Machine Learning

Regression and Its Application

Learning Objectives

By the end of this lesson, you will be able to:

- Analyze the different types of regression models
- Examine linear regression and prepare data for linear regression
- Discover linearity between variables and plot a correlation map
- Analyze the train and test a linear regression model

Learning Objectives

By the end of this lesson, you will be able to:

- Implement a logistic regression model and calculate its outcome and accuracy
- Implement polynomial, ridge, and lasso regression models

Business Scenario

A retail company aims to enhance its holiday season sales forecasting by utilizing machine learning techniques. It has historical sales data from previous years and plans to use regression analysis to create a model that estimates sales for each day of the Christmas season, taking into account factors such as promotions, time of year, and special events.

The company also plans to utilize multiple regression models, such as a ridge, polynomial, and linear regression, to identify the most accurate forecasting model. The model's assumptions will be validated by looking for missing values, outliers, and multicollinearity.

Accurately forecasting sales will enable the company to plan appropriate marketing campaigns, optimize inventory levels, prepare for the holiday rush by ensuring adequate staffing, identify sales trends, and refine sales strategies.

What Is Regression?

Discussion: Regression

Duration: 10 minutes

- What is regression, and what are the different types?
- How does ridge and lasso regression control the overfitting problem?

Regrēssion

Regression is a supervised machine-learning technique that is used to predict a continuous value based on a set of input variables.

It establishes a relationship between a dependent variable (x) and an independent variable (y).

It is the easiest and one of the most widely used machine learning algorithms.

Regression Analysis

In regression analysis, the dependent variable is the variable being predicted or explained, while the independent variables are the predictors or explanatory variables.

It predicts continuous or real values, such as:

Temperature

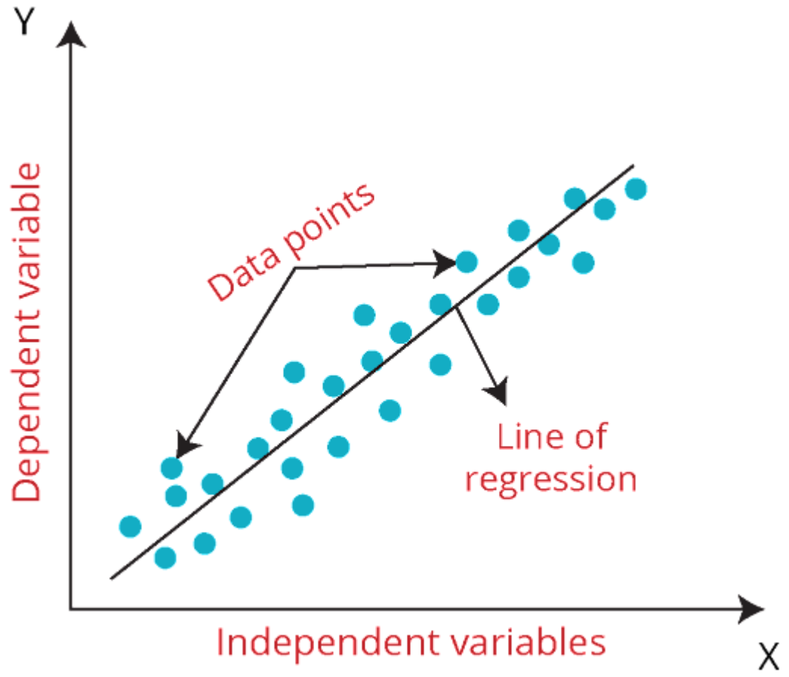
Age

Salary

Price

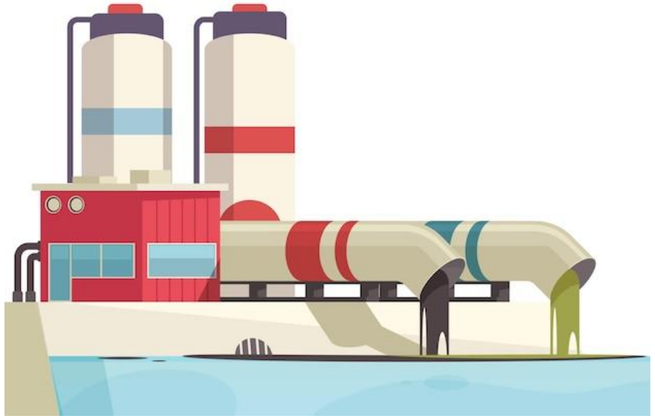
Regression Algorithms

A regression algorithm plots a best-fit line or a curve between the data.



Application of Regression: Oil and Gas Industry

Various types of data are collected in the oil and gas industry from the surface and the subsurface to understand the production and sale processes.

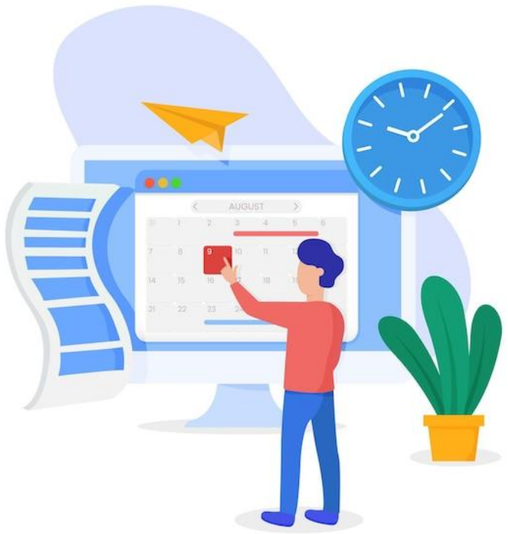


- Linear and nonlinear regression models are used to forecast global oil production.

Regression analysis was used to influence factors on the future economy of crude oil.

Application of Regression: Predicting Events

Identifying the impact of marketing campaigns



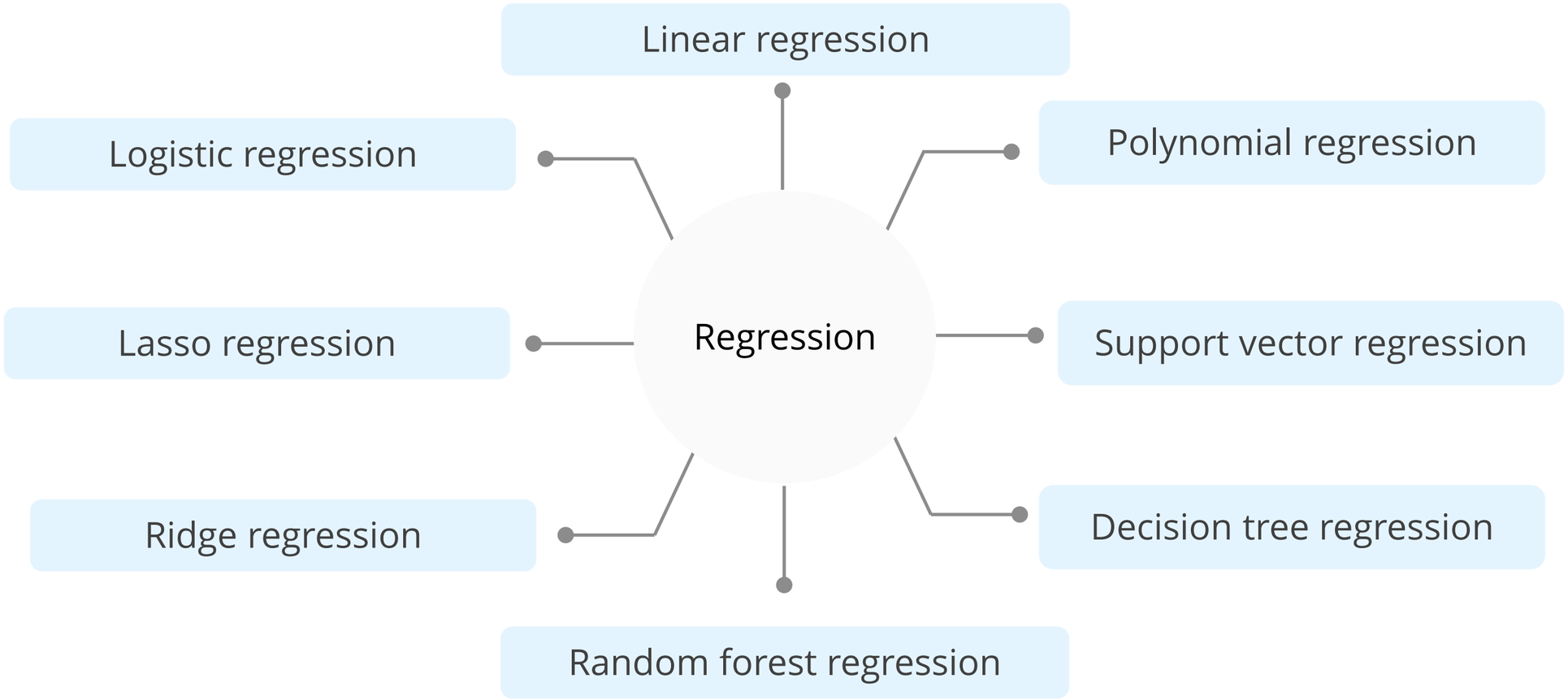
Example

Regression analysis can be used to optimize the budget for an advertising campaign by analyzing the relationship between the budget and the outcome. This information can be used to ensure that the campaign is getting the most value for money.

Regression Types

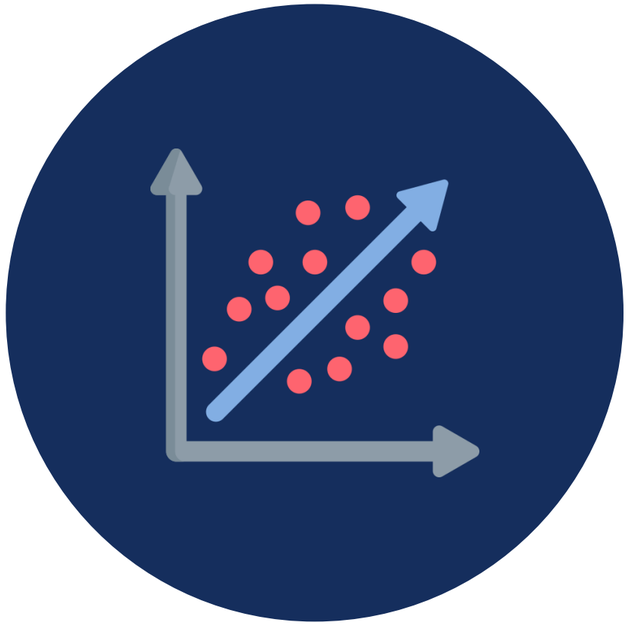
Types of Regression

Here are the most common types of regressions that are used in machine learning:



Linear Regression

Linear regression is a popular modeling technique.



- It is used to predict a continuous dependent variable based on an independent variable.
- It uses the least square criterion for estimation.
- It can be applied only if there is a linear relationship between the variables.

Polynomial Regression

Polynomial regression is a subset of linear regression that includes polynomial terms.

The relationship between an independent variable (x) and a dependent variable (y) is modeled as an nth degree polynomial.

Support Vector Regression

Support Vector Regression (SVR) is a supervised learning algorithm that extends support vector machines to regression problems.

Supervised learning algorithms use a subset of the data called support vectors to find a hyperplane that separates the data into two or more classes.

It can be used for both regression and classification problems.

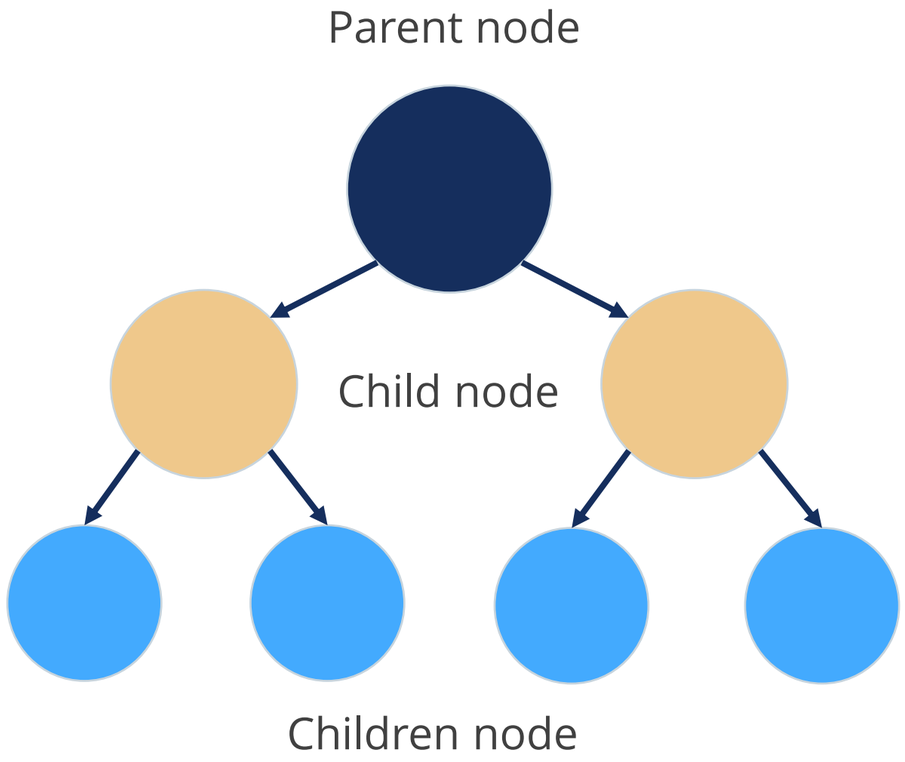
Decision Tree Regression

It is a commonly used supervised learning approach and it builds a tree-like structure.

- Internal nodes represent the test for an attribute.
- Branches represent the test results.
- Leaf nodes represent the final result or decision.

It can be used for both classification and regression problems.

Decision Tree Regression



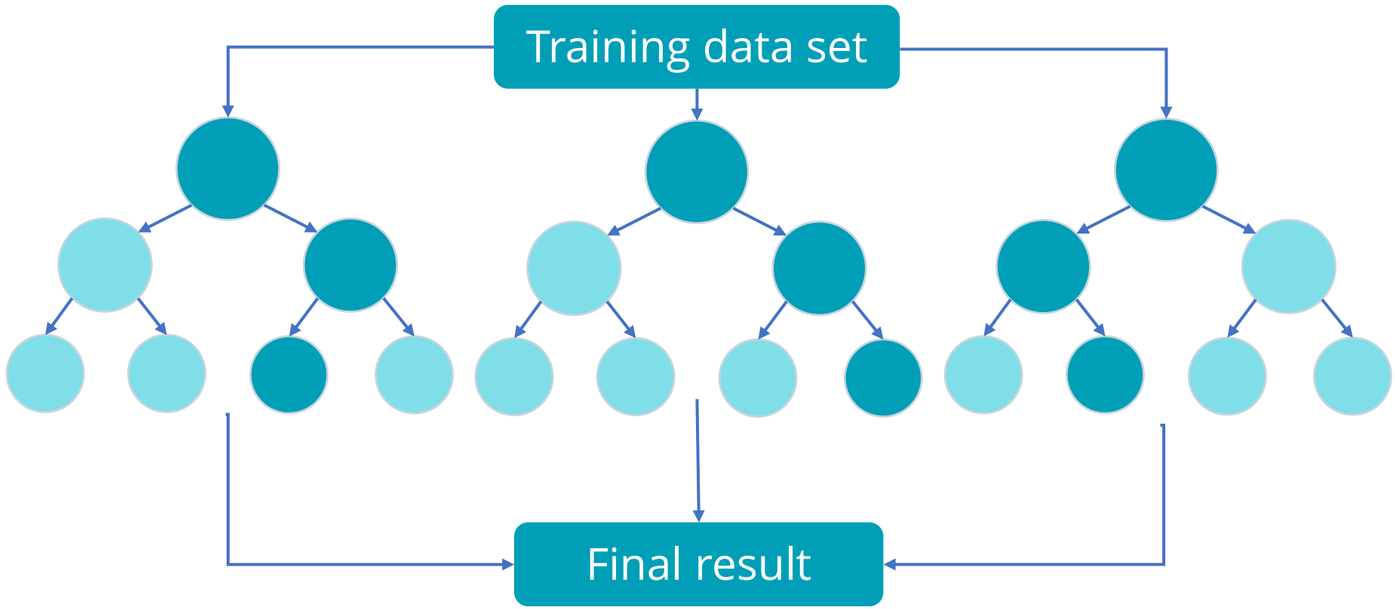
It starts from a root node (parent node) of a data set.

It then splits into left and right nodes (or subsets of the dataset), which are called child nodes.

These child nodes are further divided into children's nodes and become the parent node of these children's nodes.

Random Forest Regression

It is a form of linear regression algorithm that performs both regularization and variable selection.

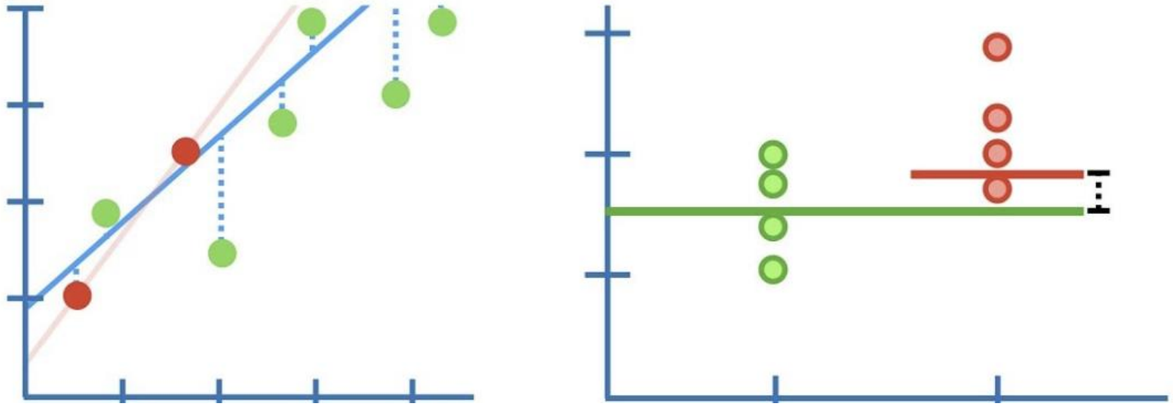


This method combines multiple decision trees to predict the final output.

The aggregated decision trees run in parallel and do not interact with each other.

Ridge Regression

Ridge regression is a regularization technique used in linear regression to prevent overfitting of the model. It is used when dealing with multicollinear data.



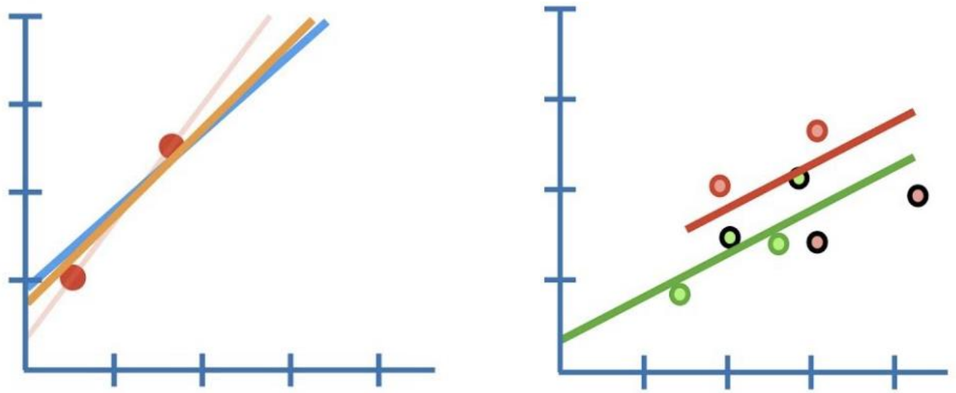
Least squares are unbiased, and variances are large.

The predicted values will be far away from the actual values.

It adds a regularization term to the loss function that penalizes large coefficients in the model and encourages the model to have smaller coefficients. It results in a simpler and more generalized model.

Lasso Regression

It is also a form of linear regression.



It is an acronym for Least Absolute Shrinkage and Selection Operator.

It uses shrinkage and performs variable selection or feature selection.

Discussion: Regression

Duration: 10 minutes

- What is regression, and what are the different types?

Answer: Regression is a supervised machine-learning technique that is used to predict a continuous value based on a set of input variables. The most common types of regression are linear regression, logistic regression, lasso regression, ridge regression, random forest regression, polynomial regression, support vector regression, and decision tree regression.

- How does ridge and lasso regression control the overfitting problem?

Answer: Ridge regression controls overfitting by adding a penalty term to the loss function, which shrinks the coefficients toward zero. Lasso regression further mitigates overfitting by not only shrinking coefficients but also performing feature selection by setting some coefficients exactly to zero.

Linear Regression

Discussion: Linear Regression

Duration: 10 minutes

- What is linear regression?
- How does multicollinearity impact regression analysis?

Linear Regression

In this the dependent variable is continuous and the independent variables can be either continuous or discrete.

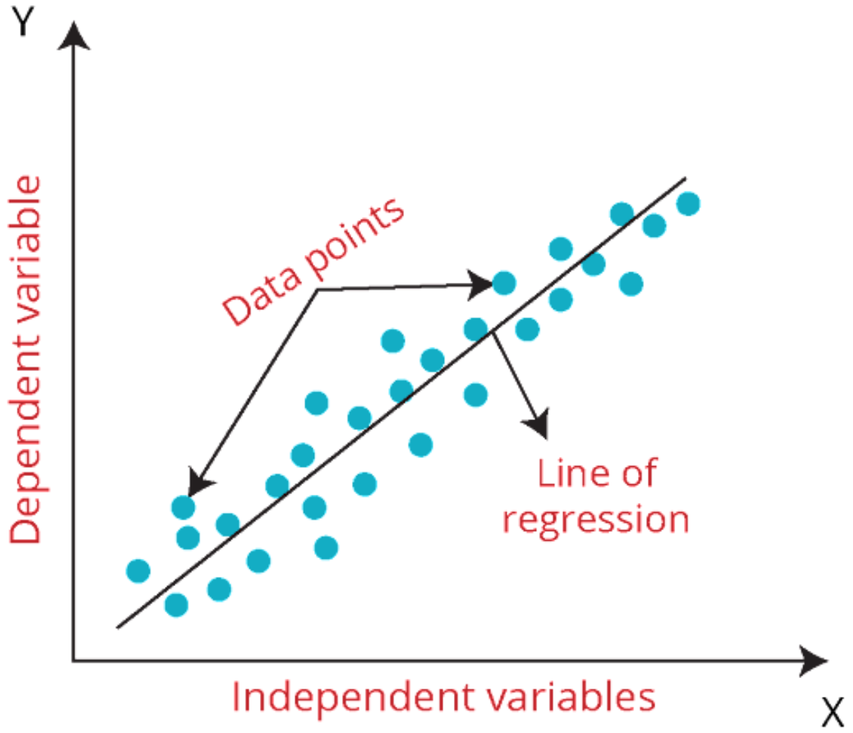
The relationship between a dependent variable (y) and one or more independent variables (x) is established using a best-fit straight line, also known as the regression line.

The nature of the regression line is linear in this technique.

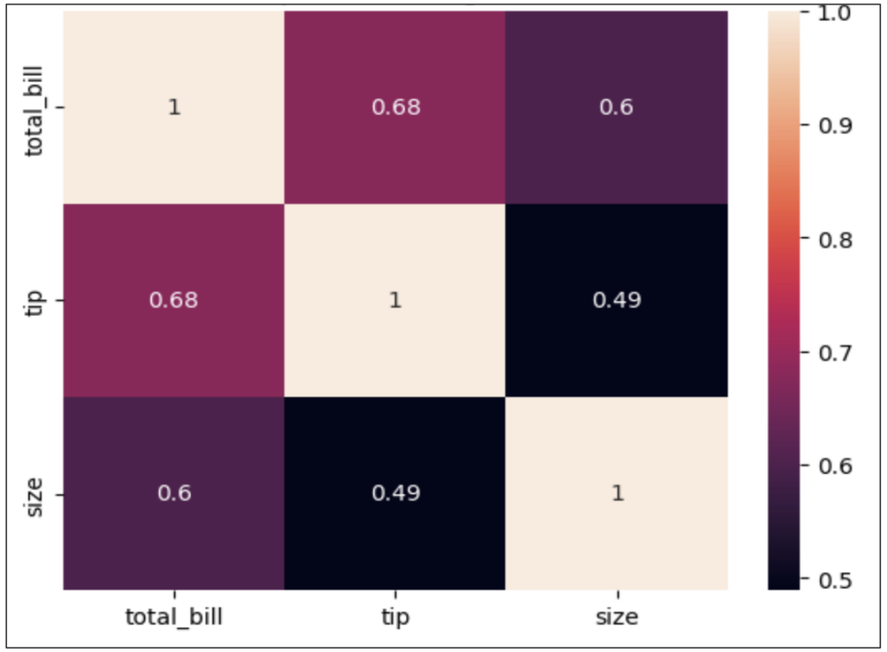
Linear Regression

There are two types of linear regression:

Simple linear regression

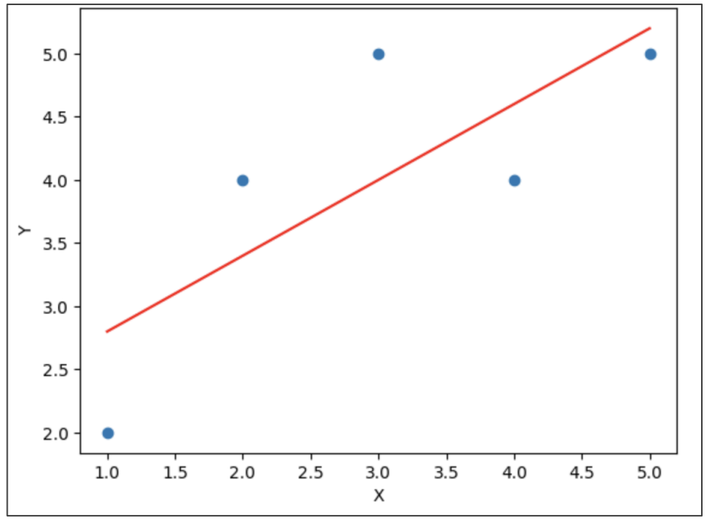


Multiple linear regression



Simple Linear Regression

In a simple linear regression, the value of a numerical dependent variable is predicted using a single independent variable.

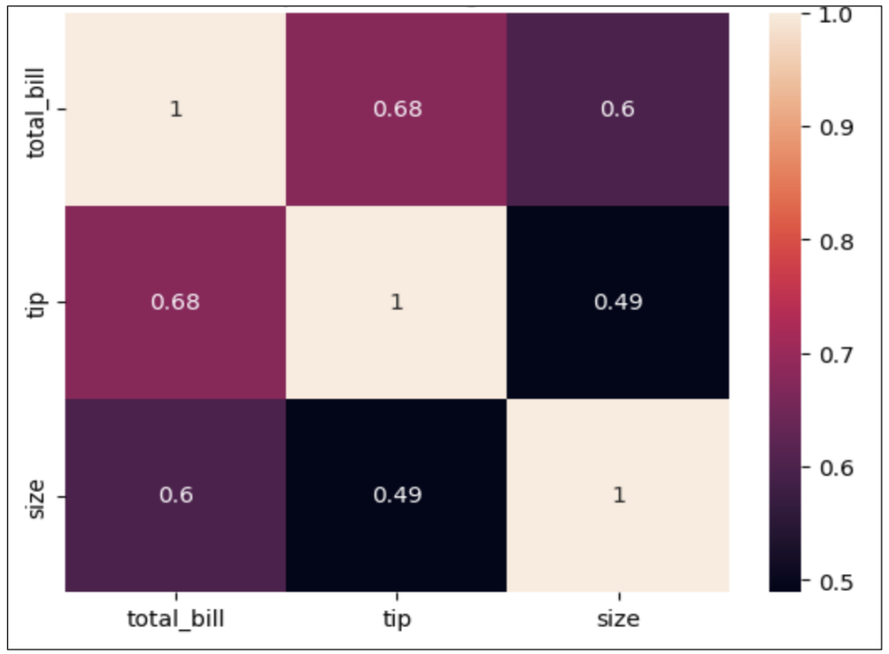


The equation used in simple linear regression is:

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression

It uses more than one independent variable to predict the value of a numerical dependent variable.



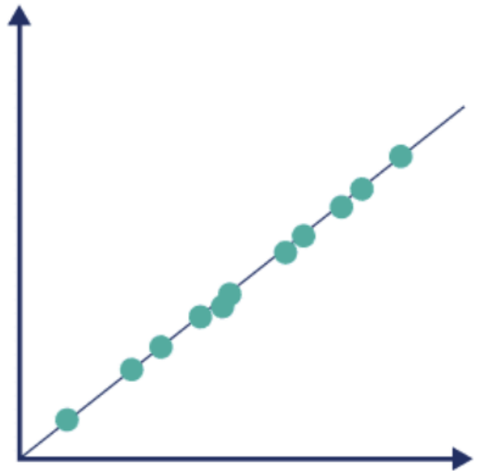
The equation used in multiple linear regression is:

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

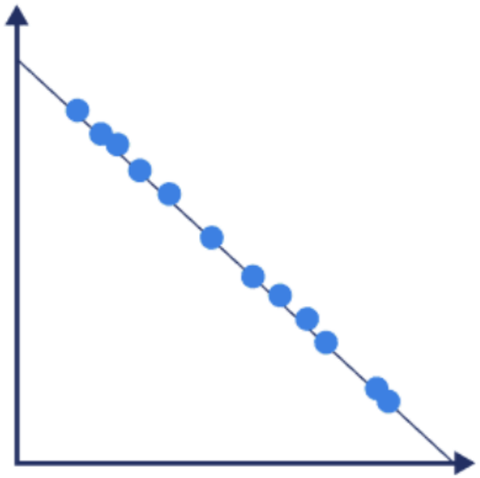
Linear Regression Line

The line showing the relationship between dependent and independent variables is called a regression line.

A regression line can depict two types of relationship:



Positive linear relationship: The dependent variable increases on the Y-axis and the independent variable increases on the X-axis.



Negative linear relationship: The dependent variable decreases on the Y-axis and the independent variable increases on the X-axis.

Applications of Linear Regression

Linear regression is used to administer the sulfur cap rule by the International Maritime Organization.

Problem

The problem is using linear regression to analyze the relationship between factors (fuel type, engine size, vessel age, and distance traveled) and sulfur emissions in the maritime industry to comply with the sulfur cap rule and reduce emissions effectively.

Solution

By applying linear regression, the shipping company identifies that engine size significantly impacts sulfur emissions, enabling it to prioritize retrofitting larger ships with more efficient engines to comply with the sulfur cap rule and reduce emissions.

Machine learning is used to measure how much sulfur is being released in the exhaust and what businesses can do to reduce the sulfur content.

Critical Assumptions for Linear Regression

Linear Regression

Regression is a parametric approach where assumptions about data are made for the purpose of the analysis.

This makes regression restrictive.

Successful regression analysis requires validation of the assumptions made.

Important Assumptions in Linear Regression

The dependent (response) variables and the independent (predictor) variables have a linear and additive relationship.

An additive relationship implies that the effect of x on the response y is independent of other variables.

For a predictor variable (x) and a response variable (y), a linear relationship implies that a change in response y due to one unit change in x remains constant, regardless of the value of x.

Important Assumptions in Linear Regression

Some other important assumptions in regression analysis are:

- The error terms are normally distributed.
- The independent variables are not correlated.
- Error or residual terms are not correlated.
- The error terms have constant variance.

Assisted Practices

Let’s understand the topic below using Jupyter Notebook.

- 4.5_Working with Linear Regression

Note: Please download the pdf files for each topic mentioned above from the Reference Material section.

ASSISTED PRACTICE

Discussion: Linear Regression

Duration: 10 minutes

- What is linear regression?

Answer: In this the dependent variable is continuous and the independent variables can be either continuous or discrete.

- How does multicollinearity impact regression analysis?

Answer: Multicollinearity in regression analysis can cause instability and unreliable estimates of the regression coefficients. It makes it difficult to determine the individual effects of predictor variables, increases standard errors, and can lead to misleading interpretations of the relationships between variables.

Logistic Regression

Discussion: Logistic Regression

Duration: 10 minutes

- Which algorithm works better in the presence of outliers, SVM or logistic regression?
- Discuss the odds ratio specific to logistic regression.

Logistic Regression

It is a statistical analysis method used to predict a data value based on prior observations of a data set.

It finds the relationship between qualitative variables and independent variables.

It is a machine learning method used to distinguish one class from another.

Logistic Regression

Example: In a binary classification, for a given question, the algorithm sorts the answers into:

A set of positive points if the answer is positive

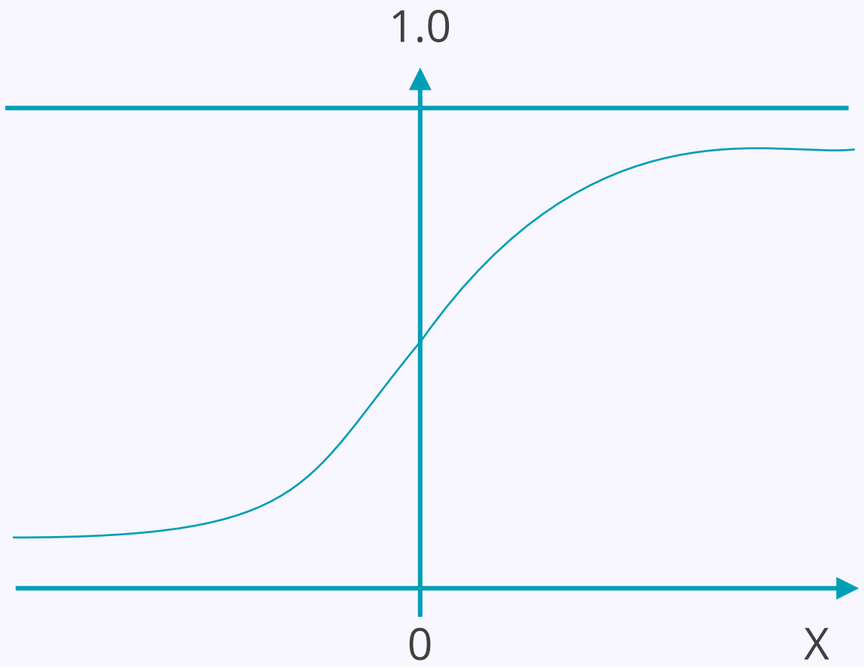
A set of negative points if the answer is negative

The main goal of logistic regression is to accurately predict the class of the two possible label data points.

A logistic regression model ensures that the output always falls between 0 and 1 as used in Sigmoid function.

Sigmoid Function

Below is the formula of a sigmoid function:



$$f(x) = \frac{1}{1 + e^{-x}}$$

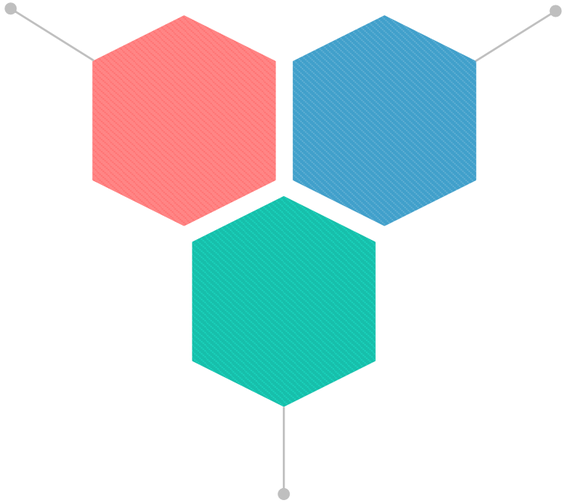
It is used to map predictions to probabilities.

It maps any real value to the other value between 0 and 1.

Types of Logistic Regressions

Binary Logistics Regression: 2 variables

The response variable can only belong to one of two categories.



Multinomial Logistic Regression: 3 or more variables

The response variable can belong to one of three or more categories and there is no natural ordering among the categories.

Ordinal Logistic Regression: 3 or more ordinal variables

The response variable can belong to one of three or more categories and there is a natural ordering among the categories.

Decision Boundary

It is a threshold that separates the predicted classes based on the predicted probability of an observation belonging to a particular class.

The probability score computed using the algorithm is between 0 and 1, based on the inputs provided.

If the probability score is more than 0.5, it will be labeled as class 1, else as 0.

Example: Consider two products with one being eco-friendly and the other being toxic. Use the threshold value to classify the given product.

Applications of Logistic Regression

Logistic regression can be used for:

Prediction of a baseball match outcome

Assessment of the presence or absence of fraud based on banking activity

Assessment of erosion of the seabed due to fishing

Prediction of a marine vessel crash based on tidal energy and icebergs

Prediction of bankruptcy

Discussion: Logistic Regression

Duration: 10 minutes

- Which algorithm works better in the presence of outliers, SVM or logistic regression?

Answer: SVM (Support Vector Machines) generally works better in the presence of outliers compared to logistic regression. SVM is less sensitive to outliers due to the use of support vectors, which prioritize correctly classifying data points near the decision boundary, while logistic regression is more influenced by outliers as it minimizes the sum of squared errors.

- Discuss the odds ratio specific to logistic regression.

Answer: In logistic regression, the odds ratio represents the change in odds of the dependent variable for a one-unit increase in the independent variable. It quantifies the strength and direction of the relationship between the independent variable and the probability of the dependent variable, with an odds ratio greater than 1 indicating a positive association and an odds ratio less than 1 indicating a negative association.

Assisted Practices

Let’s understand the topic below using Jupyter Notebook.

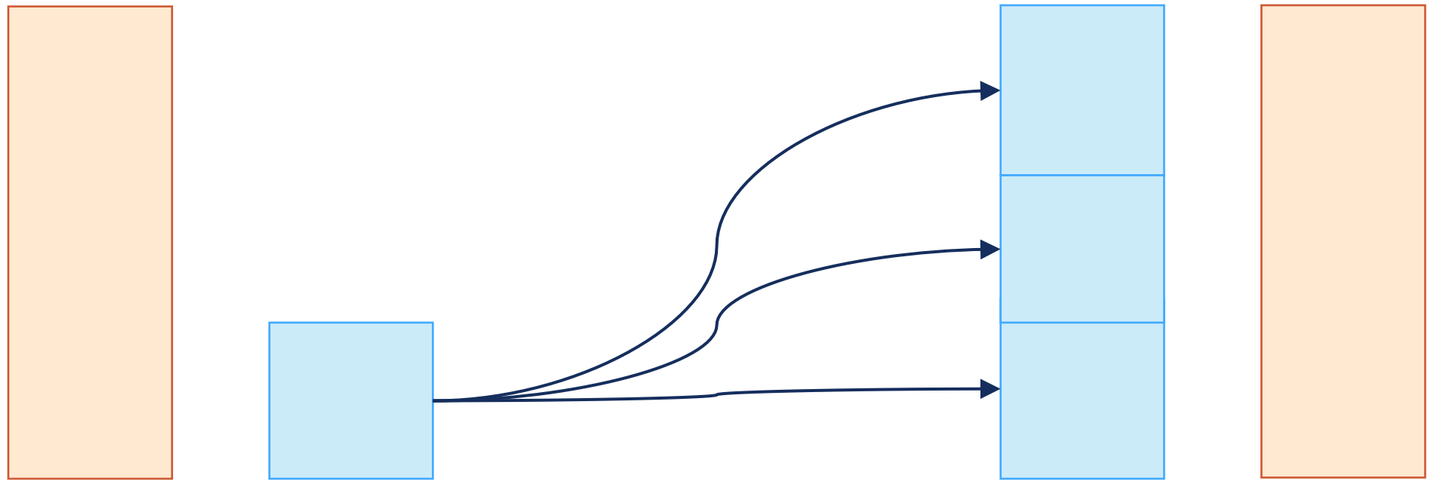
- 4.8_Data Exploration Using SMOTE

Note: Please download the pdf files for each topic mentioned above from the Reference Material section.

Oversampling using SMOTE

Imbalanced Data Set

It is a dataset where its classes are not distributed equally. This can occur when one class has a lot more instances than the other.

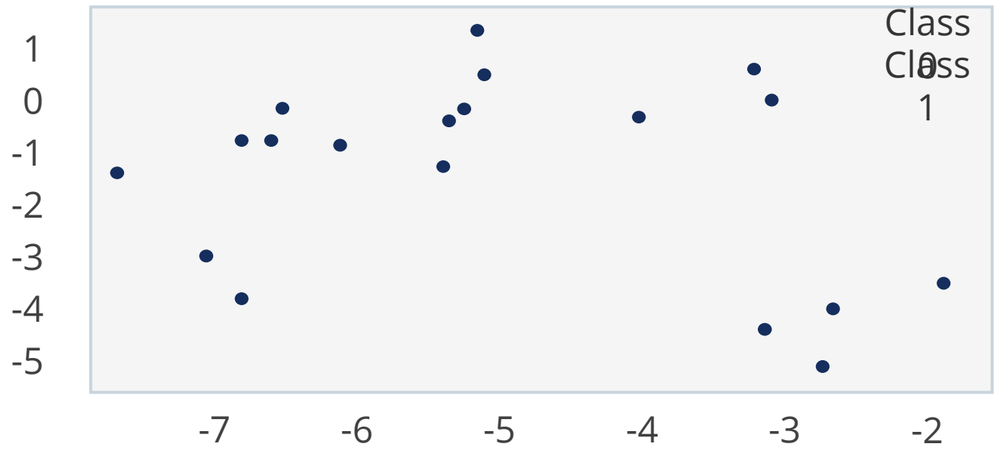


To address the data bias between classes, oversampling can be performed using a technique called Synthetic Minority Oversampling Technique (SMOTE).

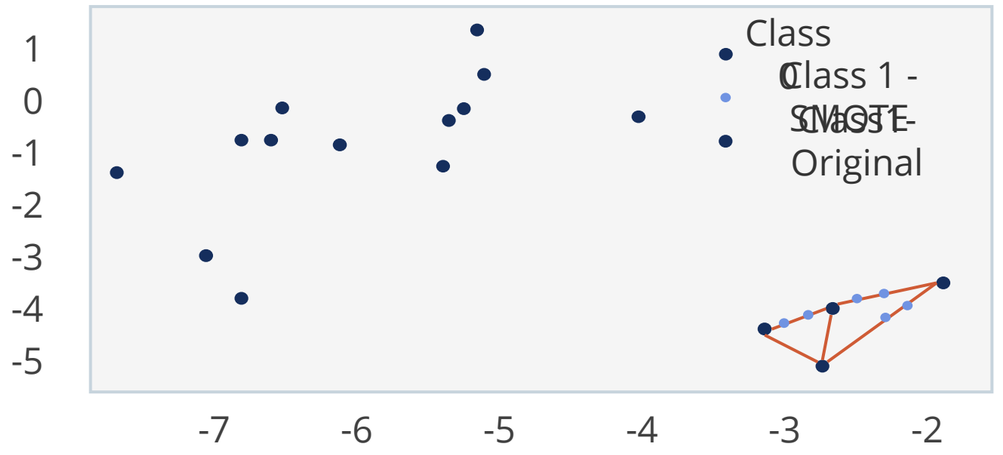
Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is a data augmentation technique that helps balance data distribution by adding to the minority class.

Consider the following visual representation of a data set:



Class 1 is a minority class with a lower number of observations.



SMOTE joins the points with the line segments, places artificial points along the lines balances data distribution.

SMOTE Algorithm

The working of the SMOTE algorithm is a four-step process:

1. Select an input vector from the minority class
2. Find its K-nearest neighbors
3. Build a line joining the point under consideration and the chosen neighbor
4. Place a synthetic point anywhere on the line drawn

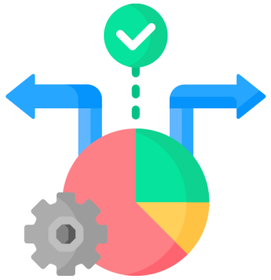
Repeat steps 1 to 4 until data is balanced.

SMOTE Applications

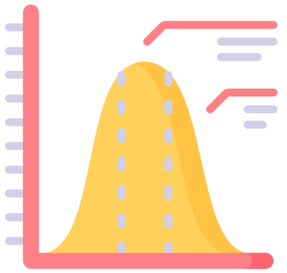
Applications of SMOTE include:

Balancing data for classification problems where one class is more than another

Classifying job roles or individuals based on the agility requirements of teams where teams are usually unbalanced data sets



Increasing decision boundaries

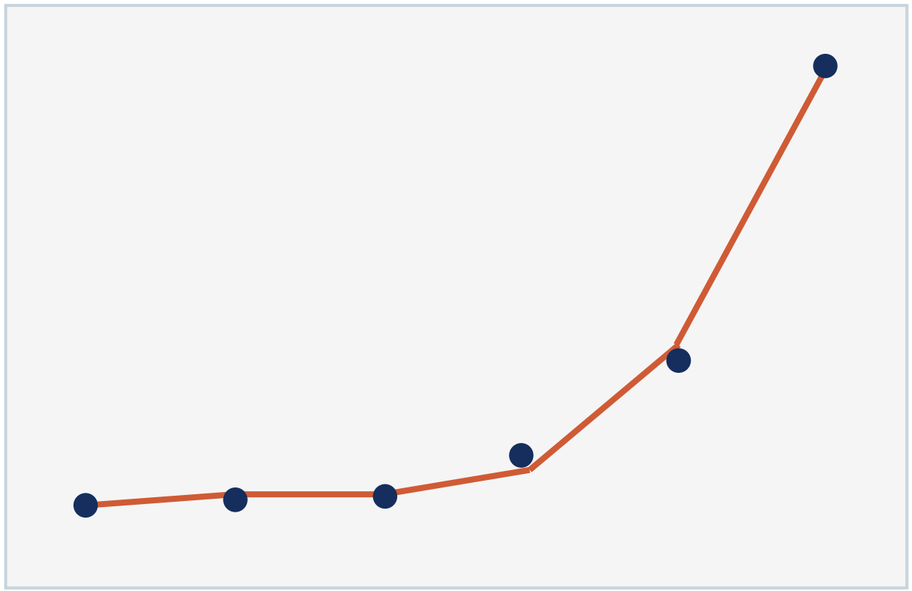


Improving classification performance by enhancing the probability distribution of data sets

Polynomial Regression

Polynomial Regression

The linear regression technique represents the connection between dependent and independent variables in a linear fashion and so cannot be utilized for complex data.



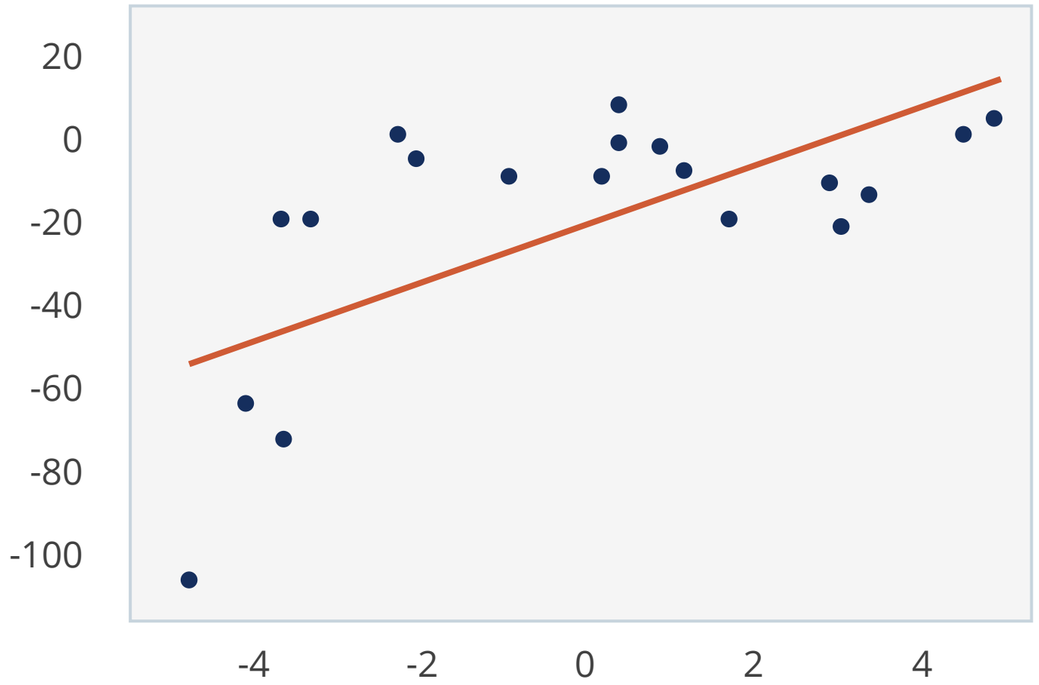
Polynomial regression is a modified linear regression for complex, nonlinear data.

It is a statistical method used in machine learning for predictive modeling and analysis.

It models the nonlinear relationship that exists between a dependent variable (y) and an independent variable (x) as an nth degree polynomial.

Polynomial Regression

From the below figure, it is seen that the straight line is unable to capture the patterns in the given data.



Polynomial Regression

To achieve a higher-order equation that captures complex data patterns, powers of the original features can be added as new features.

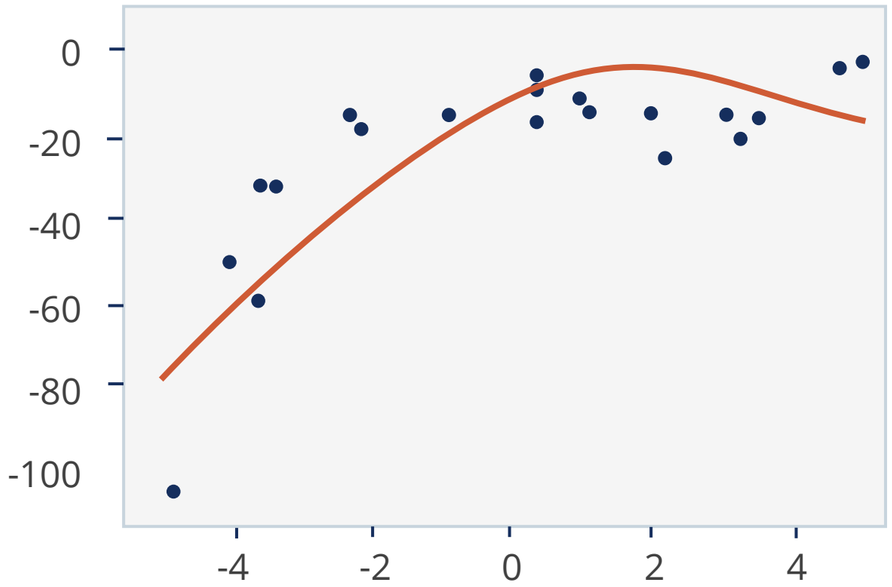
Simple linear regression: $y = b_0 + b_1 * x_1$

Multiple linear regression: $y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$

Polynomial linear regression: $y = b_0 + b_1x_1 + b_2x_1^2 + \ldots + b_nx_1^n$

Polynomial Regression

It is considered as a linear model due to the linear nature of coefficients or weights associated with the features.

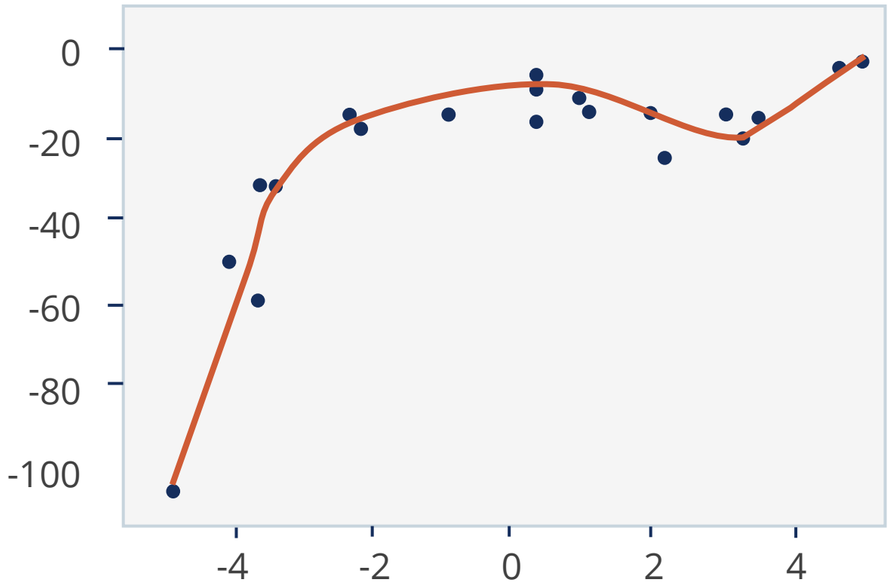


x^2 is only a feature, but the curve fitting is quadratic. It doesn't capture some data points, as can be seen in the figure.

The polynomial features class provided by scikit-learn is used to capture more data points and convert the original features into their high-order.

Polynomial Regression

The following graph is the result after applying polynomial regression:



It provides the best approximation of the relation between a dependent and an independent variable.

It captures the most data points and decreases the error between the actual data point and the predicted data point.

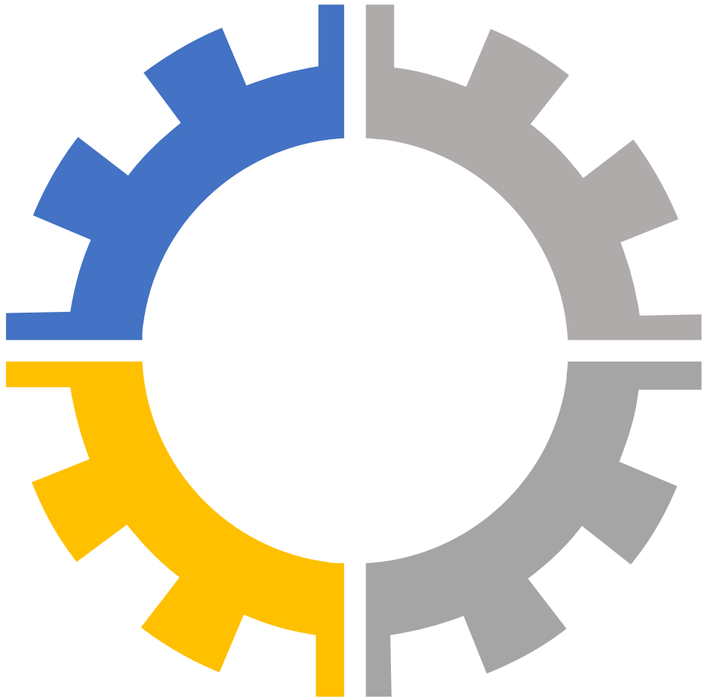
Hence, the model is now trained properly.

Applications of Polynomial Regression

Polynomial regression can be applied to:

Examine the generation of any synthesis

Develop climate model predictions that run faster and more efficiently



Study the isotopes present in sediments

Predict the rise of different diseases within populations and their spread rates

Assisted Practices

Let’s understand the topic below using Jupyter Notebook.

- 4.11_Data Preparation, Model Building, and Performance Evaluation: Part A

Note: Please download the pdf files for each topic mentioned above from the Reference Material section.

Ridge Regression

Ridge Regression

It is a form of regression that shrinks the coefficient toward zero to reduce the complexity of the data.

It prevents the use of complex models and overfitting.

It decreases the multicollinearity between features in the data set.

It minimizes the variance of the model without increasing its bias.

Ridge regression involves a loss function called the residual sum of squares or RSS.

Ridge Regression

RSS is the difference between the actual and predicted values.

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j xij \right)^2$$

Consider an example:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Y is the independent variable
β is the regression coefficient value
X is the dependent variable

Ridge Regression

λ is the constant or the tuning parameter that decides the appropriate rate to penalize the flexibility of the model.

$$Cost\ (W) = RSS(W) + \ \lambda * (sum\ of\ squares\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=1}^{M} w_j x_{ij} \right\}^2 + \lambda \sum_{j=1}^{M} {w_j}^2$$

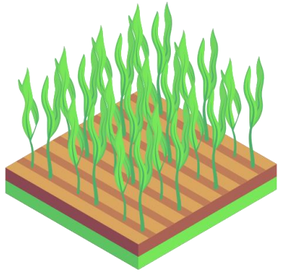
- When λ = 0, the penalty term has no effect as penalty is a multiple of λ and the sum of square of weights.
- When λ→∞, the impact of the shrinkage penalty grows the ridge regression coefficient estimates will approach zero.

This results in a fewer complex data set which in turn helps get the best-fitting model.

Applications of Ridge Regression

Ridge regression is used to eliminate multicollinearity in data models.

Used in the hospitality industry to manage seasonal fluctuations in booking prices of hotels or resorts



Used in farming as an effective method to predict grain yield under different water regimes

Assisted Practices

Let’s understand the topic below using Jupyter Notebook.

- 4.13_Data Preparation, Model Building, and Performance Evaluation: Part B

Note: Please download the pdf files for each topic mentioned above from the Reference Material section.

ASSISTED PRACTICE

Lasso Regression

Lasso Regression

Lasso regression is a form of regression that shrinks the coefficient toward zero to reduce the complexity of the data.

It forces some of the coefficient estimates to be exactly equal to zero with the help of a large tuning parameter λ.

It performs variable selection or feature selection.

It helps to reduce the learning of more complex data and overfitting of the model.

It decreases the variance of the model without an increase in bias.

Lasso Regression

The fitting procedure involves a loss function, also known as residual sum of squares (RSS), which is the difference between actual and predicted value.

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j xij\right)^2$$

Consider an example:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Where,
y = Independent variable
β = Regression coefficient value
x = Dependent variable

The variation differs from ridge regression only in penalizing the high coefficients.

Lasso Regression

The sum of the absolute value of weights is used instead of squaring the weights.

$$Cost\ (W) = RSS(W) + \lambda * (sum\ of\ absolute\ value\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} |wj|$$

Applications of Lasso Regression

Lasso regression is applied in data sets to shrink parameter estimates towards zero.

It is used in healthcare to distinguish cancer tissue from normal epithelial or stromal tissue.

It is used by surgeons to determine the spread of prostate cancer in a patient's body.

It is used by insurers to analyze social media profiles for lifestyle-based insurance offers.

It has many applications where variable selection or feature selection is required.

Assisted Practices

Let’s understand the topic below using Jupyter Notebook.

- 4.15_Data Preparation, Model Building, and Performance Evaluation: Part C

Note: Please download the pdf files for each topic mentioned above from the Reference Material section.

Key Takeaways

- Regression is a supervised learning technique that predicts continuous values.
- Linear regression predicts numerical dependent variables using one or more independent variables and can depict positive or negative linear relationships.
- Logistic regression is a statistical method used for binary or multinomial classification problems, and a sigmoid function is used to map predicted values to probabilities.
- SMOTE helps balance the data distribution by adding to the minority class.
- Ridge and lasso regression are regularization techniques used in ML to reduce the complexity of data and prevent overfitting.

Knowledge Check

Knowledge Check

1

What is the purpose of VIF in regression analysis?

A. To check the correlation between dependent and independent variables

B. To check the correlation between independent variables

C. To calculate the accuracy of the regression model

D. To calculate the variability of the regression model

Knowledge Check 1

What is the purpose of VIF in regression analysis?

A. To check the correlation between dependent and independent variables

B. To check the correlation between independent variables

C. To calculate the accuracy of the regression model

D. To calculate the variability of the regression model

The correct answer is **B**

Variance inflation factor (VIF) is used to identify the correlation between independent variables and to address multicollinearity issues in regression analysis.

Knowledge Check 2

Why is a sigmoid function used in logistic regression?

A. To plot data points on a graph

B. To cluster data points based on their distance from a centroid

C. To map real values to probabilities between 0 and 1

D. To perform classification analysis

Knowledge Check 2

Why is a sigmoid function used in logistic regression?

A. To plot data points on a graph

B. To cluster data points based on their distance from a centroid

C. To map real values to probabilities between 0 and 1

D. To perform classification analysis

The correct answer is **C**

The sigmoid function is used in logistic regression to map any real value to a probability between 0 and 1.

Knowledge Check

3

What is ridge regression?

A. A technique to reduce multicollinearity between features in the data set

B. A form of regression that shrinks the coefficient toward zero to reduce the complexity of the data

C. A technique to fit the data points to the line using polynomial features

D. A technique to predict the rise of different diseases within populations and their spread rates

Knowledge Check 3

What is ridge regression?

A. A technique to reduce multicollinearity between features in the data set

B. A form of regression that shrinks the coefficient toward zero to reduce the complexity of the data

C. A technique to fit the data points to the line using polynomial features

D. A technique to predict the rise of different diseases within populations and their spread rates

The correct answer is **B**

Ridge regression is a form of regression that shrinks the coefficient toward zero to reduce the complexity of the data.

Thank You!