Predict future sales

Pengcheng Jiang

JiLin University

2021-04-24

Overview

Problem Definition

Data Cleaning

Data analysis

Model

~~Model~~

Lightgbm

Problem Definition

Predict future sales

- ~~given: a challenging time-series dataset consisting of daily sales data, kindly provided by one of the largest Russian software firms - 1C Company.~~
- ~~target: predict total sales for every product and store in the next month~~
- ~~evaluation: Submissions are evaluated by root mean squared error (RMSE)~~

given	a challenging time-series dataset consisting of daily sales data, kindly provided by one of the largest Russian software firms - 1C Company.
target	predict total sales for every product and store in the next month
evaluate	Submissions are evaluated by root mean squared error (RMSE)

Data Cleaning

Date

- ~~item_categories.csv:item_category_name item_category_id~~
- ~~items.csv:item_id item_category_id~~
- ~~sales_train.csv:date date_block_num shop_id item_id item_price item_cnt_day~~
- ~~shops.csv:shop_name shop_id~~
- ~~test.csv:shop_id item_id~~

File	filed1	filed2	filed3	filed4	filed5	filed6
item_categories	item_category_name	item_category_id				
items	item_id	item_category_id				
sales_train	date	date_block_num	shop_id	item_id	item_price	item_cnt_day
shops	shop_name	shop_id				
test	shop_id	item_id				

Table 1: Data Infomation

Data Information

~~sales_train:~~

- ~~2935849 rows,6 columns~~
- ~~21807 items,60 shops~~
- ~~data_type~~
- ~~data: object~~
- ~~date_block_num: int~~
- ~~shop_id:int~~
- ~~item_id:int~~
- ~~item_price:float~~
- ~~item_cnt_day:float~~

sales_train	■ 2935849 rows,6 columns ■ 21807 items,60 shops ■ data_type ◆ data: object ◆ date_block_num: int ◆ shop_id:int ◆ item_id:int ◆ item_price:float

Data Information

~~test:~~

- ~~214200 rows,3 columns~~
- ~~5100 items,40 shops~~
- ~~data_type~~
- ~~ID:int~~
- ~~shop_id:int~~
- ~~item_id:int~~

~~From here you can see a lot of stores, goods in training set are not in the test set~~

test	■ 214200 rows,3 columns ■ 5100 items,40 shops ■ data_type ◆ ID:int ◆ shop_id:int ◆ item_id:int From here you can see a lot of stores, goods in training set are not in the test set

Missing Value and Non Value

~~target:Find out whether there are empty values or missing values in the dataresult:missing value:0nan~~

~~value:0~~ target	Find out whether there are empty values or missing values in the data
result	■ missing value:0 ■ nan value:0

Cartesian product

reason	The training set contains only the items that the store actually sold that month
target	for items not sold during the month, you should add them and set them to 0(Find out all the stores and merchandise, and make cartesian product with sales_trainz)

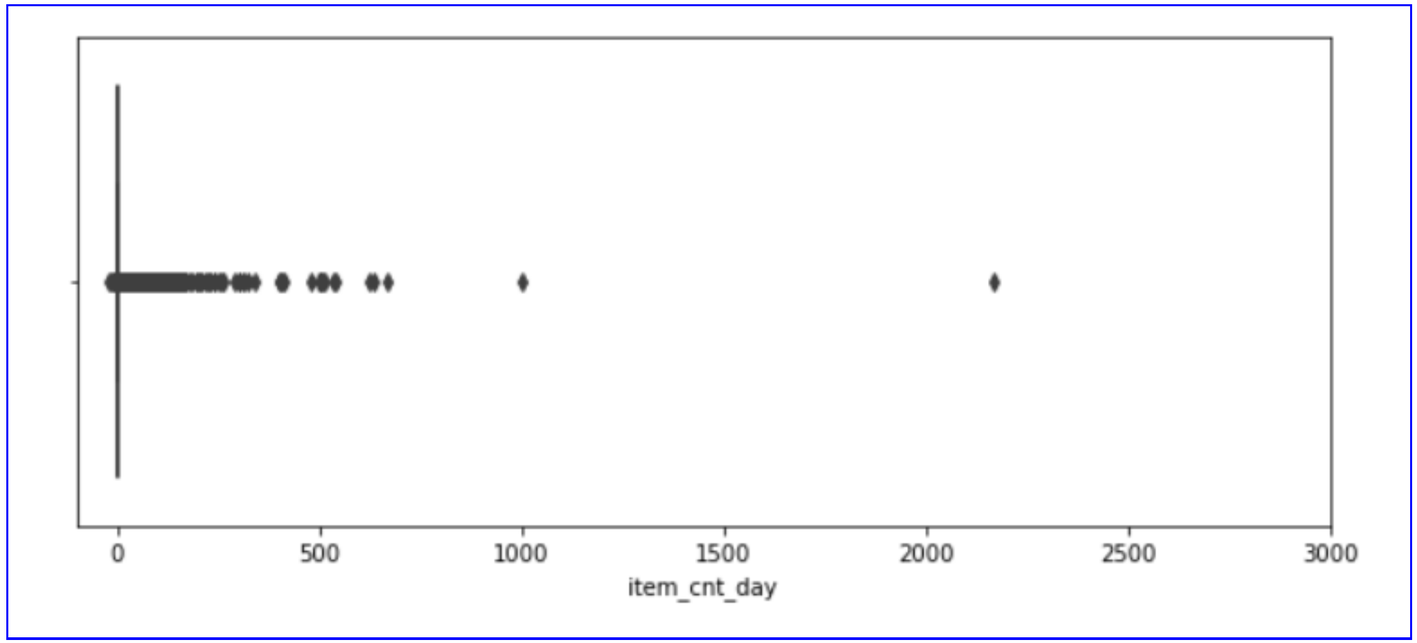
Data leakages

target	delete stores, goods in training set but not in the test set
result	■ sales_train: ■ rows:1224439 ■ items:4716 ■ shops:42

Data duplication

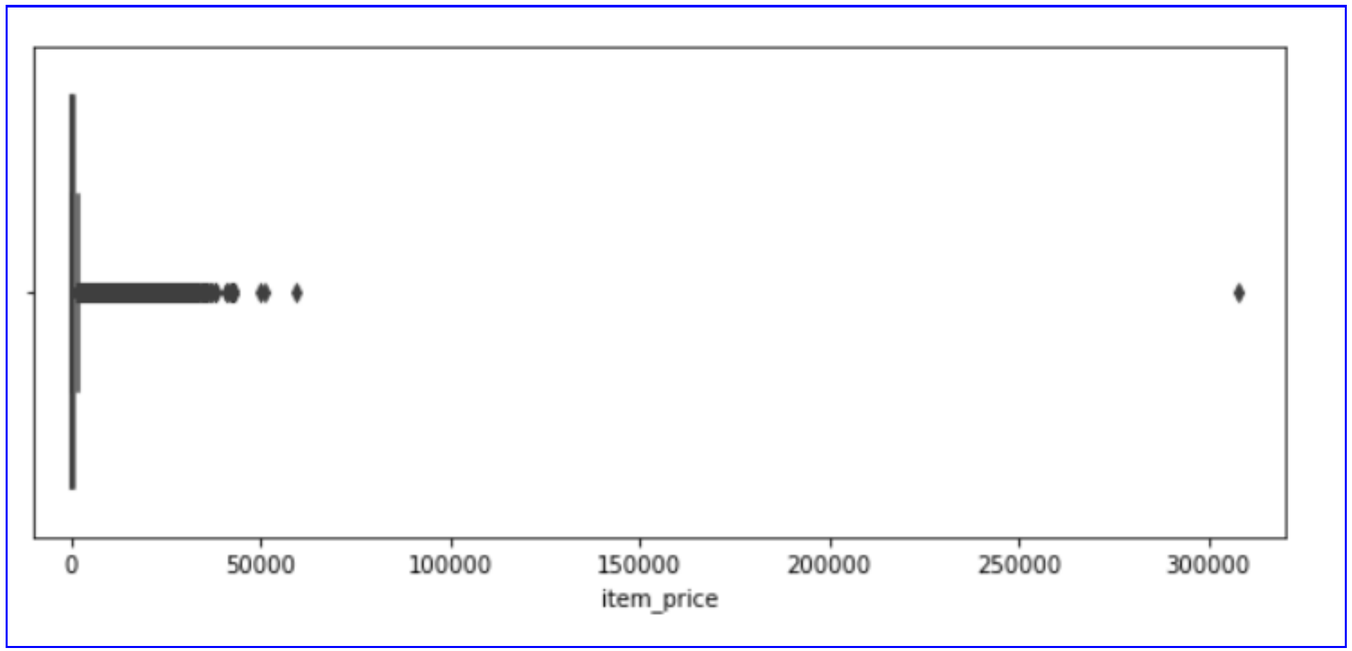
target	See if duplicate items exist in the dataset
result	■ sales_train:6 ■ test:0

Outliers

target	Calculate the outliers of item_cnt_day and item price
result	

Outliers

result



outdated items and Negative

target	Analyze how many products have not been sold in the last six consecutive months. How many of these products appear in the test set.
result	There are 12391 training sets, which have not been sold in the last six months. There are 164 test sets, which have not been sold in the last six months

Negative	Change item whose commodity price is negative to median

outdated items and Negative

Negative	Change item whose commodity price is negative to median

Data analysis

Monthly sales of goods

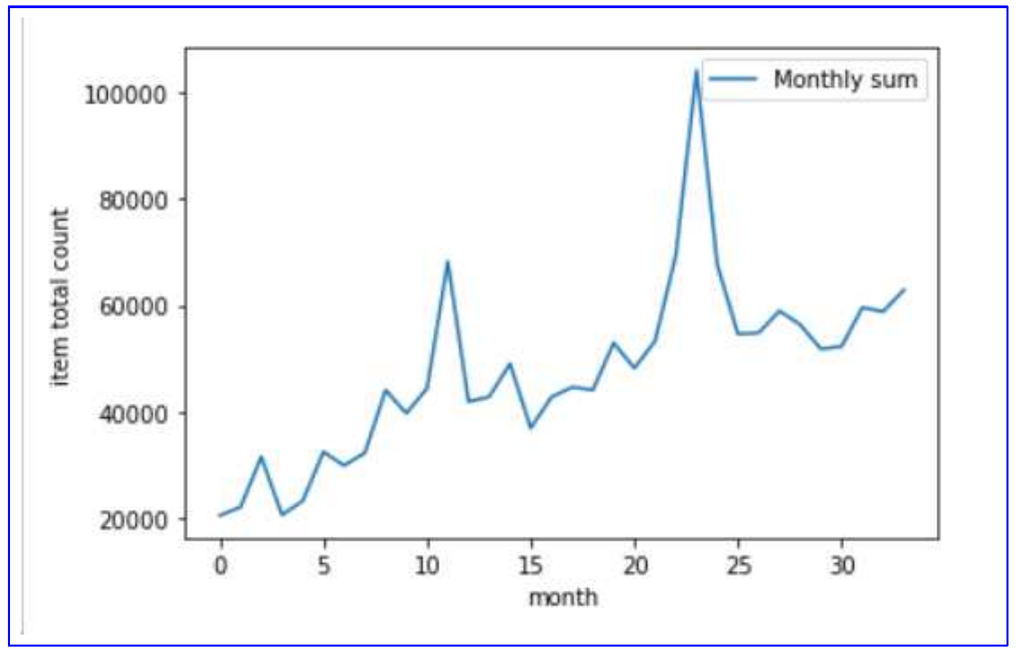
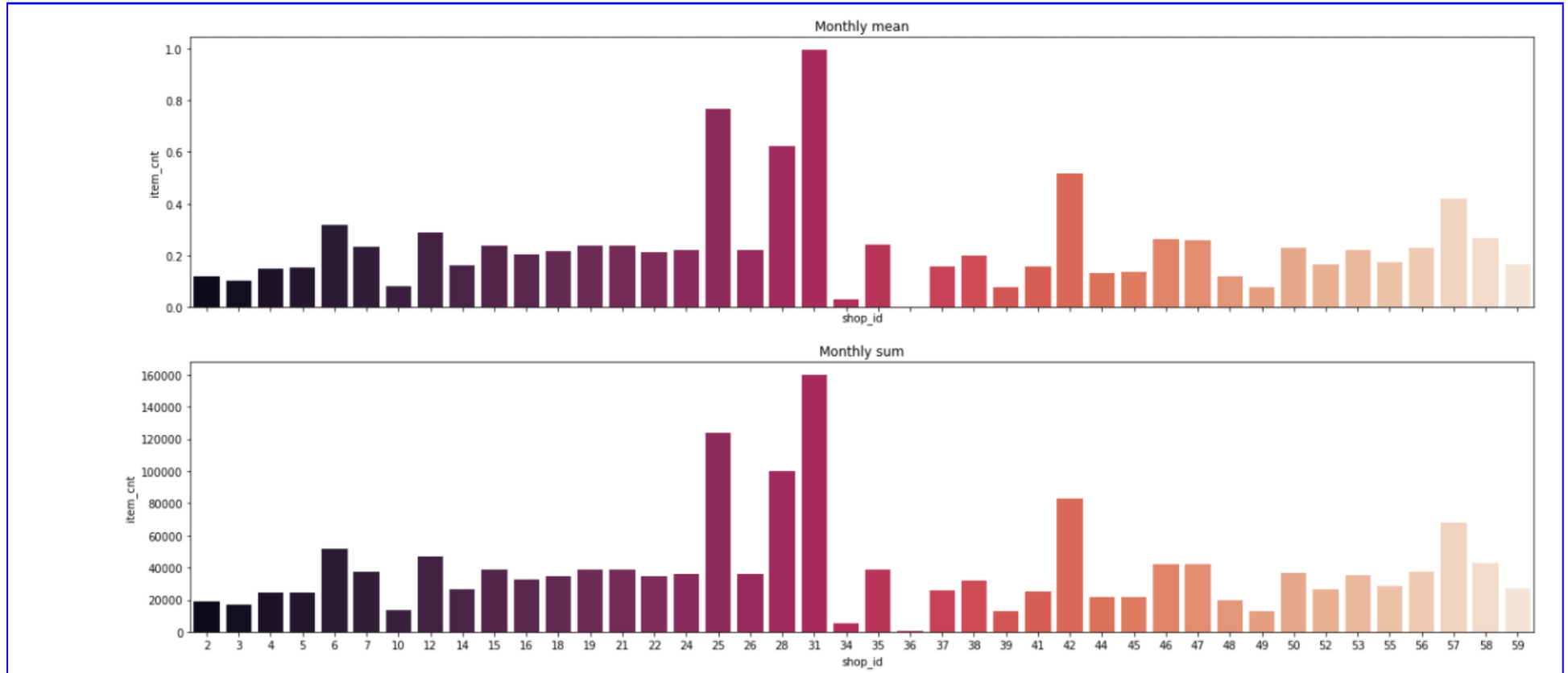


Figure 1 month_total_count.

Explain that the month is related to the sales volume of goods: the sales volume at the end of the year is increasing

Shop sales



~~Objective: To prepare for feature extraction~~ **Figure 2** shop_count.

Sales of different category

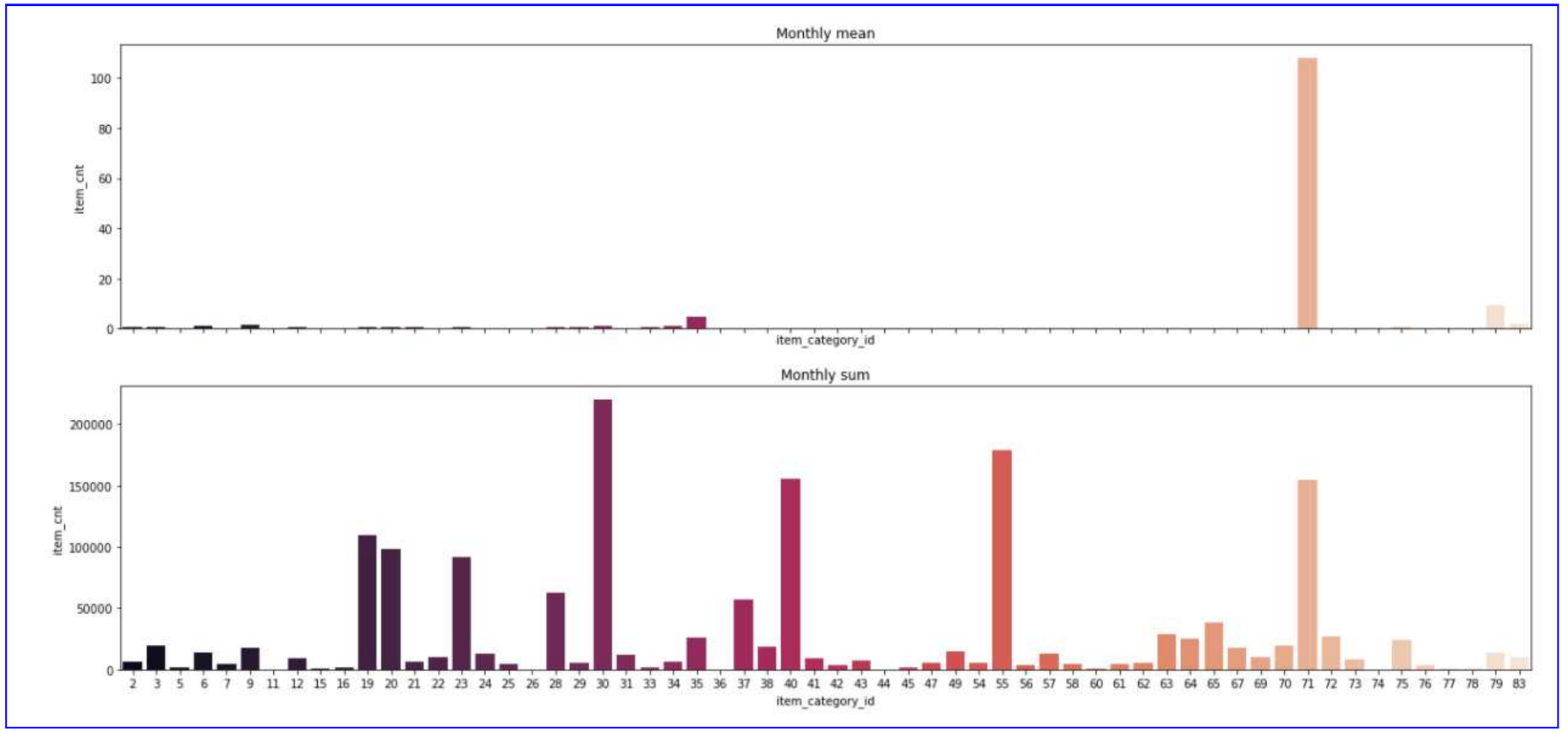


Figure 3 item_category_count.

Item and Shop Information

categories of items	large categories, small categories, we separate them, and code them separately to facilitate subsequent feature extraction
Shop information	the city where the store is located, the type of store, which we separate and encode separately for subsequent feature extraction

Model

decision tree

Dicision tree	In machine learning, decision tree is a prediction model, which represents a mapping relationship between object attributes and object values. Each node in the tree represents an object, and each branch path represents a possible attribute value, while each leaf node corresponds to the value of the object represented by the path from the root node to the leaf node. The decision tree has only a single output, if you want to have complex output, you can establish an independent decision tree to deal with different outputs. Decision tree is a frequently used technology in data mining, which can be used to analyze data, and also can be used for prediction.

~~Model~~

Model selection

- GBDT
- Xgboost
- lightgbm
- neural network

Method One

~~Method:The sales of the 34th month are regarded as the sales of the 35th monthoperation:Count the sales volume of each item in each store in the 33rd month and merge it with test Result:RMSE=1.16777~~

Method	The sales of the 34th month are regarded as the sales of the 35th month
operation	Count the sales volume of each item in each store in the 33rd month and merge it with test
Result	RMSE=1.16777

Method Two

~~Some historical information needs to be generated by delayed operations. For example, you can use the 0-33 month sales as a historical feature of the 1-34 month (one month delay).~~

Features	■ shop_id ■ item_id ■ item_cnt_month
Method	lightgbm
Result	RMSE=

Method Three

Data feature	'date_block_num', 'shop_id', 'item_id', 'item_category_id', 'cat_type_code', 'cat_subtype_code', 'shop_city_code', 'shop_type_code'
Monthly sales feature	■ item_cnt_month ■ date_avg_item_cnt ■ date_item_avg_item_cnt ■ date_shop_avg_item_cnt ■ date_cat_avg_item_cnt ■ date_cat_shop_avg_item_cnt ■ date_type_avg_item_cnt ■ date_item_type_avg_item_cnt ■ date_city_avg_item_cnt
Historica feature	delay:1,2,3,6,12

Method Three


```
print([column for column in X_train])
```

```
['date_block_num', 'shop_id', 'item_id', 'item_category_id', 'cat_type_code', 'cat_subtype_code', 'shop_city_code', 'shop_type
tem_cnt_month_lag_1', 'item_cnt_month_lag_2', 'item_cnt_month_lag_3', 'item_cnt_month_lag_6', 'item_cnt_month_lag_12', 'date_a
t_lag_1', 'date_avg_item_cnt_lag_2', 'date_avg_item_cnt_lag_3', 'date_avg_item_cnt_lag_6', 'date_avg_item_cnt_lag_12', 'date_i
em_cnt_lag_1', 'date_item_avg_item_cnt_lag_2', 'date_item_avg_item_cnt_lag_3', 'date_item_avg_item_cnt_lag_6', 'date_item_avg_
ag_12', 'date_shop_avg_item_cnt_lag_1', 'date_shop_avg_item_cnt_lag_2', 'date_shop_avg_item_cnt_lag_3', 'date_shop_avg_item_cn
'date_shop_avg_item_cnt_lag_12', 'date_cat_avg_item_cnt_lag_1', 'date_cat_avg_item_cnt_lag_2', 'date_cat_avg_item_cnt_lag_3',
avg_item_cnt_lag_6', 'date_cat_avg_item_cnt_lag_12', 'date_cat_shop_avg_item_cnt_lag_1', 'date_cat_shop_avg_item_cnt_lag_2',
hop_avg_item_cnt_lag_3', 'date_cat_shop_avg_item_cnt_lag_6', 'date_cat_shop_avg_item_cnt_lag_12', 'date_type_avg_item_cnt_lag_
type_avg_item_cnt_lag_2', 'date_type_avg_item_cnt_lag_3', 'date_type_avg_item_cnt_lag_6', 'date_type_avg_item_cnt_lag_12', 'da
pe_avg_item_cnt_lag_1', 'date_item_type_avg_item_cnt_lag_2', 'date_item_type_avg_item_cnt_lag_3', 'date_item_type_avg_item_cnt
'date_item_type_avg_item_cnt_lag_12', 'date_city_avg_item_cnt_lag_1', 'date_city_avg_item_cnt_lag_2', 'date_city_avg_item_cnt_
ate_city_avg_item_cnt_lag_6', 'date_city_avg_item_cnt_lag_12', 'date_item_city_avg_item_cnt_lag_1', 'date_item_city_avg_item_c
'date_item_city_avg_item_cnt_lag_3', 'date_item_city_avg_item_cnt_lag_6', 'date_item_city_avg_item_cnt_lag_12', 'delta_price_l
h', 'days', 'item_shop_last_sale']
```

Method Three

Result	■ training's rmse: 0.664209 ■ valid_1's rmse: 0.880256

Lightgbm