Capstone Project Submission

Team Member's Name, Email and Contribution:

Contributor Role :

1) **Data Wrangling -**
   i. Analyze the Data set

2) **Data cleaning-**
   i. Delete unnecessary data.
   ii. Null value treatment/Duplicate values treatment

Name: Simran Dapke
Email:simrandapkea99@gmail.com
Contribution:

1) **Data Visualization-**
   i. Distribution of dependent variable Close Price of stock using Distplot
   ii. Normalizing Distribution of dependent variable Close Price of stockusing Distplot
   iii. Scatter plot best fit line
   iv. Find correlation with heatmapv.

2) **VIF**
   i. Important features.

3) **Regression analysis-**
   i. Lasso
   ii. Elasticnet
   iii. SVM
   iv. KNN
   v. Ridge Regressionvi.

4) **Cross validation**
   i. Optimize hyper tuning parameter for Lasso and ElasticNet
   ii. Optimize hyper tunning parameter fro ridge

Please paste the GitHub Repo link.

GitHub Link: - https://github.com/simrandapke/Regression-Capstone-Project-

G-Drive link –

https://drive.google.com/drive/folders/1htvkVZpjkw0nO016ILuMPoDh8cCA6KPG?usp=drive_link

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

Yes Bank Limited is an Indian private sector bank headquartered in Mumbai, India, and was founded by **Rana Kapoor and Ashok Kapur** in 2004. It offers a wide range of differentiated products for corporate and retail customers through retail banking and asset management services. On 5 March 2020, in an attempt to avoid the collapse of the bank, which had an excessive amount of bad loans, the Reserve Bank of India (RBI) took control of it. When the fraud case news come out the stock price of yes bank fell from 391 to 190 and now the running price is around 13 Rs.

My First Step was to import the dataset using Pandas then data wrangling and know the features in the dataset. I did not get into the situation to remove NA values because there are 0 null values in the dataset.

The next step is EDA in that I briefly study related to the close price of the stock and find out dependent, and independent variables. After knowing the dependent variable I plot a distribution plot to check the skewness of the variable and I find the data is rightly skewed so need to use log transformation.

Now after transformation correlation has been checked with help of heatmap, there is a very high correlation among all variables so to check multicollinearity is VIF(variation inflation factor). We drop some features to prevent wrong predictions.

Prepare dependent and independent variables for the train test split method. I apply **Linear regression, Lasso regression, Ridge regression, and Elastic net regression**. As per model performance, linear regression and ridge perform well. After cross validation and Hyperparameter tuning performance increase significantly.

Conclusion:

- In EDA part we observed that

    1. There is increase in trend of Yes Bank's stock's Close, Open, High, Low price till 2018 an then sudden decrease.
    2. We observed that open vs close price graph concluded that after 2018 yes bank's stock hitted drastically.
    3. We saw Linear relation between the dependent and independent values.
    4. There was alot of multicollinearity present in data.

- Target variable(dependent variable) strongly dependent on independent variables
- We get maximum accuracy of 99%
- Linear regression and Ridge regression get almost same R squared value
- Whereas Lasso model shows lowest R squared value and high MSE,RMSE,MAE,MAPE
- Ridge regression shrunk the parameters to reduce complexity and multicollinearity but ended up affecting the evaluation metrics.
- Lasso regression did feature selection and ended up giving up worse results than ridge which again reflects the fact that each feature is important (as previously discussed).
- KNeighborsRegressor end up giving the highest R squared value. The predicted values are nearly equal to the actual values. We got 99% accuracy.
- SVM and Elastic Net showed nearly equal accuracy.