Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/59/20231135

Bicycle Sales Prediction Based on Ensemble Learning

Bin Yu[1,a,*]

[1]*Department of Industrial Engineering, Capital University of Economics and Business, Beijing, China*
a. 32021210168@cueb.edu.cn
**corresponding author*

Abstract: In the field of sales forecasting, there are still various challenges in conducting comprehensive analysis and accurate predictions for bicycle sales, including the diversity of sample data, the range of research scope, and the methods employed. This study aims to fill this research gap by applying a bicycle sales dataset and two ensemble learning methods to investigate the factors influencing bicycle sales and conduct sales predictions and analysis. The research findings indicate that cost, profit, and income are the most significant factors influencing bicycle profit predictions. Compared to the Random Forest model, the Gradient Boosting model performs better in predicting bicycle profits. This paper discusses the relevance and predictive performance of the bicycle sales dataset, providing opportunities for improvement and further optimization in future research to enhance the accuracy and reliability of bicycle sales predictions and offer valuable insights for decision-making and planning. Overall, these results shed light on guiding further exploration of sales prediction.

Keywords: sales forecasting, ensemble learning, the Random Forest model, the Gradient Boosting model

1. Introduction

The field of sales forecasting has seen significant advancements over the years, with businesses across various industries striving to accurately predict future sales and meet customer demands. In this context, bicycle sales forecasting has emerged as a crucial area of study, driven by the growing popularity of cycling as a mode of transportation, recreation, and fitness. The bicycle industry has witnessed substantial growth in recent years, with an increasing number of individuals adopting cycling as a sustainable and healthy lifestyle choice. As a result, businesses in the bicycle market face the challenge of accurately predicting sales to ensure optimal inventory management, production planning, and customer satisfaction.

In recent years, with the proliferation of data availability and advancements in machine learning techniques, significant research achievements have been made in the field of sales forecasting. Christoph and Arnd used the case and data of Canyon Bicycles, a German high-end bicycle manufacturer and online retailer, to illustrate a method of estimating demand distribution [1]. They first suggested determining the optimal team composition, and then prepared and conducted a prediction meeting. In the prediction meeting, team members can propose demand predictions by discussing and sharing their professional knowledge and experience. Then, the author introduces a method of removing bias to ensure the accuracy of team predictions. Li's work from the year 2023

DOI: 10.54254/2754-1169/59/20231135

concentrated on extracting local correlations and sequences from multivariate time series (MTS), but he rarely thought about modeling the unique information present in each time series. developed a system known as the Universality-Discrimination Mechanism (UDM) that can forecast future multi-step sales [2] Carla reasoned that sales time series or final customers' data in geographical disaggregation are unavailable for businesses planning to enter new geographic areas. As a result, he combined two literature streams (spatial marketing and sales forecasting) and proposed a new hybrid probabilistic approach: Gravitational Sales Prediction (GSP) [3]. When compared to a reliable benchmark that is based on historical sales proportions, they used sales data from two countries and more than ten economic sectors and came to the following conclusion: GSP exceeded expectations by not only matching the benchmark's performance but also outperforming it in some areas [3]. According to Lyu, sales prediction models employing big data on online word-of-mouth (eWOM) are still inaccurate. However, eWOM also incorporates the heat and sentiments of product dimensions, which can increase the accuracy of prediction models based on multi-attribute attitude theory. In order to anticipate daily sales, they therefore suggested an autoregressive heat-sentiment (ARHS) model that incorporates the heat and sentiments of dimensions into the benchmark prediction model. Scholars undertake an empirical analysis of the film business and find that the ARHS model predicts box office receipts for movies more accurately than other models [4].

Reviewing the existing literature, it is evident that there is a lack of comprehensive analysis and accurate predictions bicycle sales forecasting, caused by variations in sample data, research scope, and methodologies employed [5-9]. In light of this research gap, this study aims to examine the factors influencing bicycle sales and Utilize ensemble learning to forecast and analyze the sales. This article aims to explore the bicycle sales forecasting by examining the dataset &two ensemble learning methods, the obtained results &the implications and the limitations of the study.

2. Data and Method

2.1. Data

The dataset used in this study is sourced from a bicycle sales dataset available on the machine learning platform Kaggle. The original dataset consists of 89 observations, which were subsequently cleaned, resulting in a final dataset of 85 observations. It can be categorized as a small-scale dataset. The dataset comprises a total of 18 variables, including time-related variables such as 'Date', 'Month', and 'Year', user-related variables such as 'Customer_Age' and 'State', product-related variables such as 'Sales_Order' and 'Unit_Price', and the ultimate predictive variable, 'Profit'. The key variables utilized in the ensemble learning approach are 'Cost', 'Revenue', 'Bike_Size', 'Unit_Cost', 'Brand', 'Order_Quantity', and 'Profit'.

2.2. Method

Gradient boosting is a learning process that combines the results of numerous simple predictors to create an effective committee that performs better than the individual members. Usually, decision trees with a fixed size are employed as the basis learners in this method [10]. The primary mathematics used in GB models include gradient descent optimization and loss functions. Gradient descent is an optimization algorithm used to minimize the loss function of the model. In the context of GB, it continually adjusts new models to account for residuals or errors from earlier models. The residuals are computed by taking the difference between the predicted and actual values of the target variable. The subsequent models are trained to predict these residuals, and their predictions are added to the previous models' predictions to improve the overall model performance. Loss functions play a crucial role in determining the quality of the predictions made by the GB model. Mean squared error (MSE) and binary cross-entropy are among the loss functions frequently utilized, depending on the

DOI: 10.54254/2754-1169/59/20231135

nature of the problem being addressed. These loss functions quantify the discrepancy between the predicted and actual values and guide the optimization process towards minimizing the errors. Additionally, the GB model employs techniques such as tree pruning, regularization, and learning rate adjustment, which further contribute to its mathematical foundation and help prevent overfitting.

The Random Forest (RF) algorithm is an algorithm proposed by Leo Breiman and Adele Cutler. It uses the idea of bagging algorithm to randomly sample and select samples from the population, and randomly select a part of the features to form a CART decision tree. The final result is obtained by repeating it multiple times. Its essence is to combine multiple decision trees together, and the establishment of each tree relies on new independently extracted random samples to form a forest, which is an improved algorithm [11]. The fundamental concept behind the random forest algorithm lies in its dual random process: random sample sampling and random feature sampling. This key characteristic grants the random forest algorithm several advantages during the model fitting process. Firstly, the algorithm generates decision trees by randomly selecting subsets of samples and features, effectively mitigating the risk of overfitting and enhancing its robustness against noise. Additionally, the random forest algorithm demonstrates proficiency in handling high-dimensional data, making it well-suited for parallel computing and relatively straightforward to implement. Overall, considering the application of the random forest model within the domain of bicycle profit forecasting, it proves to be a highly suitable choice.

When evaluating Random Forest (RF) and Gradient Boosting (GB) models, two widely used indicators are R-squared (R2) and Mean Squared Error (MSE). These indicators provide valuable insights into the performance and predictive capabilities of the models. R-squared (R2) is a statistical metric that measures the extent to which the independent variables explain the variability in the dependent variable. Ranging from 0 to 1, with 1 representing a perfect fit, a higher R2 value indicates a stronger model fit. Specifically, in the context of RF and GB models, a higher R2 value suggests a greater alignment between the model and the data. This implies that the model's predictions capture a larger portion of the target variable's variability. Essentially, R2 quantifies the model's ability to explain and predict the outcome variable based on the employed features. Mean Squared Error (MSE) is a commonly used metric to assess the performance of regression models, including RF and GB. It quantifies the average squared difference between the predicted and actual values. MSE provides an objective measure of the model's accuracy and ability to predict the target variable. A lower MSE indicates better performance, as it signifies that the predicted values are closer to the actual values on average. MSE is particularly useful for comparing different models, as it provides a clear measure of the average prediction error. In summary, R2 evaluates the proportion of explained variance and indicates how well the model fits the data, while MSE quantifies the average prediction error. By considering these indicators, one can assess the predictive capabilities and accuracy of both RF and GB models in a more formal and statistically rigorous manner.

3. Results and Discussion

3.1. Correlation Analysis

In the correlation analysis, this study examined the relationships between various features (Cost, Profit, Revenue, Order_Quantity, Brand, Unit_Cost, Bike_Size) and the target variable (Profit). By calculating the correlation coefficients, this study was able to measure the strength and direction of the linear relationship between these variables. One delves deeper into these correlations and provide more insights into the relationships. Firstly, the cost of the bikes displayed a moderately strong positive correlation with profit, with a correlation coefficient of 0.44. This finding suggests that higher costs tend to lead to higher profits. Several factors could contribute to this correlation. For instance, higher-priced bikes might have higher profit margins, allowing for increased profitability.

DOI: 10.54254/2754-1169/59/20231135

Additionally, bikes with higher costs might be associated with better quality or additional features, attracting customers willing to pay a premium, thus further contributing to higher profits. Unsurprisingly, profit exhibited a strong positive correlation with itself (correlation coefficient of 1.00). This is because profit is the target variable this study is trying to predict, and it directly determines the success and financial viability of the business. Therefore, this correlation serves as a validation of our analysis. Moving on, one observed a moderately strong positive correlation between revenue and profit, with a correlation coefficient of 0.31. This implies that higher revenues generally result in higher profits. The reason behind this relationship is quite straightforward – as revenue increases, there is a greater potential for generating profit. Higher revenues can be achieved through increased sales volume, higher prices, or a combination of both.

Interestingly, the order quantity exhibited a very weak positive correlation with profit, with a correlation coefficient of 0.02. This suggests that larger order quantities may slightly contribute to higher profits. One possible explanation for this relationship is that larger order quantities may lead to economies of scale, reducing per-unit costs and increasing profit margins. However, the weak correlation suggests that other factors might have a more significant impact on profitability. Similarly, the brand of the bikes showed a very weak positive correlation with profit, with a correlation coefficient of 0.01. This indicates that certain brands may have a slight influence on profits. Brand reputation, customer loyalty, and the perceived value associated with specific brands could contribute to this slight correlation. However, other factors such as cost, and revenue may have a more substantial impact on profitability. On the other hand, both unit_cost and bike_size exhibited very weak correlations with profit, with correlation coefficients of 0.00. This suggests that these features may not have a significant impact on profits. The absence of a clear relationship may indicate that other factors such as cost, revenue, and brand play more critical roles in determining profitability.

To summarize, the correlation analysis reveals that cost and revenue have the strongest relationships with profit, indicating their importance in determining the profitability of bike sales. On the other hand, the order quantity, brand, unit cost, and bike size have relatively weaker correlations, suggesting that their impact on profit may be less significant. However, it's important to note that correlation does not imply causation, and further analysis or experiments may be necessary to establish a causal relationship between these variables. In conclusion, understanding the correlations between various features and profit is essential for businesses to make informed decisions. By identifying the factors that have the most significant impact on profitability, companies can focus their resources and strategies on maximizing those factors and, in turn, optimize their overall profitability.

3.2. Model Construction

The first crucial step in any predictive modeling project is data preparation. This study starts by encoding categorical variables using the LabelEncoder from the scikit-learn library. Categorical features like 'Month,' 'Age_Group,' 'Customer_Gender,' 'Country,' 'State,' 'Brand,' and 'Bike_Colour' are transformed into numerical values, making them suitable for machine learning algorithms. Next, this study divided the dataset into training and testing subsets to evaluate our models effectively. One keeps the essential columns, including 'Cost,' 'Revenue,' 'Bike_Size,' 'Unit_Cost,' 'Order_Quantity,' 'Brand,' and remove any other columns that are not relevant to our prediction task. The train-test split is performed using the train_test_split function from scikit-learn, with an 80-20 split ratio. One initializes two regression models: Random Forest and Gradient Boosting. These models are powerful ensemble learning techniques that have proven effective in various prediction tasks. This study aims to compare their performance to determine which one is better suited for predicting bicycle profits. The models are configured with a random seed (random_state) to ensure reproducibility.

DOI: 10.54254/2754-1169/59/20231135

Later, this study trained each model separately on the training data and evaluate their performance on the testing data. This study employs common regression evaluation metrics, including R-squared (R2) score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). - Random Forest Model: - R2 Score: 0.97 - MAE: 90.77 - RMSE: 266.56 - Gradient Boosting Model: - R2 Score: 0.98 - MAE: 56.02 - RMSE: 212.90. The results are summarized in Table 1. The results demonstrate that both models perform exceptionally well in predicting bicycle profits, with the Gradient Boosting model slightly outperforming the Random Forest model in terms of R2 score, MAE, and RMSE. This suggests that Gradient Boosting may be the preferred choice for this specific prediction task. Conclusion: In this project, one successfully constructed and evaluated regression models for predicting bicycle profits using Gradient Boosting (GBmode) and Random Forest (RFmode) algorithms. Both models exhibited high accuracy and reliability, with Gradient Boosting showcasing a slight advantage. The choice between these models may depend on specific business needs and computational resources. These models can be further fine-tuned and deployed to assist businesses in making informed decisions regarding their bicycle sales and profitability. Predictive modeling continues to be a valuable tool for organizations seeking data-driven insights to drive growth and success.

Table 1: Comparison of RFmodel and GBmodel.

Evaluation Metrics	Random Forest	Gradient Boosting
R2 Score	0.97	0.98
MAE	90.77	56.02
RMSE	26.56	212.90

3.3. Evaluation and Explanation

The evaluation of our regression models for predicting bicycle profits using Random Forest (RFmodel) and Gradient Boosting (GBmodel) has provided valuable insights into their performance. These insights are essential for understanding the reliability and accuracy of the models and can guide decision-making in practical applications. Both models demonstrated exceptional performance in predicting bicycle profits, with high R2 scores indicating their ability to explain a significant portion of the variance in the profit data. The Gradient Boosting model slightly outperformed the Random Forest model, achieving an R2 score of 0.98 compared to 0.97. This suggests that Gradient Boosting may be the preferred choice for achieving the highest predictive accuracy. In terms of error metrics, the Gradient Boosting model exhibited lower Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values compared to the Random Forest model. The Gradient Boosting model had an MAE of 56.02 and an RMSE of 212.90, while the Random Forest model had an MAE of 90.77 and an RMSE of 266.56. These results indicate that the predictions from the Gradient Boosting model were closer to the actual profit values, resulting in smaller prediction errors. Furthermore, the importance analysis identified key factors influencing bicycle profit predictions. The 'Cost' feature emerged as the most influential with an importance score of 0.44, emphasizing the significance of cost control for maximizing profits. 'Profit' and 'Revenue' were also highly important for profit predictions, with scores of 0.31 and 0.23, respectively. Considering 'Order_Quantity' and 'Brand' in decision-making, although they had lower importance scores, can still contribute to profitability.

4. Limitations & Future Outlooks

Although this study has conducted research on bicycle sales prediction using a bicycle dataset, there are still some limitations that need to be addressed. Firstly, the study only utilizes two ensemble learning methods (Gradient Boosting and Random Forest), while there are many other alternative

DOI: 10.54254/2754-1169/59/20231135

techniques available, such as lightGBM, XGBoost, and AdaBoost, which may perform better in predicting bicycle sales. Secondly, the dataset used in this study is sourced from the Kaggle platform and after cleaning, it consists of only 85 usable observations, which may not be sufficient to build accurate prediction models. Additionally, although multiple variables have been considered in this study, there are still other factors that might be relevant to bicycle sales but have not been included, such as national policy support and geographical terrain, which could have significant impacts on sales. To address these issues, future research can be conducted in the following areas:

- Explore more ensemble learning methods: In addition to the two used methods (Gradient Boosting and Random Forest), the study can further investigate other ensemble learning techniques such as lightGBM, XGBoost, and AdaBoost. By comparing the performance of different methods in predicting bicycle sales, the most suitable model can be identified.
- Increase the dataset size: The study can seek additional bicycle sales datasets to increase the sample size. This can improve the effectiveness of the models' training and make the prediction results more reliable and accurate.
- Incorporate more relevant factors: In addition to the already considered factors, the study can further explore other potential factors related to bicycle sales, such as the level of national policy support and geographical terrain. Including these factors can enhance the predictive ability of the models and provide a more comprehensive understanding of the factors influencing bicycle sales.
- Conduct more comprehensive data cleaning and preprocessing: The study can perform more detailed data cleaning and preprocessing to ensure the quality and completeness of the dataset. This can reduce the impact of data bias and noise, and improve the accuracy of the models.
- Introduce time series analysis methods: Considering the potential time-related patterns in bicycle sales, the study can explore time series analysis methods such as ARIMA, SARIMA, and Prophet to forecast future bicycle sales more accurately. These methods can capture the seasonality, trends, and cyclicality of sales.

Perform feature engineering: In addition to the existing variables, the study can conduct more in-depth feature engineering, such as creating interaction features, polynomial features, and incorporating domain-specific knowledge-related features. This can enhance the models' explanatory power for bicycle sales and better capture the underlying influencing factors. Conduct model parameter tuning: For the selected models, the study can perform more detailed model parameter tuning to find the optimal parameter combinations. Through parameter optimization, the prediction accuracy and stability of the models can be further improved.By implementing these proposed improvements, future research can enhance the accuracy and reliability of bicycle sales prediction, providing valuable insights for decision-making and planning.

5. Conclusion

In conclusion, this article aimed to explore bicycle sales forecasting by examining the dataset and utilizing two ensemble learning methods, Gradient Boosting and Random Forest. The results of the study revealed several key findings. Firstly, cost, revenue, and order quantity exhibited significant correlations with profit, indicating their importance in determining the profitability of bicycle sales. On the other hand, variables such as brand, unit cost, and bike size had relatively weaker correlations with profit, suggesting their lesser impact on profitability. The regression models constructed using Gradient Boosting and Random Forest algorithms showcased high accuracy and reliability, with Gradient Boosting slightly outperforming Random Forest in terms of prediction accuracy metrics. Additionally, the importance analysis identified cost as the most influential factor in maximizing profits, emphasizing the significance of cost control. Limitations of the study included the small sample size of the dataset and the use of only two ensemble learning methods. To enhance future research, it is suggested to incorporate more relevant factors, such as national policy support and

DOI: 10.54254/2754-1169/59/20231135

geographical terrain, and explore additional ensemble learning techniques like lightGBM, XGBoost, and AdaBoost. Overall, this research provides valuable insights for decision-making and planning in the bicycle industry and highlights the significance of understanding the factors influencing bicycle sales to optimize profitability.

References

[1] *Christoph D.,and Arnd H. (2017) Case Article—Canyon Bicycles: Judgmental Demand Forecasting in Direct Sales. INFORMS Transactions on Education 17(2):58-62.*

[2] *Li, D., Li, X., Gu, F., Pan, Z., Chen, D., and Madden, A. (2023). A Universality–Distinction Mechanism-Based Multi-Step Sales Forecasting for Sales Prediction and Inventory Optimization. Systems, 11(6), 311.*

[3] *Silveira Netto, C.F., Bahrami, M., Brei, V.A., Bozkaya, B., Balcisoy, S., and Pentland, A.P. (2023). Disaggregating Sales Prediction: A Gravitational Approach. Expert Systems with Applications, 217, 119565.*

[4] *Lyu, X., Jiang, C., Ding, Y., Wang, Z., and Liu, Y. (2019). Sales Prediction by Integrating the Heat and Sentiments of Product Dimensions. Sustainability, 11(3), 913.*

[5] *Stephens, T. (2017). A Model for Sales Forecasting and Stock Management. In Business Analytics: Progress on Applications in Asia Pacific (pp. 565-588).*

[6] *Kolkova, A., and Rozehnal, P. (2022). Hybrid demand forecasting models: pre-pandemic and pandemic use studies. Equilibrium. Quarterly Journal of Economics and Economic Policy, 17(3), 699-725.*

[7] *Hall, J. (2020). Forecasting off-street bicycle facility demands (Doctoral dissertation).*

[8] *Norang, A., Eghbali, M.A., and Hajian, A. (2010, January). Supply chain analysis model based on system dynamics approach: a case of Iranian bicycle manufacturer. In 2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM) (Vol. 3, pp. 1481-1485). IEEE.*

[9] *Whitlark, D.B., Geurts, M.D., and Swenson, M.J. (1993). New product forecasting with a purchase intention survey. The Journal of Business Forecasting, 12(3), 18.*

[10] *Biau, G., Cadre, B., and Rouvière, L. (2019). Accelerated gradient boosting. Machine Learning, 108(6), 971-992.*

[11] *Sun, Y. (2022). Research on electricity revenue prediction based on the RF-XGBoost model (Master's thesis, Yanbian University).*