# Predicting Click Behavior Based on Machine Learning Models

Xiaoyu Zhou*

Stan Richards School of Advertising & Public Relations, University of Texas at Austin, United States

*Corresponding author: rainzhou@utexas.edu

**Abstract.** As businesses spend a great amount of advertising dollars each year and hope to achieve marketing objectives, advertising practitioners often advise them to understand customer behavior and customize target audience. This study gains information about click behavior and uses those to help brands make targeting decisions. In the study, a predictive analysis on whether a viewer clicks on the ad was conducted to build a model that forecasts user clicks and forms a list of potentially interested viewers. The predictive model included variables that mainly contained viewer information, such as 'Daily Time Spent on Site', 'Age', 'Area Income', etc., and were chosen based on correlations with each other. After splitting the samples with a 7:3 ratio, machine Learning algorithms, including Logistic Regression, Random Forest, and Support Vector Machine, were used to build the model. Each model was implemented on both datasets, and with no sign of overfitting shown, Support Vector Machine model outperformed other models with a 97% prediction accuracy on training data and a 96% prediction accuracy on testing data. With this model being able to generalize well on new data points, businesses can start implementing it to filter out the audience who will click on an ad and show interests in the brand/product. This deeper understanding of the audience will help improve targeting for campaigns with different objectives and achieve higher advertising efficiency and effectiveness.

**Keywords:** Machine Learning; Predictive Analysis; Advertising; Customer Behavior.

## 1. Introduction

Each year, businesses spend a large amount on marketing, more specifically, on advertising, with the motivation of promoting their products or services. It is estimated that the total advertising expenditure in the U.S. would reach $345.3 billion in 2022, continuing to grow from 2020 [1]. Due to the rapid development of technology, an increasing number of individuals have shifted their daily media consumption to online platforms instead of traditional media such as TV, newspaper, etc. Understanding this change in customers, businesses have started putting more ad dollars into digital media advertising that mainly includes paid search, paid social, display, and programmatic ads. It is estimated that digital advertising will account for 71.8% of total US media ad spend in 2022 [2].

Given the popularity of digital advertising and its significance on businesses' advertising efforts, it is essential to keep track of how efficiently businesses are spending their ad dollars in this category to achieve their goals, and one key approach is to customize the breadth and depth of the target audience in accordance with objectives. Therefore, it is essential to understand customer behavior, especially their attitude toward the brand. One of the key metrics that are commonly used in digital advertising is "clicks", which measures every time when a user clicks on the ad. Whether the audience would click on an ad or not can potentially reveal information about their knowledge and interests towards the brand, and is especially helpful for lower-funnel campaigns with the objective of conversions when businesses want to their target audience to ones with initial interests. As a result, a predictive analysis is carried out in this study to forecast click-on-ad using an advertising dataset based on various machine learning (ML) models [3]. The predictive model selected is anticipated to assist businesses in understanding and identifying target audiences for improved campaign efficiency.

Although predictive analyses have been widely performed to understand customer behavior, the majority of them focus on purchase behavior, rather than clicks behavior that happens during the advertising process. For instance, Kim et al. used Deep Neural Networks (DNN) to predict whether a smartphone user repurchases the same-brand smartphone in 2021 [4]. These studies would help

businesses understand factors contributing to conversions, but are not as helpful for ad targeting purposes. In addition, existing studies also rarely make predictions with multiple ML models and vote for the best-performing one. In the research conducted by Subramanian et al., only one classification model called Naive Bayes, was used when making predictions about general customer behavior [5]. However, using only one model seriously limits the prediction accuracy and prevents possible optimization.

To fill this gap, this study uses advertising click data and makes predictions on whether a viewer clicks on an ad based on a variety of factors to identify interested customers with the goal of implementing the model to customize advertising targeting in the future. In order to achieve the best possible results, three different machine learning models along with model ensembling are used and compared to decide on the final predictive model. The remainder of this paper is organized as follows: in section 2, the dataset is introduced, and the methods used in this study are explained. The results of the study and their implications are provided in section 3, and lastly, section 4 covers the conclusion of this research study.

## 2. Dataset and Methods

### 2.1 Data Description and Preprocessing

#### 2.1.1 Data Description

The Advertising dataset, originally from Byrnes, is collected from Kaggle [3], and used in this study to predict click behavior. The dataset provides information of demographics, online behaviors, and click results for each viewer. The original dataset contains 1000 records and 10 variables, which include the following: 'Daily Time Spent on Site', 'Age', 'Area Income', 'Daily Internet Usage', 'Ad Topic Line', 'City', 'Male', 'Country', 'Timestamp', 'Clicked on Ad'. Of all the variables, 'Daily Time Spent on Site', 'Area Income', and 'Daily Internet Usage' are floats; 'Age', 'Male', and 'Clicked on Ad' are integers (numbers without decimals); and the rest of variables are objects. For variable 'Male', it indicates a male viewer if the value is 1 and a female viewer if the value is 0. For variable 'Clicked on Ad', it indicates that the viewer clicked on the ad if the value is 1 and did not click on the ad if it is 0.

#### 2.1.2 Data Preprocessing

In the beginning, descriptive analysis (Table 1) was conducted to gain a better understanding of the dataset. It is shown that the average age of viewers is 36 with the youngest being 19 and the oldest being 61. The average time that viewers spend on site is 65 min/day and the average time spent on the Internet is 180 min/day. Of the total 1,000 viewers, 48.1% of them are males while 51.9% are females, and their average area income is $55,000/yr.

**Table 1**. Descriptive Statistics of The Collected Dataset

| | Daily Time Spent on Site | Age | Area Income | Daily Internet Usage | Male | Clicked On Ad |
|---|---|---|---|---|---|---|
| Count | 1000.00 | 1000.00 | 1000.00 | 1000.00 | 1000.00 | 1000.00 |
| Mean | 65.00 | 36.00 | 55000.00 | 180.00 | 0.48 | 0.50 |
| Std | 15.85 | 8.79 | 13414.63 | 43.90 | 0.50 | 0.50 |
| Min | 32.60 | 19.00 | 13996.50 | 104.78 | 0.00 | 0.00 |
| 25% | 51.36 | 29.00 | 47031.80 | 138.83 | 0.00 | 0.00 |
| 50% | 68.22 | 35.00 | 57012.30 | 183.13 | 0.00 | 0.50 |
| 75% | 78.55 | 42.00 | 65470.64 | 218.79 | 1.00 | 1.00 |
| Max | 91.43 | 61.00 | 79484.80 | 269.96 | 1.00 | 1.00 |

To better understand viewer distribution for clicks, this study visualized 'Click on Ad' based on other variables, such as sex Figure 1), from which it can be found that females are slightly more likely to click on an ad. New variables, 'Hour', 'Day of Week', 'Date', and 'Month', were also introduced

from the existing variable 'Timestamp', and then took a look at how clicks distribute based on these new variables. From January to August 2016, as shown in Figure 2, no clear seasonality is shown but on February 14th (Valentine's Day), the number of clicks reached its peak. Therefore, there are potential opportunities for a holiday-specific campaign launch on Valentine's Day as it would drive more clicks compared to other days.

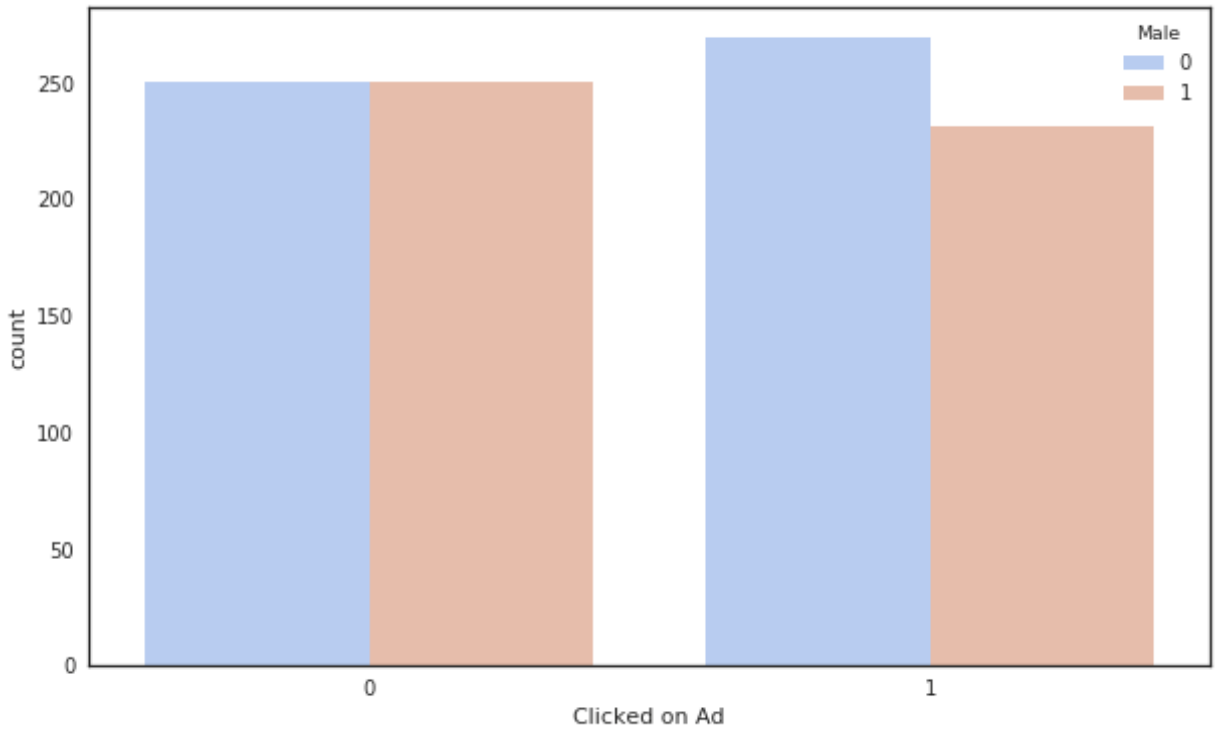


**Fig. 1** Clicked on Ad by Sex

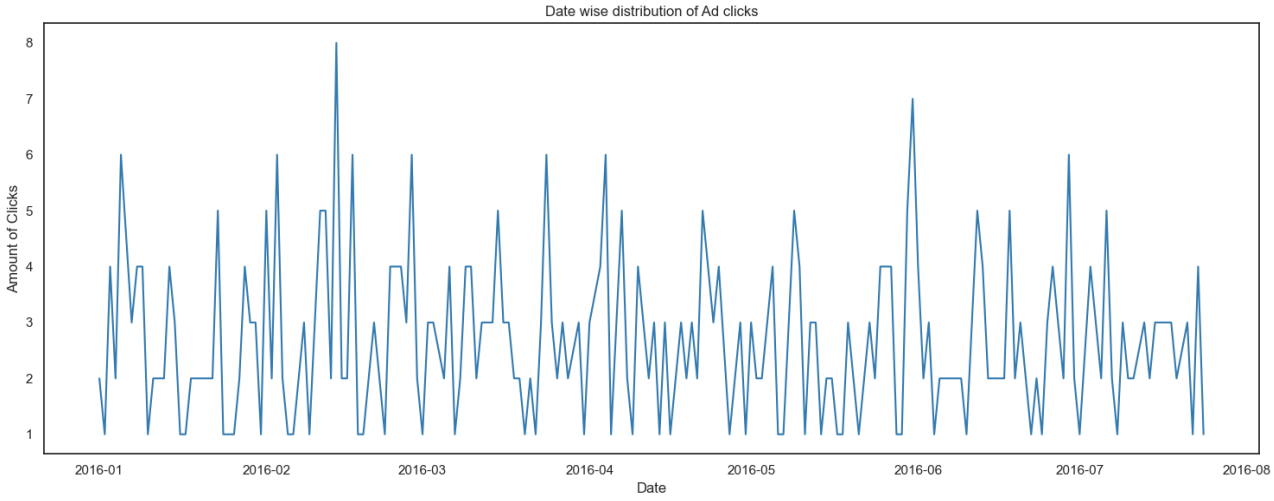


**Fig. 2** Ad Clicks by Date

In addition, this study took a look at click distribution based on viewer age (Figure 3), and it can be observed that people who are in their 40s account for a sizeable portion of clicks. Figure 4 also depicted the distribution of time people spent on the site and the Internet by age. It can be concluded that viewers around 30 years old spend more time on the site or on the Internet while viewers in their 40s are more likely to click on the ads. Lastly, as shown in Figure 5, people with lower incomes (under $60,000/year) are more likely to click on the ads than people with higher incomes. With this information, it can be concluded that females in their 40s in the lower middle/middle-class show more interest in our product/service.

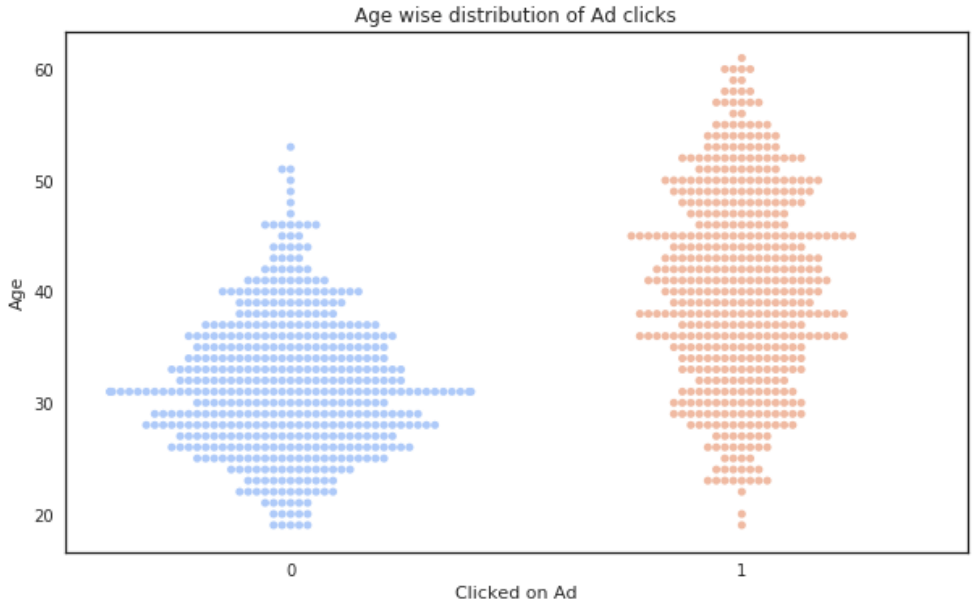


**Fig. 3** Clicked on Ad by Age

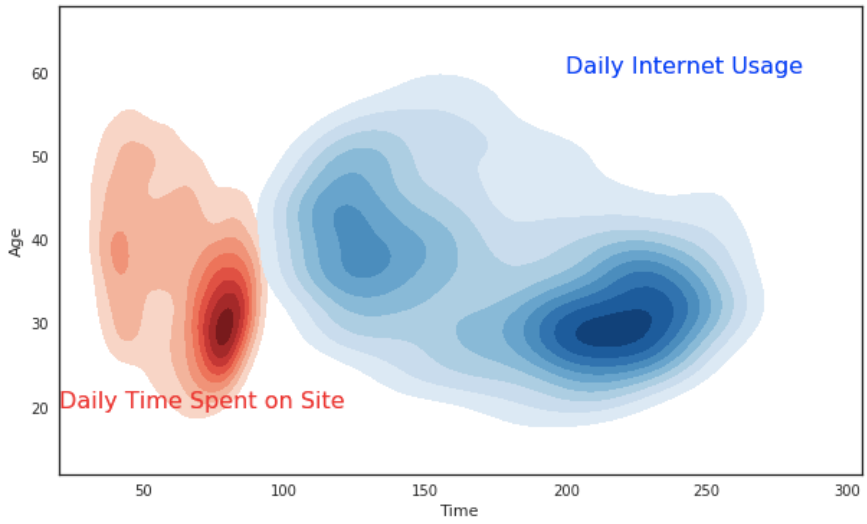


**Fig. 4** Daily Time Spent on Site (Red) & Daily Internet Usage (Blue) by Age

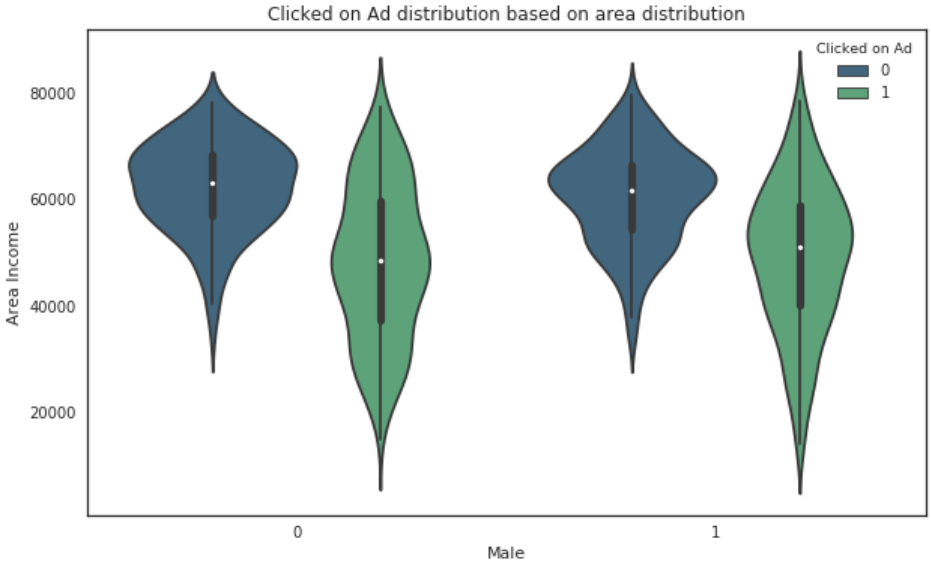


**Fig. 5** Clicked on Ad by Sex & Area Income

Before building the model to predict ad click probability, the variables that are not correlated well with the label ( i.e. 'Click on Ad') should be deleted, so correlations between each pair of variables were explored by building a correlation matrix as shown in Figure 6 [6]. From the correlation matrix, it can be found that all numerical variables in the dataset can be included in the model, and 'Daily Time Spent on Site' and 'Daily Internet Usage', specifically, have stronger inverse relationships with 'Click on Ad'. In addition, the variable "Ad Topic Line" was dropped for the model-building process because there are 1,000 unique values for this variable, making it hard to generate any insights. "Country" is also dropped because it is not reasonable to take regional differences into consideration for this analysis.

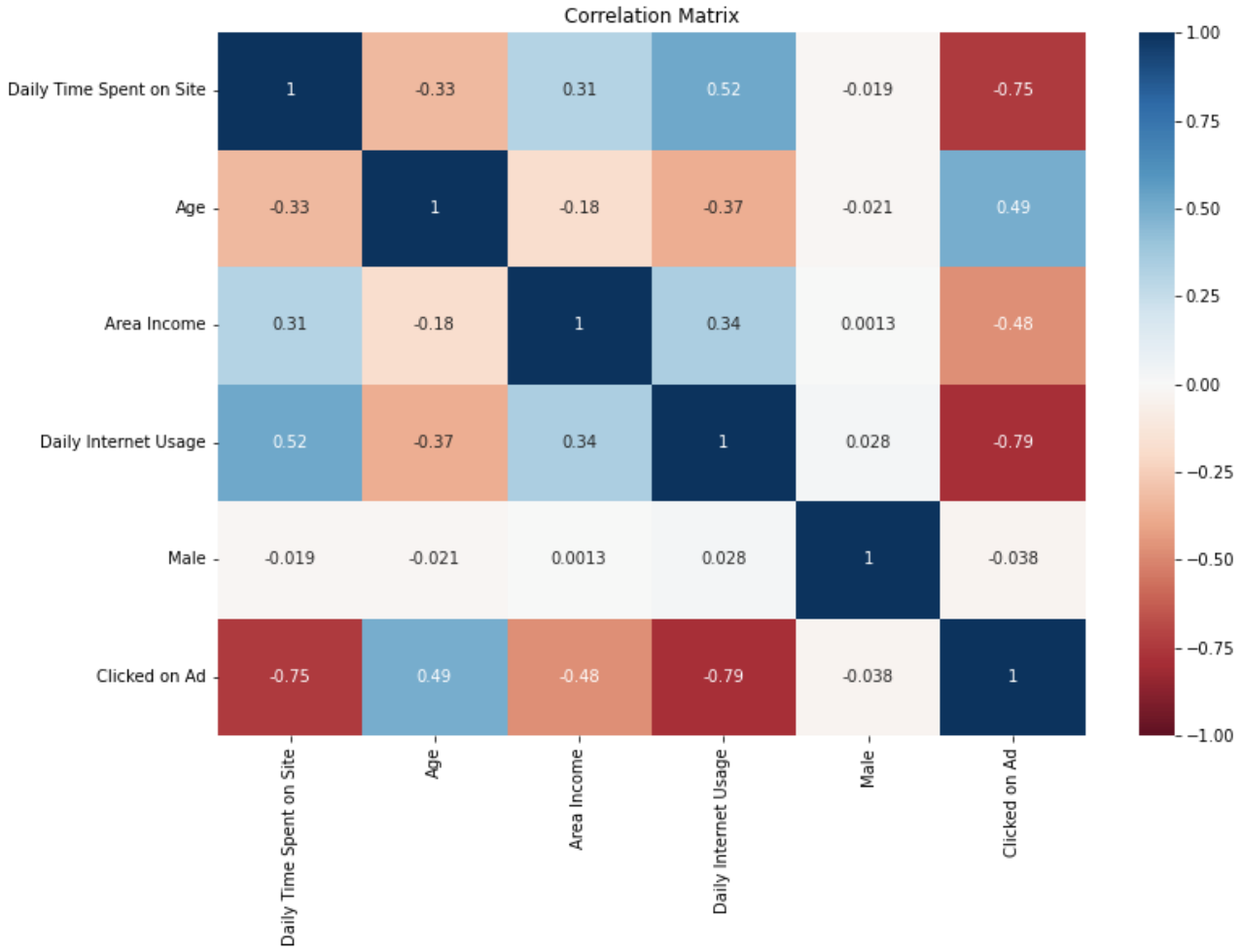


**Fig. 6** Correlation Matrix for features based on Pearson coefficient

## 2.2 Methods

In this section, it explains models built with different machine learning algorithms to predict 'Click on Ad' based on the final 5 variables: 'Daily Time Spent on Site', 'Age', 'Area Income', 'Daily Internet Usage', and 'Male'. The dataset was divided into train data and test data firstly, as train data is used to train the model, and then test data is to evaluate model performance. Since this is a binary classification problem, the following classification models were trained for this predictive analysis:

### 2.2.1 Logistic Regression Model

Logistic Regression model is a linear model that is commonly used in binary classification problems where the output y to predict is always coded with 0 and 1, indicating whether happened or not. Since a logistic regression does a logistic transformation to a linear regression surface. Therefore, as Linear Regression has the formula:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n = \sum b_ix_i \quad (1)$$

The output from the linear regression then will be passed into the sigmoid activation function for mapping and as the final output:

$$y = \frac{1}{1+e^{-\sum b_ix_i}} \quad (2)$$

### 2.2.2 Random Forest Model

Random Forest is a supervised machine learning algorithm that generally performs better for classification problems. It used the bagging technique, a model ensembling technique that combines models, as shown in Figure 7, to integrate outputs from simple decision trees on different samples in order to make more precise predictions.

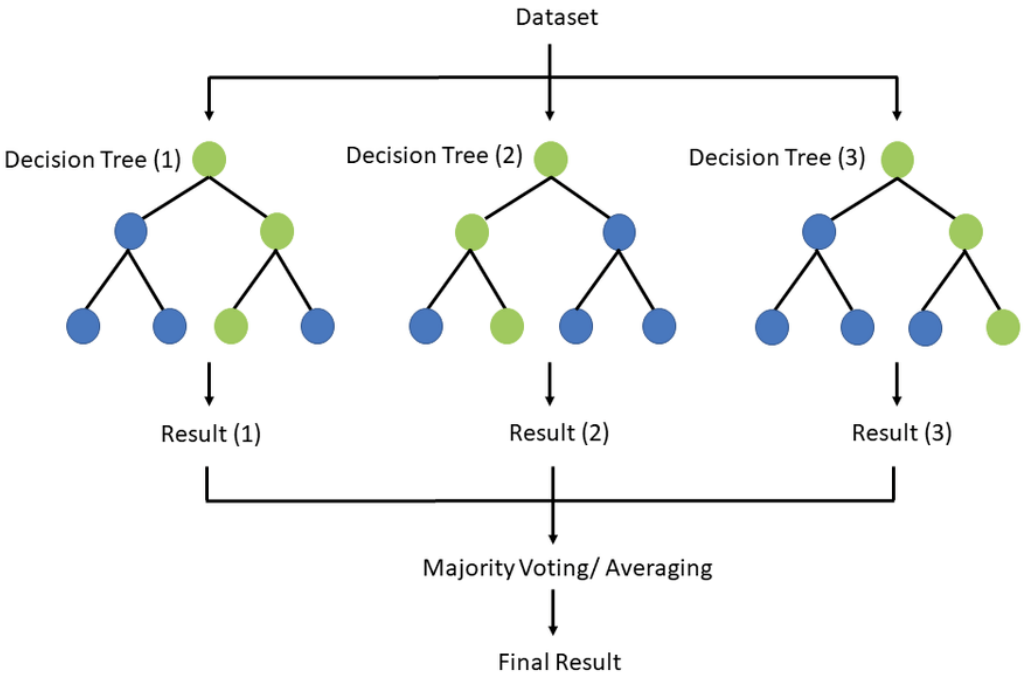


**Fig. 7** Random Forest Algorithm [7].

### 2.2.3 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning model that, although can be used for both classification and regression purposes, is more commonly used for classification problems. SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes, as shown in the following image (Figure 8):

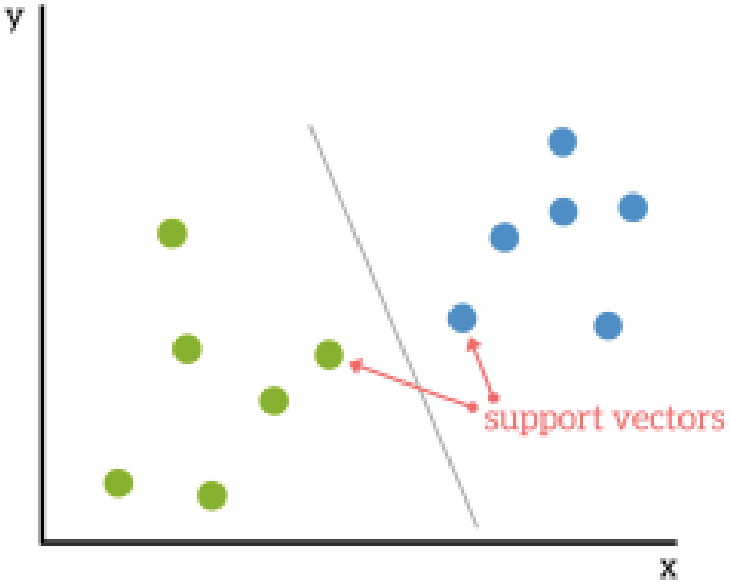


**Fig. 8** Support Vector Machine Algorithm [8].

A hyperplane as a line that linearly separates and classifies a set of data. Therefore, the further all support vectors are from the hyperplane, the most confident we are that they have been correctly classified [6].

## 3. Results and Discussion

### 3.1 Results

#### 3.1.1 Logistic Regression Model

As shown in the classification report (Table 2, 3), the logistic regression model reached 90% accuracy on both training data and testing data. It was concluded that there is no sign of overfitting due to the similar model's performance in the training set and the testing set.

**Table 2**. Logistic Regression Model Performance on Training Data

| | Precision | Recall | f1-score | Support | Accuracy |
|---|---|---|---|---|---|
| 0 | 0.88 | 0.92 | 0.90 | 354 | 0.90 |
| 1 | 0.92 | 0.88 | 0.90 | 346 | |

**Table 3**. Logistic Regression Model Performance on Testing Data

| | Precision | Recall | f1-score | Support | Accuracy |
|---|---|---|---|---|---|
| 0 | 0.85 | 0.96 | 0.90 | 146 | 0.90 |
| 1 | 0.96 | 0.84 | 0.89 | 154 | |

#### 3.1.2 Random Forest Model

The random forest model, as shown below (Table 4, 5), outperformed the logistic regression model with a prediction accuracy of 94% on training data and 93% on testing data. Similarly, there was no sign of overfitting.

**Table 4**. Random Forest Model Performance on Training Data

| | Precision | Recall | f1-score | Support | Accuracy |
|---|---|---|---|---|---|
| 0 | 0.95 | 0.94 | 0.94 | 354 | 0.94 |
| 1 | 0.94 | 0.95 | 0.94 | 346 | |

**Table 5**. Random Forest Model Performance on Testing Data

| | Precision | Recall | f1-score | Support | Accuracy |
|---|---|---|---|---|---|
| 0 | 0.96 | 0.89 | 0.92 | 146 | 0.93 |
| 1 | 0.90 | 0.96 | 0.93 | 154 | |

### 3.1.3 Support Vector Machine

Lastly, the support vector machine algorithm was implemented to predict clicks behavior and as shown in Table 6 and 7, there was a significant improvement in accuracy from the logistic regression model and the random forest model with a 97% total precision on train data and a 96% precision on test data. There was also no sign of overfitting for the SVM model.

**Table 6**. SVM Performance on Training Data

| | Precision | Recall | f1-score | Support | Accuracy |
|---|---|---|---|---|---|
| 0 | 0.96 | 0.99 | 0.97 | 354 | 0.97 |
| 1 | 0.99 | 0.95 | 0.97 | 346 | |

**Table 7**. SVM Performance on Testing Data

| | Precision | Recall | f1-score | Support | Accuracy |
|---|---|---|---|---|---|
| 0 | 0.94 | 0.97 | 0.96 | 146 | 0.96 |
| 1 | 0.97 | 0.94 | 0.96 | 154 | |

### 3.2 Discussion

Based on this predictive analysis with several machine learning models, this study is able to reach 96% prediction accuracy with the SVM model to predict ad click probability based on 5 variables that mainly contain viewer information. The SVM model performed better than other classification models in this study because this type of model is less influenced by outliers and works best with a limited number of samples (in thousands) [9, 10]. Since there was no sign of overfitting or underfitting, the model is able to generalize well when used on new data points. The analysis also helped gain a better understanding of click distributions by factors such as age, gender, income, etc. and identified characteristics of viewers who are most likely to click or most interested in the brand. With the SVM model, businesses are able to form a list of viewers who are interested in the brand and are likely to click on the ad. This information will be helpful for deciding the target audience and customizing advertising efforts for future campaigns.

## 4. Conclusion

In this study, predictive analysis was proposed to understand customer behavior, especially clicks behavior, for advertising targeting purposes. Machine learning algorithms, such as logistic regression, and random forest were applied to the advertising dataset to predict whether viewers click on the ad or not. In the end, the model generated with the support vector machine algorithm performed better than other models with the highest accuracy in both train and test data. In the future, more features of advertisements can be introduced as variables along with the existing viewer information to add more perspectives and potentially improve the model performance.

## References

[1] Industry market research, reports, and Statistics. IBISWorld. (n.d.). Retrieved December 23, 2022.

[2] Lebow, S. Digital will account for 71.8% of US media ad spend this year. Insider Intelligence. Retrieved December 23, 2022.

[3] Byrnes, T. Advertising. Kaggle. Retrieved December 23, 2022.

[4] Jina Kim, HongGeun Ji, Soyoung Oh, et al. A deep hybrid learning model for customer repurchase behavior. Journal of Retailing and Consumer Services, 2021.

[5] R. Siva Subramanian, D. Prabha. Customer behavior analysis using Naïve Bayes with bagging homogeneous feature selection approach. Journal of Ambient Intelligence and Humanized Computing, 2021, 12: 5105-5116.

[6] Vikarna. Online ad click prediction. Kaggle. Retrieved December 23, 2022, https://www.kaggle.com/code/vikarna/online-ad-click-prediction/notebook#Logistic-Regression.

[7] Wikimedia Foundation. Random Forest. Wikipedia. Retrieved December 23, 2022, https://en.wikipedia.org/wiki/Random_forest.

[8] Support Vector Machines: A simple explanation. KDnuggets. (n.d.). Retrieved December 23, 2022, https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html.

[9] Support Vector Machines (SVM) algorithm explained. MonkeyLearn Blog. Retrieved December 23, 2022, from https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/.

[10] Bhavikkmuar, S. M. Advantages of support vector machines (SVM). OpenGenus IQ: Computing Expertise & Legacy. Retrieved December 23, 2022.