

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 decrbe 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

| | |
|--|---|
| 1. Introduction | 2 |
| 1.1. Describe the Problem | 2 |
| 1.2. Interpret the Data | 2 |
| 1.3. Evaluation Criteria | 2 |
| 2. Data Processing | 2 |
| 2.1. Missing Value and NaN Value | 2 |
| 2.2. Outliers and Duplicate Data | 3 |
| 2.3. <u>Process Shops Set</u> | 3 |
| 2.4. <u>Process Items Set</u> | 3 |
| 2.5. <u>Process Categories Set</u> | 3 |
| 2.6. Sales Analysis | 3 |
| 3. Feature Selection | 3 |
| 3.1. <u>Data Feature</u> | 4 |
| 3.2. <u>Monthly Sales Feature</u> | 4 |
| 3.3. <u>Historical Feature</u> | 5 |
| 4. Experiment and Analysis | 5 |
| 5. Conclusions | 5 |
| Acknowledgment | 6 |
| References | 7 |

Date: (None).

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.

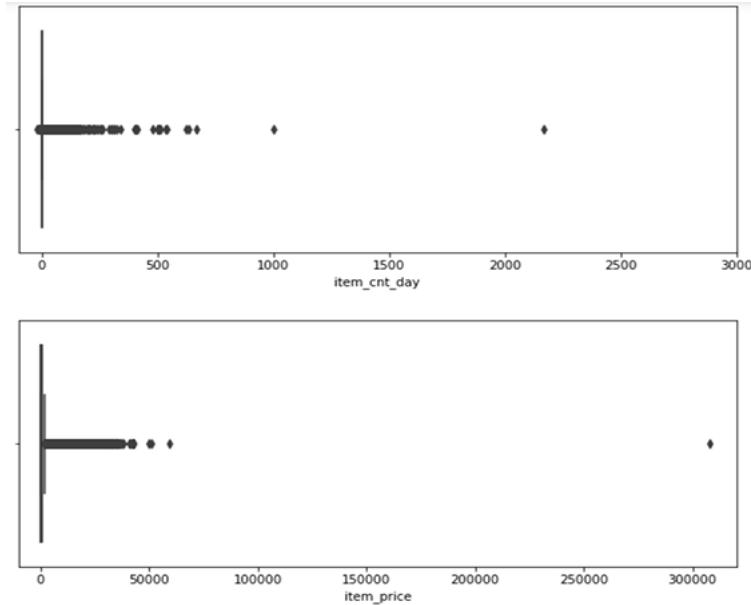


FIGURE 2. Outliers Data

2.3. Process Shops Set. Some shops have same shop name and different shop ID, such as ID of 39 and 40, 10 and 11, 0 and 57, 58 and 1. Then filter the test set and modify the shop ID based on the test. On the other hand, the report analyses shops name and finds the name can be divided into three parts: shop's city, shop's type and the shop's name. Encoding shops information is able to reduce memory usage and find a connection with sales.

2.4. Process Items Set. Some items have same item name and different item ID, such as ID of 2514 and 2558, 2968 and 2970, 5061 and 5063, 14537 and 14539, 19465 and 19475, 19579 and 19581. Then filter the test set and modify the item ID based on the test.

2.5. Process Categories Set. Analyses categories name and finds the name can be divided into two parts: category's type-category's subtype. Encoding shops information is able to reduce memory usage and find a connection with sales.

2.6. 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.

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.~~



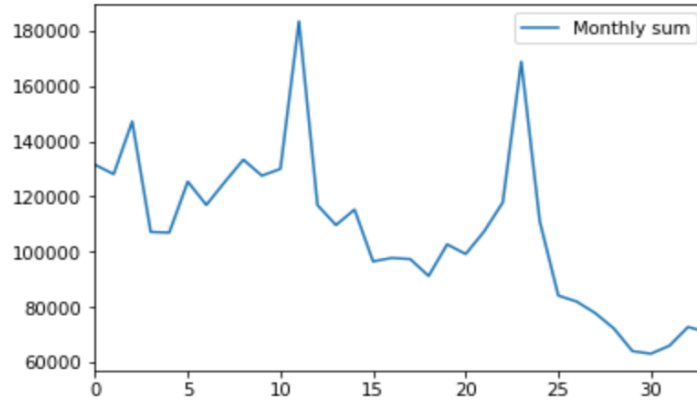


FIGURE 3. Total Sales Over Time

| 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

3.1. Data Feature. Data Feature mainly includes the features that analysing in the previous chapters, such as shop id and various codes. The specific features are shown below.

3.2. 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



Discontinued Products

| | 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 5. Data Feature

3.3. Historical Feature. From figure3 and materials searched in the Internet, historical month delays are set with 1,2,3,6 and 12, and the historical features are below. Finally, first 12 months records and NAN records should be deleted.

- 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

4. EXPERIMENT AND ANALYSIS

~~Using xgboost~~ Using lightgbm to predict the sales. And in the final database, change zero to closed shops and discontinued items.

~~XGBoost is to~~ In the midterm report, XGBoost is used to predict the sales. And in the final report, Lightgbm is used to predict the sales. Because Lightgbm has the faster speed and higher accuracy than XGBoost.

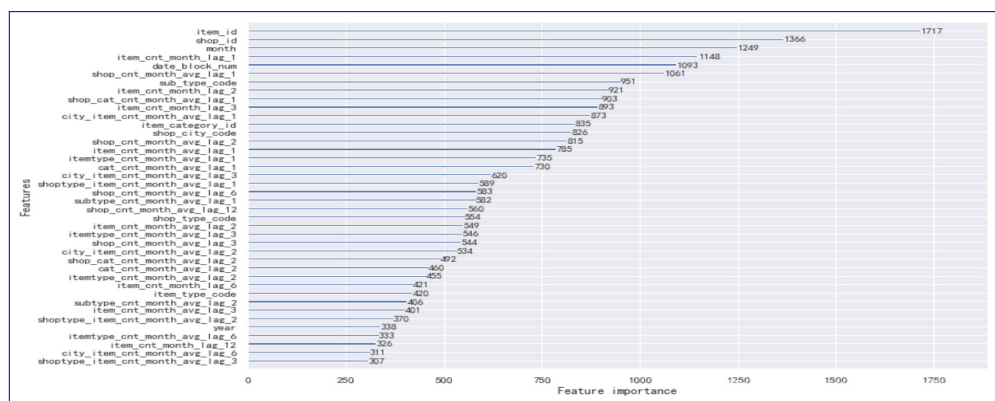
XGBoost [1] is to establish K regression trees so that the predicted value of the tree group is as close as possible to the true value (accuracy) and has the greatest generalization ability. From a mathematical point of view, this is a functional optimization, multi-target. Lightgbm[2] is a distributed gradient promotion framework based on decision tree algorithm. Lightgbm is designed to provide a data science tool with high efficiency, low memory consumption, high accuracy, supporting parallel and large-scale data processing The final score is 1.04885 and get the middle rank.

Figure6 shows the feature importance of sales.

The result:

- score:0.93740
- rank:3027/873

5. CONCLUSIONS

FIGURE 6. ~~Discontinued Products~~ Feature Importance

- ~~the features are little.~~ Exploratory data analysis and data processing is very important for the competition. Exploratory data analysis help to have a certain understanding of the overall appearance of the data, which will help later modeling and analysis. And data processing includes dealing with missing data and outliers, processing datasets and others.
- ~~The model is not trained~~ The most important thing is feature engineering. In the midterm report, several features are selected and some information in database is not used. And in this report, more features are selected, such as historical features and monthly sales feature.
- Compare to midterm presentation, RMSE decreased from 1.04885 to 0.93740. The reason mainly is more features and different modeling. In the figure of feature importance and monthly sales, some historical feature play an important role. And lightgbm model's and xgboost model's result have a marginal difference, but processing speed of lightgbm model is faster.

ACKNOWLEDGMENT

The authors would like to thank ...

REFERENCES

- [1] 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.
- [2] Y. Zhang Wang, Dehua and Y. Zhao. Lightgbm: An effective mirna classification method in breast cancer patients. *the 2017 International Conference.*, 2017.