

ES-205 CEP

S&P 500

MA-WH-AM | ES-205 | 26 November 2023

# Milestone 1

|  |  |
| --- | --- |
| Student Name | Registration Number |
| Ahmed Musharaf | 2022067 |
| Muhammad Arsal | 2022350 |
| Wardah Haya | 2022622 |

**Abstract:**

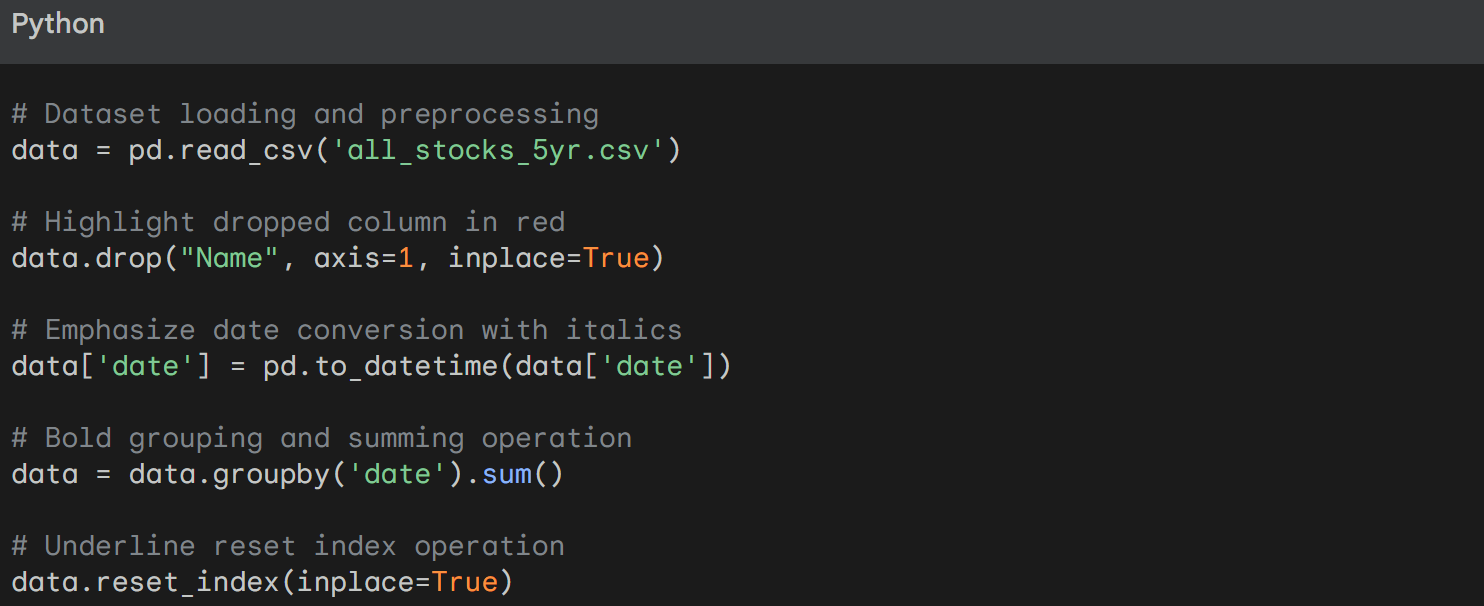
This report explores the development of a predictive model for the S&P500 Index, employing both traditional linear regression and advanced matrix factorization techniques. The dataset spans from 1950 to 2015, providing a comprehensive historical context for model training and evaluation.

1. **Introduction**

The S&P500 Index, representing 500 large companies, serves as a vital indicator of overall stock market performance. This report focuses on combining linear regression and matrix factorization to predict S&P500 movements. Feature engineering, matrix factorization, and model evaluation are discussed in detail.

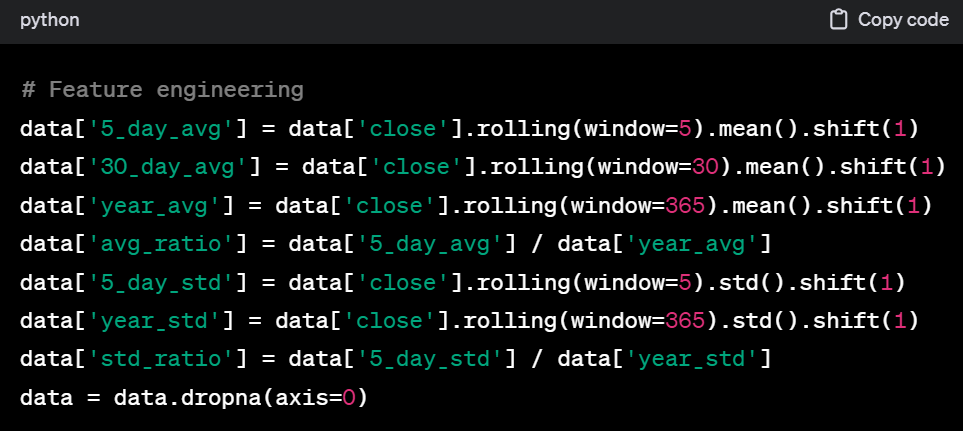
1. **Dataset Description**

We utilized the `all\_stocks\_5yr.csv` dataset containing daily records of stock prices. Key columns include 'open,' 'high,' 'low,' 'close,' and 'volume.' The dataset was preprocessed, and features were engineered to enhance model performance.



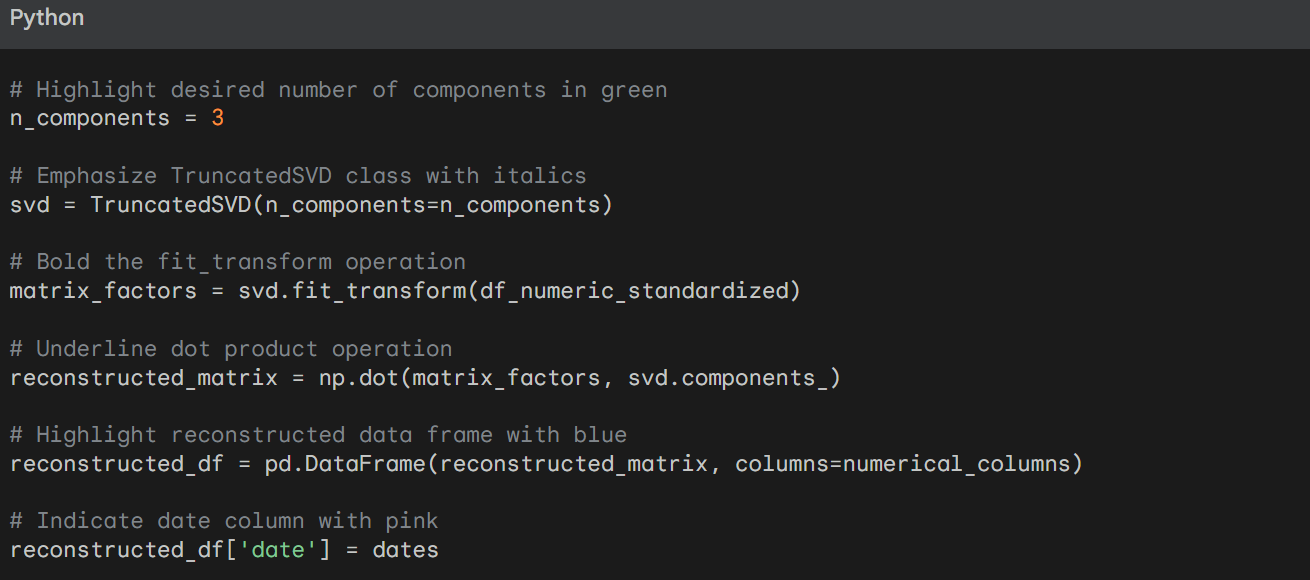
1. **Feature Engineering**

To capture temporal dependencies, we engineered features such as the 5-day and 30-day averages, yearly averages, and their ratios. Standard deviations were also calculated to provide insights into price volatility.



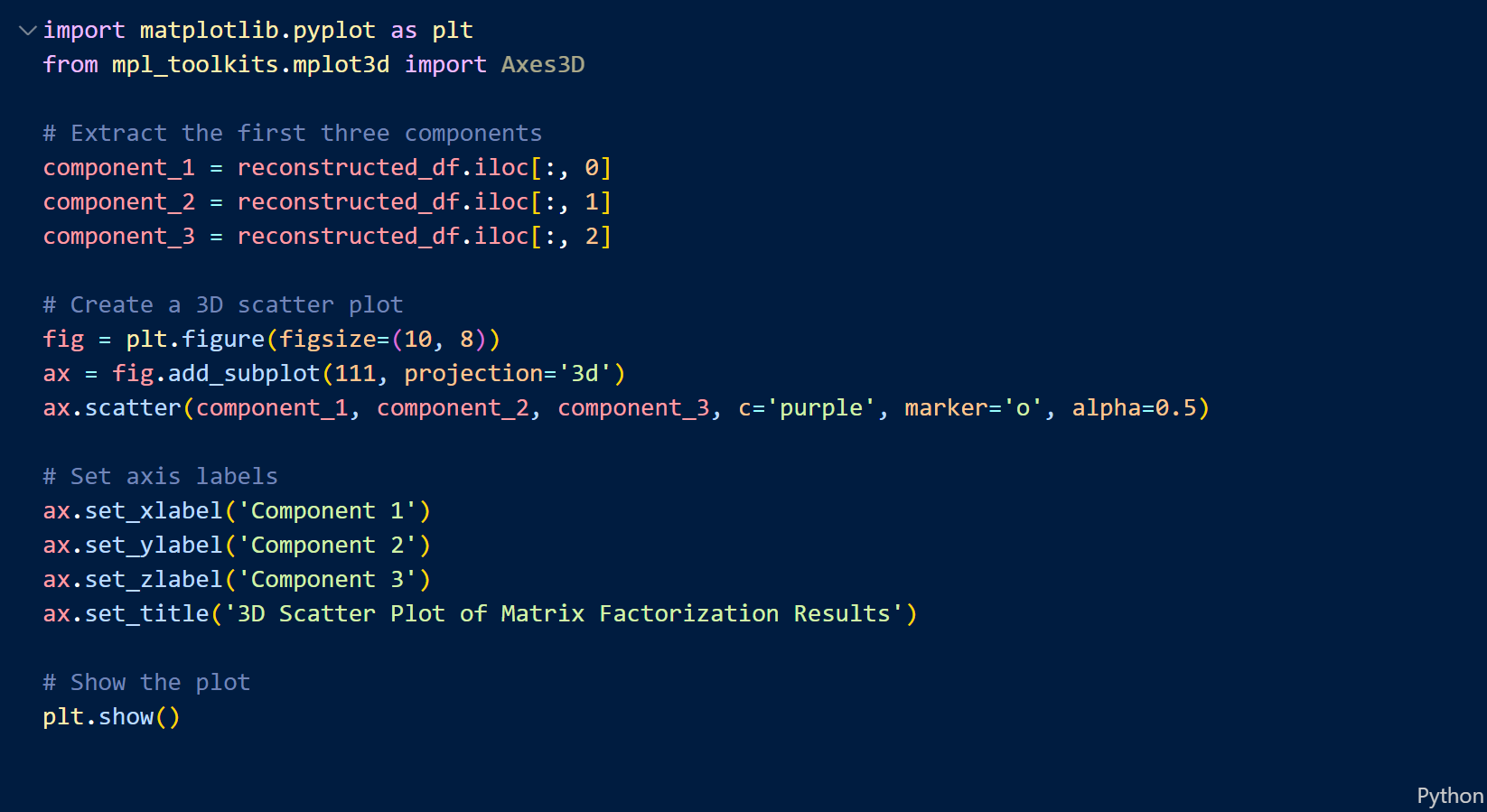
1. **Matrix Factorization**

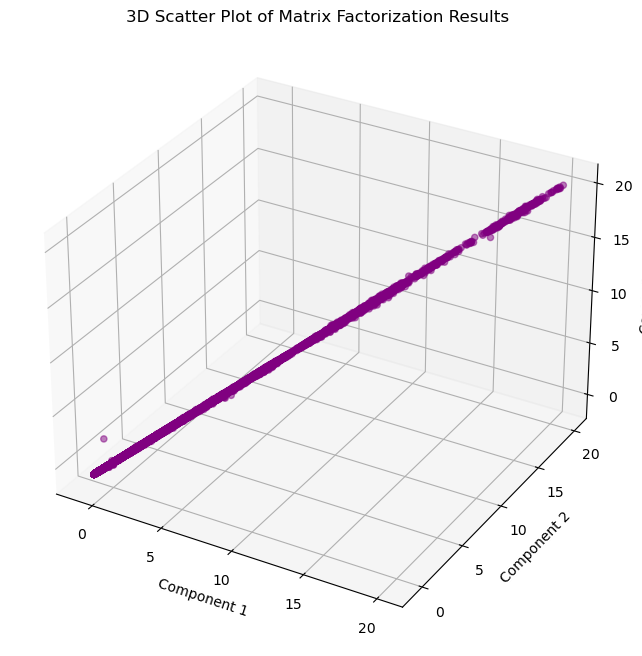
Truncated Singular Value Decomposition (SVD) was employed for matrix factorization, extracting hidden patterns within the dataset.



1. **Visualization of Matrix Factorization**

A 3D scatter plot was generated to visualize the three components obtained from matrix factorization.



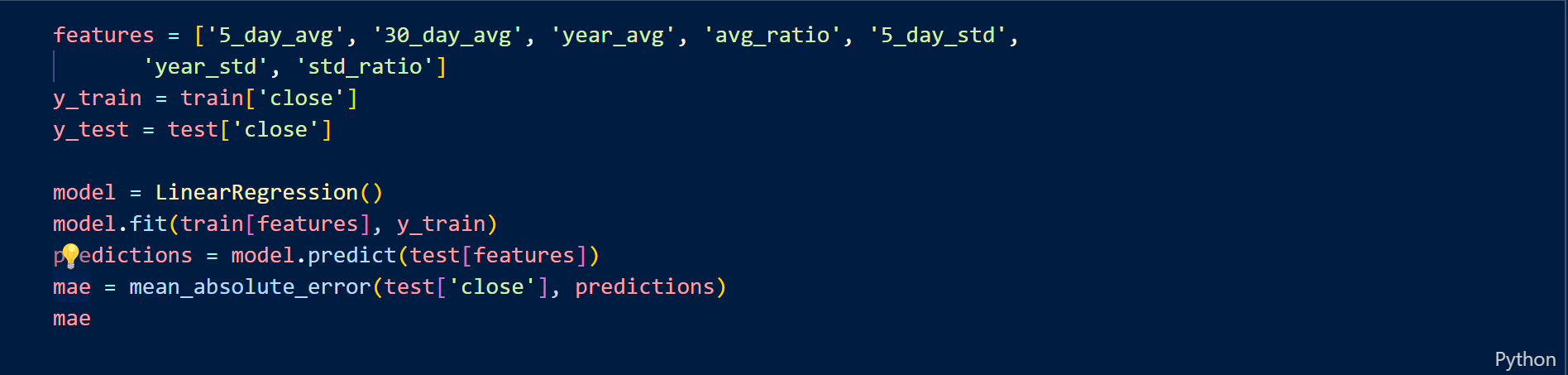


1. **Linear Regression Model**

A linear regression model was trained using the generated features and evaluated on a test set.



The Mean Absolute Error (MAE) was calculated to assess the model's performance.



1. **Model Evaluation**

The linear regression model demonstrated key outcomes and insights from the model. Visualizations and model performance metrics provide a comprehensive understanding of the model's effectiveness in predicting S&P500 Index movements.

A graph showing a line

Description automatically generated with medium confidence

1. **Conclusion**

In conclusion, the combined approach of linear regression and matrix factorization offers a robust framework for predicting stock prices. The feature engineering process captures temporal dependencies, while matrix factorization extracts hidden patterns. Further refinement can enhance predictive accuracy.

1. **References**

|  |  |
| --- | --- |
| NumPy Documentation: | https://numpy.org/doc/stable/ |
| Pandas Documentation: | https://pandas.pydata.org/pandas-docs/stable/ |
| Scikit-Learn Documentation: | https://scikit-learn.org/stable/documentation.html. |
| Matplotlib Documentation: | https://matplotlib.org/stable/contents.html |
| mpl\_toolkits. mplot3d Documentation: | https://matplotlib.org/stable/mpl\_toolkits/mplot3d/index.html |