

**Capstone project -2**

**Yes Bank Stock Price Prediction Technical Documentation**

**

**BY - ANKIT RAI**

**CONTENTS**

* **Abstract**
* **Introduction**
* **Problem Statement**
* **Data Description**
* **Exploratory Data Analysis**
* **Conclusion**

**Abstract :-**

Yes Bank, formed in 2004, provides a diverse range of distinct solutions to its corporate and retail customers through retail banking and asset management services. It is also a publicly traded corporation. Anyone can invest in Yes Bank and become a shareholder as a result of this. At the same time, it means that the company's valuation is now in the hands of investors and speculators, as share prices are frequently highly influenced by public opinion.Yes Bank stock price data set was used. This dataset has five distinct features that can be used to forecast closure prices.Machine learning is being used. For price prediction, we created a machine learning regression model. We now have Some of the better models were used.

**INTRODUCTION**

YES bank is an abbreviation for Youth Enterprise Scheme Bank. Because the stock market is one of the most popular sectors, stock market price prediction is a popular issue for academics in both financial and technical fields. Our project's goal is to create a prediction model for close price prediction. A stock market is a public market in which shares of publicly traded corporations can be bought and sold.

Stock price prediction using machine learning provides an estimate of the future worth of a company's stock and other financial assets traded on an exchange.

The main point of predicting stock prices is to make large gains.

It is difficult to predict how the stock market will fare. There are countless other possibilities factors involved in the prediction, such as the psychological

factor – namely crowd behavior etc. All these factors combine to make share prices very difficult to predict with high accuracy.

**Problem Statement:**

Yes Bank is a well-known banking institution in India. It has been in the headlines since 2018 due to the Rana Kapoor fraud case.

Because of this, it was intriguing to analyse how it affected the company's stock values and whether any forecasting models can do justice to such situations. This dataset contains the bank's monthly stock prices since its inception, including the closing, starting, highest, and lowest stock values for each month. The main goal is to forecast the stock's monthly closing price.

**Dataset Description**

It is critical to comprehend the data before executing any operations on it. We examined the dataset after loading it by inspecting a few of the initial and end rows. We examined the dataset's form and discovered that it has 185 rows and 5 feature columns.

Let's look at the features in our dataset.

• **Date:** It denotes the date of the investment (in our case we have month and year).

• **Open:** The price at which a stock began trading when the opening bell rang.

• **High:** The highest prices in a specific time period are referred to be high.

• **Low:** The lowest prices in a certain time period are referred to as low.

• **Close:** The price of a single stock at the close.

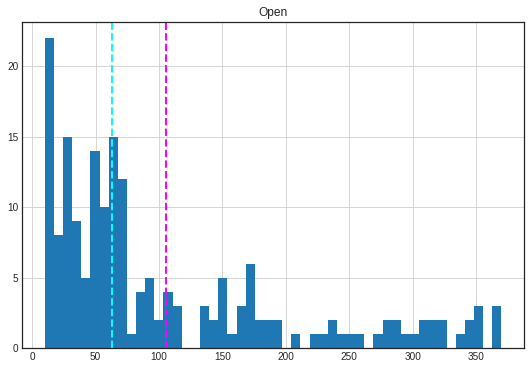
**Exploratory Data Analysis:-**

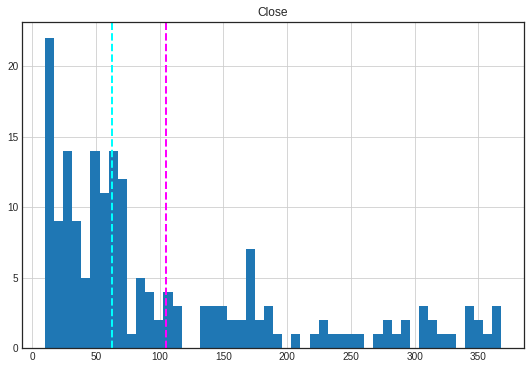
A) Data Cleaning: The supplied date in data is transformed to suitable date format of YYYY-MM-DD and given date column has dtype as object changing it into date time format.

B) Treatment of null values: Our dataset does not contain any null values, which can impair our accuracy. Depending on the situation, we might discard null numbers or replace them with mean or median values.

C) Visualization of Data:

1. Univariate Analysis: All of the parameters in our yes bank stock market dataset have favourably skewed distributions.



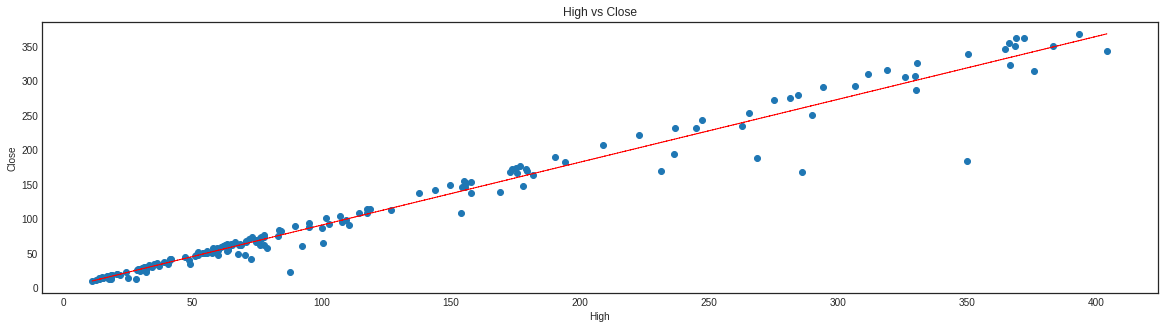


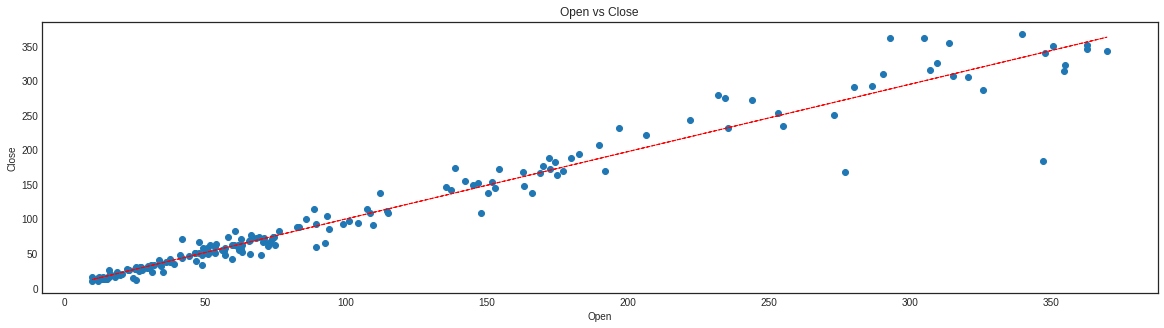
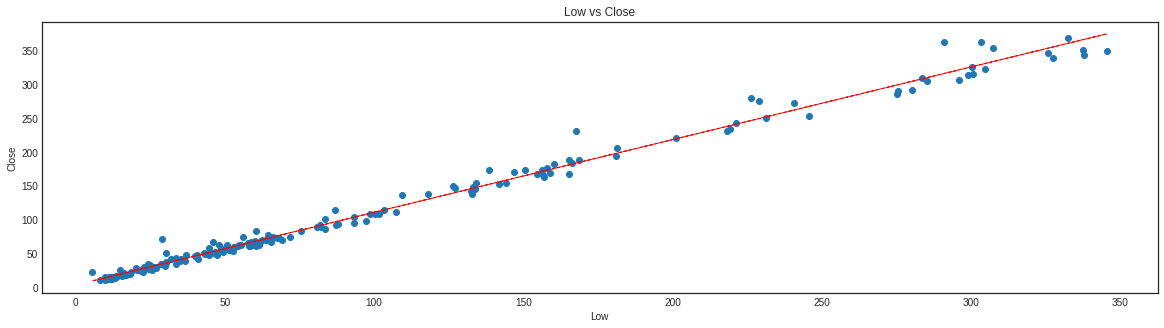
They are not regularly distributed, as evidenced by the graph above. For a perfect normal distribution curve, the mean and median should be equal.

As a result, we log convert all of the features to a normal distribution.

**Bivariate Analysis:** When performing a bivariate analysis by graphing one variable against another in the setting of supervised learning, it can aid in determining the essential predictors.

The graphs below show a strong relationship between the dependent (close) and independent variables.

****

****

**Open price and Close Price:**

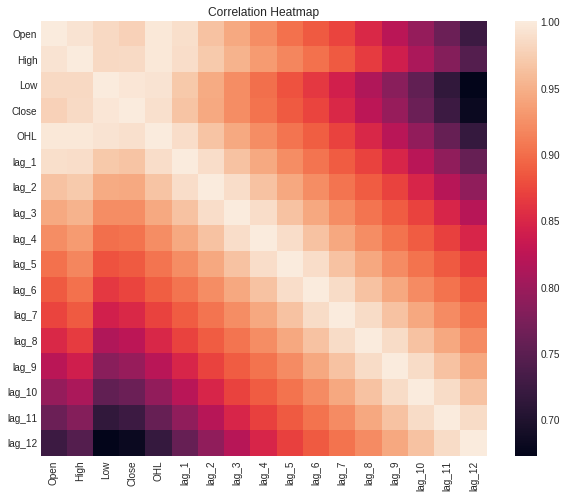
We can assume from the above line plot that the stock price will continue to rise till 2018. However, the stock price has continued to fall since 2018, owing to the Rana Kapoor fraud case.



The. graph depicts how the closing price changes with each passing year.The graph plainly shows that around 2018, when the fraud case involving Rana Kapoor came to light, there was a definite dramatic drop in the stock price of Yes Bank data

**Correlation Analysis:**

Correlation analysis is a statistical tool for determining the strength of a link between numerical variables. This heatmap depicts the relationship between all of the numerical variables in our data.

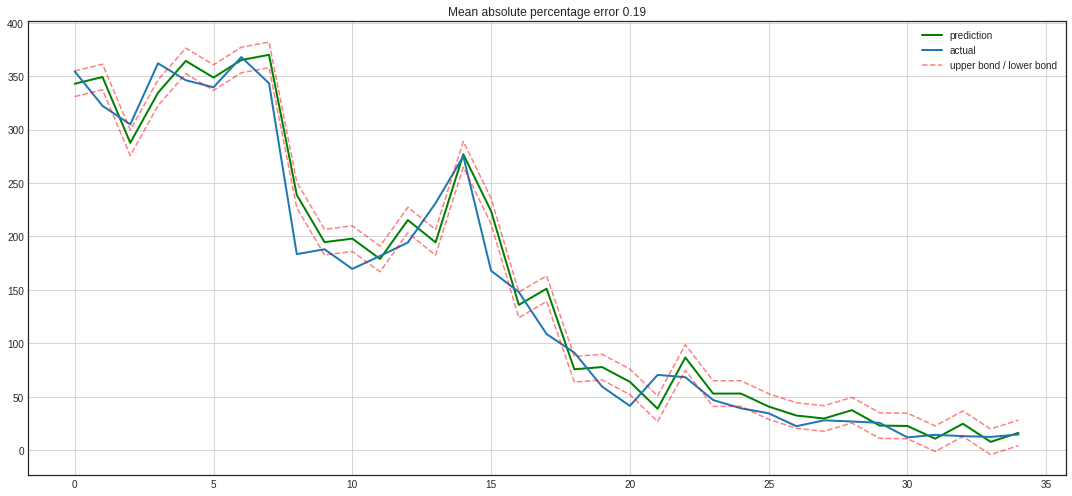


**The heatmap above shows that all of our independent variables are highly connected with one another. However, because this is a small dataset, we can't do much about it because eliminating these features or instances would result in data loss.**

**Modelling**

**Linear Regression:**

Linear regression is one of the most basic and widely used Machine Learning methods. It works best when the dependent and independent variables have a linear relationship.

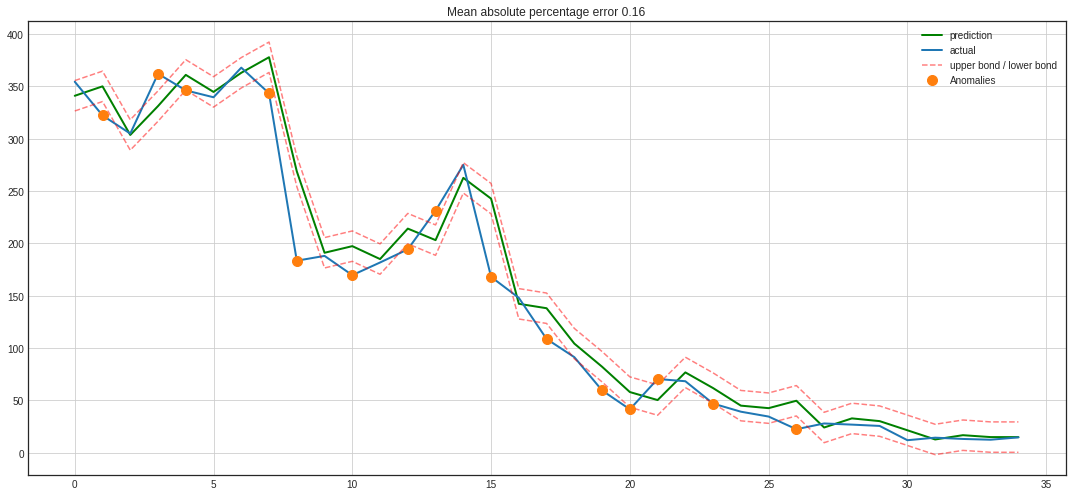


* mean absolute error: 16.93
* mean squared error: 486.35
* root mean squared error: 22.05
* r2\_score: 0.97
* mean absolute percentage error: 0.19

**Lasso Regression (with cross-validation):**

The purpose of lasso regression is to find the subset of predictors that produces the lowest prediction error for a quantitative response variable. This is accomplished by imposing a constraint on the model parameters, causing regression coefficients for some variables to shrink toward zero.

Lasso uses variable selection and regularisation to improve the predictability and interpretability of the final statistical model.

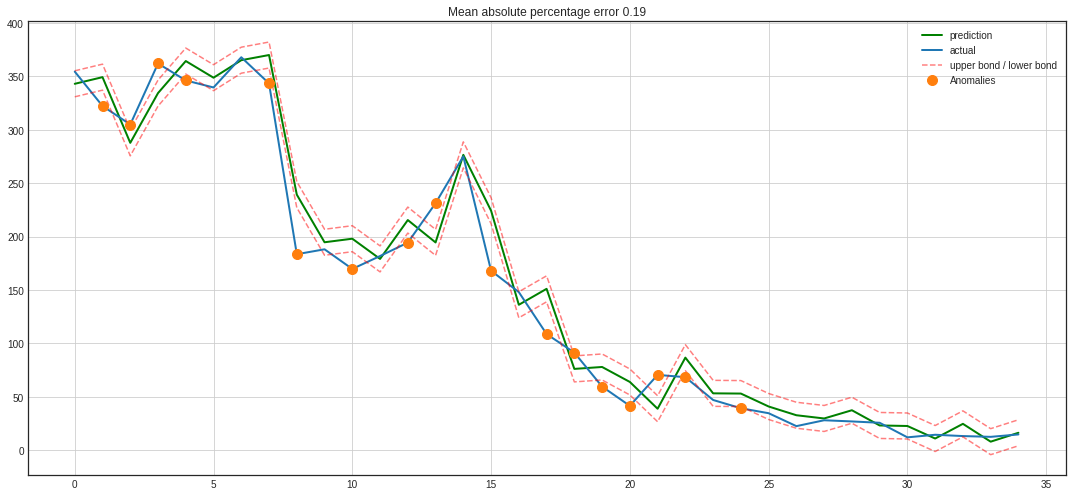


* mean absolute error: 16.87
* mean squared error: 626.48
* root mean squared error: 25.03
* r2\_score: 0.96
* Mean absolute percentage error: 0.16

**Ridge Regression with cross-validation:**

Ridge regression, like lasso, is a regularised linear regression. However, for regularisation, it employs a separate L2 penalty term.It is used to improve the prediction accuracy and interpretability of the final statistical model through regularisation.

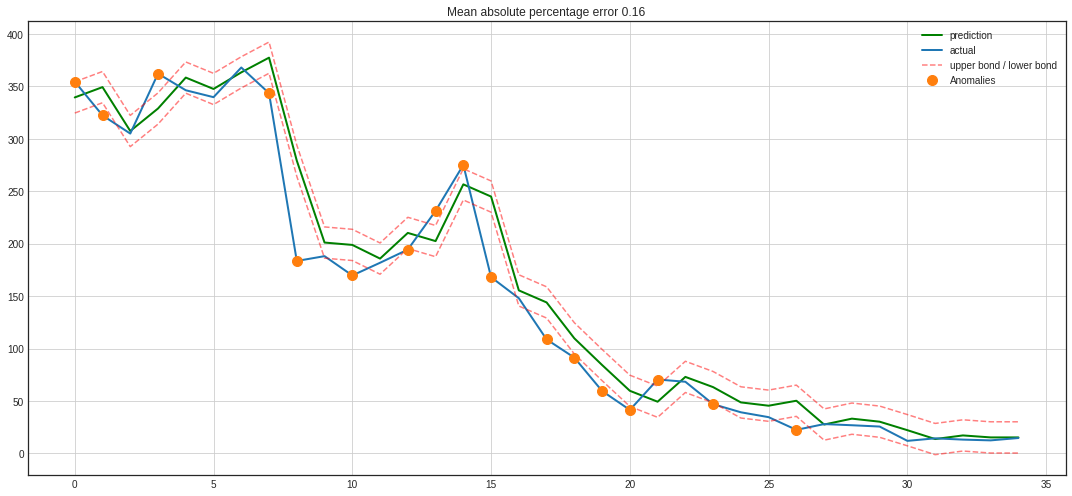
The graph below depicts the model's anticipated and actual values for the target variable.

****

* mean absolute error: 16.93
* mean squared error: 486.35
* root mean squared error: 22.05
* r2\_score: 0.97
* mean absolute percentage error: 0.19

**Elastic Net Regression with cross-validation:**

Elastic net regression combines the best features of lasso and ridge regressions. It adds the penalty terms for regularisation in lasso and ridge(L1 and L2) and uses the sum to regularise. It is used to improve the prediction accuracy and interpretability of the final statistical model through regularisation.

*  mean absolute error: 16.87
* mean squared error: 626.48
* root mean squared error: 25.03
* r2\_score: 0.96
* mean absolute percentage error: 0.16

**CONCLUSION**

* We began with data inspection, then examined the data distribution, looked for correlation, and used averaged characteristics to remove it.
* Accuracy, mean squared error, root mean squared error, r2 score, and mean absolute percentage error were used to evaluate a basic linear regression model.
* Additional features were created by employing delays and regularisation techniques such as ridge, lasso, and elastic net regression to lessen the effect of multicollinearity.
* Regression models such as the random forest regressor, the xgboost regressor, and the support vector regressor were created.
* To do time-series analysis, the time component was introduced and averaging techniques such as moving average, exponentially weighted moving average, and double exponentially weighted moving average were utilised.
* The presence of nonstationarity was recognised, and it was made stationary by taking lags and differences into account.
* Using data visualisation on our target variable, we can clearly see the impact of the Rana Kapoor fraud case in 2018, since stock prices fell drastically at that time period.
* The dependent and independent variables have a significant correlation.
* This indicates that our dependent variable is strongly dependent on our features and may be correctly predicted from them.
* We ran multiple models on our dataset to forecast the closing price and discovered that Elastic Net Regressor is the best performing model, with an Adjusted R2 score of 0.9932 and high scores on all evaluation criteria.
* All of the deployed models performed admirably on our data, with an accuracy of more than 99%.
* We looked for the existence of By graphing the residuals against the Elastic Net model projected value, we discovered that there is no Heterodasceticity in our dataset. Our model performs admirably across all data points.
* We can confidently deploy our model for future predicting tasks utilising future real data because our model predicts with such high accuracy even on unseen test data.
* There are some outliers in our features, but because this is a relatively tiny dataset, deleting those instances will result in information loss.
* We discovered a really high correlation between our independent variables. However, because the dataset is so small, multicollinearity is inescapable.
* We discovered that the distribution of all of our variables is skewed to the right. as a result, we performed log versus transformation.