User Response Prediction System using Machine Learning Techniques

Vivek Limbad, Manali Kadam, Rueben Patil, Siddhanth Ghag

Pursuing Masters in Data Science & Business Analytics.

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 improve classifier accuracy. We tend to examine information manually and conjointly use feature choice capability.

Key Words: Machine Learning, Logistic Regression.

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.

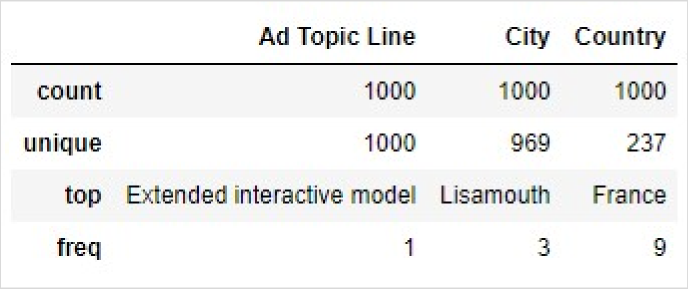
1.1 Proposed System

A) Data Collection

The dataset for this project can be downloaded from this Kaggle link. Unzip the downloaded zip file and place the "advertising.csv" file in your local drive. This is the file that we are going to use to train our machine learning model.

B) Data Pre-processing

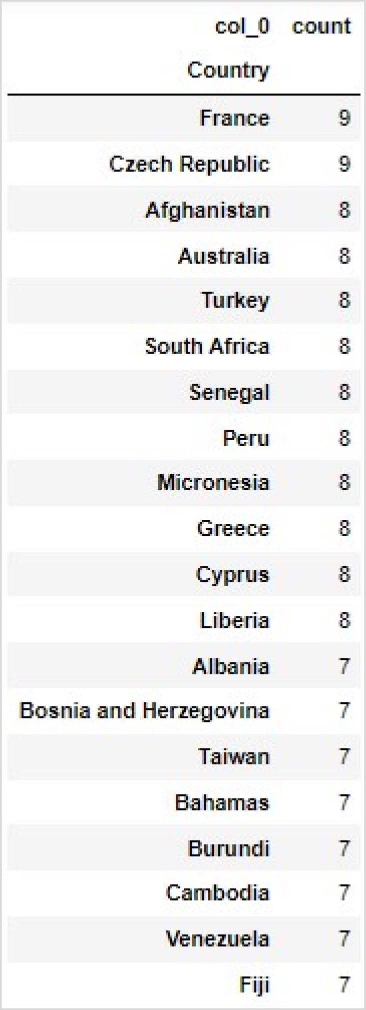
You may have noticed that **"Ad Topic Line," "City,"** and **"Country"** are categorical columns. Let plot all the unique Values for these columns. Values for these columns.



	Ad Topic Line	City	Country
count	1000	1000	1000
unique	1000	969	237
top	Extended interactive model	Lisamouth	France
freq	1	3	9

As we can see from the table above that all the values in column "Ad Topic Line" is unique, while the "City" column contains **969** unique values out of **1000** and there are too many individual elements within these two categorical columns, and it is generally difficult to perform a prediction without the existence of a data pattern. Because of that, they will be omitted from further analysis, and the third categorical variable, i.e., "Country," has a unique element (France) that repeats nine times. Additionally, we can decide on countries with the highest number of visitors.

The table shows the **20** most represented countries in our Data Frame, and we have already seen, there are **237** different unique countries in our dataset, and no single country is too dominant. A large number of individual elements will not allow a machine learning model to exist easily valuable relationships. For that variable will be excluded too



col_0	count
Country	
France	9
Czech Republic	9
Afghanistan	8
Australia	8
Turkey	8
South Africa	8
Senegal	8
Peru	8
Micronesia	8
Greece	8
Cyprus	8
Liberia	8
Albania	7
Bosnia and Herzegovina	7
Taiwan	7
Bahamas	7
Burundi	7
Cambodia	7
Venezuela	7
Fiji	7

C) Feature Extraction and Selection

The data scientist's data has several features that may or may not be relevant to the topic of interest. Also, it may not be in a suitable format. The first and foremost task to the data scientist is to extract the appropriate collection of attributes that preferably suits the learning algorithm. Before processing, it needs to be transformed to prevent relapse problems like overfitting and underfitting as presented. The following Table 1 shows the list of features present in the dataset.

Features	Description
Daily Time Spent on Site	User time spent on the website in minutes.
Age	User age in years
Area Income	Avg. Income of geographical area of user
Daily Internet Usage	Avg. minutes a day consumer is on the user.
Ad Topic Line	The headline of the advertisement
City	City of user
Male	Whether or not the user was male
Country	Country of user
Timestamp	Time at which user clicked on Ad or closed window
Clicked on Ad	0 or 1 indicated clicking on Ad

Table 1: List of features

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 develop the prototype efficiently. Some features such as Advertisement Topic Line, City, Country, Time-stamp are not human, so they are ignored from consideration. The features that are taken into consideration are shown in Table 2. All extracted attributes have been indoctrinated into a convenient form to make study easy.

Features	Description
Daily Time Spent on Site	User time spent on the website in minutes.
Age	User age in years
Area Income	Avg. Income of geographical area of user
Daily Internet Usage	Avg. minutes a day consumer is on the user.
Male	Whether or not the user was male
Clicked on Ad	0 or 1 indicated clicking on Ad

Table 2: Features taken into consideration

Next, we will analyze the **'Timestamp'** category. It represents the exact time when a user clicked on the advertisement. We will expand this category to **4** new types: month, day of the month, day of the week, and hour. In this way, we will get new variables that an ML model will process and find possible dependencies and correlations. Since we have created new variables, we will exclude the original variable "Timestamp" from the table. The "Day of the week" variable contains values from **0 to 6**, where each number represents a specific day of the week (from Monday to Sunday)

C) Train and take a look at knowledge Sets

Once the dataset is processed, we want to divide it into two components that are coaching and take a look at the set. We'll take and use the train_test_split to operate for that and every variable except 'Clicked on Ad' are the input values x for the cubic centimeter models. The variable 'Clicked on Ad' is keep in y, can represent the prediction variable and that we at randomly selected to portion thirty third of the fundamental knowledge for the coaching set.

1.2 Related Work

Much attention has been paid to advertisement research recently. The best way to maximize the commercial value of advertisements is to display the ads to people who are interested in them. However, there are some issues to be dealt with, such as matching relevant advertisements for a query, ranking the candidate advertisements, deciding how to display the advertisements on the search result page, click prediction and analysis for the presenting promotions, and pricing of the advertisements. Several machine learning algorithms such as Logistic Regression, Linear Poisson Regression, Online Bayesian Regression, Support Vector Machines, and Latent Factor Model have been adopted to predict the clicks of advertisements presented for a query. Since online data is usually massive, online data stream analysis can be beneficial in the Behavioural Targeting field. Behavioral Targeting contains three pricing models, which are Pay-Per-Click (PPC), Pay-Per-Impression (PPI), and Pay-Per-Transaction (PPT). The popular one is PPC. For the PPC model, both the advertiser and the search engine companies wish users to click the advertisements. Therefore, Behavioural Targeting is a good way to solve this problem because it reduces advertiser's cost and increase search engine companies profit simultaneously.

Multiple Criteria Linear Programming (MCLP) is promising optimization-based classification model and has extended to the family toolbox.

MCLP has many successful applications, including credit card portfolio management, credit card risk analysis, firm bankruptcy prediction, network intrusion detection, medical diagnosis and prognosis, and classification of HIV-1 mediated neuronal dendritic and synaptic damage. Multi-Criteria Linear Programming Regression (MCLPR) was firstly introduced by Zhang, which converted a classification problem to a regression one. The data can be separated into two groups to move it downward and upward by parameter and then classified by hyperplane to construct a regression model. The excellence of MCLPR is its ability to fix the ill-posed condition with a limited amount of sample, handling non-linear relationships by kernel function, and giving the global solution if it exists. MCLPR has already proved its performance in many real-life datasets.

2.1 Experimental Results

Algorithm	**Accuracy**
LogisticRegression	95.33%
RandomForestClassifier	96%
XGBClassifier	95%
Linear Support Vector Classification	96%
k Nearest Neighbors Classifier	68.85%

Table 3: Accuracy values of ML models

2.2 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 SVC 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.