Predicting the Closing Price of an S&P500 Stock

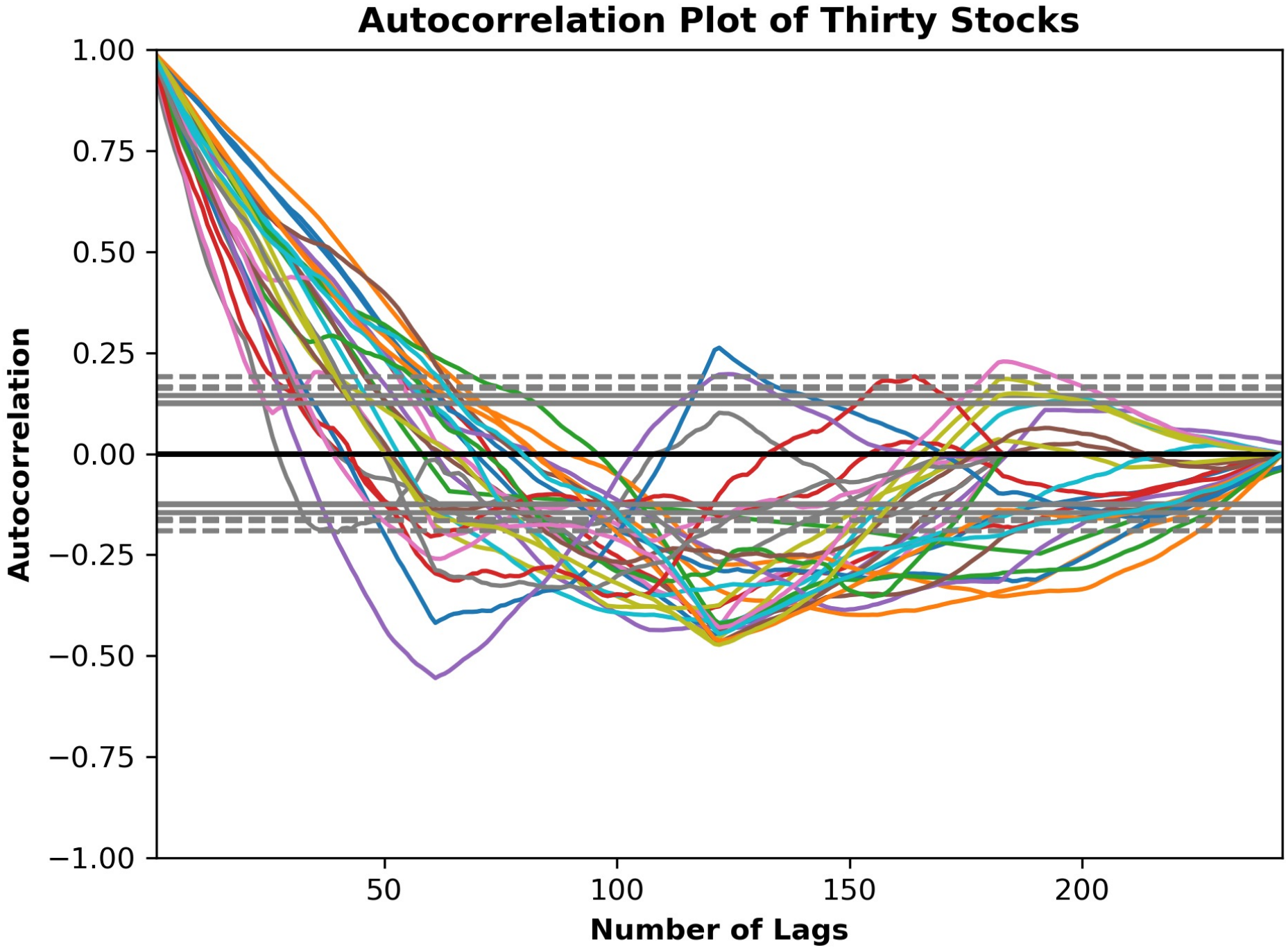
- Stefano Chiappo
- Brown University

- 6th of December 2023
- https://github.com/chistefano/Project_DATA1030.git

Introduction

- Regression problem
- The aim of this project is to predict the closing stock price based on quarterly financial metrics and previous prices
- Important to be able to invest in stocks whose price will increase

- Merged three different datasets: Prices, Securities, Fundamentals
- Data from Kaggle, sourced Prices and Securities from Yahoo Finance and Fundamentals from S&P Global


Autocorrelation Plot of Thirty Stocks
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Autocorrelation
50
100
150
200
Number of Lags

Dealing with Missing Values

- Dropped points without lagged or current closing price
- Remove some strongly corelated features

- Percentage of missing values in features:

- Quick Ratio 16.7%
- Earnings Per Share 5.91%
- Estimated Shares Outstanding 5.91%

- Iterative Imputer with Linear Regression Estimator

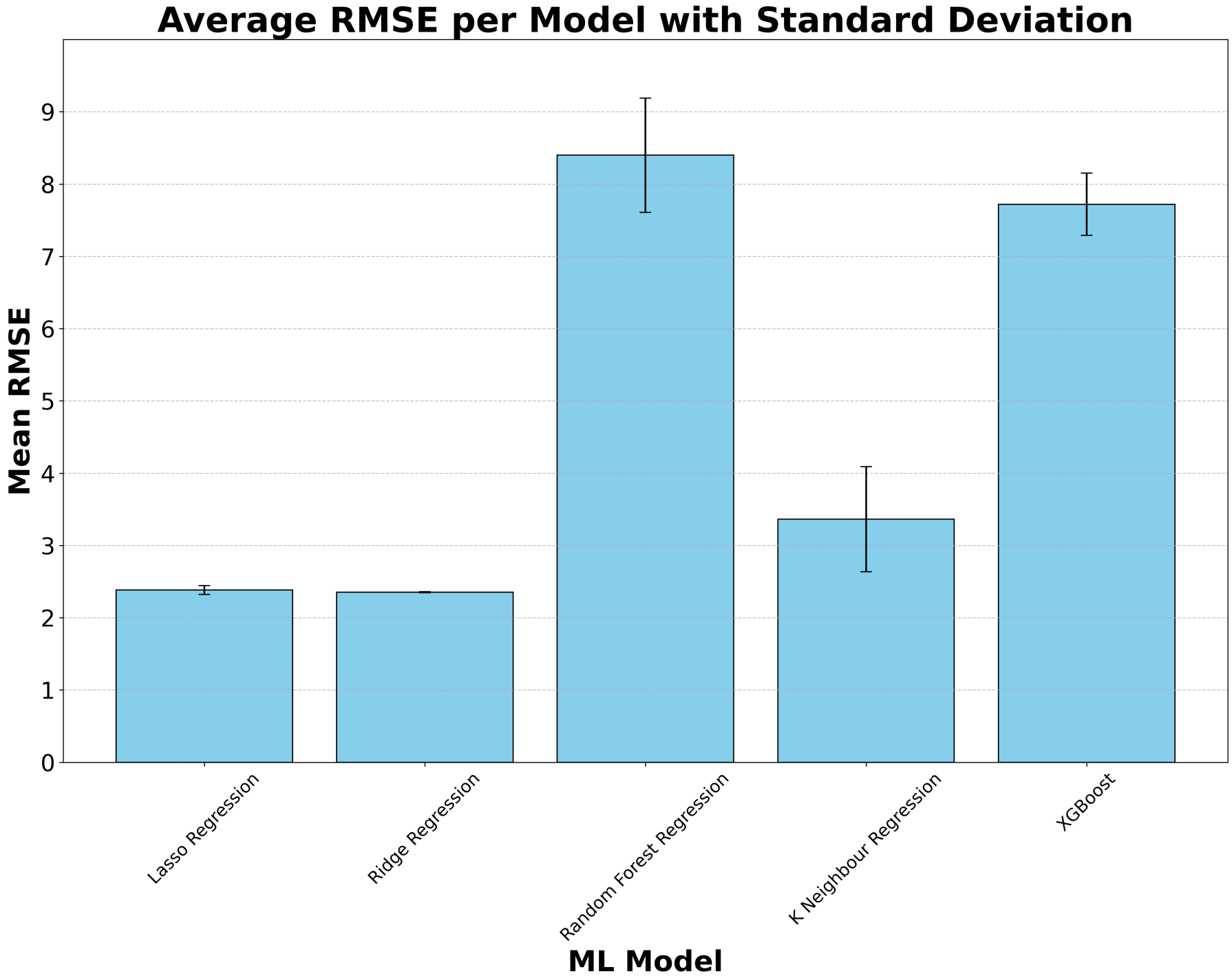
- Dataset size: (101198, 81)

Splitting and Preprocessing Data

- Time Series Data, so not iid
- Grouped by Stock Name and Date
- Used latest 20% of points for each stock as test set
- Used PredefinedSplit to create five, time-sensitive train and validation sets
- ‘GICS Sector’ only categorical feature, used OneHotEncoder
- Used Standard Scaler for all other (continuous) features

Models Used

ML Model	Hyperparameters Tuned	Average RMSE ($)	Standard Deviation Average RMSE ($)
Baseline Score	None	89.5	0.00
Lasso	max_iter, alpha	2.38	0.05
Ridge	alpha	2.35	0.08
Random Forest Regression	n_estimators, max_depth	3.36	0.73
K-neighbors Regressor	n_neighbors, weights	8.39	0.79
XGBoost	n_estimators, learning_rate, max_depth	7.72	0.43


Average RMSE per Model with Standard Deviation
Mean RMSE
0
1
2
3
4
5
6
7
8
9
Lasso Regression
Ridge Regression
Random Forest Regression
K Neighbour Regression
XGBoost
ML Model

Best Machine Learning Model

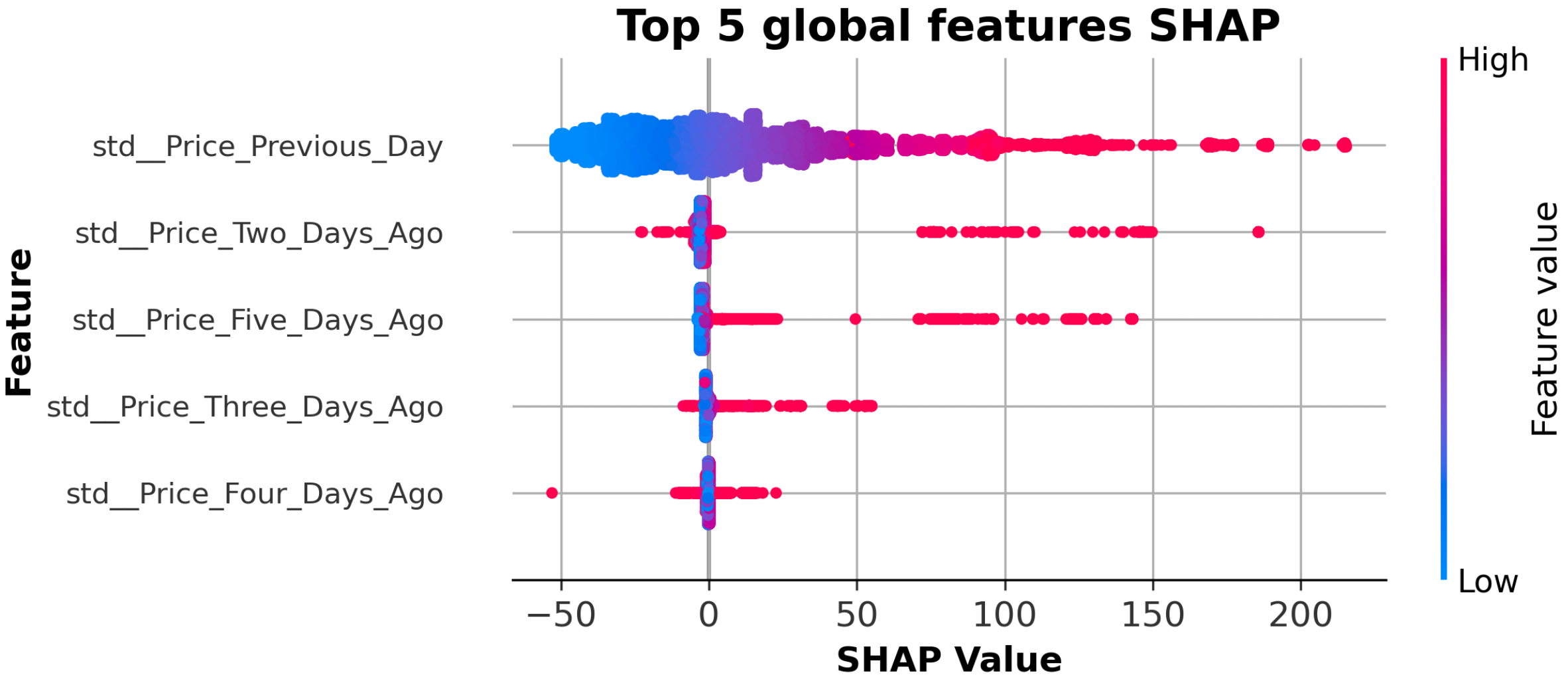
- Overall, the best model with its tuned hyperparameters was:
- Ridge Regression

- Alpha=1.67 e-09

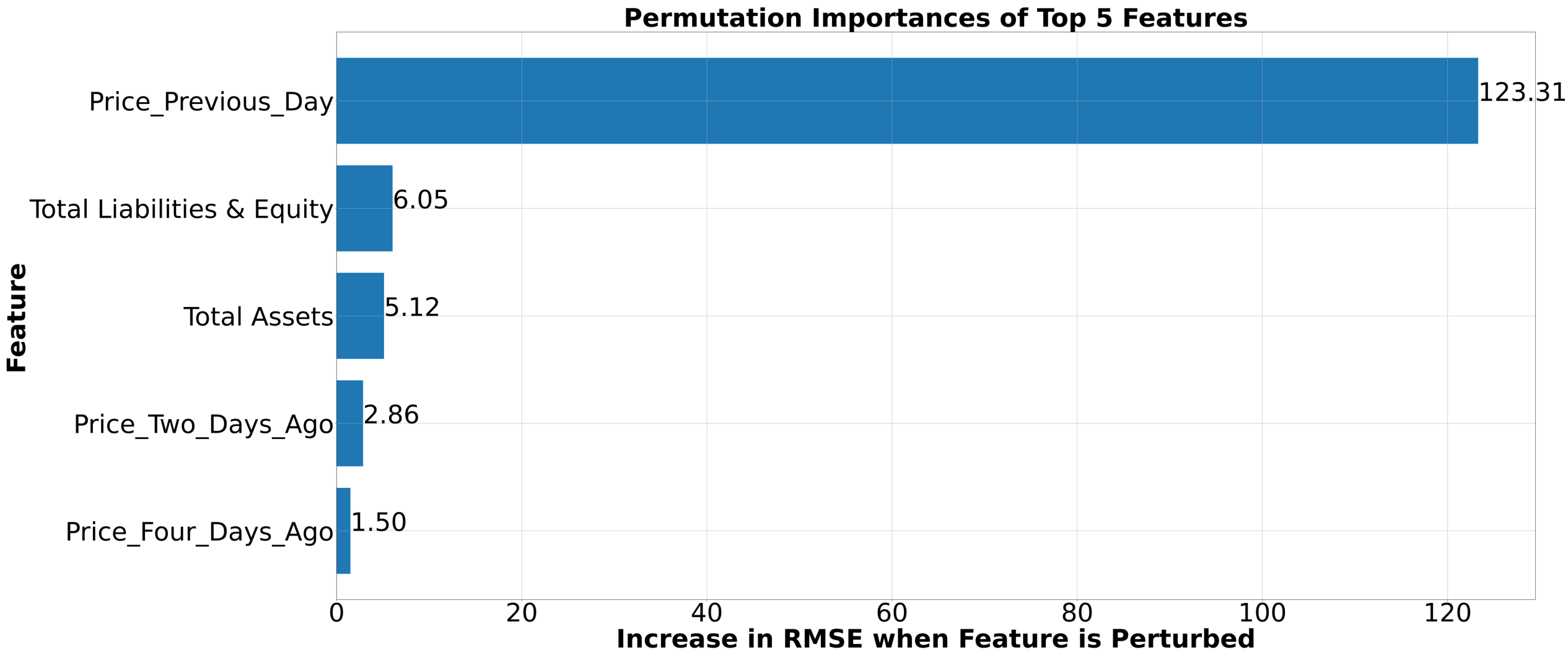
-RMSE: 2.35 ($)


Predicted Closing Price vs Actual Closing Price (in $)
Closing Price ($)
Predicted Closing Price ($)
0
200
400
600
800
1000
1200
1400

Global Feature Importance- SHAP



Global Feature Importance- Permutation



Improving Interpretability and Predictive Power

Interpretability

- Dropping highly correlated features
- Using less of the features which have a small impact on predicting the closing price of the stock

Predictive Power

- Try other ML models that are more time consuming, like SVR
- More hyperparameters can be tuned in the models
- Use more recent data to predict more accurately current prices

End of the Presentation