2020

Online Advertisement Click-Through Prediction

Real-world Big Data Analytics using Spark Framework on Alibaba E-commerce Dataset

Group Five Guys:

Ruby He

Vanessa Liang

Ruizhi Ma

Haonan Wang

Minke Wang

Table of Contents

Executive Summary

The shape of the marketing world is changing dynamically with the advent of new trends and technology. With rapid digitalization and low cost, online marketing becomes critical for business to drive traffic, leads, and sales. Displaying advertisements on various platforms is becoming one of the most efficient ways to acquire customers. Taobao, as the largest online retail platform in the world, provides billions of online display ads positions for millions of advertisers every day.

The click-through rate (CTR) is an important indicator to measure the effectiveness of the advertising display. Knowing in advance the CTR of each advertisement is critical for both advertisement platforms and advertisers. It will help advertisement platforms better allocate traffic and earn more total revenue from advertises. It will also help advertisers acquire more customers. The goal of this analysis is to accurately predict CTR for ads display and identify key features related to conversion. Basic prediction methods like logistic regression and XGboost were used at first. Then, in order to capture the complex mapping relation among different features and increase prediction accuracy, deep learning was introduced. Deep Factorization machines (DFM) and Deep Interest Network (DIN) achieved an AUC of 0.78 respectively ability of classify users in click or not click.

Further research on important features of the model illuminated the importance of customer behaviors and customer demographics. Some particular customer groups are more likely to click into advertisements. Thus, we suggest advertisement platform target ads for specific customers to increase CTR. Also, with different brand popularity and ads attractiveness, advertisements may be particularly attractive to customers. Platforms should assign different weights to different advertisers, thus achieve higher CTR and higher revenue.

Introduction

Click-through rate prediction is critical in Internet advertising and affects web publishers' profits and advertiser's payment. How to advertise to specific user groups is a key issue in the field of online advertising. In this report, we will deploy and evaluate two basic classification models - logit model and XGboost - and two advanced deep learning classification models- Factorization Machines (FMs) and Deep Interest Network (DIN). Additionally, we will examine important features that affect Click-through rate.

Formulation of Analysis Framework

Business Objective

The goal is to accurately predict CTR for ads display and identify key features related to conversion. In this way, we could help advertisers bid for specific spots and target crowds to compete for business traffic and help the platform choose the most suitable ads to display.

Key Actionable Business Initiative

Advertisers and platforms are the main benefit of this analysis. For Advertisers/vendors, based on the predicted CTR within different customer groups, they could adjust the ads quality and placement strategy to target the right customers. In this way, the ads' CTR will increase and the advertiser will acquire more customers. What's more, a high click-through rate leads to high quality scores in Taobao, which further allow advertisers to improve or maintain ad position for lower costs. For ads platforms, before allocating and launching every advertisement, they will know how it will perform and how much revenue it will generate. Thus, the platform can adjust allocation plans for each advertisement by assigning different weights to different ads and

displaying different ads to different platform customers. A higher CTR will lead to more total revenue from advertisers.

Metrics of Success

CTR (click-through rate) is a metric that measures the number of clicks advertisers receive on their ads per number of impressions. On the business side, we would ideally see an increase of average CTR on Taobao platform after the implementation of our model. On the technical side, our prediction is based on each platform user and each advertisement level, so we are predicting whether a customer will click into one advertisement. This is a classification problem after all. Thus, we could use accuracy, recall, precision, AUC/ROC and logarithmic loss to measure the model success. After research, we found that in the CTR prediction field, AUC is a widely used metric (Fawcett 2006). It measures the goodness of order by ranking all the ads with predicted CTR, including intra-user and inter-user orders.

Role of Analytics

Inappropriate advertising can lead to a decline in user experience, undesired effect for advertisers and revenue loss for the media. Thus, we would like to utilize predictive analysis to let stakeholders know the potential ads performance and use prescriptive analysis to drive business insights and raise recommendations to help both the platform and advertisers increase their revenue.

Dataset Description and Exploratory Data Analysis

The dataset for the analysis is obtained through the Alibaba Group (Largest Online E-commerce Platform in China) website throughout the time period between 2017 May 6th and 2017 May 12th. The dataset contains over 1 million users randomly selected from the entire user pool and their demographic information, user click behavior, browsing, purchase patterns, information on the ads

they click on, etc. The user behavior log contains 26 million records of user browsing, star (add an item to favorite), cart (add an item to cart) and buy items behavior over the past three weeks prior to their ads click behavior. User demographic information including not only gender, age level, user age with the company, but also multiple designed metrics indicating their shopping level, consumption level, etc.

The target outcome for prediction in the dataset is the Click or Non-click behavior of the user. And the explanatory features for the prediction can be split into three categories: user demographic metrics, user behavior metrics, and the clicked item information. Below are the features in details:

User Demographic Metrics:

'Final_gender_code': User Gender, 1 is Male, 2 is Female

'Age_level': User Age Level from 1 to 6. (User real age is hidden due to anonymization.)

'occupation': User Occupation, 1 is College student, 0 not a college student.

'New_user_class_level_': How long the user has been with the company.

User Behavior Metrics:

'Buy': User item purchase behavior.

'Cart': User adds items to cart behavior.

'Fav': User adds items to favorite behavior.

'Pv': User browsing items behavior.

'Pvalue_level': User consumption level.

'Shopping_level': User shopping level.

Click information Metrics:

'Ad_cate_id': The category of the ads that the user clicks on.

'Price': The price of the item that the user clicks on.

'Hour': The hour of the click that happens.

The exploratory analysis on the dataset tends to ask the question of how users' demographic information and their behavior are generally distributed across different categories. We started with studying the user profile and discovered that with their demographic information, female users take up as much as three times the male users. Meanwhile, the age group of users is normally distributed. Most of the users belong to non-college student occupations and have medium items purchase and consumption level and high shopping level that indicates their usage frequency (Appendix I Fig1.).

Then we started to explore how their click-through rate fluctuates across different groups or different usage patterns. Alibaba has defined their users into three categories based on their shopping behavior and shopping frequency which are shallow user, medium user and deep user. Among these shallow users, new female users have higher click through rates on the ads displayed to them. Users with different age groups have similar click through however younger shallow users tend to have very high click through. These exploratory analyses provide a general idea of how the features correlate with outcome variables. The discovered findings align with our intuition because new users are more curious and less familiar with the platform, older users and female college users are more likely to be attracted by promotions and offers displayed on the advertisements (Appendix I Fig2.).

Model Selection, Evaluation, and Interpretations

Predictive classification methods are used here. Deep learning methods like Deep Factorization machines and Deep Neural Networks perform better than traditional methods like logistic regression and XGBoost.

Traditional Classification Models

Logistic regression (LR) is a widely used shallow model for CTR prediction tasks. To build predictive models with these predictor variables, a common solution is to convert them to a set of binary features (a.k.a. feature vector) via one-hot encoding. However, the model does not take interaction into consideration automatically. In our case where raw data have a complex mapping relationship, especially for meaningful data, it doesn't have strong prediction power. However, from the features' coefficient and significance level, we could still find that the shopping level, new user, gender, and education level are all significantly related to customer click behavior. (Appendix II) And we also implemented the logit model on PySpark using the Spark ML library.

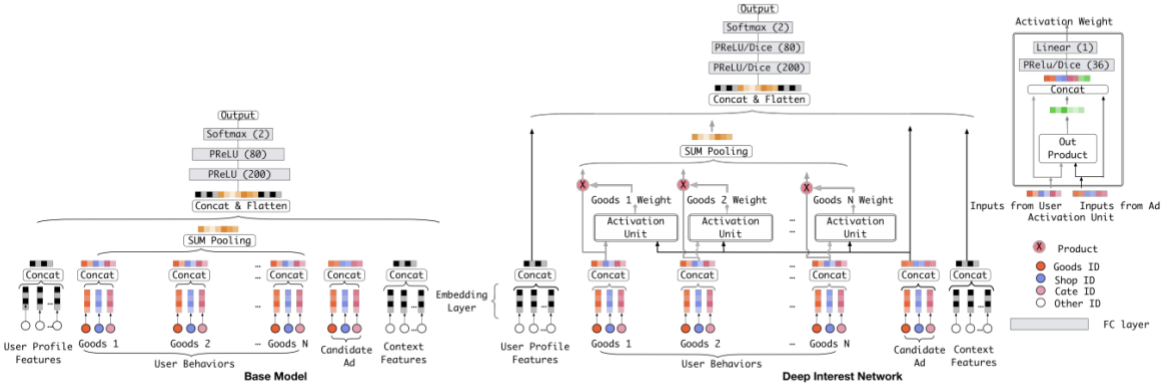
XGBoost is a widely used classification method based on gradient boosted decision trees designed for speed and performance. It is an interactive decision tree approach where the posterior tree is based on the training results of the prior tree. We trained the model using the XGBoost method with hyperparameter tuning and got a final accuracy higher than the logistic regression at 0.63.

Deep Learning Models

Since traditional CTR prediction models mainly depend on the design of features. The features of data are artificially selected and processed. However, our raw data have a complex mapping relationship, especially for meaningful data, and it is important to account for the interactions between features.

Therefore, we researched the methods in deep learning field, as a powerful approach to learning feature representation, deep neural networks have the potential to learn sophisticated feature interactions. The first model we used was Deep Factorization machines. Factorization machines [FM] is a supervised learning approach that embed features into a latent space and model the interactions between features by taking the inner product of their embedding vectors. And a Deep FM model combines the power of factorization machines for recommendation and deep learning for feature learning in a new neural network architecture.

The final deep learning model we used was Deep Interest Network (DIN). This model based on traditional Embedding & MLP models but added a local activation unit to adaptively learn the representation of user interests from historical behaviors with respect to a certain advertisement. The use of fixed-length representation in traditional deep CTR models is a bottleneck for capturing the diversity of user interests. To improve the expressive ability of the model, Deep Interest Network is designed to activate related user behaviors and obtain an adaptive representation vector for user interests which varies over different ads. Specifically, in our case, activation units are applied on the user behavior features, which performs as a weighted sum pooling to adaptively calculate user representation vector given a candidate advertisement.



Network Architecture [Zhou, G… & Gai, K. 2018]

Compared with commonly used logistic regression model, these deep learning methods can reduce a lot of feature engineering jobs and enhance the model capability greatly. We achieved a 0.78 AUC.

Implementation and Recommendation

Executing the Analytics

There are two major data sources of the analysis. The user behavior data is captured automatically by the company's internal data science tools, while the user profile and the ads feature data will be extracted from the company's databases. The marketing science team at the company is responsible for collecting and analyzing the data. The business analyst is responsible for deriving business insights from the analysis to help clients and the company improve ad effectiveness. The Data Analyst is responsible for conducting data cleaning and exploratory data analysis, and performing A/B tests with the new ad placement strategy. The data scientist is responsible for building ML models to predict ad click-through rates. The product designer is responsible for leveraging the insights from the analysis to optimize the ad placement of the platform and the user flow.

Implementing the Analytics

The results of the analysis can be implemented into an automatic system that proposes target audience and placement to clients for higher ad effectiveness.

Before proposing to the target audience, the company should first test the causality between all the customer behaviors/demographics we found significant in the model and CTR. We suggest Taobao conduct AB tests to establish the casual relationship here. Take the gender AB test for example, they should first split users into sample and test groups equally and randomly. Then determine the sample size and decide how significant the results need to be. Give the A/B test enough time to produce useful data. Use chi-square test to test if results are significantly different. Since multiple features should be tested, make sure only running one test at a time.

After guaranteeing the casualty, the recommendation system could be put into production. Through analysis of important features, we know some particular customer groups are more likely to click into ads, so we suggest ads platforms to collect thorough info about both customer behavior and customer demographic to better target customers. Also, since ads platforms charge advertisers with CTR but not number of exposures, we suggest ad platforms assigning different ad exposure to different categories or brands based on their category, brand, and popularity. Some categories/ brands may be particularly attractive to customers. By assigning higher weight to these types of ads, we could achieve higher CTR and higher revenue.

Scaling Up

Our analytics framework is recycled not only in Taobao, but also other advertisement platforms. We suggest the company realize model automation by converting static data/model into real-time data/model to achieve more accurate prediction with changes of data.

Challenges and Limitation

We assumed CTR is a two-side effect of both customers and advertisements. We have analyzed and raised feasible recommendations about the customer side. But still, for a company that

generates billions of user data every day, 15 is a small number. What's worse, there are only two features about advertisements - category and brand. Even the category feature is so specific that causes an overfit of modeling. We suggest Taobao improve data collection and database design. They should collect more customers, especially advertisement information, such as advertisement length, a more general category classification, and etc.

Another challenge is a tradeoff between CTR and user experience. Targeting only some customer groups heavily is not a long-term strategy. It will make these groups of customers quickly get tired of these advertisements or even the platform. Their CTR will be no longer high. One approach to solve the problem is to monitor users' satisfaction rate and to check how frequently these users report irrelevant ads (this is an embedded function in Taobao) on the platform.

Conclusions

In this report, with the examples of predicting CTR in the scenario of display advertising in the e-commerce industry. We demonstrated how our analytical system can be applied in advertising platforms. Such analytical systems could provide values beyond advertisement platforms like Taobao. They can also enlarge the business of advertisers. They can be improved and refined by ingesting more feature information. Customer satisfaction is a key factor to monitor in the process of implementation.

Reference:

Zhou, G., Zhu, X., Song, C., Fan, Y., Zhu, H., Ma, X., ... & Gai, K. (2018, July). Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1059-1068).

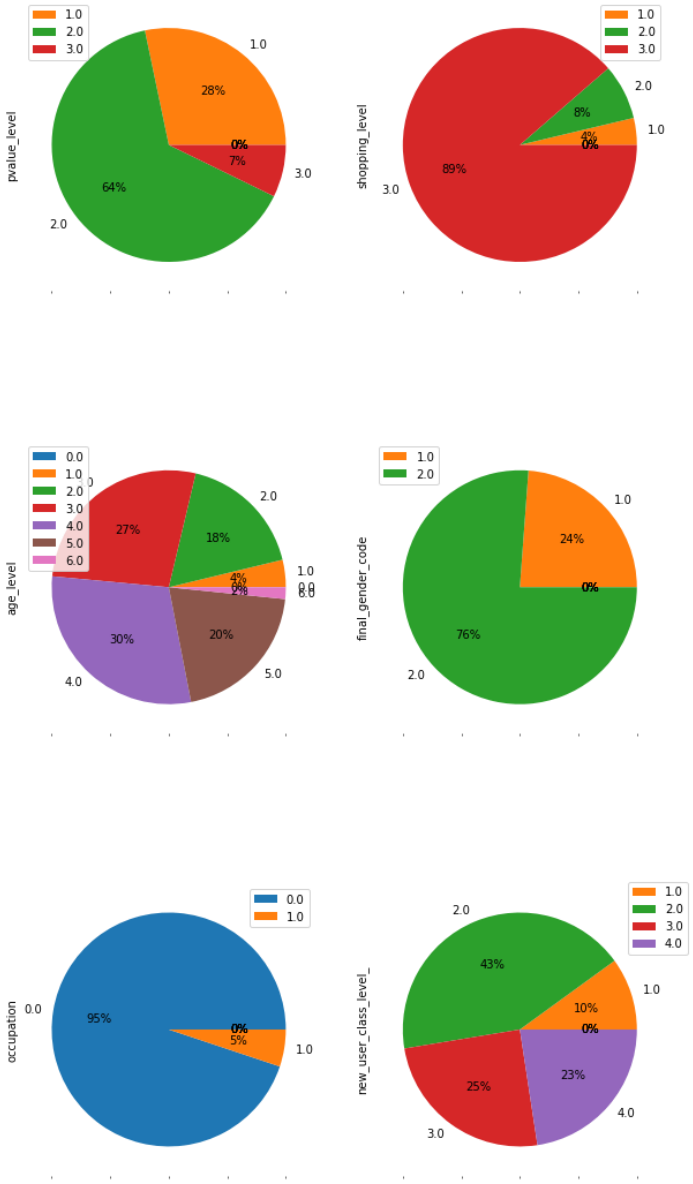
Trofimov, I., Kornetova, A., & Topinskiy, V. (2012, August). Using boosted trees for click-through rate prediction for sponsored search. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy* (pp. 1-6).

Tom Fawcett. 2006. An introduction to ROC analysis. Pattern recognition letters 27, 8 (2006), 861–874

Appendix

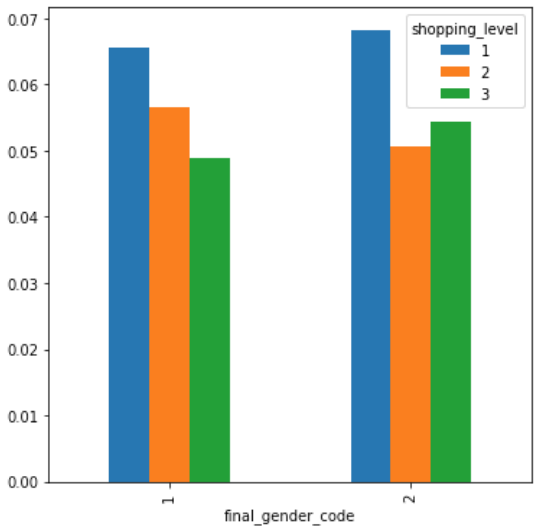
Appendix I: Exploratory Data Analysis

Appendix I Fig1. User demographics profile

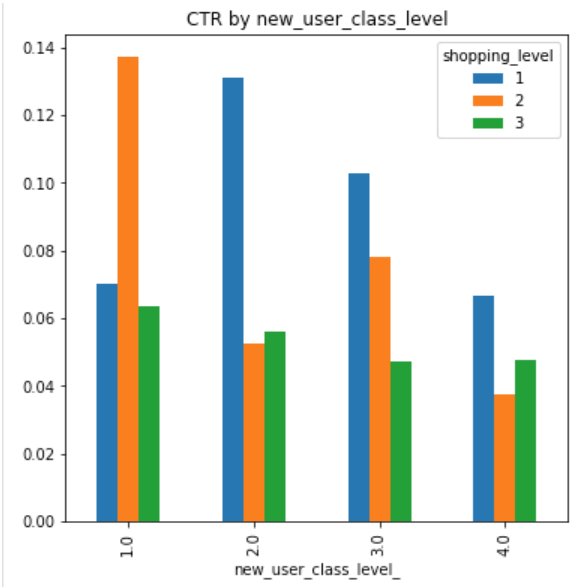


Appendix I Fig2. Click through rate by gender and shopping level.

(Gender =1 is male, Gender =2 is female)



Click through rate by new user level and shopping level. (Smaller level represents newer users.)



Appendix II: Modeling Results

```
Train on 1657564 samples, validate on 552522 samples
Epoch 1/5
1657564/1657564 [==============================] - 485s - loss: 0.1402 - binary_crossentropy: 0.1401 - val_loss: 0.6405 - val_binary_crossentropy: 0.6404
Epoch 2/5
1657564/1657564 [==============================] - 474s - loss: 0.0442 - binary_crossentropy: 0.0440 - val_loss: 0.6116 - val_binary_crossentropy: 0.6114
Epoch 3/5
1657564/1657564 [==============================] - 472s - loss: 0.0350 - binary_crossentropy: 0.0348 - val_loss: 0.5904 - val_binary_crossentropy: 0.5901
Epoch 4/5
1657564/1657564 [==============================] - 476s - loss: 0.0315 - binary_crossentropy: 0.0312 - val_loss: 0.5720 - val_binary_crossentropy: 0.5717
Epoch 5/5
1657564/1657564 [==============================] - 482s - loss: 0.0296 - binary_crossentropy: 0.0292 - val_loss: 0.5725 - val_binary_crossentropy: 0.5720
test LogLoss 0.57
test AUC 0.78
```

```
Optimization terminated successfully.
         Current function value: 0.688773
         Iterations 4
                         Results: Logit
==========================================================================
Model:                 Logit             Pseudo R-squared:  0.006
Dependent Variable:    0.0000            AIC:               4938507.8206
Date:                  2020-06-05 09:48  BIC:               4938664.9278
No. Observations:      3584986           Log-Likelihood:    -2.4692e+06
Df Model:              11                LL-Null:           -2.4849e+06
Df Residuals:          3584974           LLR p-value:       0.0000
Converged:             1.0000            Scale:             1.0000
No. Iterations:        4.0000
--------------------------------------------------------------------------
                        Coef.  Std.Err.      z      P>|z|   [0.025  0.975]
--------------------------------------------------------------------------
buy                    -0.0123   0.0004  -31.7677 0.0000 -0.0131 -0.0116
cart                   -0.0152   0.0001 -111.0782 0.0000 -0.0155 -0.0149
fav                    -0.0051   0.0001  -81.1946 0.0000 -0.0053 -0.0050
pv                      0.0006   0.0000  119.3287 0.0000  0.0006  0.0006
final_gender_code       0.0927   0.0022   42.1301 0.0000  0.0884  0.0971
age_level              -0.0250   0.0009  -27.9693 0.0000 -0.0268 -0.0233
pvalue_level            0.0412   0.0011   37.3579 0.0000  0.0391  0.0434
shopping_level         -0.0140   0.0018   -7.7854 0.0000 -0.0176 -0.0105
occupation             -0.0554   0.0051  -10.8912 0.0000 -0.0653 -0.0454
new_user_class_level_  -0.0460   0.0008  -54.4967 0.0000 -0.0476 -0.0443
price                   0.0000   0.0000    7.9014 0.0000  0.0000  0.0000
hour                    0.0014   0.0002    8.1986 0.0000  0.0011  0.0018
==========================================================================
```