1. Introduction

In the new era of the internet, marketing drastically changed, and companies extensively use the internet for advertising their product. However, it is crucial to target the right audience to manage the cost of the advertisements. Targeting the right audience via online marketing possesses many challenges. It is vital to have a detailed analysis to promote the product to the right customers efficiently and effectively to manage the cost.

In this project, the advertising marketing data is used to provide a detailed exploratory statistical and Geo-map analysis and provide a model predication based on the feature in the data such as, 'Age', 'Area Income', 'Daily Internet Usage', 'Ad Topic Line', 'City', 'Country', Gender, and if the user 'Clicked on Ad' or not. The variable of interest is 'Clicked on Ad' with two possible outcomes: 0 and 1, where 0 refers to the case where a user did not click the advertisement, while one refers to the cases where a user clicks the ad.

2. Dataset Description

The data set is from the Kaggle advertising section. This data is used for exploratory statistical analysis and used to train the machine learning model. The main objective is first to perform some statistical exploratory data analysis to see how the features such as 'Daily Time Spent on Site' affect the user's decision to click on the ad. The idea is to see which customer is more likely to Click on an Ad based on their interest feature. Or if there are any gender differences in the data distribution (whether Click in Ad or not). And finally, in the prediction part, if there is a model that can accurately predict the value 'Clicked on Ad' variable.

3. System Functionality

The dashboard is divided in the two-part as follows:

Part 1: Exploratory Analysis
Part 2: Model Prediction

Part 1) In this part, the dashboard provides interactive figures so the user can do an exploratory statistical analysis of the data as follows:

1) Entire plots in the dashboard will be interactively updated based on Age:

Since the Age of the audience is fundamental, the user is given the capability to change the entire plot with a range of Age of the audience. The user can select the Age range that provides the best correlation amount other features.

2) The entire plots in the dashboard are interactively updated based on a) the audience who clicked on the Ad, b) the audience who did not click on the Ad, c) both a and b groups.

3) Geo map plot provided (figure 1) a country location of each data. This graph helps users identify which country has to most daily internet usage on the Site. The graph also calculated the average Age of the users in each country. For example, in this figure, for the user who clicked in Ad, the avg age in Turkey is about 41 years. It also shows Norway and Germany have the average Age of 52 are the oldest, and Japan, with avg Age of 24 is the youngest country.

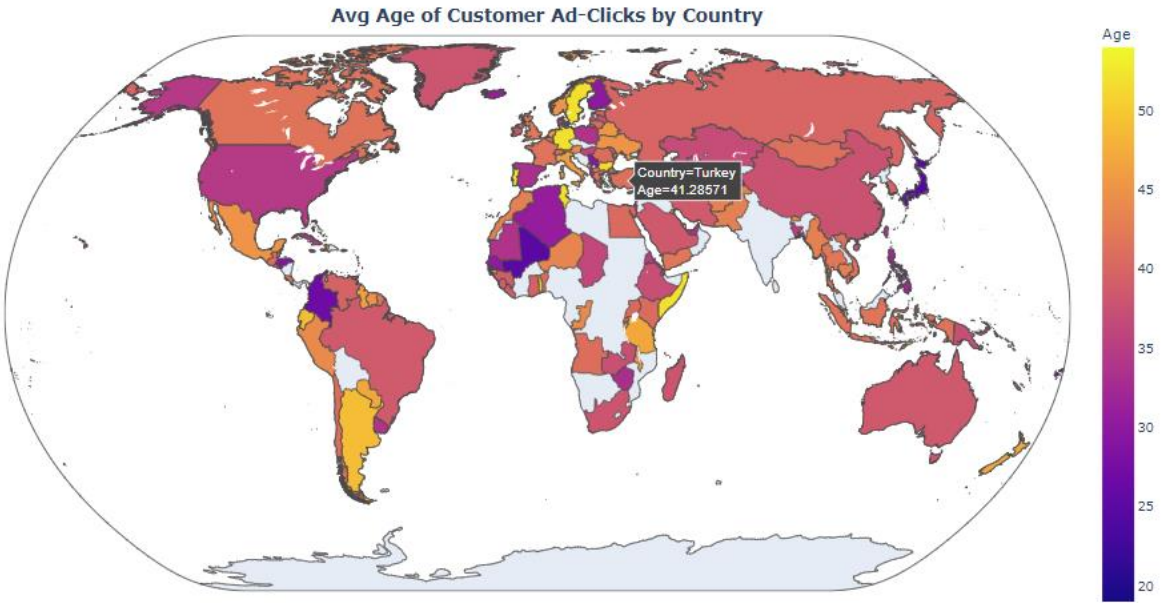

Figure 1

4) Treemap of Daily customer time Spent on Site in the figure shows the country, and the figure 2a, shows the city of each user. With this plot, a user can quickly see which county (and its cities) has the most daily time spent on Site. As the dashboard has interacted with age range and Clided Ad data, a user can see which city has the most daily time spent on

the Site per selected age range. For example, it can be seen that Turkey has the highest daily spent on Sites, along with the city of Willemstad. By hovering the mouse, the user can also see the customer's Age who clicked on the Ad is around 30 in this city.

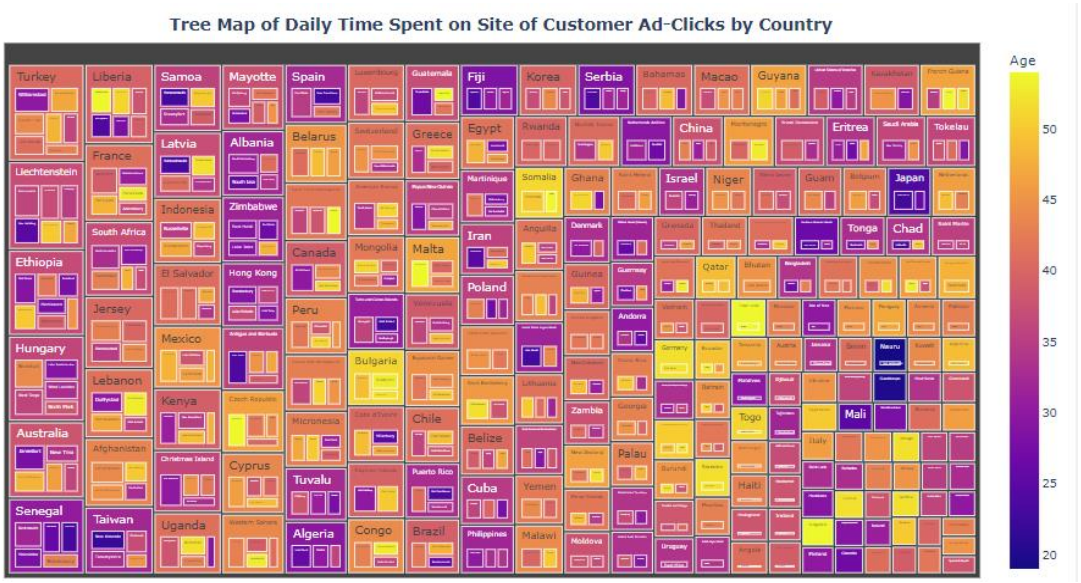

Figure 2

Figure 2a

5) Scatter Matrix and Correlation plots in Figures 3, and 4 show the scattered and the correlation between each pair in the feature data. Figure 3 mainly indicates that the data can be clustered into two groups: audiences who clicked on Ads and those who did not click on Ad.

By looking at both data sets in clicked Ad and non-clicked, in correlation plots in figure 4, we can see that generally speaking, the user who more spent on the Site are younger Age (negatively correlated), and the user who has more daily internet usage also has more spent on the Site.

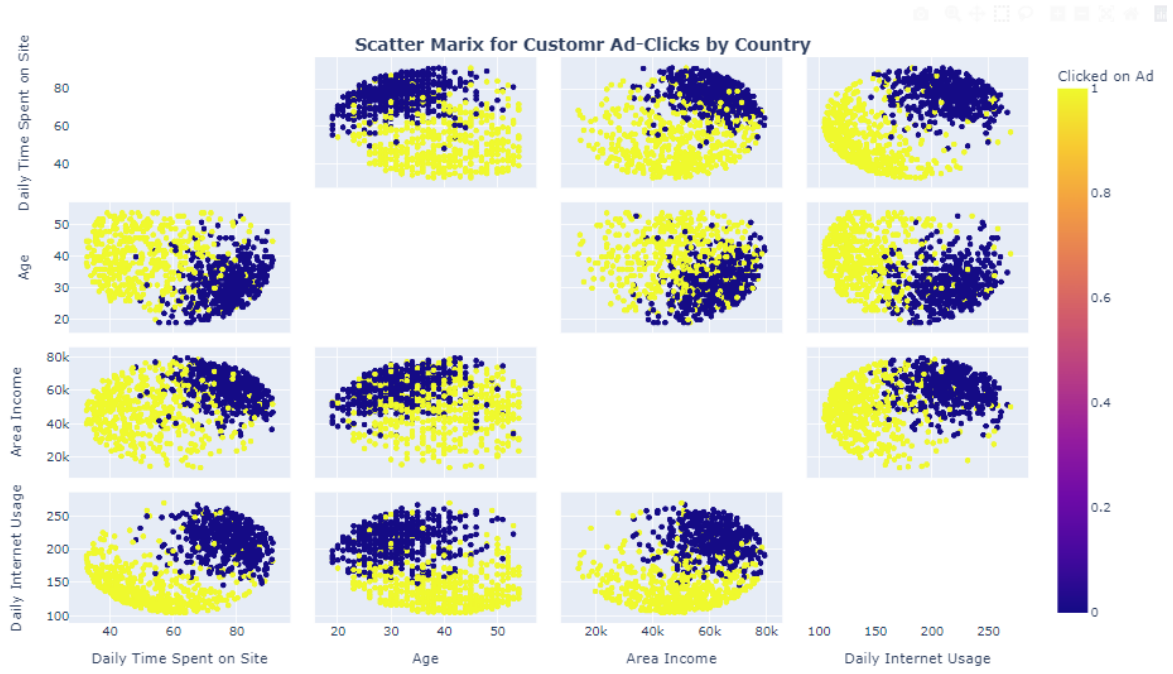

Figure 3

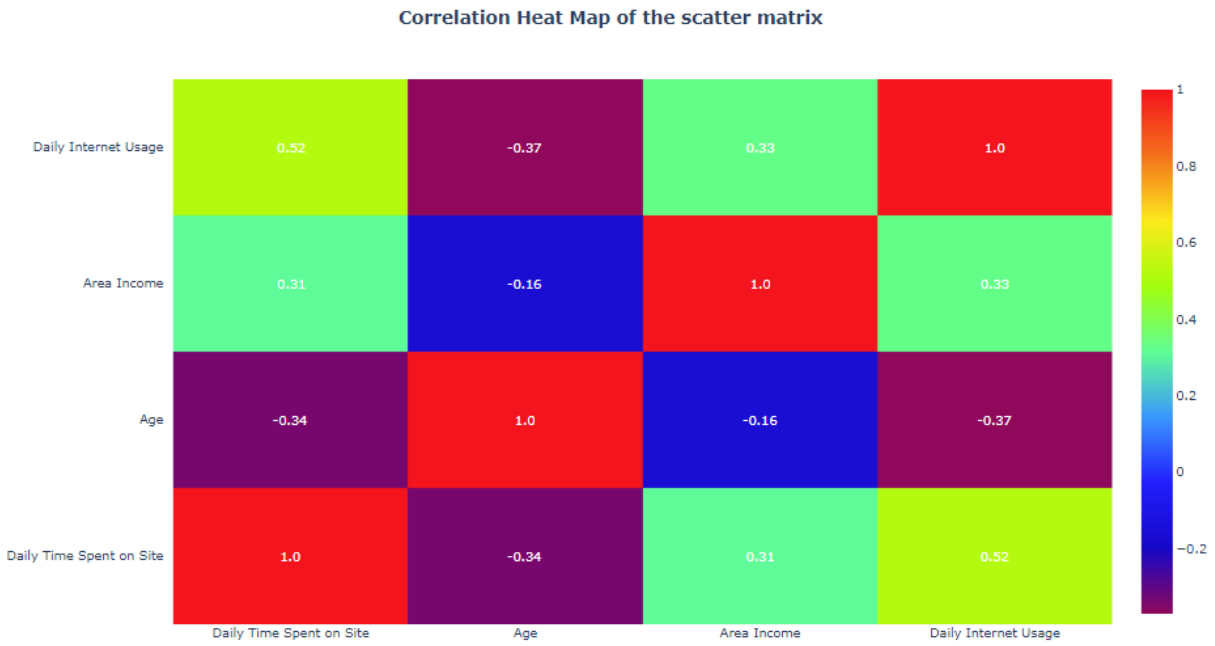

Figure 4

6) Scatter plot of Daily Internet Usage vs. Daily Spent time on Site vs. Age, and its corresponding histogram and KDE demonstrated the distribution of Daily Internet Usage and Daily Time Spent on Site provided in Figure 5. All figures suggest that the data can be clustered in two groups with two different daily spent on-site time distributions to reduce the data to two groups.

Figure 5

7) The 2D density contour shows Daily Time Spent on Site and indicates that the data can be clustered in two groups in figure 6. This plot, along with GMM in figure 7, provided two group populations based on Daily Time Spent on Site vs. Age.

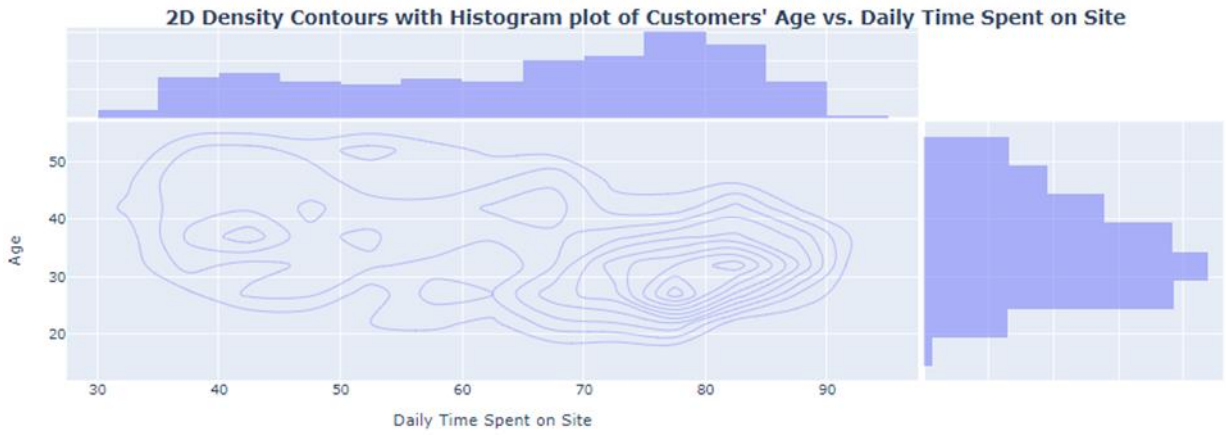

Figure 6

8) Gaussian Mixture Model method is used to see how the data can be clustered into two (or more) groups. Users can select the desired cluster, and the cluster shows with different colors. The Silhouettes method is recommended as an option for selecting the number of

cluster groups. Figure 7 shows the scatter data of Daily Time Spent on Site vs. Age, with two symbols for Clicked on Ad 0, and 1.

Figure 8 and Figure 9 interactively show the corresponding cluster. Figure 8 shows the distribution in violin plots, and Figure 9 shows the KDE of the distribution. The p-values corresponding to each pair group are calculated based on the t-Student distribution to see if there is any significant difference in the two distributions (based on some significant level criteria).

The two clusters group can give good criteria for reducing the data to two groups based on the Daily Time Spent on the Site. The idea would be to focus on the audience using less time on the Site and clicking on the Ad.

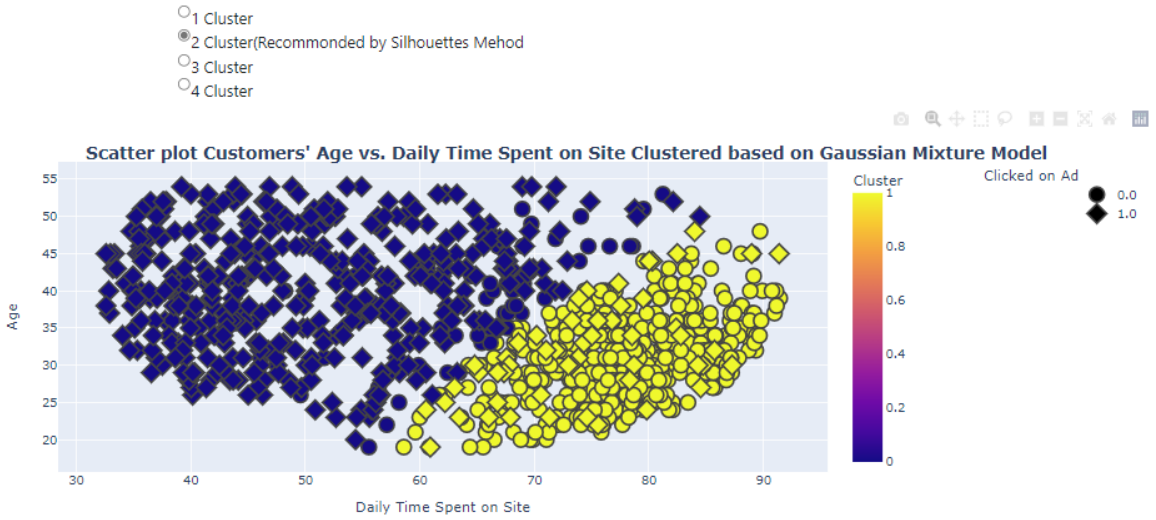

Figure 7

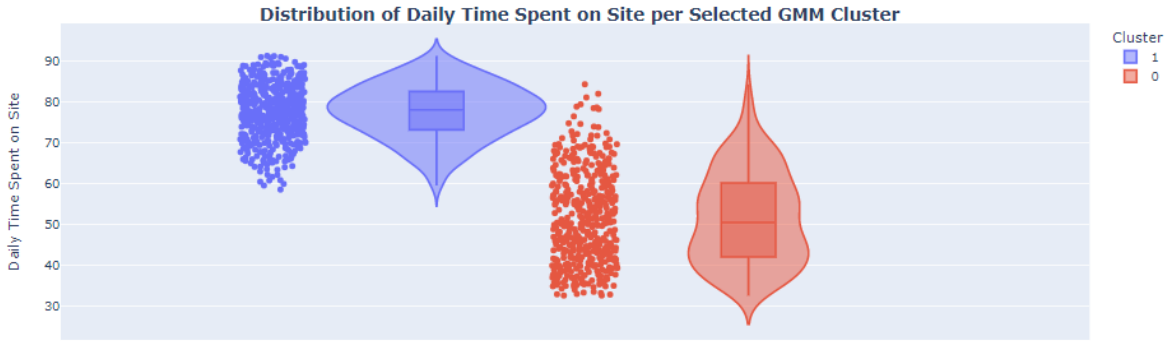

Figure 8

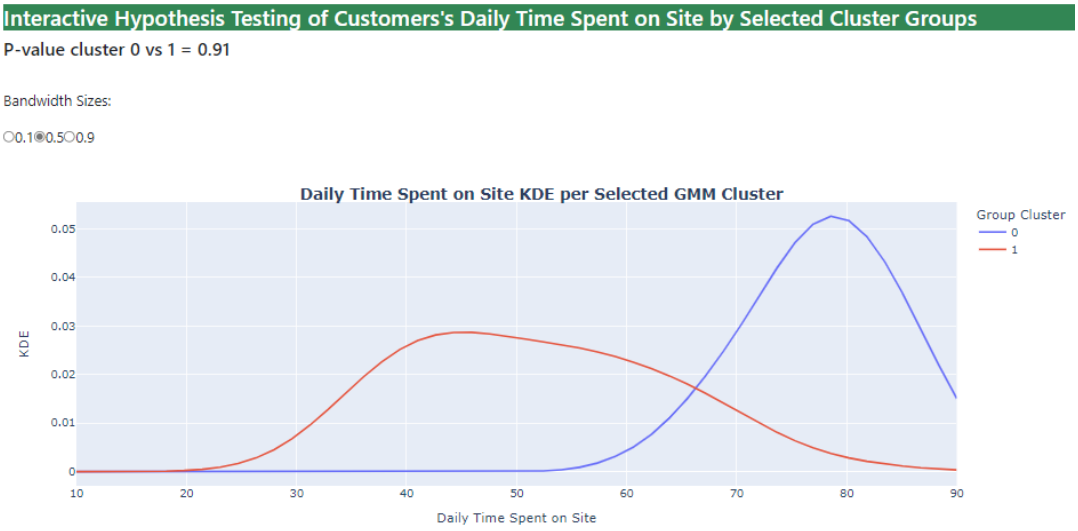

Figure 9

9) In this part of the dashboard, the gender impact in the data is analyzed. The null hypothesis is "The ave age of the male customer is the same as female customers". The alternative hypothesis is that avg Age is not the same.

Figure 10 on the left shows the distribution of males vs. females in the violin plot, which is very similar—in figure 10 on the right shows the corresponding KDE plot, which confirms the similar distribution. In the upper-right part of figure 10, the p-value of the two distributions has been calculated based on the t-Student method. As expected, the p-value is way larger than 5% significant interval level, so it is failed to reject the hypothesis, meaning there is no gender differences in using the Site.

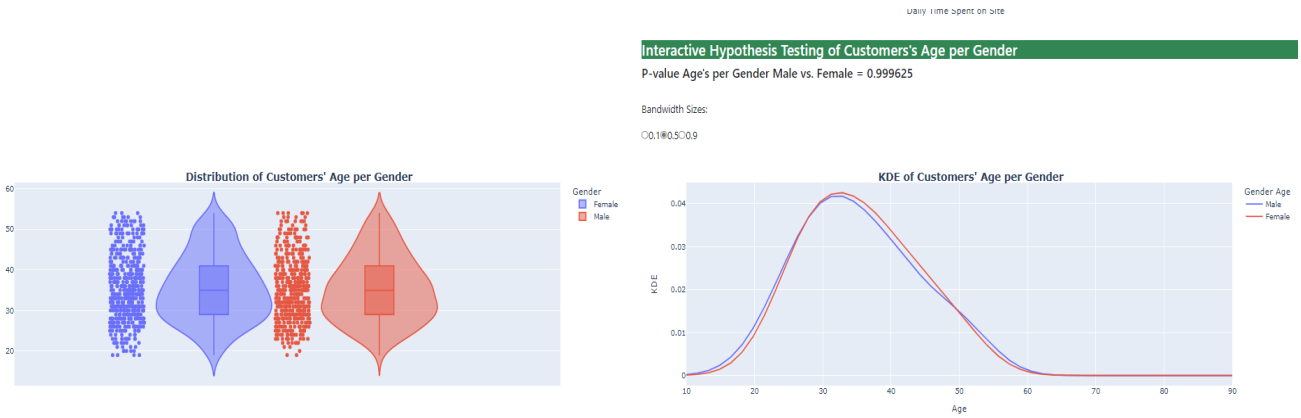

Figure 10

Part 2)

This part is dedicated to the model prediction:

A machine learning model is provided based on Logistic Regression with an accuracy of 90.6%. Figure 11 provides the performance of the model based on the confusion matrix. The total number of accurate predictions is 158+141 = 299, and the total number of incorrect predations is 27+4 = 31, which is a good performance measure for this application.

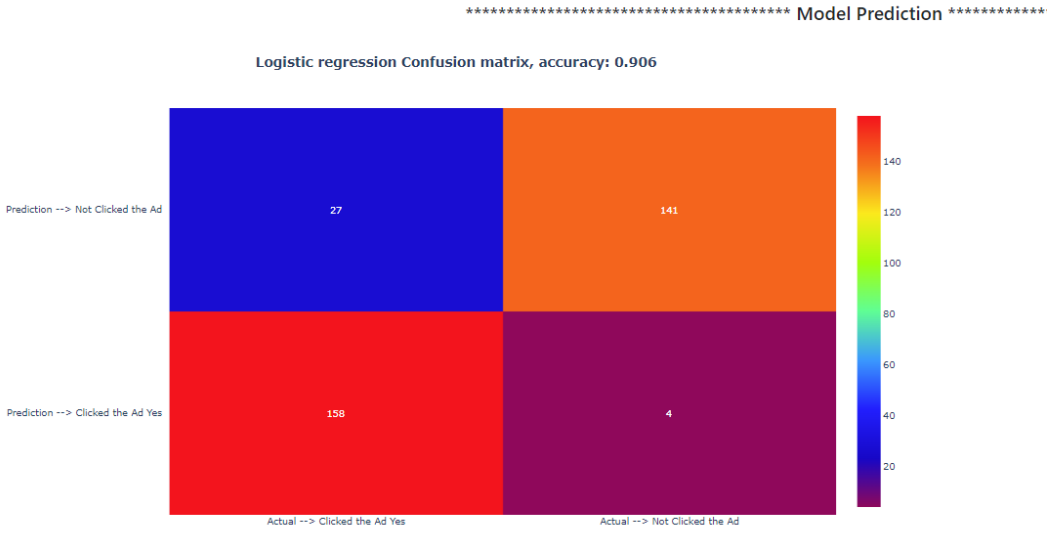

Figure 11

4. Accomplishment Summary

A dashboard has been developed for marketing Ads to target better the right audience for a product based on their interest. The dashboard has two parts; Part 1 provides exploratory statistical data analysis, and part 2 provides model predictions to predict if an audience would click on an Ad or not.

In part 1, various statistical analysis tools are provided. A Scatter matrix plot and corresponding Pearson correlation matrix were provided to show the correlation between the features. Histograms are provided along with the kennel density estimation of the distribution. The Gaussian Mixture model provided a good tool for data decomposition based on the Daily Internet Usage of the users. t-test distribution and p-value provided for each cluster pair to investigate the distribution differences and if a null hypothesis can be rejected or not. For example, the p-value of the two distributions has been calculated based on the t-Student method. As expected, the p-value is larger than the 5% significant interval level, so it failed to reject the hypothesis, meaning there are no gender differences in using the Site.

The dashboard provides a good tool to analyze the male vs. female interest in an Ad based on hypothesis analysis justification.

All plots in the dashboard interact with the range change of Age and the Clicked on Ad data options.

The dashboard also provides a Geo map of the entire world and a tree map that quickly shows the location of the counties and their cities with the highest daily spent time on the Site.\

In Part 2, the dashboard provides a model prediction based on Logistic Regression to predict which customer is more likely to click the Ad. The accuracy and confusion matrix of the model interact with the range Age. The model accuracy is 90.6% for this data which is satisfactory for this application.

Result summary:

The analysis shows the average Age of visitors is around 36 to 54 years (based on +/-2 sigma level), which concludes that the Site targets an adult audience. There is no difference in the male vs. female population in clicking the Ad, and the visitors who have more daily internet usage

also have more spent on the Site. Turkey has the highest daily spent on Sites with avg Age of 41 years, with the city of Willemstad with an avg age of 30. Model prediction based on the Logistic Regression method can predict which visitor is more likely to Click on an Ad with an accuracy of 90.6%.

The entire figure in the dashboard is interactively changed based on the Age range and labeled Clicked on Ad data which provides a good tool for analyzing different scenarios.

The model prediction of the dashboard provides the prediction label as well as the confusion matrix and accuracy of the model. This part is also interactive with Age range.