Data Science and Engineering (DSE) Record, Volume 3, issue 1.

Analysis of Sales Influencing Factors and Prediction of Sales in Supermarket based on Machine Learning Technique

Jie Yang[1] and Sakgasit Ramingwong[2]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand

jie_y@cmu.ac.th

Abstract. There is a relatively well-established process for using machine learning to predict influencing factors and sales, but for small and medium-sized enterprises, they often face problems such as low data volume and unrepresentative data types, and the large data requirements become the threshold for using machine learning methods to help business activities. The original data for this study was sourced from publicly available data from Alibaba's Tianchi platform, containing sales data from a small shop in three different branches. This paper studies the influencing factors from the correlation of data and uses random forest regression method to rank the importance of features. In order to predict sales, this paper uses a pre-training model to compare and analyze multiple machine learning models. The results show that the pre-training method has different degree of improvement or decline for different models.

Keywords: influencing factors, prediction of sales, machine learning, pre-training

1 Introduction

At this stage, the development of Internet technology and computer hardware has generated a huge amount of data, and this huge amount of data has become the sustenance for the development of machine learning, which is often used in business and industry to predict customer behavior[1], product production cycles, and the expected amount of product sales.

There is a relatively well-established process for using machine learning to predict influencing factors and sales, but for small and medium-sized enterprises, they often face problems such as low data volume and unrepresentative data types, and the large data requirements become the threshold for using machine learning methods to help business activities.

2 Literature Review

2.1 Previous Studies Using Neuro-Fuzzy Technique

In this thesis, I use the correlation coefficient method to determine the five factors affecting sales volume, and use the machine learning method to rank the importance of features and get the ranking of influencing factors. When predicting sales volume, I use pre-training method to improve prediction accuracy of small-scale data, and compared 3 different models.

Objective

- The purpose of this paper is to explore a machine learning process that can be applied to a small sample size. Using machine learning methods for sales influencing factor analysis and sales forecasting.

This article is mainly based on the guiding idea of factor analysis method[2], combined with the feature selection principle of machine learning, using random forest regression and feature importance ranking method.

In the traditional business world, people usually combine different methods to forecast sales. For example, Ramanathan et al. used a linear regression model to analyze the various factors affecting market demand and select the most effective ones for sales forecasting[3].

In NLP domain, There are well-established pre-training methods for problems with small data volumes[4]. They can learn a wealth of linguistic knowledge from a large corpus . It is beneficial for downstream tasks.

In China, some scholars have constructed BERT-based pre-training models that have achieved good results in the field of computer vision[5].

3 Data and Methodology

3.1 Research framework

This thesis is divided into two main parts, one is the study of Influencing Factors and the other is the forecast of sales as illustrated in Figure 1.

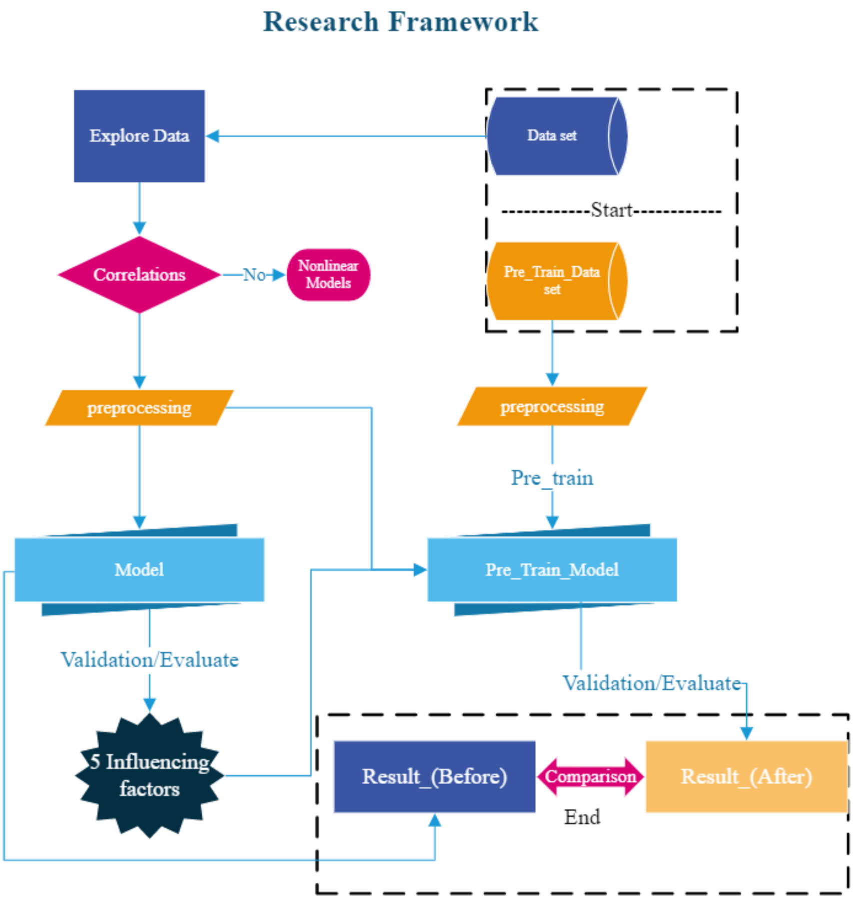


Figure 1 Research Framework

In this thesis, I first use Pearson's correlation coefficient to calculate the relationship between features, and then rank the source data with important information of random forest features to get the most important features to the model, i.e., the influencing factors; in terms of sales prediction, a pre-trained random forest model is constructed, and for the characteristics of small source data, I first use the data of relevant retail fields and train the random forest regression model, and then use the source data to fit the trained random forest regression model, and finally, I compare the results of the runs.

3.2 *Pre-processing*

Marketing data is usually a combination of continuous data and discrete data. In this thesis, for discrete features, I use Label Encoding to realize the mapping of features; for continuous features, I use normalization to scale the continuous feature data to the interval of [0,1].

3.3 *Evaluation indicators*

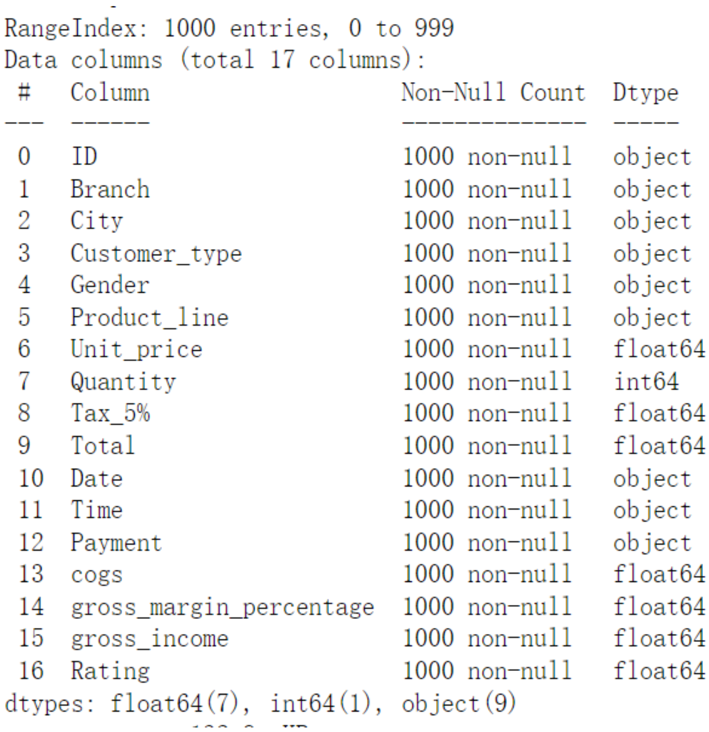
In this thesis, MSE, MAE, R^2 and EVS were chosen as indicators for the evaluation of the model. MSE and MAE were used to evaluate the error of the model predictions, and R^2 and EVS to assess the interpretability and accuracy of the model.

4 INFLUENCING FACTORS STUDY

This chapter starts from the source data, uses Pearson method to explore the correlation, and uses the machine learning method of random forest regression, the five influencing factors that affect the store sales are: COGs, gross_income, Tax_5%, Unit_price, Quantity.

4.1 *Data*

The supermarket data in this article comes from the public data of Alibaba Tianchi. Sales data includes the company's 3 branch numbers, customer types, product sales information, etc. They all faithfully reflect the company's operating conditions. The data characteristics are as follows:



```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 17 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   ID                       1000 non-null   object
 1   Branch                   1000 non-null   object
 2   City                     1000 non-null   object
 3   Customer_type            1000 non-null   object
 4   Gender                   1000 non-null   object
 5   Product_line             1000 non-null   object
 6   Unit_price               1000 non-null   float64
 7   Quantity                 1000 non-null   int64
 8   Tax_5%                   1000 non-null   float64
 9   Total                    1000 non-null   float64
 10  Date                     1000 non-null   object
 11  Time                     1000 non-null   object
 12  Payment                  1000 non-null   object
 13  cogs                     1000 non-null   float64
 14  gross_margin_percentage  1000 non-null   float64
 15  gross_income             1000 non-null   float64
 16  Rating                   1000 non-null   float64
dtypes: float64(7), int64(1), object(9)
```

Figure 2 Row Data

4.2 *Pearson correlation*

$$\rho X, Y = \frac{cov(X,Y)}{\sigma X \sigma Y} = \frac{E[(X-\mu x)(Y-\mu y]}{\sigma X \sigma Y}$$

Pearson product-moment correlation coefficient (PPMCC or PCCs) is a measure of the degree of correlation (linear correlation) between two variables X and Y, with a value between -1 and 1. **Error! Reference source not found.**In the natural sciences, this coefficient is widely used to measure the degree of linear correlation between two variables.

4.3 *Validation methods*

The cross-validation method makes full use of every piece of data, but is computationally huge. In this thesis, when studying the store dataset, the leave-one-out method is used for validation in order to maximize the use of the dataset; however, in the subsequent pre-training model, due to the relatively large dataset, this thesis uses the k-fold cross-test method.

4.4 *Random Forest regression*

Random forest is a classical machine learning algorithm proposed by Leo Breiman**Error! Reference source not found.**, which is composed of a weak model of classification and regression tree (CART) combined with Bagging method and random subspace method (RSM).

For the regression problem, the minimum mean squared deviation is used to partition the data set, and for any partition feature The corresponding partition points is divided into the left data set Dl and the right data set Dr on both sides, and the expression is:

$$min_{T,s}\left[\sum_{x_i \in D_l(T,s)} (y_i - c_i)^2 + \sum_{x_i \in D_r(T,s)} (y_i - c_i)^2\right]$$

where c1 and c2 are the mean values of the sample outputs of the data sets Dl and Dr. This principle is used to divide the data set at each branch until the threshold value is reached.

In this thesis, the random forest regression model will be mainly used in the influence factor study and sales forecasting.

4.5 *Experimental results*

According to the results of the Pearson correlation coefficient table, we can see that: Tax, COGs, and Gross-income all have a correlation coefficient of 1. It means that there is a strict linear correlation between these three features and the target feature.

Pearson correlation coefficient

	Total
Tax 5%	1.000**(p=0.0003)
Quantity	0.770**(p=0.000023)
Unit price	0.588**(p=0.00016)
cogs	1.000**(p=0.00009)
gross income	1.000**(p=0.00043)
Rating	-0.036(p=0.250)

* p<0.05 ** p<0.01

Table 1 Pearson Correlation Coefficient

The value of Rating is closed to 0. This indicates that there is almost no linear correlation between Rating and total sales, but it does not exclude the possibility that they possess a non-linear relationship with each other.

In addition, there is a correlation between Tax, COGs, gross income, Quantity and Unit price and total sales, based on this conclusion, we can perform regression methods to analyze the influencing factors.

LOOCV Evaluation Results

evaluation results	
Explained variance score(EVS)	0.5663033903621122
Mean absolute error (MAE)	167.9453475
Mean squared error (MSE)	54854.44141886966
Decidability factor (R² score)	0.11314313598303949

Table 2 LOOCV Evaluation Results

According to the evaluation index, the EVS of this model is 0.566. This result indicates that the model can explain the model in 56% of the cases. The MAE of the model is 167.9 and the mean square error is 54854, the error value shows that the actual prediction has a particularly large error with the original value, and the decidable coefficient R² is 0.11, which means that the independent variable has a very low degree of explanation of the dependent variable.

In this part of the study, the purpose is mainly to investigate that factors that have an impact on the model and do not use it to predict, therefore, the poor effect of the model is acceptable in the research on the factors that influence sales.

The results predicted by the model are compared with the true value results as in Figure 3.

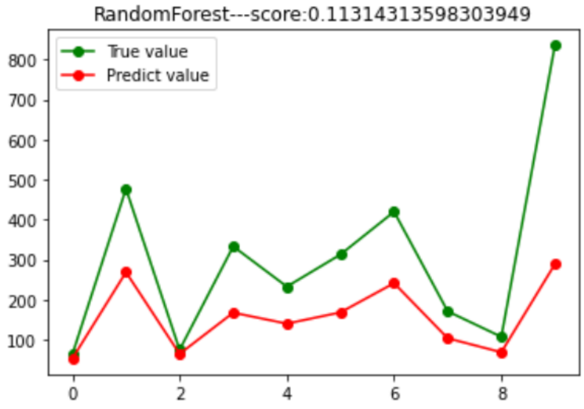


Figure 3 Comparison of Prediction Results
Abscissa: Point of prediction
Ordinate: corresponding value

In general, the predicted and true values show the same trend, which generates a large error that is acceptable in the case of low sample size.

The feature importance is visualized as in Figure 4.

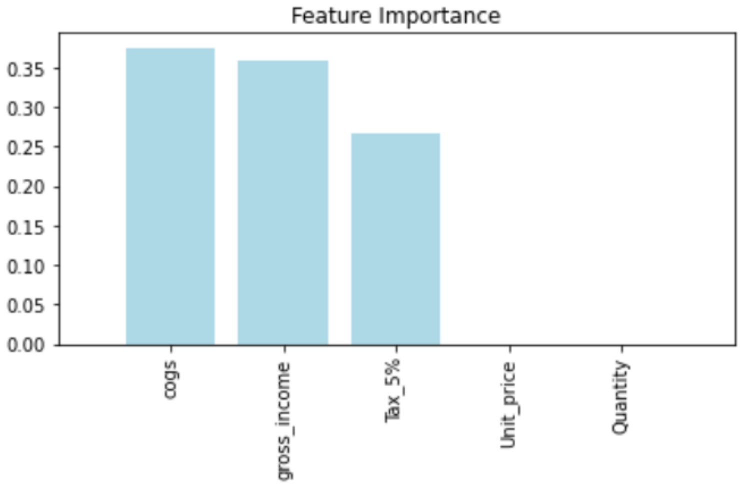


Figure 4 Feature Importance Sorting

COGs have the highest feature importance, occupying 0.37 importance, followed by gross_income, occupying 0.35 feature importance; Tax_5% feature chapter 0.26 importance, Unit_price and Quantity occupy 0.000057 and 0.000004 feature importance respectively, which is almost negligible.

It is concluded that the most important influencing factors are: COGs, gross_income, and Tax.

5 SALES PREDICTION

In this section, I propose a machine learning method based on pre-trained models to help us improve the effectiveness of model training in response to the problem of low data volume and insufficient samples of the original data and insignificant prediction results.

In this thesis, I choose several typical machine learning models and also follow the random forest model in the previous section to compare the results of the pre-trained models with the evaluation results of the directly fitted models.

5.1 *Data*

The data is from the Internet**Error! Reference source not found.**. They are listed as follows:

Region \ Country \ Item Type \ Clothes \ Sales Channel \ Order Priority \ Order Date\ Order ID\ Ship Date \ Units Sold \ Unit Price \ Unit Cost Total Revenue \ Total Cost \ Total Profit

5.2 *Machine learning methods*

According to the observation of the source data, the original data may have non-linear regression. This article uses a common regression model(bagging, boosting, lasso)

In this section, we input each of the 5 features into the model and the pre-trained model, and compare the output results. There are two main steps in pre-training the model, firstly, the model (pre_train_model) is trained by using the external data (pre_train_data_set) of the relevant domain. Second, the features of the original data set (data_set) are put into the model (pre_train_model) for training. The compared model results are as in Table 3.

Bagging	Result_Before	Result_After
Explained variance score (EVS)	0.9999882204122948	0.9999970244374887
Mean absolute error (MAE)	0.5138741999999897	0.2302135500000015
Mean squared error (MSE)	0.6623182191623067	0.18006498231750193
Decidability factor (R^2 score)	0.9999881948568896	0.9999970187491573

Table 3 Comparison of Model Results(Bagging)

Since both the EVS and the decidability coefficient of the original model reached 0.99, there is almost no margin for improvement, therefore, it is not obvious that there is an improvement on the accuracy rate However, it can be seen that the MAE is reduced from 0.51 to 0.23 and the MSE is reduced from 0.66 to 0.18, which clearly shows that the accuracy of the pre-trained model for data prediction has been effectively improved and the error was effectively reduced.

Boosting	Result_Before	Result_After
Explained variance score (EVS)	0.9999370783270306	0.999963202071521
Mean absolute error (MAE)	1.4058926735814976	1.0886992387040877
Mean squared error (MSE)	3.53048750042591	2.2225631757607434
Decidability factor (R^2 score)	0.999937072680494	0.999963202071521

Table 4 Comparison of Model Results(Boosting)

After fitting the pre-trained model, although it is not obvious, both the interpretability and R^2 of the model got a small increase, the MAE from 1.40 to 1.08, and the mean square error decreased from 3.53 to 2.22. It is indicates that the prediction results for small data in stores are significantly improved under the pre-training method.

Lasso	Result_Before	Result_After
Explained variance score (EVS)	0.999999996413399	0.9992159038864156
Mean absolute error (MAE)	0.010483871547542902	28.03707506560381
Mean squared error (MSE)	0.00020212562711673486	830.0686706545088
Decidability factor (R^2 score)	0.999999996397318	0.9852048771044063

Table 5 Comparison of Model Results(Lasso)

After fitting the original model to the dataset, there is a small decrease in the interpretability of the model, and the mean square error and absolute error both increase in different magnitudes.

For the lasso model, using the pre-trained model has no improvement on the prediction. The results are not improved, which indicates that the pre-trained model is not always effective for prediction of small-scale data.

6 SUMMARY

In this thesis, I use machine learning methods (random forest, etc.) to obtain the factors that affect the sales of a small chain store, The original data was sourced from a small shop. In this thesis, I first explored the entire data, including the shape of the distribution, and the underlying information. I used the Pearson correlation coefficient method (for quantitative information), and the Kendall method (for definite class information) for correlation analysis.

Then, I put the five characteristics with the highest correlation into the random forest regression model to rank the importance (based on the Gini coefficient to determine

Cross Entropy), and obtained the ranking of the five factors affecting sales, they are COGs, gross_income, and Tax.

In the second phase of the research, this thesis puts five features into the pre-training model for training, and compares the model results with the original results. I compared the model results of bagging, boosting and lasso methods, and the study showed that the pre-training model can solve the problem of insufficient prediction accuracy and large errors in small-scale data.

However, in this study, the different models produced different results. the Bagging model showed the most effective lift, the boosting change was not particularly significant, and the lasso model showed greater error and lower accuracy.

7 DISCUSSION

The category variable does not produce a greater effect, and conventional machine learning methods are much less sensitive to its impact.

The features of the pre-trained database are not necessarily applicable to supermarket data. We need more arguments to be sure that it can be used for Transfer learning.

The scope of the study is so large that it cannot draw definitive research conclusions in one area, but only as an empirical study, and if there are other researchers interested, I suggest narrowing the scope of the study.e.g., a study on supermarket sales prediction based on pre-training methods (using random forest)

In the following research process, I recommend using a more rigorous control variables method for comparison (subject to a precise study scope). We can compare data from different domains, different machine learning models, different evaluation metrics, etc.

According to the results, the research method in this thesis has certain effects, and the research process and method can be applied to other types of data. However, the relationship between the two data sets is not clear, that is, we are not sure that they can learn a wealth of linguistic knowledge from a large corpus. This approach should therefore be used with greater caution.

References

1. The AI-powered enterprise: Unlocking the potential of AI at scale
2. Qin Yinbing. An empirical study on the comprehensive competitiveness of China's pharmaceutical retail industry based on factor analysis [J]. Shanghai Pharmaceutical,2016,37(19):67-71.
3. Ramanathan U. Supply chain collaboration for improved forecast accuracy of promotional sales[J]. International Journal of Operations & Production Management, 2012, 32(6): 676-695.
4. Xu Shiwei . SVR-based method for predicting the sales volume of invisible eyeglasses [J]. Journal of Wenzhou Institute of Vocational Technology,2018,18(03):51-54.

5. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT.
6. https://blog.fearcat.in/a?ID=01600-6aba15bf-a618-4604-a9c7-155bcbb11e5a
7. BREIMAN L.Random forests[J].Machine Learning,2001,45(1):5-32.
8. https://eforexcel.com/wp/downloads-18-sample-csv-files-data-sets-for-testing-sales/
9.