Analysis on Worldwide Online Rating Systems by Using Multi Linear Regression Technique in Machine Learning

[1]Shashishekhar, [2]Ashish Sharma,

[1]Shashishekhar, [2]Ashish Sharma,
Department of Computer Engineering, GLA University, Mathura.
Department of Computer Engineering, GLA University, Mathura.
E-Mail: shashi.shekhar@gla.ac.in, ashishs.sharma@gla.ac.in

ABSTRACT

In today's world online rating systems are extensively used for making their decisions on a particular product or item on the web. For getting benefit, some people are always trying to manipulate corresponding systems by giving unreasonable ratings. So, resolving true ratings of such products comes extremely important and it is a crucial problem. Fake reputation is the problem that is being occurred by unreasonable ratings. In this paper we propose Multi linear regression algorithm in machine learning which will give correct reputation by eliminating unreasonable ratings.

Keywords: Machine learning, online trading

Introduction

Online Ratings plays a major role in today's life style. More number of public used internet because it provides so many services like education, gaming and also formed more number of opportunities i.e. saving money and time, delivery of the product is very fast, all items are found at a time, we buy all the products at home. Customers share their purchasing decisions in the form of online reviews and ratings on various items. These kinds of results are called rating score.

In 2013 enlargement rate of India's e-commerce industry is absorption, was repeated up to 88%. It represents least economic growth of india. The range of the market was $16 billion in 2013, and $8.5 billion in 2012. The main intension in this system is to know the performance of purchasing various customers or peoples in India. It has also made an attempt to get information about the scope at improving in online shopping website. In the year 1981 the world first direct business in the direction of business (b2b) online shopping, in the year 1984 business to consumer online shopping. Presaging the home shopping industry in 2008, the UK online residence shopping see today marketplace was significance further than $50 billion, $2.6 to $3.6 billion of that is from grocery shopping these developments are established on advance made in the 1990s such as everywhere right to use to the worldwide web, rationally protected use of debit cards over the internet.

Online marketing is the major part in present society so more number of customers will be actively participated to purchase the products**.** For any kind of e- commerce applications, trust is the key factor while parching any item on online we need to study the ratings of that particular product.

In traditional way purchasing is done in director way in that way customers went to the market and buy the product directly without any rating of particular product. The problem of the traditional way approach is waste of money and time. In olden days people living style different they have free time to shopping. The generation changes and the living style of people also changes, now a day's peoples like to purchasing goods on online shopping.

Literature Survey

R.Burke, B.Mobasher, C.Williams and R.Bhaumik, for the purpose of users in the online recommender systems they have developed many collaborative filtering techniques. To identify shilling hits and to analyze rating patterns of harmful users, they proposed several metrics. Based on the results, they proposed and evaluated algorithm for protecting rating system against shilling attacks.[1]

M.Brennan, S.wrazien, and R.greenstadt, developed online support vector relapse approach for following about engine shaft misalignment Furthermore Feedwater stream Rate, auspicious What's more correct majority of the data something like embroyonic errors in internet machines will significantly expand the improvement from claiming ideal support methods. That provision of backing vector relapse to machine wellbeing screening might have been as of late investigated; those fruition from claiming information will be relying upon bunch handling of the accessible information. [2]. V. Barnett and t. Lekuiswis, an incorporated Clustering-Based system heartiness of notoriety framework is revised, to distinguishing the nonsensical quick monies viably. In this framework they present an incorporated Clustering-Based methodology should experience outlandish testimonies for notoriety frameworks. It acknowledges grouping and acknowledges purchasing agents' information regarding offering operators. Test assessment exhibits guaranteeing comes about our approach for sifting different sorts of outlandish testimonies.[3]

P.Chirita, W.Nejdl and C.Zamfir, introducing Fuzzy a Logic Based Reputation Model against Unwarranted Ratings**,** Reputation is the most important factor in online rating systems. Due to the presence of unreasonable ratings, users have a fear of the correctness of rating systems. Many approaches was developed, in those reputation was calculated based on the reliability of the rating supplier alone without including other aspects of the rating. They calculated based on average weights. To solve this trouble, they introduced a standing representation that thinks and unites the sequential, resemblance and quantify features of the user ratings. The projected representation is extra robust compared to traditional approaches and also show the investigational outcomes depend on a set of real user data from a cyber rivalry.[4]

M.Eirinaki, M.D.Louitta, and I.varlamis, Introduced Visualizing unreasonable ratings in online reputation systems in e- commerce site, to calculate the quality of products rating systems provide a best method. But, due to their fame, rating systems are subject to many different attacks. The main problem in online rating system are unreasonable ratings which are used to unrepeatedly increase or decrease the reputation of a product. In this strategy, they hired visual analytics to identify colluding digital identities. The main advantage of this approach is the transparent revelation of the true rating of an item by interactively using both internal cause and external cause discounting methods. They extend transparency, provide greater robust rating systems and extended user experience.[5]

Problem Statement

Fake reputation or rating which means the users who give unreasonable ratings on a product is the major problem, in Online rating systems, and also find out various reasons for getting fake reputations or unwarranted ratings.

In this System, we mainly consider three types of users i.e. normal, abnormal and malicious. In traditional approaches depending upon abnormal users rating reputation calculation will be done. After calculation of reputation, finally it will generates wrong result. Due to this wrong result we are not getting accurate ratings correctly. In order to overcome that problem we are developing a solution by using an multi linear regression algorithm in machine learning to predict the correct result of the users.

Existing Method

In the existing system, we are not getting accurate ratings on true-reputation. The mainly frequent way to get ratings is to use the average which may generate unreasonable rating. Including abnormal users ratings the reputation calculation will be done, after calculation of reputation the rating of particular product may raise or fall. In traditional approach, unreasonable rating problem is solved by eliminating abnormal users, but these users are not always detected. In some cases, normal users ratings will be treated as abnormal users ratings and predicting the result. In this situation, normal users are treated as abusers. Because the correct ratings are also considered as wrong ratings and generating the result.

Proposed Method

Methodology

For developing the system, certain methodologies have been used. The methodology used in this project is Multi linear regression algorithm in machine learning, which is used to predict the true ratings of the users by eliminating unreasonable ratings or fake ratings. It follows certain different steps as follows:

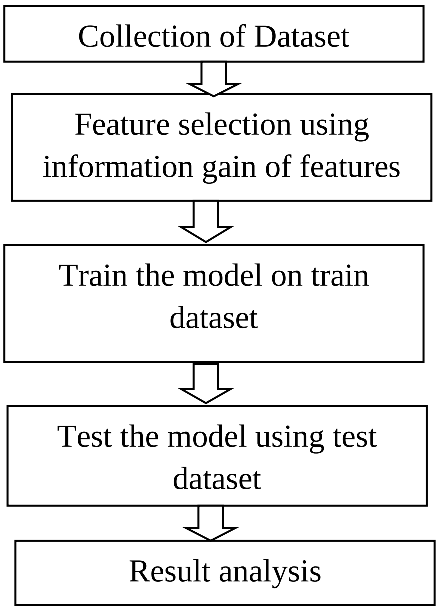

Collection of Dataset
Feature selection using information gain of features
Train the model on train dataset
Test the model using test dataset
Result analysis

Figure.1. **Steps in methodology**

Multi Linear Regression algorithm is otherwise called multiple regression, which is An measurable strategy that utilization a few logical variables should anticipate those Conclusion of a reaction variable. The objective of multi linear regression (MLR) may be with model the straight relationship between those logical (independent) variables and reaction (dependent) variable. Multiple regression dissection is a development of simple linear regression.

Implementation

Data mining is a methodology that starts with data preprocessing. A Data preprocessing is a data mining technique that engages altering unprocessed information into a comprehensible arrangement. It comprises certain steps as follows:

Step 1-Collection of Dataset

The dataset is an assortment of related sets of information that is composed of separate elements. In order to generate a machine learning model they require dataset that works on data completely using machine learning model

Step -2 introduce Libraries

For performing data preprocessing we use python imported libraries with 3 specific data preprocessors. To make requests to the particular prediction and process the returned data we will make use of a few standard libraries. Those libraries are pandas, numpy, matplotlib.

$$import\ numpy\ as\ np$$
$$import\ pandas\ as\ pd$$
$$import\ matplotlib.pyplot\ as\ plt$$

Step-3 introduce datasheet

For importing the datasheet firstly a Pandas Data Frame is produced by loading the datasets from active storage, that storage can be SQL Database, CSV (Comma Separated Value) file, or Excel file. For creating the Pandas Data Frame lists, dictionary, and from a list of dictionary etc can be used. The csv file (dataset) is read using pandas and an object to the dataset is created i.e, df (data frame).

dataset=pd.read_csv('onlinerating1.csv')
dataset

A Data frame is a two-dimensional data structure i.e., data is associated in tabular fashion in rows and columns. After uploading the dataset we need to split the dataset into two labels x (dependent variable) and y (independent variable).The splitting of x and y is done by using iloc[].

$$x = dataset.iloc[:,0:7]$$
$$y = dataset.iloc[:,-1]$$

Here unique id, crawl timestamp, product names, product id, retail price, discounted price, product rating are independent variables and overall rating is the dependent variable in which results are in binary.

Step 4-Checking for null values

The real world data may contain inconsistent, incomplete and noisy data. While applying the machine learning algorithms, the dataset should cleansed. However, it is done in preprocessing, we have to check for missing values. This can be done by using the function null(). The output value is Boolean value (true/false).

dataset.isnull().any()

Step 5-Checking for categorical data

The purpose field in the dataset consists of categorical data. As it cannot be run, it should be converted into numeric data. For converting, we have used label encoder which is imported from sklearn library. Label encoder converts the labels in purpose fields into numerics format.

Step 6-Splitting the dataset into training and test data

While working with datasets, a machine learning algorithm facilitates in two kinds of sets. We usually split the data around 20%-80% among testing and training sets. The train set is used for training and fitting the dataset and test set for testing.

$$\boldsymbol{from\ sk\ learn.model - selection\ \ import\ train\_test\_split}$$
$$\boldsymbol{x\_train, x\_test, ytrain, y\_test = train\_test\_split(x, y, tetst\_size = 0.2, random\_state = 0)}$$
$$\boldsymbol{The\ line\ test_{size} = 0.2\ suggests\ that\ the\ test\ data\ should\ be\ 20\%\ of\ the\ dataset\ and\ the\ rest\ should\ be\ train\ data.}$$

Step -7 attribute Scaling

It is a technique which is worn to normalize the assortment of self-governing variables or features of data. In data preprocessing, it is also recognized as data normalization and is usually executed throughout the data preprocessing step.

$$\boldsymbol{from\ sklearn.preprocessing\ import\ StandardScaler}$$
$$\boldsymbol{sc = StandardScaler()}$$
$$\boldsymbol{x\_train = sc.fit\_transform(x\_train)}$$
$$\boldsymbol{x\_test = sc.transform(x\_test)}$$

Step 8-Visualization

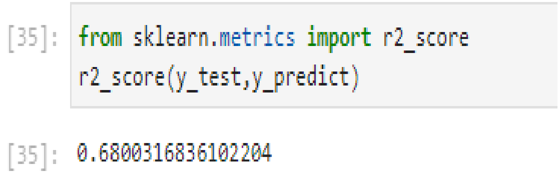
Fit the model to selected supervised data by using the matplotlib library for visualizing the independent variables and dependent variables. X label represents the different names of a products and the Y label represents the ratings given by different users.

Step 9-Model Fitting and Prediction

For giving preparing alternately fitting the model of the preparation set, we will import the straight relapse class of the sk figure out library. Then afterward importing those class, we will make a classifier item and utilize it to fit the model of the multi linear regression. Then afterward fitting, our model will be great prepared on the preparing set, something like that we will Right away anticipate the come about by utilizing test set data.

Results and Discussion

Prediction: It is worn to expect the accurate outcomes. To calculate the precise result in the code there are some calculations given in the code.



```
[35]: from sklearn.metrics import r2_score
      r2_score(y_test,y_predict)

[35]: 0.6800316836102204
```

Figgure.2. Predicted value

Accuracy is a major estimation of the categorization representations. Accuracy is the portion of prophecy that we got in our model or project. Accuracy comes out from or to 0.91, or 91%. Absolutely, let's do a earlier investigation of positives and negatives to increase more imminent model presentation.

Conclusion

In this project we describe false standing as a trouble which has occurred mostly which means in some situations true reputation or true rating are also considered as false rating and getting the result. To avoid that kind of situations we propose multi linear regression algorithm which will give accurate result.

References

1. I. Gunes, C. Kaleli, A. Bilge, and H. Polat, “Shilling attacks against recommender systems: comprehensive survey,” Artif. Intell. Rev., vol. 42, no. 4, pp. 767–799, 2014.
2. G. Häuubl and V. Trifts, “Consumer decision making in online shopping environments: The effects of interactive decision aids,” Market. Sci., vol. 10, no. 1, pp. 4–21, 2000.
3. J. Howe, “The rise of crowdsourcing,” Wired Mag., vol. 14, no. 6, pp. 1–4, 2006.
4. N. Hurley, Z. Cheng, and M. Zhang, “Statistical attack detection,” in Proc. ACM Conf. Recommender Syst. (RecSys), Vienna, Austria, 2009, pp. 149–156.
5. J. A. Konstan and J. Riedl, “Recommender systems: From algorithms to user experience,” User Model. User-Adapt. Interact., vol. 22, nos. 1–2, pp. 101–123, 2012.
6. C. Leadbeater, WE-THINK: Mass Innovation, Not Mass Production. London, U.K.: Profile Books, 2008.
7. J.-S. Lee and D. Zhu, “Shilling attack detection—A new approach for a trustworthy recommender system,” INFORMS J. Comput., vol. 24, no. 1, pp. 117–131, 2012.
8. P. Levy, L’Intelligence Collective: Pour Une Anthropologie du Cyberspace. Paris, France: La Découverte, 1997.
9. E. P. Lim, V. A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, “Detecting product review spammers using rating behaviors,” in Proc. 19th ACM Int. Conf. Inf. Knowl. Manage. (CIKM), Toronto, ON, Canada, 2010, pp. 939–948.
10. M. Limayem, M. Khalifa, and A. Frini, “What makes consumers buy from Internet? A longitudinal study of online shopping,” IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 30, no. 4, pp. 421–432, Jul. 2000.
11. S. Liu, J. Zhang, C. Miao, Y. Theng, and A. Kot, “iCLUB: An integrated clustering-based approach to improve the robustness of reputation systems,” in Proc. 10th Int. Joint Conf. Auton. Agents Multiagent Syst. (AAMAS), Taipei, Taiwan, 2011, pp. 1151–1152.
12. A. Mukherjee, B. Liu, J. Wang, N. Glance, and N. Jindal, “Detecting group review spam,” in Proc. 20th Int. Conf. World Wide Web (WWW), Hyderabad, India, 2011, pp. 93–94.
13. B. Mobasher, R. Burke, R. Bhaumik, and C. Williams, “Towards trustworthy recommender systems: An analysis of attack models and algorithm robustness,” ACM Trans. Internet Technol., vol. 7, no. 2, pp. 1–40, 2007.
14. Kumar, Manoj, and Ashish Sharma. "Mining of data stream using “DDenStream” clustering algorithm." 2013 IEEE International Conference in MOOC, Innovation and Technology in Education (MITE). IEEE, 2013.
15. Sharma, Ashish, Anant Ram, and Archit Bansal. "Feature Extraction Mining for Student Performance Analysis." Proceedings of ICETIT 2019. Springer, Cham, 2020. 785-797.
16. Sharma, Ashish, and Dhara Upadhyay. "VDBSCAN Clustering with Map-Reduce Technique." Recent Findings in Intelligent Computing Techniques. Springer, Singapore, 2018. 305-314.
17. Sharma, Ashish, Ashish Sharma, and Anand Singh Jalal. "Distance-based facility location problem for fuzzy demand with simultaneous opening of two facilities." International Journal of Computing Science and Mathematics 9.6 (2018): 590-601.
18. Agarwal, Rohit, A. S. Jalal, and K. V. Arya. "A review on presentation attack detection system for fake fingerprint." Modern Physics Letters B 34.05 (2020): 2030001.
19. Mishra, Ayushi, et al. "A robust approach for palmprint biometric recognition." International Journal of Biometrics 11.4 (2019): 389-408.

20. Singh, Anshy, Shashi Shekhar, and Anand Singh Jalal. "Semantic based image retrieval using multi-agent model by searching and filtering replicated web images." 2012 World Congress on Information and Communication Technologies. IEEE, 2012.
21. Shekhar, Shashi, et al. "A WEBIR crawling framework for retrieving highly relevant web documents: evaluation based on rank aggregation and result merging algorithms." 2011 International Conference on Computational Intelligence and Communication Networks. IEEE, 2011.
22. Varun K L Srivastava, N. Chandra Sekhar Reddy, Dr. Anubha Shrivastava, "An Effective Code Metrics for Evaluation of Protected Parameters in Database Applications", International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.1.3, 2019. doi.org/10.30534/ijatcse/2019/1681.32019
23. Varun K L Srivastava , N. Chandra Sekhar Reddy , Dr. Anubha Shrivastava, "An efficient Software Source Code Metrics for Implementing for Software quality analysis", International Journal of Emerging Trends in Engineering Research, Volume 7, No. 9 September 2019.