Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Sananda Biswas Chatterjee(sananda.uem@gmail.com)

Contribution:

1)Data Wrangling –

Analyzed the Data set

2) Data cleaning-

Null value treatment, Duplicate values treatment, outlier detection and change the data into correct format

3)EDA and Data Visualization-

Analyzed the univariate

Find the correlation between variables through heatmap

4) Calculated the VIF

5)Creating model –

i) Linear Regression

ii) Lasso Regression

iii)Ridge Regression

6) Cross validation-

Optimized the hyper parameter tuning of Lasso and Ridge Regression.

Amit Kundu (amitkundu9593548931@gmail.com)

Contribution:

1)Data Wrangling –

Analyzed the Data set

2) Data cleaning-

Null value treatment, Duplicate values treatment, outlier detection and change the data into correct format.

3)EDA and Data Visualization-

Analyzed the bivariate analysis

Find the relation between dependent and independent variables using scatter plot.

Find the correlation between variables through heatmap.

4) Calculated the VIF-

Calculated the Variance Inflation Factor (VIF) to check multicollinearity.

5) Splitting data set into train and test set

6)Creating model –

i)Elastic Net

ii)XG Boost

7)Cross validation-

Optimized the hyper parameter tuning of Elastic Net Regression.

Please paste the GitHub Repo link.

GitHub Link:- https://github.com/sananda2005/Yes-Bank-Stock-Closing-Price-Prediction

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

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.

Approaches

After importing the dataset using Pandas, data manipulation and acquaintance with the dataset's features followed. The dataset doesn't contain any null values, hence we are not required to remove any.
The EDA phase follows, when we swiftly analyse the stock's close price and pinpoint dependent and independent variables. We utilised a distribution plot to look at the skewness of the dependent variable after determining what it was. We found the data to be properly skewed, requiring the use of log transformation.
Heatmap has now been used to verify the transformation correlation, and it has revealed that there is a very high correlation between all of the variables (variation inflation factor). Prepare the dependent and independent variables for the train test split technique. Along with Lasso, Elastic net, and Ridge regression, we also employ linear regression. In terms of model performance, ridge and linear regression both perform well. Cross validation and hyper parameter adjusting result in a significant improvement in performance.

Conclusion

1. The tendency of Yes Bank's stock's Close, Open, High, and Low prices increased until 2018 and then unexpectedly decreased after fraud case of Rana Kapoor.
2. The target variable is highly dependent on input variables.
3. Each independent variable has a strong correlation with the others (Multicollinearity)
4. The R squared values for linear, lasso, and ridge regressions are nearly identical.
5. On the basis of RMSE (Root Mean Squared Error) and MAPE (Mean Absolute Percentage Error), I compared 5 models (Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression, and XG-Boost Regression).

6. The dependent variable (Closing Price) and the independent variables (High, Low, and Open) have a direct correlation.
7. With the lowest RMSE (0.0394) and MAPE (0.0196) as well as the greatest r2 score (0.9913), XG-Boost Regression is the best model.