Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

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

Our primary goal is to understand the dataset of Yes Bank and the fraud incident that happened which affect the stocks of bank and relate that values with rest all. Our secondary goal is to draw out actionable insights from our analysis and give conclusions about key aspects of dataset and performed regression models to conclude the predicted result compared with the actual. Our approach towards the problem statement was well decided by two of us. As our approach was clear and we moved in a sequence. We used 5-6 regression models to get useful insights.

As a first step we saw the dataset and divided the complete project into sub categories that is Cleaning the dataset, Handling the null values, bifurcate the Univariates, Bivariate and the regression models to get some useful predictions.

We started with Importing some of the useful machine learning libraries, needed for the project. After the basic operations of importing the dataset, seeing at the overview of the dataset we started with changing the format of date as for our data cleaning as such not required, null values were also not present then we performed univariate and bivariate analysis then all models.

As we proceed, we looked into the dataset and we have 5 columns type i.e., Date, open, high, low, and close in this close was only the dependent variable and rest all were independent variables.
Then for data pre-processing we changed the date format from MMM-YY which was converted to the proper date of YYYY-MM-DD. As our dataset was pre-processed and completed, then first we bifurcate the univariate and bivariate data.

For univariate analysis, first we performed for independent variables i.e., open, high, and low. We proceed the analysis and we conclude from the histogram plot that all were right skewed graph. Then we proceed with dependent variable and visualized that for this also the histogram plot was right-skewed.

Then we performed the log transformation, because this is the way of handle skewed data and also de-emphasizes outliers and allow a bell-shaped distribution and all the plot is transformed into a normally distributed variable to some extent.

The next we performed was bivariate analysis which used to find out the relationship between two sets of values which we performed to compare dependent variable with other three independent variables, and we concluded that they are highly collinear with each other as they are positive correlation with 98%, 99% and 98%.
After this correlation, we also conclude the multicollinearity between them and as a result open column has high multicollinearity followed by high and low.

After all this bifurcation of univariate and bivariate we performed modelling on the dataset.
First, we performed linear regression as this was one of the main regression models in machine learning algorithm, then we performed lasso, ridge, KNN followed by Elasticnet and XGBoost regression models.

Before starting this modelling, we first split the data into subsets to get insight of that. For lasso regression we have an R-square value as 81.89%, for lasso and ridge we have 82% and 81.49%. These models have almost the same percentage as they used in a very similar way.
Moving forward we performed KNN as this is one of the best regression modelling techniques, which has high accuracy in their result, and we have R-square value for KNN as 92.01% followed by Elasticnet and XGBoost regression models which have 89% and 92% as an R-square value.
KNN and XGBoost regression models were considered as the good regressor models as their accuracy percentage was too correct.

After performing all the regression models, we analyzed and compared them by metrics comparison and got clear and cut comparison between all these. KNN has the high percentage as compared with all whereas linear having the least.
We also concluded that, **we are getting error of around 10% and may not be able to reduce this using any linear models as we have small amount of data, relation between target and input variables may be nonlinear.**

Contributor's Role:

1. **Kanika Singh**
    - Introduction

- Data briefing
- Data overview
- Exploratory data analysis
  Univariate analysis
    a. Independent variables
       (High, Open, and Low)
    b. Dependent variables
       (Close)
- Multicollinearity
- Modelling
  a. KNN
  b. ElasticNet regression
  c. XGBoost regression
- **Metric comparison**
- **Conclusion**

1. **Tanjul Gohar**
   - Introduction
   - Data briefing
   - Data overview
   - Exploratory data analysis
     Bivariate analysis
       a. Independent variables
          (High, Open, and Low)
       b. Dependent variables
          (Close)
   - Correlation
   - Modelling
     a. Linear regression
     b. Lasso regression
     c. Ridge regression
   - **Metric comparison**
   - **Conclusion**

Please paste the GitHub Repo link.

GitHub Link: - https://github.com/Kanika211/Yes_Bank_Stock_Closing_price.git

Please paste the drive link to your deliverables folder, ensure that this folder consist of the project Colab notebook, project presentation and video

Drive Link -
https://drive.google.com/drive/folders/1fGY_FkwXMk_db_7gDjAOYp1Fvb3gP3pf?usp=sharing