JURNAL INFORMATIKA
PROXIES

Vol.10, No.1, July, 2026

Implementation of Regression Analysis to Improve Sales Prediction in Retail

Michael Dian Adi Wijaya[1], Yulianto Tejo Putranto[2]

[1] Teknik Informatika Fakultas Ilmu Komputer, Soegijapranata Catholic University, 22k10036@student.unika.ac.id
[2] Teknik Informatika Fakultas Ilmu Komputer, Soegijapranata Catholic University, yulianto@unika.ac.id

Corresponding Author Email: yulianto@unika.ac.id

https://doi.org/10.18280/proxies.v10i1.14988

Received: 2026-01-31
Revised: 2026-07-20
Accepted: 2026-07-20
Available online: 2026-07-23

Keywords:
Revenue Forecasting 1, Linear Regression 2, Nonlinear Regression 3, SVR 4 MSE 5, RMSE 6, Percentage Error 7, Time series 8 (no more than 8 keywords)

ABSTRACT

This study aims to compare the performance of linear and nonlinear regression algorithms in predicting monthly store revenue for the next two years to support business decision making and prevent bankruptcy risk. The models used include Linear Regression, Quadratic Non-Linear Regression, and Support Vector Regression (SVR) with Polynomial, RBF, Linear, and Sigmoid kernels, with evaluation using MSE, RMSE, and percentage as evaluation metrics. The results show that Linear Regression provides the best performance with the lowest error value, where the error percentage is in the range of 1.30% to 3.13% in five stores (S001–S005), making it effective in predicting revenue for the next two years in this case and dataset. The Quadratic Non-Linear model produces poor predictions, even producing negative values due to its mathematical nature, while Linear SVR and RBF show quite good performance although RBF tends to produce flat predictions. Polynomial SVR provides less than optimal results due to the characteristics of the data which is linear or close to linear. Overall, Linear Regression proved to be the most effective in this case, with the research limitations being that the data period was only two years and that time series methods had not been applied to capture seasonal patterns or specific events.

1. INTRODUCTION

Background

A store is a place where buying and selling goods occurs and relies on stable income to maintain sales and support business growth. A common problem that often occurs in stores is bankruptcy, where income is the most important factor in maintaining store stability, and many stores experience losses due to the inability to predict future financial conditions due to an imbalance between income and expenses. In this case, a prediction algorithm can be used to predict future income by utilizing historical sales data and can be used to predict income for approximately two years, which is useful for developing financial strategies. Based on these problems, this study was conducted to predict monthly store income for the next two years using historical sales data to prevent store bankruptcy.

Scope

This study uses the Retail Store Inventory Dataset from Kaggle containing sales transaction data for the period January 1, 2022, to December 31, 2023, to predict monthly revenue for the next two years (January 2024–December 2025). Predictions are made using Linear Regression and compared with non-linear regression algorithms such as Quadratic Regression and SVR 3 kernels, namely polynomial kernel, RBF kernel, and linear kernel to evaluate model performance. The variables used are derived from sales transaction data, with the output being predicted monthly revenue for each store (Store ID S001–S005) and model evaluation using MSE, RMSE, and RMSE percentage.

Objective

This study was conducted to predict the store's monthly revenue for the next two years using historical sales data. This prediction helps financial management and future operational planning to prevent bankruptcy.

2. LITERATURE STUDI

[1] Research by Bayu Susilo, Nur Ariesanto Ramdhan and Otong Saeful Bachri on the Application of the K-Nearest Neighbor Algorithm to Predict Digital Product Sales Obtained very good accuracy results, namely 95%. [2] Research by Salman Alfarisi, Rini Astuti, Fadhil M Basysyar on the Implementation of Data Mining to Use Linear Regression Algorithm to Predict Sales and Cash Flow at Ayam Geprek Cap Cangkir concluded that the predicted value of sales continues to increase, although the increase in value is not too large, the weakness of this research is that if the shop owner's capital decreases, it will definitely affect the predicted value. Journal created by [3] Bagus Nurislah, Dudih Gustian About Data Mining Application for Customer Pattern Analysis Using Apriori Algorithm with minimum support value of 30% and minimum confidence of 60% to obtain the rule that PUMP product purchases have a close relationship with various sports products, such as JML Volleyballs, RUBBER Basketballs, PMI Futsal Balls, LOCAL PRESS Volleyballs, and PMI Super Gold Thick Volleyballs, with a confidence level of 100%.

[4] Research by Fauzia Efita Zahra on the Application of Data Mining in Sales Analysis at the Nahdlatul Ulama Kramat Raya Executive Board Employee Cooperative Using the Apriori Algorithm using a minimum Support value of 19% and two Confidence values, namely 20% and 30%. The results obtained from the tests carried out were that the Sampoerna Mild item was the item most purchased by consumers with a value that coincided with buying 165gr of Kapal Api Coffee, then buying Sugar with a support value of 0.023 or 2.3%, with the highest confidence value of 49%. The weakness is that there is no comparative algorithm to optimize its sales strategy. The support and confidence values also affect the results, therefore an analysis is needed of the right value for this case. [5] Research by Annisa Oktavianti Hermadi,Wowon Priatna dan Allan D Alexander tentang Implementasi Algoritma K-Medoids Clustering To Find Temporary Profit in Financial Reports. This study introduces Decision making or the results of how many clusters are obtained. This study obtained the results of 2 clusters using the Davies Bouldin ndex (DBI) method then grouped with the K-Medoids Algorithm Cluster 1 is 634 and Cluster 2 is 362 and the Cluster 1 score is further from the center of gravity, so it is concluded that the initial gain level is high

The dataset used in the study [6] by Muhammad Nurdin Kurnia Ahadan, which discusses the Implementation of the Triple Exponential Smoothing (Brown) Method for Predicting Sales of Freebase and Salt Liquid Goods at CV.GRESSVAPE BALONGPANGGANG, is almost the same as what I want to use to predict goods that are often purchased by customers. The dataset contains the date and number of stock items in the inventory. The results of the Triple Exponential Smoothing (Brown) Method calculate the average MAPE value error, obtaining the smallest value, namely 23.838% and calculating the calculation of 3 test categories with reference data of 3 months, 6 months, 12 months with each using 9 different alphas, namely alpha 0.1 - 0.9 on the number of goods at CV.GRESSVAPE Balongpanggang, producing the smallest average MAPE value error. Research from [7] Dafwen Toresa, Muhammad Qadafi, Indah Muzdalifah, Fana Wiza, Rosda Syelly About Shoe Sales Data Analysis using the Apriori algorithm on Sneakers PKU obtained Association Rules, namely the first stating that if customers buy Ventela, they also buy socks, with a support value of 1 (100%) and confidence 1 (100%). The second rule states that if you buy socks, customers also buy Ventela, with the same support and confidence values, namely 1 (100%). This report can be my reference for making predictions on my dataset. Research [8] from Kurnia-Yahya, Winda Puspita Hidayanti on the Application of the K-Nearest Neighbor Algorithm for Classification of Vape (Electric Cigarette) Sales Effectiveness on "Lombok Vape On" obtained sales results of 10 pcs can be called the category of being able to achieve the target and if less than 10 pcs are sold then it will be called the category of not being able to achieve the target using k values 1,2,3,4,5,6,7,8,9,10 with cross validation 10. From the experiments that have been carried out, the results of the K-Fold Validation value 6 can be seen with the accuracy results obtained being 86.48% and AUC of 0.874.

Based on a literature review, this study aims to determine the best algorithm for predicting monthly store revenue. Linear Regression was chosen as the primary algorithm because previous research has shown it to provide good results with similar data characteristics, but that research has not yet compared it with other algorithms. Therefore, this study innovates by comparing Linear Regression with non-linear regression algorithms. Furthermore, several previous studies have demonstrated the importance of algorithms in obtaining the best model, so this calculation is expected to produce more accurate predictions and support more optimal decision-making.

3. RESEARCH METHODOLOGY

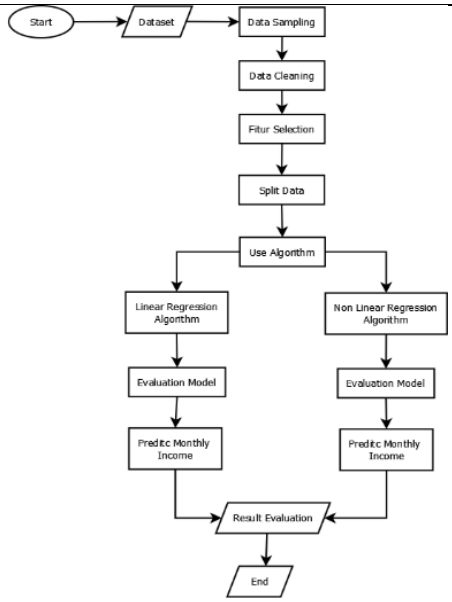

Figure 1. Research Flowchart

Dataset Collection

The dataset used in this study was taken from Kaggle entitled "Retail Store Inventory Forecasting EDA & Prediction". This dataset contains financial transaction data and inventory of goods, this dataset was taken from January 1, 2022 to January 1, 2024. The dataset has various types of columns including Date, Store ID, Product ID, Category, Region, Inventory Level, Units Sold, Units Ordered, Demand Forecast, Price, Discount, Weather Condition, Holiday/Promotion, Competitor Pricing, Seasonality. This dataset contains approximately seventy thousand rows presented in CSV format. This dataset can be downloaded from the Kaggle link below.
https://www.kaggle.com/code/beshoyemad24/retail-store-inventory-forecastin-eda-prediction/notebook

Dataset Preprocessing

The dataset preprocessing stages in this study include several main processes. First, data sampling was conducted by taking transaction data from January 1, 2023, to December 31, 2023, with approximately 73,000 data points representing almost the entire dataset from Kaggle and used to predict monthly revenue for the next two years. Second, data cleaning was performed to ensure data quality, ensuring the dataset contained no missing or incomplete data, as it consisted of real transaction and inventory data. Third, feature selection was performed by selecting relevant features, namely Inventory Level, Units Sold, Units Ordered, Demand Forecast, Price Discount, Holiday/Promotion, and Competitor Price, to improve model accuracy and prevent overfitting. Finally, data splitting was performed by dividing the data into 80% training data and 20% test data to optimally train the model and evaluate its performance in making predictions.

Linear Regression

Linear regression is a method used to model the relationship between one dependent variable and one or more independent variables. The main objective of this algorithm is to predict the value of the target variable based on the input variables. In the context of this study, linear regression is used to predict monthly revenue for two years from January 2024 to December 2025, namely by training the model on store revenue data, each month based on historical sales transaction data. The use of linear regression in predicting monthly revenue for two years was chosen because this algorithm is effective in various previous studies related to historical sales data. Sales transaction data such as product prices, purchase quantities, and discounts are examples of predictor variables that can influence revenue values linearly, making them very suitable for modelling using linear regression. Furthermore, based on several previous literature and research, the linear regression algorithm has been proven to provide good prediction results or performance in similar cases. Therefore, in this study, linear regression was chosen to be applied and its performance evaluated in the context of monthly revenue prediction using historical transaction data.

Nonlinear Regression

A nonlinear regression algorithm is a type of algorithm in statistics and machine learning used to model the relationship between independent variables (features) and dependent variables (targets) when the relationship cannot be accurately represented by a straight line. This means that changes in input variables do not produce a constant or linear change in output, so the graph of the relationship between input and output tends to be curved or follow a specific pattern. In this study, the nonlinear models used were

quadratic regression (squared) and Support Vector Regression (SVR). This second model is capable of capturing complex relationships between variables that may not be well explained by linear regression.

The purpose of using nonlinear regression is to compare its performance with that of a linear regression algorithm, thus determining which model provides more accurate monthly income predictions for the next two years. I chose to compare it with a nonlinear algorithm because this algorithm has the potential to adapt to curved or nonlinear data patterns, which may occur in sales transaction data. Thus, this comparison not only adds to the variety of algorithms used but also provides a more comprehensive understanding of the data characteristics and the effectiveness of the applied prediction model.

Metric Evaluation

Model evaluation is the result to show the performance in the following algorithm some of its performance
MSE is the Average of the square of the difference between the predicted value and the actual value. So the difference from the predicted value is reduced by the actual value, here is the RMSE formula

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

So if the smaller the predicted value obtained, the better the model is in predicting
RMSE is the square root of MSE, the meaning of RMSE is the average error value in the same unit as the target. Facilitates interpretation because it is not squared. Here is the formula for RMSE

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

4. RESULT

Store ID	Model	Metric Evaluation
S001	Linear Regression	MSE: 3,864,214,563.99, RMSE: 62,162.81, Error (%): 1.30%
S002	Linear Regression	MSE: 27,140,638,235.60, RMSE: 164,744.16, Error (%): 3.13%
S003	Linear Regression	MSE: 7,656,090,712.24, RMSE: 87,499.09, Error (%): 2.03%
S004	Linear Regression	MSE: 2,311,615,439.28, RMSE: 48,079.26, Error (%): 1.10%
S005	Linear Regression	MSE: 27,140,638,235.60, RMSE: 164,744.16, Error (%): 3.13%
S001	Nonlinear Quadratic	MSE: 22,842,082,434.47, RMSE: 151,135.97, Error (%): 2.75%
S002	Nonlinear Quadratic	MSE: 37,506,956,878.08, RMSE: 193,667.13, Error (%): 3.95%
S003	Nonlinear Quadratic	MSE: 91,007,552,670.82, RMSE: 301,674.58, Error (%): 6.03%
S004	Nonlinear Quadratic	MSE:264,799,136,034.44, RMSE: 514,586.37, Error (%): 7.97%
S005	Nonlinear Quadratic	MSE: 66,225,300,812.92, RMSE: 257,342.77, Error (%): 4.71%
S001	SVR Polynomial	MSE:102,658,635,642.35, RMSE: 320,403.86, Error (%): 6.35%
S002	SVR Polynomial	MSE:401,597,349,289.61 RMSE: 633,717.09 Error (%): 10.61%
S003	SVR Polynomial	MSE: 99,342,065,031.31, RMSE: 315,185.76, Error (%): 7.48%
S004	SVR Polynomial	MSE:109,526,206,950.87, RMSE: 330,947.44, Error (%): 7.82%
S005	SVR Polynomial	MSE: 610,341,463,727.69, RMSE:781,243.54, Error (%): 14.35%
S001	SVR RBF	MSE: 9,464,550,726.97, RMSE: 97,285.92, Error (%): 2.02%
S002	SVR RBF	MSE: 10,968,140,622.89, RMSE: 104,728.89, Error (%): 2.32%
S003	SVR RBF	MSE: 12,330,159,460.96, RMSE: 111,041.25, Error (%): 2.34%
S004	SVR RBF	MSE: 19,516,215,859.43, RMSE: 139,700.45, Error (%): 3.32%
S005	SVR RBF	MSE: 5,085,522,070.88, RMSE: 71,312.85, Error (%): 1.48%
S001	SVR LINEAR	MSE: 5,144,847,340.18, RMSE: 71,727.59, Error (%): 1.63%
S002	SVR LINEAR	MSE: 24,217,616,625.24, RMSE: 155,620.10, Error (%): 3.13%
S003	SVR LINEAR	MSE: 9,361,344,615.91, RMSE: 96,754.04, Error (%): 1.82%
S004	SVR LINEAR	MSE: 5,399,862,115.60, RMSE: 73,483.75, Error (%): 1.70%
S005	SVR LINEAR	MSE: 7,768,615,664.80, RMSE: 88,139.75, Error (%): 1.78%

4.1 Discussion

Experimental results show that Linear Regression performed best in predicting store revenue for the next two years, with a relatively small error rate, ranging from 1.10% to 3.13% across all stores (S001–S005). Although the MSE and RMSE values appear numerically large, the low error percentages indicate that the linear model is able to capture revenue patterns well, in accordance with the characteristics of revenue data that tends to be linear or near-linear.

Comparison with non-linear algorithms showed that Quadratic Regression produced less than optimal results, even producing negative revenue predictions in some periods, because negative revenue is highly unlikely in sales, and negative revenue is a mathematical value of the model. Meanwhile, SVR with various kernels produced varying performance, with the Polynomial kernel being less than optimal, the RBF kernel performing reasonably well but producing relatively flat predictions, and the Linear kernel providing results closest to the linear model. However, overall, the SVR model was still unable to surpass the performance of Linear Regression.

Thus, Linear Regression is the superior model in this study for predicting monthly store revenue over the next two years and is expected to help prevent bankruptcy and support financial planning. A limitation of this study is the limited dataset period; therefore, further research is recommended to use a longer data set.

5. CONCLUSIONS

Based on an evaluation using MSE, RMSE, and error percentages on five stores (S001–S005), the results show that Linear Regression is the best model for predicting monthly revenue for the next two years with a low error rate of around 1%–3%. This indicates that the historical transaction data used tends to have a linear or near-linear pattern. Conversely, Quadratic Regression performs less well with a higher error value and even produces negative revenue predictions; negative revenue is unlikely to occur in the store.

In the SVR model, results vary depending on the kernel used. The RBF and Linear kernels provide quite good performance with a relatively small error percentage, with the Linear kernel even able to compete with Linear Regression and being more suitable for some stores such as S003. However, the RBF kernel has the disadvantage of tending to stagnate revenue predictions. Meanwhile, the SVR with the Polynomial kernel produces less than optimal performance with a larger error value and some negative revenue predictions.

Overall, the prediction of average monthly revenue for each store over the next two years can be used as a basis for financial planning and bankruptcy prevention. The Linear Regression model proved most effective in supporting business decision-making and operational planning in this dataset. For further research, it is recommended to use a Time Series method to account for the influence of seasonal trends.

ACKNOWLEDGMENT

I would like to express my gratitude to the Lord Jesus Christ and Mother Mary for answering my prayers to complete this thesis. Furthermore, I am also grateful for God's blessings that have brought me to this point. I believe that God brought me to this point for a reason. Thank you, God. I believe that your blessings are always with me. To my parents, thank you for your support and encouragement while working on this thesis. They always listened to my complaints and worries while working on this thesis, they also gave me the encouragement to complete this thesis, without them I would not have been able to study at Soegijapranata Catholic University and be able to pursue my education this far. I dedicate this thesis to them as an expression of my gratitude, therefore I worked on it with all my heart.

REFERENCES

[1] Susilo B, Ramdhan NA, Bachri OS. Penerapan Algoritma K-Nearest Neighbor untuk Prediksi Penjualan Produk Digital: Application of the K-Nearest Neighbor Algorithm for Predicting Digital Product Sales. MALCOM: Indonesian Journal of Machine Learning and Computer Science 2024; 4: 1466–1476.

[2] Alfarisi S, Astuti R, Basysyar FM. IMPLEMENTASI DATA MINNING MENGGUNAKAN ALGORITMA REGRESI LINEAR UNTUK PREDIKSI PENJUALAN DAN CASHFLOW DI AYAM GEPREK CAP CANGKIR. JATI (Jurnal Mahasiswa Teknik Informatika) 2024; 8: 3392–3395.

[3] Nurislah B, Gustian D. PENERAPAN DATA MINING UNTUK ANALISIS POLA PEMBELIAN PELANGGAN DENGAN MENGGUNAKAN ALGORITMA APRIORI. Jurnal Rekayasa Teknologi Nusa Putra 2024; 10: 44–52.

[4] Zahra FE. PENERAPAN DATA MINING DALAM ANALISIS PEJUALAN PADAKOPERASI KARYAWAN PENGURUS BESAR NAHDLATUL ULAMA KRAMAT RAYA DENGAN MENGGUNAKAN ALGORITMA APRIORI. diploma, UNUSIA, https://si.unusia.ac.id/ (2022, accessed 15 May 2025).

[5] Hermadi AO, Priatna W, Alexander AD. Implementasi Algoritma K-Medoids Clustering Untuk Mencari Keuntungan Sementara Dalam Laporan Keuangan. JURNAL TEKNOLOGI DAN ILMU KOMPUTER PRIMA (JUTIKOMP) 2023; 6: 6–11.

[6] Hermadi AO, Priatna W, Alexander AD. Implementasi Algoritma K-Medoids Clustering Untuk Mencari Keuntungan Sementara Dalam Laporan Keuangan. JURNAL TEKNOLOGI DAN ILMU KOMPUTER PRIMA (JUTIKOMP) 2023; 6: 6–11.

[7] Toresa D, Qadafi M, Muzdalifah I, et al. Analisis Data Penjualan Sepatu Menggunakan Algoritma Apriori Pada Sneakers PKU. Technologica 2025; 4: 23–34.
[8] Yahya K-, Hidayanti WP. Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Efektivitas Penjualan Vape (Rokok Elektrik) pada â€œLombok Vape Onâ€□. Infotek: Jurnal Informatika dan Teknologi 2020; 3: 104–114.
[9] Adinarta F, Putranto YT. PREDICTING FLIGHT DELAY USING RANDOM FOREST ALGORITHM, XGBOOST ALGORITHM, AND STACKING ENSEMBLE METHOD. Proxies : Jurnal Informatika 2025; 8: 88–106.
[10] Hazanisya Priliani Hutrim , Lukman Bachtiar. Aplikasi Pendukung Keputusan Prediksi Penjualan Air Minum Isi Ulang Menggunakan Metode Trend Moment.