IOSR Journal of Business and Management (IOSR-JBM)
e-ISSN:2278-487X, p-ISSN: 2319-7668. Volume 26, Issue 4. Ser. 6 (April. 2024), PP 64-69
www.iosrjournals.org

Using Machine Learning In Sales Prediction

Martins, Emerson[1]; Galegale, Napoleão Verardi[1]
[1](Ceeteps – Centro Estadual De Educação Tecnológica Paula Souza, Brazil)

Abstract:
Machine learning (ML) has been subjected to intense study in many different industries, and fortunately, companies are gradually becoming more aware of the various ML approaches to solving their problems. However, to obtain effective results from different algorithm models, it is necessary to have a good understanding of the application of the models and the nature of the data. This work aims to investigate different approaches to obtain good results from ML algorithms applied to predict product sales in the retail segment of a Japanese multinational company in the Brazilian market. Following the DSRM guidelines, a prototype computational solution was developed in the Python language, which implements different algorithm models based on linear regression, connected directly to a database where the history of product purchases and sales movements are stored. Using different configurations, the performance results of the algorithms are compared with each other and presented in the form of graphs and indicators, signaling to the data analyst which model is best based on the R-squared coefficient of determination (R2) and the RMSE metric. To obtain such results, data pre-processing mechanisms are applied, including: outlier identification, linear correlation level analysis, normality analysis and homoscedasticity analysis. The results presented show that ML algorithms can predict sales values with better approximation compared to traditional methods such as moving averages or arithmetic averages, contributing to the organization's financial predictability and making a relevant contribution to academia when applying Design of Experiments (DOE) in conducting internal evaluations to identify the algorithm with the best performance.

Keywords*: Machine Learning, Algorithm, Big Data Analysis, Bibliometric Analysis, Predictive*

Date of Submission: 11-04-2024 Date of Acceptance: 21-04-2024

I. Introduction

Machine Learning (ML) is the branch of science where computer algorithms are developed to perform tasks without human guidance, rather than relying on codified rules. In other words, ML is the ability of computers to induce new knowledge, such algorithms have been used widely and successfully in many areas (MAXWELL *et al.*, 2015).

In the last two decades, computing power has evolved considerably, in this context we can mention: large data storage, more robust processors, faster internet connection, among other examples. Problems that seemed extremely complex or costly to solve are now within our reach. New trends such as Big Data, Cybersecurity, internet of things (IoT) and blockchain have emerged, jointly exploring the technological advances mentioned above. IoT, which aims to use embedded systems, including sensors and actuators, together with the internet, to allow control and immediate access to information in real time (Atzori, 2010; Cecchinel, 2014), represents one of these challenges, as the report from "Juniper Research", reports that by 2024 we will have more than 83 billion connected devices and sensors (IoT). Furthermore, some of these devices will have the capacity to generate a significant amount of data in the order of Zettabytes, information that can be valuable for a company's strategy, therefore, sales forecasting cannot ignore these new trends; it must use it as support for competitive advantage.

In turn, predictive analysis encompasses methods that use information to create models and carry out simulations that will provide insights into future events, allowing the most attentive executives to predict strategic actions that improve their company's performance (MARTINS et al, 2023).

By definition, the results obtained through these techniques are not 100% accurate, as no method can predict the future, therefore, a good predictive analysis is one that provides the most accurate results in a reasonable time (CASTILLO *et al.*, 2017, MARTINS et al, 2022).

One of the common uses of predictive analytics in business is sales forecasting, but there are also several other applications in domains such as: cost estimation, where (LOYER *et al.*, 2016) applied ML techniques to quickly estimate the cost of manufacturing components of aircraft engines, since classical cost estimation methods, although quite accurate, are expensive and slow. In performance assessment, (FAN *et al*, 2013) used ML to estimate supply chain performance based on the "5 Dimensional Balanced Scorecard" (5DBSC) and "Levenberg-Marquadt Back Propagation" (LMBP) with the aim of providing results quickly and avoid biased performance evaluations by managers.

The general objective of this research is to direct the development of a prototype computational solution, based on ML, where through internal variables contained in the sales history such as: product category, date of sale, quantity, sales value and external variables such as IPCA inflation index, it is possible to price future sales, helping the company to have revenue predictability, making its financial planning viable. Additionally, it is evident that ML algorithms based on linear regression present better performance compared to traditional methods such as arithmetic mean.

Companies, as productive systems, must use their resources efficiently and make strategic decisions to obtain growing and stable revenues, especially when market conditions are becoming more competitive and profit margins are increasingly under pressure. Therefore, sales forecasting is crucial to maintaining competitiveness, however, obtaining inaccurate forecasts can lead to stock shortages, causing delays in deliveries and generating customer dissatisfaction, as well as increasing inventory, increasing storage costs, forcing the "burning" of stock through promotional campaigns, directly affecting companies' profitability. (HOFMANN *et al*., 2018).

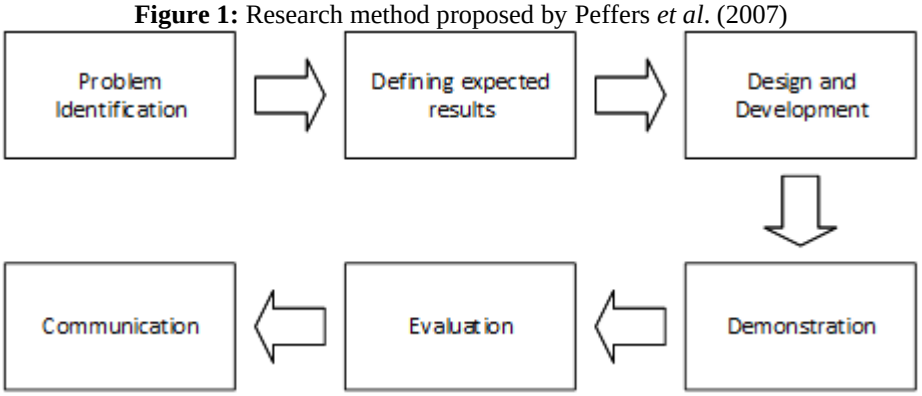
The motivation for this research is based on research on the ML approach applied to sales forecasting in the retail segment, which supports managers in decision making, contributing to supply chain operations, sales forecasts, inventory management, production planning and workforce scheduling. Thus, contributing to increased profits and reduced costs (RANGAPURAM, 2018; SALINAS, 2020).

II. Material And Methods

This article uses the Design Science Research Methodology (DSRM) in Information Systems to conduct applied research. Therefore, this study followed the DSRM guidelines based on Hevner *et al*. (2004) and Peffers *et al*. (2006, 2007). The DSRM method is a problem-solving process that allows researchers to acquire knowledge and understanding of a problem domain and its solution through the creation and application of IT artifacts (HEVNER *et al*., 2004).

Peffers *et al*. (2006, 2007) present a conceptual process for DSRM in information systems based on frameworks in design research studies. The Design Science Research Process (DSRP), that is, the design science research process consists of six distinct activities, which can be performed in sequential order or not, as the research method can be used differently, its starting point being modified according to the researcher's objectives. The expected output of each step is presented in Figure 1. The DSRP is closely related to the DSRM guidelines of Hevner et al. (2004), as it also highlights the importance of an adequate statement of the problem, the construction of a viable solution and the evaluation of results regarding their usefulness, as well as their communication.

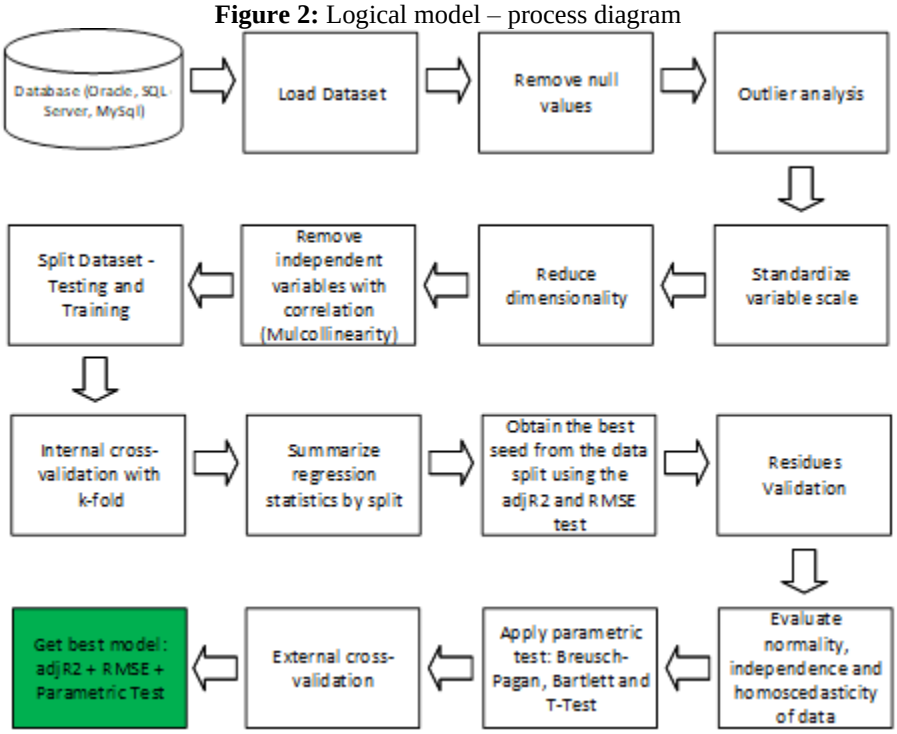
Figure 1: Research method proposed by Peffers *et al*. (2007)



Source: Adapted from Peffers *et al.* (2007)

The method proposed in this research is defined in an integrated flowchart and developed with the objective of creating and comparing multiple linear regression models, initially introduced by (Tsiliki *et al*., 2015a). This model was adapted in this research as shown in Figure 2. This diagram is implemented and tested using a computational solution developed in Python.

Figure 2: Logical model – process diagram



Source: Author

Focusing on the last step of the flowchart, the best linear regression model of this methodology was selected based on the following criteria: I) The performance of all models was evaluated (including with the test data set) according to the R- squared (R2), establishing a ranking. II) Then, the R-squared result (R2) of all models that had a variation of (0.05) was taken into consideration, with this a new ranking was reordered, in which the models that presented the lowest (RMSE) obtained in the tests.

Thus, the best model of all runs is obtained, combining the use of two performance measures with a specific criterion.

III. Result

Sales forecasting is not a trivial task, since the data required for such analysis usually has large volumes, presents noise, excess categories and several other problems, in addition to the difficulty of selecting the most appropriate model for the problem that will be addressed. analyzed.

However, such analyses can strengthen the marketing of specific products, allow for better inventory management and optimize the negotiation process with suppliers. To achieve the aforementioned benefits, it is necessary to explore a set of pre-processing techniques for the collected data, in addition to modeling and evaluating prediction tools.

This research presents an approach based on ML models, to infer sales pricing and analyse the causal relationships between the variables that influence future sales.

In this research, data were obtained from a company that operates in the Brazilian retail market, and the selection of the company for this field study did not occur randomly. By presenting a structured sales forecasting method using simple arithmetic average, the company was invited to participate in the.

This is a Japanese multinational company operating in the Brazilian retail market since 1972, in which it agreed to participate in the research as long as its identity was not revealed nor confidential information or strategic data.

To ensure that the sample of selected products effectively represents the population subject to this study, products were selected from three different lines of business, which represent 90% of the company's revenue. Within each business line, products with a movement history of over five years were selected.
The selected products are classified into three groups as shown in Table 1.

Table 1: Number of inventory transactions per product

Product	Business Line	Transactions	Period
13M1S1	Biometric Scanner	42.887	2006 a Jun/2018

0-0001	Consumable for image scanner	2.998	2008 a May/2022
0-B051	Image Scanner	1.973	2014 a May/2022
8-K011	Consumable for image scanner	2.642	2004 a May/2022
1-B301	Image Scanner	568	2016 a May/2022

Source: Research results

Once the assumptions of normality, independence and heteroscedasticity have been met, the data sets are submitted to ML algorithms using the seed that presented the best accuracy in the Cross-Validation process.

The dataset containing the historical movements of the products was separated into 70% for training and 30% for testing, in addition, the seed that presented the best accuracy in the internal cross-validation process was used.

Table 2 highlights the results obtained through a comparative matrix between the algorithms. The results presented in dark gray highlight the models with the best performance.

Table 2: Comparative matrix of models

		13M1S1	0-0001	0-B051	8-K011	1-B301
Linear Regression	R2	0.9859703	0.9655112	0.9884350	0.9698459	0.9961951
	RMSE	30.634	204	15.366	556	1.091
Random Forest Regressor	R2	0.9676439	0.9558832	0.9700913	0.9626484	0.9699347
	RMSE	46.522	231	24.711	619	3.069
Support Vector Regression	R2	0.9913870	0.9716644	0.9687610	0.9639821	0.9460529
	RMSE	24.002	185	25.254	608	4.111
Lasso Regression	R2	0.9859438	0.9643137	0.9884473	0.9702801	0.9962689
	RMSE	30.663	208	15.358	552	1.081
Ridge Regression	R2	0.9859974	0.9659043	0.9886798	0.9702801	0.9980811
	RMSE	30.604	203	15.202	552	775
ElasticNet Regression	R2	0.9862212	0.9658657	0.9896923	0.9717601	0.9971229
	RMSE	30.359	203	14.507	538	949
Simple Arithmetic Mean	R2	0.9403611	0.6257375	0.8977438	0.7912402	0.9976336
	RMSE	63.161	673	45.692	1.464	861
Normal Distribution (Shapiro-Wilk)	p-value > 0.05	No	No	Yes	No	Yes
Correlation		-	-	Pearson	-	Pearson
		Spearman	Spearman	-	Spearman	-
Homoscedasticity	Breusch-Pagan Test	Yes	No	No	No	No
	Bartlett Test	-	Yes	Yes	Yes	No
	T-Test	-	-	-	-	Yes

Source: Research results

It is evident that the SVR algorithm presented the best prediction for products 13M121 and 0-0001 with R-squared (R2) of 0.9914 and 0.9717, in addition to RMSE of 24.002 and 185 respectively. For products 0-B051 and 8-K011, the best algorithm was ElasticNet with R-squared (R2) of 0.9897 and 0.9718, in addition to RMSE of 14.507 and 538 respectively. The Ridge algorithm presented the best prediction for product 1-B301 with R-squared (R2) of 0.9981 and RMSE of 775.

In this research, the same methodology for formalizing experimental designs used by Tsiliki et al. (2015a) was adapted with cross-validation and statistical tests and proven on a set of data in the retail segment related to the regression task. The results showed excellent performance of the final models selected and demonstrated the importance of taking into account the variability of the models to choose the best one. This selection should not be based solely on the best result achieved in a single metric. In this research, the performance of all models (with the test set) is evaluated, using R2-Squared to establish a ranking of the models. In the next step, the normality, independence and homoscedasticity of the data are taken into account, selecting the model with the lowest RMSE obtained in the test.

The results also allowed us to understand that selecting the model only according to the best R2-Squared or the highest precision and taking into account only one execution is not statistically valid. An outlier or unusual value in the dataset may be an outlier and may arise due to a specific partition of the data during

model learning and should be discarded in favor of more stable methods. Aspects such as stability in results must be taken into account when making this selection.

IV. Discussion

This research proposes the combination of two profound modifications necessary in the methodology introduced by Tsiliki *et al*. (2015a), i.e., external cross-validation process in the learning phase and statistical analysis in the best model selection phase, in addition to a simple consideration of data diversity, in order to deal with counting data during the pre- data phase processing.

The findings of this research are relevant and it can be assumed that other ML models, in general, should behave in a similar way to those presented in this research, as well as the same algorithms with other data sets. Generally speaking, if the training process or algorithms are stochastic in some way, one should repeat several experimental runs to find the best results, and it is crucial to use a statistical comparison to evaluate whether differences in performance scores are relevant. before finally choosing the best model. Otherwise, the final conclusions may be wrong and the final model should be discarded.

Finally, it was evident that the methodology used in this research is applicable in other fields, as it is quite flexible for adding new phases, and can be applied to other data sets of different scopes. At the same time, according to the results presented, it can be stated that the same behavior is expected with other ML algorithms, and it is expected that further statistical studies will be relevant in this sense.

The main objective of this type of methodology in predictive modeling is to help in the initial exploration of extensive databases in order to design comprehensive screening methods in the retail market, reducing economic costs as much as possible and supporting the company's financial predictability and at the same time, ensuring the quality and reliability of the ML methods used in the process.

In order to demonstrate that the computational solution artifact developed from this research contributes significantly to sales prediction, the evaluation phase was divided between the comparative results of the models (DOE), interview with the company's user (internal evaluation) and interview with experts (external evaluation).

The interviews carried out with the experts were carried out based on the questionnaire previously sent to them, with the purpose of collecting information, so that it was possible to analyze the level of adherence of the computational solution prototype in predicting the unit sales value. Thus, respondents had to answer eleven questions, using only one of the available options, such as: “Disagree”, “Partially Agree” or “Totally Agree”. In addition to having a comment field for each question, at the end of the questionnaire the following question was made available: “Which question do you think is important and was not addressed in this research?”, offering space for the specialist to describe their suggestions.

The information acquired was organized, considering the opinion of each specialist in relation to each issue. Thus, based on the results collected, it is clear that the eleven questions were mostly classified between the criteria of “Partially Agree” and “Totally Agree”. Therefore, these criteria are considered positive answers and only two questions in particular received a response classified as “Disagree”.

In Table 3, the result of the composition of points per group of questions is presented, which met the following criteria: from 0 to 1 point for “Disagree”; from 1.1 to 2.0 points for “Partially Agree” and from 2.1 to 3.0 points for “Totally Agree”.

Table 3: Interview results by group of questions

Question Groups	Point average	Result
(1) - Quantitative analysis of the dataset	2,17	Totally Agree
(2) - Relationship of time series in sales forecasting	1,50	Partially agree
(3) - Prediction using linear regression	2,67	Totally Agree
(4) - ML Algorithms Based on Linear Regression	2,50	Totally Agree

Source: Research results

Observing the results presented in Table 3, it is noted that the computational solution prototype managed to achieve a high level of adherence among experts, with three groups of questions showing “Completely Agree” and one question group showing “Partially Agree”.

V. Conclusion

This research proposed a new multidisciplinary method for carrying out experiments in Computational Intelligence, which was implemented through some libraries available in the Python language. Six computational intelligence algorithms based on linear regression and five data sets representing the sales history of different products were used in this regard. Different critical modifications were proposed and tested in

various phases of the method previously published by Tsiliki, and the results obtained were relevant. For a better understanding of the research, all steps were applied and demonstrated to the baseline example of an experimental study on datasets.

The final phase of the statistical study was described in detail, as it is the most critical modification, as well as external cross-validation. From the five data sets, the results of which were compared with simple prediction methods such as simple arithmetic mean, it can be concluded that, following the methodology of this research, the results can be verified for statistical significance and therefore reliable in models predictive and must be proposed to the scientific community. Furthermore, with the method demonstrated in this study it is possible to solve linear regression problems in other research scopes, obtaining stable, reproducible and relevant results.

From the results presented, it is clear that there is no better algorithm than another, that is, the behaviour of the data and the correlation between its variables will directly influence obtaining the best model for a specific product. In this way, the prototype computational solution resulting from this research is useful and relevant as it indicates to the researcher which model is best for each type of product.

No method can predict the future, however with the support of new technologies it is possible to obtain an effective prediction, which is essential for several segments, especially in the retail sector which historically works with low profit margins in a highly competitive market.

Although this is a specific case study of a company reselling imported products in the image scanning and biometrics segment, this study can be generalized to other companies since the predictor variables such as "Quantity of sales", "Value of unit cost", "Percentage of gross margin" and "Percentage of inflation for the month - IPCA", may be part of the business context of other companies.

However, we recognize the limitations of this research regarding the available processing capacity, limiting the data set to a set of five products. For future research, the use of other products in different retail segments is suggested.

For future research, it is suggested to explore the relationship between profit margin and sales revenue, comparing the history of operating profit related to sales revenue, obtaining a model that can predict what the target revenue would be so that the company can maintain its sustainability financial.

References

[1]. Atzori, L.; Iera, A.; Morabito, G. The Internet Of Things: A Survey. Computer Networks. 2010.

[2]. Castillo, P. A.; Mora, A. M.; Faris, H.; Merelo, J. J.; García-Sánchez, P.; Fernández-Ares, A. J.; Cuevas, P. D.; García-Arenas, M. I. Applying Computational Intelligence Methods For Predicting The Sales Of Newly Published Books In A Real Editorial Business Management Environment - Knowledge-Based Systems, 115:133–151, 2017.

[3]. Cecchinel, C.; Jimenez, M.; Riveill, M.; Mosser, S. An Architecture To Support The Collection Of Big Data In The Internet Of Things. Ieee World Congress On Services. 2014.

[4]. Fan, X.; Zhang, S.; Wang, L.; Yang, Y.; Hapeshi, K. An Evaluation Model Of Supply Chain Performances Using 5dbsc And Lmbp Neural Network Algorithm - Journal Of Bionic Engineering. 2013.

[5]. Hevner, A.R.; March, S.T.; Park, J.; Ram, S. Design Science In Information Systems Research. Mis Q., 28, 75–105. 2004.

[6]. Hofmann, E.; Rutschmann, E. Big Data Analytics And Demand Forecasting In Supply Chains: A Conceptual Analysis - The International Journal Of Logistics Management - Vol. 29 No. 2, 2018 - Pp. 739-766 - © Emerald Publishing Limited - 0957-4093 - Doi 10.1108/Ijlm-04-2017-0088, 2018.

[7]. Juniper Research. 2020. Available In: <Https://Www.Juniperresearch.Com/Press/Iot-Connections-To-Reach-83-Bn-By-2024?Ch=Iot%20connections%20to%20grow>. Acesso Em 29 Maio 2022.

[8]. Loyer, J.L.; Henriques, E.; Fontul, M.; Wiseall, S. Comparison Of Machine Learning Methods Applied To The Estimation Of Manufacturing Cost Of Jet Engine Components. -International Journal Of Production Economics. 2016.

[9]. Martins, E.; Galegale, N. V. Sales Forecasting Using Machine Learning Algorithms. Revista De Gestão E Secretariado, V. 14, P. 11294-11308, 2023.

[10]. Martins, E ; Galegale, N. V. . Retail Sales Forecasting Information Systems: Comparison Between Traditional Methods And Machine Learning Algorithms. In: 15th Iadis International Conference Information Systems 2022, 2022, Porto.

[11]. Maxwell, A. E.; Warner, T. A.; Strager M. P.; Conley, J.F.; Sharp, A.L. Assessing Machine-Learning Algorithms And Image-And Lidar-Derived Variables For Geobia Classification Of Mining And Mine Reclamation. International Journal Of Remote Sensing Vol 36, 2015.

[12]. Peffers, K.; Tuunanen, T.; Gengler, C.E.; Rossi, M.; Hui, W.; Virtanen, V.; Bragge, J. The Design Science Research Process: A Model For Producing And Presenting Information Systems Research. In Proceedings Of The First International Conference On Design Science Research In Information Systems And Technology (Desrist 2006), Claremont, Ca, Usa, 24–25, 2006.

[13]. Peffers, K.; Tuunanen, T.; Rothenberger, M.A.; Chatterjee, S. A Design Science Research Methodology For Information Systems Research. J. Manag. Inf. Syst., 24, 45–77, 2007.

[14]. Rangapuram, S. S.; Seeger, M. W.; Gasthaus, J.; Stella, L.; Wang, Y.; Januschowski, T. Deep State Space Models For Time Series Forecasting, In: Advances In Neural Information Processing Systems, Pp. 7785–7794, 2018.

[15]. Salinas, D.; Flunkert, V.; Gasthaus, J.; Januschowski, T. Deepar: Probabilistic Fore- Casting With Autoregressive Recurrent Networks, Int. J. Forecast. 36 (3), 1181–1191, 2020.

[16]. Tsiliki, G.; Munteanu, C.R.; Seoane, J.A.; Fernandez-Lozano, C.; Sarimveis, H.; Willighagen, E. L. Rregrs: An R Package For Computer-Aided Model Selection With Multiple Regression Models. Journal Of Cheminformatics 7:1-16, (2015a).

[17]. Tsiliki, G.; Munteanu, C. R.; Seoane, J. A.; Fernandez-Lozano, C.; Sarimveis, H.; Willighagen, E. L. Using The Rregrs R Package For Automating Predictive Modelling. In: Mol2net, International Conference On Multidisciplinary Sciences, (2015b).