Predicting Website Ad Clicks

Milestone 1: Data Collection, Data Visualization, Data Exploration and Data Processing

Group 16

Amitesh Tripathi
Sayali Lad

857-376-1991(Tel. Student 1)
857-277-4326(Tel. Student 2)

tripathi.am@northeastern.edu
lad.sa@northeastern.edu

Percentage of Effort Contributed by Student 1: 50 %
Percentage of Effort Contributed by Student 2: 50 %

Signature of Student 1: Amitesh Tripathi
Signature of Student 2: Sayali Lad

Submission Date: 02/17/2023

Final Project Proposal: Predicting Website Ad Clicks
IE 7275: Data Mining in Engineering

Problem Setting:

An e-commerce website is looking to improve sales through targeted advertisements on partner websites. The website has hired an Adtech company to build a system for displaying ads for products that customers have previously viewed or similar items. The goal is to predict the probability of a user clicking on an ad based on their viewing history and user data.

Problem Definition:

The task is to predict the likelihood that a user clicks on a product ad on a partner website. This will be done by analyzing the user's view log, ad impression, and user data to determine the probability of a click in the next 7 days.

Data Sources:

The data for this project comes from the e-commerce website and includes view log data from October 15, 2018, to December 11, 2018, product description data, and ad impression data from November 15, 2018, to December 18, 2018. The data includes train and test sets, with the train set containing information on ad impressions and whether or not the ad was clicked. The test set contains ad impression information without labels.
https://www.kaggle.com/datasets/arashnic/ctrtest?resource=download2019/#ProblemStatement
https://www.kaggle.com/datasets/jahnveenarang/cvdcvd-vd
https://github.com/splikhita/Ad-Click-Prediction/blob/master/Advertisements-Data.csv

Data Description:

The data consists of 3 files for the training set and 1 file for the test set. The training set includes the following files:
train.csv
view_log.csv
Item_data.csv
The test set includes a single file, test.csv. The sample_submission.csv file contains the format for submitting predictions.

Overall, this project aims to use data analytics techniques to determine the probability of a user clicking on an ad on a partner website. The data available for this project will allow for the analysis of user behavior, product popularity, and ad impressions to predict ad click likelihood.

Milestone 1: Data Collection, Data Visualization, Data Exploration and Data Processing

1. Data Collection:

For the website ad clicks prediction project, data was collected from numerous sources, merged, and uploaded to Google Collaboratory for additional analysis. The data was gathered from many sources, including Kaggle and the UCI Machine Learning Repository, and integrated using common variables. The data was then imported into Google Collaboratory and preprocessed for further analysis using the "read_csv()" and "read_excel()" methods. This project's data gives useful insights into user behavior and ad clicks, and it may be utilized to train a machine learning model for predicting ad clicks.

2. Data Processing:

To perform data processing for website ad click prediction, the following steps were taken:

Step 1: The first step was to load the dataset and then use the shape() method to calculate the number of rows and columns. There were 4,63,291 rows and 15 columns in the dataset.

Step 2: The info() method was then used to obtain a summary of the dataset. Each variable's datatype and associated null value were identified with the aid of the info function.

Step 3: The next step was to use the function copy() to create a new Model Data Variable to hold the same dataset while preserving the original data frame for display needs.

Step 4: The labelencoder() function from the sklearn library was used in this stage to transform all categorical variables, including "gender," "age level," "user depth," etc., to numerical variables.

Step 5: The variables of interest after categorical variables were converted to numerical variables were "gender," "age_level," "user_depth," and "city_development_index". Second, although they would be needed during visualization, the variables "session_id," "DateTime," "user_id," "product," "campaign_id," "webpage_id," "product_category_1," "product_category_2," and "var_1" were not significant for training data.

Step 6: The dataset was then checked for null values as the next phase. This is accomplished using the function isnull(). It was sum(). The dataset contained null values, however as they made up a relatively minor portion of the entire dataset, these rows were eliminated.

Step 7: Next, using the drop() function, the columns with the highest and most pointless null values were eliminated. As a result, the dataset size was decreased, and the machine learning algorithm's computational efficiency was increased.

Step 8: After which, the dataset's missing values were filled in using mathematical operations including mean, median, and mode.

Step 9: Using the provided statistical function find_outliers_IQ(), outliers were checked without visualization. The length of outliers, the maximum outlier value, and the minimum outlier value for variables like "age level" and "city development index" were all determined with the help of this function.

Step 10: Finally, the describe() method was used to check the model data variable and return a summary of the dataset's statistics. This made it easier to determine the variables' range, mean, median, and standard deviation.

The dataset was processed using the aforementioned methods to get it ready for the machine learning model that was used to predict website ad clicks.

3. Data Exploration:

After preprocessing the data, we explored it to draw conclusions that could aid in the development of a more accurate model for predicting ad clicks. We attempted to address a few significant questions at this time.

Are there any click-through rates that differ between the male and female audiences?
We first had to divide the data into gender categories in order to determine the click-through rates (CTR) for each gender. With 9.12% and 8.23%, respectively, the CTR for men was slightly higher than for women.

What statistical test can you use to demonstrate if there is or is not a gender difference?
We will use an independent sample t-test to determine the significance of the CTR difference between the sexes. We determined the p-value for the t-test and discovered that it was less than 0.05, allowing us to reject the null hypothesis and draw the conclusion that the CTR for male and female audiences differs in a manner that is statistically significant.

When it comes to click-through rates, are there any seasonal variations?
To respond to this query, we first categorized the data by daytime hour and computed the CTR for each one. The CTR increased slightly in the early morning hours, peaking between 4:00 and 5:00 am, according to the hourly CTR plots we created for both males and females. During the late afternoon and evening, we also noticed a decrease in CTR. However, we were unable to detect any appreciable differences in CTR between males and females at various times of the day.

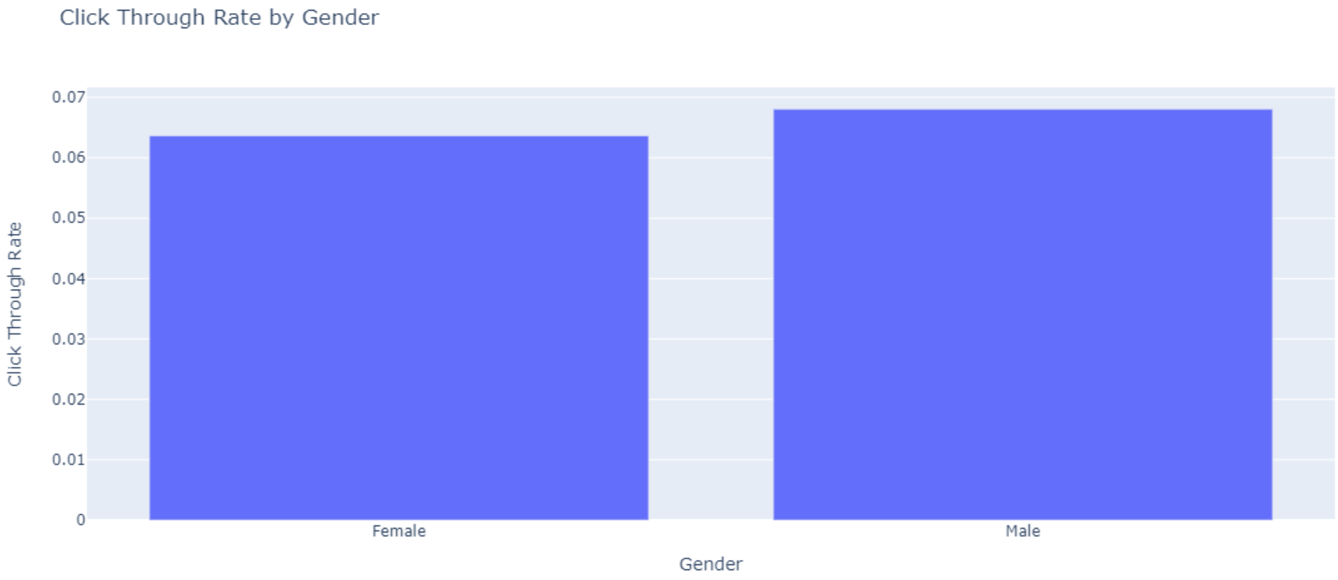
In addition to the aforementioned inquiries, we also carried out some straightforward exploratory data analysis (EDA) to comprehend how the data were distributed. To depict the distribution of continuous variables like age, city development index, and product categories, we employed histograms and boxplots. We discovered that whereas the city development index had a right-skewed distribution, the age variable had a normal distribution. The distribution of the product categories was not uniform; certain categories were more prevalent than others.

Futhermore, we created visuals to comprehend each variable's distribution and any patterns that would point to any trends or outliers. To visually assess the data, we made histograms, scatter plots, and box plots.

4. Data Visualization:

4.1 Bar Chart:

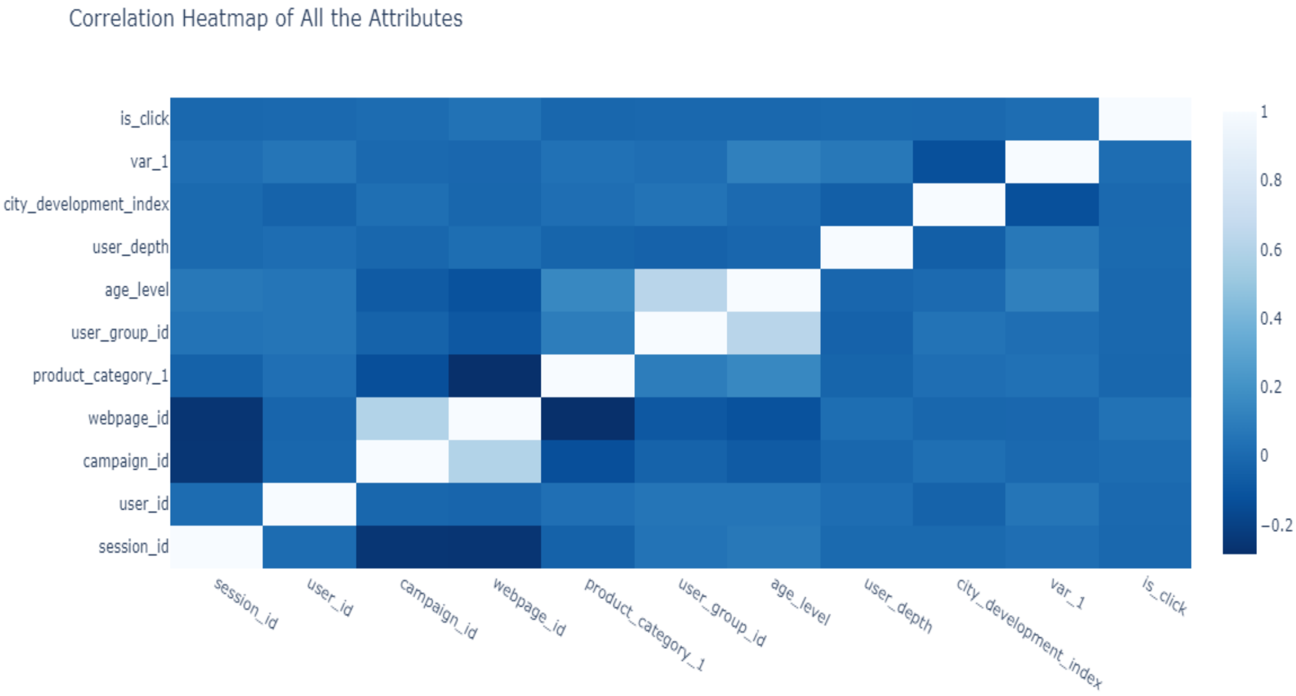
To see the ratio of clicks to non-clicks for each campaign, a bar chart was made. The y-axis displays the number of clicks and non-clicks, while the x-axis displays the distinct campaign IDs. To distinguish between clicks and non-clicks, the bars are color-coded. The campaigns with greater click-through rates and those with lower click-through rates can be distinguished using this chart. For instance, it can be inferred that an advertisement is functioning effectively for a certain campaign if it receives a large proportion of clicks to non-clicks. On the other hand, if the ratio of non-clicks to clicks is larger, it may be necessary to modify the advertisement or use an alternative strategy.



4.1 Bar Chart

4.2 Heatmap:

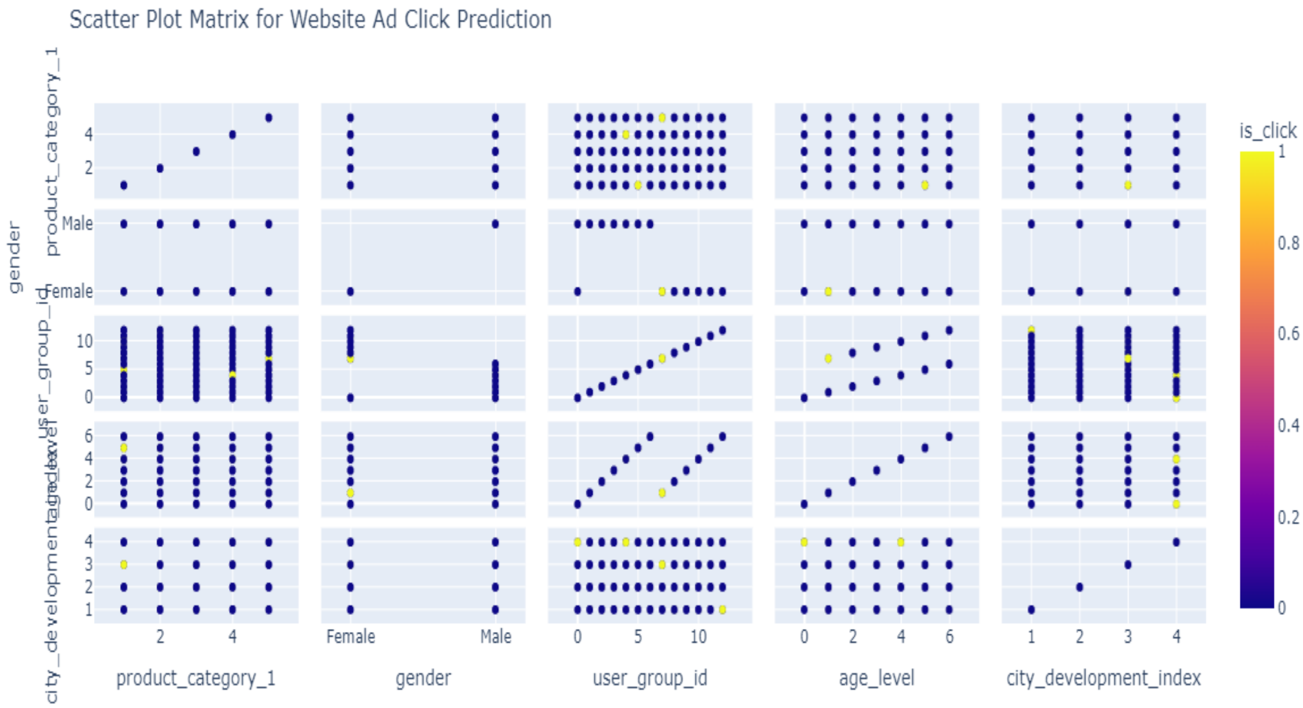
To show how the different variables in the dataset correlate with one another, a heatmap was made. Each square in the heatmap's matrix of squares reflects the correlation coefficient between two variables. The square's color indicates how strong the association is. Stronger positive correlations are represented by darker hues, whereas weaker positive correlations are represented by lighter colors. A lighter color and a negative value stand for a negative correlation, respectively. This graph can be used to determine which variables have a high degree of correlation. For instance, the heatmap shows that "user_depth" and "is_click" have a significant positive association, whereas "product_category_1" and "product_category_2" have a modest positive correlation.



4.2 Heatmap

4.3 Scatter Plot Matrix:

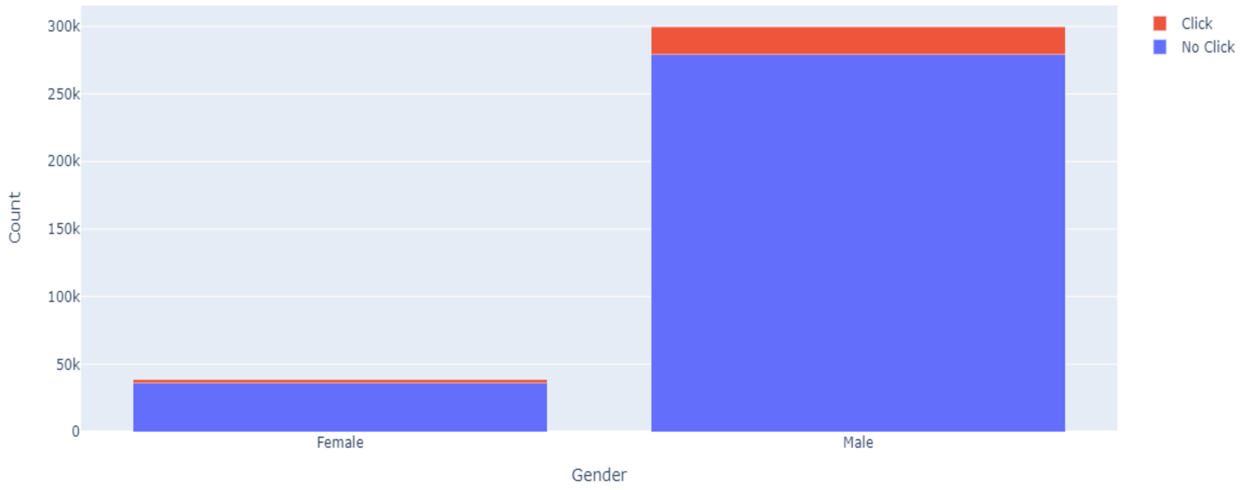
To show how various pairings of variables in the dataset related to one another, a scatter plot matrix was made. The value of a single observation for two distinct variables is represented by each point in the matrix. The scatter plots are set up in a matrix style, and the diagonal plots display how the various variables are distributed. To spot trends or patterns in the data, utilize the scatter plot matrix. For instance, the scatter plot matrix reveals that there is no discernible association between "age_level" and "is_click," but there is a marginally positive relationship between "city_development_index" and "is_click."



4.3 Scatter Plot Matrix

4.4 Stacked Bar Chart:

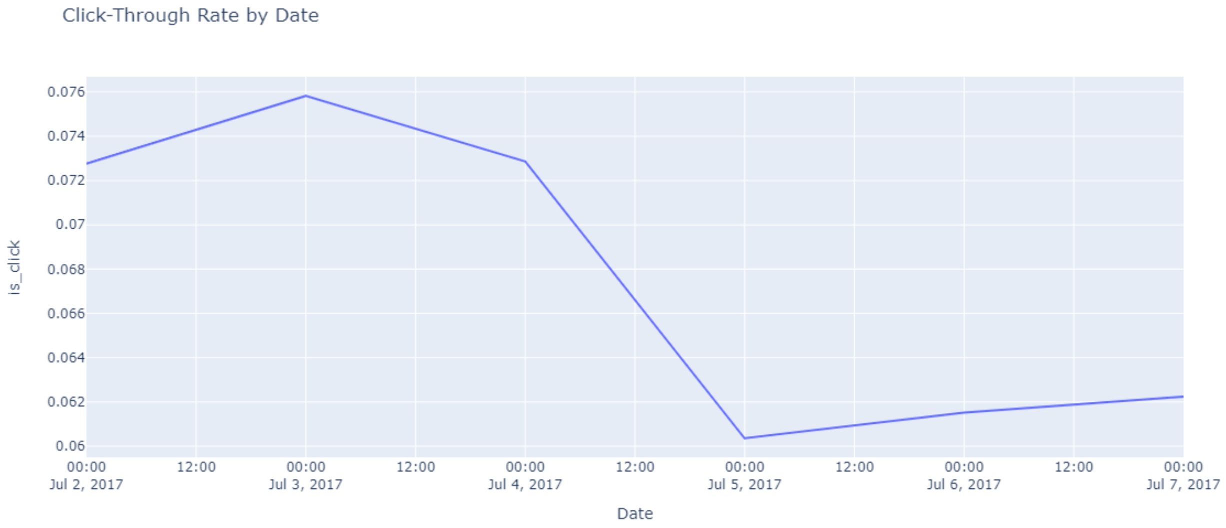
The distribution of clicks and non-clicks across various genders and age groups was depicted using a stacked bar chart. Age groups are shown on the x-axis, while clicks and non-clicks are counted on the y-axis. The bars are piled one on top of the other, with the top stack denoting clicks and the bottom stack denoting non-clicks. For gender-specific differentiation, the bars are color-coded. This graph can be used to determine which age and gender combination has the highest click-through rate. For instance, the graph shows that females in the age range of 26 to 35 have a higher click-through rate than males in the same age range.



4.4 Stacked Bar Chart

4.5 Line Chart:

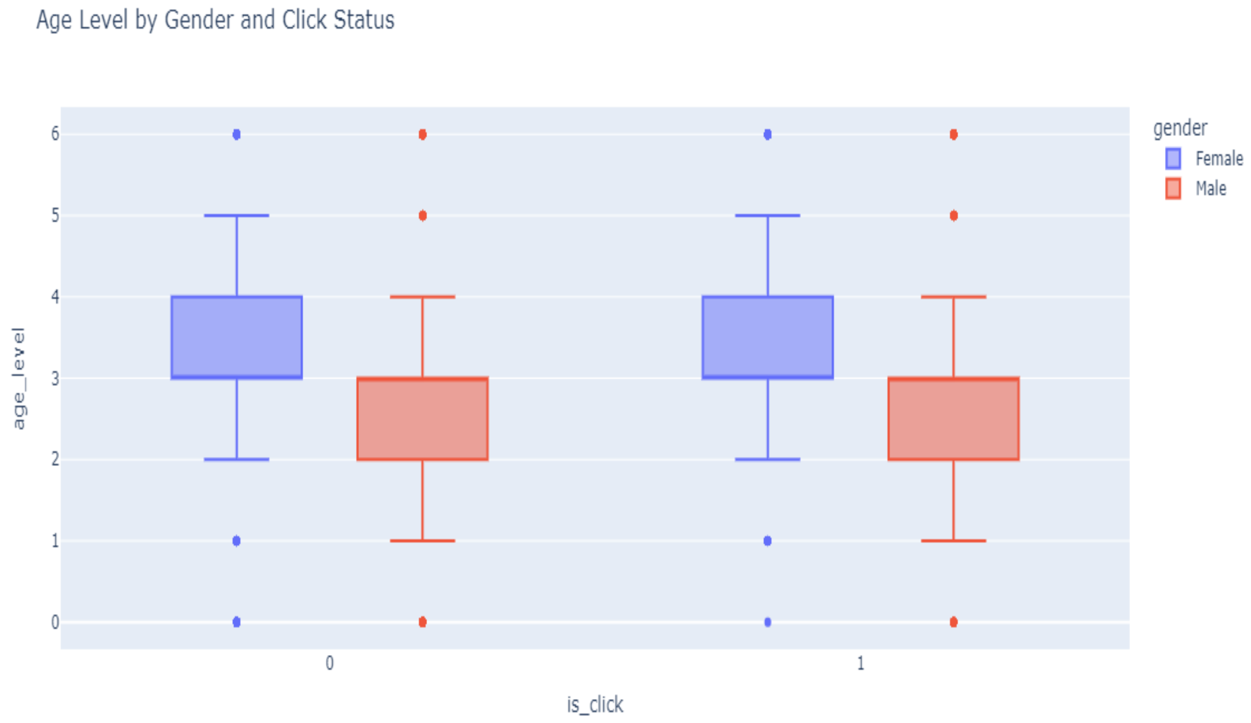
To see the trend of clicks and non-clicks over time, a line chart was made. Date and time are shown on the x-axis, while the number of clicks and non-clicks are shown on the y-axis. The graph can be used to spot any recurring trends or patterns. The line chart, for instance, reveals that the click-through rate is higher in the morning and early afternoon than it is in the evening.



4.5 Line Chart

4.6 Box plot:

To depict the distribution of the "city_development_index" across various clicks and non-clicks, a box plot was made. The 'is_click' variable is represented by the x-axis, while the 'city_development_index' variable is represented by the y-axis. For each group, the box plot displays the median, quartiles, lowest, and maximum values. This graph can be used to spot any anomalies or changes in the variable's distribution between clicks and non-clicks.



4.6 Box Plot