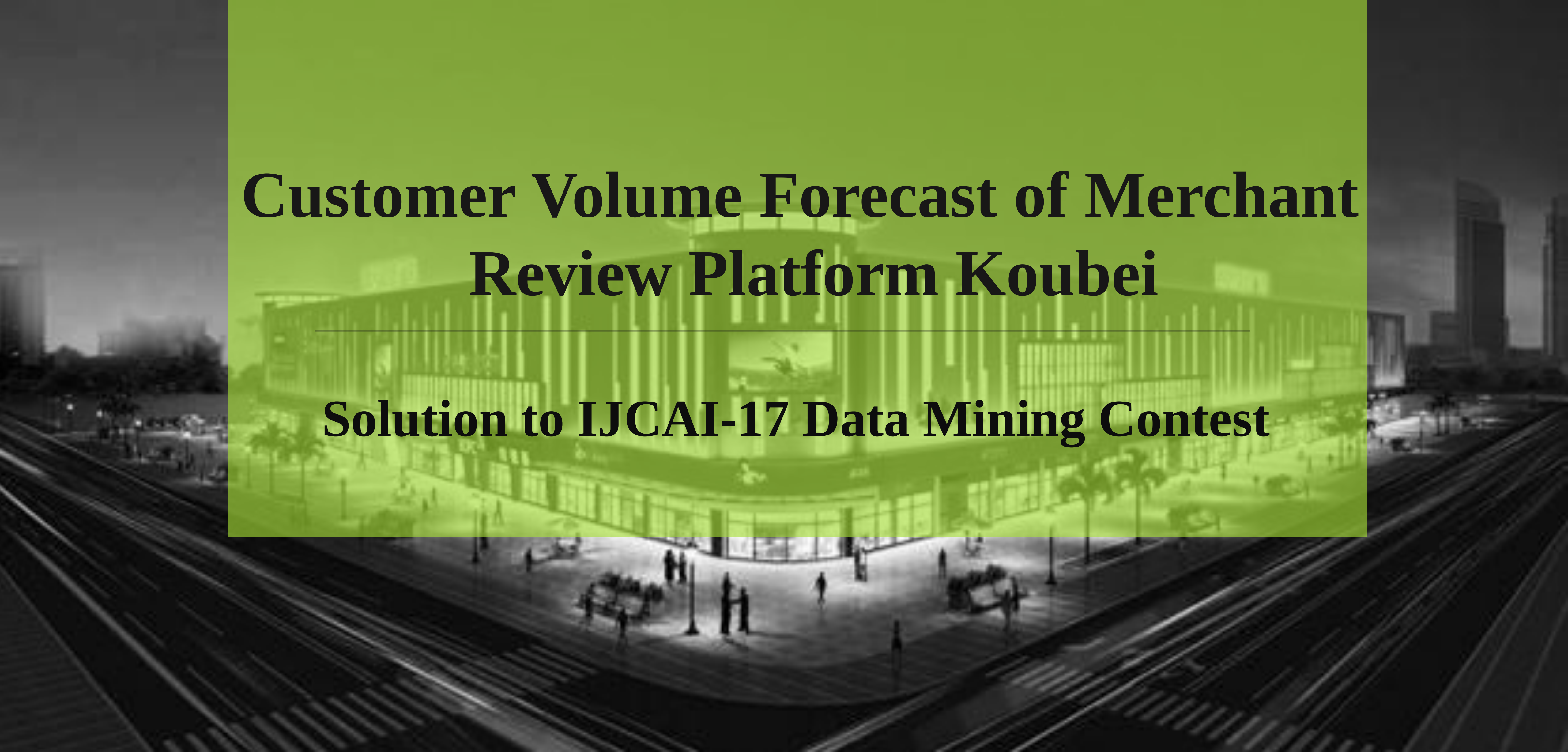


Customer Volume Forecast of Merchant Review Platform Koubei

Solution to IJCAI-17 Data Mining Contest

August 19th, 2017

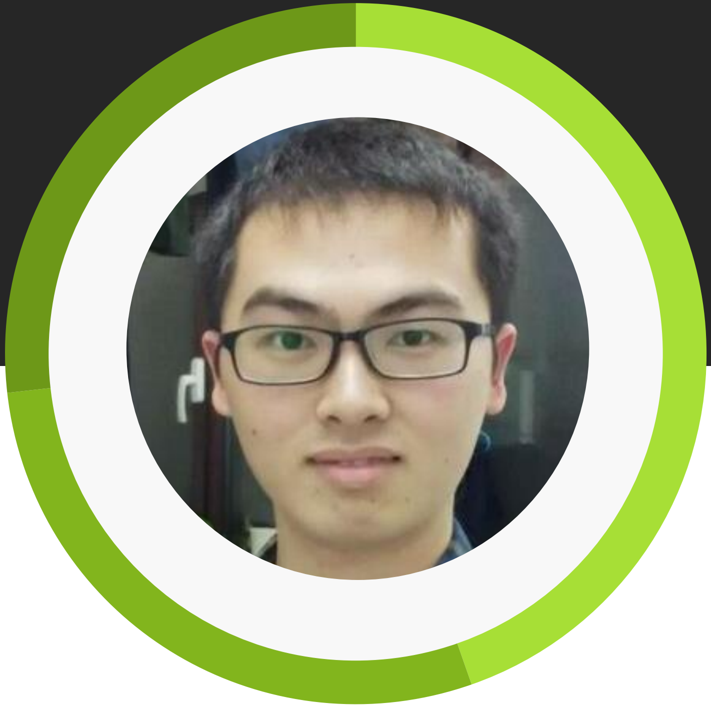
Zhongjie Li & Yichen Yao
Tsinghua University

OUR TEAM

Rank 4

ZHONGJIE LI

Ph.D. student
Department of Thermal Engineering
Tsinghua University



YICHEN YAO

Ph.D. student
Department of Engineering Mechanics
Tsinghua University

Solution Report

CONTENTS

01/ Introduction

02/ Data Analysis

03/ Data Processing

04/ Predictive Models

05/ Conclusions

FIRST

INTRODUCTION

Detail About the Competition

IJCAI-17 Data Mining Contest

Customer Volume Forecast of Merchant Review Platform Koubei

Background

- Held by Ant Financial in cooperation with IJCAI-17
- Dedicated to providing sales forecasts for each business. Based on the forecast results, businesses can optimize operations, reduce costs and improve customer experience.

Goal

- Given historical transaction logs of 2000 merchants in the past 17 months →
- Predict daily customer volume of each merchant in the next 2 weeks.

KOUBEI —— CUSTOMER VOLUME FORCAST

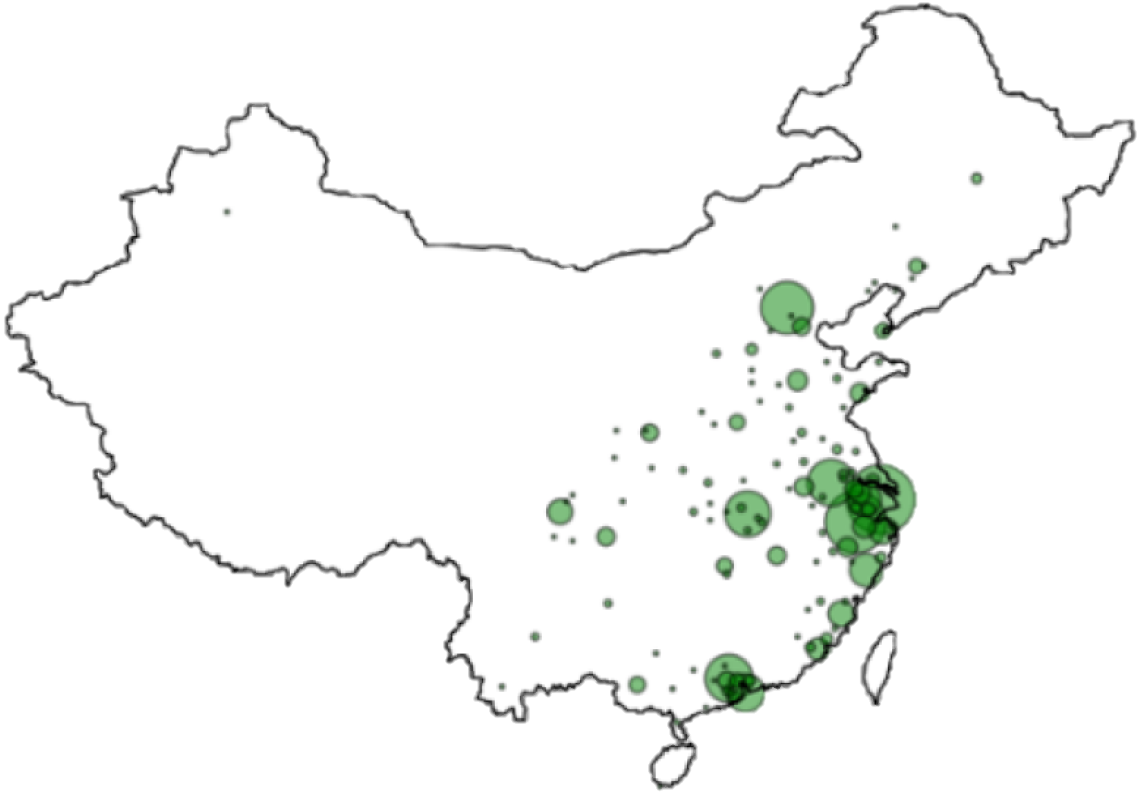
Forecasting business volume is critical to business management.

- **Customer Volume** – Defined as the number of users who use Alipay in the business hours
- **Data** – Users' browsing and payment history, and other business-related information for 2000 shops
  External data allowed.
- **Prediction** -- Daily business volume of 2000 shops in the next 14 days
- **Evaluation** -- $$L = \frac{1}{nT} \sum_{i}^{n} \sum_{t}^{T} \left| \frac{c_{it} - c_{it}^{g}}{c_{it} + c_{it}^{g}} \right|$$

SECOND

Data Analysis

Data Description



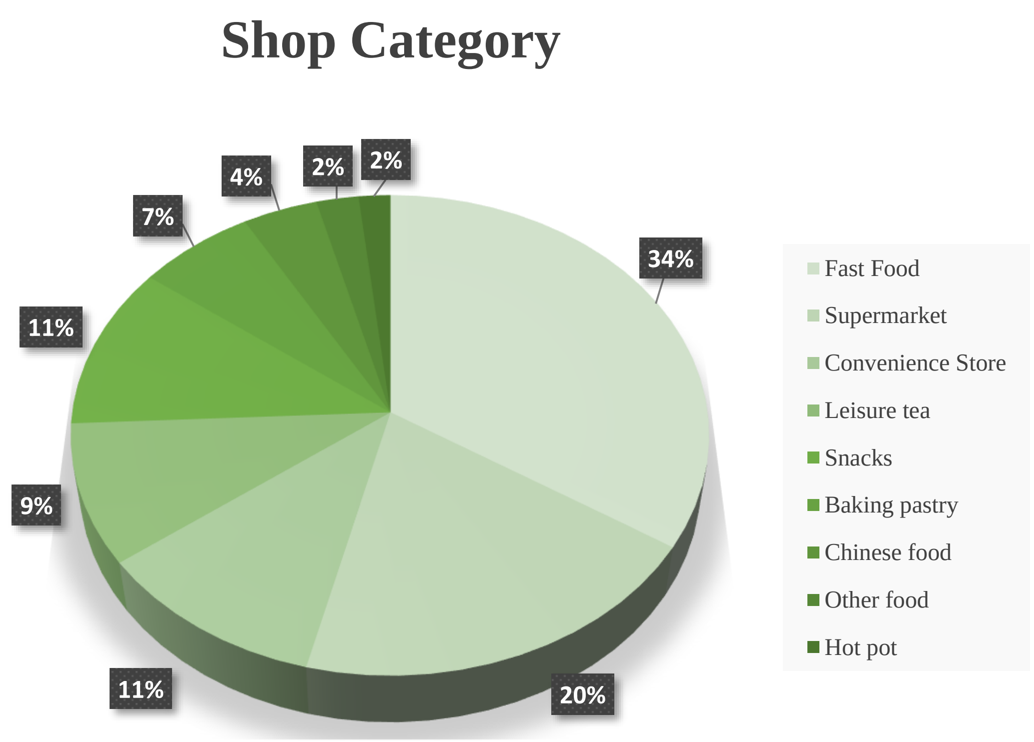
- ❑ 2000 merchants in 122 cities

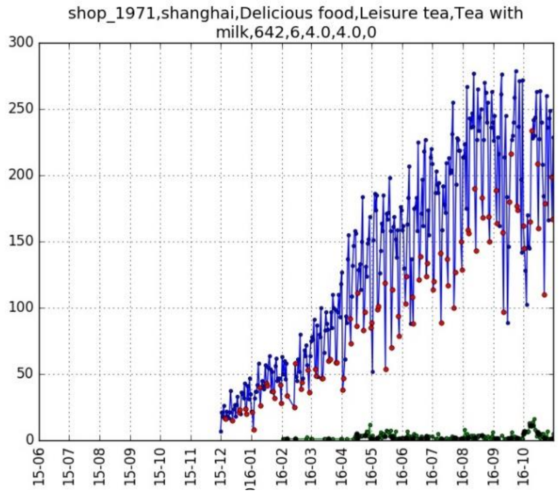
Shop Information
Shop ID
City
Location ID
Average Pay
Score by Users
Comment Count
Shop Level
1st Level Category
2nd Level Category
3rd Level Category

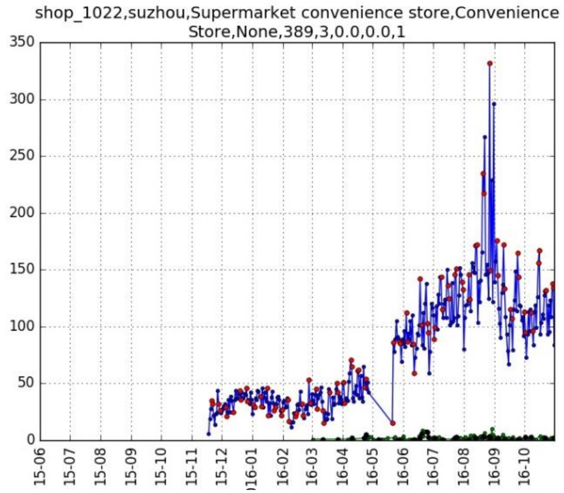
User Information
User ID
Shop ID
Pay Timestamp

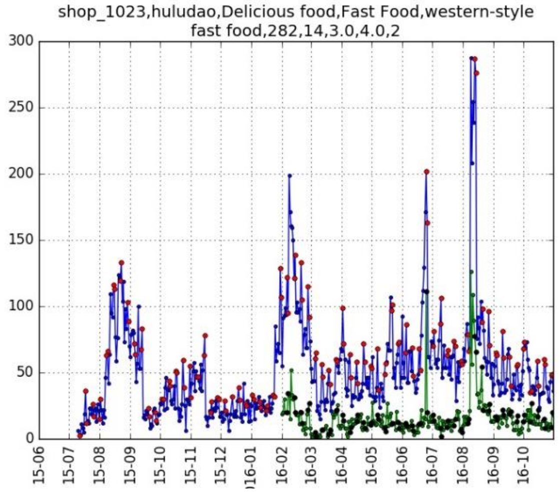
User Information
User ID
Shop ID
View Timestamp

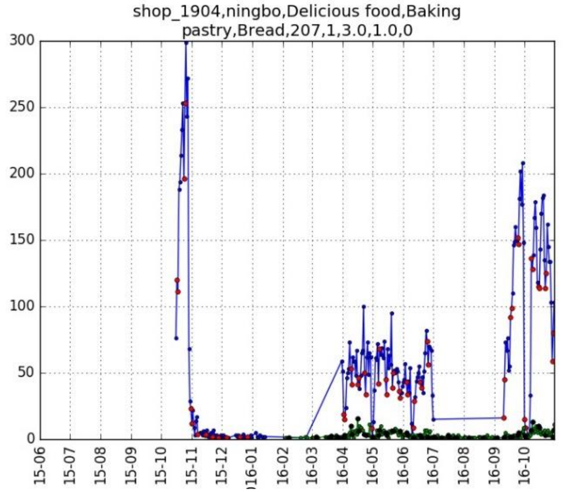
Data Description









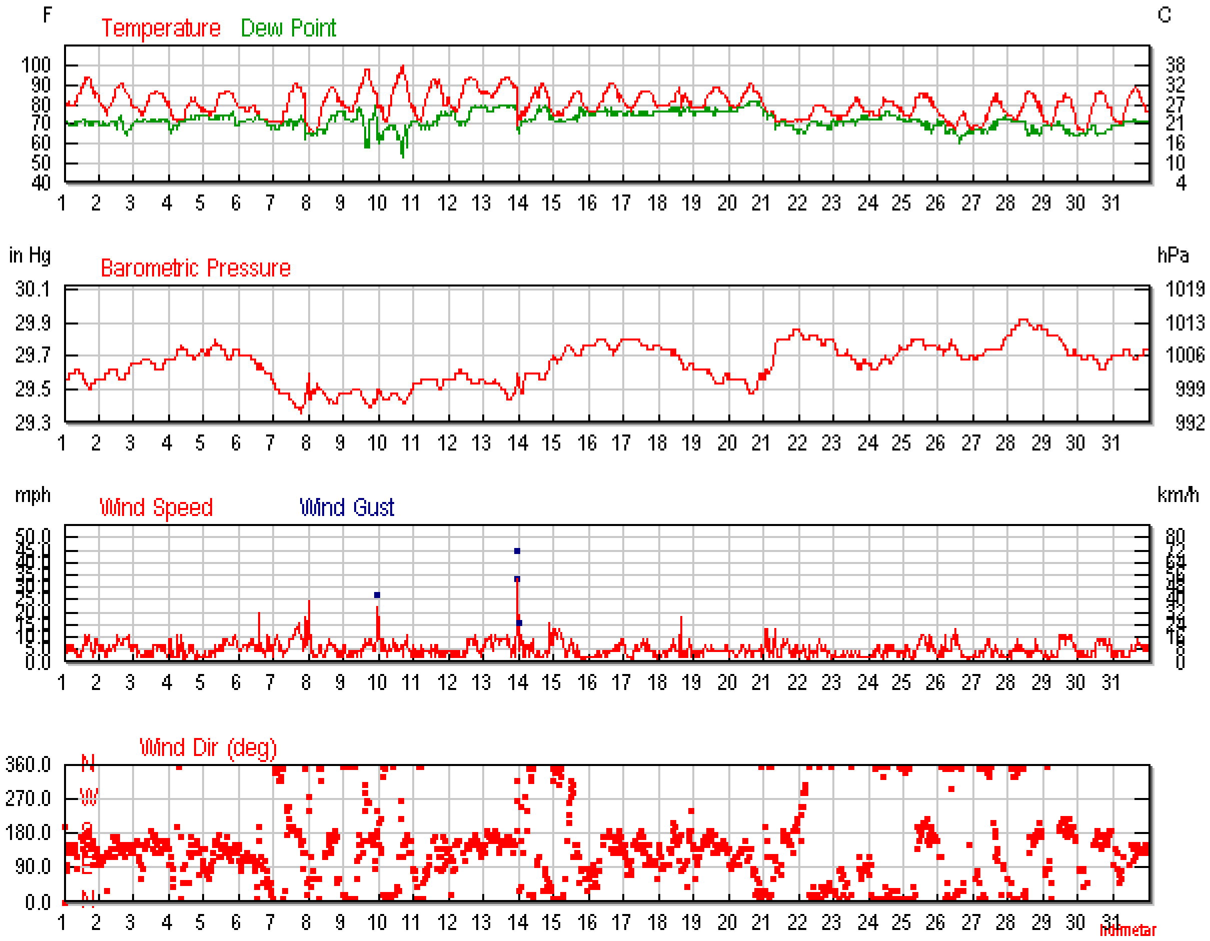


- The contest encouraged the use of external information such as weather.

Data Description

<Source: https://www.wunderground.com>

Monthly Weather History Graph



Daily Precipitation

Weather Detail

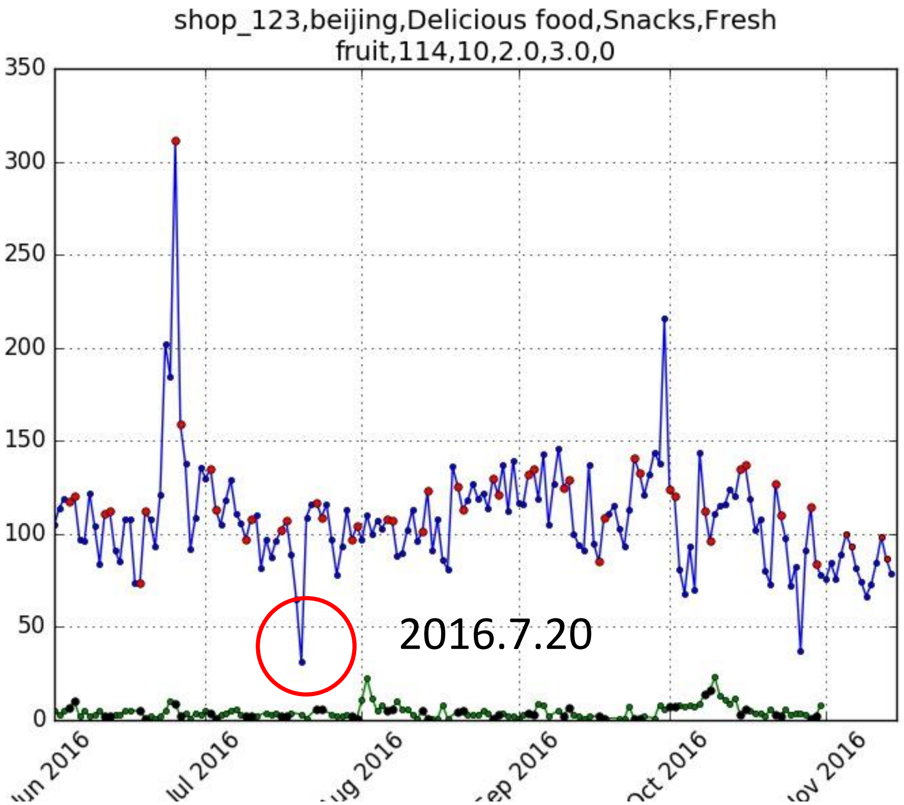
Rain & Clear Index

Comfort Index SSD

$$SSD = (1.818T + 18.18)(0.88 + 0.002F) + (T - 32)/(45 - T) - 3.2V + 18.2$$

T - Temperature
F - Humidity
V - Wind Velocity

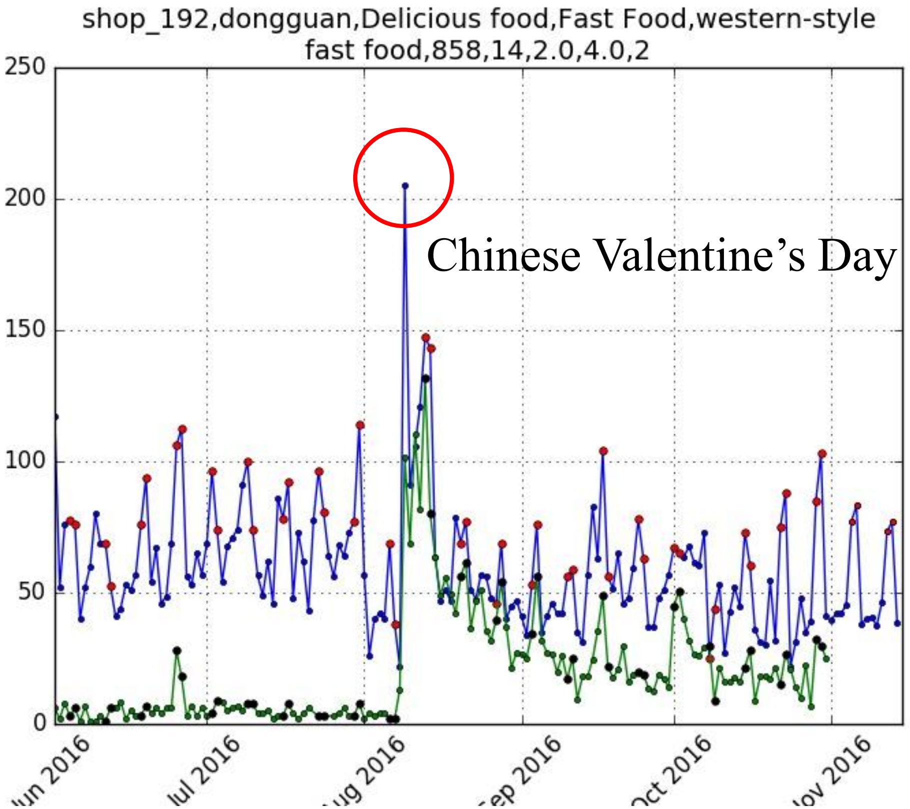
Weather Information



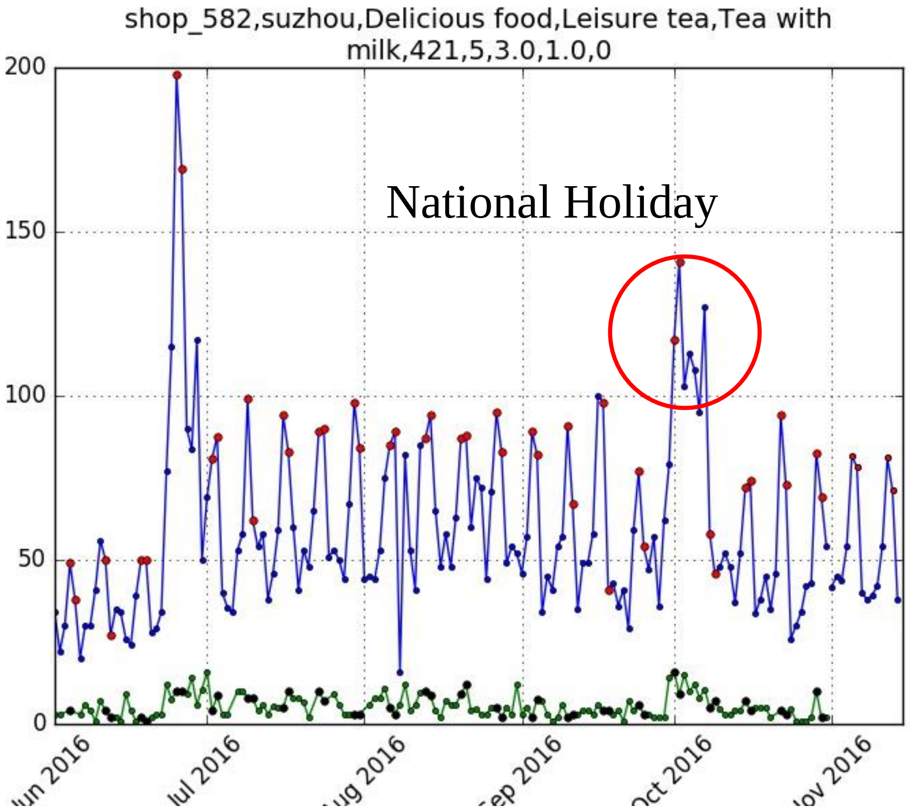
2016.7.20, Beijing, daily precipitation 210.7mm

Holiday Information

<From official forum of Tianchi Platform>



Chinese Valentine’s Day

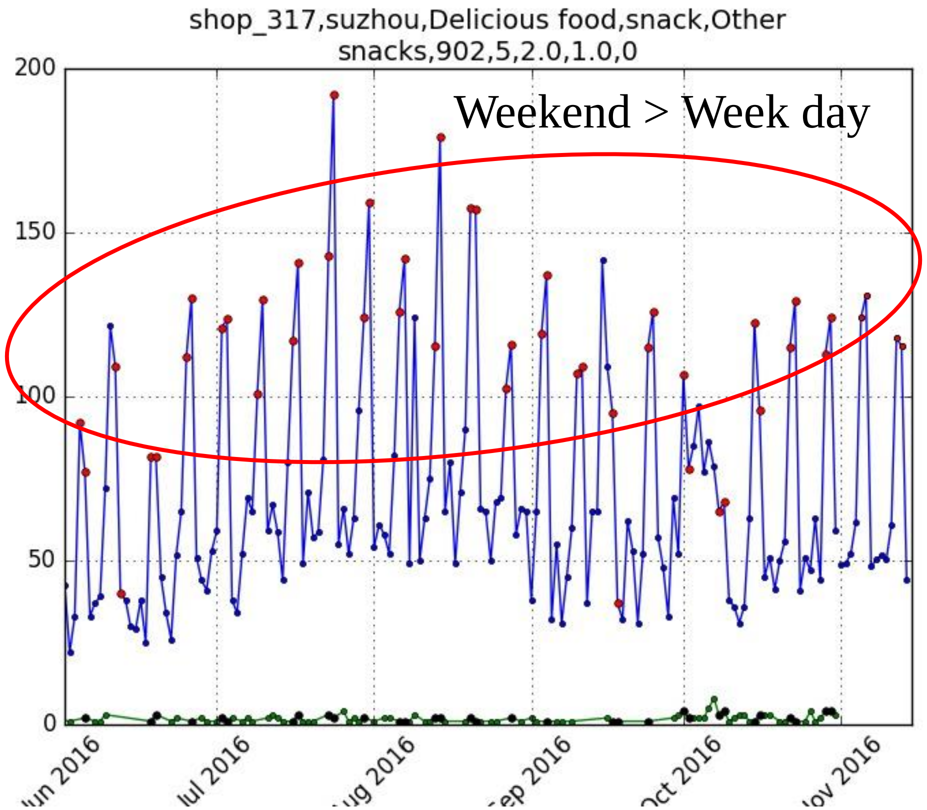


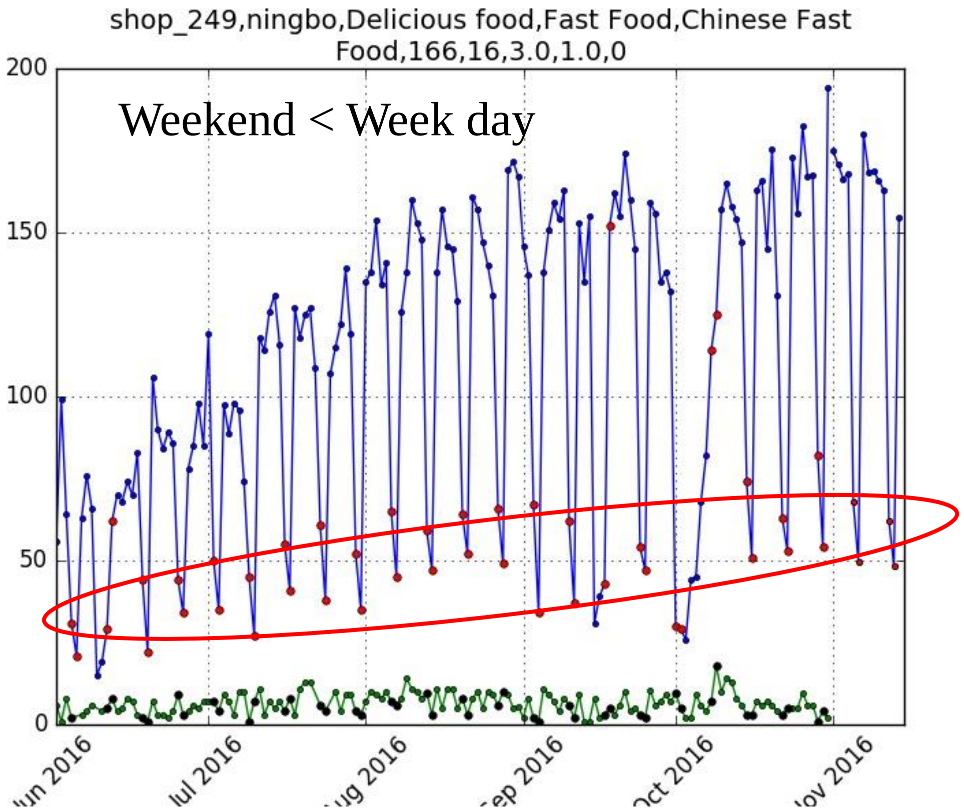
National Holiday

- Weekday: label 0; Weekend: label 1; Other holiday: label 2

Holiday Information

<From official forum of Tianchi Platform>





❑ Weekday: label 0; Weekend: label 1; Other holiday: label 2

Holiday Information

Date	Festival
2015-08-20	Chinese Valentine's Day
2015-09-27	Mid-Autumn Festival
2015-10-01	National day
2015-11-11	Singles Day
2015-12-25	Christmas Day
2016-02-08	Spring festival
2016-02-14	Valentine's Day
2016-04-04	Qing Ming Jie
2016-05-01	Labour Day
2016-06-09	Dragon Boat Festival
2016-08-09	Chinese Valentine's Day
2016-09-15	Mid-Autumn Festival
2016-10-01	National day
2016-11-11	Double 11 Festival

Train set

Test set

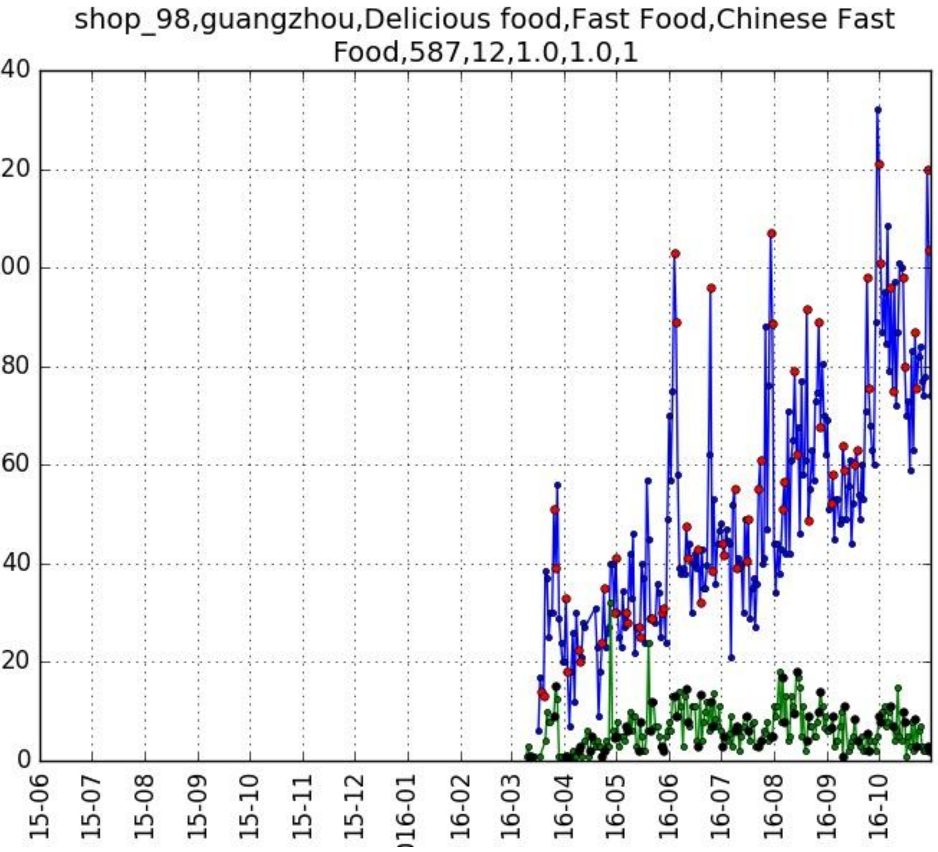
China's largest shopping festival

□ November 11th has become a special festival during recent years. With four characters of "1", this date was named as Double 11 Festival.

Challenges & Difficulties

❑ Cold Start Problem

- Rapid accumulation of new customers & new merchants

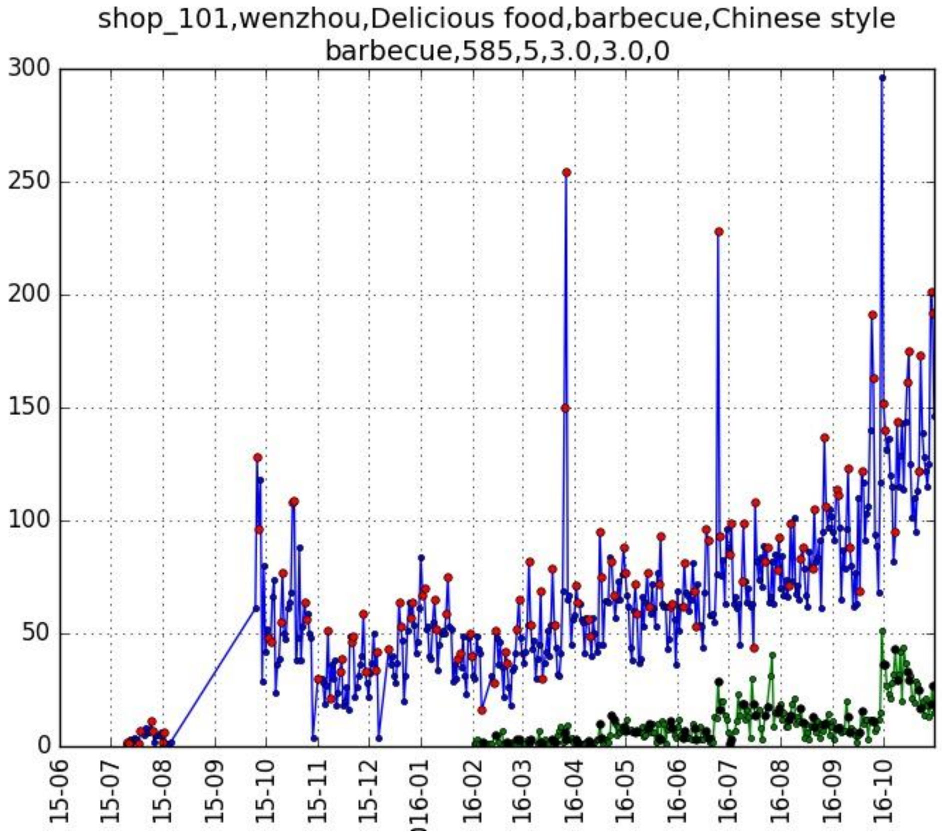


Challenges & Difficulties

❑ Cold Start Problem

- Rapid accumulation of new customers & new merchants

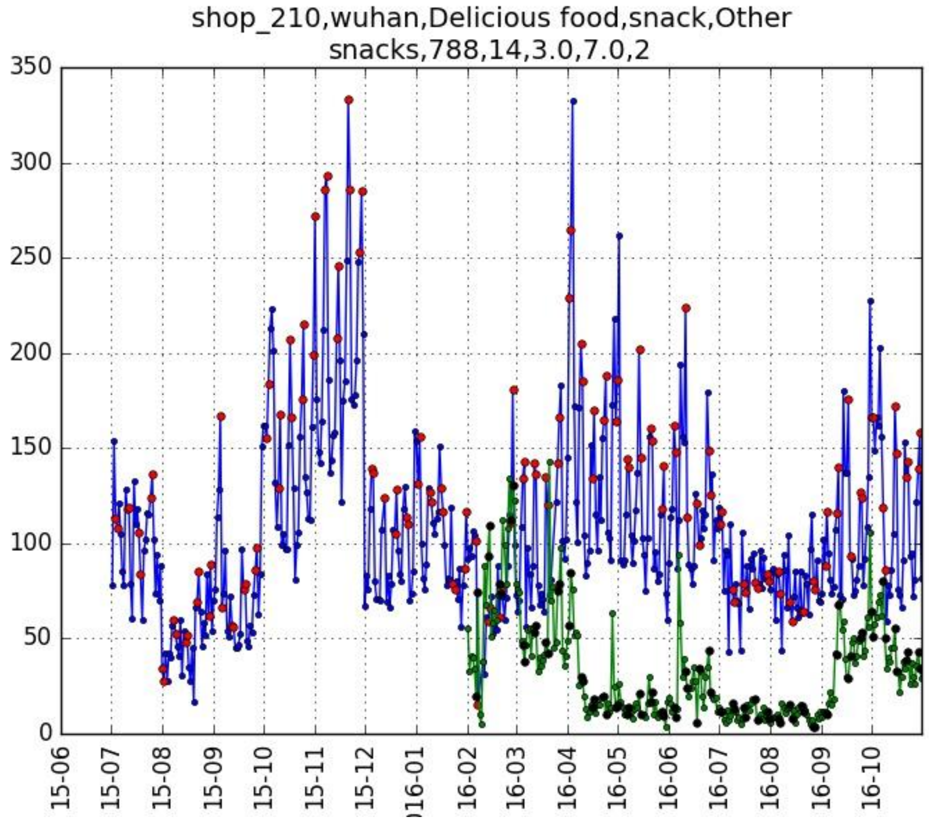
- Unsteady transaction records



Challenges & Difficulties

❑ Cold Start Problem

- Rapid accumulation of new customers & new merchants
- Unsteady transaction records
- Lack of seasonal trends

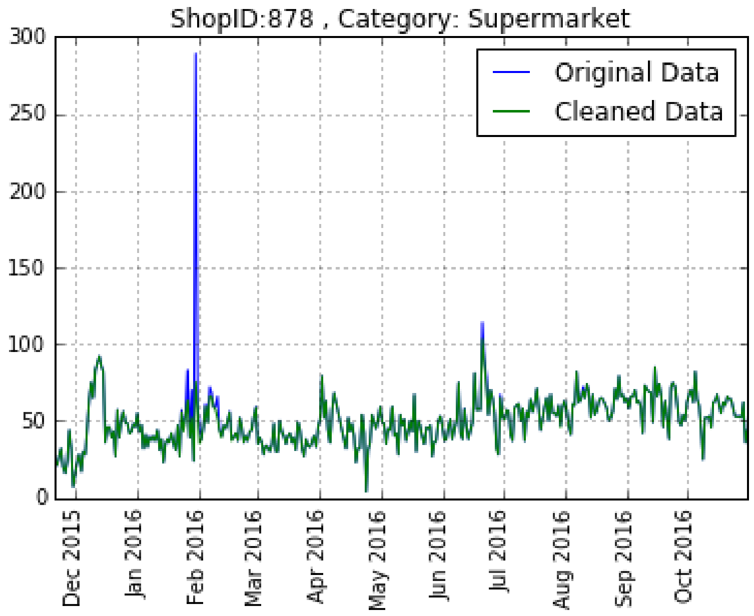


THIRD

Data Processing

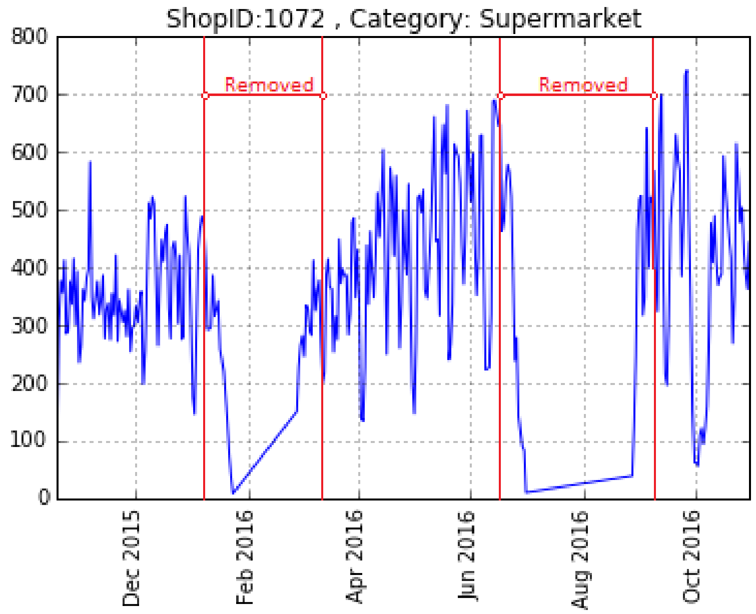
Data Cleaning

❑ Cleaning by Rules

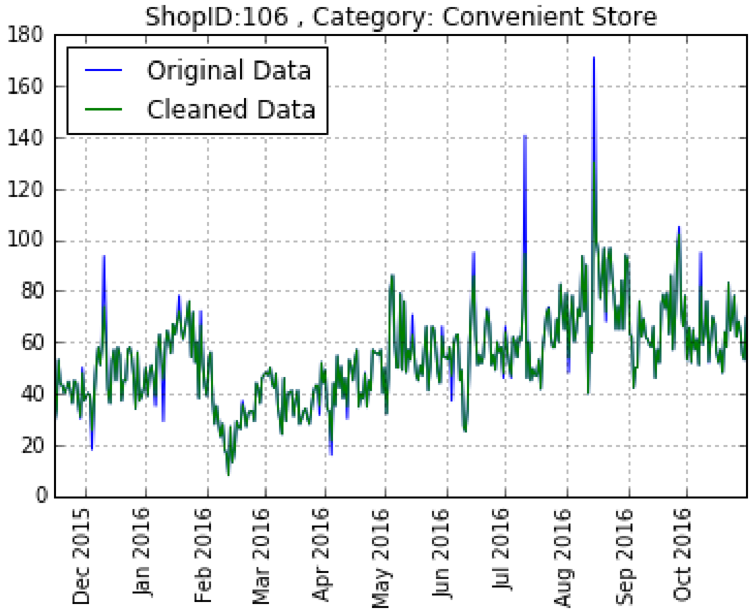


Reduction transformation

$$f(x) = 1 + \log_2 x$$



Removal of abnormal time intervals

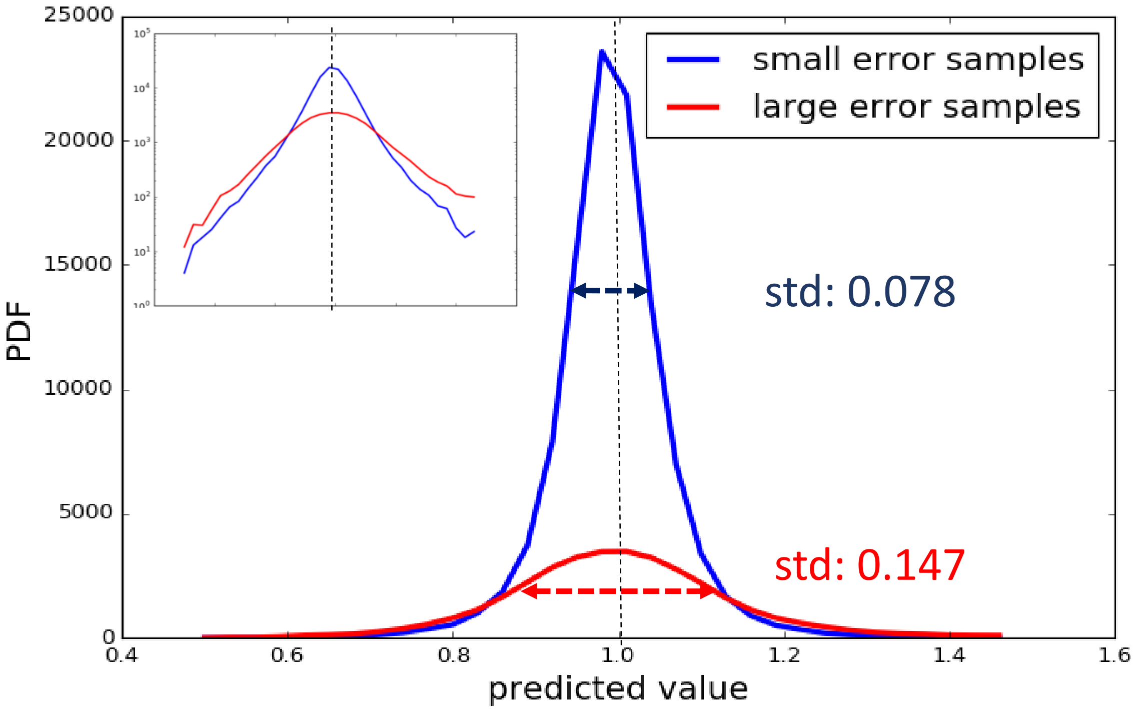


Correction of local abnormal value by $\mu \pm 2\sigma$

Data Cleaning

❑Cleaning by Residuals

- Merchant sales volume can be violated for various reasons, such as promotions, marketing strategy changes…cannot easily cleaned by rules
- Pre-training with high-bias model
- Eliminating samples with top 25% of residual error

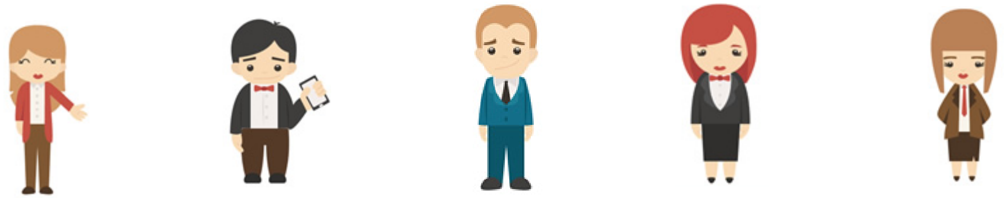


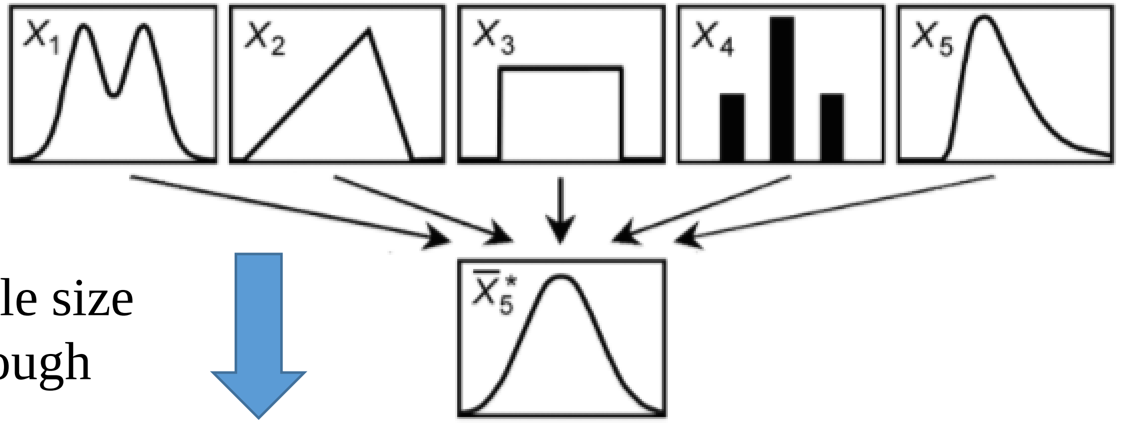
PDF of small & large error samples prediction

Data Consolidation

Central limit theorem: when independent random variables are added, their properly normalized sum tends toward a normal distribution .

purchase tendency of each individual have unique underlying distribution



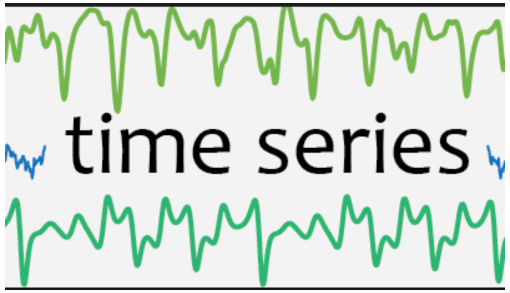


As the sample size get large enough

Sampling distribution becomes normal as the population increases

Data Consolidation

To predict the number of customers for each of the merchant **in a whole day**, the particular purchasing behavior of each individual is beyond the scope of current consideration.



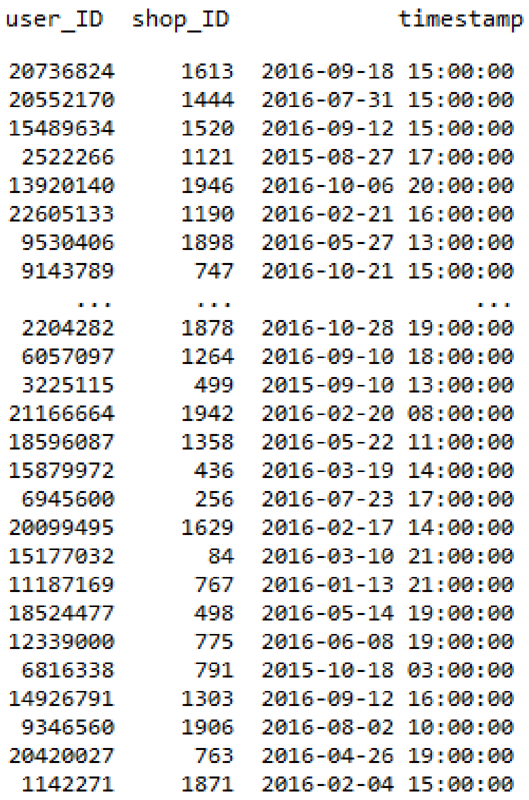
✓

✗

User id

Shop id

Pay timestamp

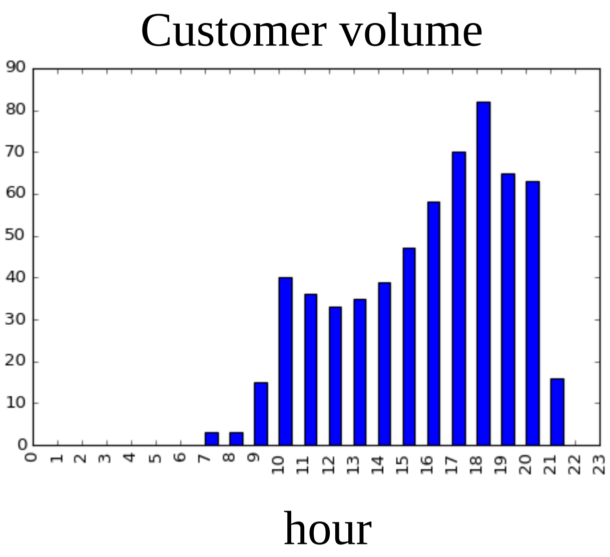


```
user_ID  shop_ID            timestamp

 20736824     1613  2016-09-18 15:00:00
 20552170     1444  2016-07-31 15:00:00
 15489634     1520  2016-09-12 15:00:00
  2522266     1121  2015-08-27 17:00:00
 13920140     1946  2016-10-06 20:00:00
 22605133     1190  2016-02-21 16:00:00
  9530406     1898  2016-05-27 13:00:00
  9143789      747  2016-10-21 15:00:00
      ...      ...                  ...
  2204282     1878  2016-10-28 19:00:00
  6057097     1264  2016-09-10 18:00:00
  3225115      499  2015-09-10 13:00:00
 21166664     1942  2016-02-20 08:00:00
 18596087     1358  2016-05-22 11:00:00
 15879972      436  2016-03-19 14:00:00
  6945600      256  2016-07-23 17:00:00
 20099495     1629  2016-02-17 14:00:00
 15177032       84  2016-03-10 21:00:00
 11187169      767  2016-01-13 21:00:00
 18524477      498  2016-05-14 19:00:00
 12339000      775  2016-06-08 19:00:00
  6816338      791  2015-10-18 03:00:00
 14926791     1303  2016-09-12 16:00:00
  9346560     1906  2016-08-02 10:00:00
 20420027      763  2016-04-26 19:00:00
  1142271     1871  2016-02-04 15:00:00
```

Hourly sales volume

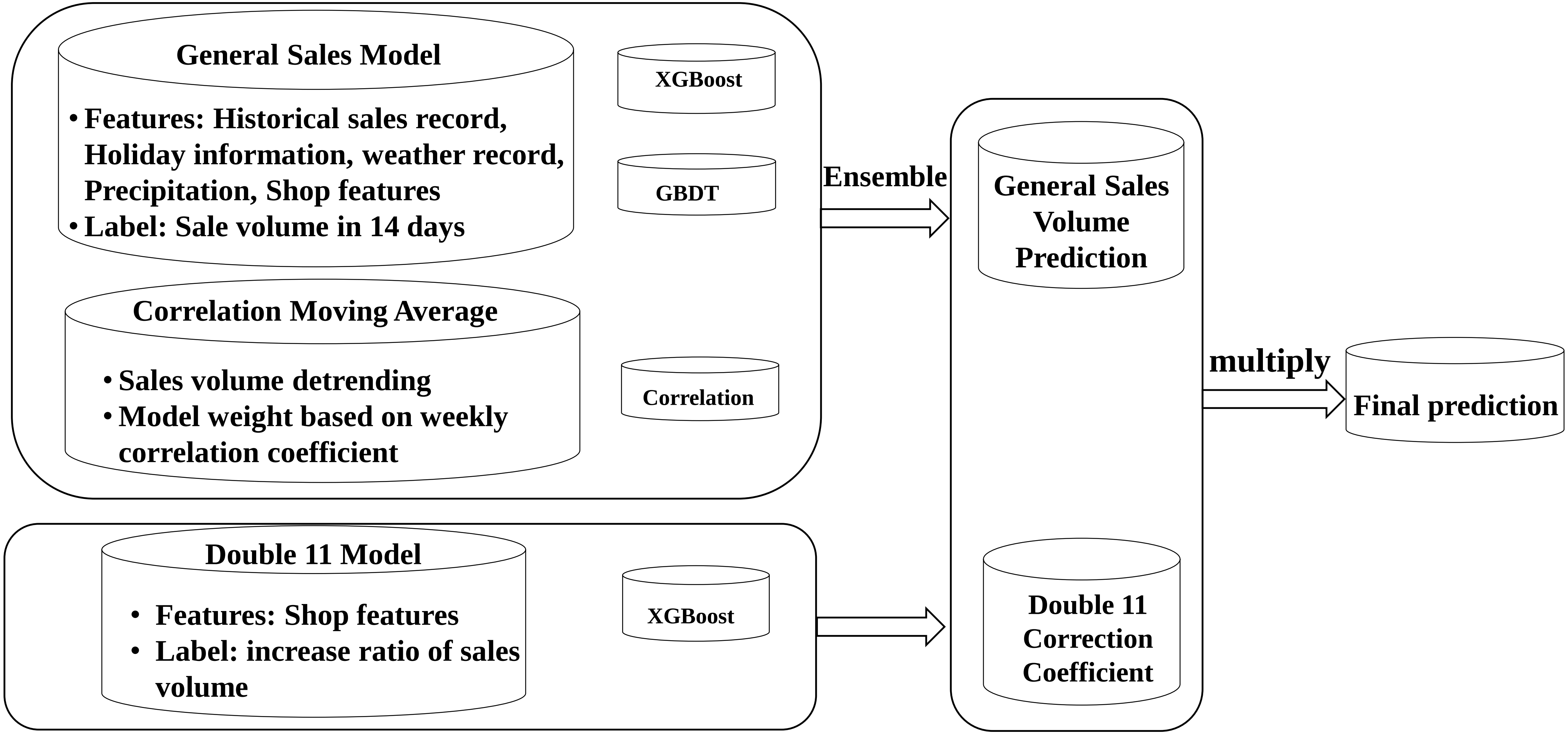
Hour stamp



FOURTH

Predictive Models

Pipeline



General Sales Model

- Features: Historical sales record, Holiday information, weather record, Precipitation, Shop features
- Label: Sale volume in 14 days

XGBoost

GBDT

Correlation Moving Average

- Sales volume detrending
- Model weight based on weekly correlation coefficient

Correlation

Ensemble

General Sales Volume Prediction

Double 11 Model

- Features: Shop features
- Label: increase ratio of sales volume

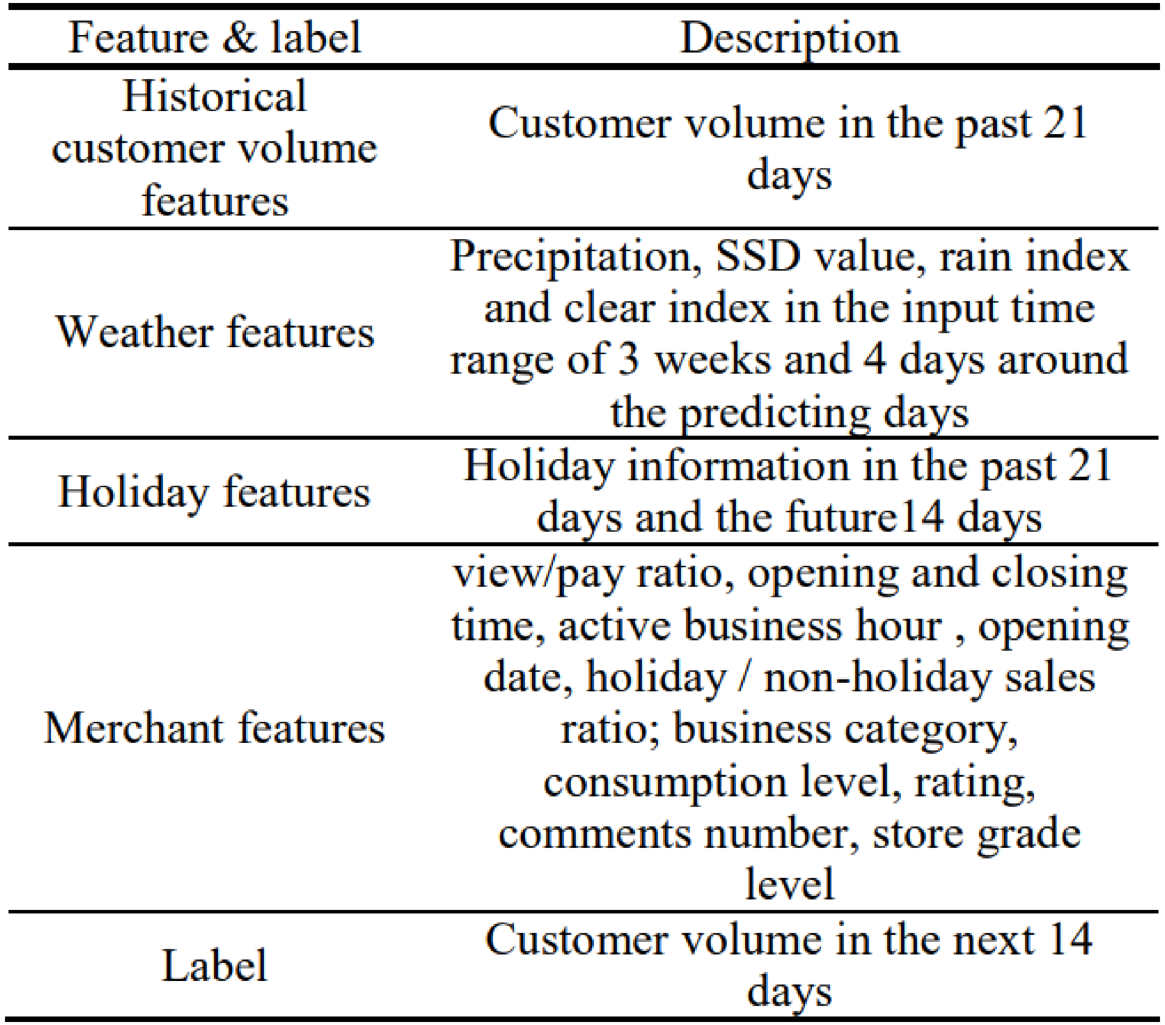
XGBoost

Double 11 Correction Coefficient

multiply

Final prediction

General Sales Model



Feature & label	Description
Historical customer volume features	Customer volume in the past 21 days
Weather features	Precipitation, SSD value, rain index and clear index in the input time range of 3 weeks and 4 days around the predicting days
Holiday features	Holiday information in the past 21 days and the future14 days
Merchant features	view/pay ratio, opening and closing time, active business hour , opening date, holiday / non-holiday sales ratio; business category, consumption level, rating, comments number, store grade level
Label	Customer volume in the next 14 days

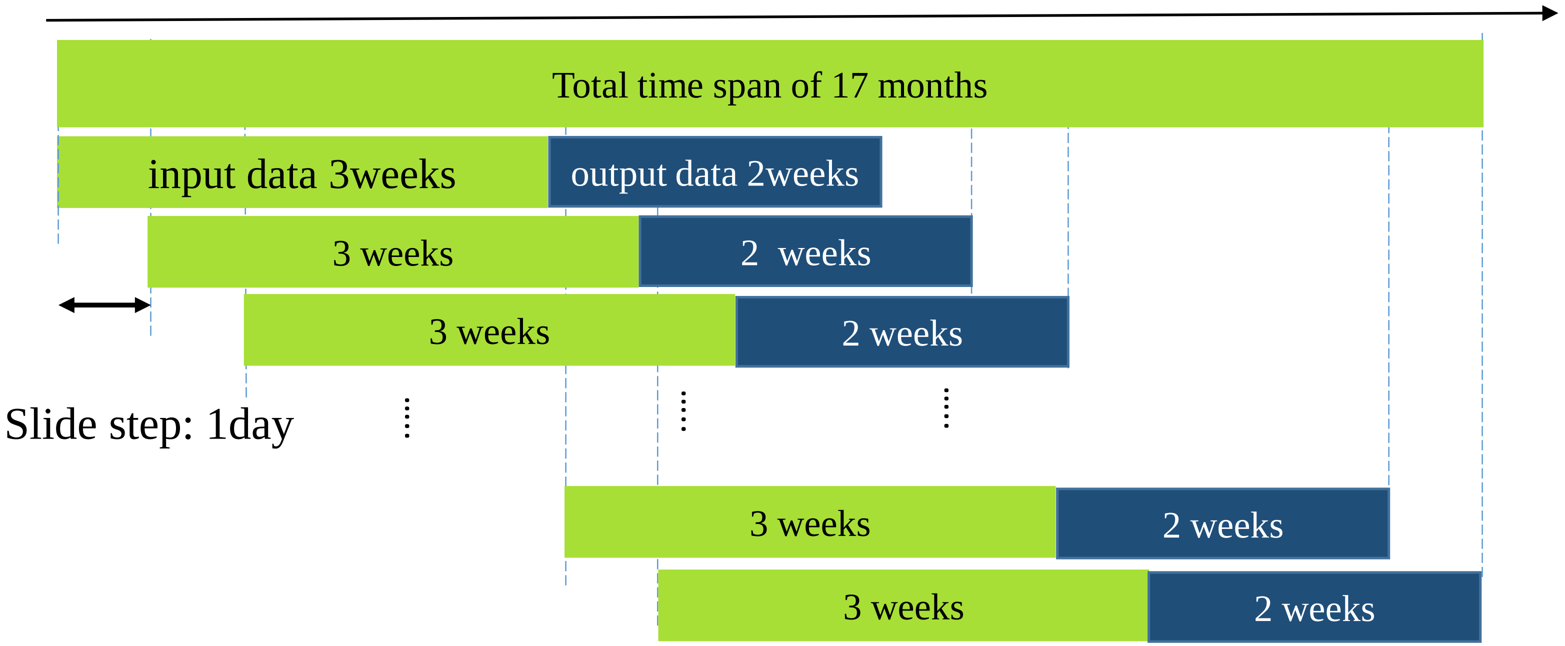
Features

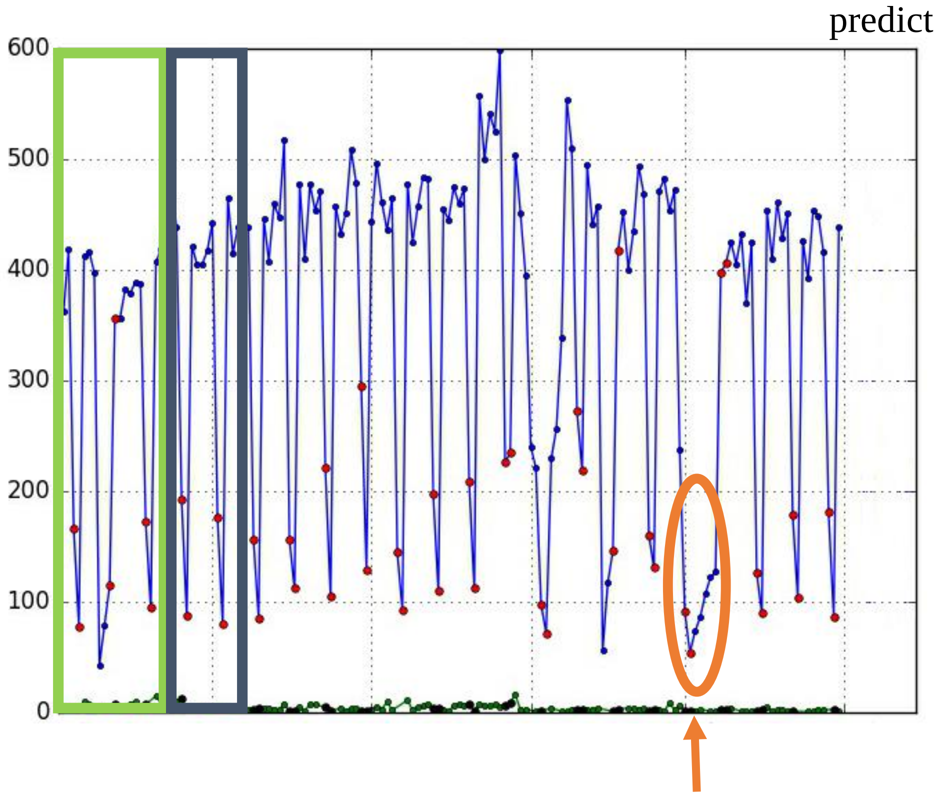
- Historical features
- Weather
- Holiday
- Shop information

Labels

- Sales volume in the next 14 days

Rolling window procedure

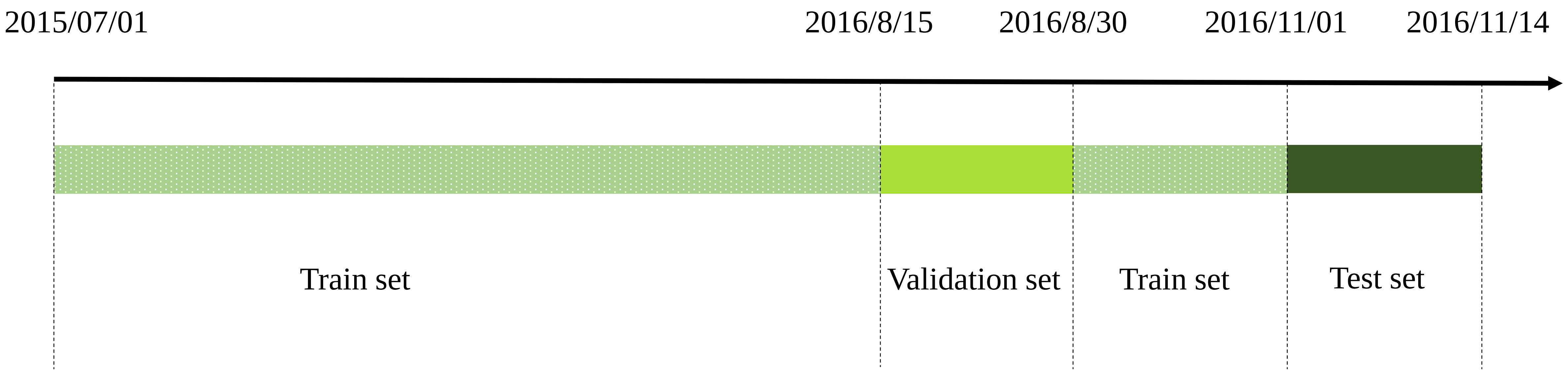




National day
golden week holiday

- Rolling windows technique is adopted along the historical timespan to accumulate training samples. Each sample is composed of the input interval of 3 weeks and the output interval of 2 weeks.
- The input length of 3 weeks could avoid the disturbance from the National holiday of China during Oct. 1st - 7th.

Data set partition

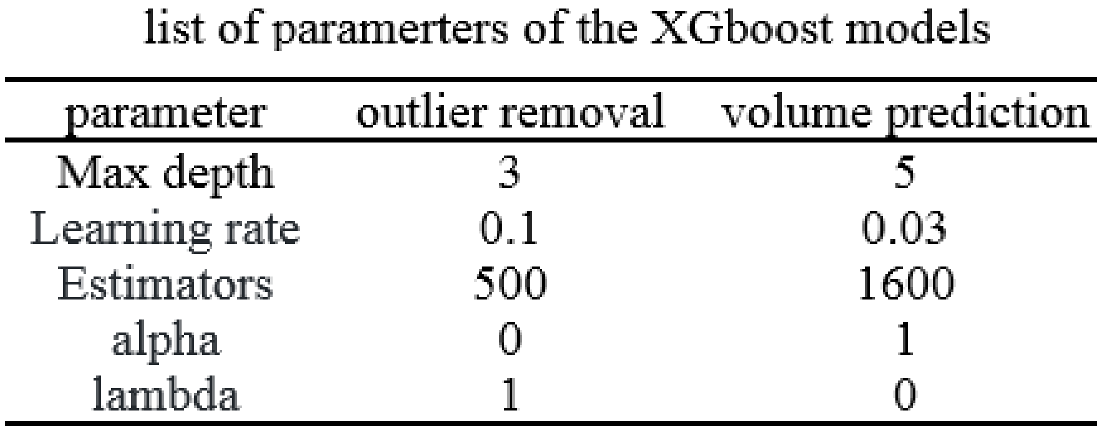


2015/07/01 2016/8/15 2016/8/30 2016/11/01 2016/11/14

Train set Validation set Train set Test set

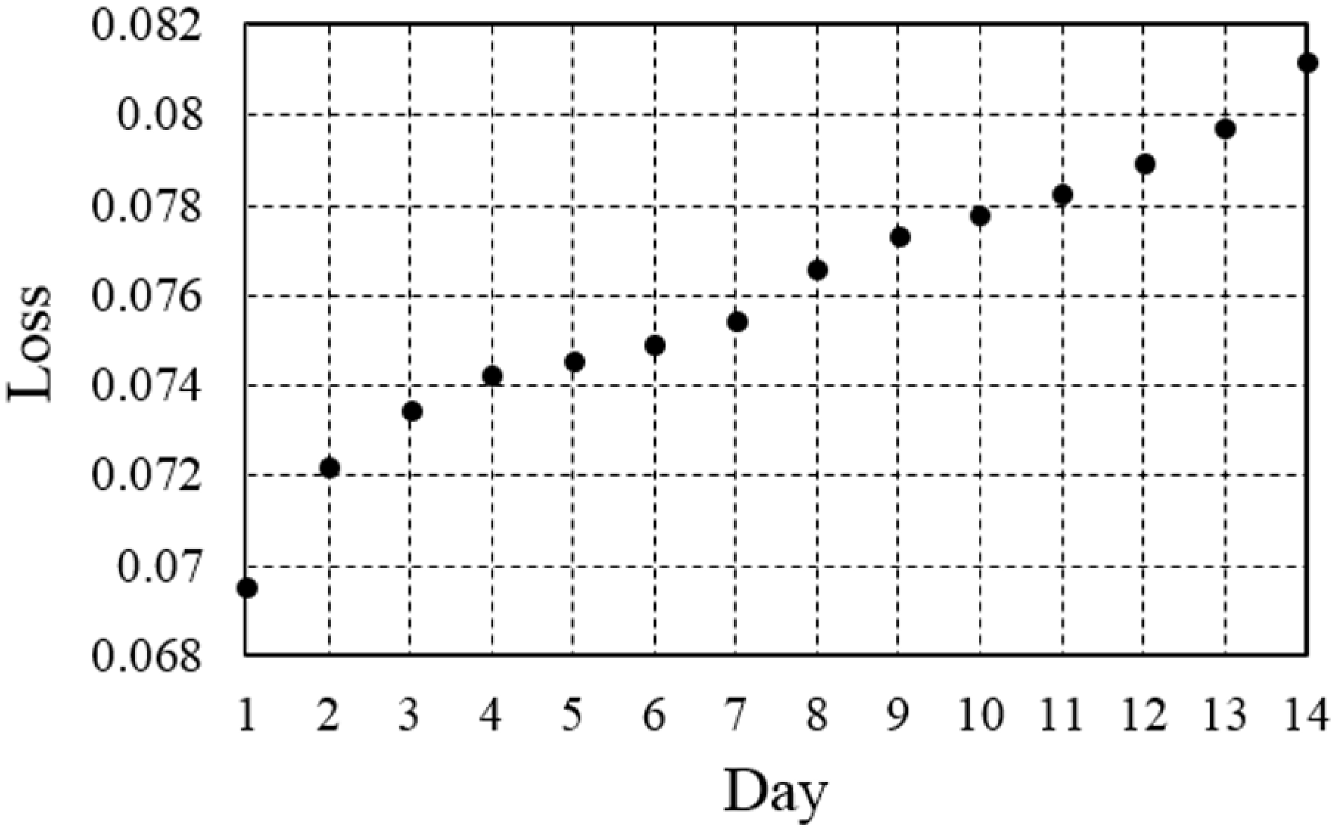
Less disturbance from holidays

General Sales Model

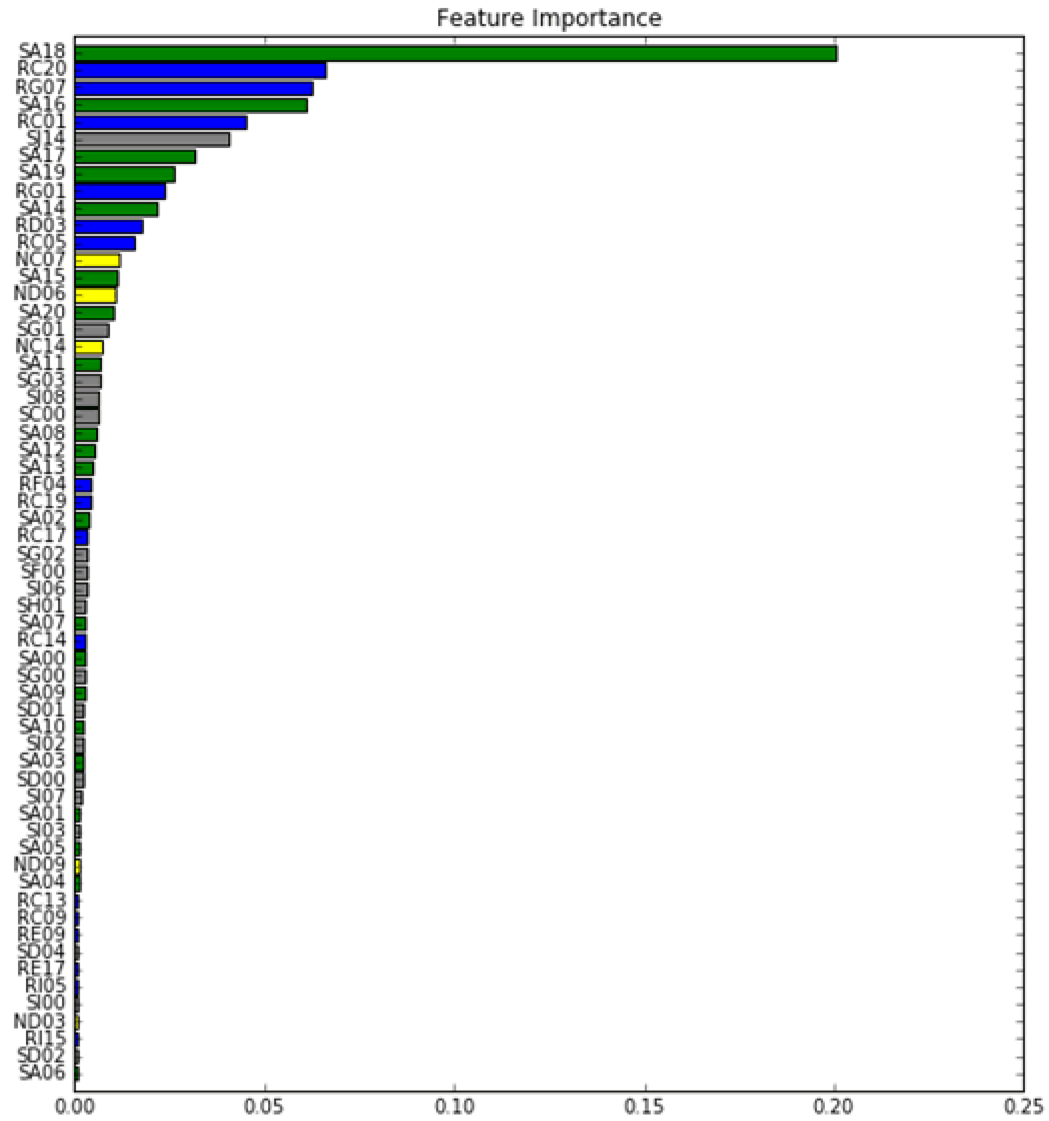


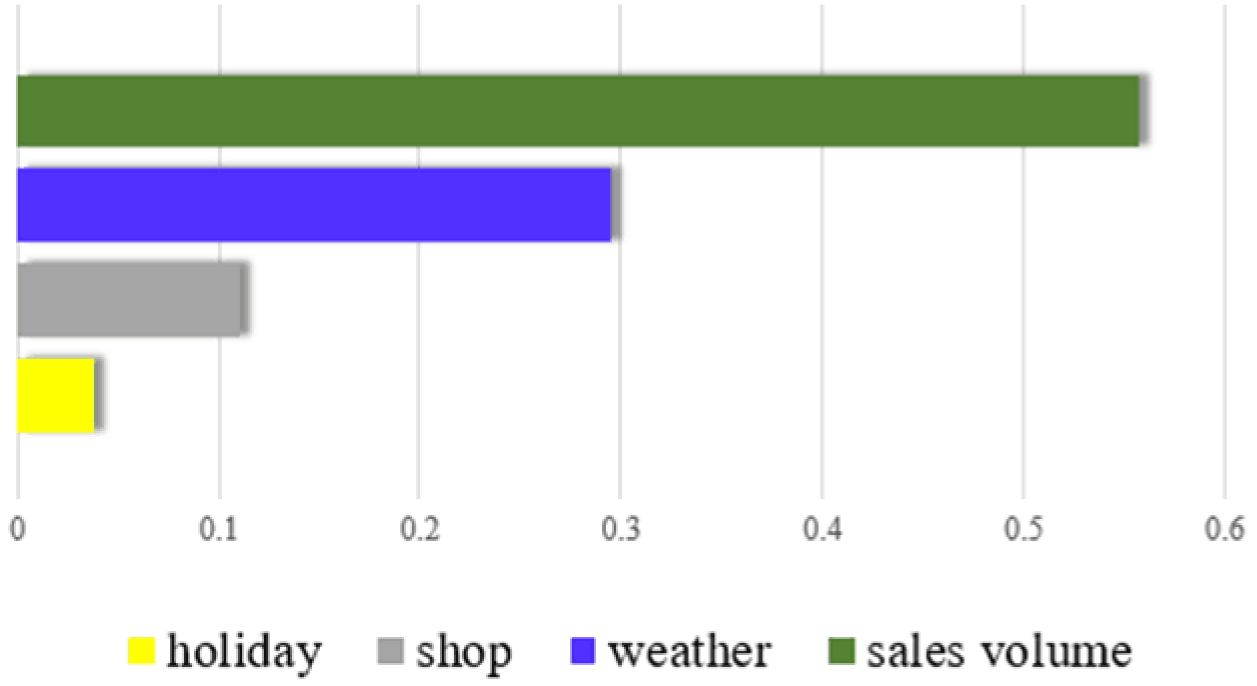
list of paramerters of the XGboost models

parameter	outlier removal	volume prediction
Max depth	3	5
Learning rate	0.1	0.03
Estimators	500	1600
alpha	0	1
lambda	1	0

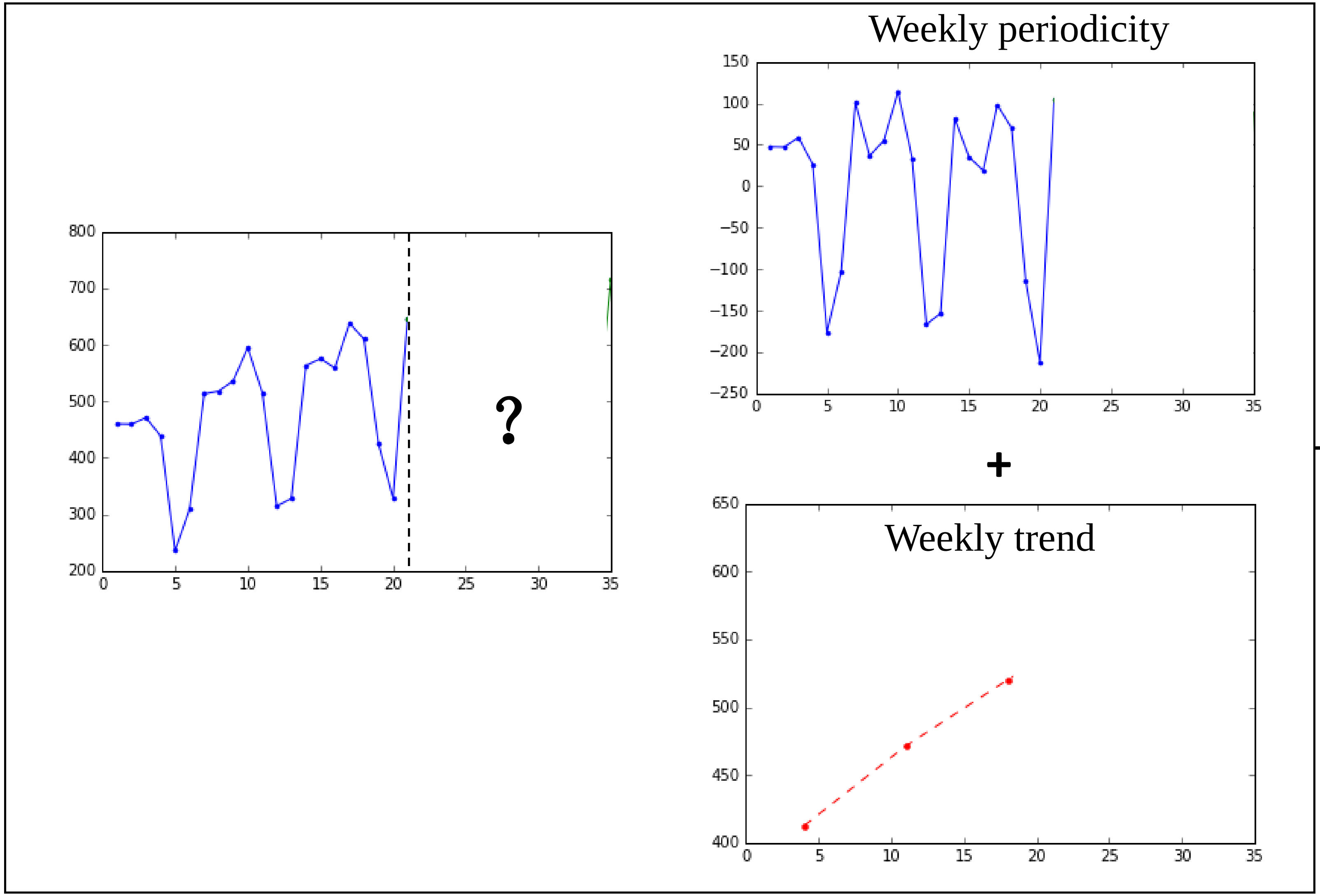


General Sales Model

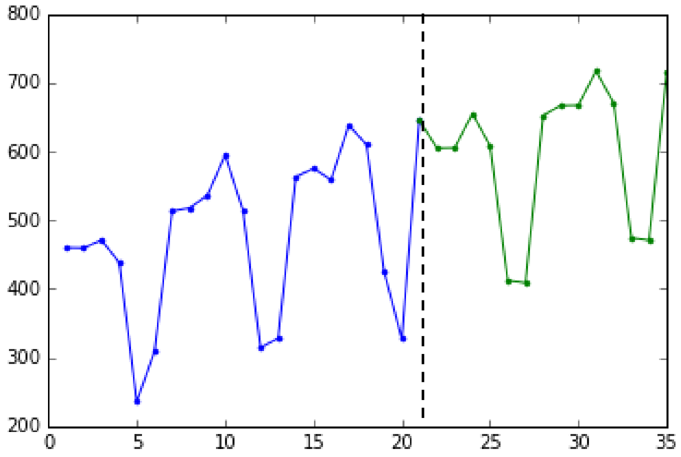




Correlation Moving Average Model



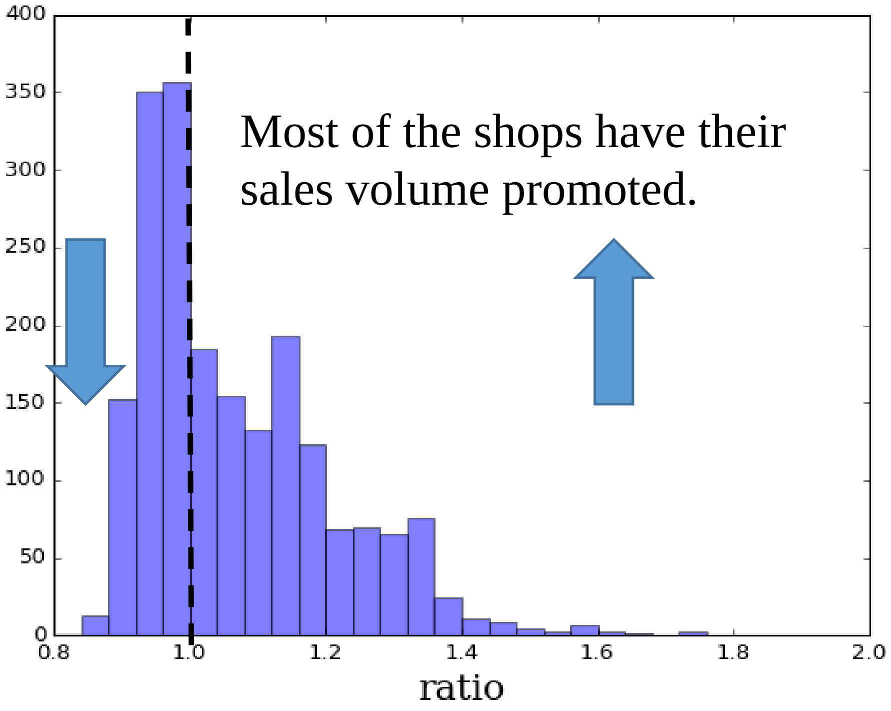
□ Weekly periodicity: based on sales correlation coefficient - a measurement of likelihood that if the historical sales pattern would occur in the future.



□ Ensemble weight of this model is proportional to the weekly correlation coefficient.

Double 11 Correction Coefficient

- □ Due to the lack of historical data, only about 1/3 of the merchants have sales record on the Double 11 in 2015.
- □ Predict the Double 11 correction coefficient for the rest 2/3 of merchants based on the features of merchant information.



Most of the shops have their sales volume promoted.

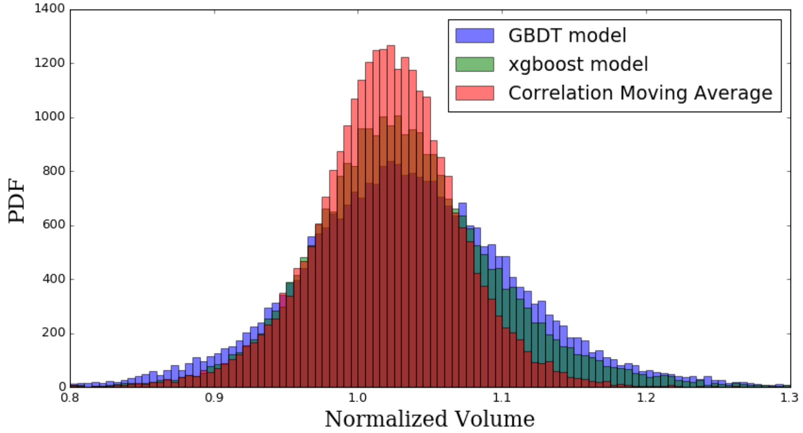
Shop features

- ◆ View/Pay ratio, opening time, closing time, business time, first opening date, median of non-holiday sales volume, median of holiday sales volume, shop category, consumption per person, score, comment number, shop level

Label

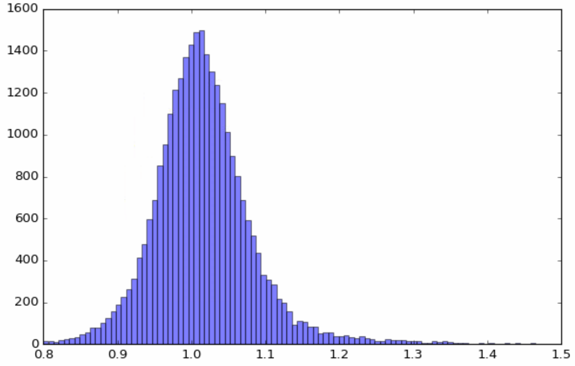
- ◆ Sales increase on historical Double 11 Festival

Model Ensemble

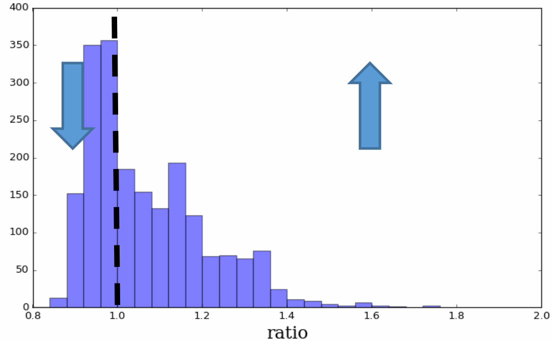


Ensemble

General Sales Volume Prediction



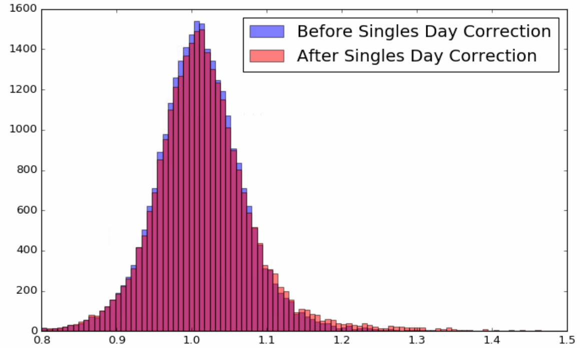
Double 11 Correction Coefficient



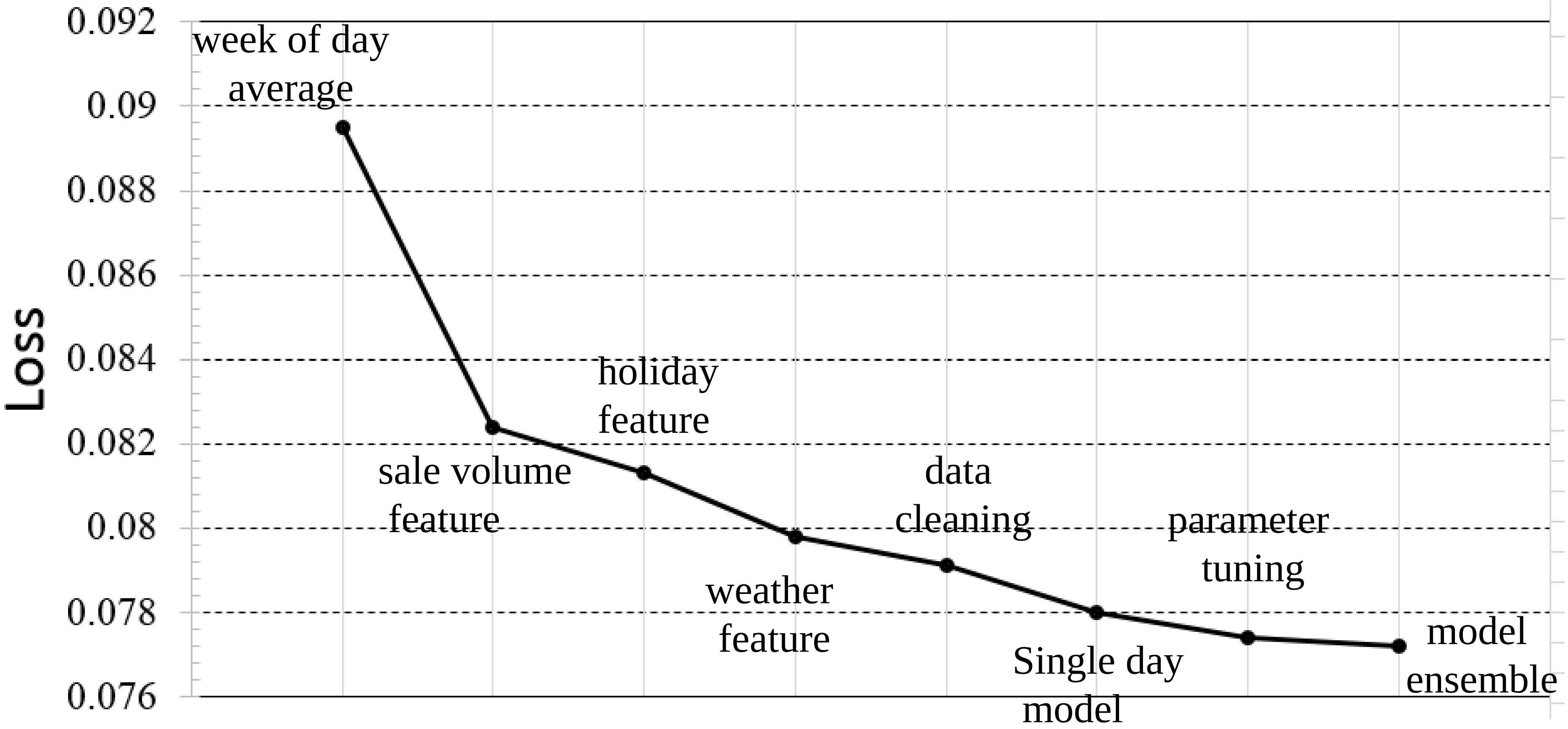
Double 11 Correlation Model

Multiply

Festival modified sales volume



Results


Loss
0.092
0.09
0.088
0.086
0.084
0.082
0.08
0.078
0.076
week of day average
sale volume feature
holiday feature
weather feature
data cleaning
Single day model
parameter tuning
model ensemble

FIFTH

Conclusions

Conclusions & Further Work

Conclusions

- □ Outliers are removed based on both empirical rules and the model residuals.
- □ General customer volume is predicted using the GBDT algorithm with a multiplication modification on the Double 11 Festival.
- □ Facilitate an improved understanding of the potential factors that may influence the customer flow, which will help merchants optimize their operations, reduce cost and improve user experience based on the forecast result.

Future work

- □ Sequential information among the time series records should be taken into considering, to discover the subtle local structure patterns that might influence the customer flow afterwards.
- □ The relative short input time span in the rolling window fail to capture long term tendency.

Codes & Solution Reports

https://github.com/Jessicamidi/IJCAI17_Tianchi_Rank4

THANKS

Acknowledgement:

TIANCHI天池 蚂蚁金服 ANT FINANCIAL 口碑 Koubei IJCAI-17