Real-Time Prediction of Online Shoppers Purchasing Intention

Model: Logit Generalized Linear Model (GLM)

ImJaviPerez. UPV-EHU

June, 2021

Contents

Subject

We have created a logit Generalized Linear Model (GLM) for *Real-Time Prediction of Online Shoppers Purchasing Intention* based on the paper (Sakar et al. 2019). The data have been downloaded from (Dua and Graff 2017) *University of California Irvine (UCI) Machine Learning Repository.*

1 Introduction

The authors of the paper (Sakar et al. 2019) tell us that the dataset consists of feature vectors belonging to 12330 sessions. 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. Moreover, among the 12330 sessions in the dataset, 84.5% (10422) were negative class samples that did not finalize a transaction and the rest (1908) were positive class samples ending with purchasing.

The dataset consists of 10 numerical and 8 categorical attributes. The `Revenue` attribute can be used as the class label. `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.

The `Bounce Rate`, `Exit Rate` and `Page Value` features represent the metrics measured by "Google Analytics" for each page in the e-commerce site. The value of `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. The value of `Exit Rate` feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session. The `Page Value` feature represents the average value for a web page that a user visited before completing an e-commerce transaction.

The `Special Day` 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.

The dataset also includes `operating system`, `browser`, `region`, `traffic type`, `visitor type` (as `returning`, `new visitor` or `other`), a boolean value indicating whether the date of the visit is `weekend`, and `month` of the year.

Table 1: Features description, by (Sakar et al. 2019).

Feature	Description	Type
`Administrative`		Numerical
`Administrative_Duration`	URL information of the pages visited by	Numerical
`Informational`	the user: number of different types of pages	Numerical
`ProductRelated`	visited by the visitor in that session and	Numerical
`Informational_Duration`	total time spent in each of these page	Numerical
`ProductRelated_Duration`	categories	Numerical
`BounceRates`	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	Numerical
`ExitRates`	Feature for a specific web page calculated as for all pageviews to the page, the percentage that were the last in the session	Numerical
`PageValues`	The average value for a web page that a user visited before completing an ecommerce transaction	Numerical
`SpecialDay`	Closeness of the site visiting time to a special day	Numerical
`Month`	Month value of the visit date	Categorical
`OperatingSystems`	Operating system of the visitor	Categorical
`Browser`	Browser of the visitor	Categorical
`Region`	Geographic region from which the session has been started by the visitor	Categorical
`TrafficType`	Traffic source by which the visitor has arrived at the website	Categorical
`VisitorType`	Visitor type as `New_Visitor`, `Returning_Visitor` and `Other`	Categorical
`Weekend`	Boolean value indicating whether the date of the visit is weekend	Categorical
`Revenue`	Class label indicating whether the visit has been finalized with a transaction	Categorical

2 Generalized Linear Model (GLM)

We want to predict whether next customer visiting a web page will be a Revenue or not. We have selected this set of uncorrelated regressors with the most statistical significance: `Administrative`, `Informational`, `PageValues`, `SpecialDay`, `Month`, `OperatingSystems`, `TrafficType`, `VisitorType`, `ProductRelated_Duration`, `ExitRates`. and we have created a Logit Generalized Linear Model (GLM) to do that prediction. Please, find the appendix to see the whole variables selection proceedings.

3 Experiments and Results

The variable `Revenue` is an imbalanced class. We have an 84.5 % of negative class samples (`Revenue = FALSE`) and 15.5 % of positives (`Revenue = TRUE`) as we can see in next table:

Table 2: The variable class `Revenue` is imbalanced. Number of negative and positive class samples ending with purchasing (`Revenue = TRUE`).

FALSE	TRUE
10422	1908

There are post hoc sampling approaches that can help attenuate the effects of the imbalance during model training, so we are going to use three kind of sampling training data to implement the models (Kuhn and Johnson 2013)(p.427):

- the original data, which is imbalanced,
- a downsampled training data and
- an oversampled training data using the Synthetic Minority Oversampling Technique (SMOTE) methodology.

We use 70 % of dataset for training and the rest for validation. Moreover, to ensure statistical significance, this procedure is repeated 100 times with random training/validation partitions. We are going to show four average metrics to test the results of the models: accuracy, sensitivity (true positive rate), specificity (true negative rate) and F1-score.

Table 3: Average results of the experiments.

Dataset	Accuracy	Sensitivity	Specificity	F1_score
imbalanced	87.36	0.31	0.98	0.5
downsample	94.41	0.17	0.99	0.27
smote	87.23	0.29	0.98	0.5

4 Appendix. How to fit the GLM

Structure of the R's data frame:

```
## 'data.frame':    12330 obs. of  18 variables:
##  $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ProductRelated         : int  1 2 1 2 10 19 1 0 2 3 ...
##  $ ProductRelated_Duration: num  0 64 0 2.67 627.5 ...
```

```
##  $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
##  $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
##  $ Month                  : Factor w/ 10 levels "Aug","Dec","Feb",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ OperatingSystems       : Factor w/ 8 levels "1","2","3","4",..: 1 2 4 3 3 2 2 1 2 2 ...
##  $ Browser                : Factor w/ 13 levels "1","10","11",..: 1 6 1 6 7 6 8 6 6 8 ...
##  $ Region                 : Factor w/ 9 levels "1","2","3","4",..: 1 1 9 2 1 1 3 1 2 1 ...
##  $ TrafficType            : Factor w/ 20 levels "1","10","11",..: 1 12 14 15 15 14 14 16 14 12 ...
##  $ VisitorType            : Factor w/ 3 levels "New_Visitor",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Weekend                : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Revenue                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

4.1 GLM 1st model with every variable

Firstly we create a model including every variable.

```
# Model formula including every variable.  mod1_formula <- paste0('Revenue ~ ',
# paste(colnames(online_shoppers_df[,-which(names(online_shoppers_df) == 'Revenue')]),
# collapse = ' + '))
mod1_formula <- "Revenue~."
```

```
## [1] "Revenue~."
```

```
# Logit model including every variable.
logitMod1 <- glm(mod1_formula, data = online_shoppers_df, family = binomial(link = "logit"))
# summary (logitMod1 )
```

4.1.1 Analysis of variance. Chi-squared (χ^2) test

We create the analysis of variance (**anova**) tables using a χ^2 test, to find out the most significant variables of the logit model.

```
# Analysis of variance
anova_mod1 <- anova(logitMod1, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Revenue
##
## Terms added sequentially (first to last)
##
##
##                           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                      12329    10624.8
## Administrative             1   208.95     12328    10415.8 < 2.2e-16 ***
## Administrative_Duration    1     1.69     12327    10414.1  0.192964
## Informational              1    19.44     12326    10394.7 1.037e-05 ***
## Informational_Duration     1     0.87     12325    10393.8  0.351719
## ProductRelated             1    91.37     12324    10302.5 < 2.2e-16 ***
## ProductRelated_Duration    1     7.14     12323    10295.3  0.007528 **
## BounceRates                1   415.97     12322     9879.4 < 2.2e-16 ***
## ExitRates                  1   305.73     12321     9573.6 < 2.2e-16 ***
## PageValues                 1  2132.22     12320     7441.4 < 2.2e-16 ***
```

```
## SpecialDay                1    21.68     12319     7419.7 3.216e-06 ***
## Month                     9   240.09     12310     7179.6 < 2.2e-16 ***
## OperatingSystems          7    12.05     12303     7167.6  0.098792 .
## Browser                  11    12.80     12292     7154.8  0.306808
## Region                    8     6.21     12284     7148.6  0.623177
## TrafficType              19    62.50     12265     7086.1 1.548e-06 ***
## VisitorType               2     5.92     12263     7080.1  0.051797 .
## Weekend                   1     1.48     12262     7078.7  0.223405
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The most significant variables have a p-value < 0.1: `Administrative`, `Informational`, `ProductRelated`, `ProductRelated_Duration`, `BounceRates`, `ExitRates`, `PageValues`, `SpecialDay`, `Month`, `OperatingSystems`, `TrafficType`, `VisitorType`.

4.2 GLM 2nd Model with most significant variables

2nd Model with most significant variables:

```
# Select most significant variables
mod2_variables <- row.names(anova_mod1[which(anova_mod1$`Pr(>Chi)` < 0.1), ])
```

```
# 2nd Model formula with most significant variables
mod2_formula <- paste0("Revenue ~ ", paste(mod2_variables, collapse = " + "))
```

```
## [1] "Revenue ~ Administrative + Informational + ProductRelated + ProductRelated_Duration + BounceRat
```

```
# 2nd Model with most significant variables
logitMod2 <- glm(mod2_formula, data = online_shoppers_df, family = binomial(link = "logit"))
# summary (logitMod2 )
```

4.2.1 Analysis of variance. χ^2 test

We create the analysis of variance (`anova`) tables using a χ^2 test for this 2nd model, to find out the most significant variables of this new logit model.

```
# anova logit model 2
anova_mod2 <- anova(logitMod2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Revenue
##
## Terms added sequentially (first to last)
##
##
##                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                     12329    10624.8
## Administrative            1   208.95     12328    10415.8 < 2.2e-16 ***
## Informational             1    20.51     12327    10395.3 5.927e-06 ***
## ProductRelated            1    92.38     12326    10302.9 < 2.2e-16 ***
## ProductRelated_Duration   1     7.62     12325    10295.3  0.005762 **
## BounceRates               1   415.97     12324     9879.4 < 2.2e-16 ***
## ExitRates                 1   305.26     12323     9574.1 < 2.2e-16 ***
## PageValues                1  2132.20     12322     7441.9 < 2.2e-16 ***
```

```
## SpecialDay                1    21.57     12321     7420.3 3.417e-06 ***
## Month                     9   240.27     12312     7180.1 < 2.2e-16 ***
## OperatingSystems          7    12.19     12305     7167.9  0.094358 .
## TrafficType              19    61.28     12286     7106.6 2.424e-06 ***
## VisitorType               2     6.02     12284     7100.6  0.049196 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

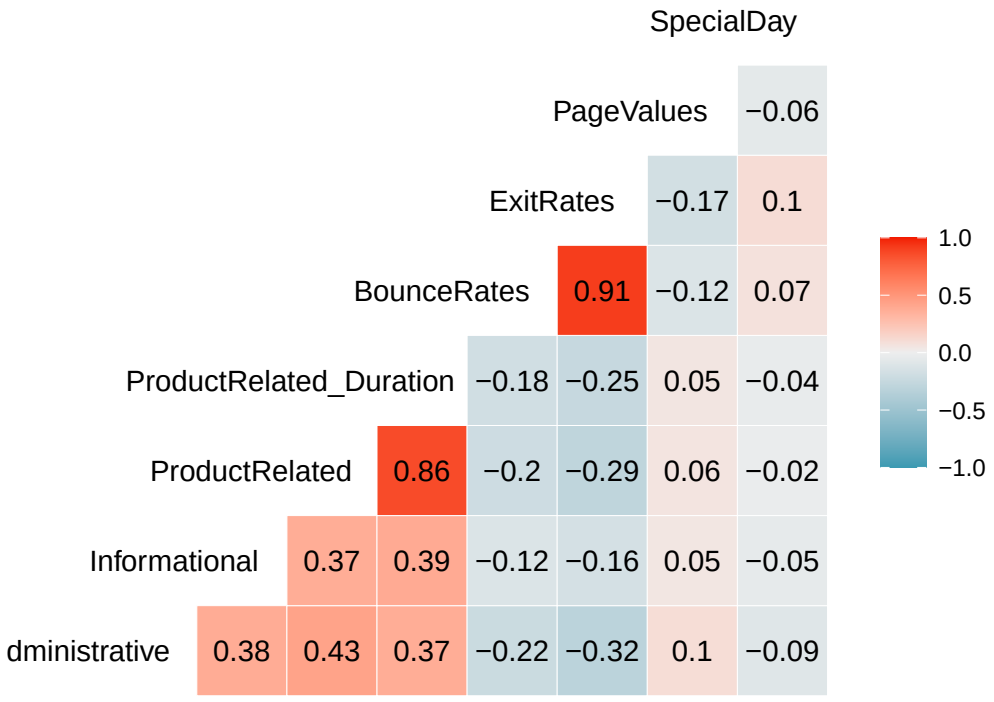
```
# Select most statistically significant variables
mod3_signif_variables <- row.names(anova_mod2[which(anova_mod2$`Pr(>Chi)` < 0.1), ])
```

Every variable have a $p-$value < 0.1, so all of them are statistically significant variables: `Administrative`, `Informational`, `ProductRelated`, `ProductRelated_Duration`, `BounceRates`, `ExitRates`, `PageValues`, `SpecialDay`, `Month`, `OperatingSystems`, `TrafficType`, `VisitorType`.

4.2.2 Avoid variables correlation

In order to improve the accuracy of the model we should avoid correlated variables (Peña 2017)(p. 429-442). We will do next tasks to find collinear variables:

1. Calculate correlation matrix.
2. Calculate Variance Inflation Factors (*VIF*) to identify the degree of multicollinearity of the predictor variables.



4.2.2.1 Correlation matrix

There is a high correlation between two couples of variables:

- Cor(`ProductRelated`, `ProductRelated_Duration`) = 0.86
- Cor(`BounceRates`, `ExitRates`) = 0.91

We have to choose only one of these variables: `ProductRelated` or `ProductRelated_Duration`, but never both of them. And we have to choose again between `BounceRates` or `ExitRates`. So we have four options:

1. (`ProductRelated`, `BounceRates`)
2. (`ProductRelated`, `ExitRates`)
3. (`ProductRelated_Duration`, `BounceRates`)

4. (`ProductRelated_Duration`, `ExitRates`).

We have selected: (`ProductRelated_Duration`, `ExitRates`).

4.2.2.2 Variance Inflation Factors (VIF)

Hence, the set of variables selected for the GLM are: `Administrative`, `Informational`, `PageValues`, `SpecialDay`, `Month`, `OperatingSystems`, `TrafficType`, `VisitorType`, `ProductRelated_Duration`, `ExitRates`. But we have to know whether these variables are collinear testing the Variance Inflation Factors (VIF):

Table 4: Variance Inflation Factors (VIF)

	GVIF	Df	GVIF^(1/(2*Df))
Administrative	1.3	1	1.14
Informational	1.3	1	1.14
PageValues	1.07	1	1.04
SpecialDay	1.26	1	1.12
Month	1.97	9	1.04
OperatingSystems	1.8	7	1.04
TrafficType	2.14	19	1.02
VisitorType	2	2	1.19
ProductRelated_Duration	1.43	1	1.2
ExitRates	1.17	1	1.08

We can see that every Generalized VIF (GVIF) is smaller than 4, so we can say that there is not collinearity in this set of predictors.

4.3 GLM 3rd Model

In this moment we have chosen en set of uncorrelated and statistically significant variables. Then, the formula of our last GLM is:

```
## [1] "Revenue ~ Administrative + Informational + PageValues + SpecialDay + Month + OperatingSystems +
```

This is the R function to create the model:

```
logitMod3 <- glm(mod3_formula, data = train_data, family = binomial(link = "logit"))
```

References

This document uses (R Core Team 2018).

Dua, Dheeru, and Casey Graff. 2017. "UCI Machine Learning Repository." University of California, Irvine, School of Information; Computer Sciences. http://archive.ics.uci.edu/ml.

Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling*. Vol. 26. Springer.

Peña, Daniel. 2017. *Regresión y Diseño de Experimentos*. Alianza editorial.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sakar, C Okan, S Olcay Polat, Mete Katircioglu, and Yomi Kastro. 2019. "Real-Time Prediction of Online Shoppers Purchasing Intention Using Multilayer Perceptron and LSTM Recurrent Neural Networks." *Neural Computing and Applications* 31 (10): 6893–6908.