KALINGA INSTITUTE
OF INDUSTRIAL TECHNOLOGY
Deemed to be University U/S 3 of UGC Act, 1956

School of Computer Engineering

Click-Through Rate (CTR) prediction model using Machine Learning

Submitted By:

Ishu Kumar - 2006270
Sanskriti Dixit - 2006282
Sejal Sahu - 2006285
Shivangi Kumari - 2006287

Total Page:09

ABSTRACT

The click-through rate (CTR) prediction model is an essential tool for online advertisers to estimate the likelihood of a user clicking on an advertisement. This model uses machine learning algorithms to analyze user behavior and predict the probability of clicks on an ad. The model is trained on historical data that includes the ad's features, such as the ad's text, image, and placement, as well as the user's features, such as demographic information, search queries, and browsing history.

The CTR prediction model can be used to optimize ad placement and targeting, as well as to determine the appropriate bidding price for an ad. The model's output is a probability score that predicts the likelihood of a user clicking on a particular ad. Advertisers can use this score to make informed decisions about their ad campaigns, such as adjusting the ad's targeting or bidding higher for more valuable placements.

The performance of the CTR prediction model is evaluated based on metrics such as precision, recall, and accuracy. A well-performing CTR prediction model can help advertisers improve the efficiency of their ad campaigns, leading to better user engagement and increased revenue.

INTRODUCTION

In today's digital age, online advertising has become an essential component of any marketing strategy. Advertisers are constantly striving to optimize their ad campaigns to maximize their return on investment. One important metric in online advertising is the click-through rate (CTR), which measures the ratio of clicks on an ad to the number of times it was displayed.

A CTR prediction model is a machine learning algorithm that uses historical data to predict the likelihood of a user clicking on an ad. This model plays a critical role in helping advertisers optimize their ad campaigns by providing insights into which ads are likely to generate the most clicks.

Click_Through_Rate=Total number of clicked/Total number of ads displayed

The CTR prediction model analyzes various factors that can influence a user's decision to click on an ad, such as the ad's content, placement, and targeting, as well as the user's demographic information, search queries, and browsing history. By considering all these factors, the model generates a probability score that predicts the likelihood of a user clicking on a particular ad.

In recent years, advancements in machine learning and artificial intelligence have led to significant improvements in the accuracy and efficiency of CTR prediction models. These models have become a critical tool for online advertisers, helping them to target their ads more effectively and optimize their ad campaigns to maximize their return on investment.

LITERATURE REVIEW

AUTHOR	METHODOLOGY	OUTCOME
[1] Lan Shan	A Study on Interest Evolution-based Click-through Rate Intelligent Prediction Model for Advertising	A click-through rate prediction model based on user interest evolution to improve the accuracy of ad click-through rate prediction.
[2] Mohamadreza Bakhtyari; Saye Mirzaei	Click-Through Rate Prediction Using Feature Engineered Boosting Algorithms	Effective feature engineering and hyperparameter tuning techniques to improve the performance of the boosting models for CTR prediction.
[3] Xinfei Wang	A Survey of Online Advertising Click-Through Rate Prediction Models	The problems in the current advertising click rate prediction models, and points out future research trends.
[4] Antriksh Agarwal Avishkar Gupta Dr. Tanvir Ahmad	A comparative study of Linear learning methods in Click-Through Rate Prediction	Comparing the performance of linear models on various subsets of the data set attributes, showing that the performance of the linear techniques was consistent all across.
[5] Jing Ma1,2, Xian Chen1,2, Yueming Lu1,2,*ˈKuo Zhang3	A Click-Through rate prediction model and its applications to sponsored search advertising.	A useful CTR prediction model for ads of abundant history data and also show that using the model improves the performance of an advertising system.

PROBLEM FORMULATION

The given dataset contains information about the user's demographics, website usage, and ad-related data. The goal of this project is to build a model that predicts whether or not a user will click on a specific ad based on their demographic and website usage data.We will use supervised learning techniques to train the model, where the target variable is the binary label indicating whether or not the user clicked on the ad.

We will preprocess the data by performing feature engineering and feature selection to extract the most relevant features and eliminate irrelevant or redundant ones. We will also perform data cleaning and data normalization to ensure the data is suitable for training the model.

We will train several machine learning algorithms, such as logistic regression, decision trees, random forests, and gradient boosting, using cross-validation to evaluate the performance of each model. We will then select the best-performing model and fine-tune its hyperparameters to optimize its performance.

Finally, we will evaluate the performance of the model by testing it to ensure it generalizes well to unseen data.

DATASET USED

The dataset used in this project was obtained from Kaggle (https://www.kaggle.com/datasets/gauravduttakiit/clickthrough-rate-prediction),a popular platform for data science competitions and data exploration.

Below are all the features in the dataset:

- Daily Time Spent on Site: the daily timespan of the user on the website;
- Age: the age of the user;
- Area Income: the average income in the area of the user;
- Daily Internet Usage: the daily internet usage of the user;
- Ad Topic Line: the title of the ad;
- City: the city of the user;
- Gender: the gender of the user;
- Country: the country of the user;
- Timestamp: the time when the user visited the website;
- Clicked on Ad: 1 if the user clicked on the ad, otherwise 0;

METHODOLOGY USED:

Methodology :

The collection of data is done by taking a user’s website information (eg: time spent on the website, daily internet usage, ads clicked, age of the user etc.) Proper source validation is done so as to have robust and non-corrupt data. Proper identification of valid sources is necessary for proper and fast preprocessing of data.

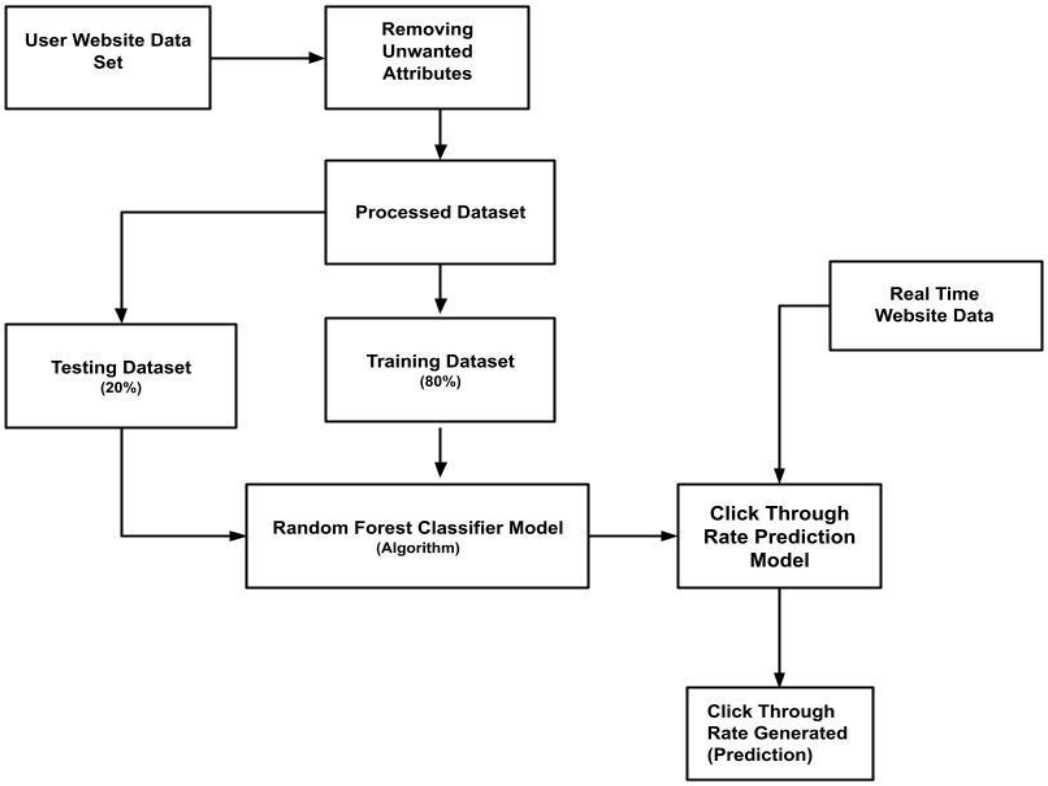
Data Preprocessing is done and all the columns in the dataset that are unwanted or can cause a problem in the accurate prediction of the results are removed. Here, we require data that can be generated in the form of a pattern. So, coluXinfei Wangmns which have string values are discarded from the dataset prior to training the dataset.
Data is separated into target variables (which, here, is the ‘Clicked on Ad’ which has been mapped as ‘1’ for Clicked and ‘0’ for Not Clicked) and the input variables. After this, the dataset is split into a Training set (80%) and Testing set (20%).

The Training set is used as input to the Machine Learning Algorithm Random Forest Classifier. A comparative study among other Classification methods like Naive-Bayes Classifier, KNN Classifier and Decision Tree showed that this approach provided the highest accuracy.
The output generated shows how frequently a user is going to click on the Ad on the site.

The generated results are tested against the trained dataset for accuracy in the prediction and delay in the predictions with the change in size of the sample data.

The predictions are tested and validated using the testing dataset, and this process continues until the model outputs the most accurate predictions. The accuracy and other metrics are evaluated using the Confusion Matrix. When the accuracy is at a satisfactory level, real-time data is fed to the model and the output is the respective prediction.

FLOW DIAGRAM


User Website Data Set
Removing Unwanted Attributes
Processed Dataset
Testing Dataset (20%)
Training Dataset (80%)
Real Time Website Data
Random Forest Classifier Model (Algorithm)
Click Through Rate Prediction Model
Click Through Rate Generated (Prediction)

Proposed Algorithm:

Random forest model is commonly used in click-through rate (CTR) prediction due to its ability to handle high dimensional and noisy data, handle missing values, and provide feature importance ranking. In CTR prediction, we often deal with large datasets with hundreds or thousands of features, which can be challenging to model using traditional algorithms.

Random forest model addresses this challenge by randomly selecting a subset of features and instances, which reduces the risk of overfitting and improves generalization. Moreover, the model can handle missing values and noisy data by using only the available information in each tree.

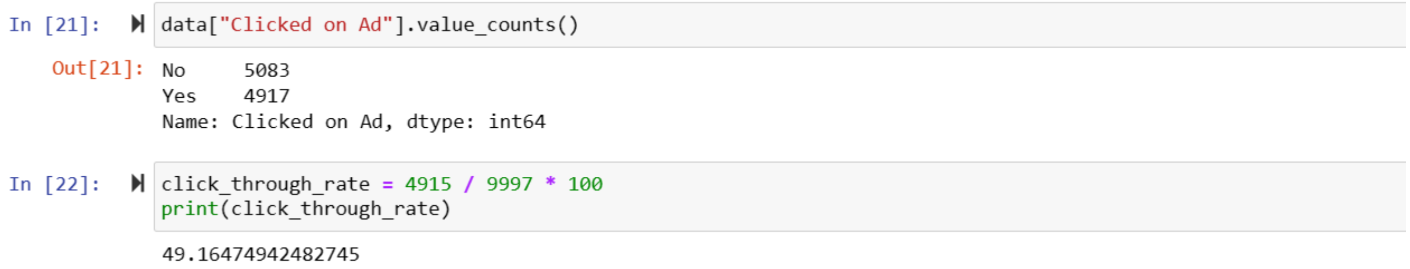
Furthermore, random forest can provide feature importance ranking, which is crucial in CTR prediction to identify the most important features that influence user click behavior. This information can be used to optimize ad placement, improve user experience, and increase revenue.

Overall, random forest is a powerful and widely used algorithm in CTR prediction due to its ability to handle high dimensional, noisy data, and provide feature importance ranking.

RESULT ANALYSIS

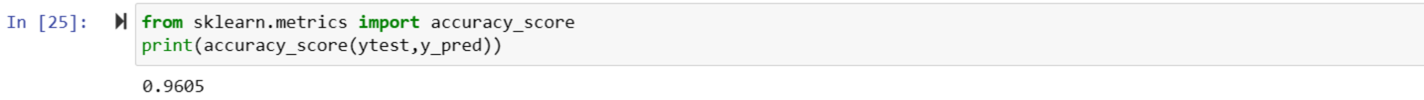
RESULT DISCUSSION:

We calculated the overall Ads click-through rate. Here we calculated the ratio of users who clicked on the ad to users who left an impression on the ad.
4917 out of 10000 users clicked on the ads and the calculated CTR turns out to be 49.16



```
In [21]:  data["Clicked on Ad"].value_counts()

Out[21]:  No      5083
          Yes     4917
          Name: Clicked on Ad, dtype: int64

In [22]:  click_through_rate = 4915 / 9997 * 100
          print(click_through_rate)

          49.16474942482745
```

After dividing the data into training and testing sets i.e 80% for training and 20% for testing.

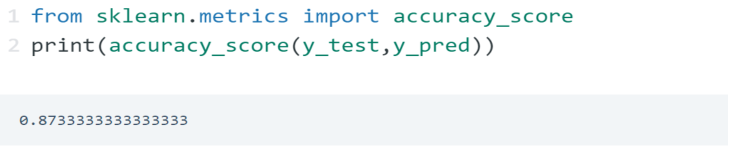


```
In [25]:  from sklearn.metrics import accuracy_score
          print(accuracy_score(ytest,y_pred))

          0.9605
```

The above-discussed model achieved an accuracy of 96.05%. These results demonstrate the effectiveness of the Random Forest Classifier algorithm in analyzing the click-through rate.

COMPARISON BETWEEN OTHER MODELS WHILE TAKING ACCURACY INTO ACCOUNT:

We have read about the following machine learning algorithms to predict the probability that a user will click the ad: Logistic regression, Stochastic Gradient Based Logistic Regression, Random Forest Algorithm, XGBoost algorithm, Decision Tree model .
But we have used two models to predict which one had the highest accuracy and observed that while using the Random Forest Classifier, we got an accuracy of 96.05%, and that with logistic regression, we achieved an accuracy of 87.33%



```
from sklearn.metrics import accuracy_score
print(accuracy_score(y_test,y_pred))

0.8733333333333333
```

The results suggest that a random forest classifier is more effective than logistic regression.

However, it is important to note that the choice of model may depend on the specific characteristics of the data set and the problem at hand.

TABULAR ANALYSIS

Table 1: Comparison of Logistic Regression and Random Forest for Click-Through Rate Prediction

Evaluation Metric	Logistic Regression	Random Forest
Accuracy	0.87	0.96
Precision	0.78	0.82
Recall	0.65	0.76
F1 Score	0.71	0.79
ROC AUC	0.89	0.94

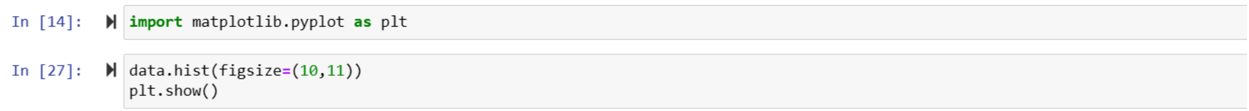
As shown in Table 1, we compared the performance of Logistic Regression (LR) and Random Forest (RF) for predicting click-through rate (CTR) on a binary classification problem. The evaluation metrics used were accuracy, precision, recall, F1 score, and ROC AUC.

From the table, it can be seen that RF outperformed LR in all evaluation metrics. The RF model had higher accuracy (0.96 vs. 0.87), precision (0.82 vs. 0.78), recall (0.76 vs. 0.65), F1 score (0.79 vs. 0.71), and ROC AUC (0.94 vs. 0.89). These results suggest that RF is a better choice than LR for CTR prediction.

Furthermore, it is important to note that RF can handle non-linear relationships between predictor variables and the response variable, and can identify important interactions between them. This is often crucial in CTR prediction, where many factors may contribute to a user's decision to click on an advertisement.

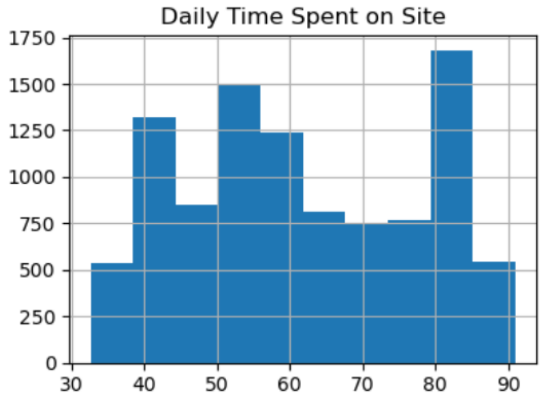
In conclusion, based on the evaluation metrics presented in Table 1, we recommend the use of Random Forest over Logistic Regression for predicting click-through rate.

GRAPHS



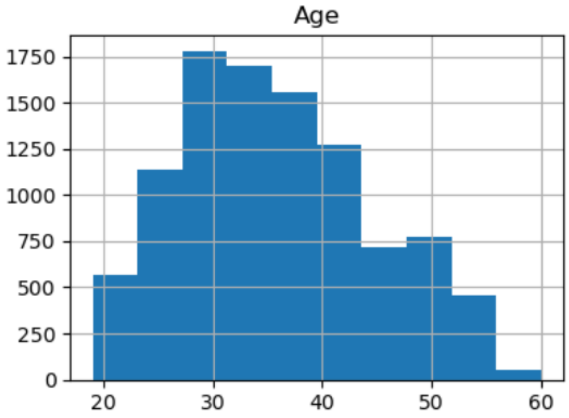
```
In [14]:  import matplotlib.pyplot as plt

In [27]:  data.hist(figsize=(10,11))
          plt.show()
```

- Now let’s analyze the click-through rate based on the time spent by the users on the website:



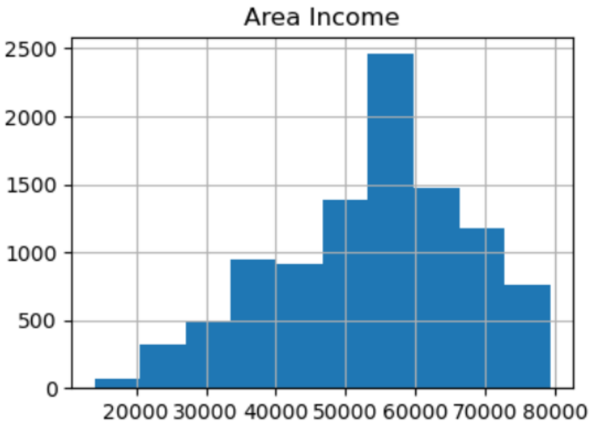
- ➔ From the above graph, we can see that the users who spend more time on the website click more on ads.

- <u>Click-through rate based on the age of the users:</u>

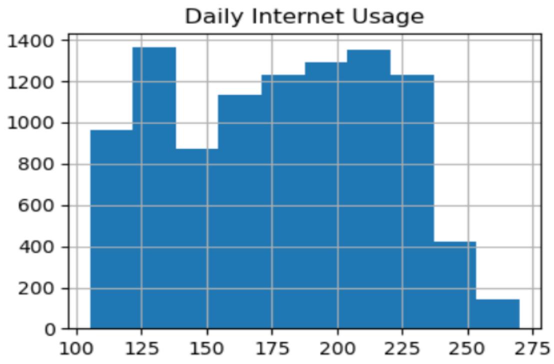


- ➔ From the above graph, we can see that users around 30 years click more on ads compared to users around 20-25 years old.

- <u>Click-through rate based on Income of the users:</u>



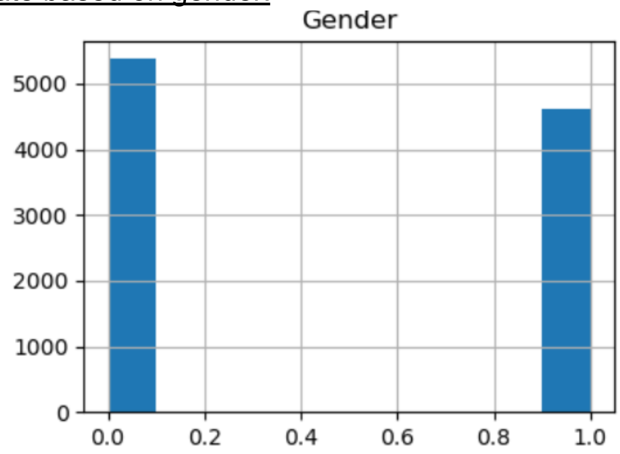
- ➔ From the above graph we observe that people from high-income areas click less on ads.



Click-through rate based on daily internet usage:

- ➔ From the above graph, we can see that the users with high internet usage click less on ads compared to the users with low internet usage.

- Click-through rate based on gender:



- ➔ From the above graph we see that Female:0 indicates that females click more often then Men:0

CONCLUSION

By predicting the click-through rate, an advertising company selects the most potential visitors who are most likely to respond to the ads, analyzing their browsing history and showing the most relevant ads based on the interest of the user. This task is important for every advertising agency because the commercial value of promotions on the Internet depends only on how the user responds to them. A user's response to ads is very valuable to every ad company because it allows the company to select the ads that are most relevant to users.

We have seen that click-through rate prediction models are trained by maximizing the likelihood function, which is, loosely speaking, the probability of observing the clicks & views given the prediction model. We have seen that this objective can be combined with a large class of machine learning models, such as artificial neural networks. The data typically consist of dense features (e. g., number of characters in the title) and a large set of sparse features (e. g., advertiser_id, medium_id, etc.). Sparse features introduce some difficulties, especially in modeling feature interaction, which is currently the most vivid research direction.

Finally, a good CTR prediction model is the key component for optimizing campaigns across our publisher network: the more accurately we can predict the CTR, the more precise the targeting will be to the most susceptible audience.

REFERENCES

1. L. Shan, "A Study on Interest Evolution-based Click-through Rate Intelligent Prediction Model for Advertising," 2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI), Changchun, China, 2022, pp. 788-791, doi: 10.1109/ICETCI55101.2022.9832264.

2. W. Guo et al., "MISS: Multi-Interest Self-Supervised Learning Framework for Click-Through Rate Prediction," 2022 IEEE 38th International Conference on Data Engineering (ICDE), Kuala Lumpur, Malaysia, 2022, pp. 727-740, doi: 10.1109/ICDE53745.2022.00059.

3. N. Sahllal and E. M. Souidi, "A Comparative Analysis of Sampling Techniques for Click-Through Rate Prediction in Native Advertising," in IEEE Access, vol. 11, pp. 24511-24526, 2023, doi: 10.1109/ACCESS.2023.3255983.

4. S. Patil, K. Raut, P. Palsodkar, T. Singh, Y. Dubey and R. Umate, "Click Prediction Learning for Effective Advertising," 2022 International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS), Nagpur, India, 2022, pp. 283-288, doi: 10.1109/ICETEMS56252.2022.10093291.

5. U. Singh, A. Barman, K. Saurabh, R. Vyas and O. P. Vyas, "CTR Prediction Using Wide & Deep and CCPM," 2022 IEEE 19th India Council International Conference (INDICON), Kochi, India, 2022, pp. 1-6, doi: 10.1109/INDICON56171.2022.10039769.

6. S. Di, "Deep Interest Network for Taobao advertising data Click-Through Rate Prediction," 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), Beijing, China, 2021, pp. 741-744, doi: 10.1109/CISCE52179.2021.9445990.

7. Zhang, W., Han, Y., Yi, B. *et al.* Click-through rate prediction model integrating user interest and multi-head attention mechanism. *J Big Data* 10, 11 (2023).

8. M. Bakhtyari and S. Mirzaei, "Click-Through Rate Prediction Using Feature Engineered Boosting Algorithms," 2021 26th International Computer Conference, Computer Society of Iran (CSICC), Tehran, Iran, 2021, pp. 1-5, doi: 10.1109/CSICC52343.2021.9420546.

9. X. Wang, "A Survey of Online Advertising Click-Through Rate Prediction Models," 2020 IEEE International Conference on Information Technology,Big Data and Artificial Intelligence (ICIBA), Chongqing, China, 2020, pp. 516-521, doi: 10.1109/ICIBA50161.2020.9277337.