北京郵電大學
Beijing University of Posts and Telecommunications

ICT 中科院计算所
INSTITUTE OF COMPUTING TECHNOLOGY,CAS

Repeat Buyers Prediction after Sales Promotion for Tmall Platform

Presenter: Bowei He*

Zhiqiang Zhang*, Jian Liu*,

Fuzhen Zhuang[I], Chuan Shi*

*Beijing University of Posts and Telecommunications

[I]Institute of Computing Technology, Chinese Academy of Sciences

Outline

- Task Description
- Framework Overview
- Data Processing
- Feature Extraction and Selection
- Model Training and Ensemble
- Experiment Results

Task Description

- Tmall.com hold big promotions on ‘Double 11’ to new buyers
  - Most users are one-time deal hunters
  - Identify the potential customers who can be converted into repeat buyers

天猫 TMALL.COM

Task Description

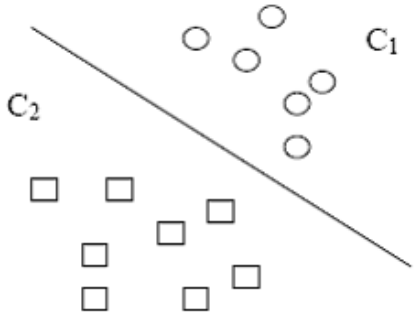
- The competition consists of two stages:
  - Stage 1
    - 212,062 users and 4,995 merchants
    - Build the model offline
    - Submit the prediction results for evaluation
  - Stage 2
    - Share the same set of merchants in Stage 1
    - Build the model on Alibaba's cloud platform
    - Submit the code to the cloud platform

Outline

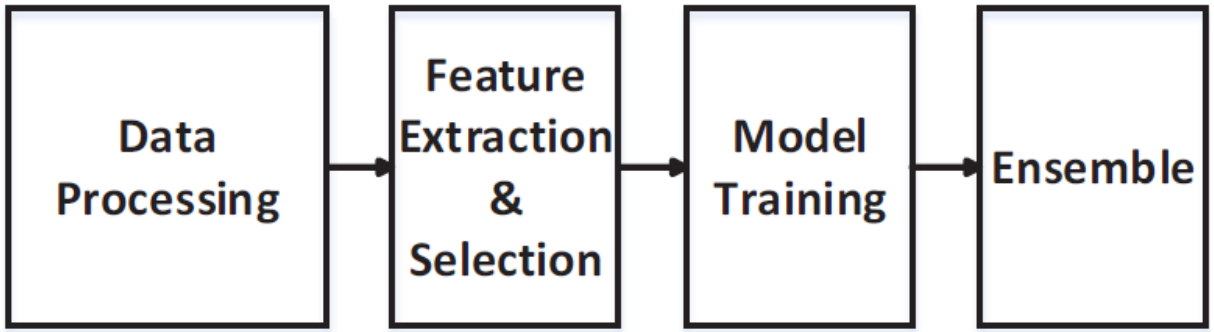
- Task Description
- Framework Overview
- Data Processing
- Feature Extraction and Selection
- Model Training and Ensemble
- Experiment Results

Framework Overview

- Formulate the task as a binary classification problem



- Our solution has four steps

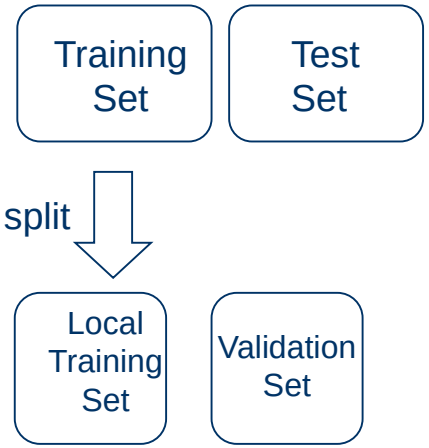


Outline

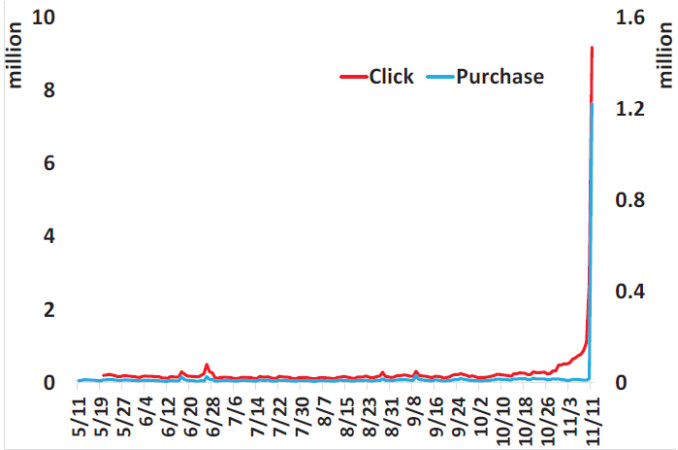
- Task Description
- Framework Overview
- Data Processing
- Feature Extraction and Selection
- Model Training and Ensemble
- Experiment Results

Data Processing

- Offline Evaluation
  - Observation:
    - Training set and test set has different users but same merchants
  - Solution:
    - Generate the local training set and validation set in the same way



- Data Statistics
  - Observation:
    - Users' behaviours grow explosively on 'Double 11' day
  - Solution:
    - Divide the users' feature into on "Double 11" day and before "Double 11" day



Outline

- Task Description
- Framework Overview
- Data Processing
- Feature Extraction and Selection
- Model Training and Ensemble
- Experiment Results

Feature Extraction

- Extract the features from three aspects:
  - User-related feature
  - Merchant-related feature
  - Interactive feature between users and merchants

Feature Category	Feature Sub-category	Feature examples
user feature	User Information Feature	age_range / gender
	User Count Feature	user_click_cnt / user_cart_cntuser_buy_cnt / user_fav_cnt / user_active_clickday_cnt …
	User Ratio Feature	user_click_11_item_ratio / user_cart_11_item_ratio / user_buy_11_item_raito …
merchant feature	Merchant Count Feature	shop_click_cnt / shop_cart_cnt / shop_buy_cnt/shop_fav_cnt / shop_click_cnt_before_10 …
	Merchant Ratio Feature	shop_11_old_user_ratio / shop_11_new_buy_ratio / shop_dup_user_before_11_ratio …
	Merchant Data Leak Feature	shop_appear_in_train / shop_label_1_in_train / shop_label_1_in_train_ratio
mutual feature	Mutual Count Feature	user_shop_click_cnt / user_shop_cart_cnt / user_shop_buy_cnt / user_shop_fav_cnt …
	Mutual Cross Feature	user_shop_brand_cross_score / user_shop_cat_cross_score / user_shop_max_dup_item_before_11_ratio …
	Mutual Data Leak Feature	user_shop_cat_label_1_in_train / user_shop_cat_appear_in_train

User-Related Feature

- User-related feature include 3 sub-categories:
  - User profile feature
    - i.e., age and gender
  - User count feature
    - i.e., the number of a user's behaviours during days before "Double 11" and on "Double 11"
  - User ratio feature
    - i.e., users' behaviours in Nov. is different from the ones before Nov.

Merchant-related Feature

- Merchant-related feature include 3 sub-categories:
  - Merchant count feature
    - i.e., times of behaviours on the merchant
  - Merchant ratio feature
    - i.e., the repeat buyer ratio of a merchant
  - Merchant data leak feature
    - All merchants appear in both training and test set
    - The ratio of a merchant that be bought again
    - Avoid over-fitting

Feature In Stage 2

Problem:

Distributed Platform: 1 master node + many slave nodes

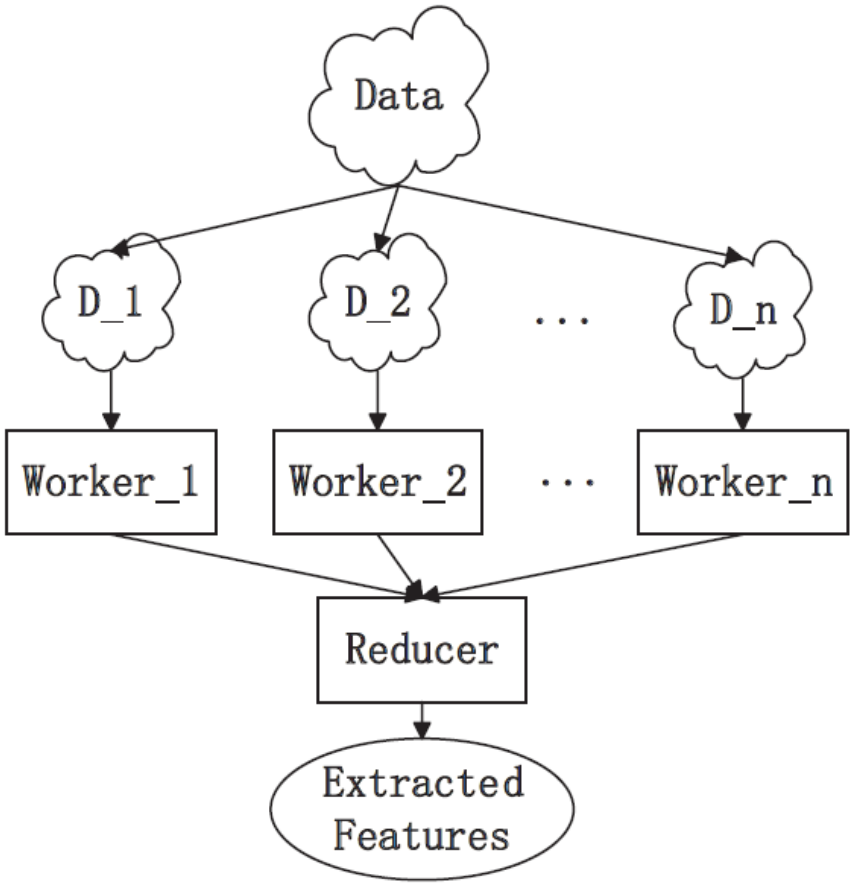
All log data of a user is divided into one slave node

All log data of a merchant is divided into many slave nodes

Solution:

The merchants in both stages are the same

Append merchant-related features in Stage 1 as an auxiliary file for Stage 2



Interactive Feature

- User-merchant interactive feature include 3 sub-categories:
  - Interactive count feature
    - i.e.Times of behaviours between user and merchant
  - Interactive crossover feature
    - User will repeatedly buy items in a merchant if the merchant sell some categories or brands that the user ever buy
  - Interactive data leak feature
    - Number of each category appearing in the training set influence the user to become a repeat user

Feature Selection

- Feature selection strategy in Stage 1
  - Cross validation
  - Extract 150 features in Stage 1

- Feature selection strategy in Stage 2
  - Running time is limited
  - Leave one out cross validation
  - 83 features out of 150 are obtained

Outline

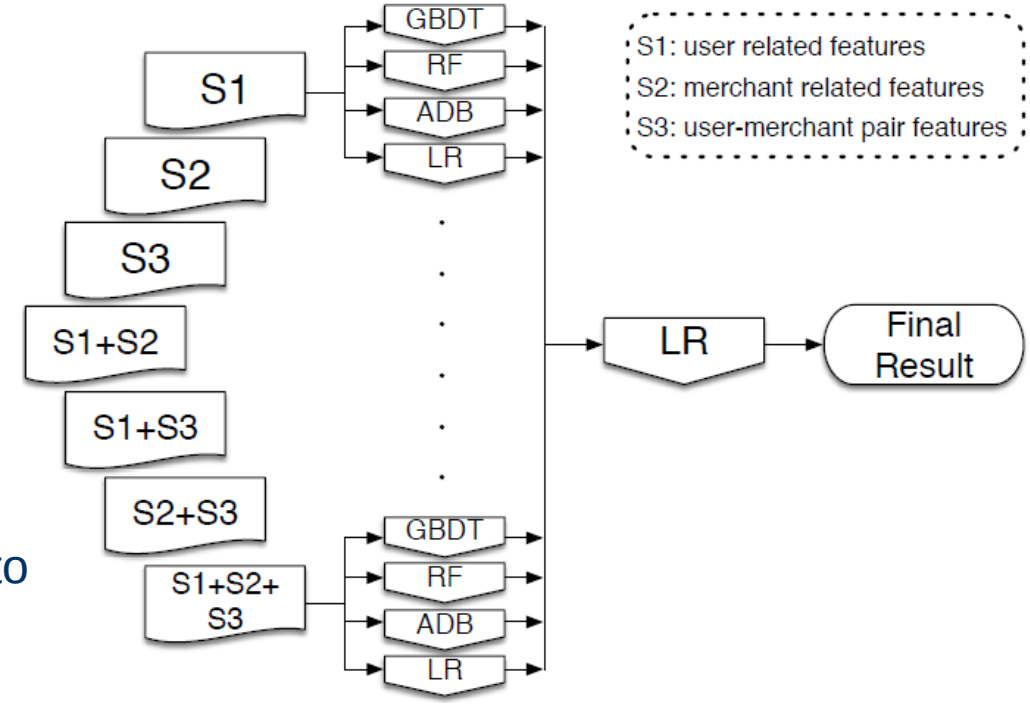
- Task Description
- Framework Overview
- Data Processing
- Feature Extraction and Selection
- Model Training and Ensemble
- Experiment Results

Model Training

- Four classifiers in Stage 1:
  - Gradient Boosting Decision Tree (GBDT)
  - Random Forests (RF)
  - AdaBoost (ADB)
  - Logistic Regression (LR)
- Only GBDT classifier in Stage 2

Ensemble

- Model Ensemble
  - Combine four models with different weights
- Feature Ensemble
  - All combinations of the three disjoint subset of features
  - Four models are trained on them independently
  - Apply the logistic regression to make final prediction



Outline

- Task Description
- Framework Overview
- Data Processing
- Feature Extraction and Selection
- Model Training and Ensemble
- Experiment Results

Experiment

- The hybrid ensemble achieves the best performance

Methods	AUC in local evaluation
GBDT	0.688379
Random Forest	0.688377
AdaBoost	0.683360
Logistic Regression	0.681488
Model Ensemble	0.691793
Model+Feature Ensemble	**0.692564**

- Stage 1 AUC = 0.699647 Rank : 17
- Stage 2 AUC = 0.711373 Rank : 1

Question

Thanks !

Q&A