Weronika Król s176057
Dawid Krefta s173355
Dariusz Kobiela s175656

PROJECT REPORT: Stock price prediction

1. Problem description

In our project, we try to forecast the average price for the next 1 day and 7 days using an artificial neural network - LSTM, based on the historical prices of a given company and external information related to a given company. To check the effectiveness of such a solution, we are comparing it with a naive autoregressive approach - ARIMA statistical model based only on the historical prices of a given company.

We collected data for two corporations Volkswagen Group and The British Petroleum Company from different sources (NASDAQ stock exchange, ycharts.com, tradingeconomics.com). There were two main problems connected with gathering daily data for dividends, earning, crude oil price and steel price.

- The first one was that data was not easily available. We had to create an account on each platform to have access to the data. If we would like to download all historical data in one file we had to pay above $400, without that we could only see the data for a particular time period. Additionally, only for access to crude oil data, we have already paid $4.
- To gather all the data without the possibility of downloading files, we had to manually webscrape the data for each available subset free of charge. It took a few hours to have all the daily data

Screenshots of prepared data frames are presented below.

BP

	timestamp	avg_price	dividend_yield	earnings_yield	enterprise_value	crude_oil_price
0	2008-01-02	73.425193	0.0346	0.0882	2.591800e+11	99.64
1	2008-01-03	74.997322	0.0339	0.0865	2.635900e+11	99.17
2	2008-01-04	74.371617	0.0345	0.0878	2.601500e+11	97.90
3	2008-01-07	74.890928	0.0338	0.0863	2.643200e+11	95.08
4	2008-01-08	74.871232	0.0344	0.0877	2.605300e+11	96.43
...	...	...	...	...	...	...
3406	2021-07-23	23.555018	0.0535	0.1087	1.276100e+11	72.07
3407	2021-07-26	24.246797	0.0518	0.1052	1.302700e+11	71.91
3408	2021-07-27	24.185154	0.0520	0.1055	1.300000e+11	71.65
3409	2021-07-28	24.334298	0.0516	0.1049	1.305100e+11	72.39
3410	2021-07-29	24.834807	0.0512	0.1039	1.312800e+11	73.62

3411 rows × 6 columns

VLKV

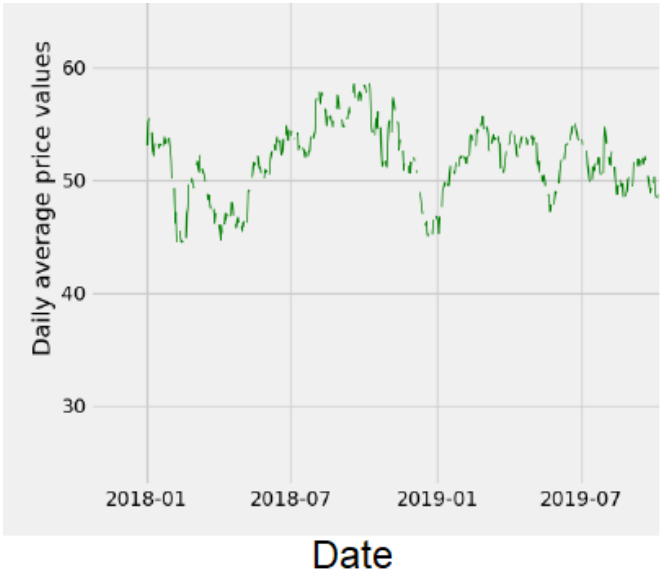
	timestamp	avg_price	dividend_yield	earnings_yield	enterprise_value	steel_price
0	2009-12-29	112.247800	0.0227	0.0498	8.907000e+10	1718.680
1	2009-12-30	112.000000	0.0228	0.0501	8.887000e+10	1707.320
2	2009-12-31	112.000000	0.0228	0.0295	1.251800e+11	1713.185
3	2010-01-04	109.500000	0.0234	0.0302	1.241700e+11	1735.305
4	2010-01-05	108.500000	0.0236	0.0305	1.237600e+11	1782.035
...	...	...	...	...	...	...
2812	2021-10-21	320.135005	0.0354	0.0354	2.255200e+11	1567.385
2813	2021-10-22	322.286857	0.0349	0.0349	2.270400e+11	1564.040
2814	2021-10-25	330.632845	0.0341	0.0341	2.291600e+11	1600.850
2815	2021-10-26	341.457074	0.0330	0.0330	2.325300e+11	1621.665
2816	2021-10-27	333.013794	0.0338	0.0338	2.301700e+11	1592.755

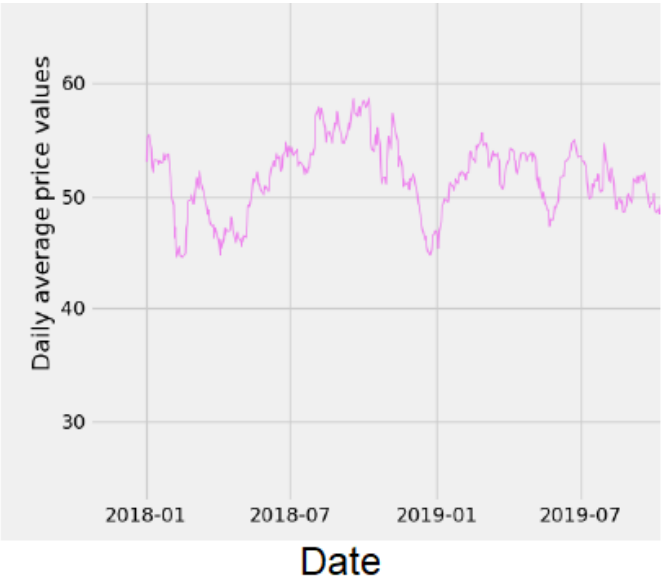
2817 rows × 6 columns

Column name	Description	Source
timestamp	Date in format YYYY-MM-DD	NASDAQ
avg_price	Average stock price in particular day	NASDAQ
divident_yield	The dividend amount due to the shareholder on account of the profit earned by the company.	ycharts.com
earning_yield	Earnings per share for the most recent 12-month period divided by the current market price per share.	ycharts.com
enterprise value	The value of the entire company in U.S. dollars	ycharts.com
crude_oil_price	Crude oil price in U.S. dollars per barrel	https://macrotrends.dpdcart.com/
steel_price	Steel price per ton in U.S. dollars	tradingeconomics.com

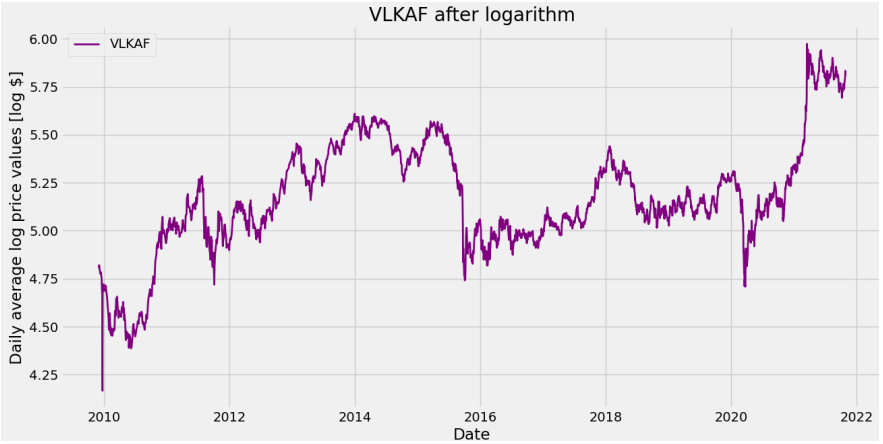
2. Data processing

The first phase of processing is to convert raw parquet data, downloaded from NASDAQ AWS bucket, into raw csv data. Data is aggregated to daily intervals for all days in the week, so from Monday to Sunday (including weekends and holidays). Reasons for choosing such an approach are motivated based on Kra, Brinwa et al. [1], whose results indicate a significant negative effect of weekends on Mondays, and the discussions with people connected with time series forecasting and different possible approaches (business days vs. all week days). Buying or selling stocks is not possible during weekends (Saturday and Sunday). Although the NASDAQ stock exchange is closed during weekends, the price of a particular company is still changing. It can be easily compared by taking the chosen Friday closing price and the following Monday opening price. The value is usually different. Not Available data were fulfilled by interpolation, which is shown in the exemplary figure below. Right chart shows data after interpolation (without missing values).

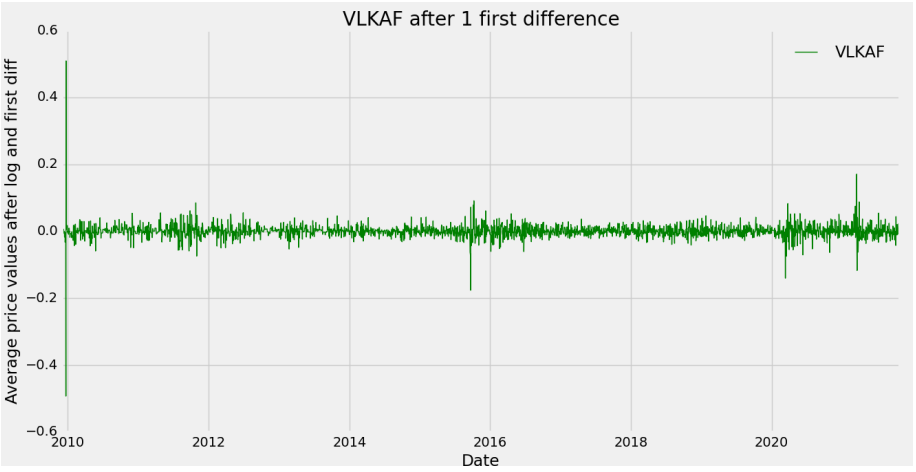




To prepare data for ARIMA models we had to logarithm them.



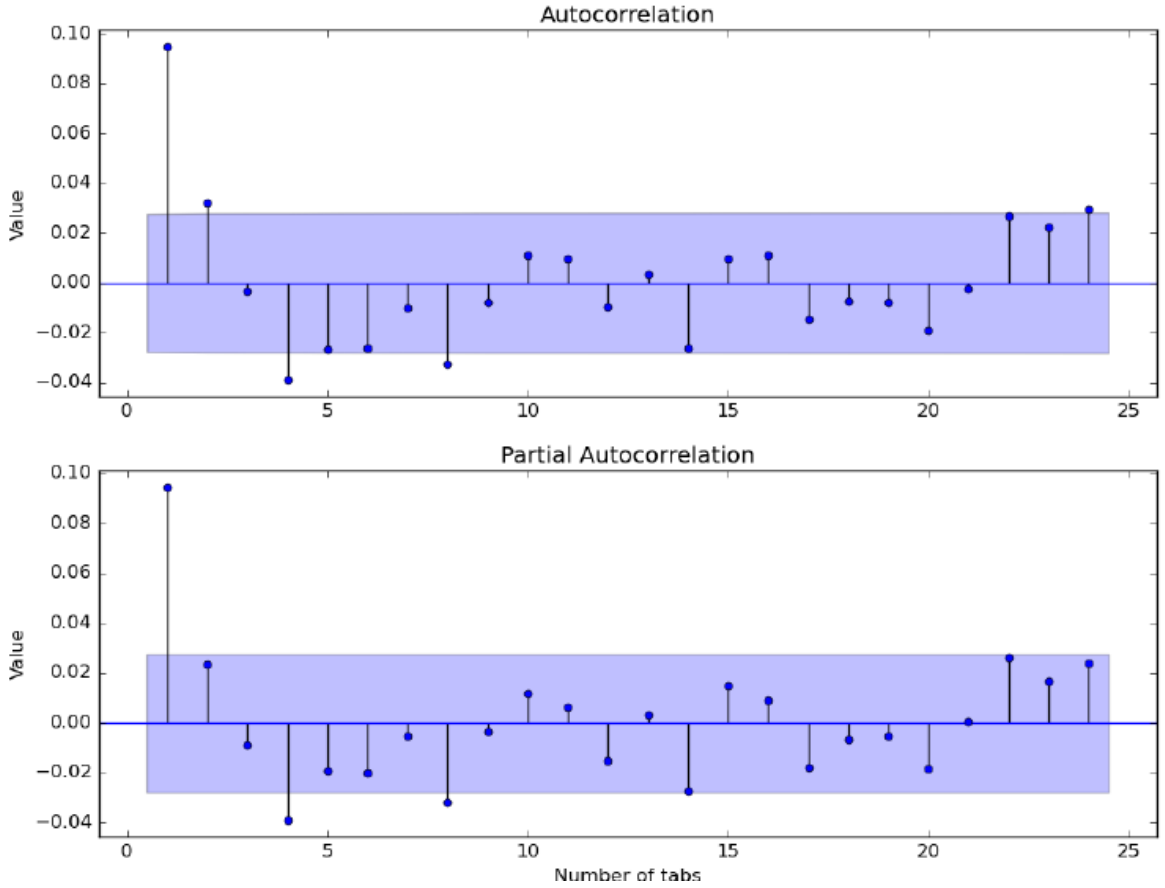
After that to get a stationarity of time series we had (in our case) to differentiate it.



3. Description of the used models

3.1 ARIMA

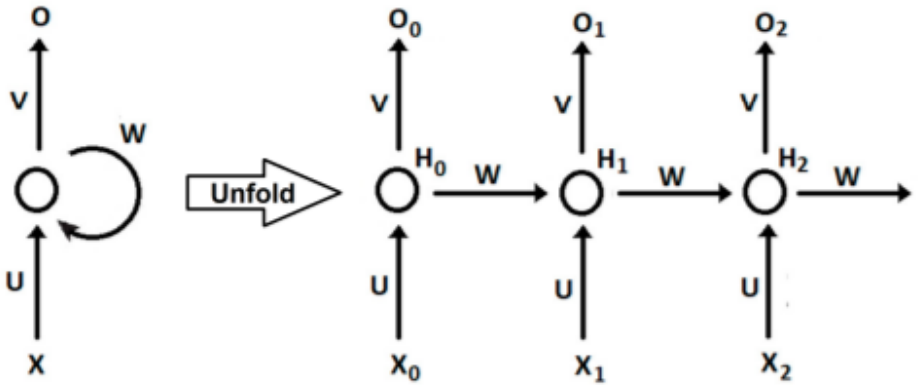
The ARIMA model is a combination of two processes - autoregressive AR and moving average MA, which is weighted delayed random components. The letter I in the model name indicates the level of integration of the analyzed variable. Integrated variables are variables that can become stationary through differentiation. The structure of ARIMA is based on the phenomenon of autocorrelation. ARIMA can be used for modeling stationary time series or non-stationary time series that can become stationary through differentiation. Stationary series are those whose expected value and variance do not change over time, the series values themselves do not deviate from the initial values, and the value of the covariance for two observations depends on the spacing between them, not the timing of their origin. The universal notation ARIMA (p, d, q) is used to describe the form of the ARIMA model. The letter p is the order of the regression, d is the order of differentiation and q is the order of the moving average. The process of building the ARIMA model can be divided into three phases, which follows the Box-Jenkins procedure. In the first phase - identification, the characteristics of the analyzed time series are checked. A decision is made about the need for data to transform and differentiate the series to stabilize its mean, variance covariance. This is done by both examining the autocorrelation function ACF and partial autocorrelation (example from our data is shown below).



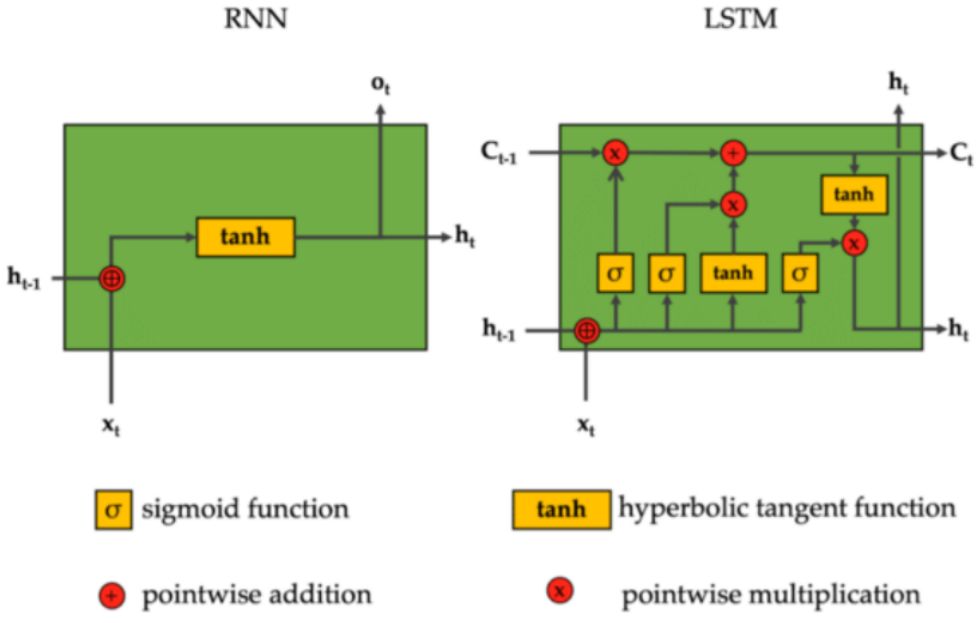
PACF and by performing statistical Dickey-Fuller and Augmented-Dickey-Fuller tests. In the second phase - estimation and tests, the parameters of selected models are estimated. The final model selection is usually based on the analysis of several criteria - the significance of model parameters, error metric and information criterion (Akaike's Information Criterion, Bayesian Information Criterion). The next step is a diagnostic check. Properties of a number of model residuals are analyzed. If the residuals of the model are a white noise process, there are no significant ACF or PACF values of a series of model residuals, the model can be used for forecasting. Otherwise, the estimation and testing phases should be repeated and a different model should be selected. In some cases, it may be necessary to return to the identification phase. In the third phase the model is used to prepare a forecast [2]. Forecast is performed using in-sample and out-of-sample periods. The given dataset is splitted into an in-sample period, used for the initial parameter estimation and model selection and an out-of-sample period used to evaluate forecasting performance. Empirical evidence based on out-of-sample forecast performance is generally considered more trustworthy than evidence based on in-sample performance, which can be more sensitive to outliers.

3.2 LSTM

Recursive neural networks (RNNs) are a broad class of networks in which, even as the input data changes over time, the same parameters are used. Recursive networks are among the most successful models that are devouring applications both in research and industry to problems related with sequential data in natural language processing, time series prediction or classification. The main differences between unidirectional and recursive networks are that RNNs see the time data in an ordered form as values in the successive steps and have a state that is preserved between successive time steps. It is this state, as well as its static parameters, that are responsible for updating the network response after providing it with new information in the next steps. Due to the fact that RNNs have a cell-based architecture, they are represented by the "build and unfold" paradigm visualized in a figure below. It describes how the same parameters are used over and over again in a network. It is this methodology that makes the number of them small despite the very long time series.



The disadvantage of RNNs is that they suffer from short-term memory. One of the solutions to that problem is a model introduced by German scientists [3] called long short-term memory (LSTM). It is capable of learning long-term dependencies using a mechanism called gates, which are different tensor operations that can learn what information to add or remove to the hidden state. Comparison of a single unit of classical RNN with LSTM is presented in the figure below.



3.3 Metrics

Models must be assessed for their predictive performance. In forecasting, errors such as Root Mean Square Error, Mean Square Error, Mean Percentage Absolute Error and Mean Absolute Error are most often used. We've decided to use MAPE in our experiments to compare obtained results. Mean Absolute Percentage Error (MAPE) expresses the percentage of the absolute mean difference between the predicted values and the actual values divided by the actual value. It is expressed by the formula:

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{N} \left| \frac{e_t}{y_t} \right|$$

where:

n - number of observations in the test set,

e_t - difference between the actual value and predicted value,

y_t- the actual value.

Mean Square Error (MSE) is a metric that evaluates the quality of the forecasting model. It is used to measure the difference between actual and predicted values. MSE takes into account both the variance, which is the dispersion of the predicted values from each other and the bias, which is the distance of the predicted value from the actual value. The error is represented by the formula:

$$MSE = \frac{1}{n} \sum_{t=1}^{n} e_t^2$$

where:

n - number of observations in the test set,

e_t - difference between the actual value and predicted value.

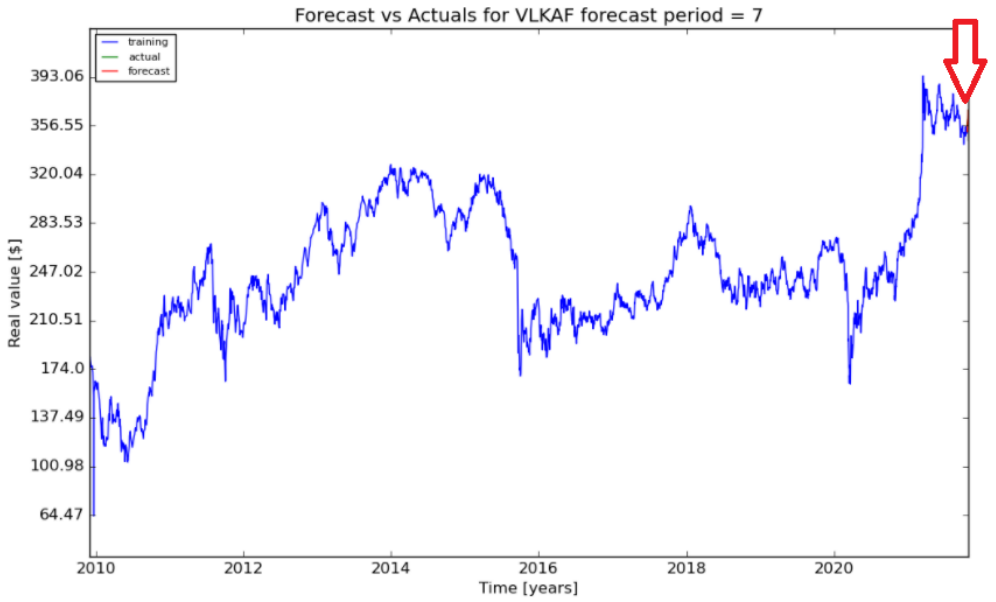
4. Experiments

4.1 ARIMA

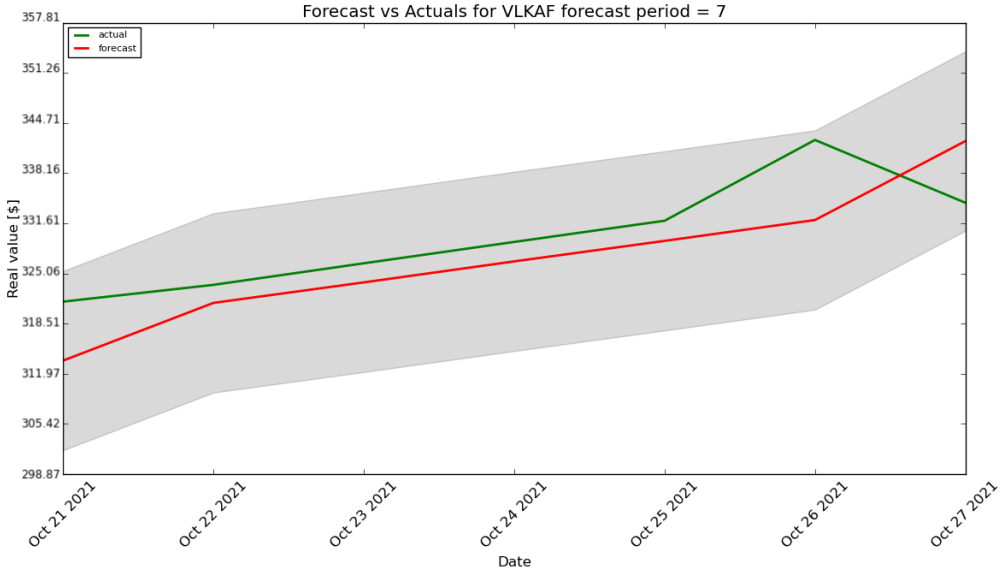
The only parameters that can be tested in the ARIMA statistical model are the parameters of p and q. However, in the initial experiments, the number of days taken for the prediction was also taken into account. Due to the fact that ARIMA can work quite differently in the cases of short-term and long-term prediction, a decision was made to test the 1-day and 7-day prediction cases. Through one particular experiment, the best p and q values were selected and the prediction effects were compared for periods of different lengths.

RESULTS:

	1 day	7 days
p	range 0-2	range 0-2
q	range 0-2	range 0-2
series averaging	False	False
output window size	1	1
execution time	0h 01min 44s	0h 01min 44s
test MAPE	2.29746	1.3209
best model	ARIMA(1, 1, 1)	ARIMA(1, 1, 1)
most frequent model	ARIMA(2, 1, 2)	ARIMA(2, 1, 2)



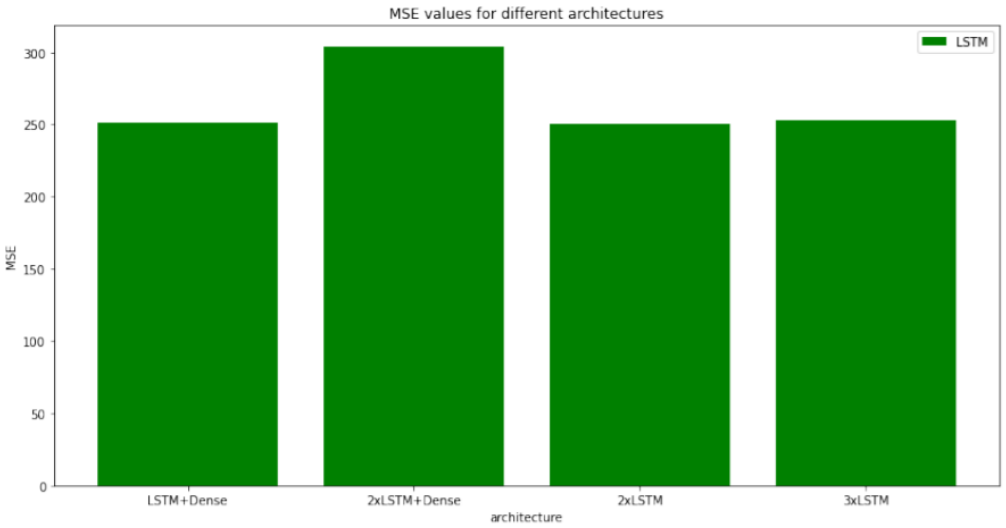
Zoom of the above chart to prediction period:



4.2 LSTM

4.2.1 LSTM architecture

We've decided to compare several architectures of LSTM to choose the best one for our future predictions. The results are shown below:

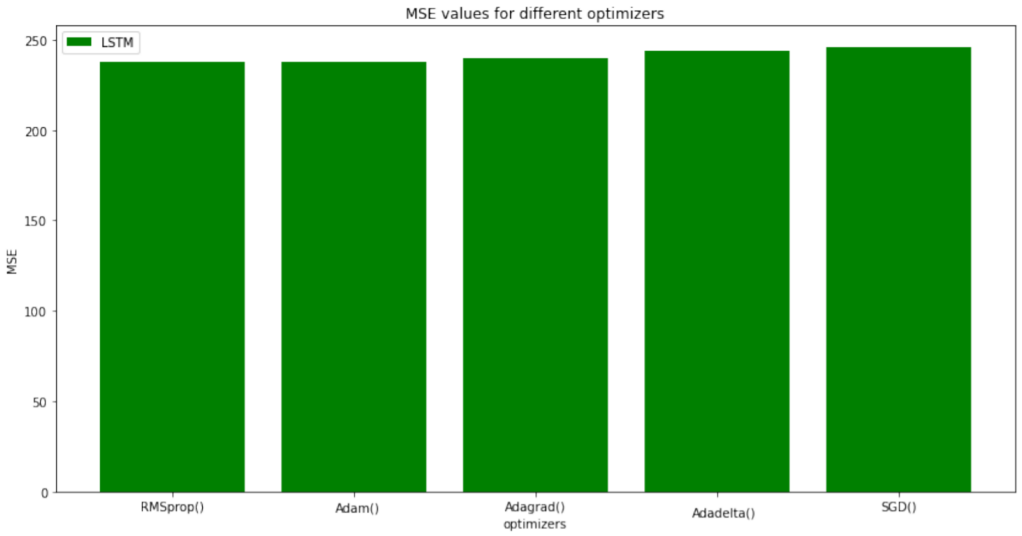


Based on obtained results we've chosen 2xLSTM architecture for the rest of the predictions

.

4.2.2 LSTM Optimizer

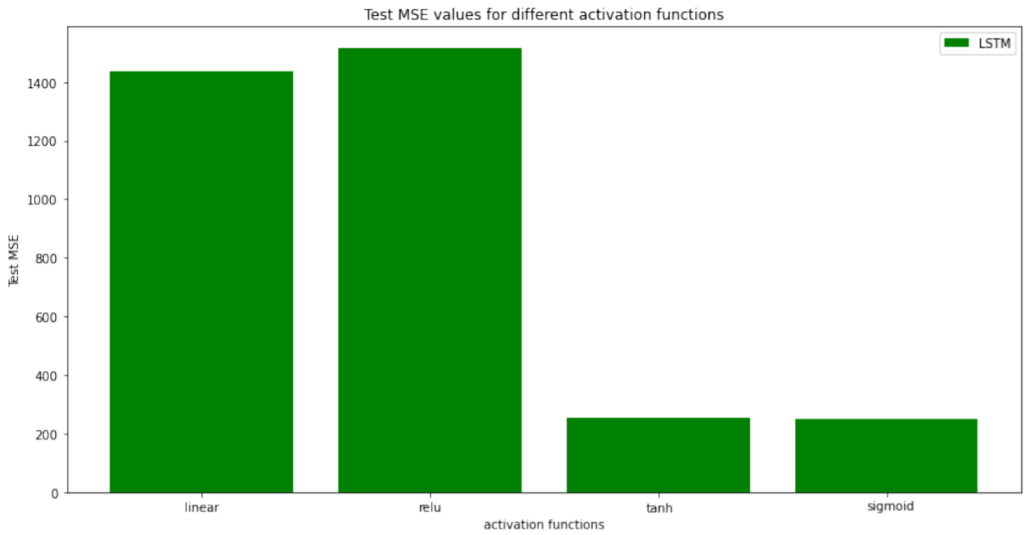
We've decided to test 5 different optimizers. The results are shown below:



We chose Adam() optimizer for further predictions.

4.2.3 LSTM Activation function

We've also tested 4 different activation functions. The obtained results are shown below.



The best activation function was sigmoid.

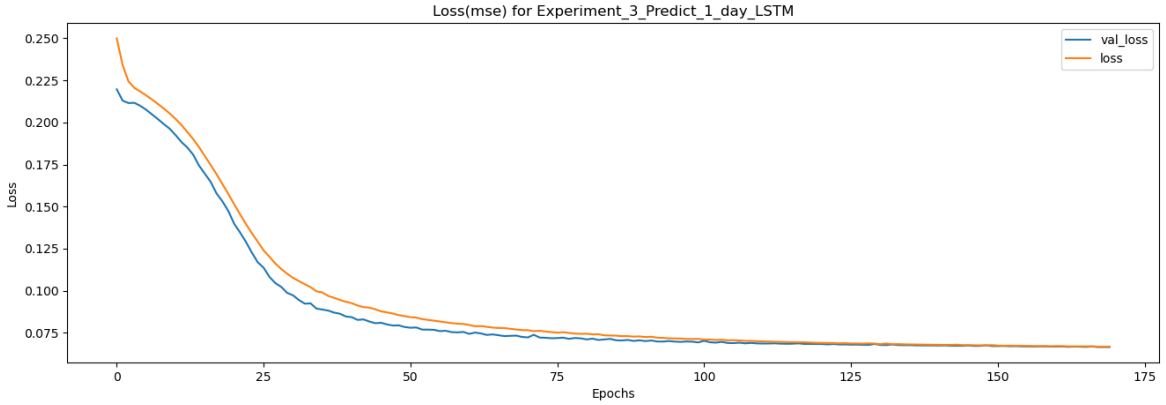
4.2.4 LSTM Training of 1-day and 7-days

A split ratio of 0.33 for the test set was applied.

LSTM Results:

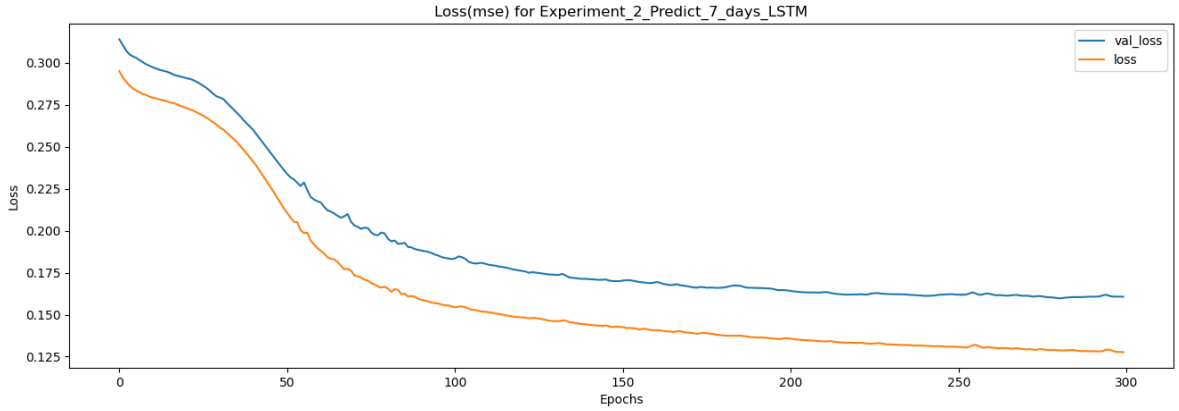
architecture	input window size	output window size	input averaging	output averaging	epochs performed	optimizer	activation function
2xLSTM	14	1	False	False	300	Adam()	sigmoid
2xLSTM	28	7	False	False	300	Adam()	sigmoid

MSE train and validation for predicting 1 day:



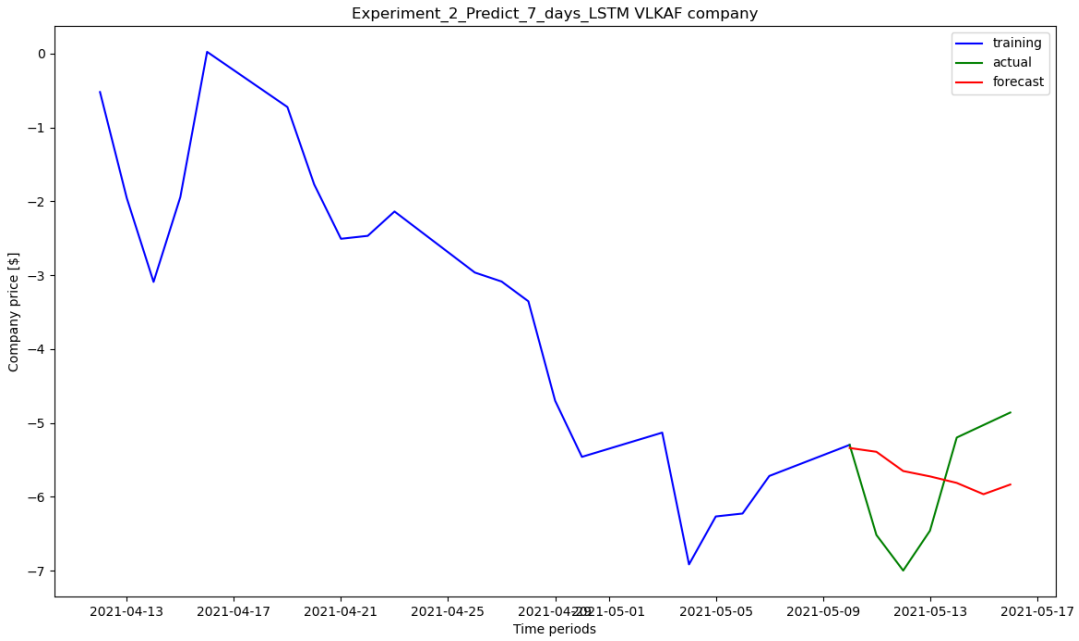
In the above figure we can see that even though we had set 300 epochs, only 168 were performed. It was caused by regularization technique used - **Early Stopping** (patience 25, min difference 0,01)

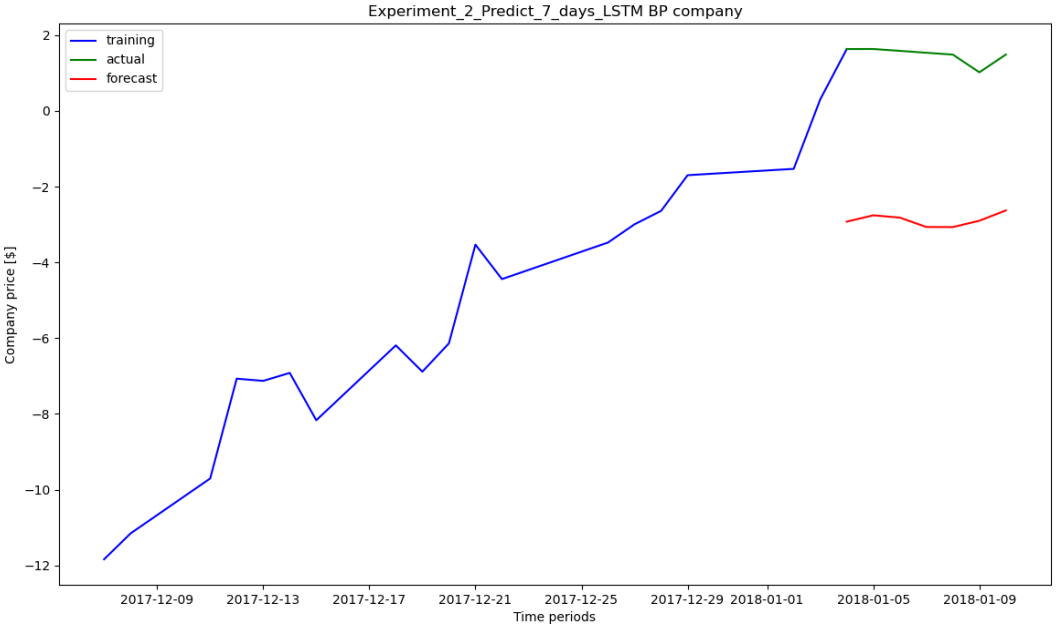
MSE train and validation for predicting 7 days:



4.2.5 LSTM Prediction for 7-days

As an examples we show two plots of 7 days prediction for Volkswagen and BP company


Experiment_2_Predict_7_days_LSTM VLKAF company
training
actual
forecast
Company price [$]
0
−1
−2
−3
−4
−5
−6
−7
2021-04-13
2021-04-17
2021-04-21
2021-04-25
2021-05-01
2021-05-05
2021-05-09
2021-05-13
2021-05-17
Time periods


Experiment_2_Predict_7_days_LSTM BP company
training
actual
forecast
Company price [$]
2
0
−2
−4
−6
−8
−10
−12
2017-12-09
2017-12-13
2017-12-17
2017-12-21
2017-12-25
2017-12-29
2018-01-01
2018-01-05
2018-01-09
Time periods

5. Summary

MAPE measure for particular models were shown below:

	ARIMA	LSTM
1-DAY	1.64	1.46
7-DAYS	2.52	9.74

As we can see, the best results for 1-day prediction were obtained for the LSTM model and surprisingly for 7-days for the ARIMA model. Our project shows that for long-time predictions statistical methods are better than deep learning methods in case of stock price data. We also agree with Strong Market Hypothesis that states that all information, public and not public, is completely accounted for in current stock prices, and no type of information can give an investor an advantage on the market.

At this point, we also would like to point out that so far the project has taken several dozen hours. Even though we made an attempt to implement the LSTNet (which is a hybrid model with statistical and deep learning methods), we did not have enough time to complete it with sufficient results.

Bibliography

[1] Brinwa Kra, Xing Lu, and Haiyan Yin. The weekend effect in African stock markets. International Journal of Business and Applied Social Science, pages 24–28, 11 2019.

[2] Spyros Makridakis and Michele Hibon. Accuracy of forecasting: An empirical investigation. Journal of the Royal Statistical Society: Series A (General), 142(2):97–125, 1979.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735– 1780, 1997.