Predicting Advertisement Clicks

Gagana Uday Kumar (WOV796)

Background:

In today's world, businesses rely heavily on online advertising to reach potential customers. You've probably noticed ads popping up while browsing the internet or using social media. But here's the thing: not everyone clicks on those ads. Companies want to know who is more likely to click on their ads so they can target those people better and save money. However, just showing ads to the right people isn't enough. Companies want to know if people actually click on their ads.

In our project, I have data from a marketing agency about how people interact with online ads. I know things like their age, how much time they spend on the internet, and whether they clicked on an ad or not. The goal is to use this data to build a model that can predict if someone will click on an ad. If we can predict this well, companies can save money and show their ads to people who are more likely to click on them. We'll use different techniques for this prediction model.

Motivation:

Businesses spend a lot of money on online ads, but they don't always get the results they want. Sometimes, ads get ignored, and money goes down the drain. That's why it's important to find better ways to target ads to the right audience. By using data to predict who's likely to click on an ad, businesses can save money and make their ads more effective. Plus, understanding what makes people click on ads can help marketers create better ads in the future. So, by diving into this dataset, we're not just helping businesses improve their advertising game—we're also uncovering insights that can shape the future of online marketing.

Here are the few analysis that I will be focusing on:

1. Build a model with good accuracy to predict which user would click on the advertisement.
2. I would like to see if there is any correlation with time spent on the site and 'clicked on Ad'.
3. What is the average time spent on the site daily and average area income of users who click on the ad.
4. I would like to investigate if mean income changes with click on the ad.
5. I will check if total daily internet use and time spent by users on the site are related to each other in some way.
6. What are the general characteristics of the users who click on the ad.
7. Which day, which month, which date do users usually click on the ad.
8. If we divide the time into morning, afternoon and night, which part of the day do users click on ad?

Description of the Data

The dataset consists of 1000 observations and 10 variables related to user interactions with online advertisements. It includes metrics such as daily time spent on site, age, area income, daily internet usage, ad topic line, city, gender, country, timestamp, and whether the user clicked on the ad. Notably,

the dataset contains no missing values, making it suitable for various analyses and predictive modeling tasks aimed at understanding factors influencing ad engagement and optimizing online advertising strategies. The variable we are trying to predict in our model is Clicked on Ad, which has two possible outcomes clicked (1) or not clicked (0)

This dataset allows for analysis of various factors that may influence user interactions with online advertisements. The variables provide insights into user demographics, behavior, and geographic locations, which can be used to develop predictive models to forecast ad clicks.

```
> str(ad)
'data.frame':  1000 obs. of  10 variables:
 $ Daily.Time.Spent.on.Site: num  69 80.2 69.5 74.2 68.4 ...
 $ Age                     : int  35 31 26 29 35 23 33 48 30 20 ...
 $ Area.Income             : num  61834 68442 59786 54806 73890 ...
 $ Daily.Internet.Usage    : num  256 194 236 246 226 ...
 $ Ad.Topic.Line           : chr  "Cloned 5thgeneration orchestration" "Monit
ored national standardization" "Organic bottom-line service-desk" "Triple-buf
fered reciprocal time-frame" ...
 $ City                    : chr  "Wrightburgh" "West Jodi" "Davidton" "West
Terrifurt" ...
 $ Male                    : int  0 1 0 1 0 1 0 1 1 1 ...
 $ Country                 : chr  "Tunisia" "Nauru" "San Marino" "Italy" ...
 $ Timestamp               : chr  "3/27/2016 0:53" "4/4/2016 1:39" "3/13/2016
20:35" "1/10/2016 2:31" ...
 $ Clicked.on.Ad           : int  0 0 0 0 0 0 0 1 0 0 ...
```

Proposed Analysis

Below is my proposed analysis for each of the question I will be focusing on:

1. Build a model with good accuracy to predict which user would click on the advertisement:

I plan to develop a predictive model to determine the likelihood of a user clicking on the advertisement based on their attributes. I will split the data into training and testing sets, train the model on the training data, and evaluate its performance on the testing data using metrics like accuracy, precision, recall, and F1-score. We will employ various machine learning algorithms, including Logistic Regression, Decision Trees, Support Vector Machines (SVM), and Random Forest, to develop a predictive model that accurately predicts whether a user will click on the advertisement. By comparing the accuracy of these models, we aim to identify the most effective algorithm for the task.

2. Investigating correlation between time spent on the site and 'clicked on Ad':

I will begin by conducting a Shapiro-Wilk test to assess the normality of the distribution of daily time spent on the site. Next, Levene's test will be employed to evaluate the equality of variances. Since the data does not follow a normal distribution and exhibits unequal variances, I will utilize the non-parametric Mann-Whitney U test (Wilcoxon rank-sum test) to compare the median daily time spent on the site between individuals who clicked on an ad and those who did not. The Mann-Whitney U test will determine if there is a statistically significant difference in time spent on the site between the two groups. This analysis aims to elucidate the influence of time spent on the site on ad interaction,

providing valuable insights into user behavior and advertising effectiveness.

3. Average time spent on the site and area income of users who click on the ad:

We aim to calculate the average time spent on the site daily and the average area income of users who clicked on the ad. This involves filtering the data for users who clicked on the ad, computing the mean values for the relevant attributes, and interpreting the results to gain insights into the behavior and demographics of the target audience.

4. Investigating if mean income changes with ad clicks:

We will begin by examining the area income of users who clicked on the ad and those who did not. Shapiro-Wilk tests will be conducted to assess the normality of the distribution of area income within each group. Additionally, Levene's test will be employed to evaluate the equality of variances between the two groups. Since the data does not follow a normal distribution and exhibits unequal variances, we will perform a non-parametric Wilcoxon rank sum test (Mann-Whitney U test). This test will determine if there is a statistically significant difference in area income between individuals who clicked on the ad and those who did not. A p-value < 0.05 will indicate a significant difference in mean income, suggesting that mean income may influence ad interaction. This analysis aims to shed light on whether mean income plays a role in ad interaction, providing valuable insights into user behavior and advertising effectiveness.

5. Exploring the relationship between total daily internet use and time spent on the site:

I will analyze the relationship between total daily internet use and time spent on the site using correlation analysis. By calculating the Pearson correlation coefficient and visualizing the data with a scatter plot, we aim to assess the degree and nature of the relationship between these variables, providing insights into user engagement and behavior.

6. General characteristics of users who click on the ad:

I plan to examine various demographic and behavioral attributes of users who click on the ad, such as age, gender, income, etc. Descriptive statistics and data visualization techniques will be used to identify common characteristics and preferences among this user group, aiding in targeted advertising strategies.

7. Timing of ad clicks:

I will analyze the timestamp data to determine when users are most likely to click on the ad. By aggregating the data based on day of the week, month, and date. I aim to identify temporal patterns in ad engagement and optimize ad placement and scheduling accordingly.

8. Part of the day users click on ad:

To investigate the preferred timing for ad clicks among users, I will categorize the timestamp data into morning, afternoon, and night segments. A chi-square test of independence will be conducted to determine if there is a significant association between the time of day and ad clicks, providing valuable insights into user behavior patterns throughout the day.

References

Kaggle Dataset: "Predict Clicks on Ads"
Link: https://www.kaggle.com/code/kamalapousajja/predict-clicks-on-ads-logistic-regression/input
Description: This dataset, sourced from Kaggle, provides information on user interactions with online advertisements, including various demographic and behavioral attributes. It serves as a valuable resource for understanding ad engagement patterns and predicting ad clicks. The dataset enables us to explore factors influencing ad performance and develop predictive models to optimize online advertising strategies.

Appendix – Variable Descriptions

Daily Time Spent on Site: The amount of time (in minutes) a user spends on the website where the advertisement is displayed.
Age: The age of the user who interacts with the advertisement.
Area Income: The income level of the area where the user resides.
Daily Internet Usage: The amount of time (in minutes) a user spends on the internet daily.
Ad Topic Line: The title or topic of the advertisement displayed to the user.
City: The city where the user resides.
Male: Binary variable indicating whether the user is male (1) or not (0).
Country: The country where the user resides.
Timestamp: The date and time when the user interacted with the advertisement.
Clicked on Ad: The target variable indicating whether the user clicked on the advertisement (1) or not (0).