Computer and Information Science; Vol. 18, No. 1; 2025
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

Utilizing Machine Learning Models to Forecast Pricing on Mechanical Components and Automate Sourcing

Manas Kumar Singh[1], Shreya Kumari[1], Pranjal Singh[2]

[1] Erik Jonsson School of Engineering and Computer Science at The University of Texas at Dallas, Texas, USA

[2] G. Narayanamma Institute of Technology and Science, Hyderabad, India

Correspondence: Shreya Kumari, Erik Jonsson School of Engineering and Computer Science at The University of Texas at Dallas, Texas, USA. E-mail: shreyautd@gmail.com

Received: January 21, 2025 Accepted: Februay 13, 2025 Online Published: March 26, 2025

doi:10.5539/cis.v18n1p102 URL: https://doi.org/10.5539/cis.v18n1p102

Abstract

In supply chain management, price quotations on machine components can be crucial in determining accurate pricing that may suit a business's reliability and improve cost control. This process emphasizes price comparisons and negotiations to enable well-informed decisions that do not compromise quality or cost. With the emergence of machine learning and artificial intelligence, companies can leverage these tools to discern reasonable prices and establish the price quotations they can present to suppliers. This study evaluates machine learning models that employ previous material prices and properties such as part family, material grade, dimensions, thread type, and coating to determine the most optimal model. However, the prototype designed within this research expands further by adjusting the hyperparameters of the chosen model to enhance its efficacy. Ultimately, the model analysis determined that Gradient Boosting had a significant predictive accuracy and thus was the best-fit model to forecast pricing on mechanical components.

Keywords: gradient boosting, machine learning model, price quotation, supply chain management

1. Introduction

Supply chain management (SCM) is an integral part of any business function, and it has become increasingly prevalent within various industries such as healthcare, e-commerce, energy, and manufacturing. A seller composes an efficacious pricing strategy, crucial to the function of SCM, and thus regulates the demand for the product, maximizes company profits, and determines the business's credibility (McCosh & Morton, 1978). Implementing a productive dynamic pricing approach has the means to gain and augment this market competitiveness. Dynamic quotations foster measures through which a buyer orders a quote, and sellers provide the specified quote. Eventually, buyers accept or reject the quote modeled by the *acceptance probability function*, which determines the price point and lead time quote a buyer is expected to accept (Gel, 2020). This implies that sellers analyze the buyer's behavior and dynamically change the product prices depending on the stability of the channels, demand, and market standing to seek opportunities to optimize profitability consistently. Ultimately, this increases the need for buyers to develop price quotations that generate more significant benefits for themselves rather than heightening the interests of profit-driven sellers. This research aims to manage costs within the manufacturing enterprise effectively and focuses on the mechanical components used in the oil and gas industry. This algorithm utilizes past datasets to identify patterns in pricing dynamics and provide price quotations comparable to those offered by suppliers. By implementing this approach, buyers can retain complete control over price quotations and inevitably automate mechanical part sourcing within the mechanical industry.

While various publications present studies on price quotations that utilize predictive machine learning models, this work solely pertains to the marketing industry. Furthermore, this research builds upon existing literature by implementing a user-friendly interface and suggesting future implementation of optical character recognition (OCR) to automate mechanical part sourcing, which remains ingenious. Studies exist on the necessity of user-interface designs within the manufacturing industry, as they enhance reliability and productivity; however, they lack implementation (Eeva, 2015). Likewise, while publications on OCR have suggested implementing the algorithm on automotive parts, this was merely utilized to identify the component's labeled part numbers (Schlüter, 2025). Other studies have discussed the advantages of employing Artificial Intelligence and Machine Learning to optimize SCM and have provided the possible challenges that may arise with the implementation. However, these

studies are enhanced by implementing this research as it produces factual evidence to corroborate its efficiency and functionality.

The objective of this study is to determine what machine-learning model, whether linear regression, decision tree, random forest, or gradient boosting, could provide the most accurate representation of the correlation between the properties of the machine parts and price. Based on industrial sourcing knowledge, there is a prevalent correlation between part family, material grade, max diameter of parts, max length of the part, weight, thread type, and coating with price. The primary hypothesis was that the gradient-boosting machine-learning model would be attributed to the best-performing model for forecasting price. This is corroborated by the theory that presents how gradient boosting is a robust approach in terms of regression and classification (Friedman, 2001). The secondary hypothesis is that the best-performing machine-learning model has 99% accuracy by utilizing a simple cost model using the available data set. In the manufacturing industry, components are bought from different companies and regions worldwide, depending on the prices offered. Since the purpose of this research is to ensure these differences can be alleviated to provide buyers with substantial savings, cost models that compare energy costs, labor costs, and varying regional machining costs are necessary to enhance the purchasing power of companies and understand which region provides an effective means in doing so (Roshan, 2024).

The hypotheses correlate with the research design as it employs various machine-learning models and determines the accuracy for each based on the calculated Mean Squared and Root Mean Squared Errors regularly used within model evaluations for regression analysis (Chicco, 2021). Additionally, to create these specific machine-learning models, a cost model is implemented, which is relevant to integrating all data points to establish a reasonable correlation that is inputted into the model. Predictions are made to create improved price forecasting prototypes through the procedures within the research design.

On a theoretical front, this study contributes to the growing amount of literature regarding the application of machine learning in the field of forecasting industrial components. It highlights the usefulness of the gradient boosting algorithm and other well-known machine learning models in predicting the prices based on complex correlations, thus reinforcing the robustness of these problem-solving approaches in the regression problems domain. To further elaborate on this point, the research highlights the utilization of cost models, data cleaning, and preprocessing, thereby depicting the best practices of machine-learning applications. Practically, the study significantly impacts the manufacturing industry by having an accurate forecasting model that will allow companies to create an improved approach to procurement strategies. Moreover, this model has the potential to be automated using optical character recognition for dataset populations to enhance operational excellence compared to manual data entry.

Additionally, Streamlit, as a front-end interface, ensures that industrial users can easily access and interact with the models, making it highly accessible to all companies within the manufacturing sector. Streamlit, a basic web application, simplifies the use of these forecasting models by eliminating a barrier that often deters users - complexity (Kannan, 2024). This application encourages companies to make informed decisions using a more user-friendly platform that combines the intricacies of machine learning with the real-world usage of front-end development.

2. Method

The methodology emphasizes using various datasets to forecast pricing on machine components and ultimately automate sourcing to represent the diverse characteristics of mechanical components crucial to the oil and gas industry. These datasets are obtained from sources such as publications on supply chain, manufacturing, and primary anecdotal interviews with industry professionals in the manufacturing and sourcing sector.

The mechanical component features highlighted in this research are material grade - indicating the type of steel utilized (e.g., stainless steel, alloy), the diameter, the length, the weight, the thread type, and the coating type. The target variable in this machine learning research is the price, which will be forecasted by assessing various models. This study assumes that part family stays consistent throughout the analysis (e.g., valves, connectors, bolts).

2.1 Machine Learning Procedure and Models Utilized

The dataset employs one-hot encoding for the various categorical data types. Approximately 170 data points are used to conduct this research to encompass all the combinations of mechanical component characteristics to avoid model bias. Additionally, to account for noise reduction, which is usually the case when solving real-world regression problems, the dataset focuses on prices in the [0,1000] range (Gupta 2019). The data for training this model is divided into 80% reserved for the training set and 20% for the testing set to maintain standard machine learning practices (Sivakumar, 2024). Visual Studio code is utilized with the Streamlit library and the ipynb python

server rendition to run the model. The ML models are trained using the training data and ultimately applied to present feasible prices based on the test set. Subsequently, Mean Squared Error, Root Mean Squares Error, R^2 and Mean Absolute Percentage Error were used to determine the best-fit model. Moreover, plots are developed for every machine learning model representing actual vs. predicted values to provide additional visuals on the accuracy of the respective models.

This study evaluates four machine learning models: Linear Regression, Decision Tree, Random Forest, and Gradient Boosting. The most effective model is selected based on RSME, R^2 and MAPE values. The pseudocode for each machine learning model follows a structured pattern that integrates data preparation, model training, evaluation, and utilization of pickle - a module responsible for serializing Python objects to save the model that will be integrated with the front-end interface developed by Streamlit as shown in Figure 1. The user can manually input the diameter and length needed to calculate the weight as the density for this preset data is set at $0.268\ lb/in^3$ with the respective material grade and the specific add-ons such as threading type and coating type. Upon clicking predict price, the output reads the price generated by the model in the back-end.

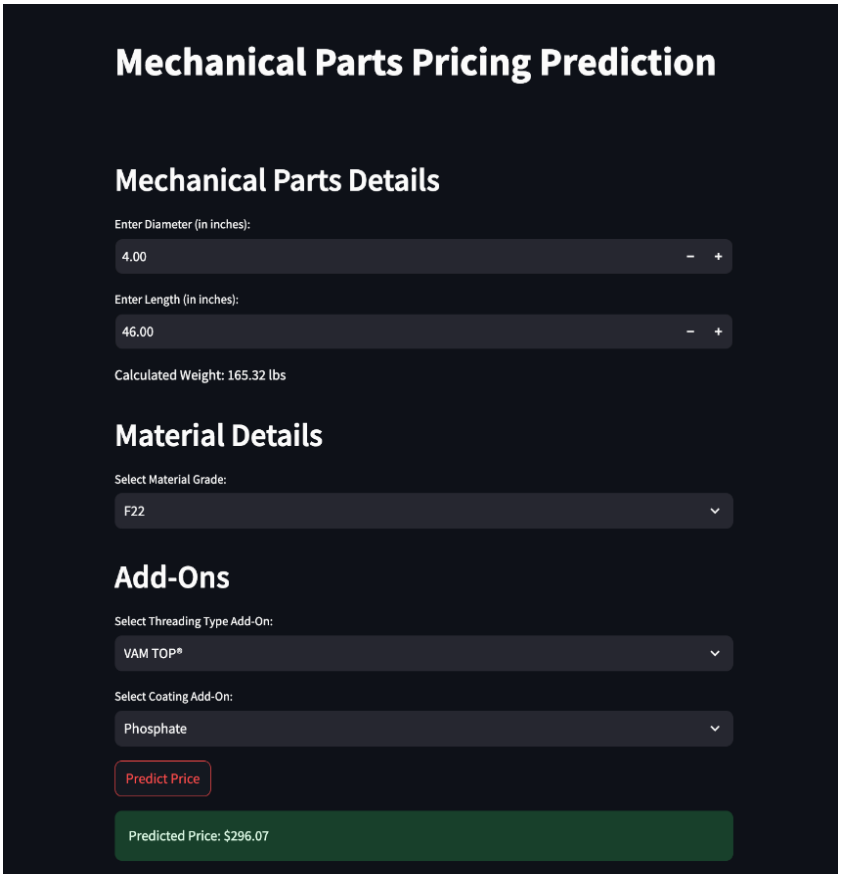

Figure 1. Front-End for Price Prediction Application

The initial procedures for all models begin with importing libraries for data manipulation through pandas, numerical operations using NumPy, plotting through Matplotlib, and various machine learning libraries utilized in the study. Such libraries, specifically NumPy, provide diverse functions that allow for the manipulation and transformation of data necessary to train the machine learning models (Rayhan, 2023).

A prepare_data function processes the target variable price, changing it from a string to a numerical float by removing '$' and '.', then selecting the features X and the target variable of Y. The code flow then goes through the function train_evaluate_model, which takes care of splitting the data into training and testing, then the set is run on the respective model for that algorithm. Thereafter, using the prediction of the test set, several model evaluation metrics are printed out, the metrics being RMSE and MAPE. The Root Mean Squared Error - represented by equation 1, augments the error between the actual and predicted value by enhancing the effects of the outlier; therefore, the Mean Absolute Percentage Error is employed to provide a more accurate representation of the error irrespective of the scale of the data.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (1)$$

After assessing the models using RMSE and MAPE, the actual vs. predicted values are plotted in a graph to complete the function. Lastly, the main function initiates the workflow, which is done by reading data from the CSV file and preparing the data through the prepare_data function. Then, training and evaluation are done utilizing the train_evaluate_model function. To end the workflow, the Python object is saved to a file using the serialization and deserialization properties of the pickle library. This ensures that the design of the code is modular, which can lead to reusable implementation of all models.

$$\text{MAPE} = \frac{\Sigma \frac{|\text{A-F}|}{\text{A}} \times 100}{\text{N}} \quad (2)$$

2.2 Overview of the Machine Learning Models

The basic strategy used within this research was a linear regression model that assumed a linear relationship between derived features and price and therefore derived a linear function to predict price. Decision tree algorithms partition data by hierarchical decision rules to capture non-linear associations and allow complex feature interactions to be modeled. Random forest as a category of ensemble methods utilizes several decision trees to combine their predictions to produce lower variance and more stable overall performance. Gradient boosting uses decision trees sequentially such that each stage is designed to correct errors in previous attempts to make it more accurate in prediction. Through using conventional regression measures to compare model performance, this research aims to determine the best algorithm to predict mechanical part prices.

2.3 Sample of Dataset

Part Family	***Material***	***Threading Type***	***Coating Type***	***Weight (lbs)***	***Length (in)***	***Diameter (in)***	***RM Cost/Lb***	***Price***
X	*F22*	*TenarisHydril Blue®*	*Xylan Coating*	*165.23936*	*46*	*4*	*0.8301*	***$281.78***
X	*4130*	*JFE Lion*	*Cu Plating*	*416.69056*	*29*	*8*	*0.6214*	***$499.19***
X	*F22*	*JFE Lion*	*Ni Plating*	*341.03069*	*31*	*7*	*0.8301*	***$561.77***
X	*4130*	*JFE Lion*	*Carbide Coating*	*758.8438*	*20*	*13*	*0.6214*	***$980.38***
X	*F22*	*TenarisHydril Blue®*	*Xylan Coating*	*517.04653*	*47*	*7*	*0.8301*	***$881.70***

3. Results

Figure 2. Actual vs Predicted Prices Plot for the Linear Regression Model

The Linear Regression Model, as represented by Figure 2, provided performance metrics of $51.25 for the RMSE, 0.9771 for the R^2 value, and 30.3304% for MAPE. This provided a reasonable assessment of the fit of the Linear Regression Model by portraying how, compared to the other machine learning models, this was determined to have relatively lower RMSE and MAPE scores; additionally, it had one of the highest R^2 values.

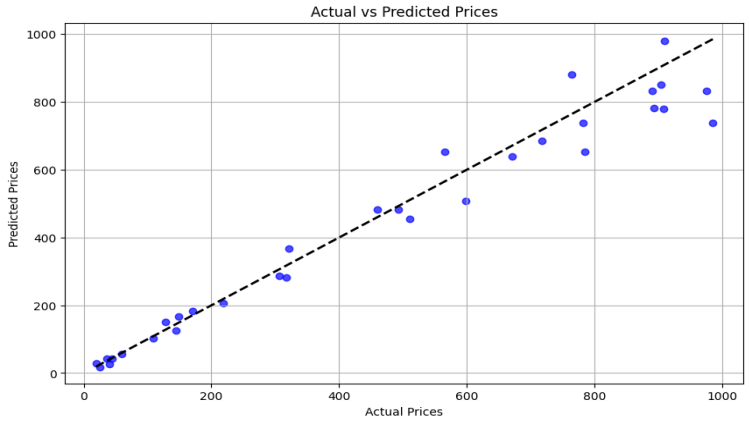

Figure 3. Actual vs Predicted Prices Plot for the Decision Tree Model

As represented by Figure 3, the Decision Tree Model provided performance metrics of $75.48 for the RMSE, 0.9503 for the R^2 value, and 12.9540% for MAPE. This provided a logical analysis of the fit of the Decision Tree Model and consequently portrayed how the model had the highest RMSE value.

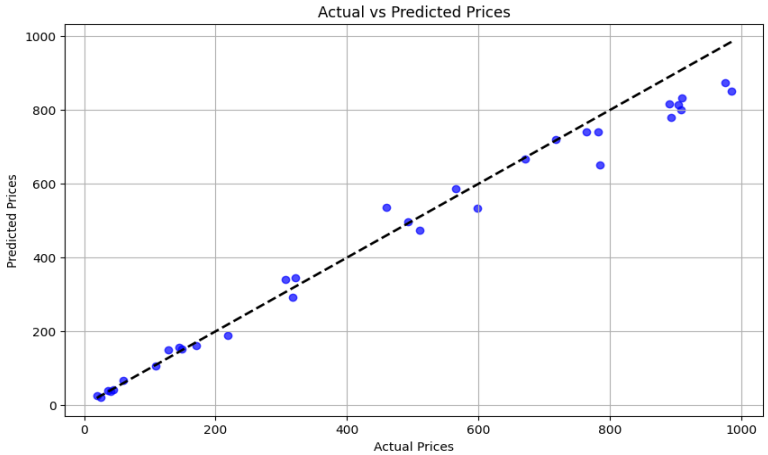

Figure 4. Actual vs. Predicted Prices Plot for the Random Forest Model

As represented by Figure 4, the Random Forest Model provided performance metrics of $58.11 for the RMSE, 0.9705 for the R^2 value, and 9.5404% for MAPE. As with the previous models, this analysis of the Random Forest Model provided insight into the model's fit and depicted it as the model with the lowest MAPE.

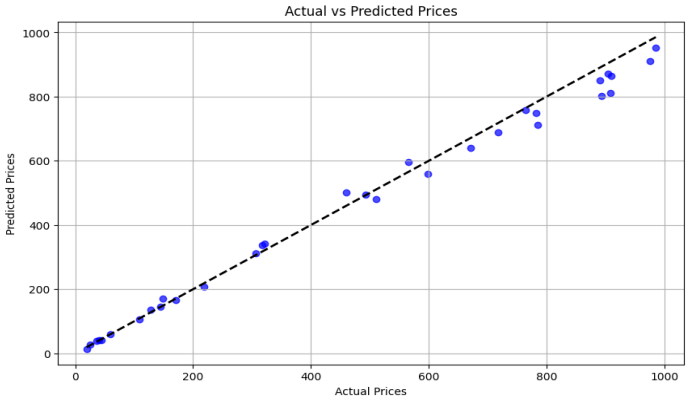

Figure 5. Actual vs Predicted Prices Plot for the Gradient Boosting Model

The Gradient Boosting Model, as represented by Figure 5, provided performance metrics of $36.56 for the RMSE, 0.9883 for the R^2 value, and 6.1585% for MAPE. This study was crucial to the research as it attributed Gradient Boosting as the most effective model by presenting the lowest RMSE and MAPE scores and having the highest R^2 value. With minimal error and greater reliability, the model had significant predictive accuracy, which is necessary for forecasting pricing on mechanical components.

4. Discussion of Results

The study's results support the research's primary hypothesis that the gradient-boosting machine-learning model provides the most accurate prediction for forecasting the prices of mechanical components. The model has the lowest RMSE of $36.56, the lowest MAPE of 6.16%, and the highest R-squared value of 0.9883, which shows that the model best fits the dataset. These metrics validate the idea that the model is ideal for handling regression problems with complex linear and non-linear relationships. Ultimately, this confirms prior research by Friedman (2001) on the potency of the gradient-boosting algorithm across similar application domains. The secondary hypothesis asserts that the model with optimal performance would attain an accuracy close to 99% received partial corroboration. Although the gradient-boosting model demonstrated significant predictive accuracy, limitations associated with the dataset and the intrinsic variability in the pricing of mechanical components hindered the model from achieving the expected benchmark.

Figure 6. Actual R-squared, MAPE, and RMSE Values of the Machine Learning Models

Machine Learning Model	R-squared value	MAPE value	RMSE value
Linear Regression Model	0.9771	30.3304 %	$51.25
Gradient Boosting Model	0.9503	12.954 %	$75.48
Random Forest Model	0.9705	9.5404 %	$58.11
Decision Tree Model	0.9883	6.1585 %	$36.56

The results of this study contribute to the inexorable growth of literature on machine learning applications within supply chain management. Utilizing predictive models for automated sourcing, this research helps link machine learning advancements with its practical application. Moreover, it emphasizes the importance of feature selection and model evaluation techniques, which are vital for creating a forecasting application (Pudjihartono, 2022). Additionally, eliminating manual labor allocated to forecasting prices of mechanical components leads to accurate prices that enable cost control integrated with operational excellence. Implementing a user-friendly interface using Streamlit ensures that despite prerequisite knowledge of machine learning, the application can still be accessible to all industry professionals. This aligns with the industry's primary focus on leveraging digital tools to automate procurement strategies and streamline supply chain operations (Olorunyomi, 2024).

Beyond the convenient user interface, this study enables numerous facets for future development. By discerning a model with high predictive accuracy and the ability to mitigate error, there is an opportunity to integrate optical character recognition (OCR) to automate sourcing. There are numerous engineering drawings to be utilized in sourcing; OCR can extract length, diameter, and other features from these drawings, which, coupled with the Gradient Boosting Model, can predict the pricing of these mechanical components without requiring manual

input. Figure 8 shows this process is initiated with the need for price quotations. Subsequently, drawings of mechanical components are inspected through OCR, and their details are captured as datasets. These datasets are then presented to the cloud, allowing suppliers to access these details and prepare quotes. The supplier quotes would then be compared to the price quotes provided by the Gradient Boosting model, using an ML algorithm, as the supplier quote is retrieved from the cloud. Depending on whether the supplier quote exceeds the price quote from the machine learning model, the business would either be awarded, or the supplier quote would have to be reevaluated.

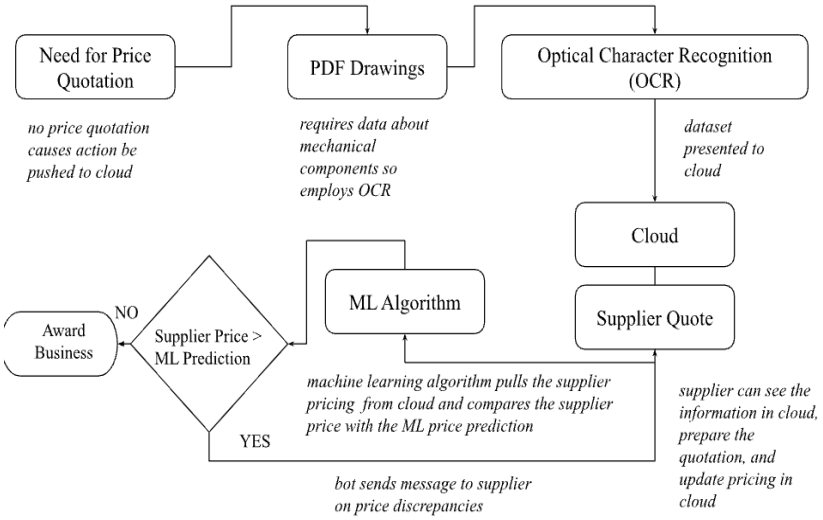

Figure 7. Use of ML Algorithms and OCR to Automate Sourcing

Nevertheless, this study has several limitations that need to be analyzed. Firstly, the study only deals with 170 data points, which is considered relatively diverse to make an adequate model. Still, it does not capture the varied nature of real-world complex sourcing situations. Additionally, the prices the study is conducted on are in a specific range, which cannot adequately consider the circumstances of real-world sourcing concerns. A larger, more reflective dataset of all these scenarios could better the model's performance and generalizability. This study also assumes the consistency of part families, which may not reflect an accurate picture of what industry professionals work with daily. Future studies could explore integrating additional features to the model, such as market trends and regional cost variations, to enhance the usability for more complex problems. Investigating these using more convoluted machine learning models, such as neural networks or other ensemble models, will improve the model's capability.

5. Conclusion

This research proves that the gradient boosting model can revolutionize price forecasting by automating sourcing within the manufacturing sector. Furthermore, it expands upon existing literature by providing a user-friendly interface that allows all industry professionals to access this model to ensure it is accessible to those who may not have prior machine learning knowledge. This not only addresses issues with accessibility but also eradicates any possibilities of human error when buyers produce price quotes to be compared with the supplier's quotes. While limitations exist, these findings emphasize a significant potential for cost saving, operational excellence, and scalability. These difficulties foster research that may prove the suggested methodology as the cornerstone of data-driven industrial supply chain management decisions.

However, despite solving various challenges that may arise with price forecasting, there is still potential for growth. This study can be further expanded with optical character recognition to alleviate the dependence on manual input of the features of the mechanical components. Inevitably, it will optimize sourcing and make supply chain management more efficient, a methodology that can be extended to other industries such as e-commerce or healthcare.

Acknowledgments

Not applicable.

Authors' contributions

Authors contributed equally to the study.

Funding

Not applicable.

Competing interests

Not applicable.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Canadian Center of Science and Education.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

References

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ. Computer science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

Eeva Järvenpää, & Lanz, M. (2015). Guidelines for Designing Human-Friendly User Interfaces for Factory Floor Manufacturing Operators. *IFIP Advances in Information and Communication Technology*, 531-538. https://doi.org/10.1007/978-3-319-22759-7_61

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189-1232. Retrieved from https://www.jstor.org/stable/2699986

Gel, E.S., Keskinocak, P., Yilmaz, T. (2020). Dynamic Price and Lead Time Quotation Strategies to Match Demand and Supply in Make-to-Order Manufacturing Environments. In: Smith, A. (Eds.), *Women in Industrial and Systems Engineering. Women in Engineering and Science.* Springer, Cham. https://doi.org/10.1007/978-3-030-11866-2_23

Gupta, S., & Gupta, A. (2019). Dealing with noise problem in machine learning data-SETS: A systematic review. *Procedia Computer Science*, *161*, 466-474. https://doi.org/10.1016/j.procs.2019.11.146

Kannan, M. K. Jayanthi., Arpit Sengar, Bhardwaj, A., Singh, A., & Shrivastava, V. (2024). 'Simplifying Machine Learning: A Streamlit Powered Interface for Rapid Model Development with PyCaret". *International Journal of Innovative Research in Computer & Communication Engineering*, *12*(5), 5857-5871. https://doi.org/10.15680/IJIRCCE.2024.1205122

McCosh, A. M., & Morton, M. S. S. (1978). *The Pricing Decision.* In Management Decision Support Systems. Palgrave Macmillan, London. https://doi.org/10.1007/978-1-349-02764-4_8

Olorunyomi, S. J., Adedoyin, T. O., Olusegun, G. O., & Oluwatobi, T. S. (2024). Leveraging Artificial Intelligence for Enhanced Supply Chain Optimization: A comprehensive review of current practices and future potentials. *International Journal of Management & Entrepreneurship Research*, *6*(3), 707-721. https://doi.org/10.51594/ijmer.v6i3.882

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, *2*. https://doi.org/10.3389/fbinf.2022.927312

Rajan, R., Abu, R., & Kinzler, R. (2023, August 19). *Exploring the Power of Data Manipulation and Analysis: A Comprehensive Study of NumPy, SciPy, and Pandas*. https://doi.org/10.13140/RG.2.2.22390.16968

Roshan, V. (2024). Enhancing Operational Efficiency and Cash Flow through Supply Chain Optimization in the Oil and Gas Sector. *International Journal of Business and Management*, *19*(3), 91-91. https://doi.org/10.5539/ijbm.v19n3p91tv

Schlüter, M., Tepper, C., Briese, C., Kroeger, O., Vicente-Garcia, R., Krüger, J. (2025). Deep Learning-Based Optical Character Recognition for Identifying On-Label Printed Part Numbers of Used Automotive Parts: A Comparative Study of Open Source and Commercial Methods. In Kohl, H., Seliger, G., Dietrich, F., Mur, S. (Eds.), *Sustainable Manufacturing as a Driver for Growth. GCSM 2023. Lecture Notes in Mechanical Engineering*. Springer, Cham. https://doi.org/10.1007/978-3-031-77429-4_58

Sivakumar, M., Parthasarathy, S., & Padmapriya, T. (2024). Trade-off between training and testing ratio in machine learning for medical image processing. *PeerJ. Computer science*, *10*, e2245. https://doi.org/10.7717/peerj-cs.2245

Appendix A: GitHub Repository

All the machine learning models relevant to this research are available on GitHub.

The repository can be accessed through this link:

GitHub Repository

Appendix B: Relevant Datasets Utilized

The datasets utilized for this research are linked here:

Mechanical Part Dataset