Repeat Buyer Prediction

-- CSCE 822 Project Proposal

Xiaoyi Liu, Jianhai Su
Department of Computer Science and Engineering
University of South Carolina

Abstract

Repeat buyer prediction is crucial for e-commerce companies to enhance their customer services and product sales. In particular, being aware of which factors or rules drives repeat purchases is as significant as knowing the outcomes of predictions in the business field. Thus, the effective deep learning techniques are used to analyze the user-behavior log in order to predict the probability that these new buyers would purchase items from the same merchants again within 6 months. In this project, different classification algorithms, such as convolutional neural network-based deep learning, support vector machine and decision tree are applied to predict.

1.Introduction

Nowadays, merchants sometimes roll out big promotions (e.g., discounts or cash coupons) on particular dates (e.g., "Black Friday" or "Double 11 (Nov 11th)", in order to attract a large number of new customers. But many of the attracted buyers are usually just one-time deal hunters. Therefore, these promotions may have little long-lasting impact on sales. To alleviate this problem, it is important for merchants to identify who can be converted into repeat buyers. Repeat buyers of a merchant are a group of customers who will regularly buy commodities from the merchant. The identification of such a special group of customers is important to the merchant. Because the successful prediction does not only increase sales, but also allow the merchant to have a good control of promotion cost. By targeting on these potential loyal customers, merchants can maximize long-term return on investment while keeping the promotion cost under control.

But it is well known that in the field of online advertising, customer targeting is extremely challenging, especially for fresh buyers. Because an effective customer targeting, bringing customers with deals that they exactly want at the time point, ideally requires a full understanding of customers' needs, personality, lifestyle, interests and so on. These kinds of customer information are difficult to obtain because of either privacy issues or measurement difficulty. While with the advent of deep learning approaches and the advancement in machine learning field, we may be able to tackle this problem by mining the long-term user behavior log accumulated by online merchant platform, like Amazon.com and Tmall.com. With the accumulated user-behavior log in the platform, a model could be built to predict the probability of if a user will become a converter for a specific a merchant in the platform. With the predicted probability information, merchants could optimize their promotion plans and release good deals to those potential converters.

In this project, we plan to work on this repeat buyer prediction problem by using a dataset provided by Tmall.com. The dataset is a set of activity events that involves merchants and their corresponding new buyers. It was

acquired during the promotion on the "Double 11" day. Our goal is to build a model to predict the probability that these new buyers would purchase items from the same merchants again within 6 months. We plan to first try DNN, SVM and Decision Tree to build classifiers to identify if a user is converter and compare their testing performances. The next step will be exploring the dataset using ensemble techniques if none of the three models give a good performance. Finally, we might try to inspect the dataset as time-series one if time is permitted Since timing is important for a customer to make a decision to purchase an item.

2.Dataset

The dataset [1] is opened to the public online by Tmall.com. Both the training and testing sets contain around 200k users. Each record in the user activity logs contain 6 attributes, while each user profile contains 3 attributes, which are listed in Table 1 and Table 2.

Table 1 User Behavior Logs

Data Fields	Definition
user_id	A unique id for the shopper.
item_id	A unique id for the item.
cat_id	A unique id for the category that the item belongs to.
merchant_id	A unique id for the merchant.
brand_id	A unique id for the brand of the item.
time_tamp	Date the action took place (format: mmdd)
action_type	it is an enumerated type {0, 1, 2, 3}, where 0 is for click, 1 is for add-to-cart, 2 is for purchase and 3 is for add-to-favourite.

Table 2 User Profile

Data Fields	Definition
user_id	A unique id for the shopper.
age_range	User' s age range: 1 for <18; 2 for [18,24]; 3 for [25,29]; 4 for [30,34]; 5 for [35,39]; 6 for [40,49]; 7 and 8 for >= 50;0 and NULL for unknown.
gender	User' s gender: 0 for female, 1 for male, 2 and NULL for unknown.

3. Approach

3.1 Data Cleaning

The original log needs to be pre-processed and then fed into the feature engineering component. A python script will be written to handle it.

3.2 Feature Engineering

The existing number of attributes for a user is 9, which is smaller. We plan to try it first and then (if necessary) dig into the dataset to discover more features for a user, like users' potential seasonal buying interest that could be inferred based on information of brand, category and time.

3.3 Build Model

We will start by building single models based on the three algorithms, DNN, SVM and decision tree, and compare their testing performances. Then we would move on to apply ensemble techniques (bagging and boosting) with these built single models to improve the prediction accuracy. The purchasing behavior is impacted by many factors. One of them is timing. A student may be interested in buying a new laptop before the semester begins, while she will instead prefer to buy vacation related service or goods, like a ticket to a theme park. With that said, it

might be helpful to massage the activity log and inspect it as time-series dataset.

4. Evaluation

Cross-validation will be applied during training phase. The best performer will be picked for testing with the provide testing dataset. The performance metric is the classification accuracy. The test performance will be reported in three groups, which are over the entire user set, sets of same gender, sets of same age ranges.

Reference

[1] Dataset of Repeat Buyers Prediction Challenge.
https://tianchi.aliyun.com/competition/entrance/231576/information