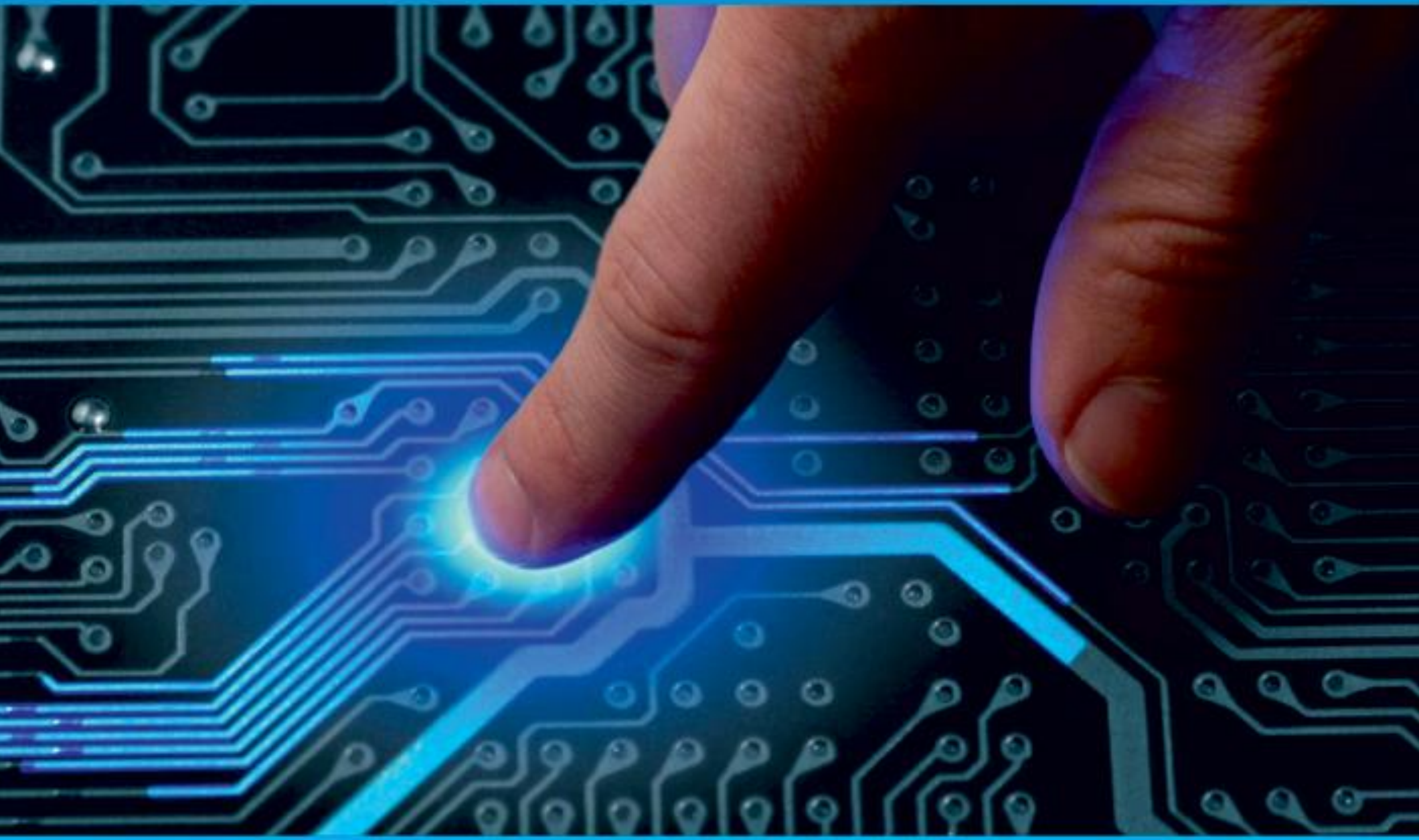
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 3, March 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

Ad Boost Maximizing Click-Trough Rates with Advanced Predictive Modelling

R.Vamsi, T. joshua, U.Anil Kumar , S. Praneeth , G. Lakshmi Narayana

Department of Computer Science and Engineering, KKR & KSR Institute of Technology and Sciences (Affiliated to JNTUK), Guntur, India

ABSTRACT*:* Online advertising relies heavily on predicting click-through rates (CTR) to optimize revenue and campaign performance. In this research, we present a data-driven methodology leveraging real-world sponsored search data and LightGBM, a gradient boosting framework tailored for large datasets. Our approach incorporates a multi-layer LightGBM architecture to capture diverse feature interactions, enhancing predictive accuracy. Experimentation with actual sponsored search data showcases the superiority of our method over existing techniques, evidenced by improved AUC, log-loss, and other key metrics. Additionally, we propose a user-friendly Flask web application that integrates advanced machine learning techniques, allowing advertisers to input ad details and obtain real-time CTR predictions. This comprehensive system offers a promising solution for the dynamic challenges of online advertising optimization.

I. INTRODUCTION

In the realm of advertising, particularly in the ever-expanding domain of online marketing, understanding and optimizing Click Through Rate (CTR) is paramount. CTR, denoting the ratio of clicks an advertisement receives to the number of times it is displayed, serves as a pivotal metric in predicting and enhancing ad engagement. As the landscape of advertising continues to evolve, with online platforms dominating the scene, businesses are increasingly reliant on efficient algorithms to gauge the likelihood of ad clicks. This necessitates a comparative analysis of various machine learning algorithms such as Decision Trees, XGB, Random Forest, and LGBM to ascertain the most effective predictor of ad click probability (CTR). Furthermore, the rise of online advertising, bolstered by its global accessibility, precise targeting capabilities, measurability, interactive features, cost-effectiveness, and adaptability, has revolutionized the marketing paradigm. Leveraging machine learning, particularly Logistic Regression and reinforcement learning models, has emerged as a promising avenue to not only estimate CTR but also optimize ad placement, mitigating issues such as ad fatigue and ensuring continued user engagement. With the market value of online advertising surpassing $50 billion and the increasing reliance on tailored advertising strategies, the intersection of machine learning and advertising holds immense potential in driving revenue growth and enhancing user experience in the digital age.

We experimented estimation of CTR by testing with Light Gradient Boosting Machine classifier. First we train this system using advertisement data sets with known target values - clicked and 0 - not clicked). The model generated after training is used to predict the probability of clicking the ad with unseen input feature values.

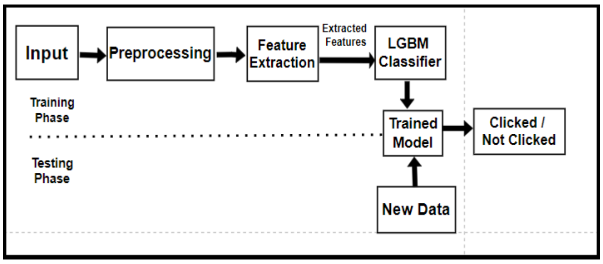


Fig. 1. System Model for CTR prediction Classifier

Figure 1 shows the process of training a classifiers and estimating the probability of clicking the new Ad.Literature Review is presented in Section 2. Feature Extraction is introduced in Section 3. Logistic Regression classifier are described in Section 4. Experimental results and discussions are shown and discussed in Section 5.

II. LITERATURE REVIEW

Estimation of click-through rates (C.T.R) have been examined from a mixed bag of points before as the most common method of advertising is on-line advertising [1] Table I shows the comparison of various advertising methods over the past five years. It can be observed that, on-line advertising is constantly increasing its growth rate and outperforming all other approaches of advertising methods. Research work has been carried on to predict the number of clicks per ad, the variables that contribute to the CTR, and predicting which advertisements ought to be shown amid web surfing.

The click-through rate (CTR) prediction in the realm of online advertising reveals a diverse array of approaches and methodologies. Lakshmanarao et al. [1] (2021) explored ad prediction utilizing CTR and machine learning, incorporating reinforcement learning techniques. Meanwhile, Kumar et al. [2] (2015) proposed a Logistic Regression classifier for CTR estimation, demonstrating promising results. Jaisinghani et al. [4] (2022) delved into CTR prediction using decision tree-based algorithms, offering insights into the efficacy of such approaches. Yu et al. [5] (2021) introduced FedDeepFM, a model for CTR prediction based on federated factorization machines, showcasing the potential of collaborative learning paradigms. Wang et al. [3] (2021) presented a novel CTR prediction model leveraging the xDeepFM network architecture, emphasizing the importance of innovative neural network architectures in enhancing prediction accuracy. Uncu [6] (2021) contributed to the field by investigating ad click prediction using various machine learning algorithms, shedding light on the comparative performance of different approaches. Li et al. [7] (2021) proposed a CTR prediction model based on DeepFM tailored for Taobao data, illustrating the adaptability of models to specific platforms or domains. Zhang [8] (2021) explored CTR prediction using xDeepFM and Bayesian optimization, highlighting the role of optimization techniques in refining prediction accuracy. Hudson et al. [9] (2020) presented smart advertisement strategies for maximizing clicks in online social networks without relying on user data, addressing privacy concerns while maintaining effectiveness. Tekin and Çebi [10] (2019) contributed a real-world application of click and sales prediction for digital advertisements, offering practical insights into ad campaign optimization. Lastly, Effendi and Abbas Ali [13] (2016) investigated CTR prediction for contextual advertisement using linear regression, providing foundational research in the field. These studies collectively contribute to the understanding and advancement of CTR prediction in online advertising, offering a diverse array of methodologies and insights for future research and application.

III. METHODOLOGY

First, we collected a “Ad_10000records” dataset from Kaggle [15]. The dataset contains information about 10000 ad clicks. After that, we checked for missing values in the data. Later, we applied feature extraction on dataset and applied Light Gradient Boosting Machine algorithm for predictions. The proposed framework is shown in figure-1.

A. Feature Extraction

Feature Extraction: Feature extraction is the procedure of reducing total features of a dataset by creating new features. The existing features are matched and combined based on user requirements. The original parameters are then discarded. In this project (Fig.2), the features of time and clicks are combined to calculate clicks per minute, hour and day.

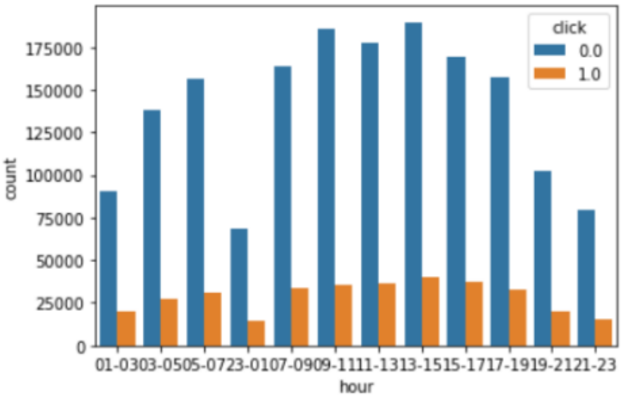


Fig.2. Graph of hours against number of clicks

B. LGBM Classsifier

In this study, we've opted for the LightGBM classifier to estimate the Click-Through Rate (CTR). LightGBM employs tree-based learning algorithms within its gradient boosting framework. It's known for its efficiency and high performance in dealing with large datasets. The predicted hypothesis in LightGBM yields probabilities, also bounded between 0 and 1. The features utilized for prediction are:

X1 → Represents the depth value (i.e., the frequency of clicks).

X2 → Indicates the position value (i.e., the display position of the ad).

Thus, the complete input feature vector for CTR prediction is denoted as:

$X = [x_0, x_1, x_2]$, where $x_0 = 1$ is the bias term.

In LightGBM, the hypothesis function is constructed using a combination of multiple weak learners (decision trees). The output of LightGBM is considered an estimation of the probability that a given input sample (advertisement) belongs to a class y = 1, which can be represented as:

$h\theta\ (X) = p(y = 1 \mid x; \theta)$

Where the probability that y = 1 given x is parameterized by θ.

The prediction of a given sample belonging to class y = 0 is given by:

$h\theta(X) = p(y = 0 \mid x; \theta)$

In LightGBM, we typically use a default threshold of 0.5 to classify advertisements as either clicked or not clicked. If the estimated hypothesis hθ(X) is greater than or equal to 0.5, it's considered clicked. Conversely, if hθ(X) is less than 0.5, it's treated as not clicked.

A. LIGHTGBM COST FUNCTION

In LightGBM, the parameter vector θ is optimized by minimizing the cost function J(θ). The cost function J(θ) for LightGBM typically involves minimizing the sum of errors between predicted and actual values. One commonly used cost function is the Mean Squared Error (MSE), given by:

$J(\theta) = 1/m * \Sigma\ (h\theta(xi) - yi)^2$, where m is the number of training samples.

However, LightGBM utilizes a different optimization approach compared to logistic regression. It employs a gradient boosting technique, which iteratively fits new models to the residuals of the previous models. This process optimizes a different objective function specific to boosting algorithms.

If we consider LightGBM's objective function, it's often related to minimizing loss, such as binary cross-entropy loss or multinomial logistic loss.The objective function might be expressed as:

$J(\theta) = 1/m * \Sigma\ \text{cost}(h\theta(xi), yi)$

Where cost(hθ(x), y) represents the specific loss function, which could be binary cross-entropy loss, multinomial logistic loss, or another suitable loss function for the problem at hand. This ensures that LightGBM iteratively minimizes the loss by fitting weak learners (decision trees) to the negative gradient of the loss function. This approach effectively leads to convergence towards a global minimum, ensuring optimal performance during training.

C. Algorithm Metrics

Four main parameters were considered for comparison of the algorithms : AUC score, F1 score, Accuracy and Precision and the one with the highest results in each category is deemed to be the best algorithm.

The parameters considered are : 1) AUC score - The Area Under the ROC Curve is an evaluation of the performance of a classifier to differentiate between classes and is used as the synopsis of the ROC curve. Primarily, it is utilized for addressing classification tasks.

$TPR = TP/(TP + FN)$ (4) $AUC = \int TPR * d(FRP)$ (1)

2) F1 score - It is a way to express the equilibrium between precision and recall, simply put it is a measure of a model's accuracy on a dataset.

$f1\ score = 2 * ((pr * rc)/(pr + rc))$ (2)

Where, : precision pr rc : recall

3) Accuracy - For any algorithm its accuracy is a measure of its ability to best predict relationships and patterns between various parameters of a dataset.

$acc = (TP + TN)/(TP + FP + TN + FN)$ (3)

TP : True positives

FP : False positives

TN : True negatives

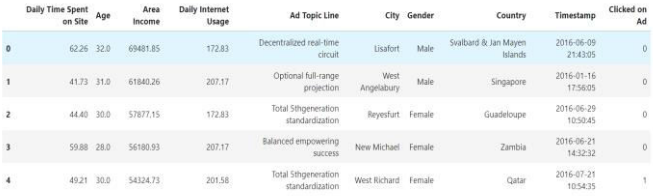
FN : False negatives

4) Precision - The measure of the quality of an algorithm to predict positives that are actually positive.

$Precision = TP/(TP + FP)$ (4)

IV. DATASET

The dataset used for this project is the ad_10000records[14]. It consists of 6 months of click-through data on online advertisements. 95 % of click-through data is used for training, and 5% of the data to test the model. It consists of 10 columns and almost 10 thousand rows and is a total of 8.99 MB size. The 10 columns include features like Daily Time Spent on Site, Age, Area Income, Daily Internet Usage, Ad Topic Line, City, Gender, Country, Timestamp, Clicked on Ad.



	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Gender	Country	Timestamp	Clicked on Ad
0	62.26	32.0	69481.85	172.83	Decentralized real-time circuit	Lisafort	Male	Svalbard & Jan Mayen Islands	2016-06-09 21:43:05	0
1	41.73	31.0	61840.26	207.17	Optional full-range projection	West Angelabury	Male	Singapore	2016-01-16 17:56:05	0
2	44.40	30.0	57877.15	172.83	Total 5thgeneration standardization	Reyesfurt	Female	Guadeloupe	2016-06-29 10:50:45	0
3	59.88	28.0	56180.93	207.17	Balanced empowering success	New Michael	Female	Zambia	2016-06-21 14:32:32	0
4	49.21	30.0	54324.73	201.58	Total 5thgeneration standardization	West Richard	Female	Qatar	2016-07-21 10:54:35	1

Fig.3. Overview of the dataset

V. EXPERIMENTS AND RESULTS

We implemented the LightGBM algorithm and compared its performance using PyCaret. We utilized the numpy library for array processing and matplotlib for plot visualizations. To optimize the cost function, we implemented a batch gradient descent algorithm with a learning rate α = 0.001.

For evaluating the robustness property of the classifier, we employed 10-fold cross-validation. In this process, we divided the dataset into 90% for training and the remaining 10% for testing, repeating this process for each part of the training sample. The final accuracy was calculated by averaging the results of all 10 iterations.

The average accuracy resulting from the classification of a dataset using 10-fold cross-validation turned out to be around 90%.Regarding the decision boundary, the decision boundary drawn by the LightGBM classifier for the advertisement dataset. From the scatter plot, it's evident that the training samples are linearly separable.

Furthermore, Figure 4 illustrates the graph of accuracy of the LightGBM classifier plotted against varying sizes of advertisement data. It's notable that as the size of the training dataset increases, the performance of the LightGBM classifier improves significantly.

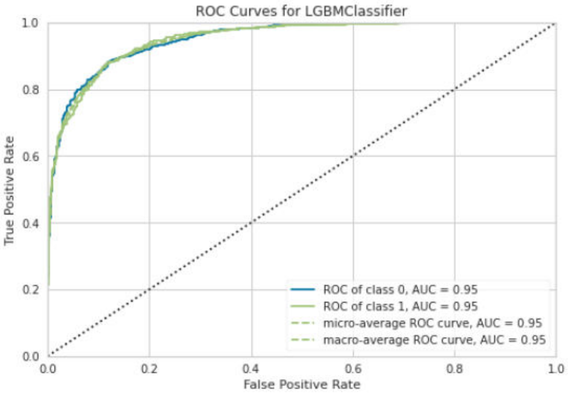
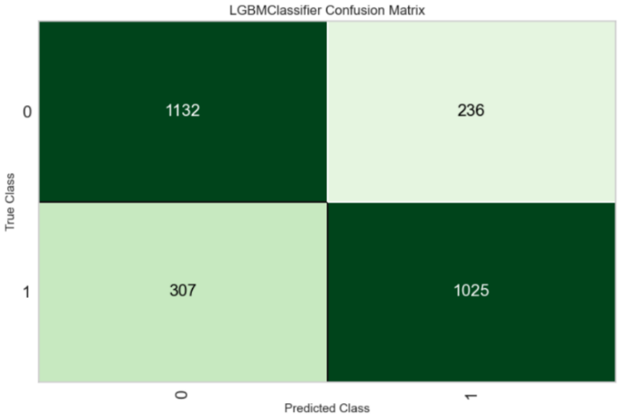


Fig.4. ROC curve for LGBM

In the above graph (Fig.7), the ROC curve for LGBM has been plotted and it has shown the best results till now. The value of 0.95 has been crossed, LGBM has the highest chance of differentiating correctly if an advertisement has been clicked or not.



Since the ROC curve for LGBM has shown the highest AUC score, it is important to look at the confusion matrix for this algorithm (Fig.8). Confusion matrix is a way to evaluate the performance of a classification model on a given dataset. It is an N*N matrix that provides a direct comparison between TP, FP, TN and FN.

The confusion matrix of LGBM shows that the algorithm correctly predicted 1025 positive values that were actually positive while predicting 1132 negative values that were actually negative. 307 false positive values were predicted, which means values which were actually negative were predicted as positive. 236 false negatives were predicted, which means the values which were actually positive were predicted as negative. Here positive is 1 (ad clicked) and negative is 0 (ad not clicked). Among the all algorithms, LGBM has shown the best results, as it has the highest AUC score, F1 score, Accuracy and Precision.

VI. Experiments and Results

In this study, we investigated the click-through rate (CTR) prediction challenge using ad_10000records dataset. CTR prediction, utilizing machine learning techniques, aims to estimate the likelihood of users clicking on an advertisement based on the number of impressions it receives. A high CTR value directly correlates with the success of an advertisement, leading to increased revenue for businesses. By accurately predicting CTR before launching advertising campaigns, businesses can optimize revenue while ensuring a positive user experience.

We used the machine learning algorithm LightGBM (LGBM), The study emphasized the importance of addressing challenges related to handling large-scale datasets and identifying high-cardinality categorical features.Through rigorous experimentation and benchmarking, we found that LGBM emerged as the top-performing model among the alternatives.

This highlights the significance of feature interaction and the incorporation of user-related data in improving model performance.

Future research avenues include enhancing efficiency and utility by leveraging all anonymized variables in the dataset. Additionally, exploring deep learning algorithms, employing diverse feature engineering techniques, and experimenting with alternative machine learning approaches can further enhance predictive accuracy and yield superior outcomes.

Moreover, the study suggests the potential development of a recommendation system tailored for advertising companies. Such a system could aid in identifying the most suitable users for specific advertisements at the optimal times, thereby maximizing the effectiveness of advertising campaigns.

REFERENCES

[1] A. Lakshmanarao; S. Ushanag; B. Sundara Leela,Ad Prediction using Click Through Rate and Machine Learning with Reinforcement Learning//International Conference on Electrical, Computer and Communication Technologies.2021

[2] Rohit Kumar; Sneha Manjunath Naik; Vani D Naik; Smita Shiralli; Sunil V.G; Moula Husain,Predicting clicks: CTR estimation of advertisements using Logistic Regression classifier//International Advance Computing Conference.2015.

[3] Peisong Wang; Minbo Sun; Zizheng Wang; Yihang Zhou, A Novel CTR Prediction Based Model Using xDeepFM Network//International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology.2021.

[4] Mayur Rattan Jaisinghani; Chirag Lundwani; Orijeet Mukherjee; Neeharika Nagori; Prerna Solanke,CTR Prediction of Advertisements using Decision Trees based Algorithms//International Seminar on Application for Technology of Information and Communication.2022.

[5] chenjia Yu;Shuhan Qi; Yang Liu,FedDeepFM: Ad CTR prediction based on Federated Factorization Machine//International Conference on Data Science in Cyberspace.2021.

[6] N. T. Uncu, “Ad click prediction using machine learning algorithms,” Capstone Project, January 2021.

[7] LinShu Li; Jianbo Hong; Sitao Min; Yunfan Xue,A Novel CTR Prediction Model Based On DeepFM For Taobao Data//International Conference on Artificial Intelligence and Industrial Design.2021.

[8] Yiying Zhang,CTR Prediction Model Using xDeepFM and Bayesian optimization//International Conference on Computer Science, Artificial Intelligence and Electronic Engineering.2021.

[9] Nathaniel Hudson; Hana Khamfroush; Brent Harrison; Adam Craig,Smart Advertisement for Maximal Clicks in Online Social Networks Without User Data//International Conference on Smart Computing.2020.

[10] A. T. Tekin and F. Çebi, “Click and Sales Prediction for Digital Advertisements: Real World Application for OTAs,” INFUS, July 2019, AISC 1029, January 2020.

[11] A.L.Rao,A.Srisaila,T.S.R.Kiran, “An Efficient Ad-Click Prediction System using Machine Learning Techniques” , International Journal of Engineering and Advanced Technology,Volume-9 Iss-3, 2020.

[12] M.J.Effendi,S.Abbas Ali, “Click Through Rate Prediction for Contextual Advertisment Using Linear Regression,” https://arxiv.org/abs/1701.08744v1-2016. Simon Breedon, ”The Rapidly Growing Digital Advertising Market,” http://viralmarketingmonsters.sharedby.co/share/8vbMff,” (Accessed: 10 February 2015).

[13] Tan Yu; Zhipeng Jin; Jie Liu; Yi Yang; Hongliang Fei; Ping Li, Boost CTR Prediction for New Advertisements via Modeling Visual Content//International Conference on Big Data.2010.

[14] www.kaggle.com/datasets/gauravduttakiit/clickthrough-rate-prediction?select=ad_10000records.csv

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING