Capstone project-2

Yes Bank Stock Closing Price Prediction

Team- Incredible data Scientist

Team Members

Ranjit Biswal

Suvendu Dey

Abhishek Kumar

Contents

Introduction

To determine the YES bank's stock's future value on the national stock exchange by making machine learning model of liner regression.

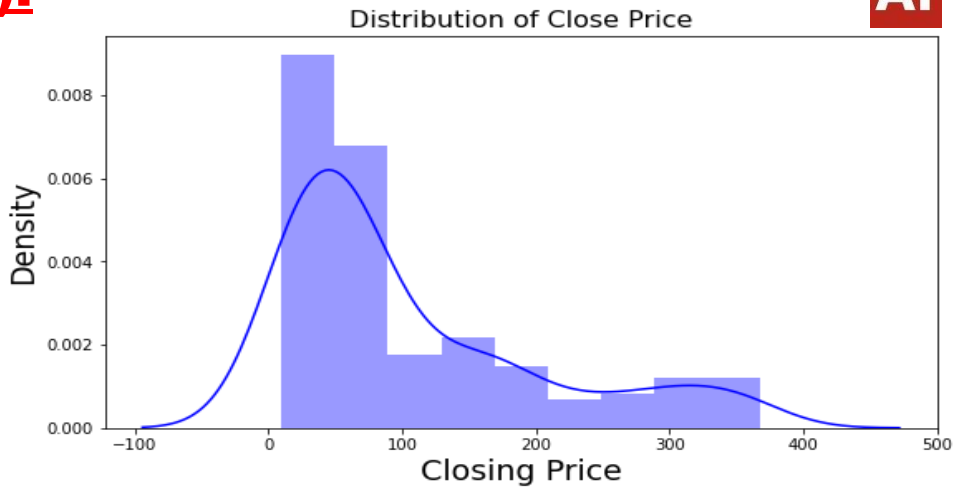
The advantage of successful prediction of stocks future price could result insignificant profit. The efficient market hypothesis recommends that stock costs mirror all right now accessible data and any value changes.

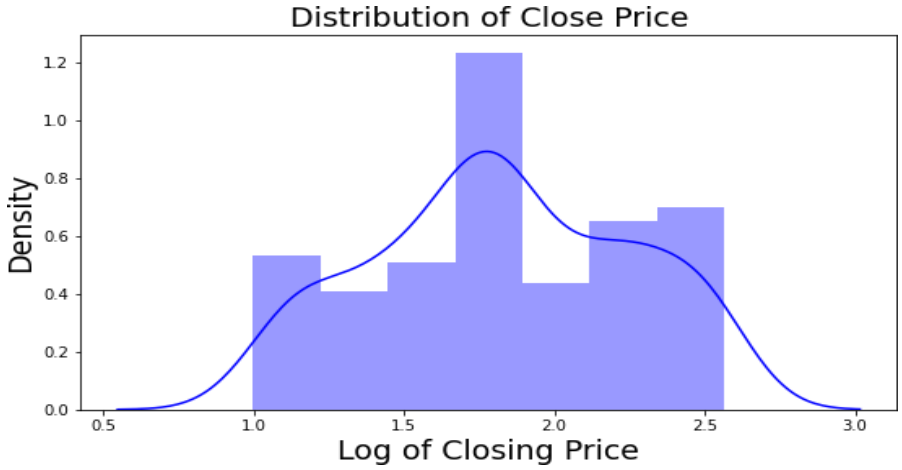
Exploratory Data Analysis (EDA):

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

We can see that the given dataset of Yes Bank Stock prices is not normally distributed, it looks like right skewed, which makes our regression model difficult to learn the pattern of the dataset.

So, we need to apply some transformation technics like log- transformation, Sqrt transformation... etc. To make the dataset normally distributed that makes our model easy to learn the pattern of the data.





Data Summary

We have Yes Bank monthly stock price dataset. It has following features (Columns):

1)Open: Opening price of the stock of particular day.

2)High: It is the highest price at which a stock traded during a period.

3)Low: It is the lowest price at which stock traded during a period.

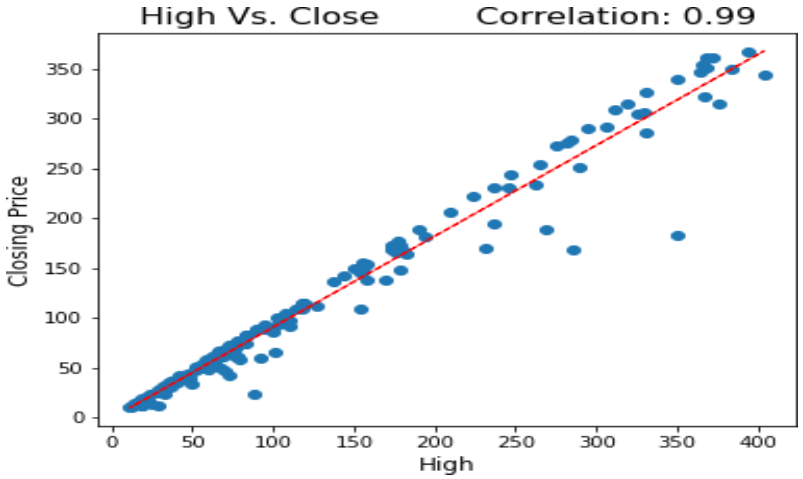
4)Close: Closing price stock at the end of a Trading Day.

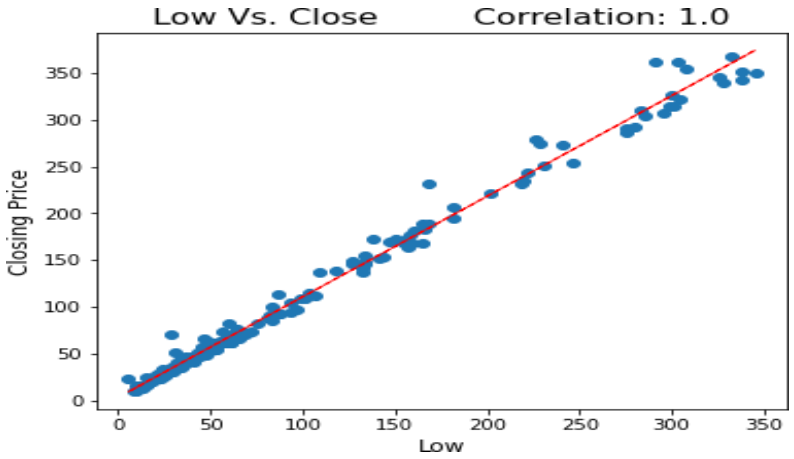
5)Date: We will use it as an index.

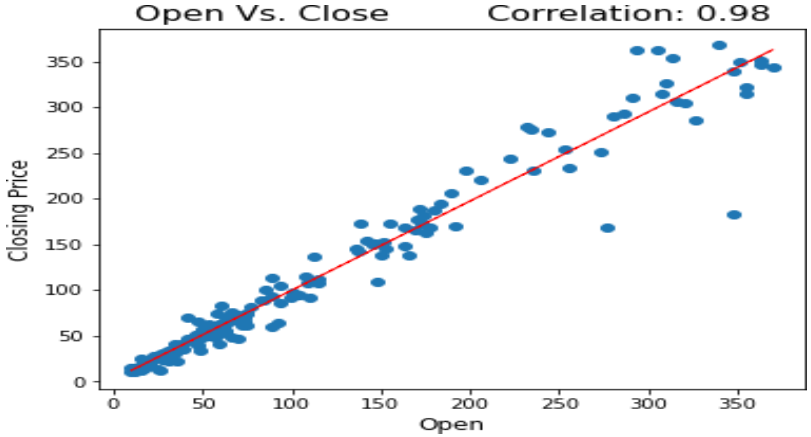
	Date	Open	High	Low	Close
0	Jul-05	13.00	14.00	11.25	12.46
1	Aug-05	12.58	14.88	12.55	13.42
2	Sep-05	13.48	14.87	12.27	13.30
3	Oct-05	13.20	14.47	12.40	12.99
4	Nov-05	13.35	13.88	12.88	13.41

Note: ‘Close’ will be our department variable and other will be independent.

Relationship Between Independent VS Dependent:

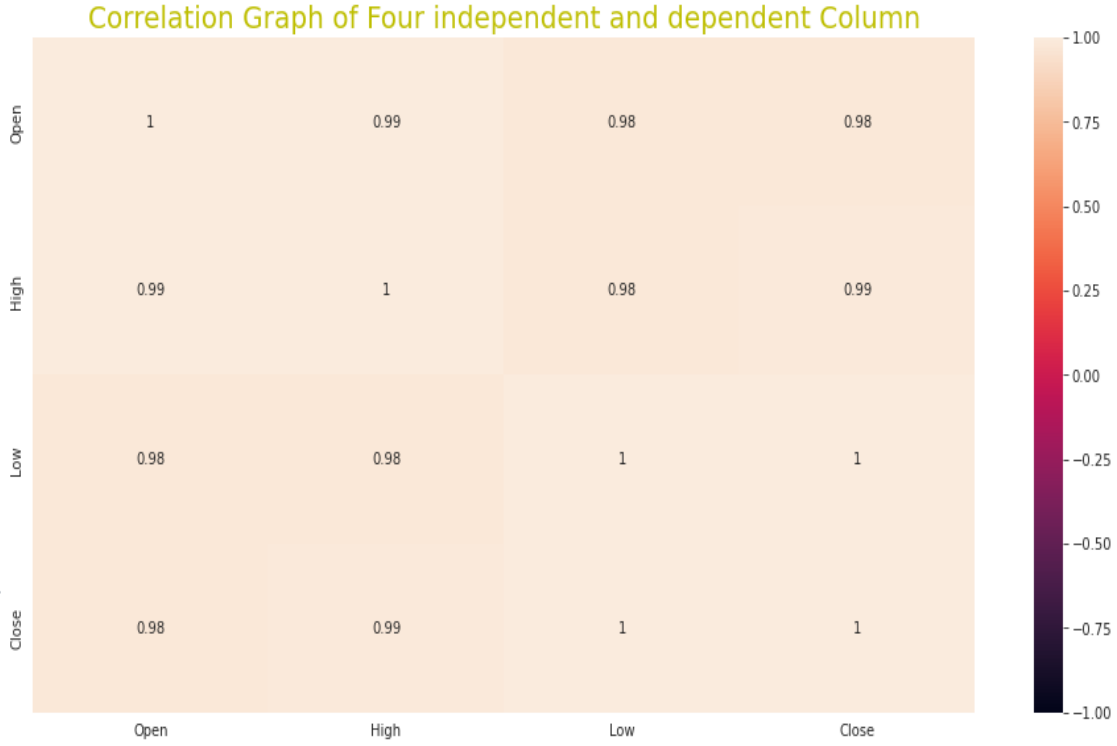






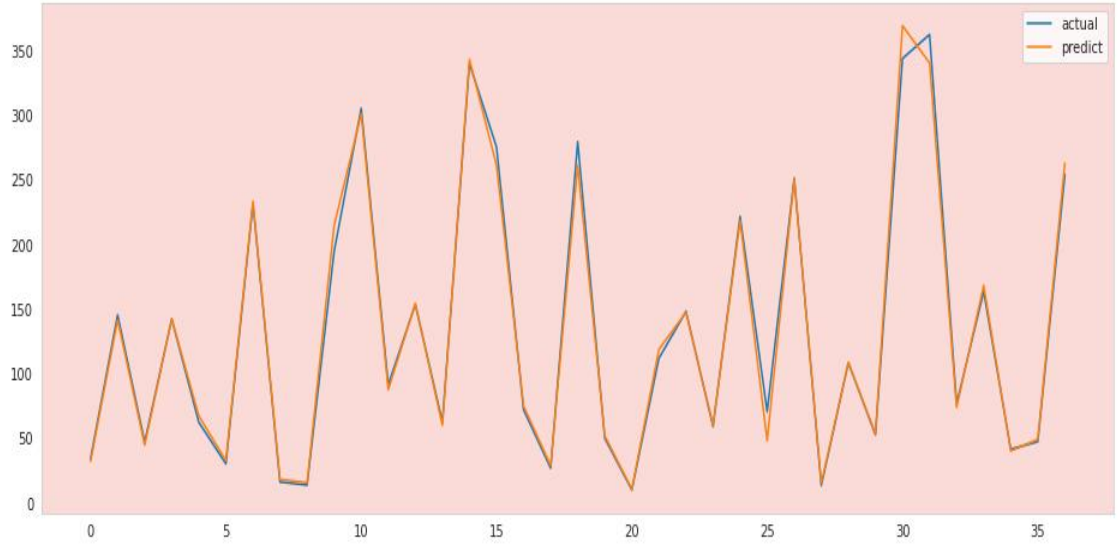
Correlation Heat-Map: Relationship between independent and dependent variable.

1. All variables show high correlationwith target variables.
2. The Heat Map helps us visualize the correlation of each parameter with respect to every other parameter.
3. The shades changes from the highest to lowest correlations.
4. We can see in the matrix on this slide that our dependent variable (Close price) is highly correlated with all the other independent variables



Linear – Regression:

Here, R2 is about 0.9930 which means model’since dependent feature is able to describe 99.30%of our dependent variable. Our adjusted r2 score for decision tree is 0.9999. It means our random forest is 99.22 correct fit in our model.

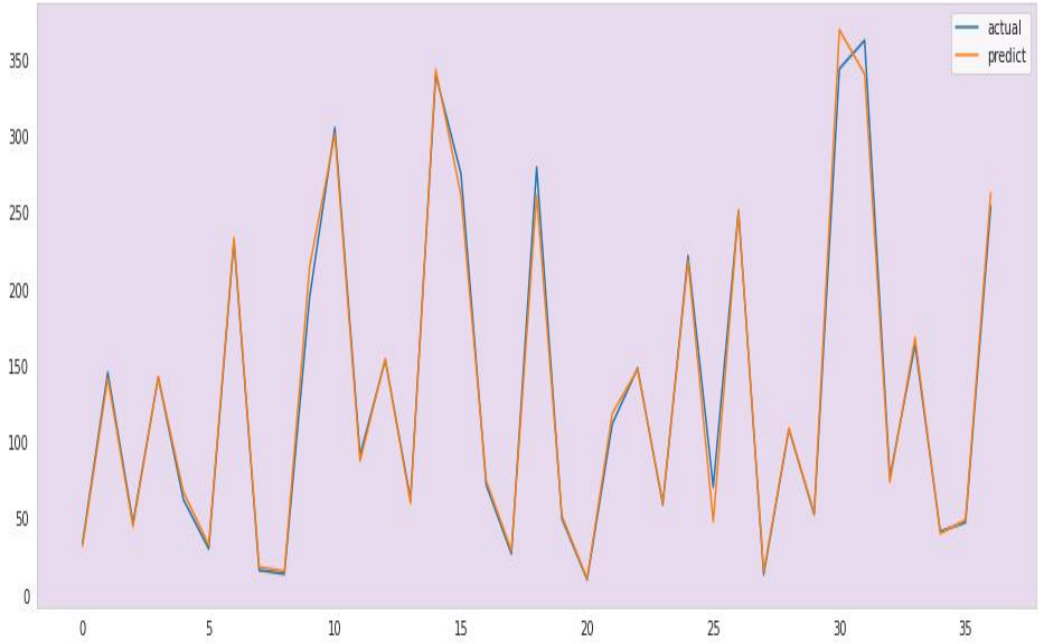


Ridge Regression:

Ridge is a l2 regularization. In ridge regression the penalty term is alpha, where alpha value is square of weights.

Here, R2 is about 0.9930 which means model's independent features is able to describe 99.30% of our dependent variable.

Our adjusted r2 score for decision tree is 0. 9922. It means our random forest is 99.22 correct fit in our model.
It means there is no change in prediction even after using cross validation (CV).

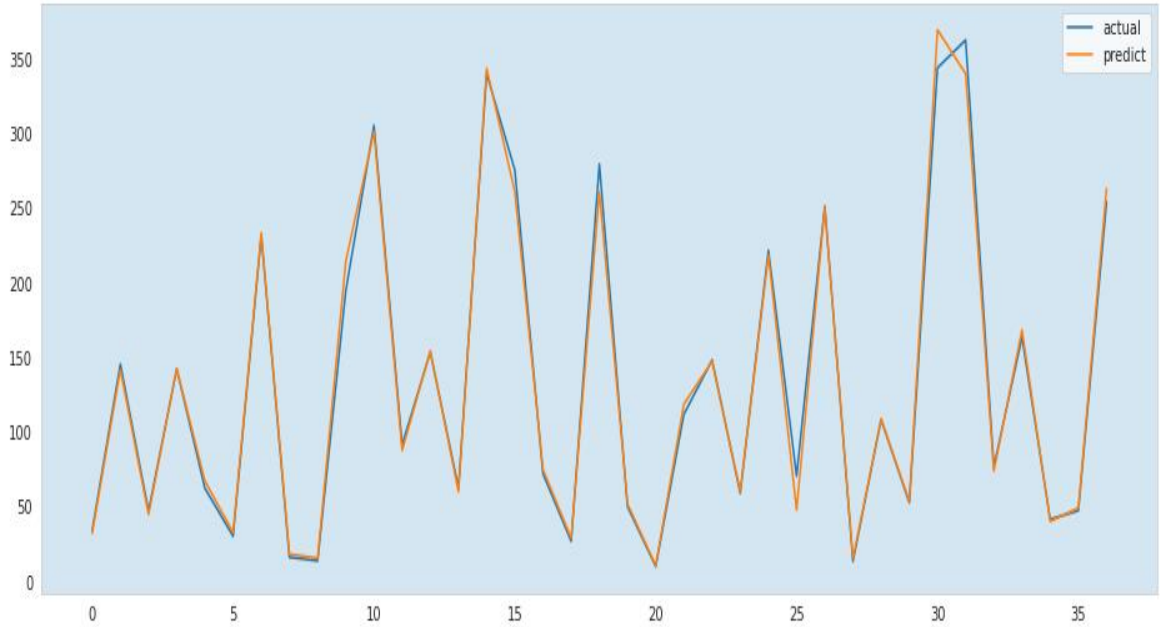


Lasso Regression:

Lasso is a l1 regularization. In lasso regression the penalty term alpha is absolute value of weights.

Here, R2 is about 0.9929 which means model's independent feature is able to describe 99.29% of our dependent variable.

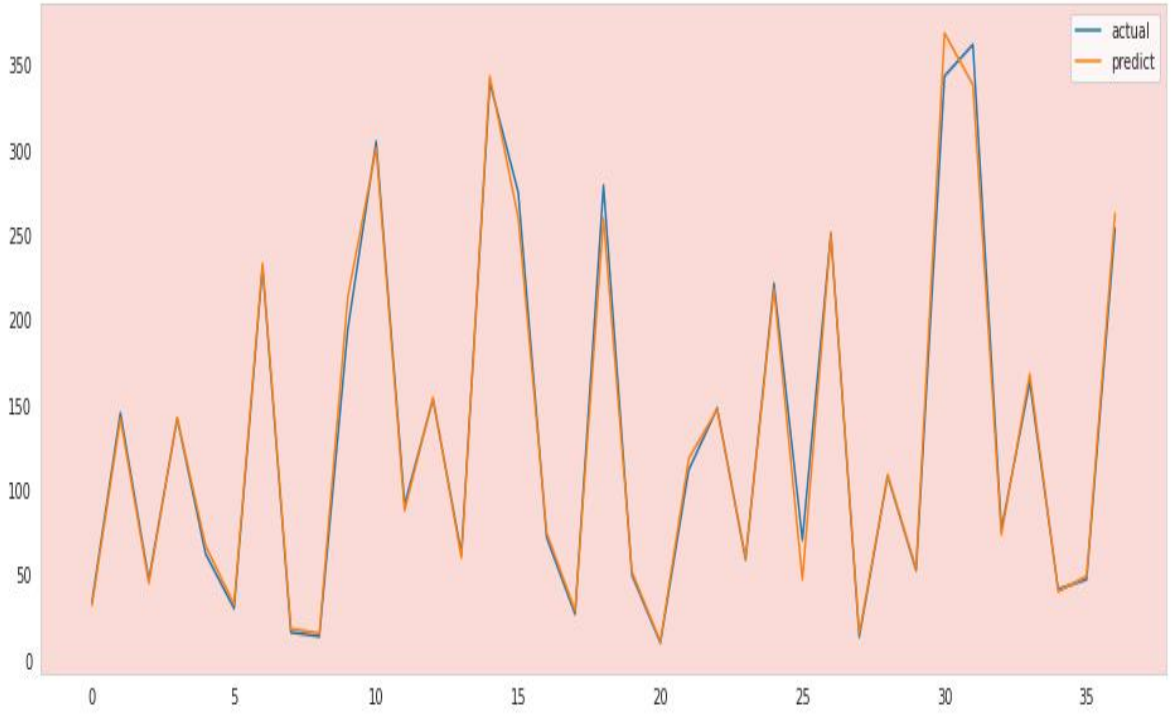
Our adjusted r2 score for decision tree is 0. 9921. It means our random forest is 99.21% correct fit in our model.



Elastic Net:

Here, R2 is about 0.9928which means model's independent feature is able to describe 99.28% of our dependent variable.

Our adjusted r2 score for decision tree is 0.9919. It means us random forest is 99.19 correct fit in our model.

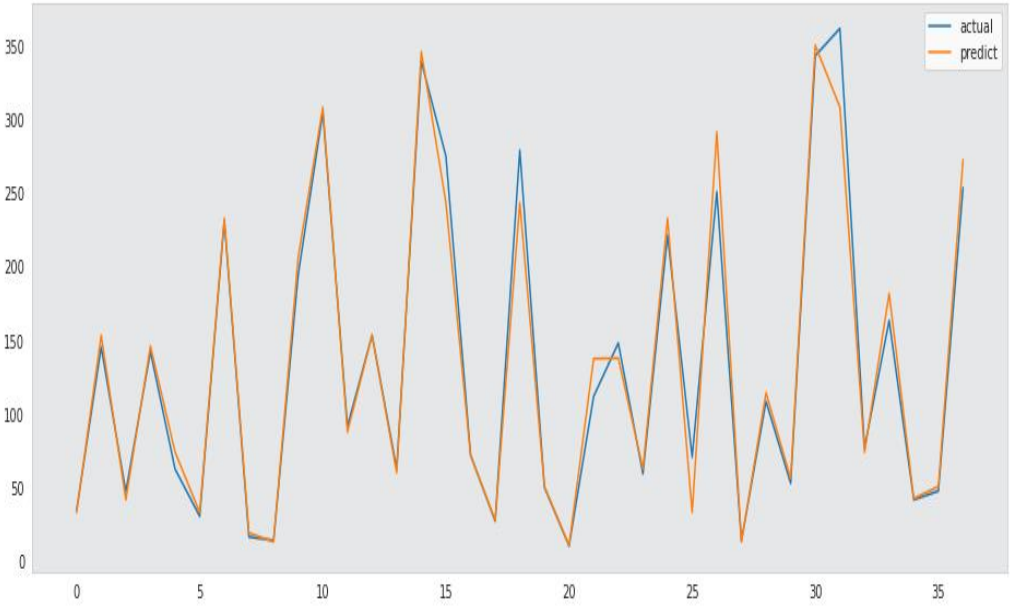


Decision tree:

Here, R2 is about 0.9764 which means model's independent features is able to describe 97.64% of ourdependent variable.

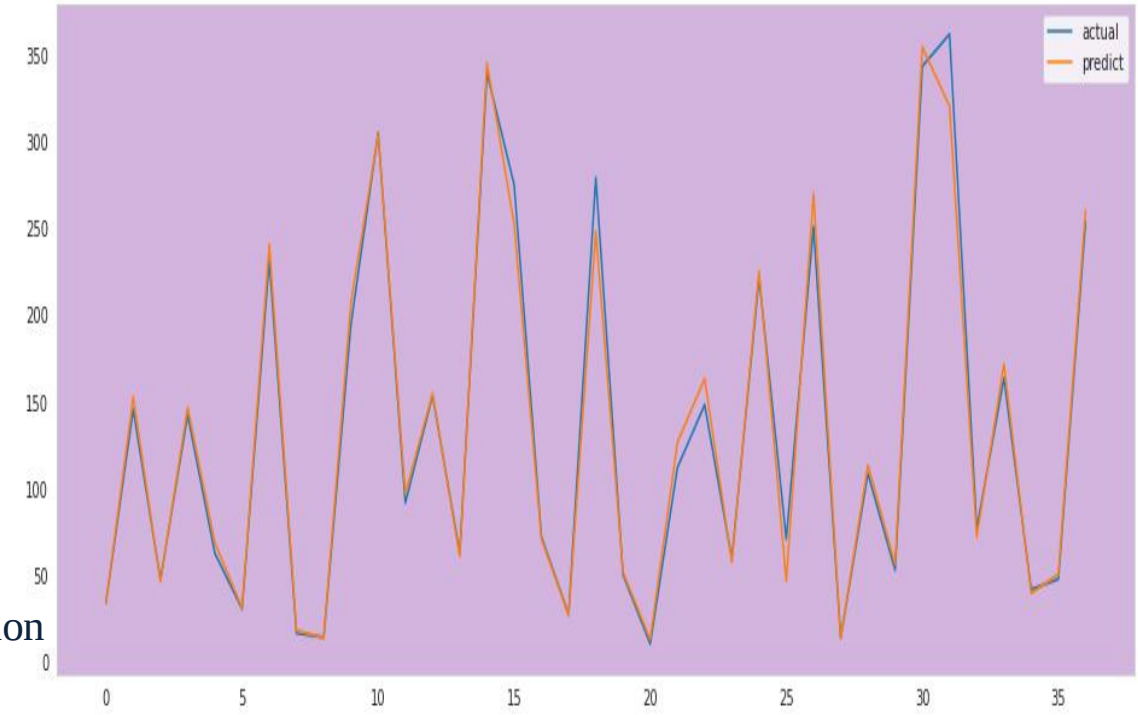
Adjusted R^2 is a corrected goodness-of-fit (model accuracy) measure for linear models. It identifies the percentage of variance in the target field that is explained by the input or inputs.

R^2 tends to optimisticallyestimate the fit of the linear regression. Our adjusted r2 score for decision tree is 0.9735 means 97.35%.



Random Forest:

Random forest is a commonly-used machine learning algorithm trademarked by Leo Barman and AdeleCutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

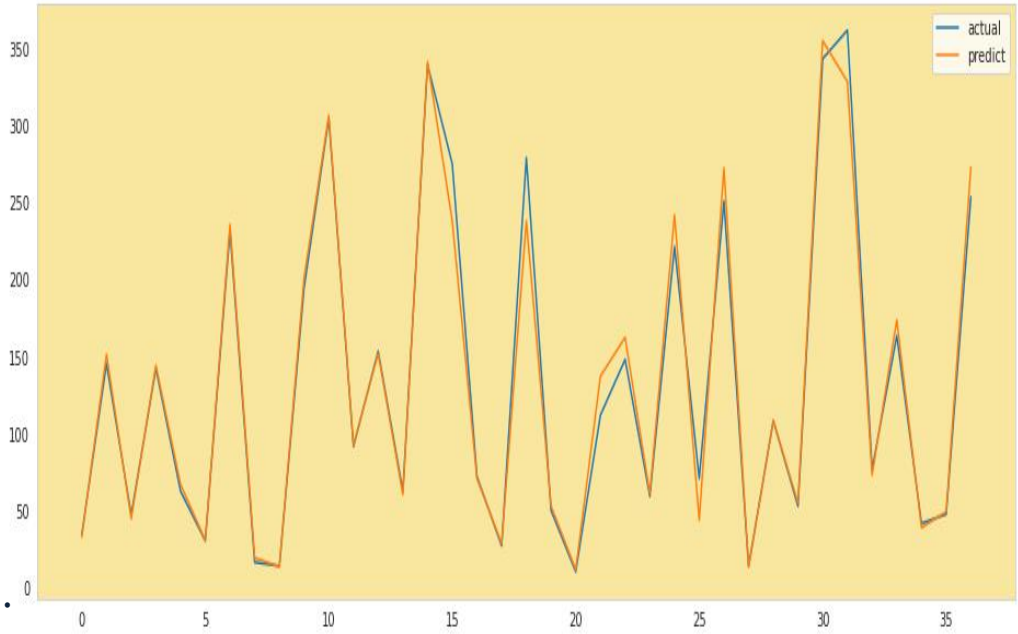


Here, R2 is about 0.9854 which means model's independent feature is able to describe 98.54% of our dependent variable.

Our adjusted r2 score for decision tree is 0.9835. It means our random forest is 98.35% correct fit in ourmodel.

GradientBoosting Regressor:

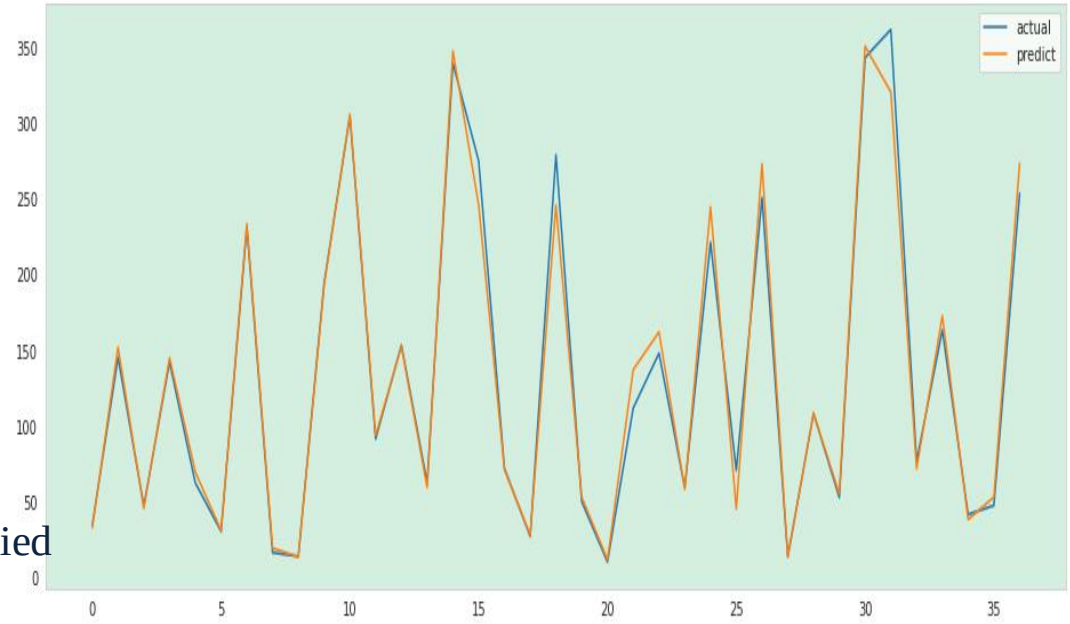
Gradient boosting is a type of machine learning Boosting. It relies on the intuition that the best Possible next model, when combined with Previous models, minimizes the overall Prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.



Here, R2 is about 0.9821 which means model's independent feature is able to describe 98.21% of ourdependent variable.

Our adjusted r2 score for decision tree is 0.9798. It means our random forest is 97.98 correct fit in ourmodel.

XGBoost:



Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effectiveimplementation of the gradient boosting algorithm.
Although other open-source implementations of the approach existed before XGBoost, the release of XGBoost appeared to unleash the power of the technique and made the applied machine learning communitytake notice of gradient boosting more generally.

Here, R2 is about 0.9830 which means model's independent features is able to describe 98.30% of our dependent variable.

Our adjusted r2 score for decision tree is 0.9798. It means our random forest is 97.98% correct fit in our model.

Evaluation Matrix for ML Regression models:

	Name	MAE	MSE	RMSE	R2_score_train	R2_score_test	adj_R2_score_train	adj_R2_score_test
0	LinearRegression:	5.393968	77.864757	8.824101	0.995455	0.993082	0.995328	0.992218
1	Lasso:	5.431330	78.971507	8.886591	0.995453	0.992984	0.995326	0.992107
2	Ridge:	5.397643	77.873453	8.824594	0.995455	0.993082	0.995328	0.992217
3	ElasticNetCV:	5.460264	81.001614	9.000090	0.995437	0.992804	0.995310	0.991904
4	DecisionTreeRegression:	10.101414	265.040114	16.280053	0.999936	0.976453	0.999934	0.973510
5	RandomForestRegressor:	7.334660	164.296212	12.817808	0.997470	0.985404	0.997399	0.983579
6	GradientBoosting Regressor:	8.651107	201.429802	14.192597	0.999711	0.982105	0.999703	0.979868
7	XGB regressor :	8.614088	190.320860	13.795683	0.999473	0.983091	0.999458	0.980978

Conclusion:

1. Target Variable is strongly dependent on Independent Variables.
2. We have seen that in our yes bank dataset there is no null values and no duplicate values are present. In this dataset we have one feature name is 'Date' which is object type, so we need to convert this into date format and apply some feature engineering methods.
4. With the help of distribution plot, we see that our data is positively skewed. So, we apply some kind of transformation i.e. Log Transformation to convert it into a normal distribution.
5. Lasso and Ridge regression models are giving the same result approximately.
6. We have implemented Cross Validation on different algorithm as CV performs better on small datasets. But, the result is nearly same.
7. We got a maximum accuracy of 99%.
8. Linear, lasso and ridge regression show almost same R squared values.
9. Whereas elastic net model shows lowest R squared value and high MSE, RMSE, MAE & MAPE.
10. Close, Open and high price of stock are strongly correlated with each other.
11. Regression models namely random forest regressor, XGBboost regressor are build.

Thank you

AlmaBetter...