```
/burl@stx null def /BU.S  /burl@stx null def  def /BU.SS
currentpoint /burl@lly exch def /burl@llx exch def burl@stx
null ne burl@endx burl@llx ne BU.FL BU.S if if burl@stx
null eq  burl@llx dup /burl@stx exch def /burl@endx exch
def burl@lly dup /burl@boty exch def /burl@topy exch def
if burl@lly burl@boty gt /burl@boty burl@lly def if  def /BU.SE
currentpoint  /burl@ury exch def dup  /burl@urx exch def
/burl@endx exch def burl@ury burl@topy lt /burl@topy burl@ury
def if  def /BU.E BU.FL def /BU.FL burl@stx null ne BU.DF
if  def /BU.DF  BU.BB [ /H /I /Border [burl@border] /Color
[burl@bordercolor]  /Action « /Subtype  /URI  /URI BU.L »
/Subtype /Link BU.B /ANN pdfmark /burl@stx null def  def
/BU.BB  burl@stx HyperBorder sub /burl@stx exch def burl@endx
HyperBorder add  /burl@endx exch def burl@boty Hyper-
Border add  /burl@boty exch def burl@topy  HyperBorder
sub /burl@topy exch def  def /BU.B  /Rect[burl@stx burl@boty
```

Predict Future Sales

~~Mingzhu~~ MingZhu Kang

Xi'an Shiyou University
Chinese Academy of Sciences

October 16, 2020

Introduction

Project Introduction

Defn

- Project Introduction
  Analyze the company's operating status, find out the relevant factors affecting the sales volume of goods, Taking the historical sales data set of convenience stores as the object of study, the data were preprocessed and feature extracted, and the model was used to train the data set to predict the sales volume of different goods in each store of the company in the next month.

Data Set Preprocessing

~~Preprocessing of project data sets~~

[=[~~.99~~!

Preprocessing Of Project Data Sets

- ~~sales_train.csv - the training set. Daily historical data from January 2013 to October 2015.~~ Data Collection
- ~~test.csv - the test set. forecast the sales for these shops and products for November 2015.~~
- ~~items.csv - supplemental information about the items/products.~~
- ~~item_categories.csv - supplemental information about the items categories.~~
- ~~shops.csv- supplemental information about the shops.~~
- ~~sample_submission.csv - a sample submission file in the correct format.~~ Download dataset from the kaggle project.

```
         date  date_block_num  shop_id  item_id  item_price  item_cnt_day
0  02.01.2013               0       59    22154      999.00           1.0
1  03.01.2013               0       25     2552      899.00           1.0
2  05.01.2013               0       25     2552      899.00          -1.0
3  06.01.2013               0       25     2554     1709.05           1.0
4  15.01.2013               0       25     2555     1099.00           1.0

                      shop_name  shop_id shop_city shop_type
0   !Якутск Орджоникидзе, 56 фран        0    Якутск    Others
1   !Якутск ТЦ "Центральный" фран        1    Якутск        ТЦ
2                Адыгея ТЦ "Мега"        2    Адыгея        ТЦ
3  Балашиха ТРК "Октябрь-Киномир"        3  Балашиха       ТРК
4        Волжский ТЦ "Волга Молл"        4  Волжский        ТЦ
5          Вологда ТРЦ "Мармелад"        5   Вологда       ТРЦ
6      Воронеж (Плехановская, 13)        6   Воронеж    Others
7          Воронеж ТРЦ "Максимир"        7   Воронеж       ТРЦ
8    Воронеж ТРЦ Сити-Парк "Град"        8   Воронеж       ТРЦ

                                           item_name  ...  item_category_id
0                  ! ВО ВЛАСТИ НАВАЖДЕНИЯ (ПЛАСТ.)         D  ...                40
1  !ABBYY FineReader 12 Professional Edition Full...  ...                76
2              ***В ЛУЧАХ СЛАВЫ   (UNV)                D  ...                40
3              ***ГОЛУБАЯ ВОЛНА  (Univ)                D  ...                40
4                  ***КОРОБКА (СТЕКЛО)                 D  ...                40
..                                                 ...  ...               ...
65                 Ядерный титбит 2 [PC, Цифровая версия]  ...                31
```

Data Cleaning

~~Forecast future trends~~

Training Set Data Cleaning

- ~~Process the training set and only keep the stores and goods with salesin the last 6 months of normal ope~~ Use a scatter plot to observe the distribution of commodity prices and daily sales.
- ~~Use the model training data set. Here, select lightGBM model for training and combine the predicted resul~~ Filter for anomalies and apparent outliers

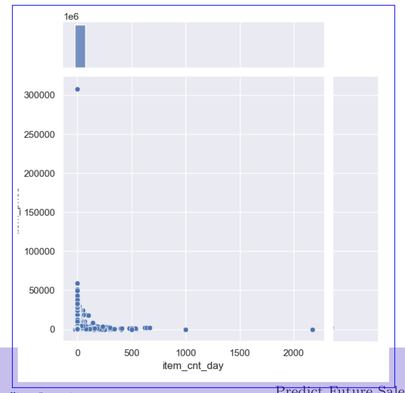


Figure 1: Distribution

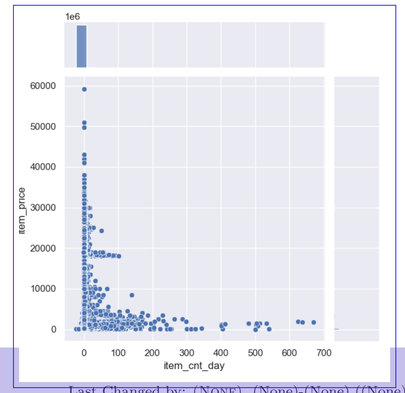


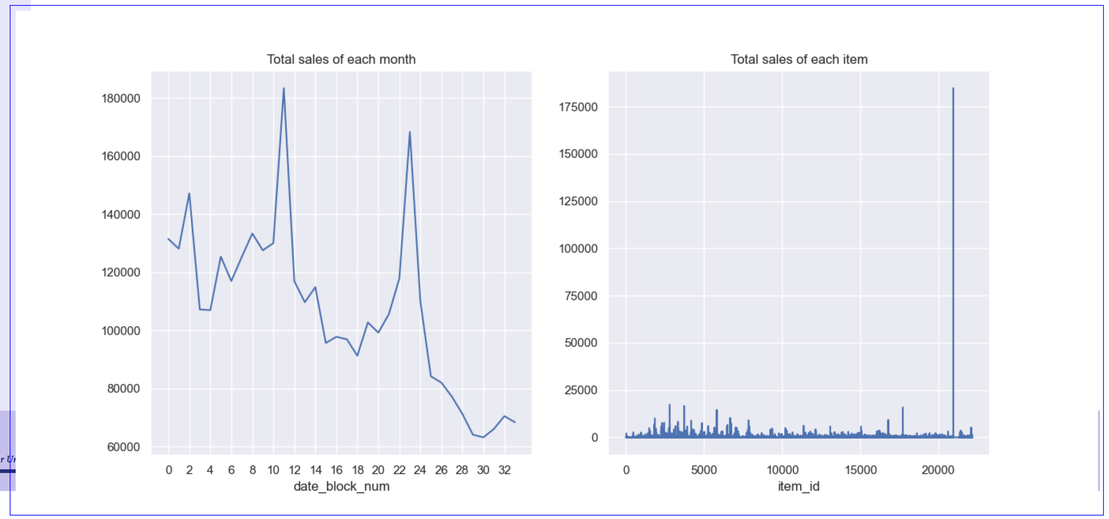
Figure 2: Filter Abnormal

Structured Data And Analysis

Sales Analysis

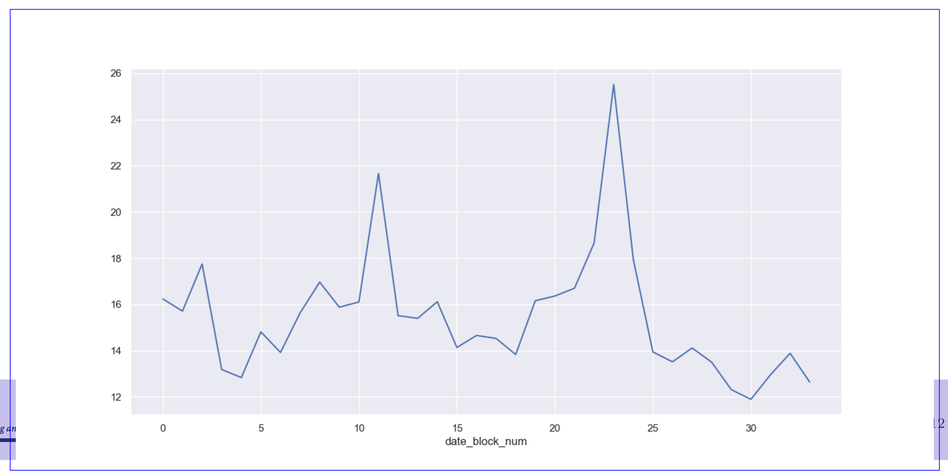
Then, we created additional features. More specifically:

- Sales analysis
  Overall sales were down, and monthly sales were mostly down year on year. One item sold exceptionally well.



Sales Analysis

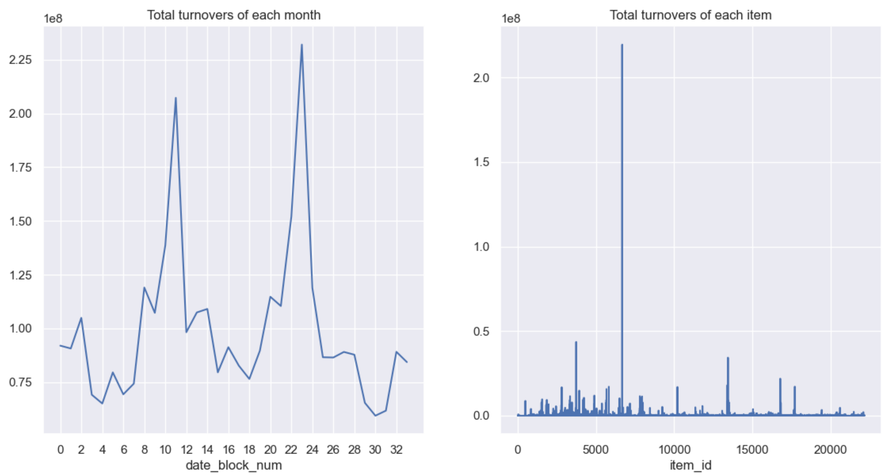
- Average number of items sold in the month
  In 2013 and 2014, the average monthly sales volume of goods under sale was basically 13-16,
  while in 2015, the average monthly sales volume of goods under sale decreased to 12-14.



Profit Analysis

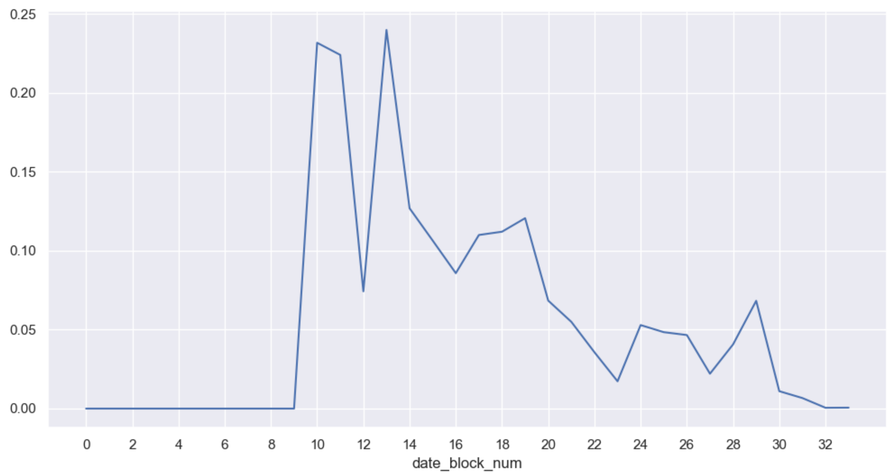
Profit Analysis

- turnover analysis



Profit Analysis

- The highest-grossing commodity
  The number one item in total revenue accounts for a percentage of monthly revenue



Predict Future Trend

- Handle closed stores and discontinued goods.
- Only keep the goods that are normally operated in the last 6 months and the goods with sale volume.

shop_id	8	9	13	17	20	23	...	32	33	36	43	51	54
date_block_num							...						
22	0	0	0	1199	0	0	...	0	814	0	2659	1098	6389
23	0	0	0	1832	0	0	...	0	1056	0	3139	1652	7677
24	0	0	0	689	0	0	...	0	1086	0	1340	976	6043
25	0	0	0	0	0	0	...	0	792	0	0	668	4221
26	0	0	0	0	0	0	...	0	585	0	0	545	4625
27	0	-1	0	0	0	0	...	0	-1	0	0	494	732
28	0	0	0	0	0	0	...	0	0	0	0	758	0

Figure 3: Closed Stores

	date	date_block_num	...	shop_city_code	shop_type_code
0	02.01.2013	0	...	29	1
1	03.01.2013	0	...	14	2
2	05.01.2013	0	...	14	2
10	03.01.2013	0	...	14	2
11	05.01.2013	0	...	14	2
...	...	...	...	...	...
2935819	10.10.2015	33	...	14	2
2935828	09.10.2015	33	...	14	2

Figure 4: Normal Opreation

- use historical sales data to predict future sales.
  Using the historical sales data as the characteristics of the model,
  this month's sales results as labels to build a model for regression analysis.

	date_block_num	shop_id	...	item_type_code	sub_type_code
0	0	59	...	18	21
1	0	59	...	12	41
2	0	59	...	12	41
3	0	59	...	12	39
4	0	59	...	12	42
...	...	...	...	...	...
11054935	34	45	...	12	38
11054936	34	45	...	13	47

Figure 5: fusion feature

Model Adopted

LightGBM Model

Since this is a typical multi-classification problem, we This project uses lightGBM model for training.
LightGBM is a fast, distributed, high-performance gradient enhancement framework based on decision tree algorithms.It supports category characteristics.
LightGBM supports category characteristics directly and natively by changing the decision rules of the decision tree algorithm, without transformation.

Summary

Project Summary

From data analysis methods to feature engineering and prediction model construction, a lot of time has been spent to study and comb.
Through this project, I have learned a lot, including the effective aspects of problem cutting, code implementation of analysis algorithm, design of analysis process, etc. which enables me to better grasp the thinking of data analysis on the whole.
In the process of predictive analysis, the theoretical and data support for feature analysis and model construction is not concise and powerful enough,which needs to be strengthened.