2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE 2021)

# Sales Forecasting Based on LightGBM

Tingyan Deng[1st]
Vanderbilt University
Nashville, United States
tingyan.deng@vanderbilt.edu

Yu Zhao[2nd]
Northeastern University
Shenyang, China,
zainzhao@outlook.com

Shunxian Wang[1st]
Chinese University of Hong Kong, Shenzhen
Shenzhen, China
shunxianwang@link.cuhk.edu.cn

Hongjun Yu[3rd]
Beijing Foreign Studies University
Beijing, China
yuhongjun_apply@163.com

2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE) | 978-1-7281-8319-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCECE51280.2021.9342445

**Abstract— The combination of data science and machine learning is making sales forecasting possible. This will help improve the competitiveness of retail companies. This paper is based on the LightGBM framework, which is an improved GBDT model to realize Wal-Mart sales fore-casting. Large amounts of data require preprocessing so feature engineering is performed in this paper. First remove some features that are not related to the model input. Then the features are extracted and classified, and the mean, standard deviation and other statistics of some features are obtained. Experiments results show that our method has an RMSE of 0.641, which is significantly better than Logistic Regression (0.803) and SVM (0.732). In addition, this paper also shows the 20 top feature importance. This is of great significance for guiding the company's sales**

***Keywords—Sales forecasting, LightGBM,*** **Feature engineering**

## I. Introduction

Sales forecasting is to predict sales in the future based on the sales data recorded in the past period. It is of great significance for companies on the retail chain to forecast the sales of goods. Accurate forecasts can help companies optimize investment, reduce inventory costs, increase sales and profits, and avoid risks.

Sales forecasting has a long history. Now, emerging technologies make it possible to record the data of goods sales process. In addition, the advancement of algorithms can help researchers extract valuable knowledge from big data, and then make accurate predictions and guide the production and sales of products.

### *A. Related Work*

Generally, the models to solve this problem are all regression ones. That is, the mapping relationship between the recorded data and the sales is first fitted, then substitute the future data into this mapping to predict future sales.

In [1], Y. Liu et al. proposed a novel demand forecasting model named WFSSVM (Wrapper Feature Selection optimized SVM) is proposed. GA based wrapper feature selection method is firstly employed to analyze the sales data of a kind product. Then, the selection result is applied to build SVM regression model. In [2], Z. Li et al. introduces the basic principles of BP artificial neural network, and constructs a sales forecasting model based on the theory of BP. Their experiments results showed a highly degree.

Gradient Boosting Decision Tree (GBDT), which has a great connection with the paper of Friedman et al. [3]. It can be said that GBDT was developed from [3][4][5]. In [3], Friedman J. H. thinks function approximation is viewed from the perspective of numerical optimization in function space, rather than parameter space. A general gradient descent "boosting" paradigm is developed for additive expansions based on any fitting criterion. Gradient boosting of regression trees produces competitive procedures for both regression and classification.

In [6], a machine learning project named XGBoost was open sourced by Chen T. et al. It implemented GBDT efficiently and made many improvements in algorithm and engineering. Now it was widely used in many ML competitions and achieved good results. But for each data feature, XGBoost needs to scan all the dataset to estimate split points, which is very time consuming. To tackle this problem, in [7], Ke G et al. propose two novel techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). A series of optimizations to GBDT in [7] constitute the framework LightGBM.

### *B. Our Contribution*

With reference to GBDT and the implementation of LightGBM mentioned above, this paper applies LightGBM to sales forecasting. The dataset used in this paper is Walmart retail goods. The data, covers stores in three US States (California, Texas, and Wisconsin) and includes time, item level, depart-ment, product categories, and store details. In addition, it has explanatory variables such as price, promotions, day of the week, and special events. Together, this robust dataset can be used to improve forecasting accuracy.

This paper trains the model based on the data of the first 1413 days in the data set, and then predicts the sales for the next 28 days. The first is the pre-processing of the dataset. The dataset contains some non-numerical data, such as "category" and "id". They need to be converted into numerical values before they can be applied to the model. Then perform feature engineering. In addition, when the regression tree is being generated, features should be selected according to the gradient of loss function. This paper removes certain tag-like features such as “state_id” and “store_id”, because these features are used to distinguish stores instead of training models. Then use

978-1-7281-8319-0/21/$31.00 ©2021 IEEE

Recursive Feature Elimination (RFE) to determine the optimal number of features through cross-validation. Some features are discarded in each training until the number of features left meets our settings.

Comparing the prediction results obtained by LightGBM in this paper with Logistic Regression and SVM, we can see that the results obtained by LightGBM are more accurate.

The remainder of this paper is organized as follows. Section II introduces GBDT and several optimi-zation strategies of LightGBM to the GBDT model. Section III explains feature engineering, and Section IV displays experimental results and analyzes. Finally, Section V concludes.

## II. FEATURE ENGINEERING

The data set provides a lot of information. In order to explain more intuitively, the data hierarchy is shown in Figure 1. This paper performs feature engineering to improve the training and calculation speed, and at the same time make the performance of the model better.

For saving memory cost, the data type should be adjusted. The floating-point numbers in original dataset take up a lot of memory. In this paper, these data are compressed into a smaller floating-point number type.

Time period has a great influence on sales fore-casting, so the more detailed the time feature extraction, the better. A lot of information about time can be extracted from the timestamp. For example, through processing, it can be determined that which week, month, quarter a certain day is in, and whether it is a weekend. You can also get the lunar date of a certain day through conversion.

Two types of statistical characteristics are calculated in program, 1) rolling characteristics, statistical characteristics within a certain period; 2) lag cha-racteristics, characteristics ending at a certain time point, such as the characteristics of the previous 7 days, the previous 14 days, and the previous 21 days. For commodity prices, statistics such as maxi-mum, minimum, and median value are calculated.

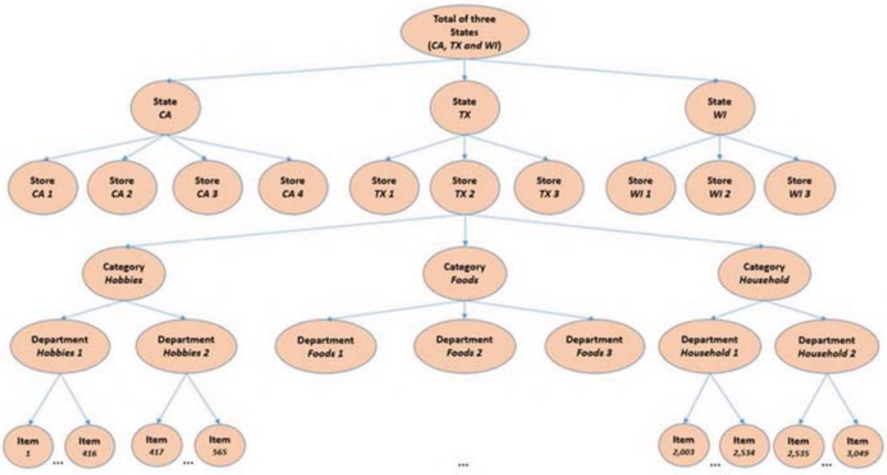

Figure 1. Data hierarchy

## III. Implementation of GBDT Based on LightGBM

### A. GBDT

GBDT is based on the idea that for a complex task, the judgment obtained by multiple experts is better than the judgment of any one of the experts alone. Let the dataset be denoted as $\mathrm{D} = \{(\mathrm{x}_1, \mathrm{y}_1), (\mathrm{x}_2, \mathrm{y}_2),$

$\ldots, (\mathrm{x}_\mathrm{n}, \mathrm{y}_\mathrm{n})\}$. The GBDT algorithm can be regard as an additive model composed of K regression trees:

$$\hat{y}_i = \sum\nolimits_{k=1}^{t} f_k(x_i), \qquad f_k \in F \qquad (1)$$

Where F is the function space composed of all trees. A regression tree corresponds to a function f. Different from general machine learning algorithms, this additive model does not learn the weights in a high-dimensional space, but directly learns a set of functions (regression trees). Because what is learned is an additive model, if you can learn only one basis function at each step from the front to the back, and gradually approximate the optimization objective function, then the complexity can be simplified. This learning process is called Boosting. The specific process is as follows:

$$\begin{gathered} \hat{y}_i^0 = 0 \\ \hat{y}_i^1 = f_1(x_i) = \hat{y}_i^0 + f_1(x_i) \\ \hat{y}_i^2 = f_1(x_i) + f_2(x_i) = \hat{y}_i^1 + f_2(x_i) \qquad (2) \\ \cdots\cdots \\ \hat{y}_i^t = \sum\nolimits_{k=1}^{t} f_k(x_i) = \hat{y}_i^{t-1} + f_t(x_i) \end{gathered}$$

The objective function of the above additive model is defined as:

$$Obj^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^t) + \Omega(f_t)$$
$$= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \qquad (3)$$

Where $\Omega(\cdot)$ represents the complexity of the deci-sion tree, such as the number of nodes in a tree, the depth of a tree, and so on. This regular term limits the complexity of the model and helps avoid overfitting.

The objective function is a function of variable $(\hat{y}_i^{t-1} + f_t(x_i))$. Using Taylor's formula to expand the objective function in the second order at $\hat{y}_i^{t-1}$, the following equation can be obtained:

$$Obj^t = \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \qquad (4)$$

$g_i$ is the first derivative of the loss function, and $h_i$ is the second derivative of the loss function.

Summarize the learning process of GBDT:

1) Before the start of each iteration, calculate the first derivative and second derivative of the loss function at each training sample point.

2) According to the greedy strategy to generate a new decision tree, the predicted value corresponding to each leaf node is calculated by Equation group (2).

3) Add the newly generated decision tree to the model.

4) Recursively execute 1-4 until certain conditions are met.

### *B. LightGBM*

LightGBM is a framework for implementing GBDT, which has the advantages of fast training speed, low memory consumption, and high accuracy.

LightGBM performs gradient calculation based on Histogram. Establish a statistical histogram for each feature, and find the optimal split point according to the discrete value of the histogram. As shown in Figure 1. The histogram of a node can be obtained from the histogram of its parent node and the histogram of its siblings, further speeding up the calculation. Based on Histogram, LightGBM adopts Leaf-wise growth strategy to create regression trees. As shown in Figure 2.

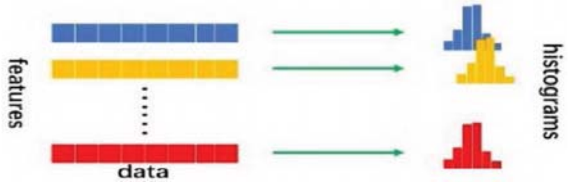

Figure 2. Establish statistical histograms

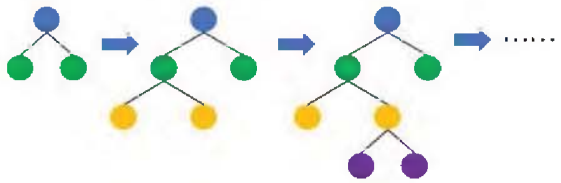

Figure 3. Leaf-wise growth strategy

LightGBM uses Gradient-based One-Side Sampling to sample and assign a weight to every sample. In addition, high-dimensional data is often sparse, and this sparsity is used to design a lossless method to reduce the dimensionality of features.

## IV. EXPERIMENTS

In order to verify the effectiveness of the model, we conducted experiments. The experiments results are shown in Table 1. The comparison index is Root Mean Square Error (RMSE). It is a commonly used index for machine learning. Calculated as follows:

$$e = \sqrt{\frac{1}{28}\sum_{i=1}^{28}(F(x_i') - y_i')} \qquad (5)$$

Where $F(x_i')$ is the output of our model. B is the ground truth.

TABLE 1 PERFORMANCE COMPARISON OF MODELS

| Models | RMSSE |
|---|---|
| Logistic Regression | 0.803 |
| SVM | 0.732 |
| LightGBM | 0.641 |

Obviously the model has obvious advantages in sales forecasting compared with two other ML algorithms and reduced the RMSE by 9-16 percen-tage points.

We output the relative importance of the features. According to the Equation (4) and the theory of Information Theory, the greater the gradient of the loss function to a feature, the more information the feature provides, the more important it is. Sorting the importance of features according to the gradient is shown in Figure 4. This is of great significance for guiding the company's sales

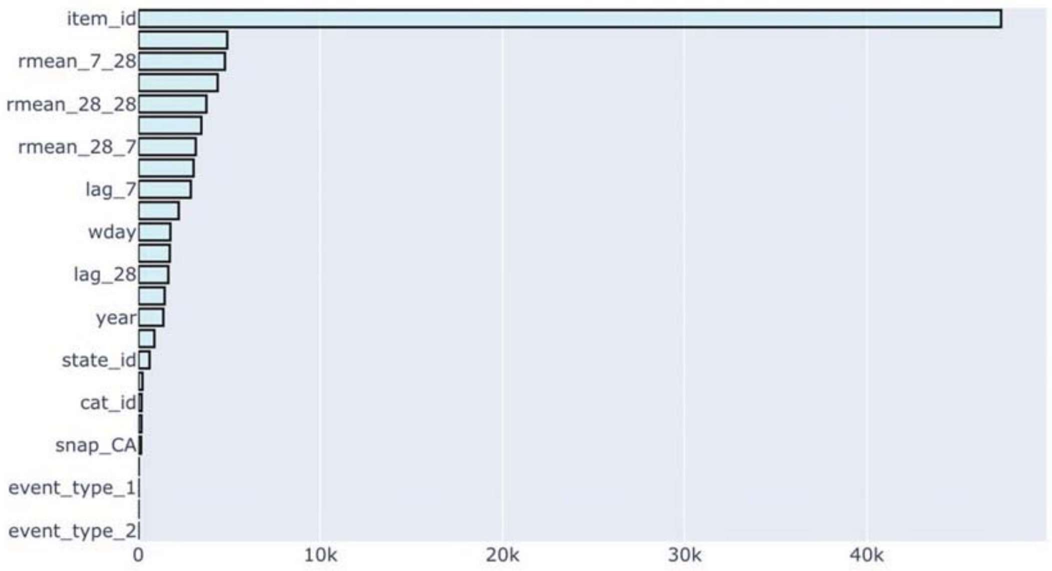

Figure 4. Feature importance ranking

## V. CONCLUSIONS

In this paper, we propose a method based on LightGBM to forecast sales of Walmart. In Section II, we introduce GBDT and framework LightGBM to the GBDT model. Section III shows feature engineering which processes data through several ways. And Section IV performs experiments. Experiments show that our model performs much better than traditional machine learning methods Logistic Regression and SVM.

## ACKNOWLEDGEMENT

The author Tingyan Deng & Wang Shunxian thanks Yu Zhao for the help in experiments, thanks Wang Shunxian for the contribution of the idea. In addition, we also thanks Hongjun Yu for fruitful discussions. Finally, we sincerely thanks to the free resources provided by Kaggle platform.

## REFERENCES

[1] Liu Y, Yin Y, Gao J, et al. Wrapper feature selection optimized SVM model for demand forecasting[C]. 2008 The 9th International Conference for Young Computer Scientists. IEEE, 2008: 953-958.

[2] Li Z, Li R, Shang Z, et al. Application of BP neural network to sale forecast for H company[C]. Proceedings of the 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE, 2012: 304-307.

[3] Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001: 1189-1232.

[4] Friedman J H. Stochastic gradient boosting[J]. Computational statistics & data analysis, 2002, 38(4): 367-378.

[5] Schonlau M. Boosted regression (boosting): An introductory tutorial and a Stata plugin[J]. The Stata Journal, 2005, 5(3): 330-354.

[6] Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting[J]. R package version 0.4-2, 2015: 1-4.

[7] Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting[J]. R package version 0.4-2, 2015: 1-4.

[8] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]. Advances in neural information processing systems. 2017: 3146-3154.

[9] Jain A, Menon M N, Chandra S. Sales forecasting for retail chains[J]. 2015.

[10] Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001: 1189-1232.

[11] Thomassey S, Fiordaliso A. A hybrid sales forecasting system based on clustering and decision trees[J]. Decision Support Systems, 2006, 42(1): 408-421.