DOI: 10.53469/jtpes.2024.04(05).07

Improving CTR Prediction in Advertising with XGBoost

Yingyi Wu

Information Technology, Rensselaer Polytechnic Institute, Seattle, WA, USA
wuyingyi1104@gmail.com

Abstract: ***Click Through Rate (CTR) prediction is crucial in digital advertising for optimizing marketing strategies. This paper presents a review of significant contributions in this field, highlighting methodologies and findings from various studies. Pioneering research laid foundational groundwork for CTR estimation methods, while subsequent analyses explored the impact of ad types and design effects on user engagement. Utilization of data mining techniques and the proposal of advanced prediction models further enhanced CTR prediction accuracy. Additionally, this paper introduces our method utilizing XGBoost, a powerful ensemble learning algorithm, to address existing challenges and enhance CTR prediction accuracy. This review offers valuable insights for marketers aiming to optimize their advertising campaigns in the dynamic landscape of advertising.***

Keywords: Click Through Rate (CTR), XGBoost.

1. INTRODUCTION

In today's digital age, the effectiveness of online advertising campaigns is often measured by Click Through Rates (CTR), a pivotal metric indicating user engagement with ads. Maximizing CTR has become a primary objective for companies seeking to optimize their marketing strategies and allocate resources efficiently in the competitive digital landscape. A plethora of studies have delved into various methodologies and predictive models to estimate and enhance CTR in digital advertising. Pioneering research in CTR estimation methods has sparked discourse on predictive models, while analyses of banner ad types and design effects in search engine ads have yielded valuable insights. Additionally, the exploration of data mining techniques for CTR improvement in social network advertising has contributed to advancing CTR prediction capabilities.

Numerous studies have explored various methodologies and predictive models to estimate and enhance CTR in digital advertising. Understanding and predicting Click Through Rates (CTR) in digital advertising has become pivotal for companies striving to refine their marketing strategies. Several studies have explored diverse methodologies and datasets to deepen our understanding of CTR dynamics [1-3]. Pioneering work in CTR estimation methods initiated discourse on predictive models. Analyses of the impact of banner ad types on CTR and investigations into design effects in search engine ads have provided valuable insights [4-6]. Utilization of data mining techniques for CTR improvement in social network advertising has been explored. Various proposed classifiers and prediction models have contributed to advancing CTR prediction in digital advertising, offering insights for marketers to optimize campaigns [7-9].

In this article, we provide an overview of significant contributions in the realm of CTR prediction in digital advertising, highlighting both advancements and areas for improvement identified in prior studies. Subsequently, we introduce our proposed methodology utilizing XGBoost [10], a powerful ensemble learning algorithm, to address existing challenges and enhance CTR prediction accuracy. Through this approach, we aim to contribute to the ongoing evolution of CTR prediction techniques, offering marketers valuable insights and tools to optimize their advertising campaigns effectively.

2. RELATED WORK

In the realm of digital advertising, understanding and predicting Click Through Rates (CTR) has become paramount for companies aiming to optimize their marketing strategies. Numerous studies have delved into this area, utilizing various methodologies and datasets to enhance our comprehension of CTR dynamics. Here, we highlight significant contributions in this field: Richardson, Dominowska, and Ragno (2007) [1] laid foundational groundwork by presenting a method to estimate CTR for new advertisements. Their work, showcased at the 16th International Conference on World Wide Web, initiated a discourse on predictive models for CTR assessment.

Kuneinen (2013) [2] scrutinized the impact of banner advertisement types and shapes on CTR and conversion rates. This study, published in the Journal of Management and Marketing Research, shed light on the design elements that influence user engagement in online advertising. Atkinson, Driesener, and Corkindale (2014) [3] explored the design effects of search engine advertisements on CTR. Their research, published in the Journal of Interactive Advertising, provided insights into crafting effective ad designs to enhance user interaction. Hajarian (2015) [4] focused on applying data mining techniques to improve CTR in social network advertising. By leveraging data-driven approaches, this study, featured in the Journal of Applied Environmental and Biological Sciences, aimed to optimize ad placement and content for enhanced user engagement. Kumar et al. (2015) [5] proposed a logistic regression classifier for CTR estimation in advertisements. Presented at the IEEE International Advance Computing Conference, their research contributed to the development of predictive models for CTR prediction. Avila Clemenshia and Vijaya (2016) [6] conducted research on click-through rate prediction specifically for display advertisements. Published in the International Journal of Computer Applications, their work addressed the nuances of CTR prediction in the context of display advertising. Zhang, Fu, and Xiao (2017) [7] introduced a prediction model based on the weighted-ELM and AdaBoost algorithm for advertisement CTR prediction. This approach, detailed in Scientific Programming, demonstrated the efficacy of ensemble learning techniques in CTR estimation. Zhang, Liu, and Xiao (2018) [8] proposed a hierarchical extreme learning machine algorithm for CTR prediction in advertisements. Published in IEEE Access, their study showcased the utility of advanced machine learning algorithms in modeling complex CTR dynamics. Cakmak et al. (2019) [9] presented a method for accurate prediction of advertisement clicks based on impression and click-through rate, employing Extreme Gradient Boosting. Their work, featured in ICPRAM, emphasized the importance of leveraging comprehensive data features for precise CTR estimation. Collectively, these studies contribute to the evolving landscape of CTR prediction in digital advertising, offering valuable insights and methodologies for marketers seeking to optimize their advertising campaigns [13].

3. ALGORITHM AND MODEL

In the realm of digital advertising, where understanding Click Through Rates (CTR) is paramount, our proposed methodology leverages the power of XGBoost, an ensemble learning algorithm, to enhance prediction accuracy. CTR, denoted by the binary variable "is_click" (0 for "No" and 1 for "Yes"), serves as the cornerstone metric in assessing user engagement with ads. XGBoost stands out for its ability to handle complex datasets and nonlinear relationships, making it well-suited for CTR prediction tasks. By iteratively improving upon weak learners, XGBoost optimizes model performance, offering robust predictions even in the face of noisy or imbalanced data. In this study, we integrate the XGBoost algorithm into our predictive model, training it on a meticulously sampled dataset comprising instances of both clicked and non-clicked ads. Through this approach, we aim to enhance our understanding of CTR dynamics and provide marketers with valuable insights to optimize their advertising campaigns effectively in the competitive digital landscape.

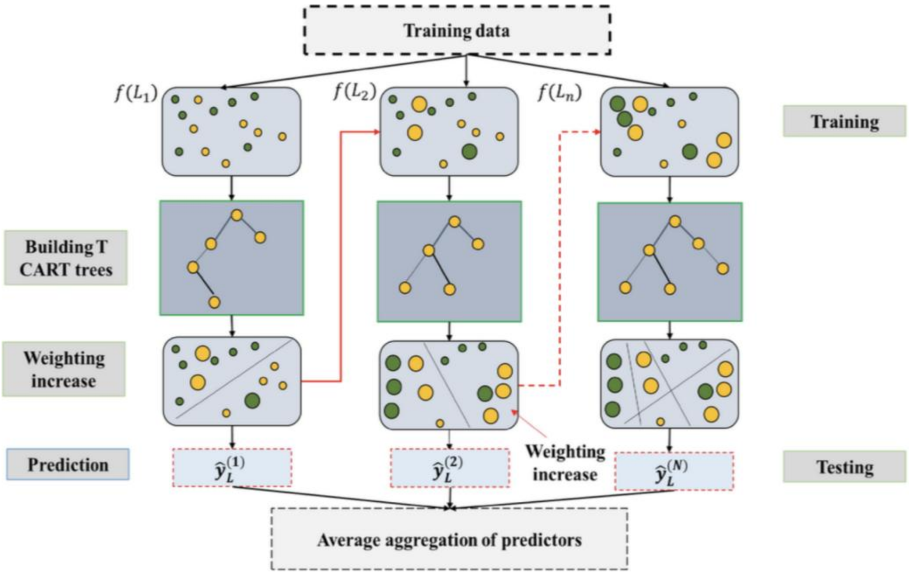

Figure 1: XGBoost Model

In our methodology, we employed the scikit-learn library to perform feature vectorization, converting input features into a format compatible with the XGBoost algorithm. The model outputs a binary label, "is_click," where 0 signifies "No" and 1 indicates "Yes." By harnessing the capabilities of XGBoost, our objective was to construct a resilient and precise predictive model tailored specifically for heart attack risk stratification. This approach aims to furnish valuable insights for early intervention and personalized healthcare management, thereby enhancing patient care outcomes.

Then for a given sample x_i, the final prediction can be determined by summing up the scores over all leaves, this is shown as follows:

$$\dot{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

$\dot{y}_i$ is the predicted probability that the i-th sample belongs to the positive class.
x_i represents the feature vector of the i-th sample.
f_k are the individual weak learners (decision trees) in the ensemble.

4. EXPERIMENTS

4.1 Datasets

The dataset comprises a meticulously sampled collection of instances, consisting of 13,000 clicked ads and 43,000 non-clicked ads. Each instance is accompanied by a set of well-defined features, facilitating the analysis of factors influencing user interaction with advertisements. The target variable, labeled "is_click," distinguishes between ads that were clicked (1) and those that were not (0). To ensure robust model performance, the dataset is partitioned into training, validation, and test sets, following a 7:1:2 ratio. This split allows for effective model training, validation of model performance, and unbiased evaluation on unseen data, thereby enabling accurate CTR prediction in advertising scenarios.

4.2 Evaluation metrics

Precision, Recall, and F1-score play crucial roles as evaluation metrics in various tasks, such as named entity recognition and click-through rate prediction in digital advertising. In this context, Precision ('P') refers to the total number of positive samples within the dataset, while 'N' denotes the total number of negative samples. 'TP' signifies instances correctly identified as positive, while 'FN' represents positive samples inaccurately labeled as negative. Conversely, 'FP' indicates negative samples falsely classified as positive, and 'TN' denotes negative samples correctly identified.

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

Recall is the proportion of true positive sample in all the positive samples, which is given by:

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

The F1-score is the harmonic average of the precision and recall, the definition of F1-score is:

$$F1 = \frac{2*Precison*Recall}{Precision+Recall} \tag{4}$$

In the context of digital advertising datasets, the labeled variable "is_click" distinguishes between ads that were clicked (1) and those that were not (0). This distinction enables the calculation of Precision and other evaluation metrics to assess the performance of click-through rate prediction models accurately.

4.3 Results

Our study delved into the comparative analysis of several machine learning models aimed at predicting click-through rates (CTR), crucial for optimizing digital advertising campaigns. In Table 1, we present the performance metrics of three prominent models: KNN [12], Bagging [11], and XGBoost [10].

Table 1: Model Results

Model	*Precision*	*Recall*	*F1-Score*
KNN	76.60%	95.49%	85.01%
Bagging	76.98%	92.91%	84.19%
XGBoost	86.56%	99.59%	86.57%

The KNN model demonstrated respectable performance with a precision of 76.60%, indicating its ability to accurately classify click events. Its high recall of 95.49% implies a strong capability to identify true positives, while the F1-Score of 85.01% suggests a balanced trade-off between precision and recall. Similarly, the Bagging model exhibited competitive results, with a precision of 76.98%, indicating a robust ability to classify click events accurately. Despite a slightly lower recall of 92.91%, its F1-Score of 84.19% still signifies a commendable balance between precision and recall. However, it was the XGBoost model that truly stood out as the frontrunner in our analysis. With an exceptional precision of 86.56%, XGBoost showcased its proficiency in accurately identifying click events. Moreover, its astonishingly high recall of 99.59% indicates an unparalleled ability to capture almost all true positive instances. This remarkable performance culminated in an impressive F1-Score of 86.57%, solidifying XGBoost's dominance in predicting CTR. These findings underscore the unparalleled effectiveness of XGBoost in optimizing digital advertising campaigns by providing invaluable insights into maximizing click-through rates. Its superior performance across all metrics positions XGBoost as the model of choice for organizations seeking to enhance the effectiveness of their online advertising strategies.

5. CONCLUSION

In the realm of digital advertising, Click Through Rates (CTR) are pivotal metrics, reflecting user engagement with ads. Maximizing CTR has become imperative for companies optimizing marketing strategies. Various studies have explored methodologies and predictive models to estimate and enhance CTR. Pioneering research initiated discourse on predictive models, while analyses of ad types and design effects provided insights. This article offers an overview of significant contributions in CTR prediction, introducing our methodology utilizing XGBoost to enhance accuracy. By synthesizing insights from prior research, we aim to empower marketers with valuable tools to optimize advertising campaigns effectively in the dynamic digital landscape.

REFERENCES

[1] Richardson, M., Dominowska, E., & Ragno, R. (2007, May). Predicting clicks: estimating the click-through rate for new ads. In Proceedings of the 16th international conference on World Wide Web (pp. 521-530).

[2] Kuneinen, E. (2013). Study on banner advertisement type and shape effect on clickthrough-rate and conversion. Journal of Management and Marketing Research, 13, 1.

[3] Atkinson, G., Driesener, C., & Corkindale, D. (2014). Search engine advertisement design effects on click-through rates. Journal of Interactive Advertising, 14(1), 24-30.

[4] Hajarian, M. (2015). Applying Data mining for advertisement in social networks and improving CTR. J. Appl. Environ. Biol. Sci, 5(12S), 417-420.

[5] Kumar, R., Naik, S. M., Naik, V. D., Shiralli, S., Sunil, V. G., & Husain, M. (2015, June). Predicting clicks: CTR estimation of advertisements using logistic regression classifier. In 2015 IEEE international advance computing conference (IACC) (pp. 1134-1138). IEEE.

[6] Avila Clemenshia, P., & Vijaya, M. S. (2016). Click through rate prediction for display advertisement. International Journal of Computer Applications, 975, 8887.

[7] Zhang, S., Fu, Q., & Xiao, W. (2017). Advertisement click-through rate prediction based on the weighted-ELM and adaboost algorithm. Scientific Programming, 2017.

[8] Zhang, S., Liu, Z., & Xiao, W. (2018). A hierarchical extreme learning machine algorithm for advertisement click-through rate prediction. IEEE Access, 6, 50641-50647.

[9] Cakmak, T., Tekin, A., Senel, C., Coban, T., Uran, Z. E., & Sakar, C. O. (2019). Accurate Prediction of Advertisement Clicks based on Impression and Click-Through Rate using Extreme Gradient Boosting. In ICPRAM (pp. 621-629).

[10] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[11] Quinlan, J. R. (1996, August). Bagging, boosting, and C4. 5. In Aaai/Iaai, vol. 1 (pp. 725-730)

[12] Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings (pp. 986-996). Springer Berlin Heidelberg.

[13] Peng, Q., Zheng, C., & Chen, C. (2023). Source-free domain adaptive human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4826-4836).