Predicting Repeated Buyers: Midterm Report

Ishmail Grady, isg9
Hao Rong, hr335
Frank Zhang, fz252

October 2017

1 Purpose

In a world where amount of revenue generated by retailers is steadily becoming more focused on online sales, improving online marketing strategies can give a retailer a huge competitive advantage. Understanding customer retention can be especially useful for large online retailers that sell products from many different merchants. Two major uses that this project can have is insight on which customers should be sent promotions as well as a metric to compare customer retention across merchants. For instance, if we are able to predict if a customer will be a repeated buyer for a given merchant, then advertising products from that merchant to the customer can result in more sales in the future. Conversely, if a merchant is not retaining customers well, this model would be able to identify the reasons why. Our project aims to predict the retention of new customers which can be very useful information to improve the online marketing strategy for our company. Specifically, our goal is to predict if a new customer of a given merchant will be a repeated buyer of that merchant.

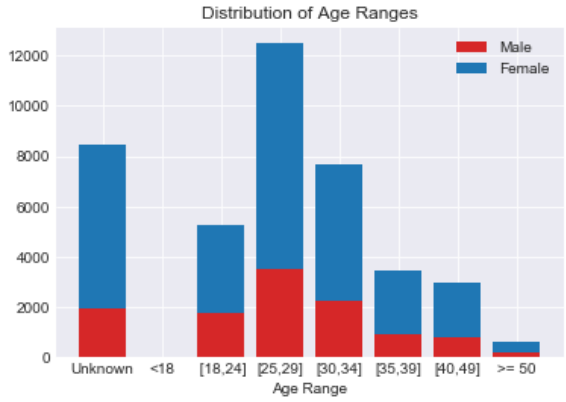
2 Data Exploration

2.1 Description

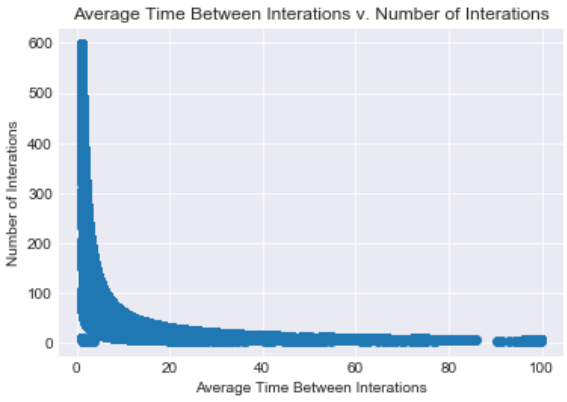
The data set used in this project is from Tmall.com, a Chinese online retailer. The data set is a collection of information about users who used the site in the 6 months that led up to a promotion including the day of. The data can be divided into two categories: user behaviour logs and user profile information. The user behavior logs include provide record of merchant each interaction a user has with the site, with the following information: item ID, category ID (of item), time of action, brand ID, and action type (add-to-cart, add-to-favorite, and purchase). The user profile information contains the gender and age of each user.

2.2 Data Cleaning

The raw data set contains 260,864 users/observations including users who interacted with all 4995 merchants on the platform. However, the amount of unique users across all of the merchants on the platform follow a power-law like distribution, thus many of the merchants included in the data set have very few users to form predictions with. For the purposes of our project we will restrict our data set to users that interacted with merchants that have at least 1000 unique users. Our final data include 44834 users divided among 29 different merchants.



(a) Figure 1



(b) Figure 2

Figure 1

2.3 Summary Statistics

Whenever trying to predict outcomes related to human behavior, demographic information can be very useful. As shown in Figure 1, the mode of the age ranges is [25,29]. Users in this age range make up about 29% of the repeated buyers in the data set, which is a disproportionate amount. From this figure we can also see that female users represent the majority of the users and the repeated buyers, about 70%. this indicates that these gender and age information could help predict our outcome.

The average amount of interactions per user is about 18 and has a median of 6. There is a relatively large amount of users with low amounts of interactions, but the data is skewed by users who have very large amount of interactions. Action type 1, clicking on an item, is make up most of these interactions, about 89%.

Another interesting relationship to investigate in this problem is the relationship between the average in between interactions and the total amount of interactions to examine how frequency affects user interaction. Figure 2 shows that these two variables have an approximately exponential relationship. Thus, according to our data, users that interact with the site a high rates are very unlikely to use it very frequently. this is important to not because one would expect that the more times a user interacts with a site, the higher the chance of them purchasing an item.

3 Baseline Model and Feature Selection

In order to get a baseline prediction for our final problem, which is predicting a user will be a repeated buyer for a particular merchant, we first tackle the problem of predicting whether a user will be a repeated buyer on the Tmall.com platform. The final product of this project will be 29 models for each merchant. This baseline model will give us a means to compare the performance of our final models. The features used in the baseline model are as follows:

Feature	Notation
Age_range	Age range of the user
Gender	Gender of the user
Action_0	Times the user clicked any item
Action_1	Times the user added any item to the chart
Action_2	Times the user purchased any item
Action_3	Times the user added any item as favorite
avg_time	The average time range between two actions
num_item	Number of items the user acted upon
num_cat	Number of categories the user acted upon
num_brand	Number of brands the user acted upon

We considered the four different actions separately because presumably different actions contribute differently to buyer's purchase behavior and thus have different influence in determining whether a buyer will be a repeated buyer.

We calculated a new feature "avg_time" based on time stamps of user actions, indicating how frequently a user acting through items. Furthermore, we grouped all the items, categories and brands a user acted upon.

Since we are looking for a general model as baseline to evaluate our more detailed model for each merchant, we implemented the Perceptron algorithm for a linear classifier. For the Perceptron task, repeated buyers are labeled as 1 and non-repeated buyers are labeled -1. We withheld 20% of the total data set as test data and applied the Perceptron algorithm to the 80% training data. The test data and train data contain 41775 and 167100 examples.

We iterated the Perceptron algorithm 100 times and updated weights w each time when the predicted value is different from the correct answer.

We tested the effectiveness of our model by evaluating the trained weights by comparing predicted buyers and the correct answers. The model worked out fairly well and is consistent for both in sample and out of sample data, and achieved accuracy of 92.7% and 92.8%, respectively.

4 Further Steps

Transform the Generalized Baseline Data The prediction goal in this midterm report model is to predict whether the buyers are retained on the Tmall platform after promotion. We want to extend this goal to predicting if a user will return to a particular merchant. The generalized model in this midterm report focuses more on the user portrait, this data will be transformed into detailed activity feature over the ¡user, merchant¿ pair.

Apply New Models The data set will be high-dimensional and sparse after we rebuild our model over the ¡user, merchant¿ pair because only a few users out of the entire user pool interact with a given merchant. The data set will also have heterogeneous data types containing categorical and Boolean features. To handle this data set we will explore Generalized Low Rank Models (GLRMs), a technique, developed by Professor Udell, used for embedding data sets like ours into a lower dimensional space [1]. The goal of this new model will be to improve our prediction performance from the benchmark.

References

[1] Madeleine Udell, Corinne Horn, Reza Zadeh, Stephen Boyd *Generalized Low Rank Models: https://arxiv.org/abs/1410.0342* (October 2017)