Paper 3811-2019

Using Cost-Sensitive Predictive Modeling in Digital Marketing

Romana Sipoldova

ABSTRACT

In the present fast-moving times, when real-time decisions are needed, data quality is highly appreciated as high-quality predictive models are created based on it. Nowadays, analytics is turning into a mature, value-adding technology. A special benefit is considering future costs and revenues. In my final thesis work, which was finalized and submitted in March 2019, I pointed out the issue around increasingly evolving digital marketing to optimize the example online campaign. The aim of my project is to describe the procedure of the application of a strategy that is sensitive to profit/loss, and to compare the models of standard logistic regression and weighted logistic regression with the application of such a strategy. When perceiving the goal of a selected digital campaign from the advertiser's point of view, achieving the best click-through rate (CTR), the media agency must also consider the potential costs of cost per click (CPC) or the most effective profit. This balance (win-win strategy) needs to be understood and needs to be reached very sensitively. The goals I tackle in my model focus on setting bids optimizing CPC (cost per click). I have already taken cost-sensitive classification techniques into account while learning the model. They optimize the predictions of the target variable by specifying the costs.

INTRODUCTION

To make your online campaign as effective as possible it is necessary to reach your target audience as accurately as possible. The more accurate the targeting of the campaign is, the better price/ratio can be achieved.

TWO TYPES OF MEDIA BUYING

When advertisers want to buy advertising inventory, they have two options: traditional buying or programmatic buying. In traditional buying, an advertiser buys advertising inventory on a specific website. Depending on the affinity index of the website for the specific target audience, the advertiser can only hope to reach as many user that fall into his target audience as possible.

Programmatic buying does not generally focus on a specific website. The advertisement can be displayed to a specific user anywhere on the internet. Publishers sell advertising inventory through advertising exchange platforms in real time. They create a supply that meets with the demand of advertisers (or media agencies) and in real time, the advertisers can bid against each other for each specific ad impression which is then displayed to a specific user. The whole principle is comparable to buying and selling stock exchange shares. The owner of the advertising inventory offers the ad impression with a set floor price to the auction, and advertisers submit their bids. The ad impression is won by the highest bidder, although he will actually pay a price close to the second highest bid, so-called second-best price auction. Finally, the winner's ad is displayed on the publisher's website. This whole process only takes milliseconds to complete.

The goal is to create a model that, based on input variables, predicts the purchase of the most effective impressions in the price / quality ratio, which are most likely to result in a click. Because, in addition to focusing on the most effective impression for a reasonable price, you still need to track one of the core KPIs set for the business goal - CTR (Click Through Rate = Clicks / Impressions).

This paper illustrates the importance of cost tracking and their involvement in the model in an example. The dataset includes data from a sample online campaign and the goal is to create two predictive models that optimize Cost Per Click (CPC) bidding. The first model is a standard logistic regression model and the second model is a weighted logistic regression model that uses cost-sensitive predictive technique.

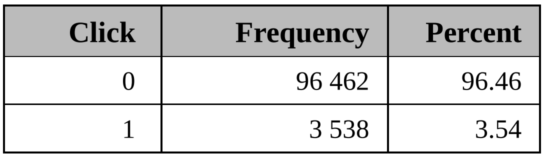
PREPARATION OF DATASET

The modeled variable - dependent variable is the binary variable Click - that is, whether the result of a bought impression is a click or not, meaning whether the user has clicked on the banner or not. Other variables entering the model - explanatory variables are:

- Day of week (Sunday = 0, *User_day*)
- Time [hours] (*User_hour*) – day of week and time in hours are variables which express the day and hour when the user encountered the ad
- Website domain of the impression (*Domain*)
- Size of the banner (*size*)
- Device (mobile, tablet, etc., *Device*)
- Placement position (*Position*)

Before you create both models, you need to edit the data file. Since it contains 100,000 observations, you must first make sure that the 0 and 1 ratio in the dependent variable is not too different, that is, whether you do not have the problem with deficiency of category 1 in dependent variable. Let's create a frequency table of the variable Click (Table 1):

```
title "Frequency of variable Click";
ods noproctitle;
proc freq data=library.input_table order=internal;
   table click / nocum;
run;
ods proctitle;
title;
```



Click	Frequency	Percent
0	96 462	96.46
1	3 538	3.54

Table 1. Frequency of variable *Click*

As you can see, the share of category 1 in dependent variable Click in the dataset is very small (only 3.54%). To solve this problem, you can use the *King and Zeng method* to edit the rare event data – take all observations where variable Click = 1 and use the Random sample method for the remaining observations to select the same number of observations for which Click = 0. This means that the result file contains 3 538 x 2 = 7 076 observations.

```
proc surveyselect data=library.input_table (where=(click = 0))
   out=work.random_sample
   method=srs
   n=3538;
run;

proc sql;
   create table library.data as
   select *
   from library.input_table
   where click=1
   union
   select *
   from
   random_sample;
quit;
```

You use new datafile DATA with 7 076 number of observations to create predictive models. Because of the complexity of next calculations, continue only with this dataset Data. The input variables are divided into the model according to the variable type:

- Dependent variable – *Click*
- Classification variables – *Position, Domain, User_day, Size* and *Device*
- Quantitative variable – *User_hour*.

Since the weights of weighted logistic regression are the costs of individual impressions, your predicted category in the dependent variable is category 0. It is because you do not just want to buy expensive clicks. Your goal is to buy the most effective click for the lowest possible impression price.

MODEL OF STANDARD LOGISTIC REGRESSION

Code for the Standard Logistic Regression Model:

```
ods graphics on;
data data_logreg1;
    set library.data;
run;

title "Logistic Regression Results";
proc logistic data=data_logreg1
    plots(only)=roc;
    class position (param=ref descending) domain (param=ref descending)
    user_day (param=ref descending) size (param=ref) device (param=ref);
    model click (event = '0')=user_hour domain user_day size device
    position          /
    selection=none
    rsquare
    link=logit;
    output out=work.pred_data_logreg1(label="Logistic Regression
    Predictions")
    predprobs=individual;
run;
title;
ods graphics off;
```

You now check the convergence criterion, the statistical significance of the model and the statistical significance of the individual variables.

Model is statistically significant, and all variables are also statistically significant (Pr > ChiSq; <.0001). Convergence criterion are satisfied so you can consider this model to be correct. The criteria / statistics that speak about model quality are in next part of this paper because comparation of created models is based on this criteria and statistics. Therefore, we now create a second model - a Weighted Logistic Regression Model.

MODEL OF WEIGHTED LOGISTIC REGRESSION

Weighted logistic regression is one of cost-sensitive learning techniques. By assigning weights to observations in the training set, weighting approaches achieve the cost-sensitive class distribution. Weights are typically determined by a frequency variable in the dataset. Our weighting variable is variable CPM.

Code for the Standard Logistic Regression Model:

```
ods graphics on;
data data_logreg2;
    set library.data;
run;
```

```
title "Logistic Regression Results";
proc logistic data=data_logreg2
    plots(only)=roc;
    class position (param=ref descending) domain (param=ref descending)
    user_day (param=ref descending) size (param=ref) device (param=ref);
    freq cpm;
    model click (event = '0')=user_hour domain user_day size device
    position           /
    selection=none
    rsquare
    link=logit;
    output out=work.pred_data_logreg2(label="Logistic regression
    predictions")
    predprobs=individual;
run;
title;
ods graphics off;
```

Again, make sure that this model is satisfied for all the necessary criteria.

MODEL COMPARISON

See the following tables for a comparison of each model. All tables and a figure in the first column belong to Model of Standard Logistic Regression (Table 2, Table 4, Table 6, Table 8, Table 10, Figure 1) and tables and a figure in the second column belong to Model of Weighted Logistic Regression (Table 3, Table 5, Table 7, Table 9, Table 11, Figure 2).

Model of Standard Logistic Regression

Criterion	Intercept Only	Intercept and Covariates
AIC	9 811.419	8 256.290
SC	9 818.283	10 871.651
-2 Log L	9 809.419	7 494.290

Table 2. Model Fit Statistics

Model of Weighted Logistic Regression

Criterion	Intercept Only	Intercept and Covariates
AIC	66 857.438	55 215.199
SC	66 866.264	58 224.838
-2 Log L	66 855.438	54 533.199

Table 3. Model Fit Statistics

R-Square	0.2790	Max-rescaled R-Square	0.3721

Table 4. R-Square and Max-rescaled R-Square

R-Square	0.2172	Max-rescaled R-Square	0.2955

Table 5. R-Square and Max-rescaled R-Square

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2 315.129	380	<.0001
Score	1 944.279	380	<.0001
Wald	1 107.724	380	<.0001

Table 6. Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	12 322.239	340	<.0001
Score	11 045.876	340	<.0001
Wald	6 619.183	340	<.0001

Table 7: Testing Global Null Hypothesis: BETA=0

Effect	DF	Wald Chi-Square	Pr > ChiSq
user_hour	1	4.879	0.0272
domain	359	602.726	<.0001
user_day	6	16.163	0.0129
size	10	70.565	<.0001
device	2	134.382	<.0001
position	2	26.896	<.0001

Table 8. Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
user_hour	1	32.325	<.0001
domain	319	4 160.985	<.0001
user_day	6	62.779	<.0001
size	10	642.130	<.0001
device	2	875.074	<.0001
position	2	200.307	<.0001

Table 9. Type 3 Analysis of Effects

Percent Concordant	79.8	Somers' D	0.596
Percent Discordant	20.2	Gamma	0.597
Percent Tied	0.0	Tau-a	0.298
Pairs	12517444	c	0.798

Table 10. Association of Predicted Probabilities and Observed Responses

Percent Concordant	76.3	Somers' D	0.526
Percent Discordant	23.7	Gamma	0.526
Percent Tied	0.0	Tau-a	0.248
Pairs	596771700	c	0.763

Table 11. Association of Predicted Probabilities and Observed Responses

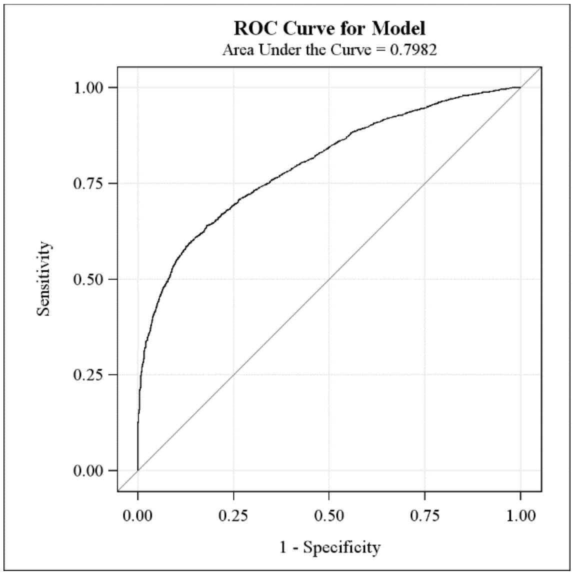


Figure 1. ROC Curve for Model of Standard Logistic Regression

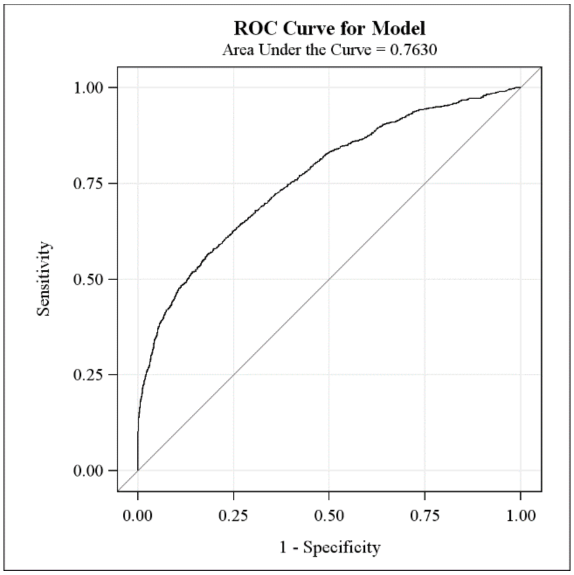
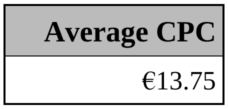


Figure 2. ROC Curve for Model of Weighted Logistic Regression

When you compare the individual models, you can see that from the statistical (analytical) point of view the model of standard logistic regression is performing better. However, is this also a better business model? Based on predicted values and total costs, you calculate the average CPC (Cost per click) for both models. First, you sum up the costs for the purchased impressions – the ones that the model has predicted to result in click. You then divide the sum by the count of correctly predicted clicks (those which are clicks in input and also in output - predicted by model). Run following code two times to generate two result tables:

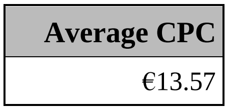
```
%let title=Standard /*Weighted */ Logistic Regression Model;
%let outputtab=pokus1 /*pokus2*/;

data &outputtab;
   set pred_data_logreg1 /*pred_data_logreg2*/;
   N+1;
   if first._INTO_="1" then Cena_CPM=0;
   if _INTO_="1" then Cena_CPM+cpm;
   if first._INTO_="1" and first._FROM_="1" then Click1_1=0;
   if _INTO_="1" and _FROM_="1" then Click1_1+1;
   CPC=Cena_CPM/Click1_1;
   format CPC euro7.2;
run;

title1 "Average CPC";
title2 &title;
proc print data=&outputtab noobs label;
var CPC;
label CPC="Average CPC";
where N=7076;
run;
```

You get two outputs – average CPC for a Model of Standard Logistic Regression (unweighted) and average CPC for a Model of Weighted Logistic Regression.



Average CPC
€13.75

Table 12. Standard Logistic Regression Model – Average CPC



Average CPC
€13.57

Table 13. Weighted Logistic Regression Model – Average CPC

From the results you can see that although the Unweighted Logistic Regression Model is performing better from the statistical (analytical) point of view according the criteria, from a business point of view considering also the cost of purchased impressions (CPM) the Model of Weighted Logistic Regression is better.

CONCLUSION

The conclusion of this modeling is that the model, which also takes into account the costs invested in the campaign, is more usable for the business strategy because it is more flexible and more sensitive to buying individual impressions.

REFERENCES

Verbeke, W., Baesens, B. and Bravo, C. 2018. *Profit-Driven Business Analytics*. Hoboken, NJ: John Wiley & Sons.

Allison, P. D. 2012. *Logistic Regression Using SAS®: Theory and Application*. 2nd ed. Cary, NC: SAS Institute Inc.

RECOMMENDED READING

- *Logistic Regression in Rare Events Data*

  https://gking.harvard.edu/files/0s.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Romana Šipoldová
romana.sipoldova@gmail.com