Report – Project of Fouille de Données

Supervised by Barbara Martin

date : 27 / 01 / 2024

Outline

1 Introduction

This report presents the outcome of the project, which involves data mining applied to a fictional advertising dataset. The objective of this project was to explore a specific data mining methodology to create a predictive model regarding the propensity of a user to click on an online advertisement.

The provided dataset includes various information such as daily time spent on the site, consumer age, average income of the geographical area, daily internet usage, consumer city and country, ad topic line, gender, and timestamp of the interaction with the ad. The target variable, 'Clicked on Ad', indicates whether the consumer actually clicked on the ad (represented by 1) or not (represented by 0).

The project progressed through several key stages, starting from the initial data exploration to the application of preprocessing specific to the chosen method, followed by encoding, model training, and in-depth analysis of results.

We will delve into detail regarding these different stages in the following sections.

2 Data Exploration

The DataFrame is stored in a CSV file named 'Advertising.csv' and contains a total of 1000 rows and 1546 columns. The columns include features such as Daily Time Spent on Site, Age, Area Income, Daily Internet Usage, Ad Topic Line, City, Male, Country, Timestamp and Clicked on Ad. The Clicked on Ad column appears to be the target variable, with binary values 0 and 1.

The data overview likes:

Daily Time Spent on Site: average of 65.00020 minutes per day, range from 32.60 to 91.43 minutes.
Age: average age of approximately 36 years, range from 19 to 61 years.
Area Income: average of 55000.00008, range from 13996.50 to 79484.80.
Daily Internet Usage: average of 180.00010, range from 104.78 to 269.96.
Male: approximately 48.1% of users are male.
Clicked on Ad: half of the users clicked on the ad.

These data provide opportunities to explore the methods that could be used to predict 'Clicked on Ad' based on the provided information.

3 Preprocessing for the chosen method

Data preprocessing was a crucial step in the process of building our model. It aimed to prepare the raw data to be suitable for further processing in our code. In this report, we will describe the different preprocessing steps performed on the dataset using the provided code.

The first step is to identify and handle missing values in the dataset. Running the code df.isna().sum() shows that there are no missing values in the dataset, indicating satisfactory data quality.

Two categorical columns, Country and City, are encoded using the label encoder technique. This allows representing these categories as numerical values, making it easier to use this information in our model.

The Timestamp column is converted into a datetime object, then subdivided into multiple columns including Year, Month, Day, Hour, Minute, and Second.

Some columns are not needed for the Clicked on Ad classification task. Thus, the Ad Topic Line, Timestamp, Country, and City columns are dropped from the dataset.

We opted for the Random Forest classification algorithm by instantiating a classifier with 100 decision trees. This choice stems from the method's robustness to non-linear relationships, its ability to reduce overfitting, automatic handling of important variables, performance with large datasets, and ease of use with few hyperparameters. These characteristics make it a versatile and effective choice for the specific problem at hand.

Finally, the data is split into training and testing sets using the train_test_split function from scikit-learn.

These preprocessing steps have provided us with a solid foundation for the subsequent classification model, ensuring that the data is properly formatted and ready to be used in the learning process.

4 Modele training

After training the RandomForestClassifier model, we proceeded with its evaluation to measure its performance. The model achieved an accuracy of 95%, demonstrating its effectiveness in predicting ad clicks. This result highlights the model's ability to correctly classify both positive and negative instances.

From the classification report obtained, the precision and recall are 0.95 which depicts the predicted values are 95% accurate. Hence the probability that the user can click on the advertisement is 0.95 which is a great precision value to get a good model.

5 In-depth analysis of results

5.1 Features influence

The visualization of feature influenced in prediction, generated from the RandomForestClassifier model, offers significant insights. The most influential features in this prediction include Daily Internet Usage (32.7%), Daily Time Spent on Site (34.2%), Area Income (13.13%) and Age (8.24%).

This analysis suggests that the daily time spent on the site and daily internet usage are the main contributors to predicting ad clicks. The influence of age and area income is also notable, indicating that these factors play a significant role in click behavior.

5.2 Prediction on new data

The model evaluation on new data shows consistent results with these observations. For data where the age is 30-42 years and the area income is high with high daily time spent on site, the model predicts a non-click on the ad (0). This can be attributed to a combination of the individual's youth and high income, indicating a lower propensity to click on the advertissement .

On the other hand, for data where the age is 72 years and the area income is low with low high daily time spent on site, the model predicts a click on the ad (1). This prediction can be interpreted as a potential reaction to an older audience associated with lower income, suggesting an increased likelihood of clicking on the ad.

Overall, the analysis of important features and predictions on new data highlights the model's ability to interpret trends and provide predictions consistent with the identified significant features.

6 Conclusion

This report has detailed the data preprocessing and the utilization of a RandomForestClassifier model for advertising classification.

The choice of the model was justified by its ability to handle complex relationships. Model training resulted in an overall accuracy of 95%, with feature analysis highlighting the importance of daily time spent on the site, daily internet usage, area income and age.

Predictions on new data demonstrated the model's ability to interpret trends. This ultimately provides us with a solid foundation for informed decision-making in the field of online advertising.