Online ad click through rate prediction using machine learning

Thoyaja Kanala
700747744
dept.Computer Science
University of Central Missouri
txk77440@ucmo.edu

Swathi Kasthuri
700747121
dept.Computer Science
University of Central Missouri
sxk71210@ucmo.edu

Ashok Bandi
700741044
dept.Computer Science
University of Central Missouri
axb10440@ucmo.edu

***Abstract*—Digital advertising industry made revenue of 31.7 billion dollars in the fiscal year of 2016. Main challenges in this industry or suggesting or recommending the right advertisement to the right people by predicting the click through rate of the advertisement.Data to predict the click through rate contains diverse data: few columns are categorical, few columns are identifiers which throws a challenge while classifying the data. Dataset is complex to use normal classification algorithms like SVM, logistic regression.The dataset file size is 1.28 GB which is considered to be the huge dataset to process. To process the complex dataset and to produce the results we are using the LightGBM classifier which uses high grade boosting methods and multiple decision trees in the learning stage. Main advantage of this algorithm is it also supports the GPU training.In addition to LightGBM classifier XGBoost algorithm also implemented to compare the performance on the dataset.To conduct the experimental analysis we collected the Avazu dataset from kaggle. Both the algorithms achieved 82 percent accuracy but the training time of LightGBM is less compared to Gradient boosting algorithm.**

[1] ***Index Terms*—CTR(Click through Rate), LightGBM(Light Gradient Boosting Machine), Gradient boosting, GPU(Graphics Processing Unit) and Avazu dataset.**

I. Introduction

Online click through rate is a gauge to measure how well the ads reach out to the people who are relevant. Usually this click through rate is calculated as follows: clicks/number of impressions for example if the number of clicks is 5 and the impressions is 100 then the click through rate will be 5 percent.

In this paper XGBoost algorithm is used to classify the binary classes into whether the user will click or not. Different feature engineering techniques are used to pick the best parameters from the dataset. XGBoost is fine tuned before selecting the best features. Exploratory Data Analysis(EDA) will show the important features rather than novel approaches. For this experiment Avazu dataset is used which is publicly available in the kaggle. Avazu is a rapidly growing multinational company. In this experiment we have used the 11 days of the Avazu work. In this work we predict whether the user will click on the mobile ad or not. In the digital advertising industry calculating the CTR for a website or advertisement is essential. Based on the CTR factor advertisements revenue is ranked. In this paper we have used machine learning to divide the dataset into train and test sets. In the dataset we have 10 categories of advertisements each company had labels as 0's and 1's here 0 means not a click and 1 means click. Based on the training parameters given, test data is classified into 2 categories whether the user will click or not. On the other hand reinforcement learning will have a negative score when the model guessing is wrong. Experiments are conducted using both the methods and compared.

Online tech giants like Google, Facebook etc., are investing in RD and concluded that most of the revenue is generated from online ads. High CTR for an advertisement is achieved by recommending the suitable categories and items. But there are few challenges in this system. One is if the user gets the similar content most of the time he loses interest in the content and in turn the rate of CTR drops. So balance of accuracy and diversity in the recommendations is important. So this is the field that demands in depth exploration. In the NLP tasks each time a new algorithm will emerge that is making the researchers explore more content in this field. This concept can be integrated with a recommendation system to yield good results as mentioned above to achieve accuracy and diversity we need to implement the multiple machine learning algorithms and a comparative study suggests which model needs to be used for future predictions.

The project flow is divided into 4 parts: one is collecting the data, feature selection, classification and evaluation. Contributions: In this project we explore various graphical representation of data using matplotlib and seaborn We use scikit-learn for model creation, training and testing Pandas and numpy libraries are used to analyze the data Python programming language is used to develop the project in the PyCharm IDE

Evaluation: In the evaluation part we analyze the model performance in various aspects of the performance measure. Accuracy measures the overall predicted performance of positive to the negative class. precision : True positive/ True positive false positive Recall: True positive/ True positive false negative F1 score: harmonic mean of precision and recall Decision surface: Decision surface of the models visualizes the separability of the classes

[1] https://github.com/txk77440/MLproject

II. Motivation

Digital advertising industry made revenue of 31.7 billion dollars in the fiscal year of 2016. Main challenges in this industry or suggesting or recommending the right advertisement to the right people by predicting the click through rate of the advertisement.

Data to predict the click through rate contains diverse data: few columns are categorical, few columns are identifiers which throws a challenge while classifying the data. Dataset is complex to use normal classification algorithms like SVM, logistic regression.

III. Objectives

The main objectives of the project are:

Simplifying the complexity of the dataset using exploratory data analysis. Few columns in the dataset are anonymous we need to explore the dataset using visualization Implementing the LightBGM classifier for binary classification of data

Main features of the project are:

The dataset file size is 1.28 GB which is considered to be the huge dataset to process. To process the complex dataset and to produce the results we are using the LightGBM classifier which uses high grade boosting methods and multiple decision trees in the learning stage. Main advantage of this algorithm is it also supports the GPU training. In addition to LightGBM classifier XGBoost algorithm also implemented to compare the performance on the dataset

IV. Related work

Online advertisements have a huge market. Google 's revenue is touching billions of dollars only with the advertisements. The existing models implemented the various machine learning models to predict the click through rate. But the accuracy of the model is not satisfactory. In this paper we are proposing a model which has improved accuracy. In this paper we are proposing a self coded Neural Network architecture. The project implementation is divided into 2 stages one is opinion mining of the user's behaviour and the other phase is predicting the click through rate whether the user clicks the ad or not [13].

In this paper we are proposing a novel algorithm to predict the click through rate. Most of the algorithms in the existing systems are supervised machine learning algorithms or unsupervised machine learning algorithms. In this paper we are proposing a reinforcement learning method to predict the click through rate of the ads. Most of the company's revenue is generated by embedding the ads on the third party websites. For the advertisement auction click through rate prediction is necessary to rank the ads. In this paper we are proposing the Upper Bound Confidence algorithm and Thompson Sampling algorithms to predict the click through rate [7].

In recent times the sales of the products increased mostly by the advertisements. The click through rate is calculated by the number of clicks for an ad divided by the total number of views. This click through rate decides which ad should be recommended to which user. In the existing system we have two kinds of prediction models one is shallow learning machine learning models and the other one is deep learning models. In this study we discuss the pros and cons of the methods [15].

Multiple machine learning algorithms are proposed in the existing work. The proposed algorithms have limitations. In the proposed algorithms the structure of the data is linear and it supports the linear data. But in general the nature of the data would be nonlinear and complex. The existing algorithms can not handle the complex data. In this paper we are proposing the Attention Network based architecture to predict the click through rate of the advertisements. In this paper CAN algorithm is proposed which gives the benefit of tuning the DNN(Deep Neural Network) . These can be tuned and the above mentioned problems will be solved using this architecture [3].

For the companies most of the revenue is generated through the advertisements. If a right advertisement hits the right person then the purchase of the products or sales of the business will be improved. To recommend the right ad for the right person, finding the click through rate is important. There are multiple factors that affect the click through rate of the ad so in this paper we are proposing a model to find the click through rate of the ads using demographic information and personal information [8].

Click through rate plays a pivotal role in ranking the advertisement system. Existing models use the deep factorization methods which use the mapping of the important features. The challenging part in click through rate prediction is the model has to predict the click through rate with limited data in limited time. To perform this task we are proposing a combination of a network of factorization models and the Boosting algorithms. Exploratory data analysis conducted to find the important features and to eliminate the redundant features from the data [2].

Click through rate is an important factor in ranking the advertisement systems. eCommerce websites mainly run based on the advertisements. These advertisements are recommended based on the user history and past searches. In this paper we are proposing the Deep Interest Network to predict the click through rate of the ads. The traditional algorithms use the linear features to map or predict the click through rate. In the proposed models we use the model to learn the non linear features also [4].

For any online business to know the right kind of recommendations are necessary. In this paper we analysed the linear algorithms to predict the click through rate of the ad. In the existing system the proposed algorithms do not show considerable efficiency. So in this study we implemented multiple linear algorithms and the factors we have selected to predict the click through rate are the site which we clicked the ad, domain of the advertisement and the app id. These are the important factors that affect the click through rate of the advertisement [1].

Online advertising industry is a huge industry that generates the maximum amount of revenue for businesses. In the existing

system the algorithms used are the basic and give monotonous results. The main challenge of the project is that we need to get accurate results in stipulated time with limited data. In this study we are proposing the state of the art FFM(Field Factorization Machine) algorithm to predict the click through rate we have optimised the algorithm using Particle Swarm Optimisation (PSO) technique [10].

Predicting the click through rate depends on the multiple factors like the device he/she is using, what kind of domain they are looking for, ad placement etc. The click through rate of any advertisement is calculated by the number of impressions of the ad and number of clicks. Here the number of impressions suggest the number of views. To conduct the experimental analysis we have used the Avazu dataset that is collected from Kaggle open source library. In this study we are proposing deep and wide neural networks. To evaluate the model we have plotted the ROC-AUC curve [6] [19].

The existing systems predicted the click through rate using the demographic parameters and site information. But in this paper we recommend the advertisements based on the similar user interests. The project is divided into 4 stages, one is extracting the features from the similar users and then we extract the latent features of the users and then segment the groups. Based on the segmented groups we recommend the ads then we calculate the click through rate.To perform this task we deployed the Deep User Segmentation Interest Network [9].

Click through rate in simple terms is defined as the probability of the ad to be clicked. In this paper we discuss the limitations of the existing models. Existing deep learning models calculate the interactions of the features using dot products the main disadvantage of this is poor interpretability. To solve this problem we implement the density matrix which improves the interpretability of the model. To implement the model we have selected the Avazu and Criteo dataset [17].

Click through rate prediction datasets contain high dimensional diverse datasets. To achieve high accuracy and to reduce the dimensions one of the possible solutions is feature selection. But the feature selection method works for the moderate amount of variables if the dataset contains a high number of variables the procedure becomes complex. So in this study we propose the Optimally Connected Deep Belief Net(OCDBN) model. In this method the redundant features are removed using the optimal mean removal method which improves the accuracy of the model and reduces the complexity of selecting the features [16] [12].

Click through rate prediction is an important factor to consider in the digital era. In the existing system the click through rate of the advertisement is predicted using images or text. In this paper we are proposing a model to predict the click through rate using banners. In this study we consider all the 3 features to predict the click through rate of an ad. To extract the feature we used deep neural networks [14].

Finding the likelihood of the click through rate of the advertisement is pivotal for digital communications. The data in general comes in large sizes and is diverse. Traditional machine learning offers feature selection methods to select the important features but these methods are time consuming and complex. To solve the above problems we are implementing the Mutual Information and feature interaction methods. To classify the CTR we have implemented the Deep Neural Network(DNN). We conduct the experimental analysis on 4 benchmark datasets [18] [5].

First, we designed a user-level distributed factorization machine that only uses the gradient information of each client (instead of the original preference data) to update the model. Secondly, in order to solve the linear aggregation model loss caused by the heterogeneity of user data, we introduced the idea of cluster federation learning in the factorization machine, and designed the framework, FedDeepFM. Finally, we implemented the prototype of FedDeepFM and evaluated it with a real ad click data set. Experimental results show that the accuracy of our proposed framework is 8 percent higher than that of the traditional federated matrix factorization algorithm, and 2.5 percent higher than the federated learning accuracy of a single global model [11].

Click through rate prediction is an important factor to consider in the digital era. In the existing system the click through rate of the advertisement is predicted using images or text. Online tech giants like Google, Facebook etc., are investing in RD and concluded that most of the revenue is generated from online ads. High CTR for an advertisement is achieved by recommending the suitable categories and items. But there are few challenges in this system. One is if the user gets the similar content most of the time he loses interest in the content and in turn the rate of CTR drops. So balance of accuracy and diversity in the recommendations is important. So this is the field that demands in depth exploration. In the NLP tasks each time a new algorithm will emerge that is making the researchers explore more content in this field.This concept can be integrated with a recommendation system to yield good results as mentioned above to achieve accuracy and diversity we need to implement the multiple machine learning algorithms and a comparative study suggests which model needs to be used for future predictions [20].

V. Data Description

Dataset is collected from the open source repository Kaggle. The dataset contains two files: train and test files. Train dataset contains 10 days of records which includes clicks and non clicks arranged in chronological order. Test file contains 1 day clicks and non clicks.

Total number of records in the dataset is 4 million records. The dataset contains two folders one train set and the other one is test set. The dataset original source is Avazu. Avazu, Inc. operates as a digital advertising company. The Company focuses on offering advertisers with solutions for real-time and data-driven display advertising across various advertisement exchanges and sell-side platforms worldwide.

- train - Training set. 10 days of click-through data, ordered chronologically. Non-clicks and clicks are sub-sampled according to different strategies.
- Test set. 1 day of ads to for testing your model predictions

Attribute description:

- id: ad identifier
- click: 0/1 for non-click/click
- hour: format is YYYYMMDDHH, so 14091123 means 23:00 on Sept. 11, 2014 UTC.
- C1 — anonymized categorical variable
- bannerpos
- site id]
- site domain
- site category
- app id
- app domain
- app category
- device id
- device ip
- device model
- device type
- device conn type
- C14-C21 — anonymized categorical variables

There are no null values in the data. The dataset contains 24 columns including one target column. There are 9 object columns in the data and 14 integer columns and 1 float data type column.

VI. Proposed Framework

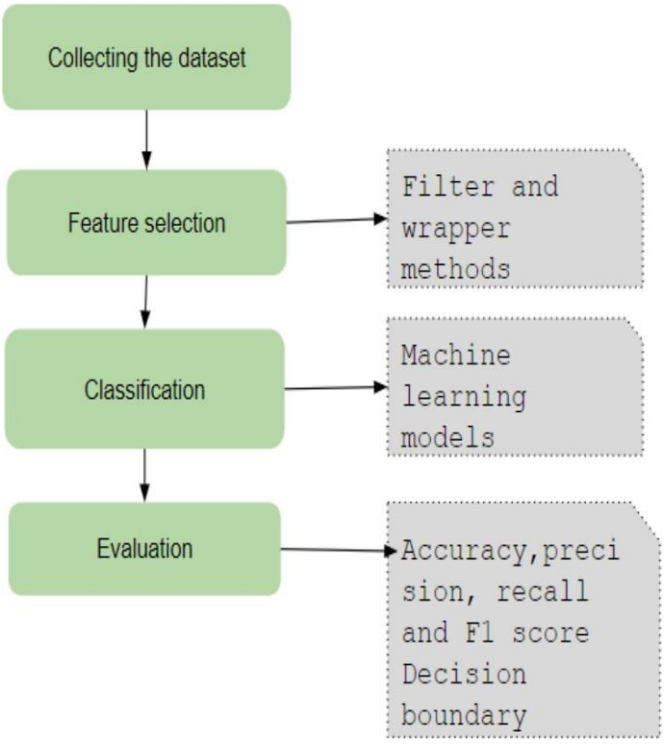

Figure 1. Workflow

This project implementation is divided into following steps:

- Data collection: data is collected from Kaggle.
- Data pre-processing: In this step we have removed the null values in the dataset and duplicates also.In the Exploratory data analysis we analyzed the underlying structure of the data.
- Data preparation : In the data preparation step label encoded the categorical values in the data and the data is split into training and testing
- Since the dataset contains 4 million records, training time is so long so we used the subset of dataset i.e. 10000 samples to train the data.In the last step we have implemented the 2 algorithms LightGBM and XGBoost algorithm.
- In the implementation step of machine learning algorithms XGBoost is implemented using scikit-learn library and LightGBM is lightgbm library. Both the algorithms achieved reasonable accuracy of 82 percent on test data

A. Methods

In the fig.1. Shows the implementation of the project. That is broken down into 4 stages, data collection, data pre-processing, LightGBM classifier and XGBoost classifier. In the fig.2. Shows the growth of nodes in XGBoost and LightGBM algorithms leafwise and levelwise.

XGBoost follows a greedy algorithm when building the tree structure. It grows levelwise. It uses regularizer parameter

to penalize while building the complex structures. It takes negative To optimize the performance of the loss fusion this algorithm takes into account negative gradients also. The advantages of Light Gradient Boosting Machine are: Less memory usage and faster training time. This algorithm grows leafwise.

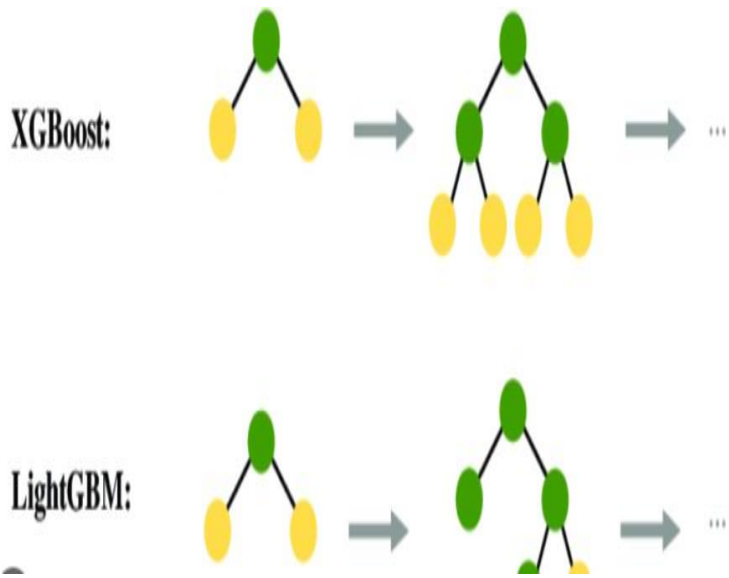

Figure 2. LightGBM vs XGBoost

B. Exploratory Data Analysis

In the exploratory data analysis we have observed the underlying structure of the data. To understand the data better we have visualized the various graphs. To visualize the data distribution and patterns we used the matplotlib and seaborn.In the first graph the distribution of targets using bar plot. The distribution plot shows the targets balance. Here we can observe that higher number of sample present for the category of not clicked and less number of samples for clicked data. Dataset is imbalanced. In the second figure we visualized the trends of clicks per hour i.e. hourly based clicks using a time series plot. The plot contains peaks and low points as the day progresses the clicks improves.

In the third plot we visualized the density plot using trends of clicks by hours in a day and count plot of hourly CTR i.e. for 24 hours. From the visualization we can conclude that the clicks are high for mid day. In the fourth graph we visualized the distribution of clicks on hourly basis separated by targets. The number of clicks are high and non clicks are low. In the final graph we represented the banner position distribution using bar plots. The banner positions are high for clicked advertisements.

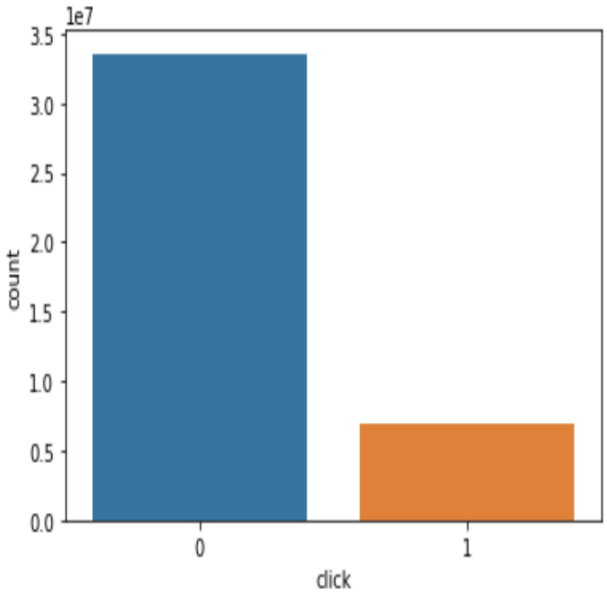

Figure 3. Distribution of targets

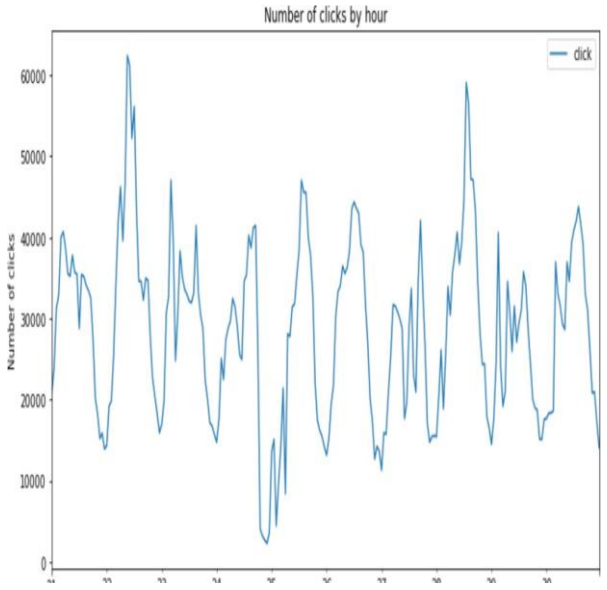

Figure 4. Trends of clicks per hour

VII. Results Analysis

A. LightGBM classifier

The size of the dataset is high and very complex to process. If the dataset size is high and complex traditional machine learning algorithms take high time to process the data and trainig time is also high. With the recent studies we can say that Boosting algorithms solve these problems but the problem is that they consume so much memory. The variant of Boosting methods LightGBM supports GPU when training the large datasets with accurate results.

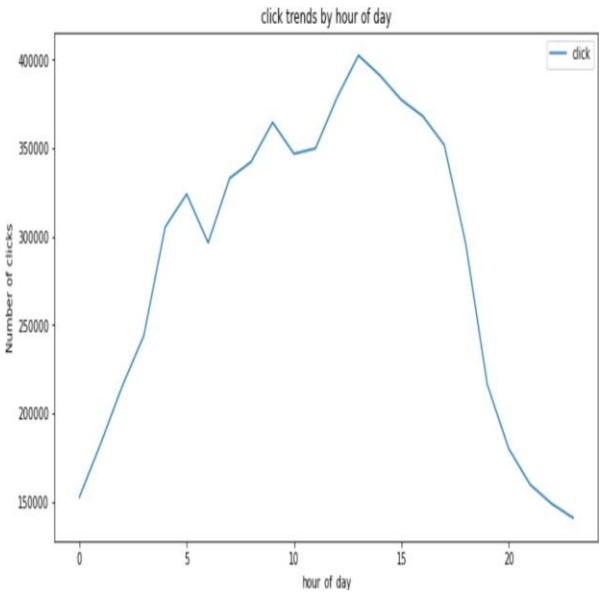

Figure 5. Trend of clicks by hour of the day

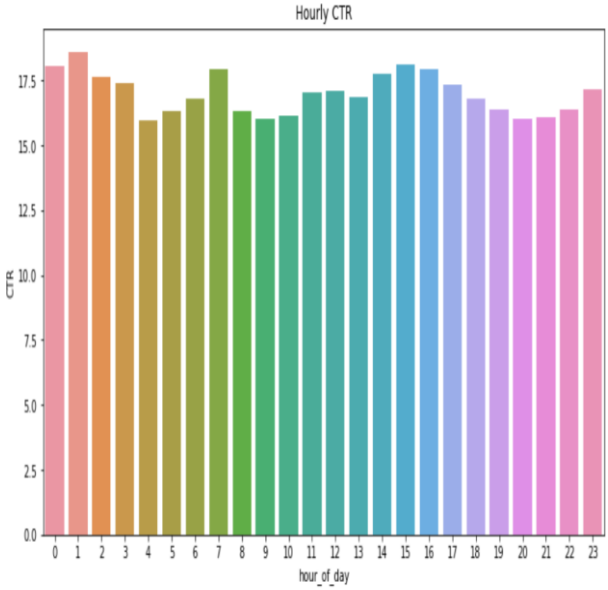

Figure 7. Hourly CTR

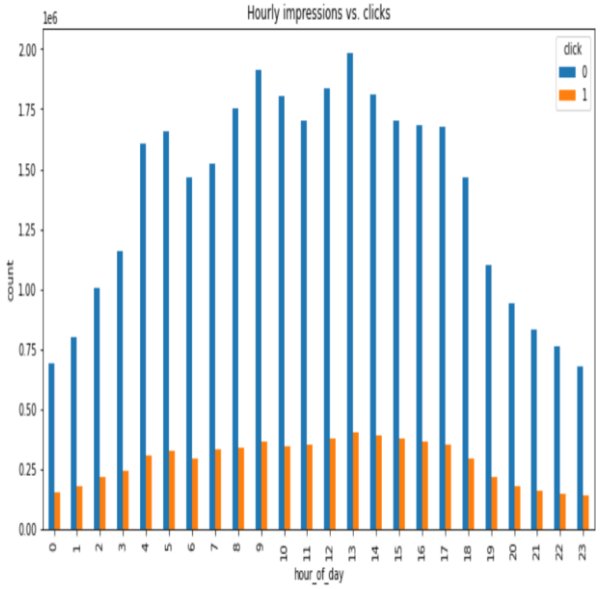

Figure 6. Countplot clicks per hour

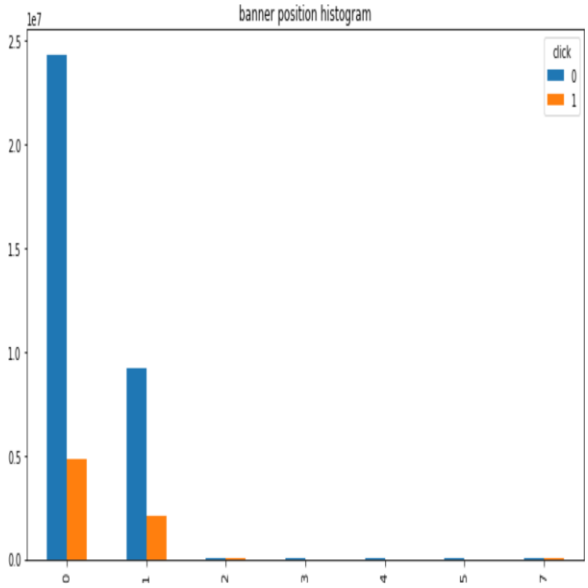

Figure 8. Banner position distribution

B. Boosting methods

Boosting methods are the ensemble methods which uses the boosting techniques in the project we have implemented the Gradient Boosting algorithm to compare with the Light-GBM classifier performance. Gradient boosting algorithm is implemented using scikit-learn library. The model achieved 82 percent accuracy.

VIII. Results summary

In this paper we have implemented two algorithms one is LightGBM and Gradient Boosting algorithm. These Gradient boosting algorithms is implemented from scikit-learn library and LightGBM is implemented using lightgbm library. These models implementation is done in Python programming language. Each object is created from the respective libraries and trained the models using training set. To evaluate the performance of the models on the test set various performance metrics are used. To evaluate the true positives and true negatives we have constructed the classification report. The

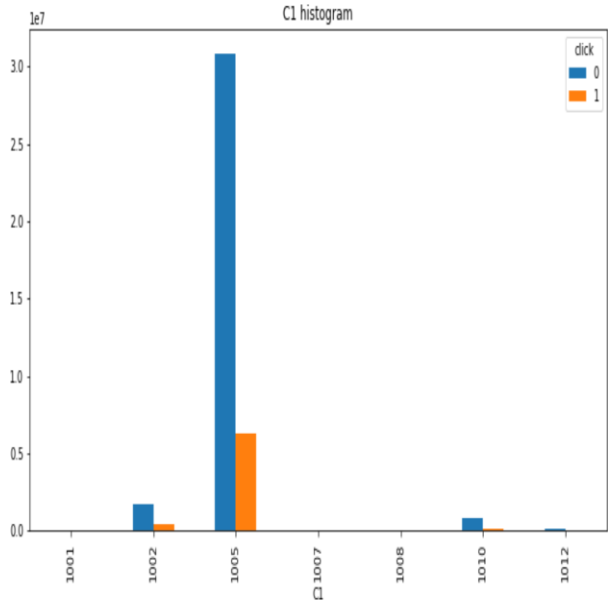

Figure 9. Anonymised C1 column distribution

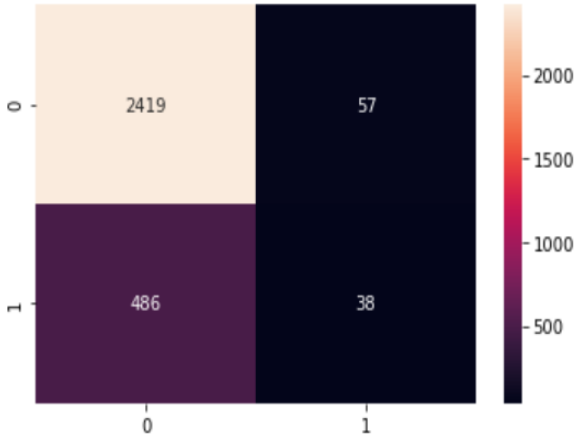

Figure 10. LightGBM confusion matrix

classification report gives the summary of precision, recall and F1 score. To further evaluate the model performance confusion matrix is constructed. The confusion matrix gives the incorrectly predicted sample counts and correctly classified sample count. Gradient boosting algorithm and LightGBM performed similarly on the test data that is with 82 percent of accuracy. If we look at the precision, recall and F1 score of both the algorithms: The three parameters are high for non click category in case of Light GBM and in case of gradient boosting it is high for non clicked category.

	precision	recall	f1-score	support
0	0.83	0.98	0.90	2476
1	0.40	0.07	0.12	524
accuracy			0.82	3000
macro avg	0.62	0.52	0.51	3000
weighted avg	0.76	0.82	0.76	3000

Figure 11. Classification report LightGBM

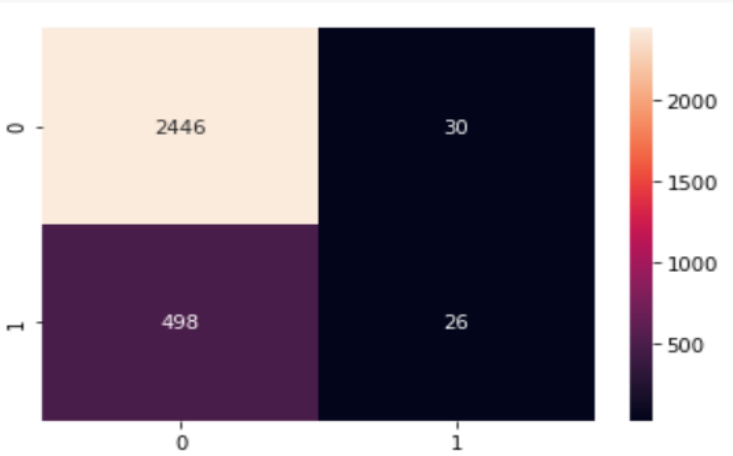

Figure 12. Confusion matrix Gradient Boosting

	precision	recall	f1-score	support
0	0.83	0.99	0.90	2476
1	0.46	0.05	0.09	524
accuracy			0.82	3000
macro avg	0.65	0.52	0.50	3000
weighted avg	0.77	0.82	0.76	3000

Figure 13. Classification report gradient boosting

REFERENCES

[1] Antriksh Agarwal, Avishkar Gupta, and Tanvir Ahmad. A comparative study of linear learning methods in click-through rate prediction. In

2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI), pages 97–102, 2015.

[2] Mohamadreza Bakhtyari and Saye Mirzaei. Click-through rate prediction using feature engineered boosting algorithms. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–5, 2021.

[3] Wenjie Cai, Yufeng Wang, Jianhua Ma, and Qun Jin. Can: Effective cross features by global attention mechanism and neural network for ad click prediction. *Tsinghua Science and Technology*, 27(1):186–195, 2022.

[4] Sun Di. Deep interest network for taobao advertising data click-through rate prediction. In *2021 International Conference on Communications, Information System and Computer Engineering (CISCE)*, pages 741–744, 2021.

[5] Nathaniel Hudson, Hana Khamfroush, Brent Harrison, and Adam Craig. Smart advertisement for maximal clicks in online social networks without user data. In *2020 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 172–179, 2020.

[6] Kyungwon Kim, Eun Kwon, and Jaram Park. Deep user segment interest network modeling for click-through rate prediction of online advertising. *IEEE Access*, 9:9812–9821, 2021.

[7] A. Lakshmanarao, S. Ushanag, and B. Sundara Leela. Ad prediction using click through rate and machine learning with reinforcement learning. In *2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–5, 2021.

[8] Jing Ma, Xian Chen, Yueming Lu, and Kuo Zhang. A click-through rate prediction model and its applications to sponsored search advertising. In *International Conference on Cyberspace Technology (CCT 2013)*, pages 500–503, 2013.

[9] Tianyuan Niu and Yuexian Hou. Density matrix based convolutional neural network for click-through rate prediction. In *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 46–50, 2020.

[10] Michael Reynaldo Phangtriastu and Sani Muhamad Isa. Optimizing field-aware factorization machine with particle swarm optimization on online ads click-through rate prediction. In *2018 3rd International Conference on Computer and Communication Systems (ICCCS)*, pages 1–5, 2018.

[11] Xianshan Qu, Li Li, Xi Liu, Rui Chen, Yong Ge, and Soo-Hyun Choi. A dynamic neural network model for click-through rate prediction in real-time bidding. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1887–1896, 2019.

[12] Nadir Sahllal and El Mamoun Souidi. A comparative analysis of sampling techniques for click-through rate prediction in native advertising. *IEEE Access*, 11:24511–24526, 2023.

[13] Lan Shan. A study on interest evolution-based click-through rate intelligent prediction model for advertising. In *2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI)*, pages 788–791, 2022.

[14] Xiaowei Wang, Hongbin Dong, and Shuang Han. Click-through rate prediction combining mutual information feature weighting and feature interaction. *IEEE Access*, 8:207216–207225, 2020.

[15] Xinfei Wang. A survey of online advertising click-through rate prediction models. In *2020 IEEE International Conference on Information Technology,Big Data and Artificial Intelligence (ICIBA)*, volume 1, pages 516–521, 2020.

[16] Bohui Xia, Xueting Wang, Toshihiko Yamasaki, Kiyoharu Aizawa, and Hiroyuki Seshime. Deep neural network-based click-through rate prediction using multimodal features of online banners. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 162–170, 2019.

[17] Rongbin Xu, Menglong Wang, and Ying Xie. Optimally connected deep belief net for click through rate prediction in online advertising. *IEEE Access*, 6:43009–43020, 2018.

[18] Chenjia Yu, Shuhan Qi, and Yang Liu. Feddeepfm: Ad ctr prediction based on federated factorization machine. In *2021 IEEE Sixth International Conference on Data Science in Cyberspace (DSC)*, pages 195–202, 2021.

[19] Sen Zhang, Zheng Liu, and Wendong Xiao. A hierarchical extreme learning machine algorithm for advertisement click-through rate prediction. *IEEE Access*, 6:50641–50647, 2018.

[20] Weijie Zhao, Peng Yang, Dong Li, Xing Shen, Lin Liu, and Ping Li. Feature fusion network for personalized online advertising systems. In *2022 IEEE International Conference on Big Data (Big Data)*, pages2160–2168, 2022.