International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

A Review of Linear Regression and Support Vector Regression

Priyanka Pawar[1], Dr. G.J. Chhajed[2], Monali Rahul Bhosale[3], Sanika Gonjari[4], Sharwari Kshirsagar [5]

[1.] *AI &DS (Computer Engineering) VP's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati.*
[2]*HOD AI &DS (Computer Engineering) VP's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati.*

[3]*Assistant Professor AI &DS (Computer Engineering) VP's Kamalnayan Bajaj Institute of Engineering and Technology,Baramati.*
[4] *AI&DS (Computer Engineering) VP's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati.*
[5]*AI&DS (Computer Engineering) VP's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati.*

Abstract - Predicting the future of a business requires analyzing key insights such as consumer behavior, product performance, and profitability using current and historical data. Statistical techniques play a vital role in generating these insights and enabling accurate forecasting, particularly with time series data. Applications like weather forecasting, financial markets, and stock analysis often integrate historical and real-time data for better accuracy. Regression models, including linear regression and support vector regression (SVR), are commonly employed to analyze time series data. Linear Regression (LR) and Support Vector Regression (SVR) are two widely used machine learning algorithms for predicting continuous outcomes. LR is a parametric method that models the relationship between independent and dependent variables by fitting a linear equation, assuming a linear dependency. It is simple, interpretable, and efficient for linearly separable data. SVR, a non-parametric technique derived from Support Vector Machines, employs kernel functions to handle non-linear relationships and constructs a decision boundary by maximizing the margin within an error tolerance (epsilon). SVR is robust to outliers and effective for complex, high-dimensional data.

Key Words: Regression, Linear Regression, Support Vector Regression, Data analytics, Supervised Machine Learning, Classification.

1.INTRODUCTION

Regression analysis plays a crucial role in the field of machine learning, particularly in Predicting continuous outcomes based on input data. Among the various regression Techniques, Linear Regression (LR) and Support Vector Regression (SVR) are two of the Most widely used methods, each offering unique advantages for different types of datasets And problems. Linear Regression, one of the oldest and most fundamental techniques, is a Statistical method that models the relationship between independent variables and a Dependent variable through a linear equation. It is widely appreciated for its simplicity, Interpretability, and efficiency, making it a go-to solution for problems where the relationship Between variables is expected to be linear. Its ability to provide direct insight into the strength And direction of the relationships between variables is particularly useful in domains like Economics, healthcare, and engineering [2]. Support Vector Regression, a more recent Advancement based on the principles of Support Vector Machines (SVM), offers a more Flexible approach for handling non-linear regression problems. SVR is designed to find the Optimal hyper plane that best fits the data within a specified margin of error, making it robust To outliers and capable of modeling complex relationships using kernel functions. Its Versatility and ability to generalize to high-dimensional spaces make SVR especially useful For applications in areas like bioinformatics, finance, and image processing[5].

2. BACKGROUND

The growth of data analytics has accelerated with the advent of Big Data technologies, enabling More efficient handling and analysis of massive datasets. This evolution has given rise to advanced Data mining tools and techniques that support predictive analytics, a key component of Business Intelligence (BI). Predictive analytics plays a significant role in market analysis by leveraging Machine learning algorithms to generate accurate forecasts [5].

Among machine learning techniques, supervised learning models are widely used due to their Ability to learn patterns from labeled data. Linear Regression (LR) is a fundamental supervised Learning algorithm that establishes a relationship between dependent and independent variables. Techniques like the Least Median Squares (LeastMedSq) function are applied to minimize the Median of squared residuals, enhancing the accuracy of linear regression models [7].

Support Vector Regression (SVR), another prominent method, extends Support Vector Machines For regression tasks. It supports both linear and non-linear kernels, making it flexible for diverse Datasets. In particular, SVR with a linear kernel can perform tasks similar to linear regression, While the Sequential

Minimal Optimization (SMO) regression function is employed to optimize Performance [11].

To assess and compare the effectiveness of these models, evaluation metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are utilized. These metrics quantify Prediction accuracy and help determine the suitability of models for specific business applications. Previous research highlights the importance of selecting appropriate regression techniques and Evaluation metrics to enhance decision-making in predictive analytics and business intelligence Systems [9].

3. SUPERVISED MACHINE LEARNING

Supervised Machine Learning learns a function from labeled training data to predict Outcomes. The training dataset comprises tuples (T) with input attributes (X) and a target value (y). Attributes can be numerical or categorical, and outcomes can be continuous or multi-valued. The algorithm uses this data to map inputs to outputs efficiently, supporting tasks like classification And regression. The goal is to find a function that, when given a new input, will predict the correct Output [1].

Process:

1. Training: A model learns from the training data (input-output pairs).
2. Testing: The model's performance is evaluated on a separate test dataset to assess hpw Well it generalizes.
3. Prediction: Once trained, the model can predict the output for new unseen data.

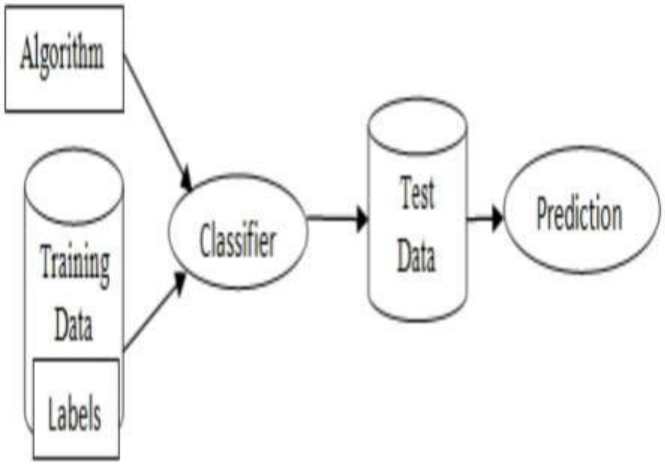

Fig. 1. Supervised Machine Learning

3.1Classification

Classification predicts categorical labels, such as determining whether a student passes or fails. A Classifier is created using a training dataset, where each tuple includes attributes and a class label. This process, called learning, builds rules to classify new data. The classifier is tested on a separate Dataset to evaluate its accuracy. Training attributes can be numerical or categorical, and target Outcomes can have multiple possible values[6].

Here are some classification algorithms:

• Logistic Regression

• Support Vector Machine

• Random Forest

• Decision Tree

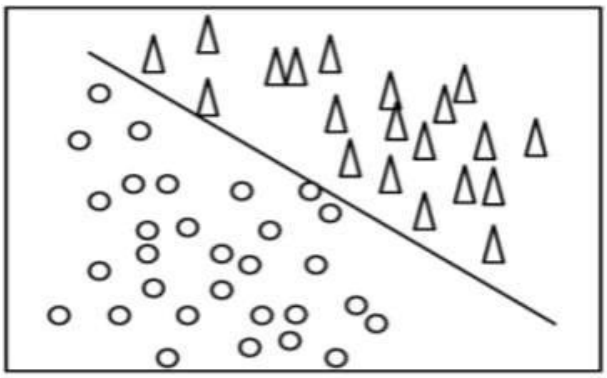

Fig 2. Classification

3.2Regression

Regression is a statistical method used to identify the relationship between variables, specifically Dependent (Y) and independent (X) variables. The goal of regression is to find a mathematical function that best approximates the mapping between the inputs and outputs. Regression is widely used in various fields, including economics, finance, healthcare, and more, to predict outcomes like stock prices, disease progression, or customer behavior. It is represented as:

$$Y=f(X,\beta)$$

Where β are unknown parameters.

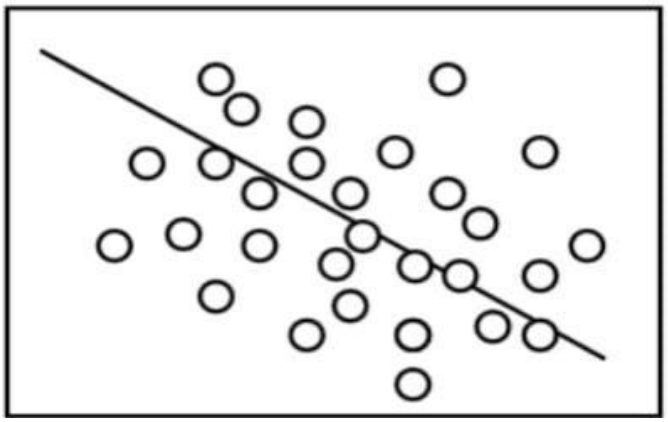

Fig 3. Regression

4.REGRESSION MODEL

A regression model is a statistical method used to understand the relationship between a dependent Variable (also called the target) and one or more independent variables (also called predictors). The goal of a regression model is to predict the value of the dependent variable based on the values Of the independent variables [1]. Here's a breakdown of the most common types of regression Models:

4.1 Linear Regression

Linear regrssion is a statistical technique used to model the relationship between a Dependent variable (Y) and one or more independent variables (X). The goal is to fit a linear Equation to observed data to predict the dependent variable based on the independent variables [2].

4.1.1 Simple Linear Regression:

Simple Linear Regression is used when there is only one independent variable and a linearRelationship exists between the dependent and independent variables. The goal is to find the Best-fitting line that describes how changes in the independent variable influence the dependent Variable. The equation is:

$$Y=a+bX+\epsilon$$

4.1.2 Multiple Linear Regression:

Multiple Linear Regression is an extension of simple linear regression used when there Are two or more independent variables. It is used when you belive that the dependent Variable is influenced by serveral facrors[13].The equation is:

$$Y=a+b1X1+b2X2+\cdots+bnXn+\epsilon$$

4.1.2.1 LeastMedSq Linear Regression

Least Median of Squares is a robust method of linear regression that aims to Minimize the median of squared residuals rather than the sum of squared residuals, As in ordinary least squares regression. It is less sensitive to outliers compared to Ordinary least squares.

4.2 Support Vector Regression

Support Vector Regression (SVR) is an extension of Support Vector Machines (SVM) [11]designed For regression tasks. While SVM is typically used for classification, SVR models a continuous Output instead of discrete labels, making it suitable for predicting numerical data. SVR aims to Find a function f(x) that has at most a deviation of ϵ\epsilon$\epsilon$ (epsilon-tube) from the true Target y for the training data, while also being as flat as possible.Unlike traditional regression That minimizes the error, SVR focuses on finding a solution within a margin of tolerance[8].

4.2.1 Linear Support Vector Regression (Linear SVR):

Linear SVR assumes a linear relationship between the input features and the target variable. In this Case, the data is approximated using a linear function, similar to linear regression, but with more Emphasis on a margin of tolerance.

- Hyperplane: In the linear SVR model, the goal is to find a hyperplane that best fits the data, With some error tolerance.
- Epsilon (ε): This defines a margin around the predicted values where no penalty is given For errors. If the data points are within this margin, the model does not penalize them.
- Regularization ©: This controls the trade-off between a smooth margin and training error Minimization. A large value of C leads to fewer support vectors, while a smaller C allows More flexibility in the model.

The model is formulated as:

$$Y=w\cdot x+b$$

4.2.1.1 SMOReg Linear Kernel

SMOReg (Sequential Minimal Optimization for Regression) is an implementation Of Support Vector Machines (SVM) for regression tasks. It is used to predict Continuous values rather than classify discrete categories. In the context of the Linear Kernel, the algorithm assumes that the relationship between the input Features and the target variable is linear.

4.2.2 Non-Linear Support Vector Regression (Non-Linear SVR):

Non-linear SVR is used when the relationship between the input features and the target Variable is not linear. In this case, the input data is transformed into a higher-dimensional Space using a kernel function, where the data becomes linearly separable.

- Kernel Trick: Non-linear SVR uses a kernel function to map the original input space into a

Higher-dimensional feature space. Common kernels include the Radial Basis Function (RBF) Kernel, polynomial kernel, and sigmoid kernel.

- Kernel Function: The kernel function computes the inner product of the data points in the Transformed feature space without explicitly computing the transformation. The most popular Choice is the RBF kernel.

5. APPLICATION

5.1 Linear Regression:

1. Finance

Stock Price Prediction: Estimating future stock prices based on historical data.

Risk Assessment: Analyzing the relationship between market factors and portfolio returns.

2.Healthcare

Disease Prediction: Predicting disease risk based on patient data (e.g., age, weight, medical history).

Healthcare Costs: Estimating treatment or insurance costs based on demographic and health variables.

3.Marketing

Sales Forecasting: Predicting sales based on factors like advertising spend, seasonality, or economic trends.

Customer Behavior Analysis: Understanding the effect of pricing or promotional strategies on sales.

4.Education

Student Performance Analysis: Predicting academic performance based on attendance, study hours, and other variables.

Resource Allocation: Assessing the relationship between funding and student outcomes.

5.Real Estate

Property Valuation: Estimating property prices based on size, location, and amenities.

Market Trends Analysis: Analyzing the impact of economic factors on real estate trends.

5.2 Support Vector Regression:

1.Stock Price Prediction

SVR can model the trends and movements of stock prices by analyzing historical data and identifying patterns.

2.Energy Consumption Forecasting

Used in energy sectors to predict electricity, gas, or water consumption based on historical data and external factors like temperature or seasonality.

3.Weather Forecasting

SVR can be used to predict temperature, humidity, or rainfall based on historical meteorological data.

4.Financial Risk Assessment

In finance, SVR is used for predicting credit scores or risk factors for loans and investments.

5.Biometric Applications

Used in face recognition or other biometric systems for creating predictive models.

6.Medical Diagnosis

SVR helps in predicting the progression of diseases or treatment outcomes based on patient data.

6. CONCLUSIONS

Data analytics and business intelligence play a crucial role in today's competitive market. When dealing with multivariate time-series data, selecting the right data model is essential for achieving accurate results. Linear regression is a widely used method for such analyses, and it includes several functions that impact its performance.In this study, we compared the performance of two functions—LeastMedSq and SMOreg—within the linear regression framework for multivariate and time-series datasets. Our analysis showed that the LeastMedSq function performs better than SMOreg for linear regression in terms of accuracy.

REFERENCES

[1] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", J. of Informatica, Vol. 31, pp.249-268, 2007.

[2] GAF Seber, AJ Lee, "Linear regression analysis", Wiley Series in Probability and Statistics,2012.

[3] DC Montgomery, EA Peck, GG Vining, "Introduction to linear regression analysis", Wiley Series in Probability and Statistics, 2015.

[4] B. Akgün and Ş. G. Öğüdücü, "Streaming linear regression on Spark MLlib and MOA," in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, 2015, pp. 1244-1247.

[5] H.-I. Lim, "A Linear Regression Approach to Modeling Software Characteristics for Classifying Similar Software," in 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), 2019, pp. 942-943.

[6] M. R. Sarkar, M. G. Rabbani, A. R. Khan, and M. M. Hossain, "Electricity demand forecasting of Rajshahi City in Bangladesh using fuzzy linear regression model," in 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2015, pp. 1-3.

[7] J. Wu, C. Liu, W. Cui, and Y. Zhang, "Personalized Collaborative Filtering Recommendation Algorithm based on Linear Regression," in 2019 IEEE International Conference on Power Data Science (ICPDS), 2019, pp. 139-142.

[8] H. Roopa and T. Asha, "A linear model based on principal component analysis for disease prediction," IEEE Access, vol. 7, pp. 105314-105318, 2019.

[9] G. A. Seber and A. J. Lee, Linear regression analysis vol. 329: John Wiley & Sons, 2012.

[10] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to linear regression analysis vol. 821: John Wiley & Sons, 2012.

[11] S. Kavitha, S. Varuna, and R. Ramya, "A comparative analysis on linear regression and support vector regression," in 2016 Online International Conference on Green Engineering and Technologies (IC-GET), 2016, pp. 1-5.

[12] Abdulazeez, A., Salim, B., Zeebaree, D., & Doghramachi, D. (2020). Comparison of VPNProtocols at Network Layer Focusing on Wire Guard Protocol
.
[13] Quiming N S, Denola N L , Saito Y ,et al.Multiple linear regression and artificial neural network retention prediction models for ginsenosides on a polyamine-bonded stationary phase in hydrophilic interaction chromatography[J]. Journal of Separation Science, 2015, 31(9): 1550-1563.DOI:10.1002/jssc.200800077.