Project Machine Learning

The goal of the project is to develop a model capable of predicting expected returns on the basis of a ranking of ratings (scores) assigned to all the evaluated stocks.

List of contents

Files

Code files

- Data preparation
- Time Series analysis
- Initial model
- Advanced modelling

Specifications

- Data preparation
- Time Series analysis
- Initial model
- Advanced modelling

Description

Data preparation

The dataset consists of date, stock index, sector, score (raiting), rate of return and closing price.

Date	symbol	sector	score	return_rate	close
2004-02-11	AEE	Utilities	0.670127	0.002350	70.309999
2004-02-11	AOS	Producer Manufacturing	0.753176	0.007533	8.005000
2004-02-11	APA	Energy Minerals	0.912117	0.005808	59.630001
2004-02-11	ARLP	Energy Minerals	0.669621	-0.011510	13.578750
2004-02-11	ATO	Utilities	0.672410	0.000765	39.230000

The main dataset is a combination of two datasets. The first set comes directly from the lecturer and includes expert assessment of the company. The second set was downloaded by the authors at Yahoo finance and used to calculate the rate of return.

Logarithmic rates of return were obtained for the daily data, and then aggregated into two-week intervals according to the dates for lecturer's set. In the further part of the project, both points and rates of return were aggregated to monthly, semi-annual and annual data, as an average for a given period for points and a sum for rates of return.

During data preparation, `512` stock indices were removed because the symbols of companies in both sets hadn't match.

```
Number of all unique symbols: 1834
Number of symbols in dataset: 1322
Number of missing symbols: 512
```

As a result of removing missing symbols and closing prices, the dataset has `30324` rows.

```
Old data frame length: 37360
New data frame length: 30324
Number of rows deleted: 7036
```

The mean score for this dataset is `0.73` , while mean closing price is `199.86` and mean return rate is `0.008` .

```
#basic statistics
data.describe()
```

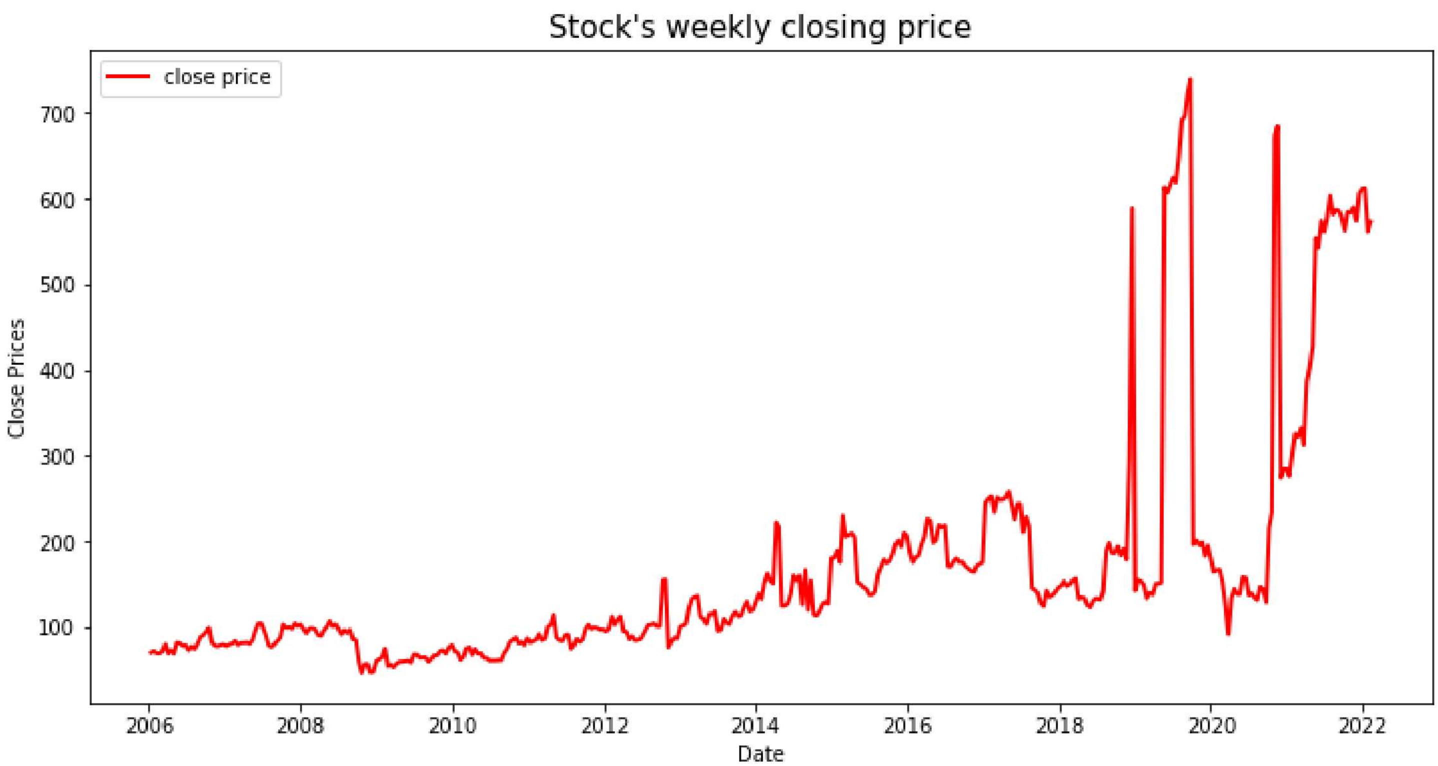
	score	return_rate	close
count	30324	30324	30324
mean	0.731377	0.008940	199.862674

```
|std   | 0.117693 | 0.152564    | 5225.047203
|min   | 0.413554 | -0.507207   | 0.030000
|25%   | 0.653702 | -0.018704   | 51.757969
|50%   | 0.741667 | 0.007707    | 89.154999
|75%   | 0.813701 | 0.034849    | 148.037234
|max   | 0.987225 | 24.600929   | 898434.375000
```

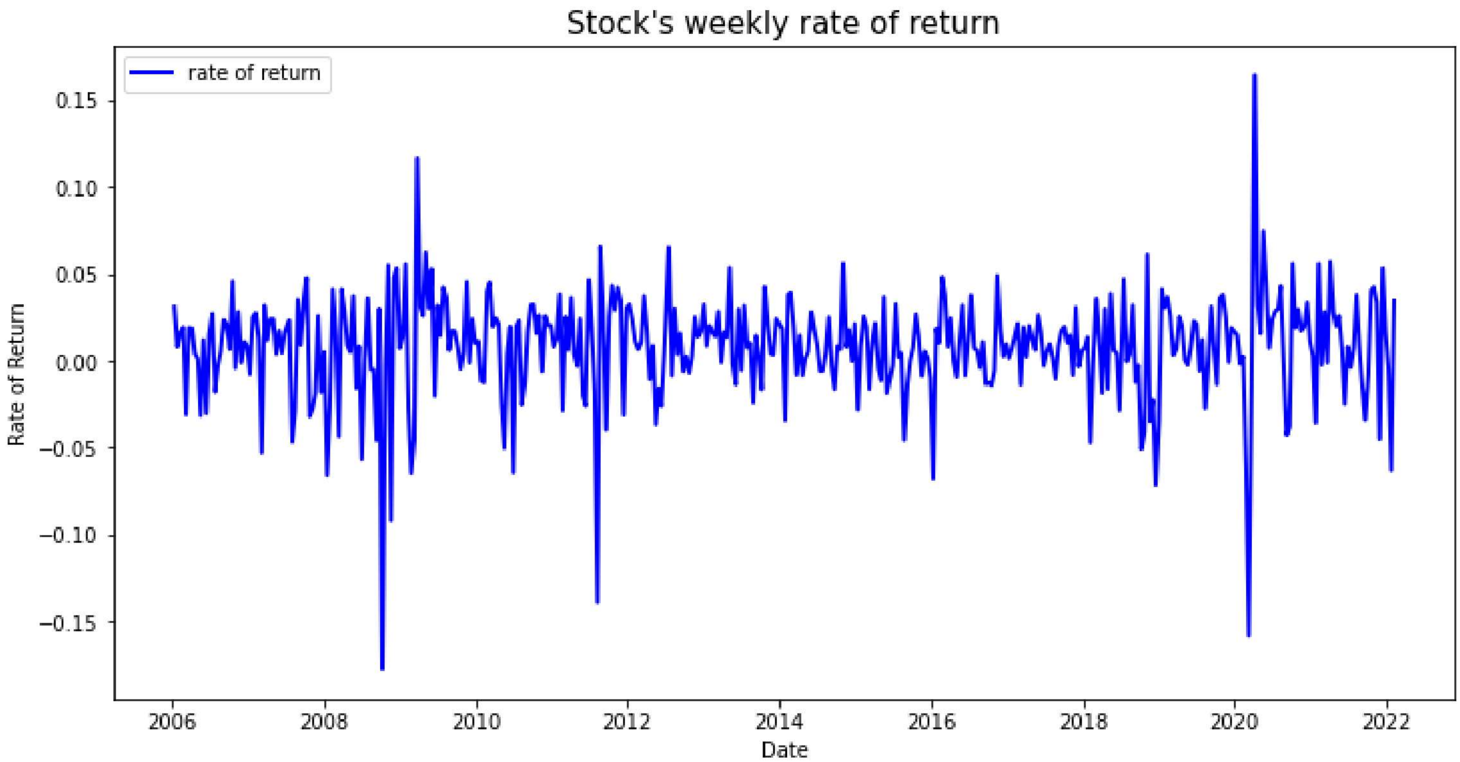
Time Series analysis

The time-series begins on `2004-02-11` and ends on `2022-02-09`. The analysis involves data from `2006`. The closing price and the rate of return are parametric measures.

Visualization of the stock’s weekly closing price and rate of return



The process above is not stationary, because the mean is not constant through time.



The rate of return has many fluctuations, while the seasonality is not observed. The highest deviance was observed in `2008` with a weekly return of `-17%` . In `2020` , the biggest fluctuations on rate of return were found out in between `-14%` and `17%` .

Dickey-Fuller test

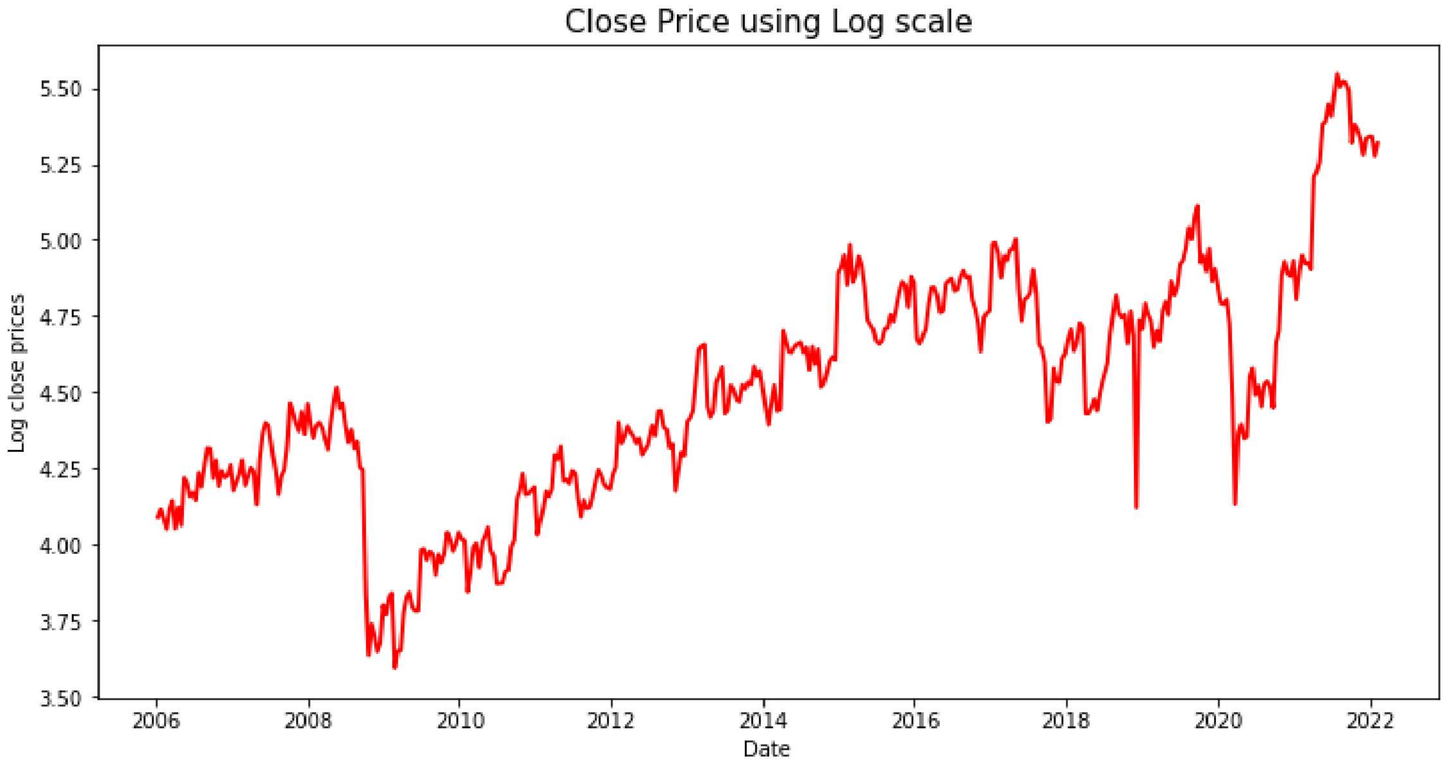
Dickey-Fuller test can be used to determine whether a series has a unit root or not, and thus whether a series is stationary (H_0) or not (H_1).

```
Results od Dickey-Fuller Test
        Values                       Metric
0    -21.177000              Test Statistics
1      0.000000                      p-value
2     35.000000             No. of lags used
3  27670.000000  Number of observations used
4     -3.430586          critical value (1%)
5     -2.861644          critical value (5%)
6     -2.566826         critical value (10%)
```

We can rule out the Null hypothesis because the p-value is smaller than 0.05. Additionally, the test statistics exceed the critical values. As a result, the data is **nonlinear**.

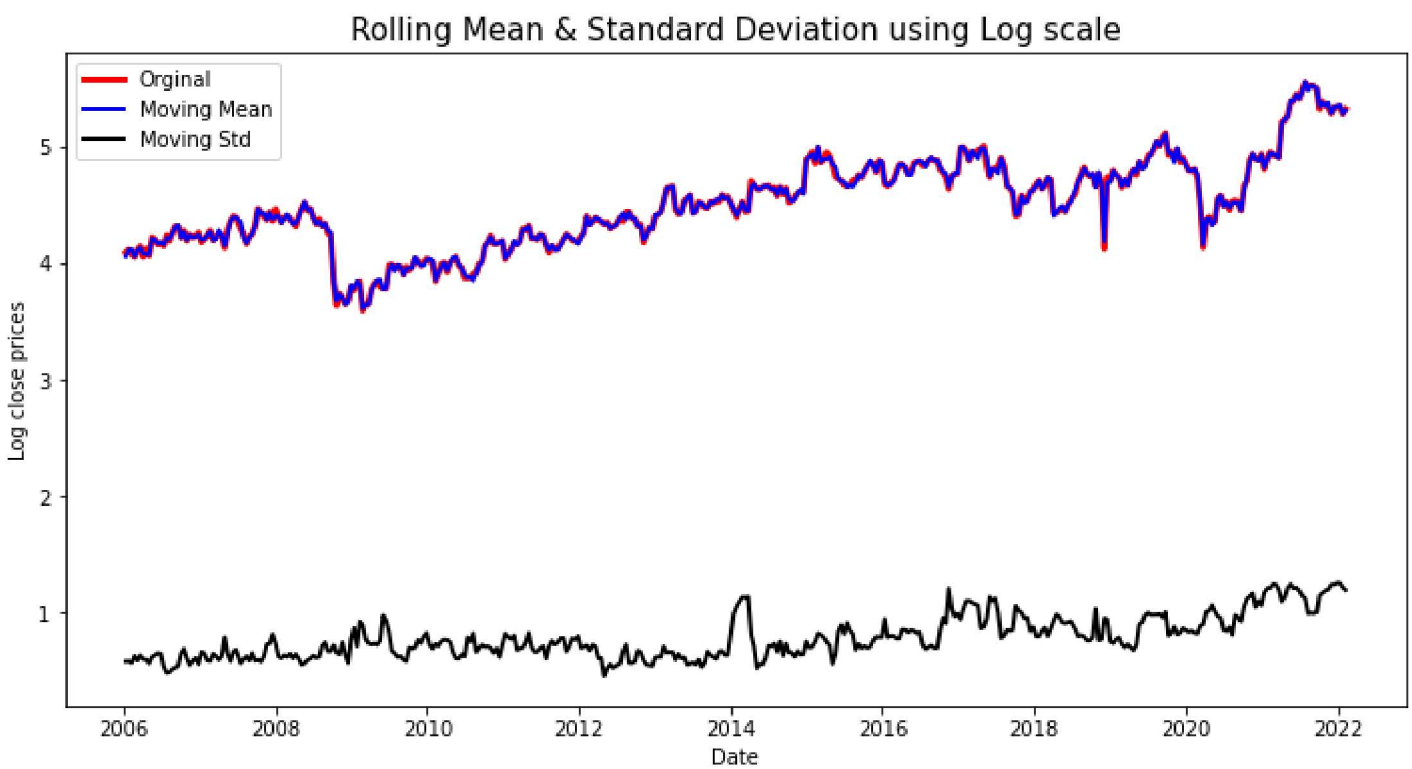
Estimating trend

The log of the series was used to reduce the magnitude of the values and the growing trend in the series.



Visualization of logarithmic closing prices. The price drops are the results of crises. The trend is growing.

Rolling statistics

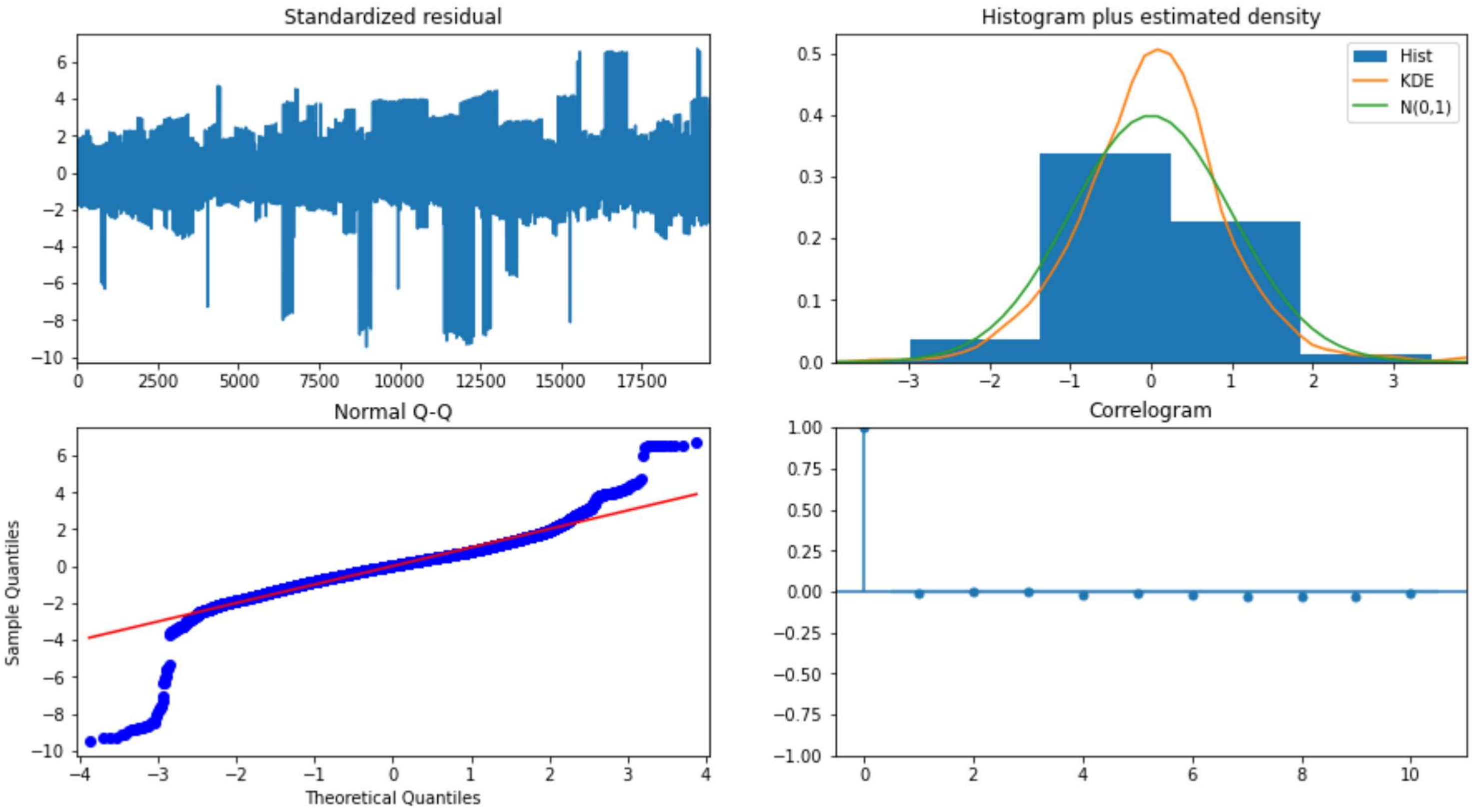


As a result of smoothing out the previous quarter, it is difficult to see the trend, as it is too close to the actual curve. In addition, a rising mean and standard deviation can be observed, indicating that our series isn't stationary.

SARIMAX (3, 0, 3) model

```
train_data = df_log['2010':'2020']
test_data = df_log['2021':'2022']
```

Model



Standardized residual

The first chart shows the grouping of volatility. The residual errors appear to have a uniform variance and fluctuate between -2 and 2.

Histogram plus estimated density

The density plot suggests a normal distribution with a mean of zero which is the excess kurtosis with long tails.

Normal Q-Q

Normal Q-Q shows deviations from the red line, both at the beginning and at the end, which would indicate a skewed distribution with long tails.

Correlogram

The fourth graph shows the linear relationships in the first lag. As a result, more Xs (predictors) have to be added to the model.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                19568
Model:               SARIMAX(3, 0, 3)   Log Likelihood              -25091.679
Date:                Tue, 07 Jun 2022   AIC                          50199.358
Time:                        21:49:30   BIC                          50262.411
Sample:                             0   HQIC                         50220.009
                              - 19568
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.2313      0.043      5.379      0.000       0.147       0.316
ar.L1         -0.8841      0.207     -4.276      0.000      -1.289      -0.479
ar.L2          0.9484      0.026     36.689      0.000       0.898       0.999
ar.L3          0.8847      0.198      4.463      0.000       0.496       1.273
ma.L1          0.9169      0.205      4.482      0.000       0.516       1.318
ma.L2         -0.8899      0.027    -33.373      0.000      -0.942      -0.838
ma.L3         -0.8590      0.189     -4.540      0.000      -1.230      -0.488
sigma2         0.7578      0.003    240.859      0.000       0.752       0.764
===================================================================================
Ljung-Box (L1) (Q):                   0.63   Jarque-Bera (JB):             99146.88
Prob(Q):                              0.43   Prob(JB):                         0.00
Heteroskedasticity (H):               1.69   Skew:                            -0.69
Prob(H) (two-sided):                  0.00   Kurtosis:                        13.94
===================================================================================
```

The best model with the lowest `AIC = 50199.358` was selected.

Is each coefficient statistically significant?

Test hypothesis:

- Null Hypothesis: each coefficient is NOT statistically significant.
- Alternate Hypothesis: the coefficient is statistically significant (p-value of less than 0.05).

Each parameter is statistically significant.

Are the residuals independent (white noise)?

The Ljung Box test is used to verify if the errors are white noise.

The probability (`0.43`) is above 0.05, so **we can’t reject the null hypothesis that the errors are white noise**.

Do residuals show variance?

Heteroscedasticity test verifies if the error residuals are homoscedastic or have the same variance.

Test statistic is `1.69` while p-value of 0.00, which means that we can reject the null hypothesis and the **residuals show variance**.

Is data normally distributed?

Jarque-Bera test verifies the normality of the errors.

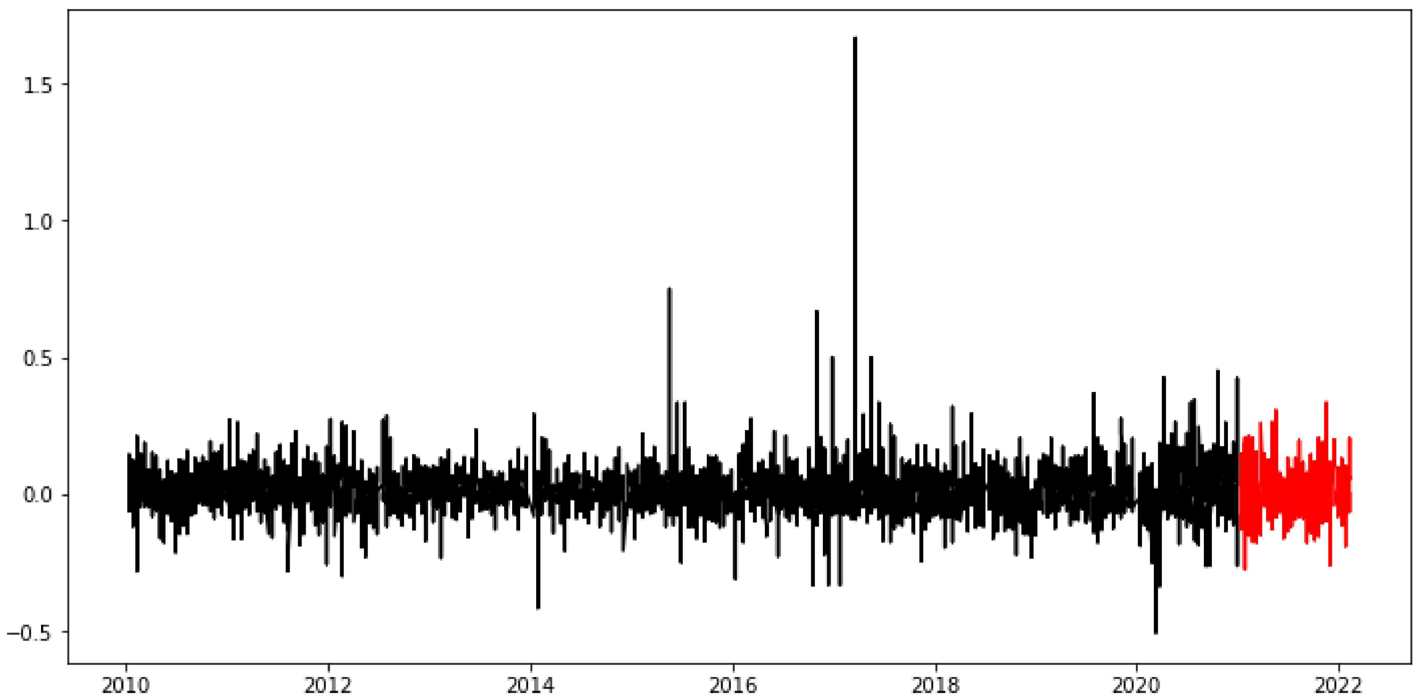
Test statistic of `99146.88` with a probability of `0` , which means THAT we reject the null hypothesis, and **the data is not normally distributed**.

In addition, the results show:

- Negative skewness - left side asymmetry (long tail on the left side).
- Excess kurtosis - results fluctuate around a mean

Initial model

Training and test sets



Training set involves data `from 2010 to 2020` while test set includes the years `2021 and 2022` .

Training set consists of `19568` observations whereas test set has `2281` observations.

Dummy regression

```
Coefficient of determination: 0.0
```

`0%` indicates that the model does not fit the training data.

```
Coefficient of determination (R2): -0.00021
Mean absolute error (MAE): 0.04415
Residual sum of squares (MSE): 0.00348
Root mean squared error (RMSE): 0.05903
```

Linear Regression

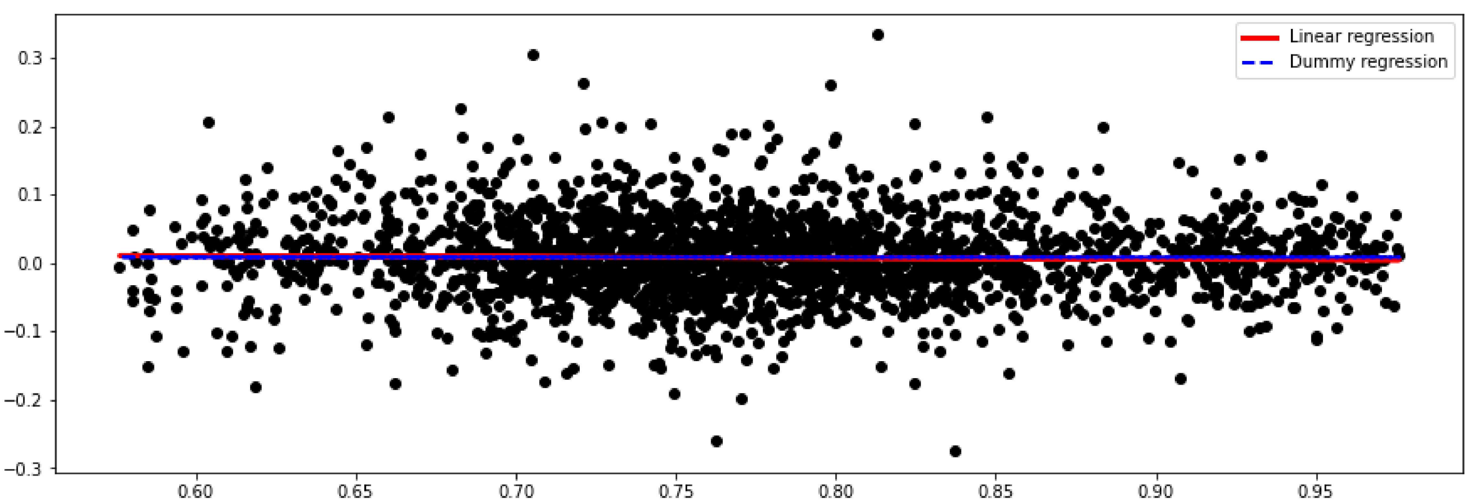
$$f(x) = - 0.015x + 0.019$$

```
Coefficient of determination: 0.001
```

`~1%` indicates that the model does not fit the training data.

```
Coefficient of determination (R2): -0.00067
Mean absolute error (MAE): 0.04414
Residual sum of squares (MSE): 0.00349
Root mean squared error (RMSE): 0.05904
```

Comparison between dummy regression and linear regression combinaed with observations from the test set.



The model does not explain any variation in the response variable around its mean.

Linear regression is marginally better than dummy regression.

Both models do not fit the variables.

Advanced modelling

We created three Machine Learning Models (SVR, Decision Tree Regressor and LASSO Regressoin) using k-fold Cross-Validation.

Comparison performance of the models against each other for the various investment horizons

We used R2 score, MAE and MSE to compare the models.

R-squared

R2

According to R2, none of the observed variation can be explained by the input data of the models.

However, R2 is not a proper statistical measure for these advanced models.

Mean Absolute Error

```
Mean absolute error (MAE) for 1M SVR Model: 0.05581
Mean absolute error (MAE) for 1M Decision Tree Regressor: 0.05678
Mean absolute error (MAE) for 1M LASSO Regression: 0.05599
```

According to MAE, the best model is SVR for a month interval.

```
Mean absolute error (MAE) for 6M SVR Model: 0.08662
Mean absolute error (MAE) for 6M Decision Tree Regressor: 0.09145
Mean absolute error (MAE) for 6M LASSO Regression: 0.08786
```

According to MAE, the best model is SVR for half a year interval.

```
Mean absolute error (MAE) for 1Y SVR Model: 0.11689
Mean absolute error (MAE) for 1Y Decision Tree Regressor: 0.12315
Mean absolute error (MAE) for 1Y LASSO Regression: 0.11999
```

According to MAE, the best model is SVR for year interval.

Mean Squared Error

```
Mean squared error (MSE) for 1M SVR Model: 0.00702
Mean squared error (MSE) for 1M Decision Tree Regressor: 0.00719
Mean squared error (MSE) for 1M LASSO Regression: 0.00706
```

According to MSE, the best model is SVR for a month interval.

```
Mean squared error (MSE) for 6M SVR Model: 0.01477
Mean squared error (MSE) for 6M Decision Tree Regressor: 0.01599
Mean squared error (MSE) for 6M LASSO Regression: 0.01507
```

According to MSE, the best model is SVR for half a year interval.

```
Mean squared error (MSE) for 1Y SVR Model: 0.03901
Mean squared error (MSE) for 1Y Decision Tree Regressor: 0.04129
Mean squared error (MSE) for 1Y LASSO Regression: 0.04023
```

According to MSE, the best model is SVR for year interval.

SVR is the best model for each investment horizons.

Comparison performance of the models against a baseline model

The mean absolute error for linear regression is `0.04414` .

This is the lowest average magnitude of the errors.

The mean square error for linear regression is `0.00349` .

This model is the closest to finding the line of best fit.

According to MAE and MSE, the basic model is better than the advanced models.

SUMMARY

We created two initial models - dummy and linear regression, and three advanced Machine Laerning Models - SVR, Decision Tree Regressor and LASSO Regressoin.

- For all models, the R2 score is close to `1%` , which means that they all do not fit well with the variables.
- The lowest value for MAE is in linear regression. This means that the average distance between the predicted and true values is `0.04414` .
- The lowest value for MSE is in linear regression, so this is the best model.

Technologies

Project is created in Python with:

- matplotlib version: 3.3.4

- numpy version: 1.20.1
- pandas version: 1.2.4
- pmdarima version: 1.8.5
- scikit-learn version: 0.24.1
- seaborn version: 0.11.1
- statsmodels version: 0.12.2
- yfinance version: 0.1.70

Authors

Wiktoria Ekwińska

Bartek Gimzicki

Agnieszka Pijaczyńska