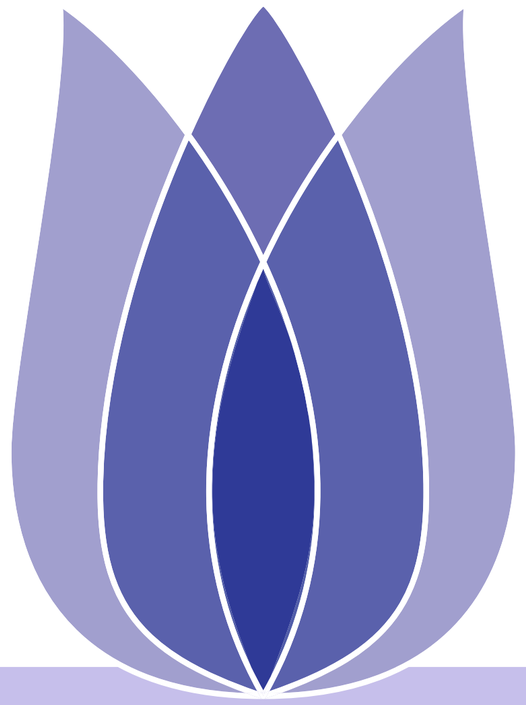
Flip00 Presentation

Huanhuan Ge

Qingdao University of Technology

2022-04-21



Overview

Problem

Description and Evaluation

Description	Predict Future Sales by giving a time-series dataset consisting of daily sales data.
Evaluation	Root mean squared error (RMSE). True target values are clipped into [0,20] range.

Data Processing

Basic Information of Data

Table 1: Data

Name	Description	Attribute
sales_train.csv	Training set(data from January 2013 to October 2015)	date,date_block_num,shop_id,item_id, item_price,item_cnt_day
test.csv	Test set(Predict sale in November2015)	ID,shop_id,item_id
items.csv	Supplementary information of products	item_name,item_id,item_category_id
shops.csv	Supplementary information of shops	shops_name,shops_id
item_categories.csv	Supplementary information of item categories	item_categories_name,item_categories_id
sample_submission.csv	Format of submission	ID,item_cnt_month

- There are 2935849 lines in train set.

  There are 214200 lines in test set.

- There are 21807 unique items in train set.

  There are 60 unique shops in train set.

  There are 5100 unique items in test set.

  There are 42 unique shops in test set.

Missing Value and NaN Value


```
-----missing value-----
date                0
date_block_num      0
shop_id             0
item_id             0
item_price          0
item_cnt_day        0
dtype: int64
--------nan value------
date                0
date_block_num      0
shop_id             0
item_id             0
item_price          0
item_cnt_day        0
dtype: int64
```

Figure 1: Missing Value and NaN Value

Outliers and Duplicate Data

- There are 2935849 lines in train set.

  There are 214200 lines in test set.

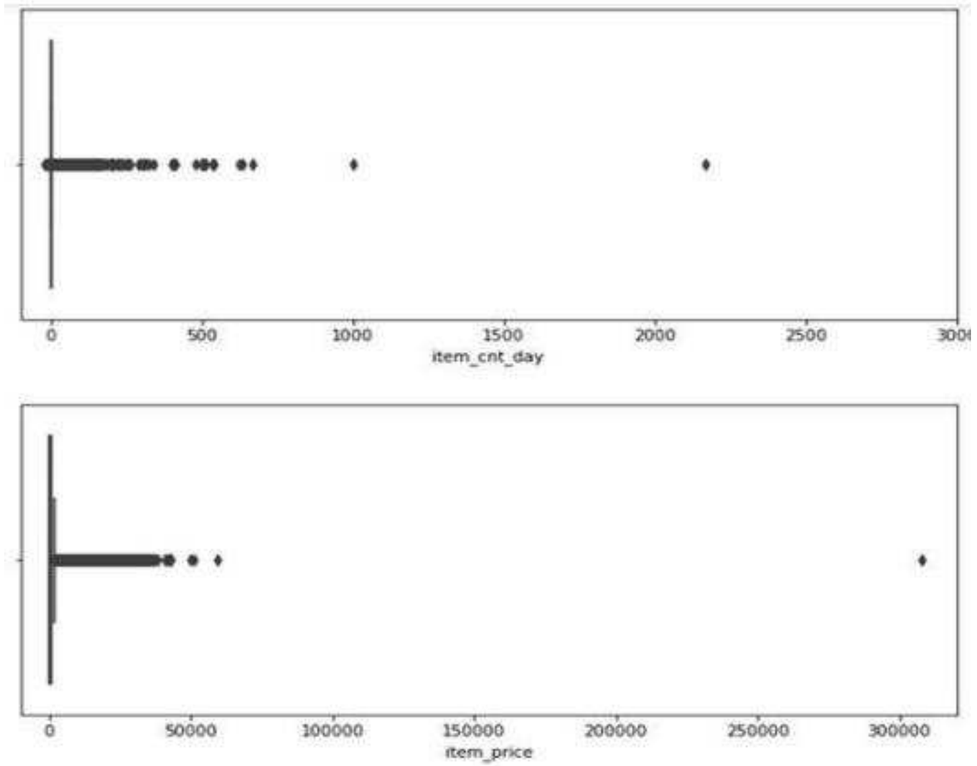


Figure 2: Outliers Data

Process Shops Set

- Same shop name, different shop ID
  39 and 40,10 and 11,0 and 57, 58 and 1
- Modify the ID based on the test
- Shop full name: shop's city-shop's type-shop's name
- Encode shops information

	shop_name	shop_id	shop_city	shop_type	shop_city_code	shop_type_code
0	!Якутск Орджоникидзе, 56 фран	0	Якутск	Others	0	0
1	!Якутск ТЦ "Центральный" фран	1	Якутск	ТЦ	0	1
2	Адыгея ТЦ "Мега"	2	Адыгея	ТЦ	1	1
3	Балашиха ТРК "Октябрь-Киномир"	3	Балашиха	ТРК	2	2
4	Волжский ТЦ "Волга Молл"	4	Волжский	ТЦ	3	1
5	Вологда ТРЦ "Мармелад"	5	Вологда	ТРЦ	4	3
6	Воронеж (Плехановская, 13)	6	Воронеж	Others	5	0

Figure 3: Encode Shops Information

Process Items Set

- Same item name, different item ID

  2514 and 2558,2968 and 2970,5061 and 5063, 14537 and 14539,19465 and 19475,19579 and 19581
- Modify the ID based on the test

Process Categories Set

- Shop full name: category's type-category's subtype.
- Encode categories information

	item_category_name	item_category_id	item_type	item_type_code	sub_type	sub_type_code
0	PC - Гарнитуры/Наушники	0	PC	0	Гарнитуры/Наушники	0
1	Аксессуары - PS2	1	Аксессуары	1	PS2	1
2	Аксессуары - PS3	2	Аксессуары	1	PS3	2
3	Аксессуары - PS4	3	Аксессуары	1	PS4	3
4	Аксессуары - PSP	4	Аксессуары	1	PSP	4

Figure 4: Encode Categories Information

Sales Analysis

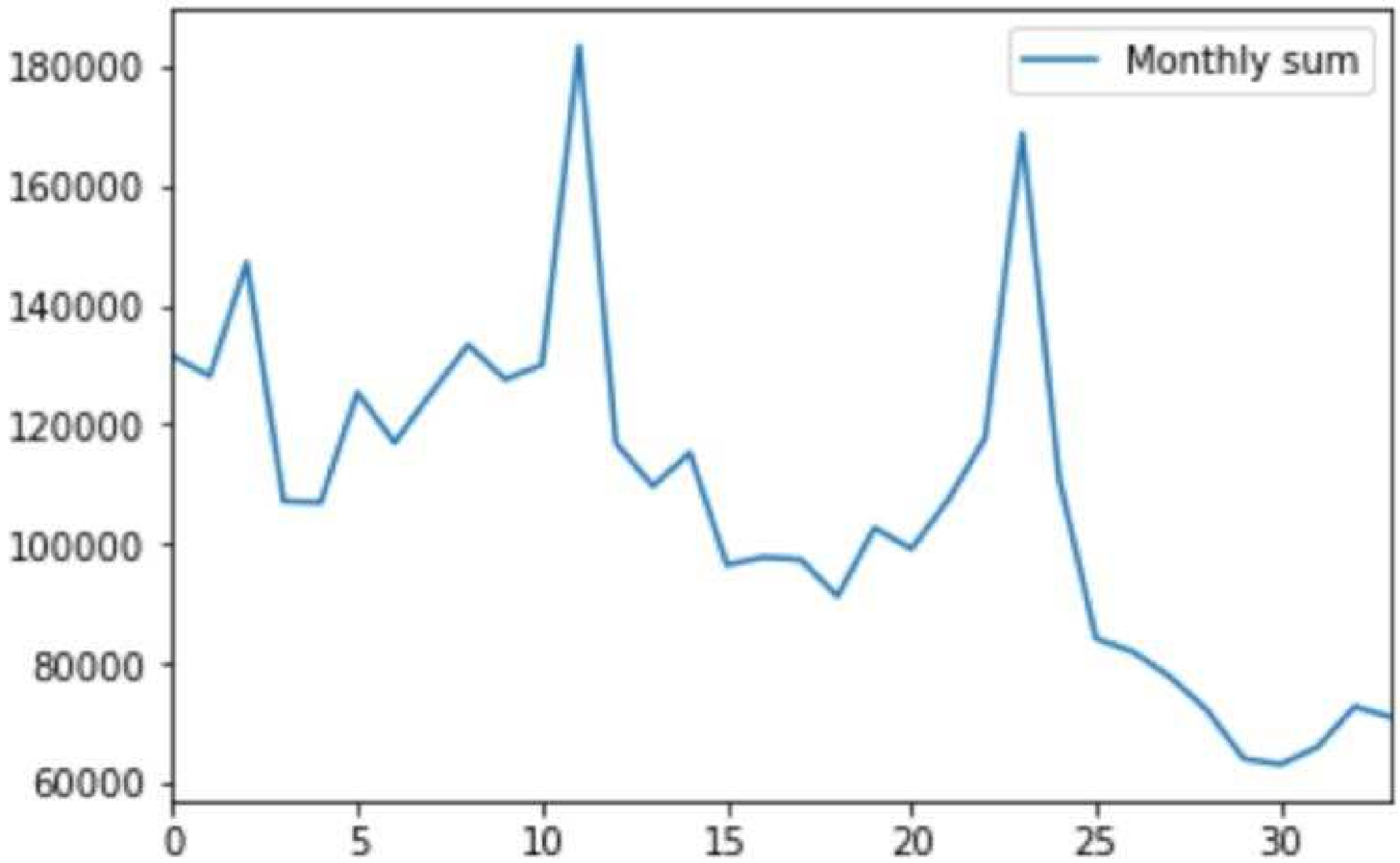


Figure 5: Total Sales Over Time

Closed Shops and Discontinued Products

- new shops:9,20,36
- closed shops:0,1,8,11,13,17,23,27,29,30,32,33,40,43,51,54

item_id	0	1	2	3	4	5	6	7	8	9	...	22150	22151	22152	22156	22157	22160	22161	22165	22168	22169
date_block_num																					
22	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	1	0	1	0	0	...	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Figure 6: Discontinued Products

Feature Selection

Data Feature

- new shops:9,20,36
- closed shops:0,1,8,11,13,17,23,27,29,30,32,33,40,43,51,54

	date_block_num	shop_id	item_id	item_cnt_month	shop_type_code	shop_city_code	item_category_id	item_type_code	sub_type_code
0	0	59	22154	1.0	1	29	37	10	21
1	0	59	2552	0.0	1	29	58	12	41
2	0	59	2554	0.0	1	29	58	12	41
3	0	59	2555	0.0	1	29	56	12	39
4	0	59	2564	0.0	1	29	59	12	42
...	...	...	...	...	...	...	...	...	...
11054935	34	45	18454	0.0	1	21	55	12	38
11054936	34	45	16188	0.0	1	21	64	13	47
11054937	34	45	15757	0.0	1	21	55	12	38
11054938	34	45	19648	0.0	1	21	40	10	24
11054939	34	45	969	0.0	1	21	37	10	21

Figure 7: Data Feature

Monthly Sales Feature

- average monthly sales of items
- average monthly sales of shops
- average monthly sales of categories
- average monthly sales of types and subtypes
- average monthly sales of shop's city-item
- average monthly sales of shop's type-item

Historical Feature

- Historical delay:1,2,3,6,12
- Historical Feature: monthly sales of items

  average monthly sales of shops

  average monthly sales of items

  average monthly sales of categories

  average monthly sales of types and subtypes

  average monthly sales of shop's city-item

  average monthly sales of shop's type-item
- Delete the records in first 12 months and NAN records

Modeling and Forecasting

Feature Engineering

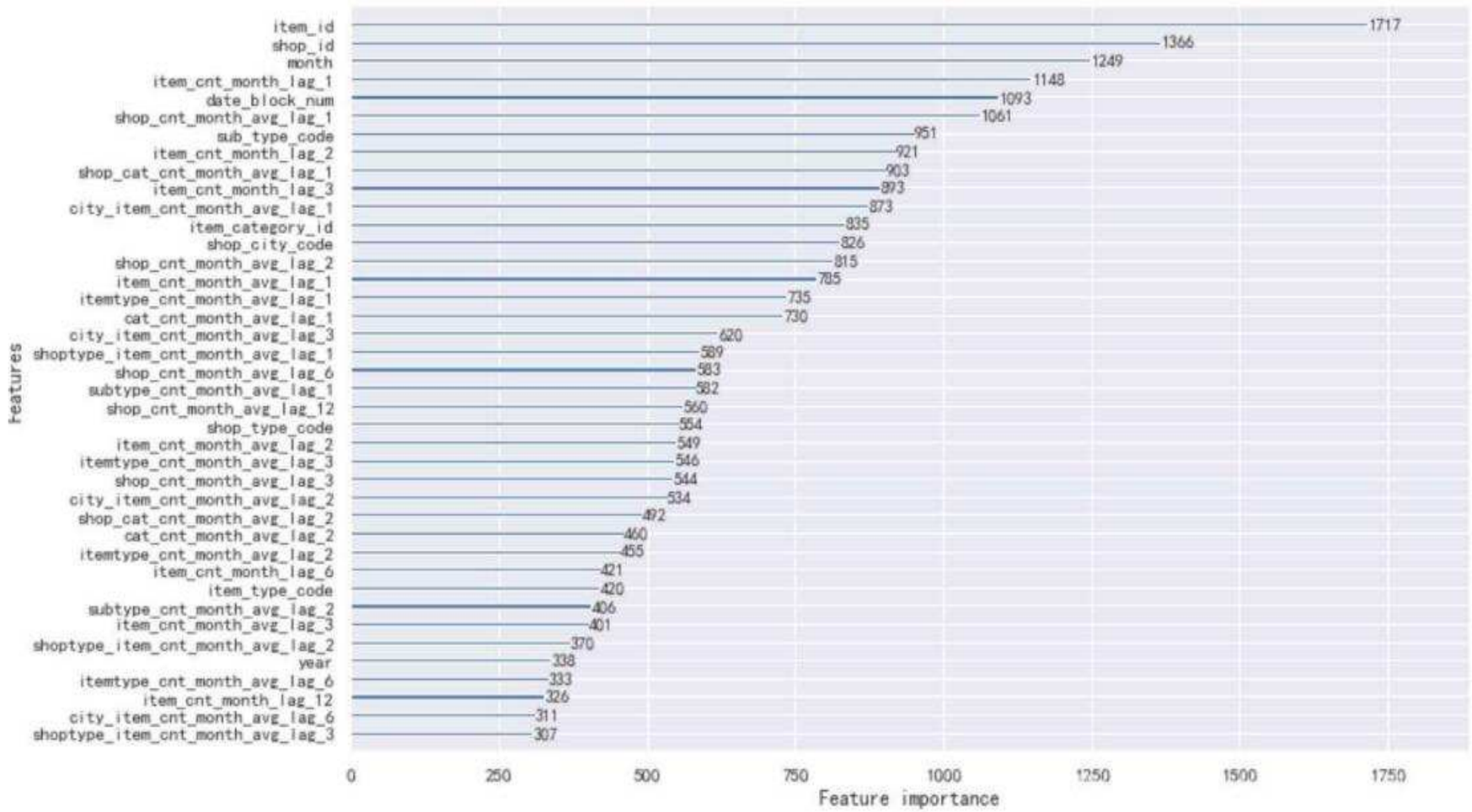


Figure 8: Feature Importance

Lightgbm

- train set:date_block_num< 33
  validation set:date_block_num == 33
  test set:date_block_num == 34
- score:0.93740
- 3027/8738

Comparison

- 1.0485→0.93740
- model and feature LightGBM and XGBoost

  Add feature:historical feature