Machine Learning applications in Call Option valuation: a case study of Crypto and Brazilian Equities markets

José Moraes de Albuquerque Neto[1]

[1]Department of Mathematics, Michigan State University, East Lansing, MI, USA

Abstract

This study investigates the use of machine learning models for pricing European-style call options in the Bitcoin and Brazilian equity markets. Traditional pricing models like Black-Scholes, while foundational, often fall short in dynamic and non-linear market conditions. By applying models such as Multi-Layer Perceptron (MLP), XGBoost, and Neural Stochastic Differential Equations (Neural SDE), this research demonstrates the potential of data-driven approaches to capture complex pricing dynamics. Results indicate that segmenting data by moneyness improves predictive performance, particularly for XGBoost models. Future work will extend this analysis to put options and American-style options while exploring model stacking techniques for enhanced accuracy. This study underscores the importance of machine learning in modern financial modeling, offering insights into adaptive pricing strategies across diverse markets.

Keywords: Machine Learning, Option Pricing, Neural Networks, XGBoost, European Options, Bitcoin, Brazilian Equities, Model Stacking, Moneyness.

1 Introduction

Options are an enormous component of global financial markets, with trading volumes exceeding 108 billion contracts in 2023 [1]. These financial derivatives grant holders the right, but not the obligation, to buy (call) or sell (put) an underlying asset. European options restrict this right to the expiration date, whereas American options allow exercise at any time up to the expiration date.

Traditional models like Black-Scholes have been instrumental in shaping our understanding of option pricing. However, their reliance on restrictive assumptions, such as constant volatility and lognormal price distribution, limits their effectiveness in dynamic and volatile markets, particularly in emerging sectors such as cryptocurrencies [2].

Advancements in machine learning have introduced new opportunities to enhance option pricing by leveraging market data to model intricate and nonlinear relationships. This research focuses on applying machine learning techniques to predict the negotiated prices of call options in the Bitcoin and Brazilian equity markets. The study incorporates market-specific features, such as implied volatility and moneyness, to capture nuanced pricing dynamics. By benchmarking machine learning models against traditional approaches, this work highlights the potential for robust and adaptive option pricing strategies that align more closely with real-world market behaviors.

2 Related Works

The field of option pricing has been profoundly shaped by the foundational work of Black and Scholes (1973), which introduced a closed-form solution for pricing European options based on a parabolic partial differential equation. Despite its influence, the Black-Scholes model is constrained by several limiting assumptions, including constant volatility, frictionless markets, and the geometric Brownian motion of underlying asset prices. These limitations have sparked widespread research into alternative methodologies that address real-world complexities.

Extensions to the Black-Scholes framework, such as Merton's jump-diffusion model [4], and Heston's stochastic volatility model [5], have attempted to capture market phenomena like the volatility smile. However, these parametric approaches often depend heavily on the precise specification of the underlying stochastic process, which makes them sensitive to model errors. As a result, researchers have turned to non-parametric methods, particularly machine learning, to bypass these constraints.

The application of neural networks in option pricing has emerged as a transformative approach. Early studies by Malliaris and Salchenberger [6] demonstrated the ability of neural networks to outperform Black-Scholes in specific cases, such as pricing out-of-the-money options. Based on this, Hutchinson, Lo, and Poggio [7] introduced neural networks trained on historical market data, showing that these models could learn the option pricing function with high accuracy. This nonparametric approach has proven effective in adapting to diverse market conditions without relying on strict parametric assumptions.

Recent advances in deep learning have further elevated the potential of neural networks in option pricing. Stark [8] compared deep feedforward networks with Black-Scholes for pricing DAX 30 index options, finding that the neural networks consistently outperformed traditional methods, particularly for medium- and long-term maturities. Similarly, Trønnes [9] applied reinforcement learning to price European options, using a neural network trained on synthetic and real-world data. This approach demonstrated superior accuracy while addressing the limitations of traditional no-arbitrage models.

In emerging markets, where volatility and dynamics often deviate significantly from standard assumptions, neural networks have shown exceptional promise. Santos and Espínola Ferreira [10] used multi-layer perceptron (MLP) networks to solve the Black-Scholes equation for Brazilian equity options, achieving improved short-term price forecasts. Their work illustrates how neural networks can adapt to region-specific market characteristics. Ke and Yang [11] further explored this potential by implementing deep learning architectures, including multi-task learning for bid/ask price predictions, demonstrating the ability of neural networks to handle both liquid and illiquid options effectively.

Building on these advances, this research applies neural networks to predict negotiated prices in Bitcoin and Brazilian equity markets, aiming to create a robust framework that bridges theoretical innovation with practical market realities.

3 Dataset and Features

Three distinct datasets were employed to develop and evaluate machine learning models for option pricing, each selected to capture unique market characteristics and assess predictive performance across different financial environments. The study focused on Bitcoin and Brazilian equity markets, while a third dataset, initially considered, was excluded due to quality issues.

The Brazilian options datasets for PETR4 and VALE3 were sourced from OpLab, a platform recommended by colleagues at XP Inc. specializing in financial market data. The benchmark interest rate for these options was the SELIC rate (*Sistema Especial de Liquidação e de Custódia*), set by the Brazilian Central Bank [13].

For the Bitcoin options dataset, data was obtained from OptionsDX, which provides comprehensive Bitcoin option chains and market data for call options. Given the global nature of Bitcoin trading, the one-year U.S. Treasury yield, obtained from the Federal Reserve Economic Data (FRED) database, was applied as the risk-free rate [15].

A fourth dataset from Wharton Research Data Services (WRDS), part of the IvyDB Europe database,

contained an option chain for Adidas from 2013. However, due to significant data quality issues—such as missing values and inconsistencies—this dataset was excluded after initial exploratory analysis, despite imputation efforts [16].

Across all datasets, key features were extracted for modeling: the underlying asset price, strike price, time to expiration (days to expiry), moneyness, and implied volatility. The target variable, the negotiated price, represented the mid-market price of the options. This feature set was chosen to capture the intricate relationships influencing option prices across different markets and contexts.

4 Methods

4.1 Methods for BTC Options

The prediction of Bitcoin option prices was investigated using three machine learning approaches: a Multi-Layer Perceptron (MLP), a Neural Stochastic Differential Equation (Neural SDE), and an XGBoost Regressor. The models were trained on a dataset of European-style Bitcoin call options using features such as the underlying price, moneyness, implied volatility, and days to expiration (DTE). Each method leverages distinct strengths to address the complex dynamics of option pricing, as described below.

4.1.1 Multi-Layer Perceptron (MLP)

The MLP model was implemented using the default architecture provided by the `MLPRegressor` class in `sklearn`. The model was trained to minimize the mean squared error (MSE) loss function using the Adam optimizer with a learning rate of 0.001. Training was conducted over a maximum of 200 iterations.

Hyperparameter tuning was conducted through grid search, optimizing parameters such as the learning rate, activation functions (ReLU and tanh), and hidden layer configurations. Regularization techniques, such as dropout or weight decay, were not applied. This decision was based on preliminary experiments that indicated reduced performance when regularization was used, likely due to the high sensitivity of option prices to all input features.

4.1.2 Neural Stochastic Differential Equation (Neural SDE)

The Neural SDE model incorporates a stochastic framework to model the random dynamics underlying Bitcoin option prices. The model is governed by the stochastic differential equation:

$$dX_t = f(X_t, t; \theta)dt + g(X_t, t; \phi)dW_t,$$

where X_t represents the state variable, such as the option price or input features. The drift term $f(X_t, t; \theta)$, parameterized by a neural network with parameters θ, captures the deterministic trends, while the diffusion term $g(X_t, t; \phi)$, parameterized by a separate neural network with parameters ϕ, models stochastic fluctuations. The stochastic component W_t represents a standard Brownian motion.

The drift and diffusion networks each consist of three fully connected layers with 256 neurons per layer and employ ELU activation functions to ensure stable gradient propagation during training. The model minimizes the MSE loss using the Adam optimizer with a learning rate of 0.001.

Hyperparameter tuning was performed through grid search, optimizing hidden layer dimensions, batch sizes, and learning rates. Training was conducted over 50 epochs for the overall dataset, with segment-specific models for in-the-money (ITM), at-the-money (ATM), and out-of-the-money (OTM) options trained for 100 epochs to account for distinct pricing dynamics in these categories.

4.1.3 XGBoost Regressor

The XGBoost Regressor models option prices using gradient-boosted decision trees, combining boosting with regularization to handle complex, non-linear relationships in the dataset. The model minimizes the squared error loss function.

Dimensionality reduction using Principal Component Analysis (PCA) retained 95% of variance, mitigating multicollinearity and stabilizing the model. Stratified sampling ensured balanced representation across moneyness categories (ITM, ATM, OTM). Hyperparameter tuning was performed in two stages: an initial randomized search explored tree depth, learning rates, and subsampling ratios, while Bayesian optimization via Optuna refined these parameters.

4.2 Methods for Brazilian Options

This section presents the methodologies for predicting the prices of European-style call options on PETR4 and VALE3 using Multi-Layer Perceptron (MLP), Neural Stochastic Differential Equation (Neural SDE), and XGBoost Regressor. The input features included spot price, strike price, days to maturity, implied volatility, and selected Greeks (rho, theta, gamma). The target variable was the option premium. Preprocessing and modeling approaches were tailored for each equity.

4.2.1 Multi-Layer Perceptron (MLP)

The MLP models were implemented using `sklearn`'s `MLPRegressor`, trained on scaled features with the premium as the target variable. The features were standardized using `StandardScaler` or `RobustScaler`, selected based on the distribution of the training data.

The training process minimized mean squared error (MSE) using the Adam optimizer. Hyperparameters such as the learning rate, activation function (ReLU or tanh), and the number of iterations (up to 500) were optimized through grid search. Regularization techniques, such as dropout or weight decay, were not applied, as the experiments showed that early stopping was sufficient to control overfitting. Premiums were capped at the 1st and 99th percentiles to mitigate the impact of extreme values.

4.2.2 Neural Stochastic Differential Equation (Neural SDE)

The Neural SDE models parameterized the drift and diffusion terms using neural networks with three fully connected layers of 256 neurons and ELU activation functions. The models were trained to minimize MSE using the Adam optimizer with a learning rate of 0.001. T

Segmentation by moneyness categories (in-the-money, at-the-money, out-of-the-money) was performed to create specialized models for each category. This segmentation intended the unique sensitivities of options within each group, such as increased volatility dependence for out-of-the-money options. Training was conducted for 50 epochs on the full dataset and 100 epochs for segmented datasets.

4.2.3 XGBoost Regressor

The XGBoost models utilized gradient-boosted decision trees. Input features were scaled using `RobustScaler`, and segmentation by moneyness categories was applied. A log transformation $\log(1 + \text{premium})$ was used to stabilize the model for small premium values, as indicated by initial experiments showing instability with raw premium data.

Hyperparameter tuning was conducted in two stages. A randomized search explored maximum depth (3 to 10), learning rate (0.01 to 0.3), and the number of estimators (100 to 500). Optuna Bayesian optimization refined the parameters further.

5 Results and Discussion

5.1 Bitcoin Data

Table 1: Model Performance on the entire Bitcoin Data

Model	RMSE	MAE	R^2	PE5 (%)	PE10 (%)	PE20 (%)
MLP	232.55	141.46	0.9995	54.57	66.05	74.05
XGBoost	594.01	311.13	0.9971	45.16	56.87	67.28
Neural SDE	469.03	156.45	0.9982	58.17	67.93	73.44

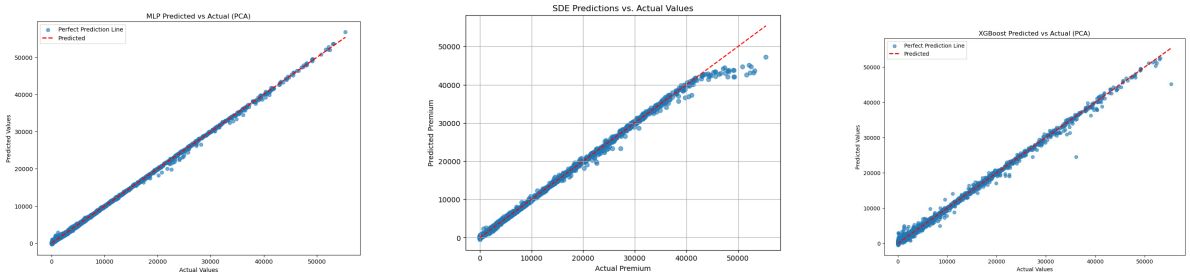

Figure 1: Comparison of MLP, SDE, and XGB models for BTC options pricing.

The MLP model stands out with the highest accuracy, leveraging its ability to learn complex patterns effectively without signs of overfitting, to PCA-reduced features. This result is in agreement with Ke and Yang [11].

Table 2: Model Performance on ITM Segment

Model	RMSE	MAE	R^2	PE5 (%)	PE10 (%)	PE20 (%)
XGBoost	699.72	276.57	0.9963	94.93	98.38	99.80
SDE	11428.78	9271.44	0.0138	5.98	10.94	21.99
MLP	3805.73	2472.74	0.8903	27.46	54.00	83.18

Table 3: Model Performance on ATM Segment

Model	RMSE	MAE	R^2	PE5 (%)	PE10 (%)	PE20 (%)
XGBoost	126.31	66.30	0.9981	68.99	81.80	90.11
SDE	2874.93	2212.70	0.0198	2.70	8.09	14.83
MLP	282.79	184.65	0.9906	39.78	53.71	67.19

Table 4: Model Performance on OTM Segment

Model	RMSE	MAE	R^2	PE5 (%)	PE10 (%)	PE20 (%)
XGBoost	87.01	30.18	0.9983	56.56	69.94	78.15
SDE	2114.96	1470.85	0.0208	1.19	2.46	6.44
MLP	75.27	39.39	0.9988	48.43	63.17	72.82

Based on these tables, it can be observed that the Neural SDE Model significantly underperforms when applied to these subsets of data, warranting further investigation. Additionally, all models show a decline in performance when fitting the ITM segment. However, both the MLP and XG-Boost models perform better with ATM and OTM data.

Further investigation will be conducted to understand the ”why” behind this behavior.

5.2 Brazilian Equities Data

Table 5: Model Performance for PETR4 and VALE3 (RMSE Metrics)

Model (Equity)	RMSE	MAE	R^2
XGB PETR4	0.6645	0.3016	0.9763
MLP PETR4	0.2538	0.0856	0.9963
NeuralSDE PETR4	0.2478	0.1184	0.9965
BS PETR4	0.0939	0.0180	
XGB VALE3	1.3906	0.6671	0.9659
NeuralSDE VALE3	0.6812	0.2574	0.9917
MLP VALE3	0.8381	0.2398	0.9874
BS VALE3	0.4333	0.0497	

From this, it then begs the question of the impact segmenting the model would have. After segmenting the model, we get the metrics:

Table 6: Model Performance for PETR4 After Segmentation

Model (Segment)	RMSE	MAE	R^2	MAPE (%)	PE5 (%)	PE10 (%)	PE20 (%)
BS ITM	0.1607		0.9985	57.80			
BS ATM	0.0001		1.0000	0.05			
BS OTM	0.0000		1.0000	0.00			
XGB ITM	0.0528		0.9998	0.59	94.93	98.38	99.80
XGB ATM	0.0032		1.0000	0.14	68.99	81.80	90.11
XGB OTM	0.0114		0.9998	8.93	56.56	69.94	78.15
MLP ITM	2.3979	1.6878	0.9039	32.03	32.03	57.03	82.03
MLP ATM	0.7924	0.6284	0.8970	20.22	20.22	37.08	67.42
MLP OTM	0.6205	0.3852	0.8419	6.09	6.09	15.23	24.37
SDE ITM	0.5680	0.4459	0.9817	48.79	48.79	77.78	96.62
SDE ATM	0.4233	0.3229	0.7537	14.67	14.67	34.67	62.67
SDE OTM	0.2576	0.1776	0.9031	4.95	4.95	8.98	16.10

A similar behavior is seen in the VALE3 data after segmentation. Asw we saw, overall the XGBoost model improves substantially and perform all of the other models when we segmented the data of these two Brazilian Equities.

6 Conclusion

This study explored the application of machine learning models to predict the negotiated prices of call options across Bitcoin and Brazilian equity markets. The analysis revealed that while models like the Multi-Layer Perceptron (MLP) and XGBoost exhibited robust predictive capabilities, their performance varied across market segments such as in-the-money (ITM), at-the-money (ATM), and out-of-the-money (OTM) options. Segmenting the data by moneyness improved model performance, particularly for the XGBoost model on Brazilian Equities, highlighting the benefits of tailored modeling approaches.

Future research will focus on extending this analysis to put options, enabling a comprehensive understanding of machine learning applications in option pricing. Additionally, investigating model stacking techniques—such as combining Random Forest with XGBoost and MLP—could enhance predictive accuracy by leveraging the strengths of each model.

It would also be valuable to evaluate model performance on American-style options, where early exercise introduces additional complexities not present in European options. This investigation will provide insights into the adaptability of machine learning techniques to a broader range of derivative instruments and trading conditions.

References

[1] FIA. Global futures and options volume hits record 137 billion contracts in 2023. *FIA.org* (2023). Available at: https://www.fia.org/fia/articles/global-futures-and-options-volume-hits-record-137-billion-contracts-2023. Accessed in December 02, 2024

[2] Black, F., & Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81(3), 637-654

[3] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *Review of Financial Studies*, 33(5), 2223-2273

[4] Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1-2), 125-144

[5] Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2), 327-343

[6] Malliaris, M., & Salchenberger, L. M. (1993). Beating the best: A neural network challenges the Black-Scholes formula. *Proceedings of the IEEE/IAFE 1993 Computational Intelligence for Financial Engineering (CIFEr)*, 348-353

[7] Hutchinson, J. M., Lo, A. W., & Poggio, T. (1994). A nonparametric approach to pricing and hedging derivative securities via learning networks. *The Journal of Finance*, 49(3), 851-889

[8] Stark, L. (2017). Machine Learning and Options Pricing: A Comparison of Black-Scholes and a Deep Neural Network in Pricing and Hedging DAX 30 Index Options. *Master's Thesis, Aalto University School of Business*

[9] Trønnes, H. A. (2018). Pricing Options with an Artificial Neural Network: A Reinforcement Learning Approach. *Master's Thesis, Norwegian University of Science and Technology*

[10] Santos, D. d. S., & Espínola Ferreira, T. A. (2024). Neural network learning of Black-Scholes equation for option pricing. *arXiv preprint arXiv:2405.05780*

[11] Ke, A., & Yang, A. (2019). Option pricing with deep learning. *CS230: Deep Learning, Fall 2019, Stanford University*

[13] Banco Central do Brasil. Taxa SELIC - Sistema Especial de Liquidação e de Custódia. Available at: https://www.bcb.gov.br/controleinflacao/historicotaxasjuros. Accessed in December 01, 2024.

[14] OptionsDX. Bitcoin Option Chains. Available at: https://www.optionsdx.com/product/btc-option-chains/. Accessed in November 28, 2024.

[15] U.S. Department of the Treasury. Daily Treasury Yield Curve Rates. Available at: https://home.treasury.gov/. Accessed in December 01, 2024.

[16] Wharton Research Data Services (WRDS). IvyDB Europe Dataset. Available at: https://wrds-www.wharton.upenn.edu/. Accessed in December 02, 2024.