

**Capstone project-2**

Yes Bank Stock Closing Price Prediction

Team- Incredible data Scientist

Team Members

Ranjit Biswal

Suvendu Dey

Abhishek Kumar

****

**Contents**

1.Introduction

2.Data Summary

3.Exploratory Data Analysis

4.Data Pre-Processing

5.Co-Relation Matrix

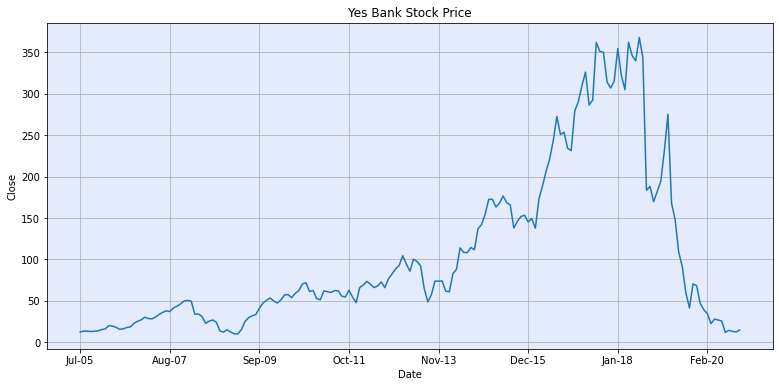
6.Model training

7.Hyper-Parameter Tuning

8.Summery of all ML models

9.Conclusion



 **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

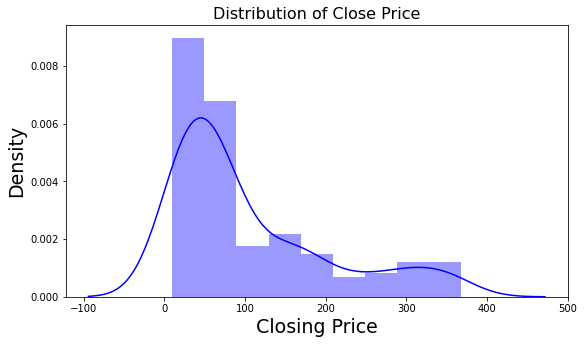
prediction of stocks future

price could result insignificant profit.

The efficient market

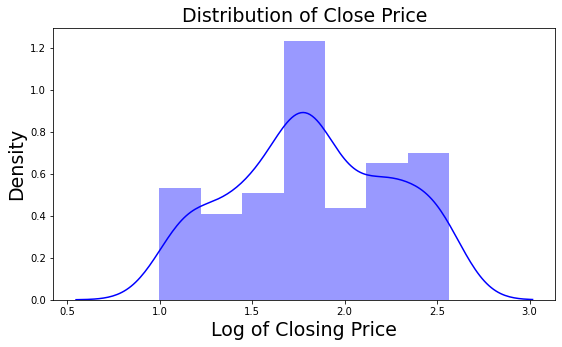
hypothesis recommends that stock costs mirror all right now accessible data and any value changes.

****

** 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.

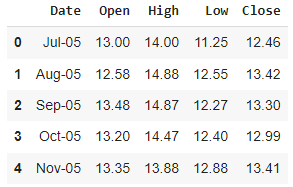
It gives us an intuition as to what is going on in the dataset.

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.

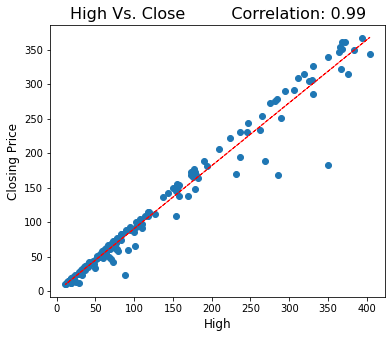
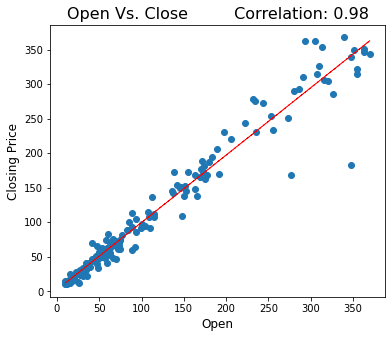
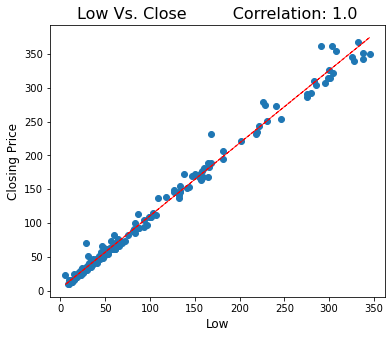
3)Low: It is the lowest price at which stock traded during a period.

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

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

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 correlation

with 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

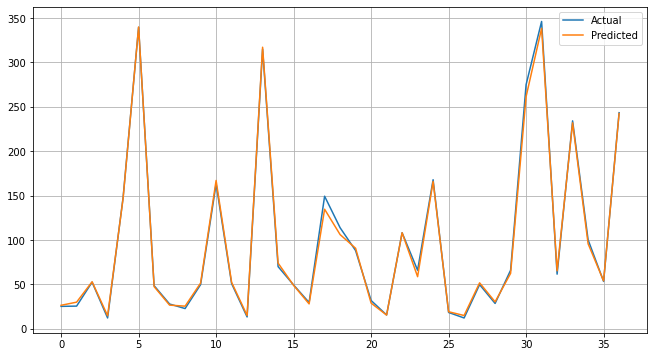
slide that our dependent variable

(Close price) is highly correlated

with all the other independent variables.



Linear – Regression:



Here, R2 is about 0.9928 which means model’s independent features is able to describe 99.16% of our dependent variable.

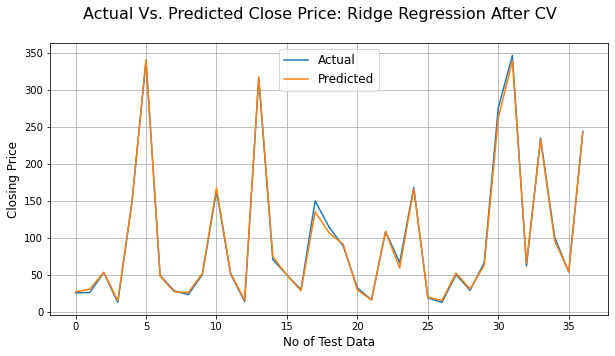
Our adjusted r2 score for decision tree

is 0.9986. It means our random

forest is 99.86 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 wights.

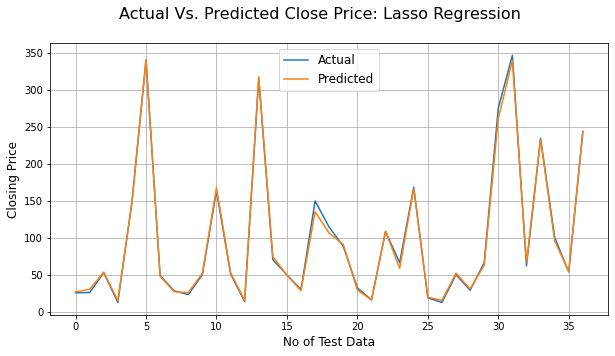
Here, R2 is about 0.9923 which means model’s independent features is able to describe 99.23% of our dependent variable.

Our adjusted r2 score for decision tree

is 0. 9911. It means our random forest is 99.11

correct fit in our model.%.

It means there is no change in prediction even after using cross validation (CV).

** Lasso Regression:**

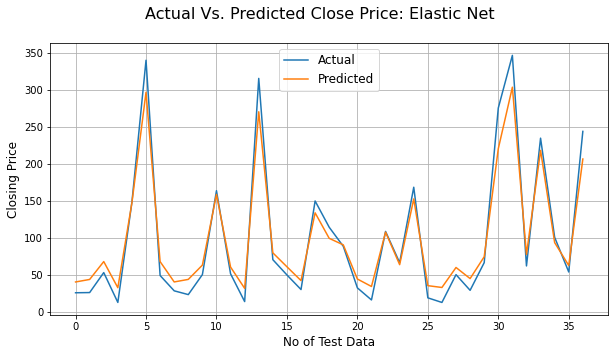
Lasso is a l1 regularization. In lasso regression the penalty term alpha is absolute value of wights.

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

Our adjusted r2 score for decision

tree is 0. 9916. It means our random

forest is 99.16% correct fit in our model.

** Elastic Net:**

Here, R2 is about 0.9689 which means model’s independent features is able to describe 96.89% of our dependent variable.

Our adjusted r2 score for

decision tree is 0.9639. It means us

random forest is 96.39 correct fit in our model.



### Decision tree:

Decision tree regression trains a model in the form of a tree to predict data in the feature and generate useful continuous output by observing the properties of an item.

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes, and leaf nodes.

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

Adjusted R2 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. R2 tends to optimistically estimate the fit of the linear regression. Our adjusted r2 score for decision tree is 0.9726. 

Random Forest:

Random forest is a commonly-used machine learning algorithm trademarked by Leo Bierman and Adele Cutler, 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 0.9850 which means model’s independent features is able to describe 98.50% of our dependent variable.

Our adjusted r2 score for decision tree is 0.9826. It means our random forest is 98.26 correct fit in our model.



XGBoost:

Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effective implementation 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 community take notice of gradient boosting more generally.

Here, R2 is about 0.9845 which means model’s independent features is able to describe 98.45% of our dependent variable.

Our adjusted r2 score for decision tree is 0.9820. It means our random forest is 98.20 correct fit in our model j.

**** **Evaluation Matrix for All ML Regression models:**

****

**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.

10Close, 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…**

