International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 01 | Jan 2020 www.irjet.net p-ISSN: 2395-0072

Ad-Click Prediction using Prediction Algorithm: Machine Learning Approach

Jitendra Mahindra Bagul[1], Tushar B. Kute [2]

[1]*Student, Dept. of Computer Engineering, KBT College of Engineering, Maharashtra, India*
[2]*Associate Director, MITU Skillologies, Pune*

Abstract - It is necessary to predict probably profitable users who can click target ads (i.e. activity Targeting), in advertising trade. The task selects the potential users that are possible to click the ads by analysing user's clicking/web browsing data and further more as displaying the foremost relevant ads to them. This paper presents associate empirical study of exploitation totally different web of things techniques to predict whether or not an advertisement are 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, SVM, Logistic Regression (LR), datasets.**

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 as yet 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 article, 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 potential results: 0 and 1 where 0 alludes to the situation where a client didn't tap the advertisement, while 1 alludes to the situation where a client taps the advertisement. We will check whether we can utilize the other 9 factors to precisely foresee the worth 'Clicked on Ad' factor. We will likewise 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 add.

1.1 Proposed System

A) Data Collection

The dataset for this article 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	Reactive bi-directional workforce	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 unique 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 9 times. Additionally, we can decide countries with the highest number of visitors.

The table below 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 unique elements will not allow a machine learning model to exist easily valuable relationships. For that variable will be excluded too.

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

Next, we will analyse the 'Timestamp' category. It represents the exact time when a user clicked on the advertisement. We will expand this category to 4 new categories: month, day of the month, day of the week, and hour. In this way, we will get new variables that an ML model will be able to 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 2 components that's coaching and take a look at set. We'll take and use the train_test_split operate for that and every one variable except 'Clicked on Ad' are the input values x for the cubic centimetre models. The variable 'Clicked on Ad' are keep in y, can represent the prediction variable and that we at random selected to portion thirty third of the whole knowledge for the coaching set.

1.2 Related Work

Much attention has been paid on the advertisement research recently. The best way to maximize the commercial value of advertisements is to display the ads to people who are interested in it. However, there are some issues to be dealt with, such as matching relevant advertisements for a query, ranking of the candidate advertisements, deciding how to display the advertisements on the search result page, click prediction and analysis for the presenting advertisements, and pricing of the advertisements. Several machine learning algorithms such as Logistic Regression, Linear Poisson Regression, Online Bayesian Profit Regression, Support Vector Machines, and Latent Factor Model have been adopted to predict the clicks of advertisements presented for a query. Since the size of online data is usually huge, online data stream analysis can be very helpful in Behavioural Targeting field. Behavioural 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 a promising optimization-based classification model and has extended to 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 in the way of moving it downward and upward by parameter and then classified by hyperplane to construct regression model. The excellence of MCLPR is its ability to fix the ill-posed condition with limited amount of sample, handling non-linear relationship by kernel function, and giving the global solution if it exists. MCLPR has already proved its performance in many real-life datasets.

2. SUMMARY OF VARIOUS PREDICTION

ALGORITHMS FOR DEPRESSION

a. logistical Regression:

Logistic regression could be a method for associate

Analysing a dataset that during which within there square measure one or a lot of freelance variables that confirm an outcome and therefore the outcome is measured with a divided variable (in which there square measure solely 2 doable outcomes). The goal of logistical regression is to seek out the most effective fitting (yet biologically reasonable) model to explain the connection between the divided characteristic of interest (dependent variable = response or outcome variable) and a collection of freelance (predictor or explanatory) variables and logistical regression generates the coefficients (and its commonplace errors and significance levels) of a formulate predict a logit transformation of the chance of presence of the characteristic of interest

b. Decision Tress:

Decision tree is that the most powerful and standard tool for classification and prediction. A call tree could be a multidimensional language like tree structure, wherever every internal node denotes a take a look at on associate attribute, every branch represents associate outcome of the take a look at, and every leaf node (terminal node) holds a category label. Provision perform employed in this model for

prediction of output is given below in equation 1-f (r)= 1/ 1+e –r (1)

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Male	Country	Clicked on Ad	Month	Day	Hour	Weekday
0	68.95	35	61833.90	256.09	Cloned 5thgeneration orchestration	Wrightburgh	0	Tunisia	0	3	27	0	6
1	80.23	31	68441.85	193.77	Monitored national standardization	West Jodi	1	Nauru	0	4	4	1	0
2	69.47	26	59785.94	236.50	Organic bottom-line service-desk	Davidton	0	San Marino	0	3	13	20	6
3	74.15	29	54806.18	245.89	Triple-buffered reciprocal time-frame	West Terrifurt	1	Italy	0	1	10	2	6
4	68.37	35	73889.99	225.58	Robust logistical utilization	South Manuel	0	Iceland	0	6	3	3	4

c. Support Vector Machine Algorithmic Rule

"Support Vector Machine" (SVM) could be a supervised machine learning algorithm which is used for each classification and regression challenges. However, it's largely used in classification issues. During this algorithmic rule, we have a tendency to plot each information item as some extent in n-dimensional area (where n is variety of options you have) with the worth of every feature being the worth of a specific coordinate. Then, we perform classification by finding the hyper-plane that differentiate the 2 categories o.k.. In SVM, it's straightforward to possess a linear hyper-plane between these 2 categories. However another burning question that arises is, ought to we'd like to feature this feature manually to possess a hyper-plane. No, SVM has a technique known as the kernel trick. These area unit functions that takes low dimensional input area and transform it to a better dimensional area i.e. it converts not divisible downside to divisible downside, these functions area unit known as kernels. It's largely helpful in non-linear separation downside. Simply put, it will some extraordinarily advanced information transformations, then conclude the process to separate the data supported the labels or outputs you've outlined.

A take a look at choice is chosen There area unit 3 major options:

Training and testing set: It evaluates category on however well it predicts the class of information, it absolutely was trained on.

Cross Validation: during this take a look at information set is mechanically ready from coaching data attack the idea no. Of fields provided to limit over fitting downside.

proportion Split

Now for prediction choice, output choice for prediction is chosen.

After this the 5 classifiers area unit compared on four parameters:

- Accuracy
- ROC Area
- Precision
- RMS Error

According to that Bayes internet Classifier was the foremost consistent and also the best classifier among the 5 taken for the proportion Split technique [10]. within the remainder of the sections summary of 2 a lot of strategies used for prediction of depression is given.

d. Random Forest Classifier

Random forests area unit supervised learning algorithmic rule. It are often used each for classification and regression. it's conjointly the foremost versatile and simple to use algorithmic rule. A forest is comprised of trees. It's aforementioned that the lot of trees it's, the lot of sturdy a forest is. Random forests creates call trees on willy-nilly hand-picked information samples, gets prediction from every tree and selects the most effective answer by means that of ballot. It conjointly provides the beautiful sensible indicator of the feature importance. Random forests have a range of applications, like recommendation engines, image classification and have choice. It are often wont to classify loyal loan candidates, determine dishonourable activity and predict diseases. It lies at the bottom of the Boruta algorithmic rule, that selects necessary options during a dataset.

It works in four steps:

1. Choose random samples from a given dataset.

2. Construct {a call|a choice|a call} tree {for every|for every} sample and obtain a prediction result from each decision tree.

3. Perform a vote for every expected result.

4. Choose the prediction result with the foremost votes because the final prediction.

Random forests conjointly supply a decent feature choice indicator. Scikit-learn provides an additional variable with the model, that shows the relative importance or contribution of every feature within the prediction. It mechanically computes the connexion score of every feature within the coaching section. Then it scales the connexion down so the total of all scores is one. This score can assist you select the foremost necessary options and drop the smallest amount necessary ones for model building. Random forest uses Gini importance or mean decrease in impurity (MDI) to calculate the importance of every feature. Gini importance is additionally referred to as the overall decrease in node impurity. This can be what proportion the model work or accuracy decreases once you drop a variable. The larger the decrease, the a lot of important the variable is. Here, the mean decrease may be a important parameter for variable choice. The Gini index will describe the overall informative power of the variables.

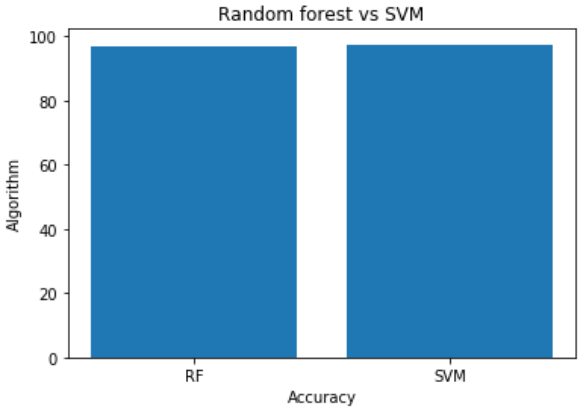
2.1 Experimental Results

As per the study, accuracy of the Support Vector Machine (SVM) amongst all the algorithms is that the highest. Amongst of these algorithms embody the supply Regression, Random Forest, KNN, and call Tree the accuracy {and the|and therefore the|and conjointly the} time potency of the Support Vector Machine is that the highest also with the time complexness and also the accuracy is ninety seven.5% with 8.000. The time prediction complexness of random forest is zero.02 and accuracy ninety five.5% that is a smaller amount than that of support vector machine. Finally, we tend to used the Support Vector Machine (SVM) algorithmic rule because the final algorithmic rule for the correct result.

Table 1: Observed result

Algorithm	Accuracy
Logistic Regression	91.5%
Random Forest	95.5%
KNN	70%
Decision Tree	93%
SVM	97.5%

Fig. Different Algorithms implemented and it's accuracy



3. CONCLUSION

The get results showed the use of both machine learning algorithms and the Random forest algorithm showed slightly better performance than the Support vector machine algorithm, but definitely, both algorithms have shown that they can be very useful in solving classification problem. The prediction results can uniquely be changed by a different approach to data analysis.

REFERENCES

[1] Fang Wang[a]*, Warawut Suphamitmongkol[a], Bo Wang[a] (2013), Advertisement Click-Through Rate Prediction using Multiple Criteria Linear Programming Regression9 Model.

[2] John Duchy, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. The Journal of Machine Learning Research, 12:2121–2159, 2011.

[3] Gouthami Kondakindi,Satakshi Rana(2014), A Logistic Regression Approach to Ad Click Prediction.

[4] Akash G. Gedam, Prof. S. G. Shikalpure. Direct Kernel Method for Machine Learning With Support Vector Machine.2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT).

[5] Mohammed A. Ambusaidi, Xiangjian He, Zhiyuan Tan, PriyadarsiNanda, Liang Fu Lu and Upasana T.Nagar, "A novel feature selection approach for intrusion detection data classification" in IEEE 13th ICTSPCC,2014.

[6] Praneeth Nskh, Naveen Varma M, Roshan Ramakrishna Naik,"Principle Component Analysis based Intrusion Detection System Using Support Vector Machine " in IEEE ICORTIEICT, May 20-21, India, 2016.

[7] Mohammed A. Ambusaidi, Xiangjian He, Priyadarsi Nanda, Zhiyuan Tan, "Building an Intrusion Detection System Using a Filter-Based Feature Selection Algorithm"in IEEE TRANSACTIONS ON COMPUTERS, VOL. 65, NO. 10, OCTOBER 2016.