Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220074

Design and Implementation of Sales Prediction Model Using Decision Tree Regressor over Linear Regression Towards Increase in Accuracy of Prediction

S. Ravi Teja Reddy[a] and P. Malathi[b,1]
[a]*Research Scholar, Department of CSE, Saveetha School of Engineering,*
[b]*Project Guide, Department of CSE, Saveetha School of Engineering,*
Saveetha Institute of Medical and Technical Sciences,
Saveetha University, Chennai, Tamilnadu, India

Abstract: The purpose is to predict future price of product and assists companies to make business strategic plans to increase overall sales and also an experiment is performed to find the best suitable algorithm among Linear Regression and Novel Decision tree regressor. Predicting future price of a product using linear regression algorithm (N=10) and Novel Decision tree regressor (N=10). Dataset used is Bigmart Sales data from kaggle. The sample size is 542 for each group. Novel Decision tree regressor produces a better accuracy of 97.5% and for Linear regression classifier is 87.6%with a statistical significance value of p is 0.03 ($p<0.05$). The results proved that the Novel Decision tree regressor is significantly better for sales forecasting than linear regression algorithm within the study's limits.

Keywords: Linear Regression, Novel Decision tree regressor, Machine learning, Supervised learning, Sales Forecasting.

1. Introduction

Sales forecasting has a critical role to playin business strategies which will increase market operations and productivity based on previous sales data and current demand of product. Statistical models used for sales forecasts are more complex due to having an impact on of internal and outside factors [1]. Sales for all Sales Retailer shops to be competitive and to be a part of global trade, forecasting is required, which will Retail-Apparel-Distribution Network Actors: Increase Production and Simplify Management[2] with varying demand from all sectors of the population. Some applications of sales forecasting are: An clever multivariate choice-making model is evolved to provide predictions for income problems by means of combining information optimization and superior modules, harmonization multivariate clever forecaster module and search-wrapper-based variable module [3]. Sales Volume

[1]P. Malathi, Department Of Computer Science And Engineering, Saveetha School Of Engineering, Saveetha Institute Of Medical And Technical Sciences,Saveetha University, E-mail: malathip.sse@saveetha.com

forecast and Long-standing sales forecast is done by most research and companies using mathematical methods, regression analysis, or sophisticated computer simulations [5]Yuan and Lee 2011 and also provides idea of product arrangements inside shop based on insights mined from previous buying pattern of customers [5]Ustundag, Cevikcan, and Kilinc 2008.

A total of 55 articles in IEEE and 234 articles in Google Scholar have been published. The previous literature compares decision trees with other computer-based classifiers such as neural networks, genetic systems. The model lets in an ordinary boomin mid-term forecasting accuracy when compared to the predictor of the mean sales profile and other classifiers. A journal by [6]Pavlyshenko, n.d. mainly focusing on time series forecasting using machine learning models and it gives a stacking approach for building a regression ensemble of unmarried models has been studied.

A study by [7] (Bhukya and Ramachandram 2010) addresses these troubles by proposing a manner to split records the usage of AVL trees, which improves best and balance. Researchers from a spread of fields such as mathematics, system learning, sample recognition, and statistics mining have looked at the hassle of building a decision-making answer from to be had records. A paper by [8]Ayomide Yusuf (Yusuf and Alawneh 2018) shows a new way of forecasting sales with integration of GPU with linear regression. (Grąbczewski 2014) found that tests performed to study management features in Decision Tree and import controls are inconsistent[9]. Sometimes, a thorough search improves results, and sometimes it is destructive. An intelligent multivariate choice-making model is advanced to provide powerful predictions for this trouble by way of combining facts optimization and superior module, harmonization seek-wrapper-a multivariate sensible forecaster module, and a based variable module(Guo, Wong, and Li 2013)[3].

Our wide portfolio in research has translated dinto publications in numerous interdisciplinary projects.[10-13] (Sekar et al. 2019; Johnson et al. 2020; Subashri and Thenmozhi 2016; Sriram, Thenmozhi, and Yuvaraj 2015). Now we are focussing on this topic.

In existing literature, sales forecasting models developed not utilized factors such as information about different vendors of each product and their marketing ideas such as promotions and discounts of products. These aforementioned factors are identified as crucial factors which have a greater impact on predicting future sales price Machine learning models are used. The aim of this current study is to include factors such as vendor information and marketing practices for building Sales forecasting Model which will improve accuracy.

2. Materials And Methods

The study was conducted at the Saveetha School of Engineering's Data Analytics Laboratory, which is part of the Department of Computer Science Engineering. Two groups are used in this study. Linear Regression is in group one, and Novel Decision Tree Regressor is in group two. Using previous study results [14](Feng and Wang 2017), a sample size of 542 was calculated with a threshold of 0.05, G power of 80%, confidence interval of 95%, and enrollment ratio of 1.

The Bigmart sales dataset is collected from Kaggle website. It contains a total of 8524 rows of food items data. Bigmart sales dataset contains 11 attributes like object Weight Outlet Identifier, Outlet establishment 12 months, Outlet size, Outlet Location

Type, Outlet type, Item fat content, Item Visibility, object type, Item MRP, Dataset needs to be processed before applying it to a machine learning model. In data collection procedure, data is collected from different stores and different cities and are stored in a CSV file for further analysis and clustering of data through the Data Mining process. Data processing includes removing missing data and replacing null values with mean or median values with standardization of data. The preprocessed data is given as an input for linear regression and novel Random Forest. Table 1 represents a sample dataset for Bigmart stores. From preprocessed data 70% is given as training data and 30% is given as testing data.

The Jupyter pocket book with Python computer language was used to evaluate Linear regression and new Random woodland techniques. The hardware setup was modified to an Intel Core i5 processor with 8 GB of RAM. A 64-bit operating system with an X64 CPU was used. Windows 10operating systems are included in the software configuration.

2.1.Group 1-Linear Regression

A linear regression line has a sum, and it is in the form of Y=a(i)+b(i)x, where x is an explanatory variable, then Y is a structured variable. One variable considers being a established variable least squares regression. The maximum commonplace method for solving a regression lineis technique for improving a regression line used to calculate technique of least squares. Pseudocode for linear regression is given in Table 1.

Table 1. Pseudocode for Linear Regression

Input - Import dataset and required packages
Step 1: Preprocess data which is the removal of unnecessary data.
Step 2: Describe the Dependent and Independent Variables. //Initialization
Step 3: Training the model - Define a LinearRegression() function - Use linearreg.fit() to fit the model between x_train and y_train.
Step 4: Testing the model - Split dataset into two parts Training(80%), Testing(20%)
Step 5: Evaluating the model - Print regression equation Output: R-Squared values,MAE,MSE,RMSE.

2.2.Group 2- Decision tree regressor

In the form of a tree structure, selection timber construct regression or type models. At the same time as it breaks down a dataset into smaller and smaller subsets, an associated decision tree is incrementally advanced. The final product is a tree with leaf nodes and selection nodes. Table 3 shows the pseudocode for the Decision Tree Regressor Algorithm.

Table 2. Pseudocode for Novel Decision tree regressor

Step 1: All training instances should be assigned to the tree's root. Set the current node to the root of the tree.
Step 2: For each attribute A. Use the value of the attribute to partition all data instances at the node. B. Calculate the information gain ratio as a result of partitioning.
Step 3: A. Determine which characteristics produce the highest information gain ratio. Set this characteristic as the splitting condition for the current node. B. If the best information gain ratio is 0, tag the current node as a leaf and return.
Step 4: Partition all instances based on the best feature's attribute value.
Step 5: Each partition should be considered a child of the current node.
Step 6: For each child node: A. Tag the child node as a leaf and return if it is "pure" for each child node (has instances from only one class). B. If it doesn't work, go back to step 2 and make the child node the current node.

2.3.Statistical Analysis

Statistical analysis is done for comparing both Linear Regression and Novel Decision tree regressor using IBM SPSS version 21 tool. Outlet Type based on dependent variable are independent variables in dataset. Item Outlet Sales are expected to be in high demand. In this study, the researchers used an independent t-test.

3. Results

The accuracy of the Novel Decision tree regressor is 97.5% and accuracy of linear regression is 87.6%. Significance level for both algorithms is less than 0.03. Table 3- shows accuracy values for Linear regression and Decision tree regressor while training model.

Table 3. Accuracy table for Linear regression and Novel Decision tree regressor

N	Linear regression	Decision tree regressor
1	87.07	96.82
2	82.51	95.54
3	82.52	97.32
4	80.88	94.65
5	87.50	95.35
6	80.98	94.54
7	91.22	92.95
8	80.85	96.58
9	83.85	93.65
10	84.68	95.65

Table 4 shows loss values for Linear regression and Decision tree regressor for different samples while training the model.

Table 4. Loss table for Linear regression and Novel Decision tree regressor

N	Linear regression	Decision tree regressor
1	0.80	0.72
2	0.10	0.69
3	0.55	0.69
4	0.43	0.69
5	0.39	0.68
6	0.35	0.67
7	0.29	0.72
8	0.01	0.71
9	0.01	0.61
10	0.01	0.01

Table 5. Statistical analysis of mean, Standard deviation, and standard error of accuracy of Novel Decision tree regressor and Linear regression algorithm. Novel Decision tree regressor had the highest accuracy.

	Algorithm	N	Mean	Std.Deviation	Std.Error Mean
Accuracy	Linear Regression	10	87.6700	5.11218	1.16006
	Decision tree regressor	10	97.5800	3.709145	1.17734
Loss	Linear Regression	10	0.2940	0.26500	0.08360
	Decision tree regressor	10	0.6250	0.21666	0.06858

Table 5gives Descriptive Statistics for Novel Decision tree regressor and Linear Regression for both Accuracy and Loss. For Accuracy, calculated mean accuracy for decision Tree Regressor is 97.58 with a standard deviation of 3.709145 and standard error is 1.17734 by comparing with Linear Regression mean accuracy is 87.6700, standard deviation is 5.11218 and standard error is 1.16006. For Loss, average loss recorded for Decision Tree Regressor is 0.6250 with a standard deviation of 0.21666 and standard error of 0.06858, for linear Regression mean loss is 0.2940 with a standard deviation of 0.26500 and standard error of 0.08360. The mean loss of Decision Tree Regressor seems to be greater than linear Regression.

Table-6 shows a significant difference between 2 groups using an independent sample T-test with a p-value of 0.03 which is less than 0.05. Novel Decision tree regressor produces a better accuracy of 97.5% than linear regression classifier 87.6%.

Table 6. Using a Novel Decision Tree regressor and Linear regression, an independent sample t-test of accuracy and loss for sales forecasting was performed. The two algorithms have statistically significant differences (p0.05). Novel Decision Tree Regressor is significantly greater than the Linear Regression algorithm.

		Levene's test for equality of variances		T-Test for equality of means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean difference	Std. error difference	95% confidence interval of the difference	
									Lower	Upper
Accuracy	Equal Variances assumed	0.008	0.003	-8.392	18	0.03	-9.914	1.891	-19.847	-11.90015
	Equal variances not assumed			-8.392	17.035	0.043	-9.914	1.891	-19.847	-11.88394
Loss	Equal Variances assumed	0.04	.025	-3.05	18	.007	-0.331	0.108	-558	-0.1035
	Equal variance not Assumed			-3.05	147	.007	-0.331	0.108	-5591	-0.1028

As illustrated in Figure 1. A comparison of mean accuracy is represented by a bar chart. of 97.5% for Novel Decision tree regressor and 87.6% for linear regression algorithm. Standard deviation among algorithms is similar.

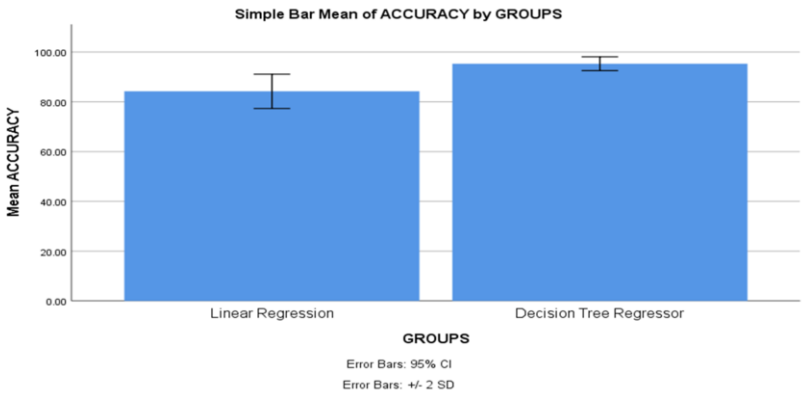

Figure.1. A bar chart with an error bar of +/- 2 SD compares the mean accuracy of the Novel Decision Tree Regressor and the Linear Regression algorithm. X-axis: Linear Regression algorithm vs. Novel Decision Tree Regressor. Y-axis: Mean detection accuracy.

4. Discussion

The Performance of Novel Decision tree regressor and Linear regression is analyzed in sales forecasting of a retail store from a dataset obtained from Bigmart. In this study, it is observed that the Novel Decision tree regressor appears to have better accuracy than Linear regression with an accuracy of 97.5% but mean loss is high when compared to linear regression. The proposed work signifies that the Novel Decision tree regressor performs better classification compared to linear regression.

[15]The study (Rahul et al. 2021) aim turned into to broaden a machine that might correctly expect the discharge of these photovoltaic cells in order to place timely orders for supply and demand for space,ensuring service quality and cost savings usage using Decision Tree Regressor and archives accuracy of 90%. [16](Tofan 2014) this study mainly focuses on optimization techniques of decision tree regressor and improved accuracy by over 5%.[17] (Goetz 2011) In this work, the author studied the impact of health care on human survival and used this decision tree regressor predicted with an accuracy of 93%. [18] (Abdullah 2019) used linear regression for forecasting electricity consumption and achieved an accuracy of 85. Machine learning algorithms, which require big datasets and are excellent at dealing with nonlinear equations, were thought to be superior to a couple of linear regressions. Similarly, because the datasets are all annual, simple linear forecasting techniques are used solely based on the character of a time plot of data that is certainly dominated by fashion and has no seasonality characteristics. [19] (Scavuzzo et al. 2018)This paintings represents an development in that situation. They evaluate assist Vector Machines, synthetic Neural Networks, okay-closest friends, and decision Tree Regressor in more than direct methods. With this, found a mechanism that contributes to Argentina's Dengue risk program that is currently in operation and achieved an accuracy of 96.8% (Zhang, n.d.)[20]. This paper by [21](Ali et al. 2021) uses Decision tree regressor with neural networks for performance optimization and has an accuracy of 90%.(Luo et al. 2021)) [22] proposed a Soft Decision Tree Regressor (SDTR),a differentiable hierarchical neural regression model SDTR is a differentiable neural network that mimics a binary decision tree and is suitable for ensemble techniques such as bagging and boosting, as well as archiving the results of 95.34%.

The limitation of current work is only curated to the retail industry as comparison of algorithms done only based on one dataset and it cannot be sure that the result will be the same for all datasets. The mean loss of Novel Decision Tree regressor also seemed to be greater than Linear regression.Future work of study is to reduce loss of Decision Tree regressor by incorporating extreme gradient boosting techniques for Sales Forecasting. These Boosting techniques increase performances of Decision Tree Regressor by reducing loss that occurred while training the Model.

5.Conclusion

Novel Decision tree regressor has obtained a better accuracy of 97.5% than Linear regression Algorithm which has an accuracy of 87.6%. The precision of Sales forecasting has been significantly increased.

References

[1] Kuo, R. J., and K. C. Xue. 1998. "A Decision Support System for Sales Forecasting through Fuzzy Neural Networks with Asymmetric Fuzzy Weights." Decision Support Systems. https://doi.org/10.1016/s0167-9236(98)00067-0

[2] Thomassey, Sébastien, and Antonio Fiordaliso. 2006. "A Hybrid Sales Forecasting System Based on Clustering and Decision Trees." Decision Support Systems. https://doi.org/10.1016/j.dss.2005.01.008

[3] Guo, Z. X., W. K. Wong, and Min Li. 2013. "A Multivariate Intelligent Decision-Making Model for Retail Sales Forecasting." Decision Support Systems. https://doi.org/10.1016/j.dss.2013.01.026.

[4] Yuan, Fong-Ching, and Chao-Hui Lee. 2011. "Sales Volume Forecasting Decision Models." 2011 International Conference on Technologies and Applications of Artificial Intelligence. https://doi.org/10.1109/taai.2011.49.

[5] Ustundag, Alp, Emre Cevikcan, and Mehmet Serdar Kilinc. 2008. "SALES FORECASTING FOR A TURKISH PAINT PRODUCER: ARTIFICIAL INTELLIGENCE BASED METHODS VERSUS MULTIPLE LINEAR REGRESSION." Computational Intelligence in Decision and Control. https://doi.org/10.1142/9789812799470_0008.

[6] Pavlyshenko, Bohdan M. n.d. "Machine Learning Models for Sales Time Series Forecasting." https://doi.org/10.20944/preprints201811.0096.v1

[7] Bhukya, Devi Prasad, and S. Ramachandram. 2010. "Decision Tree Induction: An Approach for Data Classification Using AVL-Tree." International Journal of Computer and Electrical Engineering. https://doi.org/10.7763/ijcee.2010.v2.208

[8] Yusuf, Ayomide, and Shadi Alawneh. 2018. "GPU Implementation of Sales Forecasting with Linear Regression." International Journal of Innovative Research in Computer Science & Technology. https://doi.org/10.21276/ijircst.2018.6.4.1

[9] Grąbczewski, Krzysztof. 2014. "Techniques of Decision Tree Induction." Studies in Computational Intelligence. https://doi.org/10.1007/978-3-319-00960-5_2.

[10] Sekar, Durairaj, Ganesh Lakshmanan, Panagal Mani, and M. Biruntha. 2019. "Methylation-Dependent Circulating microRNA 510 in Preeclampsia Patients." Hypertension Research: Official Journal of the Japanese Society of Hypertension 42 (10): 1647–48.

[11] Johnson, Jayapriya, Ganesh Lakshmanan, Biruntha M, Vidhyavathi R M, KohilaKalimuthu, and DurairajSekar. 2020. "Computational Identification of MiRNA-7110 from Pulmonary Arterial Hypertension (PAH) ESTs: A New microRNA That Links Diabetes and PAH." Hypertension Research: Official Journal of the Japanese Society of Hypertension 43 (4): 360–62.

[12] Subashri, A., and M. S. Thenmozhi. 2016. "Occipital Emissary Foramina in Human Adult Skull and Their Clinical Implications." Journal of Advanced Pharmaceutical Technology & Research 9 (6): 716.

[13] Sriram, Nirisha, Thenmozhi, and SamrithiYuvaraj. 2015. "Effects of Mobile Phone Radiation on Brain: A Questionnaire Based Study." Journal of Advanced Pharmaceutical Technology & Research 8 (7): 867.

[14] Feng, Youli, and Shanshan Wang. 2017. "A Forecast for Bicycle Rental Demand Based on Random Forests and Multiple Linear Regression." 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS). https://doi.org/10.1109/icis.2017.7959977.

[15] Rahul, Rahul, Aakash Gupta, Ankur Bansal, and Kshitij Roy. 2021. "Solar Energy Prediction Using Decision Tree Regressor." 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS). https://doi.org/10.1109/iciccs51141.2021.9432322.

[16] Tofan, Cezarina Adina. 2014. "Optimization Techniques of Decision Making - Decision Tree." Advances in Social Sciences Research Journal. https://doi.org/10.14738/assrj.15.437

[17] Goetz, Thomas. 2011. The Decision Tree: How to Make Better Choices and Take Control of Your Health. Rodale.

[18] Abdullah, M. K. 2019. "Forecasting Electricity Consumption In Nigeria's Commercial Sector: A Linear Regression Approach." https://doi.org/10.15405/epsbs.2019.05.02.19.

[19] Scavuzzo, Juan M., Francisco Trucco, Manuel Espinosa, Carolina B. Tauro, Marcelo Abril, Carlos M. Scavuzzo, and Alejandro C. Frery. 2018. "Modeling Dengue Vector Population Using Remotely Sensed Data and Machine Learning." Acta Tropica 185 (September): 167–75.

[20] Zhang, Qinghua. n.d. "Regressor Selection and Wavelet Network Construction." Proceedings of 32nd IEEE Conference on Decision and Control. https://doi.org/10.1109/cdc.1993.325905

[21] Ali, Ahmed M. A., Hossam M. Zawbaa, Ossama M. Sayed, Hadeer S. Harb, Haitham Saeed, Marian S. Boshra, Ahmed G. Almeldien, et al. 2021. "In Vitro and in Vivo Performance Modelling and Optimisation of Different Dry Powder Inhalers: A Complementary Study of Neural Networks, Genetic Algorithms and Decision Trees." International Journal of Clinical Practice 75 (3): e13764.

[22] Luo, Haoran, Fan Cheng, Heng Yu, and Yuqi Yi. 2021. "SDTR: Soft Decision Tree Regressor for Tabular Data." IEEE Access. https://doi.org/10.1109/access.2021.3070575