Boğaziçi University

Statistical Learning for Data Mining
IE 582

STOCK PRICE PREDICTION

Authors:
Kerem Tekeli
2022702066

15 January 2024

Introduction

In this study it is aimed to predict hourly stock prices(9:00 to 18:00) for the next day, provided data starting from 01.01.2018.Stock price forecasting has always been a topic of fascination to professionals in statistics and machine learning domains, as correct predictions ,or predictions that are sufficiently accurate to be clear, has the potential to create significant financial gains.In early 1960’s this phenomena was examined by Eugene Fama and it was presented, with strong deterministic evidence that all the past data is reflected on the current price and hence it is the best prediction for the next day.[1]The Efficient Market Hypothesis (EMH) outlines three forms that characterize the degree of information efficiency in financial markets. Weak form efficiency argues that historical price and volume data are already embedded in current market prices, rendering technical analysis ineffective in predicting future price changes. Semi-strong form efficiency extends this idea, stating that all publicly available information, including corporate disclosures and economic indicators, is swiftly and accurately incorporated into stock prices. Consequently, both technical and fundamental analyses struggle to consistently yield abnormal returns, as market prices rapidly adjust to new information. Strong form efficiency presents the strictest interpretation of the EMH, proposing that all information, whether public or private, is already reflected in stock prices. In this context, insider information provides no advantage, as it is purportedly absorbed into market prices. While the Efficient Market Hypothesis has had a significant impact on financial theory, ongoing scrutiny and debate persist, with critics citing instances of market anomalies and irrational behavior as indications of deviations from perfect efficiency[2].In this study different machine learning models with past stock price data will be trained and will be compared to naive model, which presumes no change in a given day.Time series cross validation will be implemented to decide whether any machine learning model performs better than the naive model in a statistically significant manner.In this study’s literature review part analysis of the past studies will be presented, then in the application and approaches part, methods used, the training process, modeling stage and approaches followed will be discussed.In the conclusion part results and future studies will be argued.

Literature Review

The assignment examined five articles focusing on stock prediction, each offering unique methodologies. Hsu et al. introduced a model that integrates Self-Organizing Map (SOM) and Support Vector Machine (SVM). Before SVM regression, the data undergoes clustering with SOM to enhance the representation of non-stationary behavior and structural breaks [3]. Kazem et al. presented a methodology utilizing the delay coordinate embedding method and Support Vector Regression (SVR) optimized by a chaos-based firefly algorithm [4]. The literature often explores hybridizing methods with different properties, as seen in the suggestion to use Ensemble Empirical

Mode Decomposition (EEMD) and hybridize SVR, Backpropagation Neural Network (BPNN), and Recurrent Neural Network (RNN) via Genetic Algorithm (GA) [5]. Another proposed model involves a straightforward average ensemble of Gradient-Boosted trees (GB), Deep Neural Networks (DNN), and Random Forests (RF) [6]. Textual-based predictions are also extensively studied. Nguyen et al. introduced a model implementing a sentiment-based approach, where specific topic sentiments were employed, setting it apart from other sentiment analysis-based models [7].

Approach

Validation

TimeSeries validation from sklearn library was used in this study.For training all the data till the separated day was used, for validation only the separated day was implemented.This procedure was conducted for a predefined times(n).Only a single day was used as validation set as we went to predict the next day only.Our task is best modeled with this approach.Crossvalidation techniques that requires using training data that is timestamped after validation data was avoided as it would have the potential to cause data leakage.

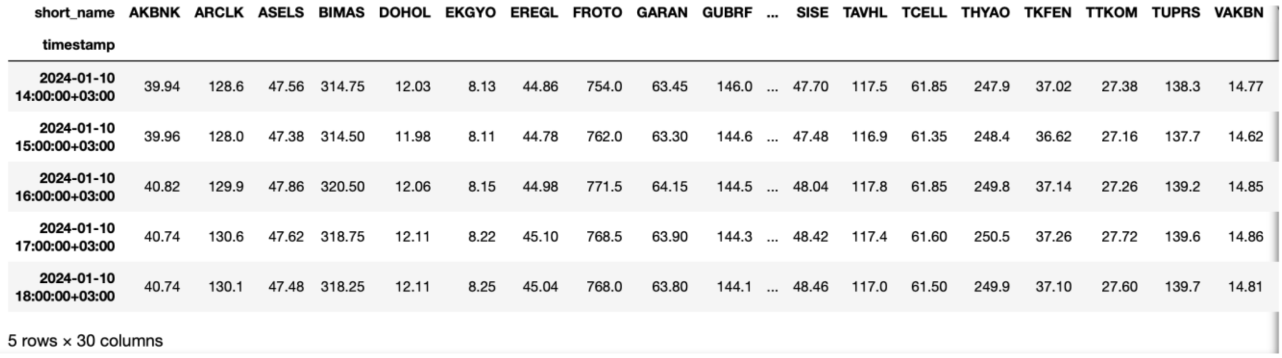
Regression

In the regression part of the study several statistical learning methods were implemented.ridge regression, lasso regression ,support vector regression and KNN regression are those stated models. Additionally custom made model that returns the average of ridge and lasso regression was implemented.The best among all these implementations were compared with the naive model and the result with lowest error metric was taken as the prediction.Naive model assumes price of a stock remains same.

APPLICATION

Preparing Data

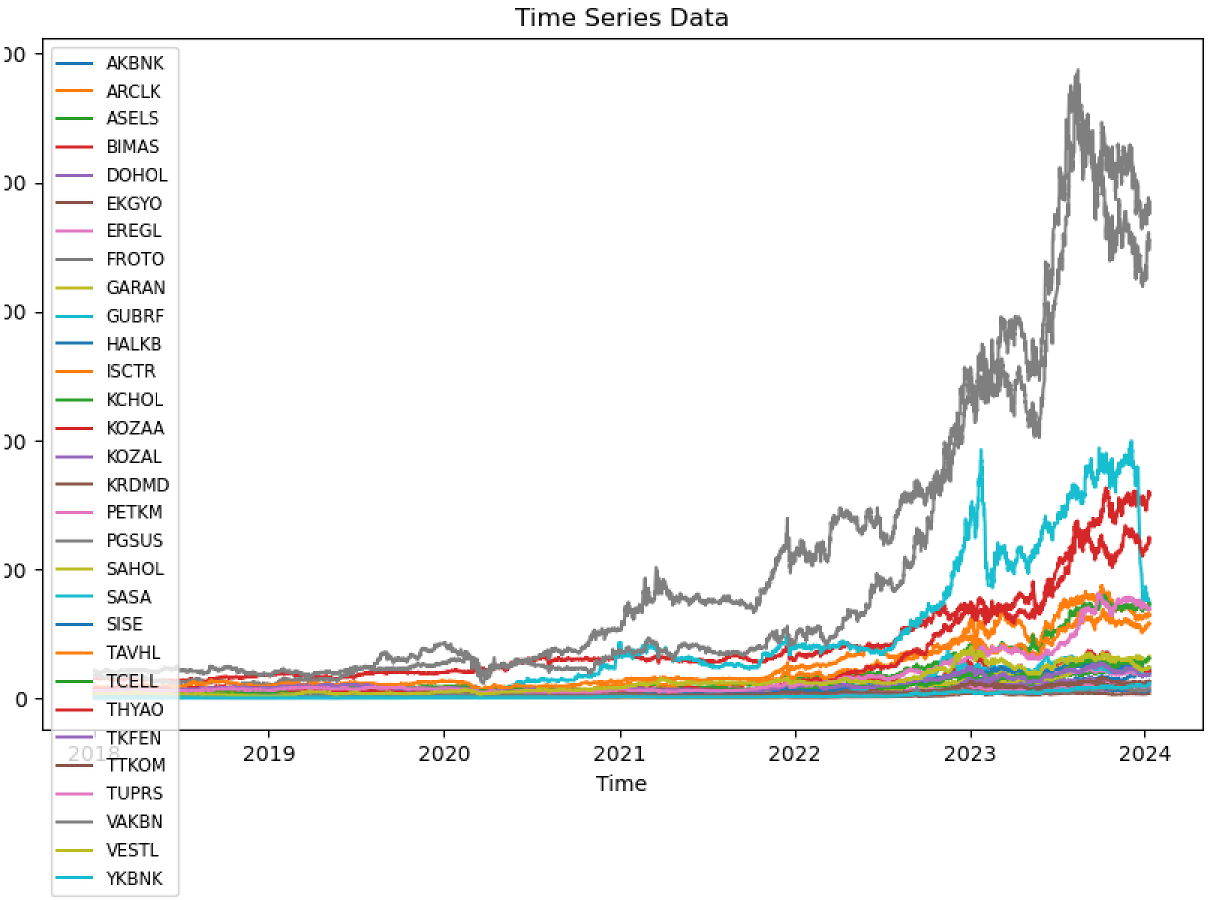
Initially data provided in the long format was transformed into short format.Down below last 5 values of the DataFrame was provided.In this dataframe every column corresponds to time series data of a stock.



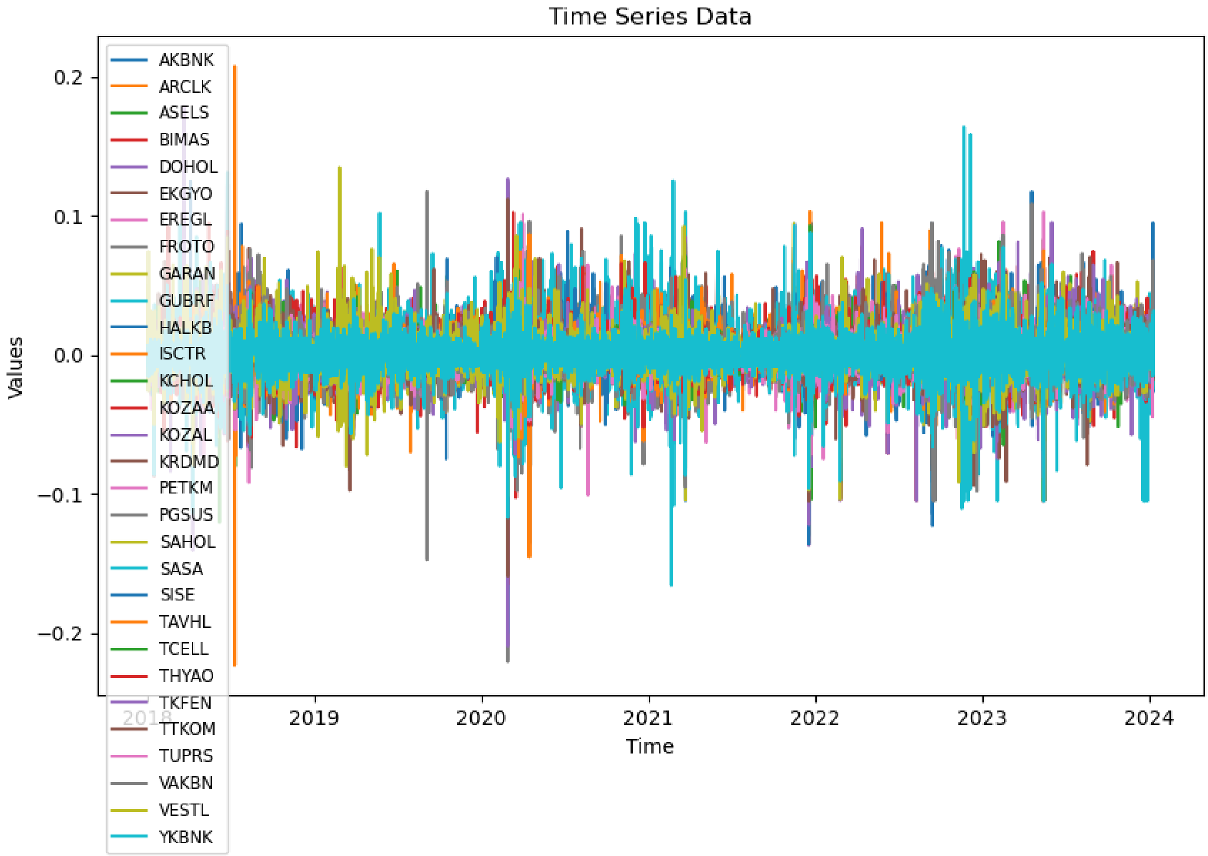
short_name	AKBNK	ARCLK	ASELS	BIMAS	DOHOL	EKGYO	EREGL	FROTO	GARAN	GUBRF	...	SISE	TAVHL	TCELL	THYAO	TKFEN	TTKOM	TUPRS	VAKBN
timestamp																			
2024-01-10 14:00:00+03:00	39.94	128.6	47.56	314.75	12.03	8.13	44.86	754.0	63.45	146.0	...	47.70	117.5	61.85	247.9	37.02	27.38	138.3	14.77
2024-01-10 15:00:00+03:00	39.96	128.0	47.38	314.50	11.98	8.11	44.78	762.0	63.30	144.6	...	47.48	116.9	61.35	248.4	36.62	27.16	137.7	14.62
2024-01-10 16:00:00+03:00	40.82	129.9	47.86	320.50	12.06	8.15	44.98	771.5	64.15	144.5	...	48.04	117.8	61.85	249.8	37.14	27.26	139.2	14.85
2024-01-10 17:00:00+03:00	40.74	130.6	47.62	318.75	12.11	8.22	45.10	768.5	63.90	144.3	...	48.42	117.4	61.60	250.5	37.26	27.72	139.6	14.86
2024-01-10 18:00:00+03:00	40.74	130.1	47.48	318.25	12.11	8.25	45.04	768.0	63.80	144.1	...	48.46	117.0	61.50	249.9	37.10	27.60	139.7	14.81

5 rows × 30 columns

To better feel the structure of our data time series were plotted for every stock our dataset posses.The obtained plot is provided below.As it can be observed visually there is a non linear trend.The trend must be gotten rid of otherwise scaling approaches will yield unstable results and our model will value error instances unevenly.(Larger values of stock price yield larger error, hence the model in question is likely to learn parameters that better predicts those larger values.)

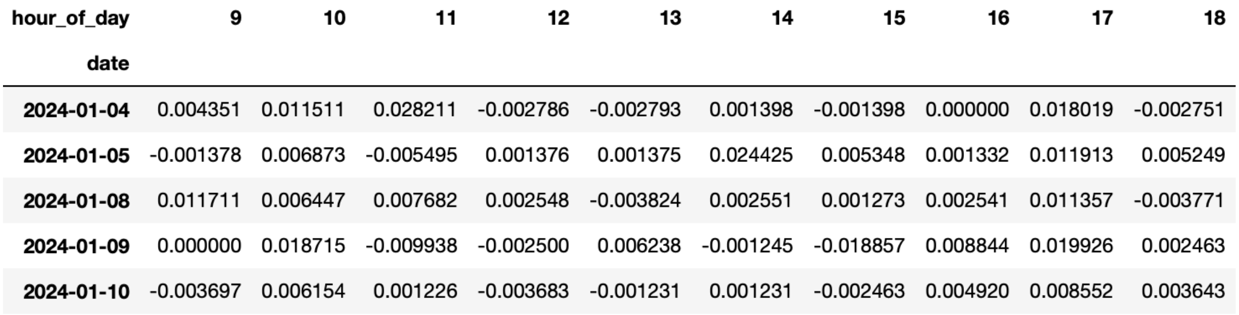


To overcome this.Time series data was converted into log returns for every stock price.By visual inspection it can be observed these time series are non stationary.Additionally Augmented-Dickey Fuller test was implemented and it was shown that the test rejected the notion that the time series in question are non-stationary.Anderson-darling test was implemented on the log return time series and the test rejected the hypothesis that the series is normally distributed, the test was conducted for each one of the series.Down Below the plot for log returns is provided.For charts, plots and results of the tests the reader can address provided codes.In addition reader could obtain summary statistics from the code provided.

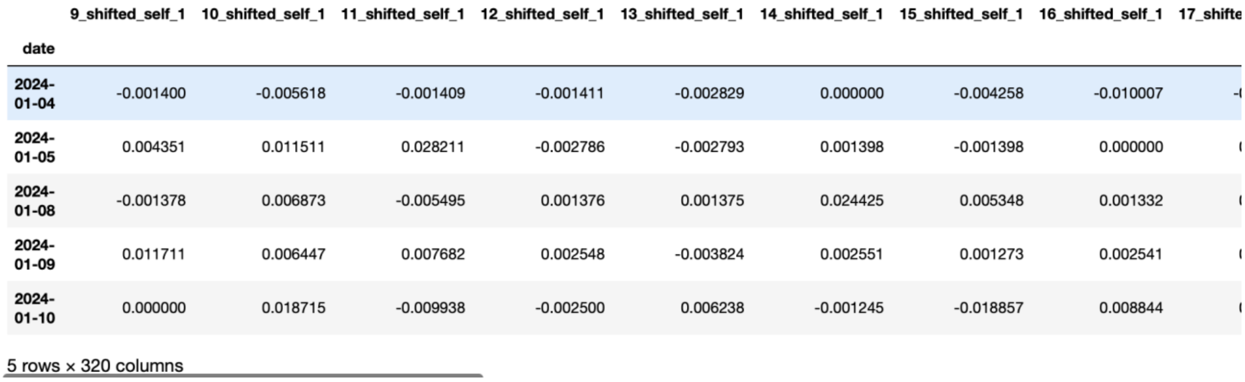


DataSet Creation

As it can be observed from the provided heatmap stock prices are correlated.In order to extract information from other stocks.One previous day stock price for other stocks was added to the features.Also three previous day stock price was added to the features as well.In the end to predict the stock price in a given hour next day.All hourly data of that stock for three days and one day past data for every other stock was used.Down below a sample input and output table was provided.Reader can access detailed codes and charts from the cde provided.First image is an example output.Second image is an example input.These outputs and inputs are created for every stock that is provided.

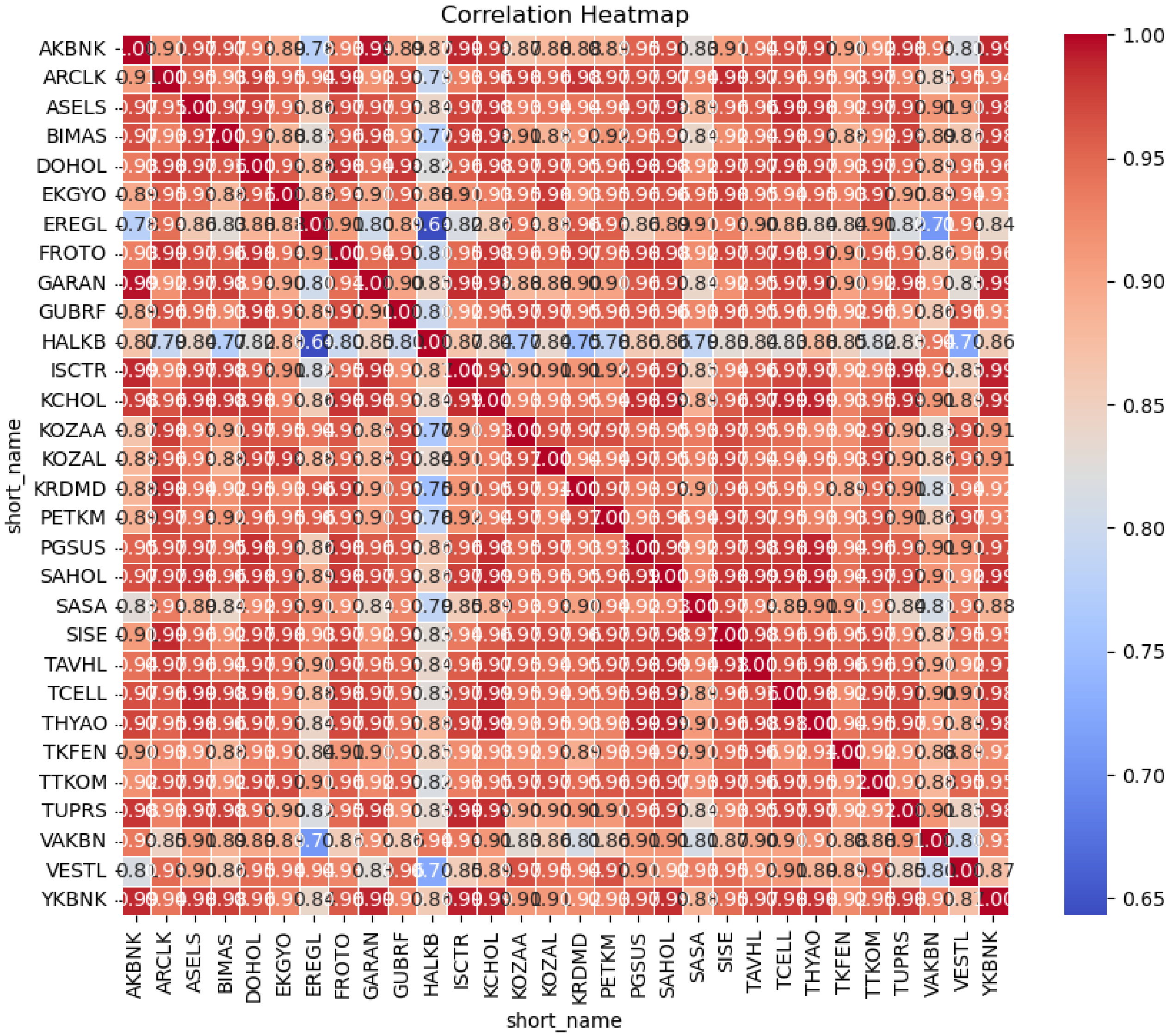


hour_of_day	9	10	11	12	13	14	15	16	17	18
date										
2024-01-04	0.004351	0.011511	0.028211	-0.002786	-0.002793	0.001398	-0.001398	0.000000	0.018019	-0.002751
2024-01-05	-0.001378	0.006873	-0.005495	0.001376	0.001375	0.024425	0.005348	0.001332	0.011913	0.005249
2024-01-08	0.011711	0.006447	0.007682	0.002548	-0.003824	0.002551	0.001273	0.002541	0.011357	-0.003771
2024-01-09	0.000000	0.018715	-0.009938	-0.002500	0.006238	-0.001245	-0.018857	0.008844	0.019926	0.002463
2024-01-10	-0.003697	0.006154	0.001226	-0.003683	-0.001231	0.001231	-0.002463	0.004920	0.008552	0.003643



date	9_shifted_self_1	10_shifted_self_1	11_shifted_self_1	12_shifted_self_1	13_shifted_self_1	14_shifted_self_1	15_shifted_self_1	16_shifted_self_1	17_shifte
2024-01-04	-0.001400	-0.005618	-0.001409	-0.001411	-0.002829	0.000000	-0.004258	-0.010007	-(
2024-01-05	0.004351	0.011511	0.028211	-0.002786	-0.002793	0.001398	-0.001398	0.000000	(
2024-01-08	-0.001378	0.006873	-0.005495	0.001376	0.001375	0.024425	0.005348	0.001332	(
2024-01-09	0.011711	0.006447	0.007682	0.002548	-0.003824	0.002551	0.001273	0.002541	(
2024-01-10	0.000000	0.018715	-0.009938	-0.002500	0.006238	-0.001245	-0.018857	0.008844	(

5 rows × 320 columns



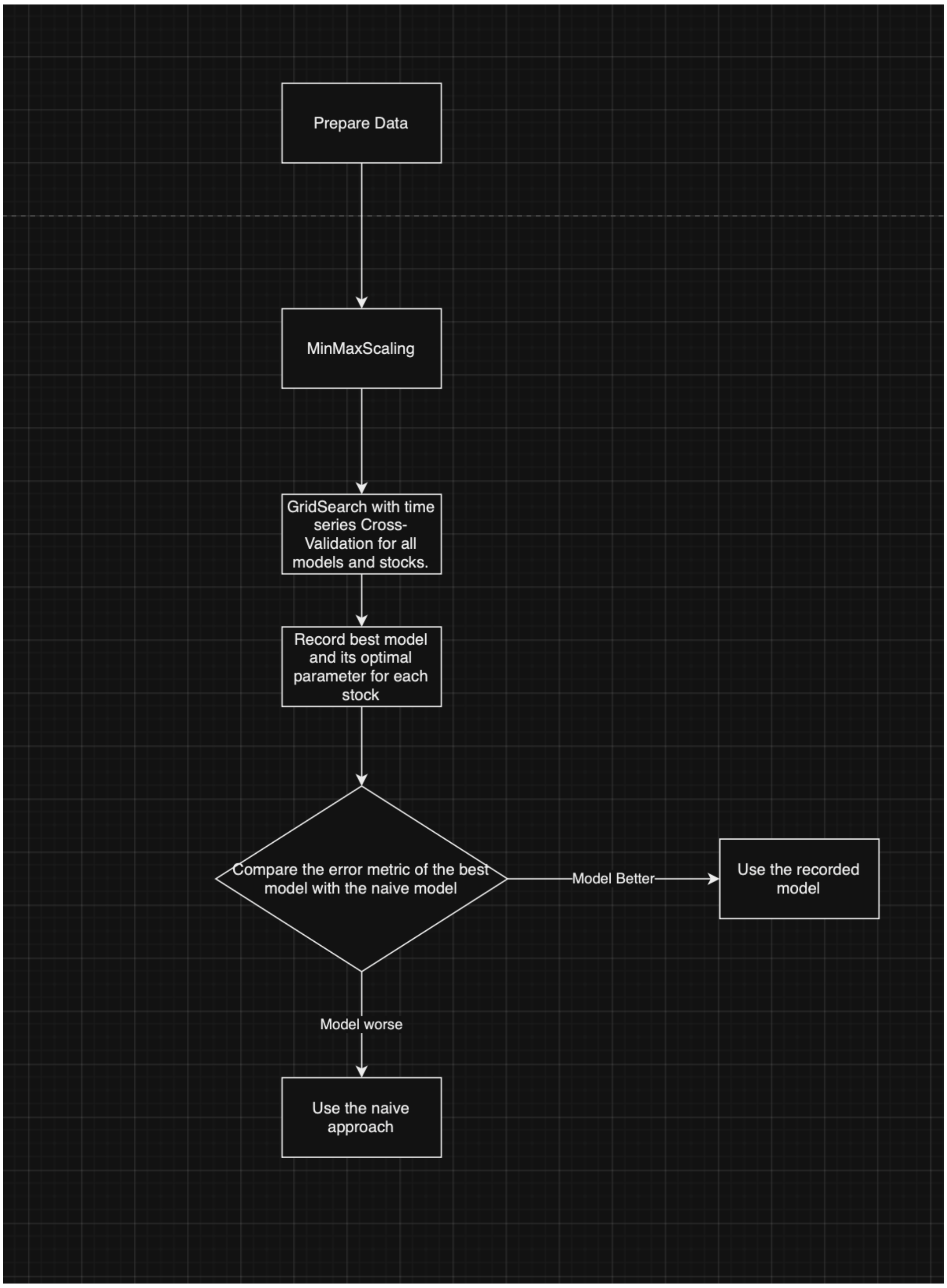
Regression

For this project multi output regressor from sklearn library was used.For every output different weights are produced by multi output regression.During parameter selection the parameter set with the lowest average error is selected.In this project a loop that conducts model and parameter selection were constructed.The constructed loop also records best parameters and the best model for every approach implemented. For every stock naive method and methods provided were implemented.The one with the lowest cross validation mse were selected as the best model.and its results were created as submission.Detailed codes of the process can be examined by checking the code file.Parameters used in models are provided below

Ridge	alpha	[0.01, 0.1, 1, 10]
Lasso	alpha	[0.01, 0.1, 1, 10]
SVR	C	[0.1, 1, 5, 10, 20]
SVR	kernel	['linear', 'rbf', 'poly']
Ensemble	svr_params	[{'C': 5, 'kernel': 'linear'}, {'C': 5, 'kernel': 'rbf'}, {'C': 1, 'kernel': 'linear'}, {'C': 1, 'kernel': 'rbf'}]
Ensemble	ridge_params	[{'alpha': 0.01}, {'alpha': 0.1}, {'alpha': 10}]
KNN	n_neighbors	[3, 5, 7]
KNN	weights	['uniform', 'distance']

Pipeline

In the pipeline initially min max scaling was implemented, in order to squeeze variables into same interval.(0,1)By conducting scaling it is ensured that in penalty terms model does not overvalue certain parameters and it is also implemented in order not to cause instability in the learning process.After scaling for the previously selected models gridsearch over the provided parameters were conducted.Model and corresponding parameter that provides results with lower error metric was saved for all stocks.Error metric from the best model was compared with the naive model, approach with the lowest error metric was implemented for final prediction.


Prepare Data
MinMaxScaling
GridSearch with time series Cross-Validation for all models and stocks.
Record best model and its optimal parameter for each stock
Compare the error metric of the best model with the naive model
Model Better
Use the recorded model
Model worse
Use the naive approach

Conclusion

Although powerful models were implemented,In all instances,apart from some isolated cases, naive method yielded superior results.The results obtained in this project supports weak efficient market hypothesis but it must be highlighted that it does not directly proves it.The results obtained does not contradict what the weak form of efficient market hypothesis suggest however it cannot be classified as proof.To obtain meaningful results sentiment data and news data can be added to the model ,by this way, contingent upon semi strong version of efficient market hypothesis is invalid meaningful results can be obtained.Statistical learning models used in this study may be implemented on exchanges that does not present efficient market behavior. Cryptocurreny exchanges is a strong candidate, considering its high volatility and non-rational agents conducting transactions in it.

REFERENCES

[1]. Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(2), 383. https://doi.org/10.2307/2325486

[2]. Malkiel, B. G. (1989). Efficient Market Hypothesis. Finance, 127–134. https://doi.org/10.1007/978-1-349-20213-3_13

[3]. Hsu SH, Hsieh JJP-A, Chih TC, Hsu KC. A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression. Expert Syst Appl. 2009;36(4):7947–7951. Available from: https://doi.org/10.1016/j.eswa.2008.10.065

[4]. Kazem A, Sharifi E, Hussain FK, Saberi M, Hussain OK. Support vector regression with chaos-based firefly algorithm for stock market price forecasting. Appl Soft Comput. 2013;13(2):947–958. Available from: https://doi.org/10.1016/j.asoc.2012.09.024

[5]. Alhnaity B, Abbod M. A new hybrid financial time series prediction model. Eng Appl Artif Intell. 2020;95:103873. Available from: https://doi.org/10.1016/j.engappai.2020.103873

[6]. Krauss C, Do XA, Huck N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. [The reference is incomplete, and the complete details were not provided. Please provide the full reference for accurate citation.]

[7]. Nguyen TH, Shirai K, Velcin J. Sentiment analysis on social media for stock movement prediction. Expert Syst Appl. 2015 Dec 15;42(24):9603-9611. Available from: https://doi.org/10.1016/j.eswa.2015.07.052