High Level Design (HLD)

User Response Prediction System using Machine Learning Techniques

Revision Number: 1.0

Last date of revision: 11/07/2021

Document Version Control

Date Issued	Version	Description	Author
11/07/2021	1	Initial HLD	Vivek

Contents

Abstract

It is necessary to predict profitable users who can click target ads (i.e., activity Targeting) in the advertising trade. The task selects the potential users that can connect the ads by analyzing users' clicking/web browsing data and displaying the foremost relevant ads to them. This project presents an associate empirical study of the exploitation of different web of things techniques to predict whether or not an advertisement is going to be clicked or not. We tend to perform click prediction on a binary scale, one for click and zero for no click. We tend to use clicks information from advertizing.csv provided as a region of Kaggle competition as our information set. We tend to perform feature choice to get rid of options that don't facilitate improved classifier accuracy. We tend to examine information manually and conjointly use feature choice capability.

1. Introduction

Internet showcasing has taken over traditional advertising methodologies in the ongoing past. Organizations like to advertise their items on websites and web-based life stages. Be that as it may, focusing on the correct crowd is a test in online advertising. Burning through millions to show the advertisement to the group of spectators that isn't probably going to purchase your items can be expensive. In This project, we will work with the advertising information of a showcasing agency to build up an AI calculation that predicts if a specific client will tap on an advertisement. The information consists of **10** factors: **'Daily Time Spent on Site'**, **'Age', 'Area Income', 'Daily Internet Usage', 'Ad Topic Line', 'City', 'Male', 'Country', Timestamp'** and **'Clicked on Ad'**. The fundamental variable we are keen on is 'Clicked on Ad.' This variable can have two possible results: 0 and 1, where 0 alludes to the situation where a client didn't tap the advertisement, while one alludes to the situation where a client taps the advertisement. We will check whether we can utilize the other nine factors to foresee the worth 'Clicked on Ad' factor precisely. Likewise, we will play out some exploratory information investigation to perceive how 'Daily Time Spent on Site' in combination with 'Ad Topic Line' influences the client's decision to tap on the ad.

1.1 Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding, and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

- Present all of the design aspects and define them in detail
- Describe the user interface being implemented
- Describe the hardware and software interfaces
- Describe the performance requirements
- Include design features and the architecture of the project
- List and describe the non-functional attributes like:
    - Security
    - Reliability
    - Maintainability
    - Portability
    - Reusability
    - Application compatibility
    - Resource utilization
    - Serviceability

1.2 Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

1.3 Definitions

Term	Description
Database	Collection of all the information monitored by this system
IDE	Integrated Development Environment
URP	User Response Prediction

2. General Description

2.1 Product Perspective

- The online advertisement industry has become a multi-billion industry, and predicting ad **CTR** (click-through rate) is now central. Nowadays, different types of advertisers and search engines rely on modelling to predict ad CTR accurately.
- We will be predicting the ad click-through rate using the machine learning approach. Before that, let us first understand few essential concepts and a general practice followed by search engines to decide which ads to display.
- **CTR**: It is the metric used to measure the percentage of impressions that resulted in a click.

$$\mathrm{CTR} = \frac{\text{Number of click-throughs}}{\text{Number of impressions}} \times 100(\%)$$

- Search ads: Advertisements that get displayed when a user searches for a particular keyword.
- Paid search advertising is a popular form of **Pay per click (PPC)** advertising in which brands or advertisers pay (bid amount) to have their ads displayed when users search for specific keywords.
- Relevance of Predicting CTR through a real-life example:

Typically, the primary source of income for search engines like Google is through advertisement. Many companies pay these search engines to display their ads when a user searches for a particular keyword. Our focus is on search ads and CTR, i.e. the amount is paid only when a user clicks on the link and redirects to the brand’s website.

2.2 Problem statement

- In this project, 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.
- 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 primary variable we are interested in is ' Clicked on Ad'.

This variable can have two possible outcomes: 0 and 1, where 0 refers to a user who didn't click the advertisement, while one refers to the scenario where a user clicks the ad.

- We will see if we can use the other **9** 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.

2.3 PROPOSED SOLUTION

- Different advertisers approach these search engines with their ads and the bidding amount to display their ads. The main objective of these search engines is to maximize their revenue. So the question is, how does a search engine decide which ads to display when a user searches for a particular keyword?
- Till now, we have seen what ad click prediction is and why is it important. Let us now explore how to calculate ad click prediction by performing machine learning modelling on a dataset. We will build a Logistic Regression model that would help us predict whether a user will click on an ad or not based on the features of that user. And hence calculate the probability of a user clicking on an ad.
- Using these probabilities, search engines could decide which ads to display by multiplying the possibilities with the bid amount and sorting it out.

2.4 Technical Requirements

This document addresses the requirements for detecting the user response prediction possibility of a customer based on his clicked history.

2.5 Data Requirements

- This data set contains the following 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 the 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

2.6 Tools used

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, Flask used to build the whole model.

pandas scikit learn

PYTHON MADE WITH JUPYTER PLOTLY PANDAS NUMPY HTML5 CSS3
JAVASCRIPT FLASK HEROKU

- Jupyter notebook is used as IDE.
- For visualization of the plots, Matplotlib, Seaborn and Plotly are used.
- Front end development is done using HTML/CSS.
- Python is used for backend development.

2.7 Constraints

The URP application must be user friendly, as automated as possible and users should not be required to know any of the workings.

2.8 Assumptions

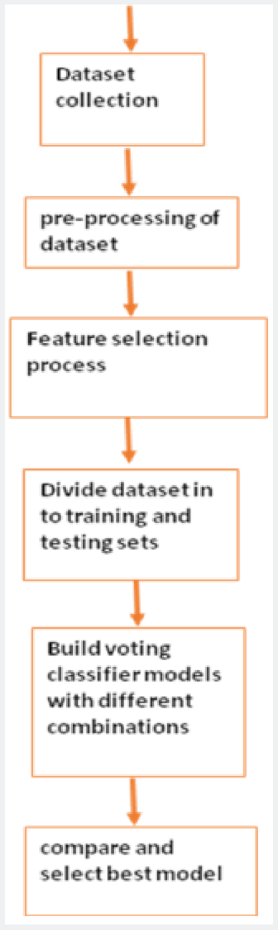
The proposed ad-click prediction model is based on human features. To adapt to this, certain human related features like Frequent Time Spent on Website, Lifetime, field Revenue, Frequent Internet Usage, and Gender are alone considered in this model. These attributes are extricated from the dataset to efficiently develop the prototype. Some features such as Advertisement Topic Line, City, Country, Time-stamp are not human features, so they are ignored from consideration. All extracted attributes have been indoctrinated into a convenient form to make study easy.

3. Design Details

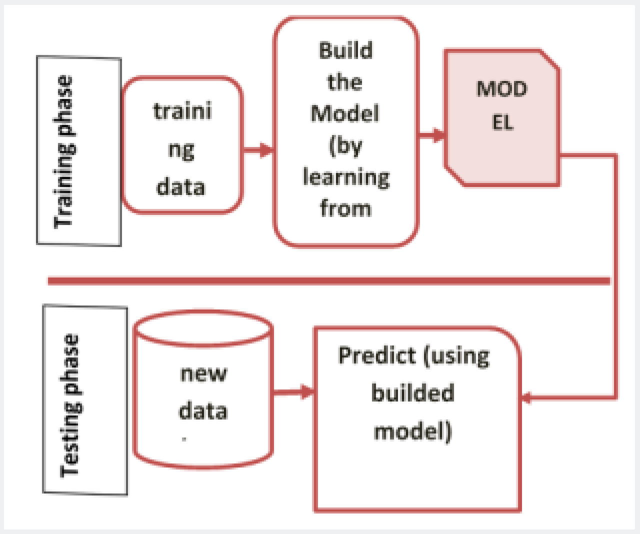
3.1 Process Flow

For predicting the possibility to click , we will use a machine learning base model. Below is the process flow diagram is as shown below.

Proposed methodology


Dataset collection
pre-processing of dataset
Feature selection process
Divide dataset in to training and testing sets
Build voting classifier models with different combinations
compare and select best model

3.1.1 Model Training and Evaluation


Training phase
traini ng data
Build the Model (by learning from
MOD EL
Testing phase
new data
Predict (using builded model)

3.2 Error Handling

Should errors be encountered, an explanation will be displayed as to what went wrong? An error will be defined as anything that falls outside the normal and intended usage.

4. Performance

The aim of this work is to predetermine the Click Through Rate (CTR) of a particular user for a particular advertisement. The CTR prediction is used to predict whether the web-site viewer will be interested in a particular advertisement (ad) or not. When an observer visits a publisher's web-site, in a period of few milliseconds the ad is being furnished established on the maximal CTR. The human attributes that have been picked through feature selection phase are now passed to learning algorithm to predict CTR. The analysis has been carried out using different learning algorithms like i) Logistic Regression, ii) Support Vector Machine, iii) RandomForestClassifier iv) XGBClassifier v) KNN. Among these SVM is the supervised learning model, was implemented and the results were tested.

4.1 Reusability

The code written and the components used should have the ability to be reused with no problems.

4.2 Application Compatibility

The different components for this project will be using Python as an interface between them. Each component will have its own task to perform, and it is the job of the Python to ensure proper transfer of information.

4.3 Resource Utilization

When any task is performed, it will likely use all the processing power available until that function is finished.

4.4 Deployment

Conclusion

While the random forest could have been tuned further, it had good precision. It did not take too much time to fit the model, which would allow for fast tuning of parameters. The linear kernel SVM took a very long time to provide the data. It is shorter at predicting than the random forest and k nearest neighbors classifiers. This time taken to fit the data is mitigated, as only one parameter must be tuned. The k Nearest Neighbors performed the worst in AUC and prediction time. This was not a good model for this data. In the end, the linear SVC should be used as it had a slightly higher AUC and faster prediction time when compared to the random forest.

The end accuracy of this project is 96%. This is not anywhere near as good as the random forest or support vector classifier from before!

Daily Internet Usage is an essential feature.

The lower Daily Internet Usage and Daily Time Spent on Site, the more likely to click the ad. Sex and Age are the least relevant feature. Area Income affects a little. Thus, targeting ads to the people who use the internet little and rarely spend time on a website is efficient to make more likely to click the ad.