Forecasting Tobacco Sales Trends Based on Combined Xgboost-Lightgbm Modeling

Ruochen Feng [1, #, *], Yunlong Xia [2, #], Weiwen Huang [3, #]

[1] School of Optics and Electronic Information, Huazhong University of Science and Technology, Hubei, China, 430070

[2] School of Electrical Engineering and Information Technology, Lanzhou University of Technology, Ganshu, China, 730050

[3] School of Economics, Huazhong University of Science and Technology, Hubei, China, 430070

* Corresponding Author Email: u202214834@hust.edu.cn

[#]These authors contributed equally.

Abstract. In the contemporary market landscape, consumers increasingly prioritize cigarette products' safety and pricing transparency. This study develops multiple time-series forecasting models to accurately predict five cigarette brands' sales volume and revenue. By integrating advanced data preprocessing techniques and feature engineering, the proposed framework addresses the challenges posed by extreme sales values and macroeconomic fluctuations. Considering that the sales in some months are "extremely large" or "extremely small", this paper introduces the Savitzky-Golay filtering algorithm for data smoothing. One of the major innovations of this paper is that in addition to product sales and average prices, other factors such as the rate of price change and household consumption index are also considered. In order to analyze the importance of the impact of different factors on sales, this paper introduces the SHAP algorithm to rank the degree of characteristics and finally selects the top three factors as the indicators of sales forecast. In this paper, two efficient gradient-boosting decision tree algorithms, XGBoost and LightGBM, are used for combined prediction, and the average of the two predictions is used as the final prediction result. The final MSE of the two cigarette brand predictions are 0.0042 and 0.0036, and the goodness of fit is 0.9056 and 0.9028, respectively, and this combination model has high accuracy.

Keywords: Tobacco Sales, XGBoost-LightGBM, Combined Prediction Coefficient of Determination.

1. Introduction

The tobacco industry constitutes a pivotal sector in China's economy, characterized by its vast market scale and complex regulatory landscape. Annual tobacco sales exhibit significant volatility influenced by multifaceted factors, including regional festivals (e.g., Spring Festival), evolving consumer preferences, stringent national tobacco control policies (e.g., advertising bans and tax hikes), and macroeconomic fluctuations such as inflation and household consumption trends. [1-3] This volatility underscores the necessity for accurate sales forecasting to optimize production planning, inventory management, and policy compliance. However, the dynamic interplay of these factors poses substantial challenges for traditional prediction models.Existing literature has explored diverse methodologies for tobacco sales forecasting. For instance, Liang Shangjian [1] employed ARIMA models to analyze pandemic-induced market disruptions, while ChenHong. [2] investigated macroeconomic impacts on regional tobacco enterprises. Recent advancements in machine learning have introduced algorithms like Long Short-Term Memory (LSTM) [4] and Genetic Algorithm-optimized Backpropagation Neural Networks (GA-BPNN) [5] to capture nonlinear temporal patterns. Despite their theoretical promise, these approaches exhibit critical limitations: LSTM models, while effective in sequence modeling, often suffer from high computational costs and sensitivity to hyperparameter tuning [4]; GA-BPNN, though capable of global optimization, struggles with local minima and overfitting in noisy datasets [5]. Consequently, reported prediction errors (e.g., MSE > 0.01, RMSE > 0.1) remain suboptimal for practical deployment.

To address these gaps, this study proposes a novel hybrid framework integrating XGBoost and LightGBM, two state-of-the-art gradient-boosting algorithms. Unlike prior methods, our approach synergizes XGBoost's regularization-driven robustness with LightGBM's computational efficiency for large-scale time-series data.

Experimental results demonstrate significant improvements over existing methods, achieving MSE values as low as 0.0036 and R^2 scores exceeding 0.90 for two major tobacco brands. This advancement not only enhances forecasting accuracy but also provides actionable insights into consumption patterns, enabling stakeholders to anticipate market shifts and align strategies with regulatory and economic dynamics.

2. Exploration of Tobacco Sales Data

2.1. Data Sources

The data comes from: http://www.nmmcm.org.cn/match_detail/33.

The data selected in this paper are the monthly sales volume (boxes) and monthly sales amount (CNY) of tobacco for Brand One from June 2014 to November 2022; and the monthly sales volume (boxes) and monthly sales amount (CNY) of tobacco for Brand Two from January 2011 to March 2022. The data selected in this paper are the monthly sales volume (boxes) and monthly sales amount (CNY) of tobacco for Brand One from June 2014 to November 2022; and the monthly sales volume (boxes) and monthly sales amount (CNY) of tobacco for Brand Two from January 2011 to March 2022. Due to space constraints, only partial data are presented here, as shown in Table.1. and Table.2. below.

Table.1. Presentation of Data Sources, Data of band one

Month	Sample ID	Name	Sales (Boxes)	Amount (CNY)
201406	42050329	A3 (Hard)	6.596	118728
201407	42050329	A3 (Hard)	24.372	438696
...	...	...	...	...
202210	42050329	A3 (Hard)	75.068	1820399
202211	42050329	A3 (Hard)	175.724	4261307

Table.2. Presentation of Data Sources,Data of brand two

Month	Sample ID	Name	Sales (Boxes)	Amount (CNY)
201101	52010212	A4(Longmach)	495.268	5571765
201102	52010212	A4(Longmarch)	421.96	4645800
...	...	...	...	...
202109	52010212	A4(Longmarch)	825.768	9847283.4
202203	52010212	A4(Longmarch)	1005.364	11988965.7

2.2. Data preprocessing

2.2.1 Savitzky-Golay filtering algorithm

Savitzky-Golay(S-G) filter algorithm is a low-pass filter designed by Savitzky and Golay, which is characterized by simple parameters, easy adjustment, good filtering effect, etc. The selection of its fitting order and filtering window length largely determines the filtering effect [6]. If the filter window length is chosen to be larger, the smoothing effect is more obvious, and vice versa, it is closer to the original curve; the smaller the polynomial fitting order is chosen, the smoothing effect is more obvious, and vice versa, it is closer to the original curve. The S-G filter minimizes the residuals between the fitted data and the real data, and its basic principle is as follows [7].

First, a polynomial is fitted to a set of discrete points centered at n=0:

$$p(n)=\sum_{k=0}^{N}a_k n^k=a_0+a_1 n+a_2 n^2+\ldots+a_k n^k \tag{1}$$

Where N is the polynomial order, a_k is the polynomial number, n is the discrete sequence value of the point, and $p(n)$ is the fitted value of the point.

Next, find the sum of the least squares fitted residuals of all the fitted data within the filter window to the original data ε_N :

$$\varepsilon_N=\sum_{x=-M}^{M}(p(n)-x(n))^2=\sum_{x=-M}^{M}(\sum_{k=0}^{N}a_k n^k-x(n))^2 \tag{2}$$

When ε_N taking the partial derivation of the above equation to get the polynomial fitting coefficients a_k , the fitting result is computed and assigned to the endpoint of this window as the output of the fitting result, i.e.:

$$y(0)=p(0) \tag{3}$$

The filter will be shifted along the time axis without panning, repeat the above steps until the end of the data, and finally realize the S-G shift smoothing. For this question, due to the presence of numerous spikes in A3 and A4 sales data, the original data smoothness is poor, in order to enhance the smoothness of the data, so set the Savitzky-Golay filter set the fitting order of 3, the length of the filtering window is 5. Take A3 (hard) as an example, the sales, sales by smoothing the effect as shown in Figure1, Figure2.

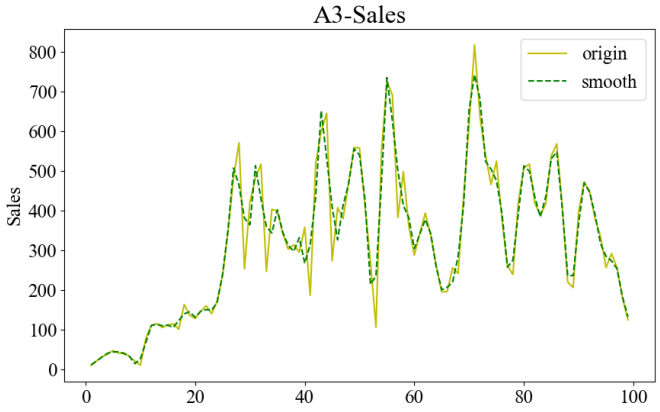

Figure 1 Sales of A3 (Hard) after smoothing

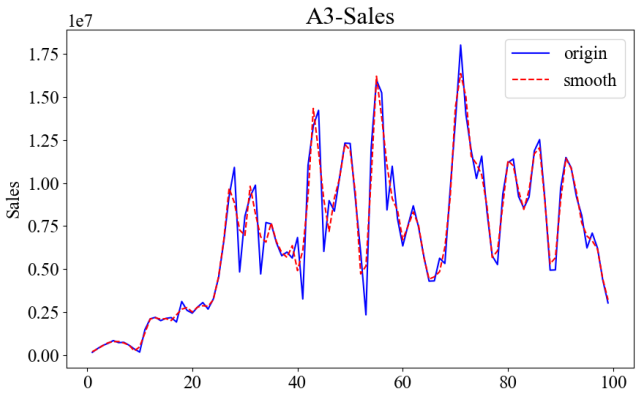

Figure 2 Smoothed A3 (Hard) Sales

By replacing the original data with the smoothed values and removing the "spikes", the deleterious effects of outliers on the model training results can be mitigated to a certain extent.

2.2.2 Feature engineering extraction

The requirement to forecast the number of sales of the A3, A4 cigarette brands, again requires the construction of two different types of time series models. The problem is essentially similar to the first one, but the forecasting objective has changed from volume to sales, so some new considerations need to be introduced. In addition to the time-related features, the creation of some derived features such as average price, rate of price change, etc. needs to be considered. Meanwhile, after checking in the National Bureau of Statistics, this paper also introduces some macroeconomic indicators as external features, such as consumer price index, and retail price index, which may have an impact on the amount of tobacco sales.

2.2.3 Calculation of average price and rate of price change

The price of a commodity may be affected by the average price and the rate of change of the price, and it is first necessary to calculate the flat A3, and A4.

The average price and the rate of price change, are calculated as follows:

$$\text{The average price of the month} = \frac{Current\ Month\ Sales}{Current\ Month\ Sales\ Volume}$$

$$\text{Rate of price chage} = \frac{price\ of\ current\ month - Price\ of\ previous\ month}{price\ of\ previous\ month}$$

The average price change for A3 and A4 was calculated as shown in Figure3. As can be seen from the figure, the average price of A3 brand and A4 brand is on the rise.

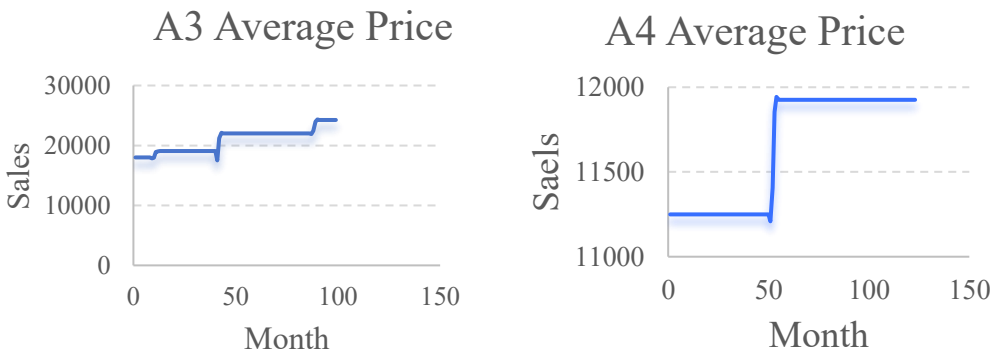

Figure 3 Average price of A3 (left) and A4 (right)

Similarly, a graph of the rate of price change for the A3 (Hard) and the A4 (Long March) can be obtained in Figure4. As can be seen from the chart, the prices of the two brands of cigarettes only change within a certain month.

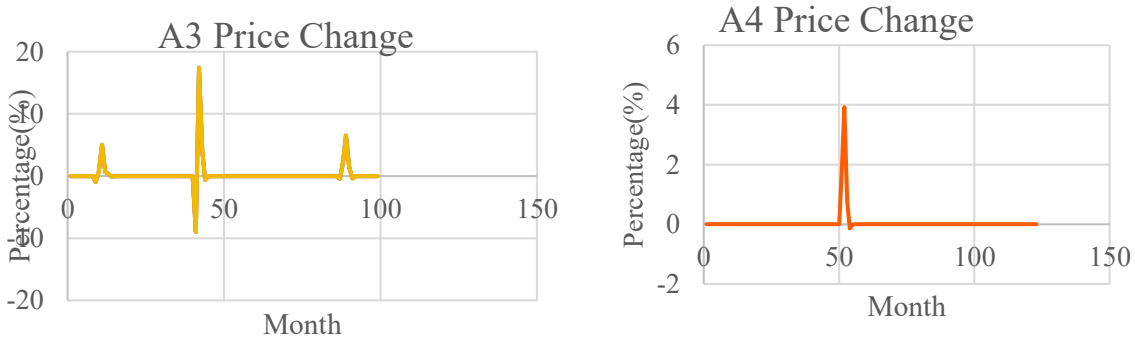

Figure 4 Rate of change in price of A3 (left) vs. A4 (right)

2.2.4 SHAP identifies the most influential features for prediction

SHAP is an algorithm for interpreting individual predictions, which is based on game theory and local interpretation, and is used to provide interpretability to the output of the model. The core idea is the marginal contribution of each feature in the game-theoretic framework, and the contribution of each feature is evaluated by calculating the contribution of the individual in the cooperation to determine the importance of that individual. Although some traditional interpretability methods can intuitively reflect the degree of influence of features on the model, they cannot reflect the relationship between features and the final prediction results, so the SHAP algorithm is widely used in the interpretability analysis of machine learning. For the XGBoost model (described in detail later in this paper), the input N (containing n features) is used to predict the output value $v(N)$, and in the SHAP algorithm, the contribution of each feature to $v(N)$ (φ_i is denoted as the contribution of feature i) is assigned based on its marginal contribution, which is computed by the formula [8]):

$$\varphi_i = \sum_{S \in N\{i\}} \frac{|S|!(n-|S|-1)}{n!}[v(S \cup \{i\}) - v(S)] \tag{4}$$

Sales volume, average price, and price change rate are used as evaluation indicators, and some macroeconomic indicators (retail price index and consumer price index) are also introduced as external indicators. In this paper, we analyze the characteristic importance of the above five indicators on product sales through SHAP and visualize the characteristic results to determine the impact of each indicator's characteristic on product sales intuitively. Due to space limitations, this paper takes A3 (Hard) as an example (see Appendix D for A4) and analyzes its SHAP chart. Where the SHAP feature importance heat map is shown in Figure5.

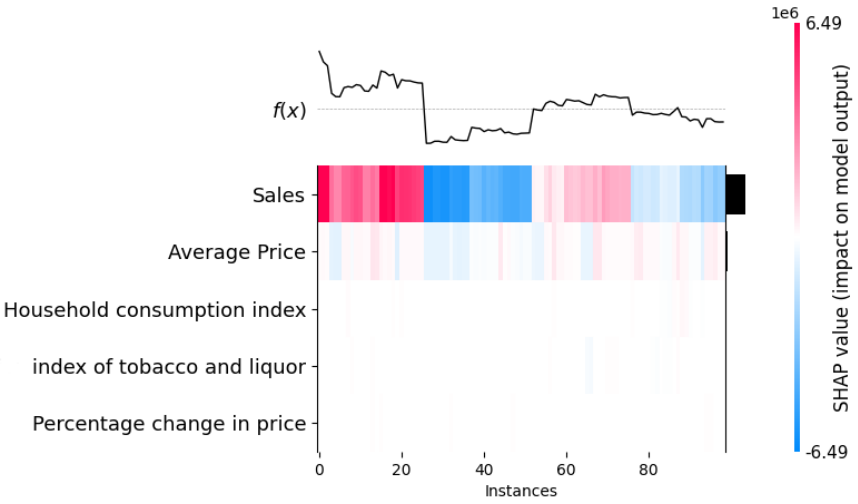

Figure 5 SHAP Feature Importance Heatmap

Based on the results of the above graph, it can be concluded that of the five indicators mentioned above, the value of sales volume has the greatest impact on sales, followed by the average price of goods and the consumer index. Both the Index of Tobacco and Liquor and the Percentage Change in Price show SHAP values close to zero, which has a very small impact on sales.

The above figure indicates the average of the absolute value of SHAP of each feature on the overall sample to indicate the importance ranking of the feature, it is not possible to know the positive or negative correlation between each indicator and sales or to analyze the impact of each indicator on the prediction results, therefore, it is necessary to make a beeswarm chart to make a more specific analysis of the features, as shown in Figure6.

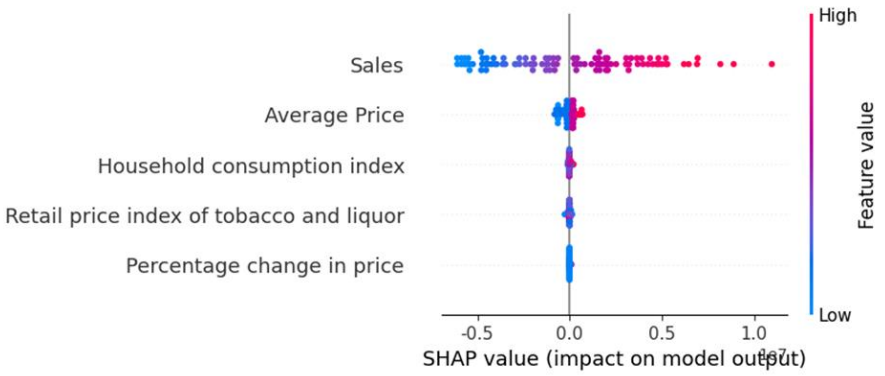

Figure 6 beeswarm diagram

As can be seen from the figure, there is a very strong positive correlation between volume and sales, a negative correlation between average price and sales, and a weak positive correlation between the household consumption index and sales. In contrast, there is almost no correlation between the retail price index for tobacco and alcohol and the percentage change in prices and sales.

Combining the importance ranking and relevance of features, this paper finally selects sales volume, average price, and household consumption index as the three features for predicting sales.

3. Tobbaco sales forecasting modeling

3.1. XGBoost prediction model establishment

XGBoost algorithm is an improved algorithm based on GBDT, which is an integrated learning method with a decision tree as a weak learner, with the advantages of easier training, higher interpretability, etc., and has a wide range of applications in the field of sales prediction, data analysis, etc. XGBoost is the first open-source system to support the training of the GBDT model in the GPU environment and is more accurate and more flexible than GBDT, and XGBoost offers better control over the learning rate. XGBoost is the first open-source system that supports training GBDT models in the image processing unit GPU environment. Compared with GBDT, XGBoost has higher accuracy, more flexibility, easier-to control-model complexity, and faster learning rate, and is widely used in various types of prediction and regression problems. Its core idea is to generate a decision tree by constantly splitting to fit the prediction residuals of the previous stage, the final prediction is the weighted sum of leaf node values.

Given a dataset with N samples, each independent variable serves as an input X_i , each variable contains a number of features, each sample corresponds to the variable y_i , and the predicted values of a model with K trees y_i are as follows [9]):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i), f_k \in F \tag{5}$$

K denotes the number of trees, F is the set space of regression trees, $f(x)$ is a regression tree, and $f_k(X_i)$ denotes the weight of the leaf in which the sample X_i is located in the kth tree.

$$F = \{f(X) = \mu_l(X)\} \tag{6}$$

Where $l(X)$ denotes the leaf node of the Xth sample and μ denotes the score of the leaf node. The prediction result after the tth iteration is shown below:

$$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(X_i) \tag{7}$$

Therefore, the objective function of XGBoost is formulated as shown in Equation (8).

$$J(f_t)=\sum_{i=1}^{n}(y_i,\hat{y}_i^{t-1}+f_t(X_i)+\Omega(f_t)) \tag{8}$$

where L is the loss function and $\Omega(f_t)$ denotes the complexity of the model.

$$\Omega(f_t)=\gamma T+\lambda\frac{1}{2}\sum_{j=1}^{T}w_j 2 \tag{9}$$

where T denotes the total number of leaf nodes, $w_j 2$ denotes the regularization term of the leaf node score $L2$, and γ denotes the penalty coefficient. From this, the objective function is linearized using a second-order Taylor expansion, yielding:

$$J(f_t)=\sum_{i=1}^{n}[g_i w_l(X_i)+\frac{1}{2}h_i w2_l(X_i)]+\gamma T+\lambda\frac{1}{2}\sum_{j=1}^{T}w2_j \tag{10}$$

$$g_i=\frac{\partial L(y_i,\hat{y}^{t-1})}{\partial\hat{y}^{t-1}} \tag{11}$$

$$h_i=\frac{\partial 2L(y_{i,}\hat{y}^{t-1})}{\partial\hat{y}^{t-1}} \tag{12}$$

Therefore, using the XGBoost algorithm can engage in better capturing the relationship between sales and other features and accurately predict the future sales of the product.

3.2. LightGBM prediction modeling

LightGBM is a gradient-boosting algorithm based on the gradient-boosting tree model. The principle of GBDT is to use gradient descent to generate a new tree based on all the previous trees and make the objective function value as small as possible. LightGBM is optimized on the basis of the GBDT algorithm [10]. By introducing innovative methods such as histogram algorithm, exclusive feature bundling, and leaf growth strategy with depth restriction, LightGBM can obtain faster training speed by occupying less memory. LightGBM's objective function is formulated as a combination of the loss function and regularization term. Giving the training dataset $T=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$, the loss function $l(y_i,\hat{y}_i)$, and the regularization parameter $\Omega(f_k)$, the objective function of LightGBM can be defined as follows:

$$L^t(\phi)=\sum_{i}^{n}l(y_i,\hat{y}_i)+\sum\Omega(f_k) \tag{13}$$

Where n is the total number of samples, y_i is the predicted value of the i th sample, f_k is the model of the kth tree, and $\sum\Omega(f_k)$ can represent the complexity of the k th tree.

Then the objective function can be transformed into the following form:

$$L^t(\phi)=\sum_{i=1}^{n}l\left(y_i,\hat{y}_i^{t-1}+f_t(x_i)\right)+\Omega(f_t)+Const \tag{14}$$

Here Const is a constant term, through the second-order Taylor expansion and the constant term, regularization expansion and removal the constant term, combined with the primary coefficients and quadratic coefficients can be obtained to the final objective function, the objective function of the second-order Taylor expansion:

$$L^t(\phi)=\sum_{i=1}^{n}\left[l(y_i,\hat{y}_i^{t-1})+g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t)+Const \tag{15}$$

Here $g_i=\partial\hat{y}^{(t-1)}l(y_i,\hat{y}_i^{(t-1)}), h_i=\partial^2\hat{y}^{(t-1)}l(y_i,\hat{y}_i^{(t-1)})$, the objective function is simplified further:

$$L^t(\phi)=\sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t)=\sum_{i=1}^{T}\left[(\sum_{i\in I_j}g_j)w_j+\frac{1}{2}(\sum_{i\in I_j}h_i+\lambda)w_j^2\right]+\gamma T$$

Let $G_j=\sum_{i\in I_j}g_j$, $H_j=\sum_{i\in I_j}h_j$, then the objective function is

$$L^t(\phi)=\sum_{j=1}^{T}\left[G_j w_j+\frac{1}{2}(H_j+\lambda)w_j^2\right]+\gamma T \tag{16}$$

The LightGBM algorithm first performs data preprocessing and feature engineering to convert the raw data into a format suitable for model training; then it constructs histograms and splits the data in preparation for gradient boosting and decision tree construction; then it improves the performance of the decision tree model by optimizing the loss function and fuses it into the final model; finally, it uses the final model to make predictions and evaluates the accuracy of the prediction results.

3.3. Model prediction results

After setting the model parameters, XGBoost and LightGBM, models are used for joint prediction and the prediction results of both are averaged. Separately, A3 February 2023 to November 2023 monthly sales, (Long March) and April 2022 to January 2024 monthly sales are predicted. the A3 prediction results are shown in Figure7.

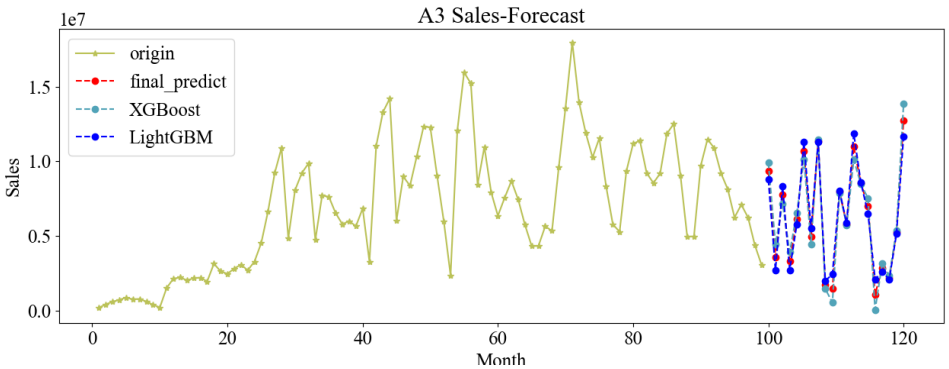

Figure 7 XGBoost and LightGBM Predictions A3 (Hard) Sales

The A4 (Long March) predictions are shown in Figure8.

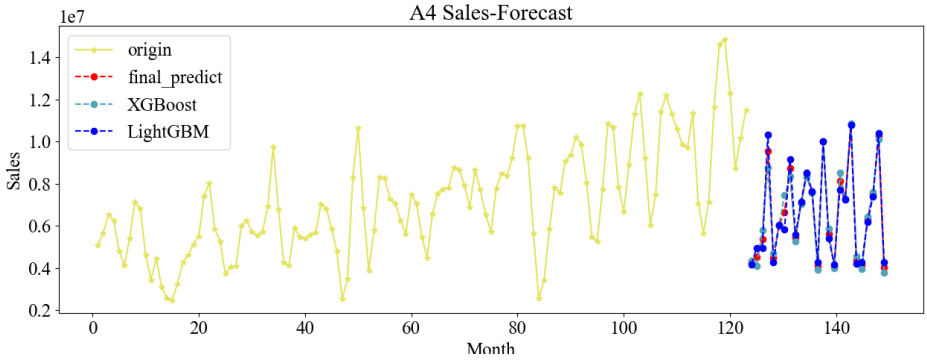

Figure 8 XGBoost vs. LightGBM Forecast A4 (Long March) Sales

Due to space constraints, only the final prediction results (i.e., the final result after averaging) are shown in the main text.

A3 (Hard) sales forecast for 2023.2 to 2023.11 is shown in Table.3.

Table.3. Prediction of Sales of A3

Month	Sales ($)	Month	Sales ($)
2023-02	9353877.81	2023-07	10712123.42
2023-03	3580342.39	2023-08	4972060.29
2023-04	7745705.83	2023-09	11379544.81
2023-05	3315494.17	2023-10	1723970.66
2023-06	6144789.08	2023-11	1468032.27

The results of A4 (Long March) sales forecast from 2022.4 to 2024.1 are shown in Table.4.

Table.4. Prediction of Sales of A4.

Month	Sales ($)	Month	Sales ($)
2022-04	4270017.84	2023-03	7607560.52
2022-05	4534070.98	2023-04	4107394.10
2022-06	5381174.18	2023-05	10000701.58
2022-07	9552551.78	2023-06	5633204.34
2022-08	4504847.55	2023-07	4084336.74
2022-09	6020225.78	2023-08	8122597.76
2022-10	6654908.31	2023-09	7278233.55
2022-11	8757548.53	2023-10	10836351.05
2022-12	5431636.65	2023-11	4391757.57
2023-01	7098763.38	2023-12	4130425.27
2023-02	8419812.22	2024-01	6316101.27

3.4. Testing of the model

The results of the joint prediction of LightGBM and XGBoost models are visualized. The horizontal axis represents the values predicted by the model, the vertical axis represents the labels of the real values, the transparency (i.e., alpha) value of each point is dynamically adjusted according to the size of the difference between the predicted value and the real value, and the larger the difference is, the lower the transparency of the point is, and the red dashed line represents the diagonal line where the predicted value is equal to the real value in the ideal case. The results of the model visualization are shown in Figure9.

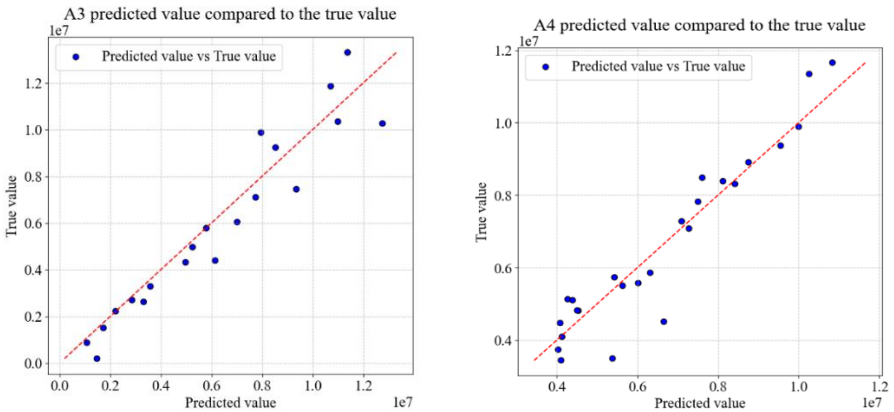

Figure 9 A3 (left) vs. A4 (right) model test

It can be seen that the scatter points are mostly distributed around the y=x straight line, then it represents a good robustness of the model.

To further evaluate the model, this paper introduces mean square error (MSE), root mean square error (RMSE), and mean absolute error (MAE) with the coefficient of determination (R-Squared) as the evaluation metrics, defined as follows:

MSE: Mean Square Error is a common statistic used to assess the difference between predicted and actual values. It is calculated as the square of the difference between each predicted value and the true value and then averaged. The higher the MSE, the greater the error of the prediction model.

RMSE: Root Mean Square Error is the square root of MSE, which is a more intuitive reflection of the size of the error, and it reflects the standard deviation form of the error.

MAE: Mean Absolute Error averages the absolute value of each prediction error and penalizes large errors less than RMSE because it focuses only on the absolute value of the prediction error and does not care about the direction.

R-Squared: also known as the coefficient of determination, usually used for goodness of fit.R_Squared is a value between 0 and 1, the closer to 1, the better the fit. However, a high R may lead to overfitting.

For A3(Hard) sales prediction, the joint prediction effect of the LightGBM and XGBoost model is shown in Table.5:

Table.5. The Evaluation of Prediction(A3)

MSE	RMSE	MAE	R-Squared
0.0042	0.0651	0.0505	0.9056

For the prediction of Long March sales, the joint prediction effect of LightGBM and XGBoost model is shown in Table.6.

Table.6. The Evaluation of Prediction(A4)

MSE	RMSE	MAE	R-Squared
0.0036	0.0600	0.0428	0.9028

It can be seen that the above model MSE, RMSE, and MAE values are very small, the prediction model error is small; and the R_Squared values were 0.9056, 0.9028, close to 1 but not 1, indicating that the degree of fit is good and there is no overfitting phenomenon, the model prediction accuracy is high.

4. Conclusion

This paper proposes a joint prediction model of XGBoost and LightGBM to predict tobacco sales. XGBoost offers robust regularization and superior prediction accuracy, whereas LightGBM excels in computational efficiency for large-scale datasets. Combining the advantages of the two prediction models, the average of their prediction results is taken as the final prediction value. According to the forecast results, it can be seen that the sales volume of A3 brand cigarettes shows a general upward trend, but there are occasional months where sales drop sharply; whereas the sales volume of A4 brand cigarettes fluctuates within a certain range, tending to be flat.Experimental results demonstrate that the proposed model achieves low prediction errors (MSE: 0.0042) and high accuracy (R^2: 0.9056).

By predicting the future sales volume and sales of each tobacco variety through this joint model, the consumption structure can be further analyzed, which is conducive to predicting the market capacity and grasping the changing trend.

References

[1] Liang Shangjian,Wang Ying. Research on the Impact of the New Crown Epidemic on China's Equity and Bond Markets--An Empirical Analysis Based on the ARIMA Model [J]. China Securities and Futures, 2024, (04): 59-66+89.

[2] CHEN Hong, WANG Yi, ZHOU Qian. Macroeconomic impact and prediction of economic operation indicators of tobacco business enterprises--Taking N city tobacco company as an example [J]. Modern Enterprise, 2024, (07): 65-67.

[3] Guo Yuanbo. Reform and Innovation of Cigarette Marketing of Tobacco Industry Enterprises under the New Retail Mode [J]. Modern Enterprise Culture, 2024, (13): 38-40.

[4] Chang Guangming,Zhao Xia. Research on wheat price prediction based on LSTM model[J]. Software,2024,45(11):42-45.

[5] Deng-Yao Zhang. Comparison of futures price prediction effects based on BPNN and optimization methods[J]. Statistics and Decision Making,2024,40(23):161-166.

[6] Ding Yi,Liu Tao,Wang Zhenya. Adaptive weighted Savitzky-Golay filtering for early bearing failure feature extraction [J]. Manufacturing Technology and Machine Tools, 2024, (06): 58-66.

[7] You YQ, Li XT, Liu H, et al. Laser self-mixing interferometric microdisplacement reconstruction based on the combination of wavelet threshold filtering and S-G filtering[J]. Intense Laser and Particle Beam,2024,36(08):13-20.

[8] WANG Yuning,ZHOU Kai,SHEN Shoufeng. A state transfer prediction model based on XGBoost[J]. Journal of Zhejiang University of Technology,2024,52(03):275-279.

[9] TAN Jie, GUI Qianjun, LIAO Chaoyang, et al. Relationship between urban form and surface temperature based on XGBoost-SHAP interpretable machine learning model[J/OL]. Journal of Applied Ecology,1-13[2025-01-17].

[10] Shanshan Li, Chaoyang Sun, Guodong Li. Depth-interpretable machine learning model of cavity effusion based on LightGBM and SHAP [J]. Mechanics Quarterly, 2024, 45 (02): 442-453.