**CAPSTONE PROJECT 2**

(Regression Model)

Yes Bank Stock Prediction

Prepared by:

PRAVEEN .S



Data Science Trainee,

**Almabetter**

**Abstract:**

A bank's business model is different from that of a manufacturing company, the method used to evaluate a bank stock is also different. ... A bank's basic business is to accept deposits and give out loans. It makes money by charging a higher rate of interest on its loans than the rate it pays its depositor.In this section,we are developing a model for one of the leading banking firm in India which is ‘Yes Bank’.

Our experiment can help understand what could be the reason for the linear regression of such labels by feature selection, data analysis and prediction with machine learning algorithms taking into account previous trends to determine the current trend.

**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 involving Rana Kapoor. Owing to this fact, it was interesting to see how that impacted the stock prices of the company and whether Time series models or 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 of the month

This dataset has around 185 observations in it with 5 columns and it is a mix between categorical and numeric values.The dataset contains the following columns for our interpretation,

* **Date :**  It denotes the Month and Year of each Observation
* **Open :**  It denotes the starting stock value of that month
* **High :**  It denotes the highest stock value of that month
* **Low :** It denotes the lowest stock value of that month
* **Close :** It denotes the Closing stock value of that month

**Introduction:**

I had presented my exploratory analysis,,visualization,Correlation plots and developed three machine learning regression models and lot of other interesting insights into the given dataset.I choose this particular dataset because of the fraudulent impact on one of the leading banks in our country.

My goal here is to build a predictive model, which could help the people that have interest in future investing in stocks based on this bank.

**Steps involved:**

* **Exploratory Data Analysis**

After loading the dataset we performed this method by comparing our target variable that is Close with other independent variables. This process helped us figuring out various aspects and relationships among 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**

Our dataset didn’t contains a single null values which might not disturb our accuracy hence we used all of them except ‘Date’ since its a date-time object in order to get a better result.

* **Data Visualization:**

The target variable is extremely correlated with all

the independent features which was found by plotting

scatter plots across each independent variables.Since all

the features follows a normal distribution, the mean and

median values are found to be approximately equal using a

histogram.Due to the fraudulent impact,the last 3 years stock values are decreased significantly which was displayed using a bar graph

* **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. This was achieved by applying logarithmic transformation across all features.

* **Fitting different models**

For modelling we tried various classification algorithms like:

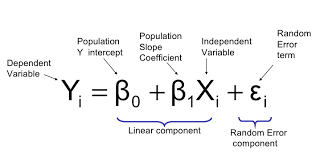
1. Linear Regression
2. Lasso Regression
3. Ridge Regression
4. Elastic Net Regression

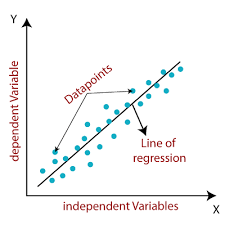
* **Tuning the hyperparameters for better accuracy**

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting and underfitting

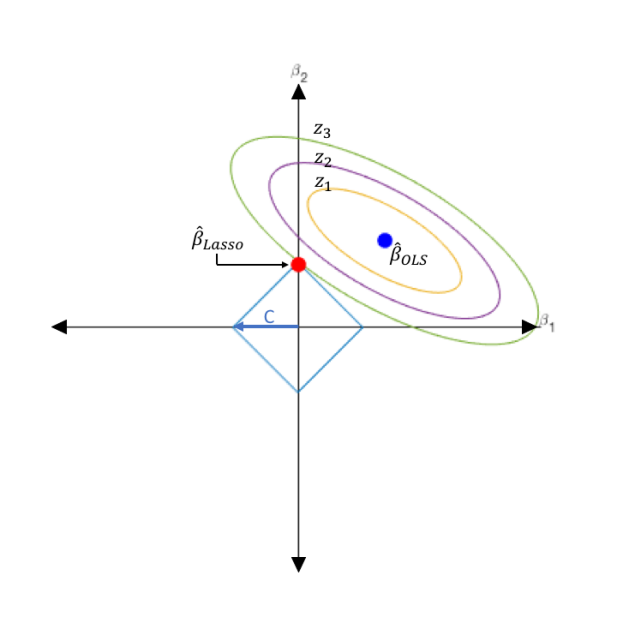
**Algorithms:**

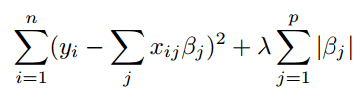
**Linear Regression:**

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. ... Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.



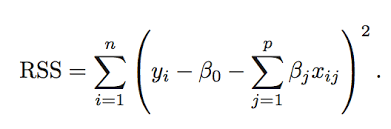
**Lasso Regression:**

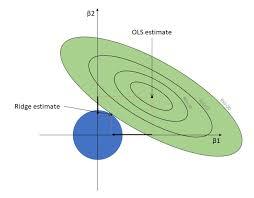
Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). ... The acronym “LASSO” stands for **Least Absolute Shrinkage and Selection Operator.**



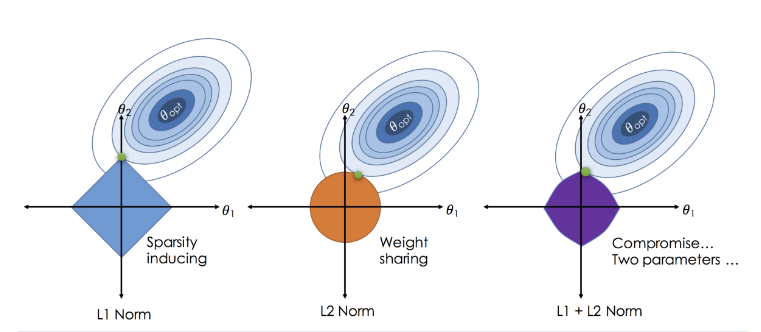
**Ridge Regression:**

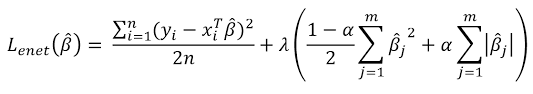
Ridge regression is the method used for the analysis of multicollinearity in multiple regression data. It is most suitable when a data set contains a higher number of predictor variables than the number of observations. The second-best scenario is when multicollinearity is experienced in a set.





**Elastic Net Regression:**

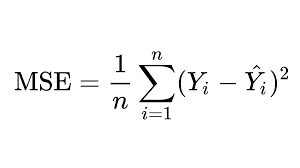
Elastic net is a popular type of regularized linear regression that combines two popular penalties, specifically the L1 and L2 penalty functions. ... Elastic Net is an extension of linear regression that adds regularization penalties to the loss function during training.



**Model performance:**

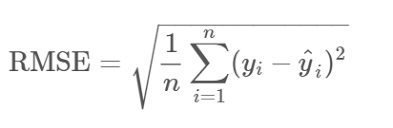
Linear model can be evaluated by various metrics such as:

Mean Squared Error:

It is a measure of how close a fitted line is to data points. For every data point, you take the distance vertically from the point to the corresponding y value on the curve fit (the error), and square the value.

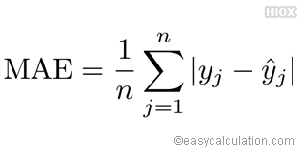
Root Mean Squared Error:

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.



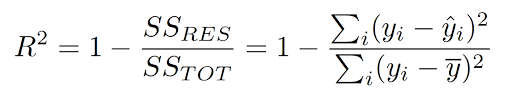
Mean Absolute Error:

Mean Absolute Error is a model evaluation metric used with regression models. The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors on over all instances in the test set



R2 Error:

R-Squared is also termed as the standardized version of MSE. R-squared represents the fraction of variance of response variable captured by the regression model rather than the MSE which captures the residual error.



2. Alpha:[1e-15, 1e-13, 1e-10, 1e-08, 1e-05, 0.0001, 0.001, 0.01, 0.1, 1, 5, 10, 20, 30, 40, 45, 50, 55, 60, 100, 0.0014]
3. Scoring: Negative mean squared error

**8. Conclusion:**

That's it! We reached the end of our exercise.

Starting with loading the data so far we have done EDA , null values treatment, encoding of categorical columns, feature selection and then model building. In all of these models our accuracy revolves in the range of 80 to 83%. And there is no such improvement in accuracy score even after hyper parameter tuning.

So the accuracy of our best model is 81% which can be said to be good for this very small dataset because of underfitting. This performance could be due to various reasons like multicollinearity of data,very small dataset as well as correct selection of models and hyperparameter tuning.

**References:**

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya
4. Wikipedia