Application of Explainable AI in Machine Learning models to identify the main determinants of Bitcoin price

Sofia Morais

Dissertation written under the supervision of Professor Ana Guedes

Dissertation submitted in partial fulfilment of requirements for the MSc in Business Analytics, at the Universidade Católica Portuguesa, December 2022

Abstract

Since 2019, Bitcoin has become one of the most popular assets in the world. However, this decentralised cryptocurrency is typically characterised by high volatility and, in that sense, creates some concerns mainly to regulatory authorities and other decision-makers, such as governments and legislators.

Furthermore, there are multiple approaches and results in the literature regarding the most relevant determinants to predict the Bitcoin price, the complexity of the Machine Learning (ML) model used to predict the Bitcoin price, and the trade-off between interpretability and the model's performance.

As a starting point, the simple model called Generalized Least Squares with Autocorrelation covariance structure (GLSAR) was found to be unrealistic to predict something as complex as the Bitcoin price. Alternatively, two more complex black box models were tested: a Long Short Term Memory neural network (LSTM) and a simple Deep Neural Network (DNN). LSTM achieved the highest R^2 *score* of 81.63% with DNN obtaining a R^2 *score* of 81.27%.

Explainability techniques were applied on DNN and the results indicate that 71% of the twenty-one most significant variables are transaction-based, although future analysis can be done for occasional events. Moreover, the three most important features are the S&P500, the Bitcoin price in the previous day and how difficult it is to mine a Bitcoin block.

Keywords: Bitcoin; Determinants; LSTM; GLSAR; DNN; Complexity; Performance; Interpretability; AI; Decision-making

Title: Application of Explainable AI in Machine Learning models to identify the main determinants of Bitcoin price

Author: Ana Sofia Rosa Morais (152021001)

Resumo

Desde 2019 que o Bitcoin se tornou um dos ativos mais conhecidos no mundo. Esta criptomoeda descentralizada é tipicamente caracterizada pela elevada volatilidade e, nesse sentido, provoca algumas preocupações, sobretudo às entidades reguladoras e a outros decisores, como governos e legisladores.

Além disso, há múltiplas abordagens e resultados na literatura relativamente aos fatores mais relevantes para prever o preço do Bitcoin; à complexidade do modelo de *Machine Learning* (ML) usado; e ao *trade-off* entre o nível de interpretação e o desempenho do modelo.

Como ponto de partida, o modelo simples designado de *Generalized Least Squares with Autocorrelation covariance structure* (GLS) revelou-se irrealista para prever algo complexo como o preço do Bitcoin. Alternativamente, dois modelos *black box* foram testados: uma rede neural *Long Short Term Memory* (LSTM) e uma simples *Deep Neural Network* (DNN). O LSTM atingiu o melhor R^2 *score* de 81,63% e o DNN obteve um R^2 *score* de 81,27%

Técnicas de explicabilidade foram aplicadas no DNN e os resultados indicaram que 71% das vinte e uma variáveis mais importantes são relacionadas com as transações, embora possam ser feitas futuras análises para acontecimentos pontuais. Moreover, the three most important features are the S&P500, the Bitcoin price in the previous day and how difficult it is to mine a Bitcoin block. Além disso, as três variáveis mais significativas são S&P500, o preço do Bitcoin no dia anterior e a dificuldade em extrair Bitcoins e *hash rate*.

Palavras-chave: Bitcoin; Determinantes; LSTM; GLSAR; DNN; Complexidade; Desempenho; Interpretação; IA; Tomada de decisão

Título: Application of Explainable AI in Machine Learning models to identify the main determinants of Bitcoin price

Autor: Ana Sofia Rosa Morais (152021001)

Glossary

Term	Definition
Cboe Volatility Index (VIX)[1]	A real-time market index representing the market's expectations for volatility over the coming 30 days.
Bias (estimator)	How far the average estimated value is from the actual value.
Data leakage[2]	Describe the situation in which the data used to teach a machine-learning algorithm contains unexpected extra information about the subject you're estimating.
Interpretability	The extent to which a cause and effect can be observed within a system.
Explainability	The extent to which the internal mechanics of a machine or deep learning system can be explained in human terms.
Mining[3]	Process in which users who provide their computing power, verify and record payments into a public ledger blockchain. In return for this service, they receive transaction fees and newly minted Bitcoins.
Standard & Poor's 500 Index (S&P500)[4]	A market-capitalization-weighted index of 500 leading publicly traded companies in the U.S.
Variance (of the estimator)	It measures how widely spread the estimates are between two data samples.

[1] Ciaian et al., 2016
[2] Singh, P. (2022, January 15). Data Leakage in Machine Learning: How it can be detected and minimize the risk. Medium. https://towardsdatascience.com/data-leakage-in-machine-learning-how-it-can-be-detected-and-minimize-the-risk-8ef4e3a97562
[3] Investopedia
[4] Investopedia

Contents

1. Introduction

1.1. Bitcoin market evolution

In 2008, Bitcoin was created as a new decentralised currency that disrupted the way transactions were settled. Furthermore, it is the most outstanding example of a blockchain-based cryptocurrency (Li & Wang, 2017).

Recently, Bitcoin has been the subject of intense discussions in newspapers, magazines, and social media and has become a new topic of study in a wide array of fields of study. For instance, among other countries, Germany and Japan have accepted Bitcoin as legal tender, and international companies such as Microsoft and Dell have also approved Bitcoin as a means of settlement for their transactions. Liang et al., 2022 explain that Bitcoin is linked to intense price volatility. Concretely, its price increased from zero value at the time of its inception in 2009 (Böhme et al., 2015) to around $1,100 at the end of 2013. Such sharp market volatility is not usual for traditional currencies, suggesting that there must be other determinants of price formation (Ciaian et al., 2016). As a consequence, the investment in Bitcoins are associated with high-levels of risk.

In addition, this cryptocurrency has the advantage of lower transaction costs than other currencies. However, Rajabi et al., 2022 state that the lack of regulation by any central bank or intermediary led some governments, such as New Zealand, to deny Bitcoin legal tender status. This issue also concerns central banks since the increase in money supply operated by private entities could hamper the implementation of monetary policies by central banks in a efficient way (Tomić et al., 2020).

For all these reasons, different covariates that might be useful to predict the Bitcoin price have been extensively analysed, as described by Bakas et al., 2022; Carbó & Gorjon, 2022; Ciaian et al., 2016; Coulter, 2022; Foley et al., 2022; Jo et al., 2020; Li & Wang, 2017; Liang et al., 2022; Liu & Tsyvinski, 2018; Lyócsa et al., 2020; Mai et al., 2018; Mudassir et al., 2020; Polasik et al., 2015; Rajabi et al., 2022; Ranjan et al., 2022; Sovbetov, 2018. Nevertheless, there are still multiple approaches and results in the literature. On the one hand, the transaction-based variables have a greater preponderance in the analysis presented by Li & Wang, 2017; Liu & Tsyvinski, 2018; Sovbetov, 2018. On the other hand, sentiment variables were emphasised in Bakas et al., 2022; Coulter, 2022; Foley et al., 2022; Jo et al., 2020; Liang et al., 2022; Lyócsa et al., 2020; Mai et al., 2018; Polasik et al., 2015.

1.2. Explainable AI

Only some studies that used complex ML models to predict the Bitcoin price apply an Explainable Artificial Intelligence (XAI, also known as interpretability and explainability techniques) in their analysis. (Mudassir et al., 2020; Rajabi et al., 2022; Ranjan et al., 2022). XAI refers to methods and techniques and allows for the results of the solution to be understood by human experts, thus solving a limitation of some ML models commonly known as "black box". If an AI technology is not explainable, ML designers cannot justify why the AI solution reached a specific decision. Legally, XAI is the solution to comply with the social right to explanation or human review of AI-based decisions, commonly invoked in data protection or creditworthiness assessment areas.

In the last place, it is worthwhile to note that the lack of transparency of these models leads relevant stakeholders such as consumers, auditors, supervisors or investors to hardly trust them ("Artificial Intelligence and Machine Learning in Financial Services Market Developments and Financial Stability Implications," 2017).

1.3. The purpose of this analysis

This project aims to answer three fundamental questions.

Firstly, this research will identify which explanatory variables are relevant to predict the Bitcoin price. On the one hand, variables intrinsically related to Bitcoin transactions will be discussed, such as transaction fees and the number of transactions. On the other hand, external covariates of a group of economic, financial and sentiment analyses will be particularly emphasised. For example, news about regulation and the demand for gold were included within the scope of the analysis. As a first approach, a group of transaction-related and external variables considered in the literature will be included in the dataset of ML models. After obtaining the models outputs, a new analysis will be done to identify its importance for Bitcoin price prediction accurately.

Secondly, this study provides three models with different degrees of complexity and that make use of time series data. A simple model, GLSAR; LSTM, a complex model able to process sequences of data points; and a DNN, a complex model for which there are implementations of XAI techniques for local and global interpretations available.

Thirdly, similar to the lack of transparency of LSTM (Carbó & Gorjon, 2022), the simpler DNN required applying the two fundamental XAI techniques on this ML model, namely, interpretability and explainability. Moreover, this analysis will discuss the trade-off between interpretability and complexity (usually associated with higher accuracy) that usually exists when using simple ML models or, instead, complex ML models.

Lastly, in order to contribute to the to the scientific community, this thesis introduces a recent topic that it is relevant in the domain of financial analysis and data science. Moreover, the code will be shared publicly so new findings can be found in the future.

In a nutshell, these three research questions constitute a deeper analysis and a new approach that enables analysts and decision-makers, such as regulators and investors, to identify the most relevant determinants of Bitcoin price beforehand.

1.4. Thesis outline

This Master's thesis is arranged into five chapters:

1. This first chapter introduces the research questions, defines the main objectives, and explains the project's significance.
2. The second chapter reviews the information available in the literature about this research and further extends this thesis' knowledge about the wide range of concepts and methods related to Bitcoin price.
3. The third chapter presents the methodology used in the research, namely, the dataset, the relevant pre-processing, and the models.
4. The fourth chapter shows the results obtained.
5. In the fifth chapter, the results are interpreted and explained. In addition, it is discussed the significance and implications of the results.
6. The last chapter provides the main conclusion of the study, indicates some limitations, and suggests opportunities for future research.

2. Literature review

The Bitcoin price determinants

In order to define the most relevant determinant of Bitcoin price and the respective methodology, this chapter provides a background knowledge of the current literature. The related literature is organised according to the main analysis of this project: (1) Identify the main determinants of Bitcoin price; (2) Find the most accurate ML models to determine the Bitcoin price; and (3) Apply interpretability and explainability techniques in ML models.

2.1. The main determinants of Bitcoin price

This project covers the related literature about the determinants of Bitcoin price with the goal of identifying the potentially relevant variables to use in the ML models. The aim was not only to select a rich dataset that contained transaction-based drivers but also other variables that could highlight the effect of external factors.

2.1.1. Transaction-based variables

First of all, transaction-based variables were emphasised by Mudassir et al., 2020; Rajabi et al., 2022; Ranjan et al., 2022; Sovbetov, 2018. Some of these articles named transaction-based variables as economic variables because they are related to the Bitcoin market. On the contrary, this thesis sets these two apart. While transaction-based variables, as the name suggests, represent all the variables related to transactions, the economic variables considers the macroeconomic indicators. This difference will be made clear in the following explanations.

While Sovbetov, 2018's conclusions hinged mostly on the determinants, the ML models' characteristics were the main discussion in Mudassir et al., 2020; Rajabi et al., 2022; Ranjan et al., 2022. In both cases, the articles analysed the explanatory variables used in order to choose the most common to use in the methodology. For example, Rajabi et al., 2022 presented a wide range of transaction variables such as daily values of the market price, transaction fees, hash rate, the number of transactions, and others. It is a complete dataset in a way that compiles many transaction-based variables used in other studies. The associated dataset would be the baseline for this project in terms of transaction variables.

Firstly, Mudassir et al., 2020 split their analysis between short-run and long-run. The reserach shows that crypto market factors such as total market prices, trading volume, and volatility are statistically significant in both the long-run and short run. The trading volume appears to have a significant long-run impact on Bitcoin at a 1% significance level.

Secondly, Ciaran et al., 2016 explained that both supply and demand are determinants of the Bitcoin price, where demand-side predictors such as the size of the Bitcoin economy have a more significant impact. When Bitcoin became more established, with a larger number of units in circulation, the estimated magnitude of the supply-side and demand-side variables on Bitcoin price tended to be larger. On the other hand, supply factors represented as mining costs, price-to-“dividend” ratio, or realised volatility were highlighted as the main drivers of the explained variable (Liu & Tsyvinski, 2018).

Additionally, Bakas et al., 2022 highlighted the total Bitcoins in circulation as one of the most important predictors to the price.

Lastly, Li & Wang, 2017 stated that both technology factors (Liu & Tsyvinski, 2018) and economic factors are relevant to determine the cryptocurrency exchange. In addition, the study shows some interesting findings regarding the mining variables. Concretely, although the market price relies on mining cost, the long-term impact of mining difficulty decreases over time as mining technology becomes more efficient. Furthermore, the predictors varied over the maturity level. Namely, early market exchange rates are driven predominantly by speculative investment and deviate from economic fundamentals. However, when the market reaches maturity, the price dynamic follows more closely with changes in economic factors, while market speculation does not cause a significant impact.

2.1.2. External predictors

First, Bakas et al., 2022; Coulter, 2022; Foley et al., 2022; Jo et al., 2020; Liang et al., 2022; Lyócsa et al., 2020; Mai et al., 2018; Polasik et al., 2015 presented another approach that hinged mainly on the external factors that could predict the Bitcoin price.

Firstly, Coulter, 2022; Lyócsa et al., 2020 analysed the volatility of Bitcoin price, whose findings highlighted that Bitcoin is utterly and negatively affected by specific risk factors such as regulation and hacking attacks. Thus, it was added these two variables to the original dataset named *news_regulation* and *theft*.

Moreover, Lyócsa et al., 2020; Bakas et al., 2022; Ciaian et al., 2016 found that, unlike traditional assets, the influence of macroeconomic news on Bitcoin price is irrelevant, and it is only weakly connected to the overall economy via the forward-looking component consumer confidence index (Bakas et al., 2022). In the case of Ciaian et al, the research found only a significant impact on global macro-financial development, captured by the Dow Jones Index, exchange rate and oil price in the short run.

Additionally, an important outcome that the article underlined is that while supporting (positive) investor sentiment seems to have a positive effect and leads to an increase in volatility, non-supporting (or negative) and neutral investor sentiment does not have a significant impact on Bitcoin volatility (Lyócsa et al., 2020). On the other hand, Polasik et al., 2015 referred unfavourable mentions matter, since they can have negative consequences on the explained variable. Liu & Tsyvinski generalised the investor's attention as a significant predictor.

Ciaian et al., 2016 split the sentiment analysis into short-run and long-run. The article showed that the increase in new information has a positive influence. Moreover, it suggested that this can be a consequence of the growing trust among users. However, the online information about Bitcoin represented as the number of views on Wikipedia has no impact on Bitcoin price in the long run.

Additionally, the statistically significant impact of Wikipedia views and new posts on Bitcoin price could be an indicator of the speculative short-run behaviour of investors, or it may capture the expansion of the demand side of the Bitcoin economy. Therefore, speculative trading is not necessarily an undesirable activity *per se* since it may absorb the excess risk from the participation of risk-averse investors and provide liquidity on the Bitcoin market. On the other hand, speculative investment can be disadvantageous in a way that it may increase price volatility and generate price bubbles. Consequently, the article argued that the success of Bitcoin depends on the reduction of potential negative consequences of such speculations and the expansion of Bitcoin trading and commerce.

Furthermore, Coulter, 2022 showed that the cryptoeconomy and markets discourse had a negative effect on Bitcoin prices over a two-year period.

In addition, Coulter, 2022; Foley et al., 2022 argued that a potentially important factor that influence on cryptomarkets is the geographical location. On the one hand, media reports, Coulter, 2022 argue that media reports may be driven and amplified dependent upon the geographical source of the news. On the other hand, Foley et al., 2022 studied the important role that culture has in investors' decisions in such high-risk assets as Bitcoin.

The approach of Bakas et al., 2022 considered 22 drivers in their study. The results showed that the most important factors for Bitcoin volatility are Bitcoin Google Trends, and the S&P500 index. Particularly, Liang et al., 2022; Polasik et al., 2015 underlined the role of Google Trends as a popularity predictor of Bitcoin price. To illustrate, Liang et al., 2022 explained that returns tend to be higher whenever newspaper articles mention Bitcoin more frequently and whenever the number of people searching for it on Google rises.

In terms of variables that might represent substitute assets to Bitcoin, the literature presented information about S&P500, VIX and gold index. Regarding the financial indexes, Liu & Tsyvinski, 2018; Liang et al., 2022 stated cryptocurrency returns have low exposures to traditional assets such as stocks, currencies, and commodities. In particular, Liang et al., 2022 explained that VIX had almost no predictive power for Bitcoin volatility. Instead, VIX contains more useful information to predict global equity markets. The research suggested that Bitcoin and the equity market are weakly connected. On the contrary, Bakas et al., 2022; Mudassir et al., 2020 indicated S&P500 is an important financial index. Based on this evidence, this thesis included the S&P500 instead of VIX as a financial driver with an influence on the price of Bitcoin. In addition, CBOE Gold ETF Volatility Index (GVZ) showed a significantly negative effect on Bitcoin volatility, according to Liang et al., 2022. The main reason is that Bitcoin and gold can be used as substitutes for investment, which means that Bitcoin's worth is opposite to gold's. On the contrary, Kang et al., 2019; Kyriazis, 2020 state that Bitcoin has a similar property of safe assets against equity and currency price fluctuations represented through, for example, *lg_SP500*. Nevertheless, both assets have high investment value.

Another predictor that Liang et al., 2022 indicated was the growing impact of the shadow economy in driving Bitcoin use, and it advised that authorities should intervene in order to restrain illegal activities facilitated by cryptocurrency.

Finally, there are occasional events in time that also conditioned the Bitcoin price. Statista is a platform that contains a wide range of statistics, in which some of which are covered by a short analysis. Although a deeper analysis will be presented in the results, Statista's indicated that the particular price hike was connected to the launch of a Bitcoin Exchange Trade Fund in the United States. Additionally, other hikes in 2021 derived from events involving Tesla and Coinbase. Concretely, Tesla's announcement in March 2021 that it had acquired $1,5 billion worth of Bitcoin, as well as the Initial Public Offer of the United States' biggest crypto exchange, fuelled mass interest in the acquisition of Bitcoin. However, close to May 2021, the Bitcoin price started declining. The same analysis of Statista explains that this trend started in April 2021 and resulted from the speculation that Bitcoin would be subject to further government regulation.

2.2. Machine Learning models

There are multiple approaches and results in terms of ML approaches applied to predict the bitcoin price.

Firstly, while some studies that aim to identify the predictors of Bitcoin price utilise conventional statistical and econometric models, experimentation in the domain of forecasting models to be used as decision-support tools in investment techniques is not so frequent (Maleki et al., 2020). The regular updating in trading policy intensifies cryptocurrencies prices fluctuations, thus weakening the predictive power of traditional models. (H. Guo et al., 2021)

Logistic Regression and Linear Discriminant Analysis models performed better for daily data with high-dimensional features. However, the XGBoost model outperformed all other ML algorithms in the case of high-frequency data (5-min intervals). (Ranjan et al., 2022)

Another method called the Prediction Hardship Factor (PHF), which was also introduced in this study to show under what conditions and with what fluctuation the calculated Mean Absolute Percentage Error (MAPE) was obtained. The evaluations proved that PHF has been able to get the MAPE of about 6.26% under harsh situations, which showed an absolutely acceptable amount (Rajabi et al., 2022).

LSTM showed the best overall performance. Nonetheless, all the developed models are satisfactory and have good performance, with the classification models scoring up to 65% accuracy for the next-day forecast. Additionally, it is possible to forecast the actual Bitcoin price with very low error rates, while it is much harder to forecast its rise and fall (Ranjan et al., 2022). Lastly, in Carbó & Gorjon, 2022, the LSTM model performed considerably well in all three periods considered. This is a particularly positive outcome considering that the method did not use lagged values of the price of Bitcoin as additional features. Also, the Root Mean Squared Error (RMSE) obtained was between 5% and 20% of the price.

In terms of performance, LSTM achieves better results in predicting the value of Bitcoin price, which is a regression problem (Ji et al., 2019; Pabuccu et al., 2020; Rather, 2022). On the other hand, DNN-based outperforms for a classification problem that intends to predict the increase or decrease in Bitcoin price. Also, this model performs best for low values of the input size. (Ji et al., 2019). Compared to DNN, LSTM achieves a greater performance. (Guo et al., 2021)

In the last place, other approaches explained in Mandeep et al., 2022 used Random Forest and XGBoost as ML models Models like the random forest and XGBoost can they reached a high accuracy of 98%. However, since they are black box models, they results are untrustworthy and explainability techniques solved this problem.

2.3. Explainable AI – Interpretability and Explainability in Machine Learning models

First, models like the random forest and XGBoost can look at the history of stock prices and make accurate predictions about their future prices. Concretely, these complex models can reach an accuracy as high as 98%, but their "black box" nature leads to a lack of transparency, thus hindering trustworthy results, fully replicable, and the ability to establish causal inference (Hoepner et al., 2021). Moreover, despite the significant performance of the DNN algorithm, it has some limitations in terms of interpretation and can be classified as a "black-box" model. Considering that the financial industry is highly regulated, it requires understanding how the decisions were made and quantifying the potential risk. (Chen et al., 2022) Therefore, a model with high accuracy is not enough for the analysis, and XAI can be the solution to this problem. (Mandeep et al., 2022).

Secondly, the XAI method is used to understand which features most influence the ML model outcome (Carbó & Gorjon, 2022; Molnar et al., 2020). For instance, one approach used to achieve explainable results is SHapley Additive exPlanations (SHAP) (Babaei et al., 2022; Carbó & Gorjon, 2022; Fior et al., 2022; Giudici & Raffinetti, 2021; Mandeep et al., 2022; Molnar et al., 2020). This approach was proposed by Lundberg and Lee (2016) as a united approach to explaining the output of any machine learning model. In addition, Local

Interpretable Model-agnostic Explanations (LIME) is another local interpretability method used (Mandeep et al., 2022). SHAP is the state-of-the-art due to the benefit of providing global interpretations when the features are correlated to interpret neural networks (Carbó & Gorjon, 2022).

One approach of XAI applied in the literature is a global XAI method which is based on Lorenz decompositions, by extending previous contributions based on variance decompositions. This allows the resulting Shapley-Lorenz decomposition to be more generally applicable and provides a unifying variable importance criterion that combines predictive accuracy with explainability, using a normalised and easy-to-interpret prediction of Bitcoin prices (Giudici & Raffinetti, 2021).

Classical feature importance tools have some limitations when used for explainability purposes, when compared to approaches like SHAP. For instance, Permutation Feature Importance can be computationally expensive, considering that an iteration through each predictor and the consequent predictions are necessary. Secondly, this tool can produce poor performance results in the presence of multicollinearity. Additionally, the scores obtained are relative to the predictive power of features in a model and not absolute values. Therefore, the predictive performance of a variable could be outstanding or the opposite, depending on the scores of the other variables. Lastly, this feature's importance does not provide statistical inference. Concretely, it does not inform about the magnitude of the explanatory variable's impact nor the nature of the relationship, which could be, for example, linear or quadratic (Molnar et al., 2020)

Concluding, when XAI methods and techniques are used alongside ML models, it is possible to get a predicting model which is supported by its empirical results and intuitive explanations (Mandeep et al., 2022). Specifically, SHAP provides is preferable compared to other XAI in terms of providing global interpretations of the features used to predict the Bitcoin price when the features are correlated.

3. Methodology

In order to contribute to the research community, the dataset is publicly available on GitHub through the link: https://github.com/s-asrmorais/MSc_thesis_code_DataScience.

3.1. Data Collection

In this study, two crucial categories of data were selected to determine the Bitcoin price. Namely, the covariates that are intrinsically related to the Bitcoin transactions mainly used in Rajabi et al., 2022 were combined with the external factors that also predict the Bitcoin price (Lyócsa et al., 2020). Thus, this dataset resulted from merging the original dataset associated with the study of Rajabi et al., 2022 with other datasets.

The original dataset is a time series data that compiled data from 2017 to 2021 and contains information about the evolution of market price, market capitalisation, trade volume, transaction fees, average confirmation time, difficulty, the highest price, the lowest price, the hash rate, the block size, miners' revenue, number of total transactions, the open price, number of payments per block, total Bitcoin in circulation, percentage cost per transaction, fees per transaction, the number of unique addresses, the number of transactions per block and the output volume. It also had the Bitcoin keyword popularity index extracted from Google Trends.

In order to incorporate financial and economic factors in the model, the gold and oil prices were collected from Statista. Moreover, it was added S&P500 from an open database of the European Central Bank.

On the other hand, five variables that aimed to integrate sentiment analysis variables in the model were extracted. Firstly, three variables were obtained from Google Trends, each representing a relative index of the importance of Bitcoin plus a keyword search. These keywords are "legal", "cryptocurrency", and "scam", and each of them represents a positive, neutral and negative sentiment, respectively. Four different words were tested to select the most relevant word. for each sentiment group.

Lastly, the explanatory variables found in the literature as insignificant, for example, macroeconomic news, were not included in the dataset.

3.2. Data Preparation

Feature Engineering

A dummy variable was created and assumed the value 1 when an article about Bitcoin regulation in Financial Times (FT) was published and 0 when the opposite is true.

Additionally, another dummy variable was included. It took the value 1 when an article about Bitcoin theft in FT was published that day and 0 when it was not.

Data transformation

Figure 1 indicates that *lg_price* shows a skewed distribution, indicating that a log transformation would be applicable for data to follow a normal distribution. Thus, the transformed distribution is shown in **Figure S1**. Hereinafter, *lg_price* will be referred to frequently as Bitcoin price for simplification. In addition, other monetary variables were also transformed into logarithms so they could be compared to *lg_price.*

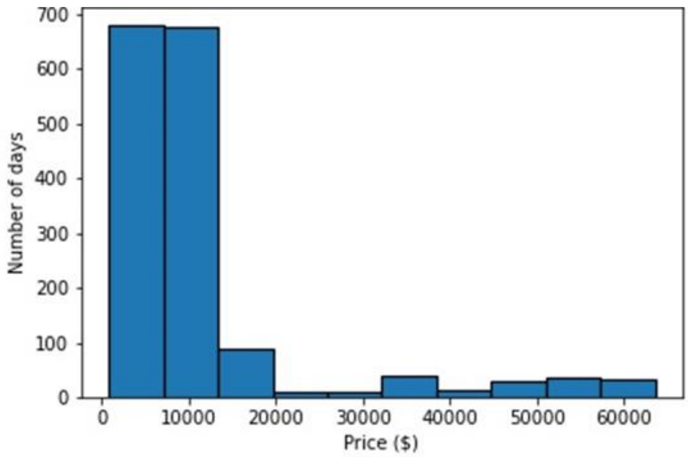


Figure 1 –Market price distribution. Histogram distribution reflecting left skew of market price distribution

Moreover, **Figure S2** presents the distribution of the other transaction-based variables. Plots number 1, 2, 3, 4, 9, 10 and 12 are skewed. Therefore, a log transformation was applied to *trade-volume, avg-confirmation-time, difficulty, hash-rate, Total-circulating-BTC, n-unique-adresses* and *output-volume*. In the other situations, the log transformation did not make the desired improvements. Consequently, this transformation was not applied to those cases. Finally, the log transformation on external variables did not normalise the distribution. Therefore, it was not applied.

Data cleaning was done in *GT_legal* so the string “<1” that indicated that the values were lower than one were converted to float equal to 0.5.

There was no missing data in the dataset.

The list and the description of the variables used in the final dataset are explicit in **Table S1** in the appendix.

Data flow

Figure 1 represents the data flow process from collecting data to obtaining the outcome data.

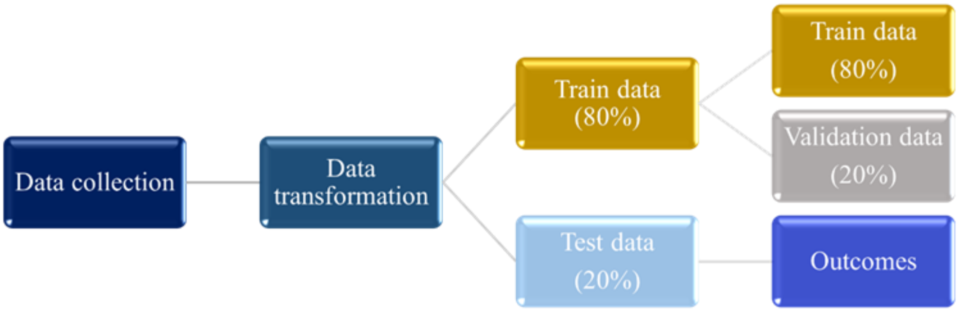


Figure 2 –Data pipeline throughout this work. Data Collection and transformation were initially performed, followed by train/test split and train/validation split avoided data leakage.

Based on the timeline, the dataset was split into the first 80% of observations to train and the remaining 20% last observations to test. Then, validation was performed.

Similar to train/test split proportions, the train data (80%) was split again into subsets: the train data (80%) and validation data (20%). This technique helps prevent data leakage. In this case, part of the information used for prediction was previously used in the model training process, leading the predictive scores to be overestimated. Therefore, this method aims to avoid biased algorithms and allow the model to generalise. Both techniques will be applied to all models.

In conclusion, a train/data split, and a train/validation split are complementary and aim to avoid overfitting and obtain predictions that are mostly approximated to the actual values.

3.3. Modelling

3.3.1. Machine Learning models

Generalized Least Square Autoregressive (GLSAR)

First, GLSAR model was built as the simplest approach.

In equation **(1)**, log_$price_t$ represents the price of Bitcoin and it is the outcome variable of the model for t between 01/01/2017 and 04/06/2021. Moreover, lg _$open_t$ and GT_legal$_t$ are two examples of the covariates that explain the value of log_$price_t$. Regarding the coefficients, β_0 is the constant and would equal log_$price_t$ if all the variables were zero; β_i for $i \geq 1$ represents the coefficients of the covariates. Finally, ε_t is the error terms and includes all the other variables that influence log_$price_t$ and that are not expressed in the model.

$$\text{log\_}price_t = \beta_0 + \beta_1 \text{lg }\_open_t + \beta_2 \text{GT\_legal}_t \ldots + \beta_i x_{it} + \varepsilon_t \quad , \; i = 0, \ldots, 29 \quad \textbf{(1)}$$

The six Gauss-Markov assumptions ensure that the estimators calculated using Ordinary Least Squares (OLS) are the Best Linear Unbiased Estimator (BLUE). When one of these assumptions is violated, it is impossible to use the sample parameters calculated using OLS to extrapolate to the population parameters. GLSAR aimed to solve the problems of heteroscedasticity and autocorrelation.

Furthermore, time series data do not follow the Gauss-Markov assumption that requires that the independent variable and the outcome variable are both random variables. Instead, it is no longer possible to assume that the covariates are random variables for times series.

In order to avoid repeated information, the ratios that included variables already used in the model, such as *n-transactions-per-block*, *lg_cost-per-transaction-percent*, *lg_fees-usd-per-transaction*, and *n-payments-per-block* were deleted. Based on the explanation in section 4.1, *lg_High*, *lg_Low* and *lg_Open* were dropped.

Moreover, Lasso will be implemented as a regularisation technique. However, if this method presents no improvement, it will be retained.

Lasso allows some estimated coefficients to be precisely zero when the lambda tuning parameter is sufficiently large. Also, it performs variable selection by keeping only the significant covariates in the model. In addition, the output model is generally easier to intercept. However, if this method presents no improvement, it will be retained.

Lastly, sometimes, regression models are too simplistic and might be far from representing more complex realities. In that case, more complex models are required.

Long Short Term Memory (LSTM)

LSTM is a variation of feedforward neural networks capable of learning the time dimension of the data (Carbó & Gorjon, 2022) by filtering the critical information and storing it in their long-term memory while forgetting the irrelevant information. Thus, LSTM is a more complex version of a DNN. **Figure 3** schematically represents how this model works.

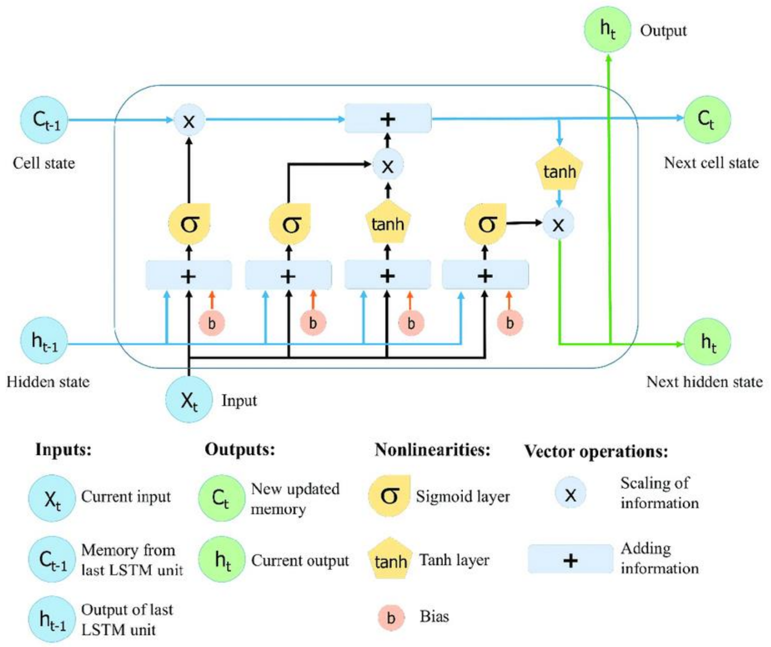


Figure 3 – LSTM model structure. LSTM has the advantage of not imposing ex-ante restrictions on the relationship between the various features and the price of Bitcoin (Source: Lee, Giha. Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting. July 2019. Research Gate. https://www.researchgate.net/publication/334268507_Application_of_Long_Short_Term_Memory_LSTM_Neural_Network_for_Flood_Forecasting.)

The LSTM layer is a variant of the Simple Recurrent Neural Network (SimpleRNN) layer. However, RNN is associated with the vanishing gradient problem, similar to non-recurrent networks (feedforward networks) with many deep layers. Thus, as one keeps adding layers to a network, the network eventually becomes untrainable. Consequently, LSTM has the advantage of solving this problem.

The key to LSTMs is the cell state, like a horizontal line on the top of the diagram representing this process. The cell state is like a conveyor belt. It moves along the entire chain with only small linear interactions. It is effortless for information to flow along it unchanged.

In the forget gate layer, the model decides what information should be removed or added to the cell state, carefully deliberated by gates, which allow information to continue or not optionally. They are composed of an activation function, in this case, a ReLu neural net layer, and a multiplication operation. The sigmoid function is commonly used because it hinges between 0 and 1, describing how much each component should be let through. Specifically, it processes and attributes a number in the range between 0 and 1 for the numbers in the cell state, where 0 allows the information to continue and 1 does not. Instead, this project used a ReLU activation function because it proceeds simpler mathematical operations requiring less power than tanh and sigmoid. Additionally, after choosing LSTM layers, a Dropout was applied to increase efficiency and eliminate useless nodes so that after each LSTM layer, it did not take the worst-performing nodes to the next step. Therefore, 20% of the total number of nodes was eliminated.

The model deliberates what new information will be stored in the cell state. This process has two parts. In this thesis' model, a ReLU layer, called the "input gate layer", decides which values will be updated. In the next step, the ReLU layer creates a vector of new potential values that could be added to the state.

In the last place, in the "output gate", the output will be decided. First, the ReLU layer aims to determine what parts of the cell state will output. Then, in the last step, a single value with the Dense layer is returned, so this estimated value will be the logarithm of the Bitcoin price predicted by our model.

This model is a flexible approach which does not impose ex-ante restrictions on the relationship between the various features and the price of Bitcoin. Therefore, the model can accommodate multiple features in non-linear and non-stationary time series (Carbó & Gorjon, 2022).

Deep Neural Network (DNN)

A DNN consists of an input layer, several hidden layers, and an output layer, as shown in **Figure 4**. DNNs are advanced and enhanced Artificial Neural Networks (ANNs) whose function is to process and predict information. An ANN is a data processing unit represented as a human brain.

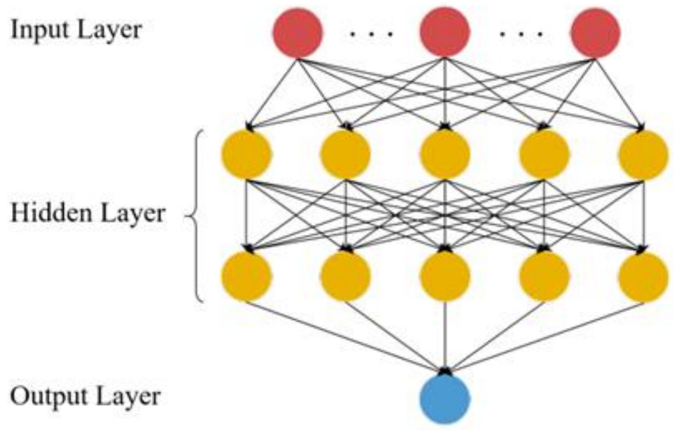


Figure 4 – DNN model structure. The values of all features on the previous day and two days ago were added so the model had the time information (Source: Suhwan, Ji, Kim Jongmin, and Im Hyeonseung. A Comparative Study of Bitcoin Price Prediction Using Deep Learning. September 25, 2019. Mathematics. https://www.mdpi.com/2227-7390/7/10/898.)

Unlike LSTM models or Recurrent Neural Network (RNN), DNN cannot learn the time component between data points. Therefore, the values of 26 different features on the previous day and two days ago were added, resulting in a total of 78 features. However, for the first two days did not have all the previous two days values, then the value on that day were used.

Considering that this model included 78 features to predict the Bitcoin price, the input layer consists of 78 nodes, in which each input node corresponds to each feature. Each column has a length of 1615 that results from the 1615 days present in the dataset.

DNN is fully-connected because the nodes are connected. Each node in the hidden layer takes input vectors from the previous layer, where n is the number of nodes present in the previous layer and m is the vector size. Its output h is then defined as:

$$h = \sigma_h(w_1 \odot x_1 + \cdots + w_n \odot x_n + b)$$

Where the weight vectors and the bias *b* are parameters to be learned by the model using training data, $\odot$ is an element-wise multiplication (Hadamard product) and is a non-linear function. For instance, the Exponential Linear Unit (ELU) and the Rectified Linear Unit (ReLU) are examples of activation functions considered when hyperparameter optimisation.

ReLU is defined as g(x) = max(0, x). ELU is similar to RELU, except it does not produce negative values.

In the last place, in the output layer, the input vectors are flattened into a single vector and then translated into a single output value using a dot product between the flattened vector and a weight vector. The final value, y is the predicted Bitcoin price for a regression problem.

Finally, a hyperparameter optimisation was executed to find the best parameters for designing the DNN model

3.3.2. Explainable AI – SHAP

In the deliberation among ML model, it is necessary to consider the performance versus interpretability trade-off represented in **Figure 5.**

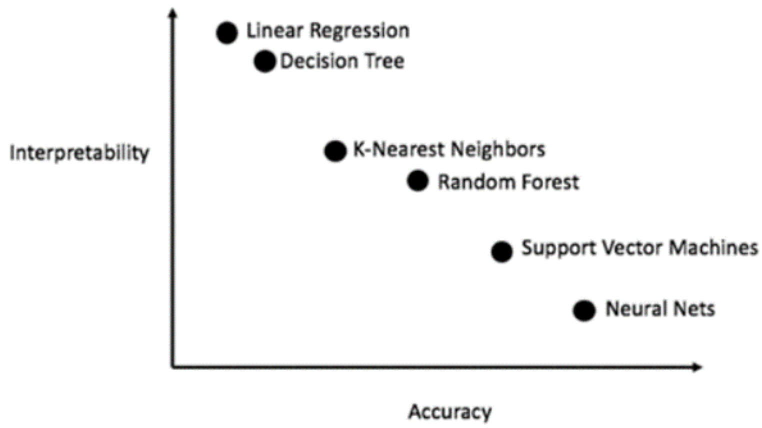


Figure 5 – The interpretability and accuracy trade-off between "White-box" models and "black-box" models. "Black-box" models have better performance, but can be hard to interpret (Source: Sajee, Ann. Model Complexity, Accuracy and Interpretability. May 7, 2020. Towards Data Science. https://towardsdatascience.com/model-complexity-accuracy-and-interpretability-59888e69ab3d.)

Usually higher accuracy is associated with "black-box" models such as neural networks, gradient boosting models or complicated ensembles. The workings of these models are very hard to understand compared to simpler models, and they do not provide an estimate of the importance of each feature in the model predictions, nor even an explanation of the interaction among different features.

"White-box" models, such as the Linear regressions presented in section 3.3.1. (or decision trees), are simpler models but provide less predictive power and sometimes are too simple to represent complex scenarios derived from a complex dataset. However, compared to "black-box" models, they are easier to explain and interpret.

SHAP is a particular benefit in several ways. Firstly, this method provides global interpretability, which means that the collective SHAP values can specify the contribution of each predictor to the outcome variable. For instance, this representation can be plotted in a variable importance graph.

SHAP is also advantageous because it provides local interpretability, which is the most critical attribute to give credibility to a model and to dramatically increase its transparency. Although traditional variable importance algorithms show the results across all population, they do not provide it for individual situations. In addition, this feature is essential to make comparisons among predictors.

Lastly, SHAP is vital in terms of visualisation. Therefore, the respective plots obtained from SHAP over DNN will be shown in section 4.

Concluding, SHAP is a valuable technique to make "black-box" models interpretable and, in turn, informative to decision-makers.

3.3.3. Error metrics

First, an estimator's MSE or Mean Squared Deviation (MSD) measures, on average, the squared difference between the predictions and the actual values. The MSE represents the second moment of the error and thus represents not only the variance of the estimator and its bias. Therefore, this metric assesses the quality of a predictor. The MSE equals to the variance of the estimator when the estimator is unbiased. Finally, MSE is measured through this formula:

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

Secondly, RMSE also evaluates the forecasting model by comparing the predicted values to the actual or observed values. Moreover, RMSE is the square root of MSE. Another difference between these two measures is that while MSE is measured in units of the square of the explained variable, RMSE is measured in the same units as the explained variable.

Consequently, MSE penalises larger errors more severely. Lastly, the formula of RMSE is the following:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Additionally, sMAPE is the average across all forecasts made for a given period and can be measured through the following formula:

$$sMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$$

This metric is an upgrade of MAPE and aims at solving the problem of large errors that occur when explained values are close to zero and the significant difference between the absolute percentage errors when the explained value is greater than the expected explained value or the other way around.

The last metric is the R^2 score, also known as the coefficient of determination, which is the proportion of the variation in the outcome variable that is predictable from the independent variables. Therefore, the larger the R^2 score, the better the model's performance.

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

The formula R^2 score represents the ratio between the Sum of Squared Residuals (SSR) and the Sum of Squared Total (SST).

3.4. Challenges

The first challenge was collecting a negative keyword associated with Bitcoin on Google Trends. After measuring the correlation between Bitcoin, followed by different negative keywords, and the Bitcoin price, they all were positive. The reason is that searching for information about Bitcoin on Google increases the demand for Bitcoin and, in turn, the price as explained by Liang et al., 2022.

The data on *News Regulation* and *Theft* should be collected with another approach. Namely, restricting the data collection to a single source can generate bias and mislead the importance of these variables due to the limited amount of news. However, since the news about regulation and theft is widespread, it can also avoid the model to see the price difference when there is news or not.

Lastly, in the section 3.3.1 the ML used to predict the Bitcoin price were explained. DNN does not have a built in time dimension. So, everytime a new feature is generated, the covariates do not have all the information of the previous days, only the previous day and the two days ago. In contrast, LSTM has a long-term memory characteristic that contributes to achieve a greater performance.

4. Results

4.1. Exploratory Data Analysis (EDA)

Bitcoin price

The evolution of Bitcoin price between 2017 and 2021 is shown in **Figure 6**.

During this period, the price varied from $1,000 in 2017 to $60,000 in 2021. In particular, in 2021, the price started to increase sharply in the beginning of the year from around $20,000 to $60,000.

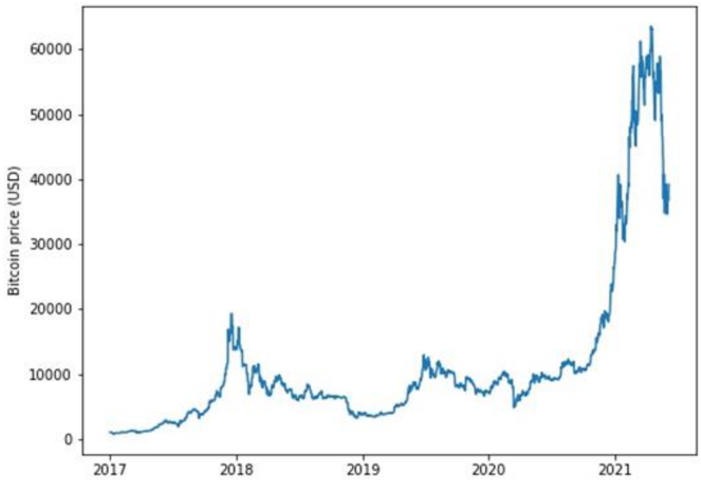


Figure 6 – Market price evolution. The Bitcoin price presented its lowest value in 2017 and reached its highest value in 2021

As explained in section 2.1.2, the price is affected by specific events that might provoke substantial variations in its value. Firstly, in March 2021, a spike resulted from Tesla's announcement that it had acquired $1,5 billion worth of Bitcoin.

Lastly, **Figure 7** and **Figure 8** inform that compared to other assets such as gold and S&P500, Bitcoin is the most volatile asset. Concretely, the variation of the logarithm of price is sharper than the variation of the other two assets.

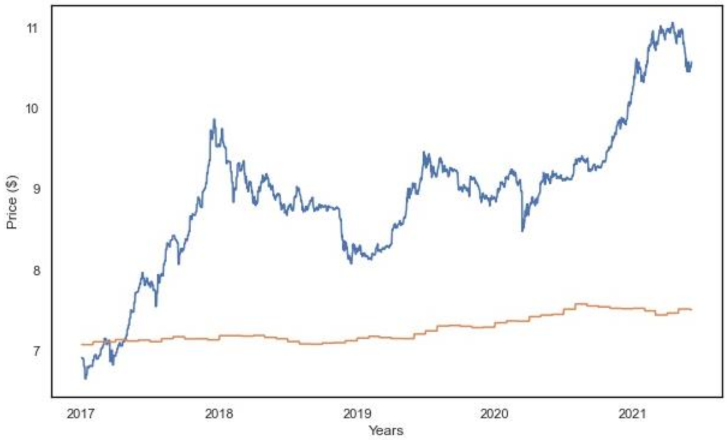


Figure 7 – The evolution of Bitcoin price vs. Gold. The Bitcoin price is considerably more volatile than gold price between 2017 and 2021

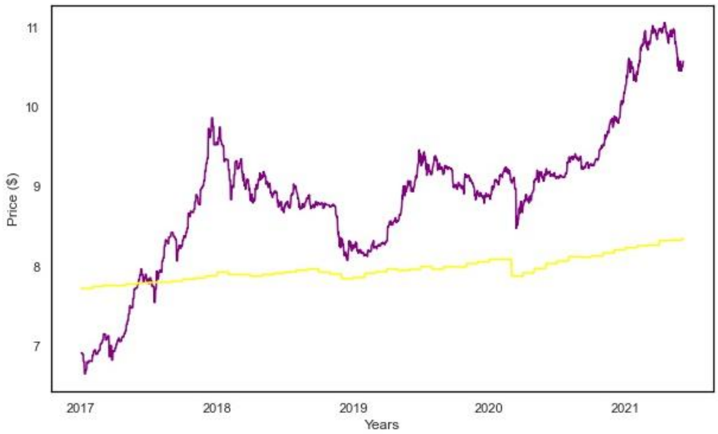
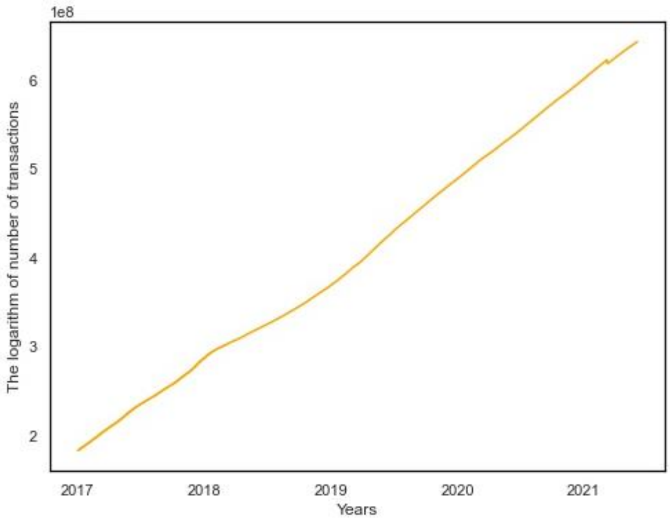


Figure 8 – The evolution of Bitcoin price vs. S&P500. The Bitcoin price is considerably more volatile than S&P500 price between 2017 and 2021

Concluding, although the factors that led to these variations in Bitcoin price will be presented later, it was highlighted that also occasional events can influence it, namely, leading to sharp variations.

Number of Transactions

The **Figure 9** shows that from 2017 to June 2021 the number of transactions increased.



z

Figure 9 – The number of Bitcoin transactions over the years. The number of transactions grew most of the time between 2017 and 2021

Figure 9 is particularly useful to conclude that both the Bitcoin price and the number of transactions tend to increase until around March 2021. This fact might result from either a growth in the demand for Bitcoin or other factors that also influence de Bitcoin price.

After March 2021, while the Bitcoin price started to decrease, the number of transactions continued to rise. This situation might derive from a decrease in supply or, instead, other external covariates that also impact the price.

Sentiment Analysis on Google Trends

First of all, **Figure S3** represents three distributions of the variables *GT_legal*, *GT_cryptocurrency* and *GT_scam*. The scale of some of the plots was adjusted in order to compare the index variable represented on the y-axis.

These graphs showed that the median of the keyword “cryptocurrency” after Bitcoin is more preponderant in comparison because 50% of the indexes are approximately 8 against 1 of *GT_legal* and 3 of *GT_scam*. This finding will be confronted with the results in section 4 to conclude if this reflects on Bitcoin price.

Concluding, the predominant sentiment regarding Bitcoin is neutral based on this approach of Google Trends.

Heatmap correlation plot

Figure S4 in the appendix presents the correlations between each predictor and the Bitcoin price.

lg_Open, *lg_High* and *lg_Low* are just statistics or the initial value of the Bitcoin price on that day. Thus, they have a correlation with the lg_price of 1 expectedly. The inclusion of these variables that report to the prediction day would lead to data leakage. Therefore, the variables will be only used in the previous two days for DNN.

Regarding the external factors, *lg_SP500, Popularity, GT_legal*, *GT_cryptocurrency, GT_scam*, *lg_Oil* and *News Regulation* has a positive correlation with the *lg_price* as it was expected based on the other articles. On the other hand, *Theft* has a negative correlation with the *lg*_price. Intuitively, this fact was already expected since negative news about Bitcoin theft inhibits individuals from investing in this asset, thus reflecting a shift downwards of the demand and, consequently, a *lg_price* decrease. These correlations are particularly important to indicate if the effect of covariates *lg_price* will be positive or negative and, in turn, monitor if the signal of the covariates in the models is correct.

GT_scam has the lowest correlation with the outcome variable compared to *GT_cryptocurrency* and *GT_legal*. Therefore, *GT_scam* will be removed in GLSAR, so the other two become dummies.

In the last place, the variables with a correlation below 0.3 can be considered irrelevant. This information was used in section 3.1 as an attempt to optimize the DNN model.

Concluding, the heatmap correlation plot helped to prepare the data in section 3.2.

Heatmap predictors correlation matrix plot

First of all, **Figure S5** represents the correlation among the predictors.

The most relevant information that the plot provides is the existence of a perfect correlation between *n-transactions-total* and *block-size*. Thus, the simultaneous presence of both variables creates the issue of multicollinearity in the GLSAR model and DNN model, so *block-size* will be dropped. Since *n-transactions-total* is extremely important as a component of market equilibrium, *block-size* will not be considered when designing the GLSAR model.

Furthermore, *lg_Open, lg_High* and *lg_Low* have a perfect correlation with other variables in the model, which also represents multicollinearity. However, these three variables were already dropped.

Time series analysis

First of all, this project intended to analyse the stationarity, the normality of the errors and the homogeneity, to conclude if the estimators have an inaccurate standard error and if they are biased and inefficient, respectively.

In the first place, to evaluate the stationarity, the time series was decomposed by viewing the Partial Auto Correlation Function (PACF) plot in **Figure S6**, which measures how much the y variable (*lg_price*) is correlated to past values of itself and how far back a statistically significant correlation exists. The PACF plot is different from the Auto Correlation Function (ACF) plot because PACF controls for correlation between past terms. In order to avoid an analysis too extensive, it was only displayed the PACF. This plot shows that the first two lags have considerably higher values than the others. This fact suggests that a statistically significant correlation may exist for up to two days in the data.

Additionally, after doing an Augmented Dickey-Fuller (ADF) test, it was obtained the approximated value of -1.346 for the statistic and 0.608 for the p-value. Therefore, the ADF test confirms the results that the plot indicated. Since the p-value is not less than 0.05, we must assume the series is non-stationary. The PACF plot and the ADF test conclusions were incorporated in the formulation of the GLSAR model in section 3.3.1.1 by correcting autocorrelation.

Secondly, a normality analysis was performed through a Jarque-Bera test, a Shapiro-Wilk test, and a Kolmogorov-Smirnov test. The three tests presented a p-value of 0.0001530, 1.2971531e e^{-8}, 2.8294370 e^{-64}, respectively. If a significance level of 0.05 is chosen, then all the three normality tests indicate the residual term does not follow normal distribution. Therefore, the estimators will be biased.

Lastly, the homogeneity was analysed by using a Goldfeld–Quandt test, a p value of Goldfeld–Quandt test, a Breusch–Pagan test, and a White test. The values obtained are 0.9999, $3.3121e^{-34}$ and $1.0374e^{-43}$, respectively. If a significance level of 0.05 is chosen, then the last two normality tests indicate the residual term does not follow normal distribution. Therefore, the estimator will be inefficient.

This information is important to conclude that the estimators are not BLUE. Concretely, they are biased, inefficient and are associated with an inaccurate standard error. Therefore, a GLSAR was used in section 3.3.1.1 in order to solve the problem of the serial correlation error.

Variance Inflation Factors (VIF) test

Firstly, a Variance Inflation Factors (VIF) test aims to identify which variables are affected by multicollinearity and the strength of the correlation. The results for the GLSAR model are presented in **Table S2.**

The minimum VIF is 1, but there is no upper limit. Therefore, 1 indicates no correlation between the independent variable and the others. Additionally, when VIF varies between 1 and 5, it suggests a moderate correlation. However, any values in that range do not require corrective measures. Lastly, if VIF is greater than the critical level of 5, there is multicollinearity, the coefficients would be poorly estimated, and the p-values would be questionable.

Table S2 indicates that News Regulation, Theft, *lg_market-cap, lg_output-volume, lg_Oil, lg_trade-volume, lg_unique-adresses* and *Popularity* are appropriate for this model because they do not present critical levels of multicollinearity. Thus, this model is not appropriate for identifying the main determinants of the Bitcoin price. In this project, GLSAR will be used for prediction, even with the issue of multicollinearity, since inadequate GLSAR coefficients will be compared to the LSTMs.

Noise analysis

In order to reduce the noise in the DNN model, the noise smoothing with the *tsmoothie* library was proceeded for the explanatory subset of the data for training, testing and validation, respectively. The blue line in **Figures S7**, **S8** and **S9** represent the final data used for modelling.

4.2. Models' performance

In order to measure the performance of this regression problem with time series, the Mean Squared Error (MSE), the RMSE, the symmetric Mean Absolute Percentage Error (sMAPE) and the R^2 score were selected. In general, **Table 1** shows that the performance of GLSAR presents the best performance for all the metrics. Specifically, the GLSAR error metrics (MSE, RMSE and sMAPE) are the lowest. Moreover, R^2 scores indicate that the covariates in GLSAR explain 96.26% of *lg_price*. The second best performance is the LSTM's performance with an R^2 score of 81.63%. Regarding the DNN model, the covariates explain 81.27% of the outcome variable.

Model	MSE	RMSE	sMAPE	R^2 score
GLS	0.016803	0.129628	0.009797	0.962613
LSTM	0.082559	0.287330	25.111598	0.816310
DNN	0.084189	0.290154	22.201072	0.812682

Table 1 – Machine Learning models' performance. LSTM achieve the best performance among the complex models with an R^2 score of 81.63%

Figure 10 shows the relationship between the actual values of the logarithm of the Bitcoin price and the predictions obtained by the GLSAR model. In particular, orange points represent individual values, and the blue line marks the ideal scenario where the actual values precisely match the predicted values. For higher actual values, the scatter points are more spread. In those cases, it is expected that the error be higher.

The evolution of the RMSE over time is shown in **Figure S10** in the appendix. When the Bitcoin price is higher, and the magnitudes of the fluctuations are more significant, the RMSE achieved the highest values as well as intense variations for both models.

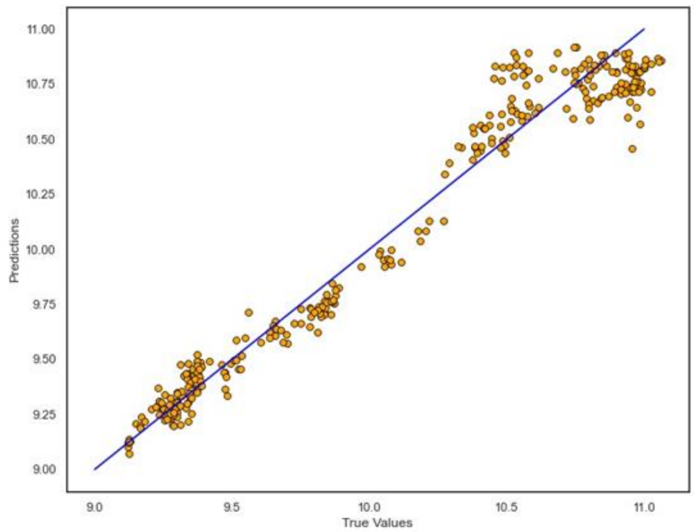


Figure 10 – The true values vs. predictions. GLSAR performs best when the true values are lower

Moreover, **Table 2** represents the GLSAR summary of all the information about the model. Firstly, the R^2 is 91.6%. Additionally, the high condition number indicates that there are issues of multicollinearity or others. Thus, the GLSAR models do not appropriately represent this business problem, and the coefficients cannot be interpreted. Also, Lasso, as a regularization technique, produced worse results. Thus, it was not applied.

Lastly, *lg_avg-confirmation-time* and *lg_circulatingBTC* were deleted from the model. since they were not statistically significant as transaction-based variables.

```
==============================================================================
Dep. Variable:               lg_price   R-squared:                       0.916
Model:                          GLSAR   Adj. R-squared:                  0.915
Method:                 Least Squares   F-statistic:                     1022.
Date:                Sat, 10 Dec 2022   Prob (F-statistic):               0.00
Time:                        18:18:40   Log-Likelihood:                 2333.9
No. Observations:                1613   AIC:                            -4632.
Df Residuals:                    1595   BIC:                            -4535.
Df Model:                          17
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                     -13.8789      1.237    -11.215      0.000     -16.306     -11.452
lg_market-cap               0.1319      0.006     23.253      0.000       0.121       0.143
lg_trade-volume             0.0152      0.003      4.467      0.000       0.009       0.022
lg_transaction-fees-usd     0.0251      0.006      4.200      0.000       0.013       0.037
lg_difficulty               0.2071      0.023      9.110      0.000       0.162       0.252
lg_hash-rate               -0.0808      0.010     -8.005      0.000      -0.101      -0.061
lg_miners-revenue           0.3207      0.012     27.763      0.000       0.298       0.343
n-transactions-total     2.112e-09   2.09e-10     10.111      0.000     1.7e-09    2.52e-09
lg_Popularity              -0.0054      0.001     -7.152      0.000      -0.007      -0.004
n-unique-addresses         -0.0420      0.016     -2.639      0.008      -0.073      -0.011
lg_output-volume           -0.1128      0.022     -5.194      0.000      -0.155      -0.070
lg_Gold                     0.4117      0.135      3.051      0.002       0.147       0.676
News Regulation             0.0016      0.010      0.160      0.873      -0.018       0.021
GT_legal                    0.0025      0.006      0.417      0.677      -0.009       0.014
GT_cryptocurrency           0.0051      0.001      6.514      0.000       0.004       0.007
lg_SP500                    1.3720      0.125     11.016      0.000       1.128       1.616
lg_Oil                      0.4481      0.046      9.660      0.000       0.357       0.539
Theft                       0.0020      0.013      0.160      0.873      -0.023       0.027
==============================================================================
Omnibus:                      166.171   Durbin-Watson:                   1.621
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1079.795
Skew:                          -0.219   Prob(JB):                    3.35e-235
Kurtosis:                       6.984   Cond. No.                     7.93e+10
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.93e+10. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Table 2 – GLSAR regression results. The high condition number suggests that at least one Gauss Markov assumption is not followed

From this point onward, an LSTM model was built, and **Figure 11** depicts how the training operation becomes stable after the second epoch. Also, it is possible to verify that on epoch number 25, the training loss equals the validation loss. Thus, 25 is the most accurate number to minimize the loss. If the number of epochs becomes too high, it might cause the model to overfit the training data.

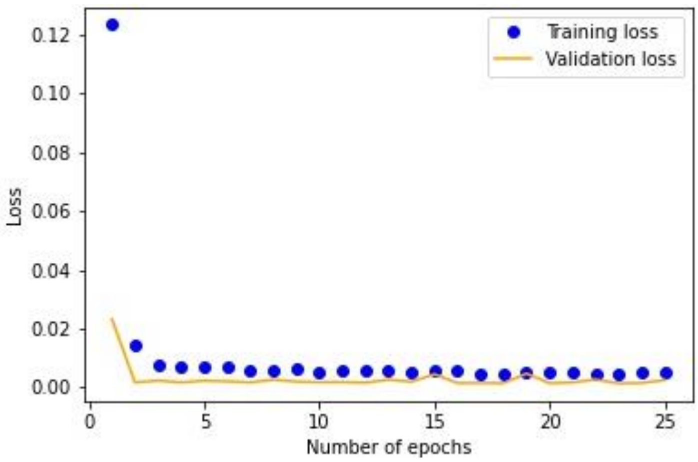


Figure 11 – The training operation become stable after the second epoch

Furthermore, **Figure 12** illustrates the predicted values compared with the actual values for the predicted period, which coincides with the period of the test set. Concretely, this plot displays that LSTM leads to undervalued predictions most of the time because the predicted values represented by the blue line are under the actual values line (orange line). However, the graph indicates that the predicted values follow the trend of the actual values, expect in the last few days.

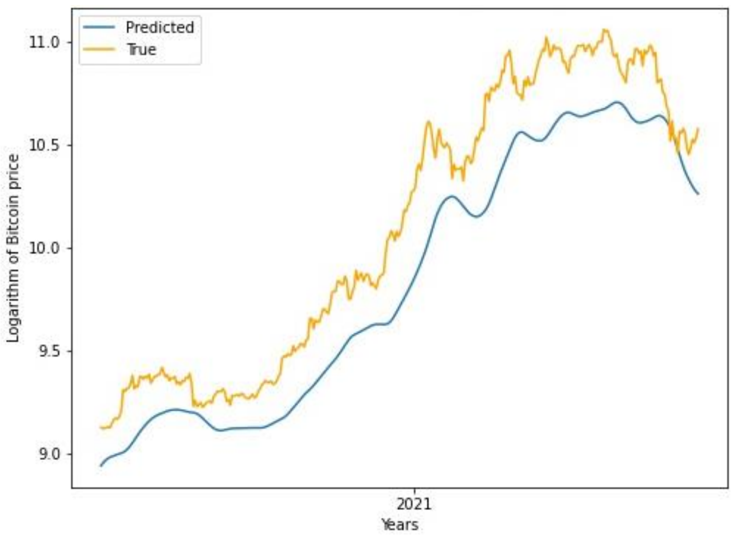


Figure 12 – Predicted values vs. true values for the predicted period of time. LSTM' predicted values are very close to the real ones.

In addition, **Figure 13** plots the difference between the predicted values and the true values of DNN for the predicted period. Concretely, the predicted values are very close or even equal to the actual ones in the first months of this period. However, slightly before 2021 the predictions start to diverge from the actual values. In comparison to LSTM**,** DNN predictions exhibit considerably more noise.

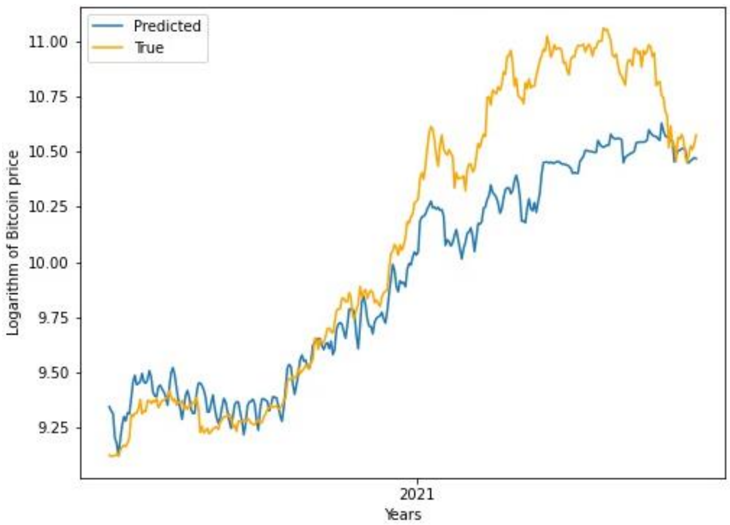


Figure 13 – Predicted values vs. true values for the predicted period of time. The noise is less present in DNN predictions for the last months

In brief, GLSAR presents the best performance, followed by LSTM.

4.3. Models' results: Bitcoin price determinants' importance

4.3.1. The GLSAR approach

First, **Figure 14** presents the impact of each variable on the logarithm of Bitcoin price measured using the GLSAR approach. The blue bars mark the transaction-based variables, while the orange marks represent the external variables.

In general, external factors significantly impact the logarithm of Bitcoin price compared to the transaction-based variables.

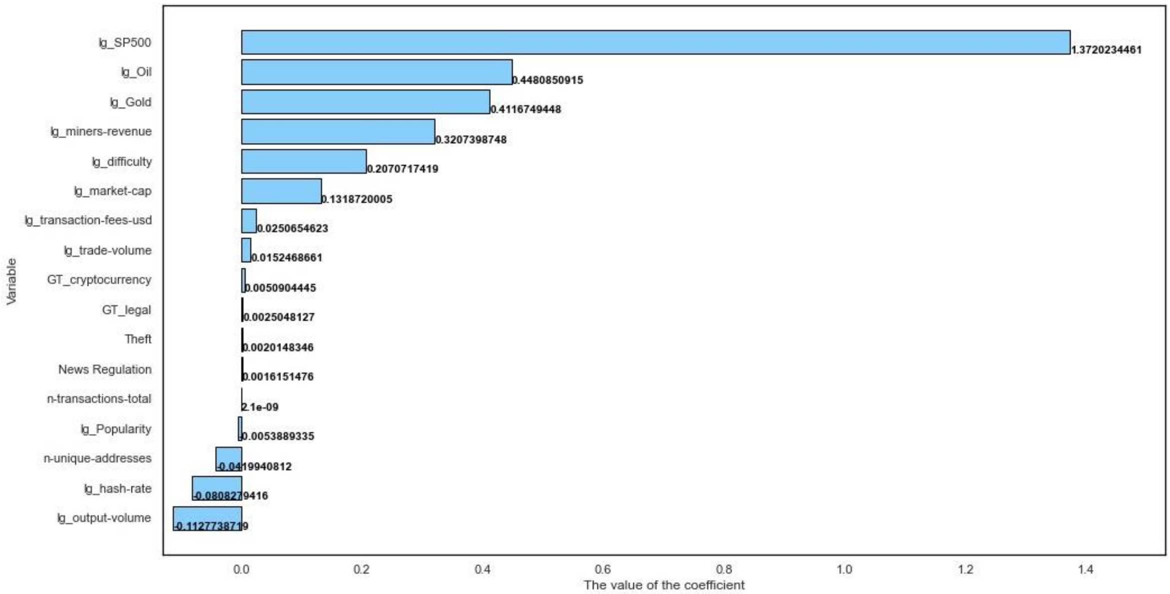


Figure 14 –The importance of each factor on the logarithm of Bitcoin price. lg_SP500, lg_Gold lg_miners-revenue are the three important features to lg_price

External factors

Since this project aims to emphasise the impact of external factors, **Figure S11** was provided in the appendix as a restriction of the previous plot to external factors.

In general, lg_SP500 is the variable with the most substantial influence on the outcome variable.

Furthermore, the two variables that mainly impact the explained variable are *lg_Oil* and *lg_Gold*. Firstly, *lg_Gold* positively impacts *lg_price,* as expected, because of the positive correlation presented in section 4.1. Moreover, as occurs with *lg_Gold*, *lg_Oil* also has a positive impact, as shown in section 4.1. However, it is less significant in comparison to *lg_Gold.*

Transaction-based variables

Regarding the transaction-based variables, the most significant one is *lg_market-cap,* followed by *lg_miners-revenue* and *lg_hash-rate.* All are statistically significant at the 5% level.

4.3.2. The SHAP applied on DNN approach

Global interpretability – Variable importance

Firstly, the bar plot in **Figure 15** indicates which covariate is the most relevant to the overall prediction. Thus, this graph presents the importance of global feature importance. Next, the SHAP values for all features obtained on each test point are aggregated to calculate the mean absolute value.

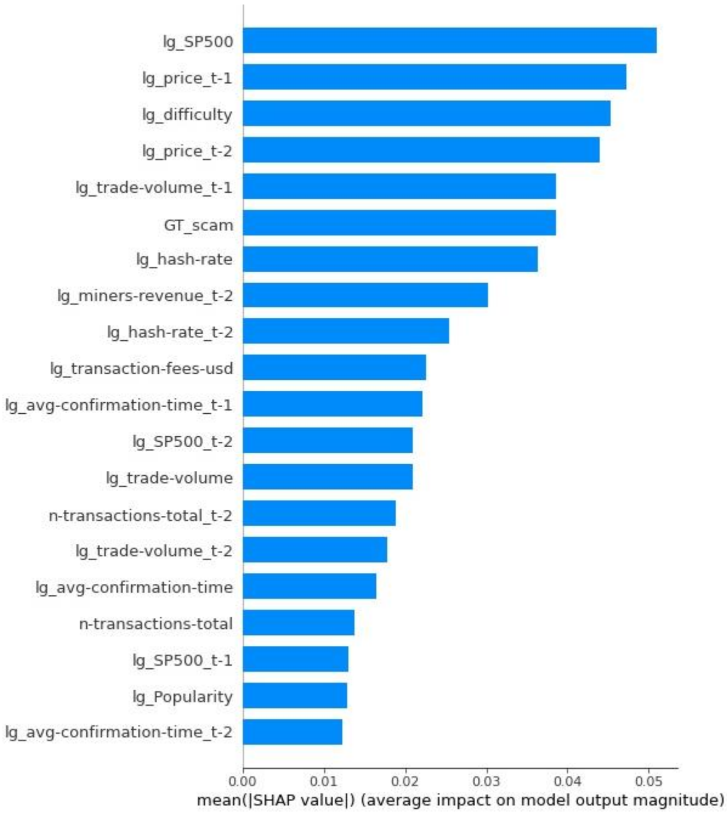


Figure 15 – The global importance of each covariate to the logarithm of Bitcoin price. The transaction-based variables are more significant than the external factors

The graph indicates that the three most significant variables to *lg_price* are *lg_SP500, lg_price_t-1* and *lg_difficulty*. Thus, these groups have the highest predictive power.

Similar to the plot in **Figure 15**, the other summary plot represented in **Figure 16** also provides the global feature importance, but for each test data point. Since the prediction days are 323, each feature corresponds to 323 data points. The plot's horizontal value indicates whether that feature's influence led to greater or lower prediction. Additionally, all the covariates in descending order by impact are named on the vertical axis. Lastly, the colour highlights the relationship between SHAP values and feature importance. For instance, for *lg_SP500*, the plot indicates that higher feature values lead to higher SHAP values. Therefore, larger values for *lg_SP500* weight cause a higher predicted *lg_price*. On the contrary, *lg_avg-confimation-time_t-1* presents the opposite relationship. Concretely, larger SHAP values are driven by lower values of this feature.

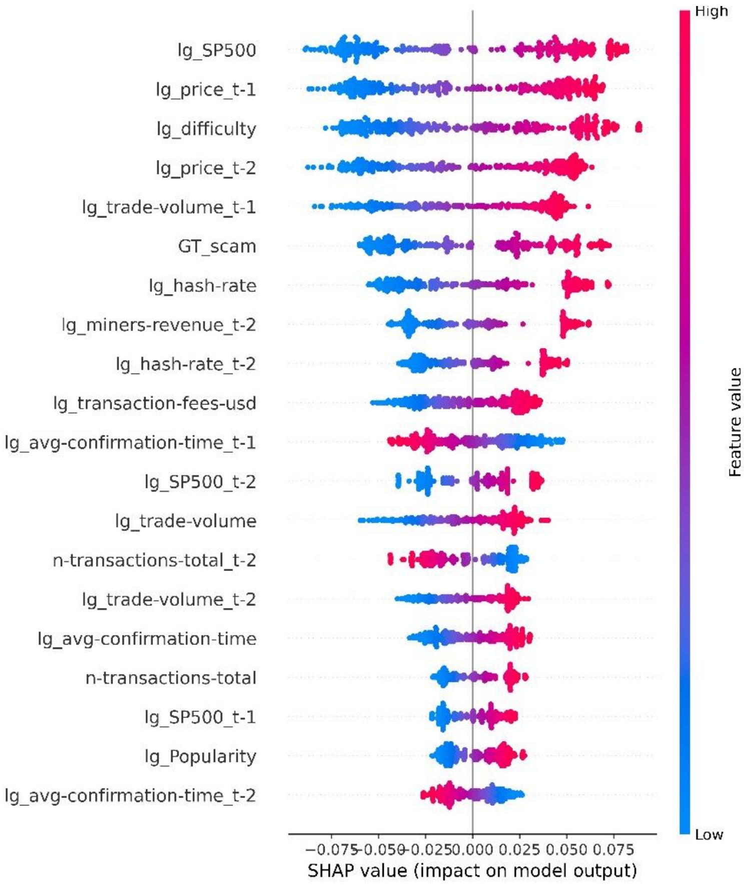


Figure 16 – The global importance of each covariate to the logarithm of Bitcoin price. All the features have a positive relationship with the SHAP value, except GT_cyptocurrency_t-1.

Global interpretability – Partial dependence

The previous explanation is illustrated in **Figures 17** and **18**. In addition, both figures inform that both marginal effects of their correspondent features on the predicted outcome are linear. Also, **Figure 17** shows that *lg_price_t-1* is the explanatory variable that *lg_SP500* most interact with. In the case of **Figure 18**, *lg_avg-confimation-time_t-1* has the strongest interaction with itself in the previous day (*lg_avg-confimation-time_t-2).*

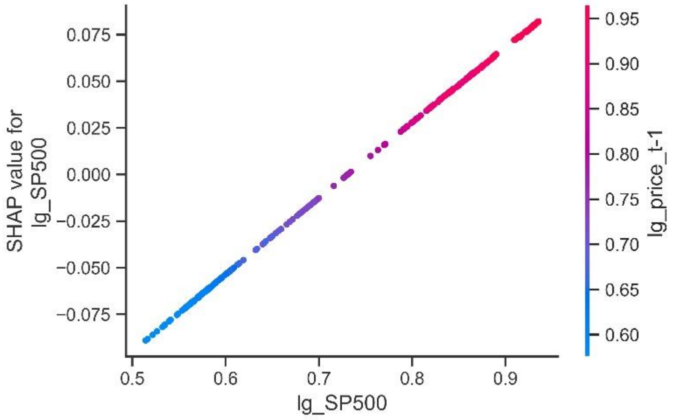


Figure 17 – The relationship between lg_SP500 and its SHAP value. lg_SP500 has a positive relationship its SHAP value.

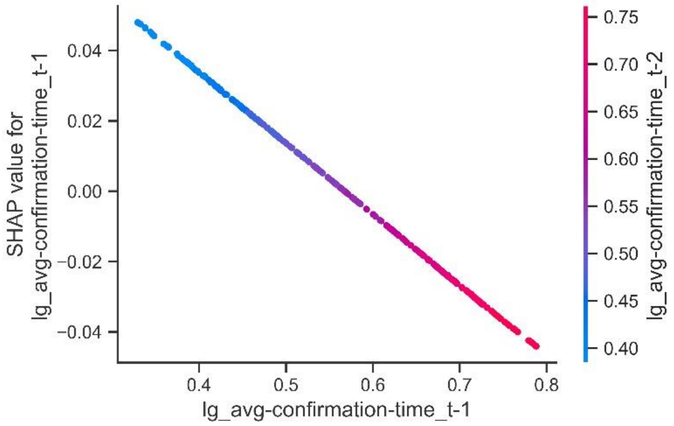


Figure 18 –The relationship between lg_avg-confimation-time_t-1 and its SHAP value. lg_avg-confimation-time_t-1 has a negative relationship its SHAP value.

Local interpretability

In terms of local interpretability, **Table 3** provides the feature values for the first five days in 202 that contribute to obtaining the predicted values (*predictions*) per day. These values do not represent precisely the values introduced in the model for prediction because they suffered noise reduction as a pre-processing pipeline. However, this is the only way to explain how the model achieved the referred predicted values.

Timestamp	lg_SP500	lg_price_t-1	lg_difficulty	lg_price_t-2	lg_trade-volume_t-1	GT_scam	lg_hash-rate	lg_miners-revenue_t-2	lg_hash-rate_t-2	lg_transaction-fees-usd	predictions
2021-01-01	8.24111	10.270094	3.737670	10.217436	20.436972	5	18.741053	17.338821	18.884601	14.190790	10.031992
2021-01-02	8.24111	10.274450	3.737670	10.270094	20.223255	5	18.754386	17.320592	18.849094	14.733514	10.046436
2021-01-03	8.24111	10.288537	4.174387	10.274450	20.012480	6	18.799702	17.185150	18.741053	14.878065	10.185022
2021-01-04	8.24111	10.379581	4.174387	10.288537	21.055376	6	18.890398	17.278383	18.754386	15.072941	10.204580
2021-01-05	8.24111	10.404286	4.174387	10.379581	20.698800	6	18.793352	17.396870	18.799702	15.117782	10.206846

Table 3 – Local interpretability - the predictors' values and predictions obtained in the first five days of 2021. The predicted value on 1st January 2021 is 10.03 USD

Moreover, the force plot is another technique to visualize the feature importance for a certain day. To illustrate, **Figure 19** presents the importance of each covariate on 1st January 2021.

Figure 19 – The impact of the features on Bitcoin price prediction on 1st January 2021. The red marks highlighted variables impact positively, while the blue marks impact negatively

The base value measures the average of the model output over the test data. **Figure 19** shows that the predicted logarithm of Bitcoin price on that day is 10.03, as indicated in bold. Interestingly, this is the approximated value that appears in the column *predictions* on that day in **Table 3**. Moreover, all the features that are highlighted in red, such as *lg_SP500,* increase the predicted value. On the other hand, blue shadow features such as *GT_legal_t-1* decrease the predicted value.

In this case, there are many predictors. Thus, the force plot does not present them well, and the decision plot in **Figure 20** and the waterfall in **Figure 21** are a good complement.

The decision plot should be interpreted from the bottom to the top. If all 57 variables were displayed, then the blue line would start in the base value represented by the vertical line. In addition, the numbers in parenthesis are the mean contribution of the covariates to the outcome variable for the test period. In the last place, the blue line achieves the value of 10.03, which represents the predicted *lg_price* on that day.

Finally, the interpretation of the waterfall should also be made in an ascending way. Since now all the features are considered, the interpretation starts at the base value of 9.922. This plot has the advantage of showing the effect of each explanatory variable on that day instead of the mean effect. In the last place, the sum of these effects leads to the predicted value of 10.03.

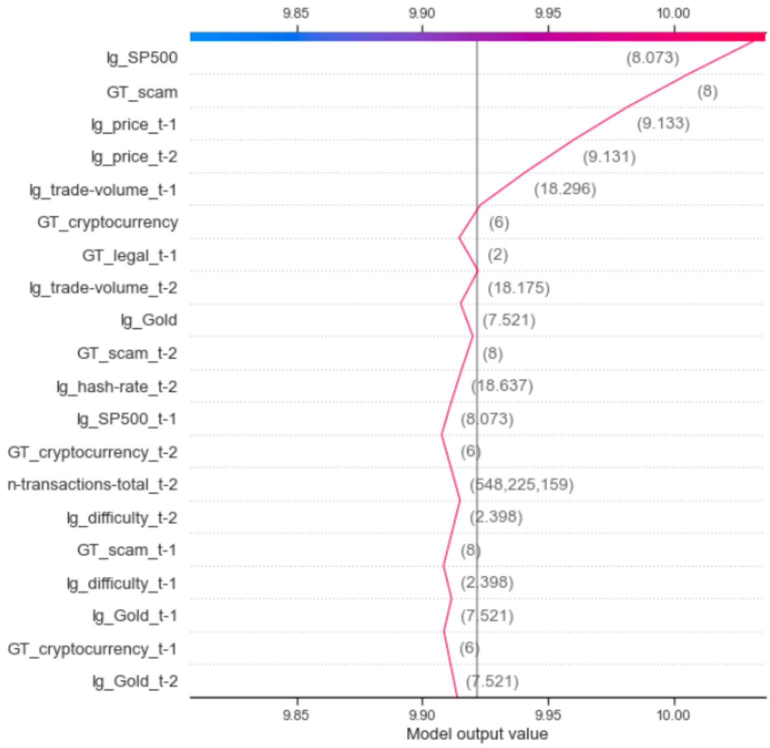


Figure 20 – The impact of the features on Bitcoin price prediction on 1st January 2021. On 1st January 2021, lg_SP500 and GT_scam had the strongest impact on Bitcoin price.

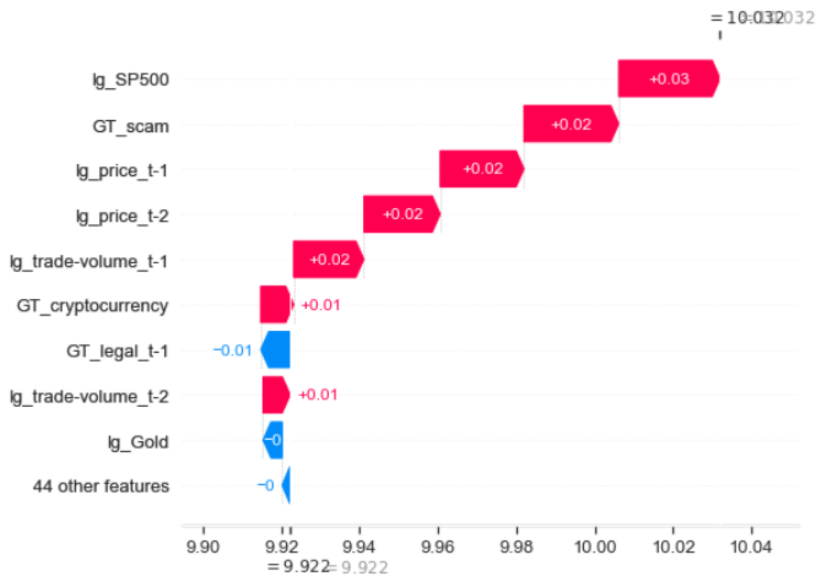


Figure 21 - The impact of the features on Bitcoin price prediction on 1st January 2021. On 1st January 2021, lg_SP500 had the strongest impact on the outcome variable

5. Discussion

5.1. Which is the most appropriate Machine Learning model to identify the determinants of Bitcoin price?

As explained before, the main purpose of this thesis is not to find the best model to identify the determinants of Bitcoin price. Instead, this study aims to determine if the additional effort of building a complex model is worthwhile in comparison to simpler models and discuss the limitations of applying XAI in some ML models.

In the first place, the GLSAR model, as well as other econometric models, are not the most appropriate to address problems of this nature because it is unrealistic to follow all the Gauss Markov assumptions. To illustrate, the GLSAR summary presented in section 4.2 warned that some issues might be present in that estimation. Even if the GLSAR solved the serial correlation problem, the requirement of linearity could be compromised. Moreover, if the model does not meet the Gauss-Markov assumptions, then it is not possible to interpret the regression coefficients anymore. Considering all these limitations, GLSAR will not be compared with other results in literature. Instead, deep learning models such as LSTM are more reliable for this difficult task of identifying the determinants that condition the Bitcoin price (H. Guo et al., 2021; Khedr et al., 2021). There are not many authors using econometric models for forecasting Bitcoin price.

Furthermore, among the ML models, LSTM reached the best performance, as referred by Carbó & Gorjon, 2022; Ji et al., 2019; Pabuccu et al., 2020; Ranjan et al., 2022; Rather, 2022. Currently there is no Application Programming Interface (API) available to apply SHAP on the top of ML models with a temporal dimension.

In order to solve this limitation, a DNN was designed. Although the DNN model does have a built in time dimension, time-derived features were added, allowing to integrate the time relationship. Specifically, the covariates in the previous day and the trend were included in the group of explanatory variables. Moreover, if strategies to improve performance are applied, for example, eliminating irrelevant variables, this model still can be improved. However, even after applying noise reduction techniques, the model still presents some variations in the performance, which make the LSTM preferable compared to DNN.

Moreover, most of the data collected for this thesis are highly non-linear and non-stationary. As explained in section 3.3.1, LSTM mitigates this limitation because the model does not require any specific *ex-ante* relationship between the outcome variable and the covariates. Thus, it can be used in non-linear and non-stationary time series scenarios. Additionally, the model can be used in multicollinear variables because the activation functions attribute it to a non-linear structure and also uses regularisation techniques. Regularisation, in this case, through the dropout method, strongly stabilises the coefficients assigned to the features, thus reducing the problem of multicollinearity among covariates. The results support the idea that LSTM achieves, most of the time, better performance as explained by Ji et al., 2019; Pabuccu et al., 2020; Rather, 2022.

Comparing the results of this thesis with the conclusions available in the literature is quite challenging since they reported different periods. However, whenever it was possible, a similar time range was selected. Also, the performance evaluations were recalculated to obtain each metric's absolute values. Compared to the article of the original dataset Rajabi et al., 2022, the LSTM presented an RMSE reduction of approximately 45% (from 19126.447. to 10573.726). In the case of DNN, the RMSE decreased around 14% (from 13642.156 to 11735.523). The results of this thesis are in line with Guo et al., 2021 regarding the R^2 scores. Concretely, Guo et al., 2021 achieved an R^2 scores for the Multilayer perceptron (MLP) and the LSTM of 87.95% and 89.83%, respectively. MLP is quite equivalent to the simple DNN. This thesis presented an LSTM with an R^2 score of 81.63% and an LSTM with an R^2 score of 81.27%.

The evolution of the RMSE over time shows that when the Bitcoin price is higher, and the magnitudes of the fluctuations are more significant, the RMSE achieved the highest values as well as intense variations for both models. This highlights the fact that Bitcoin as a strong volatile asset is hard to predict.

Concluding, this thesis provides an explainable model, a "black box" for time series data but without interpretability, and a black-box model that was explained using SHAP. The choice between LSTM and DNN will depend on the answer to the question: are the gains in performance that high to compromise interpretability?

5.2. Which technique of Explainable AI should be applied in the Machine Learning model?

Firstly, this thesis followed the approach presented by Babaei et al., 2022; Carbó & Gorjon, 2022; Fior et al., 2022; Giudici & Raffinetti, 2021; Mandeep et al., 2022; Molnar et al., 2020 and used SHAP as an XAI technique.

In general, the SHAP plots exposed in section 4.3 demonstrate that this technique is particularly useful for analysing the feature importance in the global model. This is particularly useful to the already referred decision-makers so they can generalise the determinants that mainly condition the Bitcoin price.

Lastly, it explains how features led to the predicted price on a certain day. Therefore, whoever is in charge of making decisions that involve the Bitcoin price can understand why on a particular day, especially when facing atypical events, the price reached that value. This also can be a strategy to generalise, but only for specific days in similar conditions. For example, if Tesla's announcement in March 2021 about a high investment in Bitcoin (section 2.1.2) that led to a peak in Bitcoin price was included in the covariates, this could be expressed in a local feature importance plot. Consequently, future announcements such as this one could attract investors' attention to a possible rise in the Bitcoin price.

5.3. What are the main determinants of Bitcoin price?

In the first place, the application of SHAP on DNN is the only XAI solution among all the approaches introduced in this project. Consequently, this is the method used to identify the Bitcoin price' determinants.

First, regarding the external variables, *News Regulation* and *Theft* were deleted because they have an almost insignificant correlation with the outcome variable, which improved DNN performance. However, the negative effect mentioned in Coulter, 2022; Lyócsa et al., 2020 could not be verified.

Globally, *SP500*, *GT_scam*, and *Popularity* the three most considerable features of the Bitcoin price. These results are in line with (Liang et al., 2022; Polasik et al., 2015). In particular, *SP500* is in the list of the top most relevant predictors for the current day, the previous day and two days ago, which enhance the impact of the variable. Similar to the results of CBOE Gold ETF Volatility Index (GVZ) in Liang et al., 2022, the gold price showed a negative effect on Bitcoin volatility on 1st January 2021. This derived from the fact that Bitcoin and gold can be used as substitutes for investment, which means that Bitcoin's worth is opposite to gold's. In section 4.1, a neutral sentiment was found to be more preponderant on Google Trends than negative sentiment. This result suggests that the most searched words on Google will not be necessarily the most important predictors to Bitcoin price.

Moreover, S&P500 has an important impact, as referred to in Bakas et al., 2022 and Mudassir et al., 2020. Contrary to Lyócsa et al., 2020, neutral and negative keywords after *Bitcoin* do have a significant effect on Bitcoin price. Regarding the popularity of Bitcoin on Google Trends, it is consensual that a higher number of searches raise the price (Liang et al., 2022; Polasik et al., 2015). Finally, the rest of the external variables, such as Oil, Gold, and *GT_legal* do not impact strongly on *lg_price*. However, these results diverge from Liang et al., 2022 and Ciaian et al' thesis.

In general, transaction-based variables impact more significantly on Bitcoin price than external variables. In fact, the SHAP summary plot shows that approximately 71% of the top twenty-one covariates belong to the transaction-based group. These covariates go in line with the ones used in Mudassir et al., 2020; Rajabi et al., 2022; and Ranjan et al., 2022.In addition, the three most important features are *lg_price* in the day before, *lg_difficulty*, and *lg_price* two days before.

Finally, the results of GLSAR coefficients are far from meeting the SHAP feature importance. For instance, while Gold and Oil are the second and the third most significant external covariates in GLSAR, the summary plot demonstrates that it is not true. In fact, this was expected due to GLSAR's limitations, as explained in section 5.1.

6. Conclusions

Limitations

First, the data collection process had two main limitations. On one hand, the negative keyword associated with *Bitcoin* on Google Trends did not decrease the price. Therefore, another approach should be addressed in the literature so a sentiment analysis can be done. Additionally, more newspapers are needed to indicate a post on the news about regulation and theft so we can measure these impacts on price.

Secondly, the challenges resulting from modelling time series data, for example, multicollinearity, non-linearity and non-stationarity, can be overcome if a LSTM is used. On the contrary, it is hard to ensure that these mandatory conditions are all met for linear regressions.

Lastly, even if the time-related features are incorporated in DNN, this model does not achieve a performance as high as LSTM does.

Opportunities

In order to enrich the dataset, *News Regulation* and *Theft* could be collected from a more significant range of news sources to make the coefficients of the predictors significant and unbiased. Although some criteria need to be defined, such as selecting the top 10 journals or two days a week, this leads to more generic news data.

Other web sources can be also used to do sentiment analysis with a negative-impact variable. For instance, using Web scrapers such as Scrapy or BeautifulSoup allows users to collect specified information such as blog comments from the web. Also, a Web browser plugin provides information from any public website using HTML and exports the data to the desired file format. For example, Webscraper.io is a free extension for the Google Chrome web browser that can be used for this purpose. Coulter, 2022's approach could be followed by including the geographical source of the news.

This project allows adding successively new variables to transaction-based covariates and external predictors, so the uncertainty associated with Bitcoin is minimized. For example, variable events could represent rare situations published on the news, such as Tesla's purchase in March 2021 that led to a sharp increase in Bitcoin price.

Moreover, this thesis presented some improvements in DNN in order to obtain a higher performance. Therefore, another strategies can be applied to upgrade the model.

Regarding the LSTM interpretability, other libraries such as "timeSHAP" could be employed to LSTM to explain the model.

To provide a complete XAI analysis, model explainability could also be explored in further research. This technique would be essential to understand the mechanisms of the ML model.

Finally, this thesis had the advantage of using a public dataset. Thus, the code can now be shared to contribute to future research.

Conclusions

The main contribution of this master's thesis is to introduce a recent topic that summarizes and complements the articles published related to Bitcoin price' determinants.

Firstly, this project aggregated the potentially most relevant predictors from the literature in a single dataset. All the data extracted was from public sources, and new novel variables were added. The final dataset report to the period between 2017 and 2021. The results indicated that the approach used for news about regulation and theft; and the sentiment analysis on Google Trends have some limitations.

In order to identify the main determinants of Bitcoin price, three different approaches were provided, namely, a "white-box" model that is too simplistic, a "black-box" model with high performance and a "black-box" model with an XAI technique. DNN, with the SHAP technique, was the best model to identify the determinants of Bitcoin price.

The results showed that the highest Bitcoin price, S&P500, the price in the previous day and two days ago, and the difficulty in mining a Bitcoin block are generally the essential features.

Finally, local interpretability can provide decision-makers with the most significant predictors on a particular day.

To the best of the author's knowledge, this is the first study on this topic. The discussion presented in this article could be beneficial in exploring future research problems and finding ways to solve them.

7. Bibliography

Artificial intelligence and machine learning in financial services Market developments and financial stability implications. (2017). *FSB.*

Babaei, G., Giudici, P., & Raffinetti, E. (2022). Explainable artificial intelligence for crypto asset allocation. *Finance Research Letters, 47,* 102941. https://doi.org/10.1016/J.FRL.2022.102941

Bakas, D., Magkonis, G., & Oh, E. Y. (2022). What drives volatility in Bitcoin market? *Finance Research Letters, 50*, 103237. https://doi.org/10.1016/J.FRL.2022.103237

Böhme, R., Christin, N., Edelman, B., & Moore, T. (2015). Bitcoin: Economics, Technology, and Governance. *Journal of Economic Perspectives*, *29*(2), 213–238. https://doi.org/10.1257/jep.29.2.213

Carbó, J. M., & Gorjon, S. (2022). Application of Machine Learning Models and Interpretability Techniques to Identify the Determinants of the Price of Bitcoin. *Banco de Espana*.

Chen, J.-H., Wu, C.-H., Tsai, Y.-C., & Yen-Chi Chen, S. (2022). Explainable Digital Currency Candlestick Pattern AI Learner. *2022 14th International Conference on Knowledge and Smart Technology (KST),* 91–96. https://doi.org/10.1109/KST53302.2022.9727231

Ciaian, P., Rajcaniova, M., & d'Artis Kancs. (2016). The economics of BitCoin price formation. *Applied Economics*, *48*(19), 1799–1815. https://doi.org/10.1080/00036846.2015.1109038

Coulter, K. A. (2022). The impact of news media on Bitcoin prices: modelling data driven discourses in the crypto-economy with natural language processing. *Royal Society Open Science*, *9*(4), 220276. https://doi.org/10.1098/rsos.220276

Fior, J., Cagliero, L., & Garza, P. (2022). Leveraging Explainable AI to Support Cryptocurrency Investors. *Future Internet*, *14*(9). https://doi.org/10.3390/fi14090251

Foley, S., Frijns, B., Garel, A., & Roh, T.-Y. (2022). Who buys Bitcoin? The cultural determinants of Bitcoin activity. *International Review of Financial Analysis*, 102385. https://doi.org/10.1016/J.IRFA.2022.102385

Giudici, P., & Raffinetti, E. (2021). Shapley-Lorenz eXplainable Artificial Intelligence. *Expert Systems with Applications*, *167,* 114104. https://doi.org/10.1016/J.ESWA.2020.114104

Guo, H., Zhang, D., Liu, S., Wang, L., & Ding, Y. (2021). Bitcoin price forecasting: A perspective of underlying blockchain transactions. *Decision Support Systems, 151,* 113650. https://doi.org/10.1016/J.DSS.2021.113650

Guo, Q., Lei, S., Ye, Q., & Fang, Z. (2021). MRC-LSTM: A Hybrid Approach of Multi-scale Residual CNN and LSTM to Predict Bitcoin Price. 2021 International Joint Conference on Neural Networks (IJCNN), 1–8. https://doi.org/10.1109/IJCNN52387.2021.9534453

Hoepner, A. G. F., McMillan, D., Vivian, A., & Simen, C. W. (2021). Significance, relevance and explainability in the machine learning age: an econometrics and financial data science perspective. *The European Journal of Finance*, *27*(1–2), 1–7. https://doi.org/10.1080/1351847X.2020.1847725

Ilk, N., Shang, G., Fan, S., & Zhao, J. L. (2021). Stability of Transaction Fees in Bitcoin: A Supply and Demand Perspective. *MIS Quarterly*, *45*(2), 563–692.

Ji, S., Kim, J., & Im, H. (2019). A comparative study of bitcoin price prediction using deep learning. *Mathematics*, *7*(10). https://doi.org/10.3390/math7100898

Jo, H., Park, H., & Shefrin, H. (2020). Bitcoin and sentiment. *Journal of Futures Markets*, *40*(12), 1861–1879. https://doi.org/https://doi.org/10.1002/fut.22156

Kang, S. H., McIver, R. P., & Hernandez, J. A. (2019). Co-movements between Bitcoin and Gold: A wavelet coherence analysis. *Physica A: Statistical Mechanics and Its Applications*, *536*, 120888. https://doi.org/10.1016/J.PHYSA.2019.04.124

Khedr, A. M., Arif, I., P V, P. R., El-Bannany, M., Alhashmi, S. M., & Sreedharan, M. (2021). Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. *Intelligent Systems in Accounting, Finance and Management*, *28*(1), 3–34. https://doi.org/https://doi.org/10.1002/isaf.1488

Kyriazis, N. A. (2020). Is Bitcoin Similar to Gold? An Integrated Overview of Empirical Findings. *Journal of Risk and Financial Management*, *13*(5). https://doi.org/10.3390/jrfm13050088

Li, X., & Wang, C. A. (2017). The technology and economic determinants of cryptocurrency exchange rates: The case of Bitcoin. *Decision Support Systems*, *95*, 49–60. https://doi.org/10.1016/J.DSS.2016.12.001

Liang, C., Zhang, Y., Li, X., & Ma, F. (2022). Which predictor is more predictive for Bitcoin volatility? And why? *International Journal of Finance & Economics*, *27*(2), 1947–1961. https://doi.org/https://doi.org/10.1002/ijfe.2252

Liu, Y., & Tsyvinski, A. (2018). *Risks and Returns of Cryptocurrency* (Working Paper Series, Issue 24877). https://doi.org/10.3386/w24877

Lyócsa, Š., Molnár, P., Plíhal, T., & Širaňová, M. (2020). Impact of macroeconomic news, regulation and hacking exchange markets on the volatility of bitcoin. *Journal of Economic Dynamics and Control*, *119*, 103980. https://doi.org/10.1016/J.JEDC.2020.103980

Mai, F., Shan, Z., Bai, Q., Wang, X. (Shane), & Chiang, R. H. L. (2018). How Does Social Media Impact Bitcoin Value? A Test of the Silent Majority Hypothesis. *Journal of Management Information Systems*, *35*(1), 19–52. https://doi.org/10.1080/07421222.2018.1440774

Maleki, N., Nikoubin, A., Rabbani, M., & Zeinali, Y. (2020). Bitcoin Price Prediction Based on Other Cryptocurrencies Using Machine Learning and Time Series Analysis. *Scientia Iranica*. https://doi.org/10.24200/sci.2020.55034.4040

Mandeep, Agarwal, A., Bhatia, A., Malhi, A., Kaler, P., & Pannu, H. S. (2022). Machine Learning Based Explainable Financial Forecasting. *2022 4th International Conference on Computer Communication and the Internet (ICCCI)*, 34–38. https://doi.org/10.1109/ICCCI55554.2022.9850272

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. *ArXiv E-Prints*, arXiv:2010.09337.

Mudassir, M., Bennbaia, S., Unal, D., & Hammoudeh, M. (2020). Time-series forecasting of Bitcoin prices using high-dimensional features: a machine learning approach. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-020-05129-6

Pabuccu, H., Ongan, S., & Ongan, A. (2020). Forecasting the movements of Bitcoin prices: an application of machine learning algorithms. *Quantitative Finance and Economics*, *4*, 679+. https://link.gale.com/apps/doc/A683719659/AONE?u=anon~be9bf538&sid=googleScholar&xid=fbdac6ff

Polasik, M., Piotrowska, A. I., Wisniewski, T. P., Kotkowski, R., & Lightfoot, G. (2015). Price Fluctuations and the Use of Bitcoin: An Empirical Inquiry. *International Journal of Electronic Commerce*, *20*(1), 9–49. https://doi.org/10.1080/10864415.2016.1061413

Rajabi, S., Roozkhosh, P., & Farimani, N. M. (2022). MLP-based Learnable Window Size for Bitcoin price prediction. *Applied Soft Computing*, *129*, 109584. https://doi.org/10.1016/J.ASOC.2022.109584

Ranjan, S., Kayal, P., & Saraf, M. (2022). Bitcoin Price Prediction: A Machine Learning Sample Dimension Approach. *Computational Economics*. https://doi.org/10.1007/s10614-022-10262-6

Rather, A. M. (2022). A new method of ensemble learning: case of cryptocurrency price prediction. *Knowledge and Information Systems*. https://doi.org/10.1007/s10115-022-01796-0

Sovbetov, Y. (2018). Factors Influencing CryptocurrencPrices: Evidence from Bitcoin, Ethereum, Dash, Litcoin, and Monero. *Journal of Economics and Financial Analysis*, *2*(2), 1–27.

Tomić, N., Todorović, V., & Čakajac, B. (2020). The potential effects of cryptocurrencies on monetary policy. *The European Journal of Applied Economics*, *17*(1), 37–48.

Yang, F., Du, M., & Hu, X. (2019). Evaluating Explanation Without Ground Truth in Interpretable Machine Learning. *ArXiv E-Prints*, arXiv:1907.06831.

Zhao, W., Singh, R., Joshi, T., Sudjianto, A., & Nair, V. N. (2021). *Self-interpretable Convolutional Neural Networks for Text Classification*. arXiv. https://doi.org/10.48550/ARXIV.2105.08589

Appendixes

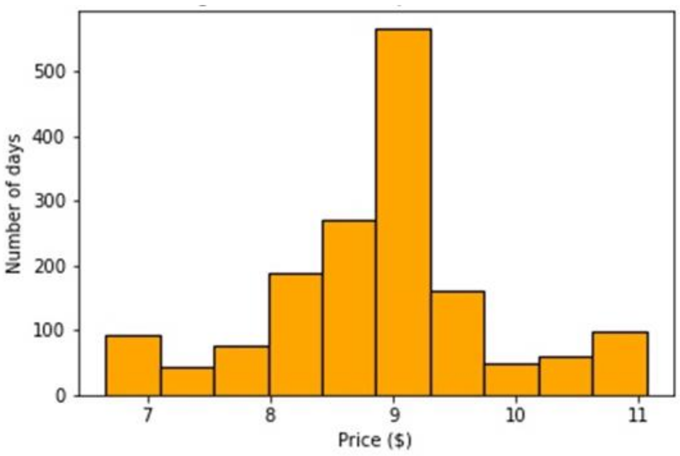


Figure S1 – Distribution of the logarithm of the market price. After the logarithm is applied on market price the distribution becomes normal Market price distribution.

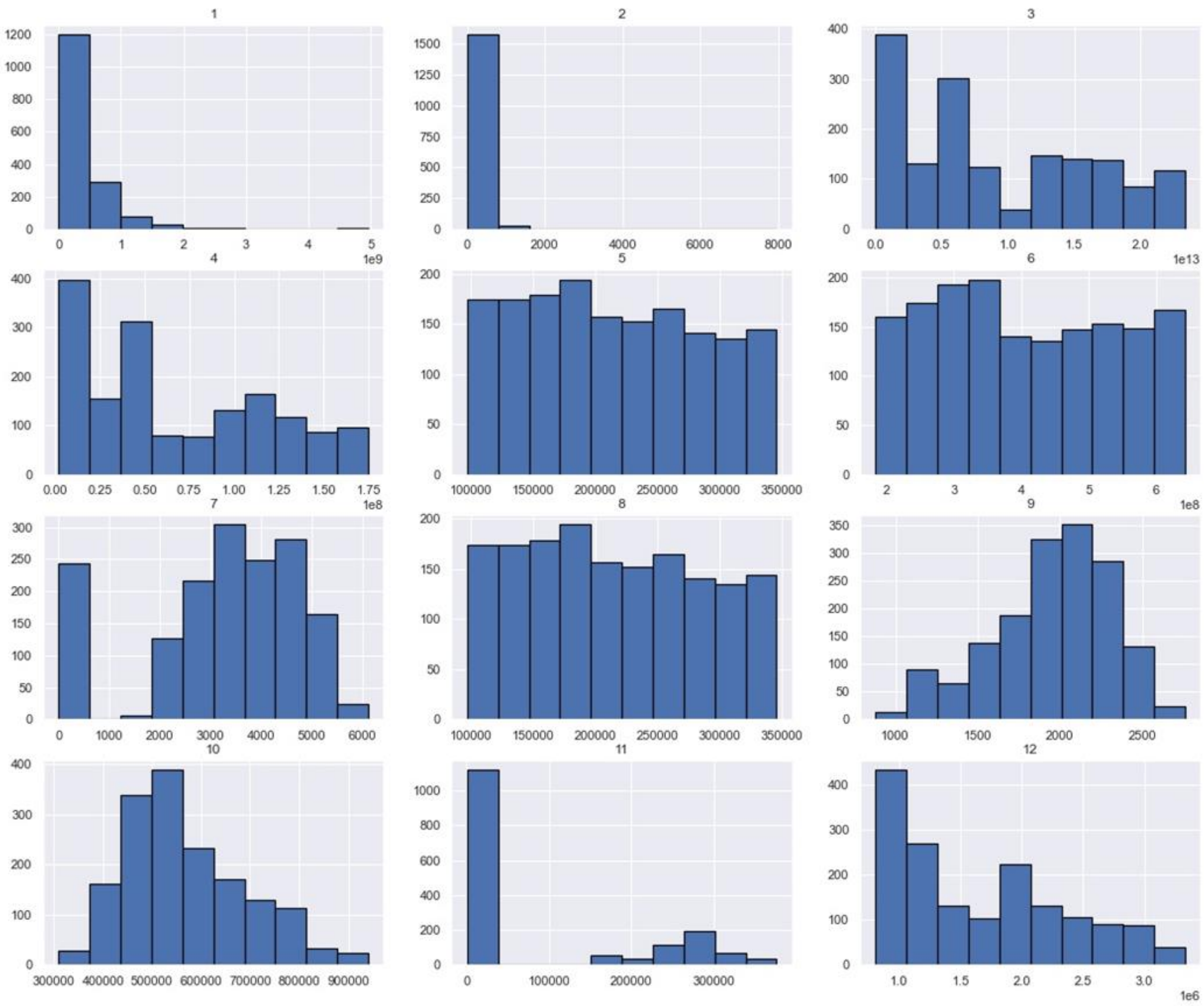


Figure S2 – Histograms of remaining transaction-based variables. The logarithm transformation will be applied on the skewed variables 1,2,3,4,9,10 and 12

Variable	*Description*
Lg_price	The logarithm of Bitcoin price.
Lg_market_cap	The logarithm of the total value of the Bitcoin market in US dollars.
lg_trade-volume	The logarithm of the total volume of Bitcoin transactions.
lg_transaction-fees-usd	The logarithm of the total cost in dollars of extracting a Bitcoin.
lg_avg-confirmation-time	The logarithm of the average time it takes for a transaction to be accepted in the extracted block and placed in the public ledger.
lg_difficulty	The logarithm of how difficult it is to mine a Bitcoin block, or in more technical terms, to find a hash below a given target. High difficulty means that it will take more computing power to mine the same number of blocks, making the network more secure against attacks.
lg_High	The logarithm of the highest price of Bitcoin per day.
lg_Low	The logarithm of the lowest price of Bitcoin per day.
lg_hash-rate	The logarithm of the performance of a minor device. In other words, the hash rate indicates the rate at which a miner succeeds in solving the hash to receive revenue.
block-size	The total size of transactions accepted in each block.
miners-revenue	The revenue received by the miners is for building each block.
n-transactions-total	The total number of transactions in the blockchain
Popularity	The index of Bitcoin searches on Google.
lg_Open	The logarithm of the initial price of Bitcoin per day.
n-payments-per-block	The total number of payments per block over the past 24 hours.
lg_circulatingBTC	The logarithm of the total number of mined Bitcoins that are currently circulating on the network.
lg_cost-per-transaction-percent	The logarithm of the percentage of miners' revenue they receive from their trading volume.
lg_fees-usd-per-transaction	The logarithm of the average cost in dollars of Bitcoin mining per trade.
n-unique-addresses	The total number of transactions, excluding those involving the network's 100 most popular addresses.
n-transactions-per-block	The total number of transactions made in each block.
Output-volume	The total value of all transaction outputs per day. This includes Bitcoins returned to the sender as change.
Confidence	Indication of future developments on households' consumption and saving, based upon answers regarding their expected financial situation, their sentiment about the general economic situation, unemployment and capability of savings.
lg_Gold	The logarithm of the gold price.
News Regulation	Dummy variable. It assumes the value 1 when there is a new in FT about regulation on that day and 0 when the opposite is true.
GT_legal	This index measures the importance of *Bitcoin* + *legal* in Google searches comparing with the other two keywords.
GT_cryptocurrency	This index measures the importance of *Bitcoin* + *cryptocurrency* in Google searches comparing with the other two keywords.
GT_scam	This index measures the importance of *Bitcoin* + *scam* in Google searches comparing with the other two keywords.
Oil	Oil price.
SP500	Standard & Poor's 500 is one of the most important stock market indexes in the world today.
Theft	The loss of Bitcoin that result from theft attacks.

Table S1 – List and description of the variables used in the final dataset

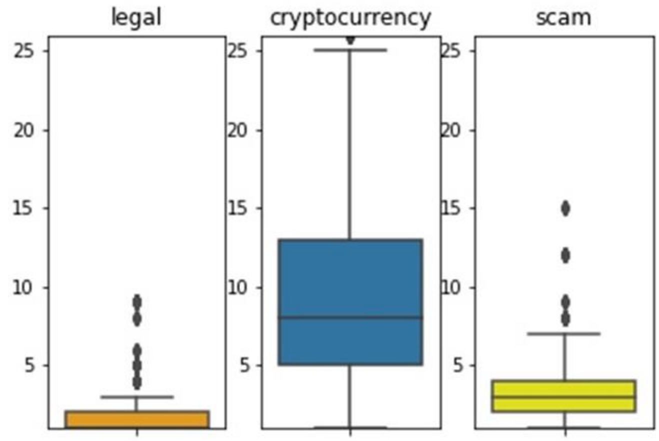


Figure S3 – The distribution of the three variables representing sentiment analysis on Google Trends. Bitcoin plus cryptocurrency variable has the highest median comparing to the other two.

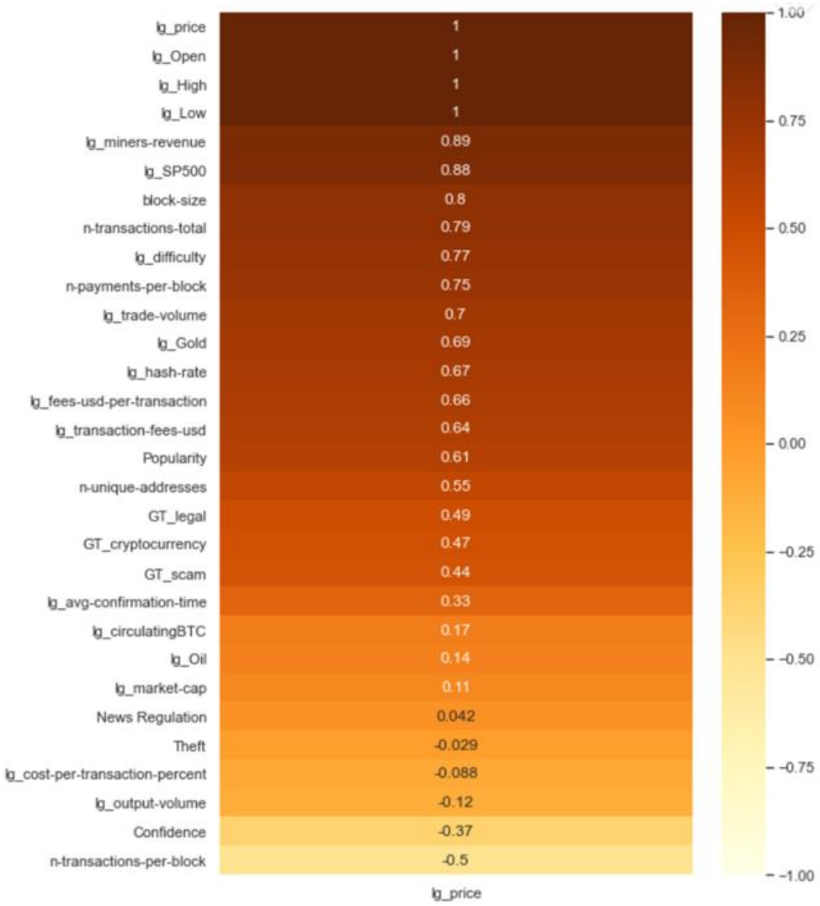


Figure S4 – Features correlating with the log of Bitcoin price. There are 7 features that are considered irrelevant with the module of correlation below 0.3.

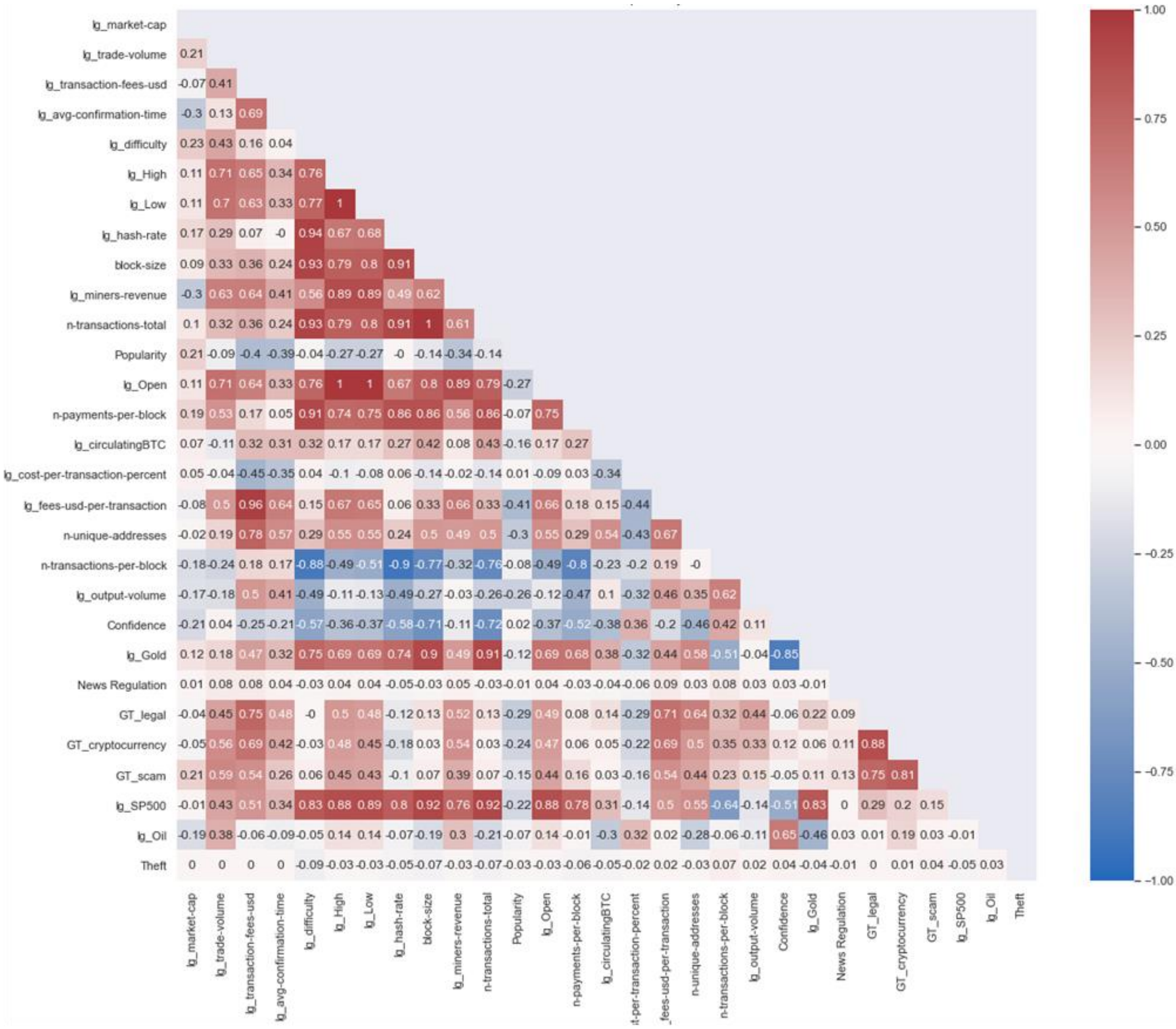


Figure S5– Correlation matrix of the explanatory variables. There are four variables that have perfect correlation (equal to 1), which represents a multicollinearity issue

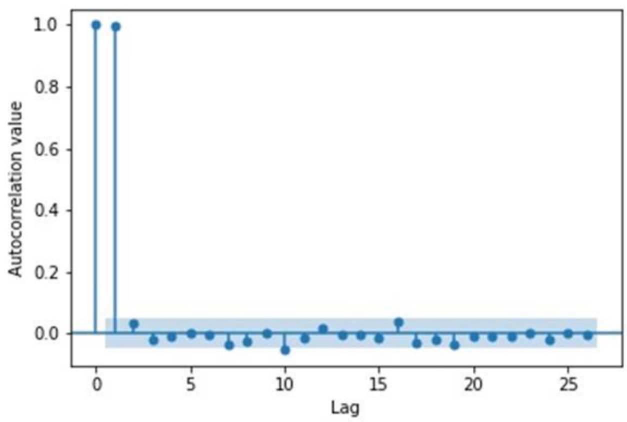


Figure S6 –Partial autocorrelation. lg_price is autocorrelated with the values of the past two days.

	Variable	VIF
11	News Regulation	1.019356
16	Theft	1.027712
0	lg_market-cap	3.137903
9	lg_output-volume	3.319067
15	lg_Oil	3.779949
1	lg_trade-volume	3.787265
8	n-unique-addresses	4.305979
7	Popularity	4.933656
12	GT_legal	7.014917
2	lg_transaction-fees-usd	7.640905
13	GT_cryptocurrency	8.756532
5	lg_miners-revenue	11.416097
14	lg_SP500	14.029970
4	lg_hash-rate	14.526925
10	lg_Gold	16.887963
3	lg_difficulty	28.196657
6	n-transactions-total	59.777472

Table S2 – The VIF results for GLSAR. All the variables above the critical value 5 have multicollinearity with another variable

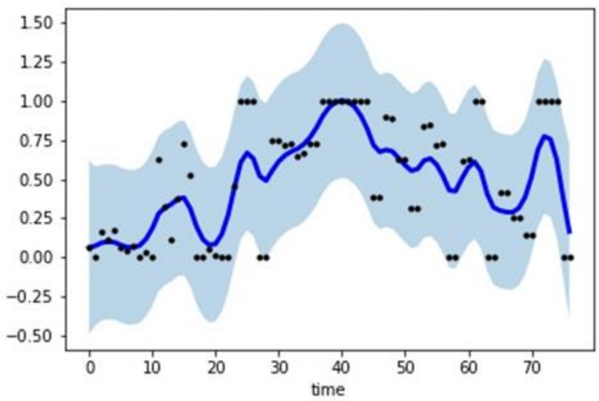


Figure S7 – Noise reduction of the covariates in the train data. The blue line represents the covariates train data being subject to noise smoothing.

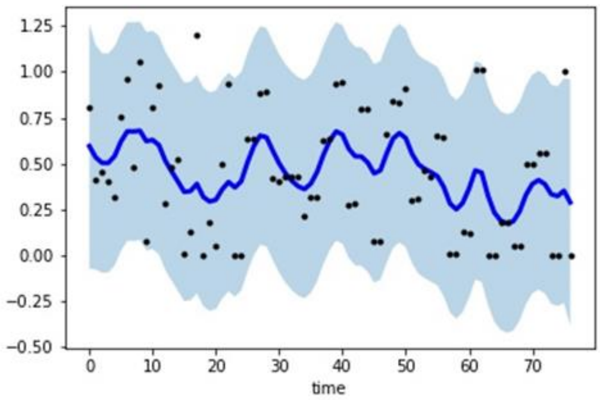


Figure S8 – Noise reduction of the covariates in the test data. The blue line represents the covariates test data being subject to noise smoothing

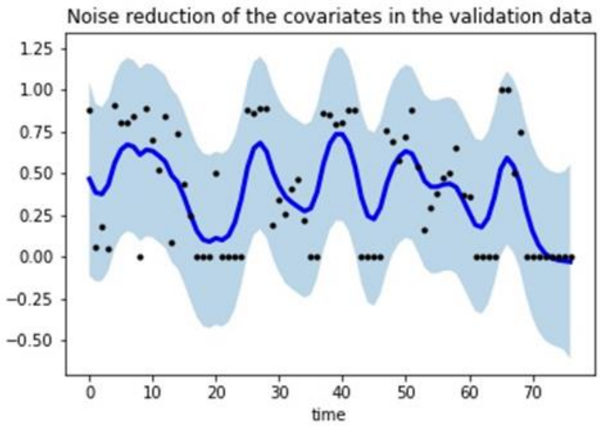


Figure S9 – Noise reduction of the covariates in the validation data. The blue line represents the covariates validation data being subject to noise smoothing.

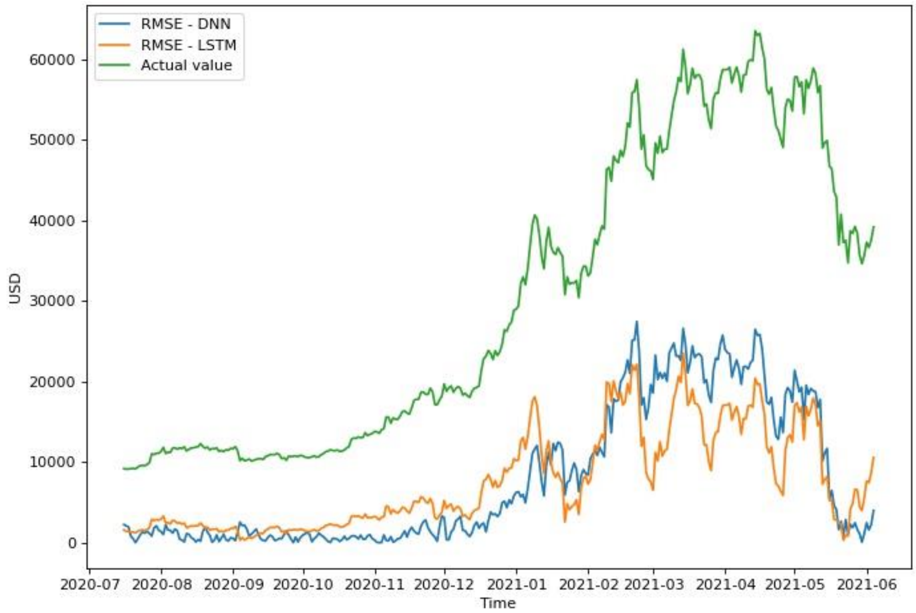


Figure S10 – The evolution of RMSE with the Bitcoin actual price for DNN and LSTM. Higher fluctuations in Bitcoin price led to higher variations of the RMSEs.

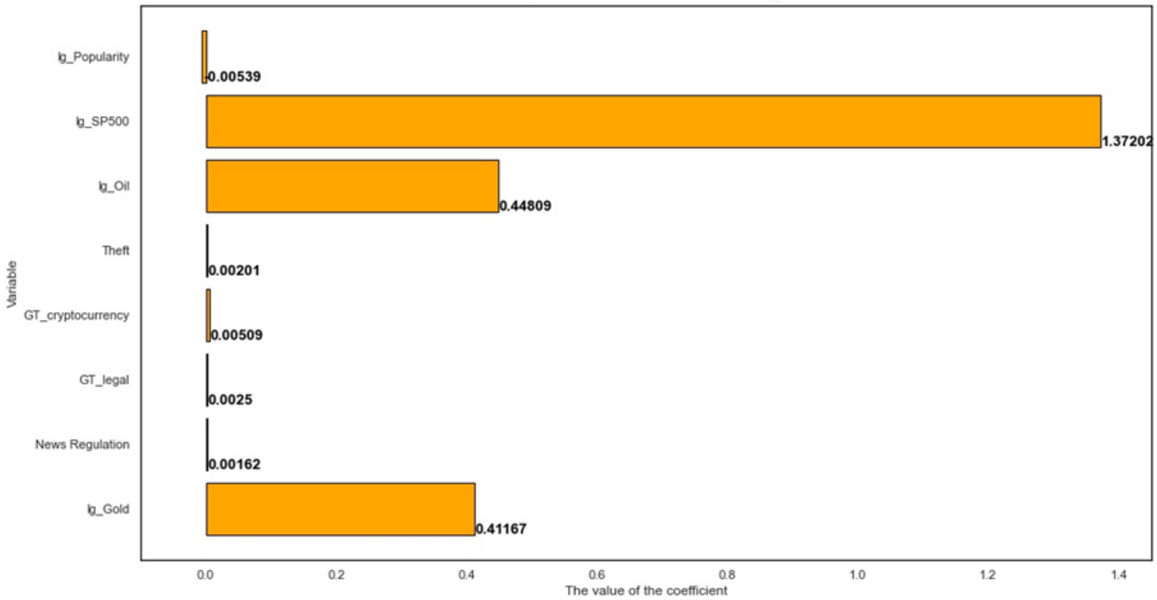


Figure S11 – The importance of each external factor on lg_price. S&P500, Oil and Gold are the most important external factors to Bitcoin price.