SEMARAK ILMU PUBLISHING

International Journal of Advanced Research in Computational Thinking and Data Science

Journal homepage:
https://semarakilmu.com.my/journals/index.php/ctds/index
ISSN: 3030-5225



Exploring the Determinants of Facebook Ad Clicks among Malaysian Users using Machine Learning

Nur Syamila Binti Ahmad Badardin[1,*], Saadi bin Ahmad Kamaruddin[1], Ch'ng Chee Keong[1]

[1] School of Quantitative Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia

ARTICLE INFO

Article history:
Received 25 September 2024
Received in revised form 9 November 2024
Accepted 9 December 2024
Available online 31 December 2024

Keywords:
Ad clicks; machine learning; logistic regression; support vector machine; artificial neural network

ABSTRACT

This research investigates the factors influencing Facebook users to click on a particular advertisement on the Facebook platform. This research has two main objectives. Firstly, it aims to analyze the relationships between Malaysian Facebook users' advertisement click behavior with age, gender, day, frequency and engagement. Secondly, this research aims to compare the performance of various machine learning models in predicting Malaysian Facebook users' advertisement click behaviour. The proposed strategy involves systematic data collection, data preprocessing, data analyzing, data splitting, data training, prediction and comparing performance of various machine learning models. By addressing these objectives, the research hopes to provide valuable insights for businesses to optimize their digital marketing. Engagement with the digital advertisement proves to be main factor in determining whether the user will click on the advertisement. Digital advertisements were also generally clicked towards the end of the week. Moreover, support vector machine proves to be the best model in predicting advertisement clicks among Malaysian Facebook users. In summary, the research was able to investigate the determinants of Facebook Ad clicks among Malaysian users using various machine learning models.

1. Introduction

In today's modern world, social media has become an important aspect of daily life. Many tasks can be done just by using smartphones. These tasks include paying bills, booking an appointment or even shopping. Thus, it just makes sense for businesses to also move their marketing from traditional ways to online or digital advertising to attract potential customers. These recent trends indicate there is a shift in marketing strategies towards online marketing, in particular social media [1]. Among the social media businesses used to put their advertisements and attract customers are Facebook, Instagram and X [2]. This research will dive deep into what attracts customers towards a particular advertisement on the social media platform and comparing machine learning models that can be used to predict customer clicking on a digital advertisement. Hence, it is crucial to understand a few key terms that are related to digital advertising. Ad clicks, for instance, are the actions performed by

* *Corresponding author.*
E-mail address: syamila91@gmail.com

https://doi.org/10.37934/ctds.4.1.19

users when they click on an advertisement, thus redirecting them to another page or prompting specific actions [16]. Moreover, frequency in terms of digital advertising is the number of times the user has seen a particular advertisement pass through their social media feed [16]. Engagement means the interaction the users have with the advertisements such as liking, commenting, sharing, saving and viewing pictures or videos [16].

In the earlier days of digital advertising, the cost of putting out advertisements on social media is small and sometimes negligible thus not affecting the businesses return on investment [17]. However, the Covid-19 pandemic coupled with Stay-At-Home orders has caused an unprecedented exponential increase in e-commerce activities and rapid acceleration of online shopping [17]. Business Insider reveals that a staggering 89% of increase in the average cost per click for Facebook Ads in just one year from 2020 to 2021. Fogarty [18] also stated that the average cost per click is in an upwards trajectory with no signs of slowing down. Data gathered by IBM's US Retail Index concludes that the pandemic has accelerated the shift towards the digital world by five times. The consequence of this situation is marketing departments of businesses are now compelled to produce digital advertisements that are effective and able to increase sales [13]. Moreover, the increase of cost per thousand views (CPM) is still on the upward trend making it crucial for businesses to have more efficient and profitable advertisements digitally [19]. Besides that, being able to target their customers demographic, predicting the time to launch new advertising campaigns and other variables that may affect user behavior towards online advertisements could be beneficial [14]. Being able to predict whether a user will click on an advertisement would also help businesses when producing their digital advertisements [14]. Thus, machine learning can be utilized to achieve these objectives. Understanding and leveraging the predictive nature of machine learning models can significantly enhance the targeting and efficiency of online advertising initiative [3,15]. This research will be able to bridge the gap between statistics and machine learning with online advertising for businesses. This research is also in line with the government's Shared Prosperity Vision 2030 (SPV 2030) that aims to create high income digital jobs and increase digital literacy.

2. Methodology

The Figure 1 shows the methodology workflow of this research. The data used for this research was collected from a local Malaysian company that uses Facebook Ads platform as its primary digital advertising channel. The data was collected over the period of five days when the advertisements were active. This timeframe was selected based on the rationale that five days were sufficient to produce insights. Moreover, prior research, notably Li and Xu [4] in 2022 also utilized about a week's worth of data collection period of eight days. The data collection period chosen was also when the digital advertisements were operational to ensure credible and dynamic data were collected. The data collection timespan was also chosen to avoid any bias or fluctuations that may arise from seasonal variations. The company selected for this research have also given their consent prior to the data collection being made. Care was taken to ensure that the data was used solely for the purpose of this research. The anonymity of the data sources was also kept ensuring compliance with data protection regulations and safeguard users' privacy and no identifiable information from the data was leaked or shared. Monitoring was also conducted during the agreed timespan of the data collection to ensure no modifications were made that may affect the credibility and reliability of the results. For data extraction, the Facebook Ads Manager platform was utilized. For this research, Google Colab platform was utilized from data cleaning to data analyzation, machine learning model implementation until comparing machine learning performance. The versatility of Google Colab and its integration with Google Drive ease the whole process from data cleaning to analyzing results in

the cloud. Moreover, the use of the Python programming language with the abundance availability of Python libraries was vital in conducting the analysis. Firstly, the data was extracted from the Facebook Ads Manager platform. The variables used in the research are listed in the Table 1.

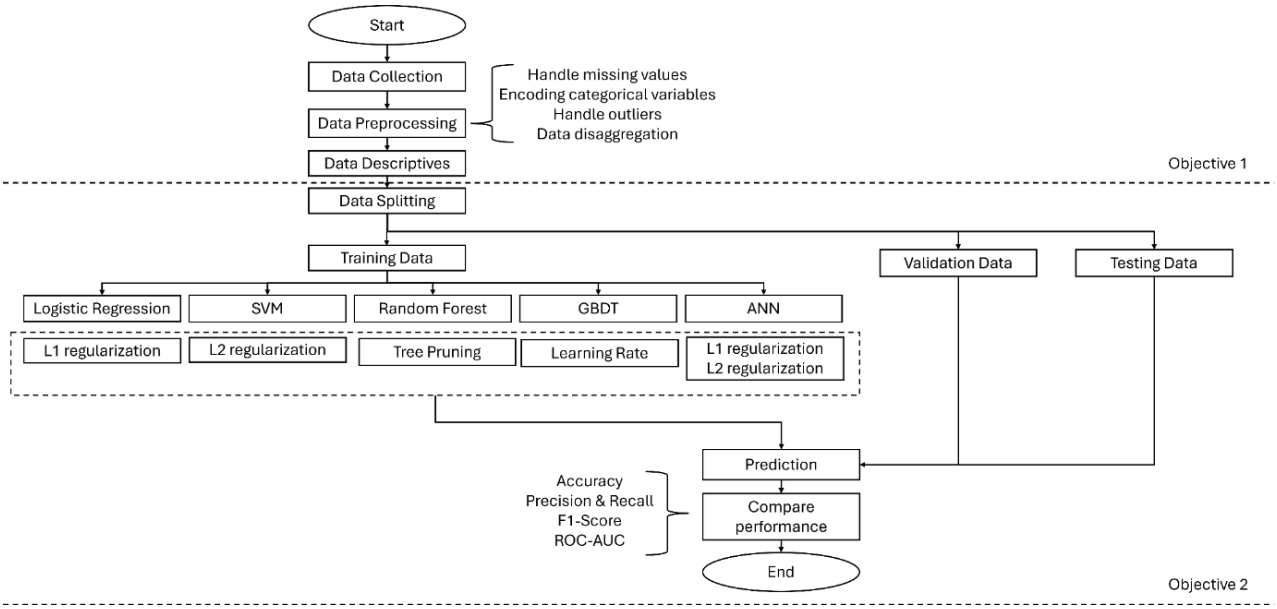


Fig. 1. Methodology workflow

Table 1
Summary of variables

Variable	Type	Measurement Level	Possible Values	Description
Age	Nominal	Categorical	'18-24', '25-34', '35-44', '45-54', '55-64', '65+'	The age range of the Facebook user
Gender	Nominal	Categorical	'male', 'female'	The gender of the Facebook user
Day	Interval	Date	Dates in the format 'YYYY-MM-DD'	The day of interest
Frequency	Ratio	Continuous	Positive real numbers	The amount of the time an ad pass through the Facebook user's timeline
Engaged	Nominal	Binary	0, 1	Engage means Facebook users' action of clicking 'Like', 'Share', comment, viewing pictures and videos on the ad. 0 means the user does not engage with the ad, 1 means the user engage with the ad
Clicked	Nominal	Binary	0, 1	Click means the Facebook users' action of clicking on the ad to the company's on or off Facebook landing page. 0 means the user does not click on the ad and 1 means the user clicks on the ad

After the data was extracted, data preprocessing was carried out. This includes data cleaning, encoding categorical variables and handling missing data and outliers. Next step was to carry out analysis to answer the first objective which was to analyze the relationships between Malaysian Facebook users' advertisement click behavior with age, gender, day, frequency and engagement. The demographic and descriptives of the data were calculated and analyzed to better understand the data at hand. The distribution of the data was examined by constructing pie charts and bar charts.

Next, a time series chart was also constructed to determine the trend and pattern of the behavior of ad click during the data collection period. The most significant visualization to answer the first objective was the correlation matrices i.e. heatmap visualization which helps in determining the factors influencing ad clicks among Malaysian Facebook users.

Next, the research proceeds to answer the second objective which is compare machine learning models performance in predicting Malaysian Facebook users' advertisement click behavior. Before deploying the machine learning models, the data was split into training, validation and testing set according to the ratio of 70%, 15% and 15% [9] respectively. The data splitting process was crucial as it enables training on the training dataset, hyperparameters tuning to prevent overfitting on the validation dataset and evaluation of performance on the unseen testing dataset [5]. This practice of splitting the data ensures that the model performance evaluation is reliable as it helps in preventing overfitting and improving model applicability to new unseen data [7]. The data at hand has an imbalance of classes between the clicked and non-clicked advertisements, therefore stratified splitting was used. Stratified splitting will prevent bias towards the majority class [6] as the proportion of clicked versus non-clicked remains the same for the training, validation and testing dataset. Then, model training was carried out for the training data. The machine learning models that are used for this research are Logistic Regression, Support Vector Machine [11], Random Forest, Gradient Boosting Decision Trees [8] and Artificial Neural Network [10]. Once training was done, validation was carried out and finally testing. All of this was done using Google Colab. To achieve the target of the second objective of this research, all the machine learning model's performance was compared. The metrics that were used to compare the performance were Accuracy, Precision, Recall, F1-score and Area Under the ROC curve. The equations to calculate Accuracy, Precision, Recall and F1-score [20] are presented in Eq. (1), Eq. (2), Eq. (3) and Eq. (4) respectively.

$$\text{Accuracy} = \frac{\text{Number of correct instances}}{\text{Number of all instances}} \tag{1}$$

$$\text{Precision} = \frac{\text{Number of correctly labelled positive instances}}{\text{All positive labelled instances}} \tag{2}$$

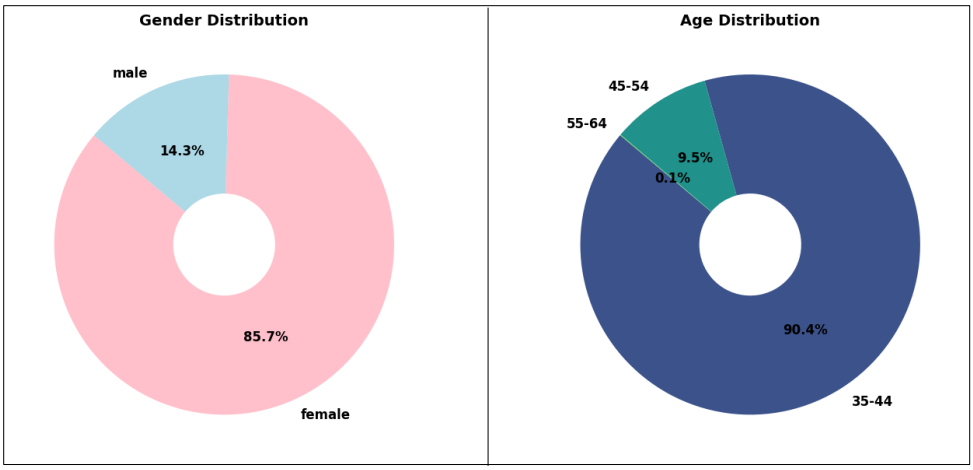
$$\text{Recall} = \frac{\text{Number of correctly lavelled positive instances}}{\text{All true positive instances}} \tag{3}$$

$$\text{F1} - \text{score} = \frac{2}{\left(\frac{1}{\text{precision}}\right)+\left(\frac{1}{\text{recall}}\right)} \tag{4}$$

To ensure accuracy and reliability of the results, cross validation was carried out and the performance was again compared [12]. Cross-validation was carried out by partitioning the datasets into smaller subsets, training, validating and testing the sets multiple times. The metrics that were compared in the cross-validation process were mean and standard deviation.

3. Results

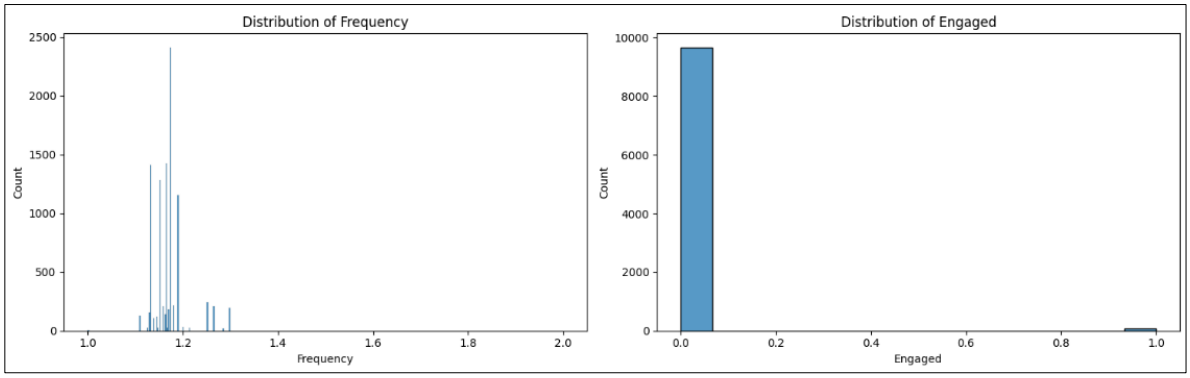
The result from the analysis is presented in this section. The Figure 2 shows the pie charts of gender and age distribution in the dataset.



(a) (b)

Fig. 2. Figure (a) is the Pie Chart of the gender distribution while figure (b) is the pie chart of the age distribution

Based on the Pie Chart in Figure 2, most of the Facebook users for the study are in the 35-44 age group. The gender distribution shows an overwhelmingly higher number of females respondents compared to males. The Figure 3 shows the bar chart of the frequency and engagement of the digital advertisements.



(a) (b)

Fig. 3. Figure (a) is the bar chart of frequency distribution while figure (b) shows the engagement distribution.

The distribution is skewed towards higher frequencies. There are low numbers of engaged users indicating that most of the users are passive or the advertisements are not interactive or interesting. This indicates an area for improvement which is to improve the ad to make it more interactive. Figure 4 shows the Average Click-through rate throughout the span of the data being collected.

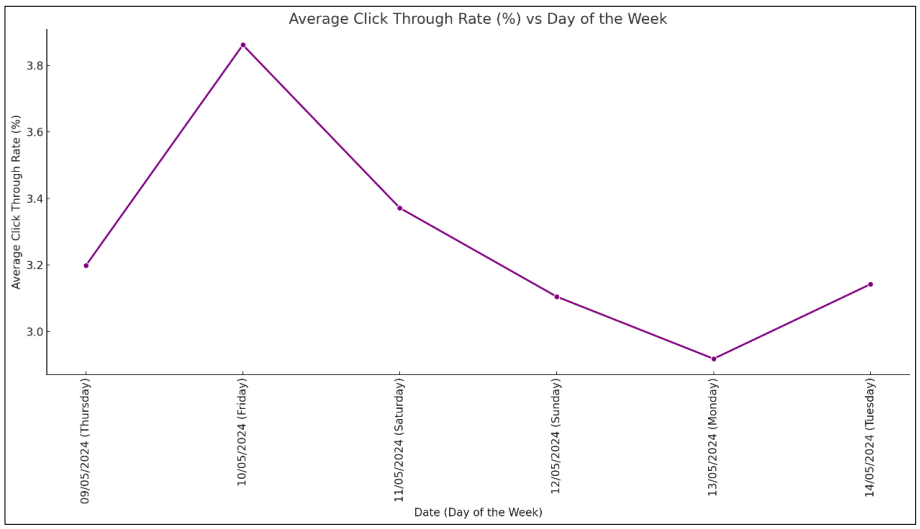


Fig. 4. Trend analysis of the average click-through rate

The trend starts approximately at 3.2% on 9th of May 2024, Thursday. There is a huge increase as it reaches peak on 10th of May, Friday with approximate value of 3.8%. After that, the trend started to decline over the weekend from 11th of May 2024, Saturday to 13th of May 2024 Monday. The trend increases again on the 14th of May 2024 on Tuesday. From this pattern, it can be concluded that there are higher responses to the ad towards the end of the week followed by a decline on the weekend, before rising again starting from Tuesday. The heatmap below in Figure 5 shows the correlation between the variable age, gender, day, frequency and engagement with clicks.

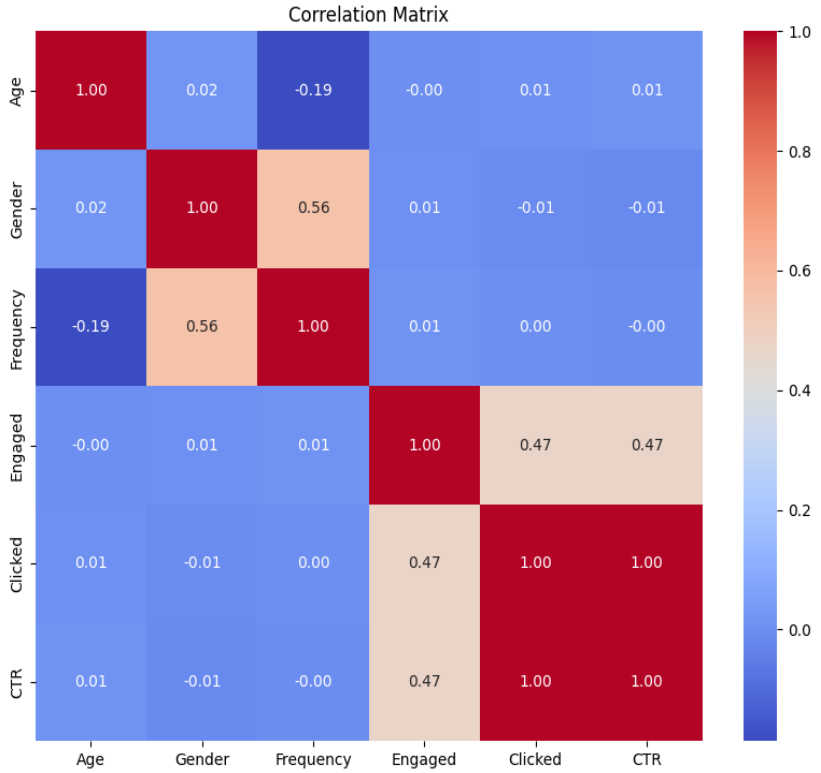


Fig. 5. Heatmap of the independent variables with clicks

From the Figure 5, age shows weak correlation with all other variables. Meanwhile, gender shows moderate positive correlation with frequency, meaning that a certain gender is more likely to view

the ad multiple times. Engaged shows approximately no correlation with other variables except for Clicked and Click Through Rate in which it has moderate positive correlation with Clicked and CTR which means that higher engagement leads to a greater number of ad clicks. Clicked is perfectly correlated with CTR as expected because they essentially measure the same outcomes in the context of ad performance. The tables below show the performance of all the machine learning model in predicting ad clicks in the validation and testing dataset.

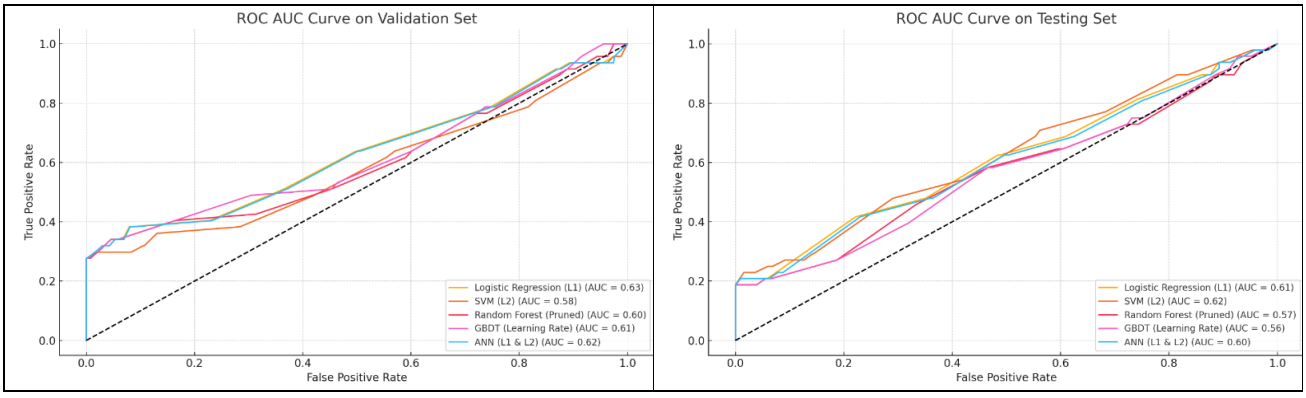
Table 2
Performance metrics of the validation dataset

Model	Accuracy	Precision	Recall	F1 Score	AUC ROC
Logistic Regression	0.97670	1.00000	0.27660	0.43333	0.62780
SVM	0.97670	1.00000	0.27660	0.43333	0.57911
Random Forest	0.97670	1.00000	0.27660	0.43333	0.59858
GBDT	0.97670	1.00000	0.27660	0.43333	0.61372
ANN	0.97670	1.00000	0.27660	0.43333	0.62487

Table 3
Performance metrics of the testing dataset

Model	Accuracy	Precision	Recall	F1 Score	AUC ROC
Logistic Regression	0.97329	1.00000	0.18750	0.31579	0.60997
SVM	0.97329	1.00000	0.18750	0.31579	0.62111
Random Forest	0.97329	1.00000	0.18750	0.31579	0.56822
GBDT	0.97329	1.00000	0.18750	0.31579	0.56453
ANN	0.97329	1.00000	0.18750	0.31579	0.60136

Comparing the metrics from the validation set and training set: accuracy, precision, recall and f1-score are the same for all the models. Attention should be given to the recall values which are relatively low suggesting that the models were able to predict clicked most of the time, however they were not able to identify all the instances in which the users clicked on the ad. The only difference in metrics is in the ROC AUC curve. From the testing metrics, the highest ROC AUC curve is by SVM with 0.62111 indicating it is the best model for ad click prediction. Other models also perform relatively well with values ranging from 0.56453 to 0.60997. In conclusion, whilst all models performed relatively well, SVM emerges as the best model for predicting Facebook ad clicks in the Malaysian context. Figures 6 shows the ROC AUC curve for the validation and testing set.



(a) (b)

Fig. 5. Figure (a) shows the AUC ROC curve for the validation set while figure (b) shows the AUC ROC curve for the testing set

In the Validation set, the highest value for AUC is 0.63 by Logistic Regression followed by 0.62 and 0.61 by ANN and GBDT respectively. This shows that Logistic Regression and ANN have the best ability to distinguish between clicked and non-clicked in the validation set. For the Testing set, the highest value for AUC is 0.62 by SVM followed next by Logistic Regression and ANN with 0.61 and 0.60 respectively. Since SVM performed better in the Testing set, it indicates that SVM is the best model for predicting as it is the best in predicting unseen data. Even though Random Forest and GBDT performed relatively well on the Validation set, its performance dropped in the Testing Set, indicating that these models have tendency at overfitting. Cross-validations were performed to get a more robust estimate of the model's performance. The results are displayed in the table as follows.

Table 4
Cross-validation performance metrics

Model	AUC ROC Mean	AUC ROC Standard Deviation
Logistic Regression	0.9435	0.0152
SVM	0.9461	0.0138
Random Forest	0.9343	0.0174
GBDT	0.9426	0.0135
ANN	0.9452	0.0139

SVM has the highest AUC ROC Mean indicating that it is the best performing model out of all the models with a value of 0.9461. This is followed closely by ANN with 0.9452. Next in line is logistic regression with 0.9435, GBDT with 0.9426 and Random Forest with 0.9461. SVM has an AUC ROC standard deviation of 0.0138 meaning that it is consistent in its prediction. GBDT and ANN also have relatively lower standard deviation indicating that they too have consistent performance overall. Random Forest has the highest standard deviation values suggesting that it has variability in its prediction across different cross validations.

4. Conclusion

The results from this study suggest that engagement plays a vital role in determining whether users will click on an ad. Thus, it is imperative for businesses to create interactive and interesting ads such as introductory catchphrases, highly shareable content, interesting photos and engaging videos among others. Besides that, the average CTR also increases towards the end of the week thus suggesting that businesses should aim to launch their ads at these times to increase ad clicks. Most of the ad viewers are female in the age range of 35-44 years indicating that ads should be geared towards these middle-aged women to increase profit. SVM emerges as the best machine learning model for predicting ad clicks based on the results from the testing set and cross validation. It outperforms other models; however, ANN and Logistic Regression is also a viable model as it follows closely by. Random Forest and GBDT although performing well in the validation set, they have low values in the validation set and the cross validation indicating they tend to overfit and inconsistent in their predictions. There were several limitations of this research as the data was from a singular company thus narrowing down the results to the Facebook users who were interested in the company's products. Thus, this might narrow the gender, and the age of the Facebook users involved in this study. Hence, the results of this study cannot be generalized as there are various products produced by various companies in different industries with different demographic of interests. Future studies should include datasets from various industries. Moreover, the research could also be

conducted for a longer period to better capture the behaviour of users towards digital advertisements.

Acknowledgement
This research was not funded by any grant.

References

[1] Erdem, Şakir, Beril Durmuş, and Osman Özdemir. "The relationship with ad clicks and purchase intention: An empiricial study of online consumer behaviour." (2021).
[2] Kumar, Ashish, Divyanshu Kishan, Harshit Kandpal, Himanshu Saraswat, and Jyotiraditya Singh. "Predicting User Click Behavior on Social Media Ads Using Machine Learning." In *2023 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1-6. IEEE, 2023. https://doi.org/10.1109/ICCCI56745.2023.10128433
[3] Dani, Yasi, and Maria Artanta Ginting. "Classification of Predicting Customer Ad Clicks Using Logistic Regression and k-Nearest Neighbors." *JOIV: International Journal on Informatics Visualization* 7, no. 1 (2023): 98-104. https://doi.org/10.30630/joiv.7.1.1017
[4] Li, Wenqi, and Ziyang Xu. "Factors Affecting User Clicks on Ads." In *2022 6th International Seminar on Education, Management and Social Sciences (ISEMSS 2022)*, pp. 2307-2318. Atlantis Press, 2022. https://doi.org/10.2991/978-2-494069-31-2_272
[5] Sarkar, Dipanjan, Raghav Bali, and Tushar Sharma. "Practical machine learning with Python." *Book" Practical Machine Learning with Python* (2018): 25-30. https://doi.org/10.1007/978-1-4842-3207-1
[6] Raschka, Sebastian, and Vahid Mirjalili. *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt publishing ltd, 2019.
[7] Breiman, Leo. "Random forests." *Machine learning* 45 (2001): 5-32. https://doi.org/10.1023/A:1010933404324
[8] Sun, Rui, Guanyu Wang, Wenyu Zhang, Li-Ta Hsu, and Washington Y. Ochieng. "A gradient boosting decision tree based GPS signal reception classification algorithm." *Applied Soft Computing* 86 (2020): 105942. https://doi.org/10.1016/j.asoc.2019.105942
[9] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. "The elements of statistical learning: data mining, inference, and prediction." (2017).
[10] Cabaneros, Sheen Mclean, John Kaiser Calautit, and Ben Richard Hughes. "A review of artificial neural network models for ambient air pollution prediction." *Environmental Modelling & Software* 119 (2019): 285-304. https://doi.org/10.1016/j.envsoft.2019.06.014
[11] Pisner, Derek A., and David M. Schnyer. "Support vector machine." In *Machine learning*, pp. 101-121. Academic Press, 2020. https://doi.org/10.1016/B978-0-12-815739-8.00006-7
[12] Nakatsu, Robbie T. "Validation of machine learning ridge regression models using Monte Carlo, bootstrap, and variations in cross-validation." *Journal of Intelligent Systems* 32, no. 1 (2023): 20220224. https://doi.org/10.1515/jisys-2022-0224
[13] Gordon, Brett R., Kinshuk Jerath, Zsolt Katona, Sridhar Narayanan, Jiwoong Shin, and Kenneth C. Wilbur. "Inefficiencies in digital advertising markets." *Journal of Marketing* 85, no. 1 (2021): 7-25. https://doi.org/10.1177/0022242920913236
[14] Santoso, Irene, Malcolm Wright, Giang Trinh, and Mark Avis. "Is digital advertising effective under conditions of low attention?." *Journal of Marketing Management* 36, no. 17-18 (2020): 1707-1730. https://doi.org/10.1080/0267257X.2020.1801801
[15] Koehn, Dennis, Stefan Lessmann, and Markus Schaal. "Predicting online shopping behaviour from clickstream data using deep learning." *Expert Systems with Applications* 150 (2020): 113342. https://doi.org/10.1016/j.eswa.2020.113342
[16] Stojanovic, Filip. 2023. "Facebook Ads Frequency Guide: How Companies Determine the Optimal Frequency Caps to Maximize Results on Each Retargeting Campaign | Databox Blog. " Databox. January 12, 2023.
[17] Rosenfeld, Brad. 2022. "How Marketers Are Fighting Rising Ad Costs." *Forbes*, November 14, 2022.
[18] Tomlinson, Max. 2022. "Why Facebook Ads Are so Expensive (and What You Can Do about It)." Conjura. April 13, 2022.
[19] Loeb, Walter. 2021. "The Rising Costs of Digital Advertising Will Force Spending Shifts." *Forbes*, August 4, 2021.
[20] Vujović, Ž. "Classification model evaluation metrics." *International Journal of Advanced Computer Science and Applications* 12, no. 6 (2021): 599-606. https://doi.org/10.14569/IJACSA.2021.0120670