International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 3, April 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Review Analysis of Products and Recommendation System

Maria Ann Toms[1], Manu P S[2], Mohammed Ashique[3], Ms. Sajitha I[4]

[4]Assistant Professor

[1,2,3,4]Computer Science Department, Jyothi Engineering College, Thrissur, Kerala, India

ABSTRACT

In this paper, we first classify the text reviews given by different users on different products. There will be a wide variety of reviews about different products in the market. Using the machine learning techniques, we can analyze this data and use the different classifiers on them to get the behavior of the reviews. Later we are performing a collaborative approach to find out the possible list of products a user tend to buy and also the potential customers who are more likely to buy a particular product.. For more expertise knowledge about the product and for its clear understanding, the most discussed features and the specifications of the product is also highlighted.

KEYWORDS: *Machine Learning; Data Analysis; Collaborative Filtering; Euclidean Distance; Pearson Score*

How to cite this paper: Maria Ann Toms | Manu P S | Mohammed Ashique | Ms. Sajitha I "Review Analysis of Products and Recommendation System" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-3, April 2020, pp.1164-1167, URL: www.ijtsrd.com/papers/ijtsrd30771.pdf

IJTSRD30771

1. INTRODUCTION

In the current scenario of digitalization and development in the business fields, people appear to be more comfortable with the online shopping rather than the conventional one. There is a very large scale of online shopping happening which is widely accepted and has become a very popular trend in the present. It is also interesting to note that the customers who purchase products online post their reviews in the corresponding sites so that the other customers can view them and understand more from a customer point of view. The customers can give ratings also for the product on the scale for which the product could satisfy the requirements and expectations of the customer. Since a huge number of products are sold online and a very large number of people buy the products online and give the reviews and their opinion regarding the product, a huge amount of data is generated on a daily basis in this background. It is possible to get this vast amount of data in terms of the reviews and ratings given to different products by different customers from the online shopping sites. The famous shopping sites like amazon, flipkart etc produces a very big quantity of data in this regard.

Now what can be done with this obtained data in the forms of reviews and ratings? There could be a very interesting answer to this question. It is understood that the new customers who wants to buy a new product studies the different reviews and ratings the other customers had given to this product before he or she gets it. In fact that is a very wise option followed by the people who do online shopping. Another interesting thing to be noticed is that there could be different frequent patterns occurring in the purchase of the products between different customers. This pattern or association found can be exploited and be used to recommend the products to the possible list of customers and in the same way around a list of potential products could be given to a customer. There are various Machine Learning techniques used for the analysis of data and classifying them and finally make predictions based on the same. This paper hence focuses on the following things: To classify the reviews of the different products using the different machine learning algorithms, Recommend a list of possible products that could be bought by the customer which is implemented using the Euclidean Distance, generating a list of potential products that the customer would tend to purchase and also identifying the most discussed features of the products. From this work it would help the customer to be smarter and make his or her choices of purchasing the products online wise and easier.

2. OVERVIEW OF BASICS

2.1. Data Analysis

In the present scenario there is data generated everywhere now and then. All of ours daily routines pays way to loads of data being accumulated. This could be observed in many different realms of the day to day living. Data Analysis is the process of cleaning, transforming, modeling and interpreting data into useful forms. From the huge amount of data used for analysis first has to undergo the process of cleaning,

further to the data integration to combine multiple data sources if required, then to the selection process and hence obtaining the pre-processed data[1]. This is then exploited and useful information is discovered from the same that could be applied in different fields. Data can be transformed and consolidated into various forms. Different intelligent methods can be used on the pre-processed data to extract data patterns from it. The extracted patterns can be used to evaluate things and derive a conclusion from it. There are also so many methods to represent the data making the visualization of the extracted knowledge possible[1]. Data Analysis is used in different areas like Market Analysis, Business Intelligence, Web Mining, etc. Data Analysis tools makes it easier for the users to process and manipulate data, analyze the relationships and correlations between the datasets and it also helps to identify the patterns and trends for interpretations. Data Analysis tools include Matlab, python, R etc. In this paper work we have done the data analysis using python language. There are lots of built in libraries in python that can be used at once on importing them rather than coding from the scratch.

2.2. Machine Learning Algorithms

2.2.1. Linear Regression

Linear Regression focuses on modeling the relationship between two variables by fitting into a linear equation to the observed data. One variable is said to be the explanatory variable and the other variable is called the dependent variable[2]. A scatter plot of the data used can be a helpful tool in determining the strength of the relationship between the variables. A valuable numerical value of the association between the two variables is the correlation coefficient which is a value between -1and 1 indicating the strength of association between the variables in the observed data.

2.2.2. Logistic Regression

Logistic Regression is a statistical method used for the analysis of a dataset in which there are one or more independent variables that determine an outcome[2]. This is an appropriate regression used when the dependent variable is dichotomous or binary. Logistic Regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal or ratio level independent variables. This is used in Predictive Analysis

2.2.3. KNN

KNN is K- Nearest Neighbors is a Machine Learning algorithm that belongs to the supervised learning domain and finds application in pattern recognition, recommendation system etc. This is carried out by comparing the given test data with the training data similar to it. KNN Classifier searches the pattern for 'K' training data that are closest to the unknown test data[2]. The distance or closeness between the test data and that of the training data is calculated using different measures like the Euclidean Distance, Manhattan Distance etc.

2.2.4. Support Vector Machine

Support Vector Machine is a discriminative classifier formally defined by a separating hyper plane. Here the given labeled training data uses the algorithm to give the optimal hyper plane which can classify new data[4]. An SVM model is the representation of data as points in space mapped so that the examples of the separate categories are divided by a clear gap that is wide as possible. In addition to this SVM's can efficiently perform a non-linear classification, implicitly mapping their inputs into high dimensional feature space.

2.2.5. Random Forest Classifier

Random Forest Classifier creates a set of decision trees from randomly selected subset the given training set. It the aggregates the votes from the different decision trees to decide the final class of the test data[3]. The reason for the random forest classifier model is that, a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

2.3. Collaborative Filtering

Collaborative Filtering comes from the idea that best recommendations happen from people with similar tastes. It uses historical item ratings of like minded people and use it for the suggestion of the same. There are two categories of collaborative filtering approach[5]. One is the memory based collaborative approach and the second is model based collaborative approach. The main difference between both these approaches is that in the case of memory based approach, it uses all the data to generate a prediction but in the case of model based approach, we use the given dataset to train the mode which is capable of being reused to make the predictions. In this paper we have worked using the user based collaborative filtering based on the ratings given by the users to the products. In this work of recommending product to users of similar taste using collaborative approach, we have used two methods: The Euclidean Distance Score and the Pearson Correlation Score. The final results for both of these methods were found to be almost similar. After finding the similarity between the users, it uses the weighted average method to assign a higher weight to the peer interest critics[5]. Finally it normalizes the score by dividing it by the similarities of the users who rated the product. Hence this turns out to be an appropriate prediction based on the interest of a particular user by collecting the taste and preferences from the other users in the system. For this work we have used the dataset from amazon site which is perfect for the objective of suggesting a product to a target user[6]. The dataset of amazon contains data about object like the brand name, image url etc and also the data about what the users think about the products in the form of ratings, textual reviews and the other products seen in the same user session. The major advantage of this work is that, based on the suggestions given for a user it can help in increasing the outcome of the reseller.

3. LITERATURE SURVEY

From the insights from the reference papers that we referred, it is noted that the existing system experimented on the structured and unstructured data with techniques like the Binomial Logit and Decision tree methods. It focused on classifying the product reviews and scoring propensities using decision tree modeling. In another existing work, it was found that the Machine Learning algorithms like the Regression learning and Q-Learning was used to predict the obsolescence of the product[7]. Obsolescence indicates the lifespan for which a product under study will be viable to sustain a particular market condition. The viability depends on various parameters, product features etc[8]. In another paper referred it used the Functional Link Artificial Neural Networking and Non Linear Regression for identifying new product growth rate for the different industries using the product reviews[9]. The input to the model is the data of various companies on different parameters related to new

product growth rate. All the inputs except the bias signal were expanded by the trigonometric expansion terms to incorporate non linearity in the form of sine and cosine terms. Also the existing recommendation system only focuses on the prediction of the possible product that could be purchased by a particular customer.

4. CHALLENGES IN THE CURRENT SYSTEM

There are few drawbacks that the existing system had:

- The system concentrated more on the meta variables and was dependent on the limited parameter tuning. The classification of the text reviews of the product tend to be biased in such cases.
- The period of obsolescence calculation of the product was not that accurate since the evaluation of obsolescence without expert data was difficult.
- The main challenge for assessing the new product growth rate is to quantify the categorical variables since the economic environment is turbulent. This also led to biased prediction when it comes to classification.
- In the case of recommendation system, the existing system could only identify the possible list of customers who are likely to get the product.

5. PROPOSED SYSTEM

In the proposed system, we are including the following features:

- Perform the product reviews classification and get the sentiment score of the texts used in the reviews of the Machine Learning algorithms like Linear Regression, Logistic Regression, KNN, Support Vector Machine and Random Forest Classifier.
- Giving the list of the possible set of products that could be bought by a customer based on his or her past purchase history and using the collaborative filtering with the similar customers.
- Giving the set of potential customers who are likely to buy the product based on the reviews they have already given and by finding their frequent purchase pattern.
- Identifying the most discussed and reviewed features of the particular product so it helps in easier and more specific understanding regarding the product. This would also help the users to identify and spot that particular feature of the product that makes it good or bad.

5.1. Product Review Classification

5.1.1. Processing the data

On importing the different in built libraries in python like pandas, matplot library, seaborn and various libraries from sklearn it would the possible to perform the Product Review analysis of the different products using python. For the purpose of classification of reviews we use the datasets which given the reviews on different products like books, music, electronic gadgets, kitchen appliances and automotive products. First these are concatenated into a mixed pandas data frame and saved. This mixed dataset contains multiple review categories. These are then to be loaded and the text body of reviews is converted to string and subject to the text pre-processing part. Later the body of the reviews is tokenized and the length of the review is calculated. The product reviews can be grouped by using the star rating which is also available in the dataset so that we get the total number of reviews which had awarded a particular star rating for the product. Now for the analysis of the product reviews we have to use the bag of words algorithm and the TfidVectorizer. Since the body of the product review is to be analyzed, the bag of algorithm is used on the same which results in the formation of the text data being turned into literally a bag of words where, it works like creating a bucket for each unique word that you want to represent. Next go over the text and put a token in the right buckets for the words encountered. The TfidVectorizer, transforms the text to feature vectors that can be used as input to the estimator. There is a dictionary that converts every token (word) to feature index in the matrix, each unique token gets a feature index. Now it is possible to get the pictorial representation of the processed data.

5.1.2. Building Models for Prediction

Now since the data to be analyzed had undergone the process of pre-processing and transformation it is now ready to be used to train the different models using the Machine Learning Algorithms. The whole data is split into two: the training data and the testing data. It is using the training data that we are able to train the different models so that it can come to a conclusion for the unknown test data. Here we have used the Linear Regression to make the prediction and it had performed the prediction of the test data with an accuracy of 58.3%. The Logistic Regression is performed on the same dataset and by making use of other independent variables like class weight, solver, multi-class, and random state. It is found that the prediction using the Logistic Regression turned out with an accuracy of 63.9%.

Sl no	Precision	Recall	F1-Score	Support
1	0.59	0.48	0.53	50
2	0.08	0.06	0.07	18
3	0.29	0.24	0.26	29
4	0.18	0.23	0.20	66
5	0.80	0.81	0.81	333
Accuracy			0.64	496

Fig 1: Classification Report using Logistic Regression

When the KNN Algorithm was used in the prediction of the test data it gave a prediction of 58%. The classification report using the KNN model is given by:

Sl no	Precision	Recall	F1-Score	Support
1	0.69	0.18	0.29	50
2	0.20	0.06	0.09	18
3	0.25	0.03	0.06	29
4	0.18	0.30	0.21	66
5	0.74	0.77	0.75	333
Accuracy			0.58	496

Fig 2: Classification Report using KNN

When using the support vector machine model for the prediction it came out with an accuracy of 67%. The classification report using the Support Vector Machine is given by:

Sl no	Precision	Recall	F1-Score	Support
1	0.67	0.20	0.31	50
2	0.00	0.00	0.00	18
3	0.00	0.00	0.00	29
4	0.33	0.06	0.10	66
5	0.69	0.97	0.81	333
Accuracy			0.68	496

Fig 3: Classification Report using SVM

When using the Random Forest Classifier model for the prediction it came out with an accuracy of 72.7%. The classification report using the Random Forest Classifier is given by:

Sl no	Precision	Recall	F1-Score	Support
1	1.00	0.04	0.08	50
2	0.00	0.00	0.00	18
3	0.00	0.00	0.00	29
4	0.00	0.00	0.00	66
5	0.67	0.99	0.80	333
Accuracy			0.72	496

Fig 4: Classification Report using Random Forest Classifier

Now that we have used different classifiers and trained different models using the training dataset and performed predictions on the test data we can now perform the sentiment analysis of the product reviews and classify the reviews as good or bad. We obtain the same using the sentiment analysis and generate the good or bad vectors for the reviews of the product. It is also possible to get the sentiment score of the desired text which have been used in the product reviews given by the customers.

5.2. Recommendation System

For building the recommendation system we are making use of the dataset from amazon. This dataset contains data about the product like the brand, specifications etc and the data regarding the opinion of the different products by different users. The users given the reviews of the product they had purchased and also a star rating for the product based on if the product had come upto their expectation level. The features selected for the process are username, rating reviews title (tag). It is possible to group by the reviews using the ratings given by the customers with respect to the unique user ID's. If the text reviews have same words used, we can get the total number of count a particular word vector had appeared in the reviews and plot them. We can get the genre count from the dataset used given the product reviews of the products purchased online. For the recommendation purpose, we need to find the similarity between different users and this is implemented using the Euclidean Distance Measurement. The distance is calculated using formula 1 divided by 1 plus sum of the squares between the two points. The inter personal correlation is obtained by calculating the Pearson Score between two person. This is basically done by checking if they have any rating in common and also if they have bought any products in common. Finally summing up the squares of rating and sum of the products bought, the Pearson Score is calculated. We can get the top similarities between the combinations of different people so that the recommendation becomes perfect.

6. CONCLUSION

For the product review classification, out of the several classifiers used for the prediction, we came to the conclusion that the Random Forest Classifier produces the prediction as its output with an accuracy of 72.7%.

In the building of the recommendation system, both the measures used the Euclidean Distance as well as the Pearson Score gave a similar output as 1 which implies that the inputs given were closely similar to each other regarding the purchasing patterns. On implementing these two measures, we are likely to get an output ranging from 0 to 1. A value closer to 1 implies more similarity. It is possible to get the best possible matches for a person from the preferred dictionary.

References:

[1] Implisit (Sales Cloud by Salesforce.com). [Online]. Available:https://www.salesforce.com/blog/2014/08/infographic-7-powerfulpredictors-closed-won-opportunity-gp.html

[2] J. Yan, C. Zhang, H. Zha, et all, "On Machine Learning towards Predictive Sales Pipeline Analytics." Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, pp. 1945-1951, 2015.

[3] M. Bohaneca, M.K. Borstnarb, M Robnik-Sikonja, "Explaining machine learning models in sales predictions." Expert Systems with Applications, no. 71, pp. 416-428, 2017.

[4] S. Gopinath, P. K. Chintagunta, and S. Venkataraman, "Blogs, Advertising, and Local-Market Movie Box Office Performance," Management Science, vol. 59, no. 12, pp. 2635–2654, 2013..

[5] M. C. A. Mesty´an, T. Yasseri, and J. Kert´esz, "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data," PLoS ONE, vol. 8, no. 8, 2013.

[6] A. Chen, "Forecasting gross revenues at the movie box office," Working paper, University of Washington, Seattle, WA, June, 2002.

[7] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state of-the-art and possible extensions. Knowledge and Data Engineering, 2005

[8] R. Burke. Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction, 2002.

[9] T. Chai and R. R. Draxler. Root mean square error (rmse) or mean absolute error (mae)? Geoscientific model development, 2014..

[10] J. Duan, X. Ding, and T. Liu, "A Gaussian Copula Regression Model for Movie Box-office Revenue Prediction with Social Media," Communications in Computer and Information Science Social Media Processing, pp. 28–37, 2015.