






Using “Machine Learning” Techniques in Increasing the Efficiency of Sales Forecasting in Albania

Valma Prifti^(✉) , Dea Sinoimeri , Armira Lazaj , Betina Dini, and Kevin Luniku

Polytechnic University of Tirana, 1010 Tirana, Albania

vprifti@fim.edu.al

Abstract. This paper investigates the utilization of a Machine Learning (ML) approach with the objective of selecting an appropriate model for sales forecasting. Within this study, three ML algorithms are examined: Simple Linear Regression, Gradient Boosting Regression, and Random Forest Regression. A comparative analysis of these algorithms is conducted using two performance metrics: Accuracy Score and Max Error. The significance of sales forecasting cannot be overstated, as it plays a critical role across various industries. Therefore, the application of ML technology is essential to mitigate potential financial losses resulting from inaccurate demand assessments. A retail company based in Albania, which provided historical data as input for the model, is utilized as a case study. The Random Forest Model demonstrates exceptional performance, characterized by minimal deviations between predicted and actual values. The findings of this research endeavor present a pioneering initiative that holds significant potential for enhancing the forecasting of future sales and delivering substantial benefits to firms operating in the Albanian market.

Keywords: Technology · Machine learning · Algorithms · Sales forecast

1 Introduction

Data Mining is described as a process of extracting valuable information from a large collection of raw data, using statistics, artificial intelligence, machine learning, and pattern recognition methods. Machine Learning (ML) is a field of study that enables machines to learn without being explicitly programmed. ML is defined as computer programs that learn from experience. There are three categories of machine learning algorithms. Figure 1 provides an overview of Machine Learning and the techniques applied to solve various tasks. In the fast-paced world of sales forecasting, accurate predictions are crucial for entrepreneurs to optimize their operations, improve inventory management, and maximize profits. Traditional sales forecasting methods often struggle to capture the complex dynamics of consumer behavior and the intricate relationships between various market factors. However, with the advent of Machine Learning (ML) techniques, there is a significant opportunity to use advanced algorithms and engineering principles to revolutionize the efficiency of sales forecasting processes. The essence of

this paper lies in the application aspects of ML techniques for sales forecasting. By employing principles such as data processing, feature engineering, model selection, and optimization, engineers can develop powerful and efficient ML models that surpass traditional forecasting approaches. This focused approach aims to address the challenges entrepreneurs face in handling large volumes of data, managing complex models [1], and ensuring scalability and reliability in real-world sales forecasting scenarios. By analyzing historical sales data, customer demographics, and relevant market variables, we aim to evaluate current forecasting practices and identify areas where ML algorithms can significantly enhance accuracy and efficiency in a real retail company.

The findings of this research will provide valuable insights for engineers, data scientists, and entrepreneurs seeking to improve their sales forecasting capabilities. By embracing machine learning techniques with a focus on engineering principles, organizations can utilize advanced analytics to gain a competitive advantage, improve resource allocation, and foster informed decision-making processes.

2 Materials and Methods

In Supervised Learning, the algorithm utilizes labeled data and attempts to find the label that corresponds to these data based on given features. If the label is continuous, it involves linear regression. In the case of categorical labels, classification algorithms are used. The Unsupervised Learning technique is employed when only feature data is available [3]. The most common application is clustering, where data with only features are divided into groups based on similarities. The Reinforcement Learning technique is used to solve much more complex problems, such as teaching a computer how to play music or drive a car. It involves three components: the Agent, which makes decisions; the Environment, with which the agent interacts, and the Action, which is taken by the agent.

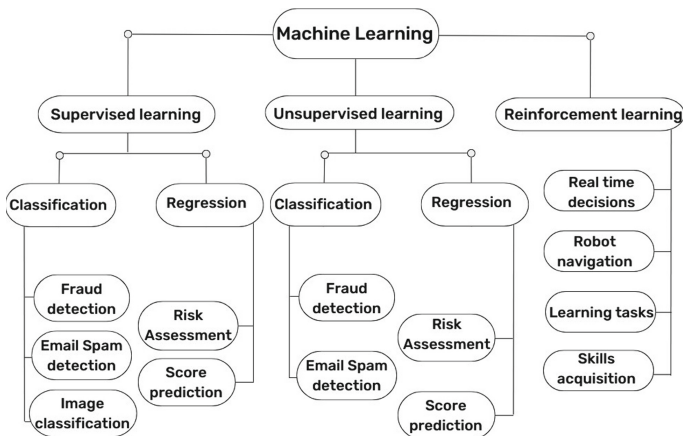


Fig. 1. Tree representation of machine learning approach, techniques, and tasks for which they are used.

2.1 Selection of Machine Learning Algorithms

Forecasting involves discovering possible future events, usually based on past data. In this paper, the algorithms used belong to the category of supervised learning, such as Simple Linear Regression, Random Forest Regression, and Gradient Boosting Regression. These algorithms [5], can facilitate finding better results compared to traditional analytical time series techniques.

- Simple Linear Regression is useful for determining the relationship between two continuous variables. This type of regression requires a non-deterministic statistical relationship.
- Gradient Boosting Regression is based on the premise that, when combined with previous techniques, iterative refinement minimizes the maximum prediction error.
- Random Forest Regression is a type of ensemble method that allows predictions by integrating decisions from a series of simple models.

2.2 Metrics for Evaluating Model Effectiveness

To assess the effectiveness of the models in a more objective manner, several metrics are used. The main objective of this study was to compare the performance of Machine Learning techniques by applying performance metrics such as accuracy score and maximum error.

Accuracy Score

This metric is known as the ratio of correct predictions to the total number of predictions (data points) (Developers 2020a), and is calculated using the formula:

$$Accuracy\ Score = \frac{TN + TP}{TN + TP + FN + FP}$$

where TN is true negative, TP is true positive, and (TN + TP + FN + FP) is the total number of predictions.

Max Error

Max Error is a metric that measures the maximum standard deviation and represents the worst- case error between the predicted value and the actual value (Developers 2020b). Max Error is calculated using the formula:

$$Max\ Error(y, x) = \max(|y_i - x_i|)$$

where y_i represents the actual values and x_i represents the predicted values.

The chosen methodology for conducting this study is a case study. To achieve the final goal, we will go through several steps:

1. Conduct an in-depth literature review to gather sufficient information on machine learning techniques.
2. Evaluate predictive models by applying performance metrics to measure their efficiency.
3. Identify the suitable algorithm for sales forecasting in the selected company.
4. Gather data from the company for the application of the ML predictive model.
5. Visualize the predicted results of the model on future sales of the company.

3 Results and Discussions

The Machine Learning predictive models considered for comparison were three: Simple Linear Regression, Gradient Boosting Regression, and Random Forest Regression. The most suitable model for achieving the objective of this research demonstrates the highest value of the selected metrics. The data collected for this research work is confidential, and to maintain the company’s privacy, the five considered articles are labeled with Arabic numerals. They are presented in Table 1, where for each article, the price in Euros, the quantity sold in each year, and the revenue generated because of sales are recorded.

Table 1. Data collected from the wholesale company for the years 2015–2018.

Year	Product	Price (Euro)	Sale (Pcs)	Income (Euro)
2015	1	15	6210	93150
	2	25	5422	135550
	3	10	6005	60050
	4	37	7115	263255
	5	99	7250	717750
<i>Total</i>			<i>32002</i>	<i>1269755</i>
2016	1	15	6821	102315
	2	25	7340	183500
	3	10	7410	74100
	4	37	7500	277500
	5	99	7415	734085
<i>Total</i>			<i>36486</i>	<i>1371500</i>
2017	1	15	7601	114015
	2	25	7589	189725
	3	10	7840	78400
	4	37	7795	288415
	5	99	8005	792495
<i>Total</i>			<i>38830</i>	<i>1463050</i>
2018	1	15	9550	143250
	2	25	9883	247075
	3	10	1058	10580
	4	37	11000	407000
	5	99	11010	1089990
<i>Total</i>			<i>42501</i>	<i>1897895</i>

The first step of the analysis is to study the dataset, which contains information on sales from the wholesale company. The graph presented in the figure illustrates the behavior of the company’s current sales (Fig. 2).

On the other hand, the revenue follows a proportional trend with the sales. As shown in Fig. 3, most of the graph is dominated by the year 2018, represented in blue, followed by 2017 in yellow, 2016 in orange, and finally 2015, depicted in green.

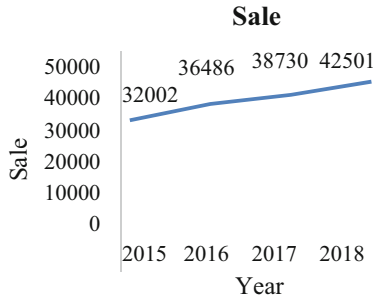


Fig. 2. The sales trend of the company from 2015 to 2018.



Fig. 3. Distribution of company revenue by year.

The final step is the use of the model for predicting the revenue from wholesale sales. The chosen [14], Machine Learning algorithm, which was deemed most suitable for the task, was Random Forest. Based on the provided historical data, the model successfully made sales predictions for the next two years, specifically 2019 and 2020.

After applying the metrics to each of the three Machine Learning algorithms, the following results were obtained.

3.1 Simple Linear Regression

The graph presented in Fig. 4 shows the accuracy score results obtained from the Simple Linear Regressor. It can be observed that the maximum accuracy score is 84.099%, the average accuracy score is 81.2%, and the minimum accuracy score is 73.95%.

The graph presented in Fig. 5 displays the Maximum Error (ME) obtained from the Simple Linear Regressor, where the maximum value is 0.5118%, the average value is 0.4917%, and the minimum value is 0.4731%.



Fig. 4. ACC graph for simple linear regression.

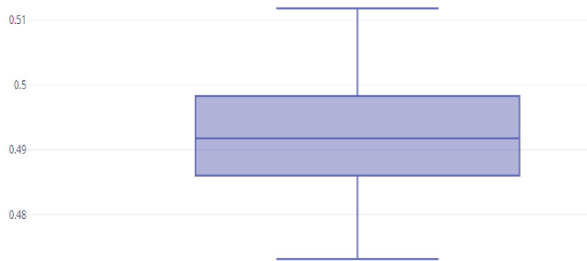


Fig. 5. ME graph for simple linear regression.

3.2 Gradient Boosting Regression

The graph presented in Fig. 6 shows the accuracy (ACC) results obtained from the Gradient Boosting Regressor, where the maximum accuracy value is 91.2%, the average accuracy is 86.27%, and the minimum accuracy is 78.3%.

In the graph of Fig. 7, the maximum error (ME) values are provided, where the highest value is 0.464, the average ME value is 0.441, and the achieved minimum value is 0.425.

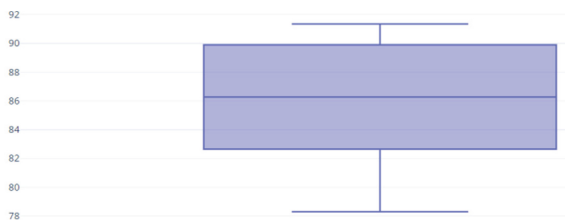


Fig. 6. ACC graph for gradient boosting regression.

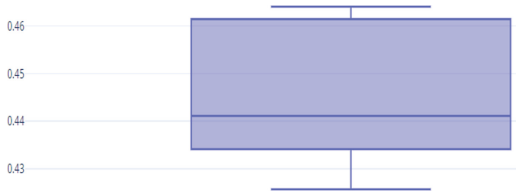


Fig. 7. ME graph for gradient boosting regression.

3.3 Random Forest Regression

Figure 8 presents the graph of accuracy (ACC) results obtained from the Random Forest Regressor, where the maximum achieved accuracy value is 91.35%, the average accuracy is 87.72%, and the minimum accuracy is 78.31%.

In the graph presented in Fig. 9, the Maximum Error (ME) obtained from the Random Forest Regressor is shown, where the maximum value achieved is 0.6568, the average ME is 0.6135, and the minimum value is 0.5964.

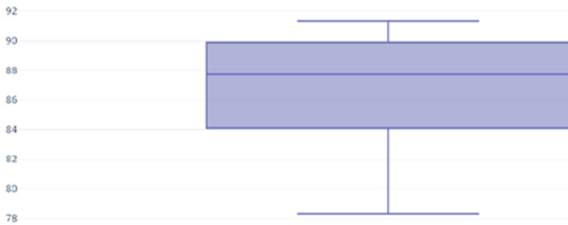


Fig. 8. ACC graph for random forest regression.

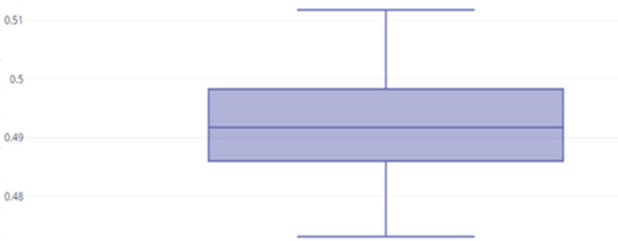


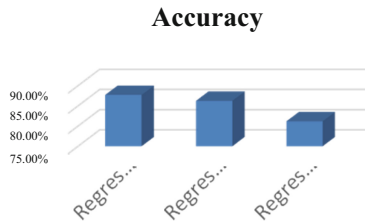
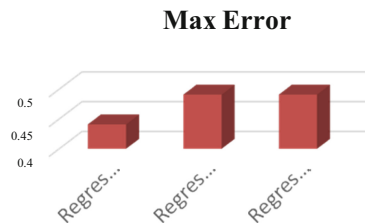
Fig. 9. ME graph for random forest regression.

To gather and simplify the understanding of readers, the obtained results from the comparison of algorithms are presented in Table 2.

Figure 10 illustrates the Accuracy Score (ACC) for Random Forest Regression, Gradient Boosting Regression, and Simple Linear Regression. It can be observed that Random Forest Regression shows the highest accuracy score. Figure 11 presents the Max Error (ME) value for each selected Machine Learning algorithm. Random Forest Regression exhibits the lowest error value.

Table 2. Results of accuracy score and max error metrics for corresponding machine learning algorithms

Algorithms	Accuracy scor (%)e	Max error
Random forest regression	87.72	0.441
Gradient boosting regression	86.27	0.491
Simple linear regression	81.20	0.491

**Fig. 10.** Accuracy values**Fig. 11.** Max serror values**Table 3.** Sales and revenue values

Year	Actual sales	Predicted sales	Actual revenue	Predicted revenue
2019	42993	43070	2000102	2003684
2020	42840	42904	1901152	1903992

The sales and revenue graphs are constructed separately because their units of measurement and value ranges are completely different from each other (Table 3).

To visualize the sales and revenue values of the company presented in the table and to understand the proximity of the actual values provided by the company to those predicted by the Random Forest Model, the following two graphs have been constructed. Figure 12 illustrates the behavior of sales, where the columns in blue represent the actual sales, while the columns in red represent the sales predicted by the Model (Fig. 13).

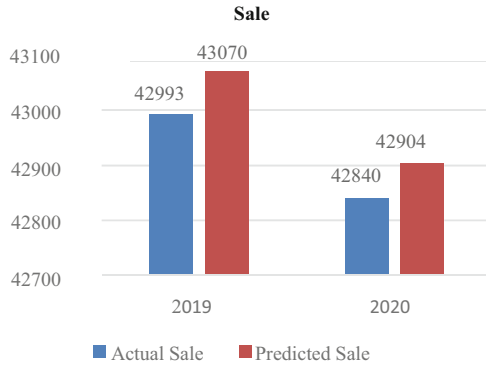


Fig. 12. Graphical representation of the behavior of actual and predicted sales for the years 2019 and 2020.

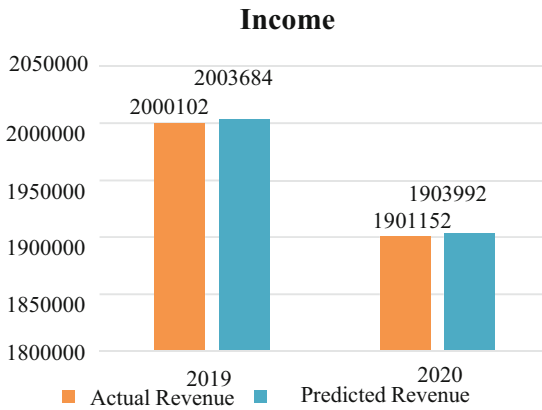


Fig. 13. Graph of the trend of actual and predicted revenues for the years 2019 and 2020.

4 Conclusion

The comparative analysis of these Machine Learning models: Simple Linear Regression Gradient Boosting Regression, and Random Forest Regression; yielded important results for sales forecasting. In terms of accuracy scores, the Simple Linear Regression model achieved a maximum score of 84.099%, with an average of 81.2% and a minimum of 73.95%. The Gradient Boosting Regression model performed even better, with a maximum accuracy score of 91.2%, an average of 86.27%, and a minimum of 78.3%. However, the Random Forest Regression model outperformed both, obtaining the highest maximum accuracy score of 91.35%, an average of 87.72%, and a minimum of 78.31%. When considering maximum error (ME), the Simple Linear Regression model exhibited a maximum error of 0.5118%, an average of 0.4917%, and a minimum of 0.4731%. The Gradient Boosting Regression model showcased a higher level of accuracy, with a maximum error of 0.464, an average of 0.441, and a minimum of 0.425. The Random Forest Regression model, however, demonstrated the lowest maximum error of 0.6568,

an average of 0.6135, and a minimum of 0.5964. Based on these findings, it can be concluded that the Random Forest Regression model offers the most accurate and precise sales forecasting capabilities among the three models evaluated. With a maximum accuracy score of 91.35% and the lowest maximum error, this model proves to be the most suitable choice for accurately predicting sales in the wholesale industry.

5 Acronyms

ML	Machine Learning
ACC	Accuracy Score
ME	Maximum Error
TN	True Negative Numbers
TP	True Positive Numbers
FN	False Negative numbers
FP	False Positive numbers

References

1. Bangdiwala, S.I.: Regression: simple linear. *Int. J. Inj. Contr. Saf. Promot.* **25**(1), 113–115 (2018)
2. Hanssens, D.M.: Order forecasts, retail sales, and the marketing mix for consumer durables. *J. Forecast.* **17**(3–4), 327–346 (1998)
3. Prifti, V., Dhoska, K.: Information systems in project management and their role in decision making. *Int. J. Tech. Phys. Prob. Eng.* **14**(4), 189–194 (2022)
4. Kahn, K., Adams, M.: Sales forecasting as a knowledge management process. *J. Bus. Forecast. Meth. Syst.* **19**(4), 19 (2001)
5. Harrison, P.J.: Short-term sales forecasting. *J. Royal Stat. Soc. Series C Appl. Stat.* **14**(2/3), 102–139 (1965)
6. Prifti, V., Aranitasi, M.: E-commerce business model in KLER enterprise for shirt manufacturing. *Int. J. Innov. Technol. Interdiscipl. Sci.* **5**(1), 858–864 (2022)
7. Prifti, V., Sinoimeri, D., Lazaj, A., Keci, J.: Impact of the information systems and technology on enterprises. *J. Integr. Eng. Appl. Sci.* **1**(1), 23–31 (2023)
8. Mentzer, J.T., Cox J.R., J.E.: Familiarity, application, and performance of sales forecasting techniques. *J. Forecast.* **3**(1), 27–36 (1984)
9. Ray, S.A.: Quick review of machine learning algorithms. In: *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 35–39 (2019)
10. Prifti, V., Dervishi, I., Dhoska, K., Markja, I., Pramono, A.: Minimization of transport costs in an industrial company through linear programming. *IOP Conf. Ser.: Mater. Sci. Eng.* **909**, 012040 (2020)
11. Morgan, M.S., Chintagunta, P.K.: Forecasting restaurant sales using self-selectivity models. *J. Retail. Consum. Serv.* **4**(2), 117–128 (1997)
12. Pavlyshenko, B.M.: Machine-learning models for sales time series forecasting. In: *Data*, vol. 4 (2019)
13. Skorikov, M., Momen, S.: Machine learning approach to predicting the acceptance of academic papers. In: *International Proceedings of Conference on Industry 4.0, Artificial Intelligence, and Communications Technology*, pp.113–117 (2020)

14. Prifti, V., Markja, I., Dhoska, K., Pramono, A.: Management of information systems, implementation, and their importance in Albanian enterprises. *IOP Conf. Ser.: Mater. Sci. Eng.* **909**, 012047 (2020)
15. Ma, X.U., Tian, Y., Luo, C.H.U., Zhang, Y.: Predicting future visitors using big data. In: *International Proceedings of Conference on Machine Learning and Cybernetics*, pp. 269–274. IEEE (2018)