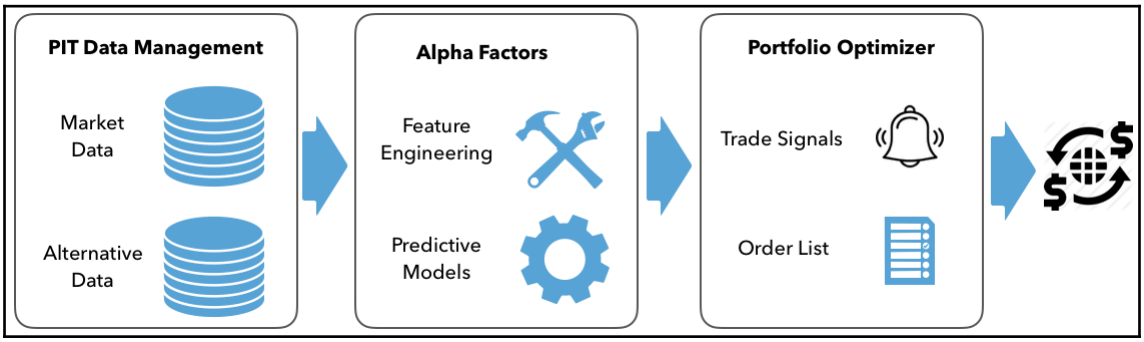
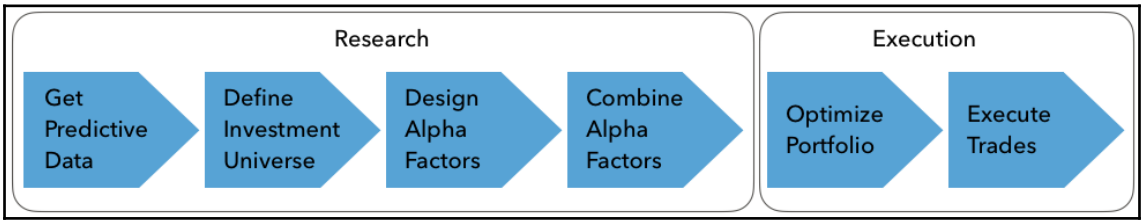
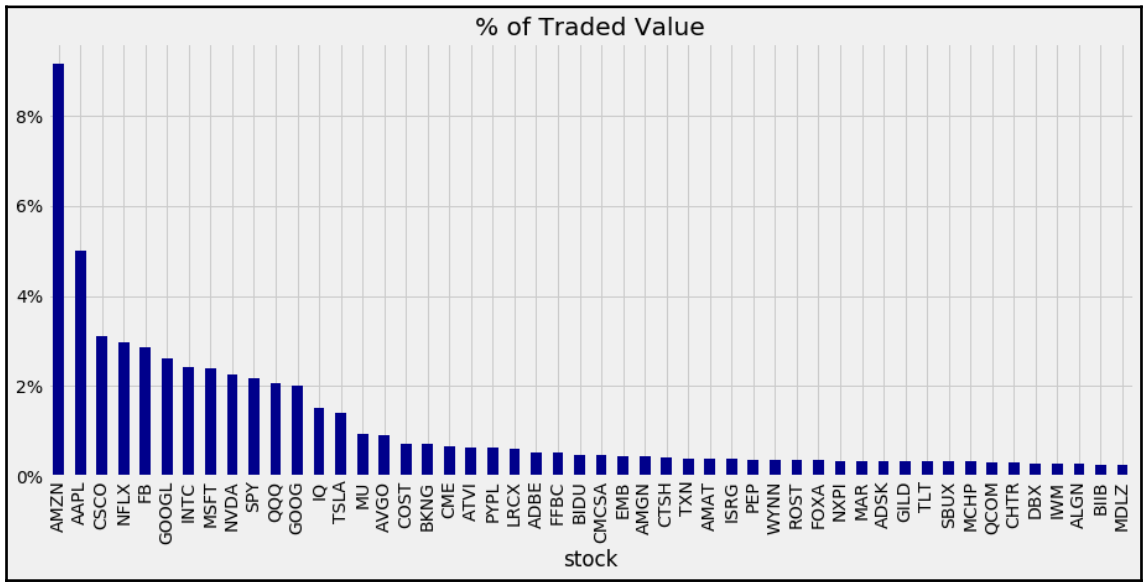
1
Graphics

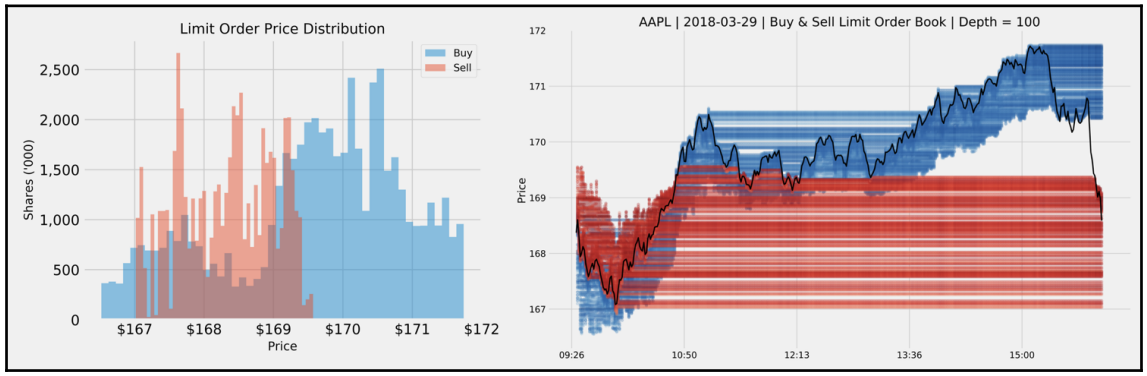
Chapter 1: Machine Learning for Trading

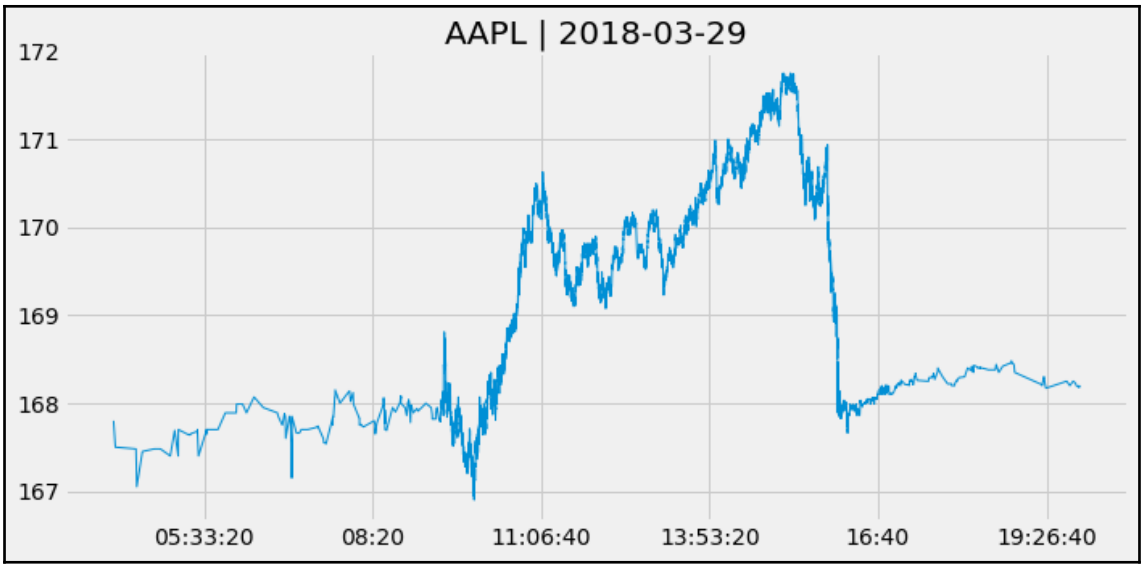


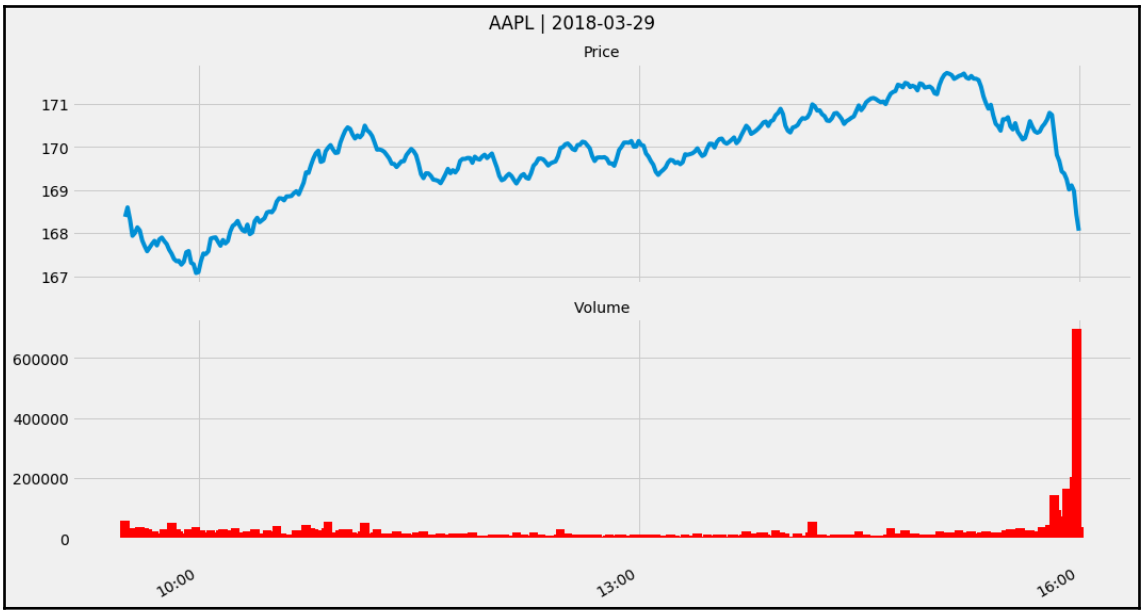


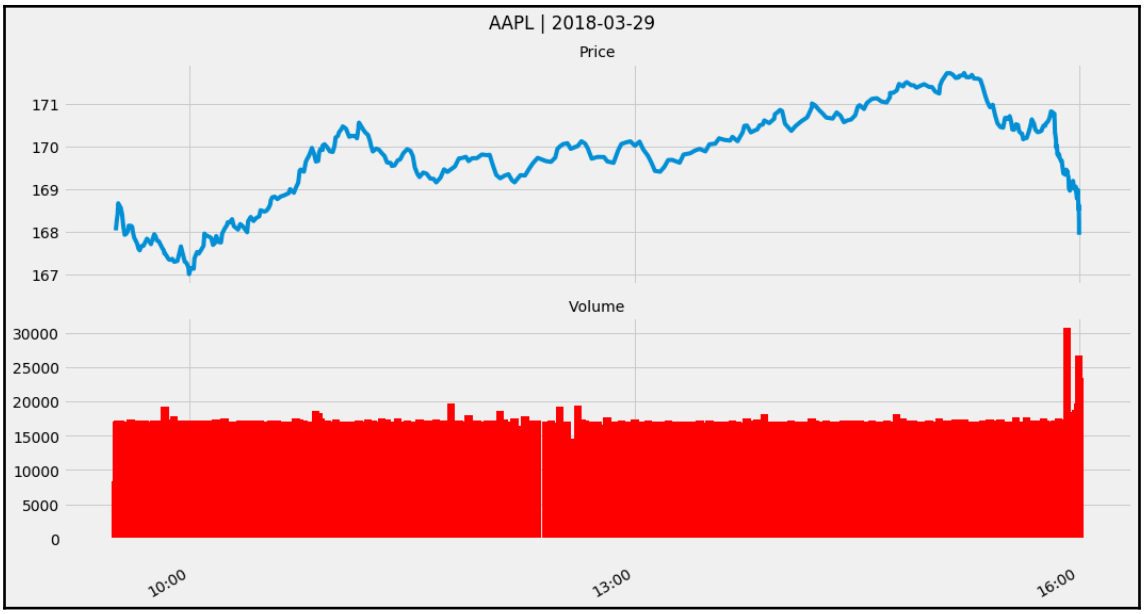
Chapter 2: Market and Fundamental Data



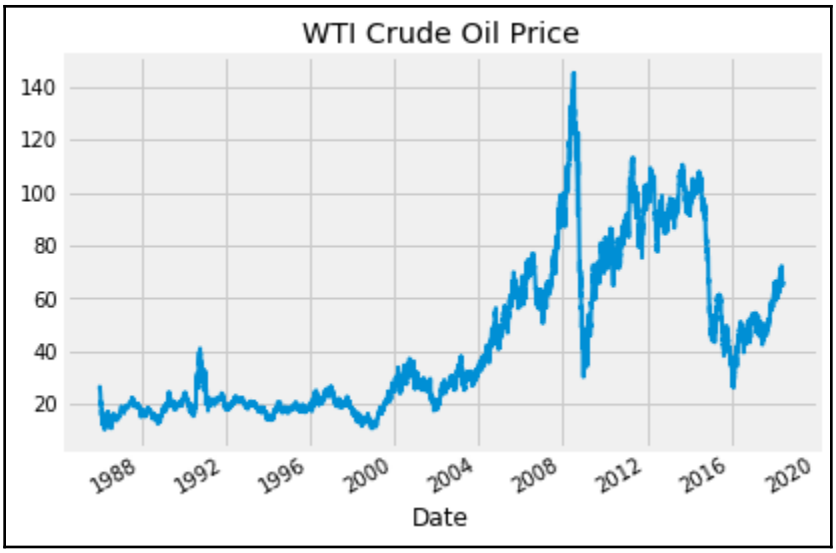


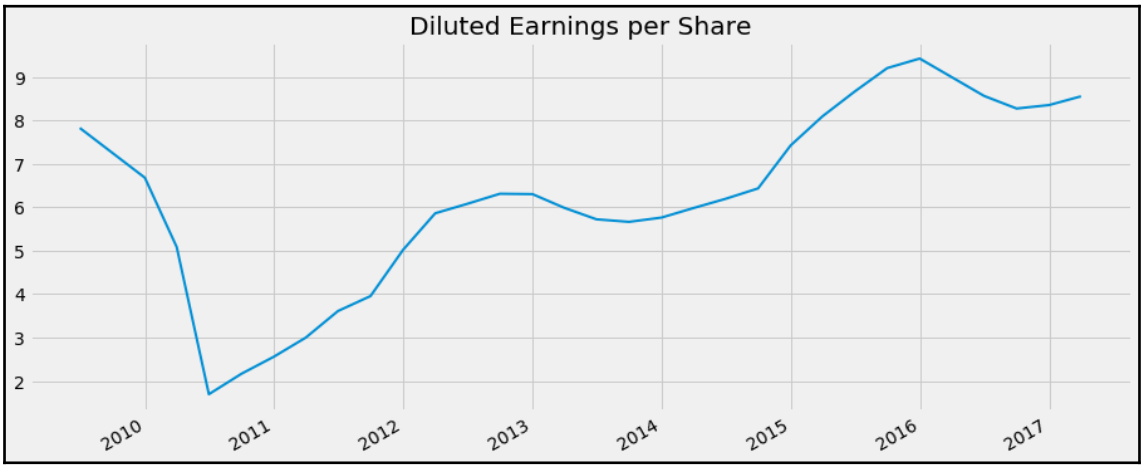

AAPL | 2018-03-29
172
171
170
169
168
167
05:33:20
08:20
11:06:40
13:53:20
16:40
19:26:40

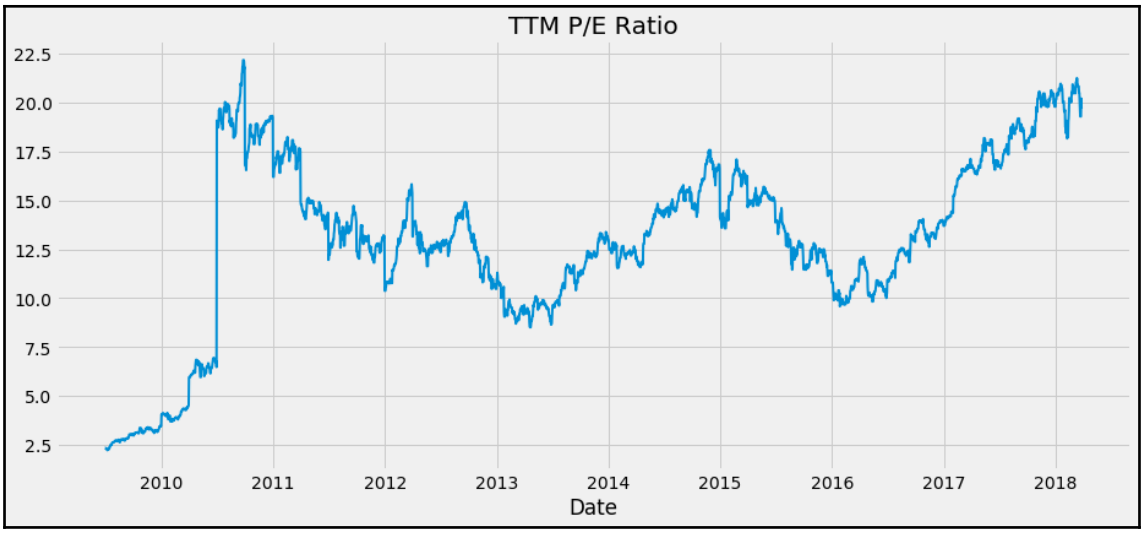

AAPL | 2018-03-29
Price
171
170
169
168
167
Volume
600000
400000
200000
0
10:00
13:00
16:00

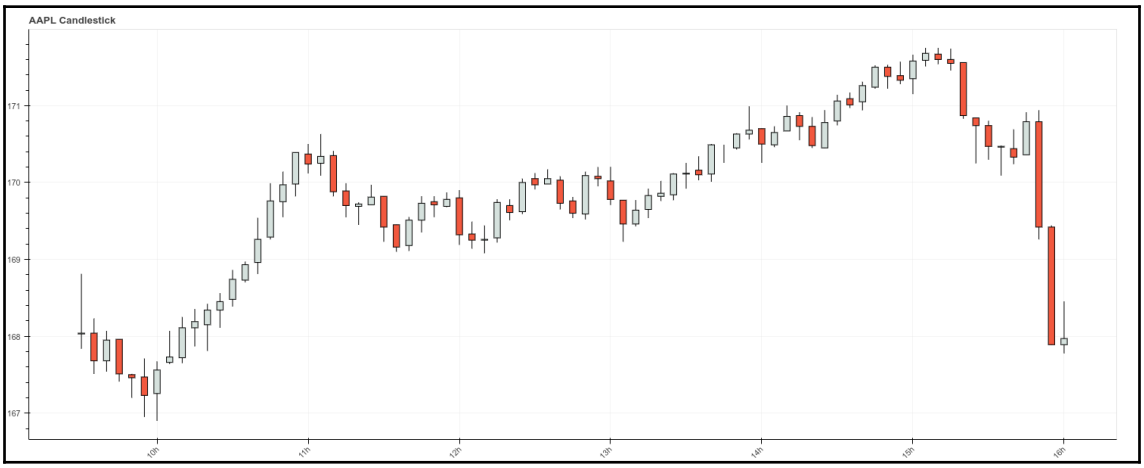

AAPL | 2018-03-29
Price
171
170
169
168
167
Volume
30000
25000
20000
15000
10000
5000
0
10:00
13:00
16:00

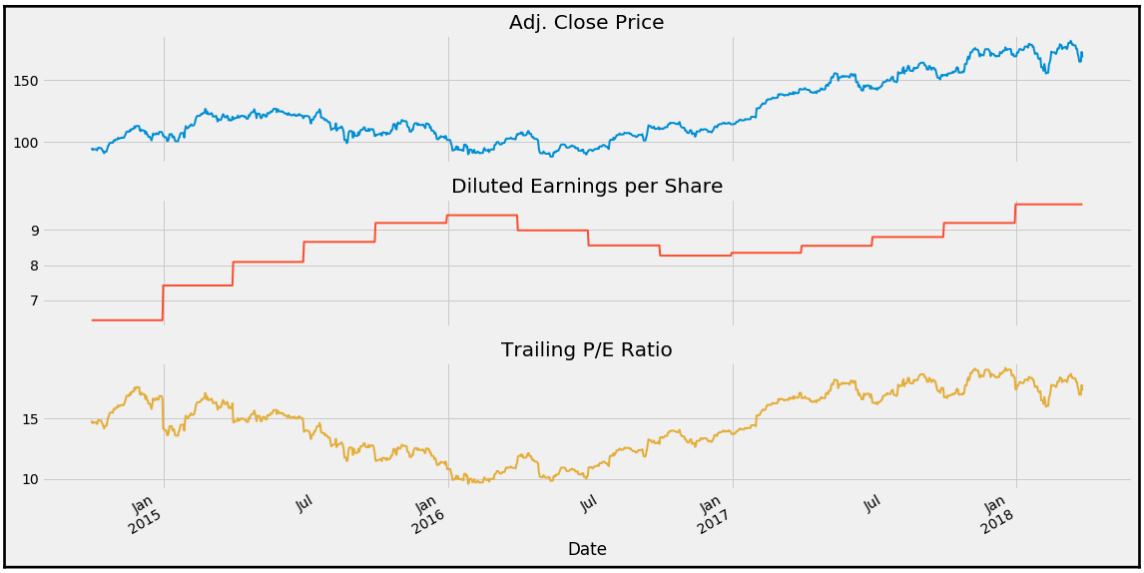

Equity(1037 [FB])
Equity(1244 [GOOG])
Equity(167 [AMZN])
1200
1000
800
600
400
200
0
2011
2012
2013
2014
2015
2016
2017
date

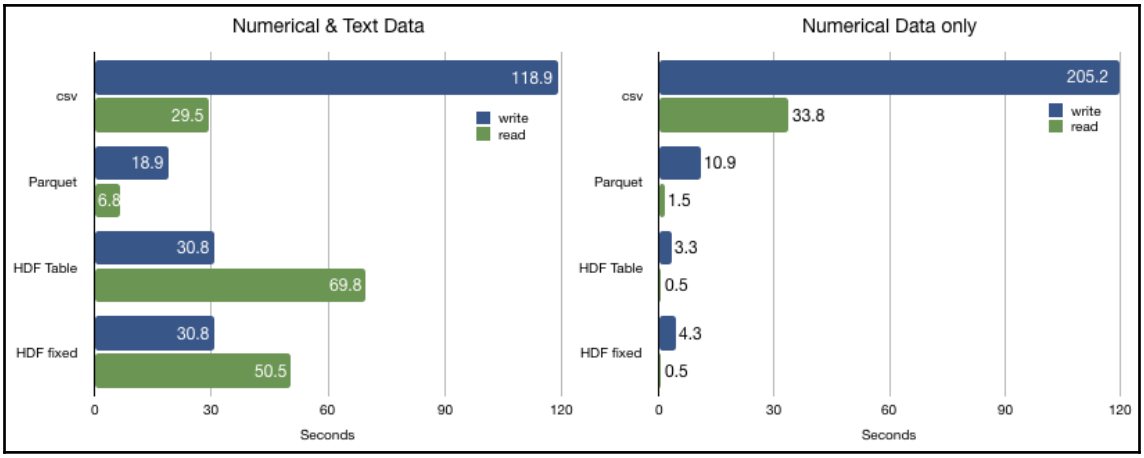

WTI Crude Oil Price
140
120
100
80
60
40
20
1988
1992
1996
2000
2004
2008
2012
2016
2020
Date


Diluted Earnings per Share
9
8
7
6
5
4
3
2
2010
2011
2012
2013
2014
2015
2016
2017

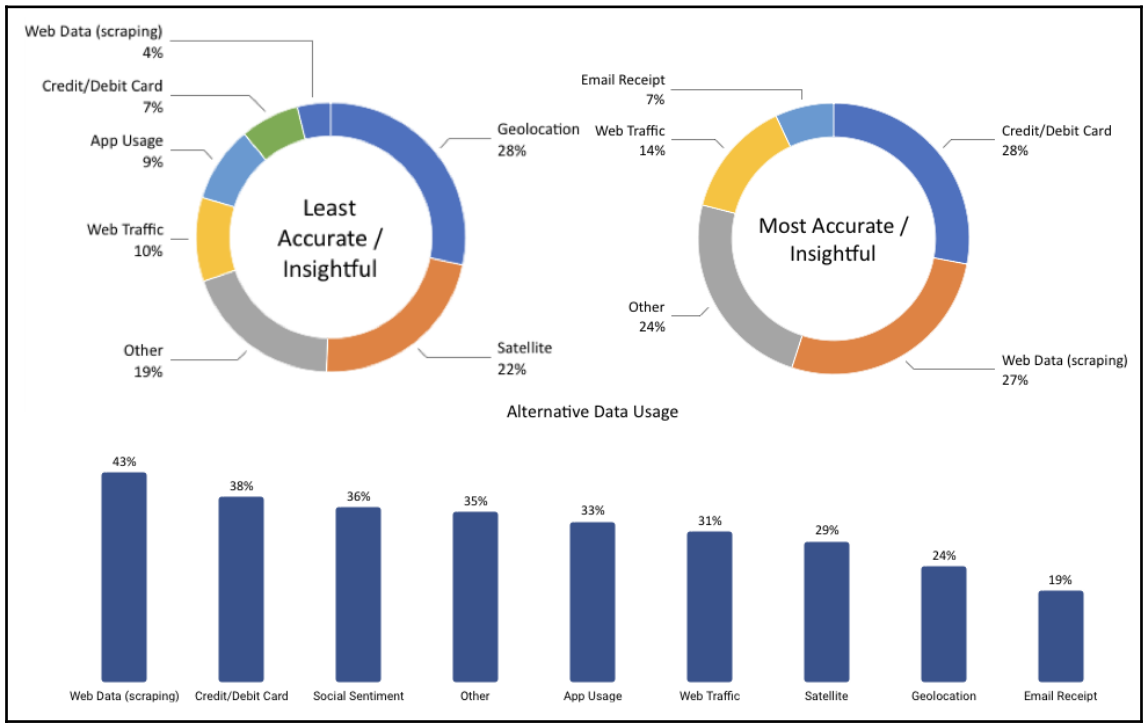

TTM P/E Ratio
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
2010
2011
2012
2013
2014
2015
2016
2017
2018
Date


AAPL Candlestick
171
170
169
168
167

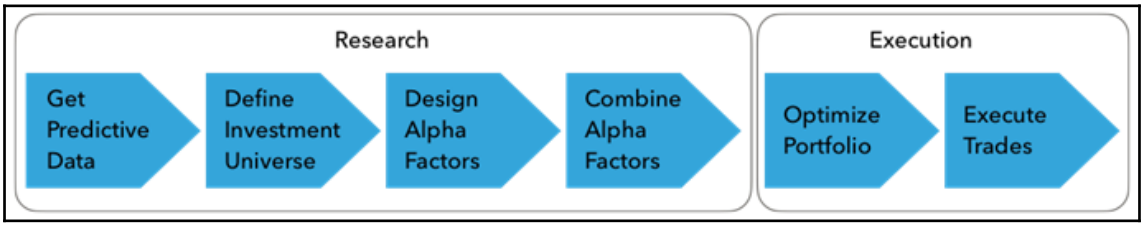

Adj. Close Price
150
100
Diluted Earnings per Share
9
8
7
Trailing P/E Ratio
15
10
Jan 2015
Jul
Jan 2016
Jul
Jan 2017
Jul
Jan 2018
Date

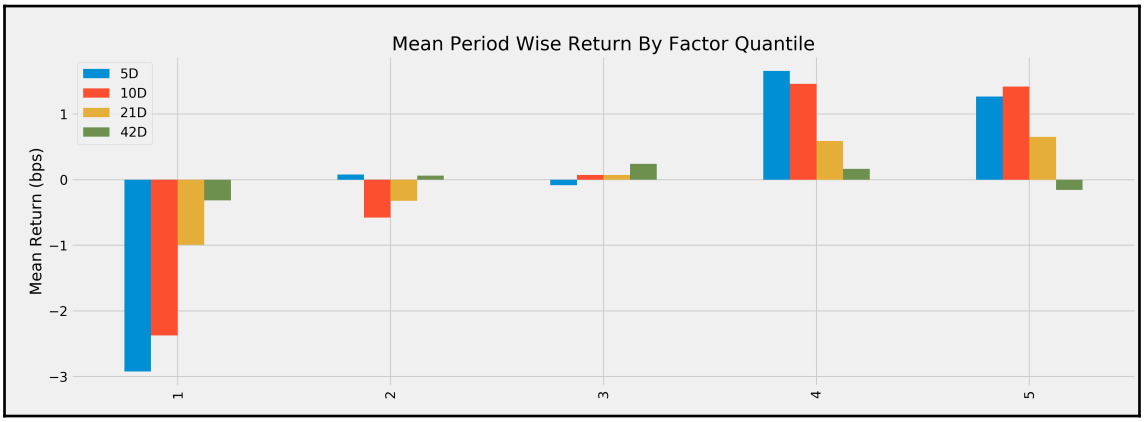

Numerical & Text Data
csv
Parquet
HDF Table
HDF fixed
118.9
29.5
18.9
6.8
30.8
69.8
30.8
50.5
write
read
0
30
60
90
120
Seconds
Numerical Data only
csv
Parquet
HDF Table
HDF fixed
205.2
33.8
10.9
1.5
3.3
0.5
4.3
0.5
write
read
0
30
60
90
120
Seconds

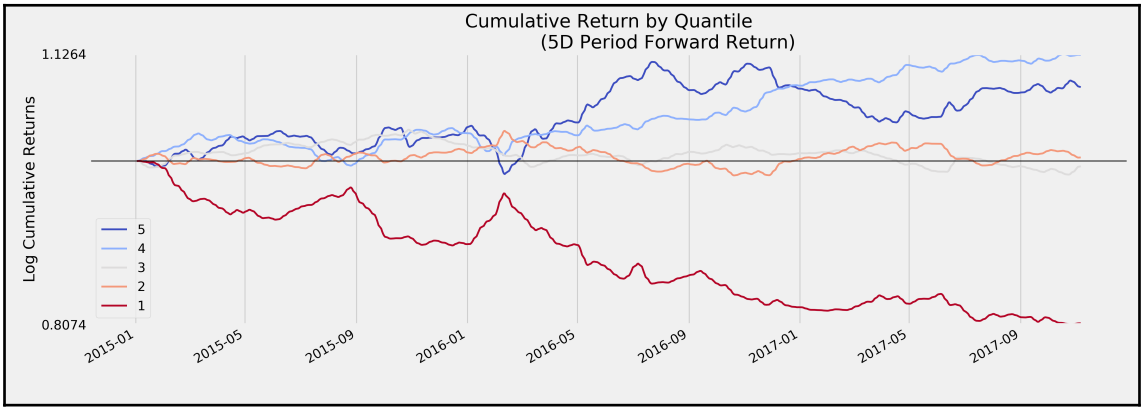
Chapter 3: Alternative Data for Finance

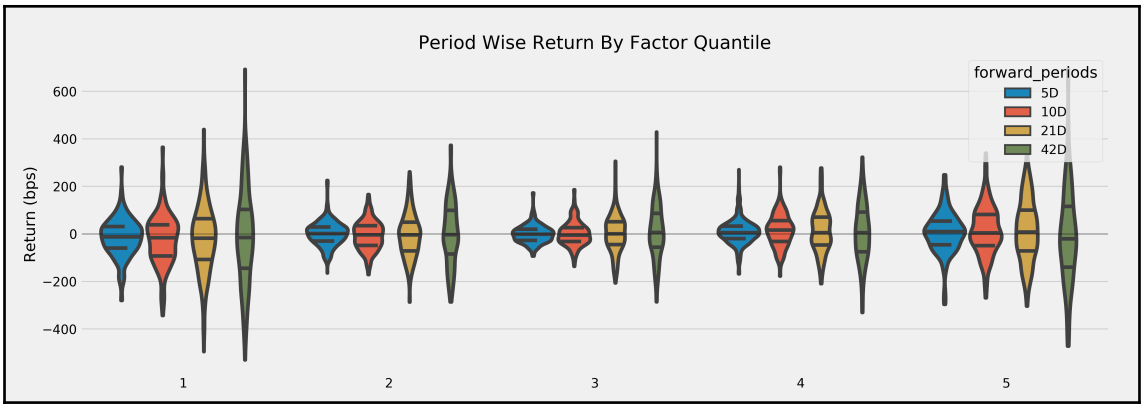


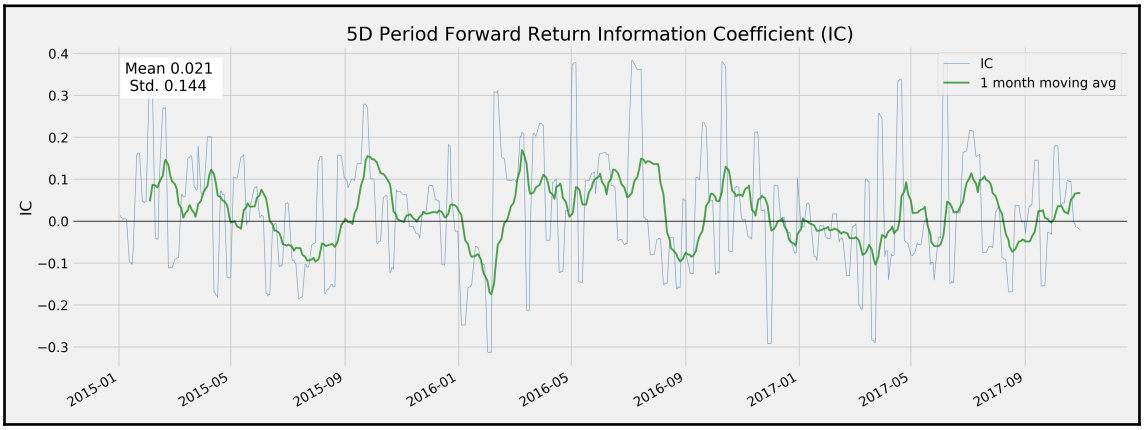
Chapter 4: Alpha Factor Research

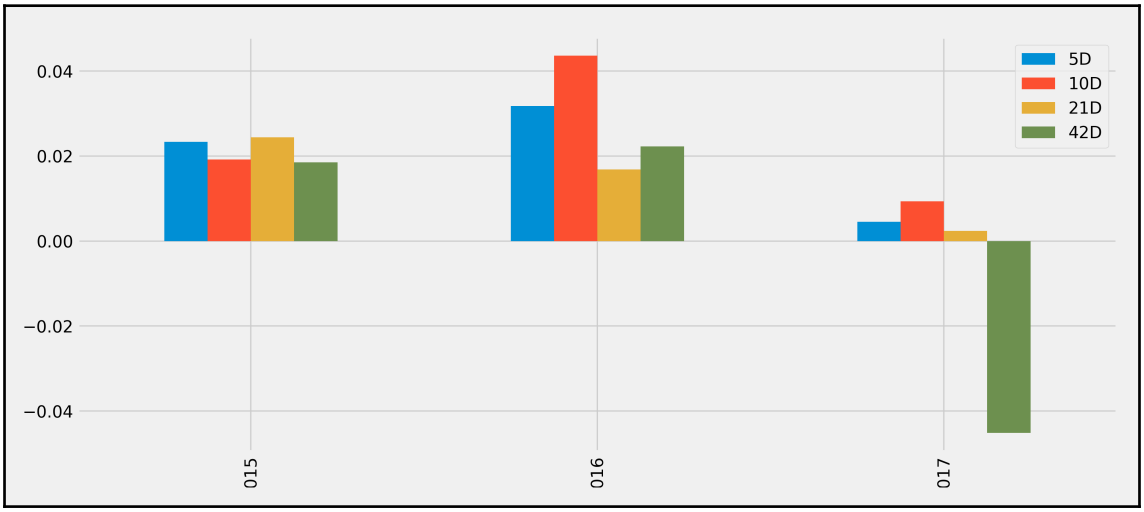




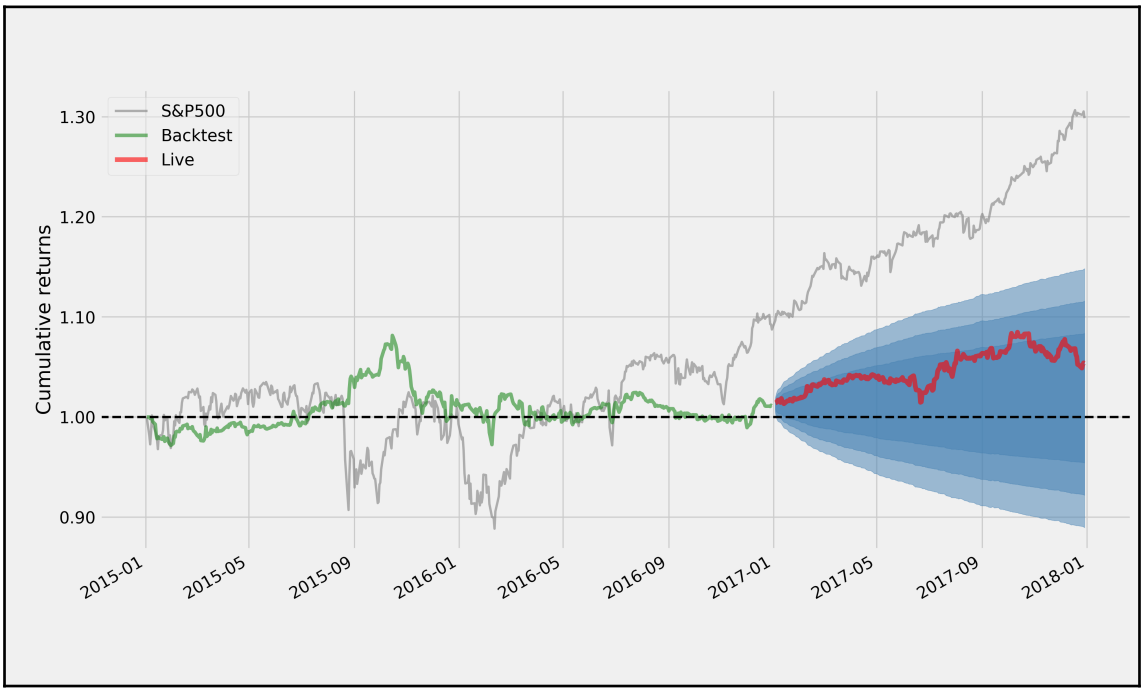


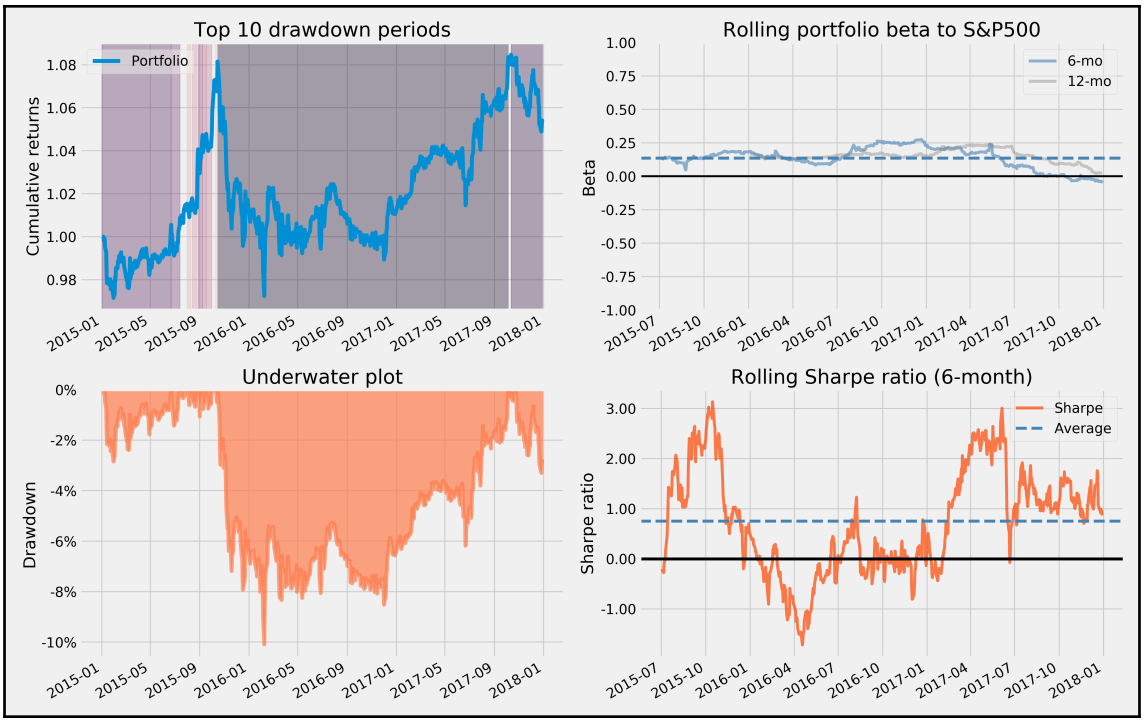

Period Wise Return By Factor Quantile
forward_periods
5D
10D
21D
42D
Return (bps)
600
400
200
0
−200
−400
1
2
3
4
5

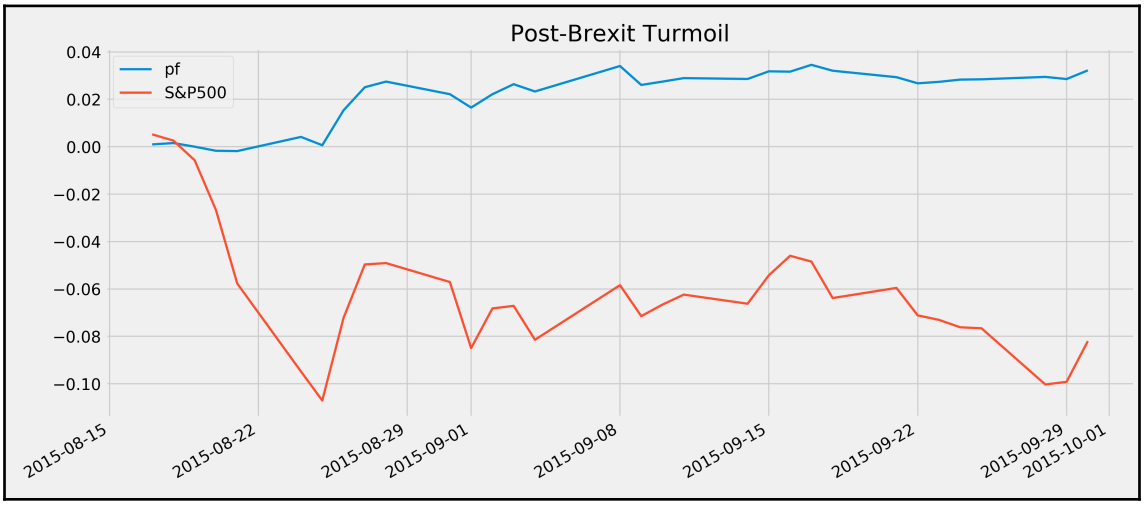

5D Period Forward Return Information Coefficient (IC)
Mean 0.021
Std. 0.144
IC
1 month moving avg
IC
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
2015-01
2015-05
2015-09
2016-01
2016-05
2016-09
2017-01
2017-05
2017-09

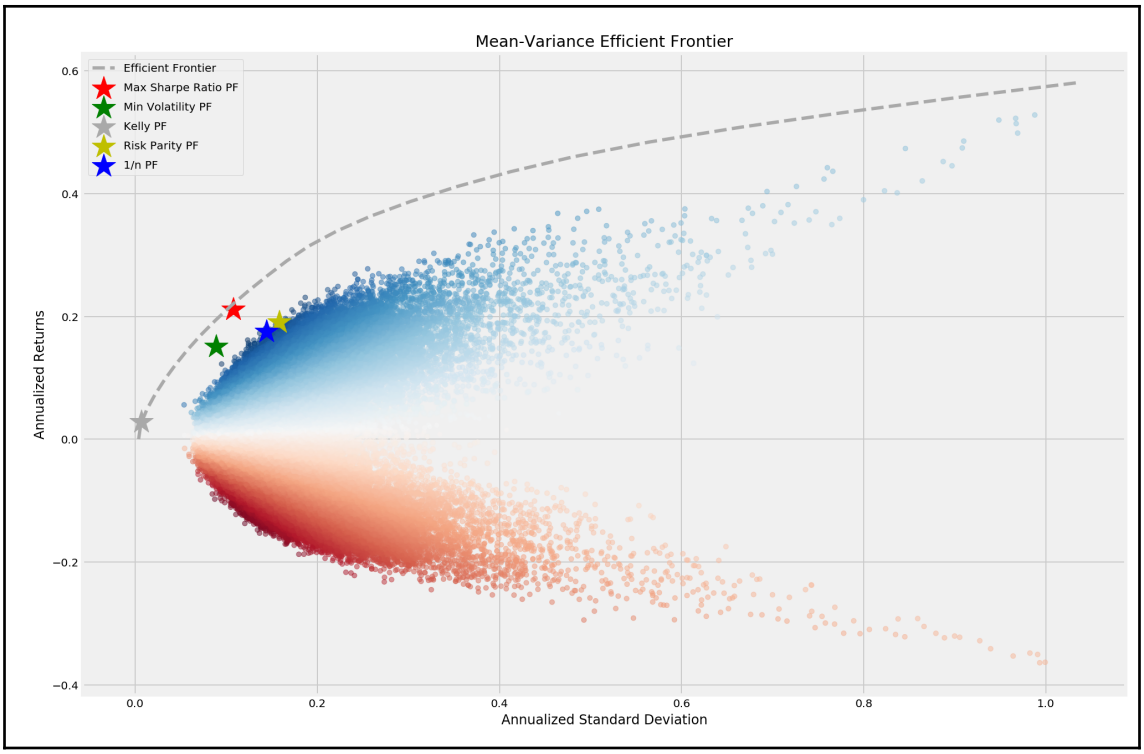

5D
10D
21D
42D
0.04
0.02
0.00
−0.02
−0.04
015
016
017

Chapter 5: Strategy Evaluation




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
pf
S&P500


Mean-Variance Efficient Frontier
Efficient Frontier
Max Sharpe Ratio PF
Min Volatility PF
Kelly PF
Risk Parity PF
1/n PF
Annualized Returns
Annualized Standard Deviation
0.6
0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0

Chapter 6: The Machine Learning Process

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

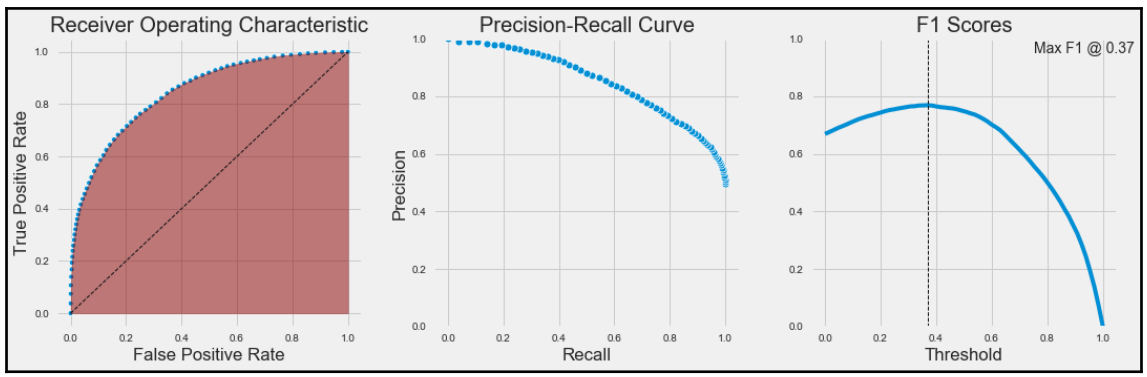
$$\text{Accuracy} = \frac{\text{\# Correct Predictions}}{\text{\# Cases}} = \frac{TP + TN}{TP + FP + TN + FN}$$

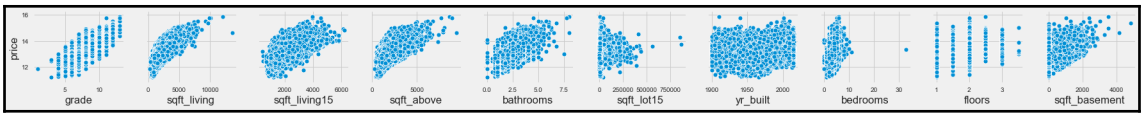
$$\text{True Positive Rate (Sensitivity, Recall)} = \frac{\text{\# Correct Positive Predictions}}{\text{\# Positive Cases}} = \frac{TP}{TP + FN}$$

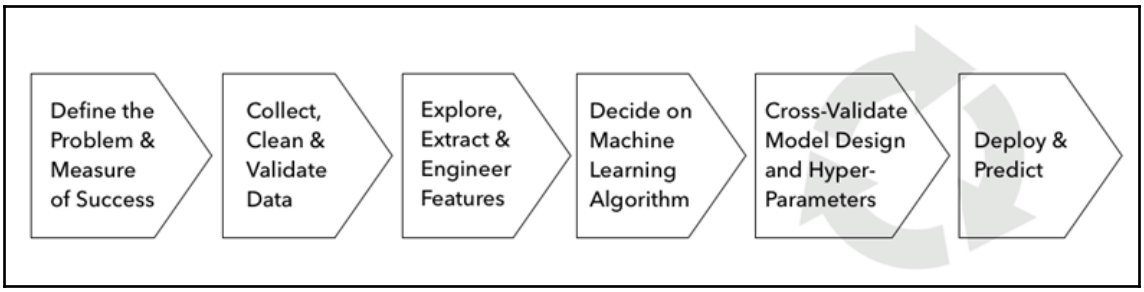
$$\text{False Negative Rate (Miss Rate)} = 1 - \text{True Positive Rate}$$

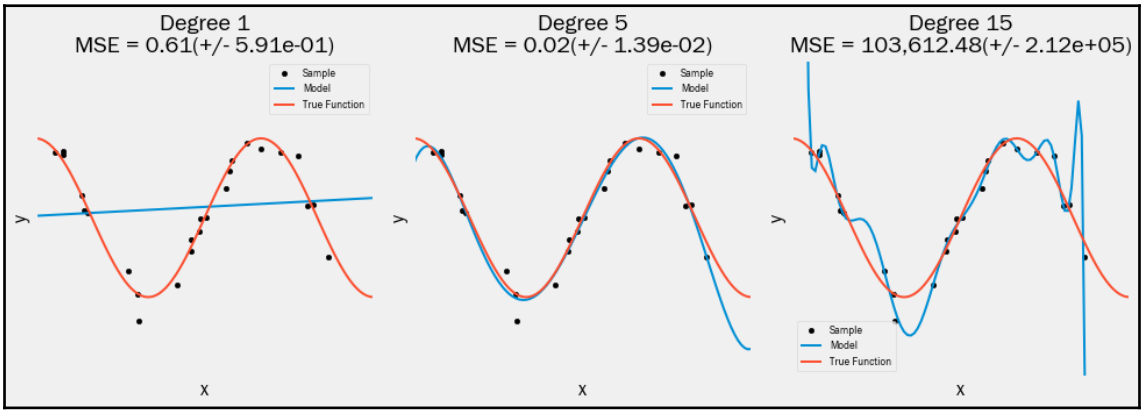
$$\text{True Negative Rate (Specificity)} = \frac{\text{\# Correct Negative Predictions}}{\text{\# Negative Cases}} = \frac{TN}{TN + FP}$$

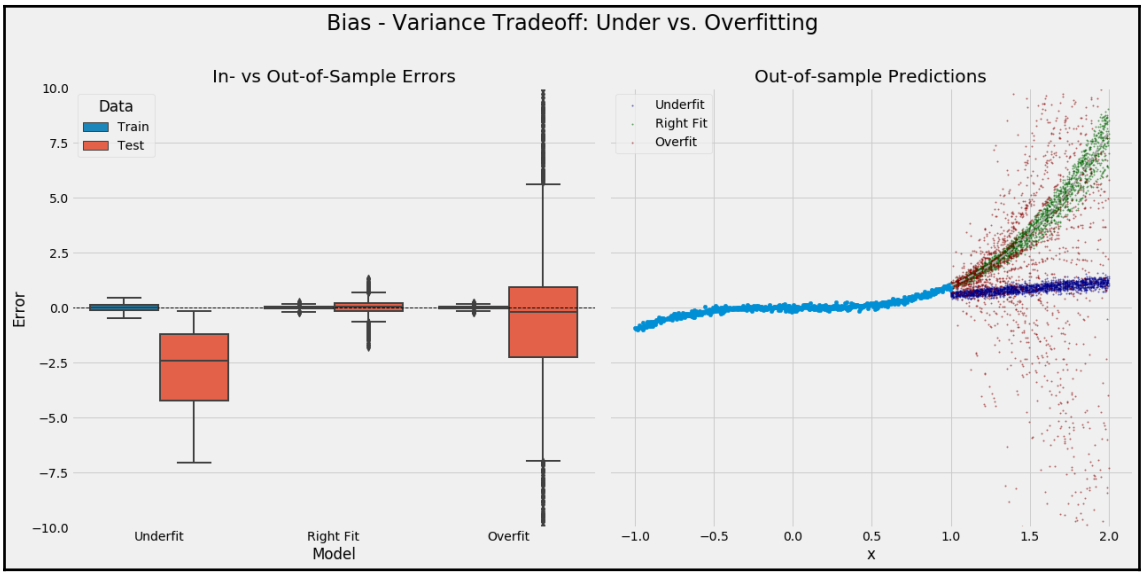
$$\text{False Positive Rate (Fall-Out)} = 1 - \text{True Negative Rate}$$

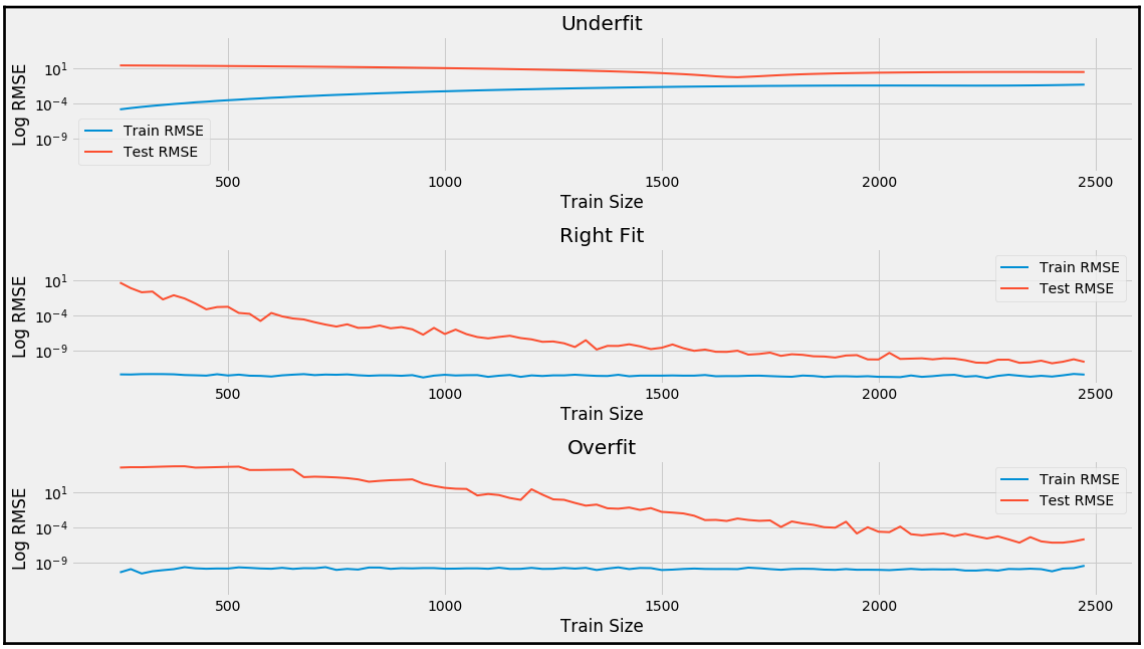


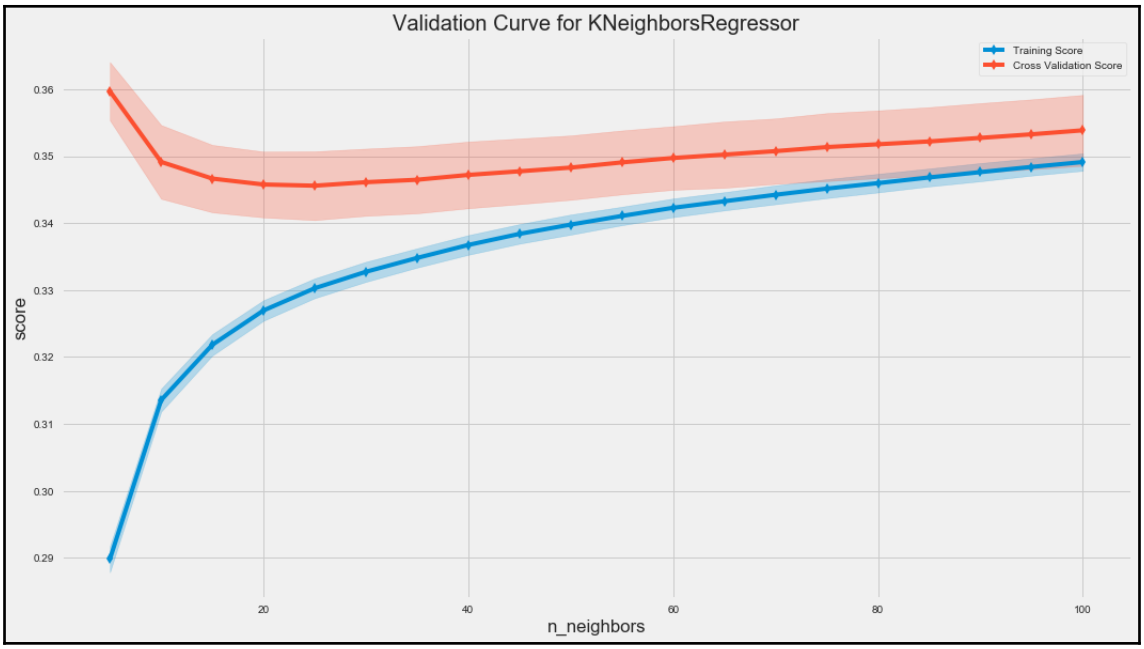

price
grade
sqft_living
sqft_living15
sqft_above
bathrooms
sqft_lot15
yr_built
bedrooms
floors
sqft_basement

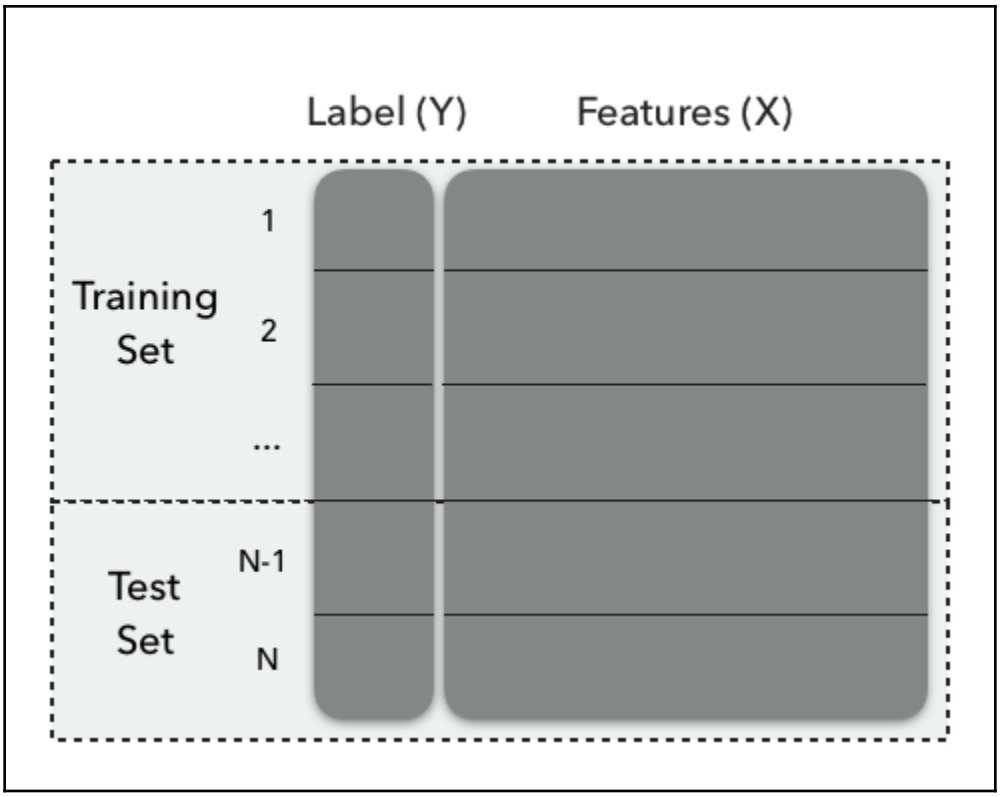

Define the Problem & Measure of Success
Collect, Clean & Validate Data
Explore, Extract & Engineer Features
Decide on Machine Learning Algorithm
Cross-Validate Model Design and Hyper-Parameters
Deploy & Predict

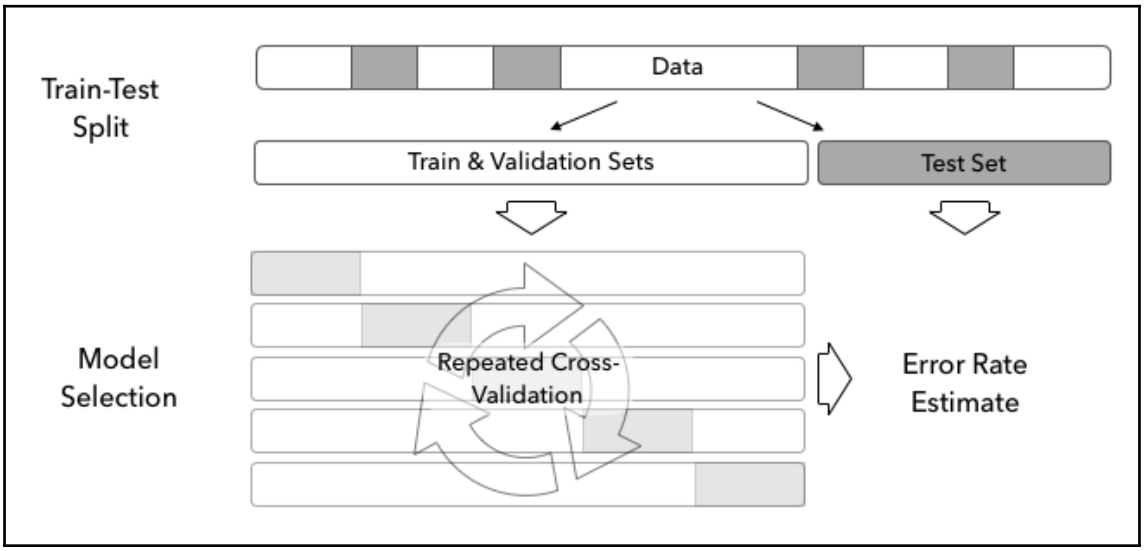

Degree 1
MSE = 0.61(+/- 5.91e-01)
Degree 5
MSE = 0.02(+/- 1.39e-02)
Degree 15
MSE = 103,612.48(+/- 2.12e+05)
Sample
Model
True Function
x
y

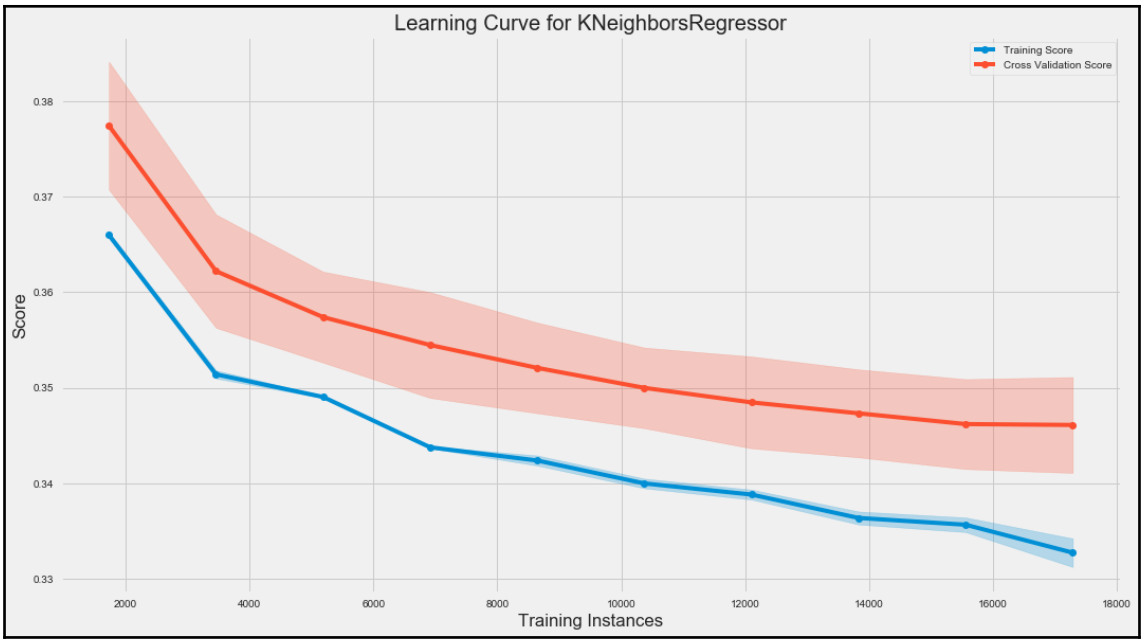

Bias - Variance Tradeoff: Under vs. Overfitting
In- vs Out-of-Sample Errors
Data
Train
Test
Error
Underfit
Right Fit
Overfit
Model
Out-of-sample Predictions
Underfit
Right Fit
Overfit
x


Underfit
Log RMSE
Train RMSE
Test RMSE
Train Size
Right Fit
Overfit


Validation Curve for KNeighborsRegressor
Training Score
Cross Validation Score
score
0.36
0.35
0.34
0.33
0.32
0.31
0.30
0.29
20
40
60
80
100
n_neighbors

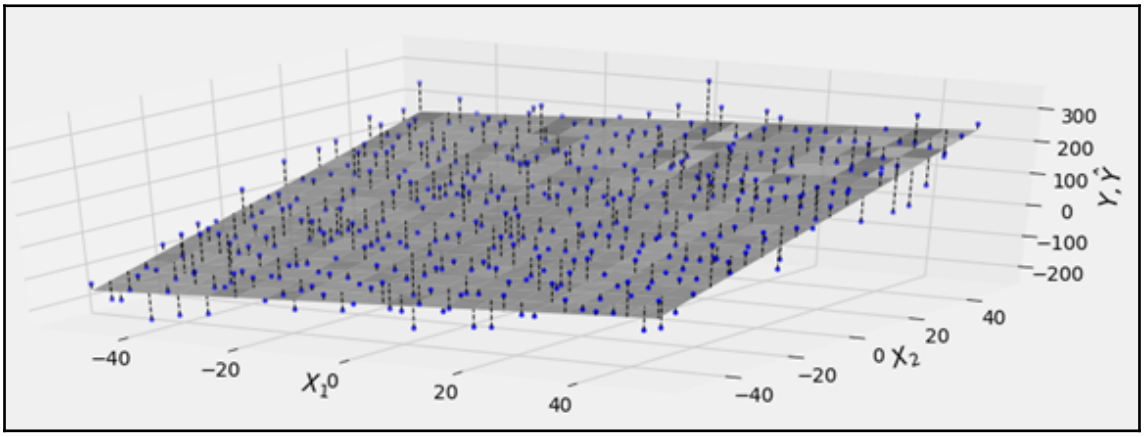

Label (Y)
Features (X)
Training Set
1
2
...
Test Set
N-1
N


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

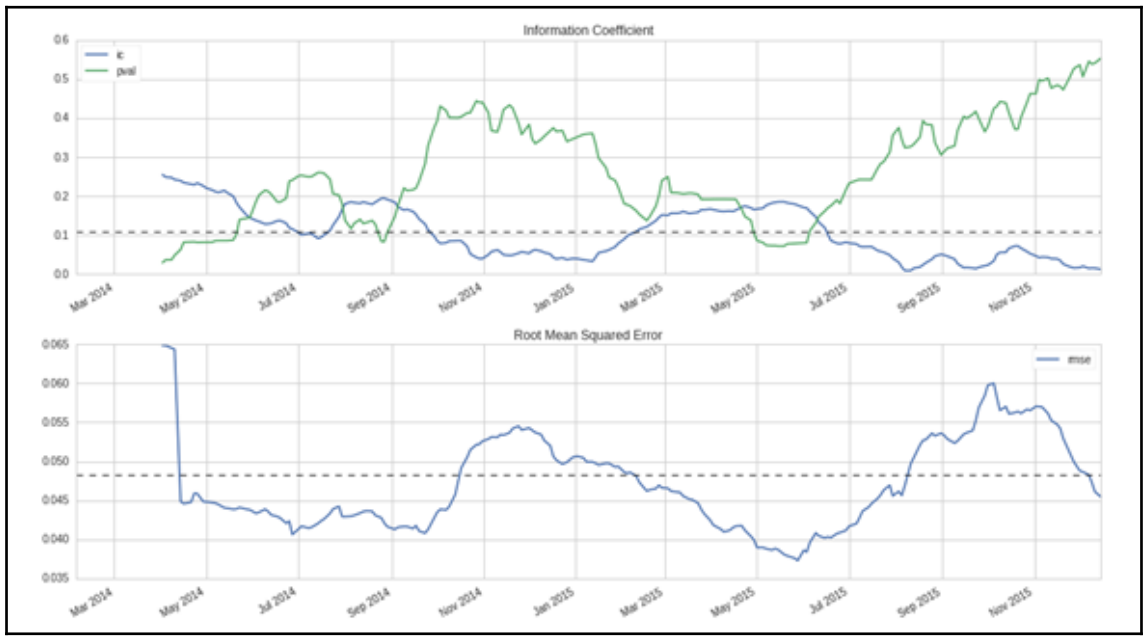

Learning Curve for KNeighborsRegressor
Training Score
Cross Validation Score
Score
0.38
0.37
0.36
0.35
0.34
0.33
2000
4000
6000
8000
10000
12000
14000
16000
18000
Training Instances

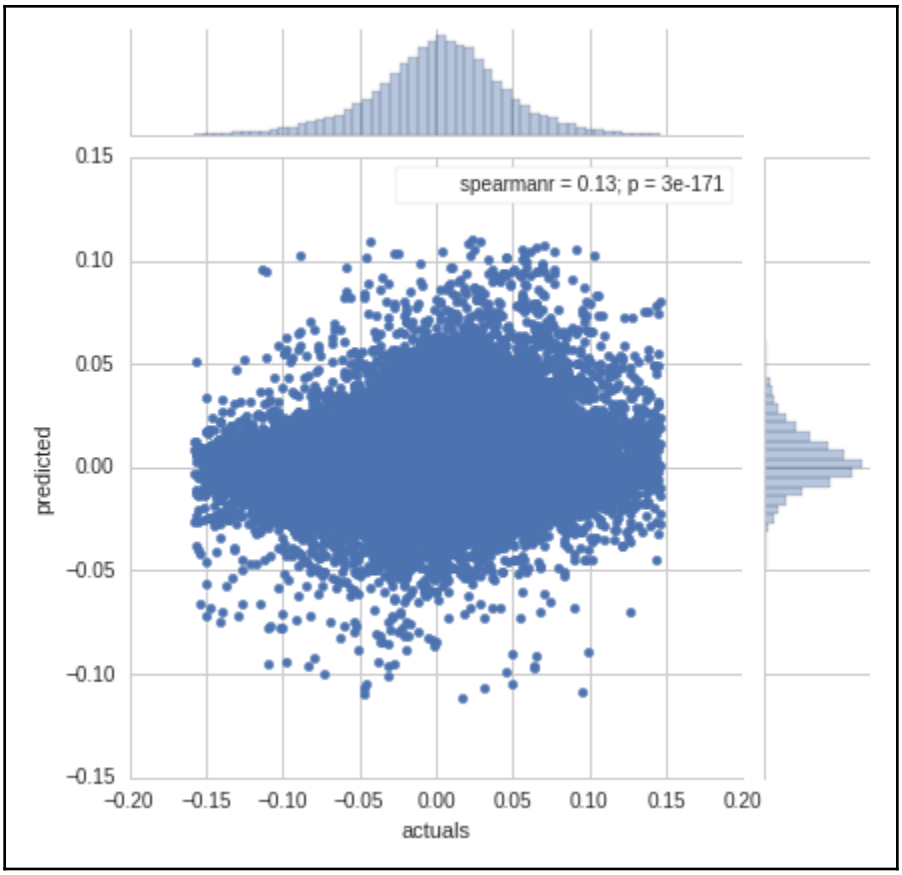
Chapter 7: Linear Models

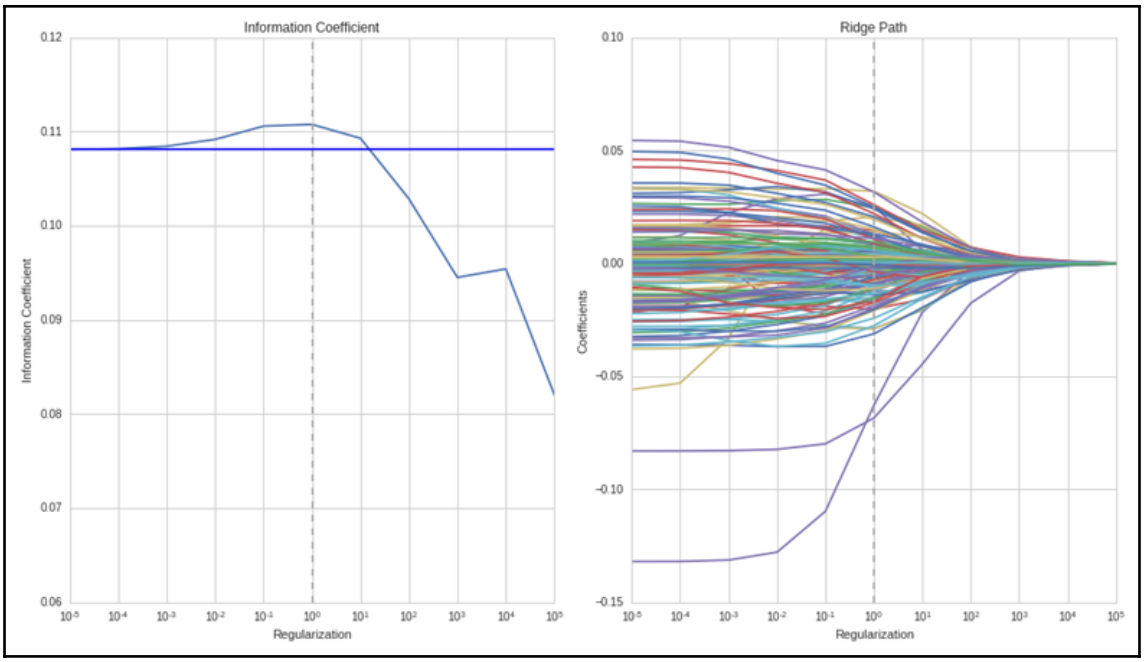
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.779
Model:                            OLS   Adj. R-squared:                  0.778
Method:                 Least Squares   F-statistic:                     1095.
Date:                Mon, 03 Sep 2018   Prob (F-statistic):          1.85e-204
Time:                        17:38:41   Log-Likelihood:                -3332.6
No. Observations:                 625   AIC:                             6671.
Df Residuals:                     622   BIC:                             6685.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         50.9371      2.007     25.376      0.000      46.995      54.879
X_1            1.0813      0.067     16.185      0.000       0.950       1.212
X_2            2.9328      0.067     43.900      0.000       2.802       3.064
==============================================================================
Omnibus:                        0.267   Durbin-Watson:                   2.140
Prob(Omnibus):                  0.875   Jarque-Bera (JB):                0.196
Skew:                           0.040   Prob(JB):                        0.907
Kurtosis:                       3.032   Cond. No.                         30.0
==============================================================================

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

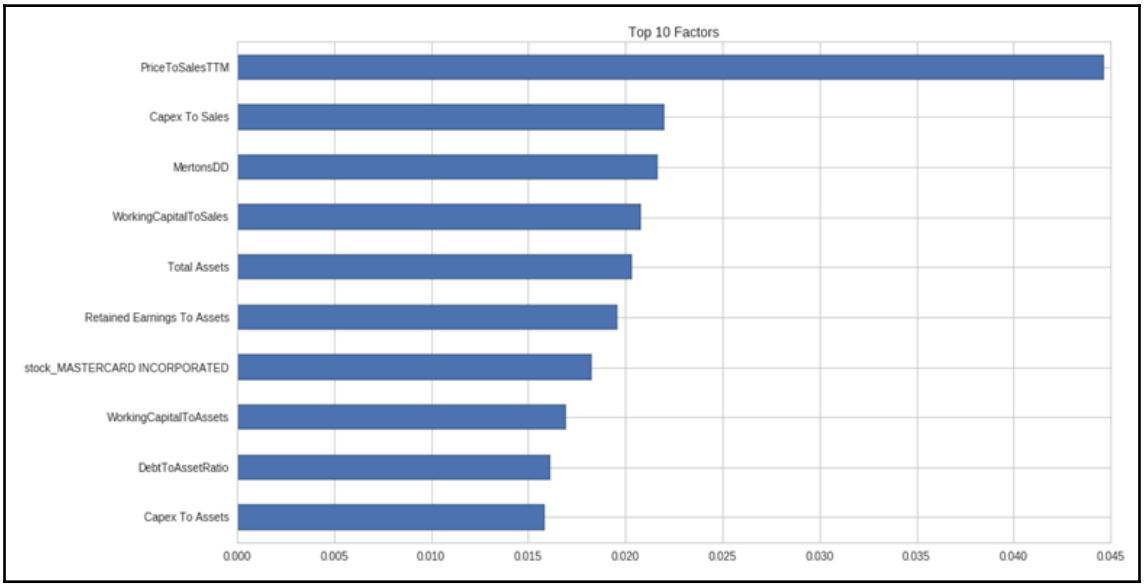


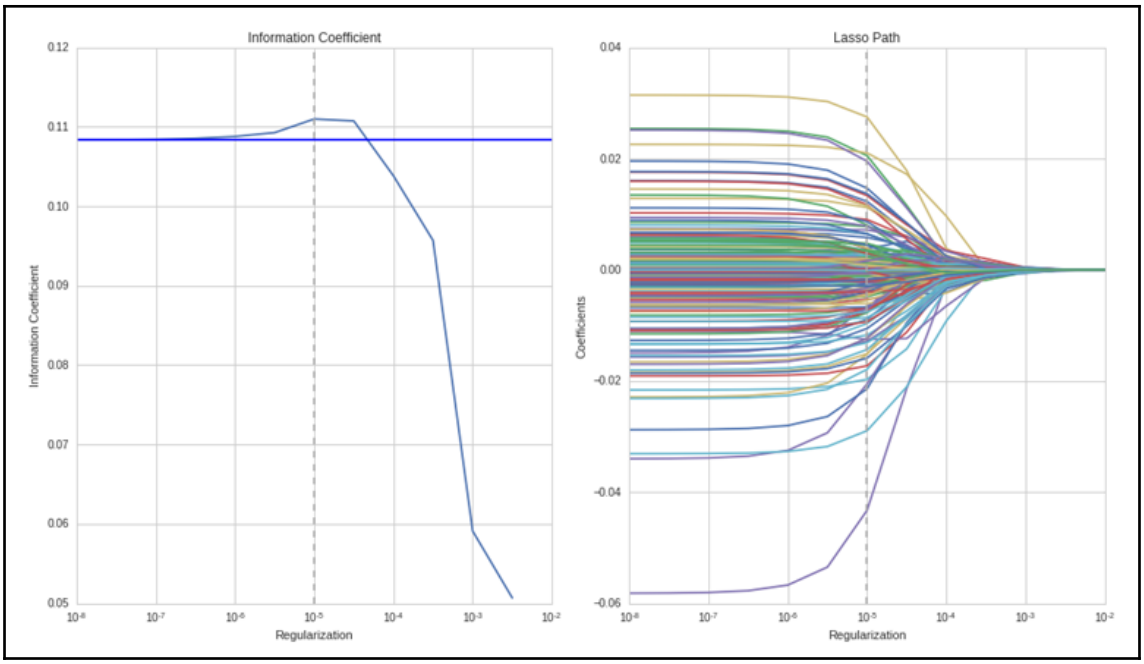
```
                    LinearFactorModel Estimation Summary
==========================================================================
No. Test Portfolios:                 17   R-squared:                0.6943
No. Factors:                          6   J-statistic:              19.155
No. Observations:                    95   P-value                   0.0584
Date:                 Wed, Oct 31 2018   Distribution:             chi2(11)
Time:                         15:15:52
Cov. Estimator:                 robust

                           Risk Premia Estimates
==========================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
--------------------------------------------------------------------------
Mkt-RF         1.2446     0.3928     3.1689     0.0015      0.4748      2.0144
SMB            0.0074     0.7055     0.0105     0.9917     -1.3753      1.3901
HML           -0.6970     0.5334    -1.3067     0.1913     -1.7424      0.3484
RMW           -0.2558     0.6888    -0.3713     0.7104     -1.6057      1.0942
CMA           -0.3086     0.4737    -0.6515     0.5147     -1.2371      0.6198
RF            -0.0133     0.0132    -1.0092     0.3129     -0.0393      0.0126
==========================================================================

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



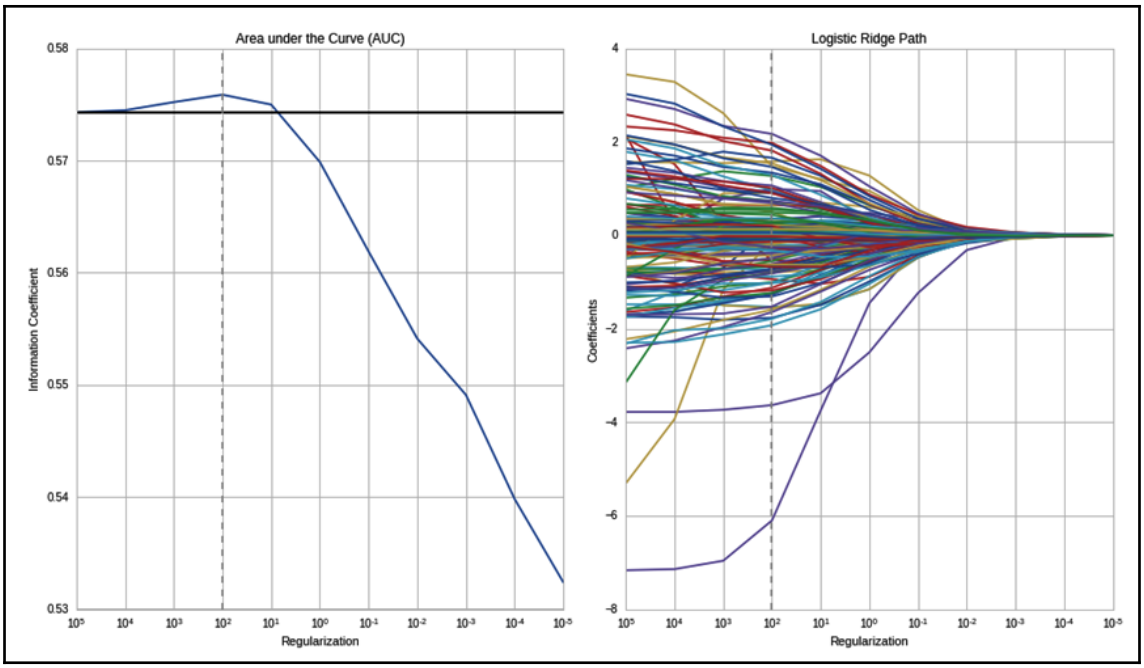

spearmanr = 0.13; p = 3e-171
predicted
actuals
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20


Information Coefficient
Information Coefficient
Regularization
Ridge Path
Coefficients
Regularization

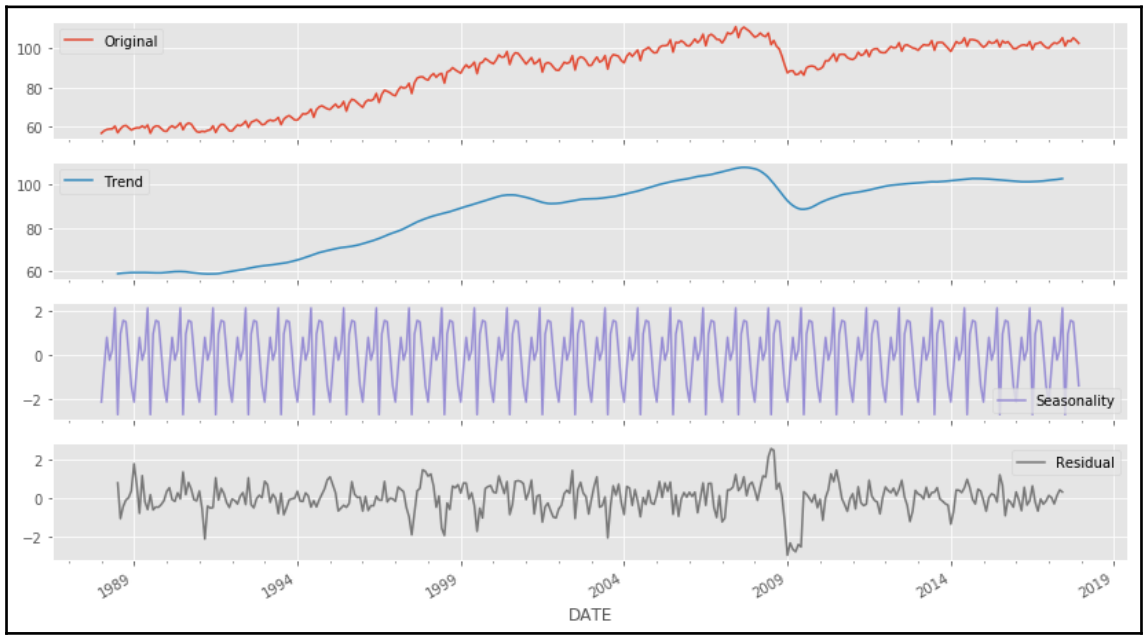

Top 10 Factors
PriceToSalesTTM
Capex To Sales
MertonsDD
WorkingCapitalToSales
Total Assets
Retained Earnings To Assets
stock_MASTERCARD INCORPORATED
WorkingCapitalToAssets
DebtToAssetRatio
Capex To Assets

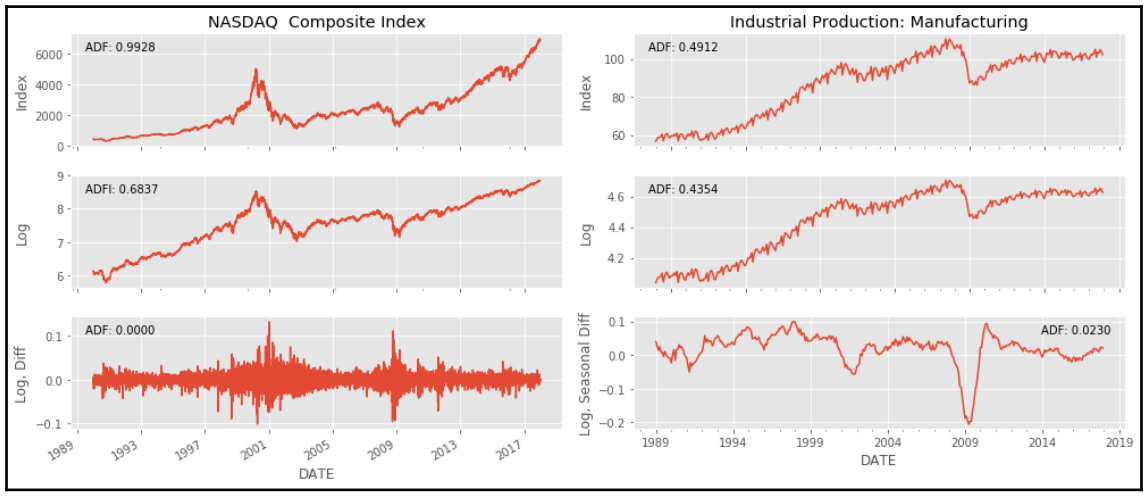

Information Coefficient
Information Coefficient
Regularization
Lasso Path
Coefficients
Regularization

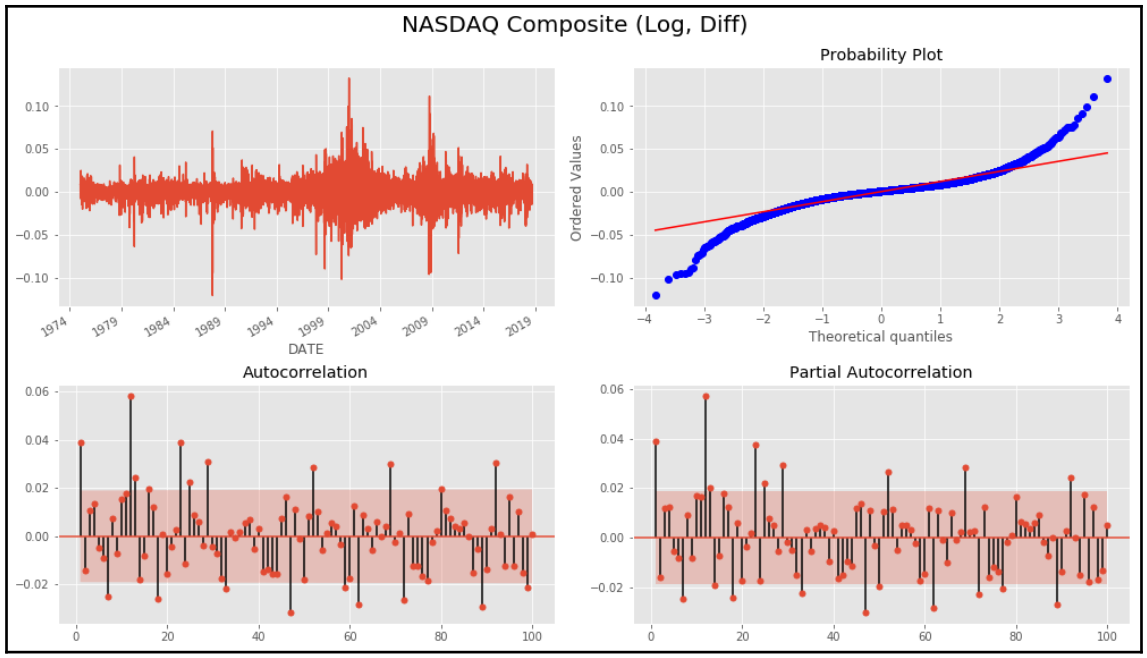
```
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:                  198
Model:                          Logit   Df Residuals:                      185
Method:                           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
==============================================================================
```

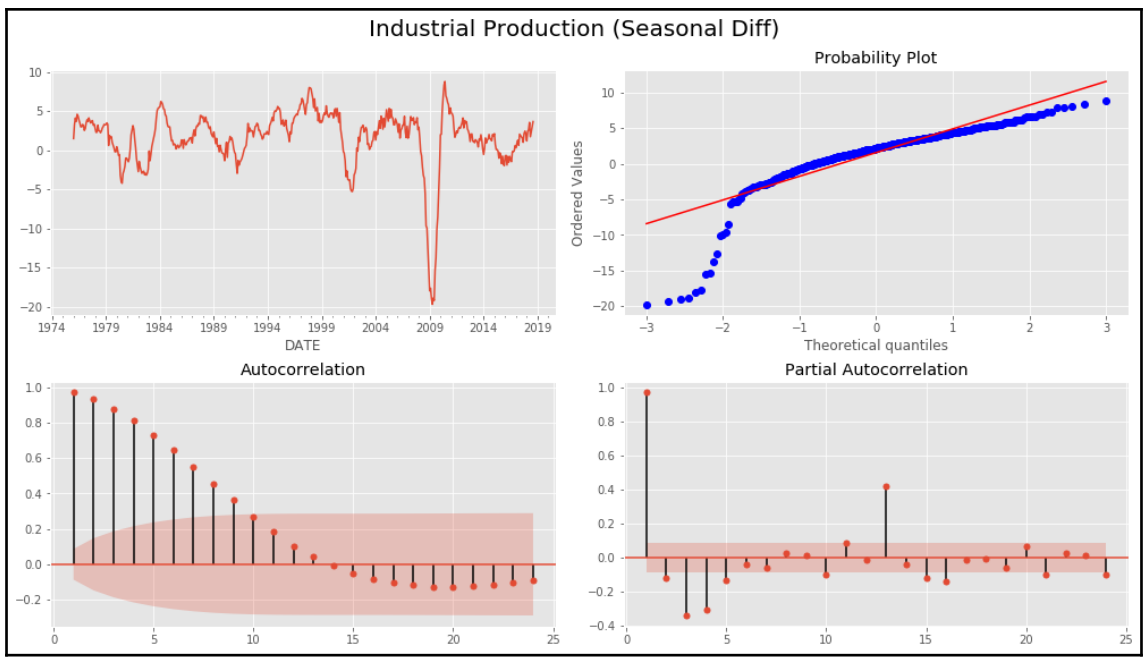

Area under the Curve (AUC)
Information Coefficient
0.58
0.57
0.56
0.55
0.54
0.53
10^{5} 10^{4} 10^{3} 10^{2} 10^{1} 10^{0} 10^{-1} 10^{-2} 10^{-3} 10^{-4} 10^{-5}
Regularization
Logistic Ridge Path
Coefficients
4
2
0
-2
-4
-6
-8
10^{5} 10^{4} 10^{3} 10^{2} 10^{1} 10^{0} 10^{-1} 10^{-2} 10^{-3} 10^{-4} 10^{-5}
Regularization

Chapter 8: Time Series Models

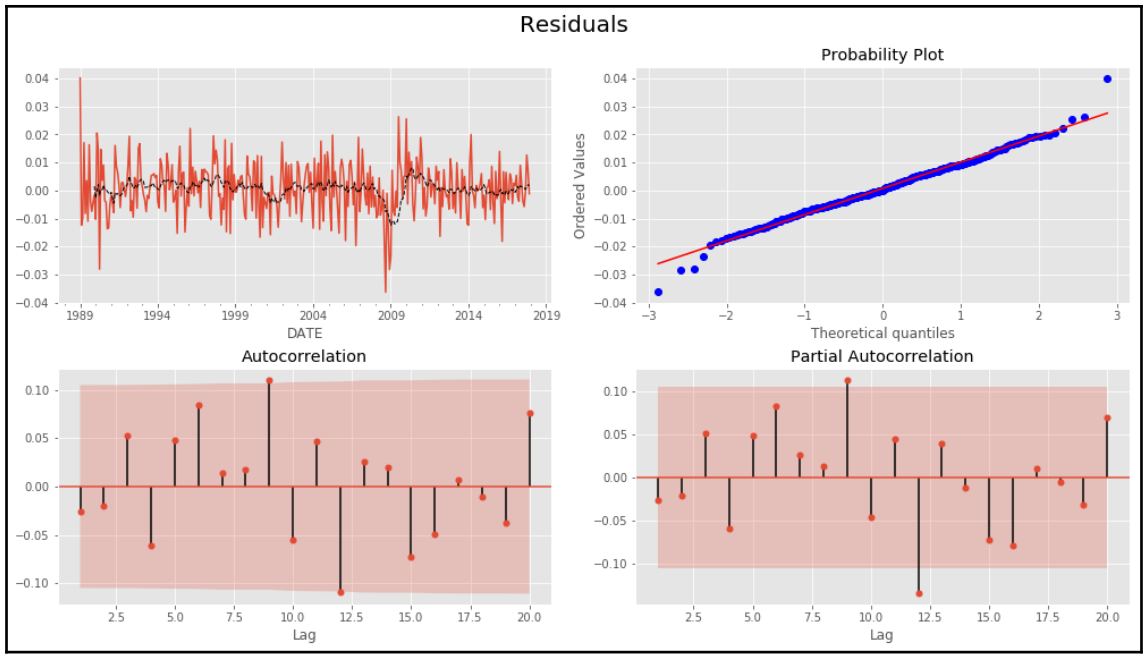


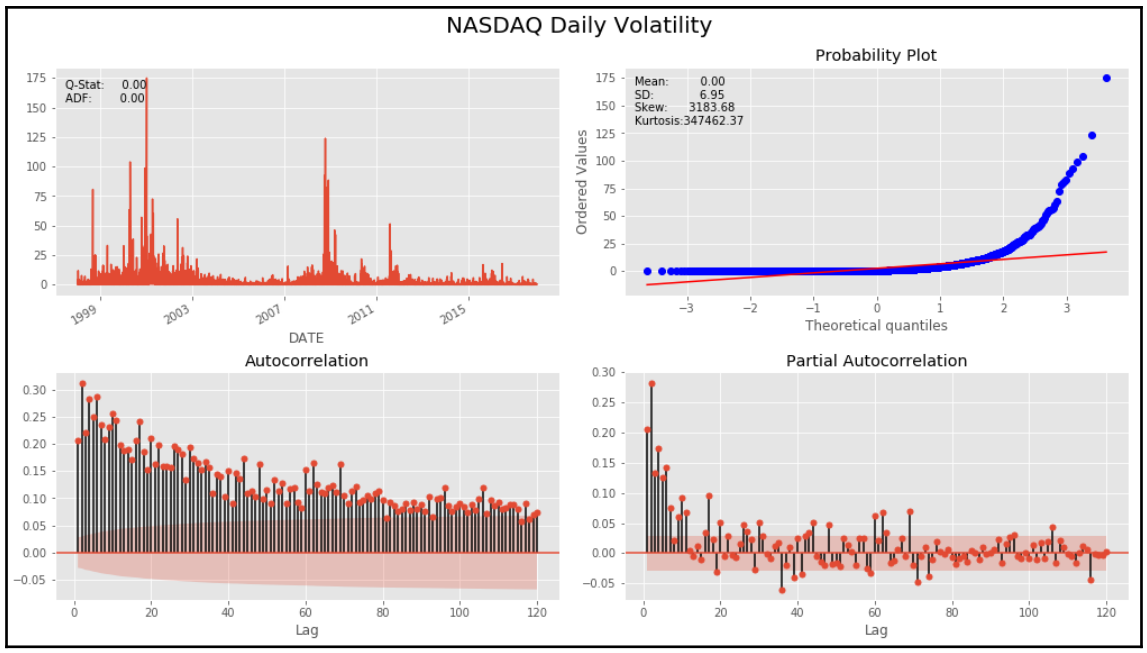



NASDAQ Composite (Log, Diff)
Probability Plot
Ordered Values
Theoretical quantiles
DATE
Autocorrelation
Partial Autocorrelation

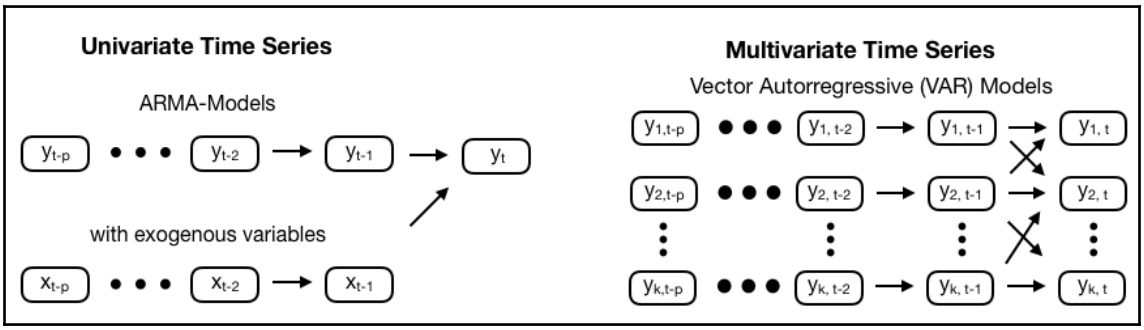

Industrial Production (Seasonal Diff)
Probability Plot
Ordered Values
Theoretical quantiles
DATE
Autocorrelation
Partial Autocorrelation

```
                              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).
```


Residuals
Probability Plot
DATE
Ordered Values
Theoretical quantiles
Autocorrelation
Partial Autocorrelation
Lag
Lag

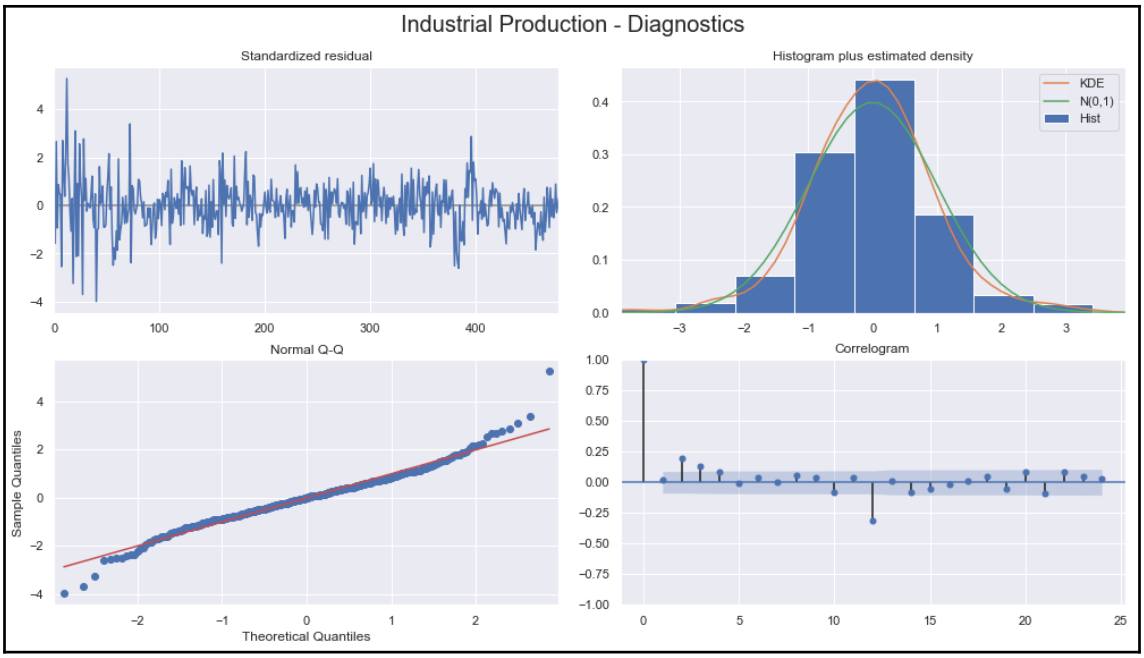

NASDAQ Daily Volatility
Q-Stat: 0.00
ADF: 0.00
Probability Plot
Mean: 0.00
SD: 6.95
Skew: 3183.68
Kurtosis:347462.37
DATE
Ordered Values
Theoretical quantiles
Autocorrelation
Partial Autocorrelation
Lag
Lag

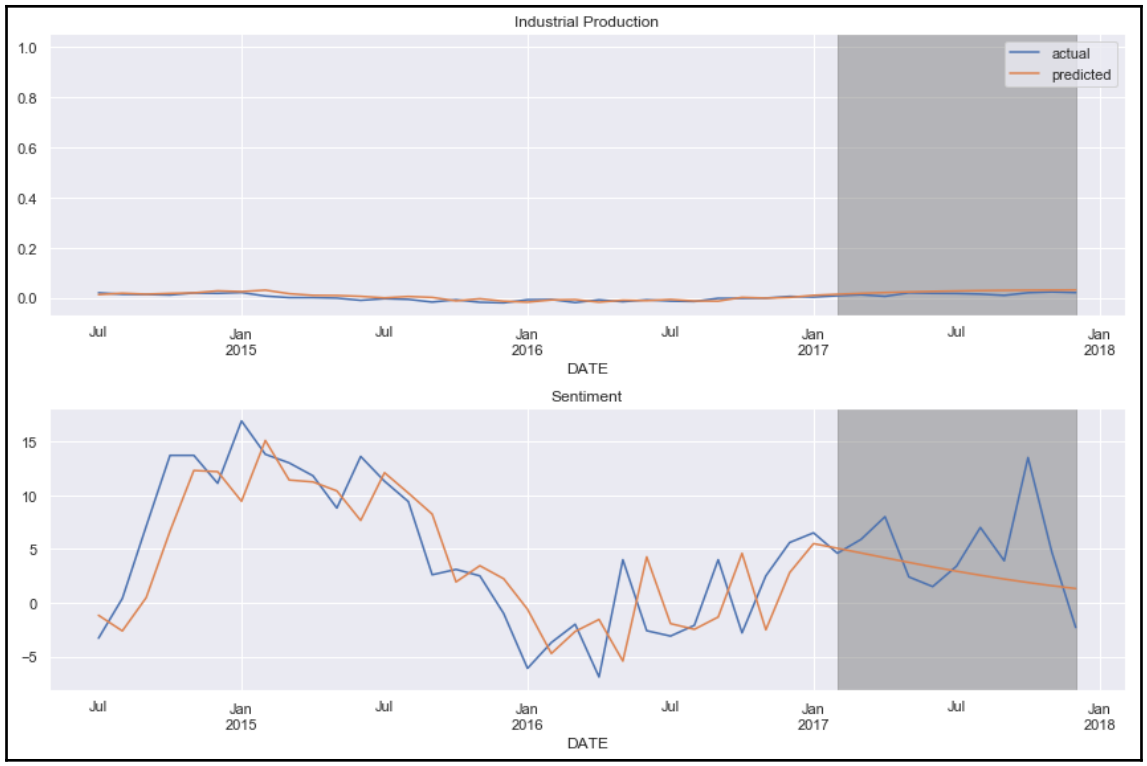
```
                    Constant Mean - GARCH Model Results
==============================================================================
Dep. Variable:              NASDAQCOM   R-squared:                      -0.001
Mean Model:             Constant Mean   Adj. R-squared:                 -0.001
Vol Model:                      GARCH   Log-Likelihood:               -7484.02
Distribution:                  Normal   AIC:                           14980.0
Method:            Maximum Likelihood   BIC:                           15019.0
                                        No. Observations:                 4852
Date:                Sun, Sep 23 2018   Df Residuals:                     4846
Time:                        15:43:41   Df Model:                            6
                              Mean Model
==============================================================================
                 coef    std err          t      P>|t|      95.0% Conf. Int.
------------------------------------------------------------------------------
mu             0.0521  1.491e-02      3.491  4.804e-04 [2.284e-02,8.130e-02]
                           Volatility Model
==============================================================================
                 coef    std err          t      P>|t|       95.0% Conf. Int.
------------------------------------------------------------------------------
omega          0.0196  8.287e-03      2.365  1.804e-02  [3.354e-03,3.584e-02]
alpha[1]       0.0247  1.470e-02      1.678  9.340e-02 [-4.148e-03,5.346e-02]
alpha[2]       0.0627  2.196e-02      2.853  4.324e-03    [1.962e-02,  0.106]
beta[1]        0.5648      0.181      3.120  1.806e-03      [  0.210,  0.920]
beta[2]        0.3337      0.180      1.853  6.393e-02   [-1.932e-02,  0.687]
==============================================================================

Covariance estimator: robust
```



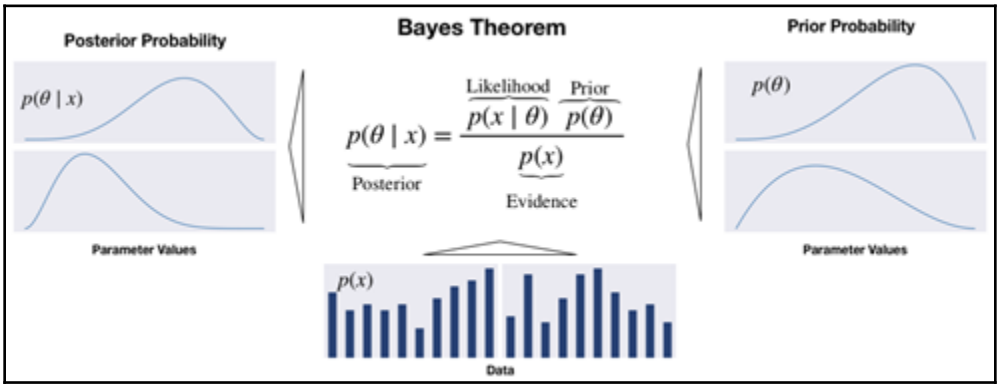

ip
sentiment
0.2
0.1
0.0
−0.1
−0.2
30
20
10
0
−10
−20
−30
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
DATE

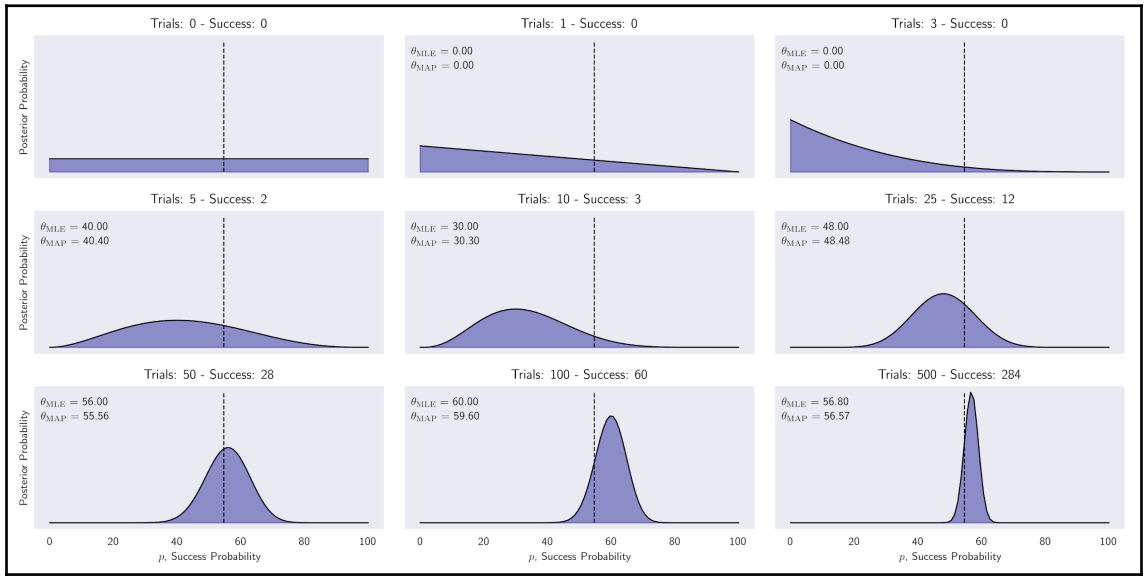
```
                           Statespace Model Results
==============================================================================
Dep. Variable:     ['ip', 'sentiment']   No. Observations:                  480
Model:                      VARMA(1,1)   Log Likelihood                -68.938
                          + intercept   AIC                            163.875
Date:                 Sun, 23 Sep 2018   BIC                            218.134
Time:                         17:53:02   HQIC                           185.203
Sample:                              0
                                 - 480
Covariance Type:                   opg
===================================================================================
Ljung-Box (Q):                129.82, 165.15   Jarque-Bera (JB):        140.59, 16.05
Prob(Q):                         0.00, 0.00   Prob(JB):                  0.00, 0.00
Heteroskedasticity (H):          0.47, 1.10   Skew:                      0.19, 0.21
Prob(H) (two-sided):             0.00, 0.55   Kurtosis:                  5.62, 3.79
                              Results for equation ip
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const              0.0016      0.001      2.531      0.011       0.000       0.003
L1.ip              0.9276      0.010     95.539      0.000       0.909       0.947
L1.sentiment       0.0006   5.92e-05     10.283      0.000       0.000       0.001
L1.e(ip)           0.0095      0.037      0.259      0.796      -0.062       0.081
L1.e(sentiment)   -0.0001      0.000     -0.836      0.403      -0.000       0.000
                           Results for equation sentiment
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const              0.3773      0.272      1.388      0.165      -0.155       0.910
L1.ip            -14.5753      5.375     -2.712      0.007     -25.109      -4.041
L1.sentiment       0.8795      0.023     37.840      0.000       0.834       0.925
L1.e(ip)          40.2063     18.695      2.151      0.032       3.565      76.847
L1.e(sentiment)    0.0411      0.051      0.800      0.424      -0.060       0.142
                            Error covariance matrix
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sqrt.var.ip          0.0128      0.000     41.131      0.000       0.012       0.013
sqrt.cov.ip.sentiment  0.0309      0.229      0.135      0.893      -0.418       0.480
sqrt.var.sentiment   5.2713      0.147     35.759      0.000       4.982       5.560
==============================================================================

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

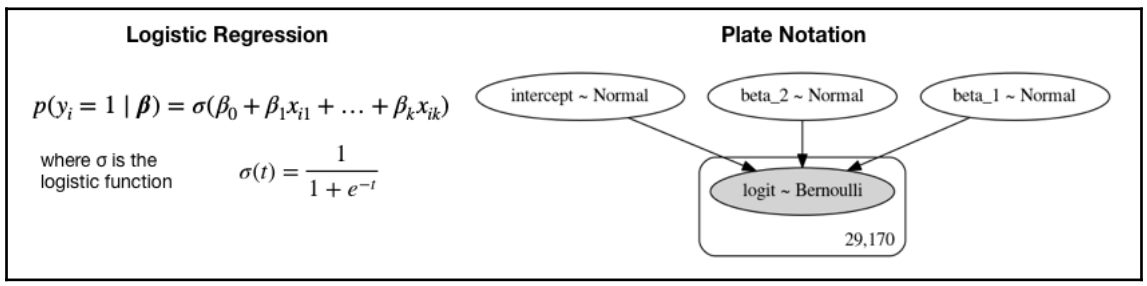

Industrial Production - Diagnostics
Standardized residual
Histogram plus estimated density
KDE
N(0,1)
Hist
Normal Q-Q
Sample Quantiles
Theoretical Quantiles
Correlogram

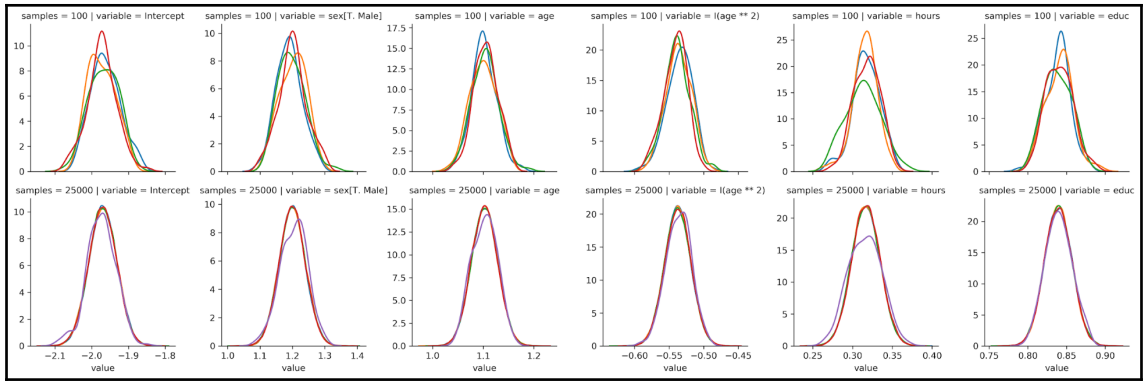

Industrial Production
actual
predicted
1.0
0.8
0.6
0.4
0.2
0.0
Jul
Jan
2015
Jul
Jan
2016
Jul
Jan
2017
Jul
Jan
2018
DATE
Sentiment
15
10
5
0
−5
Jul
Jan
2015
Jul
Jan
2016
Jul
Jan
2017
Jul
Jan
2018
DATE

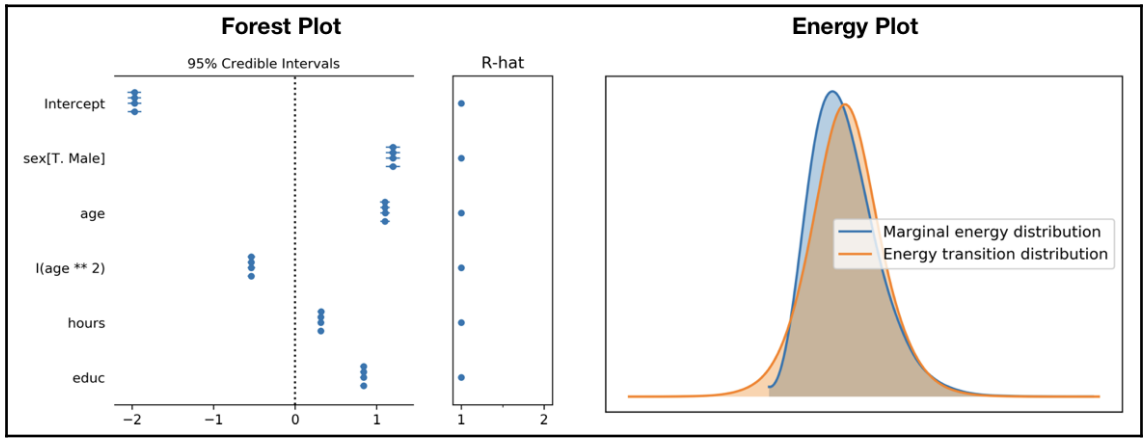
Chapter 9: Bayesian Machine Learning

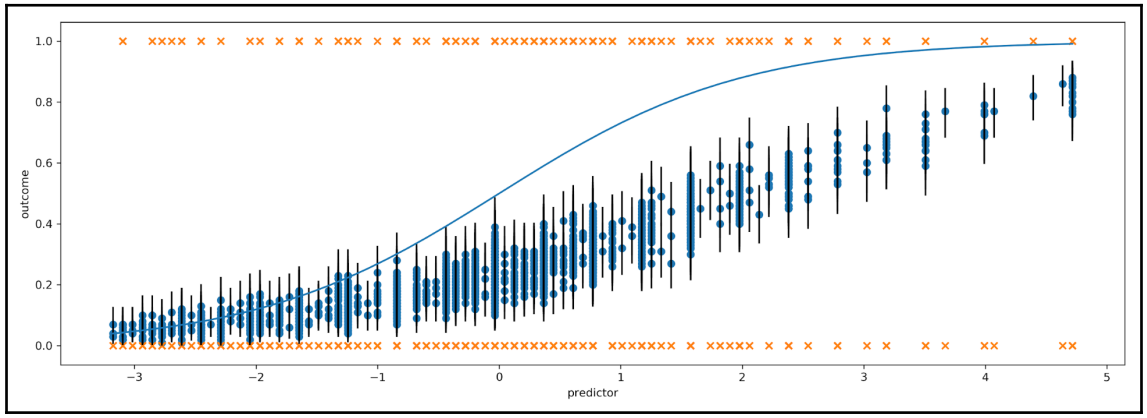


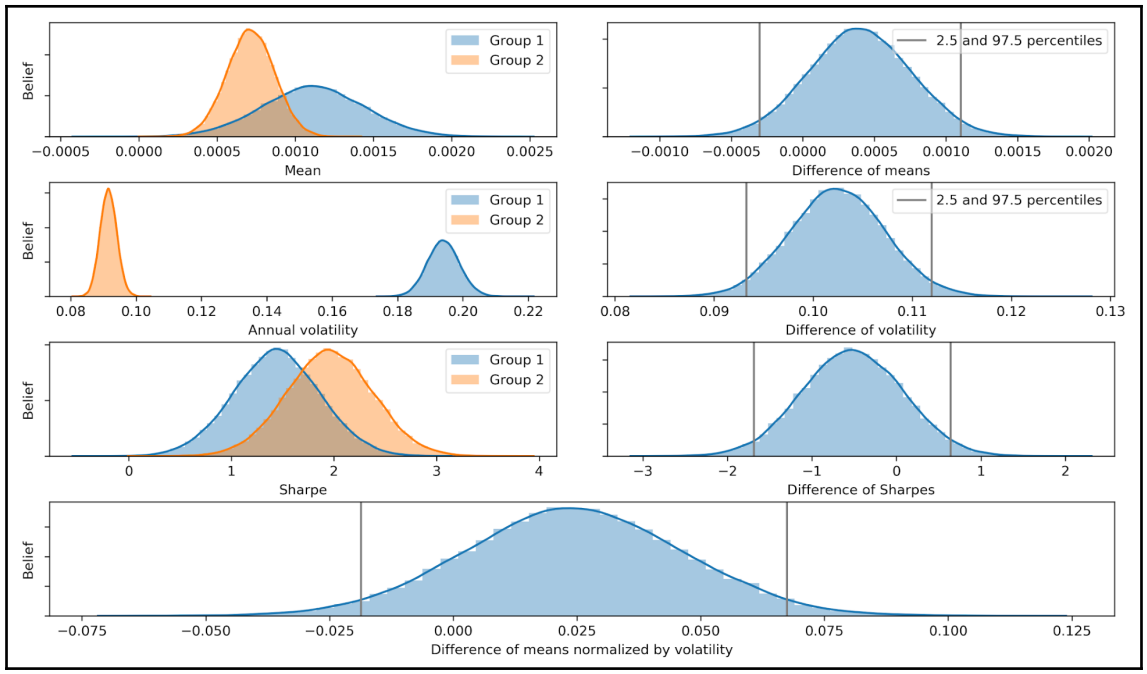




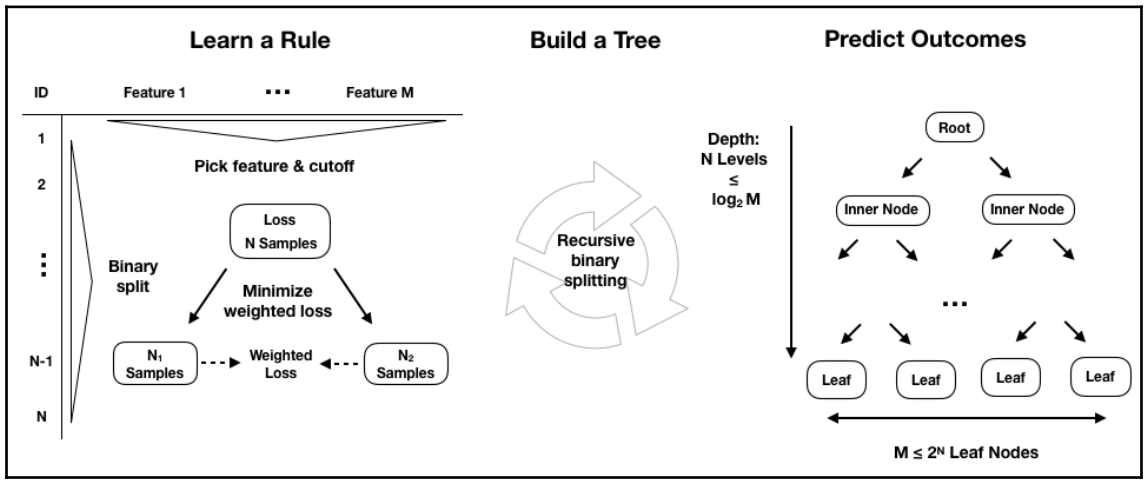

samples = 100 | variable = Intercept
samples = 100 | variable = sex[T. Male]
samples = 100 | variable = age
samples = 100 | variable = I(age ** 2)
samples = 100 | variable = hours
samples = 100 | variable = educ
samples = 25000 | variable = Intercept
samples = 25000 | variable = sex[T. Male]
samples = 25000 | variable = age
samples = 25000 | variable = I(age ** 2)
samples = 25000 | variable = hours
samples = 25000 | variable = educ
value

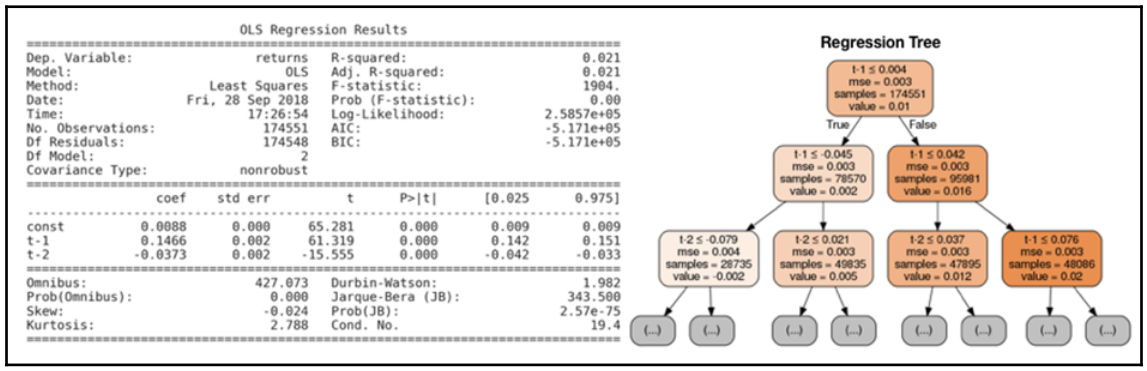

Forest Plot
95% Credible Intervals
R-hat
Intercept
sex[T. Male]
age
I(age ** 2)
hours
educ
Energy Plot
Marginal energy distribution
Energy transition distribution


outcome
predictor


Belief
Group 1
Group 2
−0.0005 0.0000 0.0005 0.0010 0.0015 0.0020 0.0025
Mean
2.5 and 97.5 percentiles
−0.0010 −0.0005 0.0000 0.0005 0.0010 0.0015 0.0020
Difference of means
0.08 0.10 0.12 0.14 0.16 0.18 0.20 0.22
Annual volatility
0.08 0.09 0.10 0.11 0.12 0.13
Difference of volatility
0 1 2 3 4
Sharpe
−3 −2 −1 0 1 2
Difference of Sharpes
−0.075 −0.050 −0.025 0.000 0.025 0.050 0.075 0.100 0.125
Difference of means normalized by volatility

Chapter 10: Decision Trees and Random Forests



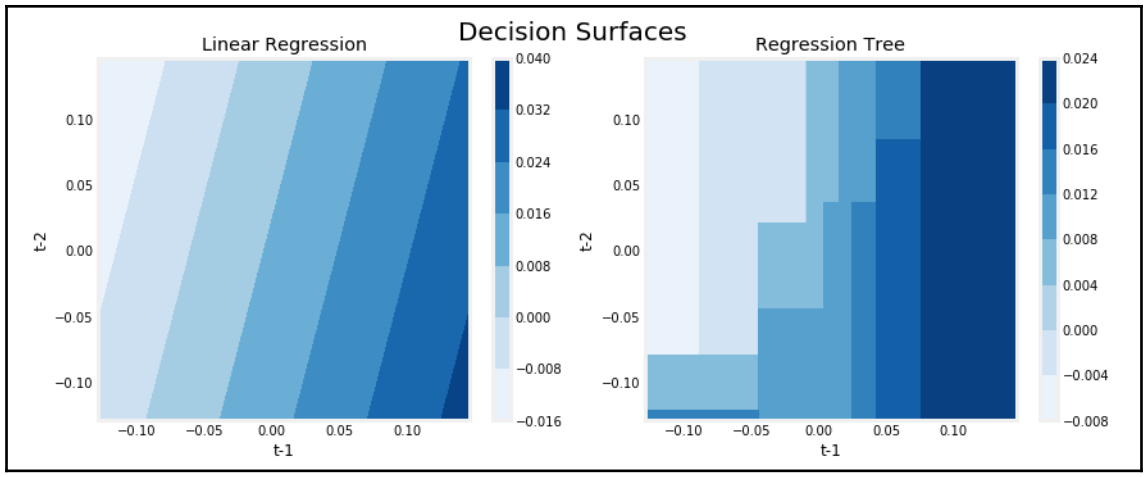


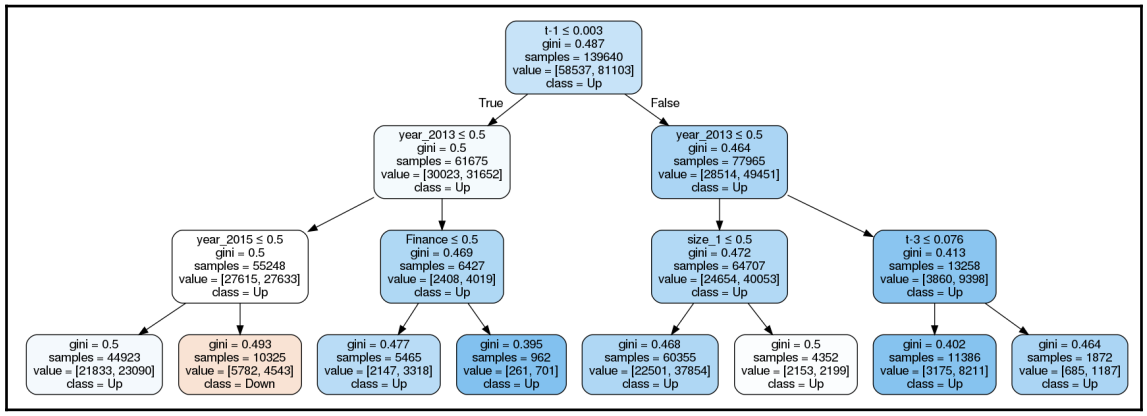
OLS Regression Results

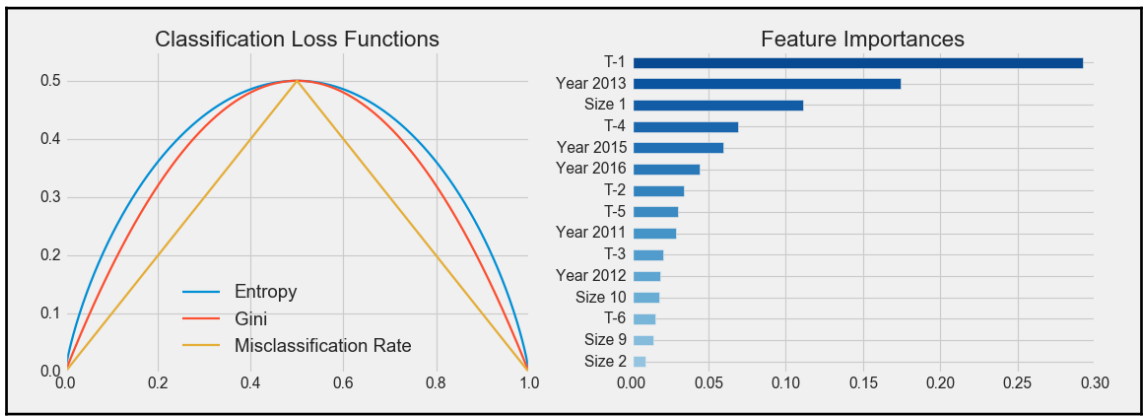
Dep. Variable:	returns	R-squared:	0.021
Model:	OLS	Adj. R-squared:	0.021
Method:	Least Squares	F-statistic:	1904.
Date:	Fri, 28 Sep 2018	Prob (F-statistic):	0.00
Time:	17:26:54	Log-Likelihood:	2.5857e+05
No. Observations:	174551	AIC:	-5.171e+05
Df Residuals:	174548	BIC:	-5.171e+05
Df Model:	2		
Covariance Type:	nonrobust		

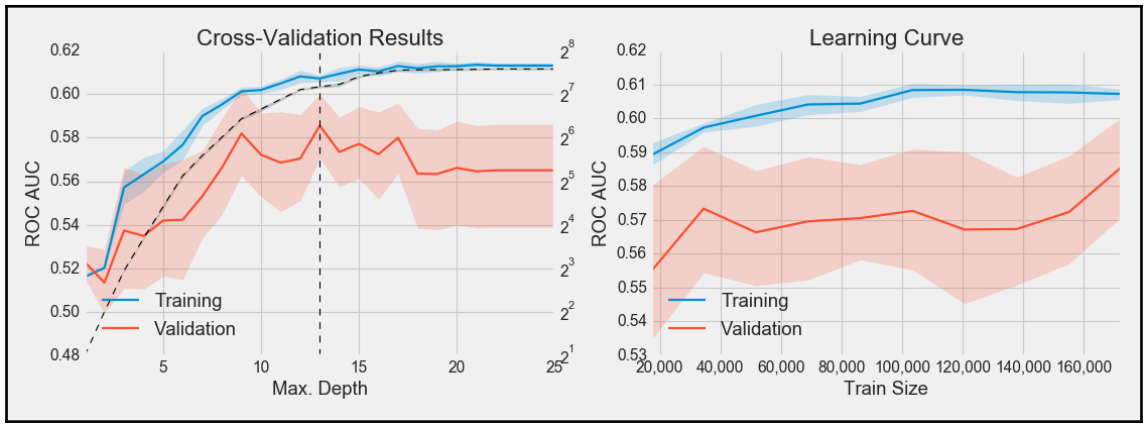
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0088	0.000	65.281	0.000	0.009	0.009
t-1	0.1466	0.002	61.319	0.000	0.142	0.151
t-2	-0.0373	0.002	-15.555	0.000	-0.042	-0.033

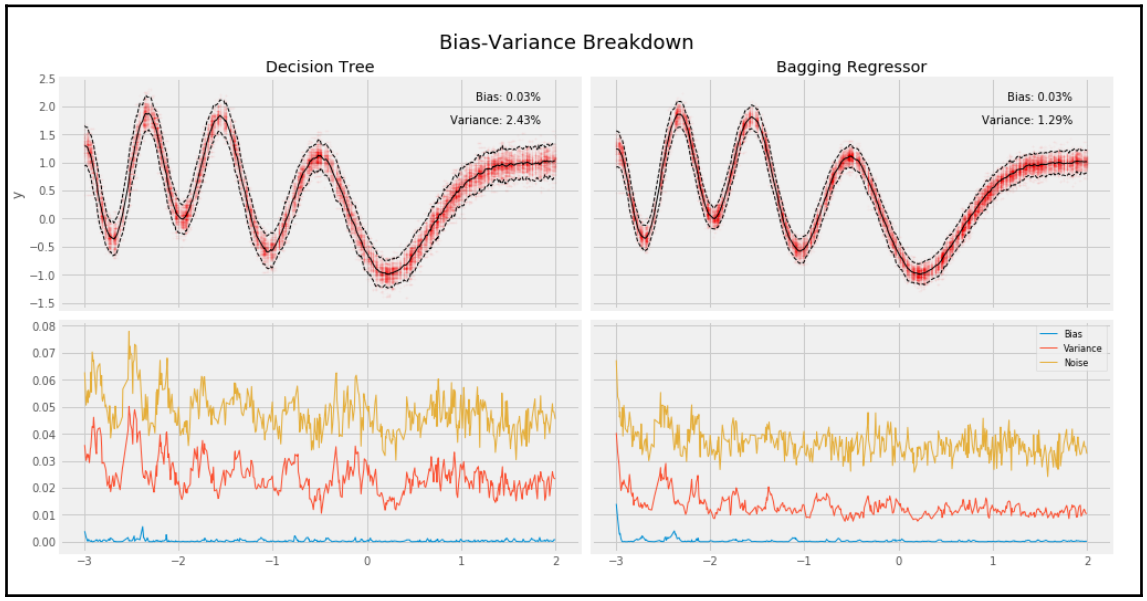
Omnibus:	427.073	Durbin-Watson:	1.982
Prob(Omnibus):	0.000	Jarque-Bera (JB):	343.500
Skew:	-0.024	Prob(JB):	2.57e-75
Kurtosis:	2.788	Cond. No.	19.4

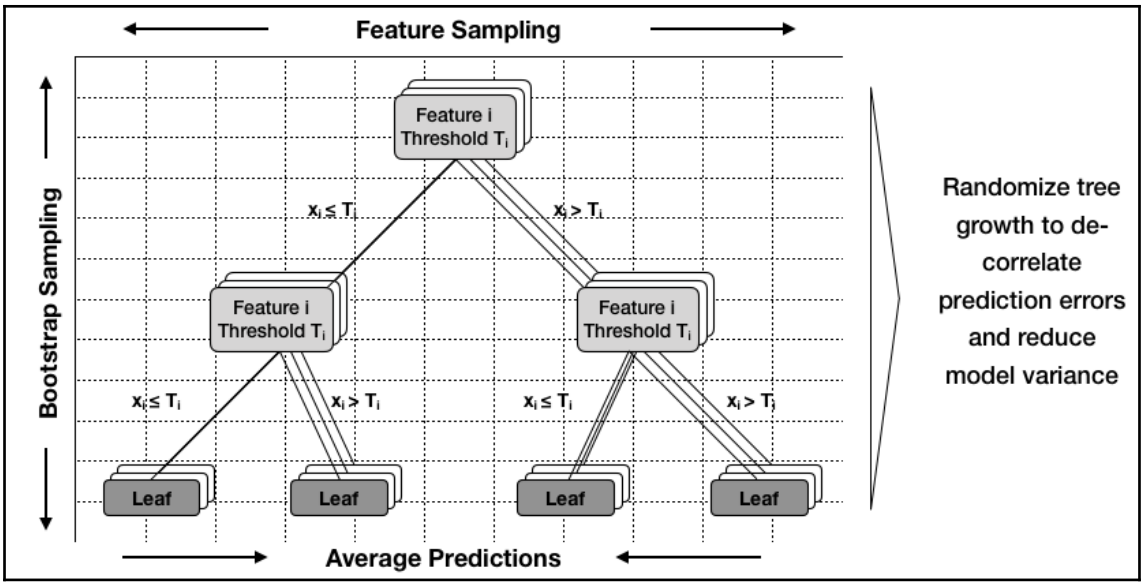

Decision Surfaces
Linear Regression
Regression Tree
t-1
t-2

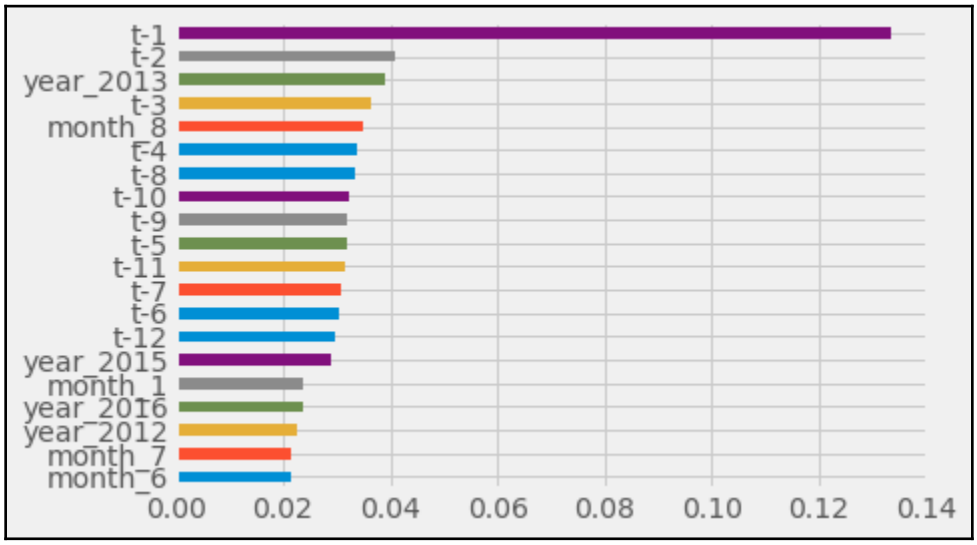

t-1 ≤ 0.003
gini = 0.487
samples = 139640
value = [58537, 81103]
class = Up
True
False
year_2013 ≤ 0.5
gini = 0.5
samples = 61675
value = [30023, 31652]
class = Up
year_2013 ≤ 0.5
gini = 0.464
samples = 77965
value = [28514, 49451]
class = Up
year_2015 ≤ 0.5
gini = 0.5
samples = 55248
value = [27615, 27633]
class = Up
Finance ≤ 0.5
gini = 0.469
samples = 6427
value = [2408, 4019]
class = Up
size_1 ≤ 0.5
gini = 0.472
samples = 64707
value = [24654, 40053]
class = Up
t-3 ≤ 0.076
gini = 0.413
samples = 13258
value = [3860, 9398]
class = Up
gini = 0.5
samples = 44923
value = [21833, 23090]
class = Up
gini = 0.493
samples = 10325
value = [5782, 4543]
class = Down
gini = 0.477
samples = 5465
value = [2147, 3318]
class = Up
gini = 0.395
samples = 962
value = [261, 701]
class = Up
gini = 0.468
samples = 60355
value = [22501, 37854]
class = Up
gini = 0.5
samples = 4352
value = [2153, 2199]
class = Up
gini = 0.402
samples = 11386
value = [3175, 8211]
class = Up
gini = 0.464
samples = 1872
value = [685, 1187]
class = Up


Classification Loss Functions
Entropy
Gini
Misclassification Rate
Feature Importances
T-1
Year 2013
Size 1
T-4
Year 2015
Year 2016
T-2
T-5
Year 2011
T-3
Year 2012
Size 10
T-6
Size 9
Size 2

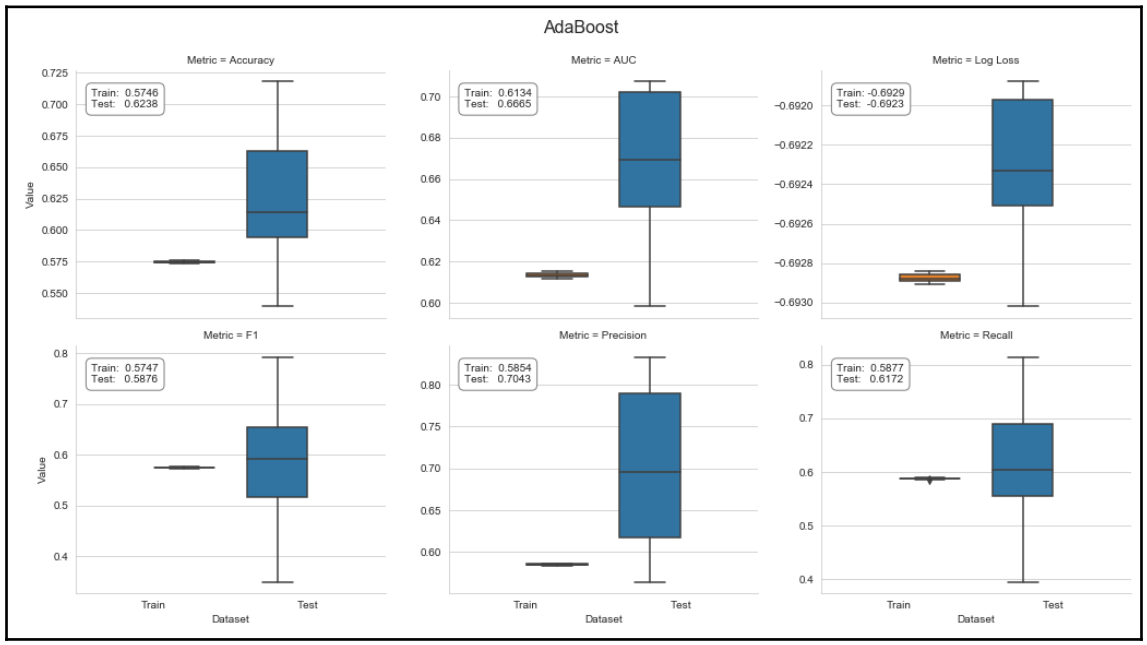

Cross-Validation Results
ROC AUC
Max. Depth
Training
Validation
Learning Curve
ROC AUC
Train Size
Training
Validation

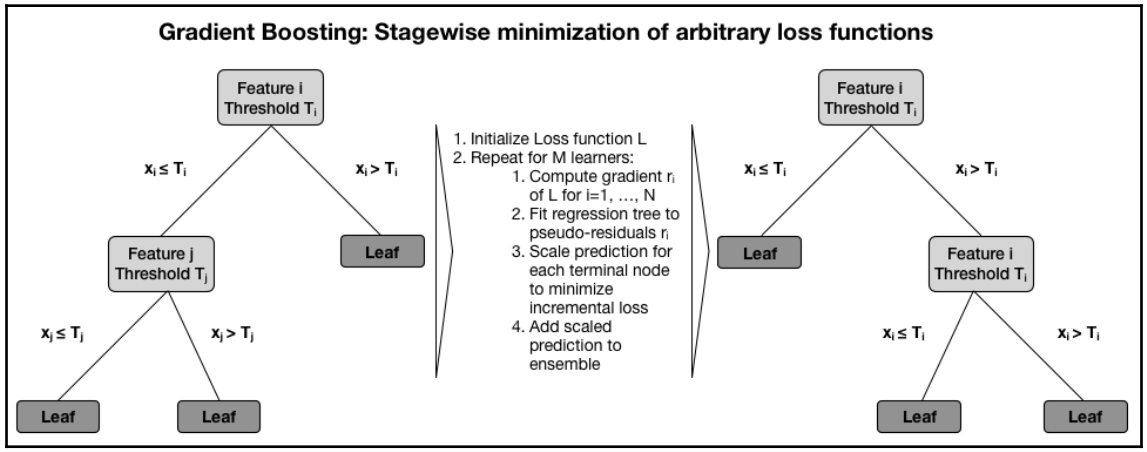

Bias-Variance Breakdown
Decision Tree
Bias: 0.03%
Variance: 2.43%
Bagging Regressor
Bias: 0.03%
Variance: 1.29%
Bias
Variance
Noise

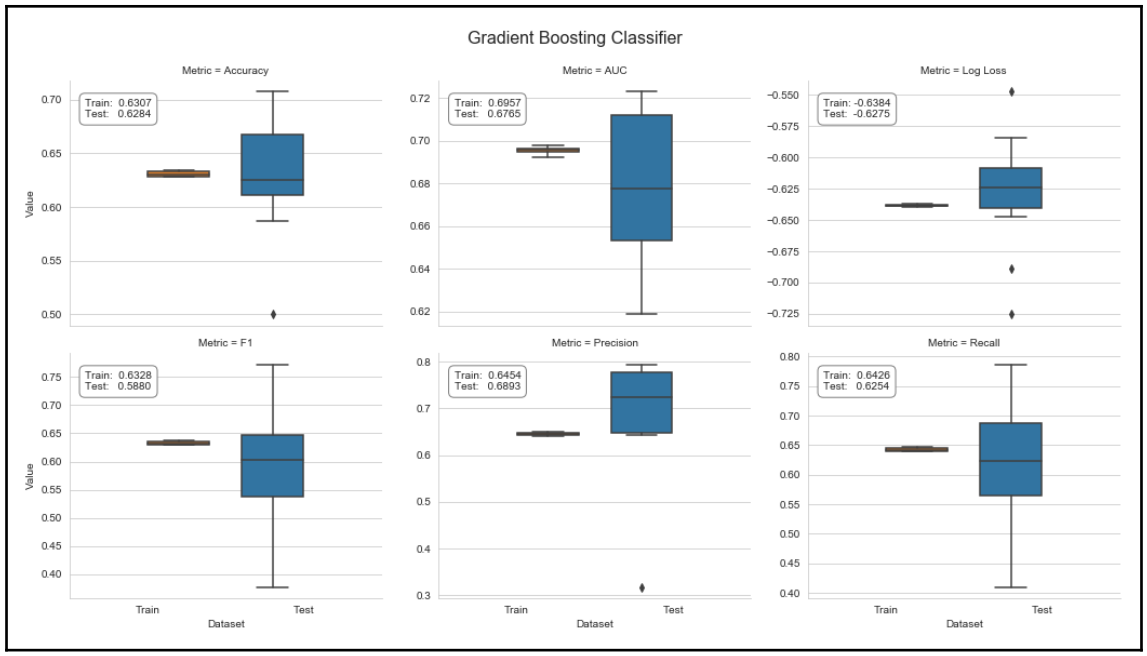

Feature Sampling
Bootstrap Sampling
Feature i
Threshold T_i
$x_i \le T_i$
$x_i > T_i$
Feature i
Threshold T_i
Feature i
Threshold T_i
$x_i \le T_i$
$x_i > T_i$
$x_i \le T_i$
$x_i > T_i$
Leaf
Leaf
Leaf
Leaf
Average Predictions
Randomize tree growth to de-correlate prediction errors and reduce model variance

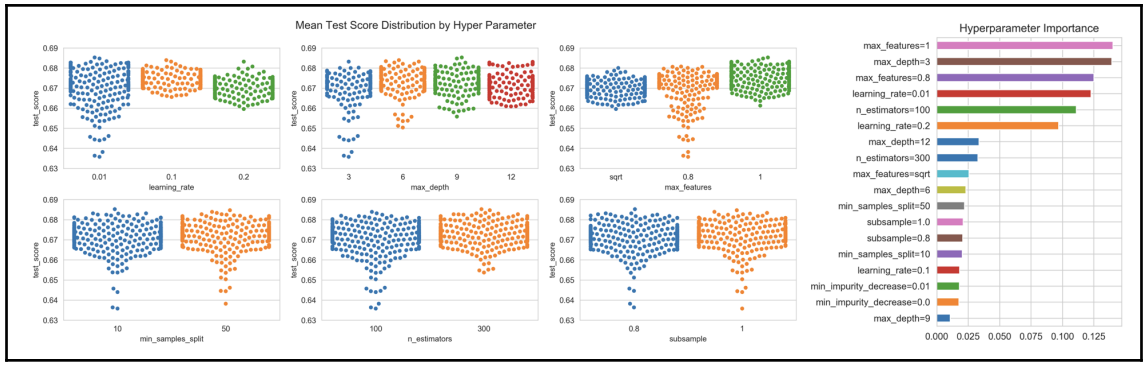

t-1
t-2
year_2013
t-3
month_8
t-4
t-8
t-10
t-9
t-5
t-11
t-7
t-6
t-12
year_2015
month_1
year_2016
year_2012
month_7
month_6
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14

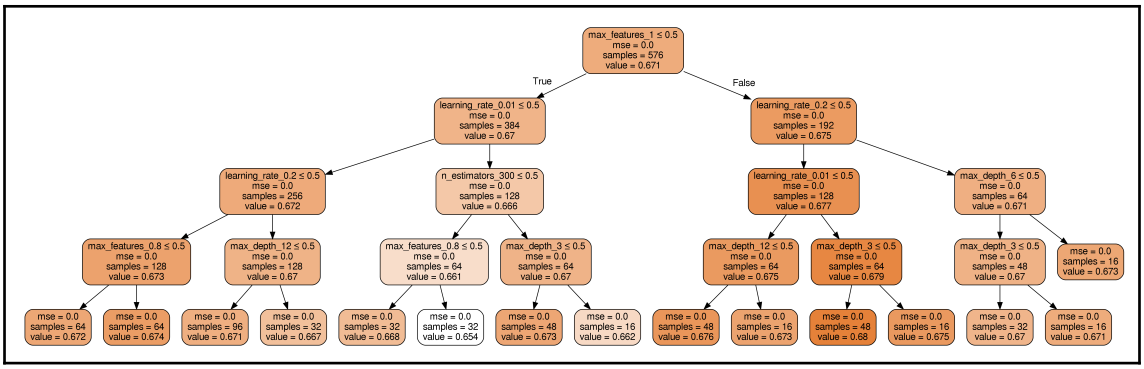
Chapter 11: Gradient Boosting Machines

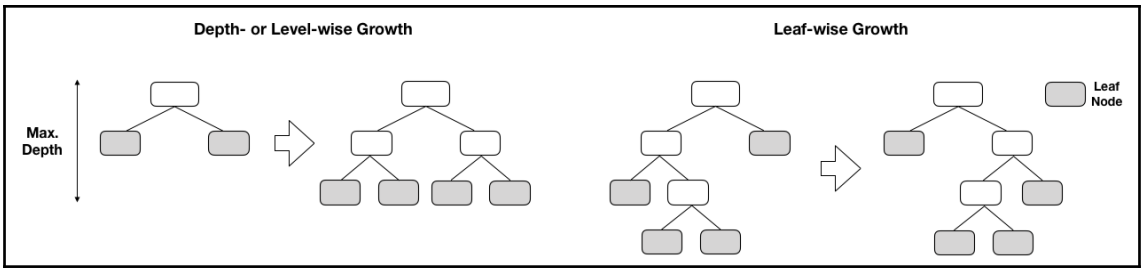


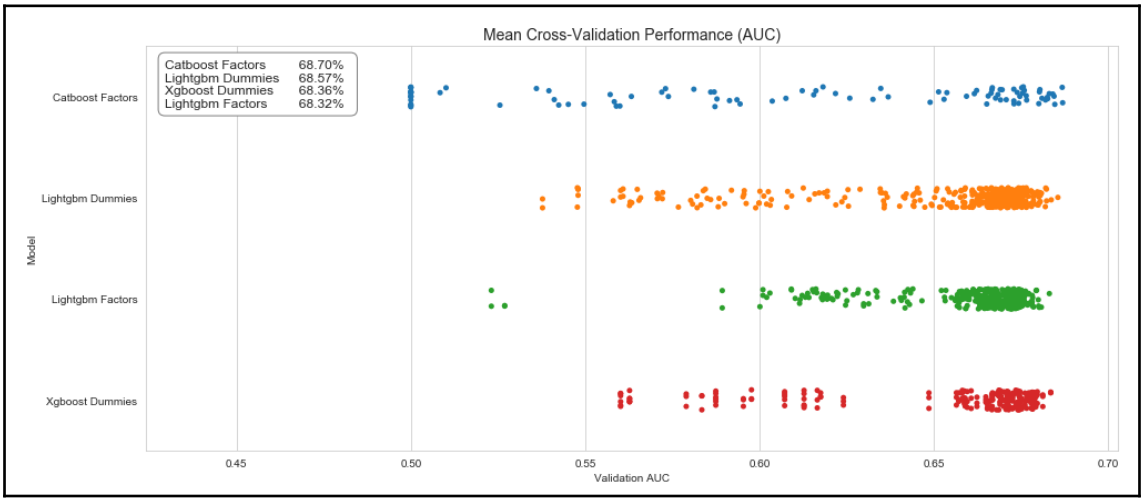


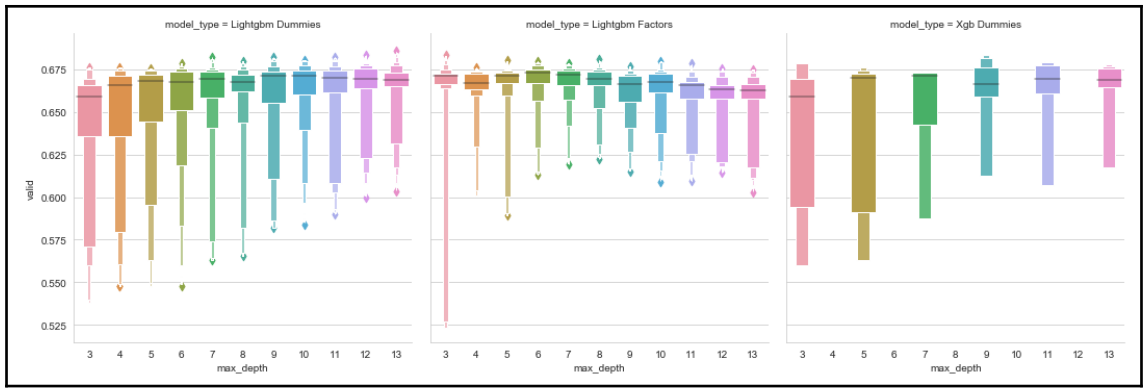

Gradient Boosting Classifier
Metric = Accuracy
Train: 0.6307
Test: 0.6284
Metric = AUC
Train: 0.6957
Test: 0.6765
Metric = Log Loss
Train: -0.6384
Test: -0.6275
Metric = F1
Train: 0.6328
Test: 0.5880
Metric = Precision
Train: 0.6454
Test: 0.6893
Metric = Recall
Train: 0.6426
Test: 0.6254
Value
Train
Test
Dataset


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

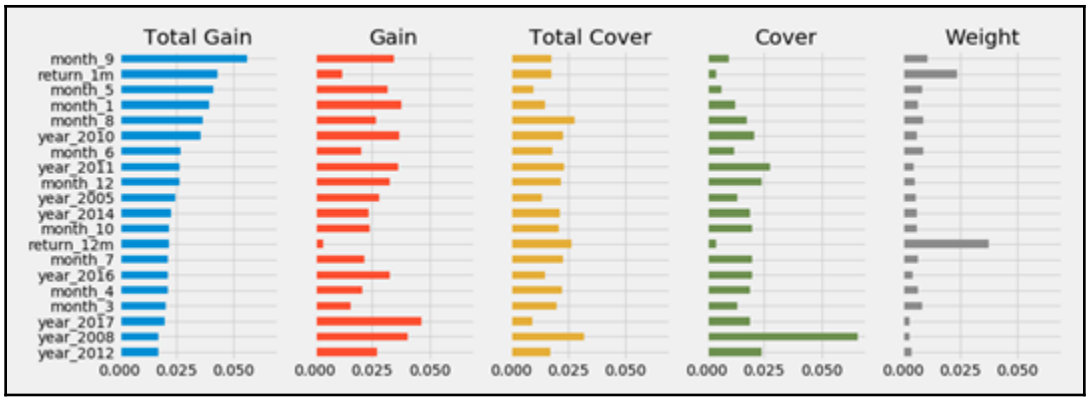

max_features_1 ≤ 0.5
mse = 0.0
samples = 576
value = 0.671
True
False
learning_rate_0.01 ≤ 0.5
mse = 0.0
samples = 384
value = 0.67
learning_rate_0.2 ≤ 0.5
mse = 0.0
samples = 192
value = 0.675
learning_rate_0.2 ≤ 0.5
mse = 0.0
samples = 256
value = 0.672
n_estimators_300 ≤ 0.5
mse = 0.0
samples = 128
value = 0.666
learning_rate_0.01 ≤ 0.5
mse = 0.0
samples = 128
value = 0.677
max_depth_6 ≤ 0.5
mse = 0.0
samples = 64
value = 0.671
max_features_0.8 ≤ 0.5
mse = 0.0
samples = 128
value = 0.673
max_depth_12 ≤ 0.5
mse = 0.0
samples = 128
value = 0.67
max_features_0.8 ≤ 0.5
mse = 0.0
samples = 64
value = 0.661
max_depth_3 ≤ 0.5
mse = 0.0
samples = 64
value = 0.67
max_depth_12 ≤ 0.5
mse = 0.0
samples = 64
value = 0.675
max_depth_3 ≤ 0.5
mse = 0.0
samples = 64
value = 0.679
max_depth_3 ≤ 0.5
mse = 0.0
samples = 48
value = 0.67
mse = 0.0
samples = 16
value = 0.673
mse = 0.0
samples = 64
value = 0.672
mse = 0.0
samples = 64
value = 0.674
mse = 0.0
samples = 96
value = 0.671
mse = 0.0
samples = 32
value = 0.667
mse = 0.0
samples = 32
value = 0.668
mse = 0.0
samples = 32
value = 0.654
mse = 0.0
samples = 48
value = 0.673
mse = 0.0
samples = 16
value = 0.662
mse = 0.0
samples = 48
value = 0.676
mse = 0.0
samples = 16
value = 0.673
mse = 0.0
samples = 48
value = 0.68
mse = 0.0
samples = 16
value = 0.675
mse = 0.0
samples = 32
value = 0.67
mse = 0.0
samples = 16
value = 0.671

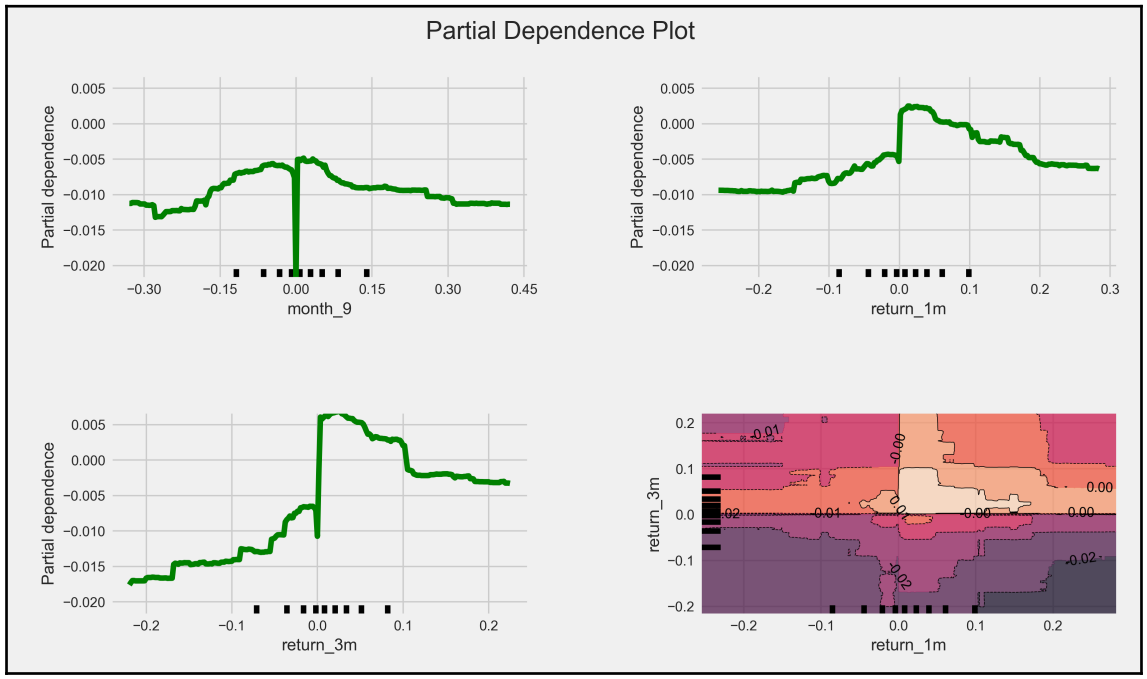

Depth- or Level-wise Growth
Leaf-wise Growth
Max. Depth
Leaf Node

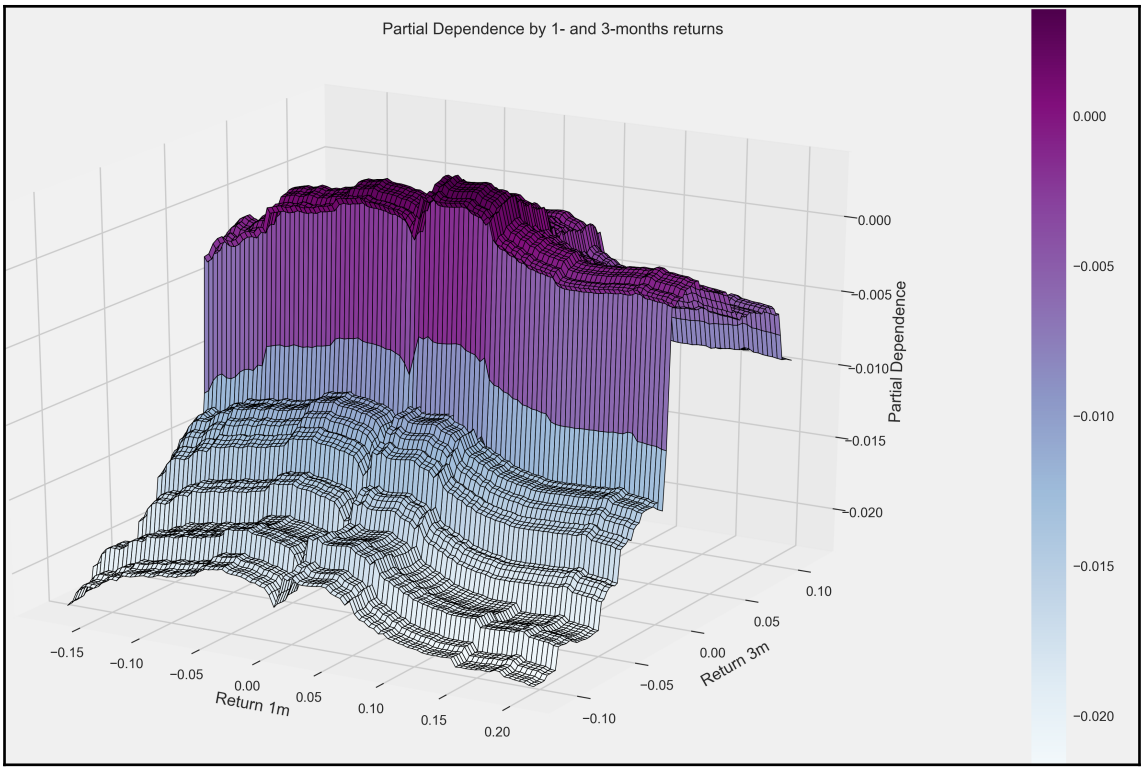

Mean Cross-Validation Performance (AUC)
Catboost Factors 68.70%
Lightgbm Dummies 68.57%
Xgboost Dummies 68.36%
Lightgbm Factors 68.32%
Catboost Factors
Lightgbm Dummies
Lightgbm Factors
Xgboost Dummies
Model
0.45
0.50
0.55
0.60
0.65
0.70
Validation AUC

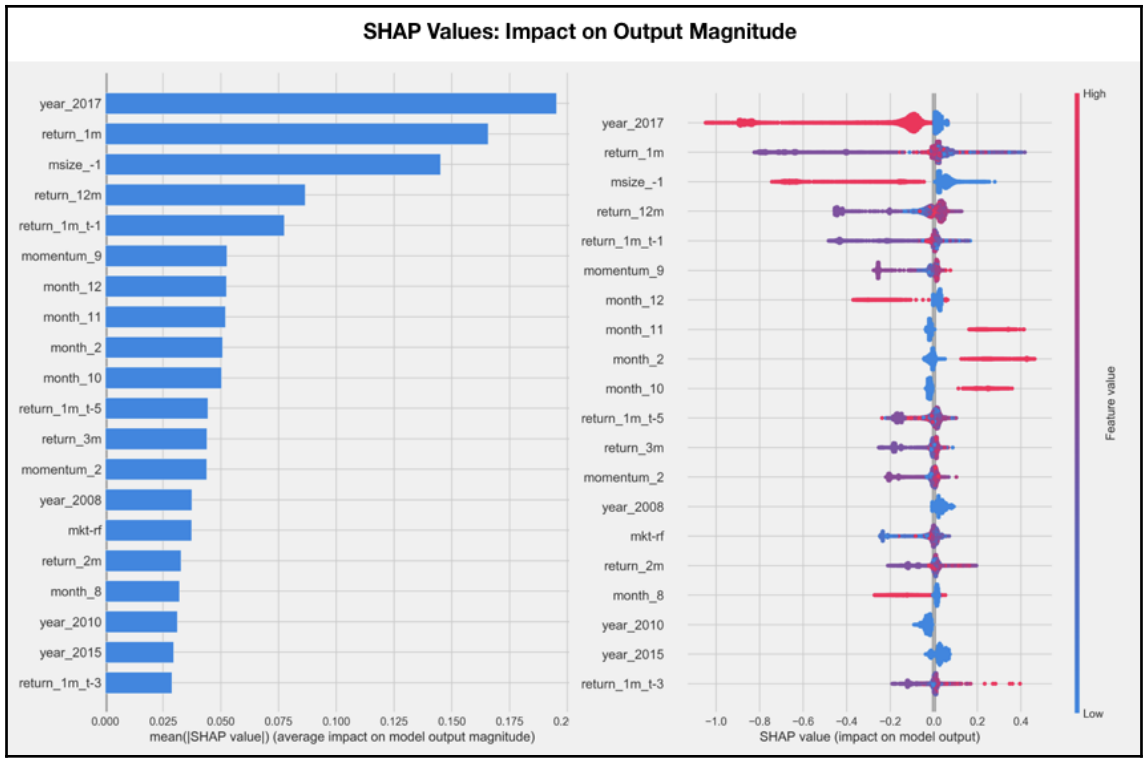

model_type = Lightgbm Dummies
model_type = Lightgbm Factors
model_type = Xgb Dummies
valid
0.675
0.650
0.625
0.600
0.575
0.550
0.525
3 4 5 6 7 8 9 10 11 12 13
max_depth

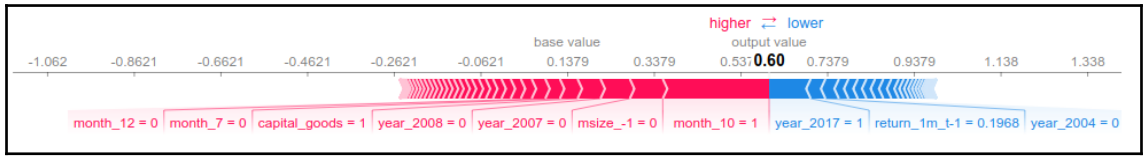
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   valid   R-squared:                       0.687
Model:                             OLS   Adj. R-squared:                  0.673
Method:                  Least Squares   F-statistic:                     26.94
Date:                 Wed, 24 Oct 2018   Prob (F-statistic):           7.92e-55
Time:                         14:03:45   Log-Likelihood:                1018.7
No. Observations:                  396   AIC:                            -2001.
Df Residuals:                      378   BIC:                            -1930.
Df Model:                           17
Covariance Type:                   HC3
==============================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                  0.6145      0.005    127.970      0.000       0.605       0.624
boosting_gbtree        0.0056      0.002      2.866      0.004       0.002       0.009
learning_rate_0.1      0.0501      0.003     18.977      0.000       0.045       0.055
learning_rate_0.3      0.0516      0.003     19.150      0.000       0.046       0.057
max_depth_4            0.0060      0.005      1.094      0.274      -0.005       0.017
max_depth_5            0.0096      0.005      1.823      0.068      -0.001       0.020
max_depth_6            0.0153      0.005      3.024      0.002       0.005       0.025
max_depth_7            0.0194      0.005      3.753      0.000       0.009       0.030
max_depth_8            0.0196      0.005      3.733      0.000       0.009       0.030
max_depth_9            0.0266      0.005      5.176      0.000       0.017       0.037
max_depth_10           0.0307      0.005      5.954      0.000       0.021       0.041
max_depth_11           0.0285      0.005      5.484      0.000       0.018       0.039
max_depth_12           0.0312      0.005      6.178      0.000       0.021       0.041
max_depth_13           0.0320      0.005      6.218      0.000       0.022       0.042
colsample_bytree_0.8  -0.0112      0.003     -4.143      0.000      -0.017      -0.006
colsample_bytree_1.0  -0.0278      0.003     -8.388      0.000      -0.034      -0.021
min_gain_to_split_1   -0.0009      0.003     -0.307      0.759      -0.006       0.005
min_gain_to_split_5   -0.0016      0.002     -0.726      0.468      -0.006       0.003
==============================================================================
Omnibus:                       11.763   Durbin-Watson:                   0.856
Prob(Omnibus):                  0.003   Jarque-Bera (JB):               11.104
Skew:                          -0.361   Prob(JB):                      0.00388
Kurtosis:                       2.609   Cond. No.                         17.1
==============================================================================

Warnings:
[1] Standard Errors are heteroscedasticity robust (HC3)
```

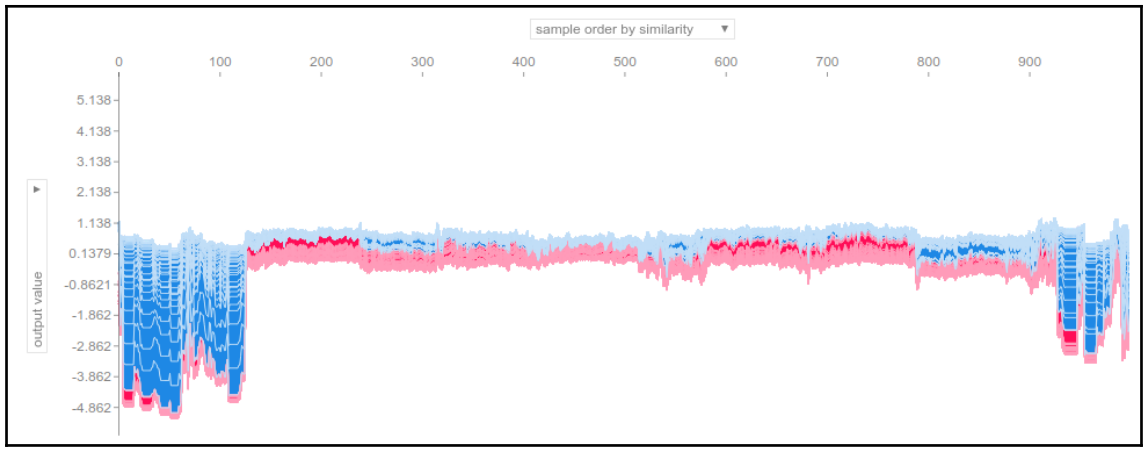

Total Gain
Gain
Total Cover
Cover
Weight
month_9
return_1m
month_5
month_1
month_8
year_2010
month_6
year_2011
month_12
year_2005
year_2014
month_10
return_12m
month_7
year_2016
month_4
month_3
year_2017
year_2008
year_2012
0.000 0.025 0.050

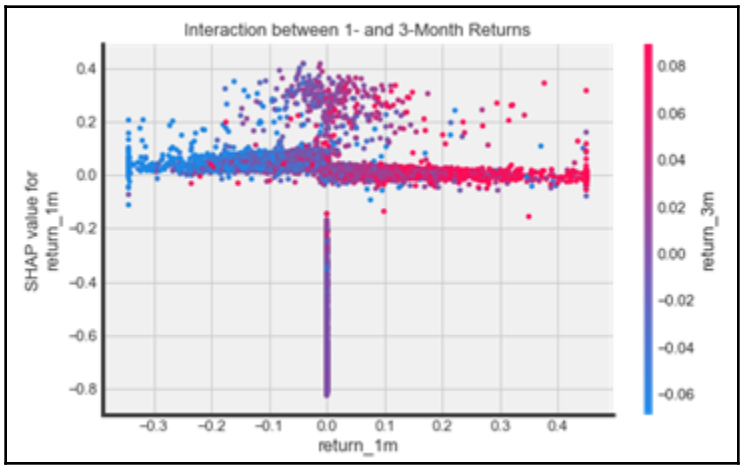

Partial Dependence Plot
Partial dependence
month_9
return_1m
return_3m
return_3m
return_1m


Partial Dependence by 1- and 3-months returns
Partial Dependence
0.000
−0.005
−0.010
−0.015
−0.020
Return 1m
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
Return 3m
−0.10
−0.05
0.00
0.05
0.10

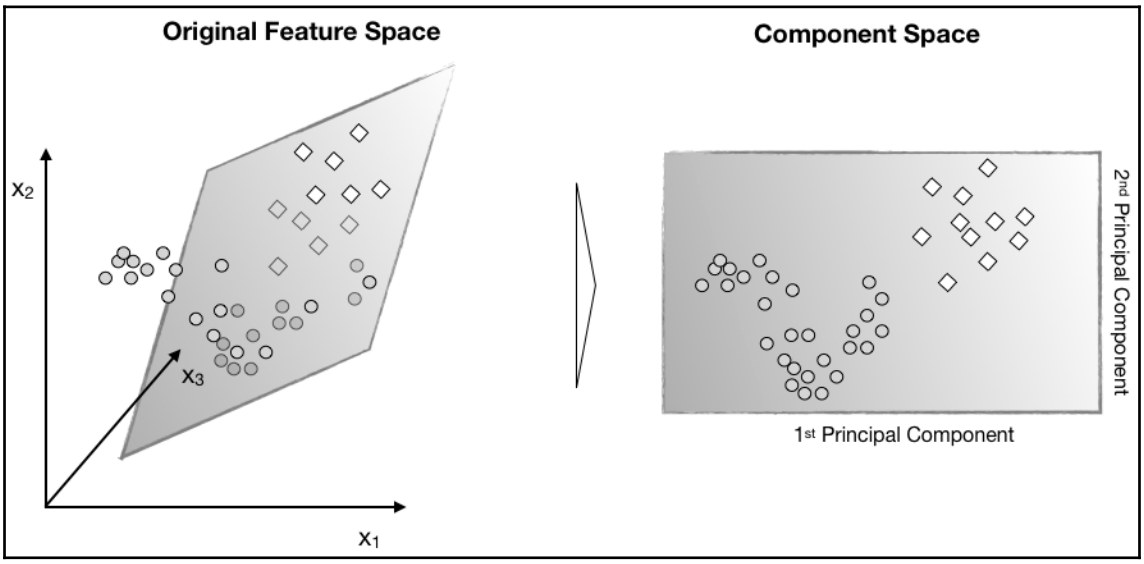

SHAP Values: Impact on Output Magnitude
year_2017
return_1m
msize_-1
return_12m
return_1m_t-1
momentum_9
month_12
month_11
month_2
month_10
return_1m_t-5
return_3m
momentum_2
year_2008
mkt-rf
return_2m
month_8
year_2010
year_2015
return_1m_t-3
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.2
mean(|SHAP value|) (average impact on model output magnitude)
−1.0 −0.8 −0.6 −0.4 −0.2 0.0 0.2 0.4
SHAP value (impact on model output)
High
Feature value
Low

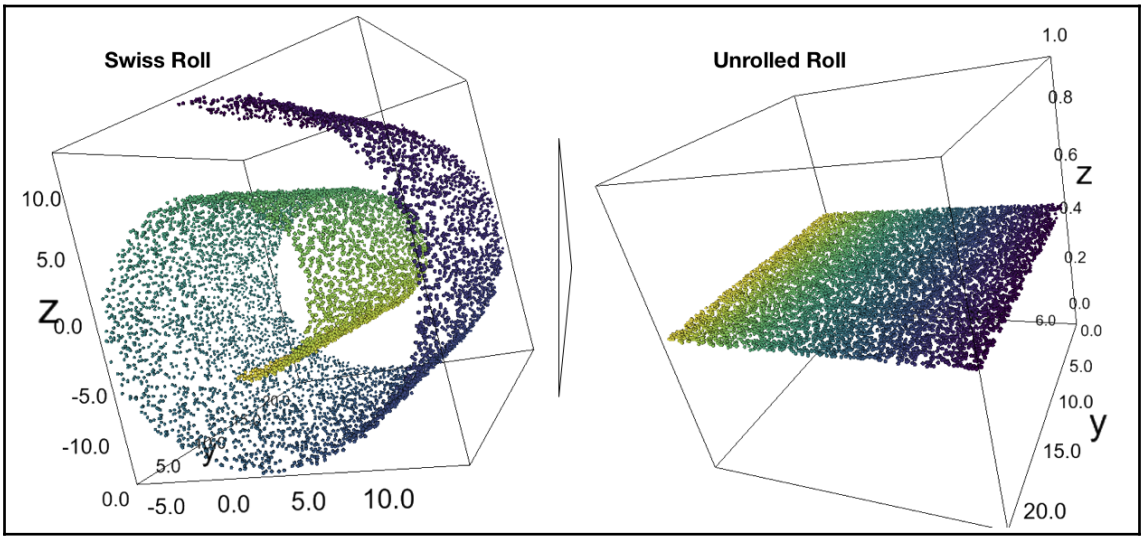

higher ⇄ lower
base value
output value
-1.062 -0.8621 -0.6621 -0.4621 -0.2621 -0.0621 0.1379 0.3379 0.537 0.60 0.7379 0.9379 1.138 1.338
month_12 = 0
month_7 = 0
capital_goods = 1
year_2008 = 0
year_2007 = 0
msize_-1 = 0
month_10 = 1
year_2017 = 1
return_1m_t-1 = 0.1968
year_2004 = 0

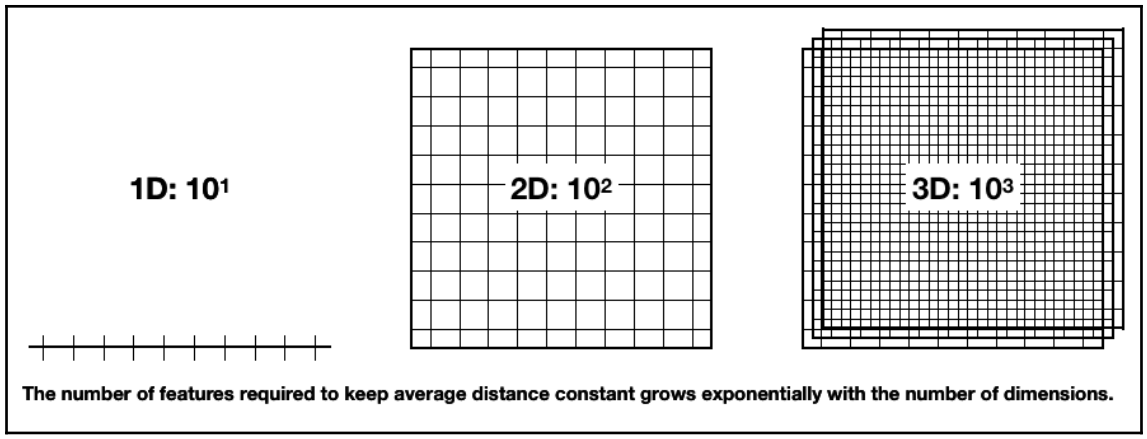

sample order by similarity
0
100
200
300
400
500
600
700
800
900
5.138
4.138
3.138
2.138
1.138
0.1379
-0.8621
-1.862
-2.862
-3.862
-4.862
output value


Interaction between 1- and 3-Month Returns
SHAP value for return_1m
return_1m
return_3m
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
0.08
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06

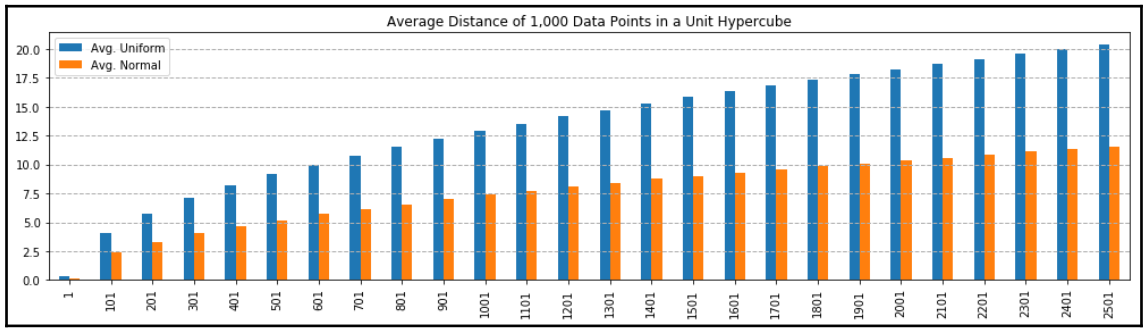
Chapter 12: Unsupervised Learning

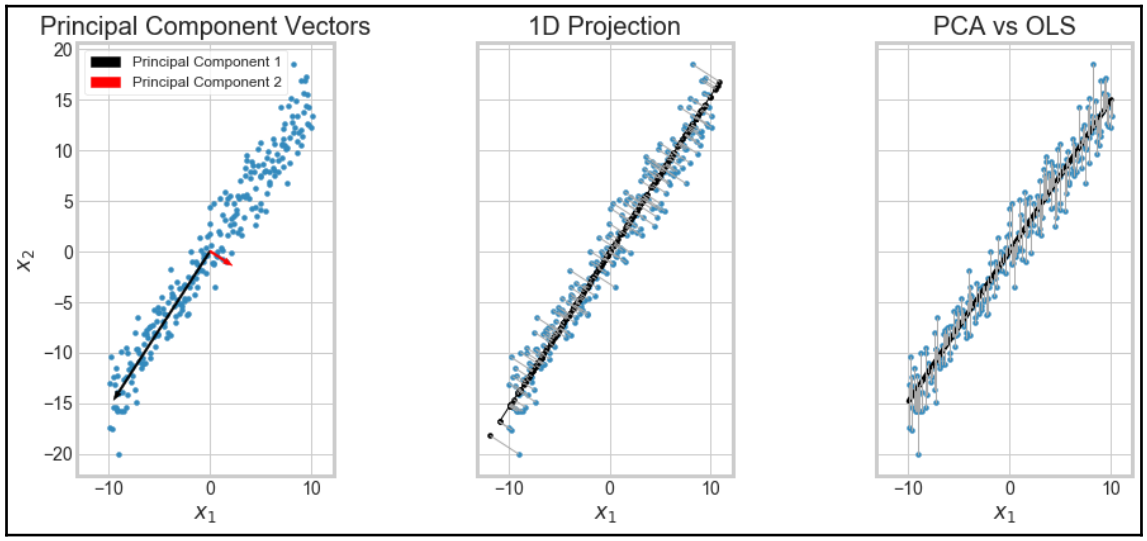


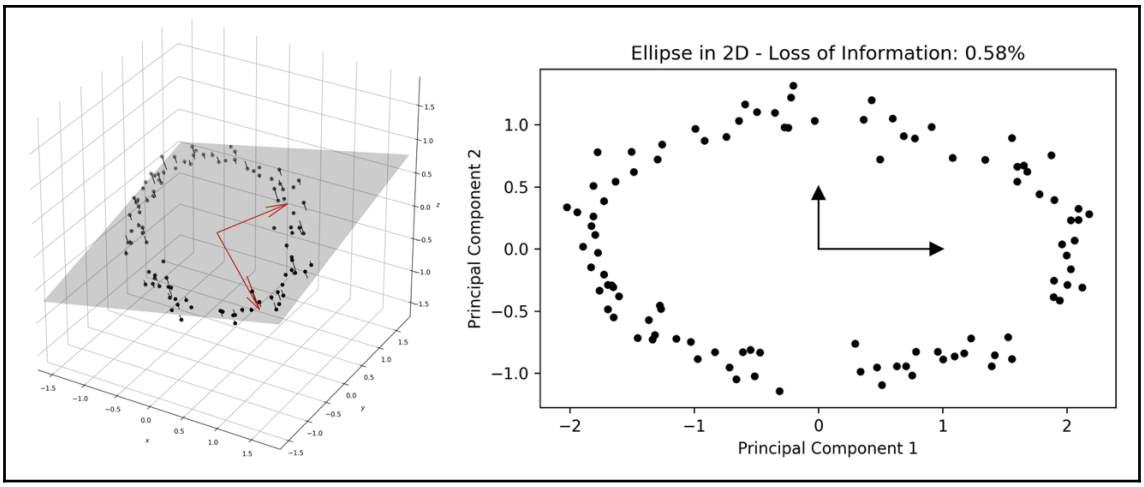


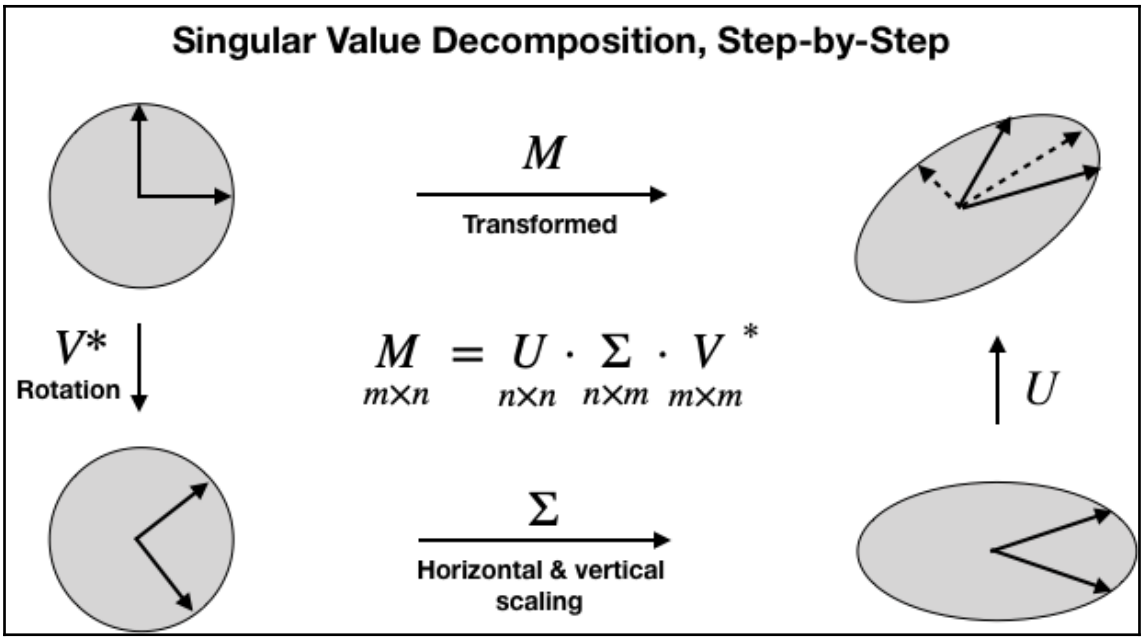


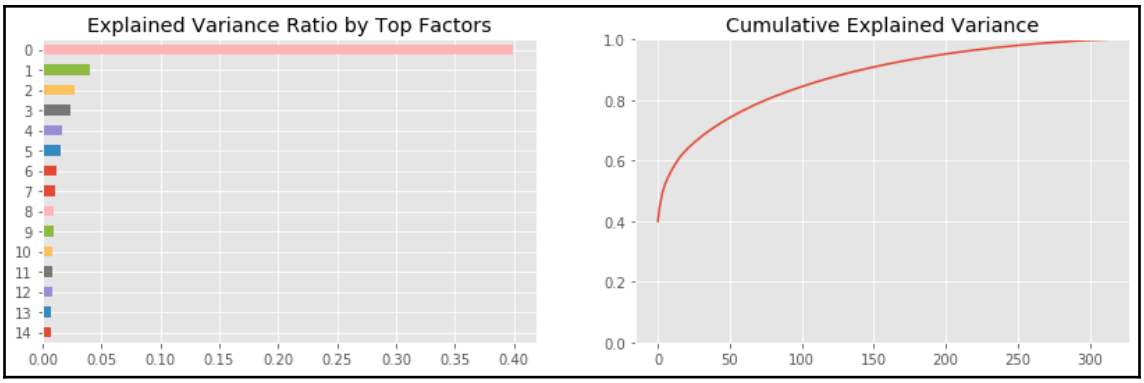
The number of features required to keep average distance constant grows exponentially with the number of dimensions.

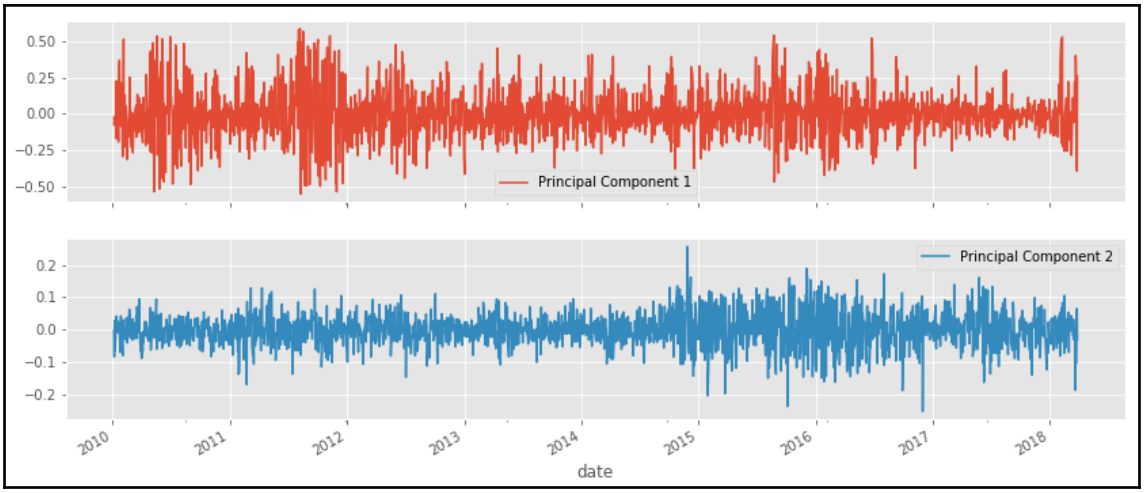


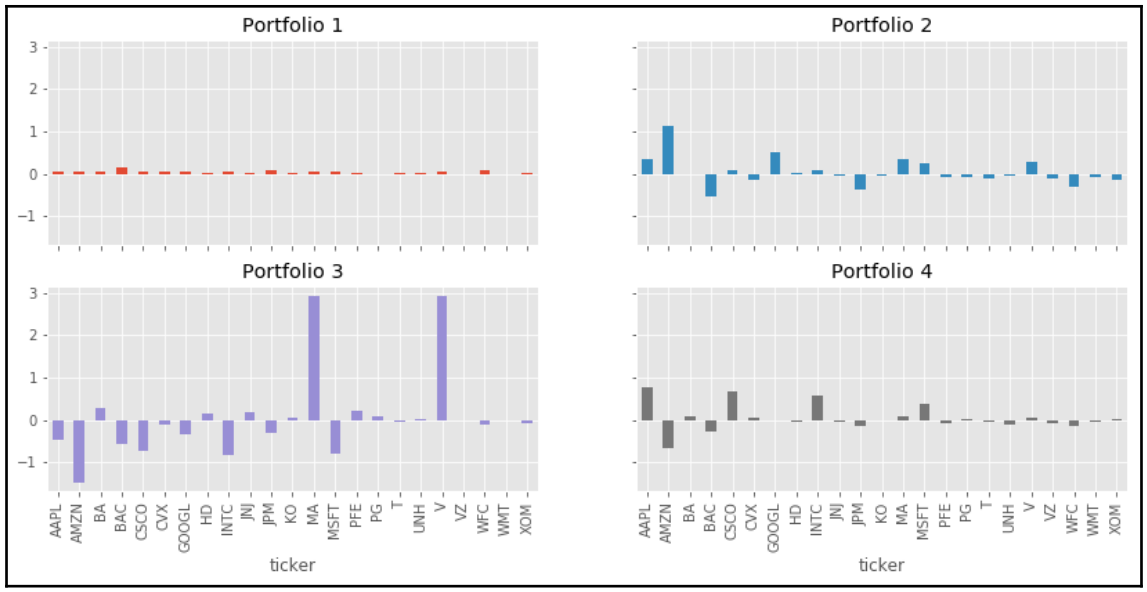


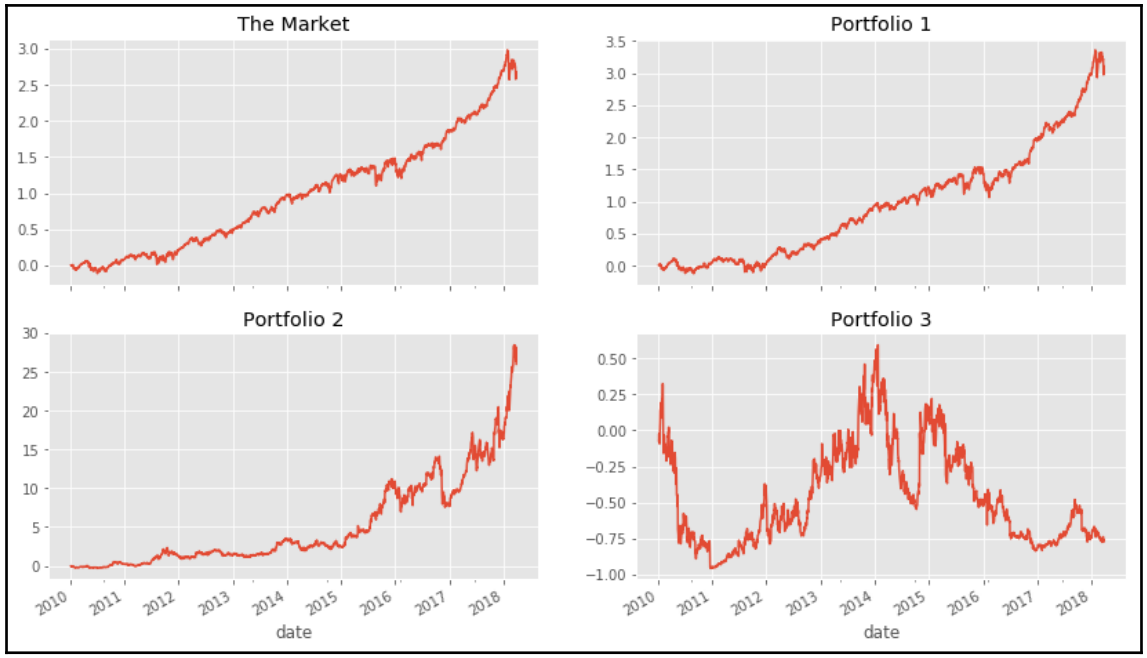

Ellipse in 2D - Loss of Information: 0.58%
Principal Component 2
Principal Component 1
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2

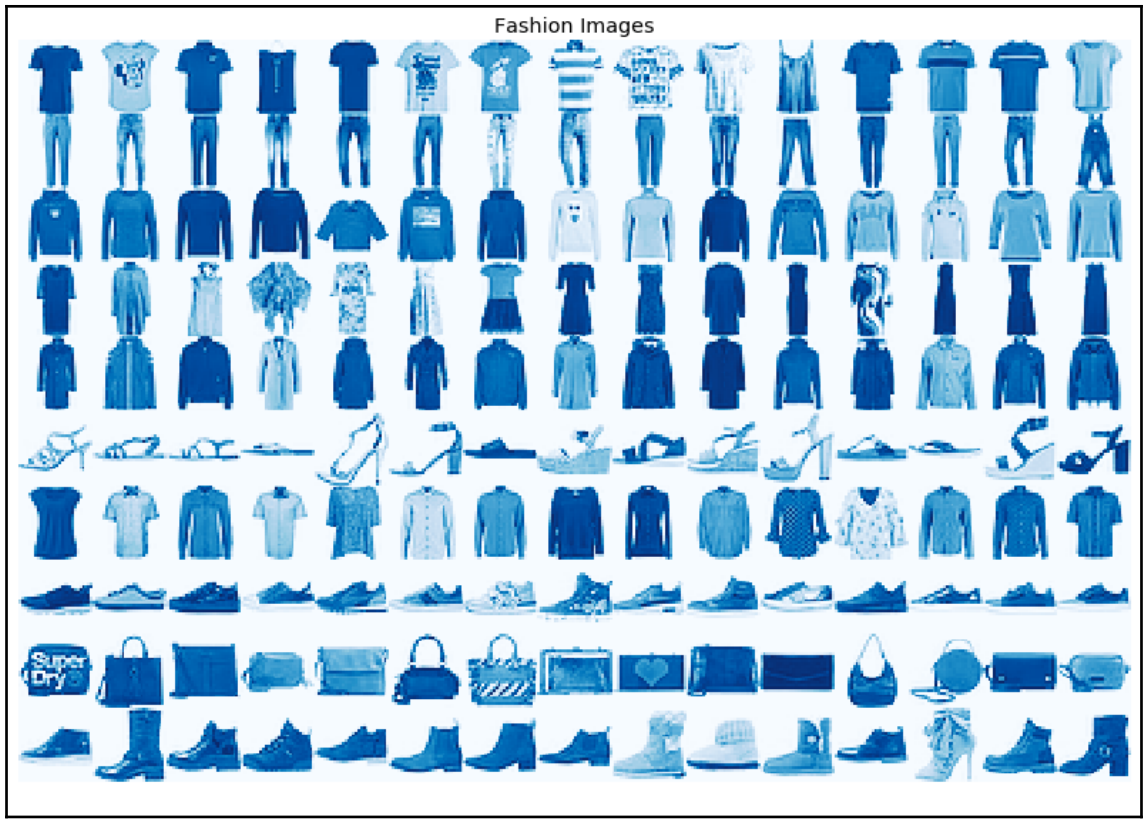

Singular Value Decomposition, Step-by-Step
M
Transformed
V*
Rotation
M = U · Σ · V*
m×n n×n n×m m×m
U
Σ
Horizontal & vertical scaling

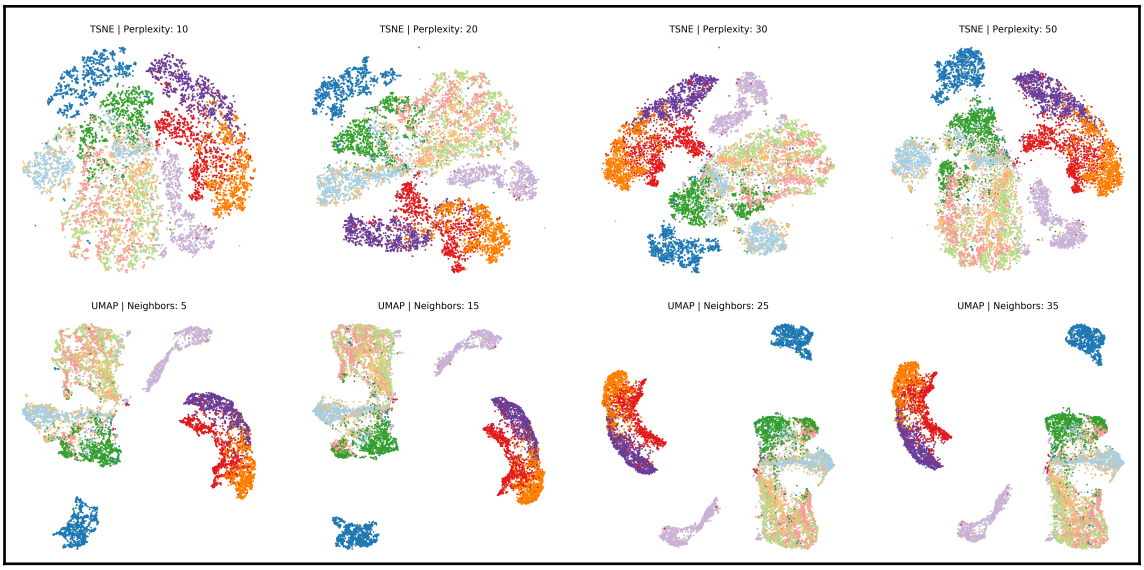

Explained Variance Ratio by Top Factors
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Cumulative Explained Variance
1.0
0.8
0.6
0.4
0.2
0.0
0
50
100
150
200
250
300

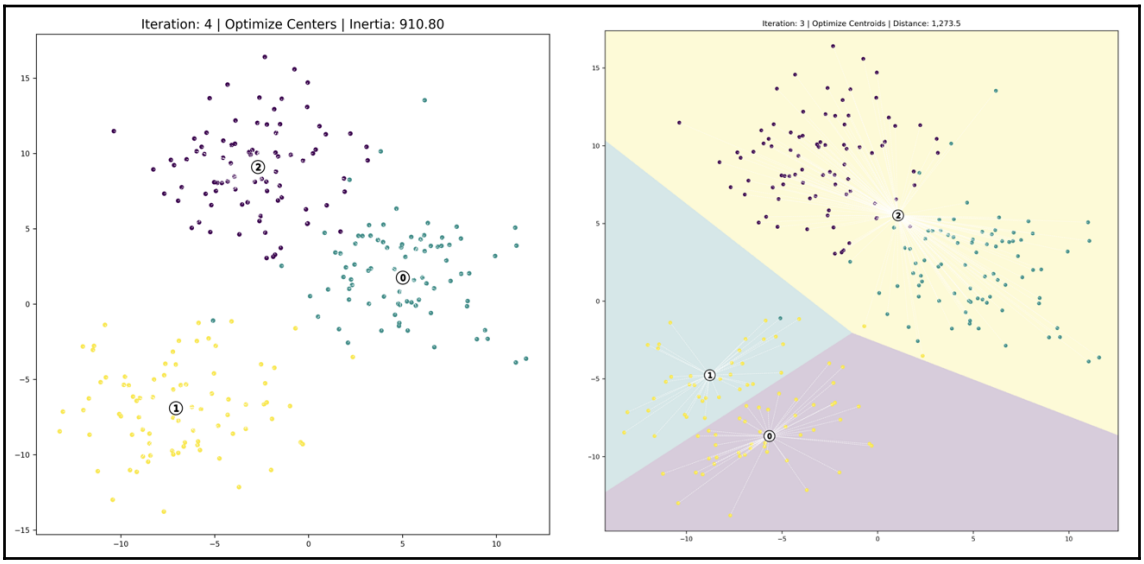

0.50
0.25
0.00
−0.25
−0.50
Principal Component 1
Principal Component 2
0.2
0.1
0.0
−0.1
−0.2
2010
2011
2012
2013
2014
2015
2016
2017
2018
date

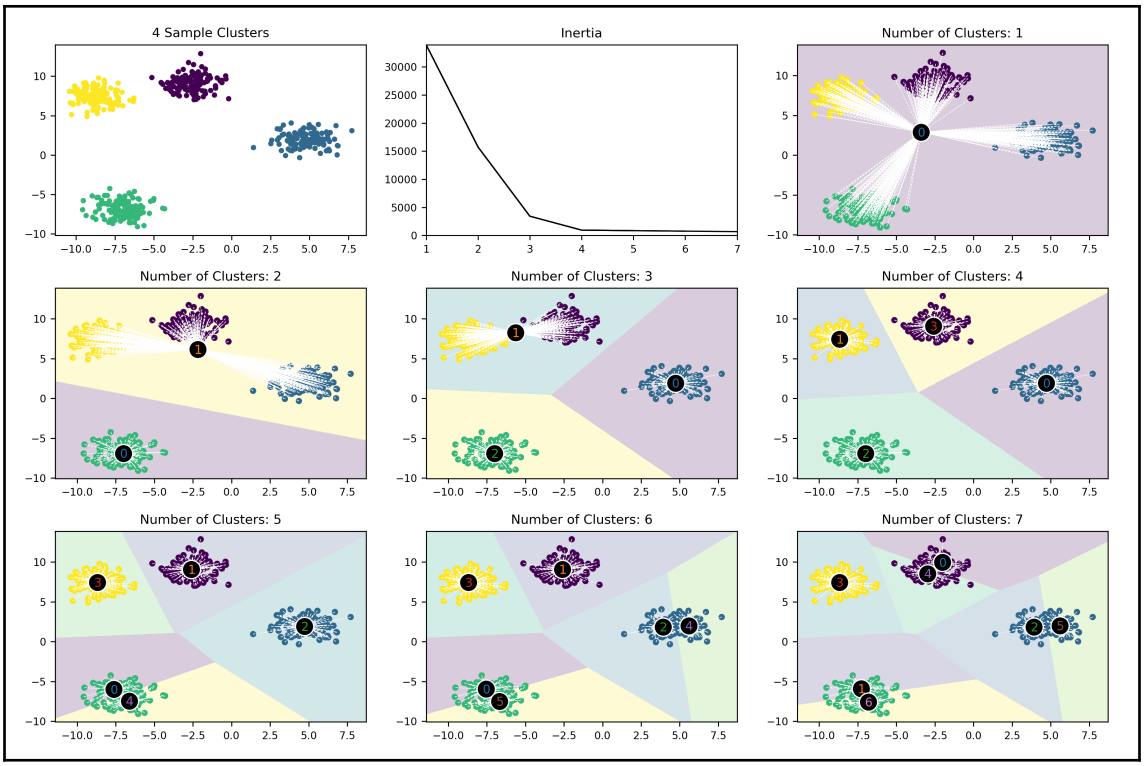

Portfolio 1
Portfolio 2
Portfolio 3
Portfolio 4
AAPL AMZN BA BAC CSCO CVX GOOGL HD INTC JNJ JPM KO MA MSFT PFE PG T UNH V VZ WFC WMT XOM
ticker

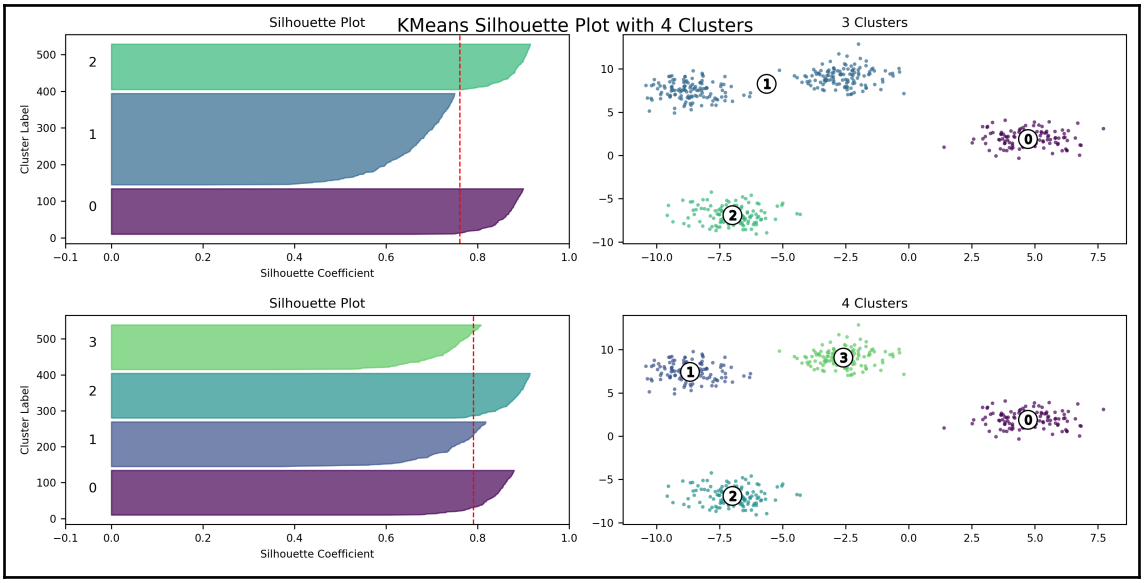

The Market
Portfolio 1
Portfolio 2
Portfolio 3
2010 2011 2012 2013 2014 2015 2016 2017 2018
date

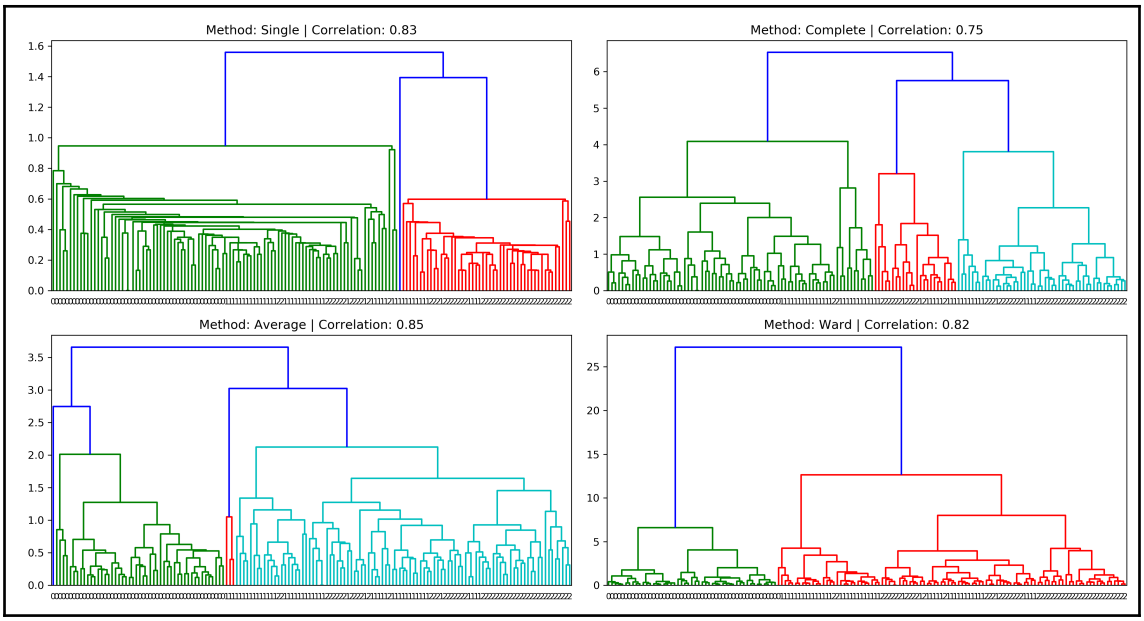

Fashion Images

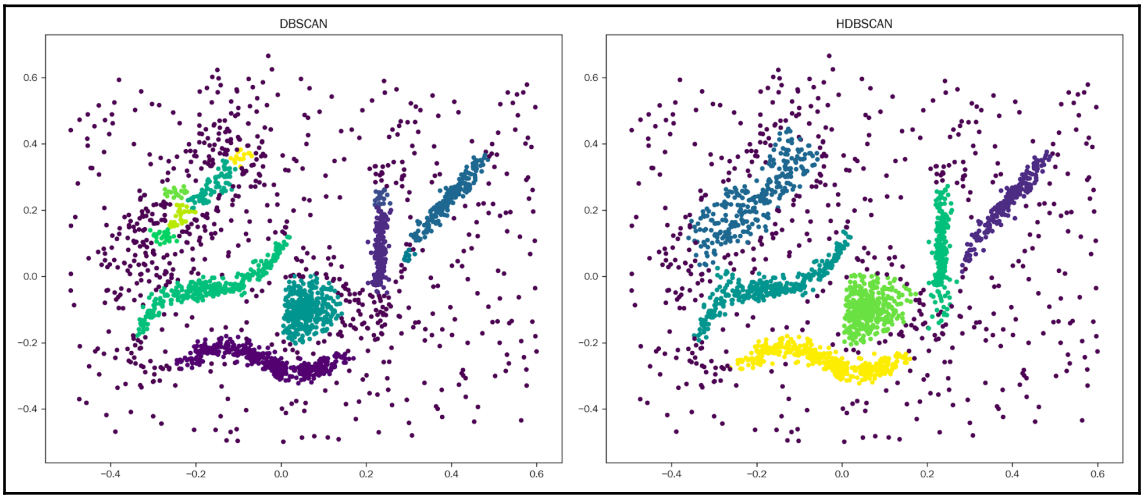

TSNE | Perplexity: 10
TSNE | Perplexity: 20
TSNE | Perplexity: 30
TSNE | Perplexity: 50
UMAP | Neighbors: 5
UMAP | Neighbors: 15
UMAP | Neighbors: 25
UMAP | Neighbors: 35

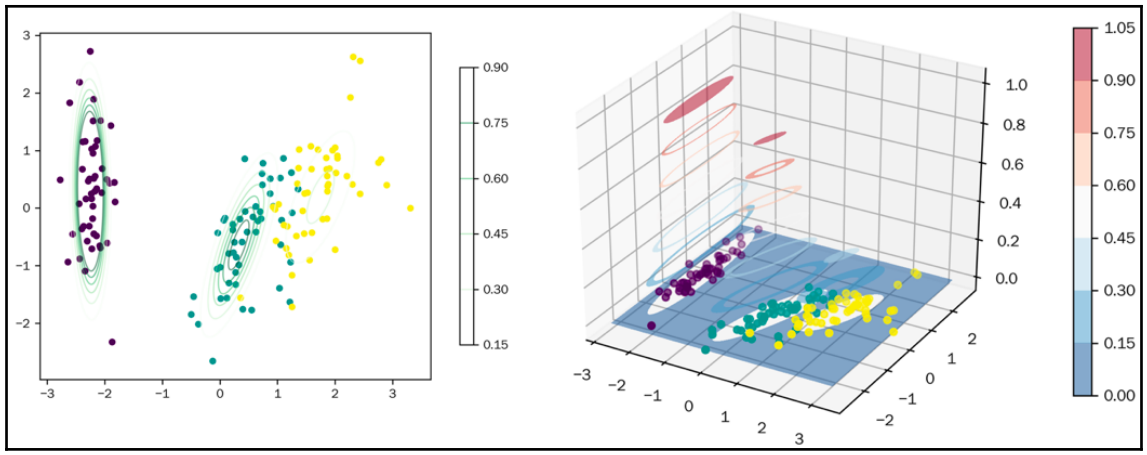

Iteration: 4 | Optimize Centers | Inertia: 910.80
Iteration: 3 | Optimize Centroids | Distance: 1,273.5

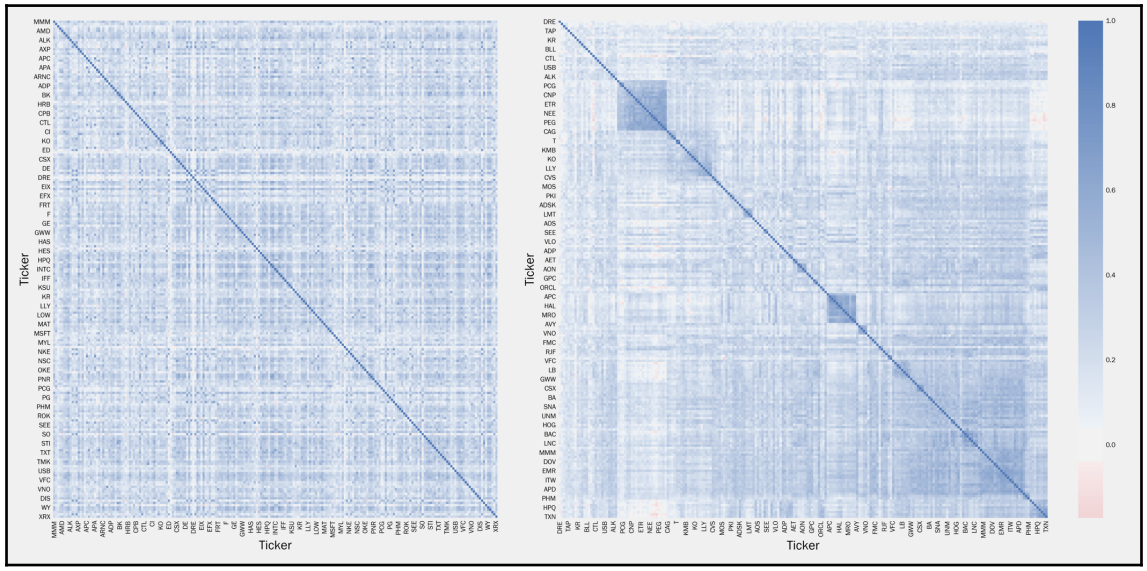

4 Sample Clusters
Inertia
Number of Clusters: 1
Number of Clusters: 2
Number of Clusters: 3
Number of Clusters: 4
Number of Clusters: 5
Number of Clusters: 6
Number of Clusters: 7


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

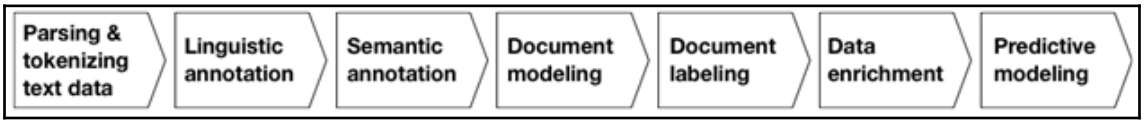

Method: Single | Correlation: 0.83
Method: Complete | Correlation: 0.75
Method: Average | Correlation: 0.85
Method: Ward | Correlation: 0.82

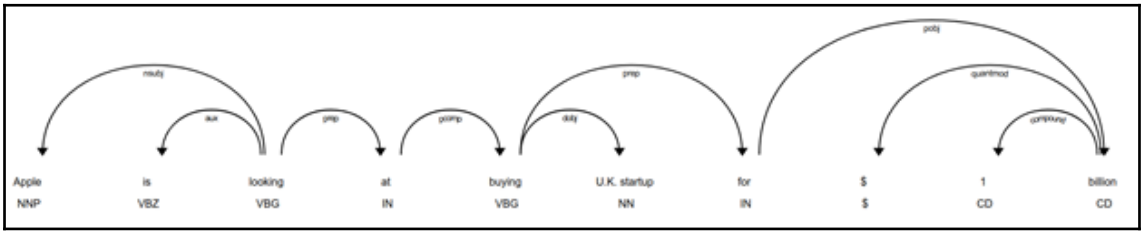

DBSCAN
HDBSCAN

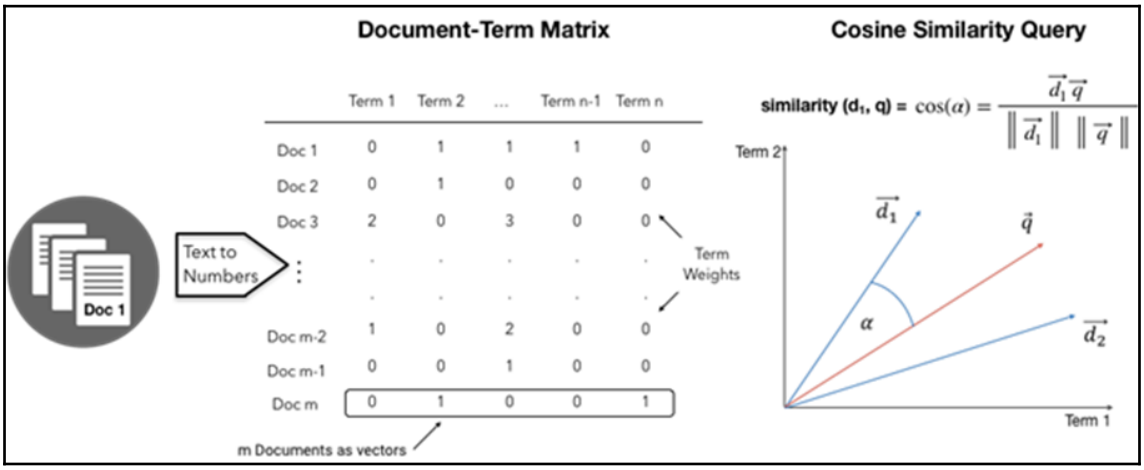


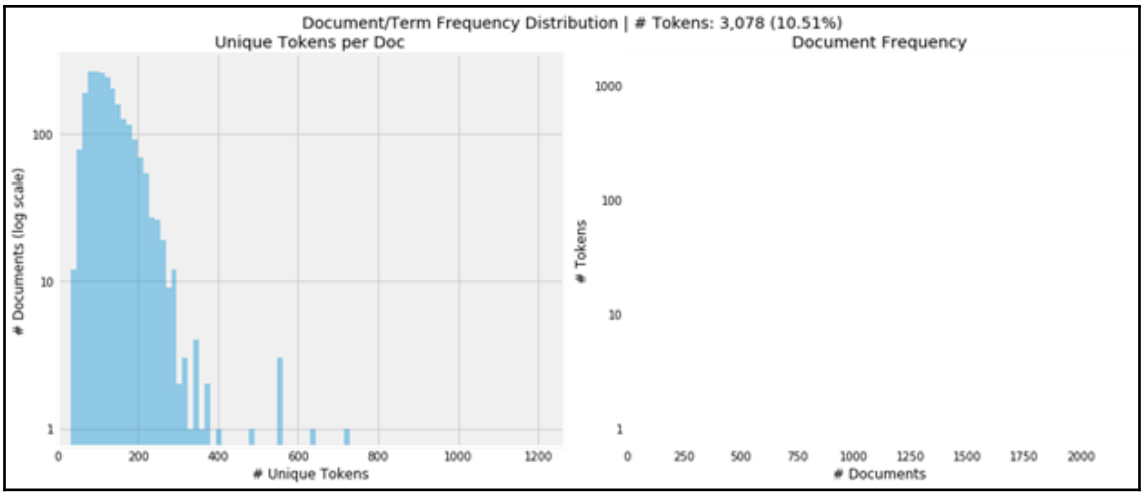

Ticker
Ticker
Ticker
Ticker

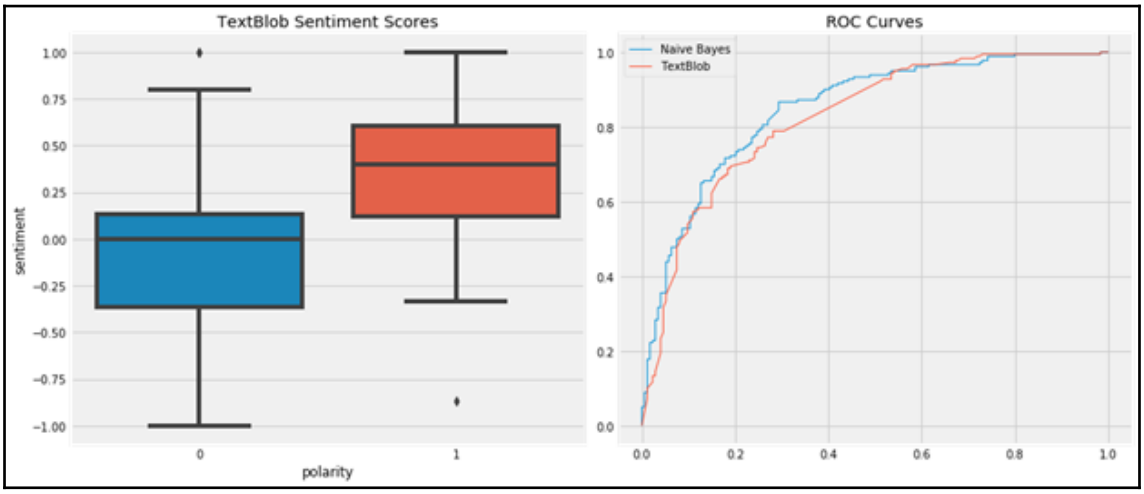
Chapter 13: Working with Text Data

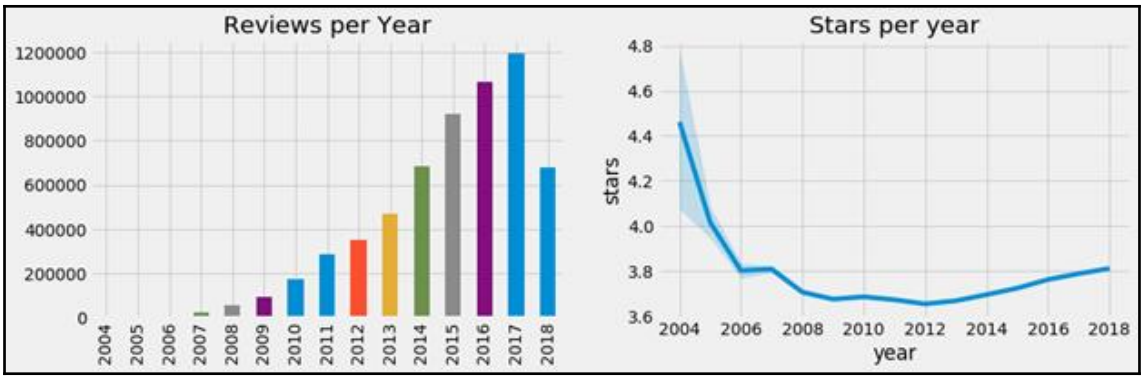




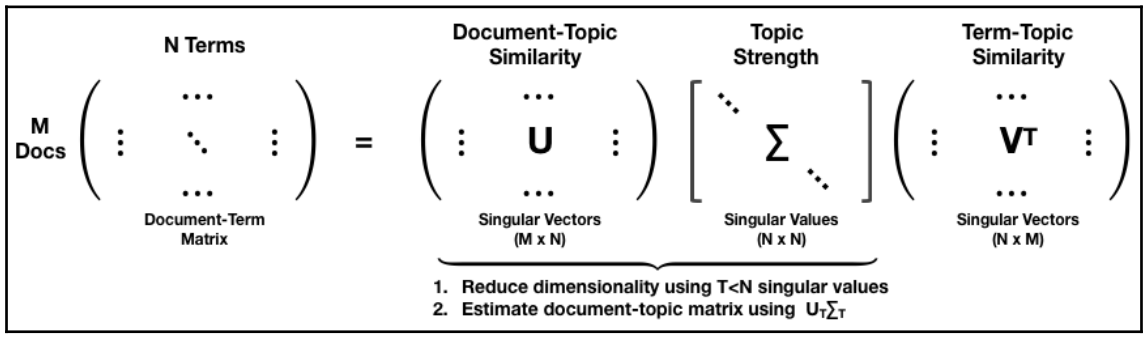


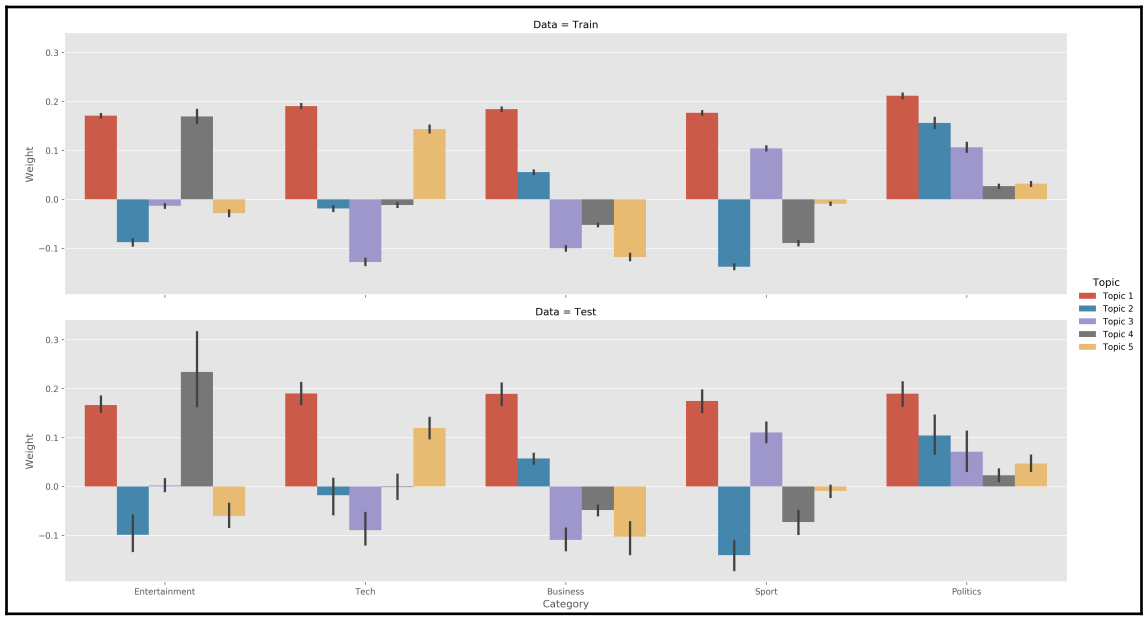

Document/Term Frequency Distribution | # Tokens: 3,078 (10.51%)
Unique Tokens per Doc
Documents (log scale)
100
10
1
0
200
400
600
800
1000
1200
Unique Tokens
Document Frequency
Tokens
1000
100
10
1
0
250
500
750
1000
1250
1500
1750
2000
Documents

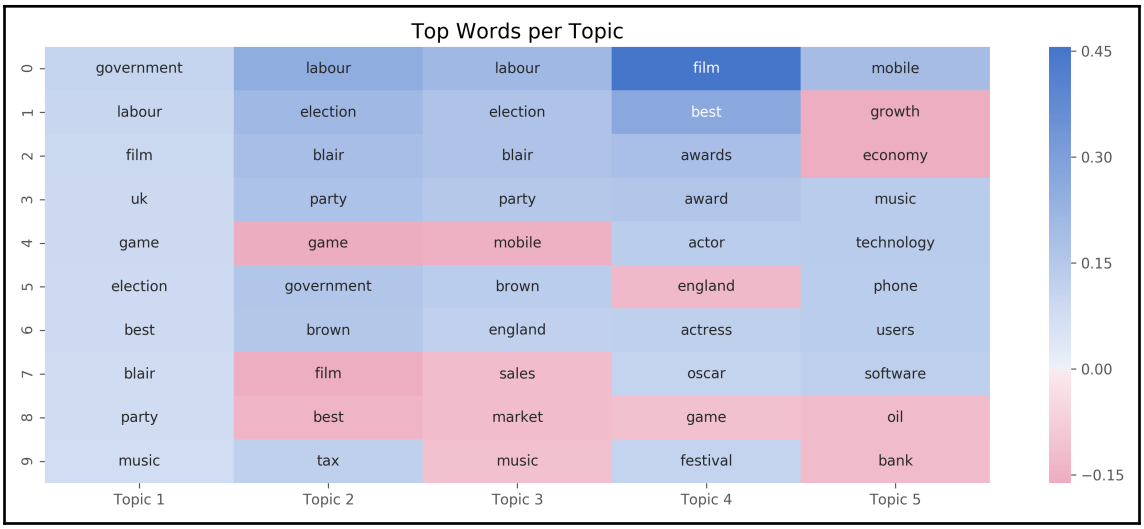

TextBlob Sentiment Scores
sentiment
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
1
polarity
ROC Curves
Naive Bayes
TextBlob
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

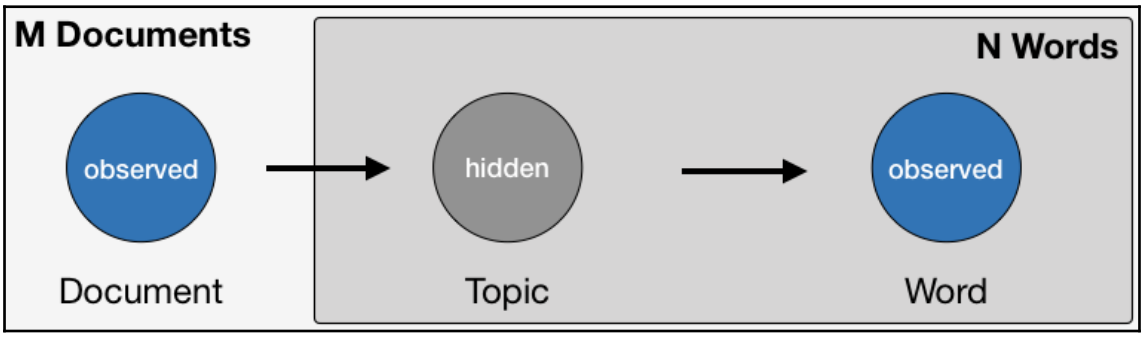

Reviews per Year
1200000
1000000
800000
600000
400000
200000
0
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Stars per year
stars
4.8
4.6
4.4
4.2
4.0
3.8
3.6
2004
2006
2008
2010
2012
2014
2016
2018
year

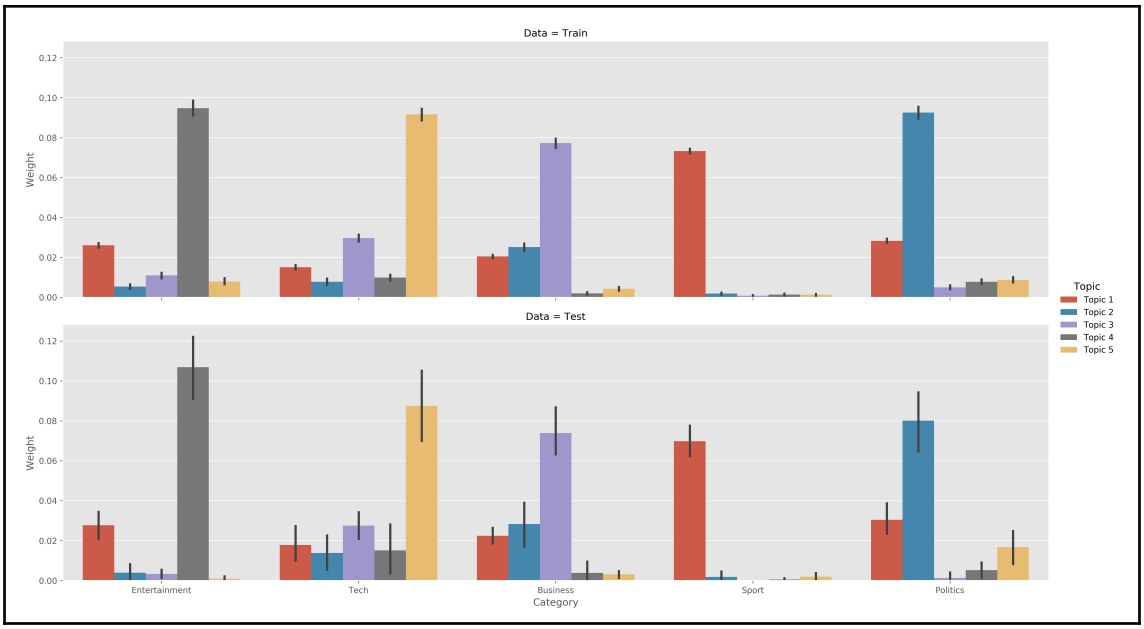
Chapter 14: Topic Modeling

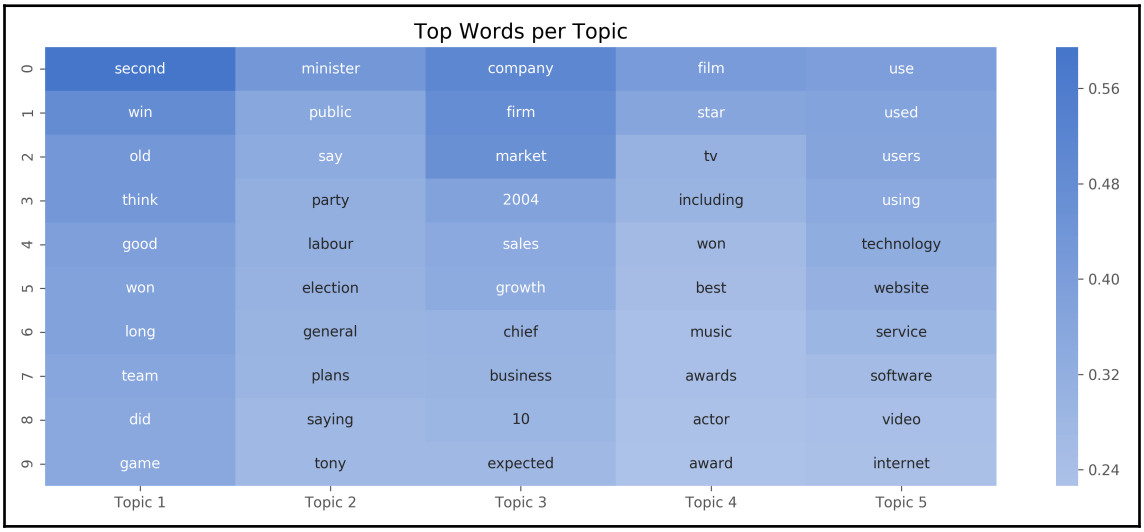


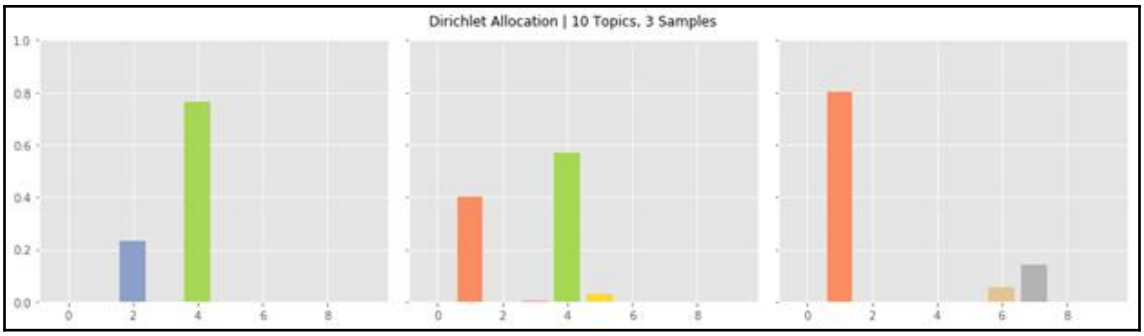



Top Words per Topic
0 government labour labour film mobile
1 labour election election best growth
2 film blair blair awards economy
3 uk party party award music
4 game game mobile actor technology
5 election government brown england phone
6 best brown england actress users
7 blair film sales oscar software
8 party best market game oil
9 music tax music festival bank
Topic 1 Topic 2 Topic 3 Topic 4 Topic 5
0.45
0.30
0.15
0.00
−0.15

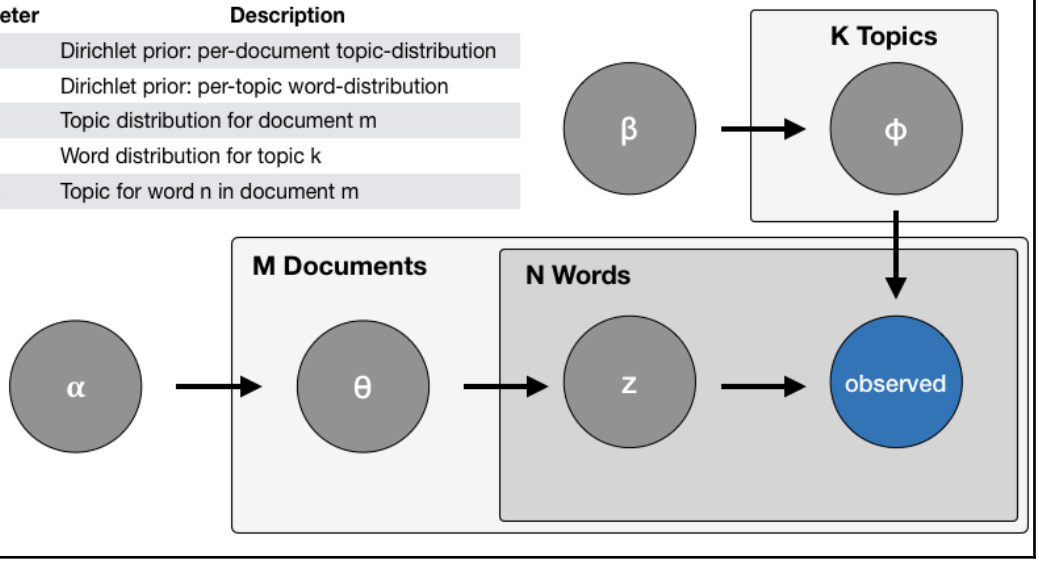

M Documents
N Words
observed
hidden
observed
Document
Topic
Word

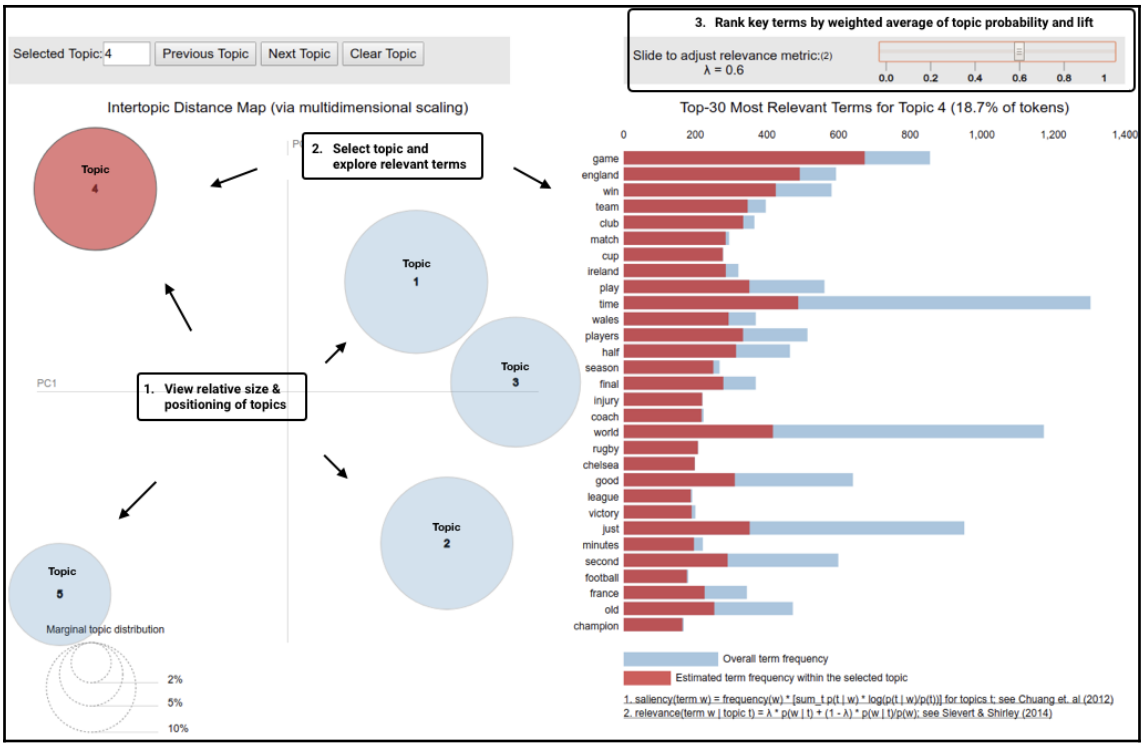

Data = Train
Data = Test
Weight
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Entertainment
Tech
Business
Sport
Politics
Category
Topic
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5

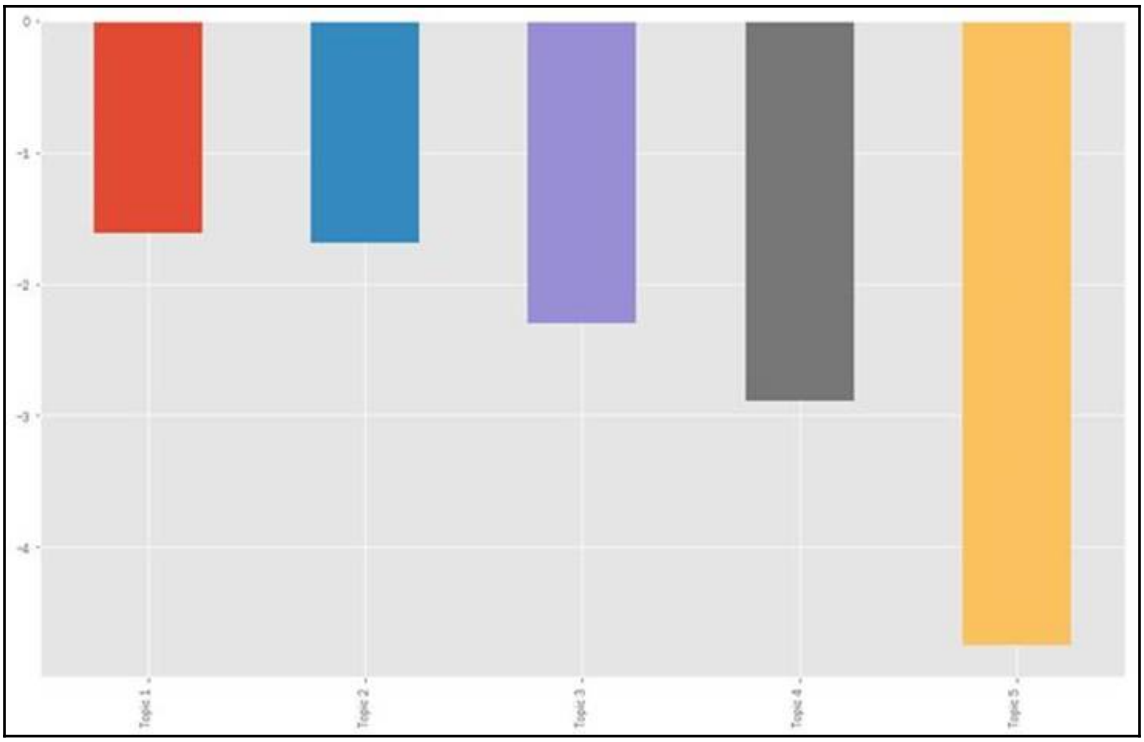

Top Words per Topic
second minister company film use
win public firm star used
old say market tv users
think party 2004 including using
good labour sales won technology
won election growth best website
long general chief music service
team plans business awards software
did saying 10 actor video
game tony expected award internet
0 1 2 3 4 5 6 7 8 9
Topic 1
Topic 2
Topic 3
Topic 4
Topic 5
0.56
0.48
0.40
0.32
0.24



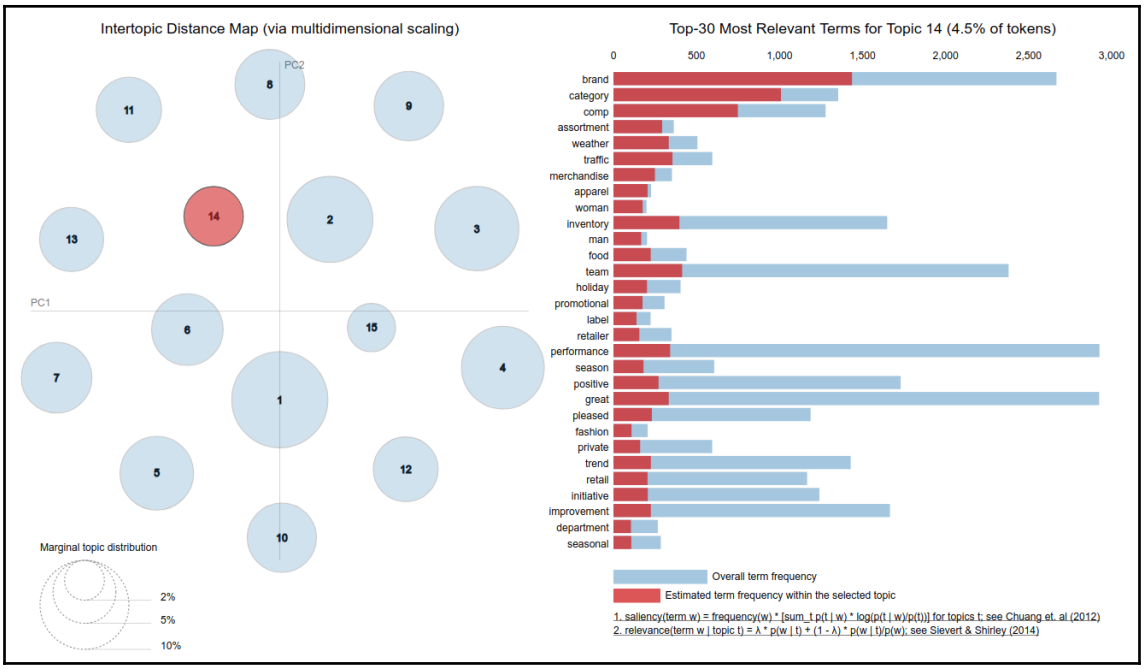
Parameter	Description
α	Dirichlet prior: per-document topic-distribution
β	Dirichlet prior: per-topic word-distribution
θ_m	Topic distribution for document m
ϕ_k	Word distribution for topic k
z_{mn}	Topic for word n in document m

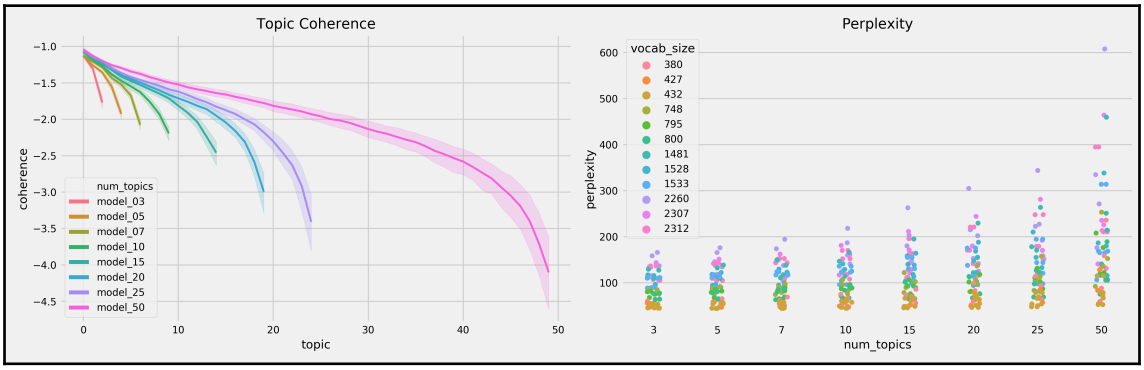


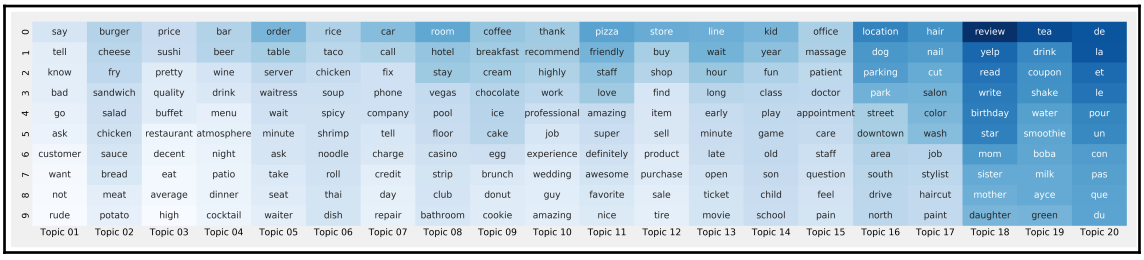

Selected Topic: 4
Previous Topic
Next Topic
Clear Topic
3. Rank key terms by weighted average of topic probability and lift
Slide to adjust relevance metric:(2)
λ = 0.6
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
2. Select topic and explore relevant terms
1. View relative size & positioning of topics
PC1
Topic 4
Topic 1
Topic 3
Topic 2
Topic 5
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 4 (18.7% of tokens)
0 200 400 600 800 1,000 1,200 1,400
game
england
win
team
club
match
cup
ireland
play
time
wales
players
half
season
final
injury
coach
world
rugby
chelsea
good
league
victory
just
minutes
second
football
france
old
champion
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)



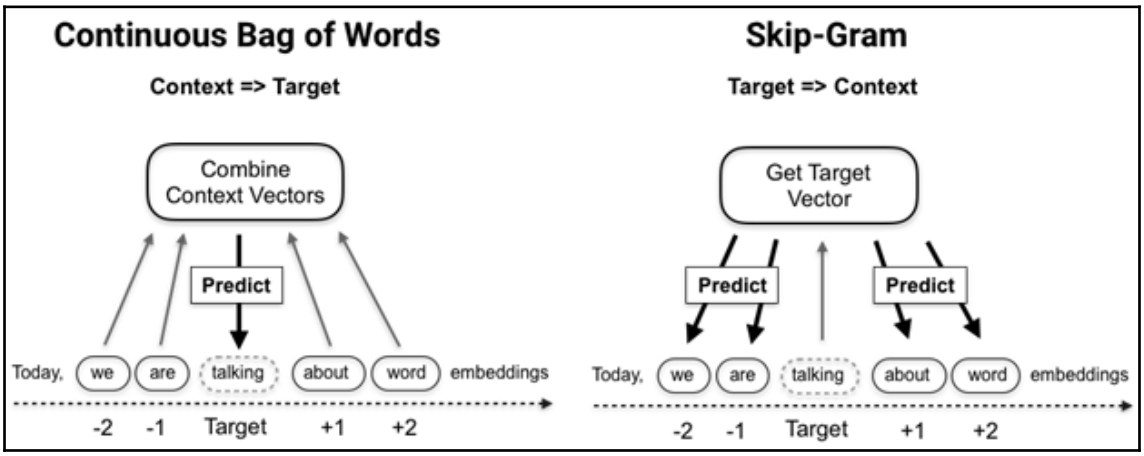
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	statement	expense	cloud	basis	patient	channel	focus	brand	want	project	lot	yes	price	maybe	bit
1	financial	total	technology	adjust	study	brand	acquisition	category	right	capital	thing	kind	production	kind	little
2	release	period	service	guidance	program	launch	investment	comp	price	investment	right	little	demand	okay	china
3	risk	income	solution	ebitda	clinical	experience	improve	team	thing	asset	mean	bit	volume	guess	loan
4	gaap	loss	platform	tax	trial	marketing	deliver	inventory	need	debt	actually	half	low	guy	service
5	officer	approximately	datum	low	phase	online	strategy	traffic	contract	portfolio	yes	pretty	capacity	sort	bank
6	chief	non	large	billion	datum	platform	invest	performance	lot	value	people	guidance	fleet	want	credit
7	conference	month	team	earning	development	digital	value	weather	say	balance	way	say	order	just	mention
8	measure	gaap	industry	gross	fda	consumer	progress	great	sure	return	different	thing	vessel	follow	tier
9	information	decrease	provide	approximately	cancer	user	performance	assortment	great	flow	obviously	low	supply	wonder	card

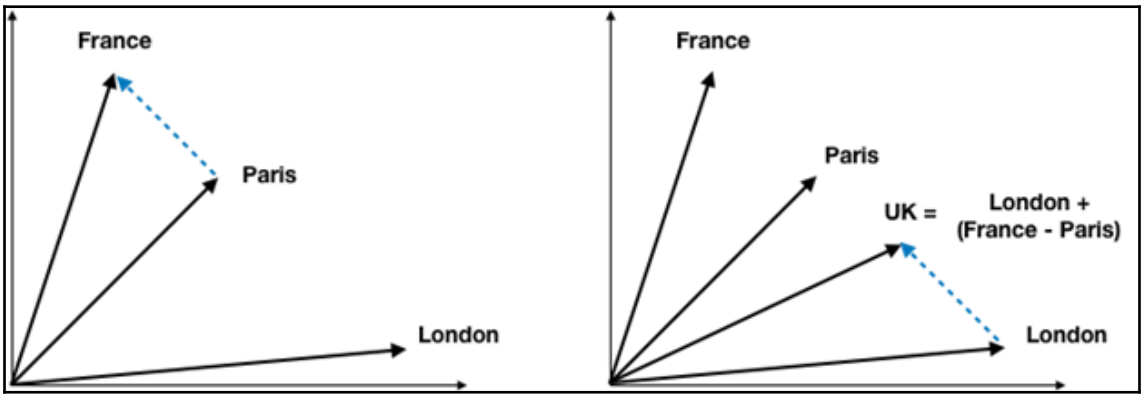

Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 14 (4.5% of tokens)
0
500
1,000
1,500
2,000
2,500
3,000
brand
category
comp
assortment
weather
traffic
merchandise
apparel
woman
inventory
man
food
team
holiday
promotional
label
retailer
performance
season
positive
great
pleased
fashion
private
trend
retail
initiative
improvement
department
seasonal
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)

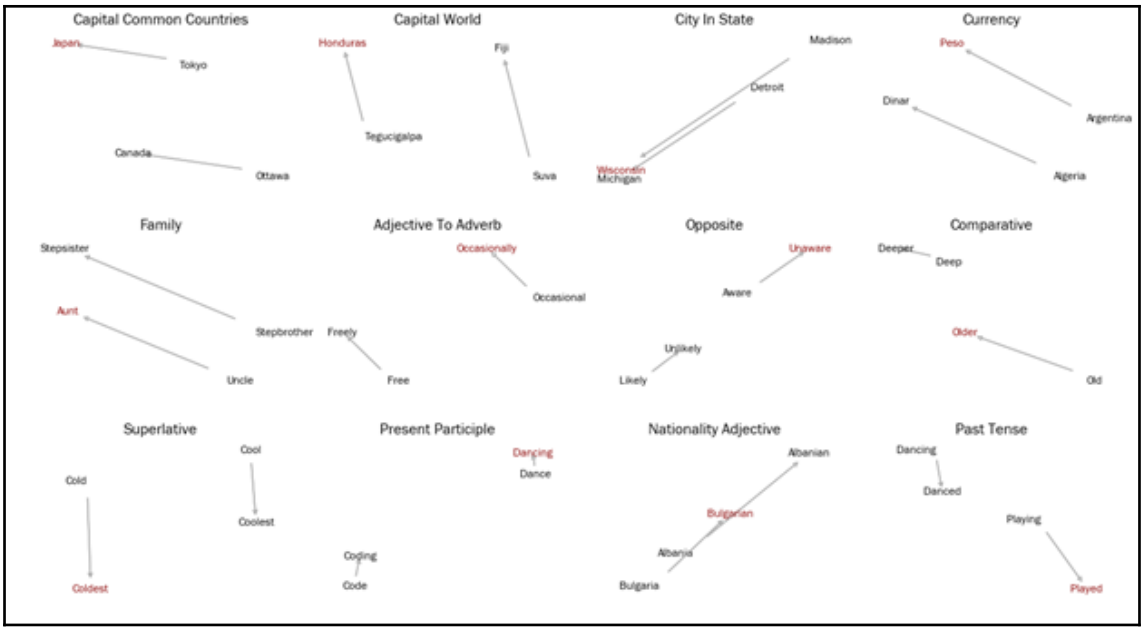

Topic Coherence
coherence
topic
num_topics
model_03
model_05
model_07
model_10
model_15
model_20
model_25
model_50
Perplexity
perplexity
num_topics
vocab_size
380
427
432
748
795
800
1481
1528
1533
2260
2307
2312

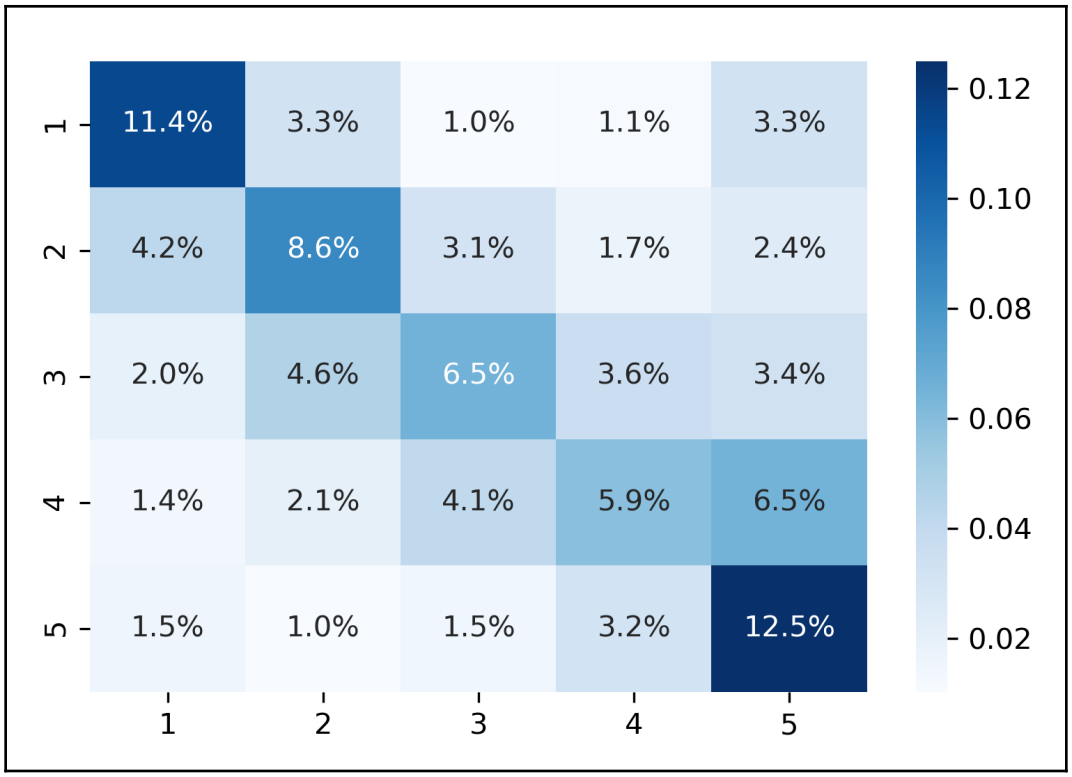

0 say burger price bar order rice car room coffee thank pizza store line kid office location hair review tea de
1 tell cheese sushi beer table taco call hotel breakfast recommend friendly buy wait year massage dog nail yelp drink la
2 know fry pretty wine server chicken fix stay cream highly staff shop hour fun patient parking cut read coupon et
3 bad sandwich quality drink waitress soup phone vegas chocolate work love find long class doctor park salon write shake le
4 go salad buffet menu wait spicy company pool ice professional amazing item early play appointment street color birthday water pour
5 ask chicken restaurant atmosphere minute shrimp tell floor cake job super sell minute game care downtown wash star smoothie un
6 customer sauce decent night ask noodle charge casino egg experience definitely product late old staff area job mom boba con
7 want bread eat patio take roll credit strip brunch wedding awesome purchase open son question south stylist sister milk pas
8 not meat average dinner seat thai day club donut guy favorite sale ticket child feel drive haircut mother ayce que
9 rude potato high cocktail waiter dish repair bathroom cookie amazing nice tire movie school pain north paint daughter green du
Topic 01 Topic 02 Topic 03 Topic 04 Topic 05 Topic 06 Topic 07 Topic 08 Topic 09 Topic 10 Topic 11 Topic 12 Topic 13 Topic 14 Topic 15 Topic 16 Topic 17 Topic 18 Topic 19 Topic 20

Chapter 15: Word Embeddings






Capital Common Countries
Japan
Tokyo
Canada
Ottawa
Capital World
Honduras
Tegucigalpa
Fiji
Suva
City In State
Madison
Detroit
Wisconsin
Michigan
Currency
Peso
Argentina
Dinar
Algeria
Family
Stepsister
Stepbrother
Aunt
Uncle
Adjective To Adverb
Occasionally
Occasional
Freely
Free
Opposite
Unaware
Aware
Unlikely
Likely
Comparative
Deeper
Deep
Older
Old
Superlative
Cool
Coolest
Cold
Coldest
Present Participle
Dancing
Dance
Coding
Code
Nationality Adjective
Albanian
Bulgarian
Albania
Bulgaria
Past Tense
Dancing
Danced
Playing
Played


1
2
3
4
5
11.4%
3.3%
1.0%
1.1%
3.3%
4.2%
8.6%
3.1%
1.7%
2.4%
2.0%
4.6%
6.5%
3.6%
3.4%
1.4%
2.1%
4.1%
5.9%
6.5%
1.5%
1.0%
1.5%
3.2%
12.5%
1
2
3
4
5
0.12
0.10
0.08
0.06
0.04
0.02