ASHISH AVDHESH GUPTA, COHORT SHIVALIK, ALMABETTER

Content -:

Acknowledgement

First of all, I would like to thank ***Almabetter*** for giving me opportunity to learn amazing things of python for data science. I would like to thank my all teachers and student success manager for helping me to solve my doubts.

Beside from my lecturer, I like to thank my friends for helping me to understand the project related questions more clearly. I thank them for their efforts.

Introduction

Yes Bank is a well-known bank in the Indian financial domain. It comes under top 10 banking in India. It was founded in 2004 and headquarters in Mumbai. It has 1000 no. of branches and 1800 no. of ATMs. It got listed on BSE, NSE in July 2005.

Our problem statement was since 2018, it has been in the news because of the fraud case involving Rana Kapoor. Owing to this fact, it was interesting to see how that impacted the stock prices of the company and whether Time series models or any other predictive models can do justice to such situations. This dataset has monthly stock prices of the bank since its inception and includes closing, starting, highest, and lowest stock prices of every month. The main objective is to predict the stock's closing price of the month.

Industry Insights

3.1 What is stock market and how it works?

Stock or share (also known as a company's 'equity') is a financial instrument that represents ownership in a Company or Stock markets are **where individual and Institutional investors come together to buy and sell shares in a public venue**. Units of stock are called "shares." Stocks are bought and sold predominantly on stock exchanges, though there can be private sales as well, and are the foundations of many individual investors' portfolios. Nowadays these exchanges exist as electronic marketplaces. Share prices are set by supply and demand in the market as buyers and sellers place orders.

3.2 Features Explanation

Since, we've the **YES BANK** Stock price dataset which has the monthly stock prices. It contains the following features:

Open - The opening price is the price at which a security first trades upon the opening of an exchange on a trading day i.e. buyers and sellers meet to make deals with the highest bidder, the opening price may not have to be the same as the last day's closing price.

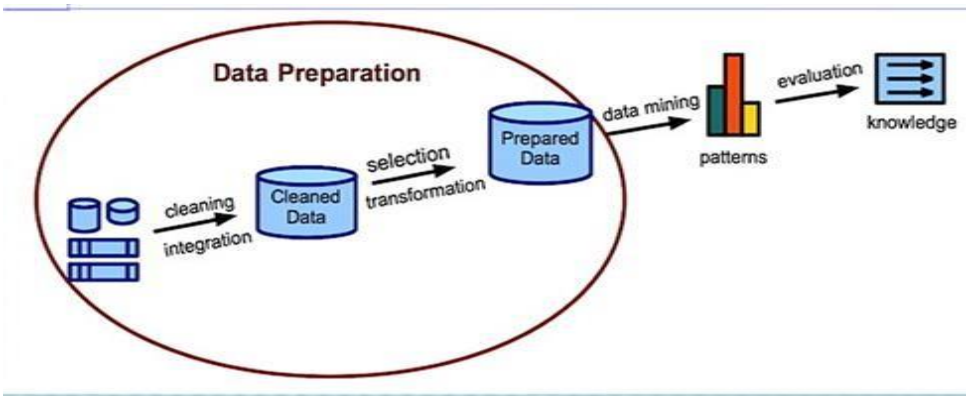
High - The high is the highest price at which a stock traded during a period.

Low -The low is the highest price at which a stock traded during a period.

Close -The closing price is a stock's trading price at the end of a trading day. This makes it the most recent price of a stock until the next trading session. The closing price is calculated as the weighted average price of the last 30 minutes, i.e. from 3:00 PM to 3:30 PM in case of equity.

Date – In date column monthly date is given.

Data Preparation and EDA



Data preparation is very important in exploratory data analysis. Because our output is depends on how our data is. First we check for the null and Na values. If very fewer amounts of data are Na or null then we just remove those data and if high amount of data is missing then we just replace it with mean and mode, depends on terminology. After this we look for duplicates. We check in data. If there are duplicates in our data then we drop those duplicate data. Now in our data there are zero null, Na and duplicates data.

4.1 Data Preparation

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making correlations to data and the combining of data sets to enrich data.

4.2 Why data preparation?

Data preparation ensures accuracy in data, which leads to accurate insights. Without data preparation, it's Possible that insights will be off due to junk data. i.e. ***No quality data, no quality mining results!***

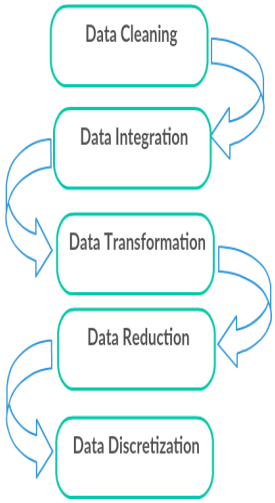
4.3 What are the techniques provided in data preparation?

1. Data cleaning: Fill in missing values, identify and/or outliers and resolve inconsistencies.

2. Data integration: integration of multiple datasets, data cubes, or files

3. Data transformation: Normalization and aggregation.

4. Data reduction: obtains reduced representation in volumes but produces the same or similar analytical result.



EDA

EDA is exploratory data analysis. In this we will analyze trends of our data more deeply. We will take each and every column and try to analyze it and get some insights from data. Same for continuous and categorical variables. **Exploratory Data Analysis** (EDA) is an approach to extract the information enfolded in the data and summarize the main characteristics of the data. EDA is essential for a well-defined and structured data science project and it should be performed before any statistical or machine learning modeling phase.

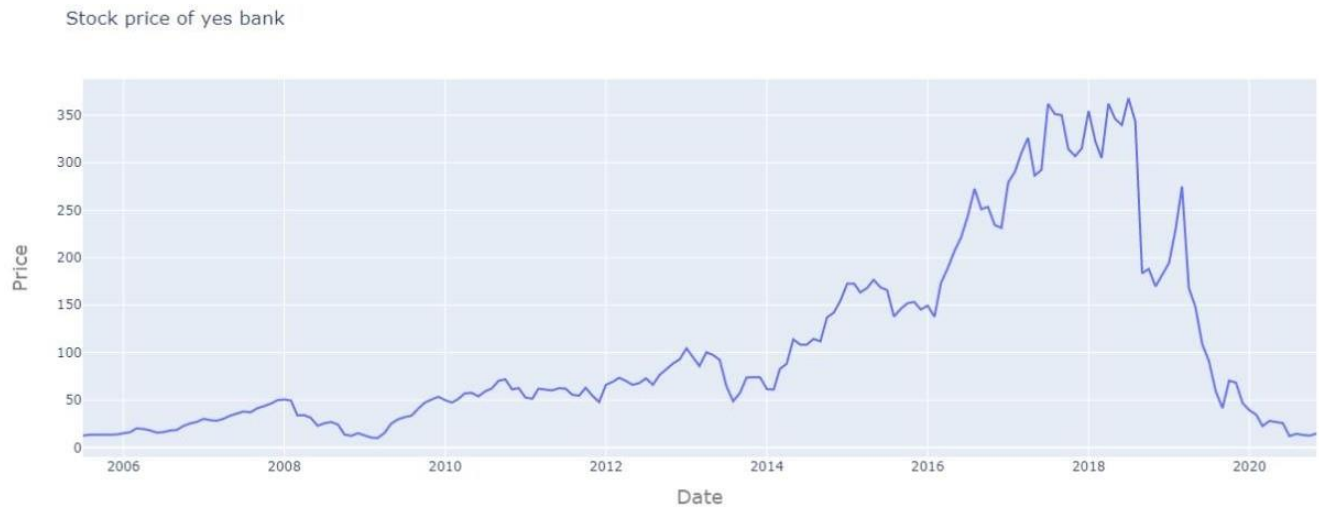
We have 185 rows and 5 columns in our data. Our date Column is of object data type we have to convert it into Date time object. This is achieved by using strptime of the datetime library. The Given Date format MMM-YY is converted to proper date of YYYY-MM-DD.

1) I checked for null values using **df.isnull().sum()** command and I found that we have zero null values in our data frame. If there are any null values in our data so we have to replace with mean or median.

2) I checked for duplicate values using **df.duplicated()** command and we have no duplicate data in our data frame. It's important to delete duplicate data because if we do not delete it then there are some chances that we can reflect from our original output.

3) The date column needs to be converted into a date time object. This is achieved by using **.strptime()** of the date time library. The Given Date format MMM-YY is converted to proper date of YYYY-MM-DD.

4.4 Visualizing the Data



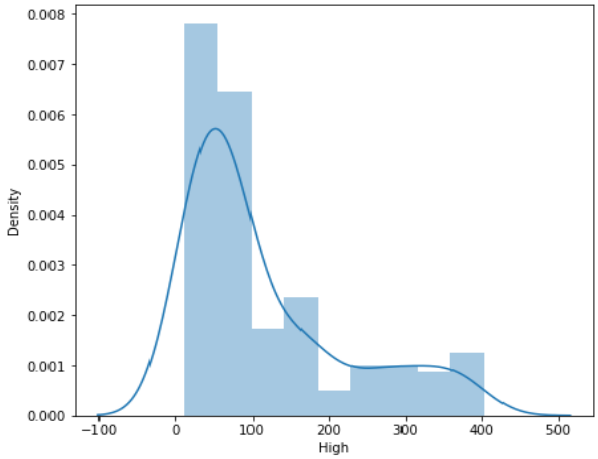
We can see that up until 2018, the stock prices more or less, kept increasing but there was a sudden dip after that. This can be attributed to the Yes bank fraud case against Rana Kapoor.

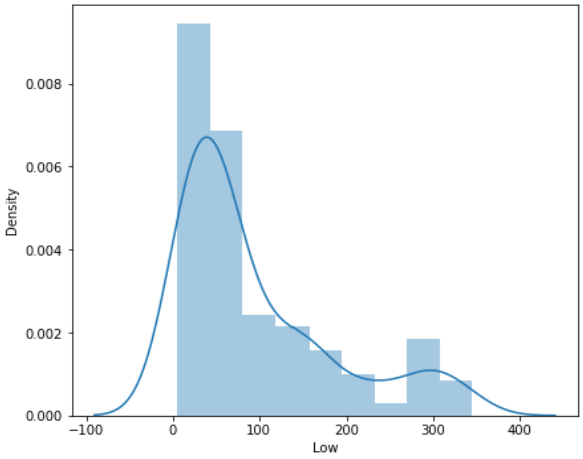
5.1 Univariate Analysis

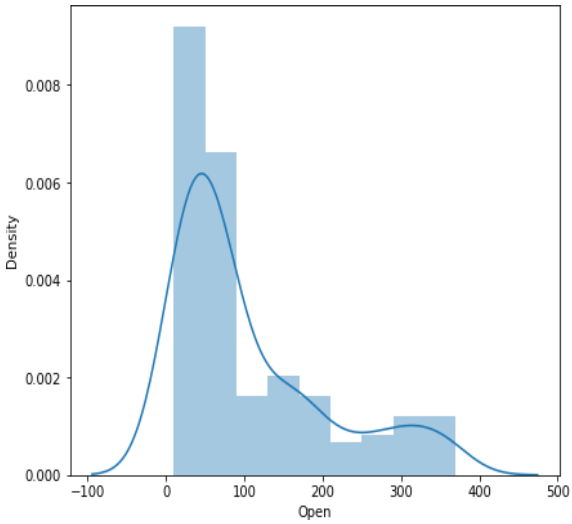
Univariate analysis means explores each variable in a data set, separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable.

Independent Variable

We have three independent variable i.e. High, Low, Open. We can see that all three independent Variable is positively skewed. We have to convert Skewed data into normal distribution. **Why do we want data to be normally distributed?** It is the most important probability distribution in statistics because it fits many natural phenomena. **Why is skewed data bad?** When these methods are used on skewed data, the answers can at times be misleading and just plain wrong. Even when the answers are basically correct, there is often some efficiency lost; essentially, the **analysis has not made the best use of all of the information in the data set**.

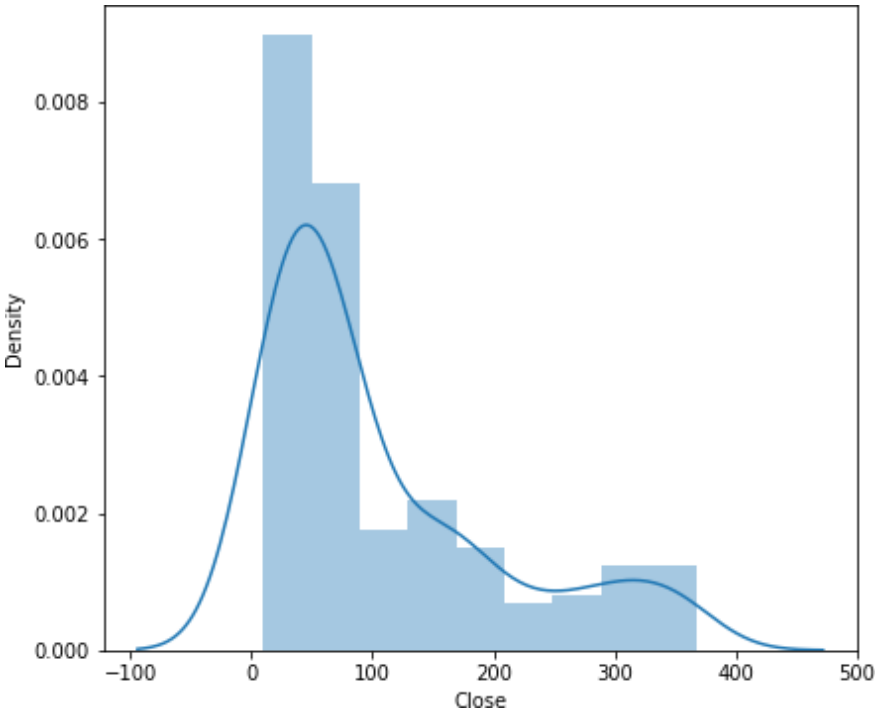






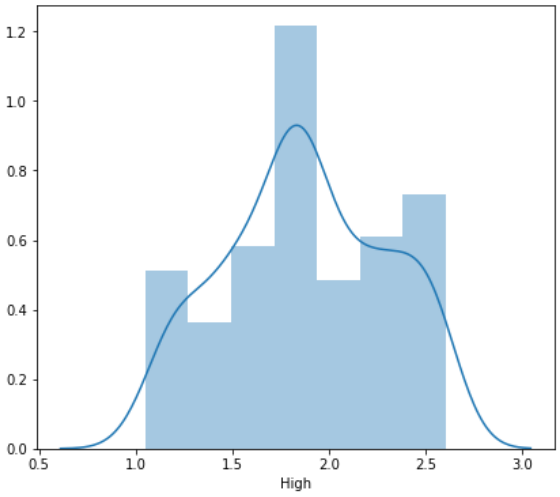
Dependent Variable

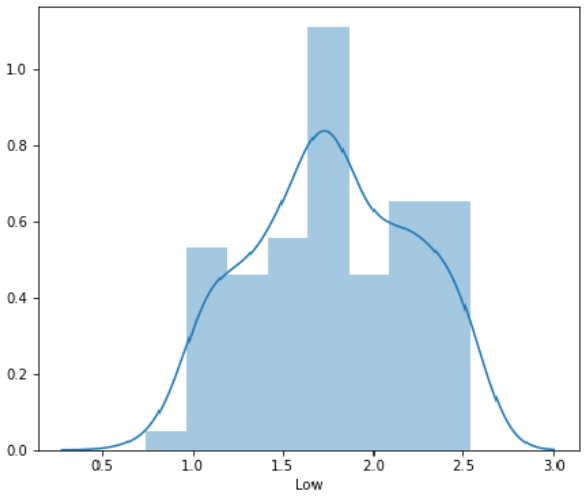
In our data dependent variable is “Close” column. The closing price is a stock's trading price at the end of a trading day. This column is also positively skewed. We have to convert it into normal distribution using log Transformation. **Why we are using log transformation?** We are using **LOGARITHMIC TRANSFORMATION** because it is most frequently used transformation is logarithmic transformation. Logarithmically transforming variables in a regression model is a very common way to handle situations where a non- linear relationship exists between the independent and dependent variables

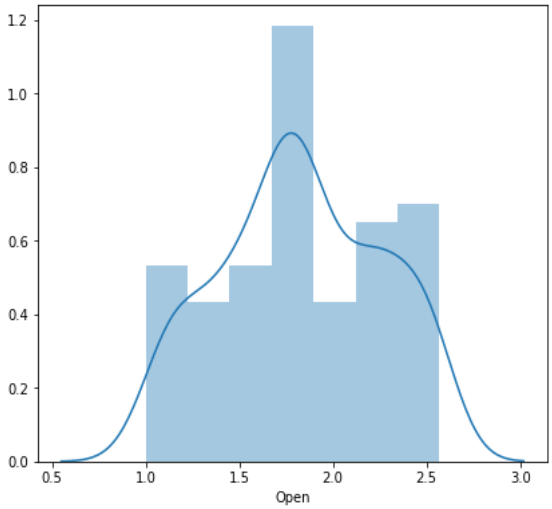


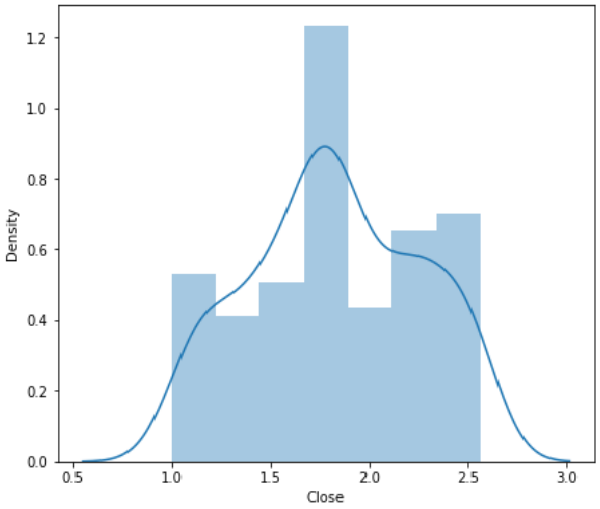
Log Transformation

We can see Log Transformation is pretty awesome. It makes our skewed original data i.e. High, Low, Open, and Close columns more normal. It improves linearity between our dependent and independent variables. It **boosts validity of our statistical analyses**. Now our all columns are almost normally distributed.



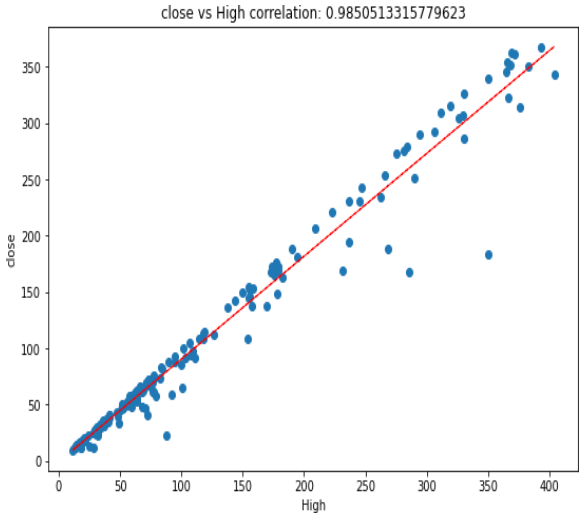


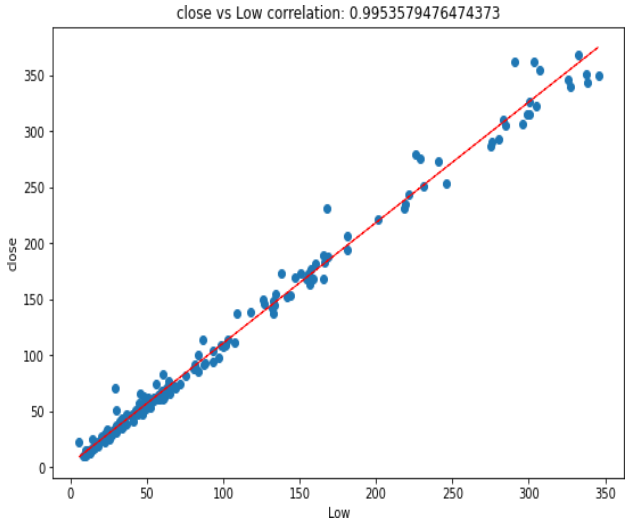


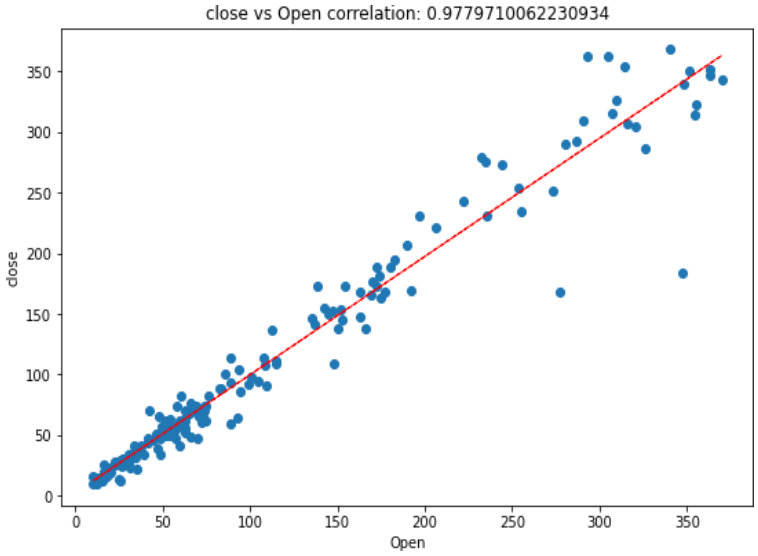


Bivariate Analysis

Bivariate analysis is performed to find the relationship between each variable in the dataset and the target variable of interest using 2 variables and finding the relationship between them. When we talk about bivariate analysis, it means analyzing 2 variables. Here we can see Close vs. High correlation is 0.98, Close vs. Low correlation is 0.99, Close vs. Open correlation is 0.97. We can see that all the independent variables are linearly related to our dependent variable.

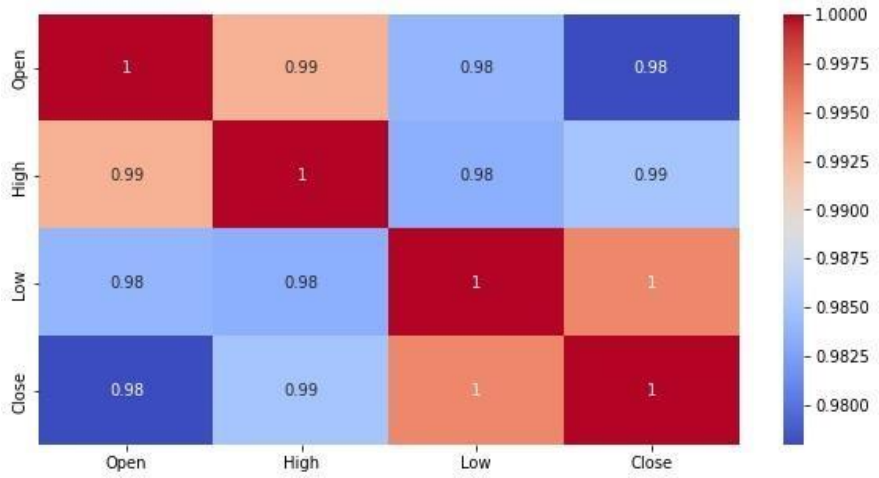





close vs Open correlation: 0.9779710062230934
close
Open
0
50
100
150
200
250
300
350

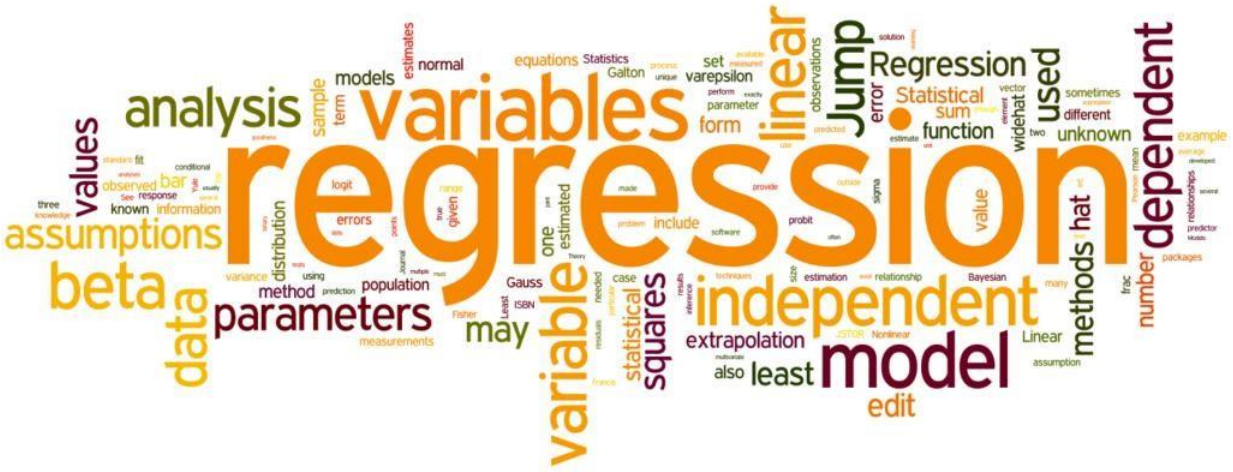
Correlation

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect. We describe correlations with a unit free measure called the correlation coefficient which ranges from -1 to +1 and is denoted by r. The closer *r* is to zero, the weaker the linear relationship. Positive *r* values indicate a positive correlation, where the values of both variables tend to increase together. Negative *r* values indicate a negative correlation where the values of one variable tend to increase when the values of the other variable decrease.



Regression

If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of y and x values. After developing such a model, if an addition value of x is then given without its accompanying value of y, the fitted in case study we choose linear regression for our analysis. First, we divide the data into two parts of training and resting. Then we use the training section for starting analysis and defining the model.

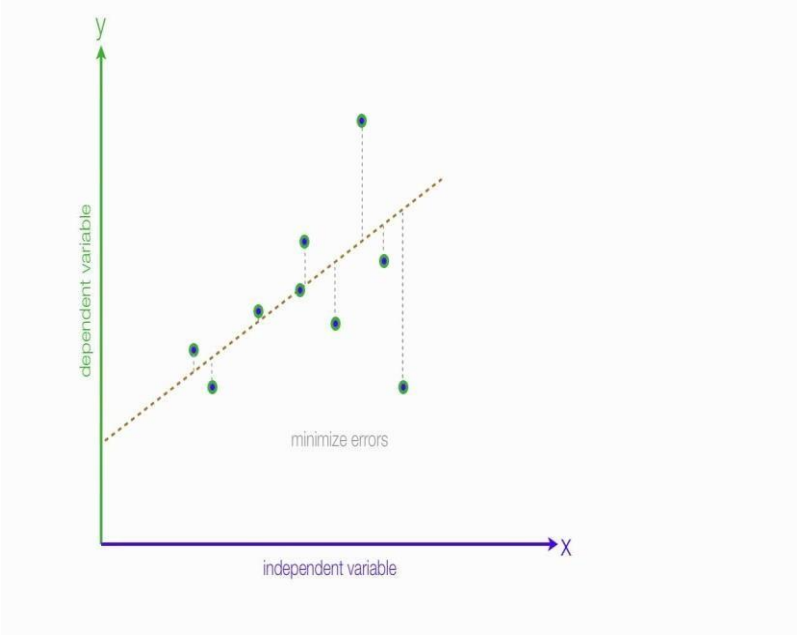


Linear Regression

What is Linear Regression?

Here is the formal definition, “Linear Regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variable denoted x.

Let me explain the concept of regression in a very basic manner, so imagine that you run a company that builds cars and you want to understand how the change in price of raw materials (let’s steel) will affect the sales of the car. The general understanding is this; the rise in the price of steel will lead to a rise in the price of the car resulting in lesser demand and in turn lesser sales. But how do we quantify this? And how do we predict how much change in sales will happen based on the degree of change in steel price. That’s when the regression comes.



Linear regression is the analysis of two separate variables to define a single relationship and is a useful measure for technical and quantitative analysis in financial markets.

Plotting stock price along a normal distribution bell curve can allow traders to see when stock is overbought or oversold.

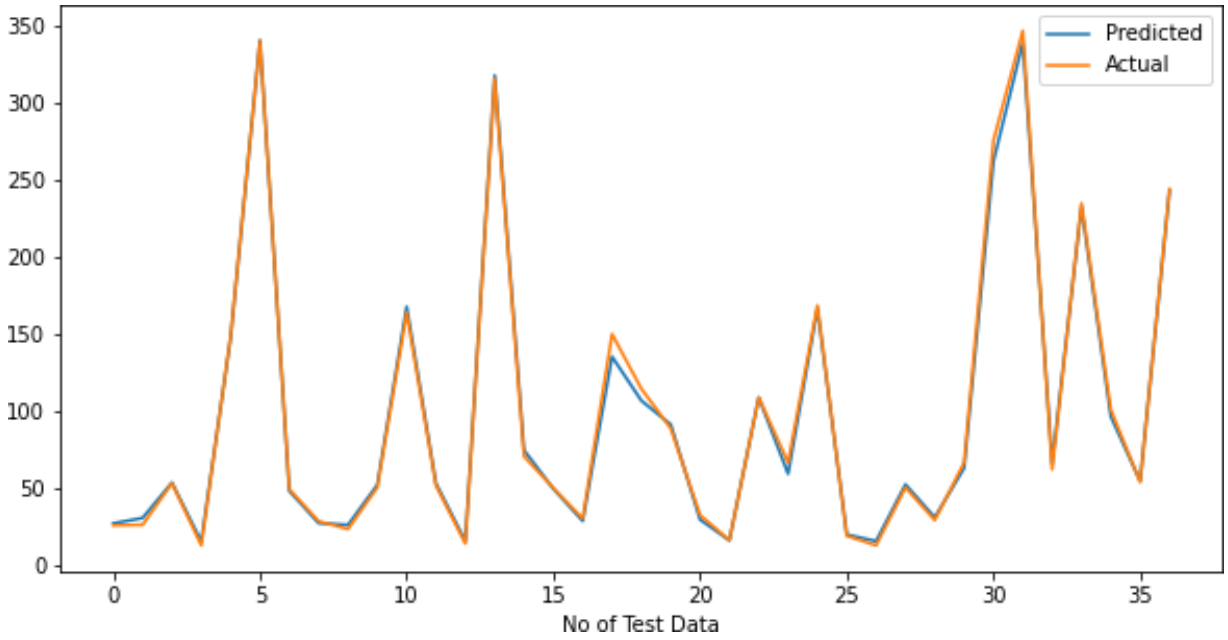
Using linear regression, a trader can identify key price point's entry price, stop Loss price, and exit prices.

A stock's price and time period determine the system parameters for linear regression, marketing the university applicable. Stock market close price is an important piece of information that is very short term trader. The close prices are very important, especially for swing traders and positional traders.

Following is our measures value:

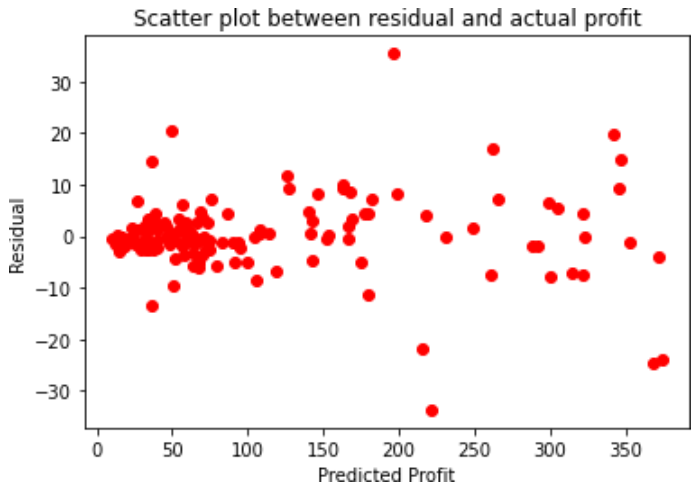
```
MSE : 19.988578593595022
RMSE : 4.470858820584142
R2 : 0.9978412541225983
Adjusted R2 : 0.9976450044973799
```

Following is the graph of linear regression:



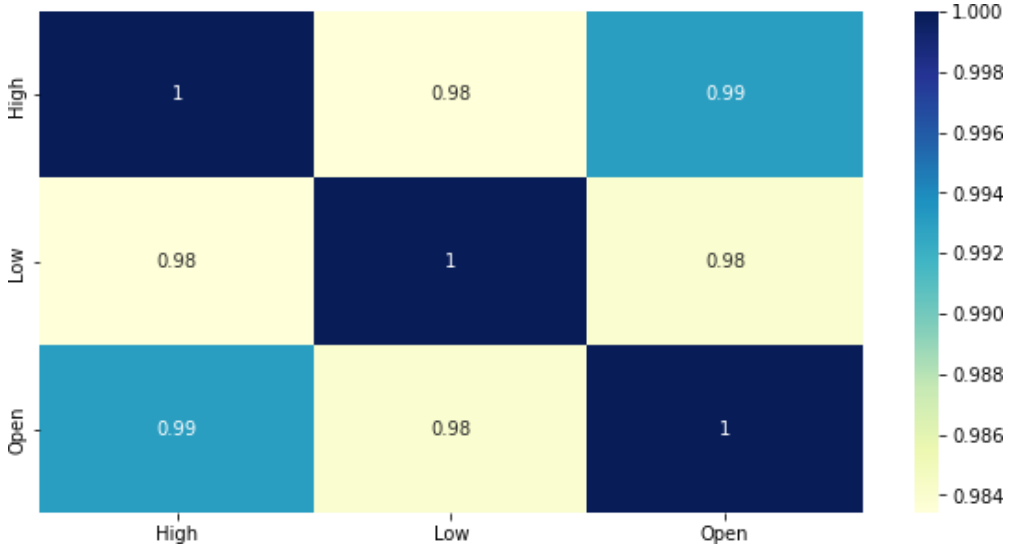
Validating Assumptions

1) Checking Hetroscedasticity



There is no significant pattern visible. So the assumption of homoscedasticity is valid.

2) Checking Multicollinearity



Since our data doesn't contain perfect multicollinearity among independent variables. We can't remove any variables from our date because each of the variables is important for our model.

Lasso Regression

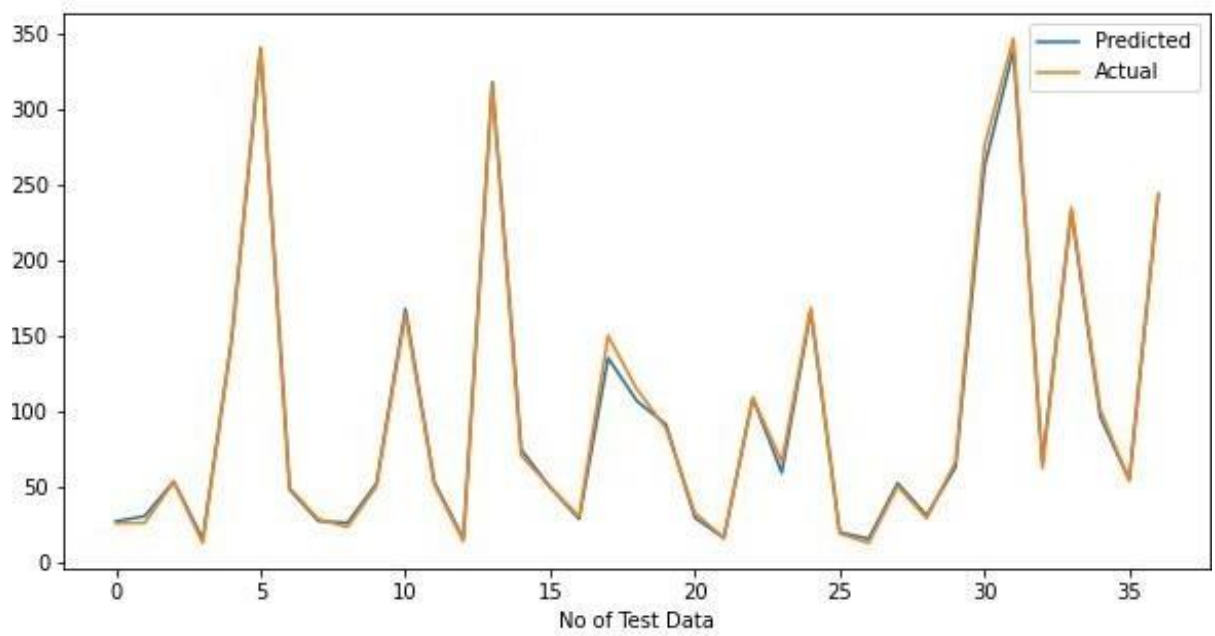
The word “LASSO” stand for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator. It is a statistical formula for the regularization of data models and feature selection. This technique used to overcome over fitting for a regression model. Residual Sum of Squares + λ * (Sum of the absolute value of the magnitude of coefficients) Where, λ denotes the amount of shrinkage. λ = 0 implies all features are considered and it is equivalent to the linear regression where only the residual sum of squares is considered to build a predictive model

λ = ∞ implies no feature is considered i.e. as λ closes to infinity it eliminates more and more features. The bias increases with increase in λ variance increases with decrease in λ

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|w_j|$$

I apply lasso regression on our model. I got the best fit lambda as 0.01 Then I check for MSE, RMSE, R^2 Adjusted R^2 to check the accuracy of the Model and I got

MSE : 20.878651216190214
RMSE : 4.569316274475889
R^2 : 0.9977451271971334
Adjusted R^2 : 0.9975401387605092


Predicted
Actual
350
300
250
200
150
100
50
0
0
5
10
15
20
25
30
35
No of Test Data

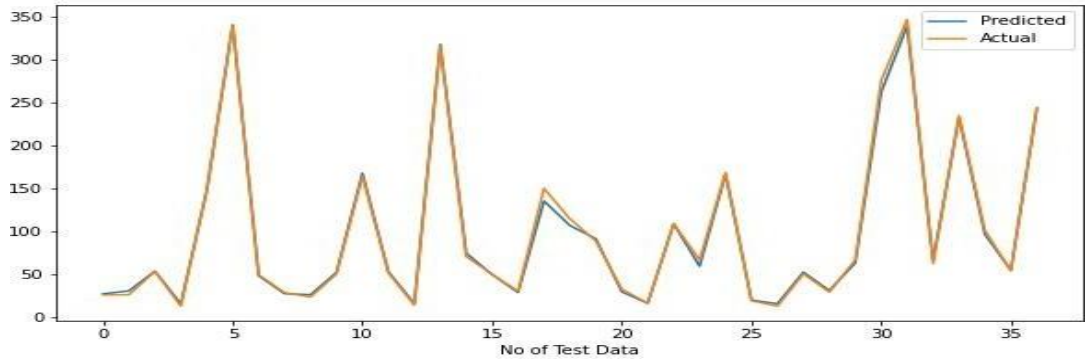
Ridge Regression

Ridge regression is a model tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values. Lambda is the penalty term. λ given here is denoted by an alpha parameter in the ridge function. So, by changing the values of alpha, we are controlling the penalty term. Higher the value of alpha, bigger is the penalty and therefore the magnitude of coefficients is reduced. λ denotes the amount of shrinkage.

$$\sum_{i=1}^{M}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}w_j^2$$

I apply Ridge regression on our model. I got the best fit lambda as 0.001 Then I check for MSE, RMSE, R^2 Adjusted R^2 to check the accuracy of the Model and I got,

MSE : 20.095425485603688
RMSE : 4.48279215284444
R^2:0.9978297147684337
Adjusted R^2 : 0.9976324161110186



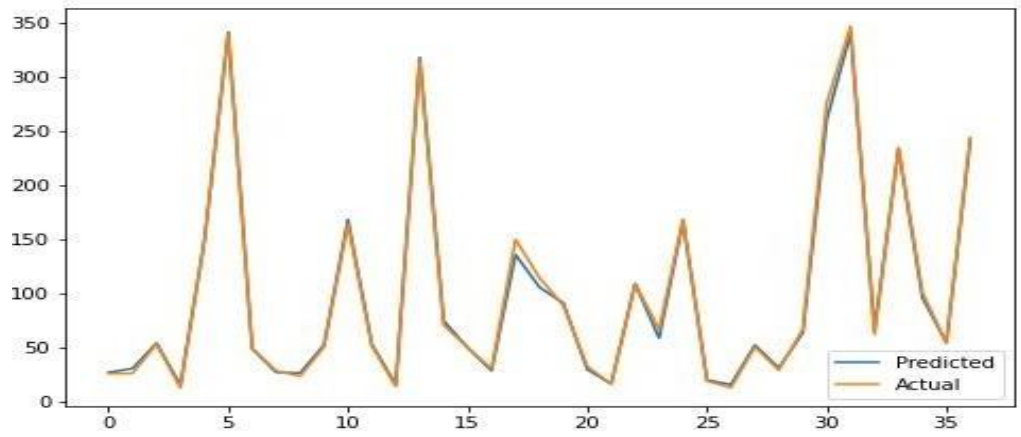
Elasticnet Regression

Elastic Net first emerged as a result of critique on lasso, whose variable selection can be too dependent on data and thus unstable. The solution is to combine the penalties of ridge regression and lasso to get the best of both worlds. Elastic Net aims at minimizing the loss function. Where α is the mixing parameter between ridge (α = 0) and lasso (α = 1).

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j-1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j-1}^{m}|\hat{\beta}_j|\right),$$

I apply the Elasticnet regression on X test of our data to get Y predicts. Then I get best fit alpha value 0.0001 and L1 ratio 0.3 Then I check for MSE, RMSE, Adjusted R^2 to check the accuracy of the Model and I got

MSE : 21.250552258088767
RMSE : 4.6098321290572795
R^2 : 0.9977049622680845
Adjusted R^2 : 0.997496322474274



Metric Comparison

Here we compare the performance of every model.

Model Name	MSE	RSME	R^2	ADJUSTED R^2
Linear regression	19.988579	4.470859	0.997841	0.997645
Lasso regression	20.095425	4.482792	0.997830	0.997632
Ridge regression	20.878651	4.569316	0.997745	0.997540
Elasticnet regression	21.250552	4.609832	0.997705	0.997496

From above metric comparison we can say that our best model is linear regression because it has minimum MSE value.

Conclusion

1) From price graph we can see until 2018 stock price was increasing but in next year July month suddenly falls. Because of rana kapoor fraud case.

2) From linear regression we can conclude that there is effect of that fraud news. Because we are getting high accuracy on linear regression.

3) In our data doesn't contain perfect Multicollinearity among independent variable. We can't remove any variable from our data because each of the variables is important for our model.

4) Target variable is highly dependent on input variable.

5) All regression models have given the best result with lowest MSE, RMSE values and highest r2, adjusted r2 value.