Developing a Predictive Model for Click Through Rate and comparing it with Real World Data

Name: Sneha Pal

CIN No: 22-300-4-07-0445

Registration No: A01-2112-0721-22

Session-2022-2025

Under the supervision of

Dr. Durba Bhattacharya

DEPARTMENT OF STATISTICS

ST. XAVIER’S COLLEGE (AUTONOMOUS), KOLKATA

30, Mother Teresa Sarani, Kolkata - 700016

DECLARATION

I affirm that I identify all my sources and that no part of my dissertation paper uses unacknowledged materials.

Sneha Pal.

Sneha Pal

Department of Statistics

St Xavier's College (Autonomous), Kolkata

Date: April 2025

ACKNOWLEDGEMENT

Throughout my research journey, I have received immense support and guidance, for which I am deeply grateful. I would like to extend my sincerest thanks and gratitude to my supervisor, Dr. Durba Bhattacharya for her priceless support and insights in the shaping of my research topic and the guiding of the process.

I would like also like to acknowledge and thank my professors at the Department of Statistics, St. Xavier's College, Kolkata. This includes Prof. Dr. Ayan Chandra, Prof. Dr. Surupa Chakraborty, Prof. Dr. Surabhi Dasgupta, Prof. Madhura Dasgupta, Prof. Dr. Sancharee Basak, Prof. Rahul Roy. Their mentorship has been crucial in cultivating a research mindset necessary to undertake this project.

My appreciation extends to the support and contributions of my peers, whose ideas and discussions enriched my research. I acknowledge all the researchers who have paved the way for future students to continue in the field of Statistics and Statistical Learning.

CONTENT

INTRODUCTION

Advertisement bridges the gap between Businesses and Consumers, facilitating their interactions. As, we navigate the bustling streets of modern cities, we are constantly surrounded by a plethora of advertising mediums, including billboards, posters, and digital displays. Advertising has become an integral component of our daily lives, exerting a profound influence on our purchasing decisions and shaping our perceptions of various brands.

However, it is essential to pause and reflect on the impact of advertising on our behaviour. Empirical research has consistently demonstrated that exposure to advertising can significantly enhance the likelihood of engagement with a particular brand or product.

In today's digital era, advertising has evolved to become increasingly targeted and personalized. When an individual clicks on an advertisement, it serves as a deliberate signal to the advertiser, indicating a genuine interest in the product or service being promoted.

Click-Through Rate (CTR) is a pivotal metric in online advertising, quantifying the probability that users will engage with an advertisement or hyperlink. By accurately predicting CTR, advertisers can optimize ad placements, enhance user experience, and ultimately maximize revenue. Effective CTR prediction enables advertisers to identify high-performing ad creatives, target relevant audiences, and allocate their budgets more efficiently.

These models can identify patterns and relationships between various data points, including user demographics, browsing history, and online behaviour, to predict the likelihood of a click.

OBJECTIVE OF THE STUDY

The aim of the study is to investigate the intricacies of Click-Through Rate (CTR) in digital advertising, seeking to understand the complex dynamics that govern user engagement with online advertisements. By delving into a rich, this research aims to illuminate the key factors that influence CTR, including demographic characteristics, internet usage patterns, and regional differences.

Additionally, a comparative analysis of CTR data with survey responses from college students, providing a nuanced understanding of ad interaction behaviours in a real-world context. By integrating these quantitative and qualitative perspectives, this research strives to develop actionable recommendations for optimizing digital advertising strategies.

Ultimately, the objective of this study is to empower advertisers with data-driven insights, enabling them to refine their targeting precision, enhance user engagement, and maximize conversion rates.

IMPORATANCE OF THE STUDY

This study on predicting Click-Through Rates (CTR) in digital advertising holds significant importance for several reasons:

1) **Optimized Advertising Strategies:** By identifying the key factors influencing CTR, advertisers can develop targeted and effective advertising strategies, which will lead to improved campaign performance and increase the return on investment (ROI).
2) **Enhanced User Experience:** Gaining insights into user interactions with online advertisements enables advertisers to design more engaging and relevant ad content, ultimately improving the overall user experience.
3) **Increased Revenue**: By accurately predicting CTR, advertisers can maximize their ad spend, leading to increased revenue and hence business growth.
4) **Contribution to Research:** This study aims to contribute to the existing body of research on digital advertising, providing new insights and perspectives on the factors influencing CTR.

SOURCES OF DATA

- **SECONDARY DATA:**

This study utilizes a publicly available dataset on Advertisement Click Prediction, sourced from Kaggle. The dataset comprises 10,000 records, providing a substantial foundation for analysis.

Dataset Characteristics:

- Source: Kaggle (https://www.kaggle.com/datasets/swekerr/click-through-rate-prediction)
- Number of Records: 10,000
- Variables: *Age (in years), Gender (Male/Female), Daily time spent on site (in minutes), Daily internet usage (in minutes), Area income (in Rupees), Timestamp, Ad topic line, City, Country, Clicked on Ad.*

- **PRIMARY DATA:**

Survey data is collected from 47 students of Department of Statistics and Postgraduate Department of Data Science from St. Xavier's College (Autonomous), Kolkata on the same variable to analyse the factors influencing Click-Through Rate (CTR) in digital advertising and compare with the survey data.

- **Survey Design:** A structured survey was developed to examine key variables affecting ad engagement. The survey also incorporated factors identified in the dataset to facilitate a comparative analysis between real-world responses and existing data.
- **Procedure for Collection:** The survey was distributed electronically using **Google Forms** to ensure broad participation. It was shared in Department of Statistics and Postgraduate Department of Data Science from St. Xavier's College (Autonomous), Kolkata.
- **Informed Consent:** Before participation, respondents were fully informed about the study's objectives, methodology, and confidentiality measures. Their consent was obtained to ensure voluntary and ethical participation. All collected responses were securely stored and systematically organized for further statistical analysis.

DESCRIPTION OF DATA

a. Secondary Data

The dataset consists data of 10000 internet users of different countries across the world to analyse the **Click Through Rate (CTR)** i.e, percentage of individuals who view a webpage and then click on a specific advertisement that appears on the webpage, measuring the success of an advertisement in capturing the user's attention.

Here the dataset contains 10 variables

Variables	***Description***
Daily Time Spent on Site	*Average time (in minutes) spent on Website each day indicating user engagement in a particular site.*
Age	*Age of customer in years; helps to understand the target audience.*
Area Income	*Average monthly income (in Rupees) of customer in that geographical area indicating economic status of target audience*
Daily Internet Usage	*Average time (in minutes) spent on internet each day indicating extent of online activity of consumer.*
Ad Topic Line	*Headline of the advertisement giving brief idea about the content of the advertisement to user*
City	*Indicates the city of consumer that's the geographical information that can be used for targeting the audience*
Gender	*Indicates the Gender – Male/ Female of the customer*
Country	*Indicates the country of consumer that's the geographical information that can be used for targeting the audience*
Time Stamp	*Digital record of time when the advertisement was shown to the consumer in day-month-year form*
Clicked on Ad	*Binary variable where 0 indicates the ad was not clicked, and 1 indicates it was clicked; a performance metrics for evaluating ad effectiveness and user engagement*

Transformation Of A Few Variables to Extract Necessary Information:

1) Ad topic line → Ad Category

Since Ad Topic Line is a vital information in understanding CTR, it was converted into a categorical variable having 11 categories and the variable was renamed as Ad Category. This conversion was possible using the concept of Bidirectional Encoder Representation from Transformers (BERT) Language Model. BERT labels sentences into different categories due to advance contextual understanding capabilities. It is developed by Google that processes text bidirectionally. The model used is BART - a transformer model combining a BERT-like bidirectional encoder with a GPT-like autoregressive decoder. Hugging Face has pre-trained it on Facebook data by corrupting text and then learning how to reconstruct it. BART does very well on text generation tasks like summarization and translation, but it also learns fairly well on comprehension tasks like classification and question answering.

The Ad Categories are as follows:

Information Technology, Software & Innovation

Finance, Banking & Investment

Marketing, Advertising & Branding

E-Commerce, Retail & Consumer Goods

Automotive, Transportation & Logistics

Environment, Sustainability & Renewable Energy

Business Strategy, Consulting & Management

Healthcare, Medicine & Wellness

Education, Learning & Skill Development

Entertainment, Media & Digital Content

Real Estate, Architecture & Construction

2) Country → Continent

The variable Country was used to understand the Geographical location and it is converted into a broader category of continents. Thus, we treat the 7 Continents as 7 different categories and use it for analysis. The 7 continents are Asia, Africa, Antarctica, North America, South America, Australia, Europe.

3) Timestamp → Time Category

Timestamp has been converted into 4 categories –

i) (0): Morning – 5 AM -12 PM

ii) (1): Afternoon – 12 PM – 4 PM

iii) (2): Evening – 4 PM – 8 PM

iv) (3): Night – 8 PM – 5 AM

4) Gender: The variable Gender is converted into binary response.

$$Gender = \begin{cases} 0, & if\ female \\ 1, & if\ male \end{cases}$$

5) **Clicked on Ad:** The variable Clicked on Ad is converted into binary response.

$$Clicked\ On\ Ad = \begin{cases} 0, & if\ No \\ 1, & if\ Yes \end{cases}$$

Hence our data now contains the following type of variables:

1) Qualitative Variable - a) Gender (Binary): Male (1), Female (0)
   b) Country → Continent (Nominal): 7 categories
   c) Clicked an Advertisement (Binary): Yes (1) or No (0).
   d) Ad topic line → Ad Category: 11 Categories
   e) City
   f) Timestamp → Time Category: 4 Categories
2) Quantitative Variable – a) Age (in years) (Continuous)
   b) Daily Time Spent on Site (in minutes) (Continuous)
   c) Daily Internet usage (in minutes) (Continuous)
   d) Area Income (in rupees)

b. Primary Data

The primary data included in this research plays a crucial role in conducting a comparative analysis between real-world user behaviour and the dataset obtained from Kaggle for Click-Through Rate (CTR) prediction. This section outlines the methodology used for primary data collection, sample selection, data characteristics, and ethical considerations.

1. Method used for Data Collection

The primary data was collected through a structured survey designed to capture key factors influencing CTR. The survey aimed to replicate the variables present in the secondary dataset while gathering real-time user responses. The survey was conducted online using Google Forms, ensuring ease of participation and accessibility.

2. Sample Selection

- **Target Population**: College students and young adults who frequently browse the internet and interact with online advertisements.
- **Sampling Method**: Convenience sampling was used, focusing on students from Department of Statistics and Postgraduate Department of Data Science at St. Xavier's College, Kolkata.
- **Sample Size**: A total of 47 respondents are been a part of the Survey, ensuring a sufficient dataset for statistical analysis.
- **Demographics**: Participants belonged to the age group 18–25 years, representing an important segment of digital consumers, belonging to Indian origin (Continent: Asia)

3. Data Characteristics

The survey questionnaire consisted to capture quantitative and qualitative insights. The key sections included:

- Demographic Information: Age, gender, location.
- Online Behaviour: Time spent on the internet and websites, frequency of ad clicks.
- Engagement with Advertisements: Preferred ad category
- Comparison with Secondary Dataset Variables: Similar variables were included to allow a direct comparison with the Kaggle dataset.
- **Informed Consent**: Respondents of the survey were informed about the purpose of the study and their voluntary participation and the data was not shared with any third parties.

4. Challenges and Limitations

- **Response Bias:** Some participants may have provided socially desirable answers rather than their actual behaviour.
- **Limited Sample Size:** While 47 respondents were surveyed, a larger sample could provide stronger statistical power.
- **Self-Reported Data:** Reliance on participants' recall of their online behaviour may introduce inaccuracies.

5. Purpose of Studying the Survey Data

The primary data is used to compare with the Kaggle dataset to determine whether real-world behaviour aligns with trends observed in pre-collected data. The study examines differences in CTR patterns, demographic influences, and the impact of online engagement across datasets. This comparative approach enhances the robustness of findings and provides practical insights into digital advertising strategies

PART-1: Building a Predictive Model

EXPLORATORY DATA ANALYSIS (EDA)

Section 1: Tabular Insights

EDA is the process of visualizing and analysing the data to discover patterns and anomalies present in data.

Table 1: Summary of Quantitative Variables

SUMMARY	DAILY INTERNET USAGE (in mins)	DAILY TIME SPENT ON WEBSITE (in mins)	AGE (in years)	AREA INCOME (in rupees)
Minimum	105.2	32.60	19	13996
1st Quartile	140.2	48.86	29	44052
Median	178.9	59.59	35	56181
Mean	177.8	61.66	35	53840
3rd Quartile	212.7	76.58	42	61840
Maximum	270.0	90.97	60	79332

Record of number of individuals in each category of the following variables:

Table 2: GENDER

CATEGORY	COUNT
MALE – 1	4624
FEMALE – 0	5376

TABLE 3: CLICKED ON AD

CATEGORY	COUNT
YES – 1	4917
NO - 0	5083

Table 4: TIME

CATEGORY	COUNT
MORNING (5 AM – 12 PM) - 0	1610
AFTERNOON (12 PM – 4 PM) – 1	1417
EVENING (4 PM – 8 PM) - 2	2471
NIGHT (8 PM – 5 AM) - 3	4502

Table 5: AD CATEGORY

CATEGORY	COUNT
AUTOMOTIVE, TRANSPORTATION & LOGISTICS – 0	145
BUSINESS STRATEGY, CONSULTING & MANAGEMENT – 1	2476
E-COMMERCE, RETAIL & CONSUMER GOODS - 2	36
EDUCATION, LEARNING & SKILL DEVELOPMENT – 3	345
ENTERTAINMENT, MEDIA & DIGITAL CONTENT – 4	1139
ENVIRONMENT, SUSTAINABILITY, & RENEWABLE ENERGY – 5	1467
FINANCE BANKING AND INVESTMENT – 6	416
HEALTHCARE, MEDICINE & WELLNESS – 7	435
INFORMATION TECHNOLOGY, SOFTWARE AND INNOVATION – 8	3308
MARKETING, ADVERTISEMENT & BRANDING – 9	9
REAL ESTATE, ARCHITECTURE & CONSTRUCTION – 10	224

Table 6: CONTINENT

CATEGORY	COUNT
ASIA (2)	2363
AFRICA (0)	2322
EUROPE (3)	2530
NORTH AMERICA (4)	1003
SOUTH AMERICA (6)	425
AUSTRALIA (5)	1262
ANTARCTICA (1)	95

Section 2: Visual Insights

a. UNDERSTANDING CONTINUOUS QUANTITATIVE VARIABLE USING HISTOGRAM & BOXPLOT

DISTRIBUTION OF DAILY TIME SPENT ON SITE (IN MINUTES)

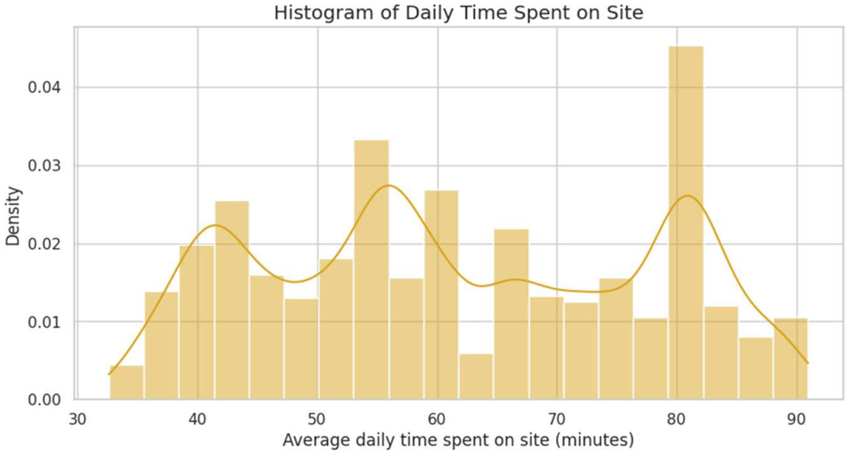


Fig1: Histogram of Daily time spent on Sites (in minutes)

This Histogram indicates a multimodal distribution with peaks at around 40-45, 55-60, 75-80 minutes. Thus, the data consists of information of individuals with different engagement on site.

Q: Do individuals who spend more time on site have higher chance on clicking ad?

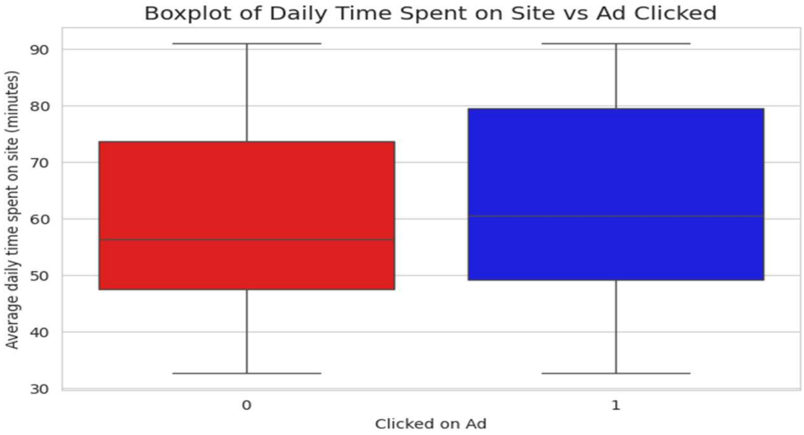


Fig2: The Boxplot represents the Daily time spent on site between users who click the ad (1) vs who did not click the ad (0).

We observe, the median time spent is slightly higher for who clicked ad (~60 mins) than those who didn’t clicked ad (~55 mins). The overall range for both categories seems more or less equal (~30 – 90 minutes). Also, the Interquartile range for category 1 is slightly higher than that of category 0, it indicates user who spend more time on site seems to more likely click on ad.

DISTRIBUTION OF DAILY INTERNET USAGE (IN MINUTES)

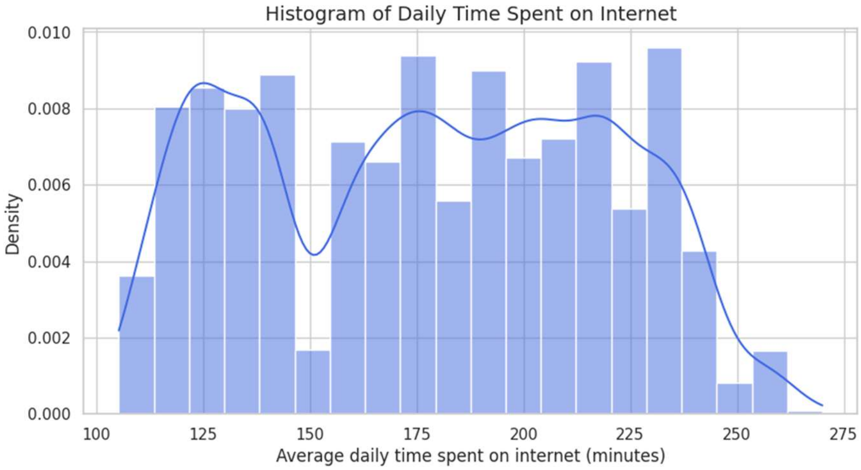


Fig 3:

This histogram represents a multimodal distribution with a spread from ~100 to ~270 minutes and a peak at ~130-140 minutes and another around ~240 minutes. It seems that there is a significant drop in frequency density around ~150 minutes segregating between low and heavy internet users.

Q: Do individuals who spend more time on internet have higher chance on clicking ad?

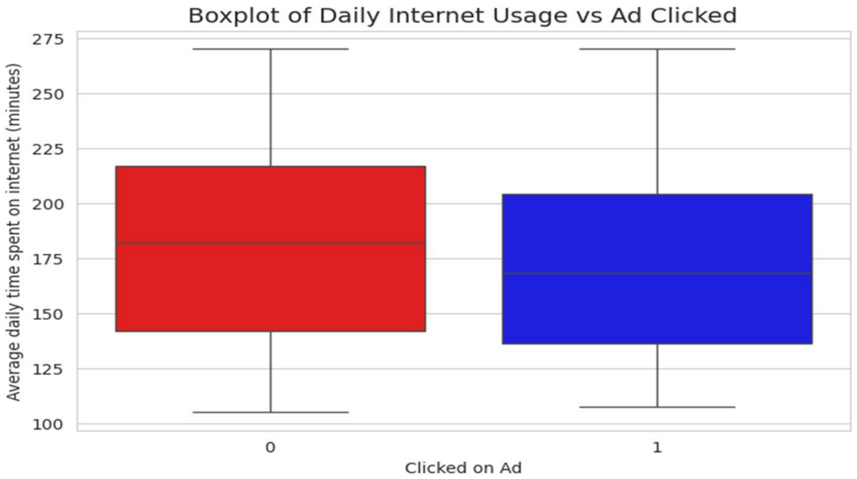


Fig4: The Boxplot represents the Daily time spent on Website between users who click the ad (1) vs who did not click the ad (0).

We observe, the median time spent is slightly higher for who didn’t clicked ad (~180 mins) than those who clicked ad (~165 mins). The overall range for both categories (~100– 270 minutes) and the interquartile range seems more or less equal. Daily Internet usage does not seem to affect ad click significantly.

DISTRIBUTION OF AGE (IN YEARS)

Q: Should advertisements be targeted toward a particular age group?

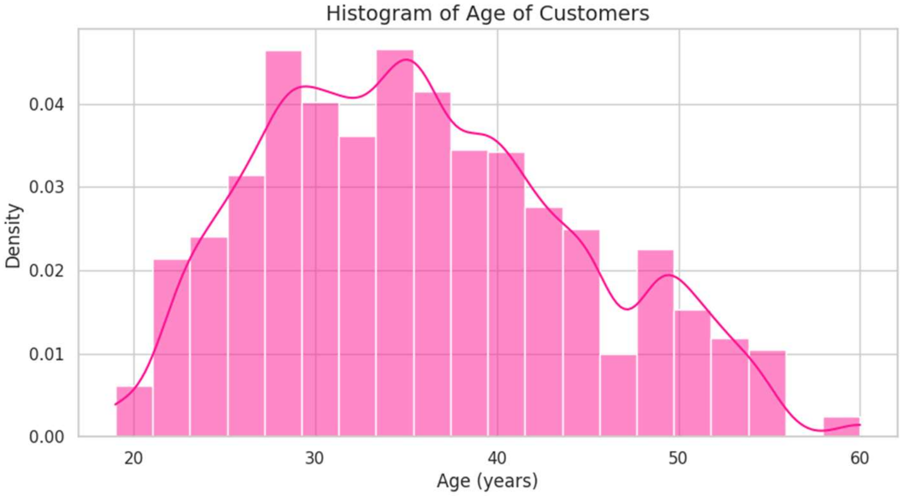


Fig5: Histogram of age of customers (in years)

This histogram depicts there are records of internet users from ~18-60 years with consecutive peaks around age 25-35 years. The distribution is slightly right skewed indicating a decrease in customer of higher age (approx. >45 years), specially very few customers above 55 years. This suggests that for better reach advertisements needs to be targeted for age group ~25-35 years.

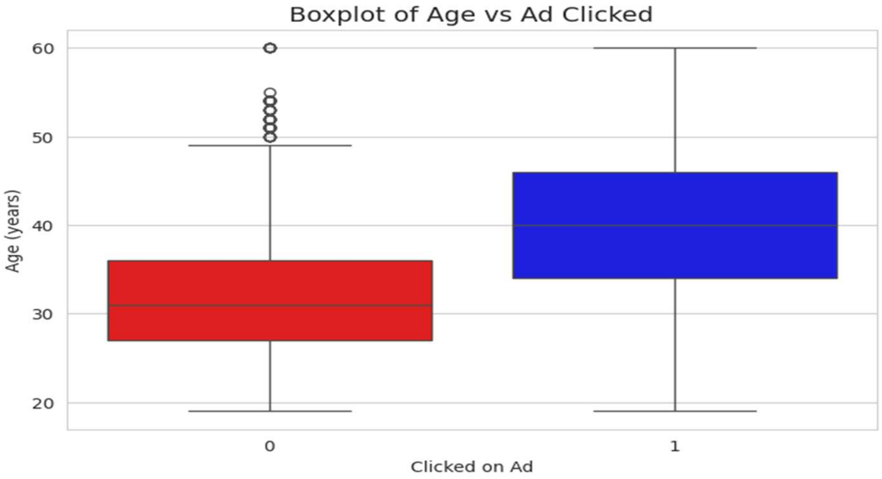


Fig6: The boxplot compares the age distribution between users who click the ad (1) vs who did not click the ad (0).

Although there are more individuals in age group 25-35 years, it seems from the boxplot that individuals of ~35-45 years tend to click on ad more often. The inter quartile range for category 1 is higher indicating more people are engaged in Ad clicking. People usually with age less than 35 years do not click on ad as evident from boxplot. Non clickers show several outliers above age 50 years indicating some users engage in ad less frequently.

DISTRIBUTION OF AREA INCOME (IN RUPEES)

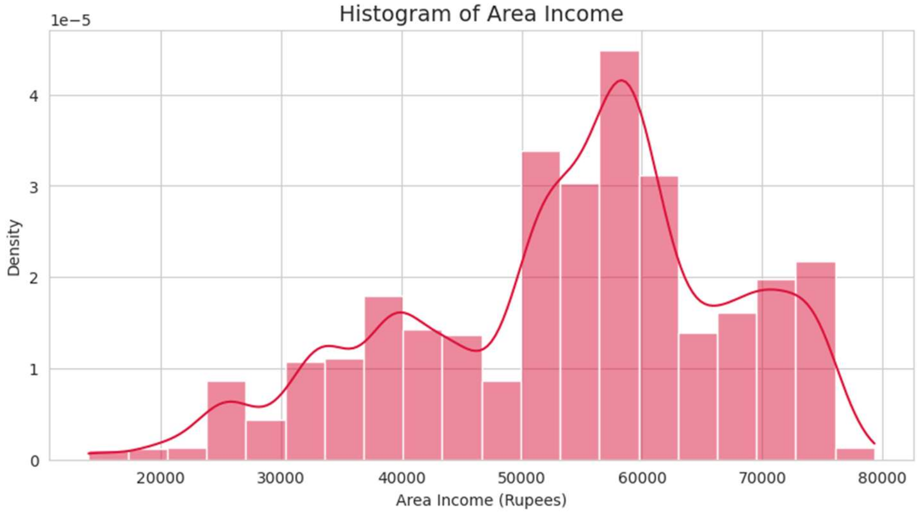


Fig7: Histogram of Area income (in rupees)

This histogram depicts the distribution of area income (in rupees) with a spread of ~18,000 – 80,000 rupees. The histogram seems to be slightly left skewed. The majority of audience falls in income group of 40,000 – 65,000 rupees. Thus, advertisements targeted for those age groups will be more effective.

Q: How does Area Income influence the response clicked on ad?

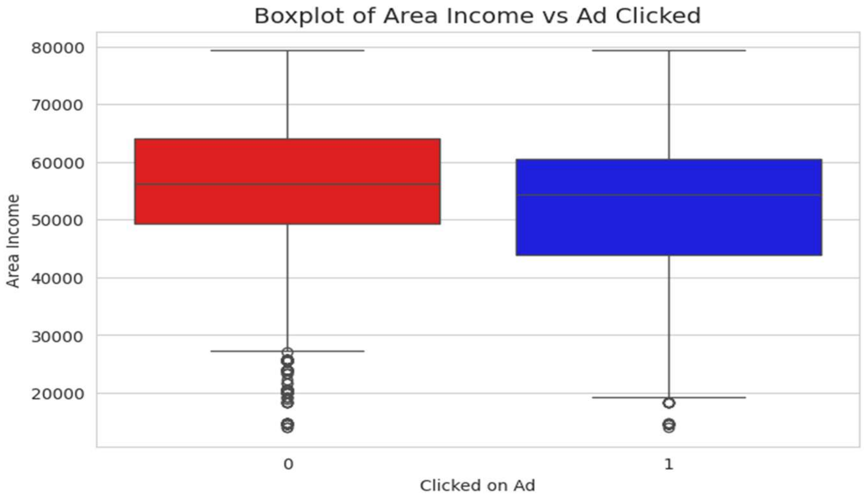


Fig8: The boxplot compares the area income between users who clicked the ad (1) vs who did not click the ad (0).

From the boxplot, it seems that Ad Clicked or not do not depend significantly on the Area income since the median and the interquartile range of the 2 categories are more or less equal. The first and third quartile of Category 1 is slightly less than Category 0. There are lower income outliers in both the group. The number of lower income outlier is more in Category 0. This could be possibly due to individuals having lower income do not click on ad due to lack of money to fulfil wishes beyond necessity.

b. UNDERSTANDING THE RELATION BETWEEN CONTINUOUS VARIABLES USING CORRELATION HEATMAP

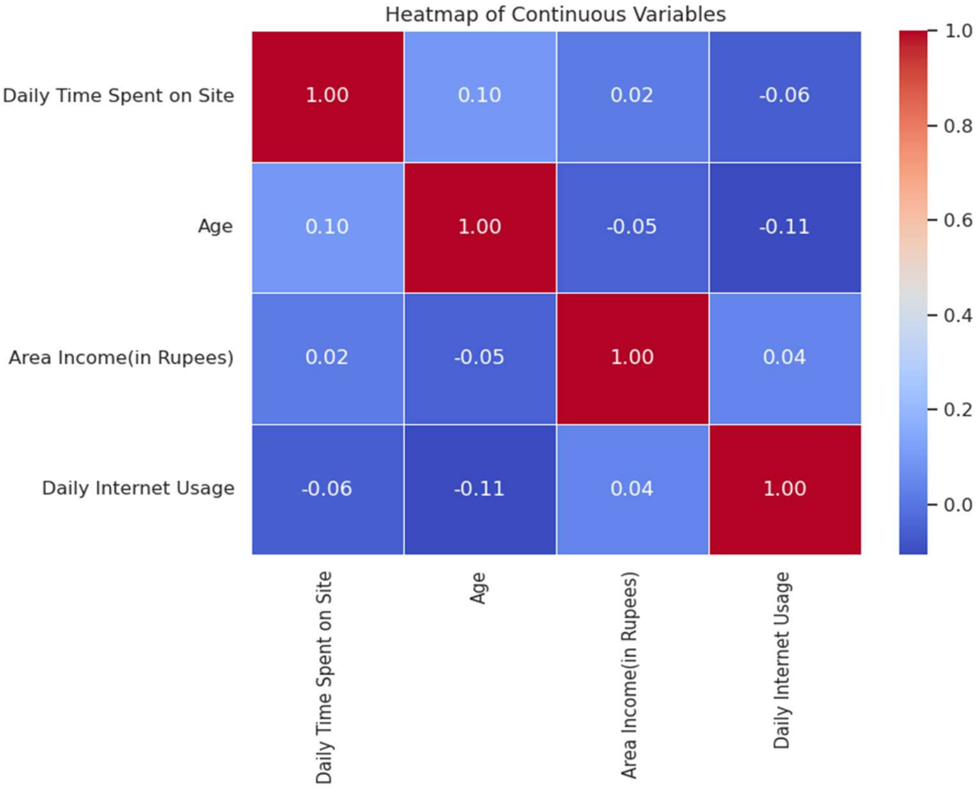


Fig9: Correlation Heatmap

Here we observe the nature of relationship between the quantitative variable of our dataset through Spearman Correlation Coefficient.

Interpretation:

1) Age displays a low negative correlation (-0.11) with daily internet usage, suggesting that younger individuals may spend a bit more time online.
2) Time spent daily and the area income of the region have an almost zero correlation coefficient, which means that the income of a region does not have a significant effect on how much time people spend on a site.
3) Daily time spent on site and age has a very weak positive correlation indicating on an average as age increases time spent on site also increases.

Overall, we observe from the correlation heatmap that the correlations are generally weak and the variables does not seem to influence each other.

c. UNDERSTANDING QUALITATIVE VARIABLE USING BARPLOT AND MOSAIC PLOT

ASSOCIATION BETWEEN GENDER AND CLICKED ON AD

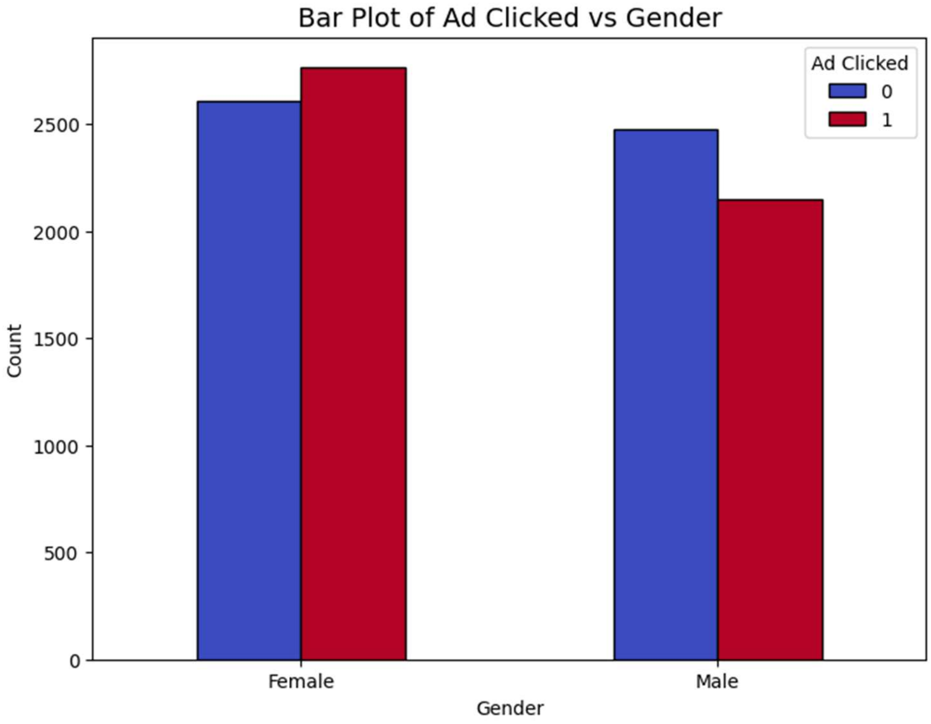


Fig10: Grouped Bar Plot of Gender and Clicked on Ad

Q: Is there any potential gender-based trend on Ad click?

From the grouped bar plot we observe that females are more responsive towards Ad click than males since the blue bar (Ad clicked = 0) is higher than the red bar (Ad clicked = 1) for female; whereas it's the opposite in case of male.

ASSOCIATION BETWEEN AGE CATEGORY AND CLICKED ON AD

We have converted the continuous variable Age into 4 categories:

- **Students:** Age 18-25 years
- **New Employees:** Age 25-35 years
- **Middle Aged:** Age 36-50 years
- **Aged:** Age 51-60 years

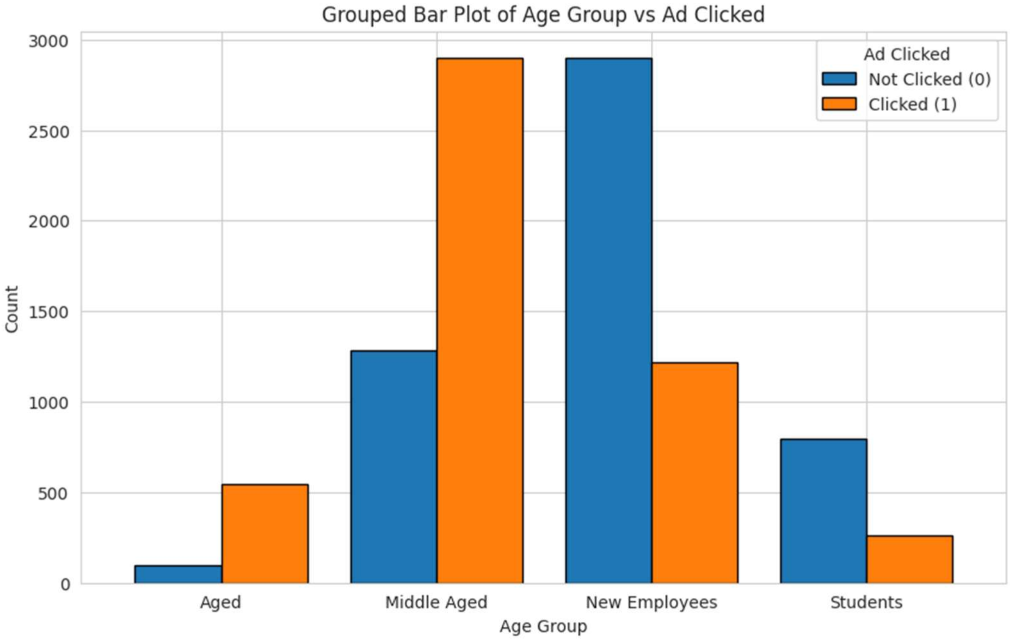


Fig11: Grouped Bar Plot of Age Category Vs Clicked on Ad

Q: Which Age Category seems to have high Ad Engagement?

From the Bar Plot, it seems that New Employees see the most Ads. They have the highest number of non-clicks but still a significant number of clicks. Middle Aged people seem to have the highest response to Ads as the number of clicks are much higher than the number of non-clicks.

Students and Aged People are comparatively less in number and seem to have a low response towards ad. This could imply lower interest or lower exposure to ads.

We will try to understand the pattern of Ad engagement in Student age group with the help of real-life data and compare with the dataset as they are supposed to be more engaged with online ads.

ASSOCIATION BETWEEN TIME CATEGORY AND CLICKED ON AD

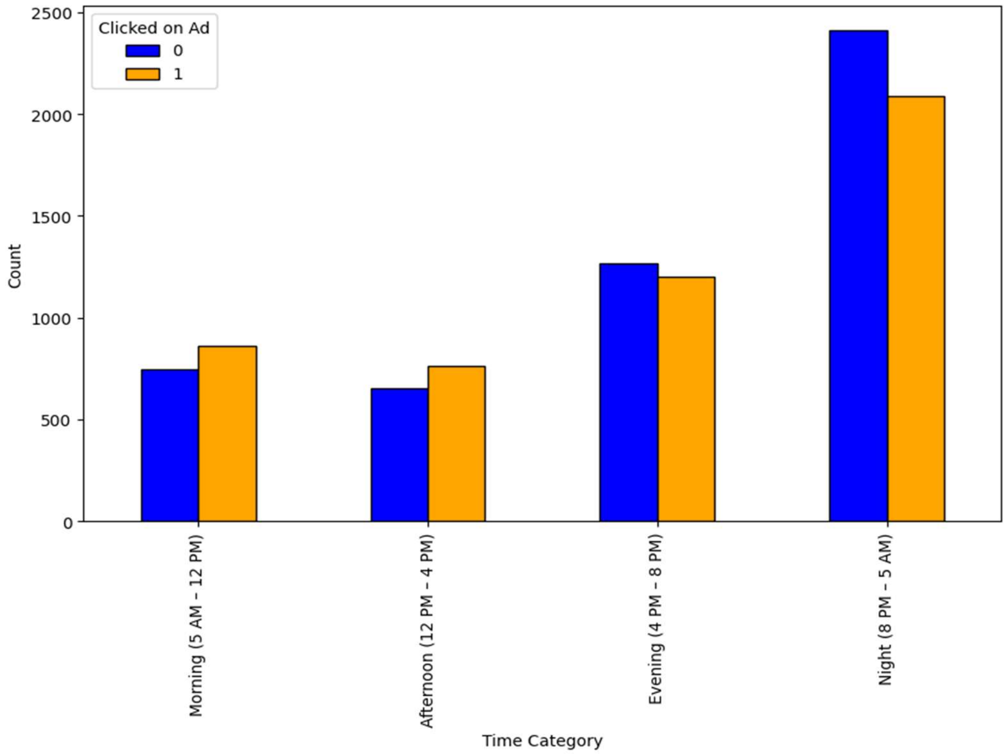


Fig12: Grouped Bar Plot of Time Category and Clicked on Ad

Interpretation:

1) Here we observe that more users are engaged on internet during night hours but the number of ad-clicked are lower than that of ad not clicked.
2) Morning and Afternoon shows a relatively better ad click through rate since more ad were clicked than ignored. Although the number of users active during Morning and Afternoon are relatively less.
3) Evening has balanced ad engagement since comparing with morning and afternoon, more people are active during this period as well as the click through rate is high.

ASSOCIATION BETWEEN CONTINENT CATEGORY AND CLICKED ON AD

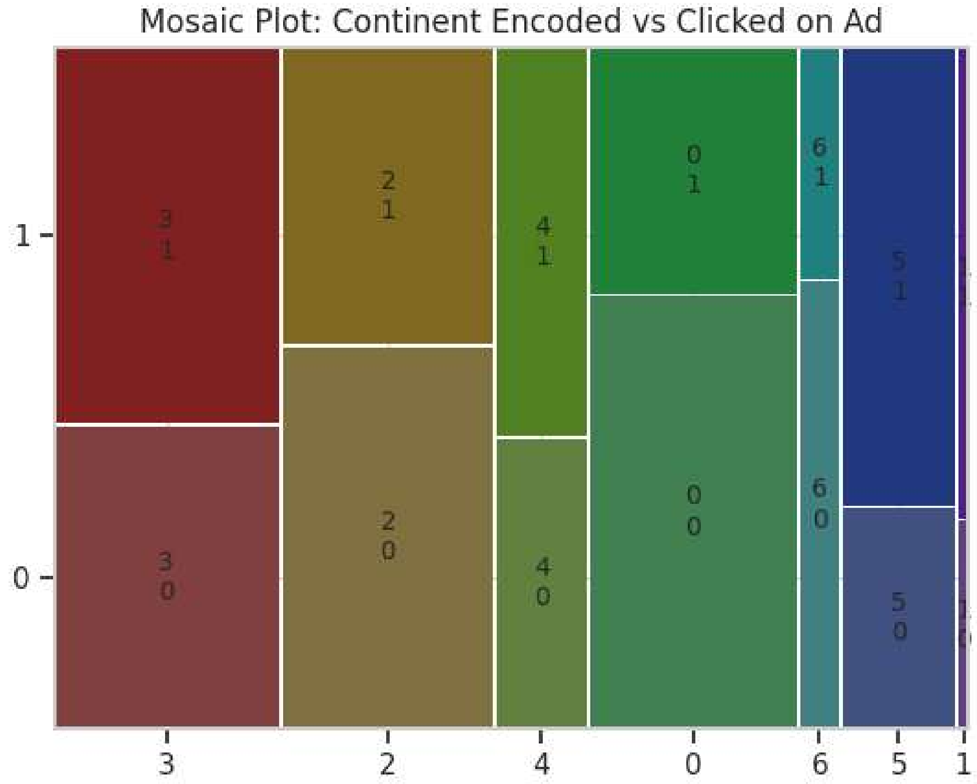


Fig13: Mosaic Plot of Continent Category and Clicked on Ads

Here, the X-axis represents different continents, categorized as per Table 6 and the Y-axis represents whether an Internet user clicked on an ad (1 = Clicked, 0 = Not Clicked).

The width of each continent's section represents the proportion of individuals from that continent in the dataset.

The height of the "Clicked" section within each continent reflects the Click-Through Rate (CTR) for that region.

Interpretation:

- **Europe (3), North America (4), and Australia (5)** show a higher proportion of clicks compared to non-clicks. This suggests that users from these continents have a higher engagement rate with ads.
- Other continents **(Asia, Africa, South America)** likely have lower CTRs, implying lower ad responsiveness.
- The size of each section suggests that some continents have a larger sample representation, which could influence overall trends.

Thus, we conclude **Ad engagement is not uniform across continents**.

PAIR PLOT

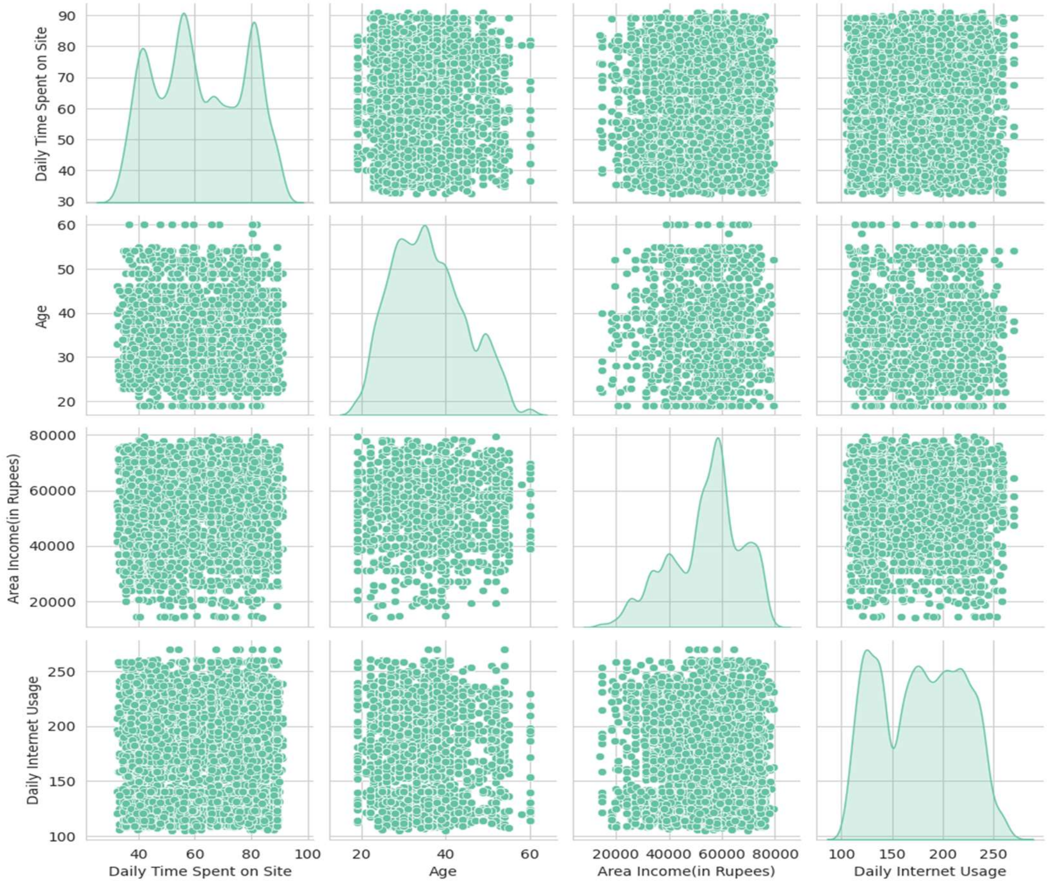


Fig14: Pair Plot of continuous variable

Diagonal Elements: These are KDE (Kernel Density Estimation) plots, showing the distribution of each variable.

- Daily Time Spent on Site and Daily Internet Usage seem to have multimodal distributions.
- Age has a right-skewed distribution.
- Area Income shows a peak around the middle-income range.

Scatter Plots:

- The off-diagonal elements show scatter plots representing relationships between variables.
- The scatter plots mostly show no strong correlations, as points appear widely spread.
- Age vs. Area Income may have some skewed trends, where younger people tend to have lower incomes.
- Daily Internet Usage vs. Age suggests that younger individuals tend to have higher internet usage.

DATA MODEL IMPLEMENTATION

Dropped Column

Columns – Country, City, Timestamp have been dropped from the dataset extracting all important and relevant information from these variables.

Splitting the data

The dataset has been divided into two parts – train and test data. 80 percent of the total dataset has been randomly assigned to training set and remaining 20 percent is allocated to testing set.

- Train data – It contains the majority of the dataset and is used to analyse the pattern and relationship of the different features with target variable. A logistic regression is fitted to the training set.
- Test data - The test dataset includes the remaining portion of the dataset and it is used to evaluate the performance of the model. The fitted logistic regression model is applied to this test data to assess its accuracy and generalization ability.

Predictor and Response Variable

To fit a multiple logistic regression model, we use the following predictor variables

Predictor variables:

- X_1: Daily time spent on site (in minutes)
- X_2: Age (in years)
- X_3: Area Income (in rupees)
- X_4: Daily internet usage (in minutes)
- X_5: Ad Category (*0,1,2,…,10*)
- X_6: Time Category (*0,1,2,3*)
- X_7: Gender (*0,1*)
- X_8: Continent (*0,1,2,…,6*)

Response variable:

- Y: Clicked on Ad

FITTING OF THE REGRESSION MODEL

The relationship between the response, or dependent variable(s), and the predictor, or independent variable(s), can be understood by regression analysis. Logistic regression is used to predict the likelihood of an event happening. Since the response variable in this study is **binary**, taking values **0 (No Click) and 1 (Click)**, **logistic regression** is used to fit the model.

We have, Y as the response variable and X_1, X_2, X_3, X_4, X_5, X_6, X_7 and X_8 are the covariates.
Let p be the probability of Y to take value 1.

i.e., $Y = \begin{cases} 1, & with\, probability\, p \\ 0, & with\, probability\, (1-p) \end{cases}$

Therefore, Y ~ Bern (p). For Y following a Bernoulli distribution with success probability p, the Probability Mass Function (PMF) is:

$$P(Y=k) = p^k\,(1-p)^{\,(1-k)}$$

which can be rewritten as

$$P(Y=k) = \begin{cases} p, & if\ k=1 \\ (1-p), & if\ k=0 \end{cases}$$

with E (*Y*) = *p* and Var(*X*)=*p(1−p)*

The logit link is given by

$$\mathrm{Ln}\left(\frac{p}{1-p}\right) = \text{f}\,(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

where $f\,(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = \beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+\sum_{j=1}^{10}\beta_{5j} X_{5j}+\sum_{j=1}^{3}\beta_{6j} X_{6j}+\beta_7 X_7+\sum_{j=1}^{6}\beta_{8j}\ X_{8j}+\varepsilon$

$$X_{5j} = \begin{cases} 1, & if\ jth\ category\ of\ Ad \\ 0, & Otherwise \end{cases} \quad \forall\ \text{j= 1,2,…,9,10}$$

$$X_{6j} = \begin{cases} 1, & if\ jth\ Time\ Category \\ 0, & Otherwise \end{cases} \quad \forall\ \text{j= 1,2,3}$$

$$X_{8j} = \begin{cases} 1, & if\ jth\ Continent \\ 0, & Otherwise \end{cases} \quad \forall\ \text{j= 1,2,…,5,6}$$

Here, 0th Category is considered to be the reference point. $\mathcal{E}$ is the random error associated with the model

$f\,(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon$

where $\mathbf{X}^{\mathrm{T}} = [X_1, X_2, X_3, X_4, X_{51}, X_{52}, \ldots, X_{59}, X_{5\,10}, X_{61}, X_{62}, X_{63}, X_7, X_{81}, \ldots, X_{85}, , X_{86}]$

$\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_4, \beta_{51}, \beta_{52}, \ldots, \beta_{59}, \beta_{5\,10}, \beta_{61}, \beta_{62}, \beta_{63}, \ \beta_7, \beta_{81}, \ldots, \beta_{85}, \beta_{86}]^{\mathrm{T}}$

β_0, β_1, β_2, β_4, β_{5j} (j ϵ {1,2,3,4,5,6,7,8,9,10}), β_{6j} (j ϵ {1,2,3}), β_7, β_8 (j ϵ {1,2,3,4,5,6}) are the regression parameters.

There are 8000 (say, n) observations in the training dataset. The logit link for ith observation is given by:

$$\ln\left(\frac{pi}{1-pi}\right) = \mathbf{X}^{\mathrm{T}}{}_i\boldsymbol{\beta}_i + \varepsilon_i \qquad \forall \ \ i=1(|)n \qquad \text{-(a)}$$

where p_i denotes the probability of ith individual to click on an Ad, $\mathbf{X}^{\mathrm{T}}{}_i$ denotes the value corresponding to vector ***X*** for ith individual and $\boldsymbol{\beta}_i$ be the corresponding parameter vector. ε_i denotes the random error corresponding to ith individual.

In deterministic form, equation (a) can be rewritten as

$$\Rightarrow \qquad E(Yi) = pi = \frac{\exp(\mathbf{x}_i{}'\boldsymbol{\beta}_i)}{1+\exp(\mathbf{x}_i{}'\boldsymbol{\beta}_i)} \quad ; \forall \ \ i=1(|)n$$

The probability mass function of y_i is given by:

$$f_i(y_i) = p_i^{y_i}\,(1-p_i)^{(1-y_i)} \ ; \forall \ i=1(|)n \ \& \ \ y_i \in \{0,1\}$$

The likelihood function is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(y_i)$$

Taking Logarithm on both sides we get,

$$l = ln(L(\boldsymbol{\beta})) = ln(\ \prod_{i=1}^{n} f_i(y_i))$$
$$= \sum_{i=1}^{n} y_i ln(p_i) \ + \ \sum_{i=1}^{n}(1-y_i)ln(1-p_i)$$
$$= \sum_{i=1}^{n} y_i ln(\frac{p_i}{1-p_i}) \ + \ \sum_{i=1}^{n} ln(1-p_i)$$

We fit the logistic regression model in the training dataset consisting of 8000 observations.

The parameters are estimated using Fischer's Scoring method.
The score equations are:

$$\sum_{i=0}^{n}(\mathrm{yi}-\mathrm{pi})=0$$

$$\sum_{i=0}^{n}(\mathrm{yi}-\mathrm{pi})X_{1i}=0$$

$$\sum_{i=0}^{n}(\mathrm{yi}-\mathrm{pi})X_{2i}=0$$

$$\sum_{i=0}^{n}(\mathrm{yi}-\mathrm{pi})X_{3i}=0$$

$$\sum_{i=0}^{n}(\mathrm{yi}-\mathrm{pi})X_{4i}=0$$

$$\sum_{i=0}^{n}(\mathrm{yi}-\mathrm{pi})X_{5\ 10i}=0$$

$$\sum_{i=0}^{n}(\mathrm{yi}-\mathrm{pi})X_{51i}=0$$

$$\vdots$$

$$\sum_{i=0}^{n}(\mathrm{yi}-\mathrm{pi})X_{59i}=0$$

$$\sum_{i=0}^{n}(\mathrm{yi}-\mathrm{pi})X_{63i}=0$$

$$\sum_{i=0}^{n}(\mathrm{yi}-\mathrm{pi})X_{61i}=0$$

$$\sum_{i=0}^{n}(\mathrm{yi}-\mathrm{pi})X_{62i}=0$$

$$\sum_{i=0}^{n}(\mathrm{yi}-\mathrm{pi})X_{7}=0$$

$$\sum_{i=0}^{n}(\mathrm{yi}-\mathrm{pi})X_{86i}=0$$

$$\sum_{i=0}^{n}(\mathrm{yi}-\mathrm{pi})X_{81i}=0$$

$$\vdots$$

$$\sum_{i=0}^{n}(\mathrm{yi}-\mathrm{pi})X_{85i}=0$$

We now obtain the Standard error of our estimates using **Fisher Information matrix.** The (24+ 1) × (24 + 1) matrix is obtained as:

$$\begin{bmatrix} I_{11} & \cdots & I_{1\ 25} \\ \vdots & \ddots & \vdots \\ I_{25\ 1} & \cdots & I_{25\ 25} \end{bmatrix}$$

Where I_{kl} = - $E[\frac{d^2\, l}{d\beta_k\beta_l}]$; *k, l* = 1(|)25

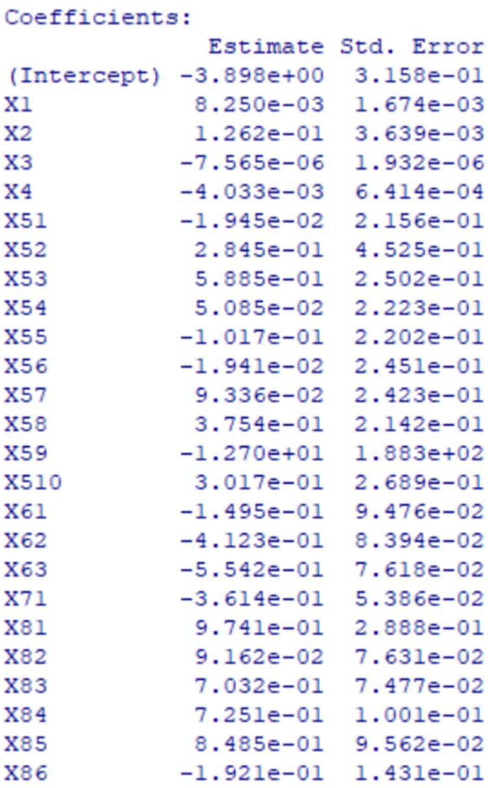
On inverting the Fisher Information matrix (by using numerical method) we get,

$$\begin{bmatrix} I^{11} & \cdots & I^{1\ 25} \\ \vdots & \ddots & \vdots \\ I^{25\ 1} & \cdots & I^{25\ 25} \end{bmatrix}$$

The standard errors of the estimates are given by

SE (β_{k}) = $\sqrt{I^{kk}}$ ∀ k= 1(|)25

The estimate of the parameters and standard errors for the fitted logistic regression is given by:



```
Coefficients:
              Estimate Std. Error
(Intercept) -3.898e+00  3.158e-01
X1           8.250e-03  1.674e-03
X2           1.262e-01  3.639e-03
X3          -7.565e-06  1.932e-06
X4          -4.033e-03  6.414e-04
X51         -1.945e-02  2.156e-01
X52          2.845e-01  4.525e-01
X53          5.885e-01  2.502e-01
X54          5.085e-02  2.223e-01
X55         -1.017e-01  2.202e-01
X56         -1.941e-02  2.451e-01
X57          9.336e-02  2.423e-01
X58          3.754e-01  2.142e-01
X59         -1.270e+01  1.883e+02
X510         3.017e-01  2.689e-01
X61         -1.495e-01  9.476e-02
X62         -4.123e-01  8.394e-02
X63         -5.542e-01  7.618e-02
X71         -3.614e-01  5.386e-02
X81          9.741e-01  2.888e-01
X82          9.162e-02  7.631e-02
X83          7.032e-01  7.477e-02
X84          7.251e-01  1.001e-01
X85          8.485e-01  9.562e-02
X86         -1.921e-01  1.431e-01
```

Thus, the estimated parameters are as follows:

TABLE 7: ESTIMATED PARAMETERS OF LOGISTIC REGRESSSION

PARAMETER	ESTIMATES	PARAMETER	ESTIMATES	PARAMETER	ESTIMATES
$\widehat{\beta_0}$	-3.898	$\widehat{\beta_{54}}$	0.5085	$\widehat{\beta_{62}}$	-0.4123
$\widehat{\beta_1}$	0.00825	$\widehat{\beta_{55}}$	-0.1017	$\widehat{\beta_{63}}$	-0.5542
$\widehat{\beta_2}$	0.1262	$\widehat{\beta_{56}}$	-0.01941	$\widehat{\beta_7}$	-0.3614
$\widehat{\beta_3}$	-0.00000756	$\widehat{\beta_{57}}$	0.09336	$\widehat{\beta_{81}}$	0.9741
$\widehat{\beta_4}$	-0.004033	$\widehat{\beta_{58}}$	0.3754	$\widehat{\beta_{82}}$	0.09162
$\widehat{\beta_{51}}$	-0.01945	$\widehat{\beta_{59}}$	-12.70	$\widehat{\beta_{83}}$	0.7032
$\widehat{\beta_{52}}$	0.2845	$\widehat{\beta_{5\ 10}}$	0.3017	$\widehat{\beta_{84}}$	0.7251
$\widehat{\beta_{53}}$	0.5885	$\widehat{\beta_{61}}$	-0.1495	$\widehat{\beta_{85}}$	0.8485
				β_{86}	-0.1921

Hence the fitted model is

$$\widehat{Y} = \widehat{\beta_0}+\widehat{\beta_1}X_1+\widehat{\beta_2}X_2+\widehat{\beta_3}X_3+\widehat{\beta_4}X_4+\widehat{\beta_{51}}X_{51}+\widehat{\beta_{52}}X_{52}+\widehat{\beta_{53}}X_{53}+\widehat{\beta_{54}}\ X_{54}+\widehat{\beta_{55}}\ X_{55}+\widehat{\beta_{56}}\ X_{56}+\widehat{\beta_{57}}\ X_{57}$$
$$+\widehat{\beta_{58}}\ X_{58}+\ \widehat{\beta_{59}}\ X_{59}+\widehat{\beta_{5\ 10}}\ X_{5\,10}+\widehat{\beta_{61}}\ X_{61}+\widehat{\beta_{62}}\ X_{62}+\widehat{\beta_{63}}\ X_{63}+\widehat{\beta_7}\ X_7+\widehat{\beta_{81}}\ X_{81}+\widehat{\beta_{82}}X_{82}+\widehat{\beta_{83}}X_{83}$$
$$+\widehat{\beta_{84}}X_{84}+\widehat{\beta_{85}}X_{85}+\beta_{86}X_{86}$$

Interpretation of the estimates of the parameters

(a) Intercept: $\widehat{\beta_0}$=−3.898

- When all predictor variables are absent, the log-odds of the event occurring is -3.898.
- This corresponds to an actual probability of:

  P(*Y*=1|***X***=**0**) = p = $\frac{exp(-3.898)}{1+ex\ \ (-3.898)}$ ≈ 0.0199 (very low probability)

  Meaning, the baseline probability of the event occurring is **about 1.99%**.

(b) Individual Predictors: Each coefficient ($\widehat{\beta_i}$) represents the effect of that predictor on the log-odds of the outcome.

For continuous variable, β_i represents the **log-odds change** for a **unit increase** in Xi, keeping all other variables constant.

- The estimated value of β_1 is 0.00825 i.e., Keeping other variables constant, for one minute increase in **Daily Time Spent on Site,** the log-odds of **Clicking an Ad** *slightly increases.*
- The estimate of β_2 is 0.1262 i.e., Keeping other variables constant, for one year increase in **Age,** the log-odds of **Clicking an Ad** increases significantly. Thus, age seems to be a significant variable influencing the click through rate.
- The estimate of β_3 is -0.00000756 i.e., Keeping other variables constant, for one rupee increase in **Area Income**, the log-odds of **Clicking an Ad** *very slightly decreases, almost insignificant.*
- The estimate of β_4 is -0.004033 i.e., Keeping other variables constant, for one minute increase in **Daily Internet Usage**, the log-odds of **Clicking an Ad** *very slightly decreases.*

  The **continuous variables** (X_1, X_2, X_3, X_4) generally have **small** effects on the odds of the event occurring. Their coefficients are close to **zero**, meaning that their influence on the outcome is relatively weak. However, some may still be statistically significant.

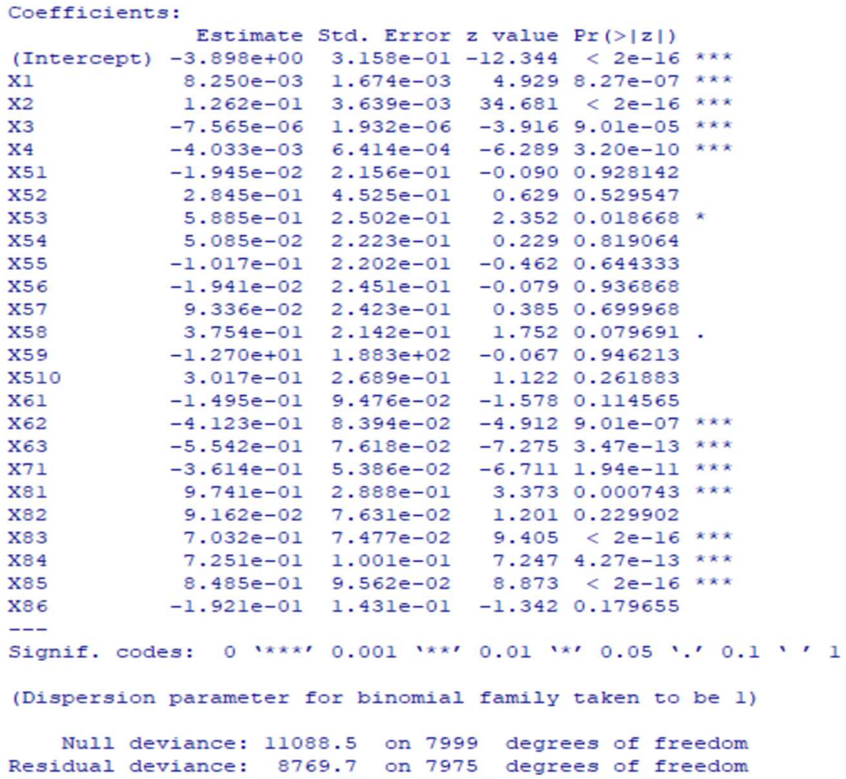
  For categorical variables (dummy-coded), β_i represents the log-odds change for being in a particular category (compared to the reference category), keeping other variables constant.

- The **categorical variables** (X_{51}, X_{52} ,…,X_{86}) show **stronger effects** compared to the continuous variables. Since categorical variables are encoded using dummy variables, their coefficients represent the effect of belonging to a particular category compared to a reference category. Some categorical variables have **positive coefficients**, indicating that being in that category **increases the likelihood** of the event occurring, while others have **negative coefficients**, meaning they **decrease the likelihood**.

- β_{59} has a **highly negative estimate (-12.70**), suggesting that the **9th Ad Category** (with reference to the 0th Category of Ad) is strongly associated with **very low odds** of **Clicking on Ad**. In practical terms, this means that individuals belonging to this category are much less likely to experience the outcome compared to the reference category. This could be due to the fact that very less number of observation (=9) corresponds to Ad category – 9, Marketing Advertisement and Branding (as observed from table 5).

- Conversely, variables like β_{81}, β_{83}, β_{84} and β_{85} have **relatively high positive coefficients**, meaning that observations in these categories have significantly **higher odds** of the event occurring.

GOODNESS OF FIT



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.898e+00  3.158e-01 -12.344  < 2e-16 ***
X1           8.250e-03  1.674e-03   4.929 8.27e-07 ***
X2           1.262e-01  3.639e-03  34.681  < 2e-16 ***
X3          -7.565e-06  1.932e-06  -3.916 9.01e-05 ***
X4          -4.033e-03  6.414e-04  -6.289 3.20e-10 ***
X51         -1.945e-02  2.156e-01  -0.090 0.928142    
X52          2.845e-01  4.525e-01   0.629 0.529547    
X53          5.885e-01  2.502e-01   2.352 0.018668 *  
X54          5.085e-02  2.223e-01   0.229 0.819064    
X55         -1.017e-01  2.202e-01  -0.462 0.644333    
X56         -1.941e-02  2.451e-01  -0.079 0.936868    
X57          9.336e-02  2.423e-01   0.385 0.699968    
X58          3.754e-01  2.142e-01   1.752 0.079691 .  
X59         -1.270e+01  1.883e+02  -0.067 0.946213    
X510         3.017e-01  2.689e-01   1.122 0.261883    
X61         -1.495e-01  9.476e-02  -1.578 0.114565    
X62         -4.123e-01  8.394e-02  -4.912 9.01e-07 ***
X63         -5.542e-01  7.618e-02  -7.275 3.47e-13 ***
X71         -3.614e-01  5.386e-02  -6.711 1.94e-11 ***
X81          9.741e-01  2.888e-01   3.373 0.000743 ***
X82          9.162e-02  7.631e-02   1.201 0.229902    
X83          7.032e-01  7.477e-02   9.405  < 2e-16 ***
X84          7.251e-01  1.001e-01   7.247 4.27e-13 ***
X85          8.485e-01  9.562e-02   8.873  < 2e-16 ***
X86         -1.921e-01  1.431e-01  -1.342 0.179655    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11088.5  on 7999  degrees of freedom
Residual deviance:  8769.7  on 7975  degrees of freedom
```

NULL DEVIANCE - Value: 11,088.5 (on 7,999 degrees of freedom)

The null deviance represents the model with only the intercept (no predictor variables). It shows how well the response variable can be predicted without any independent variables. A higher null deviance suggests a lot of variation in the dependent variable.

RESIDUAL DEVIANCE - Value: 8,769.7 (on 7,975 degrees of freedom)

The residual deviance can be interpreted with respect to the null deviance. The null deviance is the deviance of the null model and is used to assess how well the predictor variables in the model predict the response. A lower residual deviance compared to the null deviance indicates that the predictor variables have improved the model fit.

DEVIANCE REDUCTION – Value: 11088.5 – 8769.7 =2,318.8 (difference between the null deviance and residual deviance)**.**

This reduction suggests that independent variables explain a substantial portion of the variation in dependent variable. In other words, the model is performing well in capturing the underlying patterns of the data. By examining these values, we gain insight into the effectiveness of our logistic regression model and the importance of our predictor variables.

TEST OF SIGNIFICANCE OF PREDICTORS - WALD TEST

We have obtained the estimates of the parameter, now our objective is to test which predictors are significant in determining whether an individual will click on an advertisement or not.

The testing problem is given by:

$$Hoj: \beta_j = 0 \ vs \ H1j: \beta_j \neq 0 \ \forall \ j = 2(1)25$$

The test statistic under Hoj is given by:

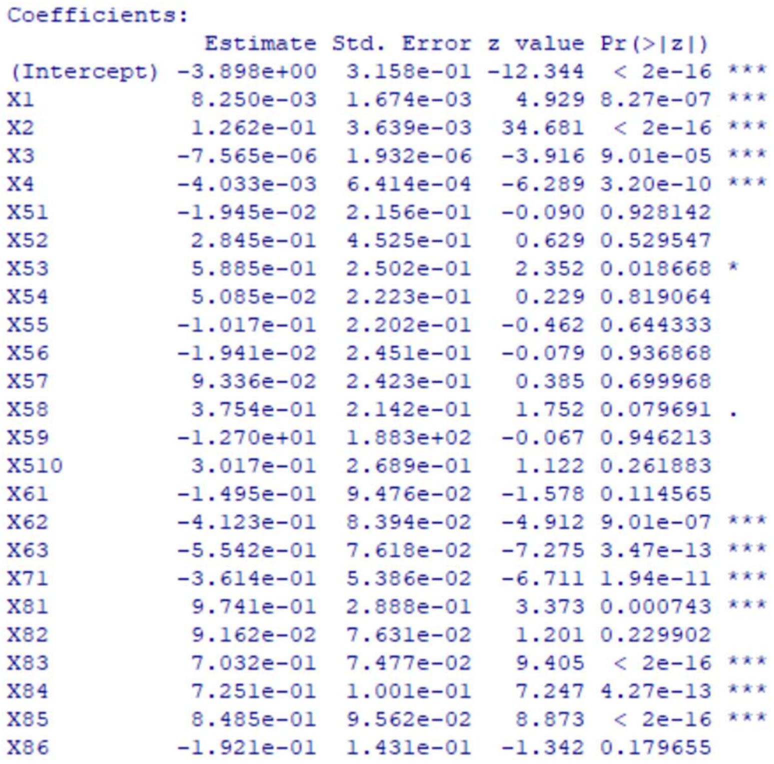
$$T_j = \frac{\widetilde{\beta_j}}{\boldsymbol{SE}(\widetilde{\beta_j})} \sim AN\,(0,1)$$

Critical Region: We will reject Hoj at α level of significance iff

$$|\ T_{j\ obs}| > \tau_{\alpha/2}$$

$$\text{For } \alpha = 0.05,\ \tau_{\alpha/2} = 1.96$$

The snapshot below gives the results obtained using R,



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.898e+00  3.158e-01 -12.344  < 2e-16 ***
X1           8.250e-03  1.674e-03   4.929 8.27e-07 ***
X2           1.262e-01  3.639e-03  34.681  < 2e-16 ***
X3          -7.565e-06  1.932e-06  -3.916 9.01e-05 ***
X4          -4.033e-03  6.414e-04  -6.289 3.20e-10 ***
X51         -1.945e-02  2.156e-01  -0.090 0.928142    
X52          2.845e-01  4.525e-01   0.629 0.529547    
X53          5.885e-01  2.502e-01   2.352 0.018668 *  
X54          5.085e-02  2.223e-01   0.229 0.819064    
X55         -1.017e-01  2.202e-01  -0.462 0.644333    
X56         -1.941e-02  2.451e-01  -0.079 0.936868    
X57          9.336e-02  2.423e-01   0.385 0.699968    
X58          3.754e-01  2.142e-01   1.752 0.079691 .  
X59         -1.270e+01  1.883e+02  -0.067 0.946213    
X510         3.017e-01  2.689e-01   1.122 0.261883    
X61         -1.495e-01  9.476e-02  -1.578 0.114565    
X62         -4.123e-01  8.394e-02  -4.912 9.01e-07 ***
X63         -5.542e-01  7.618e-02  -7.275 3.47e-13 ***
X71         -3.614e-01  5.386e-02  -6.711 1.94e-11 ***
X81          9.741e-01  2.888e-01   3.373 0.000743 ***
X82          9.162e-02  7.631e-02   1.201 0.229902    
X83          7.032e-01  7.477e-02   9.405  < 2e-16 ***
X84          7.251e-01  1.001e-01   7.247 4.27e-13 ***
X85          8.485e-01  9.562e-02   8.873  < 2e-16 ***
X86         -1.921e-01  1.431e-01  -1.342 0.179655    
```

Here the value of Z scores is equivalent to $T_{j\ obs}$

Decision Table:

PARAMETER	DECISION	PARAMETER	DECISION	PARAMETER	DECISION
X_1	REJECT	X_{55}	ACCEPT	X_{63}	REJECT
X_2	REJECT	X_{56}	ACCEPT	X_7	REJECT
X_3	REJECT	X_{57}	ACCEPT	X_{81}	REJECT
X_4	REJECT	X_{58}	ACCEPT	X_{82}	ACCEPT
X_{51}	ACCEPT	X_{59}	ACCEPT	X_{83}	REJECT
X_{52}	ACCEPT	$X_{5\ 10}$	ACCEPT	X_{84}	REJECT
X_{53}	REJECT	X_{61}	ACCEPT	X_{85}	REJECT
X_{54}	ACCEPT	X_{62}	REJECT	X_{86}	ACCEPT

Interpretation:

In the light of given data, it seems that Variables such as X_1, X_2, X_3, X_4, X_{63}, X_7, X_{81}, X_{83}, X_{84}, and X_{85} have a significant contribution to the model. The remaining variables does not seem to contribute significantly to the model.

Since most dummy variables associated with different ad categories (X_5 group: X_{51}, X_{52}, X_{53}, etc.) fail to reject the null hypothesis, it suggests that the type of advertisement may not have a significant influence on the likelihood of a user clicking on an ad.

We will now analyse the trained logistic regression model on the Testing Dataset and evaluate the different metrics in presence and absence of Ad Category.

EVALUATING MODEL PERFORMANCE ON TEST DATASET

We assess the performance of the logistic regression model using a confusion matrix, a 2×2 table that compares the model's predicted values with the actual values from the test dataset.
To determine the optimal probability threshold that maximizes accuracy, we use the optimal function in R.

- **In presence of variable Ad category:**

Optimal cut-off probability: 0.4924.

CONFUSION MATRIX

Predicted Actual	Ad not Clicked (0)	Ad not Clicked (1)
Ad not Clicked (0)	775	247
Ad Clicked (1)	272	706

From the confusion matrix we obtain,
TRUE POSITIVE RATE (TPR) : Measures the proportion of actual positives correctly identified.
TPR= 706/(706+272) = 0.7218

FALSE POSITIVE RATE (FPR) : Measures the proportion of actual negatives, incorrectly classified as positive.
FPR=247/ (247+775) = 0.2417

ACCURACY: Measures proportion of correctly classified observations.
Accuracy= (706+775) / (706+247+272+775)
= 0.7405

COMMENT:

1) TPR indicates that the model correctly identifies **72.18% of actual positive cases**. A reasonably high TPR suggests that the model is effective at identifying positive instances.
2) FPR indicates that **24.17% of actual negative cases** are incorrectly classified as positive. A lower value of FPR is desirable, and while this rate is not extremely high.
3) The model correctly classifies **74.05% of all observations**, which indicates a fairly good overall performance.

The model performs well, with a decent **true positive rate** and **overall accuracy**.

RECIEVER OPERATING CHARACTERISTIC (ROC) CURVE:

The graph is obtained by plotting TPR on vertical axis and FPR on horizontal axis**.** The line **TPR = FPR** indicates the **chance line** for a threshold with equal TPR and FPR where prediction is left to chance.

An ideal situation is one for which **TPR = 1 and FPR = 0.** However, in practice it is difficult to get such a predictor model.
In practice, we get a curve above the chance line and the model for which the ROC curve is the steepest is best for prediction purpose.

The steepness is usually measured by the Area Under Curve (AUC) i.e., Area covered by the curve and the Chance line.

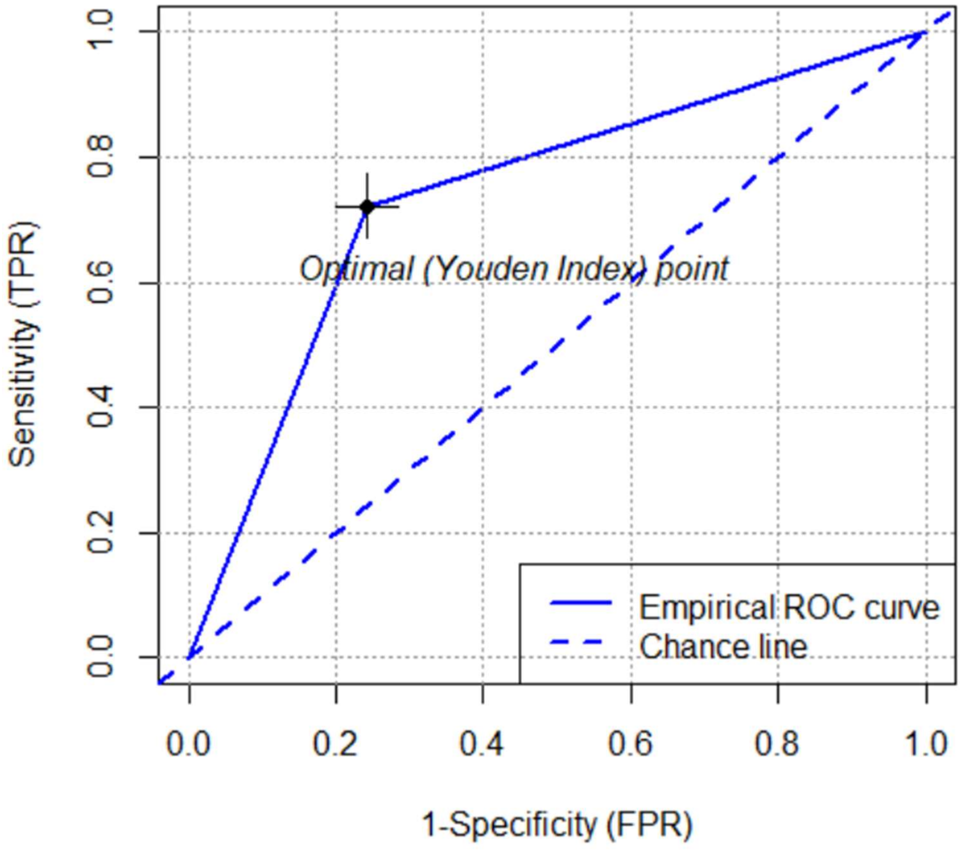


Fig15: ROC Curve for model-1

Area under curve (AUC) = 0.7401

COMMENT: An AUC of 74.01% suggests that the model has a moderately good ability to differentiate the two classes- Ad Clicked (1) & Ad not Clicked (0). The model seems to be significantly better than a random classifier (AUC = 0.5)

• In absence of variable Ad category:

Optimal cut-off probability: 0.4919

CONFUSION MATRIX

Predicted Actual	Ad not Clicked (0)	Ad not Clicked (1)
Ad not Clicked (0)	768	254
Ad Clicked (1)	271	707

From the confusion matrix we obtain,
TRUE POSITIVE RATE (TPR) : Measures the proportion of actual positives correctly identified.
TPR= 707/(707+271)= 0.7229

FALSE POSITIVE RATE (FPR) : Measures the proportion of actual negatives, incorrectly classified as positive.
FPR= 254/(768+254) = 0.2485

ACCURACY: Measures proportion of correctly classified observations.
Accuracy= (707+768) / (254+768+271+707)
= 0.7375

COMMENT:

4) TPR indicates that the model correctly identifies **72.29% of actual positive cases**. A reasonably high TPR suggests that the model is effective at identifying positive instances. The TPR is slightly higher than previous model.
5) FPR indicates that **24.85% of actual negative cases** are incorrectly classified as positive. The FPR is slightly higher than previous model.
6) The model correctly classifies **73.75% of all observations**, which indicates a fairly good overall performance, though not better than the previous model.

ROC CURVE

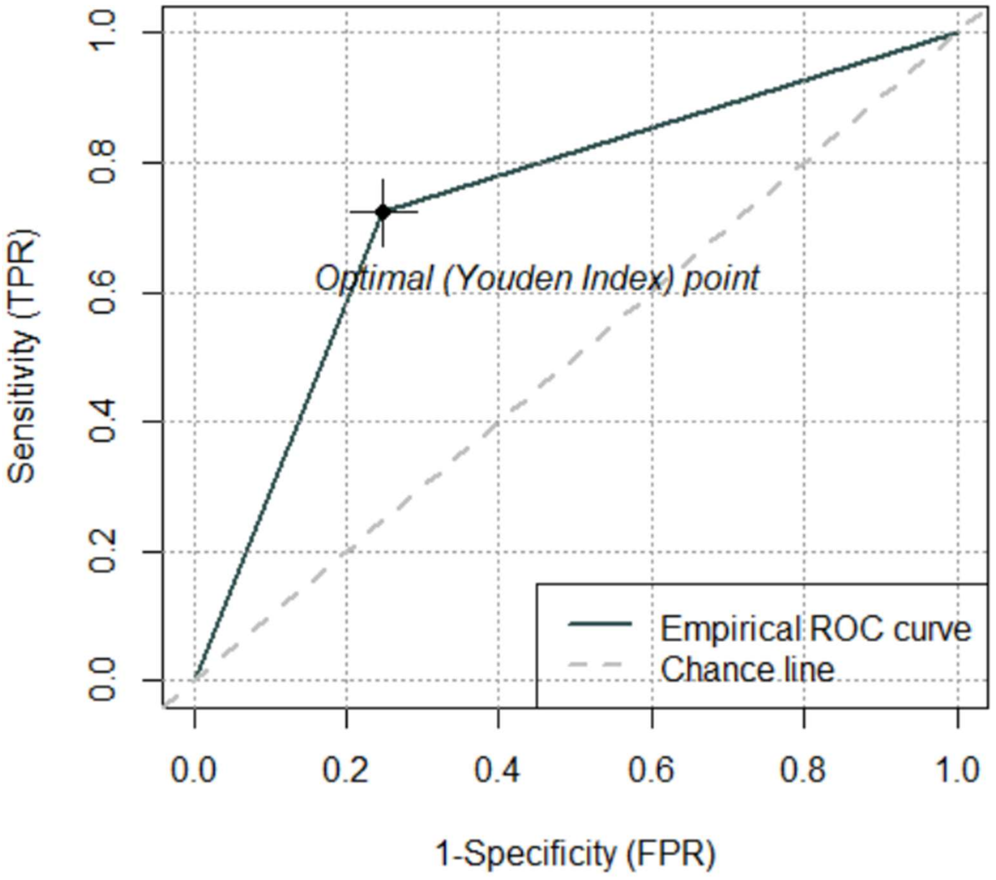


Fig16: ROC Curve for model-2

Area under curve (AUC) = 0.7372

COMMENT: An AUC of 73.72% suggests that the model has a moderately good ability to differentiate the two classes- Ad Clicked (1) & Ad not Clicked (0). The model seems to be significantly better than a random classifier (AUC = 0.5).

From the TPR, FPR, Accuracy and AUC, it seems that Model 1: In presence of Ad Category, performs slightly better than Model 2: In absence of Ad Category.

PART-2:

Studying the Primary Dataset to Understand Real World Behaviour

In this section, we study the ad click behaviour of students from the Department of Statistics and the Postgraduate Department of Data Science at St. Xavier's College (Autonomous), Kolkata and compare them with suitable individuals from the secondary dataset obtained from Kaggle.

To ensure a meaningful comparison, we subset the Kaggle dataset based on age and continent, selecting data points that align with the demographics of our surveyed students. Our analysis revealed that age plays a crucial role in determining ad click behaviour, as younger individuals tend to exhibit different online engagement patterns compared to older users. Additionally, geographical differences across continents provide valuable insights into variations in online advertising responsiveness.

Reasons for Comparison:

1) We aim to uncover similarities and discrepancies between real-world student responses and secondary data trends, thereby enhancing our understanding of CTR prediction across different demographic groups.
2) It helps to determine if insights from Kaggle Dataset applies to a specific demography.

First, we will analyse the similarities and dissimilarities in Kaggle data subset and Survey Data using Graphical Tools. Then using suitable a test, we will understand how the Click Through Rate changes based on the dataset.

Graphical Analysis:

1) Daily Time Spent on Internet

- **Subset of Student Age group belonging to Asian subcontinent from Kaggle Dataset**

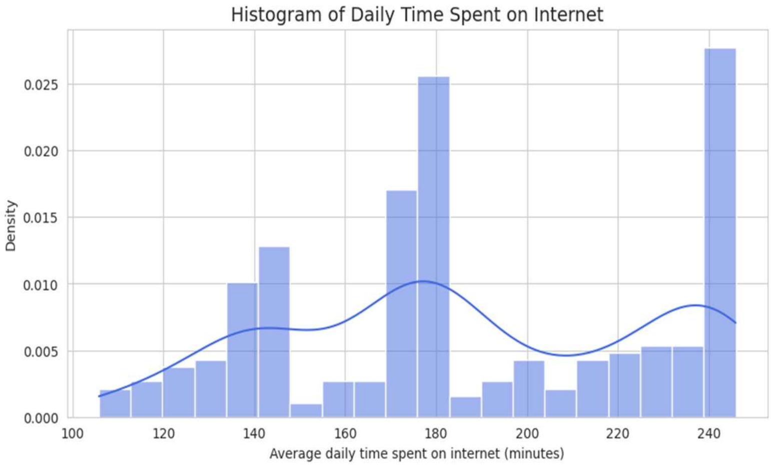


Fig17: Histogram of Daily Time Spent on Internet (in minutes)

- **Primary dataset**

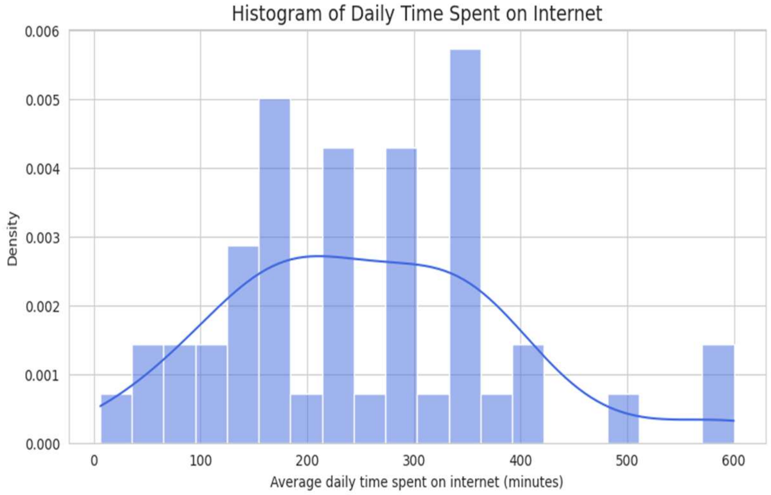


Fig18: Histogram of Average Daily Time spent on Internet (in minutes)

COMPARISON:

1) **Range**: First histogram (100–250 min), Second Histogram (0–600 min).
2) **Distribution**: Both multimodal; second is more right-skewed.
3) **Variability**: First is more concentrated; second has higher dispersion.
4) **CTR Insight**: First group has consistent usage; second has diverse patterns.

Thus, we observe from the histograms that Survey data includes more diverse group of internet user (0-600 minutes) – both heavy and light whereas the Kaggle dataset includes internet users belonging to a certain range (100-250 minutes).

2) **Daily time spent on site (in minutes)**

- **Subset of Student Age group belonging to Asian subcontinent from Kaggle Dataset**

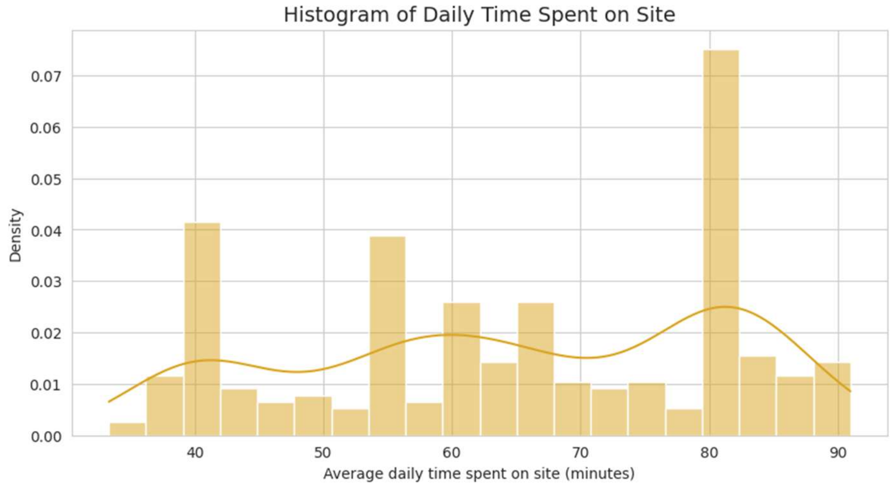


Fig19: Histogram of Daily time spent on site (in minutes)

- **Primary Dataset**

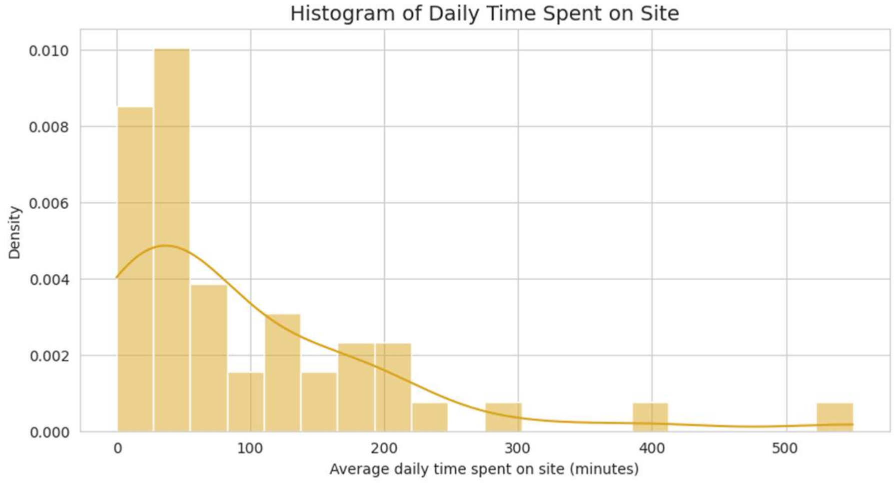


Fig20: Histogram of Daily time spent on internet (in minutes)

COMPARISON:

5) **Range**: First histogram (0-90 min), Second Histogram (0–500 min).
6) **Distribution**: First Histogram is multimodal; Second one is right skewed.
7) **Variability**: First is more concentrated; second has higher dispersion.
8) **CTR Insight**: First group has consistent usage; second has diverse patterns.

Thus, we observe from the histograms that Survey data includes more diverse group of internet user having wide range of time spent on site whereas the Kaggle dataset includes internet users belonging to a certain range (0-90 minute).

3) Area Income (in Rupees)

- **Subset of Student Age group belonging to Asian subcontinent from Kaggle Dataset**

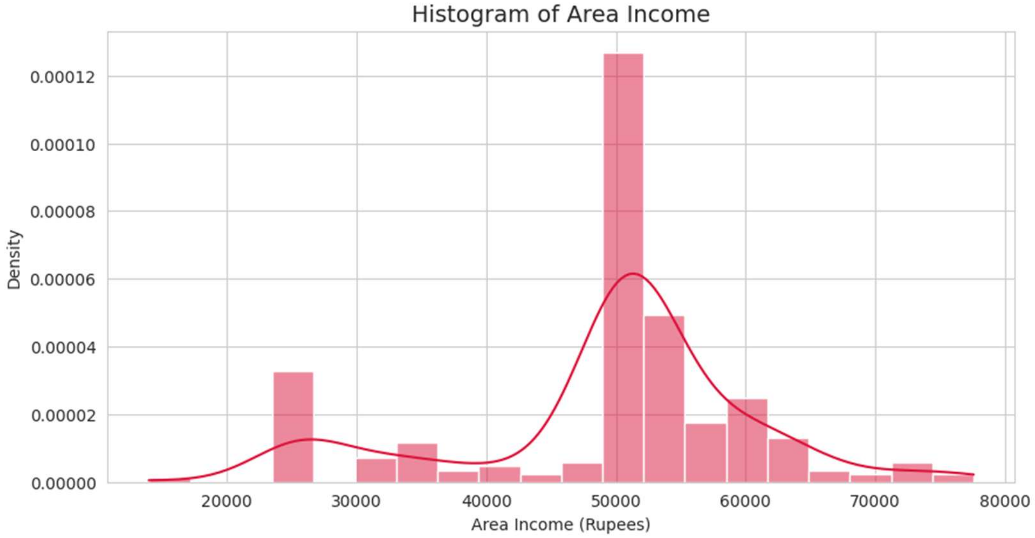


Fig21: Histogram of Area Income (in Rupees)

- **Primary Dataset**

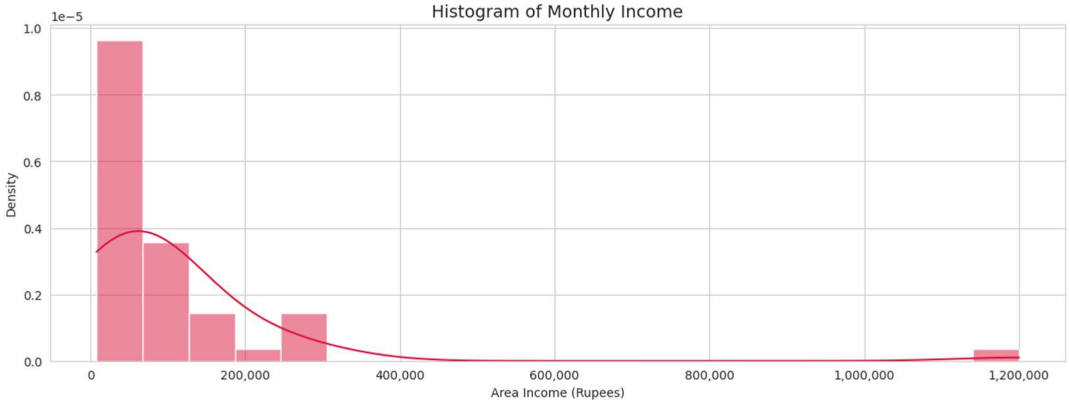


Fig22: Histogram of Area Income (in rupees)

COMPARISON:

1) **Range**: First histogram (Rs.10,000-80,000), Second Histogram (Rs.12,000-12,00,000).
2) **Distribution**: First histogram has a peak around Rs.50,000; Second one is right skewed.
3) **Variability**: First is more concentrated; second has higher dispersion.
   The survey data shows a wider range of Area Income; although the bulk of the distribution is within Rs. 1,00,000, yet some area income has significantly higher income as compared to the Data subset from Kaggle.

4) Gender vs Clicked on Ad

- **Subset of Student Age group belonging to Asian subcontinent from Kaggle Dataset**

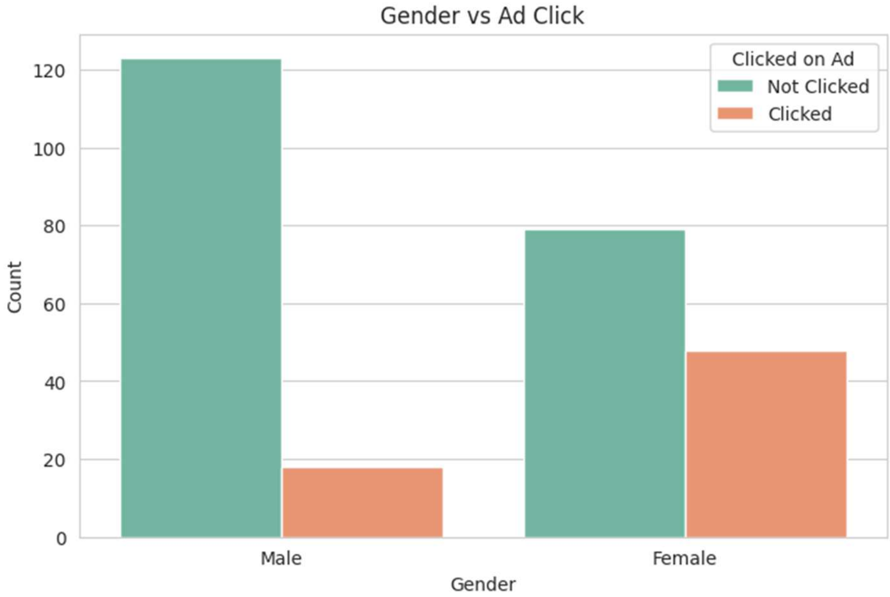


Fig23: Bar-Plot of Gender Vs Clicked on Ad

- **Primary Dataset**

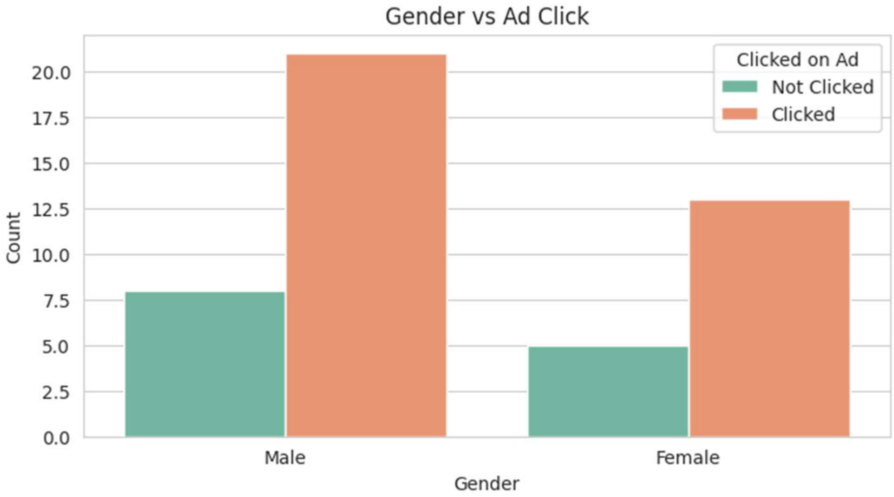


Fig24: Bar-Plot of Gender Vs Clicked on Ad

COMPARISON:

1) In both Kaggle dataset and Survey dataset, it seems that frequency of males are higher than that of females.
2) The Click Through Rate (CTR) is significantly higher for both gender in Survey Dataset.
3) The CTR pattern in the Survey Dataset is reversed compared to the Kaggle data.

5) Time Category Vs Clicked on Ad

- **Subset of Student Age group belonging to Asian subcontinent from Kaggle Dataset**

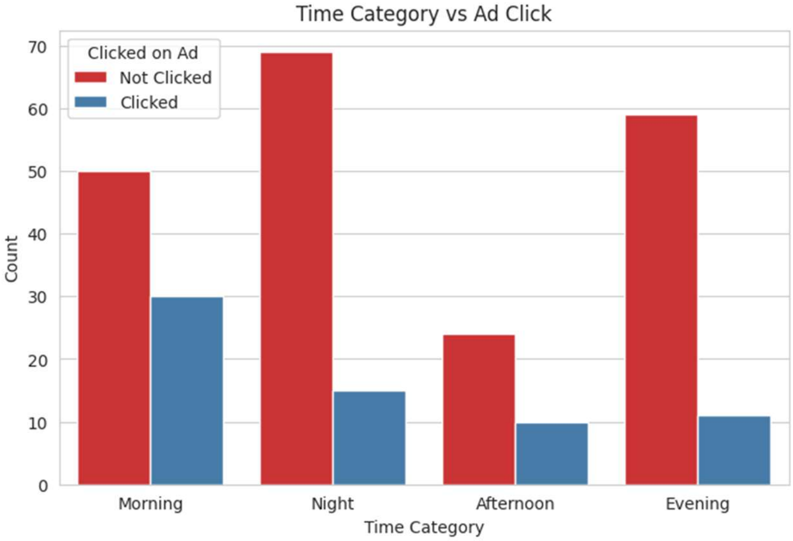


Fig25: Bar Plot of Time category vs Clicked on Ad

- **Primary data**

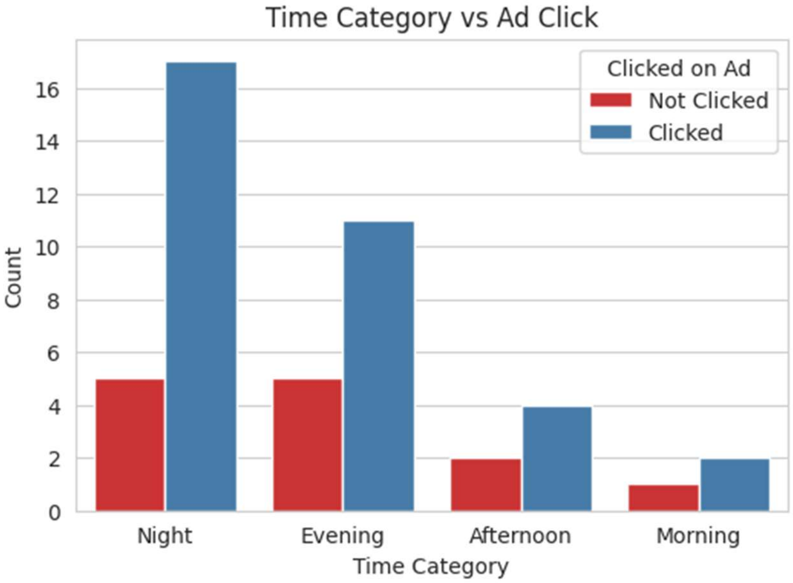


Fig26: Bar Plot of Time Category vs Clicked on Ad

COMPARISON:

1. Both datasets show high ad engagement in the Night and Evening, but the CTR is significantly higher in the Survey Dataset, indicating that survey participants are more active ad clickers.
2. Morning internet usage is notably higher in the Kaggle dataset than in the survey dataset.

Thus, from graphical analysis, we observe the Click Through Rate is comparatively higher in Survey Data. We need to perform statistical test to validate our observation.

Two-Proportion Z-Test (for Large Samples)

Hypotheses:

- **H_0 (Null Hypothesis):** The proportion of students who clicked an ad in Survey is equal to the proportion of individuals in the Kaggle dataset who clicked an ad.
- **H_1 (Alternative Hypothesis):** The proportions are different.

Test Statistics:

$$Z= \frac{\widehat{p_1}-\widehat{p_2}}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}}$$

where:

- $\widehat{p_1}$ and $\widehat{p_2}$ are the sample proportions of ad clicks in Survey dataset and Kaggle data subset respectively.
- n_1 and n_2 are the sample sizes corresponding to survey data and Kaggle dataset respectively
- $\hat{p}$ is the pooled proportion across both groups.

Under the **Central Limit Theorem (CLT)**, if the sample sizes are **large enough**, the difference between two independent sample proportions approximates a **normal distribution**.

Under H_0, Z~AN(0,1)

Critical Region: We will Reject H_0 at ∝ Level of significance iff

$$|Z_{obs}| > \tau_{\frac{\alpha}{2}}$$

Calculation:

We have $n_1 = 47$, $n_2 = 268$, $\widehat{p_1} = 0.7234$, $\widehat{p_2} = 0.2463$, $\hat{p} = 0.3181$

$Z_{obs} = 0.4771/0.07365 = 6.478$

Decision: For ∝=0.05, $Z_{obs} = 6.478 > 1.96 = \tau_{\frac{\alpha}{2}}$

Conclusion: In the light of given data, it seems that the proportion of students who clicked on ads in the survey is significantly higher than the proportion observed in the corresponding subset of the Kaggle dataset. This suggests that the Click-Through Rate (CTR) is notably higher among individuals aged 18-25 years in the survey conducted among students of a specific college.
This Conclusion also supports our observations using different graphs used to compare.

However, this observation may not accurately reflect the broader reality, as the survey data is limited to students from a single institution.

Possible Reasons for Difference

The difference in Click-Through Rate (CTR) between the survey data and the Kaggle dataset could be attributed to several factors:

1. **Demographic Restriction**: The survey data is limited to students from a particular college, whereas the Kaggle dataset represents a more diverse population, including individuals from various backgrounds, professions, and age groups.
2. **Small Sample Size of survey data**: It includes data on Ad click behaviour of only 47 individuals, restricted to more or less similar lifestyle.
3. **Income Disparity**: College students often belong to a relatively homogenous income bracket, whereas the Kaggle dataset may include individuals from varying economic backgrounds, influencing ad engagement.
4. **Ad Relevance**: The advertisements shown in the survey might have been more aligned with student interests compared to those shown to the broader audience in the Kaggle dataset.
5. **Sampling Bias**: The survey respondents may have been more active online, or generally more inclined to interact with ads compared to the general population.
6. **Survey Environment**: The controlled setting of the survey (e.g., academic environment, peer influence) might have subtly encouraged higher engagement with ads compared to real-world browsing behaviour.

Thus, if we extend the analysis to a larger student population across different income groups, the results may align more closely with those from the Kaggle dataset.

Benefits of Adding a Survey Data Comparison

The survey data shows a very different CTR trend from the Kaggle dataset. – This variation is likely because the survey focuses on a **specific segment of the population** with **similar internet usage behaviours**.

How This Assists Ad Agencies:
1) **Fine-Grained Audience Segmentation-** Ad agencies can tailor advertisements identifying high engagement groups.

2) **Cost-Saving Advertising** – Rather than a mass effort, targeting high CTR groups minimizes wasted ad budget and maximizes Return on Investment.

3) **Maximized Ad Timing** – Where surveying indicates higher engagement during particular time periods (e.g., night/evening), agencies can time ads to have the greatest effect.

4) **Tailored Content Development** – Knowing audience behaviour makes it possible to tailor ad creatives for higher engagement rates.
5) **Data-Driven Budgeting** – Through CTR trend analysis, agencies can budget effectively, prioritizing platforms and demographics with the greatest conversion potential.

For Example - Imagine you are a large brand offering a variety of products and services, with insights from a global dataset (like Kaggle) showing key factors influencing Click-Through Rate (CTR). These insights help to shape general ad strategy.

However, you decide to sponsor a college fest at a specific university and need to engage students effectively. Since your global data may not fully capture the behaviour of this localized audience, you conduct a survey among the students to refine your ad strategy.

The survey results reveal significant differences—for example, while your global data suggests that discount offers drive high CTR, the survey shows that students respond more to interactive contests and gamified ads. By leveraging this insight, you shift your strategy from generic discount-based ads to interactive engagement-driven campaigns, maximizing participation and brand visibility. This approach ensures that your ad spend is optimized, engagement is higher, and your campaign resonates with the specific audience rather than relying solely on broad, generalized insights

BIBLIOGRAPHY

Logistic Regression. Retrieved from- https://www.geeksforgeeks.org/understanding-logistic-regression/

Advertising Dataset. Retrieved from- https://www.kaggle.com/datasets/swekerr/click-through-rate-prediction

Visualizations. Retrieved from – https://matplotlib.org/

Research for Logistic regression. Retrieved from – https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Google Ads. (2023). *Understanding Click-Through Rates and Their Impact on Ad Performance.*

Retrieved from. https://support.google.com/google-ads/answer/2615875?hl=en

Survey Questionnaire. Google form link - https://docs.google.com/forms/d/e/1FAIpQLScKrRZgjFTEDVeqw-1GqAmhkiQfDTNN53icAHEVVMdv1IN01g/viewform?usp=sharing

Fundamental of Statistics, Volume 1: A.M. Goon, M.K. Gupta, B Dasgupta.

Fundamental of Statistics, Volume 2: A.M. Goon, M.K. Gupta, B Dasgupta.

Montgomery, D. C., & Runger, G. C. (2018). *Applied Statistics and Probability for Engineers* (7th ed.). Wiley.

Cochran, W.G. (1984) Sampling Techniques (3rd Ed.), Wiley Eastern