Play Store Apps Review Analysis

Rajat Chaudhary, Anukriti Shakyawar,

Raman Kumar, Deepmala Srivastava

Team: Web Crawlers

Abstract:

YES BANK Limited is a private-sector bank. The Bank is engaged in providing banking services, including corporate and institutional banking, financial markets, investment banking, corporate finance, branch banking, business and transaction banking, and wealth management. The Company's segments include Treasury, Corporate/Wholesale Banking, Retail Banking, and Other Banking Operations. Its Treasury segment includes investments and financial markets activities undertaken on behalf of the Bank's customers, trading, maintenance of reserve requirements, and resource mobilization. Corporate/Wholesale Banking includes lending, deposit taking, and other services offered to corporate customers. Retail Banking includes lending, deposit taking, and other services offered to retail customers. The Other Banking Operations segment includes para-banking activities, such as third-party product distribution and merchant banking, among others.

1. Introduction:

YES BANK has been recognized among the Top and Fastest Growing Banks in various Indian Banking League Tables by prestigious media houses and Global Advisory Firms and has received several national and international honors for our different Businesses including Corporate Investment Banking, Treasury, Transaction Banking, and Sustainability at Yes Bank.

Stock market prediction is the act of trying to determine the future value of company stock or other financial instruments traded on an exchange. Successful estimation of a future stock price can yield significant profits. Several financial investment decisions are based on such forecasts and analyses.

The aim of this project is to estimate the closing price of stocks using various algorithms to aid in investment decisions with regard to purchasing or selling of stocks.

2. Data Description:

This data set consists of stock price data of YES BANK for 15 years from July 2005(the year it was listed) to November 2020 is collected. Data is extracted using the Pandas package and was taken from the Almabatter website. The data in this project consists of the monthly closing price of stocks. This dataset consists of 185 rows and 5 columns- Date, Open, High, Low, and Close

Date- This column is the Month and Year of the time the prices were recorded and we will use this column as an index of our dataset.

Open- This column contains the opening stock price of the month.

High- Highest stock price of the month.

Low- Lowest stock price of the month.

Close- Closing stock price of the month.

3. Analysis Methodology:

In this project, we performed various methods starting with data cleaning to data filtering, data visualization, and data transformation making it our three-step strategy for the analyses. We started by performing basic data inspection and found out that various info about our dataset such as data type, the shape of data, and total rows and columns present in our dataset. In this, we also found out if there are any null values or missing values present or not but fortunately, we didn't have any null or missing values in our dataset. While pursuing data cleaning we also found out that there are zero duplicates which makes our data cleaning way easier. After, this we pursued data visualization for both the dependent and independent variables and find various relations between them by plotting various types of graphs between them.

3.1. Data Cleaning:

We started our data cleaning by finding missing, null, and duplicate values but there are none present in our dataset. While performing the info command on our dataset we found that we have one object type column in our dataset which we converted into a date datatype and in describe command we found out that we have high variance between features like High, Low, and Close.

3.2. Data Visualization:

In this section performed various data visualization techniques to find the relations in our dataset. We started our visualization with the relation between the dependent variables "Closing price" and "date".

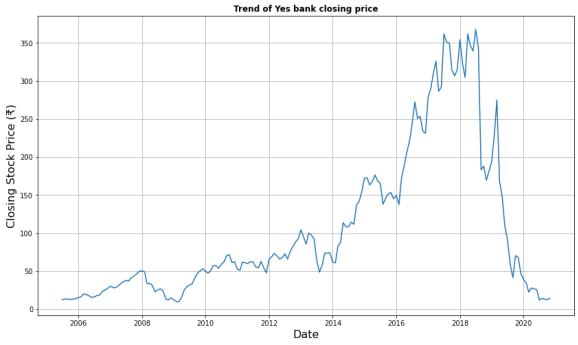


Fig.1: Trend of Yes Bank Closing Price

In this graph, we can see that there is constant growth in stock prices till 2018 but after that, we can see that there is a sudden fall in stock prices which we can relate to the Rana Kapoor incident which also occurred after 2018 and this has gravely affected the stock prices of YES BANK.

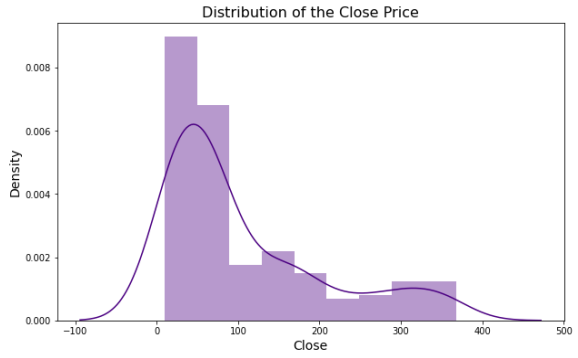


Fig.2: Distribution of Closing Price

In this graph, the distribution of Stock Closing Price is a rightly skewed distribution. It may lead us to misleading results in view of the statistical hypothesis. It can be corrected by applying Log Transformation after that we'll check how this data behave.

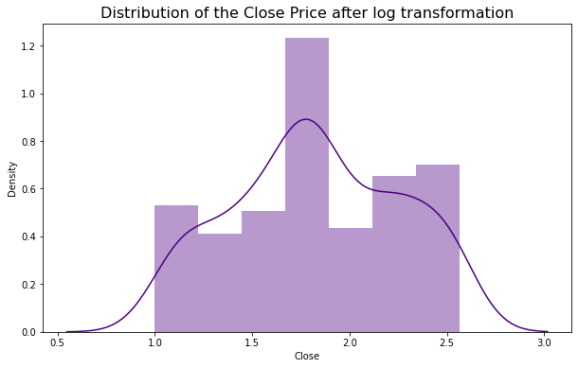
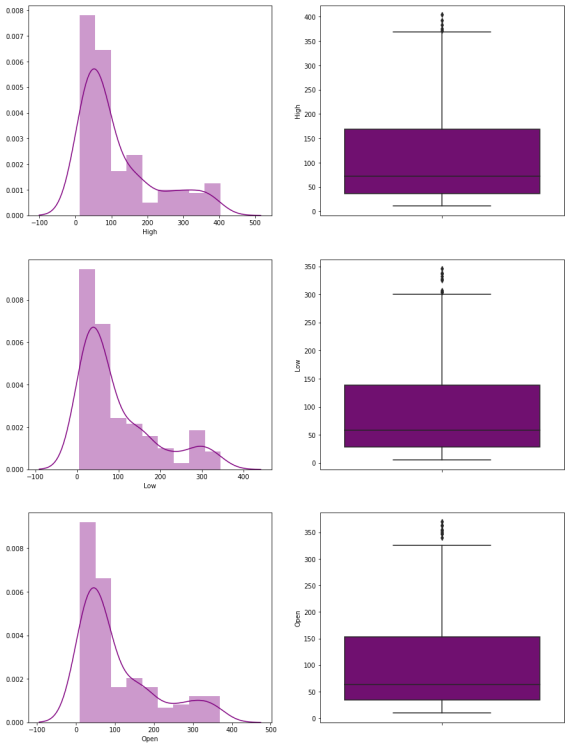


Fig.3: Distribution of the Close Price after log transformation

After applying log transformation we can see that the data is good and not rightly skewed and will result in good results in views of the statistical hypothesis.



In these graphs, we can see that all the graphs are rightly skewed and will lead us to mislead information and needed log transformation in all of them so that they can be further used for the analysis.

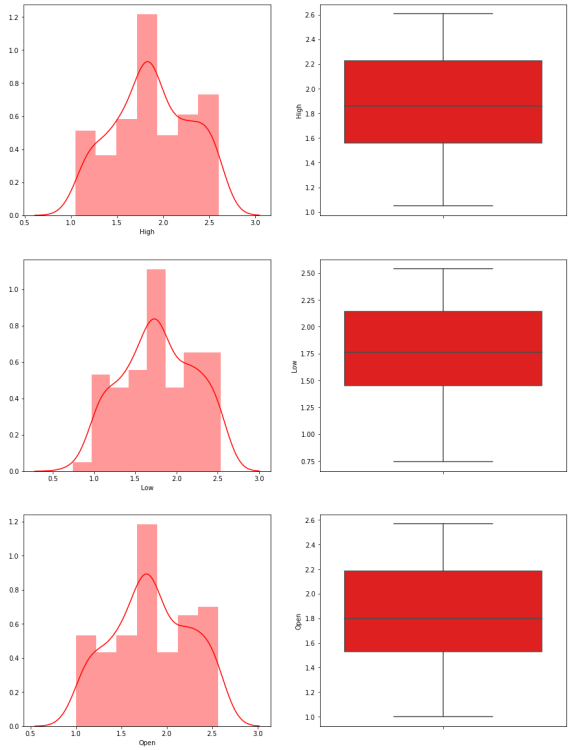


Fig.4: Distribution of High, Low & Open before and after log transformation

After the log transformation, we can see the improvement in graphs of the High, Low, and Open features with respect to closing price.

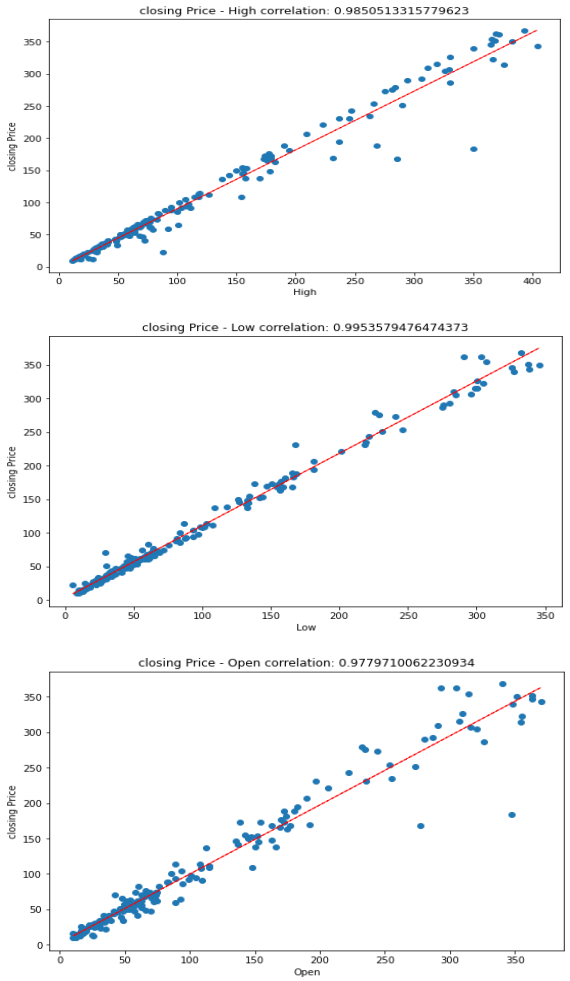


Fig.5: Correlation between Independent variables and Dependent variables

As we can see that there is linear relation and high correlation between each independent variable(High, Low, and Close) and dependent variable(Date).

4. Feature Engineering:

This is the pre-processing step of machine learning, which is used to transform raw data into features that can be used for creating a predictive model using Machine learning or statistical Modelling. Feature engineering in machine learning aims to improve the performance of models.

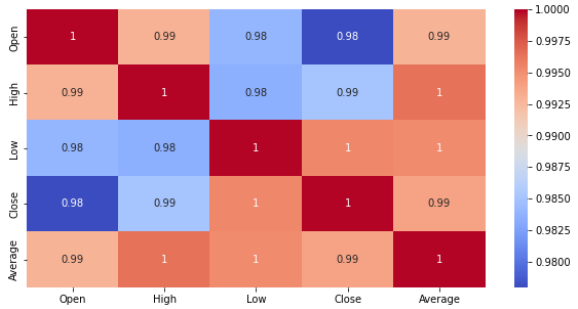


Fig.6: HeatMap for correlation between variables

In this graph, we can see that there is a very high correlation between independent variables which leads us to multicollinearity. High multicollinearity is not good for fitting models 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). So, we can decide which variable we should keep in our analysis and prediction model and which should be removed from the datasets.

VIF: Variance Inflation Factor (VIF) is used to detect the presence of multicollinearity.

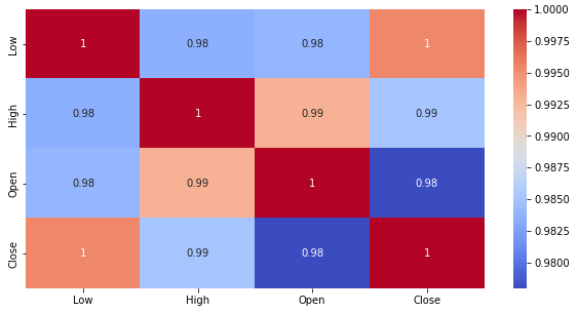
	Variables	**VIF**
0	Open	175.185704
1	High	inf
2	Low	inf
3	Average	inf

Through this table we can see that have very high VIF in our dataset so, we have to drop one of them which is least correlated with the dependent variable.

Now, let's check the correlation in all independent variables.

	Variables	**VIF**
0	Open	175.185704
1	High	167.057523
2	Low	71.574137

In this table we can see that the "Open" variable has a very high correlation which leads to less reliability in our regression result. But, before deleting any variable again we have to plot a heatmap between the left independent variables and the dependent variable.
So, we can decide which variable we can drop.



Our final dropping variable will be the High feature because it has less correlation with the dependent variable in comparison with the dependent variable(Close). We've dropped 3 features from our dataset it can affect our model efficiency but neglecting high VIF is far more dangerous than dropping features. So, we preferred to drop the features and move forward with the Low Variable.

5. Model Building:

For our Machine Learning model we have divided our data set in an 80:20 ratio with 80% of our data in the training module and 20% of our data in the testing module.

In this project, we are going to apply 4 models to train our dataset-

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

In terms of handling bias, Elastic Net is considered better than Ridge and Lasso regression, Small bias leads to the disturbance of prediction as it is dependent on a variable. Therefore Elastic Net is better at handling collinearity than the combined ridge and lasso regression.

Also, When it comes to complexity, again, Elastic Net performs better than ridge and lasso regression as in both ridge and lasso, the number of variables is not significantly reduced. Here, the incapability of reducing variables causes declination in model accuracy.

Ridge and Elastic Net could be considered better than the Lasso Regression as Lasso regression predictors do not perform as accurately as Ridge and Elastic Net. Lasso Regression tends to pick non-zero as predictors and sometimes it affects accuracy when relevant predictors are considered as non-zero.

5.1. Linear Regression Model:

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis.

Mathematically, we can represent a linear regression as

$$y= a0+a1x+ \varepsilon$$

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

It also has a cost function-

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

After using it in our model we get the following results-

Mean Squared Error: 0.008378716531125619
Root Mean Squared Error: 0.09153532941507131
R2: 0.9550214108859424
Adjusted R2: 0.9147774100996803

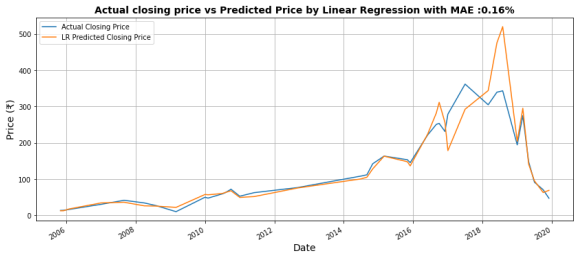


Fig.7: Actual closing price vs Predicted Price by Linear Regression

This graph predicts closing price with a training accuracy of 95.5% introducing dummy variables and removing columns which are highly correlated and that lead to multicollinearity.

5.2. Ridge Regression Model:

Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions. It is a regularization technique, which is used to reduce the complexity of the model. It is also called L2 regularization. In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called the Ridge Regression penalty. We can calculate it by multiplying the lambda by the squared weight of each individual feature.

The equation for the cost function in ridge regression will be:

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}\beta_j{}^2$$

In the above equation, the penalty term regularizes the coefficients of the model, and hence ridge regression reduces the amplitudes of the coefficients that decrease the complexity of the model. As we can see from the above equation, if the values of λ tend to zero, the equation becomes the cost function of the linear regression model. Hence, for the minimum value of λ, the model will resemble the linear regression model. A general linear or polynomial regression will fail if there is high collinearity between the independent variables, so to solve such problems, Ridge regression can be used.

After using it in our model we get the following results-

Mean Squared Error: 0.009365359873489867
Root Mean Squared Error: 0.0967747894520565
R2: 0.9497249164487018
Adjusted R2: 0.9047419469554351

After applying Cross Validation and Hyperparameter Tuning in Ridge we can get-
By Using {'alpha': 3} Negative mean squared error is: -0.012538304166010358

After applying this to our equation we get the following result-

Mean Squared Error: 0.008847513525776934
Root Mean Squared Error: 0.09406122222136461
R2: 0.9525048169276678
Adjusted R2: 0.9100091268103179

In this, we can see that after applying results from cross-validation our model has improved very much and is much more reliable than before.

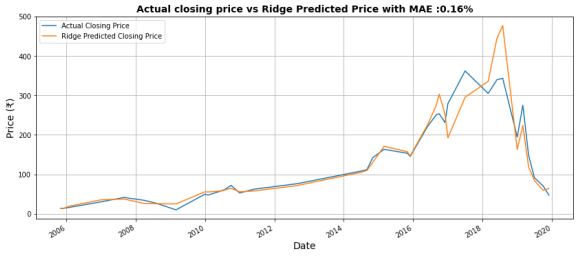


Fig 8. Actual closing price vs Predicted Price by Ridge Regression

This graph predicts the closing price with a training accuracy of 94.58% after hyperparameter tuning and cross-validation.

5.3. Lasso Regression Model:

Lasso regression is another regularization technique to reduce the complexity of the model. It stands for Least Absolute and Selection Operator. It is similar to the Ridge Regression except that the penalty term contains only the absolute weights instead of a square of weights. Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0. It is also called L1 regularization.

The equation for the cost function of Lasso regression will be:

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}|\beta_j|$$

Some of the features in this technique are completely neglected for model evaluation.

Hence, the Lasso regression can help us to reduce the overfitting in the model as well as the feature selection.

After using this method we get the following results-

Mean Squared Error: 0.00940109189845321
Root Mean Squared Error: 0.09695922802112861
R2: 0.9495330999499494
Adjusted R2: 0.9043785051683252

After applying Cross Validation and Hyperparameter Tuning in Lasso we can get-

By Using {'alpha': 0.0014} Negative mean squared error is: -0.0126307155201585

After applying this to our equation we get the following result-

Mean Squared Error: 0.009376701436556797
Root Mean Squared Error: 0.09683336943717696
R2: 0.9496640327198869
Adjusted R2: 0.9046265883113646

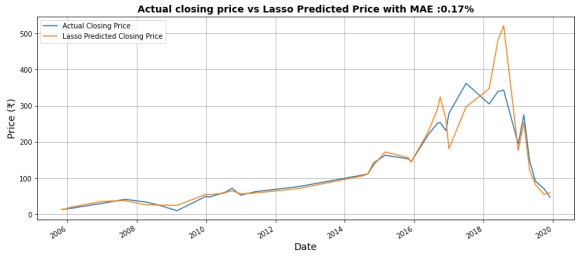


Fig 9. Actual closing price vs predicted price by Lasso Regression

This graph predicts the closing price with a training accuracy of 94.58%.

5.4. Elastic Net Regression:

Coefficients to the variables are considered to be information that must be relevant, however, ridge regression does not promise to remove all irrelevant coefficients which is one of its disadvantages over Elastic Net Regression(ENR)

It uses both Lasso as well as Ridge Regression regularization in order to remove all unnecessary coefficients but not the informative ones.

ENR = Lasso Regression + Ridge Regression

The equation for ENR is given below-:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - (mx_i + z))^2 + \lambda\sum_{i=1}^{p}(mx_i + z)^2 + \lambda\sum_{i=1}^{p}(mx_i + z)$$

Mean Squared Error: 0.030334217104865353
Root Mean Squared Error: 0.17416721018855802
R2: 0.8371599895774178
Adjusted R2: 0.6914610328835284

After applying Cross Validation and Hyperparameter Tuning in Elastic Net we can get the-
{'alpha': 0.01, 'l1_ratio': 0.3} Negative mean squared error is: -0.0123333508481652 9

Mean Squared Error: 0.009096377849834535
Root Mean Squared Error: 0.09537493302663198
R2: 0.9511688645612937
Adjusted R2: 0.9074778486424512

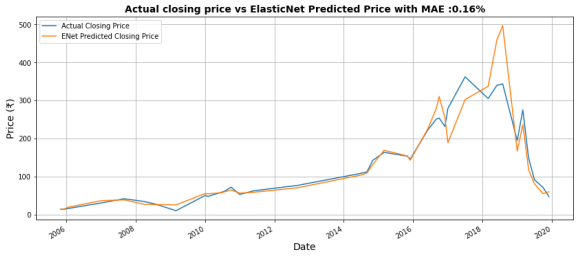


Fig 10. Actual closing price vs predicted price by ElasticNet Regression

This graph predicts closing price with a training accuracy of 94.58% after introducing dummy variables and removing columns that are highly correlated and that lead to multicollinearity.

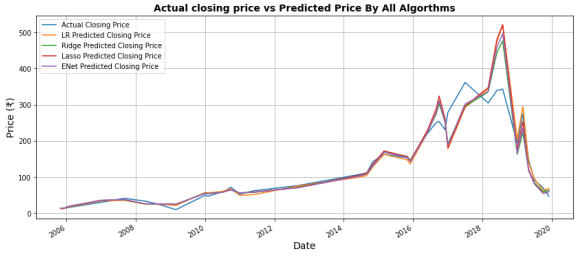


Fig 11. Actual closing price vs predicted by all algorithms

This graph shows the combined result of the predicted price of all models in comparison to the actual closing price.
Linear Regression and Lasso are performing better than other models with training accuracy of 94.0359% and 94.45777% respectively.
Apart from Linear Regression and Lasso, Ridge and Elastic Net are also performing better but they have less training accuracy.

Conclusion:

- Target Variable is strongly dependent on Independent Variables.
- Linear Regression and Lasso are performing better than other models with training accuracy of 94.0359% and 94.45777% respectively.
- Apart from Linear Regression and Lasso, Ridge and Elastic Net are also performing better but they have less training accuracy.
- Ridge and ElasticNet are performing far much better after Applying Hyperparameter Tuning and Cross-validation, it is because we have a small set of datasets. 5. R2 and Adjusted R2 are around 95 and 91% in each model.

Future Work:

We can apply more models in our dataset like Random Forest or XGBoost or we can also explore this project with the time series analysis method to get more accurate results.

References:

1. https://www.kaggle.com/lava18/google-play-store-apps
2. https://jovian.ml/ritz1602-rs/course-project-google-play-store-dataset
3. https://jovian.ai/learn/data-analysis-with-python-zero-to-pandas
4. https://seaborn.pydata.org/examples/index.html
5. https://matplotlib.org/3.1.1/index.html