**Capstone Project Technical Documentation:**

|  |
| --- |
| **Name, Email and Contribution:** |
| 1. Abhishek Kumar Mishra ([Abhishekkumarmishra364@gmail.com](mailto:Abhishekkumarmishra364@gmail.com))  * Objective * Introduction * Problem Statement * Data Summary * EDA * Modelling * Conclusion |
| **GitHub Repo link.** |
| Github Link:- https://github.com/Abhishek709mis/Yes-bank-stock-predictions  Drive Link:- https://drive.google.com/drive/u/0/folders/1ar8Qc5YRVeCDX9Ysi4ftkMLmQdSXXR1b |
| * **Writing a short summary of your Capstone project and its components.** |
| **Objective:**  The main objective is to build a prediction model for the stock’s closing price of the month.  **Problem Statement:**  Yes Bank is known bank in the Indian financial domain. Since 2018, 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.  This dataset has monthly stock prices of the bank since its inception and includes closing, starting, highest, and lowest stock prices of every month.  **Importing Libraries:**  Some of the libraries like Num-Py for numerical operations, Pandas for data manipulation, matplotlib and seaborn for data visualization were loaded. In addition to these libraries, pycountry library was installed and loaded.  **Reading Data:**  After drive was mounted, data from csv file was read and store in a pandas dataframe.   * Date * Open * High * Low * Close   **Data analysis:**   * **Date**: It denotes date of investment done. * **Open**: Open means the price at which a stock started trading when the opening bell rang. * **High**: The maximum prices in a given time period. * **Low**: The minimum prices in a given time period. * **Close**: The price of an individual stock at the end of the considered time period   **Data Cleaning** :  Dealing with null values, duplicate data and outliers present in data.  **Exploratory Data Analysis:**   * Our dataset does not contain null values which tend to affect our accuracy. * Plotting the dependent variable and independent variables. * Checking and visualizing the correlation between our dependent and independent variables. * Visualizing the relationship between each pair of our variables. * **Data Visualization:** * **Univariate Analysis**:   Yes bank stock market dataset all the features have positively skewed distribution.   * **Bivariate Analysis:**   Supervised learning, it determine the essential predictors when the bivariate analysis is done by plotting one variable against another.  **Data Preprocessing & Feature Engineering** :   * Checking for and Dealing with multicollinearity present in our dataset. * Applying the log transform to deal with positively skewed data. * Scaling the data and splitting it into train and test sets       **Correlation Analysis:**  **heat.png** **Modelling:**  * **Linear** * Linear regression is one of the easiest and most popular Machine Learning algorithms. * It works best when there is a linear relationship between dependent and independent variables. * Root Mean Square Error is approximately 8.3917. * Adjusted R Square is approximately 0.8212. * **Lasso Regression (with cross-validation):**  The main of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable.  * RMSE for lasso regression is 8.3864. * Adjusted R-squared value is 0.9932.  **Ridge Regression with cross-validation:**  * Ridge regression is a regularized linear regression similar to lasso. * RMSE for Ridge regression is 8.3824 * Adjusted R-squared value for Ridge regression is 0.9932. * **Elastic Net Regression with cross-validation:** * Elastic net regression works in a manner that takes the best of lasso and ridge regressions. * RMSE for Elastic Net regression is 8.3760. * Adjusted R-squared value for Elastic Net regression is 0.9932   Model Implementation :-   * Fitting various models on our data and optimizing them via cross-validation. * Using these models to make predictions on test and train data. * The Models implemented are :-   1. Linear Regression  2. Lasso Regression  3. Ridge Regression  4. Elastic Net Regression Conclusion  * By data visualization on our target variable, clearly see the impact of 2018 fraud case involving Rana Kapoor as the stock prices decline dramatically during that period. * After loading the dataset, found that there are no null values in our dataset. * The distribution of all our variables is positively skewed, so performed log transformation on them. * There is a high correlation between the dependent and independent variables. * This is a signal that our dependent variable is highly dependent on our features and can be predicted accurately from them. * There is a rather high correlation between our independent variables. * This multicollinearity is however unavoidable here as the dataset is very small. * Implemented several models on our dataset in order to be able to predict the closing price and found that all our models are performing remarkably well and Elastic Net regressor is the best performing model with Adjusted R2 score value of 0.9932 and scores well on all evaluation metrics. * All of the implemented models performed quite well on our data giving us the Adjusted R-square of over 99%.   . |
|  |