Machine Learning Methods for Predicting the Popularity of Forthcoming Objects

Gulab Sah, Rajat Subhra Goswami, Sunit Kumar Nandi

Abstract: ***Now a day, product ratings are very much essential for the product available online so that customers can view a product's actual rating before they are going to buy it. This is only the primary source of information for a product, and it is also essential for manufacturers, retailers to improve product quality in terms of production and sale.A rating can make it easy for consumers to figure out how much they enjoy the product. Now in case of new arrival products which have not been used by any customers or any users, the ratings not available online. We have tried to find ratings for new arrival products in this research work by identifying the quality of that product, which will assist customers before buying it. We have also examined different method that will predict the rating of the newest arrival product based on product features, description, information that are available on the e-commerce platform like Amazon, Flipchart. To achieve the defined goal, we have worked on existing data that are available for products already arrived in the market and already used by a customer. The main objective of this research is to help the customer who is going to purchase new arrival products. This is done by comparing different existing Machine Learning methods with the help of the existing data set. We have tried to find out the best method among the existing Machine learning methods and applied that method to predict the rating of the newest arrival product based on the available features.***

Keywords—Product rating, Amazon, classifiers, Support Vector Classifier, K-Nearest Neighbors , Naive Bayes classifier , Random forest Classifier , Neural network, Decision tree, Multinomial logistic regression ,Confusion Matrix.

I. INTRODUCTION

Consumer facing difficulty in making a purchase decision and also consumer often seeks others' opinion about a product. They search rating of product in internet such as amazon.com, read magazines such as consumer report. But it is difficult to find rating of newest arrival product for consumers because the newest arrival products is neither used by customer nor rated by customers. So, this create doubt in consumer mind that product is good or not. Consumers consistently check out multiple online rating and observe the difference. To learn the important of the variance of ratings, a meaningful beginning would be to consider the imbalance of ratings. A major information source for consumer is online product rating which tells the consumer about the product quality that is product is good or not. The rating of upcoming product is not available for consumer which makes hard to take decision about the product in this condition the feature of product will help consumers in decision making.

Revised Manuscript Received on December 15, 2019.

Gulab Sah, Research Scholar, Department of CSE, NIT Arunachal Pradesh, India Email:gulabsah15@gmail.com

Rajat Subhra Goswami, Assistant Professor, Department of Computer Science & Engineering, National Institute of Technology, Arunachal Pradesh, India.

Sunit Kumar Nandi, Department of computer Science & Engineering, National Institute of Technology, Arunachal Pradesh, India.

Our finding provide rating forconsumer and they should keep in mind that a product
Will going to be high rating or low rating or average rating which help them in decision making about the products. To achieve our goal first we work on existing data that is ratings are available for products that are already arrived on market and used by customer. The main objective of this paper is to compare different method and find best method based on existing data.

II. LITERATURE REVIEW

In this section, we survey the related research in rating, price and review of product. We mainly discuss the works which utilize both reviews and ratings for rating prediction.
Product rating matter: In this paper, they check-up the unofficial important of the generalization of product rating by focus on the variation of rating and for a product with a lowest average rating, a highest variation of rating communicates to feasible buyers that well matched consumer would like the product ,which in turn increases demand[1]. Performance measurement of classification algorithm like Naive Bayes and J48: In this paper, they implement classification using these two algorithms on bank data-train.arff dataset in weka tool. They used these two algorithms for comparing. Differentiation is made usually on accuracy and they use confusion matrix to illustrate accuracy of the algorithms to solve a classification problem [2]. Probabilistic Machine Learning for Sentiment Analysis: in this paper they used an Amazon and other e-commerce website to extract a data, then they done feature extraction, extracting a review and performed the classification using Naive Bayes and decision list. At last, comparing the result of algorithms [3]. Understanding online product ratings: In this paper, shown that explanation of score of ratings than orthodox quality-centered explanation for describing customer feast models of online products ratings is best suited and alsoan online product rating is very important because it is major information source for manufacture, retailer and customers [4]. Implicit Rating and Filtering: according to this paper, some evidence available informs that inherent ratings have great feasible but their impact remains unproven. As with many technologies inherent rating may first be joined with existing rating systems to form a mixture system. One method is to use inherent data as a cover on explicit ratings; example if an inspector is explicitly rating an object then there should be some corresponds inherent data to sure that she/he has actually check it. If there is no evidence to inform that the inspector has checked an object then perhaps their rating should be neglect, or minimize in importance.According to this paper technology like machine learning, artificial intelligence play important role in classification
of data [5].

Random Forest classifier builds a forest of random trees. For training this
Model, the bagsizepercent parameter is set to100 [6].Random Treeclassifier constructs a tree by considering K attributes chosen randomly for every node without performing any pruning.The minimum total weight of the instances in each leaf is set to 1 and the minVarianceProp parameter is set to 0.001,that gives the minimum portion of the variance on splitting data [7]. The PART classifier generates a decision list using separate-and-conquer. In every, iteration a C4.5 decision tree is built and the leaf with "best performance" is used to define the rule. The confidence factor parameter, for training this classifier on Weka, is set to 0.25 and the number of folds used is 3[8].NaïveBayesclassifier uses estimator classes. After analyzing the training data, the values of numeric estimator precision values are determined. The parameter kernelEstimator for numeric attributes is set to False for this classifier [9].Multi-Layer Perceptronclassifier uses back propagation to train a multilayer perceptron for classification. This network can be hand coded as well but Weka's tailored GUI helped us to create the network by specifying the parameters.ThebatchSize is used to train this classifier as 100, with learning rate tuned at 0.3, momentum of 0.2 and a validation Threshold of 20[10].
Logistic classifier builds a model having a ridge estimator. In order to train the model on the software; we set the ridge parameter to log likelihood of 10^{-8}. The classifier capabilities are checked before the classifier was built [11]. Decision Table classifier builds a simple decision table which makes the decision using maximum voting technique. We set the parameters batchSize to a 100 and the search method to 'Best First Search' method [12].Decision Stump classifier is trained after tuning the parameter batchSize set as 100[13].J48 classifier creates C4.5 decision trees. The c-factor is set to 0.25, with 2 instances per leaf and numFold to 3[14].REP Tree Abbreviation for Reduced Error Pruning Tree, classifier is considered to be a quick learner. It builds a decision tree using information gain.TheminVarianceDropis the same as that of the previous classifier [15].

III. METHODLOGY

A. Dataset

For the purpose of the research, Amazon.com appears to be a good source to collect data. We know Amazon.com is a most popular online shopping website and it is the largest e-commerce platform, which sells a variety of consumer product. We created a panel of dataset of product belonging to Smartphone product category. We built a crawler or web spider that visit web page to extract product reviews, product information, product features etc. The information related to specific product (example Smartphone) get saved in a binary file that is then convert into a list of tuples and attributes consisting of the product features, product information, product name, product price etc.

B. Data pre-processing

Several data preprocessing steps are performed on the dataset before it used.

- Change to lower case: we first transform our texts into lower case. This avoids having multiple copies of the same words.
- Removing punctuation: it does not enlarge any needless information while accept text data.
- Removal of duplicate product: we do this because we need identical products.
- Removal of stop words: common words should be deleting from the text data.
- Spelling correction: we do spelling correction for minimize duplicate words.
- Rare words removal: we delete rarely present words from the text because they are so less; the compatibility between them and other words is dominated by detonation.
- Common word removal: We can delete repeated words from our text data.
- Similarity and dissimilarity: to fit our dataset into various machine learning techniques we need only integer values so in order to get integer values we first find similarity between texts of each product let say x. second we find dissimilarity that is 1-x.
- We add additional column "rating class" to our dataset. Based on this rating class we classified our data.

C. Data Classifier

In this paper, k-nearest neighbors algorithm, naive bayesalgorithm,decision tree algorithm,random forest algorithm, Multinomial logistic regression algorithm and neural network algorithm,support vector classifier, are use for comparison., is done on accuracy, precision recall and F1-score using correct prediction and incorrect prediction in confusion matrix produce by the particular algorithms..

1) Support vector Classifier (SVC): SVC is fit to data to enable, returning "best fit" hyperplanesthat categorizes or divides our data. To look what the "predicted" class is, we feed some features to our classifier after getting hyper planes. The implementation is based on libsvm. The multiclass support is handled according to a one vs. one scheme.

2) K-Nearest Neighbors: K-Nearest Neighbors can be used for both classification and Regression predictive problems. Here we used KNN for classification. KNN is a easiest algorithm that learn data and classifies new state based on a similarity calculated (e.g., distance function).

3) Naive Bayes classifier:Naive Bayes classifier is a collectionof classification algorithms based on Bayes' Theorem.It is not a single algorithm but a family of algorithms where allof them share a common principle, i.e. every pair of features

4) Decision tree: Decision Tree Classifier is a easiest and widely used classification algorithms. It applies a clear idea to solve the classification problem. This Classifier pretence a series of sharply crafted questions about the property of the test record. Every time it accepts an answer, a follow-up question is asked until a result about the class state of the record is obtained.

5) Random forest Classifier (RFC): RFC work as Meta predictor that fits number of decision tree classifier (it may be one or more) on many sub-part samples of dataset and utilize average to raise predictive (estimated) accuracy and alsocontrol over-fitting. The sub-part of sample size is always similar as the prime input sample size but samples are warped with substitute if

bootstrap is equal to True (by default bootstrap is true).

6) Multinomial logistic regression: Multinomial logistic regression is used to model suggestive resulted variables, in which the log odds of the resulted are develop as a linear organization of the predictor variables. This classifier is used to estimate a suggestive subject variable produce one or more self-sufficient variables. It periodically known as an extended version of binomial logistic regression to permit for a subject t variable with many groups where as other types of regression, it can suggestive and/or continuous self-sufficient variables and provide interaction between self-sufficient (independents) variables to estimate the subject variable. Here estimate means predict.

7) Neural network: Neural network provide multi-label classification in which a pattern can belong to more than one group (also called class). For each group, the incomplete output routing upon the logistic function. Values largest or equal to 0.5 are taken approximate as 1, otherwise to 0. For a estimated output of a pattern, the indices where the value is 1shows the assigned classes of that pattern. The Fig.1 shows the block diagram of the proposed work.

D. Model

F

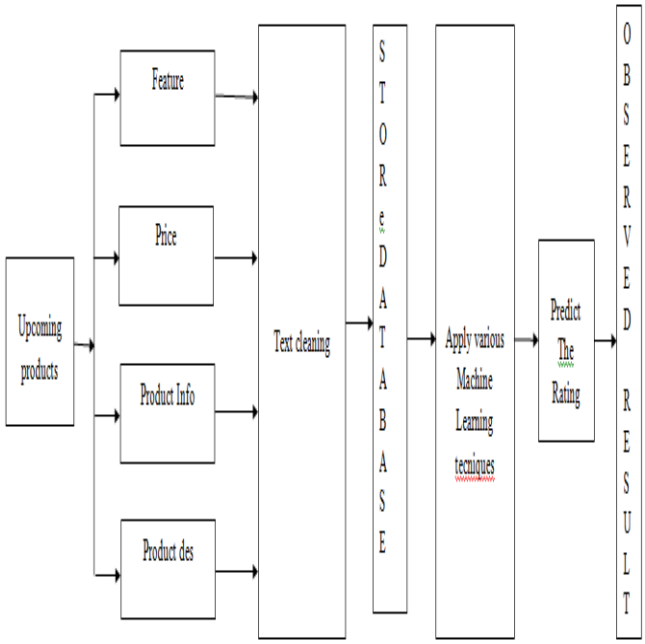

Fig.1. Block diagram

IV. MEASURING PERFORMANCE

Based on accuracy of the classification algorithm, the performance of algorithm is measured so that we can find the best algorithm.

Classification Accuracy: Classification Accuracy is the ratio of number of correct prediction to the total number of input samples.

$$\text{Accuracy}=\frac{Numberofcorrectpredictions}{Totalnumberofpredictionsmade}$$

Confusion Matrix:

- **Accuracy:** accuracy for the matrix can be calculated by

$$\text{Accuracy}=\frac{TruePositives\ +FalseNegatives}{TotalNumberofsamples}$$

- **Precision:** It is the number of correct positive result divided by the number of positive result predicted by the classifier.

$$\text{Precision}=\frac{TruePositives}{TruePositives\ +FalsePositives}$$

- **Recall:** It is the number of correct positive result divided by the number of all relevant samples(all samples that should have been identified as positive)

$$\text{Recall}=\frac{TruePositives}{TruePositives\ +FalseNegatives}$$

- **F1 Score:** F1 score for the matrix can be calculated by

$$\text{F1 score}=\frac{2*Recall\ *Precision}{Recall\ +Precision}$$

- **Micro:**Measure metrics generally by tally the overall true positives (correct prediction), false positive and false negatives

$$\text{Precision}_{micro}\text{Avg}=\frac{TP1+\cdots+TPK}{TP1+\cdots+TPK+FP1+\cdots+FPK}$$

$$\text{Recall}_{micro}\text{Avg}=\frac{TP1+\cdots+TPK}{TP1+\cdots+TPK+FN1+\cdots+FNK}$$

$$\text{F1-score}_{micro}\text{Avg}=\frac{2*Precision_{\ micro\ Avg}*Recall_{\ micro\ Avg}}{Precision_{\ micro\ Avg}+Recall_{\ micro\ Avg}}$$

- **Macro:**Measurethe metrics for each step and search their unweighted mean. This does not take stepvariance into account.

$$\text{Precision}_{macro}\text{Avg}=\frac{precision\ 1+\cdots+pricisionK}{K}$$

$$\text{Recall}_{macro}\text{Avg}=\frac{Recall\ 1+\cdots+RecallK}{K}$$

$$\text{F1-score}_{macro}\text{Avg}=\frac{2*Precision_{macro}\ Avg*Recall_{macro}\ Avg}{Precision_{macro}\ Avg+Recall_{macro}\ Avg}$$

- **Weighted:**Measure metrics for each step and identify their average weighted by support (the number of true object for each step). This update 'macro' to account for step variance; it can outcome in F-score that notamongPrecision and Recall.

$$\text{F1score}_{weighted}\text{Avg}=\frac{2*Weighted_{Precision}\ Avg*Weighted_{Recall}\ Avg}{Weighted_{Precision}\ Avg+Weighted_{Recall}\ Avg}$$

V. EXPERIMENTAL WORK AND RESULTS

Table 1:support vector classifier result

Support vector classification Report					
	classes	precision	recall	f1-score	support
	eight	0.98	1.00	0.99	65
	Five	0.73	0.67	0.70	12
	Four	0.67	0.33	0.44	6
	Nine	0.97	1.00	0.98	29
	Seven	1.00	0.99	0.99	82
	Six	0.93	1.00	0.97	43
	Ten	1.00	0.96	0.98	26
	Three	0.50	1.00	0.67	1
	two	1.00	1.00	1.00	8
Micro avg.		0.96	0.96	0.96	272
Macro avg.		0.86	0.88	0.86	272
Weighted avg.		0.96	0.96	0.96	272
Accuracy	0.9632				

Table 2: Results for classification using K-Nearest Neighbors:

K-Nearest Neighbors Classification Report					
	classes	precision	recall	f1-score	support
	eight	0.33	0.45	0.38	65
	Five	0.38	0.25	0.30	12
	Four	0.00	0.00	0.00	6
	Nine	0.24	0.21	0.22	29
	Seven	0.49	0.49	0.49	82
	Six	0.37	0.40	0.38	43
	Ten	0.53	0.38	0.44	26
	Three	0.00	0.00	0.00	1
	two	0.75	0.38	0.50	8
Micro avg.		0.40	0.40	0.40	272
Macro avg.		0.34	0.28	0.30	272
Weighted avg.		0.40	0.40	0.39	272
Accuracy	0.3970				

Table 3: Results for classification using Naive Bayes:

Naive Bayes classification Report					
	classes	precision	recall	f1-score	support
	eight	0.98	0.97	0.98	65
	Five	0.69	0.75	0.72	12
	Four	0.33	0.17	0.22	6
	Nine	0.97	0.97	0.97	29
	Seven	0.93	0.99	0.96	82
	Six	0.95	0.91	0.93	43
	Ten	0.96	0.96	0.96	26
	Three	0.00	0.00	0.00	1
	two	1.00	1.00	1.00	8
Micro avg.		0.93	0.93	0.93	272
Macro avg.		0.76	0.75	0.75	272
Weighted avg.		0.93	0.93	0.93	272
Accuracy	0.9338				

Table 4: Results for classification using Decision Tree (Entropy):

Decision Tree Classification Report using Entropy					
	classes	precision	recall	f1-score	support
	eight	1.00	1.00	1.00	65
	Five	0.44	1.00	0.62	12
	Four	0.00	0.00	0.00	6
	Nine	1.00	1.00	1.00	29
	Seven	1.00	1.00	1.00	82
	Six	1.00	1.00	1.00	43
	Ten	1.00	1.00	1.00	26
	Three	0.00	0.00	0.00	1
	two	0.00	0.00	0.00	8
Micro avg.		0.94	0.94	0.94	272
Macro avg.		0.60	0.67	0.62	272
Weighted avg.		0.92	0.94	0.93	272
Accuracy	0.944				

Table 5: Results for classification using Random Forest:

Random forest Classification Report					
	classes	precision	recall	f1-score	support
	0	1.00	1.00	1.00	65
	1	0.83	0.83	0.83	12
	2	0.60	0.50	0.55	6
	3	1.00	1.00	1.00	29
	4	0.99	1.00	0.99	82
	5	0.93	0.98	0.95	43
	6	1.00	1.00	1.00	26
	7	0.00	0.00	0.00	1
	8	0.86	0.75	0.80	8
Micro avg.		0.97	0.97	0.97	272
Macro avg.		0.80	0.78	0.79	272
Weighted avg.		0.96	0.97	0.96	272
Accuracy	0.9669				

Table 6: Results for classification using Multinomial Logistic Regression

Multinomial logistic regression Classification Report					
	classes	precision	recall	f1-score	support
	0	0.78	1.00	0.88	65
	1	0.00	0.00	0.00	12
	2	0.00	0.00	0.00	6
	3	0.92	0.41	0.57	29
	4	0.98	0.96	0.97	82
	5	0.73	0.95	0.83	43
	6	0.93	0.96	0.94	26
	7	0.00	0.00	0.00	1
	8	0.80	1.00	0.89	8
Micro avg.		0.85	0.85	0.85	272
Macro avg.		0.57	0.59	0.56	272
Weighted avg.		0.81	0.85	0.81	272
Accuracy	0.8455				

Table 7: Results for classification using neural network

Neural network classification report					
	classes	precision	recall	f1-score	support
	0	1.00	1.00	1.00	65
	1	0.73	0.92	0.81	12
	2	1.00	0.33	0.50	6
	3	0.97	1.00	0.98	29
	4	0.99	1.00	0.99	82
	5	0.98	0.98	0.98	43
	6	1.00	0.96	0.98	26
	7	0.00	0.00	0.00	1
	8	0.89	1.00	0.94	8
Micro avg.		0.97	0.97	0.97	272
Macro avg.		0.84	0.80	0.80	272
Weighted avg.		0.97	0.97	0.97	272
Accuracy	0.9705				

Table 8: Overall comparison of precision of different method

SL No	Method	Precision		
		Macro Avg.	Weighted Avg.	Micro Avg.
1	Support vector classification	96%	86%	96%
2	K-Nearest Neighbors	40%	34%	40%
3	Naive Bayes	93%	76%	93%
4	Decision Tree using Gini Index	94%	60%	92%
5	Decision Tree using Entropy	94%	60%	92%
6	Random forest	97%	80%	96%
7	Multinomial logistic regression	85%	57%	81%
8	Neural network	97%	84%	97%

Table 9: Overall comparison of recall of different method

SL No	Method	Recall		
		Macro Avg.	Weighted Avg.	Micro Avg.
1	Support vector classification	96%	88%	96%
2	K-Nearest Neighbors	40%	28%	40%
3	Naive Bayes	93%	75%	93%
4	Decision Tree using Gini Index	94%	67%	94%
5	Decision Tree using Entropy	94%	67%	94%
6	Random forest	97%	78%	97%
7	Multinomial logistic regression	85%	59%	85%
8	Neural network	97%	80%	97%

Table 10: Overall comparison of F1-score of different method

SL No	Method	F1-score		
		Macro Avg.	Weighted Avg.	
1	Support vector classification	96%	86%	96%
2	K-Nearest Neighbors	40%	30%	39%
3	Naive Bayes	93%	75%	93%
4	Decision Tree using Gini Index	94%	62%	93%
5	Decision Tree using Entropy	94%	62%	93%
6	Random forest	97%	79%	96%
7	Multinomial logistic regression	85%	56%	81%
8	Neural network	97%	80%	97%

Table 11: Results for classification using Decision Tree(Gini-index):

Decision Tree Classification Report using Gini Index					
	classes	precision	recall	f1-score	support
	eight	1.00	1.00	1.00	65
	Five	0.44	1.00	0.62	12
	Four	0.00	0.00	0.00	6
	Nine	1.00	1.00	1.00	29
	Seven	1.00	1.00	1.00	82
	Six	1.00	1.00	1.00	43
	Ten	1.00	1.00	1.00	26
	Three	0.00	0.00	0.00	1
	two	0.00	0.00	0.00	8
Micro avg.		0.94	0.94	0.94	272
Macro avg.		0.60	0.67	0.62	272
Weighted avg.		0.92	0.94	0.93	272
Accuracy	0.9448				

Table 12: Overall comparisons of accuracy of different method

SL NO.	Method	Accuracy
1	Support vector classification (SVC)	96.32%
2	K-Nearest Neighbors (KNN)	39.70%
3	Naive Bayes	93.38%
4	Decision Tree C using Gini Index	94.48%
5	Decision Tree using Entropy	94.4%
6	Random forest	96.69%
7	Multinomial logistic regression	84.55%
8	Neural network	97.05%

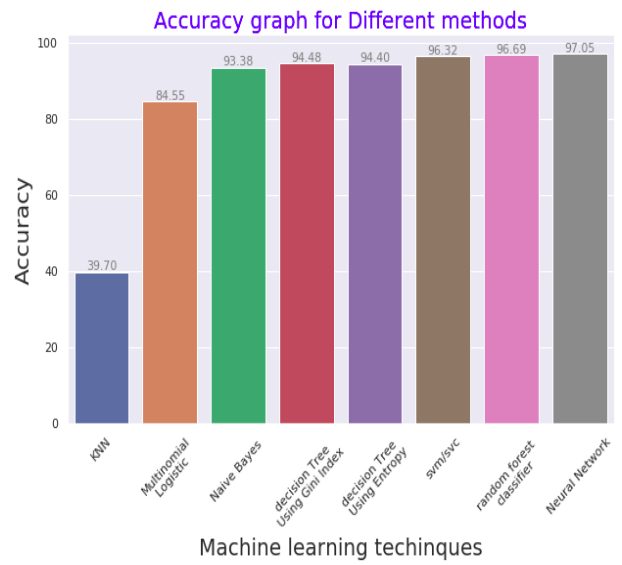

Fig.2. Accuracy graph for different methods

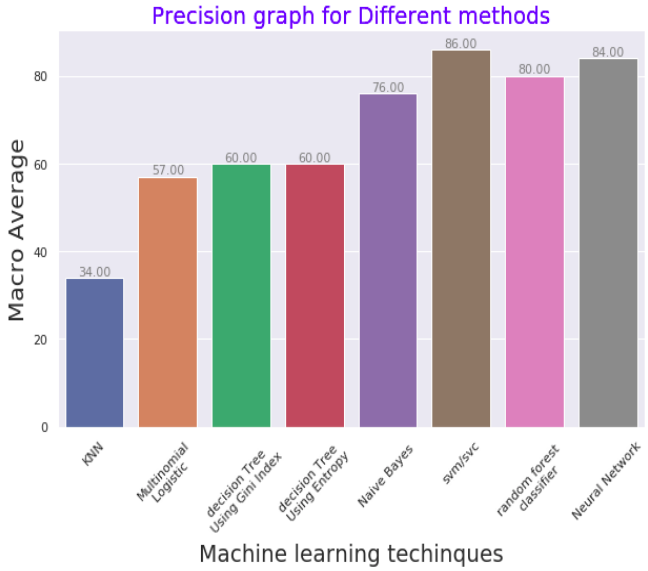

Fig.3.Precision in terms of Macro average

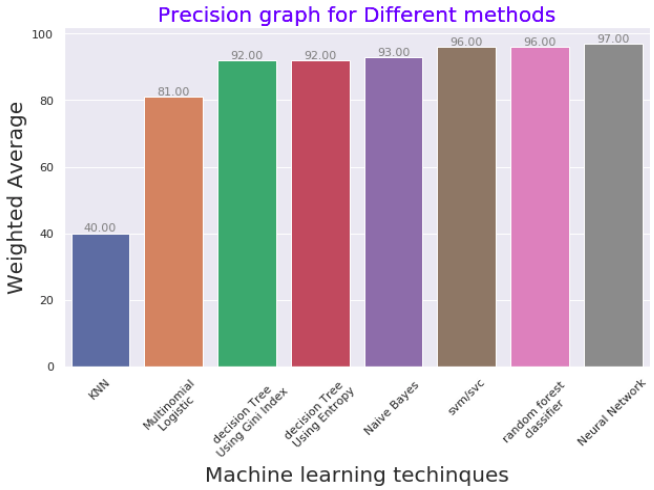

Fig.4. Precision in terms of weighted average

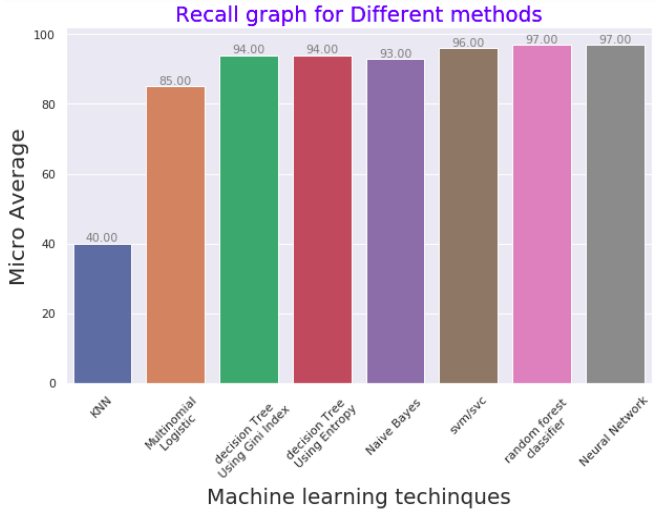

Fig.5. Precision in terms of Micro average

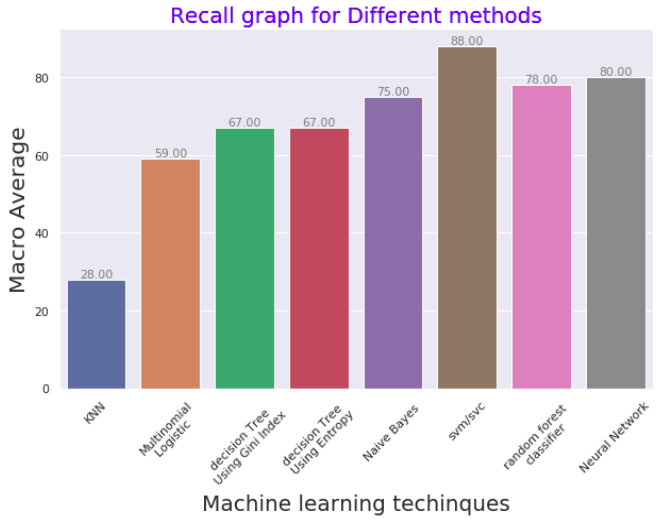

Fig.6. Recall in terms of Macro average

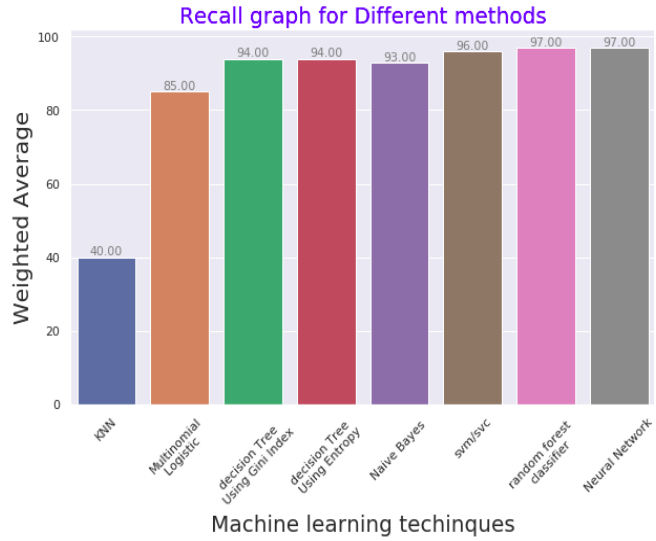

Fig.7. Recall in terms of weighted average

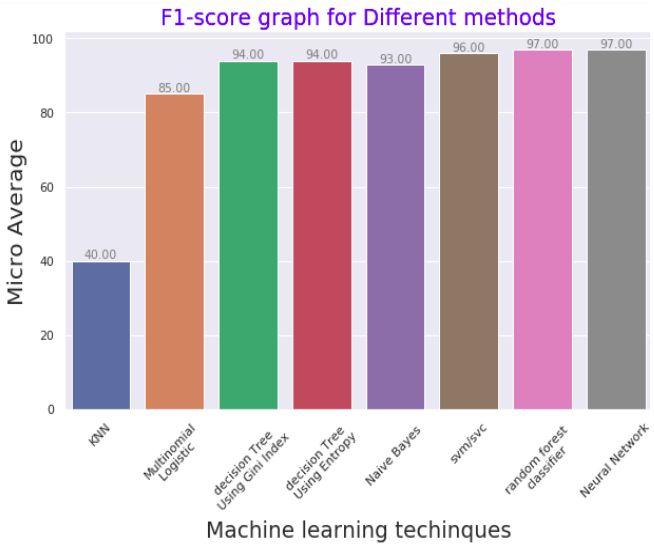

Fig.8.Recall in terms of Micro average

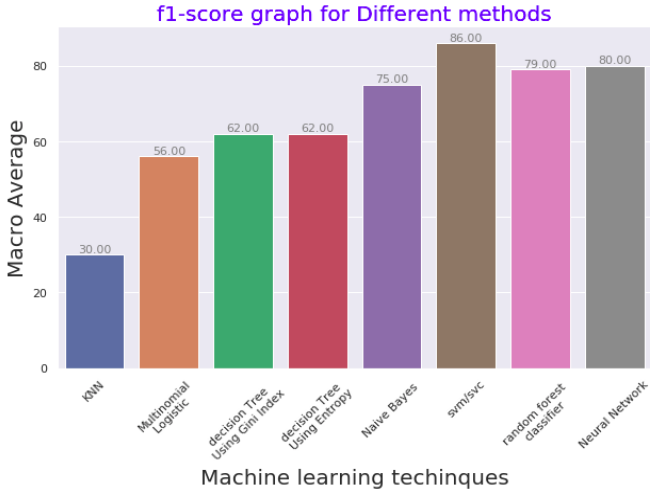

Fig.9. F1-score in terms of Macro average

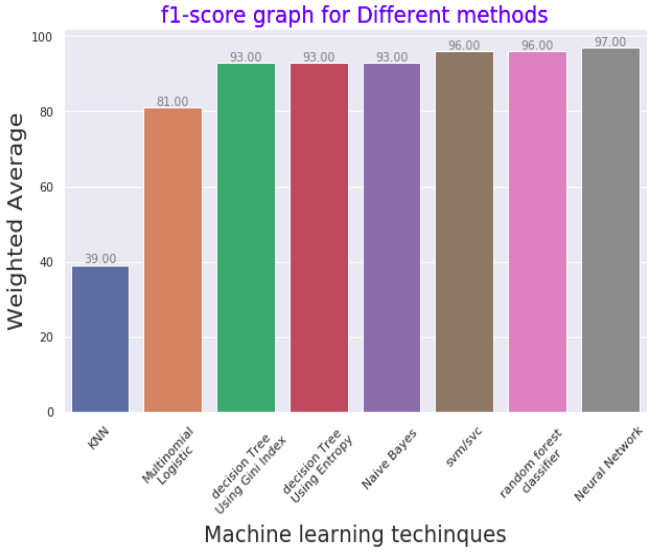

Fig.10.F1-score in terms of weighted average

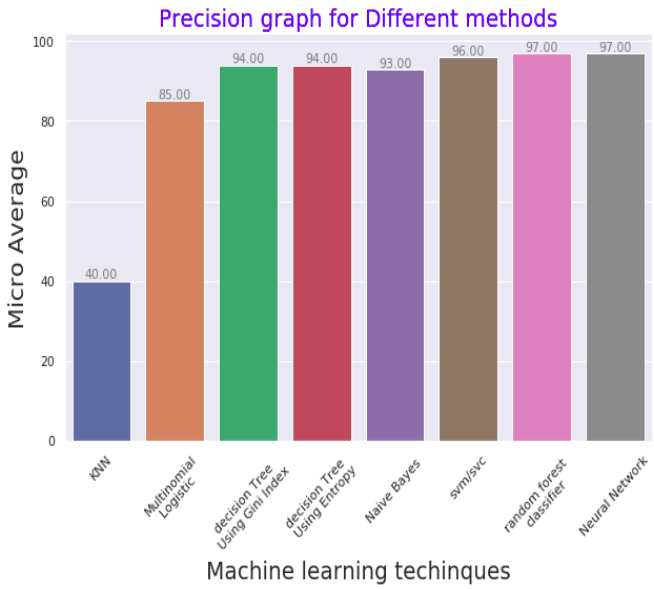

Fig.11.F1-score in terms of Micro average

VI. CONCLUSIONS

In this paper, we have examined the different methods that predicted the rating of the upcoming products. We have found that neural network classifier accuracy is reasonable compared to other classifierslike support vector classifier,K-nearest neighbors,Naive Bayes classifier, Random forest classifier,decisiontree, multinomial logistic regression. Support vector classifier is obtained the best f1-score macro average compared to other classifiers. It also concludes that based on experimental results our proposed method neural network classifier performs better in terms of generalization performance and training cost in case of a multiclass problem.

FUTURE WORK

Product price plays a very important role in e-commerce platform. Consumer always wants to know that what price of all upcoming products. One direction of our research is to predicting the price of newest arrival product. In some case, price of upcoming product is not available on e-commerce platform therefore it makes hard consumer to take decision about product.

ACKNOWLEDGEMENT

I am thankful to TEQIP III for providing financial support.

REFERENCES

[1] Sun, M. (2012). How does the variance of product ratings matter?. *Management Science*, *58*(4), 696-707.
[2] Patil, T. R., &Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International journal of computer science and applications*, *6*(2), 256-261.
[3] Rain, C. (2013). Sentiment analysis in amazon reviews using probabilistic machine learning. *Swarthmore College*.
[4] Engler, T. H., Winter, P., & Schulz, M. (2015). Understanding online product ratings: A customer satisfaction model. *Journal of Retailing and Consumer Services*, *27*, 113-120.

[5] Nichols, D. (1998). Implicit rating and filtering. ERCIM.
[6] L. Breiman, Random Forests. Machine Learning. 45(1), pp. 5-32
[7] E. Frank, R. Kirby (October 2019) (Revision 13864) [Online]Available:http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/RandomTree.html
[8] E. Frank, I. H. Witten, Generating Accurate Rule Sets Without Global Optimization. In: Fifteenth International Conference on Machine Learning, 1998, pp. 144-151
[9] G. H. John, P. Langley, Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 1995, pp. 338-345,
[10] Malcolm Ware (October 2019) (Revision: 14886) [Online] Available:http://weka.sourceforge.net/doc.dev/weka/classifiers/functions/MultilayerPerceptron.html.
[11] S. le Cessie, J.C. van Houwelingen, Ridge Estimators in Logistic Regression. Applied Statistics. 41(1), pp. 191-201.
[12] R. Kohavi, The power of decision tables. European conference on machine learning. Springer, Berlin, Heidelberg, 1995.
[13] I. Wayne and P. Langley. Induction of one-level decision trees. Machine Learning Proceedings 1992. Morgan Kaufmann, 1992, pp. 233-240.
[14] R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.
[15] E. Frank (October 2019) (Revision: 12887) [Online] Available:http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/REPTree.html

AUTHORS PROFILE

Gulab Sah is a research scholar in Computer Science & Engineering at National Institute of Technology, Arunachal Pradesh. He received Diploma in Computer Science and Engineering from Rajiv Gandhi government polytechnic. He also received Bachelor of Engineering from Visvesvaraya Technology University in 2016. He has done Master of technology in computer science and Engineering from National Institute of technology Arunachal Pradesh. His research interests include Network security, Machine learning, Predictive Analytics, Data Mining and other Data Science topics.

Rajat Subhra Goswami is Assistant Professor in the Department of Computer Science & Engineering at National Institute of Technology, Arunachal Pradesh.. He received the PhD degree in Computer Science & Engineering from the National Institute of technology Arunachal Pradesh, India. He also received his Master of Technology degree from Jadavpur University, West Bengal in 2009. His research interest includes information Security, cryptography, image processingand Network traffic classification

Sunit Kumar Nandi is Trainee teacher in the Departmentof computer Science & Engineering at National Institute of Technology, Arunachal Pradesh.. He received Bachelor of Technology from Assam Engineering College in 2016. He also received Master of technology fromIndian Institute of Technology Guwahati and pursing PhD from Indian Institute of Technology Guwahati. His research area is Computer Network and security, machine learning and data Science.