International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.089

Cost Prediction in Acquiring Customers Using Machine Learning

NarayanaDarapaneni
Director - PGDCSAI
Great Learning/Northwestern University, Illinois, USA
darapaneni@gmail.com

Anwesh Reddy Paduri
Senior Data Scientist
Great Learning,
Hyderabad, India
anwesh@greatlearning.in

Shiva Shukla
Student - PGDCSAI
Great Learning,
Bangalore, India
sshukla.iitk@gmail.com

KuldeepDwivedi
Student - PGDCSAI
Great Learning,
Bangalore, India
kuldeep.dwivedi1985@gmail.com

Shreya Sharma
Student - PGDCSAI
Great Learning,
Bangalore, India
s26shreyasharma@gmail.com

SubhadipSaha
Student - PGDCSAI
Great Learning,
Bangalore, India
subhadipsaha123@gmail.com

Sandeep Kumar
Student - PGDCSAI
Great Learning,
Bangalore, India
sksingh.netcradle@gmail.com

Ankit Gupta
Student - PGDCSAI
Great Learning,
Bangalore, India
ankit123.kc@gmail.com

Abstract—Cost prediction for customer acquisition is a very significant assignment because it has evolved into a critical business metric that aids companies in determining the resources they will need to continue to expand and acquire new clients. So, customer acquisition cost is the sum of money that a business pays to bring on a new client. Nevertheless, developing an appropriate cost prediction model for customer acquisition is not a straightforward process because there are numerous independent variables in both the presented Food Mart X data set and real-world situations, as well as numerous categorical data with high cardinality. The hardest challenge that draws the attention of researchers and practitioners is estimating the required cost.

In order to estimate the cost of acquiring a customer, this paper uses a machine learning approach. The findings demonstrate that, in comparison to conventional estimating methods, machine learning techniques may be utilised to anticipate Cost on Acquiring Customers with High Accuracy Rate.

The proposed model in this research was trained on the Food Mart X data set, having the records of 48k customers. An analysis of the results of the implementation for the proposed methods showed that the cost Predicting process using Decision Tree algorithm (DT), Random Forest algorithm (RF) and Bagging Forest algorithm (BF) has the ability to predict the costs required to acquire the customer. Out of those, Random Forest algorithm (RF) has shown highest accuracy inpredicting the required cost to acquire customer compare to decision tree and bagging method. Literature also shows the development of neural network model for predicting the cost but it's not cover here in this study.

Keywords— Decision Tree (DT), Random Forest (RF), Bagging Forest (BF), Deep learning neural Network (DNN), Root Mean Squared Error (RMSE).

I. INTRODUCTION

Customer Acquisition cost is being vastly used by the business to improve the vision and positioning of the resources and capital. "Even in well-managed organisations, there can be a high level of client churn," it has been noted. We must find replacements for these lost clients.Clients could leave as they get older and move through the family life cycle. If their personal circumstances change and they no longer require or appreciate your product, we could lose our current clientele. These unpredictable reasons of client loss suggest that customer acquisition will always be required to replace natural attrition. Hence in today's world, there is a need of accurate model through which a businessman can simulate those condition and can able to forecast the cost. These types of models can be used at various places such aspredicting the cost of goods, raw material, commodities such as steel price etc as these prices directly hit the final product prices as well the overall business.

II. RELATED WORKS - LITERATURE SURVEY

Regression analysis, which is nothing more than a statistical approach used to assess the relationship between a dependent/target variable (here, the cost of customer acquisition.) and one or more independent (interdependent) factors, may generally be used to create cost prediction tasks.

Cost forecasting is typically carried out using descriptive, predictive, and prescriptive analytics.

*Descriptive Analytics:*Statistical techniques used in descriptive analytics [19] include data gathering, analysis, interpretation, and presenting of results. In essence, this form of analytics aids in determining what occurred.

*Predictive analytics:*In the context of price predictions, predictive analytics [19] involves analysing recent and previous data to estimate the likelihood of upcoming occurrences, outcomes, or values. Many statistical methods, including machine learning and data mining (the detection of patterns in data), are needed for predictive analytics.

A. Cost Prediction methods and techniques

The development of considerably more efficient and precise predictive algorithms has been made feasible by the developments in statistics during the past 20 years. What potential uses are there for cost modelling and estimation?

Traditional Costing Models

DOI: rp-9788770040723.089

*1. **Analog Method:***

By evaluating similar products that have been manufactured or purchased in the past, this method calculates the cost of a new product. Although this strategy is unreliable, it can be utilised in very early stages (such as a feasibility study) when the details of the project or service are unknown. In this study, we won't linger on this kind of fundamental estimate.*Analytical Method*

By simulating the industrial production process, it calculates a product's cost. This approach is based on the product's cost structure, and it estimates each intermediate component using the materials & components used, the process costs (labour and machine), as well as any additional structural expenses.

This kind of approach necessitates a thorough understanding of the subject; thus it occasionally fails when particular precautions are not taken or when methods are changed abruptly.

2. Parametric Method

This technique uses statistical modelling to determine the price of a good or service. This method models the evolution of the cost as a function of specific factors known as "cost drivers" by using historical data from similar goods and services to build equations or statistical laws.

These models are typically based on regressions that are linear, multilinear, polynomial, or logarithmic.

- Major problems with this type of method are as below:They don't handle missing data well, so highly clean databases are needed.
- They manage "breaks" or threshold effects ineffectively. Because the manufacturing process can change, the price, for instance, may behave linearly up to a certain point before drastically changing (size, weight, volume, etc.).

3. Non-Parametric Method

The shortcomings of conventional parametric approaches have mostly been resolved by the advancements made in the fields of algorithms and machine learning, which have also improved their performance and scope of application.

The "random forests" approach, which Leo Breiman and Adèle Cutler formally proposed in 2001 (Breiman, L., Random Forests. Machine Learning. 45, 5-32 (2001)) is a statistical method. nonparametric that uses "Bootstrap" approaches to create numerous decision trees that are trained on slightly different data subsets.

In both regression and classification, Pierre Geurts [2] introduced and described how bias-variance trade-off is accomplished. The bias and variance of the statistical model are significantly influenced by the choice of variables and attributes.

In order to assist farmers in predicting the amount of rain that will fall, Wanie M. Ridwan et al. [1] employed four distinct regression algorithms: Bayesian Linear Regression (BLR), Boosted Decision Tree Regression (BDTR), Decision Forest Regression (DFR), and Neural Network Regression (NNR).

A. Krishna et al. [2] attempted to identify the best algorithm by using various machine learning techniques to predict the sales of a retail store. They used both conventional regression approaches and boosting techniques, and discovered that the boosting algorithms produced superior results to the conventional regression algorithms.

Based on a number of independent variables, a regression model is used to forecast or predict the value of the dependant variable—cost.With the evolution of Machine Learnings and advancement in internet speeds, Machine learnings are now being extensively used to predict the continuous numerical as well classification problem. The two giant E-commerce Industry- Amazon &Flipkart are extensively using the machine learning algorithms in predicting their customer acquisition cost.

B. Machine Learning Algorithms focussed here are as follows:

1. Decision Tree Regression
2. Random Forest
3. Bagging Forest

Some of the case studies has been done for improvement of sales forecasting problems by Deep Neural Network also.

III. PROPOSED WORK

A. Materials and methods used

This dataset has sourced from https://www.kaggle.com/datasets/ramjasmaurya/medias-cost-prediction-in-foodmart. It contains the information regarding demographics of existing customers. This dataset has two files, one for training (36,256 rows & 41 columns) and other for testing (12086 rows & 41 column). Also, one additional file was provided, which contain information about feature values and their description.

In these features, Cost is a dependent variable and the remaining features are called as independent variable. Here we need to predict the value of dependent variable using independent variable. For validation of model in the last, we have created a validation data set from the training data set.

From the literature survey & preliminary study, we applied following 3 regression models on the dataset:

1. Decision Tree Regression
2. Random Forest Regression
3. Bagging Forest Regression.

IV. IMPLEMENTATION

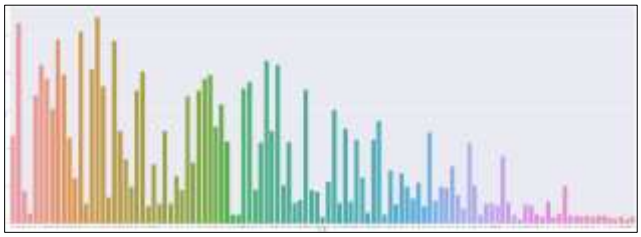
A. Dataset, data collection, data pre-processing

After collection of datasets- we started to understand the datasets. During data pre-processing, we faced challenges like “high Cardinality”, “Non-Linearity”, “Collinearity” as these data set has 16 categorical type data. Out of these 16 categorical data, 6 categorical data have a high unique value, such as., “Media Type”, “Store City”, “Brand Name”, “Promotion Name”, “Food Department”, “Food Category” thus suffering from high cardinality issue.

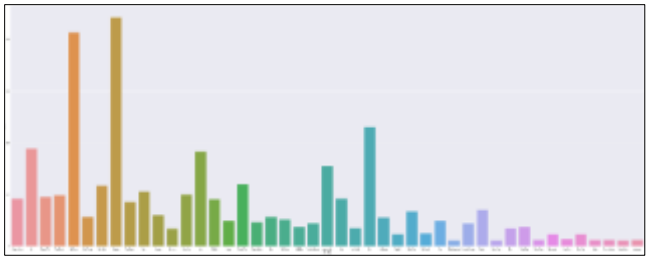
DOI: rp-9788770040723.089

Some of the categorical type data count plot has been plotted to understand the curse of dimensionality which are as follows:

1. Brand Column Count plot is as follows to shows the number of unique values:

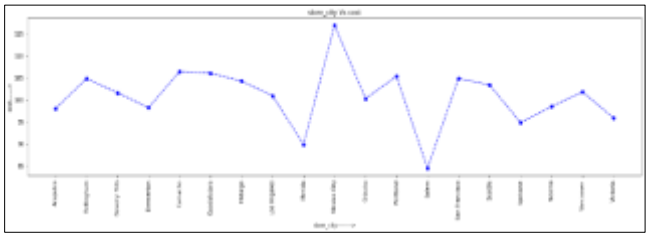


2. Similarly, food column category count plot is as follows:

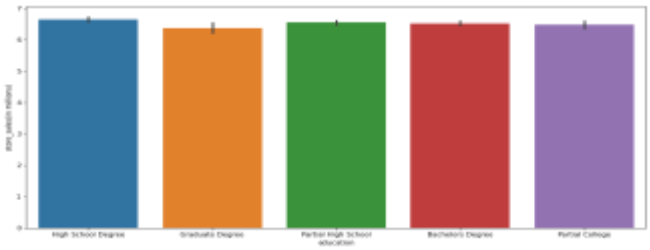


Through "One Hot Encoding" – when have converted all these categorical types data & it leads to final 298 column leading to large dimensions. We have also tried to check the relation between dependent and independent columns. Some of the relation is as follows:

- *Relation between Store city and Cost*



- *Education vs Cost*



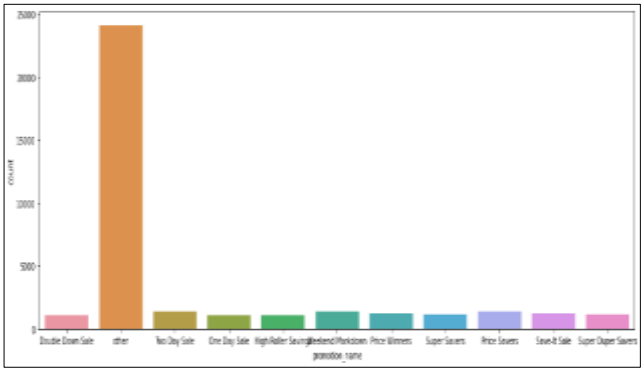
To reduce the dimensions, we have also calculated another attributes from some of the related columns, for e.g., we have calculated & made a new column for "Profit" from the two given columns namely store_sales& store cost to overcome the problem of large dimensions.

B. Exploratory Data Analysis and Evaluation of Models

1. EDA for Analysis-1:

Model has been made using only Top 10 features & grouping rest into "OTHERS". Thereafterconverting Category data to numerical type using Label encoders.

- Top 10 features for Promotions is as below:



- Top 10 features for Food Department is as below:

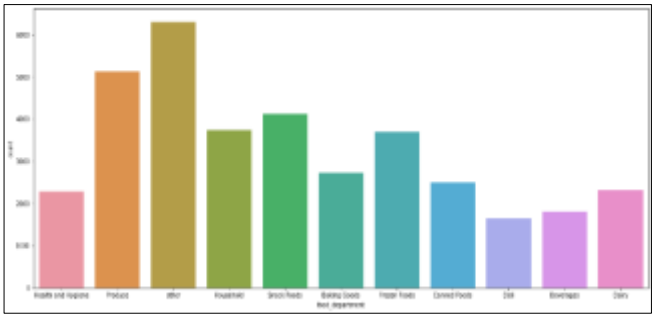


TABLE-1: RMSE AND SCORE FOR ANALYSIS-1

Model	Test RMSE	Train RMSE	Test R2 Score	Train R2 Score
Decision Tree	28.22	28.18	0.11	0.11
Random Forest	14.60	7.41	0.76	0.93
Bagging	15.11	7.95	0.74	0.92

Result of test score for both random forest and Bagging reveals that the model is overfitting. Overfitting might happen due to the high variance which might have introduced. Hence another analysis is envisaged where instead of label encoding, one-hot encoding shall be applied.

Joint Plot for analysis-1 is shown in fig 1, 2 & 3:

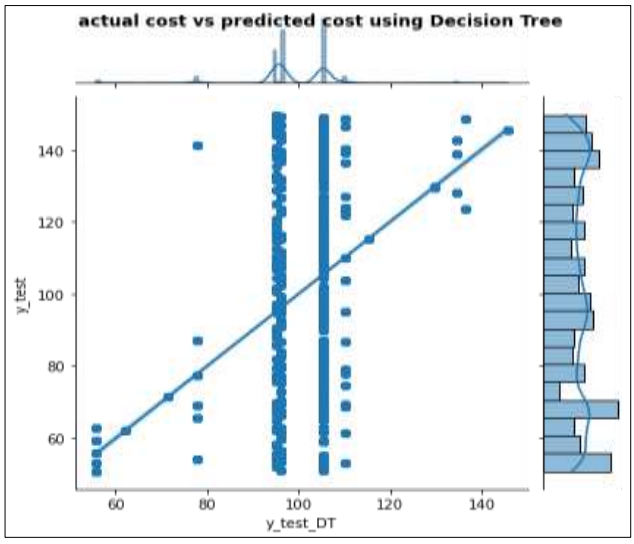


Fig.1. DT Based Model

DOI: rp-9788770040723.089

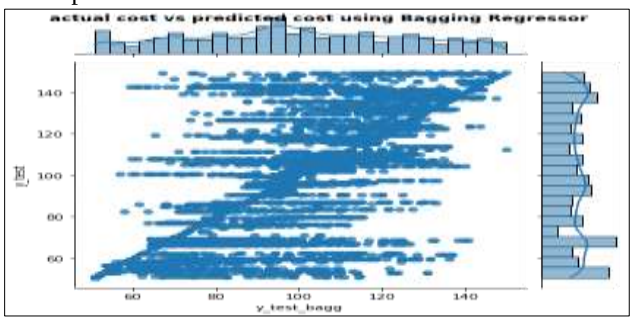


Fig.2. Bagging Based Model

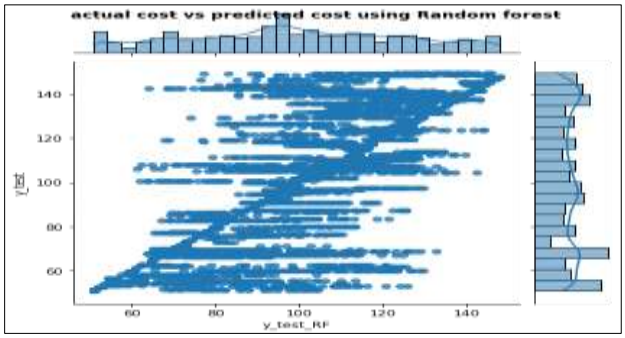


Fig.3. Random Forest Model

2. Analysis-2

Model has been made using all features but this time instead of using Label Encoder, One hot encoding has been used to remove the bias error which might have come due to Label encoder in Analysis-1.

Results are mentioned under section-5.

However, since frequency encoding was applied on only one feature, it is not sure whether this will work in real life problems as well. So, - Analysis-3 with another method of encoding called as target encoding has been done.

Target encoding involves replacing a categorical feature with average target value of all data points belonging to the category. *Python Code for Target Encoding which shall be used in Analysis-3:*

```
fromcategory_encoders import TargetEncoder
te=TargetEncoder ()
te.fit(X,y)
```

3. Analysis-3

Model has been made using all features but this time instead ofLabel Encoder & One hot encoding, "Target Encoding" has been applied on the categorical columns and results were further found to be improved compared to the results of Analysis-2.

TABLE3: RMSE AND SCORE FOR ANALYSIS-3

Model	Test RMSE	Train RMSE	Test R2 Score	Train R2 Score
Decision Tree	1.49	0.34	0.99	0.99
Random Forest	1.18	0.59	0.99	0.99
Bagging	1.38	0.59	0.99	0.99

Joint Plot for analysis-3 is shown in fig 4,5&6:

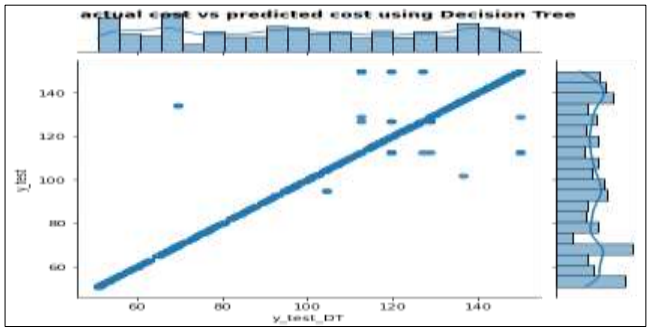


Fig.4. DT Based Model

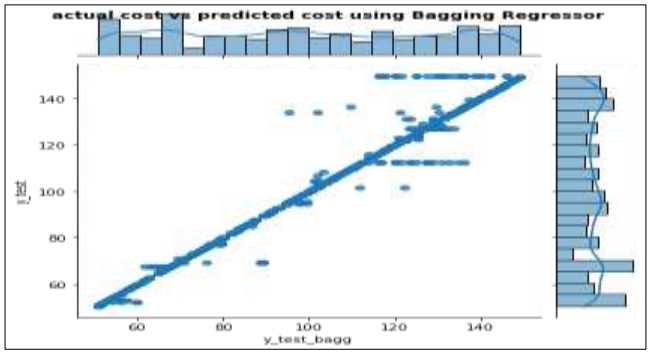


Fig.5. Bagging Based Model

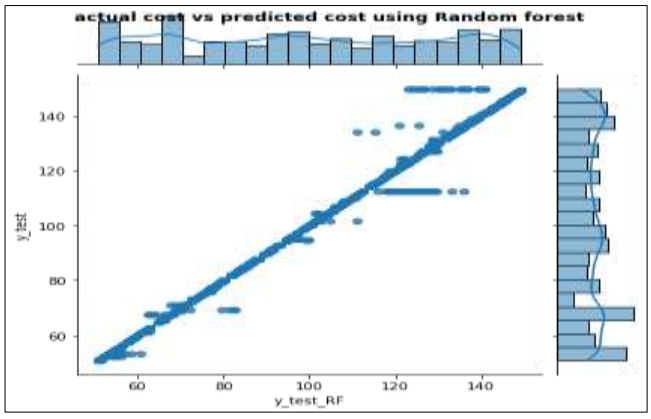


Fig.6. Random Forest Model

V. RESULTS AND DISCUSSION

The methods and procedures utilised for machine learning as applied to observational datasets that can provide information on customer acquisition cost were thoroughly studied and described in this work. The summary of this work does not necessarily apply to all machine learning-based investigations because machine learning approaches have been utilised much more broadly throughout observational studies than in the context of individual decision making. This study focuses on a topic that is still relatively unexplored: how to leverage massive datasets in a way that can enhance customer acquisition cost prediction outcomes.

With the application of Frequency Encoding Technique on categorical column having highest unique values, results improved compare to Analysis-1.

Result of RMSE & Score are tabulated below after the application of Frequency Encoding:

TABLE2: RMSE AND SCORE AFTER FREQUENCY ENCODING APPLICATION ON "BRAND COLUMN" IS AS BELOW:

Model	Test RMSE	Test R2 Score	Train R2 Score
Decision tree Regression	1.52	0.9975	1.0
Random Forest	1.23	0.998	0.999
Bagging Forest	3.23	0.988	0.988

Similarly with the application of Target Encoding Technique on all categorical column, results found even better then Table-2 and has been tabulated below:

TABLE3: RMSE AND SCORE FOR ANALYSIS-3

Model	Test RMSE	Test R2 Score	Train R2 Score
Decision tree Regression	1.49	0.997	0.999
Random Forest	1.18	0.998	0.999
Bagging Forest	1.38	0.997	0.999

VI. CONCLUSION

As can be seen from this study, problem(data) having high number of categorical data and that too with high unique values, then Target encoding is working best among popular encoding method such as Label Encoder and One hot Encoding.

Among the models, after applying the target encoding during pre-processing steps, random Forest model performs well with respect to other model in terms of root mean square error.Performance analysis for all the three-analysis done here are as below:

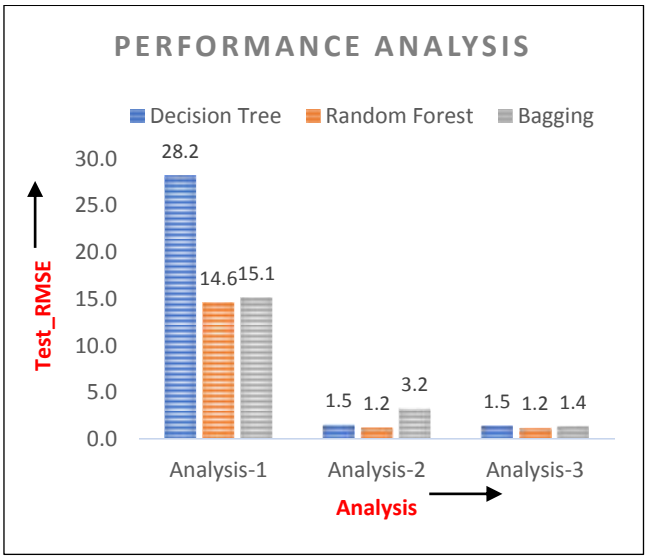


Fig.7. Performance analysis of different models

By doing various analysis, we found out that target encoding works better.

Additionally, seeing the performance of various modelsbeing studied here, Random Forest gives better accuracy compare to other two models. Hence, we propose to use Random Forest with target encoding for such high category type data set in future in predicting the cost.

Yet, as evidenced by our literature review, we believe external validation is also necessary to guarantee model fidelity but is rarely done. There may be a number of causes for this, including a dearth of pertinent datasets or a failure to recognise the significance of external validation. Before applying models to any cost forecast, there is an increasing necessity for external validation as machine learning model development rises.

REFERENCES

[1] Wanie M. Ridwan, Michelle Sapitang, Awatif Aziz, KhairulFaizalKushiar, Ali Najah Ahmed, and Ahmed El-Shafie, "Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia," Ain Shams Engineering Journal, vol. 12, issue 2, pp. 1651-1663, ISSN 2090-4479, 2021, https://doi.org/10.1016/j.asej.2020.09.011.https://www.sciencedirect.com/science/article/pii/S2090447920302069

[2] A. Krishna, A. V, A. Aich and C. Hegde, "Sales-forecasting of Retail Stores using Machine Learning Techniques," 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), pp. 160-166, 2018, doi: 10.1109/CSITSS.2018.8768765.

[3] Jeffrey M. Stanton,"Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors," Journal of Statistics Education, vol. 9, p. 3, 2001, DOI: 10.1080/10691898.2001.11910537.

[4] M. Li, S.Ji and G. Liu, "Forecasting of Chinese E-Commerce Sales: An Empirical Comparison of ARIMA Nonlinear Autoregressive Neural Network and a Combined ARIMA-NARNN Model[J]", Mathematical Problems in Engineering 2018.

[5] https://www.kaggle.com/datasets/ramjasmaurya/medias-cost-prediction-in-foodmart.

[6] Y.M. Khaing, M.M. Yee, E. Ei, forecasting stock market using multiple linear regressionAung. Int. J. Trend Sci. Res. Dev. (IJTSRD),vol. 3, no. 5, pp. 2019.

[7] G. Nguyen, Kedia, Jai, Snyder, Ryan, Pasteur, R., Wooster, R. Sales Forecasting Using Regression and Artificial Neural Networks, 2013.

[8] A. Aima, WALMART sales data analysis & sales prediction using multiple linear regression in R programming Language, [Online], Available: https://medium.com/@arneeshaima/walmart-sales-data-analysis-sales-prediction-using-multiple-linear-regression-in-r-programming-adb14afd56fb (March 19).

[9] E. Bank, "How to develop & use a regression model for sales forecasting," Updated September 26, 2017. https://bizfluent.com/how-7298496-develop-regression-model-sales-forecasting.html. Accessed 4 Oct 2019.

[10] R.R. Shelke, R.V. Dharaskar, and V.M. Thakare, "Data mining for supermarket sale analysis using association rule," Int. J. Trend Sci. Res. Dev., vol. 1, no. 4. ISSN: 2456-6470.

[11] A.L.D. Loureiro, V.L. Miguéis, and Lucas F.M. da Silva, "Exploring the use of deep neural networks for sales forecasting in fashion retail," Decision Support Systems, vol. 114, pp. 81-93, ISSN 0167-9236, 2018,https://doi.org/10.1016/j.dss.2018.08.010.(https://www.sciencedirect.com/science/article/pii/S0167923618301398).

[12] Rajesh, M., &Sitharthan, R. (2022). Image fusion and enhancement based on energy of the pixel using Deep Convolutional Neural Network. Multimedia Tools and Applications, 81(1), 873-885.

[13] B. Nithya, and V. Ilango,"Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems ICICCS, 2017, 978-1-5386-2745-7/17/$31.00 ©2017 IEEE.

[14] A S. Temür, M. Akgün, and G. Temür, "Predicting Housing Sales in Turkey Using Arima, Lstm and Hybrid Models," J. Bus. Econ. Manag., vol. 20, no. 5, pp. 920–938, 2019, doi: 10.3846/jbem.2019.10190.

[15] J. M. Keller, M. R. Gray, J. A. Givens Jr., "A Fuzzy K-Nearest Neighbor Algorithm", IEEE Transactions on Systems, Man and Cybernetics, vol. SMC-15, no. 4, August 1985.

[16] Pazhani. A, A. J., Gunasekaran, P., Shanmuganathan, V., Lim, S., Madasamy, K., Manoharan, R., &Verma, A. (2022). Peer–Peer Communication Using Novel Slice Handover Algorithm for 5G Wireless Networks. Journal of Sensor and Actuator Networks, 11(4), 82.

DOI: rp-9788770040723.089
[17] R. S. Sutton, "Introduction: The Challenge of Reinforcement Learning", Machine Learning, Kluwer Academic Publishers, Boston, vol. 8, pp. 225-227, 1992.

[18] P. Harrington, "Machine Learning in action", Manning Publications Co., Shelter Island, New York, 2012.

[19] Business Analytics The Science of Data-driven Decision Making by U. Dinesh Kumar (z-lib.org)