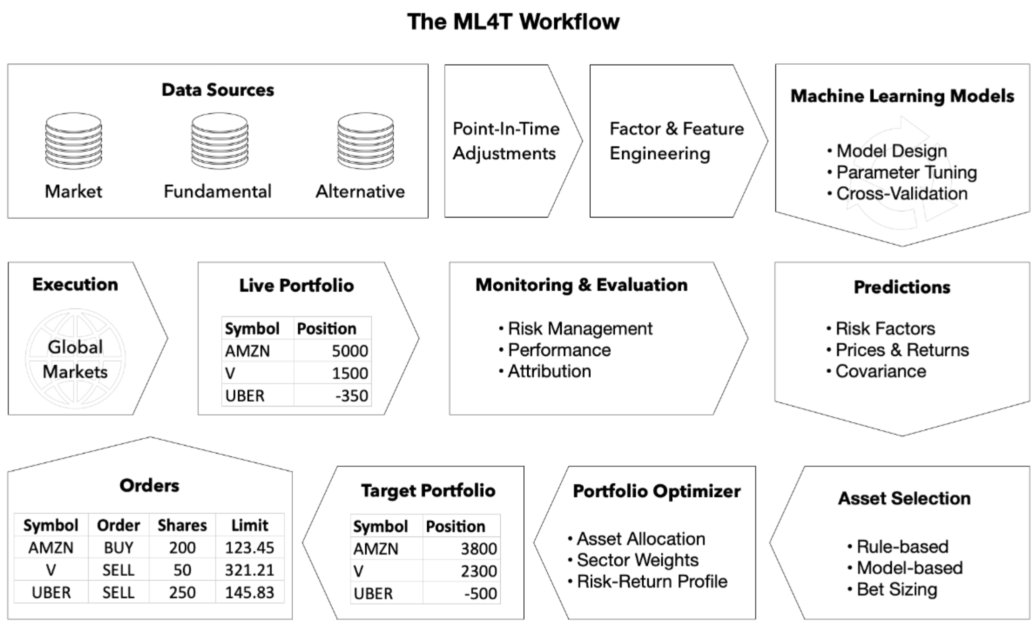
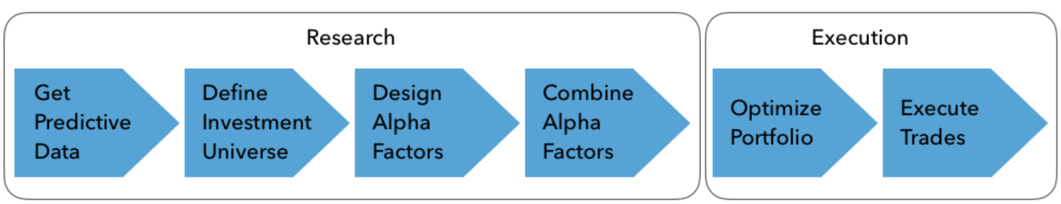
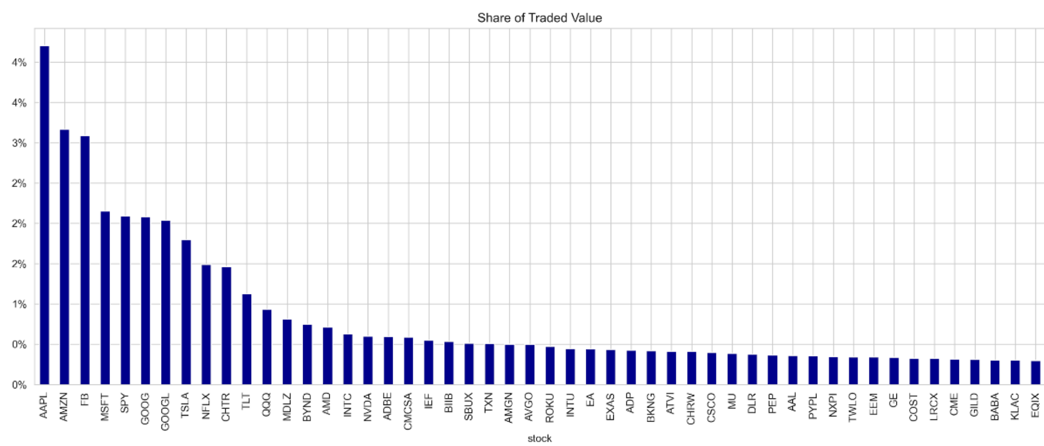
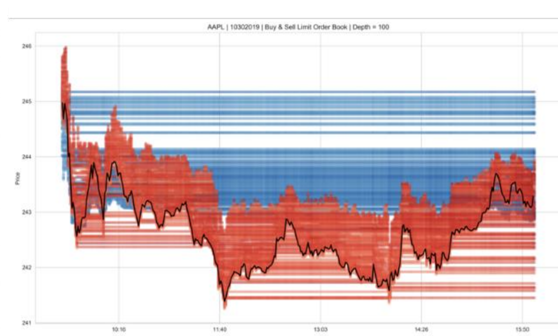
Chapter 1: Machine Learning for Trading – From Idea to Execution





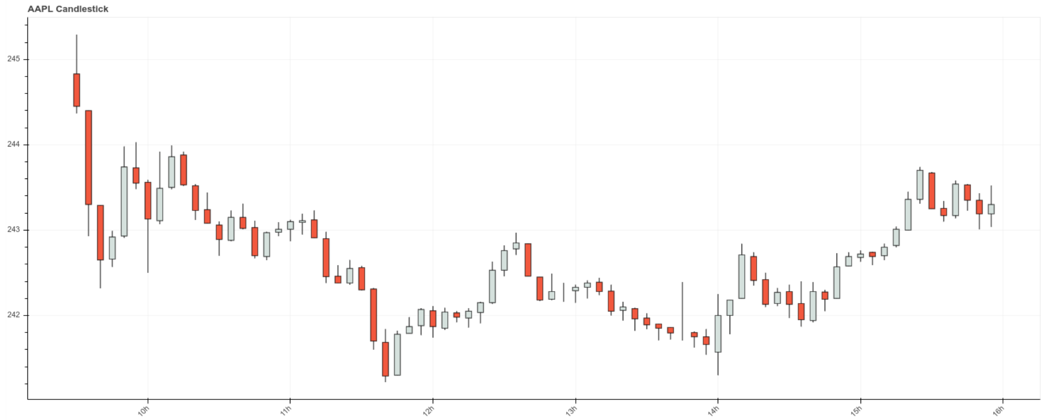
Chapter 2: Market and Fundamental Data – Sources and Techniques



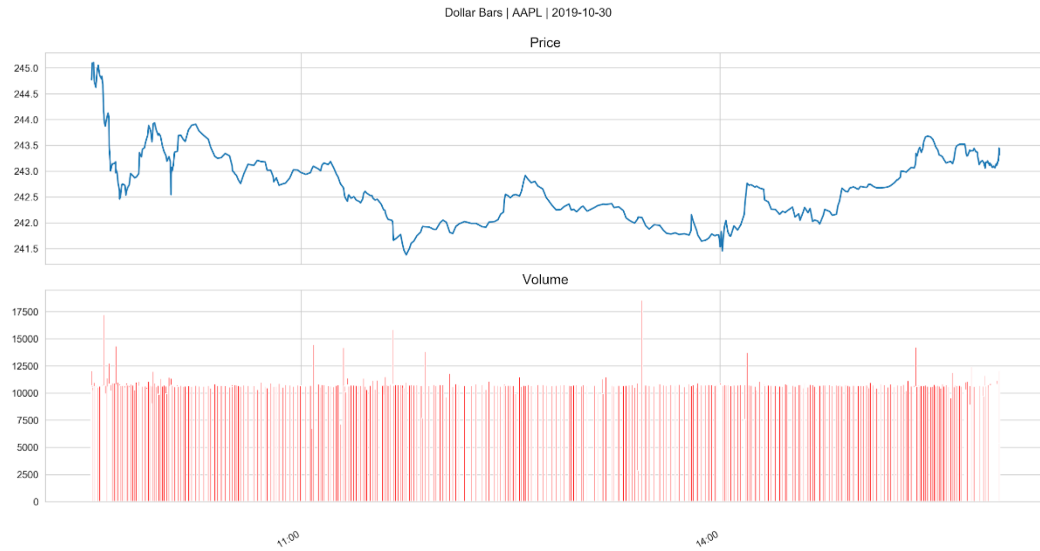



Tick Bars | AAPL | 2019-10-30
245.0
244.5
244.0
243.5
243.0
242.5
242.0
241.5
09:43:20
11:06:40
12:30
13:53:20
15:16:40


Time Bars | AAPL | 2019-10-30
Price
Volume
120000
100000
80000
60000
40000
20000
0
11:00
14:00

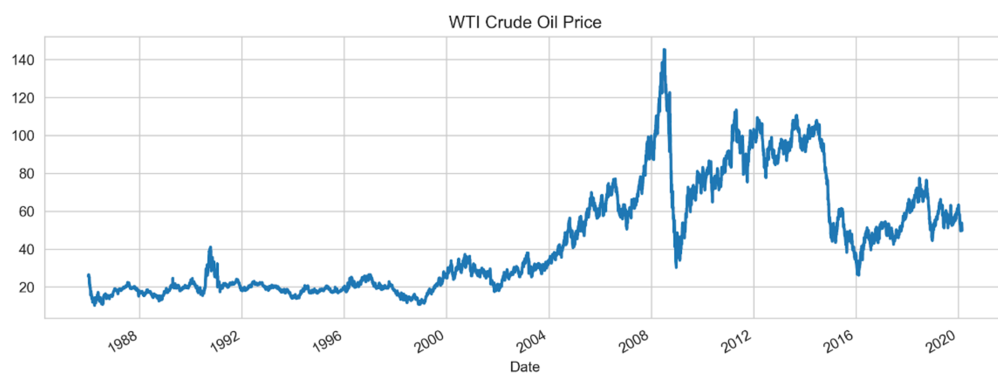

AAPL Candlestick
245
244
243
242
10h
11h
12h
13h
14h
15h
16h


Volume Bars | AAPL | 2019-10-30
Price
245.0
244.5
244.0
243.5
243.0
242.5
242.0
241.5
Volume
20000
17500
15000
12500
10000
7500
5000
2500
0
11:00
14:00

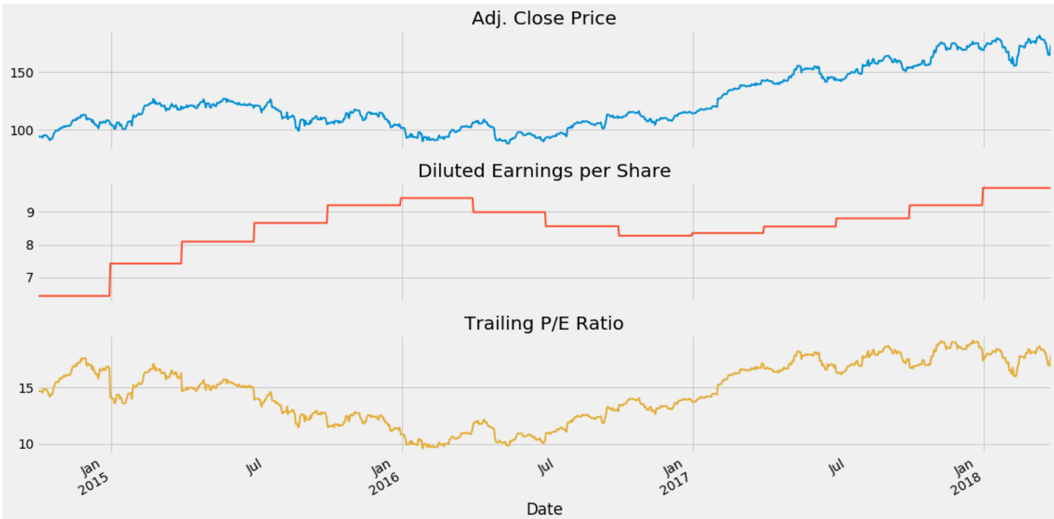

Dollar Bars | AAPL | 2019-10-30
Price
245.0
244.5
244.0
243.5
243.0
242.5
242.0
241.5
Volume
17500
15000
12500
10000
7500
5000
2500
0
11:00
14:00


171
170.5
170
169.5
169
06:00
Dec 29, 2017
09:00
12:00
15:00
18:00

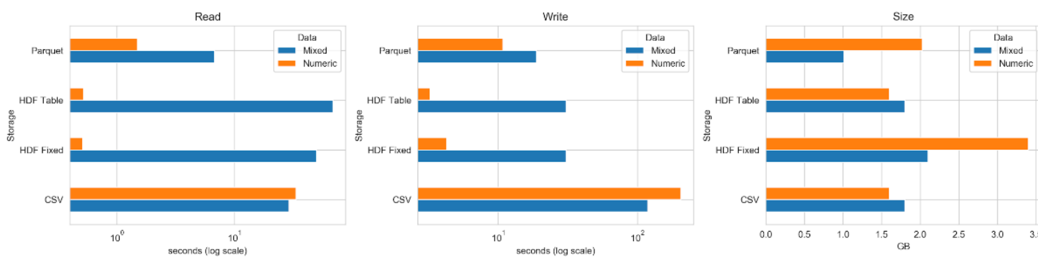
Equity(1037 [FB])
Equity(1244 [GOOG])
Equity(167 [AMZN])
1200
1000
800
600
400
200
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
date



WTI Crude Oil Price
140
120
100
80
60
40
20
1988
1992
1996
2000
2004
2008
2012
2016
2020
Date



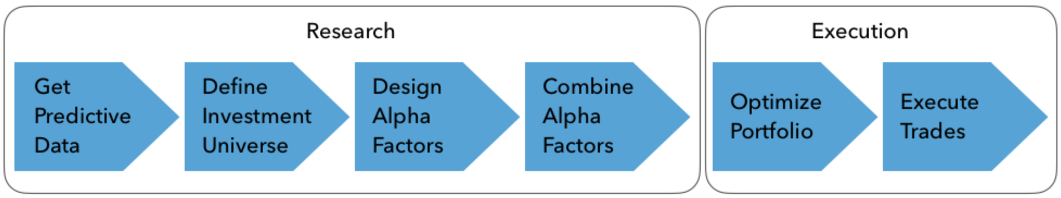
Adj. Close Price
150
100
Diluted Earnings per Share
9
8
7
Trailing P/E Ratio
15
10
Jan 2015
Jul
Jan 2016
Jul
Jan 2017
Jul
Jan 2018
Date

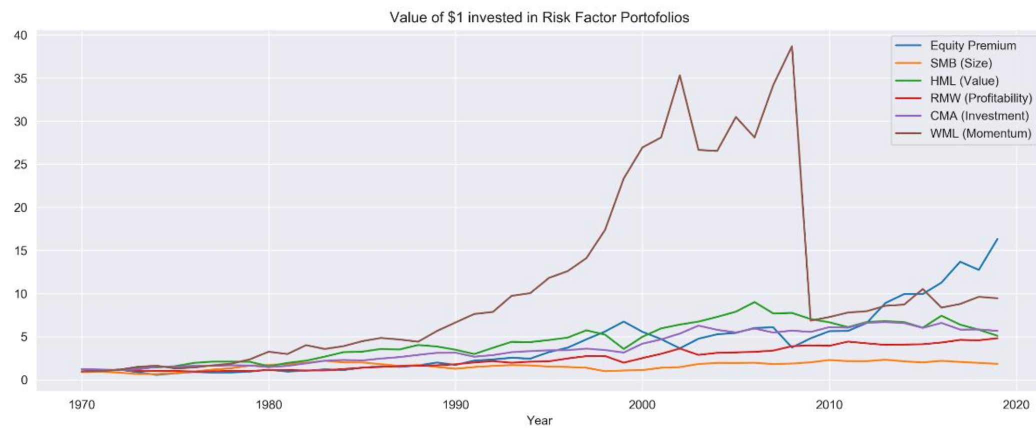


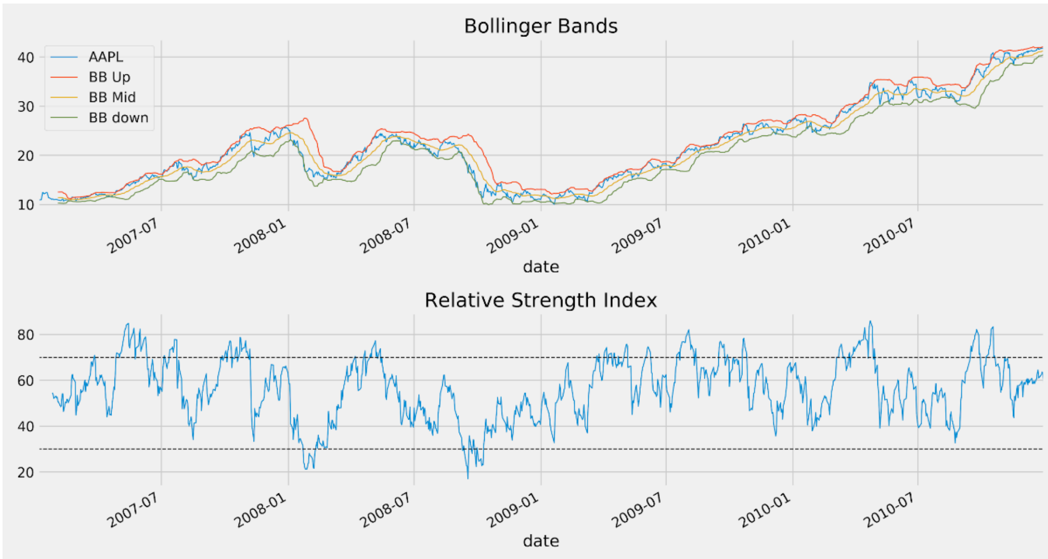
Read
Write
Size
Data
Mixed
Numeric
Parquet
HDF Table
HDF Fixed
CSV
Storage
seconds (log scale)
GB

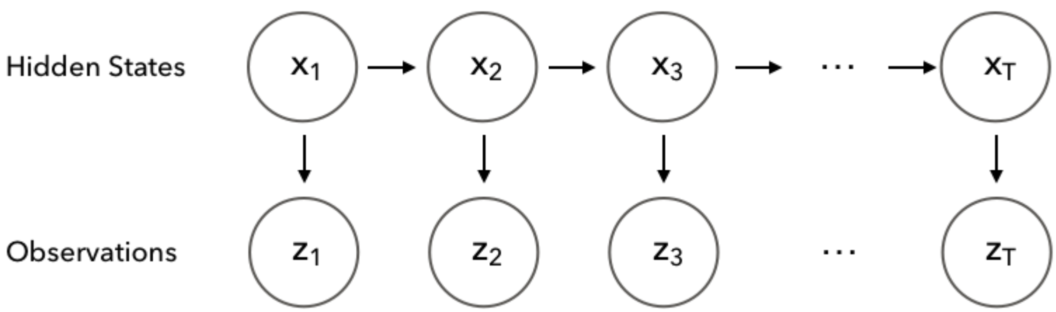
Chapter 3: Alternative Data for Finance – Categories and Use Cases

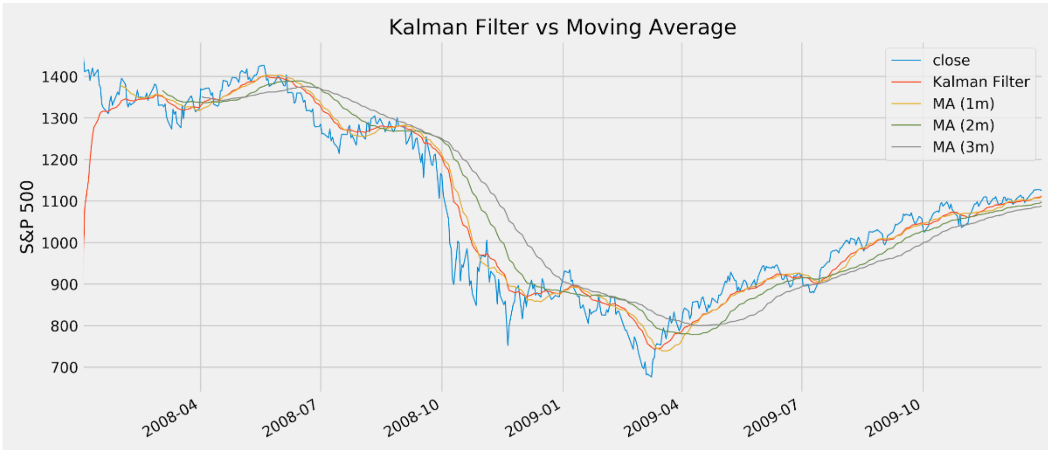
Chapter 4: Financial Feature Engineering – How to Research Alpha Factors



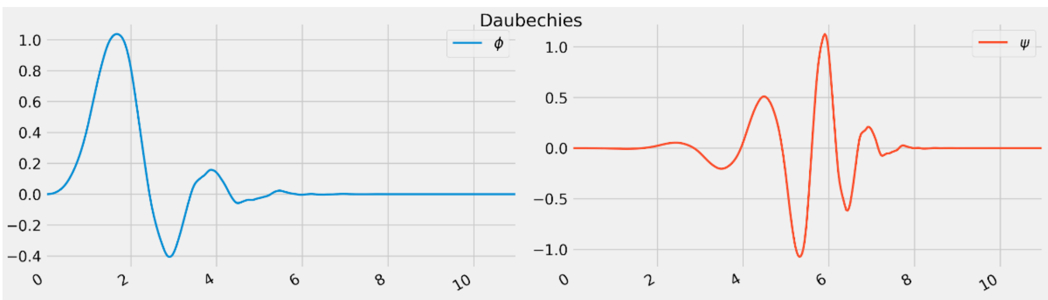



Bollinger Bands
AAPL
BB Up
BB Mid
BB down
40
30
20
10
2007-07
2008-01
2008-07
2009-01
2009-07
2010-01
2010-07
date
Relative Strength Index
80
60
40
20
2007-07
2008-01
2008-07
2009-01
2009-07
2010-01
2010-07
date

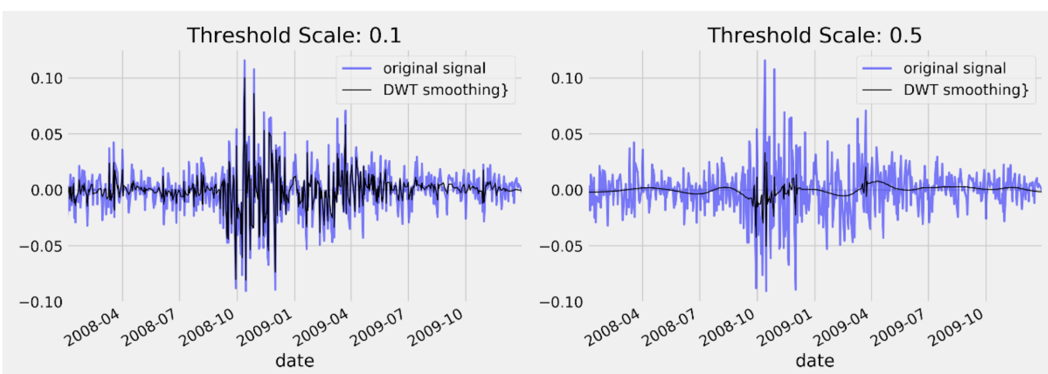

Hidden States
x_1
x_2
x_3
...
x_T
Observations
z_1
z_2
z_3
...
z_T



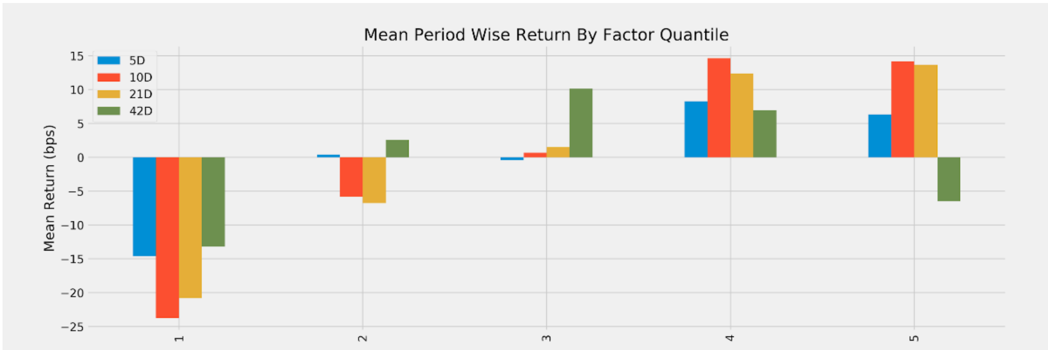
Kalman Filter vs Moving Average
close
Kalman Filter
MA (1m)
MA (2m)
MA (3m)
S&P 500
1400
1300
1200
1100
1000
900
800
700
2008-04
2008-07
2008-10
2009-01
2009-04
2009-07
2009-10

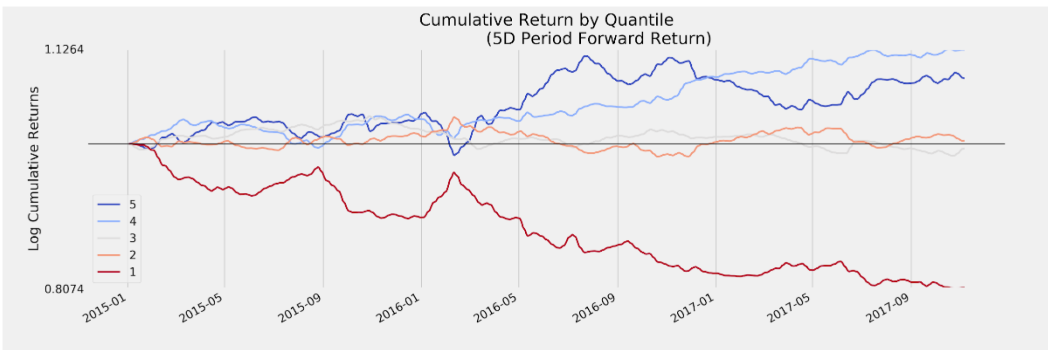


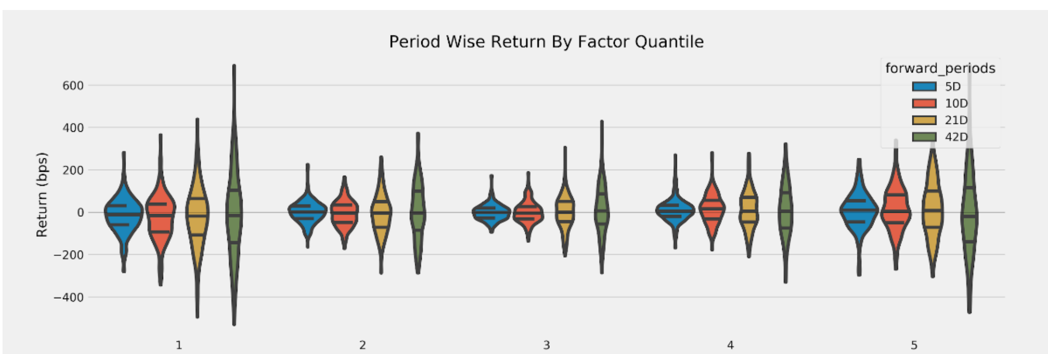
Daubechies
ϕ
ψ

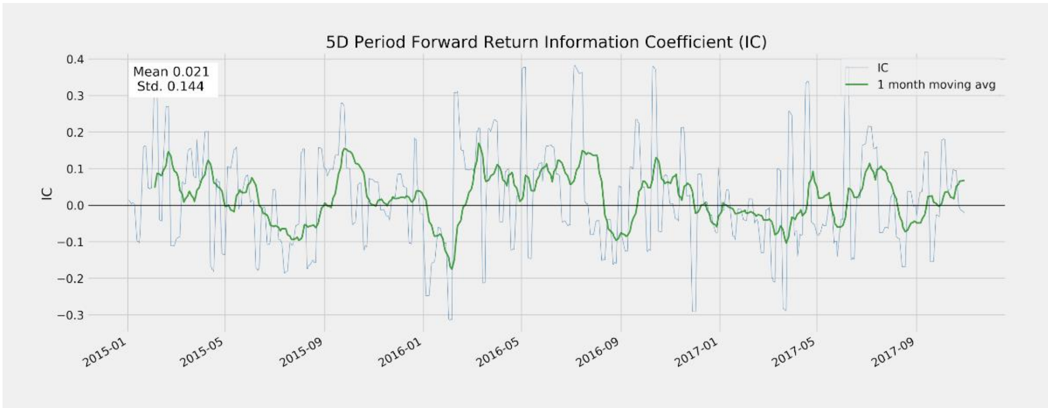


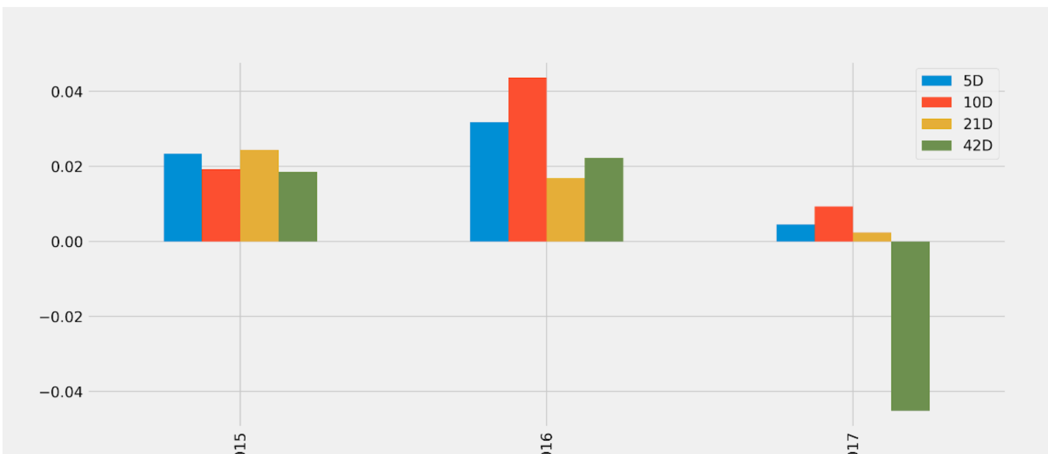
Threshold Scale: 0.1
Threshold Scale: 0.5
original signal
DWT smoothing}
date


Mean Period Wise Return By Factor Quantile
5D
10D
21D
42D
Mean Return (bps)
15
10
5
0
−5
−10
−15
−20
−25
1
2
3
4
5

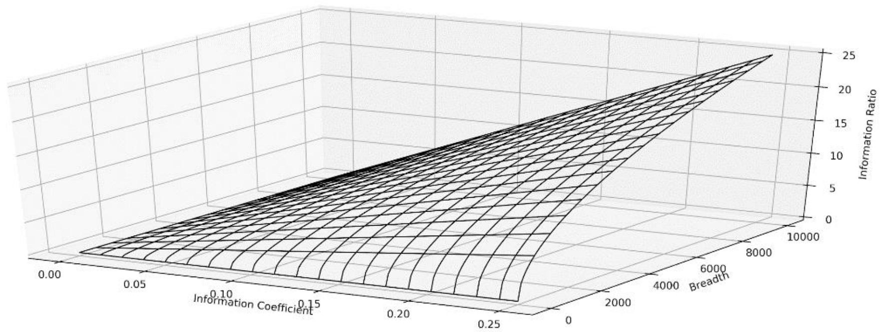

Cumulative Return by Quantile
(5D Period Forward Return)
1.1264
0.8074
Log Cumulative Returns
5
4
3
2
1
2015-01
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09

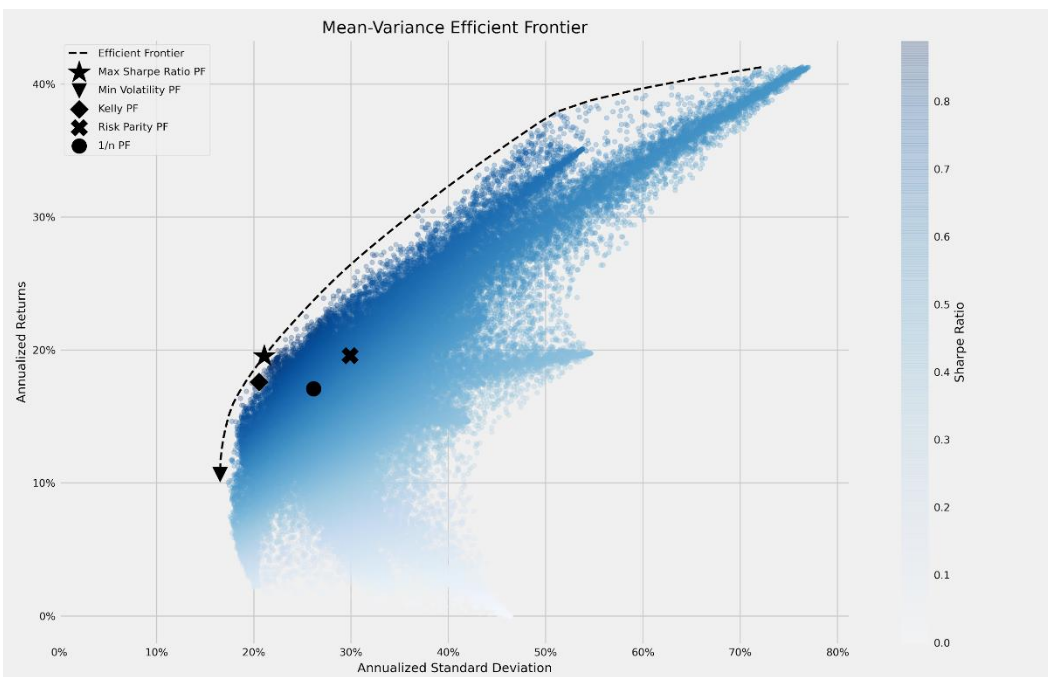

Period Wise Return By Factor Quantile
forward_periods
5D
10D
21D
42D
Return (bps)
600
400
200
0
−200
−400
1
2
3
4
5

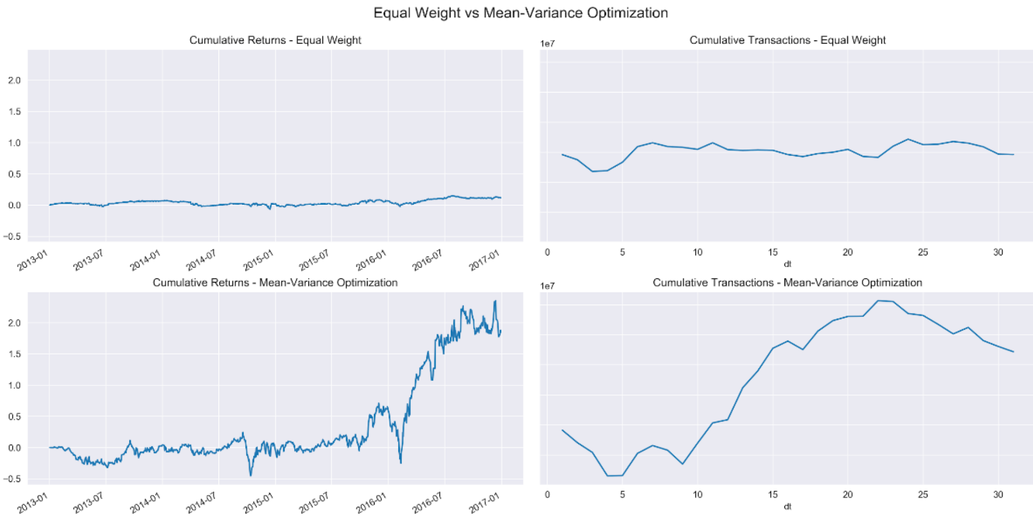

5D Period Forward Return Information Coefficient (IC)
Mean 0.021
Std. 0.144
IC
1 month moving avg
2015-01
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09


5D
10D
21D
42D
015
016
017

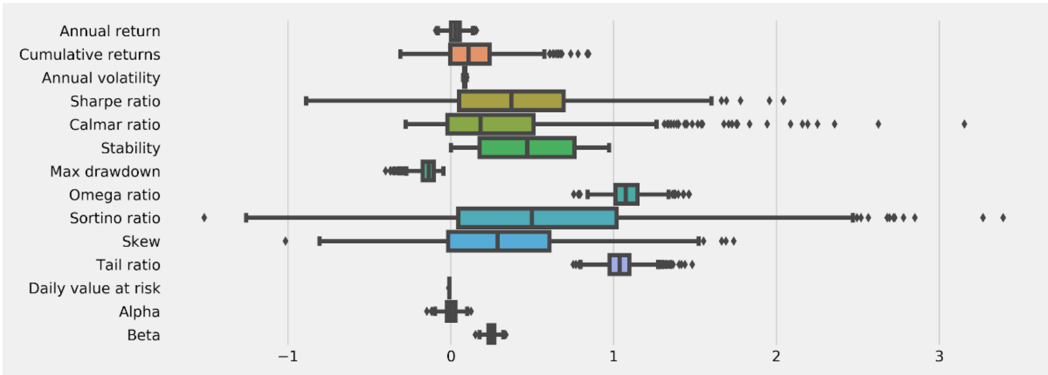
Chapter 5: Portfolio Optimization and Performance Evaluation

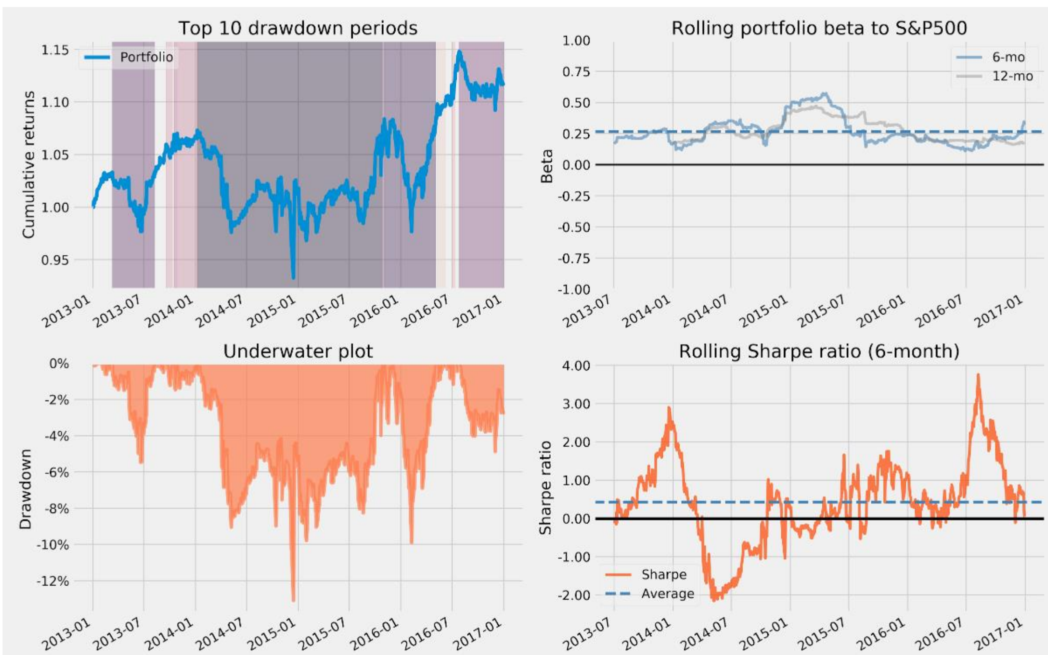


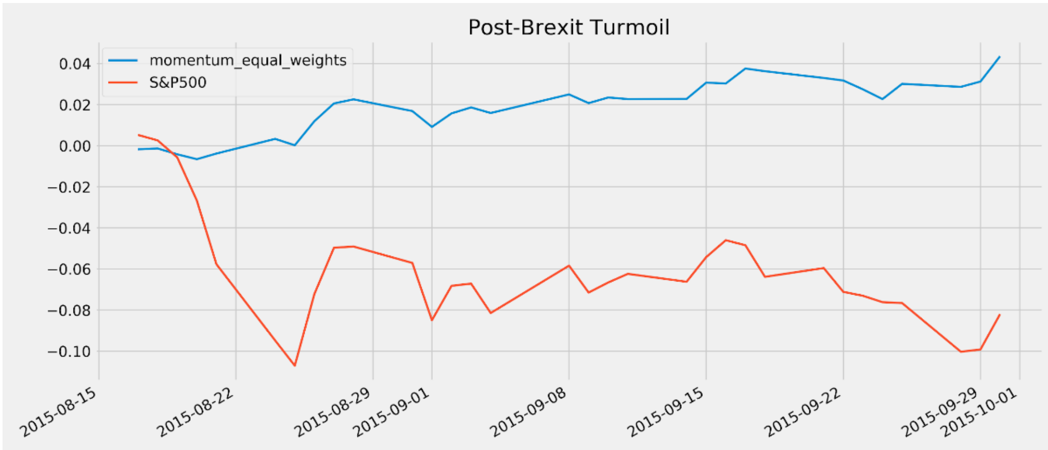



Equal Weight vs Mean-Variance Optimization
Cumulative Returns - Equal Weight
Cumulative Transactions - Equal Weight
Cumulative Returns - Mean-Variance Optimization
Cumulative Transactions - Mean-Variance Optimization
1e7
dt


S&P500
Backtest
Live
Cumulative returns

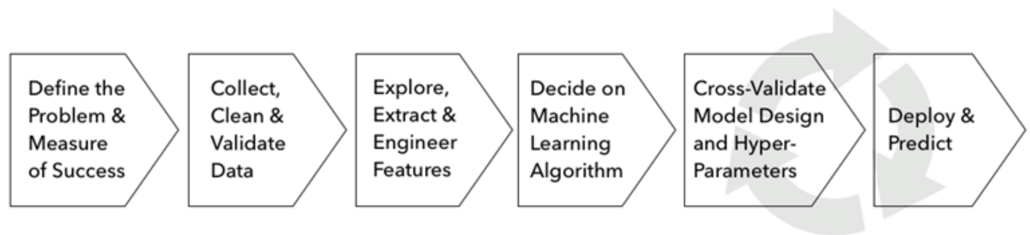

Annual return
Cumulative returns
Annual volatility
Sharpe ratio
Calmar ratio
Stability
Max drawdown
Omega ratio
Sortino ratio
Skew
Tail ratio
Daily value at risk
Alpha
Beta
−1
0
1
2
3

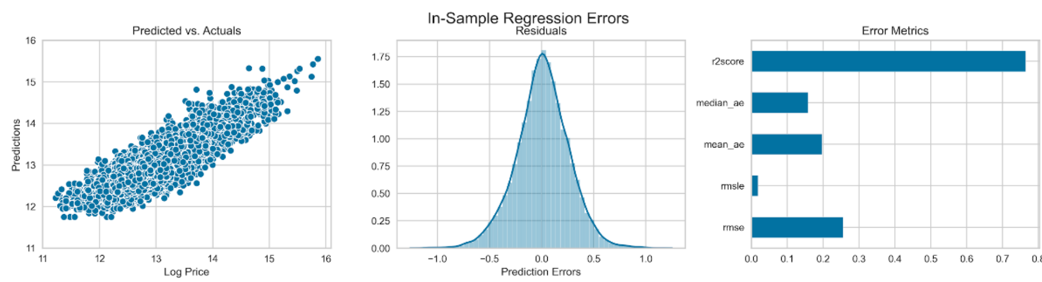

Top 10 drawdown periods
Portfolio
Cumulative returns
Rolling portfolio beta to S&P500
6-mo
12-mo
Beta
Underwater plot
Drawdown
Rolling Sharpe ratio (6-month)
Sharpe
Average
Sharpe ratio



Post-Brexit Turmoil
momentum_equal_weights
S&P500
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
−0.10
2015-08-15
2015-08-22
2015-08-29
2015-09-01
2015-09-08
2015-09-15
2015-09-22
2015-09-29
2015-10-01

Chapter 6: The Machine Learning Process





		Actual (Truth): Positive	Actual (Truth): Negative
Prediction	Positive	True Positive (TP)	False Positive (FP)
Prediction	Negative	False Negative (FN)	True Negative (TN)

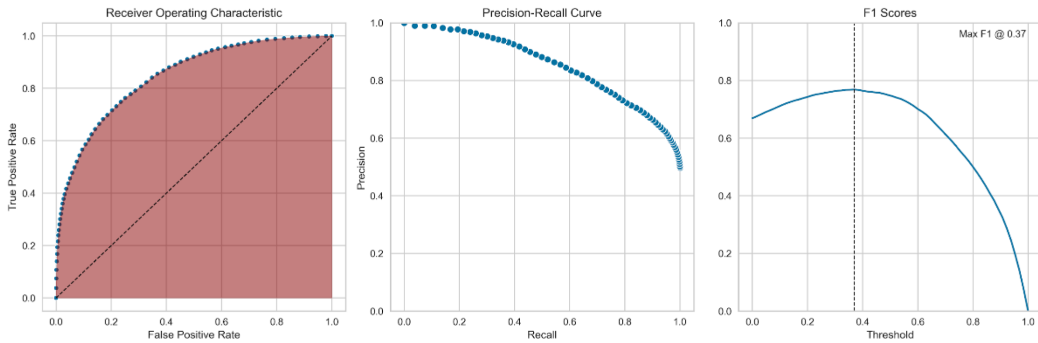
$$\text{Accuracy} = \frac{\text{\# Correct Predictions}}{\text{\# Cases}} = \frac{TP + TN}{TP + FP + TN + FN}$$

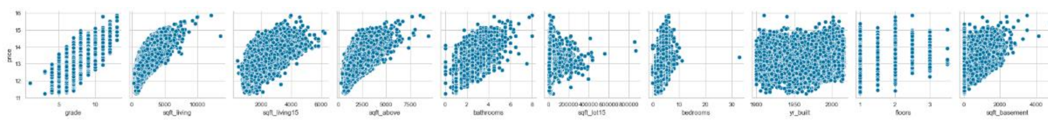
$$\text{True Positive Rate (Sensitivity, Recall)} = \frac{\text{\# Correct Positive Predictions}}{\text{\# Positive Cases}} = \frac{TP}{TP + FN}$$

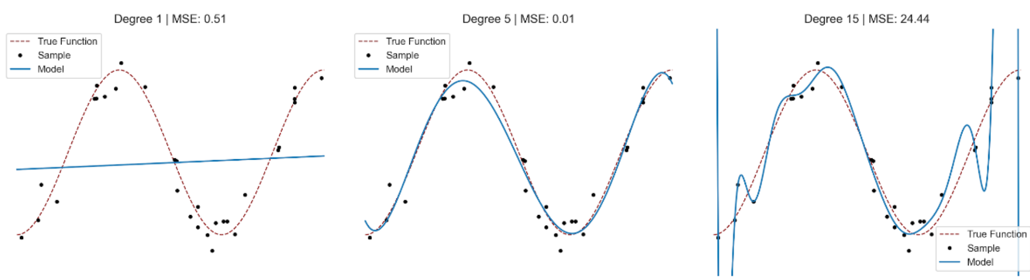
$$\text{False Negative Rate (Miss Rate)} = 1 - \text{True Positive Rate}$$

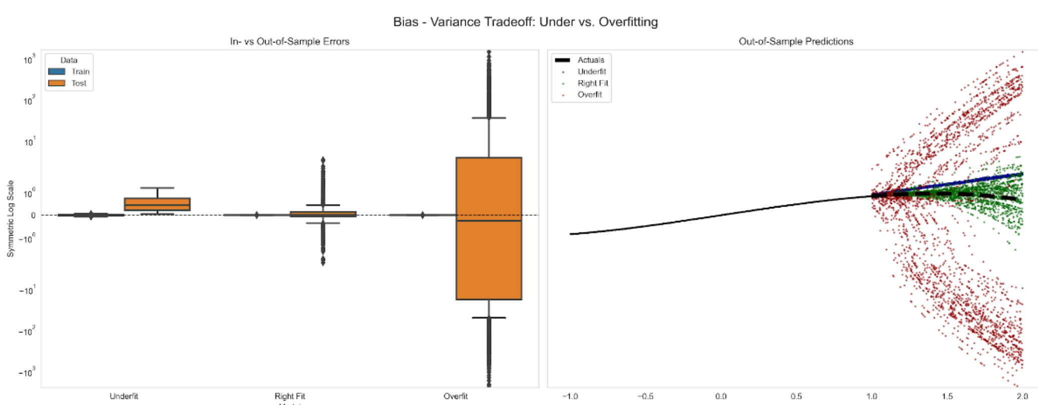
$$\text{True Negative Rate (Specificity)} = \frac{\text{\# Correct Negative Predictions}}{\text{\# Negative Cases}} = \frac{TN}{TN + FP}$$

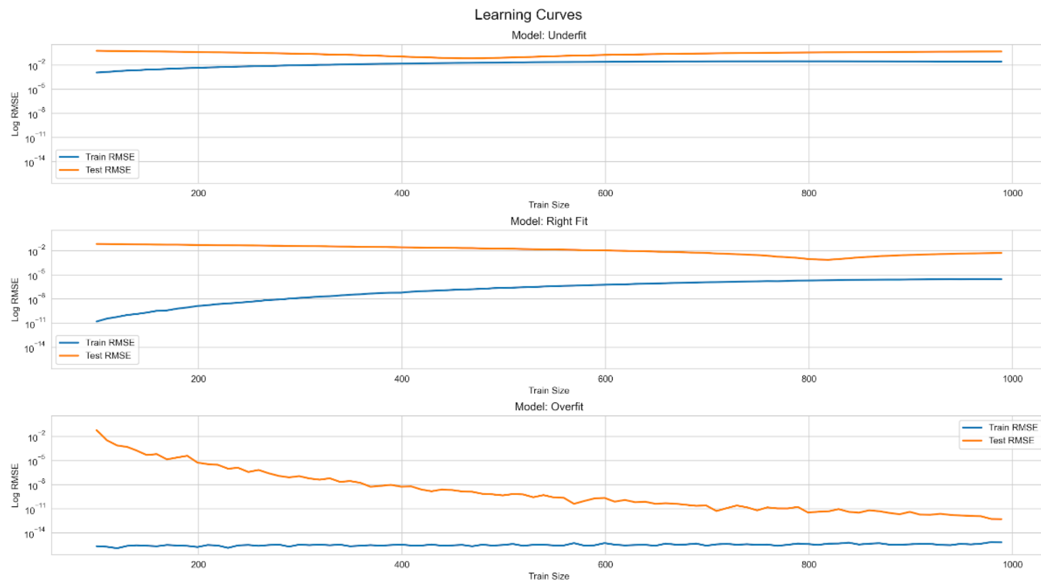
$$\text{False Positive Rate (Fall-Out)} = 1 - \text{True Negative Rate}$$

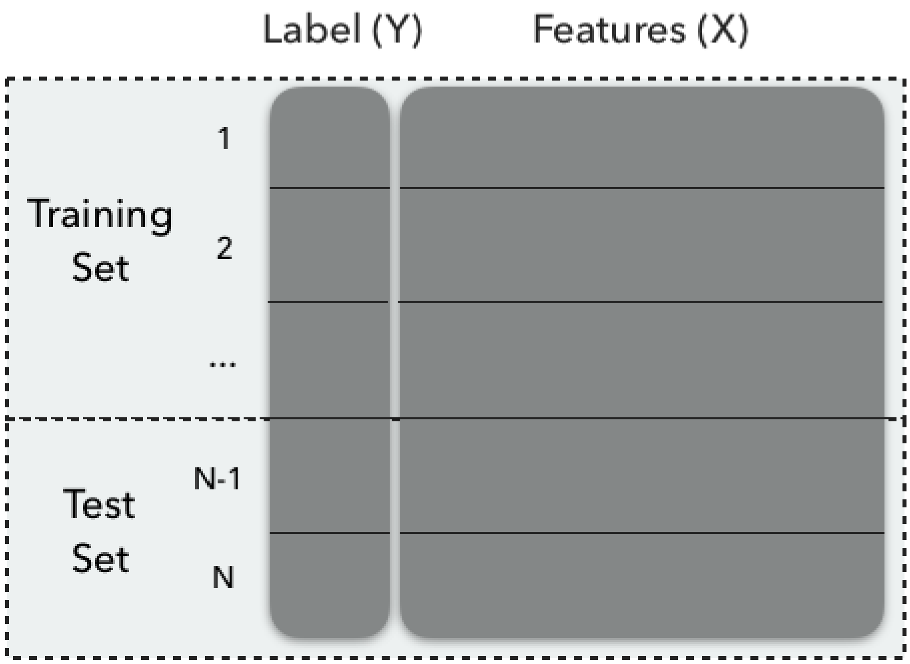

Receiver Operating Characteristic
True Positive Rate
False Positive Rate
Precision-Recall Curve
Precision
Recall
F1 Scores
Max F1 @ 0.37
Threshold

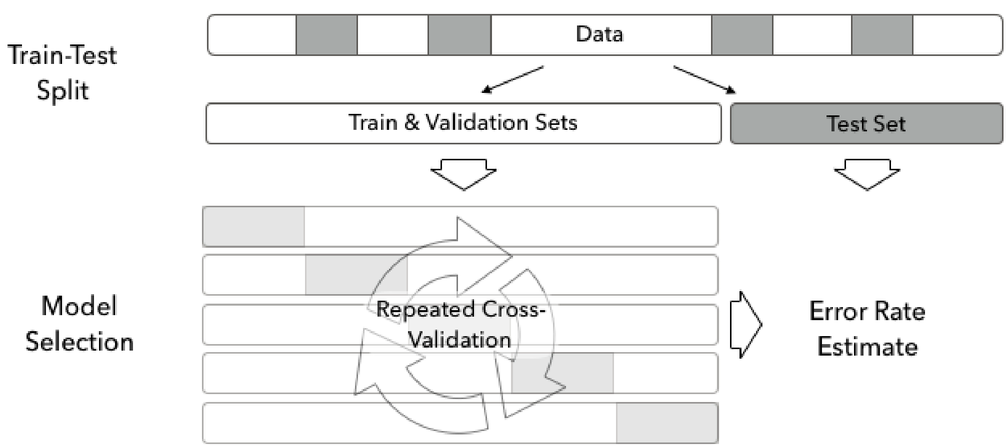

price
grade
sqft_living
sqft_living15
sqft_above
bathrooms
sqft_lot15
bedrooms
yr_built
floors
sqft_basement

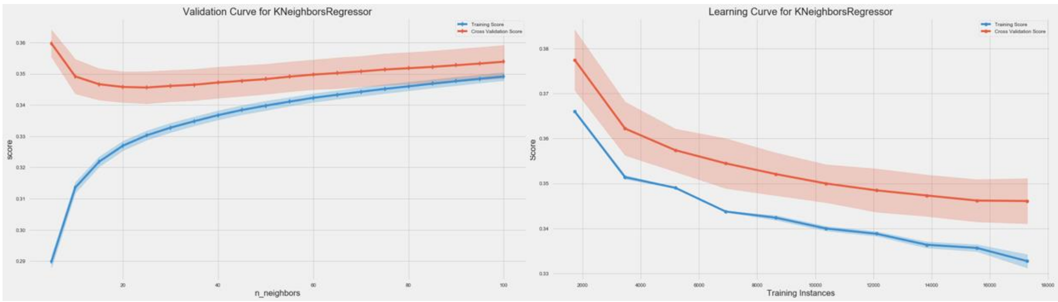

Degree 1 | MSE: 0.51
Degree 5 | MSE: 0.01
Degree 15 | MSE: 24.44
True Function
Sample
Model


Bias - Variance Tradeoff: Under vs. Overfitting
In- vs Out-of-Sample Errors
Data
Train
Test
Symmetric Log Scale
Underfit
Right Fit
Overfit
Model
Out-of-Sample Predictions
Actuals
Underfit
Right Fit
Overfit
x


Learning Curves
Model: Underfit
Log RMSE
Train RMSE
Test RMSE
Train Size
Model: Right Fit
Log RMSE
Train RMSE
Test RMSE
Train Size
Model: Overfit
Log RMSE
Train RMSE
Test RMSE
Train Size

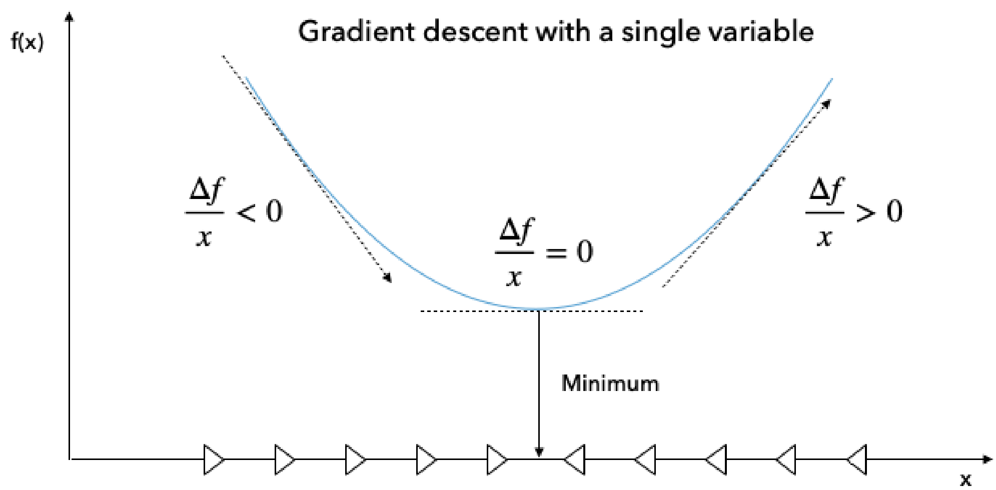

Label (Y)
Features (X)
Training Set
1
2
...
Test Set
N-1
N


Train-Test Split
Data
Train & Validation Sets
Test Set
Model Selection
Repeated Cross-Validation
Error Rate Estimate

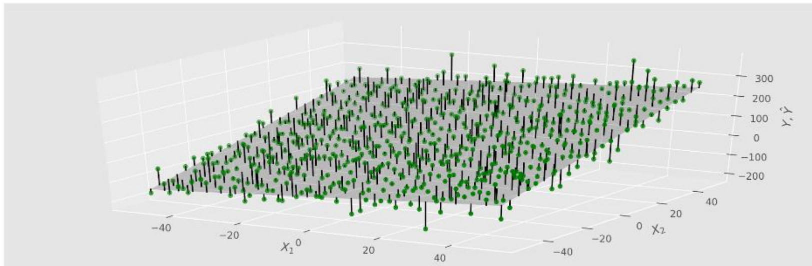


Validation Curve for KNeighborsRegressor
Training Score
Cross Validation Score
score
n_neighbors
Learning Curve for KNeighborsRegressor
Training Score
Cross Validation Score
Score
Training Instances

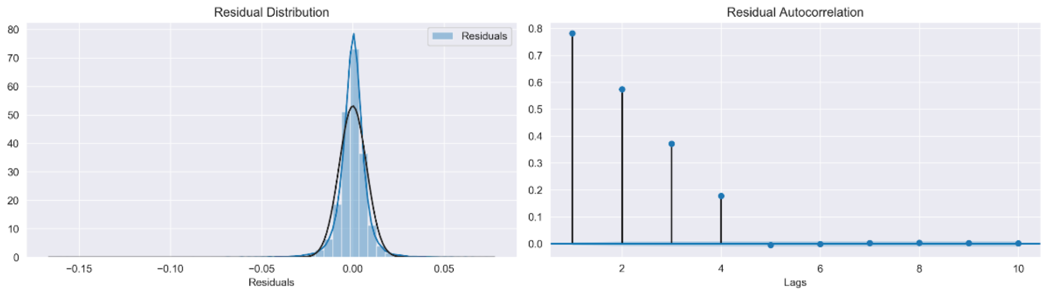
Chapter 7: Linear Models – From Risk Factors to Return Forecasts

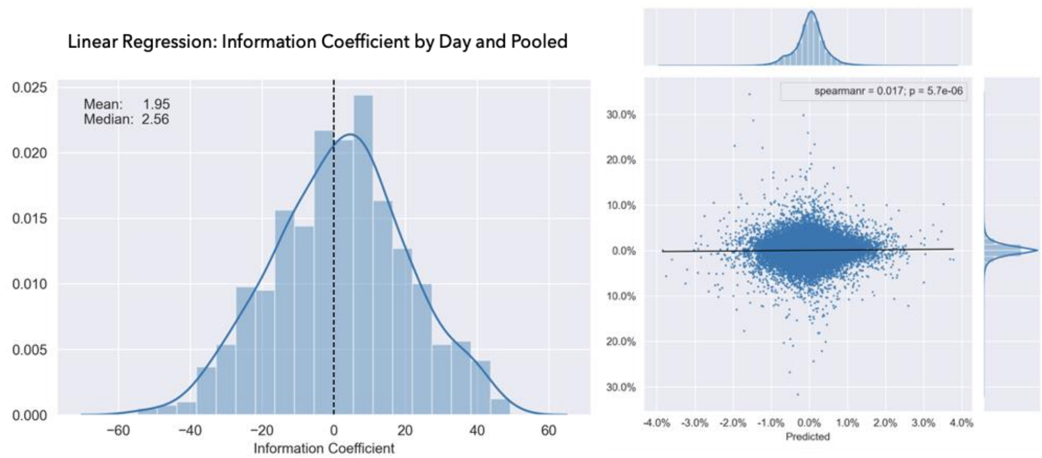


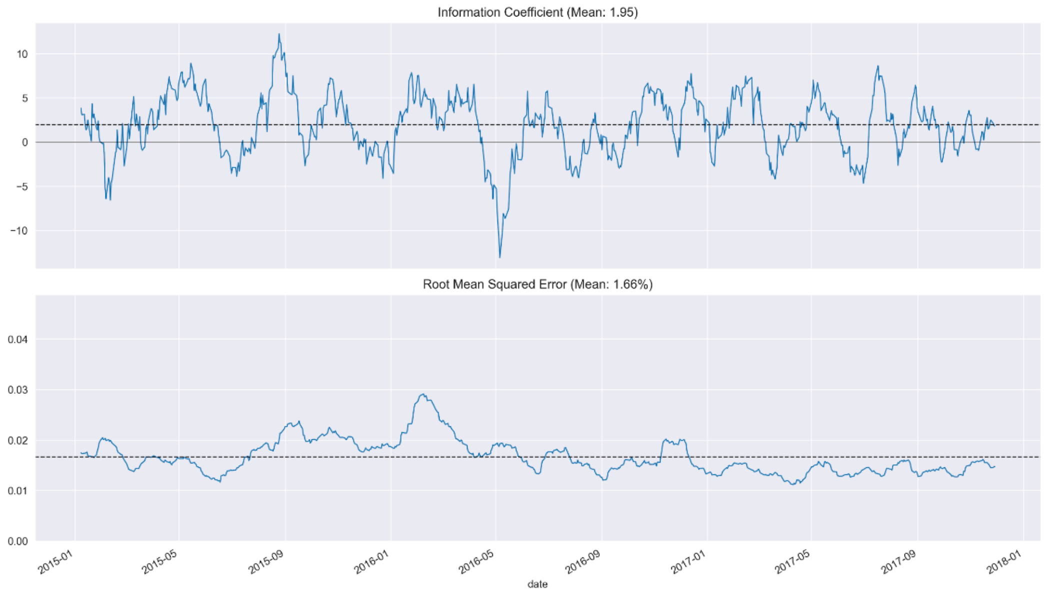

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.791
Model:                            OLS   Adj. R-squared:                  0.790
Method:                 Least Squares   F-statistic:                     1176.
Date:                Thu, 14 Nov 2019   Prob (F-statistic):         4.33e-212
Time:                        18:58:15   Log-Likelihood:                -3309.2
No. Observations:                 625   AIC:                             6624.
Df Residuals:                     622   BIC:                             6638.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         53.2923      1.934     27.561      0.000      49.495      57.089
X_1            0.9904      0.064     15.390      0.000       0.864       1.117
X_2            2.9600      0.064     45.996      0.000       2.834       3.086
==============================================================================
Omnibus:                        0.267   Durbin-Watson:                   2.148
Prob(Omnibus):                  0.875   Jarque-Bera (JB):                0.149
Skew:                           0.014   Prob(JB):                        0.928
Kurtosis:                       3.071   Cond. No.                         30.0
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

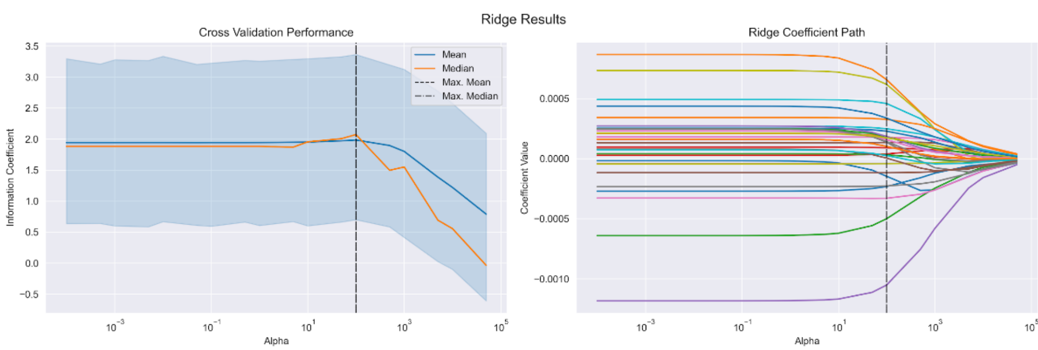


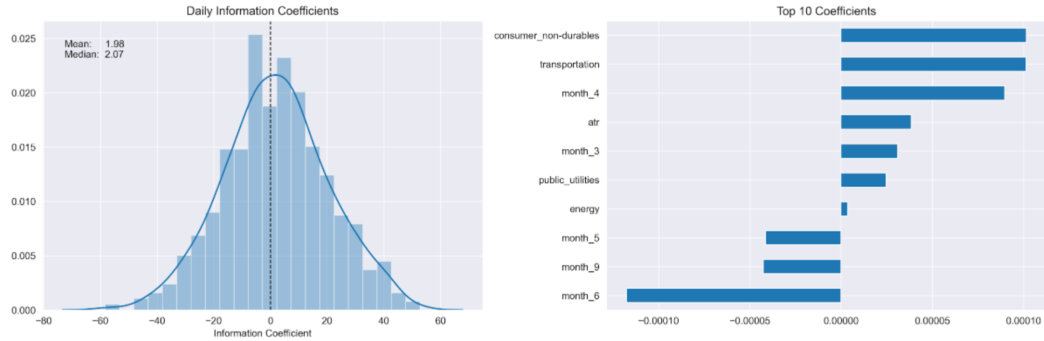

```
                    LinearFactorModel Estimation Summary
================================================================================
No. Test Portfolios:                  17   R-squared:                      0.6944
No. Factors:                           6   J-statistic:                    19.501
No. Observations:                     95   P-value                         0.0527
Date:                   Thu, Nov 14 2019   Distribution:                chi2(11)
Time:                           19:34:04
Cov. Estimator:                   robust

                            Risk Premia Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
Mkt-RF         1.2436     0.3928     3.1662     0.0015      0.4738      2.0135
SMB           -0.0049     0.6993    -0.0070     0.9945     -1.3754      1.3657
HML           -0.6882     0.5360    -1.2838     0.1992     -1.7388      0.3625
RMW           -0.2373     0.6729    -0.3527     0.7243     -1.5562      1.0815
CMA           -0.3181     0.4633    -0.6865     0.4924     -1.2261      0.5900
RF            -0.0133     0.0132    -1.0026     0.3161     -0.0392      0.0127
==============================================================================

Covariance estimator:
HeteroskedasticCovariance
See full_summary for complete results
```


Residual Distribution
Residuals
Residual Autocorrelation
Lags

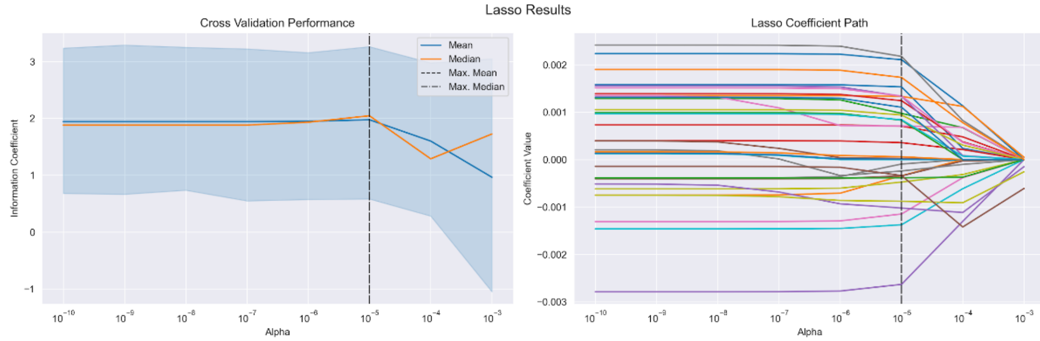

Linear Regression: Information Coefficient by Day and Pooled
Mean: 1.95
Median: 2.56
Information Coefficient
spearmanr = 0.017; p = 5.7e-06
Predicted

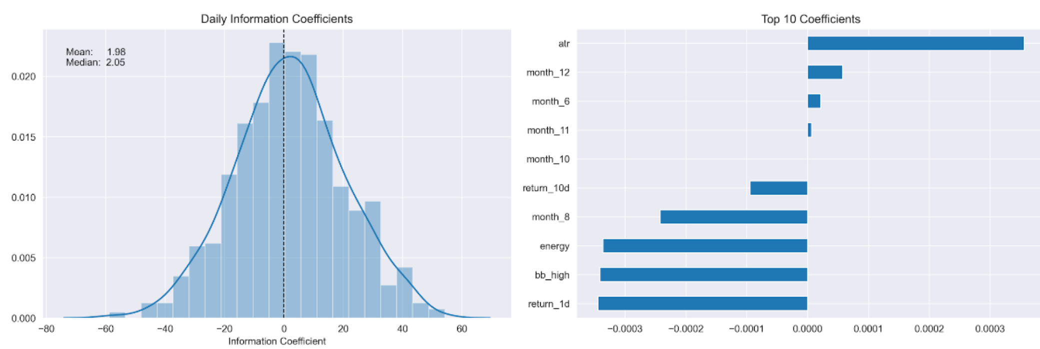

Information Coefficient (Mean: 1.95)
Root Mean Squared Error (Mean: 1.66%)
date
2015-01
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01

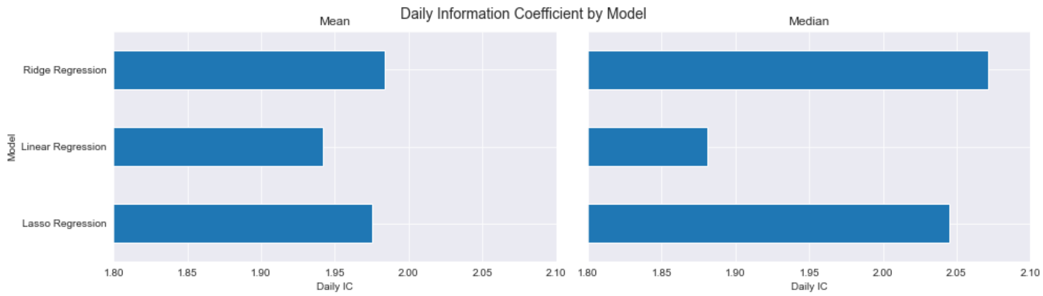

Ridge Results
Cross Validation Performance
Mean
Median
Max. Mean
Max. Median
Information Coefficient
Alpha
Ridge Coefficient Path
Coefficient Value
Alpha



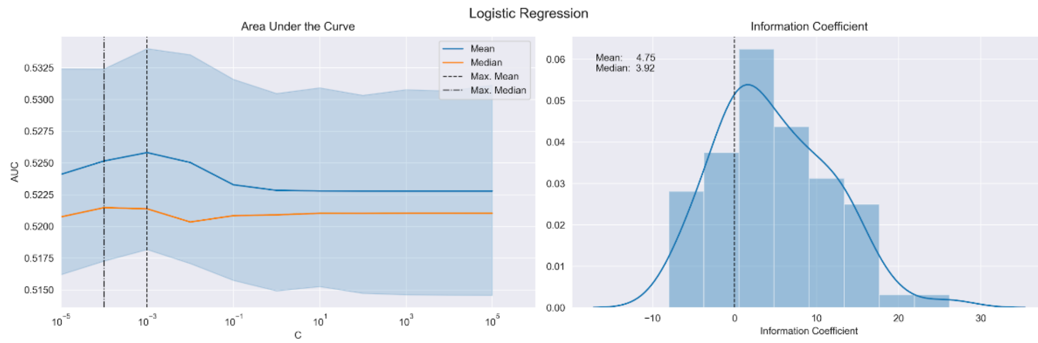
Lasso Results



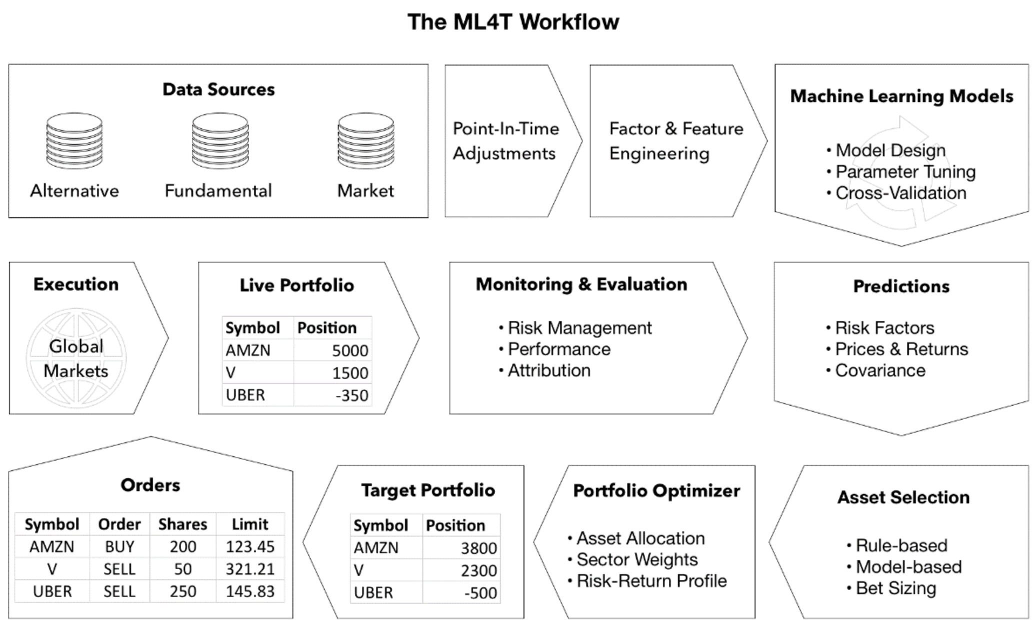


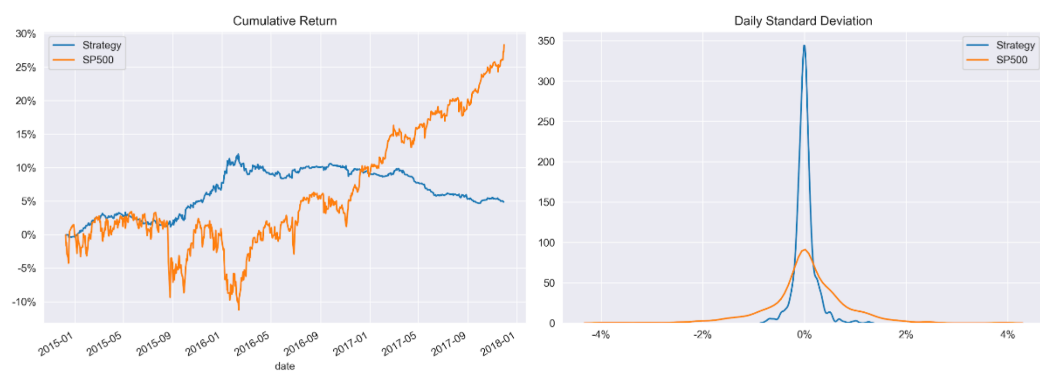



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:                  198
Model:                          Logit   Df Residuals:                      185
Method:                           MLE   Df Model:                           12
Date:                Mon, 10 Sep 2018   Pseudo R-squ.:                  0.5022
Time:                        20:27:53   Log-Likelihood:                -67.907
converged:                       True   LL-Null:                       -136.42
                                        LLR p-value:                 2.375e-23
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -8.5881      1.908     -4.502      0.000     -12.327      -4.849
realcons     130.1446     26.633      4.887      0.000      77.945     182.344
realinv       18.8414      4.053      4.648      0.000      10.897      26.786
realgovt     -19.0318      6.010     -3.166      0.002     -30.812      -7.252
realdpi      -52.2473     19.912     -2.624      0.009     -91.275     -13.220
m1            -1.3462      6.177     -0.218      0.827     -13.453      10.761
tbilrate      60.8607     44.350      1.372      0.170     -26.063     147.784
unemp          0.9487      0.249      3.818      0.000       0.462       1.436
infl         -60.9647     44.362     -1.374      0.169    -147.913      25.984
realint      -61.0453     44.359     -1.376      0.169    -147.987      25.896
quarter_2      0.1128      0.618      0.182      0.855      -1.099       1.325
quarter_3     -0.1991      0.609     -0.327      0.744      -1.393       0.995
quarter_4      0.0007      0.608      0.001      0.999      -1.191       1.192
==============================================================================
```

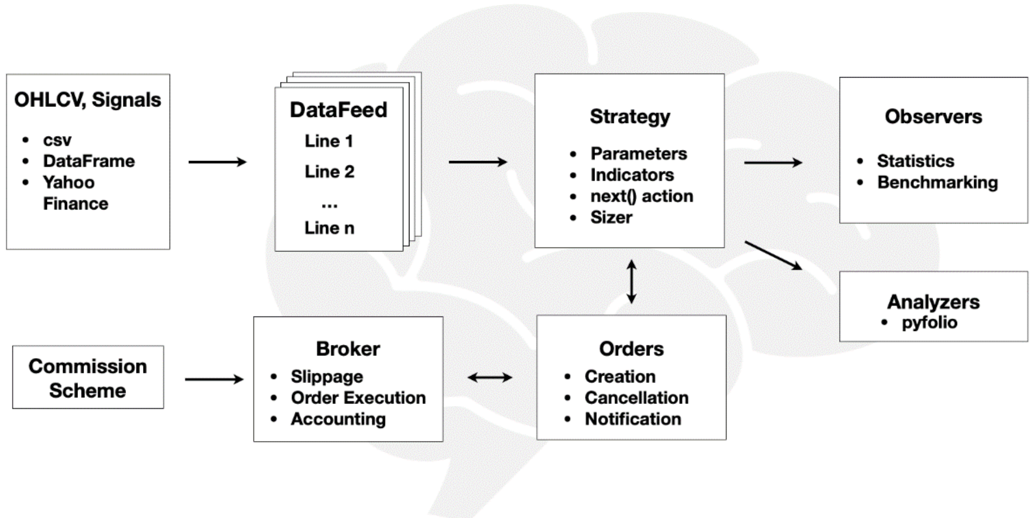

Logistic Regression
Area Under the Curve
Mean
Median
Max. Mean
Max. Median
AUC
C
Information Coefficient
Mean: 4.75
Median: 3.92
Information Coefficient

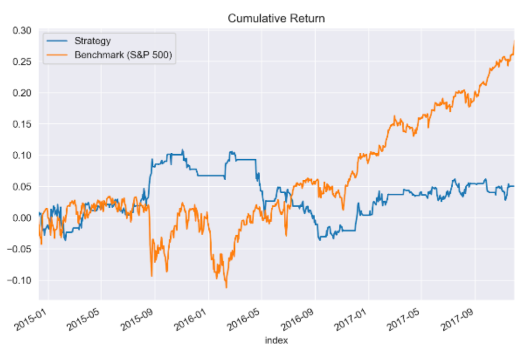
Chapter 8: The ML4T Workflow – From Model to Strategy Backtesting

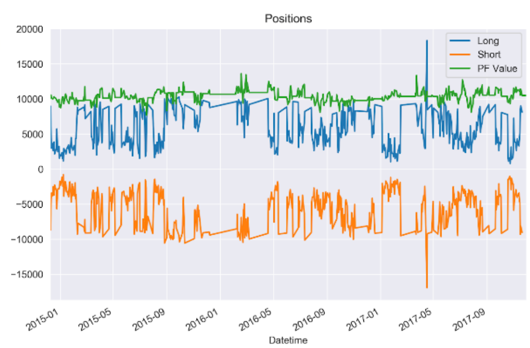




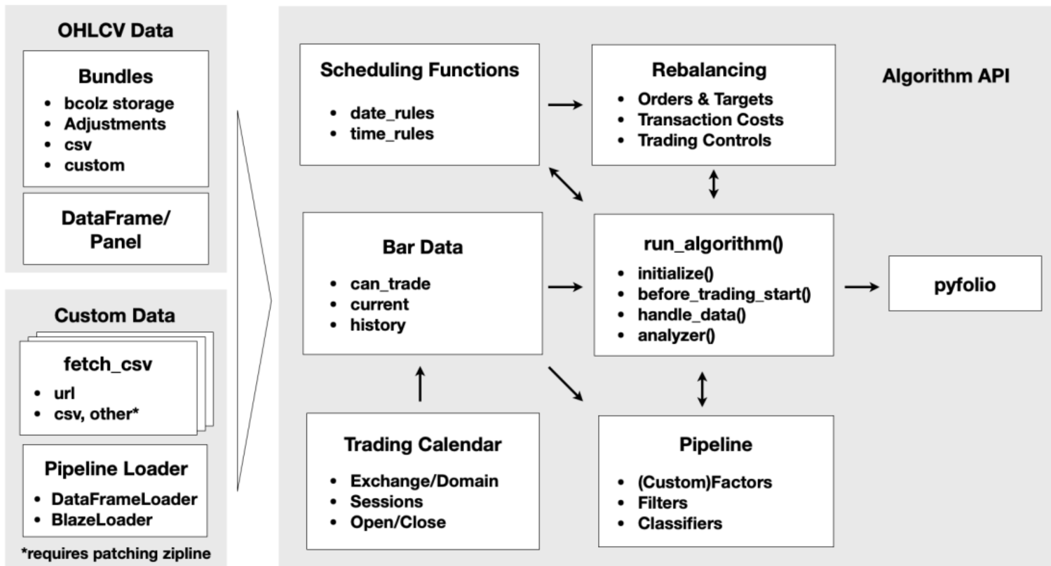
backtrader “Cerebro” Architecture

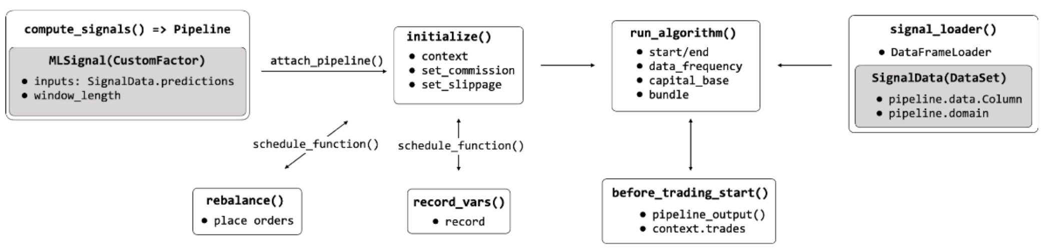


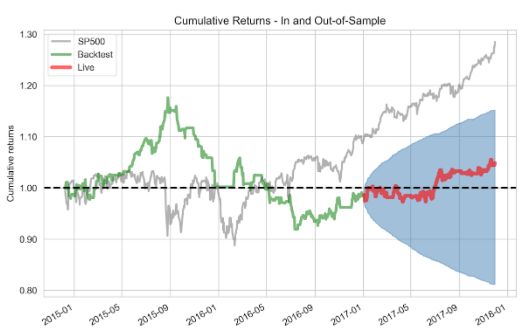


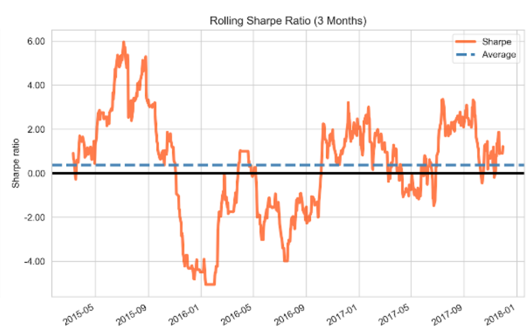


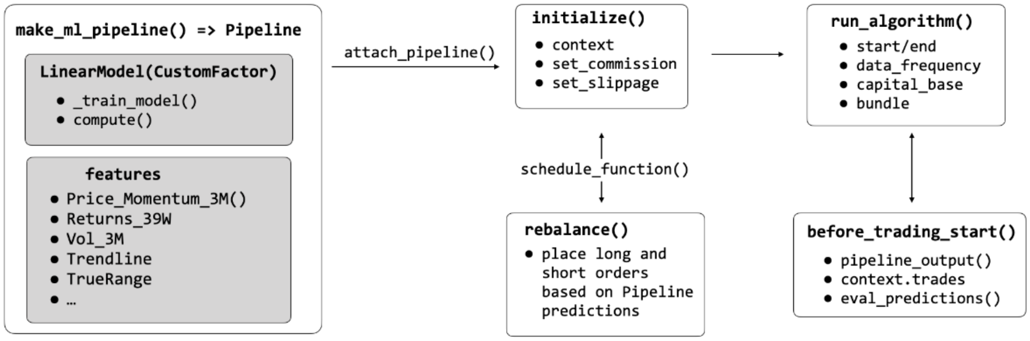
Zipline Architecture

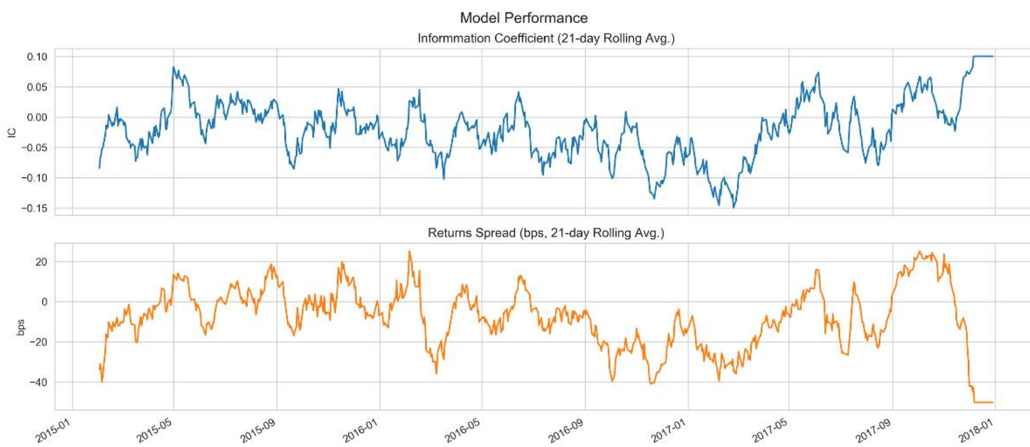


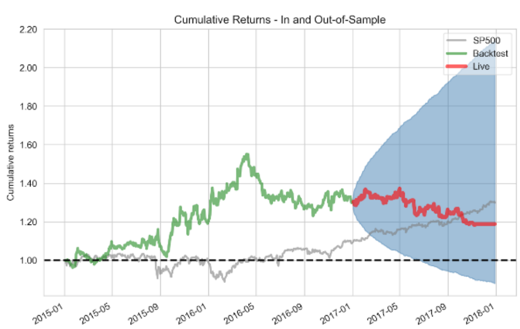


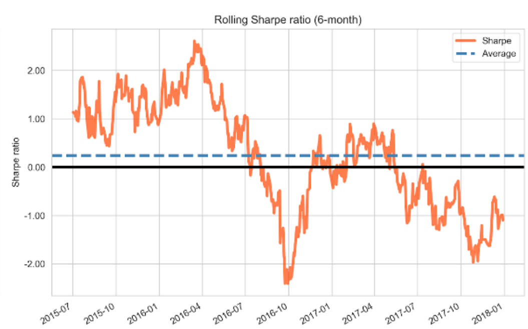




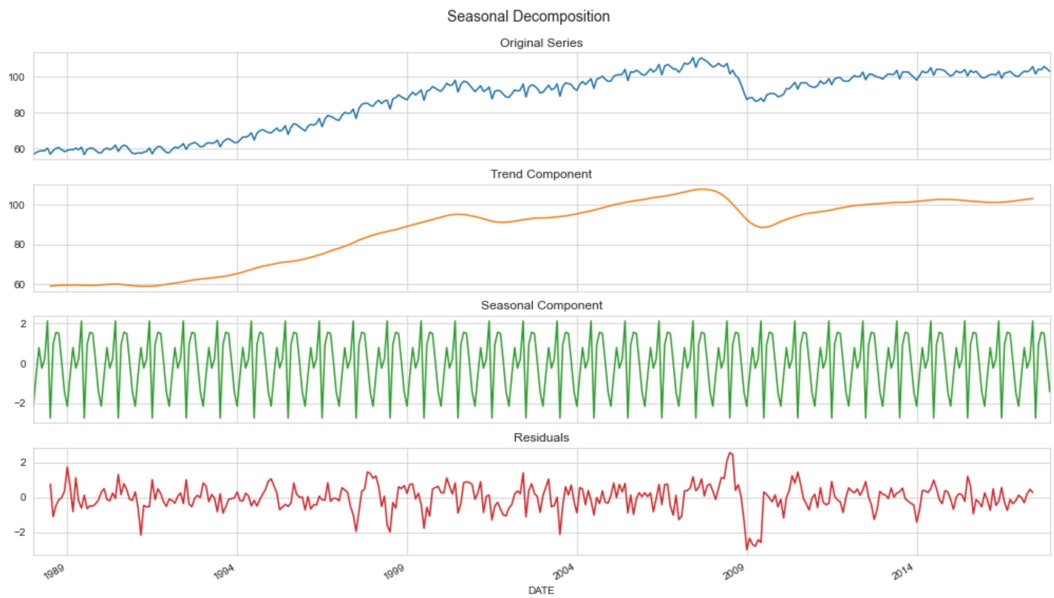

make_ml_pipeline() => Pipeline
LinearModel(CustomFactor)
_train_model()
compute()
features
Price_Momentum_3M()
Returns_39W
Vol_3M
Trendline
TrueRange
…
attach_pipeline()
initialize()
context
set_commission
set_slippage
run_algorithm()
start/end
data_frequency
capital_base
bundle
schedule_function()
rebalance()
place long and short orders based on Pipeline predictions
before_trading_start()
pipeline_output()
context.trades
eval_predictions()

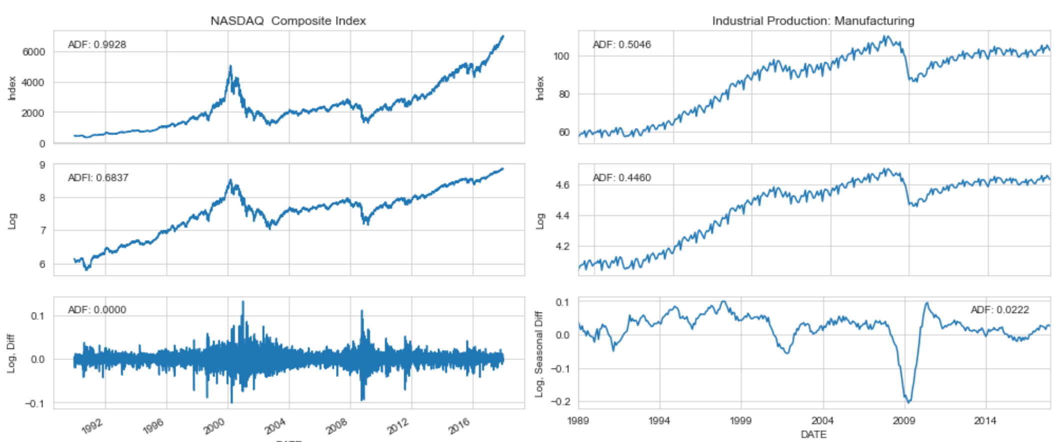

Model Performance
Information Coefficient (21-day Rolling Avg.)
IC
Returns Spread (bps, 21-day Rolling Avg.)
bps

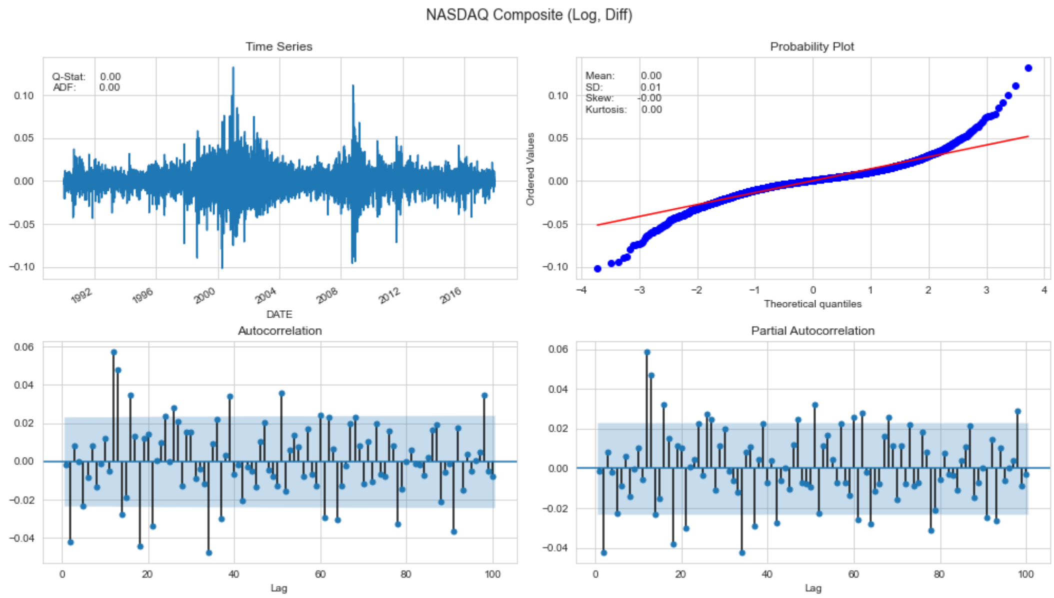

Cumulative Returns - In and Out-of-Sample
SP500
Backtest
Live
Cumulative returns

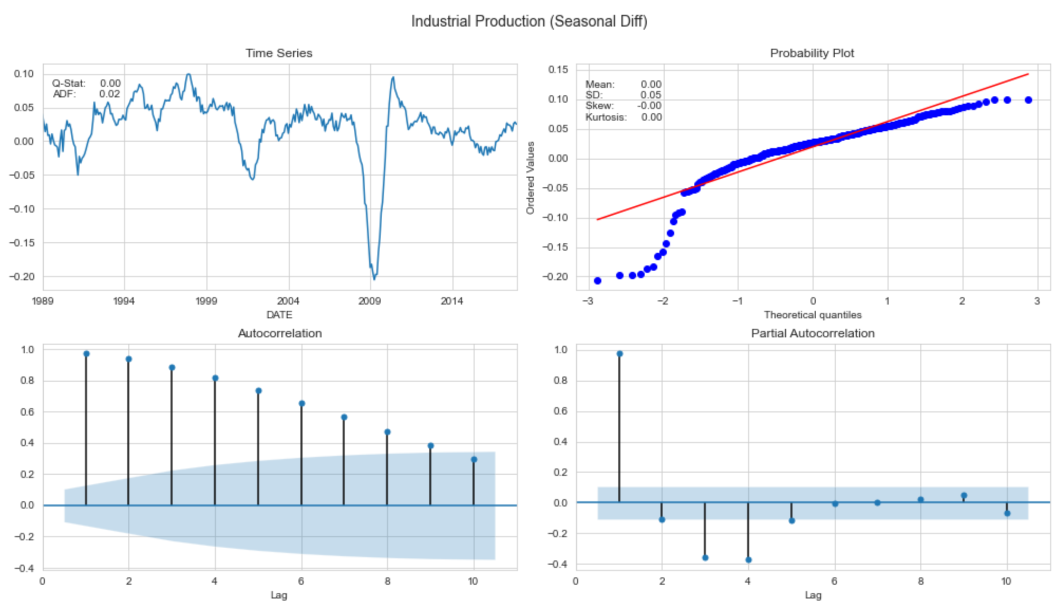

Rolling Sharpe ratio (6-month)
Sharpe
Average
Sharpe ratio

Chapter 9: Time-Series Models for Volatility Forecasts and Statistical Arbitrage

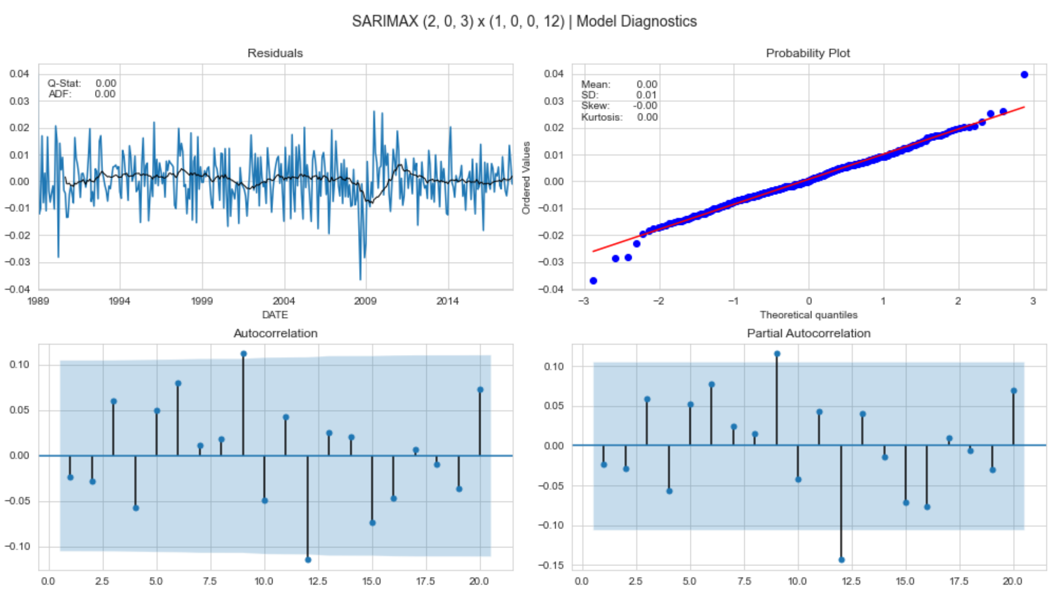


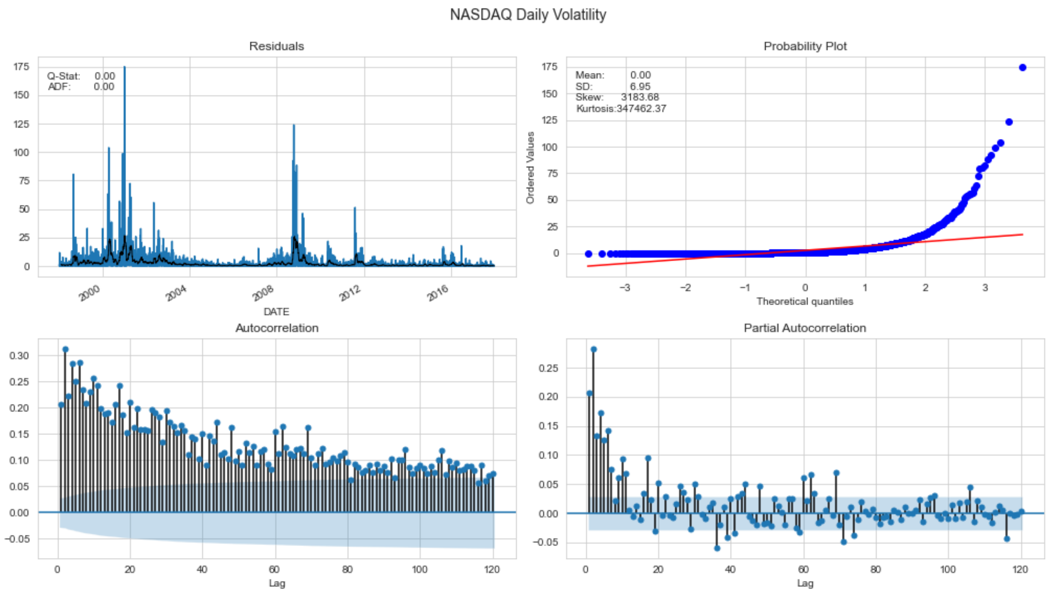



NASDAQ Composite (Log, Diff)
Time Series
Q-Stat: 0.00
ADF: 0.00
DATE
Probability Plot
Mean: 0.00
SD: 0.01
Skew: -0.00
Kurtosis: 0.00
Ordered Values
Theoretical quantiles
Autocorrelation
Lag
Partial Autocorrelation
Lag

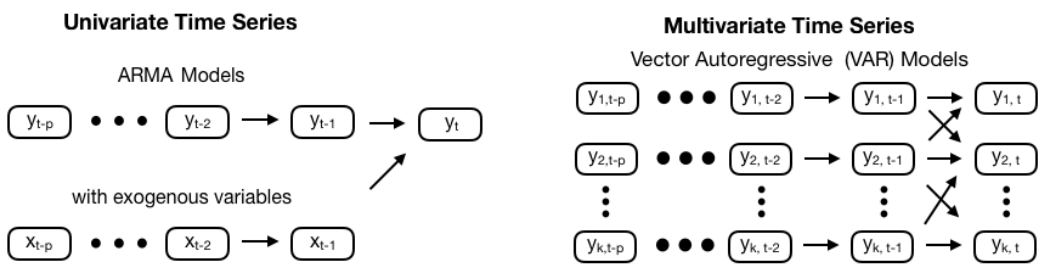

Industrial Production (Seasonal Diff)
Time Series
Q-Stat: 0.00
ADF: 0.02
DATE
Probability Plot
Mean: 0.00
SD: 0.05
Skew: -0.00
Kurtosis: 0.00
Ordered Values
Theoretical quantiles
Autocorrelation
Lag
Partial Autocorrelation
Lag

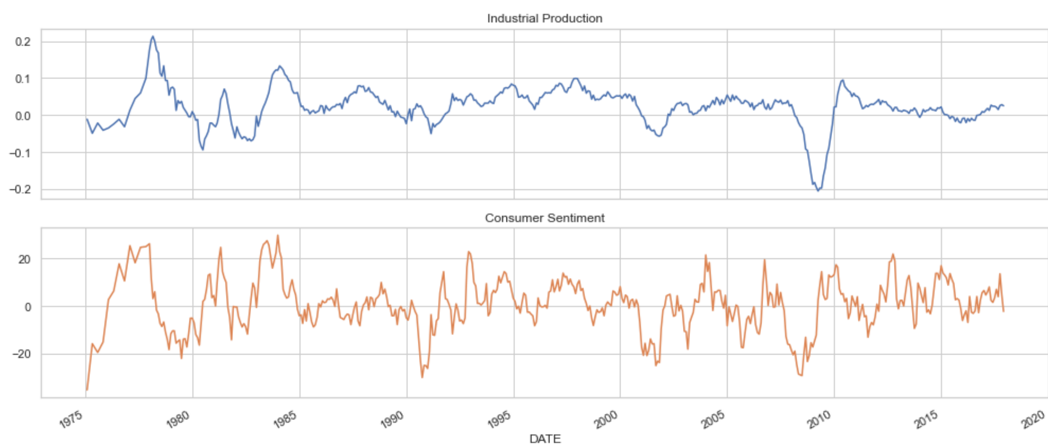
```
                                 Statespace Model Results
==========================================================================================
Dep. Variable:                             IPGMFN   No. Observations:                  348
Model:             SARIMAX(2, 0, 3)x(1, 0, 0, 12)   Log Likelihood                1139.719
Date:                            Sat, 22 Sep 2018   AIC                          -2265.438
Time:                                    17:48:17   BIC                          -2238.472
Sample:                                01-01-1989   HQIC                         -2254.702
                                     - 12-01-2017
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.4934      0.104     14.351      0.000       1.289       1.697
ar.L2         -0.5159      0.102     -5.083      0.000      -0.715      -0.317
ma.L1         -0.5499      0.114     -4.813      0.000      -0.774      -0.326
ma.L2          0.2872      0.062      4.662      0.000       0.166       0.408
ma.L3          0.1815      0.070      2.589      0.010       0.044       0.319
ar.S.L12      -0.4486      0.047     -9.533      0.000      -0.541      -0.356
sigma2      8.141e-05   5.65e-06     14.399      0.000    7.03e-05    9.25e-05
===================================================================================
Ljung-Box (Q):                       61.58   Jarque-Bera (JB):                 9.97
Prob(Q):                              0.02   Prob(JB):                         0.01
Heteroskedasticity (H):               1.07   Skew:                            -0.20
Prob(H) (two-sided):                  0.71   Kurtosis:                         3.73
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```



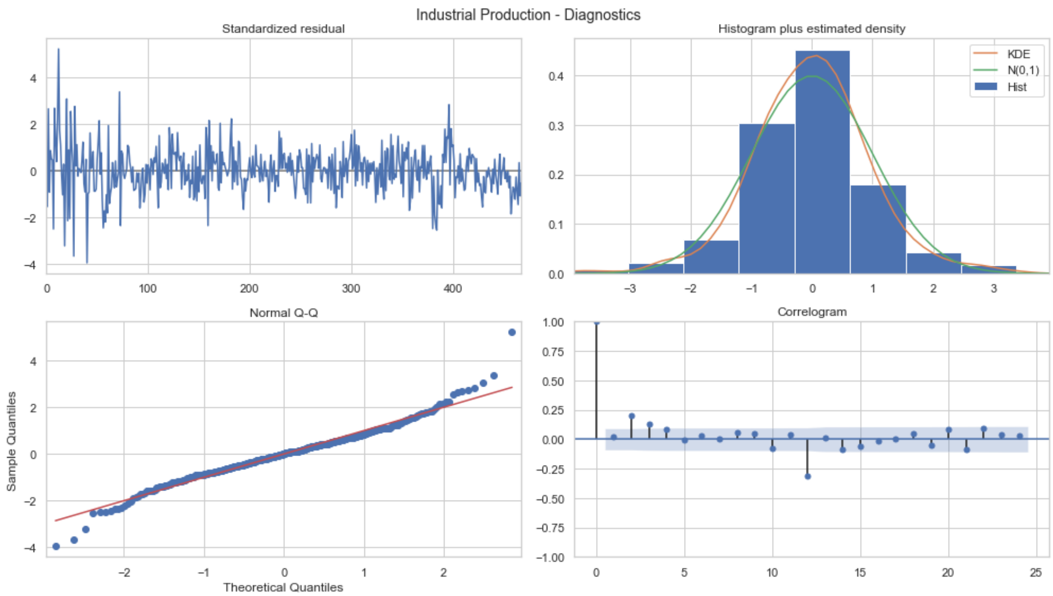


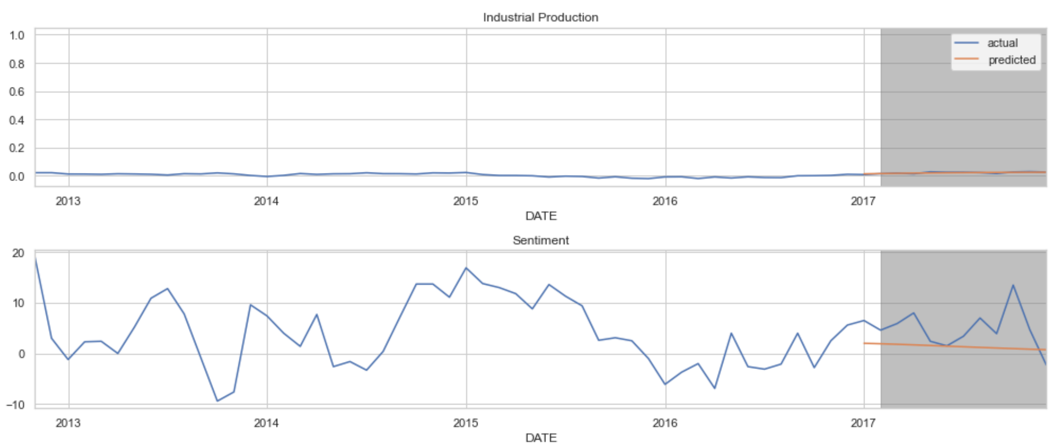

```
                     Constant Mean - GARCH Model Results
==============================================================================
Dep. Variable:              NASDAQCOM   R-squared:                      -0.001
Mean Model:             Constant Mean   Adj. R-squared:                 -0.001
Vol Model:                      GARCH   Log-Likelihood:               -7244.08
Distribution:                  Normal   AIC:                           14500.2
Method:            Maximum Likelihood   BIC:                           14539.1
                                        No. Observations:                 4851
Date:                Thu, Apr 16 2020   Df Residuals:                     4845
Time:                        22:41:39   Df Model:                            6
                                 Mean Model
============================================================================
                 coef    std err          t      P>|t|      95.0% Conf. Int.
----------------------------------------------------------------------------
mu             0.0526  1.416e-02      3.714  2.043e-04 [2.484e-02,8.036e-02]
                              Volatility Model
============================================================================
                 coef    std err          t      P>|t|      95.0% Conf. Int.
----------------------------------------------------------------------------
omega          0.0270  1.047e-02      2.574  1.005e-02 [6.430e-03,4.748e-02]
alpha[1]       0.0350  1.581e-02      2.215  2.678e-02 [4.027e-03,6.601e-02]
alpha[2]       0.0581  3.943e-02      1.473      0.141  [-1.919e-02,  0.135]
beta[1]        0.8675      0.535      1.622      0.105     [ -0.181,  1.916]
beta[2]        0.0179      0.495  3.618e-02      0.971     [ -0.952,  0.987]
============================================================================

Covariance estimator: robust
```

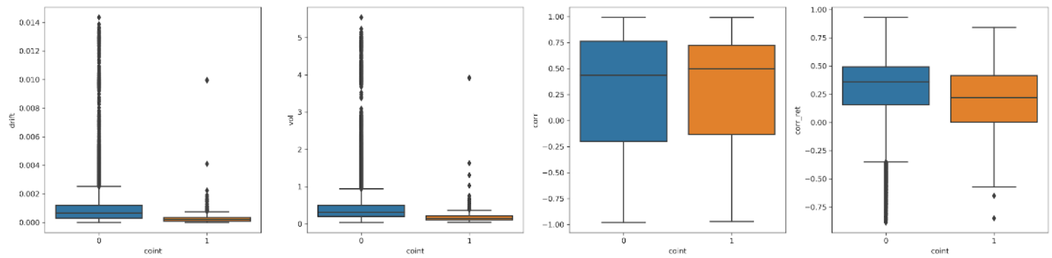

Univariate Time Series
ARMA Models
with exogenous variables
Multivariate Time Series
Vector Autoregressive (VAR) Models


Industrial Production
Consumer Sentiment
DATE

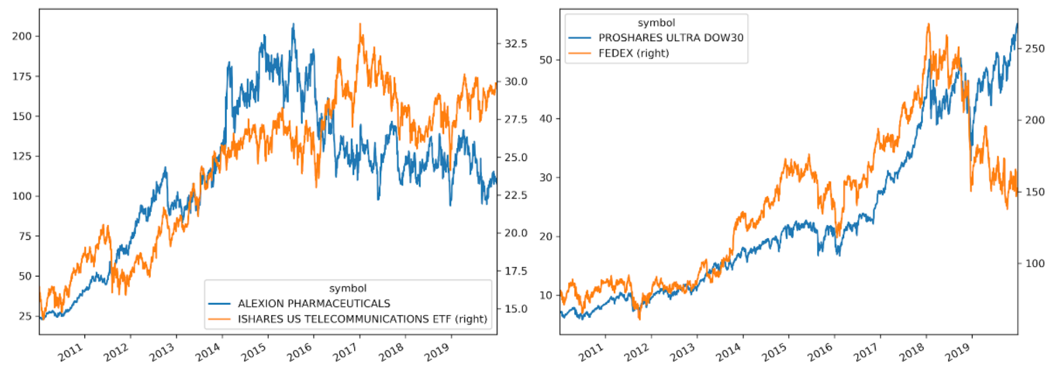
```
                           Statespace Model Results
==============================================================================
Dep. Variable:     ['ip', 'sentiment']   No. Observations:                  468
Model:                       VARMA(1,1)   Log Likelihood                 -71.870
                            + intercept   AIC                            169.741
Date:                  Thu, 16 Apr 2020   BIC                            223.671
Time:                          22:55:23   HQIC                           190.962
Sample:                               0
                                  - 468
Covariance Type:                    opg
==================================================================================
Ljung-Box (Q):            127.93, 161.51   Jarque-Bera (JB):       128.70, 17.04
Prob(Q):                      0.00, 0.00   Prob(JB):                  0.00, 0.00
Heteroskedasticity (H):       0.48, 1.10   Skew:                      0.19, 0.21
Prob(H) (two-sided):          0.00, 0.57   Kurtosis:                  5.54, 3.83
                            Results for equation ip
==================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
intercept           0.0015      0.001      2.401      0.016       0.000       0.003
L1.ip               0.9284      0.010     93.628      0.000       0.909       0.948
L1.sentiment        0.0006   6.03e-05     10.059      0.000       0.000       0.001
L1.e(ip)            0.0116      0.037      0.311      0.756      -0.062       0.085
L1.e(sentiment) -9.925e-05      0.000     -0.814      0.415      -0.000       0.000
                         Results for equation sentiment
==================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
intercept           0.3374      0.279      1.208      0.227      -0.210       0.885
L1.ip             -14.3677      5.450     -2.636      0.008     -25.049      -3.687
L1.sentiment        0.8801      0.023     37.598      0.000       0.834       0.926
L1.e(ip)           39.6834     18.798      2.111      0.035       2.839      76.528
L1.e(sentiment)     0.0509      0.052      0.983      0.326      -0.051       0.152
                              Error covariance matrix
==========================================================================================
                            coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
sqrt.var.ip               0.0129      0.000     40.298      0.000       0.012       0.014
sqrt.cov.ip.sentiment     0.0368      0.231      0.159      0.873      -0.416       0.489
sqrt.var.sentiment        5.2738      0.148     35.519      0.000       4.983       5.565
==========================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```


Industrial Production - Diagnostics
Standardized residual
Histogram plus estimated density
KDE
N(0,1)
Hist
Normal Q-Q
Sample Quantiles
Theoretical Quantiles
Correlogram

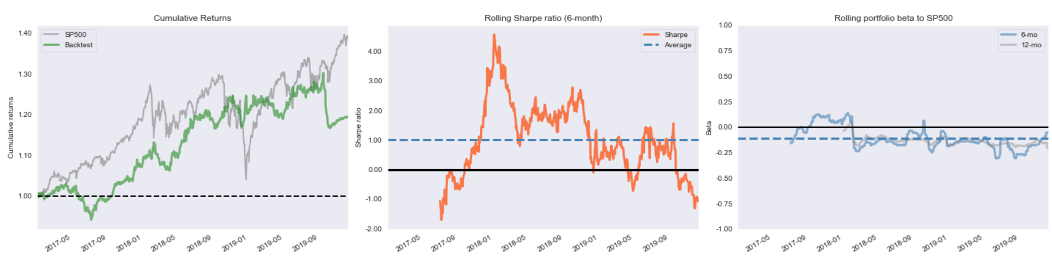

Industrial Production
actual
predicted
DATE
Sentiment
DATE



drift
vol
corr
corr_ret
coint
0
1

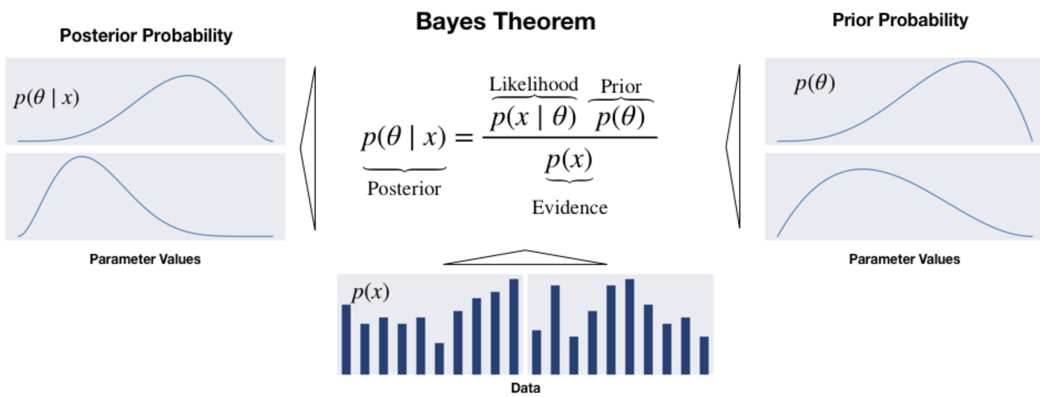


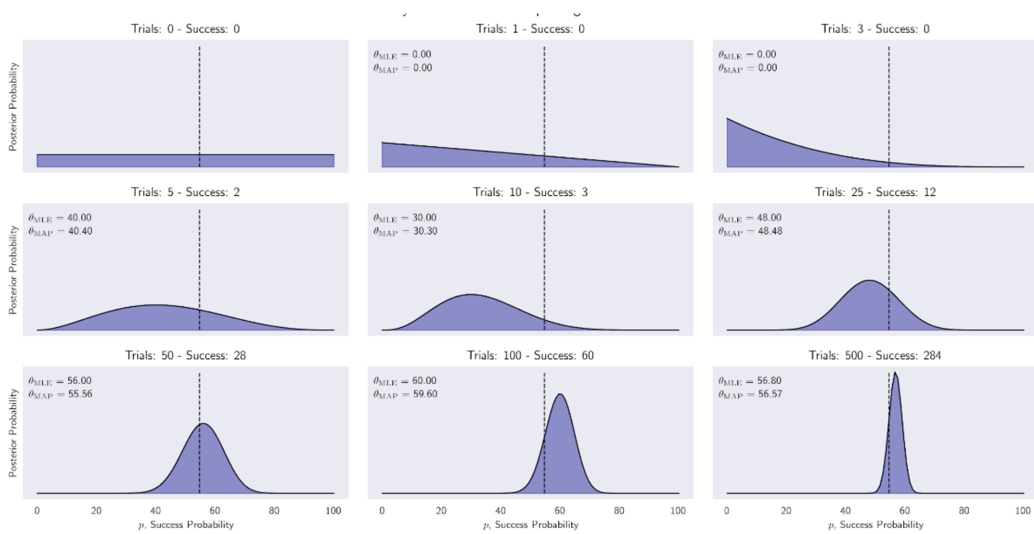
symbol
ALEXION PHARMACEUTICALS
ISHARES US TELECOMMUNICATIONS ETF (right)
symbol
PROSHARES ULTRA DOW30
FEDEX (right)
2011
2012
2013
2014
2015
2016
2017
2018
2019

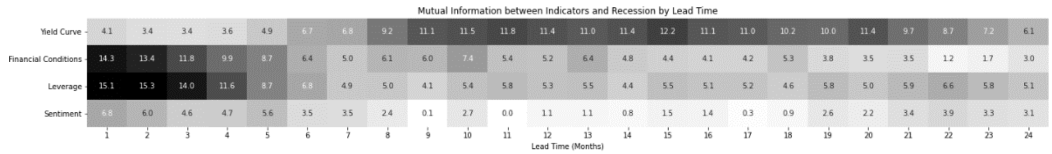


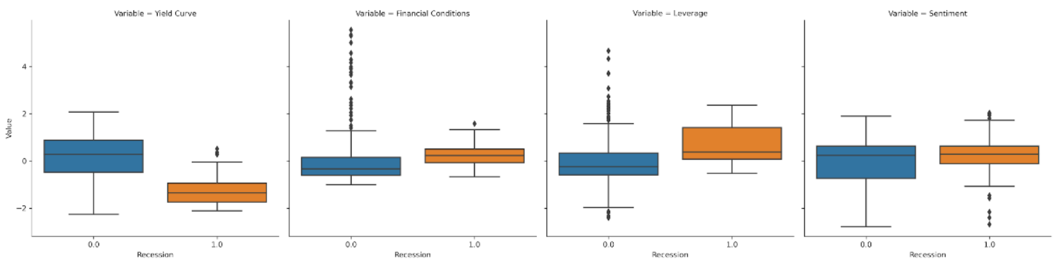
Cumulative Returns
SP500
Backtest
Cumulative returns
index
Rolling Sharpe ratio (6-month)
Sharpe
Average
Sharpe ratio
Rolling portfolio beta to SP500
6-mo
12-mo
Beta

Chapter 10: Bayesian ML – Dynamic Sharpe Ratios and Pairs Trading









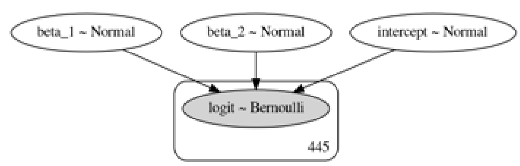
Logistic Regression

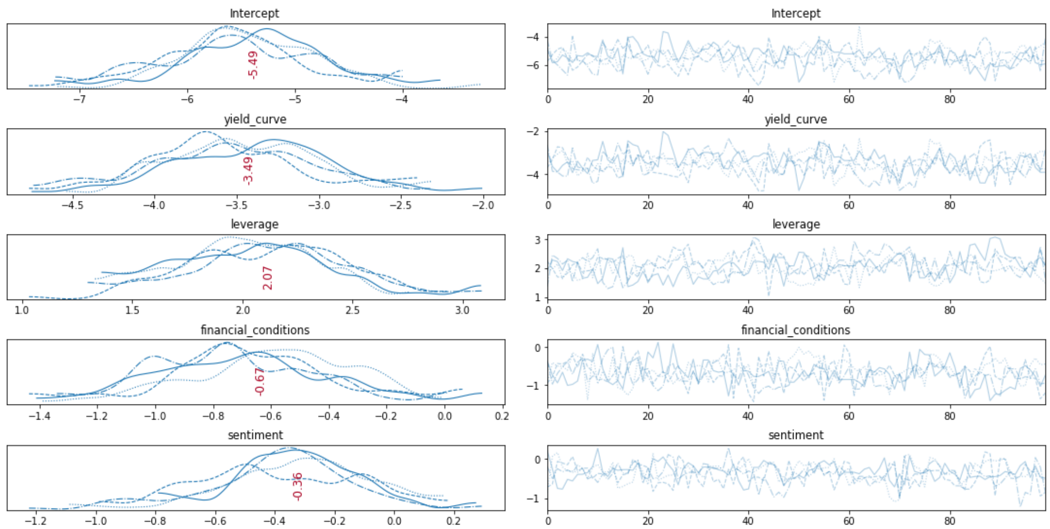
$$p(y_i = 1 \mid \boldsymbol{\beta}) = \sigma(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})$$

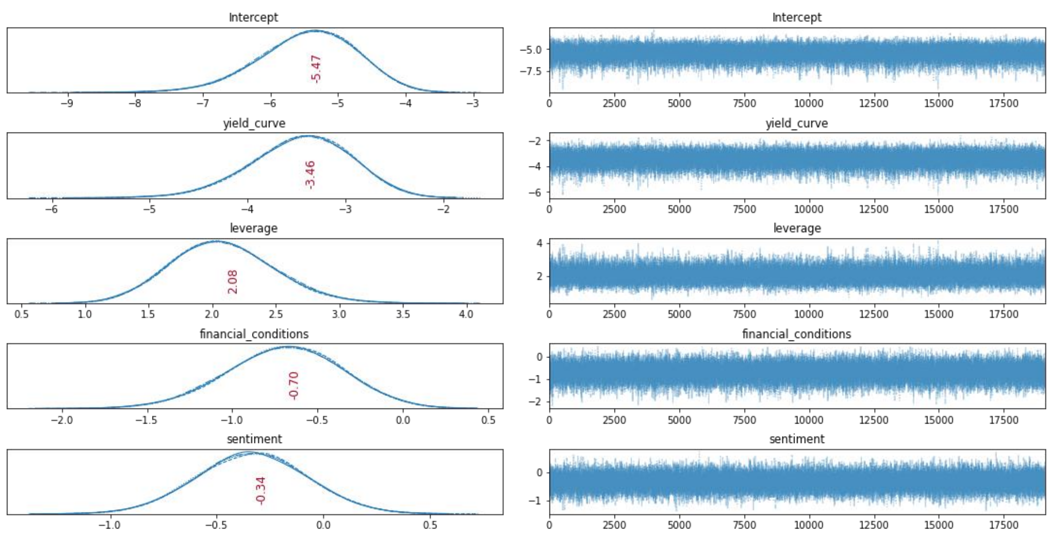
where σ is the logistic function

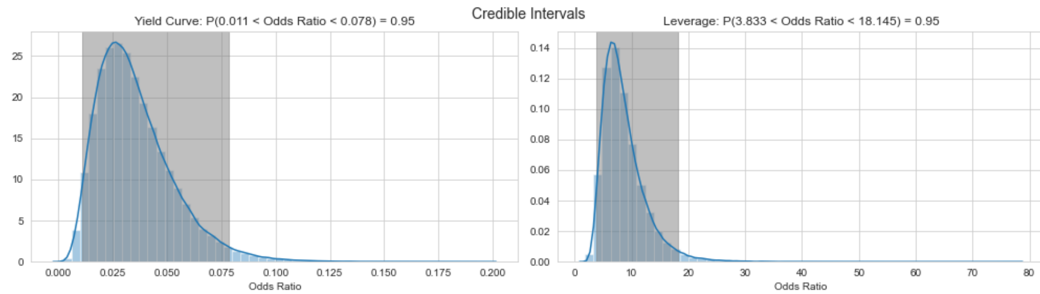
$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

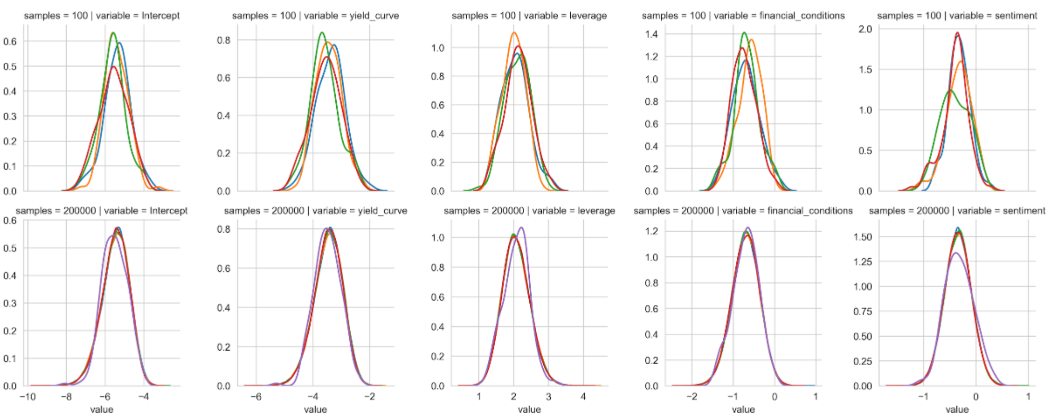
Plate Notation

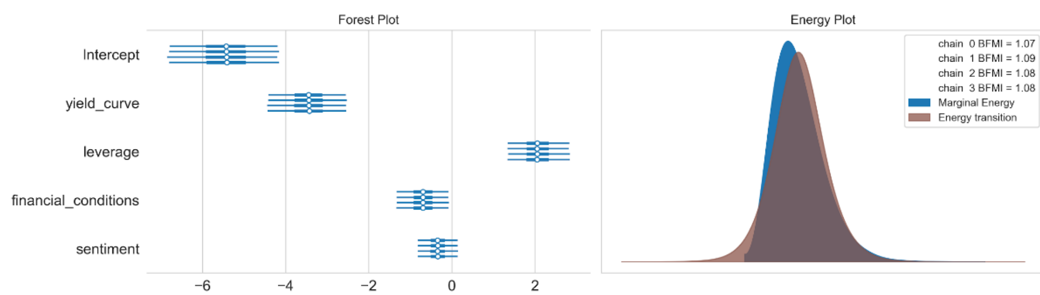


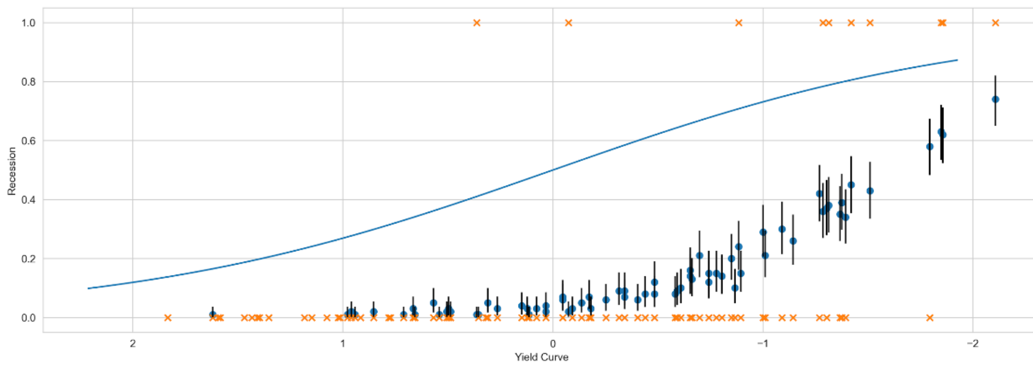

Intercept
-5.49
yield_curve
-3.49
leverage
2.07
financial_conditions
-0.67
sentiment
-0.36


Intercept
-5.47
yield_curve
-3.46
leverage
2.08
financial_conditions
-0.70
sentiment
-0.34

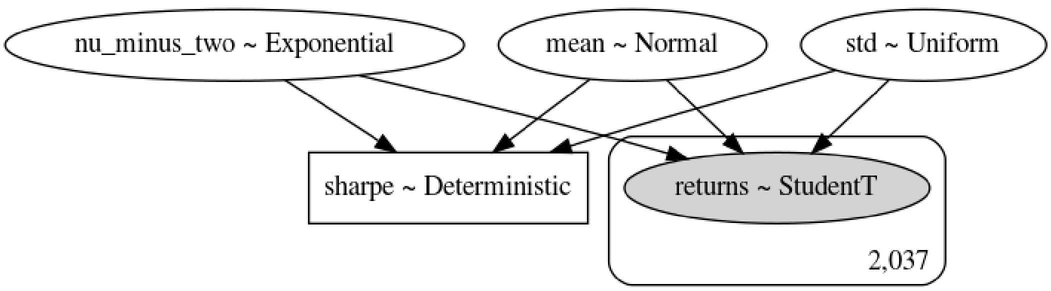

Credible Intervals
Yield Curve: P(0.011 < Odds Ratio < 0.078) = 0.95
Leverage: P(3.833 < Odds Ratio < 18.145) = 0.95
Odds Ratio
Odds Ratio


samples = 100 | variable = Intercept
samples = 100 | variable = yield_curve
samples = 100 | variable = leverage
samples = 100 | variable = financial_conditions
samples = 100 | variable = sentiment
samples = 200000 | variable = Intercept
samples = 200000 | variable = yield_curve
samples = 200000 | variable = leverage
samples = 200000 | variable = financial_conditions
samples = 200000 | variable = sentiment
value

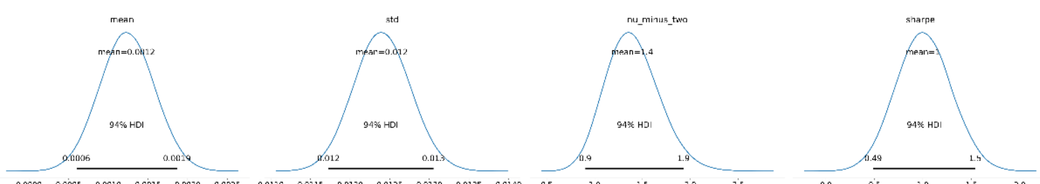

Forest Plot
Intercept
yield_curve
leverage
financial_conditions
sentiment
Energy Plot
chain 0 BFMI = 1.07
chain 1 BFMI = 1.09
chain 2 BFMI = 1.08
chain 3 BFMI = 1.08
Marginal Energy
Energy transition



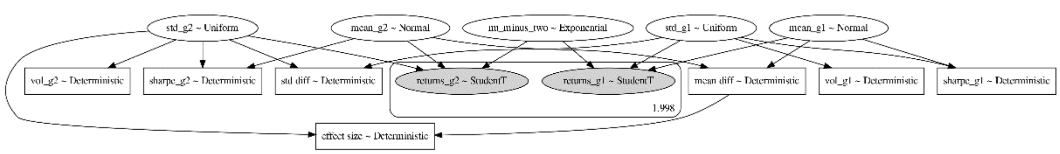
Recession
Yield Curve
1.0
0.8
0.6
0.4
0.2
0.0
2
1
0
−1
−2



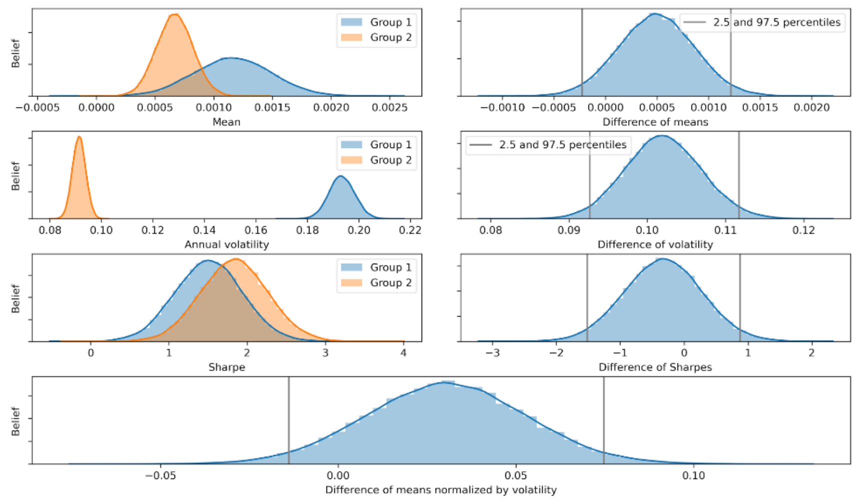
nu_minus_two ~ Exponential
mean ~ Normal
std ~ Uniform
sharpe ~ Deterministic
returns ~ StudentT
2,037

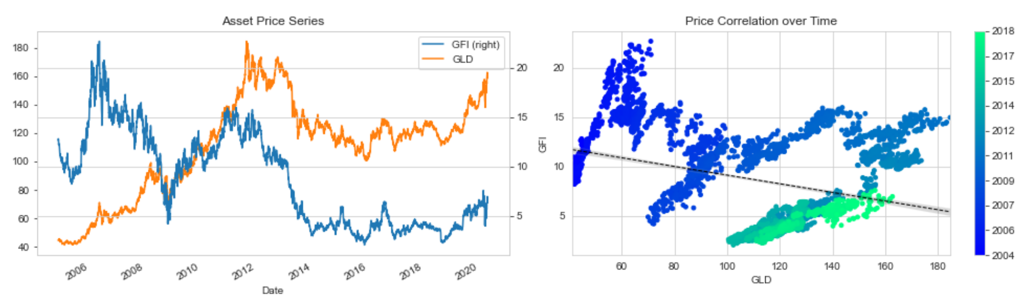


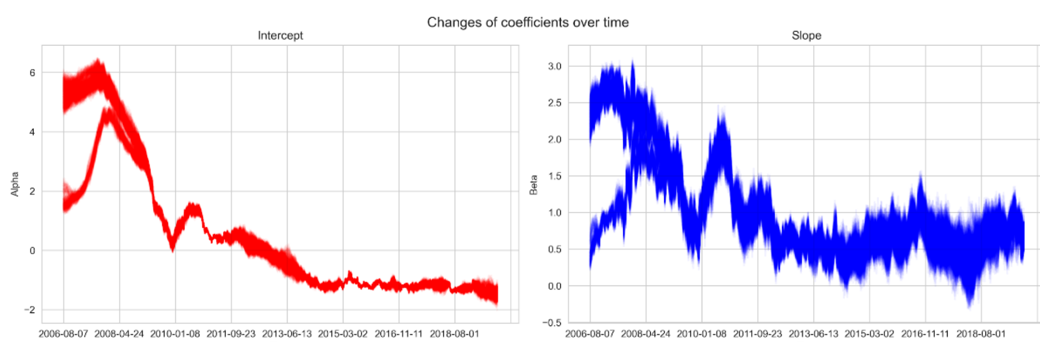
mean
std
nu_minus_two
sharpe
94% HDI

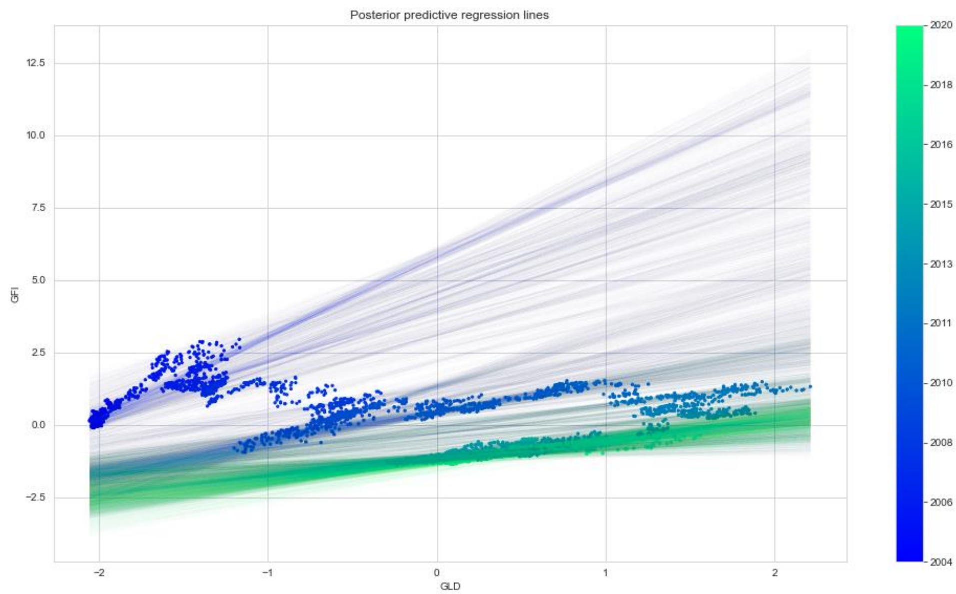


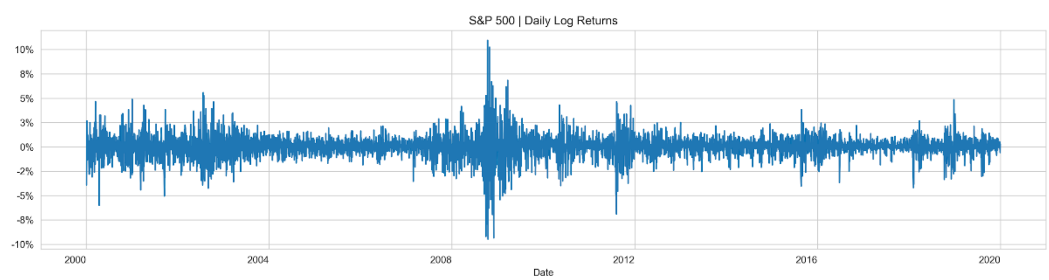
std_g2 ~ Uniform
mean_g2 ~ Normal
nu_minus_two ~ Exponential
std_g1 ~ Uniform
mean_g1 ~ Normal
vol_g2 ~ Deterministic
sharpe_g2 ~ Deterministic
std diff ~ Deterministic
returns_g2 ~ StudentT
returns_g1 ~ StudentT
mean diff ~ Deterministic
vol_g1 ~ Deterministic
sharpe_g1 ~ Deterministic
1,998
effect size ~ Deterministic

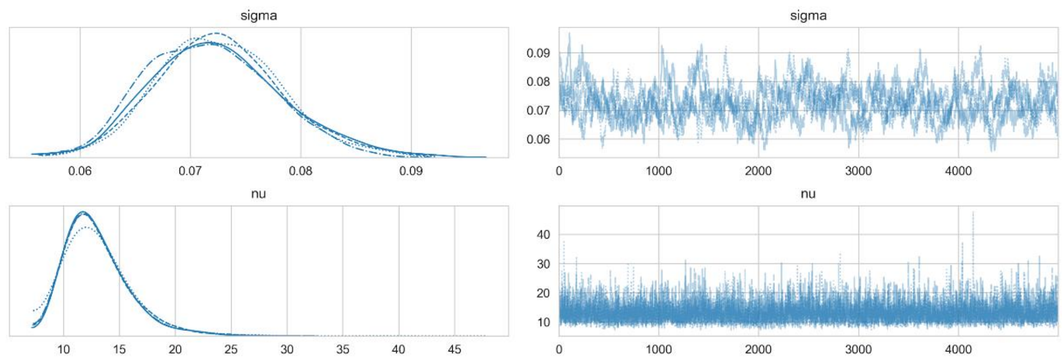

Belief
Group 1
Group 2
Mean
2.5 and 97.5 percentiles
Difference of means
Annual volatility
Difference of volatility
Sharpe
Difference of Sharpes
Difference of means normalized by volatility

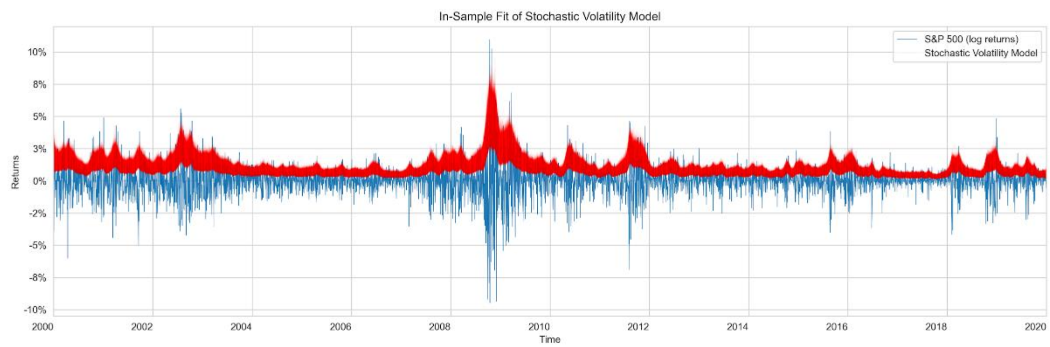

Asset Price Series
GFI (right)
GLD
Date
Price Correlation over Time
GFI
GLD


Changes of coefficients over time
Intercept
Alpha
Time
Slope
Beta
Time

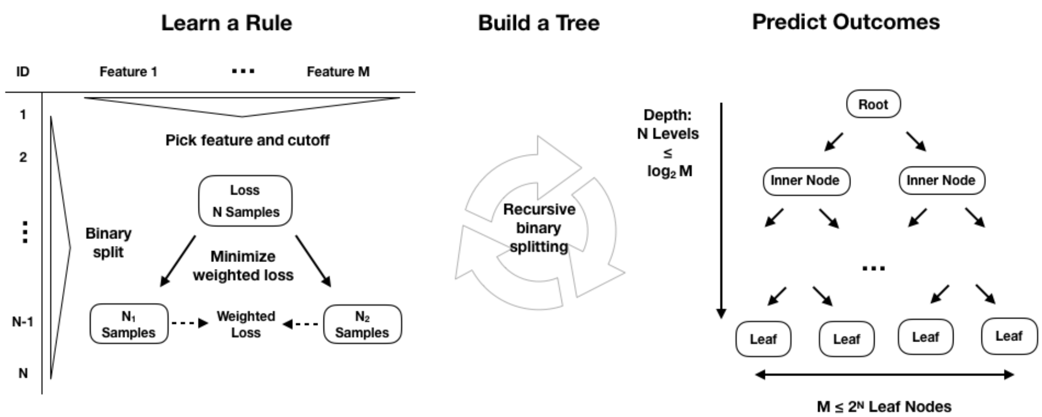

Posterior predictive regression lines
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
GFI
−2
−1
0
1
2
GLD
2020
2018
2016
2015
2013
2011
2010
2008
2006
2004

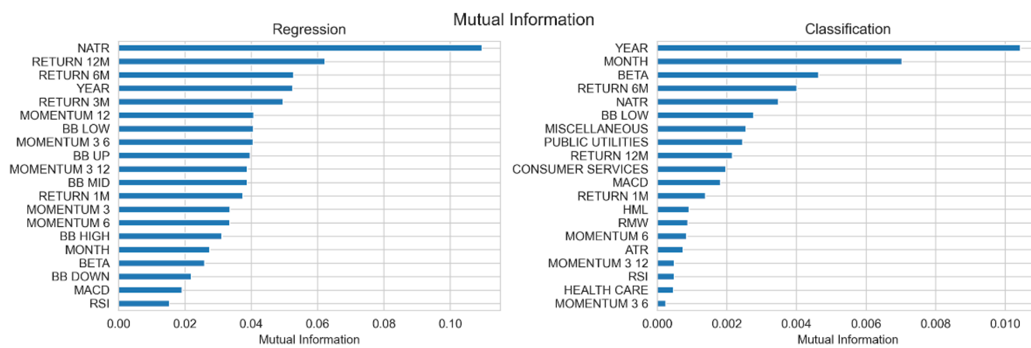

S&P 500 | Daily Log Returns
10%
8%
5%
3%
0%
-2%
-5%
-8%
-10%
2000
2004
2008
2012
2016
2020
Date


sigma
sigma
nu
nu

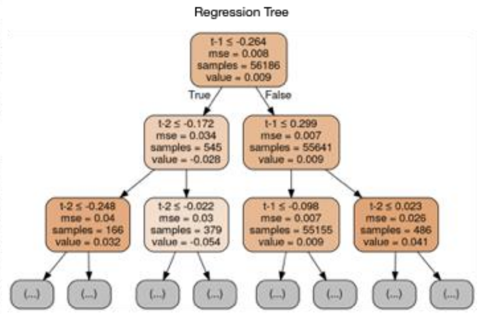

In-Sample Fit of Stochastic Volatility Model
S&P 500 (log returns)
Stochastic Volatility Model
Returns
Time

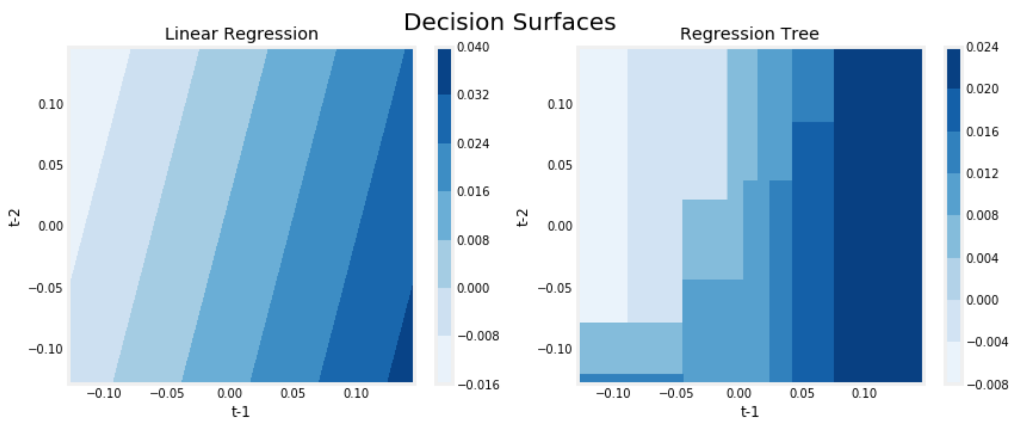
Chapter 11: Random Forests – A Long-Short Strategy for Japanese Stocks

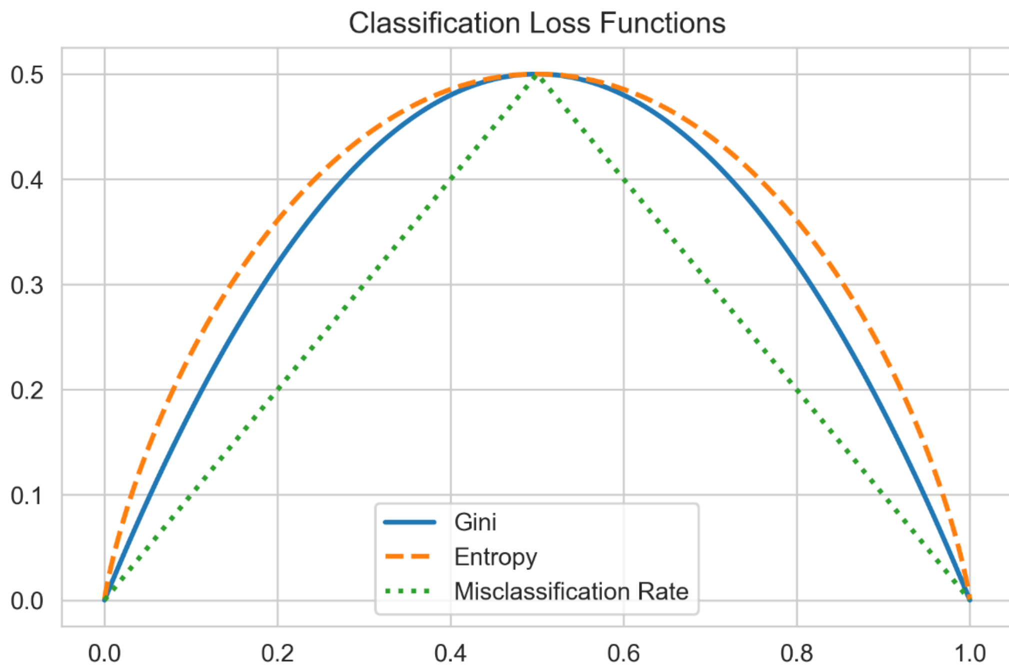


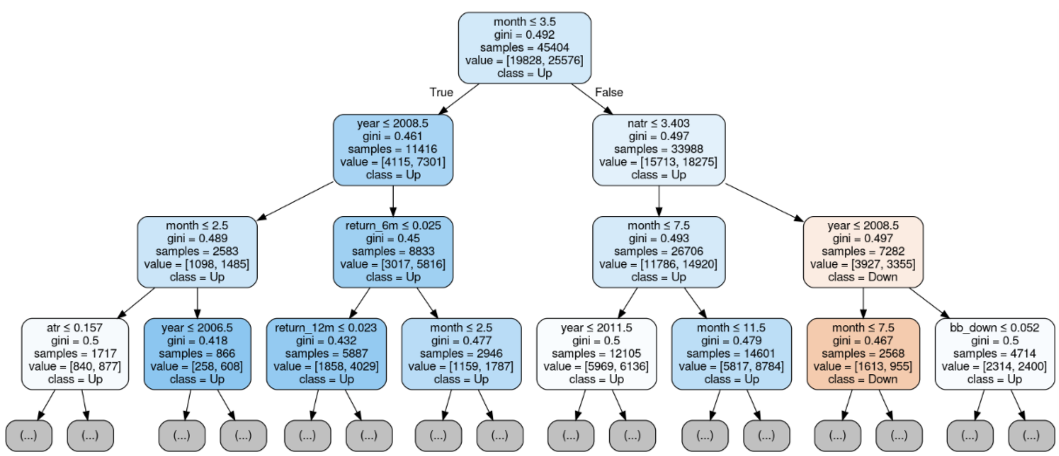


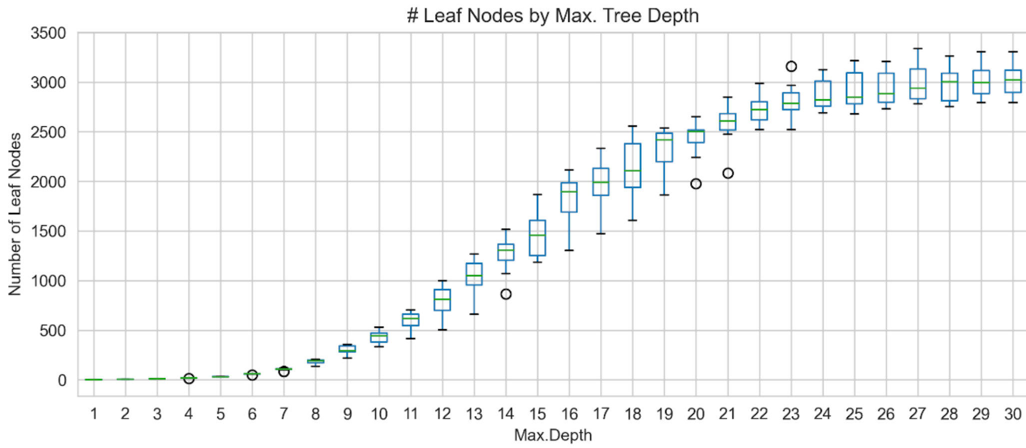
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     39.02
Date:                Wed, 22 Apr 2020   Prob (F-statistic):           1.17e-17
Time:                        19:18:07   Log-Likelihood:                 56967.
No. Observations:               56186   AIC:                        -1.139e+05
Df Residuals:                   56183   BIC:                        -1.139e+05
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0088      0.000     23.412      0.000       0.008       0.009
t-1            0.0327      0.004      7.761      0.000       0.024       0.041
t-2           -0.0187      0.004     -4.437      0.000      -0.027      -0.010
==============================================================================
Omnibus:                     2103.126   Durbin-Watson:                   1.999
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6483.607
Skew:                           0.018   Prob(JB):                         0.00
Kurtosis:                       4.664   Cond. No.                         11.5
==============================================================================
```






Classification Loss Functions
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Gini
Entropy
Misclassification Rate

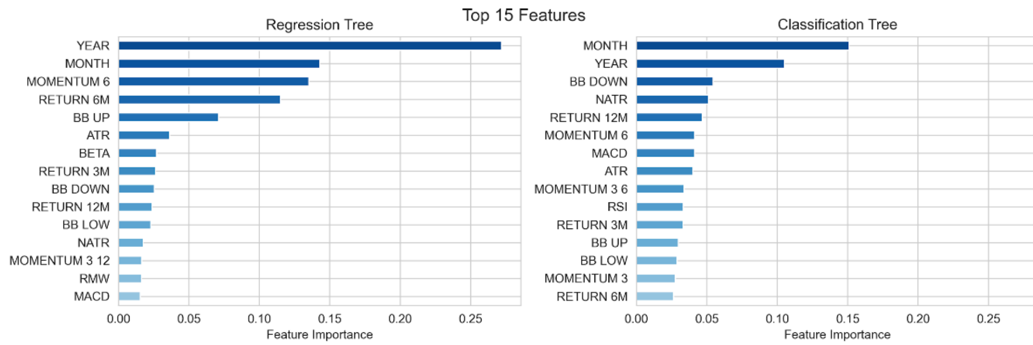

month ≤ 3.5
gini = 0.492
samples = 45404
value = [19828, 25576]
class = Up
True
False
year ≤ 2008.5
gini = 0.461
samples = 11416
value = [4115, 7301]
class = Up
natr ≤ 3.403
gini = 0.497
samples = 33988
value = [15713, 18275]
class = Up
month ≤ 2.5
gini = 0.489
samples = 2583
value = [1098, 1485]
class = Up
return_6m ≤ 0.025
gini = 0.45
samples = 8833
value = [3017, 5816]
class = Up
month ≤ 7.5
gini = 0.493
samples = 26706
value = [11786, 14920]
class = Up
year ≤ 2008.5
gini = 0.497
samples = 7282
value = [3927, 3355]
class = Down
atr ≤ 0.157
gini = 0.5
samples = 1717
value = [840, 877]
class = Up
year ≤ 2006.5
gini = 0.418
samples = 866
value = [258, 608]
class = Up
return_12m ≤ 0.023
gini = 0.432
samples = 5887
value = [1858, 4029]
class = Up
month ≤ 2.5
gini = 0.477
samples = 2946
value = [1159, 1787]
class = Up
year ≤ 2011.5
gini = 0.5
samples = 12105
value = [5969, 6136]
class = Up
month ≤ 11.5
gini = 0.479
samples = 14601
value = [5817, 8784]
class = Up
month ≤ 7.5
gini = 0.467
samples = 2568
value = [1613, 955]
class = Down
bb_down ≤ 0.052
gini = 0.5
samples = 4714
value = [2314, 2400]
class = Up
(...)

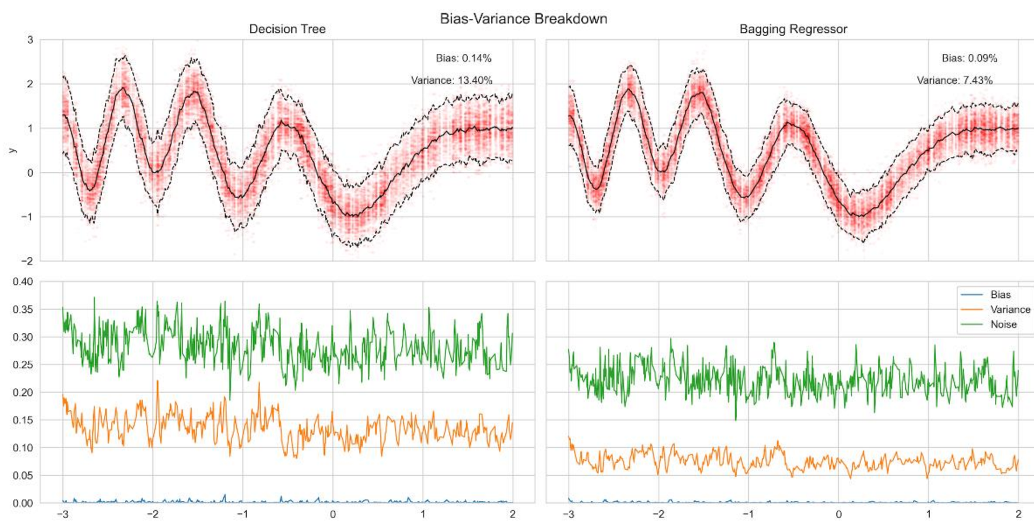


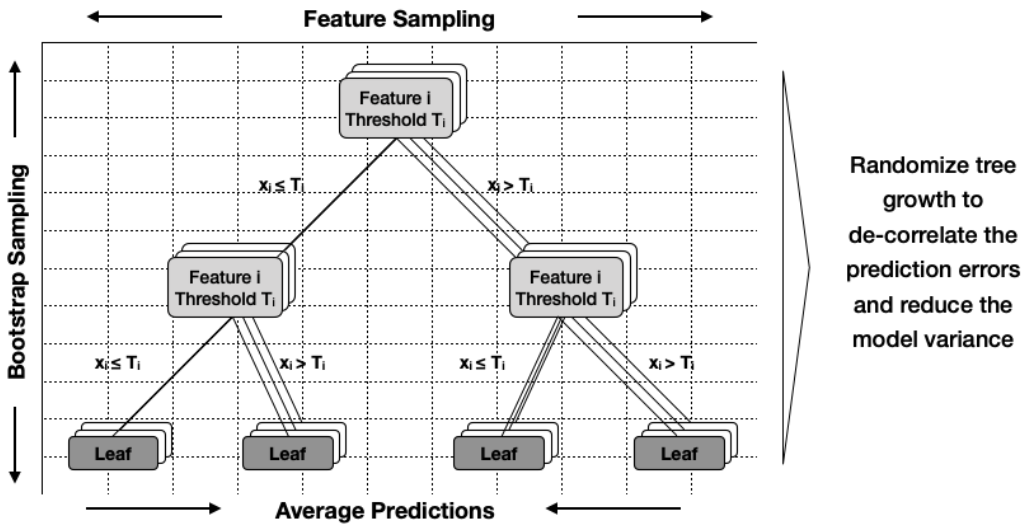
Leaf Nodes by Max. Tree Depth
Number of Leaf Nodes
Max.Depth

Train-Validation Scores
Regression Tree
Classification Tree
Data
Train
Valid
IC
ROC AUC
Max. Depth

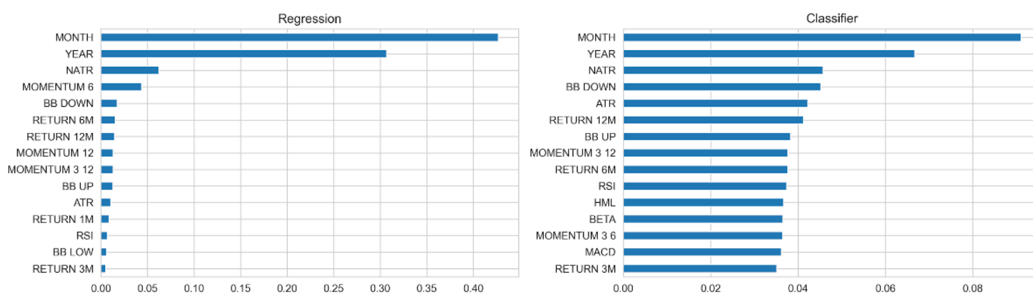
Learning Curves
Best Regression Tree
Best Classification Tree
Data
Train
Valid
IC
ROC AUC
Train Size


Top 15 Features
Regression Tree
YEAR
MONTH
MOMENTUM 6
RETURN 6M
BB UP
ATR
BETA
RETURN 3M
BB DOWN
RETURN 12M
BB LOW
NATR
MOMENTUM 3 12
RMW
MACD
Feature Importance
Classification Tree
MONTH
YEAR
BB DOWN
NATR
RETURN 12M
MOMENTUM 6
MACD
ATR
MOMENTUM 3 6
RSI
RETURN 3M
BB UP
BB LOW
MOMENTUM 3
RETURN 6M
Feature Importance

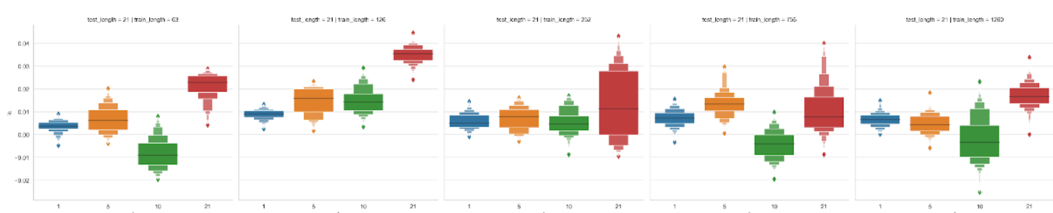

Bias-Variance Breakdown
Decision Tree
Bias: 0.14%
Variance: 13.40%
Bagging Regressor
Bias: 0.09%
Variance: 7.43%
y
Bias
Variance
Noise

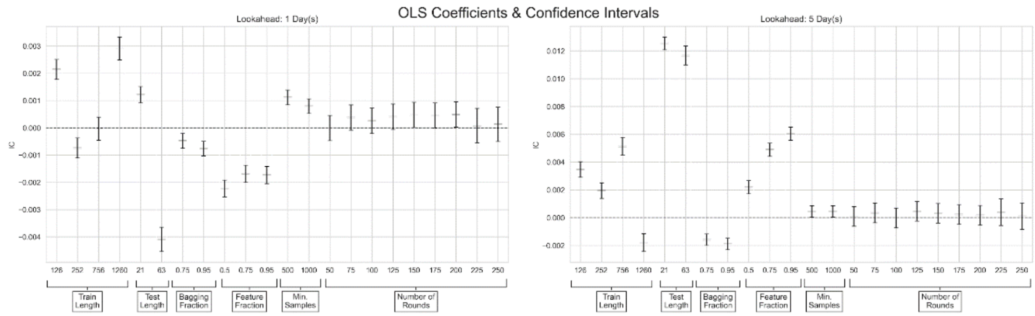


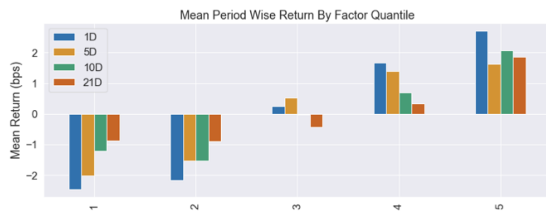
Feature Sampling
Bootstrap Sampling
Feature i
Threshold T_i
$x_i \leq T_i$
$x_i > T_i$
Leaf
Average Predictions
Randomize tree growth to de-correlate the prediction errors and reduce the model variance

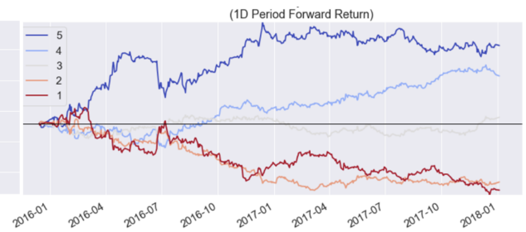


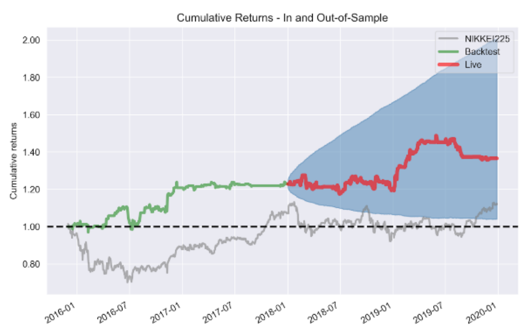
Regression
MONTH
YEAR
NATR
MOMENTUM 6
BB DOWN
RETURN 6M
RETURN 12M
MOMENTUM 12
MOMENTUM 3 12
BB UP
ATR
RETURN 1M
RSI
BB LOW
RETURN 3M
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40
Classifier
MONTH
YEAR
NATR
BB DOWN
ATR
RETURN 12M
BB UP
MOMENTUM 3 12
RETURN 6M
RSI
HML
BETA
MOMENTUM 3 6
MACD
RETURN 3M
0.00 0.02 0.04 0.06 0.08

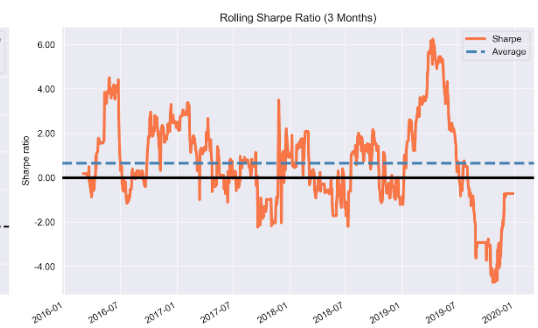



OLS Coefficients & Confidence Intervals
Lookahead: 1 Day(s)
Lookahead: 5 Day(s)
IC
Train Length
Test Length
Bagging Fraction
Feature Fraction
Min. Samples
Number of Rounds

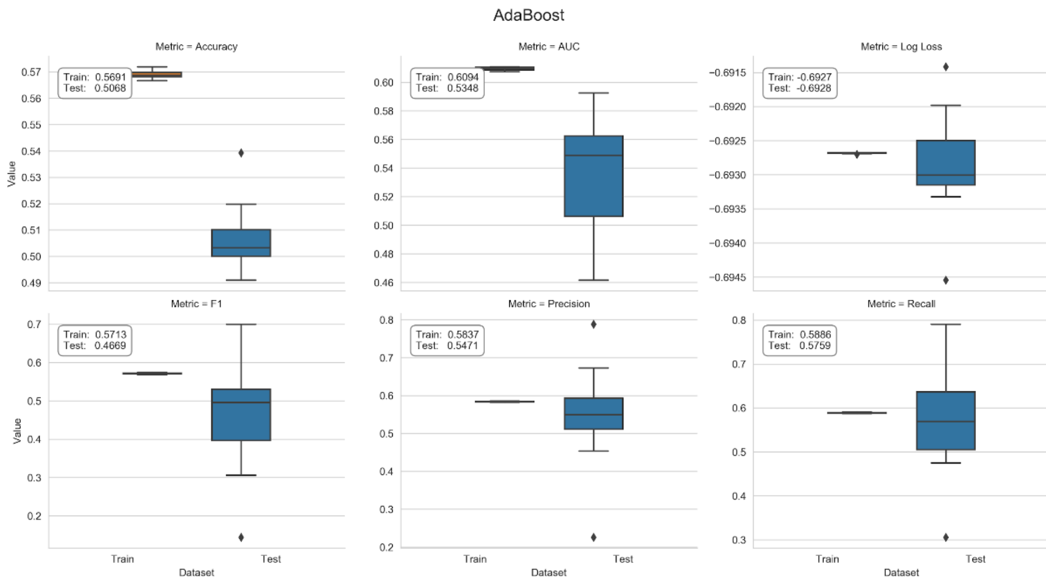

Mean Period Wise Return By Factor Quantile
Mean Return (bps)
1D
5D
10D
21D

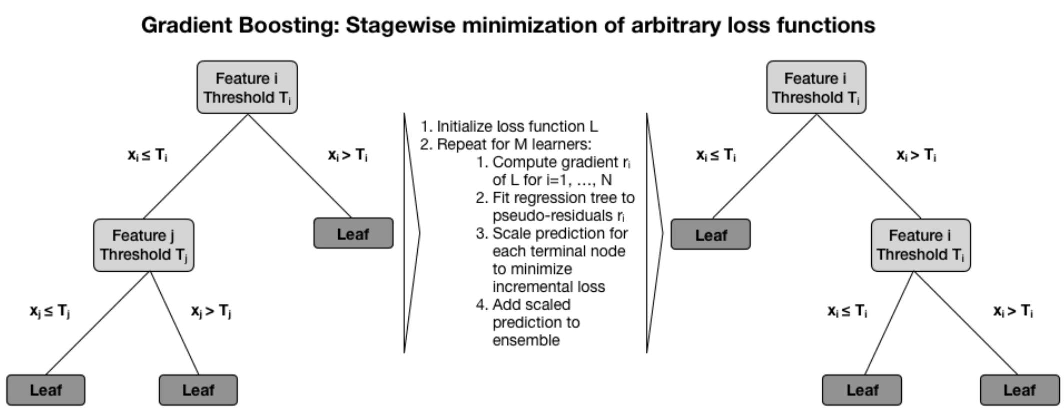

(1D Period Forward Return)

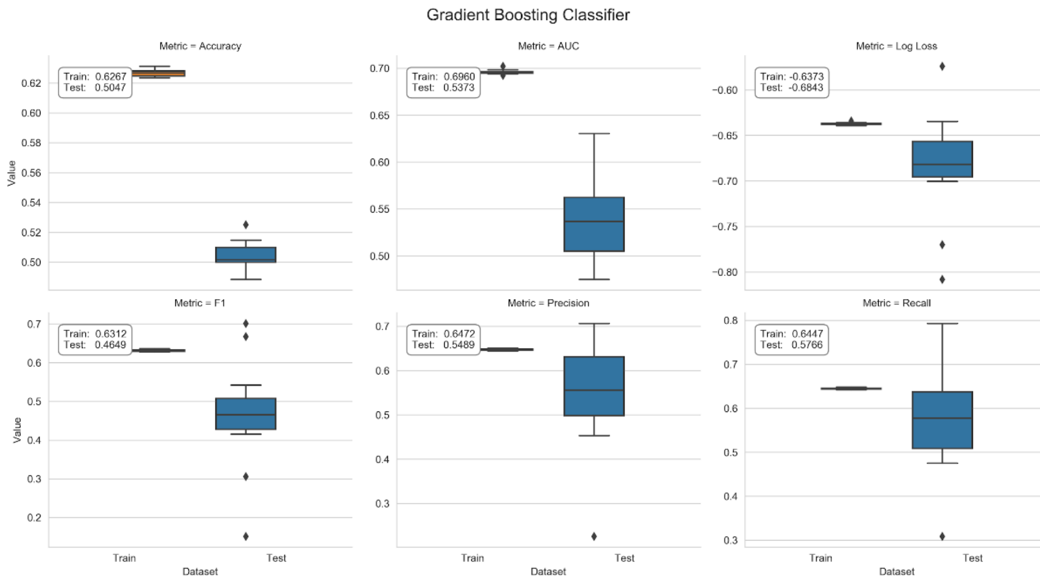

Cumulative Returns - In and Out-of-Sample
Cumulative returns
NIKKEI225
Backtest
Live

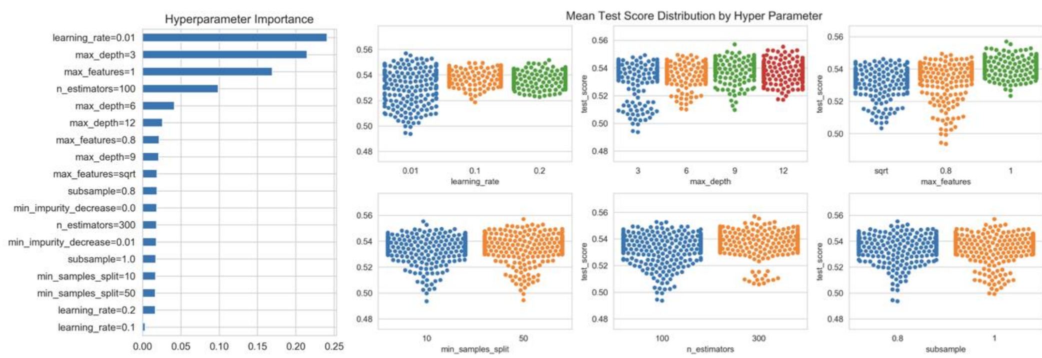

Rolling Sharpe Ratio (3 Months)
Sharpe ratio
Sharpe
Average

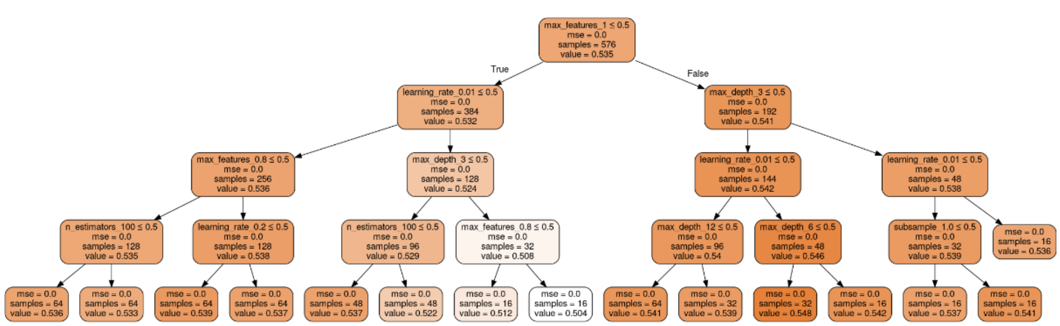
Chapter 12: Boosting Your Trading Strategy

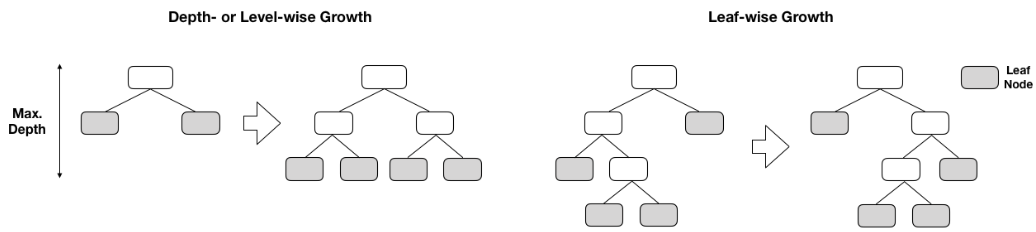


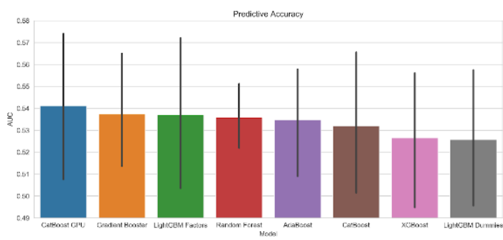


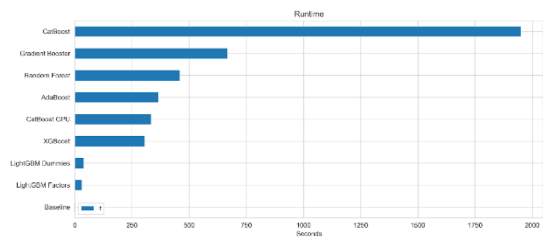

Gradient Boosting Classifier
Metric = Accuracy
Train: 0.6267
Test: 0.5047
Metric = AUC
Train: 0.6960
Test: 0.5373
Metric = Log Loss
Train: -0.6373
Test: -0.6843
Metric = F1
Train: 0.6312
Test: 0.4649
Metric = Precision
Train: 0.6472
Test: 0.5489
Metric = Recall
Train: 0.6447
Test: 0.5766
Value
Train
Test
Dataset

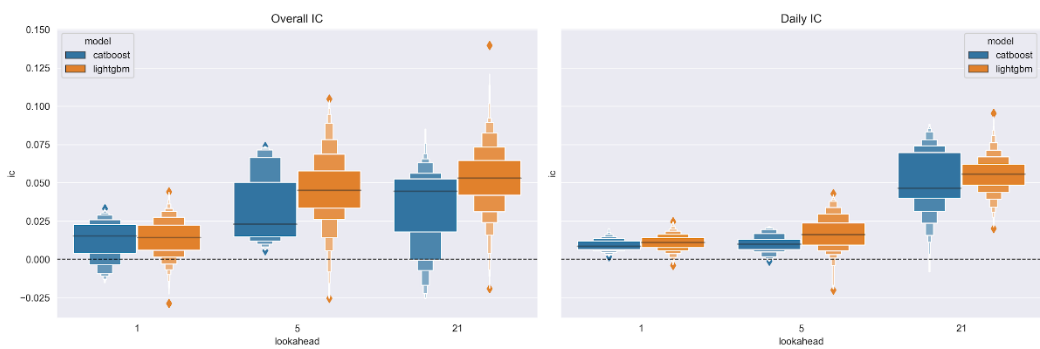

Hyperparameter Importance
learning_rate=0.01
max_depth=3
max_features=1
n_estimators=100
max_depth=6
max_depth=12
max_features=0.8
max_depth=9
max_features=sqrt
subsample=0.8
min_impurity_decrease=0.0
n_estimators=300
min_impurity_decrease=0.01
subsample=1.0
min_samples_split=10
min_samples_split=50
learning_rate=0.2
learning_rate=0.1
Mean Test Score Distribution by Hyper Parameter
test_score
learning_rate
max_depth
max_features
min_samples_split
n_estimators
subsample

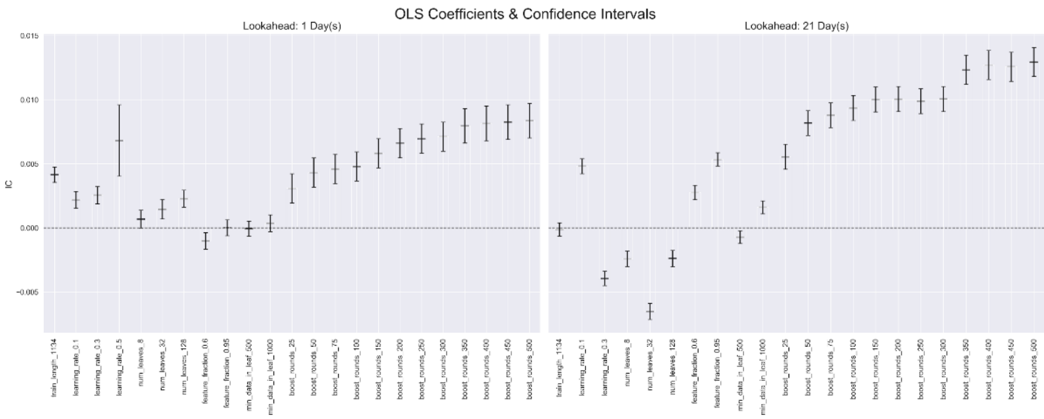

max_features_1 ≤ 0.5
mse = 0.0
samples = 576
value = 0.535
True
False
learning_rate_0.01 ≤ 0.5
samples = 384
value = 0.532
max_depth_3 ≤ 0.5
samples = 192
value = 0.541
max_features_0.8 ≤ 0.5
samples = 256
value = 0.536
max_depth_3 ≤ 0.5
samples = 128
value = 0.524
learning_rate_0.01 ≤ 0.5
samples = 144
value = 0.542
learning_rate_0.01 ≤ 0.5
samples = 48
value = 0.538
n_estimators_100 ≤ 0.5
samples = 128
value = 0.535
learning_rate_0.2 ≤ 0.5
samples = 128
value = 0.538
n_estimators_100 ≤ 0.5
samples = 96
value = 0.529
max_features_0.8 ≤ 0.5
samples = 32
value = 0.508
max_depth_12 ≤ 0.5
samples = 96
value = 0.54
max_depth_6 ≤ 0.5
samples = 48
value = 0.546
subsample_1.0 ≤ 0.5
samples = 32
value = 0.539
samples = 16
value = 0.536

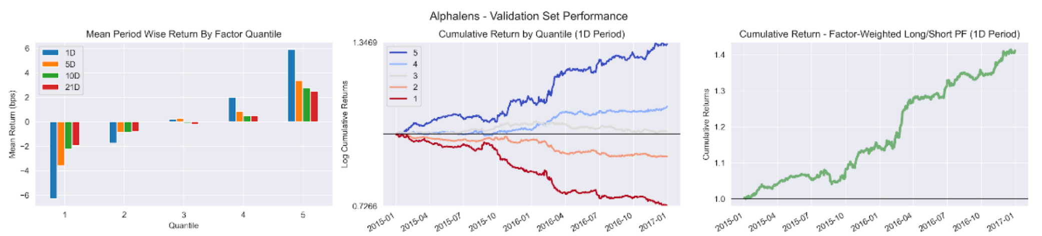

Depth- or Level-wise Growth
Leaf-wise Growth
Max. Depth
Leaf Node

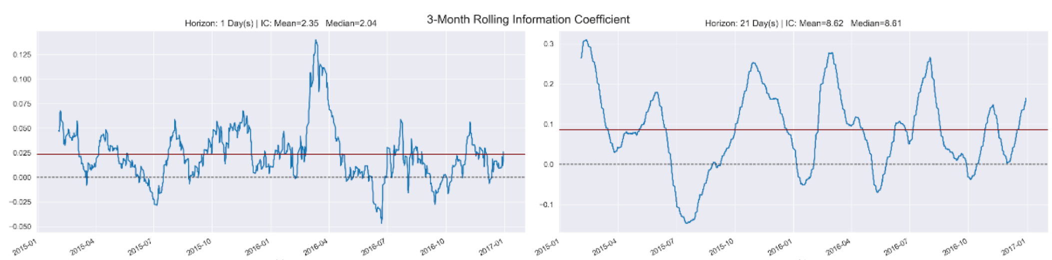

Predictive Accuracy
AUC
0.58
0.57
0.56
0.55
0.54
0.53
0.52
0.51
0.50
0.49
CatBoost GPU
Gradient Booster
LightGBM Factors
Random Forest
AdaBoost
CatBoost
XGBoost
LightGBM Dummies
Model

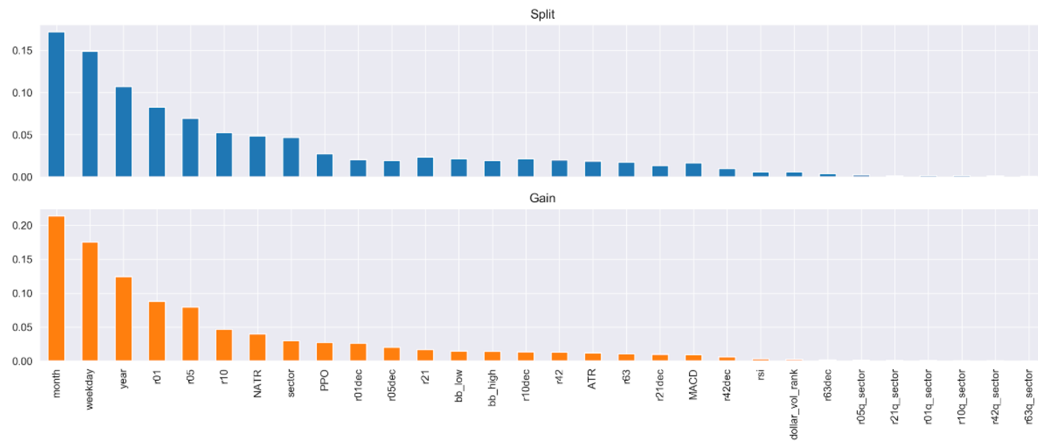

Runtime
CatBoost
Gradient Booster
Random Forest
AdaBoost
CatBoost GPU
XGBoost
LightGBM Dummies
LightGBM Factors
Baseline
t
0
250
500
750
1000
1250
1500
1750
2000
Seconds

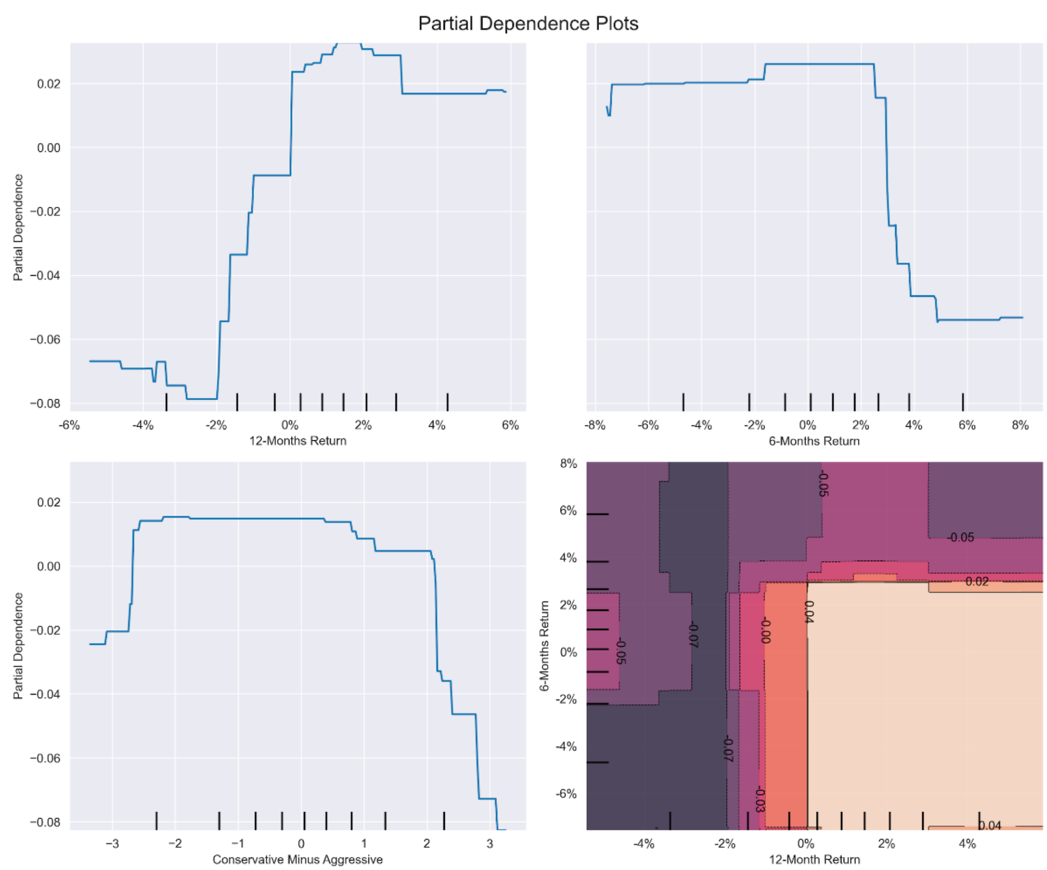

Overall IC
Daily IC
model
catboost
lightgbm
ic
0.150
0.125
0.100
0.075
0.050
0.025
0.000
−0.025
1
5
21
lookahead

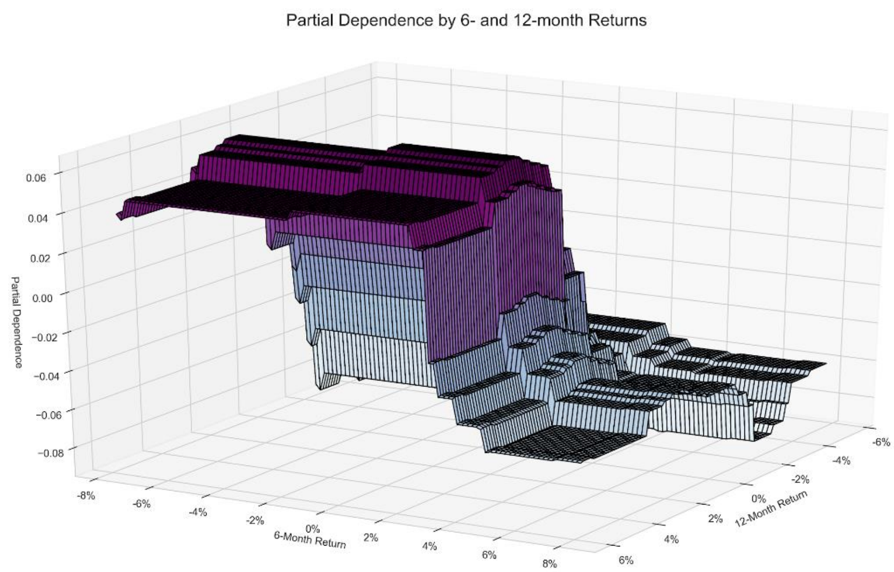

OLS Coefficients & Confidence Intervals
Lookahead: 1 Day(s)
Lookahead: 21 Day(s)
IC


Alphalens - Validation Set Performance
Mean Period Wise Return By Factor Quantile
Mean Return (bps)
Quantile
1D
5D
10D
21D
Cumulative Return by Quantile (1D Period)
Log Cumulative Returns
Cumulative Return - Factor-Weighted Long/Short PF (1D Period)
Cumulative Returns


3-Month Rolling Information Coefficient
Horizon: 1 Day(s) | IC: Mean=2.35 Median=2.04
Horizon: 21 Day(s) | IC: Mean=8.62 Median=8.61
date

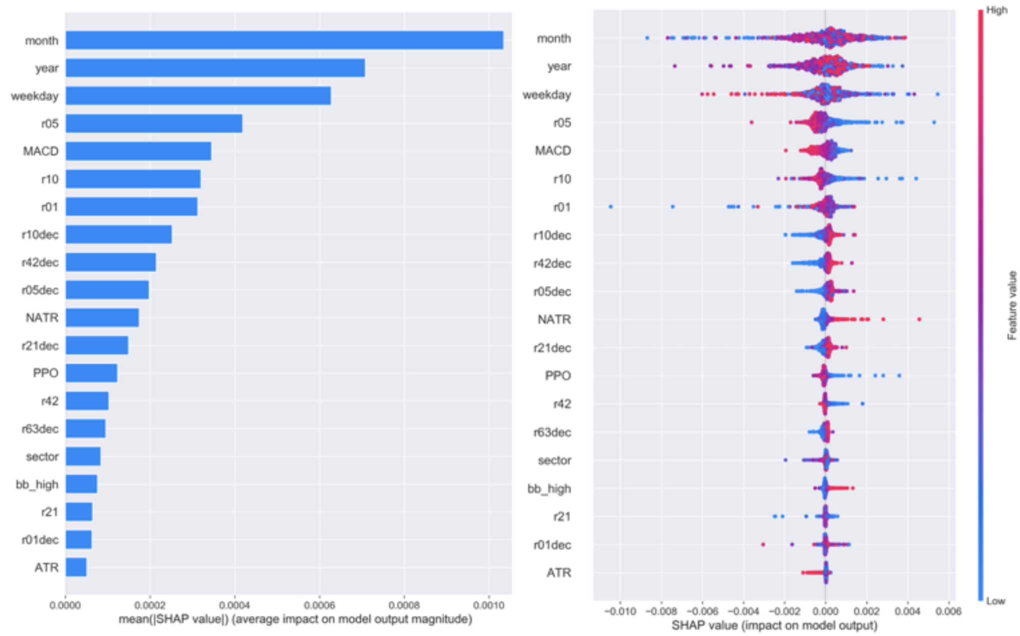

Split
Gain
month
weekday
year
r01
r05
r10
NATR
sector
PPO
r01dec
r05dec
r21
bb_low
bb_high
r10dec
r42
ATR
r63
r21dec
MACD
r42dec
rsi
dollar_vol_rank
r63dec
r05q_sector
r21q_sector
r01q_sector
r10q_sector
r42q_sector
r63q_sector

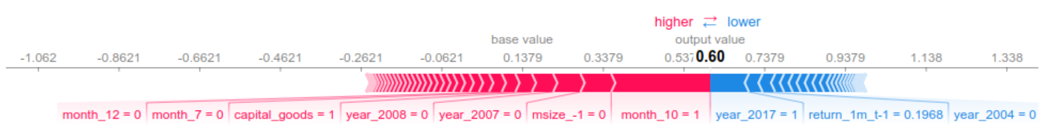

Partial Dependence Plots
Partial Dependence
12-Months Return
6-Months Return
Partial Dependence
Conservative Minus Aggressive
6-Months Return
12-Month Return



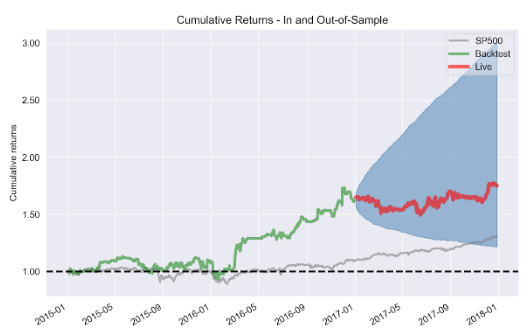
Partial Dependence by 6- and 12-month Returns
Partial Dependence
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
-8%
-6%
-4%
-2%
0%
2%
4%
6%
8%
6-Month Return
-6%
-4%
-2%
0%
2%
4%
6%
12-Month Return

SHAP Values: Impact on Output Magnitude

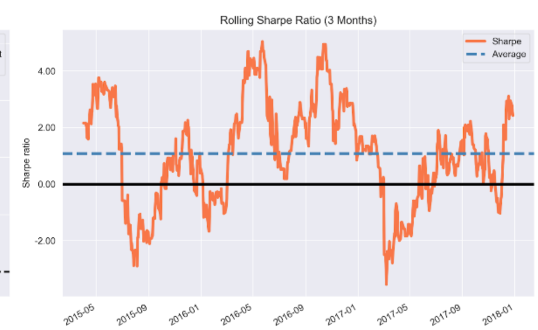




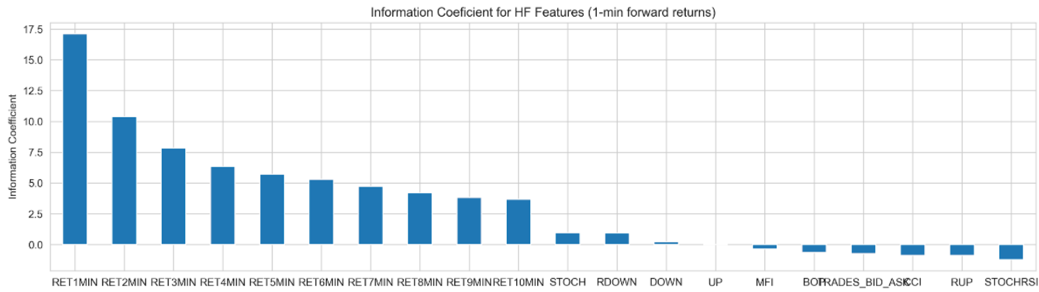
SHAP value for
r01
r01
r05
0.010
0.005
0.000
−0.005
−0.010
−0.015
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.04
0.02
0.00
−0.02
−0.04

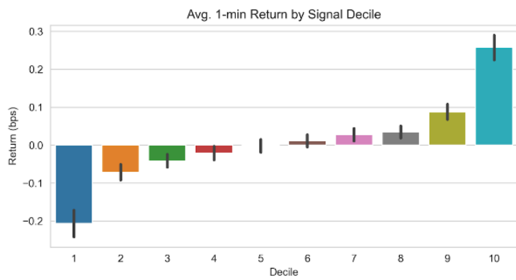


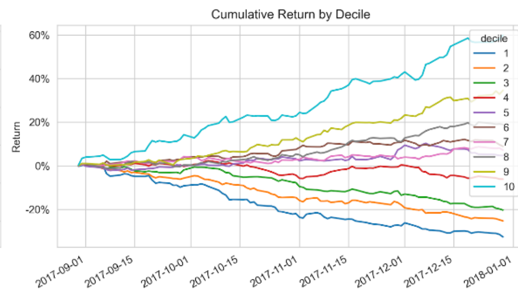
Cumulative Returns - In and Out-of-Sample
SP500
Backtest
Live
Cumulative returns



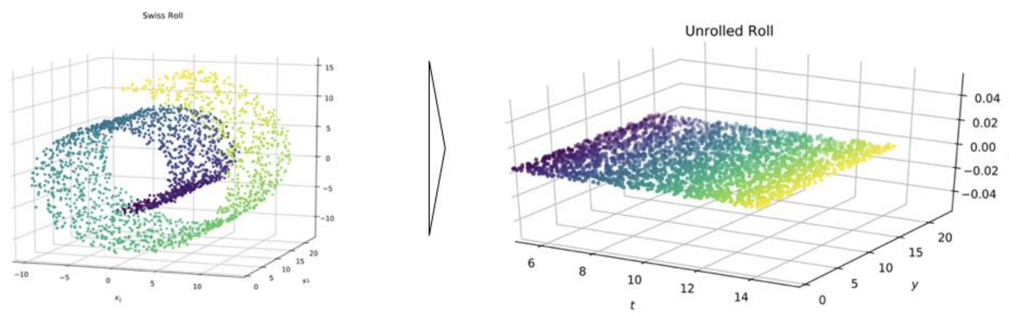
Rolling Sharpe Ratio (3 Months)
Sharpe
Average
Sharpe ratio

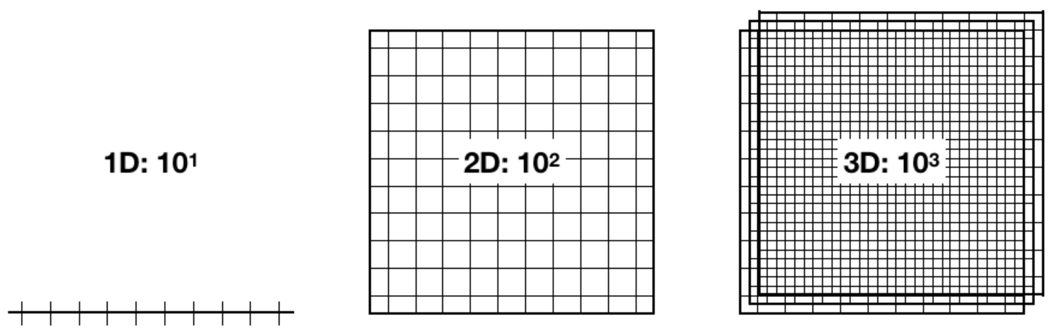

Information Coeficient for HF Features (1-min forward returns)
Information Coefficient
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
RET1MIN
RET2MIN
RET3MIN
RET4MIN
RET5MIN
RET6MIN
RET7MIN
RET8MIN
RET9MIN
RET10MIN
STOCH
RDOWN
DOWN
UP
MFI
BOP
TRADES_BID_ASK
CCI
RUP
STOCHRSI


Avg. 1-min Return by Signal Decile
Return (bps)
0.3
0.2
0.1
0.0
−0.1
−0.2
1
2
3
4
5
6
7
8
9
10
Decile

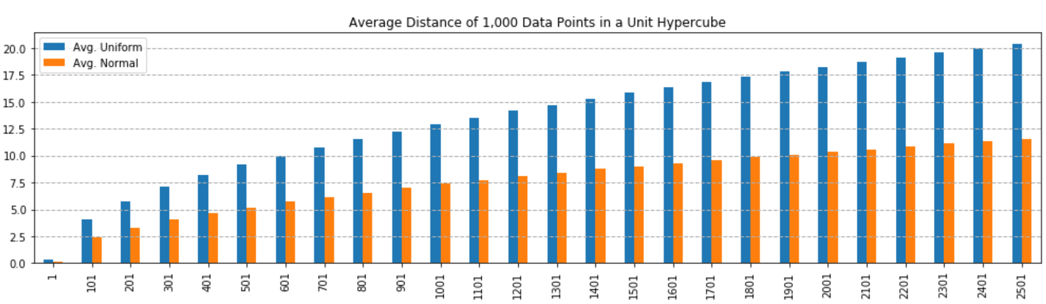

Cumulative Return by Decile
Return
60%
40%
20%
0%
-20%
decile
1
2
3
4
5
6
7
8
9
10
2017-09-01
2017-09-15
2017-10-01
2017-10-15
2017-11-01
2017-11-15
2017-12-01
2017-12-15
2018-01-01

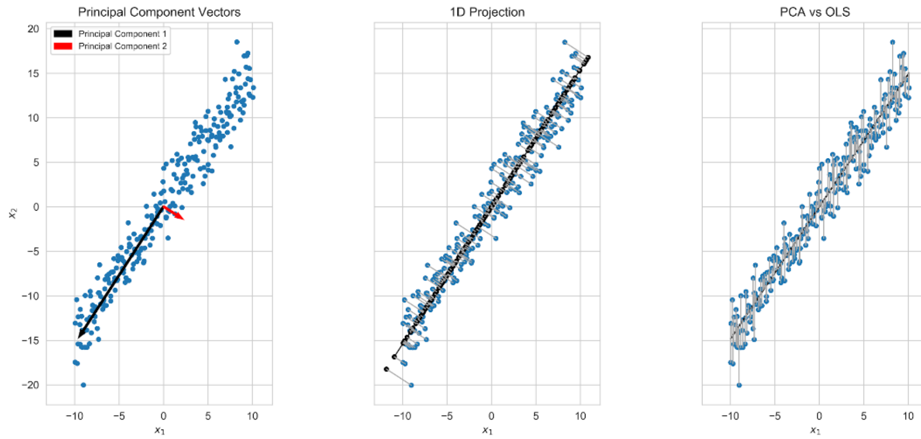
Chapter 13: Data-Driven Risk Factors and Asset Allocation with Unsupervised Learning



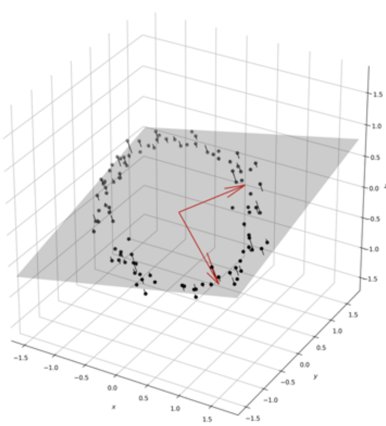


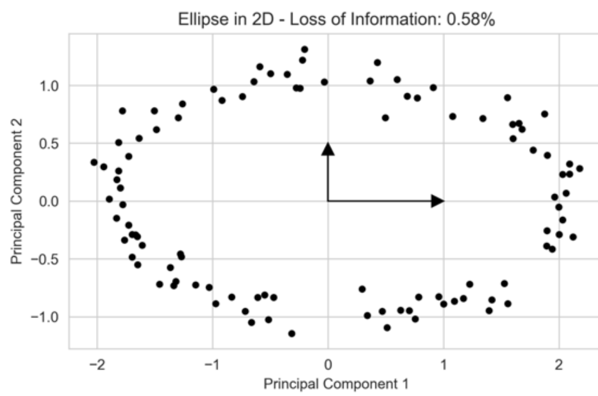
The number of features required to keep average distance constant grows exponentially with the number of dimensions.



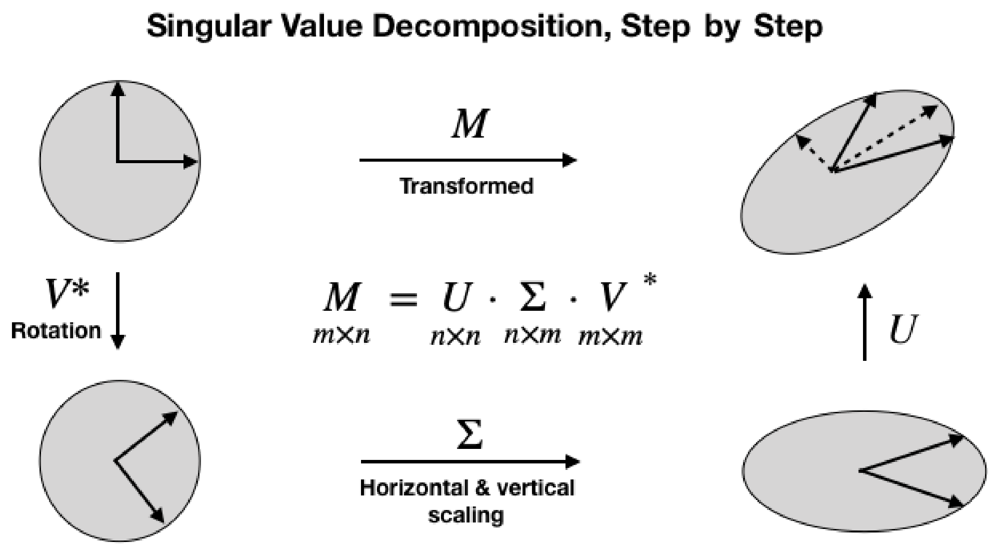


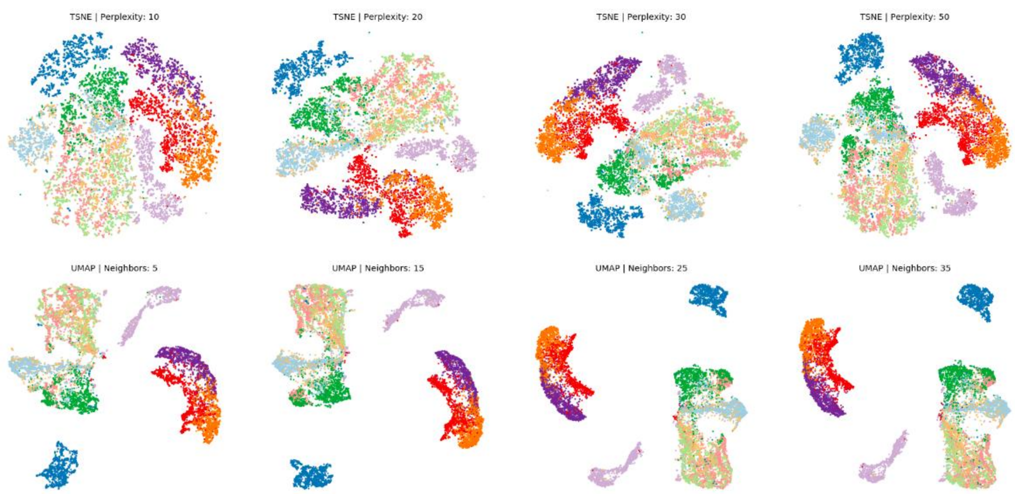
Principal Component Vectors
Principal Component 1
Principal Component 2
x_2
x_1
1D Projection
x_1
PCA vs OLS
x_1

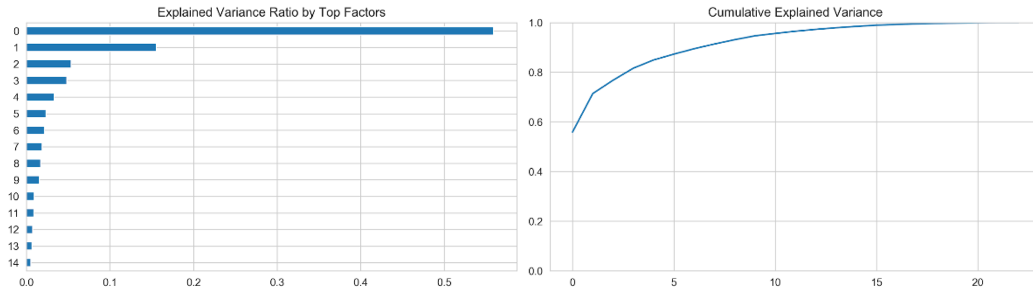




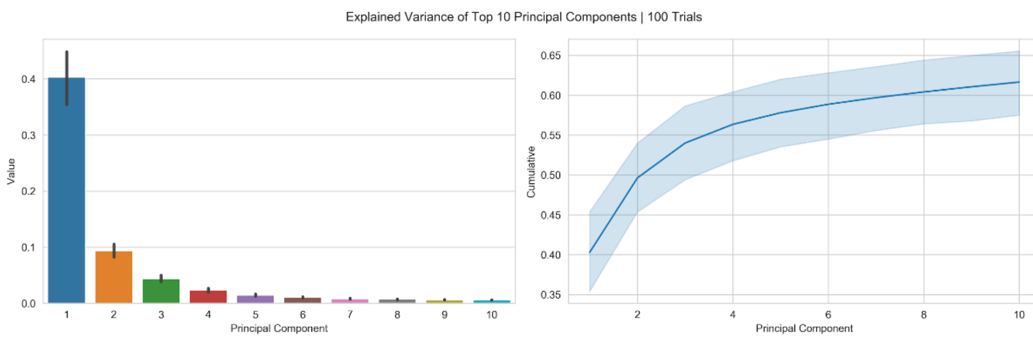
Ellipse in 2D - Loss of Information: 0.58%
Principal Component 2
Principal Component 1


Singular Value Decomposition, Step by Step
M
Transformed
V^*
Rotation
$\underset{m\times n}{M} = \underset{n\times n}{U} \cdot \underset{n\times m}{\Sigma} \cdot \underset{m\times m}{V}^*$
U
Σ
Horizontal & vertical
scaling

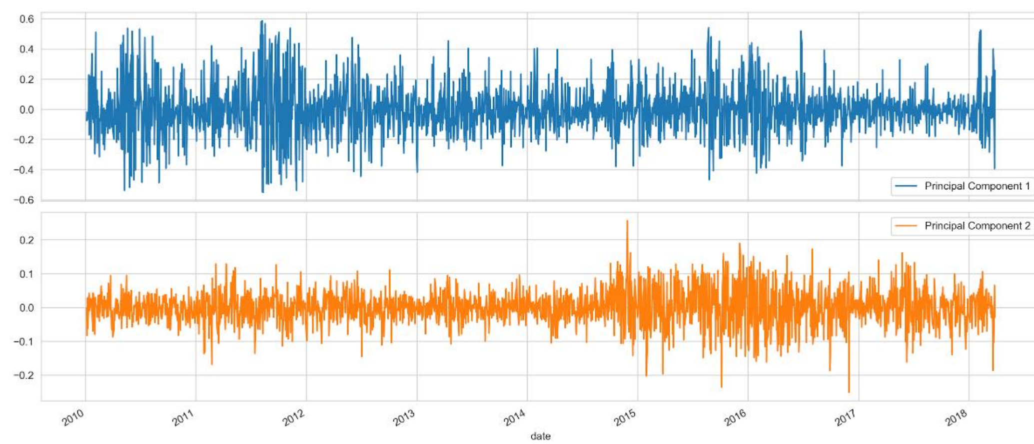

TSNE | Perplexity: 10
TSNE | Perplexity: 20
TSNE | Perplexity: 30
TSNE | Perplexity: 50
UMAP | Neighbors: 5
UMAP | Neighbors: 15
UMAP | Neighbors: 25
UMAP | Neighbors: 35



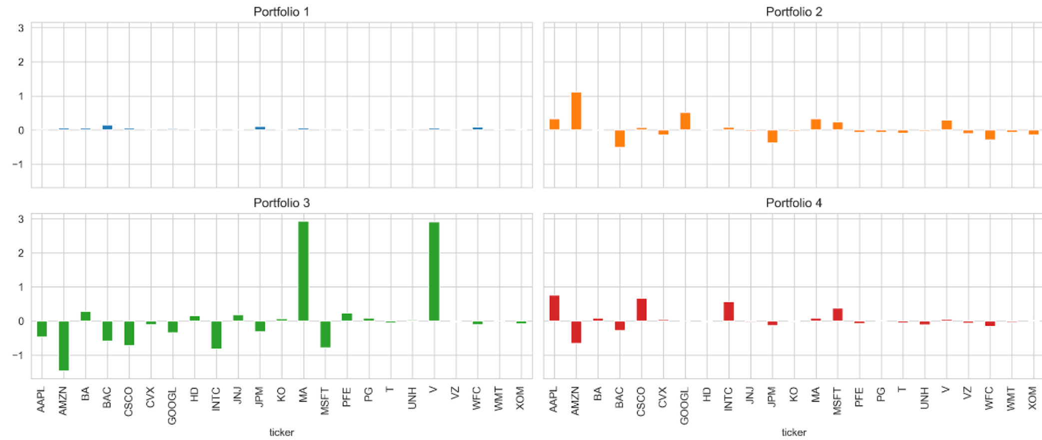
Explained Variance Ratio by Top Factors
Cumulative Explained Variance

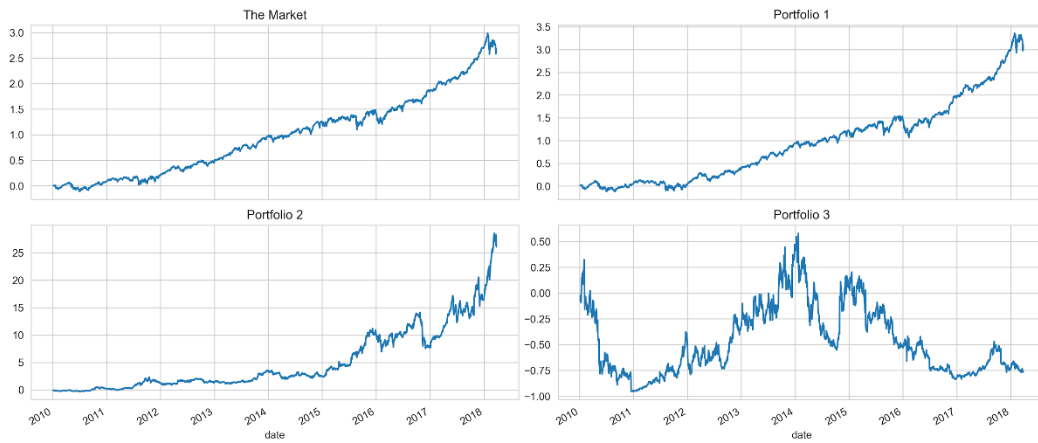


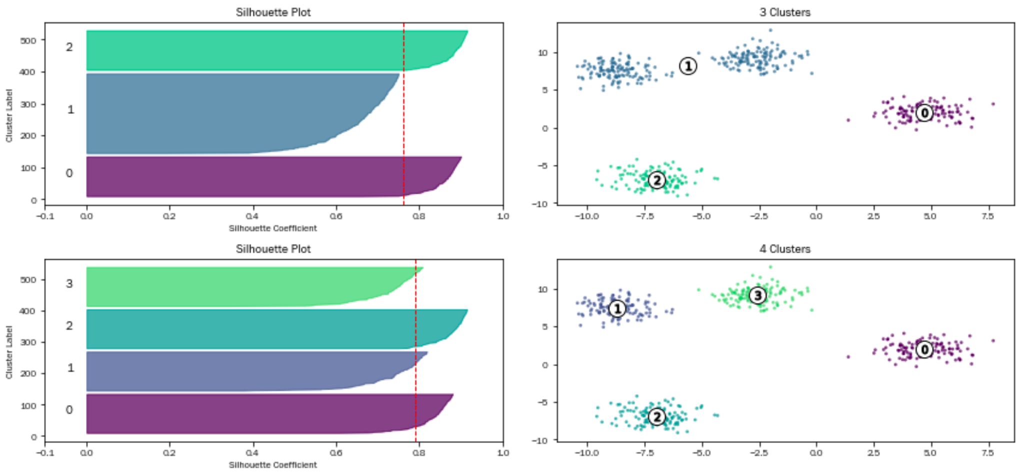
Explained Variance of Top 10 Principal Components | 100 Trials
Value
Principal Component
Cumulative

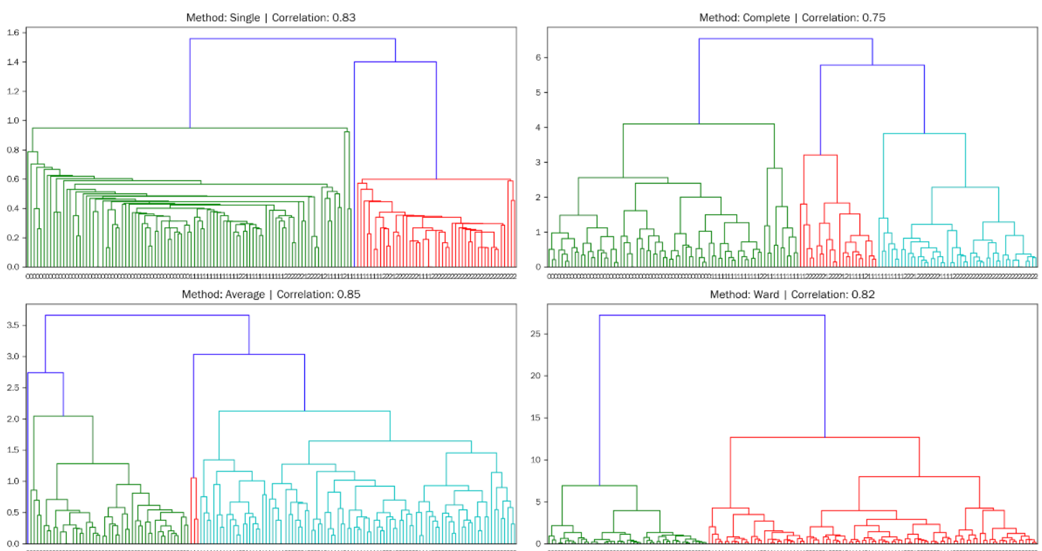


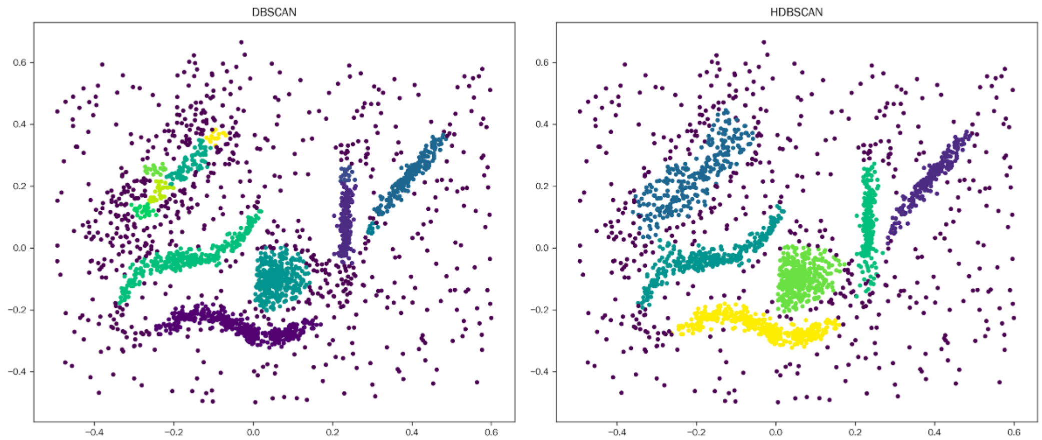
Principal Component 1
Principal Component 2
date

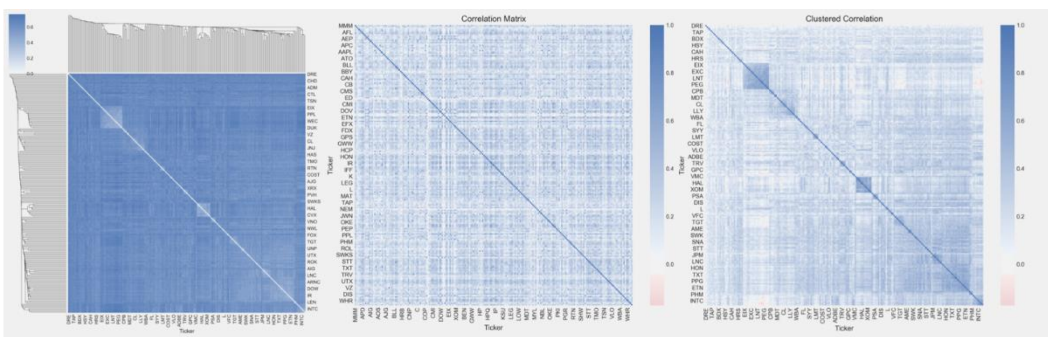

Portfolio 1
Portfolio 2
Portfolio 3
Portfolio 4
ticker
AAPL
AMZN
BA
BAC
CSCO
CVX
GOOGL
HD
INTC
JNJ
JPM
KO
MA
MSFT
PFE
PG
T
UNH
V
VZ
WFC
WMT
XOM

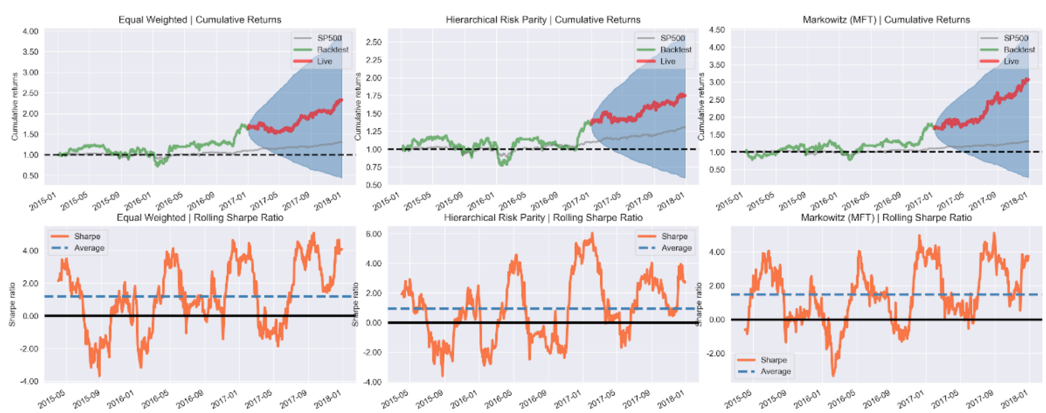

The Market
Portfolio 1
Portfolio 2
Portfolio 3
date


Silhouette Plot
Cluster Label
Silhouette Coefficient
3 Clusters
Silhouette Plot
Cluster Label
Silhouette Coefficient
4 Clusters

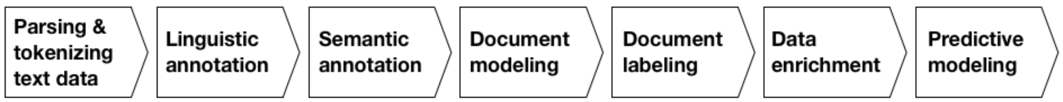

Method: Single | Correlation: 0.83
Method: Complete | Correlation: 0.75
Method: Average | Correlation: 0.85
Method: Ward | Correlation: 0.82

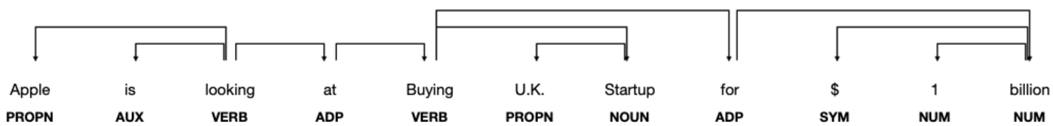

DBSCAN
HDBSCAN

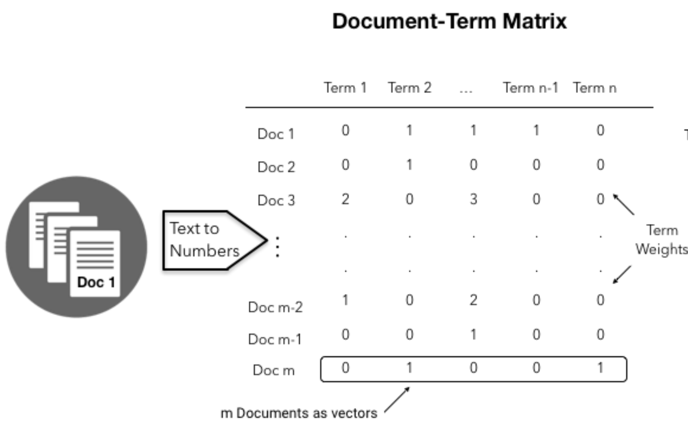

Correlation Matrix
Clustered Correlation
Ticker

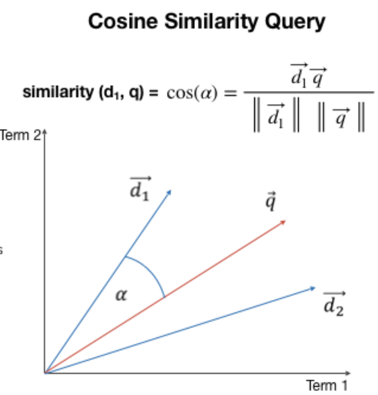

Equal Weighted | Cumulative Returns
Hierarchical Risk Parity | Cumulative Returns
Markowitz (MFT) | Cumulative Returns
SP500
Backtest
Live
Cumulative returns
Equal Weighted | Rolling Sharpe Ratio
Hierarchical Risk Parity | Rolling Sharpe Ratio
Markowitz (MFT) | Rolling Sharpe Ratio
Sharpe
Average
Sharpe ratio

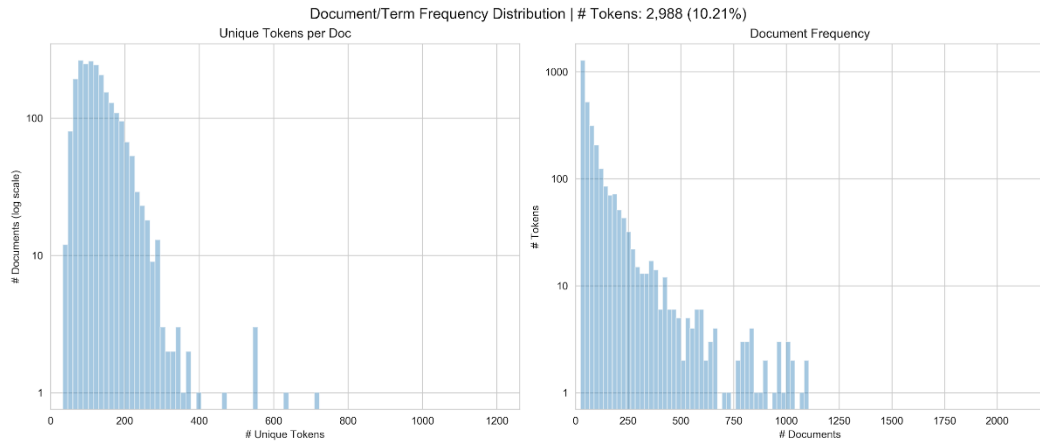
Chapter 14: Text Data for Trading – Sentiment Analysis

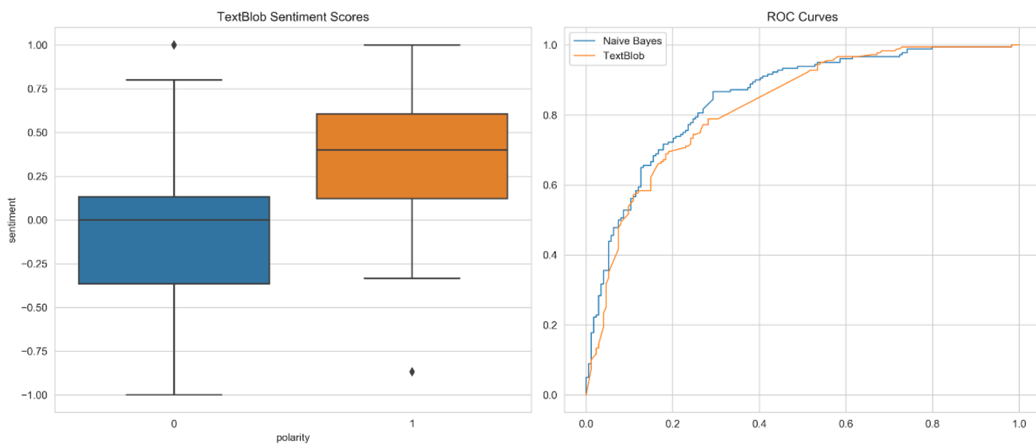


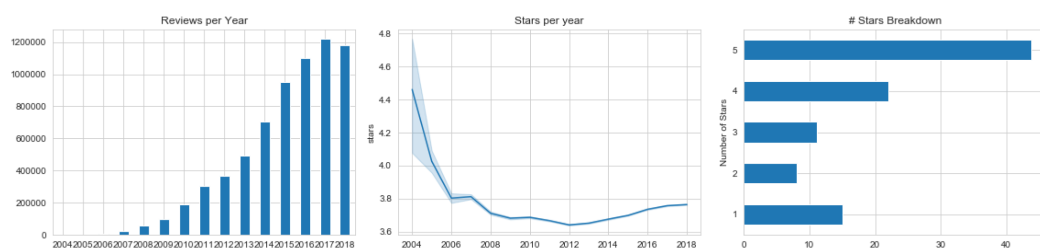


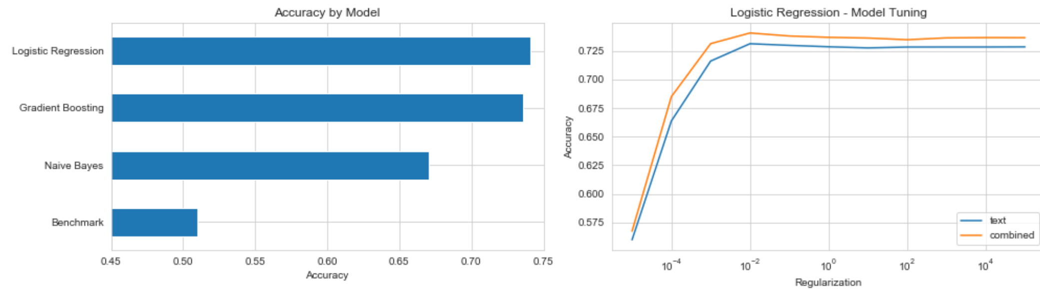




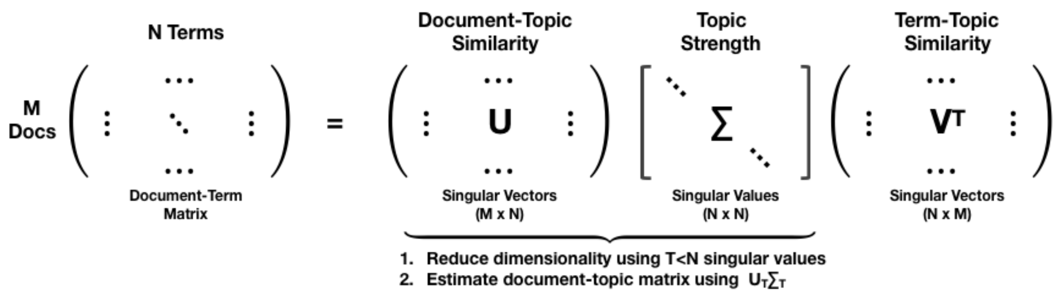

Document/Term Frequency Distribution | # Tokens: 2,988 (10.21%)
Unique Tokens per Doc
Documents (log scale)
Unique Tokens
Document Frequency
Tokens
Documents

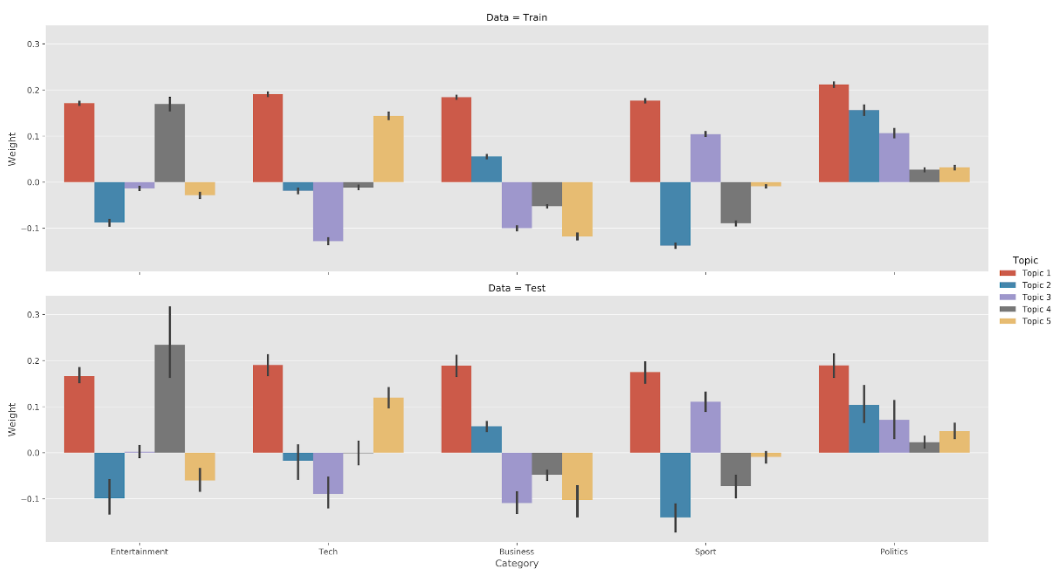

TextBlob Sentiment Scores
sentiment
polarity
ROC Curves
Naive Bayes
TextBlob

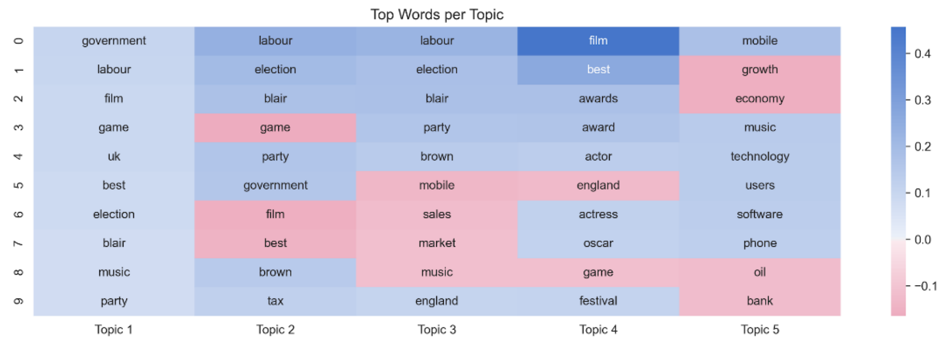

Reviews per Year
Stars per year
stars
year
Stars Breakdown
Number of Stars
Share of all Ratings (%)

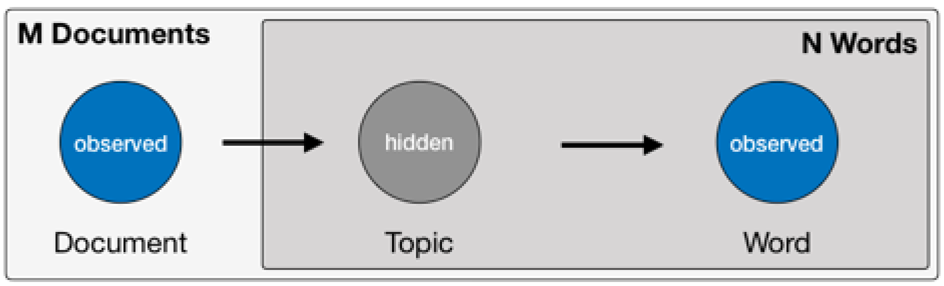

Accuracy by Model
Logistic Regression
Gradient Boosting
Naive Bayes
Benchmark
0.45
0.50
0.55
0.60
0.65
0.70
0.75
Accuracy
Logistic Regression - Model Tuning
Accuracy
0.725
0.700
0.675
0.650
0.625
0.600
0.575
10^{-4}
10^{-2}
10^{0}
10^{2}
10^{4}
Regularization
text
combined

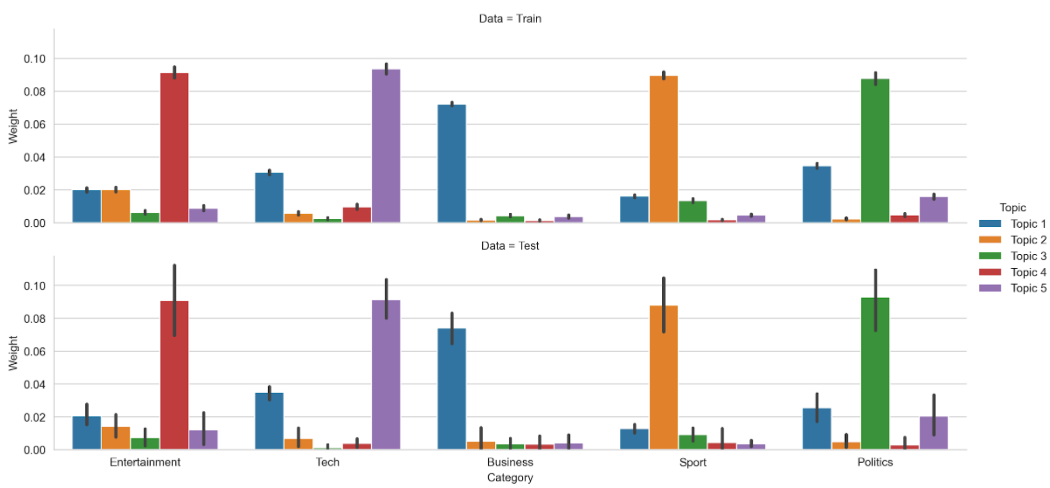
Chapter 15: Topic Modeling – Summarizing Financial News

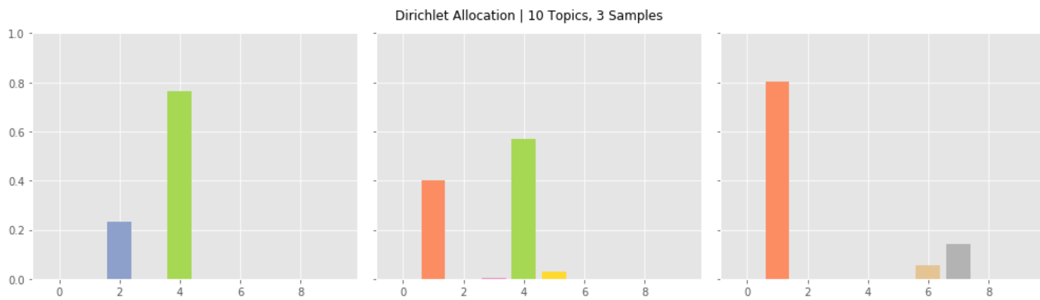




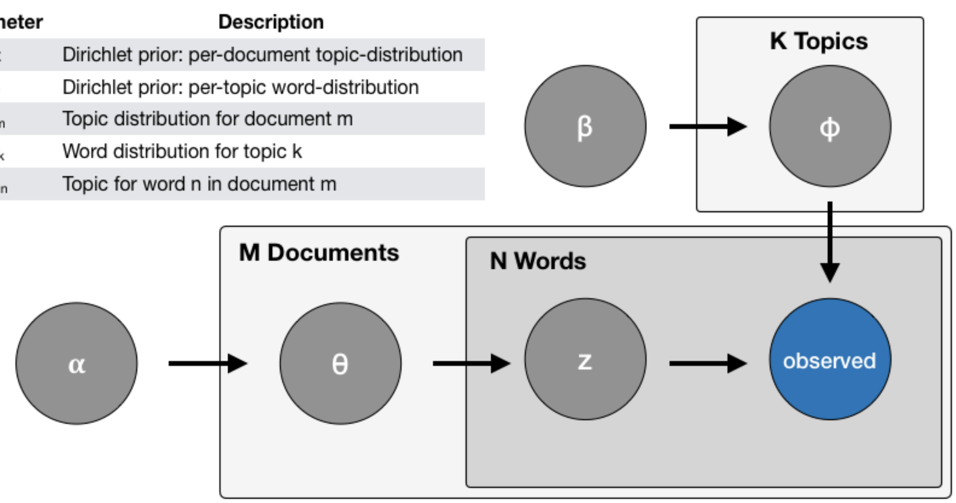

Top Words per Topic
Topic 1: government, labour, film, game, uk, best, election, blair, music, party
Topic 2: labour, election, blair, game, party, government, film, best, brown, tax
Topic 3: labour, election, blair, party, brown, mobile, sales, market, music, england
Topic 4: film, best, awards, award, actor, england, actress, oscar, game, festival
Topic 5: mobile, growth, economy, music, technology, users, software, phone, oil, bank
0 1 2 3 4 5 6 7 8 9
0.4 0.3 0.2 0.1 0.0 −0.1

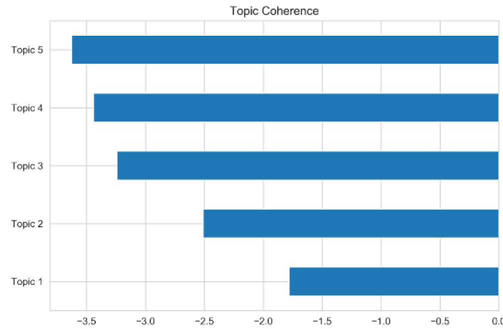

M Documents
N Words
observed
hidden
observed
Document
Topic
Word

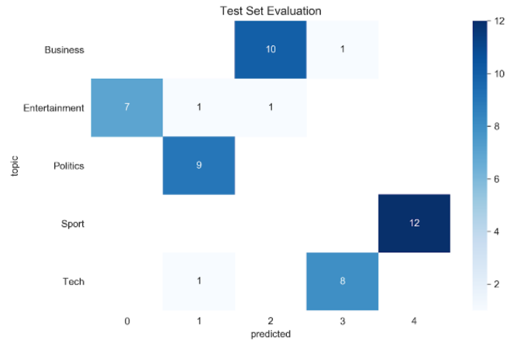

Data = Train
Data = Test
Weight
0.00 0.02 0.04 0.06 0.08 0.10
Entertainment
Tech
Business
Sport
Politics
Category
Topic
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5

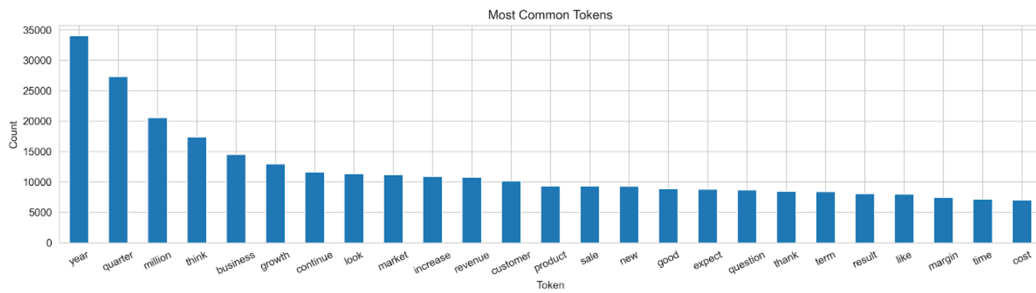


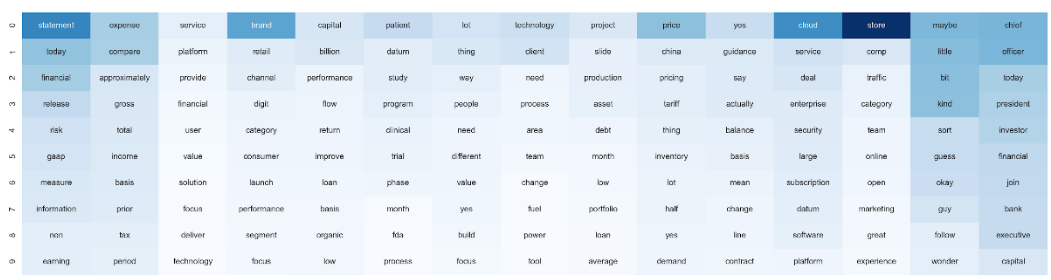
Parameter	Description
α	Dirichlet prior: per-document topic-distribution
β	Dirichlet prior: per-topic word-distribution
θ_m	Topic distribution for document m
ϕ_k	Word distribution for topic k
z_{mn}	Topic for word n in document m

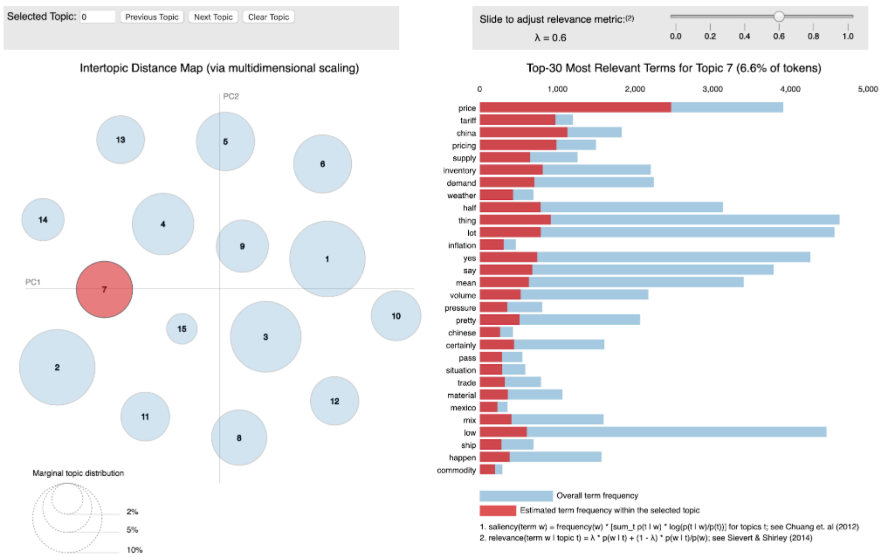



Topic Coherence
Topic 5
Topic 4
Topic 3
Topic 2
Topic 1
−3.5 −3.0 −2.5 −2.0 −1.5 −1.0 −0.5 0.0

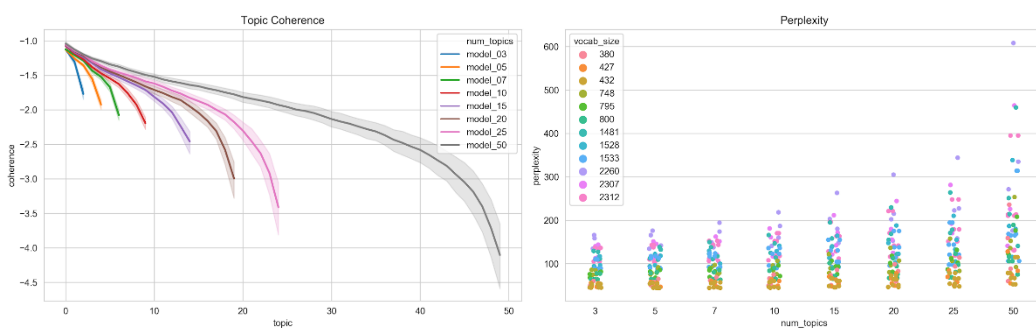

Test Set Evaluation
Business
Entertainment
Politics
Sport
Tech
topic
10 1
7 1 1
9
12
1 8
0 1 2 3 4
predicted
12 10 8 6 4 2


Most Common Tokens
35000 30000 25000 20000 15000 10000 5000 0
Count
year quarter million think business growth continue look market increase revenue customer product sale new good expect question thank term result like margin time cost
Token

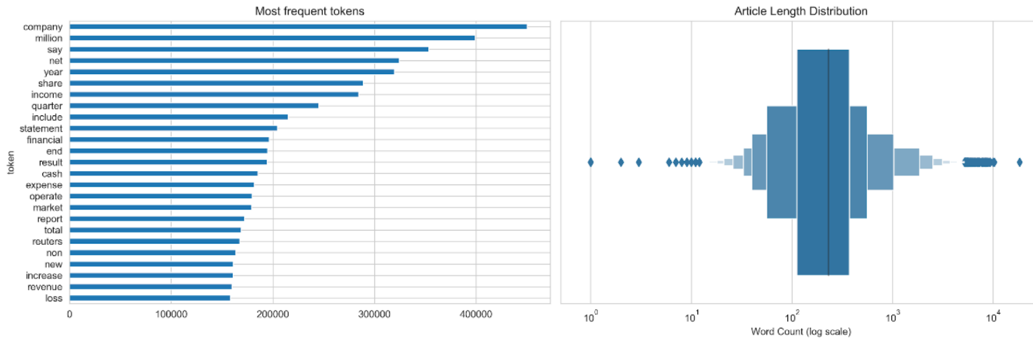

0 statement expense service brand capital patient lot technology project price yes cloud store maybe chief
1 today compare platform retail billion datum thing client slide china guidance service comp little officer
2 financial approximately provide channel performance study way need production pricing say deal traffic bit today
3 release gross financial digit flow program people process asset tariff actually enterprise category kind president
4 risk total user category return clinical need area debt thing balance security team sort investor
5 gaap income value consumer improve trial different team month inventory basis large online guess financial
6 measure basis solution launch loan phase value change low lot mean subscription open okay join
7 information prior focus performance basis month yes fuel portfolio half change datum marketing guy bank
8 non tax deliver segment organic fda build power loan yes line software great follow executive
9 earning period technology focus low process focus tool average demand contract platform experience wonder capital
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14

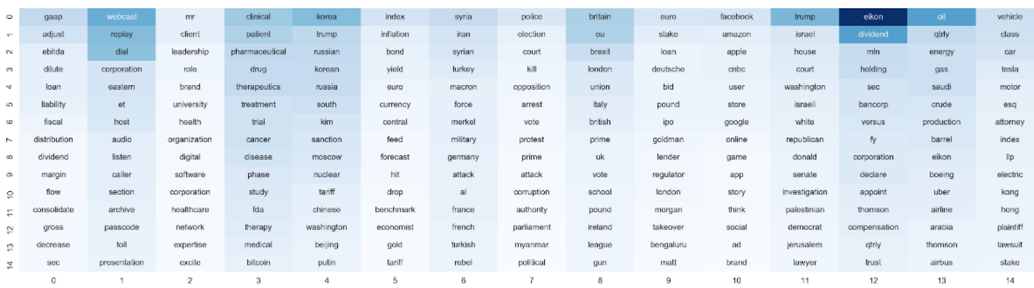


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0.6
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 7 (6.6% of tokens)
0
1,000
2,000
3,000
4,000
5,000
price
tariff
china
pricing
supply
inventory
demand
weather
half
thing
lot
inflation
yes
say
mean
volume
pressure
pretty
chinese
certainly
pass
situation
trade
material
mexico
mix
low
ship
happen
commodity
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

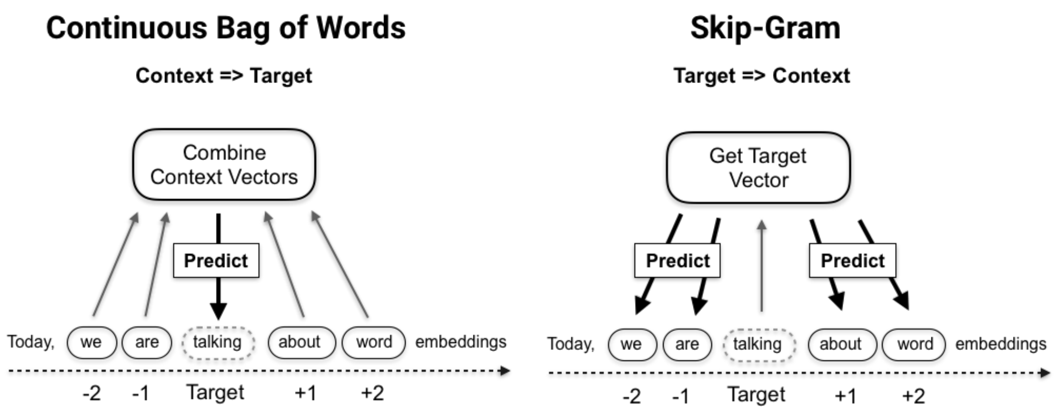


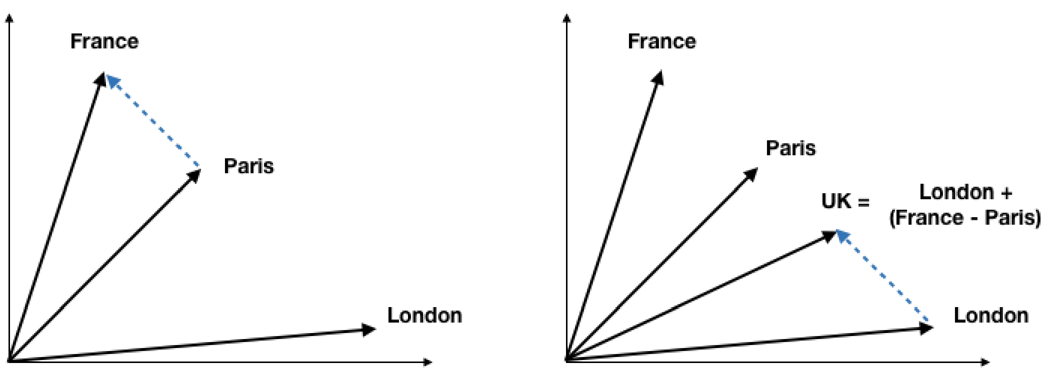
Topic Coherence
coherence
topic
num_topics
model_03
model_05
model_07
model_10
model_15
model_20
model_25
model_50
Perplexity
perplexity
num_topics
vocab_size
380
427
432
748
795
800
1481
1528
1533
2260
2307
2312

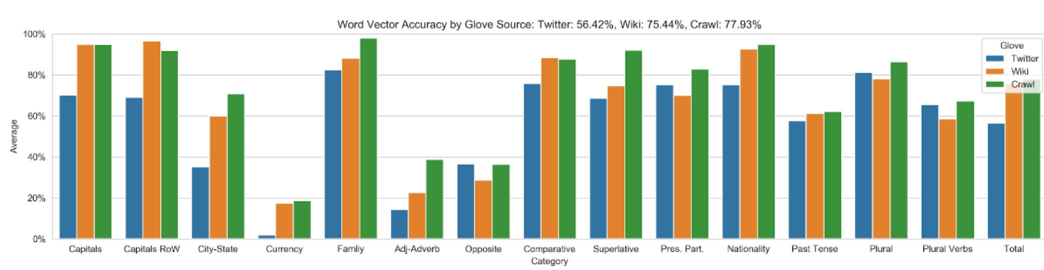

Most frequent tokens
company
million
say
net
year
share
income
quarter
include
statement
financial
end
result
cash
expense
operate
market
report
total
reuters
non
new
increase
revenue
loss
token
0
100000
200000
300000
400000
Article Length Distribution
10^0
10^1
10^2
10^3
10^4
Word Count (log scale)

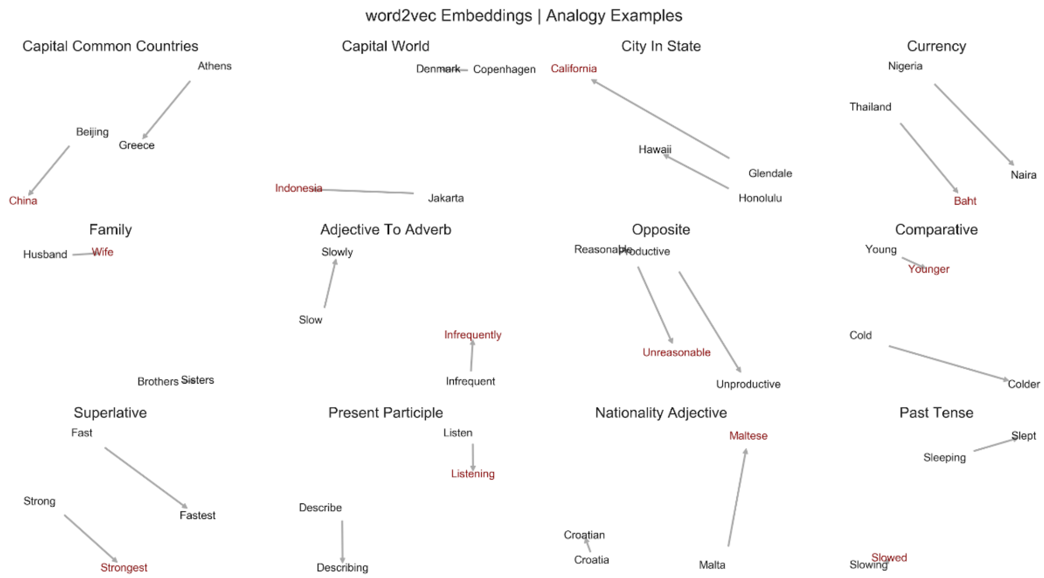

gaap webcast mr clinical korea index syria police britain euro facebook trump eikon oil vehicle
adjust replay client patient trump inflation iran election eu stake amazon israel dividend qtrly class
ebitda dial leadership pharmaceutical russian bond syrian court brexit loan apple house mln energy car
dilute corporation role drug korean yield turkey kill london deutsche cnbc court holding gas tesla
loan eastern brand therapeutics russia euro macron opposition union bid user washington sec saudi motor
liability et university treatment south currency force arrest italy pound store israeli bancorp crude esq
fiscal host health trial kim central merkel vote british ipo google white versus production attorney
distribution audio organization cancer sanction feed military protest prime goldman online republican fy barrel index
dividend listen digital disease moscow forecast germany prime uk lender game donald corporation eikon llp
margin caller software phase nuclear hit attack attack vote regulator app senate declare boeing electric
flow section corporation study tariff drop al corruption school london story investigation appoint uber kong
consolidate archive healthcare fda chinese benchmark france authority pound morgan think palestinian thomson airline hong
gross passcode network therapy washington economist french parliament ireland takeover social democrat compensation arabia plaintiff
decrease toll expertise medical beijing gold turkish myanmar league bengaluru ad jerusalem qtrly thomson lawsuit
sec presentation excite bitcoin putin tariff rebel political gun matt brand lawyer trust airbus stake
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14

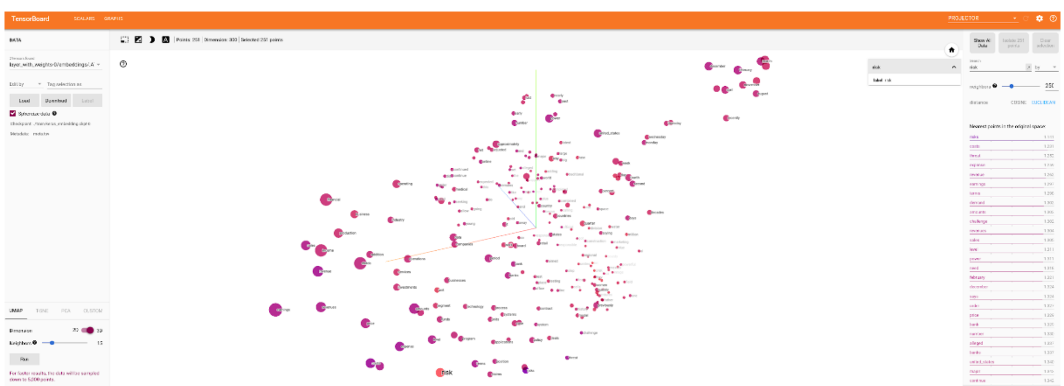
Chapter 16: Word Embeddings for Earnings Calls and SEC Filings

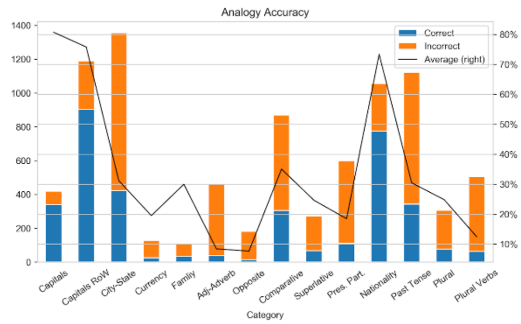


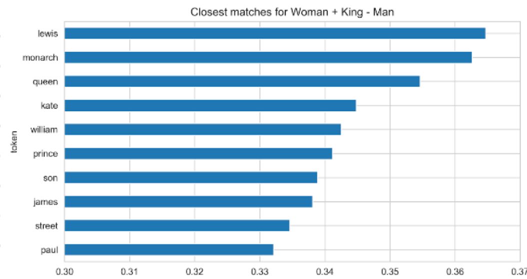


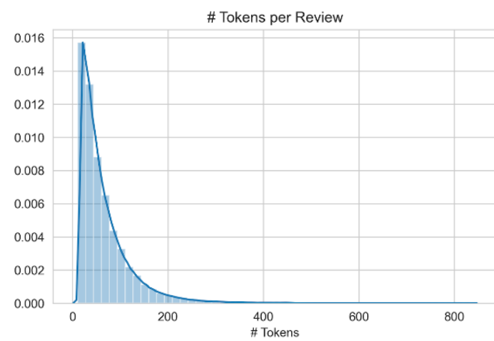


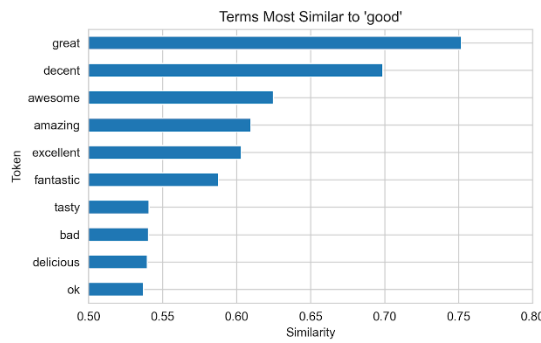

word2vec Embeddings | Analogy Examples
Capital Common Countries
Athens
Beijing
Greece
China
Capital World
Denmark
Copenhagen
Indonesia
Jakarta
City In State
California
Hawaii
Glendale
Honolulu
Currency
Nigeria
Thailand
Naira
Baht
Family
Husband
Wife
Brothers
Sisters
Adjective To Adverb
Slowly
Slow
Infrequently
Infrequent
Opposite
Reasonable
Productive
Unreasonable
Unproductive
Comparative
Young
Younger
Cold
Colder
Superlative
Fast
Strong
Fastest
Strongest
Present Participle
Listen
Listening
Describe
Describing
Nationality Adjective
Maltese
Croatian
Croatia
Malta
Past Tense
Slept
Sleeping
Slowed
Slowing

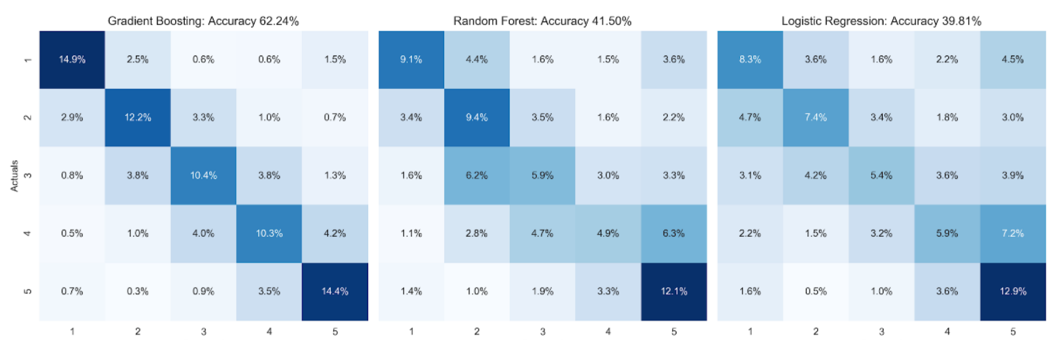



Analogy Accuracy
Correct
Incorrect
Average (right)
Capitals
Capitals RoW
City-State
Currency
Family
Adj-Adverb
Opposite
Comparative
Superlative
Pres. Part.
Nationality
Past Tense
Plural
Plural Verbs
Category

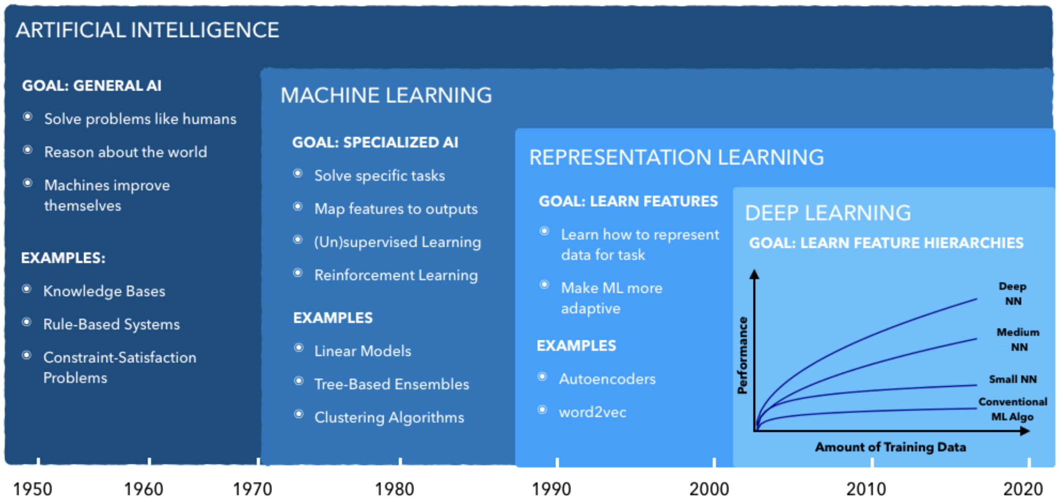

Closest matches for Woman + King - Man
token
lewis
monarch
queen
kate
william
prince
son
james
street
paul

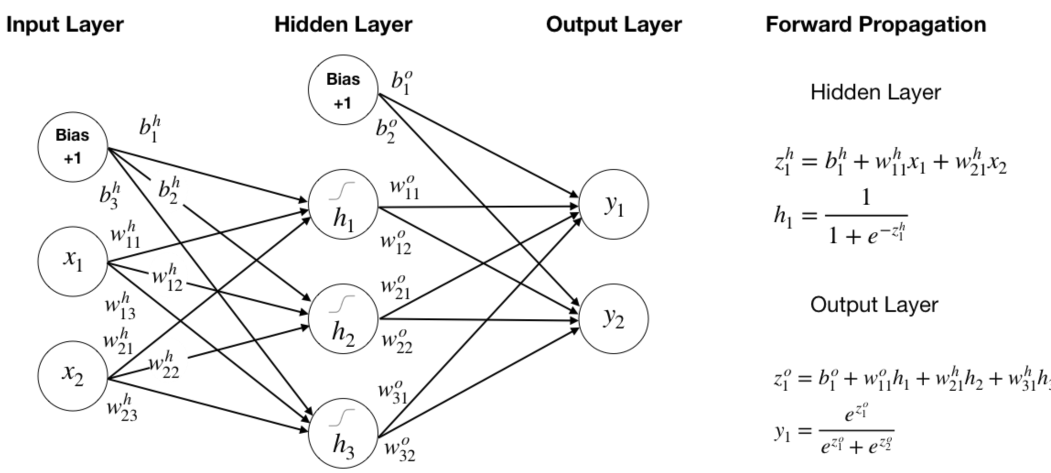

Tokens per Review
Tokens

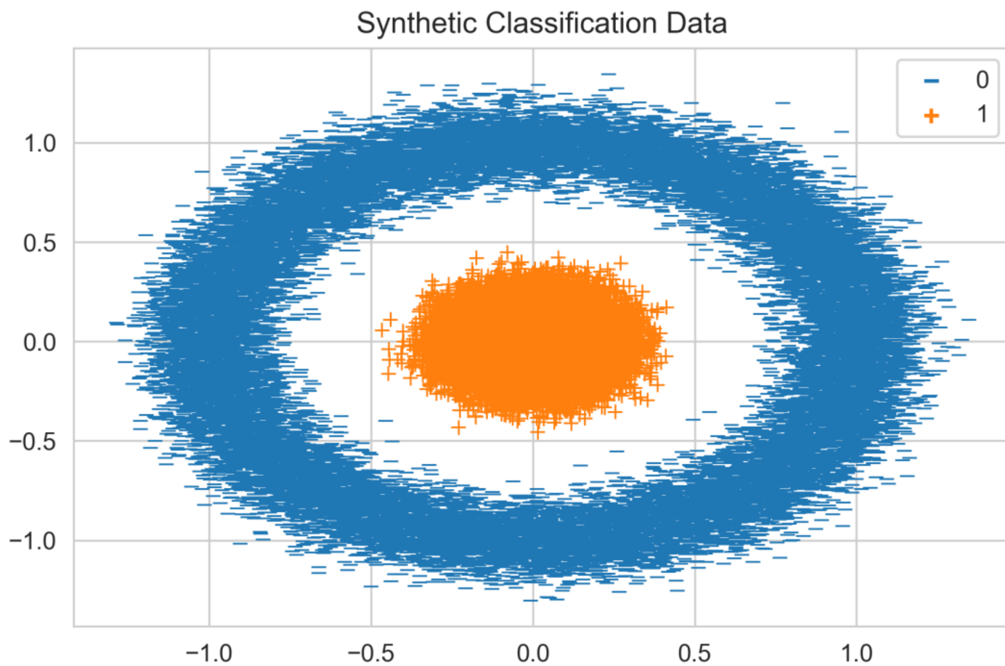

Terms Most Similar to 'good'
Token
great
decent
awesome
amazing
excellent
fantastic
tasty
bad
delicious
ok
Similarity

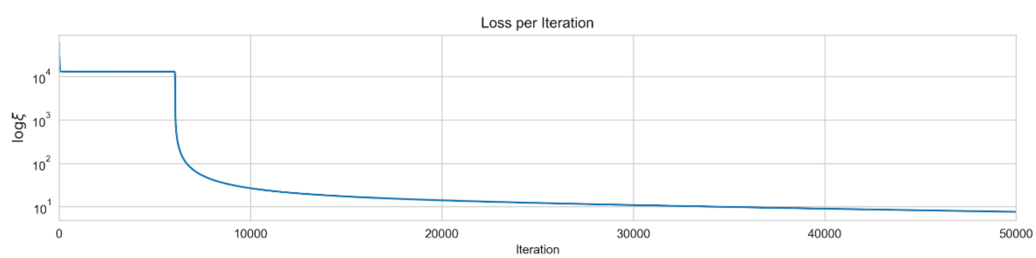

Gradient Boosting: Accuracy 62.24%
Random Forest: Accuracy 41.50%
Logistic Regression: Accuracy 39.81%
Actuals
Predicted

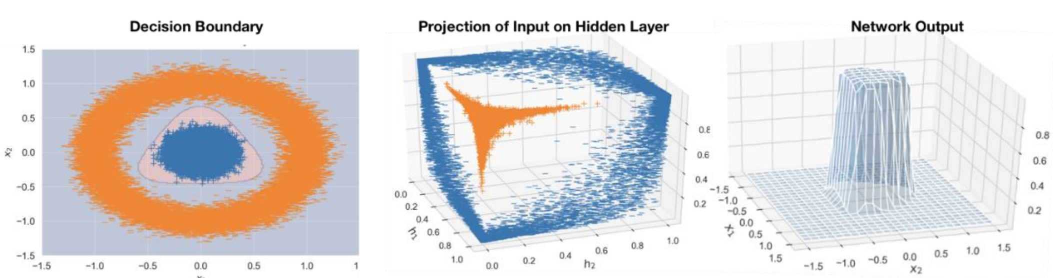
Chapter 17: Deep Learning for Trading

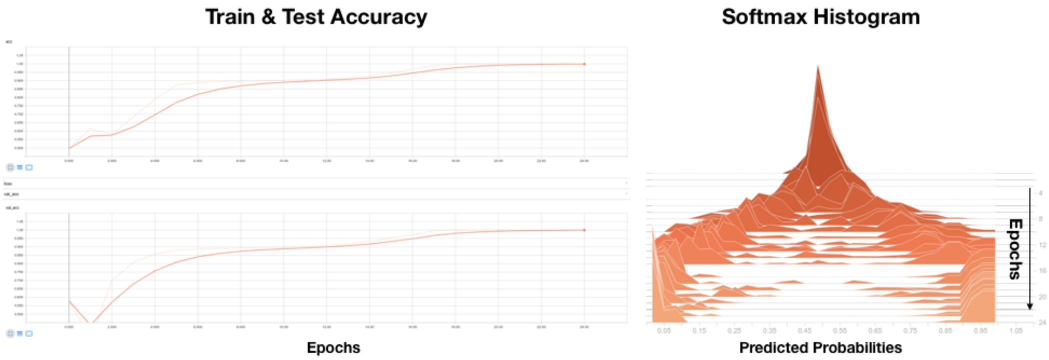


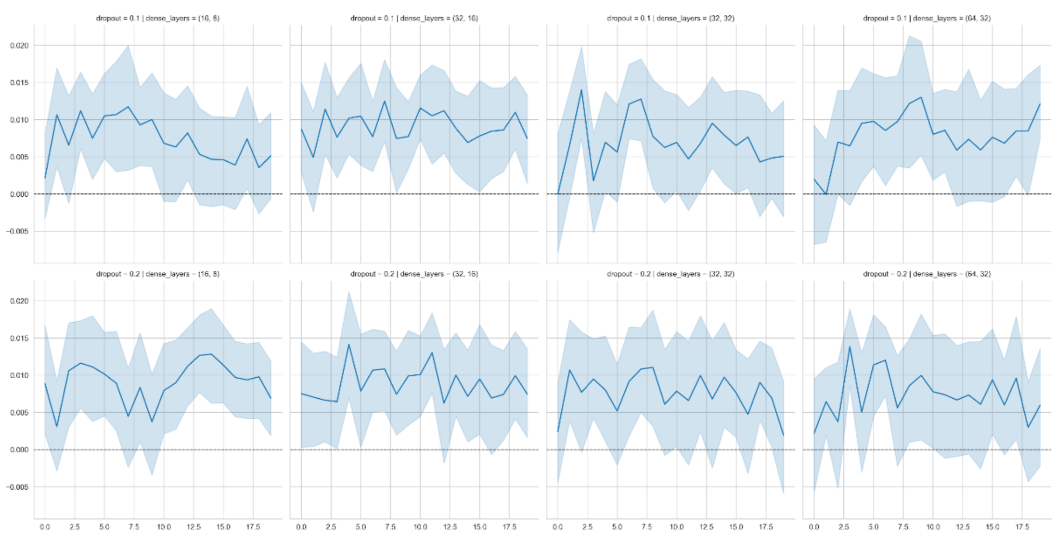


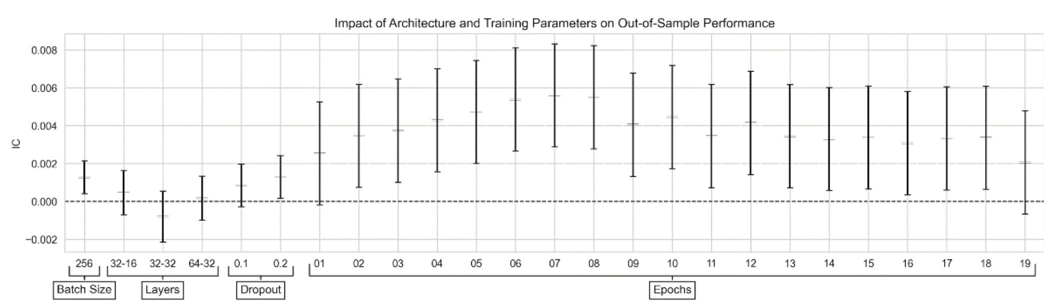

Synthetic Classification Data
0
1
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

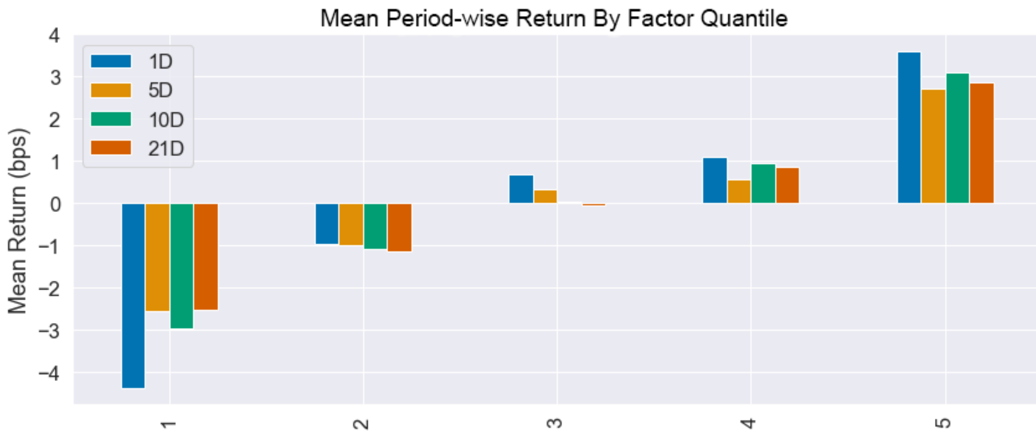

Loss per Iteration
log ξ
10^4
10^3
10^2
10^1
0
10000
20000
30000
40000
50000
Iteration

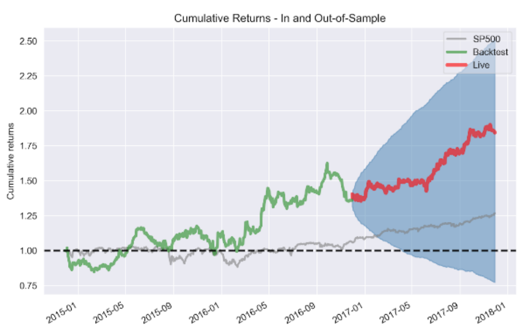

Decision Boundary
Projection of Input on Hidden Layer
Network Output
x1
x2
h1
h2

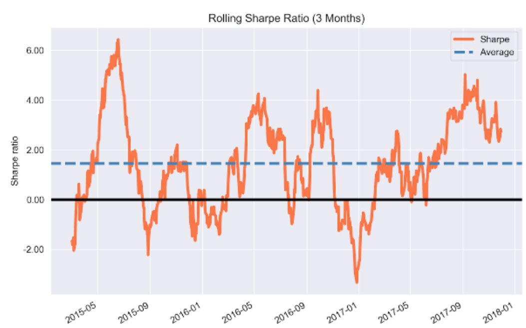

Train & Test Accuracy
Softmax Histogram
Epochs
Epochs
Predicted Probabilities


dropout = 0.1 | dense_layers = (16, 8)
dropout = 0.1 | dense_layers = (32, 16)
dropout = 0.1 | dense_layers = (32, 32)
dropout = 0.1 | dense_layers = (64, 32)
dropout = 0.2 | dense_layers = (16, 8)
dropout = 0.2 | dense_layers = (32, 16)
dropout = 0.2 | dense_layers = (32, 32)
dropout = 0.2 | dense_layers = (64, 32)

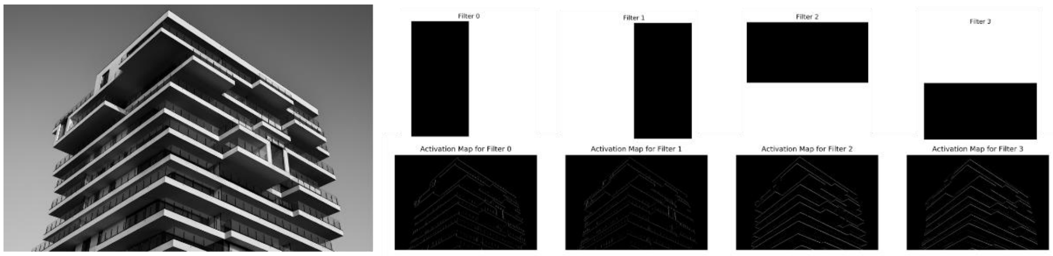

Impact of Architecture and Training Parameters on Out-of-Sample Performance
IC
0.008
0.006
0.004
0.002
0.000
−0.002
256
32-16
32-32
64-32
0.1
0.2
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
Batch Size
Layers
Dropout
Epochs

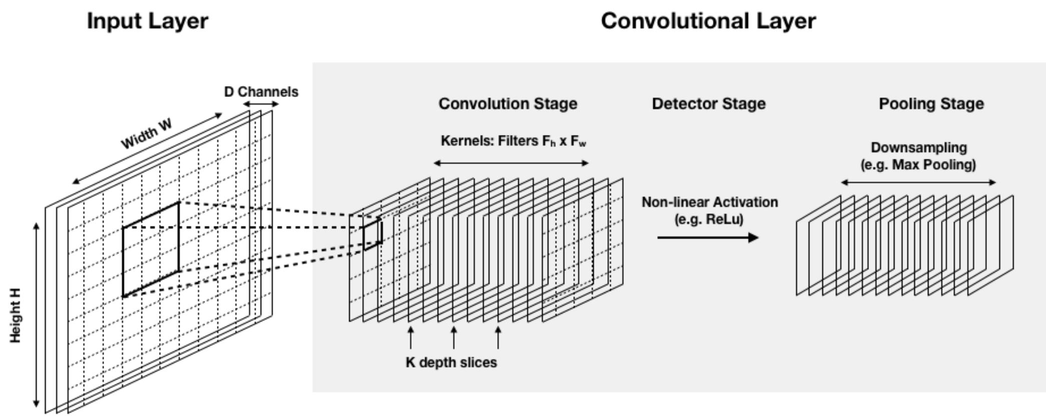

Mean Period-wise Return By Factor Quantile
Mean Return (bps)
1D
5D
10D
21D
4
3
2
1
0
−1
−2
−3
−4
1
2
3
4
5

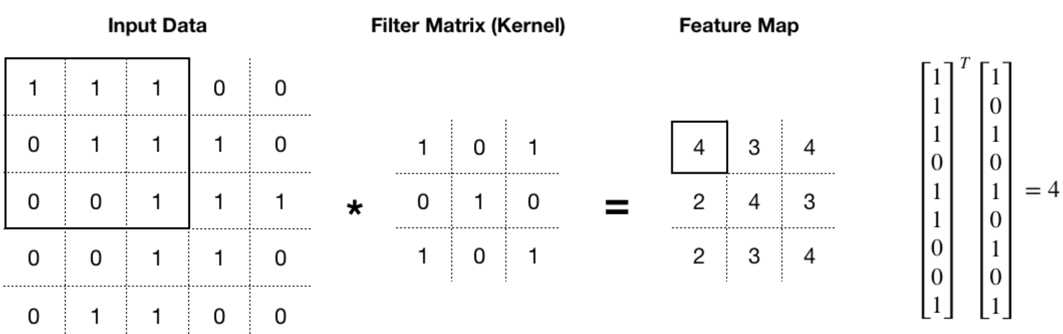

Cumulative Returns - In and Out-of-Sample
Cumulative returns
SP500
Backtest
Live
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.75
2015-01
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01

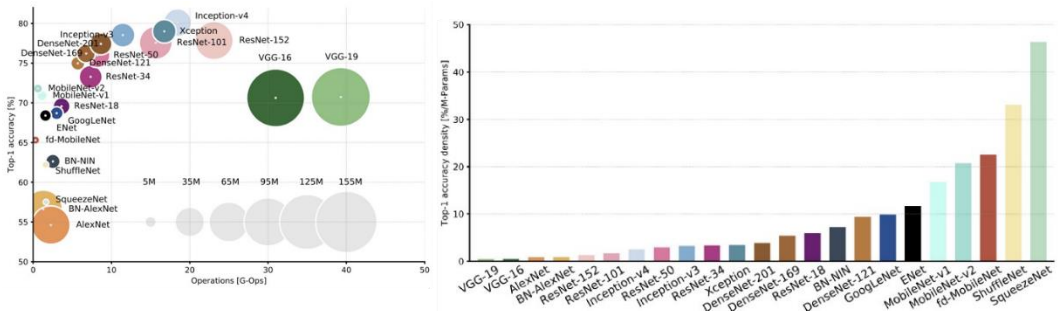

Rolling Sharpe Ratio (3 Months)
Sharpe ratio
Sharpe
Average
6.00
4.00
2.00
0.00
-2.00
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09
2018-01

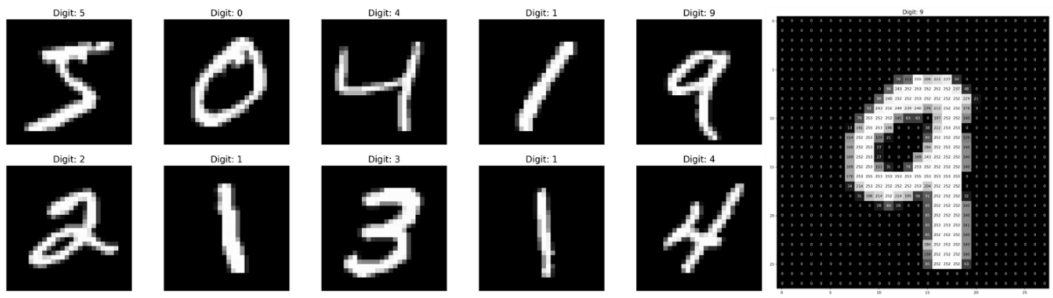
Chapter 18: CNNs for Financial Time Series and Satellite Images



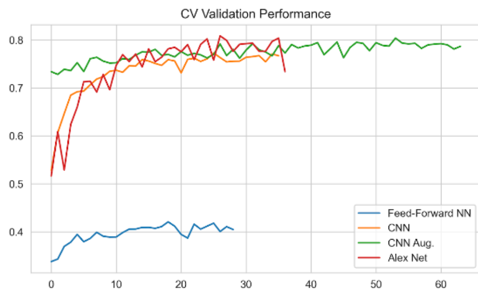


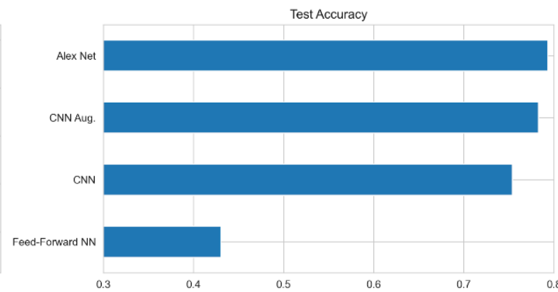


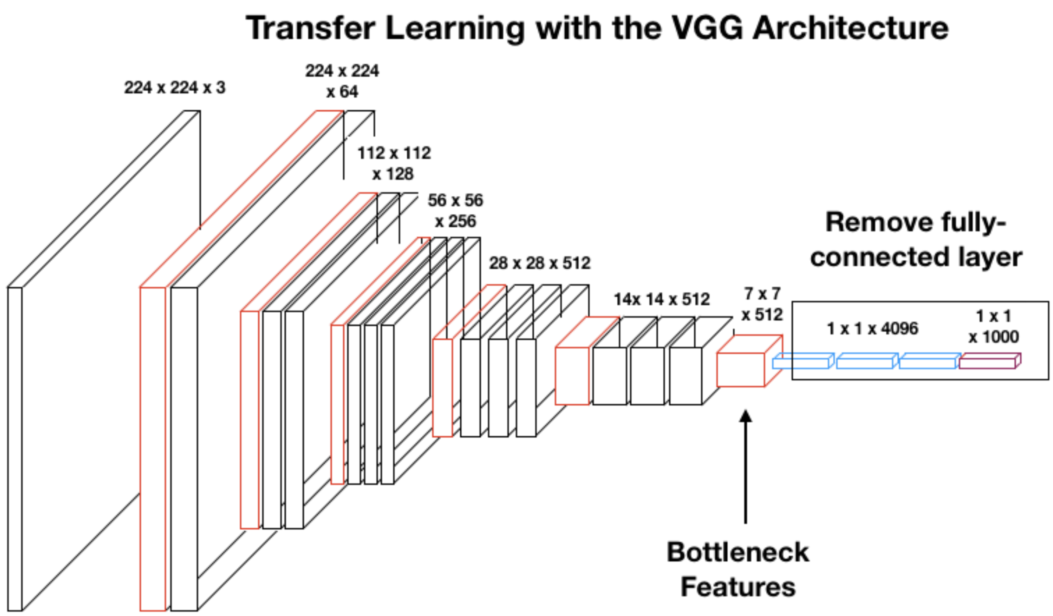

Top-1 accuracy [%]
Operations [G-Ops]
Inception-v4
Xception
ResNet-152
ResNet-101
ResNet-50
Inception-v3
DenseNet-201
DenseNet-169
DenseNet-121
ResNet-34
VGG-16
VGG-19
MobileNet-v2
MobileNet-v1
ResNet-18
GoogLeNet
ENet
fd-MobileNet
BN-NIN
ShuffleNet
SqueezeNet
BN-AlexNet
AlexNet
5M
35M
65M
95M
125M
155M
Top-1 accuracy density [%/M-Params]
VGG-19
VGG-16
AlexNet
BN-AlexNet
ResNet-152
ResNet-101
Inception-v4
ResNet-50
Inception-v3
ResNet-34
Xception
DenseNet-201
DenseNet-169
ResNet-18
BN-NIN
DenseNet-121
GoogLeNet
ENet
MobileNet-v1
MobileNet-v2
fd-MobileNet
ShuffleNet
SqueezeNet

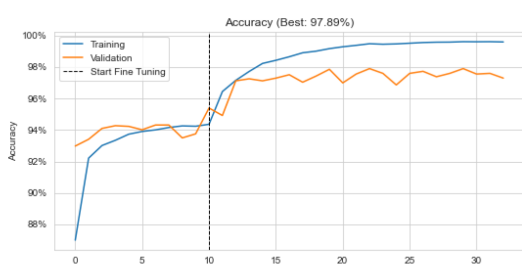

Digit: 5
Digit: 0
Digit: 4
Digit: 1
Digit: 9
Digit: 9
Digit: 2
Digit: 1
Digit: 3
Digit: 1
Digit: 4


Subset of Original Training Images
Augmented Images


CV Validation Performance
Feed-Forward NN
CNN
CNN Aug.
Alex Net

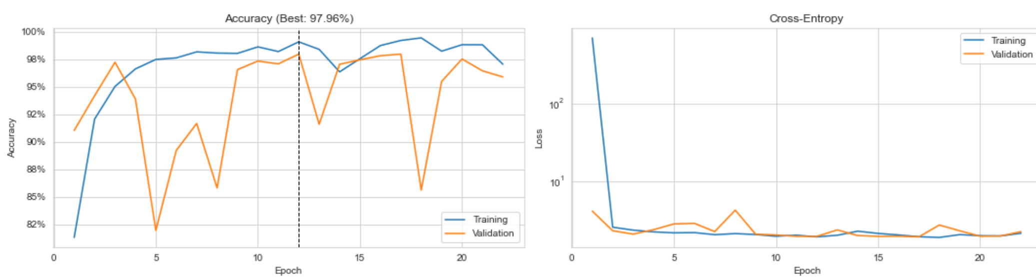

Test Accuracy
Alex Net
CNN Aug.
CNN
Feed-Forward NN


Transfer Learning with the VGG Architecture
224 x 224 x 3
224 x 224
x 64
112 x 112
x 128
56 x 56
x 256
28 x 28 x 512
14x 14 x 512
7 x 7
x 512
Remove fully-
connected layer
1 x 1 x 4096
1 x 1
x 1000
Bottleneck
Features

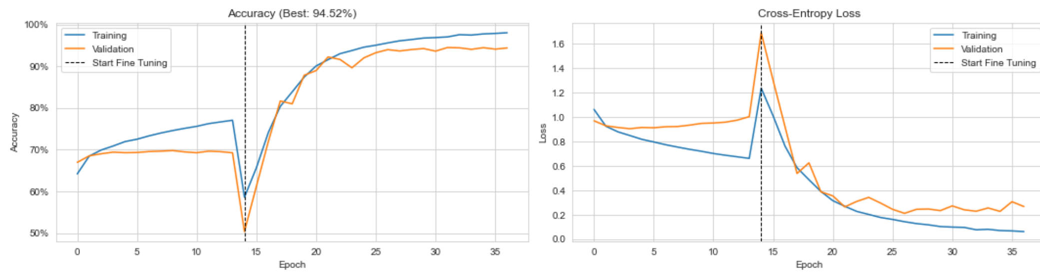

Accuracy (Best: 97.89%)
Training
Validation
Start Fine Tuning
Accuracy
Epoch

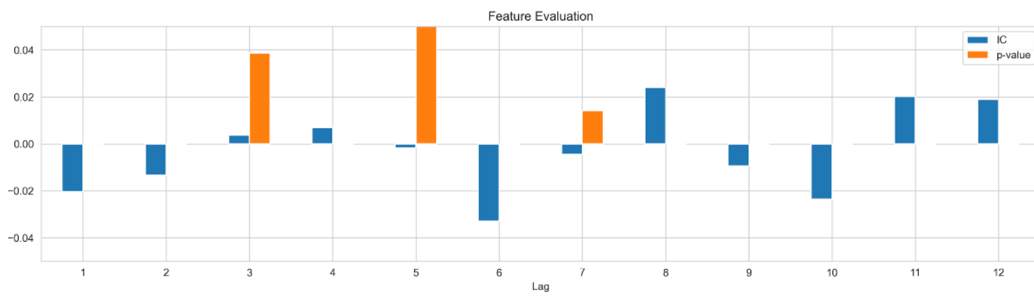

Cross-Entropy Loss
Training
Validation
Start Fine Tuning
Loss
Epoch

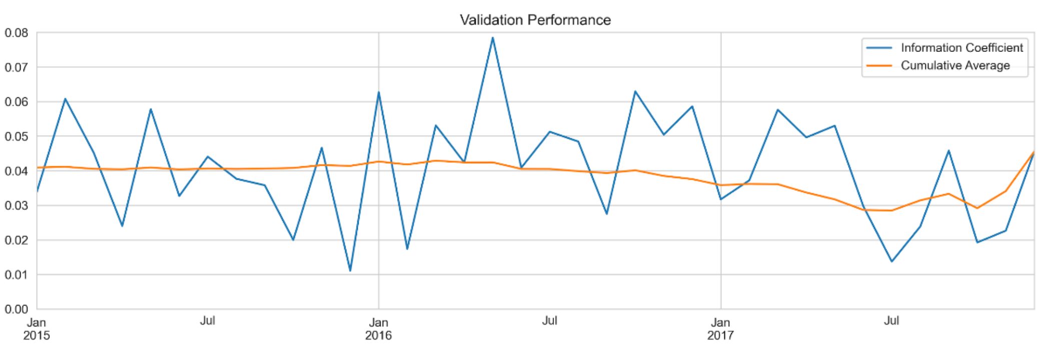

EuroSat Satellite Images
HerbaceousVegetation
Highway
Pasture
River
Residential
SeaLake
Forest
Industrial
PermanentCrop
AnnualCrop


Accuracy (Best: 97.96%)
Accuracy
100%
98%
95%
92%
90%
88%
85%
82%
Training
Validation
Epoch
Cross-Entropy
Loss
Training
Validation
Epoch

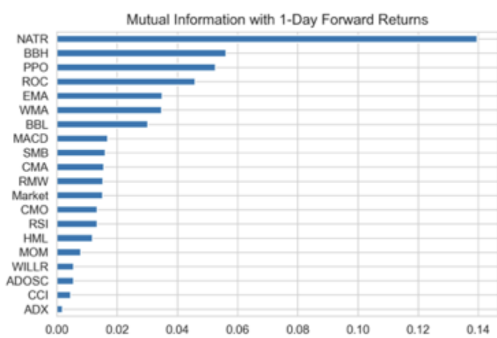

19
23
25
93
31
33
28
744
128
16
23
63
42
58
16
23

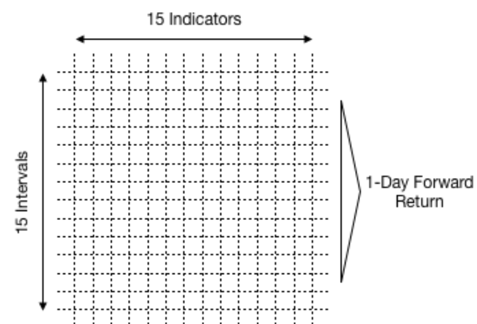

Accuracy (Best: 94.52%)
Training
Validation
Start Fine Tuning
Accuracy
Epoch
Cross-Entropy Loss
Training
Validation
Start Fine Tuning
Loss
Epoch

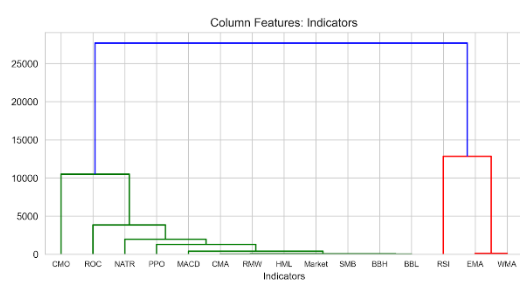

Feature Evaluation
IC
p-value
Lag

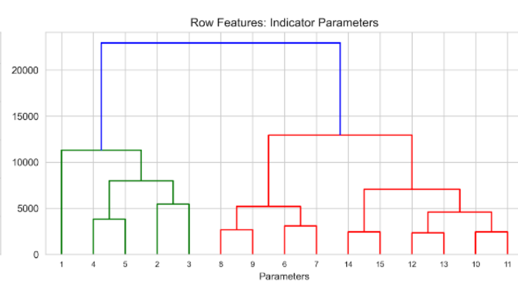

Validation Performance
Information Coefficient
Cumulative Average
Jan
2015
Jul
Jan
2016
Jul
Jan
2017
Jul

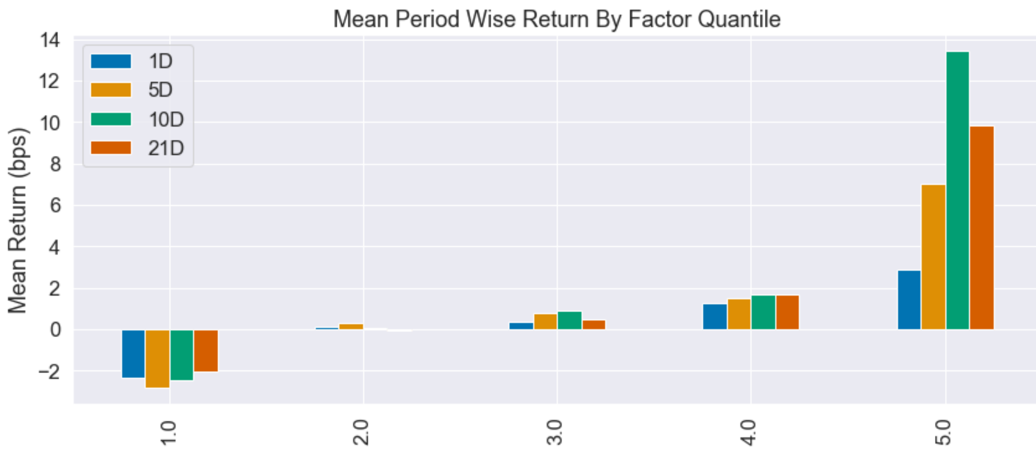
Indicator	Name	Formula
Relative Strength Index (RSI)	Oscillates in [0, 100] range; below 30: oversold, over 70: overbought	See Chapter 4
Williams %R	Momentum-based in [-100, 0] range, below -80: oversold, above -20: overbought	$R = \frac{\max(\mathbf{high}) - \mathbf{close}}{\max(\mathbf{high}) - \min(\mathbf{low})}$
Bollinger Bands	20-day moving average plus/minus daily standard deviation; prices above/below these bands indicate overbought/sold	See chapter 4.
Normalized Average True Range (NATR)	Avg. true range: max of current high-low, current high-prev.close or absolute of prev. close - current low, averaged over t days.	$\mathbf{NATR} = \frac{\mathbf{ATR}(t)}{\mathbf{Close}}$
Percentage Price Oscillator (PPO)	Momentum: compares two exponential moving averages (EMA) in percentage terms	$\mathbf{PPO} = \frac{\mathbf{EMA}_{12} - \mathbf{EMA}_{26}}{\mathbf{EMA}_{26}}$
Commodity Channel Index (CCI)	Momentum-based: difference between current and simple moving average (SMA) of the historical average price, normalized by their mean difference	$p^{\mathbf{hist}} = \sum_{t=1}^{P} (\mathbf{high} + \mathbf{low} + \mathbf{close})/3$ $\mathbf{CCI} = \frac{p^{\mathbf{hist}} - \mathbf{SMA}\left(p^{\mathbf{hist}}\right)}{0.15 \times \sum_{t=1}^{P} (\lvert p^{\mathbf{hist}} - \mathbf{SMA}\left(p^{\mathbf{hist}}\right) \rvert)/P}$

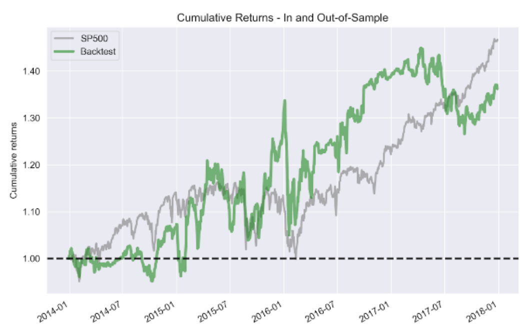


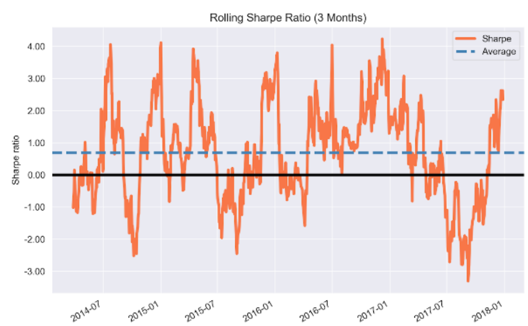




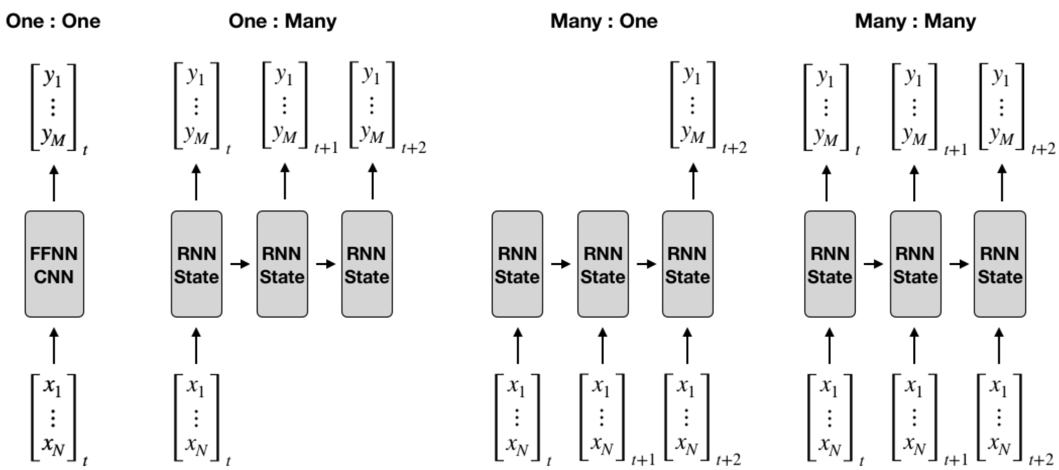


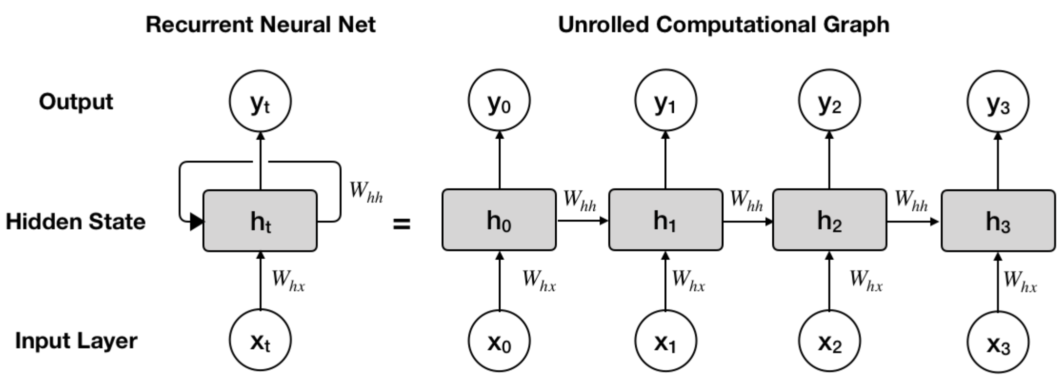

Mean Period Wise Return By Factor Quantile
1D
5D
10D
21D
Mean Return (bps)
14
12
10
8
6
4
2
0
−2
1.0
2.0
3.0
4.0
5.0

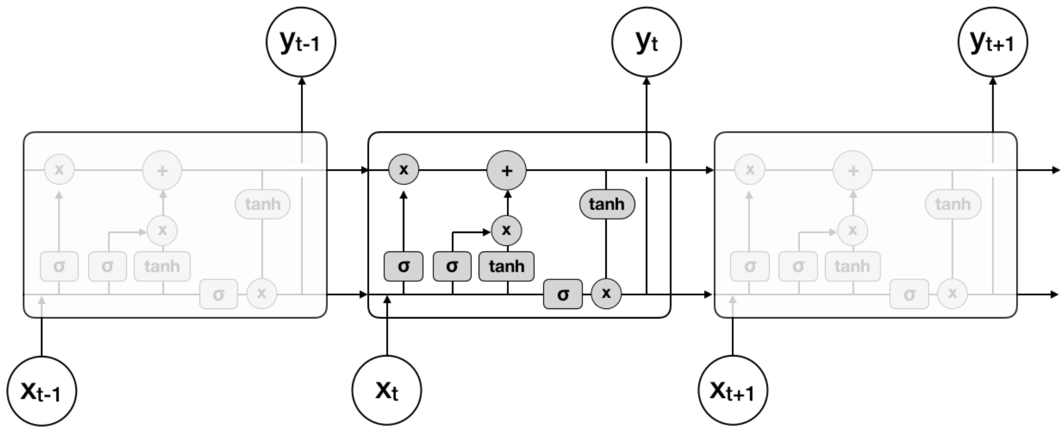

Cumulative Returns - In and Out-of-Sample
SP500
Backtest
Cumulative returns
1.40
1.30
1.20
1.10
1.00
2014-01
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01

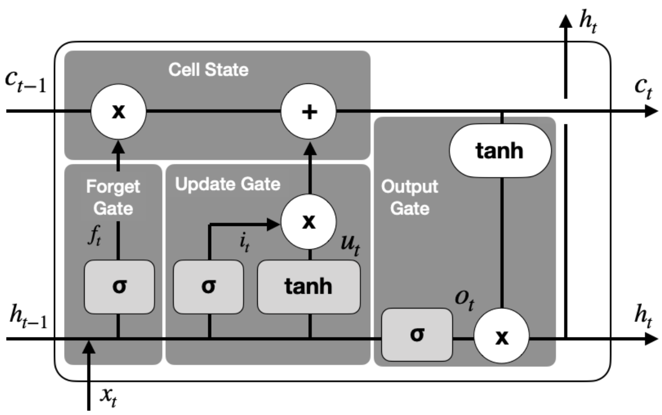

Rolling Sharpe Ratio (3 Months)
Sharpe
Average
Sharpe ratio
4.00
3.00
2.00
1.00
0.00
-1.00
-2.00
-3.00
2014-07
2015-01
2015-07
2016-01
2016-07
2017-01
2017-07
2018-01

Chapter 19: RNNs for Multivariate Time Series and Sentiment Analysis









LSTM Equations

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$

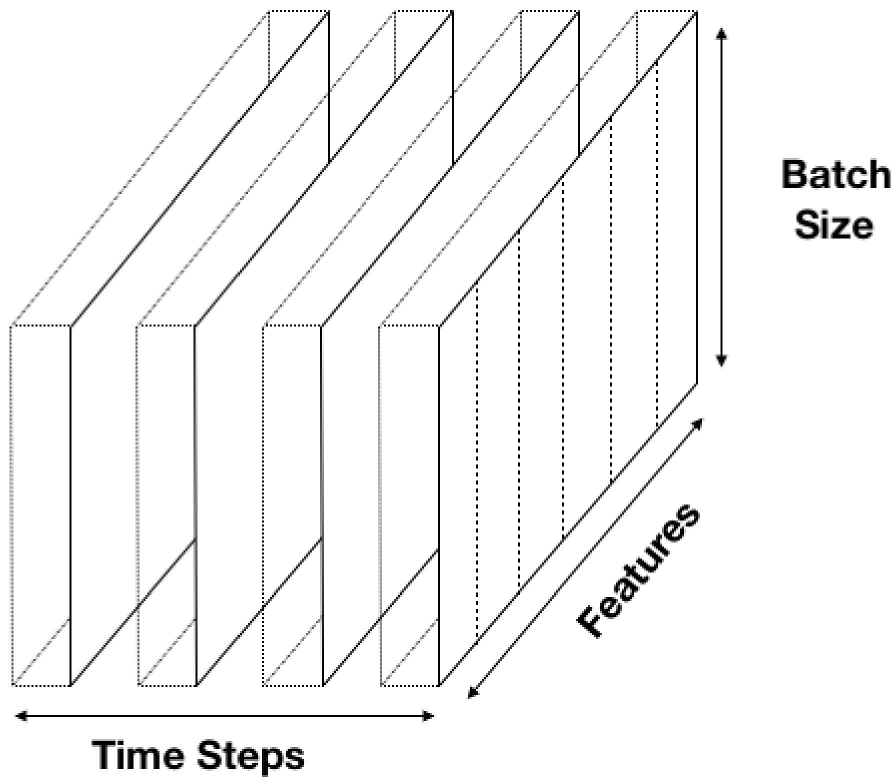
$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$

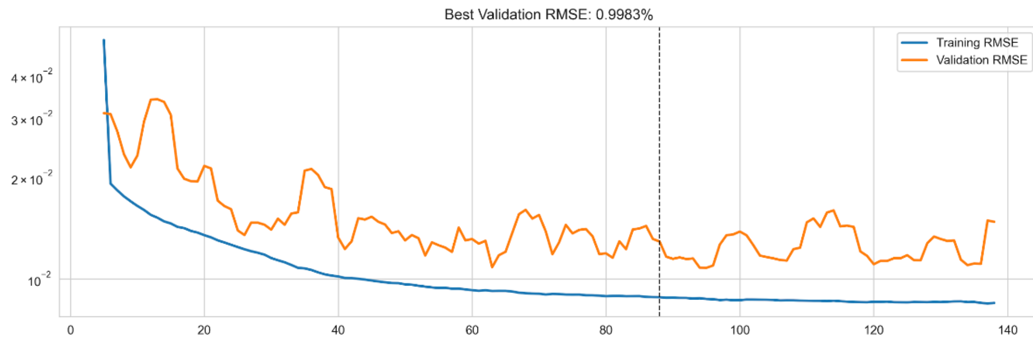
$$u_t = \tanh(W_c x_t + U_c h_{t-1} + b_c)$$

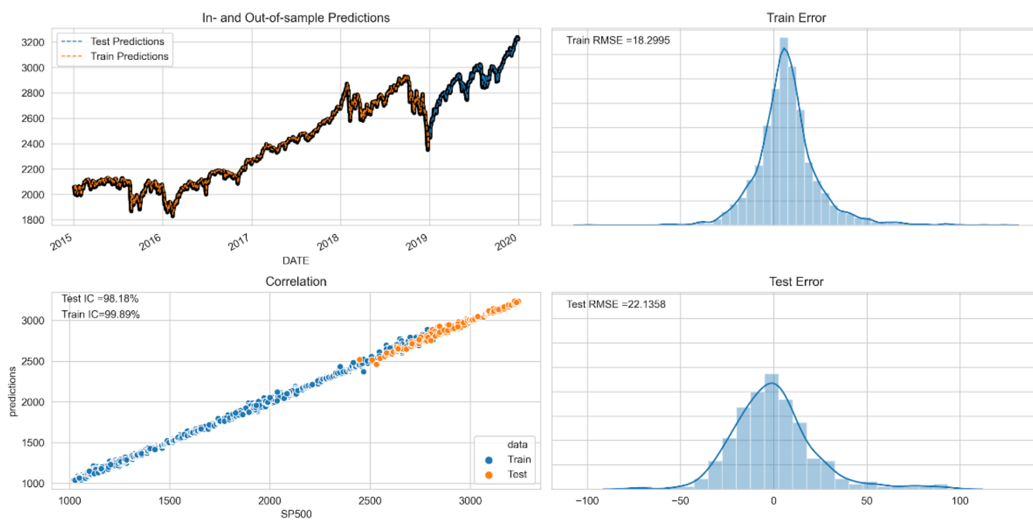
$$c_t = f_t \odot c_{t-1} + i_t \odot u_t$$

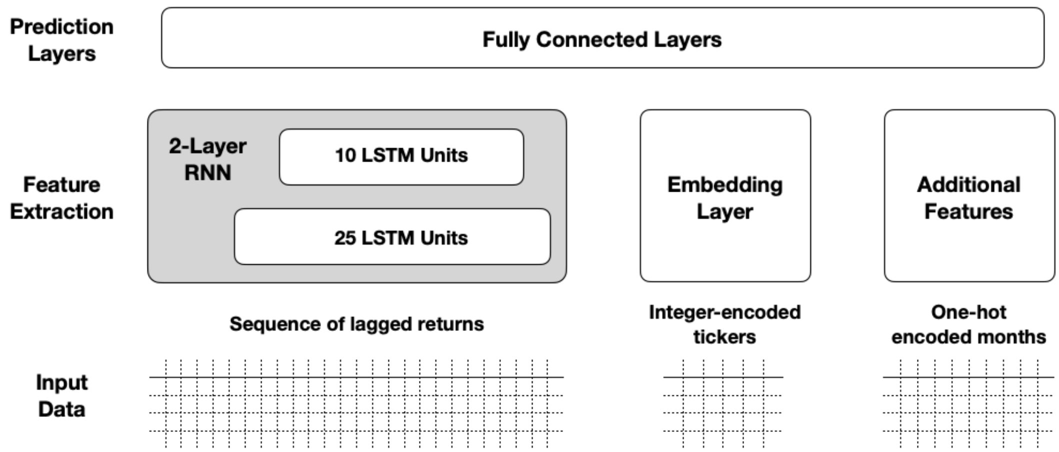
$$h_t = o_t \odot \tanh c_t$$

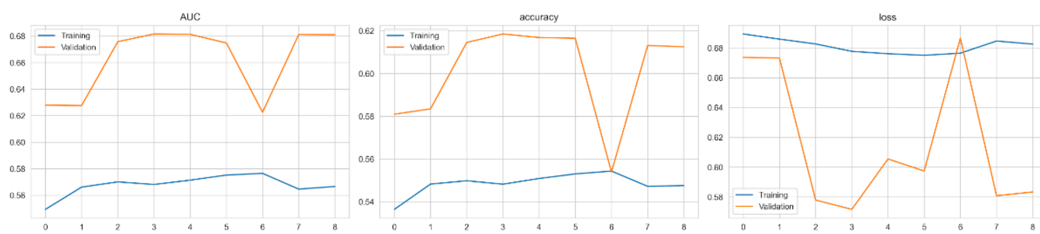


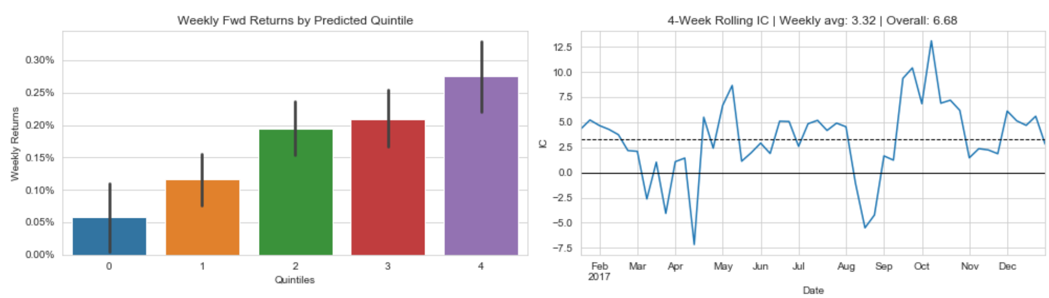
Input	Output
$\langle x_1, x_2, x_3, x_4, x_5 \rangle$	x_6
$\langle x_2, x_3, x_4, x_5, x_6 \rangle$	x_7
$\vdots$	$\vdots$
$\langle x_{T-5}, x_{T-4}, x_{T-3}, x_{T-2}, x_{T-1} \rangle$	x_T

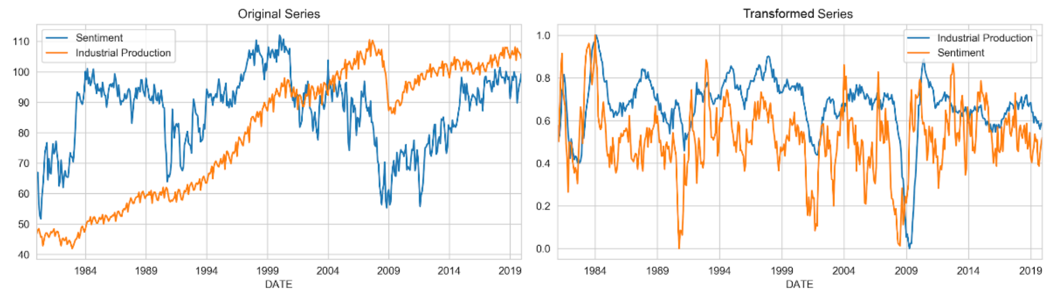

Best Validation RMSE: 0.9983%
Training RMSE
Validation RMSE
4×10^{-2}
3×10^{-2}
2×10^{-2}
10^{-2}
0
20
40
60
80
100
120
140


In- and Out-of-sample Predictions
Test Predictions
Train Predictions
DATE
Train Error
Train RMSE =18.2995
Correlation
Test IC =98.18%
Train IC=99.89%
predictions
SP500
data
Train
Test
Test Error
Test RMSE =22.1358

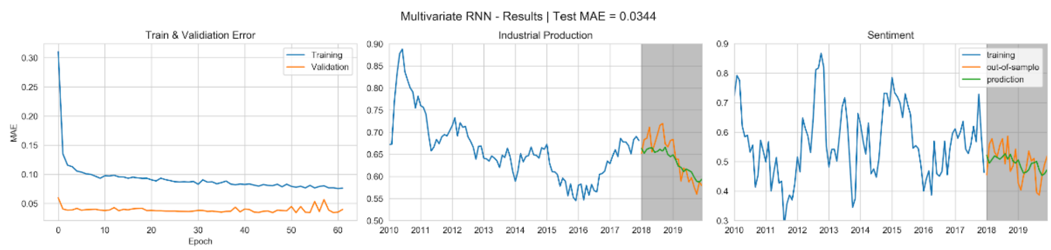

Prediction Layers
Fully Connected Layers
Feature Extraction
2-Layer RNN
10 LSTM Units
25 LSTM Units
Embedding Layer
Additional Features
Sequence of lagged returns
Integer-encoded tickers
One-hot encoded months
Input Data


AUC
Training
Validation
accuracy
Training
Validation
loss
Training
Validation

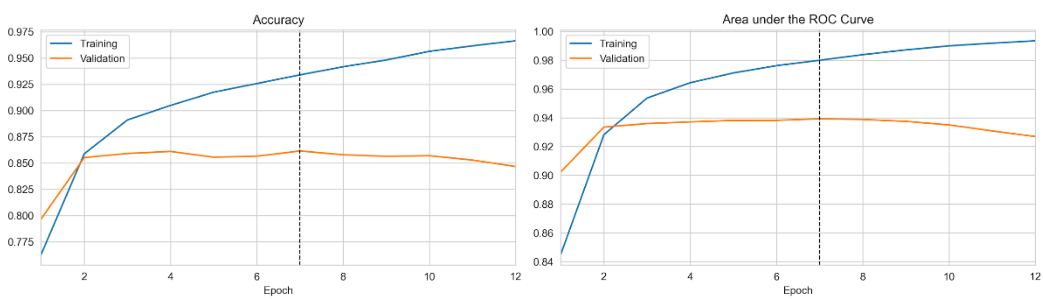

Weekly Fwd Returns by Predicted Quintile
Weekly Returns
Quintiles
4-Week Rolling IC | Weekly avg: 3.32 | Overall: 6.68
IC
Date



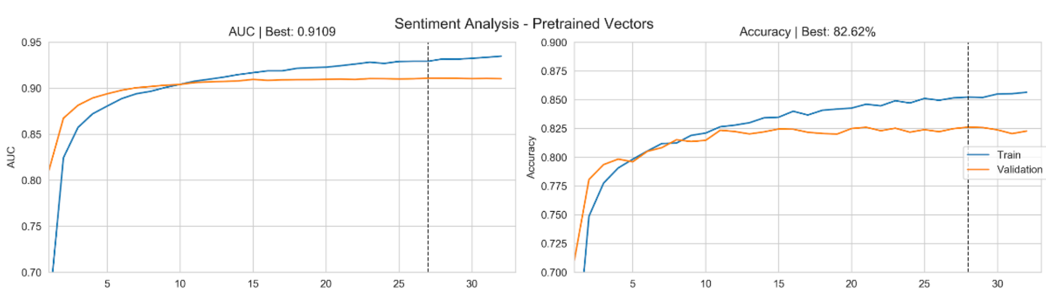
Original Series
Sentiment
Industrial Production
DATE
Transformed Series
Industrial Production
Sentiment
DATE



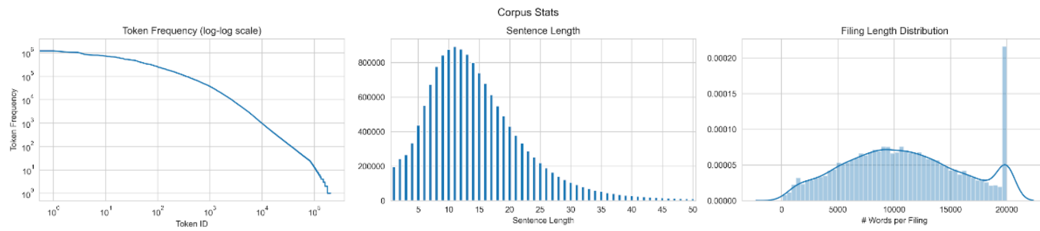
Multivariate RNN - Results | Test MAE = 0.0344
Train & Validiation Error
Training
Validation
MAE
Epoch
Industrial Production
Sentiment
training
out-of-sample
prediction

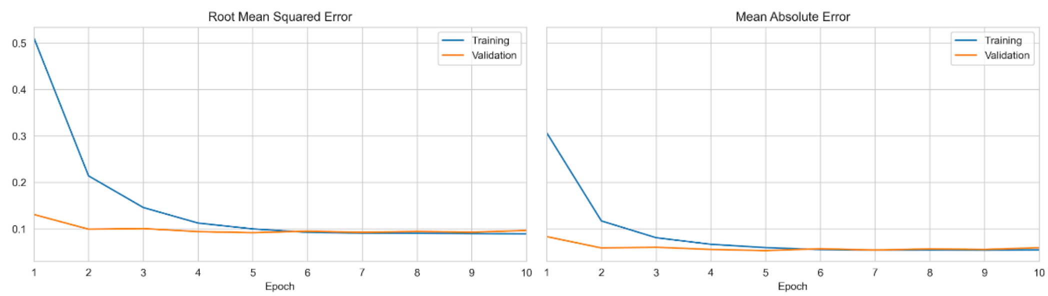


Accuracy
Training
Validation
Epoch
Area under the ROC Curve
Training
Validation
Epoch

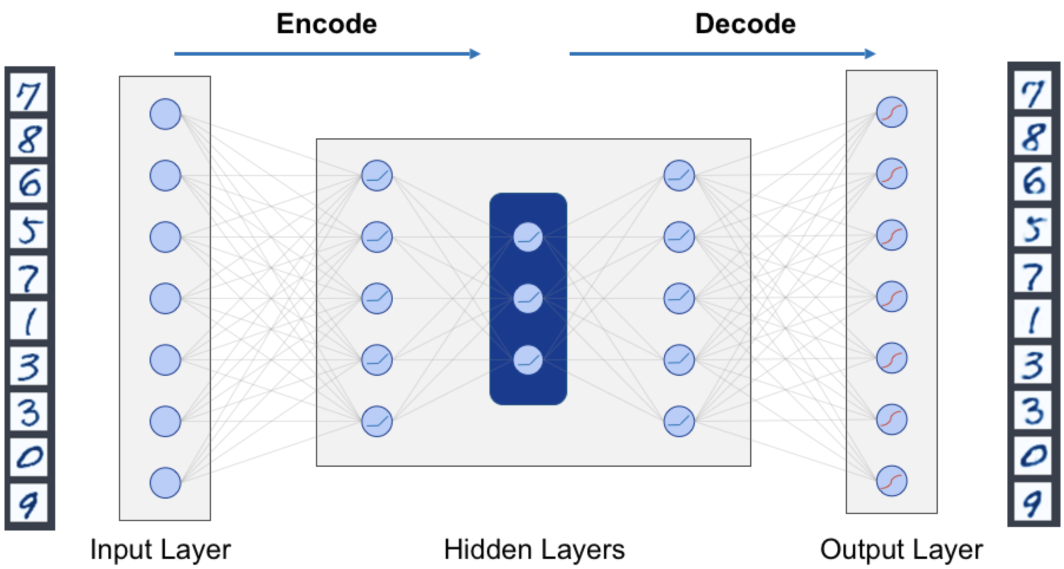


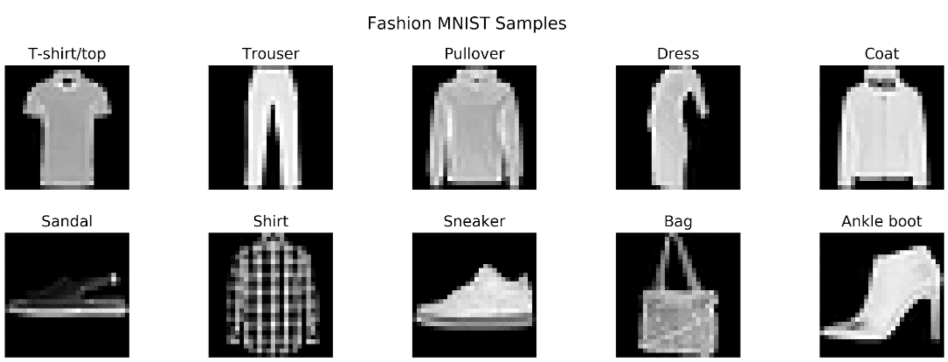
Sentiment Analysis - Pretrained Vectors
AUC | Best: 0.9109
AUC
Epoch
Accuracy | Best: 82.62%
Accuracy
Train
Validation
Epoch

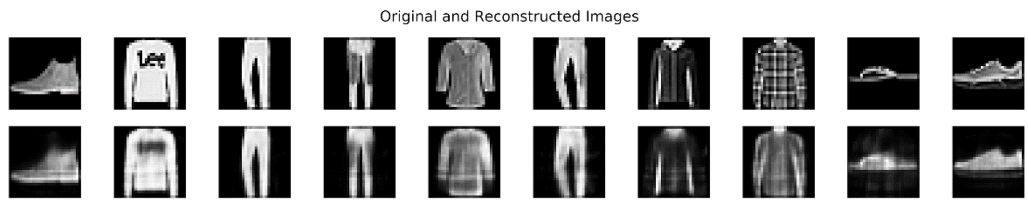

Corpus Stats
Token Frequency (log-log scale)
Token Frequency
Token ID
Sentence Length
Sentence Length
Filing Length Distribution
Words per Filing

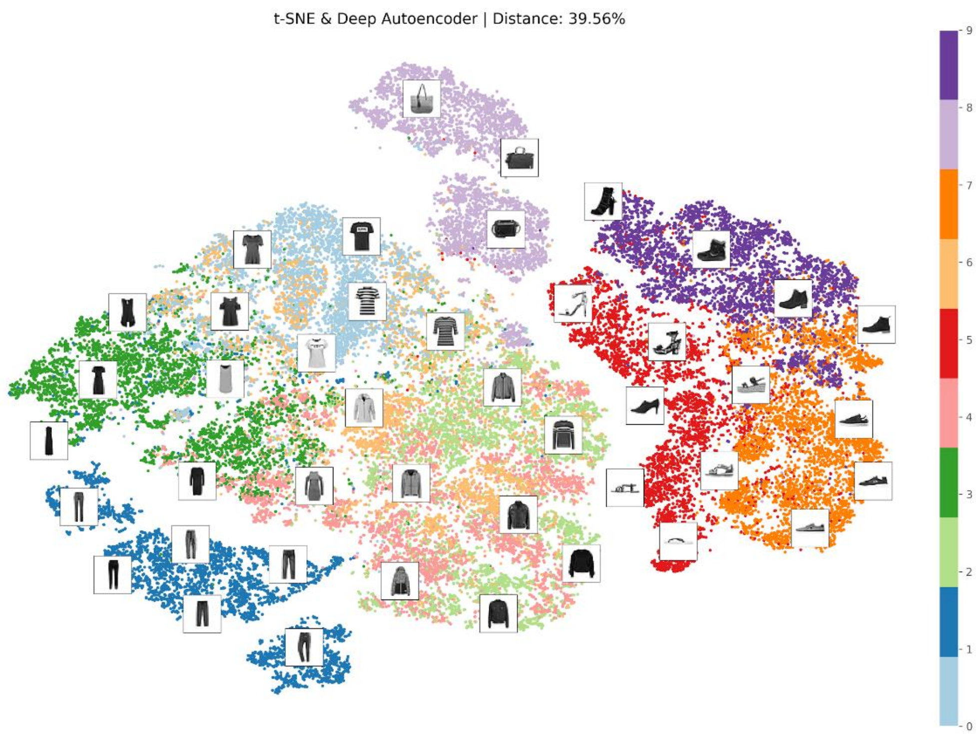

Root Mean Squared Error
Training
Validation
Epoch
Mean Absolute Error
Training
Validation
Epoch

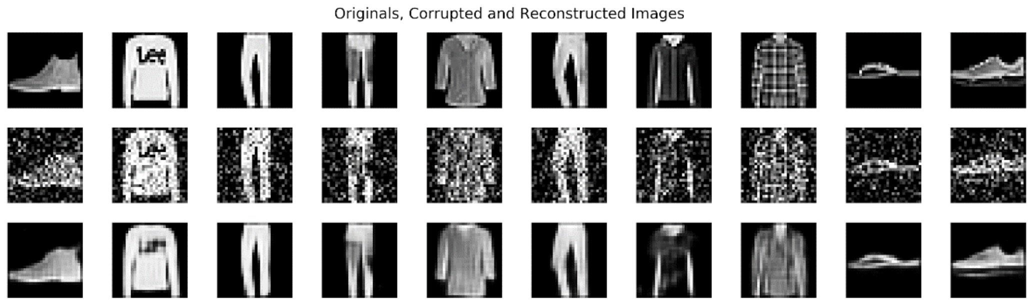
Chapter 20: Autoencoders for Conditional Risk Factors and Asset Pricing

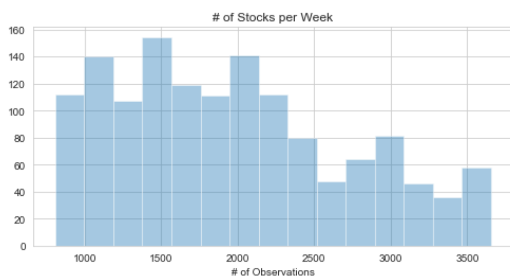


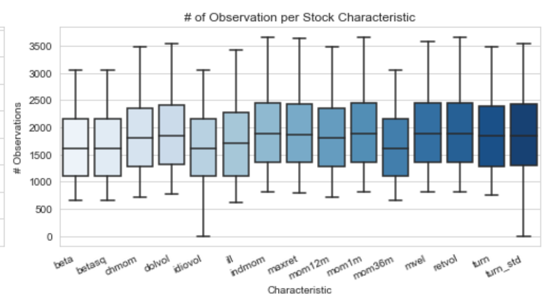


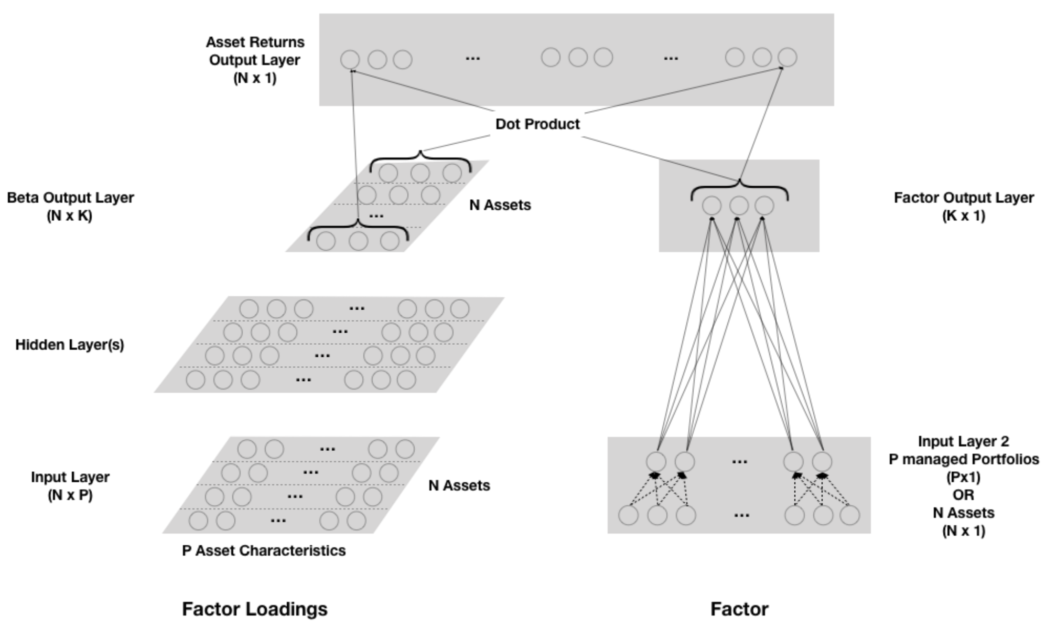

Original and Reconstructed Images

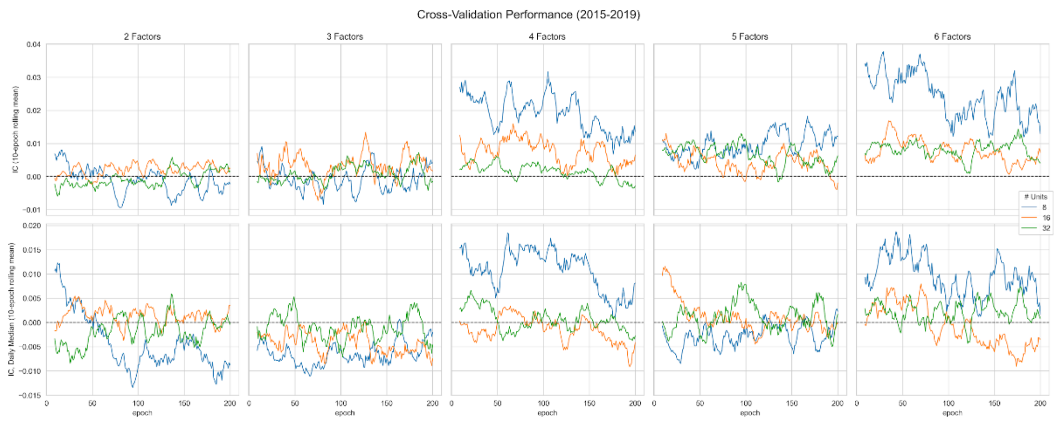

t-SNE & Deep Autoencoder | Distance: 39.56%
9
8
7
6
5
4
3
2
1
0

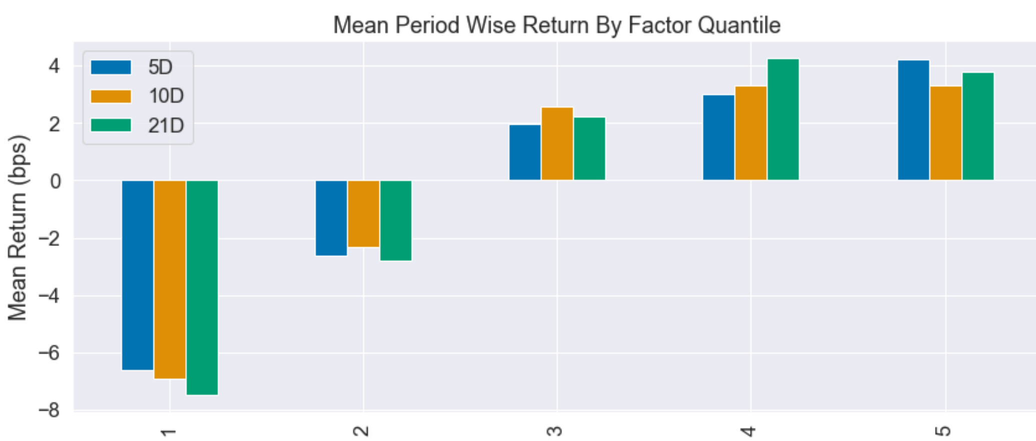

Originals, Corrupted and Reconstructed Images

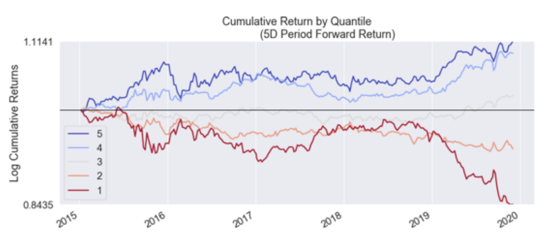

# of Stocks per Week
# of Observations
0
20
40
60
80
100
120
140
160
1000
1500
2000
2500
3000
3500


# of Observation per Stock Characteristic
Observations
Characteristic
0
500
1000
1500
2000
2500
3000
3500
beta
betasq
chmom
dolvol
idiovol
ill
indmom
maxret
mom12m
mom1m
mom36m
mvel
retvol
turn
turn_std


Asset Returns
Output Layer
(N x 1)
Dot Product
Beta Output Layer
(N x K)
N Assets
Factor Output Layer
(K x 1)
Hidden Layer(s)
Input Layer
(N x P)
N Assets
P Asset Characteristics
Input Layer 2
P managed Portfolios
(Px1)
OR
N Assets
(N x 1)
Factor Loadings
Factor

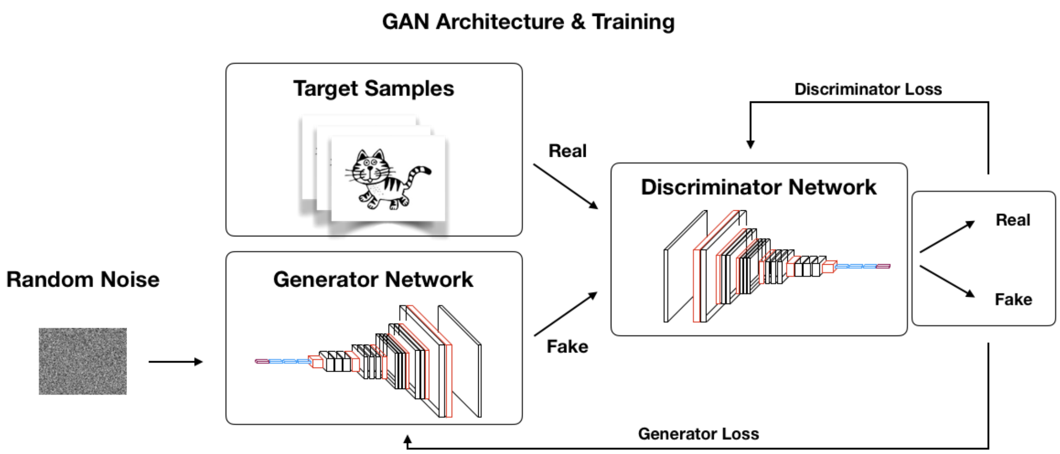

Cross-Validation Performance (2015-2019)
2 Factors
3 Factors
4 Factors
5 Factors
6 Factors
epoch

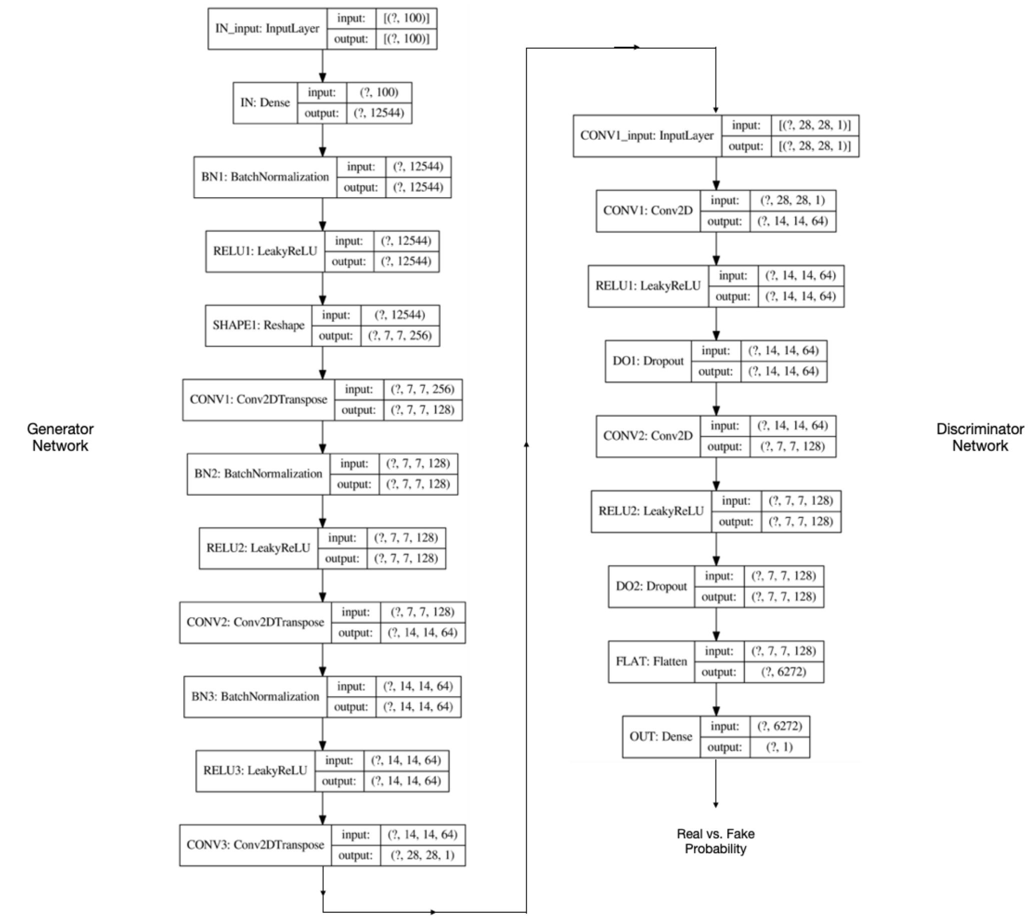

Mean Period Wise Return By Factor Quantile
5D
10D
21D
Mean Return (bps)
4
2
0
-2
-4
-6
-8
1
2
3
4
5

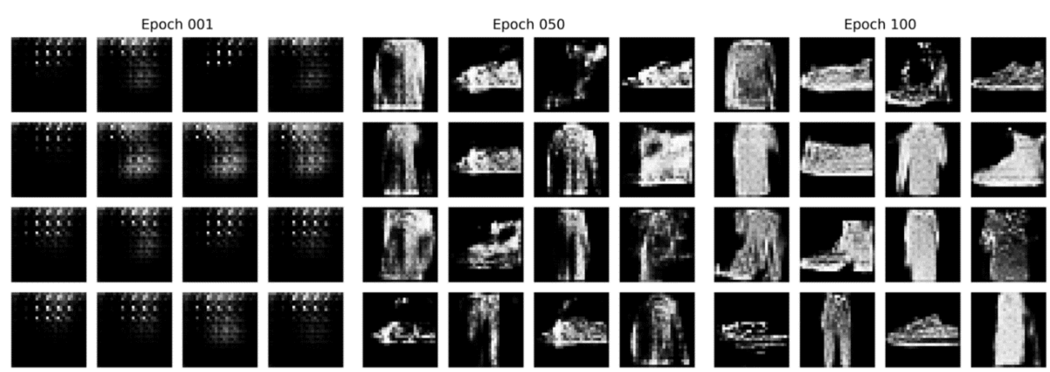

Cumulative Return by Quantile
(5D Period Forward Return)
Log Cumulative Returns
1.1141
0.8435
5
4
3
2
1
2015
2016
2017
2018
2019
2020

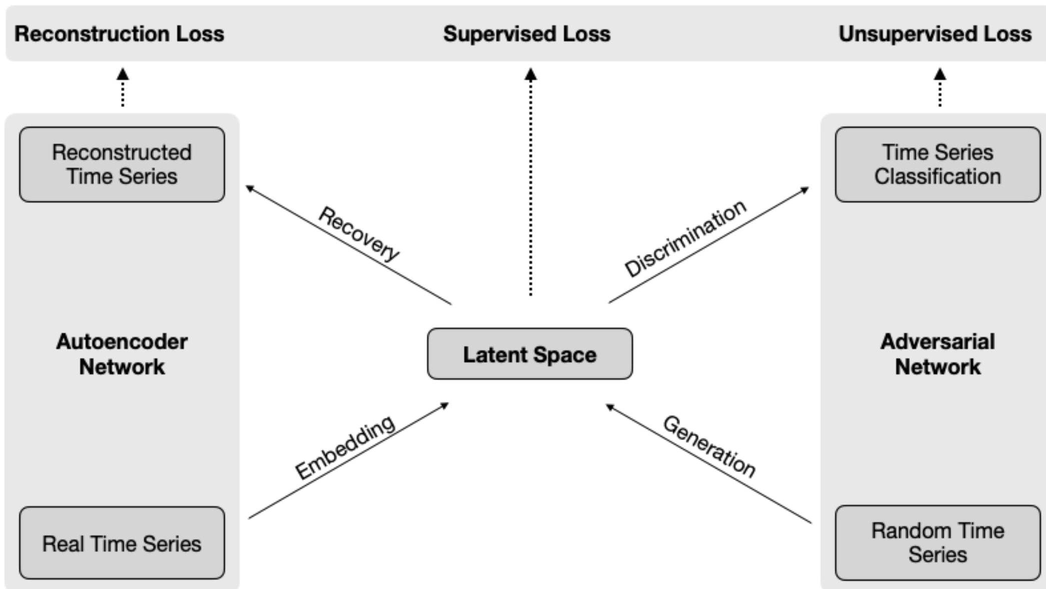

Factor Weighted Long/Short Portfolio Cumulative Return (5D Period)
Cumulative Returns
1.12
1.10
1.08
1.06
1.04
1.02
1.00
2015
2016
2017
2018
2019
2020

Chapter 21: Generative Adversarial Networks for Synthetic Time-Series Data

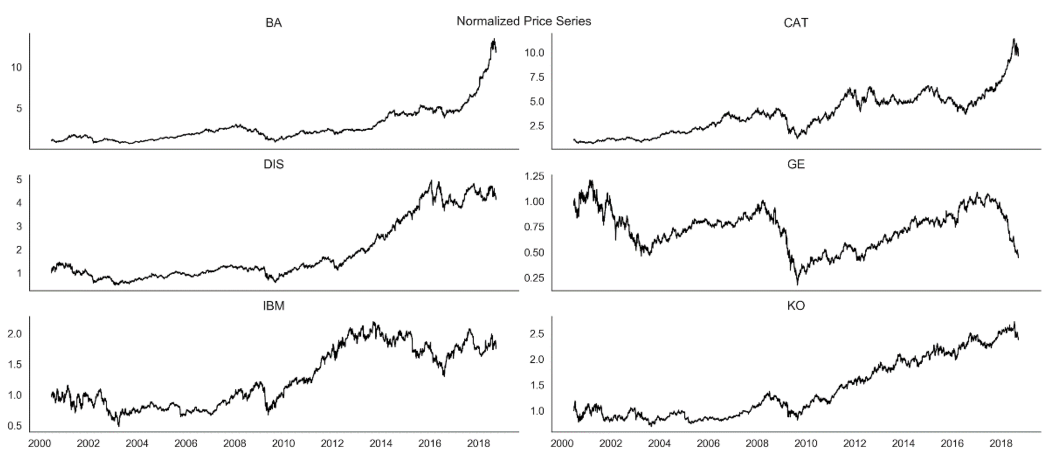



Generator
Network
IN_input: InputLayer | input: [(?, 100)] | output: [(?, 100)]
IN: Dense | input: (?, 100) | output: (?, 12544)
BN1: BatchNormalization | input: (?, 12544) | output: (?, 12544)
RELU1: LeakyReLU | input: (?, 12544) | output: (?, 12544)
SHAPE1: Reshape | input: (?, 12544) | output: (?, 7, 7, 256)
CONV1: Conv2DTranspose | input: (?, 7, 7, 256) | output: (?, 7, 7, 128)
BN2: BatchNormalization | input: (?, 7, 7, 128) | output: (?, 7, 7, 128)
RELU2: LeakyReLU | input: (?, 7, 7, 128) | output: (?, 7, 7, 128)
CONV2: Conv2DTranspose | input: (?, 7, 7, 128) | output: (?, 14, 14, 64)
BN3: BatchNormalization | input: (?, 14, 14, 64) | output: (?, 14, 14, 64)
RELU3: LeakyReLU | input: (?, 14, 14, 64) | output: (?, 14, 14, 64)
CONV3: Conv2DTranspose | input: (?, 14, 14, 64) | output: (?, 28, 28, 1)
Discriminator
Network
CONV1_input: InputLayer | input: [(?, 28, 28, 1)] | output: [(?, 28, 28, 1)]
CONV1: Conv2D | input: (?, 28, 28, 1) | output: (?, 14, 14, 64)
RELU1: LeakyReLU | input: (?, 14, 14, 64) | output: (?, 14, 14, 64)
DO1: Dropout | input: (?, 14, 14, 64) | output: (?, 14, 14, 64)
CONV2: Conv2D | input: (?, 14, 14, 64) | output: (?, 7, 7, 128)
RELU2: LeakyReLU | input: (?, 7, 7, 128) | output: (?, 7, 7, 128)
DO2: Dropout | input: (?, 7, 7, 128) | output: (?, 7, 7, 128)
FLAT: Flatten | input: (?, 7, 7, 128) | output: (?, 6272)
OUT: Dense | input: (?, 6272) | output: (?, 1)
Real vs. Fake
Probability

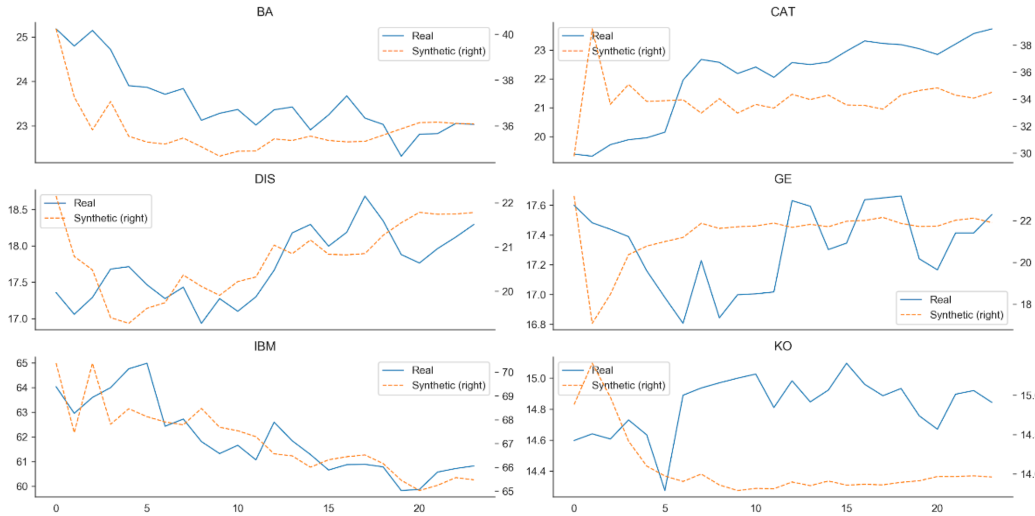

Epoch 001
Epoch 050
Epoch 100

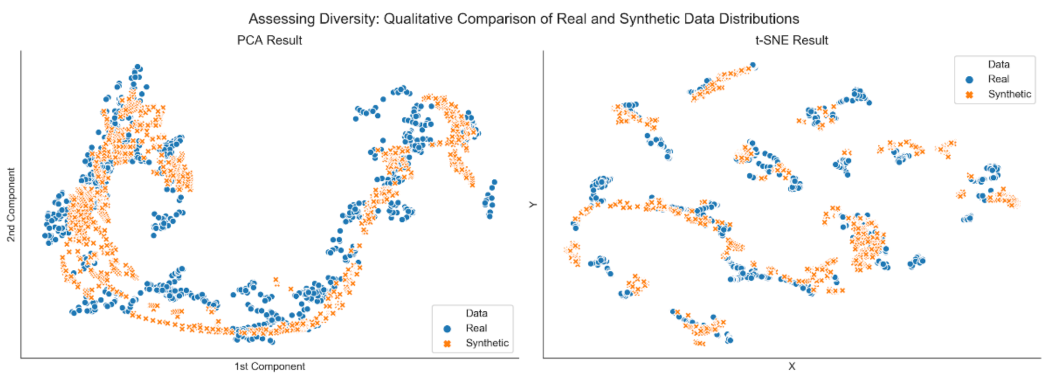


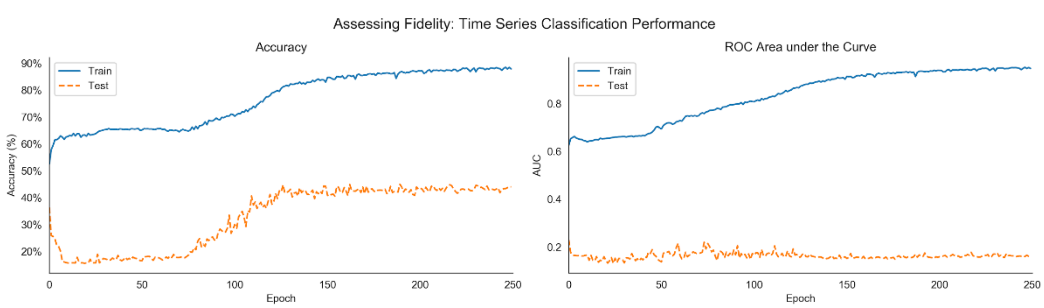
Reconstruction Loss
Supervised Loss
Unsupervised Loss
Reconstructed Time Series
Time Series Classification
Recovery
Discrimination
Autoencoder Network
Latent Space
Adversarial Network
Embedding
Generation
Real Time Series
Random Time Series

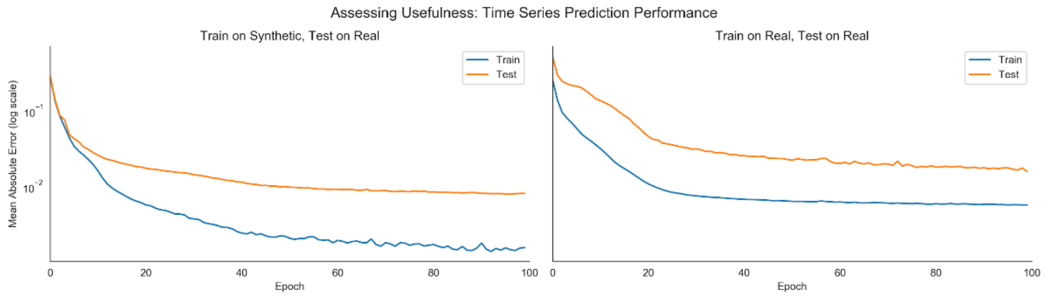


Normalized Price Series
BA
CAT
DIS
GE
IBM
KO


BA
Real
Synthetic (right)
CAT
DIS
GE
IBM
KO
0
5
10
15
20

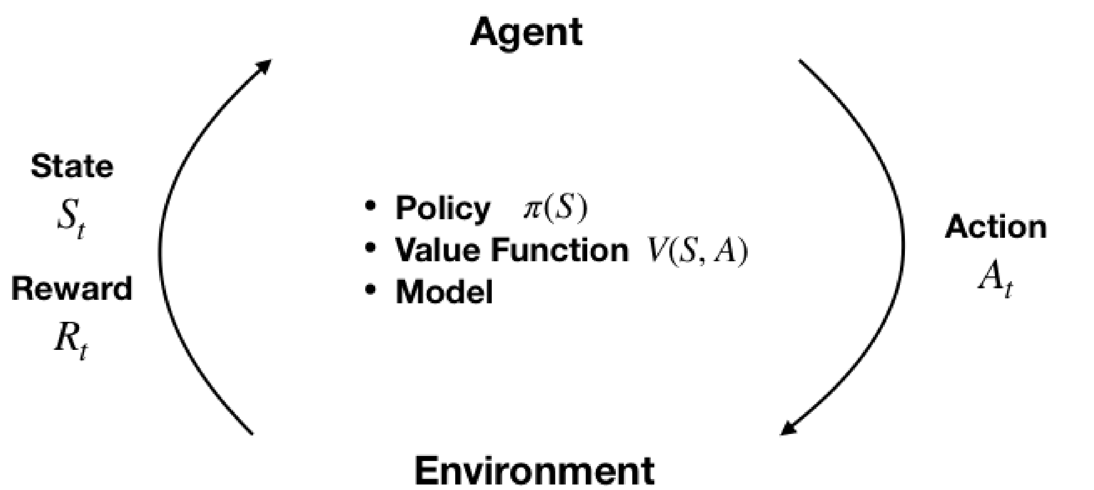

Assessing Diversity: Qualitative Comparison of Real and Synthetic Data Distributions
PCA Result
t-SNE Result
2nd Component
1st Component
Y
X
Data
Real
Synthetic

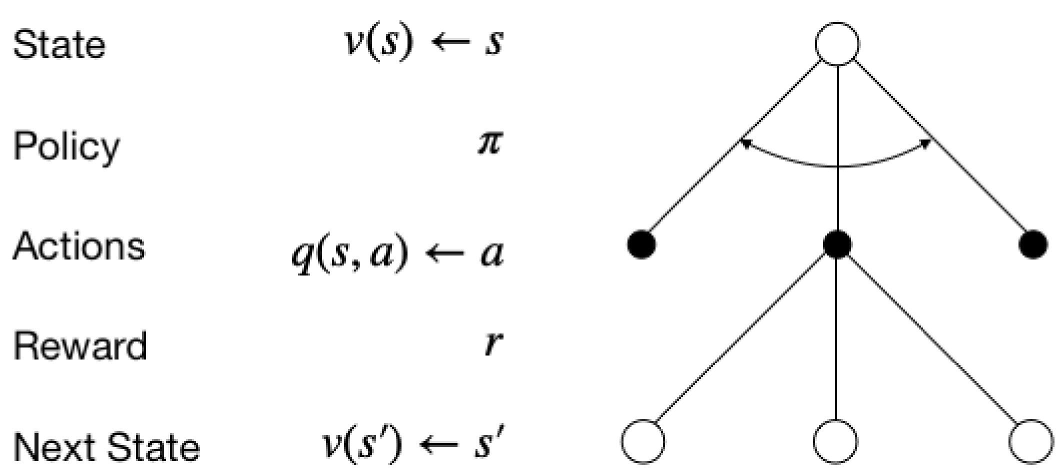

Assessing Fidelity: Time Series Classification Performance
Accuracy
ROC Area under the Curve
Train
Test
Accuracy (%)
AUC
Epoch

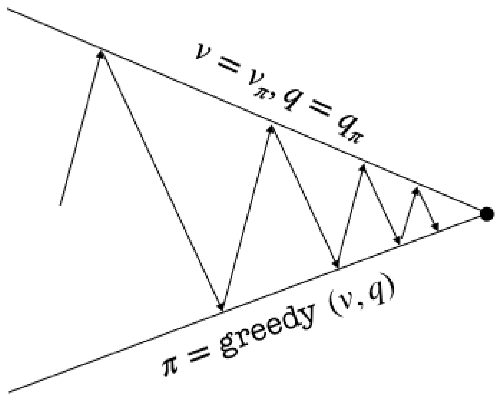


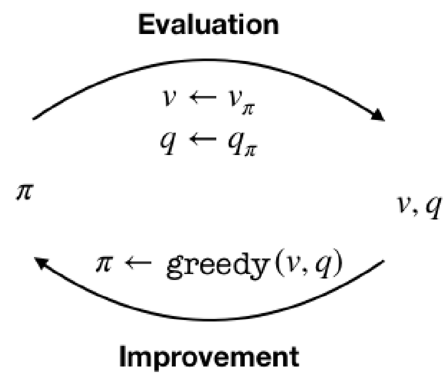
Assessing Usefulness: Time Series Prediction Performance
Train on Synthetic, Test on Real
Train on Real, Test on Real
Train
Test
Mean Absolute Error (log scale)
10^{-1}
10^{-2}
0
20
40
60
80
100
Epoch

Chapter 22: Deep Reinforcement Learning – Building a Trading Agent





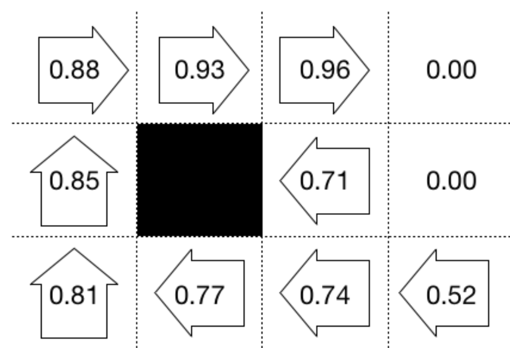


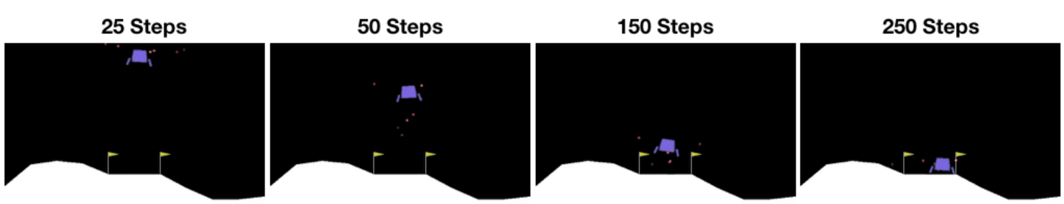


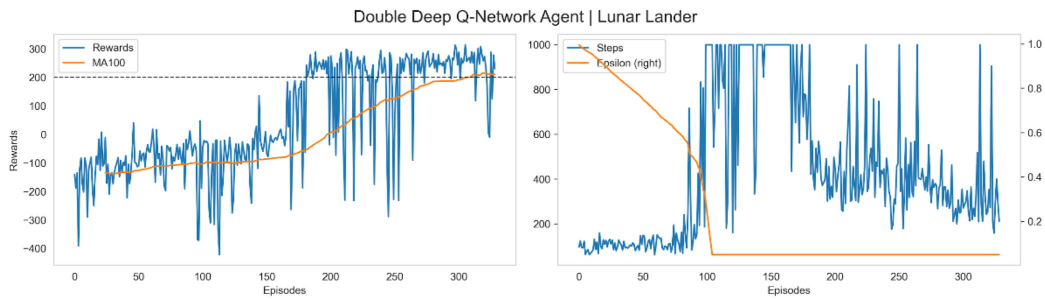
Rewards

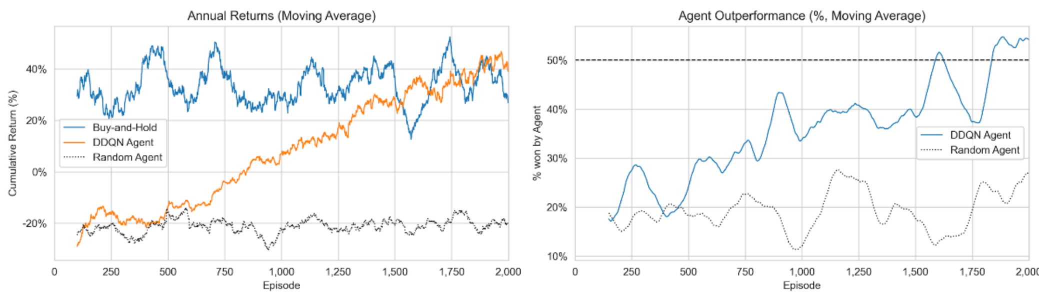
-0.02	-0.02	-0.02	1
-0.02		-0.02	-1
-0.02	-0.02	-0.02	-0.02

Optimal Values & Policy ($\gamma = .99$)

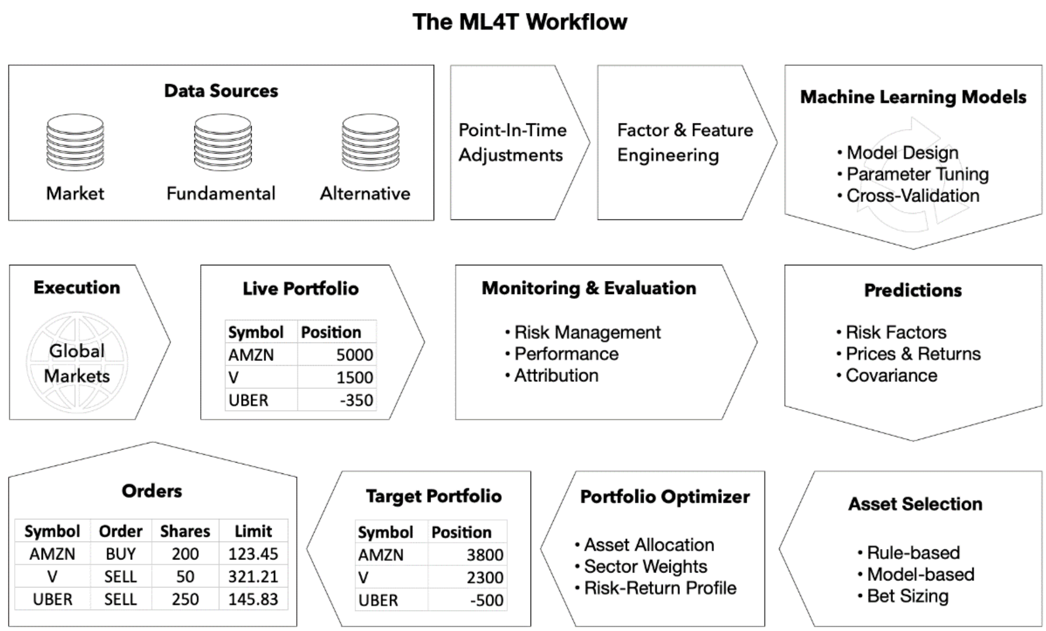


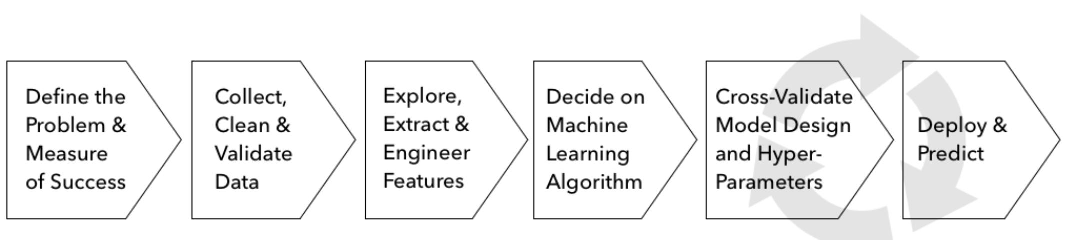



Double Deep Q-Network Agent | Lunar Lander
Rewards
MA100
Rewards
Episodes
Steps
Epsilon (right)
Episodes

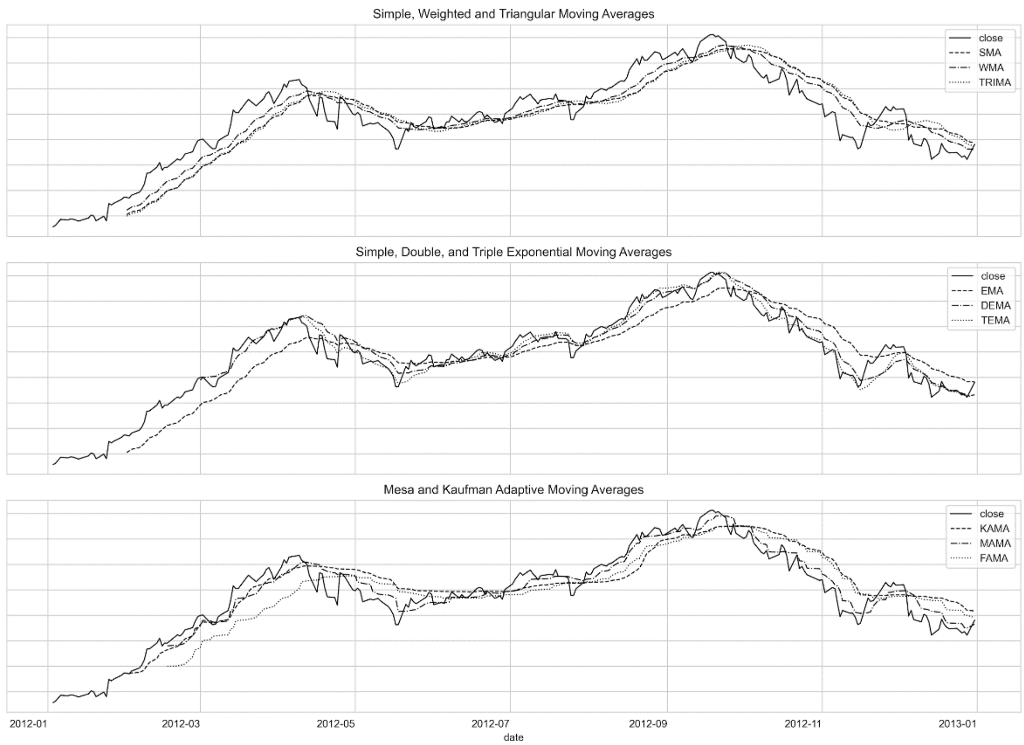

Annual Returns (Moving Average)
Buy-and-Hold
DDQN Agent
Random Agent
Cumulative Return (%)
Episode
Agent Outperformance (%, Moving Average)
DDQN Agent
Random Agent
% won by Agent
Episode

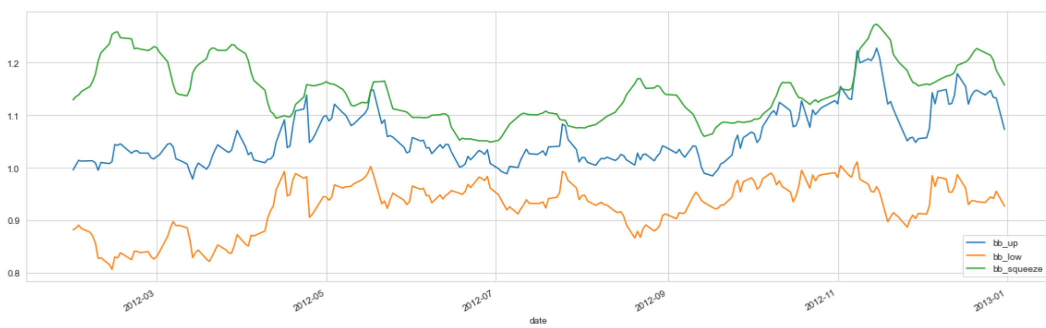
Chapter 23: Conclusions and Next Steps



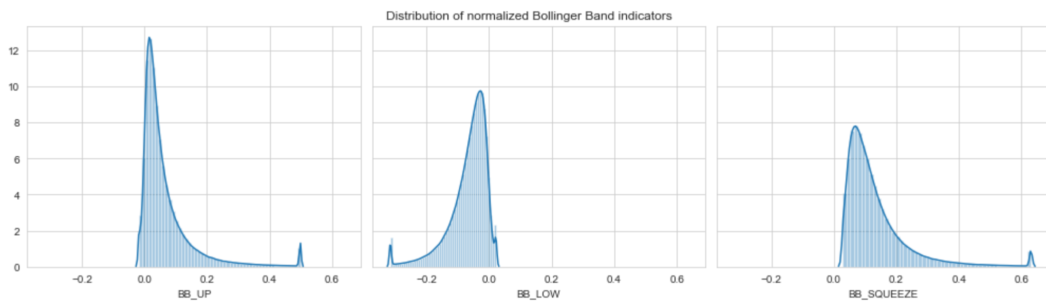


Appendix: Alpha Factor Library

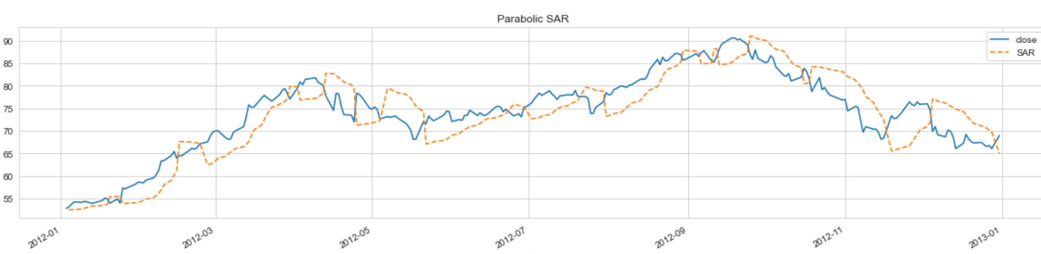




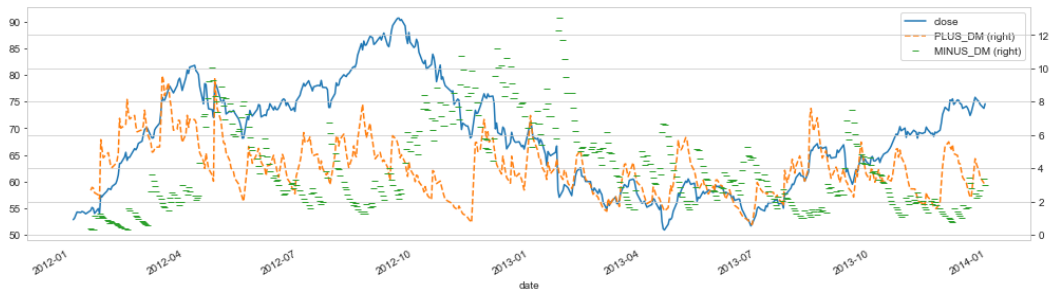
bb_up
bb_low
bb_squeeze
date

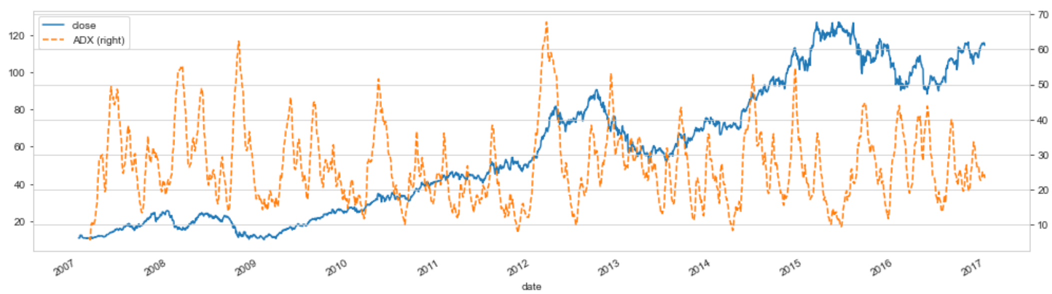


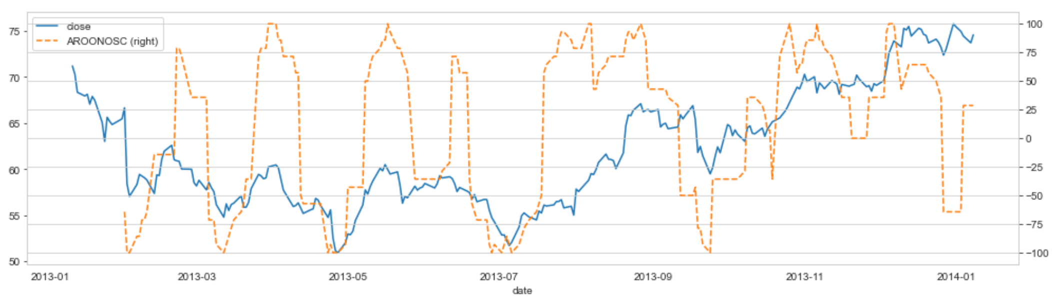
Distribution of normalized Bollinger Band indicators
BB_UP
BB_LOW
BB_SQUEEZE

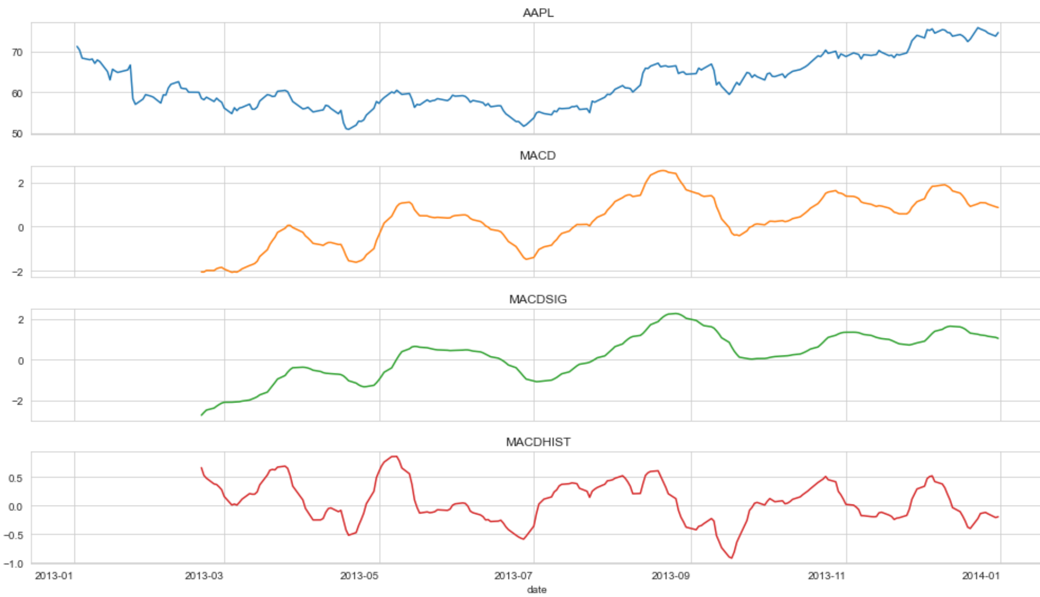


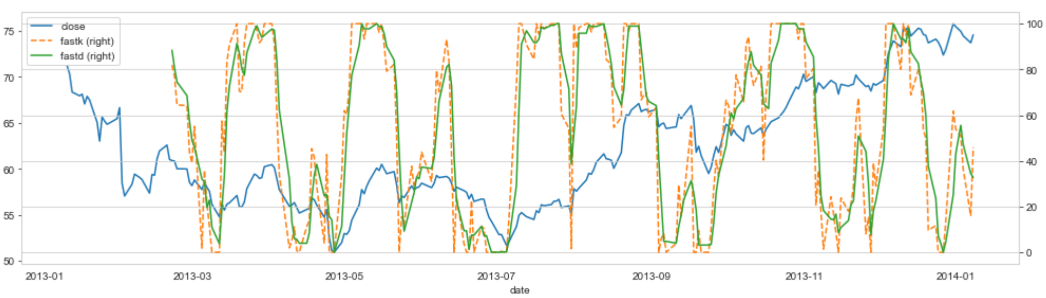
Parabolic SAR
close
SAR
date

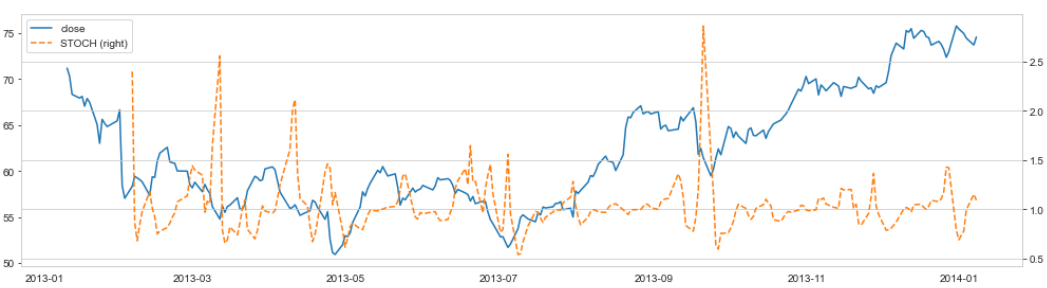

close
PLUS_DM (right)
MINUS_DM (right)
date

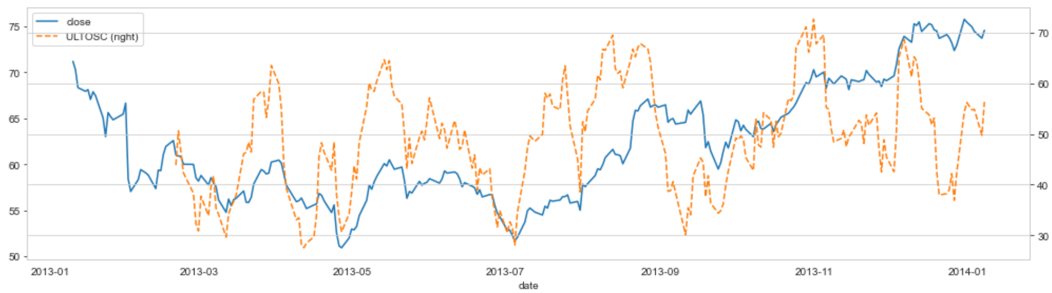

close
ADX (right)
date

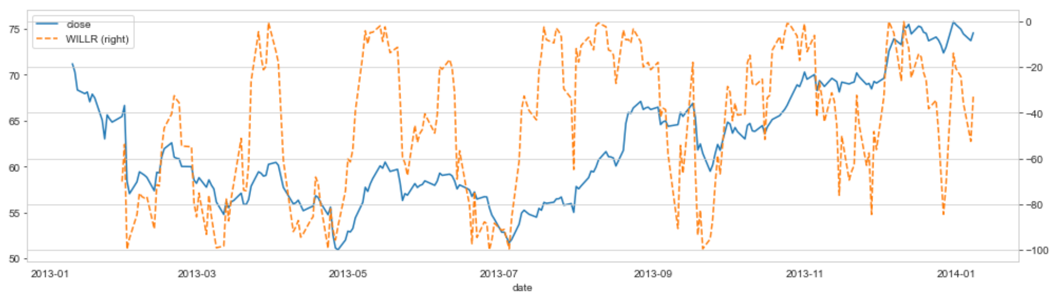

close
AROONOSC (right)
date

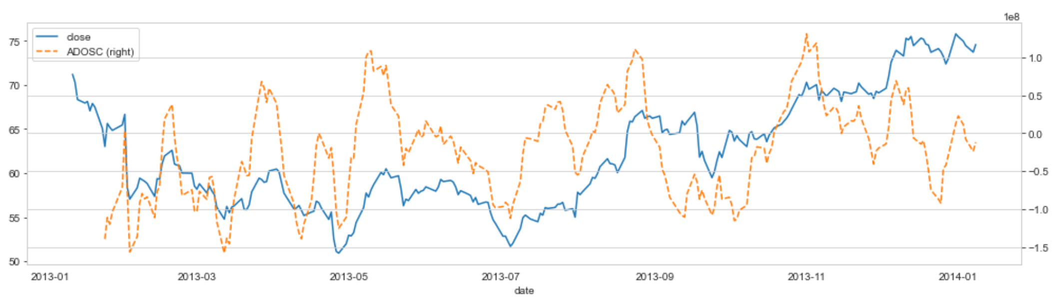

AAPL
MACD
MACDSIG
MACDHIST
2013-01
2013-03
2013-05
2013-07
2013-09
2013-11
2014-01
date

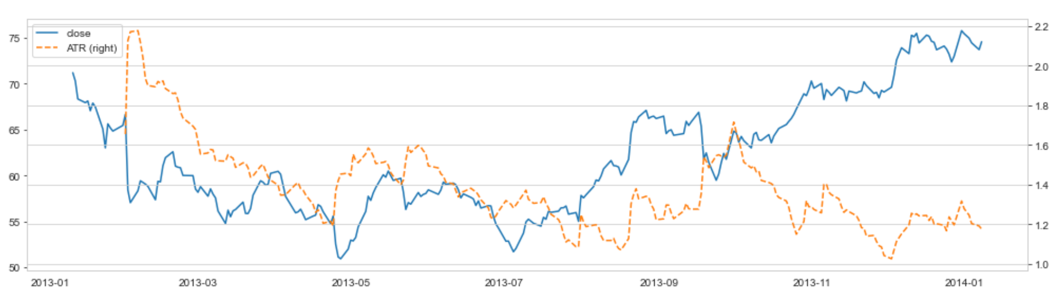

close
fastk (right)
fastd (right)
2013-01
2013-03
2013-05
2013-07
2013-09
2013-11
2014-01
date

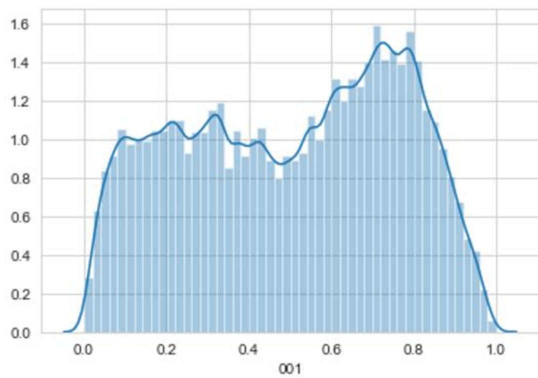

close
STOCH (right)
2013-01
2013-03
2013-05
2013-07
2013-09
2013-11
2014-01
date


close
ULTOSC (right)
date

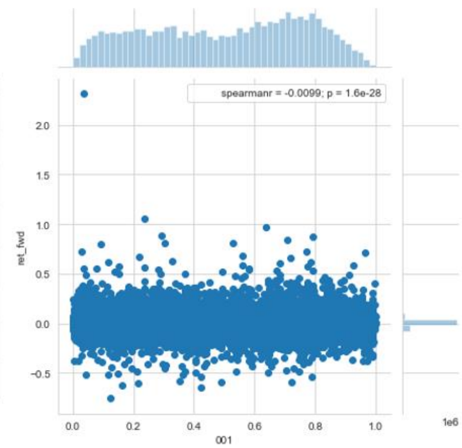

close
WILLR (right)
date


1e8
close
ADOSC (right)
date

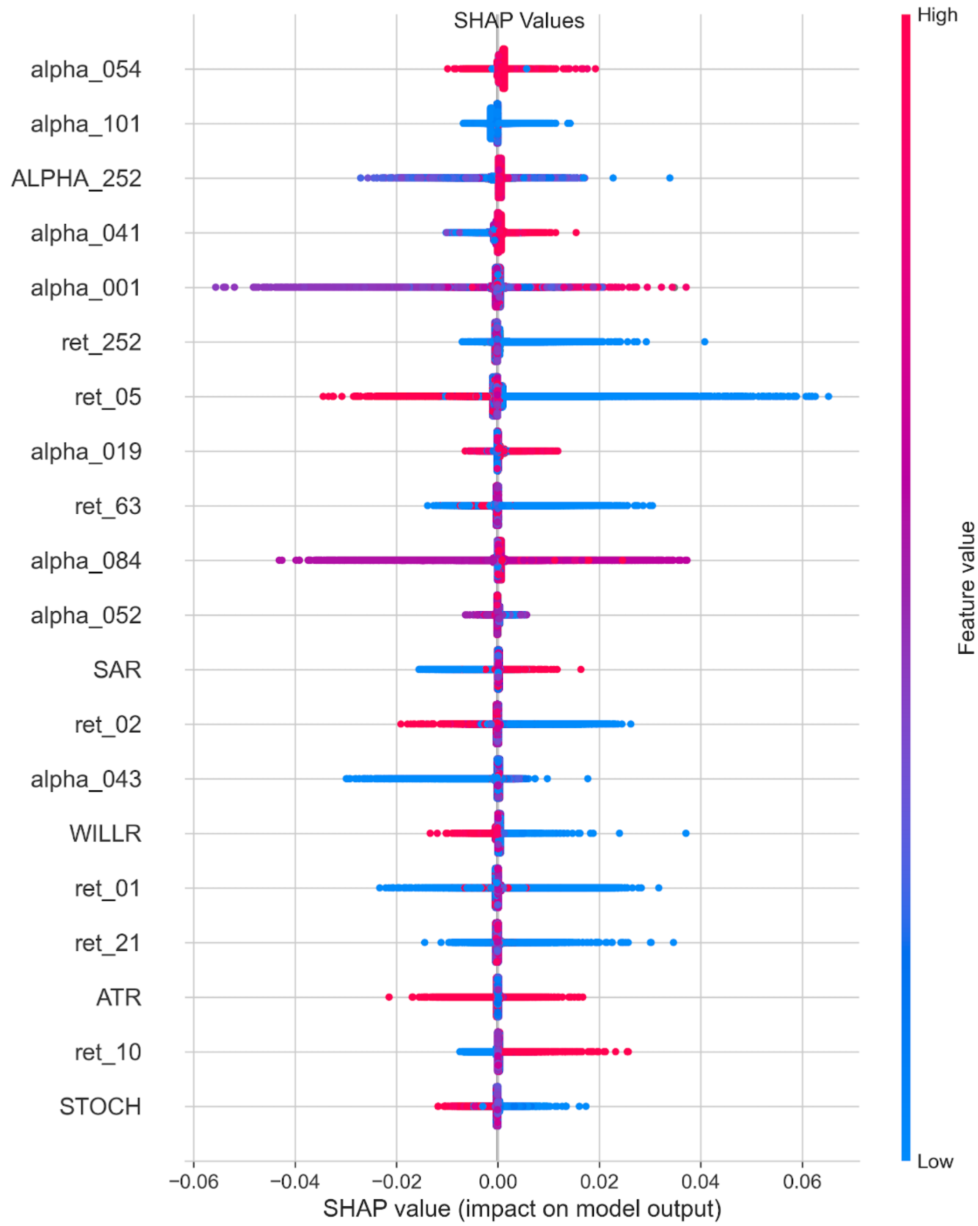

close
ATR (right)
date



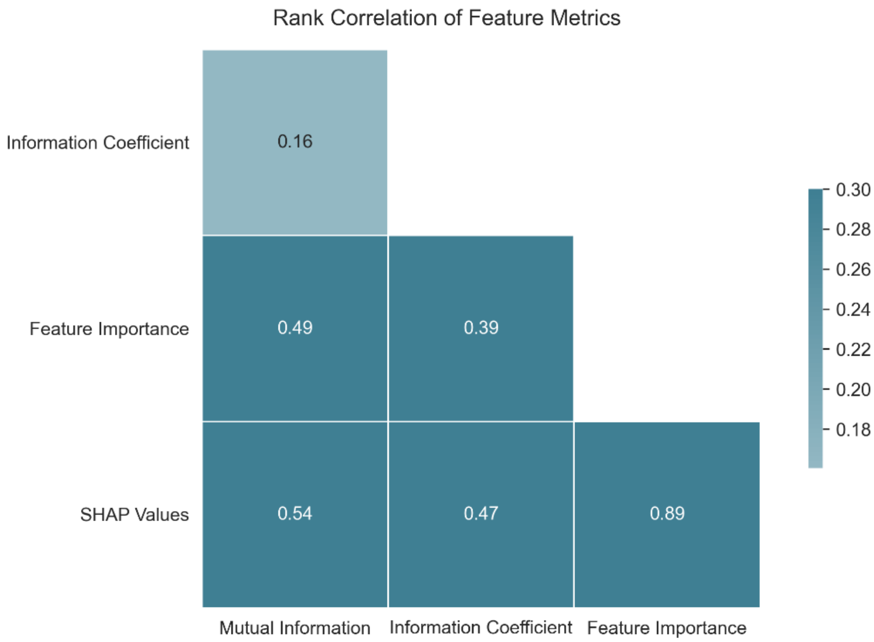
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
001



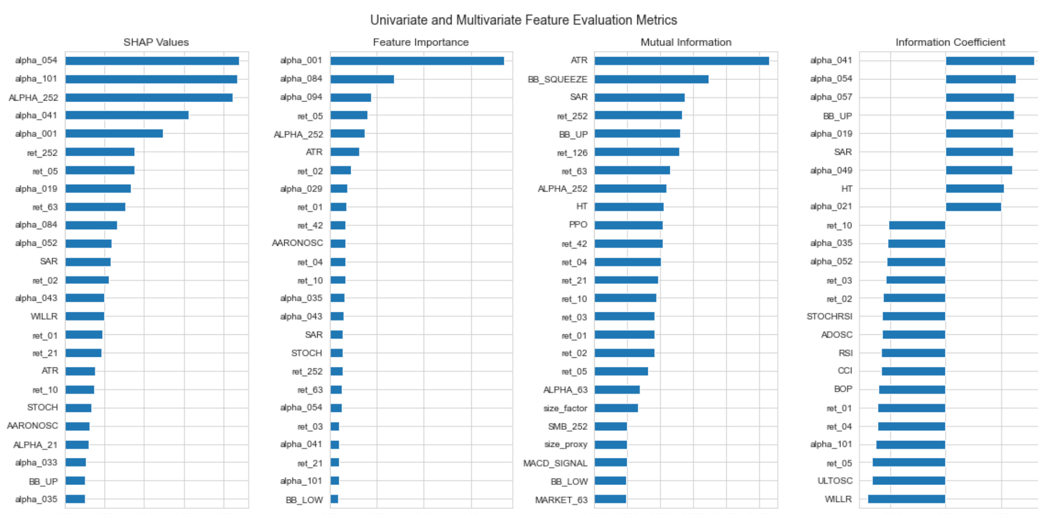
spearmanr = -0.0099; p = 1.6e-28
ret_fwd
2.0
1.5
1.0
0.5
0.0
-0.5
0.0
0.2
0.4
0.6
0.8
1.0
001
1e6



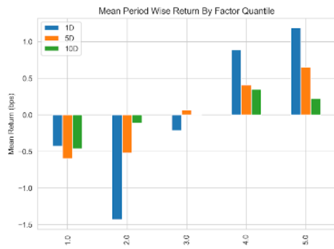
SHAP Values
alpha_054
alpha_101
ALPHA_252
alpha_041
alpha_001
ret_252
ret_05
alpha_019
ret_63
alpha_084
alpha_052
SAR
ret_02
alpha_043
WILLR
ret_01
ret_21
ATR
ret_10
STOCH
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
SHAP value (impact on model output)
High
Low
Feature value

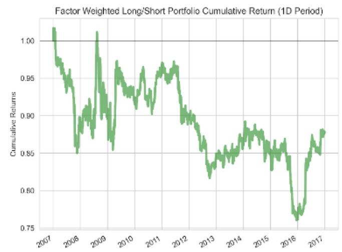


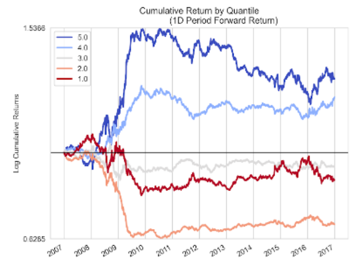
Rank Correlation of Feature Metrics
Information Coefficient
Feature Importance
SHAP Values
0.16
0.49
0.39
0.54
0.47
0.89
Mutual Information
Information Coefficient
Feature Importance
0.30
0.28
0.26
0.24
0.22
0.20
0.18



Univariate and Multivariate Feature Evaluation Metrics
SHAP Values
Feature Importance
Mutual Information
Information Coefficient


Mean Period Wise Return By Factor Quantile
1D
5D
10D
Mean Return (bps)
1.0
0.5
0.0
−0.5
−1.0
−1.5
1.0
2.0
3.0
4.0
5.0


Factor Weighted Long/Short Portfolio Cumulative Return (1D Period)
Cumulative Returns
1.00
0.95
0.90
0.85
0.80
0.75
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017


Cumulative Return by Quantile
(1D Period Forward Return)
5.0
4.0
3.0
2.0
1.0
1.5366
0.6265
Log Cumulative Returns
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017