Capstone Three Presentation
"Predicting Stock Returns"

Thanh Nguyen

1

Introduction

- Being able to predict stock returns means better trading decisions
- Goal: at the end of trading today to predict the return tomorrow
- Tools: traditional time series models (ARIMA), machine learning models
- Stocks: Zions bank (ZION, 4.4 billion market capitalization) and Fifth Third bank (FITB, 17.8 billion market capitalization)

2

Quick summary and agenda

- The process
  - Models: ARIMA, RandomForest, Gradient Boosting, Ridge, KNN
  - Cross-validation with TimeSeriesSplit()
- Outcomes
  - Achieved marginal improvement compared to naïve model
  - KNN models show potential to predict direction of price movement
- Agenda
  - Data source
  - Exploratory data analysis (EDA)
    - Target, features, correlation analysis
  - Modeling and performance on unseen data
  - Conclusion
  - Future work

3

Data source

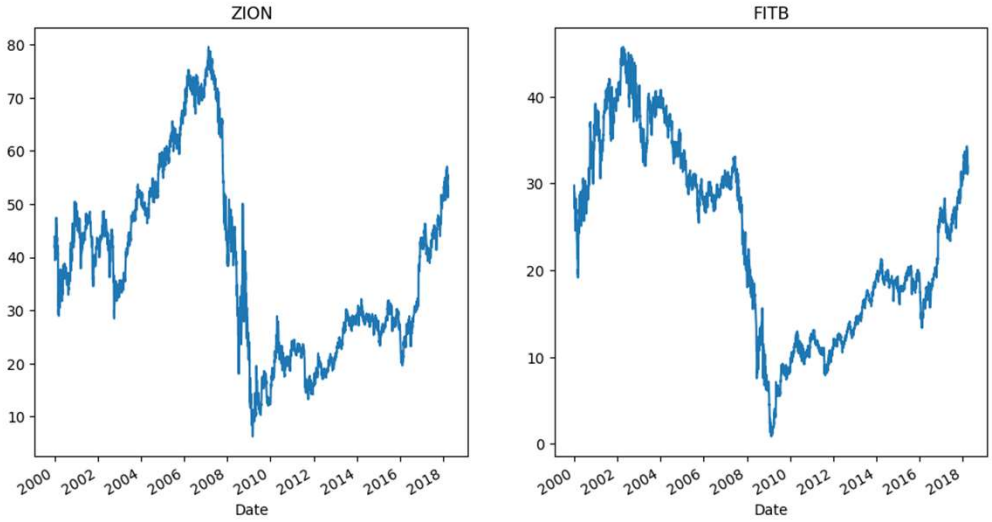
- Nasdaq Data Link, WIKI Prices database
  - 3000 US stocks (prices, dividends, splits)
  - Coverage ends March 2018.
- High-quality data, no need for data wrangling

Date	Open	High	Low	Close	Volume	Ex-Dividend	Split Ratio	Adj. Open	Adj. High	Adj. Low	Adj. Close	Adj. Volume
2000-01-03	59.03	59.12	53.44	55.50	1199600.0	0.0	1.0	46.614284	46.685355	42.200023	43.826745	1199600.0
2000-01-04	54.63	55.00	52.50	52.81	816100.0	0.0	1.0	43.139731	43.431910	41.457732	41.702530	816100.0
2000-01-05	52.75	53.25	51.06	53.06	1124700.0	0.0	1.0	41.655150	42.049985	40.320606	41.899948	1124700.0
2000-01-06	52.75	54.94	52.38	53.50	1112100.0	0.0	1.0	41.655150	43.384529	41.362971	42.247403	1112100.0
2000-01-07	53.75	54.25	53.31	53.63	782000.0	0.0	1.0	42.444821	42.839656	42.097366	42.350060	782000.0

4

Data source

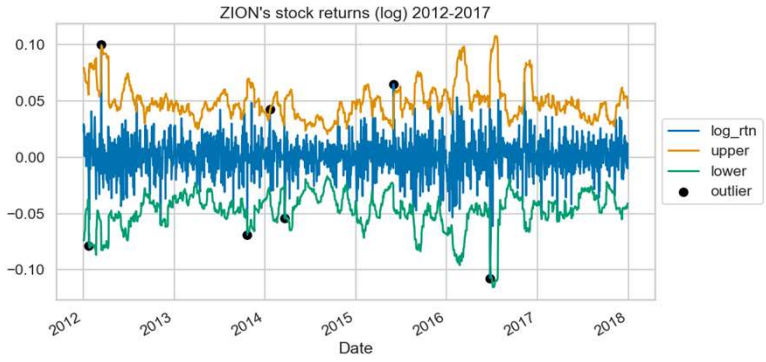
ZION and FITB Prices



5

EDA: Target (log returns)

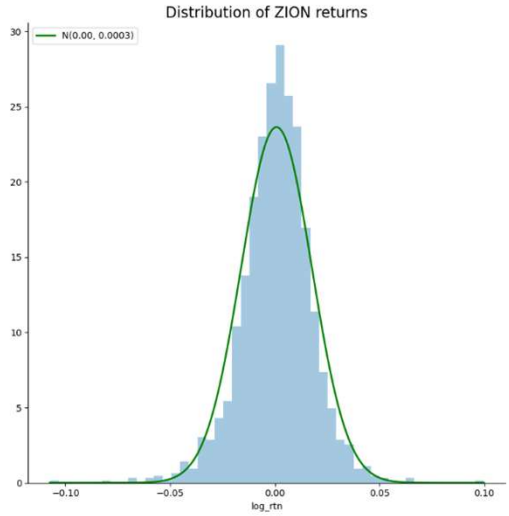
$$log\_rtn_t = log(\frac{P_t}{P_{t-1}})$$

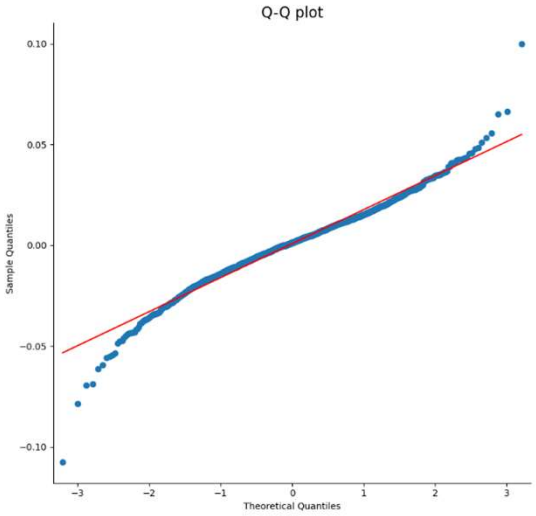


- Series is stationary
  - Augmented Dickey-Fuller test (ADF), Kwiatkowski-Phillips-Schmidt-Shin test (KPSS).

6

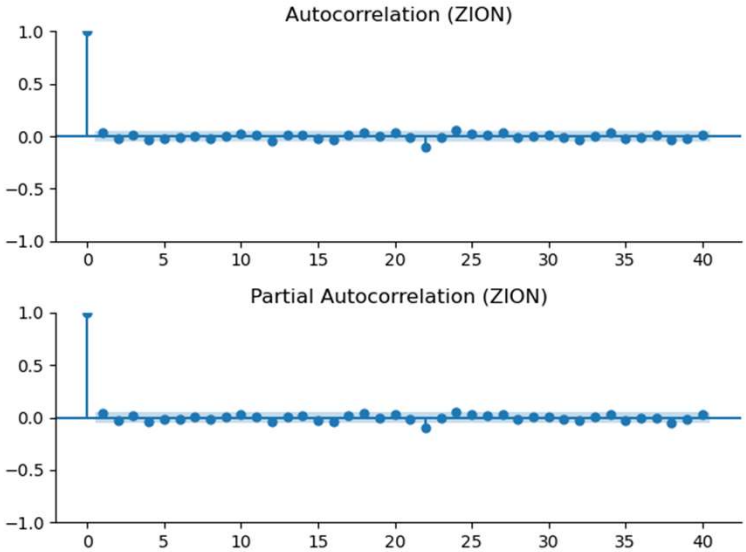
EDA: Target (log returns)





7

EDA: Target (log returns)



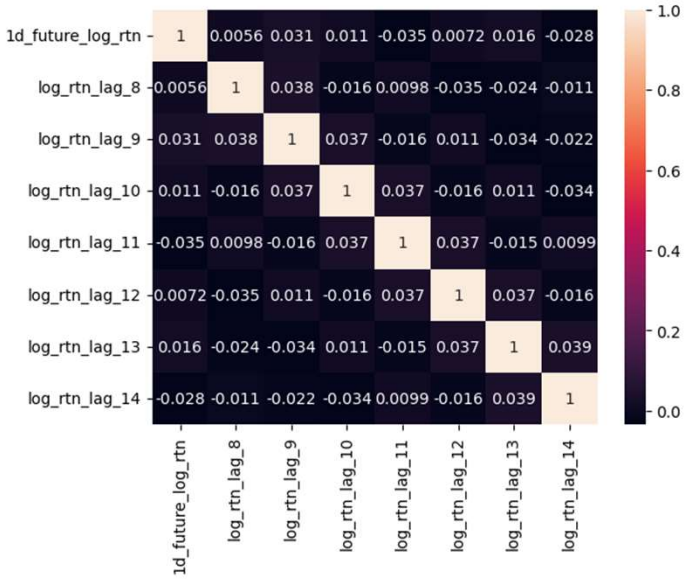
8

EDA: Features

- Features are for machine learning models (4 groups of features)
- Lag0 to lag14 of returns
- Features based on trading volume
  - Today's one-day change in trading volume
  - 10-day moving average of trading volume change
- Technical analysis indicators
  - Moving average (14-day, 50-day, and 200-day)
  - Relative strength index (14-day, 50-day, and 200-day)
- Features based on time
  - Month of the year (sine and cosine transformation)

9

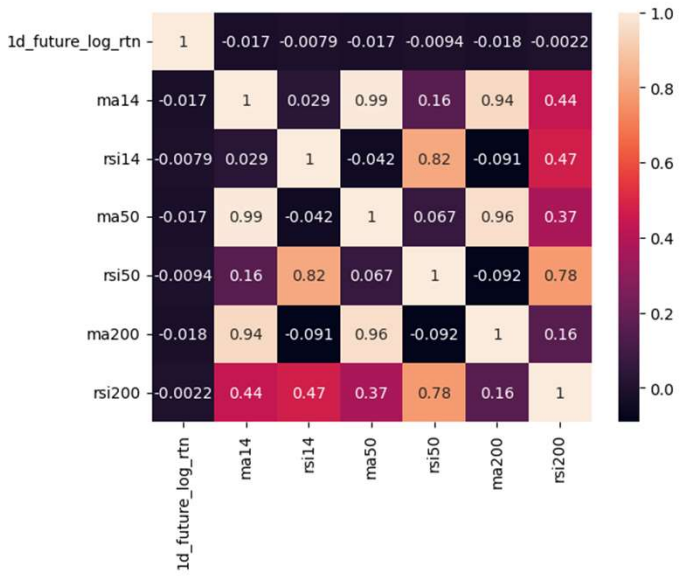
EDA: Correlation analysis



10

EDA: Correlation analysis

- High correlations among moving average features
- High correlations among relative strength index features
- For modeling, use only ma50 and rsi50



11

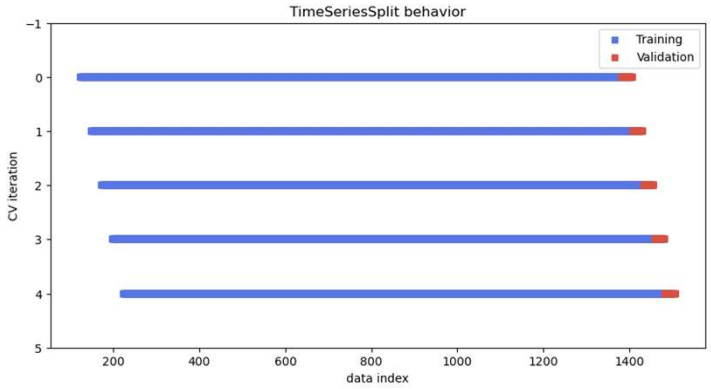
Modeling

- Performance measure: Mean Squared Error (MSE)
  - Mean Absolute Percentage Error (MAPE) not preferred
- 3-year and 5-year training period for all models
- ARIMA (p, d, q)
  - p and q: from 0 to 4
  - d: 0 or 1
- Machine learning models: tune corresponding hyperparameters

12

Modeling (Machine Learning models)

- Data: daily returns 2012-2017
- Test set: first 25 trading days of 2018
- Cross validation: TimeSeriesSplit()
  - Train set: 1260 trading days (5 years) or 756 (3 years)
  - Validation set: 25 trading days



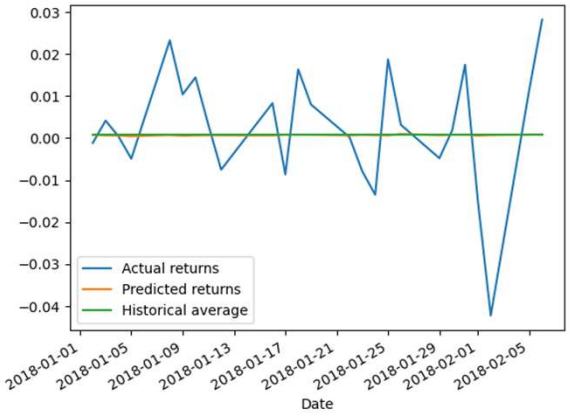
13

Hyperparameter Tuning for ZION Prediction

Model Name	Best Parameters		Cross-Validation Score (mse)
Random Forest	'max_depth': 1 'n_estimators': 500	'max_features': 0.1	0.000172
Gradient Boosting	'learning_rate': 0.005 'max_features': 0.1 'n_estimators': 50	'max_depth': 6 'min_samples_split': 8 'subsample': 0.8	0.000171
Ridge	alpha: 7250		0.000172
KNN	n_neighbors: 14		0.000185
KNN (include month sine and cosine)	n_neighbors: 16		0.000177
ARIMA	(4, 1, 3)		0.000172

14

Performance on unseen data (ZION)



	Best model performance		Naïve model performance	
	MSE	R-square	MSE	R-square
Predicting ZION returns	0.0002087	-1.35%	0.0002092	-1.60%

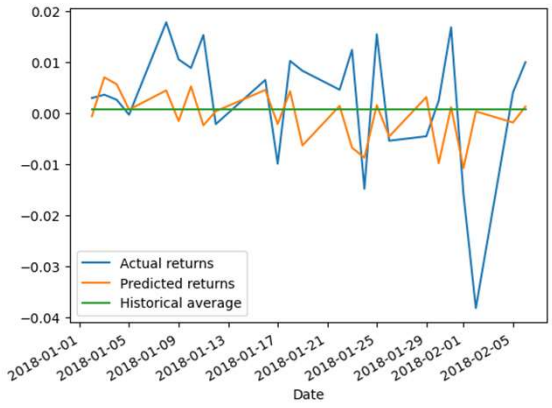
15

Hyperparameter Tuning for FITB Prediction

Model Name	Best Parameters		Cross-Validation Score (mse)
Random Forest	'max_depth': 20 'n_estimators': 100	'max_features': 0.1	0.000128
Gradient Boosting	'learning_rate': 0.08 'max_features': 0.1 'n_estimators': 50	'max_depth': 3 'min_samples_split': 6 'subsample': 1	0.000129
Ridge	alpha: 40000		0.000130
KNN	n_neighbors: 11		0.000132
KNN (include month sine and cosine)	n_neighbors: 6		0.000128
ARIMA	(1, 0, 2)		0.000131

16

Performance on unseen data (FITB)



	Best model performance		Naïve model performance	
	MSE	R-square	MSE	R-square
Predicting FITB returns	0.000147	1.28%	0.000152	-2.12%

17

Conclusion

- Achieved marginal improvement compared to naïve model
- KNN models show potential to predict direction of price movement
  - Can develop trading strategy based on this signal and carry out back testing

18

Future Work

- Longer training period (longer than 5 years)
- ARIMA with seasonality and exogenous features
- Overfitting of Random Forest
- Neural network models (RNN)

19

References

- Lewinson, E. 2022. *Python for Finance Cookbook*. <packt>

20