YES Bank Stock Closing Price Prediction

Sameer Satpute
Data science trainee,
AlmaBetter, Bangalore

Abstract:

Accurate prediction of stock market returns is a very challenging task due to the volatile and non-linear nature of the financial stock markets. With the introduction of artificial intelligence and increased computational capabilities, program methods of different predictions have proved to be more efficient in predicting stock prices. Here I used regression to find relationships. The financial data: Open, High, Low, and Close prices of stock are used for variables that are used as inputs to the model. The models are evaluated using standard strategic indicators: RMSE, MAPE, and R2. The low values of these indicators show that the models are efficient in predicting stock closing price.

Keywords: Stock market price prediction, Algorithm, Strategic indicators

1. Problem Statement

Yes, Bank is a well-known bank in the Indian financial domain. Since 2018, it has been in the news because of the fraud case of Rana Kapoor. To this fact, it was interesting to see how that impacted the stock prices of the company and whether Time series models or different model ANN and 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 for the month. Analyze and check which model performs well on a particular algorithm.

2. Introduction

The stock market is characterized as dynamic, unpredictable, and non-linear in nature. Predicting stock prices is a challenging task as it depends on various factors including but not limited to political conditions, global economy, company's financial reports and performance, etc. Thus, to maximize the profit and minimize the losses, techniques to predict the values of the stock in advance by analysing the trend over the last few years, could prove to be highly useful for making stock market movements. There are two main approaches that have been proposed for predicting the stock price of an organization. Technical analysis method uses the historical price of stocks like closing and opening price, the volume traded, adjacent close values, etc. of the stock for predicting the future price of the stock. The second type of analysis is qualitative, which is performed on the basis of many factors like company profile, market situation, political news, and economic factors, global news, textual information in the form of financial new

articles, social media, and even blogs by the economic analyst. Nowadays, advanced intelligent techniques based on either technical or fundamental analysis are used for predicting stock prices. Particularly, for stock market analysis, the data size is large and also non-linear. To deal with this variety of data efficient model is needed that can identify the hidden patterns and complex relations in this large data set. Machine learning techniques in this area have proved to improve efficiencies by 60-86 percent as compared to the past methods.

2.1 Yes bank stock Dataset

We have been provided with a dataset of the monthly stock price details of Yes Bank. The data has been provided from July 2005 till November 2020. The bank has been making headlines due to its recent default. We analysed the dataset and worked on predicting the stock closing price for the bank using other given parameters.

2.3 Python

Most of the info scientists use python due to the good built-in library functions and therefore the decent community. Python now has 70,000 libraries. Python is a simple programing language for select compared other languages. The most reason data scientists use python more often, for machine learning and data processing data analyst want to use some language which is easy to use. That's one among the most reasons to use python. Specifically, for data scientists, the foremost popular data inbuilt open-source library is named panda. As we've seen earlier in our previous assignment once we got to plot scatterplot, heat maps, graphs, and 3-dimensional data python built-in library comes very helpful choose to wait a few minutes to see if the rates go back down.

3. Steps involved:

- **Exploratory Data Analysis**
  After loading the dataset I performed the method by comparing our target variable which is the closing price of the stock with other independent variables open, low, and high prices of stocks. This process helped us figure out various aspects and relationships between the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.
- **Null values Treatment**
  In the dataset yes bank stock price there is no any null value present also no any missing value I can see there.

- **Establish basic observation**
  The basic observation I find out that highest price of stock around 350-360 Rs. After sudden fraud case news yes bank consecutive fall I see from 350 to 14-15 Rs. More seller we expect because of fraud case

news. When news disturb stock prices that time high and low value difference increases because of volatility in that stock price.

- **Data mining**

  Data mining process of extracting discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, database systems. From dataset I derive some meaningful insights from closing price of stock. Data mining technique have been effectively revealed to produce high forecasting accurateness of the movement of stock price of bank. Now a days,. Data mining methods have been introduced for forecasting of movement indication of stock market index. Data mining techniques have a more successful act predicting various fields such policy, economy and engineering compared to usual statistical techniques by discovering unknown information of data .

- **Standardization of features**

  Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.

  The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

- **Fitting different models**

  For modelling we tried various classification algorithms like:

  - ❖ Linear Regression
  - ❖ Lasso Regression
  - ❖ Ridge Regression
  - ❖ Elastic Net Regression
  - ❖ KNN
  - ❖ SVM

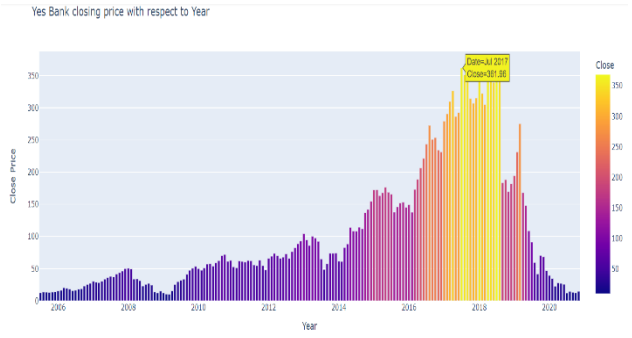
- **Cross-Validation on model**

  Cross-Validation is a resampling technique that helps to make our model sure about its efficiency and accuracy on the unseen data. It is a method for evaluating Machine Learning models by training several other Machine learning models in subsets of the available input data set and evaluating them on the subset of the data set.

3.1 EDA

For visualization we firstly check closing price of stock with time.

3.1.1 Closing price vs time



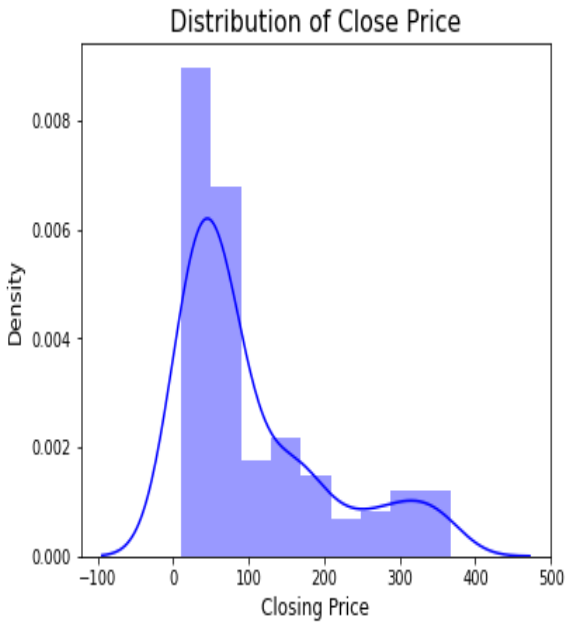
Stock price shows upside from 2014 but after fraud news case price fallen in 2018.

3.1.2 Combination plot

We combine open, high, low and close price vs time in single plot

After applying log transformation it looks like normally distributed

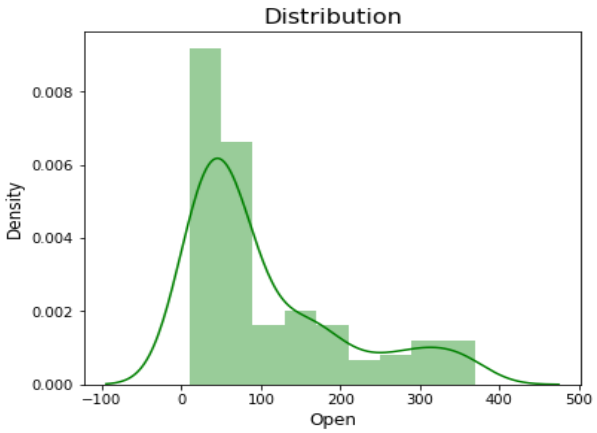
3.1.4 Distribution of close price

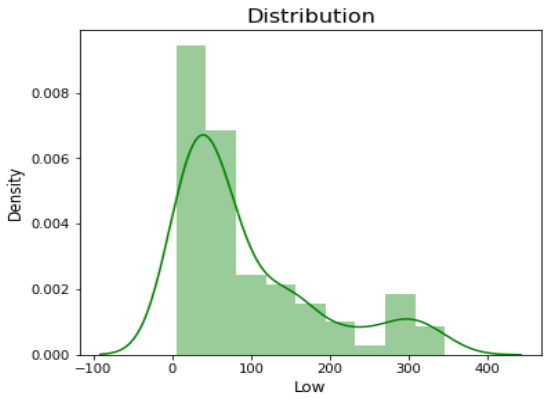


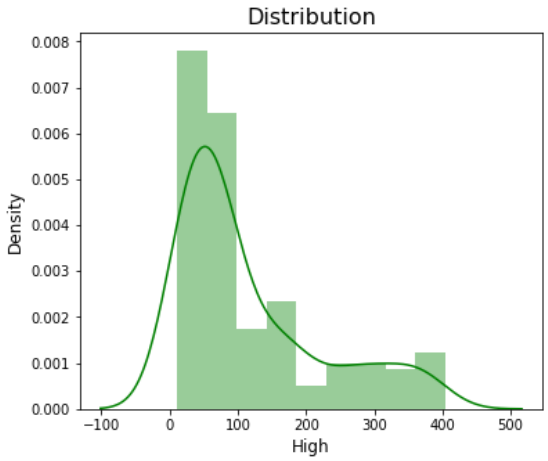
This is distribution of close price here we can see Positive skewed distribution.

Distribution of high, low, open price

All high low open price is distributed rightly skewed and all need to use log transform to make it in normally distributed.



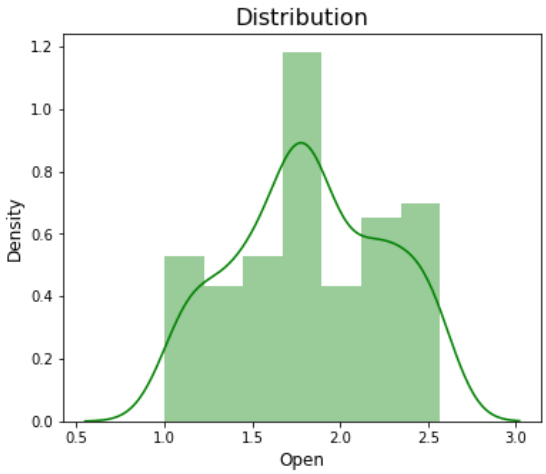


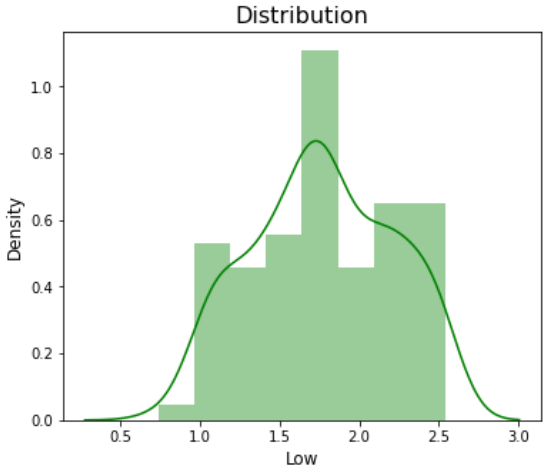


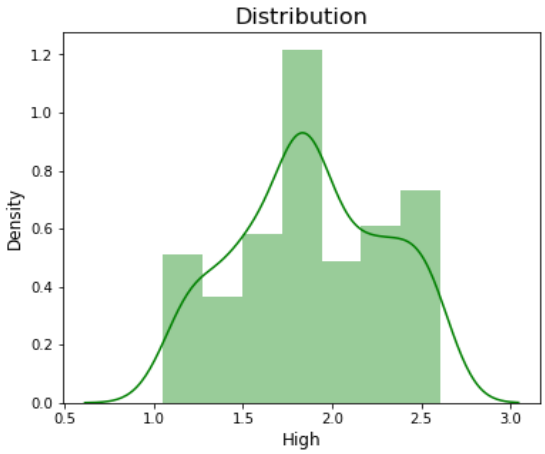
Transformation

Below are some types of methods or ways to deal above type of problem.

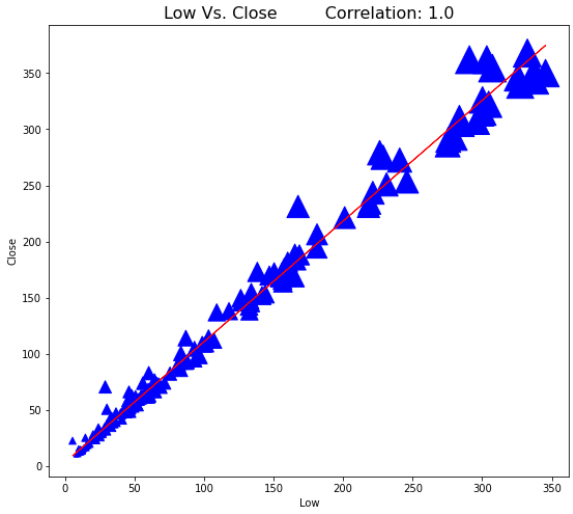
- **Square-root for moderate skew:** sqrt(x) for positively skewed data, sqrt(max(x+1) - x) for negatively skewed data
- **log for greater skew:** log10(x) for positively skewed data, log10(max(x+1) - x) for negatively skewed data
- **Inverse for severe skew:** 1/x for positively skewed data 1/(max(x+1) - x) for negatively skewed data
- **Linearity and heteroscedasticity:** First try log transformation in a situation where the dependent variable starts to increase more rapidly with increasing independent variable values If your data does the opposite – dependent variable values decrease more rapidly with increasing independent variable values – you can first consider a square transformation.



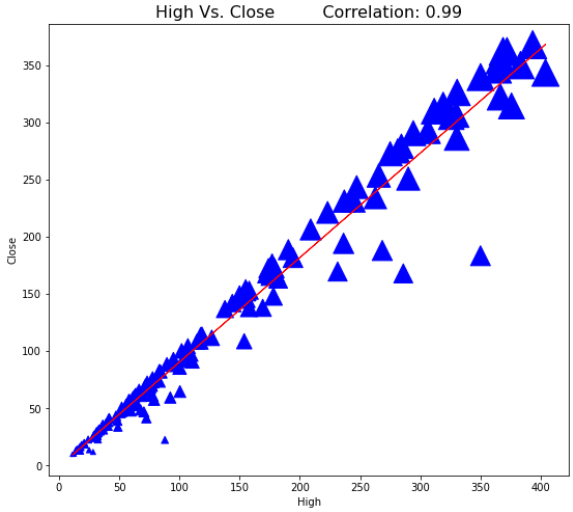


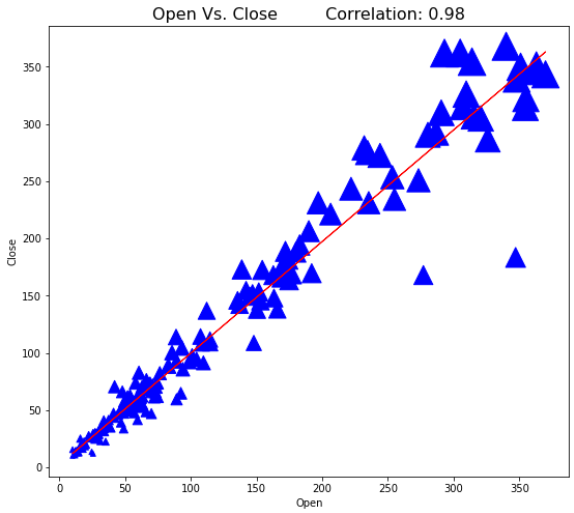


3.1.6 Best fit line and correlation



Low price and close price see highly correlated with 1.0



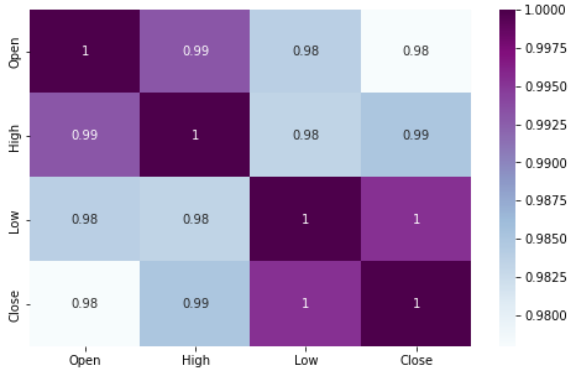


Open price and high price of stock also shows highly correlated.

3.1.6 Correlation

Correlation refers to a process for establishing the relationships between two variables.

Positive correlation: A positive correlation would be 1. This means the two variables moved either up or down in the same direction together.



here are very high correlation between independent variables which lead us to multicollinearity. High multicollinearity is

not good for fitting model and prediction because a slight change in any independent variable will give very unpredictable results.

To check multicollinearity and how much it is in our dataset, we have to calculate VIF(Variation Inflation Factor).

Variation Inflation Factor:

	Variables	VIF
0	Open	175.185704
1	High	167.057523
2	Low	71.574137

- In general case, Any variable having VIF above 5 is considered to be multicollinear.
- The thumb rule is to drop the highest VIF variable.
- However, you may choose to select the variable to be dropped based on business logic
- Here all feature are equally important and we have very limited features.

3.2 Data Transformation

Splitting data

X = Independent variable

Y = Dependent variable

Splitting train-test data with 80-20

data must be normally distributed before apply normalization..

Normalization is one of the feature scaling techniques. We particularly apply normalization when the data is skewed on the either axis i.e. when the data does not follow the Gaussian distribution. In normalization, we convert the data feature of different scales to common scale which further makes it easy for data to be processed for modelling. Thus, all the data features tend to have a similar impact on the modelling portion.

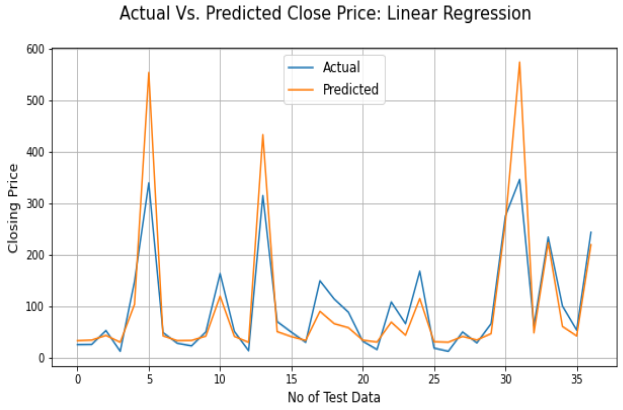
4.1 Algorithms

Its time to apply different models on given dataset as follows.

1) Linear Regression

Linear regression is one of the easy and popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous or numeric variables such as **sales, salary, age, product price,** etc.

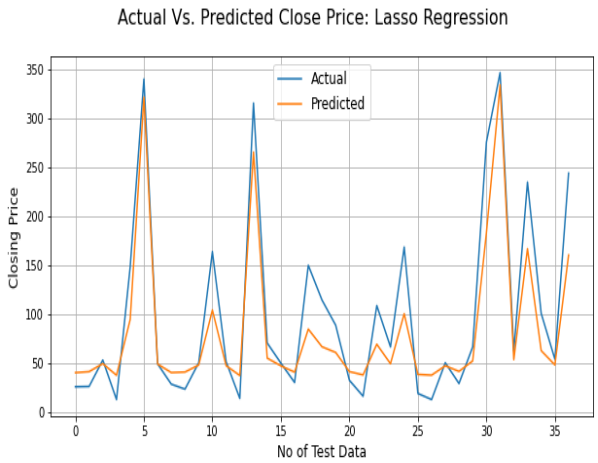
Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.



2) Lasso Regression

Lasso regression is linear regression, but it uses a technique **"shrinkage"** where the coefficients of determination shrink to towards **zero**. Linear regression gives you regression coefficients as observed in the dataset. The lasso regression allows to shrink or regularize coefficients to avoid overfitting and make them work better on different datasets.

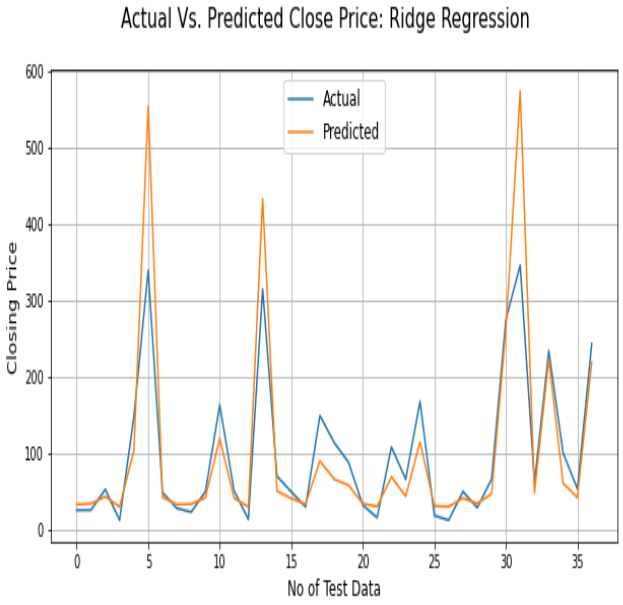
This type of regression is used when the dataset shows high multicollinearity or when you want to automate variable elimination and **feature selection**.



3) Ridge Regression

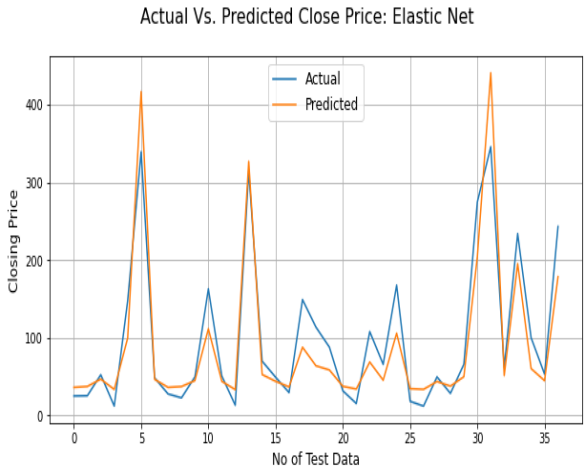
Ridge regression is a regularized version of linear least squares regression. It works by shrinking the coefficients or weights of the regression model towards zero. This is achieved by imposing a squared penalty on their size.

This is one of the method of regularization technique which the data suffers from multicollinearity. In this multicollinearity ,the least squares are unbiased and the variance is large and which deviates the predicted value from the actual value. Equation have an error term.



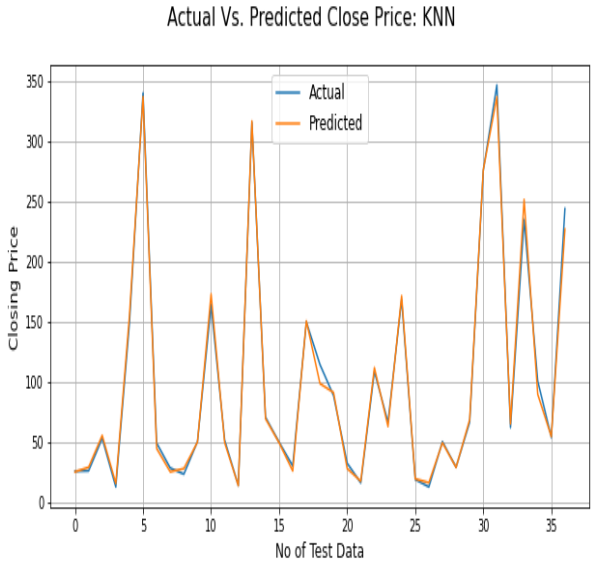
4) Elastic Net Regression

Elastic Net Regression is the third type of Regularization technique. It came into existence due to the limitation of the Lasso regression. Lasso regression cannot take correct alpha and lambda values as per the requirement of the data. The solution to the problem is to combine the penalties of both ridge regression and lasso regression.



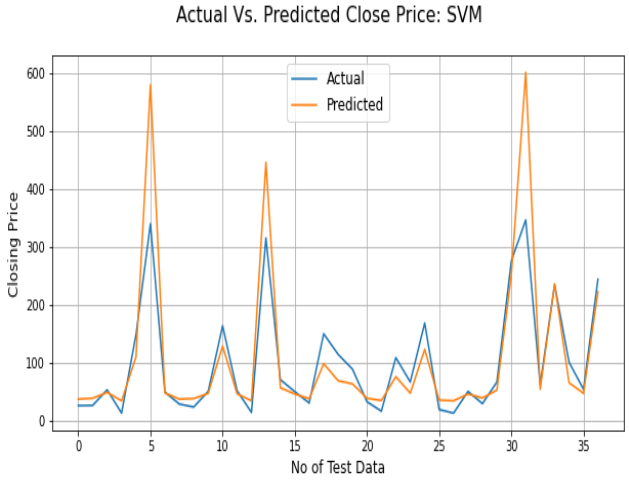
5) K Nearest Neighbours (KNN)

K-NN stands for K-Nearest Neighbours. It is an algorithm used for the prediction of a continuous variable. A non-parametric and a prediction problem; it does not care about the relationship between the predictor x the response variable y. It takes k nearest neighbours whose distances form that point are minimum and computes the average of those values.



6) Support Vector Regressor (SVR)

Support Vector regression is a type of Support vector machine that supports linear and non-linear regression. As it seems in the below graph, the mission is to fit as many instances as possible between the lines while limiting the margin violations.



4.2 Hyper parameter tuning

Parameter- A model parameter is a configuration variable that is internal to the model and whose value can be estimated from the given data.

Hyper-parameter- A model hyperparameter is a configuration that is external to the model and whose value cannot be estimated from data.

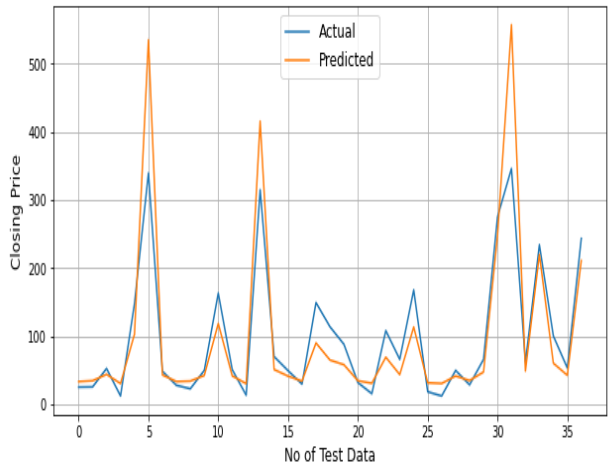
Grid Searching of hyperparameter- Grid search is an approach to hyper parameter tuning that will methodically build and evaluate a model for each combination of algorithm parameters specified in a grid.

Grid Search combines a selection of hyper parameters established by the scientist and runs through all of them to evaluate the model's performance. This is a simple technique that will go through all the programmed combinations. The biggest disadvantage is traverses a specific region of the parameter space and not understand which movement or which region space is important to optimize the model.

Cross validation on Lasso regression:

Actual Closing Price	lasso Predicted Closing Price
25.32	33.471548
25.60	34.648004
52.59	43.984987
12.26	30.530694
147.95	102.907521

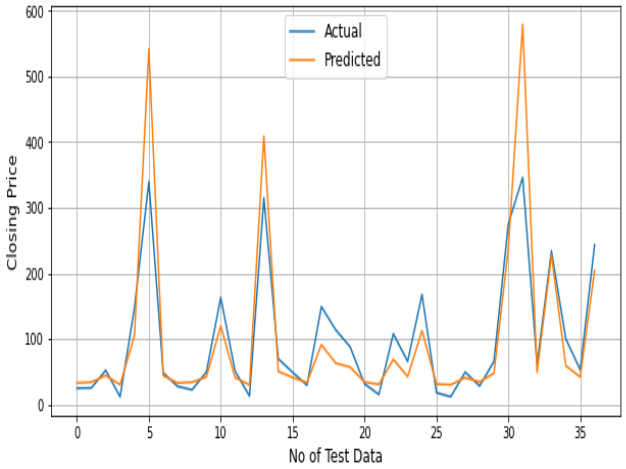
Actual Vs. Predicted Close Price: Lasso Regression After CV



Cross validation on Ridge regression:

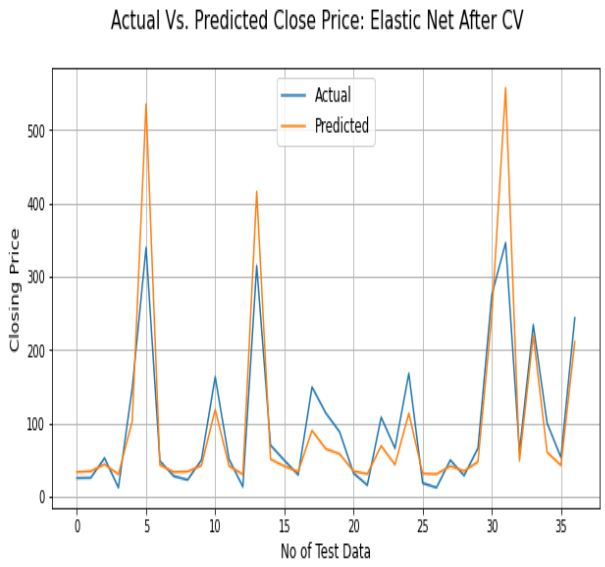
Actual Closing Price	Ridge Predicted Closing Price
25.32	33.214715
25.60	34.457302
52.59	44.607474
12.26	30.472487
147.95	105.605617

Actual Vs. Predicted Close Price: Ridge Regression After CV



Cross validation on Elastic Net regression:

Actual Closing Price	Elastic net Predicted Closing Price
25.32	33.471548
25.60	34.648004
52.59	43.984987
12.26	30.530694
147.95	102.907521



5.Conclusion

Evaluation Metrics Comparison:

	Model	MSE	RMSE	MAE	MAPE	R2
5	Lasso	0.0436	0.2088	0.1672	0.1099	0.7550
4	ElasticNet	0.0364	0.1908	0.1574	0.1024	0.7955
3	SVR	0.0347	0.1864	0.1489	0.0976	0.8048
2	Ridge	0.0316	0.1777	0.1513	0.0954	0.8225
1	LinearRegression	0.0316	0.1777	0.1513	0.0954	0.8226
0	KNeighborsRegressor	0.0015	0.0389	0.0274	0.0182	0.9915

- Target variable(dependent variable) strongly dependent on independent variables
- There is increase in trend of Yes Bank's stock till 2018 an then sudden decrease.
- We saw Linear relation between the dependent and independent value.
- KNeighborsRegressor giving the highest R squared value. The predicted values are nearly equal to the actual values. We got 99% accuracy.
- Linear regression and Ridge regression get almost same R squared value
- Whereas Lasso model shows lowest R squared value and high **MSE,RMSE,MAE,MAPE**

6. Reference

I. Stack overflow
II. GeeksforGeeks
III. Jovian
IV. Research paper based on Stock price prediction using ANN
V. Analytics Vidhya
VI. Towards data science