FLIPOO PRESETANTION REPORT

BAOBAO SONG

Abstract. This report mainly explain, analyse and solve the problem of predicting the future sales with the database which is given. And the report contains five parts. The first part decribe the problem, interpret the data and evaluation criteria. Moreover,the report do some data processing, which includes filter duplicate data and others. Third, select some features to construct feature matrix. Then, experiment and analyze the performance of the lightgbm model based on experimental result. The last one is conclusion.

Contents

Date: 2020-09-18.
1991 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. Machine Learning, Data Mining, ...

1. Introduction

1.1. **Describe the Problem.** This is a problem with time-series prediction. Many information are given about daily sales data.The raw dataset contains train set with 2935849 samples and 214200 unlabeled samples as test set. Through the train data, predict total sales for every product and store in the next month.

1.2. **Interpret the Data.** Here's the data in the dataset.

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 November 2015)	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

1.3. **Evaluation Criteria.** Before experiment, determine the evaluation methods to assess the model performance is very important, usually it has the RMSE methods to evaluate.

2. Data Processing

2.1. **Missing Value and NaN Value.** There are no missing value and none 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

2.2. **Outliers and Duplicate Data.** Filter duplicate data, outliers and data with price less than zero.

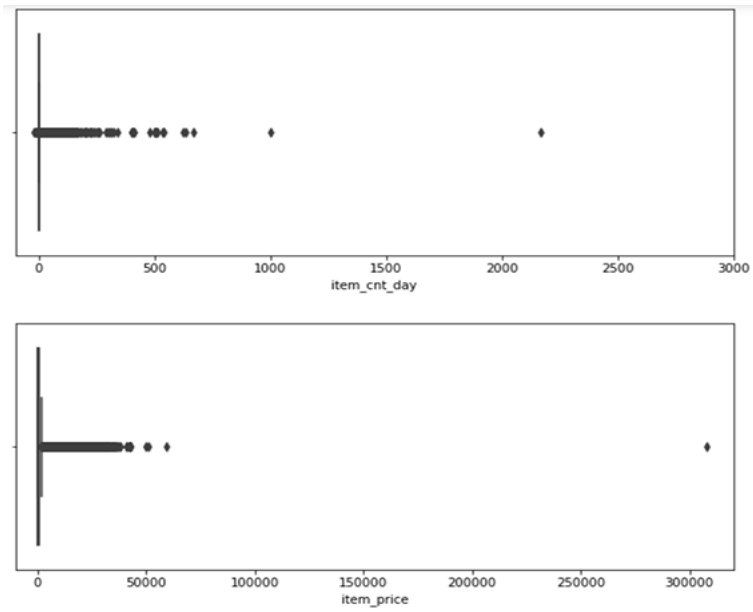


FIGURE 2. Outliers Data

2.3. **Sales Analysis.** Figure3 shows that total sales every month are decreased over time. This reason probably is shops and items are decreased. By analyzing the data, there are many discontinued items in figure4 and these shops are closed:closed shops:0,1,8,11,13,17,23,27,29,30,32,33,40,43,51,54.

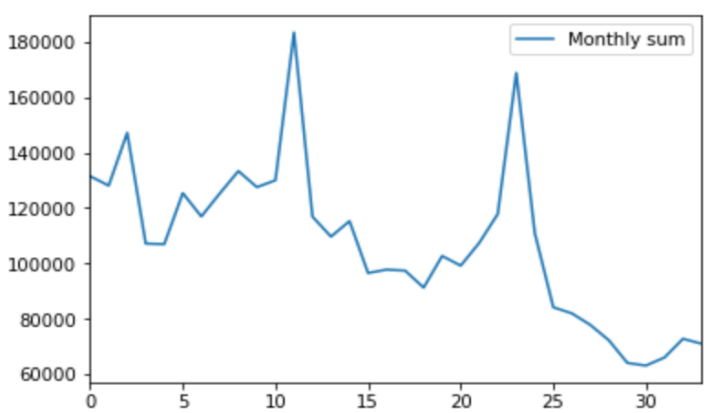


FIGURE 3. Total Sales Over Time

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 4. Discontinued Products

3. Feature Selection

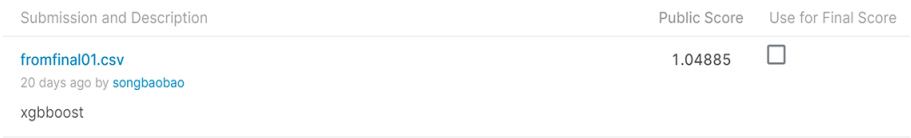
This report simply counts monthly sales of every items, and choose each item every month sales and item categories as feature, final matrix is figure5.

	ID	shop_id	item_id	0	1	2	3	4	5	6	...	25	26	27	28	29	30	31	32	33	item_category_id
0	0	5	5037	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	1.0	1.0	1.0	3.0	1.0	0.0	19.0
1	1	5	5320	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	2	5	5233	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	3.0	2.0	0.0	1.0	3.0	1.0	19.0
3	3	5	5232	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	23.0
4	4	5	5268	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	5	5	5039	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	3.0	0.0	0.0	0.0	1.0	1.0	23.0
6	6	5	5041	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	2.0	20.0
7	7	5	5046	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	55.0
8	8	5	5319	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	2.0	3.0	2.0	2.0	4.0	3.0	2.0	3.0	0.0	55.0
9	9	5	5003	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	10	5	4806	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	3.0	6.0	2.0	6.0	6.0	5.0	5.0	2.0	3.0	30.0

FIGURE 5. Discontinued Products

4. Experiment and Analysis

Using xgboost to predict the sales. And in the final database, change zero to closed shops and discontinued items.
XGBoost is to establish K regression trees so that the predicted value of the tree group is as close as possible to the true value (accuracy) and has the greatest generalization ability. From a mathematical point of view, this is a functional optimization, multi-target. The final score is 1.04885 and get the middle rank.



Submission and Description	Public Score	Use for Final Score
fromfinal01.csv 20 days ago by songbaobao xgbboost	1.04885	☐

Figure 6. Discontinued Products

5. Conclusions

- the features are little.
- The model is not trained.

Acknowledgment

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

The authors would like to thank …

References

[1] Gleb Beliakov and Gang Li. Improving the speed and stability of the k-nearest neighbors method. *Pattern Recognition Letters*, 33(10):1296–1301, 2012.

[2] Gleb Beliakov, Simon James, and Gang Li. Learning choquet-integral-based metrics for semisupervised clustering. *Fuzzy Systems, IEEE Transactions on*, 19(3):562–574, 2011.

List of Todos