Proceedings of the 3rd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/128/2024.18672

Comparative Analysis of Machine Learning Algorithms for Sales Forecasting in the Russian Toy Retail Sector

Hao Liu[1,a,*]

[1]*College of Big Data and Information Engineering, Guizhou University, Guizhou, China*
a. ie.hliu21@gzu.edu.cn
**corresponding author*

Abstract: This research provides a detailed examination of various machine learning approaches for the sake of forecasting sales within the Russian toy retail industry. This sector is marked by fluctuating consumer preferences and a pressing need for precise inventory control, making accurate sales predictions vital. The study aims to determine which forecasting technique yields the most reliable results for this dynamic market. To achieve this, several predictive models were assessed, including Linear Regression, Random Forest, Light Gradient Boosting Machine (LightGBM), and Extreme Gradient Boosting (XGBoost). The analysis involved a comprehensive process encompassing data preprocessing, exploratory data analysis (EDA), feature engineering, and rigorous model training, using a dataset sourced from Kaggle. The results highlight that Random Forest consistently delivers superior predictive performance compared to its counterparts. This model excels in balancing accuracy with generalization, making it particularly effective for sales forecasting in environments characterized by variability and uncertainty. The practical implications of these results are substantial, offering retailers a robust tool for refining their operational strategies, optimizing inventory management, and enhancing customer experience through improved forecasting accuracy.

Keywords: Machine Learning, Sales Forecasting, Random Forest, Inventory Management.

1. Introduction

Sales forecasting plays a vital role in the governance of retail operations, particularly within niche markets such as Russian toy stores [1]. Reliable sales predictions are essential for optimizing stock levels, boosting customer satisfaction, and ensuring financial stability. Given the significant impact that accurate forecasting has on business performance, this study focuses on assessing the effectiveness of different machine learning models—Linear Regression, Random Forest, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM)—in predicting sales within this sector [2-4].

The research aims to evaluate how well each of these models performs in forecasting sales trends, emphasizing their practical relevance to the retail industry. By conducting a comprehensive comparative analysis, this study seeks to promote advancements in predictive analytics and offer actionable insights to retail professionals. The insights gained are intended to refine decision-making processes, enhance inventory management strategies, and improve overall operational efficiency in a market characterized by its volatility and complexity. The findings of this study are anticipated to

DOI: 10.54254/2754-1169/128/2024.18672

bridge the divide between practical application and theoretical research, providing retailers with a more comprehensive understanding of how various forecasting models can be utilized to achieve more accurate and effective sales predictions. By integrating advanced predictive modeling techniques, the research aims to empower retailers with the tools for better decision-making and strategic decisions.

Sales forecasting has been a longstanding focus in various industries, with early works laying the foundation for predictive techniques and systems [5]. Judgmental forecasting, involving human judgment, has also been studied for its effectiveness in practical applications [6]. Organizational factors, including company culture, have been shown to impact forecasting accuracy [7]. Recent developments have been privy to the rise of artificial intelligence (AI) and machine learning (ML) methodologies in the realm of sales prediction. These technological innovations are increasingly being adopted to enhance forecast accuracy and provide deeper insights into market trends. Venkataramanan and Sadhu demonstrated the superiority of AI in predictive accuracy [8], while Sohrabpour et al. utilized genetic programming for high-accuracy export sales forecasting [9]. He et al. introduced an approach that integrates Long Short-Term Memory (LSTM) networks optimized using Particle Swarm Optimization (PSO) to predict sales volumes in e-commerce. This hybrid model combines the temporal sequence analysis capabilities of LSTMs with the optimization strengths of PSO, aiming to improve forecasting accuracy in dynamic online retail environments [10]. Ma and Fildes proposed a meta-learning framework for customizing forecasting models [11]. Shilong designed a machine learning-based sales forecasting model optimized through feature engineering [12]. Gandhi et al. emphasized the strategic importance of precise sales forecasting [13]. These studies collectively indicate a clear trajectory toward the integration of AI and ML in sales forecasting, emphasizing personalized, real-time, and adaptive models. The field is rapidly evolving, with researchers and practitioners increasingly focusing on leveraging cutting-edge technologies to address the intricate dynamics of modern markets.

This research investigates a broad spectrum of machine learning approaches to optimize sales forecasts for toy retailers in Russia. The goal is on discovering the most efficient algorithm for improving inventory control, customer satisfaction, and financial forecasting. Initially, the study uses Linear Regression as a baseline model on the basis of its straightforward nature and comprehensibility, which helps elucidate the relationship between sales metrics and influencing factors. Following this, the study employs more highly developed models, including Random Forest, XGBoost, and LightGBM, to uncover non-linear associations and intricate patterns that simpler models may miss. LightGBM, in particular, is highlighted for its superior performance in handling large-scale datasets and effectively managing categorical data, which is crucial for intricate retail environments. The models are assessed using real-world sales data, with performance evaluated through metrics such as R-squared, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), providing a comprehensive view of prediction accuracy and model reliability. Additionally, a feature importance analysis is carried out to identify key drivers of sales trends. This analysis offers actionable insights that can inform inventory management and marketing strategies, enabling retailers to make more data-driven decisions.

To further optimize model performance, hyperparameter tuning is applied to each algorithm, ensuring that prediction accuracy is maximized. This tuning process not only enhances model efficiency but also refines the predictive power of each approach, making the results more reliable for practical business applications in the retail sector. The use of robust forecasting methods enables improved demand anticipation, waste reduction, and profitability enhancement, supporting data-driven strategies in the dynamic market environment of Russian toy stores.

DOI: 10.54254/2754-1169/128/2024.18672

2. Methodology

2.1. Dataset Description and Preprocessing

The data leveraged in this study were sourced from Kaggle, a respected data analytics platform that provides a broad spectrum of real-world datasets commonly used for data science competitions and projects [1]. Specifically, the dataset was released by a Russian toy retailer and contains authentic sales records from January 2013 to October 2015. It is organized in a tabular format, with several key columns capturing essential sales information. The "date" column records the transaction date in the DD.MM.YYYY format, while "date_block_num" is an integer representing the time period, such as the month number as DBM. The dataset also includes "shop_id," identifying the shop where the transaction took place as SI, and "item_id," specifying the product sold as II. Additionally, "item_price" reflects the price of the product at the time of sale as IP, and "item_cnt_day" denotes the sales volume for a particular day as IC. Table 1 provides a sample of the dataset to demonstrate its structure and content. This dataset is the key component in model training and evaluation, offering a rich source of information for predictive analysis and insights into retail sales patterns. The detailed transaction data allows for comprehensive exploration and modeling of key factors influencing sales trends.

Table 1: Performance of different algorithms.

Date (2013)	DBM	SI	II	IP	IC
02.01.2013	0	59	22154	999	1
03.01.2013	0	25	2552	899	1
05.01.2013	0	25	2552	899	-1
06.01.2013	0	25	2554	1709.05	1
15.01.2013	0	25	2555	1099	1
10.01.2013	0	25	2564	349	1
02.01.2013	0	25	2565	549	1
04.01.2013	0	25	2572	239	1
11.01.2013	0	25	2572	299	1

The utilization of this dataset offers a detailed record of sales across various item types different points of sale and precise transaction timestamps. This wealth of information deepens the training of models and enables a fine-grained analysis of market trends enhancing the precision and applicability of the predictions. Thorough preprocessing was conducted to ensure the data was appropriate for modeling. This entailed methodical data cleansing to remove inconsistencies, managing missing values to preserve data integrity, and tackling potential data leakage to avert biased results. Ensuring high data quality was critical, as it has a direct impact on the reliability of the derived conclusions. Multiple data sources were integrated into a unified dataset facilitating seamless analysis and modeling.

During the exploratory data analysis (EDA) phase, the data were meticulously examined to uncover hidden patterns and trends. This involved identifying seasonal variations sales peaks and other significant insights informing feature engineering and model selection processes. Features engineered included price fluctuation capturing the volatility of product prices through computations time window utilizing past sales data to reflect trends with statistical indicators and lag leveraging historical sales data for forecasting future sales. These features enhanced the predictive power of models by providing nuanced insights into sales behavior.

DOI: 10.54254/2754-1169/128/2024.18672

2.2. Proposed Approach

The purpose of this research is to analyze the effectiveness of multiple machine learning algorithms in predict sales for toy retailers in Russia. The study's primary focus is to pinpoint the most dependable and robust forecasting method that can assist retailers in refining inventory management, boosting customer satisfaction, and enhancing financial planning. To achieve this, the research adopts a holistic methodology that includes thorough data preparation, EDA, feature engineering, and the application of several machine learning algorithms. The methodology is illustrated in Figure 1, which outlines the key stages of the research process. This systematic strategy allows for a comprehensive evaluation of each model's performance and delivers actionable insights for improving sales predictions. By leveraging advanced analytical techniques, this study seeks to provide retailers with strategic recommendations that can help optimize their operations and better respond to market fluctuations. The ultimate goal is to offer practical solutions that enhance the accuracy of sales forecasts, enabling retailers to make informed strategic decisions and maintain their competitive advantage in a changing market.

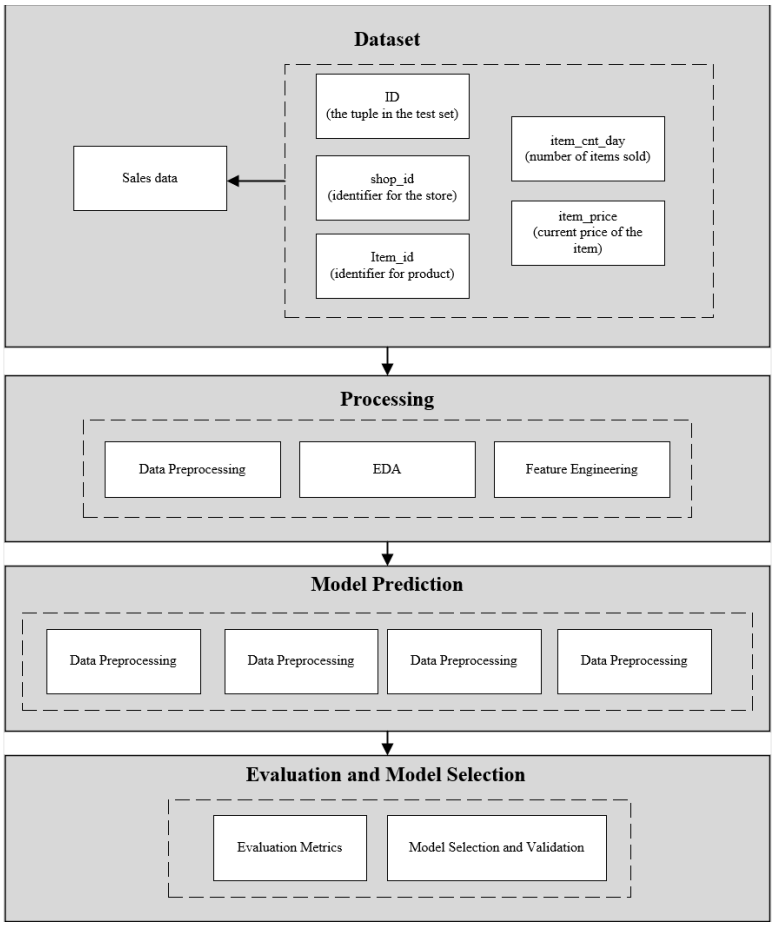

Figure 1: Pipeline of the proposed approach.

The procedure starts with the collection of sales data, after which preprocessing is carried out to enhance the quality and consistency of the dataset. A comprehensive EDA then reveals underlying patterns and trends. Feature engineering creates new features that capture the nuances of sales dynamics. The engineered features are then input into a series of machine learning models for prediction. The performance is evaluated, and the best model is chosen based on predefined performance measures.

DOI: 10.54254/2754-1169/128/2024.18672

Linear regression serves as a foundational benchmark due to its straightforward design and interpretive ease, providing a baseline for comparing more finely crafted and elaborate models. Following this, Random Forest, XGBoost, and LightGBM are employed to handle non-linear associations and identify sophisticated patterns in the data. Ensemble methods like Random Forest and XGBoost are particularly effective in handling extensive datasets with numerous variables, making them well-suited for this analysis. LightGBM is selected for its high efficiency in processing large-scale data and its adeptness at managing categorical features.

With those models, the performance of these models is tested real-world sales data from Russian toy retailers, with key metrics including R-squared, RMSE, and MAE. Feature importance analysis is conducted to identify the critical factors driving sales, enabling a better understanding of which features are most influential in accurate forecasting. This analysis helps refine inventory management and marketing strategies. Additionally, hyperparameter tuning is performed to enhance each algorithm's performance by systematically adjusting parameters to achieve the best configuration and ensure optimal predictive accuracy. This thorough evaluation and optimization process aims to provide actionable insights and improve the reliability of sales forecasts.

By identifying the most robust and reliable method for sales forecasting, retailers can anticipate market demands more accurately, reducing waste and enhancing profitability. This research supports the development of data-driven strategies that help businesses navigate the complexities of demand forecasting in dynamic market conditions.

2.2.1. Machine Learning Methods

Linear regression is a key statistical technique used to model the relationship between a continuous target variable and one or more input variables, providing a basis for more complex models. In this study, it is selected as the baseline model for sales forecasting. Its simplicity and interpretability provide a clear framework for understanding how various factors influence sales. Linear regression assumes a linear association between the independent variables and the sales outcome. For this research, the linear regression is applied to a preprocessed dataset using the least squares method. This technique aims to minimize the sum of the squared discrepancies from observed and predicted sales figures. By offering insights into how each variable affects sales, linear regression establishes a benchmark for comparing more complex models.

Random Forest is a method that constructs multiple decision trees during training to improve the model's predictive accuracy. Each tree is built from a random subset of the data, and predictions are made using the majority rule for classification tasks and the average for regression tasks. This method reduces overfitting and enhances generalization by combining the outputs of various trees. In this study, Random Forest is employed to refine sales forecasting accuracy. The model is trained on a dataset that has been thoroughly preprocessed, with each tree trained on a different segment of the data. This approach mitigates overfitting and bolsters the model's performance on new dataset, making Random Forest a robust choice for sales prediction.

XGBoost, or eXtreme Gradient Boosting, is a highly developed gradient boosting algorithm known for its efficiency and outstanding performance. Due to its proficiency in managing large datasets and complex models, it is ideal for advanced sales forecasting. XGBoost builds trees in a step-by-step manner, with each new tree targeting the errors of the previous tree.

For this study, XGBoost is trained on a well-prepared dataset with the goal of minimizing a loss function that measures prediction errors. Its iterative approach, combined with its effectiveness in addressing missing data and optimize feature importance, makes XGBoost a powerful tool for generating accurate sales forecasts. High efficiency and scalability, particularly with large datasets, are achieved by LightGBM, a gradient boosting framework. It is equipped with several features such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to enhance

DOI: 10.54254/2754-1169/128/2024.18672

performance and reduce computational demands. In this research, LightGBM is utilized to leverage its efficient processing capabilities and robust performance. The model is trained on a meticulously prepared dataset, focusing on minimizing the loss function between predicted and actual sales figures. LightGBM's advanced features contribute to its effectiveness in sales forecasting.

2.3. Implementation Details

The predictive models were implemented on a hardware setup featuring an AMD Ryzen 9 5900HX with integrated Radeon Graphics and an NVIDIA GeForce RTX 3060 Laptop GPU with 6GB memory, ensuring efficient computation. The software environment used Python 3.7 for developing and deploying the models. Hyperparameters were tuned to optimize model performance. For the LightGBM model, key parameters included a maximum depth of 8, 255 leaves, 500 estimators, and a minimum child sample size of 1000 to prevent overfitting, with early stopping after 40 rounds without improvement. The XGBoost model had similar settings, including a maximum depth of 8, 500 estimators, and early stopping after 20 rounds of no improvement. The Random Forest Regressor was configured with 30 estimators and a maximum depth of 5 to balance complexity and generalization.

3. Result and Discussion

This chapter explores the analysis and discussion of the experimental results from the use of four different machine learning models—Linear regression, Random Forest, XGBoost, and LightGBM—on the sales forecasting dataset. The analyses focus on the performance metrics of RMSE, MAE, and R-square for both validation and training sets.

Table 2: Comparison of RMSE in models.

	Validation RMSE	Train RMSE
linear regression	1.52	2.27
random forest	1.38	2.08
XGBoost	1.64	2.47
lightGBM	1.55	2.35

Table 2 presents a comparison of RMSE values for the training and validation datasets across different models. The Random Forest model achieves the lowest RMSE values (Validation RMSE: 1.38; Train RMSE: 2.08), demonstrating the most balanced performance and suggesting it effectively captures training data patterns while maintaining strong generalization capabilities for unseen data.

However, despite efforts to optimize hyperparameters, the differences in RMSE scores between training and validation sets indicate that the model's performance may be influenced by the intrinsic complexity and variability of the dataset. These discrepancies are not solely indicative of overfitting but rather suggest that certain data characteristics, such as noise or non-linear relationships, could be affecting the model's predictive ability. This points to the importance of further refining feature selection or considering alternative approaches to capture more nuanced patterns in the data.

Table 3: Comparison of MAE in models.

	Validation MAE	Train MAE
linear regression	0.32	0.27
random forest	0.33	0.28
XGBoost	0.36	0.66
lightGBM	0.35	0.25

DOI: 10.54254/2754-1169/128/2024.18672

The MAE values, which measure the average magnitude of the discrepancies between the forecasted and true values, are showcased in Table 3. This statistic provides insight into the model's accuracy by evaluating the size of the errors without considering their direction. Similar to RMSE, Random Forest provides the best performance (Validation MAE: 0.33; Train MAE: 0.28). The larger gap between training and validation MAE for XGBoost (Training: 0.66; Validation: 0.36) could be attributed to the nature of the dataset, suggesting that certain patterns within the data are more challenging for the models to capture accurately.

Table 4: Comparison of R-square in models.

	Validation R-square	Train R-square
linear regression	0.62	0.6
random forest	0.68	0.66
XGBoost	0.55	0.53
lightGBM	0.6	0.57

Table 4 shows the R-square values, indicating how well the models fit the data. Once again, Random Forest leads with the highest scores (Validation R-square: 0.68; Train R-square: 0.66), suggesting it effectively captures the variability in the sales dataset. The lower R-square values for other models, especially XGBoost, might indicate that the dataset contains features that are not fully exploited by the models, leading to suboptimal predictions.

4. Conclusion

This study centers on the use of machine learning algorithms to forecast sales for toy retailers in Russia, utilizing a multi-step methodology. The process involves several key phases, including data collection, cleaning, and preprocessing to verify for accuracy and readiness for analysis. A thorough EDA follows, aiming to identify hidden trends and relationships in the sales data. Feature engineering is crucial in this phase, where new variables are created to better represent the complexities of sales patterns, ultimately boosting the models' performance. Multiple machine learning models are applied, including Linear Regression, Random Forest, XGBoost, and LightGBM, with the enhanced features serving as input to these algorithms. The comprehensive approach allows for an in-depth comparison of each model's forecasting abilities. Extensive testing reveals that ensemble models, particularly Random Forest, deliver the highest predictive accuracy. This suggests that these models are well-suited for the retail environment, offering significant advantages in inventory management and decision-making strategies. Future research will focus on the impact of external factors like seasonality and promotional activities on demand fluctuations, aiming to further refine forecasting accuracy and improve real-world retail applications.

References

[1] *Stanford. (2015) Predict Future Sales. Kaggle. Retrieved on 2024, Retrieved from: https://www.kaggle.com/c/competitive-data-science-predict-future-sales.*
[2] *Breiman, L. (2001) Random forests. Machine learning, 45: 5-32.*
[3] *Chen T, Guestrin, C. (2016) Xgboost: A scalable tree boosting system, Proceedings of the acm sigkdd international conference on knowledge discovery and data mining, 785-794.*
[4] *Ke, G., Meng, Q., Finley, T., et al. (2017) Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.*
[5] *Moon, M.A., Mentzer, J.T., Smith, C.D. (2003) Conducting a sales forecasting audit. International Journal of Forecasting, 19(1): 5-25.*
[6] *Lawrence, M., O'Connor, M., Edmundson, B. (2000) A field study of sales forecasting accuracy and processes. European Journal of Operational Research, 122(1): 151-160.*

DOI: 10.54254/2754-1169/128/2024.18672

[7] *Davis, D.F., Mentzer, J.T. (2007) Organizational factors in sales forecasting management. International Journal of Forecasting, 23(3): 475-495.*

[8] *Venkataramanan, S., Sadhu, A.K.R., Gudala, L., et al. (2024) Leveraging Artificial Intelligence for Enhanced Sales Forecasting Accuracy: A Review of AI-Driven Techniques and Practical Applications in Customer Relationship Management Systems. Australian Journal of Machine Learning Research & Applications, 4(1): 267-287.*

[9] *Sohrabpour, V., Oghazi, P., Toorajipour, R., et al. (2021) Export sales forecasting using artificial intelligence. Technological Forecasting and Social Change, 163, 120480.*

[10] *He, Q.Q., Wu, C., Si, Y.W. (2022) LSTM with particle Swam optimization for sales forecasting. Electronic Commerce Research and Applications, 51, 101118.*

[11] *Ma, S., Fildes, R. (2021) Retail sales forecasting with meta-learning. European Journal of Operational Research, 288(1): 111-128.*

[12] *Shilong, Z. (2021) Machine learning model for sales forecasting by using XGBoost. IEEE international conference on consumer electronics and computer engineering, 480-483.*

[13] *Gandhi, M.A., Maharram, V.K., Raja, G., et al. (2023) A novel method for exploring the store sales forecasting using fuzzy Pruning LS-SVM approach. International Conference on Edge Computing and Applications, 537-543.*