Senior Thesis

A Machine Learning Approach For Forecasting Financial Bubbles

Daniel S. Londoño Bohórquez

Advisor: Silvia Takahashi, Ph.D

This thesis is submitted in partial fulfillment of the requirements for a degree in Systems and Computer Engineering.

Department of Systems and Computing Engineering
Universidad de Los Andes
Bogotá, Colombia
2022

A Machine Learning Approach For Forecasting Financial Bubbles

Systems and Computing Engineering
Universidad de Los Andes
Bogotá, Colombia

For Ermias.

Abstract

Machine Learning can be used for predicting the likelihood of modern era securities having bubble like behavior, particularly stocks. An asset is considered '*bubbled'* when its intrinsic value does not reflect at all the price to which is traded at. In order to accomplish this a wide variety of libraries, including Google's renowned Tensorflow and a well founded and updated stock market dataset can be used. The data gathering process for this project was without a doubt the biggest challenge of all. This is basically due to the fact that we are studying dead companies. The 2001 Dotcom Crash forced hundreds of companies file for chapter 11, forcing their financial data to become unavailable, even in large databases such as Bloomberg's or the SEC's.

The loss function used for evaluating the proposed neural network model was categorical cross entropy, a specialized type for one-hot encoded outputs. This model follows a pretty uncomplicated architecture, including a couple of Rectified Linear Unit neurons (ReLU), and some Softmax activation neurons. Trained with the stochastic gradient algorithm, this model performs optimally when it comes to evaluating aggressive and unfounded price volatility fluctuations of a particular stock.

Table of Contents

1. Introduction

Humanity has witnessed disasters of all kind. From wars to natural catastrophes, almost all of them have a common denominator. The roots of their inception are somehow visible before they even start. Take a look at the geopolitical conflicts around the world prior to World War II. World War I immediate aftermath introduced several financial restrictions on Germany, a country who was facing the rise of fascism, on its purest form. The U.S was challenged by an infamous economic crisis, Italian soldiers were lining up to invade Albania, and Asian skies were battered with air blasts from the Japanese army. Although there's no equation capable of modeling the course of a world war, the collection of these misfortunate events gave evident insights of following mayhem.

Financial bubbles were introduced on the 17th century in the Netherlands, one of the most advanced countries in finance at the moment. It started off with an unexpected growth in the popularity of tulips as the Ottoman Empire sent a couple of them around the time. As demand for them grew, thus its price, gardeners were getting paid insanely high amounts of money to cultivate these rare plants wherever they could. *Tulipmania*[1] was the nickname given to this crisis, the first speculative asset bubble recorded in history. Somewhere along the way, close to when the Bubonic plague landed in Dutch soil, tulip prices fell off abruptly. Just as if they suddenly realized these flowers were worth nothing.

An asset is considered 'bubbled' when its intrinsic value does not reflect at all the price to which is traded at. Tulip bulb contracts, for example, were commerced for over 600% its actual value. There's no need to have financial expertise to notice these sort of discrepancies on a plant's worth. But is it the same on the modern era?

The outset of the 21st century was sabotaged by a financial crisis mainly on the tech sector. Booming technologies, particularly the internet, played a fundamental role on a stock's market performance. A trending technology who everyone wanted to be part of, including the old-fashioned Wall Street bankers. The U.S exchanges flooded with IPOs, and tens of ordinary companies reintroduced themselves with the *.com* suffix, from this the name of the crisis (Dotcom)[2]. With weak corporate records and almost no real added value, dotcom companies portrayed themselves as the next big thing. Investors bought that idea, and for the beginning of the year 2000, tech markets rose over 500%, the highest returns ever recorded in history.

Years went by, and the housing market reached a shady peak, with a declining ABX index for the second quarter of '07. Lewis Ranieri's private label MBS jeopardized worldwide markets as

thousands of Americans defaulted on their loans. Despite the fact of speculation carrying a decisive role on the evolution of the recession, corrupt officials triggered this bomb even more, intentionally[3].

We know that in Machine Learning, Artificial Neural Networks (ANN) are basically the computational equivalent of the actual human brain. Their approach is similar to the way neurons operate in our nervous system. ANNs have the capacity to learn from datasets, and subsequently generalize unknown input data[4]. Without deep diving into specifics, it is important to state that Machine Learning models by contrast to ANN models, have a single activation function. For example, common Linear Regression can be viewed as a Neural Network with just a single node (neuron), with a linear activation function. This is a crucial concept to understand before following this document, since it's the main reason for following the present network architecture. These networks are pretty handy for these financial problems. This context, specifically, operates with absurdly large extents of data, ANN will outperform common Machine Learning models regarding the amount of needed human intervention, that's why companies rely more on them. The goal of this document is to categorically foresee bubble like behavior in stocks.

The Random Walk hypothesis, firstly introduced by the French broker Jules Regnault in 1863, is a financial theory that stipulates market fluctuations are completely random, meaning they can't be predicted[5]. A wide number of finance professionals back up this theory, while others don't. From my personal experience, it depends. The role of market *hype* (speculation, but positive) has been witnessed before in the markets. The cryptocurrency market for example, is the clearest example of all. If over two thousand people are on the waitlist for buying some new and revolutionary cryptocurrency, it is likely — extremely likely — a bullish season will come up. Or if Elon Musk suddenly tweets he'll be buying some meme coin just for fun, one can sense it's time to buy. However, these approaches for predicting market fluctuations solely rely on news, *hype*.

The following model operates under input data of financial nature. Meaning, it does not contain labels such as how much *hype* it received on Twitter, or how the general public is feeling about it. If we were able to quantify market speculation, for example, worldwide finance will have a massive turn. Because we wouldn't be no longer generalizing out of financial inputs only, but also on how people feel.

The rest of this document is organized as follows. Firstly with the objectives section, where we briefly list and describe the general and specific objectives for the present investigation. Shortly after, section 3 summarizes the theoretical background followed by a detailed explanation of the architecture and topology used for the neural network. Finally we close with the model results, conclusions of it and future work motivation.

2. Objectives

- General Objective

Develop a neural network capable of categorically determining bubble like behavior in a given stock.

- Specific Objectives

    - Identify optimal hyper parameters for maximizing performance.
    - Explore limitations on the prediction of financial bubbles.
    - Evaluate the impact of quantified speculation on the model.

3. Theoretical Background

1. Artificial Neural Networks

a. General Definition

Artificial Neural Networks are a computational system consisting of interconnected artificial neurons (as nodes). Between them they transmit signals. Input data, commonly known as tensors (an array of n-dimensions), travels through the network. Each value a neuron outputs, is multiplied by a specific weight. These weights determine the state of activation of the following neuron, and when passed through a neuron, an activation function is triggered. This function adjusts the input data depending on the function[6].

ANNs learn by themselves. In contrast with traditional Machine Learning models, these networks require much less —almost none— human intervention.

The output of each neuron is the dot product of the input tensor and the weights, plus a bias constant b following. This means the size of them must be equal. The mathematical equation for following up

this aspect goes as follows. Given an input tensor X and a set of weights W, the respective output would be:

$$output = X_0 \cdot W_0 + X_1 \cdot W_1 + \ldots + X_{n-1} \cdot W_{n-1} + b$$

Now that's for each neuron. Each output emits a signal to other connected neuron that can either trigger or not the activation function in order to pass through. The activation functions used for this model were two. The first one is the Rectified Linear Unit activation (or ReLU). This activation function, used for continuous output, yields the positive part of the given argument. Given an input x[7]:

$$ReLU(x) = x^+ = max(0, x)$$

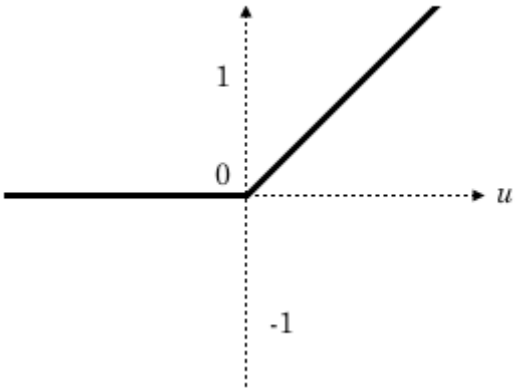

Figure 1. ReLU Activation Function.

As you can see, this activation function discards the negative section of the input data. On the other hand, we used the Softmax activation function. Since we are categorically evaluating bubble like behavior in stocks, we need a classifier to do so. Softmax is specifically for categorical features, meaning qualitative. This activation function basically renders the input tensor normalizing it into a probability distribution. The standard Softmax (*smx*) activation function equation goes as follows. Given an input tensor z with n features[8]:

$$smx(z) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \text{ for } i = [1, n] \text{ and } z = (z_1, z_2, z_3, \ldots z_n) \in \mathbb{R}^n$$

Where smx: $\mathbb{R}^n \rightarrow (0, 1)^n$. These two functions were crucial for the development of the network, the criteria for their choice was purely based on the wide number of times these are used in the industry,

countless resources for its development. The *ReLU* activation function was used for the hidden layers of the network, while the Softmax function just for the last, due to the categorical classification we are inquiring.

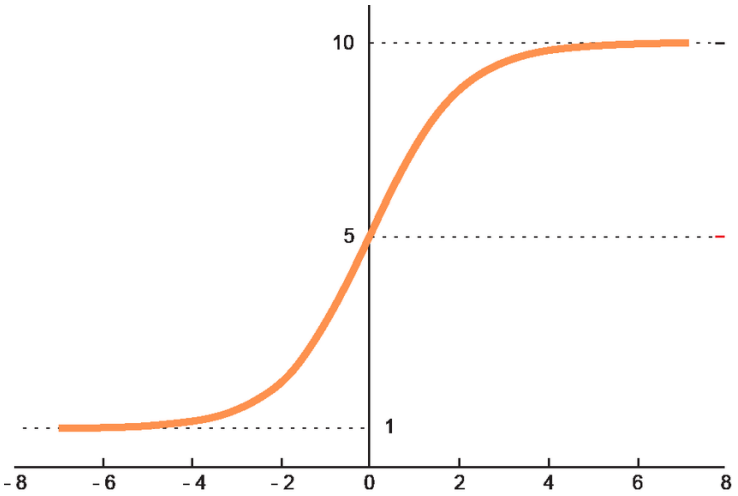

Figure 2. Softmax Activation Function.

b. Training Algorithm

This neural network is trained with SGD (short for stochastic gradient descent algorithm). Just like the gradient descent algorithm, the SGD purpose is to optimize parameter settings. The SGD calculates the gradient with aid of a small set of random samples from the dataset, this optimizes time consumption and it is widely recognized as a powerful algorithm in the industry[9].

$$Q(w) = \frac{1}{n}\sum_{i=1}^{n} Q_i(w)$$

For n observations, where we are willing to optimize parameter w. The i-th $Q(w)$ constitutes the value of the loss function at the given index. This summation is averaged and it results in the desired result.

c. Loss Function

The loss function used was categorical cross entropy. This loss function is pretty handy when it comes to one-hot encoded labels, such as ours. Although sparse categorical cross entropy worked —with almost the same results— we opted for the loss function that optimized memory and time. The loss function is shown below.

$$-\frac{1}{N}\sum_{s\in S}\sum_{s\in C} 1_{s\in c} log p(s \in c)$$

Where S are the samples, C are the classes, and s and c are subsets of their respective sets[10]. With an appropriate comprehension, the model is all set to run. Python was chosen as the development programming language due to its easy scalability, flexibility, and of course because it hosts the world's most recognized machine learning libraries.

2. Training Data

a. Capital Asset Pricing Model (CAPM)

During the ideation process, questions regarding a stock's relation with the overall market arised constantly. After all, general hype is a strong indicator that these securities possibly share something in common, beyond market speculation. The Capital Asset Pricing Model (CAPM) introduces the beta indicator, a measure of volatility that translates to the systematic risk assumed by a particular portfolio or security[11]. Since exploring relationships between the dataset's stocks was key for the project, this measure of market sensitivity aided in the establishment of these relations.

Beta is variant through time, that's why it was considered a relevant benchmark to contemplate the stock's first recorded beta since its IPO and on its ATH date, crucial moments in the life cycle of a bubble. It may be considered a naïve approach to study these indicators just at two given moments in time. However, these are not randomly chosen moments. These moments are the conception of the asset to the market, and its biggest moment ever. Some other indicators clearly follow the random walk hypothesis, since their overall fluctuations have absolutely zero relation to anything.

b. Price (and fluctuation)

This dataset was designed considering endogenous and exogenous trading strategies, it contains labels regarding technical and corporate indicators. A bubbled asset by definition has strong fluctuations, commonly in short periods of time. These short periods usually ranged from the IPO of the stock till its all time high (ATH) recorded price. This two dates were used in the dataset alongside its respective price at the moment. At the bottom line, the algorithm will study these price fluctuations with respect to time, and identify patterns in order to generalize unknown data and predict bubble like behavior.

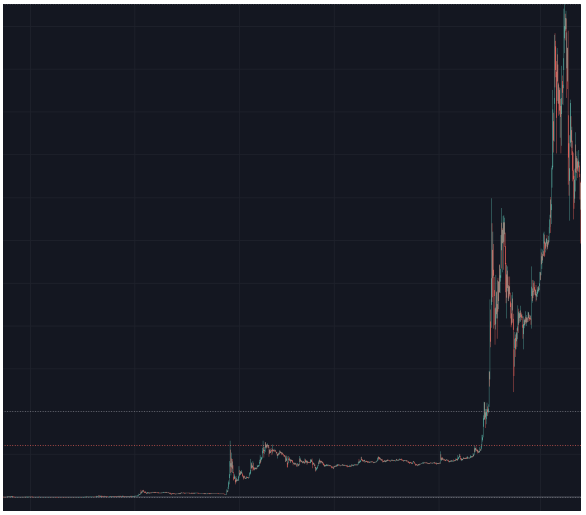

Figure 3. Dogecoin (DOGE) insane overvaluation for May 2021 in the crypto market.

c. PE Ratio (and fluctuation)

The PE Ratio marker is an indicator that stipulates a company's current share price relative to its earnings per share. It basically measures how much the investors are willing to pay for one dollar of earnings. This relative value estimate helps integrate a more fundamental approach on the model, bringing in some corporate financials. This is a key factor for an exogenous approach. We consider its fluctuation because it is extremely relevant for the algorithm to study this valuation metric at the inception of the asset on the market and on its peak. These relationship could be somehow related to other market indicators and in a future be considered for further financial analysis.

d. Criteria for bubble behavior

Bubbles have a very simple definition, one that makes it almost subjective to thousands of investors, an overvalued asset. It doesn't necessarily specify for how much, for how long, or under what qualitative and quantitative conditions the asset is overvalued. The important thing is that it is. Now, under the current market circumstances (partially leaving the COVID-19 crisis, announcing almost 100 days of the Ukrainian-Russian war and a —very— likely following recession provoked by the U.S Federal Reserve) almost everything is overvalued. It is no mystery that the balance sheets of the top-dog stocks have some shady speculation tweaking its market price. It's not the fault of the company, sometimes it's just how the world works. Like GameStop (NYSE: GME), a falling company that suddenly got heaven blessed by some Reddit users, who used their influence on social media to pump GME's stock up to an unbelievable all time high. The same could have happened in a different proportion to other stocks. This is the power of speculation, it can't be tracked nor quantified properly. And the best part is, since there's no equation for it, it might make some investors feel like their shares are priced as they should.

Figure 4. GME's stock price for January 2021. The massive spike grew in less than a month.

Since it's a pretty undeveloped and subjective definition, we will give one based on our prior experience. Fluctuations in price over or equal to 70% in less than 5 months (for *overnight wonders*)

are strong enough to be considered as a bubbled asset. This is due to the fact that no company in the world has enough liquidity, financial capacity, cash flow, and means to back up that fluctuation in such a short period of time, and account it for internal reasons such as a ‘good administration’ or a ‘good financial move’. There’s clearly something behind, and that’s reflected on its share price. This model considers bubbles as assets who suffered the previously mentioned fluctuation in less than 5 months. Examples of these fluctuations can be seen in the cryptocurrency market (almost every month) and during the second quarter of the dotcom bubble year (2001). The S&P 500 Index illustrates the behavior of the overall market and makes it quite simple to identify possible periods of general hype[13].

Essentially, the criteria followed in order to select the previous variables was based on two main considerations. Firstly the prevention of overfitting in the model. It is inefficient to contemplate too many variables on this context, bearing in mind the high correlation between some of these. For example, a dataset with a security’s call option prices and traded volume (for some industries) will add up more noise to the model, partially due to their high variance. The core purpose of this dataset is to feed this baseline model with well balanced data, combining endogenous and exogenous features, using the most relevant metrics used by finance professionals when evaluating securities.

4. Neural Network Architecture & Topology

The architecture followed in order to accomplish the main goals is highly influenced by the Grid Search Algorithm[14]. This algorithm computes all possible combinations of hyperparameters, given an initial random grid, in order to maximize the outcome of the neural network. The random grid is just some sort of estimator for the algorithm to follow by.

Although for this particular problem we desire to identify bubble like behavior in stocks —meaning its a classification problem (yes/no output)— during the investigation process it was discovered that the intrinsic qualities of the origin of a bubbled asset vary. This variation is mainly due to the time component, starting since its IPO to its all time high price (ATH). Some assets because of the eventual economic conditions at the time (such as a highly optimist period, like the dotcom bubble) suffered stronger fluctuations that granted them the bubble title, while some others didn’t. This means, that the initial algorithm treated two bubbled assets differently, even if both were overvalued. For this reason it was decided to optimize the initial dataset, adding this factor. This was accomplished by feeding the algorithm with an additional classification measure, *overnight wonder* (an asset that reached its peak too fast), meaning it no longer had one output layer but two.

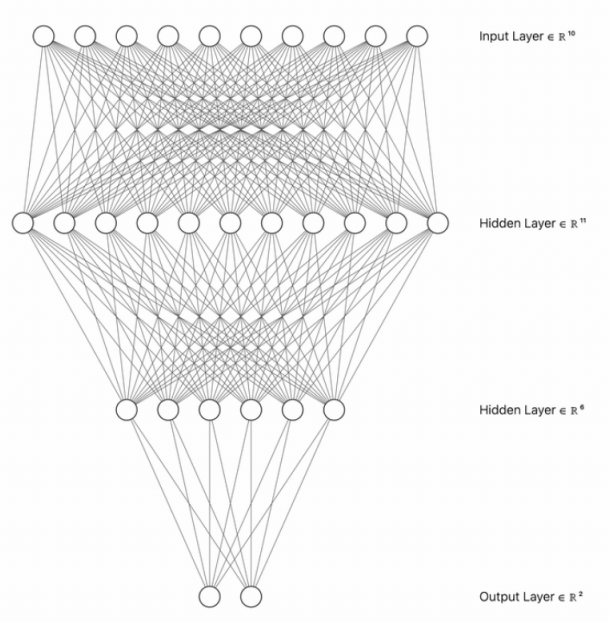

Figure 5. Neural Network representation.

The GridSearchCV (for Cross-Validation) Algorithm when executed suggested the previous configuration. The initial layer, with the same number of nodes as the previously mentioned fields, and two hidden layers with eleven and six nodes respectively. The final layer constitutes the quality of the bubble mentioned previously (An *overnight wonder*, or a common one, one that essentially took its time to be badly pumped). Two neurons are triggered by a softmax activation function that evaluates these conditions for the given stock.

5. Results

After running the model, two metrics where used in order to evaluate its accuracy. We did not rely on just one because not all mean the same. An overfitted model can score over 90% accuracy for predicted labels. These two metrics are the ROC AUC score and the known accuracy score.

ROC AUC is a performance measurement for classification problems at various threshold settings[15]. ROC is a probability curve and AUC represents the degree or measure of separability. It indicates how much the model is capable of distinguishing between classes. In this case, if the stock has bubble behavior or not. For this case, the model returned an optimal result, with a 82% ROC AUC score. It is important to mention that this score relied considerably on the partition for train and test samples. A model trained with few amounts of data will have problems generalizing, but if its trained with too much it will become overfitted. The optimal split for the model was 65:35 (train:test). Other

ratios resulted in higher scores, however it was intuitively clear the risk of overfitting considering the epochs vs loss graph (see Figure 6). This score could have been better if the initial dataset was larger, but as stated before, this limitation was key for the development of the project. Dead companies have almost no financial records available, some of them had, but there was no possible automation on their extraction process, making it hard to populate the dataset.

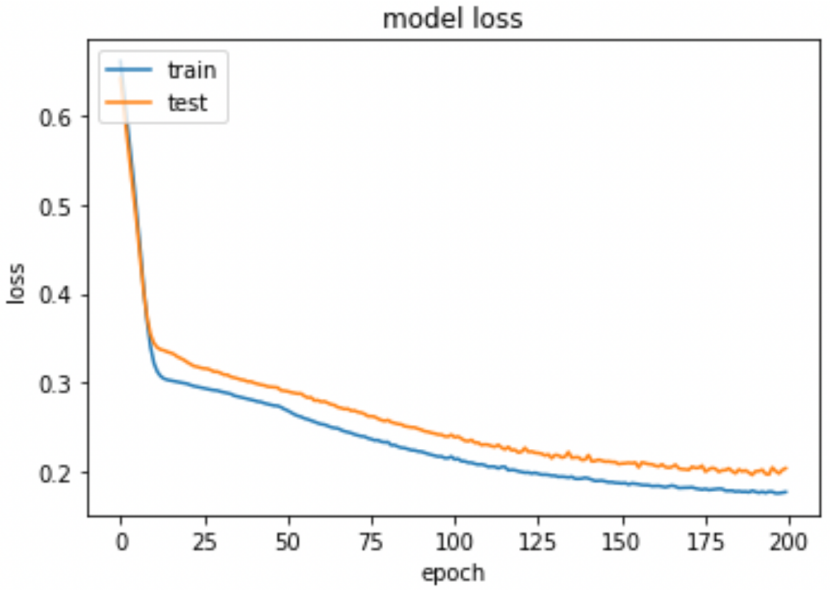

Figure 6. Epochs vs Model Loss

The model isn't perfect, but is still very accurate. Distinguishing the 'bubleness' of an asset barely based on ten input fields is not an easy task for any algorithm, considering the fact that a lot of them could have been born out of pure speculation, a field we wish this model could have had. However technology is not as developed to appropriately quantify these sort of aspects.

If we take a quick look at the following sample of results, it's possible to have an idea of what could have been the influence of a speculation variable in the initial dataset. This sample was taken out of the predicted labels of the so called classifier, and out of this we can recall one of the first limitations described for the project.

Asset Ticker	Predicted	Actual
ETYS	1	1
AAPL	1	0
ARB	0	1
NVDA	0	0
NVR	0	1
AMC	1	1
VW	0	1

SRT	0	1
MNST	1	0

Table 1. Sample of predicted labels

False negatives have a crucial role understanding what could have been the reach of this project if speculation was quantified. The assets who were evaluated as not *bubbled* and indeed were, have a very meaningful financial background. And as stated before, results like these arise from the limitations mentioned previously (the incapacity to quantify factors such as speculation).

6. Conclusions

As a result of the previous investigation, it is possible to conclude that some limitations during the development process were overcome, and still got favorable results. The model is capable of correctly distinguishing between bubbled assets 82% of the time, even with a small dataset. The initial length of this dataset forced the model to be slightly overfitted, meaning that the results could be somehow biased, but since the definition of a bubble is partially subjective, it can be a great aid for investors to follow up their financial strategies.

It was clear the impact speculation could have had in the final results, some assets were classified as not bubbled even if they actually were. This was due to the fact these assets stocks fluctuated by the news, and as stated before, there's no measure for that.

VolksWagen (NYSE: VW) for example, the famous car manufacturer had pretty strong financial records and for the algorithm, its stock popped out of nowhere. This strong price decline back at 2014 was due to the EPA (U.S Environmental Protection Agency) scandal, where VW allegedly equipped thousands of vehicles with "defeat devices", cheating on federal emission tests[16]. The question arises, how bad was it? If a number could somehow describe this scandal, what would it be? And here's one of the main limitations of the project, quantifying speculation is still beyond our computational capability. And even if we could, a Machine Learning model would treat these numbers exactly the same, not considering its background. This means it would also be necessary to normalize these numbers in a special manner, because a scandal for the EPA might not be as severe as one concerning national security.

The second biggest challenge had nothing to do with the actual computational limitations for quantifying speculation. It was the fact that a lot of financial records for hundreds of companies were non existent. Some of them vanished on 9/11, some others vanished in acquisition mergers, and others

simply due to chapter 11 filling where lost over the course of time. If a centralized dataset of historical financial records and indicators was available, a more accurate and efficient model could have been built. Still, results are solid for the few data we had and the few considerations we built the model under.

7. Future Work

Limitations arise out of the lack of available data on the web. Over the course of time, if these resources are more accessible to financial researchers, this worldwide problem could be addressed in a different manner. Various approaches have been successful when predicting some market bubbles, but they don't fully take advantage of the power of data. Every company in the world suffers the risk of events they can barely control, that's why is our duty as engineers to help these people evaluate the markets in a data driven way, and when events like the Dotcom bubble occur, proceed to contingency measures and infringe the less amount of harm.

Under this model, a financial expert can tweak some variables, adjust some parameters, and with his/her knowledge come up with a more developed architecture. Our approach is a baseline model, we invite financial experts pour in their knowledge in markets, economy and sociology into it, in order to come up with better ideas aiming to address this infamous economic issue. Recent market fluctuations are an implicit call praying that work must be done fast, and now. Interest rates are increasing, and we all know what that means.

"It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so."

- Mark Twain

8. References

1. Shiller, Robert J. (2005), Irrational Exuberance (2nd ed.), Princeton: Princeton University Press, ISBN 0-691-12335-7
2. Here's Why The Dot Com Bubble Began And Why It Popped. (2010). Retrieved 9 June 2022, from https://web.archive.org/web/20200406151705/https://www.businessinsider.com/heres-why-the-dot-com-bubble-began-and-why-it-popped-2010-12
3. Zandi, Mark (2010). Financial Shock. FT Press. ISBN 978-0-13-701663-1
4. Chollet, F. (2018). Deep learning with Python ISBN 9781617294433.
5. Smith, T. (2020). Random Walk Theory. Retrieved 9 June 2022, from https://www.investopedia.com/terms/r/randomwalktheory.asp
6. Chollet, F. (2018). Deep learning with Python ISBN 9781617294433.
7. Brownlee, J. (2020). A Gentle Introduction to the Rectified Linear Unit (ReLU). Retrieved 9 June 2022, from https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/#:~:text=The%20rectified%20linear%20activation%20function,otherwise%2C%20it%20will%20output%20zero.
8. Brownlee, J. (2020). Softmax Activation Function with Python. Retrieved 9 June 2022, from https://machinelearningmastery.com/softmax-activation-function-with-python/
9. Srinivasan, A. (2019). Stochastic Gradient Descent—Clearly Explained !!. Retrieved 9 June 2022, from https://towardsdatascience.com/stochastic-gradient-descent-clearly-explained-53d239905d31
10. Gomez, R. (2018). Understanding Categorical Cross-Entropy Loss, Binary Cross-Entropy Loss, Softmax Loss, Logistic Loss, Focal Loss and all those confusing names. Retrieved 9 June 2022, from https://gombru.github.io/2018/05/23/cross_entropy_loss/
11. Kenton, W. (2021). Beta. Retrieved 9 June 2022, from https://www.investopedia.com/terms/b/beta.asp
12. Wendorf, M. (2022). P/E Ratio: What It Is & How It Works (Video) | Seeking Alpha. Retrieved 9 June 2022, from https://seekingalpha.com/article/4494467-pe-ratio?gclid=CjwKCAjwtIaVBhBkEiwAsr7-c336ZSj5f8WfCDyRmtkKGws6g2LdOrWN3_yg43IZ0WRLPuXcHDDUFxoCE9IQAvD_BwE&internal_promotion=true&utm_campaign=14926960698&utm_medium=cpc&utm_source=google&utm_term=127894704186%5Eaud-1457157706279%3Adsa-1427141793820%5E%5E552341146729%5E%5E%5Eg

13. Beers, B. (2022). Why Do Investors Use the S&P 500 as a Benchmark?. Retrieved 9 June 2022, from https://www.investopedia.com/ask/answers/041315/what-are-pros-and-cons-using-sp-500-benchmark.asp
14. Brownlee, J. (2020). How to Grid Search Hyperparameters for Deep Learning Models in Python With Keras. Retrieved 9 June 2022, from https://machinelearningmastery.com/grid-search-hyperparameters-deep-learning-models-python-keras/
15. Classification: ROC Curve and AUC | Machine Learning Crash Course | Google Developers. (2022). Retrieved 9 June 2022, from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=es_419#:~:text=AUC%20represents%20the%20probability%20that,has%20an%20AUC%20of%201.0.
16. Hotten, R. (2015). Volkswagen: The scandal explained. Retrieved 9 June 2022, from https://www.bbc.com/news/business-34324772