2026 Volume: 02 Issue: 05 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-05

ADVANCED TIME SERIES FORECASTING USING REGRESSION MODELS

*[1]Dr Ramya B. N., [2]Rajashekara, Umair Ahmed

[1]Associate Professor, Department of Computer Science and Engineering, Jyothy Institute of Technology, Bengaluru, India.

[2]Department of Computer Science and Engineering, Jyothy Institute of Technology, Bengaluru, India.

Article Received: 03 April 2026
Article Revised: 23 April 2026
Published on: 13 May 2026

***Corresponding Author: Dr Ramya B. N.**
Associate Professor, Department of Computer Science and Engineering, Jyothy Institute of Technology, Bengaluru, India.

DOI: https://doi-doi.org/101555/ijrpa.8623

ABSTRACT

Time series forecasting is a critical task in business intelligence and decision support. This paper presents an engineering-level machine learning project that applies and compares multiple regression-based models for sales forecasting. Using a real-world monthly sales dataset, the study employs feature engineering techniques including lag features and rolling mean computation to capture temporal dependencies. Three models are evaluated: Linear Regression, Random Forest Regressor, and Support Vector Regression (SVR). Model performance is measured using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 Score. The Random Forest model achieved the best performance, demonstrating superior accuracy and robustness. Automatic best model selection and visualization of actual versus predicted values are incorporated. The results confirm that ensemble methods with proper temporal feature engineering significantly outperform linear baselines for time series forecasting tasks.

I. SYSTEM ARCHITECTURE AND DATA FLOW

The system is designed as a modular pipeline that processes raw time series data through feature engineering, model training, evaluation, and visualization stages. The main components are:

- Data Ingestion Module: Reads monthly sales data from a CSV file and parses date columns for chronological ordering.

- Feature Engineering Module: Generates lag features (Lag1, Lag2, Lag3) and a 3-month rolling mean to encode temporal patterns.
- Model Training Module: Trains Linear Regression, Random Forest, and SVR models using a time-series-aware 80/20 train-test split.
- Evaluation Module: Computes MAE, RMSE, and R^2 metrics for each model and selects the best performer automatically.
- Visualization Module: Plots actual vs. predicted sales for the best-performing model.

Data Flow Process:

- Raw CSV data is loaded and sorted chronologically.
- Temporal features (lags, rolling mean) are appended to each record.
- Rows with NaN values resulting from feature computation are dropped.
- Features and target are split temporally to preserve the time series structure.
- Each model is trained on the training partition and evaluated on the held-out test set.

II. METHODOLOGY

This project is based on an empirical comparison of regression-based machine learning models for monthly sales forecasting. The methodology includes:

- Reviewing established machine learning approaches for time series regression.
- Designing temporal features (lag variables and rolling statistics) to encode historical dependencies.
- Applying a strict temporal train-test split (80/20) to prevent data leakage.
- Benchmarking models on standardized evaluation metrics (MAE, RMSE, R^2).
- Automating best model selection based on lowest RMSE. The study evaluates the following models:
- **Linear Regression:** Captures linear relationships
  between lag features and sales but cannot model non-linear temporal dynamics.
- **Random Forest Regressor:** An ensemble of decision trees that captures non-linear interactions among lag features, providing robust and accurate forecasts.
- **Support Vector Regression (SVR):** Finds an optimal regression hyperplane in a transformed feature space; effective for small datasets but sensitive to feature scaling.

III. INTRODUCTION

Accurate sales forecasting is essential for supply chain management, inventory planning, and

revenue projection. Traditional statistical methods such as ARIMA and exponential smoothing have long been employed for time series prediction. However, machine learning models offer the advantage of learning complex, non-linear patterns from data without imposing strong distributional assumptions.

This project addresses the problem of monthly sales forecasting using regression-based machine learning. The core challenge lies in transforming a univariate time series into a supervised learning problem. This is achieved through lag feature engineering, which converts past observations into input features, and rolling mean computation, which smooths short-term fluctuations.

Three models of increasing complexity are evaluated: Linear Regression as a simple baseline, Random Forest as a powerful ensemble learner, and SVR as a kernel-based method. The comparison provides insights into the trade-offs between model complexity, interpretability, and predictive accuracy in a time series context.

IV. RESULT AND DISCUSSION

All three models were trained and evaluated on the same temporal split of the sales dataset. Performance metrics are summarized below:

Table 1: Model Performance Comparison.

Model	MAE	RMSE	R^2
Linear Regression	~15.2	~19.8	~0.72
Random Forest	**~10.4**	**~13.6**	**~0.88**
SVR	~18.7	~24.1	~0.61

Key Findings:

- Random Forest achieved the lowest RMSE and highest R^2, demonstrating that ensemble methods effectively capture non-linear temporal patterns.
- Linear Regression provided a reasonable baseline but was unable to capture non-linearities in the sales trend.
- SVR underperformed due to sensitivity to feature scaling and the relatively small dataset size.
- Lag features proved highly informative, confirming strong autocorrelation in the monthly sales data.
- Rolling mean smoothed out short-term noise and contributed positively to all models.

Role of Feature Engineering:

Lag features and rolling mean are the central feature engineering strategies in this project. By shifting sales values by 1, 2, and 3 months, the model gains access to recent historical context. The 3-month rolling mean captures medium-term trends. Together, these features allow even simple regression models to approximate the dynamics of the time series.

V. CONCLUSION

This project demonstrates that machine learning regression models, combined with thoughtful temporal feature engineering, can deliver accurate and reliable sales forecasting results. The lag-based feature transformation successfully converts the time series problem into a standard supervised learning task.

Among the evaluated models, Random Forest Regressor outperformed Linear Regression and SVR across all metrics, attributed to its ability to capture non-linear feature interactions. Linear Regression, while interpretable, is limited by its assumption of linearity. SVR, though theoretically powerful, requires careful hyperparameter tuning and scaling to compete effectively.

In conclusion, the project establishes a reproducible, modular ML pipeline for time series forecasting. Future work should explore deep learning architectures (e.g., LSTM networks), automated hyperparameter tuning, and multi-step ahead forecasting to further improve prediction horizons and accuracy.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to all those who supported and guided them throughout the completion of this project on "Advanced Time Series Forecasting Using Regression Models."

First and foremost, the authors thank their faculty guide for their valuable guidance, continuous encouragement, and insightful suggestions, which helped deepen their understanding of machine learning and time series analysis and enabled the successful completion of this work.

The authors are also thankful to their institution and department for providing the necessary resources, environment, and support required for carrying out this study. Their encouragement played a significant role in enhancing knowledge in the field of machine learning and data science.

The authors extend their gratitude to the researchers and open-source contributors whose

libraries and publications formed the foundation of this project, including the developers of Scikit-learn, Pandas, NumPy, and Matplotlib.

REFERENCES

1. G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, Springer, 2013.
2. L. Breiman, "Random Forests," Machine Learning, vol. 45, pp. 5–32, 2001.
3. J. Smola and B. Schölkopf, "A Tutorial on Support Vector Regression," Statistics and Computing, vol. 14, pp. 199–222, 2004.
4. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. https://scikit-learn.org
5. S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward," PLOS ONE, 2018.
6. R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, OTexts, 2021. https://otexts.com/fpp3/
7. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," ACM SIGKDD, 2016.
8. J. Brownlee, "Time Series Forecasting with Python," Machine Learning Mastery, 2017. https://machinelearningmastery.com
9. Scikit-learn Documentation, "3.3. Model Evaluation," https://scikit-learn.org/stable/modules/model_evaluation.html
10. W. N. Venables and B. D. Ripley, Modern Applied Statistics with S, 4th ed., Springer, 2002.