Comparative Analysis of Logistic Regression, Random Forest, and XGBoost for Click-Through Rate Prediction in Digital Advertising

Jiacheng Lou

Nanjing University of Information Science and Technology, Nanjing, Jiangsu Province, China
em804562@student.reading.ac.uk

Abstract. In the current research landscape, there remain gaps in predicting Click-Through Rates (CTR) for online advertisements. The study addresses these shortcomings by scrutinising three prediction models: logistic regression (LR), random forests (RF), and extreme gradient boosting (XGBoost). This methodology involves the same preprocessing of the data for all models. The performance metrics reveal that XGBoost shows the highest accuracy. For instance, XGBoost achieved a notable accuracy percentage of 94.10%, with RF and LR at 93.52% and 93.23% respectively. XGBoost also recorded the highest area under the curve, indicating its proficiency in distinguishing clicks from non-clicks. The study goes beyond mere numbers by delving into the strengths and weaknesses of each model. While LR is prized for its simplicity and interpretability, RF is valued for its robustness and accuracy over a range of data. However, XGBoost excels at handling complex data structures more efficiently. This study provides a theoretical basis for strengthening digital marketing strategies. It can guide advertisers and platform managers to optimize marketing activities. For example, it helps develop more sophisticated prediction tools for online advertising.

Keywords: Click-Through Rates, Prediction models, XGBoost.

1 Introduction

The Click-Through rate (CTR) is an essential standard to assess the efficacy of online advertisements, including the degree of user engagement. Ads with relatively low CTR can only be eliminated. Today's famous Internet companies, such as Google and Huawei, attach great importance to CTR.

CTR is critical across social media, search engines, and websites. Effective advertising is now key to attracting potential customers and increasing brand awareness. Or it's one of the best ways to make money. Digital advertising now occupies more than 70% of the global ad market. In 2020, U.S. online ad revenue grew by 12.2%, about $139.8 billion. There is a projection suggesting it will reach $982.82 billion by 2025 [1]. The origin of CTR prediction traced back to the birth of web advertising [2]. Click-through rate has not only become a point of focus for those involved in online

B. Siuta-Tokarska et al. (eds.), *Proceedings of the 2024 2nd International Conference on Management Innovation and Economy Development (MIED 2024)*, Advances in Economics, Business and Management Research 300,
https://doi.org/10.2991/978-94-6463-542-3_54

advertising, search engine optimization, and managing sponsored searches but also a financial instrument for budgeting and forecasting revenues [3].

Initially, CTR prediction methods were mainly based on basic statistical models and heuristic rules. Nowadays, the emergence of deep learning methods has offered new perspectives and tools for CTR prediction. The ability of deep learning to handle large datasets can fit straight lines and even higher-order curves. Compared with the traditional method, the prediction accuracy is improved. Machine learning is at the intersection of computer science and statistics, as well as artificial intelligence and data science. With the development of new algorithms and theories, machine learning has developed rapidly. The wide range of applications of data science machine learning methods can be found in various industries such as healthcare, manufacturing, education, and even stocks. A primary benefit of machine learning lies in its capability to process data autonomously once it is trained [4]. It's like a gear that never stops.

CTR prediction has advanced yet still confronts obstacles. Machine learning offers promise for this task, but selecting the appropriate model, handling data, and crafting features present complexities. These steps significantly influence accuracy and demand thorough experimentation [5]. A key issue is overfitting, where a model fits too precisely to training data, impairing its performance on new data. Causes of overfitting include noisy data, limited training data, and intricate classifiers [6]. Additionally, the variable nature of user behavior and digital contexts necessitates continual updates to models for precise CTR forecasts.

In addressing these challenges, this research investigates deep learning for CTR prediction. It processes data uniformly and then evaluates the precision of Logistic Regression (LR), Random Forest (RF), and XGBoost models. The study identifies challenges in current CTR prediction, contrasts these models, and explores feature management to bolster accuracy following standardized data preprocessing. Its goal is to pioneer novel and efficient models and strategies within digital marketing, aiming to augment the efficacy and impact of advertising efforts.

2 Method

2.1 LR

Although LR is referred to as a regression task, it is used for binary classification. LR is widely used in industry due to its simplicity, parallelism, and interpretability [7]. The fundamental principle of LR involves presuming a specific distribution for the data and employing maximum likelihood estimation for parameter determination. For example, the decision boundary shown in Fig. 1 can be expressed as:

$$w_1 x_1 + w_2 x_2 + b = 1 \tag{1}$$

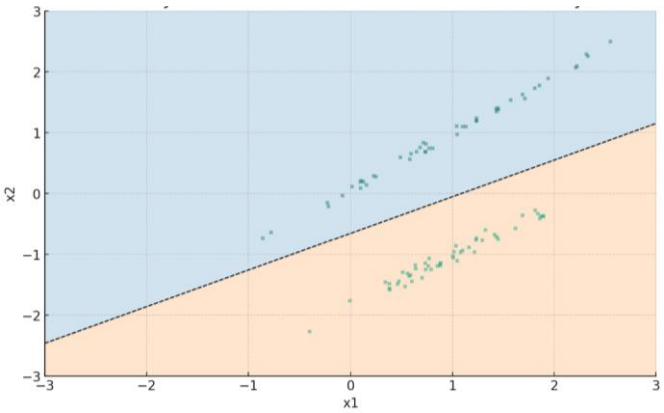

Fig. 1. Binary Classification with a Linear Decision Boundary (Picture credit: Original).

2.2 RF

RF is an ensemble learning classifier. It consists of multiple decision trees, each a simple model. While a single decision tree may struggle with complex problems, its aggregation forms a more powerful model. Fig. 2 illustrates how each tree arrives at its own decision; these are then combined to yield the final output. This collective decision-making process typically results in greater performance than any single tree could achieve alone [8].

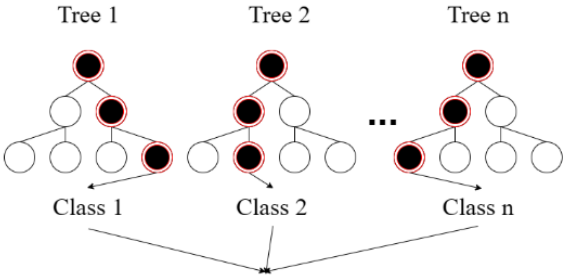

Fig. 2. Ensemble learning (Picture credit: Original)

2.3 XGBoost

Gradient Boosting Decision Trees (GBDT) are machine learning algorithms that improve prediction accuracy and stability by building and integrating various models, each aiming to rectify the errors made by its predecessor. XGBoost is based on the principles of GBDT, but it is heavily optimized internally for performance, speed, and functionality using more advanced algorithms. At the same time, it uses more advanced algorithms to control the model complexity. Moreover, its computation capability is optimized, such as speeding up the tree-building process through parallel processing [9].

3 Data

3.1 Dataset Introduction

Data sets from Kaggle and the dataset contain 23 key features and the objective feature 'click'. Each feature captures a different aspect of the ad environment or user behavior, which can influence the likelihood of a click. Other non-anonymous variables are explained in Table 1.

Table 1. Feature name inference

Feature	Description	Feature	Description
site_id	the site that is displaying the AD	banner_pos	location of the AD on the web page
site_domain	domain name of the website	device_model	model of the user device
site_ category	the category of the website, such as entertainment, news, etc.	device_type	type of device, such as a phone, tablet, or desktop
app_id:	AD is displayed in a mobile app	device_ conn_type	the connection type of the device, such as Wi-Fi, 4G, and so on
app_domain	domain name of the mobile application	device_ip	IP address of the user device
app_ category	the category of the mobile app	device_id	unique identifier for the user's device
id	unique identifier for each ad	click	represents the binary outcome of an ad being clicked (1) or not clicked (0)
hour	For example, '14091123' corresponds to 23:00 on September 11, 2014 UTC.		

These different types of data give us various ways to examine what influences someone to click on an ad, which is the essence of CTR.

3.2 Data Preprocessing

First, the key indicators such as the mean, median, maximum, and minimum value of all features are calculated. For the anonymized features 'C15', 'C16', 'C19', and 'C21', the maximum values exhibit a sparse distribution and present significant outliers. So the outliers above the 95th percentile are adjusted to align with the 95th percentile.

Specifically, removing extreme scores can greatly improve accuracy and significantly reduce inference errors [10].

In the preprocessing phase, the dataset's attributes were bifurcated into numerical and categorical variables. Numerical features inherently carry quantitative information that is directly computable within machine learning algorithms. Categorical variables, conversely, encompass qualitative data that represent distinct categories or groups that the models do not intrinsically interpret as numbers. So, categorical features necessitate an additional step of encoding.

The Pearson correlation coefficient quantifies the extent of linear relationship between two variables, ranging from -1 to 1. A value of 1 indicates a flawless positive correlation, and -1 denotes a flawless negative correlation. It is one of the most commonly used methods to describe linear relationships between variables [11]. From Fig. 3, the correlation between features can be observed, with darker colors representing a higher correlation.

The visual analysis presented in Fig. 4 indicates that the 'month' feature is uniformly distributed across the observational period. This uniformity suggests a lack of variability about temporal factors, thereby indicating that the 'month' feature does not contribute discernible information regarding the likelihood of an ad click. This observation leads to the inference that 'month' may be an extraneous variable in the context of predictive modeling for click-through rates.

To improve the models used to predict whether someone will click on an ad, a new feature was created. Take two pieces of information that are related to each other and combine them into one. This helps the model understand the relationship better. If a piece of information didn't help predict clicks, it was removed to make the model simpler and more focused.

For the information in the data that falls into categories, a technique was used to turn them into numbers. This is because the models work best with numbers. For example, if there are categories like "banner pos" or "ip address", each category is assigned a number based on how often clicks happen in each category.

Also, all the different pieces of information used to predict clicks were adjusted to be within the same range, from 0 to 1. This makes sure the model treats all information fairly, improving its ability to predict clicks.

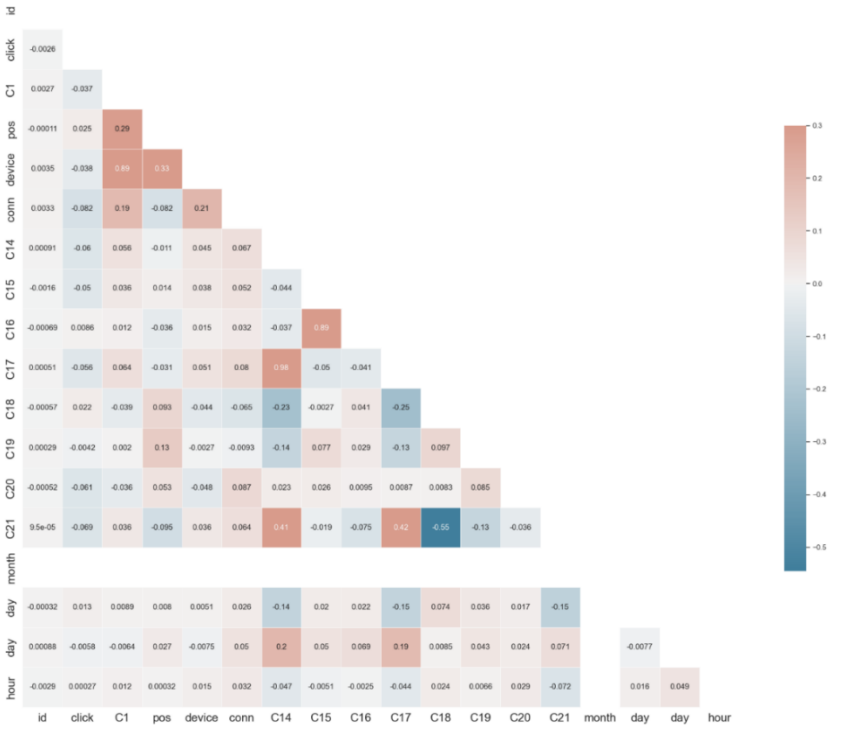

Fig. 3. Pearson's correlation coefficient (Picture credit: Original)

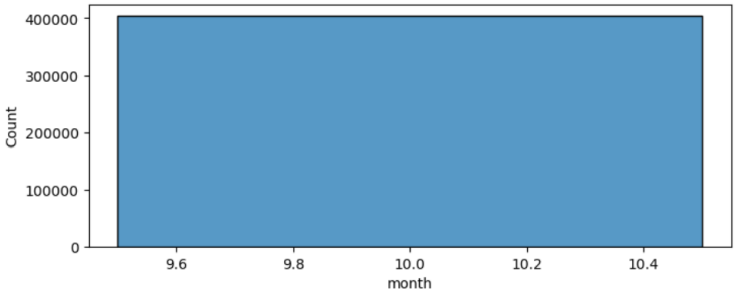

Fig. 4. Distribution of feature 'month' shows it is irrelevant to 'click' (Picture credit: Original)

3.3 Model Performance Evaluation and Validation

The dataset is split into the training set and testing set first. This split is even, with the same types of data in both parts.

Cross-validation, specifically employing a five-fold approach in this study, assesses the model's capacity to generalize to data it has not encountered before. The dataset is divided into five distinct, non-overlapping sections, with the model undergoing training and validation sequentially on each segment.

Next, the confusion matrix is computed using the model_eval function. The confusion matrix serves as a critical instrument for assessing a classification model's effectiveness. It offers an in-depth perspective on the outcomes of the model's predictions via four essential metrics: positive correct predictions, negative correct predictions, incorrect positive predictions, and incorrect negative predictions. Based on the confusion matrix, the recall and precision of the model are further calculated, which measure the ability of the model to identify positive examples and the accuracy of predicting positive examples, respectively. The Receiver Operating Characteristic (ROC) curves and accuracy metrics distinctly showcase each model's diagnostic capabilities.

4 Results and Discussion

This study looked at three models for predicting CTR: LR, RF, and XGBoost. The RF model was slightly more accurate than LR by about 0.3%, shown in Table 2. Yet, RF needs more computing power and is much slower on big datasets. It is more than ten times slower than LR. XGBoost did the best, with accuracy 1% higher than the others. Adding new interactive features and object encodings to the dataset greatly improved these models.

Also, cleaning data and making features consistent are key to better predictions. If skipping these steps, it would make the LR model's accuracy fall by 3%. Illustrated in Figure 5, the area under the ROC curve for the XGBoost model is also the largest.

Table 2. Model accuracy and confusion matrix

Model	LR	RF	XGB
Accuracy	0.9323	0.9352	0.9410
Confusion matrix	[98105 2595]	[98402 2298]	[99108 1592]
	[5610 14977]	[5551 15036]	[5558 15029]

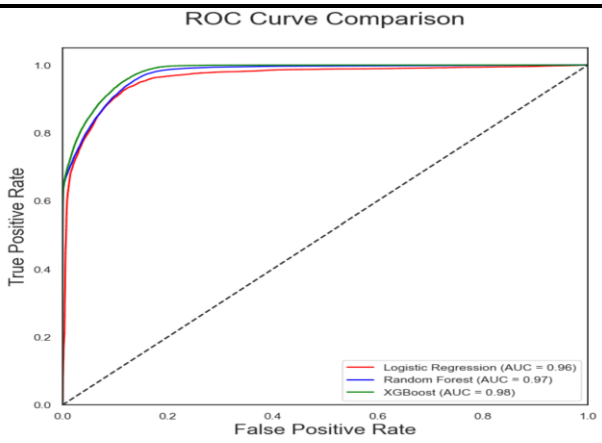

Fig. 5. ROC curves for the three models (Picture credit: Original)

5 Conclusion

The study's analysis of CTR prediction models provides insight into their differential performance. XGBoost emerges with the highest AUC value of 0.98, signifying its exceptional ability to discriminate between 'click' and 'no-click' instances. LR and RF exhibit commendable performance with AUCs of 0.96 and 0.97, respectively, though slightly trailing behind XGBoost.

Accuracy percentages align with the ROC findings, placing XGBoost at the forefront with a 94.10% accuracy rate, succeeded by RF at 93.52% and LR at 93.23%. Analyzing the confusion matrices reveals XGBoost's superior precision in minimizing false positives, a crucial factor in optimizing advertisement spending and targeting accuracy.

Delving into operational specifics, the RF's slower performance is attributed to its intrinsic method of constructing numerous decision trees, which involves aggregating the outcomes of various tree predictions to determine the final class. This process, although robust, is computationally demanding, leading to slower operational speeds. In contrast, XGBoost's speed and high accuracy are the results of its gradient boosting framework, which efficiently combines weak predictive models to strengthen the overall prediction and reduces error iteratively, contributing to its expedited processing time and enhanced predictive accuracy.

The impact of data preprocessing on LR's accuracy is significant due to the model's reliance on the assumption that predictors are linearly related to the log odds of the outcome. Effective preprocessing, such as feature scaling and handling of outliers, can align the data more closely with these assumptions, which is pivotal for LR's performance but less critical for tree-based methods like RF and XGBoost that are naturally robust to different data distributions.

In summary, this paper validates the efficacy of LR, RF, and XGBoost in CTR predictions, with XGBoost showing the most promise due to its sophisticated underlying mechanisms. The comparative analysis underscores the need for future enhancements in predictive modeling, emphasizing computational efficiency, model transparency, and practical application within the fast-evolving sphere of digital marketing. Future research will explore an extended range of features and innovative feature selection and engineering techniques to enhance model performance.

References

1. Yang Y., Zhai, P.: Click-through rate prediction in online advertising: A literature review, Information Processing and Management, 59, 102853 (2022).
2. Regelson, M., Fain, D.: Predicting click-through rate using keyword clusters. In Proceedings of the Second Workshop on Sponsored Search Auctions, Vol. 9623, pp. 1-6 (2006).
3. Zhou, G., Zhu, X., Song, C., et al.: Deep interest network for click-through rate prediction. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining pp. 1059-1068 (2018).
4. Kamal, M., Bablu, T. A.: Machine learning models for predicting click-through rates on social media: Factors and performance analysis. International Journal of Applied Machine Learning and Computational Intelligence, 12(4), 1-14 (2022).

5. Zhang, W., Qin, J., Guo, W., et al.: Deep learning for click-through rate estimation. arXiv preprint arXiv:2104.10584 (2021).
6. Ying, X.: An overview of overfitting and its solutions. In Journal of Physics: Conference series, Vol. 1168, pp. 022022, IOP Publishing (2019).
7. Boateng, E. Y., Abaye, D. A.: A review of the logistic regression model with emphasis on medical research. Journal of data analysis and information processing, 7(4), 190-207 (2019).
8. Ao, Y., Li, H., Zhu, L., et al.: The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. Journal of Petroleum Science and Engineering, 174, 776-789 (2019).
9. Ustuner, M., Sanli, F. B., Abdikan, S., et al.: A booster analysis of extreme gradient boosting for crop classification using PolSAR imagery. In 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), pp. 1-4, IEEE (2019).
10. Osborne, J. W., Overbay, A. The power of outliers (and why researchers should always check for them). Practical Assessment, Research, and Evaluation, 9(1), 6 (2019).
11. Karim, A., Azam, S., Shanmugam, B., et al.: A comprehensive survey for intelligent spam email detection. IEEE Access, 7, 168261-168295 (2019).