Forecasting directional movements of stock prices for intraday trading using LSTM and random forests

Akira Matsue

Nakatsuma Seminar
Keio University

2022/5/9

Thesis Information

- Title: Forecasting directional movements of stock prices for intraday trading using LSTM and random forests
- Author: Pushpendu Ghosh, Ariel Neufeld, Jajati Keshari Sahoo
- Date: May 2022

Abstract

- We employ both random forests and LSTM networks as training methodologies to analyze their effectiveness in forecasting out-of-sample directional movements of constituent stocks of the S&P 500 from 1993 till 2018 for intraday trading.
- On each trading day, we buy the 10 stocks with the highest probability and sell short the 10 stocks with the lowest probability to outperform the market in terms of intraday returns.

Contents

Introduction

Background

In the last decade, machine learning methods have exhibited distinguished development in financial time series prediction.

Krauss et al. (2017)

Deep learning method : Random forest

Single-feature setting

Fischer and Krauss (2018)

Deep learning method : LSTM network

Single-feature setting

Introduction

Benchmark

the results in Krauss et al. (2017) and Fischer and Krauss (2018)

Trading strategy

- the 10 stocks with the highest probability are bought
- the 10 stocks with the lowest probability are sold short

Training methodology

- Random forests
- LSTM networks (CuDNNLSTM)

Data set

All stocks of the S&P 500 from the period of January 1990 until December 2018

Data and technology

Data

Adjusted closing prices and opening prices of all constituent stocks of the S&P 500 from the period of January 1990 until December 2018 collected from Bloomberg
(For each day, stocks with zero volume were not considered for trading at this day.)

Technology

Codes and simulations : Python 3.6.5,
TensorFlow 1.14.0, scikit-learn 0.20.4

Visualization and statistical values : MATLAB R2016b

Methodology

Dataset creation with non-overlapping testing period

1. Dataset creation with non-overlapping testing period
2. Features selection
3. Target selection
4. Model training specification
5. Prediction and trading methodology

Methodology

Dataset creation with non-overlapping testing period

Dividing the dataset of 29 years using a 4-year window and 1-year stride

Each study period is divided into
a training period of approximately 756 days($\fallingdotseq 3 years$)*and*
a trading period of approximately 252 days($\fallingdotseq 1 year$)
$\rightarrow$ 26 study periods with non-overlapping trading part

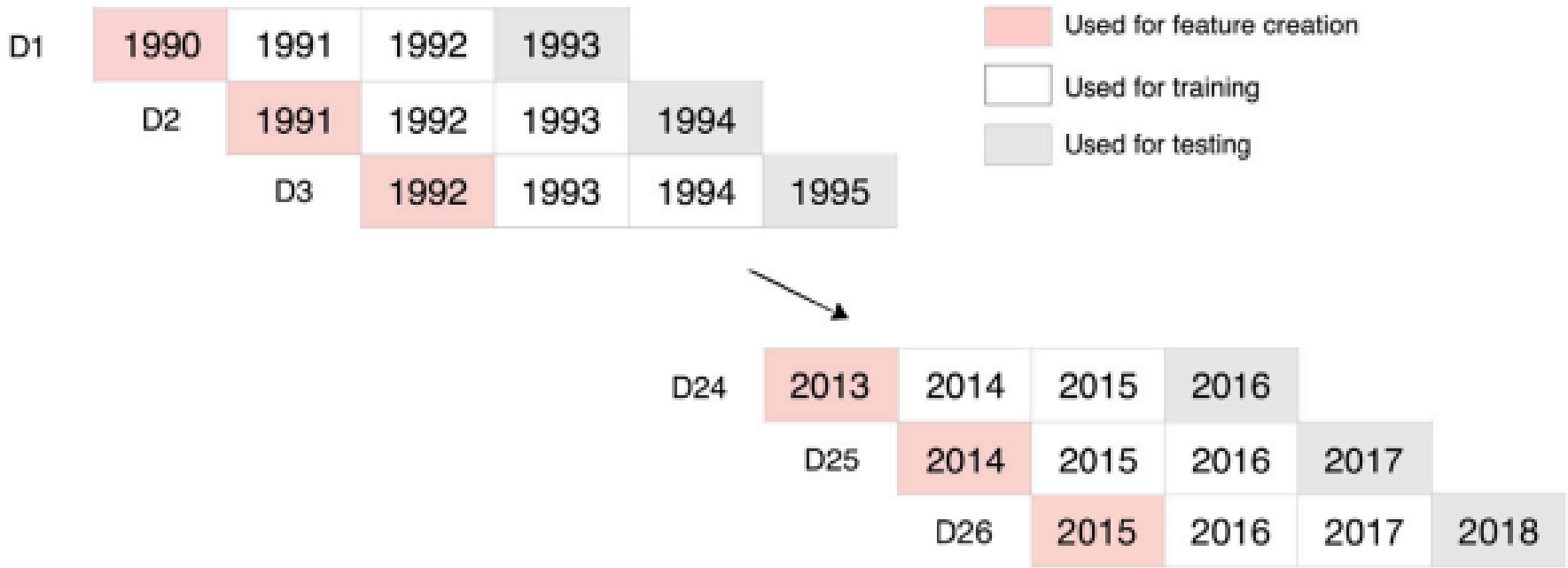


Fig. 1. Dataset creation with non-overlapping testing period.

Methodology

Features selection

T_{study} : the amount of days in a study period

i : each study period

n_i : the number of stocks s in S having complete historical data available at the end of i

$cp_t^{(s)}$: closing price of any stock $s \in S$ at time t

$op_t^{(s)}$: opening price of any stock $s \in S$ at time t

Methodology

Features selection

Input :

$cp_t^{(s)},\ t \in \{0,\ 1,\ ...,\ \tau - 1, \tau\}$

$op_t^{(s)},\ t \in \{0,\ 1,\ ...,\ \tau - 1\}$

Task : Out of all n stocks, predict stocks with the highest and stocks with the lowest intraday return $ir_{\tau,0} := \frac{cp_\tau}{op_\tau} - 1$

Methodology

Feature generation for random forest

1. Intraday returns : $ir_{t,m}^{(s)} := \frac{cp_{t-m}^{(s)}}{op_{t-m}^{(s)}} - 1$
2. Returns with respect to last closing price: $cr_{t,m}^{(s)} := \frac{cp_{t-1}^{(s)}}{cp_{t-1-m}^{(s)}} - 1$
3. Returns with respect to opening price: $or_{t,m}^{(s)} := \frac{op_{t}^{(s)}}{cp_{t-m}^{(s)}} - 1$

where $m \in \{1, 2, 3, \ldots, 20\} \cup \{40, 60, 80, \ldots, 240\}$

$\rightarrow$ 93 features

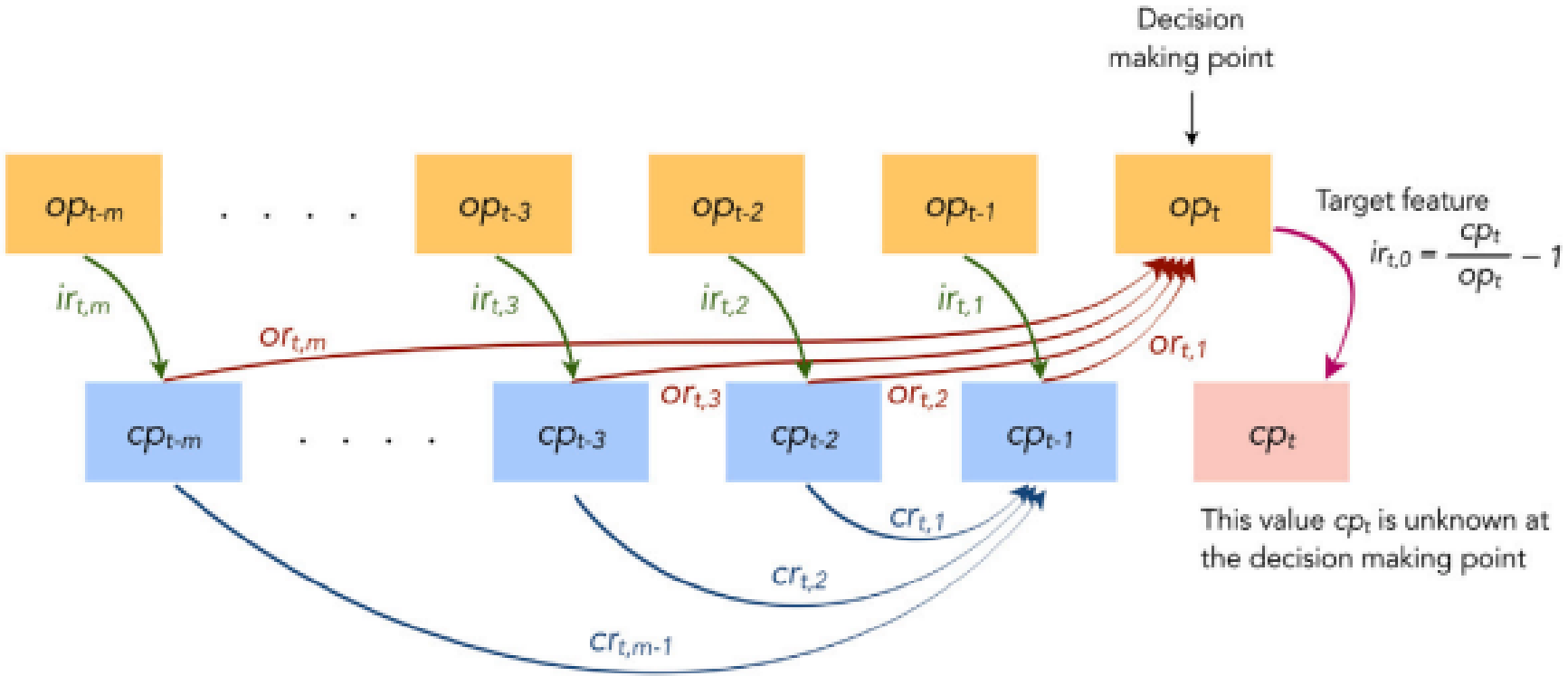


Fig. 2. Feature generation for random forest.

Methodology

Feature generation for LSTM

1. Inputting the model with 240 time steps and 3 features, and training it to predict the direction of the 241st intraday return
2. Robust Scaler standardization : Removing the median and scaling the data using the inter-quantile range
   $\rightarrow$ making it robust to outliers
3. $t \in \{241, 242, ..., T_{study}\}$
   $\tilde{F}^{(s)}_{t-i,1} := (\tilde{ir}^{(s)}_{t-i,1}, \tilde{cr}^{(s)}_{t-i,1}, \tilde{or}^{(s)}_{t-i,1})$, $i \in \{239, 238, ..., 0\}$
   $\rightarrow$ 240 consecutive, three-dimensional standardized features
   $\{\tilde{F}^{(s)}_{t-239,1}, \tilde{F}^{(s)}_{t-238,1}, \cdots, \tilde{F}^{(s)}_{t,1}\}$

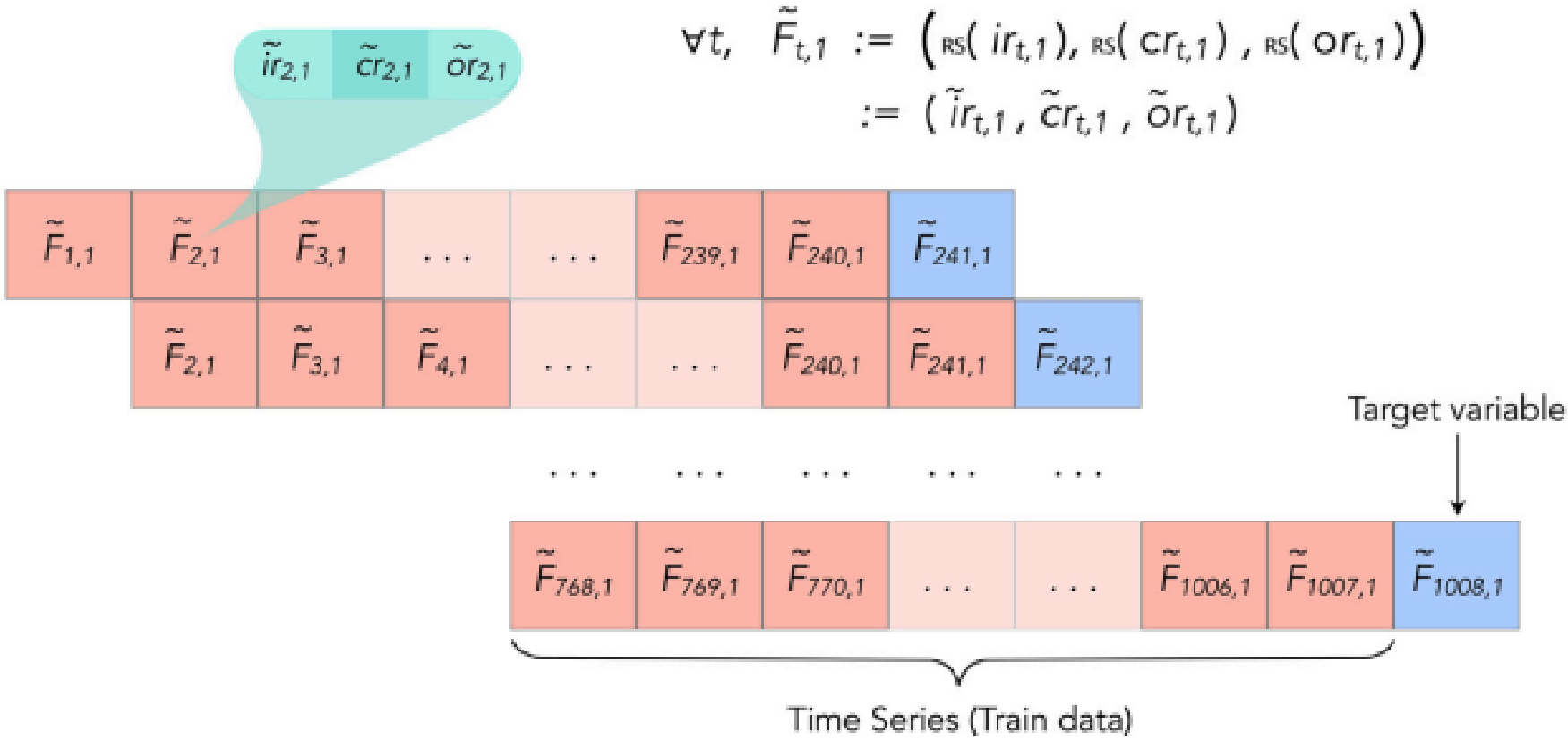


Fig. 3. Feature generation for LSTM.

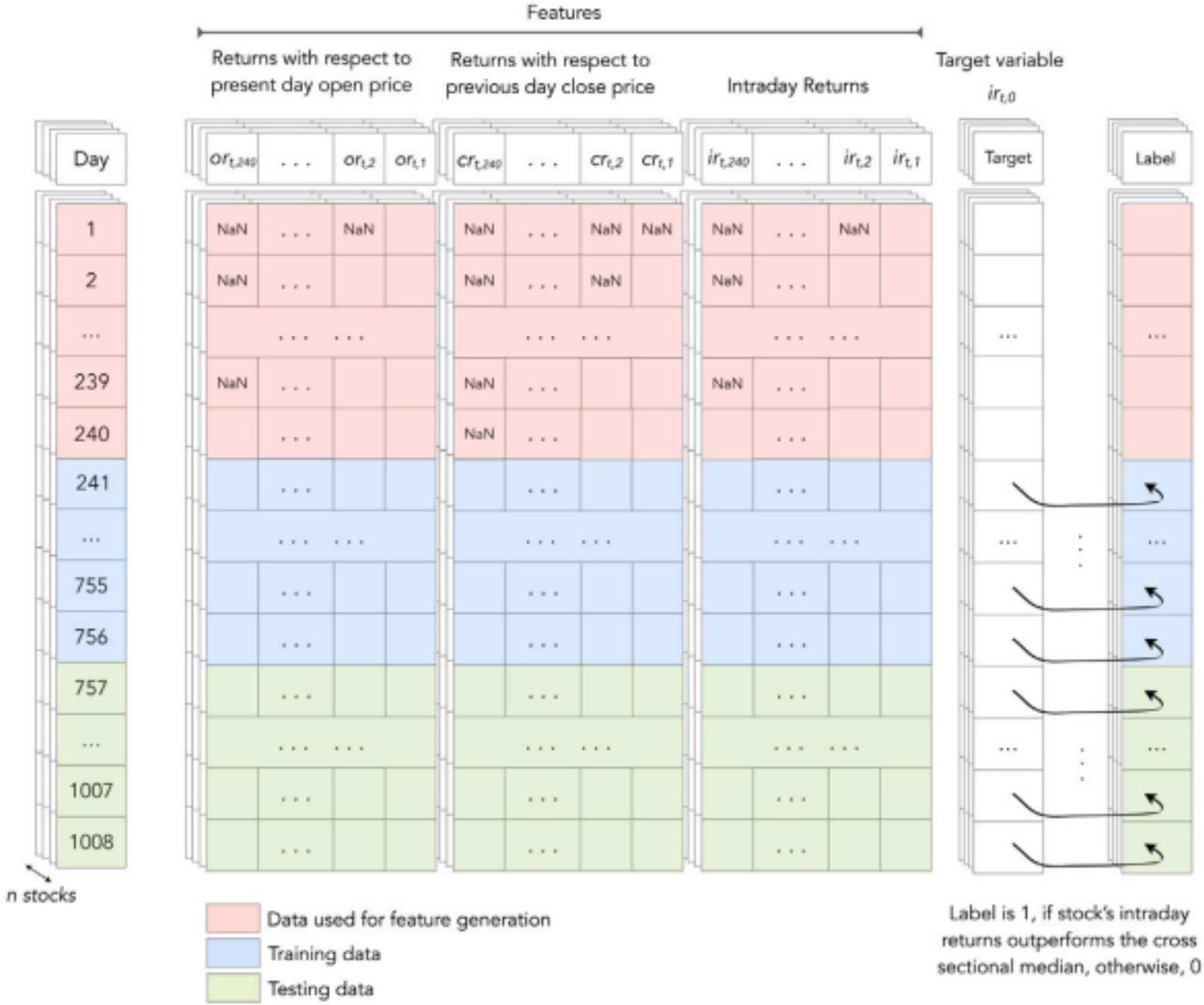
Methodology

Train-test split

Not defined : $t \leqq 240$

Training : from $t = 241$ to $t = 75$

Testing : from $t = 757$ to $t = T_{study}$


Features
Returns with respect to present day open price
Returns with respect to previous day close price
Intraday Returns
Target variable
$ir_{t,0}$
Day
$or_{t,240}$ . . . $or_{t,2}$ $or_{t,1}$
$cr_{t,240}$ . . . $cr_{t,2}$ $cr_{t,1}$
$ir_{t,240}$ . . . $ir_{t,2}$ $ir_{t,1}$
Target
Label
1
2
239
240
241
755
756
757
1007
1008
NaN
n stocks
Data used for feature generation
Training data
Testing data
Label is 1, if stock's intraday returns outperforms the cross sectional median, otherwise, 0

Methodology

Target selection

Dividing each stock at time t into 2 classes of equal size

- Class 0 : $ir_{t,0}^{(s)}$ of stock s is smaller than the cross-sectional median intraday return of all stocks at time t
- Class 1 : $ir_{t,0}^{(s)}$ of stock s is bigger than the cross-sectional median intraday return of all stocks at time t

Methodology

Model training specification

Model specification for random forest

- Number of decision trees in the forest = 1000
- Maximum depth of each tree = 10

Model specification for LSTM

- Loss function : categorical cross-entropy
- Optimizer : RMSProp (with the keras default learning rate of 0.001)
- Batch size : 512
- Early stopping : patience of 10 epochs, monitoring the validation loss
- Validation split : 0.2

Table 2
Average performance metrics of daily returns before transaction cost.

Metric of daily returns	3-Feature IntraDay LSTM	3-Feature IntraDay RF	1-Feature NextDay LSTM	1-Feature NextDay RF	1-Feature IntraDay LSTM	1-Feature IntraDay RF	SP500 Index
Mean (long)	0.00332	0.00273	0.00257	0.00259	0.00094	0.00104	0.00033
Mean (short)	0.00312	0.00266	0.00158	0.00130	0.00180	0.00187	0.00000
Mean return	0.00644	0.00539	0.00414	0.00389	0.00274	0.00290	0.00033
Standard error	0.00019	0.00020	0.00024	0.00023	0.00021	0.00021	0.00014
Minimum	−0.1464	−0.1046	−0.1713	−0.1342	−0.1565	−0.1487	−0.0903
Quartile 1	−0.0017	−0.0028	−0.0052	−0.0051	−0.0054	−0.0050	−0.0044
Median	0.00559	0.00462	0.00352	0.00287	0.00242	0.00221	0.00056
Quartile 3	0.01433	0.01306	0.01294	0.01161	0.01086	0.01036	0.00560
Maximum	0.14101	0.14153	0.19884	0.28139	0.13896	0.16064	0.11580
Share > 0	0.69663	0.65857	0.60598	0.59479	0.58405	0.58937	0.53681
Std. deviation	0.01572	0.01597	0.01961	0.01831	0.01713	0.01683	0.01133
Skewness	0.15599	0.28900	0.36822	1.41199	−0.1828	0.12051	−0.1007
Kurtosis	9.71987	8.32627	10.8793	19.8349	10.1893	11.7758	11.9396
1-percent VaR	−0.0352	−0.0364	−0.0492	−0.0432	−0.0461	−0.0448	−0.0313
1-percent CVaR	−0.0519	−0.0528	−0.0712	−0.0592	−0.0678	−0.0660	−0.0451
5-percent VaR	−0.0157	−0.0170	−0.0234	−0.0208	−0.0214	−0.0197	−0.0177
5-percent CVaR	−0.0284	−0.0297	−0.0401	−0.0345	−0.0377	−0.0357	−0.0270
Max. drawdown	0.22345	0.19779	0.42551	0.23155	0.35645	0.43885	0.56775
Avg return p.a.	3.84750	2.75103	1.68883	1.53806	0.91483	1.00281	0.06975
Std dev. p.a.	0.24957	0.25358	0.31135	0.29071	0.27193	0.26719	0.17990
Down dev. p.a.	0.17144	0.17301	0.21204	0.18690	0.19270	0.18530	0.12970
Sharpe ratio	6.34253	5.20303	3.22732	3.23339	2.39560	2.59188	0.24867
Sortino ratio	62.7403	49.6764	27.8835	30.0753	19.0217	21.2964	1.77234

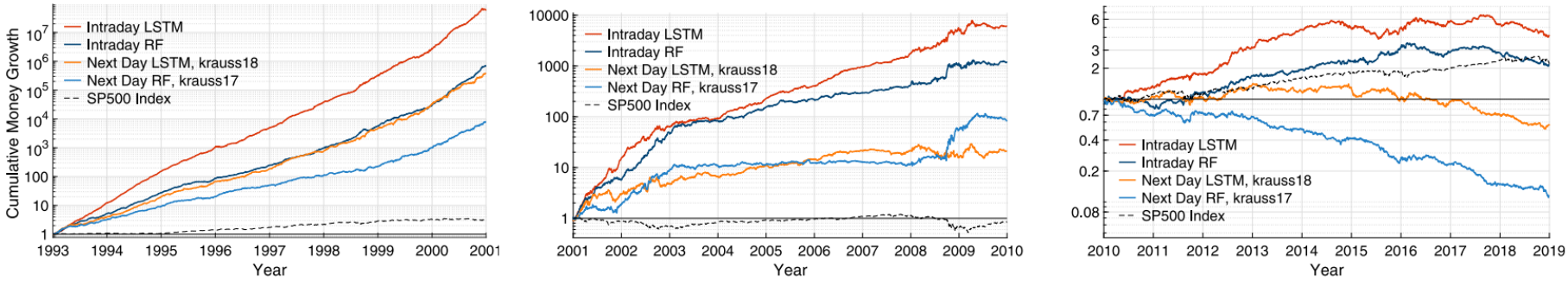


Fig. 5. Cumulative money growth with US$1 initial investment, after deducting transaction cost.

Results and discussion

Results

1. The multi-feature setting outperforms the single feature setting of Krauss(2017, 2018)
2. LSTM outperforms random forests
   $\rightarrow$ LSTM has an advantage compared to the memory-free methods

Results and discussion

The first time-period

Great performance
$\rightarrow$ the dot-com-bubble

The second time-period

moderation with the bursting of the dot-com bubble and the financial crisis of 2008

The last time-period

Deterioration of performance
$\rightarrow$ machine learning algorithms are broadly available
$\rightarrow$ decreasing the opportunity of creating statistical arbitrage having a technological advantage

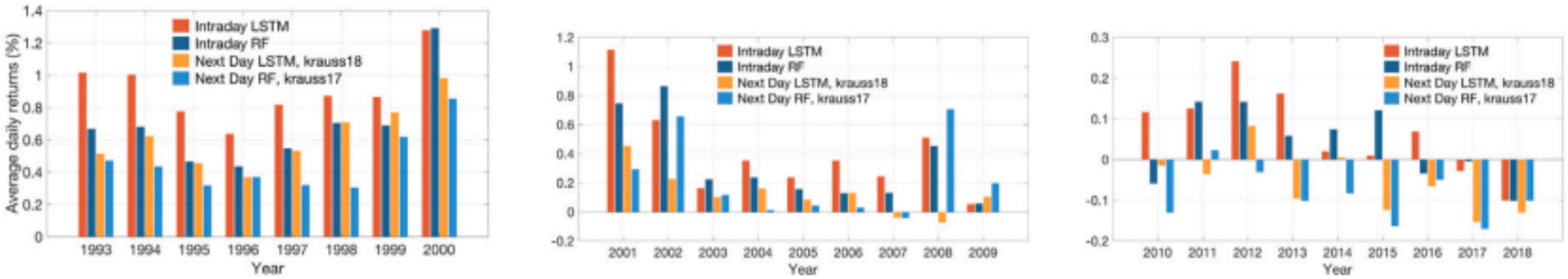


Fig. 6. Average of daily mean returns, after deducting transaction cost.

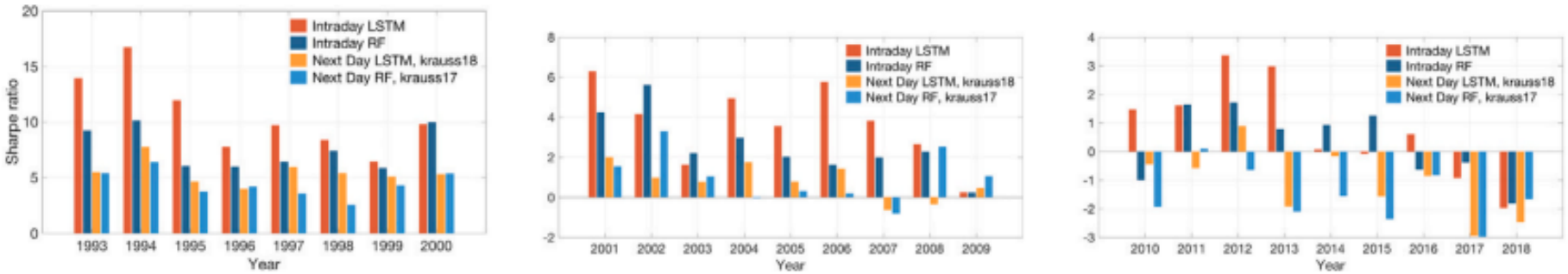


Fig. 7. Annualized sharpe ratio, after deducting transaction cost.