FUNDAÇÃO GETULIO VARGAS

ESCOLA DE ADMINISTRAÇÃO DE EMPRESAS DE SÃO PAULO

LUCIAN ANDREAS FELIX DIETSCHE

USING DEEP LEARNING IN STOCK PRICE FORECASTING

SÃO PAULO

2019

LUCIAN ANDREAS FELIX DIETSCHE

USING DEEP LEARNING IN STOCK PRICE FORECASTING

Thesis presented to Escola de Administração de Empresas de São Paulo of Fundação Getulio Vargas, as a requirement to obtain the title of Master in International Management (MPGI).

Knowledge Field: Economics, International Finance

Adviser: Prof. Dr. Rafael SCHIOZER

SÃO PAULO

2019

Dietsche, Lucian Andreas Felix.

Using deep learning in stock price forecasting / Lucian Andreas Felix Dietsche. - 2019.

112 f.

Orientador: Rafael Felipe Schiozer.

Dissertação (mestrado profissional MPGI) – Fundação Getulio Vargas, Escola de Administração de Empresas de São Paulo.

1. Ações (Finanças) - Preços. 2. Empresas - Avaliação. 3. Inteligência artificial. 4. Aprendizado do computador. 5. Redes neurais (Computação). I. Schiozer, Rafael Felipe. II. Dissertação (mestrado profissional MPGI) – Escola de Administração de Empresas de São Paulo. III. Fundação Getulio Vargas. IV. Título.

CDU 336.763.2

Ficha Catalográfica elaborada por: Isabele Oliveira dos Santos Garcia CRB SP-010191/O
Biblioteca Karl A. Boedecker da Fundação Getulio Vargas - SP

LUCIAN ANDREAS FELIX DIETSCHE

USING DEEP LEARNING IN STOCK PRICE FORECASTING

Thesis presented to Escola de Administração de Empresas de São Paulo of Fundação Getulio Vargas, as a requirement to obtain the title of Master in International Management (MPGI).

Knowledge Field: Economics, International Finance

Approval Date

____/____/_____

Committee members:

_____

Prof. Dr. Rafael Felipe SCHIOZER

_____

Prof. Dr. Hsia Hua SHENG

_____

Prof. Dr. Vinicius Augusto Brunassi SILVA

ACKNOWLEDGMENT

I would like to thank my supervisor Prof. Rafael Schiozer and my family for their tremendous support during the time I wrote my thesis.

ABSTRACT

The following paper investigates the possibility of using artificial intelligence, in particular a long short-term memory Network (LSTM), to forecast stock prices. As input data 59 different variables are chosen based on desk research and include: fundamental, technical, and macroeconomic data. The objective of the study is to use the selected independent variables to predict the stock return of the subsequent quarter of five retail companies listed on the Brazilian stock exchange (IBVOESPA). The research showed, that LSTM can be used to forecast stock price changes and an investment strategy based on the forecasts outperforms a buy and hold strategy of the same stock. Nevertheless, it should be said, that such an investment strategy is unlikely to have the same return in a real environment like it had in the backtesting. The reason for that is, that the number of data entries for each individual variable was not sufficiently large and the LSTM was not able to generalize the relationships. In other words, the superior performance of the algorithm may be due to overfitting of the model.

KEY WORDS: Finance and Accounting, Valuation, Equity Valuation, Firm Valuation, Financial Ratios, Artificial Intelligence, Neural Network, Recurrent Neural Network, LSTM.

RESUMO

O artigo a seguir investiga a possibilidade de usar inteligência artificial, em particular um long-term memory network (LSTM), para prever os preços das ações. Dados considerados são 59 variáveis diferentes escolhidas com base em pesquisas e incluem: dados fundamentais, técnicos e macroeconômicos. O objetivo do estudo é usar as variáveis selecionadas para prever o retorno das ações do trimestre subsequente de cinco empresas de varejo listadas na bolsa de valores brasileira (IBVOESPA). A pesquisa mostrou que o LSTM pode ser usado para prever mudanças nos preços das ações e uma estratégia de investimento baseada em previsões supera a estratégia de investimento “buy and hold” da mesma ação. No entanto, deve-se dizer que é improvável que essa estratégia de investimento tenha o mesmo retorno em um ambiente real do que no backtesting. O motivo disso é que o número de entradas de dados para cada variável individual não era grande suficientemente e o LSTM não foi capaz de generalizar os relacionamentos. Em outras palavras, o desempenho superior do algoritmo pode ser devido ao ajuste excessivo do modelo.

PALAVRAS CHAVE: Finanças e contabilidade, avaliação, avaliação de ações, avaliação de empresa, índices financeiros, inteligência artificial, rede neural, rede neural recorrente, LSTM.

Contents

Abbreviations

ADF Augmented Dickey Fuller Test

AI Artificial Intelligence

ANN Artificial Neural Network

BCB Banco Central do Brasil

BRDT3 Petrobras Distribuidora SA

BRL Brazilian Real

BTOW3 B2W Companhia Digital

capex Capital Expenditures

CCI Consumer Confidence Index

CNI Confederação Nacional da Indústria

CPI Consumer Price Index

DSO Days Sales Outstanding

EPS Earnings per Share

FDI Foreign Direct Investment

GDP Gross Domestic Product

GICS Global Industry Classification Standard

IBGE Instituto Brasileiro de Geografia e Estatística

INEC Índice Nacional de Expectativa do Consumidor

LAME3 Lojas Americanas SA

LREN3 Lojas Renner SA

LSTM Long short-term memory

MAE Mean Absolute Error

MGLU3 Magazine Luiza SA

ML Machine Learning

NOA Net Operating Assets

REER Real Effective Exchange Rate

ReLU Rectified Linear Unit

RGB Relative long-term government bond yield

RMSE Root-Mean-Square Error

RNN Recurrent Neural Network

RNOA Return on Net Operating Assets

ROA Return on Assets

ROE Return on Equity

Tanh Hyperbolic Tangent Activation Function

TSP Term Spread

USD United States Dollar

VMA Variable-Length-Moving-Average

VVAR3 Via Varejo SA

List of Figures

List of Tables

1 Introduction

Almost all real-word data has a temporal aspect, which explains why time series analysis has been a well researched field for decades and due to the increased availability of data in recent years gained further in popularity (Längkvist, Karlsson, & Loutfi, 2014, p. 11). Even though a lot of research has been conducted in the area of time series analysis, there is no unifying framework that is suitable for different data mining tasks. Due to international capital flow and more volatile financial markets, forecasting has never been more important in the decision making process of decision makers such as politicians or financial analysts. Forecasting in this sense describes the process of understanding which independent variables may help to predict the respective dependent variable that the decision maker is interested in. The results help to gain valuable insights into the relationship amongst the variables considered and their mutual behavior in the past (McNelis, 2005, pp. 1-3).

According to Yang and Wu (2006) the complexity is the reason, why time series analysis remains one of the most challenging problems in data mining research. Artificial intelligence (AI) and deep learning have gained much attention in the recent years. This surge in popularity is mainly due to significant progress in certain areas such as image processing, natural language processing, and time series analysis. Growing interest in other areas also raised the question whether deep learning can be applied in finance. The current literature is still limited, but there are several authors that have used deep learning in time series forecasting, especially stock price prediction. The idea behind this stock prediction approach is that past data will be utilized to forecast future prices and deep learning models are created in order to determine hidden nonlinear relationships between data and future stock returns (Chong, Park, & Han, 2017). Predicting stock prices based on past data is something that has been around for decades and is a method used by finance professionals. The issue however is, that complex world problems such as stock price changes are hard to express with analytical equations and require the development of a complex model to forecast stock prices. A potential candidate for solving such complex problems are Artificial Neural Networks (ANN). They are structured the same way like a human brain and thus allows to model a human like thinking considering different variables at the same time in a non-linear fashion (McNelis, 2005, pp. 21-23).

The objective of this study is to analyse whether an ANN can be used to make stock return predictions. Other than existing studies, this study will not only use technical indicators to forecast stock price development, but combine fundamental, technical, and economic data. Besides that, a longer investment horizon will be considered. The purpose of the study is therefore to explore if ANN can be used to forecast stock prices and thus can be used to create superior returns.

As the universe of stocks is enormous, the focus of the predictability of the stock prices will be on retail companies that are part of the Bovespa Index and constituents of the GICS Industry Group *2550 Retailing*: LREN3, MGLU3, BTOW3, VVAR3, LAME3. The downsizing to these five companies was done during the research process, because it became visible that on a single company level no ANN can be programmed due to the lack of accounting data. As a result of that, the largest GICS Industry Group in which companies had comparable business models was chosen.

2 Usage of AI/ML in Practice

First and foremost, the interviews revealed that investors, financial analysts and advisors in the Brazilian market still mainly rely on classical approaches to stock picking and stock price forecasting. They consider technical, fundamental, and macro data, but rely on man-made models and traditional valuation methods. Some of the people interviewed said, that they are starting to use ML or AI in some of their models or decision making process, but its usage is still in its early days.
That being said, it can be seen in the answers of the people, that AI and ML are gaining importance in assisting the analysts in their stock selection and price forecasting. Emy Shayo (2019) from JPMorgan said that they are increasingly using computer models for their quantitative research, so analysts can focus more on the thematic part of the company analysis. She said, that the creation of quantitative reports are already fully automated. She further pointed out, that they worked with a startup to use web-scraping for information gathering and integration in their research process. This gives them an advantage, because they are able to monitor real time development and provide this research to their own clients. On the contrary, smaller asset managers and private investors, do not have access to these type of solutions yet and thus monitor developments manually. This became clear as many people from smaller investment firms replied, they could not help me with an interview, because they do not employ AI or ML in their investment process.
Pandhora is a fund employing quantitative investment strategies and thus has a strong focus on data analytics. Even they are cautious with employing AI/ML, because according to Eduardo Ferrari Malgalhães (2019), a quantitative analyst at Pandhora, many deep learning models cannot be understood well. This is a big challenge, because it makes it impossible to locate the error and improve the model if it stops working well in reality. In addition to that, he pointed out that many models perform well in the training phase, but poorly on out of sample data. As a result of that, they still prefer man-made models to assist the analysts in the investment decision process, but use some ML to search for important variables or model some events.

Even though ML and AI still has some problems, many of the interviewees states that algorithms in general have different advantages. The main advantage of computer-based models using AI/ML according to Markus Jordi (2019) from theScreener is, that they do not have emotions and are thus not prone to make necessary adjustments to their forecasts. Analysts, or humans in general, often have difficulties in admitting they were wrong. This creates opportunities in the market, because one might be able to bet against the market consensus. Nevertheless he said, that the reaction of quantitative signals often have a time-lag, meaning that the algorithms take some time to gather enough data to be able to update their forecasts. Eduardo further pointed out that algorithms are consistent, meaning that given the same input they always produce the same output. This is important, especially during crisis when humans often overreact due to their emotions. He however said, that ML and AI is used to help the analyst make better investment decisions and not replace him. This supports the point mentioned by Emy Shayo earlier, that currently computer-based models are here to assist the analyst to make superior investment decisions rather than making the investment decision themselves.
All of the people interviewed stated that they believe AI and ML will become more important in the future, because of advantages mentioned before. That being said, we are still in the early stages and only just beginning to use the power of machines in the investment process. Brazil in particular seems to be a little further behind some more developed countries such as the US, but startups are here to close the gap in this area as well (Catalan, 2019; Koh, 2019).

3 Literature Review

Stock price predictions have always been a hot topic and are widely covered in academic research. The recent boom of machine learning has led to an increased research coverage in this area. The reason for its popularity is the fact that machine learning algorithms are able to process large amounts of data and may help to explain the complexity of the stock market and its value drivers. Deep learning in particular is very suitable for forecasting stock prices according to Heaton, Polson, and Witte (2016), because stock prices are influenced by a large number of factors and ANN potentially can detect relationships among these data points that have not yet been considered by classical financial theory. In their examination of the recent literature of stock price forecasting using artificial intelligence, Krollner, Vanstone, and Finnie (2010) found that the usage of ANN seems to be predominant in the area of stock price forecasting. Chatzis, Siakoulis, Petropoulos, Stavroulakis, and Vlachogiannakis (2018) developed a systematic approach to select the most important variables to forecast a stock market

crisis using various machine learning approaches. When comparing the efficacy of the different algorithms, they found that deep learning is able to create a superior performance prediction compared to other machine learning approaches. Also Chong et al. (2017) use deep learning networks for stock analysis and forecast by relying on a vast set of high frequency intraday stock returns to build their prediction model for the Korean stock market. Out of all the models they constructed, the deep neural network's performance is higher compared to a linear autoregressive model in the training set, but in the test set the difference vanishes. Similarly, Zhang, Cui, Xu, Li, and Li (2017) use the sliding window method to first dissect and assign stock transaction data into four major classes (Up, Down, Flat, Unknown) by applying unsupervised learning. Then they further split the classes Up and Down according to their magnitude of price appreciation respectively depreciation to train their prediction model with and without feature selection.

Heaton et al. (2016) also use deep learning models to forecast financial data and further use this data to construct a smart portfolio to measure their relative performance against its benchmark index. Like other authors the three authors mainly depend on historical returns as input data for the construction of the model. Ticknor (2013) was among the first one to use Bayesian regularized artificial neural network in order to predict financial market behavior. As input factors he considered daily market prices and technical parameters with the goal of forecasting the next day's closing price of a stock. A Bayesian regularization has the advantage to punish an overly complex neural network and makes models more robust. Also Hedayati Moghaddam, Hedayati Moghaddam, and Esfandyari (2017) try to predict the price movements of the NASDAQ using an artificial neural network, besides short term historic returns, the researchers further used the day of the week as data input to create their algorithm. Likewise Dai, Wu, and Lu (2012) used nonlinear independent component analysis combined with backpropagation neural networks to do their time series forecasting of Asian stock markets. According to the authors the forecasting probability of the combination is higher than just using a backpropagation neural network as a forecast model. Another hybrid model was developed by Zhou, Yang, Zhou, and Yang (2019) in which they used a two step approach to integrate Empirical Mode Decomposition and a Factorization Machine based Neural Network to forecast stock market closing prices. They use the decomposed time series data as input data for their neural network in order to exploit non-linear relationships for the stock price forecasting.

Patel, Shah, Thakkar, and Kotecha (2015) used four different models amongst them an ANN to predict the direction (up or down) of Indian stocks and indices. In order to train the model the authors use ten technical parameters of the stock and index prices. As the technical parameters are continuous numbers the authors suggest to use a Trend Deterministic Data Preparation Layer in order to convert it to a value +1 or -1, which

should indicate an up respectively a down movement. They were able to improve the overall accuracy of the model by integrating a Trend Deterministic Data Preparation Layer instead of using the technical parameters directly as a data input.
Nayak, Pai, and Pai (2016) built a supervised model to forecast stock market trends for the next day. The data input considered by the authors were also technical indicators such as continuous up/down and recent transaction volume, but also recent sentiment about the company. One of the challenges faced by the researchers was the fact that for some companies they did not have daily news/tweets to conduct a sentiment analysis. Their findings show that the accuracy of the next day's forecast can be improved by including intraday price movements, in longer term models sentiment analysis provides a better forecast accuracy.
Rather, Agarwal, and Sastry (2015) use an autoregressive moving average, an exponential smoothing model and a recurrent neural network in order to predict stock returns. In order to improve the accuracy they use a hybrid model, which combines the predictions of the three models by determining the weight by a genetic algorithm. Using a combination of models and an optimal calibration of their forecasts the authors were able to create an exceptional forecasting model. Göçken, Özçalıcı, Boru, and Dosdoğru (2016) also used a hybrid ANN to forecast stock prices and were able to show that a mixed model performs better. Further Hsu (2013) used an approach to combine three different models with each other in order to forecast the stock price of the Taiwan Stock Exchange. First, Hsu created a backpropagation neural network to construct a stock price forecasting model based on the complex, non-linear relationship of technical indicators and stock prices. He then used a feature selection technique to find the features that influence the stock price the most in order to then use these indicators to optimize the neural network. His research shows, that short term stock price predictions with reasonable forecasting accuracy can indeed be done using a neural network.
Safi and White (2017) created a model based on an ANN as well in order to forecast stock prices and compared its performance to an autoregressive integrated moving average (ARIMA). They found that an ANN has better forecasting accuracy independent from the granularity. A potential explanation from the authors is, that an ANN can detect nonlinear and non-stationary relationships which are common in stock data. Based on their findings, the authors recommend using ANN to forecast economic data, especially if the model is non-linear. They further mention that the accuracy of the ANN increases if the data input grows.
In their paper, Kara, Acar Boyacioglu, and Baykan (2011) used an ANN and a model based on support vector machines (SVM) to predict the direction of the stock price development. They found that both models are able to forecast the direction with decent accuracy, but the ANN has a superior forecasting accuracy in a direct comparison. The authors further suggest, that the accuracy of their models could be increased by in-

cluding other variables such as macro-economical data.
As can be seen in the literature review, stock price forecasting using deep learning is already a well covered area in academics, nevertheless there is so far, no author that included fundamental data such as free cash flow, gross margin, or Price to Book Ratio as input data for the construction of the model.

4 Definition of a Neural Network

Before starting to set up a neural network, it is important to have a look at the definition and an illustration in order to better understand what a neural network is. The characteristics of a neural network depends on three main elements: the architecture, the process of determining the weights, and the activation function (da S. Gomes, Ludermir, & Lima, 2011, p. 419). *"Similar to linear and polynomial approximation methods, a neural network relates a set of input variables* $\{x_i\}, i = 1, ..., n$ *to a set of one or more output variables,* $\{y_i\}, i = 1, ..., n$*. The difference between a neural network and the other approximation methods is that the neural network makes use of one or more hidden layers, in which the input variables are squashed or transformed by a special function."* (McNelis, 2005, p. 21). These hidden layers enable an efficient modelling of nonlinear relationships amongst the variables (McNelis, 2005, p. 21). It is the nonlinearity that makes neural networks suitable for real world problems, the reason for that is, that the assumption of linearity in traditional econometric models if often not valid in reality. Building non-linear models is therefore crucial (Marwala, 2013, p. 6).

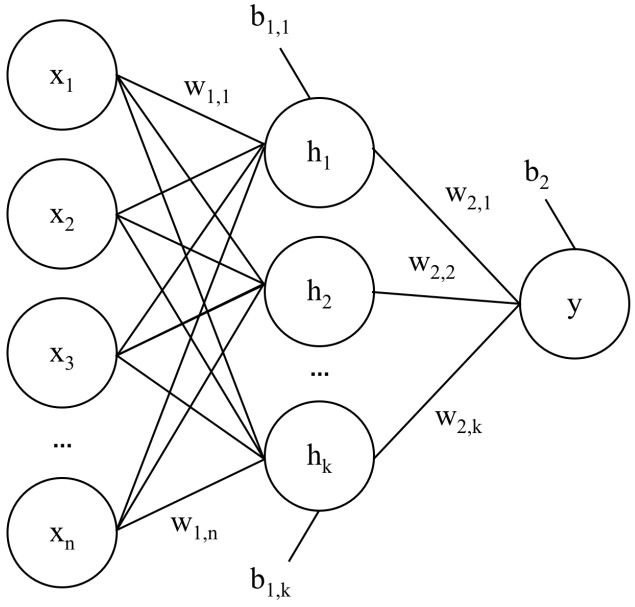


Figure 1: Illustration of a Neural Network, own illustration

Figure 1 is an illustration of a neural network with input variables$\{x_i\}, i = 1, ..., n$ containing one hidden layers with the nodes $\{h_i\}, i = 1, ..., k$ and a single output variable y. As can be seen in the illustration, a neural network has both parallel and sequential processing. Each connection between the nodes in a neural network is associated with a specific weight $\{w_{i,j}\}, i = 1, 2, j = 1, ..., n$ that regulates the flow of the data. Input variables are passed through the connections of a neural network under the consideration of the respective weights. The output of each neuron is the weighted sum of the inputs and the addition of a constant, referred to as the bias (in Figure 1 labelled as $\{b_{i,j}\}, i = 1, 2, j = 1, ..., k$) and can be written mathematically as follows:

$$y = \sum(weight * input) + bias$$

After the linear combination of the input variable is calculated, the data enters into an activation function before transmitting the output to the next neuron. The activation function is responsible for deciding, whether the neuron will be considered. In other words, it is the responsibility of the activation function to decide whether the input information is considered to be relevant or not. The activation function is also one of the main differences between a linear regression model and a neural network, because it performs a nonlinear transformation of the input data and thus enables a neural network to perform more complex tasks compared to a linear regression model (McNelis, 2005, pp. 21-24). This describes the fundamentals of any neural network, depending on the complexity of the task, the models may vary however significantly and it should be said, that there are many different types of neural networks available. Before selecting the most suitable model it is therefore important to clearly state the objectives and what should be achieved with the model. Depending on the anticipated outcome, one model may be more suitable than another one.

4.1 Feedforward Networks

Feedforward neural networks were the first type of neural networks and are the simplest among the models. As described above, the data flows from the input layer through the hidden layers without loops to the output layer. In financial and economic applications a single-layer feedforward with a single hidden layer is the most commonly used neural network. This is often also referred to as a shallow neural network. Models get much more complicated in areas such as neural linguistic, which includes the recognition of hand written letters or machine translation. In these cases, several hidden layers are used and referred to as deep neural networks (McNelis, 2005, pp. 21-24).

Participants of the financial markets base their expectations on memories such as experiences, culture, and education. They therefore establish a relationship amongst

variables in order to predict the future development of things such as share prices or exchange rates. The hidden layers of neural networks allow the modelling of relationships between data points in a human fashion by establishing an interconnected relationships amongst the input variables (McNelis, 2005, pp. 21-24).
Neural networks are according to Hornik, Stinchcombe, and White (1989, p. 363) *"capable of approximating any measurable function to any desired degree of accuracy, in a very specific, and satisfying sense"*. In other words, neural networks can be used to represent any measureable function and a potential failure could be traced back to *"inadequate learning, insufficient numbers of hidden layers, or the lack of a deterministic relationship between the input and the target"* (Hornik et al., 1989, p. 363). Not only are neural networks universal approximators, but two hidden layers with trainable weights are sufficient to approximate any nonlinear function. This characteristic in combination with how neural networks are trained allows the process of the creation of the functional form and the fitting of function at the same time. This better adaptability gives neural networks an advantage over traditional regression models (Dayhoff & DeLeo, 2001, p. 1624). The flexibility of neural networks make them very suitable for stock price forecasting if the right input variables are chosen.

4.2 Recurrent Neural Network (RNN)

Recurrent Neural Networks differ largely from Feedforward Networks, as output and input nodes can be connected with each other. As a result of that, the network is sometimes referred to have a memory, because it does not only depend on current inputs but takes into consideration their own lagged values, like depicted below in figure 2. Such feature makes RNN suitable for tasks where previous inputs may also contain information for subsequent outputs. This approach is similar to the process of moving averages in time series analysis, in which the dependent variable describes a function of the current and lagged value plus some random error. Due to the reliance on lagged values, RNN are mainly suitable for datasets that have a temporal aspect, such as stock prices (Yadav, Yadav, & Kumar, 2015, pp. 25–26; McNelis, 2005, p. 34; Brownlee, 2019, pp. 9-10). In RNN there are two types of feedback that can be included in the model. On the one hand global feedback, which connects the network output and the network input with each other and on the other hand a local feedback, which relates lagged values within the hidden layer or between single neurons (Mandic & Chambers, 2001, p. 43). Due to this characteristics, an RNN will be chosen to create the neural network in the second part of the paper.

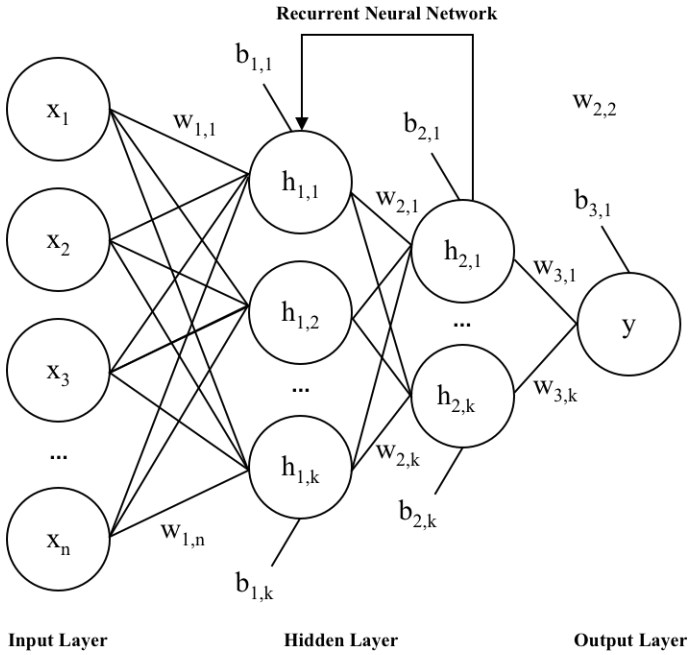


Figure 2: Recurrent Neural Network Illustration, own illustration

Recurrent neural networks are powerful machine learning tools that allow modelling of time sequences with dynamic temporal behavior. Despite that, RNN have shown to face difficulties when gradient descent is used for optimization of parameters as the temporal span of the data increases (Yoshua, Patrice, & Paolo, 1994, pp. 26-27). The reason for that is the gradient descent problem, which is a phenomena arising because of the way neural networks find the optimal weights of the neurons. The neural network gets optimized by changing the value of the parameters slightly and then calculates the impact on the forecasting accuracy (Bengio, Boulanger-Lewandowski, & Pascanu, 2012, pp. 1-2). One way to address this problem is by choosing an activation function that has a larger output range and is thus not prone to the gradient descent problem to the same extent or using a specific RNN such as long short-term memory (LSTM).

4.3 Long Short-Term Memory (LSTM)

Long short-term memory is one specific form of recurrent neural networks and is well suited for time series forecasting. LSTM is a particularly relevant model, if the order of the observations in a sequence should be preserved. The reason for that is, that the order of observations contains relevant data and should be kept in the training phase as well as when predicting the next value in the sequence. The type of model that is required in the stock price prediction is called *Sequence Prediction*. In this type of model an input is given sequence in order to forecast the next observation in the sequence (Brownlee, 2019, pp. 3-4). Businesses are complex and can only be understood fully when considering their historic development (Marshall, 2017, p. 39). As mentioned

previously, an LSTM helps as it considers previous neurons as an input factor in the forecasting model.
The problem faced by many different forms of recurrent neural networks is that they take a lot of time to be optimized and often do not work very well, especially when the range of the time lags is large. Such large time lags often make past error signals blow up or vanish as they depend on the magnitude of the weights (Hochreiter & Schmidhuber, 1997, pp. 1735-1737). LSTM tries to address the issue of error backflow by *"enforcing constant error flow through constant error carousels within special units"* (Hochreiter & Schmidhuber, 1997, p. 1735). In other words, one of the main advantages of LSTM is that it is able to handle greater range dependencies compared to other RNNs, as it is less prone to the vanishing gradient problem (Bengio et al., 2012, pp. 2-3; Brownlee, 2019, pp. 9-10). Because of LSTM's superior ability to exploit long range content it has frequently been used for speech recognition (Graves, Mohamed, & Hinton, 2013, pp. 1-2) by big tech companies and makes it an optimal solution for stock time series analysis. The reason for that is, that LSTM can include longer-term trends and process lags of various durations. Besides overcoming the problem of vanishing and exploding gradients that RNN often faces, it contains a memory for long-term temporal relationship of the input sequence, and it processes the sequence in chronological order (Brownlee, 2019, pp. 11-12).
Of course, LSTM also have their limitations and are not suitable for all time series forecasting. In cases in which the forecast is based on a short sequence of input data, simpler models may be more suitable and produce equally accurate results. It is therefore important, that the individual problem is considered and the suitability of LSTM is examined (Brownlee, 2019, pp. 14-15).

4.4 Activation Function

As described earlier in this section, neurons process data first through the formation of a linear combination of the input data and then enter the output into an activation function, which decides whether the input is relevant or not. An activation function needs to have a threshold behavior, meaning if the outcome is above a certain limit the node will be activated otherwise the node will stay inactive. It is the activation function that gives a neural network the nonlinear feature (McNelis, 2005, pp. 21-24). The threshold feature was described by Kuan and White as: *"tendency of certain types of neurons to be quiescent of modest levels of input activity, and to become active only after the input activity passes a certain threshold, while beyond this increases in input activity have little further effect"* (1994, p. 2). As discussed earlier, a neuron essentially multiplies the inputs with the respective weights and adds a bias to it. The output of this equation will then be inserted into the activation function to determine whether a neuron was

activated. Mathematically this can be represented as follows:

$$y = f(net) = f(\sum(weight * input) + bias)$$

whereas f is the activation function of the neuron (Yadav et al., 2015, p. 21). There are different activation functions used in the creation of neural networks and depending on the characteristics the result may differ significantly. One of the necessary conditions is however that the function is differential, if a gradient-descent learning algorithm is used. In practice, monotonic functions that satiate at two extremes are often chosen as activation functions. When designing a neural network, different activation functions can be used at different neurons, important is that they contain a free parameter, so called gain, which needs to be set by trial and error (Azoff, 1994, pp. 51-55; Mandic & Chambers, 2001, p. 47).
There is a large range of requirements an activation function has to fulfill in order to be considered suitable for the nodes in a neural network, the detailed analysis of these requirements and the mathematical implications would however exceed the purpose of this paper (Mandic & Chambers, 2001, pp. 47-52). Some of the commonly used activation functions from recent years are:

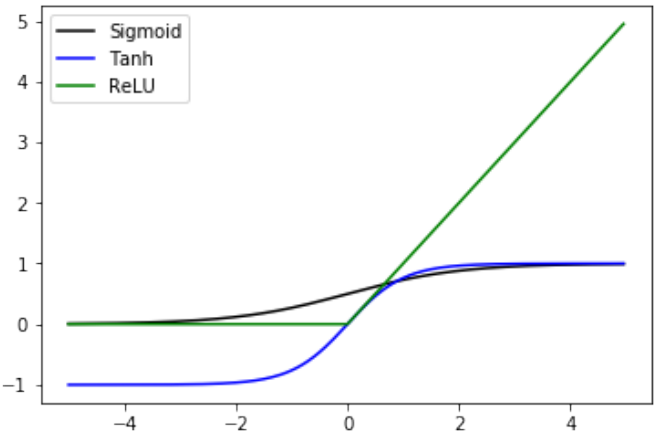


Figure 3: Activation Functions, own illustration

- **Sigmoid**
  The Sigmoid function is still a commonly used function that accepts input values (I) from $-\infty$ to $+\infty$ and creates outputs(O) in the range $O \in [0, 1]$ (Azoff, 1994, p. 52). The Sigmoid function is still a popular activation function in neural networks, but it is subject to the gradient vanishing problem.
  Each of the neural network's weights receives an update proportional to the partial derivative of the error function with respect to the current weight in each iteration of training. The problem is that in some cases, the gradient will be vanishingly small, effectively preventing the weight from changing its value. In the

worst case, this may completely stop the neural network from further training (Pascanu, Mikolov, & Bengio, 2012, p. 52; da S. Gomes et al., 2011, pp. 47-52).

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$$

See figure 3 for graphical illustration of a Sigmoid function.

- **Swish**

  The Swish function is defined as

  $$y = x * \sigma(\beta x) = x * \frac{1}{1+e^{-\beta x}}$$

  It should be noted that $\sigma(\beta x)$ is the regular Sigmoid function that and β represents a trainable parameter or can be set as a constant. Further it should be noted that as β approaches ∞ the Swish function becomes similar to the ReLU activation function and with β equal to 1 it represents the the Sigmoid-weighted Linear Unit (SiL), which is another activation function. As a result of that the Swish activation function can be seen as *"a smooth function which nonlinearly interpolates between the linear function and the ReLU function"* (Ramachandran, Zoph, & Le, 2017, p. 5; Elfwing, Uchibe, & Doya, 2018). Setting β as a trainable parameter allows the control of the degree of interpolation. In their paper Ramachandran et al. (2017, pp. 4-5) found that Swish in the case of automatic search techniques was able to outperform other activation functions including ReLU when comparing forecasting accuracy. This makes Swish a great candidate as an activation function for neural networks. So far however Swish has not been tested on time series forecasting or stock market prediction.

- **Hyperbolic Tangent Function (Tanh)**

  The Tanh function is a Sigmoid function (S) that is symmetrical about the origin with the output range $[-S, S]$ (Azoff, 1994, p. 52).

  $$f(x) = tanh(x) = \frac{2}{1+e^{-2x}} - 1$$

  See figure 3 for graphical illustration of a Tanh function.

- **Rectified Linear Unit(ReLU)**

  The ReLU is currently the most widely used activation function and is mathematically defined as (Ramachandran et al., 2017, p. 1):

  $$y = f(z) = f(max(z, 0))$$

  Recently ReLU has gained much attention, because of its superior learning speed in networks with many layers compared to other popular functions such as Sigmoid or Tanh. As a result of its simplicity and effectiveness the ReLU has established itself as the standard activation function amongst many people of the deep

learning community. There have been attempts to replace the ReLU with other functions, but practitioners still favor ReLU, because it is simple and reliable. Many activation functions are inconsistent when comparing different models or datasets, which can negatively impact the forecasting power (Ramachandran et al., 2017, p. 1). Research around activation functions is constantly evolving and it is probable that the ReLU function sooner or later will be replaced by an even more effective activation function. One possible example is Swish, the activation function discovered by Ramachandran et al. (2017, p. 10). See figure 3 for a graphical illustration of a ReLU function.

The list of potential activation functions is enormous and many of them have positive aspects for particular applications, but may be less suitable for others. The 21 novel activation functions suggested by Ramachandran et al. (2017) were evaluated by Eger, Youssef, and Gurevych (2018, pp. 4420-4423) in different settings. Their findings were that ReLU and a similarly shaped Swish function are overall able to achieve the best performance in various tasks, but they lack a stable performance. On the other hand the performance of the penalized version of Tanh is much more stable across datasets and tasks, and due to that stability the high performance can be achieved using different hyper-parameters. As a result of that, the authors recommend a penalized Tanh for more sophisticated networks including long short-term memory networks (LSTM), in which activation functions such as ReLU do not perform very well.
It should be said, that all of them but the regular Sigmoid function were among the top performers overall in the paper written by Eger et al. (2018, pp. 4420-4423) and thus the popularity amongst practitioners reflects the actual performance in the academic context.
Last but not least, it should be said, that for the output neuron it is crucial to use an activation function with the desired output distribution or preprocess the input values in order to get the needed range (Mandic & Chambers, 2001, pp. 47-48). This allows for a direct interpretation of the output variable and an inverse transformation of the forecasted dependent variable. This is required in order to estimate the accuracy of the ANN. The process of approximating the input-output relationship using a neural network is essentially the approximation of an unknown function based on the input data points (Mandic & Chambers, 2001, p. 51).
The previously mentioned information will be taken into account when preparing the data and designing the ANN, starting in section 6.

5 Selection of Data Input

There are many different ways to value a stock and identify whether it is priced fairly. Amongst the various methods there is a distinction between a relative and absolute valuation. Absolute approaches include ways to value a company from the ground up based on financial reports. The issue with these valuation approaches is that they are often based on a lot of assumptions and are troublesome to use. That is the reason, why relative valuation methods such as multiples are often used to simplify the assessment of a company valuation and compare companies amongst each other (Rossi & Forte, 2016, pp. 5-7). As a consequence, the LSTM will not try to value the company with an absolute approach, but will consider relative valuation and the development of selected accounting ratios over time. Looking at the company within the economic environment and across time should give a profound understanding of the company and hopefully its fair valuation respectively the development thereof.

Generally you can feed as many variables into an ANN as possible and the activation function will decide which ones should be considered. Although this is a common approach, the model of an ANN is, unlike a regression model, hard to interpret. For that reason, the relevance of the input data will be explained before constructing the ANN to ensure that the data is actually relevant for the profitability of a company and could potentially have an impact on the stock price.

As mentioned in section 1 of the paper, the construction of the LSTM will be based on Brazilian retail companies that are traded on The Bovespa Index and are part of the GICS Industry Group *2550 Retailing*: LREN3, MGLU3, BTOW3, VVAR3, LAME3. BRDT3 is left out on purpose, because it is owned to 71.25% by Petroleo Brasileiro SA Petrobras, which is included in the corruption scandal Lava-Jato. As a result of that, the stock price was heavily influenced by the political scandal. In addition to that, the business of petroleum retail is very different from regular retail.

Taking a single Brazilian company and constructing an LSTM just on this specific company would be ideal, because each company is unique and should be treated accordingly. This is however simply impossible, as accounting numbers are only released quarterly and there is thus not enough reliable data available to train such a model. For this reason, the GICS industry with the largest number of constituents that have a similar business model was chosen.

When valuing a company, fundamental investors consider not only company data, but also a broader view of the development of the economy. The reason behind that is, that the success of a company is linked to the overall economic development. Depending on the industry the company operates in, the impact of business cycles may be greater (Zoran, Sinisa, & Suzana, 2013). The importance of considering macroeconomic variables when forecasting stock prices is further supported by various authors such as

Flannery and Protopapadakis (2002), who show that some economic variables are able to influence the development of stock prices.
The idea behind the data selection of the LSTM is summarized illustrative in figure 4. The graphic shows the concept behind the data selection process. The company operates in its environment and has a constant exchange of resources with it. Because of this relationship, high level, macroeconomic variables should be considered. These variables affect companies to a different degree, but has a similar effect on companies within the same industry. The capital market is considered part of the environment, but is overlapping with the corporation. This is based on the fact that companies raise equity and debt on the capital market and its valuation changes daily. On the company level there are not only dynamic ratios such as the market value ratios that change daily, but also accounting data that is only reported quarterly. Importance is given to all elements that impact business operations and it will be up to the LSTM to find out the connection between the variables and the future stock price and the relative importance thereof. Since there is no way to find out which variables are relevant to forecast the stock price, the following section of the paper will discuss in detail the three different categories of input variables.

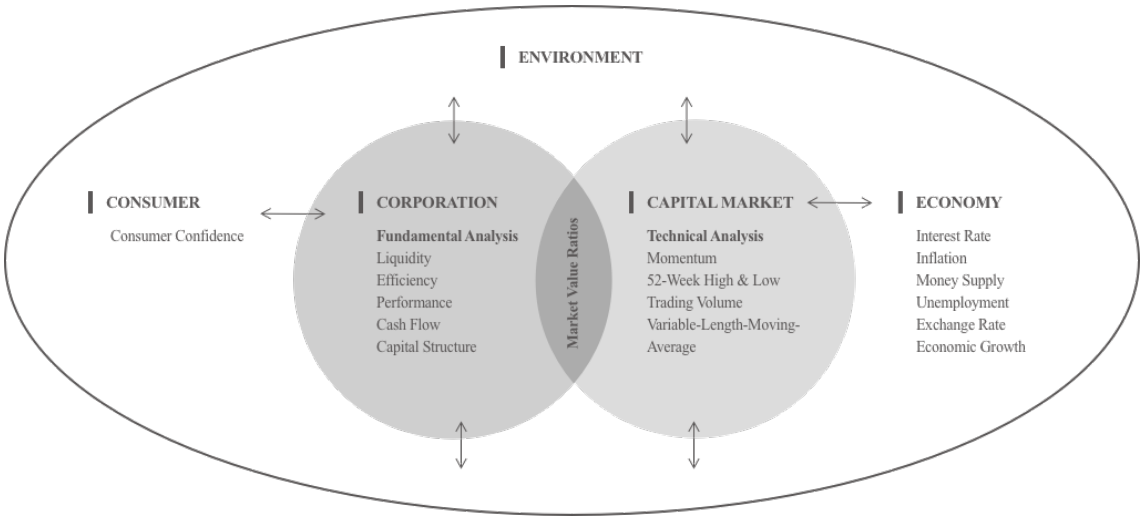


Figure 4: Illustration of the Data Input Variables, own illustration

5.1 Fundamental Parameters

Financial ratios are a mixture between accounting data and (Charles, Darne, & Kim, 2017, pp. 13) Using fundamental parameters as input data for the stock price forecasting was already done by Ou and Penman (1989). They found that fundamentals often include information that was not entirely reflected in the stock prices. Even though the paper was published in 1989 and markets have changed significantly since then, several papers including Piotroski (2000) or Wahlen and Wieland (2011) have found that fundamental analysis can still be used to generate abnormal returns over longer

investment horizons. Wahlen and Wieland (2011) even showed in their research that investors would be better off focusing entirely on information available in the financial statements while ignoring the analysts' consensus.
For the above mentioned reasons, relevant fundamental parameters will be included as input data for the construction of the ANN.
It should be mentioned, that for accounting ratios, which include both figures from the balance sheet and the income statement, the ratio will be computed using the average of the number from the balance sheet (meaning it considers the beginning and the ending balance of the respective figure). This is necessary in order to improve the significance of the ratio, because numbers from the balance sheet consider a certain point in time, while figures from the income statement include a time period.
Matsumoto, Shivaswamy, and Hoban (1995, pp. 45-49) conducted a survey with over 220 financial analysts to find out which financial ratios are considered the most important when analyzing retailers and manufacturers. This survey is very important for the selection of the ratios, because the selected GICS Industry Group is *2550 Retailing*. The final selection of the ratios is thus based on the survey results of Matsumoto et al. (1995) and further desk research. The following paragraphs talk about the selected ratios and will list the computation of the most important ones, because as Damodaran (2006, p. 238) points out, even though analysts may use the same multiples, they often use different formulas.

5.1.1 Market Value Ratios

Market value ratios measure how expensive a company's stock price is relative to profit or accounting values (Rist & Pizzica, 2015, p. 2).

Dividends

The dividend yield and the dividend payout ratio have some ability to predict stock returns. A potential explanation is that dividends contain certain information about the future development of dividends and may thus be relevant for investment decisions. It should however be said, that the dividend yield and the dividend payout ratio are only relevant for short term stock price developments (Owen, 1998, pp. 1582-1582).
Nevertheless, Graham and Bogle (2005, pp. 294-295) states, that continuous dividend payments over a long time horizon is one of the most compelling tests of a high quality stock. The ANN will account for this as an LSTM includes feedback connections with previous values.
The selection of this value is supported by the findings of Matsumoto et al. (1995, p. 49), they showed that the Dividend Yield is the most important ratio in the category of *Dividend Ratios* when analyzing retail companies.

Price to Book Ratio (P/B Ratio)

Even though the book value is not the most important ratio according to the study conducted by Matsumoto et al. (1995, p. 49), they still found that the ratio is given a significant weight in the analysis of retailers.

$$P/B\ Ratio = \frac{Market\ Value\ of\ Equity}{Book\ Value\ of\ Equity}$$

The P/B Ratio puts the share price relative to the book value of the company, which represents the accounting value of net assets (Rist & Pizzica, 2015, p. 83). The ratio gives insights about the market's expectations of a company's future earning power. If a company is constantly loosing money and generates negative earnings the book value of equity can become negative. In such a case, the Price to Book Ratio cannot be calculated (Damodaran, 2006, pp. 262-264).

In their study about the European stock market, Bird and Casavecchia (2007, pp. 237-240) found that a portfolio long in value stocks based on only the Price to Book Ratio yields higher returns compared to a portfolio with growth stock, which was defined as high price to book ratio. Fama and French (2006, pp. 495-501) support the findings that companies with a lower P/B tend to have higher stock returns in the subsequent period. One of the potential explanations of this superior return, is that companies with lower P/B values tend to have higher profitability. Rafael La, Josef, Andrei, and Robert (1997, pp. 863-869) further found that the higher performance of value stock even persists when comparing companies of a similar size.

Price Earnings Ratio (P/E Ratio)

The responses of Matsumoto et al. (1995, p. 49) show that the P/E Ratio is the most important ratio, security analysts consider on average when looking at an investment opportunity in the retail sector.

$$P/E\ Ratio = \frac{Market\ Value\ of\ Equity}{Equity\ Earnings}$$

The P/E Ratio compares the current stock price relative to a company's earnings in order to give an idea of how much one is paying for future earnings (Rist & Pizzica, 2015, p. 82). Amongst the portfolio strategies studied by Santos and Montezano (2011, pp. 195-201), the portfolio strategy going long on value stocks (measured by a low P/E) outperformed in the last 20 years portfolio strategies with a focus on growth stocks on a risk adjusted basis. Their findings support the argument that in the Brazilian stock market value stocks are on average less risky and have higher yields compared to growth stocks, even after risk adjustments. The authors show a similar outperformance in the case of a portfolio composed of stocks with low P/B ratios.

As a result of these different studies, the P/E ratio and the P/B ratio will be used in the process of constructing the ANN. Not only the P/B and P/E ratios itself will be considered, but also the relative change over time shall be used as an input variable in order to give a complete overview over the two variables. As LSTM use past values as inputs it is not necessary to include both the number and the relative change.

Price to Sales Ratio (P/S Ratio)

In the analysis of retailers, analysts often also consider the P/S Ratio (Matsumoto et al., 1995, p. 49).

$$P/S\ Ratio = \frac{Market\ Value\ of\ Equity}{Revenue}$$

The Price to Sales Ratio looks at the relative market valuation of a certain company compared to the revenue generated over a specific period (Rist & Pizzica, 2015, p. 85). The advantage of the Price to Sales Ratio is that it is comparably little affected by accounting standards and choices and thus allows for a better comparison across different companies from different markets using different accounting standards (Damodaran, 2007, pp. 756-757). It can also be computed for almost all firms even though they might have negative earnings. Only in rare cases of banks and insurances, where the recognition of revenue streams may be unclear or start up companies without revenue, does the calculation of the P/S Ratio become difficult (Damodaran, 2006, pp. 262-264). Hsieh and Walkling (2006, pp. 2439-2464) found that companies with extremely high market to sales ratios tend to underperform their peers with similar price to book ratio and market capitalization both in the short and long run. Similar findings of Bird and Casavecchia (2007, pp. 237-260) in the European market further support the negative relationship between a high market to sales ratio and future stock returns. The ratio should however not be considered in isolation, because sales do not tell us anything about the profitability of the company (Montier, 2009, p. 239). As the input variables contain many different data inputs for each company, such a concern can be disregarded.

5.1.2 Profitability Ratios

The income statement gives insights about the profitability of a company and using profitability ratios can help compare the profitability across time and companies (Damodaran, 2002, p. 23).

Earnings

As Graham and Bogle (2005, pp. 310-311) states in his book *The Intelligent Investor*, *"Don't take a single year's earnings seriously"*, it is important to look at earnings over time. Long-term developments are being included by an LSTM as output of the past

neurons are reused as an input factor. Earnings can reveal more about a company if compared across similar companies in the same industry and across time (Monks, Lajoux, & LeBaron, 2011, pp. 129-130).

Earnings per Share (EPS)

The overall most important financial ratio that Matsumoto et al. (1995, p. 51) found in their survey about retail companies was the growth rate of EPS.

$$Earning\ Per\ Share = \frac{Profits\ After\ Tax}{Average\ Number\ of\ Shares\ Outstanding}$$

EPS shows the profit earned on every share after the deduction of all costs, taxes, and financial charges. It therefore represents the amount that would be available for shareholder allocation. Investors pay close attention to this ratio as EPS and the growth in EPS gives information about the possibility of a dividend payout and can be used to calculate the P/E ratio (Friend & Zehle, 2004, pp. 189-190; Vause, 2005, pp. 141-142).

Operating Income

There are four types of incomes in an income statement, but it is important to consider operating income, because it describes the income that was generated in a particular period by the core business. It is the topmost measure, meaning that it does not consider other business activities, the capital structure, or the respective tax regime. This is important for comparison, as it allows us to compare a company across time and between similar companies at the same point of time. Operating income is therefore a good indicator of the income-generating power of a company (Marshall, 2017, pp. 65-66).

Gross Profit Margin

The most important profitability ratio according to Matsumoto et al. (1995, p. 49) in the analysis of retailers is the Gross Profit Margin.

$$Gross\ Profit\ Margin = \frac{Net\ Sales - COGS}{Net\ Sales}$$

The gross margin shows how much profit a company makes relative to its sale. Margins differ largely across industries and can thus only be compared amongst peers or changes of the company itself through time (Rist & Pizzica, 2015, p. 62). The Gross Profit Margin is a good indicator of the overall profitability of a business. The ratio naturally changes if the product portfolio, the production efficiency, or the price of a product changes (Vause, 2005, pp. 132-133). Leibowitz (1997, pp. 43-45) further argues that the drivers of a longer-term competitive advantages and higher valuation are superior margins across products. The reason for his argument is based on the

assumption, that a more cost efficient production can be replicated, but it is harder to duplicate a higher perceived value.

Operating Profit Margin

$$Operating\ Profit\ Margin = \frac{Operating\ Earnings}{Net\ Sales}$$

The Operating Profit Margin is the second margin from the top of the income statement. When computing the Operating Profit Margin, companies include further operating activities that cannot be associated directly with a specific good such as research and development, distribution expenses, administration, and general overhead. As a result, the ratio gives more detailed insights about the profitability of a certain company when considering regular costs associated with the providing of its goods or services. It is important that the Operating Profit Margin is considered together with the Gross Profit Margin in order to understand the complete picture of the operating profitability of the analyzed company. As more items are included in the calculation of the Operating Profit Margin, usually the notes should be considered when analyzing this accounting figure (Vause, 2005, pp. 133-134). Evaluating the notes to make the necessary adjustments to be able to compare a company across time with a separate algorithm or manually is optimal, but would exceed the purpose of this thesis.
As mentioned earlier, the Operating Profit Margin is part of the DuPont model, in which the RNOA is decomposed into the Asset Turnover Ratio and the Operating Profit Margin to analyze the operating profitability further (Penman, 2013, pp. 374-375; Soliman, 2008, pp. 849-850).

Return on Assets (ROA)

$$Return\ On\ Assets = \frac{Net\ Income + Interest\ Expense\ after\ tax}{Total\ Assets}$$

ROA describes the profitability of a company relative to its assets and thus gives an understanding of the company's operational efficiency - how well the company uses its assets to generate revenue (Rist & Pizzica, 2015, p. 91; Damodaran, 2002, p. 23). The ROA is an extensive ratio as it includes all assets, hence it considers both the return from operations and financial assets. It is therefore helpful to include further ratios that isolate the return of operating activities (Penman, 2013, pp. 371-372). Nevertheless ROA excludes financing expenses and as a result of that gives a more realistic overall return on assets (Damodaran, 2002, p. 23).

Return on Net Operating Assets (RNOA)

$$Return\ on\ Net\ Operating\ Assets = \frac{Operating\ Income}{Net\ Operating\ Assets}$$

As mentioned previously, ROA does not separate return from operating and financing activities. This is the reason why RNOA is an important ratio to consider, as it distinguishes between income from operations and financial activities RNOA (Penman, 2013, pp. 371-372). According to Burns, Sale, and Stephan (2008, p. 40) *"Changing the amount of debt does not affect the operating assets or the profit before interest expense and, therefore, does not affect RNOA"*.
In order to calculate RNOA, the net operating assets (NOA) have to be calculated first. Normally it is required to create a reformulated balance sheet for the company in order to estimate NOA, in which all elements of the balance sheet are reorganized into either operating or financial assets (Penman, 2013, pp. 241-242; Burns et al., 2008, p. 40). In this paper the standard RNOA calculated by Bloomberg will be used (Bloomberg, 2019).

Return on Equity (ROE)

$$ROE = \frac{Net\ Income}{Book\ Value\ of\ Common\ Equity}$$

Return on equity is a good indicator about the profitability from a shareholder perspective. As the ANN will consider only equity investments of the respective company, the ROE will be an important factor that will give the LSTM a historic perspective of the profitability over time. The ratio shows the relationship between net income and the book value of common equity. As preferred shares have different claims on the company, the calculation of the ROE should be computed using net income after preferred dividends and without the book value of preferred shares (Damodaran, 2002, pp. 25-26).

5.1.3 Efficiency Ratios

Efficiency ratios help to explain the effectiveness of usage of a company's assets and liabilities to create sales (Damodaran, 2002, pp. 29-30).

Inventory Turnover

$$Inventory\ Turnover = \frac{Costs\ of\ Goods\ Sold}{Average\ Inventory}$$

Inventory turnover describes the number of times a company was able to sell and replace the inventory over a given period of time (Damodaran, 2002, p. 30; Rist & Pizzica, 2015, p. 66). A greater inventory turnover is associated with a more efficient

usage of the inventory to generate revenue. Even though a higher inventory turnover is favorable, too high of a ratio is unfavorable as it can indicate a lack of product availability (Seitz, 1976, p. 8).

Receivable Turnover

$$Receivable\ Turnover = \frac{Sales}{Average\ Accounts\ Receivable}$$

Receivable turnover calculates how many times a company is able to convert its receivable into cash in a given period (Damodaran, 2002, p. 29; Rist & Pizzica, 2015, p. 88).

Days Sales Outstanding (DSO)

$$Days\ Sales\ Outstanding = \frac{Days * Account\ Receivable}{Credit\ Sales}$$

Days Sales Outstanding is also referred to as the average collection period and describes the number of days it takes a company on average to transform its receivables into cash (Rist & Pizzica, 2015, p. 38). The development of DSO gives insights about whether the company sends inventory to customers to boost its sales (Montier, 2009, p. 262). This is critical, because an increasing collection period may generate cash flow problems due to the maturity mismatch between payables and receivables.

Asset Turnover Ratio

According to Matsumoto et al. (1995, p. 51) the Asset Turnover Ratio is one of the more important ratios in the analysis of a retail company.

$$Asset\ Turnover\ Ratio = \frac{Sales}{Assets}$$

The asset turnover ratio indicates how effective a company uses its assets to generate revenue. A higher ratio therefore demonstrates a greater efficiency. Margins should be considered in combination with asset turnover, as a company may be able to increase their sales and thus have a higher asset turnover, but if their profit margin decreases, the overall performance may suffer (Fairfield & Yohn, 2001, pp. 372-373; Rist & Pizzica, 2015, p. 14). Penman (2013, p. 374) further summarizes this negative relationship by stating that *"industries with low asset turnovers tend to have high profit margins, and industries with high asset turnovers tend to have low profit margins."*. The inclusion of this ratio is further supported by Soliman (2008, pp. 849-850), who found, that a change in Asset Turnover can help to predict future profitability and longer term stock returns are positively related with a change in the Asset Turnover Ratio.

The asset turnover ratio is part of the DuPont model that states a company's profitability is driven by two sources: First, the company's ability to increase its profit margin; second, the efficiency of usage of a company's operating assets. The analysis thus decomposes the RNOA into the Asset Turnover Ratio and the Operating Profit Margin (Penman, 2013, pp. 374-375; Soliman, 2008, pp. 849-850).

5.1.4 Cash Flow Ratios

Accounting profit and its importance can be summarized with the statement of Vause (2005, p. 89): *"Profit is part fact, part opinion and occasionally part hope. Different assumptions, views and accounting treatment will produce different profits."*. In other words, profit is important for the company, but it does not matter as much as the availability of cash. If a company has sufficient cash flow available it can survive even when making losses. As a result, it is therefore important to not look at the profit or the cash flow in isolation, but consider both when analyzing a business (Mills & Yamamura, 1998, p. 53; Vause, 2005, pp. 89-97). According to Montier (2009, p. 262) a comparison of the income and the cash flow from operations also helps to identify earnings management. As earnings are based on a lot of subjective estimates, it is far easier to influence the numbers compared to the cash flow.

Operating Cash Flow

The Operating Cash Flow determines the cash created by a company's regular business activity during the period analyzed. It further is an important determinant of whether a company can finance growth internally or it depends on external funding (Vause, 2005, pp. 97-102; Rist & Pizzica, 2015, p. 77).

Free Cash Flow

Marshall (2017, pp. 71-72) calculates free cash flow by *"subtracting capital expenditure from cash flow from operations"*. He however mentions capex should be split into maintenance capex and growth capex, the reason is that the estimation should be about the present operations, including growth capex would therefore underestimate the performance of the business. Finding the respective number of maintenance capex can be tricky and if no such number can be found in the detailed breakdown of the capex, one can use the total capex, but one has to consider that the actual free cash flow in reality is higher than the calculated free cash flow. Depreciation can give some guidance about the overall maintenance capex, but because of inflation and technological advancement there often is a big difference between depreciation and actual maintenance capex. The formula above describes the levered free cash flow as the interest payments are included in the calculation Marshall (2017, pp. 71-73). Such as detailed

breakdown in this analysis is however not possible and thus the input of the ANN will use the Free Cash Flow as reported in the quarterly reports.

Cash Flow to Debt Ratio

The results of Matsumoto et al. (1995, p. 49) show that the Cash Flow to Debt Ratio is considered by a large part of the analysts when analyzing retailers.

$$Cash\ Flow\ to\ Debt\ Ratio = \frac{Operating\ Cash\ Flow}{Total\ Debt}$$

The Cash Flow to Debt Ratio calculates the company's ability to cover its liabilities using its Operative Cash Flow. A higher ratio therefore indicates a low probability of the company not being able to meet its debt obligations (Mills & Yamamura, 1998, pp. 55-56; Rist & Pizzica, 2015, p. 25).

Cash Flow Expenditure Ratio

$$Capital\ Expenditure\ Ratio = \frac{Operating\ Cash\ Flow}{Capital\ Expenditure}$$

Not only the Cash Flow to Debt Ratio, but also the Cash Flow Expenditure Ratio can help to identify whether a company will be able to finance internal growth. The ratio considers the financial strength of a company with a longer-term view, as it measures the capital available for repayments of existing liabilities and potential internal investments. A ratio exceeding 1.0 is preferable, as the company cannot only finance reinvestment, but also has some spare cash to meet its obligations. Therefore the greater the ratio, the more money the company has available for debt service and repayment (Mills & Yamamura, 1998, p. 58; Damodaran, 2002, pp. 31-32).

5.1.5 Liquidity Ratios

When looking at the value of a company one should also include some risk measure of the investments and operations of the company itself. In the annual report there are many footnotes and disclosures regarding the risk of the company itself (Damodaran, 2002, pp. 27-28). For the algorithm however ratios are required in order to quantify these risks.
The two liquidity ratios, which analysts consider as being most important when analyzing the liquidity of a retail company in the survey of Matsumoto et al. (1995, p. 49) are:

Current Ratio

$$Current\ Ratio = \frac{Current\ Assets}{Current\ Liabilities}$$

The current ratio gives an estimate of the company's ability to finance short-term operations. It shows whether current liabilities can be covered by current assets. A liquidity risk generally arises if the company has a current ratio lower than one. Even though a liquidity ratio should be above one, it should not be too high either. The reason for that is the trade-off between reducing the liquidity risk and having cash accumulated in net working capital (Damodaran, 2002, pp. 28-29; Monks et al., 2011, pp. 88-89).
Even though working capital gives important insights about a company's short-term financial health, it cannot be considered isolated as the ratio can be influenced by the company around the reporting date (Damodaran, 2002, pp. 28-29).

Quick Ratio

$$Quick\ Ratio = \frac{Cash + Marketable\ Securities}{Current\ Liabilities}$$

The quick ratio is a variation of the current ratio and further separates current assets according to their financial liquidity. As a result of that, only assets that can be converted into cash quickly (cash and marketable securities) will be considered for the computation of the ratio (Damodaran, 2002, pp. 28-29; Rist & Pizzica, 2015, p. 9).

5.1.6 Capital Structure

It is important to see changes in the capital structure of a company in order to estimate the default risk and the long-term solvency. In this category it is important to consider two factors: the company's capacity to meet the interest rate payments and pay back the principle (Montier, 2009, pp. 31-35).

Interest Coverage Ratio

$$Interest\ Coverage\ Ratio = \frac{EBIT}{Interest\ Expenses}$$

The interest coverage ratio calculates the company's capacity to service interest payments on their outstanding debt. A lower ratio therefore increases the risk that the company will not be able to meet their interest obligations. A ratio below 1 would therefore indicate that the company is unable to generate enough earnings to pay all interests of their outstanding debt (Damodaran, 2002, p. 31; Rist & Pizzica, 2015, p. 64).

Debt to Equity Ratio

$$Debt\ to\ Equity\ Ratio = \frac{Debt}{Equity}$$

The Debt to Equity Ratio calculates the amount of debt outstanding relative to the company's equity. The ratio gives some insights about the leverage of the company (Damodaran, 2002, p. 33; Rist & Pizzica, 2015, p. 67).

Debt to Capital Ratio

$$Debt\ to\ Capital\ Ratio = \frac{Debt}{Debt + Equity}$$

The Debt to Capital Ratio gives further insights about the financing structure of a company and shows the proportion of a company's assets that are financed by debt. A higher ratio consequently means that a bigger part of the company is financed with debt (Damodaran, 2002, p. 33; Rist & Pizzica, 2015, p. 66).

Book Value of Equity

The book value of equity shows the value of a company from a pure accounting perspective. The change over time is important, because it helps to get a clearer picture about retained earnings, dividend payments, issuance of new stock, and stock repurchases (Marshall, 2017, pp. 79-81).

5.2 Technical Parameters

Other than fundamental parameters, technical analysis does not take the intrinsic value of the company represented by accounting figures into consideration. Technical analysts believe that prices move in trends, which reflect the investors' opinion about a variety of factors such as economic, psychological, monetary, or political forces (Park & Irwin, 2007, pp. 786-787). Pring (2014, p. 4) defines a trend as *"a period in which a price moves in an irregular but persistent direction"*. Trends therefore have a time span and a direction. The time horizon examined is irrelevant for the factors considered, important are patterns and these emerge in the short run as well as in the long run. That means that we can use technical factors as inputs for our neural network, even if the forecasting horizon is longer term and not day trading. Longer term trends are often referred as primary and intermediate trends. Primary trends are investors' attitudes about business cycle so they last from around 9 months until around 2 years. Intermediate trends are a little more short term and last from a couple of weeks up until 9 months. Depending on the investment horizon you should consider different lengths of trends (Pring, 2014, pp. 3-6).

Trading Volume

Trading volume can tell a technical analyst different things about the trend in the stock market. The most important principle when looking at trade volume is that it typi-

cally goes in line with the development of the price, meaning that the trade volume increases with an upwards trending market and falls in declining markets. It should be mentioned, that a comparison of volume should be done between periods that do not differ in time too much. The reason is that more companies are listed over time and the overall market participation increased in recent years compared to for instance the early twentieth-century (Pring, 2014, pp. 98-99). Zhu, Wang, Xu, and Li (2008, pp. 3053-3054) found that trading volume was not able to increase the forecasting power of their ANN in the short-term term, but was able to improve the forecasting performance significantly in the mid- to long-term. As accounting data is considered, the investment horizon will be mid- to long-term as the relevant data is only available quarterly.

Momentum

Different authors including Grinblatt and Moskowitz (2004), Jegadeesh and Titman (1993), and Fama and French (1996) have studied momentum strategies, which are based on buying past winners and short selling losers. Grinblatt and Moskowitz (2004, pp. 551-558) found in their study, that past winner of different time horizons tend to outperform their peers in subsequent periods both in value-weighted and equal-weighted portfolios.

Additionally, Grinblatt and Moskowitz (2004, pp. 551-557)'s findings show that consistency over time is an important factor to consider, this means the relevant returns are not only the ones from the last period, but several previous periods. The relevance of persistence in a momentum based portfolio strategy has been conformed more recently by Chen, Chou, and Hsieh (2018, p. 888), who discovered that only around *"60% of winner and loser stocks can be persistent winners and loser"*. As LSTM use feedback connections between the different data entries of an individual variable, they are very well suited to remembering longer term dependencies and automatically account for the potential significance of a persistent positive return in the past if it were relevant for the model.

52-Week High and 52-Week Low

$$52 - Week\ High = \frac{Current\ Share\ Price}{52 - Week\ High}$$

When George and Hwang (2004, pp. 2149-2175) compared different momentum strategies in their study, they found that an investment strategy based on 52-week high outperformed the other two strategies in an equally weighted portfolio. The two authors computed the relative valuation of the current stock price and its corresponding 52-week high and ranked the stocks in ascending order and afterwards attributed the highest ranked to their winner portfolio and the lowest ranked to their loser portfolio.

With this simple strategy they were not only able to generate positive returns, but their investment strategy outperformed more traditional momentum strategies in a head-to-head comparison. A potential explanation of the generated profits is according to the authors, that a stock whose price is close to the 52-week high has been surrounded by good news and *"this may be the time when biases in how traders react to news, and hence profits to momentum investing, are at their peaks* (George & Hwang, 2004, p. 2146). When a stock trades close to its 52-week high traders are often subject to the adjustment and anchoring bias, as George and Hwang (2004, p. 2146) puts it: *"traders use the 52-week high as a reference point against which they evaluate the potential impact of news. When good news has pushed a stock's price near or to a new 52-week high, traders are reluctant to bid the price of the stock higher even if the information warrants it."*. At some point in time, the good news however prevails and pushes the stock price above the 52-week mark. Their findings thus suggest that the 52-week high is a more important variable to consider than the past performance. The same holds true for stocks that are trading close to their 52-week low.

Hao, Chou, Ko, and Yang (2018, pp. 169-183) further investigated the outperformance of the 52-week high momentum strategy proposed by George and Hwang (2004) and found that part of it can be attributed to the sentiment of the market participants. This is also the reason why the performance of the 52-week high strategy is superior in periods where investor's sentiments, measured by the sentiment index proposed by Malcolm and Jeffrey (2006), are high.

The importance of considering the level of the current price in relation to the 52-week high and 52-week low is further supported by the findings of Huddart, Lang, and Yetman (2009, p. 30). The authors found that *"stocks breaking outside prior trading ranges appear to earn positive excess returns after the event"*. Their findings are consistent with the theory of Barber and Odean (2008, pp. 812-813) that state, that the return anomaly around the 52-week high or 52-week low can be explained using a model of attention-based decision making. According to the authors, it is impossible for an individual investor to evaluate a large range of stocks as potential investments without the help of the computer. As a result of that, individual investors consider mainly stocks that have gathered their attention recently. This explains why individual investors are net buyers on trading days following both negative and positive extreme returns. Professional investors have more resources available and may run algorithms to do a preselection of stocks. Consequently, institutional investors are less prone to purchases based on attention.

As daily prices or returns are not included in the input variables, a comparison between the 52-week high and the current stock price can still give some insights whether the stock may be undervalued or overvalued based on momentum strategy.

Variable-Length-Moving-Average

In their research, Hendrik and Kalok (1998, pp. 6-7) found that Variable-Length-Moving-Average (VMA) rules can be used as a buy or sell signal respectively in order to create an excess return. The buy (sell) signals are triggered when the curve of the short-term average crosses the curve of the long-term average and its deviation exceeds a specified percentage band. The buy signal is emitted when the short-term curve exceeds the long-term curve, for the sell signal the opposing is true, meaning that when the short-term average drops below the long-term average curve. The notation of an VMA (S, L, B) rule specifies all the necessary factors, S refers to the number of days in the short-term moving average whereas L stands for the number of days in the long-term average. B represents the percentage boundary that triggers the buy or sell signal when exceeded. For example VMA (1, 200, 0) would calculate the VMA using a short period of one day, a long period of 200 days and a band of zero percent. The mathematical equation for a VMA (S, L, B) would be represented by the following term:

$$VMA = \frac{MA(S) - MA(L)}{MA(L)} = \frac{\frac{1}{S}\sum_{i=0}^{S-1} p_{M-i} - \frac{1}{L}\sum_{j=0}^{L-1} p_{M-j}}{\sum_{j=0}^{L-1} p_{M-j}}$$

Whereas VMA $\leq B$ would be a sell signal and VMA $\geq B$ a buy signal respectively. The variables can be determined by using any number, but we will use the main rule used by William, Josef, and Blake (1992), a short term period of one day, and a long period of 200 days. For the training of the neural network a band does not need to be defined, because the input will be the raw percent difference between the two moving averages.

5.3 Economic Parameters

Besides technical and fundamental factors, one should also include some basic macroeconomic variables as input factors of the model. The reason for that is, that as firms operate in the macro environment, they are influenced by business cycles and affected by changes in the macroeconomic environment such as future investment opportunities, consumption, or interest rates. The connection between stock price prediction and macro variables has been researched in depth and some evidence has been found that macroeconomic variables can help forecast future stock prices. It should be mentioned however that there is no strong consensus amongst researchers and it is therefore hard to identify which particular macro variable can reliably be used to forecast stock returns (Rapach, Wohar, & Rangvid, 2005, pp. 137-138).

Amongst many macroeconomic variables the interest rate, rate of inflation, industrial production, and real exchange rate has been found to have a significant impact on stock prices in emerging markets (Chandrashekar, Sakthivel, Sampath, & Chittedi,

2018, pp. 98-99). Analogous results were obtained by Mohapatra and Rath (2015a, pp. 150-153) when analyzing the relationship between the stock exchanges of India, Brazil, and China and the macroeconomic variables real effective exchange rate, industrial production, inflation, and interest rate. Consequently, the input parameter for the ANN will be based on variables in these categories.
Especially in emerging countries such as Brazil, India, and China it is not only important to consider the country isolated but as a participant of a more global economy and therefore has an exposure not only to its own but also to the global business cyles.

Interest Rate

Leigh, Hightower, and Modani (2005, pp. 7-8) found that the interest rate is an important factor when determining the price of the New York Stock Exchange Composite Index. A potential explanation mentioned by the authors is that the interest rate is important in investors valuation of individual stocks as it is used as the discount rate. There is no fixed relationship between stock prices and the interest rate, but generally if the interest rate goes up the stock prices go down, because stocks are worth less due to a higher discounting of future cash flows. In addition to that, the purchase of bonds and holding cash becomes more attractive if the interest rate goes up. The opposite relationship holds also true most of the time (Leigh et al., 2005, pp. 7-8).
While for developed countries studies have shown a negative relationship between stock returns and interest rates, Mohapatra and Rath (2015a, pp. 150-152) found a positive relationship for emerging countries. To put it differently, an increase in the interest rate of the Treasury bill leads to an appreciation of the stock market.
Rapach et al. (2005, p. 158) found the relative long-term governmental bond yield (RGB) to have a predictive power. This is the *"difference between between the long-term government bond yield and a 12-month backward-looking moving average"* (Rapach et al., 2005, p. 158). The authors found that the predictive ability of RGB is independent of the forecast horizon they considered. In other words, it influences both long but also short term stock price movements.

Industrial Production

Among the economic variables Mohapatra and Rath (2015b, pp. 150-153) studied, an increase in industrial production is associated with an increase in stock prices. This study has thus confirmed findings of previous studies such as Huang and Yang (2004, pp. 359-360) who found an indirect causal relationship between the industrial production and stock returns in the US, Japan, and Taiwan. The explanation behind this variable is, that industrial production output is a proxy of economic activity and can thus give insights about the development of GDP. In other words, an increase in industrial production results in higher economic growth and as a result of that higher

overall earnings of companies (Mohapatra & Rath, 2015b, pp. 150-153). The relevance of industrial production on the stock market returns was also confirmed by Humpe and Macmillan (2009, pp. 150-153). In their study the authors considered the long term interactions between the Japanese and US stock returns, and the macroeconomic variables industrial production, CPI, monetary supply M1, and the 10 year US treasury bond. From these variables, industrial production had a significant positive relationship for both countries, the US and Japan.

Retail Sales

McCormack (2007) and Vigna and Horner (2019) report in their article that stock prices dropped, after retailers announced weaker sales. The explanation behind it is that weaker retail sales is associated with a slowdown of economic activity and as a result has an impact on asset prices. The relationship between retail sales and stock return is therefore indirect, because retail sales are seen as an important figure of economic growth.

In addition to that, aggregate retail sales are important, because the companies that are considered in the training phase of the ANN are all retailers (GICS Industry Group: *2550 Retailing*).

Unemployment Rate

The overall state of the economy and the labor market are highly correlated, because an expansion requires more workers in order to meet the additional demand, assuming the productivity is held constant. On the contrary, a decline in the economic output leads to a rising unemployment rate until the economy starts to grow again. The relationship between the economic output and the unemployment rate is amongst the strongest factors and therefore one of the most important coincident indicators of a business cycle. As a result of that, unexpected changes of the indicator can cause larger swings in stock prices (Yamarone, 2017, pp. 69-71).

Often the employment report gives further insights to important indicators including but not limited to average hourly wage, total hours worked, and overtime worked (Yamarone, 2017, pp. 69-71). The IBGE unfortunately stopped at the end of 2015 with the corresponding surveys in Brazil and the data available is thus limited to the monthly unemployment rate (IBGE, 2015).

Term Spread

The second economic variable that was found to be effective in forecasting stock price was the term spread (TSP). This economic factor however showed forecasting capabilities mainly in longer terms (Rapach et al., 2005, p. 158). Similarly Yamarone (2017, pp. 27-29) states that the slope yield curve is *"one of the most accurate forecasting measures*

of economic activity". In other words it is crucial to incorporate several term spreads as input variables for the ANN in order to include information about the term spread and the yield curve. Interest-rate spreads are positive under normal economic conditions, because investors demand a higher risk premium for longer maturities. On the contrary, an inversion of the yield curve, meaning negative interest-rate spreads, is associated with economic slow downs (Yamarone, 2017, pp. 27-29). For that reason, the interest rates of different maturities will be included as input variables.

Money Supply

There has been some evidence that an increase in money supply is associated negatively with subsequent stock returns especially for longer holding periods. A potential explanation of this relationship is that an increase in money supply leads to a decreased risk premium, because current prices go up and future expected returns decrease. This is due to the lower interest rate that is associated with monetary expansion (McMillan, 2017). Similar findings have been found by Maio (2014), he however found that value stocks, defined by a low P/B ratio, are more responsive to changes in money supply compared to growth stocks. As the input variables include fundamental data, it is beneficial to include money supply as well, because of potential cross-relations. On the contrary to the two previously mentioned study, Rogalski and Vinso (1977) states in his research, that information about money supply is priced in already in the stock price and therefore only has an immediate effect of stock returns.

Foreign Direct Investment (FDI)

In his research Sanvicente (2014, pp. 101-104), found that cash flows from direct investments influence the returns of the principal Brazilian stock index BMFBovespa. A positive net inflow of capital may therefore be related with an overall increase of the stock returns. The research further showed that a depreciation of the BRL and positive stock market returns in developed countries lead to an increase of capital flows from outside Brazil. Likewise Reis, Meurer, and Da Silva (2010, pp. 1359-1360) found that these three variables are able to explain a significant part of the Bovespa return in the time period of 1995 until 2005. In this context it should be mentioned, that FDI significantly change throughout business cycles and decrease significantly during recessions, which impacts stock returns (Stoddard & Noy, 2015, pp. 396-398). As a result of that, netflows of FDI will be considered as an input variable to reflect the demand of outside investors and the stock returns associated.

Inflation Rate

Depending on the country, the authors found that the inflation rate, measured by the increase in consumer prices, was able to explain some of the subsequent stock price

movements in the following 12 months. The magnitude and statistical significance however strongly dependent on the respective country (Rapach et al., 2005, p. 158). Confirming the findings of Rapach et al., Kaul (1990, pp. 311-318) found a substantial negative relationship between stock returns and changes in the inflation expectations in the four developed countries he analyzed. Similarly, Humpe and Macmillan (2009, pp. 115-116) found a negative relationship between inflation measured by the Consumer Price Index CPI and subsequent stock movements. Analogous to Rapach et al. (2005) the statistical significance varied amongst the examined countries. In the case of the United States, the relationship was statistically significant, but in Japan the relationship between the CPI and stock returns was not statistically significant. In the context of this research, it is therefore important to look at the relationship between inflation and stock prices in Brazil particular. In their recent study, Chaves and Silva (2018, pp. 530-538) explored exactly this relationship and found that the real IBOVESPA returns exhibit a negative correlation with the inflation forecasts of the proceeding 12 months. For this reason, the inflation outlook will be included as a variable to train the ANN. The data will be obtained from the Focus Survey (Focus - Relatório de Mercado), a weekly survey conducted by the Brazilian Central Bank (BCB) that collects market expectations from industry industry professionals (Banco Central do Brasil, 2019).

Exchange Rates

Foreign exchange rates have a big impact on the development of the overall economy of the country. Consequently, exchange rates have a direct and indirect influence on the stock price of companies. The direct influence is that the value of a company is determined by the present value of future cash flows, consequently the expectations of the exchange rate is important especially for international investors. Indirectly it influences the company's business environment and the exposure of the cash flows to currency fluctuations is a risk factor that investors consider. The magnitude and direction of the causality between the exchange rate and the stock returns however depends on individual countries and may be due to different stages of economic development, regulations, or other factors (Hamori, 2003, pp. 62-79). In the case of India, Brazil, and China, Mohapatra and Rath (2015a, pp. 150-151) and Gay (2016, pp. 4-5) have independently shown a positive relationship between the Real Effective Exchange Rate (REER) and stock prices. In other words, a depreciation of the country's domestic currency is associated with an appreciation of the local stock market index. As it would exceed the scope of this paper to include all exchange rates, only the exchange rates of the most important trade partners of Brazil will be considered. According to trade volume in USD, China was the most important trade partner in 2017, followed by the United States and then Argentina (WITS - World Integrated Trade Solution, 2017). Hence the exchange rates BRL/CNH, BRL/USD, BRL/ARS will be used as data input for the

ANN.

GDP Growth Rate

One of the most common ways to measure the overall economic activity of a country and the business cycles related to it, is the gross domestic product, GDP. Accordingly changes in the GDP help economists determine the overall business cycles and in which stage the economy currently is in. Even though the GDP gives a good impression about the overall health of the economy, its forecasting power is limited, as it is only released on a quarterly basis (Yamarone, 2017, pp. 17-19). In order to estimate the development of an economy, the real GDP growth rate rather than the absolute value should be considered. Many studies show a relationship between the GDP growth and an appreciation of the stock market. Alexius and Spang (2018, pp. 122-123) show in their study that GDP and stock prices cointegrate on a domestic level in all of the countries included in their research. It can therefore be said that the development of the GDP is an important aspect when looking at what economic variables drive stock prices in the long run (Humpe & Macmillan, 2009, pp. 122-123).

Consumer Confidence Index (CCI)

Consumer confidence is an important factor in the prediction of economic activity, because it gives insights about the current economic perception of the population. For the construction of an ANN the relationship between CCI and the stock market is important. Fisher and Statman (2003, pp. 116-120) found that changes in the CCI and investors sentiments exhibit a positive relationship, meaning that they tend to change in a similar direction. The relationship mainly holds true among individual investors. One possible explanation is, that high stock returns make individuals feel wealthier, which results in an increased level of confidence. It should be said though that an increase in income has a comparably higher impact on consumption than an increase in wealth. The authors not only found a positive relationship between investor's sentiment and the CCI, but also between the CCI and the returns of the SP 500 index (Fisher & Statman, 2003, pp. 116-120). The impact of investor's sentiment is bigger amongst stocks of small firms, one possible explanation is that small stocks are mainly monitored by individuals rather than institutional investors (Lee, Shleifer, & Thaler, 1991, pp. 81-84).

For the development of the model we will use the Índice Nacional de Expectativa do Consumidor (INEC), which is the consumer confidence index of Brazil. The index is published quarterly by the Confederação Nacional da Indústria (CNI) and INEC consists of six different elements. It incorporates future expectation of inflation, unemployment, and personal income and information about the current financial situation including indebtedness and the intention to purchase a good with higher value (CNI,

2015, pp. 12-14). In order to use the (INEC) as an input variable the relative change will be considered.

6 Data Preprocessing

Figure 5: Process Overview Construction LSTM, own illustration

The following sections will outline the necessary steps to create and optimize the LSTM. All of the steps were implemented using a Jupyter Notebook, Python version 3.6.4, and several Python packages.
As acronyms of the variables are used in the code, a list with the respective names and descriptions is available in the appendix in table 4.

6.1 Data Collection Import

All of the data required to calculate the input variables mentioned above were downloaded from the Bloomberg Terminal using the Excel Add-In (Bloomberg, 2019) or Thompson Reuters Eikon using their Python API (Eikon, 2019) and stored locally in order to avoid redownloading the data every time the algorithm is run. Most of the variables were directly available on Bloomberg and could thus be imported without calculations into Python. Only the 52-Week High and Low, and the Variable-Length-Moving-Average needed to be calculated first before merging with the main dataframe. The reason for that is, that these values change daily as they always consider a certain, moving time frame in the past. Upon calculating the necessary values, the data frames were merged based on their quarterly dates in order to proceed with the analysis.
It should be mentioned, that for the stock price forecast the relative log-return will be considered. The reason for this is, that a neural network should not be used to forecast the actual price of the next period but rather the relative price change. Forecasting actual prices is not only prone to errors as the price might be outside of the historic range, but relative price movements are more important when forecasting stock price changes (Marwala, 2013, pp. 13).

6.2 Data Preparation

After downloading the data, it is important to preprocess the datasets. According to Yu, Wang, and Lai (2006, p. 222) and Azoff (1994, pp. 21–22) the first step in this phase

is *Data Inspection*, which has the objective to find potential issues with the data set. The two main features that are considered are the quality and the quantity of the data set. When it comes to data quantity, the amount of available data is restricted to the release of accounting data, which is maximum every quarter. As Bloomberg's Database (Bloomberg, 2019) is not able to provide reliable data before 1998 for the Brazilian stock market, the quantity is further restricted. In order words, data entries per variable are restricted, so there may only be too many different variables that have been used as an input. As there is no general rule for a data minimum for an LSTM, all data available will be used. When interpreting the results at the end, it should however be kept in mind, that the data set available may be too small and the model may not be able to generalize sufficiently. This conclusion can however only be drawn upon analyzing the performance of the LSTM. The quality inspection, also known as cross-validation, is done manually using the actual values on the Bloomberg Terminal (Bloomberg, 2019) including the linked annual or quarterly reports respectively (Azoff, 1994, pp. 21-22). In addition to that, the scatter plots(available in section C) were individually analyzed to see whether there may be suspicious data entries.
Following *Data Inspection* is *Data Preprocessing*. There are many issues that should be addressed when preprocessing the data, the two most relevant issues that are present in the dataset used are (Yu et al., 2006, pp. 222-224):

6.2.1 Missing Data and Data Repairing

Accounting data is released quarterly and the majority of the variables chosen are available since the listing of the company on the IBOVESPA. Even though companies' report their data quarterly, the dataset has some missing entries that should be addressed.
The first one is related with negative earnings and discontinuation of dividend payment. Due to that, ratios such as the Dividend Yield, and the P/E Ratio cannot be computed anymore. Based on that, the missing values of the the 12-month Dividend Yield were replaced with a 0. The reason for that is, that the decision itself, not to pay dividends in a financial year/quarter contains information that is beneficial for the algorithm. The same holds true for the missing entries of the P/E Ratio, if a company has negative earnings and thus no P/E Ratio, it also sends important signals.
Second, are the missing values that can be calculated. They are missing, either because of some error in the database and thus were not downloaded correctly when creating the Excel spreadsheets with the Bloomberg Add-In or the company actually did not report a certain figure. If the ratio is available or can be calculated, it was added manually using the values available on the Bloomberg Terminal (Bloomberg, 2019). In the case the values were not available on the Bloomberg Terminal, the values were treated

as missing. In such a case, there are three options: *"removing the rows with the missing values, mark and learn the missing values, and mask and learn without the missing values"* (Brownlee, 2017).
For the dataset used in this analysis, the missing data entries were addressed in two ways. First, the starting point of the time series was determined by the availability of data. For instance, as the company *B2W - Companhia Digital* was established by a merger of *Americanas.com* and *Submarino*, there are pro forma balance sheets available of the time before the merger, but when analyzing the historical data after the merger, it can be concluded that they are of no value as the values are too different. In this case only the data after the merger will be used. The amount of data used as an input always faces one important problem described by Moody (2012, p. 345) as: *"using a short training window results in too much model variance or estimation error due to noise in limited training data, while using a long training window results in too much model bias or approximation error due to nonstationarity"*. This is also called the noise/nonstationarity tradeoff.
Second, individual missing data entries were calculated using KNN Imputer. K-nearest neighbor is an algorithm that estimates the missing data based on data points that are close to it. A KNN Imputer thus considers the relationship between different variables in order to estimate the missing one. Even though k-nearest neighbor is able to estimate missing data entries fairly well, its performance decreases when estimating values for a time series with a global trend. The reason behind that is that the values may be out of the historic range and it thus makes sense to remove the trend with strategies such as differencing (Martínez, Frías, Pérez, & Rivera, 2019). Despite the findings of Martínez et al. (2019), the KNN Imputer will be used before differencing. The reason is, that in the case at hand k-nearest neighbor is not used to actually forecast the next value of a time series but just missing values. As a result of that, the algorithm will not need to forecast values out of the historical range and can be seen as suitable for this purpose.

6.2.2 Trending Data

Even though LSTMs do not require stationarity of data, trending data can negatively influence the performance and efficiency of neural networks. As a result of that, Yu et al. (2006, p. 223) recommend to do data detrending. An effective and widely used way is to take the difference or the log-difference of the values.
When the dependent variable is a time series one always has to verify that the time series displays covariance stationary. In other words, the first and second moment need to be constant through time. In order to test the stationarity of the means, variances, and covariances of a time series a Dickey-Fuller Test is used (McNelis, 2005, pp. 59-60).

As the number of input variables is large it is difficult to verify stationarity visually, for that reason the Augmented Dickey-Fuller Test was performed. The results of the analysis can be found in appendix C.1. The Augmented Dickey Fuller Test is a unit root test and has the Null Hypothesis (H_0) that the time series contains a unit root and thus is non-stationary. A rejection of H_0 leads to the acceptance of the alternative hypothesis H_1, which indicates that the time series is stationary. The more negative the ADF Statistics is the higher the likeliness of a rejection of H_0. When comparing the ADF statistics to the critical values stated in appendix C.1 , one can identify at which level of significance one can reject H_0 (Neusser, 2016, pp. 147–148; Brownlee, 2016a).
In the case at hand, many variables fail to reject H_0 and are likely to be non-stationary time series. A possible explanation of the non-stationarity is, that inflation has a direct or indirect influence on many of the values and other values increased or decreased over time, because companies changed operative aspects of the business. Baltagi (2011, pp. 383-383) differentiates between two types of stationarity, trend stationarity, and difference stationarity. Among which trend stationary describes a time series with a deterministic trend and difference stationary is defined as a random walk with a drift. The Augmented Dickey Fuller Test gives insights in whether a trend is difference or trend stationary and based on that the correct form of detrending or differencing can be chosen.

Company-Level Data

The analysis of the results from the Augmented Dickey-Fuller Test show, that depending on the company some variables may be stationary, but for others it may not be stationary. As it is important to treat all companies the same, the decision was taken per variable. The choice was made on the basis of the individual Augmented Dickey-Fuller Test for each company, but also across companies. Stationarity was assumed for variables for which, H_0 can be rejected at a significance level of less than 5%. Upon analyzing the input variables the following variables are likely to be non-stationary and one should perform differencing on the values to achieve a better performance of the LSTM: *Price to Sales Ratio, Current Ratio, Quick Ratio, Inventory Turnover, Receivables Turnover, Asset Turnover, Net Operating Profit After Tax, Earnings per Share, Operating Income, Free Cashflow, Cashflow from Operations, Short- and Long-term Debt, Interest Expenses, Capital Expenditure, Sales, Gross Margin, Operating Margin, Return on Asset, Net Operating Assets, RNOA, Return on Common Equity, EBIT, Interest Coverage Ratio, Debt to Equity, Debt to Capital, Long-term Debt to Capital, and Total Equity*. Most of these variables probably are non-stationary, because of either the direct impact impact of inflation, or the advancement of business operations.
For the input variables that have been identified as potential non-stationary, a scatter plot has been created to further examine stationarity visually (scatter plots are available

in appendix section C.2). Comparing the scatter plots of each variable amongst companies showed that the following input variables should be considered non-stationary, because they either have a trend or seasonality (Brownlee, 2016a): *Asset Turnover, Net Operating Profit After Tax, Operating Income, Short- and Long-term Debt, Interest Expenses, Sales, Net Operating Assets, EBIT, Total Equity*. The other input variables do not seem to have clear seasonalities or trends. Some of the companies may have a trend in an input variable, but this can most likely be attributed to alterations in business operations. In other words, if an input variable was considered as stationary despite a high ADF and a visible trend/seasonality, it is because the trend in the variable is company rather than variable specific.
After differencing the time series data, another Augmented Dickey-Fuller Test was performed. Comparing the two data tables and figures shows that differencing was successful and most of the variables are now stationary according to the results of the Augmented Dickey Fuller Test. As mentioned before stationarity is not a requirement of the input variables of an LSTM, despite that it is used to improve the network performance and efficiency. As a result of that, it is sufficient that most, but not all of the variables are stationary at first difference.

Economic Data

Equivalent to company-level data, the economic data was analyzed using the results of the ADF of the raw data and the scatter plots. As economic data has fewer variables, a scatterplot for all of them was created (see appendix C.2.3).
After examination of the individual ADF statistics and comparing the results to the respective scatter plot, only two time series can be considered as stationary at significance level of 5% : *BZFDTMON Index, BZPIIPCM Index*. These two variables represent the net foreign direct investments into Brazil in USD (BZFDTMON Index) and the Brazilian growth rate of the consumer price index (BZPIIPCM Index) respectively. Based on this analysis, all other time series will be differenced.
After differencing the data once, most time series become stationary. Even though in some of the cases, for instance money supply M1, and M2, the time series still be may non-stationary, the data will not be further processed. The reason for that is, that stationarity of data input is not a requirement for an LSTM. The process is done in order to improve the overall performance of the model.

6.3 Transformation of the Data

Before starting with the training of the neural network it is important to transform the input data in into a range matching the range of the activation function. The reason for that is, that activation functions are centered around a specific value of the output

range and it is required to have a similar range for the input variables in order for a neural network to work efficiently. In addition to that, depending on the activation function, a significant amount of information will be lost. For instance in the case of logsigmoid or transigmoid neurons, scaling is crucial, because otherwise the neurons will set relatively large numbers to 1 and comparably low numbers to 0, which leads to a big loss of information, because most of the values will be assigned to the extremes of the functions (Mandic & Chambers, 2001, p. 23; McNelis, 2005, pp. 64-65). Transforming the data is therefore crucial not only to improve the performance of the model, but also increase the learning speed of the network (Brownlee, 2019, pp. 27-28).
There are several approaches that can be used to perform such a transformation (Mandic & Chambers, 2001, p. 23):

1. **Normalization**
   Takes input data and converts it into values within the range of 0 and 1 or any other selected range. One of the prerequisites of normalization is the knowledge of the minimum and maximum of the input data or at least an accurate estimate thereof. In many cases the range of the input values can be estimated from the data available, but in the case of time series with a trend such an estimate may be difficult and therefore inaccurate. In this case, normalization is not the best method for data transformation. The issue arises if new values exceed the maximum or the minimum and the conversion does not result between 0 and 1. In such a case, one could remove such data or limit it to the predefined maximum or minimum value (Brownlee, 2019, p. 28).

2. **Rescaling**
   Rescaling refers to the transformation process in which the input data is divided or multiplied by a constant, and a constant is also potentially subtracted or added to the individual data points. A real world practical example would be the transformation from Fahrenheit to Celsius (Mandic & Chambers, 2001, pp. 23-22).

3. **Standardisation**
   In order to get reliable results, standardisation requires the distribution of the input values to fit the Gaussian distribution and have an accurate estimation of the mean and the standard deviation. Often these values can be estimated using the training data set, especially if it is large enough. In many cases the standard deviation and the mean are more robust measures compared to the minimum and the maximum and in such cases are thus the preferred method of data transformation (Brownlee, 2019, pp. 29-30).

In this context it is important to point out that each data series should be treated separately with the right method chosen (Brownlee, 2019, p. 31). In other words, for each

input variable the right data transformation method has to be chosen. In addition to that, it should be noted that transformation of input values has a direct effect on the training process of the algorithm. The reason for that is, that steepest descent algorithms react sensitively to scaling, as changes in the weight are proportional to the gradient and the input data (Mandic & Chambers, 2001, pp. 23-22).
As mentioned previously it is important to transform the input data of the ANN for two reasons. First, it needs to be scaled into the range to match the input range of the respective *activation function* to keep the largest information content possible. Second, it is crucial for the performance and speed of the ANN. As the hyperbolic tangent activation function (Tanh) will be used, the data will be transferred into the range [-1,1] in order to match the input result of the activation function. After analyzing the distribution of the different variables it became clear that standardisation is not possible, because a normal distribution cannot be assumed for most of the input values. The most suitable way is therefore normalization. The function *MinMaxScaler* from the Python package sklearn will be used in order to transform the features column-wise into the desired range (Pedregosa et al., 2011).

7 Designing the Neural Network

As outlined in section 5 the goal is to use many different input variables in order to forecast the price change of a stock in the subsequent quarter. The LSTM model used for Multivariate Time Series Forecasting is called *Many-to-One Model*, as many input variables (series) with several time steps are used to forecast the output variable (Brownlee, 2019, pp. 57-60). The *Many-to-One Model* is not only suitable, because the model should consider different input variables, but also because it is important to consider the change of an input variable through time.
Figure 6 graphically illustrates a *Many-to-One Model* with many input sequence values (X_t), several hidden states (u_t), and one output sequence value (y_t), whereas t indicates the time step.

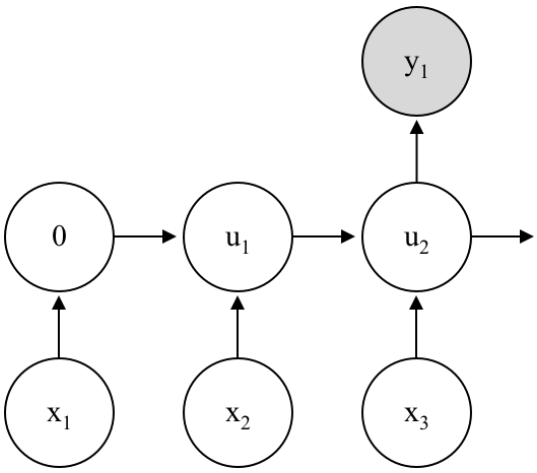


Figure 6: Many-to-One Prediction Model, based on Brownlee, 2019

Stock price forecasting is a challenging prediction task and therefore requires several hidden LSTM layers with numerous memory cells. This decision is further supported by the assumption that deep, hierarchical models are exponentially more efficient and faster in depicting more complex functions than shallow networks (Brownlee, 2019, pp. 78-79; Pascanu, Gulcehre, Cho, & Bengio, 2013, pp. 2-3). Put differently, as stock prices are based on countless factors that are interdependent a deeper model will be more efficient in approximating the forecasting function.
The five retail companies will be treated as separate samples and used to train the model. This is necessary, because the time series of each company should be considered as an independent sequence of data and cannot be mixed with data from the other companies. The sequence is however never passed in its entirety, but rather batch wise. This makes the computation and optimization of the LSTM faster. In the case at hand, the datasets are not that big and could theoretically be passed through in one batch. Nevertheless, batches were created, because the relevance of accounting data from more than a couple of quarters ago might not be that relevant anymore.

Model Summary

The neural network is programmed using the Python Library Keras (Chollet et al., 2015) and a hyperparameter optimizer, that was programmed according to the idea of random grid search outlined in the paper of Bergstra and Bengio (2012).
As mentioned previously, Dayhoff and DeLeo (2001, p. 363) found in their research, that a neural network with only two hidden layers with trainable weights, is able to approximate any nonlinear function. As a result of that, the ANN will have up to two hidden LSTM layers based on Hochreiter and Schmidhuber (1997) and one Dense layer to change the dimension of the time series back to one, because the ANN makes a single value prediction by forecasting the subsequent stock return. The activation function of the last neuron is either a Tanh, because of its output range [-1,1] or no acti-

vation function, because the output range after the Dense layer is already in the range [-1,1]. The model summary can be found in table 1 below.

Layer (type)	**Activation Function**	**Output Shape**	**Param #**
lstm_11 (LSTM)	Tanh	(None, 64)	31744
dense_21 (Dense)	None	(None, 1)	2
Total params:		31'809	
Trainable params:		31'809	
Non-trainable params:		0	

Table 1: LSTM Model Summary

7.1 Training and Sampling

In order to assure an unbiased estimate and avoid over-fitting, the input data is split randomly into two non-overlapping subsets of data: the training set, and the validation set. Each of these subsets have a different role in the process of building a neural network. As the name implies, data from the training set is used to train the model in order to estimate the weights. After that the validation set, also known as test set, is used to test the accuracy of the estimators (Dayhoff & DeLeo, 2001, p. 1626). The splitting of the data into two parts is very important in order to test the model on data that has not been used to train the model. This gives a true accuracy estimation, because the data used for the estimation has not been used for anything before. If such a data split is ignored, the model will function well on the available data, but it's performance will decrease in the case of new data due to over-fitting (Marwala, 2013, pp. 11-12). It should be mentioned, that all of the individual samples are split into a test and validation set. Training the LSTM across different companies further helps to avoid overfitting, because they all have slightly different relationships between the data points.

7.1.1 K-Fold Cross Validation

While training the model it became clear that the datasets even when aggregating the five companies is too small to avoid overfitting of the model and a classical train-test split is not the best choice. For that reason, k-fold cross validation was used to replace simple train-test split. Cross-validation describes a resampling procedure that can be used to estimate the out-of-sample forecasting accuracy of a model even if the data sample is small. K-fold cross validation splits the sample data into k subgroups, also referred to as fold, that are then randomly used to train and test the model (James, Witten, Hastie, & Tibshirani, 2013, pp. 180-184). In the case of time series it is impor-

tant that between the iterations the order of the individual folds get shuffled, but the data inside a specific fold remains unchanged. This helps to maintain the relationship between the data entries of a specific fold. In the ANN that is developed, it is further important, that the folds are created within the specific company data and not across them, because this would mix data points of two samples and train the algorithm on wrong relationships.

7.2 Hyperparameters

When programming a neural network, you have two types of parameters: First, hyperparameters, which are elements that you have to set before the model are created. In the optimization process, neural networks will not change these values, because they define the fundamental outline and structure of the network itself. On the contrary, parameters which are coefficients of the model are optimized and updated during the learning process of the ANN. These parameters are chosen based on the underlying optimization strategy. In the case of stock price forecasting, the chosen metric to be minimized is the mean absolute error (MAE) between the predicted and the actual quarterly log-return. The MAE thus gives insights about the forecasting power of the ANN.
When compiling the model in Keras, several arguments can be defined. The three arguments that will be specified is the *optimizer*, the *loss function*, and the *metrics*. The metric that should be tracked are both *mean absolute error (MAE)* and *root-mean-squared-error (RMSE)*. These two metrics allow us to evaluate the forecasts of the LSTM (Neusser, 2016, pp. 249-250). The mean absolute error is chosen as the loss function, because it is less sensitive to outliers than the root-mean-squared-error. In addition to that, the number of epochs was fixed to 50, because in the initial simulation the improvements after around 50 epoch were marginal. These are the only hyperparameters that are fixed, for the rest of the hyperparameters a hyperparameter optimizer function based on random grid search is used in order to select the best model in a systematic way.
The hyperparameter optimizer is built on top of the Keras model and instead of the actual parameter, a Python dictionary with the corresponding list of hyperparameter options is inserted. This allows the optimizer to change the hyperparameters randomly, estimate its accuracy, and optimize the LSTM accordingly. In the end of the optimization, the model is specified according to the suggested hyperparameters.
The input range of the hyperparameter optimizer are chosen manually, based on some initial manual tests of the LSTM.

Parameter	Range	Increment
Batch Size	[8, 16, 32, 64]	8
Sequence Length	[4, 8, 12, 16]	4
Neurons of Hidden Layers	[8, 16, 32, 64]	8
Number of Hidden Layers	[1, 3]	1
Dropout Rate	[0.05, 0.2]	0.05
Learning Rate	[0.001, 0.01]	0.001
Optimizer	Adam, Nadam	
Loss Function	mae, mse	
Last Activation Function	Tanh, None	

Table 2: Input Range for Hyperparameter Optimization

7.2.1 Hyperparameter Optimization

When programming an ANN, hyperparameters are often stil optimized manually. There are however more systematic ways to optimize ANNs, which are less prone to biases. One of the methods is using a random grid search algorithm, that randomly chooses the hyperparameter from a given range and optimizes the parameters of the model using the chosen values. After that the result of the model is stored, compared with the other randomly chosen combination, and used to evaluate the best combination of the hyperparameters. The choice of the random grid search is based on the research of Bergstra and Bengio (2012) who showed that random optimization is more efficient than regular grid search and requires significant less computation power.
One of the hyperparameters that needs to be chosen are the *optimizers*, that fine tune the LSTM itself and are used to find the minimum of the loss function. According to Kingma and Ba (2014, pp. 9-10) the suggested Adam optimizer is an *"computationally efficient algorithm for gradient-based optimization of stochastic objective functions"*. The algorithm can be used for a wide range of applications in machine learning and is suitable for problems with high-dimensional parameter or large datasets. Dozat (2016) developed adam further and incorporated Nesterov Momentum into the algorithm. As the Hyperparameter Optimizer can include several *optimizer*, both Adam and Nadam are included and the one with the better performance will be used.

8 Results and Discussion

The best configuration of hyperparameters for the LSTM based on the outputs of the k-fold cross validation are:

Hyperparameter	Configuration
Batch Size	16
Sequence Length	16
Epoch	50
Neurons of Hidden Layers	64
Number of Hidden Layers	1
Dropout	0.2
Learning Rate	0.01
Optimizer	Nadam
Loss Function	MAE
Last Activation Function	None

Table 3: Best Hyperparameter for optimized LSTM

The hyperparameters were found using a randomized grid search and then optimizing the parameters using k-fold cross validation to minimize the objective function, MAE. Comparing the results to initial manual trainings runs clearly show that the performance of the ANN improves as the model complexity increases. This intuitively makes sense, because stock prices are influenced by many different factors and their relationship is complicated. That being said, a more complex model increases the risk of overfitting.

8.1 Performance

After optimizing the ANN with randomized grid search using k-fold cross-validation the best five configurations were used to train the ANN again for 50 epochs. The reason for that is, that k-fold cross validation helps us to identify the best hyperparameter configuration for the ANN and therefore shows, which configurations should generalize well on out-of-sample data.
After selecting the best hyperparameter configuration, the model still needs to be trained. The result of the training of the models is shown in figure 7, which shows the development of the MAE for each of the configurations. Based on this graphical output the best configuration is chosen to make the forecast of the stock prices.
In this case, configuration of *Rank 5* was the best and will be used to forecast the stock prices in section 8.2.

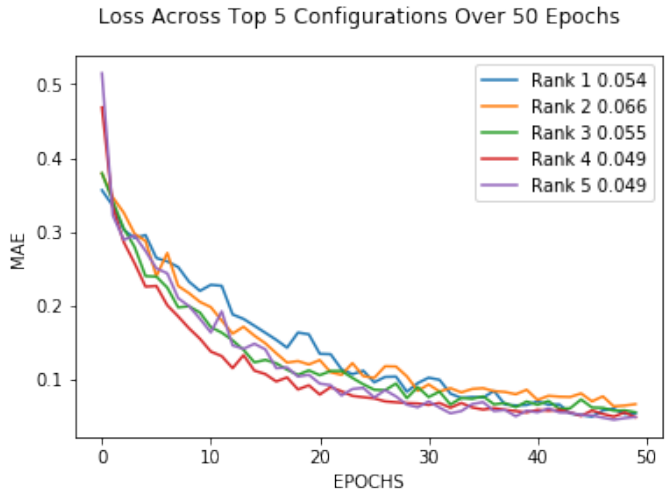


Figure 7: MAE over Epochs

8.2 Return Forecast

In order to evaluate the forecasting ability of the ANN, a comparison between the actual and the predicted value has to be drawn. Figure 8 shows the predictions of all of the data entries. The results show that an ANN can establish a relationship between the variables and forecast the stock returns of the subsequent quarter relatively well. It however also becomes visible, that the optimization process led to an overfitting of the model to the dataset available. The aforementioned is one of the largest restrictions in this research. Even though this concern was addressed in the optimization process by using a k-fold cross-validation, it increased generalization to a certain extent, but most certainly not enough to be able to make stock price predictions for new data points. The presence of overfitting is further confirmed by the large MAE of the best k-fold validation before training the model using the determined configuration.

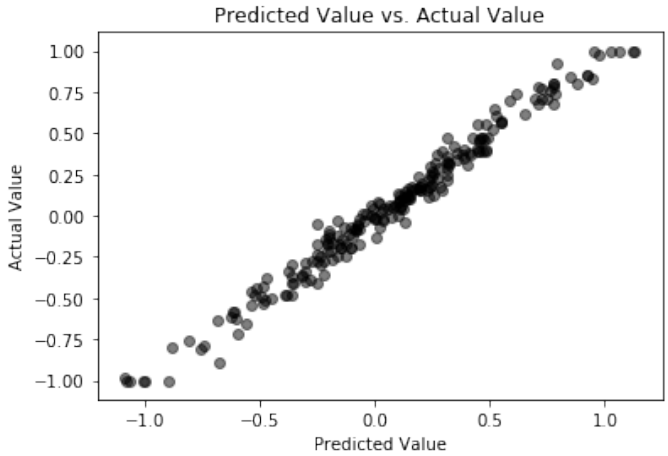
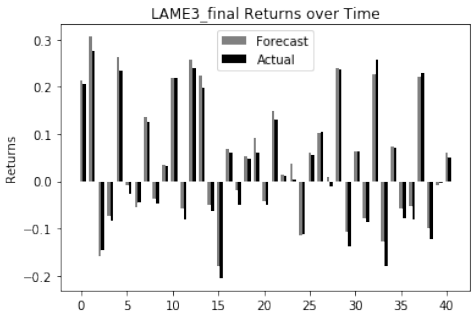


Figure 8: Predicted Values vs. Actual Values

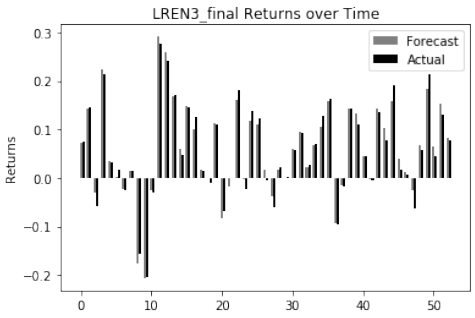
Figure 9 further breaks down the forecasts to the individual company level and

shows an overview of the forecasted and actual stock return for the time horizon studied. The forecasts were done with the model created in the previous sections. When comparing the forecasts between the different companies, one can see that the model has a better performance for some firms than others. This can further be confirmed with the MSE scores of the companies. As a result, it can be said, that an ANN for stock price forecasting should ideally be optimized at the company rather than the industry level. This was however not possible, because of data availability. In addition to that, when looking at a specific company, one might be able to forecast the prices more accurate, as there are still major differences amongst them in their business, their strategy, and thus also what variables impact stock prices. This could however also result in an overfitting of the model and the forecasting power of the model on nonseen data might suffer.

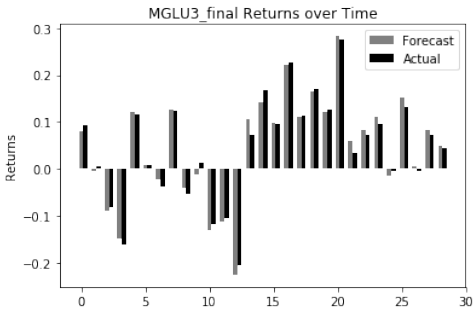
The ANN forecasts the price change mostly in the right direction, but depending on the company the forecast magnitude is off. It can therefore be said, that the usage of such an ANN would be more to assist an analyst to confirm his results rather than replace a detailed company-level analysis. Besides that, the overall forecast might be integrated with a stock price forecasting model using more immediate data inputs such as tweets, blog posts, news, or investor sentiment. Such a combination of immediate and longer-term data might be able to recognize big changes that are caused by such a single unforeseeable event.



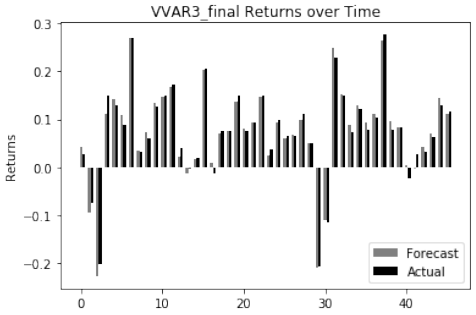
(a) LAME3



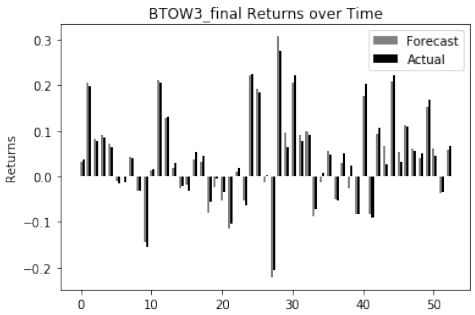
(b) LREN3



(c) MGLU3



(d) VVAR3

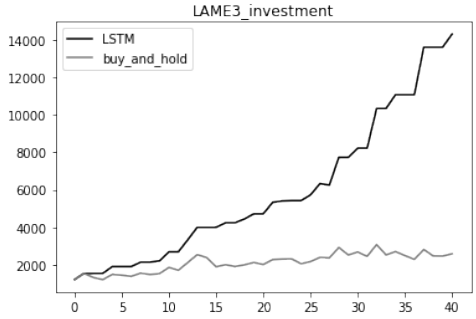


(e) BTOW3

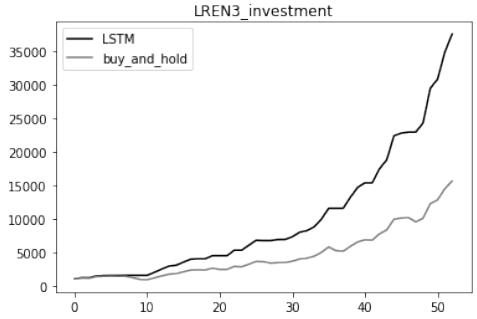
Figure 9: Comparison of Forecasts and Actual Stock Returns

8.3 Investment Comparison

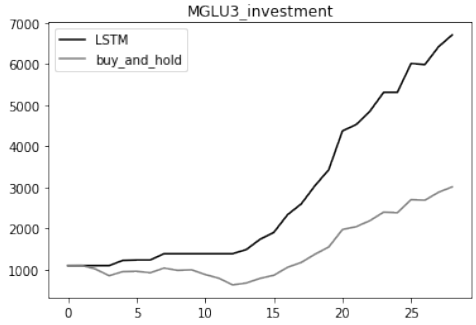
In the following section a comparison between a buy-and-hold strategy and a strategy that buys or sells shares based on the forecast of the ANN will be done. It is important to mention, that in this return forecast no transaction costs have been included and the return of the algorithmic approach is therefore overestimating the real return of implementing such a trading strategy. In addition to that, the investment strategy is long-only, meaning that it does not short-sell companies if the forecast is negative.



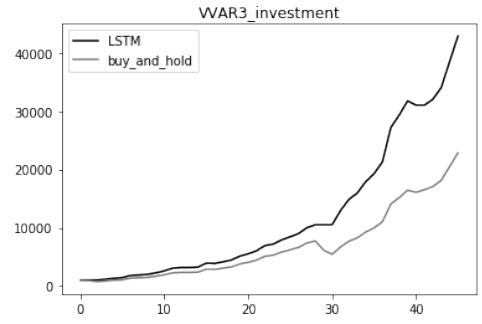
(a) LAME3



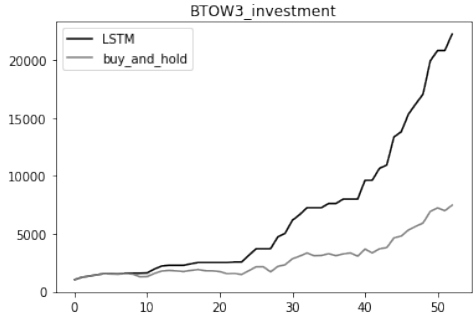
(b) LREN3



(c) MGLU3



(d) VVAR3



(e) BTOW3

Figure 10: Comparison of Investments

As one can see in figure 10 the ANN was able to outperform the buy and hold strategy in all cases. When doing this comparison one however needs to consider again, that the ANN is subject to overfitting and trading according to the forecast of the model will most likely not result in such accurate forecasts. That being said, it is likely that the model can give some assistance to the financial analysts to incorporate the price forecast of an ANN in the decision making process.

8.4 Limitations

The biggest limitation of the forecasting power of the LSTM is caused by the restricted availability of input data. In total there are on average around 50 time steps for each input variable distributed for each of the five companies. Even though this yields almost 15'000 data points, the availability for each variable can be considered as too little.

During the training phase this was a big challenge, because having only a limited amount of data available, ANN tend to overfit the model and the target function is thus not able to generalize well on new data. This means the model performs very well in the dataset available, but will have poor performance with out of sample data - data not seen by the model while training. As mentioned before, the issue of overfitting was addressed by including k-fold cross validation (Brownlee, 2016b). Despite the implementation of k-fold cross validation, the ANN optimization overfitted the model and it is most likely going to have a poor performance on unseen data. A possible way to address this problem in the future, would be to increase the size of the dataset is by including all constituents of Bovespa Index. As the different industries have significant differences in the accounting ratio a training across all companies would probably lead to a bad overall performance and in order to compensate for that, some industry specific component should be integrated.
Another important factor to consider is that the algorithm was trained in an ideal environment, meaning that all of the 59 chosen input variables were available at the end of each quarter. This is unlikely to happen in reality, because companies release their quarterly results a long time after the ending of the calendar quarter. Similarly, economic data is not released at the end of the time period, but rather after some time. If the ANN would be used in reality to forecast stock prices, variables would be available at different points in time and there would be the risk that the information of previous variables is already priced in when the rest is released. In other words, even if the trading strategy could be used in theory to create abnormal returns, it is unlikely that it could be used in reality due to the different release dates of the variables.

9 Conclusion

There is no doubt that AI, especially deep learning is already playing an important role in stock market predictions and its importance will continue to increase in the near future. As seen in the short overview of the current market situation, we are still in a transition, meaning that there are some early adaptors, but there is still a significant amount of asset managers that do not consider artificial intelligence or machine learning in their stock selection process. People that use it, currently use it to assist them in the decision making process rather than taking the decisions for them. It will be interesting to see how the picture changes over the next couple of years, as some of the people contacted replied, that they will start using AI/ML soon in the investment process. Of course, there are many ways to incorporate AI/ML into the selection process and companies will probably try different approaches especially in early stages.
Predictions of AI/ML models may not necessarily be used to replace traditional fi-

nancial analysts, but could be of assistance to them. Similar to how Ray Dalio (2017) describes the investment strategies at Bridgewater Associates, they forecast developments themselves and see if the computer comes up with something similar. If the team forecasts something different than the algorithms, this would mean they probably overlooked something. The reason why the inclusion of AI models in financial analysis will become more important is due to the computer's ability to simultaneously process almost an indefinite amount of variables. The human brain's ability is limited to processing only four variables at the same time and figuring out relationships amongst a large set of variables is therefore impossible (Halford, Baker, McCredden, & Bain, 2005). The LSTM used for this paper for instance considered the development of 59 variables through time in order to forecast the stock returns - a task that is simply impossible to do for any financial analyst.
The literature review and the construction of the LSTM showed that AI may be used to create superior returns. The biggest challenge is however data availability, because is only with enough data that the algorithm will be able to generalize well enough to be able to forecast returns in the future. This is the reason, why the inclusion of fundamental, technical, and economic data is difficult, unless the model is optimized using a large group of companies. As data availability increases continuously it is likely to also improve the out of sample performance of stock prediction algorithms using deep learning such as an LSTM.

9.1 Further Research

During the research process of this thesis I identified different research areas that remain untouched by academic literature and could be used as the basis of further research. As seen in section 8, LSTM can be used to predict stock returns to a certain degree in an theoretical environment. It would now be interesting to expand this research and see if any of these findings could be transferred into a real environment, meaning that the different release dates of the input variables are considered. One possible way of doing that is to only work with accounting data.
Furthermore, it would be interesting to examine the performance of a portfolio that is entirely based on the forecasts of an LSTM. Here an LSTM would need to be established per industry or even for each individual company. A reduction of input variables to variables that are available on a daily basis may be suitable, because otherwise there will not be enough data available to train the models at a company- or industry-level.
Last but not least, one could use the suggested approach as a baseline and further include shorter-term factors such as Tweets, blogs, news, or investor sentiments. This would allow the algorithm to find potential short-term fluctuations around a longer term trend.

References

Alexius, A., & Spang, D. (2018). Stock prices and gdp in the long run. *Journal of Applied Finance and Banking*, *8*(4), 107 - 126.

Azoff, E. M. (1994). *Neural network time series forecasting of financial markets. e. michael azoff.* Chichester <etc.> Wiley, 1994.

Baltagi, B. H. (2011). *Econometrics. badi h. baltagi.* Springer.

Banco Central do Brasil. (2019). *Relatório focus.* Retrieved 2019-05-21, from `https://www.bcb.gov.br/controleinflacao/relatoriofocus`

Barber, B. M., & Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors.

Bengio, Y., Boulanger-Lewandowski, N., & Pascanu, R. (2012). Advances in optimizing recurrent networks.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2), 281 - 305.

Bird, R., & Casavecchia, L. (2007). Value enhancement using momentum indicators: the european experience. *International Journal of Managerial Finance*(3), 229.

Bloomberg. (2019). Bloomberg terminal.

Brownlee, J. (2016a). *How to check if time series data is stationary with python.* Retrieved 2019-10-08, from `https://machinelearningmastery.com/time-series-data-stationary-python/`

Brownlee, J. (2016b). *Overfitting and underfitting with machine learning algorithms.* Retrieved 2019-10-30, from `https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/`

Brownlee, J. (2017). *How to handle missing timesteps in sequence prediction problems with python.* Retrieved 2019-10-07, from `https://machinelearningmastery.com/handle-missing-timesteps-sequence-prediction-problems-python`

Brownlee, J. (2019). *Long short-term memory networks with python: Develop sequence prediction models with deep learning*. Jason Brownlee. Retrieved from `https://books.google.com.br/books?id=ONpdswEACAAJ`

Burns, D. C., Sale, J. T., & Stephan, J. A. (2008). A better way to gauge profitability. *Journal of Accountancy*, *206*(2), 38 - 42.

Catalan, R. (2019, August 2). *Personal interview.* Private Investor.

Chandrashekar, R., Sakthivel, P., Sampath, T., & Chittedi, K. R. (2018). Macroeconomic variables and stock prices in emerging economies: A panel analysis. *Theoretical Applied Economics*, *25*(3), 91 - 100.

Charles, A., Darne, O., & Kim, J. H. (2017). International stock return predictability: Evidence from new statistical tests. *INTERNATIONAL REVIEW OF FINANCIAL ANALYSIS*, *54*, 97 - 113.

Chatzis, S. P., Siakoulis, V., Petropoulos, A., Stavroulakis, E., & Vlachogiannakis, N. (2018). Forecasting stock market crisis events using deep and statistical machine learning techniques. *Expert Systems with Applications*, *112*, 353 - 371.

Chaves, C., & Silva, A. C. (2018). Inflation and stock returns at b3. *Revista Brasileira de Finanças*, *16*(4), 521 - 544.

Chen, H., Chou, P., & Hsieh, C. (2018). Persistency of the momentum effect. *European Financial Management*, *24*(5), 856.

Chollet, F., et al. (2015). *Keras.* `https://keras.io`.

Chong, E., Park, F., & Han, C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications*, *83*, 187-205.

CNI. (2015). *Nec Índice nacional de expectativa do consumidor - metodologia* (1.2 ed.).

Dai, W., Wu, J.-Y., & Lu, C.-J. (2012). Combining nonlinear independent component analysis and neural network for the prediction of asian stock market indexes. *Expert Systems With Applications*, *39*, 4444 - 4452.

Dalio, R. (2017). *Principles : Life and work.* (Vol. First Simon Schuster hardcover edition). Simon Schuster.

Damodaran, A. (2002). *Investment valuation : Tools and techniques for determining the value of any asset.* (Vol. 2nd ed). Wiley.

Damodaran, A. (2006). *Damodaran on valuation. security analysis for investment and corporate finance. aswath damodaran.* J. Wiley.

Damodaran, A. (2007). Valuation approaches and metrics: A survey of the theory and evidence. *Foundations and Trends in Finance*, *1*(8), 693-784.

da S. Gomes, G., Ludermir, T., & Lima, L. (2011). Comparison of new activation functions in neural network for forecasting financial time series. *Neural Computing Applications*, *20*(3), 417 - 439.

Dayhoff, J. E., & DeLeo, J. M. (2001). Artificial neural networks: opening the black box. *Cancer*, *91*(8 Suppl), 1615 - 1635.

Dozat, T. (2016). Incorporating nesterov momentum into adam.

Eger, S., Youssef, P., & Gurevych, I. (2018, October). Is it time to swish? comparing deep learning activation functions across nlp tasks. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4415–4424). Association for Computational Linguistics.

Eikon. (2019). Thomson reuters eikon.

Elfwing, S., Uchibe, E., & Doya, K. (2018). Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, *107*, 3-11.

Fairfield, P. M., & Yohn, T. L. (2001). Using asset turnover and profit margin to forecast changes in profitability. *Review of Accounting Studies*, *6*(4), 371-385.

Fama, E. F., & French, K. R. (1996). Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, *51*(1), 55.

Fama, E. F., & French, K. R. (2006). Profitability, investment and average returns. *Journal of Financial Economics*, *82*(3), 491 - 518.

Fisher, K. L., & Statman, M. (2003). Consumer confidence and stock returns. *Journal of Portfolio Management*, *30*(1), 115.

Flannery, M. J., & Protopapadakis, A. A. (2002). Macroeconomic factors do influence aggregate stock returns. *Review of Financial Studies*, *15*(3), 751 - 782.

Friend, G., & Zehle, S. (2004). *Guide to business planning*. EIU: Economist Intelligence Unit.

Gay, R. D. (2016). Effect of macroeconomic variables on stock market returns for four emerging economies: Brazil, russia, india, and china.

George, T. J., & Hwang, C.-Y. (2004). The 52-week high and momentum investing. *The Journal of Finance*, *59*(5), 2145.

Graham, B., & Bogle, J. C. (2005). *The intelligent investor. the classic text on value investing. by benjamin graham ; foreword from john c. bogle.* HarperBusiness.

Graves, A., Mohamed, A., & Hinton, G. E. (2013). Speech recognition with deep recurrent neural networks. *CoRR*, *abs/1303.5778*.

Grinblatt, M., & Moskowitz, T. J. (2004). Predicting stock price movements from past returns: the role of consistency and tax-loss selling. *Journal of Financial Economics*, *71*(3), 541 - 579.

Göçken, M., Özçalıcı, M., Boru, A., & Dosdoğru, A. T. (2016). Integrating metaheuristics and artificial neural networks for improved stock price prediction. *Expert Systems With Applications*, *44*, 320 - 331.

Halford, G. S., Baker, R., McCredden, J. E., & Bain, J. D. (2005). How many variables can humans process? *Psychological Science*, *16*(1), 70-76.

Hamori, S. (2003). *An empirical investigation of stock markets. the ccf approach. by shigeyuki hamori.* Springer US.

Hao, Y., Chou, R. K., Ko, K.-C., & Yang, N.-T. (2018). The 52-week high, momentum, and investor sentiment. *International Review of Financial Analysis*, *57*, 167 - 183.

Heaton, J. B., Polson, N. G., & Witte, J. H. (2016). Deep learning for finance: deep portfolios. *APPLIED STOCHASTIC MODELS IN BUSINESS AND INDUSTRY*, *33*(1), 3 - 12.

Hedayati Moghaddam, A., Hedayati Moghaddam, M., & Esfandyari, M. (2017). Stock market index prediction using artificial neural network ; predicción del índice del mercado bursátil utilizando una red neuronal artificial.

Hendrik, B., & Kalok, C. (1998). Market efficiency and the returns to technical analysis. *Financial Management*, *27*(2), 5.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*,

9(8), 1735.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359-366.

Hsieh, J., & Walkling, R. A. (2006). The history and performance of concept stocks. *Journal of Banking and Finance*, *30*(9), 2433 - 2469.

Hsu, C.-M. (2013). A hybrid procedure with feature selection for resolving stock/futures price forecasting problems. *Neural Computing Applications*, *22*(3/4), 651 - 671.

Huang, B.-N., & Yang, C.-W. (2004). Industrial output and stock price revisited: an application of the multivariate indirect causality model. *Manchester School (1463-6786)*, *72*(3), 347 - 362.

Huddart, S., Lang, M., & Yetman, M. H. (2009). Volume and price patterns around a stock's 52-week highs and lows: Theory and evidence. *Management Science*, *55*(1), 16-31.

Humpe, A., & Macmillan, P. (2009). Can macroeconomic variables explain long-term stock market movements? a comparison of the us and japan. *Applied Financial Economics*, *19*(2), 111-119.

IBGE. (2015). *Pesquisa industrial mensal de emprego e salário - pimes.* `https://www.ibge.gov.br/estatisticas/sociais/trabalho/9298-pesquisa-industrial-mensal-de-emprego-e-salario.html?=&t=o-que-e`. (Accessed: 18-07-2019)

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning. with applications in r.* Springer.

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, *48*(1), 65.

Jordi, M. (2019, August 8). *Personal interview.* theScreener.

Kara, Y., Acar Boyacioglu, M., & Baykan, K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert Systems With Applications*, *38*, 5311 - 5319.

Kaul, G. (1990). Monetary regimes and the relation between stock return and inflationary expectations. *Journal of Financial Quantitative Analysis*, *25*(3), 307 - 321.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization.

Koh, F. (2019, November 7). *Personal interview.* Proprietary Desk.

Krollner, B., Vanstone, B., & Finnie, G. (2010). Financial time series forecasting with machine learning techniques: A survey.

Kuan, C.-M., & White, H. (1994). Artificial neural networks: An econometric perspective: Reply. *Econometric Reviews*, *13*(1), 139 - 143.

Lee, C. M. C., Shleifer, A., & Thaler, R. H. (1991). Investor sentiment and the closed-end

fund puzzle. *The Journal of Finance*, *46*(1), 75.

Leibowitz, M. L. (1997). Franchise margins and the sales-driven franchise value. *Financial Analysts Journal*, *53*(6), 43 - 53.

Leigh, W., Hightower, R., & Modani, N. (2005). Forecasting the new york stock exchange composite index with past price and interest rate on condition of volume spike. *Expert Systems With Applications*, *28*(1), 1 - 8.

Längkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling.

Maio, P. (2014). Another look at the stock return response to monetary policy actions*. *Review of Finance*, *18*(1), 321 - 371.

Malcolm, B., & Jeffrey, W. (2006). Investor sentiment and the cross-section of stock returns. *The Journal of Finance*, *61*(4), 1645.

Malgalhães, E. (2019, November 7). *Personal interview.* Pandhora.

Mandic, D. P., & Chambers, J. A. (2001). *Recurrent neural networks for prediction. learning algorithms, architectures, and stability. danilo p. mandic, jonathon a. chambers.* Chichester New York John Wiley, 2001.

Marshall, K. J. (2017). Good stocks cheap : value investing with confidence for a lifetime of stock market outperformance.

Martínez, F., Frías, M. P., Pérez, M. D., & Rivera, A. J. (2019). A methodology for applying k-nearest neighbor to time series forecasting. *Artificial Intelligence Review: An International Science and Engineering Journal*, *52*(3), 2019.

Marwala, T. (2013). *Economic modeling using artificial intelligence methods. by tshilidzi marwala.* London Springer London, 2013.

Matsumoto, K., Shivaswamy, M., & Hoban, J., James P. (1995). Security analysts' views of the financial ratios of manufacturers and retailers. *Financial Practice and Education*, *5*(2), 44 - 55.

McCormack, K. (2007). Stocks sink on weak retail sales. *BusinessWeek Online*, 15.

McMillan, D. G. (2017). Does money supply growth contain predictive power for stock returns? evidence and explanation. *International Journal of Banking, Accounting and Finance*(2), 119.

McNelis, P. D. (2005). *Neural networks in finance. gaining predictive edge in the market. paul d. mcnelis.* Amsterdam Elsevier Academic Press, 2005.

Mills, J. R., & Yamamura, J. H. (1998). The power of cash flow ratios. *Journal of Accountancy*, *186*(4), 53 - 61.

Mohapatra, S. M., & Rath, B. N. (2015a). Do macroeconomic factors matter for stock prices in emerging countries? evidence from panel cointegration and panel causality. *International Journal of Sustainable Economy*(2), 140.

Mohapatra, S. M., & Rath, B. N. (2015b). Do macroeconomic factors matter for stock prices in emerging countries? evidence from panel cointegration and panel

causality. *International Journal of Sustainable Economy*(2), 140.

Monks, R. A. G., Lajoux, A. R., & LeBaron, D. (2011). *Corporate valuation for portfolio investment. analyzing assets, earnings, cash flow, stock price, governance, and special situations. robert a.g. monks, alexandra reed lajoux ; foreword by dean lebaron.* Bloomberg Press.

Montier, J. (2009). *Value investing. tools and techniques for intelligent investment.* J. Wiley Sons.

Moody, J. (2012). *Neural networks: Tricks of the trade: Second edition*. Springer Berlin Heidelberg.

Nayak, A., Pai, M. M. M., & Pai, R. M. (2016). Prediction models for indian stock market.

Neusser, K. (2016). *Time series econometrics. by klaus neusser.* Springer International Publishing.

Ou, J. A., & Penman, S. H. (1989). Financial statement analysis and the prediction of stock returns. *Journal of Accounting Economics*, *11*(4), 295 - 329.

Owen, L. (1998). Earnings and expected returns. *The Journal of Finance*, *53*(5), 1563.

Park, C.-H., & Irwin, S. H. (2007). What do we know about the profitability of technical analysis?. *Journal of Economic Surveys*, *21*(4), 786 - 826.

Pascanu, R., Gulcehre, C., Cho, K., & Bengio, Y. (2013). *How to construct deep recurrent neural networks.*

Pascanu, R., Mikolov, T., & Bengio, Y. (2012). Understanding the exploding gradient problem. *CoRR*, *abs/1211.5063*.

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. *Expert Systems With Applications*, *42*, 2162 - 2172.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Penman, S. H. (2013). *Financial statement analysis and security valuation. stephen h. penman, columbia university.* McGraw Hill Education.

Piotroski, J. D. (2000). Value investing: The use of historical financial statement information to separate winners from losers. *Journal of Accounting Research*, *38*(3), 1 - 41.

Pring, M. J. (2014). *Technical analysis explained - the successful investor's guide to spotting investment trends and turning points* (5th ed.). McGraw-Hill Education.

Rafael La, P., Josef, L., Andrei, S., & Robert, V. (1997). Good news for value stocks: Further evidence on market efficiency. *The Journal of Finance*, *52*(2), 859.

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *CoRR*, *abs/1710.05941*.

Rapach, D. E., Wohar, M. E., & Rangvid, J. (2005). Macro variables and international stock return predictability. *International Journal of Forecasting*, *21*, 137 - 166.

Rather, A. M., Agarwal, A., & Sastry, V. (2015). Recurrent neural network and a hybrid model for prediction of stock returns. *Expert Systems With Applications*, *42*, 3234 - 3241.

Reis, L., Meurer, R., & Da Silva, S. (2010). Stock returns and foreign investment in brazil. *Applied Financial Economics*, *20*(17), 1351 - 1361.

Rist, M., & Pizzica, A. J. (2015). *Financial ratios for executives*. Berkeley, CA Apress, 2015.

Rogalski, R. J., & Vinso, J. D. (1977). Stock returns, money supply and the direction of causality. *Journal of Finance (Wiley-Blackwell)*, *32*(4), 1017 - 1030.

Rossi, E., & Forte, G. (2016). *Assessing relative valuation in equity markets. bridging research and practice. by emanuele rossi, gianfranco forte.* Palgrave Macmillan UK.

Safi, S., & White, A. (2017). Short and long-term forecasting using artificial neural networks for stock prices in palestine: A comparative study. *Electronic Journal of Applied Statistical Analysis*, *10*(1), 14-28.

Santos, L. d. R., & Montezano, R. M. d. S. (2011). Value and growth stocks in brazil: risks and returns for one - and two-dimensional portfolios under different economic conditions / ações de valor e de crescimento no brasil: riscos e retornos de carteiras uni e bidimensionais em diferentes estados da economia. *Revista Contabilidade Finanças*(56), 189.

Sanvicente, A. Z. (2014). The foreign capital flows and the behavior of stock prices at bmfbovespa. *BAR - Brazilian Administration Review*, *11*(1), 86 - 106.

Seitz, N. (1976). *Financial analysis: A programmed approach*.

Shayo, E. (2019, June 24). *Personal interview.* JP Morgan.

Soliman, M. T. (2008). The use of dupont analysis by market participants. *THE ACCOUNTING REVIEW*, *83*(3), 823 - 853.

Stoddard, O., & Noy, I. (2015). Fire-sale fdi? the impact of financial crises on foreign direct investment. *Review of Development Economics*, *19*(2), 387 - 399.

Ticknor, J. L. (2013). A bayesian regularized artificial neural network for stock market forecasting. *Expert Systems With Applications*, *40*, 5501 - 5506.

Vause, B. (2005). *Guide to analysing companies*. EIU: Economist Intelligence Unit.

Vigna, P., & Horner, W. (2019). U.s. stocks slip after retail sales unexpectedly fall. *Wall Street Journal - Online Edition*, 1.

Wahlen, J., & Wieland, M. (2011). Can financial statement analysis beat consensus analysts' recommendations?. *Review of Accounting Studies*, *16*(1), 89 - 115.

William, B., Josef, L., & Blake, L. (1992). Simple technical trading rules and the stochastic properties of stock returns. *The Journal of Finance*, *47*(5), 1731.

WITS - World Integrated Trade Solution. (2017). *Trade summary for brazil 2017.*

Retrieved 2019-05-22, from https://wits.worldbank.org/CountryProfile/en/Country/BRA/Year/LTST/Summarytext

Yadav, N., Yadav, A., & Kumar, M. (2015). *Introduction to neural network methods for differential equations.* Dordrecht, [Netherlands] Springer, 2015.

Yamarone, R. (2017). *The economic indicator handbook. how to evaluate economic trends to maximize profits and minimize losses. richard yamarone.* Wiley.

Yang, Q., & Wu, X. (2006). 10 challenging problems in data mining research. *International Journal of Information Technology Decision Making*, *5*(4), 597 - 604.

Yoshua, B., Patrice, S., & Paolo, F. (1994). Learning long-term dependencies with gradient descent is difficult.

Yu, L., Wang, S., & Lai, K. (2006). An integrated data preparation scheme for neural network data analysis. *IEEE Transactions on Knowledge and Data Engineering, Knowledge and Data Engineering, IEEE Transactions on, IEEE Trans. Knowl. Data Eng*(2), 217.

Zhang, J., Cui, S., Xu, Y., Li, Q., & Li, T. (2017). A novel data-driven stock price trend prediction system. *Expert Systems with Applications*, *97*, 60 - 69.

Zhou, F., Yang, Z., Zhou, H.-M., & Yang, L. (2019). Emd2fnn: A strategy combining empirical mode decomposition and factorization machine based neural network for stock market trend prediction. *Expert Systems with Applications*, *115*, 136-151.

Zhu, X., Wang, H., Xu, L., & Li, H. (2008). Predicting stock index increments by neural networks: The role of trading volume under different horizons. *Expert Systems With Applications*, *34*(4), 3043 - 3054.

Zoran, I., Sinisa, B., & Suzana, B. (2013). Strategy of stock valuation by fundamental analysis.

A Appendix

B Input Variables - Overview

Tag	Variable	Description
Economic Variables		
BZSTSETA Index	Brazil Selic Target Rate	A target interest rate set by the central bank in its efforts to influence short-term interest rates as part of its monetary policy strategy.
BZAD3M Index	Anbima Brazil Govt Bond Fixed Rate 3 Months	Anbima's Estimated Term Structure of Interest Rate, based on fixed rate sovereigns zero coupon bonds (LTNs).
GEBR02Y Index	Brazil Government Generic Bond 2 Year	The rates are comprised of Generic Brazilian real government bills and bonds.
GEBR10Y Index	Brazil Government Generic Bond 10 Year	The rates are comprised of Generic Brazilian real government bills and bonds.
BZRTRTSA Index	Brazil Retail Sales Volume SA	Retail sales (also referred to as retail trade) tracks the resale of new and used goods to the general public, for personal or household consumption. This concept is based on the volume of goods sold.
EHUPBR Index	Brazil Unemployment Rate (%)	Both the quarter and yearly values are reported on an annualised basis to avoid seasonal issues, using either a 4-quarter or 12-month moving sum of national sourced indices. Refer to the underlying tickers for their description:
BZTWBALW Index	Brazil Trade Balance Weekly Balance	The international trade balance measures the difference between the movement of merchandise trade leaving a country (exports) and entering a country (imports). This measure tracks the volume of the merchandise trade balance.
BZMS1 Index	Brazil Money Supply M1 Brazil M1	The money supply measures the total amount of money in circulation in a country or group of countries in a monetary union.

Tag	Variable	Description
BZMS2 Index	Brazil Money Supply M2 Brazil M2	The money supply measures the total amount of money in circulation in a country or group of countries in a monetary union.
BZMS3 Index	Brazil Money Supply M3 Brazil M3	The money supply measures the total amount of money in circulation in a country or group of countries in a monetary union.
BZFDTMON Index	Brazil Foreign Direct Investment Net	The financial & capital accounts are part of the balance of payments. The financial account tracks all of a country's external financial assets and liabilities. Major components of the financial account include direct investment, portfolio investment, other investments and reserve assets. The capital account tracks transfers of ownership across borders of fixed assets and acquisition or disposal of nonproduced, nonfinancial assets.The balance of payments is a record of a country's overall international transactions with the rest of the world (i.e. transactions between residents of a country and nonresidents). The balance of payments is divided into current, capital and financial accounts.
BZPIIPCY Index	Brazil CPI IPCA YoY	Consumer prices (CPI) are a measure of prices paid by consumers for a market basket of consumer goods and services. The yearly (or monthly) growth rates represent the inflation rate.
BRLUSD CURNCY	BRLUSD	Spot Exchange Rate - Price of 1 BRL in USD
BRLCNH CURNCY	BRLCNH	Spot Exchange Rate - Price of 1 BRL in CNH
BRLARS CURNCY	BRLARS	Spot Exchange Rate - Price of 1 BRL in ARS

Tag	Variable	Description
BZGDQOQ% Index	Brazil GDP QoQ SA 1995=100	Gross domestic product (GDP) measures the final market value of all goods and services produced within a country. It is the most frequently used indicator of economic activity. The GDP by expenditure approach measures total final expenditures (at purchasers' prices), including exports less imports. This concept is adjusted for inflation.
BZFGCCSA Index	Brazil FGV Consumer Confidence Index SA	The target sample size factors in the ranges of income across the different municipalities in Brazil to allow for an even distribution of responses.Target Audience: households, survey preferably to be taken by head of household who is >18yrs old. Sample Size: 2045 households. Date of Survey: first 2 weeks of the month
BZIPTLSA Index	Brazil Real Industrial Production SA 2012=100	Industrial production measures the output of industrial establishments in the following industries: mining and quarrying, manufacturing and public utilities (electricity, gas and water supply). Production is based on the volume of the output.
BRCCOUHT Index	Brazil Credit Outstanding Households Total	This concept tracks the outstanding amount of credit (or loans) extended to businesses and consumers.

Table 4: Overview - Input Variables

C Statistical Outputs

C.1 Augmented Dickey Fuller Test

C.1.1 ADF - Accounting Data - Original

LREN3		MGLU3		BTOW3		VVAR3		LAME3	
LN_NEXT		**LN_NEXT**		**LN_NEXT**		**LN_NEXT**		**LN_NEXT**	
ADF Stat:	-1.7646	ADF Stat:	-2.1459	ADF Stat:	-7.1975	ADF Stat:	-4.6157	ADF Stat:	-3.2034
p-value:	0.3982	p-value:	0.2265	p-value:	0.0000	p-value:	0.0001	p-value:	0.0198
LN_LAST_Q		**LN_LAST_Q**		**LN_LAST_Q**		**LN_LAST_Q**		**LN_LAST_Q**	
ADF Stat:	-1.5994	ADF Stat:	-2.0233	ADF Stat:	-6.9161	ADF Stat:	-4.7980	ADF Stat:	-4.1897
p-value:	0.4839	p-value:	0.2765	p-value:	0.0000	p-value:	0.0001	p-value:	0.0007
LN_LAST_12M		**LN_LAST_12M**		**LN_LAST_12M**		**LN_LAST_12M**		**LN_LAST_12M**	
ADF Stat:	-3.2071	ADF Stat:	-1.7955	ADF Stat:	-2.4501	ADF Stat:	-2.9754	ADF Stat:	-4.0842
p-value:	0.0196	p-value:	0.3827	p-value:	0.1281	p-value:	0.0373	p-value:	0.0010
EQY_DVD_YLD_12M		**EQY_DVD_YLD_12M**		**EQY_DVD_YLD_12M**		**EQY_DVD_YLD_12M**		**EQY_DVD_YLD_12M**	
ADF Stat:	-2.1151	ADF Stat:	-2.3209	ADF Stat:	-2.0662	ADF Stat:	-3.6098	ADF Stat:	-2.9500
p-value:	0.2385	p-value:	0.1653	p-value:	0.2583	p-value:	0.0056	p-value:	0.0398
PX_TO_BOOK_RATIO		**PX_TO_BOOK_RATIO**		**PX_TO_BOOK_RATIO**		**PX_TO_BOOK_RATIO**		**PX_TO_BOOK_RATIO**	
ADF Stat:	-3.8070	ADF Stat:	-1.1289	ADF Stat:	-4.3219	ADF Stat:	-2.6130	ADF Stat:	-3.0307
p-value:	0.0028	p-value:	0.7034	p-value:	0.0004	p-value:	0.0903	p-value:	0.0321
PE_RATIO		**PE_RATIO**		**PE_RATIO**		**PE_RATIO**		**PE_RATIO**	
ADF Stat:	-2.6249	ADF Stat:	-1.8696	ADF Stat:	-3.3731	ADF Stat:	-2.9474	ADF Stat:	-2.0158
p-value:	0.0880	p-value:	0.3465	p-value:	0.0119	p-value:	0.0401	p-value:	0.2797
PX_TO_SALES_RATIO		**PX_TO_SALES_RATIO**		**PX_TO_SALES_RATIO**		**PX_TO_SALES_RATIO**		**PX_TO_SALES_RATIO**	
ADF Stat:	0.1629	ADF Stat:	0.6161	ADF Stat:	0.5195	ADF Stat:	-6.5233	ADF Stat:	-1.5044
p-value:	0.9701	p-value:	0.9880	p-value:	0.9855	p-value:	0.0000	p-value:	0.5313
CUR_RATIO		**CUR_RATIO**		**CUR_RATIO**		**CUR_RATIO**		**CUR_RATIO**	
ADF Stat:	1.0472	ADF Stat:	-1.4501	ADF Stat:	-2.5580	ADF Stat:	-0.8316	ADF Stat:	-0.7650
p-value:	0.9947	p-value:	0.5580	p-value:	0.1020	p-value:	0.8097	p-value:	0.8292
QUICK_RATIO		**QUICK_RATIO**		**QUICK_RATIO**		**QUICK_RATIO**		**QUICK_RATIO**	
ADF Stat:	1.3450	ADF Stat:	-2.2274	ADF Stat:	-2.5773	ADF Stat:	-2.1916	ADF Stat:	-2.1714
p-value:	0.9968	p-value:	0.1964	p-value:	0.0978	p-value:	0.2093	p-value:	0.2168
INVENT_TURN		**INVENT_TURN**		**INVENT_TURN**		**INVENT_TURN**		**INVENT_TURN**	
ADF Stat:	-4.5509	ADF Stat:	-3.2513	ADF Stat:	-1.6825	ADF Stat:	-1.4275	ADF Stat:	-0.1943
p-value:	0.0002	p-value:	0.0172	p-value:	0.4401	p-value:	0.5690	p-value:	0.9392
ACCT_RCV_TURN		**ACCT_RCV_TURN**		**ACCT_RCV_TURN**		**ACCT_RCV_TURN**		**ACCT_RCV_TURN**	
ADF Stat:	-4.2555	ADF Stat:	-1.2234	ADF Stat:	0.8728	ADF Stat:	-4.2073	ADF Stat:	-1.1452
p-value:	0.0005	p-value:	0.6635	p-value:	0.9927	p-value:	0.0006	p-value:	0.6967
ACCT_RCV_DAYS		**ACCT_RCV_DAYS**		**ACCT_RCV_DAYS**		**ACCT_RCV_DAYS**		**ACCT_RCV_DAYS**	
ADF Stat:	-3.3120	ADF Stat:	-1.5628	ADF Stat:	-0.7480	ADF Stat:	-3.0082	ADF Stat:	-1.1312
p-value:	0.0143	p-value:	0.5023	p-value:	0.8339	p-value:	0.0341	p-value:	0.7025
ASSET_TURNOVER		**ASSET_TURNOVER**		**ASSET_TURNOVER**		**ASSET_TURNOVER**		**ASSET_TURNOVER**	
ADF Stat:	-1.1379	ADF Stat:	0.8151	ADF Stat:	-0.9943	ADF Stat:	-2.1741	ADF Stat:	1.8400

p-value:	0.6997	p-value:	0.9919	p-value:	0.7553	p-value:	0.2158	p-value:	0.9984
NET_OPER_PROFIT_AFTER_TAX		**NET_OPER_PROFIT_AFTER_TAX**		**NET_OPER_PROFIT_AFTER_TAX**		**NET_OPER_PROFIT_AFTER_TAX**		**NET_OPER_PROFIT_AFTER_TAX**	
ADF Stat:	1.9252	ADF Stat:	0.3737	ADF Stat:	-4.0386	ADF Stat:	-1.8478	ADF Stat:	-5.7783
p-value:	0.9986	p-value:	0.9805	p-value:	0.0012	p-value:	0.3570	p-value:	0.0000
IS_EPS		**IS_EPS**		**IS_EPS**		**IS_EPS**		**IS_EPS**	
ADF Stat:	1.8032	ADF Stat:	-5.3032	ADF Stat:	-2.5482	ADF Stat:	-3.6574	ADF Stat:	-2.3036
p-value:	0.9984	p-value:	0.0000	p-value:	0.1042	p-value:	0.0048	p-value:	0.1708
IS_OPER_INC		**IS_OPER_INC**		**IS_OPER_INC**		**IS_OPER_INC**		**IS_OPER_INC**	
ADF Stat:	1.0178	ADF Stat:	1.1745	ADF Stat:	-1.8392	ADF Stat:	-3.2948	ADF Stat:	-1.9218
p-value:	0.9945	p-value:	0.9958	p-value:	0.3612	p-value:	0.0151	p-value:	0.3219
CF_FREE_CASH_FLOW		**CF_FREE_CASH_FLOW**		**CF_FREE_CASH_FLOW**		**CF_FREE_CASH_FLOW**		**CF_FREE_CASH_FLOW**	
ADF Stat:	-2.0751	ADF Stat:	-1.5163	ADF Stat:	-2.1025	ADF Stat:	-3.1699	ADF Stat:	-6.3501
p-value:	0.2547	p-value:	0.5254	p-value:	0.2435	p-value:	0.0218	p-value:	0.0000
CF_CASH_FROM_OPER		**CF_CASH_FROM_OPER**		**CF_CASH_FROM_OPER**		**CF_CASH_FROM_OPER**		**CF_CASH_FROM_OPER**	
ADF Stat:	-1.7407	ADF Stat:	-0.2878	ADF Stat:	-2.0259	ADF Stat:	-2.9976	ADF Stat:	-6.0082
p-value:	0.4103	p-value:	0.9272	p-value:	0.2753	p-value:	0.0351	p-value:	0.0000
SHORT_AND_LONG_TERM_DEBT		**SHORT_AND_LONG_TERM_DEBT**		**SHORT_AND_LONG_TERM_DEBT**		**SHORT_AND_LONG_TERM_DEBT**		**SHORT_AND_LONG_TERM_DEBT**	
ADF Stat:	-0.0463	ADF Stat:	-1.9454	ADF Stat:	0.9504	ADF Stat:	-3.6465	ADF Stat:	0.9897
p-value:	0.9545	p-value:	0.3110	p-value:	0.9937	p-value:	0.0049	p-value:	0.9942
DEBT_TO_OCF		**DEBT_TO_OCF**		**DEBT_TO_OCF**		**DEBT_TO_OCF**		**DEBT_TO_OCF**	
ADF Stat:	-3.2963	ADF Stat:	8.7788	ADF Stat:	-2.7996	ADF Stat:	-3.1049	ADF Stat:	-5.6324
p-value:	0.0150	p-value:	1.0000	p-value:	0.0583	p-value:	0.0262	p-value:	0.0000
IS_INT_EXPENSE		**IS_INT_EXPENSE**		**IS_INT_EXPENSE**		**IS_INT_EXPENSE**		**IS_INT_EXPENSE**	
ADF Stat:	-1.5020	ADF Stat:	-0.8528	ADF Stat:	-1.9826	ADF Stat:	1.5271	ADF Stat:	-0.9292
p-value:	0.5325	p-value:	0.8031	p-value:	0.2942	p-value:	0.9976	p-value:	0.7782
OCF_TO_INTEREST_EXPENSE		**OCF_TO_INTEREST_EXPENSE**		**OCF_TO_INTEREST_EXPENSE**		**OCF_TO_INTEREST_EXPENSE**		**OCF_TO_INTEREST_EXPENSE**	
ADF Stat:	-2.5834	ADF Stat:	1.6355	ADF Stat:	-6.4772	ADF Stat:	-2.9877	ADF Stat:	-8.3868
p-value:	0.0965	p-value:	0.9980	p-value:	0.0000	p-value:	0.0361	p-value:	0.0000
CAPITAL_EXPEND		**CAPITAL_EXPEND**		**CAPITAL_EXPEND**		**CAPITAL_EXPEND**		**CAPITAL_EXPEND**	
ADF Stat:	-1.4582	ADF Stat:	-2.3166	ADF Stat:	-0.7251	ADF Stat:	-3.7930	ADF Stat:	-3.4402
p-value:	0.5541	p-value:	0.1667	p-value:	0.8401	p-value:	0.0030	p-value:	0.0097
OCF_EXPEND		**OCF_EXPEND**		**OCF_EXPEND**		**OCF_EXPEND**		**OCF_EXPEND**	
ADF Stat:	-4.3110	ADF Stat:	0.4074	ADF Stat:	-6.9165	ADF Stat:	-3.0797	ADF Stat:	-4.7631
p-value:	0.0004	p-value:	0.9818	p-value:	0.0000	p-value:	0.0281	p-value:	0.0001
SALES_REV_TURN		**SALES_REV_TURN**		**SALES_REV_TURN**		**SALES_REV_TURN**		**SALES_REV_TURN**	
ADF Stat:	0.2487	ADF Stat:	0.8400	ADF Stat:	-1.5970	ADF Stat:	-2.9546	ADF Stat:	-1.5385
p-value:	0.9749	p-value:	0.9923	p-value:	0.4851	p-value:	0.0394	p-value:	0.5144
GROSS_MARGIN		**GROSS_MARGIN**		**GROSS_MARGIN**		**GROSS_MARGIN**		**GROSS_MARGIN**	
ADF Stat:	-2.8453	ADF Stat:	-2.4665	ADF Stat:	-2.1071	ADF Stat:	-3.9417	ADF Stat:	-0.6325
p-value:	0.0521	p-value:	0.1238	p-value:	0.2417	p-value:	0.0017	p-value:	0.8635
OPER_MARGIN		**OPER_MARGIN**		**OPER_MARGIN**		**OPER_MARGIN**		**OPER_MARGIN**	
ADF Stat:	-1.6475	ADF Stat:	-0.4486	ADF Stat:	-1.2861	ADF Stat:	-1.1284	ADF Stat:	-6.3391
p-value:	0.4584	p-value:	0.9017	p-value:	0.6355	p-value:	0.7036	p-value:	0.0000
RETURN_ON_ASSET		**RETURN_ON_ASSET**		**RETURN_ON_ASSET**		**RETURN_ON_ASSET**		**RETURN_ON_ASSET**	
ADF Stat:	-2.0912	ADF Stat:	1.0293	ADF Stat:	-3.3955	ADF Stat:	-2.1538	ADF Stat:	-0.5595
p-value:	0.2481	p-value:	0.9946	p-value:	0.0111	p-value:	0.2234	p-value:	0.8799
NET_OPERATING_ASSETS		**NET_OPERATING_ASSETS**		**NET_OPERATING_ASSETS**		**NET_OPERATING_ASSETS**		**NET_OPERATING_ASSETS**	
ADF Stat:	0.3950	ADF Stat:	0.3389	ADF Stat:	-1.3281	ADF Stat:	-3.7532	ADF Stat:	-0.3267

p-value:	0.9813	p-value:	0.9791	p-value:	0.6162	p-value:	0.0034	p-value:	0.9216
RNOA		**RNOA**		**RNOA**		**RNOA**		**RNOA**	
ADF Stat:	-2.2057	ADF Stat:	-2.7471	ADF Stat:	-2.6471	ADF Stat:	-2.1690	ADF Stat:	-0.5679
p-value:	0.2042	p-value:	0.0662	p-value:	0.0837	p-value:	0.2177	p-value:	0.8781
RETURN_COM_EQY		**RETURN_COM_EQY**		**RETURN_COM_EQY**		**RETURN_COM_EQY**		**RETURN_COM_EQY**	
ADF Stat:	-3.3767	ADF Stat:	2.0171	ADF Stat:	-4.4108	ADF Stat:	-1.7806	ADF Stat:	-2.3007
p-value:	0.0118	p-value:	0.9987	p-value:	0.0003	p-value:	0.3901	p-value:	0.1717
EBIT		**EBIT**		**EBIT**		**EBIT**		**EBIT**	
ADF Stat:	1.0178	ADF Stat:	1.1745	ADF Stat:	-1.8392	ADF Stat:	-3.2948	ADF Stat:	-1.9218
p-value:	0.9945	p-value:	0.9958	p-value:	0.3612	p-value:	0.0151	p-value:	0.3219
INT_COVERAGE_RATIO		**INT_COVERAGE_RATIO**		**INT_COVERAGE_RATIO**		**INT_COVERAGE_RATIO**		**INT_COVERAGE_RATIO**	
ADF Stat:	-1.7216	ADF Stat:	-1.5526	ADF Stat:	-0.9071	ADF Stat:	-1.9221	ADF Stat:	-0.7188
p-value:	0.4200	p-value:	0.5073	p-value:	0.7856	p-value:	0.3217	p-value:	0.8418
TOT_DEBT_TO_TOT_EQY		**TOT_DEBT_TO_TOT_EQY**		**TOT_DEBT_TO_TOT_EQY**		**TOT_DEBT_TO_TOT_EQY**		**TOT_DEBT_TO_TOT_EQY**	
ADF Stat:	-2.9066	ADF Stat:	1.0980	ADF Stat:	-2.3904	ADF Stat:	-2.1650	ADF Stat:	-0.9490
p-value:	0.0446	p-value:	0.9952	p-value:	0.1445	p-value:	0.2192	p-value:	0.7714
TOT_DEBT_TO_TOT_CAP		**TOT_DEBT_TO_TOT_CAP**		**TOT_DEBT_TO_TOT_CAP**		**TOT_DEBT_TO_TOT_CAP**		**TOT_DEBT_TO_TOT_CAP**	
ADF Stat:	-3.5145	ADF Stat:	-2.4313	ADF Stat:	-2.3977	ADF Stat:	-2.2294	ADF Stat:	0.0076
p-value:	0.0076	p-value:	0.1331	p-value:	0.1424	p-value:	0.1958	p-value:	0.9592
LT_DEBT_TO_TOT_CAP		**LT_DEBT_TO_TOT_CAP**		**LT_DEBT_TO_TOT_CAP**		**LT_DEBT_TO_TOT_CAP**		**LT_DEBT_TO_TOT_CAP**	
ADF Stat:	-2.3276	ADF Stat:	-2.2204	ADF Stat:	-3.4833	ADF Stat:	-10.3468	ADF Stat:	-1.9794
p-value:	0.1632	p-value:	0.1989	p-value:	0.0084	p-value:	0.0000	p-value:	0.2957
TOTAL_EQUITY		**TOTAL_EQUITY**		**TOTAL_EQUITY**		**TOTAL_EQUITY**		**TOTAL_EQUITY**	
ADF Stat:	2.8247	ADF Stat:	2.9965	ADF Stat:	-1.2063	ADF Stat:	-2.4440	ADF Stat:	5.5995
p-value:	1.0000	p-value:	1.0000	p-value:	0.6709	p-value:	0.1297	p-value:	1.0000
PX_VOLUME		**PX_VOLUME**		**PX_VOLUME**		**PX_VOLUME**		**PX_VOLUME**	
ADF Stat:	-1.7924	ADF Stat:	-1.4908	ADF Stat:	-1.2653	ADF Stat:	-0.0131	ADF Stat:	-3.3144
p-value:	0.3842	p-value:	0.5381	p-value:	0.6449	p-value:	0.9574	p-value:	0.0142
52w_max_CP		**52w_max_CP**		**52w_max_CP**		**52w_max_CP**		**52w_max_CP**	
ADF Stat:	-3.8271	ADF Stat:	-2.4259	ADF Stat:	-3.7670	ADF Stat:	-4.3741	ADF Stat:	-4.1944
p-value:	0.0026	p-value:	0.1346	p-value:	0.0033	p-value:	0.0003	p-value:	0.0007
52w_min_CP		**52w_min_CP**		**52w_min_CP**		**52w_min_CP**		**52w_min_CP**	
ADF Stat:	-5.1462	ADF Stat:	-1.8479	ADF Stat:	-2.0249	ADF Stat:	-2.4129	ADF Stat:	-0.9967
p-value:	0.0000	p-value:	0.3570	p-value:	0.2758	p-value:	0.1381	p-value:	0.7545
VMA		**VMA**		**VMA**		**VMA**		**VMA**	
ADF Stat:	-4.5905	ADF Stat:	-0.8266	ADF Stat:	-4.6370	ADF Stat:	-4.8615	ADF Stat:	-4.8209
p-value:	0.0001	p-value:	0.8112	p-value:	0.0001	p-value:	0.0000	p-value:	0.0000

Table 5: ADF Output - Accounting Data, Original

C.1.2 ADF - Accounting Data - Treated

LREN3		MGLU3		BTOW3		VVAR3		LAME3	
LN_NEXT		**LN_NEXT**		**LN_NEXT**		**LN_NEXT**		**LN_NEXT**	
ADF Stat:	-1.0143	ADF Stat:	-2.1459	ADF Stat:	-4.8230	ADF Stat:	-4.4953	ADF Stat:	-3.2379
p-value:	0.7480	p-value:	0.2265	p-value:	0.0000	p-value:	0.0002	p-value:	0.0179
LN_LAST_Q		**LN_LAST_Q**		**LN_LAST_Q**		**LN_LAST_Q**		**LN_LAST_Q**	
ADF Stat:	-1.2015	ADF Stat:	-2.0233	ADF Stat:	-6.7418	ADF Stat:	-11.0035	ADF Stat:	-2.5877
p-value:	0.6729	p-value:	0.2765	p-value:	0.0000	p-value:	0.0000	p-value:	0.0956
LN_LAST_12M		**LN_LAST_12M**		**LN_LAST_12M**		**LN_LAST_12M**		**LN_LAST_12M**	
ADF Stat:	-4.4392	ADF Stat:	-1.7955	ADF Stat:	-3.4926	ADF Stat:	-3.2954	ADF Stat:	-2.8782
p-value:	0.0003	p-value:	0.3827	p-value:	0.0082	p-value:	0.0151	p-value:	0.0479
EQY_DVD_YLD_12M		**EQY_DVD_YLD_12M**		**EQY_DVD_YLD_12M**		**EQY_DVD_YLD_12M**		**EQY_DVD_YLD_12M**	
ADF Stat:	-2.4528	ADF Stat:	-2.3209	ADF Stat:	-1.9304	ADF Stat:	-2.4944	ADF Stat:	-2.3049
p-value:	0.1274	p-value:	0.1653	p-value:	0.3179	p-value:	0.1168	p-value:	0.1704
PX_TO_BOOK_RATIO		**PX_TO_BOOK_RATIO**		**PX_TO_BOOK_RATIO**		**PX_TO_BOOK_RATIO**		**PX_TO_BOOK_RATIO**	
ADF Stat:	-3.7067	ADF Stat:	-1.1289	ADF Stat:	-3.7813	ADF Stat:	-2.8821	ADF Stat:	0.0541
p-value:	0.0040	p-value:	0.7034	p-value:	0.0031	p-value:	0.0475	p-value:	0.9628
PE_RATIO		**PE_RATIO**		**PE_RATIO**		**PE_RATIO**		**PE_RATIO**	
ADF Stat:	-3.0873	ADF Stat:	-1.8696	ADF Stat:	-3.7941	ADF Stat:	-2.8738	ADF Stat:	-1.6849
p-value:	0.0275	p-value:	0.3465	p-value:	0.0030	p-value:	0.0485	p-value:	0.4389
PX_TO_SALES_RATIO		**PX_TO_SALES_RATIO**		**PX_TO_SALES_RATIO**		**PX_TO_SALES_RATIO**		**PX_TO_SALES_RATIO**	
ADF Stat:	-0.1904	ADF Stat:	0.6161	ADF Stat:	1.3320	ADF Stat:	-2.7687	ADF Stat:	-1.4160
p-value:	0.9397	p-value:	0.9880	p-value:	0.9968	p-value:	0.0629	p-value:	0.5746
CUR_RATIO		**CUR_RATIO**		**CUR_RATIO**		**CUR_RATIO**		**CUR_RATIO**	
ADF Stat:	-0.7659	ADF Stat:	-1.4501	ADF Stat:	-1.9536	ADF Stat:	-0.9727	ADF Stat:	-0.5119
p-value:	0.8289	p-value:	0.5580	p-value:	0.3073	p-value:	0.7631	p-value:	0.8897
QUICK_RATIO		**QUICK_RATIO**		**QUICK_RATIO**		**QUICK_RATIO**		**QUICK_RATIO**	
ADF Stat:	-0.0670	ADF Stat:	-2.2274	ADF Stat:	-2.0606	ADF Stat:	-1.9182	ADF Stat:	-0.5194
p-value:	0.9526	p-value:	0.1964	p-value:	0.2606	p-value:	0.3235	p-value:	0.8882
INVENT_TURN		**INVENT_TURN**		**INVENT_TURN**		**INVENT_TURN**		**INVENT_TURN**	
ADF Stat:	-1.2954	ADF Stat:	-3.2513	ADF Stat:	-2.0108	ADF Stat:	3.8081	ADF Stat:	-0.1298
p-value:	0.6313	p-value:	0.0172	p-value:	0.2819	p-value:	1.0000	p-value:	0.9464
ACCT_RCV_TURN		**ACCT_RCV_TURN**		**ACCT_RCV_TURN**		**ACCT_RCV_TURN**		**ACCT_RCV_TURN**	
ADF Stat:	-4.2893	ADF Stat:	-1.2234	ADF Stat:	1.1828	ADF Stat:	-3.8730	ADF Stat:	-1.3434
p-value:	0.0005	p-value:	0.6635	p-value:	0.9959	p-value:	0.0022	p-value:	0.6091
ACCT_RCV_DAYS		**ACCT_RCV_DAYS**		**ACCT_RCV_DAYS**		**ACCT_RCV_DAYS**		**ACCT_RCV_DAYS**	
ADF Stat:	-3.5996	ADF Stat:	-1.5628	ADF Stat:	-1.7945	ADF Stat:	-2.1842	ADF Stat:	-3.4708
p-value:	0.0058	p-value:	0.5023	p-value:	0.3832	p-value:	0.2120	p-value:	0.0088
ASSET_TURNOVER		**ASSET_TURNOVER**		**ASSET_TURNOVER**		**ASSET_TURNOVER**		**ASSET_TURNOVER**	
ADF Stat:	-3.5623	ADF Stat:	-2.5650	ADF Stat:	-8.5747	ADF Stat:	-3.7329	ADF Stat:	-3.4378
p-value:	0.0065	p-value:	0.1004	p-value:	0.0000	p-value:	0.0037	p-value:	0.0097
NET_OPER_PROFIT_AFTER_TAX		**NET_OPER_PROFIT_AFTER_TAX**		**NET_OPER_PROFIT_AFTER_TAX**		**NET_OPER_PROFIT_AFTER_TAX**		**NET_OPER_PROFIT_AFTER_TAX**	
ADF Stat:	0.3781	ADF Stat:	-6.7619	ADF Stat:	-6.8132	ADF Stat:	-4.8465	ADF Stat:	-10.3222
p-value:	0.9807	p-value:	0.0000	p-value:	0.0000	p-value:	0.0000	p-value:	0.0000
IS_EPS		**IS_EPS**		**IS_EPS**		**IS_EPS**		**IS_EPS**	
ADF Stat:	2.2797	ADF Stat:	-5.3032	ADF Stat:	-2.6800	ADF Stat:	-3.6096	ADF Stat:	-7.5455

p-value:	0.9989	p-value:	0.0000	p-value:	0.0776	p-value:	0.0056	p-value:	0.0000
IS_OPER_INC		**IS_OPER_INC**		**IS_OPER_INC**		**IS_OPER_INC**		**IS_OPER_INC**	
ADF Stat:	1.2297	ADF Stat:	1.1745	ADF Stat:	-1.8704	ADF Stat:	-3.3804	ADF Stat:	-2.2806
p-value:	0.9962	p-value:	0.9958	p-value:	0.3461	p-value:	0.0116	p-value:	0.1783
CF_FREE_CASH_FLOW		**CF_FREE_CASH_FLOW**		**CF_FREE_CASH_FLOW**		**CF_FREE_CASH_FLOW**		**CF_FREE_CASH_FLOW**	
ADF Stat:	-1.5463	ADF Stat:	-1.5163	ADF Stat:	-2.1597	ADF Stat:	-3.2341	ADF Stat:	-5.2212
p-value:	0.5105	p-value:	0.5254	p-value:	0.2212	p-value:	0.0181	p-value:	0.0000
CF_CASH_FROM_OPER		**CF_CASH_FROM_OPER**		**CF_CASH_FROM_OPER**		**CF_CASH_FROM_OPER**		**CF_CASH_FROM_OPER**	
ADF Stat:	-0.7069	ADF Stat:	-0.2878	ADF Stat:	-2.0910	ADF Stat:	-3.3037	ADF Stat:	-3.2448
p-value:	0.8449	p-value:	0.9272	p-value:	0.2482	p-value:	0.0147	p-value:	0.0175
SHORT_AND_LONG_TERM_DEBT		**SHORT_AND_LONG_TERM_DEBT**		**SHORT_AND_LONG_TERM_DEBT**		**SHORT_AND_LONG_TERM_DEBT**		**SHORT_AND_LONG_TERM_DEBT**	
ADF Stat:	-2.8163	ADF Stat:	-2.5071	ADF Stat:	-4.5414	ADF Stat:	-2.8773	ADF Stat:	-6.6839
p-value:	0.0560	p-value:	0.1138	p-value:	0.0002	p-value:	0.0481	p-value:	0.0000
DEBT_TO_OCF		**DEBT_TO_OCF**		**DEBT_TO_OCF**		**DEBT_TO_OCF**		**DEBT_TO_OCF**	
ADF Stat:	-2.2250	ADF Stat:	8.7788	ADF Stat:	-3.1014	ADF Stat:	-2.4284	ADF Stat:	-4.6275
p-value:	0.1973	p-value:	1.0000	p-value:	0.0264	p-value:	0.1339	p-value:	0.0001
IS_INT_EXPENSE		**IS_INT_EXPENSE**		**IS_INT_EXPENSE**		**IS_INT_EXPENSE**		**IS_INT_EXPENSE**	
ADF Stat:	-6.8668	ADF Stat:	-0.1475	ADF Stat:	-6.0061	ADF Stat:	-7.4038	ADF Stat:	-5.5336
p-value:	0.0000	p-value:	0.9445	p-value:	0.0000	p-value:	0.0000	p-value:	0.0000
OCF_TO_INTEREST_EXPENSE		**OCF_TO_INTEREST_EXPENSE**		**OCF_TO_INTEREST_EXPENSE**		**OCF_TO_INTEREST_EXPENSE**		**OCF_TO_INTEREST_EXPENSE**	
ADF Stat:	-1.8409	ADF Stat:	1.6355	ADF Stat:	-6.4185	ADF Stat:	-3.8781	ADF Stat:	-4.8285
p-value:	0.3603	p-value:	0.9980	p-value:	0.0000	p-value:	0.0022	p-value:	0.0000
CAPITAL_EXPEND		**CAPITAL_EXPEND**		**CAPITAL_EXPEND**		**CAPITAL_EXPEND**		**CAPITAL_EXPEND**	
ADF Stat:	-0.8524	ADF Stat:	-2.3166	ADF Stat:	-0.7714	ADF Stat:	-3.8219	ADF Stat:	-3.7844
p-value:	0.8032	p-value:	0.1667	p-value:	0.8274	p-value:	0.0027	p-value:	0.0031
OCF_EXPEND		**OCF_EXPEND**		**OCF_EXPEND**		**OCF_EXPEND**		**OCF_EXPEND**	
ADF Stat:	-1.2322	ADF Stat:	0.4074	ADF Stat:	-6.7204	ADF Stat:	-2.8559	ADF Stat:	-4.8191
p-value:	0.6596	p-value:	0.9818	p-value:	0.0000	p-value:	0.0507	p-value:	0.0000
SALES_REV_TURN		**SALES_REV_TURN**		**SALES_REV_TURN**		**SALES_REV_TURN**		**SALES_REV_TURN**	
ADF Stat:	-4.5259	ADF Stat:	-0.3282	ADF Stat:	-2.7231	ADF Stat:	-1.5309	ADF Stat:	-5.4705
p-value:	0.0002	p-value:	0.9214	p-value:	0.0701	p-value:	0.5182	p-value:	0.0000
GROSS_MARGIN		**GROSS_MARGIN**		**GROSS_MARGIN**		**GROSS_MARGIN**		**GROSS_MARGIN**	
ADF Stat:	-0.7856	ADF Stat:	-2.4665	ADF Stat:	-2.3950	ADF Stat:	-3.9501	ADF Stat:	-3.2553
p-value:	0.8233	p-value:	0.1238	p-value:	0.1431	p-value:	0.0017	p-value:	0.0170
OPER_MARGIN		**OPER_MARGIN**		**OPER_MARGIN**		**OPER_MARGIN**		**OPER_MARGIN**	
ADF Stat:	-1.1195	ADF Stat:	-0.4486	ADF Stat:	-1.4843	ADF Stat:	-3.2295	ADF Stat:	-0.7152
p-value:	0.7072	p-value:	0.9017	p-value:	0.5413	p-value:	0.0183	p-value:	0.8428
RETURN_ON_ASSET		**RETURN_ON_ASSET**		**RETURN_ON_ASSET**		**RETURN_ON_ASSET**		**RETURN_ON_ASSET**	
ADF Stat:	-4.1180	ADF Stat:	1.0293	ADF Stat:	-2.7130	ADF Stat:	-4.0290	ADF Stat:	-0.9548
p-value:	0.0009	p-value:	0.9946	p-value:	0.0718	p-value:	0.0013	p-value:	0.7694
NET_OPERATING_ASSETS		**NET_OPERATING_ASSETS**		**NET_OPERATING_ASSETS**		**NET_OPERATING_ASSETS**		**NET_OPERATING_ASSETS**	
ADF Stat:	-1.4654	ADF Stat:	-8.9420	ADF Stat:	-1.9303	ADF Stat:	-7.4177	ADF Stat:	-5.0007
p-value:	0.5506	p-value:	0.0000	p-value:	0.3180	p-value:	0.0000	p-value:	0.0000
RNOA									
RNOA		**RNOA**		**RNOA**		**RNOA**			
ADF Stat:	-4.6911	ADF Stat:	-2.7471	ADF Stat:	-2.7756	ADF Stat:	-2.8594	ADF Stat:	0.0386
p-value:	0.0001	p-value:	0.0662	p-value:	0.0618	p-value:	0.0503	p-value:	0.9616
RETURN_COM_EQY		**RETURN_COM_EQY**		**RETURN_COM_EQY**		**RETURN_COM_EQY**		**RETURN_COM_EQY**	
ADF Stat:	-3.1478	ADF Stat:	2.0171	ADF Stat:	-3.0993	ADF Stat:	-2.4733	ADF Stat:	-3.4202

p-value:	0.0232	p-value:	0.9987	p-value:	0.0266	p-value:	0.1221	p-value:	0.0103
EBIT		**EBIT**		**EBIT**		**EBIT**		**EBIT**	
ADF Stat:	-20.5749	ADF Stat:	0.6589	ADF Stat:	-9.8174	ADF Stat:	-5.2013	ADF Stat:	-5.9154
p-value:	0.0000	p-value:	0.9890	p-value:	0.0000	p-value:	0.0000	p-value:	0.0000
INT_COVERAGE_RATIO		**INT_COVERAGE_RATIO**		**INT_COVERAGE_RATIO**		**INT_COVERAGE_RATIO**		**INT_COVERAGE_RATIO**	
ADF Stat:	-0.4689	ADF Stat:	-1.5526	ADF Stat:	-1.1996	ADF Stat:	-2.0358	ADF Stat:	-1.9635
p-value:	0.8979	p-value:	0.5073	p-value:	0.6738	p-value:	0.2711	p-value:	0.3028
TOT_DEBT_TO_TOT_EQY		**TOT_DEBT_TO_TOT_EQY**		**TOT_DEBT_TO_TOT_EQY**		**TOT_DEBT_TO_TOT_EQY**		**TOT_DEBT_TO_TOT_EQY**	
ADF Stat:	-2.2004	ADF Stat:	1.0980	ADF Stat:	-2.4357	ADF Stat:	-2.4276	ADF Stat:	-1.9586
p-value:	0.2061	p-value:	0.9952	p-value:	0.1319	p-value:	0.1341	p-value:	0.3050
TOT_DEBT_TO_TOT_CAP		**TOT_DEBT_TO_TOT_CAP**		**TOT_DEBT_TO_TOT_CAP**		**TOT_DEBT_TO_TOT_CAP**		**TOT_DEBT_TO_TOT_CAP**	
ADF Stat:	-1.9458	ADF Stat:	-2.4313	ADF Stat:	-2.6304	ADF Stat:	-2.1853	ADF Stat:	-2.1660
p-value:	0.3108	p-value:	0.1331	p-value:	0.0869	p-value:	0.2116	p-value:	0.2188
LT_DEBT_TO_TOT_CAP		**LT_DEBT_TO_TOT_CAP**		**LT_DEBT_TO_TOT_CAP**		**LT_DEBT_TO_TOT_CAP**		**LT_DEBT_TO_TOT_CAP**	
ADF Stat:	-2.7003	ADF Stat:	-2.2204	ADF Stat:	-2.6916	ADF Stat:	-7.5088	ADF Stat:	-2.0407
p-value:	0.0740	p-value:	0.1989	p-value:	0.0755	p-value:	0.0000	p-value:	0.2690
TOTAL_EQUITY		**TOTAL_EQUITY**		**TOTAL_EQUITY**		**TOTAL_EQUITY**		**TOTAL_EQUITY**	
ADF Stat:	0.4252	ADF Stat:	23.3467	ADF Stat:	-0.7480	ADF Stat:	-3.6626	ADF Stat:	-3.8360
p-value:	0.9824	p-value:	1.0000	p-value:	0.8339	p-value:	0.0047	p-value:	0.0026
PX_VOLUME		**PX_VOLUME**		**PX_VOLUME**		**PX_VOLUME**		**PX_VOLUME**	
ADF Stat:	-1.8523	ADF Stat:	-1.4908	ADF Stat:	-1.1786	ADF Stat:	0.2502	ADF Stat:	-4.3022
p-value:	0.3548	p-value:	0.5381	p-value:	0.6828	p-value:	0.9749	p-value:	0.0004
52w_max_CP		**52w_max_CP**		**52w_max_CP**		**52w_max_CP**		**52w_max_CP**	
ADF Stat:	-4.4883	ADF Stat:	-2.4259	ADF Stat:	-4.6685	ADF Stat:	-2.7089	ADF Stat:	-5.1380
p-value:	0.0002	p-value:	0.1346	p-value:	0.0001	p-value:	0.0725	p-value:	0.0000
52w_min_CP		**52w_min_CP**		**52w_min_CP**		**52w_min_CP**		**52w_min_CP**	
ADF Stat:	-4.8532	ADF Stat:	-1.8479	ADF Stat:	-5.0181	ADF Stat:	-2.2811	ADF Stat:	-0.4549
p-value:	0.0000	p-value:	0.3570	p-value:	0.0000	p-value:	0.1781	p-value:	0.9005
VMA		**VMA**		**VMA**		**VMA**		**VMA**	
ADF Stat:	-1.4326	ADF Stat:	-0.8266	ADF Stat:	-5.3643	ADF Stat:	-5.5335	ADF Stat:	0.1867
p-value:	0.5665	p-value:	0.8112	p-value:	0.0000	p-value:	0.0000	p-value:	0.9715

Table 6: ADF Output - Accounting Data, Treated

C.1.3 ADF - Economic Data - Original

BZSTSETA INDEX	
ADF Stat:	-1.8412
p-value:	0.3602
BZAD3M INDEX	
ADF Stat:	-1.5480
p-value:	0.5096
GEBR02Y INDEX	
ADF Stat:	-0.6056
p-value:	0.8697
GEBR10Y INDEX	
ADF Stat:	-1.2205
p-value:	0.6647
BZRTRTSA INDEX	
ADF Stat:	-1.7057
p-value:	0.4282
EHUPBR INDEX	
ADF Stat:	-2.1068
p-value:	0.2418
BZTWBALW INDEX	
ADF Stat:	-2.7383
p-value:	0.0676
BZMS1 Index	
ADF Stat:	-0.1193
p-value:	0.9475
BZMS2 Index	
ADF Stat:	-0.5529
p-value:	0.8813
BZMS3 Index	
ADF Stat:	0.2359
p-value:	0.9742
BZFDTMON INDEX	
ADF Stat:	-4.8775
p-value:	0.0000
BZPIIPCM Index	
ADF Stat:	-4.2236
p-value:	0.0006
BRLUSD CURNCY	
ADF Stat:	-1.4464
p-value:	0.5598
BRLCNH CURNCY	
ADF Stat:	-1.9399
p-value:	0.3135
BRLARS CURNCY	
ADF Stat:	3.6842
p-value:	1.0000
BZGDQOQ% Index	
ADF Stat:	-2.9650
p-value:	0.0383
BZFGCCSA Index	
ADF Stat:	-2.2790
p-value:	0.1788
BZIPTLSA Index	
ADF Stat:	-2.8305
p-value:	0.0541
BRCCOUHT Index	
ADF Stat:	-0.0383
p-value:	0.9552

Table 7: ADF Output - Economic Data, Original

C.1.4 ADF - Economic Data - Treated

BZSTSETA INDEX	
ADF Stat:	-2.0340
p-value:	0.2719
BZAD3M INDEX	
ADF Stat:	-2.0211
p-value:	0.2774
GEBR02Y INDEX	
ADF Stat:	-3.5685
p-value:	0.0064
GEBR10Y INDEX	
ADF Stat:	-4.3939
p-value:	0.0003
BZRTRTSA INDEX	
ADF Stat:	-3.2322
p-value:	0.0182
EHUPBR INDEX	
ADF Stat:	-1.7348
p-value:	0.4133
BZTWBALW INDEX	
ADF Stat:	-6.8048
p-value:	0.0000
BZMS1 Index	
ADF Stat:	-1.4681
p-value:	0.5492
BZMS2 Index	
ADF Stat:	-1.8181
p-value:	0.3715
BZMS3 Index	
ADF Stat:	-4.4111
p-value:	0.0003
BZFDTMON INDEX	
ADF Stat:	-2.8521
p-value:	0.0512
BZPIIPCM Index	
ADF Stat:	-2.4677
p-value:	0.1235
BRLUSD CURNCY	
ADF Stat:	-4.9063
p-value:	0.0000
BRLCNH CURNCY	
ADF Stat:	-1.4231
p-value:	0.5711
BRLARS CURNCY	
ADF Stat:	-4.4646
p-value:	0.0002
BZGDQOQ% Index	
ADF Stat:	-2.6747
p-value:	0.0785
BZFGCCSA Index	
ADF Stat:	-1.6314
p-value:	0.4668
BZIPTLSA Index	
ADF Stat:	-4.2217
p-value:	0.0006
BRCCOUHT Index	
ADF Stat:	-1.9866
p-value:	0.2925

Table 8: ADF Output - Economic Data, Treated

C.2 Scatter-plots of Input Variables

C.2.1 Scatter-plots of Input Variables, Original Data

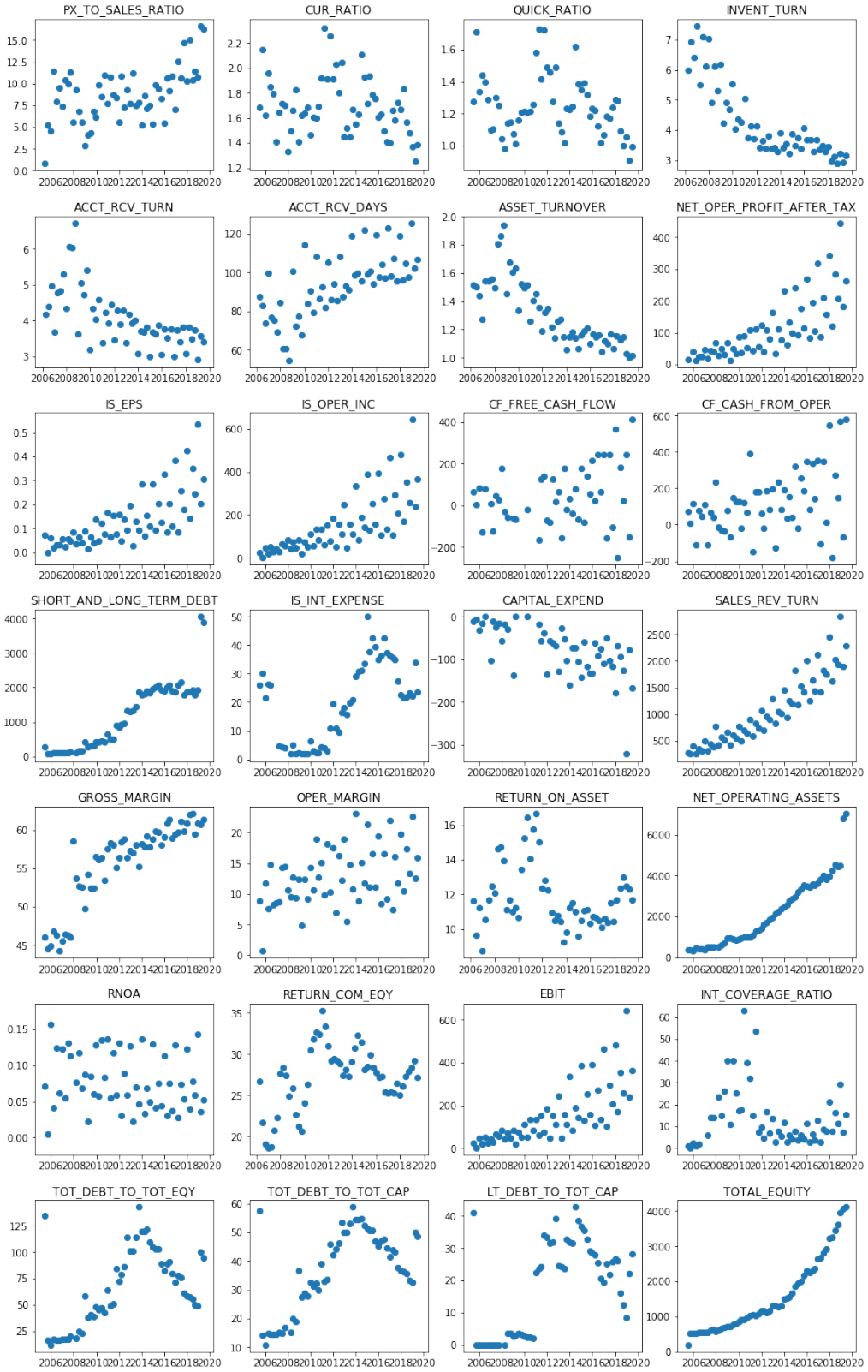


Figure 11: Scatterplot of the Input Variables, LREN3

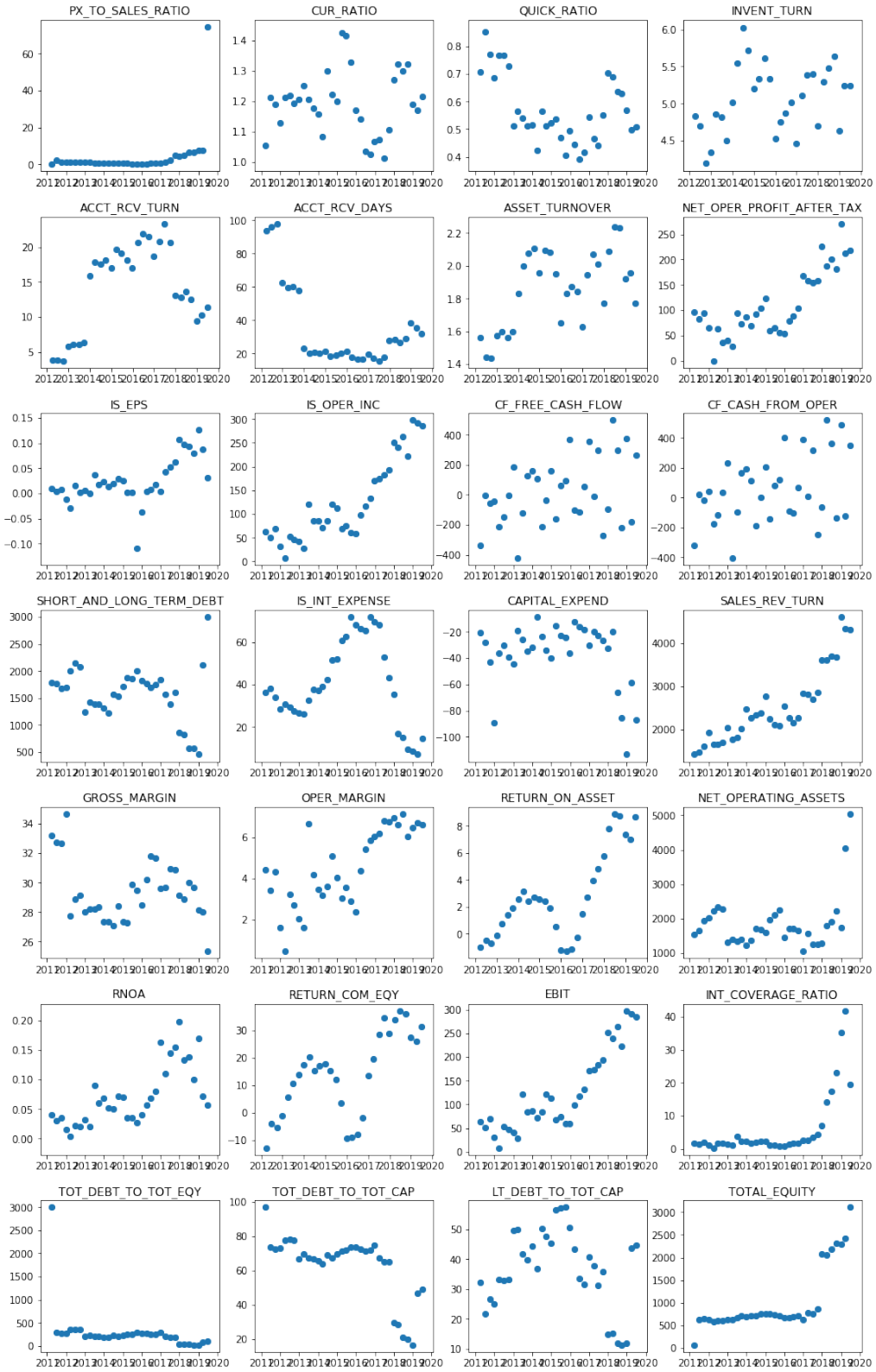


Figure 12: Scatterplot of the Input Variables, MGLU3

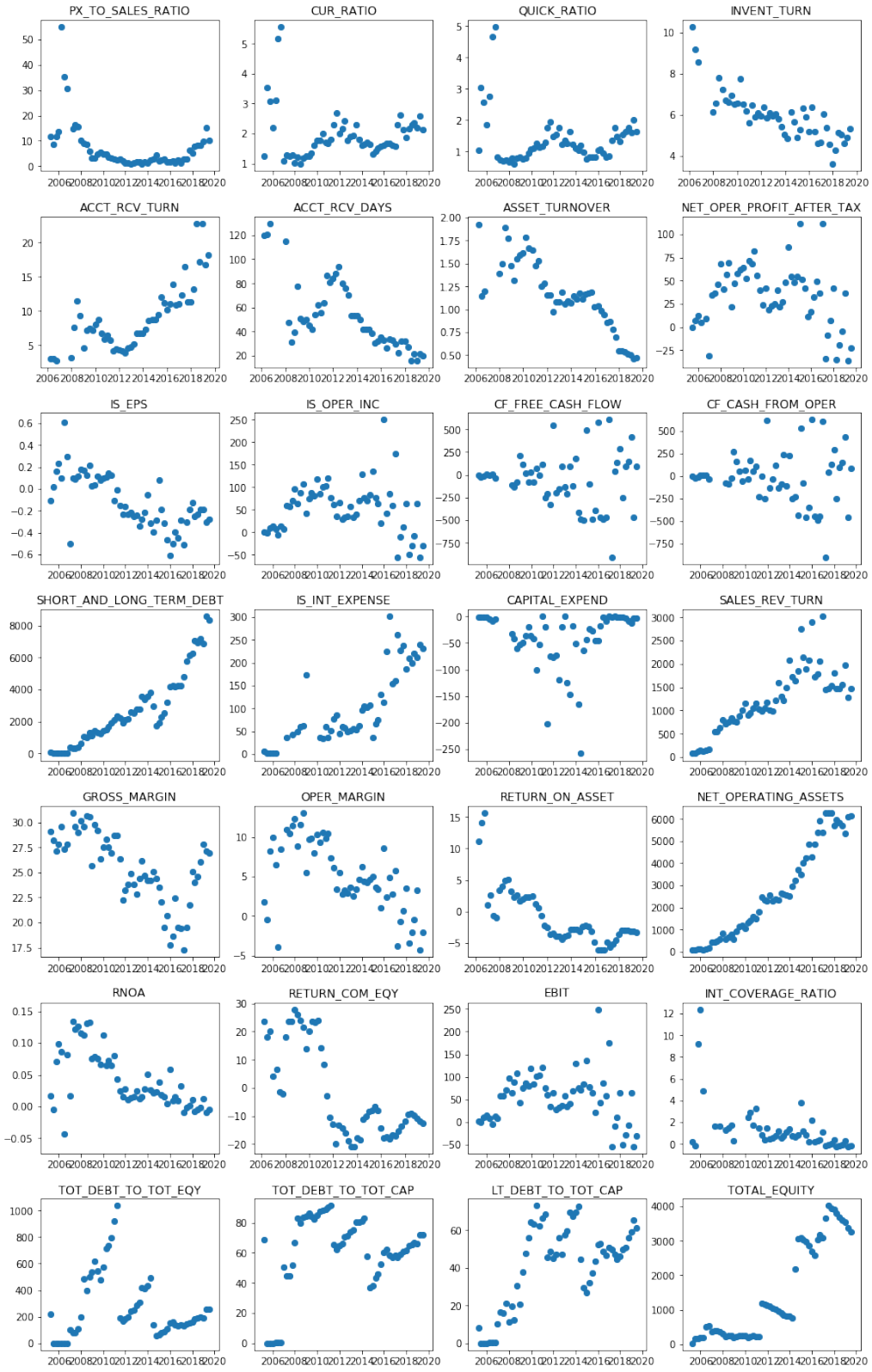


Figure 13: Scatterplot of the Input Variables, BTOW3

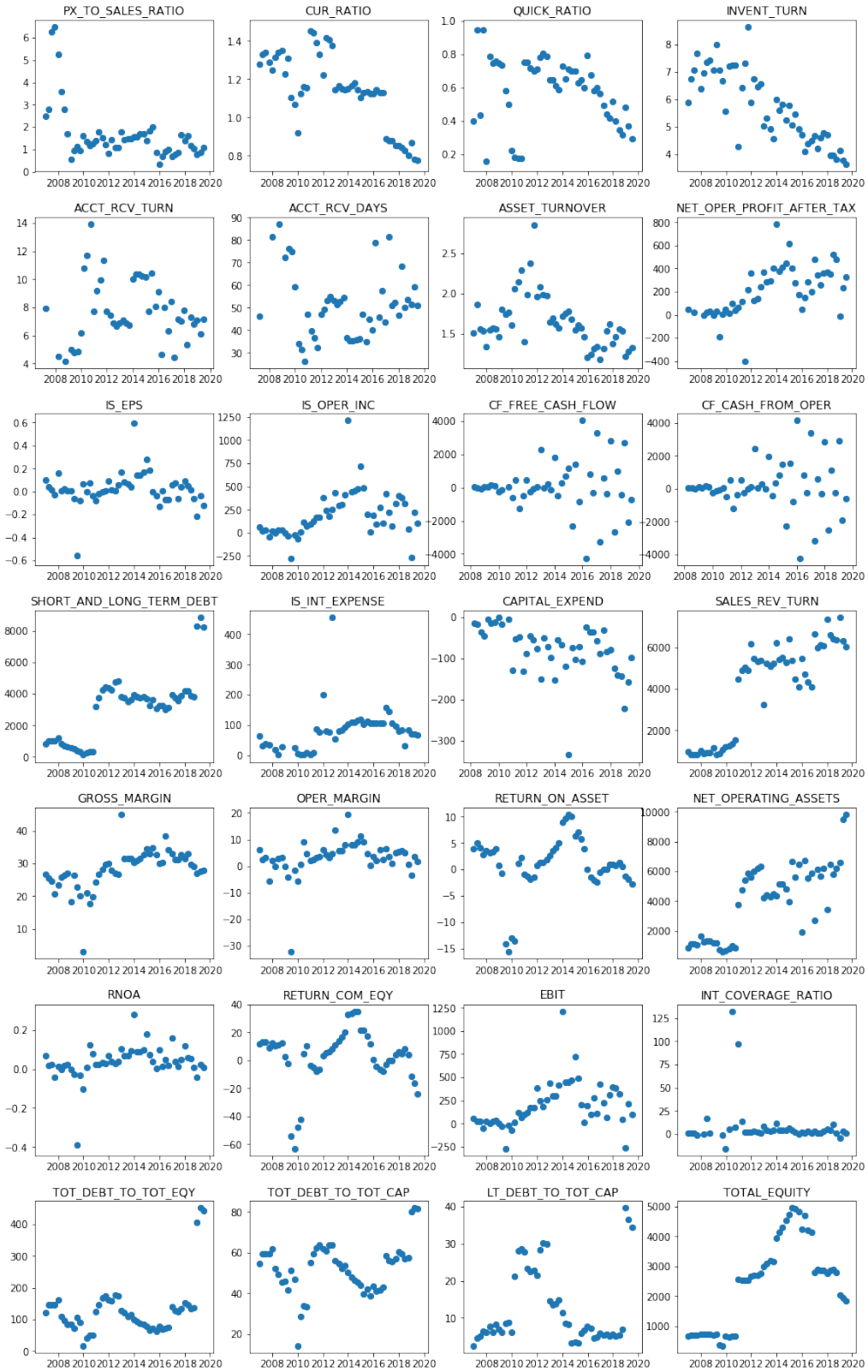


Figure 14: Scatterplot of the Input Variables, VVAR3

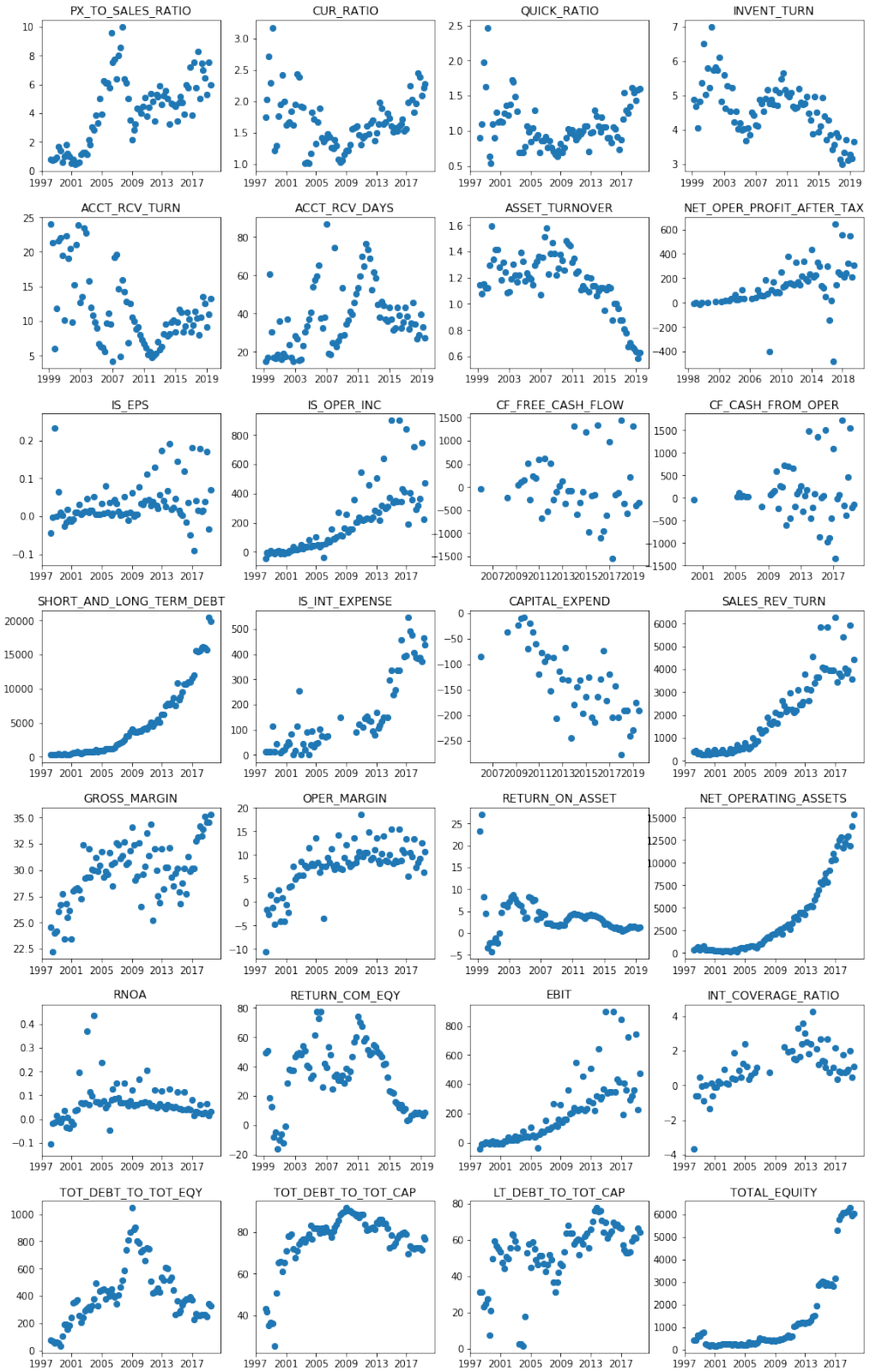


Figure 15: Scatterplot of the Input Variables, LAME3

C.2.2 Scatter-plots of Input Variables, Company-Level Treated Data

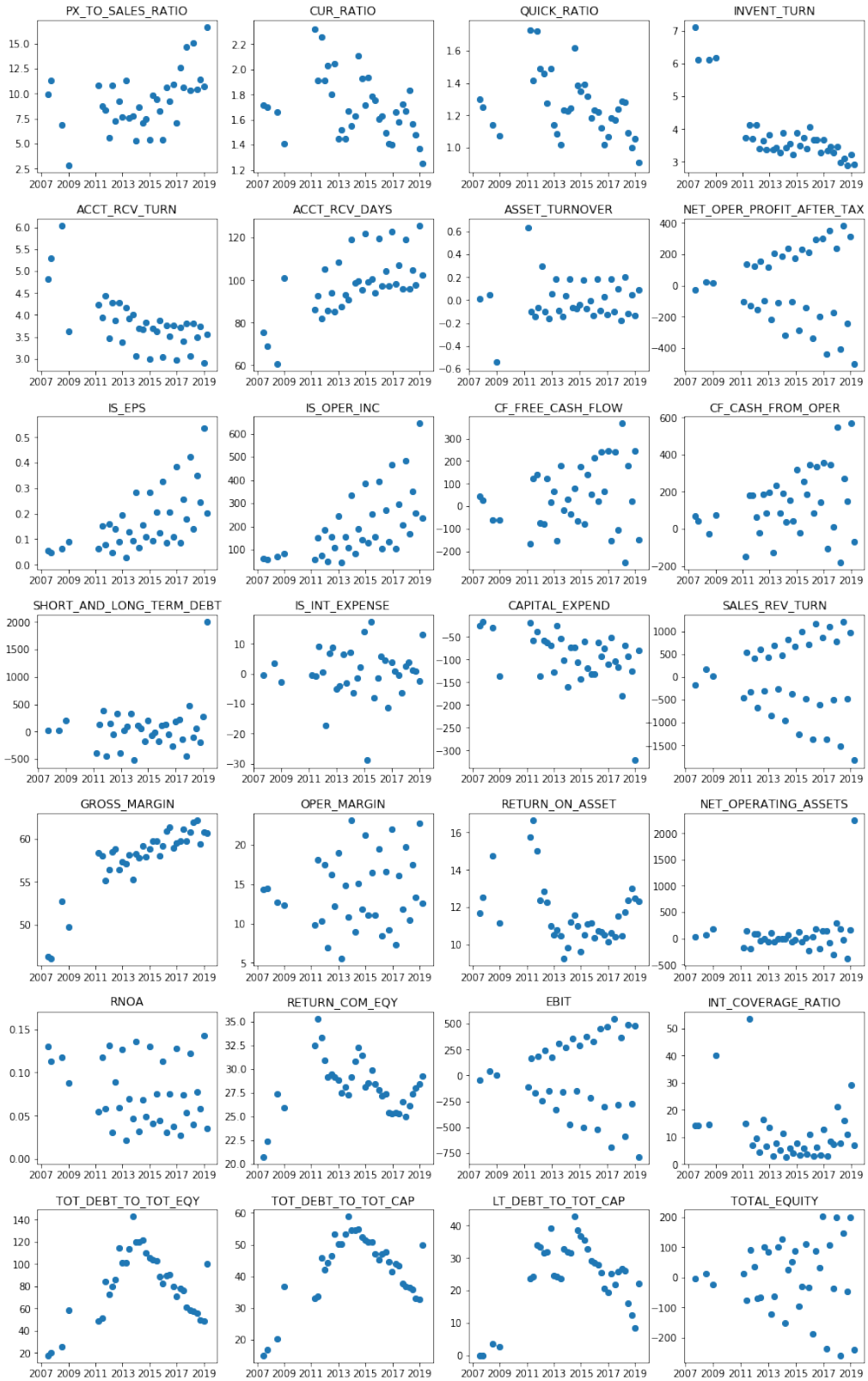


Figure 16: Scatterplot of the Input Variables, LREN3_treated

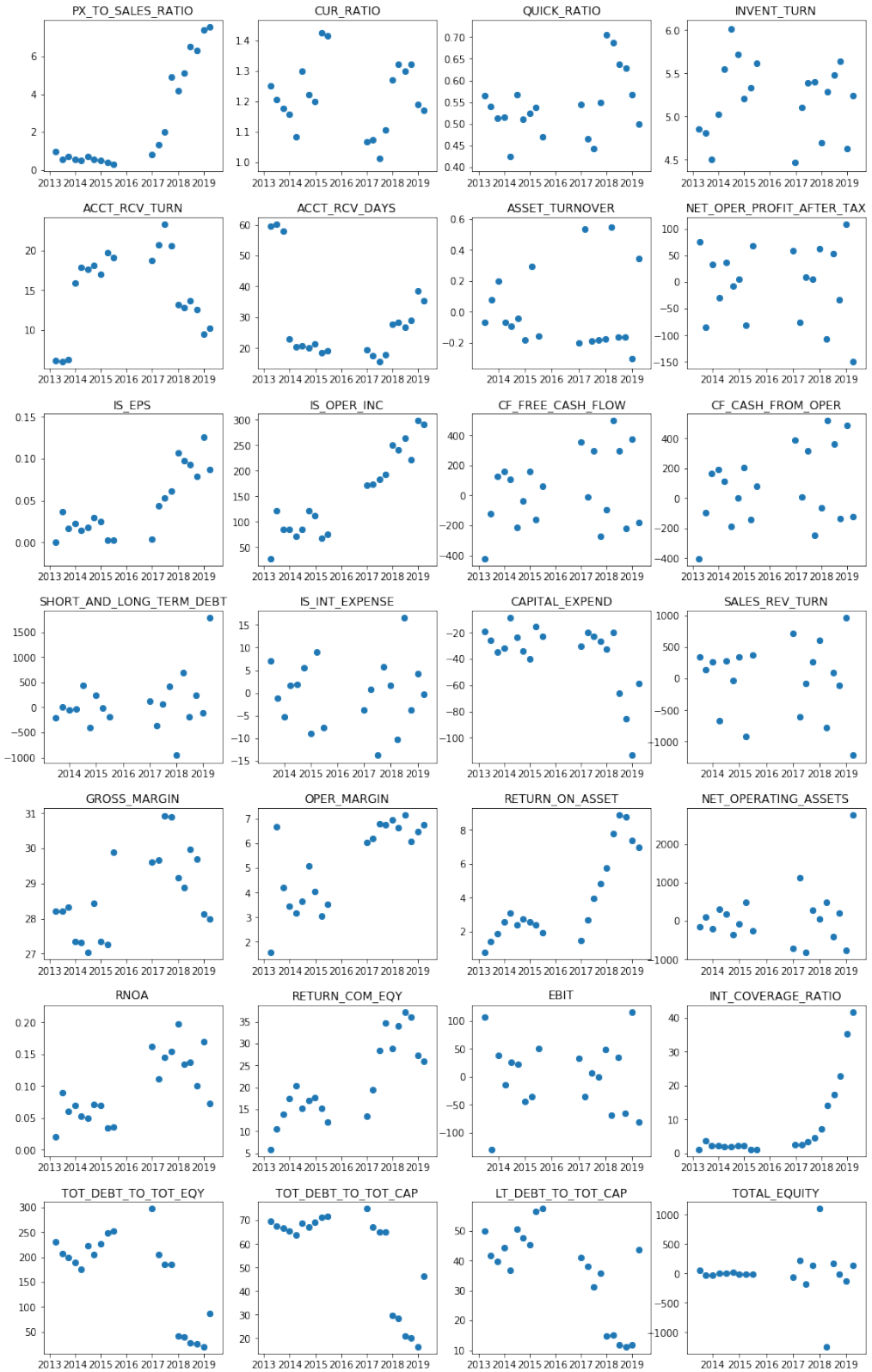


Figure 17: Scatterplot of the Input Variables, MGLU3_treated

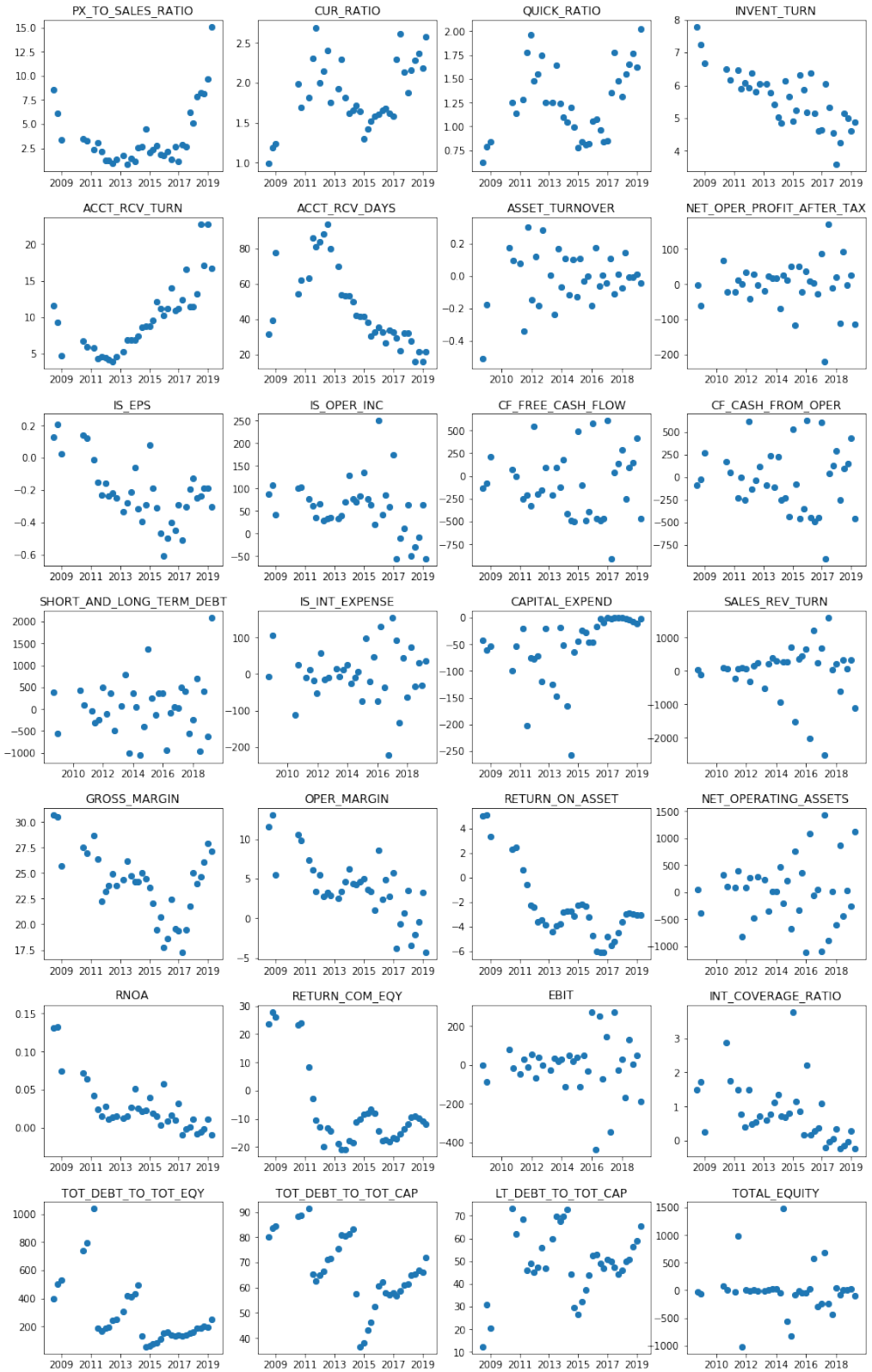


Figure 18: Scatterplot of the Input Variables, BTOW3_treated

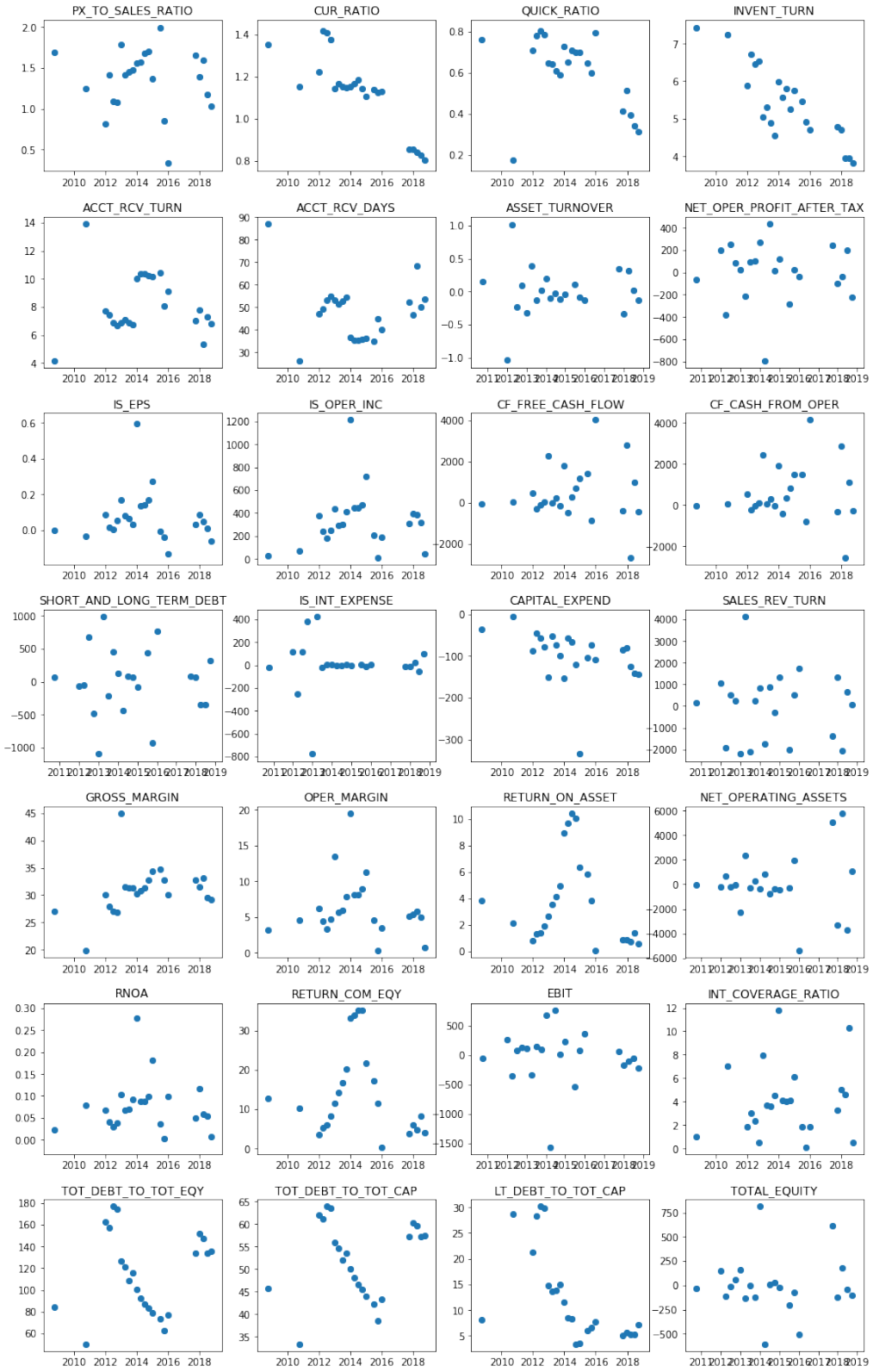


Figure 19: Scatterplot of the Input Variables, VVAR3_treated

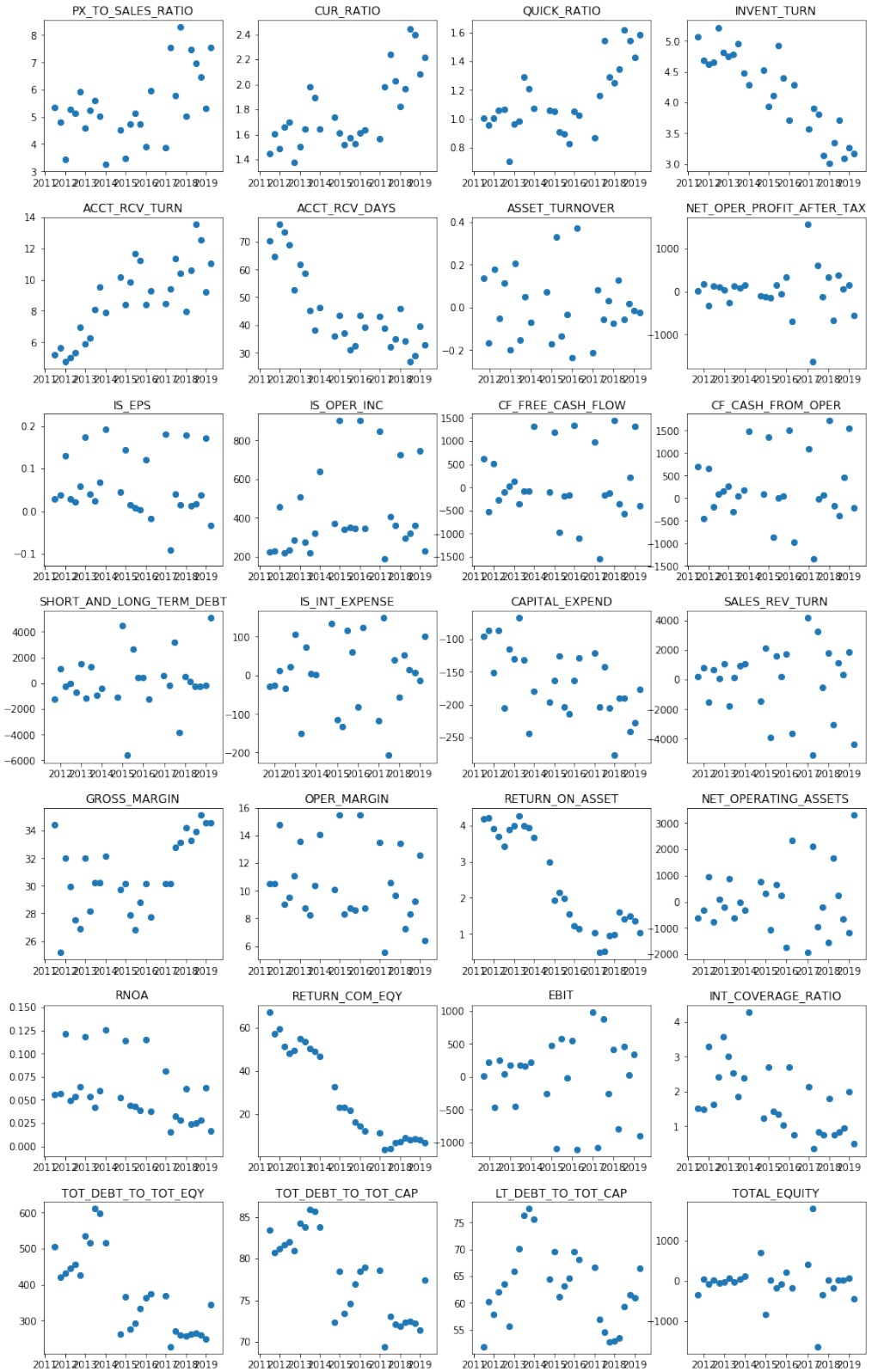


Figure 20: Scatterplot of the Input Variables, LAME3_treated

C.2.3 Scatter-plots of Input Variables, Original Economic Data

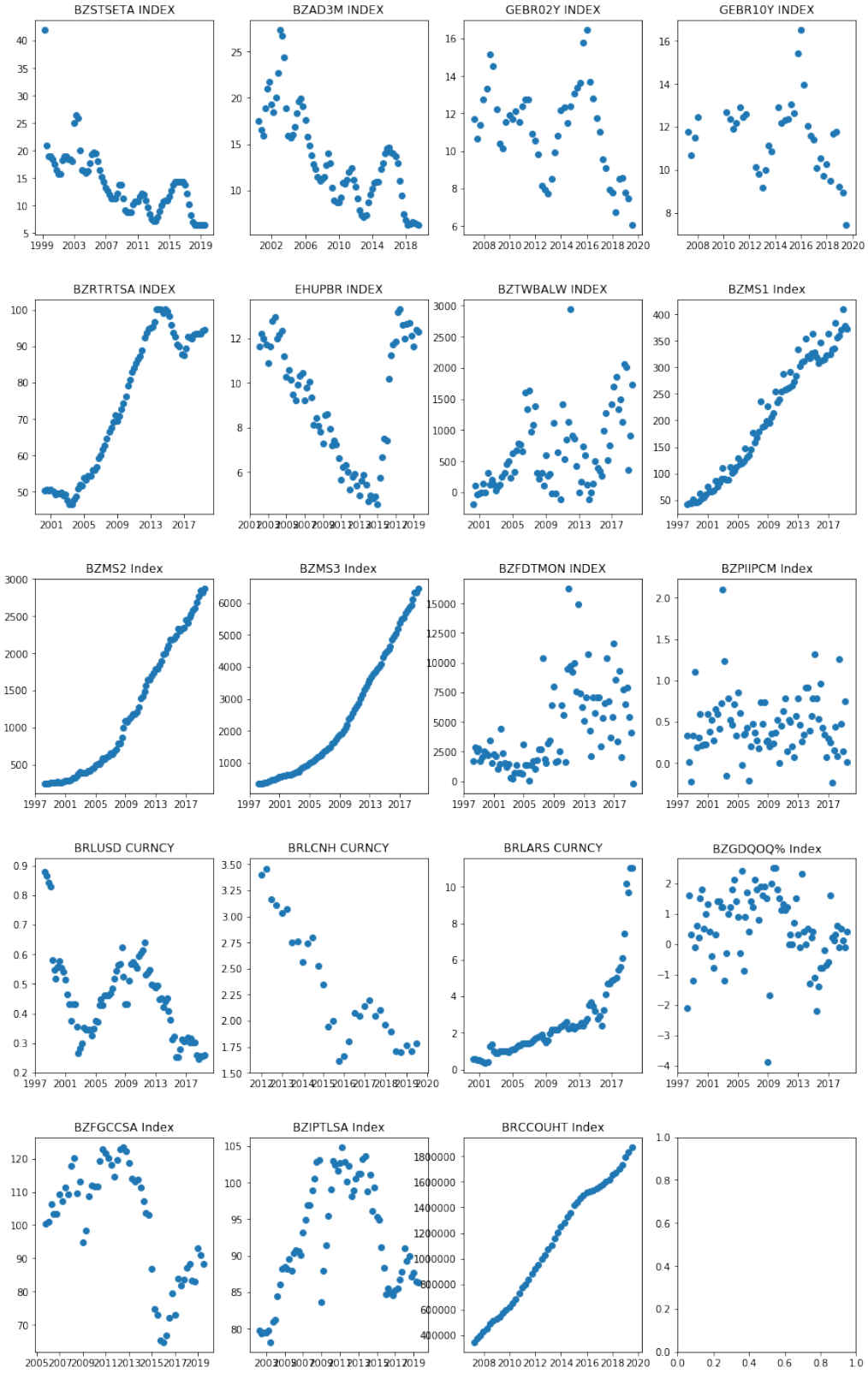


Figure 21: Scatterplot of the Economic Input Variables, Original

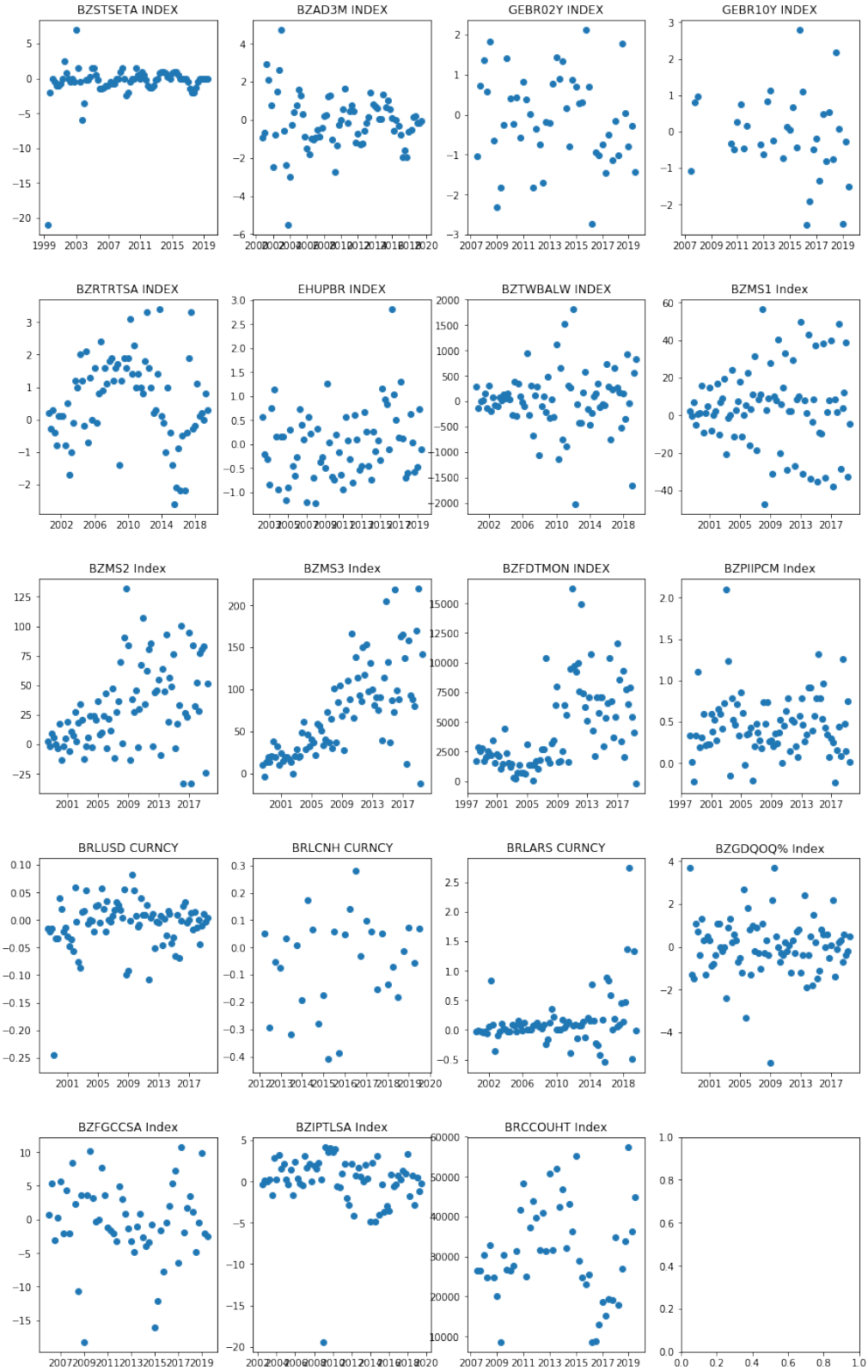


Figure 22: Scatterplot of the Economic Input Variables, Treated

D Interview Questions

1. **Is AI/ML currently used? How?**

2. **In which areas is AI/ML currently used to improve decision making?**

3. **Has AI/ML changed the structure of the organization? Have new departments emerged?**

4. **How is AI/ML used to serve the clients? How has it changed the way you do business with clients? Do you offer products to your clients that are solely based on AI/ML?**

5. **In what areas do you see the biggest improvements AI/ML bring to your company and the financial industry?**

6. **Where do you see the advantage of an AI/ML solution over an analyst? Do they work in symbiosis?**

7. **Do you use AI/ML for stock price forecasting? What models do you use?**

8. **What type of data do you use as input? What data sources do you use?**

9. **What are the drawbacks/challenges faced by AI/ML solutions in the financial industry?**

10. **Do you believe the importance of AI/ML in stock price forecasting will increase? Why?**

11. **How advanced is the development of the usage of AI/ML in the Brazilian financial industry compared to the rest of the world?**

E Interview Transcripts

E.1 Interview Transcript: E. Shayo, JP Morgan

Date: 24.06.2019
Interviewer: Lucian Dietsche
Interviewee: Emy Shayo Cherman, LatAM and Brazil Equity Strategist, JP Morgan

---

Interviewer: Hi Amy How are you?
Interviewee: I'm okay I'm sorry I wasn't able to meet you. Do you want me to call you?
Interviewer: Yeah sure if that's more convenient for you but for me it doesn't matter.
Interviewee: No, okay. Are you Brazilian?
Interviewer: No, I'm Swiss but I speak some Portuguese but I would prefer to do the interview in English because of the technical language.
Interviewee: You already have a Brazilian accent
Interviewer: Really?
Interviewee: Anything you want to tell me and I'll help you as much as possible
Interviewer: Thank you very much, first of all for taking your time and doing an interview with me. As you know I am currently writing my master's thesis in the area of artificial intelligence and portfolio construction. So the first question that I would like to ask you is, how do you currently use machine learning and artificial intelligence at JP Morgan?
Interviewee: What a tough question before answering your question, I think one important information is that is it will change our research that I mentioned. There has been a very important, push to bring more machine learning and artificial intelligence to our research being in terms of more automation, or bringing a web scraping or looking for alternative sets of data. So the whole idea is to a) differentiate our research, and to allow the analyst to focus on things that are more thematic. To me that does more than for example to write, resolve, and things that require not so much of our own brains. Let's put it this way. In my own research basically what we are trying to do is 1), this question of automation ahead of everything, so most of our quantitative reports they are today are mostly automated using different data sources. And the second thing is that we are introducing when possible, some, for example, web scraping for example when you know the Social Security reform was a very important/is a very important driver for the market. So, we got, we partnered, with a startup that dive into the web on how each congressperson is talking about the social security reform. And we were able to publish this data.
Interviewer: Okay, so basically in your team and your department you use AI and ma-

chine learning more to assist the analyst in stuff that in real time manners than actually replacing the person itself.
Interviewee: Yes. I don't think the idea is to replace the person, but is to make the person more focused on thematic reports that you know have more of talking, researching, looking for information then just *[phone line cut off]*.
Interviewer: You have you have mentioned that already a little bit in the, in the previous answer, but the structure of the department has changed, right? So you have more people that have a tech background that just the financial background.
Interviewee: We have added people with a tech background. So, there are people who, who are specialized in this. So let's say there is a smaller department within research in general, not necessarily Latam. So within the research department as a whole came the research department as a whole we have people who are specialized in this that are providing us training of different methods that are more geared to tech, to artificial intelligence for us to be able to see data in a different way and so on. So yes.
Interviewer: Okay so but is not expected from your analysts to know for instance coding themselves?
Interviewee: Not yet, of course it's a welcome thing, but what is expect is that each of us should be able to identify how to include the tech component in our own research. So we see that you know if we don't do that. Our research is going to be obsolete.
Interviewer: Has artificial intelligence and machine learning changed the way you interact with clients?
Interviewee: Yeah, so that's a very important question I find. I'm not sure if the answer is yes or no. We are always thinking about new ways, or better ways to connect to the to bring our research to the forefront of our clients. So, you know, as time passes, we believe that our will be able to pick and choose better in that we will also be able to know better, using technology, but in the interest of our clients. So for example, that client that looking at my research basically, because they might want only a quant table. So they want to look and compare valuation compare across across different factors and countries. There are other clients that are going to look for our research just because they want to go know our opinion about the country's marco policy and so on. So this is one way that artificial intelligence is very important to know your clients better. To provide better what they want. And the second thing is how really to access the client. A lot of the research today, actually all the research today is distributed via email. And you know we are still in the first steps about this, but maybe they want to go to our website or push that out of Bloomberg or some day it may be some social media. At some point, you know, this is gonna be changing more and more. But we are still also, you know, thinking on how to do this in a better way and less time to do. It exactly the same selection of what is the project that the client reads the most and personalize the research articles that the client is going to see.

Interviewer: okay so it's more used for personalization of products then to actually make a product, solely based on artificial intelligence and machine learning?
Interviewee: No, because for example, our machine they know what the client likes, so that we can personalize better. It is like when you go to Amazon, they show you what you like the most.
Interviewer: Correct.
Interviewee: So, our thing is gonna be kind of the same. We are moving into this direction.
Interviewer: Make sense so it's you're trying to customize the research reports towards the client's needs based on the data you gathered about the client in the past.
Interviewee: Yes.
Interviewer: Ok.
Interviewee: and we already get this data. Okay? *[Audio unclear]* on how our data is accessed if it is email, internet, how is it or if they are getting it from someone else such as a forward e-mail. If they opened the e-mail or not, what are they reading, who are they. We already get a lot of data. On a weekly basis every analyst gets a report about that.
Interviewer: Ok, so the you based on the answer so far the biggest improvements that AI and machine learning brings to the company is the personalization of the products that you offer.
Interviewee: Yes.
Interviewer: Okay, so the next question is do you see an advantage of AI and machine learning solutions over an analyst, or do you think that even in the future they will work in symbiosis meaning more together than one or the other?
Interviewee: I still think that you will always need an analyst. Because, it is difficult to say. Maybe? I work with Latin America. *[Audio unclear]* In the Latin American market they are policy direction. It is funny, because if you go to any conference that you go, any event that you go, with an investment in Latin America, there is a very big policy component. And I think it's difficult for us all technology and etc to really be able to do this analysis of how policy is going, what are the different elements and so on, of politics, policy and so on. That is permeating investments. I might be totally wrong about this but from my point of view of Latin America strategy. I think we will still need an analyst in the future.
Interviewer: Okay, do you currently use AI and machine learning for stock price forecasting meaning that you actually have models that are doing some price prediction?
Interviewee: We have models that do price prediction for indexes and so on. I am not sure if they have a integrated component of machine learning and AI, not yet.
Interviewer: Do you think this is something that will change in the near future?
Interviewee: Yes, maybe.

Interviewer: Next question would be, what type of data do you use as an input for your machine learning tools? You touched a little bit on that already. Is there something that you could add to that question?
Interviewee: I don't have machine learning things just yet. We are incorporating this. I don't know how to answer this question.
Interviewer: It goes more into the direction.
Interviewee: What was it, you said that I already mentioned?
Interviewer: You said that you used web scrapping to analyze the policy makers sentiment about us specific law.
Interviewee: Yeah, but that is not something that we did, right? This is someone who did it for us.
Interviewer: Ah, so it was a third party product that you you added into your report, or that you used in order to make forecasts?
Interviewee: Yes, exactly.
Interviewer: Okay, next question would be, what are the challenges faced by AI and machine learning solutions in the financial industry?
Interviewee: The challenges are a lot more in terms of mentality and learning more about this proposal. Okay so many of us are not the generation that grew up incorporating technology into our research. So I think the biggest challenge by far has been the mindset of embracing technology and bringing it on a daily basis really, to the work that we do, not just once in a while. But you know fully incorporate in the business. And I think that we are especially the most senior analyst we have a view that is not going to go back and we need to incorporate this. But the biggest challenge is the question of mindset. We are being provided with many tools, with many trainings, with many resources to embrace this. It is just a matter for us to have the mindset, and understand how we can change and improve our research.
Interviewer: Are there initiatives in the bank already that are embracing like cross to cross team interactions? Because where I see difficulties in stuff like that is that you have an analyst team and you have a tech team but they don't they don't necessarily talk to same language or they might not be in the same team. Are there initiatives to try to put a data scientist in all of the analyst teams.
Interviewee: I think it would be very costly to put a data scientist in every analyst team, but they are available for us. A very large team who are there to are there to help the analysts and the interaction is being stimulated every time more. So I don't think this is a challenge.
Interviewer: The question was more in that direction there are initiatives, trying to improve that cross-team relationship?
Interviewee: Yes, there are initiatives. Of course the relationship is actually because we have within research a tech hub. It is not only that we have to tech team inside the

bank, we have a tech team within the research department that we can be be interacting with them.

Interviewer: Okay makes sense. And then the last question that I would like to know if you think there is a difference between the development or the stage of development of AI and machine learning usage in Brazil, compared to the financial industry of the rest of the world?

Interviewee: So I just know about us right. I don't think we are behind to be honest with you, I think that that whatever is presented globally is presented to us at the same time and our analysts are always very flexible in terms of adopting new technologies they like to adopt new technologies. So I see that we are actually relatively late, to be honest with you. No, I don't think so. Especially, I think that JPMorgan being a global bank. We emerged more to what happens outside of here, it makes it different that we are already incorporated in the whole framework that is being brought to research in terms of technology here and abroad at the same time.

Interviewer: So the the roll outs JP Morgan of new initiatives and products are mostly done globally not first in the US or testing in the market and then done in the rest of the world.

Interviewee: No, it is done globally. Then everyone adopts it at a different time every new technique we have now available at the same time. As I said, the challenges is the analyst itself more than the framework, the tools available, or how to implement it and so on.

Interviewer: Okay, make sense. Great, so that is it say from my end. Do you have any questions for me?

Interviewee: No, I have a few issues. Lucian, I don't know if I want to publish it. I might be providing information to you that I am not sure if I can share it with you. *[Audio unclear].* So when you are writing this paper or when you are going to publish it, let me know. I don't think there will be a problem, but just to be sure. *[Audio unclear]* So please share with me whatever it is.

Interviewer: Okay, so you would like to have the the part in, in case I'm using this interview in my thesis you would like to have a look at the part where I'm using it before publishing?

Interviewee: Yes, you know, as I said, just let me take a look. I don't think it is going to be a problem. I actually cleared our internal Compliance before, but considering that the topic goes well beyond what is my own coverage, you know, which I thought that we were going to talk more about Brazilian stocks than anything else. So let me know.

Interviewer: Okay, great. No, I can send you also summary of the talk if you want and then you say, no, I didn't mean that or stuff like that. So, then you can clear that beforehand. And the way it's going to be incorporated in my master's thesis is more in a way that I'm, I'm doing a general introduction of stock forecasting and machine learning,

and then I'm going to make a summary of where the industry in Brazil currently is at. And then I am using academic research in order to build an AI and machine learning tool to try to make stock price forecasting.

Interviewee: Wonderful.

Interviewer: It's like a snapshot of okay here the industry is and now let me try to come up with something that might help to do some stock price forecasting, that's the that's the general idea that's the outline of the paper.

Interviewee: I would love to see your master thesis as a whole, who knows. *[Audio unclear]*

Interviewer: Depending on how the research goes I will be sharing it with you for sure.

Interviewee: I would love that and next time when we talk you come here. I'm sorry, because Monday is very complicated. I know you offered Tuesday and Wednesday, but I'm not gonna be in São Paulo. Next time we see each other.

Interviewer: Perfect, sounds good to me. Thank you very much for your time and for your helpful information, I wish you a great morning and the rest of the day. Thank you so much. Bye, bye.

Interviewee: All the best.

Interviewer: To you as well.

E.2 Interview Transcript: R. Catalan, Private Investor

Date: 02.08.2019 *translated from Portuguese*
Interviewer: Lucian Dietsche
Interviewee: Rodrigo Catalan

Interviewer: Hi Rodrigo, how are you?
Interviewee: Hey Lucian, how are you?
Interviewer: Do you prefer the interview in English or Portuguese?
Interviewee: I can try to explain some things in English, but I prefer Portuguese.
Interviewer: Ok, then I'll try to ask the questions in Portuguese.
Interviewer: Thank you very much for doing this interview with me.
Interviewee: No worries.
Interviewer: Can I record the conversation to do a transcript?
Interviewee: Yes, you can.
Interviewer: Immo told me, that you are developing some trading algorithms. Can you tell me a little bit about how they work and which data you are using?
Interviewee: I am not developing the algorithms, I am a user. There are many different trading services in Brazil. I chose one of them and I educated myself about the parameters that I can choose.
Interviewer: So the trading system is rules-based, so it does the orders automatically?
Interviewee: Yes, the system executes orders automatically.
Interviewer: So you only have a couple of configurations that you have to do.
Interviewee: Exactly.
Interviewer: In which financial products are you investing? Are these only shares, or do you also trade bonds or currencies?
Interviewee: I operate in the futures market. Futures of the Bovespa Index, and Futures of Dollars.
Interviewer: Are there any analysis that the algorithm does and shows to you afterwards to improve the configuration of the algorithm or does he do everything automatically?
Interviewee: No the algorithm does not show any indication. Every user has to study and backtest the configuration himself.
Interviewer: So you chose the configurations, do the backtesting, and then based on that the algorithm does the investing?
Interviewee: Exactly, I can give you an example. Do you know an indicator MACD?
Interviewer: Yes.
Interviewee: These indicators are applied using candlestick graphics. The example that I'll give is very simple, so in reality it can be way more complex. So for instance

a robot that operates RSI. I configure my options for the algorithm. So for instance I can choose, that I would like to operate in a graphic of five minutes. Other people might choose to operate in graphics of ten minutes, twenty minutes. After choosing the time I choose if the RSI is above 80% he sells the stock, others might choose the RSI until 90% to make a sale. In my configuration I might sell 10 futures, another person might sell 100. So what I am trying to say is that the robot is just doing the executions according to the configurations that I have done. I can change the time interval of the graphic, I can change the metric, I can choose the quantity of futures - the amount of risk I'm willing to take, configure stop loss or stop gain. These are all parameters that I can define in the setup of the strategy, then do a backtesting, and based on that I can see how profitable this strategy would have been in the past. After the backtesting I can further put the same configurations in a live-environment without actually doing orders.

Interviewer: So there are three phases: First, backtesting. Second, to gain confidence you do a live testing without investment and then if you think you have a working algorithm you start investing.

Interviewee: Exactly like that.

Interviewer: There is a checkbox in the configuration that lets you decide whether you would like to execute an order or not. So I can put the algorithm live but without execution of the order. The software then creates a log with the theoretical purchase overviews in case the algorithm is live but without execution. If the checkbox is selected, the order will be executed according to the rules.

Interviewee: Can you change part of the code in the trading software?

Interviewer: No, there is no coding involved, you can only select the parameters.

Interviewee: Did you do systematic tests to do your initial configuration? Meaning did you create a robot on top of the robot to find out with AI which are the best parameters or did you chose them randomly and them optimized them according to the backtesting?

Interviewee: I did a thing called "caseira" for myself, other people operate differently though. So I downloaded a dataset of historic prices minute for minute and then in Excel I create some plots to analyze what happens if I change the parameters a little bit. So I do my analysis in Excel based on historic prices. This was the way I used to do it.

Interviewer: Makes sense.

Interviewee: I know other people that use metatraders that allow you do write some pseudocode, in which you can choose the parameters and the metatrader itself has a database of the prices and executes the backtest accordingly. But I don't have the metatrade, I do it in Excel.

Interviewer: Just out of curiosity, how was your perfromance in the past year?

Interviewee: This is a very interesting question, because in the beginning of the year, January, February, I had spectacular performances like doubling my capital. So 100% of profit per month. March and April were complicated months, because especially in Brazil we have market variables that are not exact. What I would like to say with that is that for the Brazilian stock exchange passing a bill in the parliament can either have a positive or negative impact that cannot be foreseen. I don't know if you are accompanying the political development here in Brazil.
Interviewer: Yes, some of it I accompanying.
Interviewee: We are currently passing a reform of the government and that reform has created a lot of unpredictable market volatility. So even if all indicators are showing one logic, the determination of a congressman can change everything. So March and April were horrible months.
Interviewer: I understand.
Interviewee: In my opinion the movements were not due to logical factors. In the more recent months May, June, we have a very high volatility because of some stuff the president said. As a result increased market volatility, because of some news, it is possible, that the stop loss is triggered within a couple of minutes. For example, it is possible, that you have BRL 100-200 profit and then because of some conversation of five minutes you lose 1000. In these moments I tell even to my friends, exact variables have a small impact, because the political variables have a huge impact. We have some groups on WhatsApp and Telegram of investors that exchange some ideas and in the last couple of months everyone said that this is the logic, but what will be reality may differ significantly. So what happened is that everyone thinks a lot more conservative in their investment strategies, because we cannot forecast the development. So it can take only five minutes to lose all of your money. So it has been very complicated these days. **Interviewer:** The financial market has some trends that can be forecasted, but it is like that. If Trump sends out a Tweet about some import taxes, the market already falls. So this is actually a signal of market efficiency, because the news are incorporated in the price very quick.
Interviewee: Exactly. If you are a professional trader and are online all the time you might even benefit from these market swings, but if you are a robot you don't have that opportunity, because the robot is programmed. It executes exactly the order that you put. If you say that you would like to buy at 9 and it only goes until 9.01 it doesn't execute the order, because it is exact. This is a difficulty, but also an advantage of the robots, they take out the emotions of the trader and execute. But in markets in which you have political surprises, this can be dangerous. This is the good and the bad side to operate with robots.
Interviewer: Makes sense. So have used these algorithms now for some time. I would like to know what you learned during that time?

Interviewee: Let me know how I can explain that. I learned things in various areas. So for example. I learned a lot about trading techniques like how to analyze a candle stick chart or indicators. So my technical learning was very rich in this area. The technological learning was also interesting, because I could learn how a trading platform like smartbot works and how to program stuff with it. As I have a technological background, that was very interesting. There is a third group that is very interesting to understand that a robot doesn't do magic and there is no such thing as easy money. That's were things like political aspects enter that can create surprises in the market and make you lose money. So I learned things in many different areas. Does that answer your question?
Interviewer: Yes. I think these learnings are very normal when you invest in the financial market. It hurts when you lose money, but then you have to have a strong mind to believe in your rules and continue to invest. It is unfortunately normal to now and then have big losses.
Interviewee: Exactly.
Interviewer: Now some more general questions. What do you think are the biggest opportunities of AI/ML in the financial industry?
Interviewee: I think there are many possibility. [Audio unclear]. All of the robots that we have available at the moment are exact, meaning that if you tell them 10 it is 10. It would be great to have some sort of AI that could make adjustments due to political news. There are many news broadcasting services, it would be great to have a robot that could evaluate these news and decide whether this news will have a positive or negative impact on the financial market. If I had these indications, I'm sure it can be done with AI, this news robot could give me triggers to continue or stop with the investment robots. I don't know if I was able to explain myself.
Interviewer: Makes sense. So for you it would be helpful to have an AI/ML tool on top of your robot that decided when to invest and when not.
Interviewee: No, I would like to create a new robot with AI that interprets news broadcasting. Only for that. That is where I think AI could come in. This new robot would send out triggers based on past data and these triggers would then create opportunities to buy and sell and deviate to what was programmed in the beginning of the day. During the day the robot will send triggers about irregular market activity. In normal market environment the robots work very well and do not need AI. Where we need AI is in irregular, unforeseeable market environments. So if the AI could interpret news that would be great. If I am not mistaken, today is the day they publish the payroll in the US. This news is a bomb. The Brazilian market will in a window of five minutes it will oscillate five points up and five points down in a very short amount of time. This will all depend on whether the notification of the payroll is good or not. So here is where I am saying that a robot with machine learning could read the text that was

published at 11am and give a return whether the news is optimistic or pessimistic that would be all I need. But because of the fast reaction of the market this would need to be interpreted automatically to be able to act fast enough and take advantage of it.
Interviewer: There are already news traders, but they are obviously not as fast as a robot that already does the execution.
Interviewee: But do these news trader give you a feedback whether the news was positive or negative or is it just a feeder?
Interviewer: No they are actual traders that look at news and then act upon it.
Interviewee: This is what I call professional traders.
Interviewer: Yes, of course.
Interviewer: Some of the challenges of the robots you have already mentioned. He is logical and does everything right, this is an advantage, but also in some cases a disadvantage. Is there another thing that you think a human being is better at?
Interviewee: I have never thought about that.
Interviewer: Ok then let's go to the last questions.
Interviewee: But the difference between a robot and a human is the factor of intelligence and the second difference is velocity. A robot is fast and stupid a human is slow and intelligent. If we can put together both we have something fast and intelligent. In my opinion this is the perfect combination for the financial market.
Interviewer: Do you think that in the future these robots will help the professional investors or will there be solutions that are more intelligent and are going to substitute analysts?
Interviewee: I don't think they will replace analysts. Have you ever heard of quantico, that would be very interesting if you read about that. These are quantitative funds. I did a project when I worked as a consultant for the Itaú Bank. For this consultancy I planned for them a couple of quantitative funds. This is very interesting. Quantitative funds are operated by robots and are the high frequency traders, HFT. These quantitative funds that operate as HFT have algorithms that are extremely fast and act on arbitrage, not artificial intelligence. For example he could buy a BRL/USD future in Brazil, while selling it in the US at the same time if there is a price difference. Because in a fraction of a second there is a difference in price. HFT sells and buys at the same time. They make millions of trades per second. They have a lot of technology in there and I know that they are testing things with artificial intelligence. Maybe they already have it live by now. That would be interesting to do some research on it.
Interviewer: That would be very interesting. The last question I have, do you think the development of robo trading and the usage of AI in the financial industry is as advanced as in other countries?
Interviewee: Until where I had contact with this topic, yes. In some cases Brazil is even further. But the difference is the difference in the number of users. In Brazil we

have amazing people to develop these kind of algorithms and robots, but the social situation of Brazil does not allow the usage thereof. There are a lot of many poor people in Brazil. So they cannot take part of the financial market. If you compare the number of automated trades in the US and in Brazil you will see that degree of automation in the US is far greater than in Brazil. Americans use automated trading a lot more than Brazilians. This however does not talk about the development but more about the usage.

Interviewer: But this is leads to opportunities, because if there are many people already doing something there are not that many opportunities anymore. If there are a lot of robots the market will be even more efficient. So if there are not that many people using automated trading there are more opportunities.

Interviewee: I think there are more opportunities in Brazil, yes.

Interviewer: Thank you very much for you help and giving me these insights. It was very interesting.

Interviewee: I think it is a very interesting work that you are doing. A very interesting topic.

Interviewer: I hope that I will be able to create my own robot with a long-term investment strategy. Let's see if I can find something interesting or not.

Interviewee: Very interesting.

Interviewer: Perfect. Have a great day. Bye.

Interviewee: Bye

E.3 Interview Transcript: M. Jordi, theScreener

Date: 08.08.2019
Interviewer: Lucian Dietsche
Interviewee: Markus Jordi, theScreener by e-mail

Interviewer: Does the screener currently use AI/ML? How/In which area?
Interviewee: Yes, language generated on the basis of quantitative input.
Interviewer: Are all of the reports, ratings, and risk calculations done by computers?
Interviewee: Yes, but prior to letting the computers do the calculations, we check and improve the data quality personally. This is very important as the data feed always contains mistakes that our special programs flags as suspicious and then a person checks if it is really a mistakes and corrects if necessary.
Interviewer: In what areas do you see the best performance improvement of AI/ML at theScreener?
Interviewee: For the ratings themselves we use classic computer algorithms without self learning. This is important to avoid following the crowd into bubbles.
Interviewer: Where do you see the advantage of an AI/ML solution over an analyst? Do they work in symbiosis with theScreener?
Interviewee: Computers work without emotions. This is the main advantage. Our algos have no problem in changing a rating. There is no psychological problem to admit we were wrong. This is very important to dare being sometimes against the mass and interestingly enough this is statistically often the better way.
Interviewer: Do you use AI/ML for stock price forecasting? What models do you use?
Interviewee: No stock price forecasting.
Interviewer: What type of data do you use as input?
Interviewee: We purchase a global data feed from Refinitiv (former Thomson Reuters)
Interviewer: What are the drawbacks/challenges faced by AI/ML solutions in the financial industry? When is the forecasting power of theScreener low?
Interviewee: The quantitative systems typically react with delay to trend changes. If a long bull market turns bear or vice versa often quant signals need some time (theScreener typically some weeks) to adjust.
Interviewer: Do you believe the importance of AI/ML in stock price forecasting will increase? Why?
Interviewee: Yes. The percentage of computer based trading volume clearly increases, so the importance to understand what is going on will increase too.
Interviewer: How advanced is the development of the usage of AI/ML in the Brazilian financial industry compared to the rest of the world?
Interviewee: The leading market for AI/ML is the US. All other markets are behind

the US, including Brazil.

E.4 Interview Transcript: F. Koh, former Citigroup

Date: 07.11.2019
Interviewer: Lucian Dietsche
Interviewee: Felipe Jiman Koh, former Citigroup, currently proprietary desk, Investment Research Sr. Associate

---

Interviewer: Is AI/ML currently used? How?
Interviewee: Used mostly for arbitration using ML, but the evolution would be AI.
Interviewer: In which areas is AI/ML currently used to improve decision making?
Interviewee: ML is not used to make decision making. There was a project to develop ML historical data window in specific similar time periods to make a decision. However, project was not completed and halted.
Interviewer: Has AI/ML changed the structure of the organization? Have new departments emerged?
Interviewee: ML has changed as a new arbitration business mentioned previously is based on ML.
Interviewer: How is AI/ML used to serve the clients? How has it changed the way you do business with clients? Do you offer products to your clients that are solely based on AI/ML?
Interviewee: The firm is a proprietary desk, with no clients. As such, there was no impact yet.
Interviewer: In what areas do you see the biggest improvements AI/ML bring to your company and the financial industry?
Interviewee: Arbitration and speculation. I know some project to use in investment research recommendation, but nothing concrete on the radar yet.
Interviewer: Where do you see the advantage of an AI/ML solution over an analyst? Do they work in symbiosis?
Interviewee: Haven't seen nothing working yet. However, I would believe it would help the analyst to make a decision rather than making the decision itself.
Interviewer: Do you use AI/ML for stock price forecasting? What models do you use?
Interviewee: No, everything still old school. Valuation, growth forecast, macro and micro analysis.
Interviewer: What type of data do you use as input? What data sources do you use?
Interviewee: Always reliable data from original source. We create our own data base as there is no proper data to acquire in the Brazilian exchange market.
Interviewer: What are the drawbacks/challenges faced by AI/ML solutions in the financial industry?

Interviewee: I'm not aware of.
Interviewer: Do you believe the importance of AI/ML in stock price forecasting will increase? Why?
Interviewee: Yes I believe, yet there will be a long run until it works in my vision.
Interviewer: How advanced is the development of the usage of AI/ML in the Brazilian financial industry compared to the rest of the world?
Interviewee: Brazil is a small market, in our day to day analysis. No major international player is operating in the region.

E.5 Interview Transcript: E. Magalhães, Pandhora

Date: 07.11.2019
Interviewer: Lucian Dietsche
Interviewee: Eduardo Ferrari Magalhães, Quantitative Analyst Pandhora

---

Interviewer: Is AI/ML currently used? How?
Interviewee: Currently our use of Machine Learning and Deep Learning is limited to helping us understand and model some phenomena. Mainly pointing out important variables, or helping us to verify if the available data is sufficient to model some events. We also implement some Deep Learning algorithms to verify if a complex model could bring better results, or if a simpler man-made model could be enough to explain the market behavior.
Overall, we tend to avoid making models that are based in complex machine learning algorithms because they can be fairly unpredictable with data from outside train/test environment and can be very hard to track. This is especially important because even if some models perform well, when they don't it becomes very hard to understand what is going wrong and how to make it better.
Interviewer: In which areas is AI/ML currently used to improve decision making?
Interviewee: The ML models have been used mostly in decision making on the RD area, we use them to help us improve some already deployed man-made models and to bring inspiration on other approaches to market modeling by creating some complex models' baselines and sensibility analysis to verify important variables.
Interviewer: Has AI/ML changed the structure of the organization? Have new departments emerged?
Interviewee: In general no, every researcher is under the RD department working mostly individually so despite there being researchers focused in ML and model development (such as myself) we did not change our structure.
Interviewer: How is AI/ML used to serve the clients? How has it changed the way you do business with clients? Do you offer products to your clients that are solely based on AI/ML?
Interviewee: ML for market analysis is still a new theme in Brazil, just by being a quantitative fund we already have to have a very dedicated process to introduce the investor to the "trading robots" world. That being said, we keep our ML models as a tool to make better man-made models.
Interviewer: In what areas do you see the biggest improvements AI/ML bring to your company and the financial industry?
Interviewee: Overall ML can be a very powerful tool to help us understand the market and automate some internal affairs such as social media investor profiling. But in

general, even as a quantitative fund, we tend to avoid models based purely on ML because of the unpredictability of some models.

Interviewer: Where do you see the advantage of an AI/ML solution over an analyst? Do they work in symbiosis?

Interviewee: The biggest advantage of computational models in general over analysts is the consistency. Computation models, despite of any endogenous factor, if given the same inputs will yield the same outputs, and having such impartiality is very useful specially in times of crisis. But as stated before, here in Pandhora ML models are tools for the analyst to develop a more understandable, straight-foward market model, so they do work in symbiosis.

Interviewer: Do you use AI/ML for stock price forecasting? What models do you use?

Interviewee: When we develop a baseline for such models, we tend to use LSTM Deep Learning models as they can grasp tendencies much better than some classical models such AR, ARMA, ARIMA etc.

Interviewer: What type of data do you use as input? What data sources do you use?

Interviewee: This is a very sensible information and vary a lot depending on the stock category, but overall, for equities for instance prices, indicators derived from prices, company fundaments and so on. What can be disclosed in this matter is that, so far, we use no "alternative data" for models, no satellite images or NLP approaches.

Interviewer: What are the drawbacks/challenges faced by AI/ML solutions in the financial industry?

Interviewee: The biggest drawback is the unpredictability and the incapability of understand complex models, especially Deep Learning ones. Even simpler models such as decision trees sometimes need to scale up in size to have good results, and by doing so become completely unreadable.

Interviewer: Do you believe the importance of AI/ML in stock price forecasting will increase? Why?

Interviewee: Yes, I personally believe that in couple years ML models will grow in importance, as a well calibrated ML/DL model can detect subtleties that a human or man-made model could miss.

Interviewer: How advanced is the development of the usage of AI/ML in the Brazilian financial industry compared to the rest of the world?

Interviewee: Some funds such as Giant Steps (Old Visia) use this kind of technology and are somewhat open about it, but I still think that this kind of technology is currently is in early stages of development.