Research Article

Click Prediction for Advertisement in Websites using Linear Regression

D. Ramya [(1)], S. K. Manigandan [(2)], J. Deepa [(3)], V. UmaRani [(4)]
[1] [3] Assistant Professor, VelTech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology
[2] Assistant Professor, Vel Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College
[4] Associate Professor, Jaya Engineering College

Article History: Received:11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: Prediction is one of the most powerful and effective method used nowadays for improvement in business. Machine Learning Algorithms plays a vital role in predicting the future of business. It is widely used in the field of Marketing and Advertising fields also. The Commercial Value for the advertisement is gained based on the user click on the website. Digital advertisement and marketing play very important role in influencing the profit of business. Many Machine Learning algorithms were used for predicting and analyzing the online advertisement. In this paper, Linear Regression is used for predicting the user click on the advertisement.

Key Words: Prediction, Linear Regression, Advertising, Machine Learning

1. Introduction

Data Mining is the widely used technology in the recent trends. It is used to extract the useful information from available large set of data. The extracted information is useful for taking future decisions. Machine learning algorithms were used to analyse the data[1].

Data is the most important thing in the world of Business. Data are stored by the organization for analysing the important information used in the current business and also for future use [2]. Every data is crucial in the point of view of organization which can be used for taking future decisions in the Organization.
Data mining is the technology that is used for finding out the hidden and valid data's from huge data sets. It is the combination of Statistics, Data base techniques, Artificial Intelligence and Machine Learning.

1.1 Data Mining Implementation Process

The Data mining implementation process consists of the following six processes.

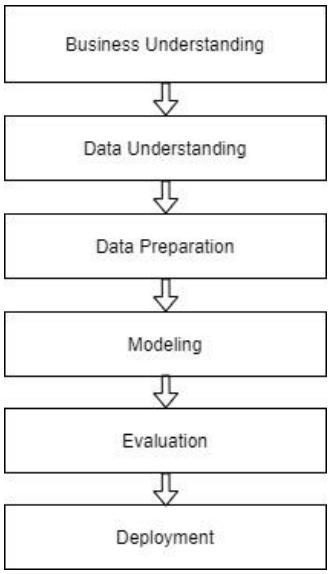

Fig 1: Data Mining Implementation Process

Step 1: Business Understanding

At first, the objectives of business and customer should be clearly defined and understood. They should find the business needs, resources, constraints and other factors which should be considered to analyse the current situation. A detailed data mining plan has to be formed to achieve business goals.

Step 2: Data Understanding

Data understanding starts with initial data collection from multiple sources such as databases, flat filter and data cubes. Data exploration needs to be done by addressing the questions. Finally, the data quality is analysed.

Step 3: Data Preparation
The outcome of this phase is the final data set. Once available data source is identified, it should be cleaned and formatted into desired format. Data exploration is done to explore the patterns from business understanding.

Step 4: Modelling
Modelling techniques have to be selected. Then, the test scenario must be created to validate the quality of the model. Finally, models have to be analysed to make sure that the models met business objectives.

Step 5: Evaluation
In this step, the models must be evaluated in terms of business objectives. New business objectives which arise later and existing objectives have to be reviewed thoroughly.

Step 6: Deployment
This phase presents the information collected through the data mining process to the stakeholders. Based on the business requirements, the deployment can be either simple or complex.

1.2 Artificial Intelligence

Artificial Intelligence is the branch of Computer science which focuses the development of intelligent machines [3]. It is otherwise called as Machine Intelligence. It refers to the programming of machines to think like human and exhibit their actions. Machines that exhibits human traits such as learning, reasoning, problem-solving and predicting the future also said to be Artificial Intelligence.

There are three types of Artificial Intelligence. They are Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI).

The main goal of the Artificial Intelligence is that to create technology which makes computer and machines to think and functions like a human in intelligent manner. The AI is applied in all sectors such as Finance, marketing, Banking, Trading, Healthcare Industries, etc.

1.3 Machine Learning

Machine Learning is a subset and application of Artificial Intelligence (AI) which makes the system the ability to learn by itself and from its own experience without being programmed. It develops the computer programs that can access data and use that data to learn by it. The main aim of the Machine Learning is to allow the computers learn automatically and adjust the process and actions without the intervention of human.

The categories of Machine learning are the following.

1.Supervised Learning
Supervised Learning is the algorithm which is used to learn the mapping function from input variables (X) and an output variable (Y). The relation is given by Y=f(X)

The output (Y) can be easily predicted for any value of input variable (X). Both input and output variables are known.

2.Unsupervised learning
Unsupervised Learning is the algorithms which have only input data (X) and there are no corresponding output variables. Unlike supervised learning, there is no predicted output. Algorithms are left to find their own decisions to represent the structure of data.

3.Semi Supervised Learning
It is the combination of small amount of labelled data and large amount of unlabelled data. There are three types of assumptions namely Continuity assumption, Cluster assumption and Manifold assumption which is assumed by the algorithm about the data.

1.4 Machine Learning in the field of Advertising

People were widely using the Internet nowadays. Internet has become the key which connects the people all over the world. Companies aim to post the advertisement in the relevant website in order to get more revenue for the business. Users clicking on the advertisement helps to identify the most relevant advertisement for each user. The advertisers pay for the advertisement when the user click on it [4].

Click-through rate is the ratio of number of clicks to the point at which the ad is viewed once by the user. This is one of the metrics used to calculate the commercial value of an advertisement. CTR helps to find out the amount that should be paid by the advertisers since this CTR leads to pay-per-click (PPC) success. This

advertisement on websites is purely an auction based. Highest bidder will get the opportunity to post the advertisement.

2. Proposed Methodology

2.1 Linear Regression

Machine Learning is considered to be the field of predictive modelling. Most accurate predictions can be done. Linear regression is one of the simplest supervised machine learning algorithms that is used for predictive modelling. It is one of the common methods which is used for predicting the future. Linear Regression is both a statistical algorithm and a machine learning algorithm [5]. It is an attractive model for defining the relationship between input variables and a single output variable.

The representation of the model is very simple. It combines a set of input values (x) to the predicted output (y). Both input and output are numerical and continuous values. Input variables x is called predictor or independent variable and output variable y is called dependent or response variable.

The relation is given by

$$y = a0 + a1 * x$$

The values of a0 and a1 must be taken such that they minimize error. The a0 is intercept and a1 is co-efficient. Intercept a0 is given by

$$a0 = \bar{y} - a1\bar{x}$$

The Co-efficient formula is given by

$$a1 = \frac{\sum_{i=1}^{n}(x - \bar{x})(y - \bar{y})}{\sum_{i=1}^{n}(x - \bar{x})^2}$$

Error calculation is given by

$$Error = \sum_{i=1}^{n} A - P$$

Where A is actual output and P is Predicted output respectively.

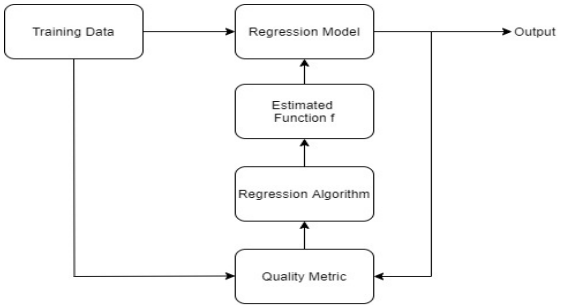

Fig 2: Block diagram for Linear Regression

Fig 2 and Fig 3 gives the block diagram and flow chart of Linear Regression algorithm. The training data is pre-processed and sent to Regression model [6]. It gets the input from predefined quality metrics and functions of Linear Regression algorithm. The data is processed and performance is analysed.

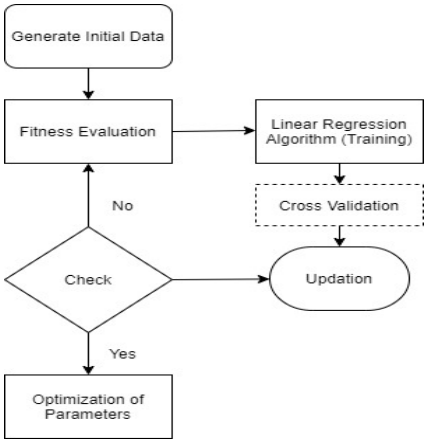

Fig 3: Flow chart for Linear Regression

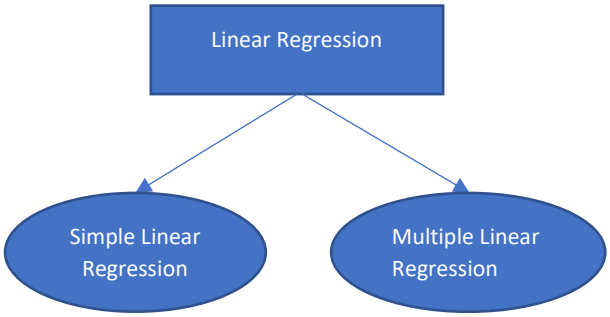

Fig 4: Types of Linear Regression

2.2 Simple Linear Regression

It is the model which exhibits the relationship between a dependent variable (Continuous or real value) and a single independent variable (continuous or categorial values).[7] The diagram shows sloped straight line. So, it is called Simple Linear Regression.

The main advantages of Linear Regression algorithm are the following

(i) Gives the relationship between the two variables
(ii) Future forecasting

The Simple Linear Regression model can be represented by the equation:

$y= a_0+a_1x+e$

Where,

a0= Intercept of the Regression line

a1= It is the slope of the regression line, which explains whether the line is decreasing or increasing. e = The error term.

2.3 Multiple Linear Regression

It is the model which exhibits the relationship between a single variable (dependent Continuous variable) and a more than one independent variable (continuous or categorial values). It fits a regression line through a multidimensional space of data points.

The equation is given by:

Y = a0 + a1 * x1 + a2 * x2 + a3 * x3 + …… an * xn
Y = Dependent variable and x1, x2, x3, …… xn are multiple independent variables

3. Result Analysis

The data pre-processing techniques were applied and prepared the dataset in the correct format. Then the filter is applied to select only the relevant features. In the advertising example, the dataset contains the following fields. [8] Time spent by the consumer spent on website in minutes, Age of the consumer, Average income of the area of the consumer, Daily Internet usage of the consumer, heading of the advertisement, City, Country and Gender of consumer, Time at which the consumer clicked on Ad and Whether they click on Ad or not is notified. [9] [10]

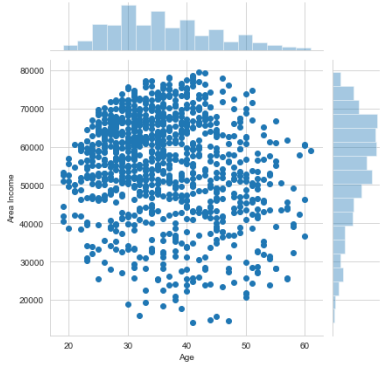

Fig 4: Jointplot showing Area Income versus Age

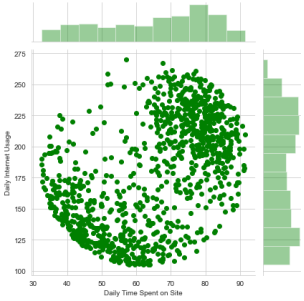

Fig 5: Jointplot showing the kde distributions of Daily Time spent on site vs. Age.

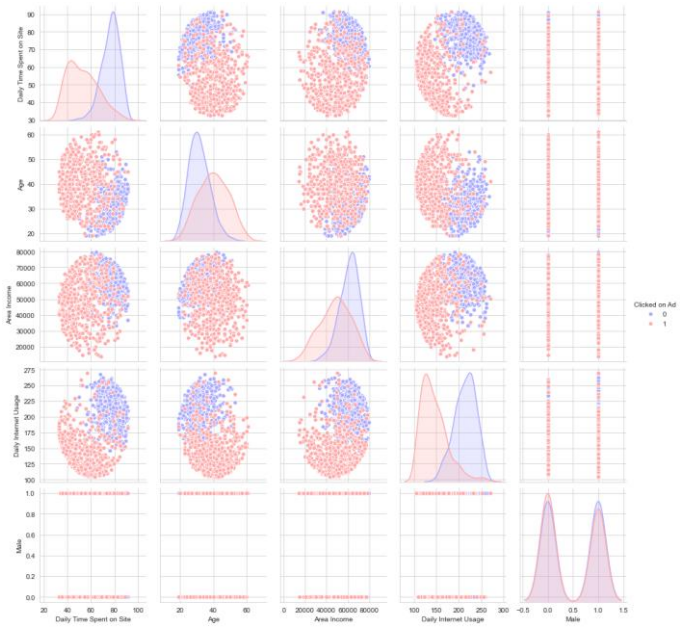

Fig 6: Pair plot with the hue defined by the 'Clicked on Ad' column feature.

	Precision	**Recall**	**F1-score**	**Support**
0	0.87	0.96	0.91	162
1	0.96	0.86	0.91	168
avg/total	0.91	0.91	0.91	330

Table: Performance analysis for Click through rate

The four-performance metrics are considered and the Linear Regression Algorithm gives the performance of 0.91% which is considered to be a higher.

4. Conclusion

In this paper, the click prediction is done for an advertisement based on the criteria of age, income, Internet usage, etc. The advertisement should be related to the website concepts and products. An advertiser can gain more profit, if number of users click on the advertisement is higher. The performance of Machine Learning algorithm is analysed based on number of clicks on the advertisement. The Linear Regression algorithm is used and the performance is achieved by 91%. The performance can be calculated by using other Machine Learning algorithms and comparative study also can be done.

References

1. G. Broussard, "How Advertising Frequency Can Work To Build Online Advertising Effectiveness," international Journal oj Market Research, vol. 42, no. 4, Jan. 2000.
2. R. L. Zeff and B. Aronson, Advertising on the internet, 2nd ed. New York, NY, USA: John Wiley & Sons, Inc., 1999.
3. A. M. Masucci and A. Silva, "Advertising competitions in social networks," http://www.dim.uchile.cl/∼alsilva/ACC17/acc2017.pdf, 2017.
4. S. Bharathi, D. Kempe, and M. Salek, "Competitive influence maximization in social networks," in Internet and Network Economics, ser. Lecture Notes in Computer Science, X. Deng and F. Graham, Eds. Springer Berlin Heidelberg, 2007.
5. Chen, T., Guestrin, C.: Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
6. eMarketer. Online ad spending to total $19.5 billionvin 2007. http://www.emarketer.com/Article.aspx?1004635,Feb. 2007.
7. A. Metwally, D. Agrawal, and A. E. Abbadi. Using association rules for fraud detection in web advertising networks. In VLDB 2005.
8. Simon Breedon, "The Rapidly Growing Digital Advertising Market," http://viralmarketingmonsters.sharedby.co/share/8vbMff," (Accessed: 10 February 2015).
9. Weiss, D. (2008). Predicting ads click-through rate with decision rules.
10. Zhang, Y., Jansen, B., Spink, A. (2008). Identification of factors predicting click-through in web searching using neural network analysis. Journal of the American Society for Information Science, 60(3), 1-15.
11. Chen, T., He, T., Benesty, M.: Xgboost: extreme gradient boosting. R package version 0.4-2, 1–4 (2015)
12. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. 63(1), 3–42 (2006)
13. Balfer, J., Bajorath, J.: Systematic artifacts in support vector regression-based compound potency prediction revealed by statistical and activity landscape analysis. PLoS One 10(3), 0119301 (2015)
14. Teng-Kai Fan, Chia-Hui Chang, "Sentiment-oriented contextual advertising" Knowledge and Information Systems, June 2010, Volume 23, Issue 3, pp 321–344
15. Sandra Soroa-Koury, Kenneth C.C. Yang, "Factors affecting consumers' responses to mobile advertising from a social norm theoretical perspective", Telematics and Informatics, 27 (2010) 103–113
16. Kai Li , Timon C. Du , "Building a targeted mobile advertising system for location-based services", Decision Support Systems, v. 54, 2012, pp. 1-8