“ ชื่อกลุ่ม ”

Stock Return Prediction

Outline

1. Problem Setup
2. Metrics
3. Methods Tried
4. Results
5. Error Analysis

Problem Setup

Objectives

- To develop machine learning models to **predict stock return** using historical data
- To evaluate and compare their results using various metrics and **back testing** algorithm

Dataset

S&P500 Stock Prices (Yahoo Finance)

- 10 Years
- Start Date: 2014-05-01
- End Date: 2024-05-01

Assumptions

IID

Independent and Identically Distributed random variables

- Data
  - stationary time series
- Regression Problem
- Classification Problem
  - Binary

Time Series

- Data
  - multivariate time series
  - stationary
- Regression Problem

Data Preprocessing

Download data

- from *yfinance* library
- format to: pandas.**MultiIndex** object
  - index: Ticker, Date
  - columns: OHLCV
    - use *Adj_Close* instead of *Close* Price

	Price	Open	High	Low	Close	Volume
Ticker	Date					
AAPL	2014-05-01	21.142857	21.242857	20.941429	18.581392	2.440480e+08
	2014-05-02	21.155001	21.221430	21.061071	18.615948	1.915144e+08
	2014-05-05	21.076429	21.464287	21.071428	18.879210	2.870672e+08
	2014-05-06	21.492857	21.586071	21.228930	18.673445	3.745644e+08
	2014-05-07	21.258928	21.331785	20.990356	18.608095	2.828644e+08
...	...	...	...	...	...	...
^GSPC	2024-04-24	5084.859863	5089.479980	5047.020020	5071.629883	3.656740e+09
	2024-04-25	5019.879883	5057.750000	4990.580078	5048.419922	3.958050e+09
	2024-04-26	5084.649902	5114.620117	5073.140137	5099.959961	3.604140e+09
	2024-04-29	5114.129883	5123.490234	5088.649902	5116.169922	3.447450e+09
	2024-04-30	5103.779785	5110.830078	5035.310059	5035.689941	4.082470e+09

OHLCV

IID

- Return
  - lag: 1d, 1w, 2w, 1m, 2m, 3m
- Dollar Volume
- Date Indicator
  - Year & Month
- Technical Indicator
  - Relative Strenght Index (RSI)
  - Bollinger Band
  - Average True Range
    - ATR & Normalize ATR
  - Moving Average Convergence/Divergence (MACD)
  - Momentum
    - Lag

Time Series

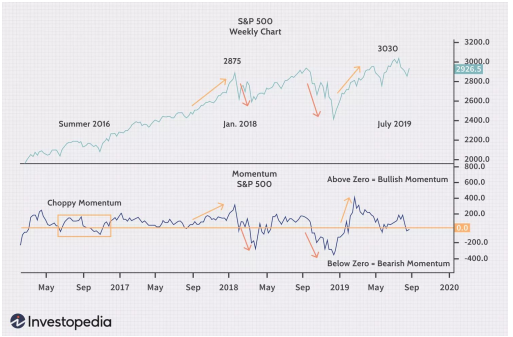
Log Return of OHLCV

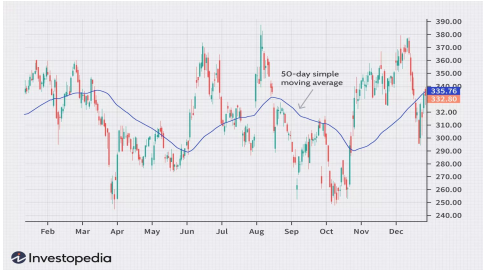
- Log Return Close Price
- Technical Indicator
- Month

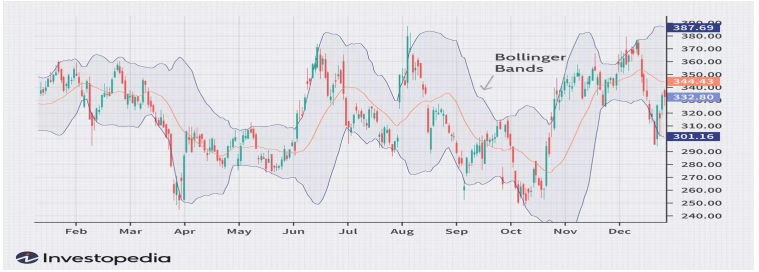
Technical Indicators

To preserve time series properties when sampling using IID

- Return
- Dollar Volume
- Relative Strenght Index (RSI): overbought or oversold
- Bollinger Band: high and low price relative to each other
- Average True Range: volatility of the market
    - ATR & Normalize ATR
- Momentum: price trend

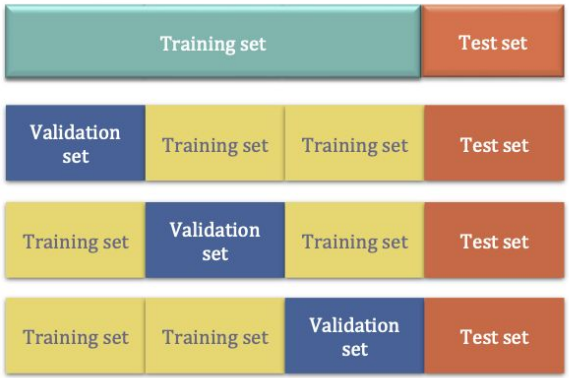


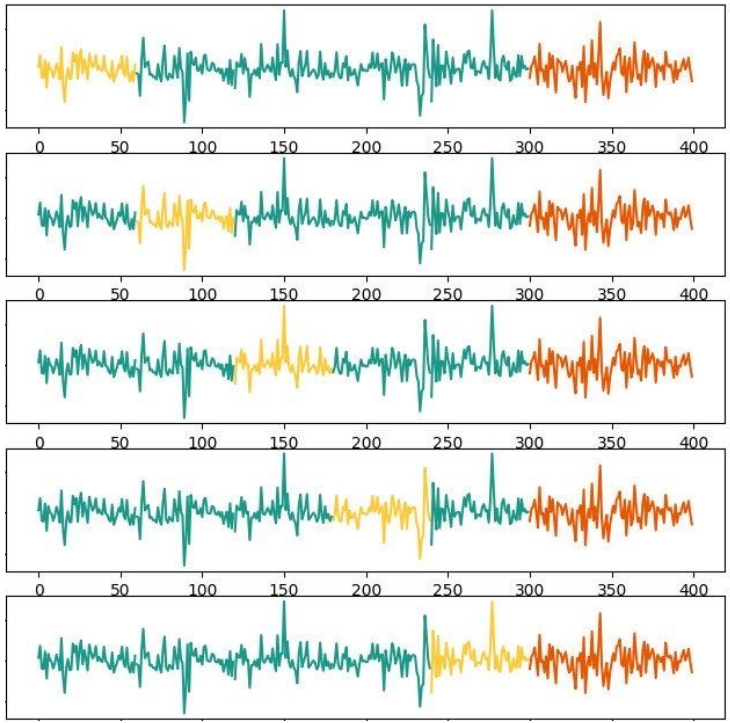




Data Split

- **Train and Test Set:** 2014/05 - 2023 (9 years)
- **Trading Evaluation:** 2023/05 - 2024 (1 year)





Metrics

Metrics

Regression

- MAE
- MSE
- RMSE
- Direction

Classification

- Accuracy
- Precision
- Recall
- ROC

Backtest

- Cumulative Return
- Sharpe Ratio

Methods Tried

Models

Baseline Model: Naive Forecasting

Using today's price as a prediction for tomorrow.

Classical

Regression

- Linear Regression
- Support Vector Regression (SVR)
- Random Forest Regression (RFR)
- Extreme Gradient Boosting (XGBR)

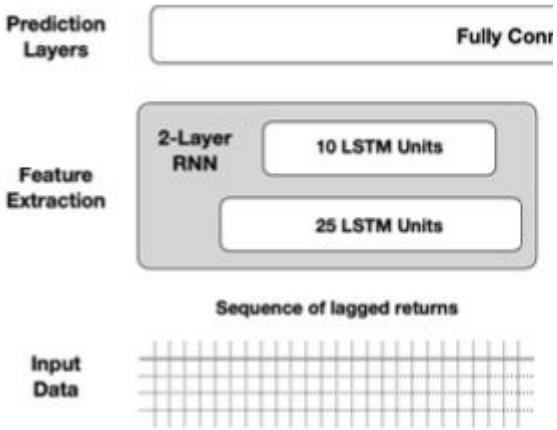
Classification

- Logistic Regression
- Support Vector Classification (SVC)
- Random Forest Classification (RFC)
- Extreme Gradient Boosting (XGBC)
- K-Nearest Neighbor (KNN)

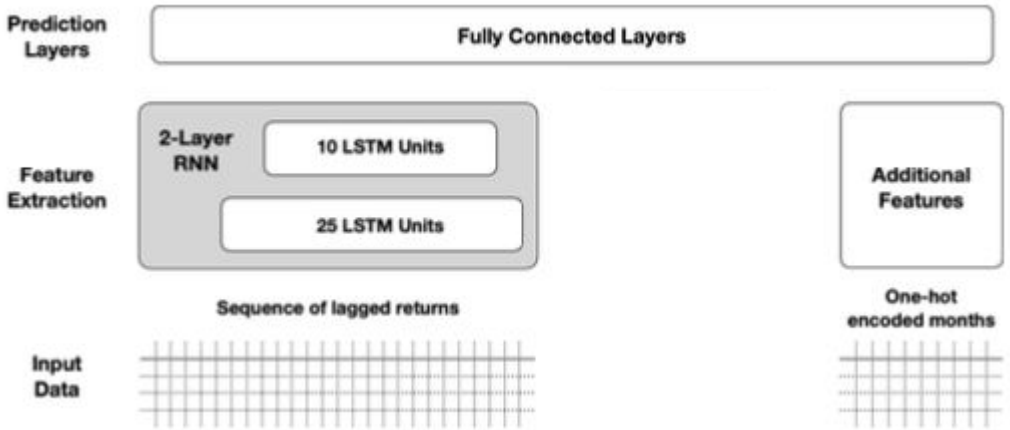
Neural Network

- Long Short Term Memory (LSTM)
- Gated Recurrent Unit (GRU)

NN Architecture

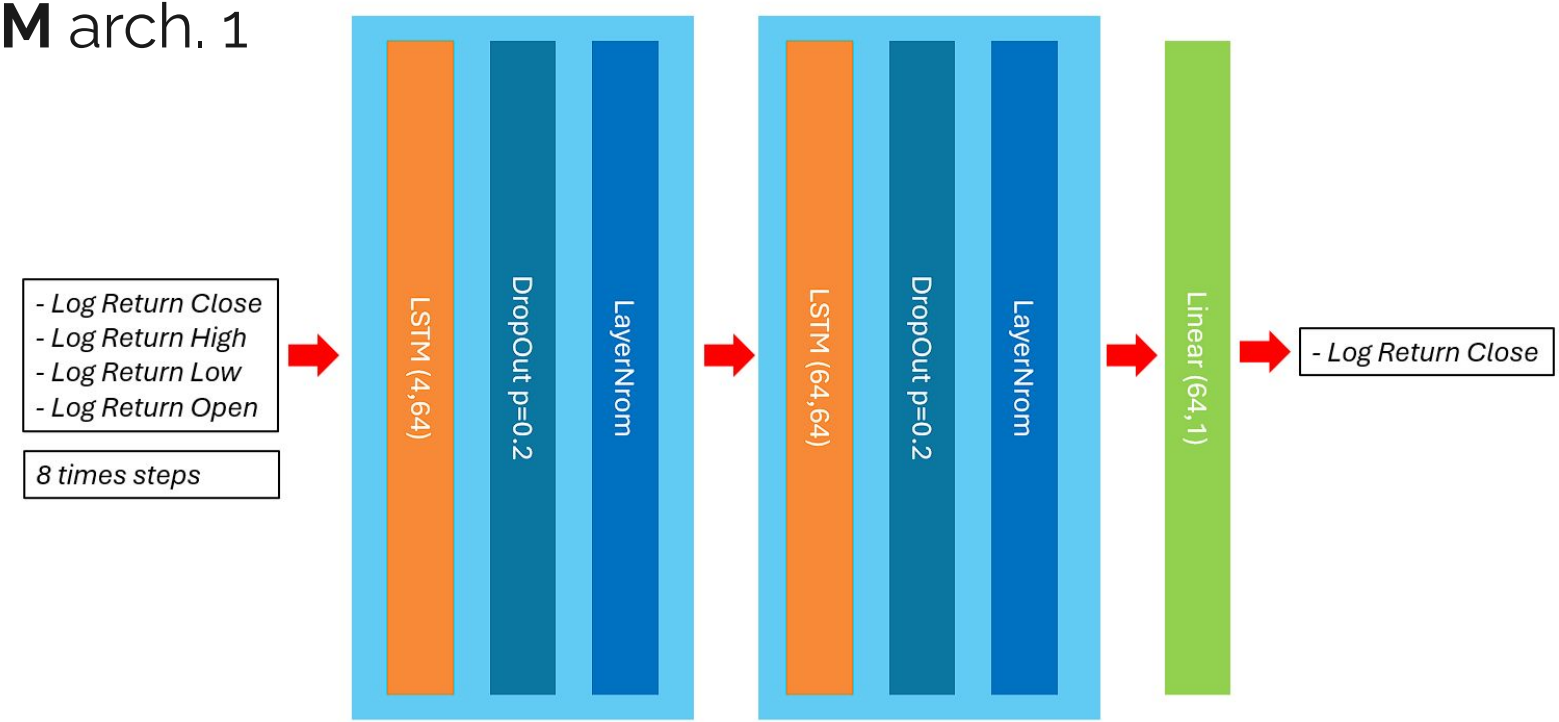


Arch. 1

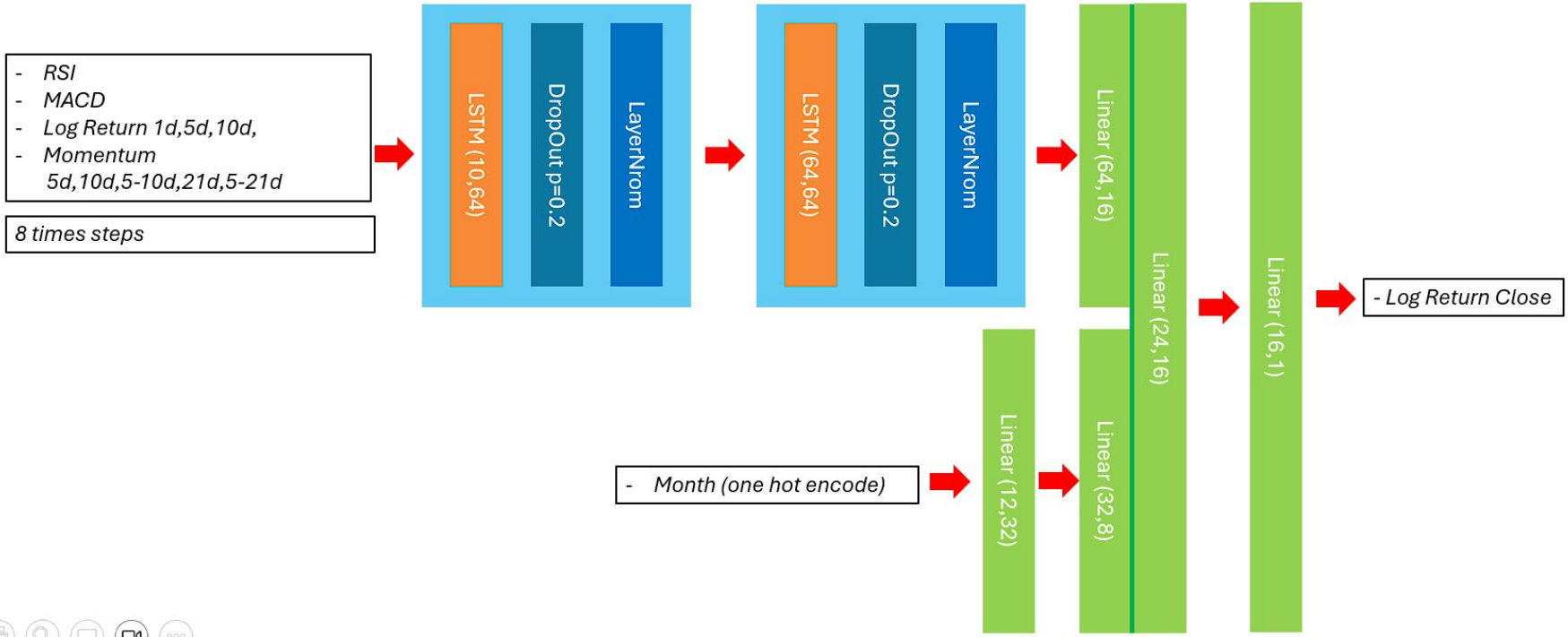


Arch. 2

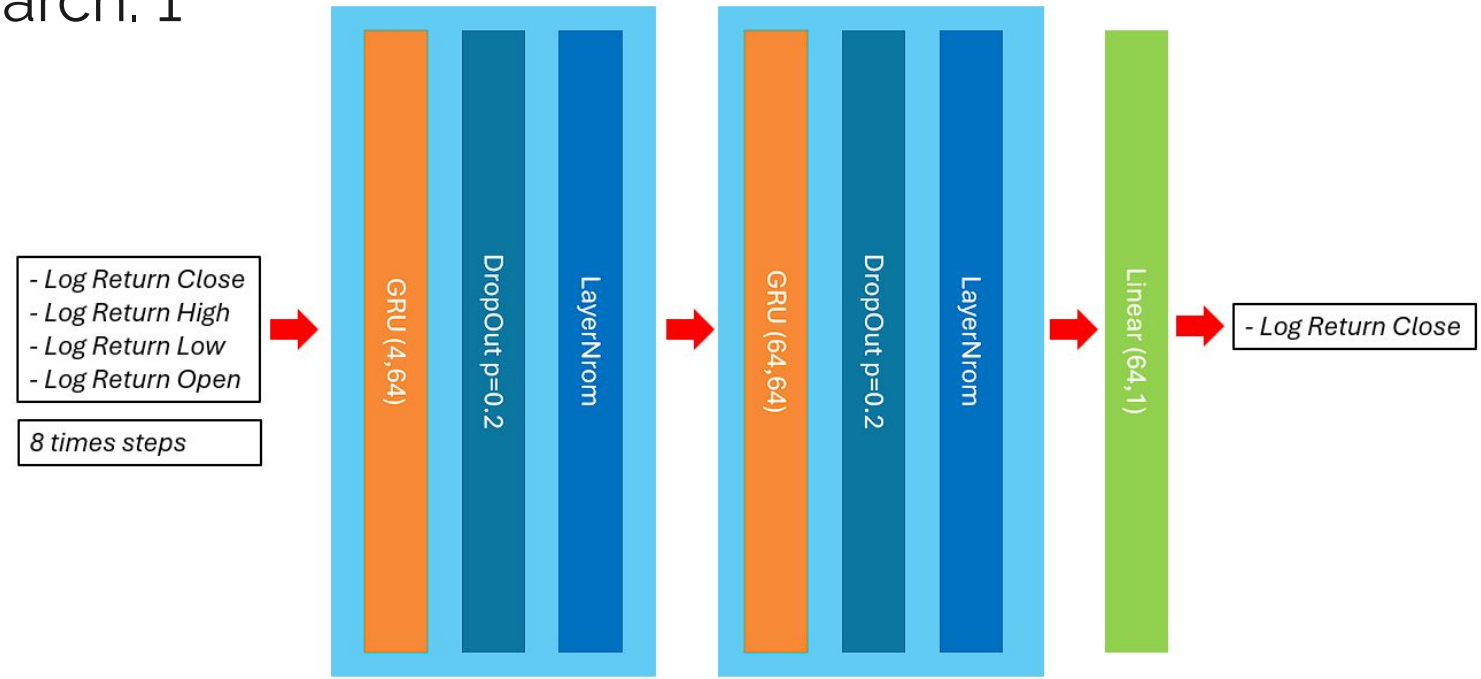
LSTM arch. 1


- Log Return Close
- Log Return High
- Log Return Low
- Log Return Open
8 times steps
LSTM (4,64)
DropOut p=0.2
LayerNrom
LSTM (64, 64)
DropOut p=0.2
LayerNrom
Linear (64, 1)
- Log Return Close

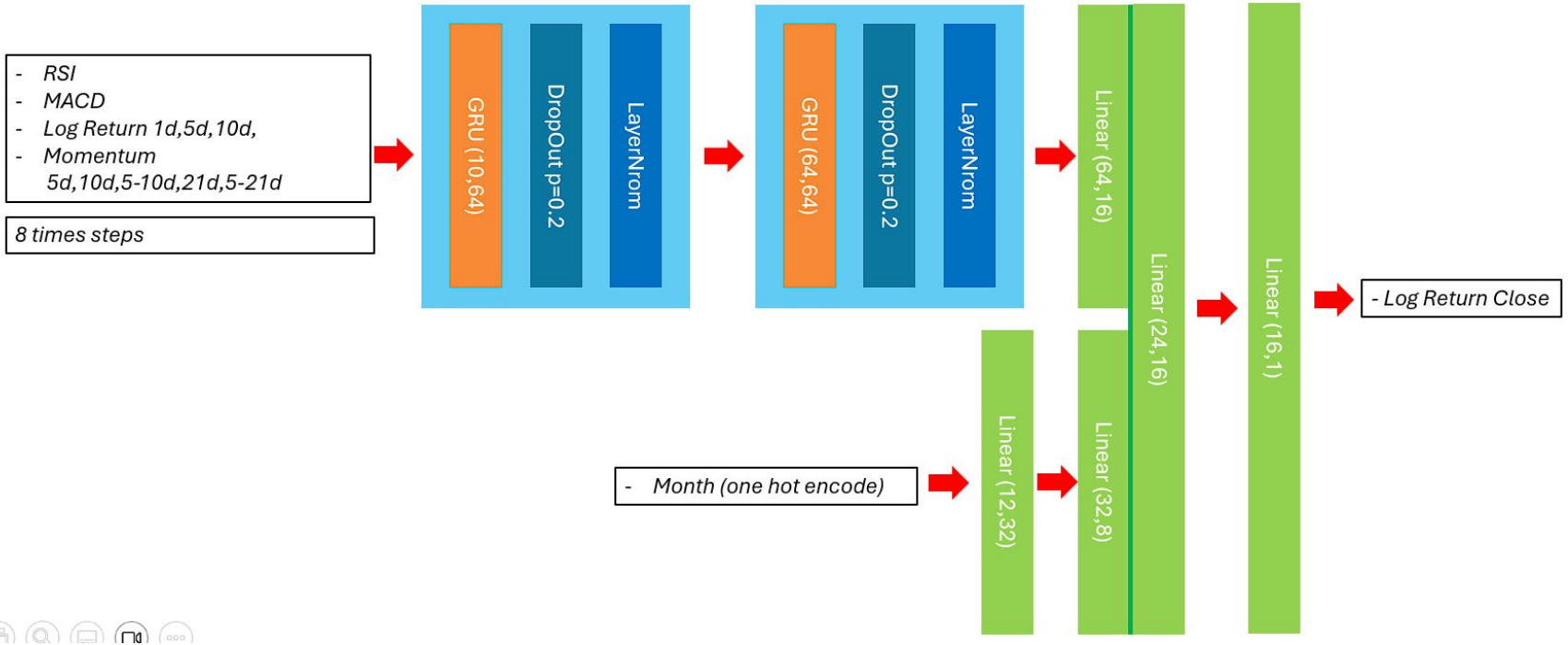
LSTM arch. 2


- RSI
- MACD
- Log Return 1d,5d,10d,
- Momentum
5d,10d,5-10d,21d,5-21d
8 times steps
LSTM (10,64)
DropOut p=0.2
LayerNrom
LSTM (64,64)
DropOut p=0.2
LayerNrom
Linear (64,16)
- Month (one hot encode)
Linear (12,32)
Linear (32,8)
Linear (24,16)
Linear (16,1)
- Log Return Close

GRU arch. 1


- Log Return Close
- Log Return High
- Log Return Low
- Log Return Open
8 times steps
GRU (4,64)
DropOut p=0.2
LayerNrom
GRU (64,64)
DropOut p=0.2
LayerNrom
Linear (64,1)
- Log Return Close

GRU arch 2.


- RSI
- MACD
- Log Return 1d,5d,10d,
- Momentum
5d,10d,5-10d,21d,5-21d
8 times steps
GRU (10,64)
DropOut p=0.2
LayerNrom
GRU (64,64)
DropOut p=0.2
LayerNrom
Linear (64,16)
Linear (24,16)
Linear (16,1)
- Log Return Close
- Month (one hot encode)
Linear (12,32)
Linear (32,8)

Hyperparameter Tuning

GridSearchCV

A “brute force” approach to hyperparameter optimization. We fit the model using all possible combinations after creating a grid of potential discrete hyperparameter values.

K-Fold

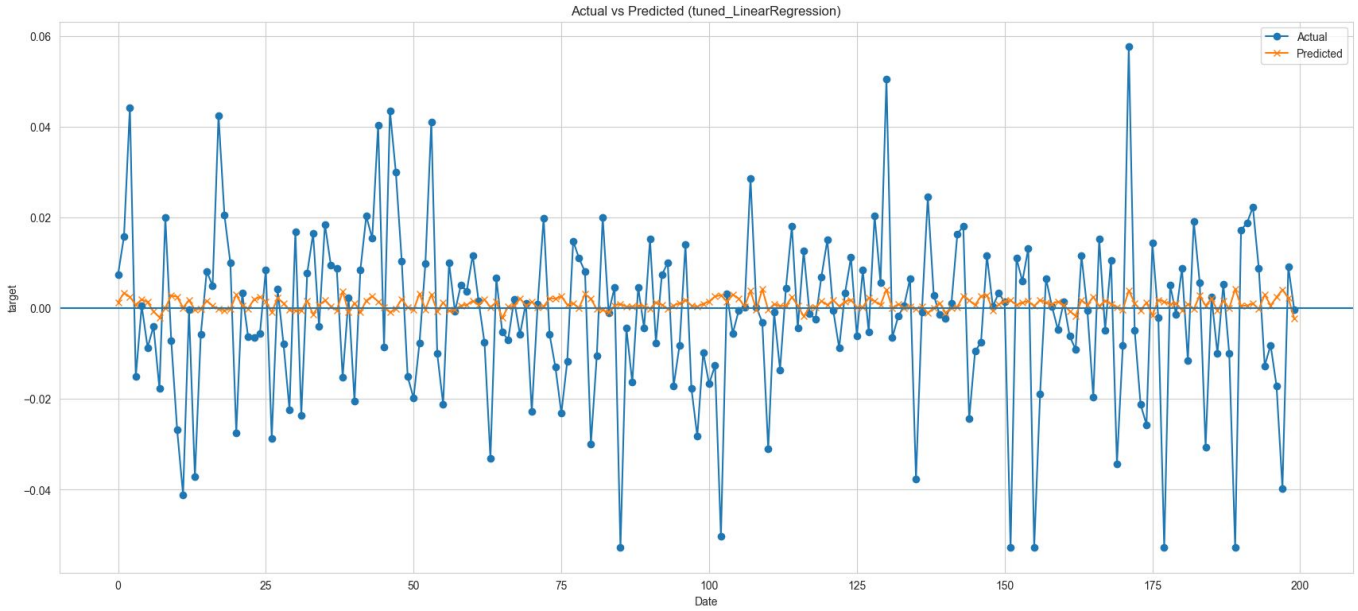
An approach for estimating the skill of the models. (Set K = 5)

Hyperparameter Tuning

- **Linear Regression:** fit_intercept
- **Logistic Regression:** penalty, C
- **SVM:** kernel, degree, epsilon
- **RF:** min_samples_leaf, max_dept
- **XGB:** eta, max_depth, subsample
- **KNN:** algorithm

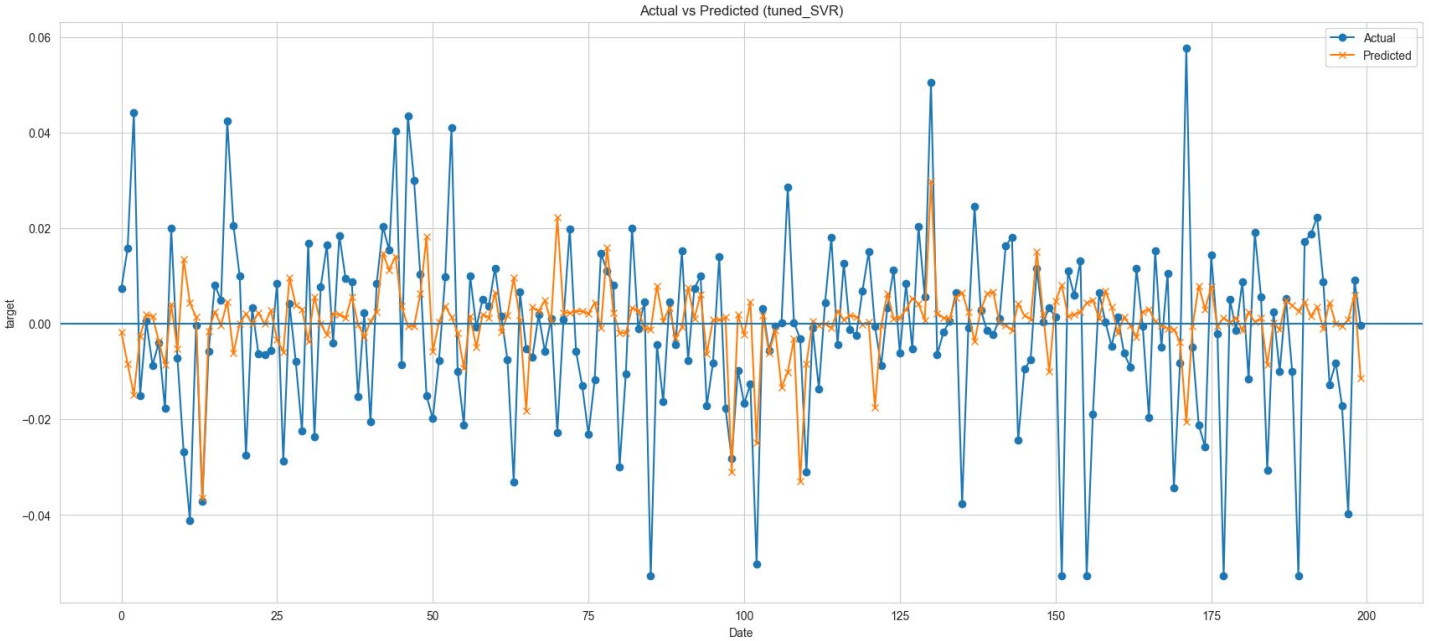
Results

Linear Regression



MAE: 0.012494
MSE: 0.000307
RMSE: 0.017518
Direction: 0.525280

SVR



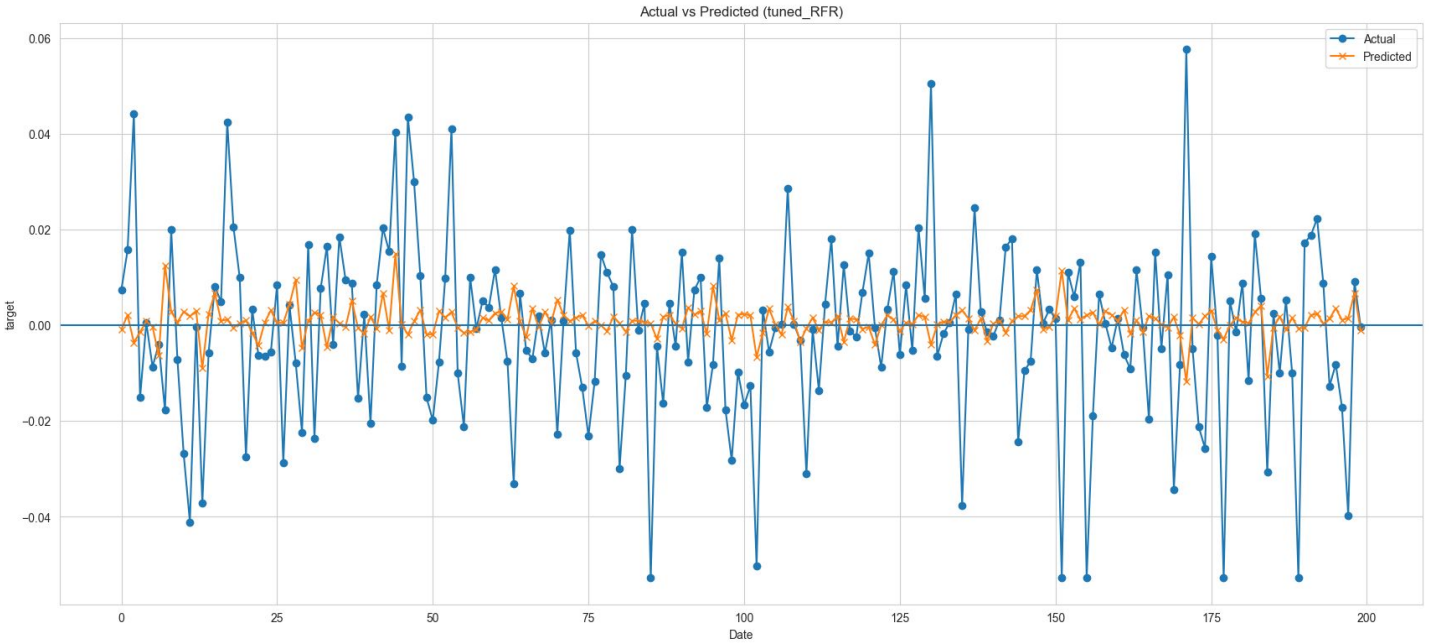
MAE: 0.012559

MSE: 0.000313

RMSE: 0.017682

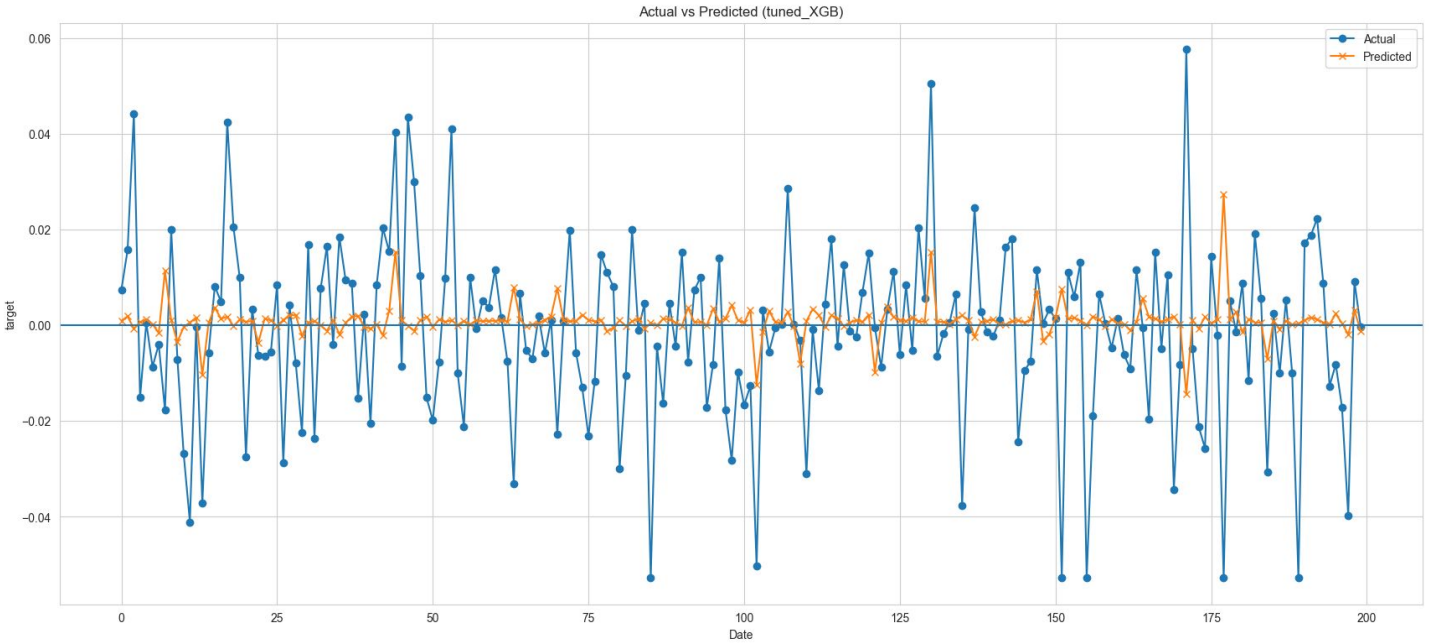
Direction: 0.538574

RFR



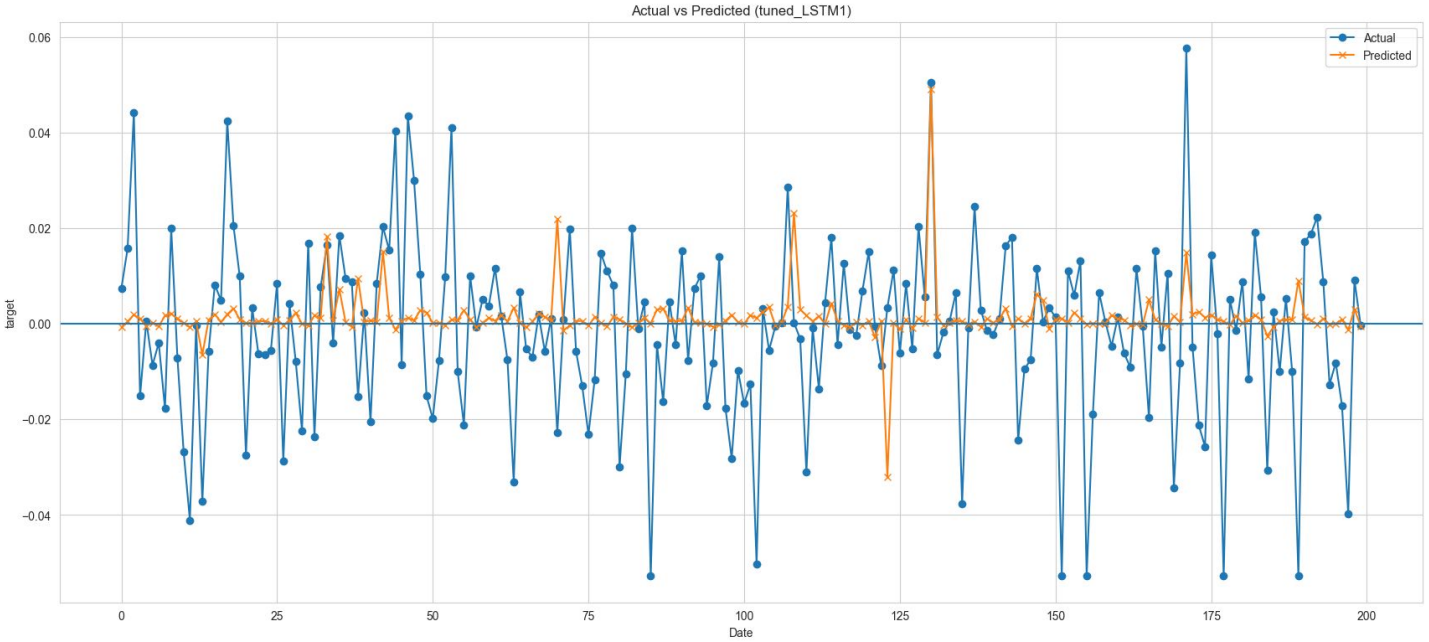
MAE: 0.012417
MSE: 0.000302
RMSE: 0.017374
Direction: 0.535577

XGB



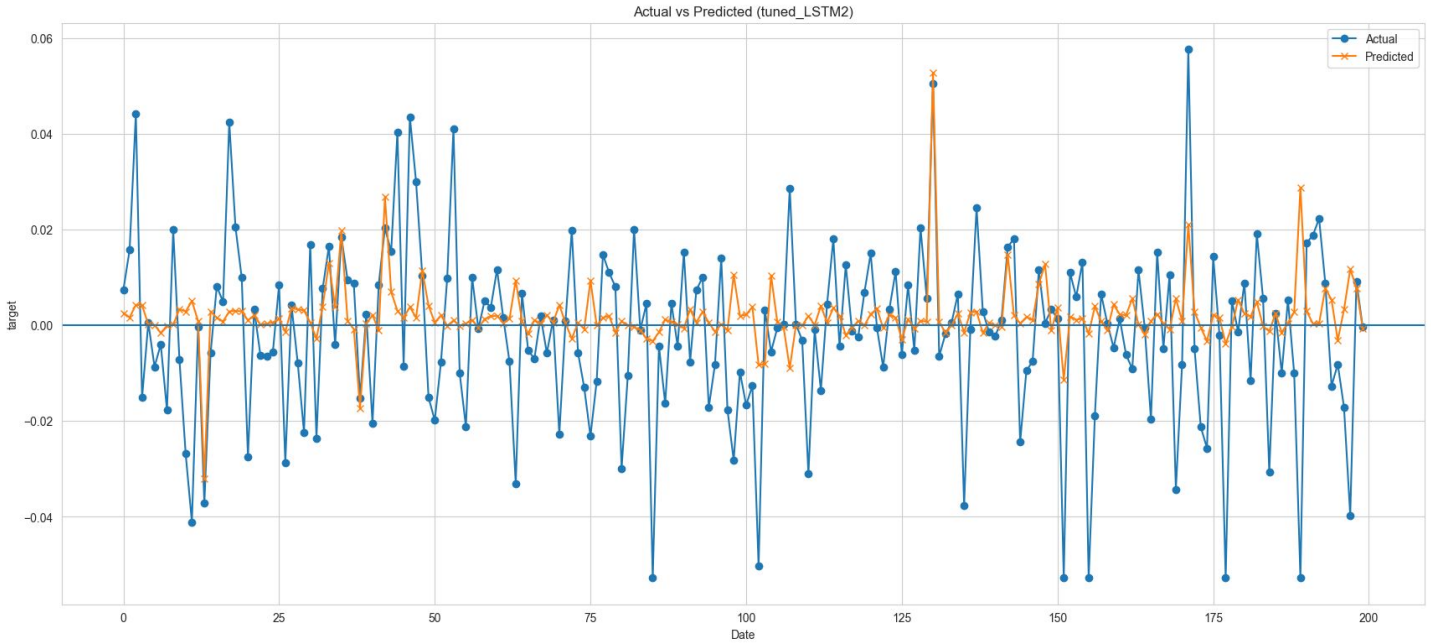
MAE: 0.012368
MSE: 0.000301
RMSE: 0.017336
Direction: 0.535270

LSTM model1



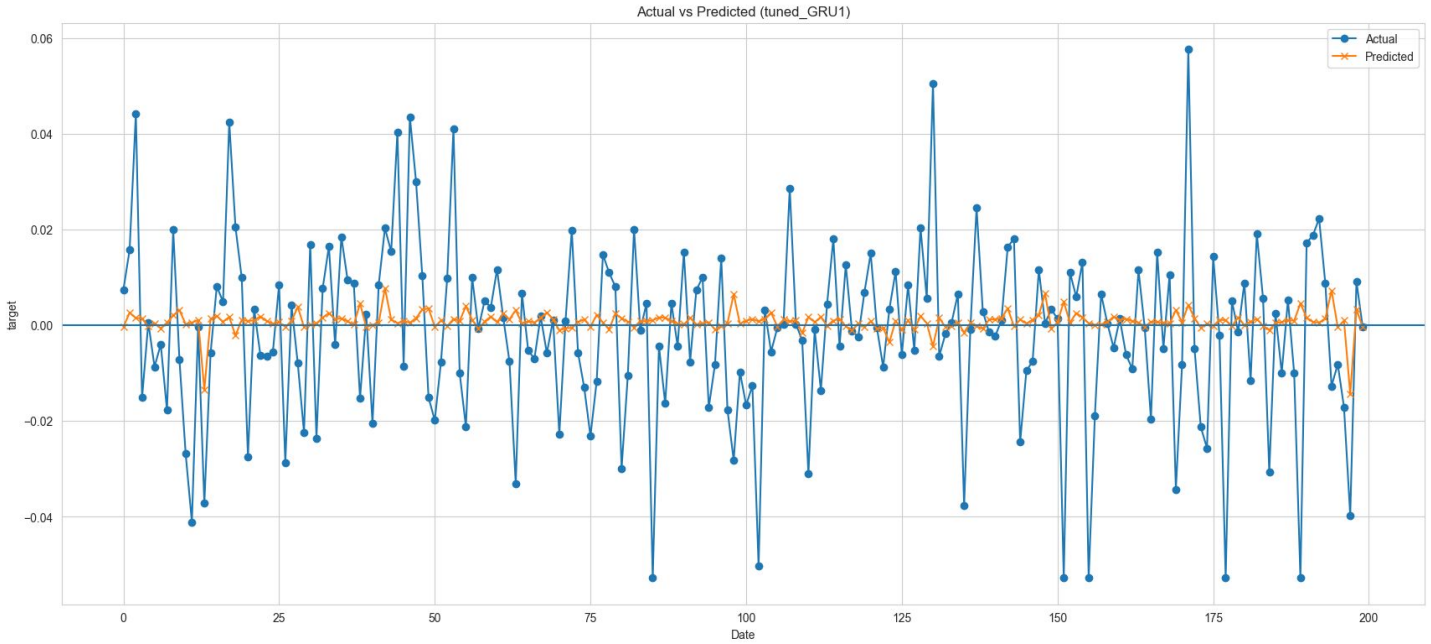
MAE: 0.012215
MSE: 0.000295
RMSE: 0.017176
Direction: 0.531812

LSTM model2



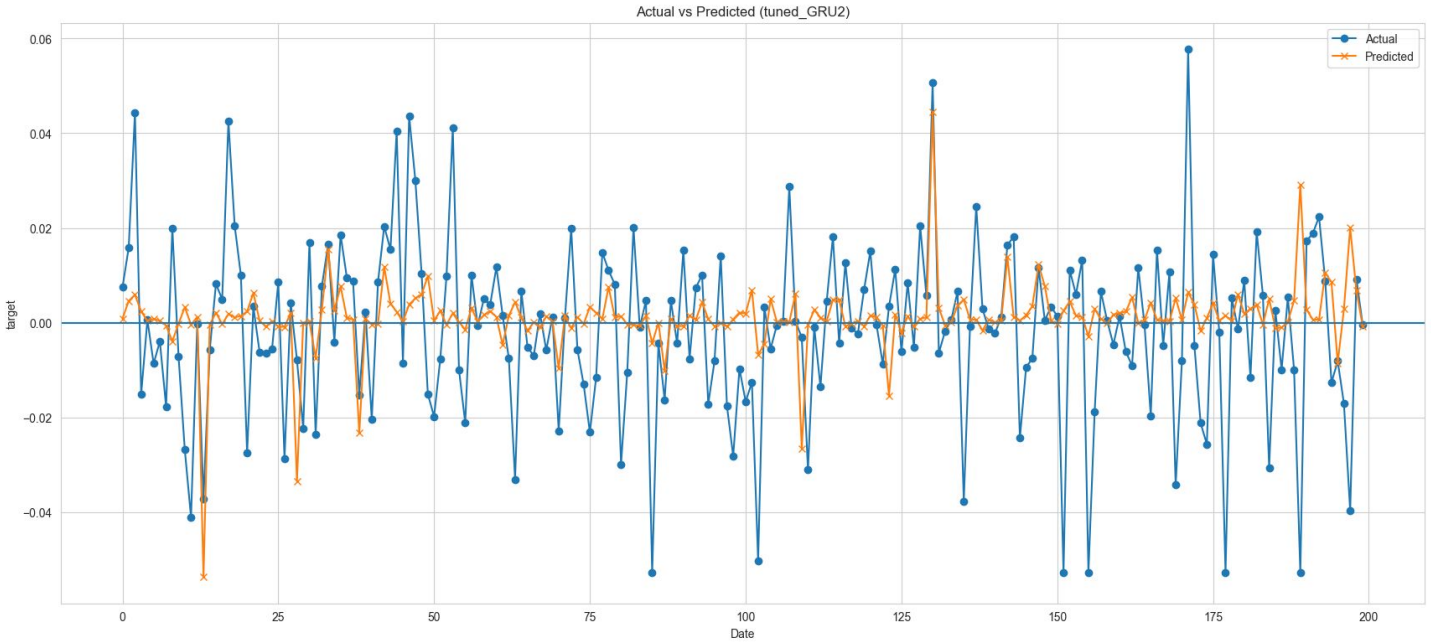
MAE: 0.012227
MSE: 0.000300
RMSE: 0.017310
Direction: 0.545566

GRU model1



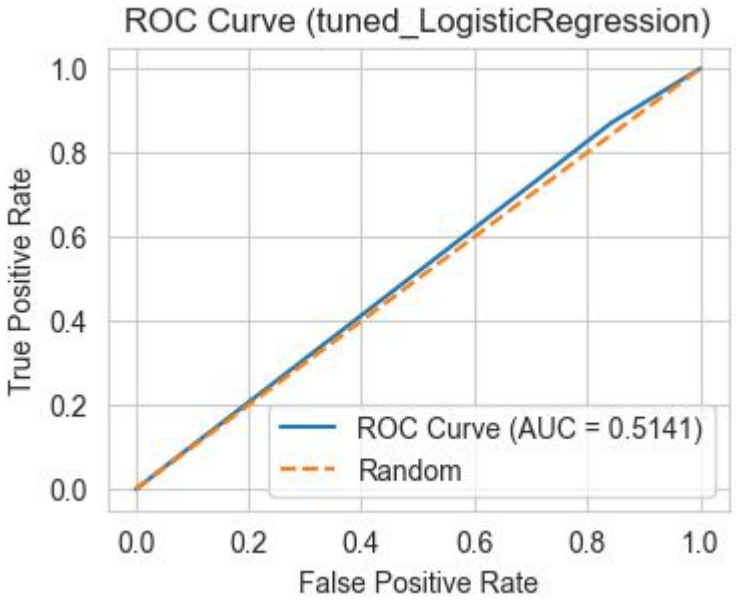
MAE: 0.012285
MSE: 0.000297
RMSE: 0.017232
Direction: 0.536115

GRU model2



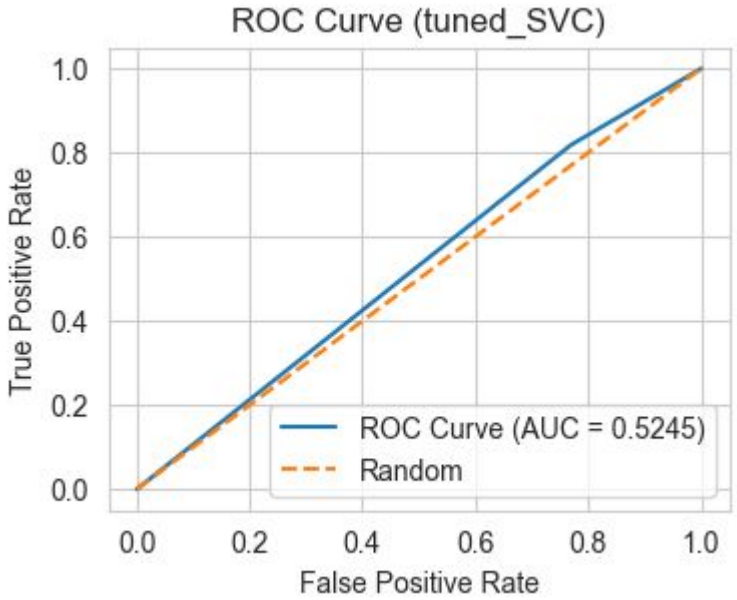
MAE: 0.012125
MSE: 0.000294
RMSE: 0.017142
Direction: 0.553327

Logistic Regression



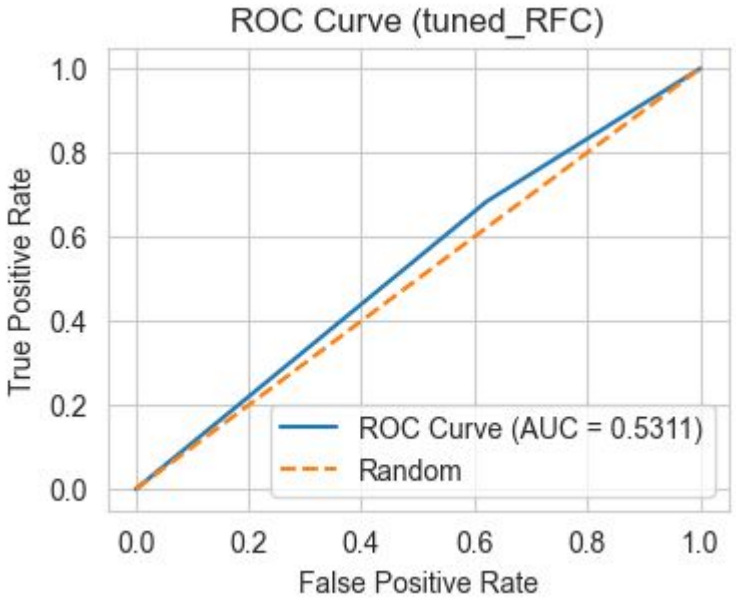
Accuracy: 0.5327

SVC



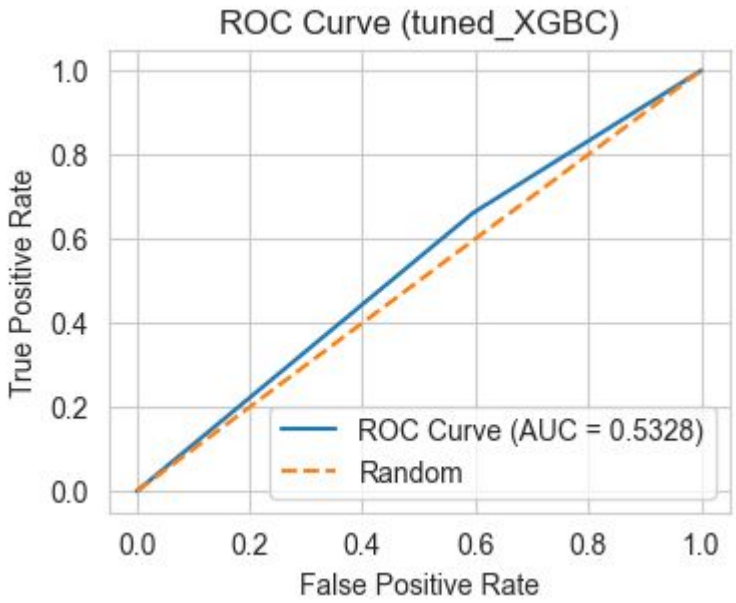
Accuracy: 0.5397

RFC



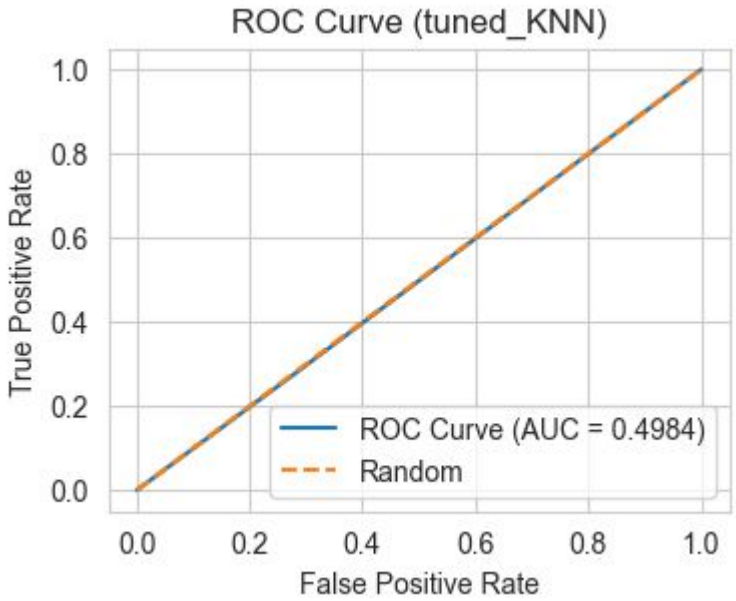
Accuracy: 0.5391

XGBC



Accuracy: 0.5395

KNN



Accuracy: 0.4869

Comparison

	MAE	MSE	RMSE	Direction
Linear Regression	0.012494	0.000307	0.017518	0.52528
SVR	0.012559	0.000313	0.017682	0.538574
RFR	0.012417	0.000302	0.017374	0.535577
XGB	0.012368	0.000301	0.017336	0.53527
LSTM model1	0.012215	0.000295	0.017176	0.531812
LSTM model2	0.012227	0.0003	0.01731	0.545566
GRU model1	0.012285	0.000297	0.017232	0.536115
GRU model2	0.012125	0.000294	0.017142	0.553327

Comparison

Model	Accuracy
Logistic Regression	0.5327
SVC	0.5397
RFC	0.5391
XGBC	0.5395
KNN	0.4869

Backtest

Strategy

Baseline Strategy (Benchmark): Buy and Hold

- Position (Buy / Sell / do nothing) based on the predicted”Return” of the next day
  - return > 0 : buy
  - return = 0 : do nothing
  - return < 0 : sell
- Position applies to the entire portfolio
- No short positions are allowed

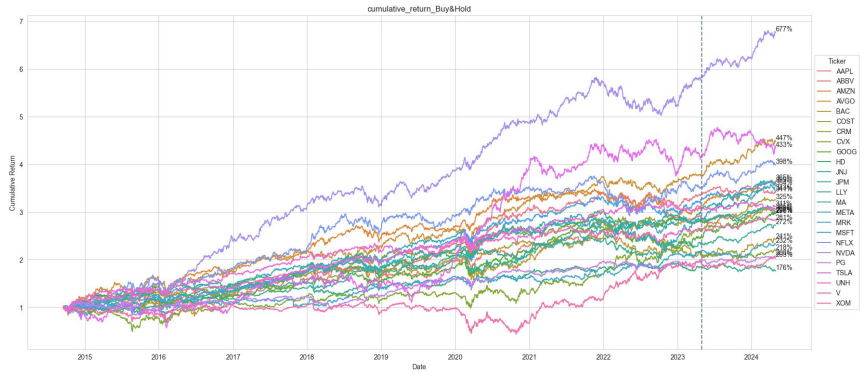
Max Cumulative Return on latest day

NaiveForecast: 'NVDA': 339.07 %
cumulative_return_tuned_LinearRegression 598.279644
cumulative_return_tuned_SVR 1522.292688
cumulative_return_tuned_RFR 1939.127261
cumulative_return_tuned_XGB 1279.966290
cumulative_return_tuned_LSTM1 671.538562
cumulative_return_tuned_LSTM2 677.932993
cumulative_return_tuned_GRU1 705.343564
cumulative_return_tuned_GRU2 759.423074
cumulative_return_tuned_LogisticRegression 645.769544
cumulative_return_tuned_SVC 1052.699327
cumulative_return_tuned_RFC 2007.084402
cumulative_return_tuned_XGBC 1656.404031
cumulative_return_tuned_KNN 1124.579390
cumulative_return_Buy&Hold 676.597023

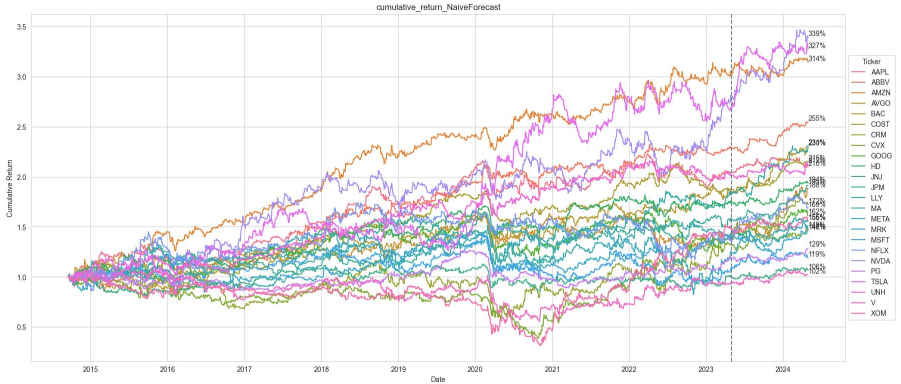
Max Cumulative Return on latest day

MAXreturn	NaiveForecast	nearRegressic	SVR	RFR	XGB	LSTM1	LSTM2	GRU1	GRU2	gisticRegressi	SVC	RFC	XGBC	KNN	Buy&Hold
1310.27	0.15990487	0.277302	0.63293513	0.93370109	0.50106335	0.29720654	0.33335016	0.31358616	0.324492855	0.29371112	0.44502264	1	0.671555067	0.51563671	0.25999548
1148.21	0.221674675	0.27699669	0.47642992	0.93587132	0.44554336	0.25390923	0.28486662	0.25683338	0.277343224	0.24182584	0.36838136	1	0.726417124	0.58682943	0.26136095
1392.07	0.22574042	0.25896544	0.63024093	0.95001616	0.54090568	0.27120838	0.2762811	0.24852547	0.306957055	0.28085095	0.4269743	1	0.775267868	0.54354543	0.25824142
1517.13	0.124003914	0.37223307	0.61415187	0.95052406	0.56918755	0.35474779	0.37148483	0.31466032	0.358392455	0.31609843	0.42756241	1	0.787627048	0.54703262	0.29483495
1295.41	0.177706355	0.18610723	0.5631209	0.98275908	0.48077575	0.26928868	0.39731102	0.2602209	0.394636873	0.1692232	0.37432503	1	0.725461153	0.51100803	0.18619181
1013.03	0.212213285	0.36629611	0.49316304	0.94144523	0.43971464	0.361256	0.33844917	0.31524411	0.354143482	0.36058469	0.37702566	1	0.624406437	0.53419312	0.32078074
1413.92	0.121726361	0.20593886	0.51947323	0.99893911	0.52797286	0.249001	0.28751977	0.30259989	0.344511846	0.24731453	0.39940769	1	0.749711339	0.55444764	0.20901778
1207.84	0.121301312	0.20638127	0.53940887	0.9606562	0.41256028	0.23251283	0.31911146	0.2239717	0.312570654	0.18627859	0.37247407	1	0.682086188	0.52063973	0.1808043
1224.1	0.132040901	0.23054943	0.56409713	0.95385291	0.47696357	0.28164127	0.37590656	0.30314667	0.361091738	0.23937189	0.38628627	1	0.725969724	0.60369672	0.24194101
1069.09	0.181226413	0.26296915	0.52117075	0.94349115	0.43235099	0.34194743	0.34661721	0.32190716	0.374119154	0.26286907	0.41233121	1	0.597266323	0.55634705	0.27804254
789.229	0.134774176	0.22544214	0.36031651	0.96315833	0.34836559	0.27844901	0.35508018	0.24708165	0.357727924	0.24023014	0.30698097	1	0.566639804	0.53222389	0.22274523
1145.69	0.16686737	0.20788544	0.53943445	0.95209225	0.4561679	0.31260547	0.40545838	0.28918505	0.44422184	0.21254127	0.36342159	1	0.671858179	0.53180764	0.23734008
1141.89	0.202166922	0.32948905	0.46662473	0.96017268	0.42002223	0.31028645	0.34549388	0.26256259	0.361045951	0.32638095	0.37444545	1	0.623298741	0.56947538	0.31945535
1212.43	0.128122777	0.30469731	0.56559161	0.98116003	0.48533039	0.33143991	0.41041719	0.36269935	0.463721063	0.28992737	0.43269749	1	0.645053257	0.57176145	0.25126563
1386.58	0.107130727	0.25726752	0.5832301	0.95694764	0.53660864	0.31718802	0.27743675	0.26945711	0.283734646	0.27309481	0.39449914	1	0.694095881	0.55053575	0.24750561
962.193	0.151670182	0.2922515	0.37093031	0.95085415	0.34398664	0.27794383	0.30110597	0.25976524	0.317199064	0.30073354	0.32420072	1	0.627409831	0.55845443	0.24149509
1254.91	0.094831944	0.30107637	0.58287907	0.95657356	0.53698024	0.31275682	0.3847746	0.34496435	0.418478004	0.32280263	0.4216367	1	0.651581128	0.52375551	0.28185908
1637.58	0.103017282	0.25556839	0.63114231	0.93030524	0.61196625	0.28176981	0.26273815	0.28369856	0.25414099	0.25581429	0.37026693	1	0.775267088	0.56277362	0.24315539
2007.08	0.168938463	0.29808395	0.69597764	0.96614136	0.62944379	0.33458412	0.33777005	0.35142696	0.37837127	0.32174509	0.52449181	1	0.772159737	0.52539329	0.33710442
801.177	0.16082566	0.29846296	0.37733939	0.98459981	0.36770169	0.32083646	0.32658641	0.30883283	0.329399843	0.31124169	0.35281478	1	0.544622517	0.61450438	0.25914779
1885.01	0.173690453	0.20219105	0.80757723	0.98666007	0.67902293	0.18273842	0.21672946	0.21417906	0.268494654	0.15843312	0.52331336	1	0.878723386	0.59659008	0.22981215
1110.1	0.191337307	0.29309014	0.43399981	0.96056878	0.46390305	0.28543771	0.33512517	0.30631762	0.385779647	0.31204205	0.37283812	1	0.641430188	0.5174702	0.27981041
1075.94	0.094944524	0.28618543	0.62001457	1	0.49858113	0.34110252	0.44066476	0.39027787	0.450242295	0.30703493	0.42619838	0.98965575	0.634471839	0.53706483	0.26157474
1158.25	0.134783947	0.21233221	0.52693594	0.98573065	0.4155258	0.25730373	0.32297909	0.25276843	0.303850342	0.21711414	0.38409716	1	0.738477347	0.60954974	0.17521789

Cumulative Return (Baseline Model)

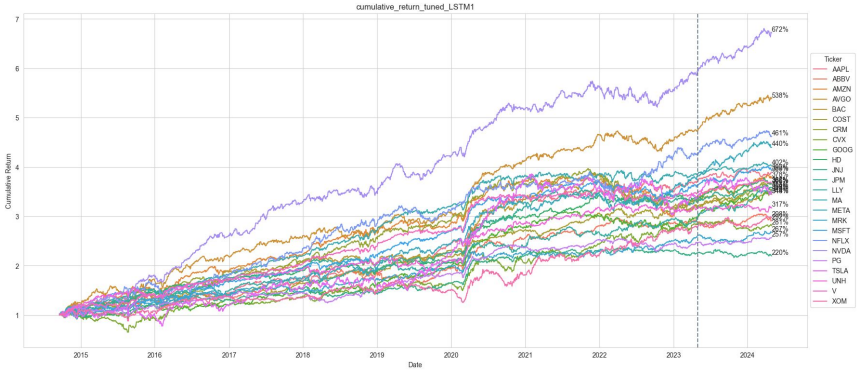


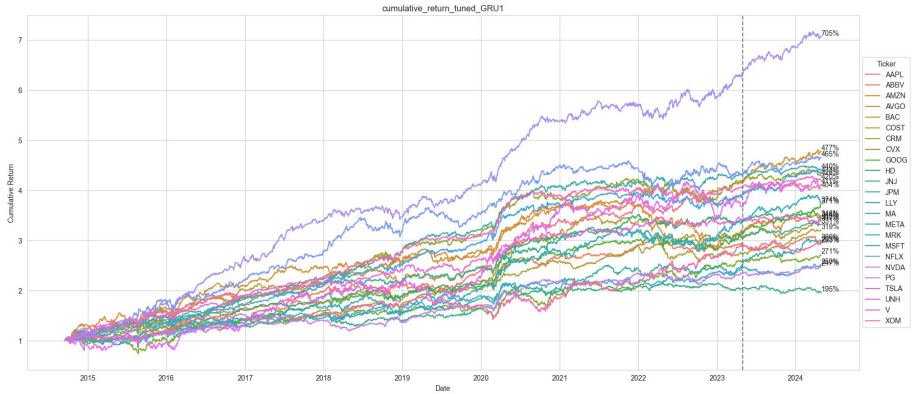
Buy & Hold

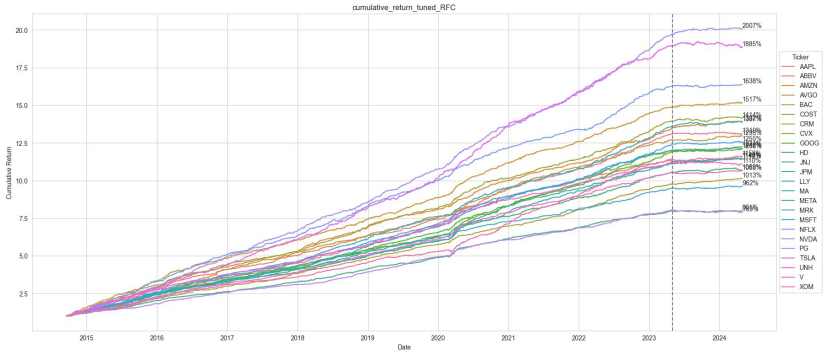


Naive Forecast

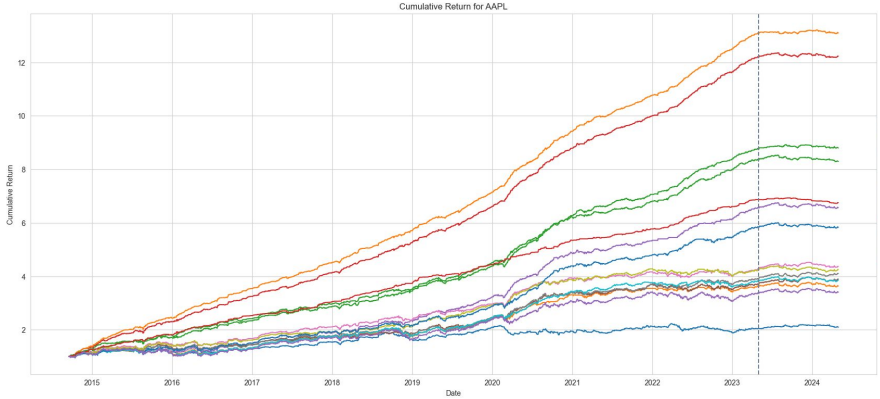
Cumulative Return (Our Models)

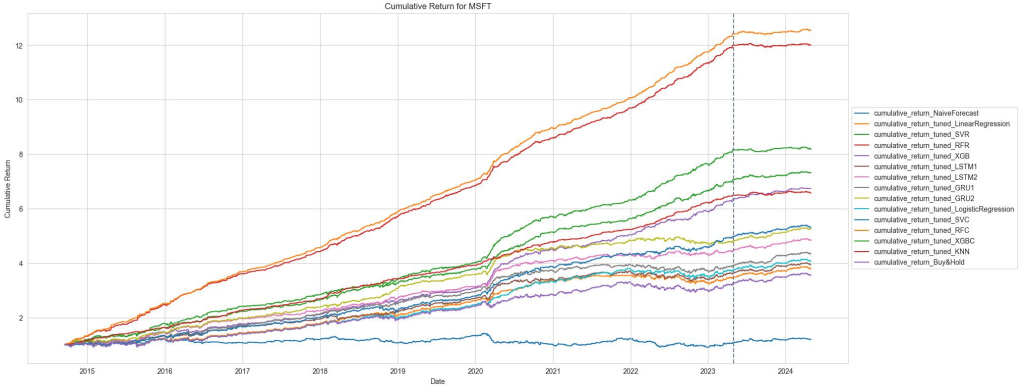




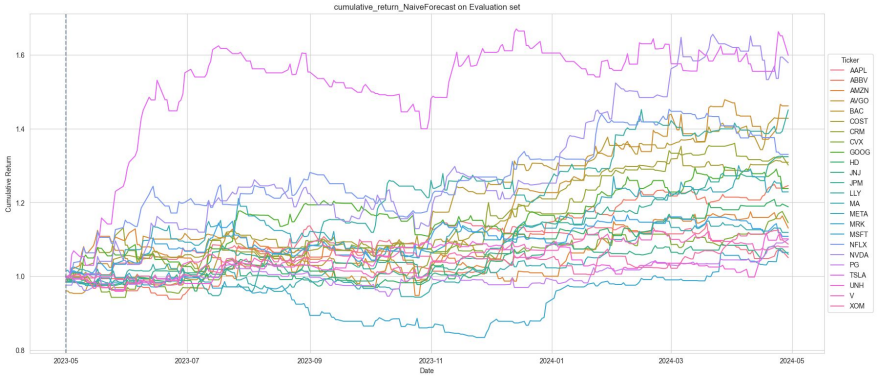


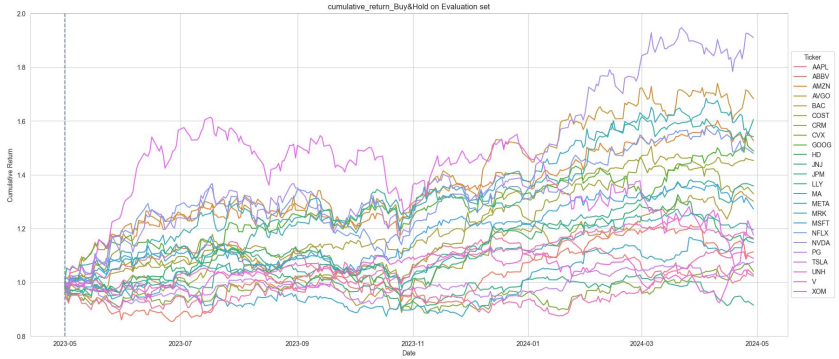
Cumulative Return Entire Data (AAPL & MSFT)



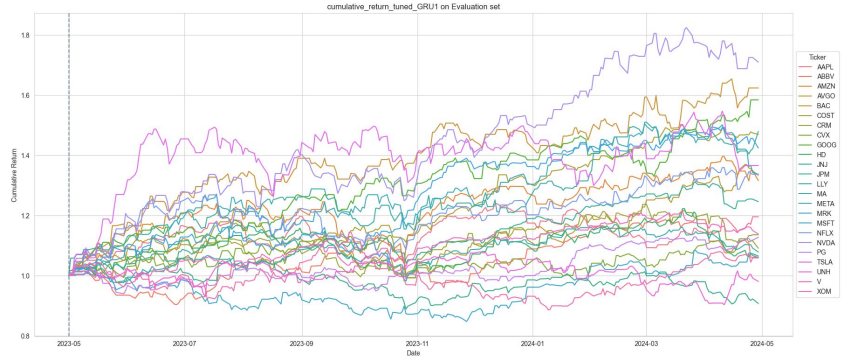


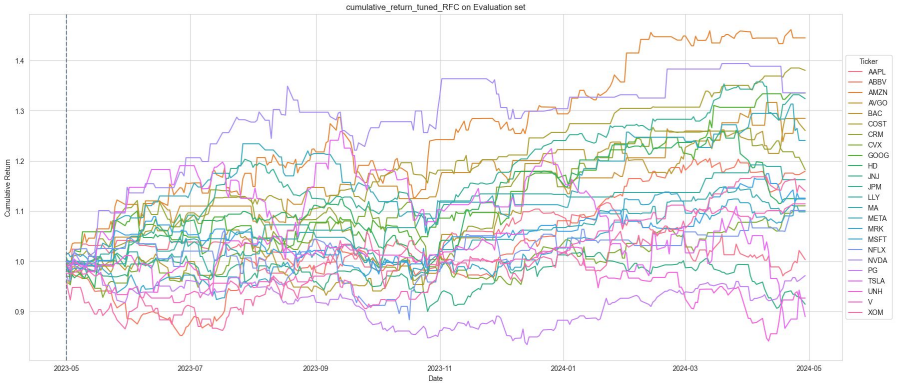
Cumulative Return Evaluation Set (Baseline Models)

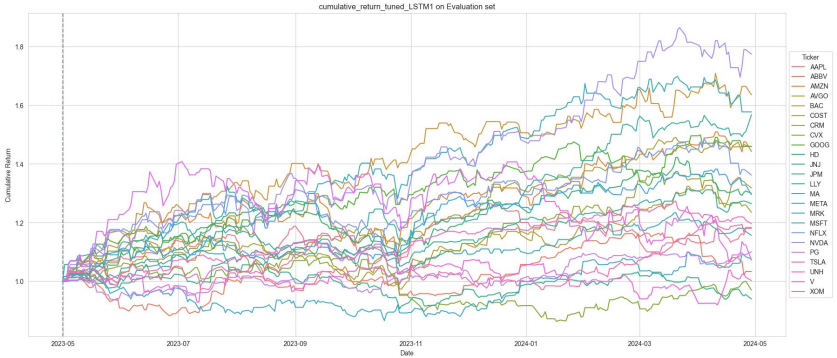




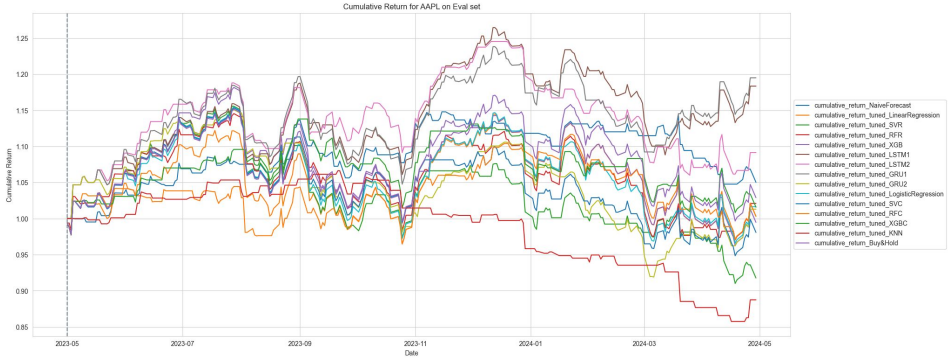
Cumulative Return Evaluation Set (Our Models)

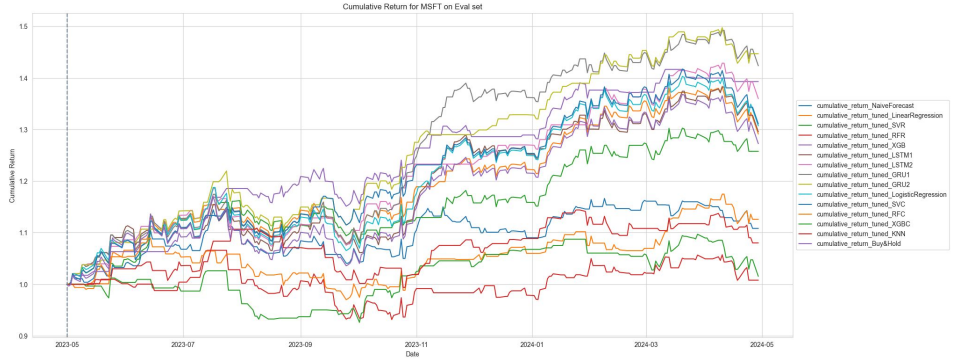




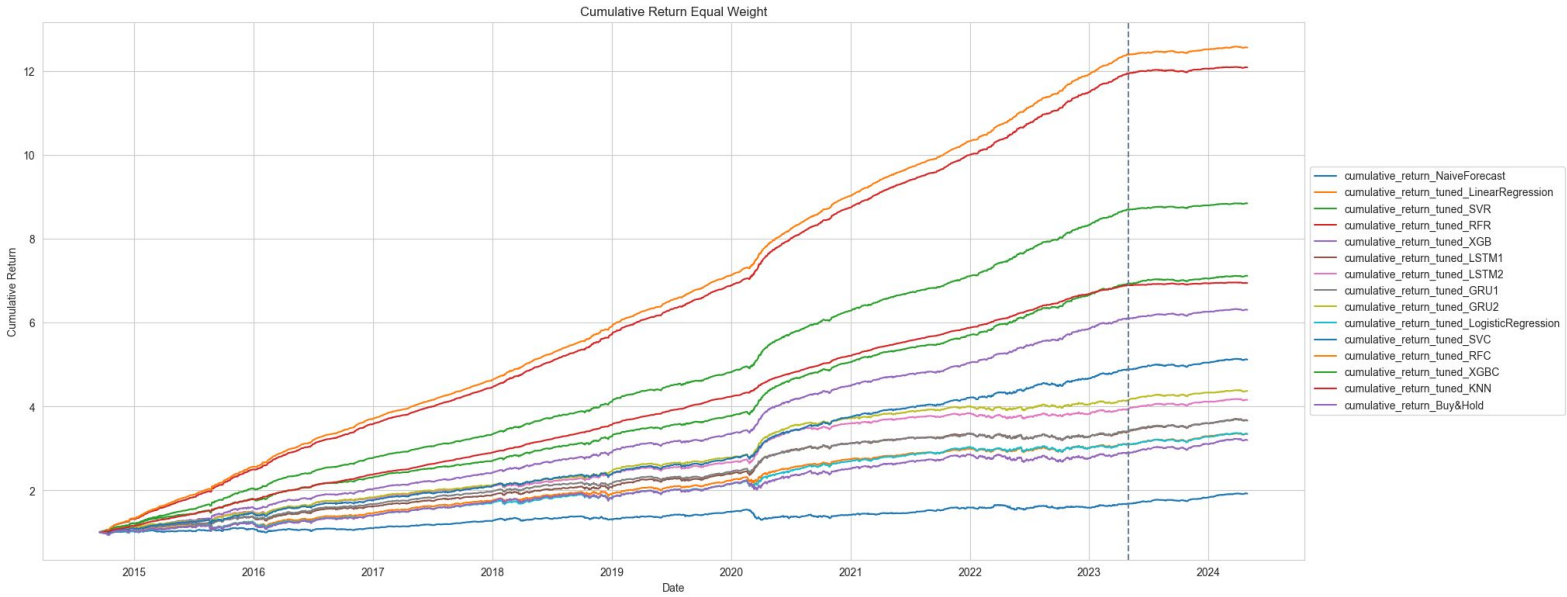


Cumulative Return Evaluation Set (Ticker)

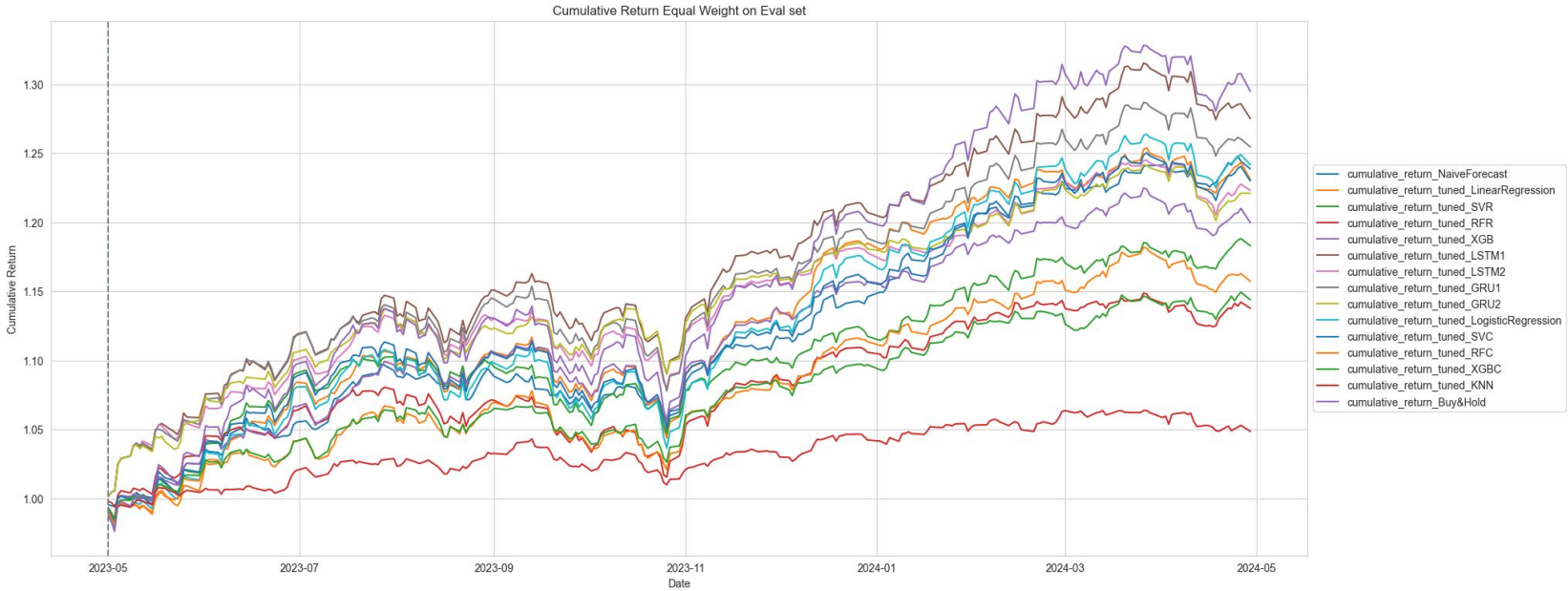




1/N Portfolio Entire Data (equally-weighted portfolio)



1/N Portfolio Evaluate Set (equally-weighted portfolio)



Sharpe Ratio

Evaluate Set

- Usually, any Sharpe ratio greater than 1.0 is considered acceptable to good by investors.
- A ratio higher than 2.0 is rated as very good.
- A ratio of 3.0 or higher is considered excellent.
- A ratio under 1.0 is considered sub-optimal.

NaiveForecast	1.63681	Good
Buy&Hold	1.68712	Good
LinearRegression	1.68135	Good
SVR	2.06761	Very Good
RFR	1.96252	Good
XGB	1.88035	Good
LSTM1	2.17567	Very Good
LSTM2	2.40786	Very Good
GRU1	2.24918	Very Good
GRU2	2.5062	Very Good
LogisticRegression	1.6108	Good
SVC	1.76543	Good
RFC	1.53012	Good
XGBC	1.74951	Good
KNN	1.74017	Good

Error Analysis

Models Comparison

From 1/N Porfolio

Entire set

Win

1. RFC
2. RFR
3. XGBC

Lose

1. Naive
2. BH
3. LinearReg

Evaluate set

Win

1. Buy and Hold
2. LSTM1
3. GRU1

Lose

1. KNN
2. RFR
3. XGBC

Data

In the past: Only Log Return Close

Now: Technical indicators retain historical data for more than a single day, capturingcapture trends and patterns over time

In the past
Net Return (per year)
Buy & Hold: 19.05%
NaiveForecast: 4.47%
LinearRegression: 17.43%
SVR: 0.00%
RFR: 15.17%
LSTM: 30.52%

Now
Net Return (per year)
Buy & Hold : 19.05%
NaiveForecast: 4.47%
LinearRegression: 23.11%
SVR: 18.30%
RFR: 15.77%
LSTM: 42.52%

Feature Engineering

Top Best Feature

- Momentum_21
- Return_63d
- Momentum_5_21
- Return_10d
- Low_Return
- Open_Return
- Return_1d
- Momentum_5_63
- Return_21d
- Return_42d

Worse Feature

- Dollar_Volume
- Month
- NATR
- BB_Mid
- Year
- RSI
- Momentum_5_42
- MACD
- ATR

If momentum is considered the most important factor, using only Open, Low, High, Close, and Volume (OLHCV) data might not perform well.

NN Architecture

- 512 nodes/layer -> 64nodes/layer
- LayerNorm
- Use all high,open,low,close
- With indicator is slightly better than just close,high,open,low

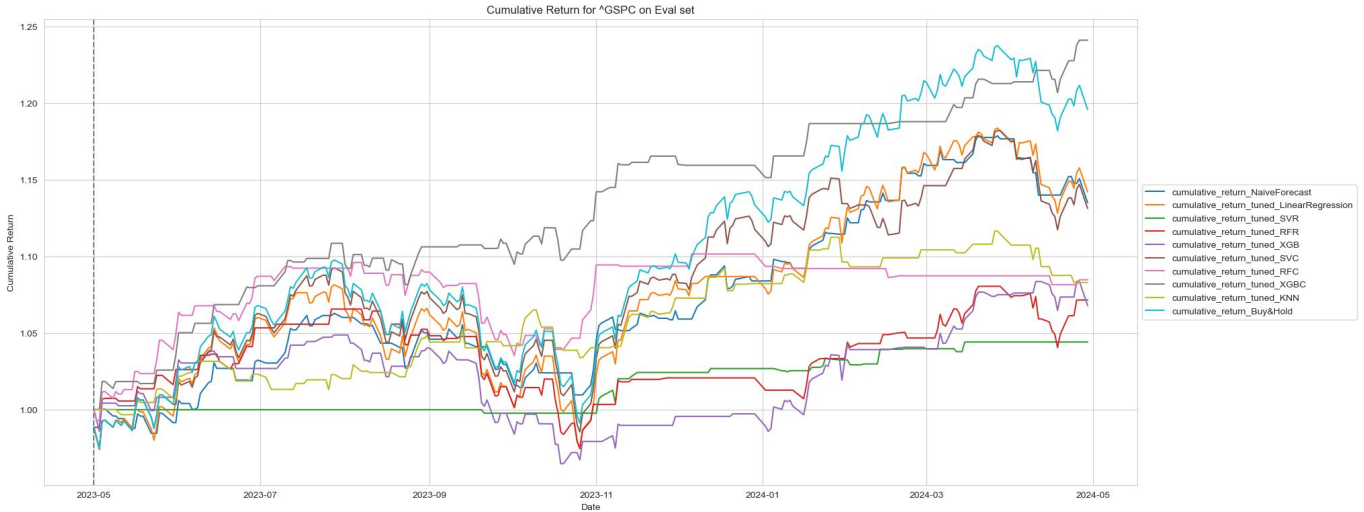
Backtest

- Overfitting: Tree base
- Lose to Benchmark: buy&hold is mostly better
- Sharpe ratio
  - almost every model has sharpe ratio more than benchmark
  - model can prevent risk in same level of return

NaiveForecast	1.63681	Good
Buy&Hold	1.68712	Good
LinearRegression	1.68135	Good
SVR	2.06761	Very Good
RFR	1.96252	Good
XGB	1.88035	Good
LSTM1	2.17567	Very Good
LSTM2	2.40786	Very Good
GRU1	2.24918	Very Good
GRU2	2.5062	Very Good
LogisticRegression	1.6108	Good
SVC	1.76543	Good
RFC	1.53012	Good
XGBC	1.74951	Good
KNN	1.74017	Good

Stock

- Train on the 25 largest market cap companies, mostly in the tech sector.
  However, performance may not be as good when applied across different sectors.



Thank you