IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 7, July 2021

DOI 10.17148/IJARCCE.2021.10733

E-COMMERCE PRODUCT RATING BASED ON CUSTOMER BASED REVIEW MINING

Bodduluri Hemacharita[1], Amshini Magaima[2], Mr.D.Anand Joseph Daniel, M.E, (Ph.D)[3]

UG Scholar, Computer science Engineering, Anand Institute Of Higher Technology, Chennai, India[1,2]

UG Scholar, Computer science Engineering, Anand Institute Of Higher Technology, Chennai, India[3]

Abstract: The numerous activities related to e-commerce are carried out in social networks, in which trust plays an important role in decision making of customers. Suggestion by a friend is a common service that has been provided by almost all of the social networks, and evaluation of trust between users improves the quality of suggestions. Social networks have become the main infrastructure of today's daily activities of people during the last decade. In these networks, users interact with each other, share their interests on resources and present their opinions about these resources or spread their information. Since each user has a limited knowledge of other users and most of them are anonymous, the trust factor plays an important role on recognizing a suitable product or specific user. The inference mechanism of trust in social media refers to utilizing available information of a specific user who intends to contact an unknown user. Next, fuzzy logic is incorporated to rank the membership of trust to a specific class, according to two-, three- and five-classes classification. Finally, to classify the trust values of users, three machine learning techniques, namely Linear Regression, Decision Tree (DT), and Random forest, are used instead of traditional weighted sum methods, to express the trust between any two users in the presence of a special pattern.

Keywords: Data mining, Pre processing, Transformation, Data Mining.

I. INTRODUCTION

Online reviews have become an important factor when people make purchase and business decisions. Seller selling products on the web often ask or take reviews from customers about the products that they have purchased. As e-commerce is growing and becoming popular day-by-day, the number of reviews received from customer about the product grows rapidly. For a popular product, the reviews can go upto thousands. The increasing popularity of online reviews also stimulates the business of fake review writing, which refers to paid human writers producing deceptive reviews to influence readers' opinions. Our project tackles this problem by building a classifier that takes the review text and the basic information of its reviewer as input and outputs whether the review is reliable. This creates difficulty for the potential customer to read them and to make a decision whether to buy or not the product. Problems also arise for the manufacturer of the product to keep track and to manage customer opinions. And also, additional difficulties are faced by the manufacturer because many other merchants' sites may sell the same product at good ratings and the manufacturer normally produces many kinds of products.

II. ANALYSIS

System analysis is a problem-solving technique that decomposes a system into its component pieces for the purpose of the studying how well those component parts work and interact to accomplish their purpose along with the accurate measurement of performance delivered by the system.

To calculate and analyzing the patterns of the input data, and building a model, machine learning methods are very flexible in evaluating the test data. Furthermore, different classifiers are proper solutions to classify the trust in social networks. It extracts the feature vector from a dataset of raw information after pre-processing. Feature vector is fed into a classifier, and the extra information have been removed from it. For a better granularity, the trust value in feature vectors is converted into fuzzy values using membership functions

(i) using hybrid methods (graph structure of the network and interactions of the users) for a better performance.

(ii) using a dataset with basic and limited information, because evaluating trust with datasets with comprehensive information is much simpler and yields better results, but datasets such as the one used are more challenging; however, the processing speed of these methods are very high.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 7, July 2021
DOI 10.17148/IJARCCE.2021.10733

(iii) extraction of new feature based on the dataset by processing the basic features; in fact, the proposed features are not present in the previous works to the best of our knowledge.
(iv) using machine learning methods for evaluating trust instead of using weighted sums, to obtain a model that can decide whether there is trust or not, based on the feature vector; three methods are used for obtaining reliability, namely Random Forest, decision tree, and Linear Regression.

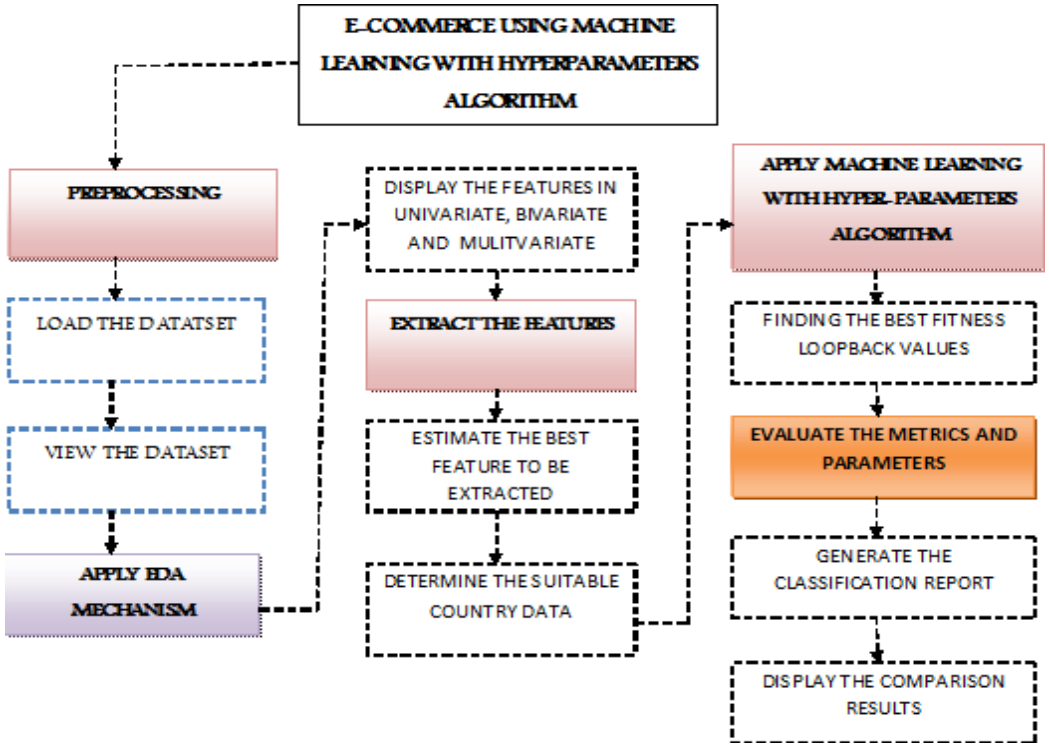

Fig 3.1 System Architecture

III. SYSTEM REQUIREMENTS

3.1 DATA MINING

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets ("big data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves data management model and inference considerations-interestingness-metrics, complexity considerations,post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining). It involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms data dredging, data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 7, July 2021
DOI 10.17148/IJARCCE.2021.10733

It concerns large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. The data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

3.2 DATA COLLECTION

Data collection is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools. The challenges include capture, curation, storage, search, sharing, analysis, and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions.

Put another way, big data is the realization of greater business intelligence by storing, processing, and analyzing data that was previously ignored due to the limitations of traditional data management technologies

What to do with the data

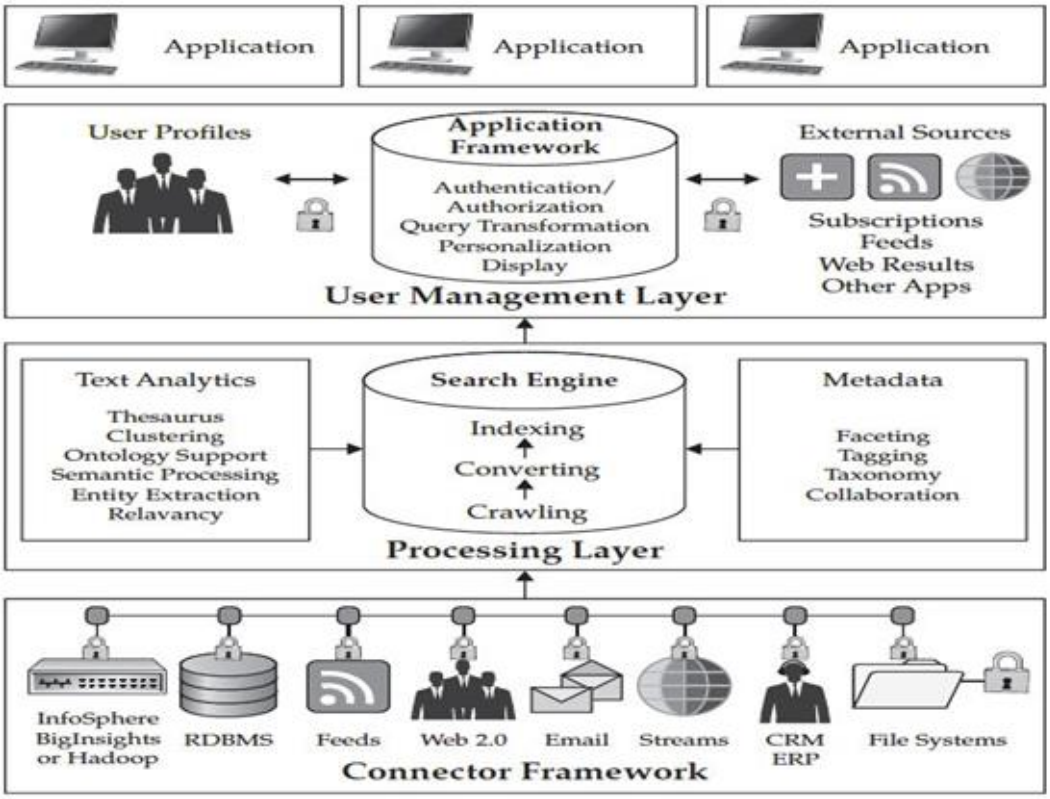

Fig 3.2 Overall architecture

IV. TECHNOLOGY

4.1 PYTHON

Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 7, July 2021
DOI 10.17148/IJARCCE.2021.10733

Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of Python's other implementations. Python and CPython are managed by the non-profit Python Software Environment.
Python is a multi-paradigm programming language. Object-oriented programming and structured programming are fully supported, and many of its features support functional programming and aspect-oriented programming (including by metaprogramming and meta-objects). Many other paradigms are supported via extensions, including design by contract and logic programming.

Most Python implementations (including CPython) include a read–eval–print loop (REPL), permitting them to function as a command line interpreter for which the user enters statements sequentially and receives results immediately.

Other shells, including IDLE and IPython, add further abilities such as auto-completion, session state retention and syntax highlighting.

As well as standard desktop integrated development environments, there are Web browser-based IDEs; SageMath (intended for developing science and math-related Python programs); Python Anywhere, a browser-based IDE and hosting environment; and Canopy IDE, a commercial Python IDE emphasizing scientific computing.

Python uses dynamic typing, and a combination of reference counting and a cycle-detecting garbage collector for memory management. It also features dynamic name resolution(late binding), which binds method and variable names during program execution.

The language's core philosophy is summarized in the document The Zen of Python (PEP 20), which includes aphorisms such as:[49]

- Beautiful is better than ugly
- Explicit is better than implicit
- Simple is better than complex
- Complex is better than complicated
- Readability counts

Rather than having all of its functionality built into its core, Python was designed to be highly extensible. Compact modularity has made it particularly popular as a means of adding programmable interfaces to existing applications. Van Rossum's vision of a small core language with a large standard library and easily extensible interpreter stemmed from his frustrations with ABC, which espoused the opposite approach.

Python's developers strive to avoid premature optimization, and reject patches to non-critical parts of the CPython reference implementation that would offer marginal increases in speed at the cost of clarity. When speed is important, a Python programmer can move time-critical functions to extension modules written in languages such as C, or use PyPy, a just-in-time compiler. Cython is also available, which translates a Python script into C and makes direct C-level API calls into the Python interpreter.
An important goal of Python's developers is keeping it fun to use. Now it
is reflected in the language's name—a tribute to the British comedy group Monty Python[52]—and in occasionally playful approaches to tutorials and reference materials, such as examples that refer to spam and eggs (from a famous Monty Python sketch) instead of the standard foo and bar.
A common neologism in the Python community is pythonic, which can have a wide range of meanings related to program style. To say that code is pythonic is to say that it uses Python idioms well, that it is natural or shows fluency in the language, that it conforms with Python's minimalist philosophy and emphasis on readability. In contrast, code that is difficult to understand or reads like a rough transcription from another programming language is called unpythonic.
Python is meant to be an easily readable language. Its formatting is visually uncluttered, and it often uses English keywords where other languages use punctuation. Unlike many other languages, it does not use curly brackets to delimit blocks, and semicolons after statements are optional. It has fewer syntactic exceptions and special cases than C or Pascal.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 7, July 2021
DOI 10.17148/IJARCCE.2021.10733

Python has extensive built-in support for arbitrary precision arithmetic. Integers are transparently switched from the machine-supported maximum fixed-precision (usually 32 or 64 bits), belonging to the python type int, to arbitrary precision, belonging to the Python type long, where needed. The latter have an "L" suffix in their textual representation. (In Python 3, the distinction between the int and long types was eliminated; the behavior is now entirely contained by the int class.) The decimal type/class in module decimal (since version 2.4) provides decimal floating-point numbers to arbitrary precision and several rounding modes. The fraction type in module fractions (since version 2.6) provides arbitrary precision for rational numbers

Python's large standard library, commonly cited as one of its greatest strengths, provides tools suited to many tasks. For Internet-facing applications, many standard formats and protocols such as MIME and HTTP are supported. It includes modules for creating graphical user interfaces, connecting to relational databases, generating pseudorandom numbers, arithmetic with arbitrary precision decimals, manipulating regular expressions, and unit testing.

Some parts of the standard library are covered by specifications (for example, the Web Server Gateway Interface (WSGI) implementation follows PEP 333), but most modules are not. They are specified by their code, internal documentation, and test suites (if supplied). However, because most of the standard library is cross-platform Python code, only a few modules need altering or rewriting for variant implementations.

4.2 MACHINE LEARNING

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on models and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in the applications of email filtering, detection of network intruders, and computer vision, where it is infeasible to develop an algorithm of specific instructions for performing the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

Classification algorithms and regression algorithms are types of supervised learning. Classification algorithms are used when the outputs are restricted to a limited set of values. For a classification algorithm that filters emails, the input would be an incoming email, and the output would be the name of the folder in which to file the email. For an algorithm that identifies spam emails, the output would be the prediction of either "spam" or "not spam", represented by the Boolean values true and false. Regression algorithms are named for their continuous outputs, meaning they may have any value within a range. Examples of a continuous value are the temperature, length, or price of an object.

In unsupervised learning, the algorithm builds a mathematical model of a set of data which contains only inputs and no desired outputs. Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points. Unsupervised learning can discover patterns in the data, and can group the inputs into categories, as in feature learning. Dimensionality reduction is the process of reducing the number of "features", or inputs, in a set of data.

V. ALGORITHM DESIGN

5.1 RANDOM FOREST

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance. Random forests are frequently used as "blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration in packages such as scikit-learn.

5.2 LINEAR REGRESSION

Adding one further step of randomization yields extremely randomized trees, or Extra Trees. While similar to ordinary random forests in that they are an ensemble of individual trees, there are two main differences: first, each tree is trained using the whole learning sample (rather than a bootstrap sample), and second, the top-down splitting in the tree learner is randomized. Instead of computing the locally optimal cut-point for each feature under consideration (based on, e.g., information gain or the Gini impurity), a random cut-point is selected. The value is selected from a uniform distribution within the feature's empirical range (in the tree's training set). Then, of all the randomly generated splits, the split that yields the highest score is chosen to split the node. Similar to ordinary random forests, the number of randomly selected

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 7, July 2021
DOI 10.17148/IJARCCE.2021.10733

features to be considered at each node can be specified. Default values for the parameter are {\displaystyle {\sqrt {p}}}{\sqrt {p}} for classification and {\displaystyle p}p for regression, where {\displaystyle p}p is the number of features in the model.

5.3 DECISION TREE

Decision trees are a popular method for various machine learning tasks. Tree learning "come[s] closest to meeting the requirements for serving as an off-the-shelf procedure for data mining", say Hastie et al., "because it is invariant under scaling and various other transformations of feature values, is robust to inclusion of irrelevant features, and produces inspectable models. However, they are seldom accurate".

In particular, trees that are grown very deep tend to learn highly irregular patterns: they overfit their training sets, i.e. have low bias, but very high variance. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. Here comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model.

Forests are like the pulling together of decision tree algorithm efforts. Taking the teamwork of many trees thus improving the performance of a single random tree. Though not quite similar, forests give the effects of a K-fold cross validation.

VI. CONCLUSION

To propose a machine learning model to categorize E-Commerce consumers, identify their sales characteristics and people are a trend to shop their daily needs in e-commerce sites and here the product recommendation takes a major role in every e-commerce sites to overcome their failures. It is one kind of marketing process by which we can advertise for many products and make the customers feel comfort while purchasing into the sites. Product recommendation will analysis the existing things where it can find the frequently purchased products which the customer like most and wish to buy will be recommended for them, it increases the sale percentage. To highly focus on effective product recommendation system using Linear Regression, Random Forest and Decision Tree method.

VII. ACKNOWLEDGEMENT

The recent overview report says that greater part of business client is giving their phony survey among the great items. It may prompt negative shade among the item and it has been immense downside among the administration side Furthermore, contributing tremendous sum and they confronting substantial misfortune in market. To defeat the issue, the proposed framework, the client will be able to give review only after purchasing the product. It is primarily to dodge the phony clients among the items.

REFERENCES

[1]. Han, J. Pei, J. and Kamber, M. (2011) 'Data Mining: Concepts and Techniques'.
[2]. Djenouri, Y. Belhadi, A. Fournier Viger, P. and Fujita, H. (2018) 'Mining diversied association rules in big datasets: A cluster/GPU/genetic approach', Inf. Sci., vol. 459, pp. 117134.
[3]. K.Gouda, M. Hassaan, and M. J Zaki (2010) 'Prism: An effective approach for frequent sequence mining via prime-block encoding', J. Computer Syst. Sci., vol. 76, no. 1, pp. 88102.
[4]. V. Codocedo, G. Bosc, M. Kaytoue, F. Boulicaut, and A. Napoli (2017) 'A proposition for sequence mining using pattern structures', in Proc. Int. Conf. Formal Concept Anal., pp. 106121.
[5]. B. Vo, T. Le, T.-P. Hong, and B. Le (2014) 'An effective approach for maintenance of pre-large-based frequent-itemset lattice in incremental mining', Appl. Intell., vol. 41, no. 3, pp. 759775.
[6]. E. Ya, A. S. Al-Hegami, M. A. Alam, and R. Biswas (2012)'Yami: Incremental mining of interesting association patterns', Int. Arab J. Inf. Technol., vol. 9, no. 6, pp. 504510.
[7]. L. Duan andW. N. Street (2016) 'Speeding up maximal fully-correlated itemsets search in large databases', Int. J. Mach. Learn. Cybern., vol. 7, no. 5, pp. 741751.
[8]. J. C.-W. Lin, S. Ren, and P. Fournier-Viger (2018) 'Memu: More efficient algorithm to mine high average-utility patterns with multiple minimum average-utility thresholds', IEEE Access, vol. 6, pp. 75937609.
[9]. U.Y. Bhatt and P.A. Patel (2015)'Mining interesting rare items with maximum constraint model based on tree structure', in Proc. 5th Int. Conf. Commun. Syst. Netw. Technol.,pp. 10651070.
[10]. P.Qiu, L.Zhao, and X.Dong (2017) 'NegI-NSP: Negative sequential pattern mining based on loose constraints', in Proc. 43rd Annu. Conf. IEEE Ind. Electron. Soc., pp. 34193425.