PREDICTION FOR DISPLAY ADVERTISING

OBJECTIVE

The objective of the proposed system is to predict if an ad will be clicked by the user or not.

PROBLEM

In the recent time, internet and social media users have grown tremendously. As a result of this, many companies prefer to advertise their products on websites and social media platforms. However, targeting the right audience is still a challenge in online marketing.
Companies spend lot of money to display the advertisement of their products and if the consumers are not likely to buy the product, the advertisement cost can be a burden for the company.

This comes as a challenge for the company to analyze online internet users who view advertisements on their web pages. Always, a higher value of Click Through Rate plays a crucial role in increasing the revenue of the business. Showing the user an Ad that is relevant to his/her need greatly improves user's satisfaction. It's important to predict the Click Through Rate of ads accurately.

Unsuccessful online advertising leads to a variety of problems. First, it has a bad influence on user experience, especially when the user is searching on the website. That's because the users of search engine always have clear searching purpose and needs. Second, bad recommendation of advertising will reduce the revenue of both the advertisers and the search engine company. That is why advertisers need to know if an ad will be clicked by the users or not.

In order to advertise successfully predicting whether a particular consumer click on an advertisement displayed plays a crucial role.

PRIOR WORK

Title: **Simple and scalable response prediction for display advertising**
Authors: OLIVIER CHAPELLE,EREN MANAVOGLU, ROMER ROSALES
https://people.csail.mit.edu/romer/papers/TISTRespPredAds.pdf

Title: **Contextual advertising by combining relevance with click feedback**
Authors: CHAKRABARTI, D., AGARWAL, D., AND JOSIFOVSKI

This paper gives an overview of certain machine learning algorithms that can be used to predict how likely it is that an advertisement that is displayed will be chosen by the user
This paper focuses on designing a supervised classification algorithm based on logistic regression and compares the analysis with decision tree algorithm.

It also gave me an idea to experiment on two important supervised classification algorithms and observe that Decision Tree algorithm gives better classification accuracy than logistic regression for the data set I have considered.

DATA

https://www.kaggle.com/fayomi/advertising/version/1#

The data consists of 10 variables:

- Daily Time Spent on Site
- Age
- Area Income
- Daily Internet Usage
- Ad Topic Line
- City
- Male
- Country
- Timestamp
- Clicked on Ad

METHODOLOGY

DATA PREPROCESSING

Feature Selection is a very important step in data pre-processing that determines the accuracy of a classifier. It helps remove unnecessary correlation in the data that might decrease accuracy.

Initially I tried to implement Principal Component Analysis to reduce the number of attributes in train data. PCA is often useful to measure data in terms of its Principal Components rather than on a normal x-y axis. Principal Components are the underlying structure in the data. They are the directions where there is most variance, the directions where the data is most spread out. PCA works best for discrete features. But upon observation, we found that all attributes were categorical attributes. So, I was unable to use PCA for data reduction.

As an alternative to PCA for dimensionality reduction, I decided to observe the data fields manually and remove fields that are not very helpful for classification task.

Step 1:

- I notice that each feature contains 1000 data and they don't have null values, If my data consisted null values then further check would be required how it had to replace the null value.

Step 2:

- Getting the description of the features from data frame to understand how the data is spread
- I also notice that there are 4 categorical features and 6 quantitative features

Step 3:

- Summarizing the Numerical Values
  OBSERVATIONS:

1) Daily time Spent on site is between 32 min to 91 min
2) Age --> mean is 36 and ranges between 19 to 61 --> Hence site is targeting adults
3) Male --> 48% are male who visit the site and rest 52% are female --> Since we see almost equal value, I can say that the site is targeting both.
4) Area Income --> Since the mean is 55000 , the site is targeting rich customers

Step 4:

- Summarizing the Categorical Values
  OBSERVATIONS:
  1) Ad Topic Line --> All the data values are unique
  2) City ---> Consists of 969 unique values out of 1000
  3) Country ---> Consists of 237 unique elements and the country **France** is repeated 9 times, hence needs to be further investigated
  4) Timestamp ---> Can be used to create new features

Step 5:

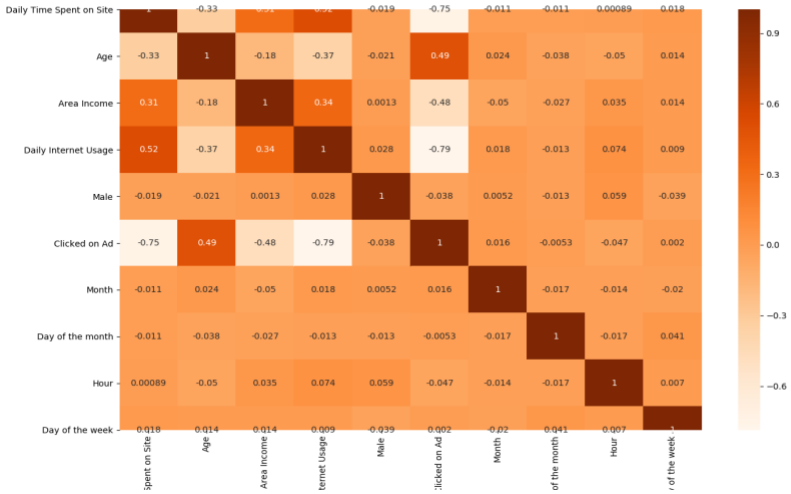
- Analyzing Country Variable
  OBSERVATION:
  1) Already 237 unique elements with the France having maximum frequency
  2) So these unique elements won't help to establish valuable relationships, Hence I can ignore this column too
  3) So out of 4 categorical values I will only use Timestamp feature to further create new features and ignore the other 3

Step 6:

- Feature Creation
  1) Timestamp field was expanded to four new categories: 'Month', 'Day of the month', 'Hour', 'Day of the week',
  2) In this way, we will get new variables that an ML model will be able to process and find possible dependencies and correlation
  3) Now I have created new features I have dropped "Timestamp" feature

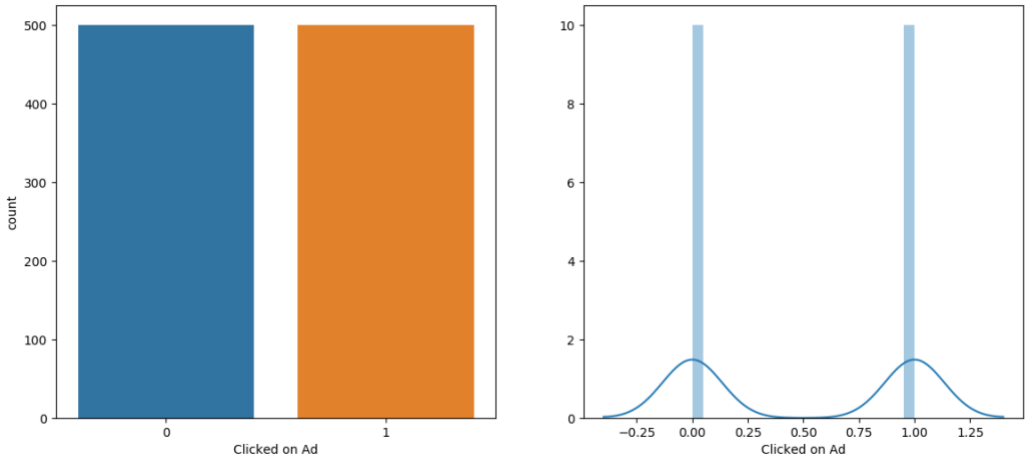
Step 7:

- Finding correlation between values
  OBSERVATIONS:
  1) Daily time spent and Daily Internet usage has highest correlation and hence these features can be merged



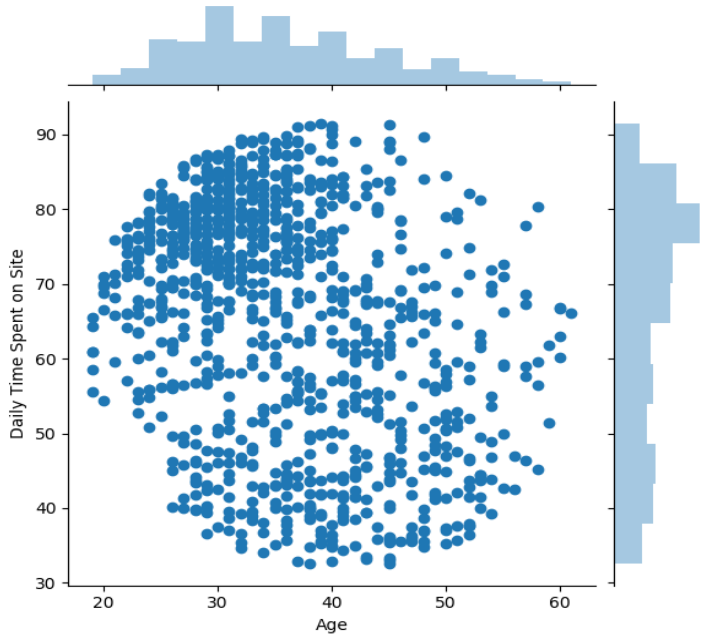
Step 8:

- Visualizing the target variable Clicked on Ad
  OBSERVATION:
  1) Interesting thing here is number of ppl clicked on ad and who did not click, are equal



Step 9:

- Age --> mean is 36 and ranges between 19 to 61 --> Hence site is targeting adults
  Plotting against Daily time spent
  OBSERVATIONS
  1) People age between 30 and 40 are the ones who spend lot of time on the site daily



From the above analysis, 'Daily Time Spent on Site', 'Age', 'Area Income', 'Male', 'Month', 'Day of the month', 'Day of the week', 'Daily Internet Usage' features can be used to train our model.

TRAIN AND TEST DATA SETS

Once the dataset is processed:

1) Divided it into two parts: training and test set.
2) The variables 'Daily Time Spent on Site', 'Age', 'Area Income', 'Male', 'Month', 'Day of the month', 'Day of the week', 'Daily Internet Usage' will be the input values X for the ML models.
3) The variable 'Clicked on Ad' will be stored in a variable and will represent the prediction variable.
4) I have arbitrarily chosen to allocate 30% of the total data for the training set.

MODEL DEVELOPMENT AND THE RESULTS

I am mainly experimenting two ML methods

1) Logistic Regression
2) Decision Trees

Logistic Regression

The Logistic Regression model is an algorithm that uses a logistic function to model binary dependent variables. It is a tool for predictive analysis, and it is used to explain the relationships between multiple variables.

Once the model is fit, Accuracy and Confusion Matrix is obtained

The accuracy of the logistic regression model is 88.66%.

```
Logistic regression accuracy: 0.8866666666666667

Confusion matrix:
[[140   6]
 [ 28 126]]
```

As can be observed, the performance of the model is also determined by the confusion matrix. The condition for using this matrix is to be exploited on a data set with known true and false values. Confusion matrix tells us that the total number of accurate predictions is 140 + 126 = 266. On the other hand, the number of incorrect predictions is 28 + 6 = 34.

Decision Trees

The Decision Tree is one of the most commonly used data mining techniques for analysis and modeling. It is used for classification, prediction, estimation, clustering, data description, and visualization. The advantages of Decision Trees compared to other data mining techniques are simplicity and computation efficiency.

Once the model is fit, Accuracy and Confusion Matrix is obtained
The accuracy of the logistic regression model is 92.66%.

```
Decision Tree accuracy: 0.9266666666666666

Confusion matrix:
[[136  10]
 [ 12 142]]
```

As can be observed, the performance of the model is also determined by the confusion matrix. The condition for using this matrix is to be exploited on a data set with known true and false values. Confusion matrix tells us that the total number of accurate predictions is 136 + 142 = 278. On the other hand, the number of incorrect predictions is 12 + 10 = 22.

CONCLUSION

Both models have shown that they can be very successful in solving classification problems.
But the Decision Tree model showed slightly better performance than the Logistic Regression model, but definitely, both models have shown that they can be very successful in solving classification problems.