Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Huế, ngày 07-08/6/2019
DOI: 10.15625/vap.2019.0007

BLACK FRIDAY SALE PREDICTION VIA EXTREME GRADIENT BOOSTED TREES

Nghia Duong Trung[1], Tan Dang Thien[2], Tien Dao Luu[3], Hiep Xuan Huynh[4]

[1]Can Tho University of Technology, Can Tho city, Vietnam
[2]Can Tho University of Technology, Can Tho city, Vietnam
[3]Can Tho University Software Center, Can Tho city, Vietnam
[4]Can Tho University, Can Tho city, Vietnam

dtnghia@ctuet.edu.vn,duong-trung@ismll.de; dttan.khmt0115@student.ctuet.edu.vn; ltdao@ctu.edu.vn; hxhiep@ctu.edu.vn

ABSTRACT: *The largest shopping day of the year in America is the Friday following the Thanksgiving holiday. It is recognized as the ignition of one of the busiest shopping seasons in a year. From the computer science point of view, one of the most interesting applications of machine learning in the retail industry is to effectively predict how much a customer is probably to spend at a store based on historical purchasing patterns. If retailers comprehensively understand their customers in terms of characteristics, behaviors and motivations in the previous shopping seasons, they can implement and develop more effective marketing strategies for specific customers categories. This study proposes an empirical implementation of extreme gradient boosted trees algorithm for addressing an interesting challenge in the retail industry. From the experimental results, the authors can conclude that the applications of bagging and boosting techniques can achieve great performance and be further improved by a proper combination of models' hyperparameters tuning and feature engineering.*

Keywords*: Ensemble Learning, Bagging and Boosting, XGBoost, Sales Prediction, Black Friday.*

I. INTRODUCTION

For a long history of several decades, Black Friday has been recognized as the largest shopping day of the year in the US. It is the Friday after Thanksgiving and for American consumers, it ignites the Christmas holiday shopping. For most retailers, it is the busiest day of the year. Black Friday is traditionally known for long lines of customers waiting outdoors in cold weather before the open hours. Sales are so high for Black Friday that it has become a crucial day for stores and the economy in general with approximate 30% of all the annual retail sales occurring in the time from Black Friday through Christmas making it the kick-off day for the busiest and most profitable season for many businesses. It is unofficially a public holiday in more than 20 states and is considered the start of the US Christmas shopping season. In 2018, US shoppers expected to drop $483.18 on the shopping holiday of holidays, which equates to $90.14 billion [1]. Although Black Friday is originally from America, it has become a universal recognition worldwide.

Because consumers are eager to spend so much money during this period, retailers seriously look forward to good preparation for the shopping holiday [2]. In preparation for this day, retailers will typically hire more employees, stock their commodities, prepare new promotions, and decorate store layouts. Retailers rely on designing advertising campaigns to attract more customers into their stores and/or their online shops. In order to maximize their efforts and revenues, retailers enthusiastically understand how the consumers make shopping decisions that will assist them to achieve the most profits during the shopping season [3]. Many possible parameters that have been considered are presented in Figure (1). If retailers comprehensively understand their customers in terms of characteristics, behaviors and motivations in the previous shopping seasons, they can implement and develop more effective marketing strategies for specific customers categories [4,5,6].

The authors place the Black Friday challenge is an interesting opportunity to investigate the performance of several machine learning models. We decide to apply boosting-based models to the problem and see how would they perform. The objective is to predict the amount of purchase a consumer is willing to pay given several categorical and numerical features.

The rest of this paper is as follows. First of all, in Section II, we briefly discuss the overview of technical background including ensemble learning, bagging and boosting that summarizes critical materials existing in the literature that is essential to solving the problems. In Section III, we evaluate and perform the approach to our experiment dataset. In Section IV, the authors discuss some important remarks of the proposed approach. Finally, Section V recapitulates the approaches, discuss achievements and further investigation.

II. MATERIAL AND METHODS

A. Ensemble Learning

In practice, it may not be effective to entirely rely upon the performance of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners referring to learning a weighted combination of base models of the form

$$f(y|\mathbf{x},\sigma) = \sum_{m\in\mathcal{M}} w_m f_m(y|\mathbf{x}) \ , \tag{1}$$

where w_m is tunable parameters and f_m is called a base model.

The resultant is a single model which gives the aggregated output from several models. The aggregation models could be either from the same learning algorithm or different learning algorithms. Bagging and boosting are two widely used ensemble learners. Though these two techniques can be used with several statistical models, the most predominant usage has been with decision trees. One of the obvious reason is that we can grow as how many trees as needed with ease of computation.

Figure 1. Possible parameters that influence the customers' expenditure in Black Friday

B. Bagging and Boosting

While decision trees are one of the most easily interpretable models, they exhibit highly variable behavior. Consider a single training dataset that we randomly split into two parts. Now, let's use each part to train a decision tree in order to obtain two models. When we fit both these models, they would yield different results. Decision trees are said to be associated with high variance due to this behavior. Bagging or boosting aggregation helps to reduce the variance in any learner. Several decision trees which are generated in parallel, form the base learners of bagging technique. Data sampled with replacement is fed to these learners for training. The final prediction is the averaged output from all the learners.

In boosting, the trees are built sequentially such that each subsequent tree aims to reduce the errors of the previous trees. Each tree learns from its predecessors and updates the residual errors. Hence, the tree that grows next in the sequence will learn from an updated version of the residuals. The base learners in boosting are weak learners in which the bias is high, and the predictive power is just a tad better than random guessing. Each of these weak learners contributes some vital information for prediction, enabling the boosting technique to produce a strong learner by effectively combining these weak learners. The final strong learner brings down both the bias and the variance. In contrast to bagging techniques like Random Forest, in which trees are grown to their maximum extent, boosting makes

use of trees with fewer splits. Such small trees, which are not very deep, are highly interpretable. Parameters like the number of trees or iterations, the rate at which the gradient boosting learns, and the depth of the tree, could be optimally selected through validation techniques like k-fold cross validation.

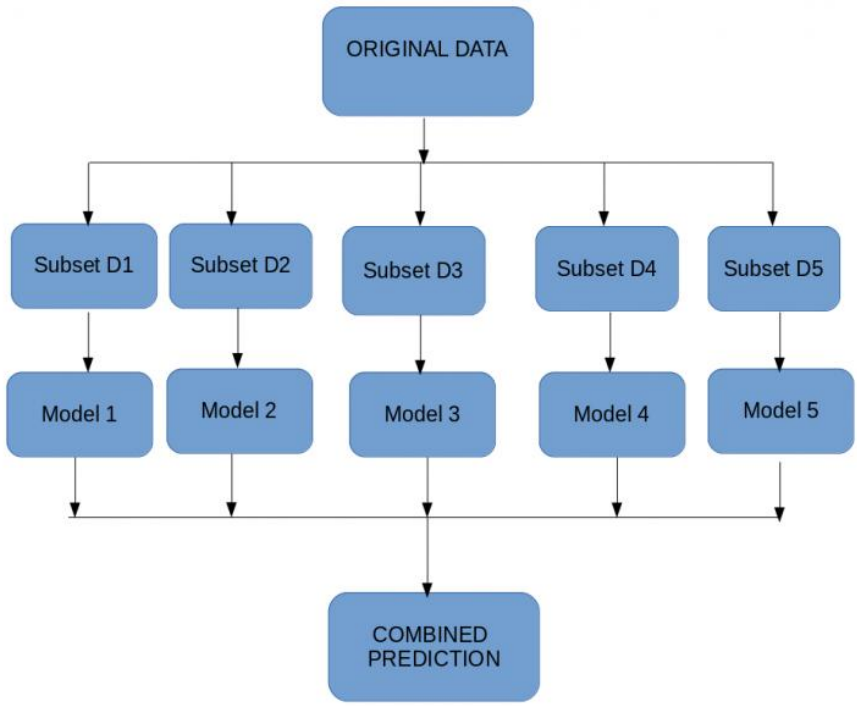

Figure 2. Combination of different models in bagging

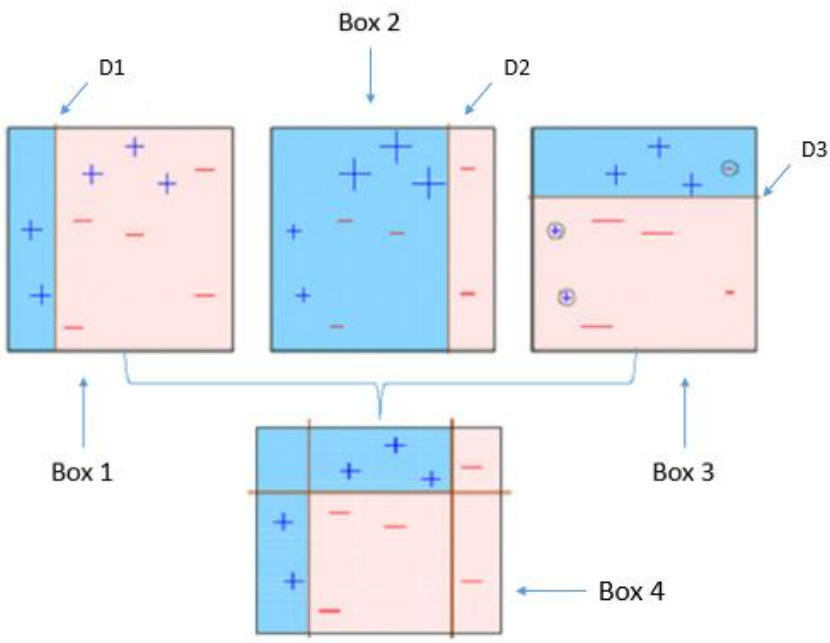

Figure 3. Combination of weak learners to form a strong learner in boosting

C. XGBoost Learning Model

XGBoost [7] is considered one of the most powerful and efficient implementations of the Gradient Boosted Trees algorithm to tackle all the tasks of supervised learning. XGBoost has proved to be a highly effective ML algorithm, extensively used in machine learning competitions and hackathons. XGBoost has high predictive power and is almost 10 times faster than the other gradient boosting techniques. It also includes a variety of regularization which reduces overfitting and improves overall performance. It is based on function approximation by optimizing specific loss functions as well as applying several regularization techniques. Before discussing what are loss functions and

regularization techniques, let us define some notations and settings. By $x_i \in \mathbb{R}^d$, we denote the i-th instance with an associated label, e.g. in case of classification, or real value, e.g. in case of regression. We denote $\hat{y}$ as the prediction given x_i. Assume we have *K* trees, we define:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \ , \ f_k \in \mathcal{F} \ . \tag{2}$$

Machine learning is basically the procedure of learning parameters $\Theta = \{w_j \mid \mathrm{j} = 1, \dots, \mathrm{d}\}$ from data. The overall objective function is follows:

$$Obj(\Theta) = L(\Theta) + \Omega(\Theta) \ , \tag{3}$$

where $L(\Theta)$ is the training loss measuring how well a model fit on the training data. $\Omega(\Theta)$ is the regularization configuration measuring the model's complexity. While optimizing the training loss part inspires the prediction accuracy of the model, optimizing the regularization seeks for a simple model. Expanding Equation (3), we define the XGBoost objective function at iteration t that we need to optimize is as following:

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \ . \tag{4}$$

From Equation (2) we can see that we cannot optimize Equation (4) by using traditional optimization methods in Euclidean space because XGBoost objective is a function of functions. In order to be able to use traditional optimization techniques, we need to transform the original objective function to a function in the Euclidean domain. One solution is to use Taylor approximation [8]. By taking Taylor expansion of the objective, we define Recall as follows:

$$f(x + \Delta x) \simeq f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2 \ . \tag{5}$$

We define the first and second order gradient statistics of the loss function as follows:

$$g_i = \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}) \ , \ h_i = \partial^2_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1}) \tag{6}$$

Then Equation (4) can be written as follows:

$$Obj^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{7}$$

The above is a sum of simple quadratic functions of one variable and can be minimized by using known techniques. More mathematical details can be found in [7].

III. EXPERIMENTS

A. Dataset

A retail company "ABC Private Limited" wants to understand the customer purchase behavior (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for a selected high volume products from last month. They want to build a model to predict the purchase amount of customer against various products which will help them to create a personalized offer for customers against different products. The data, as well as discussion, are publicly available on Kaggle [9]. The dataset also comes from a current competition hosted by Analytics Vidhya [10]. More statistical details of Black Friday dataset can be found in Table (1) and Figure (4).

Table 1. The Black Friday data made available by "ABC Private Limited"

Variable	**Definition**	**Value Examples**
User_ID	User ID	1000001, 1000032
Product_ID	Product ID	P00332342, P00032842
Gender	Sex of user	F,M
Age	Age in bins	0-17, 26-35, 55+
Occupation	Occupation (masked)	1,9,20
City_Category	Category of the user's city	A,B,C

Variable	Definition	Value Examples
Stay_In_Current_City_Years	Duration in years a user stays in current city	2,4+
Marital_Status	Marital status	0,1
Product_Category_1	Product category (masked)	2,8,15
Product_Category_2	Product category (masked)	11,15,16
Product_Category_3	Product category (masked)	14,15,16
Purchase	Purchase amount (target variable)	1057,19215

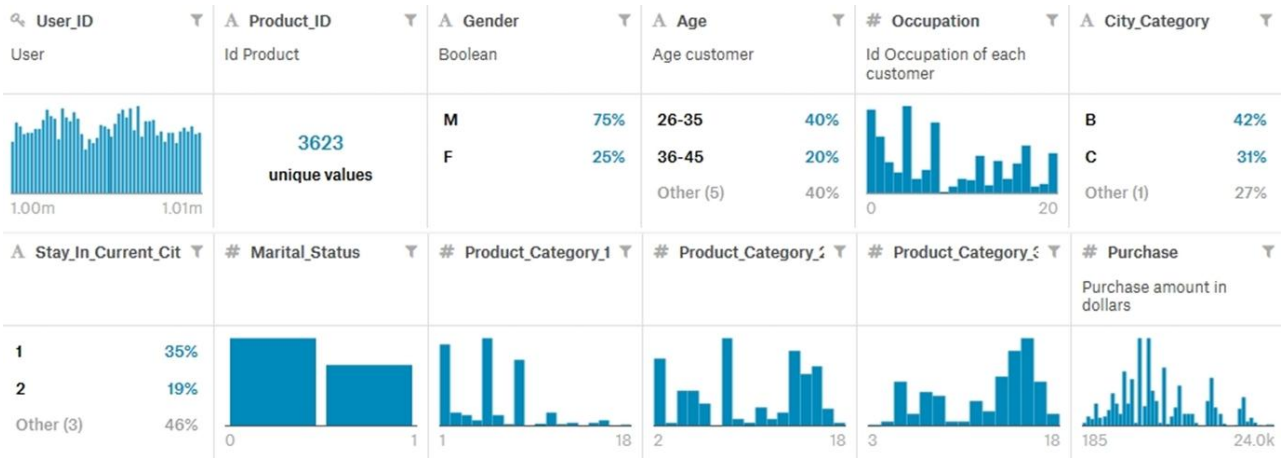

Figure 4. Statistics on the Black Friday dataset

B. Data Splitting Schemes

Machine learning models have the fundamental goal of making accurate predictions on unseen instances beyond those appeared in the training set. To estimate the quality of models' predictions with data it has not seen, we can split a portion of the data for which we already know the answer as a proxy for the unseen data. Then we evaluate how well the model predicts for that data. Typically, training dataset contains observations used to fit a learning model and tune hyperparameters. Test dataset includes samples of data used to provide an unbiased evaluation of the final learning model fit on the training dataset. The authors randomly shuffle the data into training and test sets without replacement in every experiment. In our experiment, we set up three different dataset splitting schemes by tuning various split ratios. We divide the obtaining data into three different splitting schemes which the readers can see a summation in Table (2).

1. Dataset splitting scheme 1: The proportion of training and test sets are 70% and 30% respectively. We denote it as 70|30 hereafter.

2. Dataset splitting scheme 3: The proportion of training and test sets are 90% and 10% respectively. We denote it as 80|20 hereafter.

3. Dataset splitting scheme 3: The proportion of training and test sets are 90% and 10% respectively. We denote it as 90|10 hereafter.

Table 2. Summary of several splitting schemes

Black Friday data	Splitting schemes		
	70 | 30	80 | 20	90 | 10
A total of 537577 instances	376304 | 161273	430062 | 107515	483819 | 53758

C. Evaluation Metrics

Models' performance will be evaluated on the basis of prediction of the purchase amount. The performance measure is root mean square error (RMSE) as defined below:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}\ . \tag{8}$$

D. Implementation and Results

1. Scenario 1: Experiment results on dataset splitting scheme 1

In this scenario, we evaluated models on dataset splitting scheme 1. We randomly shuffle dataset without replacement 3 times and execute. Then we take an average in the end. The best RMSE score is 2646 $\pm$ 0.00 and 2607 $\pm$ 0.00 on the test and training respectively. The results of this scenario is presented in Table (3 and 4), column 70 | 30 specifically.

2. Scenario 2: Experiment results on dataset splitting scheme 2

Similar to scenario 1, we randomly shuffle dataset without replacement 3 times, execute four models and take an average of classification accuracy in the end. The best RMSE score is 2647 $\pm$ 0.00 on the test data while RMSE score is 2605 $\pm$ 0.00 on the training data. The results of this scenario is presented in Table (3 and 4), column 80 | 20 specifically.

3. Scenario 3: Experiment results on dataset splitting scheme 3

The third scenario is investigated to examine the models' performance in the case of the highest proportion of training set, e.g. 90%. We apply a similar experiment configuration as the previous two scenarios. The best RMSE score is 2602 $\pm$ 0.00 and 2600 $\pm$ 0.00 which evidently confirms the effectiveness of extreme gradient boosting. The results of this scenario is presented in Table (3 and 4), column 90 | 10 specifically.

Table 3. The RMSE score of evaluated models on the training sets. The best score in each column is in **bold**

Models	RMSE score (in US dollar)		
	70 | 30	80 | 20	90 | 10
SVM	4569 ± 0.10	4566 ± 0.03	4564 ± 0.15
Random forest	2902 ± 0.23	2901 ± 0.14	2899 ± 0.20
Gradient Boosting	2616 ± 0.00	2614 ± 0.11	2611 ± 0.00
XGBoost	**2607 ± 0.00**	**2605 ± 0.00**	**2600 ± 0.00**

Table 4. The RMSE score of evaluated models on the training sets. The best score in each column is in **bold**

Models	RMSE score (in US dollar)		
	70 | 30	80 | 20	90 | 10
SVM	4611 ± 0.41	4610 ± 0.06	4585 ± 0.22
Random forest	2911 ± 0.58	2910 ± 0.18	2910 ± 0.40
Gradient Boosting	2673 ± 0.05	2668 ± 0.01	2662 ± 0.08
XGBoost	**2646 ± 0.00**	**2647 ± 0.00**	**2602 ± 0.00**

E. Reproducibility

In order to encourage readers participating in the Black Friday challenge, we would like to provide our grid search strategy on tuning models' hyperparameters. So interesting readers can reproduce our experimental results and brainstorm from there. For all tree-based models, we investigate the two most important settings. e.g. the number of estimators and the allowable depth of the trees. In the case of SVM, we take into account the effect of C and gamma configuration. Other hyperparameters leave default settings by scikit-learn library[1] [11]. The coding environment is as follows: windows 10 64 bit, Anaconda Python 3.6 ecosystems[2], scikit-learn machine learning library. Table (5) presents our hyperparameters settings.

Table 5. Models' grid search for the best hyperparameters configuration. The best settings are in **bold**

Models	Hyperparameters	Settings
Random forest	# estimators	1, 3, 10, 30, 100, 150, **300**, 450
	Max depth	1, 3, 5, 7, **9**, 11
Gradient boosting	# estimators	1, 3, 10, 30, 100, **150**, 300, 450
	Max depth	1, 3, 5, **7**, 9, 11
XGBoost	# estimators	5, **6**, 7
	Max depth	400, **500**, 600
SVM	C	0.001, **0.01**, 0.1, 1.0
	gamma	50, **60**, 70, 80

[1] https://scikit-learn.org/stable/
[2] https://www.anaconda.com/distribution/

IV. REMARKS AND DISCUSSION

At the current time that our paper is written, there are a total of 1524 participants participating in the Black Friday challenge [10]. The current best RMSE score is 2405 while the least score is 15267, e.g. the achievement is set on data splitting scheme 1. The analysis has underlined the effectiveness and robustness of gradient boosting in general and XGBoost in specific in a real-world regression problem. The Black Friday dataset is considered interestingly challenging with four over eleven attributes are masked, e.g. occupation, product category 1, product category 2, and product category 3. Moreover, noise in data is considerable that makes the challenge is hard to effectively solved. For example, 69.4% and 31.1% of product category 3 and product category 2 are missing respectively. This existence outweighs the performance of SVM as finding strong support vectors is difficult. And from the nature of SVM, it encodes sparsity in the loss function rather than the prior. The advantages of SVM become its disadvantages when data's noise is extremely high. The experimental results in Table (3 and 4) have proved this point.

Nevertheless, ensemble-learning-based models perform well in this kind of data. Taking the performance of random forest as an example, .e.g. the bagging-based model, we can agree that by splitting the dataset into different chunks and executing models on them have provided better results. A huge improvement has been made from around RMSE of 4611 by SVM down to that of 2911 by random forest in the test set, e.g. the 70 | 30 splitting scheme. Similar improvements have been witnessed across all other experimental results. However, if we take the average of multiple weak learners in bagging, we would result in a weak combined model. That is where the boosting idea comes into place where each subsequent learner aims to reduce the errors of the previous learners. The approximate RMSE of 300 has been reduced in every experimental result when the gradient boosting model takes place on the same dataset. Overall, the results of our study confirm the ideas ensemble learning on an interesting challenging problem.

V. CONCLUSION

The key purpose of this study is to investigate the performance of gradient boosting technique in extremely noise data. More specifically, the current results have confirmed the effective strategy of applying extreme gradient boosted trees to predict the amount of purpose. The Black Friday challenge is still operating, so much further consideration can be made to improve the RMSE score. The challenge's leaderboard shows that the RMSE difference of our score to that of the first position is approximate 240, e.g. 2646 versus 2405, which strongly indicates more advanced model's tuning and feature engineering.

REFERENCES

[1] J. McDermott, "Black Friday spending statistics 2018," Apr 2019. [Online]. Available: https://www.finder.com/black-friday-statistics

[2] E. Swilley and R. E. Goldsmith, "Black Friday and cyber Monday: Understanding consumer intentions on two major shopping days," Journal of retailing and consumer services, vol. 20, no. 1, pp. 43-50, 2013.

[3] J. Boyd Thomas and C. Peters, "An exploratory investigation of black Friday consumption rituals," International Journal of Retail & Distribution Management, vol. 39, no. 7, pp. 522-537, 2011.

[4] L. Simpson, L. Taylor, K. O'Rourke, and K. Shaw, "An analysis of consumer behavior on black Friday," American International Journal of Contemporary Research, 2011.

[5] G. C. Bell, M. R. Weathers, S. O. Hastings, and E. B. Peterson, "Investigating the celebration of black Friday as a communication ritual," Journal of Creative Communications, vol. 9, no. 3, pp. 235-251, 2014.

[6] H. J. Kwon and T. M. Brinthaupt, "The motives, characteristics and experiences of us black Friday shoppers," Journal of Global Fashion Marketing, vol. 6, no. 4, pp. 292-302, 2015.

[7] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, pp. 785-794, 2016.

[8] A. Guzman, Derivatives and integrals of multivariable functions. Springer Science & Business Media, 2012.

[9] M. Dagdoug, "Black Friday: A study of sales trough consumer behaviors," Jul 2018. [Online]. Available: https://www.kaggle.com/mehdidag/black-friday

[10] "Practice problem: Black Friday sales prediction | knowledge and learning," Jul 2016. [Online]. Available: https://datahack.analyticsvidhya.com/contest/black-friday/

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

DỰ ĐOÁN KINH DOANH NGÀY BLACK FRIDAY SỬ DỤNG PHƯƠNG PHÁP HỌC CÂY TĂNG CƯỜNG ĐỘ DỐC LỚN

Nghia Duong Trung, Tan Dang Thien, Tien Dao Luu, Hiep Xuan Huynh

***TÓM TẮT**: Ngày mua sắm lớn nhất trong năm ở Mỹ là thứ Sáu sau ngày lễ Tạ ơn. Nó được công nhận là sự bắt đầu của một trong những mùa mua sắm bận rộn nhất trong một năm. Từ quan điểm khoa học máy tính, một trong những ứng dụng thú vị nhất của máy học trong ngành bán lẻ là dự đoán hiệu quả số tiền khách hàng có thể chi tiêu tại một cửa hàng dựa trên lịch sử mua hàng. Nếu các nhà bán lẻ hiểu toàn diện khách hàng của họ về các đặc điểm, hành vi và động lực trong các mùa mua sắm trước đó, họ có thể thực hiện và phát triển các chiến lược tiếp thị hiệu quả hơn cho các danh mục khách hàng cụ thể. Nghiên cứu này đề xuất một triển khai thực nghiệm của thuật toán cây tăng cường độ dốc lớn để giải quyết bài toán trong ngành bán lẻ. Từ các kết quả thử nghiệm, các tác giả có thể kết luận rằng các ứng dụng của kỹ thuật cây tăng cường có thể đạt được hiệu suất cao và được cải thiện hơn nữa bằng sự kết hợp đúng đắn giữa điều chỉnh siêu tham số của mô hình và kỹ thuật tính năng.*

***Từ khóa:** Học toàn bộ, đóng gói và tăng cường, XGBoost, Dự đoán bán hàng, Black Friday.*