# Web dataset

Logistic regression

**Objective:**

The goal of the case study is to Predict who is likely going to click on the Advertisement so it can contribute to the more revenue generation to the organization.

**Steps:**

Imported the dataset and identified the Target variable which is clicked:

**1: clicked**

**0: not clicked**

**Summary of dataset:**

* An interesting fact from the summary table is that the smallest area income is 13,996 and the highest is 79,485. This means that site visitors are people belonging to different social classes.
* Since users spend between 32 and 91 minutes on the website in one session. These are big numbers!
* It can also be concluded that we are analysing a popular website.
* Furthermore, the average age of a visitor is 36 years. We see that the youngest user has 19 and the oldest is 61 years old. We can conclude that the site is targeting adult users.

**Removing the Garbage variables:**

* **VistID:** There are too many unique elements within these columns, and it is generally difficult to perform a prediction without the existence of a data pattern. Because of that, it will be omitted from further analysis.
* **Country\_Name:** We have already seen, there are 237 different unique countries in our dataset and no single country is too dominant. Many unique elements will not allow a machine learning model to establish easily valuable relationships. For that reason, this variable will be excluded too.
* **Year:** it is also not going to help us in prediction because it has only one-year 2020 repeating all the time.

**Data pre-processing:**

1. **Checking the missing values:** We didn’t find any missing values in this dataset
2. **Checking the outliers:** We haven’t found outliers as well
3. **Encoding concept:** We had moved the categorical variables into factor which has not that much unique values.

(“Ad\_Topic","City\_code","Male","Time\_Period","Weekday","Month","Clicked")

1. **Feature scaling:** We have done feature scaling because the Avg\_Income column is really huge number as compared to other columns in dataset

**Univariate and Bivariate analysis**

Univariate analysis:

* for continuous variables we will plot it by histogram graph
* for categorical variables we will plot it by bar graph

Bivariate analysis:

* continuous variables to continuous variables ----> scatter plot
* continuous variables to categorical variables ----> Box Plot
* categorical variables to categorical variables ---> group bar plot
* In Month of July and March clicked are less as compared to other months.
* Almost all these days clicked are same on Wednesday slight less.
* In the evening clicked are bit more as compared to other time.
* In the city 1 has max no of clicked as compared to another city.
* Female has more clicked on Ad as compared to male Finally, if we are wondering whether the site is visited more by men or women, we can see that the situation is almost equal (52% in favour of women).

**Statistical Relationship between target variable (Categorical) and predictors**

Continuous Vs Categorical relationship strength: Analysis of Variance (ANOVA)

* H0: Variables are NOT correlated
* Small P-Value <5%--> Variables are correlated (H0 is rejected)
* Large P-Value--> Variables are NOT correlated (H0 is accepted)

All variables are good for prediction

Categorical variables Vs Categorical variables -- Chi-square test

* Column weekdays: p-value = 0.7226 and Column Month: p-value = 0.4229, We will not be considering these two variables in further analysis because the p- value is greater than 0.05

**Build the Logistic model:**

* Installed packages (CaTools) and splitting the Train sample (70%) and test sample (30%)
* glm() is used for wide variety of modelling activities. Logistic regression is one of the models that you can create using glm (). In order to tell glm() that you have to perform logistic regression, you have to say family= 'binomial"
* We used step () to get better AIC value and for better model
* We haven’t found the multicollinearity in these dataset
* Then last step is summary of the final model

Some fact about good model:

* Deviance is a measure of goodness of fit of a generalized linear model.
* residual deviance must be lesser than null deviance to have a good model.
* Unlike R-squared- this AIC value is relative as you run different models you see how the AIC value is changing lower it is, better is the model.

**Predictions of the model**

* Using the obtained model, generate the prediction probabilities on the test data. Considering the threshold, obtain the predictions on the test data set.
* considering the thresholds value 0.50
* our model is 91% accuracy based on 95% of confidence interval accuracy lies between (0.8991, 0.9244)
* Sensitivity/Recall: 0.9000
* Specificity: 0.9229
* Pos Pred Value/precision: 0.9089
* we will calculate F1-score= 2\*(recall\* precision)/recall + precision
* f1<- 2\*(0.9000\* 0.9089)/ (0.9000+0.9089)
* f1 =0.9044281

**Business Recommendation:**

* It can be concluded that the variable 'Age' has a normal distribution of data. we can conclude that younger users spend more time on the site. This implies that users of the age between 20 and 40 years can be the main target group for the marketing campaign.
* Hypothetically, if we have a product intended for middle-aged people, this is the right site for advertising.
* Conversely, if we have a product intended for people over the age of 60, it would be a mistake to advertise on this site.