Supervised ML-
Regression Capstone Project
Yes Bank-Stock Price Prediction

Abhishek Kumar
Amitha Sridhar
Mohita Rathour
Mukesh Sablani
Sanjay Paul

Introduction

“All there is to investing is picking good stocks at good times and staying with them as long as they remain good companies.”

-Warren Buffett

The stock market is known for being volatile, dynamic, and nonlinear. Accurate stock price prediction is extremely challenging because of multiple (macro and micro) factors, such as politics, global economic conditions, unexpected events, a company’s financial performance, and so on.

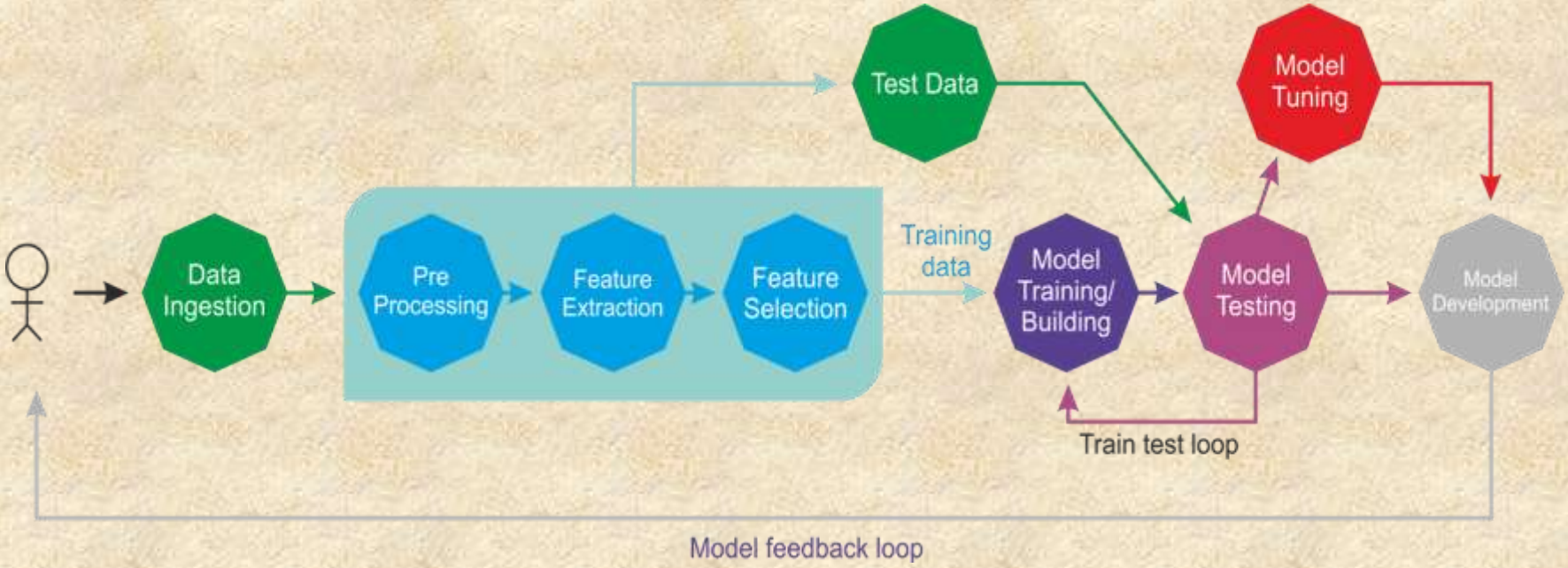
In the era of big data, deep learning for predicting stock market prices and trends has become even more popular than before.

Yes Bank is an Indian bank headquartered in Mumbai, India and was founded by Rana Kapoor and Ashok Kapoor in 2004.

In 2018 the Enforcement Directorate (ED) has alleged that Yes Bank co-founder Rana Kapoor and Dewan Housing Finance Limited (DHFL) promoters Kapil and Dheeraj Wadhawan siphoned off funds worth ₹ 5,050 crore through suspicious transactions.

In this project we will be studying the stock price pattern from the inception of the Yes Bank Stock and the effects of the alleged 2018 fraud case of Rana Kapoor on the stock price.

Workflow


Test Data
Model Tuning
Data Ingestion
Pre Processing
Feature Extraction
Feature Selection
Training data
Model Training/ Building
Model Testing
Model Development
Train test loop
Model feedback loop

Features

This data has 185 rows and 5 columns

-> **Date:-**A trade date refers to the month, day, and year that an order is executed in the market.

-> **Open:-**It is the price at which the financial security opens in the market when trading begins.

-> **High:-** the highest price at which a stock traded during the course of the trading day

-> **Low:-**the lowest price that a stock trades in that day.

-> **Close:-**the last level at which it was traded on any given day.

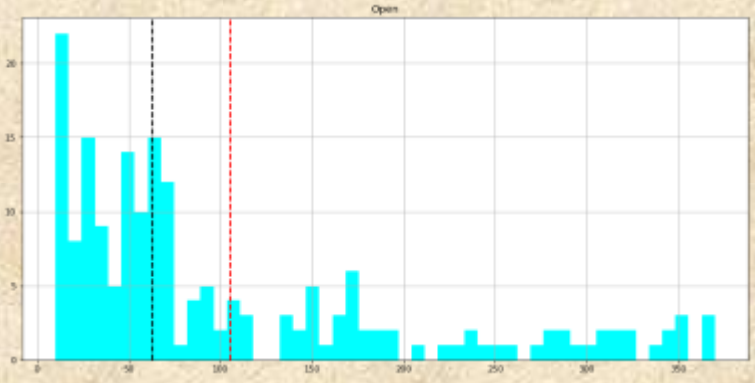
Year wise Study Open/Close

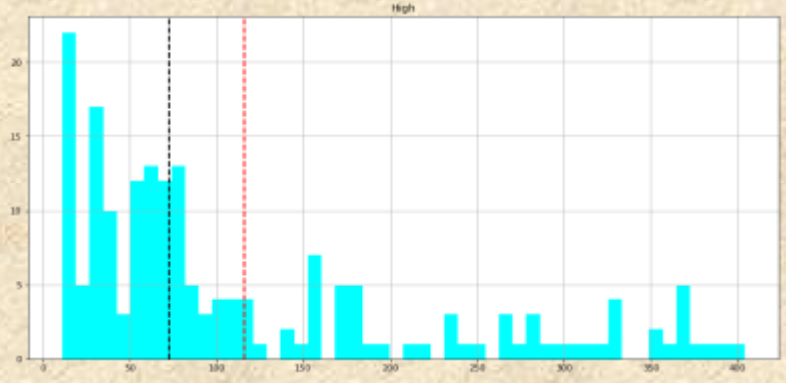
It’s seen that the Opening value of the stock Price has a steady growth since its inception till 2018. A total growth of around 900%. Then after the fraud case there is steep fall post 2018.

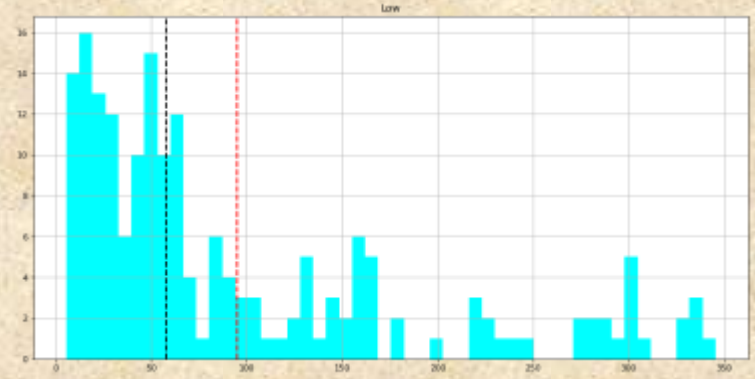
Year wise Study High/Low

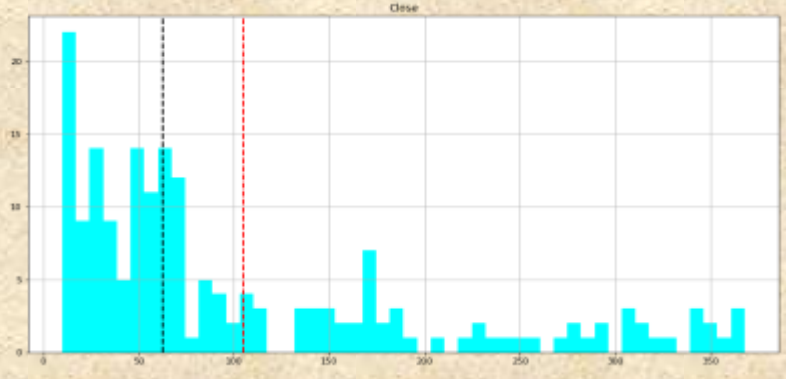
The Growth pattern shadows that of the Open/Close pattern. A steady growth since its inception till 2018 and then a steep fall post the alleged fraud case.

Mean and Median Study



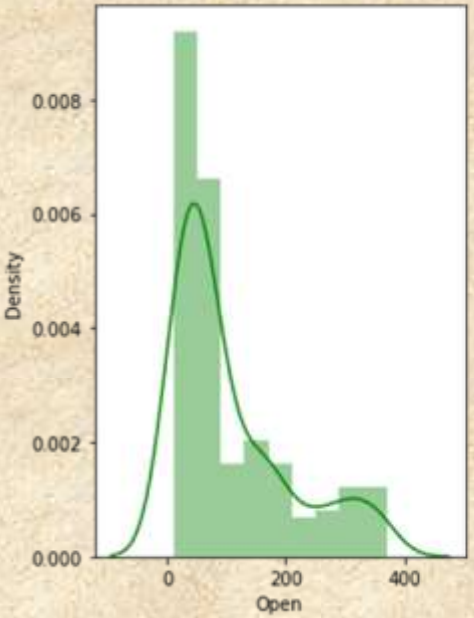


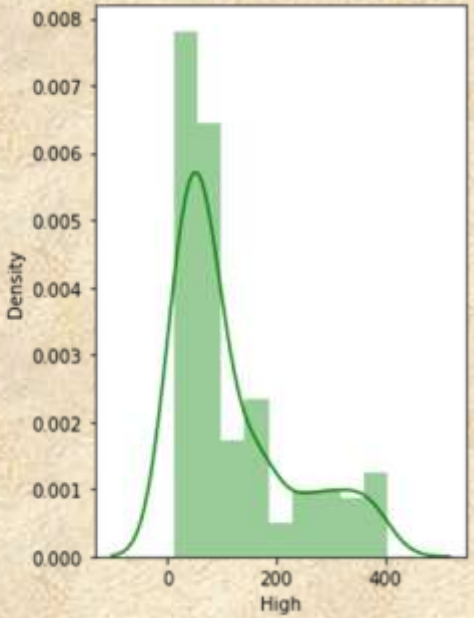


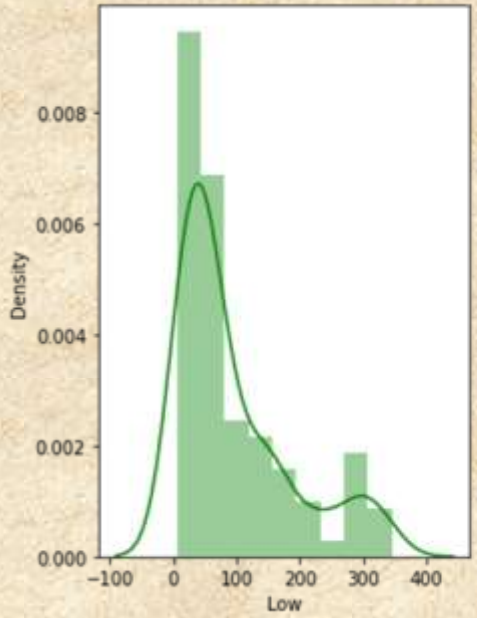


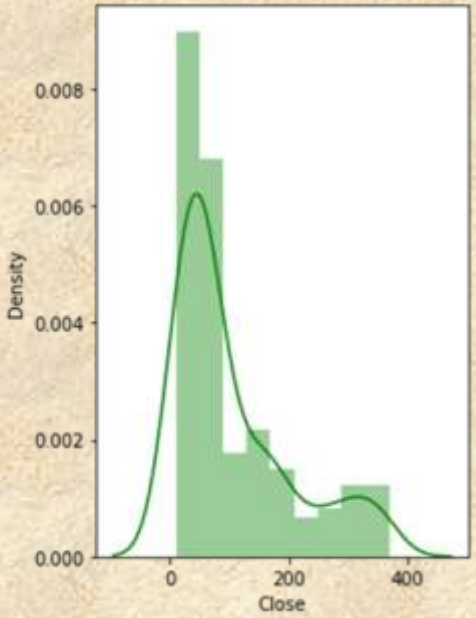
Data isn't evenly distributed. The difference between the mean and the median is high. Thus normalization of data is needed.

Data Distribution Graph









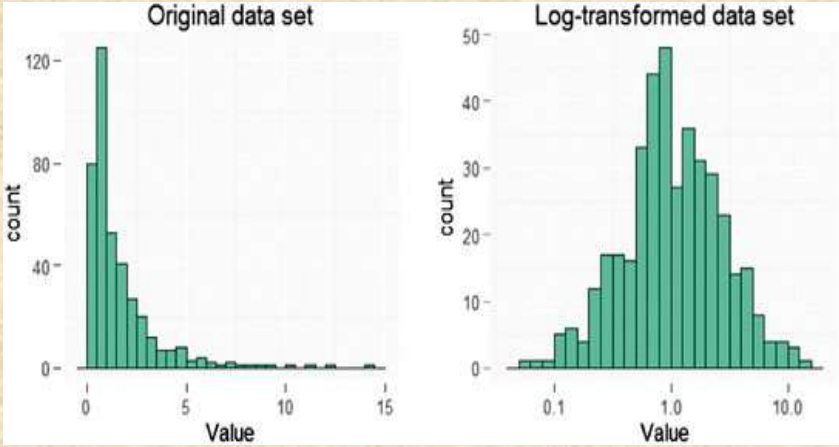
Data is Positively Skewed

Data Transformation

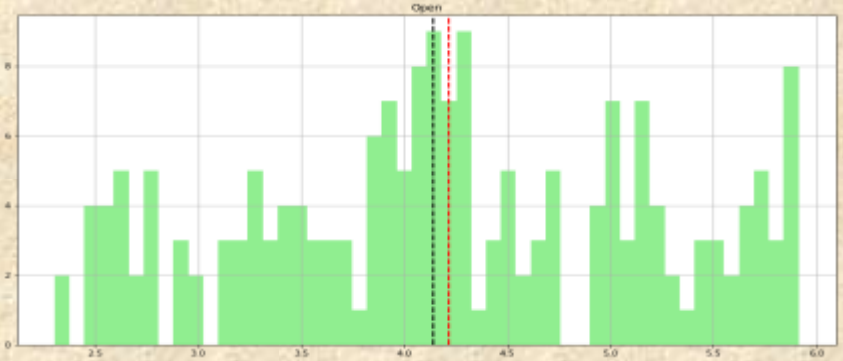
Data transformation is the process of converting, cleansing, and structuring data into a usable format that can be analyzed to support decision making processes.

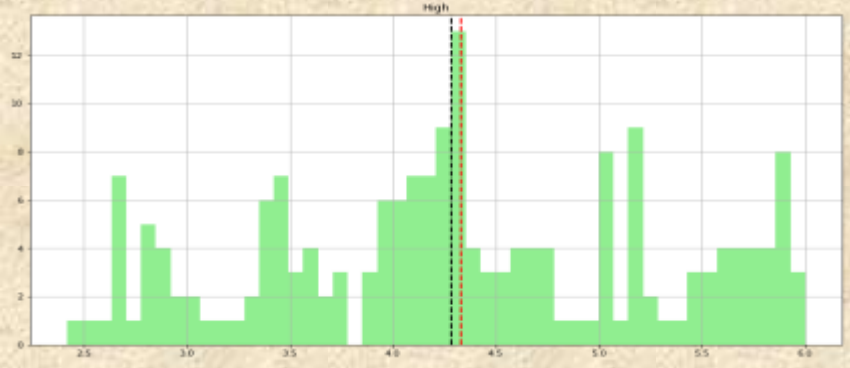
Log transformation is a data transformation method in which it replaces each variable x with a log(x).

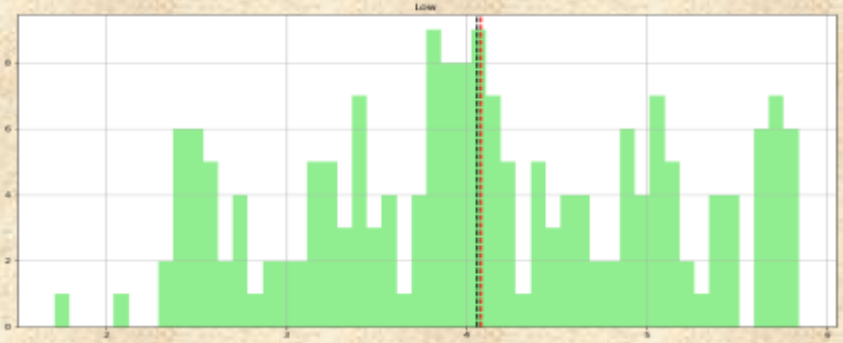
It is primarily used to convert a skewed distribution to a normal distribution/less-skewed distribution.
In this transform, we take the log of the values in a column and use these values as the column instead.



Mean and Median after Normalization

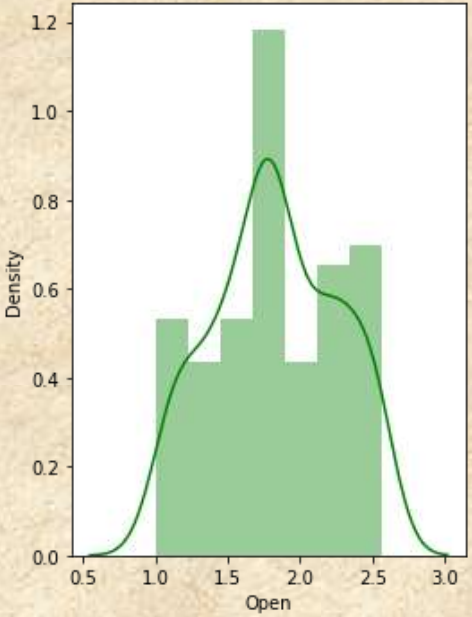


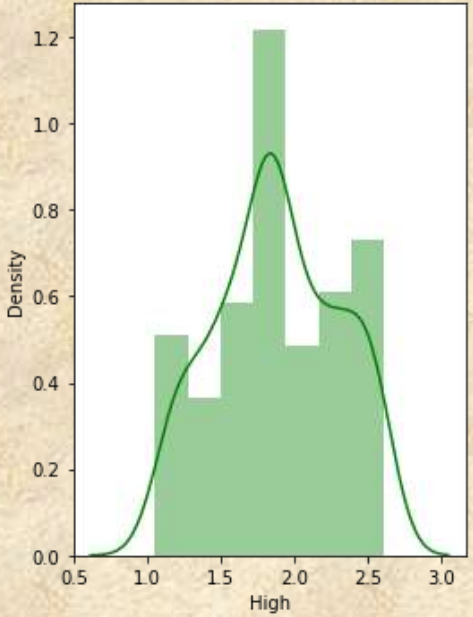


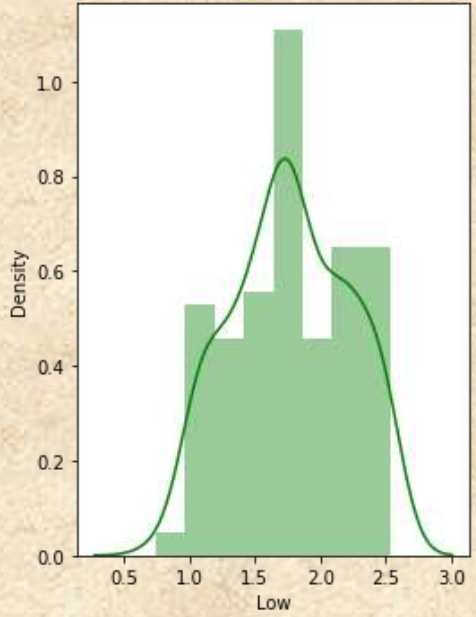


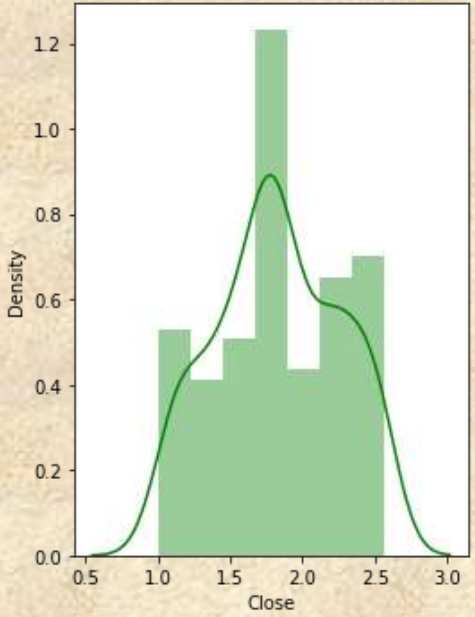
Normalized the data using the Log Transformation.Also the difference between Mean-Median is minimalized.

Data distribution graph after Normalization



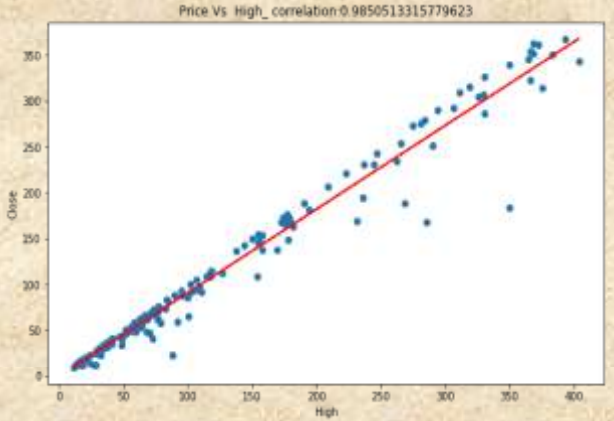


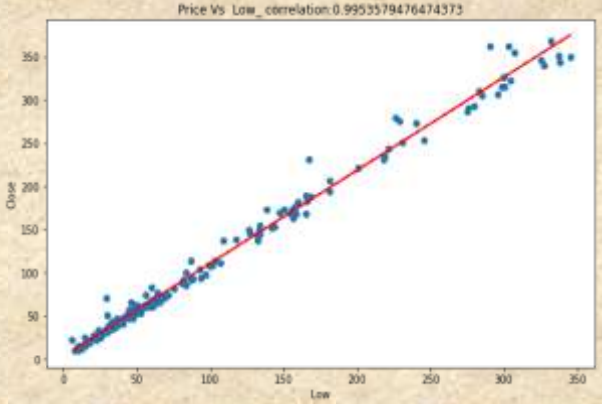


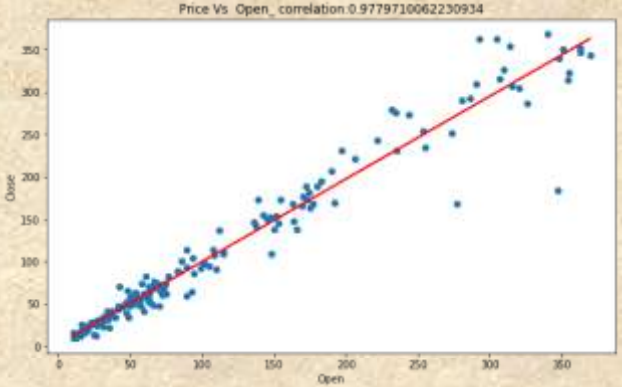


Post Log Transformation, data is uniformly distributed. Normal Bell curve achieved.

Correlation Graphs







Independent Features:- High , Low, Open
Dependent Feature:- Close

High Correlation between the dependent and independent Variables.

Multicollinearity And VIF

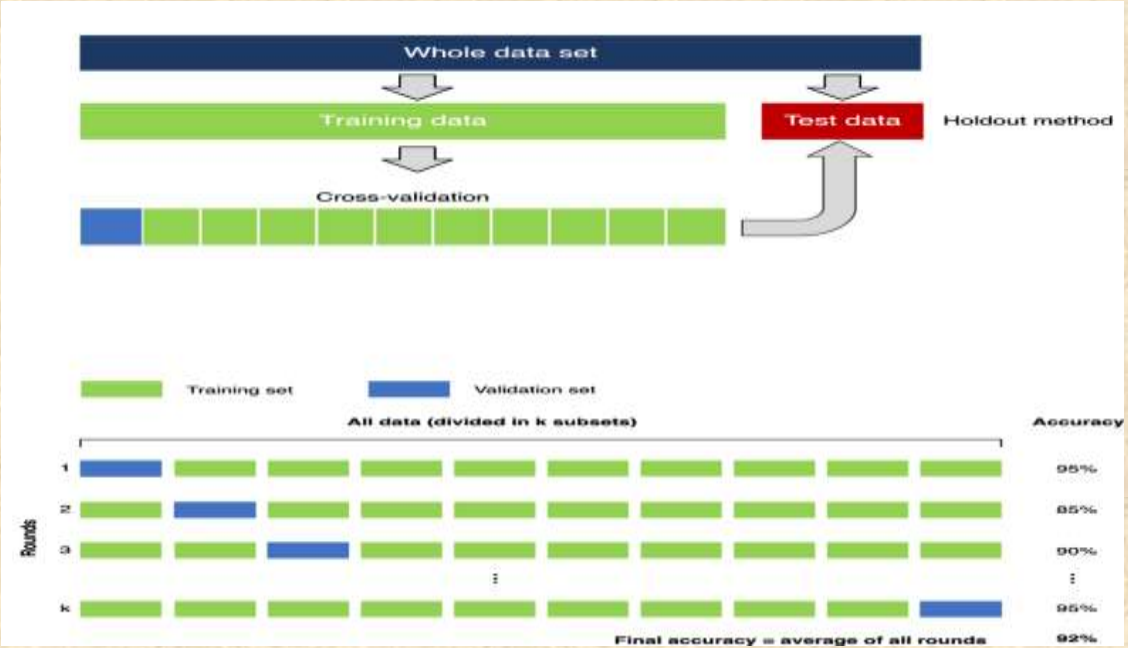
Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a regression model.

$$VIF_i = \frac{1}{1 - R_i^2}$$

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis.

Feature	VIF
Open	175.1857041
High	167.0575232
Low	71.574137

Train Test Model



Train-Test Split is a procedure that allows to simulate how a model would perform on given dataset.
Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model.
Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a model argument whose value is set before the learning process begins.

Normalization

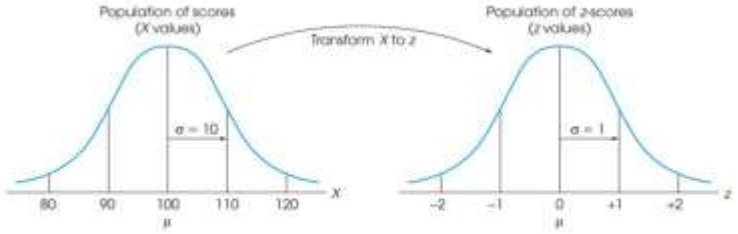
Normalization is a scaling technique used to change the values of numeric columns in the dataset to use a common scale.

It also improves the performance and accuracy of machine learning models using various techniques and algorithms.

$$z = \frac{x - \mu}{\sigma}$$

z	→	Standard (Normal) or Z score
x	→	member element of group
μ	→	mean of expectation
σ	→	standard deviation

Z-score is a variation of scaling that represents the number of standard deviations away from the mean. We should use z-score to ensure your feature distributions have mean = 0 and std = 1.

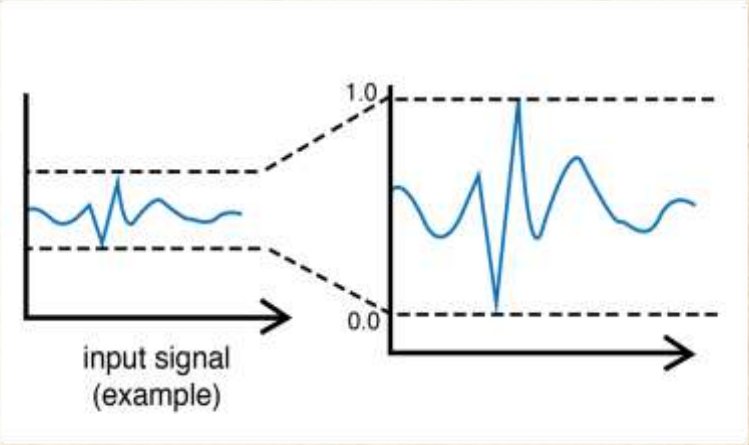


Z-Score Normalisation

Feature Scaling

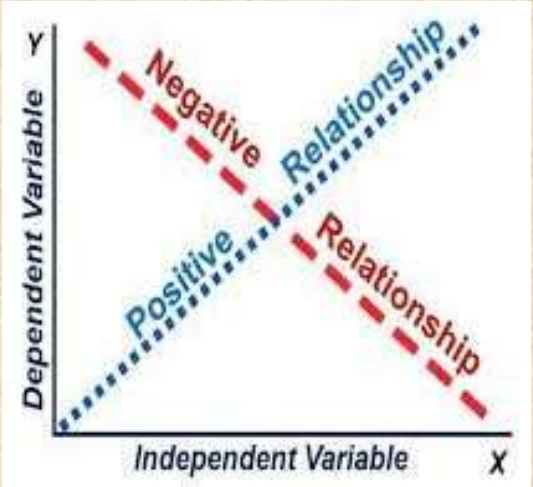
Feature Scaling is the process of scaling or converting all the values in our dataset to a given scale.

The MinMaxscaler is a type of scaler that scales the minimum and maximum values to be between 0 and 1 respectively.

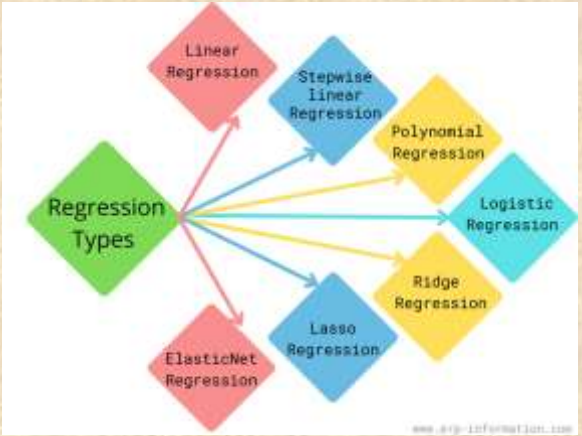


Model Training

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (Close) and a series of other variables (High,Low,Open).

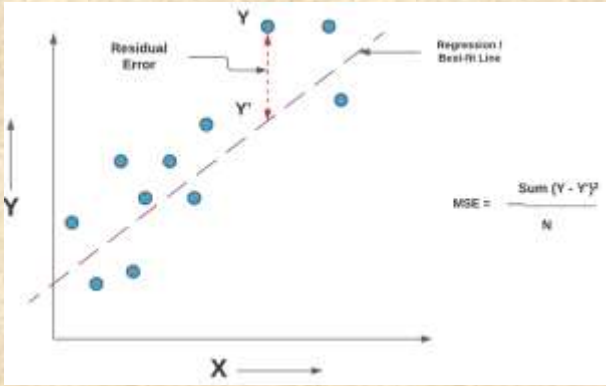


Model training is the phase in the data science development lifecycle where practitioners try to fit the best combination of weights and bias to a machine learning algorithm to minimize a loss function over the prediction range.



Metric Comparison Features

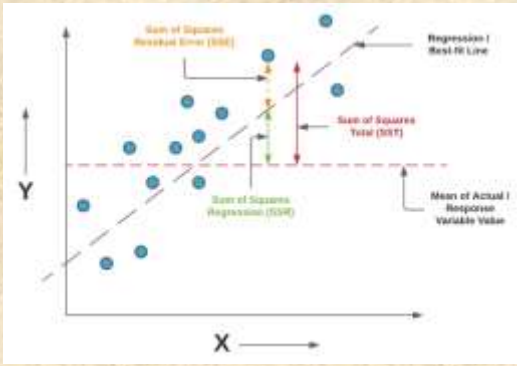
➔ **MSE:-** Mean squared error (MSE) measures the amount of error in statistical models.It assesses the average squared difference between the observed and predicted values.



➔ **RMSE:-**Root-mean-square Error(RMSE) is used to measure the differences between values predicted by a model and the values observed.

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

Matric Comparison Features



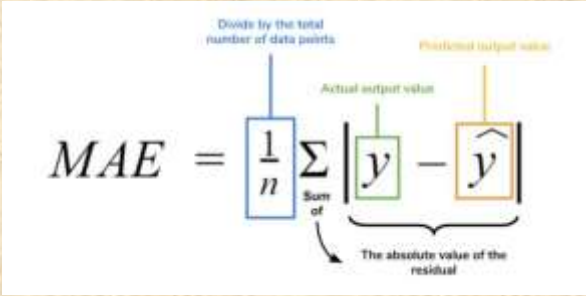
➔ **R2:-**R-squared (R2) represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

➔ **Adjusted R2:-**Adjusted R2 is a corrected goodness-of-fit (model accuracy) measure for linear models.

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

$$\text{Adjusted } R^2 = 1 - \frac{SS_{residuals}/(n-K)}{SS_{total}/(n-1)}$$

➔ **MAE:-**Mean Absolute Error (MAE) to the magnitude of difference between the prediction of an observation and the true value of that observation.RMSE is the Root of the Mean of the Square of Errors

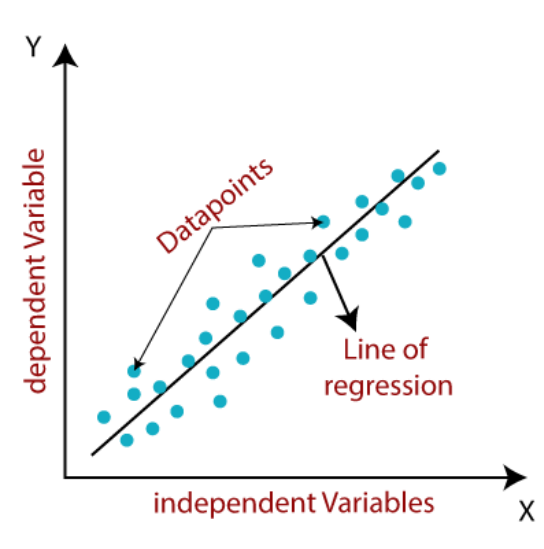


$$MAE = \frac{1}{n}\sum \left| y - \hat{y} \right|$$

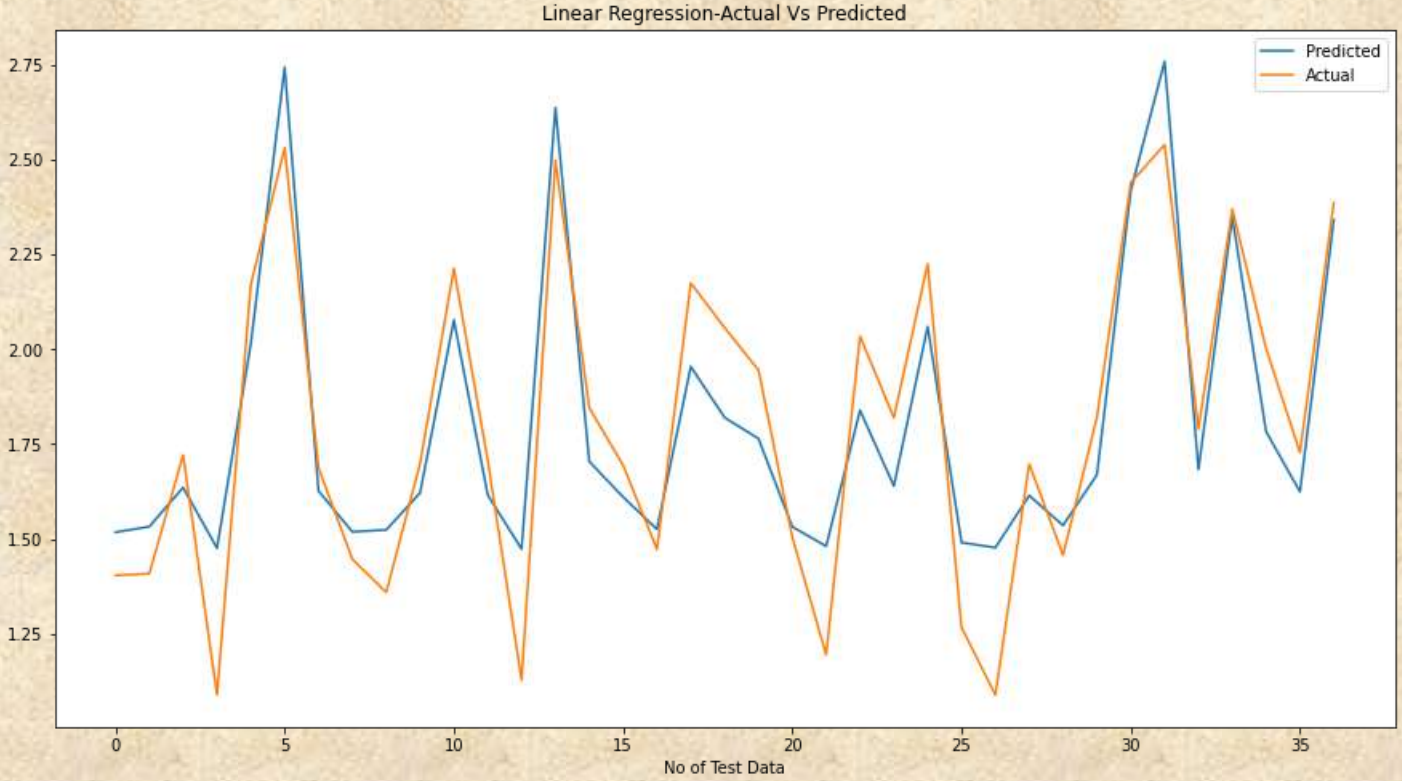
Linear Regression

Linear regression is used to model the relationship between two variables and estimate the value of a response by using a line-of-best-fit.

This calculator is built for simple linear regression, where we study the independent features Open, High and Low and how the dependent feature Close response to it.



Linear regression graph



MSE:-0.031583

MAE:-0.151285

RMSE:-0.177715

R2:-0.822570

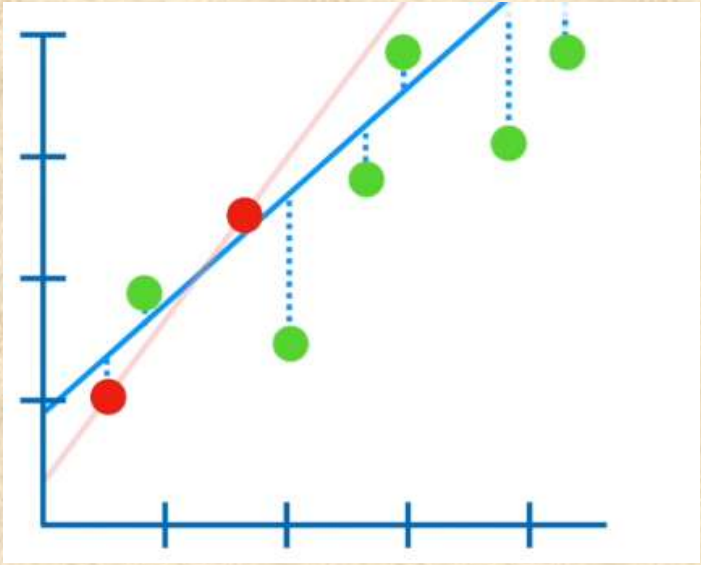
Adjusted R2:-0.806440

Ridge regression

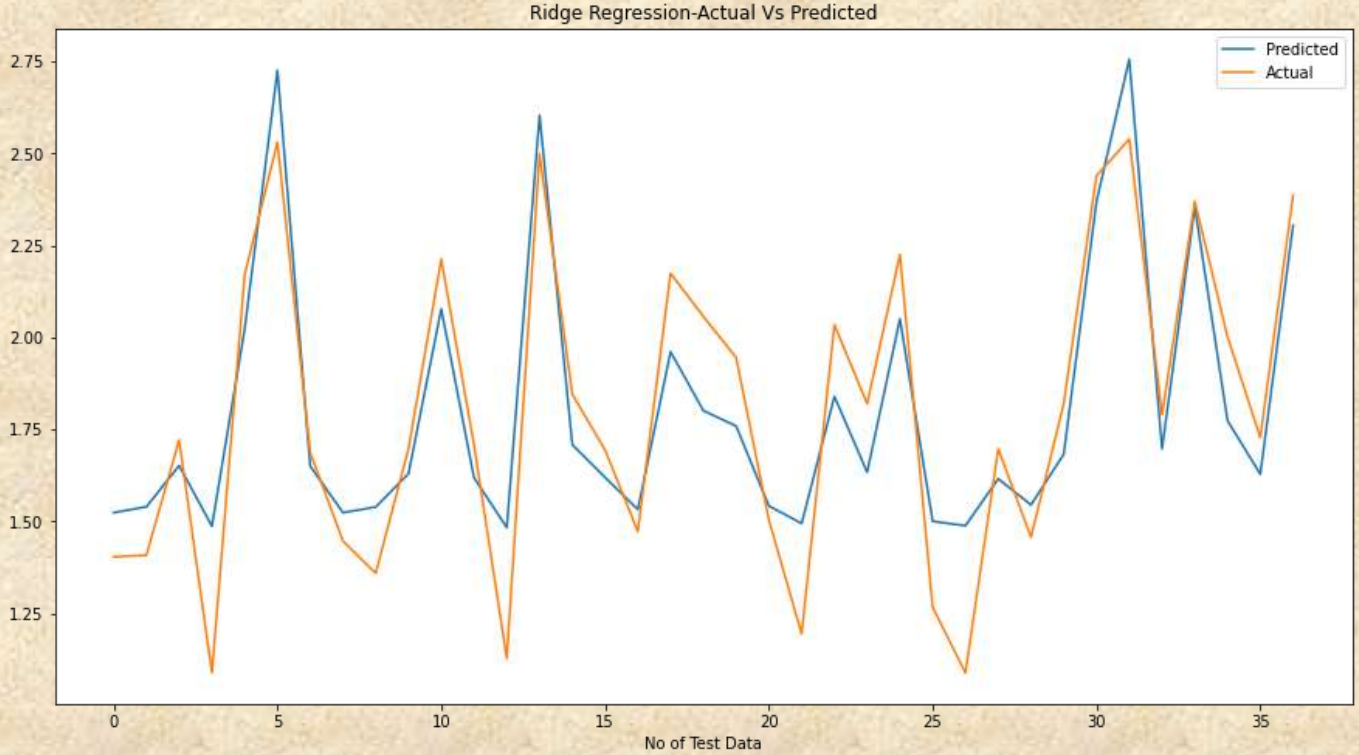
Ridge regression is a specialized technique used to analyze multiple regression data that is multicollinear in nature.

Regression stays the same, but in regularization, the way the model coefficients are determined is different.

The main idea of ridge regression focuses on fitting a new line that does not fit.



Ridge regression graph



MSE:- 0.031685

MAE:- 0.151477

RMSE:- 0.178001

R2:- 0.821997

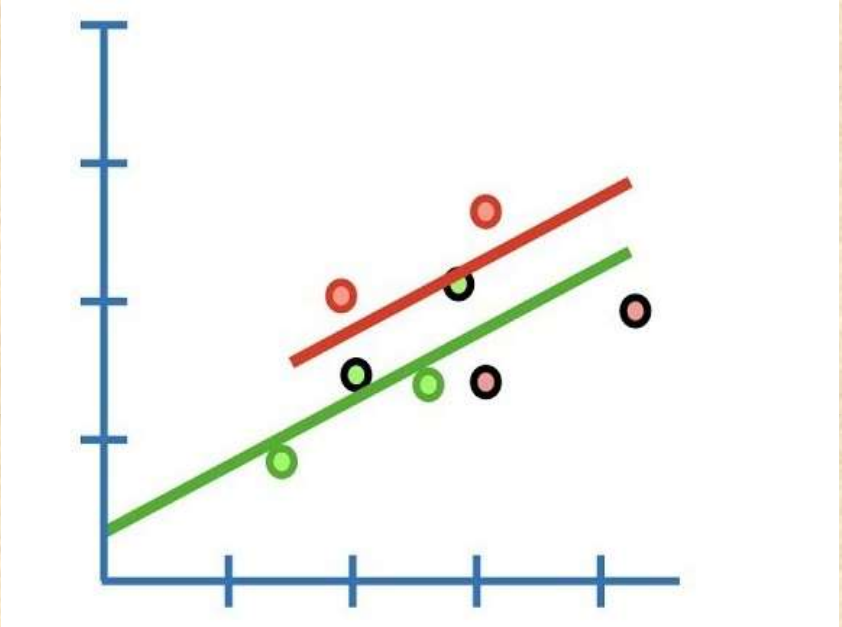
Adjusted R2:-0.805815

Lasso Regression

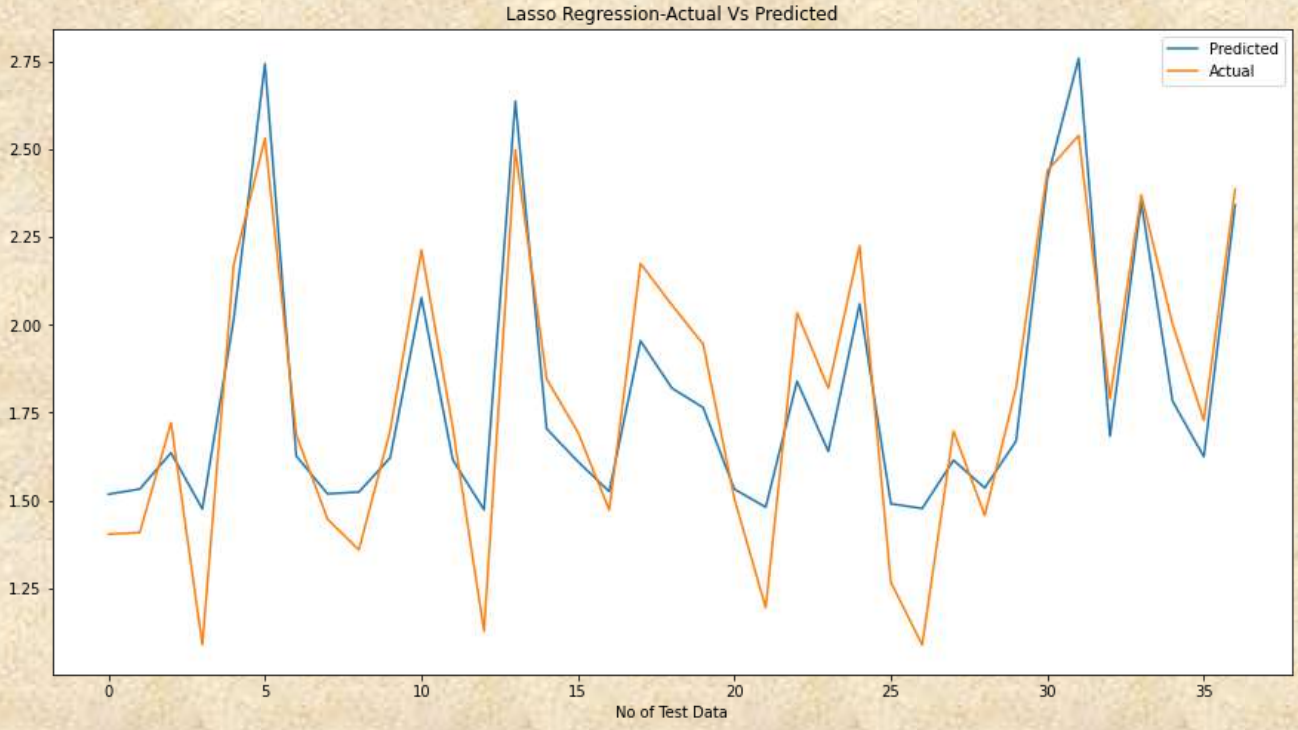
Lasso regression is a regularization technique over regression methods for a more accurate prediction.

This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean.

The lasso procedure encourages simple models with fewer parameters .



Lasso Regression Graph



MSE:- 0.032040

MAE:- 0.151477

RMSE:- 0.178996

R2:- 0.820001

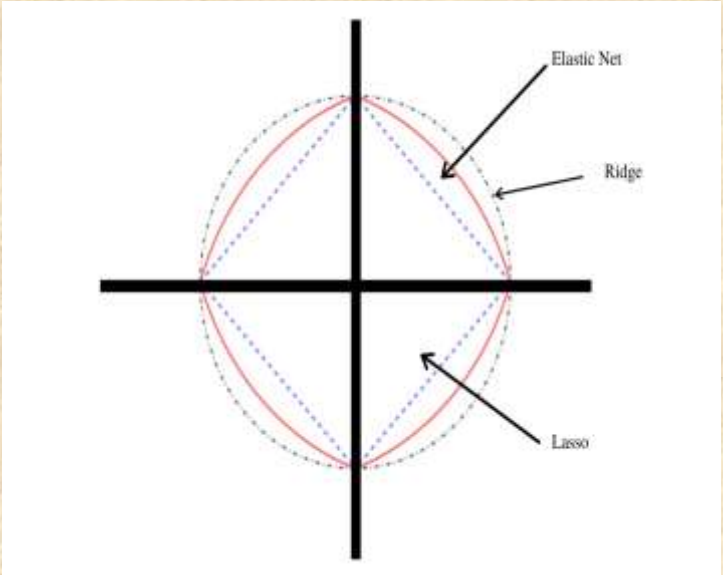
ADJUSTED R2:- 0.803638

Elastic Net regression

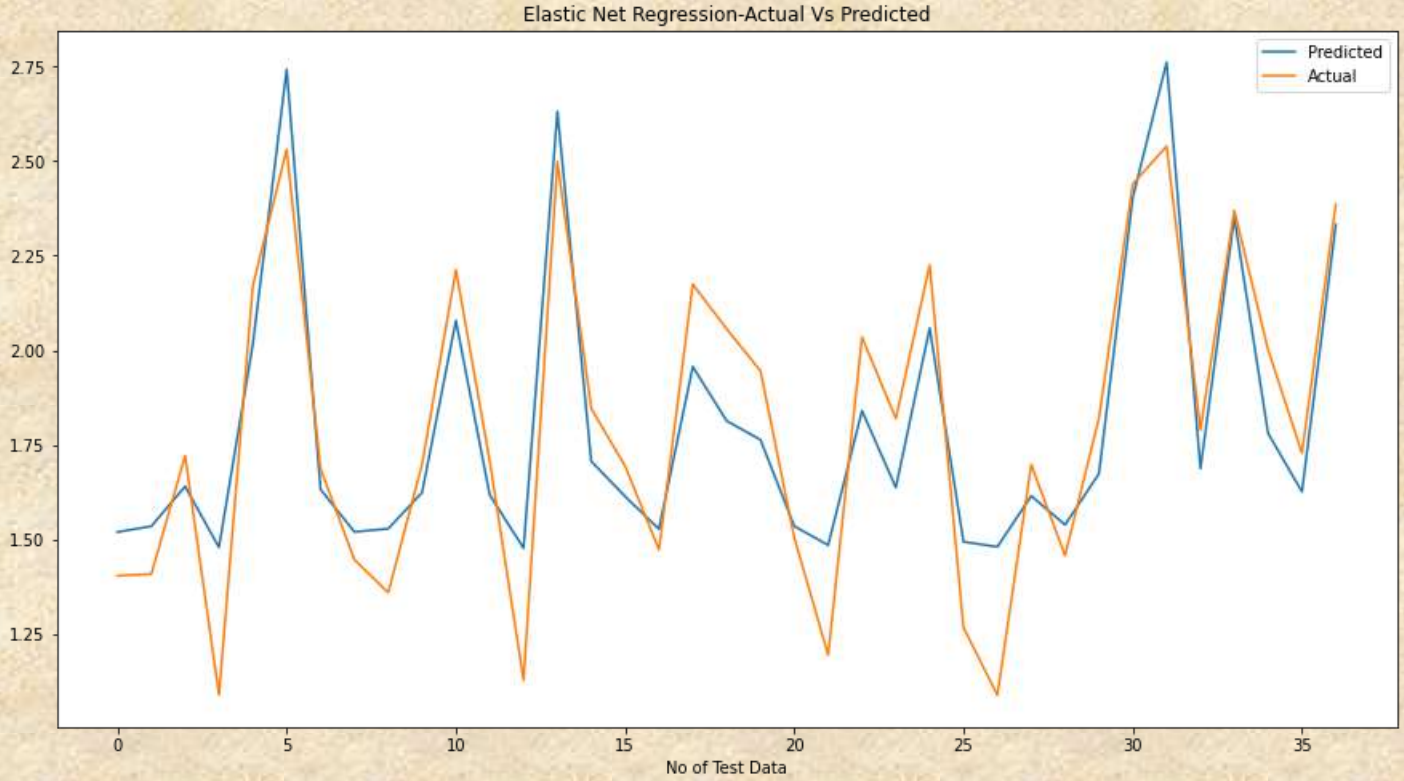
The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

Elastic Net, a convex combination of Ridge and Lasso.

It first finds the ridge regression coefficients and then conducts the second step by using a lasso sort of shrinkage of the coefficients.



Elasticnet regression Graph



MAE:- 0.031957

MSE:- 0.152095

RMSE:- 0.178764

R2:- 0.820468

ADJUSTED R2:- 0.804147

Matric Comparison

Model_Name	MSE	MAE	RMSE	R2	Adjusted R2
Linear Regression	0.031583	0.151285	0.177715	0.822570	0.806440
Ridge Regression	0.031685	0.151477	0.178001	0.821997	0.805815
Elastic Net Regression	0.031957	0.152095	0.178764	0.820468	0.804147
Lasso Regression	0.032040	0.151477	0.178996	0.820001	0.803638

From the above study we can say that Linear regression has the best R2 and Adjusted R2 score hence Linear Regression is the best model.


Thank
You