Classified Instances   |
|                                                 | TRUE          | 0.733     | 0.690  |                                         |
|                                                 | Weighted Avg. | 0.911     | 0.913  | 8.68% Incorrectly Classified Instances  |
| Excluding Bounce Rate, Exit Rate and Page Value | FALSE         | 0.861     | 0.984  | 85.26% Correctly Classified Instances   |
|                                                 | TRUE          | 0.609     | 0.136  |                                         |
|                                                 | Weighted Avg. | 0.822     | 0.853  | 14.74% Incorrectly Classified Instances |
| Using Only the six attributes                   | FALSE         | 0.848     | 0.995  | 84.58% Correctly Classified Instances   |
|                                                 | TRUE          | 0.541     | 0.030  |                                         |
|                                                 | Weighted Avg. | 0.801     | 0.846  | 15.42% Incorrectly Classified Instances |

Table 13: Weka Multilayer Perceptron Classifier Results

| MULTILAYER PERCEPTRON                           |               |           |        |                                         |
|-------------------------------------------------|---------------|-----------|--------|-----------------------------------------|
|                                                 | Class         | Precision | Recall | Summary                                 |
| Using All Attributes                            | FALSE         | 0.901     | 0.986  | 89.67% Correctly Classified Instances   |
|                                                 | TRUE          | 0.841     | 0.411  |                                         |
|                                                 | Weighted Avg. | 0.892     | 0.897  | 10.33% Incorrectly Classified Instances |
| Excluding Bounce Rate, Exit Rate and Page Value | FALSE         | 0.906     | 0.936  | 88.78% Correctly Classified Instances   |
|                                                 | TRUE          | 0.718     | 0.454  |                                         |
|                                                 | Weighted Avg. | 0.877     | 0.888  | 11.22% Incorrectly Classified Instances |
| Using Only the six attributes                   | FALSE         | 0.845     | 1.000  | 84.51% Correctly Classified Instances   |
|                                                 | TRUE          | ?         | 0.000  |                                         |
|                                                 | Weighted Avg. | ?         | 0.845  | 15.49% Incorrectly Classified Instances |

Table 14: Weka Naive Bayes Classifier Results

| NAÏVE BAYES                                     |               |           |        |                                         |
|-------------------------------------------------|---------------|-----------|--------|-----------------------------------------|
|                                                 | Class         | Precision | Recall | Summary                                 |
| Using All Attributes                            | FALSE         | 0.937     | 0.854  | 82.75% Correctly Classified Instances   |
|                                                 | TRUE          | 0.462     | 0.684  |                                         |
|                                                 | Weighted Avg. | 0.863     | 0.827  | 17.25% Incorrectly Classified Instances |
| Excluding Bounce Rate, Exit Rate and Page Value | FALSE         | 0.865     | 0.916  | 80.78% Correctly Classified Instances   |
|                                                 | TRUE          | 0.322     | 0.218  |                                         |
|                                                 | Weighted Avg. | 0.781     | 0.808  | 19.22% Incorrectly Classified Instances |
| Using Only the six attributes                   | FALSE         | 0.861     | 0.916  | 80.45% Correctly Classified Instances   |
|                                                 | TRUE          | 0.300     | 0.196  |                                         |
|                                                 | Weighted Avg. | 0.774     | 0.805  | 19.55% Incorrectly Classified Instances |

Table 15: Weka J4.8 Classifier Results

| J4.8                                            |               |           |        |                                         |
|-------------------------------------------------|---------------|-----------|--------|-----------------------------------------|
|                                                 | Class         | Precision | Recall | Summary                                 |
| Using All Attributes                            | FALSE         | 0.946     | 0.969  | 92.70% Correctly Classified Instances   |
|                                                 | TRUE          | 0.804     | 0.699  |                                         |
|                                                 | Weighted Avg. | 0.924     | 0.927  | 7.30% Incorrectly Classified Instances  |
| Excluding Bounce Rate, Exit Rate and Page Value | FALSE         | 0.867     | 0.991  | 86.41% Correctly Classified Instances   |
|                                                 | TRUE          | 0.775     | 0.173  |                                         |
|                                                 | Weighted Avg. | 0.853     | 0.864  | 13.59% Incorrectly Classified Instances |
| Using Only the six attributes                   | FALSE         | 0.845     | 1.000  | 84.51% Correctly Classified Instances   |
|                                                 | TRUE          | ?         | 0.000  |                                         |
|                                                 | Weighted Avg. | ?         | 0.845  | 15.49% Incorrectly Classified Instances |

From the above, my conclusion to Question 1 remains. Namely, we could say that data mining techniques can be used to automatically predict online shoppers' intention. The prediction is higher when all the attributes including operating system, browsers, region etc are included. However, before we can say this conclusively, more trials with new data are required.

### 5.3 Question 2: Is there any co-occurrence between browsing behaviour (pattern) and the transaction ending in a purchase?

In this second question, I want to investigate if there is any co-relation with browsing behaviour and transaction ending in a purchase. By browsing behaviour here, I am referring specifically to the combined values of Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration attributes only. In other words, whether the user visited the Administrative pages and how long he/she spent there, whether he/she visited the Informative pages and what was the duration and finally the duration on Product Related pages.

I am not interested in the other attributes of Bounce Rate, Exit Rate and Page Value as well as the attributes of Special Day, Operating system, Browser, Region, Traffic type, Visitor type, Weekend and Month.

The Classifiers we have used above do not do so well, when using only these six attributes with the Revenue Class label. However, using the RandomTree Classifier produced the following results.

```

=== Summary ===

Correctly Classified Instances      8626           99.9421 %
Incorrectly Classified Instances      5           0.0579 %
Kappa statistic                    0.9978
Mean absolute error                  0.0006
Root mean squared error              0.0178
Relative absolute error              0.2433 %
Root relative squared error          4.9334 %
Total Number of Instances           8631

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
          1.000    0.004    0.999     1.000    1.000     0.998    1.000    1.000    FALSE
          0.996    0.000    1.000     0.996    0.998     0.998    1.000    1.000    TRUE
Weighted Avg.    0.999    0.003    0.999     0.999    0.999     0.998    1.000    1.000

=== Confusion Matrix ===

  a    b  <-- classified as
7294    0 |    a = FALSE
  5 1332 |    b = TRUE

```

Figure 19: Weka RandomTree Classifier Results

The result is very good and way better than those obtained using the other classifiers of JRIP, J4.8, Multilayer Perceptron and Naïve Bayes.

I ran the RandomTree Classifier again, this time I used the Cross-Validation 10 folds and I got the following result:

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      6488           75.1709 %
Incorrectly Classified Instances      2143           24.8291 %
Kappa statistic                    0.0594
Mean absolute error                  0.2484
Root mean squared error              0.4983
Relative absolute error              94.869 %
Root relative squared error          137.7169 %
Total Number of Instances           8631

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
          0.851    0.791    0.854     0.851    0.853     0.059    0.530    0.853    FALSE
          0.209    0.149    0.205     0.209    0.207     0.059    0.530    0.165    TRUE
Weighted Avg.    0.752    0.692    0.754     0.752    0.753     0.059    0.530    0.747

=== Confusion Matrix ===

  a    b  <-- classified as
6209 1085 |    a = FALSE
1058  279 |    b = TRUE

```

Figure 20: Weka RandomTree Classifier Result using Cross-Validation 10 folds

I saved the models from the two result. The size of the tree from the models is 3735. When I tested the models using the Test set, I obtained the same result as follows:

```

3683  1:FALSE  1:FALSE  1
3684  1:FALSE  1:FALSE  1
3685  1:FALSE  2:TRUE   +  1
3686  1:FALSE  1:FALSE  1
3687  1:FALSE  1:FALSE  1
3688  2:TRUE   2:TRUE   1
3689  1:FALSE  1:FALSE  1
3690  1:FALSE  1:FALSE  1
3691  1:FALSE  1:FALSE  1
3692  1:FALSE  1:FALSE  1
3693  1:FALSE  1:FALSE  1
3694  1:FALSE  1:FALSE  1
3695  1:FALSE  1:FALSE  1
3696  1:FALSE  1:FALSE  1
3697  1:FALSE  1:FALSE  1
3698  1:FALSE  1:FALSE  1
3699  2:TRUE   1:FALSE  +  1

=== Summary ===

Correctly Classified Instances      2824           76.345 %
Incorrectly Classified Instances    875           23.655 %
Kappa statistic                     0.0747
Mean absolute error                 0.2369
Root mean squared error             0.4865
Total Number of Instances          3699

```

Figure 21: Prediction Result on Test Set using the two models from RandomTree

The two models each produced a 76.345% accurate prediction. It is interesting that the two models should produce the same result, even if their Precision and Recall during training were very different. I would definitely need new data to stretch the two models and demonstrate their validity. Until then, on the strength of the test I have carried out, I can say that the prediction of whether a session would end in purchase or not on the basis of browsing behaviour alone stands at 76.35% accuracy.

#### 5.4 Question 3: Is it possible to identify the pages that are winners (i.e. those that contribute to purchases) and pages that are losers (i.e. those that do not encourage purchase)?

This question is very critical for any serious online store. If they are able to identify pages that convert visitors to buyers, they could improve such pages and propose them to prospective buyers. Also, if they can identify pages that do not encourage visitors to be buyers, such pages could be taken down or improved upon to boost their conversion rates.

The values that can indicate this are the Google Analytics values of Bounce Rate, Exit Rate and Page Values. Unfortunately, our dataset only includes the Average Bounce Rate, Average Exit Rate and Average Page Values of the pages visited during a session. It does not have the individual Bounce Rate, Exit Rate and Page Values of each of the pages visited during each session. Thus, we would not be able to answer this question, given the limitation of our dataset. Nonetheless, this is a crucial question that online stores must constantly answer and improve their site, if they are to remain successful as a business.

## 6 Reflection

My major difficulty was not getting all the information that I needed on the dataset. Particularly, what the labels on the Operating System, Browser, Region and Traffic Type Attributes stood for. While this might not have hampered my data analysis, it certainly affected my confidence in interpreting the results. It would be wonderful to know what 1, 2, etc stood for in the Region and Traffic Type attributes.

At the time I chose the dataset, I was hoping I would write to the uploaders of the dataset and get a response to these questions. Little did I know then that I would not get any response to my email. Probably, they are not longer using the email since the email is an institutional one. In the future, I would not make such presupposition. I would try to gain all the information I need on a dataset before embarking on its project.

From this project, it is clear that online shoppers' behaviours often do not follow any clear patterns for one to clearly say this would end in purchase or not. This is why even when a model was built to predict purchase from online shoppers' behaviour above, the size of the tree from the models is 3735. This could not be easily read or calculated by a human being. Nevertheless, despite the complex online shoppers' behaviour more needs to be done in this area of interpretation because of the prevalent nature of online shopping today.

## 7 Conclusion

Online Shoppers' behaviours are complex and might not be easy to predict sessions ending in a purchase. Nonetheless, the analysis of online shoppers' behaviour should be done because of the prevalent nature of online shopping. Such analysis is possible as shown in this project and must be done, if online shoppers are to be given more user-fulfilling experience.

Apart from the shoppers' browsing behaviours, the web pages of the online shop need to be continually analysed to identify pages that are not productive in conversion. Such pages should either be taken down or improved upon to give users more fulfilling experience.

## References

- [1] The UCI Machine Learning Repository, "Online Shoppers Purchasing Intention Dataset Data Set," [Online]. Available: <http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset>. [Accessed 14 October 2019].
- [2] Springer Link, "About this article," [Online]. Available: <https://link.springer.com/article/10.1007%2Fs00521-018-3523-0#citeas>. [Accessed 31 December 2019].
- [3] Springer Link, "Neural Computing and Applications," [Online]. Available: <https://link.springer.com/journal/521>. [Accessed 31 December 2019].

- [4] C. O. Sakar, S. O. Polat, M. Katircioglu and Y. Kastro, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks," *Neural Computing and Applications*, vol. 31, no. 10, p. 6893–6908, October 2019.
- [5] Google Analytics Help, "Bounce rate," [Online]. Available: <https://support.google.com/analytics/answer/1009409?hl=en>. [Accessed 30 December 2019].
- [6] Google Analytics Help, "Exit Rate vs. Bounce Rate," [Online]. Available: <https://support.google.com/analytics/answer/2525491?hl=en>. [Accessed 30 December 2019].
- [7] Google Analytics Help, "How Page Value is calculated," [Online]. Available: <https://support.google.com/analytics/answer/2695658?hl=en>. [Accessed 30 December 2019].
- [8] Microsoft, "Data Mining Algorithms (Analysis Services - Data Mining)," 5 January 2018. [Online]. Available: <https://docs.microsoft.com/en-us/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining>. [Accessed 10 January 2020].
- [9] H. A. Abbass, R. Sarker and C. S. Newton, *Data Mining: A Heuristic Approach*, Herchey, PA, PA: Idea Group Publishing, 2001, p. 310.
- [10] T. T. Goh, "Data Mining: A Heuristic Approach," *Online Information Review*, vol. 27, no. 5, pp. 364-365, 2003.
- [11] Google: Machine Learning Crash Course, "Classification: Precision and Recall," [Online]. Available: <https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall>. [Accessed 10 January 2020].