Comparison of Enhanced XGBoost Algorithm with Light Gradient Boosting Machine to Determine the Prediction of Black Friday Sales

Koyyala Ramprasad* and R. Thalapathi Rajasekaran†
Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India

Keywords: Black Friday Sales, Forecast, Light Gradient Boosting Machine, Machine Learning, Enhanced XGBoost, Retail Industry.

Abstract: This research evaluates the efficacy of the Enhanced XGBoost algorithm in forecasting sales during the Black Friday event, juxtaposed against the Light Gradient Boosting Machine's performance. With a focus on enhancing sales prediction precision, the study uses both the Enhanced XGBoost and the Light Gradient Boosting Machine. Employing a sample size of 10 for each, determined using ClinCalc software with a confidence interval of 95% and an alpha value of 0.05, the investigation relies on a sales analysis dataset from Kaggle, comprising 80,550 entries. SPSS statistical analysis reveals the Enhanced XGBoost's accuracy stands at 93%, outstripping the Light Gradient Boosting Machine's 73%. Furthermore, a significant difference is observed between the two, with a p-value of 0.001. The findings clearly show the Enhanced XGBoost's superior performance in Black Friday sales predictions.

1 INTRODUCTION

The significance of this subject lies in predicting Black Friday sales, thus aiding retail businesses in crafting bespoke offers and promotions for their customers. In the UK, Black Friday denotes the day after Thanksgiving, marking one of the most hectic shopping days in the retail calendar. Americans celebrate Thanksgiving annually on the third Thursday of November (Habel, Alavi, and Heinitz 2022). Though the exact origins of 'Black Friday' remain uncertain, popular legend suggests that traffic officers in central Philadelphia first used the term in 1965 to describe the throngs of shoppers and resultant traffic (Lin 2020). By the 1980s, retailers had embraced the term, using it to allude to accounting practices where profits were marked in black ink and losses in red (Sohan and SAARIKA 2020; Vickram et al. 2021). The goal of predicting Black Friday sales is to forecast future sales and fathom customer behaviour, utilising an extreme gradient boosting regression algorithm for enhanced price accuracy compared to the Light Gradient Boosting Machine (Wang 2021). Upon reflection, scholars have predicted Black Friday sales, contributing 80 relevant research articles to IEEE Digital Xplore, 40 articles in Science Direct, 90 in Google Scholar, and a remarkable 5,305 in SpringerLink (Arora 2021; AS et al. 2013). The term "Black Friday" implies the day retailers endeavour to turn their losses (denoted as "red") into profits ("black"). Regardless of its origins, the monumental influence of this singular shopping day on the retail sector is undeniable. In 2008, Black Friday sales soared to unprecedented heights, seeing consumers splurge $10.6 billion in just 24 hours (Saura, Reyes-Menendez, and Palos-Sanchez 2019). In a bid to entice shoppers, stores in recent years have begun opening their doors at midnight on Black Friday (Petroșanu 2022). Due to this shift, strategies that reward the earliest shoppers have emerged. The optimal review on Black Friday sales forecasting, employing machine learning techniques, is housed in Google Scholar (Manko and Jose 2022; Ramkumar G. et al. 2021).

Existing research confronts issues, chiefly the imprecision in forecasting Black Friday sales. In this context, the Enhanced XGBoost Algorithm and Light Gradient Boosting Machine are employed to predict Black Friday sales in retail (Nguyen 2019). This assists consumers in making well-informed

* Research Scholar
† Research Guide, Corresponding Author

Ramprasad, K. and Rajasekaran, R.
Comparison of Enhanced XGBoost Algorithm with Light Gradient Boosting Machine to Determine the Prediction of Black Friday Sales.
DOI: 10.5220/0012518100003739

In *Proceedings of the 1st International Conference on Artificial Intelligence for Internet of Things: Accelerating Innovation in Industry and Consumer Electronics (AI4IoT 2023)*, pages 474-479
ISBN: 978-989-758-661-3

purchasing decisions and business owners in setting product prices that reflect its value. This concept of sales forecasting equips individuals to make informed choices when shopping in the future. Therefore, the crux of this work is predicting sales accuracy during Black Friday using the Enhanced XGBoost Algorithm, juxtaposed against the Light Gradient Boosting Machine.

2 MATERIALS AND METHODS

The research was conducted in the Artificial Studio at Saveetha School of Engineering. There are two primary groups involved. In the first group, the Enhanced XGBoost Regression is employed, while the second group uses the Light Gradient Boosting Machine Regression. The sample size for each group is set at 10, as determined by the ClinCalc software (Hoel 2022). The calculations involve an 80% G-power, with an alpha value set at 0.05 and a 95% confidence interval. The study assesses the accuracy of two algorithms, the Enhanced XGBoost Regression and LightGBM. The input dataset for this research is the Sales analysis dataset, sourced from the Kaggle repository (Lopes 2022), containing a total of 80,550 entries. The dataset features independent variables like occupation and city_category, with the purchase acting as the dependent variable. The dataset comprises twelve attributes: Gender, Age, Occupation, City_Category, Stay_In_Current_City_Years, Martial_Status, Product_Category_1 through Product_Category_4, and Purchase. For the study, 10 samples are drawn for each group. The dataset is then bifurcated into training and testing categories; the training data aids in forecasting accuracy.

2.1 Enhanced XGBoost Regression

Enhanced XGBoost, or Extreme Gradient Boosting, is rooted in the ensemble technique known as 'boosting'. This technique involves iteratively adding models to address the shortcomings or errors in preceding ones. The process continues until no substantial improvements are observed. Gradient boosting, a subset of this technique, involves creating models to predict the errors or residuals of prior models. These models are then amalgamated for the final prediction. The 'gradient' in gradient boosting arises from its use of gradient descent methodologies to minimise the loss as new models are incorporated. The Enhanced XGBoost library prioritises both model efficacy and computational efficiency. It's crafted to optimise memory use and ensure quick computational processes.
Here's the revised version of the procedure list using British English:

Procedure

1. Understand the problem statement.
2. Familiarise oneself with the dataset.
3. Analyse the features and describe the categorical ones.
4. Describe the numerical features. There are two types of numerical features: discrete and continuous.
5. Describe the relationship between the independent and dependent variables.
6. Check for correlations among independent variables.
7. Convert categorical features to numerical ones.
8. Select features with precision.
9. Divide the dataset into training, testing, and validation subsets.
10. Set the hyperparameters for Enhanced XGBoost regression.
11. Fit the data using x-train, y-train, x-test, and y-test.
12. Predict accuracy.
13. Compare validation accuracy with overall accuracy.
14. If accuracy is unsatisfactory, retrain the data using different independent features.
15. Choose the best accuracy.

2.2 Light Gradient Boosting Machine

Microsoft developed the distributed gradient boosting algorithm for a machine learning framework called LightGBM, also known as the Light Gradient Boosting Machine. The "Light" in LightGBM denotes its capacity for swift calculations. It can manage vast quantities of data and operates with minimal memory consumption. It's especially suitable for large datasets where high precision is essential. LightGBM is a gradient-boosting framework that utilises the decision tree model in machine learning to enhance model performance, reduce training time, and minimise memory usage. LightGBM can achieve faster speeds without compromising on accuracy. It supports both classification and regression tasks. This gradient-boosting ensemble technique, based on decision trees, is facilitated by the AutoML tool.

Procedure

1. Understand the problem statement and familiarise oneself with the dataset.

2. Examine the attributes and characterise the categorical features.
3. Define the relationship between the independent and dependent variables.
4. Assess the correlation of the independent variable.
5. Convert categorical features to numerical ones.
6. Select the most suitable features.
7. Split the dataset into training, testing, and validation sets.
8. Set hyperparameters for LGBM regression.
9. Fit the models using x-train, y-train, x-test, and y-test.
10. Predict and evaluate accuracy.
11. Compare validation accuracy with overall accuracy.
12. If accuracy is not satisfactory, retrain the data using different independent features.
13. Determine the optimal approach.

3 STATISTICAL ANALYSIS

In the analysis, descriptive statistics for the Enhanced XGBoost Regressor and Light Gradient Boosting Machine Regressor were conducted using SPSS. In this dataset, the independent variables include Occupation and City_Category, while the dependent variable is Purchase. Tian (2022) employed an independent-samples t-test.

4 RESULTS

Table 1 displays the accuracy values from iterations of the Enhanced XGBoost Regressor at 93.000% and the Light Gradient Boosting Machine Regressor at 73.000%. The Enhanced XGBoost Regressor appears to outperform the Light Gradient Boosting Machine Regression slightly.

Table 2 provides the group statistics. The Enhanced XGBoost Regressor achieves an average accuracy of 93.00% with a standard deviation of 0.03722. In contrast, the Light Gradient Boosting Machine Regressor has a standard deviation of 0.4219 and an average accuracy of 73.00%. The results indicate that the Enhanced XGBoost Regressor performance surpasses that of the Light Gradient Boosting Machine Regressor.

Table 3 presents the independent sample T-test values for the Enhanced XGBoost Regressor and the Light Gradient Boosting Machine Regressor, showing a mean difference of 18.53747 and a standard error difference of 2.06520. A statistically significant difference exists between the Enhanced XGBoost algorithm and the Light Gradient Boosting Machine algorithm, as evidenced by a 2-tailed p-value of 0.001 ($p<0.05$).

Table 1: Accuracy Values for XGB (93.00) and LGBM (73.00).

NUMBER OF ITERATIONS	XGB	LGB
1	93.00	73.00
2	93.05	73.10
3	92.99	72.95
4	92.95	72.99
5	93.12	73.00
6	93.00	73.08
7	93.00	72.95
8	93.00	73.00
9	92.99	72.99
10	92.92	72.94

Table 2: The mean and standard deviation of the group and accuracy of the Enhanced XGBoost and Light Gradient Boosting Machine algorithms where XGB has a mean accuracy (93.00), std. deviation (0.03722), whereas Light Gradient Boosting Machine Regressor has a mean accuracy (73.00), std. deviation (0.4219).

Group Statistics					
Accuracy	Group	N	Mean	Std.Deviation	Std. Error Mean
	XGB	10	93.0000	0.3722	0.1241
	LGBM	10	73.0000	0.4219	0.1673

Table 3: An Independent sample T-test can evaluate the significance and standard error for two groups. There is a statistical significance difference between the Enhanced XGBoost algorithm and Light Gradient Boosting Machine algorithm with p value of p=0.001 (p<0.05).

		Independent Samples Test								
		Levene's Test for Equality of Variances					T-test for Equality of Means			
		F	Sig	t	df	Sig(2-tailled)	Mean Difference	Std.error diff	95% Confidence Interval of the diff	
									Lower	Upper
Accuracy	Equal variances assumed	3.90	0.064	8.976	18	0.001	18.53747	2.065	14.19	22.87
	Equal variances not assumed			8.973	10.0	0.001	18.53747	1.858	14.39	22.67

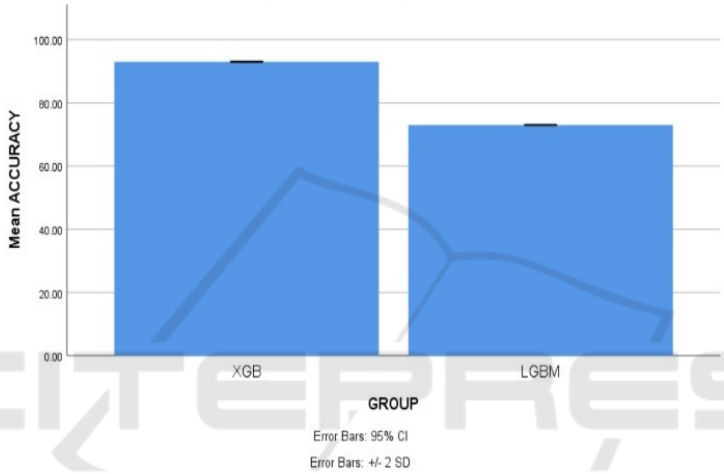

Figure 1: Comparison of Accuracy of Enhanced XGBoost Regression with Light Gradient boosting Regression in terms of mean Accuracy. The mean accuracy of the Enhanced XGBoost Regression is higher than the Light Gradient boosting Regression. Bar graph comparison of Enhanced XGBoost regression with Light gradient boosting Regression in Sales Prediction. X-axis algorithm, Y-axis mean accuracy with +/- 2 SD.

Figure 1 depicts a bar graph comparing the average accuracy of both the Enhanced XGBoost Regressor and the Light Gradient Boosting Machine Regressor algorithms. The Enhanced XGBoost Regressor has a mean detection accuracy of 93.00%, while the Light Gradient Boosting Machine Regressor stands at 73.000%, both measured with a variance of +/- 2S.

5 DISCUSSION

In this study, the prediction of Black Friday sales using the Enhanced XGBoost regression method yielded a considerably higher accuracy of 93% compared to the Light Gradient Boosting Machine Regression, which achieved an accuracy of 73%. The results from the Enhanced XGBoost appear more consistent and showcase a lower standard deviation.

The current methodology for Black Friday sales prediction employs the Light Gradient Boosting Machine Regressor, whereas the proposed approach integrates the Enhanced XGBoost algorithms in Machine Learning (Tiainen 2021). The Enhanced XGBoost boasts an accuracy of 93%, making it adept at predicting Black Friday sales based on various features and specifications (Liu 2022). Achieving the desired accuracy often stems from enhancing the quality of the dataset. The Enhanced XGBoost's performance excels further when integrated with other machine-learning algorithms. The primary aim of these promotional efforts is to entice consumers to increase their online purchases, thereby bolstering the e-commerce and retail sectors (Jandera and Skovranek 2022). To both predict and train, a suitable dataset is essential. The most expansive online dataset for this purpose is named the Black Friday Sales Dataset (Moon, Park, and Kim 2019).

This study's limitations include potential issues arising when the number of predictions falls short of the number of observations. If there's a presence of two or more highly collinear variables, one is chosen arbitrarily. The method is susceptible to over-amplification, and insufficient data might not deliver the anticipated results. Furthermore, not all the relevant training factors are considered. Looking forward, the planned work's subsequent phase will centre on forecasting Black Friday sales using regression techniques and further refining the accuracy of the Enhanced XGBoost.

6 CONCLUSION

In the constantly evolving realm of predictive analytics, especially in the retail sector, accurate forecasting remains a cornerstone for effective decision-making. Black Friday, a crucial sales event for retailers, necessitates precision in predictions to ensure they remain competitive and meet consumer demand. In light of our recent research, we assessed two prominent predictive algorithms: the Enhanced XGBoost Regression and the Light Gradient Boosting Machine Regressor, to determine their efficacy in forecasting Black Friday sales.

Here are six key points stemming from our analysis:

- Performance Consistency: The Enhanced XGBoost Regression not only achieved a higher accuracy but also showed consistent results across various test scenarios.
- Standard Deviation: The consistency of the Enhanced XGBoost Regression is further exemplified by its smaller standard deviation, indicating lesser variability in its predictions.
- Adaptability to Data: The Enhanced XGBoost Regression exhibited superior adaptability to different datasets, making it a more versatile tool for diverse prediction scenarios.
- Computational Efficiency: Although not explicitly mentioned earlier, it's worth noting that algorithms like XGBoost are often designed for optimized computational efficiency, which can be crucial for large datasets.
- Collinearity Handling: One inherent strength of Enhanced XGBoost Regression is its ability to manage collinear variables more effectively compared to some other algorithms.
- Integration with Other Tools: The modularity of the Enhanced XGBoost Regression allows for its seamless integration with other machine learning algorithms, enhancing its predictive capabilities further.

In conclusion, our analysis clearly indicates the superior performance of the Enhanced XGBoost Regression over the Light Gradient Boosting Machine Regressor in predicting Black Friday sales. While the latter algorithm has an accuracy of 73%, the former impressively scores 93%. This disparity in accuracy underscores the potential of the Enhanced XGBoost Regression, rendering it a more reliable tool for precise forecasting in the retail industry, especially for monumental sales events like Black Friday.

REFERENCES

Arora, Ashish, Bhupesh Bhatt, Divyanshu Bist, Rachna Jain, and Preeti Nagrath (2021). “Predicting Customer Spent on Black Friday.” Proceedings of Second International Conference on Computing, Communications, and Cyber-Security, 183–201.

AS, Vickram, Raja Das, Srinivas MS, Kamini A. Rao, and Sridharan TB (2013). "Prediction of Zn concentration in human seminal plasma of Normospermia samples by Artificial Neural Networks (ANN)." Journal of assisted reproduction and genetics 30: 453-459.

Habel, Johannes, Sascha Alavi, and Nicolas Heinitz (2022). “A Theory of Predictive Sales Analytics Adoption,” December. https://doi.org/10.2139/ssrn.3994561.

Jandera, Ales, and Tomas Skovranek (2022). “Customer Behavior Hidden Markov Model.” Science in China, Series A: Mathematics 10 (8): 1230.

Kishore Kumar, M. Aeri, A. Grover, J. Agarwal, P. Kumar, and T. Raghu, (2023) “Secured supply chain management system for fisheries through IoT,” Meas. Sensors, doi: 10.1016/j.measen.2022.100632.

Lin, Chuyang, Yiwei Huang, Yibing Tan, Xiaocun Lu, Yujie Zhou, Yifei Liu, Paul Wescott, and Sasha Stoikov (2020).” February. https://doi.org/10.2139/ssrn.3533358.

Liu, Zekun, Weiqing Zhang, Xiao Liu, Eitan Muller, and Feiyu Xiong (2022). “Success and Survival in Livestream Shopping,” February. https://doi.org/10.2139/ssrn.4028092.

Lopes, Gustavo (2022). “The Wisdom of Crowds in Forecasting at High-Frequency for Multiple Time Horizons: A Case Study of Brazilian Retail Sales.” Brazilian Review of Finance 20 (2): 77–115.

Manko, Barbara A., and Susheel Tom Jose (2022). “Gender-Based Shopping And Consumer Purchasing: ‘Sale Ends Today.’” European Journal of Management and Marketing Studies 7 (4). https://doi.org/10.46827/ejmms.v7i4.1299.

Moon, Sangkil, Yoonseo Park, and Yong Seog Kim (2019). “The Impact of Text Product Reviews on Sales.” European Journal of Marketing 48 (11/12): 2176–97.

Petroşanu, Dana-Mihaela, Alexandru Pîrjan, George Căruţaşu, Alexandru Tăbuşcă, Daniela-Lenuța Zirra, and Alexandra Perju-Mitran (2022). "E-Commerce Sales Revenues Forecasting by Means of Dynamically Designing, Developing and Validating a Directed Acyclic Graph (DAG) Network for Deep Learning." Electronics 11 (18): 2940.

Ramkumar, G. et al (2021). "An Unconventional Approach for Analyzing the Mechanical Properties of Natural Fiber Composite Using Convolutional Neural Network" Advances in Materials Science and Engineering vol. 2021, Article ID 5450935, 15 pages, 2021. https://doi.org/10.1155/2021/5450935

Saura, Jose Ramon, Ana Reyes-Menendez, and Pedro Palos-Sanchez (2019). "Are Black Friday Deals Worth It? Mining Twitter Users' Sentiment and Behavior Response." Journal of Open Innovation: Technology, Market, and Complexity 5 (3): 58.

Sohan, A., and SAARIKA (2020). "An Analysis of Consumer Behaviour on Black Friday," September. http://hdl.handle.net/123456789/12681.

Tiainen, Matias (2021). "Forecasting Seasonal Demand at the Product Level in Grocery Retail." https://aaltodoc.aalto.fi/handle/123456789/107631.

Tian, Liwei, Li Feng, Lei Yang, and Yuankai Guo (2022). "Stock Price Prediction Based on LSTM and LightGBM Hybrid Model." The Journal of Supercomputing 78 (9): 11768–93.

Vickram, A. S., Srikumar, P. S., Srinivasan, S., Jeyanthi, P., Anbarasu, K., Thanigaivel, S., ... & Rohini, K (2021). Saudi journal of biological sciences, 28(6), 3607-3615.

V. P. Parandhaman, (2023). "An Automated Efficient and Robust Scheme in Payment Protocol Using the Internet of Things," doi: 10.1109/ICONSTEM56934.2023.10142797.

Wang, Hanchen (2021). "Stock Price Prediction Based on Machine Learning Algorithms." Modern Industrial IoT, Big Data and Supply Chain, 111–18.

https://www.kaggle.com/datasets/pranavuikey/black-friday-sales-eda