KAGGLE PROJECT REPORT

PENGCHENG JIANG AND GANG LI

Abstract. I finished the project and got good results.

Contents

Date: 2021-04-26.
1991 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. kaggle,slides,notebook.

1. Problem Definition

We are given a challenging time-series dataset consisting of daily sales data, kindly provided by one of the largest Russian software firms - 1C Company. Our target is predict total sales for every product and store in the next month Submissions are evaluated by root mean squared error (RMSE).This report introduces how to participate in this competition, and gives the whole process of how to use python language for data cleaning, feature engineering extraction and model construction.

2. Data Cleaning

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

The above table is some information about the datas.

Next, let's look at the data types of the training set which will be conducive to the subsequent processing:

- 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

Next, let's look at the data types of the test set:

- 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

Next, we do some common processing on the data:

- Missing Value and Non Value:Find out whether there are empty values or missing values in the data
- Cartesian product: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:delete stores, goods in training set but not in the test set

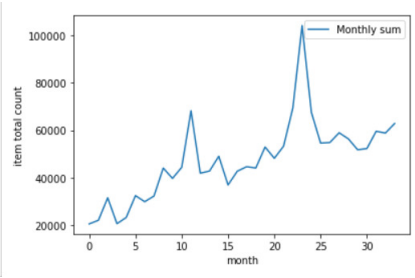

FIGURE 1. month˙total˙count

- Data duplication:See if duplicate items exist in the dataset
- Outliers:Calculate the outliers of item˙cnt˙day and item price

3. DATA ANALYSIS

First, look at the monthly sales of goods at figure 1

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

Next, take a look at the sales of each store in figure 2

Finally, let's look at the sales of different kinds of goods in figure 3

Item and Shop Information:

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

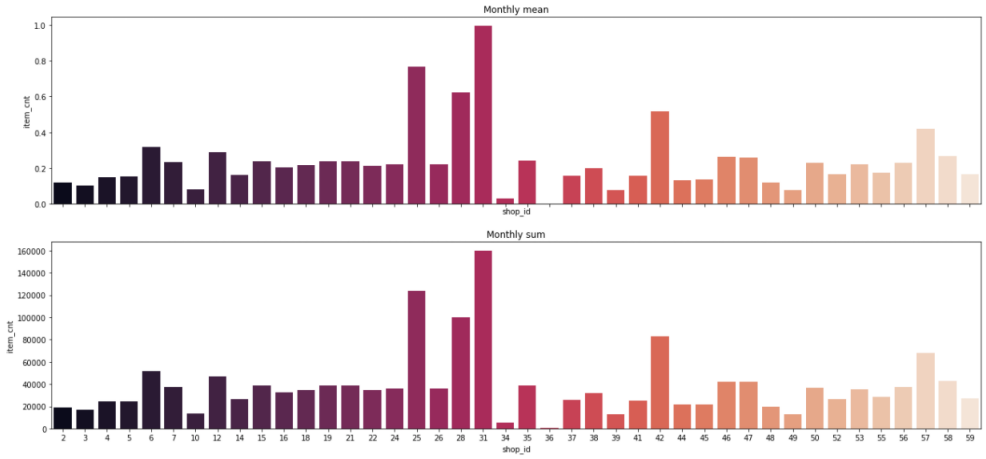

FIGURE 2. shop˙count

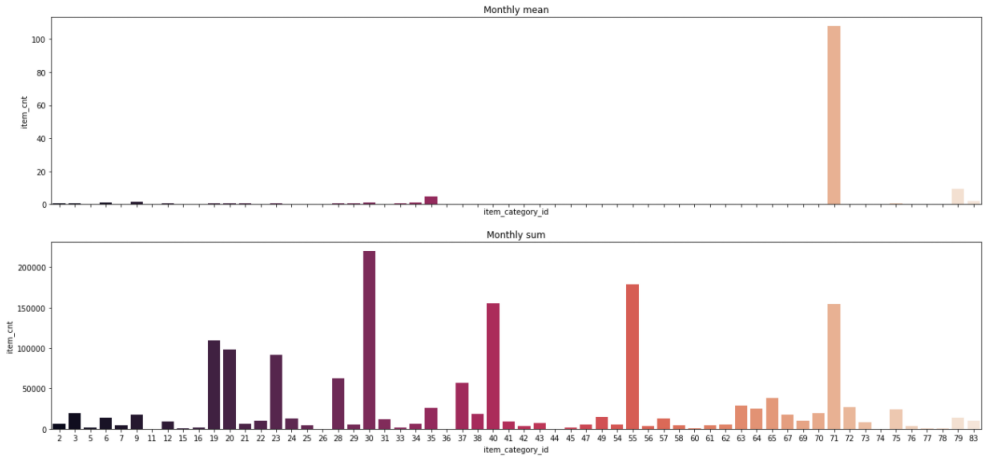

FIGURE 3. item˙category˙count

4. MODEL

4.1. **decision 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.

4.2. **Candidate model.**

- GBDT
- Xgboost
- lightgbm
- neural network

4.3. **Method One.**
The sales of the 34th month are regarded as the sales of the 35th month,Count the sales volume of each item in each store in the 33rd month and merge it with test The result is **RMSE=1.16777**

4.4. **Method Two.**
Features:

- shop˙id
- item˙id
- item˙cnt˙month

The model we choosed is lightgbm,and The result is **RMSE=**

```
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_code', 'i
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_avg_item_cn
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_item_avg_it
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_item_cnt_l
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_cnt_lag_6',
'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', 'date_cat_
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', 'date_cat_s
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_1', 'date_
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', 'date_item_ty
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_lag_6',
'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_lag_3', 'd
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_cnt_lag_2',
'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_lag', 'mont
h', 'days', 'item_shop_last_sale']
```

FIGURE 4. final features

4.5. **Method Three.**

Adding historical information is good for prediction. We can see the features after adding historical information in Figure 4. Finally, through experiments, it can be proved that the prediction results are improved. **training's rmse: 0.664209,valid˙1's rmse: 0.880256**

5. CONLUSION

Through data processing and adding delay information, the model achieves good results

(A. 1) SCHOOL OF COMPUTER SCIENCE,, JILIN UNIVERSITY, JILIN 130000, CHINA
Email address, A. 1: pcjiang@tulip.academy

(A. 2) SCHOOL OF INFORMATION TECHNOLOGY, DEAKIN UNIVERSITY, GEELONG, VIC 3216, AUSTRALIA
Email address, A. 2: gang.li@deakin.edu.au