**Stock prediction using Machine learning methods**

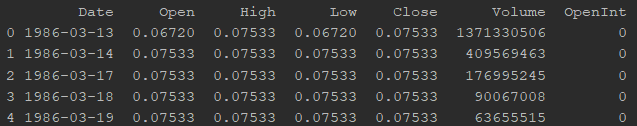
**Miloš Pantić, Mihajlo Jovković, Dejan Šorgić**

1. Motivation

Shares represent a proportional part of the corporation, they allow their owner access to funds and because of that part of the corporation. Investing in successful corporate actions means a constant increase in the value of our work, while on the other hand, investing in a collapse of corruption will result in our financial collapse. As trade in shares is no longer reserved for the elite, a "ordinary" man can also be involved in the game, software that can predict the future success of a certain corporation as well as its possession would be invaluable. That is precisely the goal of our project, to make a program that will predict the prices of its actions, or its success in the future based on the history of the operation of a certain corporation.

2. Research questions

The field of machine learning, or more specifically, deep learning, offers us a wide range of different approaches to our problem. In this project, we will implement several different methods and prediction algorithms in order to finally find the most effective approach for our data set. The data set we will use is downloaded from the Kaggle site [1]

 Figure 1.

In Figure 1, we can see an example of the appearance of a single file in our data set. Columns "Open" and "Close" represent the starting and ending price of the action during the specific day "Date". "High" and "Low" represent the highest and lowest price of the action during that day, while in the column "Volume" we can find the total number of purchased or sold shares during that day. Since in the "Close" column we find the final price of the action for that day, it will be the date and the column "Date" that will be the most important in our project.

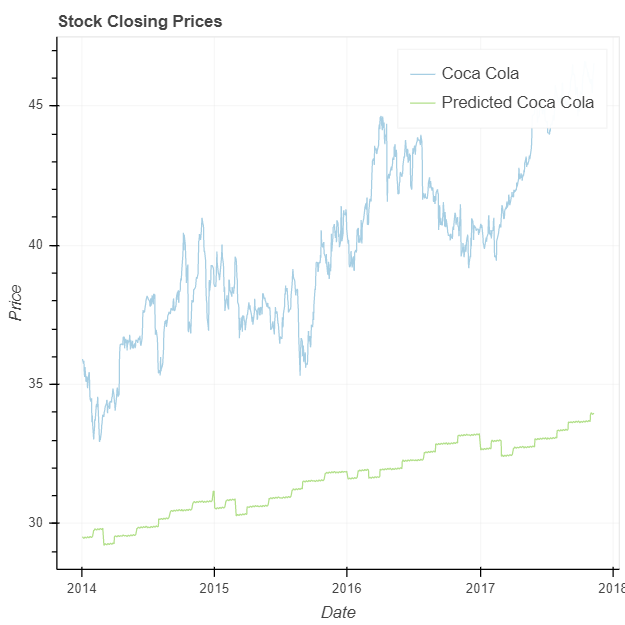
3. Related work

We have found a lot of other works trying to solve this problem, and use a lot of different factors, such as news related to company. Our goal was to make prediction just using prices and dates, and compare different method while doing so.

4. Metodologije

4.1 Linear Regression with multiple variables

Our data are of such a format that we can conclude that the desired solution is influenced by a number of factors, as one of the methodologies we used linear regression with multiple variables. Multiple linear regression represents the case of linear regression, where several independent variables are received as input parameters. It was then necessary to remove the variables that we concluded that they were not needed in the analysis. After pre-processing, using the SKLearn package we calculated the values ​​of "closing" actions representing the requested output of the application. As this method is one of the simpler, it is expected that the RMSE error is greater than the others.

****Figure 2. Predicting Coca Cola prices using Linear regression

4.2 Facebook Prophet

In order to try as many methodologies as possible and to compare as many solutions as possible, we decided to use the FacebookProphet algorithm, which specializes in making predictions that depend on the time periods. This would mean that the algorithm looks at annual changes, seasonal, monthly, and daily. It also observes the so-called effect of holidays on the stretches. The algorithm is resistant to missing data and trend changes, and also handles outliers very well.

4.3 K-Nearest Neighbors (KNN)

Although KNN is, in most cases, used as an algorithm for solving the classification problem, it can also be used for regression. It performs the prediction by taking K's closest neighbors and setting the mean value as a prediction. In our project, during the preprocessing of the data set, we will "break" the column "Date" into several smaller columns that may be relevant to our prediction: "Year", "Month", "Week", "Day", "Dayofweek", "Dayofyear "," Month\_start\_end "," Week\_start\_end ". Then we divide the set of data into the training section and the validation part in relation to 80/20 and scale them. We still have to determine the optimal number of neighbors, K, for this we use GridSearchCV. Now that we have everything we need, we model the model and make a price prediction.

4.4 Auto ARIMA

A utoregressive integrated moving average (ARIMA) is a very popular model especially when it comes to data analysis that changes over a certain period of time. It is used for better understanding of data or predicting future values. Arima is defined by three nonnegative integers p, d, q, where p represents the past values ​​used to predict the future. D represents the degree of differentiation (the number of times the data has been taken in the past), while q represents past errors in the prediction.

Determining the three parameters of the arima, can take a long time. That's why in our project we use auto arima, which automatically selects the combination of parameters, which will return the best result, with the smallest error.

Although the fitting process of the Arima model is the longest, it achieves the most convincing results.

  
Figure 3. Predicting Google prices using Auto ARIMA

5. Discussion

We have tested each algorithm with several different data sets, for different companies, and examined how well the predictions are made, and how well do these methods follow different price trends. The most success, out of used methods, had the Auto ARIMA method.

6. References

[1] <https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs?fbclid=IwAR3NhhZpOl_d9HwFPUSDp7vh2j8QqY6oF-CJ3J7BR1vGcdXhRXacf5Wmol0>