Predict Future Sales

— exploring time-series data

Yufei Gao 001814582

1 What the project had done?

Time-series dataset (with textual variables)

Preprocessing (tf-idf & tendency features)

Model training (Linear Regression, Random Forest, XGBoost, LightGBM)

Predict specific item sales in specific shop in next month

2 Preprocessing

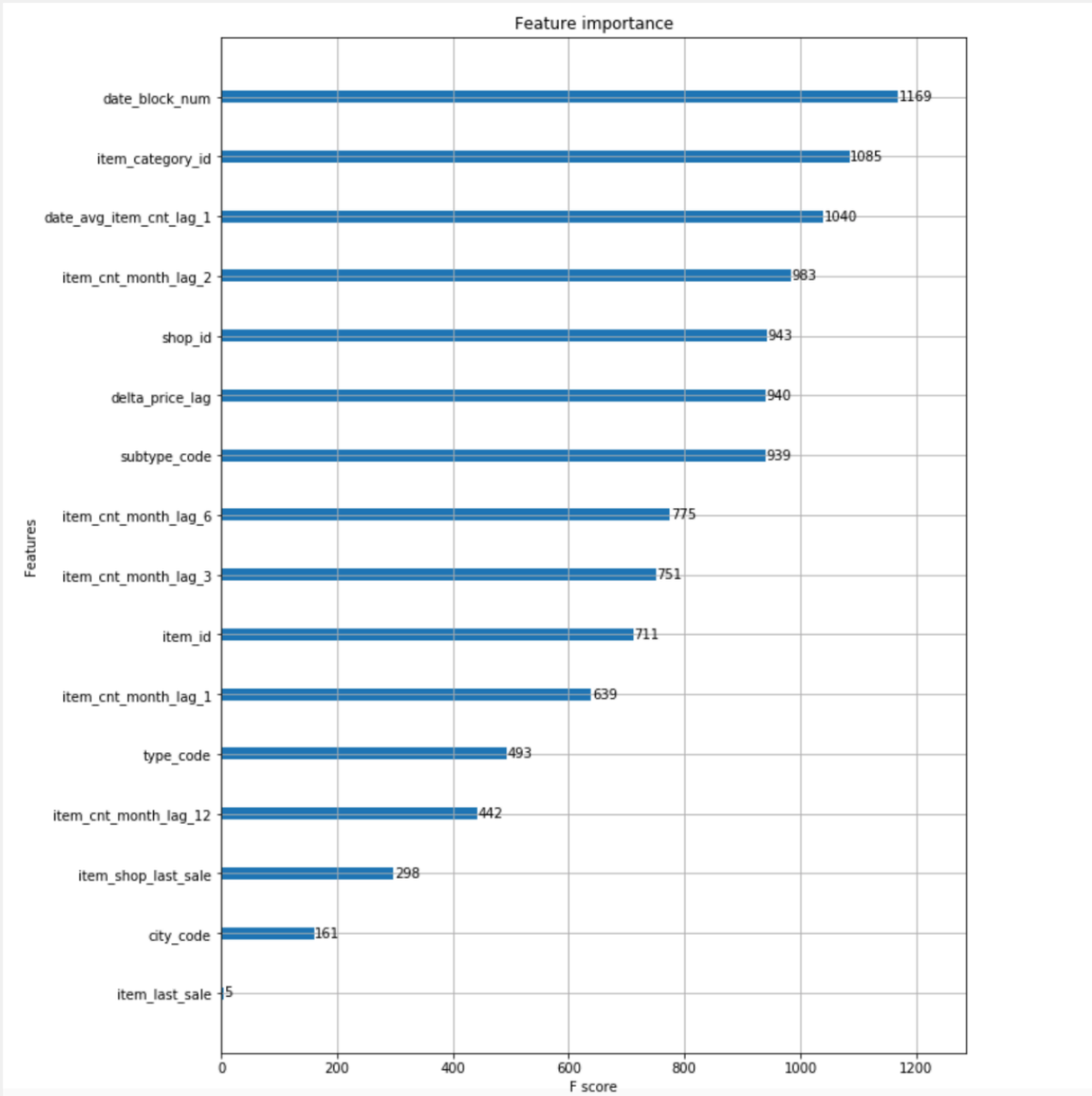
tf-idf

- term frequency–inverse document frequency
- “phoneix” is more important than “as”
- based on corpus

tendency features

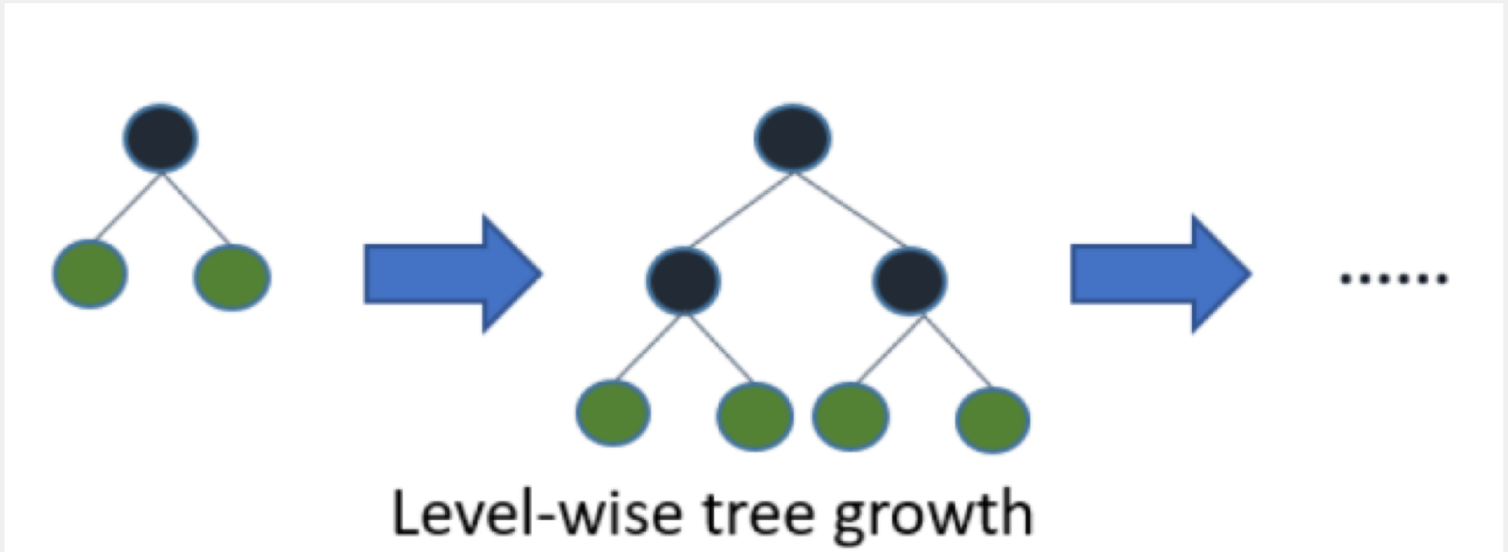
- combine shop and item as an object
- find the tendency of sold, price, revenue
- generate large matrix

ignore some

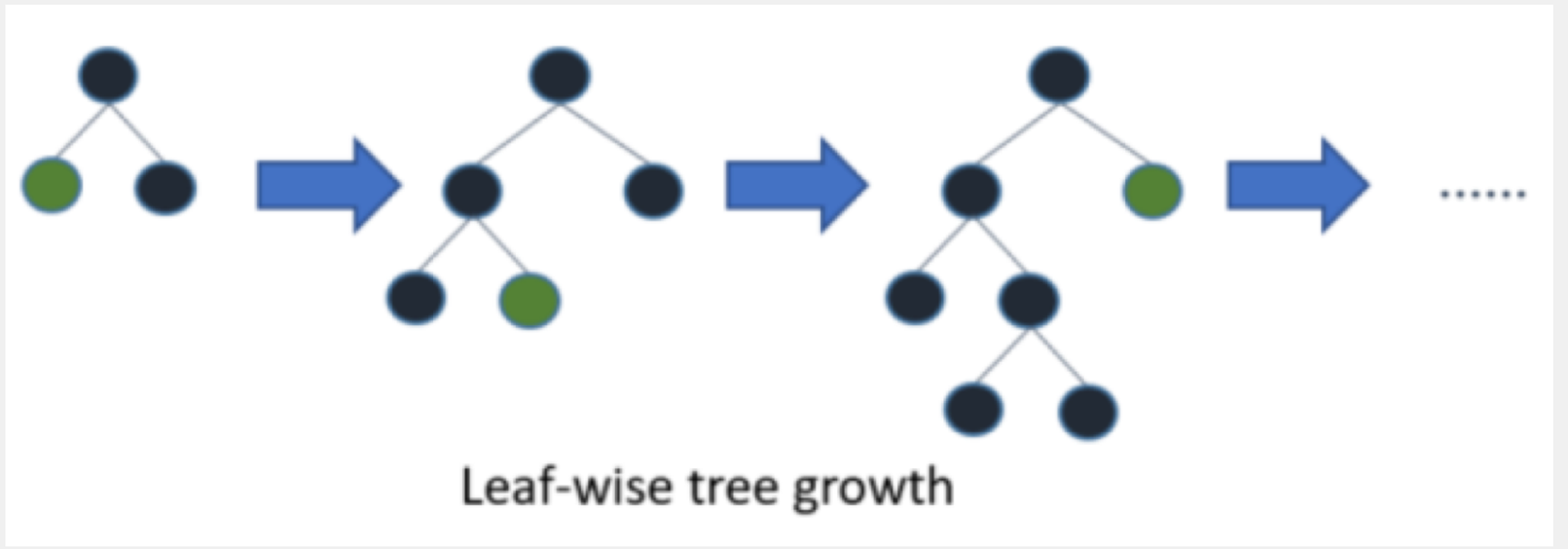


3 Models

1. Linear Regression
2. Random Forest
3. XGBoost
4. LightGBM



level-wise growth strategy used in XGBoost



leaf-wise growth strategy used in lightGBM

Result

Method	RMSE(the lower, the better)
Linear Regression	1.16
Linear Regression with tf-idf	9.23
Random Forest	1.20
Random Forest with tf-idf	1.23
Random Forest with most important variables	1.21
Random Forest tf-idf, important variables	1.22
LightGBM	1.06
LightGBM with tf-idf	1.20
LightGBM with trend features	0.96
XGBoost	1.13
XGBoost with tf-idf	1.19
XGBoost with trend features	0.90

Conclusion

- **tf-idf not good for this dataset**

Especially for the linear regression, the result is ridiculous. It generated a lot of noise which couldn't provide valid information for model to predict

- **Trend features make sense**

This dataset is time-series, and this method seems could help to predict the future data

- **Ignore some variables seems didn't make sense**

Instead, should weight or find more valuable features

- **XGBoost has the best performance**

Linear regression make sense to some extends, random forest could handle more different variables. LightGBM fits the training set faster than XGBoost and occupied less computing resource, but it accelerate at the expense of losing accuracy. The result may change with the parameters become larger.

ID	item_cnt_month
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
11	0
12	0.730749409
13	0
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	0
22	-53.08917104
23	0
24	0
25	0
26	0
27	-67.69830348

ridiculous result by linear regression

Thank you!