# A Review on Data Mining Models for Click-Through-Rate Prediction

**Umashree Patidar[1], Prof. Pankaj Raghuwanshi[2]**

***Abstract: During Online advertising is a multi-billion dollar industry that has served as one of the great success stories for machine earning. Sponsored search advertising, contextual advertising, display advertising, and real-time bidding auctions have all relied heavily on the ability of learned models to predict ad click–through rates accurately, quickly, and reliably. Predicting ad click–through rates (CTR) is a massive-scale learning problem that is central to the multi -billion dollar online advertising industry. Search engine advertising has become a significant element of the web browsing experience. Choosing the right ads for a query and the order in which they are displayed greatly affects the probability that a user will see and click on each ad. Accurately estimating the click-through rate (CTR) of ads has a vital impact on the revenue of search businesses; even a 0.1% accuracy improvement in production would yield hundreds of millions of dollars in additional earnings. An ad's CTR is usually modelled as a classification problem, and thus can be estimated by machine learning models. The training data is collected from historical ads impressions and the corresponding clicks. A comprehensive review is presented in the paper pertaining to the supervised learning architecture for the prediction model.***

***Keywords: Online Advertising, Click Through Rates (CTR), Sponsored Search Advertising, Real Time Bidding, Supervised Learning.***

## I. INTRODUCTION

The Online advertising is one of the most effective ways for businesses of all sizes to expand their reach, find new customers, and diversify their revenue streams [1].

With so many options available – from PPC and paid social to online display advertising and in-app ads – online advertising can be intimidating to newcomers, but it doesn't have to be. Online advertising, also called online marketing or Internet advertising or web advertising is a form of marketing and advertising which uses the Internet to deliver promotional marketing messages consumers [2]. Consumers view online advertising as an unwanted distraction with few benefits and have increasingly turned to ad blocking for a variety of reasons. When software is used to do the purchasing, it is known as programmatic advertising [3].

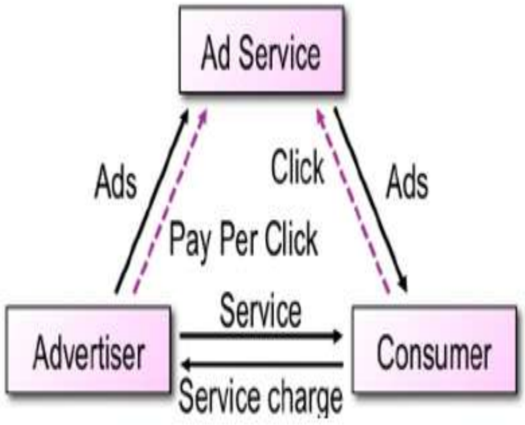

**Fig.1 The pay per click model**

Display Advertising conveys its advertising message visually using text, logos, animations, videos, photographs, or other graphics. Display advertisers frequently target users with particular traits to increase the ads' effect. Online advertisers (typically through their ad servers) often use cookies, which are unique identifiers of specific computers, to decide which ads to serve to a particular consumer. Cookies can track whether a user left a page without buying anything, so the advertiser can later retarget the user with ads from the site the user visited.

As advertisers collect data across multiple external websites about a user's online activity, they can create a detailed profile of the user's interests to deliver even more targeted advertising. This aggregation of data is called behavioral targeting. Advertisers can also target their audience by using contextual to deliver display ads related to the content of the web page where the ads appear [4].

Retargeting, behavioral targeting, and contextual advertising all are designed to increase an advertiser's return on investment, or ROI, over untargeted ads. The click probability is thus a key factor used to rank the ads in appropriate order, place the ads in different locations on the page, and even to determine the price

that will be charged to the advertiser if a click occurs. Therefore, ad click prediction is a core component of the sponsored search system.
Computation-heavy tasks to nearby more capable UEs using links. Considerable research has gone into design of offloading technique [5]

## II. PREVIOUS WORK

This section presents the previous work in the domain.

**Zhang et al. [6]** proposed a Target Attention (TA) to Target Pattern Attention (TPA) to model pattern-level dependencies. Furthermore, three critical challenges demand attention: the inclusion of unrelated items within patterns, data sparsity of patterns, and computational complexity arising from numerous patterns. To address these challenges, we introduce the Deep Pattern Network (DPN), designed to comprehensively leverage information from behavior patterns.

**Xiao et al. [7]** proposed a novel method named as Deep Multi-Interest Network (DMIN) which models user's latent multiple interests for click-through rate prediction task. Specifically, authors have designed a Behavior Refiner Layer using multi-head self-attention to capture better user historical item representations. Then the Multi-Interest Extractor Layer is applied to extract multiple user interests. The paper evaluates the method on three real-world datasets. Experimental results show that the proposed DMIN outperforms various state-of-the-art baselines in terms of click-through rate prediction task.

**Craswell et al. [8]** proposed a "generalized second-price" (GSP) auction, a new mechanism used by search engines to sell online advertising. Although GSP looks similar to the Vickrey-Clarke-Groves (VCG) mechanism, its properties are very different. Unlike the VCG mechanism, GSP generally does not have an equilibrium in dominant strategies, and truth-telling is not an equilibrium of GSP. To analyze the properties of GSP, authors describe the generalized English auction that corresponds to GSP and show that it has a unique equilibrium. This is an ex post equilibrium, with the same payoffs to all players as the dominant strategy equilibrium of VCG.

**Fain el al. [9]** proposed empirical analysis of search advertising strategies. IMC, Accurate estimation of the click-through rate (CTR) in sponsored ads significantly impacts the user search experience and businesses' revenue, even 0.1% of accuracy improvement would yield greater earnings in the hundreds of millions of dollars. CTR prediction is generally formulated as a supervised classification problem. In this paper, authors represent the experience and learning on model ensemble design and our innovation. Specifically, authors present 8 ensemble methods and evaluate them on our production data. Boosting neural networks with gradient boosting decision trees turns out to be the best. With larger training data, there is a nearly 0.9% AUC improvement in offline testing and significant click yield gains in online traffic. In addition, authors share our experience and learning on improving the quality of training.

**Galen et al. [10]** proposed that click through and conversation rates estimation are two core predictions tasks in display advertising. Authors present in this paper a machine learning framework based on logistic regression that is specifically designed to tackle the specifics of display advertising. The resulting system has the following characteristics: it is easy to implement and deploy; it is highly scalable (authorshave trained it on terabytes of data); and it provides models with state-of-the-art accuracy.

**Graepel et al. [11]** proposed that tree boosting is a highly effective and widely used machine learning method. In this paper, authors describe a scalable end-to-end tree boosting system called XGBoost, which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges. Authors propose a novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning. More importantly, authors provide insights on cache access patterns, data compression and build a scalable tree boosting system. By combining these insights, XGBoost scales beyond billions of examples using far fewer resources than existing systems.

**Graepel et al. [12]** proposed a probabilistic latent variable model, namely GaP (Gamma-Poisson), to ad targeting in the contexts of sponsored search (SS) and behaviorally targeted (BT) display advertising. Authorsalso approach the important problem of ad positional bias by formulating a one-latent-dimension GaP factorization. Learning from click-through data is intrinsically large scale, even more so for ads. Authors scale up the algorithm to terabytes of real-world SS and BT data that contains hundreds of millions of users and hundreds of thousands of features, by leveraging the scalability characteristics of the algorithm and the inherent structure of the problem including data sparsity and locality. Specifically, authors demonstrate two somewhat orthogonal philosophies of scaling algorithms to large-scale problems, through the SS and BT implementations, respectively. Finally, authors report the experimental results using Yahoo's vast datasets, and show that our approach substantially outperform the state-of-the-art methods in prediction accuracy. For BT in particular, the ROC area achieved by GaP is exceeding 0.95, while one prior approach using Poisson regression yielded 0.83. For computational performance, authors compare a single-node sparse implementation with a parallel implementation using Hadoop MapReduce, the results are counterintuitive yet quite interesting.

**He et al. [13]** showed that sponsored search is a multi-billion dollar business that generates most of the revenue for search engines. Predicting the probability that users click on ads is crucial to sponsored search because the prediction is used to influence ranking, filtering, placement, and pricing of ads. Ad ranking, filtering and placement have a direct impact on the user experience, as users expect the most useful ads to rank high and be placed in a prominent position on the page. Pricing impacts the advertisers' return on their investment and revenue for the search engine. The objective of this paper is to present a framework for the personalization of click models in sponsored search. Authors develop user-specific and demographic-based features that reflect the click behavior of individuals and groups. The features are based on observations of search and click behaviors of a large number of users of a commercial search engine. Authors add these features to a baseline non-personalized click model and perform experiments on offline test sets derived from user logs as well as on live traffic. Our results demonstrate that the personalized models significantly improve the accuracy of click prediction.

**Juan et al. [14]** proposed that the success of many computer games depends on designing a robust and adaptable AI opponent that would ensure the games continue to challenge, immerse and excite the players at any stage. The outcomes of card based games like ``Heartstone: Heros of Warcraft", aside the player skills heavily depend on the initial composition of player card decks. To evaluate this impact authors have developed an ensemble prediction model that tries to predict the average win-rates of the specific combination of bot-player and card decks. Our ensemble model consists of three sub- models: two Logistic Regression models and one Deep Learning model. The models are trained with both provided data and additional data about the cards, their health, attack power and cost. To avoid over fitting, authors employ a trick to generate predictions for all possible combinations of opponent players and decks and obtain the result as the average of all these predictions.

**Juan et al. [15]** give a detailed statistical analysis of the relationship between the AUC and the error rate, including the first exact expression of the expected value and the variance of the AUC for a fixed error rate. Our results show that the average AUC is monotonically increasing as a function of the classification accuracy, but that the standard deviation for uneven distributions and higher error rates is noticeable. Thus, algorithms designed to minimize the error rate may not lead to the best possible AUC values. Authors show that, under certain conditions, the global function optimized by the Rank Boost algorithm is exactly the AUC. Authors report the results of our experiments with Rank Boost in several datasets demonstrating the benefits of an algorithm specifically designed to globally optimize the AUC over other existing algorithms optimize.

**Maas et al. [16]** proposed engine click logs provide an invaluable source of relevance information, but this information is biased. A key source of bias is presentation order: the probability of click is influenced by a document's position in the results page. This paper focuses on explaining that bias, modeling how probability of click depends on position. Authors propose four simple hypotheses about how position bias might arise. Authors carry out a large data-gathering effort, where authors perturb the ranking of a major search engine, to see how clicks are affected. Authors then explore which of the four hypotheses best explains the real-world position effects, and compare these to a simple logistic regression model. The data are not well explained by simple position models, where some users click indiscriminately on rank 1 or there is a simple decay of attention over ranks.

**McMahan et al. [17]** showed that the success of sponsored search has radically affected how people interact with the information, websites, and services on the web. Sponsored search provides the necessary revenue streams to web search engines and is critical to the success of many online businesses. However, there has been limited academic examination of sponsored search, with the exception of online auctions. In this paper, authors conceptualize the sponsored search process as an aspect of information searching. Authors provide a brief history of sponsored search and an extensive examination of the technology making sponsored search possible. Authors critique this technology, highlighting possible implications for the future of the sponsored search process.

## III. THE SUPERVISED LEARNING MODEL FOR AD-CLICK PREDICTION

The supervised learning models are extremely effective for regression learning problems where the data and the targets are continuously marked. The implementation of neural network is defined in two phases' first training and second prediction: training method utilizes data and designs the data model. By this data model next phase prediction of values is performed. Traditional data mining techniques such as logistic regression have played a vital role in early CTR prediction systems. Logistic regression is popular because of its computational efficiency and interpretability. It models the click-through probability as a function of input features such as user demographics, device type, ad position, and historical click patterns. Feature engineering is critical for this method, including one-hot encoding, cross-feature generation, and sampling strategies to handle class imbalance. Although simple, logistic regression often struggles with nonlinear relationships present in user-ad interactions, making it less competitive for complex datasets [18].

**Training Vector Space:**

1. Prepare two arrays, one is input and hidden unit and the second is output unit.

2. Here first is a two dimensional array $W_{ij}$ is used and output is a one dimensional array $Y_i$.

3. Original weights are random values put inside the arrays after that the output [19].

$$\boldsymbol{x_j = \sum_{i=0} y_i W_{ij}} \tag{1}$$

Where,

$y_i$ is the activity level of the jth unit in the previous layer and

$W_{ij}$ is the weightof the connection between the ith and the jth unit.

4. Next, action level of $y_i$ is estimated by sigmoid function of the total weighted input.

$$\boldsymbol{y_i = \left[\frac{e^x - e^{-x}}{e^x + e^{-x}}\right]} \tag{2}$$

When event of the all output units have been determined, the network calculates the error (E).

$$\boldsymbol{E = \frac{1}{2}\sum_i (y_i - d_i)^2} \tag{3}$$

5. Calculation of error for the back propagation algorithm is as follows:

Error Derivative $(EA_j)$ is the modification among the real and desired target:

$$\boldsymbol{EA_j = \frac{\partial E}{\partial y_j} = y_j - d_j} \tag{4}$$

Here,
E represents the error
y represents the Target vector
d represents the predicted output

The supervised regression model is designed mathematically as [20]:

These techniques are based on the time series approach based on the fitting problem that accurately fits the data set at hand. The approach generally uses the auto-regressive models and means statistical measures. They can be further classified as:

a) Linear
b) Non-Linear

Mathematically:

Let the time series data set be expressed as:

$$\boldsymbol{Y = \{Y1, Y2 \ldots\ldots\ldots\ldots\ldots\ldots.. Yt)} \tag{5}$$

Here,
Y represents the data set
t represents the number of samples
Let the lags in the data be expressed as the consecutive differences.
The first lag is given by:

$$\boldsymbol{\Delta Y_1 = Y_{t-1}} \tag{6}$$

Similarly, the jth lag is given by:

$$\boldsymbol{\Delta Y_j = Y_{t-j}} \tag{7}$$

The graphical description of the model is given by:

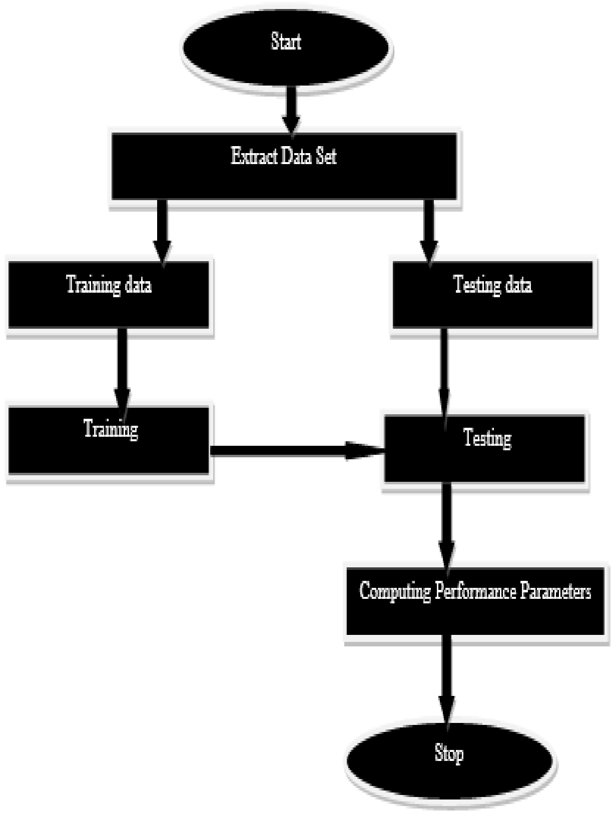

**Fig.2 The Supervised Regression Model**

Data mining techniques for CTR prediction have evolved from simple statistical methods to highly sophisticated deep learning architectures. Each technique—logistic regression, ensembles, factorization machines, clustering, NLP, and online learning—offers unique strengths and trade-offs depending on data complexity, scale, and application requirements. As user behavior becomes increasingly dynamic across devices and digital platforms, future CTR prediction systems will integrate multimodal data (text, images, video), personalized reinforcement learning, and explainable AI to deliver more accurate, transparent, and real-time decision-making in digital advertising ecosystem

## IV. EVALUATION PARAMETERS

The performance metrics of the machine learning based classifier is generally done based on:
The parameters which can be used to evaluate the performance of the ANN design for time series models is given by:
1) Mean Absolute Error (MAE)
2) Mean Absolute Percentage Error (MAPE) and
3) Mean square error (MSE)
The above mentioned errors are mathematically expressed as [21]:

$$MAE = \frac{1}{N}\sum_{t=1}^{N} |V_t - \hat{V}_t| \quad (8)$$

Or

$$MAE = \frac{1}{N}\sum_{t=1}^{N} |e_t| \quad (9)$$

$$MAPE = \frac{100}{N}\sum_{t=1}^{N} \frac{|V_t - \hat{V}_t|}{V_t} \quad (10)$$

The mean square error (MSE) is given by:

$$MSE = \frac{1}{N}\sum_{t=1}^{N} {e_t}^2 \quad (11)$$

Here,

N is the number of predicted samples
V is the predicted value
$\hat{V}_t$ is the actual value
e is the error value

**Conclusion: It can be concluded form the previous discussions that advances presented in this study, such as supervised regression learning can be utilized for ad-click prediction. Essentially, the proposed methods can be utilized in any task where one needs to find a good match among the instances from two distinct sources of free text data. Prominent examples of such tasks are online recommender systems, where best match of product description and user's query should be found; professional networking services where one needs to match appropriate job opportunities and prospective employees based on requirements and skills in textual form; or online dating sites where users should be matched based on the textual descriptions of themselves. The prominent work in the domain and evaluation parameters have also been presented.**

**References:**

1. F. Arbab, A. Iftikhar and M. S. Awan, "XGBDeepFM for CTR Predictions in Mobile Advertising," *Complexity*, Wiley Online Library, vol. 2020, Article ID 1747315, Apr. 2020.

2. H. Pande, "Field-Embedded Factorization Machines for Click-through rate prediction," *arXiv preprint arXiv:2009.09931*, Sep. 2020. arXiv

3. W. Zhang and J. Wang, "Deep Field-Aware Interaction Machine for Click-Through Rate Prediction," *Complexity*, Wiley Online Library, vol. 2021, Article ID 5575249, Apr. 2021.

4. H. Zhang, H. Du, F. Du, J. Zhu and J. Li, "An Attention-based Deep Network for CTR Prediction," in *Proceedings of the 43rd*

*International ACM SIGIR Conference*, . ACM Digital Library, 2020

5. F. Wang, H. Hu, L. Li et al., “FRNet: Enhancing CTR Prediction with Context-Aware Feature Interaction,” in *Proceedings of SIGIR 2022 (short paper / workshop paper)*, 2022.

6. W. Zhang, R. Liu, J. Zhang et al., “Graph Attention Interaction Aggregation Network for Click-Through Rate Prediction (GAIAN),” *Applied Sciences / PMC*, 2022.

7. Z. Luo, “Click-Through Rate Prediction Models based on Interest Modeling,” *ACM Transactions / Proc.* (full article), 2023

8. Jelena Gligorijevic·Djordje Gligorijevic1·Ivan Stojkovic1·Xiao Bai1· Amit Goyal2·Zoran Obradovic, “Deeply supervised model for ad-click-through prediction for sponsored research, Springer 2019.

9. Zhabiz Gharibshah1, Xingquan Zhu, Arthur Hainline, and Michael Conway, “Deep learning for online display advertising user clicks and interest prediction, Springer 2019

10. H Zhang, J Pan, D Liu, J Jiang, X Li, “Deep pattern network for click-through rate prediction”, Proceedings of the 47th International ACM SIGIR Conference on Research and and Development in Information Retrieval, ACM,2024, pp. 1189 – 1199

11. Z Xiao, L Yang, W Jiang, Y Wei, Y Hu, “Deep multi-interest network for click-through rate prediction”, CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, ACM, 2020, pp.2265-2268.

12. N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey “An experimental comparison of click position-bias models”, WSDM, 2008.

13. D. C. Fain and J. O. Pedersen, “ Sponsored search: A brief history.” ,Bulletin of the American Society for Information Science and Technology,2006.

14. A. Galen and G. Jianfen, “Scalable training of l1-regularized log-linear models.”,ICML 2007.

15. T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich, “ Web-scale bayesian click-through rate prediction for sponsored search advertising in Microsoft’s Bing search engine” ICML 2010.

16. T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. “Web-scale bayesian click-through rate prediction for sponsored search advertising in Microsoft’s bing search engine “, ICML 2010.

17. X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, and J. Q. n. Candela. Practical lessons from predicting clicks on ads at facebook. ADKDD, 2014..

18. Y. Juan, D. Lafortier, and O. Chapelle. Field-aware factorization machines in a real-world online advertising system. In WWW, 2017.

19. Y. Juan, Y. Zhuang, W.-S. Chin, and C.-J. Lin. Field-aware factorization machines for CTR prediction. RecSys, 2016.

20. A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models., 2013..

21. H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. M. Hrafnkelsson, T. Boulos, and J. Kubica. Ad click prediction: a view from the trenches. In KDD, 2013.

.