High-Level Design (HLD)

User Response Prediction System using Machine Learning Techniques

Document Version Control

Date Issued	Version	Description	Author
03/08/2023	1	Initial HLD	Nabanita

Contents

Abstract:

This project aims to predict profitable users who are likely to click on target ads through activity targeting in the advertising industry. By analyzing users' clicking and web browsing data, relevant ads are displayed to potential users. We approach click prediction as a binary classification task, where 1 indicates a click and 0 indicates no click. We use the advertising data from 'advertizing.csv' obtained from a Kaggle competition as our dataset. The project involves feature selection to enhance classifier accuracy by eliminating irrelevant features. Both manual examination and feature selection techniques are employed for data analysis.

Introduction:

In the digital era, internet marketing has surpassed traditional advertising methods. Businesses prefer to promote their products on websites and social media platforms. However, targeting the right audience remains a challenge in online advertising. Displaying ads to an audience unlikely to purchase products can lead to substantial costs. This project involves working with advertising data from a marketing agency to develop an AI algorithm capable of predicting if a particular user will click on an advertisement. The dataset comprises 10 variables, including 'Daily Time Spent on Site', 'Age', 'Area Income', 'Daily Internet Usage', 'Ad Topic Line', 'City', 'Male', 'Country', 'Timestamp', and 'Clicked on Ad.' The primary focus is on the 'Clicked on Ad' variable, which represents whether a user clicked on the ad (1) or not (0). We will explore the predictive potential of the other nine variables to accurately forecast the 'Clicked on Ad' outcome. Additionally, we will conduct exploratory data analysis to understand how the combination of 'Daily Time Spent on Site' and 'Ad Topic Line' influences the user's decision to click on the ad.

Why this High-Level Design Document?

The High-Level Design (HLD) Document provides detailed information to support the project description and serves as a reference for coding the system. It helps detect contradictions before coding and describes how modules interact at a high level. The HLD includes design aspects, user interface, hardware and software interfaces, performance requirements, architecture, and non-functional attributes like security, reliability, maintainability, portability, reusability, application compatibility, resource utilization, and serviceability.

1.1 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 ofthe system.

1.2 Definitions

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

General Description:

2.1 Product Perspective:

The online advertising industry has become a multi-billion-dollar sector, and accurate prediction of ad CTR (click-through rate) is essential for advertisers and search engines.

In this project, we will predict the ad click-through rate using a machine learning approach. Understanding key concepts, such as CTR and search ads, is crucial for this task.

CTR measures the percentage of impressions resulting in ad clicks.

Search ads are displayed when users search for specific keywords.

Paid search advertising involves advertisers bidding to have their ads displayed when users search for particular keywords.

The primary focus is on CTR for search ads, where payment occurs only when a user clicks on the ad and redirects to the brand's website.

2.2 Problem Statement:

The project involves developing a machine learning algorithm to predict if a user will click on an advertisement using advertising data from a marketing agency.

The dataset consists of 10 variables, with the primary variable of interest being 'Clicked on Ad.'

'Clicked on Ad' can have two outcomes: 0 (user did not click the ad) and 1 (user clicked the ad).

The objective is to use the other 9 variables to accurately predict the 'Clicked on Ad' outcome.

Exploratory data analysis will also be performed to understand how 'Daily Time Spent on Site' in combination with 'Ad Topic Line' influences ad clicks.

2.3 Proposed Solution:

The search engines aim to maximize revenue by displaying relevant ads to users when they search for specific keywords.

To achieve this, we will use machine learning modeling, particularly building a Logistic Regression model, to predict the likelihood of a user clicking on an ad based on their features.

By calculating these probabilities, search engines can determine which ads to display by considering the bid amount and sorting them accordingly.

2.4 Technical Requirements:

This document outlines the requirements for detecting the user response prediction possibility based on the customer's click history.

2.5 Data Requirements:

The dataset contains the following features: 'Daily Time Spent on Site,' 'Age,' 'Area Income,' 'Daily Internet Usage,' 'Ad Topic Line,' 'City,' 'Male,' 'Country,' 'Timestamp,' and 'Clicked on Ad.'

Each feature provides relevant information, such as consumer time on-site, age, area income, daily internet usage, ad headline, city, male gender, country, timestamp of ad click or window closure, and whether the consumer clicked on the ad (1) or not (0).

Tools used

- Jupyter notebook is used as IDE.
- Python programming language and frameworks such as NumPy, Pandas,Scikit-learn, Flask used to build the whole model.
- For visualization of the plots, Matplotlib, Seaborn and Plotly are used.

1.3 Constraints

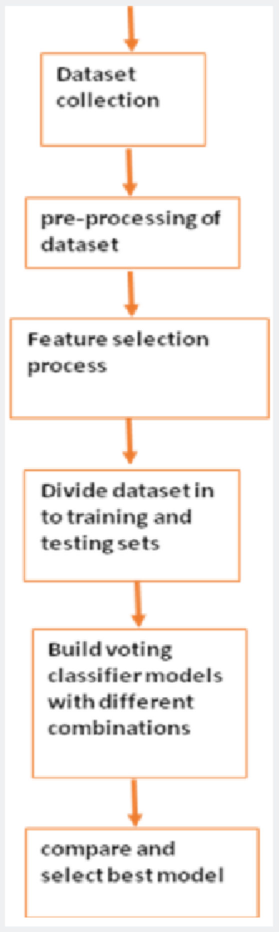
The URP application must be user friendly, as automated as possible and usersshould not be required to know any of the workings.

1.4 Assumptions

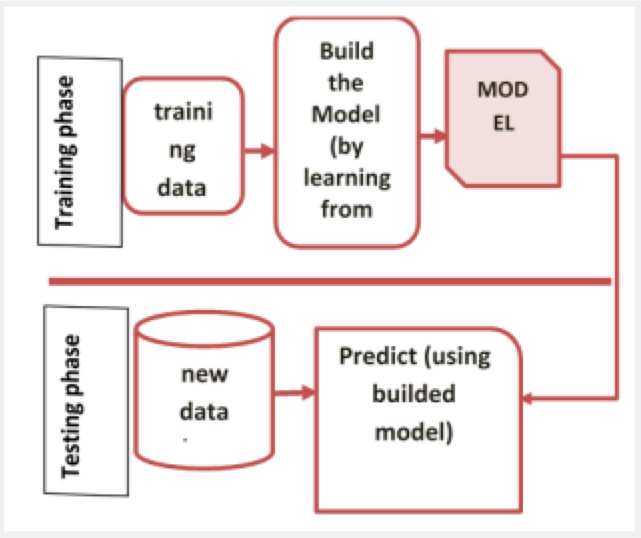
The proposed ad-click prediction model is based on human features. To adapt tothis, certain human related features like Frequent Time Spent on Website, Lifetime,field Revenue, Frequent Internet Usage, and Gender are alone considered in thismodel. These attributes are extricated from the dataset to efficiently develop theprototype. 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 studyeasy.

DESIGN FLOW.

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)

1.5 Error Handling:

In case of encountering errors during the project execution, informative explanations will be provided to clarify the issue.

Errors are defined as any deviations from normal and intended usage.

Performance:

The objective of this project is to predict the Click Through Rate (CTR) of a user for a specific advertisement. CTR prediction helps determine whether a web-site viewer is likely to be interested in a particular ad. When a user visits a publisher's website, the ad is quickly displayed based on the maximal CTR. The human attributes selected during the feature selection phase are used as input to the learning algorithm for CTR prediction. Different learning algorithms, including Logistic Regression, Support Vector Machine (SVM), RandomForestClassifier, XGBClassifier, and KNN, were analyzed. SVM, a supervised learning model, was implemented, and the results were evaluated.

2.1 Reusability:

The code and components used in this project are designed to be reusable without any issues.

2.2 Application Compatibility:

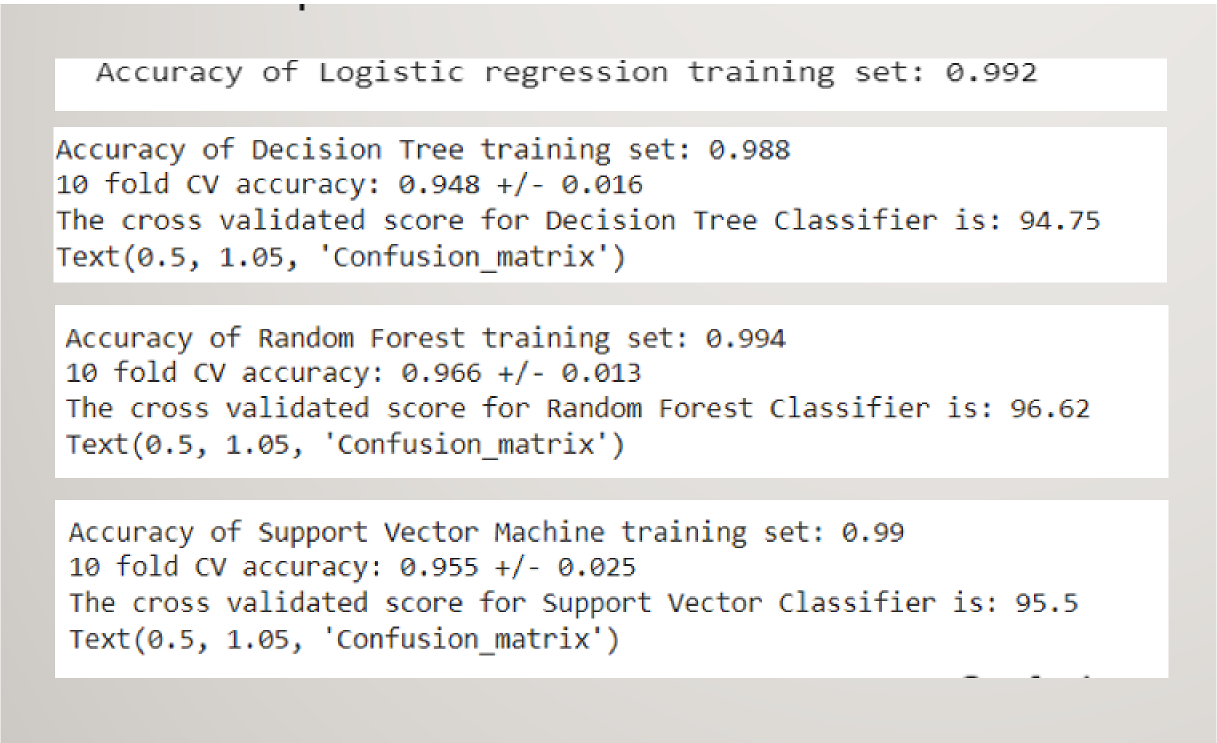
The different components in this project communicate using Python as an interface.

Each component performs specific tasks, and Python ensures the proper transfer of information between them.

2.3 Resource Utilization:

During task execution, all available processing power will likely be utilized until the function is completed.

Conclusion:



```
  Accuracy of Logistic regression training set: 0.992

Accuracy of Decision Tree training set: 0.988
10 fold CV accuracy: 0.948 +/- 0.016
The cross validated score for Decision Tree Classifier is: 94.75
Text(0.5, 1.05, 'Confusion_matrix')

Accuracy of Random Forest training set: 0.994
10 fold CV accuracy: 0.966 +/- 0.013
The cross validated score for Random Forest Classifier is: 96.62
Text(0.5, 1.05, 'Confusion_matrix')

Accuracy of Support Vector Machine training set: 0.99
10 fold CV accuracy: 0.955 +/- 0.025
The cross validated score for Support Vector Classifier is: 95.5
Text(0.5, 1.05, 'Confusion_matrix')
```

While the random forest model could have been further tuned, it demonstrated good precision and had a relatively fast model fitting time, facilitating quick parameter tuning.

The linear kernel SVM, on the other hand, took a considerable amount of time to provide results. However, its prediction time was shorter than both the random forest and k-nearest neighbors classifiers.

The k-nearest neighbors model performed poorly in terms of AUC and prediction time, making it less suitable for this dataset.

Ultimately, the linear SVC model was chosen due to its slightly higher AUC and faster prediction times compared to the random forest.

The project achieved an accuracy of 96%, which is not as high as the random forest or support vector classifier from earlier analyses.

Daily Internet Usage is identified as an essential feature, and users with lower Daily Internet Usage and Daily Time Spent on Site are more likely to click on the ad.

Gender and Age are the least relevant features, while Area Income has a relatively minor impact.

Targeting ads to users who use the internet infrequently and spend less time on websites is more likely to result in ad clicks.