ADVERTISEMENT CLICK-THROUGH RATE PREDICTION

[1] A.SudhaVali, M.Tech, Assistant Professor, Department of CSE, Eluru College of Engineering And Technology, Duggirala, Andhra Pradesh-534004.
[2] Bhargavi Venkata Nagalakshmi Dimmeta, B.Tech, Department of CSE, Eluru College of Engineering And Technology, Duggirala, Andhra Pradesh-534004.
[3] Mounika Nerusu, B.Tech, Department of CSE, Eluru College of Engineering And Technology, Duggirala, Andhra Pradesh-534004.
[4] Suvarna Agollu, B.Tech, Department of CSE, Eluru College of Engineering And Technology, Duggirala, Andhra Pradesh-534004.
[5] Girish Annepu, B.Tech, Department of CSE, Eluru College of Engineering And Technology, Duggirala, Andhra Pradesh-534004.

Abstract: Advertisement click-through rate prediction and recommendation system is proposed in this document. The system is composed of three main parts: data cleaning, prediction model training and front end recommendation system designing. One-hot approach is used to catch the useful features of the user data and advertisement data. After that, cleaned data is used to training the prediction model. In order to achieve as much as higher prediction accuracy, Deep FM algorithm is used after doing many literature review and comparison. Finally, a front-end system is designed for advertiser and advertising platform to use. The system not only predicts the CTR of a certain advertisement for advertiser but also offer the information and percentage of the target user who's predicted CTR is higher than the threshold to advertiser. The document describes the design procedure in details, including introduction, theoretical and literature review, hypothesis and goals of the system, methodology and concrete implementation of the predicted system and the result analysis and future scope of the projects.

1. INTRODUCTION

The Internet has promised a variety of online advertising forms leveraging many digital media vehicles (e.g., search portals, social media platforms, e-commerce platforms, online games, mobile apps, online videos, banners, etc.) to deliver marketing messages to potential consumers (Yang et al., 2017). Online advertising has become a dominant sector in the advertising industry. According to the Statista report (Statista, 2021), online advertising revenue in the United States grew by 12.2 percent in 2020 compared to 2019, from $124.6 billion to $139.8 billion. The online advertising market is expected to reach $982.82 billion by 2025 (Mordor Intelligence, 2021). In online advertising, click-based performance indexes, e.g., clicks and click-through rate (CTR) reflect the relevance of advertisements from users' perspective. As recognized by both researchers and practitioners, improving CTR is an effective way to realize the sustainable development of online advertising ecosystems (Robinson et al., 2007; Rosales et al., 2012; Tan et al., 2020). Hence, advertising CTR prediction has attracted much research attention in the past decades (Yan et al., 2014; Chapelle et al., 2014; McMahan et al., 2013; Richardson et al., 2007). As a hot research frontier driven by industrial needs, recent years have witnessed more and more novel learning models employed to improve advertising CTR prediction. Although extant research provides sufficient details on algorithmic design for addressing a variety of specific problems related to this topic, the methodological evolution and connections between modeling frameworks are precluded. However, to the best of our knowledge, there are few comprehensive surveys on CTR prediction in the context of online advertising. The objectives of our review are two-fold. First, we aim to make a systematic literature review on existing CTR prediction research, with a special focus on modeling frameworks. Second, we identify current research trends, main challenges and potential future directions worthy of further explorations. This review complements review articles recently published on users' responses (including CTR, conversion rate and user engagement) prediction (Gharibshah and Zhu, 2021) and CTR prediction (Wang, 2020; Zhang et al., 2021a). More specifically, Gharibshah and Zhu (2021) focused on online advertising platforms, data sources and features, and typical methods for user response prediction; Wang (2020) briefly introduced several classical methodsfor CTR prediction; and Zhang et al. (2021a) concentrated on the transfer from shallow to deep learning models for CTR prediction, explicit feature interaction modules and automated methods for architecture design. Our review is different from existing ones in the following aspects. First of all, we provide a comprehensive review on CTR prediction by organizing the development of extant research in a synthetic framework emphasizing connections between CTR prediction models. In particular, we give preliminaries on the problem of advertising CTR prediction, present stateof-the-art models, and outline major challenges and research perspectives in this area. For each category of CTR prediction models, we introduce the basic framework, its variants and ensemble models, and discuss advantages and disadvantages, and performance evaluation on various datasets. Second, thisreview focuses on CTR prediction modelsin the context of online advertising, excluding those in other contexts such as recommender system and Web search. Although Web search, recommender system and online advertising are

three popular information-seeking mechanisms to mitigate the information overload problem (Zhao et al., 2019), they are differ with respect to inputs, outputs and goals (see Appendix A.1). Third, this review covers existing research on CTR prediction almost without reservation, by searching over six major academic databases. In other words, this review contains a complete map of state-of-the-art and latest models proposed for advertising CTR prediction. The rest of this paper is organized as follows. Section 2 describes the search procedure for identifying articles covered in this review. Section 3 gives preliminary knowledge on the topic of advertising CTR prediction, including problem definition, foundational concepts, procedure, features and evaluation metrics. Section 4 gives a classification of state-of-the-art CTR prediction models in the extant literature and presents modeling frameworks, advantages and disadvantages, and performance assessment for CTR prediction. Section 5 summarizes CTR prediction models with respect to the complexity and the order of feature interactions, and performance evaluation on various datasets. Section 6 discusses current research trends, main challenges and future directions in the domain of advertising CTR prediction. Section 7 concludes this review.

2. LITERATURE SURVEY

2.1 R. Xu, M. Wang and Y. Xie, "Optimally Connected Deep Belief Net for Click Through Rate Prediction in Online Advertising,"in IEEE Access, vol. 6, pp. 43009-43020, 2018.

2.2 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.

2.3 Tan, Bin, et al. "Multi-task learning for click-through rate prediction." Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017.

2.4 Richard, Nicolas, et al. "Neural collaborative filtering." Proceedings of the 26th International Conference on World Wide Web. 2017.

2.5 Juan, Yuchin, et al. "Field-aware factorization machines for CTR prediction." Proceedings of the 10th ACM Conference on Recommender Systems. 2016.

2.6 He, Xinran, et al. "Practical lessons from predicting clicks on ads at Facebook." Proceedings of the Eighth International Workshop on Data Mining for Online Advertising. 2014.

2.7 Grbovic, Mihajlo, et al. "Scalable semantic matching of queries to ads in sponsored search advertising." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013.

2.8 McMahan, H. Brendan, et al. "Ad click prediction: a view from the trenches." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. 2013.

2.9 McMahan, H. Brendan, et al. "Follow-the-regularized-leader and mirror descent: Equivalence theorems and L1 regularization." Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. 2011.

2.10 Graepel, Thore, Joaquin Quinonero Candela, and Thomas Borchert. "Web-scale Bayesian click through rate prediction for sponsored search advertising in Microsoft's Bing search engine." Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010.

3. EXISTING SYSTEM

In order to accomplish this project better, we did some research in the relevant field. There have been a lot of research carried out in developing the prediction models using data mining and machine learning for click-through rate forecast. In this paper, they discovered the relationship between personal user behaviour and the clicking behaviour of their friends on social networks. The comparison of the actual CTR and predicted CTR including the prediction models based on GCM, DBN and CCM. This approach is that it only introduced two newly-designed features in the model, but it just considered the ad blocks and ignore user click behaviours in other blocks. In other words, this simplification of this model may sacrifice some useful information in other blocks.

DISADVANTAGES OF EXISTING SYSTEM

- Privacy Concerns:
- Predictive models rely on user data, raising privacy issues.
- Balancing personalization with privacy is challenging.
- Data Quality and Bias:
- Biased or incomplete data can lead to inaccurate predictions.

4. PROPOSED SYSTEM

This paper has an examination hypothesis which assumed that because general users almost will not click the advertisement if it is at the lower ranks, so if one certain advertisement is clicked then it must be examined and relevant which means that the higher the URL is ranked in the advertisement is at, the higher probability that the advertisement will be clicked. In order to get a more accurate prediction rate, the paper gave some probability of the binary click event C. the paper proposed a cascade model which is different to cascade model.

ADVANTAGES OF PROPOSED SYSTEM

- Cost Efficiency:
- Accurate CTR prediction helps allocate resources effectively by targeting ads to users who are more likely to click, reducing wasted impressions.
- Improved User Experience:
- Relevant ads enhance user experience by showing content aligned with user interest

SYSTEM ARCHITECTURE

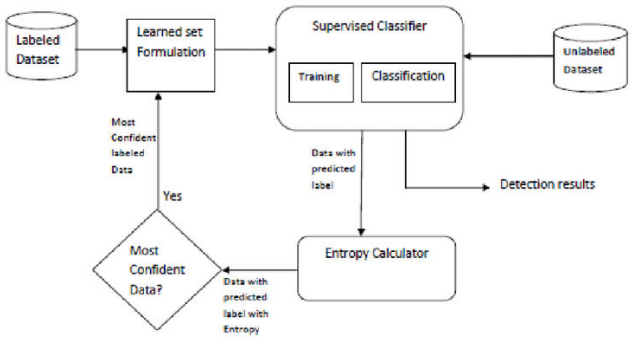

Fig 1: System Architecture

5. UML DIAGRAMS

1. CLASS DIAGRAM

Class diagram is a static diagram. It represents the static view of an application. Class diagram is not only used for visualizing, describing, and documenting different aspects of a system but also for constructing executable code of the software application. Class diagram describes the attributes and operations of a class and also the constraints imposed on the system. The class diagrams are widely used in the modelling of object oriented systems because they are the only UML diagrams, which can be mapped directly with object-oriented languages. It is also known as a structural diagram. Class diagram contains • Classes • Interfaces • Dependency, generalization and association.

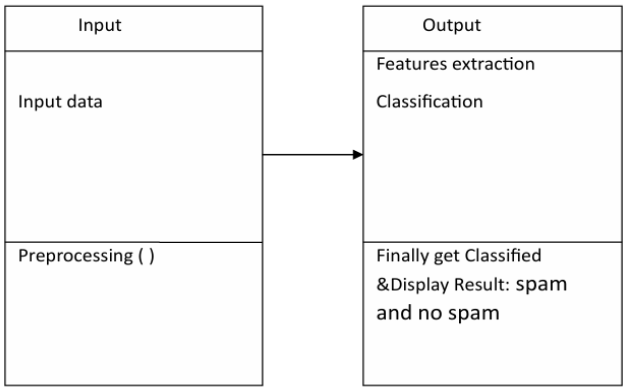

Fig 5.1 shows the class diagram of the project

2. USECASE DIAGRAM:

Use Case Diagrams are used to depict the functionality of a system or a part of a system. They are widely used to illustrate the functional requirements of the system and its interaction with external agents (actors). In brief, the purposes of use case diagrams can be said to be as follows

• Used to gather the requirements of a system.

• Used to get an outside view of a system.

• Identify the external and internal factors influencing the system.

Use case diagrams commonly contains

• Use cases

• Actors

• Dependency, generalization and association relationships.

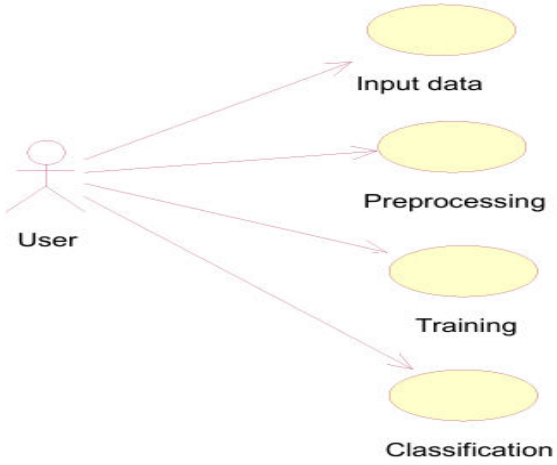

Fig 5.2 Shows the Use case Diagram

3. SEQUENCE DIAGRAM:

A sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function. Sequence diagrams are used to formalize the behaviour of the system and to visualize the communication among objects. These are useful for identifying additional objects that participate in the use cases. These diagrams are widely used by businessmen and software developers to document and understand requirements for new and existing systems.

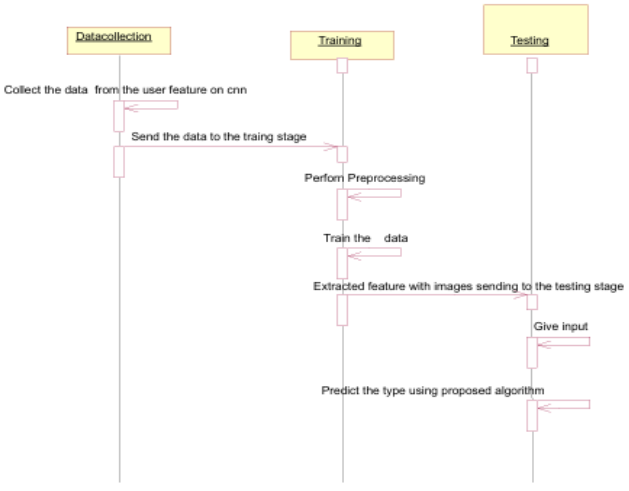

Fig 5.3 Shows the Sequence Diagram

6. RESULTS

6.1 Output Screens

In below screen takes the data from the dataset

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0

Fig 6.1 Upload the Dataset

After upload the dataset we can do preprocess and the get the below graph

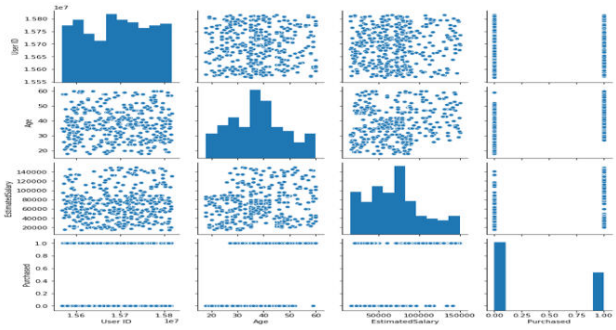

Fig 6.2 Pre-process the data

After pre-processing to run the K-Nearest Neighbor Algorithm and the below accuracy result

	precision	recall	f1-score	support
0	0.93	0.96	0.94	80
1	0.92	0.85	0.88	40
accuracy			0.93	120
macro avg	0.92	0.91	0.91	120
weighted avg	0.92	0.93	0.92	120

Fig 6.3 Run K-Nearest Neighbor Algorithm

In above screen we calculated k-nn accuracy, precision , recall and f-measure and k-nn got 93% prediction accuracy. Now click on 'Run Decision Tree Algorithm' button to train MLP model and to calculate its accuracy.

	precision	recall	f1-score	support
0	0.85	0.95	0.90	80
1	0.87	0.68	0.76	40
accuracy			0.86	120
macro avg	0.86	0.81	0.83	120
weighted avg	0.86	0.86	0.85	120

Fig 6.4 Run Decision Tree Algorithm

In above screen Decision Tree Algorithm got 86% prediction accuracy and in below screen we can see accuracy of Random Forest Algorithm

	precision	recall	f1-score	support
0	0.93	0.95	0.94	80
1	0.89	0.85	0.87	40
accuracy			0.92	120
macro avg	0.91	0.90	0.91	120
weighted avg	0.92	0.92	0.92	120

Fig 6.5 Run Random Forest Algorithm

In above screen to rum random forest algorithm got 92% prediction accuracy and now clicks on Gausian Naive Bayes Classifier Algorithm to get the accuracy in below screen

	precision	recall	f1-score	support
0	0.88	0.95	0.92	80
1	0.88	0.75	0.81	40
accuracy			0.88	120
macro avg	0.88	0.85	0.86	120
weighted avg	0.88	0.88	0.88	120

Fig 6.6 Gausian Naive Bayes Classifier Algorithm

In above screen we calculated Naive Bayes Classifier Algorithm accuracy, precision , recall and f-measure and Naïve Bayes got 88% prediction accuracy. Now click on 'Run Logistic regression Algorithm' button to to calculate its accuracy.

	precision	recall	f1-score	support
0	0.86	0.96	0.91	80
1	0.90	0.68	0.77	40
accuracy			0.87	120
macro avg	0.88	0.82	0.84	120
weighted avg	0.87	0.87	0.86	120

Fig 6.7 run logistic regression algorithm

In above screen we calculated Logistic Regression Algorithm accuracy, precision , recall and f-measure and Logistic regression got 87% prediction accuracy.

.

7. CONCLUSION

The project almost achieves the proposed goals and hypothesis. The main work of this project is composed of the following main three part: massive data cleaning and data feature catching, prediction model training and front end recommendation system designing. Firstly we cleaned the massive data and caught the data features from them. We use real-world advertising dataset collect from taobao users by Alibaba. The data is formed by randomly sampled 1140000 users from the website of Taobao for 8 days advertising display and click. The whole raw data contains advertising information, user profile information, user behavior logs and advertising clicking information. Before we use them as the input data of the training model, we analyze the data in details and do a lot of clean work for them. We firstly discuss about how many features should be kept for the user input data. The features of the user data is used to demonstrate the preference and interest of one certain user, so finally we choose six typical features from the user profile: sex, age, consumption level, shopping dependency, whether is a student or not, the city they live and combined them with the user's history shopping and browsing list and click history record in the previous seven days to entirely demonstrate the preference of one user. Then all of the user datas are processed and store as a row of the csv file. Similarly, we keep some main features for the advertisement, including the advertiser name, category and brand of the product, price of the product and some keywords of the advertisement. In a word, we did the data clean for not only the user information, but also the advertisement information. Secondly we choose an appropriate algorithm and proposed and suitable prediction model then trained it. After comparing other algorithm and other approaches used in advertisement CTR prediction, DeepFM is chosen by our project. Considering that the quality of the predicted modelwill have the biggest influence on the final predicted result, so we spend a lot of time to train the model by use as much training data as possible. After several forward and backward of batches' and several iteration of epochs, the parameters, such as the number of hidden layer, the number of the node in each layer, the initial value of the functions are determined and the predicted model is used to do real life prediction work. Thirdly we designed the front end recommendation system. The front end recommendation system includes two main part, one is advertiser part which is used by the advertiser who want to put the advertisement on the platform, another is platform part which is used by various advertising platform who want to attract more advertiser to put their advertisement on their platform. The advertiser is required to enter the category, brand, price and other features in the home page, then they also have the personal choice to choose one specific platform to do the prediction. The advertising platform is required to submit their user data which is already processed in the specific format. If more and more advertiser and advertising platform are willing to our prediction system, then the advertiser will have more choice when deciding to use which platform and the platform will know more information about coming advertisement which achieve a win-win result. In the recommendation system, In a word, our project successfully achieved an advertisement CTR prediction and recommendation system which has a relative high prediction accuracy. The system is beneficial to user, advertiser and also to the advertising platform by decreasing the advertisement pollution on the Internet. The goals and hypothesis are almost achieved in our project and even get a better result.

FUTURE SCOPE

The AUC is a used to evaluate the prediction model performance. The higher AUC is, the better the prediction model perform. Thus, to improve the AUC of our prediction model through adjusting coefficients and choose the data features would be a direction in the future study. Our project will provide a recommendation system which demonstrates the potential targeting user group based on the click probability calculated by the prediction model . In our target user recommendation system, CTR threshold is used to decide whether a user should be a target user. If the click probability of a certain user, which is forecasted from the prediction model, is larger than the CTR threshold, then this user is recommended as a target user by our system. It is obvious that a reasonable CTR threshold is of great importance for the effectiveness and economical efficiency recommendation system. A reasonable CTR threshold should ensure a desirable actual ad click through rate of target users, meanwhile it should also ensure the advocacy of advertisers, which keeps the appearance on public and increases the awareness of companies and products.If CTR threshold is set to be too high, the target users will have a very high probability to click the ad. However, a high CTR threshold will make the target user pool too small, so that this ad make only a small coverage. On the other hand, a low CTR threshold will guarantee this ad a high coverage, but make the this ad a low actual CTR. Thus, the contributing factors that should be considered for setting a reasonable CTR includes the desired advertisements coverage and the desired actual CTR wanted. Additionally,

the current statistic data shows that the CTR of different industries vary greatly (shown in Table 6). Thus, it is reasonable to consider different CTR threshold for different industry category when setting the benchmark to recommendation. The decision of a reasonable CTR threshold could be studied in the future work.

8. REFERENCES

[1] Minka, T. P. (2001). A family of algorithms for approximate Bayesian inference (Doctoral dissertation, Massachusetts Institute of Technology).

[2] Richardson, M., Dominowska, E., & Ragno, R. (2007, May). Predicting clicks: estimating the click-through rate for new ads. In Proceedings of the 16th international conference on World Wide Web (pp. 521-530). ACM.

[3] Craswell, N., Zoeter, O., Taylor, M., & Ramsey, B. (2008, February). An experimental comparison of click position- bias models. In Proceedings of the 2008 international conference on web search and data mining (pp. 87-94). ACM.

[4] Zhu, Z. A., Chen, W., Minka, T., Zhu, C., & Chen, Z. (2010, February). A novel click model and its applications to online advertising. In Proceedings of the third ACM international conference on Web search and data mining (pp. 321-330). ACM.

[5] Wang, T., Bian, J., Liu, S., Zhang, Y., & Liu, T. Y. (2013, August). Psychological advertising: exploring user psychology for click prediction in sponsored search. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 563-571). ACM.

[6] Chapelle, O. (2015, May). Offline evaluation of response prediction in online advertising auctions. In Proceedings of the 24th International Conference on World Wide Web (pp. 919-922). ACM.

[7] Zhang, W., Du, T., & Wang, J. (2016, March). Deep learning over multi-field categorical data. In European conference on information retrieval (pp. 45-57). Springer, Cham.

[8] JG. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, (2017,May),‘‘Learning from class-imbalanced data: Review of methods and applications,’’ *Expert Syst. Appl.*, vol. 73, pp. 220–239.

[9] Pan, J., Xu, J., Ruiz, A. L., Zhao, W., Pan, S., Sun, Y., & Lu, Q. (2018). Field-weighted Factorization Machines for Click-Through Rate Prediction in Display Advertising. Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW 18.

[10] J. M. Johnson and T. M. Khoshgoftaar, ‘‘Survey on deep learning with class imbalance,’’ *J. Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[11] Readpeak—*The Nordic’s Fastest-Growing Native Advertising Platform*, Helsinki, Finland, Jul. 2022.