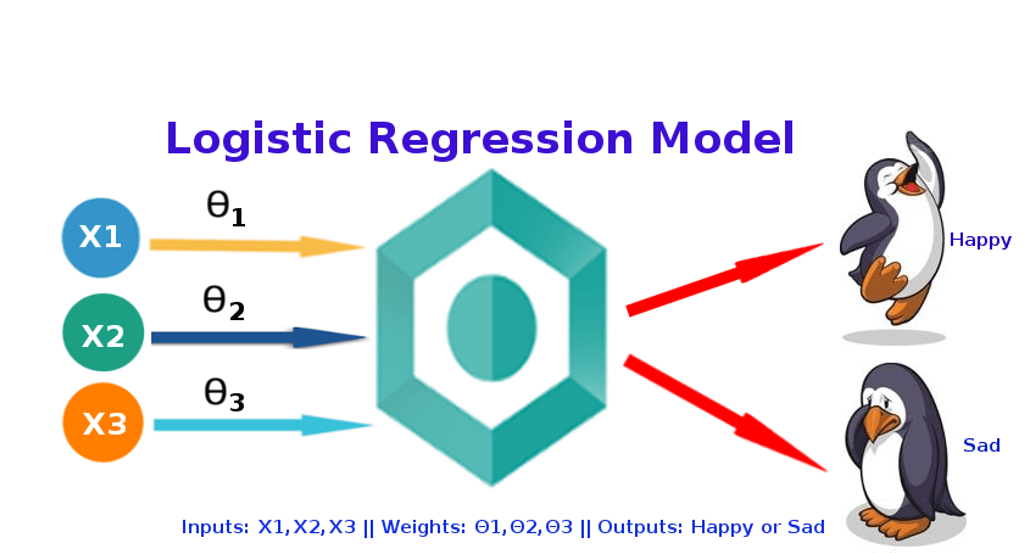
**Project On**

**Logistic Regression**

****

**Name:** Harshal Devidas Pagar.

**Class:** MSc Part 1

**Topic:** Ad-Click-Prediction

**INDEX**

|  |
| --- |
| **Introduction** |
| **Motivation** |
| **Problem Statement** |
| **About Data** |
| **Exploratory Data Analysis** |
| **About Model Analysis** |
| **Conclusion** |

**Motivation**

The main motivation behind the project is “Targeted Advertising”. At its most basic, targeted advertising can just mean that ads are chosen for their relevance to site content, in the assumption that they will then be relevant to the site audience as well. Online advertisers can use different methods to target a particular advertisement on the user based on its traits. Most of company do this as part of social media like Facebook, LinkedIn etc. But most of the times the process goes wrong and the advertisement does not reach its target audience because it is sent out without actually understanding the probability of the occurring click.

**Problem Statement::**

Publicizing is a multi-billion-dollar industry that goes about as a scaffold among organizations and their clients. While the vast majority are aware of the promotions around them, they represent the intensity of those advertisements and the impact of publicizing all in all. Research proposes that basically making somebody mindful of items, occasions, and brands expands the chances of that individual really purchasing those items, going to those occasions, or supporting those brands. Further, if an advertisement catches a man's thoughtfulness regarding the degree that he or she has a prompt, positive response to it, those chances of direct item commitment spikes significantly.

In this project, we are going to work on an advertising dataset, indicating whether or not a particular internet user has clicked on an Advertisement.

The goal is to predict if a user would click on an advertisement based on the features of the user. Few assumptions made as a part of this project is:

1. User taken into consideration are between the age group of 19 to 61.
2. There is almost equal ratio of male and female internet users.

**Using Python :**

**For this I used these libraries and pacakges ,**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

**Using ML Techniques:**

1. Exploratory Data Analysis (EDA)
2. Logistic Regression Analysis

**About Data Set::**

**Data Deploy On Github Repository:**

[**https://github.com/Harsh1800/Datasets/blob/main/advertisement%20dataset.txt**](https://github.com/Harsh1800/Datasets/blob/main/advertisement%20dataset.txt)

The dataset consists of below features:

Daily Time Spent on Site: consumer time on site in minutes

Age: Customer age in years

Area Income: Avg. Income of geographical area of consumer

Daily Internet Usage: Avg. minutes a day consumer is on the internet

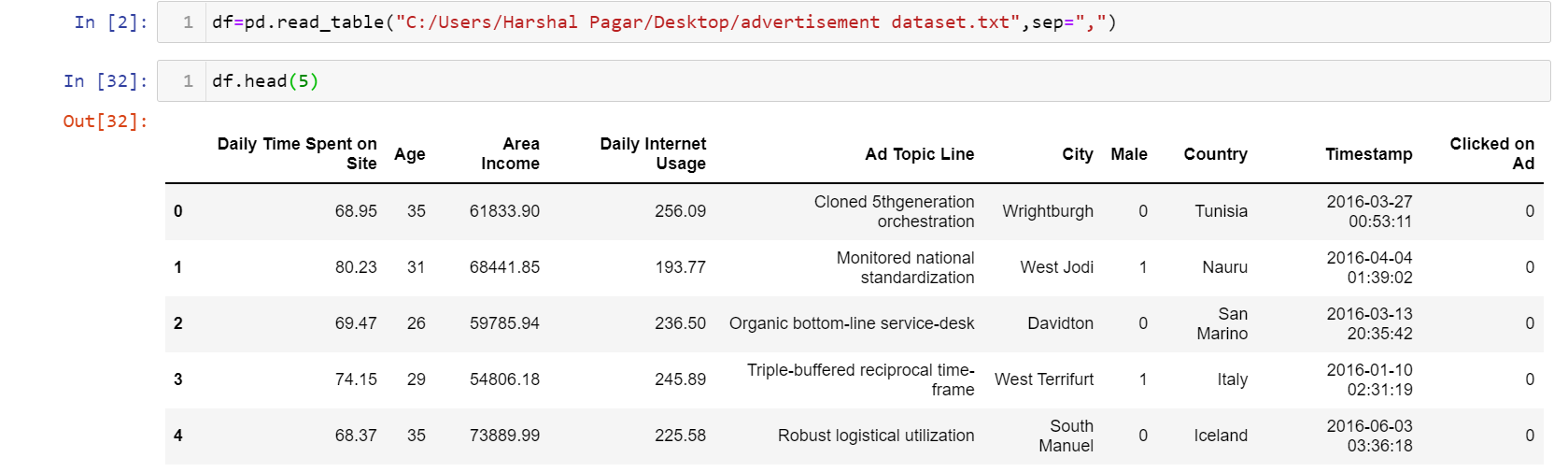
Ad Topic Line: Headline of the advertisement

City: City of consumer

Male: Whether or not consumer was male

Country: Country of consumer

I have done preliminary EDA(mention below) .

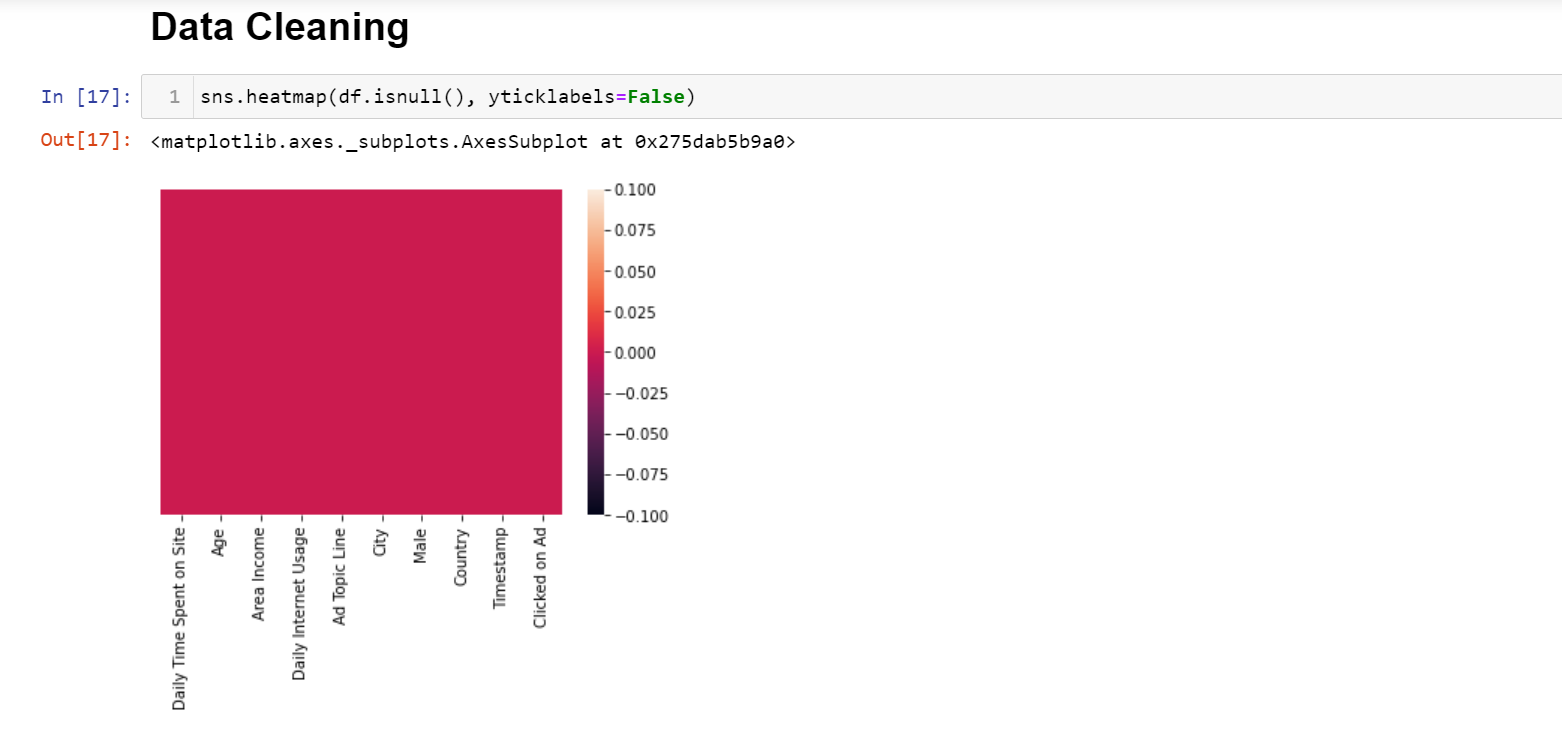


**Cleaning and Approaches::**

Considering the 'Advertisement Topic Line', I decided to drop it. In any case, it if we need to extract any form of interesting data from it, I can use Natural Language Processing.

As to 'City' and the 'Nation', I can supplant them by dummy variables with numerical features, Nonetheless, along these lines I got such a large number of new highlights.

Another methodology would be thinking about them as a categorical features and coding them in one numeric element. Changing 'Timestamp' into numerical value is more complicated. So, I can change ‘Timestamp’ to numbers or convert them to spaces of time/day and consider it to be categorical and afterwards I converted it into numerical values. And I selected the month and the hour from the time stamp as features.



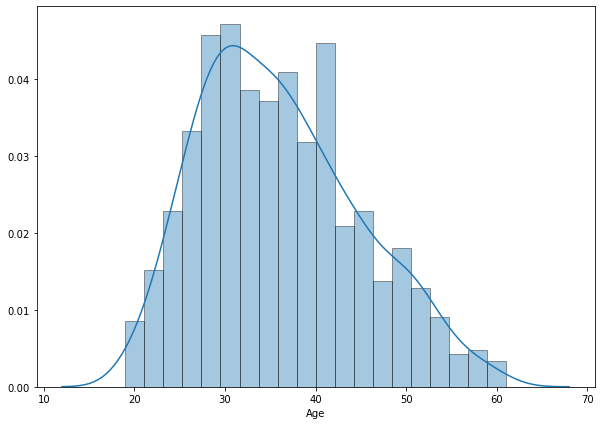
As we see ,here we don’t have missing data.

**Exploratory Data Analysis(EDA) ::**

I came up with some interesting questions on the dataset and I tried to find answers for the same during EDA process.

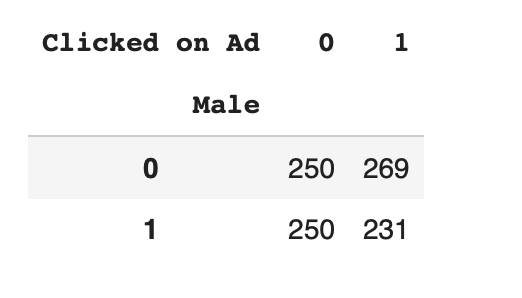
**What age group does the dataset majorly consist of ?**

I observe that the oldest person in the dataset is 61 years old and the youngest person is 19 years old and the average age as per dataset is 36 years old.



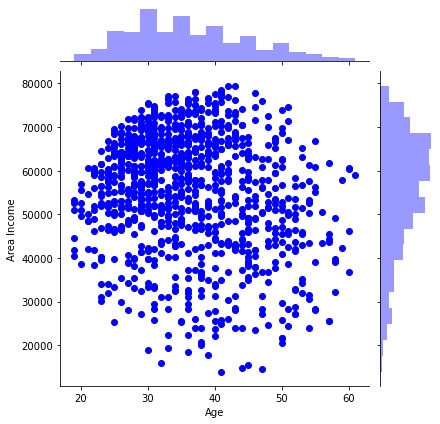
#### **Which gender has clicked more on online ads?**

Based on below data I can see that a greater number of females have clicked on ads compared to male.

[](https://user-images.githubusercontent.com/22437872/49623712-94f64680-f983-11e8-8572-ac20a0239c4d.png)

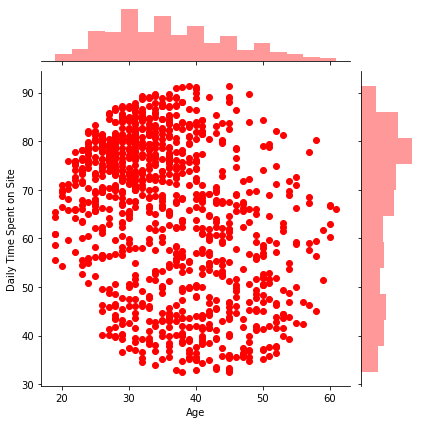
**What is the income distribution in different age groups ?**

Teenagers are higher earners with age group of 25-35 earning 58k-68k.



**Which age group is spending maximum time on the internet ?**

Age group of 25-40 is most active on the internet.



**Model Fitting::**

* **Logistic Regression::**

Model:



**What is Logistic Regression?:**

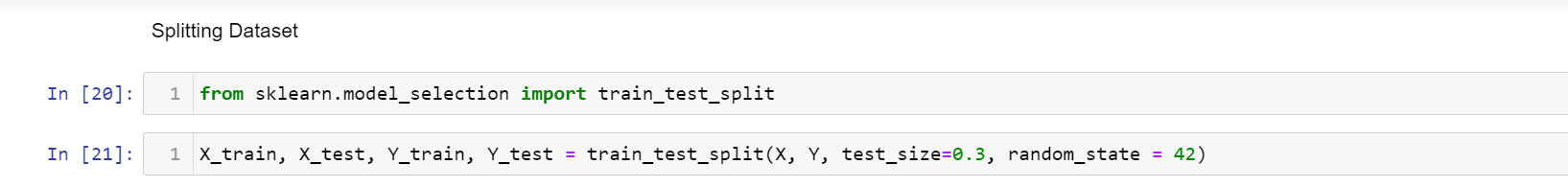
The logistic regression technique involves the dependent variable, which can be represented in the binary (0 or 1, true or false, yes or no) values, which means that the outcome could only be in either one form of two. **For example**, it can be utilized when I need to find the probability of a successful or fail event.

**Why we used Logistic Regression ?**

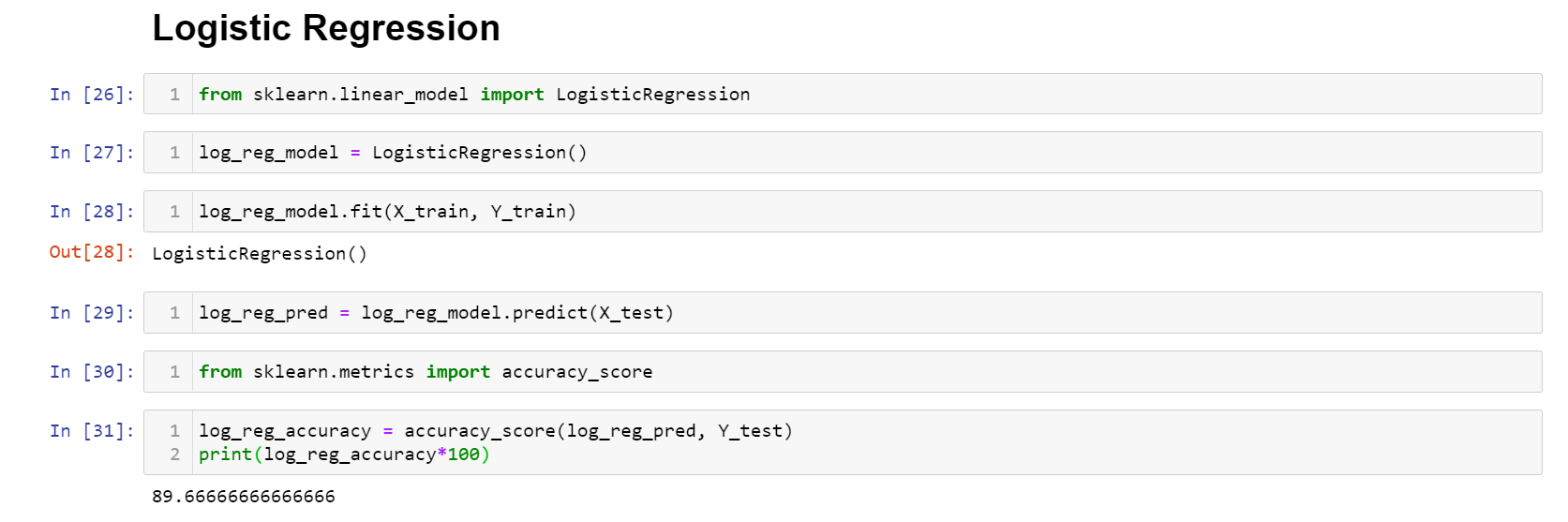
Logistic Regression is incredibly easy to implement and very efficient to train. So, it's always better start with a Logistic Regression model as a benchmark and try using more complex algorithms from there on. Convenient probability scores for observations

**Spliting data into train and test ..**

The train-test split is used **to estimate the performance of machine learning algorithms that are applicable for prediction-based Algorithms.**



#### **What accuracy are getting from Logistic Regression?**



**Conclusion::**

The above implementation model is good fit ,so I conclude that Logistic regression Algorithm gives perfect accuracy for determining the click probability. I believe in future there will be fewer ads, but they will be more relevant. And also these ads will cost more and will be worth it.

**Reference:**

[**https://en.wikipedia.org/wiki/Logistic\_regression**](https://en.wikipedia.org/wiki/Logistic_regression)