A Survey on Regression Algorithms in Machine Learning

Prashant Kalyane[1], Jamshed Damania[2], Harsh Patil[3], Mahadev Wardule[4], Prof. Priyanka Shahane[5]

[1, 2, 3, 4] Student, SCTR's Pune Institute of Computer Technology, Pune.
[5] Assistant Professor, SCTR's Pune Institute of Computer Technology, Pune.

Abstract: This review paper provides a detailed analysis and comparison of eight popular regressor algorithms: Polynomial, Random Forest, LASSO, Decision Tree, Linear, and Neural Network regression. This study aims to provide a comprehensive overview of contemporary regression analysis approaches by examining the algorithms' complexities, advantages, disadvantages, and practical applications across several domains. The performance of several regression methodologies is thoroughly evaluated on a range of datasets and real-world use cases in order to facilitate an educated choice of the optimal regression methodology for specific analytical tasks. This study enhances regression analysis approaches in this way.

Keywords: Regression algorithms, Decision tree regression, Linear regression, LASSO regression, Random forest regression, Neural network, Polynomial regression.

1 Introduction

In contemporary data analysis, regression algorithms are crucial instruments for forecasting and modeling continuous variables. Because there are so many different regression approaches available, it can be challenging for practitioners and researchers to select the optimal one for a particular set of applications. To address this, in our survey study we provide an in-depth analysis of six popular regression algorithms, aiming to elucidate their basic principles, advantages, disadvantages, and applications. The polynomial, random forest, LASSO, decision tree, neural network, logistic, and ridge regression models are the ones we focus on. Every model has a benefit over the others and may be applied to different types of data analysis. When determining the most effective way to use these algorithms in various fields, it is essential to grasp the nuances that are present in them. This study analyses the performance and applicability of multiple methods over a range of datasets and use cases to help researchers and analysts select the appropriate regression methodology for their specific analytical needs. Through investigating the intricacies of these regression algorithms, we hope to further regression analysis techniques and their practical application in the social sciences, engineering, finance, and healthcare areas. By presenting the benefits and drawbacks of each approach, we hope to empower practitioners to choose wisely for their data analysis endeavours.

2 Related Work

[1] In recent years, machine learning has gained immense popularity due to its ability to train models to perform complex tasks. Machine learning algorithms are one of the cornerstones of artificial intelligence, which is currently ubiquitous in many aspects of our lives. For machine learning algorithms to be effective, training datasets are essential. For machine learning algorithms to achieve reasonable accuracy, well-prepared input datasets must be used during training. "Data preparation" refers to a series of procedures that enhance a dataset's fit for machine learning. The aim of the paper is to present a summary of several techniques for preparing data and analyze their impact on the accuracy of the final model. Various machine learning strategies are considered and assessed to train a model to predict numerical variables that are not based on neural networks.

[2] The authors of this paper demonstrate that an appropriate strategy for forecasting algorithm performance is multi-variable linear regression based on trees. By taking into account prior machine learning

experiences, authors construct meta-knowledge for supervised learning. The idea is to use summary data about these datasets along with previous algorithmic performance on them to build this meta-knowledge. The authors find that transformed datasets obtained by taking a high dimensional feature space and reducing it to a smaller dimension still retain important characteristic knowledge required to predict algorithm performance. They achieve this by adding descriptive features and a misclassification cost to pure statistical summaries. When applied to nominal and numerical data gathered from real-world contexts, the author's method works incredibly well.

[3] The aim of this effort is to enhance comprehension of the real-world uses of different machine learning models. This study looks at data from throughout the world as well as the current trend or pattern of Covid-19 transmission in India. With the help of data from the Indian Ministry of Health and Family Welfare, this study shows a range of global trends and patterns. The data for the study was gathered between January 22, 2020, and June 24, 2020, a total of 154 days. By analysing the data more thoroughly, more conclusions can be drawn for later use.

[4] AdaBoost, K-Nearest Neighbours, and random forest algorithms are used in this paper's investigation. The study looks at the association between the innovation index and the gross regional product using the three machine learning regression methodologies previously described. Based on this large, complex, multidimensional dataset, a substantial relationship is discovered between the seven primary economic parameters. The three approaches combined yielded predictions with an accuracy of more than 0.85; the random forest method had the highest accuracy of 0.95. The number of trademark authorizations is the most important characteristic, according to a feature importance study. The machine learning algorithm model can be improved and new application scenarios can be added with the help of this effort.

[5] The authors of this work evaluate many widely used and well-liked machine learning approaches for regression in the energy disaggregation task in this study. In particular, the K-Nearest-Neighbors, Support Vector Machines, Deep Neural Networks, and Random Forest methods were evaluated on five datasets comprising seven different sets of statistical and electrical characteristics. Additionally, the Non-Intrusive Load Monitoring method was examined. The experiment's results demonstrated how important it is to select appropriate features and regression techniques. For energy disaggregation accuracy, the Random Forest regression method yielded the best results.

3 Regression Algorithms

1) Decision tree regressor

Within the realm of supervised learning, the Decision Tree Regressor is a well-liked machine learning algorithm. Regression analysis is mostly used to solve regression issues, where the goal is to predict a continuous numerical value from input features.

1.1) Algorithm

Choose a Feature: The characteristic that divides the dataset into two subsets is the best choice. This is accomplished by analysing different splitting criteria, like mean squared error, to identify the feature that significantly lowers the target variable's variation.

Divide the Info: Based on the threshold value of the chosen feature, divide the data into subsets. Making subgroups with the least amount of volatility in the target variable is the aim.

Repeat: On each subset, recursively repeat steps 1 and 2 until a stopping condition is satisfied. This could be a minimum number of samples per leaf, a maximum depth, or another criterion.

Establish Leaf Nodes: Establish leaf nodes and give them the average goal value of the samples in that subset after the stopping requirement is satisfied.

A Decision tree regressor can be used to make predictions by going through the tree from the root to a leaf node and returning the target value connected to that leaf.

1.2) Benefits

Simplicity: Decision tree regressors offer insights into the decision-making process of the data and are simple to comprehend and interpret, making them appropriate for non-experts.

Managing Non-linearity: Unlike linear models, they are able to simulate non-linear connections between features and the goal variable.

Can handle mixed data: Decision trees are adaptable for a variety of applications since they can handle both numerical and categorical data.

Feature Importance: Decision trees have the ability to produce feature importance scores, which are useful for selecting features and identifying the most important factors.

1.3) Drawbacks

Overfitting: Decision trees have a tendency to overfit the training set, particularly when the data is complicated and deep. Pruning is one method used to lessen this problem.

Instability: Minor modifications to the training set can produce noticeably different tree architectures, which can cause instability.

Lack of global optimality: At each node, the algorithm selects locally optimal options, but these options may not result in the ideal tree structure overall.

1.4) Use

The Decision Tree Regressor is frequently employed in a number of domains, such as:

Finance: Credit risk assessment and stock price prediction.

Healthcare: Disease diagnosis and patient outcome prediction.

Marketing: Forecasting sales and segmenting customers.

Climate modelling and environmental parameter prediction are aspects of environmental science.

1.5) Applications

Housing Price Prediction: Estimate a home's cost by taking into account factors like location, size, and amenities.

Estimate crop yields by taking into account variables such as weather, soil conditions, and agricultural techniques.

Forecasting Energy Consumption: Take into account variables like temperature, occupancy, and time of day to estimate energy consumption in buildings.

1.6) Used Functions

Mean squared error: A decision tree regressor is examining the possibility of splitting a node that has a set of data points (samples) with real target values of Y_1, Y_2, ..., Y_n into two child nodes (left and right), each with a series of data points and forecast values of Y_{left}. Yes, that split's MSE may be computed as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 \qquad (1)$$

2) Linear Regression

A basic and popular machine learning method for resolving regression issues is linear regression. It uses one or more input features to forecast a continuous numerical target variable in an easy-to-understand manner.

2.1) Algorithm

The goal of the linear regression algorithm is to create a linear relationship between the target variable (dependent variable) and the input features (independent variables). The algorithm looks for the linear equation that minimizes the sum of squared differences between the target values and the predicted values and finds the best match. The linear regression model can be expressed mathematically as follows:

$Y = B_0 + B_1 * X_1 + B_2 * X_2 + ... + B_n * X_n$ (2)

Where

The target variable is denoted by Y.

The intercept, or bias, is B_0.

The coefficients of the input features X_1, X_2,..., X_n are B_1, B_2,..., B_n.

During the training phase, the coefficients (B_0,B_0,B_2,...,B_n) are found, usually with the use of techniques like gradient descent or Ordinary Least Squares (OLS).

2.2) Benefits

Simplicity: Both novices and experts can benefit from linear regression's ease of comprehension and interpretation. It seems to make sense that the variables have a linear connection.

Interpretable Coefficients: A linear regression model's coefficients can help you comprehend how each attribute affects the target variable because they are easy to understand.

Efficiency: Large datasets may be effectively trained and predicted using linear regression due to its processing efficiency.

2.3) Drawbacks

Assumption of Linearity: The linearity of the connection between variables is the foundation of linear regression. When dealing with non-linear relationships, it might not function properly.

Sensitive to Outliers: Unpredictable outcomes may arise from outliers' substantial impact on the model's coefficients and predictions.

Limited Complexity: Complex relationships, such as interactions between variables or non-linear patterns, are outside the scope of linear regression.

2.4) Use

There are many different uses for linear regression, such as:

Economics: Depicting how variables such as GDP and unemployment rate relate to one another.

Finance: Models for asset pricing and stock price prediction.

Medicine: Using medical indicators to forecast patient health outcomes.

Marketing: Examining how advertising expenditure affects sales.

2.5) Application

Predicting Home Prices: Determine a home's estimated cost by taking into account attributes like location, square footage, and number of bedrooms.

Sales forecasting: Project future sales using past performance information and promotional activities.

Risk assessment: Using information from credit and financial histories, determine the likelihood of a loan default.

3) LASSO Regression

A regularization method called Lasso Regression, which stands for "Least Absolute Shrinkage and Selection Operator," is used in linear regression to weed out unimportant features and avoid overfitting. By adding a penalty term that promotes feature selection and results in sparsity in the model's coefficients, it expands on the idea of linear regression.

3.1) Algorithm

Lasso regression adds an L1 regularization term to linear regression, making it resemble linear regression. The Lasso Regression objective function can be expressed as follows:

$$\sum_{i=1}^{n} (Y_i - B.X_i)^2 + \lambda.\sum_{j=1}^{p} |B_j| . L(B) \quad (3)$$

Where

With the loss function denoted by L(B).

Y_i stands for the intended variable.

The coefficient vector is denoted by B.

The input features for the i^{th} sample are represented by X_i.

The regularization parameter λ regulates the penalty term's strength. By picking a subset of significant features and reducing the coefficients of less relevant features to zero, the L1 penalty term $\lambda . . \sum_{j=1}^{p} |B_j|$ promotes sparsity in the coefficient vector.

3.2) Benefits

Choose Features: Because Lasso Regression tends to set the coefficients of redundant or unimportant features to zero, it is useful for feature selection. This can enhance interpretability and simplify the model.

Preventing Overfitting: Lasso's regularization term reduces the model's complexity, which makes it appropriate for situations involving high-dimensional data and helps prevent overfitting.

Interpretability: By emphasizing the most significant features, Lasso's sparsity introduces increased interpretability to the model.

In line with "Small n, Large p" Issues: Lasso works well in "small n, large p" problems, or scenarios in which there are many more features than data points.

3.3) Drawbacks

Unstable Coefficient Estimates: Lasso may produce unstable coefficient estimates since it is sensitive to even minute changes in the data.

Selection Bias: When significant variables are left out of the model, Lasso's feature selection procedure may result in selection bias.

Hyperparameter Tuning Difficulty: Choosing the right value for the regularization parameter λ can be difficult and may need cross-validation.

3.4) Use

Lasso There are several uses for regression, such as:

Economics: Calculating the influences on economic indicators, such as inflation.

Biology: Determining significant genes or characteristics from data on gene expression.

Finance: Developing risk assessment and asset pricing prediction models.

Image analysis by feature selection and denoising.

3.5) Application

Healthcare: Choosing the most pertinent features and forecasting patient outcomes using a range of medical markers.

Marketing: Examining how certain aspects of marketing campaigns affect consumer behavior and sales.

Climate science is the study of the elements that affect weather forecasts and climate models.

4) Random Forest Regression

This potent machine learning technique applies the idea of Decision trees to regression issues. It falls under the group of ensemble learning, which uses a number of Decision trees combined to produce reliable and accurate predictions.

4.1) Algorithm

An ensemble of Decision trees makes up a Random Forest Regression model. The algorithm operates in the following manner:

Bootstrap Sampling: Choose training data subsets at random using replacement. We refer to these selections as "bootstrap samples."

Decision Tree Construction: Create a Decision tree for every bootstrap sample. Nevertheless, only a random subset of features is taken into account (feature subspace sampling) at each node of the tree for determining the optimal split. This gives each individual tree more variation and randomness.

forecast: Every tree in the forest generates an output, and the ultimate forecast is the average of those outputs (for regression) or the majority vote (for classification).

Ensemble Aggregation: By pooling Decision trees, the risk of overfitting is minimized, and more reliable predictions are produced.

4.2) Benefits

High Accuracy: The robustness and high accuracy of Random Forest Regression are well-known. It is able to provide accurate forecasts by integrating several Decision trees.

Resistance to Overfitting: Random Forests are appropriate for complicated datasets with noise and outliers because of their ensemble nature, which minimizes overfitting.

The most significant features in the dataset can be found by using Random Forests to compute feature importance scores.

Non-stationarity Handling: Random Forests offer modelling versatility by capturing nonlinear interactions between features and the target variable.

4.3) Drawbacks

Complexity: Because Random Forests require a large number of Decision trees, they can be difficult to interpret and computationally demanding.

Lack of Transparency: It might be difficult to grasp the decisions made by individual trees, which makes it difficult to comprehend the logic behind the model.

Hyperparameter tuning: For the best results, Random Forests require careful adjustment of a variety of hyperparameters, including the number of trees and tree depth.

4.4) Use

Random Forest Regression is frequently employed in a number of fields, such as:

Finance: Risk assessment and stock price prediction.

Healthcare: Disease diagnosis and patient outcome prediction.

Climate modeling and environmental parameter prediction are aspects of environmental science.

Sales forecasting and recommender systems in e-commerce.

4.5) Application

Predicting Home Prices: Determine the cost of a home by factoring in amenities, location, and size.

Predict crop yields in agriculture by taking into account variables such as soil condition, weather, and farming techniques.

Energy Consumption: Estimate how much energy a facility will use depending on temperature, occupancy, and time of day.

5) Neural Network Regressor

Regression problems can be solved using neural network regression, a machine learning technique. It entails modelling and predicting continuous numerical values (such as real numbers) as the output using artificial neural networks. Neural network regression predicts a continuous range of values, as opposed to discrete categories as in classification problems. This makes it appropriate for problems involving price prediction, time series forecasting, and any other situation where you wish to estimate a real-valued output based on input features.

5.1) Algorithm

Enter Data: A dataset containing input features and associated continuous goal values is available to you.

Model Architecture: Create a neural network with the right architecture, which should include hidden layers in addition to input and output layers. The intricacy of the issue can determine how many neurons and layers are used.

Forward Propagation: Each neuron in the neural network computes its output using a predetermined activation function after the input data is passed through it. The anticipated continuous value is generated by the last output layer.

Loss Function: Determine the difference between the true target values and the forecasted values to compute a loss (error). Mean Absolute Error (MAE) and Mean Squared Error (MSE) are typical regression loss functions.

Backpropagation and Training: To minimize the loss, adjust the network's parameters (weights and biases) using the backpropagation method and optimization strategies like gradient descent. For many cycles, this process is performed iteratively.

Prediction: After been trained, the neural network may be used to new, unobserved data sets to provide predictions.

5.2) Benefits

Flexibility: Neural networks are appropriate for a variety of regression problems due to their ability to grasp intricate, non-linear relationships in the data.

Generalization: Neural networks with proper training may adapt well to new data.

Feature Engineering: They don't require as much feature engineering because they can automatically extract pertinent features from raw data.

Scalability: Neural networks are scalable, allowing them to handle complicated issues and big datasets.

5.3) Drawbacks

Complexity: The architecture and hyperparameter tuning of neural networks can be rather intricate, requiring meticulous attention to detail.

Overfitting: If a neural network is not appropriately regularized or if the dataset is small, it may overfit.

Computational Resources: Deep neural network training can be a computationally demanding process that calls for strong hardware.

5.4) Usage and Applications

Neural network regression is a tool used in a variety of predictive modeling applications, including demand forecasting, sales forecasting, and stock price prediction.

Finance: It is used in asset pricing, investment forecasting, and credit risk assessment financial modeling.

Healthcare: Predicting patient outcomes, the course of a disease, and the approximate cost of medical care may all be done with neural network regression.

Environmental modeling: It's used to forecast aspects of the environment such as the climate, air quality, and weather.

Economics: GDP forecasting, and inflation analysis are two applications of neural networks in economic forecasting.

Engineering: Applications include quality assurance, monitoring structure health, and anticipating equipment breakdowns.

Manufacturing and Supply Chain: Quality control, supply chain optimization, and demand forecasting are all done with it.

5.5) Used Functions

5.5.1) Propagation Forward

Weighted Sum = (Input * Weight) + Bias

5.5.1.1) Linear Activation Function

Since it does not introduce non-linearity, the linear activation function is an easy choice for regression. It can be used for applications where you want the network to predict a continuous value because it outputs the weighted sum of the inputs directly.

Linear (X) = X (4)

5.5.1.2) Scaled Exponential Linear Unit (SELU)

It is a useful tool for regression tasks; however, it is mainly recognized for its advantages in deep neural networks. It can hasten convergence and aid in normalizing activations.

SELU(x)=scale*(xif (x>0) else (α *(e^x - 1))) (5)

Scale and α are hyperparameters that require careful selection.

5.5.2) Backward Propagation

5.5.2.1) Gradient Descent

This technique essentially descends over the loss surface in the direction of a local or global minimum by iteratively changing the network's weights and biases in response to the computed gradients. The magnitude of the steps in this descent is determined by the learning rate (η).

New weight = Old weight - (η*($\frac{\delta J(w)}{\delta w}$)) (6)

6) Polynomial Regression

This type of regression technique represents the relationship as an nth degree polynomial between an independent variable (x) and a dependent variable (y). The formula for polynomial regression is as follows:

$Y = B_0 + B_1 * X_1 + B_2 * X_2 + B_3 * X_3 + \ldots\ldots + B_n * X_n$ (7)

In machine learning, it's also known as the multiple linear regression special case. In order to transform the Multiple Linear Regression equation into Polynomial Regression, we must add certain polynomial terms to it. It is a linear model that has been adjusted to improve accuracy. The polynomial regression training dataset is non-linear in character. To suit the intricate and nonlinear functions and datasets, it employs a linear regression model. Therefore, "In Polynomial regression, the original features are converted into Polynomial features of required degree (2, 3, …, n) and then modeled using a linear model."

6.1) Algorithm

Compile and Get Ready Data: Gather the dataset, which needs to contain the dependent variable as well as the independent variable or variables. Prepare your data by addressing any outliers and missing numbers as needed.

Decide on the Polynomial's Degree: Choose the polynomial whose degree (n) you wish to utilize. This establishes the maximum complexity of the curve. More complex curves can be created to a greater degree, although overfitting is also a possibility.

Feature Development: Raise the independent variable(s) to powers of '1' to 'n' to add more features. To generate new features for a degree 2 polynomial, you would use notation like X^2, X^3, etc.

Divide the Info: Make two sets of your dataset: a test set and a training set. The polynomial regression model is trained on the training set, and its performance is assessed on the test set.

Model Choice: Select the suitable model for polynomial regression. Typical options include specialized polynomial regression techniques or linear regression using polynomial features.

Model Instruction: Utilizing the training data, train the selected polynomial regression model. This entails determining each polynomial equation term's ideal coefficients.

Model Assessment: Utilize suitable measures, such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), or R-squared(R^2), to assess the model's performance on the test data.

Visualization of Models: To see how well the model matches the data, plot the data points beside the polynomial curve.

Adjusting: You can experiment with other polynomial degrees or regularization strategies to avoid overfitting if the model's performance is not up to par.

From Assumptions: You can use your model to forecast new data if you're satisfied with it.

Interpret Findings: To comprehend the link between the independent and dependent variables, interpret the coefficients and the polynomial equation.

Validation of Models: To make sure the model performs well and generalizes effectively to new data, cross-validation can be employed.

Normalization (Selective): Large coefficients in the polynomial equation can be penalized using regularization techniques like Ridge or Lasso regression if overfitting is an issue.

6.2) Benefits

The polynomial regression fits a wide range of curvatures because of its flexibility.

It is easily capable of housing a wide variety of functions.

The best approximation of the relationship between the two dependent and independent variables is provided by polynomial regression.

6.3) Cons

The outcome of the nonlinear analysis may suffer if there are one or more outliers in the data.

The outliers have a significant impact on polynomial regression.

Compared to the tools available for linear regression, there are far fewer model validation techniques available for nonlinear regression that aid in the detection of outliers.

6.4) Uses and Applications

This equation is utilized to generate the desired result in a variety of experimental approaches.

It gives the link between the independent and dependent variables a clear definition.

It is used to study the isotopes of sediments.

It is employed to investigate the emergence of various illnesses in any given community.

It is employed to investigate how any synthesis is created.

7) Logistic Regression:

7.1) Algorithm:

An approach for binary classification called logistic regression is used to forecast the likelihood that an instance will belong to a specific class. The logistic function, often known as the sigmoid function, is used to convert a linear combination of input data into a probability value between 0 and 1. The following is a summary of the algorithm:

$$P(Y = 1X) = \frac{1}{1 + e^{(B_0 + B_1X_1 + B_2X_2 + ... + B_nX_n)}}$$

Where:

P(Y=1|X) is the probability of the target variable Y being 1 given input features X.

B_0, B_1, B_2, ..., B_n are the coefficients to be learned during training.

X_1, X_2, ..., X_n are the input features.

7.2) Benefits:

Simple Interpretability: The outcomes of logistic regression are easily interpreted and make it simple to comprehend how each feature affects the likelihood that is predicted.

Efficient Training: It is computationally efficient and works well with a large number of features.

Probabilistic Predictions: Outputs probabilities, allowing for nuanced interpretation and decision-making.

7.3) Drawbacks:

Assumes Linearity: Logistic regression, like to linear regression, postulates a linear correlation between the input features and the target variable's log-odds.

Sensitive to Outliers: Outliers can have a significant impact on the learned coefficients.

Limited to Binary Classification: Extensions are required for multi-class classification because the original design was intended for binary classification problems.

7.4) Use and Application:

Medicine: Predicting the likelihood of a patient having a certain disease based on medical indicators.

Marketing: Predicting the probability of a customer making a purchase based on historical data.

Credit Scoring: Assessing the risk of default in credit applications.

Customer Churn Prediction: Predicting whether a customer is likely to churn based on usage patterns and customer demographics.

Fraud Detection: Identifying potentially fraudulent transactions based on transaction history and user behaviour.

7.5) Used Functions:

Sigmoid Function (Logistic Function): The logistic function transforms the linear combination of input features into a range between 0 and 1. It is defined as:

$$P(Y=1X) = \frac{1}{1+e^{-x}}$$

8) Ridge Regression:

8.1) Algorithm:

Ridge Regression is a linear regression approach that adds a regularization component to the ordinary least squares (OLS) objective function. It is sometimes referred to as Tikhonov regularization or L2 regularization. It penalizes big coefficients in an attempt to prevent overfitting. The objective function of the Ridge Regression can be written as follows:

$$J(B) = MSE(B) + \lambda \sum_{j=1}^{p} B_j^2$$

Where:

J(B) is the Ridge Regression objective function.

MSE(B) is the Mean Squared Error, measuring the difference between predicted and actual values.

B is the vector of coefficients.

λ is the regularization parameter, controlling the strength of the penalty term.

8.2) Benefits:

Handles Multicollinearity: By reducing coefficients, Ridge Regression works well when there is multicollinearity, or a high degree of correlation between features.

Prevents Overfitting: The regularization term penalizes large coefficients, reducing the risk of overfitting.

Works Well for High-Dimensional Data: Suitable for datasets with more features than observations.

8.3) Drawbacks:

Not Sparse Solution: Ridge Regression does not lead to a sparse solution, meaning it does not eliminate irrelevant features but rather shrinks their coefficients.

Difficulty in Interpretation: The interpretation of coefficients becomes more challenging due to the regularization term.

8.4) Use and Application:

Economics: Predicting economic indicators while handling correlated features.

Finance: Asset pricing models where multicollinearity is a concern.

Stock Price Prediction: Predicting stock prices by considering various financial indicators.

Climate Modelling: Handling correlated environmental parameters in climate models.

Gene Expression Analysis: Dealing with high-dimensional gene expression data in biology.

Sr.No	Name of Algorithm	Equations	Advantages	Disadvantages	Application
1	Decision tree Regression	$MSE = \frac{1}{N}\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$	Simplicity, Managing Non-linearity, Can handle mixed data, Feature Importance	Overfitting, Instability, Lack of global optimality	Housing Price Prediction, Forecasting Energy Consumption.
2	Linear regression	$Y = B_0 + B_1 * X_1 + B_2 * X_2 + ... + B_n * X_n$	Simplicity, Interpretable Coefficients, Efficiency	Assumption of Linearity, Sensitive to Outliers, Limited Complexity	Predicting Home Prices, Sales forecasting, Risk assessment, Finance, Medicine

3	LASSO regression	$\sum_{i=1}^{n} (Y_i - B.X_i)^2 + \lambda.\sum_{j=1}^{p} \|B_j\|.L(B)$	Choose Features, Preventing Overfitting, Interpretability	Unstable Coefficient Estimates, Selection Bias, Hyperparameter Tuning Difficulty	Economics, Finance, Healthcare, Marketing,
4	Random forest regression	$Y = \frac{1}{N}\sum_{i=1}^{N} fi(\mathbf{x})$	High Accuracy, Resistance to Overfitting, Non-stationarity Handling	Complexity, Lack of Transparency, Hyperparameter tuning	Finance, Predicting Home Prices, Predict crop yields in agriculture , Energy Consumption
5	Neural network regression	SELU(x)=scale*(xif (x>0) else (α *(e^x - 1)))	Flexibility, Generalization, Scalability	Complexity, Overfitting, Computational Resources	Environmental modelling, Engineering, Manufacturing and Supply Chain
6	Polynomial regression	$Y = B_0 + B_1 * X_1 + B_2 * X_2 + B_3 * X_3 + \ldots\ldots + B_n * X_n$	Best approximation, Polynomial regression fits a wide range	Sensitivity to Outliers, Limited Model Validation Techniques	Overfitting, Computational Complexity, Sensitivity to Outliers
7	Logistic regression	$P(Y=1X) = \frac{1}{1+e^{-x}}$	Simple Interpretability, Efficient Training, Probabilistic Predictions	Assumes Linearity, Sensitive to Outliers, Limited to Binary Classification	Medicine, Marketing, Credit Scoring, Fraud Detection
8	Ridge regression	$J(B) = MSE(B) + \lambda \sum_{j=1}^{p} B_j^2$	Handles Multicollinearity, Prevents Overfitting	Not Sparse Solution, Difficulty in Interpretation	Stock Price Prediction, Climate Modelling, Gene Expression Analysis

Table 3.1 Comparison Table

4 Conclusion

In conclusion, our research endeavors to shed light on the comparative performance of various regression algorithms, namely Ridge, Decision Tree, Random Forest, Neural Network, Logistic, Linear, Polynomial, and LASSO. Through a systematic evaluation and analysis, we have gained valuable insights into the strengths and limitations of each algorithm in addressing different types of regression problems. Following a thorough examination and contrast of various algorithms, we have determined some important findings:

LASSO regression, with its feature selection capabilities through sparsity-inducing regularization, proves beneficial in situations where a subset of features significantly contributes to the predictive accuracy.

Linear and polynomial regression, while relatively simple, offer interpretability and ease of implementation. Linear regression is particularly useful when the underlying relationship between variables is linear, while polynomial regression accommodates non-linear trends by introducing higher-degree polynomial terms.

Decision tree regression proved useful for identifying interactions and non-linear correlations within the feature set. Educational institutions can benefit from the interpretability and simplicity of this algorithm's implementation.

Random forest regression using the strength of an ensemble of decision trees has stood out as a dependable and accurate option. Its capacity to handle complex data, resistance to overfitting, and robustness make it a good choice.

Ridge regression, with its regularization parameter, exhibits robustness against multicollinearity and helps prevent overfitting, making it a suitable choice for datasets with high-dimensional features.

Neural network regression, with its ability to model complex relationships, demonstrates exceptional performance, particularly in large-scale datasets. However, it may require careful tuning and consideration of computational resources.

Logistic regression proves to be effective in binary classification tasks, providing a straightforward yet powerful approach for scenarios where the output is categorical.

In conclusion, the choice of a regression algorithm should be guided by the specific characteristics and requirements of the dataset at hand. Researchers and practitioners must consider factors such as dataset size, linearity, complexity, and interpretability when selecting the most suitable algorithm for their regression tasks. Moreover, a combination of algorithms or ensemble methods, such as Random Forest, may offer improved predictive performance by leveraging the strengths of individual models. Our study contributes to the broader understanding of regression algorithms and provides valuable insights that can aid researchers and practitioners in making informed decisions based on the specific nuances of their regression problems. Further exploration and experimentation in diverse domains can enhance our understanding and contribute to the continuous refinement of regression modelling techniques.

4 References

1. Kinaneva, G. Hristov, P. Kyuchukov, G. Georgiev, P. Zahariev and R. Daskalov, "Machine Learning Algorithms for Regression Analysis and Predictions of Numerical Data," 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2021, pp. 1-6, doi: 10.1109/HORA52670.2021.9461298.
2. T. Doan and J. Kalita, "Selecting Machine Learning Algorithms Using Regression Models," 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 2015, pp. 1498-1505, doi: 10.1109/ICDMW.2015.43.
3. E. Gambhir, R. Jain, A. Gupta and U. Tomer, "Regression Analysis of COVID-19 using Machine Learning Algorithms," 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2020, pp. 65-71, doi: 10.1109/ICOSEC49089.2020.9215356.
4. X. Qu, F. Zhao, L. Gao and Z. Zhang, "The application of machine learning regression algorithms and feature engineering in practical application," 2022 10th International Conference on

Information Systems and Computing Technology (ISCTech), Guilin, China, 2022, pp. 259-263, doi: 10.1109/ISCTech58360.2022.00048.

5. P. A. Schirmer, I. Mporas and M. Paraskevas, "Evaluation of Regression Algorithms and Features on the Energy Disaggregation Task," 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), Patras, Greece, 2019, pp. 1-4, doi: 10.1109/IISA.2019.8900695.