# A Survey on Data Mining Models For Click-Through-Rate Prediction

**Ashutosh Chaturvedi[1], Prof.Priyanka Prajapati[2]**
[1, 2] Dept of CSE
[1, 2] AIT, Ujjain

***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 ThroughRates (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].

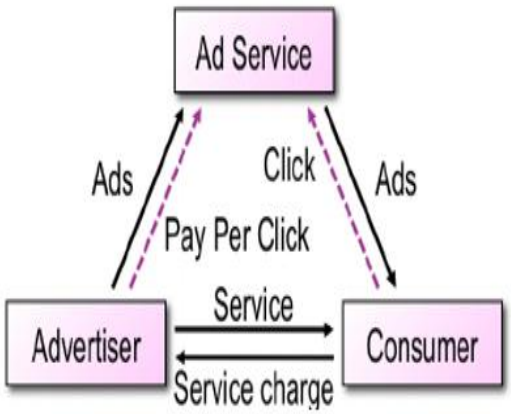


***Figure.1 The pay per click model***

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 [4].

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 [5].

Retargeting, behavioral targeting, and contextual advertising all are designed to increase anadvertiser'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 [6].

## II. PREVIOUS WORK

This section presents an overview of the existing work in the domain.

The analysis of existing approaches show that ML-powered systems exist on both sides of this transaction, boosting yield for publishers and performance for advertisers. While click-through rates are the focus of this piece and one of the most frequently-used and easy-to-measure key performance indicators, it should be noted that a diverse array of models exist to optimize each problem with a specific machine learning approach.

**Xiao et al. in [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.

**Gligorijevic et al. in [8]** proposed a deeply supervised architecture that jointly learns the semantic embeddings of a query and an ad as well as their corresponding CTR. The authors showed that search it is critical to match ads that are relevant to a query and to accurately predict their likelihood of being clicked. Commercial search engines typically use machine learning models for both query-ad relevance matching and click-through-rate (CTR) prediction. However, matching models are based on the similarity between a query and an ad, ignoring the fact that a retrieved ad may not attract clicks, while click models rely on click history, limiting their use for new queries and ads. The paper also puts forth a novel cohort negative sampling technique for learning implicit negative signals. The model trained the proposed architecture using one billion query-ad pairs from a major commercial web search engine. This architecture improves the best-performing baseline deep neural architectures by 2% of AUC for CTR prediction and by statistically significant 0.5% of NDCG for query-ad matching.

**Zhang et al. in [9]** showed that recommendation systems and computing advertisements are of great value for commercial applications. Click-through rate (CTR) prediction is a critical issue because the prediction accuracy affects the user experience and the revenue of merchants and platforms. Feature engineering is usually used to improve the click-through rate prediction; however, it heavily relies on user experience. It is difficult to construct a feature combination that can describe the complex patterns implied in the data. This paper combines the traditional feature combination methods and the deep neural networks to automate the feature combinations to improve the accuracy of the click-through rate prediction. We propose a mechanism named Field-aware Neural Factorization Machine (FNFM). This paper can have strong second-order feature interactive learning ability, such as Field-aware Factorization Machine; on this basis, a deep neural network is used for higher order feature combination learning. This experiment shows that the model has stronger expression ability than previous deep learning feature combination models, such as the DeepFM, DCN, and NFM.

**Qu et al. in [10]** proposed a dynamic CTR prediction model designed for the Samsung demand-side platform (DSP). From our production data, authors identify two key technical challenges that have not been fully addressed by the existing solutions: the dynamic nature of RTB and user information scarcity. To address both challenges, authors develop a Dynamic Neural Network model. Our model effectively captures the dynamic evolutions of both users and ads and integrates auxiliary data sources (e.g., installed apps) to better model users' preferences. The paper puts forward a novel interaction layer that fuses both explicit user responses (e.g., clicks on ads) and auxiliary data sources to generate consolidated user preference representations. The papers evaluates the model using a large amount of data collected from the Samsung advertising platform and compare our method against several state-of-the-art methods that are likely suitable for real-world deployment. The evaluation results demonstrate the effectiveness of our method and the potential for production. In addition, we discuss how to address a few practical engineering challenges caused by big data toward making our model in readiness for deployment.

**Zhang et al. in [11]** proposed a weighted output extreme learning machine (WO-ELM) to learn the imbalanced data. A hierarchical extreme learning machine (H-C-ELM) is proposed based on the proposed WO-ELM and the weighted extreme learning machine (W-ELM). The H-C-ELM has two levels in its structure. In the first level, the WO-ELM and the

W-ELM are trained on different combined fields of the CTR (each field has some attributes). The two extreme learning machines (ELMs) output their predicted scores of the corresponding combined fields of the CTR. The WO-ELM and the W-ELM have different predicted results on the same combined fields because of the difference of the two ELMs. Therefore, in the second level, another ELM is applied based on the outputs of the two ELMs in the first level and the actual outputs in order to improve the prediction accuracy.

**Dhanani et al. in [12]** showed that majority of Web users utilize search engines to locate Web site links. Based upon the search queries provided by the users, search engines display sponsored advertisements together with actual Web site link results to procreate monetary benefits. However, users may click the concerned sponsored advertisements that generate revenue for the search engines based upon a predefined pricing model. Furthermore, by analyzing previous information of users, advertisements, and queries; search engines estimate click-through rate (CTR) for predicting users' clicks. CTR is a ratio of clicks to number of impressions associated with a particular advertisement. In this paper, we propose a model, based on CTR, to estimate probabilities of clicks using logistic regression that determines parameters using stochastic gradient ascent method (SGA). Moreover, this paper also summarizes the comparative analysis of SGA and batch gradient ascent (BGA) methods, in terms of accuracy and learning time.

**She et al. in [13]** proposed that a single structure model does not take into account the characteristics including highly nonlinear association for features. Aiming at this problem, this paper presents a click through rate prediction model based on CNN (Convolutional Neural Networks) and FM (factorization machine). This model uses CNN to extract high-impact features, and predicts and classifies them by FM, which can learn the relationship between mutually distinct feature components. The experimental results show that compared with the single structure model, the CNN-FM hybrid model can effectively improve the accuracy of advertising click through rate prediction.

**Liu et al in [14]** proposed that due to real-time response nature of the online digital advertising eco-systems, it is vital to accurately estimate the CTR in real-time. In this paper, we propose a URL truncation based fast page grouping for real-time CTR estimation (ULTR-CTR). Our hypothesis is that web pages under the same URL folder have similar page style and semantic content, and will share similar CTR values. While grouping web pages based on the page content is computationally expensive and hardly scalable to real-time applications, we use simple URL truncation to estimate CTR values of different site-folder combinations. Our empirical study and A/B test carried out on a commercial bidding engine confirm that ULTR-CTR based bidding achieves 2.0% performance gain in CTR estimation, and 1.4% lift in Gross Profit (GP) gain.

**Hao et al. in [15]** proposed that computational ddvertising aims to advertise to specific group of audience and has been a hotspot issue in the field of emerging internet applications. The key problem is to predict the Click Through Rate(CTR) of an ad and it is usually done by machine learning ways. The proposed a method based on feature engineering and online training to predict the CTR of Search Ads. We use the Field-aware Factorization Machine(FFM) to abstract highly sparse feature vectors from the original ones and trained it with Follow-the-Regular-Leader(FTRL). Experiment results show that the method we proposed is 0.65%~6.44% more accurate than common prediction model, LR, and 29.72% more efficient than normal training methods.

**Zhu et al. in [16]** proposed a Softmax-based Ensemble Model, SEM, which adopts only a few key features after feature hashing for CTR estimation. The information of the bid request and the ad contains categorical attributes (such URL) and numerical attributes (such ad size). To vectorize the information for the input of regression-based approaches, the categorical attributes will be expanded to several binary features in general. However, some categorical attributes have infinite possible values (such as URL). Thus, for these attributes, only observed values in training will be transformed into binary features. If there is a new attribute or value in online environment, this information will be lost after vectorization. The experimental results demonstrate that our proposed approach is able to adapt to the harsh environments in RTB, and outperforms the state-of-the-art approaches effectively when only less than 50 features are adopted in two real datasets.

**Bai et al. in [17]** showed that that language representation can be learnt from scratch at character level when trained on enough data. Through extensive experiments using billions of query-advertisement pairs of a popular commercial search engine, we demonstrate that both approaches significantly outperform a baseline model built on well-selected text features and a state-of-the-art word2vec-based approach. Finally, by combining the predictions of the deep models introduced in this study with the prediction of the model in production of the same commercial search engine, we significantly improve the accuracy and the calibration of the click-through rate prediction of the production system.

## 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 [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 andoutput is a one dimensional array $Y_i$.

3. Original weights are random values put inside the arrays after that the output**.**

$$x_j = \sum_{i=0} y_i W_{ij} \quad (1)$$

Where,
$y_i$ is the activity level of the $j^{th}$ unit in the previous layer and $W_{ij}$ is the weightof the connection between the $i^{th}$ and the $j^{th}$ unit.

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

$$y_i = \left[\frac{e^x - e^{-x}}{e^x + e^{-x}}\right] \quad (2)$$

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

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

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:

$$EA_j = \frac{\partial E}{\partial y_j} = y_j - d_j \quad (4)$$

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

The supervised regression model is designed mathematically as:

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 [19]:

a) Linear
b) Non-Linear

Mathematically:

Let the time series data set be expressed as:

$$Y = \{Y1, Y2 \ldots\ldots\ldots\ldots\ldots Yt) \quad (5)$$

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:

$$\Delta Y_1 = Y_{t-1} \quad (6)$$

Similarly, the $j^{th}$ lag is given by:

$$\Delta Y_j = Y_{t-j} \quad (7)$$

The graphical description of the model is given by:

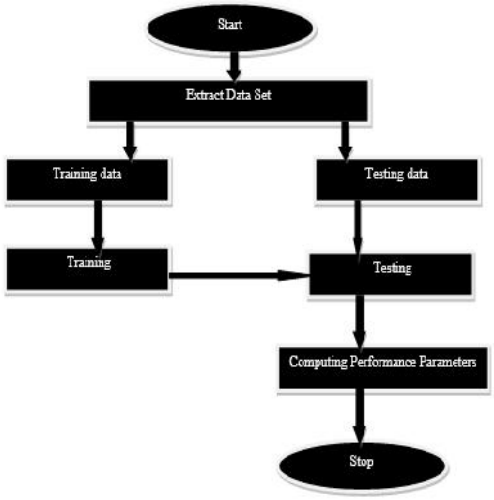


***Figure.2 The Supervised Regression Model***

## 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:

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

Or

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

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

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
$\widehat{V}_t$ is the actual value
e is the error value

## V. 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] JA Choi, K Lim, "Identifying machine learning techniques for classification of target advertising", ICT Express, Elsevier 2022, vol. 6, no. 3, pp. 175-180.
[2] M. Gan and K. Xiao, "R-RNN: Extracting User Recent Behavior Sequence for Click-Through Rate Prediction," in IEEE Access, 2021, vol. 7, pp. 111767-111777.
[3] Q. Wang, F. Liu, P. Huang, S. Xing and X. Zhao, "A Hierarchical Attention Model for CTR Prediction Based on User Interest," in IEEE Systems Journal, 2019., vol. 14, no. 3, pp. 4015-4024.
[4] L. Y. Akella, "Ad-Blockers — Rising threat to digital content: Business analytics study," 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), 2017, pp. 324-32.
[5] G. Chauhan and D. V. Mishra, "Evaluating deep learning based models for predicting click through rate," 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), 2019, pp. 1-5.
[6] X Wang, G Hu, H Lin, J Sun, "A novel ensemble approach for click-through rate prediction based on factorization machines and gradient boosting decision trees", APWeb-WAIM 2019: Web and Big Data, Springer 2019, pp 152–162.
[7] 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, 2021, pp.2265-2268.
[8] J Gligorijevic, J .Gligorijevic., D. Stojkovic, "Deeply supervised model for click-through rate prediction in sponsored search" Data Min Knowledge Discovery, Springer 2019, vol. 33, pp: 1446–1467.
[9] L. Zhang, W. Shen, J. Huang, S. Li and G. Pan, "Field-Aware Neural Factorization Machine for Click-Through Rate Prediction," in IEEE Access, 2019, vol. 7, pp. 75032-75040.
[10] X. Qu, L. Li, X. Liu, R. Chen, Y. Ge and S. -H. Choi, "A Dynamic Neural Network Model for Click-Through Rate Prediction in Real-Time Bidding," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 1887-1896.
[11] S. Zhang, Z. Liu and W. Xiao, "A Hierarchical Extreme Learning Machine Algorithm for Advertisement Click-Through Rate Prediction," in IEEE Access, 2018, vol. 6, pp. 50641-50647.
[12] J Dhanani, K Rana, "Logistic Regression with Stochastic Gradient Ascent to Estimate Click Through Rate", Information and Communication Technology for Sustainable Development, Springer 2018, pp.319-326.

[13] X. She and S. Wang, "Research on Advertising Click-Through Rate Prediction Based on CNN-FM Hybrid Model," 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2018, pp. 56-59.

[14] H. Liu, X. Zhu, K. Kalish and J. Kayne, "ULTR-CTR: Fast Page Grouping Using URL Truncation for Real-Time Click Through Rate Estimation," 2017 IEEE International Conference on Information Reuse and Integration (IRI), 2017, pp. 444-451.

[15] C. Jie-Hao, L. Xue-Yi, Z. Zi-Qian, S. Ji-Yun and Z. Qiu-Hong, "A CTR prediction method based on feature engineering and online learning," 2017 17th International Symposium on Communications and Information Technologies (ISCIT), 2017, pp. 1-6.

[16] W. -Y. Zhu, C. -H. Wang, W. -Y. Shih, W. -C. Peng and J. -L. Huang, "SEM: A Softmax-based Ensemble Model for CTR estimation in Real-Time Bidding advertising," 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), 2017, pp. 5-12.

[17] B Edizel, A Mantrach, X Bai, “Deep character-level click-through rate prediction for sponsored search”, SIGIR '17: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2017 pp. 305–314.

[18] K. Ren, W. Zhang, K. Chang, Y. Rong, Y. Yu and J. Wang, "Bidding Machine: Learning to Bid for Directly Optimizing Profits in Display Advertising," in IEEE Transactions on Knowledge and Data Engineering, 2017, vol. 30, no. 4, pp. 645-659.

[19] A Kumar, A Nayyar, S Upasani, A Arora, “Empirical Study of Soft Clustering Technique for Determining Click Through Rate in Online Advertising”, Data Management, , Analytics and Innovation, Springer 2020, vol.42, pp 3-13.