Predicting Bitcoin prices

The effect of interest rate, search on the internet, and energy prices

Joao Tiago Aparicio
INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, 1049-001 Lisbon, Portugal
joao.aparicio@tecnico.ulisboa.pt

Mario Romao
Advance/ISEG (Lisbon School of Economics &Management), Universidade de Lisboa, 1200-109 Lisbon, Portugal
mario.romao@iseg.ulisboa.pt

Carlos J. Costa
Advance/ISEG (Lisbon School of Economics &Management), Universidade de Lisboa, 1200-109 Lisbon, Portugal
cjcosta@iseg.ulisboa.pt

***Abstract* — The current study's goal is to explain the price of bitcoins. We examined the effect of Web search statistics, energy prices, and alternative investment (or cost of opportunity) on bitcoin prices in particular. The second goal is to find the algorithm with the best predictive power. Data were obtained from public and open data. We use a variety of machine learning algorithms to accomplish this. Statistical results were coherent according to the expectation.**

Keywords -criptocurrency, machine learning, bitcoins, financial market.

I. INTRODUCTION

Prediction has been one of the main objectives of pursuit science is or at least creating models that may help understand reality and further help prediction. [12] In recent years Machine Learning has had a very positive impact [1] on this journey, unlike in the distant past. [2] Like many other applications, we have seen rising interest in the development of price prediction methods in financial assets. [39] This is especially true when the possibility of obtaining financial profit is involved. So, it is fairly common to see new and inventive approaches to predict Bitcoin (BTC) prices that are focused on trading profit using blockchain data. [13] In this sense, the purpose of the present research is to explain the price of bitcoins. Specifically, we analyzed the impact of Web search statistics, energy price, and alternative investment (or cost of opportunity) on bitcoins prices. The second purpose is to identify the algorithm with better predicting power. To do it, we use several machine learning algorithms.

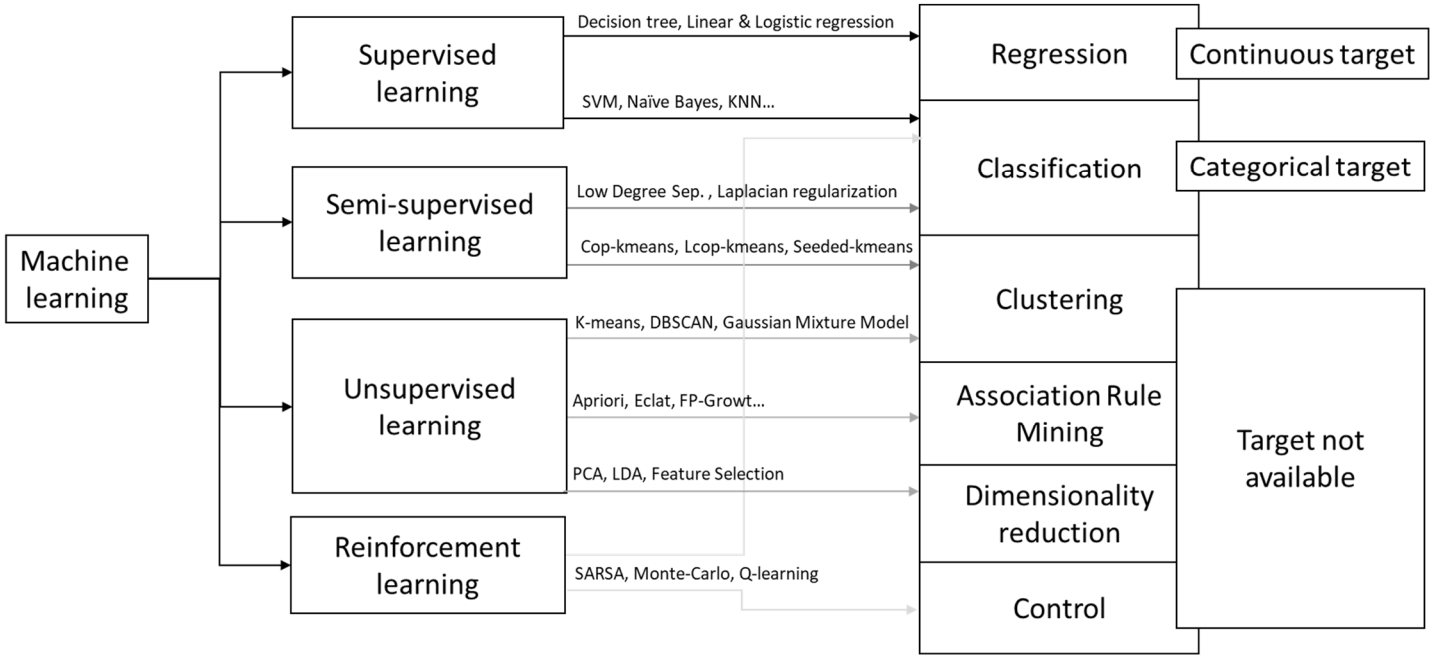

Figure 1. Machine Learning Algorithm types and tasks

2022 17th Iberian Conference on Information Systems and Technologies (CISTI)
22 – 25 June 2022, Madrid, Spain
ISBN: 978-989-33-3436-2

II. BACKGROUND

Understanding the price of assets can be done in a variety of ways using machine learning techniques (Figure 1). Usually, we may have dependent variables that could be explained by one or more independent factors. We might also use just time series. The premise of machine learning is that a computer program can learn and adapt to new data without the need for human intervention. Artificial intelligence is a field that includes machine learning. The following techniques classes are included in machine learning: Unsupervised, supervised, semi-supervised, and reinforcement learning. These all constitute several subfamilies of statistical tools to perform varied tasks depending on the available data. Supervised and unsupervised learning are two of the most common machine learning methodologies in the literature. Bellow, we explore the main ideas within each task using each high-level algorithm family, with examples on the study of bitcoin pricing prediction.

Supervised learning encompasses algorithms to classify data and predict outcomes using labeled data (data with known answers). Based on known samples, supervised learning can identify generate answers based on seen examples. [4]

Firstly, regarding regression, one method is to forecast prices based on prior data with models such as ARIMA [10,11]. In this case, a time series analysis is used to forecast future values based on its own. Explanatory variables can also be used, leading to other types of regression that may help understand the relationship between price and other external observations. These can be, for instance, Linear and Logistic regressions. There are numerous algorithms that can be employed in the case of regression. For example, OLS and LASSO. [6]

Other models can also be both used for Regression and Classification, such as CARTs. [13] This is a particular case where we may use a model that can use both continuous and categorical targets. In this case, this can be both the currency's price as a target or the low and high price as a binary target for instance. [14]. This principle can also be applied in the classification using models such as SVM and Neural Network based models.

Recently we have seen the rise of the deep learning revolution. This has solved a slew of previously "unsolvable" issues. The deep learning revolution did not begin with a single breakthrough. It essentially happened when a number of necessary variables were in place: computers were fast enough, computer storage was large enough, better training methods and tuning methods were developed. [20] These highly diverse sets of statistical representations have been studied to devise highly powerful and flexible representations regarding prediction tasks, many of the use categorical tasks. [12]

When some labeled data is particularly challenging to get buy, one may use a semi-supervised or unsupervised approach. This is the case where we may have some labeled features such as simple datapoints and other features based on information in raw and unlabeled text. In this case, we have seen this in the usage of text comments from social media platforms to further extract sentiment from text. [17,18] This is the data is fed without a label to the model, and then a sentiment analysis dictionary could be applied for further prediction. [15] This task can be used to either aggregate similar behaviors (clustering) or to attribute a label to a particular set of labeled circumstances (classification).

Unsupervised learning is used for predicting undefined characteristics in the available data when there are no labels to define the relevant information for the model to assimilate. Meaningful patterns can thus be extracted using clustering and association rule mining. This is particularly important if, for instance, we want to classify anomalies in bitcoin prices. [16] Generating a set of rules based on co-occurrence of certain characteristics in the data or devising clusters based on previously known trading heuristics or even set similarities within the available data can be extremely useful to understand the reality at hand may help decision making. Within this type of learning, we can also try to eliminate irrelevant dimensions on the data to be able to compute it in a timely way. Dimensionality reduction techniques such as PCA and LDA can thus be used to reduce the redundant information available. Let's say the goal is to relate the price of bitcoins based on five other cryptocurrencies. However, three of them have an incredibly similar evolution. The information of the said three dimensions could thus be summarized in just one dimension, drastically reducing the size of the data being used to their most important features.

Reinforcement learning (RL) is also an option where we may use unlabeled information. This is a particularly interesting solution when we know what kinds of overall results from a behavior we want to penalize and which we want to reward. This is a very common perspective in agent behavior modeling. For instance, if the goal of a certain research is to program a bot (or agent) to trade a particular type of asset, one may use a simulation of agents with different strategies and penalize those who perform the least amount of gains. [19] This is usually the case when we want to control the overall direction of the behavior, but we are not sure of every particular step of the process. This is a control task. Classification tasks can also arise from RL. After devising the different states and succession of states, the agent may be in at each time step one may devise patterns and classify them. The price of other cryptocurrencies is a feasible feature in an exploratory model whose focus is the price of a certain cryptocurrency. However, numerous writers claim that there are still more variables that can be utilized to explain cryptocurrency prices. [12, 13].

Recently we have seen that analyzing social media to understand the rise in prices of cryptocurrencies has yielded positive results. [21] However, it would be interesting to understand if the results obtained before the wild fluctuations, we have seen recently still hold true.

Other important assets such as energy directly affect the bitcoin price. [22] This is a very interesting assertion that may be relevant to study and further justify the sudden change in pricing due to increased energy prices.

Lastly, the role of alternative investments and lower yields may further push the market towards the cryptocurrency market. Previously, it has been postulated that periods of

2022 17th Iberian Conference on Information Systems and Technologies (CISTI)
22 – 25 June 2022, Madrid, Spain
ISBN: 978-989-33-3436-2

higher BTC price fluctuation may provide benefits as a diversification tool. [23] Nevertheless, does this mean that agents are investing in BTC as a main alternative to other safer options such as Treasury yields, and is there in fact a relationship between these?

III. METHOD

To test the above-mentioned hypotheses, we used CRISP-DM [24] for the data mining and applied POST-DS [8,9] for the project management. To understand the relationship between the BTC price and other entities mentioned we used energy prices from Eurostat [28] from 2009 to the beginning of 2022. Figure 2. Shows a plot of the data available

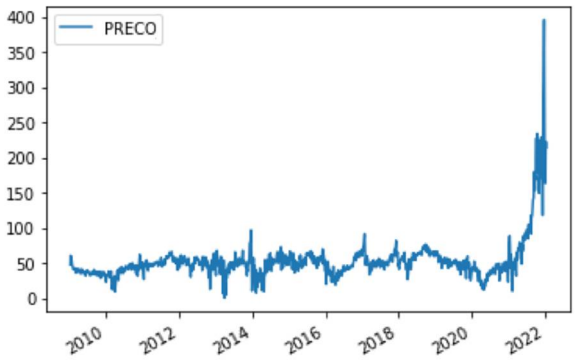

Figure 2. Energy Price Series

We chose the Treasury Note Yield Index closing to represent the cost opportunity since it was previously seen as an indicator for trust in future monetary policy. [26] Since BTC is a viable speculative investment to diversify portfolios with higher risk tolerance and not a currency by itself. [25] In Figure 3, we see the time series relative to the Treasury Note Yield Index at closing.

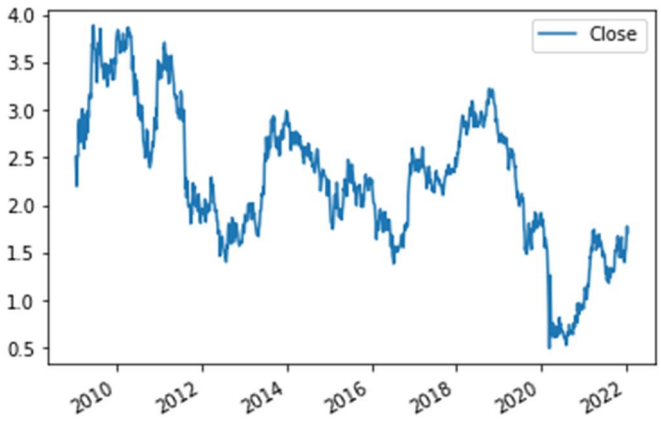

Figure 3. Treasury Note Yield Index at closing

Since previews studies look at the impact of social media and news outlets for understanding their impact on BTC prices, [27] we use google stats on bitcoins search trends over time (Figure 4).

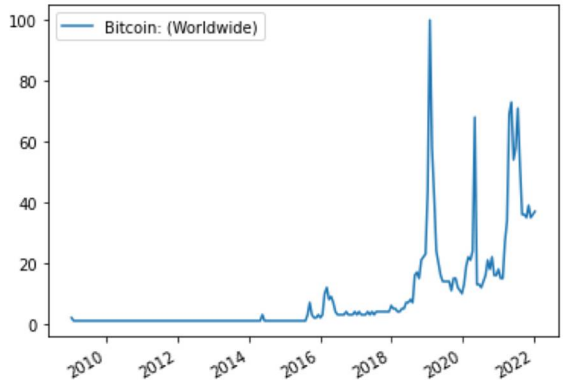

Figure 4. Google search results for the Bitcoins

To understand the relationships between the entities, we used different regressions since we have a continuous label, the BTC price itself. We use linear regressions as a baseline. Then we used Ridge regression so to combat multicollinearity issues. [33] This method, along with Lasso regression, is also used to understand the overall change in weights and if any of the independent variables tends to a zero weight. Since a sparse solution may be desirable. [34] Complementarily in the Bayesian ridge regression, all regression coefficients are regarded to have the same variance. [35]

On another note, Gradient Boosting, Random Forests, and Multi-Layered Perceptron regressors are much more flexible and powerful learners. [5,7] These usually provide a better predicting power. In that sense, we employ them to better understand if the current variables not only explain the BTC prices but also if they are adequate to predict the dependent variable. Gradient Boosting is capable of fitting a weak learner to the residual recursively in order to enhance model performance with a greater number of cycles, finding complex data structures, including nonlinearity and high-order interactions. [36] The Multi-layered Perceptron is a type of neural network which is based on backpropagation for increasingly learning the relationships between the feature space and the resulting variable. [37] And finaly, the Random Forest encompasses a set of arbitrary decision trees and averages their estimates. [38] And All of them have had fairly good results as predictors with these kinds of data. [36-38]

IV. RESULTS

Firstly, the data was preprocessed by grouping the said time series into a dataframe, and the unknown values were dropped. This implementation was based on Python. [3] The values were also standardized between one and zero so that the regression coefficients could be directly compared as a means to understand which variable had the most impact on the BTC price.

2022 17th Iberian Conference on Information Systems and Technologies (CISTI)
22 – 25 June 2022, Madrid, Spain
ISBN: 978-989-33-3436-2

The Alpha value for the Lasso regression was 0.5. The MLP Regression was done on 4x8 hidden layers with ReLu activations and gradient descent for 5000 iterations. The random Forest used 98 estimators with a maximum depth of 3. These values could be further fine-tuned.

The first set consisted in estimating a regression model using OLS (ordinary least square), or simply Linear regression. The Python language was also used in this step along with Pandas [31], and the *statsmodel* module [29][30].

As shown in Figure 5, all of the independent variables studied were statistically significant. This research confirms that Google searches (Bitcoin_google) and energy costs (PRECO) positively impact bitcoin market values. Google searches for cryptocurrency speculating are common. The price of bitcoin transactions is inversely proportional to Treasury Note Yield Index at closing (Close). This leads us to believe that the price of bitcoin has an inverse relationship with investment results. Additionally, we can also state that the weight of the google search cardinality of bitcoin has the highest weight for the regression among the tree, followed by Treasury Note Yield Index at closing and energy price respectively. This is the case since the module of negative impact is higher. This means that the same yield and energy price variation do not weigh one another.

OLS Regression Results

Dep. Variable:	market-price-log	R-squared:	0.868
Model:	OLS	Adj. R-squared:	0.868
Method:	Least Squares	F-statistic:	360.9
Date:	Sun, 27 Feb 2022	Prob (F-statistic):	1.00e-156
Time:	02:17:53	Log-Likelihood:	-1525.4
No. Observations:	961	AIC:	3059.
Df Residuals:	957	BIC:	3078.
Df Model:	3		
Covariance Type:	HAC		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.6810	0.331	5.080	0.000	1.032	2.330
PRECO-log	0.2496	0.081	3.073	0.002	0.090	0.409
Close-log	-0.7252	0.154	-4.707	0.000	-1.027	-0.423
Bitcoin_google-log	2.2474	0.072	31.159	0.000	2.106	2.389

Figure 5. OLS Regression Results

As referred previously, our second objective consists of identifying the better algorithm for forecasting tasks. Several algorithms were employed to identify the one that could have better-predicting power. We used the scikit-learn [32] To split the sample into training and test subsamples. Then, the following regressors were estimated: linear Ordinary least squares (OLS), Ridge regression, Lasso regression, Bayesian ridge, gradient boosting, multi-layer perceptron regressor, and random Forest.

As we can see in the table above, the Gradient Boosting had by far the best performance in predicting power. Achieving a Correlation (R2) and Explained Variance Scores of 0.962 and also the lowest Mean Square, Average and Median Errors (MSE and MAE and MdAE) as well. This shows us that it may be a viable candidate for future predictions using these kinds of data. Random Forest also has a good performance. This allows to verify that ensemble methods outperformed, which is in line with similar research in other fields.

TABLE I. TABLE TYPE STYLES

Model	EVS	MAE	MSE	MdAE	R2
Linear	0.778	4167.83	50032942.2	2467.3	0.776
Ridge	0.778	4165.61	50036560.9	2462.8	0.776
Lasso	0.778	4167.64	50033294.7	2466.3	0.776
Bayesian Ridge	0.777	4151.52	50063279.9	2435.9	0.776
Gradient Boosting	0.962	1221.71	8591626.4	189.9	0.962
MLP Regressor	0.802	3601.63	45170620.8	1466.4	0.798
Random Forest	0.921	1835.65	17717700.3	256.0	0.921

The results show us that not only there is a strong relationship between the variables analyzed, with strong prediction. To further understand this, we could devise a multiagent system and simulate the demand for bitcoin in different scenarios [40].

V. CONCLUSIONS

The purpose of this research is to figure out why bitcoins are so expensive in accordance with a few variables. In particular, we looked at the impact of Web search data, energy costs, and alternative investments on bitcoin pricing. Finding the algorithm with the highest prediction capability was the second aim. To do so, we employed a number of machine learning methods. The OLS regression was employed to achieve the first aim. All the factors studied were found to be significant. This research confirms that Google searches and energy costs have a beneficial influence on bitcoin market values. Cryptocurrency speculation is the subject of Google searches. The cost of bitcoin transactions is correlated with the price of electricity. On the other hand, investment returns are inversely connected to the price of bitcoins. This might be linked to other forms of investments changing their preferences when returns are too low. The Gradient Boosting had the best performing regression model on the present data, leaving us with a 0.962 on both Explained Variance Score and Correlation. In terms of future study, it would be fascinating to investigate the impact of sentiment concerning material provided over the internet to the bitcoin pricing as an addition to the current model.

ACKNOWLEDGMENT

We gratefully acknowledge financial support from FCT - Fundação para a Ciência e a Tecnologia (Portugal), national funding through research grant UIDB/04521/2020..

REFERENCES

[1] S. Aparicio, J. T. Aparicio and C. J. Costa, "Data Science and AI: Trends Analysis," 2019 14th Iberian Conference on Information Systems and Technologies (CISTI), 2019, pp. 1-6, doi: 10.23919/CISTI.2019.8760820.

[2] A. Samuel. "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal of Research and Development. 3 (3), 1959 pp. 210–229. CiteSeerX 10.1.1.368.2254. doi:10.1147/rd.33.0210

[3] C. Albon, C. Machine Learning with Python Cookbook : Practical Solutions from Preprocessing to Deep Learning, O'Reilly Media, Inc, 2018.

[4] E. Alpaydin, Introduction to Machine Learning (3rd ed.). The MIT Pres, 2014.

[5] A. Müller, & S. Guido, S. Introduction to Machine Learning with Python. O'Reilly Media, Inc. 2017.

[6] J. D. Kelleher, B. Mac Namee, & A. D'Arcy, A. Fundamentals of Machine Learning for Predictive Data Analytics. MIT Press, 2015

[7] T. M. Mitchell. Machine Learning (1st ed.). McGraw-Hill, 1997.

[8] C. J. Costa and J. T. Aparicio, "POST-DS: A Methodology to Boost Data Science," 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), 2020, pp. 1-6, doi: 10.23919/CISTI49556.2020.9140932.

[9] C. Costa and J. T. Aparicio J.T. A Methodology to Boost Data Science in the Context of COVID-19. In: Arabnia H.R. et al. (eds) Advances in Parallel & Distributed Processing, and Applications. Transactions on Computational Science and Computational Intelligence. Springer, Cham. 2021, https://doi.org/10.1007/978-3-030-69984-0_7

[10] M. Amjad and D. Shah. Trading bitcoin and online time series prediction. In *NIPS 2016 time series workshop* , 2017, pp. 1-15). PMLR.

[11] A. Azari, Bitcoin price prediction: An ARIMA approach. *arXiv preprint arXiv:1904.05315,* 2019.

[12] A. Clauset, D. Larremore, and R. Sinatra. Data-driven predictions in the science of science. *Science*, *355*(6324), 2017, pp. 477-480.

[13] L. Y. Chang and W. Chen. Data mining of tree-based models to analyze freeway accident frequency. *Journal of safety research*, *36*(4), 2005, pp. 365-375.

[14] Z. Chen, C. Li, and W. Sun. Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, *365*, 112395, 2020.

[15] F. Akba, I. Medeni, M. Guzel, and I. Askerzade. Assessment of iterative semi-supervised feature selection learning for sentiment analyses: Digital currency markets. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)* . IEEE. 2020, pp. 459-463

[16] G. Arya, K. Harika, D. Rahul, S. Narasimhan, and A. Ashok. Analysis of Unsupervised Learning Algorithms for Anomaly Mining with Bitcoin. In *Machine Intelligence and Smart Systems*, Springer, Singapore, 2021, pp. 365-373.

[17] C. Costa, M. Aparicio, and J. Aparicio. Sentiment Analysis of Portuguese Political Parties Communication. In *The 39th ACM International Conference on Design of Communication*, 2021, pp. 63-69.

[18] Aparicio, J. T., de Sequeira, J. S., & Costa, C. J. (2021, June). Emotion analysis of Portuguese Political Parties Communication over the covid-19 Pandemic. In *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-6). IEEE.

[19] K. Lee, S. Ulkuatam, P. Beling, and W. Scherer. Generating synthetic bitcoin transactions and predicting market price movement via inverse reinforcement learning and agent-based modeling. *Journal of Artificial Societies and Social Simulation*, *21*(3). 2018.

[20] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[21] B. Süssmuth. The mutual predictability of Bitcoin and web search dynamics. *Journal of Forecasting*. 2021

[22] S. Küfeoğlu, and M. Özkuran. Bitcoin mining: A global review of energy and power demand. *Energy Research & Social Science*, *58*, 101273, 2019.

[23] D. Koutmos. Market risk and Bitcoin returns. *Annals of Operations Research*, *294*(1), 2020, pp. 453-477.

[24] R. Wirth and J. Hipp, J. CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Vol. 1, 2000, pp. 29-40.

[25] D. Baur, A. Lee, and K. Hong. Bitcoin: currency or investment?. *Available at SSRN*, *2561183,* 2015.

[26] S. Corbet, C. Larkin, B. Lucey, A. Meegan, and L. Yarovaya. The impact of macroeconomic news on bitcoin returns. *The European Journal of Finance*, *26*(14), 2020, pp 1396-1416.

[27] A. Dutta, S Kumar, asn M. Basu. A gated recurrent unit approach to bitcoin price prediction. *Journal of Risk and Financial Management*, *13*(2), 23, 2020

[28] "Electricity price statistics." https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Electricity_price_statistics (accessed Feb. 25, 2022).

[29] S. Seabold, and J. Perktold. "statsmodels: Econometric and statistical modeling with python." Proceedings of the 9th Python in Science Conference. 2010.

[30] "Introduction statsmodels." https://www.statsmodels.org/stable/ (accessed Feb. 25, 2022).

[31] W. McKinney "Data structures for statistical computing in python." In Proceedings of the 9th Python in Science Conference, vol. 445, no. 1, 2010, pp. 51-56.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2011. pp.2825-2830.

[33] V.Mahajan, A. Jain, & M. Bergier,. Parameter estimation in marketing models in the presence of multicollinearity: An application of ridge regression. *Journal of Marketing Research*, *14*(4), 1977, 586-591.

[34] A. B. Owen A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, *443*(7),2007, 59-72.

[35] F.A. da Silva, A.P. Viana, C Correa, E. Santos, J. de Oliveira, . ., Andrade & L. Glória, Bayesian ridge regression shows the best fit for SSR markers in Psidium guajava among Bayesian models. *Scientific Reports*, *11*(1), 2021, 1-11.

[36] Z. Zhang, Y. Zhao, A. Canes, D. Steinberg, & O. Lyashevska, O. Predictive analytics with gradient boosting in clinical medicine. *Annals of translational medicine*, *7*(7).2019

[37] T. Deepika, & P. Prakash Power consumption prediction in cloud data center using machine learning. *Int. J. Electr. Comput. Eng.(IJECE)*, *10*(2), 2020, 1524-1532.

[38] G. Biau, & E. Scornet. A random forest guided tour. *Test*, *25*(2), 2016, 197-227.

[39] A. Sharma, D. Bhuriya, & U. Singh, (2017, April). Survey of stock market prediction using machine learning approach. In *2017 International conference of electronics, communication and aerospace technology (ICECA)*, Vol. 2, 2017, pp. 506-509

[40] Aparicio, J. T., Trinca, M., Castro, D., & Henriques, R. (2021, June). Vehicle Smart Grid Allocation using Multi-Agent Systems sensitive to Irrational Behavior and Unstable Power Supply. In 2021 16th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-6). IEEE.

ISBN: 978-989-33-3436-2