**Activity Report**

**On**

**Advertisement Prediction using Logistic Regression**

**For**

***(Data Analytics)***

**Submitted by**

**1930014 Devam Chakorborty**

**1930032 Ravi Prakash Mishra**

**1930033 Ravi Raj**

****

**B. Tech in Electronics & Computer Science Engineering**

**School of Electronics Engineering**

**Kalinga Institute of Industrial Technology Deemed to be University**

**Bhubaneswar, India**

**November 2021**

# **OBJECTIVE / AIM:**

This project aims to use previous advertisement engagement data to create a model that will predict whether a user will click on an ad or not in the future. This would help the advertisers use the appropriate technique to increase their overall revenue through targeted advertising. As this is a binary, classifying decision, we will be using logistic regression to model this data.

# **THEORY**

## INTRODUCTION

The Online Advertising Market is expected to grow at a CAGR (**compound annual growth rate**) of 14.3% over the forecast period (2021 - 2026). The Covid-19 pandemic had a positive effect on the Online Advertising Market. People became more attracted onwards online portals and social networking sites, which resulted in increased online population and streaming.

This report is basically about the advertising data of a marketing agency for the development of an algorithm that predicts if a user will click on an advertisement or not. Given data consists of some variables like **'Daily Time Spent on Site'**, '**Age**', ' **Daily Internet Usage'**, **'Area Income'**, **'Ad Topic Line'**, **'Timestamp'**, **'Male'**, **'Country'**, **City '** and **'Clicked on Ad'**.

SO, we will be looking at all other variables to accurately predict the value 'Clicked on Ad' variable. We will also perform some exploratory data analysis to see how 'Daily Time Spent on Site' in combination with 'Ad Topic Line' affects the user's decision to click on the ad.

## Concept

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no). The logistic regression equation has a very similar representation to linear regression. The difference is that the output value being modeled is binary in nature.

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection, etc.

# **Types of Logistic Regression** [**[1]**](#oone)

Generally, logistic regression means binary logistic regression having binary target variables, but there can be two more categories of target variables that can be predicted by it. Based on those number of categories, Logistic regression can be divided into the following types –

## Binary or Binomial

In such a kind of classification, a dependent variable will have only two possible types either 1 and 0. For example, these variables may represent success or failure, yes or no, win or loss etc.

## Multinomial

In such a kind of classification, the dependent variable can have 3 or more possible ***unordered*** types or the types having no quantitative significance. For example, these variables may represent “Type A” or “Type B” or “Type C”.

## Ordinal

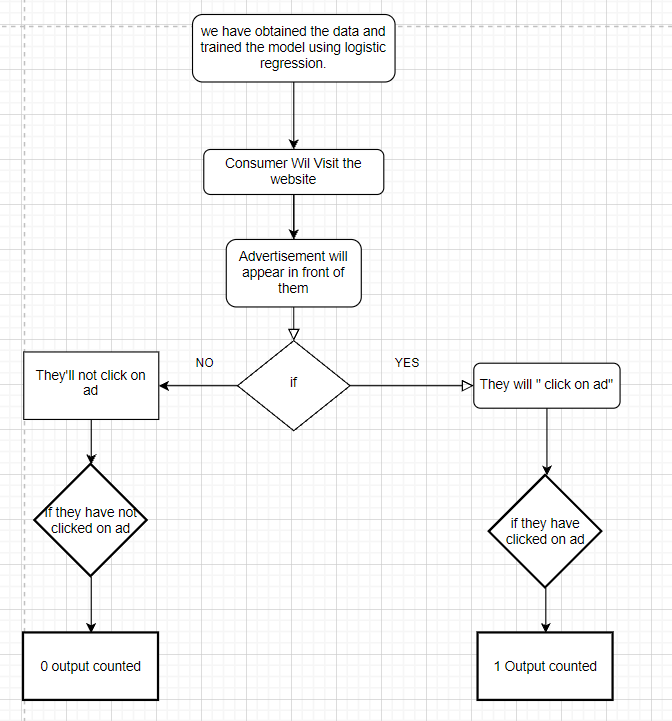
In such a kind of classification, the dependent variable can have 3 or more possible ***ordered*** types or the types having a **quantitative significance**. For example, these variables may represent “poor” or “good”, “very good”, “Excellent” and each category can have the scores like 0,1,2,3.

# **Logistic Regression Assumptions** [**[2]**](#oone)

Before diving into the implementation of logistic regression, we must be aware of the following assumptions about the same −

* In the case of binary logistic regression, the target variables must be binary always and the desired outcome is represented by the factor level 1.
* There should not be any multicollinearity in the model, which means the independent variables must be independent of each other.
* We must include meaningful variables in our model.
* We should choose a large sample size for logistic regression.

# **FLOWCHART**



# **REQUIREMENTS**

* A system with an installed **python environment** and some **libraries** to work upon it like:

**padas**,

**NumPy**,

**seaborn**,

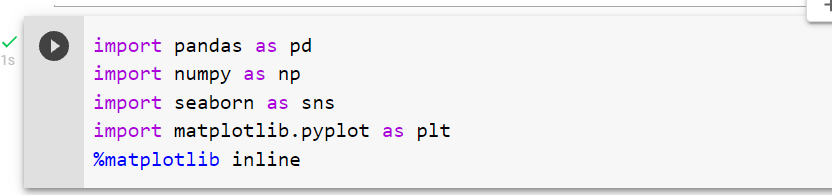
**matplotlib.pyplot**,

**sklearn.linear\_model, etc.**

* **Data Set** “asvertising.csv” [**[3]**](#tthree) (to work upon).

# **CODE**

Now, Importing the libraries and retrieving Data,



Loaded the data into variable **“ad\_data”**, and viewed the dataset first 8 columns, the dataset is having few information regarding each column like

'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

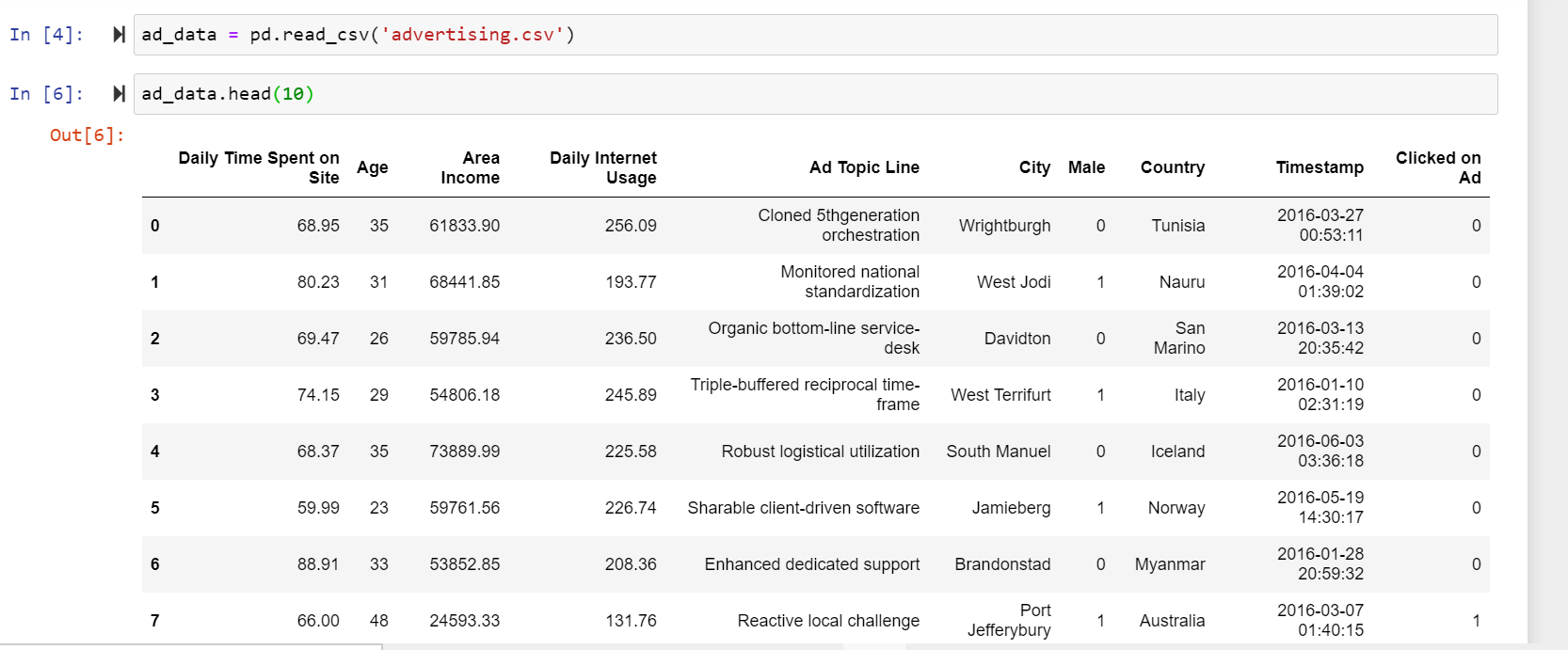
'City': City of consumer

'Male': Whether or not consumer was male

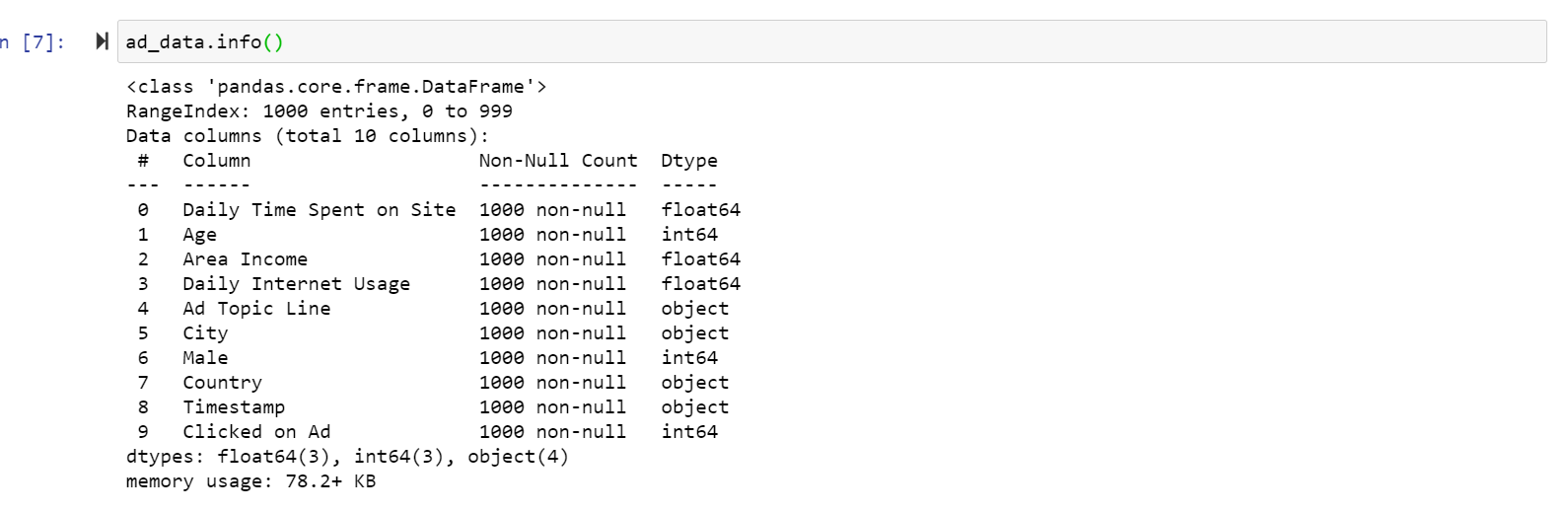
'Country': Country of consumer

'Timestamp': Time at which consumer clicked on Ad or closed window

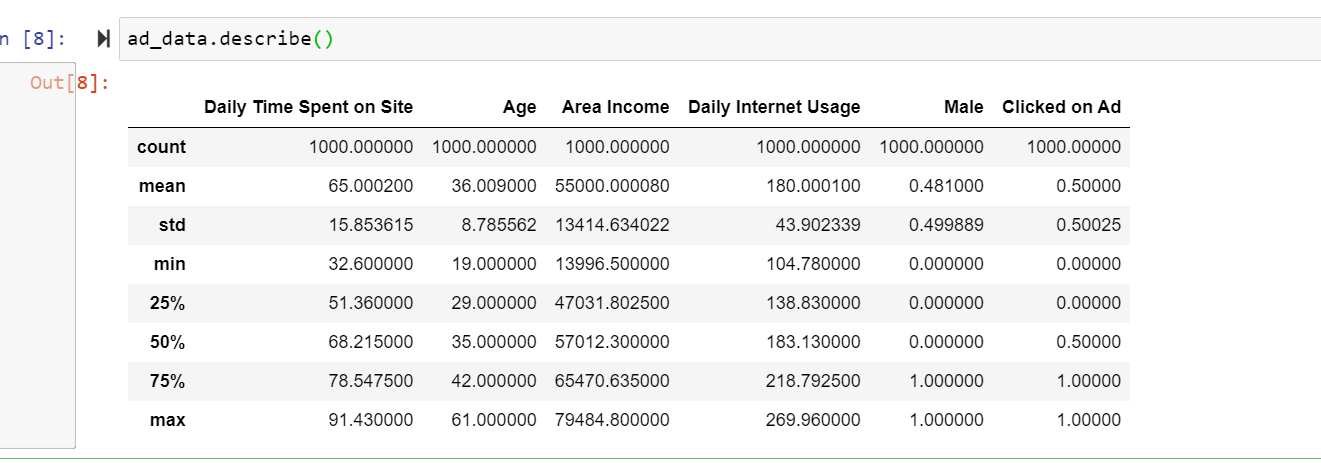
'Clicked on Ad': 0 or 1 indicated clicking on Ad



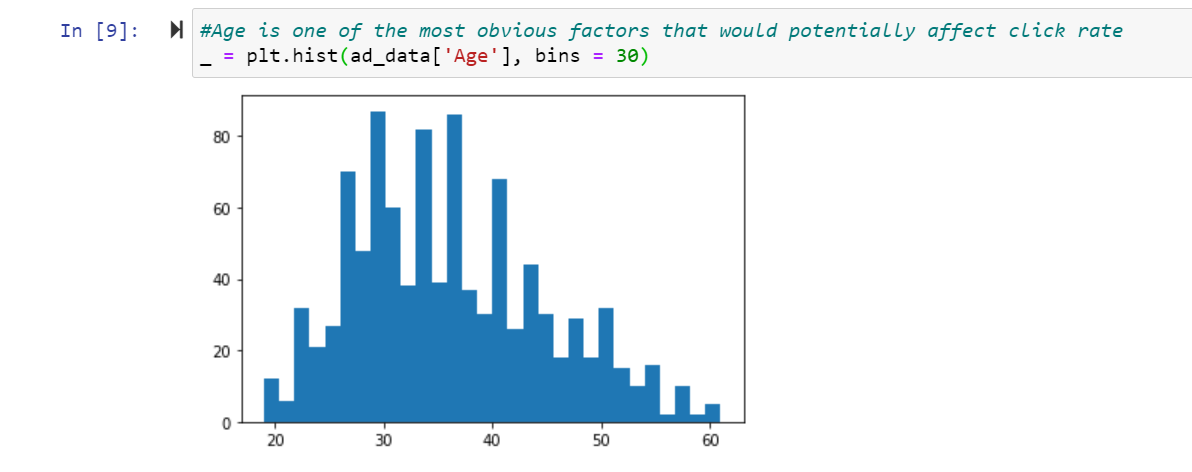
**ad\_data.info()** info about data so that we must not miss, as all the 1000 entries are present and we need not do any data cleaning.

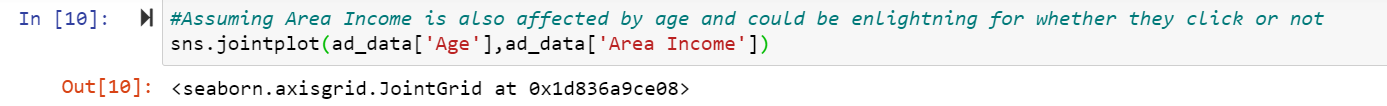


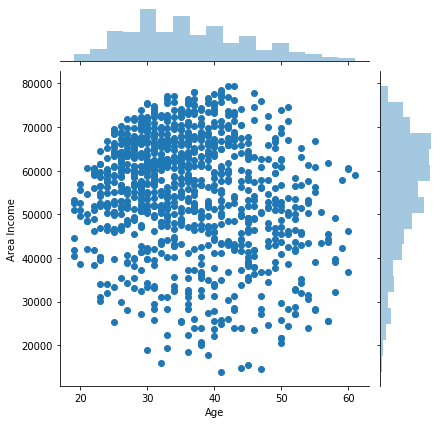
**Describe()** function to gain insight into the ranges in which variables change



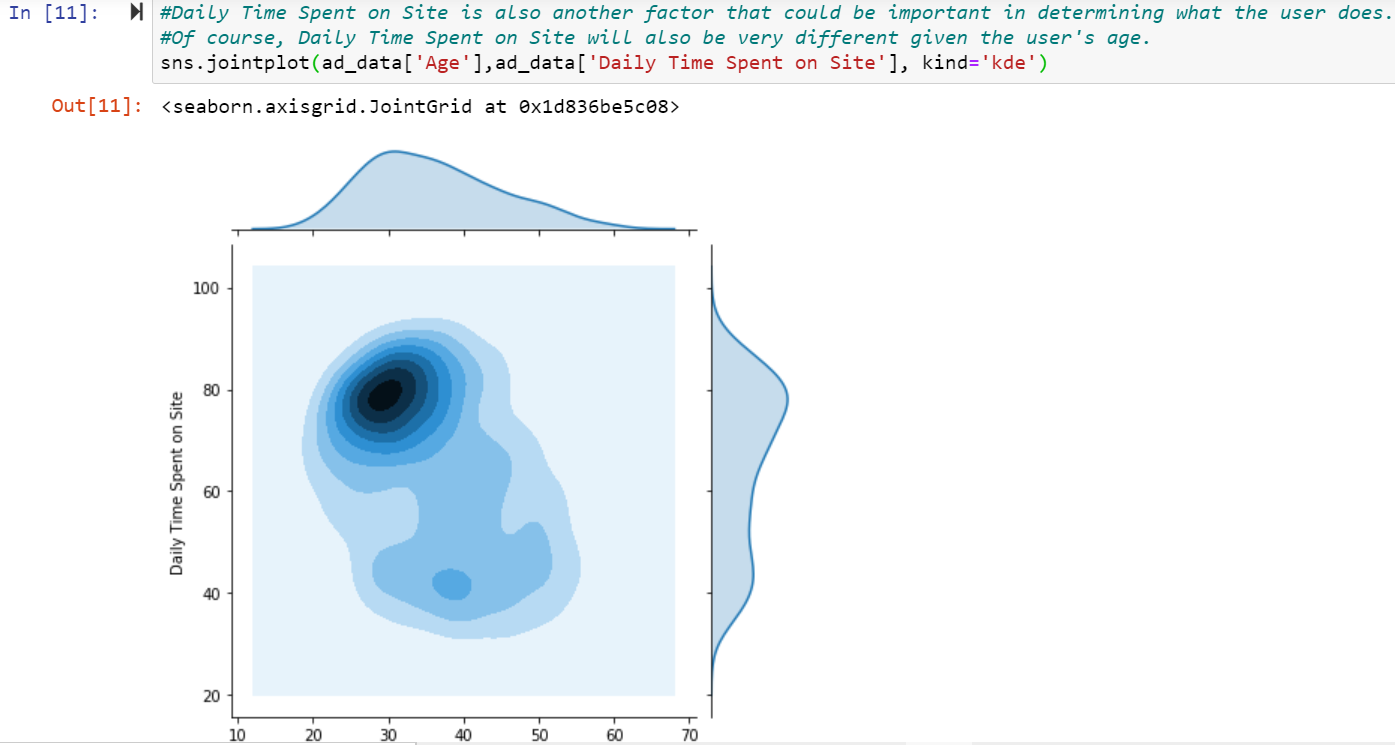
Now first, we'll **explore the data** to get a better understanding of what is happening. With Clicked on Ad being the variable in question, we'll be looking for the factors that could impact this the most.



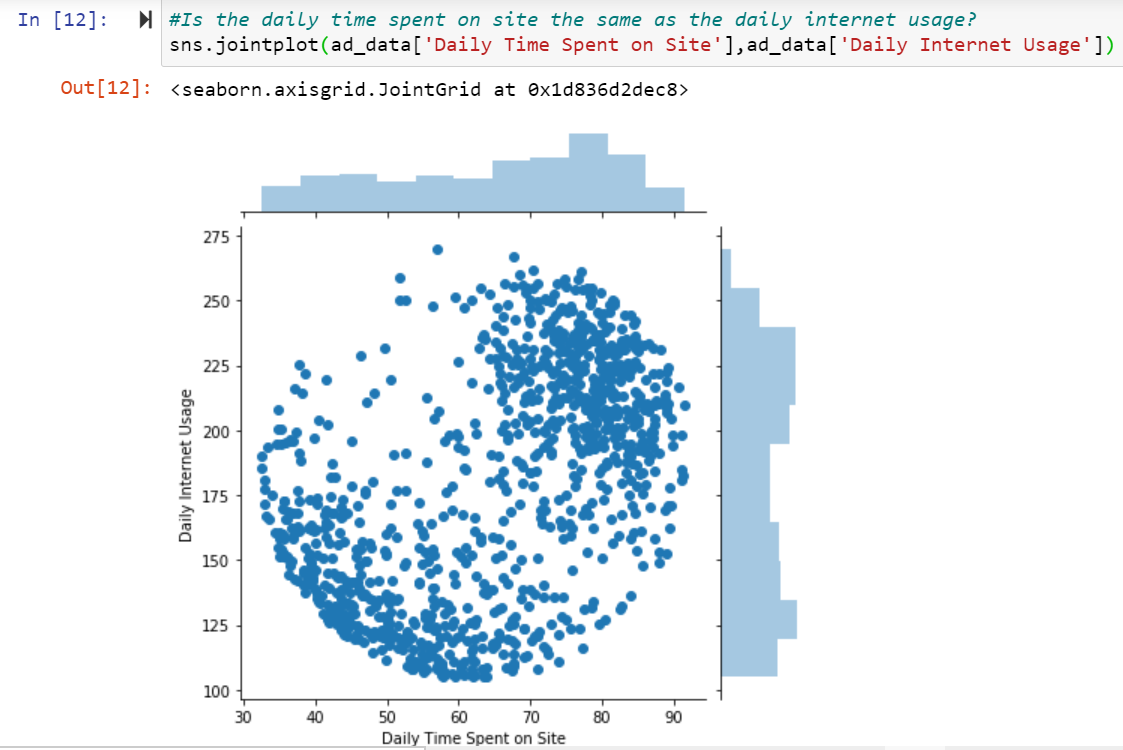




We can see a **larger concentration in the top left**. Knowing there are more users of younger age, it makes sense that there would be a higher density on the left side of the graph. However, It looks like the majority of those younger people have an income above $45,000.

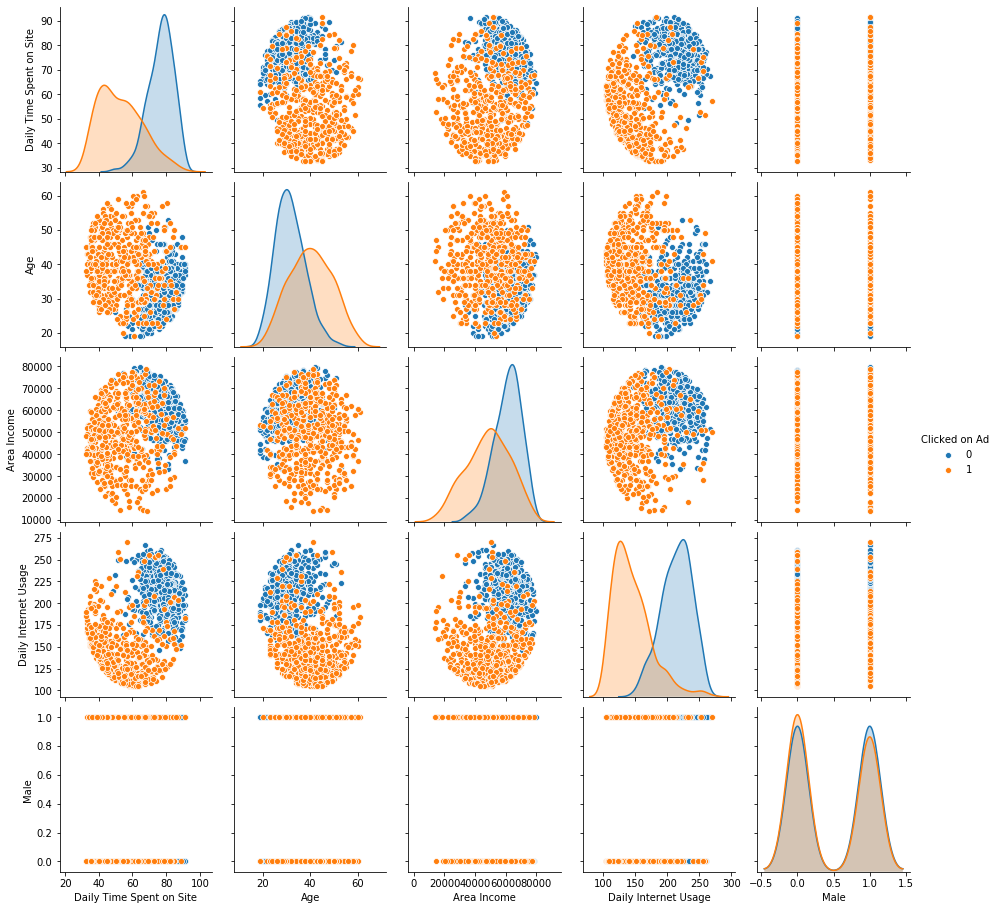


**We** see from this that most of these users are between 25 and 40 and spend over 60 minutes on the site each day, which means young ones are spending more time.



With almost a circular dataset, users' usage is pretty spread out, with concentrations roughly being in high internet usage and high daily time spent on site and low internet usage and low daily time spent on site.

In the below figure, with a better understanding of what our users are like, let's plot them all together in regards to whether they clicked on the Ad or not to give us some visual understanding of what's happening.



What's clear from the above is that there is a higher concentration of people who clicked on the ad if they:

1. Spent less time on the Site
2. Spent less time on the Internet every day

Visually, these seem to be the identifying factors in whether they will click on an ad. Now I'll move on to performing a logistic regression to determine if these variables are as important as I hypothesize.

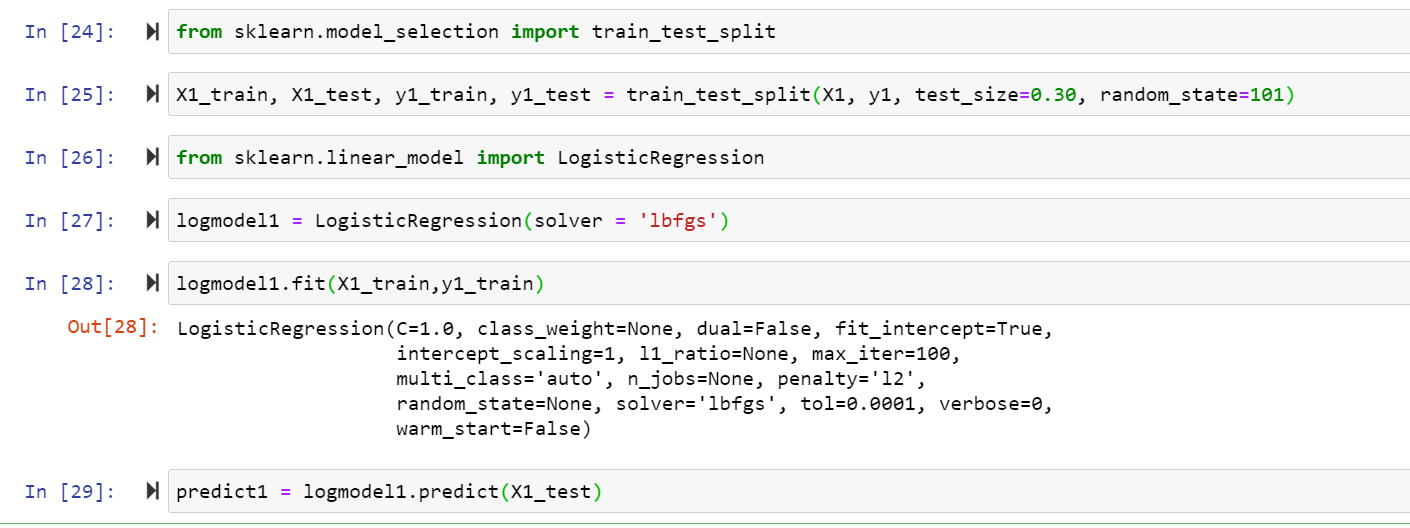
# **OBSERVATIONS/RESULTS**

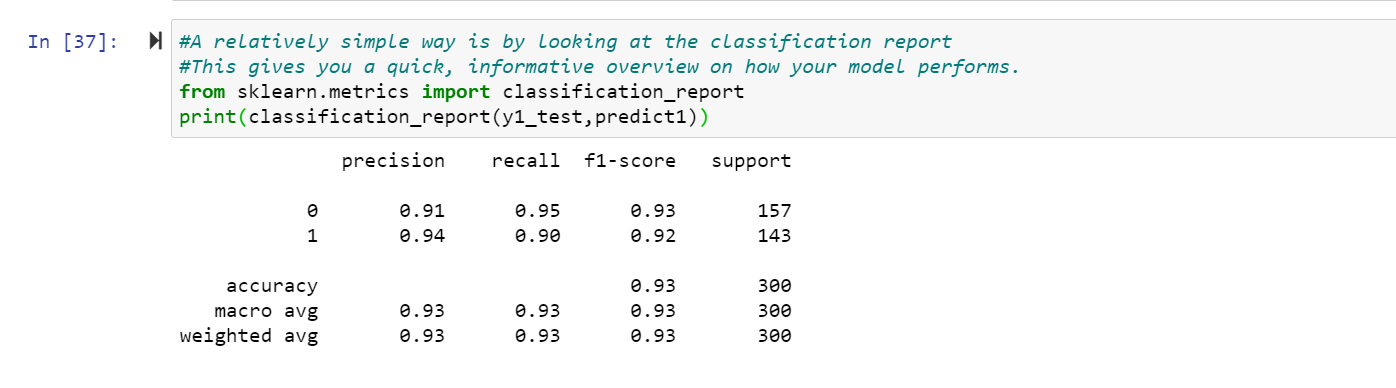
## Building a Logistic Regression Model, Training and Predicting.

It's time to split the data set into two, train the model, and test how it performs. We’ll be dropping the **Ad Header** and **City columns** for the first model.

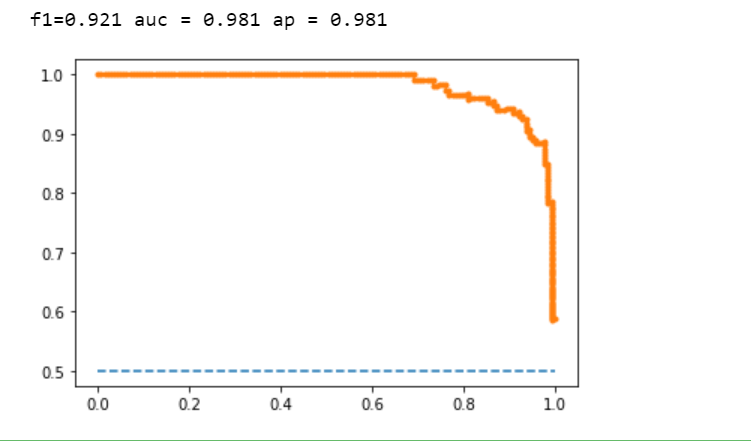
I'm assuming, for this relatively small data set, the ***city will not be impactful*** enough to make much of a difference in my analysis, though I will be keeping the country in as a dummy variable. We'll be changing the timestamp data into months.

After both our X data and y data are formatted correctly, I'll split the data into training and test data sets to cross-validate our model when we want to see how it performs.





We can also draw a diagram of a **precision-recall curve** as,



# **DISCUSSION OF RESULTS**

**Precision**, **recall**, and **f1-score** are used to understand how your data performs, especially on an imbalanced dataset. Our data is balanced, meaning the column to be classified has a 50/50 chance of being clicked on or not.

**Precision** is the ratio of correctly predicted positive outcomes to all positively predicted outcomes - it essentially tells us "How many people clicked after being labeled they would?".

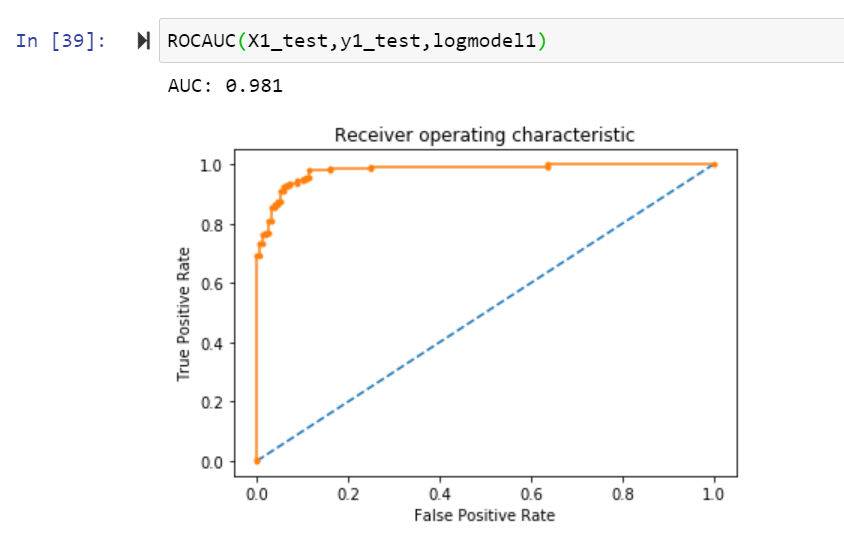
The **recall**is the ratio of how many correctly predicted positive outcomes to how many people survived in total (thus including rows that the model labeled that they wouldn't click on the ad even though they did) - this answers the question "Of those who clicked on the ad, how many did we predict to do so?"

The **F1-score** is a weighted average of Precision and recall. It's harder to interpret but allows us to combine precision and recall and understand what the overall performance is to be used in comparison with other models.

Remember that a completely random model would get 0.5 (50/50 chance) for these metrics, and a perfect model would be 1.0

It's also important to note that the classification report and the precision-recall curve metrics are different, even though they measure the same thing, due to how they are measured. The classification report only uses a binary yes/no outcome to measure these metrics, whereas the precision-recall curve breaks each row into a probability of being yes, causing the differences. Both are important to understand because both give slightly different answers depending on the question you are asking.

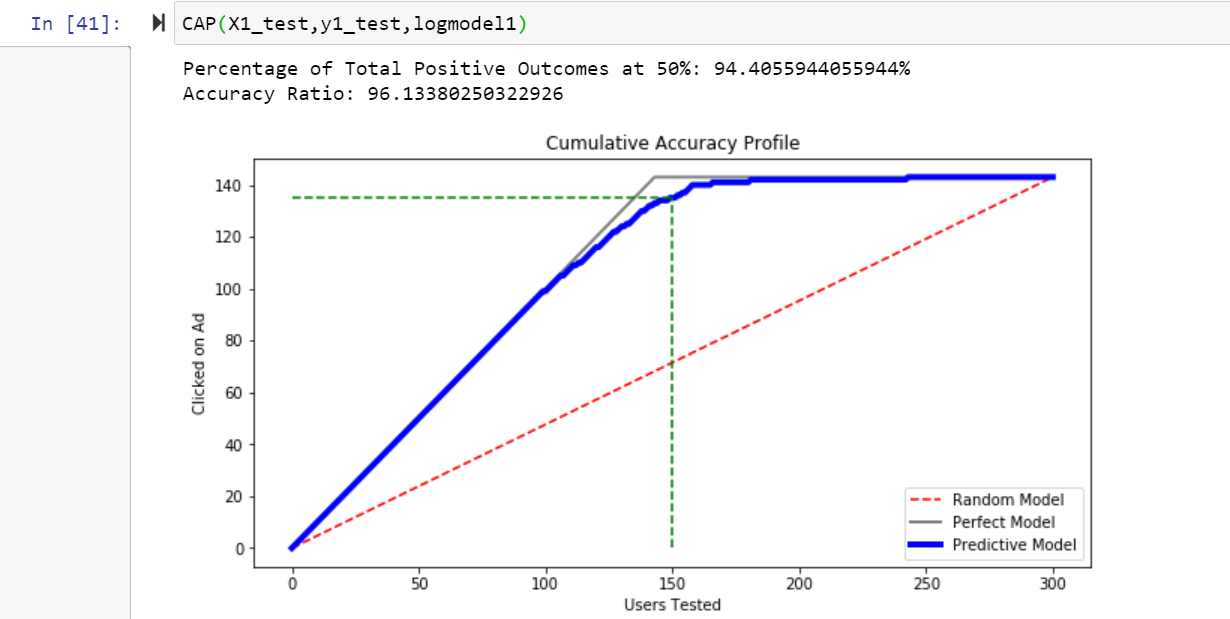
**With** this said, **anything over 0.9 is very good**! But we'll also want to see what else we can learn about our model



This is the **ROC curve**. The better the model, the larger the Area Under the Curve (AUC), and the farther away to the top left the model will be compared to the dotted line (a completely random model). Ours with an AUC of 0.98 is very good!

As we've observed so far, our model seems to be good. But how can we apply this?

Well, for starters, we can look at the **Cumulative Accuracy Profile (CAP)** Curve to help us understand how well if we predicted a dataset, we could determine if they would click on an ad or not.



We see that our percentage of **total positive outcomes** at 50% is **94%**, which means that we can identify 94% of the people who will **click on our ad** by just applying our model to the top 50% of the users in our test data. Typically, anything **above 90% is an exceptional model**, meaning this model is very good. This is valuable information for budgeting for a given user set. To apply this concept, you could apply this model directly to each user coming in and only show the ad to users with a high probability rating or you could apply the model to a new data set of users and you could order the data by the probability that they'll click the ad and then use this information to choose which users you want to contact out of the list.

But what we're primarily interested in is what variables/attributes of users are important for them to click on the ad according to our model. For that, we'll create a coefficient report.

# **CONCLUSION:**

In this analysis, we made a model using logistic regression to help us try to predict whether or not users will click on our ads. We made an excellent model (at least partly attributed to having a fictional dataset). We also saw that, by splitting the country into a dummy variable, we were unable to use Recursive Feature Elimination and we saw the effects of that on our model.

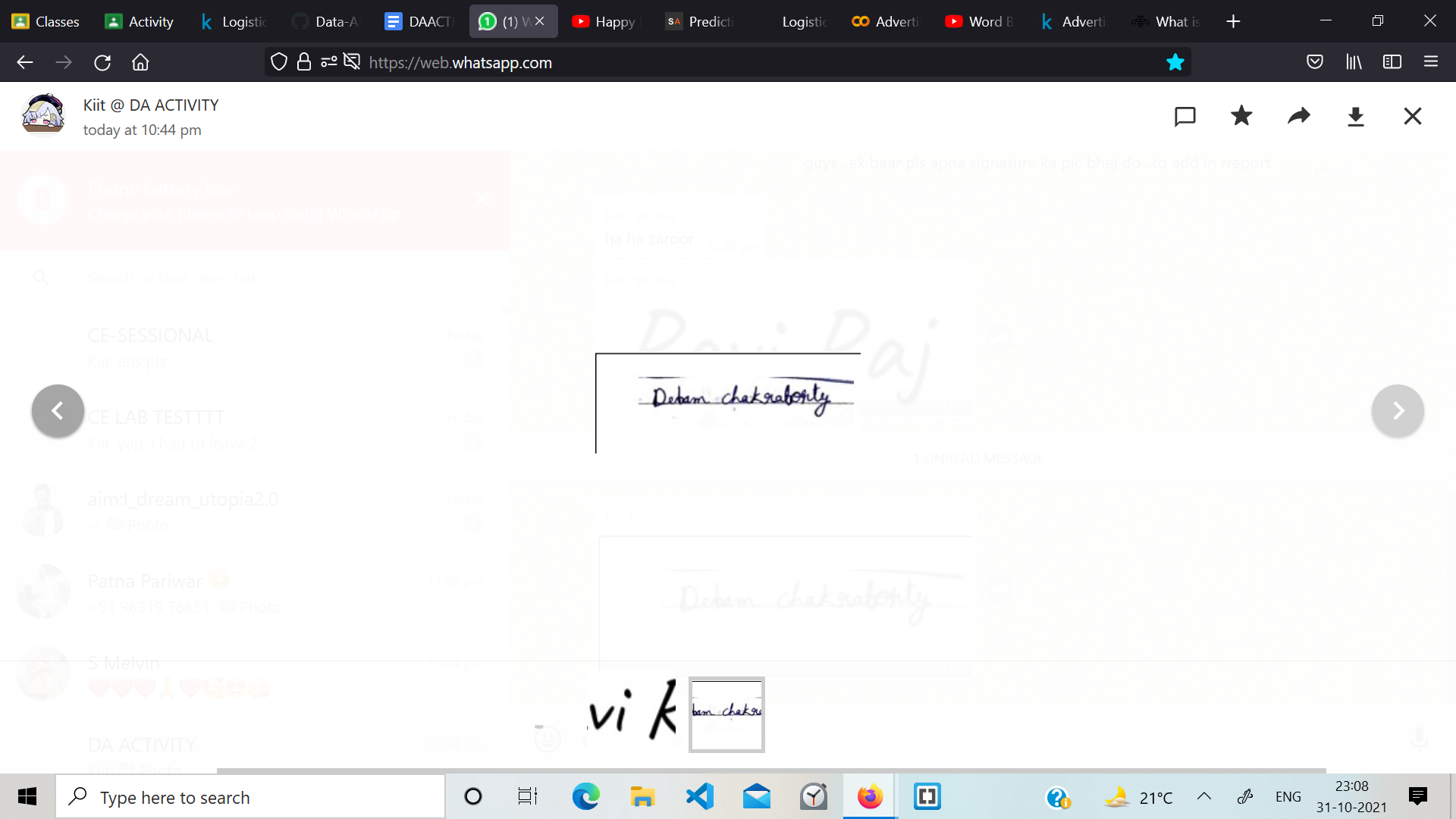
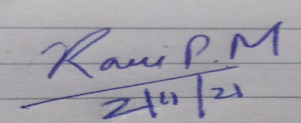
We were able to interpret our model to help us understand what variables affect whether someone will click on an ad; namely, we found that an increase in Age had the largest impact on the odds of someone clicking our Ad.

**Finally**, we noted along the way that our model could be used not only in predicting whether a future user will click on an ad or not, but also in other applications, from budgeting how much we should spend on advertising to identifying what kind of users we should target with our ad.

# **References**

|  |  |
| --- | --- |
| [1] | S. Swaminathan, "Towards Data Science," 15 March 2018. [Online]. Available: <https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc>. |
| [2] | IBM, "IBM," [Online]. Available: <https://www.ibm.com/topics/logistic-regression.> |
| [3] | "Kaggle," [Online]. Available: <https://www.kaggle.com.> |

**STUDENT SIGNATURES**



**SIGNATURE OF THE CONCERNED FACULTY**