Grocery Sales Forecasting

Yixu Liu

Qing Teng academy, Canton, Guangdong province, China, 510000
**Corresponding author. Email: yixuliu558@gmail.com*

ABSTRACT
Forecasting grocery items can not only avoid excessive stockpiling but also meet customer demand. This can reduce the grocery store's losses and increase the grocery store's turnover. On the other hand, considering features that affect sales is also the core of feature engineering. This article makes a forecast about the sales of merchandise in a large chain store. In this paper, two cores of time series and LGBM are mainly used to complete the model establishment. A model that can predict the sales of goods is built. The data used for training is analyzed and processed. The accuracy of the model is measured using the mean squared error, which gives a final accuracy of 0.35069696616549817. At the end of the paper, it proposes an improved method for this model, and how to solve the same problem under other conditions (such as when the data is particularly small).

Keywords: *Machine learning, sale prediction, time series, LGBM, regression*

1. INTRODUCTION

Brick-and-mortar grocery stores are always closely tied to purchasing and sales forecasts. An incorrect prediction will cause over-purchasing, which will lead to overstock and spoilage. On the other hand, an insufficient purchase will cause a shortage of merchandise available to customers. Therefore, it is very important for grocery stores to accurately predict the purchase volume of goods. Sales prediction is an important part of modern business intelligence [1]. Accurate forecasts can bring huge benefits to a businessman or a business. In the last decade, machine learning has been used for various business predictions, such as in the financial industry, stock forecasting [2], etc.

The research topic of this paper is sales forecasting. Further, in this paper, the specific research problem that needs to be solved is how to correctly predict the sales volume of each different item in different grocery stores. Corporación Favorita, a large Ecuadorian-based grocery retailer, operates hundreds of supermarkets with over 200,000 different products on the shelves. Every month, retailers add new locations with unique needs, new products, ever transitioning seasonal tastes, and unpredictable product marketing, all of which complicate the problem, so machine learning needs to take into account associations in many different features. To accurately predict sales, it is critical to take into account a wide range of factors. Corporación Favorita gives important data to make relevant predictions for this model training. The purpose of this study was to address the inventory problems that exist in most grocery stores, such as overstocking and not having enough items for customers to purchase. In this paper, based on the above problem, a related prediction model is studied. The model was able to predict the sales of different items in each store. Stores can purchase goods at different times according to the forecast, which can reduce unreasonable purchases in the store and increase the turnover of the grocery store.

2. METHODOLOGY

2.1. Data resource

The dataset used in this article was obtained through a competition published by the Corporación Favorita community [3] on the Kaggle website platform.

2.2. Data analysis and processing

As a regression machine learning problem, analyzing the data is very important. Because extracting a part of the data set and selecting a part of the features for training can effectively improve the accuracy of the data and make the model run more quickly. This dataset contains 7 sub-datasets of different kinds. The training dataset contains 23808261 rows × 7 columns of data.

	Unnamed: 0	id	date	store_nbr	item_nbr	unit_sales	onpromotion
0	101688779	101688779	2017-01-01	25	99197	1.0	False
1	101688780	101688780	2017-01-01	25	103665	7.0	False
2	101688781	101688781	2017-01-01	25	105574	1.0	False
3	101688782	101688782	2017-01-01	25	105857	4.0	False
4	101688783	101688783	2017-01-01	25	106716	2.0	False
...	...	...	...	...	...	...	...
23808256	125497035	125497035	2017-08-15	54	2089339	4.0	False
23808257	125497036	125497036	2017-08-15	54	2106464	1.0	True
23808258	125497037	125497037	2017-08-15	54	2110456	192.0	False
23808259	125497038	125497038	2017-08-15	54	2113914	198.0	True
23808260	125497039	125497039	2017-08-15	54	2116416	2.0	False

23808261 rows × 7 columns

Figure 1 The part of the dataset

The period of the data set is very large, and it spanned the 4 years from 2013 to 2017. But in fact, such a large dataset is not needed. Because with the change of time, people's demand for goods will also change. For example, 20 years ago, CDs were very popular, so their sales were very high. However, as time has gone by, more and more new audio and video products have appeared, and the sales of CDs have become less and less. If too large a period is used, the model's predictions will also be less accurate because too much noise is absorbed.

After analyzing the training dataset, the dataset was processed. 2017 was chosen to improve the training accuracy of the model. In addition, the data set was divided into different periods (in years) to verify the conjecture. It can be seen from the data that the ID feature in this training dataset is not very useful. Although it is a unique value, it is the same as the previous index. In addition, although store ID and item ID are two columns of data, they are essentially combined data classes, because they are essential to predict the sales information of the specified product in the specified store.

In the test data set, the specified store corresponds to a prediction of the sales of the specified product. So, it is not a simple time-series regression in essence, but a relatively complex time series regression analysis prediction. Apart from that, considering multiple datasets, this problem can also be classified as a multiple regression problem at the same time. Because, it is not only necessary to consider the impact of different products on sales in the sales data set in the training dataset, but also consider the impact of other data sets on sales, such as the impact of different periods on product sales.

In essence, the store information table plus the commodity information table are essential to forming the curve. That is to say, these two are the inputs, and the transaction information, oil information, and holiday information will have a great impact on commodity sales. Therefore, there is more than one input dependent variable, but the result of combining multiple dependent variables. This means that this is a multiple regression model. To sum up, the current data information problem is a regression problem about time series first. Secondly, this is a multiple regression problem because there is more than one dependent variable. Finally, the prediction is a large cycle, so this problem is cyclical.

2.3. Model architecture

To solve this problem, the author uses machine learning to finish this regression problem. Machine learning is an evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment [4].

LGBM aims to make gradient boosting on decision trees faster. The idea is that instead of checking all of the splits when creating new leaves, only some of them are checked: the model first sorts all of the attributes and buckets the observation by creating discrete bins. When there needs to be a split of a leaf in the tree, instead of iterating over all of the leaves, it simply iterates over all of the buckets. This implementation is called histogram implementation by its authors. The trees are grown depth-first (or leaf-wise) keeping the presorted state instead of level-wise like other gradient boosting methods.

The algorithm chooses the leaf with the maximum delta loss to grow and does not grow the whole level. LightGBM uses the histogram algorithm to convert feature values into bin values and does not need to record the index of features to samples, which greatly reduces memory consumption [5]. According to the characteristics of LGBM, this is a very good model to train on a very large dataset. That is why LBGM is used as the main model to predict grocery sales.

According to the analysis, time series is needed to predict the sale of groceries. Time series is the core of this model, according to the previous data analysis which predicts that sales of every item are cyclical. Therefore, it is necessary to use the idea of time series to divide the data into the data set into multiple sections, which are divided according to the length of different times. In the training of this model, the period was divided into the data according to, 1 day, one week, two weeks, one month, 60 days, 140 days, and then the results of different periods were obtained, and finally, the mean regression performed.

```
 def prepare_dataset(t2017, is_train=Tru
e):
   X = pd.DataFrame({
     "day_1_2017": get_timespan(df_2017,
t2017, 1, 1).values.ravel(),
     "mean_3_2017": get_timespan(df_2017,
 t2017, 3, 3).mean(axis=1).values,
     "mean_7_2017": get_timespan(df_2017,
 t2017, 7, 7).mean(axis=1).values,
     "mean_14_2017": get_timespan(df_201
7, t2017, 14, 14).mean(axis=1).values,
     "mean_30_2017": get_timespan(df_201
7, t2017, 30, 30).mean(axis=1).values,
     "mean_60_2017": get_timespan(df_201
7, t2017, 60, 60).mean(axis=1).values,
     "mean_140_2017": get_timespan(df_201
7, t2017, 140, 140).mean(axis=1).values,
     "promo_14_2017": get_timespan(promo_
2017, t2017, 14, 14).sum(axis=1).values,
     "promo_60_2017": get_timespan(promo_
2017, t2017, 60, 60).sum(axis=1).values,
     "promo_140_2017": get_timespan(promo
_2017, t2017, 140, 140).sum(axis=1).value
s
  })
```

2.4. Experiment

```
params = {
    'num_leaves': 31,
    'objective': 'regression',
    'min_data_in_leaf': 300,
    'learning_rate': 0.1,
    'feature_fraction': 0.8,
    'bagging_fraction': 0.8,
    'bagging_freq': 2,
    'metric': 'l2',
    'num_threads': 4}
```

This prediction uses parallelism. The main idea of feature parallelism is to find the optimal segmentation points on different feature sets on different machines and then synchronize the optimal segmentation points between machines.

Num_leaves is to adjust the maximum number of leaves, and the maximum number of leaves can be adjusted to prevent overfitting. Here, num_leaves is adjusted according to the default value of 31 to LGBM.

Figure 2 Light GBM grows tree vertically

Because LightGBM uses a leaf-wise algorithm, num_leaves is used instead of max_depth when adjusting the complexity of the tree. num_leave and max_depth can be converted into the following formula: num_leaves = 2^(max_depth). Its value should be set less than 2^(max_depth), otherwise, it may lead to overfitting.

According to the above analysis, this problem is a regression problem. On the other hand, to avoid underfitting, other factors need to be considered. Min_date_in_leaf whose value is determined according to the sample data and the number of samples. Increasing this value can avoid underfitting, but it should also be noted that too deep numbers will lead to overfitting. LightGBM considers all features in a dataset during the training process. Setting LightGBM randomly selects 80% of features at the beginning of constructing each tree to increase the total number of splits that have to be evaluated to add each tree node.

3. RESULT

```
 [LightGBM] [Info] Total Bins 5598
 [LightGBM] [Info] Number of data points
in the train set: 1005090, number of used
 features: 40
 [LightGBM] [Info] Start training from sc
ore 0.951286
 Training until validation scores don't i
mprove for 50 rounds
 [100] training's l2: 0.357337 valid_1's
l2: 0.364934
 [200] training's l2: 0.353194 valid_1's
l2: 0.360464
 [300] training's l2: 0.350275 valid_1's
l2: 0.3574
 [400] training's l2: 0.347735 valid_1's
```

```
l2: 0.354746
 [500] training's l2: 0.345476 valid_1's
l2: 0.352438
 Did not meet early stopping. The best it
eration is:
 [500] training's l2: 0.345476 valid_1's
l2: 0.352438
 mean_30_2017: 1274146.02
 mean_14_2017: 628060.77
 mean_60_2017: 408910.31
 mean_7_2017: 180058.20
 promo_15: 133881.93
 mean_20_dow1_2017: 129704.40
 mean_4_dow1_2017: 19523.54
 mean_20_dow2_2017: 18210.
 day_1_2017: 17458.11
 promo_14_2017: 14440.82
 mean_3_2017: 13047.28
 promo_60_2017: 10967.97
 promo_140_2017: 8512.83
 promo_14: 8011.99
 mean_20_dow4_2017: 5903.63
 mean_140_2017: 5418.23
 mean_20_dow0_2017: 5239.94
 mean_4_dow6_2017: 4325.53
 mean_4_dow2_2017: 4048.12
 mean_4_dow0_2017: 4003.44
 mean_20_dow6_2017: 3866.71
 mean_4_dow4_2017: 3251.47
 mean_20_dow5_2017: 3224.08
 promo_10: 3223.94
 mean_4_dow3_2017: 3059.59
 mean_20_dow3_2017: 2923.79
 mean_4_dow5_2017: 2918.48
 promo_13: 2542.46
 promo_12: 2266.04
 promo_7: 1789.76
 promo_0: 1754.63
 promo_9: 1615.56
 promo_11: 1236.23
 promo_8: 1015.74
 promo_6: 769.47
 promo_2: 746.09
 promo_4: 647.41
 promo_3: 646.71
 promo_1: 625.12
 promo_5: 444.90
```

The training is complete and the results are as shown above. To test the accuracy of the model, mean_squared_error is called to predict the exact value of the model. The final accurate value is Validation_mse: 0.35069696616549817. Going back to the original purpose of the model, this model can predict the sales of each item in the grocery store, thereby reducing the occurrence of excessive stockpiling of goods and not enough goods for customers.

According to the accuracy of the model, the current model can do its job. Because the purchase of grocery items does not need to be accurate to single digits, it only needs to predict a percentage value to complete the work. In a broader perspective, this model can be used not only for pre-sales of grocery store sales but also for other purposes. For example, this model can be used in supermarkets, in some economic sectors, etc.

4. DISCUSSION

4.1. Experimental process

At the beginning of the experiment, it was found that this data set is not suitable for model training due to the limited computing power of personal computing. Such a large dataset cannot be processed. (The original dataset was over 100 million columns) After discovering this problem, a decision was made to pick the most suitable data in this dataset, (selected according to the 2.2 data analysis section) to be able to complete the model building. The data set is too large, so the author chooses the LGBM model to make predictions because public datasets show that LightGBM can accelerate the training process by up to over 20 times while achieving almost the same accuracy [6]. However, the disadvantage of LGBM is that it is prone to overfitting. Therefore, in the parameter adjustment part of the model later, the parameters are adjusted according to the situation that overfitting is less likely to occur.

4.2. The way of improving results

In the construction of this model, there is a lack of comprehensive data analysis. For example, oil prices in the dataset, and regional factors, were not factored into the model's characterization. This time, only the impact of intuitive factors on the data, such as the impact of the periodicity of time on commodities, is considered. However, there is a lack of external large-scale influencing factors. For example, the price of oil is an important influencing factor for a country [7]. In addition, early stopping [8] can be added to the program. Adding early stopping can reduce the occurrence of overfitting.

4.3. The comparison with other models

Different ideas were used in the programs that predict sales. It is a model from a different idea. The solution is based on a three-level model. The first level used many single models, most of which were based on the XGBoost machine-learning algorithm [9]. The second stacking level used two models from Python scikit-learn package —ExtraTree model and the linear model from, as well as the Neural Network model. The results from the second level were summed with weights on the third level [10]. Although this model can also predict grocery

store sales, the model does not consider the relationship between the various features.

5. CONCLUSION

This paper experiments on grocery sales forecasting and obtains a model that can be used to predict sales. In this paper, the Lgbm algorithm is used, and the time series is used to divide the data into different periods. The model is established based on the above algorithm. Finally, using the MSE measurement method, the final accuracy of the model is 0.35069696616549817. However, the disadvantage of this model is that training requires a very large data set, but not all conditions can meet this requirement. In the case of fewer data, the author hypothesized that two problems occur. The training error is much smaller than the test error; the training error is similar to the test error, but the test error is very large.

The occurrence of the first situation indicates that the model has over-fitted, which means that the number of samples of the model is not enough. But in general, it is impossible to supplement the sample. At this time, dimensionality reduction can be used to solve such problems, such as PCA, LDA. The occurrence of the second situation proves that the model is too simple (not enough complexity). Therefore, more complex models can be used, such as models with hidden variables, Neural Network, ensemble-like methods GBDT, Random Forest, etc.

REFERENCES

[1] Zhang, G. P. Business forecasting with artificial neural networks: An overview. Neural networks in business forecasting, 2004, 1-22.

[2] Shen, S., Jiang, H., & Zhang, T. Stock market forecasting using machine learning algorithms. Department of Electrical Engineering, Stanford University, Stanford, CA, 2012, 1-5.

[3] Corporación Favorita Grocery Sales Forecasting, competition from the Kaggle community.

[4] El Naqa, I., & Murphy, M. J. What is machine learning?. In machine learning in radiation oncology (pp. 3-11). 2015, Springer, Cham.

[5] About Omar, K. B. (2018). XGBoost and LGBM for Porto Seguro's Kaggle challenge: A comparison. Preprint Semester Project.

[6] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

[7] Campbell, C J. The coming oil crisis. The United Kingdom.

[8] Prechelt, L. Early stopping-but when? In Neural Networks: Tricks of the trade (pp. 55-69). 1998, Springer, Berlin, Heidelberg.

[9] Kaggle Competition 'Grupo Bimbo Inventory Demand' #1 Place Solution of The Slippery Appraisals Team. Online: https://www.kaggle.com/c/grupo-bimbo-inventory-demand/discussion/23863 (accessed on 3 November 2018).

[10] Pavlyshenko, B. M. Machine-learning models for sales time series forecasting. Data, 4(1), 2019, 15.