iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5329

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

An Effective Approach for Sales Forecasting

Mr. Faraz Hariyani [1], Mrs. Haripriya V [2]
[1]*MSc IT Student, Dept of MSc IT, JAIN (Deemed-to-be University), INDIA*
[2]*Assistant Professor. Department of MSc IT, JAIN (Deemed-to-be University)*

Abstract: Sales Forecasting plays an important role in business. It provides appropriate and dependable information about the past and present events and likely future events. Nowadays, predicting sales has become a common method but not many known approaches are applied. Sales forecasting works as a vital data input to organizational development. Machine Learning techniques are very effective tools in extracting hidden knowledge from vast dataset to boost the precision and effectiveness of estimating sales. Supervised Machine Learning is being used by many organizations to identify and solve business problems. The proposed methodology is to use regression to do a comparative study on sales forecasting. Here, three regression techniques, which are Linear Regression, Ridge Regression, and LASSO Regression are discussed in detail. The accuracy in sales prediction offers a big positive impact on business. All the regression techniques are giving excellent outcome for R-squared and RMSE value. The perfect result would be an RMSE value of zero and R-squared value of 1, but that's almost impossible in real economic datasets.

Keywords: Regression, Lasso, Ridge, Linear, Sales Forecasting

I. INTRODUCTION

Big data is defined by the size of a dataset. Data sets are mostly vast measuring tens of terabytes. Big data analytics is the process of mining useful information by evaluating different types of data sets. It is used to find hidden patterns, market trends, consumer preferences, and a lot more for benefiting organization's decision making.

Businesses are concentrating more on agility and uniqueness rather than consistency and implementing the big data technologies help the companies achieve that in no time. Big data analytics has not only allowed the firms to stay updated with the changing aspects but has also let them predict the future trends, sales giving a competitive edge.

Patterns can then be found in historic and transactions of data and can be used to categorize risks and opportunities in the future. Main role of predictive analytics in relation to predicting is models can find relationships among various factors and measure risk with a specific set of detailed conditions, and then allocate a score to the risk assessment. When this is used business get two outcomes:

1) Complete Coverage involves considering all data sources, combining it all to give an accurate prediction.
2) Detecting changes, using Machine Learning businesses can determine the repeating patterns in the enormous datasets. All the business organizations are alerted if any change is detected so as to improve their performance.

"Any sales prediction should be thought of as a best guess about customer demand for a vendor's goods or services, during a particular time horizon, given a set of assumptions about the environment." (Moon & Mentzer, 1999)

Sales forecasting is, as the statement above tells, a best guess about customer demand for a vendor's goods in a particular time period. How this is made depends on whether one using a qualitative or a quantitative method. Its' purpose is to, as accurately as possible, try to predict what quantity of goods or services will be sold, and by doing that, try to decrease the costs for inventory and conveyance. A prediction works as a management control system and has almost the same attributes as a budget, although there are relevant differences between the two. A prediction can be expressed in both financial and physical units whereas a budget is expressed only in financial units. A prediction can be for any period and has not an obligation to meet the prediction outcomes.

A cost/benefit analysis is a useful instrument when deciding the timeliness of predictions within the organization. A constant estimating schedule cost money and if they are not done correctly, the cost will exceed the benefits and it will therefore not prove to be a beneficial investment. Even though the predictions are made in a good way there might still be problems.

One of the big issues with the sales predictions is regarding the accuracy of it, especially if the organization sets its plans based on what the prediction predicts. Whether the organization uses judgmental techniques or a more statistical approach to prediction their sales, the important thing is that they areaccurate.

If an organization is collecting data in a particular pattern, and they want to add some new features in the existing data they have to wait for the financial year to end or they can purchase the missing data. Figure 1.1 shows how the existing data can still be used with the missing data, but it will affect the accuracy, it will also be more time consuming, it will require more assumptions and will have to model the missing data.

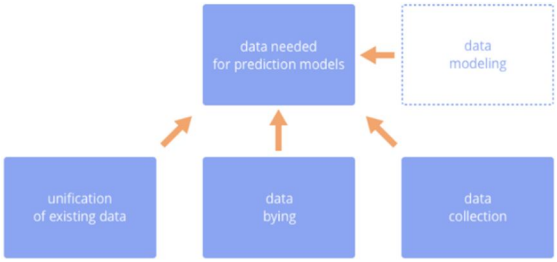


Figure 1.1 Working with Existing Data [15]

Some methods can predict the missing values of other variables with existing data on specific variables. Let's take an example, where an organization has its prices and sales for two years, and they have a history of its competitors' prices for 1.5 years, a simulation can determine the missing competitive prices. Classifiers foresee the missing values based on other independent values whose data is provided.

A Prediction model is used to get an estimate of the missing values, the data set has to be divided into two parts; first part being the existing data while the second being the missing data. The first part will become the train set, while the missing values in the second part become the forecast target.

II.LITERATURE SURVEY

Pavlyshenko [1] has used machine-learning models for sales predictive analytics. The main goal of his paper was to consider main approaches and case studies of using machine learning for sales forecasting. The accuracy on the validation set is an important indicator for choosing an optimal number of iterations of machine-learning algorithms. Author suggested as the next level model, Lasso regression can be used. Huang et.al [2] proposed a Dependency SCOR-topic Sentiment (DSTS) model to analyze the online textual reviews and predict sales performance. Their analysis is limited to online users who leave reviews at a Chinese review website. Hence, this analysis focuses on review texts written in Chinese. It would be interesting if future research expands the study to a global context. Sunitha et.al [3] said Intelligent Decision Analytical System requires integration of decision analysis and predictions. The results are summarized in terms of reliability and accuracy of efficient techniques taken for prediction and forecasting. The studies found that the best fit model is Gradient Boost Algorithm, which shows maximum accuracy in forecasting and future sales prediction. Loureiro et.al [4] used Decision Trees, Random Forest, Support Vector Regression, Artificial Neural Networks and Linear Regression. The model employing deep learning was found to have good performance to predict sales in fashion retail market, however for part of the evaluation metrics considered, it does not perform significantly. The results demonstrate that the use of DNN and other data mining techniques for performing sales forecasting in the fashion retail industry when there is no historical sales data is very promising. Nunnari et.al [5] have presented a case study concerning the forecasting of monthly retail time series recorded by the US Census Bureau from 1992 to 2016. The numerical results obtained, show that between the two neural networks approaches, the NF slightly outperform the NN models for the considered task. Kui Zhao et.al [6] have tested their approach on a real-world dataset from CaiNiao.com and it demonstrates strong performance. There are several interesting problems to be investigated in our further works: Is it possible to find the most important indicators for sales forecast from the raw log data by deep neural networks; It will be very appealing to find a unified framework for extracting features automatically from all types of data. Frank M. Thiesing et.al [7] have said that Neural networks trained with the back-propagation algorithm are applied to predict the future values of time series that consist of the weekly demand on item Their program runs a prototype and handles only a small subset of the supermarket's inventory. Future work will concentrate on the integration of our forecasting tool into the whole enterprise data flow process. Oliver Vornberger et.al [8] have said artificial neural networks are applied to a short-term forecast of the sale of articles in supermarkets. The times series of sales, prices and advertising campaigns are modelled to fit into feedforward multilayer perceptron networks that are trained by the backpropagation algorithm. The aim of their research is to develop a forecasting system for supermarkets. The system will reduce stock-keeping costs by flexible adaptability to changing circumstances. Efendigil et.al [9] presented a comparative forecasting methodology regarding too uncertain customer demands in a multi-level supply chain (SC) structure via neural techniques. Future research will perform various ANN

types and aforementioned neuro-fuzzy systems to make a similar approach. Doganis et.al [10] have said that for the food industry, successful sales forecasting systems can be very beneficial, due to the short shelf-life of many food products and the importance of the product quality which is closely related to human health. In a future study they will show how additional information, like price, promotions, etc. can be explicitly taken into account in the development of the time series model. Cerqueira et.al [11] presented an approach based on arbitrating, in which several forecasting models are dynamically combined to obtain predictions Finally, since diversity is a fundamental component in ensemble methods, they proposed a method for explicitly handling the inter-dependence between experts when aggregating their predictions. The main point for improvement is the scalability of the method. Papacharalampous et.al [12] conducted each simulation experiment twice; the first-time using time series of 100 values and the second time using time series of 300 values. The empirical investigation shows that in the given finite space, formed by simulated and annual river discharge time series, the no free lunch theorem is still satisfied. Abudu et.al [13] used two different types of monthly streamflow data (original and depersonalized data) were used to develop time series and Jordan-Elman ANN models using previous flow conditions as predictors. The forecasting models used by them presented in their study is limited to applications within the study basin and with observed specific conditions. Future enhancement should be conducted to identify potential predictors and build forecasting models using ANN models and/or other more advanced modeling methods, to improve forecasting. Ahmed et.al [14] considered multilayer perceptron models, Bayesian neural networks, radial basis functions, generalized regression neural networks (also called kernel regression), K-nearest neighbor regression, CART regression trees, support vector regression, and Gaussian processes. The study reveals significant differences between the different methods. The best two methods turned out to be the multilayer perceptron and the Gaussian process regression. In addition to model comparisons, they had tested different preprocessing methods and have shown that they have different impacts on the performance.

III.METHODOLOGY

The regression method of forecasting involves examining the relationship between two different variables, known as the dependent and independent variables. Suppose that you want to forecast future sales for your firm and you've noticed that sales rise or fall, depending on whether the gross domestic product goes up or down. Using statistical formulas, predictive analytics might predict the sales for a period of time. The regression method of forecasting means studying the relationships between data points, which can help you to:

1) Predict sales in the near and long term.
2) Understand inventory levels.
3) Understand supply and demand.
4) Review and understand how different variables impact all of these things.

A comparative study using the regression techniques will be studied here to predict the error rate of the business sales.

Projection of past sales approach is easy to implement. It's also the most used approach, and also a safe method for organizations involved in more or less steady industries. Though, in many cases, this is not reliable and also this can't be adopted in the case of new products or by uprising companies.

Products in use analysis, is based on two norms, firstly, the future marketplace for a product will vary in direct share to the quantity already in use, and secondly, it also gives an estimate that the present and past users of the product of a concern will continue to utilize the same in future.

Industry prediction and distribution of sales to the industry estimates its sales by applying a certain amount to the sales forecast of the entire industry. This method is very simple and also it is fast and is not costly.

Statistical demand analysis, the important aspects which are probably to cause variations in the sales, such as the population, disposable income in the hands of the people, the prices of the products, advertising programs, etc., are evaluated, and on the base of such an evaluation, sales estimation is made by a firm for its brand.

A. Regression

Regression analysis is one of the most vital fields for research in the area of ML and statistics. It is based on identifying for relations in between variables. It is a form of predictive modelling technique which examines the relationship between a dependent and independent variable (s). This method is used for forecasting, for example, relation between rash driving and number of road accidents by a driver is best examined through regression. There are multiple benefits of using regression analysis. They are as follows:

1) It indicates the significant relations between dependent and independent variable in the dataset used.
2) It indicates the strength of impact of multiple independent variables on a dependent variable.

Regression analysis also allows to compare the effects of variables measured on different scales, such as the effect of price changes and the number of publicity activities. These benefits help market researchers / data analysts / data scientists to remove and estimate the best set of variables to be used for developing predictive models.

B. Linear Regression

It is one of the most widely known modeling techniques. The flow of this technique is shown in Figure 3.1. In this technique, the dependent variable is continuous whereas the independent variable(s) can be continuous or discrete. It establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as the line of regression). It is represented by an equation,

Y = a + b * X + e [17]

where a is intercept, b is slope of the line and **e** is error term. This equation can be used to predict the value of target variable based on given predictor variable(s). It tries to bond a relationship between two variables by fitting a linear equation to observed data, where, one variable is considered to be a descriptive variable, and the other a dependent variable.

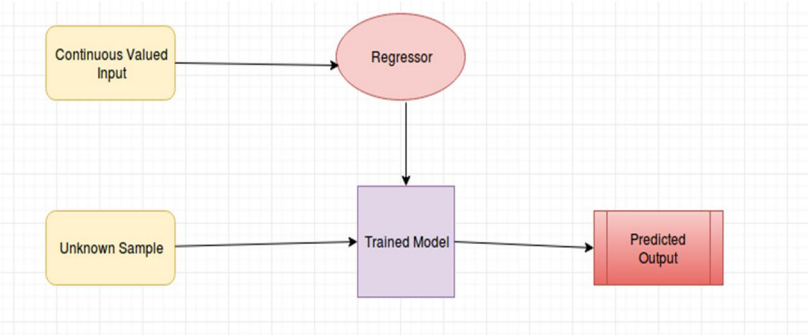


Figure 3.1 Linear Regression Model [18]

C. Ridge Regression

It performs L2 regularization where in it adds a factor of the sum of squares of coefficients in the optimization objective. It is an extension of linear regression where the loss function is improved to minimize the complexity of the model. This alteration is done by adding a penalty parameter that is correspondent to the square of the magnitude of the coefficients.

Loss function = OLS + alpha * summation (squared coefficient values) [17]

Above, Ordinary least squares (OLS) method, works by minimizing the sum of squares of residuals (actual value - predicted value) and loss function, alpha is the parameter we need to select. A low alpha value can lead to over-fitting, whereas a high alpha value can lead to under-fitting.

D. LASSO Regression

Lasso regression, or the Least Absolute Shrinkage and Selection Operator, is also an alteration of linear regression. Here there are 2 key words with utmost importance – 'absolute' and 'selection'.

It performs L1 regularization, i.e., it adds a factor of sum of absolute value of coefficients in the optimization objective.

Here, the loss function is modified to minimize the complexity of the model by limiting the sum of the absolute values of the model coefficients (also called the l1-norm).

The loss function for Lasso Regression can be expressed as below:

Loss function = OLS + alpha * summation (absolute values of the magnitude of the coefficients) [17]

Above, Ordinary least squares (OLS) method, works by minimalizing the sum of squares of residuals (actual value - predicted value) and loss function, alpha is the penalty parameter which needs to be selected. Using an L1 norm constraint forces some weight values to zero to allow other coefficients to take non-zero values.

Lasso Regression holds all the norms of the Linear Regression, such as:

1) The response variable is normally distributed.
2) There is a linear relationship between the response and the explanatory variables.
3) The random errors are normally distributed, have constant (equal) variances at any point in X, and are independent.

IV.RESULTS AND DISCUSSION

The structure of the dataset containing – 1500000 observations of 14 variables, and giving the summary statistics of the variables is shown in Figure 5.1. Also, there is no missing values because all the variables have 1500000 as 'count' which is equal to the number of records in the dataset.

```
In [2]: dataset = pd.read_csv("C:/Users/Faraz/Desktop/RESEARCH_PYTHON/newdataset.csv")
   ...: print(dataset.shape)
   ...: dataset.describe()
(1500000, 14)
Out[2]:
            Order ID    Units Sold      ...         Total Cost  Total Profit
count  1.500000e+06  1.500000e+06      ...       1.500000e+06  1.500000e+06
mean   5.500681e+08  4.999305e+03      ...       9.374880e+05  3.923999e+05
std    2.599834e+08  2.885556e+03      ...       1.149108e+06  3.789181e+05
min    1.000012e+08  1.000000e+00      ...       6.920000e+00  2.410000e+00
25%    3.246863e+08  2.501000e+03      ...       1.618176e+05  9.506640e+04
50%    5.497922e+08  4.998000e+03      ...       4.673575e+05  2.813704e+05
75%    7.756279e+08  7.498000e+03      ...       1.196572e+06  5.654252e+05
max    9.999999e+08  1.000000e+04      ...       5.249600e+06  1.738700e+06

[8 rows x 7 columns]
```

Figure 4.1 Displaying details of orginal dataset.

The training set of independent variables after splitting the data, as shown in Figure 4.2. Here 'x_train' data is displayed which contains Sales, Quantity, Discount, subcategory, productname columns. It contains 1125000 rows and 6 columns.

```
In [8]: x_train
Out[8]:
         UnitPrice  UnitCost  TotalRevenue   TotalCost  region  item
962323        9.33      6.92      59292.15    43976.60       3     0
1127992     154.06     90.93     164844.20    97295.10       5     9
1064282      81.73     56.67     105758.62    73330.98       0     7
566879      154.06     90.93    1096290.96   647057.88       0     9
278004       47.45     31.79      73262.80    49083.76       0     3
969855      437.20    263.33    3524706.40  2122966.46       1     5
1070545     421.89    364.69     683039.91   590433.11       3     2
429241      109.28     35.84     987563.36   323886.08       5     1
412429      205.70    117.11     551481.70   313971.91       0    11
714140      154.06     90.93     782008.56   461560.68       0     9
1191654     668.27    502.54    1301789.96   978947.92       2     8
1438814     421.89    364.69     297432.45   257106.45       0     2
843786       47.45     31.79     414760.45   277876.39       4     3
1036478     205.70    117.11    1065937.40   606864.02       6    11
841907      109.28     35.84     906696.16   297364.48       3     1
474571      154.06     90.93     605917.98   357627.69       4     9
1337395     437.20    263.33    2883334.00  1736661.35       4     5
562332      668.27    502.54     370221.58   278407.16       5     8
1158781     255.28    159.42    1890348.40  1180505.10       1    10
1235074     154.06     90.93     531661.06   313799.43       3     9
```

Figure 4.2 Training set of x

Displaying 'y_train' data after splitting which contains only Profit column, as shown in Figure 4.3. It has TotalProfit column from the dataset and contains 1125000 columns.

```
In [10]: y_train
Out[10]:
962323        15315.55
1127992       67549.10
1064282       32427.64
566879       449233.08
278004        24179.04
969855      1401739.94
1070545       92606.80
429241       663677.28
412429       237509.79
714140       320447.88
1191654      322842.04
1438814       40326.00
843786       136884.06
```

Figure 4.3 Training set of y

The below output shows that the RMSE, one of the two evaluation metrics, is more than 2000 for train data and more than 1000 for test data, as shown in Figure 4.4. On the other hand, R-squared value is 100 percent for train data and 100 percent for test data, which is an excellent performance.

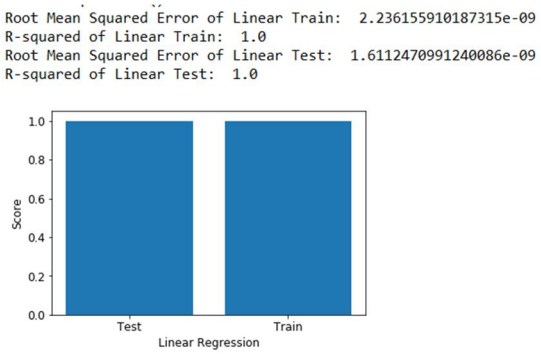


Figure 4.4 Linear Regression (Train and Test)

The below output shows that the RMSE and R-squared values for the Ridge Regression model on the training data is more than 2000 and 100 percent, respectively, as shown in Figure 4.5. For the test data, the result for these metrics is more than 1000 and 100 percent, respectively.

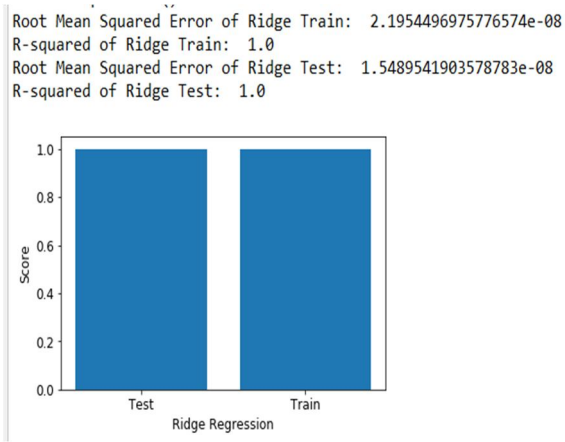


Figure 4.5 Ridge Regression (Train and Test)

The below output shows that the RMSE and R-squared values for the Lasso Regression model on the training data is 727 thousand and 99 percent, respectively, as shown in Figure 4.6. The results for these metrics on the test data are 741 thousand and 99 percent, respectively. Lasso Regression can also be used for feature selection because the coefficients of less important features are reduced to zero.

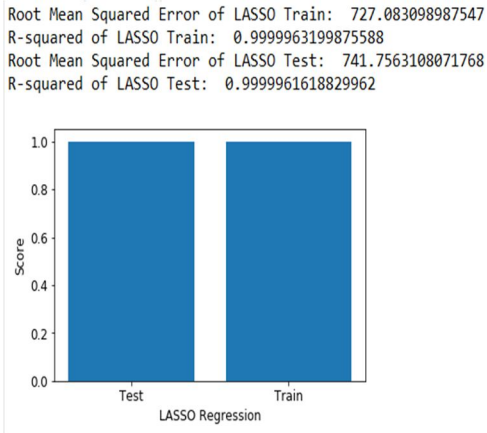


Figure 5.6 LASSO Regression (Test and Train)

Here, the comparision between the three regression techniques using their testing sets have been graphically represented, as shown in Figure 4.7. There is just a little difference at all between the three accuracy scores.

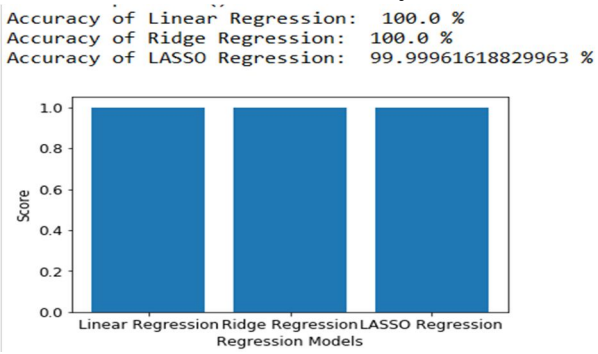


Figure 4.7 Comparision of Regression Techniques

V.CONCLUSION

Sales forecasting is a critical part of the strategic planning process and allows a company to predict how their company will perform in the future. The performance of the models is summarized below:

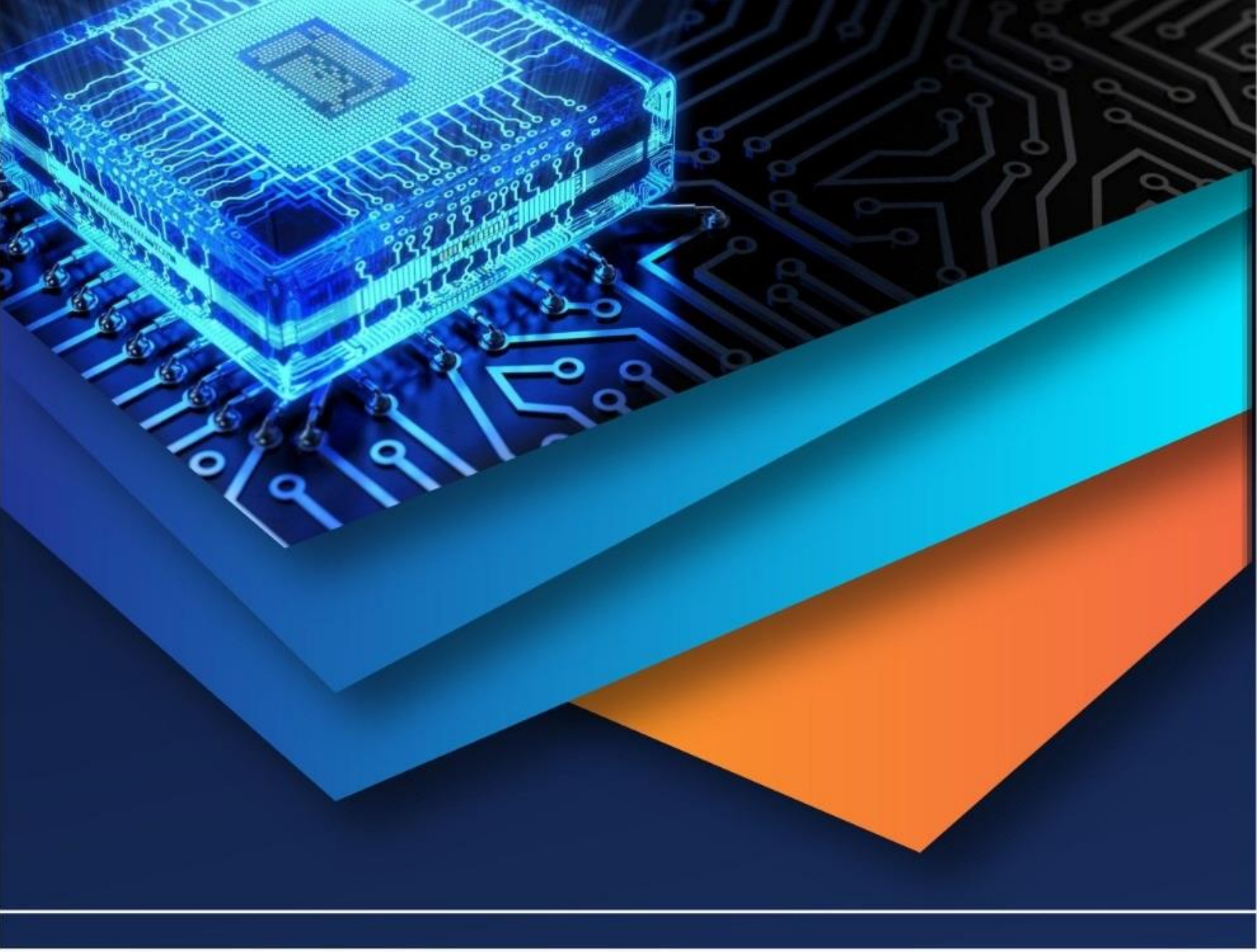
1) *Linear Regression Model:* Test set RMSE of more than one thousand and R-square of 100 percent.
2) *Ridge Regression Model:* Test set RMSE of more than one thousand and R-square of 100 percent.
3) *Lasso Regression Model:* Test set RMSE of 741 thousand and R-square of 99.99 percent.

All the regression models are performing with an excellent R-squared and stable RMSE value. The most ideal result would be an RMSE value of zero and R-squared value of 1, but that's almost impossible in real economic datasets.

There are other iterations that can be done to improve model performance. The value of alpha is assigned to be 0.01, but this can be altered to arrive at the optimal alpha value. Cross-validation can also be tried along with feature selection techniques. Various Regression models using the scikit-learn library have been implemented here. As for the future enhancement, elastic net regression can be used.

REFERENCES

[1] Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. Data, 4(1), 15.
[2] Huang, L., Dou, Z., Hu, Y., & Huang, R. (2019). Online Sales Prediction: An Analysis with Dependency SCOR-topic Sentiment Model. IEEE Access.
[3] Cheriyan, S., Ibrahim, S., Mohanan, S., & Treesa, S. (2018, August). Intelligent Sales Prediction Using Machine Learning Techniques. In 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE) (pp. 53-58). IEEE.
[4] Loureiro, A. L., Miguéis, V. L., & da Silva, L. F. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. Decision Support Systems, 114, 81-93.
[5] Nunnari, G., & Nunnari, V. (2017, July). Forecasting Monthly Sales Retail Time Series: A Case Study. In 2017 IEEE 19th Conference on Business Informatics (CBI) (Vol. 1, pp. 1-6). IEEE.
[6] Zhao, K., & Wang, C. (2017). Sales Forecast in E-commerce using Convolutional Neural Network. arXiv preprint arXiv:1708.07946.
[7] Thiesing, F. M., & Vornberger, O. (1997, June). Sales forecasting using neural networks. In Proceedings of International Conference on Neural Networks (ICNN'97) (Vol. 4, pp. 2125-2128). IEEE.
[8] Thiesing, F. M., Middelberg, U., & Vornberger, O. (1995). Short term prediction of sales in supermarkets. In Proceedings of ICNN'95-International Conference on Neural Networks (Vol. 2, pp. 1028-1031). IEEE.
[9] Efendigil, T., Önüt, S., & Kahraman, C. (2009). A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy models: A comparative analysis. Expert Systems with Applications, 36(3), 6697-6707.
[10] Doganis, P., Alexandridis, A., Patrinos, P., & Sarimveis, H. (2006). Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing. Journal of Food Engineering, 75(2), 196-204.
[11] Cerqueira, V., Torgo, L., Pinto, F., & Soares, C. (2019). Arbitrage of forecasting experts. Machine Learning, 108(6), 913-944.
[12] Papacharalampous, G., Tyralis, H., & Koutsoyiannis, D. (2019). Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. Stochastic Environmental Research and Risk Assessment, 33(2), 481-514.
[13] Abudu, S., Cui, C. L., King, J. P., & Abudukadeer, K. (2010). Comparison of performance of statistical models in forecasting monthly streamflow of Kizil River, China. Water Science and Engineering, 3(3), 269-281.
[14] Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. Econometric Reviews, 29(5-6), 594-621.
[15] Mentzer, J. T., & Moon, M. A. (2004). Sales forecasting management: a demand management approach. Sage
[16] https://bigdata-madesimple.com/machine-learning-sales-forecasting-tackle-insufficient-data-issue/
[17] https://www.pluralsight.com/guides/linear-lasso-ridge-regression-scikit-learn
[18] https://bigdata-madesimple.com/machine-learning-sales-forecasting-tackle-insufficient-data-issue/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY