JIEET (Journal of Information Engineering and Educational Technology)

Volume 9, Number 2, 2025 pp. 80-88
E-ISSN : 2549-869X
Open Access: https://journal.unesa.ac.id/index.php/jieet

Predicting Software Sales Performance Using Support Vector Regression (SVR) and Linear Regression Algorithms

Muhammad Athallah Rafi[1], Alvin Adam Anton Suryadarma[2], Hazbie Alfarhizi Syahwadana[3], Aji Setiawan[4]
[1,2,3,4] Teknologi Informasi Fakultas Teknik, Universitas Darma Persada, Jakarta, Indonesia

Corresponding author: atallahrafi93@gmail.com

ARTICLE INFO

Article history:
Submitted 10-11-2025
Accepted 5-12-2025
Available online 6-12-2025

Keywords:
Amazon Software Sales, Global Sales Dataset, Model Evaluation, Linear Regression, Support Vector Regression (SVR)

DOI:
https://doi.org/10.26740/jieet.v9n2.p80-88

ABSTRACT

Software has become an essential part of everyday life, both in the workplace and in education. Various applications, such as Microsoft Office and Google Workspace, are widely used to enhance productivity. As public demand for digital solutions continues to rise, software distribution through global platforms such as Amazon has also grown significantly. However, not all software products achieve high sales figures due to a lack of effective strategies for understanding consumer behaviour and market demand. Therefore, accurate sales prediction plays a crucial role in supporting successful software marketing strategies.
This study aims to predict the best-selling software on Amazon by applying two algorithms: Linear Regression and Support Vector Regression (SVR). Before implementing these algorithms, several stages were conducted, including identifying the research object, preprocessing the data—reducing the original dataset of 2,424 rows to 1,338—followed by splitting the dataset into 80% training, 10% validation, and 10% testing sets. The final stage involved developing and comparing prediction models using both Linear Regression and SVR. The results of this study are expected to contribute to determining the most suitable algorithm for predicting software sales and to serve as a reference for future research in this field.

INTRODUCTION

The daily life of society today cannot be separated from software. Whether in the workplace or in education, software plays an essential role in improving productivity and efficiency. Applications such as Microsoft Office (Word, Excel, PowerPoint) or Google Workspace (Docs, Sheets, Slides) help users create documents, analyse data, and deliver presentations more efficiently. As society's demand for efficient digital solutions continues to increase, Amazon, as one of the largest global marketplaces, plays a significant role in the distribution and sale of software worldwide. Many software products achieve high sales figures, while others struggle to compete due to ineffective strategies for understanding market needs and consumer behaviour.

Therefore, accurately predicting software sales is crucial for software marketing and business decision-making.

A study by N. Widya Utami and J. Juinor Soplantila analysed 210 book data on Amazon using the Support Vector Machine (SVM) method. Using the Orange data mining application, the study identified two best-selling book types — Fiction and Non-Fiction, each having its own top-selling titles (Ismail, M. et al. 2025)

Research by N. N. F. Adzani et al. examined video game sales using a public dataset from Kaggle.com. The dataset was based on user reviews from online game platforms. The study divided the dataset into 80% for training and 20% for testing. The results showed a Recall of 99.3%, a Precision of 99.7%, and an Accuracy of 99.4%, indicating a high level of model performance in predicting game sales (Pavlyshenko, B. M., 2019).

Another study by Y. Syakir et al. analysed the sales prediction process for the Ariqa Collection boutique on the Shopee marketplace using the Linear Regression algorithm. The researchers collected sales data from May 2020 to April 2022 through Shopee's Seller Centre. Of the 730 available data rows, 574 were used, and 156 were removed. Using RapidMiner Studio software for computation, the study achieved MSE = 5,172,628,212,404, RMSE = 2,274,341.27, and MAPE = 4.34% (Xu, Q. et al. 2019).

In another study, A. Setiawan and R. Mulyanti aimed to assist Toko Busana Muslim Trendy in developing its online business competitiveness. Using the Apriori algorithm, they analysed customers' purchasing patterns from six months of transaction data, filtered by support and confidence values. The study achieved a 75% confidence level, indicating that customers who purchased Anindya Syar'i were most likely to buy Gloria Syar'i next. The results provided product recommendations to customers and insights for administrators to analyse buying trends (Cheriyan, S. et al. 2018).

Research by R. Ishak focused on determining relevant attributes from a student graduation dataset. Initially, 13 attributes (gender, class, age, SKS1–SKS5, IPS1–IPS5, status) were included; only nine were deemed relevant. This study extended previous research by E. P. Rohmawan, titled "Predicting Student Graduation Timeliness Using Decision Tree and Artificial Neural Network Methods" (Lu, C. J., et al. 2014).

Meanwhile, Z. Rani and B. K. Khotimah conducted research analysing public opinions on the traditional Karapan Sapi (bull racing) in Madura. The study aimed to assess public sentiment, evaluate the combined effectiveness of the K-Means and Support Vector Machine (SVM) methods, and apply the Synthetic Minority Oversampling Technique (SMOTE) to address dataset imbalance. A total of 647 Indonesian-language tweets were collected via web crawling. Using GridSearchCV, the study determined the optimal SVM configuration with a linear kernel, C = 1.0, and gamma = 1.0, achieving 92% accuracy. However, the model showed limitations in handling independent variables, as reflected in low recall scores for some variables (Gumus, M., & Kiran, M. S., 2017).

Based on the studies above, this research aims to determine which algorithm—Support Vector Regression (SVR) or Linear Regression—is more suitable for predicting software sales on Amazon. It is expected that this study will provide insight into the performance comparison of

both algorithms and serve as a reference for future researchers to enhance predictive modelling in software sales forecasting.

LITERATURE REVIEW

1. Support Vector Regression

A machine learning algorithm used for both classification and regression tasks is known as the Support Vector Machine (SVM). Support Vector Regression (SVR) is a regression method derived from SVMs, well-suited for datasets prone to overfitting and regression problems.

$$X\ 2, \ldots\ldots, Xn+\leq Rn \quad (1)$$

$$f(x) = wT\varphi(x) + b \quad (2)$$

Equation (1) represents the training dataset used for the implementation of the Support Vector Machine algorithm.

In Equation (2), a residual term is included to minimise the output scale y relative to the resistance $f(x)$.

2. Regresi Linier

Linear Regression describes the functional relationship between a dependent variable and one or more independent variables. The least-squares method is used to find the best-fitting regression line for the observed sample data. When a dataset contains two or more variables, it is essential to understand how they relate to and influence one another (Tarta, E. N., et al., 2021).

$$Y = a + bX \quad (3)$$

In Equation (3), YYY is the dependent variable, XXX is the independent variable, aaa is the intercept (the point where the line crosses the Y-axis), and bbb is the regression coefficient or slope.

3. Python

Python is a programming language first introduced in 1991 by Guido van Rossum and has since evolved under the supervision of the Python Software Foundation. Its uniqueness lies in being a versatile, interpreted programming language that is highly readable and easy to understand (Schneider, P., & Gupta, A., 2016).

4. Scikit-Learn

Scikit-Learn is a Python library designed to simplify the implementation of machine learning algorithms, featuring a consistent, user-friendly API. It is widely used in both industry and academia. In this research, Scikit-Learn is particularly relevant, as it provides a range of predictive algorithms, such as Linear Regression, Random Forests, and Support Vector Regression, that can be applied to historical sales data (Wang G et al., 2022).

5. Correlation

Correlation analysis is a set of techniques used to measure the relationship between two variables.

$$\frac{r_{xy} = N\ \sum xy - (\sum x)\ (\sum y)}{\sqrt{[N\ \sum x^2 - (\sum x)^2]\ [N\ \sum y^2 - (\sum y)^2]}} \quad (4)$$

Equation (4) represents the formula for determining the strength of the relationship between two variables, where variable xxx is the independent variable and variable yyy is the dependent variable (Vapnik, V. 1995).

METHOD

This study employs a quantitative, comparative approach to evaluate the effectiveness of two predictive algorithms, Support Vector Regression (SVR) and Linear Regression, for forecasting software sales on the Amazon e-commerce platform (García, S., Luengo, J., & Herrera, F., 2016).

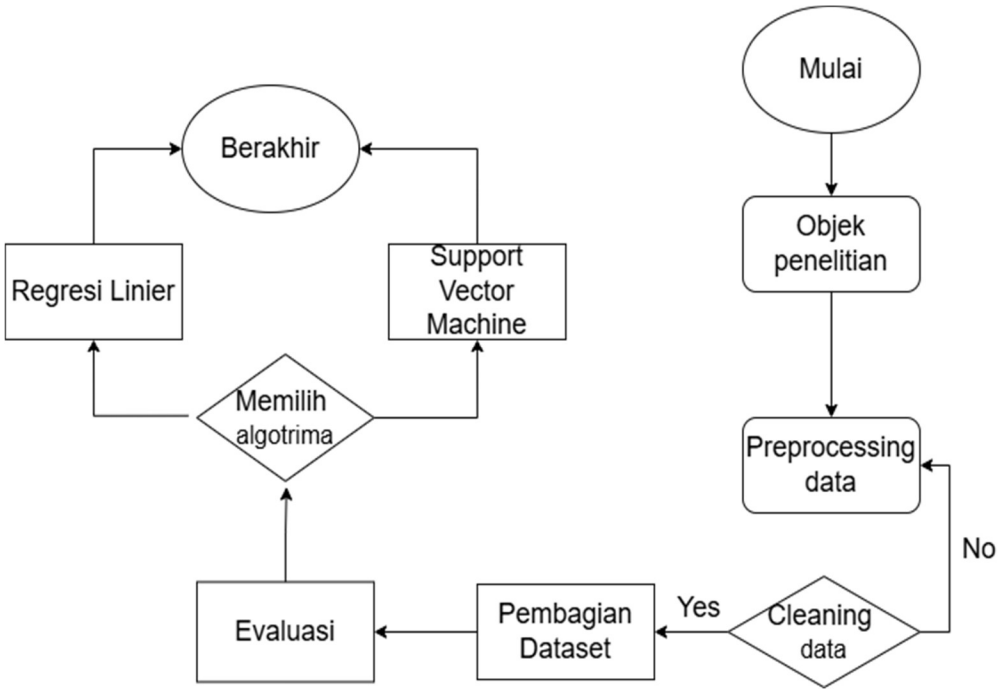

Figure 1. Research Flow Diagram

1. Research Object

Figure 1 illustrates the flow of this research. The research object is a secondary historical dataset of software sales available on the Kaggle platform: https://www.kaggle.com/datasets/kaverappa/amazon-best-seller-softwares/data. The dataset focuses on products sold on Amazon, one of the largest global marketplaces. This dataset was chosen for its relevance to the research objective of predicting sales volume based on essential attributes, including product category, user rating, price, number of sales, and rank.

Table 1. Secondary Historical Data of Software Sales

No	Product_title	product_price	product_star_rating	product_num_ratings	Rank	country
1	TurboTax Deluxe 2024 Tax Software, Federal & State Tax Return [PC/MAC Download]	$55.99	4.2	6511	1	US
2	TurboTax Premier 2024 Tax Software, Federal & State Tax Return [PC/MAC Download]	$82.99	4.1	2738	2	US
3	TurboTax Home & Business 2024 Tax Software, Federal & State Tax Return [PC/MAC Download]	$95.99	4.2	1672	3	US

2. Research Object

At this stage, data preprocessing was performed to remove invalid, irrelevant, and redundant values. This process includes eliminating duplicate, inconsistent, and null (empty) data. Null values are invalid entries with no value, which can cause errors during modelling. For instance, in the *product_price* attribute, about 11% of entries were missing, requiring appropriate handling. Additionally, the "$" symbol in the price column was removed to allow numeric conversion. The original dataset contained 2,424 rows, and after cleaning, it was reduced to 1,211 rows (Kotsiantis, S.B., Kanellopoulos, D., & Pintelas, P. E., 2006).

3. Dataset Splitting

After preprocessing, the dataset was divided into two main parts: training data and testing data. In this research, the train-test split method was applied, with 80% for training, 10% for validation, and 10% for testing (Chicco, D., Warrens, M. J., & Jurman, G., 2021).

4. Evaluation

The evaluation phase involved conducting correlation analysis across the dataset to classify and understand the relationships between variables before applying the two supervised machine learning algorithms, Support Vector Regression (SVR) and Linear Regression.

The purpose of this stage is to compare the effectiveness of both methods in predicting the number of software sales based on attributes such as product price, star rating, and number of user ratings (Botchkarev, A. 2019).

RESULTS

Microsoft Excel was used for data cleaning, while Python was used to implement the Linear Regression and Support Vector Regression (SVR) algorithms. Using these tools, the research process began with dataset cleaning and continued through the application of both regression algorithms.

1. Research Object

The data used in this study originates from a public dataset available on the Kaggle platform. The dataset initially contained 2,424 rows, which were reduced to 1,211 after data cleaning.

2. Data Preprocessing

Before performing Linear Regression and Support Vector Regression modelling, data preprocessing was carried out. Table 2 presents the data columns used in the modelling process, including product_price, product_star_rating, product_num_ratings, and rank.

Table 2. Data Columns

product_price	product_star_rating	product_num_ratings	Rank
143.99	4	389	4
47.99	4.1	1416	6
140.99	3.8	470	7
41.97	4	312	8

3. Correlation Heatmap

The correlation heatmap shows the relationships among the variables in the dataset.

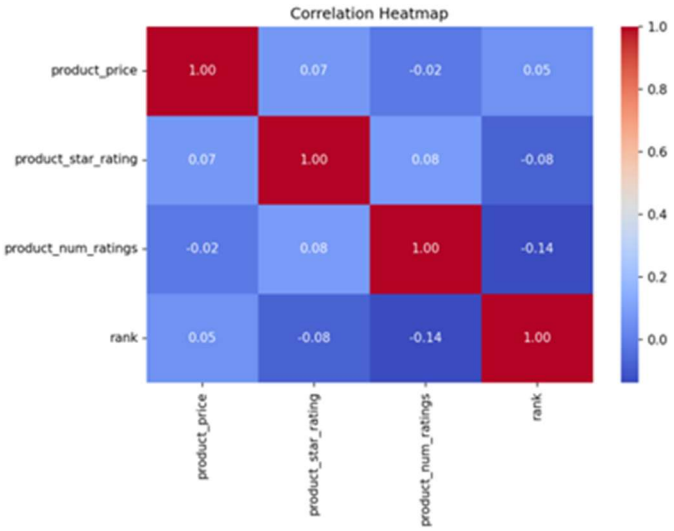

Figure 2. Correlation Heatmap

Figure 2 shows that the correlation coefficient between the independent and dependent variables is 1.0, indicating that the dataset's columns are strongly interrelated (B. Moharana et al., 2023).

4. Linear Regression

Based on the correlation results shown in Figure 2, the Linear Regression algorithm was implemented. The dataset was divided into 80% for training, 10% for validation, and 10% for testing. The Rank column was used as the dependent variable, while product_price, product_star_rating, and product_num_ratings served as independent variables.

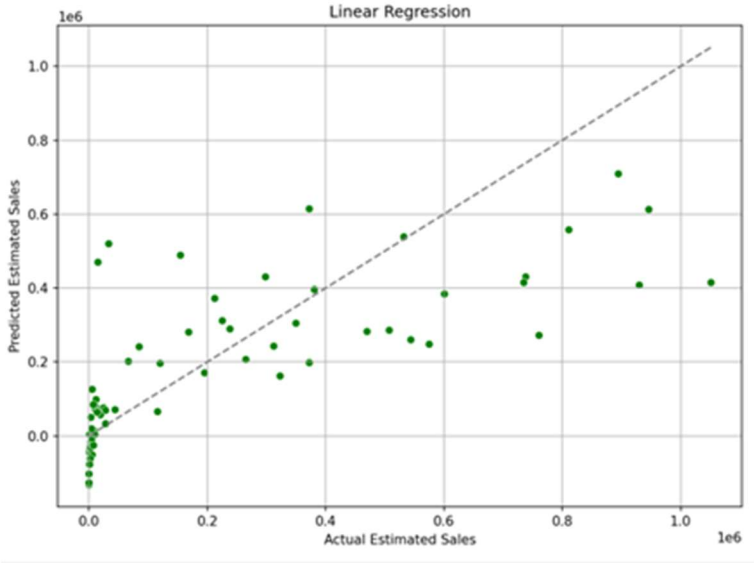

Figure 3. Linear Regression Results

Figure 3 presents the predicted sales rankings of best-selling products based on the Rank column. The Linear Regression model achieved the following results:

Table 3. Linear Regression Results

Validation	Test
MSE = 33172188341.96	MSE = 38720403274.18
R^2 = 0.48	R^2 = 0.55
Accuracy = 48.38%	Accuracy = 54.93%

5. Support Vector Regression

Based on the correlation analysis in Figure 2, the Support Vector Regression (SVR) algorithm was applied using the same data partitioning: 80% for training, 10% for validation, and 10% for testing. The Rank column served as the dependent variable, while product price, product_star_rating, and product_num_ratings were independent variables.

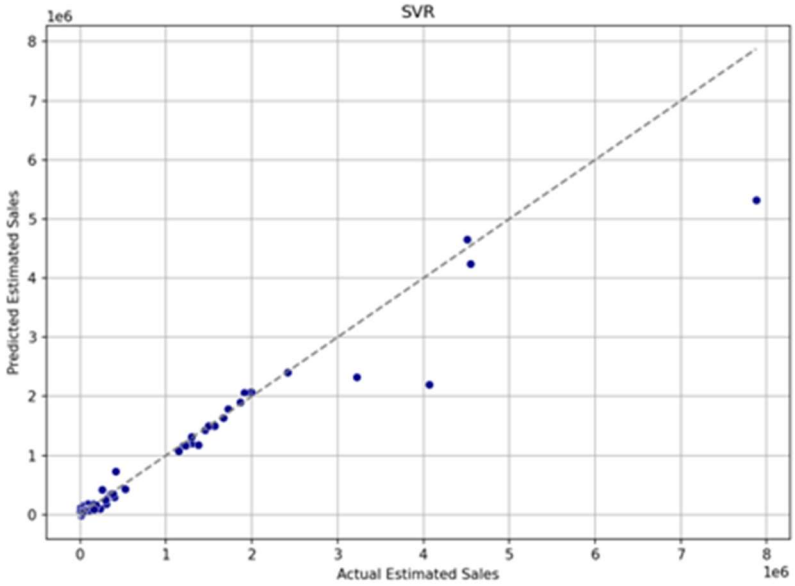

Figure 4. Support Vector Regression Results

Figure 4 shows the predicted rankings of best-selling software products based on the Rank column. The SVR model achieved the following results:

Table 4. Support Vector Regression Results

Validation	Test
MSE = 52,194,167,890.25	MSE = 151,971,280,828.22
R^2 = 0.97	R^2 = 0.91
Accuracy = 97.10%	Accuracy = 91.10%

DISCUSSION

The comparison between the two algorithms, Linear Regression and Support Vector Regression (SVR), shows a clear performance difference when applied to the Amazon software sales dataset. Linear Regression achieved a validation accuracy of 48.38% with an R^2 of 0.48 and a test accuracy of 54.93% with an R^2 of 0.55, indicating it could explain only about half of the data variance. This suggests that Linear Regression struggled to capture the nonlinear relationships among variables such as product price, user ratings, and sales rank. In contrast, Support Vector Regression achieved validation accuracy of 97.10% and test accuracy of 91.10%, with R^2 values of 0.97 and 0.91, respectively, demonstrating a much stronger predictive capability. The kernel-based approach of SVR enabled it to model complex, nonlinear relationships more effectively, avoiding overfitting while maintaining generalisation. Based on these findings, SVR is considered more suitable for predicting software sales on e-commerce platforms like Amazon, as it offers higher accuracy and robustness than Linear Regression.

CONCLUSION

This study concludes that, based on the observed case of predicting software sales performance, the Support Vector Regression (SVR) algorithm performs better and is more suitable than Linear Regression. For future research, it is recommended to develop both algorithms further or explore other machine learning approaches to determine the most effective model for predicting the best-selling software on the Amazon platform.

REFERENCES

B. Moharana, B. B. Biswal, S. Dey, M. K. Rath and S. Banerjee, "Play Store App Analysis & Rating Prediction Using Classical ML Models & Artificial Neural Network," 2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2023, pp. 1-5, doi: 10.1109/ICCUBEA58933.2023.10391960.

Botchkarev, A. (2019). "A new typology design of performance metrics to measure errors in machine learning regression algorithms." *Interdisciplinary Journal of Information, Knowledge, and Management*, 14, 45-76. https://doi.org/10.28945/4184.

Cheriyan, S., et al. (2018). "Intelligent Sales Prediction Using Machine Learning Techniques." 2018 International Conference on Computing, Power and Communication Technologies (GUCON). IEEE.DOI: 10.1109/iCCECOME.2018.8659115

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

García, S., Luengo, J., & Herrera, F. (2016). "Data preprocessing in data mining." Springer International Publishing, 1-3.

Gumus, M., & Kiran, M. S. (2017). "Crude oil price forecasting using XGBoost, support vector regression and artificial neural networks." International Journal of Energy Economics and Policy, 7(6), 46-55. https://doi.org/10.1109/UBMK.2017.8093500

Ismail, Mustapha & Tukur, Hafsat & Friday, Mamudu. (2025). Sales Prediction using Ensemble Machine Learning Model. International Journal of Scientific Research and Modern Technology. 4. 24-35. 10.38124/ijsrmt.v4i3.350.

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). "Data preprocessing for supervised learning." International Journal of Computer Science, 1(2), 111-117.

Lu, C. J., et al. (2014). "Sales forecasting for computer products based on a variable selection scheme and support vector regression." Neurocomputing, 128, 491-499. https://doi.org/10.1016/j.neucom.2013.08.012

Pavlyshenko, B. M. (2019). "Machine-learning models for sales time series forecasting." Data, 4(1), 15. https://doi.org/10.3390/data4010015

Schneider, P., & Gupta, A. (2016). "Forecasting sales of new and existing products using consumer reviews: A machine learning approach." Information Systems Frontiers, 18, 247-263. https://doi.org/10.1016/j.ijforecast.2015.08.005

Tarta, E. N., et al. (2021). "Comparison of Linear Regression and Random Forest Algorithms for Sales Forecasting." 2021 IEEE International Conference on Automation/XXIV Congress of the Chilean Association of Automatic Control (ICA-ACCA).

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer, New York.

Xu, Q., et al. (2019). "Product adoption and sales forecast for e-commerce: A review." Electronic Commerce Research and Applications, 36, 100869.

Wang G. (2022). Sales Forecasting for Firms based on Multiple Regression Model. In *Proceedings of the International Conference on Big Data Economy and Digital Management - Volume 1: BDEDM,* ISBN 978-989-758-593-7, pages 628-633. DOI: 10.5220/0011198600003440