PREDICTION ANALYSIS FOR US STOCK MARKET

Petr Šild

Abstract

The goal of this paper is to use prediction analysis for managing the risk of the selected portfolio of US shares, especially to prevent massive loses during unfavourable situation on stock market and otherwise to maximize the profit on bull market. The aim is not to predict which share would be the best to buy, but to predict the sector which would be the best to hold and in what proportion of the entire portfolio. As we can experience in recent days a coronavirus crisis, we will try to learn model from past crisis and then try to evaluate, if the model managed to avoid massive loses during crisis. For this purpose was selected US stock market because there is the longest series of data available. For the realization of the prediction analysis was selected statistical method of logistic regression and the more advanced methods like decision trees, random forest and neural nets. The suitable programming language for prediction analysis is Python, therefore, it was also used in the analyses presented in the article.

Key words: Modelling, Python, Investments, Prediction analysis

JEL Code: C53, C58, G11

Introduction

In this article reader is going to be acquainted with three quite popular topics in these times and one topic which follows a humanity for centuries.

First topic is data analytics and data science. Everybody had a chance to meet it directly or indirectly thanks to companies, which have an access to huge volume of data and they are trying to find a way how to use it effectively. Therefore also teams of data analysts and data scientists are one of the most rapidly growing teams. The next topic is robotics and automatics which is again topic which is often mentioned. In this time of coronavirus crisis it could be driver how companies can reduce their costs. The last of three popular topics is data visualization. Thanks to visualization it is much easier to understand data and it possible to find patterns which people can’t see just from tables.

Last but not least it is necessary to mention a topic of investments. We can remember a famous quote of great Isac Newton: „I can calculate the motion of heavenly bodies, but not the madness of people.“

So the main purpose of this article is to automatically collect data by robots, store them in database, then create a model which purpose is to predict combinations of sectors which are best to hold or predict, when the best option is to sell portfolio and hold cash only. The last step is to visualize outputs to make easier to read recommendations of model. Recently we can see quite a big fall of all stocks due to coronavirus crisis. In time series which are available we can see two major crisis (2000 internet crisis and 2008 great recession), so the question is: is it possible to teach model from past crisis to predict future crisis (specifically this recent crisis)?

1 Approach

Despite the length of time series, number of companies and volume of data we want to use, it is necessary to use more robust tools than only Excel. For this purpose, the storage of data will be SQL database. But the first step is to collect data, which will not be so difficult in case of macroeconomic data, but it would be worse in case of 505 companies which are in index S&P500, which we used because of availability of macroeconomic data of USA and of data about companies. So for this purpose we used RPA platform UI Path, where we configure a robots which download current data and load them into the SQL database.

The most important part is model by itself. We will create it Python (Stewart, John M, 2014), because it is more suitable language for modelling and it facilitate very good compatibility with another programming languages, so it makes no problems to connect model to the SQL database.

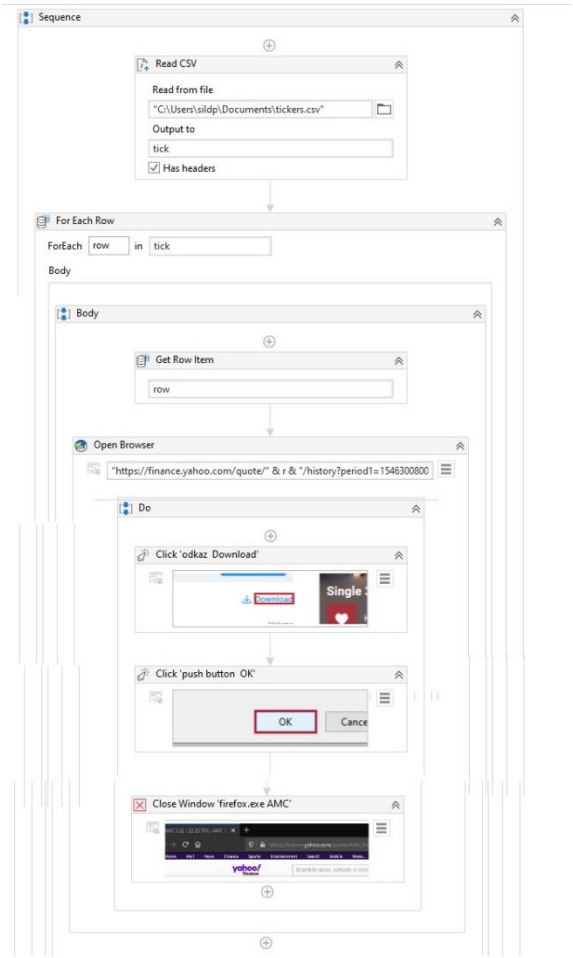
1.1 Collection of data

Market data

First thing, which is necessary to do, is to obtain a list of tickers of 505 companies which compose index S&P 500. Then we created an algorithm, which load those tickers and one by one put it into the URL: "https://finance.yahoo.com/quote/"&row&"/history?period1=473382000&period2=1554933600&interval=1d&filter=history&frequency=1d". Then it open the browser, open this page and

downloads trade data about company. Algorithm in environment of UI Path can looks like this:

Fig. 1: **UI Path algorithm**



Source: Own processing in UI Path

Macroeconomic data

We can get macroeconomic data by similar algorithm, but it is not necessary to repeat that cycle 505 times, but for each indicator just once. Specifically we use Monetary Aggregate M2 (Federal Reserve Bank of St. Louis, 2020), Consumer Confidence Index (CCI) (Leading indicators, 2020) and interest rate defined by yield curves of US bonds (U.S. Department of the Treasury, 2020).

Other data

Another data which are suitable to use are trade data of S&P 500 (Slick Charts, 2020) and it's percentage distribution to sectors, volume of consumer and commercial loans (Federal Reserve Bank of St. Louis, 2020) and index of gold price (Goldhub, 2020).

1.2 Data warehouse

The next step is to upload data into the database. We use an environment of Microsoft SQL Server Management Studio. Need to mention, that before uploading trading data about 505 companies, it is necessary to merge them to one file and create new variable, which define company to which are data related.

1.3 Data preparation

Despite the fact, that data are not in raw state and quite clean, this step will not take so much time, than it is usual. Necessary is, that we need to interpolate data, which are reported on another, than daily basis, for example volume of consumer and commercial loans or consumer confidence index.

1.4 Designing the model

Level of individual stocks

Model is divided into the three levels. The lowest is level of individual stocks and here we perform technical analysis. Specifically we use Bollinger bands, Relative strength index (RSI), Moving average convergence divergence (MACD), Stochastic and Money flow index (MFI). (Márton, P., Adamko, N., 2011)

First, we have to create auxiliary indicators for each stock, like moving averages, exponential moving averages etc. and then identify buy and sell signals which determine when those stocks will be held and when they should be sold. Then we have to create binomial target variable which says from known data, if it is conveniently to hold the stock, or not. Because it is undesirable to change holding/selling stock very often it was set frontier of change of ± 5 % stock value in the next month.

Then we use logistic regression (Meerschaert, Mark M., 2013). where input is results of six methods of technical analysis (0/1 variable saying if I should hold stock or not) and target variable is again 0/1 variable which says if it was really convenient to hold/not hold that stock. As a result we have model which says how to treat results of our five methods of technical analysis to get the best result of profit. Outputs of the model, <0;1> variable, we aggregate to sectors and we can move to the next level.

Level of sectors

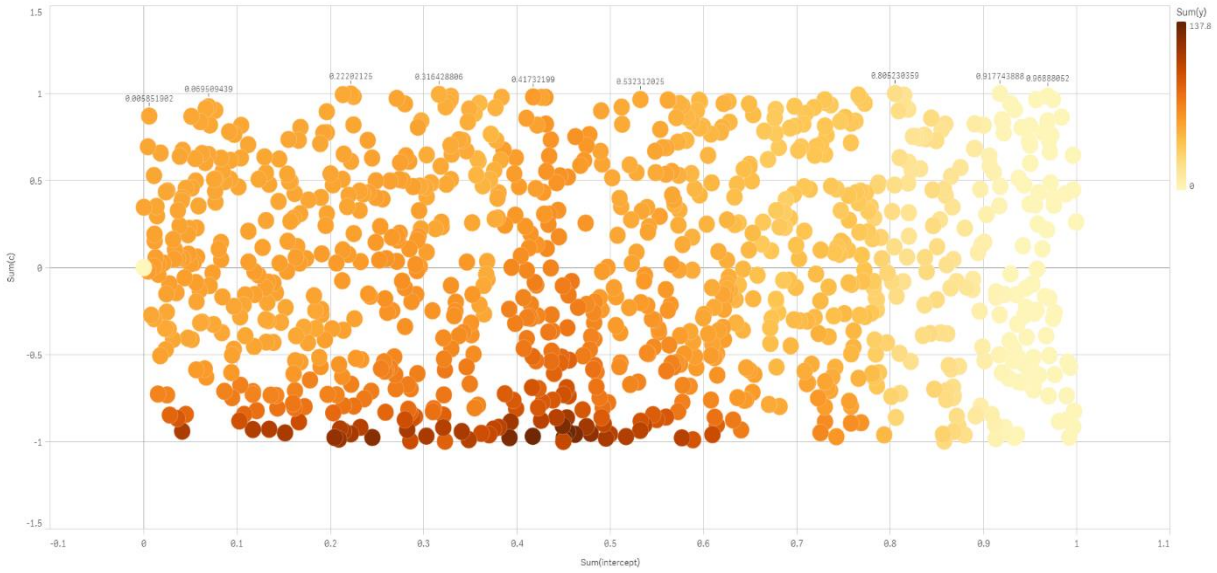
Level of sectors is similar like previous level, but we have to make few changes. Target variable is going to be made from yield of whole sectors which we can get by weighted average

of yields of stocks by their percentage representation in index S&P 500. Then we make sector model where we use our macroeconomic data and outputs from previous model as predictors. The question is which method would be best fit to our purpose? So we tried 3 methods: decision trees, random forests and neural nets and then we decided which method would be best. From the results, the best method has been shown for this purpose the random forests. It is necessary to add, that we have to split our data to training and testing sample. As was said in the beginning, we want to test if our model can learn from previous crisis to predict our recent coronavirus crisis, so our learn sample is our whole dataset except last year, and out test sample contains our data from the last year.

Level of whole market

This is the final level, where we have to put together outputs from the previous level and transform probabilities to benefit from holding the sector to portfolio distribution of sectors. For this purpose we used Monte Carlo simulation to catch the cases, when probabilities of sectors are very low and it is best to hold only cash. Another usable transformation could be adding a constant to those probabilities <0;1> interval, to determine how much our final model can rely on our previous model. If the constant will be 1, the model will be close to naïve diversification. Otherwise, if constant will be -1, then our final model strongly rely on outputs from previous model because it hardly penalize low probabilities. Determining factor will be profit we can get with consideration of our constants.

Fig. 2: Constant visualization



Source: own processing in Qlik Sense

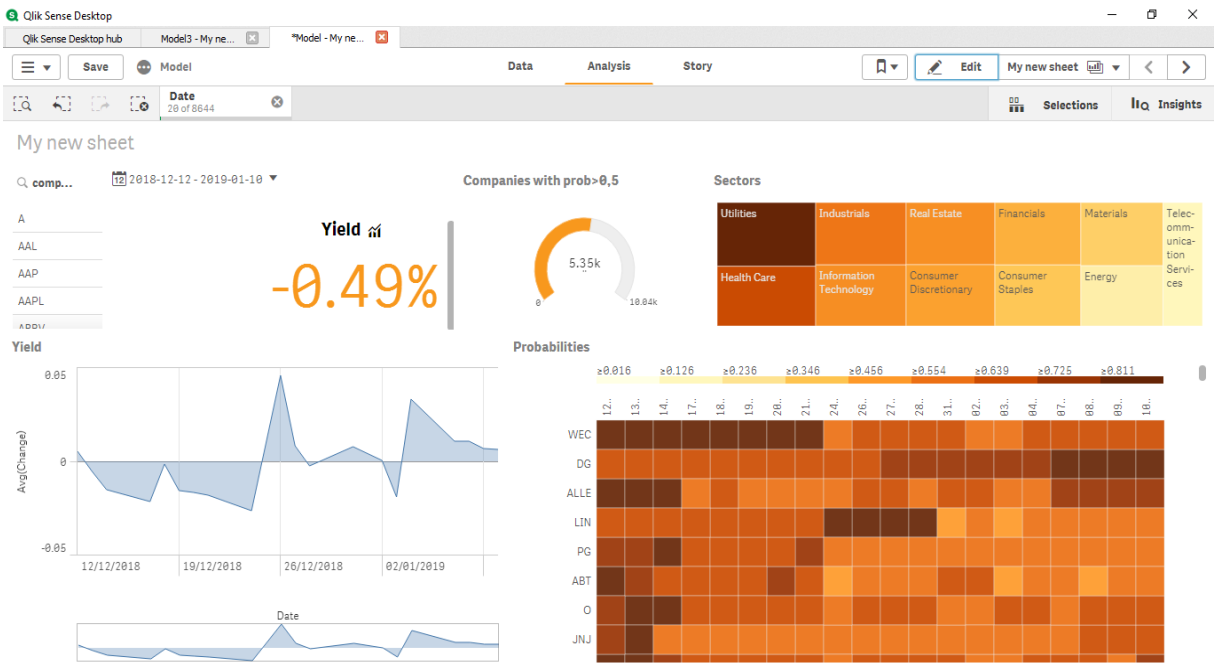
On this chart let's take a look on dark points which shows higher profit. As we can see, the frontier where we should sell stocks and hold only cash is about 0,45 and constant we should add to those probabilities is -1 (of course there is limitation, that probability can't be less than 0).

1.5 Visualization

The last, but very important thing is visualization of the whole model. For this purpose we used Qlik Sense, which is one of the best three and mostly used BI tools.

Level of individual stocks

Fig. 3: Level of individual stocks - visualization

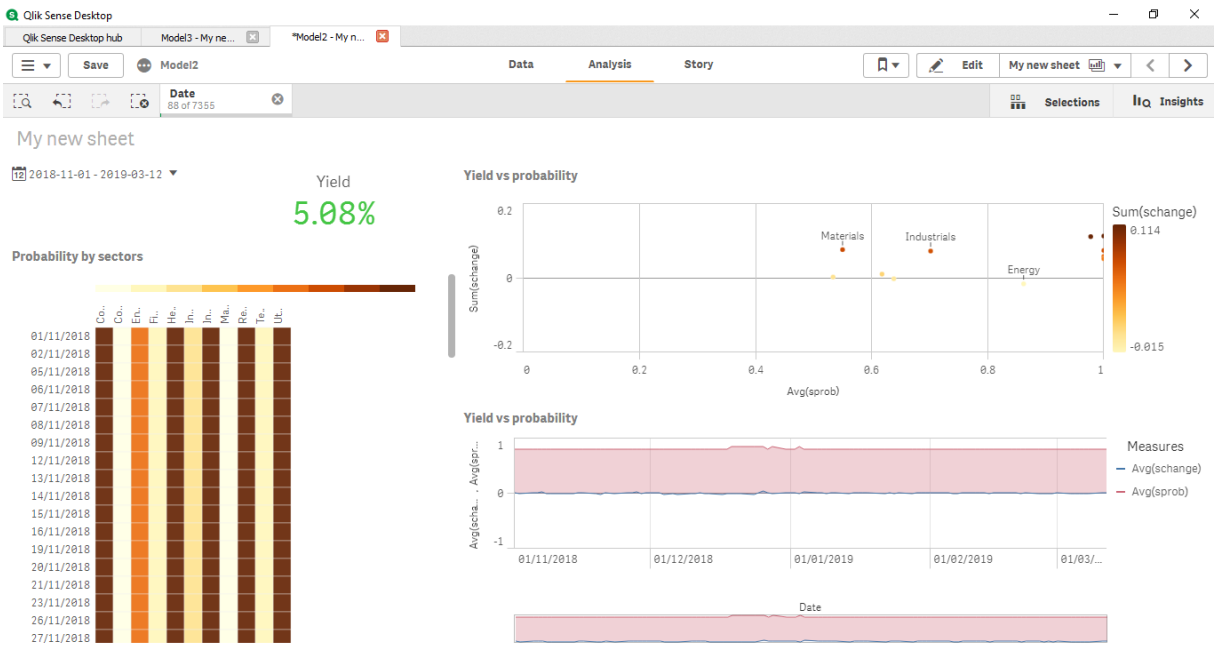


Source: own processing in Qlik Sense

In a right bottom corner you can see a heatmap, where is by shade visualized probability from the first level model, so from technical analysis of each stock. When you click on any name of stock, you can place a filter and whole dashboard is related only to that one stock. Then in a right top corner, you can see average probabilities by sectors. Then you can see daily changes od stock price and yield for the selected period.

Level of sectors

Fig. 4: Level of sectors - visualization



Source: own processing in Qlik Sense

On this dashboard you can see outputs from sector model. You can see here heatmap again, but instead of individual stocks, you can see whole sectors. The new element here is scatter chart, where you can see dependency between probability and yield.

Level of whole market

Fig. 5: **Level of whole stock market - visualization**

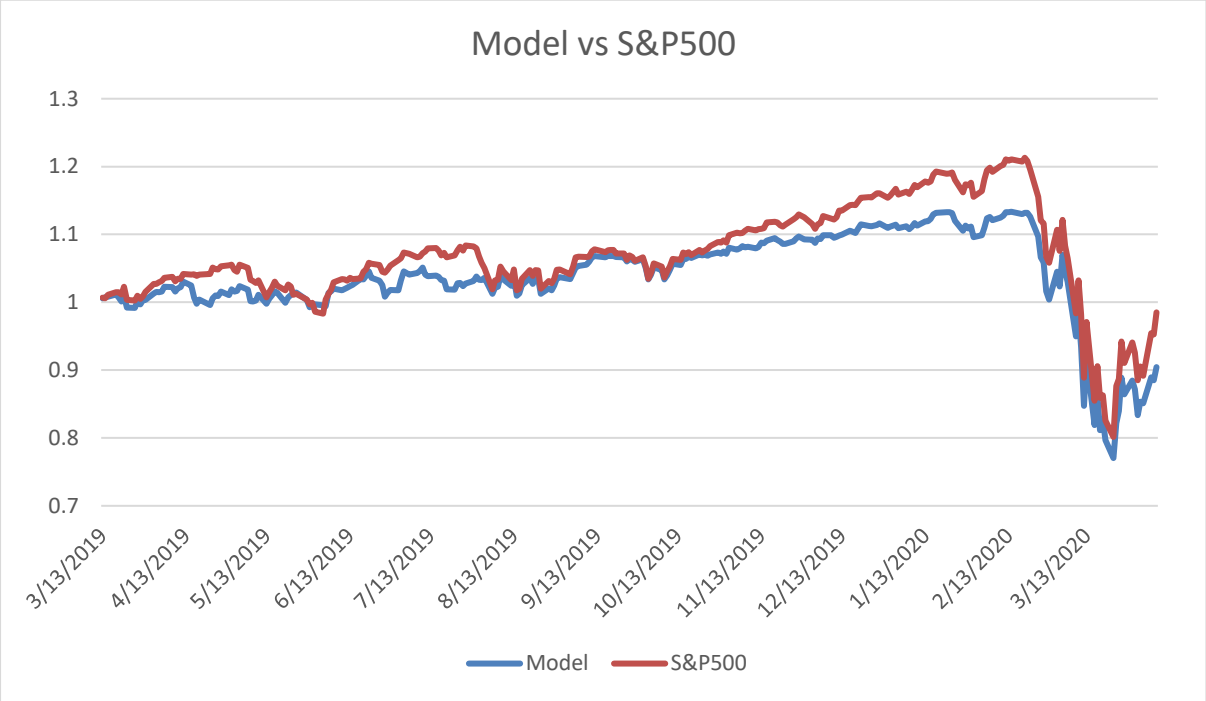
Source: own processing in Qlik Sense

In the end, we have model of whole market, where we can see pie chart, which defines diversification of our portfolio to sectors. Of course, there needs to be a comparison with a benchmark of the model, index S&P 500.

Conclusion

In conclusion it is necessary to evaluate how this model performed in the time of coronavirus crisis. In a chart below there is comparison between index S&P 500 and our portfolio which is output from our model.

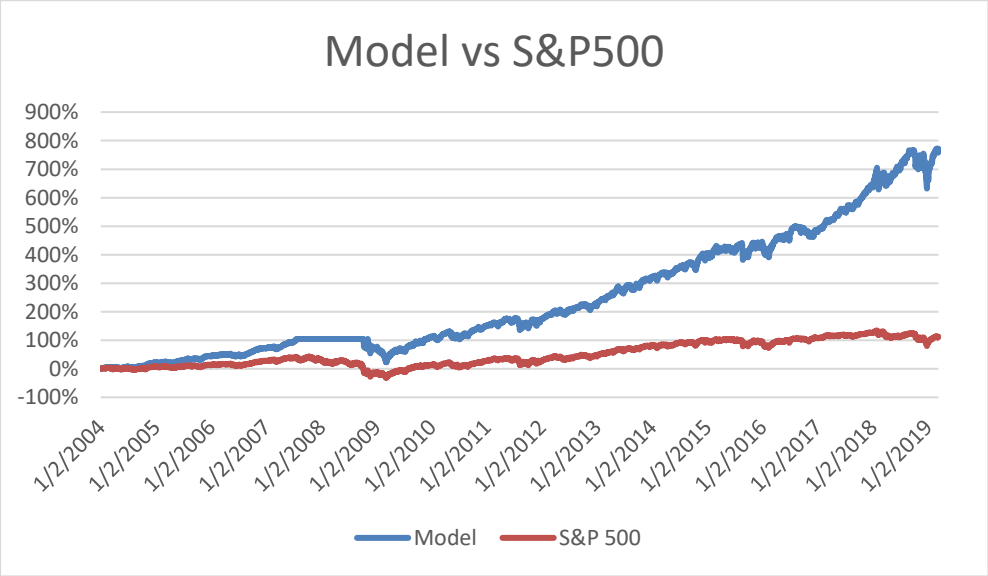
Fig. 6: **Comparison between model and benchmark - crisis**



Source: own processing

Unfortunately, even if we took quite a lots of data, tried more advanced predictive methods, our model didn't performed better, than his benchmark in the time of coronavirus crisis. As a small solace could be fact, that in past the model outperformed his benchmark by a lot.

Fig. 7: Comparison between model and benchmark – before crisis



Source: own processing

The reason, why we didn't manage to create model, which would predict the crisis could be, that it is simply impossible, because no one, could have predict that crisis. Maybe in the future some skillful team could create so advanced predictive model and maybe some team even have already created one, but I am afraid not.

Acknowledgment

This work was supported by SGS project University of Pardubice, Faculty of Economics and Administration, No. SGS_ 2020_016.

References

Leading indicators - Consumer confidence index (CCI) - OECD Data. (2020). Retrieved May 10, 2020, from https://data.oecd.org/leadind/consumer-confidence-index-cci.htm

Goldhub - Gold prices (2020). Retrieved May 10, 2020, from: https://www.gold.org/goldhub/data/gold-prices

Federal Reserve Bank of St. Louis - M2 Money Stock. (2020). Retrieved May 10, 2020, from: https://fred.stlouisfed.org/series/M2

Federal Reserve Bank of St. Louis - Commercial and Industrial Loans, All Commercial Banks. (2020). Retrieved May 10, 2020, from: https://fred.stlouisfed.org/series/BUSLOANS

Slick Charts - S&P 500 Companies by Weight. (2020). Retrieved May 10, 2020, from: https://www.slickcharts.com/sp500

U.S. Department of the Treasury - Daily Treasury Yield Curve Rates. (2020). Retrieved May 10, 2020, from:https://www.treasury.gov/resource-center/data-chartcenter/interestrates/pages/TextView.aspx?data=yieldAll

Márton, P. & Adamko, N. (2011). Praktický úvod do modelovania a simulácie, EDIS - vydavatelstvo ŽU, Žilina, ISBN 978-80-554-0387-8

Meerschaert, M. M. (2013) Mathematical modeling. 4th ed. Waltham: Academic Press, Cambridge, Massachusetts, USA, ISBN 978-0-12-386912-8

Stewart, J. M. (2014). Python for scientists. Cambridge: Cambridge University Press, ISBN 978-1-107-68642-7

Gogola, J. & Šild, P. (2019). Robotic process automation for investment modelling. *European Financial Systems: Proceedings of the 16th International Scientific Conference*. 126-132.

Contact

Ing. Petr Šild
University of Pardubice
Faculty of Economics and Administration
Studentská 84
532 10 Pardubice 2
Czech Republic
st44571@student.upce.cz