Linear Regression vs. Support Vector Machine: A Formal Enhancement in Online Purchase utilizing Novel Customer Feedback Segmentation

A.Bola Pavan Kumar[1], K. Malathi[2]*

[1]Research Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Tamil Nadu, India, PinCode: 602105

[2]*Project Guide, Corresponding Author, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Tamil Nadu, India, PinCode: 602105

ABSTRACT

Aim: The aim of the research is to improve the customer segmentation in Online purchases using Novel Customer Feedback. **Materials and methods:** The Categorizing is performed by adopting a sample size of n=10 in Linear Regression and sample size n=10 in Support Vector Machine algorithms with a sample size = 10 was iterated 20 times for efficient and accurate analysis on labeled images with G power in 80% and threshold 0.05%, CI 95% mean and standard deviation. **Results and Discussion:** The analysis of the results shows that the Linear Regression has a high accuracy of (88.56%) in comparison with the Support Vector Machine (81.30%). There is a statistically significant difference between the study groups with ($p<0.001$). **Conclusion:** Prediction in customer segmentation in E-Commerce shows that the Linear Regression appears to generate better accuracy than the Support Vector Machine algorithm.

Keywords: Novel Customer Feedback, Clustering Technique, Linear Regression, Support Vector Machine, Machine learning, Data Segmentation.

INTRODUCTION

The purpose of this research is to classify the customer segmentation in E commerce using Novel Customer feedback (Kashwan and Velu 2013). In this research the customers are divided according to their purchase of the products. It is found important in today's world since the pandemic everyone is willing to buy the necessary things online mostly (Tsiptsis and Chorianopoulos 2011). The application or websites they are using will show the good suggestion and the analysis made will give the good suggestion to the customers which will attract the customers through videophones, personal computers, consumer electronics like television and game boxes, digital assistants like voice-driven, pen-based computing, etc (Segovia, Szczepaniak, and Niedzwiedzinski 2013). The customer's data is collected from the E-commerce website and that data is analyzed by taking the products of the customer he bought and the review given to that product by using the different algorithms (Han, Pei, and Kamber 2011). The importance of this research work is that the customer data segmentation for e-commerce sites is done using the linear regression and Support Vector machine and comparing both

algorithms to get the highest accurate algorithm for customer data segmentation (Dolnicar, Grün, and Leisch 2018). This helps the e-commerce website to give the particular suggestions to that customer about the products.

In distinguishing and forecasting customer data segmentation using Clustering Technique in E-Commerce using the customer feedback technique by comparing over 962 journals from IEEE Xplore digital library, 563 articles from ScienceDirect, 1156 articles from Google Scholar and 459 articles from Springer and Customer feedback using rating of products is highly correlated parts in customer feedback (Shastri and Pathak 2018). Among all the articles and journals the most cited paper is (Carbajal 2021) nearly 1000 researchers cited these articles and this article for various e-commerce customer feedbacks and it will be a real source for further research (Shine et al. 2018). The most sequential and best research work on e-commerce with respect to various online platform studies helped (Yao, Eklund, and Back 2010) because this work shows the accurate analysis of the customers by analyzing the products that they are buying. The main objective of this research work is showing the linear regression accuracy of the customer data segmentation from sequential cloud environments. (Parakh et al. 2020; Pham et al. 2021; Perumal, Antony, and Muthuramalingam 2021; Sathiyamoorthi et al. 2021; Devarajan et al. 2021; Dhanraj and Rajeshkumar 2021; Uganya, Radhika, and Vijayaraj 2021; Tesfaye Jule et al. 2021; Nandhini, Ezhilarasan, and Rajeshkumar 2020; Kamath et al. 2020)

The methods which were used before have less accuracy on predicting customer data segmentation in e-commerce. The staging of the Support Vector Machine method is lagging to find the customer's division of a particular product with a particular unit price (Faraone, Gorgoglione, and Palmisano 2008) . In order to sequence the methods and techniques in this research study linear regression model generally fares better than Support Vector Machine. It also takes more time to train a random approach for analyzing the customer datasets which are based on the model to increase with the size of the datasets (Faraone, Gorgoglione, and Palmisano 2008; Wu and Chou 2011). Its estimates and calculations are probably done using feedback submitted by the customer. The aim of the research work is to classify the customer segmentation using Novel Customer feedback with Linear regression (“Customer Segmentation: Part 1” 2017).

MATERIALS AND METHODS

The research work was performed in the Image Processing Laboratory, Department of Computer Science and Engineering, Saveetha School Of Engineering, SIMATS. Basically it is considered that two groups of classifiers are used namely Linear Regression and Support Vector Machine, which is used to classify the customer segmentation in E commerce. Group 1 is the Linear Regression algorithm with the sample size of 10 and the Support Vector Machine is group 2 with the sample size of 10 and they are compared for more accuracy score and precision score values for choosing the best algorithm. The Pre- test analysis has been prepared by having G power of 80% and threshold of 0.05%, CI 90% mean and

standard deviation. Sample size has been calculated and it is identified that 10 sample groups in total 20 samples with a standard deviation for Linear regression = 88.56% and Support Vector Machine = 81.30% (Stamp 2017).

Linear Regression

Linear regression is a supervised learning machine learning algorithm. Linear regression algorithm shows a linear relationship between a dependent (x) and one or more independent (x) variables, hence called linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

Pseudocode for Logistic Regression

```
Import LinearRegression as LR
Import pandas as pd
Import Matplotlib.pyplot as plt
Compare from sklearn.ensemble import
Linearregression
from sklearn.tree import
DecisionTreeClassifier
Data extraction from sklearn
Initiate sklearn.metrics import accuracy
score
Calculate sequence
sklearn.mode_selection import
train_test_split
Find results from
sklearn.feature_extraction.value import
CountVectorizer
Count_vectorizer=CountVectorizer()
Cv.shape
lrg=LRclassifer(max_depth=8,estimators
=100,nthread=3)
lrgc.fit(x_train,y_train)
prediction_lrg=lrgc.predict(x_test)
print(accuracy_score(prediction_lrg,y_te
st)
```

Support Vector Machine

Support Vector Machine is one of the most popular supervised learning algorithms, which is used for classification as well as Regression problems. However, primarily, it is used for classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called qs support vectors, and hence the algorithm is termed as Support Vector Machine.

Pseudocode for Support Vector Machine

```
Import svm
Import svm as SVM
Compare from kaggle.ensemble
Import multi svm
Data extraction from kaggle
Calculate sequence from kaggle.metrics
Import accuracy score
Calculate sequence from
kaggle.model_selection
Import train_test_split
Import vectorizer
Count_vectorizer=Count_vectorizer(sto
p_words='english')
SVM = Multisvm
SVM.fit(X_train,Y_train)
Prediction_SVM=SVM.predict(X_test)
Print(accuracy_score(prediction SVM
action_Dt,y_test))
```

Statistical Analysis

The analysis was done using IBM SPSS version 21. It is a statistical

software tool used for data analysis. For both proposed and existing algorithms 10 iterations were done with a maximum of 20 samples and for each iteration the predicted accuracy was noted for analyzing accuracy. The value obtained from the iterations of the Independent Sample T-test was performed. The independent data sets are groups, accuracy and loss. The Dependent values are products and price. A detailed analysis has been done on these values for finding customer segmentation in E commerce.

RESULTS

The dataset is provided which selects the random samples from a given dataset for customer segmentation. The data is collected for a period of 30 days. As the sample sets are executed for a number of iterations the accuracy and precision values of Linear Regression and Support Vector Machine for customer segmentation with a mean value = 88.5690 %, Standard deviation = 1.13933. Thus the model is able to work efficiently to analyze customer segmentation. Table 1 Shows the customers data set with four attributes which selects the random samples from a given dataset. Implement Linear Regression train steps for each data point. Table 2 shows the group Statistics of Linear Regression with Decision Tree by grouping the iterations with Sample size 10, Mean = 88.5690 , Standard Derivation = 1.13933 , Standard Error Mean = 0.36029. Descriptive Independent Sample Test of Accuracy and Precision is applied for the dataset in SPSS. Here it specifies Equal variances with and without assuming a T-Test Score of two groups with each sample size of 10. The mean difference, standard deviation difference and significant difference of Linear Regression based customer segmentation and Support Vector Machine based Customer Segmentation is tabulated in Table 3, which shows there is a significant difference between the two groups since $P<0.001$ with an independent sample T-Test. The dependent variables in customer segmentation are predicted with the help of the independent variables. Figure 1 shows the comparison of Linear Regression over Support Vector Machine in terms of mean accuracy. It explores that the mean accuracy is slightly better than Support Vector Machine and the standard deviation is moderately improved compared to Support Vector Machine.

DISCUSSION

In this research customer feedback based customer segmentation has proven to be a highly effective and versatile approach (Collica 2017). The applied unsupervised learning with R explains clustering methods, distribution analysis, data encoders, and features of R that enable us to understand the given data better and get answers to our point of segmentation problems. Clustering algorithms like k-means, divisive, and agglomerative will be used for segmentation (Malik and Tuckfield 2019). These techniques are improvised and made to an extent of implementing accurate customer segmentation in e-commerce.

It was observed that the important factors and the concept of customer value has been analyzed based on the matrix for the customer segmentation (Dibb and Simkin 2013). The discussion offers insights into the reasons for the development of the practical approach, a concrete methodology for its implementation and strategic and tactical

applications of the customer segmentation (Cabrera 2011). In this the linear regression is also used for predicting the output of the given data. It is found that the recent research works on retail customers who are getting the products frequently from a store or warehouse. Thus the linear regression shows the more accurate results in it (Daniel 2015). The main purpose of the customer segmentation is to divide the target market into subsets that share similar characteristics, needs and properties (Dixit 2017; Hruschka and Natter 1992). In this the customers are being separated based on the parameters like product reviews,buying patterns by their feedback on the products. The Novel Customer Feedback technique along with the predictive neural network to give the accurate results for the data. A comprehensive model based on generalized association rules and Support Vector Machine technology (Ma 2015). In this the customer segmentation is done using Support Vector Machine technology which will not produce good results because it is taking the customers decision for making new predictions which is a little bit more process for the customer segmentation. Finally it is observed to find the results that validate Linear regression as higher in comparison with the Support Vector Machine.

Hence the study results produce clarity in performance with both experimental and statistical analysis, but it has some limitations to the proposed work such as threshold and precision. The accuracy level of customer segmentation using Novel Customer feedback can still be improved by implementing artificial intelligence techniques to predict and analyze results while comparing with existing machine learning algorithms. In the future, the large dataset for customers can be considered to validate our proposed model with respect to recent scenarios.

CONCLUSION

The current study focused on machine learning algorithms, Linear Regression over Support Vector Machine for higher classification in customer segmentation. It can be slightly improved based on the Support Vector Machine data sets analysis in future. The outcome of the study shows Linear Regression 88.5690% higher accuracy than Decision tree 81.3050%.

DECLARATION

Conflict of Interest

No conflict of interest

Author Contributions

Author ABPK was involved in data collection, data analysis, manuscript writing. Author DV was involved in the Action process, Data verification and validation, and Critical review of manuscript.

Acknowledgments

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this research work successfully.

Funding: We thank the following organization for providing financial support that enabled us to complete the study.

1. SNEW.AI Technologies, Chennai.
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

REFERENCES

1. Cabrera, Delia. 2011. "A Review of Algorithms for Segmentation of Retinal Image Data Using Optical Coherence Tomography." *Image Segmentation*. https://doi.org/10.5772/15833.
2. Carbajal, Santiago García. 2021. "Customer Segmentation through Path Reconstruction." *Sensors* 21 (6). https://doi.org/10.3390/s21062007.
3. Collica, Randall S. 2017. *Customer Segmentation and Clustering Using SAS Enterprise Miner,Third Edition*. SAS Institute.
4. "Customer Segmentation: Part 1." 2017. https://doi.org/10.4135/9781529733433.
5. Daniel, Ines. 2015. "MARKET SEGMENTATION USING COLOR INFORMATION OF IMAGES." *International Journal of Electronic Commerce Studies*. https://doi.org/10.7903/ijecs.1400.
6. Devarajan, Yuvarajan, Beemkumar Nagappan, Gautam Choubey, Suresh Vellaiyan, and Kulmani Mehar. 2021. "Renewable Pathway and Twin Fueling Approach on Ignition Analysis of a Dual-Fuelled Compression Ignition Engine." *Energy & Fuels: An American Chemical Society Journal* 35 (12): 9930–36.
7. Dhanraj, Ganapathy, and Shanmugam Rajeshkumar. 2021. "Anticariogenic Effect of Selenium Nanoparticles Synthesized Using Brassica Oleracea." *Journal of Nanomaterials* 2021 (July). https://doi.org/10.1155/2021/8115585.
8. Dibb, Sally, and Lyndon Simkin. 2013. *Market Segmentation Success: Making It Happen!* Routledge.
9. Dixit, Abhishek. 2017. *Neural Network in Market Segmentation: A Descriptive Study of Market Segmentation Using Artificial Neural Networks*.
10. Dolnicar, Sara, Bettina Grün, and Friedrich Leisch. 2018. *Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful*. Springer.
11. Faraone, M. F., M. Gorgoglione, and C. Palmisano. 2008. "Using Contextual Information in Transactional Segmentation: An Empirical Study in E-Commerce." *2008 IEEE International Conference on Data Mining Workshops*. https://doi.org/10.1109/icdmw.2008.64.
12. Han, Jiawei, Jian Pei, and Micheline Kamber. 2011. *Data Mining: Concepts and Techniques*. Elsevier.
13. Hruschka, Harald, and Martin Natter. 1992. *Using Neural Networks for Clustering-Based Market Segmentation*.
14. Kamath, S. Manjunath, K. Sridhar, D. Jaison, V. Gopinath, B. K. Mohamed Ibrahim, Nilkantha Gupta, A. Sundaram, P. Sivaperumal, S. Padmapriya, and S. Shantanu Patil. 2020. "Fabrication of Tri-Layered Electrospun Polycaprolactone Mats with Improved Sustained Drug Release Profile." *Scientific Reports* 10 (1): 18179.
15. Kashwan, Kishana R., and C. M. Velu. 2013. "Customer Segmentation Using Clustering and Data Mining Techniques." *International Journal of Computer Theory and Engineering*. https://doi.org/10.7763/ijcte.2013.v5.811.
16. Ma, Haiying. 2015. "A Study on Customer Segmentation for E-

Commerce Using the Generalized Association Rules and Decision Tree." *American Journal of Industrial and Business Management*. https://doi.org/10.4236/ajibm.2015.512078.

17. Malik, Alok, and Bradford Tuckfield. 2019. *Applied Unsupervised Learning with R: Uncover Hidden Relationships and Patterns with K-Means Clustering, Hierarchical Clustering, and PCA*. Packt Publishing Ltd.
18. Nandhini, Joseph T., Devaraj Ezhilarasan, and Shanmugam Rajeshkumar. 2020. "An Ecofriendly Synthesized Gold Nanoparticles Induces Cytotoxicity via Apoptosis in HepG2 Cells." *Environmental Toxicology*, August. https://doi.org/10.1002/tox.23007.
19. Parakh, Mayank K., Shriraam Ulaganambi, Nisha Ashifa, Reshma Premkumar, and Amit L. Jain. 2020. "Oral Potentially Malignant Disorders: Clinical Diagnosis and Current Screening Aids: A Narrative Review." *European Journal of Cancer Prevention: The Official Journal of the European Cancer Prevention Organisation* 29 (1): 65–72.
20. Perumal, Karthikeyan, Joseph Antony, and Subagunasekar Muthuramalingam. 2021. "Heavy Metal Pollutants and Their Spatial Distribution in Surface Sediments from Thondi Coast, Palk Bay, South India." *Environmental Sciences Europe* 33 (1). https://doi.org/10.1186/s12302-021-00501-2.
21. Pham, Quoc Hoa, Supat Chupradit, Gunawan Widjaja, Muataz S. Alhassan, Rustem Magizov, Yasser Fakri Mustafa, Aravindhan Surendar, Amirzhan Kassenov, Zeinab Arzehgar, and Wanich Suksatan. 2021. "The Effects of Ni or Nb Additions on the Relaxation Behavior of Zr55Cu35Al10 Metallic Glass." *Materials Today Communications* 29 (December): 102909.
22. Sathiyamoorthi, Ramalingam, Gomathinayakam Sankaranarayanan, Dinesh Babu Munuswamy, and Yuvarajan Devarajan. 2021. "Experimental Study of Spray Analysis for Palmarosa Biodiesel-diesel Blends in a Constant Volume Chamber." *Environmental Progress & Sustainable Energy* 40 (6). https://doi.org/10.1002/ep.13696.
23. Segovia, Javier, Piotr S. Szczepaniak, and Marian Niedzwiedzinski. 2013. *E-Commerce and Intelligent Methods*. Physica.
24. Shastri, Niket, and Kamlesh Pathak. 2018. "Estimation of Perceptible Water Vapor of Atmosphere Using Artificial Neural Network, Support Vector Machine and Multiple Linear Regression Algorithm and Their Comparative Study." https://doi.org/10.1063/1.5033289.
25. Shine, Philip, John Upton, Ted Scully, Laurence Shalloo, and Michael D. Murphy. 2018. "Comparing Multiple Linear Regression and Support Vector Machine Models for Predicting Electricity Consumption on Pasture Based Dairy Farms." *2018 Detroit, Michigan July 29 - August 1, 2018*. https://doi.org/10.13031/aim.201801140.
26. Stamp, Mark. 2017. *Introduction to Machine Learning with Applications in Information Security*. CRC Press.
27. Tesfaye Jule, Leta, Krishnaraj Ramaswamy, Nagaraj Nagaprasad, Vigneshwaran Shanmugam, and

Venkataraman Vignesh. 2021. "Design and Analysis of Serial Drilled Hole in Composite Material." *Materials Today: Proceedings* 45 (January): 5759–63.

28. Tsiptsis, Konstantinos K., and Antonios Chorianopoulos. 2011. *Data Mining Techniques in CRM: Inside Customer Segmentation*. John Wiley & Sons.
29. Uganya, G., Radhika, and N. Vijayaraj. 2021. "A Survey on Internet of Things: Applications, Recent Issues, Attacks, and Security Mechanisms." *Journal of Circuits Systems and Computers* 30 (05): 2130006.
30. Wu, Roung-Shiunn, and Po-Hsuan Chou. 2011. "Customer Segmentation of Multiple Category Data in E-Commerce Using a Soft-Clustering Approach." *Electronic Commerce Research and Applications*. https://doi.org/10.1016/j.elerap.2010.11.002.
31. Yao, Zhiyuan, Tomas Eklund, and Barbro Back. 2010. "Using SOM-Ward Clustering and Predictive Analytics for Conducting Customer Segmentation." *2010 IEEE International Conference on Data Mining Workshops*. https://doi.org/10.1109/icdmw.2010.121.

TABLES AND FIGURES

Table 1. Customers data set with four attributes which selects the random samples from a given dataset. Implement Linear Regression train steps for each data point.

Target	Product_ID	Date	Customer Name	Price (in Rs)
1	1A772S15	05-09-2021	Johar	1000
2	1A772S20	06-09-2021	Malhotra	4586
3	1A772P28	06-09-2021	Greeshma	1568
4	1O668H96	10-09-2021	Mayuri	1999
5	1O668H76	14-09-2021	Hitesh	9999
6	1B2774Y67	16-09-20201	Phani	1599
7	1B4563H89	21-09-2021	Vayusuth	7599
8	1G3245U76	22-09-2021	Nainekka	15339
9	1G4563U13	25-09-2021	Tejaswini	6299
10	1H6734Y99	29-09-2021		799

Table 2. Group Statistics of Linear Regression with Support Vector Machine by grouping the iterations with Sample size 10, Mean = 88.5690 , Standard Derivation = 1.13933 , Standard Error Mean = 0.36029. Descriptive Independent Sample Test of Accuracy and Precision is

applied for the dataset in SPSS. Here it specifies Equal variances with and without assuming a T-Test Score of two groups with each sample size of 10.

	Group	**N**	**Mean**	**Std.Deviation**	**Std error mean**
Accuracy	Linear Regression	10	88.5690	1.13933	.36029
	SVM	10	81.3050	.71719	.22680
Precision	Linear Regression	10	1.9220	.56829	.17971
	SVM	10	2.9460	.50507	.15972

Table 3: Independent Sample Test of Accuracy and Loss (calculate P- Value < 0.001 and significant value = 0.075, Mean difference = 7.26400 and confidence interval = (8.158420 - 6.36958).

		F	**Sig**	**T**	**Df**	**Sig(2 .tailed**	**Mean difference**	**Std error difference**	**lower**	**upper**
Accuracy	Equal Variance Assume	3.575	0.075	17.063	18	<.001	7.26400	0.42573	6.36958	8.15842
	Equal variance not assumed			17.063	15.165	<.001	7.26400	0.42573	6.35744	8.17056
Precision	Equal Variance Assumed	0.189	0.669	-4.259	18	<.001	-1.02400	0.24043	-1.52912	-0.51888
	Equal variance not assumed			-4.259	17.755	<.001	-1.02400	0.24043	-152961	-0.51839

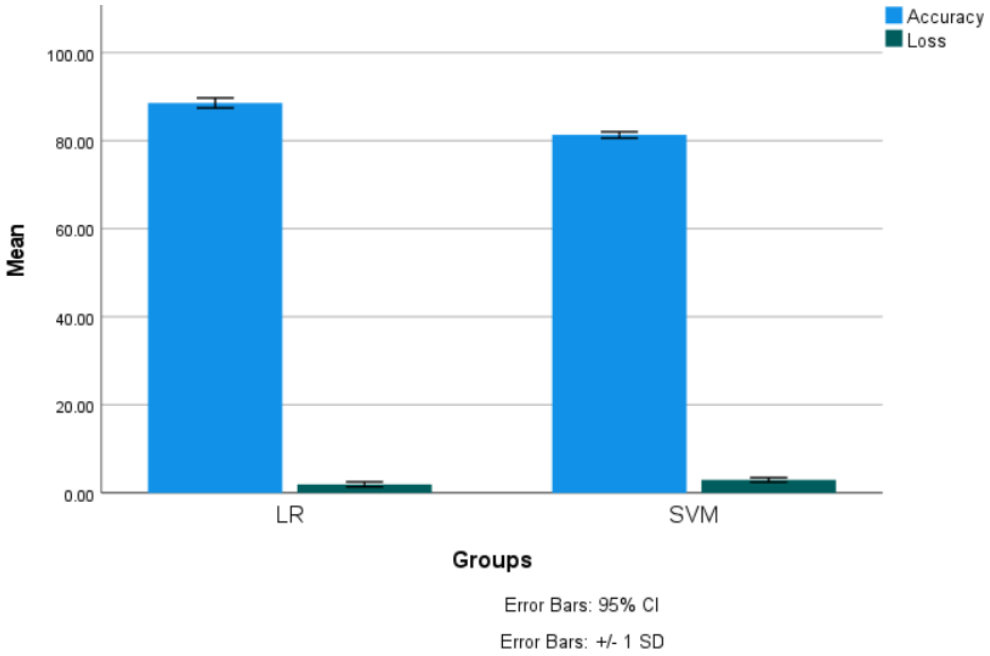

Fig. 1. Comparison of Linear Regression over Support Vector Machine in terms of mean accuracy. It explores that the mean accuracy is slightly better than Support Vector Machine and the standard deviation is moderately improved compared to Support Vector Machine. Graphical representation of the bar graph is plotted using groupid as X-axis LR vs SVM, Y-Axis displaying the error bars with a mean accuracy of detection +/- 1 SD.