Data Science and Engineering (DSE) Record, Volume 5, Issue 1

# Prediction of Customer Response in Online Advertising

Thanawat Lukuan[1], Prompong Sungunnasil[2], Waranya Mahanan[2], and Sumalee Sangamuang*[2]

[1] Data Science Consortium, Chiang Mai University, Thailand
thanawat_lukuan@cmu.ac.th
[2] College of Art, Media and Technology, Chiang Mai University, Thailand
{prompong.sugunnasil,waranya.m,sumalee.sa}@cmu.ac.th

**Abstract.** Typically, businesses might considerably benefit from user behavior when developing their advertising tactics. Click-through Rate (CTR) is one of the most efficient metrics that provide insights into advertising effectiveness. Moreover, CTR analysis is also used to develop advertising tactics for online marketing. Since a person's lifestyle has changed from offline to online during the COVID-19 pandemic, online-to-offline (O2O) commerce has emerged. O2O commerce is an efficient business model that links offline business activities with online platforms, e.g., Facebook ads. In online situations, CTR analysis can predict the state or fact of something's being likely, the probability that something on an online review and website advertisements will be clicked. Firstly, this paper considers a problem of customer response in online advertising based on CTR prediction. Afterward, a research framework for CTR prediction based on customer response in online advertising using regression models, i.e.linear regression, support vector regression, multi-layer perceptron regression, and random forest regression, is proposed.Such methods only use certain parameters for learning and ignore temporal variance and changes in user behavior The experiments evaluate the regression model's accuracy using R-squared. The experimental results are visualized on scatter plots to describe the relationship between the number of predicted likes and actual likes. The R-squared of the random forest regression model is higher than the others, so the random forest regression model outperforms the other models in analyzing customer response in a tech company's Facebook ads.

**Keywords:** Social network analysis · CTR Prediction · Click-through-rate · Regression.

* corresponding author

.

## 1 Introduction

During the COVID-19 pandemic, people's lifestyles have been changed from offline to online, i.e., online-to-offline (O2O) [12]. O2O commerce is one of the popular business models linking offline business activities with online platforms, e.g., Facebook Ads. In O2O commerce, consumer behavior is the most complicated of the other traditional business models because it involves online platforms. However, two practical popular advertising tools, i.e., sponsored search and social media endorsement, can increase traffic and sales for online sellers at a retail e-commerce platform. According to Sun et al., 2020, sponsored search and social media endorsement can significantly increase seller traffic. [8]

Raising awareness is the most important to create ads targeting the right customer group. Digital marketing tools are online applications that allow a marketing team to execute an effective strategy to sell goods or services, such as The MarTech Wheel. It is a summary of marketing technologies to support lifecycle marketing activities. According to The MarTech Wheel lifecycle, there are four stages, i.e., reach, act, convert, and engage. Finally, it transforms into a conversion rate. Nevertheless, the Conversion rate should be considered relative to a conversion window during which conversions based on a click are measured [1]. Click-through Rate (CTR) [11] measures ad performance by measuring two factors: display and ad click. Ad clicks prediction [4], or click-through rate prediction, mainly represents the customer as a static feature set and trains machine learning regressor to predict customer interest and attention. Such approaches do not consider temporal variance and changes in customer behaviors and solely rely on given features for learning. I propose regression frameworks such as linear, random forest, and MLP regressor for customer click prediction. The goal is to predict customer ad response and campaign-specific ad clicks over a specific period.

Firstly, this paper considers a problem of customer response in online advertising based on CTR prediction. Afterward, a research framework for CTR prediction based on customer response in online advertising using regression models, i.e.linear regression, support vector regression, multi-layer perceptron regression, and random forest regression, is proposed. The methodology for this framework is as follows. Firstly, the data frame is obtained from a tech company's Facebook page, such as Jib Computer Group, IT City, Advice, Banana IT Shop, etc. After that, it is scraped and cleaned to gather a series of the Facebook posts, i.e., *post_id*, *texts*, *time*, *likes*, *comments*, *shares*, *links*, *images*, *video_thumbnail*, *user_id*, and *username*. Next, data patterns, i.e., text analysis and image processing, are explored.

After exploratory data analysis, the data frame consists of 56 features, and there are only 52 features that correlate to each other. For setting up the experiment, there are two parts, i.e., considering all 52 features that correlate to each other and considering only features that correlate to the *like* feature.

Subsequently, machine learning models such as linear regression, support vector regression, multi-layer perceptron regression, and random forest regression are used to predict customer response regarding advertising and campaign-specific advertising. The experiments evaluate the regression model's accuracy using R-squared. This statistical measure indicates how much of the variation of a dependent variable is explained by an independent variable in a regression model. The experimental results are visualized on scatter plots to describe the relationship between the number of predicted likes and actual likes. The R-squared of the random forest regression model is higher than the others, so the random forest regression model outperforms the other models in analyzing customer response in a tech company's Facebook ads.

## 2 Related Works

Online advertising is a form that leverages the Internet to deliver promotional messages to consumers [14]. Enables sellers to reach out to customers seamlessly in many digital media vehicles (e.g., search portals, social media platforms, e-commerce platforms, online games, mobile apps, videos, banners, etc.) [11]. Advertising boosts the awareness of the shop/product to get customers to click. In online advertising, click, ad click prediction or click-through-rate (CTR) prediction is the goal of measuring the performance of the ads, that is, the suitable ads for the right customer to maximize the revenue. Suitable ads bring more visitors (traffic) to the shop/product. Hence, more traffic resulting more engagement and leads to conversion. Ads can be divided into many types. For instance, an ad campaign is most known for promoting. It is a set of ads with a common theme that targets similar groups of customers. It is not only accessible for sellers to create a campaign that covers many ads promoting the same theme but also lets customers quickly focus on what they are interested in and want to click on such ads. From what I have read in most articles, updating advertising strategies and optimizing consumer revenue is key. However, most existing works focus on maximizing ads' income while ignoring ads' negative influence on user experience [14]. Displaying ads too often may disturb consumers, while displaying little may affect income.

The objectives of the review are two main topics. Firstly, I decided to conduct a systematic literature review on existing CTR prediction research and, secondly, identify the current trends and potential future direction worth further exploration. The review of attentive capsule network for click-through-rate [6] is a more specific transformer method for feature interaction from user behaviors, and another topic is the user's response of [4], focusing on several classical methods for CTR prediction.

I have reviewed the paper about computational approaches, such as machine learning methods for user response prediction [3]. It is the result of the analysis from multiple types of research. Most research focuses on algorithmic-driven design to solve specific tasks, such as the classification approach. This paper will comprehensively review user response prediction in online advertising. Our goal is to provide a thorough understanding of online advertising platforms and the typical way of predicting user response. I propose a regression method to analyze and make predictions.

# 3 The Proposed Research framework

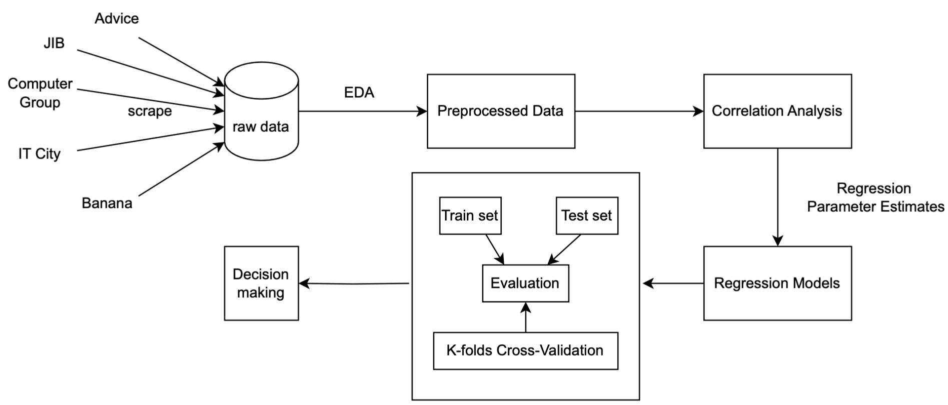


Fig. 1: The overview of the proposed framework's methodology.

Fig. 1 shows the proposed framework's methodology overview. Firstly, the data frame is obtained from a tech company's Facebook page, such as Jib Computer Group, IT City, Advice, Banana IT Shop, etc., selling IT products from 2019 to 2022. After that, it is scraped and cleaned to gather a series of the Facebook posts, i.e., *post_id*, *texts*, *time*, *likes*, *comments*, *shares*, *links*, *images*, *video_thumbnail*, *user_id*, and *username*. Next, data patterns, i.e., text analysis and image processing, are explored. Subsequently, machine learning models such as linear regression, random forest regression, and Multi-layer Perceptron Regression are used to predict customer response in terms of advertising and campaign-specific advertising. Finally, accurately detect the formulated model to analyze the customer response in online advertising.

# 4 Exploratory Data Analysis

In this phase, Facebook posts, i.e., *post_id*, *texts*, *time*, *likes*, *comments*, *shares*, *links*, *images*, *video_thumbnail*, *user_id*, and *username*, can be gathered and scraped using the Facebook Scraper API[3] where the example is shown in Fig.2.

[3] https://github.com/kevinzg/facebook-scraper?tab=readme-ov-file

```
RangeIndex: 11793 entries, 0 to 11792
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   post_id          11793 non-null  int64
 1   text             10257 non-null  object
 2   time             11793 non-null  object
 3   likes            11793 non-null  int64
 4   comments         11793 non-null  int64
 5   shares           11793 non-null  int64
 6   link             8473 non-null   object
 7   images           10937 non-null  object
 8   video_thumbnail  1596 non-null   object
 9   user_id          11791 non-null  float64
 10  username         11793 non-null  object
dtypes: float64(1), int64(4), object(6)
```

Fig. 2: An example of Facebook's posts.

The posts mentioned above come from a tech company's Facebook page, e.g., Jib Computer Group, IT City, Advice, Banana IT Shop, in the category of selling IT products. These scraped features are stored in Python's DataFrame. However, some features are inappropriate, and the Facebook posts must be cleaned to explore a relationship in the data.

Therefore, some features of the Facebook posts must be modified into numerical values, such as the *post_id*, the *comments*, the *shares*, the *user_id* and the *username* are eliminated. In the series of the *likes*, the data will be removed from the DataFrame unless the number of *likes* exceeds zero. For the timestamps, the series of *time* are changed to be the date and time format in Python in which they are used this particular date and time to analyze a big campaign event on the e-commerce platform as a guideline for selecting the effective date, such as double-digit day [5], e.g., *Shopee 11.11* or *Shopee 12.12*. Moreover, they are also used to create a mid-month campaign and payday for the date and transform it using One-Hot encoding. In this paper, first of all, the *date* is separated into five categories as follows the proposed method from [5], i.e., the *normal day*, the *mid-month day*, the *payday day*, the *pre-hype day*, and the *spike day*, which is transformed using One-Hot encoding. Likewise, the *time* is separated into nine categories based on [2], i.e., *midnight*, *early morning*, *morning*, *late morning*, *noon*, *afternoon*, *evening*, *late*, and *very late*, and transformed into a sparse matrix using One-Hot encoding as shown in Fig.??b. Moreover, One-Hot encoding is used to transform the *links* and the *video_thumbnail* to numerical format.

Next, in the content of the *texts*, the number of hashtags is considered by counting all hashtags in each post. Likewise, the words in the Thai language are changed to English using Google Translate to track sentiment analysis on both polarity and subjectivity. The CharCount method is used to count and analyze the content of the *texts*.

The values of *sentiment_polarity* and *sentiment_subjectivity* are obtained by installing and using a Python library called *TextBlob*[4]. The polarity is floating points from -1 to 1, where -1 represents very negative sentiment, 0 represents neutral sentiment, and 1 represents very positive sentiment. The subjectivity is floating points from 0 to 1, where 0 represents objective text (facts), and 1 represents highly subjective text (opinion). Likewise, the values of *lowercase_ratio* are formulated by the ratio of a number of the lowercase (a-z) characters and a character length in each row. Meanwhile, the values of *uppercase_ratio* are formulated by the ratio of a number of the uppercase (A-Z) characters and a character length in each row. Otherwise, the *digit_ratio* is formulated by the ratio of a number of the digit (0-9) and a character length in each row. The *symbol_ratio* is formulated by the ratio of the number of the symbol characters and a character length in each row. The *log1p_emoji_count* is formulated by applying the logarithm function to the total number of emoji(s) plus one in each row. Moreover, The CharCount method is also used to determine the key importance of the brand's names in the content of the *texts*. The *log1p_CharCount* is analyzed from the content of the *texts*, and the numeric contents can be formulated by applying the logarithm function to the total number of character(s) plus one in each row.

On the other hand, the impacts of the colors of the images on the *post* can be considered by finding the correlation of each color that is more important in this work. The gathered images are extracted into RGB image format, which consists of pixels containing three color channels: red, green, and blue. The RGB images can be resized to 0-255 using the *VGG-16* model, a convolutional neural network 16 layers deep. For each color, mean, standard deviation, skewness, and excess kurtosis (exkurt) are calculated to find the correlation of each color, including tri-correlation impacts. Moreover, human detection is essential in considering the impacts of user attention on the images. The YOLO model can extract the images from the post to learn more about whether the result varies and whether the image consists of humans or not [2].

## 5 Correlation Analysis

According to Section 4, the extracted data consists of 52 features, and relationships between them are analyzed using the Pearson correlation coefficient. Typically, there are two types of relationships: positive correlation and negative correlation. A score between +0.5 and +1 can indicate a positive correlation. Meanwhile, a score between -0.5 and -1 indicates a high negative correlation. Fig.3 shows the Pearson correlation coefficient of all features.

[4] https://textblob.readthedocs.io/en/dev/

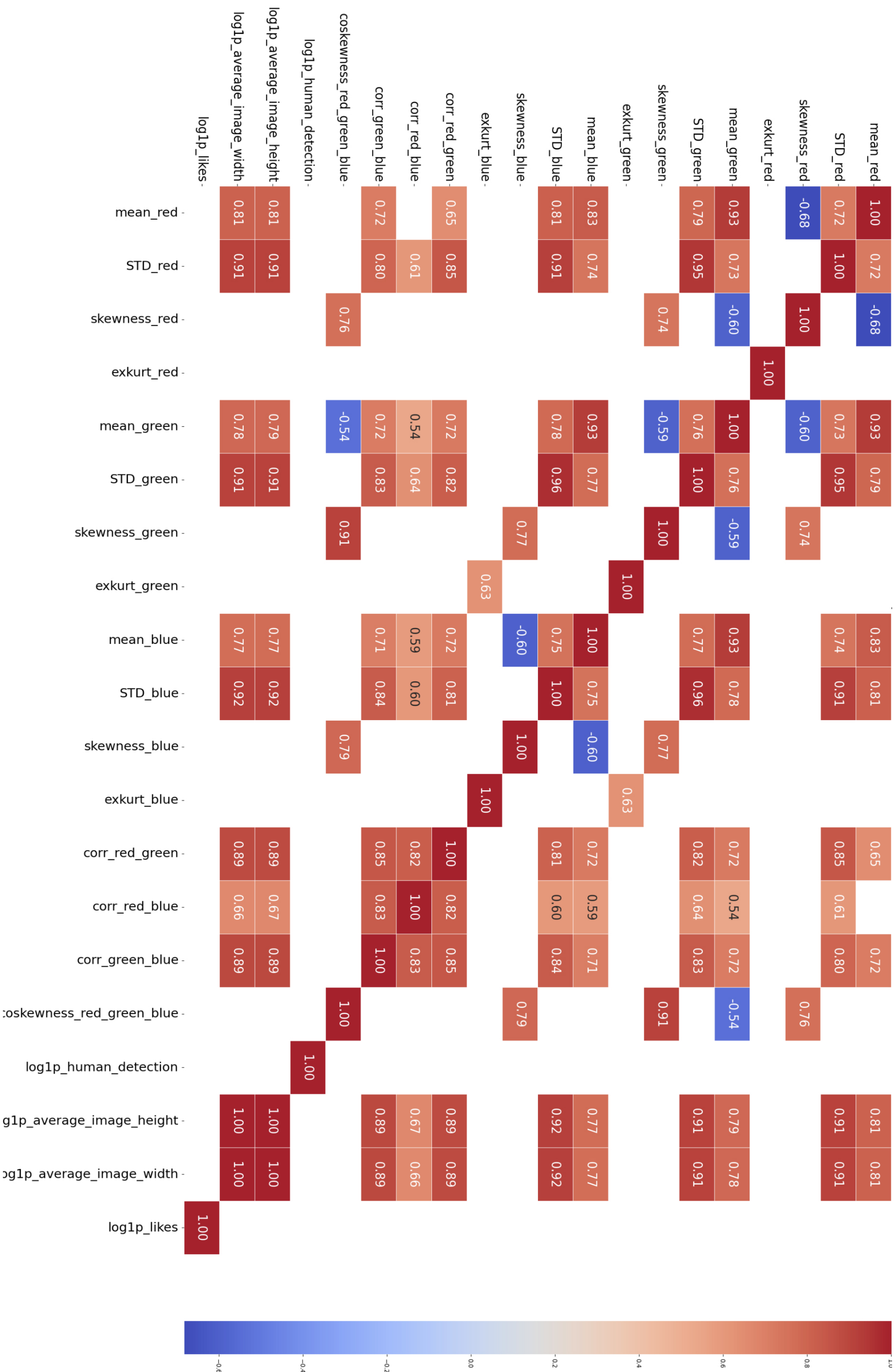


Fig. 3: Pearson correlation coefficient of all features

# 6 Regression Model Formulating

## 6.1 Linear Regression

A linear relationship between one or more factors can be found using linear regression. This study uses the most effective technique to examine the performance of various regression models, including linear regression, to maximize prediction and precision based on [7].

## 6.2 Support Vector Regression

Support vector regression is one method for handling regression issues in supervised machine learning. Analyzing the relationship between a dependent variable and one or more predictor variables can be done via regression analysis. To learn a regression function that maps input predictor variables to output observed response values, Support vector regression formulates an optimization problem. It works well with high-dimensional data because it balances prediction inaccuracy and model complexity based on [13].

## 6.3 Multi-layer Perceptron Regression

A multi-layer perception regression is a neural network with one or more hidden layers. Its Applications for regression and classification tasks can be found in many domains. One of the main goals of neural network research is choosing the right neural network architecture. Therefore, this paper decided to use the regression method. Nowadays, people's interest in internet advertising has increased due to the increasing use of social media platforms. Therefore, this paper intends to assess the effectiveness of the Neural Network Algorithm and other artificial intelligence techniques in forecasting click-through-rate popularity, given the growing relevance of online advertising based on [9].

## 6.4 Random Forest Regression

Random Forest Using ensemble learning techniques, regression is a supervised learning algorithm. Ensemble learning aims to improve prediction accuracy above that of a single model by combining predictions from several machine learning methods. However, most already written studies concentrate on classification techniques, typically restricted to Breiman's original approach for regression and classification problems. However, many advancements may help resolve various real-world issues with online ad prediction. Regression models are used to obtain quantitative numbers for click-through-rate estimates, where the gap in online ads is found, as previously stated. The dependent variable in classification algorithms is qualitative, while the dependent variable in regression methods is quantitative based on [10]

# 7 Experiment Evaluation

According to Section 6, the data input must be normalized before processing each regression model to reduce the effect of outliers using the min-max scalar. the experiments evaluate the regression model's performance using R-squared. This statistical measure indicates how much of the variation of a dependent variable is explained by an independent variable in a regression model. Its values range from 0 to 1, where 1 indicates a perfect fit of the model to the data; thus, if the R-squared value is large, the model's performance will also be high; otherwise, the model's performance will also be low if the R-squared value is small. Since the model's performance in training and testing sometimes does not behave the same in production, e.g., overfitting or underfitting will be experienced during the production environment. This paper uses k-fold cross-validation to estimate the model's performance on new data. The dataset is divided into ten folds (k=10). The regression models are trained and evaluated ten times, using a different fold as the validation set. This experiment for virtualization, the scatter plots show the relationship between the number of predicted likes of linear regression, support vector regression, multi-layer perceptron regression, and random forest regression versus the number of actual likes, respectively, on logarithmic scales. The regression models can describe the relationship between the number of predicted likes and actual likes based on a range of correlations analyzed in Section 5.

## 7.1 Setup

The data in this experiment was gathered from a tech company's Facebook page, e.g., Jib, Computer Group, IT City, Advice, and Banana IT Shop, in the category of selling IT products, and 2,019 posts that were scraped and cleaned to explore a relationship and stored in data storage. The regression models were implemented with Google Colab[5] and analyzed by the Python libraries such as *Numpy*, *Pandas*, *Sklearn*, and *Matplotlib*. The correlation coefficient ($\rho$) determines the degree of each feature's correlation. After exploratory data analysis, the data frame consists of 52 features correlating Facebook posts. For setting up the experiment, there are two parts:

**Part 1**: Consider all 52 features correlating by focusing on a correlation score between +0.5 and +1 and a correlation score between -0.5 and -1.

**Part 2**: Consider only features that correlate to the *logp1like* feature by focusing on a correlation score between 0.05 and 1.

## 7.2 The model's performance

This section shows the relationship between the number of predicted likes of the regression models using the scatter plots on logarithmic scales. Each figure contains the scatter plot of (a) linear regression (labeled as **Linear Reg.**), (b)

[5] https://colab.research.google.com/

support vector regression (labeled as **SVR**), (c) multi-layer perceptron regression (labeled as **MLP Reg.**), and random forest regression (labeled as **RFM Reg.**). The x-axis of all scatter plots is the number of actual likes on a logarithmic scale, while the y-axis is the number of predicted likes on a logarithmic scale.

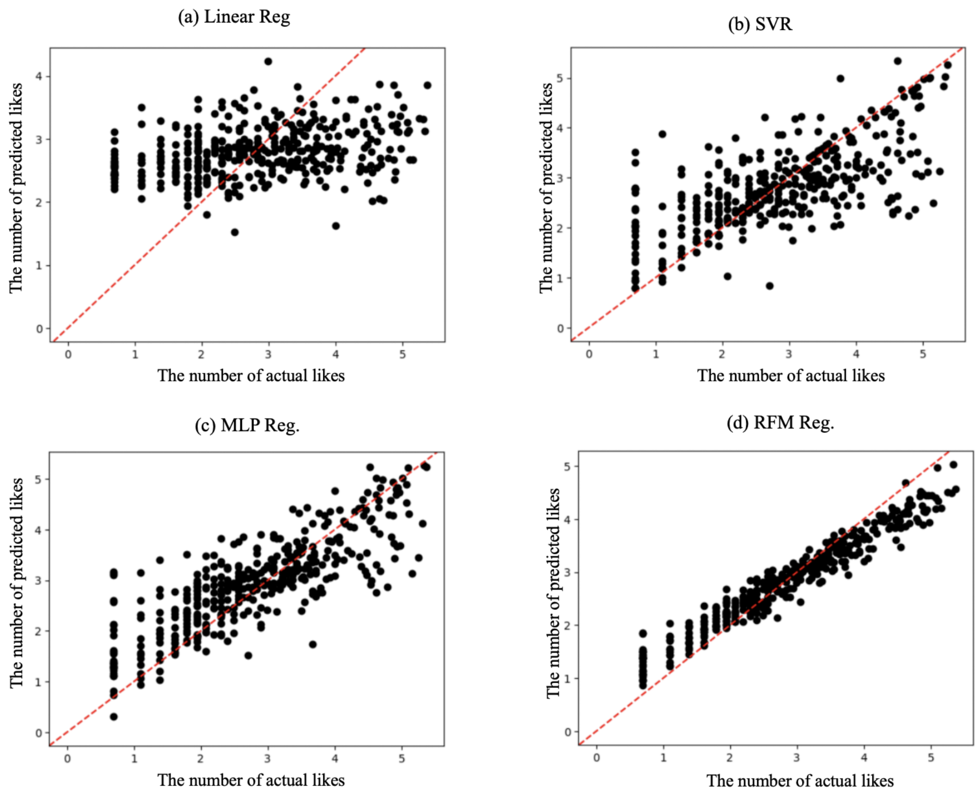


Fig. 4: The scatter plots of the four regression models using the height degrees of the feature's correlation ($-0.5 < \rho < -1$ and $+0.5 < \rho < +1$)

Fig.4 shows the scatter plots of the four regression models in which the height degrees of the feature's correlation ($-0.5 < \rho < -1$ and $+0.5 < \rho < +1$) are used to train and test each regression model. According to the scatter plots for visualization, this experiment used the R-squared value to evaluate the regression model's performance. The train's R-squared values of all models are very similar, i.e., Linear Reg. is 0.860, SVR is 0.842, MLP Reg. is 0.833, and RFM Reg. is 0.793. On the other hand, The test's R-squared values of all models are very different, i.e., Linear Reg. is 0.116, SVR is 0.483, MLP Reg. is 0.641, and RFM Reg. is 0.888. Thus, the RFM Reg's performance outperforms that of the other models.

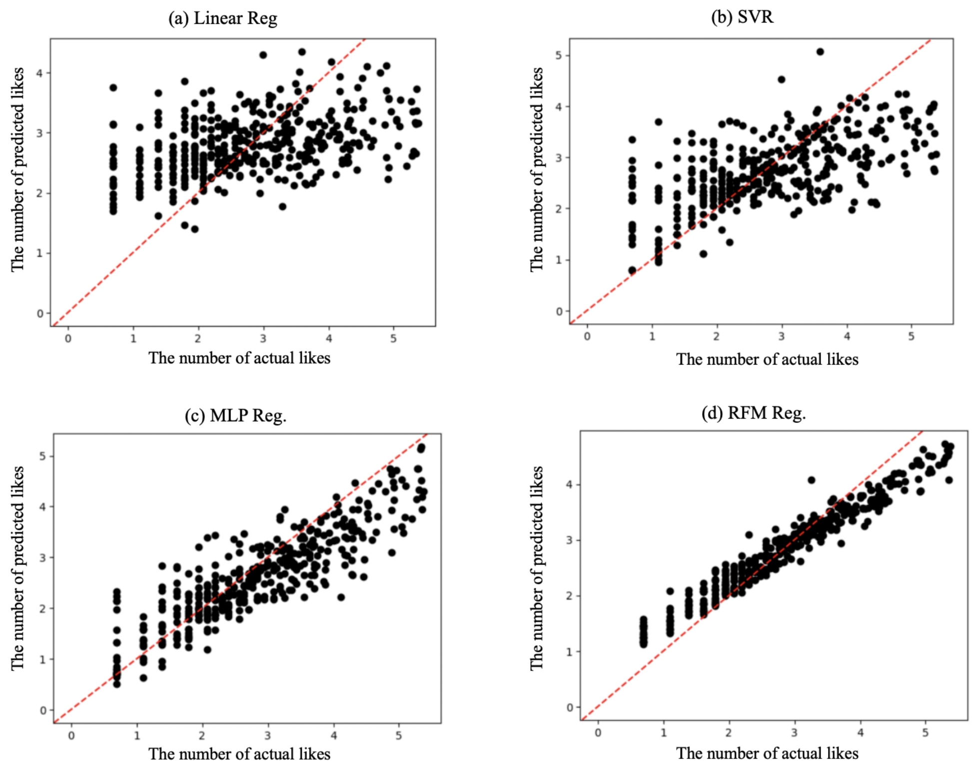


Fig. 5: The scatter plots of the four regression models using only features that correlate to the *logp1like* feature $(0.05 < \rho < 1)$

Similarly, Fig.5 shows the scatter plots of the four regression models that only features that correlate to the *log1plike* feature $(0.05 < \rho < 1)$ are focused. According to the scatter plots for visualization, this experiment used the R-squared value to evaluate the regression model's performance. The train's R-squared values of all models are very similar, i.e., Linear Reg. is 0.837, SVR is 0.791, MLP Reg. is 797, and RFM Reg. is 733. On the other hand, The test's R-squared values of all models are very different, i.e., Linear Reg. is 0.178, SVR is 0.382, MLP Reg. is 0.717, and RFM Reg. is 0.888. Thus, the RFM Reg's performance outperforms that of the other models.

## 8 Contribution

Predicting click-through rates (CTR) with regression is useful for online ads. It helps in the following ways.

Firstly, with better results from the campaign, CTR prediction helps improve ad targeting by finding the people most likely to click. There are more conversions (like purchases) and a better return on ad spend (ROAS) for advertisers when marketers get more clicks on their ads.

Secondly, there is a Better experience for users by guessing the CTR; platforms can keep users from seeing ads that aren't relevant to them. This makes the experience better for users by showing ads more likely to be useful and interesting, making them more engaged and happy.

Lastly, Personalized Delivery of Ads in Regression analysis in CTR prediction helps figure out what about the person and the ad makes people click. This makes it possible to personalize ads and make them more relevant and effective by focusing on specific groups of users.

## 9 Conclusion

Since a person's lifestyle has changed from offline to online during the COVID-19 pandemic, a business model links offline business activities with online platforms, e.g., Facebook ads. In online situations, CTR analysis can predict the state or fact of something's being likely, the probability that something on an online review and website advertisements will be clicked. This paper considers a problem of customer response in online advertising based on CTR prediction. A research framework for CTR prediction based on customer response in online advertising using regression models, i.e.linear regression, support vector regression, multi-layer perceptron regression, and random forest regression, is proposed. The methodology for this framework is as follows. Firstly, the data frame is obtained from a tech company's Facebook page. After exploratory data analysis, the data frame consists of correlated features. For setting up the experiment, there are two parts, i.e., considering all 52 features that correlate to each other and considering only features that correlate to the *like* feature. Subsequently, machine learning models such as the Linear Reg. model, the SVR model, the MLP Reg. model, and the RFM Reg. model are used to predict customer response regarding advertising and campaign-specific advertising. The experiments evaluate the regression model's accuracy using R-squared, and the results are visualized on scatter plots to describe the relationship between the number of predicted likes and actual likes. The R-squared of the random forest regression model is higher than the others, so the random forest regression model outperforms the other models in analyzing customer response in a tech company's Facebook ads.

## 10 Acknowledgement

We thank the Data Science Consortium, Chiang Mai University, for supporting us in this research.

## References

1. Chaffey, D., Smith, P.: Digital Marketing Excellence. Routledge (6 2022). https://doi.org/10.4324/9781003009498

2. Dahlen, M.: It's time for tga: Truly good advertising. Journal of Current Issues Research in Advertising **42**, 123–131 (4 2021). https://doi.org/10.1080/10641734.2020.1780999
3. Gharibshah, Z., Zhu, X.: User response prediction in online advertising. ACM Computing Surveys **54**, 1–43 (4 2022). https://doi.org/10.1145/3446662
4. Gharibshah, Z., Zhu, X., Hainline, A., Conway, M.: Deep learning for user interest and response prediction in online display advertising. Data Science and Engineering **5**, 12–26 (3 2020). https://doi.org/10.1007/s41019-019-00115-y
5. Ginanjar, F.P.: Objektifikasi perempuan sebagai elemen persuasif dalam iklan : Kritik terhadap iklan shopee 11.11 big sale versi tukul arwana. CITRAWIRA : Journal of Advertising and Visual Communication **3**, 59–68 (6 2022). https://doi.org/10.33153/citrawira.v3i1.4621
6. Li, D., Hu, B., Chen, Q., Wang, X., Qi, Q., Wang, L., Liu, H.: Attentive capsule network for click-through rate and conversion rate prediction in online advertising. Knowledge-Based Systems **211**, 106522 (1 2021). https://doi.org/10.1016/j.knosys.2020.106522
7. Maulud, D., Abdulazeez, A.M.: A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends **1**, 140–147 (12 2020). https://doi.org/10.38094/jastt1457
8. Sun, H., Fan, M., Tan, Y.: An empirical analysis of seller advertising strategies in an online marketplace. Information Systems Research **31**, 37–56 (3 2020). https://doi.org/10.1287/isre.2019.0874
9. Syulistyo, A.R., Agustin, V.M., Puspitasari, D.: Predicting news article popularity with multi layer perceptron algorithm. Journal of Applied Intelligent System **7**, 193–205 (9 2022). https://doi.org/10.33633/jais.v7i2.6826
10. Tyralis, H., Papacharalampous, G., Langousis, A.: A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water **11**, 910 (4 2019). https://doi.org/10.3390/w11050910
11. Yang, Y., Zhai, P.: Click-through rate prediction in online advertising: A literature review. Information Processing Management **59**, 102853 (3 2022). https://doi.org/10.1016/j.ipm.2021.102853
12. Yao, P., Osman, S., Sabri, M.F., Zainudin, N.: Consumer behavior in online-to-offline (o2o) commerce: A thematic review. Sustainability **14**, 7842 (6 2022). https://doi.org/10.3390/su14137842
13. Zhang, F., O'Donnell, L.J.: Support vector regression. In: Machine learning, pp. 123–140. Elsevier (2020)
14. Zhao, X., Gu, C., Zhang, H., Yang, X., Liu, X., Tang, J., Liu, H.: Dear: Deep reinforcement learning for online advertising impression in recommender systems. Proceedings of the AAAI Conference on Artificial Intelligence **35**, 750–758 (5 2021). https://doi.org/10.1609/aaai.v35i1.16156