**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

| **Team Member’s Name, Email and Contribution:** |
| --- |
| **Name -** Ansh Srivastava  **Email -** [anshsri2001@gmail.com](mailto:anshsri2001@gmail.com)  1. Question framing(Based on problem statement)  2. Data inspection & pre-processing  3. Exploratory Data Analysis(EDA)  4. Model implementation  Linear regression  Lasso regression  Ridge regression  KNN  XGboost regression  5. Conclusions |
| **Please paste the GitHub Repo link.** |
| **Drive Link:-** [**Yes Bank Stock Closing Prediction - Google Drive**](https://drive.google.com/drive/folders/1tQ-bweldcC-2Bqm3qSD5zbqqg1Zza0HT)  **Github Link:-** [Ansh-srivastav/Yes--Bank--Stock--Closing--Price--Prediction (github.com)](https://github.com/Ansh-srivastav/Yes--Bank--Stock--Closing--Price--Prediction)  **Email:-** anshsri2001@gmail.com |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| First Step was to import the dataset through Pandas ‘read\_csv’ then data wrangling and feature engineering in our dataset. We did not get into the situation to remove NA values because there are 0 null values in the Yes Bank dataset.  Next, EDA(exploratory Data Analysis) in which trends of stock closing price, distribution of dependent variables have been examined. Plotted histogram of all variables with mean and median to check measures of central tendency is close to each other or far. Then log transformation has been applied on each variable, which led to a conclusion: to normalize right-skewed data perform log transformation.  the correlation has been checked among each other through heatmap, there was a very high correlation among independent features means high multicollinearity in our model, so to check how high multicollinearity is VIF(Variation Inflation Factor) has been checked based on VIF, three features had to drop from the dataset to prevent the wrong prediction.  Prepared independent and dependent variables for the train test split method. Applied Linear model, Ridge regression, Lasso regression and ElasticNet all the models are performing in a better way but Linear Model and Lasso are performing in a better way in comparison to Ridge and ElasticNet but Ridge regression and ElasticNet performance improved by applying cross-validation and Hyperparameter tuning. |
| **Conclusions:**  1. Target variable is highly dependent on input variables.  2. Linear Regression has given the best results with lowest MAE, MSE, RMSE and MAPE score.  3. Ridge regression shrunk the parameters to reduce complexity and multicollinearity, but ended up affecting the evaluation metrics.  4. Lasso regression did feature selection and ended up giving up worse result than ridge which again reflects to the fact that each feature is important(as previously discussed).  5. KNN AND XGBoost have given similar results.  6. The accuracy for each model is more than 90%. |