Repeat Buyers Prediction: a Feature Engineering Approach

Presented by Jianxun Lian

University of Science and Technology of China

Introduction

- To predict which shoppers would become repeat buyers after sales promotion in Tmall.com
- User behavior logs

User	Item	Category	Merchant	Brand	Time stamp	Action

- Problem

User	Merchant	Repeat purchase?

Binary classification problem

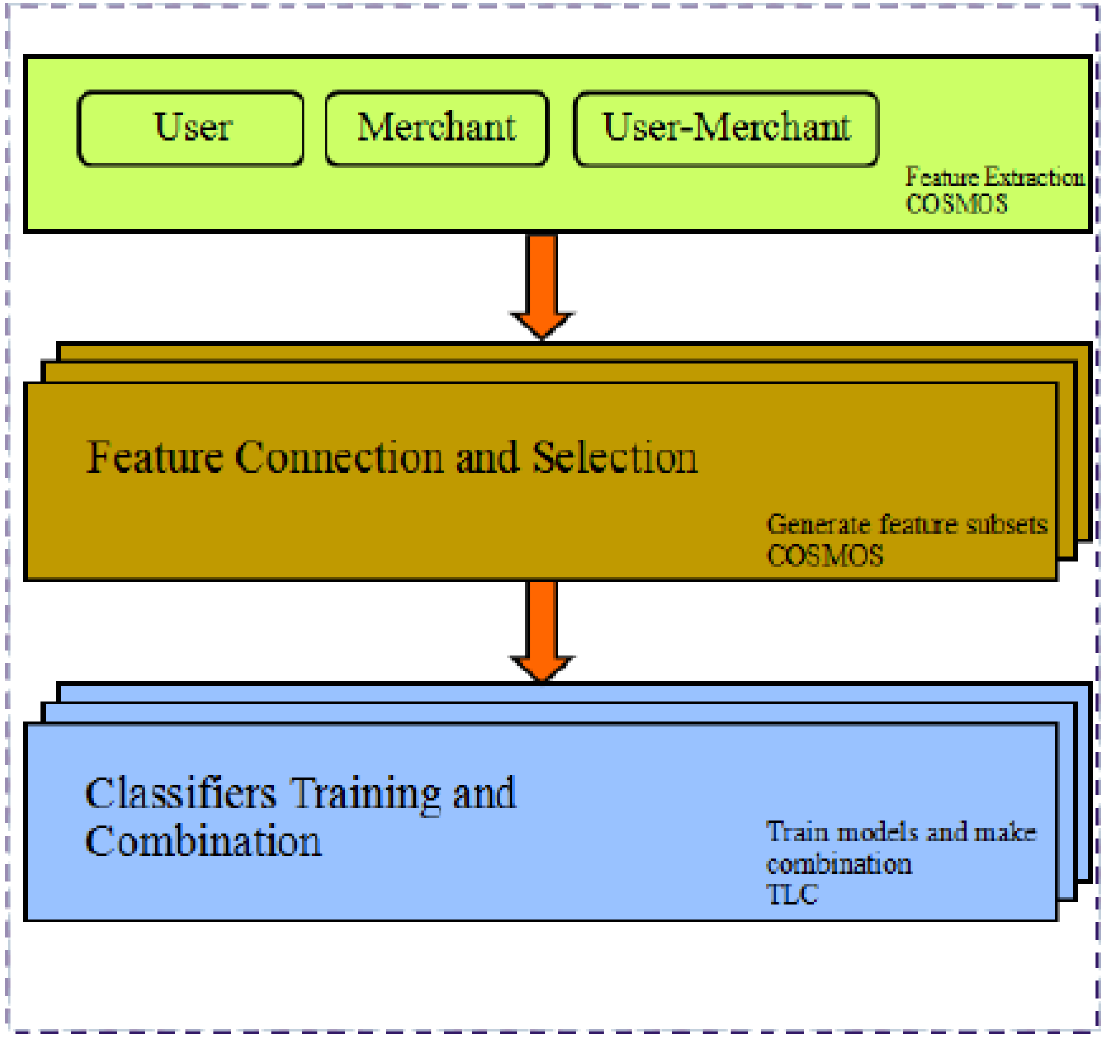
- The simplest (and perhaps most efficient) way to model the problem

User	Merchant	Repeat purchase?

User & Merchant	Feature_1, Feature_2, … Feature_n	label

- The features come from 3 pillars:
  - User level
  - Merchant level
  - User-merchant level
- Labels
  - 1: repeat buyer
  - 0: not repeat buyer

The framework(mainly for stage 1)



Feature engineering

- User pillar
  - Overall statistics
    - Some general aggregation value of the user's past shopping
    - E.g., total_number_of_purchase, total_number_of_categories
  - Lifespan
  - Buy-to-Click ratio
    - The total number of purchases divided by total number of clicks
    - Further refined into category, brand, merchant and item level
  - Temporal behavior
    - How users' behavior change with date
    - E.g. $lift_{festival} = \frac{number\ of\ double\ 11\ purchase}{number\ of\ purchase\ from\ the\ other\ days}$,
    - E.g. the last week, the last 3 months
  - Repeat behavior
    - Historical repeat purchase (before Double 11 day)
    - Assumption: users' behavior would be consistent over the time
    - E.g. $RBR_{item} = \frac{number\ of\ repeat\ purchased\ items}{number\ of\ total\ purchased\ items}$
  - Behavior Entropy
    - BE describes the amount of variation within a user's activity
    - $BE = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$, $p(x_i) = \frac{number\ of\ actions\ on\ x_i}{number\ of\ total\ actions}$
  - Demography

Feature engineering

- Merchant pillar
  - Overall statistics
  - Lifespan
  - Buy-to-Click ratio
  - Temporal behavior
  - Repeat behavior
  - Promotion Frequency
    - How often the merchant would have sales campaigns
    - Count the spikes in the sales curve
  - Prior repeat user(category, brand) ratio
    - The merchant id set of the training and test are identical
    - Calculate each merchant’s repeat user ratio based on training data
    - Fight with overfitting
      - AUC is poor if used this feature directly
      - Simulate the random splitting process: improve a little
      - Cut the low frequency merchants, and set their value to -1 for generalization

Feature engineering

- User-Merchant pillar
  - Overall statistics
  - Lifespan
  - Buy-to-Click ratio
  - Temporal behavior
  - Remaining items
    - Some items(categories/brands) has been clicked, but not purchased yet
  - Merchant attraction
    - The current merchant’s relative rank among all the user’s historical merchants
    - How many categories/brand are in the intersection of user’s favorite ones with merchants’ top sellers

Model selection and combination

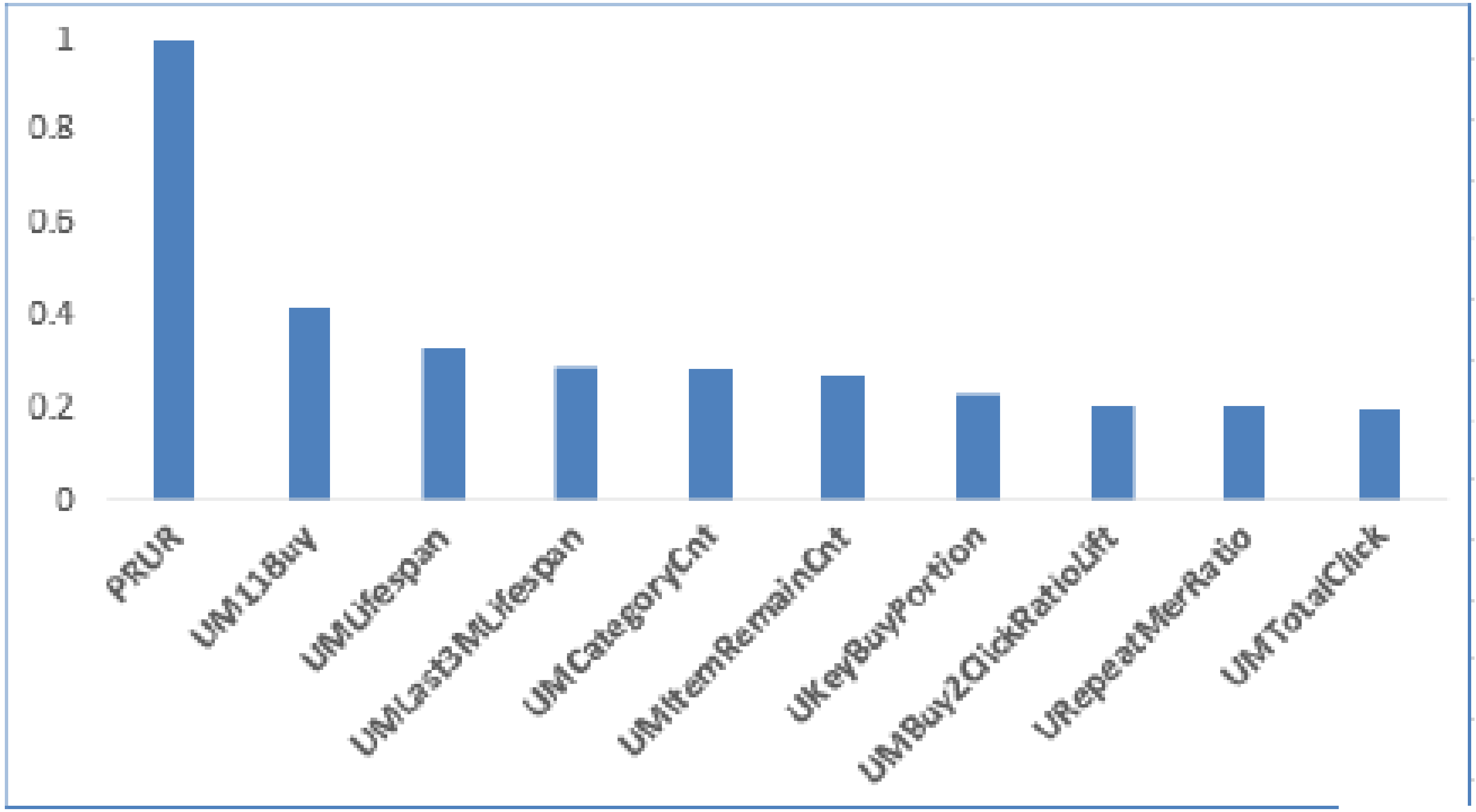
- We tried 6 kinds of classifiers
  - All were done by Microsoft's internal ML tool TLC

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

  - Grid search the parameters
  - Finally choose to linearly combine GBDT, LR, and NN with weights 0.82, 0.09, 0.09

Feature importance

- Top 10 most important features from GBDT



Feature combination

- Train models based on different feature subset and then combine the results together could improve the performance
  - Learn from last year's TianChi competition
  - Combine features from 3 pillars:
    - Totally $2^3 - 1 = 7$ kinds of sub-features
  - With or without Prior Repeat Ratio
    - 3 feature sets:
      - without prior repeat ratio
      - with prior user repeat ratio
      - with prior user repeat ratio and prior category/brand repeat ratio
    - Improve the AUC significantly

Results

- Stage 1: AUC 0.701789 , rank 15th

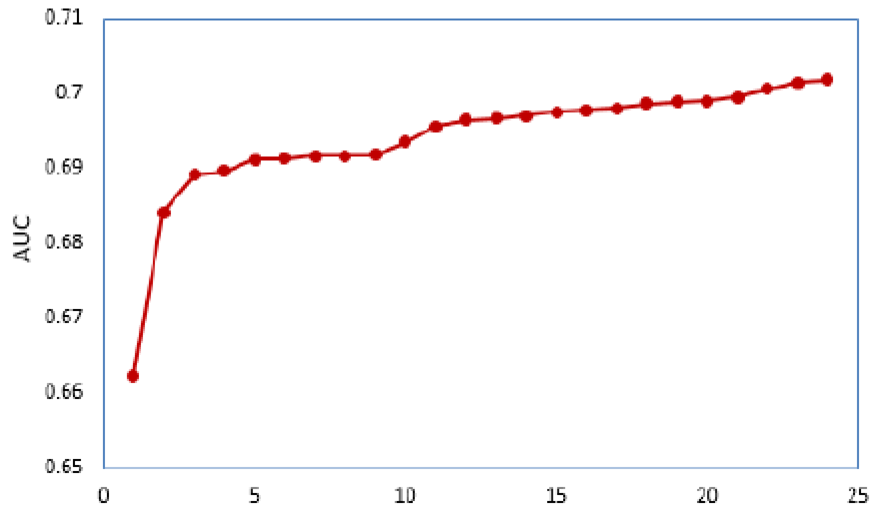


Figure 3: AUC improvement history in stage 1

- Stage 2: AUC: 0.711287

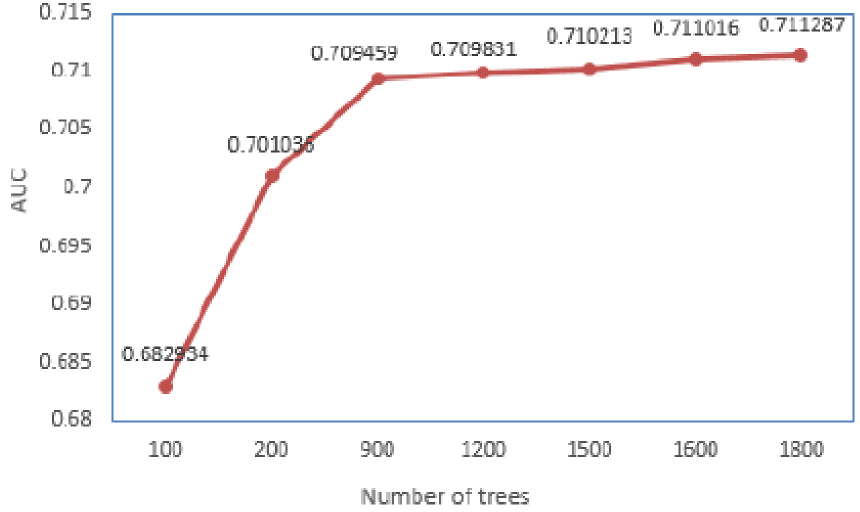


Figure 4: AUC improvement with the number of trees in GBDT

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

Table 2: GBDT parameters at stage 2

Conclusions

- Binary classifications
- Extract features from user, merchant, user-merchant pillars
- Training models on sub-features
- Linear combination of different classifiers

Thanks