POST-CLICK CONVERSION RATE PREDICTIVE MODEL ON E-COMMERCE RECOMMENDER SYSTEM

Yuhe Ding

Abstract

This report discusses about how different features influence customers' decision on their online purchase after click behavior. The dataset is gathered from real-world traffic log of the recommender system in e-commerce. Logistic Regression and Extreme Gradient Boosting are used as main machine learning approaches for predictive analysis and modeling. In this study, features from users' profile, shops' profile and context are tested to see to what extent they may exert influence on customers' purchase intention. Based on the experiment results, this paper also proposes some possible improvement for e-commerce platform in personalized recommendation in order to increase conversions and discusses about potential approaches to improve conversion rate prediction performance.

Introduction

As the Internet develop, the e-commerce industry has gradually entered a period of prosperity. People today rely much on shopping online, which brought a huge amount of

consumer behavioral data. Finding out what lead to their final consumption behavior is of great importance for improving total sales of e-commerce. Personalized search is a common way to better user experience and persuade user consumption. With personalized search, websites will provide tailor-made recommendations and rank the results based on their conversion rate prediction. Given recommended items when visiting e-commerce websites, users might click interested ones before making a further purchase, which follow a sequential pattern of impression → click → conversion. Along this long, the project experiment on a way of combining personalized search related data and choice model to predict users' purchase intention after their click behavior.

Methodology

The goal of experiment is to predict post-click conversion rate of advertised search result using a set of related features. A promising approach is to train predictors which learn how to predict post-click conversion rate based on factors that influence customers' online purchase.

Since there are known results and there are only two types of them, traded or not traded, the problem essentially can be solved with classification methods. In machine learning, predictive algorithm is called hypothesis. The dataset is gathered from real-word traffic log of the recommender system in Taobao which is the largest online retail platform in China. Training and test set are split along time sequence, which is a traditional industrial setting. There are 478138 clicked instances in the training set and 18371 in the test set, sharing 26 possible features in the both sets for analysis. And the only difference between these two sets is that there is one more label in the training set indicating trade status but not in the test set. In order to provide a clearer explanation and gain a better understanding, we divide these features into five tables about clicked samples, advertising items, users, context and shops separately. The algorithms we use in the experiment are logistic regression, XGBoost.

1) Logistic Regression:

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

Probability of a single sample is:

$$p(y|\mathbf{x}, \beta_0, \beta) = \frac{1}{1 + \exp\{-y(\beta_0 + \mathbf{x}^T\beta)\}}$$

Likelihood function is:

$$\mathcal{L}(\beta_0, \beta|\mathbf{y}, \mathbf{x}) = \prod_i \frac{1}{1 + \exp\{-y_i(\beta_0 + \mathbf{x}_i^T\beta)\}}$$

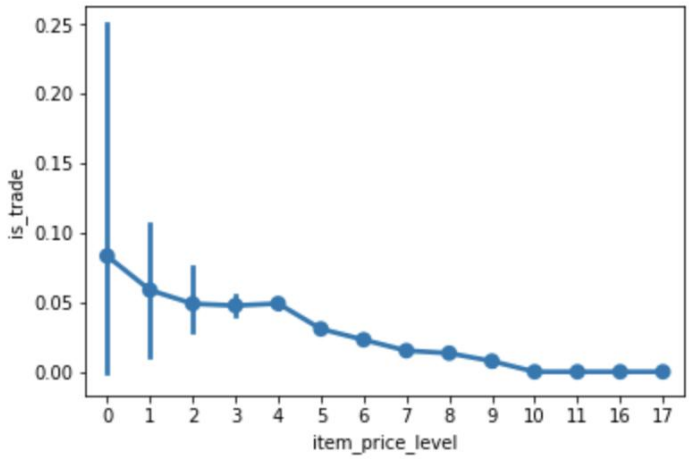
Log-likelihood function is:

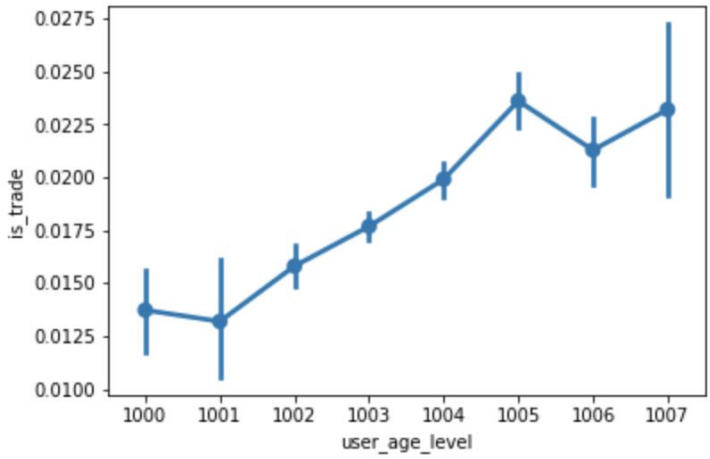
$$\log \mathcal{L}(\beta_0, \beta|\mathbf{y}, \mathbf{x}) = -\sum_i \log\{1 + \exp\{-y_i(\beta_0 + \mathbf{x}_i^T\beta)\}\}$$

2) XGBoost is an algorithm that has recently been dominating applied machine learning. There are many advantages of XGboost algorithm. For example, it can help to reduce overfitting based on its regularization, allows to define custom optimization objectives and evaluation metrics and etc. The most important one is that the trained XGBoost algorithm can also be used to encode numerical values for other machine learning algorithms, such as

Logistic Regression. Thus, in the study, experiment will be conducted to combine both of these models.

For some of continuous features, if they do not show clear relation with target label, it may be useful to divide them into sections in order to find out some connections. We can see that both item price and user age are clearly related to target value in figure 1 and figure 2 which indicates these two features could be strong predictors.

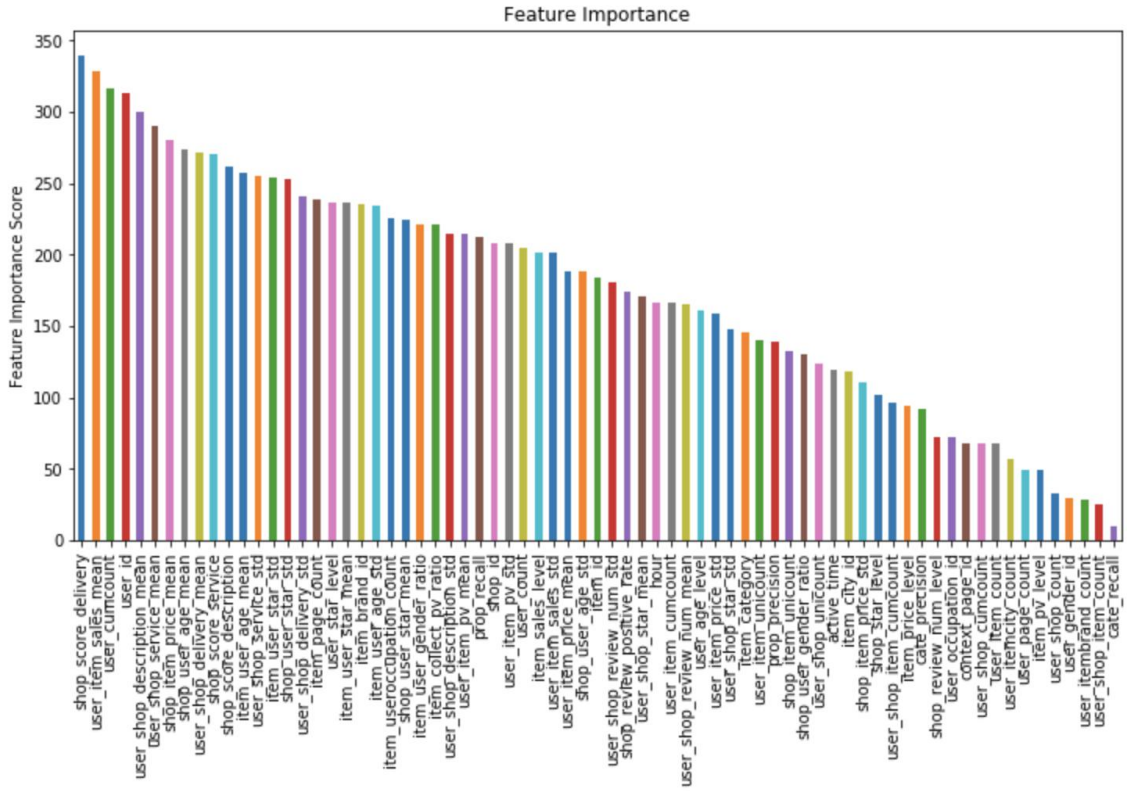




Results

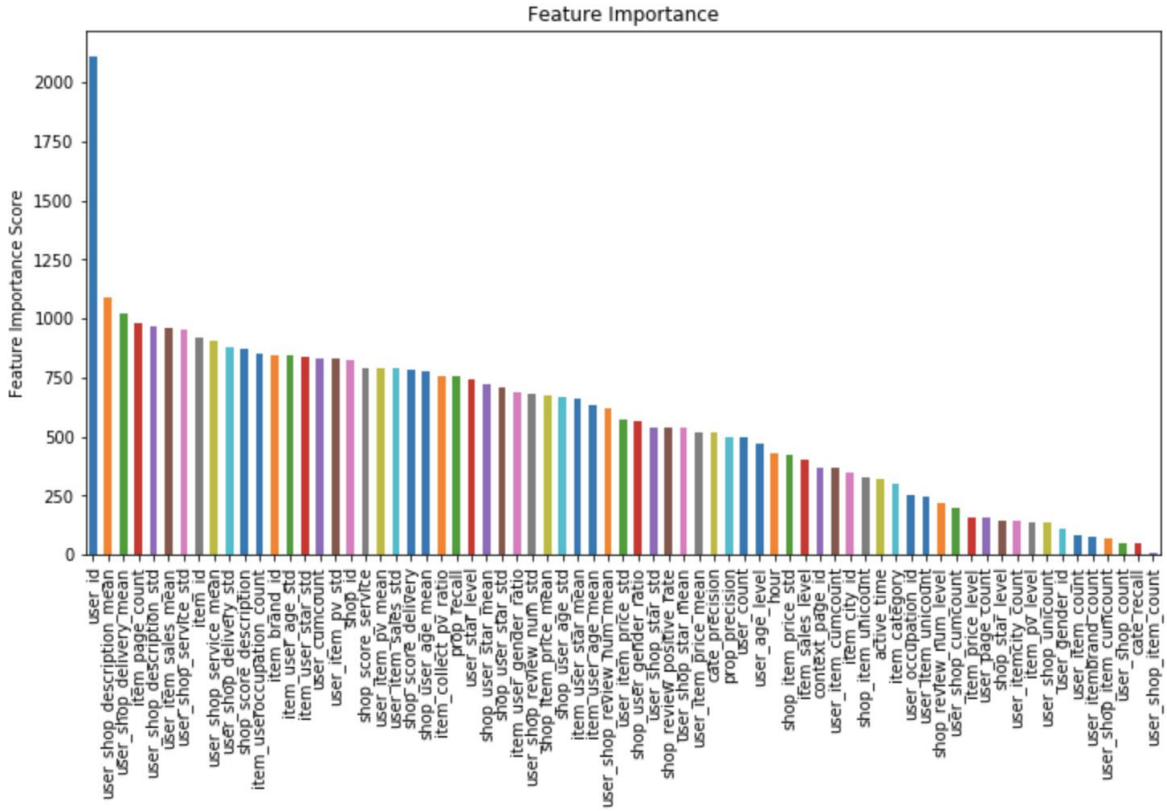
1) Logloss (training): 0.78390

Logloss (validation): 0.084535



2) Logloss (training): 0.072778

Logloss (validation): 0.77509


Feature Importance
Feature Importance Score
0
250
500
750
1000
1250
1500
1750
2000
user_id
user_shop_description_mean
user_shop_delivery_mean
item_page_count
user_shop_description_std
user_item_sales_mean
user_shop_service_std
item_id
user_shop_service_mean
user_shop_delivery_std
shop_score_description
item_useroccupation_count
item_brand_id
item_user_age_std
item_user_star_std
user_cumcount
user_item_pv_std
shop_id
shop_score_service
user_item_pv_mean
user_item_sales_std
shop_score_delivery
shop_user_age_mean
item_collect_pv_ratio
prop_recall
user_star_level
shop_user_star_mean
shop_user_star_std
item_user_gender_ratio
user_shop_review_num_std
shop_item_price_mean
shop_user_age_std
item_user_star_mean
item_user_age_mean
user_shop_review_num_mean
user_item_price_std
shop_user_gender_ratio
user_shop_star_std
shop_review_positive_rate
user_shop_star_mean
user_item_price_mean
cate_precision
prop_precision
user_count
user_age_level
shop_item_price_std
item_sales_level
context_page_id
user_item_cumcount
item_city_id
shop_item_unicount
active_time
item_category
user_occupation_id
user_item_unicount
shop_review_num_level
user_shop_cumcount
item_price_level
user_page_count
shop_star_level
user_itemcity_count
item_pv_level
user_shop_unicount
user_gender_id
user_item_count
user_itembrand_count
user_shop_item_cumcount
user_shop_count
cate_recall
user_shop_item_count