**Project Proposal: Alibaba O2O shops’ coupon usage prediction**

MIS 382 Fall 2017

Jack Cheng Sandra Cui, Christine Mulcahy, Elena Reynolds, and Steven Zheng

**Problem Description**

In nowadays business environment, the Internet has become the major part of a business organization and almost all companies start to use their webpage as well as apps to do marketing to attract more customers. O2O, which stands for ‘Online to Offline’ is just one of these strategies. A typically offline store uses the online platform to do marketing and branding, trying to reach out more customers through sharing information, providing coupon and other methods. These shops can offer the online coupon, offline coupon or both of them. Understanding for some certain type of shop, which method would be better and trying to predict the usage rate for these coupons would create great profit for both the platform as well as the shop itself. The O2O marketing is typically based on a large amount of users’ behavior and location information recorded by various APPs, thus admits an ideal application of big data and machine learning.

**Dataset Description**

The dataset is about the O2O coupon usage on Alibaba.com, through Tianchi.com[[1]](#footnote-1), one of the biggest open dataset for Alibaba([BABA](https://www.google.com/search?q=NYSE:BABA&stick=H4sIAAAAAAAAAONgecRozi3w8sc9YSm9SWtOXmPU4OIKzsgvd80rySypFJLiYoOyBKT4uHj00_UNC82STNIMC9J4AGV2bqI8AAAA&sa=X&ved=0ahUKEwiQqYSNxpLXAhVKziYKHUVeAZMQsRUIxAEwEw) (NYSE)), which is the biggest e-commerce platform in China. It consists of two parts transaction data, both online and offline. The online part has 11,429,826 instances and 7 attributes. And the offline part has 1754884 instance with 7 attributes. Since this whole dataset is not originally build for competition use, so all of these data are in raw database forms which means our team needs to do more feature engineering tasks. Each row represents a transaction or online action made by a certain user with a certain merchant within the January and June of 2016. It has the columns as follow:

1. User\_id: The unique user id for every customer
2. Merchant\_id: The unique merchant\_id for every shop
3. Coupon\_id: The coupon\_id this customer received, null for not receiving any coupon
4. Discount\_rate: A real number x in [0,1] denotes the discount rate, while `x:y’represents the discount is CNY[[2]](#footnote-2) (Chinese currency unit) `y’Yuan only if the transaction costs at least CNY `x’ Yuan. `Fixed’ denotes a fixed and discounted price transaction.
5. Date\_received: Date the customer received the coupon.
6. Date: Date the transaction occurred.

If Date = null & Coupon\_id != null: The user gets a coupon without usage

If Date != null & Coupon\_id = null: The user bought without a coupon

If Date != null & Coupon\_id != null: This transaction is with a coupon

For the 7th attribute, it is different in the online data and the offline data:

In online data, we can capture the user action of:

1. Action: 0 for click, 1 for purchase, 2 for getting a coupon.

With the offline data, we can know the distance of the merchant and the user’s location that he or she often presents.

1. Distance: x in[0,10], which denotes the distance between the shop and the nearest location that the user often presents is x\*500m. ‘null’ means not applicable，0 for less than 500m, and 10 for father than 5000m

I have uploaded the dataset to google drive in case it is hard to open the original Alibaba.com in the US.

https://drive.google.com/open?id=0BztghCAqaCFQUmh3RVZJZG9BSkE

**Some key advantage of this dataset**

Firstly, this dataset is large enough with over 10 million instances but still be able to process just with pandas in python. And also, this dataset is not built fixed for certain type of prediction, one possible way our team found is to predict the usage of a coupon given the customer and the merchant, but the usage is not even a dummy in the dataset and need our team to build out of the raw data, which really enable us to have a practice in the feature extraction skills. Third, this dataset is recently posted (actually just posted on 2017.08.03), so there is no such kind of ‘Winning Strategy’ that might limit our approaches. The dataset is combined with online and offline data which we can compare and get some insight for the O2O industry. Last but not the least, this data is the open data by Alibaba, with specific data metadata description (That is something a lot of kaggle data doesn’t have) and from real business environment with a rather long user history data of half a year, our team can also do other type of prediction such as the expected days between a customer receives a coupon and uses that coupon or which day of the week is best for delivering a coupon. Our team can use the first 5 month as a training set and the 6th month as the test set and therefore simulate a real business problem.

**Potential Approaches**

The very first step of analyzing this dataset would be feature extraction and feature engineering. Since the whole dataset is in raw data form, our team need to build new interpretable features out of it, here are some features that we thought might be useful and available from this raw data:

Customer related features:

1. Number of time this customer gets a coupon
2. Number of time this customer uses a coupon
3. The relative ratio of a customer’s online and offline purchasing within those months
4. The average amount of coupon the customer got a week.
5. The average distance of the shops this customer uses a coupon.

Merchant related features:

1. Number of time this merchant sends out a coupon
2. Number of time this merchant’s coupon be used
3. Number of different type of coupon this merchant provides
4. The average distance of the merchant to the receivers of coupons.

Customer-Merchant related features:

1. Number of time this customer gets the coupon from this merchant
2. Number of time this customer uses the coupon from this merchant
3. The ratio of this merchant’s coupon to all the coupons the customer receives
4. The ratio of customer doesn’t use the coupon to purchase at this merchant
5. If this customer goes to both online and offline stores of this merchant

Coupon related features:

1. Type of that coupon (directly discount or X China Yuan off once purchase up to Z China Yuan)
2. The equivalent discount rate of the coupon
3. The minimum required amount of purchase to use the coupon (Z China Yuan)
4. Which day of week is the coupon delivered
5. Which day of month is the coupon delivered
6. Number of times this kind of coupon is provided
7. The total number of times of certain customer uses this type of coupon

We will develop models to predict whether the customer would use this coupon within 15 days after receiving this coupon (the time of a coupon is valid in Alibaba.com) and maybe also the expected days between receiving the coupon and the purchasing action. These are the possible models that might be useful:

*Linear Regression*

The goal of using linear regression is to obtain an interpretable model. To help maintain interpretability, we will utilize feature selection techniques such as forward and backward selection or LASSO regression to limit the number of features and complexity of the model.

*Decision Trees*

In developing our decision tree model, we will tune the model using measures such as node purity, number of leaves, and depth of branches. Especially the random forest and the XGBoost which are really famous in Kaggle competitions.

*Neural net*

Though the neural net is hard to interpret, we still want to try if the neural net can provide some precise prediction about this problem and practice what we learned in the class.

*K-Nearest Neighborhood*

It might be useful to compare the customer and merchant to some similar neighbors to get a better prediction. We might either try this method or integrate some features of the similar neighbors to the other prediction models.

*Ensemble Methods*

We will ensemble different method by majority voting or weighted average and a threshold, also the famous Adaboost and GBDT models.

**References**

1. Data’s description page:

<https://tianchi.aliyun.com/datalab/dataSet.htm?spm=5176.100073.888.30.62b8a501JD1dj8&id=23>

1. The google drive of that data:

<https://drive.google.com/open?id=0BztghCAqaCFQUmh3RVZJZG9BSkE>

1. A paper in the related field of topic that we referenced when developing this proposal:

Daskalova N, Bentley F R, Andalibi N. It's All About Coupons: Exploring Coupon Use Behaviors in Email[C]// CHI Conference Extended. 2017:1152-1160.

1. Data’s Description is on this website: (Might be slow when open in the US)

   https://tianchi.aliyun.com/datalab/dataSet.htm?spm=5176.100073.888.30.62b8a501JD1dj8&id=23 [↑](#footnote-ref-1)
2. Chinese currency is called ‘China Yuan’ where ‘Yuan’ means dollar. [↑](#footnote-ref-2)