E-Commerce Marketing Key Performance Indicators Analysis

YuTzu (Marie) Chen

Report submitted for the class project of
Introduction to Data Science

UCLA Extension
March 2020

Table of Contents

1. INTRODUCTION

1.1 Project Description

E-commerce (electronic commerce) is the activity of electronically buying or selling products on online services or over the Internet. It has been expanding every year and delivering huge contribution to the enterprises' total revenue. There are several key performance indicators, such as bounce rate, visitor type, and pageviews per visit, for marketing to help understand how good the company is doing in relation to marketing and advertising goals.

The hypothesis of this project is that it is more likely that customers end up with shopping when the page value is higher. Also, we would like to know how other features will affect the result(if the customers end up with shopping or not).

1.2 Dataset Description

The data set was downloaded from UCI Repository-Online Shoppers Purchasing Intention Dataset Data Set (https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset). There are 10 numerical features and 8 categorical features with 12,330 observations in the dataset(C. Okan Sakar, 2018). Features are shown in the following tables.

Table 1. Numerical features in the dataset

Feature Name	Description
Administrative	Number of pages visited by the visitor about account management
Administrative duration	Total amount of time (in seconds) spent by the visitor on account management related pages
Informational	Number of pages visited by the visitor about Web site, communication and address information of the shopping site
Informational duration	Total amount of time (in seconds) spent by the visitor on informational pages
Product related	Number of pages visited by visitor about product related pages
Product related duration	Total amount of time (in seconds) spent by the visitor on product related pages
Bounce rate*	Average bounce rate value of the pages visited by the visitor.
Exit rate*	Average exit rate value of the pages visited by the visitor
Page value*	Average page value of the pages visited by the visitor
Special day	Closeness of the site visiting time to a special day

*Bounce rate, exit rate, and page value are the metrics measured by “Google Analytics” for each page in the E-commerce site:

A. Bounce rate: 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

B. Exit rate: for all pageviews to the page, the percentage that were the last in the session

C. Page value: the average value for a page that a user visited before landing on the goal page or completing and E-commerce transaction (or both). Page value = (E-commerce revenue + total goal value) ÷ number of unique pageviews for given page

Table 2. Categorical features in the dataset

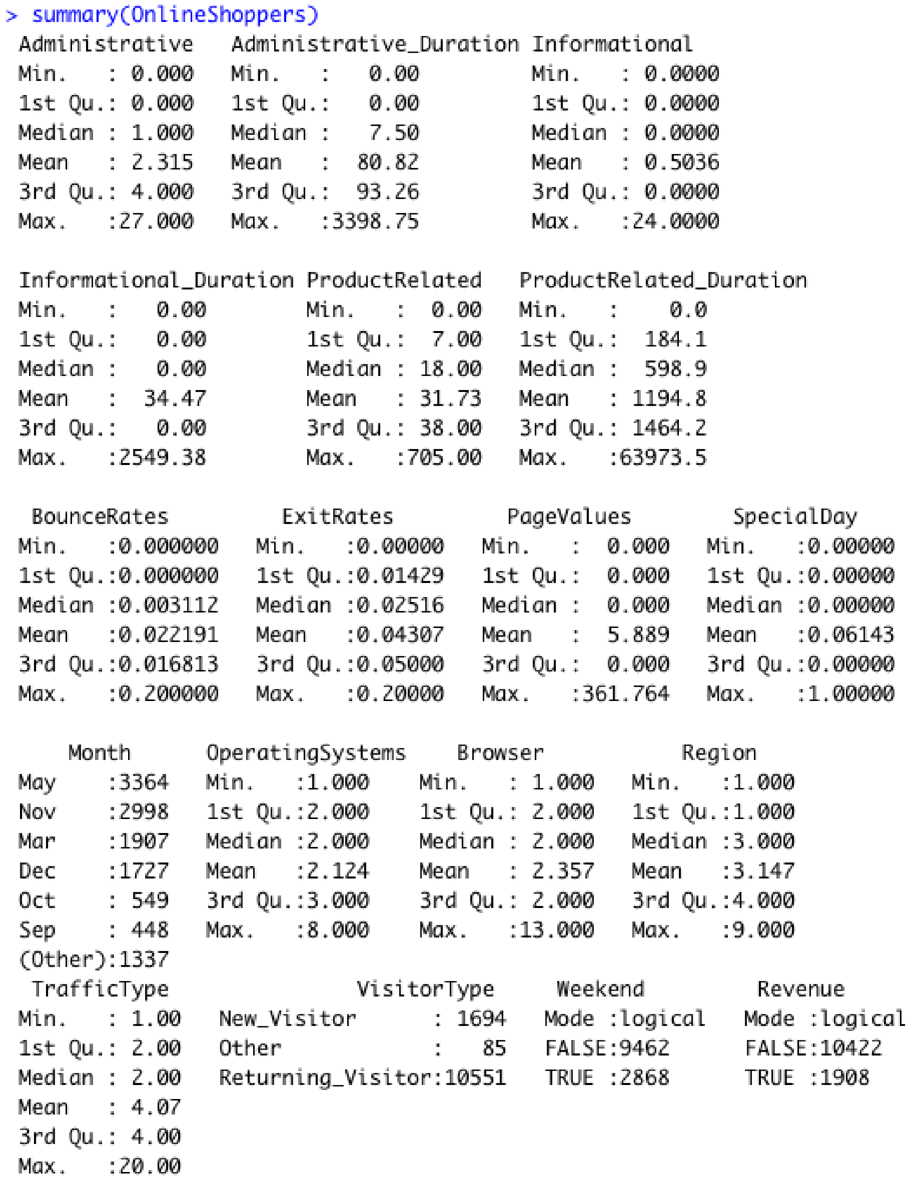
Feature Name	**Description**
Operating systems	Operating system of the visitor
Browser	Browser of the visitor
Region	Geographic region from which the session has been started by the visitor
Traffic type	Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)
Visitor type	Visitor type as ‘‘New Visitor,’’ ‘‘Returning Visitor,’’ and ‘‘Other’’
Weekend	Boolean value indicating whether the date of the visit is weekend
Month	Month value of the visit date
Revenue*	Class label indicating whether the visit has been finalized with a transaction

*Revenue: “True” indicates the positive class samples ending with shopping; while “False” indicates the negative class samples that did not end with shopping.

2. PREPARATION

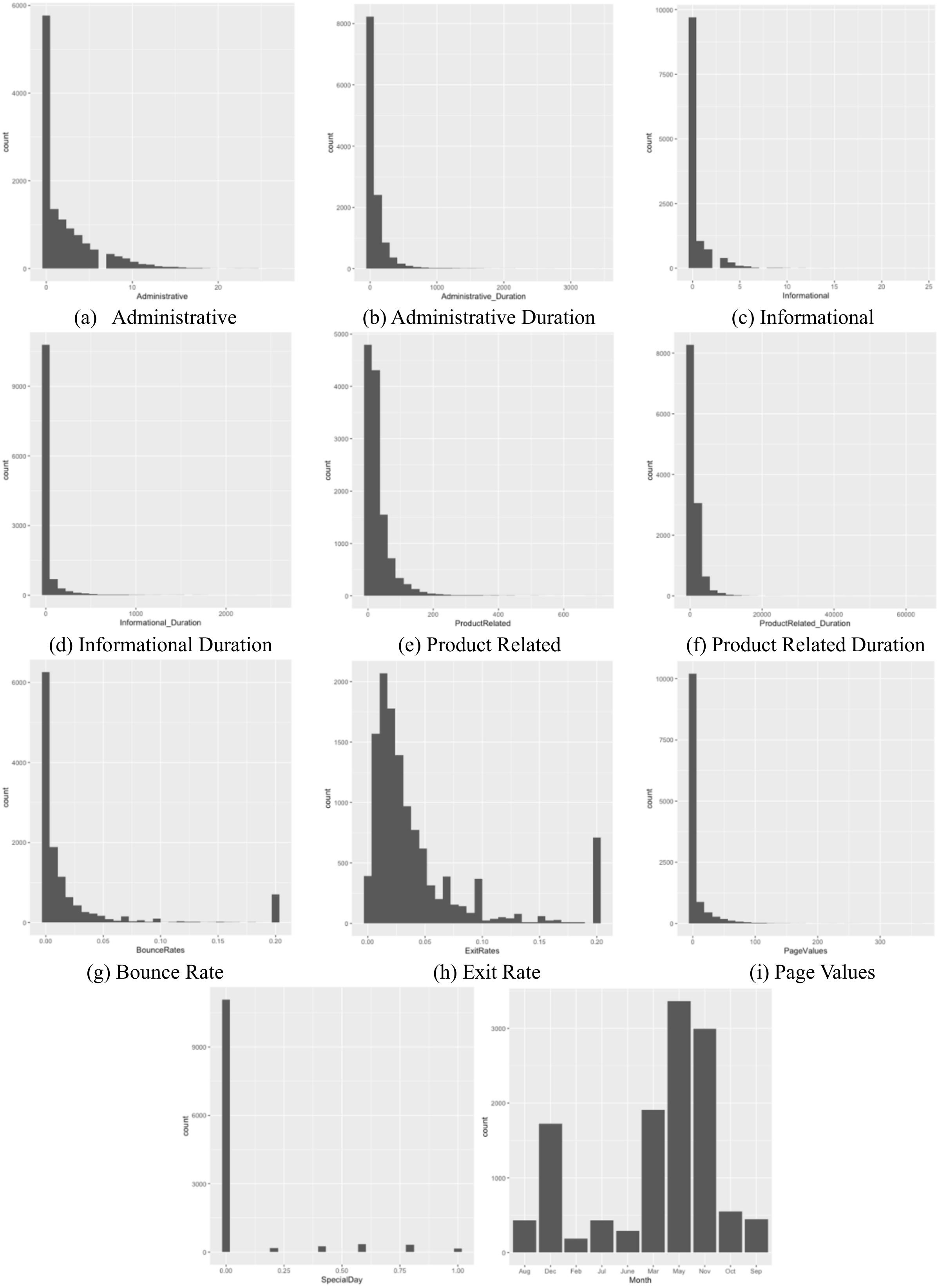
2.1 Basic Understanding of the Dataset

First of all, we would like to explore the basics of the distribution of numerical features and the levels of categorical features. “summary” is used to display the properties of the features and the response variable(Revenue).



```
> summary(OnlineShoppers)
 Administrative    Administrative_Duration Informational
 Min.   : 0.000   Min.   :   0.00         Min.   : 0.0000
 1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.0000
 Median : 1.000   Median :   7.50         Median : 0.0000
 Mean   : 2.315   Mean   :  80.82         Mean   : 0.5036
 3rd Qu.: 4.000   3rd Qu.:  93.26         3rd Qu.: 0.0000
 Max.   :27.000   Max.   :3398.75         Max.   :24.0000

 Informational_Duration ProductRelated    ProductRelated_Duration
 Min.   :   0.00        Min.   :  0.00   Min.   :    0.0
 1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  184.1
 Median :   0.00        Median : 18.00   Median :  598.9
 Mean   :  34.47        Mean   : 31.73   Mean   : 1194.8
 3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1464.2
 Max.   :2549.38        Max.   :705.00   Max.   :63973.5

  BounceRates         ExitRates         PageValues        SpecialDay
 Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   Min.   :0.00000
 1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   1st Qu.:0.00000
 Median :0.003112   Median :0.02516   Median :  0.000   Median :0.00000
 Mean   :0.022191   Mean   :0.04307   Mean   :  5.889   Mean   :0.06143
 3rd Qu.:0.016813   3rd Qu.:0.05000   3rd Qu.:  0.000   3rd Qu.:0.00000
 Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000

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

According to the result, there are no missing values(NAs) in the dataset. The levels of operating systems, browser, region, traffic type are not specified in the data repository nor C. Okan Sakar’s paper. Thus, the categorical features mentioned above will not be discussed in this project. Exploratory graphs can help us get a more comprehensive understanding of the data and guide us when making early decisions. Figures 1(a) to 1(n) present the histograms of numerical variables and bar plots of categorical variables.

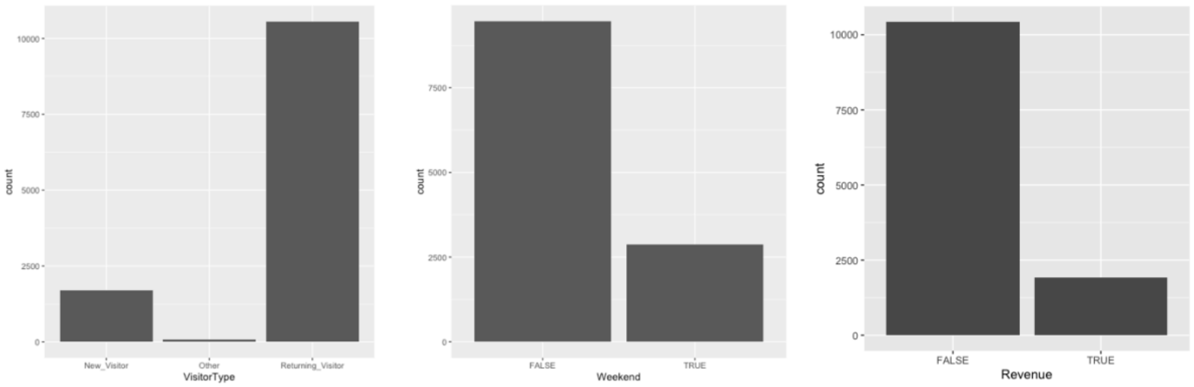


(a) Administrative (b) Administrative Duration (c) Informational

(d) Informational Duration (e) Product Related (f) Product Related Duration

(g) Bounce Rate (h) Exit Rate (i) Page Values

(j) Special Day (k) Month



(l) Visitor Type (m) Weekend (n) Revenue

Figure 1. Histograms of Numerical Variables and Bar Plots of Categorical Variables

2.2 Data Transformation

In C. Okan Sakar's paper, one of the Visitor Type, "Other", is not specified as well. We want to decrease ambiguity, so all the rows with "Other" in Visitor Type are deleted. In addition, the data of Administrative Duration, Informational Duration, and Product related Duration are converted to "minutes" instead of "seconds". Bounce Rates and Exit Rates will be expressed in percentage.

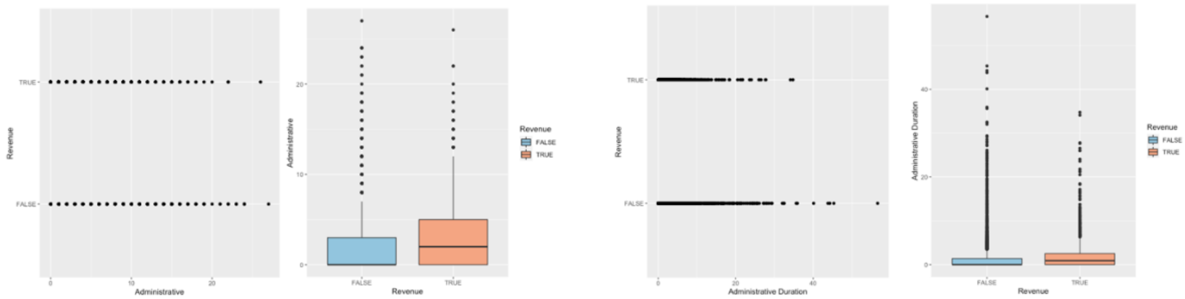
The histograms shown in Figure 1 indicate that all the numerical variables are positively-skewed distribution. Normally, data transformation, e.g. log transformation, square-root transformation and arcsine transformation, can be used to make data distribution conform to normality. However, after log-transformed, the distribution is still skewed to the right. Also, transformation might lead to misuse(Changyong FENG, 2014). The skewed distribution will not be transformed in this project.

3. DATA EXPLORATION

3.1 Relationship between Single Numerical Predictor and Response Variable

We would like to explore the relationship between each numerical predictor and Revenue in this part. Scatterplots and boxplots are useful for investigating the association between response variable and the other features. In addition, we will utilize logistic regression to see the relationship between Revenue and single feature.

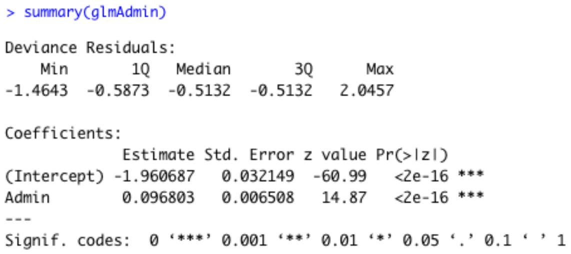
3.1.1 Administrative and Administrative Duration



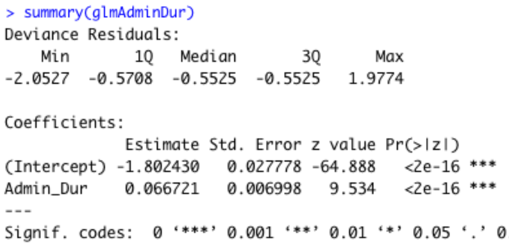
(a) Administrative vs. Revenue (b) Administrative Duration vs. Revenue

Figure 2. Scatterplots and Boxplots of Administrative and Administrative Duration vs. Revenue

Based on Figure 2, boxplots (Revenue=True and False) on the same scale are different from each other(different medians, quartiles and inter-quartile range) in both Administrative and Administrative Duration plots. It shows that Administrative and Administrative Duration have relationship with Revenue. In order to give a more quantitative clarification about the relationship, we use logistic regression.



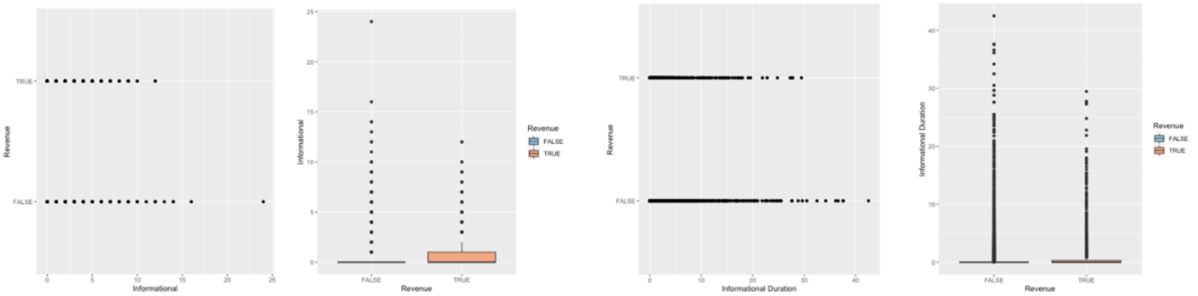
```
> summary(glmAdmin)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4643  -0.5873  -0.5132  -0.5132   2.0457

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.960687   0.032149  -60.99   <2e-16 ***
Admin        0.096803   0.006508   14.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
> summary(glmAdminDur)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0527  -0.5708  -0.5525  -0.5525   1.9774

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.802430   0.027778 -64.888   <2e-16 ***
Admin_Dur    0.066721   0.006998   9.534   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For every one unit change in Administrative, the log odds of Revenue=1 increases by 0.097. Also, for every one unit change in Administrative Duration, the log odds of Revenue=1 increases by 0.067. As the number of pages visited by the visitor about account management and amount of time spent by the visitor on account management related pages increase, the probability of Revenue=True increases. The relationship between Administrative and Revenue and the relationship between Administrative Duration and Revenue are significant(p-value < 2e-16). The result might because customer who intends to purchase the product will visit more pages and spend more time on account management. Thus, we will take these two variables into consideration for final model.

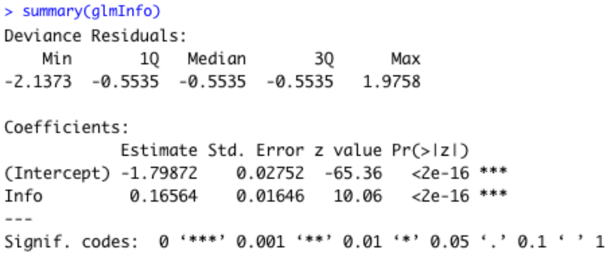
3.1.2 Informational and Informational Duration



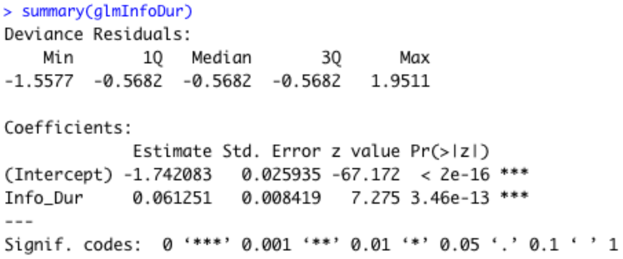
(a) Informational vs. Revenue (b) Informational Duration vs. Revenue

Figure 6. Scatterplots and Boxplots of Informational and Informational Duration vs. Revenue

Revenue=True and Revenue=False are quite different on the same scale in Informational. However, it hardly shows the inter-quartile box in Informational Duration. We will take a closer look at logistic regression for more information.



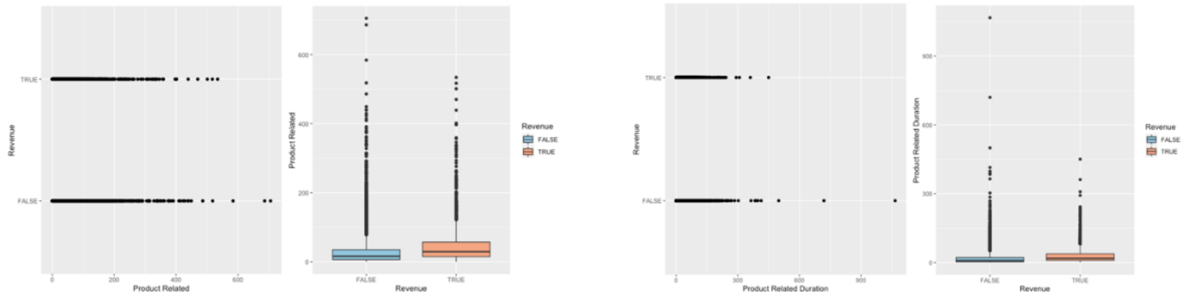
```
> summary(glmInfo)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1373  -0.5535  -0.5535  -0.5535   1.9758

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.79872    0.02752  -65.36   <2e-16 ***
Info         0.16564    0.01646   10.06   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
> summary(glmInfoDur)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5577  -0.5682  -0.5682  -0.5682   1.9511

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.742083   0.025935 -67.172  < 2e-16 ***
Info_Dur     0.061251   0.008419   7.275 3.46e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

For every one unit change in Informational, the log odds of Revenue=1 increases by 0.165. For every one unit change in Informational Duration, the log odds of Revenue=1 increases by 0.061. The number of pages visited about Web site, communication and address information of the shopping site and amount of time spent by the visitor on informational related pages increase, the probability of Revenue=True increases. The reason might also be similar to Section 3.1.1. We will also put these two variables into our final model.

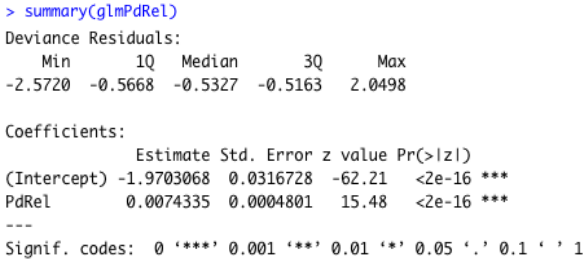
3.1.3 Product Related and Product Related Duration



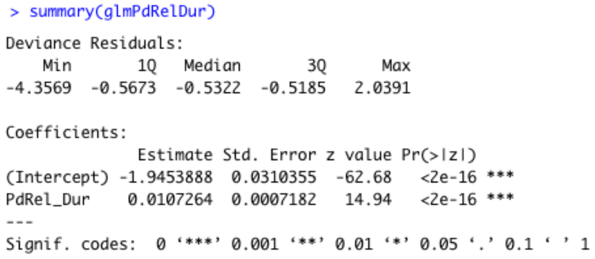
(a) Product Related vs. Revenue (b) Product Related Duration vs. Revenue

Figure 9. Scatterplots and Boxplots of Product Related and Product Related Duration vs. Revenue

In Figure 9(a), it is obvious that Revenue=True and Revenue=False are different; while, it's hardly to see the difference in the boxplot of Product Related Duration. We are still going to see the quantitative clarification.



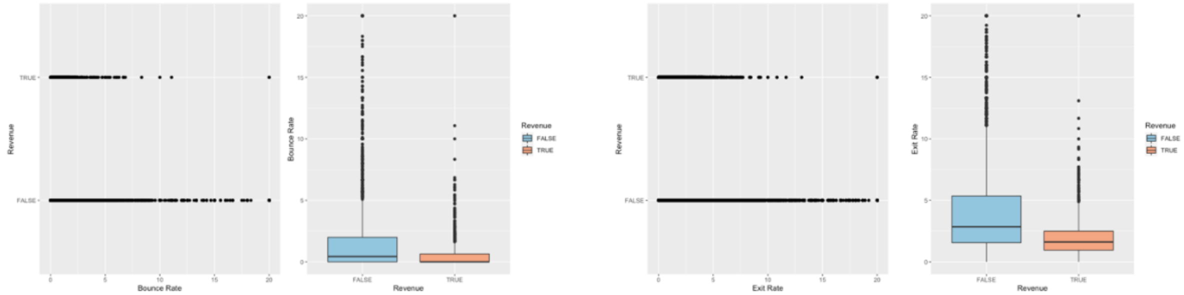
```
> summary(glmPdRel)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5720  -0.5668  -0.5327  -0.5163   2.0498

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.9703068  0.0316728  -62.21   <2e-16 ***
PdRel        0.0074335  0.0004801   15.48   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
> summary(glmPdRelDur)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.3569  -0.5673  -0.5322  -0.5185   2.0391

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.9453888  0.0310355  -62.68   <2e-16 ***
PdRel_Dur    0.0107264  0.0007182   14.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Every one unit change in Product Related, the log odds of Revenue=1 increases by 0.007. In addition, every one unit change in Product Related Duration, the log odds of Revenue=1 increases by 0.0107. As the number of pages visited by visitor about product related pages increases and amount of time spent by the visitor on product related pages increases, the probability of Revenue increases. Customer who is interested in a specific product may spend more time on product related information and will be likely to purchase the product. The p-values of two predictors are very significant(<2e-16), so two predictors will be put into the final model.

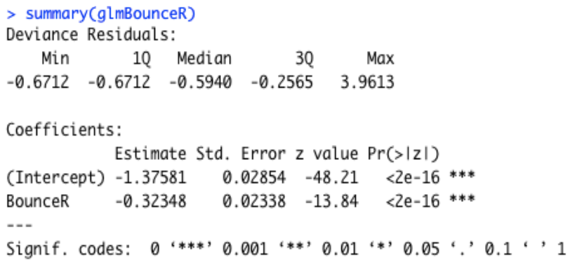
3.1.4 Bounce Rates and Exit Rates



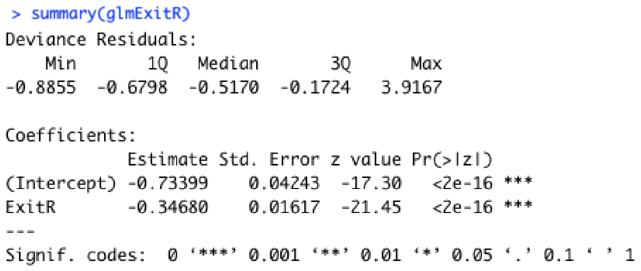
(a) Bounce Rates vs. Revenue (b) Exit Rates vs. Revenue

Figure 12. Scatterplots and Boxplots of Bounce Rates and Exit Rates vs. Revenue

From both scatterplots and boxplots of Bounce Rates and Exit Rates, there are significant difference between if the customers are going to purchase the product or not.



```
> summary(glmBounceR)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6712  -0.6712  -0.5940  -0.2565   3.9613

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.37581    0.02854  -48.21   <2e-16 ***
BounceR     -0.32348    0.02338  -13.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
> summary(glmExitR)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8855  -0.6798  -0.5170  -0.1724   3.9167

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.73399    0.04243  -17.30   <2e-16 ***
ExitR       -0.34680    0.01617  -21.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Every one unit change in Bounce Rates, the log odds of Revenue=1 decreases by 0.323, and every one unit change in Exit Rates, the log odds of Revenue=1 decreases by 0.347. When Bounce Rates and Exit Rates increase, the probability of customer purchasing the product will decrease. It is very easy to confuse these two rates. There is one significant difference. Exit rate is the percentage of visits that were the last in the session whereas bounce rate is the percentage of visits that were the only one of the session. In any case, lower bounce rates and lower exit rates will be better for E-commerce. Since the p-values of these two predictors are pretty small, we will also include Bounce Rates and Exit Rates in our final model.

3.1.5 Page Values

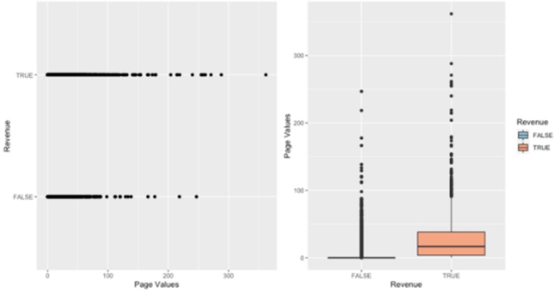
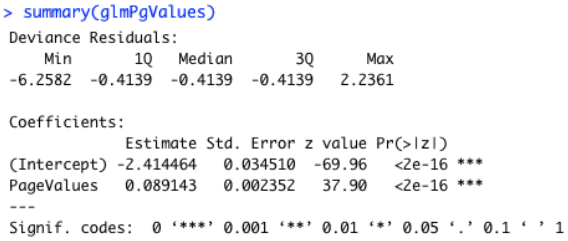


Figure 15. Scatterplots and Boxplots of Page Values vs. Revenue



```
> summary(glmPgValues)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.2582  -0.4139  -0.4139  -0.4139   2.2361

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.414464   0.034510  -69.96   <2e-16 ***
PageValues   0.089143   0.002352   37.90   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In the hypothesis, we assume that customers are likely to purchase the item if the Page Values are higher. According to the result, when Page Values increase, the probability of Revenue=1(buying the product) will increase. For every one unit change in Page Values, the probability of purchasing the product increases by 0.089. Page Values will be included in the final model as well.

3.1.6 Special Day

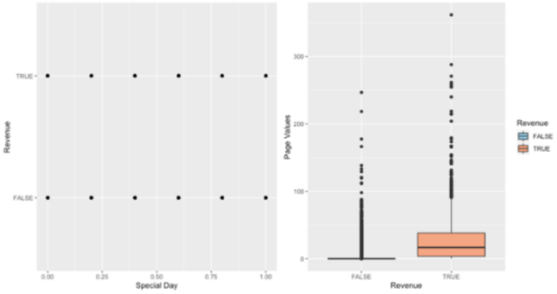
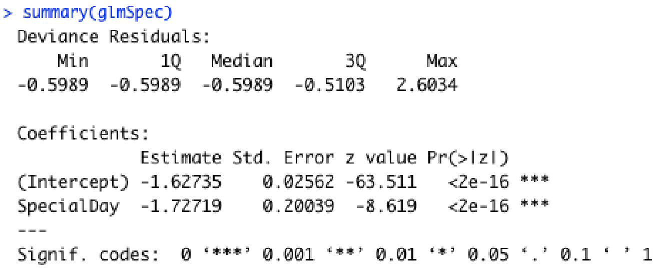


Figure 17. Scatterplots and Boxplots of Special Day vs. Revenue

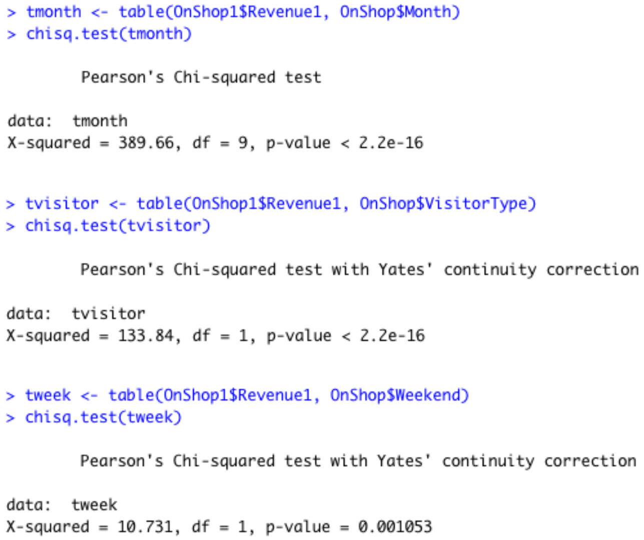


```
> summary(glmSpec)
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-0.5989  -0.5989  -0.5989  -0.5103   2.6034  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.62735    0.02562 -63.511   <2e-16 ***
SpecialDay  -1.72719    0.20039  -8.619   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The value of Special Day means the closeness of the site visiting time to a special day. The maximum value is 1. According to the data definition, the value is determined by considering the dynamics of E-commerce, such as the duration between the order date and delivery date. Thus, the higher value of Special Day doesn’t mean that it is closer to the Special Day. The only thing we know is that in this dataset, as Special Day value increases, the probability of Revenue=1(end up with shopping) decreases. The variable will be included in the final model since the p-value is very small.

3.2 Relationship between Single Categorical Predictor and Response Variable

There are three categorical predictors in the dataset that we’re interested in. Chi-square test will be employed to test the relationship between categorical predictors and response variable.



```
> tmonth <- table(OnShop1$Revenue1, OnShop$Month)
> chisq.test(tmonth)

	Pearson's Chi-squared test

data:  tmonth
X-squared = 389.66, df = 9, p-value < 2.2e-16

> tvisitor <- table(OnShop1$Revenue1, OnShop$VisitorType)
> chisq.test(tvisitor)

	Pearson's Chi-squared test with Yates' continuity correction

data:  tvisitor
X-squared = 133.84, df = 1, p-value < 2.2e-16

> tweek <- table(OnShop1$Revenue1, OnShop$Weekend)
> chisq.test(tweek)

	Pearson's Chi-squared test with Yates' continuity correction

data:  tweek
X-squared = 10.731, df = 1, p-value = 0.001053
```

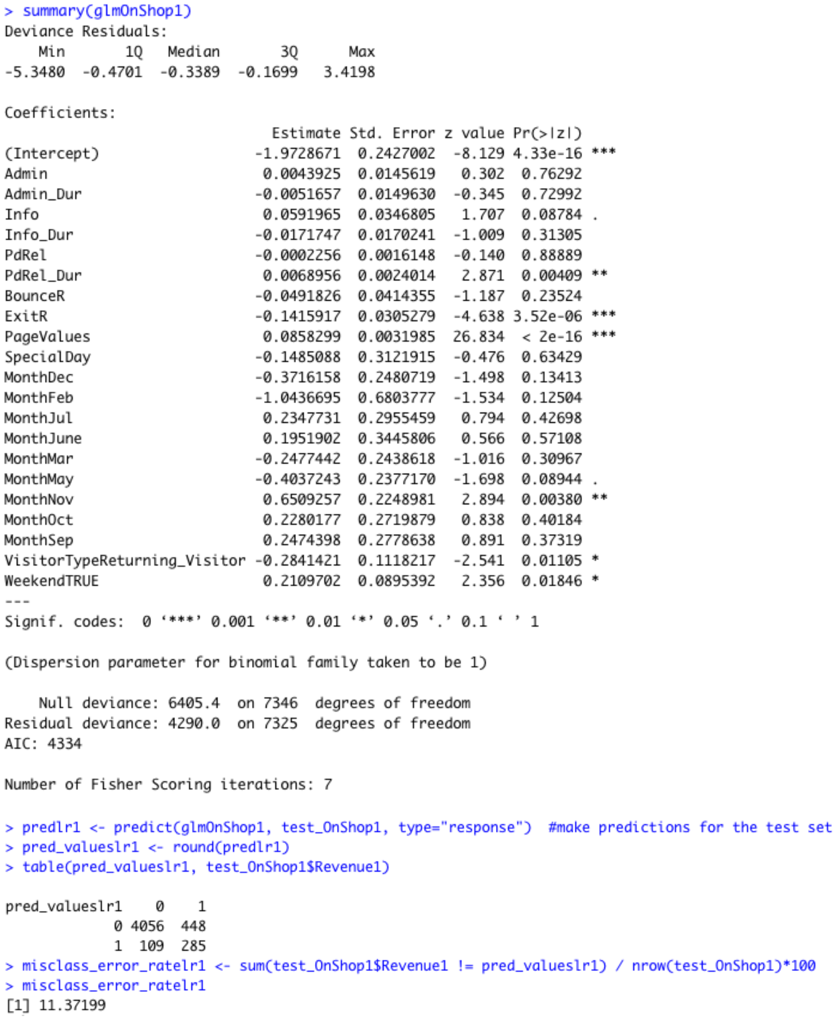
The chi-square test reveals that the relationship between Month and Revenue, Visitor Type and Revenue are significant, since p-values are both extremely small(<2e-16). The information gives us a hint that we can include Month and Revenue in the final model. As for “Weekend”, the p-value is 0.001053, which is still smaller than 0.05. It will be included in the model as well.

4. MACHINE LEARNING

In this part, we will predict Revenue by using machine learning algorithms based on the predictors we included from Section 3. First, we split the data into train dataset and test dataset. Three machine learning algorithms will be used to compare the error rates in this project.

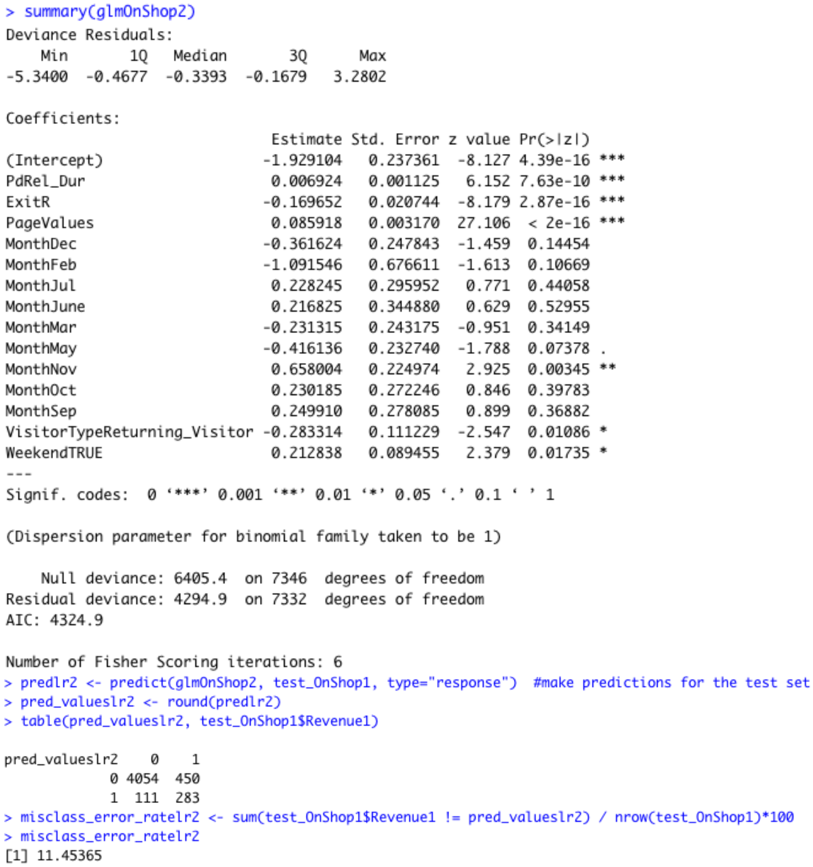
4.1 Logistic Regression

Logistic Regression can be used with categorical variables and with multiple predictors. The algorithm is often considered the most popular machine learning technique applied in solving classification problems.



```
> summary(glmOnShop1)
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-5.3480  -0.4701  -0.3389  -0.1699   3.4198  

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)    
(Intercept)               -1.9728671  0.2427002  -8.129 4.33e-16 ***
Admin                      0.0043925  0.0145619   0.302  0.76292    
Admin_Dur                 -0.0051657  0.0149630  -0.345  0.72992    
Info                       0.0591965  0.0346805   1.707  0.08784 .  
Info_Dur                  -0.0171747  0.0170241  -1.009  0.31305    
PdRel                     -0.0002256  0.0016148  -0.140  0.88889    
PdRel_Dur                  0.0068956  0.0024014   2.871  0.00409 ** 
BounceR                   -0.0491826  0.0414355  -1.187  0.23524    
ExitR                     -0.1415917  0.0305279  -4.638 3.52e-06 ***
PageValues                 0.0858299  0.0031985  26.834  < 2e-16 ***
SpecialDay                -0.1485088  0.3121915  -0.476  0.63429    
MonthDec                  -0.3716158  0.2480719  -1.498  0.13413    
MonthFeb                  -1.0436695  0.6803777  -1.534  0.12504    
MonthJul                   0.2347731  0.2955459   0.794  0.42698    
MonthJune                  0.1951902  0.3445806   0.566  0.57108    
MonthMar                  -0.2477442  0.2438618  -1.016  0.30967    
MonthMay                  -0.4037243  0.2377170  -1.698  0.08944 .  
MonthNov                   0.6509257  0.2248981   2.894  0.00380 ** 
MonthOct                   0.2280177  0.2719879   0.838  0.40184    
MonthSep                   0.2474398  0.2778638   0.891  0.37319    
VisitorTypeReturning_Visitor -0.2841421  0.1118217  -2.541  0.01105 *  
WeekendTRUE                0.2109702  0.0895392   2.356  0.01846 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6405.4  on 7346  degrees of freedom
Residual deviance: 4290.0  on 7325  degrees of freedom
AIC: 4334

Number of Fisher Scoring iterations: 7

> predlr1 <- predict(glmOnShop1, test_OnShop1, type="response")  #make predictions for the test set
> pred_valueslr1 <- round(predlr1)
> table(pred_valueslr1, test_OnShop1$Revenue1)
              
pred_valueslr1    0    1
             0 4056  448
             1  109  285
> misclass_error_ratelr1 <- sum(test_OnShop1$Revenue1 != pred_valueslr1) / nrow(test_OnShop1)*100
> misclass_error_ratelr1
[1] 11.37199
```

Once the model has been successfully trained, we can use the predict() function to make predictions for the test set and calculate the misclassification error rate. The error rate is 11.37%. From the logistic regression results, it can be noticed that some variables, such as Admin, Info_Dur, and PdRel, are not statistically significant. Keeping them in the model may contribute to overfitting. In regression, p-values less than the significance level indicate that the term is statistically significant. “Reducing the model” is the process of including all candidate variables in the model, and then repeatedly removing the single term with the highest non-significant p-value until the model contains only significant terms. In our revised logistic model, we exclude Admin, Admin_Dur, Info, Info_Dur, PdRel, BounceR, and SpecialDay.

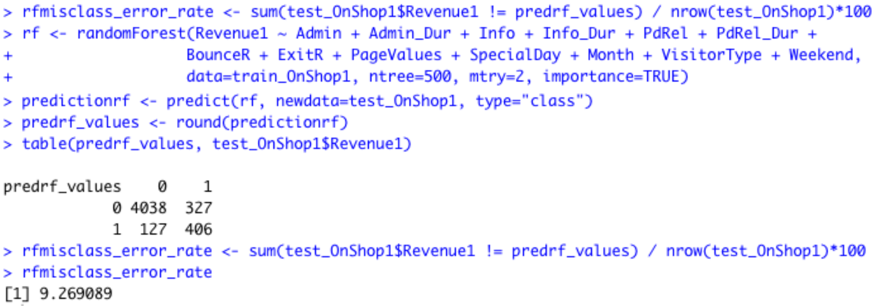


```
> summary(glmOnShop2)
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-5.3400  -0.4677  -0.3393  -0.1679   3.2802  

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)    
(Intercept)                  -1.929104   0.237361  -8.127 4.39e-16 ***
PdRel_Dur                     0.006924   0.001125   6.152 7.63e-10 ***
ExitR                        -0.169652   0.020744  -8.179 2.87e-16 ***
PageValues                    0.085918   0.003170  27.106  < 2e-16 ***
MonthDec                     -0.361624   0.247843  -1.459  0.14454    
MonthFeb                     -1.091546   0.676611  -1.613  0.10669    
MonthJul                      0.228245   0.295952   0.771  0.44058    
MonthJune                     0.216825   0.344880   0.629  0.52955    
MonthMar                     -0.231315   0.243175  -0.951  0.34149    
MonthMay                     -0.416136   0.232740  -1.788  0.07378 .  
MonthNov                      0.658004   0.224974   2.925  0.00345 ** 
MonthOct                      0.230185   0.272246   0.846  0.39783    
MonthSep                      0.249910   0.278085   0.899  0.36882    
VisitorTypeReturning_Visitor -0.283314   0.111229  -2.547  0.01086 *  
WeekendTRUE                   0.212838   0.089455   2.379  0.01735 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6405.4  on 7346  degrees of freedom
Residual deviance: 4294.9  on 7332  degrees of freedom
AIC: 4324.9

Number of Fisher Scoring iterations: 6
> predlr2 <- predict(glmOnShop2, test_OnShop1, type="response")  #make predictions for the test set
> pred_valueslr2 <- round(predlr2)
> table(pred_valueslr2, test_OnShop1$Revenue1)
              
pred_valueslr2    0    1
             0 4054  450
             1  111  283
> misclass_error_ratelr2 <- sum(test_OnShop1$Revenue1 != pred_valueslr2) / nrow(test_OnShop1)*100
> misclass_error_ratelr2
[1] 11.45365
```

The misclassification error rate increases (from 11.37% to 11.45%) might because of deleting non-significant variables from the previous model.

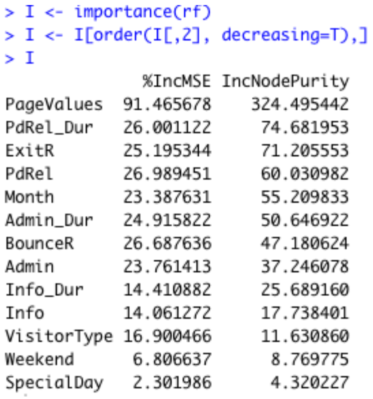
4.2 Random Forest

The random forest algorithm is one of the most popular and commonly used bagging methods. It is a tree-based algorithm and considered an “ensemble” method because it uses a set of classification trees that are calculated on random subsets of the data. According to the result, the misclassification error rate of Random Forest is 9.27%.



```
> rfmisclass_error_rate <- sum(test_OnShop1$Revenue1 != predrf_values) / nrow(test_OnShop1)*100
> rf <- randomForest(Revenue1 ~ Admin + Admin_Dur + Info + Info_Dur + PdRel + PdRel_Dur +
+                    BounceR + ExitR + PageValues + SpecialDay + Month + VisitorType + Weekend,
+                    data=train_OnShop1, ntree=500, mtry=2, importance=TRUE)
> predictionrf <- predict(rf, newdata=test_OnShop1, type="class")
> predrf_values <- round(predictionrf)
> table(predrf_values, test_OnShop1$Revenue1)
             
predrf_values    0    1
            0 4038  327
            1  127  406
> rfmisclass_error_rate <- sum(test_OnShop1$Revenue1 != predrf_values) / nrow(test_OnShop1)*100
> rfmisclass_error_rate
[1] 9.269089
```

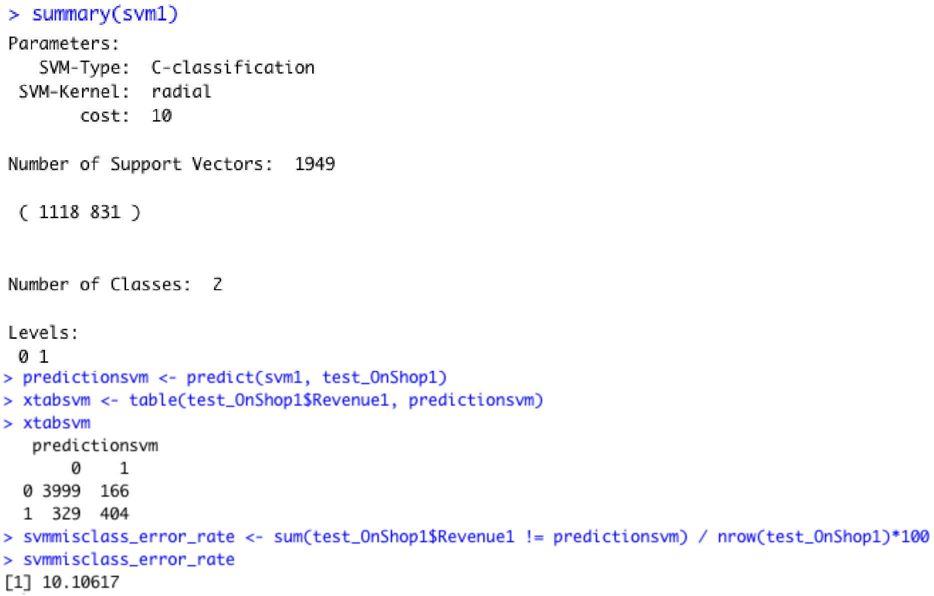
Moreover, we can use importance() function to figure out the importance of each predictors in the model. As we can see, PageValues plays a critical role in the model. Similar to logisitc regression, PdRel_Dur and ExitR are the next features to consider, though the importance is not as compelling as PageValues.



```
> I <- importance(rf)
> I <- I[order(I[,2], decreasing=T),]
> I
              %IncMSE IncNodePurity
PageValues  91.465678    324.495442
PdRel_Dur   26.001122     74.681953
ExitR       25.195344     71.205553
PdRel       26.989451     60.030982
Month       23.387631     55.209833
Admin_Dur   24.915822     50.646922
BounceR     26.687636     47.180624
Admin       23.761413     37.246078
Info_Dur    14.410882     25.689160
Info        14.061272     17.738401
VisitorType 16.900466     11.630860
Weekend      6.806637      8.769775
SpecialDay   2.301986      4.320227
```

4.3 Support Vector Machine (SVM)

SVMs have become quite popular in the data science community and tend to perform well in a variety of problem domains. We use the default type=C-classification, since it is for a classification machine. The kernel argument has a variety of possible types, including linear, polynomial, radial, and sigmoid. We use kernel="radial" (the default) for this multi-class classification problem. The result shows that the misclassification error rate is 10.16% when we use support vector machine algorithm.



```
> summary(svm1)
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  10

Number of Support Vectors:  1949

 ( 1118 831 )


Number of Classes:  2

Levels:
 0 1
> predictionsvm <- predict(svm1, test_OnShop1)
> xtabsvm <- table(test_OnShop1$Revenue1, predictionsvm)
> xtabsvm
   predictionsvm
       0    1
  0 3999  166
  1  329  404
> svmmisclass_error_rate <- sum(test_OnShop1$Revenue1 != predictionsvm) / nrow(test_OnShop1)*100
> svmmisclass_error_rate
[1] 10.10617
```

5. CONCLUSIONS

We can gain abundant information and conclusions from the analysis we did above. Most noteworthy of all, the feature we are interested in, PageValues, is a key part in the machine learning models. When we use PageValues as a single variable to predict that if customer is going to purchase the product, the p-value is statistically-significant. The probability of purchasing the product increases by 0.089, for every one unit increase in PageValues. In logistic regression model with multiple predictors, the probability of purchasing increases by 0.086 for one unit of PageValues increases. In addition, PageValues is the most essential feature in random forest model.

The definition of PageValues is the average value for a page that a user visited before landing on the goal page or completing and E-commerce transaction (or both). The PageValues is calculated by adding E-commerce revenue and total goal value, and divide the value by the number of unique pageviews for given page. The higher PageValues means the higher revenue; thus, it is highly related to our response variable, Revenue.

Last but not least, there are a few limitations in this project. Since we did not get the definition of some variables(e.g. operating systems, browser, region), we could not get the whole picture of the dataset. Also, we remove all rows with VisitorType=Other because it is not described. We would not know the effect of VisitorType=Other on the final model.

6. REFERENCE

(1) C. Okan Sakar, S. Olcay Polat, Mete Katircioglu, Yomi Kastro, *Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks*, Neural Computing and Applications, 2018

(2) Google Analytics: How Page Value is calculated; Exit Rate vs. Bounce Rate (https://support.google.com/analytics/answer/2695658?hl=en&ref_topic=6156780)

(3) Changyong FENG, Hongyue WANG, Naiji LU, Tian CHEN, Hua HE, Ying LU, and Xin M. TU, *Log-transformation and its implications for data analysis*, Shanghai Arch Psychiatry, 2014