Predicting the Click Through Rate Using Machine Learning Methodologies

[1]Sukrant Kathuria, [2]Anjesh Kumar,[3]Shikha Gupta, [4]Atul Mishra
[1,2]Research Scholar, [3]Asst. Professor, [4]Professor
[1]Department of Information Technology,
[1]J.C. Bose University of Science and Technology, YMCA, Faridabad, India

Abstract: Computational Advertising is the currently emerging model in the advertising industry. Web pages visited per user every day is considerably increasing day by day which results in the vast access to display advertisements (ads). The main metric facilitates the measurement of the effectiveness of an advertisement is termed as Click Through Rate (CTR) [3], it's the rate at which the ad is clicked by the user. The placement of ads in appropriate location leads to the rise in the CTR value that influences the growth of customer access to advertisement which is beneficial for publisher and advertiser. Thus, it is very important to predict the CTR metric in order to formulate efficient ad strategy for placement. This paper proposes a predictive model that generates the click through rate based on different dimensions of ad placement for display advertisements using machine learning techniques such as decision trees, random forest, SGD-based logistic regression. The experiment result reports that SGD-based logistic regression-based click model outperforms in predicting CTR. Further this paper is divided into 5 sections, section-1 contains introduction, section-2 contains literature review, section-3 contains ctr prediction models, section-4 contains experiments and result, section-5 contains conclusion and future work and section-6 contains references.

Index Terms **- Click Through Rate (CTR), Contextual Advertisements, Machine Learning, Web advertisements, Decision tree, random forest, SGD-logistic regression.**

1.INTRODUCTION

Online advertisement is a widely considered component in the current marketing industry that acts as a major resource provider for the web users. Advertisement is broadly categorized [5] into display advertisements or banner ads, sponsored search, and contextual advertisements. In context search the ad is selected based on the match with the page content and in sponsored search ad is selected based on search query. The tedious task is about the selection of ad in display advertisements which is decided by the ad exchanges. The percentage at which the user clicks the ad out of the impression is termed as click through rate. If the user visits the application or the website that gives impact on the CTR of the ad. Position of the ad also plays a important role in increasing the CTR of an ad. There is the whole process from ad server to publisher to ad network that involves administration of the ads and its distributing ways. Before receiving the ad files the ad server allocates them in different websites.

As the user visit some webpage the ad gets displayed, the challenging task is of predicting the click through rate. There are various factors which are affecting the value of CTR. The features that are going to contribute in the evaluation of CTR gets analyzed. In the online advertisement the prediction of the click through rate is the important metric. When an ad is fetched from the source its impression gets counted each time, its click or not click is recorded to measure the impression. To measure the ad campaign performance, CTR is the metric behind it. Impression [5] is defined as the number of times the ad is displayed to the user. CTR is used to measure the number of clicks ad of the advertiser receive on the ads per the number of times it is displayed.

The CTR of an ad is calculated as No. of clicks on an ad divided by the No. of times ad is shown, and it can be expressed in the form of percentage. This vision to the advertisement agencies of identifying the most viewed ads is being done through predicted value of CTR. Predicting the future response of the ads helps in generating the relevant and the better quality and revenue also

2. LITERATURE REVIEW

Various analogous research works are reviewed and analyzed to understand the nature and circumstances of the work. The purpose has been well studied and based on the literature survey the proposed work is identified.

Syed Abbas Ali et al., presented a model [1] that is innovative and unique way of solving the advertisement prediction problem which is considered as a learning problem over the past several years. Their main goal of this research is to enhance CTR of the contextual advertisements using Linear Regression along with some dynamically added feature known as the keyword. They propose a new technique in their research to predict the CTR which will increase overall revenue of the system by serving the ads more suitable to the viewers with the help of feature extraction and displaying the ads based on context of the publishers. From a different angle their proposed technique helps to calculate the CTR despite causing a minor decrease in efficiency. In their

research, it also proved that CTR is also dependent on the keywords, the accuracy found is 83% and removing some feature could take this accuracy to 95% which fits the model perfectly and a major increase.

Michael Young et al., presented a model [2] which include furtherance in the context of traditional supervised learning based on an FTRL-Proximal online learning algorithm which has excellent sparsity and combining properties and the use of per-coordinate learning rates. They also covered some of the challenges that arise in a real-world system that may appear at first to be outside the domain of classic machine learning research, that include beneficial tricks for methods for inspection, memory savings and practical methods for providing confidence estimates for predicted probabilities, assessment methods, and methods for automated management of features. Their goal was to highlight the close connection between theoretical advances and practical engineering in this industrial setting, and to show the depth of challenges that appear when applying classic machine learning methods in a complex dynamic system.

Lihue Shi et al., presented a model [3] in which they tried to predict the CTR and average CPC of a keyword using some machine learning methods. They used cross validation to evaluate the results and find the optimal predictors for the CTR and average CPC. A very crucial task for advertisers is to use the historical data to predict the CTR and average cost per click (CPC) for a keyword with a set of features. The CTR and average CPC are two fundamental metrics to measure the paid search performance on keyword level, so that the advertisers will be able to optimize the bids and attain the highest profits for their Google AdWords accounts. They investigate the keywords' CTR and average CPC prediction problems on Google AdWords using a wide range of performance features. Different ML methods, including regression, random forest and gradient boosting are applied to evaluate the prediction performance on both metrics. They found that random forest turns out to be the best method for both the CTR prediction and the average CPC prediction, while the gradient boosting gives the most inaccurate results. Though it was interesting to determine the best features, and how much each feature may overlap with other features, they believed that ultimately, the best habit is to take as many as possible features in the final model.

3. CTR PREDICTION MODEL

CTR prediction is the sustainable model for assessing the performance of the display advertisement. The importance of the CTR prediction in the field of display ad led to the proposal of this work with greater reliability. The research is initiated by data collected from an ad server which contains click log details. The dataset is then analyzed to identify the relevant features which are going to contribute to predict the CTR on the selected advertisement. From the raw dataset, the training dataset is prepared through data transformation and feature extraction. The CTR models are build by employing ML algorithms namely Decision Tree, Random forest, SGD- Based Logistic regression. The proposed model and the various phases involved in this research are illustrated in fig.1.

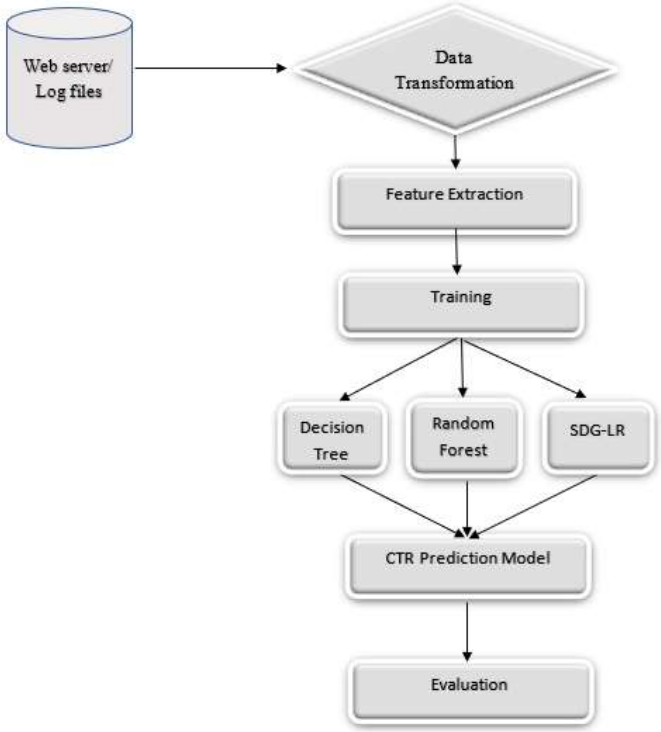

Fig 1: Architecture of proposed CTR prediction model

3.1 Dataset

The dataset is acquired from https://www.kaggle.com/c/avazu-ctr-prediction. The dataset contains 10 days of click log data ordered with attributes like id, click, hours, banner pos, site id, site domain, site category, app id, app domain, app category, device id, device ip, device model, device type, device connection type and C14 to C21 which are categorized as anonymized variables. The data initially is not appropriate for building model for predicting CTR. Hence data transformation and feature extraction are performed to attain max prediction accuracy.

3.2 Data Pre-processing

To ensure meaningful data mining result data pre-processing is necessary. Initially data collected are pre-processed so that they are within the span of the problem framed. The irrelevant features like device id, device ip, device model, device connection type, hour, record id are found to be out of scope of the anticipated model and hence removed.

3.3 Features Extraction

The feature extraction phase has the important responsibility in building the CTR model and predicting the CTR metrics efficiently, from the various dimensions of the ad location the features are extracted, such as site domain, web site, application, neighboring ads like previous ad and next ad in each web page and application domain information.

3.3.1 Component-1, Site Domain and its Metrics

The several websites with the same domain attract frequent visitors. The ad that is displayed on these websites influences the repeated users of a domain to click on the advertisements. The number of clicks, impression and the CTR of the site domain where the ad is placed is included in the feature vector on this motive. The appropriate click attribute and the site domain id in the pre-processed dataset are computed and the number of clicks, impression and CTR are extracted.

3.3.2 Component-2, Websites and its Metrics

The website on which the ad is placed has to be considered. If a site has been visited more often, then it is certain that the ad is displayed to the users many times and this increases the number of times an ad being viewed and clicked. The site CTR highly stimulates the ad CTR. So, the number of clicks, impression and the CTR of the site where the ad is placed is contemplated in the feature vector. The related click attribute and site id in the dataset are processed to derive the impression, CTR and number of clicks, on website.

3.3.3 Component-3, Application Domain and its Metrics

The various applications having the same domain have regular visitors and the ads are displayed to them more frequently. The users are motivated to click on the delivered ads. The number of clicks, impression and the CTR of the application domain where the ad is placed is included in the feature vector for the prediction model. The corresponding click attribute and application domain id are considered for generating these features.

3.3.4 Component-4, Previous, next Ads and its Metrics

In display advertisement, the specific traits and commercial message is conveyed are used to target the audience. The source dataset with click attribute is utilized with respect to the mean values for number of clicks, impression, next ads, and CTR are computed. The customers obtaining an ad during the regular browsing activity are mostly induce by the ad that is located next to the ad in focus. There are many ads that are in the next locality to the ad under reflection in different web sites. The user who prone to click the ad has the high probability of viewing the adjacent ads. This might steer them to click on the ad and go for that ad service provided. The ad placed in previous location to the selected ad is referred to as previous advertisement. The previous advertisement varies for every different website in which the advertisement is placed. Also, there are more than one previous ad for the ad under review. The respective, impression, clicks and CTR of the previous advertisement have serious impact on the selected advertisement. The preceding ad metrics are thus generated by handling the click attribute of the source dataset with respect to the previous id. Increase in CTR metric gets highly affected, which motivates the metrics namely impression, number of clicks and CTR of the next ad to be included as features for the proposed model. The mean values for no. of clicks, CTR and impression of preceding ads are extract forming the features for analysis.

3.3.5 Component-5, Current Ads and its Metrics

From the transformed dataset the unique ads are found to be distributed in different setting and grouped to obtain corresponding impression, no. of clicks and the CTR which is the response variable is produced using the formula [4],
CTR=Clicks/Impression *100

3.4 Model Generation

The CTR metric of the targeted ad is taken as the response for generating the prediction models. The three variants of machine learning namely Decision tree, Random forest and SGD-logistic regression algorithms are implemented for building the models.

3.4.1 Decision Tree

Decision tree is one of the simple non-linear supervised learning method [5] which can be used for estimation of CTR. The tree model that predicts the label of a given bid request is done by learning a simple sequential, tree structured (binary), decision rule from the training data, a bid request instance x is parsed through the tree on the basis of its attributes, and at the end arrived at one of the leaves. The weight assigned to the leaf is used as the prediction.

3.4.2 Random Forest

One of the most power full and fully automated machine learning technique is random forest, it is combined learning method for regression and classification, this is done by constructing a multitude of decision tree at training time and outputting the class i.e., classification or regression. It's a collection of decision tree that together performs prediction and detailed insights into structure of data. This out of the box data is used to get a running unbiased estimate of the classification error as trees are added to the forest. Estimation of Variable importance can also be done.

3.4.3 SGD-Regression

For the regression and classification problems another machine learning technique can be use ie., gradient boosting, it produces prediction model in the form of an ensemble of other weak prediction models, specially decision trees. In the stage- wise fashion model is built and generalization is done by optimization of an arbitrary differential loss function. It is the integration of "gradient descent" plus "boosting". Here, the learning procedure progressively fits new models to provide a more accurate estimate of the variable response. It is widely used in many applications due to its efficiency, easy use, feasibility and accuracy.

4. EXPERIMENTS AND RESULT

Experiments are carried out in Python environment to build the CTR model for display ad by implementing machine learning techniques. Feature extraction and Preprocessing processes have been carried out as described on section 3. Decision tree, random forest, SGD- logistic regression are the three algorithms that are trained for the model generation and the models are evaluated for their performance. For the above model AUC score is examined for accuracy and SGD based logistic regression outperforms as shown in fig2.

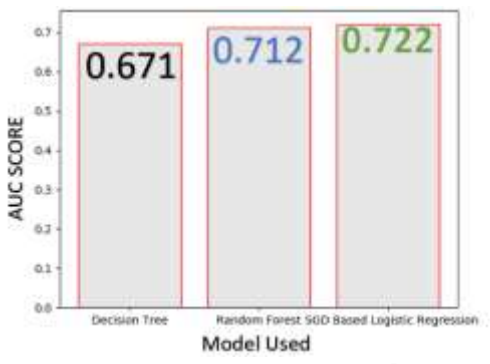

Fig 2: AUC score result

5. CONCLUSION AND FUTURE SCOPE

This research work demonstrates the modeling and implementation of CTR prediction for displaying advertisement.

The features for contribution are identified based on various component of the target ad in the web page and extracted to develop the model. Supervised machine learning algorithms namely decision tree, random forest and SDG based logistic regression are adopted to construct the models. It is observed from the experimental results that SGD based logistic regression is more efficient than the other two model in predicting CTR. As the scope for future research, the model can be generated using more related features and large dataset and the models that were built can be integrated to improve the prediction result by creating ensemble model.

6. REFERENCES

[1] Muhammad Junaid Effendi and Syed Abbas Ali,"Click Through Rate Prediction for Contextual Advertisment Using Linear Regression".
[2] H. Brendan McMahan et all," Ad Click Prediction: a View from the Trenches", KDD'13.
[3] Lihui Shi and Bo Li, " Predict the Click-Through Rate and Average Cost Per Click for Keywords Using Machine Learning Methodologies", Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management Detroit, Michigan, USA, September 23-25, 2016.
[4] Avila Clemenshia P and Vijaya M.S.," Click Through Rate Prediction for Display Advertisement", International Journal of Computer Applications (0975 – 8887) Volume 136 – No.1, February 2016
[5] Breiman, L., Friedman, J. H., Stone, C. J., and Olshen, R. A. (1984). Classification and regression trees. CRC press.