Winning the Second Place of IJCAI-15 Repeat Buyers Prediction Contest: a Feature Engineering Approach

Jianxun Lian†, Xing Xie‡, Guangzhong Sun†
†University of Science and Technology of China, Hefei, China
‡Microsoft Research, Beijing, China
jianxun.lian@outlook.com, xingx@microsoft.com, gzsun@ustc.edu.cn

Abstract

This paper describes the detailed method of team LeavingSeason's approach to IJCAI-15 competition. The task of the competition is to identify which new buyers for given merchants will become loyal customers in the future. Merchants sometimes run big promotions on particular dates in order to attract new buyers. It is important for merchants to identify who can be converted into repeated buyers. In our approach, we split the raw information into user, merchant and user-merchant pillars. From each pillar we extracted features based on the analysis of its exclusive properties. We further refined the framework by selectively combining the features and ensembled different models as well. Our approach achieved second place in the second stage.

1 Introduction

Merchants sometimes run big promotions(e.g., discounts or cash coupons) on particular dates (e.g., Boxing-day Sales, Black Friday or Double 11). While it could attract a large number of new buyers, many of them are one-time deal hunters. The task of IJCAI-15 competition[1] was to identify who could be converted into repeated buyers. By targeting on these potential loyal customers, merchants can greatly reduce the promotion cost and enhance the return on investment(ROI).

1.1 Problem Definition

In this challenge, participants were provided with a set of merchants and their corresponding new buyers acquired during the promotion on the "Double 11" day. The task was to predict the probability of the new buyers for given merchants to purchase from the same merchants again within six months.

1.2 Dataset

The competition consisted of two stages. In the first stage, the data set contained anomymized shopping logs from around 200k users and 5k merchants in the past 6 months before and on the "Double 11" day . Each log was an action record containing user id, merchant id, user age, user gender, item id, category id, brand id, time stamp, and action type. There were four kinds of activity in the logs: click, add-to-cart, purchase and add-to-favourite. In the second stage, the size of the data set was more than five times larger than that of stage 1.

In this paper, we describe the detailed information of our approach, including how we designed features from raw data and how we combined different models in the first stage. Our approach achieved fifteenth in the first stage and then second place in the second stage.

2 Feature Engineering

The task of this competition was to predict whether a new buyer would become a repeat buyer in the future for a given merchant, and each instance of training and test set was a {user_id, merchant_id} dyad. We assumed that both the user' preference and the merchant's trait would have a marked impact on the probability to become a repeat buyer. So we designed features from three pillars: user level, merchant level, and user-merchant level.

2.1 User Properties

User features contained the properties of the user's overall shopping behavior. Over the 6 months in the shopping logs, a user might have viewed and purchased different items from multiple merchants. We tracked his behavior properties from all available action types, i.e. click, add-to-cart, purchase and add-to-favourite. For each action type we extracted the overall statistics and some temporal analysis. Most of the feature extracting were trivial, here we only list some of the representative ones.

Overall Statistics First of all we assumed the simplest aggregation value of the user's past shopping behavior could reflect his/her properties. For example, the more items the user has bought, the higher probability that he/she is a shopaholic and therefore more likely to purchase in the future. We counted the total number of items, categories, brands, and merchants under each of his 4 different action types.

[1] http://ijcai-15.org/index.php/repeat-buyers-prediction-competition

Lifespan User's lifespan is the number of days since the user's first action date in the shopping logs till the end date. A longer lifespan might indicate the user enjoy shopping online.

Buy-to-Click Ratio The total number of purchases divided by the total number of clicks. We further refined this feature into category, brand, merchant and item level buy-to-click ratio. E.g., item level buy-to-click ratio was the total number of purchased items divided by the total number of clicked items from this user.

Temporal Behavior Besides overall aggregation values, we extracted some features considering the temporal effect. The last date, i.e. the "Double 11" day, is very different from the other days because of the big promotion. For each user we calculated his last day purchase lift as (1). A higher lift value indicates the user is more easy to be attracted by sales promotion.

$$lift_{festival} = \frac{number\ of\ last\ day\ purchase}{number\ of\ purchase\ from\ the\ other\ days} \tag{1}$$

During the several months' shopping logs, if the user's behavior pattern changed, we assumed that user's recent activity reflected his future activity better. So next we re-calculated most of the above features based on user's last week and last 3 months' shopping logs respectively as another features group.

Repeat Behavior When a user purchased a same item at two or more different dates, we recognized this item as a repeat purchased item from this user. Some users are loyal to some items and prefer to buy multiple items from the same merchant, or of the same brand, while some users like trying new things and consuming from different merchants. The latter is known as "novelty-lover"[Zhang *et al.*, 2014]. We calculated the repeat behavior ratio(RBR) indicator for each user as (2).

$$RBR_{item\_purchase} = \frac{number\ of\ repeat\ purchased\ items}{number\ of\ total\ purchased\ items} \tag{2}$$

We modeled repeat behavior into features making an assumption that the behavior properties from the same user would be consistent over the time. So if a user was a loyal consumer in the past, he/she would also be loyal in the future. the *item* and *purchase* in (2) could be replaced by *merchant/category/brand* and *click/add-to-cart/add-to-favourite*.

Activity Entropy Behavior entropy(BE) is defined as (3). Entropy describes the amount of variation within a user's activity, therefore provides another means to gauge consuming novelty. We calculated the item, category, brand and merchant entropy on click and purchase action separately.

$$BE = -\sum_{i=1}^{n} p(x_i) log\, p(x_i) \tag{3}$$

$$p(x_i) = \frac{number\ of\ actions\ on\ x_i}{number\ of\ total\ actions} \tag{4}$$

Demography Including user's gender and age range.

2.2 Merchant Properties

Merchant features aim to portray the merchant's general sales status. We extracted these features based on all the shopping logs belong to the merchant.

Overall Statistics Some overall aggregation values such as how many customers have purchased from the merchant, how many items/categories/brands have been sold/clicked/added-to-favourite, and the merchant's sales rank. These features indicate the popularity of the merchant.

Lifespan Similar to the definition of user lifespan.

Buy-to-Click Ratio Similar to the definition of user's buy-to-click ratio. A higher ratio means a better transition from click to purchase. It's a useful means to gauge how the merchant could attract the users. We refined this feature into category, brand, user and item level buy-to-click ratio.

Temporal Behavior We considered each merchant's overall aggregation, last 3 months' aggregation, and last week's aggregation. We also precessed "Double 11 Day" specially, and calculated the sales lift on this key data comparing with the ordinary days.

Promotion Frequency The most important factor attracting more customers was whether there was a sales promotion. So the more sales campaigns a merchant had, the more chance it could retain the customers. We assumed each promotional campaign would lead to a sales spike in the merchant's daily sales curve. Then we measure the number of spikes as a feature indicating the merchant's sales campaign plan.

Repeat Behavior Similar to user's repeat behavior analysis, we assumed that a merchant's popularity and credit would be consistent over several months' time range. So we calculated each merchant's repeat user ratio. If a merchant had a very low repeat user ratio in the past logs, it might be unlikely that it could retain the new buyers in the future.

One tricky thing deserved to be mentioned is that we also calculated repeat user ratio exclusive of the "Double 11 Day", which turned out to be one of the key features. The reason is straightforward: the test time range covers 6 months after the "Double 11". During this period most of the merchant would not carry on big sales promotions as they did on "Double 11 Day".

2.3 User-Merchant Properties

Besides modeling a user or a merchant's overall properties, we also extracted features based on a user's activity logs from a specific merchant. These user-merchant properties reflects the user's specific taste on a merchant.

Overall Statistics, Lifespan, Buy-to-Click Ratio, Temporal Behavior Similar to the analysis of user pillar and merchant pillar.

Remaining Items The number of item has been clicked or added to cart or added to favourite but not purchased yet. We assumed that some users would like to finish their investigation and purchase the items already in their shopping list.

Merchant Rank The current merchant's rank among all the merchants visited by the current user. We ranked the merchants by their sales volume.

3 Modeling Techniques

3.1 Data Analysis

In data pre-processing step, we compared the training set with test set, and found that while the training set and test set had different user id set, they shared the same merchant id set. Since we knew the true label in the training set, we were able to calculate the real repeat user ratio for each merchant based on training labels. We named this feature with *post repeat user ratio* (PRUR) and after we ran a 5-fold cross validation experiment we got an AUC of above 0.73. However by doing this we overfitted the training set, the score in test set was poor. So we tried to generalize this feature. There are totally 1995 merchants in the training set, and about half of them has less than 50 new buyers. We assumed that the PRUR for small merchants were volatile and were very sensitive to the random split process. Here we set a threshold and categorized those merchants with new buyer number less than this threshold as small merchants. For small merchants, we set their PRUR to a constant, i.e. -1. By trying different thresholds, we found 50 to be the best choice for it.

Besides PRUR, we calculated the post repeat probability of each item, category and brand as well. The process were just slightly different from PRUR's.

3.2 Framework

Figure 1 shows our framework. As described in Section 2, we extracted features from three pillars: user level, merchant level, and user-merchant level. In contest stage 1 this step was done with Microsoft's big data platform named COSMOS using language SCOPE[Chaiken *et al.*, 2008]. Next we joined the features and got the complete information for each user-merchant instance. In step 3 we tried different kinds of binary classifiers. While gradient descent decision tree (GBDT) was the best single classifier among them, a simple combination of different classifiers' output could lead to better performance.

3.3 Classifier Selection

We tried 6 different classifiers and used grid searching to find the optimal parameters. All the experiments were done with Microsoft's machine learning tool TLC. The results are shown in Table 1. We investigated a few solutions to the past data mining contest, and found that most of the solutions used a ensemble of different models as the final approach[Niculescu-Mizil *et al.*, 2009]. In our experiment we used a simple linear ensemble of GBDT, Logistic Regression and Neural Network with weights 0.82, 0.09, 0.09 respectively.

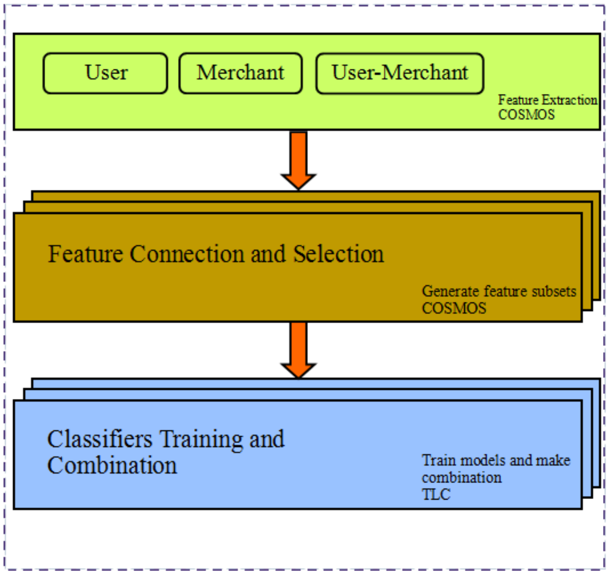
User
Merchant
User-Merchant
Feature Extraction
COSMOS
Feature Connection and Selection
Generate feature subsets
COSMOS
Classifiers Training and Combination
Train models and make combination
TLC

Figure 1: Framework

Classifier	AUC
GBDT	**0.6956**
LR	0.6839
SVM	0.620
Random Forest	0.678
Neural Network	0.6819
Averaged Perceptron	0.6716

Table 1: AUC of different classifiers

3.4 Feature Analysis

Figure 2 shows the top 10 key features learned from GBDT. feature name starting with "UM" means it comes from User-Merchant pillar. *post repeat user ratio* (PRUR) was the most important feature, and it verifies our assumption that a merchant's properties would be consistent between training set and test set. How many items the user purchased at "Double 11" day from this merchant was the second key feature. And the feature *URepeatMerRatio* verifies our assumption that a user's behavior would be consistent over the months. If he liked to repeatedly shop from his visited merchants in the past, he would also likely to repeatedly shop from his visited merchants in the future. PRUR feature also inspired us to consider the problem of overfitting. To fight against overfitting, we tried selecting a subset from the features and trained a model based on the feature subset. Then we combined these models' prediction as the final output. The way we selected feature subset including:

Pillar Selection We assumed that features come from different pillars were orthometric. So we trained a model based on each possible combination of different pillar, there were totally 7 ($2^3 - 1$) models.

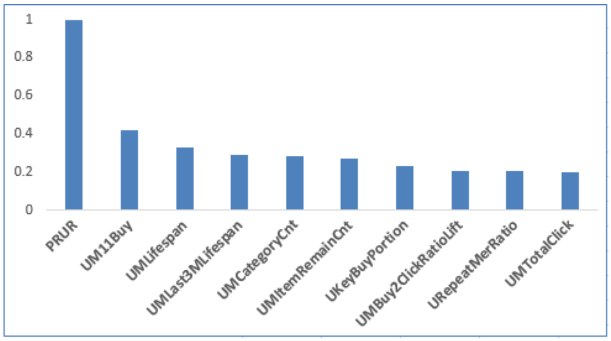

Figure 2: Top 10 most important features from GBDT

Depth of tree	5
Number of tree	1800
Learning rate	0.025
Min Leaf Number	32

Table 2: GBDT parameters at stage 2

PRUR Selection *post repeat user ratio* was the biggest overfitting factor. We trained one model inclusive PRUR and trained another one exclusive PRUR, then combined their prediction.By doing this we achieved our final stage 1 score of 0.701789.

3.5 Stage 2 Process

In stage 2 the participants need to work with a new big data platform. The feature engineering work was almost the same as stage 1, while we didn't make model ensemble any more. We use GBDT as the sole classifier with parameters in Table 2.

4 Results

Figure 3 shows our AUC improvement over the first stage. We finally achieved fifteenth at this stage.

In stage 2, we mainly focused on tuning the number of trees in GBDT. Figure 4 shows the AUC change with different GBDT parameters. Our last submission was with NumOfTree=2000, but it took more than 5 hours to run so was killed by the system. Finally we achieved the second place.

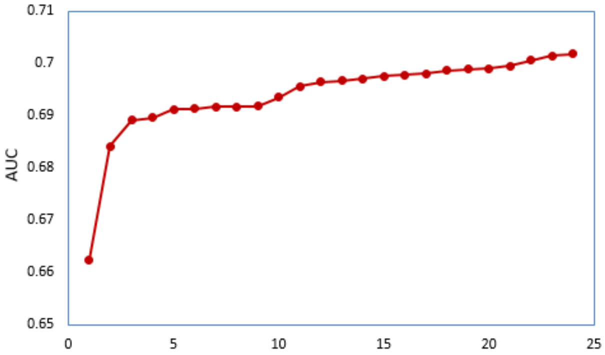

Figure 3: AUC improvement history in stage 1

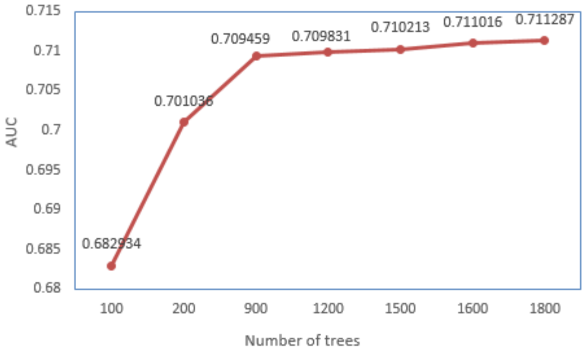

Figure 4: AUC improvement with the number of trees in GBDT

5 Discussions

During the competition we kept designing new features and refining models and then made progress. But not all the new features we ever tried could improve the AUC. One feature we tried but not helpful came from recommendation model. We used collaborative filtering method to predict which item/category/brand the user might purchase in the future. Then we count how many items/categories/brands were included in the merchant. One possible explanation is that our task is about repeat buyer prediction, not a traditional recommendation problem.

Another feature we have tried is the post repeat buyer distribution for each merchant. As mentioned before, training data and test data share a same merchant id set. So from training data we could know the ratio of repeat buyers for each merchant. Next we designed algorithms to make the model generate prediction with the distribution that only a specific portion of users are predicted as repeat buyers. We didn't manage to improve the AUC by doing this, however we still believe this method is quite promising.

In the future, the performance could be further improved by enriching the data set. E.g, the sponsor could provide users' reviews to the merchants or items. By conducting sentiment analysis we could know users' preference better.

References

[Chaiken *et al.*, 2008] Ronnie Chaiken, Bob Jenkins, Per-Åke Larson, et al. Scope: easy and efficient parallel processing of massive data sets. *Proceedings of the VLDB Endowment*, 1(2):1265–1276, 2008.

[Niculescu-Mizil *et al.*, 2009] Alexandru Niculescu-Mizil, Claudia Perlich, Grzegorz Swirszcz, et al. Winning the kdd cup orange challenge with ensemble selection. *The 2009 Knowledge Discovery in Data Competition (KDD Cup 2009) Challenges in Machine Learning, Volume 3*, page 21, 2009.

[Zhang *et al.*, 2014] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, and Xing Xie. Mining novelty-seeking trait across heterogeneous domains. In *Proceedings of the 23rd international conference on World wide web*, pages 373–384. ACM, 2014.