Performance Comparison of ML Regression Algorithms in Predicting Supermarket Sales

Balaji Jayakrishnan, Gunja Pandey

Nitika Verma, Ritika Sarkar

Muskan Dhingra, Palak Tande

Vellore Institute of Technology, Chennai

Corresponding author: Ritika Sarkar, Email: ritika.sarkar2019@vitstudent.ac.in

The ability of regression algorithms to reliably identify the influencing factors of any data on the desired result is irrefutable. With the available techniques, we can investigate the main reason behind the influence of distinguishing factors on a supermarket's sales. We'll be building a machine learning model that can accurately predict the sales in millions of units for a given product. Our work will investigate the ability of some of the most popular ML regression algorithms to provide this information. Seven regression algorithms will be trained using data collected through supermarket sales. To gain key insights, the algorithms are compared along two axes, prediction quality and usefulness of output. This class of algorithms produces models that can be used to predict performance in sales and indicate the sources of potential market problems and quantify the potential gain.

Keywords: Machine Learning, Regression Algorithm, Supermarkets, Sales Analysis, Prediction.

2021. In Raju Pal & Praveen K. Shukla (eds.), *SCRS Conference Proceedings on Intelligent Systems*, 181–187. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-93-91842-08-6-16

1 Introduction

Analyzing algorithms and their performance in various applications is an important research topic across the Computer Science domain. It helps us pinpoint reasons for the success or failure of these algorithms, and also gives us sufficient data to carry out further investigations on other related or unrelated applications as well. Machine Learning is one of the most popular technologies in today's era that one would have ever come across. To capitalize on the benefits of this approach, we attempt to model it to precisely predict the sales in millions of units in any supermarket. We will be working with machine learning algorithms; we had many algorithms to choose from given the diverse nature of ML algorithms. There are a plethora of pros and cons for every algorithm. Although one algorithm may not always be better than another, there are certain features of every algorithm that we will use as a guide in choosing the right one quickly by tuning the hyper parameters. In the following sections, we will be going through some standard algorithms used for regression and discuss their application and usefulness for our problem.

1.1 Literature Survey

In [1] multiple linear regressions have been used to predict the market values of players in the 2017-2018 seasons. They achieved a good model with 0.86 R squared score using 52 attributes. The limitation is that they could not remove multi-collinearity. In [2] a probability-based approach to perform linear regression is presented. The merits of this paper include the presentation of a marketing application helpful for evaluating the performance of trade show and the identification of potential future research applications. The cross validation and external validation in order to compare models using the average annual concentrations of NO2 in [3]. It was observed that different algorithms having statistical foundations performed identically when spatial variation in the annual average pollution was modelled. In [4], they have discussed the simulation process for evaluating the characteristic of Bayesian Ridge regression parameter estimates. This method gives better performance for small size of the sample, and uses weights for as the dependent variables. This study [5] explores the impacts of congestion in the occurrence of crash in urban expressways. The merits include the development of multiple measures for combating congestion. The limitation was that multi-collinearity was observed, during certain time period which was further solved through ridge regression. The objective of this paper [7] is based on fuzzy neural networks and attempts to derive learning algorithms from them. A useful finding presented here is that this method is better in estimation than squared error sums. The article [11] is devoted to the contrast of Ridge and LASSO estimators. The authors have used custom-generated data to investigate the advantages of every two regression analysis methods. All the necessary calculations are performed by the R software for arithmetic computing. Elastic net regression is a mixture of LASSO and Ridge Regression techniques toward which numeric, categorical, and image outline data can be specified to the regression. This paper [14] describes a method of finding confidence information about predictions. Results on artificial datasets show that the confidence intervals generated by the technique are robust to changes in the amount of noise in the data and across different parameters for the underlying regression. This paper [15] has provided an efficient solution to the sparsity problem in logistic regression.

2 Methods

We have tried to implement the algorithms which are going to look into the capacity of a number of the most famous Machine Learning regression algorithms. For our regression models, we can use the fit-predict method of the scikit-learn library. We have implemented Lasso Regression, Decision Tree

Regressor, K Neighbors Regressor, Ridge Regression, Polynomial Regression, Bayesian regression and Linear Regression to offer this information. As these Machine Learning algorithms come under the class of supervised learning, we can find the relationship between sales (target) and the other attributes (predictor variables) by minimizing the prediction error and hence, maximize the prediction accuracy. We have used the R2 score as our metric for indicating the efficiency of the model.

2.1 Data Collection

For the purpose of this research, the data used has been obtained from an open-source data repository. The dataset contains sales data of Superstore in Canadian locations (provinces / regions) with details about the customer, postal region, products, sales and profit.

2.2 Dataset Description

We have used an open-source data set that has been collected by a supermarket with details of their sales. Our dataset consists of 19 different attributes namely- Order ID, Order Date, Product ID, Product Name, Customer ID, Customer Name, Shipment Mode, Segment, Country, State, City, Region, Postal code, Category, Sub-Category, Quantity, Discount, Profit and Sales. It is a large dataset consisting of 9994 row entries.

Figure 1 shows the various columns present in our dataset with their data types, null value counts, and the total number of entries present for each column. Such visualization helped us to plan our future approach easily and efficiently, by removing irrelevant columns not required for our study to predict the sales such as Order ID, Customer ID, Customer Name, etc.

```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Order ID        9994 non-null   object
 1   Order Date      9994 non-null   datetime64[ns]
 2   Ship Date       9994 non-null   datetime64[ns]
 3   Ship Mode       9994 non-null   object
 4   Customer ID     9994 non-null   object
 5   Customer Name   9994 non-null   object
 6   Segment         9994 non-null   object
 7   Country         9994 non-null   object
 8   City            9994 non-null   object
 9   State           9994 non-null   object
 10  Postal Code     9994 non-null   int64
 11  Region          9994 non-null   object
 12  Product ID      9994 non-null   object
 13  Category        9994 non-null   object
 14  Sub-Category    9994 non-null   object
 15  Product Name    9994 non-null   object
 16  Sales           9994 non-null   float64
 17  Quantity        9994 non-null   int64
 18  Discount        9994 non-null   float64
 19  Profit          9994 non-null   float64
dtypes: datetime64[ns](2), float64(3), int64(2), object(13)
```

Fig. 1. Dataset information

2.3 Data Preprocessing

Several columns were dropped from the dataset due to their irrelevance. A correlation analysis was performed between the columns in our dataset and irrelevant columns were eventually dropped. The attributes eventually under study are 'Segment', 'Region', 'Quantity', 'Discount', 'Category', 'Sub-Category', 'Profit' as predictor variables and 'Sales' attribute as the target variable. Furthermore, since a lot of columns have string data and the machine learning algorithms require integer data for processing, therefore the columns containing string data are encoded by creating dummy variables for each unique value in the columns with a string value.

Algorithms under study

We have deployed 7 regression algorithms for our comparative analysis, namely, Linear Regression, Ridge Regression, Lasso Regression, Polynomial Regression, Baye's Ridge Regression, Decision Tree Regressor, and K Neighbors Regressor. All the algorithms were trained using the dataset obtained after pre-processing as specified above and they were finally compared alongside each other using R-squared score as the evaluation parameter. The statistical measure, R-squared (R2) , determines the proportion of variance for a variable which is dependent, that in turn, is described by a(n) independent variable(s) in a regression model.

Figure 2 shows the stepwise methodology employed in the analysis of the 7 regression algorithms in this paper using the dataset.

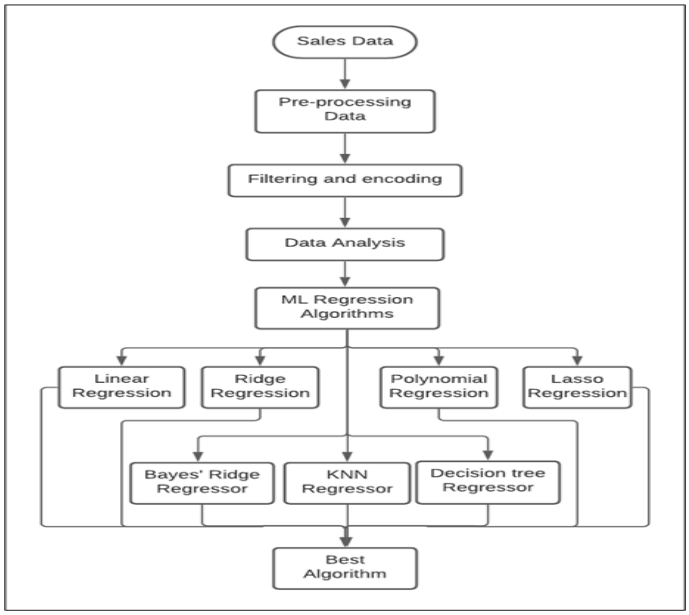

Fig. 2. Schema of experimental methodology.

3 Results

Table 1. R2 score for regression algorithms.

Serial Number	Algorithm Name	R2-Score
1	Linear Regression	0.384600
2	Ridge Regression	0.384605
3	Polynomial Regression	0.384617
4	Lasso Regression	0.384600
5	Bayes' Ridge Regression	0.630116
6	KNeighbors Regressor	0.503177
7	Decision Tree Regressor	0.630116

Table 1 shows the R2-score of the algorithms under study. We conclude that Baye's Ridge Regression performs the best amongst the traditional regression algorithms and Decision Tree Regressor (CART), a modern regression algorithm derived from classifier algorithm, provides a similar R2 score as Baye's Ridge Regression proving the efficiency of traditional algorithms as well as new and modified algorithms in the predictions made on our dataset.

The 7 algorithms were compared using a box plot. 5-fold cross validation has been performed to achieve a better model while avoiding over fitting. From figure 3, we see that Baye's model and Decision tree perform the best but Decision tree model has some outliers in the negative region due to over-fitting in this model, hence Baye's Bridge Regression performs the best in this case.

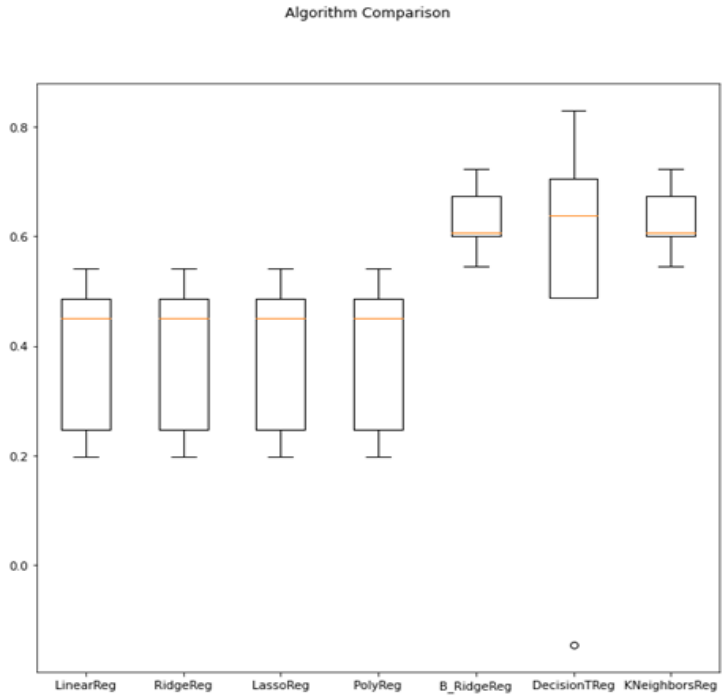

Fig. 3. Results of 5-fold cross validation

4 Discussion

4.1 Conclusion

The best algorithm was Bayes Ridge Regression after the performance analysis of ML regression algorithms in predicting supermarket sales. Even though the R2-score of Decision Tree Regression was the same as that of Bayes' Ridge Regression, the outliers present in the analysis due to over fitting in the Decision tree model make it a less likely choice in the case of our dataset. We can safely say that our dataset cannot be modeled properly by linear algorithms like linear regression, lasso, and ridge, as well as polynomial regression. With our study, we have predicted sales and by using this model and

analyzing the final cleaned dataset, the potential gain of the supermarket can be derived as well as the factors that lead to potential market problems.

4.2 Limitations of Research

The proposed method was performed using an already available dataset in which the values were obtained using the sales of a specific region in America. In order to increase the accuracy of our analysis and prediction in the proposed system, we plan to validate our approach by using actual data collected under proper observations using the system architecture proposed in this paper and use that in our prediction for predicting sales of supermarkets in India. With the continuous growth in the field of Machine learning, we believe soon there will be an incorporation of these methodologies in all areas of business and we plan to keep modifying our model as new and better technology arises in the ever-growing field of scientific research.

References

[1] Y. Kologlu, H. Birinci, S. I. Kanalmaz and B. Ozyilmaz, "A Multiple Linear Regression Approach For Estimating the Market Value of Football Players in Forward Position", *arXiv Appl.*, 2018.

[2] W. S. DeSarbo and W. L. Cron, "A maximum likelihood methodology for clusterwise linear regression", *J. Classific.*, vol. 5, no. 2, pp. 249-282, 1988.

[3] J. Chen et al., "A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide", *Env. Int.*, vol. 130, 104934, 2019.

[4] A. Efendi and Effrihan, "A simulation study on Bayesian Ridge regression models for several collinearity levels", in *AIP Conference Proceedings*, 2017, pp. 020031.

[5] Q. Shi, A. M. Abdel and J. Lee, "A Bayesian ridge regression analysis of congestion's impact on urban expressway safety", *Accident Anal. Preven.*, vol. 88, pp. 124-137, 2016.

[6] E. Burnaev and V. Vovk, "Efficiency of conformalized ridge regression", in *Conf. on Learning Theory*, 2014, pp. 605-622.

[7] M. Mosleh, M. Otadi and S. Abbasbandy, "Fuzzy polynomial regression with fuzzy neural networks", *Appl. Math. Model.*, vol. 35, no. 11, pp. 5400-5412, 2011.

[8] B. Sun, H. Liu, S. Zhou and W. Li, "Evaluating the performance of polynomial regression method with different parameters during color characterization", *Math. Prob. Eng.*, Id 418651, 2014.

[9] O. Giustolisi and D. A. Savic, "A symbolic data-driven technique based on evolutionary polynomial regression", *J. Hydro Inform.*, vol. 8, no. 3, pp. 207-222, 2006.

[10] R. Tibshirani, R, "Regression shrinkage and selection via the lasso: a retrospective", *J. Royal Statist. Society: Series B (Statistical Methodology)*, vol. 73, no. 3, pp. 273-282, 2011.

[11] M. Anjali, M. R. Shivhare and M. K. P. M. M. Dixit, "A Survey ON REGRESSION ESTIMATE WITH LASSO METHOD", *Int. J. Adv. Tech. Eng. Res.*, vol. 8, no. 2, pp. 18-23, 2018.

[12] O. Kohannim et al., "Discovery and replication of gene influences on brain structure using LASSO regression", *Frontiers in Neuroscience*, vol. 6, 115, 2012.

[13] C. Saunders, A. Gammerman and V. Vovk, "Ridge regression learning algorithm in dual variables", in *Proceeding of the 15th Int. Conf. on Machine Learning,* 1998, pp. 515-521.

[14] I. Nouretdinov, T. Melluish, and V. Vovk, "Ridge regression confidence machine" in *Proceeding of the 18th Int. Conf. on Machine Learning*, 2001, pp. 385-392.

[15] S. K. Shevade and S. S. Keerthi, "A simple and efficient algorithm for gene selection using sparse logistic regression", *Bioinformatics*, vol. 19, no. 17, pp. 2246-2253, 2003.

[16] T. Mythili, D. Mukherji, N. Padalia and A. Naidu, "A heart disease prediction model using SVM-Decision Trees-Logistic Regression (SDL)", *Int. J. Comp. Appl.*, vol. 68, no. 16, pp. 11-15, 2013.

[17] Z. Bursac et al., "Purposeful selection of variables in logistic regression", *Source code for biology and medicine*, vol. 3, no. 1, pp. 1-8, 2008.

[18] P. Harrington, *Machine learning in action,* Simon and Schuster, 2012.

[19] X. D. Zhang, "Machine learning", in *A Matrix Algebra Approach to Artificial Intelligence*, Springer, Singapore, 2020. pp. 223-440.

[20] Sammut, Claude and G. I. Webb, Eds. Encyclopedia of machine learning. Springer Science & Business Media, 2011.

[21] E. Naqa, Issam,and M. J. Murphy, "What is machine learning?", in *Machine Learning in Radiation Oncology,* Springer, Cham, 2015, pp. 3-11.
[22] Murphy and P. Kevin, *Machine learning: A Probabilistic Perspective*, MIT Press, 2012.
[23] G. Carleo et al., "Machine learning and the physical sciences", *Rev. Modern Phy.*, vol. 91, no. 4, 045002, 2019.