International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V72I4P121

Volume 72 Issue 4, 199-208, April 2024
© 2024 Seventh Sense Research Group®

Original Article

A Novel Sales Promotional Schemes based on Clustering and Linear Regression Analysis

Atul O. Thakare[1], Soora Narasimha Reddy[2], Omprakash W. Tembhurne[3], Parag S. Deshpande[4]

[1]STME, SVKM's NMIMS University, Navi Mumbai, Maharashtra, India.
[2]CSE (Networks) Department, Kakatiya Institute of Technology & Science, Warangal, Telangana, India.
[3]CSE Department, MIT School of Engineering, MIT Art Design and Technology University, Pune, Maharashtra, India.
[4]Visvesvaraya National Institute of Technology, Nagpur, Maharashtra, India.

[1]Corresponding Author : aothakare@gmail.com

Received: 12 January 2024 Revised: 03 April 2024 Accepted: 11 April 2024 Published: 24 April 2024

Abstract *- Different sales promotion schemes are used to promote sales among customers. Depending upon the sales promotion mix, the response of customers changes. One of the most common sales promotion schemes is an advertisement. For example, customers of mid-age may respond more positively to an advertisement in the newspaper than an advertisement on social media. Generally, the relationship between dependent (sales) and independent variables (television advertisement, newspaper advertisement, social media advertisement) is inferred using regression techniques. However, suppose the number of customers is from different groups. In that case, one generalized equation may not represent the relationship accurately due to various groups in data because one single line may not fit the data well. In the proposed work, we present an algorithm for grouping customers having similar dependency relationships and then extracting such relationships for each group. The dependency relationship is extracted by using regression and the accuracy is measured using Mean Square Error (MSE). Experimentation on standard and synthetic datasets proves that the proposed method enhances accuracy to a very large extent and it will be practically applicable to extract more strategic information from the dataset.*

Keywords *- Regression, MSE, R-square, Clustering, Segmented Regression.*

1. Introduction

Sales Promotion is an important component in gaining market share in a highly competitive business environment. Many tactics for sales marketing are used to promote sales among customers [1]. A sales promotion plan may receive different responses from different clients. One of the most common techniques is an advertisement. It is best known for its pervasiveness. Consider the following example. A company wants to sell its product by promoting it in various mediums like social media, television, and newspapers. Now, the company obviously wants to know the impact of advertisements on sales of the product from the historical data. This result can be easily obtained using various regression techniques. A regression equation can be drawn with sales as the dependent variable and advertising techniques as the independent variable. But consider the real-life situation where customers belong to different groups, which may be formed due to age variation. Then, for each such group, the advertisement impact will be different. Customers belonging to the age group of 15 to 25 respond more positively to social media advertisements, while customers in the age group of 26 to 40 respond positively to a television advertisement. On the other hand, customers aged 41 and above find newspaper advertisements more attractive. In such a case, one single regression equation may not represent the relationship between sales and advertisement techniques accurately. Therefore, the problem is to find a mechanism to find a dependency relation among data that consists of distinct groups. One way could be to cluster the data into different groups and find dependency relations for each group. However, clustering algorithms find similarities based on spherical distance; therefore, it works well when data is roughly spherical. In real life, data is generally messy, i.e., it is not isotropic data, which makes it hard to cluster it using clustering algorithms. Therefore, the challenge is to divide this data into groups and find dependency relationships among them, considering the scalability of data and outliers. To handle the above-mentioned challenges in the problem, we propose a scalable algorithm providing the following essential features.

- The proposed algorithm can extract dependency relationships by grouping data with very high accuracy as compared to the existing state-of-the-art methods.
- The proposed algorithm is scalable and runs almost independently of the size of the data. The complexity of

the algorithm is n^{2,} where n is the number of sample points.

- The algorithm detects outliers in the data and ignores them to provide results independent of them.
- The accuracy of the algorithm is experimentally proved on standard datasets and synthetic datasets with RMSE as the performance measure.

1.1. Literature Review

Clustering aims at discovering similarities and differences among patterns as well as drawing insightful and useful conclusions. The similarity and dissimilarity are determined by the proximity measure between the two patterns. The kind of clusters anticipated to be hidden in the data set is determined by the clustering criterion. Having adopted a proximity measure and a clustering criterion, the clustering algorithm [2] refers to the selection of a certain algorithm scheme that reveals the clustering outline of the data set [3]. Clustering algorithms may be looked at as procedures that, by considering just a small portion of the set containing all viable divisions of the pattern set, provide us with a meaningful grouping [4]. The output is determined by both the specific algorithm and the selection criteria used.

The most popularly used algorithm is K-means which is a simplistic partition-based clustering method. Reducing the squared error across all K clusters is the main objective of K-means. K-means begins with an initial split of K clusters and assigns patterns to the clusters to minimize the squared error. When the number of clusters rises, naturally, the squared error will gradually decrease. Hence, only for a predetermined number of clusters, the squared error can be minimized. One of the values K-means asks the user for is the number of clusters [5].

The main idea of density-based clustering methods is that the data points situated in the region of the high density are assumed to be in the same cluster. The most popular density-based clustering algorithm is DBSCAN. DBSCAN's core idea is that a cluster consists of points with at least X nearby points within a radius of Y units. X & Y are input parameters. OPTICS is an improvement over DBSCAN, and it overcomes DBSCAN sensitivity to these two parameters: the neighbourhood radius and the minimum number of points in the neighbourhood [6]. Authors in [7] proposed a hierarchical divisive version of K-means, called bisecting K-means, for clustering a large number of text documents. This algorithm combines the efficiency of K-means and cluster quality of agglomerative hierarchical clustering.

The success of ensemble techniques for supervised learning prompted the development of the same with unsupervised learning. The core concept is that we can create several clustering partitions through different parameter initialization, and by merging the generated ensemble of partitions, we can get the clusters which are rightly packed as well as sufficiently dispersed [8]. Sequential algorithms typically lead to compact and hyperspherical or hyperellipsoidally shaped clusters based on the chosen distance measure. At each step, the agglomerative algorithms generate a series of clustering with a smaller number of clusters. Each step's clustering is the outcome of merging multiple previous clusters into one. These techniques work well for recovering elongated clusters (single link algorithm) and compact clusters (full link algorithm). Clustering algorithms are built on cost function-based optimization algorithms, where "sensible" is measured by a cost function, J, based on which clustering is assessed. The number of clusters, m, is typically set. The majority of these algorithms employ techniques from differential calculus and create consecutive clustering while attempting to maximize J. They come to an end when J's local optimum is identified. These kinds of algorithms are also known as iterative function optimization techniques.

The borders of the areas where clusters are located are adjusted repeatedly using boundary detection techniques. These algorithms are distinct from the ones mentioned above, even though they also follow a cost-function optimization philosophy. All these approaches aim to locate cluster representatives in space in an optimal way. The "kernel trick" is utilized by kernel-based techniques to execute a mapping from the original space, X, into a high-dimensional space, Y, in the context of nonlinear support vector machines. This technique is typically used when clusters have nonlinear forms. None of the above-mentioned algorithms can create clusters when patterns are connected to established relationships.

Dependency algorithms like linear regression can infer the relationship between regressand (dependent) and regressor (independent) variables. But when data is divided among different groups then one equation may not represent the relationship accurately. In such cases, only one generalized regression cannot give the desired result. Hence, linear regression cannot be used to find dependency relationships in such kinds of data. Another way could be to use Polynomial Regression rather than Linear Regression as in polynomial regression, the relationship is a curvilinear representation, but using this kind of regression for such data can be very complex as data here is divided into different groups, and the regression equation may be very complex which would not express the relationship very clearly. Also, as data is divided into distinctive groups, the curve may not fit the data points well. Therefore, these approaches fail to give accurate results, and hence, in the proposed approach dependency relationship is used to group data. Segmented regression is regression where the interval relationships are derived by linear regression. Support vector regression uses “Kernel Trick” to obtain nonlinear relationships. Now, since data is divided into groups, a generalized regression cannot be used. Therefore, we first tend to cluster the data and then find a dependency relation for each cluster. Grouping algorithms like clustering can be used, but, in general, clustering algorithms find similarities based on the distance between two data points [9]. Regarding the nearby centroid, the algorithms try to reduce the within-cluster variation. This works perfectly well when the data is generally spherical, that is when the data points are evenly spaced over the two-dimensional plot, normally distributed, and isotropic (that is, they have the same variance in all directions). But the real-life data is generally messy. Moreover, it typically has a Gaussian distribution. Mostly, it is not isotropic. Therefore, it is difficult for clustering algorithms to determine which centroid each data sample is nearest to. Hence, clustering algorithms cannot be used in such types of data.

Tharwat, A [10] illustrates the behaviour of the SVM classifier using mathematical and geometrical interpretations, visualizations, and experimenting with different values of kernel parameters and penalty parameters. It is demonstrated that the choice of these two parameters decides the complexity of the classification model, as well as the impact of the over-fitting / under-fitting problem on the performance of the classifier. The paper [11] presents a survey of 77 popular regression models, which belong to 19 different categories. To name a few categories, they are linear and non-linear regression, LASSO and ridge regression, bagging and boosting, quantile regression, support vector regression, neural network regression, etc.

The study by Mussol, S. [12] looks at the range of circumstances in which sales promotion activities might lead to relational advantages and better consumer-brand connections. These advantages might be advantageous for brand expression and client loyalty. The findings of the study demonstrate that non-cash sales promotions, such as in-store raffles, competitive gaming, and giveaways, have a greater positive effect on brand expression than cash sales promotions, like in-store discounts. Paper [13] offers the SPKC algorithm, which fuses the established K-means algorithm with the self-paced learning strategy. To discriminate between noisy and normal data when locating clusters, this approach employs a linear self-paced regularisation factor.

The paper from [14] gives a thorough and methodical overview and comparison of major clustering methods with uses in statistics, computer science, and machine learning. This paper analyses various approaches for clustering data having different types of structures. It also discusses issues like cluster analysis, selection of the number of clusters, distance and similarity measures, the importance of feature selection and feature extraction, feature standardization and normalization, cluster validation, etc. Authors in the paper [15] conceive of the Clusterwise Linear Regression (CLR) method, a combination of clustering and linear regression, for the forecasting of monthly rainfall. In addition, the suggested method's outcomes were contrasted with those of other approaches already in use, including multiple linear regression, artificial neural networks, and support vector machines.

The goal of [16] aims to present a detailed evaluation of several clustering algorithms in data mining in the context of two distinct marketing communication instruments, advertisement and sales promotion. Authors [17] suggest a K-means and support vector machine-based clustering-based sales forecasting approach. For locating the clusters, three similarity metrics—mini-max, median, and mean are suggested.

Authors W. A. Boland et al. in [18] explain how children's reactions to sales promotions create a big influence on how they purchase. While sales promotion campaigns may arouse interest and desire amongst kids, they may also trigger impulsive purchases and affect the development of brand loyalty. Authors Y. Duan et al. in [19] state that sales and prices in the e-commerce industry are greatly influenced by online reviews and discounts. While bad evaluations can result in reduced pricing and fewer sales, positive reviews strengthen the reputation of the product and support higher prices. The authors also discussed different scenarios in which online reviews and coupons can work together to influence consumer behaviour.

Paper [20] aims to quantify the impact of various locational metrics on sales dynamics over a broad spectrum of product categories. Along with calculating geographical impacts on sales at the product-category level, authors also discover and assess clusters of product categories that have similar sales patterns. A real-time sales prediction model is proposed by authors D. Li et al. in [21] and is based on a variety of different influential components. A stage future vision mechanism is created by combining a two-stage Long Short-Term Model (LSTM) with a two-layer Convolutional Neural Network (CNN). This mechanism is used to predict dependent correlations between possible future important variables and product sales.

The analysis of the relationships and outcomes of competing organizations in a market is presented together in the paper [22] with the dynamics of an advertising competition model. The model often considers elements like advertising costs, sales incentives, market share, and customer reaction. Paper [23] discusses the importance of social media marketing. Authors said that businesses may use social media for marketing purposes more effectively, respond to customers swiftly, foresee grievances and communicate product information more efficiently than they could with traditional media. The paper emphasises the fact that activities like advertising, sales promotions, social media, and conventional publicity have a larger role in generating income from company sales. Sales promotions are typically thought of as a strategy to increase direct sales, whereas advertising is seen as a crucial instrument for building brand equity, brand loyalty, and brand attitude. Advertising that invests a large sum of money in sales promotions must comprehend the attitude that the brand will take on as a consequence of these efforts [24, 25].

According to Yang and Peterson [26], customer satisfaction also has an impact on customer loyalty in addition to a direct effect from customer perceived value. The study in [27] looks at the interconnections between variables and identifies and analyzes seven key aspects that affect customer loyalty. While a customer's expectations are shaped by their impression of the business, the customer's evaluation of the quality affects how satisfied they are with the service [28].

2. Research Methods

2.1. Terms and Definitions

2.1.1. Coefficient of Determination (R^2)

R^2 (R-square) is used for evaluating the goodness of fit of the regression equation. R-square, also known as the coefficient of determination assesses how well the data truly matches the linear approximation given by the least-squares regression line.

2.1.2. Mean Square Error (MSE)

The regression estimated accuracy is gauged by looking at the mean square error. It measures the discrepancy between the expected and actual response values.

2.1.3. Direction Cosines (DC)

The direction cosines (also known as directional cosines) of a vector are the sines of the angles between the vector and the system's orthogonal coordinate axes. These are, in essence, each basis component's contributions to a unit vector pointing in that direction. Normally, they are used for 3-dimensional axes in analytical geometry, but in the proposed work, their definition is extended to the n-dimensional system.

2.2. Proposed Solution

In the proposed solution, similar customers are grouped based on their dependency relationship, and then each group can be represented by their dependency relationship. We first divide the data into the required number of groups, and regression is applied to each group. Each group has a regression equation. The idea used here is derived from segmented regression in which data is first divided, and a regression equation for each division is obtained. The main task here is to find such groups where the similarity of the members is based on the similarity of their dependency relationship. To find groups, sample data is selected out of whole data and direction cosines of all possible combinations of data points are found. Using the concept that points belonging to similar planes have similar direction cosines, and the data is divided based on direction cosines. Collinear points are ignored as collinear points belong to one plane, so they may give false clustering results. Therefore, only non-collinear points are allowed, and then a dependency relationship for each group is found which is then used for dividing the whole data set. Once we have the final clusters, we need to update the dependency relationship for each cluster as new data points are added to the cluster.

Therefore, we find the regression equation for each cluster and use regression statistics to analyse the performance of the regression. To provide scalability and effect independent of outliers, the following processing is added to the design.

- To handle large data, instead of running the algorithm on whole data, we randomly select a set of sample points; the algorithm is run only on the sample set of points and regression equations for those clusters are obtained. These equations can now be used to cluster whole data by looking for points satisfying respective equations. These equations are checked against all the points and points satisfying any of the equations are put in the respective group.
- Since the algorithm divides the data based on similar direction cosines, then the points which are having a substantial difference in direction cosines are considered outliers. Such points are not added to any of the existing clusters. These points are checked to see if they form a separate cluster among themselves. If they form a cluster, one more group is added, and its regression equation is obtained. If not then these points are left as outliers, and they are not included in any of the clusters.
- As we have used the concept that points lie in a plane belonging to one cluster, therefore direction cosines of those points lying in one plane are the same. In case certain points may not lie exactly on the plane but near it, those points must also be included in that cluster. Those points may not have the same direction cosines but direction cosines near to that of the plane. We have handled this situation by assigning a range for approximately similar direction cosines.
- Since collinear points lie on the same plane, these points will have similar direction cosines, and thus, they will be falsely clustered into one group. Therefore, it is important to remove collinear points first; the condition for collinearity of points is checked (direction ratios of collinear points are proportional), and only non-collinear points are allowed.

2.3. Segmented Regression Algorithm

K-Means Regression Algorithm

Input: Data, NumberOfClusters(k), NumberOfAttributes(d).

Output: Clusters, Regression equation, Regression statistics for each cluster

```
– Randomly select a sample data set from the given data
– Find all possible combination sets of sample data points of
  size d
– For each combination
```

```
•   If points are collinear
                do nothing
    Else
•   find direction cosine
•   count the frequency of unique direction cosine
    considering similarity using Euclidean distance
•   record the points corresponding to each DC
    End If

End For
–   find the k most frequent DC's
–   assign k most frequent DCs to initial seed points for K
    means clustering algorithm
–   apply k means algorithm on all possible combinations
–   get the clusters
–   apply regression to each cluster
–   divide whole data based on the previously obtained
    dependency relationship
–   check for outlier points which don't belong to any cluster
–   If outlier points form a cluster among themselves
        •   consider a new cluster
–   Else
        •   ignore these points
–   EndIf
–   find regression equation and regression statistics for each
    group
```

3. Analysis

The performance of an algorithm is measured by its complexity. We compute the complexity of this algorithm as the complexity of the part responsible for dividing the sample data points into the required number of groups. There are various steps involved in this algorithm. The challenging part of the algorithm is to divide the data into the required number of clusters.

Therefore, the complexity of the algorithm can be derived from the complexity of this part of the algorithm. In this part of the algorithm, we initially sample the data and find direction cosines of all possible combinations of sample data points. Then we find k most frequent direction cosines, where k is the number of required clusters, and initialize the centroids for K means clustering algorithm. We get the sample data divided into groups, we find a dependency relationship for each cluster and hence, we can use this dependency relationship to divide the whole data. The complexity is computed in the following way:

Assume r is the number of attributes and n is the number of sample data points. For obtaining all possible combinations, we have $^{n}C_{r}$ choices. Therefore, the complexity of this step is $O(^{n}C_{r})$. (assuming $r < n$) The complexity of finding the direction cosine for each combination is O(r), and the complexity of finding it for all possible combinations becomes $O\ (^{n}C_{r} *r)$. The complexity of the K-means algorithm is n^2. The overall complexity becomes $O\ (^{n}C_{r} *r + n^{2})$. Practically number of distinct groups is much less, so n can be chosen so that the algorithm is executed in real time.

4. Experimentation

Since there is no standard dataset available for testing and comparing, the data is generated by considering multiple dependencies and inducing outliers. The algorithm experiments on customer data showing a response of advertisements on sales which is dependent on groups.

4.1. Data Generation

The data is used to observe the impact of advertisements on sales of the company. The data has independent variables such as Newspaper advertisements, TV advertisements, social media advertisements, and sales as dependent variables. The data is categorized into 3 groups based on age. The groups are
Group 1: 15 - 25 age
Group 2: 26 - 40 age
Group 3: 41 and above age

Since this data consists of distinct groups, i.e. they are responding differently to advertisement mix, a single regression line may not fit the data well. Therefore, data is generated for each group assuming some dependency, and all the combined data is given as input to the algorithm. In order to analyze the performance of the algorithm, the data is checked to see if it has a linear dependency, i.e. the data is also checked against generalized linear regression.

In order to check if data points lying near the cluster are included in that cluster or not. The points are generated by slightly modifying the equation with a 1% to 5% change in the equation coefficients. In order to check if outlier points were detected and formed a group among themselves or not, a random equation was added to the set of equations. This equation was not included as the equation of existing dependency points satisfying this equation was included in the input dataset. Data is generated using three hypothetical equations, one for each group. Each equation has three independent (TV advertisement, newspaper advertisement, social media advertisement) and one dependent (sales) variable. Points satisfying these equations are generated and taken as input datasets to the algorithm. The data is shown in Table 1.

Table 1. Customer advertisement data

Sr.No.	X1	X2	X3	X4
1	3.0	5.0	1.0	5.25
2	6.0	6.0	2.0	2.25
3	8.0	3.0	2.0	12.75

Abbreviations X1 ... X3 in the above figure represent the advertisements in the following media.
X1 - social media, X2 - Newspaper, X3 - TV, X4 – Sales.

4.2. Results and Discussions

The result of the algorithm is the three dependency relations, one for each group, and it is measured using regression statistics like Mean Square Value (MSE) value and Coefficient of Correlation(R^2). The outlier points were also detected and they formed a group among themselves. The points which lie near the cluster were also included in that cluster which gave more accurate clustering. This result is equivalent to the desired output. The results of the experiment are shown in Table 2. The algorithm mines the three groups based on their dependency relationship and it is observed that the dependency of sale is different for each group. For example, in group 3, which has more people of higher ages, sales are dependent more on newspaper advertisements. The value of R^2 and the lower value of MSE also indicate that the dependency relationship is more accurate. If the dependency relationship is obtained for all data without considering groups, then the result is obtained, as shown in Table 3. From Table 3, it can be inferred that the relationship without considering the group has having very high value of MSE and a lower value of R^2. This indicates that the sales dependencies are not well captured, and analysis done without considering the existence of groups in data is less accurate.

4.3. Experimentation Results on Real Data Set

There are various applications of this algorithm especially when data consists of multiple invisible groups which may have different dependency relationships between dependent and predictor variables. Considering the problem of predicting movie ratings, which is a measure of the excellence and popularity of a movie using the Internet Movie Database (IMDB). The data consists of various factors on which the rating of a movie depends. These factors are actor, director, writer, producer, and duration. In this case, different types of people may respond differently to the same movie which may depend on their age, gender, and region. So, in order to find the effect of various factors on movie ratings single equation may not be sufficient. Hence, we first mined different groups, and then, for each group, we obtained dependency relationships using linear regression. The sample data is shown in Table 4.

Table 2. Results with segmented regression on Customer Advertisement Data

Grp	Relation	MSE	R^2
1	Sales=(0.1330*$x1^1$) + (0.0987*$x2^1$) + (0.4974*$x3^1$) + (0.0001)	34.93	0.99
2	Sales=(2.9168*$x1^1$) + (1.6312*$x2^1$) + (0.0712*$x3^1$) + (0.0002)	48.29	0.99
3	Sales= (0.6902*$x1^1$) + (2.2498*$x2^1$) + (1.6666*$x3^1$) + (-0.0002)	96.76	0.98

Table 3. Results with generalized (unsegmented) regression on Customer Advertisement Data

Grp	Relation	MSE	R^2
1	Sales= (0.9953*$x1^1$) + (0.6083*$x1^1$) + (0.8978* $x1^1$) + (0.0707)	17795.51	0.8644

Table 4. IMDB Data

Sr.No.	X1	X2	X3	X4	Y
1	6.52	6.34	6.93	101	6.8
2	5.96	6.34	6.28	91	7.5
3	7.03	7.5	6.48	113	7.4

Abbreviations X1 ... X4, and Y in the above figure represent the advertisements in the following factors.

X1 – Director, X2 – Producer, X3 – Writer, X4 – Duration, Y – Rating

Table 5. Results with segmented regression on the IMDB dataset

Grp	Relation	MSE	R^2
1	Rating=(1.2465*$x1^1$) + (0.4723*$x2^1$) + (0.0708*$x3^1$) + (0.0213*$x4^1$)	0.35	0.95
2	Rating=(-0.1375*$x1^1$) + (0.4182*$x2^1$) + (0.2333*$x3^1$) + (0.0234*$x4^1$)	0.35	0.93

Table 6. Results with generalized (unsegmented) regression on IMDB Dataset

Grp	Relation	MSE	R^2
1	Rating=(-.1375*$x1^1$) + (0.4182*$x1^1$) + (0.2333*$x1^1$) + (0.0234*$x1^1$)	67.33	0.5044

Here director, producer and writer are represented using a grade score out of 10 depending upon their success from the IMDB database. The algorithm mines the desired number of groups, and for each group dependency relationship is obtained using linear regression.

The dependency relationship obtained by mining groups has having very low MSE and a high value of $R^{2,}$ so it provides good accuracy for prediction. The results are shown in Table 5. If groups are not mined and the dependency relationship is obtained, then the results obtained are shown in Table 6.

5. Proof of Concept

To confirm the variation in the user response (to sales advertisement) with respect to different age groups of people on different advertisement platforms, we generated a synthetic sales dataset by normalization and cleaning the web-crawled advertisement data and then dividing the dataset based on age attribute. The results confirm the fact that different age groups of people (mentioned in section 4.1) are responding differently to different advertisement media. For different groups of users based on age attributes, we derived a correlation of sales with the various advertisement platforms. When we run our linear regression model on the cluster of Group-1 users, we obtain an accuracy score of 0.8829524241395408. For Group 1, the correlation of sales is highest on social media marketing, followed by TV marketing (Figure 1). This is also reflected in the correlation values of sales with different media parameters, which are reported as SocialMedia: 0.880576, TV: 0.861283, Radio: 0.344886, Newspaper: 0.222601. This clearly indicates the high popularity and following of social media and TV amongst the users of Group 1.

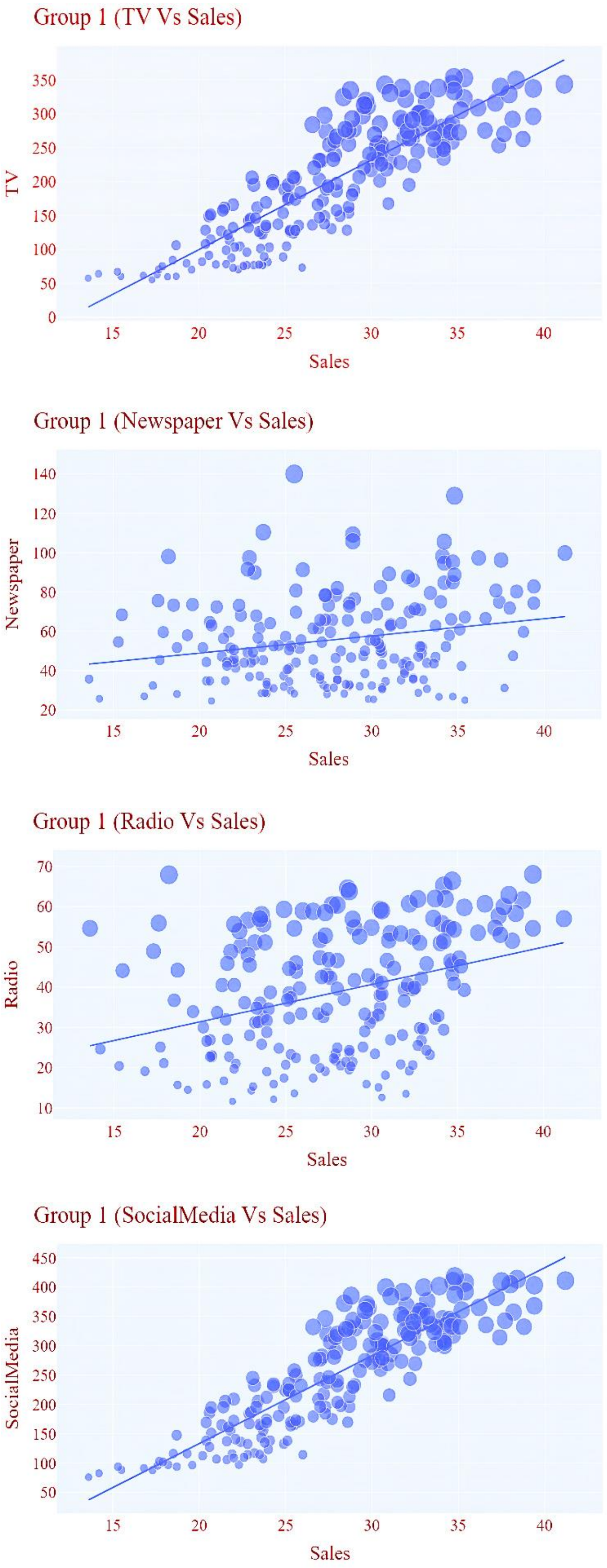

Fig. 1 Dependency Relationship of Sales with different Advertisement media for Group-1 Users (15 - 25 age)

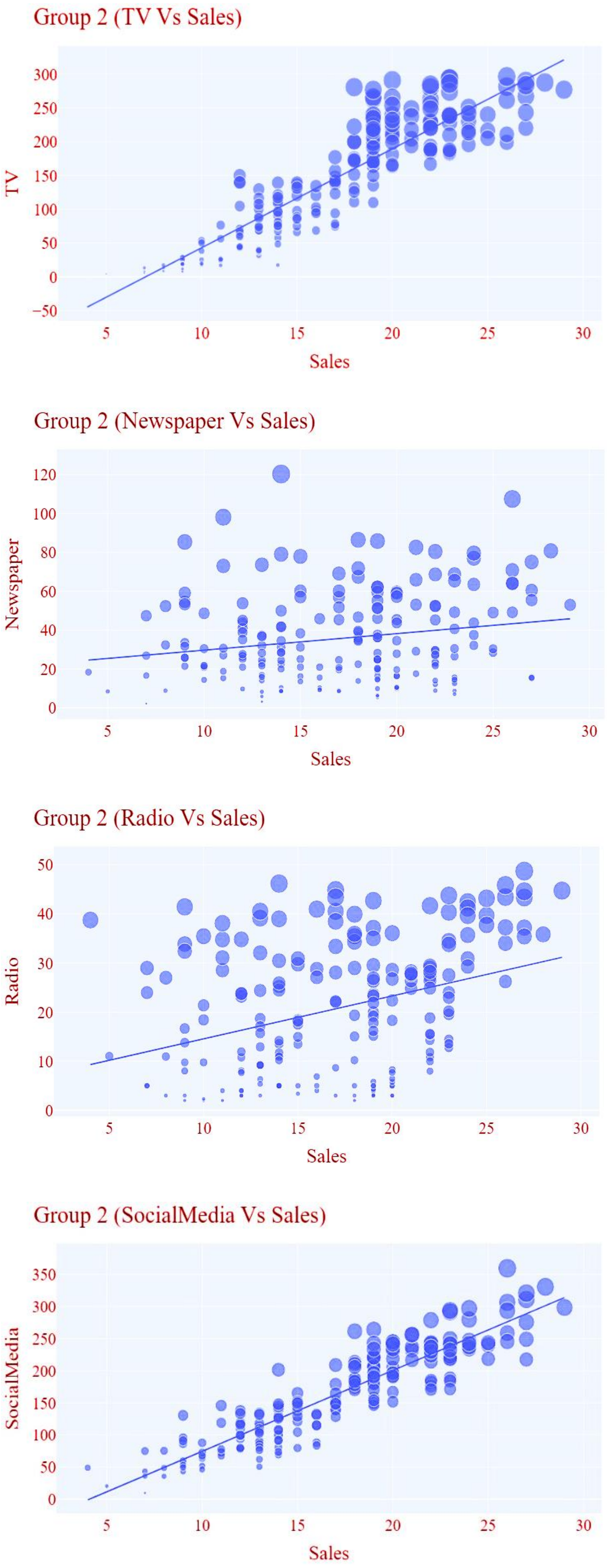

Fig. 2 Dependency Relationship of Sales with different Advertisement media for Group-2 Users (26 - 40 age)

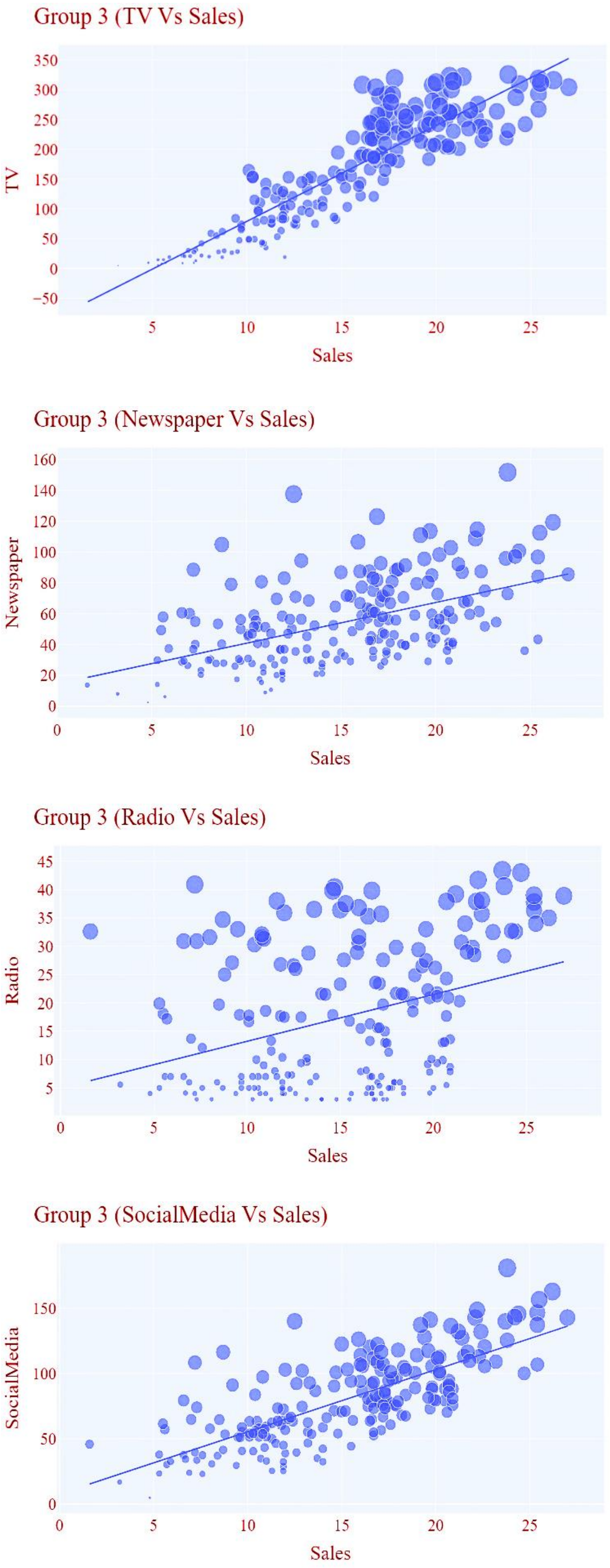

Fig. 3 Dependency Relationship of Sales with different Advertisement media for Group-3 Users (41 and above age)

In Group-2 users, TV marketing and social media marketing has a high correlation with sales, whereas sales attributes are less correlated with newspaper marketing and radio marketing attributes. This can be concluded that in this group of users, TV marketing and social media marketing are creating a high positive impact on sales.

Results can be interpreted as newspaper marketing, and radio marketing is nonsuitable or, outdated or nonproductive marketing media for this group of users (Figure 2). For Group 2, the accuracy score observed is 0.9010689815682468, whereas correlation values of sales with different media parameters are TV: 0.898809, SocialMedia: 0.909112, Newspaper: 0.196335, Radio: 0.341401.

For Group-3 users, the accuracy score observed is 0.907495179754958, whereas correlation values of sales with different media parameters are TV: 0.901208, SocialMedia: 0.753564, Newspaper: 0.511181, Radio: 0.351000. For Group 3, users' highest sales correlation is on TV marketing, followed by social media marketing. However, considering the 0.51 correlation value, we can conclude that Newspaper marketing is slightly relevant to this group of users (Figure 3).

6. Conclusion

Finding accurate dependency relationships from the social and economic data is a challenging problem. As different customers respond to different advertisement media and different promotional schemes differently, it is hard to represent the dependency relationship among such data by a single regression line. Representing such a relationship by polynomial regression is complex, and accurate results are not achieved.

Therefore, one solution is to group the data based on similarity measures on the most relevant attribute(s) and find a dependency relationship for each group. We mean that there exist clusters within the dataset, and each cluster has its own individual dependency relationship. Therefore, the idea is to first divide the data into the required number of clusters and then find the dependency relation for each cluster. Traditional clustering algorithms find similarities on the basis of spherical distance, but our data is not isotropic.

Hence, the idea is to cluster the data on the basis of their dependency relationship. This algorithm can handle large data as we first divide the sample data points, it handles the outlier points and checks if there is a relationship between those points which don't belong to any cluster. As it can be seen through experimentation this algorithm gives accurate results which R-square values of generalized regression and segmented regression algorithms can analyze.

We have focused on predicting IMDB ratings of the movie as different people may respond differently to different movies, the algorithm is used for grouping customer data based on dependency relationships.

We get accurate results as it is verified by the R^2 value, which is 0.55 for the generalized regression algorithm and 0.97 for the segmented regression algorithm. Higher R^2 indicates clearly that the model fits the data more accurately in the case of a segmented regression algorithm.

Acknowledgments

All the authors are grateful to Dr. Parag. S. Deshpande, Professor & Head of CSE Department, Visvesvaraya National Institute of Technology, Nagpur, for his valuable guidance for this work.All the authors are thankful to Miss. Rupal Chavan, Student, VNIT, Nagpur for being a major contributor in implementing designed algorithms and helping us several times during our research work.

References

[1] Peter Duchessi, and Eitel J.M. Lauría, "Decision Tree Models for Profiling Ski Resorts' Promotional and Advertising Strategies and the Impact on Sales," *Expert Systems with Applications,* vol. 40, no. 15, pp. 5822–5829, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[2] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim, "Efficient Algorithms for Mining Outliers from Large Data Sets," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 427-438, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[3] Sergios Theodoridis et al., *Introduction to Pattern Recognition: A Matlab Approach,* Academic Press, 2010. [Google Scholar] [Publisher Link]

[4] Rui Xu, and Donald Wunsch, "Survey of Clustering Algorithms," *IEEE Transactions on Neural Networks,* vol. 16, no. 3, pp. 645-678, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[5] Kristina P. Sinaga, and Miin-Shen Yang, "Unsupervised K-means Clustering Algorithm," *IEEE Access,* vol. 8, pp. 80716-80727, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Ki Hyun Kim, and Jun Geol Baek, "A Prediction of Chip Quality using OPTICS (Ordering Points to Identify the Clustering Structure)-Based Feature Extraction at the Cell Level," *Journal of Korean Institute of Industrial Engineers,* vol. 40, no. 3, pp. 257-266, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[7] Michael Steinbach, George Karypis, and Vipin Kumar, "*A Comparison of Document Clustering Techniques,*" Technical Reports, Department of Computer Science and Engineering, University of Minnesota, 2000. [Google Scholar] [Publisher Link]

[8] Ana L.N. Fred, and Anil K. Jain, "Data Clustering using Evidence Accumulation," *International Conference on Pattern Recognition*, IEEE, Quebec City, QC, Canada, vol. 4, pp. 276-280. 2002. [CrossRef] [Google Scholar] [Publisher Link]

[9] Pavel Berkhin, "A Survey of Clustering Data Mining Techniques," *Grouping Multidimensional Data: Recent Advances in Clustering,* Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 25-71, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[10] Alaa Tharwat, "Parameter Investigation of Support Vector Machine Classifier with Kernel Functions," *Knowledge and Information Systems*, vol. 61, pp. 1269-1302, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Manuel Fernández-Delgado et al., "An Extensive Experimental Survey of Regression Methods," *Neural Networks*, vol. 111, pp. 11-34, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Sarah Mussol, Philippe Aurier, and Gilles Séré de Lanauze, "Developing In-Store Brand Strategies and Relational Expression through Sales Promotions," *Journal of Retailing and Consumer Services,* vol. 47, pp. 241-250, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[13] Hao Yu et al., "Self-Paced Learning for K-Means Clustering Algorithm," *Pattern Recognition Letters,* vol. 132, pp. 69-75, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[14] Adil M. Bagirov, Arshad Mahmood, and Andrew Barton, "Prediction of Monthly Rainfall in Victoria, Australia: Clusterwise Linear Regression Approach," *Atmospheric Research,* vol. 188, pp. 20-29, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[15] Hui Jiang, Zhaosheng Feng, and Guirong Jiang, "Dynamics of an Advertising Competition Model with Sales Promotion," *Communications in Nonlinear Science and Numerical Simulation,* vol. 42, pp. 37-51, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[16] Dai, Wenseng, Yang-Yu Chuang, and Chi-Jie Lu, "A Clustering-Based Sales Forecasting Scheme Using Support Vector Regression for Computer Server," *Procedia Manufacturing,* vol. 2, pp. 82-86, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[17] Buil, Isabel, Leslie De Chernatony, and Eva Martínez. "Examining the Role of Advertising and Sales Promotions in Brand Equity Creation," *Journal of Business Research,* vol. 66, no. 1, pp. 115-122, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[18] Wendy Attaya Boland, Paul M. Connell, and Lance-Michael Erickson, "Children's Response to Sales Promotions and Their Impact on Purchase Behavior," *Journal of Consumer Psychology,* vol. 22, no. 2, pp. 272-279, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[19] Yongrui Duan, Tonghui Liu, and Zhixin Mao, "How Online Reviews and Coupons Affect Sales and Pricing: An Empirical Study Based on E-Commerce Platform," *Journal of Retailing and Consumer Services,* vol. 65, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[20] Tomáš Formánek, and Ondřej Sokol, "Location Effects: Geo-Spatial and Socio-Demographic Determinants of Sales Dynamics in Brick-and-Mortar Retail Stores," *Journal of Retailing and Consumer Services,* vol. 66, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[21] Daifeng Li et al., "A Multiple Long Short-Term Model for Product Sales Forecasting based on Stage Future Vision with Prior Knowledge," *Information Sciences*, vol. 625, pp. 97-124, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Hui Jiang, Zhaosheng Feng, and Guirong Jiang, "Dynamics of an Advertising Competition Model with Sales Promotion," *Communications in Nonlinear Science and Numerical Simulation*, vol. 42, pp. 37-51, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[23] Harlan E. Spotts et al., "The Role of Paid Media, Earned Media, and Sales Promotions in Driving Marcom Sales Performance in Consumer Services," *Journal of Business Research*, vol. 152, pp. 387-397, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[24] Kapil Bawa, and Robert W. Shoemaker, "The Effects of a Direct Mail Coupon on Brand Choice Behavior," *Journal of Marketing Research*, vol. 24, no. 4, pp. 370-376, 1987. [CrossRef] [Google Scholar] [Publisher Link]

[25] Sunil Gupta, "Impact of Sales Promotions on When, What, and How Much to Buy," *Journal of Marketing Research*, vol. 25, no. 4, pp. 342-355, 1988. [CrossRef] [Google Scholar] [Publisher Link]

[26] Zhilin Yang, and Robin T. Peterson, "Customer Perceived Value, Satisfaction, and Loyalty: The Role of Switching Costs," *Psychology & Marketing*, vol. 21, no. 10, pp. 799-822, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[27] Richa Agrawal, Sanjaya S. Gaur, and Archana Narayanan, "Determining Customer Loyalty: Review and Model," *The Marketing Review*, vol. 12, no. 3, pp. 275-289, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[28] Tor Wallin Andreassen, and Bodil Lindestad, "Customer Loyalty and Complex Services: The Impact of Corporate Image on Quality, Customer Satisfaction and Loyalty for Customers with Varying Degrees of Service Expertise," *International Journal of Service Industry Management*, vol. 9, no. 1, pp. 7-23, 1998. [CrossRef] [Google Scholar] [Publisher Link]