Capstone Project -2

Yes bank Stock Closing Price Prediction

Done By:

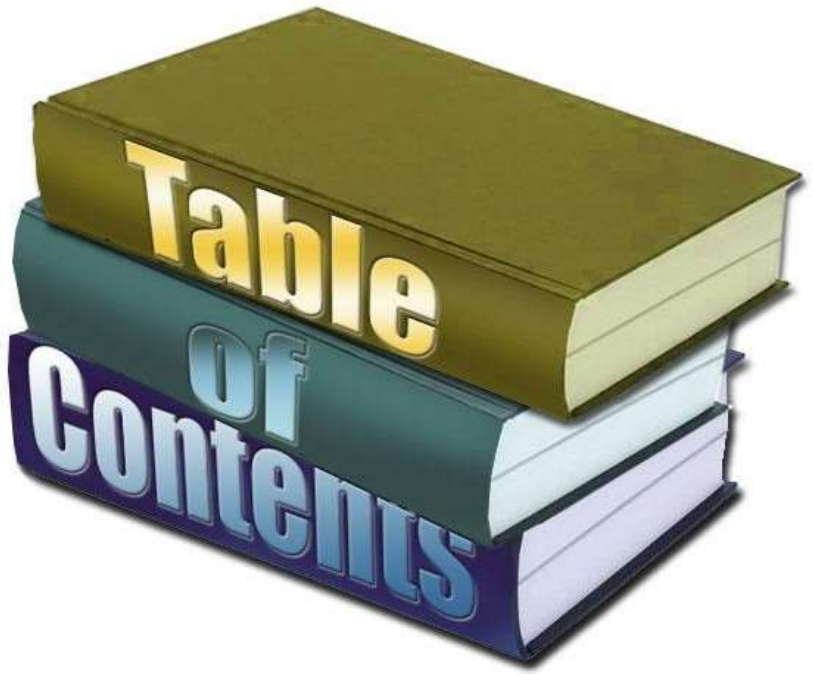
Saquib Neyaz

Pranali Dongre

Rajni Shukla

Contents

1. Problem Statement
2. Introduction
3. Data Pipeline
4. Data Cleaning
5. Exploratory Data Analysis (EDA)
6. Transforming Data
7. Splitting Data In Train & Test
8. Fitting Different Model
9. Cross Validation & Hyperparameter Tuning
10. Conclusion



Problem Statement

- Perform regression analysis on Yes Bank Stock Price dataset using multiple models to predict the closing price of Yes Bank stock at the end of every month and compare the evaluation metrics for all of them to find the best model.
- Prediction of Yes Bank stock closing price.
- Getting accuracy score of several machine learning model.

Introduction

- Data set of Yes Bank Stock Prices contains observations regarding open,close, high and low prices of the yes bank stock from July 2005 - November 2020.

- Date: Monthly observation of stock prices since its inception.
- Open: The price of a stock when stock exchange market open for the day.
- Close: The price of a stock when stock exchange market closed for the day.
- High: The maximum price of a stock attained during given period of time.
- Low: The minimum price of a stock attained during given period of time.

Data Pipeline

Data Preprocessing : At this Stage,

- we check for duplicate values and missing values and treat them if any.
- Detecting the outliers and removed it.
- We check the datatype of the features present in our dataset, transform them if necessary.

Exploratory Data Analysis(EDA): At this stage, we conduct an EDA on the selected features in order to better understand their spread, pattern and relationship with the other features. It gives us an intuition as to what is going on in the dataset.

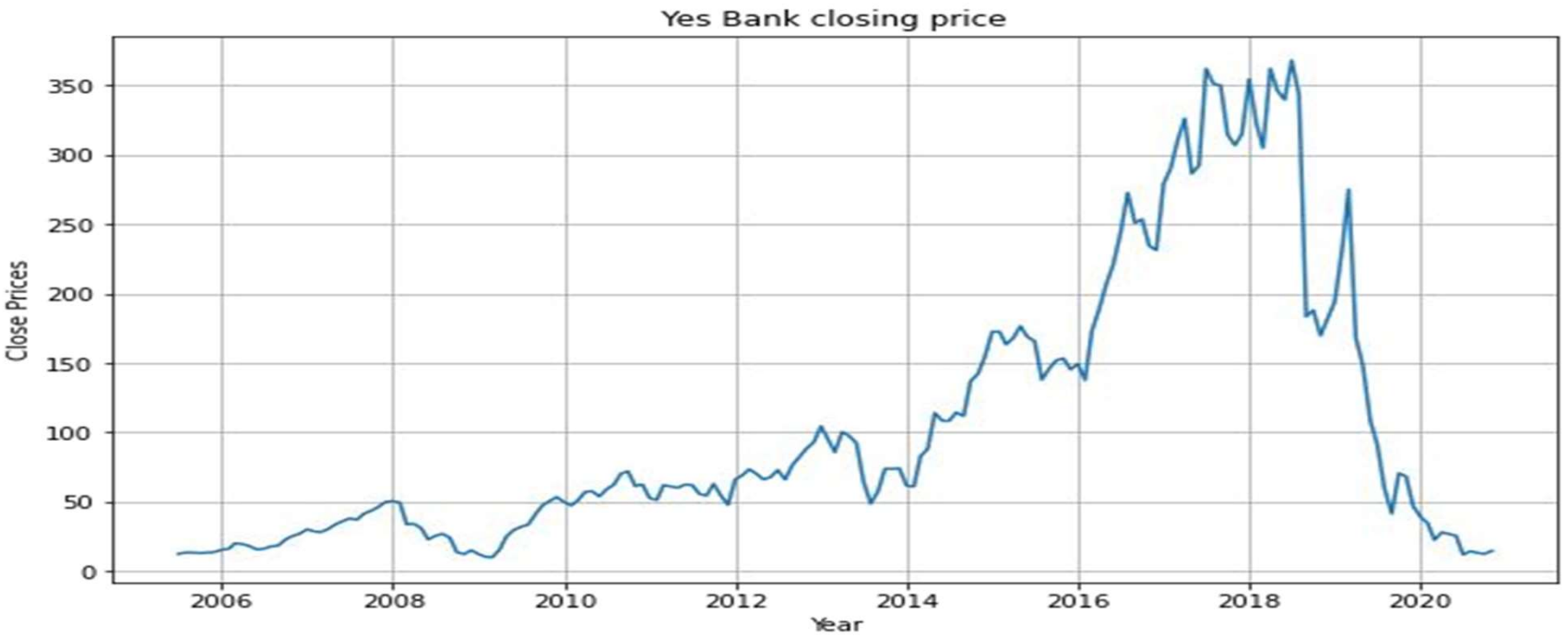
Model Building : At this stage, we apply various models to understand which one will give us the best result.

Data Cleaning

- Brief summary of the dataset
- Null Values Treatment
- Duplicated Values Treatment
- Date Format Change (i.e from Jul-05 to 2005-07-01)
- Checking outliers
- So after successfully cleaning the dataset we have 5 columns and 185 rows in the dataset

Exploratory Data Analysis

Visualising The Data



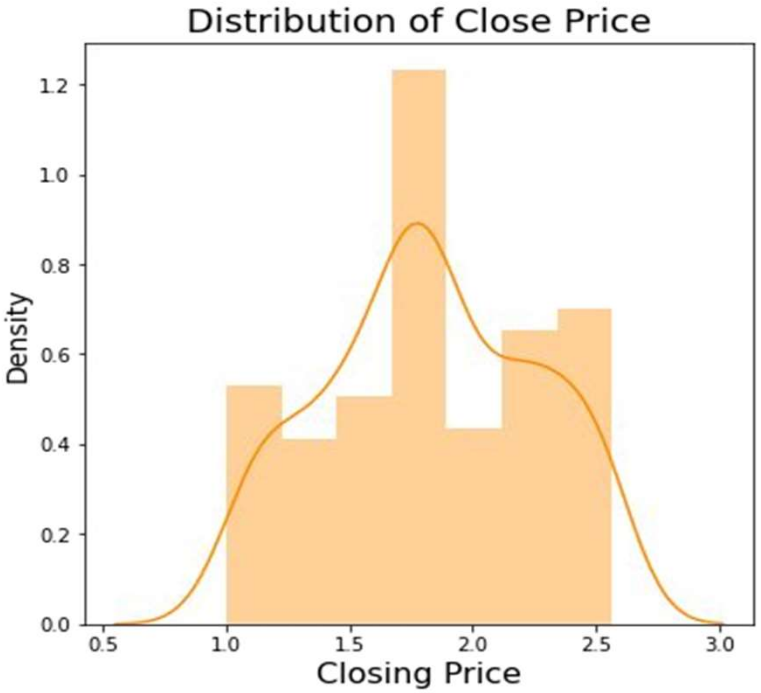
This plot of Closing prices of different dates give a very vivid picture of fluctuation in prices regarding different time-duration. After 2018 there is sudden fall in the stock closing price. It makes sense how severely Rana Kapoor case fraud affected the price of Yes bank stocks.

EDA (Continued…)

- **Distribution of Closing Price :**

- Distribution of closing price is right skewed.
- We need this distribution to be normal distribution for training algorithm.

- **After Log Transformation :**

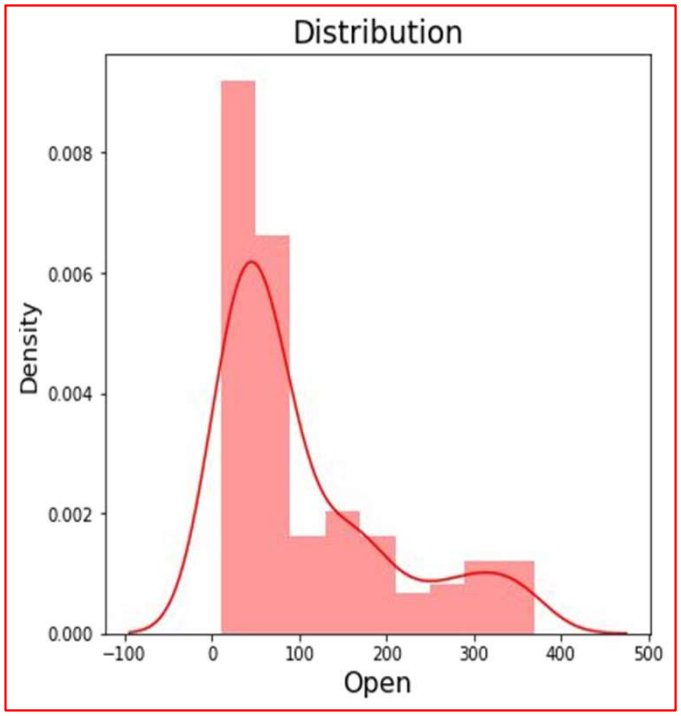


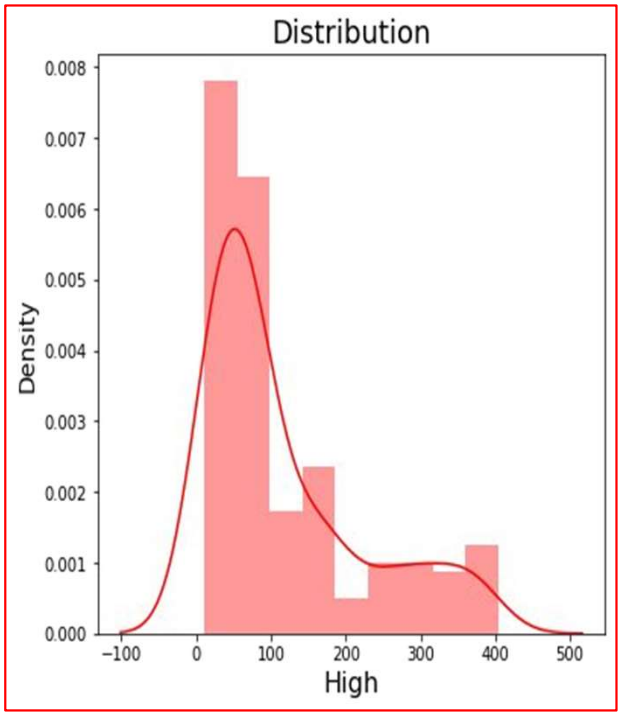
- Distribution of closing price is normal distribution.

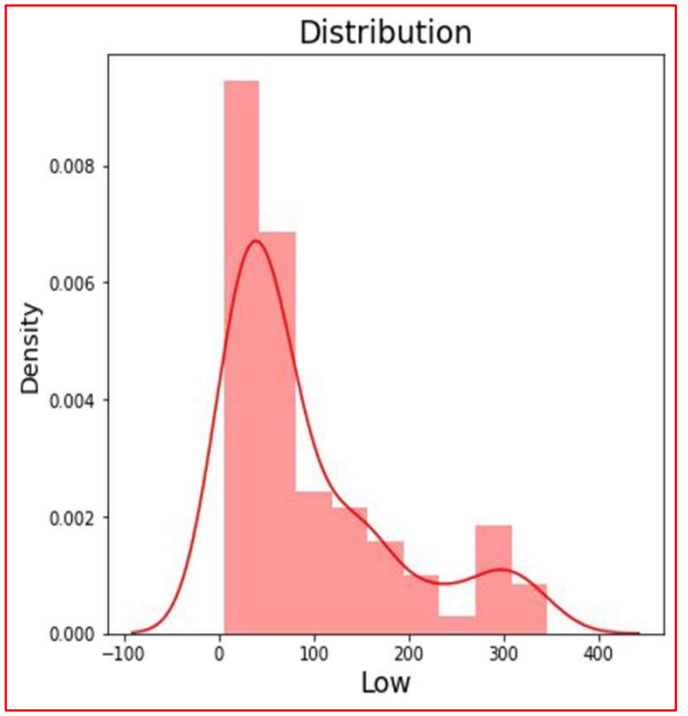
EDA (Continued…)

AI

- **Distribution of Open, High & Low Price of a stock :**





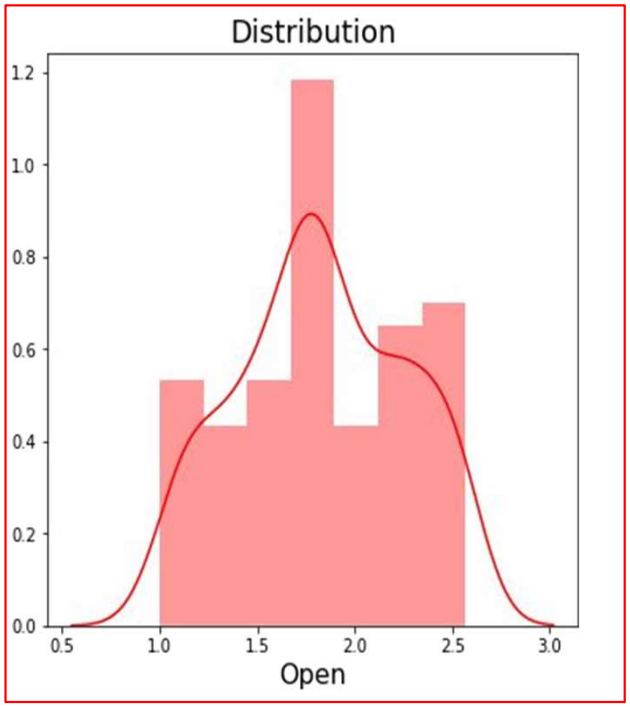


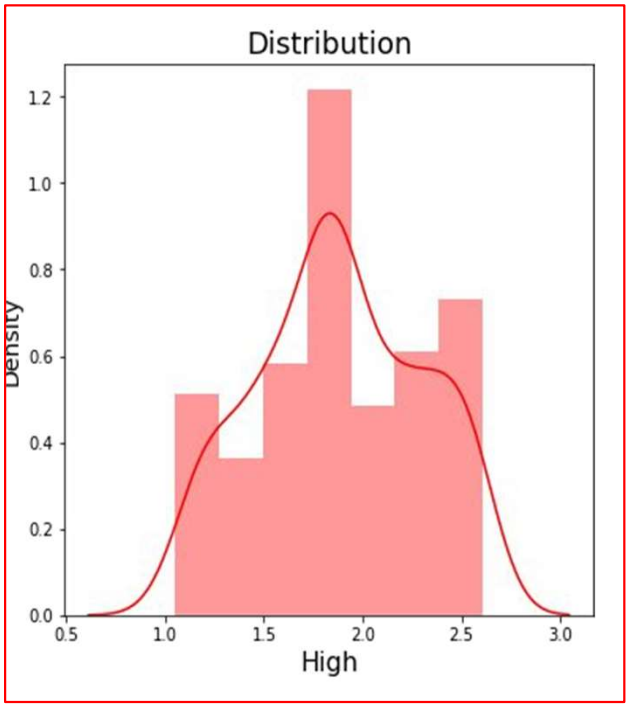
- Distribution of opening price, high price and low price are also right skewed.
- Log transformation applied to make this distribution normal.

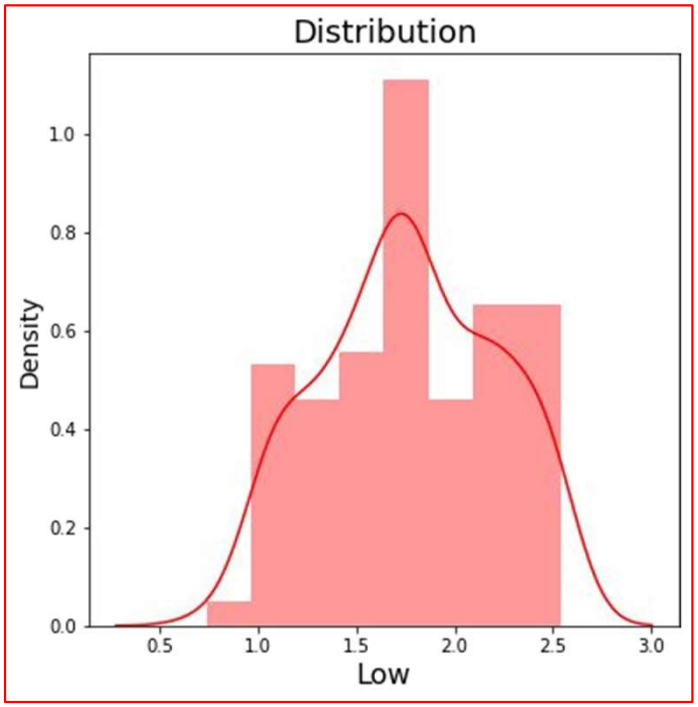
EDA (Continued..)

AI

- **Distribution of Open, High & Low Price of a stock after Log Transformation :**



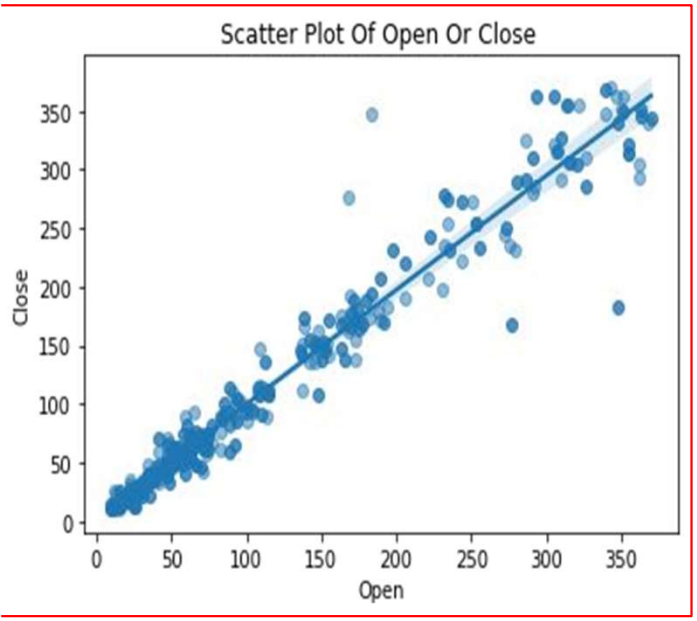


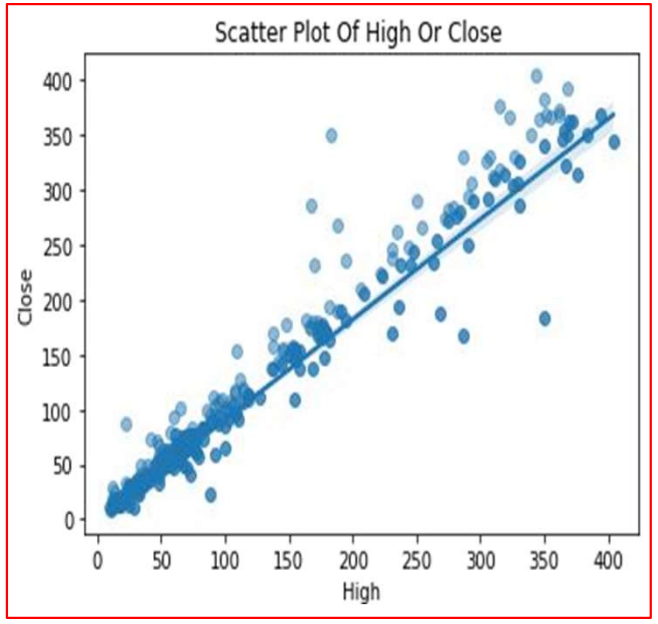


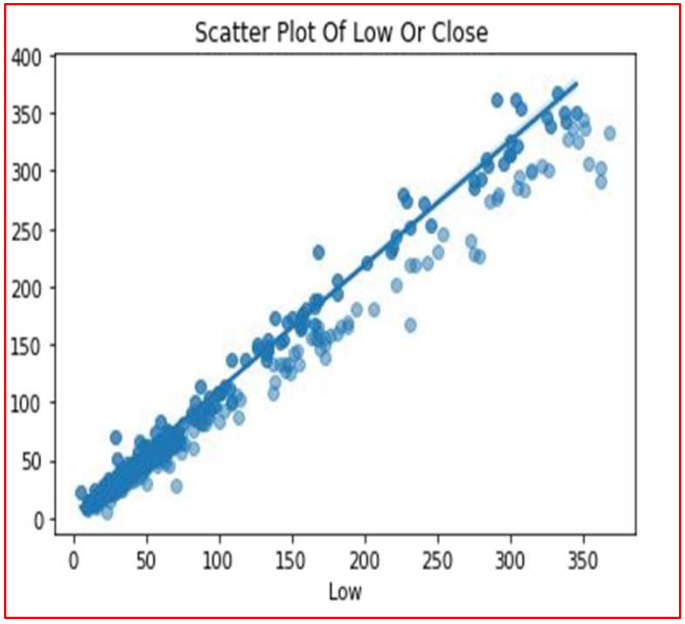
- Distribution of opening price, high price and low price are now normal distribution.

Bivariate Analysis Plots

scatter plot to see the relationship between dependent & independent variables

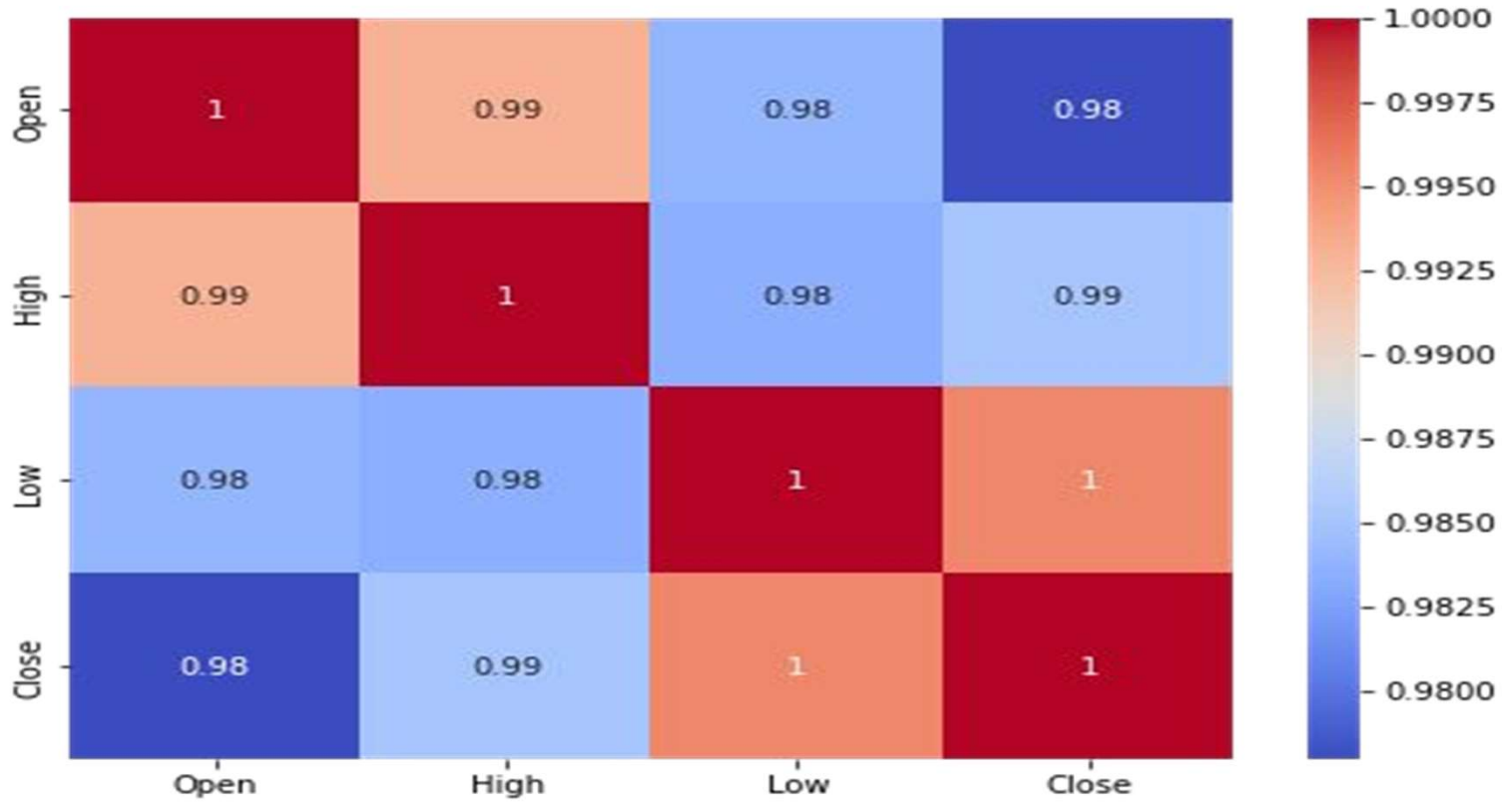






EDA (Continued…)

- **Correlation :**



- All the features are strongly correlated with each other.

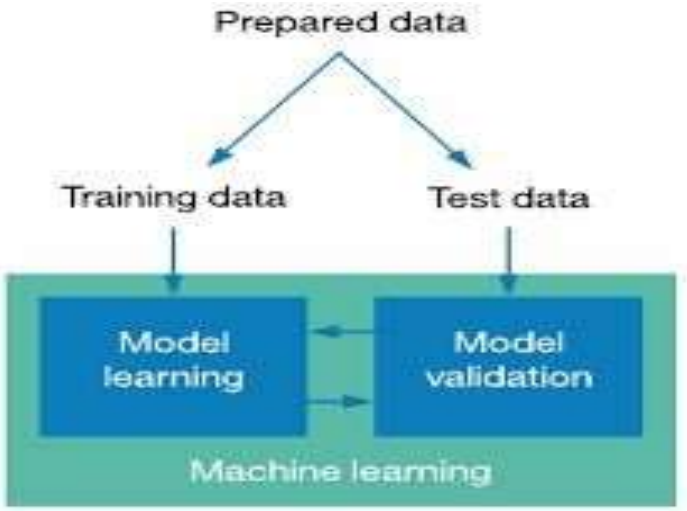
Transformation of Data

- To scale data into a uniform format that would allow us to utilize the data in a better way.
- For performing fitting and applying different algorithms to it.
- The basic goal was to enforce a level of consistency or uniformity to dataset.

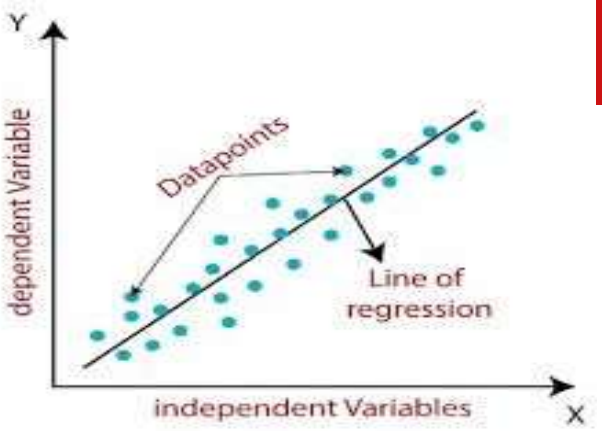
Splitting data in Train and Test

AI

- Data splits into training dataset and testing dataset.
- Training dataset is for making algorithm learn and train model.
- Test dataset is for testing the performance of train model.
- Here 80% of data taken as training dataset & remaining 20% of dataset used for testing purpose.

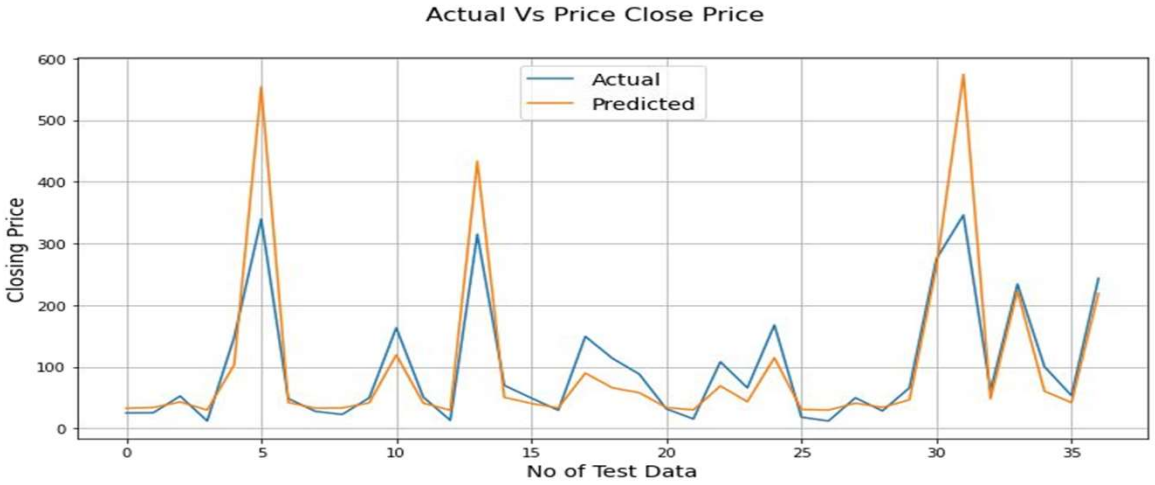


Fitting Different Model



Linear Regression

- Linear regression is one of the easiest and most popular Machine Learning algorithms.
- It is a statistical method that is used for predictive analysis.
- Linear regression algorithm shows a linear relationship between a dependent and independent variable; hence it is called as linear regression.

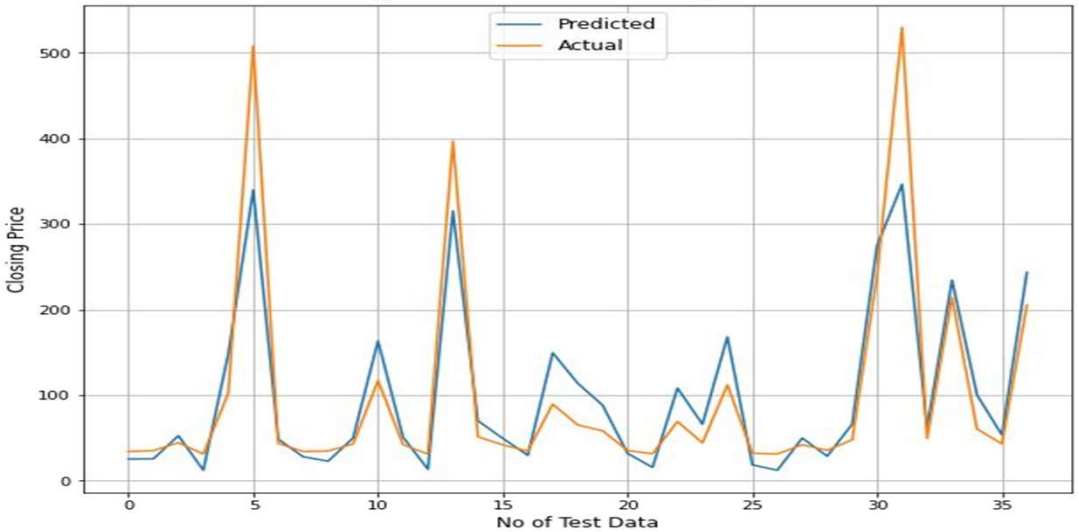


Evaluation Metrics: Linear Regression				
MSE	RMSE	MAE	MAPE	R2
0.0316	0.1777	0.1513	0.0954	0.8226

Lasso Regression

AI

- Lasso: Least Absolute Shrinkage and Selection operator
- It is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.
- This method performs L1 regularization.

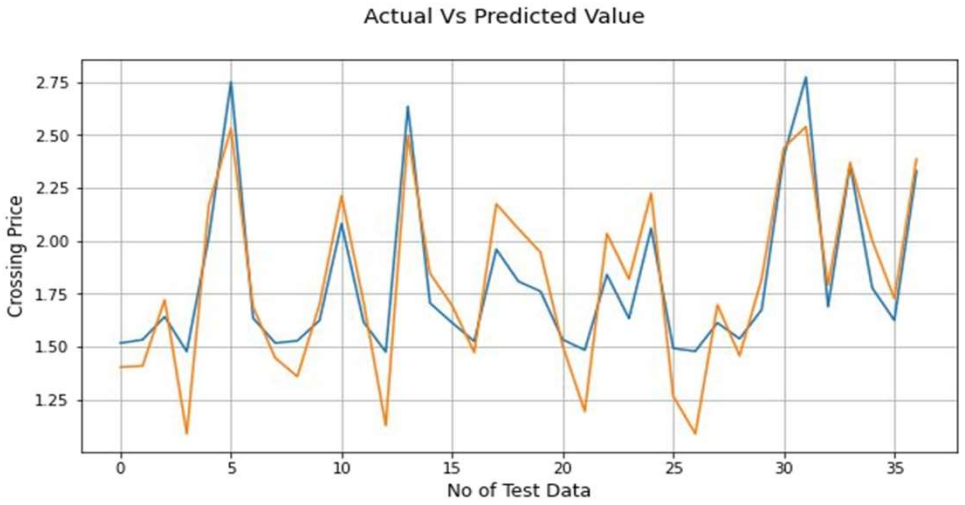


Evaluation Metrics: Lasso Regression				
MSE	RMSE	MAE	MAPE	R2
0.0326	0.1806	0.1535	0.0978	0.8168

Ridge Regression

AI

- Ridge regression is a model tuning method that is used to analyses any data that suffers from Multicollinearity.
- 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.
- This method performs L2 regularization.



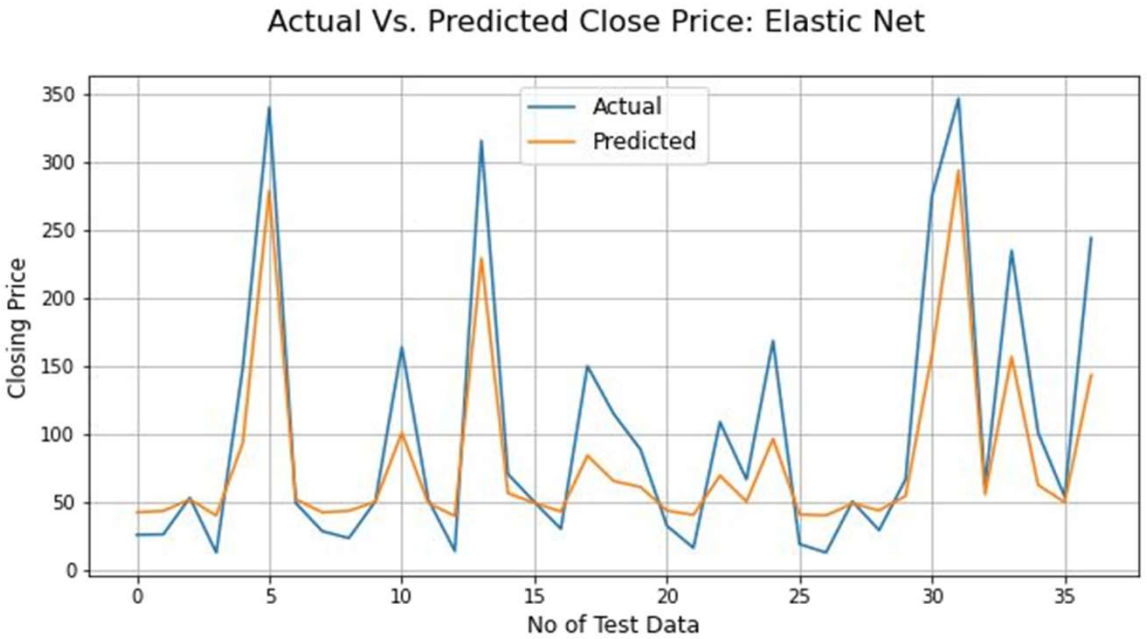
Evaluation Metrics: Ridge Regression				
MSE	RMSE	MAE	MAPE	R2
0.0321	0.1791	0.1522	0.0959	0.8197

Elastic Net

AI

- Elastic net is a popular type of regularized linear regression that combines two popular penalties,specifically the L1 and L2 penalty functions.

- Elastic Net is an extension of linear regression that adds regularization penalties to the loss function during training.

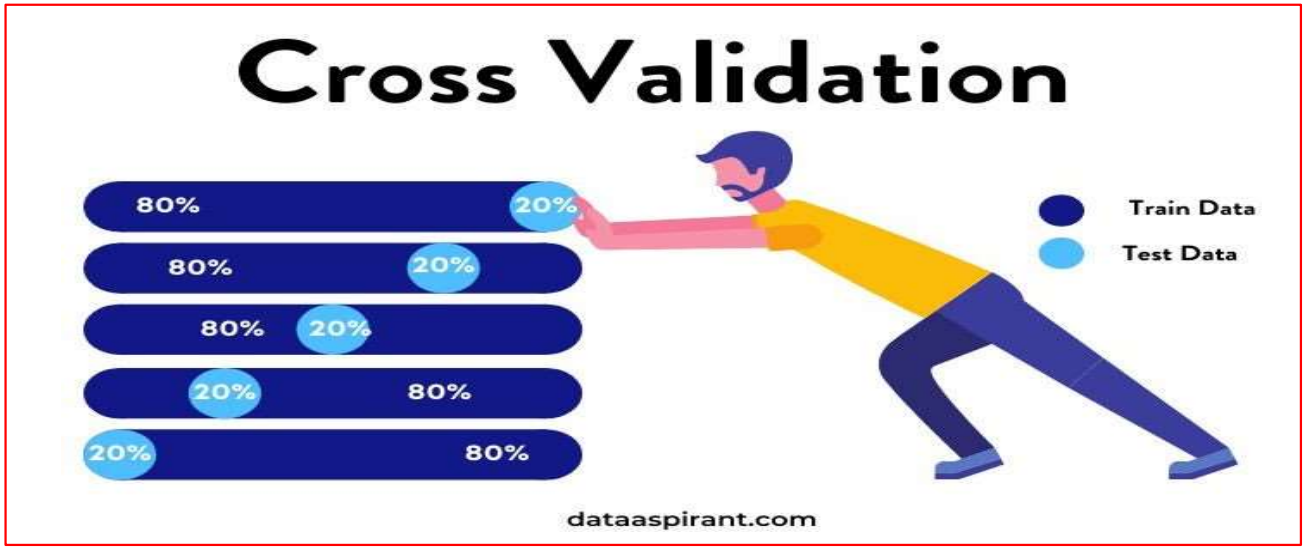


Evaluation Metrics: Elastic Net				
MSE	RMSE	MAE	MAPE	R2
0.051	0.226	0.182	0.118	0.714

Cross Validation & Hyperparameter Tuning

AI

- It is a resampling procedure used to evaluate machine learning models on a limited data sample.
- Basically, Cross Validation is a technique using which Model is evaluated on the dataset on which it is not trained that is it can be a test data or can be another set as per availability or feasibility.
- Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting.



- **Cross Validation & Hyperparameter tuning on Lasso Regression**

Evaluation Metrics :- CV & tuning on Lasso Regression				
MSE	RMSE	MAE	MAPE	R2
0.0326	0.1806	0.1535	0.0978	0.8168

- **Cross Validation & Hyperparameter tuning on Ridge Regression**

MSE	RMSE	MAE	MAPE	R2
0.0327	0.1808	0.1534	0.0971	0.8164

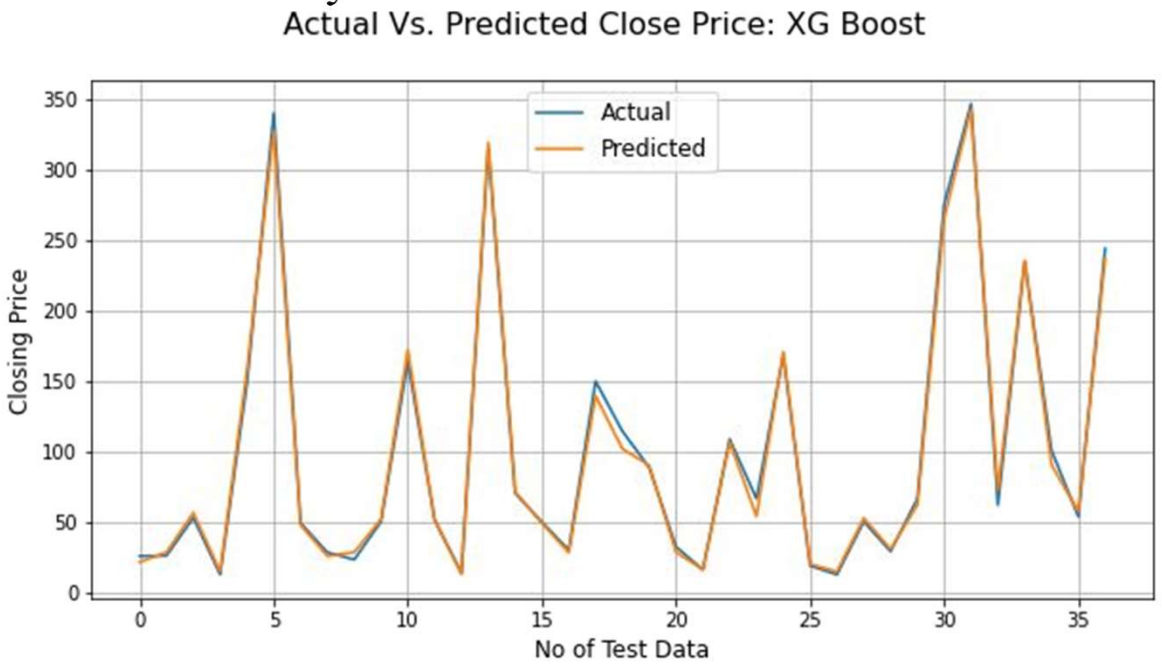
- **Cross Validation & Hyperparameter tuning on Elastic Net**

MSE	RMSE	MAE	MAPE	R2
0.032	0.1789	0.1522	0.0961	0.8202

XGBoost Regressor

AI

XGBoost stands for “Extreme Gradient Boosting”.XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning algorithms under the Gradient Boosting framework. It provides a parallel tree boosting to solve many data science problems in a fast and accurate way.



Evaluation Metrics: XGBoost Regression				
MSE	RMSE	MAE	MAPE	R2
0.0016	0.0394	0.0303	0.0196	0.9913

Conclusion

AI

- The target variable is highly dependent on input variables
- Close, Open and high price of stock are strongly correlated with each other.
- This technique is used for prediction is not only helpful to researchers to predict future stock closing prices or any fraud happen or not but also helps investors or any person who dealing with the stock market in order to prediction of model with good accuracy.
- In this work we use linear regression technique, lasso regression, ridge regression, elastic net regression and XGBoost Regression technique. these five models gives us the following results

a) Linear, lasso and ridge regression show almost same R squared values.
b) Independent variables (High, Low and Open) are directly correlated with Dependent variable (Closing Price)
c) Xgboost regression results as best model for yes bank stock closing price data with very less mean square error i.e. 0.0016

THANK YOU