Content

- Introduction
- Data Description
- Feature Extraction
- Obstacles
- Method & Result
- Conclusion & Limitation

Introduction

- Business Problem
  - Merchant: run big promotions
  - Customers: one-time deal hunters
  - Purpose: predict loyal customers for given merchant
- Transactional Data
  - The sales data of the “Double 11” shopping event in 2014 at Tmall.com
- Methods
  - Classification problem
  - Logistic Regression, Decision Tree, Random Forest, Gradient-Boosted Trees

Data Description

Test_table
User_id
Merchant_id
Probability

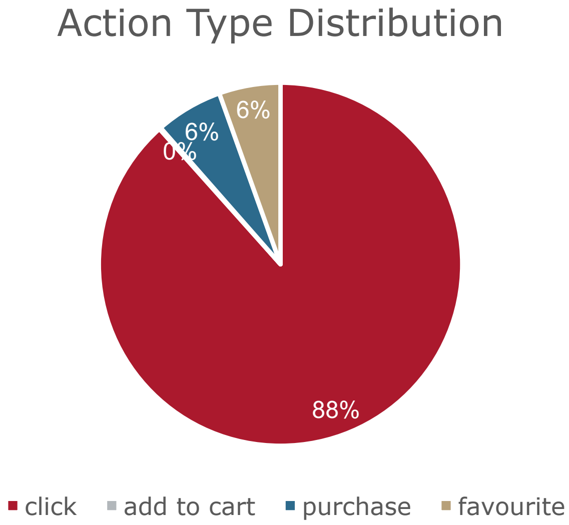
Trian_table
User_id
Merchant_id
label

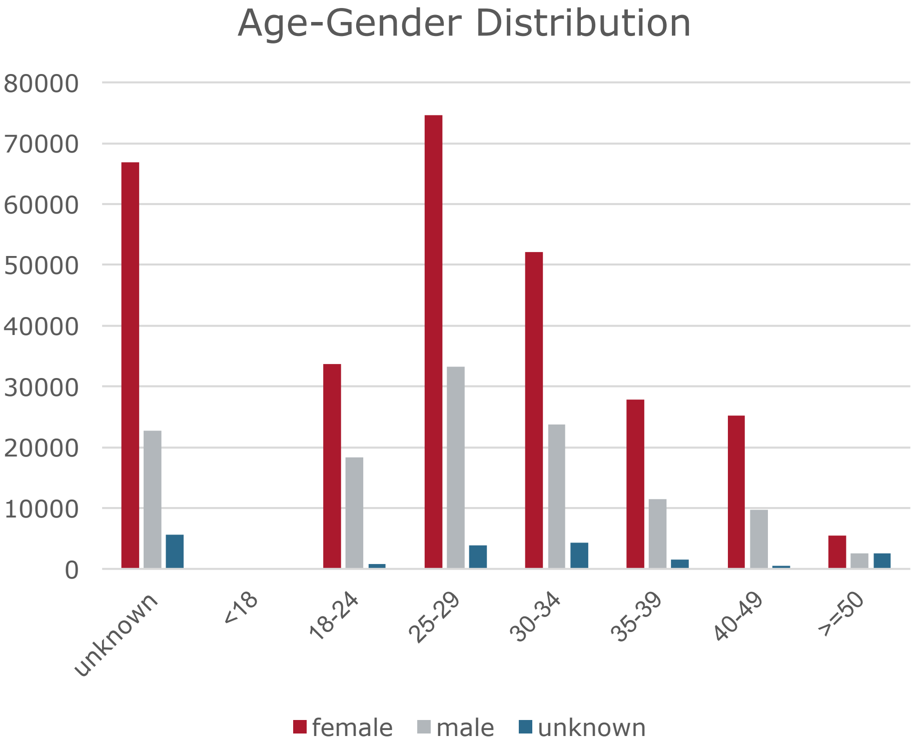
User_info
User_id
Age: [0-8]
Gender: [0,1,2]

data	train	test
users	212,062	212,108
merchants	1,993	1,993
pairs	260,864	261,477
Positive pairs	15,952	16,037
Positive %	6.12%	-

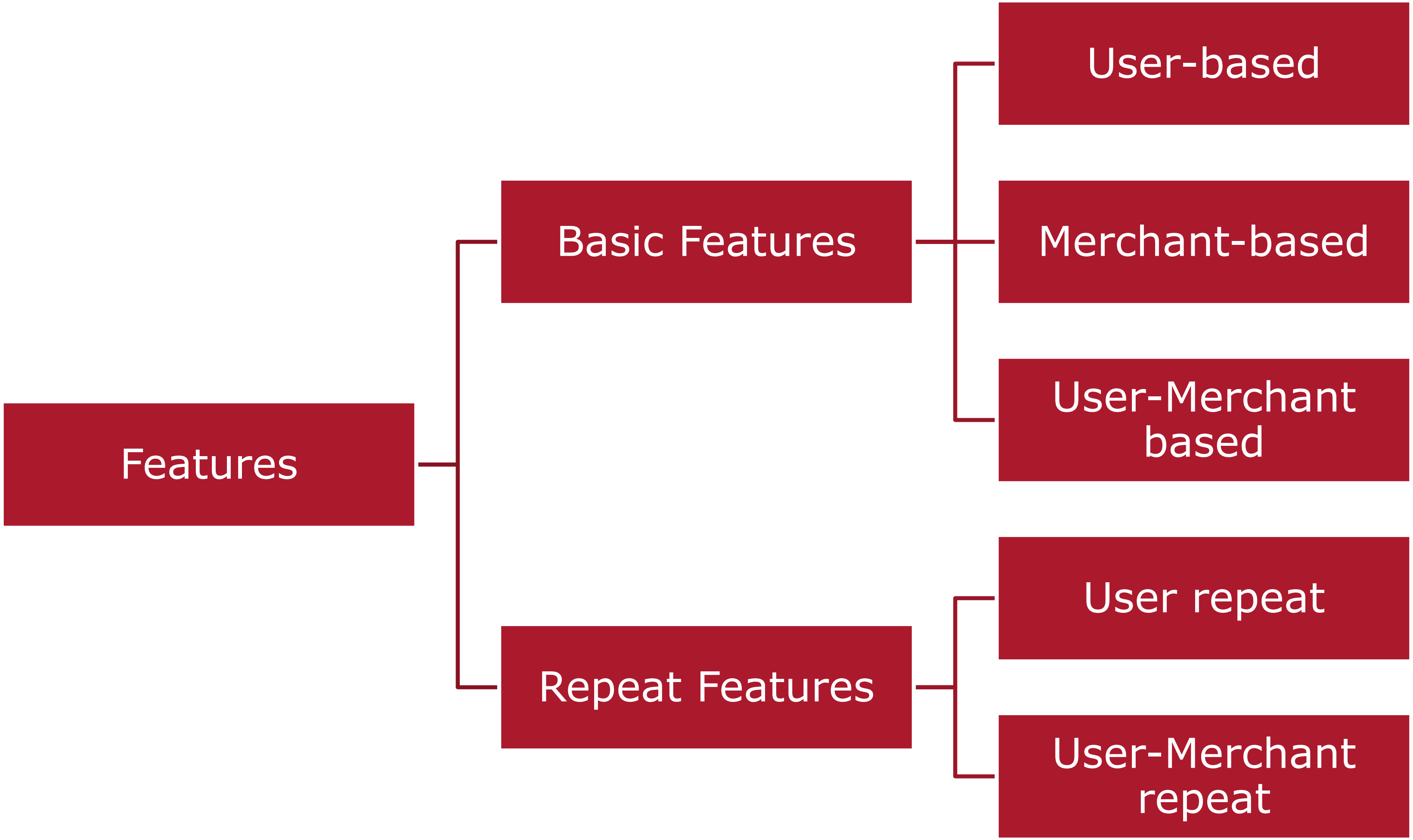
User_activity_log
User_id
Merchant_id
Item_id
Cat_id
Brand_id
Action_type: [0,1,2,3]
Time_stamp

Data Description





Feature Extraction



Basic Features

- User-based features
  - Capture a user’s overall buying behavior in terms of total actions made, number of merchants/items a user click/purchased/favored from, etc.
- Merchant-based features
  - Explore a merchant’s overall characteristics
- User-Merchant interaction based features
  - Capture affinity of a user to the merchant

Repeat Features

- User repeat features
  - Average span between any two actions
  - Average span between two purchases
  - The number of days since the last purchase
- User-Merchant repeat features
  - Average active days for a merchant
  - Ratio of merchants with repeated actions
  - Ratio of the number of merchants that the user made a purchase from to the total number of merchants that the user took some actions
  - How many times a user purchased again in a merchant

Features: # 75

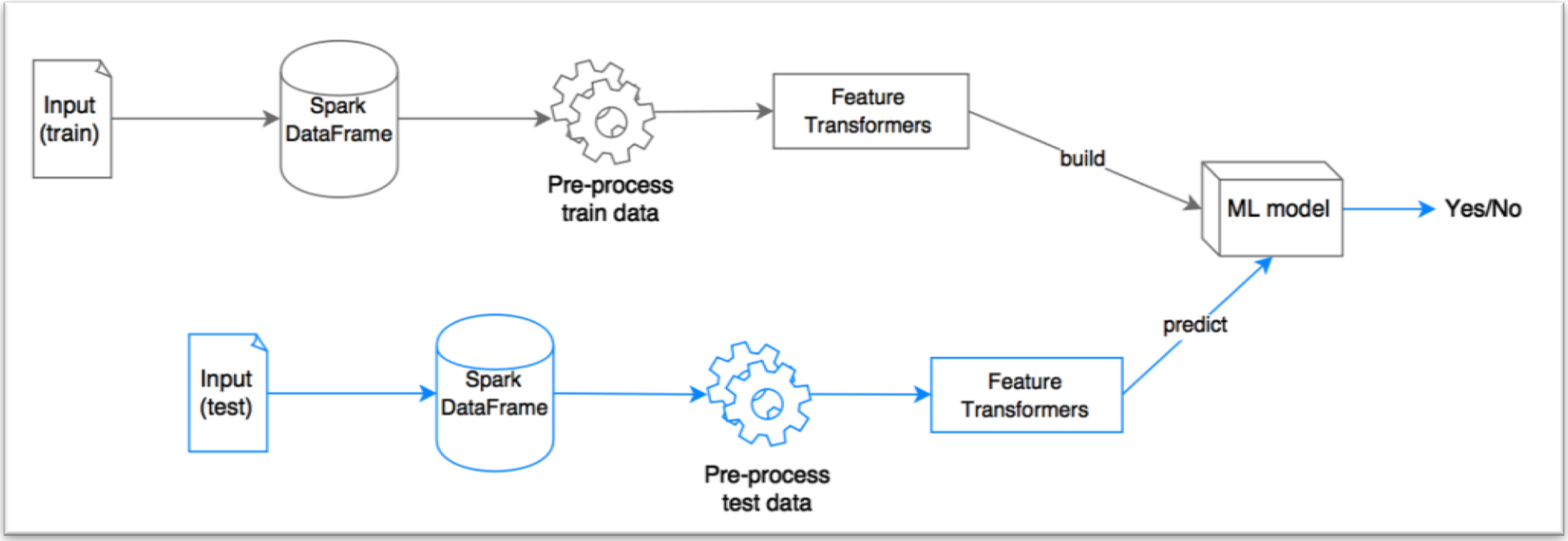
	Basic: #66		Repeat: #9
User-based	50	User repeat	3
Merchant-based	10	User-Merchant	6
User-Merchant	6		

Dataset	Features
10 schema	75

Obstacles

- Unbalanced data
  - Sparse metric
  - Class-imbalance ratio of 1:15
- Running time
  - Useless Row: join Training data and User Log table first
  - Out of Memory: save Feature data frame as csv file and reload it
  - Large Dataset: use small sample data to test

Method



Logistic Regression
Decision Tree
Random Forest
GBT

Results

Validation dataset		Balanced	Unbalanced
Basic Features		AUC	AUC
	LR	**61.74%**	59.45%
	DT	57.82%	78.86%
	RF	61.71%	**96.95%**
	GBT	59.70%	81.70%
Basic & Repeat Features		AUC	AUC
	LR	59.97% ↓	51.53% ↓
	DT	59.43% ↑	85.25% ↑
	RF	**60.93%** ↓	**96.99%** ↑
	GBT	59.59% ↓	83.35% ↑

Conclusion & Limitation

- Conclusion
  - Basic + repeat features
  - Random forest
  - Unbalanced data
  - 30th place in 971 teams
- Limitation
  - Ensemble method
  - More complicated features, like similarity, etc.,