Robotic process automation for investment modelling

Ján Gogola, Petr Šild

University of Pardubice
Faculty of Economics and Administration,
Institute of Mathematics and Quantitative Methods
Address: University of Pardubice
Faculty of Economics and Administration
Studentská 84, 532 10 Pardubice
E-mail: jan.gogola@upce.cz, st44571@student.upce.cz, sild.petr@gmail.com

Abstract: *The developments in data science, machine learning and artificial intelligence force us to revisit the question "What should be automated and what should be done by humans?" The main objective of our contribution is to apply Robotic Process Automation (RPA) to create a model which identifies risk situations on the model based on the market prices and external data such as M2 Money Stock, Consumer Confidence Index (CCI), Daily Treasury Yield Curve etc., and recommend a proportion of assets in the portfolio. Our goal is to build a model that will beat its benchmark, the S&P 500 index, for that purpose we create a portfolio composed of individual stock titles contained in the S&P 500 index and compare the model rate of return with the real rate of the S&P 500 for the period from 1.1. 2004 to 1.1. 2019. As a result we can show that the cumulative yield of the model beats its benchmark approx. 7 times during the period under review.*

Key words: robotic process automation (RPA), theory of portfolio, data mining,

JEL Classifications: C55, C61, G11

1 Introduction

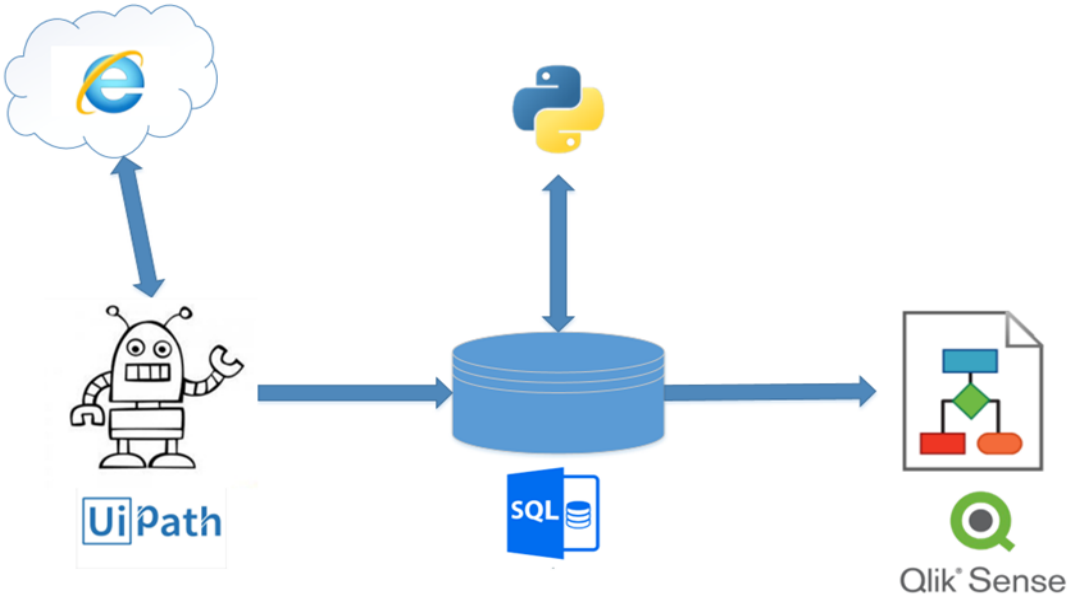
The main goal of this paper is to apply Robotic Process Automation (RPA) to create a model that determines the ratio of assets that it is appropriate to hold in the portfolio. The development in data science, machine learning and Artificial Intelligence (AI) force us to ask the question: „What should be automated and what should be done by human?"
RPA is one of these developments. It is a tool that operate on the user interface of other computer systems in the way a human would do.
When robot inclusion occurred in manufacturer industry, empowering factories with robots that are more capable, reliable, and with 24–7 working capacity. (Willcocks, L. P., Lacity, M., & Craig, A., 2015)
What could be robotized? Simply put, any process that could be documented, which can be considered as being repetitive. In a more technical language, any process that can grab or introduce data via a desktop application or web page could be robotized, as well as manipulation of data, persistence in excel worksheets or interactions with 3rd party systems and emailing. (van der Aalst, W. M., Bichler, M., Heinzl, A., 2018).
RPA aims to replace human by automation done in an „outside-in" manner. RPA provides agents that interact with different information systems thus interact partly replacing humans. Using Artificial Intelligence (AI) and Machine learning (ML), this can be done in a fairly robust manner. There are many vendors offering RPA tools we decided to use UiPath.
UiPath is a global software company that develops a platform for Robotic Process Automation (RPA), a pretty cool concept that intends to automate repetitive tasks made by humans allowing them to focus on work that requires intelligence and judgment.
UI Path comes with open source version for medium business and a more complete enterprise solution with server software for multiple robots execution in background.
Consider the fact that in most cases, actively managed funds are lagging behind their benchmark our goal is to build a model that will beat its benchmark, the S&P 500 index.

2 Methodology and Data

Since the types of assets are innumerable, let's define the individual sectors of the US economy. Therefore, the model will model how to have the most preferably distributed portfolio across sectors. We can then imagine the assets either as funds that replicate the development of the sector or purchased individual titles from the S&P 500. Stocks are selected from the US market as the US market is the most effective market and have the longest available relevant data series. The ratio of stocks in the portfolio will determine the risk shareholding due to overbought market or overvaluation of stocks. We will then model this risk from two perspectives: the aggregation of stock titles and the investment horizon. As part of the aggregation of stock titles, we will look at the market as a whole, the individual sectors and further to the level of individual stock titles contained in the S&P 500 index. If someone is looking for a very complex approach to analyzing the stock, there is a fundamental analysis for him. (Fanta, J. 2001) Indeed, fundamental analysis has a great deal of focus, not just purely corporate factors such as debt, historical gains, dividends, profitability, or liquidity. In addition to these factors, it also examines global factors that affect the market as a whole or sectoral factors that affect a particular industry in which it operates. Therefore, our approach is to model at three levels: stock, sector and market.

Since our model will use a large amount of data and the model will need to be updated on a daily basis, it would be very time-consuming to do this manually. (Petr, P., 2014) and (Dietrich, D. (Ed.)., 2015). For this purpose, Ui Path robots will be configured to retrieve, download, and load the required data into the model into. The entire system will look like in the Figure 1.

Figure 1 Scheme of our model



Source: Own processing

First of all, you needed to get a list of 505 tickers of companies that make up the S&P 500, and save that list in a csv file. The robot then loads the file and downloads company data one by one. The source is the finance.yahoo.com. (Yahoo Finance, 2019) By a similar algorithm, as we gained market data, we can also obtained macroeconomic data such as M2 Money Stock, Consumer Confidence Index (CCI), Daily Treasury Yield Curve, Commercial and Industrial Loans, Gold prices.

The SQL database was chosen as data storage and analysis will be performed in Python programming (Stewart, John M, 2014) environment using Numpy, Pandas and SciPy modules. The outputs will then be visualized into the dashboard using Qlik Sense for better clarity.

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. NumPy is open-source software and has many contributors. Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software.

- **Model of individual stocks**

Since the linking of macroeconomic factors directly to an individual stock is relatively weak, it does not make sense to include these indicators in the stock model. We will try to predict the profitability of holding stocks through purchase and sale signals of technical analysis. (Márton, P., Adamko, N., 2011) and (Meerschaert, Mark M., 2013). For this purpose, we were chosen 5 methods of technical analysis: Bollinger bands, Relative Strength Index, MACD, Stochastic Oscilator and Money Flow Index.

So how do we define the profitability of holding a stock? As the methods of technical analysis generally focus on a short period of time, we will model whether a month later we will realize a return when holding a stock. In the first step, it is necessary to create auxiliary indicators such as moving averages, exponential moving averages and others, and then identify the purchasing and sales signals for the entire time series for each company. Then determine the sales and purchasing signals and decide when to hold the stock. Our modelled variable will take values 0 and 1, depending on whether the shares are worth or not worth to keep. The modeled variable is obtained by twenty-one day moving average of daily changes. The result is a logistic regression model that, based on technical analysis indicators, models the likelihood that each of the 505 stocks of the S&P 500 index is beneficial to keep.

- **Sector model**

The influence of macroeconomic indicators can already be reflected in the development of the share price of the entire sector, so it makes sense to include it here. Besides them, we will use the knowledge from the technical analysis of the previous model and include its output among the input variables. But it is necessary to first obtain a modeled variable, i.e. a yield by sector. So we need a weighted average of corporate earnings by their weightings in the S&P 500 index within each sector. The output of this model is a probability vector, which determines the likelihood that the acquisition of the values is preferable to keep the sector index.

- **Model of the entire market**

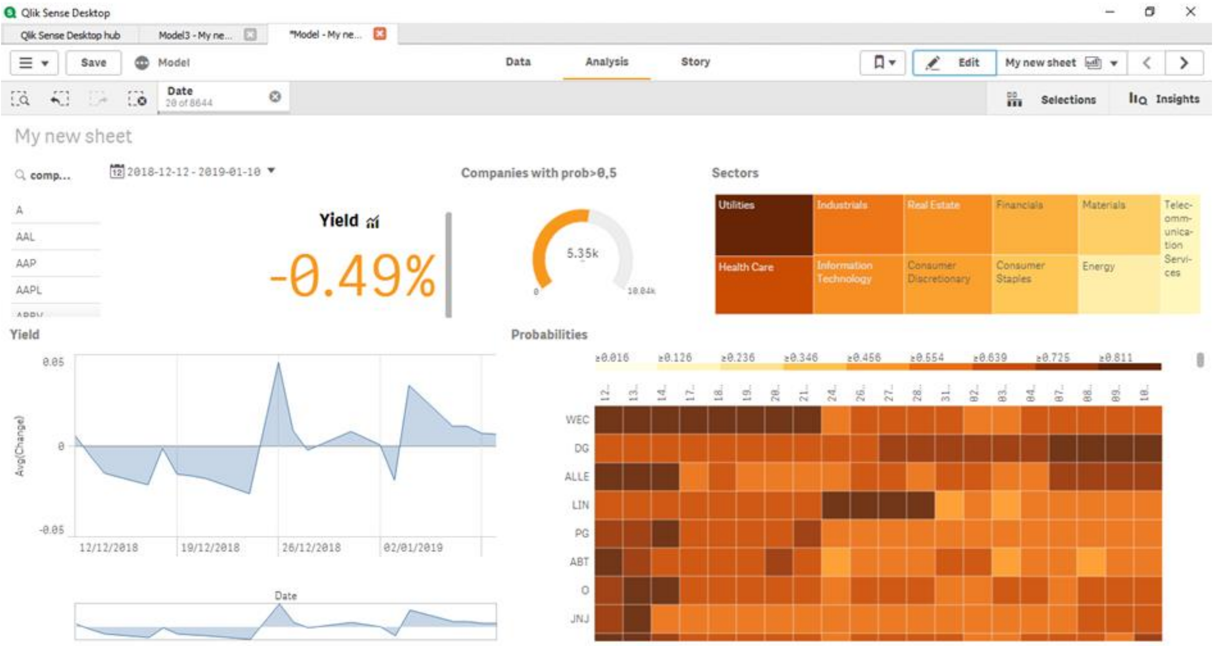
This model is based on the results of the previous model, which models the probability of profitability of holding the sector's shares by day. We transform these probabilities into portfolio portions.

So we have a model, in three levels, that is bound together and at the end of which is the composition of the investor's portfolio. The investor can further analyze individual risks within sectors, markets or directly at the level of individual stocks.

3 Results

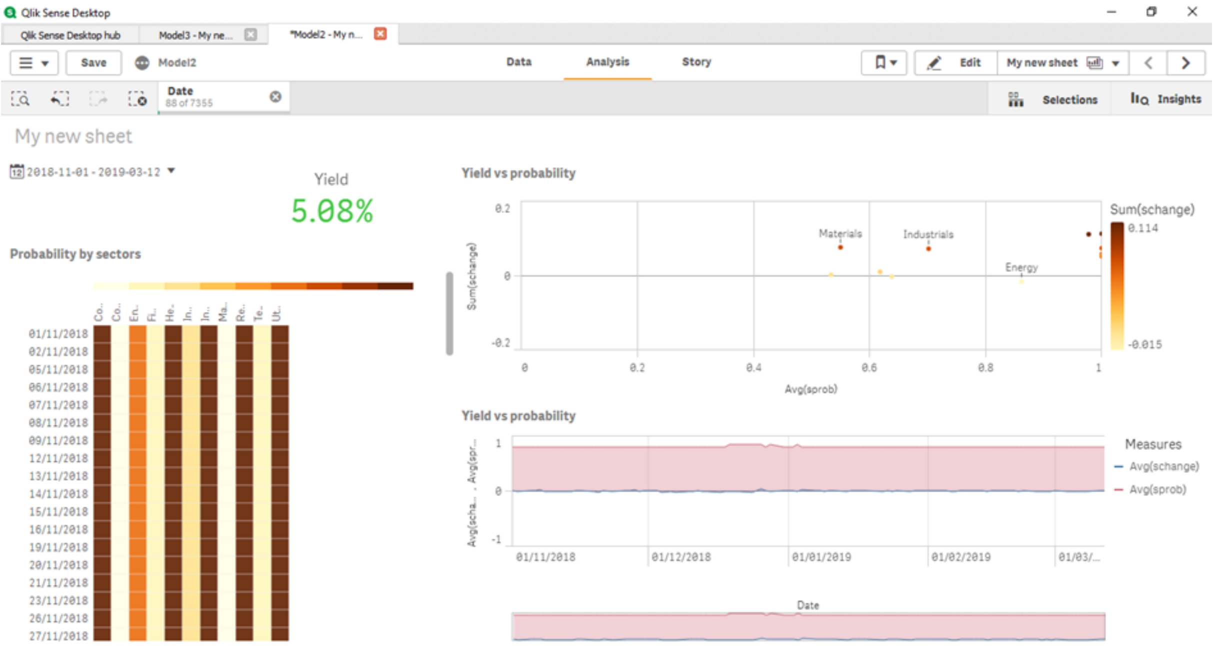
We will visualize our results in Qlik Sense. Figure 2 shows a visualization of the stock model. At the bottom right, we see a heat map where the color of the paint shows for each stocks the probability of holding them for the selected period. These titles can either be manually scrolled or filtered using the filter at the top left. Another important filter is next to it, which is used to select the time period. Bottom left we can see the development of average daily changes for all stock titles. In addition to filters, we can find revenue or loss for the selected time period. The last indicator is the division of sectors where the size and shade of the field indicate the average probability. The advantage is that almost every object can filter by clicking. For example, if we click on the health care sector, we will only see data for health care companies.

Figure 2 Result visualisation - Stock model



Source: Own processing in Qlik Sense

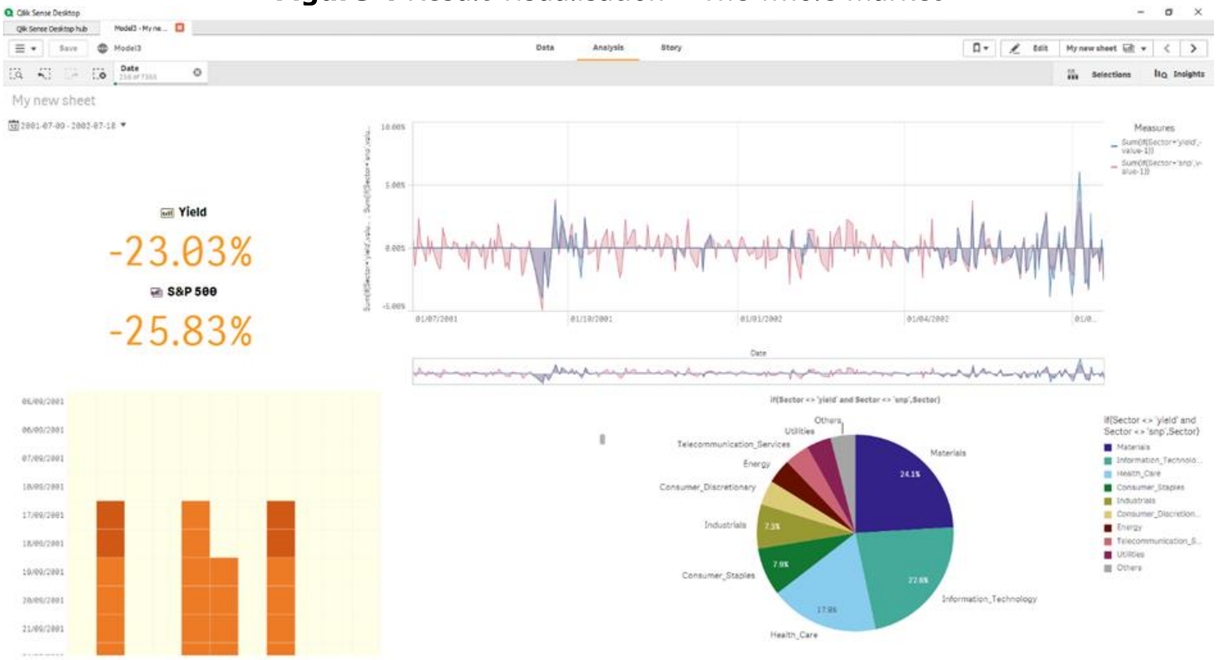
Figure 3 Result visualisation - Sector model



Source: Own processing in Qlik Sense

Figure 3 contains outputs from the sector model. On the bottom left there is a heat map again, but the stock titles have replaced the sectors here. Again, we have a yield and filter period.

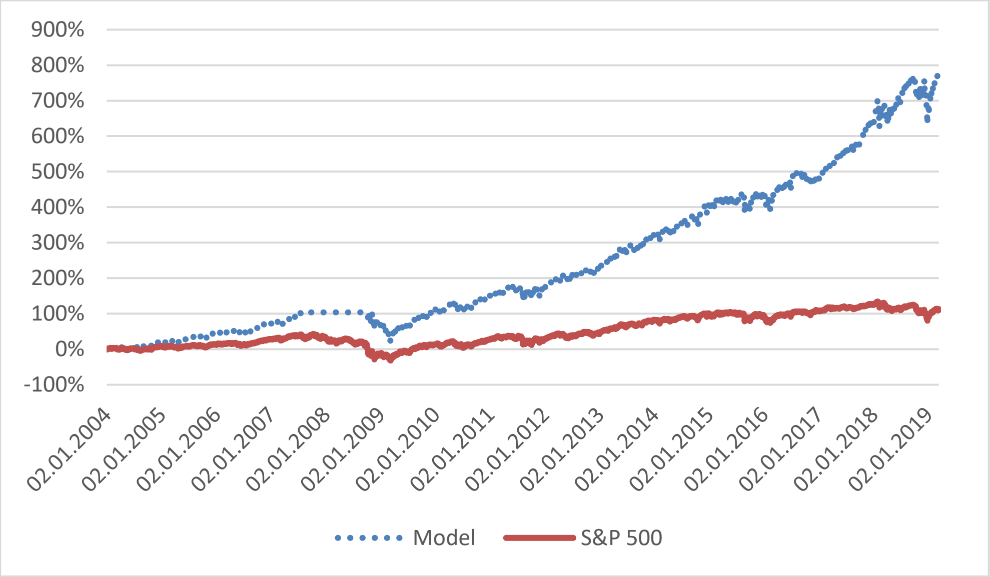
Figure 4 Result visualisation – The whole market



Source: Own processing in Qlik Sense

The final result is a model that determines the proportion of assets in the portfolio. The new feature is the pie chart. This chart makes the most sense, if one day is selected, then it shows the exact composition for that day. Part of the output is a comparison of the model with its benchmark, i.e. the S&P 500 index, both in graphical form and in the form of yield for the given period. The particular period here is chosen for the course of 2001, so we can notice a very bright heat map and very negative profits because of the technological crisis. Finally, the remaining question is whether our model beats its benchmark, ie the S&P 500. As can be seen in Figure 5 - the cumulative yield, the model really beat its benchmark.
It is also important to take into account the risk of this yield. Although the risk (shown in Figure 6) varies considerably over time, it cannot be claimed to be significantly higher than the S&P 500.

Figure 5 Cumulative yield since 2004



Source: Own processing

4 Conclusions

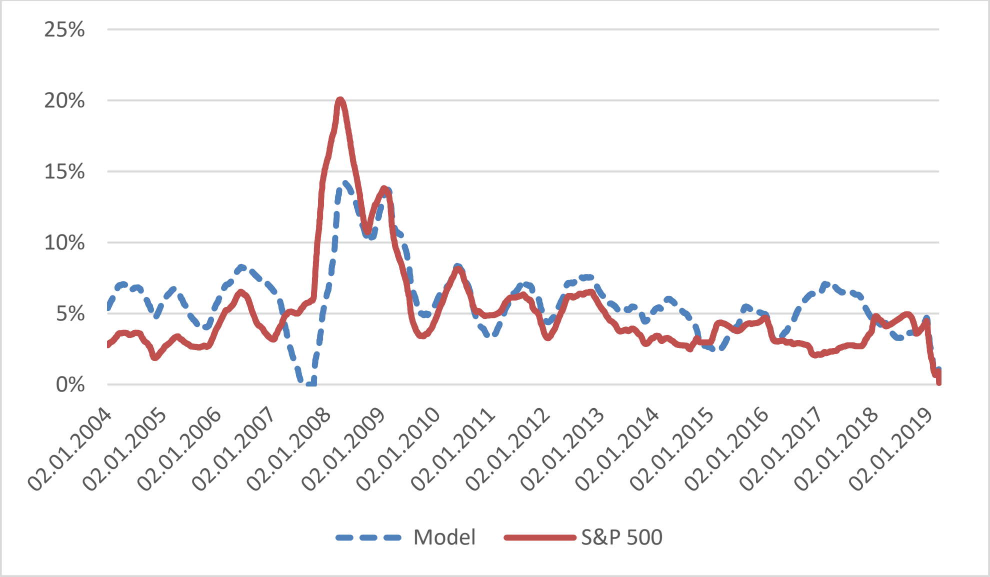
Our conclusion contains more open questions than information about our findings. Most of the times, the effort to robotize every client processes could be significantly higher than the potential gains to him, so this topic requires negotiation. Another topic that must be addressed during a negotiation phase is related with the criticality of tasks made by the robot: what happens if the robot fails? What is the client business impact? Robots are accurate but the systems that they deal with aren't. Imagine that the robot is extracting data from a page and that page server fails exactly in the moment the robot is executing. You could spend time making your robot more robust and error prone, but that represents a cost. The use of RPA brings many interesting questions. How to control RPA agents and avoid security, compliance and economic risk? Who is responsible when RPA agent ”misbehave”?

As part of data preparation, it is necessary to add that when modeling future development, it is necessary to take into account the time delay with which we obtain the data. In the case of the market, it is only one day, but in the case of the consumer expectation index it can be up to two months.

There is also a need for awareness of model risk.

Last but not least, it should be added that, although technology permeates the world of finance increasingly, the human factor cannot be completely replaced. There will always be a need for experts to make decisions based on their own experience and on the recommendations of the system.

Figure 6 Comparison of volatility since 2004



Source: Own processing

Acknowledgments

This research could be performed due to the support of the University of Pardubice student project grant no. SGS_2019_018, Dynamika ekonomických aspektů veřejných politik v kontextu efektivnosti a udržitelnosti. (Faculty of Economics and Administration).

References

Dietrich, D. (Ed.). (2015). *Data science & big data analytics: discovering, analyzing, visualizing and presenting data*, Wiley, ISBN 978-1-118-87613-8

Fanta, Jiří (2001). *Psychologie, algoritmy a umělá inteligence na kapitálových trzích*. Grada, Finance, Praha, ISBN 80-247-0024-7

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

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

Petr, P. (2014). *Metody Data Miningu*. Univerzita Pardubice, Pardubice, ISBN 978-80-7395-872-5

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

van der Aalst, W. M., Bichler, M., Heinzl, A. (2018). *Robotic process Automation*, online on https://link.springer.com/content/pdf/10.1007%2Fs12599-018-0542-4.pdf

Willcocks, L. P., Lacity, M., & Craig, A. (2015). *The IT function and robotic process automation*, online on http://eprints.lse.ac.uk/64519/1/OUWRPS_15_05_published.pdf

Yahoo Finance (2019) *Bussines Finance, Stock Market, Quotes, News* [online]. Retrieved from: https://finance.yahoo.com/