Machine Learning Engineer Nanodegree
Capstone Proposal

Guilherme A. Camara

Domain Background

The future value of a company stock over a period might be predictable, and the technique developed to do so is known as *Stock Market prediction*[1]. Given some metrics over a period, such as opening stock price, volume of stocks traded and highest stock price traded, predictions can be done using an algorithm. The analyses of all possible inputs are done in order to predict the Adjusted Close stock price, in other words, the future value of the company.

Accurate future value prediction is an objective pursued by many companies. This project will not be sensible to all data available, trying to perform data-mining in twitter and other tools to predict if a company will have a drastically rise/fall on their stock price. I intend to investigate the financial market, understanding how it works and the variables responsible of its changes, not to create the best predictive tool on the market.

Problem Statement

Some tools, available on-line and with no costs, provide a significant amount of information about the stock market. Thus, historical information can be downloaded about it in order to analyze how the stock price of a company behaved over the past years. All the data available is based on dates and prices, and is therefore quantifiable, measurable and replicable, given that data from the past can be analyzed. To sum up: The problem stated is to determine the *adjusted close stock price*, and it will be done through an analysis of historical and recent data from stock values.

Datasets and Inputs

All data will be achieved on-line by a free open source. Firstly, the *Yahoo! Finance* tool will be used. There is a module available for python to get the data on https://pypi.python.org/pypi/yahoo-finance . There is a list of available methods implemented by the *Yahoo! Finance* tool that will provide all the inputs needed for this project, at least in the initial part of the project. If the algorithm requires some new data, it will be described on the final report.

The data achieved from the *Yahoo! Finance* will be stored in a local database system, such as *sqLite*, to prevent downloading always the same data and in order to analyze the data off-line. The inputs that will be used at first are: *Volume* stock trading, *Open* stock value, *Low* stock price, *High* stock price, *Close* stock price and the *Adjusted Close* stock price.

Solution Statement

As listed in *Datasets and Inputs*, the solution is to analyze all the input data, which are 5, and by their behavior predict the *Adjusted Close* price. The challenge here might be to find the ideal period to perform the analysis, since there are many variables that might affect its behavior, such as a Market crisis. One solution to this could be performing the analyzes over a short period, as a couple of months, of a stable company, such as Apple. A supervised learning algorithm will be applied in order to achieve the result. Since the output of this solution will be continuous, a linear regression algorithm will be used. The algorithms available achieve a better accuracy depending of the size of data used and number of features. Some algorithms that will be implemented in a first time, in order to evaluate its performance are: *SGD Regressor, SVR(multiple kernels) and Ensemble Regressors.* As the solution becomes more defined and clear, some solutions with deep learning might be studied and implemented.

Benchmark Model

The result can be easily evaluated, since it will be quantifiable and compared to the already existing value. Therefore, the evaluation of the success of the model can be done by comparing the final predicted value of the *Adjusted Stock price* with the real one. The benchmark value, considering the $R^2 score$ (next topic) to this project, that will be pursed initially will be of 0.5. However, a value of 0.5 to predict stock prices does not seems to be reliable, since you cannot evaluate if the value of the stock is rising or decreasing over a period. Therefore, a reliable benchmark value to achieve when finishing this project would be around 0.8. With this value, the client using this software can interpret the values and decide if it is a good scenario or not to keep/buy/sell stocks from this company.

Evaluation Metrics

The evaluation will compare the stock price estimated with the one achieved. Therefore, an evaluation metric as $R^2 Score$ should be a good fit to this. The $R^2 score$ is a statistical measure of the closeness between the data and the fitted regression line [2]. The result will be between 0 and 1, where 0 indicating that the model doesn't explain the data, and 1 indicating that the model completely explains the data.

Project Design

This project is designed to be developed in three parts. The first one consists of an algorithm responsible of archiving the data, given a company name and a period, and will store all the data in a database. This algorithm is essential to future improvements on the code, such as adding new features. Organizing it in a database, such as sqLite, will improve the performance of the application, allowing it to avoid archiving the data every time. Also, storing data in a database is better than doing so in a file, since it is easily scalable and accessible. Therefore, it doesn't require lots of parsings over files.

The second is an algorithm responsible of getting the data from the database and applying the selected predictive model. All the analyses will be done in this code. This algorithm will be coded in Python 2.7, and it is intended to use *TensorFlow* as a tool to create the model. Other tools, such as *seaborn* will be used to analyze the results and write the report. The use of the *TensorFlow* instead of *sk-learn* is to learn a new tool with new solutions and to try to implement deep learning, after achieving the benchmark of 0.5.

Then, an interface to the user is intended to be provided. Depending on the time available after the implementation, a web framework could be used with some solution with python's flask or Angular.js.

Finally, the planned workflow intended to this project is the following:

- Determine the tables' format of the database.
- Determine the period and companies that will be used to begin this project.
- Code the algorithm responsible to download the data and store it in the database.
- Analyze the data, understand it and find the principal components and/or the proper model.
- Evaluate the results, plot it and start the report.
- Develop an user interface.
- Finalize the report

References

[1] Wikipedia, Available from World Wide Web at (https://en.wikipedia.org/wiki/Stock_market_prediction), on 15/01/2017

[2] Minitab, Available from World Wide Web at (http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit), on 15/01/2017.