CLICK THROUGH RATE PREDICTION OF ADVERTISEMETS

INTRODUCTION

Click-through rate (CTR) is the ratio of the number of clicks on a specific link or call to action (also known as CTA, for example the ‘Learn More’ text at the bottom of an email marketing campaign) to the number of times people were exposed to the link .

CTR can be used to measure the success of pay-per-click (PPC) search results (for example with Google AdWords or other search engines), CTAs on a landing page, or hyperlinks in blog posts and email campaigns.

In cost-per-click (CPC) advertising system, advertisements are ranked by the eCPM (effective cost per mille), which is the product of the bid price and CTR (click-through rate), and CTR needs to be predicted by the system. Hence, the performance of CTR prediction model has a direct impact on the final revenue and plays a key role in the advertising system.

IMPORTANCE OF CTR

CTR is an important metric because it helps you understand your customers—it tells you what works (and what doesn’t work) when trying to reach your target audience. A low CTR could indicate that you’re targeting the wrong audience or that you’re not speaking their language persuasively enough to convince them to click.An online advertisement’s CTR lets you know how effective the ad is at drawing in potential customers

DATA SOURCE :

The dataset used in the project is obtained from Kaggle .
It contains 1000 rows and following 10 columns

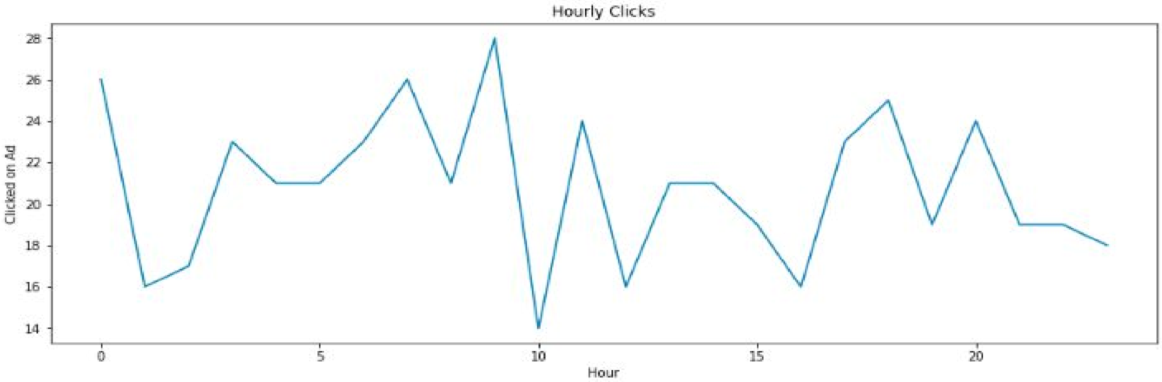
1. Daily Time Spent on a Site : Time spent by the user on a site in minutes.
2. Age : Customer's age in terms of years.
3. Area Income : Average income of geographical area of consumer.
4. Daily Internet Usage : Avgerage minutes in a day consumer is on the internet.
5. Ad Topic Line : Headline of the advertisement.
6. City : City of the consumer.
7. Male : Whether or not a consumer was male.
8. Country : Country of the consumer.
9. Timestamp : Time at which user clicked on an Ad or the closed window.
10. Clicked on Ad : 0 or 1 is indicated clicking on an Ad.

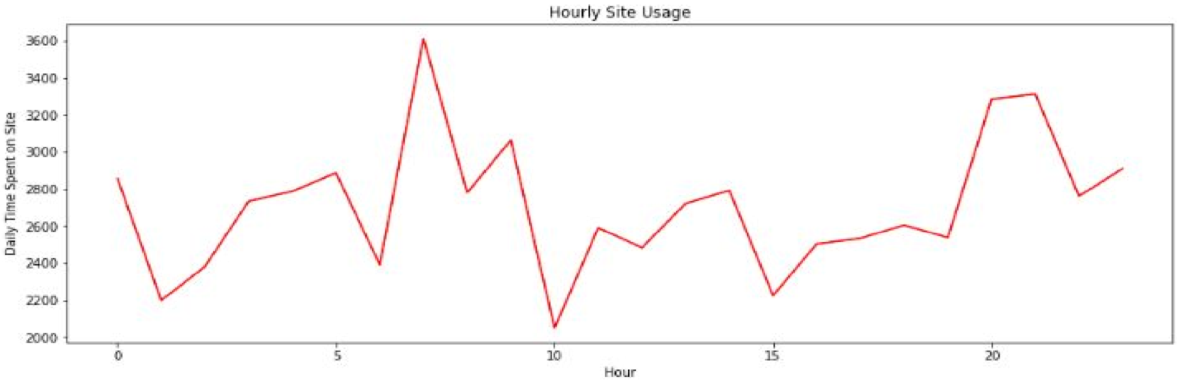
The target variable considered here is the ‘ Clicked on Ad ‘ which has two classes : 0 for not clicked and 1 for clicked .

The dataset is perfectly balanced with 500 records of each class .It is clean and has no null values or duplicates

EXPLORATORY DATA ANALYSIS (EDA) :

- ➔ The daily amount of internet used by a customer does seem to have a proportional increase in the number of clicks which is a good sign . The maximum number of clicks are in the morning at 9 am , followed by 7 am and 12 am , then in the evening at 6pm .

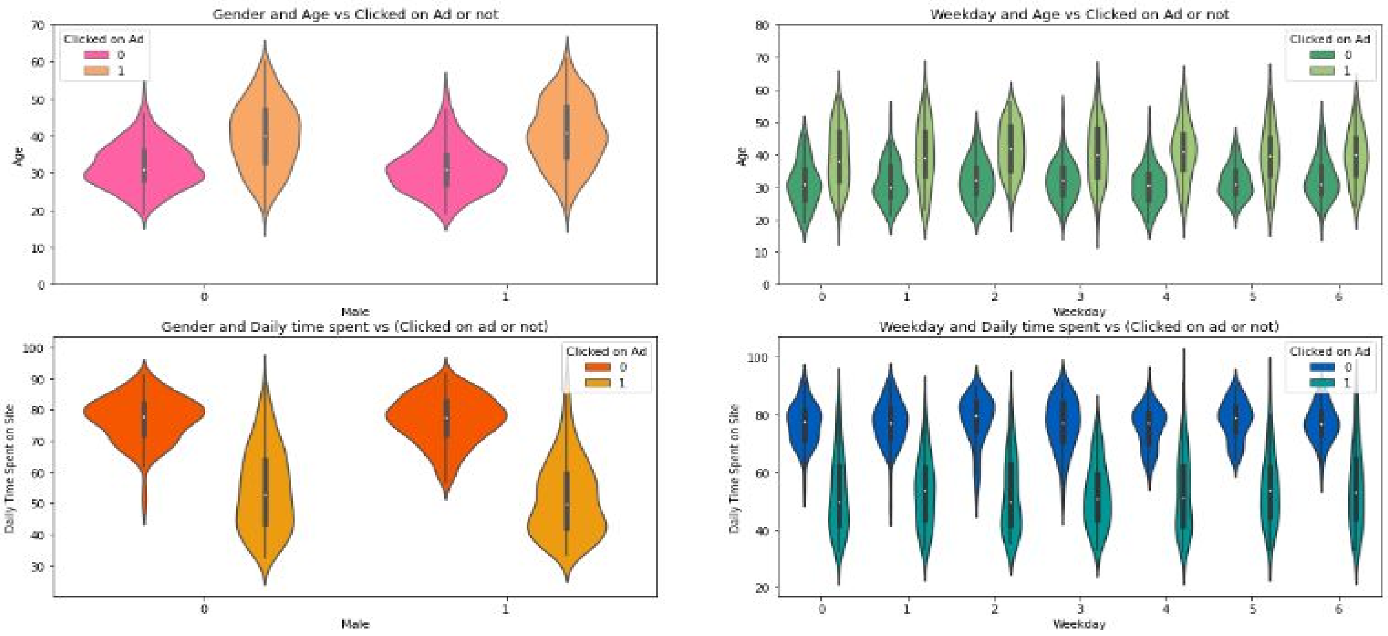




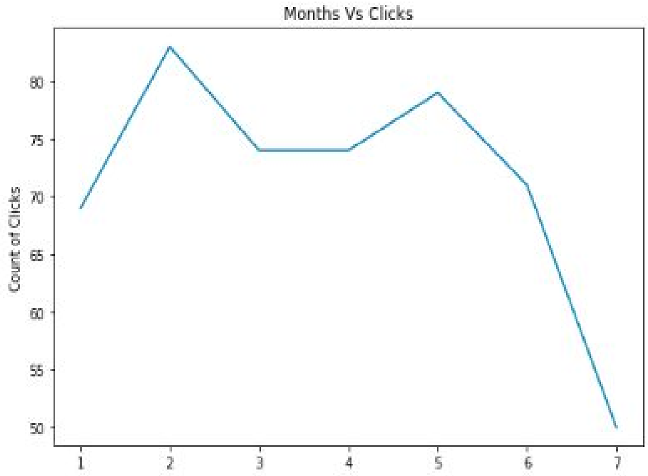
- ➔ The gender and age of the user does have an impact on the ad clicks whereas the day of the week doesn’t really have a distinct impact

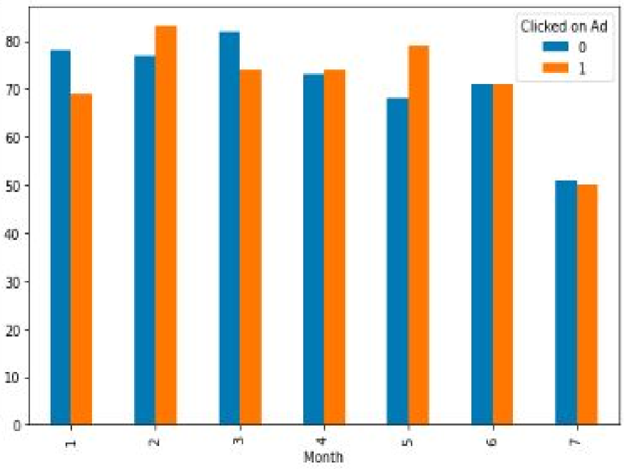
Male and Female of age group 32-48 years tend to click the ad with the median age being 40 years

Male and Female of who spend 70-95 mins do not click on the add Females who spend 42-65 mins are more likely to click through , basically around an hour Males who spend 42-60 mins are more likely to click through



➔ The dataset had a 7 month timeframe and most of the clicks have happened from the month of February to May





MODELS AND METRICS

The dataset has been trained with two models :

1) Logistic Regression
2) Random Forest

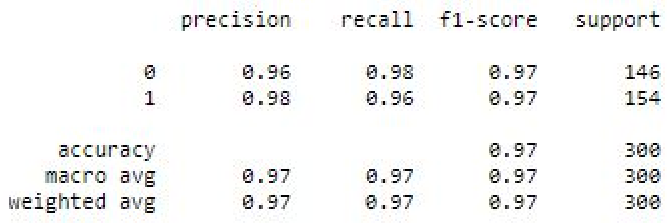
LOGISTIC REGRESSION

The data has been standardised using StandardScaler() and fed into the LogisticRegression model with the help of pipeline. The model has given an accuracy score of 0.97

The recall of the clicked class is 0.96 which means that the model has been able to reciprocate 96 % of the actual ad clicks .

The precision of the clicked class is 0.98 which means 98 % of the models prediction of the clicks is correct

Below is the detailed classification report



```
              precision    recall  f1-score   support

           0       0.96      0.98      0.97       146
           1       0.98      0.96      0.97       154

    accuracy                           0.97       300
   macro avg       0.97      0.97      0.97       300
weighted avg       0.97      0.97      0.97       300
```

As the dataset does not consist of any imbalance , the macro avg and weighted avg do not make any sense

RANDOM FOREST CLASSIFIER

The dataset is trained with two criterions : Gini and Entropy and max depth till 9 and subjected to GridSearch Cross Validation process to obtain the best parameters

The gini impurity validation for the trees have resulted in the model performance of 96 % at max depth 3 and 200 estimators on the training data .
On running the model with the test data , a R squared score of 94 % is achieved

We can see the Logistic regression has performed better with a good accuracy

FUTURE SCOPE :

The information about the medium (for eg ., Facebook , Email campaign , Website , Chat assistant , etc) in which the Ads have been put up would provide an additional scope to research the best medium for the organisation to invest .

SUMMARY :

The CTR prediction can be implemented in different uses cases for different industry domains . Here a company's Ad on its website has been studied and recommendations have been provided to the company in order to improve its click through score .