ISSN 2959-6157

Research on sales forecasting in e-commerce industry for imbalanced classification data

Taiyu Lu

University Putra Malaysia, 43400 Serdang, Selangor, Malaysia
Email: 18087385747@163.com

Abstract:

The background of this study is that with the advent of the big data era, e-commerce sales forecasting has become a key factor in improving the market competitiveness and economic benefits of enterprises. To solve this problem, we used machine learning technology to build a comprehensive sales forecasting system. By processing massive sales data, including data cleaning, label encoding, outlier processing and other steps, we established a complete data set. In terms of model selection, we tried multiple regression models, such as RandomForestRegressor [1], ExtraTreesRegressor, etc., and evaluated their performance through cross-validation. In order to solve the problem of data imbalance, a combination of oversampling technology (RandomOverSampler)[2] and normalization processing is used. Finally, we selected ExtraTreesRegressor as the best model and evaluated it on the training set. The research results show that the accuracy and reliability of sales forecasts can be improved by comprehensively processing sales data and selecting appropriate machine learning models. The contribution of this study in the field of e-commerce sales forecasting is to provide a comprehensive and practical solution, which provides important decision-making support for enterprises in market competition. Combining machine learning technology and data processing methods, we provide e-commerce companies with an effective sales forecasting strategy that is expected to have a positive impact in improving market competitiveness, reducing risk costs, and accelerating revenue growth [3]. Future research directions can be carried out in deeply exploring the characteristics of sales data, optimizing model parameter adjustment, and combining professional knowledge in more fields. Introducing more emerging machine learning algorithms and technologies to adapt to the changing market demands in the e-commerce field is expected to further improve the performance and adaptability of the sales forecasting system.

Keywords: Random Over Sampler, Extra Trees Regressor, Sales Forecast, imbalanced classification data

1. Introduction

1.1 background

With the advent of the big data era, enterprises are facing an unprecedented surge in data volume. As an important part of the commercial field, the e-commerce industry is experiencing vigorous development driven by the popularity of the Internet and emerging business formats such as live broadcast e-commerce. However, this explosive growth in data volume also brings new challenges for enterprises, namely how to extract valuable information from massive data [3]. In this context, e-commerce sales forecasting has become one of the most important tools for enterprises to improve market competitiveness and economic benefits [4].

The popularity of the Internet and the rise of live streaming e-commerce have provided more data sources for e-commerce sales forecasting, making the forecasting model more complex and larger. Traditional statistical forecasting methods are often unable to handle such large, complex and noisy sales data. Therefore, researchers have turned to machine learning, a powerful data mining technology, to achieve prediction and decision-making on unknown data by learning patterns and features from data [5].

The extensive application of machine learning in the field of e-commerce sales forecasting enables companies to better understand the market and customer needs, optimize sales strategies and plans, and improve market competitiveness and economic benefits. Through in-depth analysis of sales data, machine learning models can predict product sales in the short term based on current data, providing companies with timely and accurate decision support in production, marketing, pricing, and investment [6]. For the e-commerce industry, this not only means better un-

derstanding of the market shape and reducing risk costs, but also accelerating the company's revenue growth.

Against this background, this study aims to build a combined prediction model that can solve the problem of data imbalance through machine learning technology to further improve the accuracy and reliability of e-commerce sales prediction. At the same time, researchers are also committed to exploring how to effectively handle large, complex, and noisy sales data, and how to optimize models to adapt to special situations in the e-commerce industry, such as sales fluctuations [7]. This will provide e-commerce companies with strong support and competitive advantages in the current fiercely competitive market environment.

1.2 Problem Statement

With the vigorous development of the e-commerce industry and the advent of the big data era, e-commerce sales forecasting has become a key factor in improving the market competitiveness and economic benefits of enterprises. However, when dealing with massive, complex, and noisy sales data, traditional statistical forecasting methods face a series of challenges, and they have limitations in data processing and model adaptability.

The main problem is that e-commerce sales data often show imbalance, that is, the sales volume of different products or categories varies greatly. This results in that when training a machine learning model, the model is more likely to be biased towards the side with a larger number of learning samples, while it is easy to ignore products with relatively small sales volume. This imbalance will affect the accuracy of forecast results, especially for products with low sales volume, where traditional models often struggle to achieve satisfactory results.

In addition, the complexity of e-commerce sales data is also reflected in the high volatility of sales, which may be affected by factors such as promotions and cause abnormal fluctuations. Traditional methods may prove inadequate when dealing with this volatility, and more flexible and adaptable models are needed to address this challenge [7].

Therefore, the questions of this research focus on how to effectively deal with the imbalance problem in e-commerce sales data, improve the prediction accuracy of products with less sales, and how to optimize the model to adapt to special situations such as sales fluctuations in the e-commerce industry. Solving these problems will provide enterprises with more reliable sales forecast results, help make accurate decisions, and improve market competitiveness and economic benefits.

1.3 Research Questions

When facing the imbalance problem and sales volatility challenge in e-commerce sales forecasting, this study aims to solve the following key issues:

(1) How to deal with the category imbalance problem in e-commerce sales data, especially in product categories with low sales volume, to improve the model's prediction accuracy for a few categories?

(2) How to optimize the model so that it can better adapt to the volatility of sales data in the e-commerce industry, reduce excessive sensitivity to abnormal fluctuations, and thereby improve the stability and reliability of the overall forecast?

(3) While solving imbalance problems and volatility challenges, how to maintain the model's high sensitivity to overall sales data to ensure rapid response and accurate predictions to market changes?

Through in-depth research and answers to the above questions, this study aims to provide e-commerce companies with a more reliable and accurate sales forecast model, provide strong support for their decision-making, and enhance market competitiveness [8].

1.4 Research Objective

The main purpose of this study is to:

(1) Propose a combined prediction model based on machine learning to effectively solve the category imbalance problem existing in e-commerce sales data. Through this goal, it aims to achieve accurate forecasts of products with lower sales volumes, thereby helping companies better understand market demand and optimize production and inventory strategies.

(2) Optimize the model structure and algorithm to enhance the model's adaptability to the volatility of e-commerce sales data. Through this goal, it aims to reduce the model's over-sensitivity to abnormal sales fluctuations, improve the stability of the overall sales forecast, and make it more consistent with the actual operating conditions of the e-commerce industry.

(3) Maintain the model's high sensitivity to overall sales data to ensure timely response to market changes. Through this goal, the model is designed to solve imbalance problems and volatility challenges while still being able to flexibly respond to dynamic changes in the market and achieve accurate predictions of future sales trends.

By achieving the above research goals, this study aims to provide e-commerce companies with more reliable and accurate sales forecasting tools to help them make more informed business decisions and improve market competitiveness and economic benefits .

2. Literature Review

Research in the field of e-commerce sales forecasting has made significant progress both at home and abroad, but

there are some differences in methods and focus.

2.1 Research state in China:

Chinese researchers pay more attention to processing outliers and noise in e-commerce sales data. For example, Hu Bowen's research [9] processed abnormal discrete values, weakened the adverse effects, and improved the quality of data by deleting abnormal points. In terms of feature engineering, Chinese researchers place more emphasis on the processing of text attributes and the application of principal component analysis (PCA). Yang Yuxin's [10] research uses LabelEncoder to convert text values into numeric codes, and principal component analysis is used to reduce the dimensionality of feature sets. Chinese researchers have used a variety of machine learning algorithms, including support vector machines, LSTM network models, and random forests. Among them, the research on combination model is relatively sufficient, and the accuracy of sales forecast [11] is improved through the combination of different algorithms.

2.2 Current status of foreign research:

In foreign research, more attention is paid to standardization methods for data preprocessing. For example, the StandardScaler method used in literature [12] eliminates the dimensional relationship between variables by normalizing feature data. Foreign researchers place more emphasis on the extraction of time series and various features related to time, users, products, promotions, etc. This is reflected in the extensive use of LSTM network models to better capture timing information. Foreign researchers also use a variety of machine learning algorithms, but place more emphasis on deep learning methods, such as the LSTM network model. In addition, there are relatively many studies on time series models to deal with the time series characteristics in sales data.

2.3 Comparative analysis:

In general, both domestic and foreign research on e-commerce sales forecasting have emphasized the importance of data preprocessing, feature engineering, model selection and optimization. However, there are some differences in the focus of the two methods. China focuses more on the processing of outliers and research on combination models, while foreign countries place more emphasis on the application of time series models and deep learning methods. This difference may stem from the differences in data background, industrial development stage, and subject preferences between researchers in the two places. In future research, we can draw on their respective advantages to form a more comprehensive e-commerce sales forecast research system.

3. Research methods

3.1 Data analysis model

In order to solve the challenges and problems in e-commerce sales forecasting, we built a comprehensive data analysis model, covering key steps such as data processing, feature engineering, outlier processing, model selection, and data imbalance processing. The following are the main components of our data analysis model:

(1) Data collection and preprocessing: For existing e-commerce sales data, use the Pandas library for data reading and preliminary observation. Ensure data integrity and accuracy by performing preprocessing steps such as filling missing values and handling outliers.

(2) Feature engineering: By analyzing the characteristics of sales data, select characteristics related to time, users, products, etc. LabelEncoder is used to label some features for subsequent model training. By performing feature engineering on sales data, we select features that are closely related to sales forecasts, including time, users, products and other information, which helps to improve the sensitivity and accuracy of the model to sales data.

(3) Outlier processing: Use statistical methods and graphical methods to identify and process outliers. Use methods such as Grubbs test to clean the data and improve the stability of the model. Outlier processing: Improve the robustness of the model to interference and enhance the stability of the model through reasonable outlier processing.

(4) Model selection and comparison: Choose a variety of machine learning algorithms including ExtraTreesRegressor, AdaBoostRegressor, etc. for sales forecast modeling. Compare the performance of different models on the data set through methods such as cross-validation and select the model with the best performance [13]. We evaluate the performance of individual models on the dataset and select the best performing model for further analysis and prediction.

(5) Data imbalance processing: In order to solve the problem of imbalanced sales data distribution, we introduced RandomOverSampler for oversampling processing. This step helps balance the samples of each category and improve the model's coverage of the entire data distribution.

Through the comprehensive application of the above research methods, it aims to improve the accuracy and stability of e-commerce sales forecasts, provide a more reliable reference for corporate decision-making, and promote the development of the e-commerce industry and the improvement of e-commerce levels. Economic benefits.

3.2 Data analysis plan

3.2.1 Import libraries and data:

Import the necessary libraries, including pandas, numpy,

seaborn, etc.Read e-commerce sales data from Excel files.

3.2.2 Data exploration:

Print column names with unique value 1.

```
has_urgency_banner
theme
crawl_month
currency_buyer
badges_count
badge_local_product
badge_product_quality
badge_fast_shipping
shipping_is_express
```

Figure.1 column names with unique value 1

Print column names whose values differ by greater than 1200.

```
0.0    1421
1.0     138
2.0      11
3.0       2
Name: badges_count, dtype: int64
0.0    1543
1.0      29
Name: badge_local_product, dtype: int64
0.0    1455
1.0     117
Name: badge_product_quality, dtype: int64
0.0    1552
1.0      20
Name: badge_fast_shipping, dtype: int64
0.0     1568
1.0        4
50.0       1
Name: shipping_is_express, dtype: int64
```

Figure.2 column names whose values differ by greater than 1200

The columns are then processed, removing unnecessary columns.

3.2.3 Label encoding:

Encode tags for 'product_color', 'origin_country', 'merchant_id', 'product_id'.

3.2.4 Handling missing values:

Fill the 'has_urgency_banner' column with missing values of 0. Fill other missing values using the most frequent value. Outlier handling: Use Grubbs tests to detect and handle outliers[14].

3.2.5 Data exploration:

Plot a histogram of the processed data set.

(1) Data distribution observation: Histograms can display the data distribution of each feature and help observe the overall shape of the data, such as whether it presents a normal distribution, a skewed distribution, etc.

(2) Outlier detection: Histograms can help you find out whether there are outliers. Outliers usually appear as extreme values that are relatively isolated or deviate from the main body of the entire distribution.

(3) Feature engineering: Observing the histogram helps to select appropriate feature engineering methods, such as logarithmic transformation, normalization, standardization, etc., to improve the distribution of data.

(4) Model selection: For some machine learning models, such as models that require features to obey a normal distribution, the histogram can help you determine whether the features need to be transformed to meet the assumptions of the model.

(5) Optimize data processing steps: If the histogram shows that the distribution of some features is not ideal, you may need to adjust the data processing steps, such as modifying the outlier processing strategy, choosing a different method of filling missing values, etc.Through the histogram below, we can see that after data processing, there is no abnormality in the data distribution.

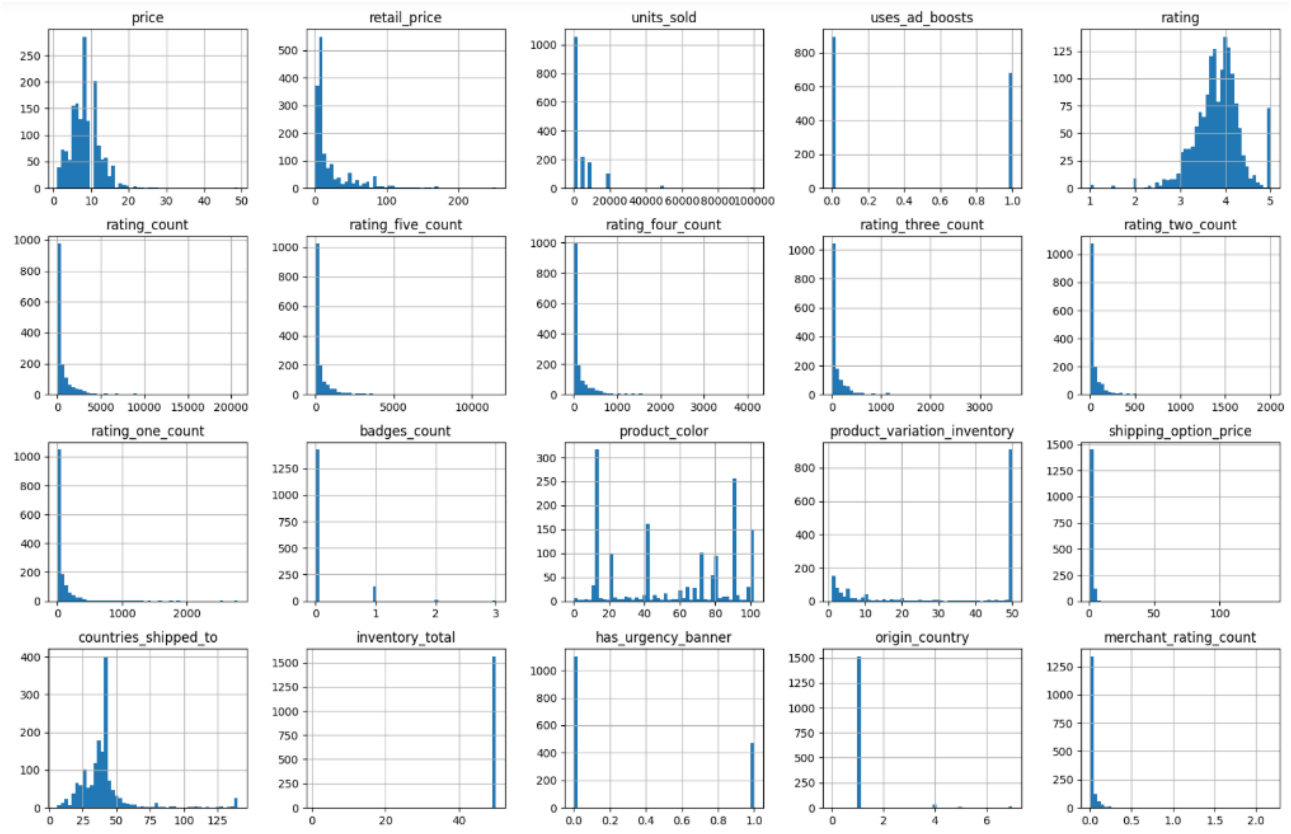

Figure.3 data distribution

(6) Data splitting:

Use StratifiedShuffleSplit to split the dataset to ensure that the class distributions in the training and test sets are similar. Print data set information: Print the size of the training set and test set. Print the distribution of values for the 'badges_count' and 'merchant_has_profile_picture' columns. Finally, the training set is used as the final data set to model the model.

3.2.6 correlation analysis

Guide feature selection: By analyzing correlations, you can understand which features have a greater impact on the target variable (here, the number of units sold), thereby guiding the feature selection process [15].

Identify multicollinearity: Correlation analysis can help detect multicollinearity between features, i.e. one feature can be predicted by other features.

Predicting the target variable: Features with high correlation may have a stronger impact on predicting the target variable, which helps in selecting appropriate features for modeling.

Discover underlying relationships: Scatter matrices and scatter plots can reveal underlying relationships between features and help understand the patterns and structure of your data.

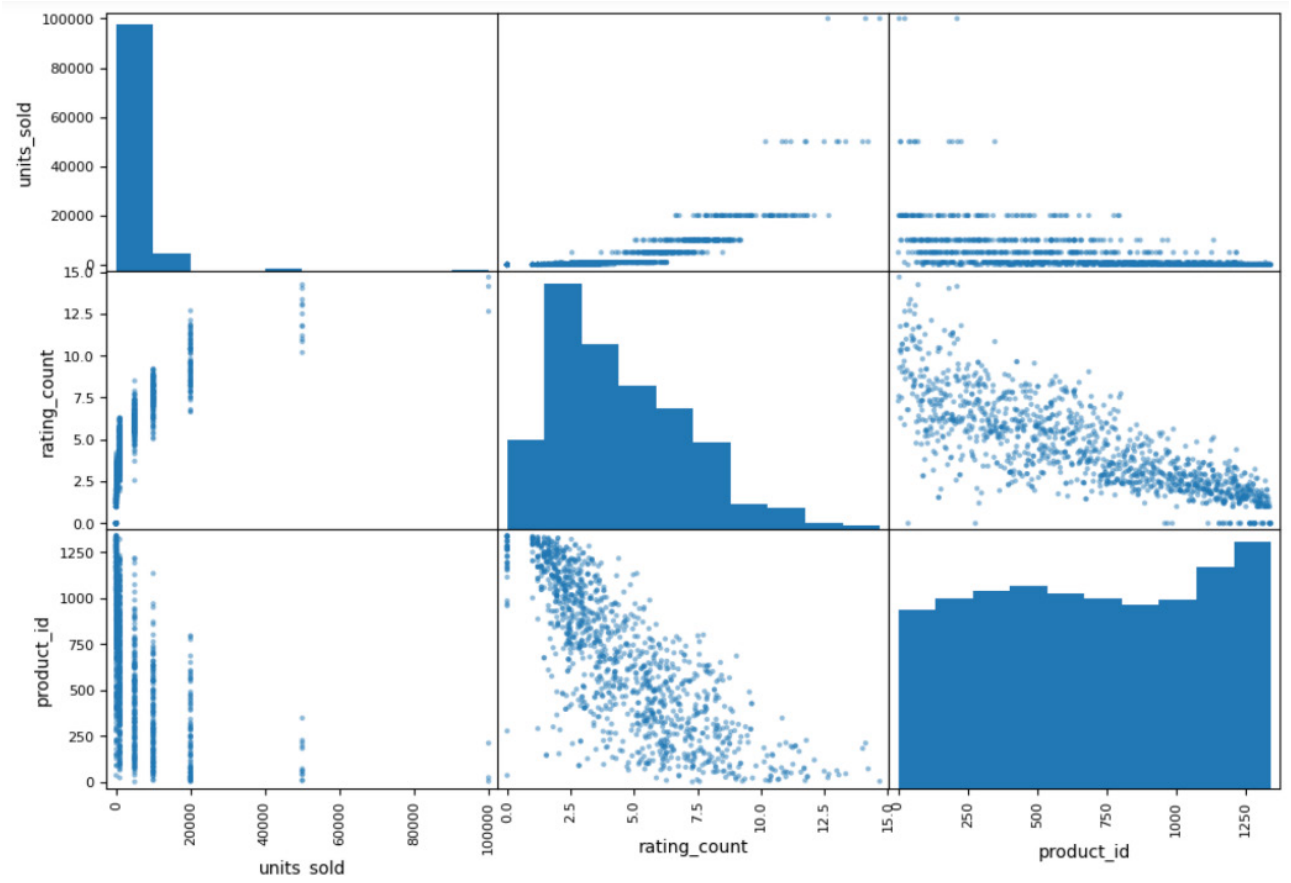

Figure.4 relationships between features

Correlation between 'units_sold' and 'rating_count': 0.7571739492352604

Correlation between 'units_sold' and 'product_id': -0.4728532073705615

Correlation between 'rating_count' and 'product_id': -0.7923500900252624

Compare the data modeling before and after balancing:

We use the processed data and then split the data. The training set is used to train the model, and the test set is used to verify the accuracy of the model. For the unbalanced data units_sold, we use oversampling combined with standardization to perform the data After balancing, the data is reduced to a certain range. This can firstly balance the data, and secondly, reduce the dispersion of the data, which will help greatly reduce the mean square error of the model and make the results develop in a good direction.

Before balancing:

	model_name	mean_rmse	rmse_std
0	<class 'sklearn.ensemble._forest.ExtraTreesReg...	3291.187146	1317.705056
2	<class 'sklearn.ensemble._bagging.BaggingRegre...	3364.972395	1436.383866
4	<class 'sklearn.ensemble._forest.RandomForestR...	3404.678581	1286.256885
3	<class 'sklearn.ensemble._gb.GradientBoostingR...	3417.484981	1363.662789
1	<class 'sklearn.ensemble._weight_boosting.AdaB...	3986.815679	1616.322468

Figure.5 data modeling before balancing

After balanc ing:

```
Model comparison results after balancing the data:
                                                model_name  mean_rmse  rmse_std
0  <class 'sklearn.ensemble._forest.ExtraTreesReg...   0.033275  0.013086
4  <class 'sklearn.ensemble._forest.RandomForestR...   0.033477  0.012576
3  <class 'sklearn.ensemble._gb.GradientBoostingR...   0.034594  0.013976
2  <class 'sklearn.ensemble._bagging.BaggingRegre...   0.036010  0.010974
1  <class 'sklearn.ensemble._weight_boosting.AdaB...   0.044266  0.016550
```

Figure.6 data modeling before after balancing

In the first result, the evaluation index of model performance is the root mean square error (RMSE), which is used to measure the prediction error of the model. The average RMSE and RMSE standard deviation of different models are given in Result 1. The performance of the model is sorted according to the size of the average RMSE. The smaller the value, the better the performance of the model. From the first result, it can be seen that the ExtraTreesRegressor model performs best in terms of root mean square error, and the AdaBoostRegressor model performs relatively poorly.

In the second result, the evaluation index of model performance is the root mean square error (RMSE) after data balancing. Data balancing can be achieved through methods such as oversampling or undersampling, aiming to solve the problem of imbalanced data sets. The performance evaluation results in Result 2 show that after data balancing processing, the RMSE of all models is significantly reduced, and the standard deviation is also reduced accordingly, indicating that the performance of the model has been significantly improved when processing balanced data sets. In this result, the ExtraTreesRegressor model still leads in performance, while the AdaBoostRegressor model is still relatively poor, but the overall performance has improved.

Taken together, the first result mainly shows the performance difference of the model on the original data, while the second result highlights the improvement of the model performance after balancing the data.

4. Conclusion and Discussion

This study addresses the critical challenges faced by the e-commerce industry in sales forecasting, driven by the surge in data volume in the big data era. The study focuses on leveraging machine learning technology to enhance the accuracy and reliability of e-commerce sales predictions, considering issues such as data imbalance and sales volatility.

The background section highlights the importance of e-commerce sales forecasting in improving market competitiveness and economic benefits for enterprises. The study emphasizes the need for advanced predictive models due to the complexity and size of modern e-commerce data.

The identified problems include data imbalance, especially in product categories with low sales volume, and the challenge of adapting models to the high volatility of e-commerce sales data. These problems are crucial as they can impact the accuracy of sales forecasts, particularly for less-sold products.

The research questions guide the study toward solving these key issues:

Addressing category imbalance in e-commerce sales data to improve prediction accuracy for low-sales categories.

Optimizing models to adapt to sales data volatility and enhance overall forecast stability.

Maintaining high sensitivity to overall sales data while solving imbalance and volatility challenges.

The research objectives are designed to achieve these goals by proposing a combined prediction model based on machine learning. The model aims to effectively handle category imbalance, improve adaptability to sales volatility, and maintain sensitivity to overall sales data.

The literature review provides insights into the current state of e-commerce sales forecasting research, highlighting differences between domestic and foreign approaches. Chinese research emphasizes outlier processing and combination models, while foreign research focuses more on standardization methods and deep learning, especially LSTM network models. The study suggests drawing on the strengths of both approaches for a more comprehensive research system.

The research methods involve a comprehensive data analysis model covering data processing, feature engineering, outlier processing, model selection, and data imbalance processing. The data analysis plan includes steps such as data exploration, label encoding, outlier handling, model selection, and data splitting.

The correlation analysis helps guide feature selection, identify multicollinearity, and understand the relationships between features. The comparison of data modeling before and after balancing demonstrates the significant improvement in model performance after addressing data imbalance.

The results showcase the model performance before and

after balancing. In the initial results, ExtraTreesRegressor outperforms other models, while AdaBoostRegressor performs relatively poorly. After data balancing, all models exhibit significantly reduced RMSE, indicating improved performance. ExtraTreesRegressor remains the top-performing model.

In summary, this study provides a robust framework for enhancing e-commerce sales forecasting using machine learning. By addressing data imbalance and volatility challenges, the proposed model aims to offer e-commerce companies more reliable and accurate sales forecasts. This, in turn, can support informed decision-making, improve market competitiveness, and enhance economic benefits in the dynamic e-commerce landscape [16].

References

[1] Li Xinhai. (2013). Application of random forest model in classification and regression analysis. Journal of Applied Entomology (04), 1190-1197.

[2] Fang Yu, Zheng Huyu, Cao Xuemei. Three-way oversampling imbalanced data classification method [J]. Journal of Shandong University (Science Edition), 2023(012):058.

[3] Ou Jiequan. Data mining analysis based on e-commerce platform product information [J]. Electronic Technology and Software Engineering, 2016, No. 93(19): 209.

[4] Lin Muxing. Research on commodity sales forecast model using large-scale data Gaussian process regression under demand uncertainty [D]. Jinan University, 2020.

[5] Jiang Yanmei, Bu Qingkai. Supermarket Commodity Sales Forecast Based on Data Mining[J]. 2018.

[6] Pu Jiapeng. Application of machine learning in commodity sales forecast[J]. Electronic Production, 2018(22):3.DOI:CNKI:SUN:DZZZ.0.2018-22-039.

[7] Sun Puyang, Zhang Yan, Huang Jiuli. Export behavior, marginal cost and sales fluctuation - a study based on Chinese industrial enterprise data [J]. Financial Research, 2015(9):15. DOI:CNKI:SUN:JRYJ.0.2015 -09-011.

[8] Chen Yun, Wang Huanchen, Shen Huizhang. Research on price competition between e-commerce retailers and traditional retailers [J]. Systems Engineering Theory and Practice, 2006, 26(1):7.DOI:10.3321/j.issn:1000- 6788.2006.01.005.

[9] Hu Bowen. (2022). Research on e-commerce sales prediction based on deep learning (Master's thesis, Qingdao University). https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202301&filename= 1022773119.nh.

[10] Yang Yuxin. (2021). Accurate market description and forecast based on big data analysis (Master's thesis, Beijing Jiaotong University). https://link.cnki.net/doi/10.26944/d.cnki.gbfju.2021.000991doi:10.26944/d.cnki.gbfju.2021.000991.

[11] Yin Chunwu. Application of GM(1,1) in commodity sales forecast[J]. China Business and Trade, 2010(28):2.DOI:10.3969/j.issn.1005-5800.2010.28.160.

[12] Raju V N G , Lakshmi K P , Jain V M ,et al.Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification[C]//2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT).2020.DOI:10.1109/ICSSIT48917.2020.9214160.

[13] Zhou Yuduan Yongrui. Retail product sales forecast based on clustering and machine learning [J]. Computer System Applications, 2021, 30(11):188-194.

[14] Liu Ying, Wei Gong, LIUYing, et al. Improvement of GRUBBS method in outlier detection [J]. Henan Science, 2006, 24(5):641-644.DOI:10.3969/j.issn.1004-3918.2006 .05.006.

[15] Tang Liang, Duan Jianguo, Xu Hongbo, et al. Feature selection algorithm and application based on mutual information maximization [J]. Computer Engineering and Applications, 2008, 44(13):4.DOI:10.3778/j.issn .1002-8331.2008.13.039.

[16] Duan Lili. Research on the impact of order flow imbalance on the yield and volatility of agricultural product futures market [D]. Harbin Institute of Technology, 2016. DOI: 10.7666/d.D01099911.