**Online Shopping Predictability**

**Prepared by : Jinping Bai**

**Krishna Kiruba**

**Leolein Paopchi**

**Paul Flemming**

**Sam Fawzi**

**Prepared for : Hashmat Rohian**

[**CSDA1010 Basic Methods of Data Analytics - Winter 2020 Section III**](https://learn.continue.yorku.ca/course/view.php?id=3641)

Contents

[Abstract 3](#_Toc39841628)

[Introduction 4](#_Toc39841629)

[Background 5](#_Toc39841630)

[Objective 6](#_Toc39841631)

[Data Understanding 7](#_Toc39841632)

[Target Variable 8](#_Toc39841633)

[Preparation 8](#_Toc39841634)

[Importing and reviewing the dataset 8](#_Toc39841635)

[Preview of data 9](#_Toc39841636)

[Summary of attributes 9](#_Toc39841637)

[Missing Data 10](#_Toc39841638)

[Data Visualization 10](#_Toc39841639)

[Outliers 16](#_Toc39841640)

[Feature Engineering and Data Transformation 18](#_Toc39841641)

[Modelling 19](#_Toc39841642)

[Classification 19](#_Toc39841643)

[Logistic Regression 19](#_Toc39841644)

[Random Forest 24](#_Toc39841645)

[Clustering 27](#_Toc39841646)

[k-means clustering 27](#_Toc39841647)

[k-means clustering with scaled data 29](#_Toc39841648)

[Hierarchical Clustering 31](#_Toc39841649)

[Summary 32](#_Toc39841650)

[References 33](#_Toc39841651)

# Abstract

Online shopping has become an alternative revenue stream for some retailers such as Canadian Tire, Costco, and other large retailers. Online shopping is the major revenue stream for other retailers such as Amazon and Alibaba. Being able to predict online purchases and being able to profile online shoppers would be a benefit to these companies. This report addresses these issues by using historical data and testing a number of classification and clustering models to determine the best algorithm for profiling online shoppers and being able to predict a shopper will make a purchase or not. A total of three classification scenarios and four clustering scenarios were modeled and based upon the success criteria the Random Forest model achieved the required results for the classification portion of this report. The report discusses the methodology that was followed and the other models that were tested. This report includes a portion of the coding and a portion of the visualizations that were created. Attached with this report is a pdf file created with R studio which includes all the coding used for data understanding, feature engineering and modelling. This report plus the R Markdown file and the data can be found on Github at the following web address <https://github.com/Group4CSDA1010/Online-Shopping-Predictability>

# Introduction

Retail ecommerce sales, globally, have increased from 1.3 trillion US dollars in 2014 to 2.8 trillion US dollars in 2018 and are projected to be 4.9 trillion US dollars in 2021 (Orendorff, 2019). Total global retail sales during the same period were 22.5 trillion US dollars in 2014, 24 trillion US dollars in 2018 and are projected to be 27.25 trillion US dollars in 2021 (MarketingCharts, 2015) (Lipsman, 2019). In 2014, retail ecommerce sales accounted for 5.8% of total retail sales and by 2021 are projected to be 18% of total retail sales. This increase will provide online retailers, new and existing, opportunities in new and emerging markets. Existing online retailers will need to understand the buying patterns of existing customers as well as understand why customers purchase or decide not to purchase online.

The team assembled for this project has been tasked with two tasks. The first is understanding why an online shopper follows through with a purchase and the second task is to create a profile of an online shopper who makes a purchase.

The participants of the team involved in this study are Sam Fawzi, Jinping Bai, Krishna Kiruba, Leolein Paouchi, and Paul Flemming. The team brings together a wide variety of experience from the fields of business, finance, logistics, engineering and IT. This combined experience enables the team to analyze the project from different perspectives.

# Background

The dataset (Online Shoppers Purchasing Intention Dataset Data Set) that has been provided for this study are statistics describing the web pages visited previously, the length of time browsing the web pages and various other metrics measured by Google Analytics. The data consists of data collected over a one year period. Although the data provides a variety of features, there are issues with some of the features. Most issues will not affect the outcome of the modelling but a better insight into the data would be available if some of the issues are resolved.

# Objective

One of objectives of this study is to develop an algorithm for predicting the probability of an online shopper making a purchase. The second objective of the study is to create a profile of the typical online shopper who makes a purchase. This process will include looking at the variables and deciding if they are required in the modelling process. Looking for outliers in the data. Checking for missing data. Checking the independence of the variables to each other. A variety of models will initially be tested in determining the best way to develop the desired algorithm. The models for the first objective will include classification modelling and the second objective will include clustering models.

All coding for data understanding, data preparation and modelling is provided in a separate pdf file that has been generated from R Studio. A Shinyapp has also been created and can be accessed at the following web address.   <https://data4all-toronto.shinyapps.io/OnlineShopping/>  Select tables and graphs are included in this report. This report and the R markdown file are also available on github at the following address. <https://github.com/Group4CSDA1010/Online-Shopping-Predictability>

# Data Understanding

**Data Set Information:**

The dataset (Online Shoppers Purchasing Intention) consists of features belonging to 12,330 sessions. The data set was formed so that each session would belong to a different user in a one year period to avoid any tendency to a specific campaign, special day, user profile or period. Furthermore, all 18 feature variables are listed with descriptions.

**Data Understanding**

Data generated from online shoppers purchasing intentions makes it interesting to determine if it’s possible to predict buying behaviors of a person visiting a website during various times during a year. With companies like Amazon that primarily focus on E-commerce, it’s important for businesses to understand what factors will influence a site visitor from making an online purchase. Likewise, companies in marketing research can use the models developed to better target future ads that could lead to an increase in sales.

**Data Characteristics and Attribute Information:**

The dataset consists of 10 numerical and 8 categorical attributes:

Numerical Features  
  
**Administrative, Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration** - Represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another.

**Bounce Rate, Exit Rate and Page Value** - Features represent the metrics measured by "Google Analytics" for each page in the e-commerce site.

**Bounce Rate** - Feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session.

**Exit Rate** - Feature for a specific web page is calculated as for all page views to the page, the percentage that were the last in the session.

**Page Value** - Feature represents the average value for a web page that a user visited before completing an e-commerce transaction.

**Special Day -** Feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina’s day, 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.

Categorical Attributes

The dataset also includes 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, and month of the year.

*Source:* [*http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset*](http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset)

Target Variable

Has the online shopper purchased a product from the website? Binary (TRUE, FALSE)

Preparation

To analyse the data, various R libraries (i.e. dplyr, ggplot2, tidyverse, etc.) were used. Before starting the analysis, the R packages were made available by running the following code.

library(dplyr)

library(tidyverse)

…

Before preparing the data for modeling we need to better understand our data, we will start by finding the correlations between our target and other features.

Importing and reviewing the dataset

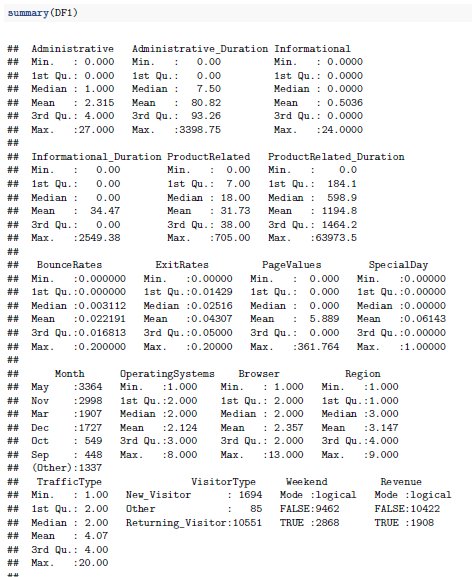
The dataset was loaded using the following code:

DF1 <- read.csv("online\_shoppers\_intention.csv")

Preview of data

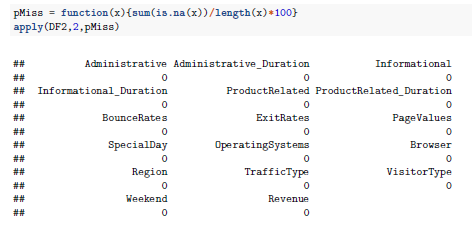
Summary of attributes

For each feature, the table below displays the Min, 1st and 3rd Quartiles, Median, Mean and Max

****

Missing Data

The R statement below was used to determine if the dataset has any missing values. There were no missing values found.



## Data Visualization

1. *Distribution of Visitor Type*

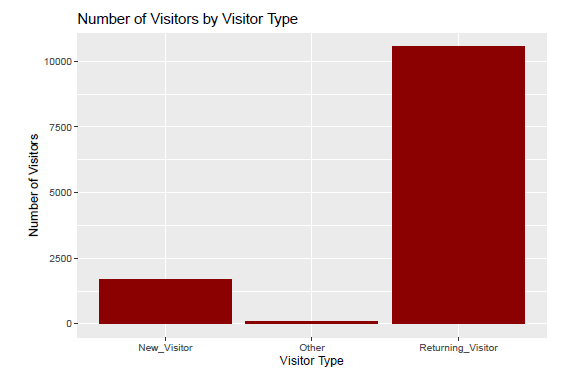


Figure 2

1. Correlation

cor(DF2)

Administrative Administrative\_Duration Informational Informational\_Duration ProductRelated

Administrative 1.000000000 0.601583342 0.376850429 0.255848140 0.431119340

Administrative\_Duration 0.601583342 1.000000000 0.302709709 0.238030789 0.289086621

Informational 0.376850429 0.302709709 1.000000000 0.618954862 0.374164291

Informational\_Duration 0.255848140 0.238030789 0.618954862 1.000000000 0.280046268

ProductRelated 0.431119340 0.289086621 0.374164291 0.280046268 1.000000000

ProductRelated\_Duration 0.373939013 0.355421954 0.387505306 0.347363577 0.860926836

BounceRates -0.223562630 -0.144170410 -0.116113616 -0.074066610 -0.204577633

ExitRates -0.316482998 -0.205797757 -0.163666061 -0.105275683 -0.292526283

PageValues 0.098989585 0.067608481 0.048631692 0.030860874 0.056281794

SpecialDay -0.094777598 -0.073303725 -0.048219254 -0.030576549 -0.023958175

OperatingSystems -0.006347063 -0.007343418 -0.009526668 -0.009578676 0.004289621

Browser -0.025034572 -0.015391527 -0.038234678 -0.019284981 -0.013145721

Region -0.005486805 -0.005560563 -0.029168638 -0.027144112 -0.038121842

TrafficType -0.033560713 -0.014376431 -0.034490754 -0.024674908 -0.043064304

VisitorType -0.025819710 -0.023939717 0.055827573 0.044676760 0.126655811

Weekend 0.026416750 0.014990142 0.035784725 0.024078486 0.016091964

Revenue 0.138917094 0.093586719 0.095200343 0.070344502 0.158537984

ProductRelated\_Duration BounceRates ExitRates PageValues SpecialDay OperatingSystems

Administrative 0.373939013 -0.223562630 -0.316482998 0.09898959 -0.094777598 -0.0063470633

Administrative\_Duration 0.355421954 -0.144170410 -0.205797757 0.06760848 -0.073303725 -0.0073434175

Informational 0.387505306 -0.116113616 -0.163666061 0.04863169 -0.048219254 -0.0095266679

Informational\_Duration 0.347363577 -0.074066610 -0.105275683 0.03086087 -0.030576549 -0.0095786764

ProductRelated 0.860926836 -0.204577633 -0.292526283 0.05628179 -0.023958175 0.0042896206

ProductRelated\_Duration 1.000000000 -0.184541115 -0.251984097 0.05282306 -0.036379845 0.0029757898

BounceRates -0.184541115 1.000000000 0.913004396 -0.11938603 0.072702253 0.0238231825

ExitRates -0.251984097 0.913004396 1.000000000 -0.17449831 0.102241802 0.0145667353

PageValues 0.052823063 -0.119386026 -0.174498310 1.00000000 -0.063541272 0.0185079466

SpecialDay -0.036379845 0.072702253 0.102241802 -0.06354127 1.000000000 0.0126522347

OperatingSystems 0.002975790 0.023823182 0.014566735 0.01850795 0.012652235 1.0000000000

Browser -0.007380440 -0.015772209 -0.004442355 0.04559192 0.003498747 0.2230128882

Region -0.033090520 -0.006485347 -0.008907006 0.01131530 -0.016097975 0.0767754856

TrafficType -0.036377170 0.078285541 0.078616331 0.01253169 0.052301443 0.1891536121

VisitorType 0.119329172 0.135536393 0.179143931 -0.11122783 0.085556612 0.0015042220

Weekend 0.007310614 -0.046513997 -0.062587048 0.01200164 -0.016767155 0.0002842506

Revenue 0.152372611 -0.150672912 -0.207071082 0.49256930 -0.082304598 -0.0146675596

Browser Region TrafficType VisitorType Weekend Revenue

Administrative -0.025034572 -0.0054868053 -0.033560713 -0.025819710 0.0264167503 0.138917094

Administrative\_Duration -0.015391527 -0.0055605628 -0.014376431 -0.023939717 0.0149901419 0.093586719

Informational -0.038234678 -0.0291686379 -0.034490754 0.055827573 0.0357847251 0.095200343

Informational\_Duration -0.019284981 -0.0271441124 -0.024674908 0.044676760 0.0240784862 0.070344502

ProductRelated -0.013145721 -0.0381218417 -0.043064304 0.126655811 0.0160919642 0.158537984

ProductRelated\_Duration -0.007380440 -0.0330905198 -0.036377170 0.119329172 0.0073106138 0.152372611

BounceRates -0.015772209 -0.0064853474 0.078285541 0.135536393 -0.0465139965 -0.150672912

ExitRates -0.004442355 -0.0089070060 0.078616331 0.179143931 -0.0625870480 -0.207071082

PageValues 0.045591919 0.0113152995 0.012531693 -0.111227826 0.0120016392 0.492569295

SpecialDay 0.003498747 -0.0160979746 0.052301443 0.085556612 -0.0167671553 -0.082304598

OperatingSystems 0.223012888 0.0767754856 0.189153612 0.001504222 0.0002842506 -0.014667560

Browser 1.000000000 0.0973928492 0.111938224 -0.021866988 -0.0402608638 0.023984289

Region 0.097392849 1.0000000000 0.047520231 -0.036190794 -0.0006906703 -0.011595068

TrafficType 0.111938224 0.0475202313 1.000000000 -0.002839178 -0.0022212292 -0.005112971

VisitorType -0.021866988 -0.0361907939 -0.002839178 1.000000000 -0.0436792493 -0.104725722

Weekend -0.040260864 -0.0006906703 -0.002221229 -0.043679249 1.0000000000 0.029295368

Revenue 0.023984289 -0.0115950678 -0.005112971 -0.104725722 0.0292953680 1.000000000

now let’s visualize the correlation

cor.matrix <- cor(DF2, method = "pearson", use = "complete.obs")

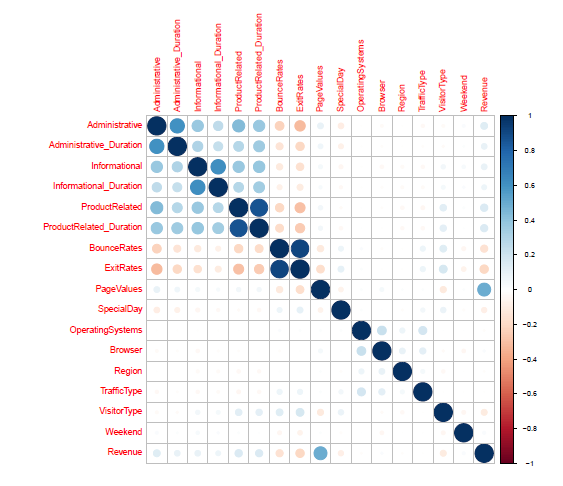
corrplot(cor.matrix)

Figure 3

The figure above shows that there is a correlation between the BounceRates and ExitRates and between ProductRelated and ProductRelated\_duration.

1. Relationship between ProductRelated and ProductRelated\_Duration

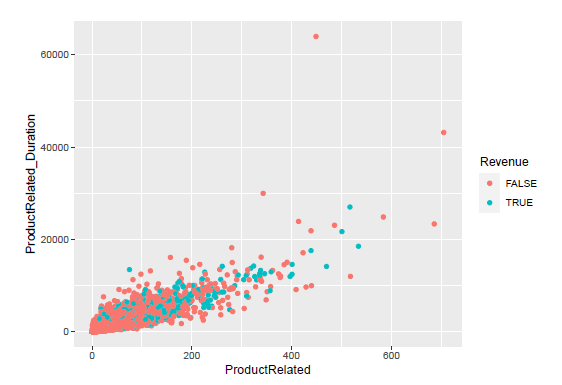


Figure 4

1. Relationship between ExitRates and BounceRates

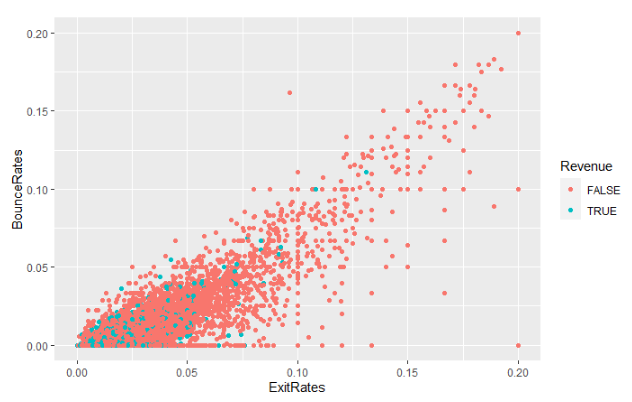
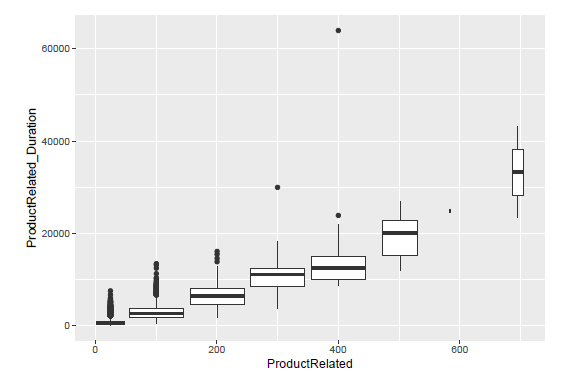
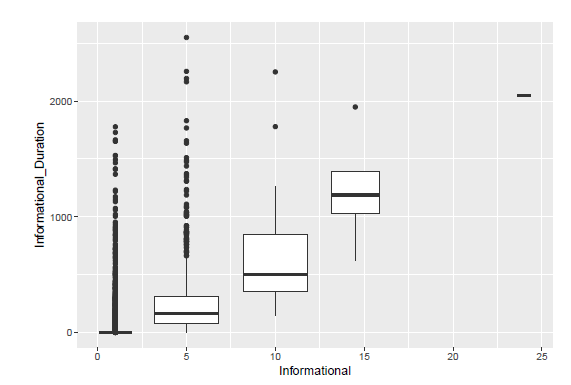


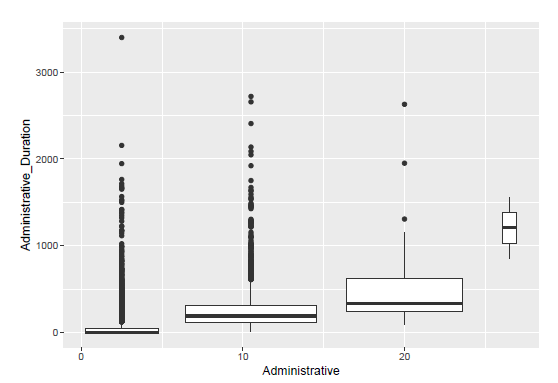
Figure 5

# Outliers

An outlier is an observation that is unlike the other observations. It is rare, or distinct, or does not fit in some way. Let's have a look at our data set using boxplot.







From the previous plots a couple of outliers can be seen but will not immensely affect the modelling.

# Feature Engineering and Data Transformation

Based on the above analysis and visualizations there are a few features that will need transformation before use in the models. The feature “Month” will be deleted because there are two months(January and April) that are missing from the dataset. VisitorType, Weekend and Revenue will also be transform into number fields to allow for ease of modelling.

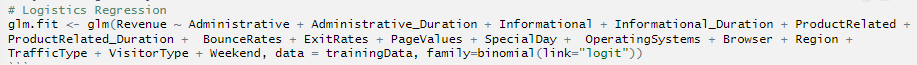
# Modelling

For this project two different types of modelling were utilized, classification and clustering. Principal Component Analysis(PCA) was modelled but did not provide a useful set of data. Results of the PCA modelling can be found in the accompanying pdf modelling file.

## Classification

### Logistic Regression

The first model that was run was Logistic Regression. The code to run this model in R is presented below.



The following is a summary of the model run:

Call:

glm(formula = Revenue ~ Administrative + Administrative\_Duration +

Informational + Informational\_Duration + ProductRelated +

ProductRelated\_Duration + BounceRates + ExitRates + PageValues +

SpecialDay + OperatingSystems + Browser + Region + TrafficType +

VisitorType + Weekend, family = binomial(link = "logit"),

data = trainingData)

Deviance Residuals:

Min 1Q Median 3Q Max

-6.7990 -0.7790 -0.0685 0.6928 2.7585

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -6.723e-01 2.557e-01 -2.629 0.00857 \*\*

Administrative 3.390e-02 1.897e-02 1.787 0.07399 .

Administrative\_Duration -4.102e-04 3.468e-04 -1.183 0.23688

Informational 5.601e-02 4.727e-02 1.185 0.23599

Informational\_Duration -1.023e-06 4.312e-04 -0.002 0.99811

ProductRelated 1.350e-03 2.225e-03 0.606 0.54419

ProductRelated\_Duration 1.832e-04 5.624e-05 3.258 0.00112 \*\*

BounceRates -3.530e+00 4.120e+00 -0.857 0.39156

ExitRates -1.049e+01 3.507e+00 -2.992 0.00277 \*\*

PageValues 1.103e-01 6.189e-03 17.824 < 2e-16 \*\*\*

SpecialDay -1.030e+00 3.157e-01 -3.262 0.00111 \*\*

OperatingSystems -1.381e-02 5.875e-02 -0.235 0.81422

Browser 1.002e-02 3.060e-02 0.327 0.74338

Region -5.756e-03 2.101e-02 -0.274 0.78406

TrafficType 1.285e-03 1.329e-02 0.097 0.92300

VisitorType -9.871e-02 7.163e-02 -1.378 0.16819

Weekend 1.451e-01 1.163e-01 1.248 0.21215

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

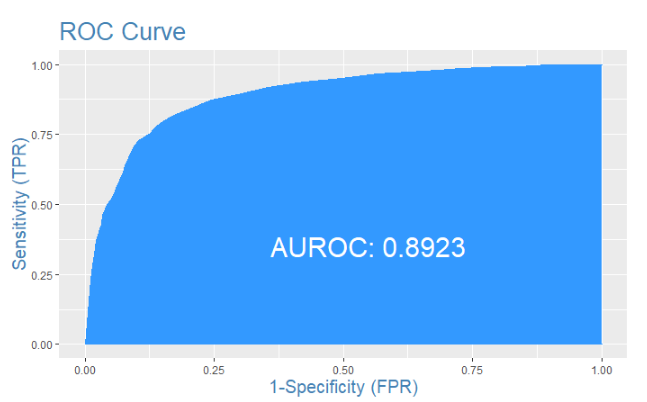
Null deviance: 3701.4 on 2669 degrees of freedom

Residual deviance: 2402.1 on 2653 degrees of freedom

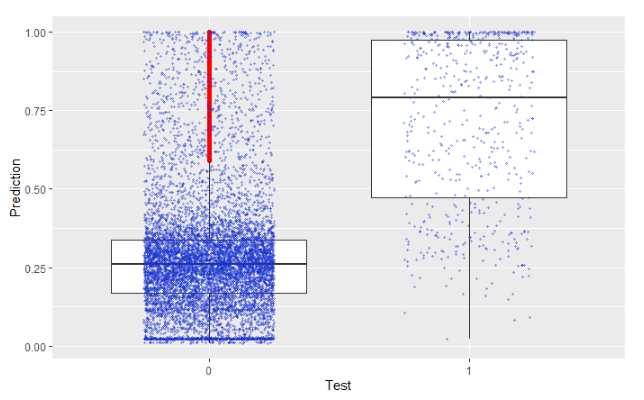
AIC: 2436.1

Number of Fisher Scoring iterations: 7

Here is a ROC curve for the modelling run.



The boxplot below shows the predicted value of the test dataset versus the training dataset.



The next modelling run is also a Logistic Regression but with balanced data.

Here is the summary of this run:

Call:

glm(formula = Revenue ~ Administrative + Administrative\_Duration +

Informational + Informational\_Duration + ProductRelated +

ProductRelated\_Duration + BounceRates + ExitRates + PageValues +

SpecialDay + OperatingSystems + Browser + Region + TrafficType +

VisitorType + Weekend, family = "binomial", data = trainingData\_Half)

Deviance Residuals:

Min 1Q Median 3Q Max

-7.0534 -0.8000 0.1475 0.6733 2.5272

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.1147605 0.3068514 -0.374 0.7084

Administrative 0.0336951 0.0245609 1.372 0.1701

Administrative\_Duration -0.0003363 0.0004471 -0.752 0.4520

Informational 0.1264286 0.0640812 1.973 0.0485 \*

Informational\_Duration -0.0003108 0.0005525 -0.562 0.5738

ProductRelated 0.0061021 0.0030889 1.976 0.0482 \*

ProductRelated\_Duration 0.0001459 0.0000738 1.977 0.0480 \*

BounceRates -7.6806237 4.6157768 -1.664 0.0961 .

ExitRates -5.4810975 3.9868023 -1.375 0.1692

PageValues 0.1162501 0.0082113 14.157 <2e-16 \*\*\*

SpecialDay -0.8604806 0.3733133 -2.305 0.0212 \*

OperatingSystems -0.0061384 0.0684751 -0.090 0.9286

Browser 0.0018834 0.0370022 0.051 0.9594

Region -0.0331115 0.0249131 -1.329 0.1838

TrafficType 0.0047324 0.0157099 0.301 0.7632

VisitorType -0.1423616 0.0855985 -1.663 0.0963 .

Weekend 0.1402929 0.1401397 1.001 0.3168

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

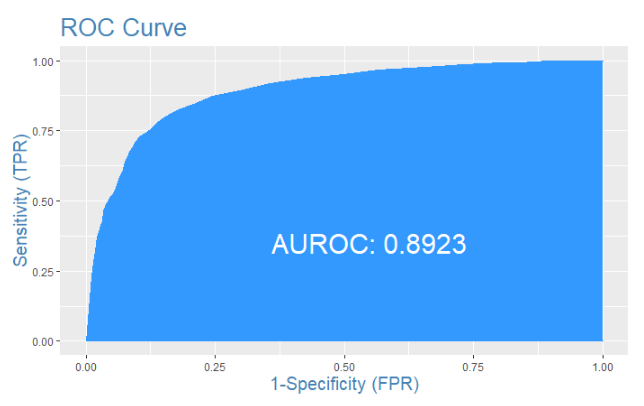
Null deviance: 2548.1 on 2001 degrees of freedom

Residual deviance: 1657.6 on 1985 degrees of freedom

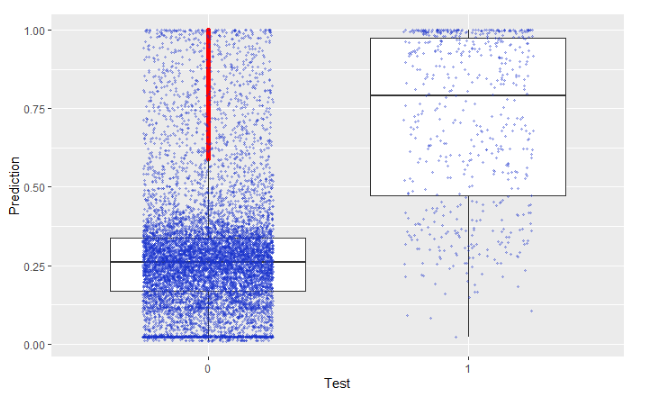
AIC: 1691.6

Number of Fisher Scoring iterations: 7

And the ROC Curve:



The boxplot below shows the predicted value of the test dataset versus the training dataset.



Here is a comparison of the confusion matrices:

First Run:

|  | **0**  <int> | **1**  <int> |  |  |
| --- | --- | --- | --- | --- |
| 0 | 8989 | 441 |  |  |
| 1 | 98 | 132 |  |  |

Second Run:

|  |
| --- |
|  |
|  | **0**  <int> | **1**  <int> |  |  |
| 0 | 9755 | 572 |  |  |
| 1 | 0 | 1 |  |  |

### Random Forest

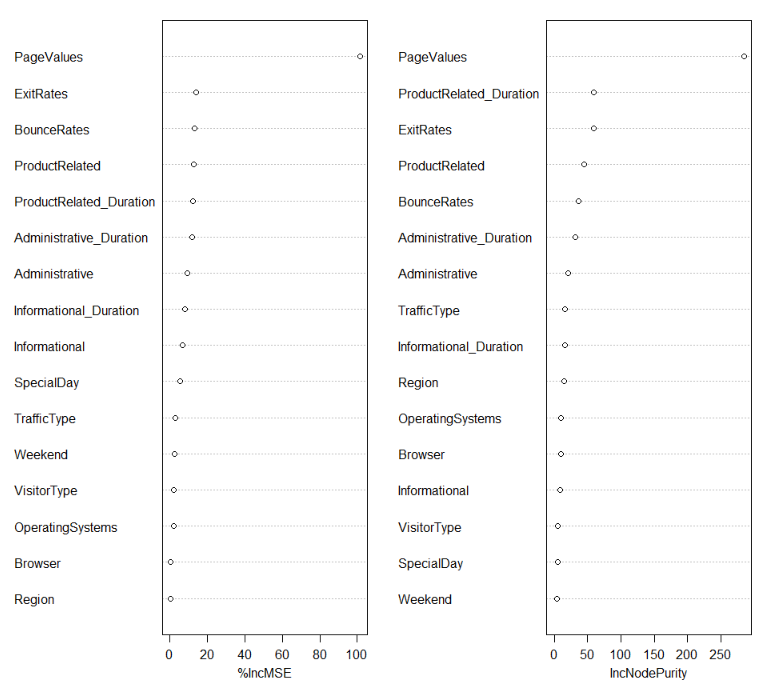
The next model that was tested with the data was a Random Forest model. The R code to run this model is shown below.

fitRF1 <- randomForest(

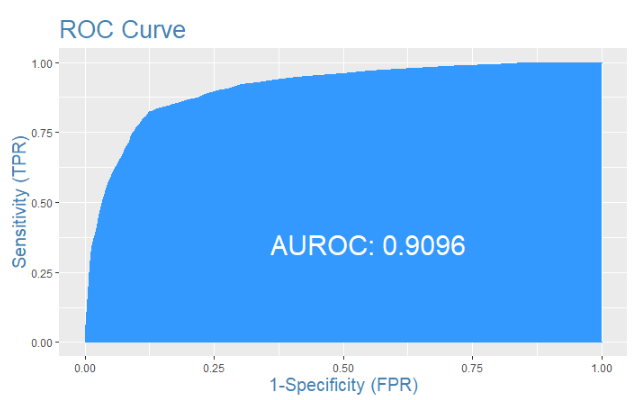
Revenue ~ ., method="anova",

data=trainingData[1:17], importance=TRUE, ntree=100)

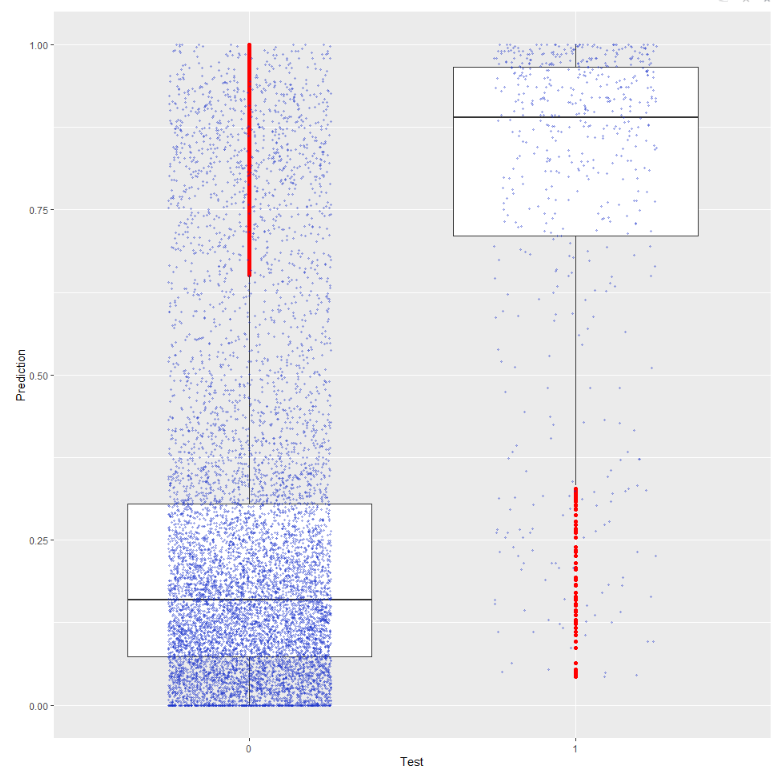
The following Figure shows the importance of Pagevalues in the outcome of Revenue.



And the ROC Curve:



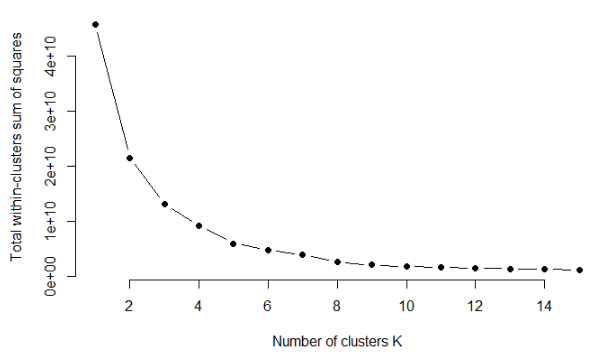
The boxplot below shows the predicted value of the test dataset versus the training dataset.



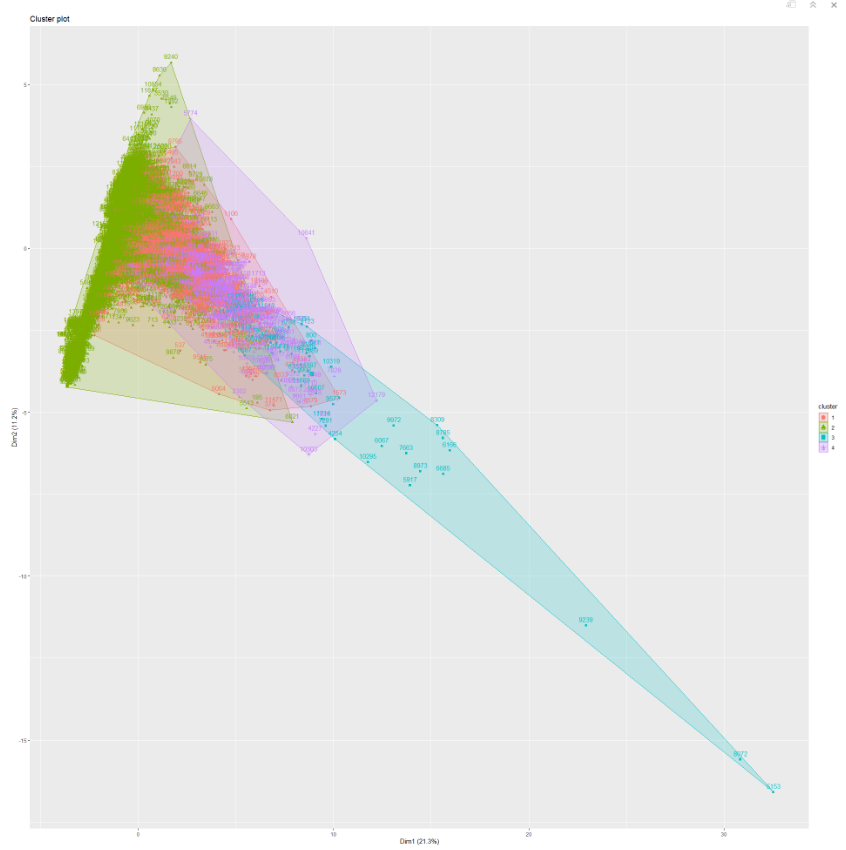
## Clustering

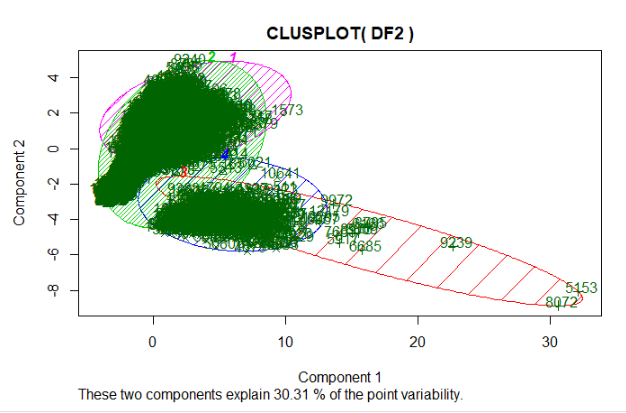
### k-means clustering

k-means was the first model to be run for the clustering component. Below is the graph used to determine the number of clusters for this model. The optimal number of clusters is 4.



Various plots showing distribution of k-means cluster modelling.

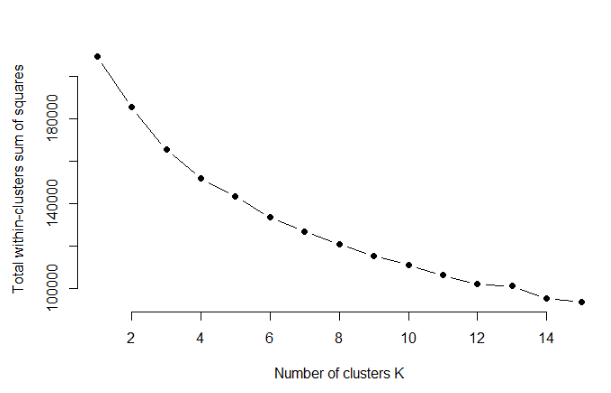




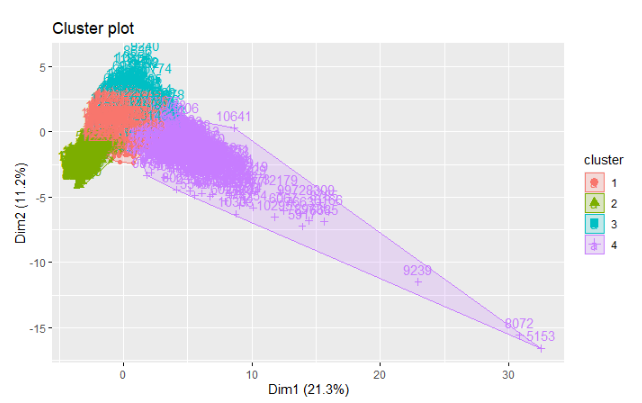
### k-means clustering with scaled data

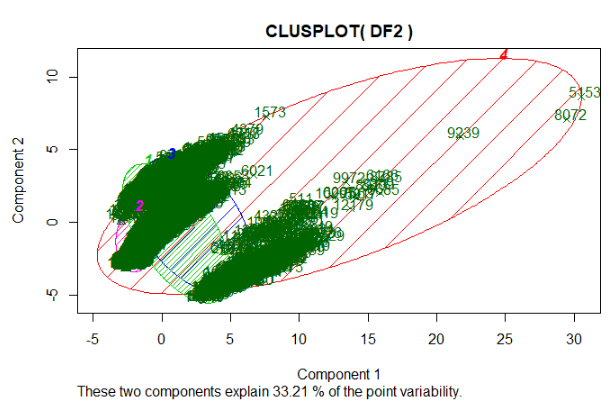
The second k-means model was run scaled data. Below is the graph used to determine the number of clusters for this model. The optimal number of clusters is 4.

.



Various plots showing distribution of k-means cluster modelling.





### Hierarchical Clustering

Hierarchical Clustering was also run as a means to determine the optimal clustering. The results of this model can be found in the associated R markdown file.

# Summary

After a thorough investigation of the data provided and a series of model runs, the Random Forest model provided the most accurate result for the classification portion of this report. At this time, the Random Forest will need to be refined before it can be implemented in a production environment. More study will also need to be undertaken before any recommendation about a clustering model can be made.

The Random Forest model would benefit from a more complete data set. This would include filling in the data for the missing months. Data from a longer time period will also help in producing a more accurate classification model.

Overall, this project was a partial success. First, we produced an algorithm that will, with some degree of accuracy, be able to predict if an online shopper would make a purchase or not. Second, we were also able to see where there are limitations in the whole process. Some of these limitations include the experience and knowledge of the team members. One benefit of this project was working as a group and learning how to overcome challenges that present themselves and also sharing in the successes of the project. The second benefit of this project was learning from each other and sharing the knowledge that we learned with other members of the group.

# References

Lipsman, A. (2019, June 27). *Global Ecommerce 2019*. Retrieved from eMarketer: https://www.emarketer.com/content/global-ecommerce-2019

MarketingCharts. (2015, January 2). *Global Retail and E-Commerce Sales Forecast, 2013-2018*. Retrieved from marketingcharts.com: https://www.marketingcharts.com/industries/retail-and-e-commerce-49733/attachment/emarketer-global-retail-and-ecommerce-sales-forecast-2013-2018-jan2015

*Online Shoppers Purchasing Intention Dataset Data Set.* (n.d.). Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

Orendorff, A. (2019, February 14). *Global Ecommerce Statistics and Trends to Launch Your Business Beyond Borders*. Retrieved from shopify.com: https://www.shopify.com/enterprise/global-ecommerce-statistics