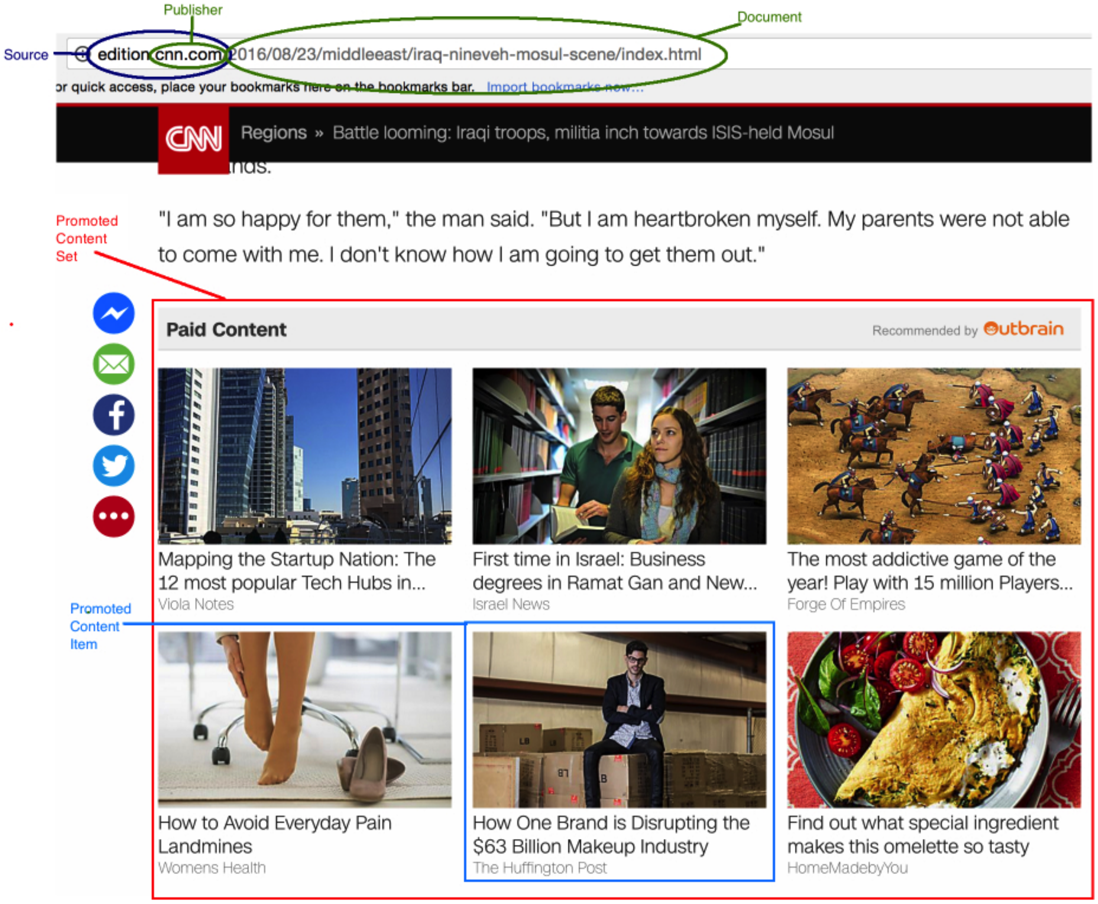
Click or not

group 2

(1)Problem Description

- The internet is a stimulating treasure trove of possibility. Every day we stumble on news stories relevant to our communities or experience the serendipity of finding an article covering our next travel destination.
- We are challenged to predict which pieces of content its global base of users are likely to click on.



edition.cnn.com/2016/08/23/middleeast/iraq-nineveh-mosul-scene/index.html

or quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...

Regions » Battle looming: Iraqi troops, militia inch towards ISIS-held Mosul

nds.

"I am so happy for them," the man said. "But I am heartbroken myself. My parents were not able to come with me. I don't know how I am going to get them out."

Paid Content

Recommended by Outbrain

Mapping the Startup Nation: The 12 most popular Tech Hubs in...
Viola Notes

First time in Israel: Business degrees in Ramat Gan and New...
Israel News

The most addictive game of the year! Play with 15 million Players...
Forge Of Empires

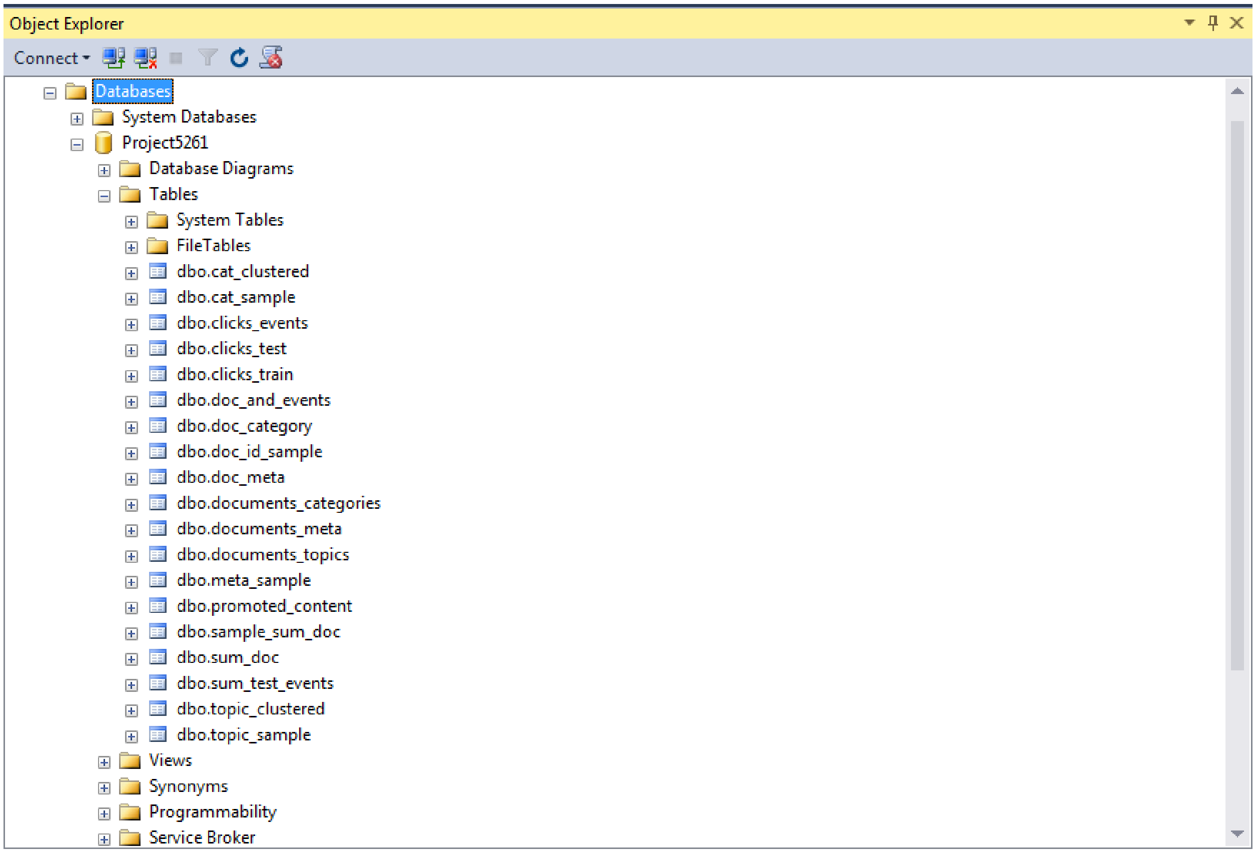
How to Avoid Everyday Pain Landmines
Womens Health

How One Brand is Disrupting the $63 Billion Makeup Industry
The Huffington Post

Find out what special ingredient makes this omelette so tasty
HomeMadebyYou

(2)Data

- Data comes from Kaggle.com
- Raw data includes Content of websites, information of users who browsed the website and information of ads on the websites.
- Data contains all kinds of information, which is challenging and interesting.
- Given the information of website, users and ads on the website, We will predict the probility of each ad being clicked and use Mean Average Precision to evaluate the result.



```
⊟ Databases
    ⊞ System Databases
    ⊟ Project5261
        ⊞ Database Diagrams
        ⊟ Tables
            ⊞ System Tables
            ⊞ FileTables
            ⊞ dbo.cat_clustered
            ⊞ dbo.cat_sample
            ⊞ dbo.clicks_events
            ⊞ dbo.clicks_test
            ⊞ dbo.clicks_train
            ⊞ dbo.doc_and_events
            ⊞ dbo.doc_category
            ⊞ dbo.doc_id_sample
            ⊞ dbo.doc_meta
            ⊞ dbo.documents_categories
            ⊞ dbo.documents_meta
            ⊞ dbo.documents_topics
            ⊞ dbo.meta_sample
            ⊞ dbo.promoted_content
            ⊞ dbo.sample_sum_doc
            ⊞ dbo.sum_doc
            ⊞ dbo.sum_test_events
            ⊞ dbo.topic_clustered
            ⊞ dbo.topic_sample
        ⊞ Views
        ⊞ Synonyms
        ⊞ Programmability
        ⊞ Service Broker
```

Figure 1:

$$MAP@12 = \frac{1}{|U|} \sum_{u=1}^{|U|} \sum_{k=1}^{min(12,n)} P(k)$$

Figure 2:

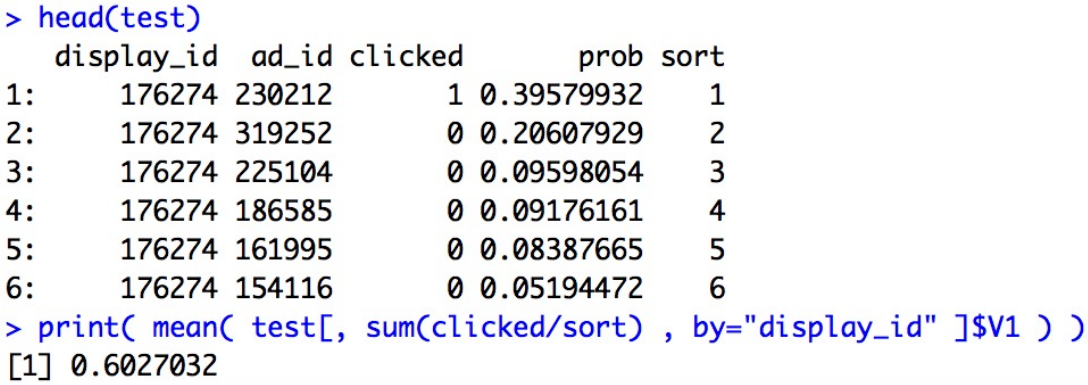
(3)First Attempt

- Firstly we tried a relatively Bayesian method which focus on the click itself.

$$W = \frac{Rv + Cm}{v + m}$$

Figure 3:

- R is the average clicked rate of an ad.
- v is the times an ad was displayed.
- C is the mean clicked rate around all ads in the data.
- m is the minimum displaying times required for an ad.
- The method is very efficient and the result is not bad.



```
> head(test)
   display_id  ad_id clicked       prob sort
1:     176274 230212       1 0.39579932    1
2:     176274 319252       0 0.20607929    2
3:     176274 225104       0 0.09598054    3
4:     176274 186585       0 0.09176161    4
5:     176274 161995       0 0.08387665    5
6:     176274 154116       0 0.05194472    6
> print( mean( test[, sum(clicked/sort) , by="display_id" ]$V1 ) )
[1] 0.6027032
```

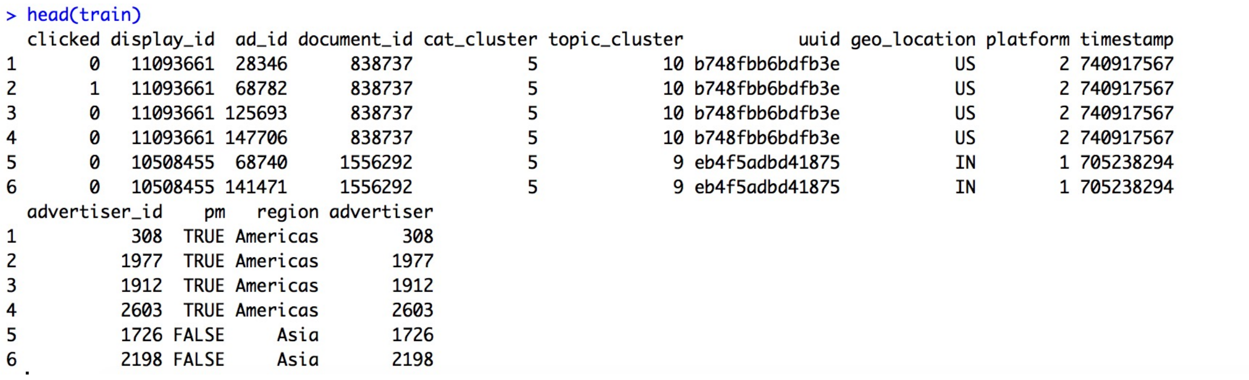
Figure 4:

(4)Advanced Exploation

- We want to use more information in the dataset to get a more precise result.
- Logistics Regression is a good way to predict probility.

Data Processing

- The raw data is way too large to process in R. SQL is the only choice.
- Delete unrelative information
- Use k-means cluster to reduce the number of catagories in data.

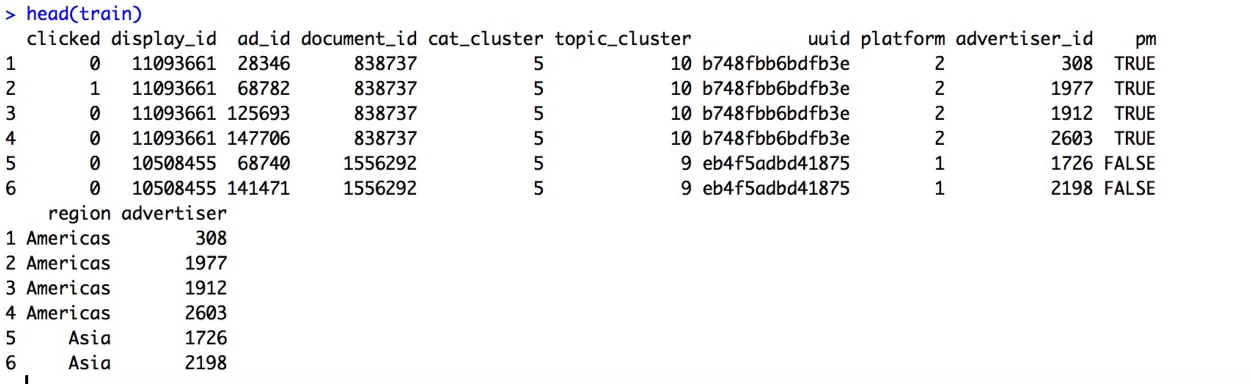


```
> head(train)
  clicked display_id  ad_id document_id cat_cluster topic_cluster           uuid geo_location platform timestamp
1       0   11093661  28346      838737           5            10 b748fbb6bdfb3e           US        2 740917567
2       1   11093661  68782      838737           5            10 b748fbb6bdfb3e           US        2 740917567
3       0   11093661 125693      838737           5            10 b748fbb6bdfb3e           US        2 740917567
4       0   11093661 147706      838737           5            10 b748fbb6bdfb3e           US        2 740917567
5       0   10508455  68740     1556292           5             9 eb4f5adbd41875           IN        1 705238294
6       0   10508455 141471     1556292           5             9 eb4f5adbd41875           IN        1 705238294
  advertiser_id    pm   region advertiser
1           308  TRUE Americas        308
2          1977  TRUE Americas       1977
3          1912  TRUE Americas       1912
4          2603  TRUE Americas       2603
5          1726 FALSE     Asia       1726
6          2198 FALSE     Asia       2198
```

Figure 5:

Variable Selection

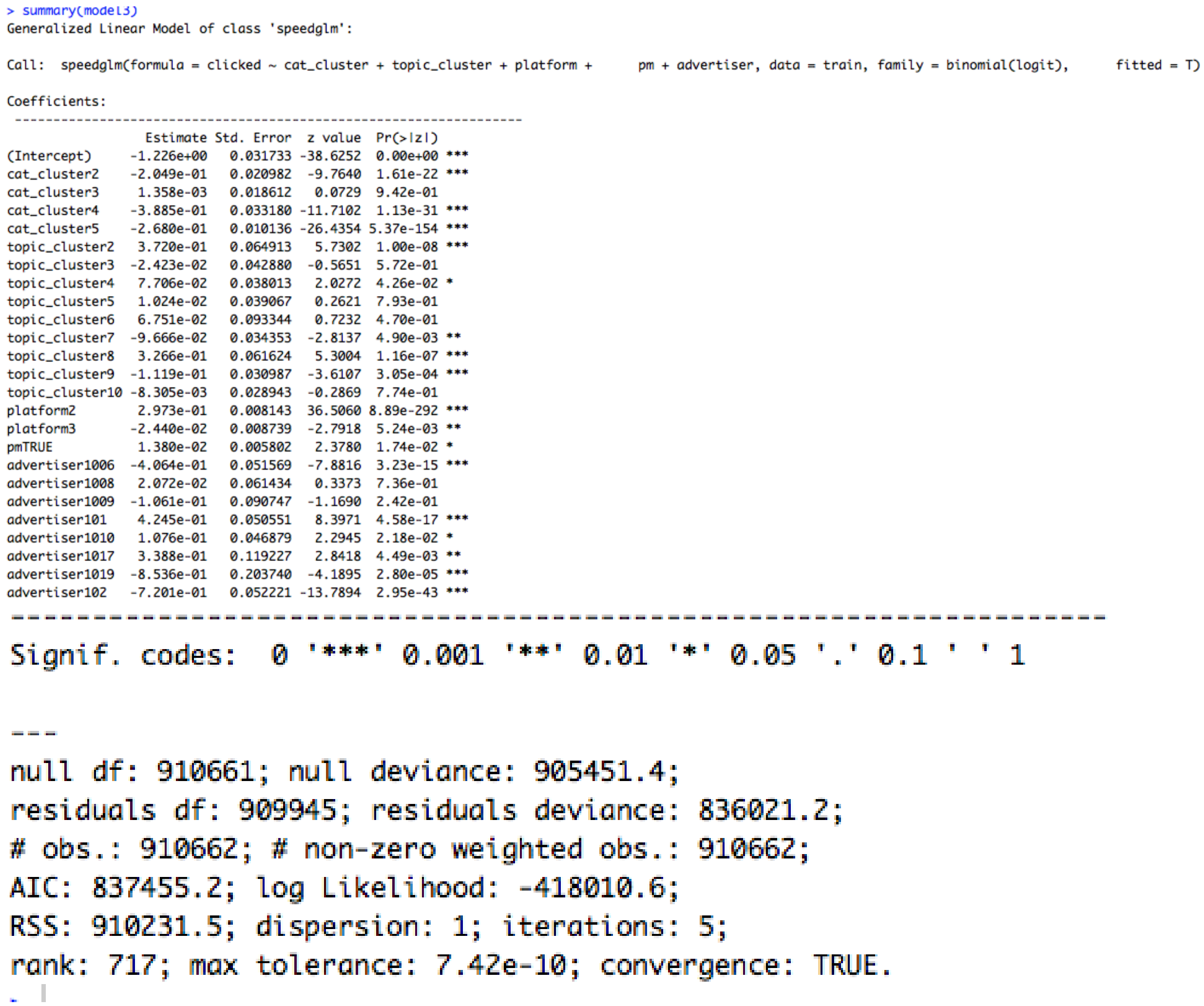
- After processing, there still are plenty of variables.
- Since regression includes matrix calculation, we still need to select necessary variables to reduce calculation time.
- Backward selection is applied to select variables.



```
> head(train)
  clicked display_id  ad_id document_id cat_cluster topic_cluster           uuid platform advertiser_id    pm
1       0   11093661  28346      838737           5            10 b748fbb6bdfb3e        2           308  TRUE
2       1   11093661  68782      838737           5            10 b748fbb6bdfb3e        2          1977  TRUE
3       0   11093661 125693      838737           5            10 b748fbb6bdfb3e        2          1912  TRUE
4       0   11093661 147706      838737           5            10 b748fbb6bdfb3e        2          2603  TRUE
5       0   10508455  68740     1556292           5             9 eb4f5adbd41875        1          1726 FALSE
6       0   10508455 141471     1556292           5             9 eb4f5adbd41875        1          2198 FALSE
    region advertiser
1 Americas        308
2 Americas       1977
3 Americas       1912
4 Americas       2603
5     Asia       1726
6     Asia       2198
```

Figure 6:

Result

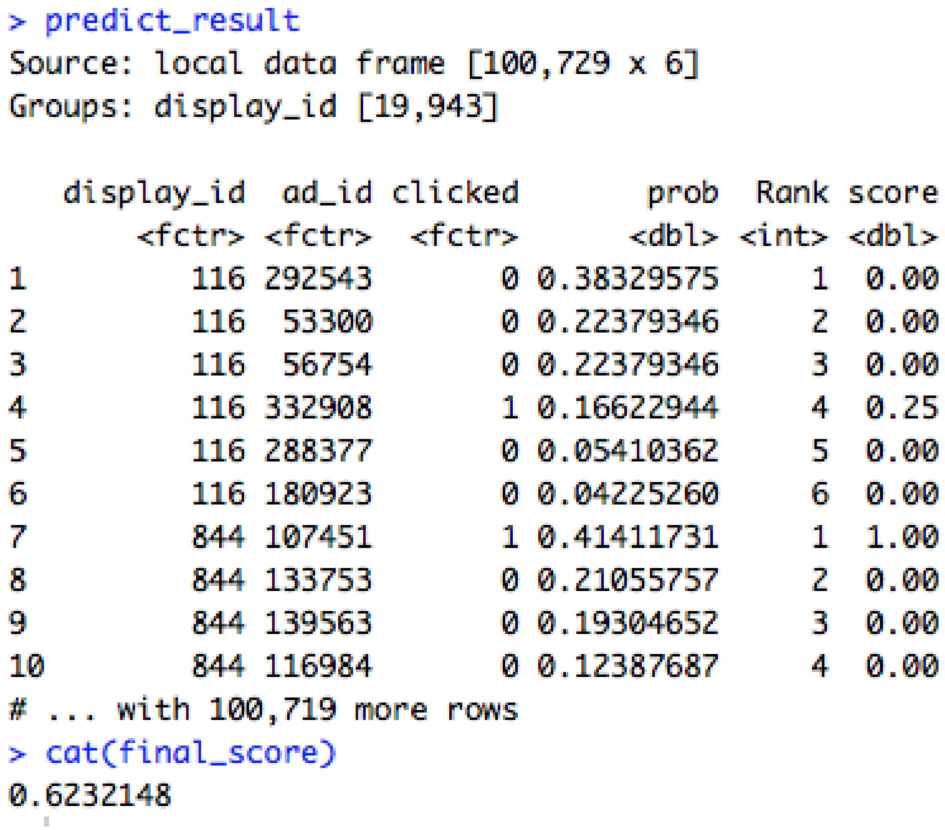


```
> summary(model3)
Generalized Linear Model of class 'speedglm':

Call:  speedglm(formula = clicked ~ cat_cluster + topic_cluster + platform +      pm + advertiser, data = train, family = binomial(logit),      fitted = T)

Coefficients:
 ------------------------------------------------------------------
                  Estimate Std. Error  z value  Pr(>|z|)
(Intercept)      -1.226e+00   0.031733 -38.6252  0.00e+00 ***
cat_cluster2     -2.049e-01   0.020982  -9.7640  1.61e-22 ***
cat_cluster3      1.358e-03   0.018612   0.0729  9.42e-01
cat_cluster4     -3.885e-01   0.033180 -11.7102  1.13e-31 ***
cat_cluster5     -2.680e-01   0.010136 -26.4354 5.37e-154 ***
topic_cluster2    3.720e-01   0.064913   5.7302  1.00e-08 ***
topic_cluster3   -2.423e-02   0.042880  -0.5651  5.72e-01
topic_cluster4    7.706e-02   0.038013   2.0272  4.26e-02 *
topic_cluster5    1.024e-02   0.039067   0.2621  7.93e-01
topic_cluster6    6.751e-02   0.093344   0.7232  4.70e-01
topic_cluster7   -9.666e-02   0.034353  -2.8137  4.90e-03 **
topic_cluster8    3.266e-01   0.061624   5.3004  1.16e-07 ***
topic_cluster9   -1.119e-01   0.030987  -3.6107  3.05e-04 ***
topic_cluster10  -8.305e-03   0.028943  -0.2869  7.74e-01
platform2         2.973e-01   0.008143  36.5060 8.89e-292 ***
platform3        -2.440e-02   0.008739  -2.7918  5.24e-03 **
pmTRUE            1.380e-02   0.005802   2.3780  1.74e-02 *
advertiser1006   -4.064e-01   0.051569  -7.8816  3.23e-15 ***
advertiser1008    2.072e-02   0.061434   0.3373  7.36e-01
advertiser1009   -1.061e-01   0.090747  -1.1690  2.42e-01
advertiser101     4.245e-01   0.050551   8.3971  4.58e-17 ***
advertiser1010    1.076e-01   0.046879   2.2945  2.18e-02 *
advertiser1017    3.388e-01   0.119227   2.8418  4.49e-03 **
advertiser1019   -8.536e-01   0.203740  -4.1895  2.80e-05 ***
advertiser102    -7.201e-01   0.052221 -13.7894  2.95e-43 ***
 -------------------------------------------------------------------
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

---
null df: 910661; null deviance: 905451.4;
residuals df: 909945; residuals deviance: 836021.2;
# obs.: 910662; # non-zero weighted obs.: 910662;
AIC: 837455.2; log Likelihood: -418010.6;
RSS: 910231.5; dispersion: 1; iterations: 5;
rank: 717; max tolerance: 7.42e-10; convergence: TRUE.
```



```
> predict_result
Source: local data frame [100,729 x 6]
Groups: display_id [19,943]

   display_id  ad_id clicked       prob  Rank score
       <fctr> <fctr>  <fctr>      <dbl> <int> <dbl>
1         116 292543       0 0.38329575     1  0.00
2         116  53300       0 0.22379346     2  0.00
3         116  56754       0 0.22379346     3  0.00
4         116 332908       1 0.16622944     4  0.25
5         116 288377       0 0.05410362     5  0.00
6         116 180923       0 0.04225260     6  0.00
7         844 107451       1 0.41411731     1  1.00
8         844 133753       0 0.21055757     2  0.00
9         844 139563       0 0.19304652     3  0.00
10        844 116984       0 0.12387687     4  0.00
# ... with 100,719 more rows
> cat(final_score)
0.6232148
```

(5)Summary

- We use two methods to predict which ad will be clicked, both of them give us ideal result.