International Journal of Scientific Research in Computer Science and Engineering
Research Paper
Vol.10, Issue.5, pp.45-51, October (2022)
E-ISSN: 2320-7639

Click Through Rate Prediction Employing Wavelet Tree and Regression Learning

Meghna Chandel[1*], Sanjay Silakari[2], Rajeev Pandey[3], Smita Sharma[4]

[1,2,3,4]Department of Computer Science and Engineering, UIT-RGPV, BHOPAL- 460236, India

Available online at: www.isroset.org

Received: 10/Aug/2022, Accepted: 20/Sept/2022, Online: 31/Oct/2022

***Abstract*:** Click through rates have proven to be a critical factor in deciding the effectiveness of online advertising models. 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. Forecasting 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 forecasting problem, and thus can be estimated by machine learning models. The training data is collected from historical ads impressions and the corresponding clicks. An estimate of click through prior to fetching an add for a query is important for the accurate decision in the context. In this work a recursive binary partitioning algorithm is used along with support vector regression to estimate the bipolar nature of add clicks. A comparative analysis has also been made with exiting baseline techniques and it has been found that the proposed approach attains better performance metrics compared to baseline techniques.

***Keywords*:** Online Advertising, Data Mining, Click Through Rates (CTR), Wavelet Tree, Regression Learning, Support Vector Regression, Prediction Accuracy.

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.

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. 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. There has been a metamorphosis in the advertising realm with the conventional techniques such as billboards, print media and television media facing extremely large competition from the online advertising platform by dint of the fact that a multitude of users purchase online which has increased with the increase in the infrastructure and reliability of online marketing. This has resulted in a necessary requirement to churn out ads specific and apt to the queries entered. The outcome may be a possible click or not based on the experience of the customer. An estimate of click through prior to fetching an add for a query is important for the accurate decision in the context. 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 [1].
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 [2].

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.

1.2 Overview of Click through rate Management

Click through rate management is the management of the flow of goods and services and includes all processes that transform raw materials into final products. It involves the active streamlining of a business's supply-side activities to maximize customer value and gain a competitive advantage in the marketplace. SCM represents an effort by suppliers to develop and implement click through rates that are as efficient and economical as possible [3].

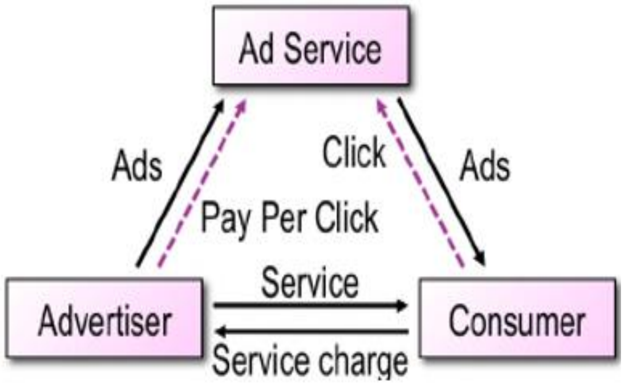


Figure 1.1. The pay per click model

1.3 The Pay per Click Model

The new age advertising models try and leverage the much larger audience which is constantly on the internet and new branding and advertising methods have some up such as [4]:

- Sponsored search advertising
- Contextual advertising
- Display advertising
- Affiliate marketing
- Online brand influencing
- Real-time bidding auctions etc.

It is extremely important to choose the correct or apt ads for a quarry to maximize the probability of clicks. Is the estimates of click through can be made accurately; they may materialize into staggeringly large profits. For instance an accuracy increase of 0.1% may increase the chances of increasing the profits by a million dollars depending on the diaspora of the audience the add is catering to. In some instances, click baits are also employed to increase the click through rate, which is a bottom-line for the pay per click model. Typically, the pay per click is measured as:

$$PPC = \frac{Cost_{tot}}{N} \quad (1.1)$$

Here,
PPC is the pay per click
$Cost_{tot}$ corresponds to the total add cost.
N is the number of measured clicks

From a business point of view, lesser pay per click is profitable for companies while add designers try to increase the number of clicks using the users online trails, search trends, cookies etc. from the add server [3]. The multitude of data to be managed in this case is staggeringly large and hence effective techniques to manage the same is challenging. The tasks are generally computation heavy and hence machine learning based approaches are needed to accomplish the same [4]

II. PROPOSED METHODOLOGY

- The most critical challenges associated with ad click through rate predictions are:
- The multivariate data is often extremely complex and user preference dependent to find exact patterns.
- Click baits often manipulate the accuracy.
- Just the **2-way polarity** of data makes the data (Click: Yes or No) extremely prone to errors, unlike continuously changing data.
- Due to the complexity of the data, the applications become very computationally heavy with high time complexities making relatively accurate systems slow.

It is generally extremely difficult to achieve high prediction accuracy for complex bi-polar data such as add clicks. The flow chart of proposed methodology has been given below.

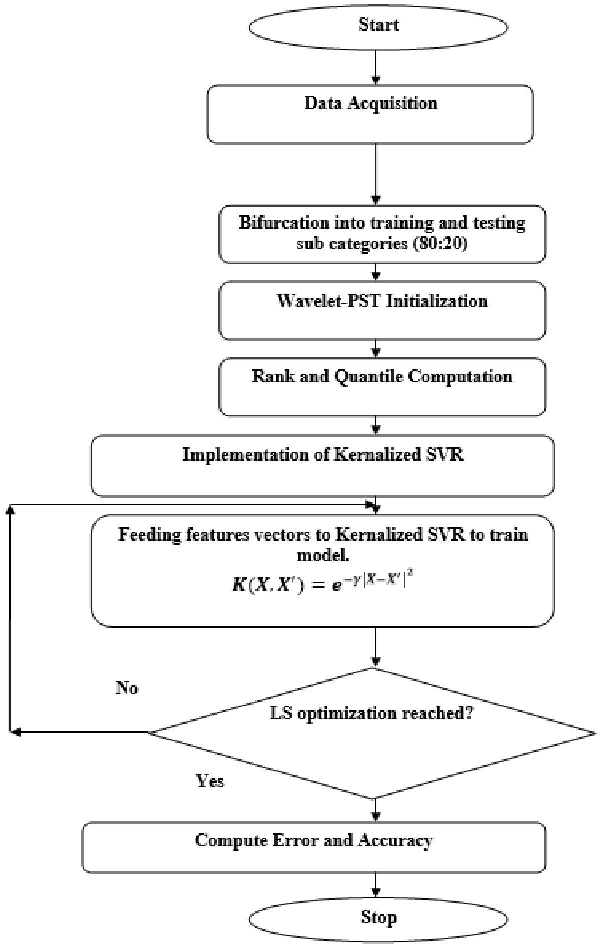


Figure 1. Flowchart of Proposed System

III. NEED FOR CTR PREDICTION

Sponsored search typically uses keyword based auction. Advertisers bid on a list of keywords for their ad campaigns. When a user searches with a query, the search engine matches the user query with bidding keywords, and then selects and shows proper ads to the user. When a user clicks any of the ads, the advertiser will be charged with a fee based on the generalized second price. A typical system involves several steps including selection, relevance filtration, CTR prediction, ranking and allocation [20].

- The input query from the user is first used to retrieve a list of candidate ads (selection). Specifically, the selection system parses the query, expands it to relevant ad keywords and then retrieves the ads from advertisers' campaigns according to their bidding keywords.
- For each selected ad candidate, a relevance model estimates the relevance score between query and ad, and further filters out the least relevant ones (relevance filtration).
- The remaining ads are estimated by the click model to predict the click probability given the query and context information (click prediction). In addition, a ranking score is calculated for each ad candidate here *bid* is the corresponding bidding price.
- These candidates are then sorted by their ranking score (ranking). Finally, the top ads with a ranking score larger than the given threshold are allocated for impression (allocation), such that the number of impressions is limited by total available slots. Click Probability

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 techniques.

IV. SUPPORT VECTOR REGRESSION

The support vector regression (SVR) model is a modified version of the support vector machine (SVM) with a modification in the objective or loss function. The advantages of support vector machines are [23]:

1) Effective in high dimensional spaces.
2) Still effective in cases where number of dimensions is greater than the number of samples.
3) Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
4) Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:
1) If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.
2) They generally exhibit saturation in performance with addition of data.

SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see Scores and probabilities, below).The classification using the support vector machine (SVM) is depicted in figure 4.7.

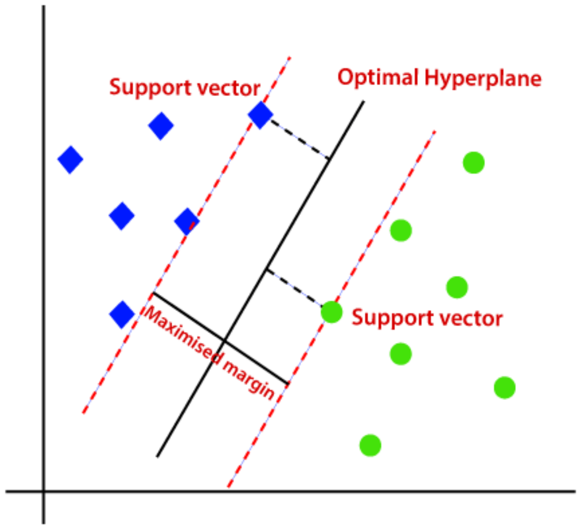


Figure 4.7 The SVM model [8]

The support vector regression can be designed as a least squares optimization (LS optimization) as:
for (i=1 : n)
{
Update weights and bias
And

$$Minimize\{\frac{e_1^2+e_2^2+\cdots\ldots\ldots e_n^2}{n} \tag{4.8}$$

}
The SVR can be used for both linearly separable and non-linearly separable data.

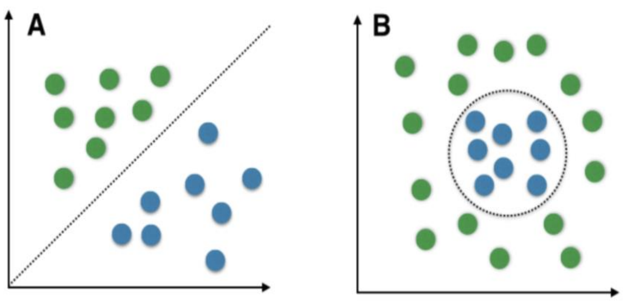


Figure 4.8 The linearly and non-linearly separable data

The least squares minimization approach is the fastest and most stable approach to convergence. The iterative update of the support vectors keeps changing the bias and weights to minimize the least squares objective function. The parameters to be evaluated for the CTR prediction are:

1) Date and time
2) Product
3) Campaign
4) Product category
5) Webpage
6) User group
7) Gender
8) Age level
9) City (location)

The Kernalized SVR is modelled as:
To analyze non-linear data sets, linear SVR model is not applicable. Hence, Kernelized SVR is needed.

The Kernel is typically a non-linear function. The Radial Basis Function Kernel (RBF) is the similarity between two points in the transformed feature space. Mathematically the SVR-RBF is defined as:

$$K(X,X') = e^{-\gamma|X-X'|^2} \tag{4.9}$$

$$\gamma = \frac{1}{2\sigma} \tag{4.10}$$

Here,
$\boldsymbol{\gamma}$ is called the free parameter of RBF
$\boldsymbol{\sigma}$ is called the feature factor
K represents the RBF Kernel
$\boldsymbol{X}$ ***and*** $\boldsymbol{X'}$ are the samples in an input feature space
$|\boldsymbol{X}-\boldsymbol{X'}|$ is termed as the Euclidean Distance

The loss function is computed as:
The support vector regression can be designed as a least squares optimization (LS optimization) as:
Minimize $\{\frac{1}{n}\Sigma(predicted\ value - actual\ value)^2\}$

Or

Minimize$\{\frac{1}{n}\Sigma(error)^2\}$

The dependent variable is chosen as the occurrence of click (1) or non-click (0). The evaluation parameters are the accuracy and percentage errors given by:

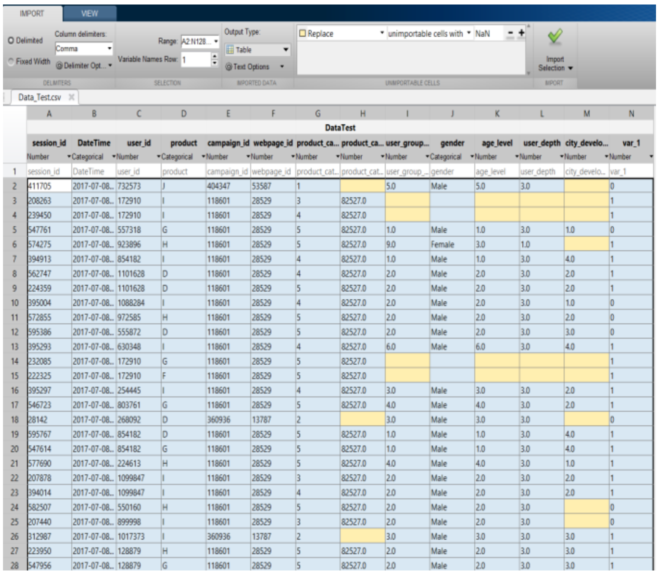
$$error\% = \frac{false\ classifications}{total\ classifications} * 100 \tag{4.11}$$

$$Ac = 100 - error\% \tag{4.12}$$

V. RESULTS

The results of the proposed system are evaluated in terms of the error % and the classification accuracy.
The training vector is designed as the training parameters along with the rank and quantile values of the outcomes to feed the SVR model.
Once the system is trained, the testing is done based on the testing data. The data division has been done in the ratio of 80:20 based for training and testing. The add sample and click polarity are recorded for the computations.

The tokenization (target formation) of the clicks are done as:
1: Expected Click
0: No Click
-1: Diverted click



Figrue 5.1 Raw Data

Once the system is trained, the testing is done based on the testing data. The data division has been done in the ratio of 80:20 based for training and testing. The add sample and click polarity are recorded for the computations.

The tokenization (target formation) of the clicks are done as:
1: Expected Click
0: No Click
-1: Diverted click

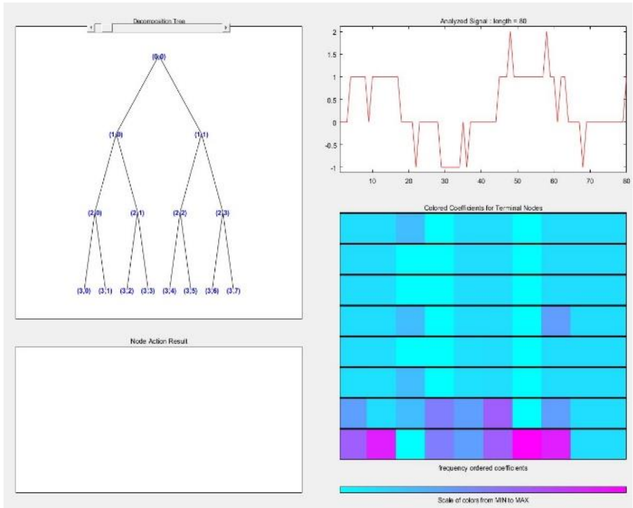


Figure 5.2 Initial Tree

Figure 5.2 depicts the initial tree for binary classification. The subsequent steps are to find the best level and the best tree for the given dataset.

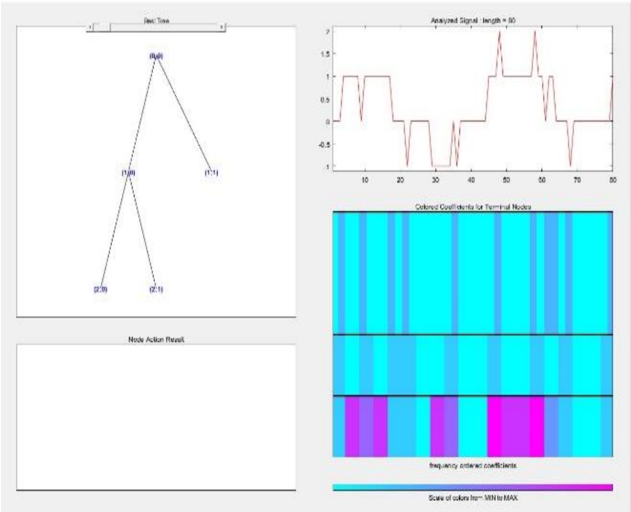


Figure 5.3 Best Tree

Figure 5.3 depicts the best tree among the possible bifurcations.

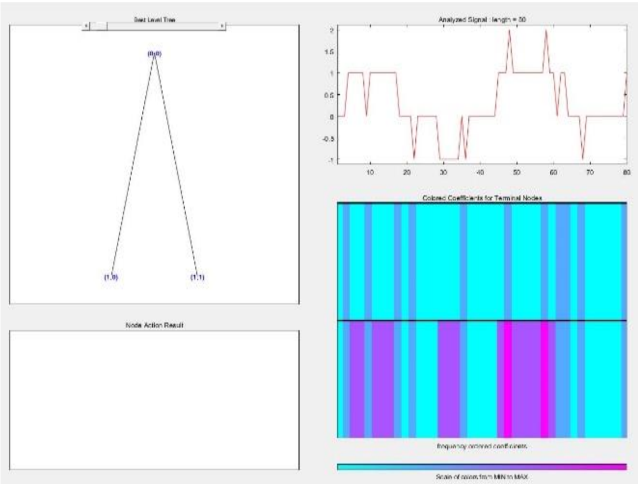


Figure 5.4 Best Level

Figure 5.4 depicts the best level among the possible bifurcations.

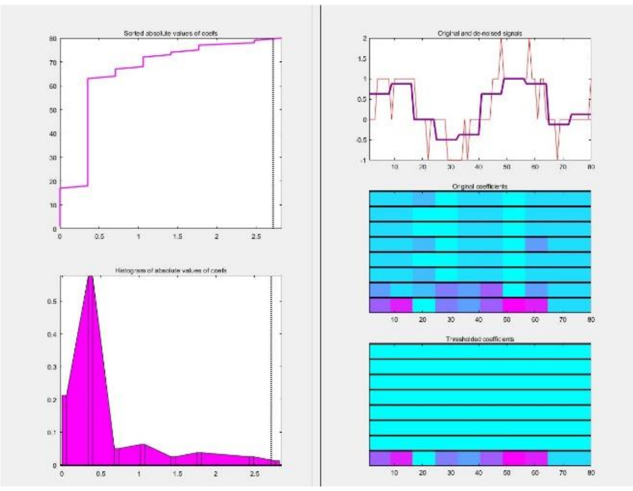


Figure 5.5 Denoised data using Shannon Entropy

Figure 5.5.depicts the denoised version of the data based on the Shannon entropy wherein the entropy is considered for smoothening operation.

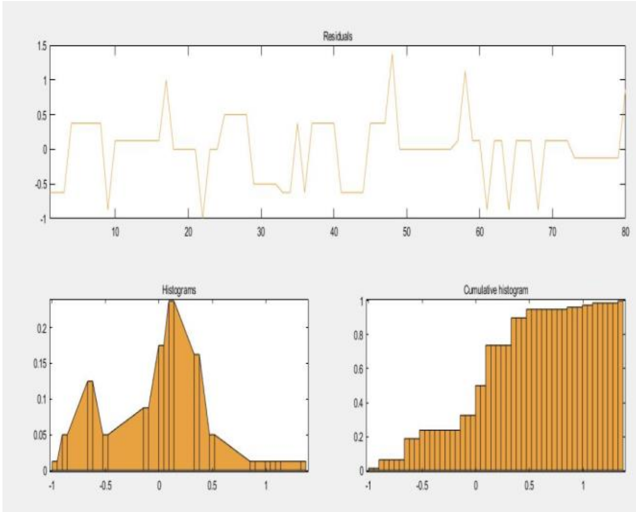


Figure 5.6 Histogram Analysis of Residuals

Figure 5.6 depicts the normal and cumulative histogram for the residuals of the decomposition.

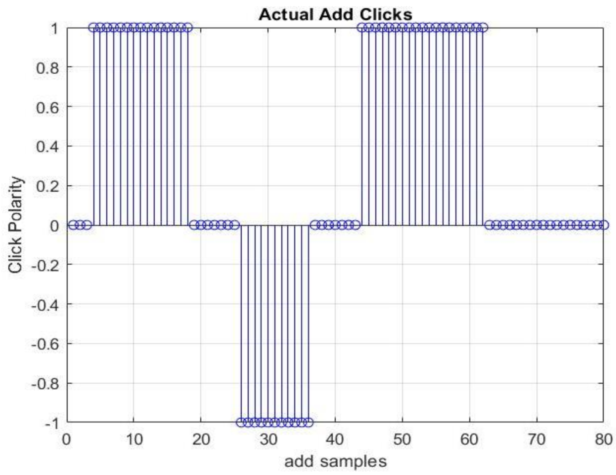


Figure 5.7 Actual Add Clicks

Figure 5.7 depicts the actual add click with three polarities of 1,-1 and 0.

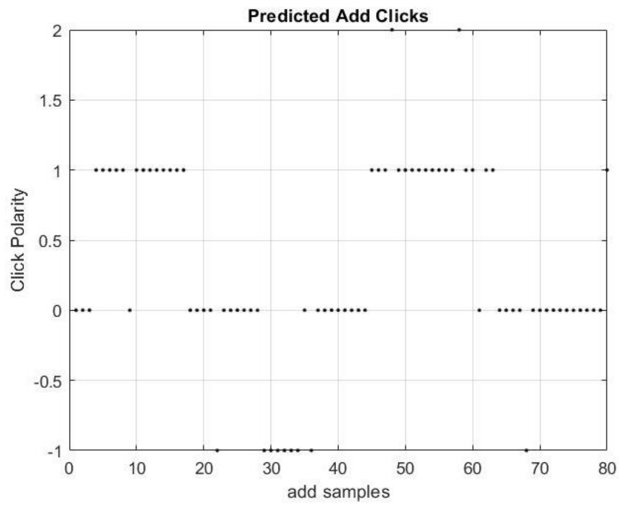


Figure 5.8 Predicted Add Clicks

Figure 5.8 depicts the predicted add clicks with three polarities of 1,-1 and 0.

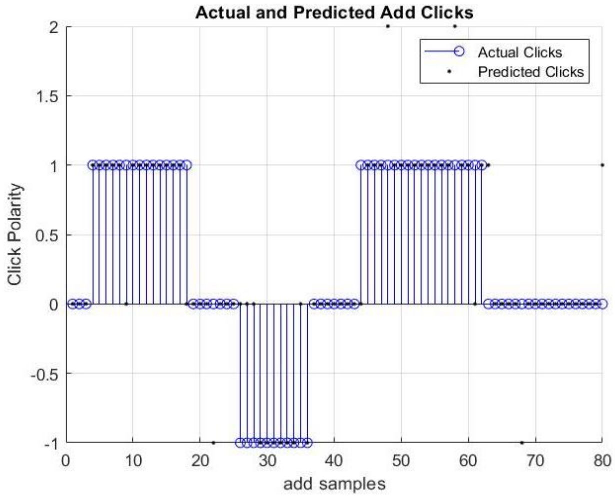


Figure 5.9 Predicted and Actual Clicks

Figure 5.9 depicts the predicted and actual clicks.

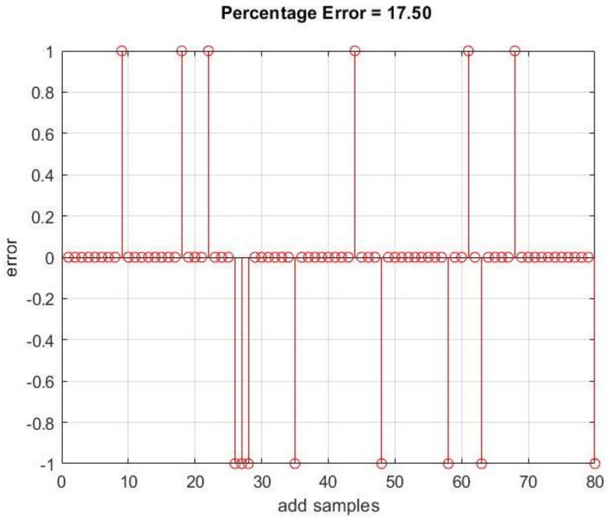


Figure 5.10 Errors and Percentage Error

Figure 5.10 depicts the sample wise errors and percentage errors. The errors are estimated with the condition of:

$$actual\ click\ \neq predicted\ click$$

The percentage error obtained in this approach is 17.50% and hence the accuracy of the system is 82.50%. This is significantly higher compared to the average accuracy of previous benchmark approach of 79%.

Table 1

S.No.	Parameter	Value
1.	Dataset	Query Add NDCG dataset. https://www.kaggle.com/wendykan/ndcg.
2.	Pre-Processing	PST
3.	Regression Model	SVR
4.	Kernel	Exponential RBF
5.	Error	17.50%
6.	Accuracy	82.50%
7.	Accuracy of Previous Work [1]	79%

VI. CONCLUSION AND FUTURE SCOPE

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. 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. 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. This work presents a recursive binary tree partition algorithm employing wavelet trees for data preparation. Subsequently the data is fed to a support vector regression model to estimate add click through rate (CTR). Previous discussion have emphasized upon the CTR and its estimation for online advertising models. The performance of the designed system has been evaluated in terms of the error% and classification accuracy. It has been shown that the error% of the system is 17.5% and the classification accuracy is 82.5% which is higher compared to the existing benchmark approaches .

Further advances in the proposed work can be thought of as:

1) Use of dynamic fractional cascading of trees to generate dynamic trees.

Use of Neural Backed Trees (NBT)

REFERENCES

[1] JA Choi, K Lim, "Identifying machine learning techniques for classification of target advertising", ICT Express, Elsevier, Vol. **6**, Issue.**3**, pp.**175-180**, **2020.**

[2] M. Gan and K. Xiao, "R-RNN: Extracting User Recent Behavior Sequence for Click-Through Rate Prediction," in IEEE Access, Vol. **7**, pp.**111767-111777, 2019.**

[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, Vol.**14**, Issue.**3**, pp.**4015-4024, 2019.**

[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), pp.**324-332, 2017.**
[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), pp.**1-5, 2019.**
[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, pp.**152–162, 2019.**
[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, pp.**2265-2268, 2020.**
[8] J Gligorijevic, J .Gligorijevic., D. Stojkovic, "Deeply supervised model for click-through rate prediction in sponsored search" Data Min Knowledge Discovery, Springer, Vol.**33**, pp:**1446–1467, 2019.**
[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, Vol.**7**, pp.**75032-75040, 2019.**
[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), pp.**1887-1896, 2019.**
[11] S. Zhang, Z. Liu and W. Xiao, "A Hierarchical Extreme Learning Machine Algorithm for Advertisement Click-Through Rate Prediction," in IEEE Access, Vol.**6**, pp.**50641-50647, 2018.**
[12] J Dhanani, K Rana, "Logistic Regression with Stochastic Gradient Ascent to Estimate Click Through Rate", Information and Communication Technology for Sustainable Development, Springer, pp.**319-326, 2018.**

