Predicting Stock Prices

Capstone Project. N.Kairakbayev

Project Goal

> Build, train and test models which could be used to predict public companies stock prices

Key Questions

01 How to group companies into specific clusters?

02 Which ML algorithm suitable for forecasting?

03 Is it possible to make qualitative long-term forecasts using the ML algorithms?

04 Can the quality of forecasting be improved by adding data from the financial statements to the input?

05 Can a model which trained on a cluster’s center time serie be used for forecasting on data from other time series in that cluster?

Dataset

All necessary and publicly available data were uploaded from free and open https://financialmodelingprep.com with API-service

Data types

- Daily close stock prices from 2016-01 to 2020-01
- Table with Company profile:
  - ticker
  - name
  - industry
  - sector
  - exchange
  - market capitalization (on uploaded date)
  - descriptive data (description, ceo, logo etc.)

Data volumes

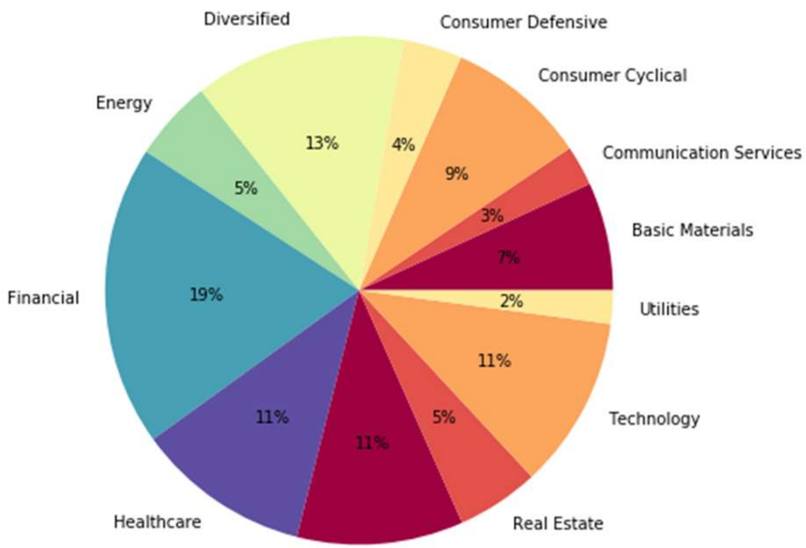
- Initially ~ 13 900 tickers loaded
- Dropped tickers:
  - dublicates
  - composite indexes
  - ETF indexes
  - Mutual Funds
  - Crypto Currencies
- Remained tickers: ~ 6 900

Quotes Timeline: first 80% - used for training, 20% - used for testing and building forecasts

2016 **2017** **2018** **2019** **01-2020**

Groups and Composites

2 CLUSTER SETS for FURTHER MODELS: BASED on SECTORS and BASED on CORRELATION of stock prices

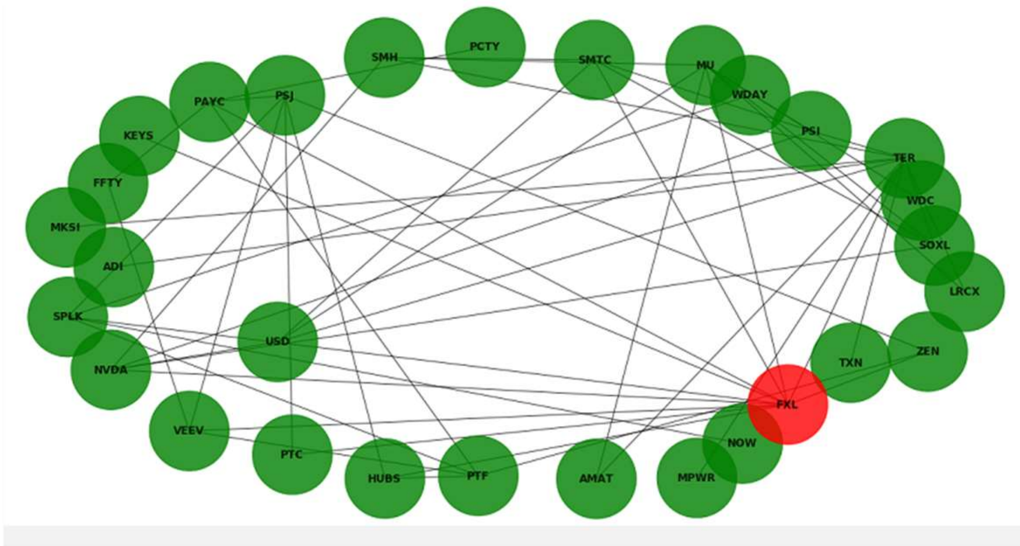


CLUSTERS based on SECTOR

- 12 clusters
- Covers 6 899 companies
- Average number of companies per cluster: 575
- Cluster Center: Median stock price (synthetic composite)

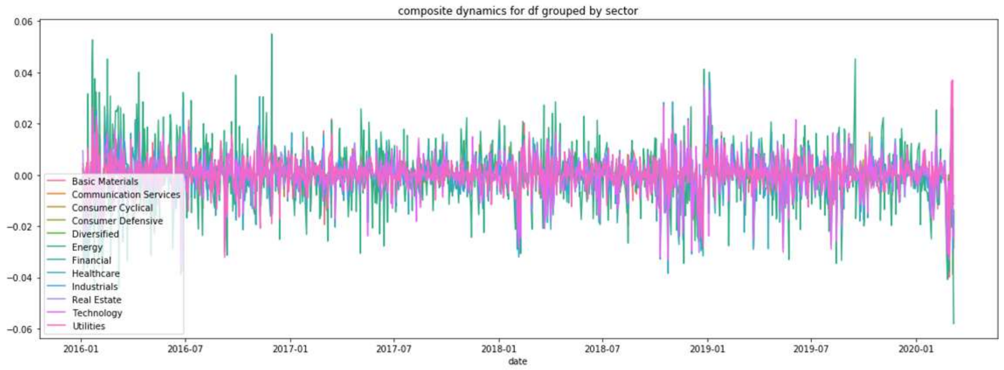
CLUSTERS based on CORRELATION between prices changes (Net Price Margin)

- 14 clusters (formed by 0.65 corr threshold)
- Covers 856 companies
- Average number of companies per cluster: 61
- Cluster Center: center of Graph

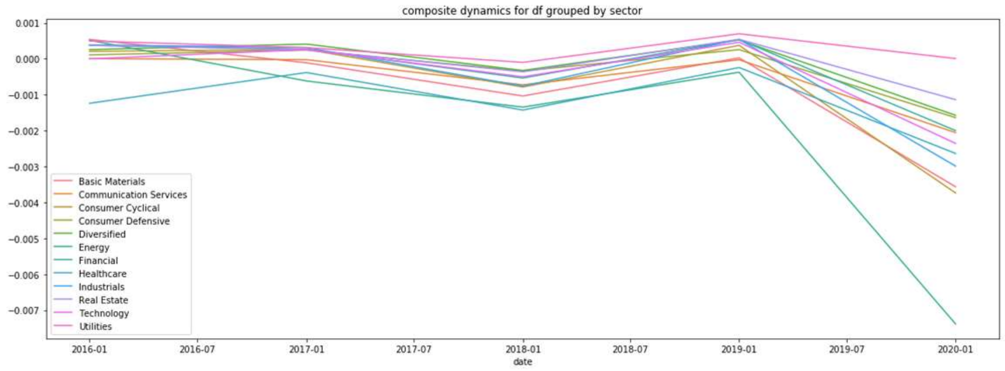


Sector Composite Dynamics. Return Rates p.d.

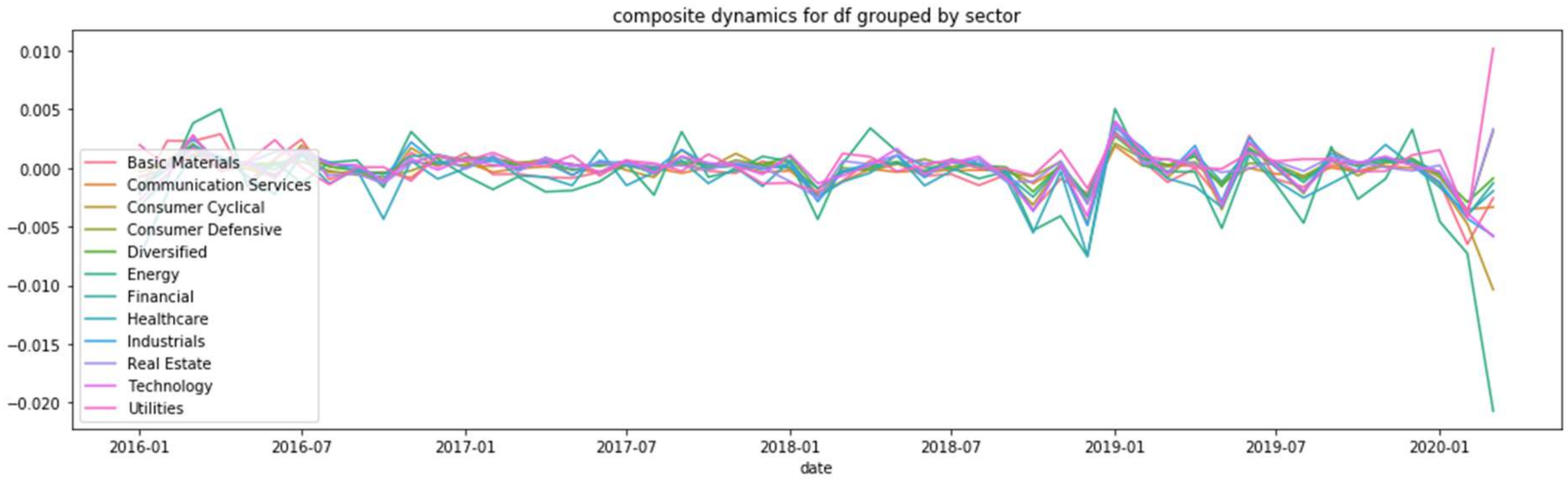
Daily Net Price Margins



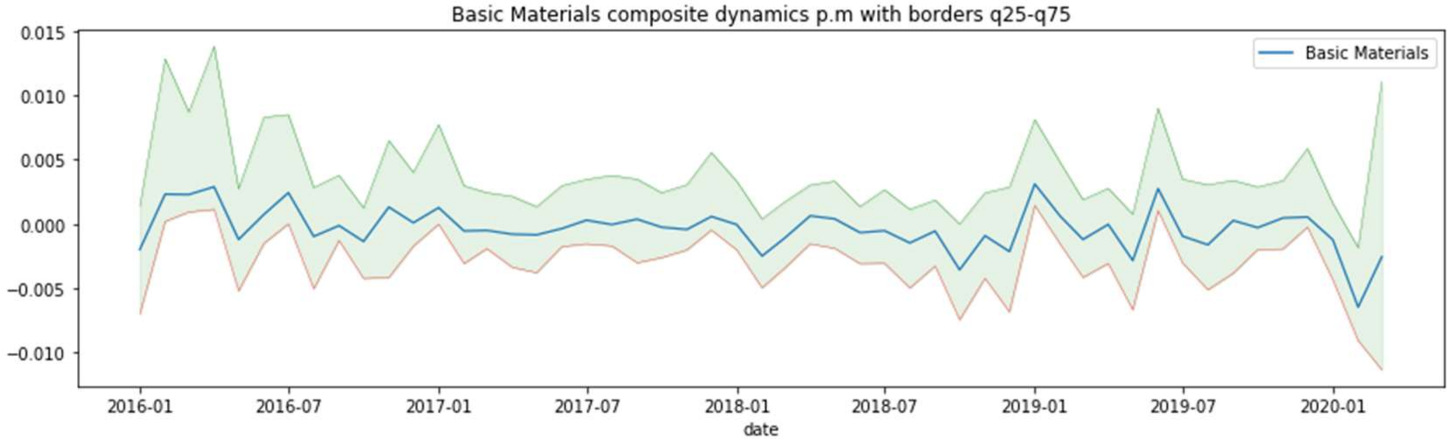
Daily Net Price Margins averaged on YEAR basis



Daily Net Price Margins averaged on MONTH basis



Clusters centers. Return Rates p.d.

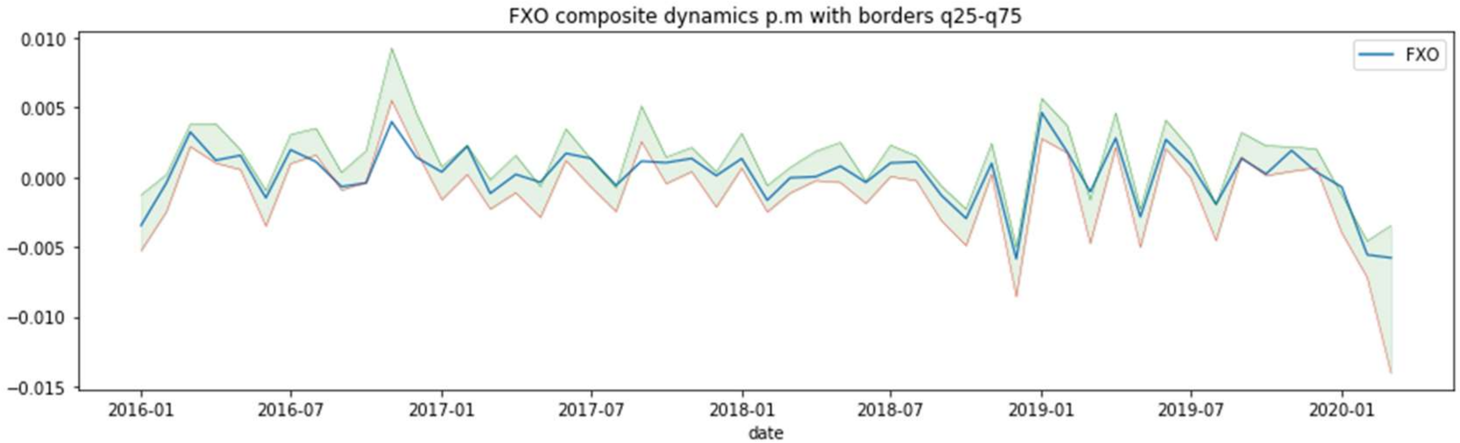


Sector Composite:

- Sector: Basic Materials
- Center: Median values

Corr-Cluster center:

- Center: FXO



Approached models

1. ARIMA

Seasonal autoregression models

- INPUT: only stock prices

2. Recurrent Neural Network

LSTM-layer based models

- INPUT: stock prices and date features

3. Recurrent Neural Network with financial statements

LSTM-layer based models

- stock prices, date features, data from FS

Project Stages

1. Data Loading
- GET-request from API
- Tickers filtering
- Merge to DataFrame

2. Data Cleaning
- Timeline select
- Missing values
- PCT-conversion

3. Clustering
- EDA
- Sector composites
- Kmeans tests
- Correlation clusters

4. ARIMA
- Norm Price vs NetPriceMargin
- Stat.tests (ACF, PACF, ADF)
- SARIMAX param search
- Fit, Predict, R2

5. RNN
- EDA on Dates
- Data Preparation
- Architecture select
- Fit, Predict, R2
- Test on cluster items

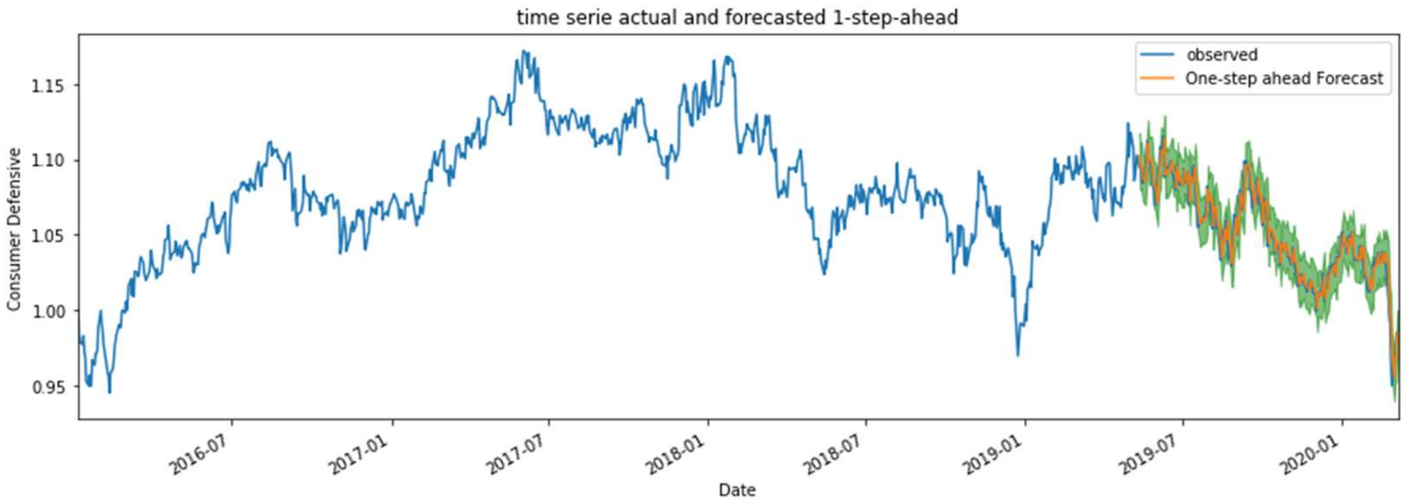
6. RNN with FS
- FS loading
- Architecture select
- Fit, Predict, R2
- Comparison

Forecast Results

	ARIMA Individually Optimized Params	RNN Unified 1 Architecture	RNN Improved Improved for 1 ticker	RNN with FS For 1 ticker
R2 range for Sector Clusters	**0.86 – 0.97**	**Negative – 0.88**	-	-
R2 range for Corr-Clusters	**0.92 – 0.95**	**Negative – 0.90**	-	-
R2 for selected TICKER (NBTB)	**0.94**	**0.81**	**0.91**	**0.61**
Long-Term R2 range	**Negative – 0.34**	**Negative – 0.03**	-	-
Avg positive R2 (median)	**0.93**	**0.77**	**0.91**	**0.61**

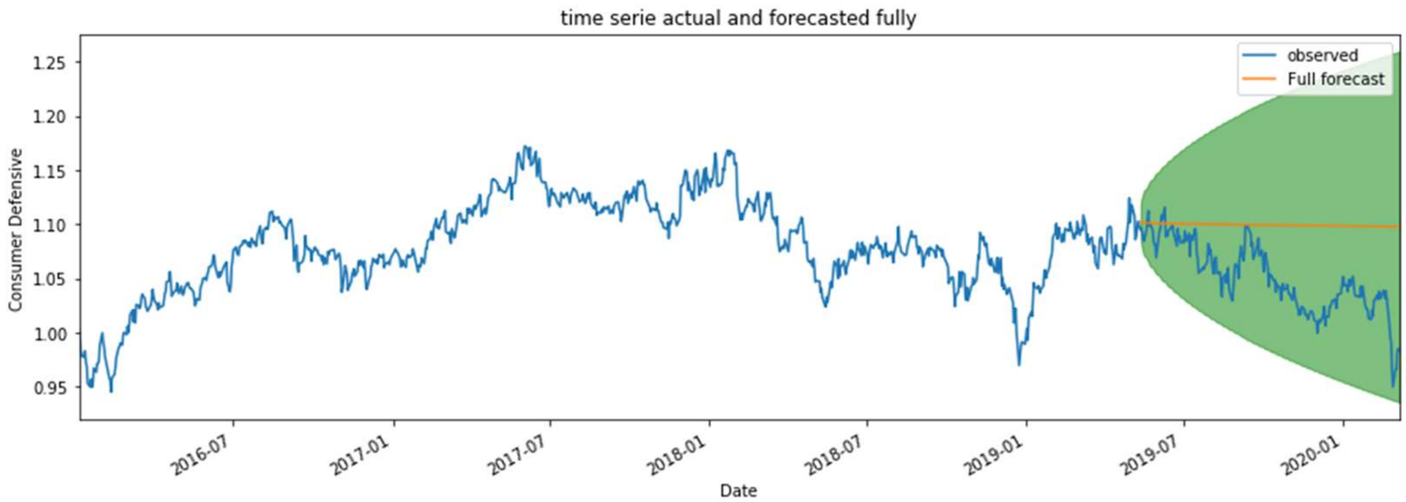
For all Time Series and applied algorithms FULL-CYCLE LONG-TERM forecasts: R2 - unsatisfactory

ARIMA forecast charts example

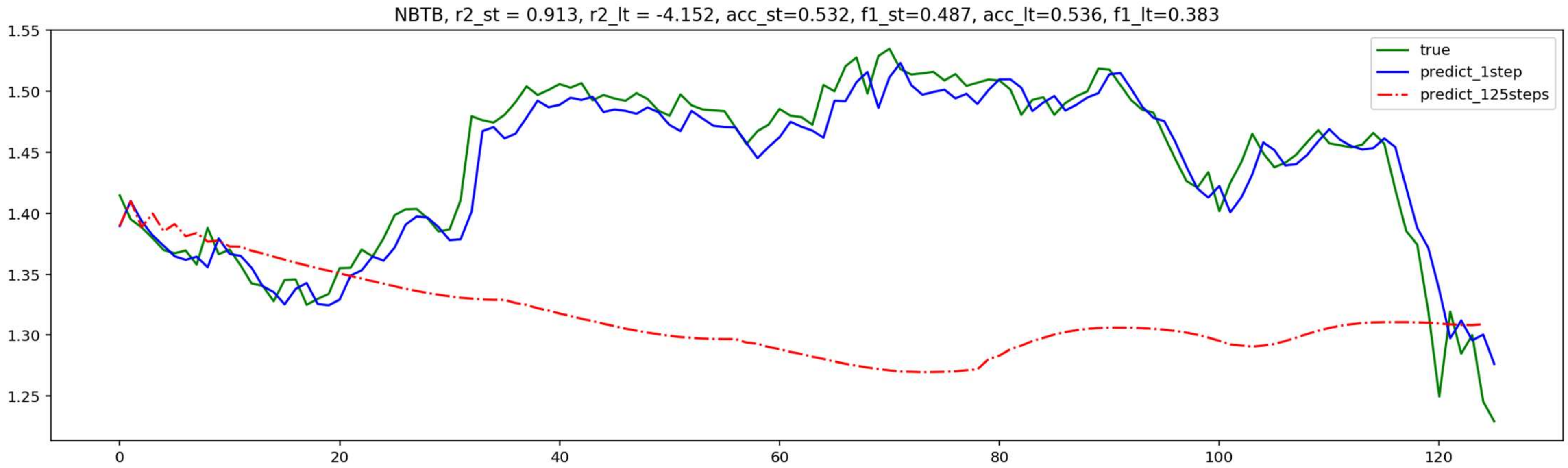


1-day-forward forecast shows robust forecast ability and good R2 metrics with using linear models

Using linear ARIMA models for LT-forecast on time series with high volatility – not good idea

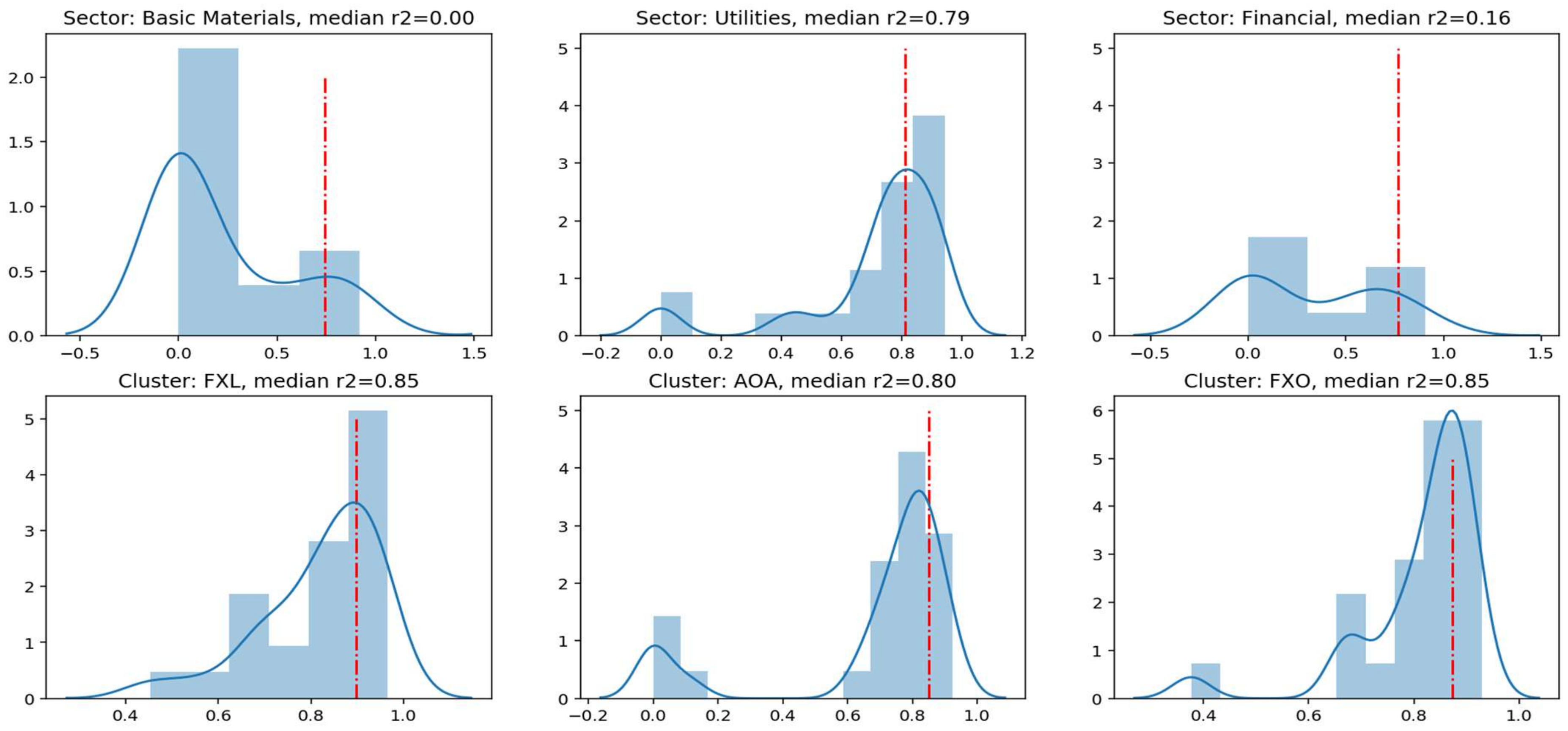


RNN forecast charts example



- RNN models comparable with ARIMA for short-term forecasting
- LT forecast looks weak (as for ARIMA), but has some potential for improvement

RNN tests on random comps from cluster



Simple random simulations show good result of model usage for clusters based on correlation approach

Issues and Recommendations

RESEARCH ISSUES

- The correlative clustering method helps identify homogeneous companies.
- Qualitative short-term forecasting of share prices using ML-algorithms is possible.
- In order to achieve maximum quality, each model must be tuned for the specific company.
- The task of long-term prediction of daily quotes - is not achievable only at these quotes as inputs.
- Use of sparse (quarterly) financial reporting data - does not improve the quality of predictions for daily prices.

BENEFITS for BUSINESS

- Trained models with high forecast accuracy - powerful tools for making investment decisions
- Correlation clusters with different construction criteria - can be used to form effective investment portfolios

Future Work

- **Add daily macro and commodities data for model improvements**
- **Develop clustering methods for forming different types of clusters**
- **Create approach for finding optimal portfolio based on correlation data**


Thank You
Insert the Sub Title of Your Presentation

APPENDIX

Project stages in details

Loading Cleaning
- GET API
- Tickers Filtering
- Files Consolidation
- Timeline
- Missing Values
- PCT calculation

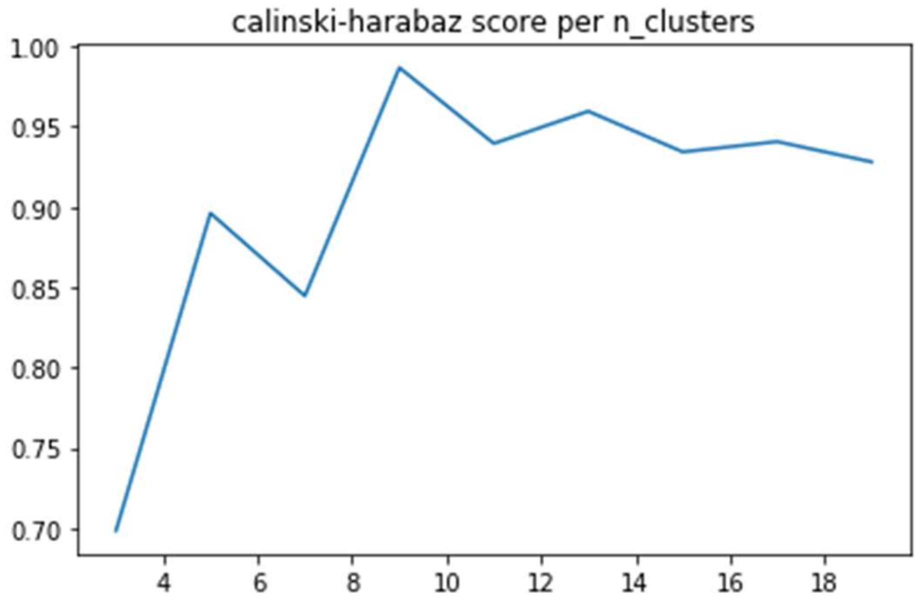
Clustering
- EDA on profiles
- EDA on prices
- Composite clusters (groupby)
- Kmeans
- Corr-clusters as Graphs

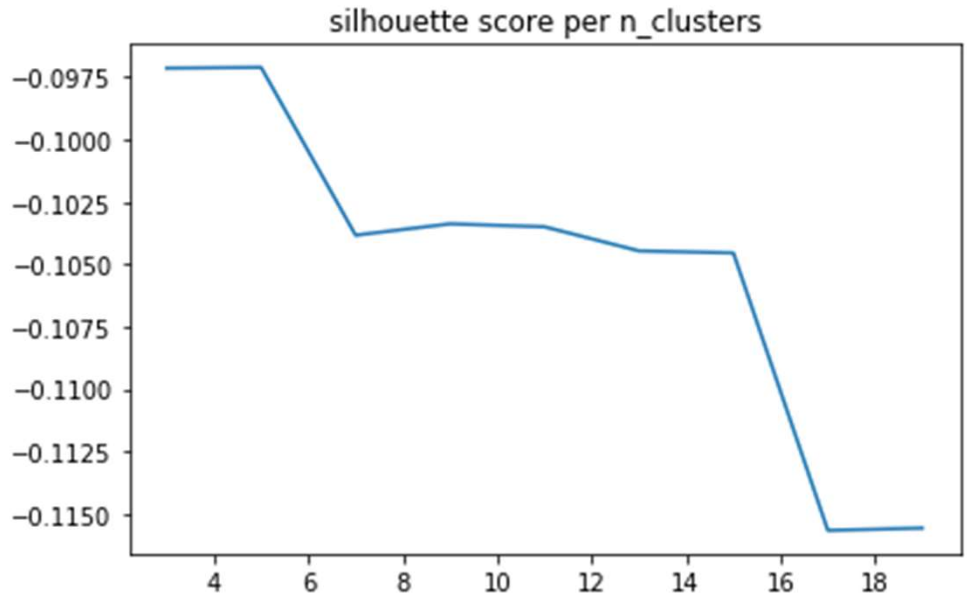
ARIMA
- EDA (rolling plots)
- ACF, PACF, ADF
- HyperParam search
- Fitting models
- ST and LT predicts
- R2 scores

LSTM
- EDA (date vars)
- Model Architecture
- Fitting models
- ST and LT predicts
- R2 scores
- Testing models on cluster companies

LSTM FS
- GET API (FS)
- EDA (chosen ticker)
- Models Architectures (with/without FS)
- Fitting models
- ST and LT predicts and comparison
- R2 scores

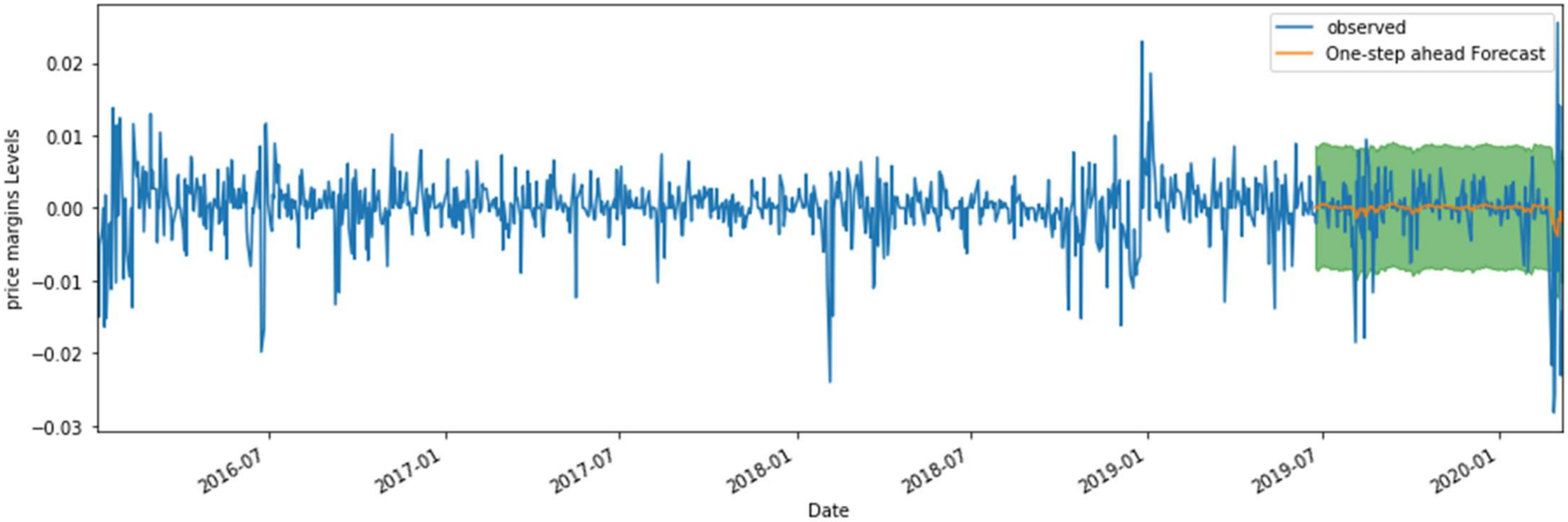
Kmeans applied on Net Price Margins





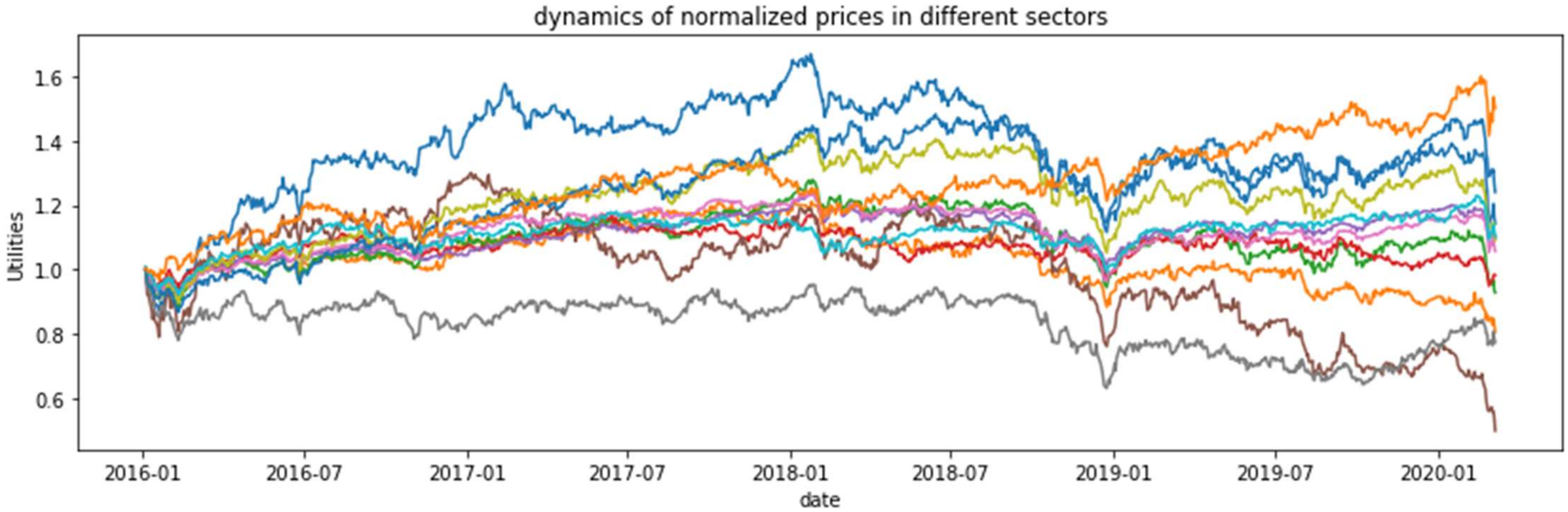
Negative silhouette score indicates impossibility to apply unsupervised learning algorithm here

ARIMA forecast based on Net Price Margins

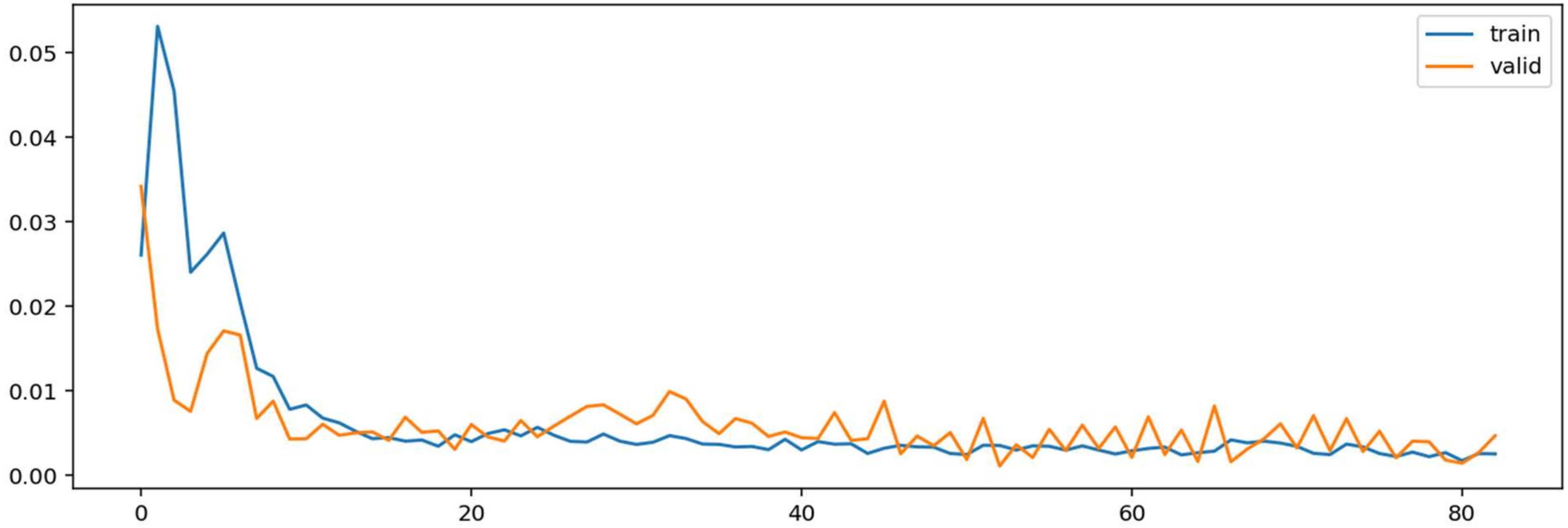


Price Return rates mainly distributed around zero and close to White Noise process
Such forecasts have weak R2 scores

Normalized Price dynamics for sector composites



Example of learning curves for LSTM models



Finance Composite and chosen company