Data and Web Mining Course

Predicting whether a user will click on a advertisement based on the features of the user

-BY
Sreeya Raavi

ABSTRACT

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:

- Users taken into consideration are basically based on their age. Different age groups will click their respective ads.
- There is an almost equal ratio of male and female internet users.
- The ad topic is limited to what is given in the dataset.

The algorithms we used are Logistic regression, Support Vector Machine, Linear Regression. We will work with the advertising data of marketing agency to develop a machine learning algorithm that predicts if a particular user will click on an advertisement. The data consists of 10 variables: 'Daily Time Spent on Site', 'Age', 'Area Income', 'Daily Internet Usage', 'Ad Topic Line', 'City', 'Male', 'Country', 'Timestamp' and 'Clicked on Ad'.

The acquired outcomes demonstrated the utilization estimation of both machine learning models.

The logistic regression model demonstrated somewhat preferred execution over the other two, however unquestionably, all the three models have indicated that they can be successful in solving classification problems.

Keywords — Machine Learning, Classification, prediction, Logistic , SVM, Linearregression.

INTRODUCTION

We are visiting take a shot at a advertising dataset, demonstrating whether a chose web client has tapped on an ad. The objective is to anticipate if a user would tap on an announcement upheld the highlights of the user. Not many presumptions made as a segment of this venture are Clients thought about are essentially upheld their age. Diverse age gatherings will click their particular advertisements, there's a practically equivalent proportion of male and lady like web clients and furthermore the promotion theme is confined to what exactly given inside the dataset.

Internet (Online) promoting is moreover a boundless business worth more than 50 billion dollars. The income that sponsors procure is expanding a result of their act of focused promoting. numerous explores are depleted the earlier years, from promotion click guess to advertisement serving. The promotion click prediction has been utilized in history in a wide range of ad design like, web indexes, printed and relevant notices, video commercials and so forth. With the fast increment of publicizing organization, click prediction needs immense information investigation. The commercial prediction is mulled over to be one through and through the preeminent worthwhile stories inside the space of MACHINE LEARNING. In general, the methodology, technique in machine learning, brings a change to the ad-click prediction method. The strategies incorporate information arrangement, quality removal, advertisement click anticipation and promotion serving. The highlights utilized during this work are altogether human based which contributes parts to created and-click prediction framework it's acclimated decide the value of perpetual variable. The Logistic Regression model is a calculation that utilizes a logistic capacity to demonstrate dependent factors. it's an apparatus for prescient examination. Logistic regression is generally utilized for order purposes. In contrast to regression, the variable can take a predetermined number of qualities just i.e., the number is all out. At the point when the quantity of potential results is only two it's called Binary Logistic Regression. We should examine how logistic regression are utilized for order assignments. In regression, the yield is that the weighted amount of information sources. In the event that we take the weighted amount of information sources on the grounds that the yield as we squander measurable technique, the value is more than 1 however we wish a cost somewhere in the range of 0 and 1. That is the reason regression toward the mean can't be utilized for characterization undertakings.

PROBLEM SURVEY:

Our team conducted a survey in which questions such as "Does the user know about Machine learning?" "Does the user think machine learning approaches are good or bad?" and the reasons behind having such an opinion. About 72% of those who took the survey were aware of Machine learning. 45% of the respondents thought that machine learning approaches were good and bad.

51% thought they were good and 21% thought they were bad.

Respondents to the survey believed that Machine learning approaches were adequate because companies tend to promote their goods on blogs and social media channels. However, in online marketing, finding the right audience remains a challenge. It can be expensive to spend millions on advertising to an audience that is unlikely to purchase your goods. Since few algorithms in previous works showed less accuracy, respondents thought machine learning approaches were poor.

DATASET DESCRIPTION:

In this project we have selected an advertising dataset namely advertising.csv. The dataset consists of 1000 observations, and the dataset consists of the following attributes:

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

DATASET PREPROCESSING:

Data processing in data mining is the process which is used to transform raw data in a useful and appropriate format, so that we can use the data for further process easily. In out project we will process the data in dataset to make it perfect i.e out of all attributes mentioned we used only-daily time spent on site, age, area income, male and clicked on ad to make it easy and efficient to the code.

IMPLEMENTATION:

The principle inspiration driving the undertaking is "Focused on Advertising". At its most basic, targeted advertising can just mean that ads are chosen for their relevance to site content, in the suspicion that they will at that point be applicable to the site crowd also. Online promoters can utilize various strategies to focus on a specific

commercial on the client dependent on its qualities. The majority of organizations do this as a feature of online media like Facebook, LinkedIn and so forth Yet, the greater part of the occasions the cycle turns out badly and the ad doesn't arrive at its intended interest group since it is conveyed without really understanding the likelihood of the happening click. Web advertising has taken over customary promoting systems in the ongoing past. Organizations like to publicize their items on sites and online media stages. Nonetheless, focusing on the correct crowd is as yet a test in internet promoting. We will work with the advertising data of a marketing agency to develop a machine learning algorithm that predicts if a particular user will click on an advertisement.

Mathematical Model and equations:

- Logistic Regression: It is used when the dependent variable(target) is categorical. It is also appropriate regression analysis to conduct when the dependent variable is dichotomous(binary). Binary outcomes can be predicted from the independent variable.

To predict the class of binomial target features we use logistic regression function.

$$f(z)=1/1+(e^{(-z)})$$

- Support Vector Machine: It is one of the most popular supervised learning algorithms, which is used for classification as well as regression problems. However, primarily, it is used for classification problems in Machine Learning. In an SVM model, each data item is represented as points in an n-dimensional space where n is the number of features where each feature is represented as the value of a coordinate in the n-dimensional space.

SVM is the classifier that maximizes the margin or it is a frontier which best separates the two classes.

$$g(x) = ((w^T) *x) + b$$

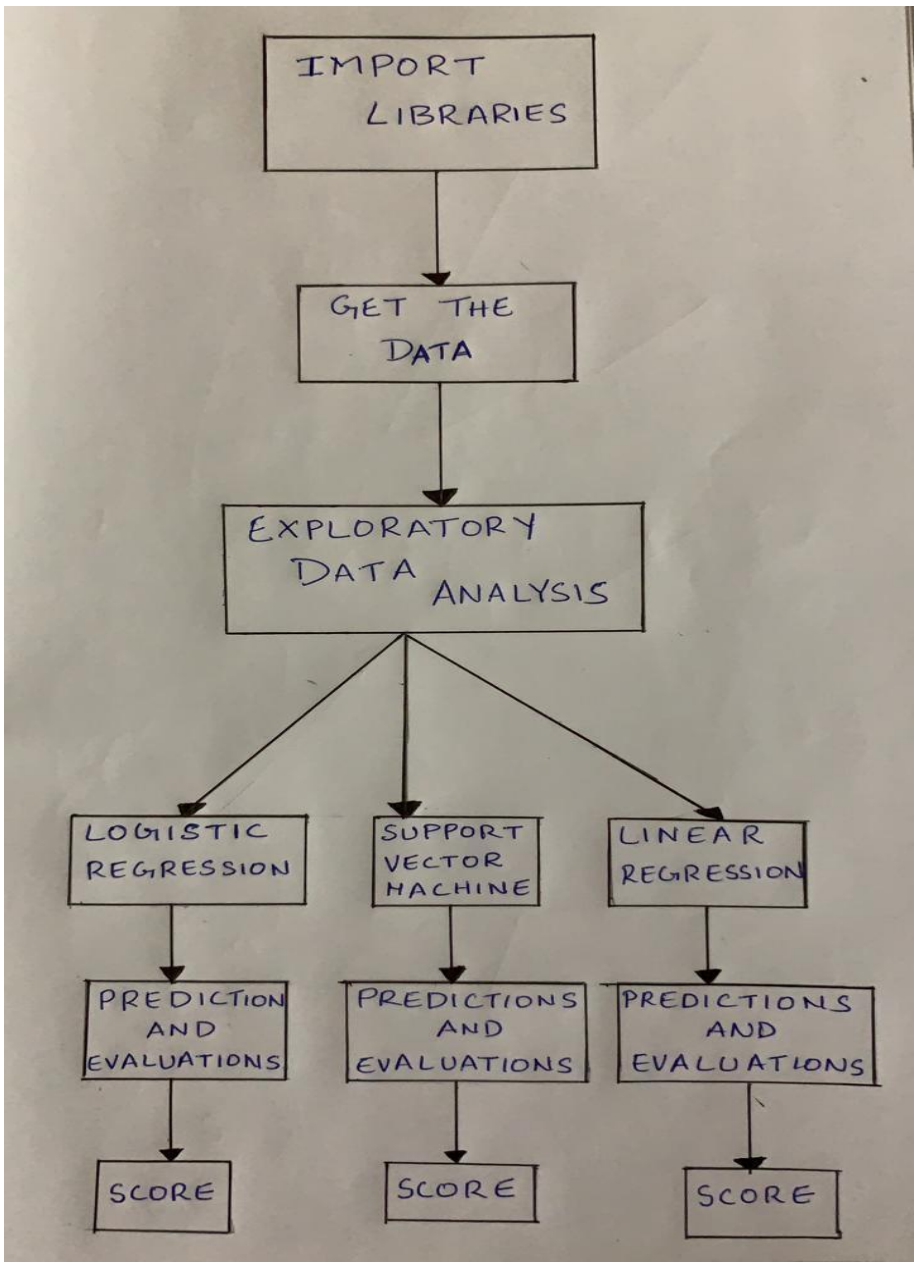
- Linear Regression: It attempts to model a relationship between two variables by fitting a linear equation to observe data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

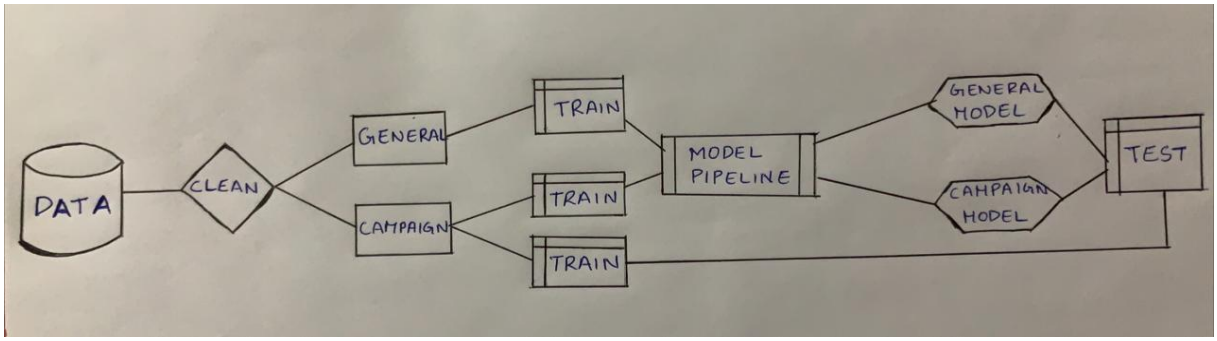
The linear regression algorithm represents a

relationship between one or more dependent variable(s).

$y = a + a x + \varepsilon$

FLOW CHARTS:





EXPERIMENTAL SETUP :

Libraries used:

- Numpy
- Pandas
- Seaborn
- Matplotlib

Software used:

- Jupyter Notebook is the IDE used to launch and run the program.

Programming language used:

- Python

Dataset used:

- Advertisement.csv

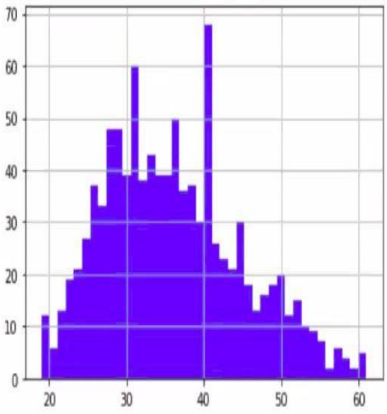
Attributes used:

- Daily Time Spent on Site’:consumer time on site in minutes
- 'Age': cutomer 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

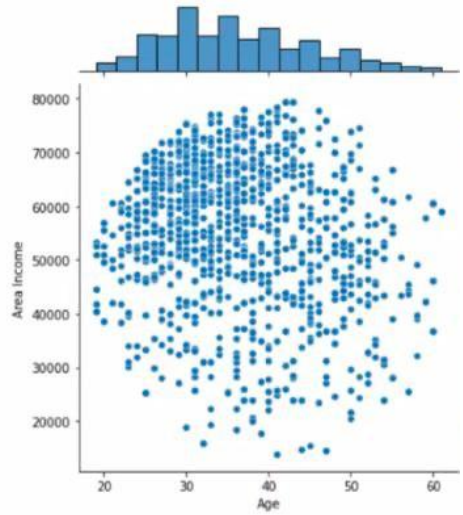
Graph Screen shots:

The graph screen shots which we have implemented in this project. They are:

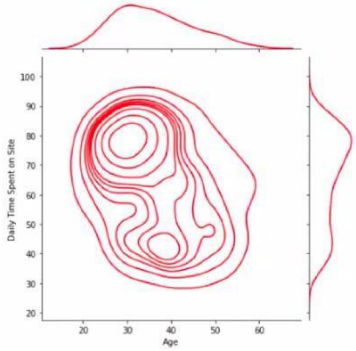
- Creating histogram of the age

-
```
ad_data['Age'].hist(bins=40,color='blue',alpha=0.7)
```
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f97646e3b00>
```


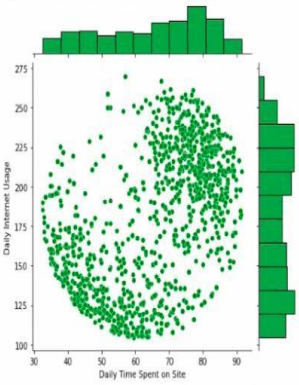
- Creating a joint plot showing area income versus age

-
```
sns.jointplot(x="Age",y="Area Income",data=ad_da
```
```
<seaborn.axisgrid.JointGrid at 0x7f97646980b8>
```


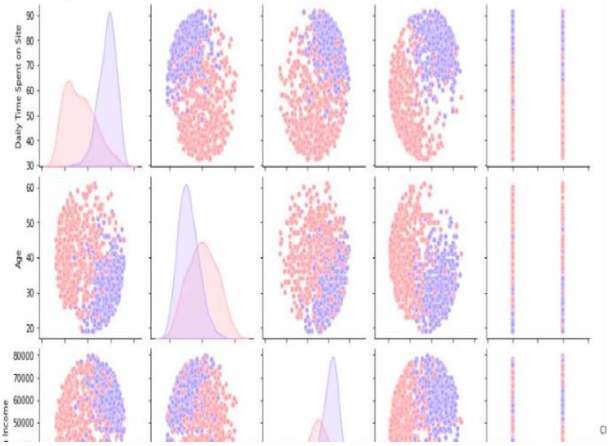
- Create a joint plot showing the ked distribution of daily time spent on site vs age

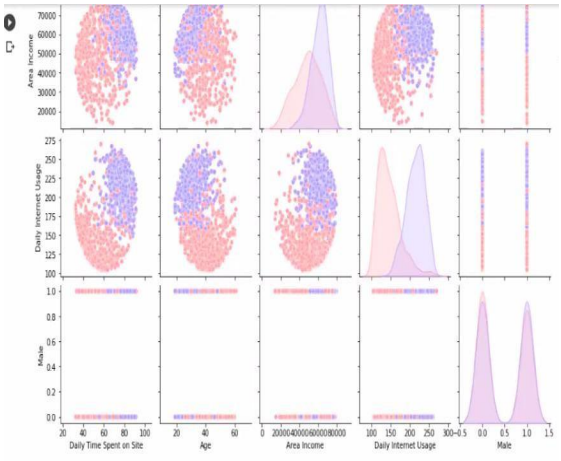
-
```
sns.jointplot(x="Age",y="Daily Time Spent on Site",data=ad_data,kind='kde',color='red')
```
```
<seaborn.axisgrid.JointGrid at 0x7f9764767eb8>
```


- Creating a joint plot 'daily time spent on site vs daily internet usage

-
```
sns.jointplot(y="Daily Internet Usage",x="Daily Time Spent on Site",data=ad_data,color='green')
```
```
<seaborn.axisgrid.JointGrid at 0x7f9764420438>
```


- Create a pair plot with hue defined by the 'clicked on ad' column feature

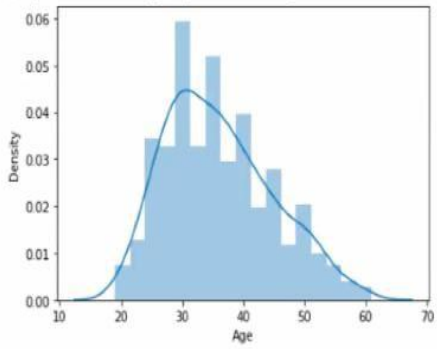
-
```
sns.pairplot(ad_data,hue='Clicked on Ad',palette='bwr')
```
```
<seaborn.axisgrid.PairGrid at 0x7f97647f8e10>
```




- Create a dis plot showing age

```
sns.distplot(ad_data['Age'])
```

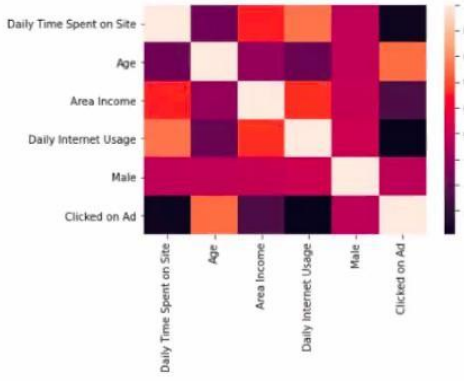
```
/usr/local/lib/python3.6/dist-packages/seaborn/distrib
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f076264b8
```



- Heat map

```
sns.heatmap(ad_data.corr())
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0762582fd0>
```



Logistic Regression:

Three cases are implemented in the project-

Test case: 45%

```
[675] print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
0	0.51	1.00	0.68	231
1	0.00	0.00	0.00	219
accuracy			0.51	450
macro avg	0.26	0.50	0.34	450
weighted avg	0.26	0.51	0.35	450

Test case: 60%

```
print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
0	0.88	0.89	0.88	300
1	0.89	0.87	0.88	300
accuracy			0.88	600
macro avg	0.88	0.88	0.88	600
weighted avg	0.88	0.88	0.88	600

Test case: 42%

```
[675] print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
0	0.51	1.00	0.68	231
1	0.00	0.00	0.00	219
accuracy			0.51	450
macro avg	0.26	0.50	0.34	450
weighted avg	0.26	0.51	0.35	450

Support Vector Machine:

Test case: 42%

```
print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
0	0.66	0.84	0.74	21
1	0.78	0.57	0.66	20
accuracy			0.70	42
macro avg	0.72	0.70	0.70	42
weighted avg	0.72	0.70	0.70	42

Linear Regression:

Test case: 25%

```
[729] y_train_prediction=model2lr.predict(X_train)

[730] y_test_prediction=model2lr.predict(X_train)

[731] score = ('Accuracy of linear regression classify on test:{:.2f}'.format(model2lr.score

score

'Accuracy of linear regression classify on test:0.58'
```

CONCLUSION:

Comparing all the above implementation models, we conclude that Logistic regression(accuracy 91%) gives us the maximum accuracy for determining the click probability.

We 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. In the other hand we also implemented Support Vector Machine which has an accuracy of 70% and linear regression which has an accuracy of 58% which did not give better accuracy compared to the logic regression. So, we conclude among all the algorithms implemented logistic regression is the best one to get the output.