IJIRCCE

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

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 12, December 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165|

|| Volume 10, Issue 12, December 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1012008 |

Click Through Rate Prediction

Nikhil Chaube[1] , Akash Gawai[2] , Atharv Sankpal[3] , K N Attarde [4]

UG Students, Dept. of Computer Engineering, Theem COE, Mumbai University, Maharashtra, India [1,2,3]

HOD, Dept. of Computer Engineering, Theem COE, Mumbai University, Maharashtra, India [4]

ABSTRACT: An industry worth more than 50 billion is online advertising. Due to the targeted advertising to certain users, internet advertisers are currently expanding exponentially.Past research in this area has covered everything from ad click prediction to ad serving. The problem of click prediction has gotten worse with the rapid expansion of ad networks.One of the most notable Machine Learning success stories is the one involving the advertisement serving.A real-time bidding solution has also been made possible by the development of ad serving technologies, where advertising are chosen in accordance with the characteristics of the publishers and the viewers.This project aids in the analysis of ad click prediction using various methods, including logistic regression and boosting algorithms, as well as the examination of the impact of various parameters on click rate.Additional features are also developed based on user usage patterns and other activities

KEYWORDS: Click Through Rate Prediction, Advertisement, XGBoost, LGBM, Logistic Regression, Machine learning.

I. INTRODUCTION

The online serving of adverts has benefited greatly from machine learning; extensive study has been done in this area. Observed in this field.The methodology and strategy employed cause a variance in the predicted number of ad clicks. technique.the extraction of features. For the regression, this phase has been carried out in a variety of methods. classification difficulties, linear discriminant analysis (LDA), which is utilised. We have seen a variety of solutions to the prediction of advertisement clicks. Numerous academics have done it using logistic regression, and naive bayes has been crucial in the development of ad click prediction.Researchers are looking for various methods to get targeted adverts based on viewers ' interests due to the consumers' rapidly increasing online activity.An other study in this area compares various machine learning methods, including logistic regression.

II. RELATED WORK

By utilising these characteristics of advertising, phrases, and advertisers, Matthew Richardson, Ewa Dominowski, and Robert Ragno learned a model that precisely predicts the click-through rate for new ads intheir paper, "Predicting clicks: Estimating the clicks-through rate for new ad."Using the model, we can further demonstrate that.We suggested updating and enhancing an ad system's performance and convergence.As a result, our model boosts customer satisfaction and revenue. Similar to search results, the likelihood that a person will click on an advertisement decreases fast with display position, by as much as 90%.The search engine should therefore prioritise the best-performing advertisements. Note that due to the likelihood of selecting an.The accuracy of our estimates of its CTR can have a substantial impact because advertise reduces so dramatically with advertise position.A huge impact on the amount of money made

III. PROPOSED ALGORITHM

In order to increase advertise quality (as measured by user clicks on the advertise) and total revenue, mostsearch engines today order them advertises primarily based on expected revenue:

Ead[revenue]=Pad(click) x CPCad

The most notable exception to this is Yahoo, which orders ads based on advertiser bids alone, but plans to switch to using expected revenue soon. The CPC for an advertise is its bid (in a first price auction) or the bid of the next-highest bidder (in

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165|

|| Volume 10, Issue 12, December 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1012008 |

a second-price auction), normalized by advertise performance. Every detail of the relation between CPC and bid are not prominent to this paper, but are the study of various works on search engine auction models.

Baseline CTR:

P(click|ad,pos) = p(click|ad,seen)p(seen|pos)

Let the CTR of an ad be defined as the probability it would be clicked if it was seen, or p(click | ad, seen). From the CTR of an ad, and the discounting curve p(seen | pos), we can then estimate the probability an ad would be clicked at any position. This is the worth we want to estimate, since it gives us a simple basis for comparison of competing advertises.

Light GBM [2] :

Light GBM is a gradient boosting framework that utilizes tree-based learning algorithms. It is made to be distributed and efficient with the following advantages:

- Quick training speed and high efficiency.
- Low memory usage.
- Increased accuracy.
- Supports parallel and GPU learning.
- Can hand large-scale data.

XG Boost:

XGBoost is an optimized distributed gradient boosting library designed to be much more efficient, flexible and portable. It uses machine learning algorithms under the Gradient Boosting framework. XGBoost provides us with a parallel tree boosting (also known as GBDT, GBM) that solves many data science problems in a quick and accurate way.

I. FLOWDIAGRAM

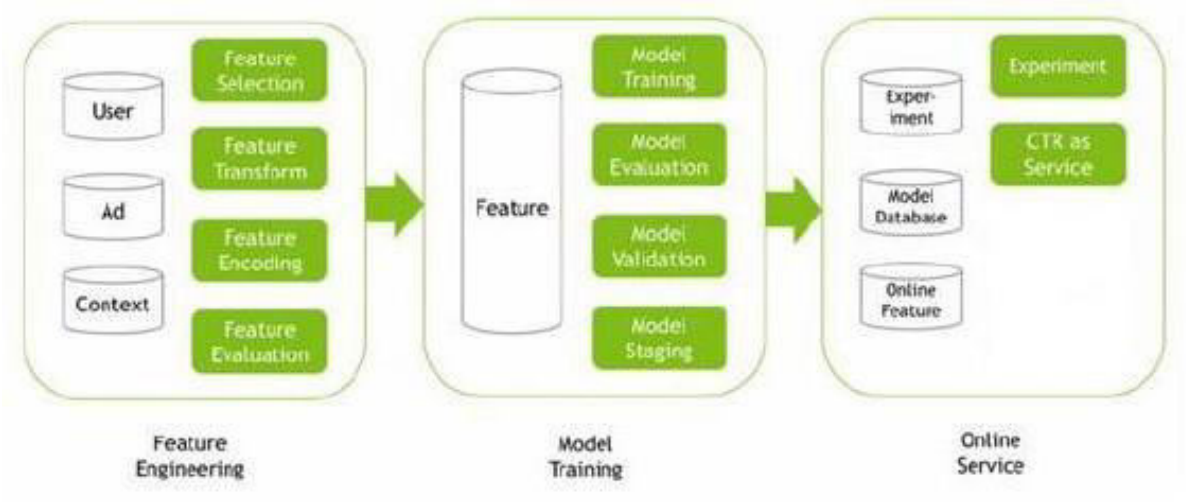

Figure 1:Flow diagram

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165|

|| Volume 10, Issue 12, December 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1012008 |

V. RESULT

Firstly, data visualization techniques yielded how the data was distributed with important features being highlighted, most of the data visualization suggested a non-linear, or not so strong correlation between any of the attributes and target variable. Logistic Regression model was trained and tested with the data and yielded very poor results of precision and recall both were 0, area under roc-curve was observed at 0.5, accuracy however in this case was higher due to imbalanced target variable click.

Logistic regression ROC-curve and classification report[1]

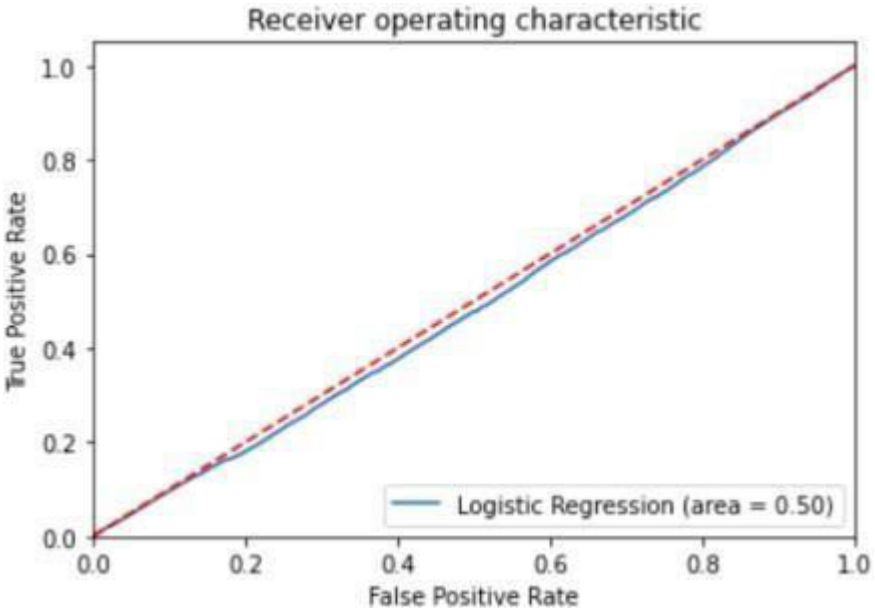

Figure2:Receiveroperatingcharacteristics

Comparison between XG boost and Light gbm

Next light gradient boosting (lgb) was used to train the data, it yielded a log-loss of 0.2545 after 100 num boost rounds and other performance metrics were 0.7 (precision-1), 0.99 (precision-0), 0.69 (recall-1), 0.99 (recall-0). Confusion matrix had False negative - 12,523,206, True positives-32,529, False positive- 13,834, True negatives- 14,600.

Finally, XGboost was tried with different parameters with objective: binary logistic and eval metric as log loss. It gave a log loss of 0.057569, other performance metrics were 0.75 (precision-1), 0.99 (precision-0), 0.68 (recall-1), 0.99 (recall-0). Confusion matrix had False negative - 12,556,104, True positives-31,857, False positive- 10,658, True negatives- 15,163.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165|

|| Volume 10, Issue 12, December 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1012008 |

Table 1: Comparison between different algorithms

	Log loss	**Confusion matrix values**			
		False Negative	True Positive	False Positive	True Negative
Light GBM	0.2545	12,523,206	32,529	13,834	14,600
XG Boost	0.0057569	12,556,104	31,857	10,658	15,163

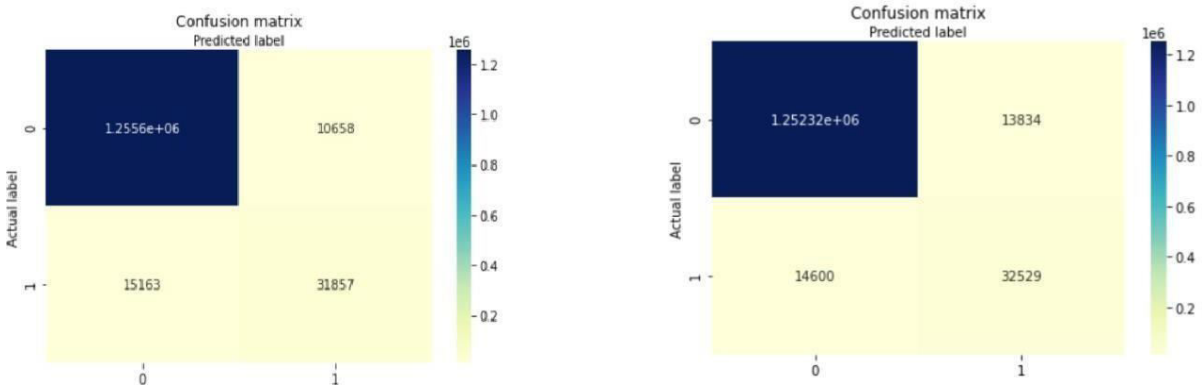

Figure3:Confusion matrix of Light GBM and XG Boost

VI. CONCLUSION AND FUTURE WORK

Thereby light gradient boosting though performed comparatively well with higher number of false positives, its log loss was much higher and tuned xgboost produced better results overall which may be improved further. Logistic regression and other classifiers which require some correlation cannot separate such data because of its non-linearity and high imbalance in the number of labels.

Lgb log loss comparatively was higher and hence its performance on real time data can decline significantly from xgboost. Furthermore, roc curve of former (0.97), and latter (0.98) though don't show any significant difference however the steep change in roc on left side can help prompt to have permissible amount of true positive with small false positives for the same and as a result is better for xgboost than light gbm. As future work, we plan to tune the parameters for xgboost for even better performance, some variable introduction/stratified sampling of the data for logistic regression analysis to see if it can perform better just on the basis of sampling. Feature engineering of some attributes with parameters tuning may be applied.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165|

|| Volume 10, Issue 12, December 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1012008 |

REFERENCES

1. “Predicting clicks: estimating the click- through rate for new ads,” M. Richardson, E. Dominowska, and R.Ragno, in event of the 16th international conference on the Internet, pp. 521–530, 2007.
2. "Advertise Click Prediction in Sequence with Long Short-Term Memory Networks: an area aware Model.." Deng, Weiwei, Ling, Xiaoliang, Qi, Yang, Tan, Tunzi, Manavoglu, Eren and Zhang, Qi. Paper presented at the meeting of the SIGIR, 2018.
3. “Click through rate prediction for contextual advertisement using linear regression,” M. J. Effendi and S. A. Ali, CoRR, vol. abs/1701.08744, 2017.
4. “Machine Learning in the Real World,” V. Chaoji, R. Rastogi, and G. Roy, Proc. VLDB Endow., Vol. 9, No. 13, 2016.
5. “Feature extraction for regression problems and an example application for pose estimation of a face,” N. Kwak, S.-I. Choi, and C.-H. Choi, in International Conference Image Analysis and Recognition, pp. 435–444. 2008.
6. “Ad click prediction: a view from the trenches,” H. B. McMahan et al., in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining,, pp. 1222–1230,2013.
7. “Predicting clicks: estimating the click- through rate for new ads,” M. Richardson, E. Dominowska, and R. Ragno, in Proceedings of the 16th international conference on the World Wide Web, pp. 521–530, 2007.
8. “Linear Regression,” A. Ng, CS229 Lect. Notes, vol. 1, no. 1, pp. 1–3, 2000. [1]
9. https://seaborn.pydata.org/
10. https://github.com/microsoft/LightGBM [2]

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details