International Journal of Finance and Investment
ISSN: 3078-6002 (Print), ISSN: 3078-2295 (Online) | Vol. 5, No. 2, 2026

Mispricing Detection in the S&P 500 ETF Options Market Based on an XGBoost Model

Chenyi Wu

School of Mathematics, Southwestern University of Finance and Economics, Chengdu, Sichuan, 611130, China
cywu003@gmail.com

Abstract: Based on high-frequency data for S&P 500 index options (SPX) obtained from the WRDS database, this paper builds an empirical framework for identifying option pricing mismatches. First, taking the Black–Scholes model as the theoretical pricing benchmark and combining it with actual market quotes, we introduce a relative mispricing measure and adopt a quantile standardization strategy to construct mispricing labels, which are divided into three categories: overpriced, underpriced, and fairly priced. Then, we construct a mispricing identification model using an XGBoost multiclass classifier, trained and predicted with multidimensional features including contract attributes, trading indicators, and volatility. In data preprocessing, we clean both option and underlying information and strictly align them using a dual primary key of trading date and contract identifier to ensure the accuracy and consistency of label construction. The empirical results show that the model achieves good identification performance in the out-of-sample test period (2023), with an overall accuracy of 77.6%. The precision and recall for the underpriced category are particularly strong, indicating that the model can effectively identify pricing deviations with potential arbitrage value. At the same time, the results reflect the existence of certain structural inefficiency in the options market. Although the Black–Scholes framework relies on idealized assumptions, it still has important reference value in mispricing identification. This study not only verifies the feasibility of using machine learning to identify option mispricing, but also provides a methodological foundation for subsequent quantitative trading strategy design.

Keywords: Option pricing, mispricing identification, XGBoost, machine learning.

1. Introduction

1.1. Research Background

With the continuous development and maturation of financial derivatives markets, options have become an indispensable tool in the modern financial system, widely used in risk management, asset allocation, and the construction of derivative strategies. However, in actual trading, option prices often exhibit significant deviations from their theoretical values, i.e., "mispricing." This phenomenon may result from multiple factors, including biases in volatility estimation, sentiment-driven trading, and the failure of model assumptions to hold [1].

Although a substantial body of literature has sought to improve option pricing models (e.g., by introducing stochastic volatility, jump–diffusion, or GARCH processes) to reduce theoretical errors, research on the structural identification and modeling of mispricing behavior itself remains extremely limited [2]. In particular, within the Chinese academic community, there is almost no systematic research that treats mispricing as an independent object and builds supervised learning models for quantitative identification and prediction.

Meanwhile, the application of machine learning in financial modeling has become increasingly widespread, demonstrating strong fitting capabilities for complex nonlinear structures. Statistical learning methods represented by XGBoost have achieved remarkable results in fields such as credit scoring, fraud detection, and market volatility forecasting [3]. However, research on applying these methods to the identification and classification modeling of option pricing errors remains blank.

1.2. Research Significance

As a key derivatives market within the financial system, the pricing efficiency of the options market directly affects asset allocation, risk management, and market stability. How to effectively identify and quantify mispricing behavior in the market has become an important direction at the intersection of financial engineering and behavioral finance [4].

In terms of theoretical significance, this study steps beyond the traditional research path of model modification and takes "identifying mispricing" as the core objective, proposing a new approach of classification identification using statistical learning methods. The study employs the XGBoost model and, within a supervised learning framework, constructs labels based on deviations between historical trading data and Black–Scholes theoretical prices to identify option samples that are overpriced or underpriced. This data-driven approach to mispricing identification does not rely on complex mathematical modeling assumptions and is better suited to the high-frequency, high-dimensional, and nonlinear characteristics of modern financial markets, with strong practical applicability.

In terms of practical significance, the mispricing identification system constructed in this study can be used for trading strategy generation, helping investors filter potentially overpriced or underpriced options, thereby realizing arbitrage opportunity identification and risk control. Moreover, identifying and quantifying deviations in option pricing helps market participants understand option price formation mechanisms and enhances the rationality and transparency of market pricing. The findings can also be embedded in option trading platforms, robo-advisory systems, or risk-control engines to implement pricing-signal detection and dynamic risk alerts, provide quantitative tool support for regulatory

authorities to detect market irregularities and identify potential systemic risks, and help formulate more scientific policy response mechanisms.

Therefore, this study possesses a certain degree of innovation and practical relevance in its research object, methodology, and objectives, achieving an effective integration of theoretical exploration and practical application while maintaining rigor.

1.3. Research Objectives and Structure

The overall objective of this study is to construct a mispricing identification framework for options based on statistical learning models, with a specific focus on the S&P 500 ETF options market. By introducing the XGBoost algorithm and combining historical option trading data with deviations from Black–Scholes theoretical pricing, the model is trained to capture high-probability mispricing samples and to analyze their feature structures and distributional patterns. The research not only mines characteristics of market inefficiency from a data-driven perspective but also aims to provide practical support for intelligent quantitative trading systems and market regulation.

The structure of the paper is as follows: Chapter 1 is the introduction, outlining the research background, significance, and objectives; Chapter 2 presents the literature review and theoretical foundations, summarizing advances in option mispricing research and statistical learning methods; Chapter 3 describes the research design and methodology, including data sources, label construction, and model-building procedures; Chapter 4 reports empirical results and analysis; Chapter 5 concludes with research findings and offers practical recommendations and directions for future research.

2. Literature Review and Theoretical Foundations

This chapter reviews research related to option pricing and mispricing identification, introduces the application of statistical learning methods in the financial domain, clarifies the positioning and innovations of this study, and specifies its theoretical foundations and research value.

2.1. Option Mispricing Phenomena and Research Progress

Option pricing has long been a core issue in financial engineering. The Black–Scholes model proposes a classic pricing formula but is built on a series of idealized assumptions—such as constant volatility and no transaction costs —which leads to certain deviations in real markets [5]. Subsequent studies have improved upon it by introducing jump–diffusion models and volatility dynamics modeling, yet pricing errors remain widespread in the market [6, 7].

Option mispricing refers to the systematic deviation between actual market prices and theoretical model prices, a deviation that reflects factors such as incomplete market information, investor sentiment fluctuations, and liquidity issues. Existing research shows that during periods of high volatility or severe market turbulence, mispricing becomes more pronounced, which also creates possibilities for arbitrage strategies. Mispricing occurs particularly frequently in deep out-of-the-money options or short-maturity options [8].

In recent years, an increasing number of studies have shifted focus from merely improving pricing models to identifying mispricing itself. This shift has also propelled the application of machine learning methods in finance. Statistical learning methods represented by XGBoost can handle large-scale, high-dimensional, and nonlinear financial data, showing substantial potential in identifying regularities in pricing errors. Some literature points out that due to its efficiency and strong interpretability, the XGBoost algorithm has been widely applied to prediction tasks in the financial field [3].

2.2. Applications of Statistical Learning Methods in Financial Research

With the rapid growth of financial market data, traditional econometric models have gradually revealed limitations in handling high-dimensional and nonlinear data. To better extract patterns from complex market information, statistical learning methods—especially machine learning techniques—have been widely introduced into financial research in recent years.

In asset pricing, credit risk prediction, quantitative investing, and other areas, researchers employ methods such as support vector machines, random forests, and neural networks to address nonlinear structures that traditional models struggle to capture. Among these, machine learning methods have attracted growing attention. Gu et al. (2020) find that machine learning methods can achieve better predictive performance when dealing with large financial samples, and are particularly well-suited to financial markets characterized by numerous features and complex structures [2].

Specifically for options markets, some studies have used machine learning for implied volatility forecasting or option price regression, with most research focusing on improving pricing models and narrowing the gap between theoretical and market prices. For example, many studies enhance pricing accuracy by introducing jump components, volatility structures, or nonparametric methods [9].

2.3. Positioning and Innovations of This Study

Research on the structural identification of pricing deviations themselves—especially research that achieves automated, market-level identification and modeling—remains relatively scarce. Mispricing identification involves not only judging model errors but also constructing a reasonable labeling system that incorporates market behavior, which makes traditional regression-based approaches ill-suited to the task. The classification capabilities of XGBoost are well suited to learning and identifying mispricing labels. Therefore, employing this model to identify mispricing samples in the S&P 500 ETF options market is not only technically feasible but also reflects a degree of innovation and practicality in research methodology.

This study takes the S&P 500 ETF options market as its empirical setting and proposes to use the error between Black–Scholes model prices and market prices as mispricing labels, combined with an XGBoost classification model for estimation. This approach differs from traditional pricing-improvement lines of research and from past arbitrage-strategy constructions based on subjective partitions. It emphasizes identifying and understanding mispricing behavior through a data-driven approach, offering good extensibility and practical potential.

3. Research Design and Methods

This study draws on the index options database and companion security-level data provided by the Wharton Research Data Services (WRDS), constructs relative mispricing labels with reference to the Black–Scholes (BS) theoretical price, and trains a gradient boosting tree model to classify option prices as overvalued/reasonably priced/undervalued, followed by out-of-sample validation [10]. The sample spans 2020–2023, with 2020–2022 used as the training period and 2023 as the test period.

3.1. Data Sources and Processing

The raw data comprise two parts: (i) an option daily (tick-by-tick) quote table (denoted opprcd), including trading date, contract identifier (secid), call/put flag, strike, bid, ask, volume, and open interest; and (ii) a security-level reference table (denoted secprd), containing implied volatility, time to maturity (in days), and premium for the same contract on the same trading day. To ensure one-to-one correspondence at the contract level across the two tables, we first harmonize field definitions and value scales within each dataset (e.g., automatically converting percentage-form implied volatility to decimals), and then perform a strict inner join using "trading date (date) × contract identifier (secid)" as the composite primary key [11]. During cleaning, we apply robust processing to key variables: standardize the call/put flag to {C, P}, and compute the mid-quote and the relative bid–ask spread [9].

$$\text{mid} = \frac{\text{bid}+\text{ask}}{2}, \text{rel\_spread} = \frac{\text{ask}-\text{bid}}{\text{mid}} \tag{1}$$

and filter abnormal observations under liquidity and quality constraints (e.g., bid, ask, mid>0, rel_spread≤0.5). The time to maturity is taken from the field *days*; if missing, it is estimated by the difference between the expiration date and the trading date, with a one-day lower bound imposed to avoid numerical instability:

$$\text{t} = \frac{\max(\text{days},1)}{365} \tag{2}$$

To reduce the impact of extreme values on subsequent computations, we apply mild winsorization to implied volatility σ and value variables Δ (e.g., $\sigma \in [10^{-4}, 5], \Delta \in [-0.999, 0.999]$). The above processing ensures comparability of samples from different sources and years in terms of contract definitions, maturity conventions, and price scales, while primary-key alignment prevents "cross-contract" mismatches.

3.2. Construction of Mispricing Labels

This study takes the BS theoretical price as the no-arbitrage reference. Rather than directly using the spot price, we infer an "implied spot consistent with the contract," S_{est}, from the contract's same-day Δ and implied volatility, thereby standardizing the moneyness convention and reducing errors caused by timing mismatches of the underlying asset. Under the Black–Scholes framework with continuous dividends, let the annualized time to maturity be τ, the strike be K, and the risk-free rate and dividend yield be r and q, respectively (in the empirical analysis, considering the sample maturity distribution and data availability, we approximate $r \approx q \approx 0$). For calls/puts, we have,

$$d_1 = \begin{cases} \emptyset^{-1}(\Delta), \text{call} \\ \emptyset^{-1}(1+\Delta), put \end{cases} \tag{3}$$

$$\ln\frac{S}{K}\, d_1\sigma\sqrt{\tau} - \frac{1}{2}\sigma^2\tau - (r-q)\tau \tag{4}$$

Accordingly, we obtain

$$S_{est} = K \cdot \exp(d_1\sigma\sqrt{\tau} - 1/2\sigma^2\tau - (r-q)\tau) \tag{5}$$

Substituting into the standard BS pricing formula and denoting $d_2 = d_1 - \sigma\sqrt{\tau}$, the theoretical price can be written as [5]

$$C = Se^{-q\tau}\phi(d_1) - Ke^{-r\tau}\phi(d_2),$$
$$P = Ke^{-r\tau}\phi(-d_2) - Se^{-q\tau}\phi(-d_1) \tag{6}$$

Comparing the market mid-quote with the theoretical price, we define the relative mispricing to eliminate dimensional effects:

$$e^{BS} = \frac{P_{mid}-P^{BS}}{\max(P^{BS},\varepsilon)} \tag{7}$$

Where is a very small positive number to avoid a zero denominator [12]. Considering that the distribution of mispricing varies systematically with the term structure and option type, we first estimate the conditional distribution of e^{BS} during the training period (2020–2022) under the partition "Call/Put × remaining-maturity buckets," and take the quantile thresholds within each bucket

$$(q_{low}, q_{high}) = (Q_{0.25}(e^{BS}), Q_{0.25}(e^{BS})) \tag{8}$$

Accordingly, we label samples into three classes:

$$\text{y} = \begin{cases} -1, e^{BS} < q_{low}\,(\text{Uderpriced}) \\ 0, q_{low} \le e^{BS} \le q_{high}(Fairly\ Priced) \\ +1, e^{BS} > q_{high}(Overpriced) \end{cases} \tag{9}$$

The thresholds are estimated using training-period information only and extrapolated to the test period (2023). If a particular bucket lacks thresholds in the test period, we backstop with overall training-period quantiles. This labeling strategy combines theoretical grounding with statistical robustness: the relative mispricing measure ensures comparability across price levels and moneyness samples; the bucketed conditioning controls for heterogeneity in maturity and type; quantile thresholds are robust to heavy tails and outliers; and out-of-time extrapolation avoids information leakage while explicitly reflecting annual distributional drift [2].

3.3. Construction and Training of the XGBoost Model

After constructing the labels, we extract from the aligned samples a set of leakage-free features that cover trading microstructure information (mid-quote, bid/ask, relative bid–ask spread), include key pricing factors (remaining time to maturity, strike, Delta, implied volatility, call/put indicator), and incorporate derived variables related to core BS quantities, such as the implied spot and implied moneyness:

$$\log moneyness_{est} = ln(S_{est}/K) \tag{10}$$

In particular, the $\hat{P}_{BS}$ used in label definition and the raw premium are excluded from the features to avoid directly leaking target information to the model [9]. The model adopts a multiclass gradient boosting tree (XGBoost), using "multiclass cross-entropy" as both the objective function and

the evaluation metric (implemented with multi: softprob and mlogloss). To interface with the multiclass API, the three-class labels are mapped to the set of nonnegative integers $\{-1,0,1\} \hookrightarrow \{0,1,2\}$ for training, and are decoded back to the original labels after prediction. Training is conducted on the 2020–2022 samples, with approximately 15% drawn from the training period via stratified holdout as a validation set to monitor loss and enable early stopping; given sample size and computational resources, we apply stratified down-sampling to the training set (e.g., capped at 600,000 observations) without altering class proportions. The main hyperparameters adopt relatively conservative tree depth and learning rate (e.g., about 400 trees, maximum depth 6, learning rate 0.05, subsample and column subsample 0.8, L_2 regularization 1.0) to balance fitting capacity and generalization performance.

After training, we report precision, recall, and F1 on the validation set, and then use the fixed model to evaluate out-of-sample data for 2023 to test robustness under out-of-time extrapolation. Throughout the workflow, intermediate artifacts (cleaned/aligned data, bucket thresholds, labeled data, and the final model) are persisted at each stage to ensure the reproducibility of the empirical results.

4. Empirical Analysis

After completing data cleaning, mispricing label construction, and model training, this study conducts a systematic identification and evaluation of option mispricing based on the XGBoost model. The analysis proceeds from three aspects: model identification performance, regularities in mispricing identification, and stability and applicability.

4.1. Evaluation of Model Identification Performance

Judging from the model's performance on the training and validation sets, XGBoost achieves a relatively robust identification effect in the three-class classification task.

Table 1. Results for the 2020-2022 training set

Category	Precision	Recall	F1-score	Support
-1 (Underpriced)	0.934	0.895	0.914	22624
0 (Fairly Priced)	0.898	0.937	0.917	44948
1 (Overpriced)	0.936	0.893	0.914	22428
Accuracy			0.916	90000
Macro Average	0.922	0.908	0.915	90000
Weighted Average	0.916	0.916	0.915	90000

The results for the training set (2020–2022) reported in Table 1 show that the model achieves an overall accuracy of 91.6%. The macro-averaged precision, recall, and F1 scores are 0.922, 0.908, and 0.915, respectively, indicating strong performance in identifying underpriced, fairly priced, and overpriced samples, with no pronounced class bias. Meanwhile, the confusion matrix reveals that the decision boundary between the overpriced and underpriced categories is effectively separated, and the recognition rate for the fairly priced class is high, suggesting that the model can effectively capture feature differences across the various mispricing states.

Table 2. Results for the 2023 test set

Category	Precision	Recall	F1-score	Support
-1 (Underpriced)	0.669	0.965	0.790	326514
0 (Fairly Priced)	0.891	0.614	0.727	482537
1 (Overpriced)	0.836	0.861	0.849	193799
Accuracy			0.776	1002850
Macro Average	0.799	0.813	0.789	1002850
Weighted Average	0.808	0.776	0.771	1002850

Table 2 reports the test results on the independent 2023 sample, with an overall accuracy of 77.6% and a macro-averaged F1 score of 0.789. Notably, the recall for underpriced samples (−1) reaches 0.965, indicating that the model is highly sensitive to underpricing and can effectively identify potential underpriced opportunities; by contrast, the recall for fairly priced samples (0) is only 0.614, suggesting a certain degree of misclassification. This disparity to some extent reflects that signals of price deviations from the fair-value range are more salient in the options market and thus more easily captured by the model.

4.2. Analysis of Mispricing Identification Patterns

Further examination of the distribution of predicted labels and class-wise metrics reveals certain regularities in the model's response to mispricing signals. The high recall for underpriced samples indicates that when the market exhibits clear underpricing, the price deviations highlighted by the BS pricing model can be accurately amplified and exploited by the machine learning algorithm. Although the identification rate for overpriced samples is slightly lower than that for underpriced samples, both precision and recall remain above 0.83, demonstrating that the model likewise possesses stable recognition capability for overpriced cases. By comparison, fairly priced states are more prone to being misclassified as overpriced or underpriced, reflecting the presence of small fluctuations and bounded rationality in actual market prices that blur the boundaries of the "fair" state.

These patterns are consistent with real market characteristics. Factors such as volatility smiles, supply–demand imbalances, and trading frictions frequently cause market prices to deviate systematically from BS theoretical values in the options market [1]. The model can identify and capture these deviation trends, thereby providing investors with potential signals for arbitrage or risk management.

4.3. Stability and Applicability

The model exhibits strong stability across samples from different years. The high accuracy during the training phase (2020–2022) and the solid performance during the testing phase (2023) indicate that the model not only fits mispricing patterns in-sample but also possesses cross-period predictive capability. This cross-year robustness validates the rationality of the mispricing label construction and reflects the advantage of XGBoost in handling large-scale nonlinear features.

However, the results also suggest certain limitations in identifying "fairly priced" samples, manifested as insufficient recall. This implies that, in practical applications, the model may misclassify some fair states as deviating states, thereby triggering additional trading signals. Therefore, in practice, it is advisable to combine market liquidity, transaction costs, and richer pricing factors to further screen and calibrate the

identification results.

Overall, the empirical findings of this study indicate that constructing mispricing labels based on BS pricing and introducing XGBoost modeling can, to a considerable extent, identify pricing deviations in the options market, demonstrating good predictive performance and cross-period robustness.

5. Result

Based on high-frequency data from the S&P 500 ETF options market, this paper proposes and implements an empirical framework that uses the Black–Scholes theoretical price as a reference and combines it with machine learning to identify market mispricing. In terms of data processing, option and underlying information are obtained from the WRDS database, rigorously cleaned and aligned, and organized into a structured sample spanning 2020–2023. For label design, a relative mispricing measure is introduced and combined with a bucketed-quantile strategy to construct three-class labels—overpriced, underpriced, and fairly priced—thereby improving the comparability and robustness of mispricing judgments. In modeling, an XGBoost classification model is employed to automatically identify different mispricing states, with testing conducted on out-of-sample data from 2023.

The empirical results show that the model delivers solid classification performance in both the training and test periods. In the test data, the model attains an overall accuracy of 77.6%; in particular, the precision and recall for underpriced samples remain high, indicating that the constructed mispricing labels exhibit a stable structure and that the model can capture the statistical features and market regularities underlying mispricing signals. The classification results also reveal structural pricing deviations in the options market. Although the BS theoretical price relies on simplifying assumptions, it still provides a valuable no-arbitrage benchmark for mispricing identification. Therefore, the proposed framework not only offers a data-driven solution to the problem of option mispricing identification but also supplies a foundational model for subsequent research such as arbitrage signal extraction and dynamic trading strategy optimization.

6. Conclusion

Although this study achieves encouraging empirical results in the task of mispricing identification, several limitations remain and warrant further improvement and extension in future work.

In constructing theoretical prices, the BS model is still used as the benchmark. Although robustness has been enhancedvia implied-spot back-out and quantile standardization, the structural assumptions of the BS model (e.g., constant volatility, continuous trading) do not fully hold in real markets, which may cause the constructed labels to deviate from truly reasonable pricing [5]. Future research could consider introducing more flexible pricing benchmarks—such as contract pricing models based on the market implied-volatility surface—to further improve the rationality of label characterization. In addition, feature design in this study is primarily based on trading attributes and BS-derived quantities, without incorporating information from macro variables, market volatility factors, or trader-behavior indicators, which may limit the breadth of model identification [13]. Subsequent work may integrate multi-level, multi-frequency data sources to build a more interpretable multi-factor feature system and enhance the model's ability to adapt to complex mispricing phenomena.

Regarding model choice, this paper adopts a static XGBoost classifier, which, while offering good fit and interpretability, is somewhat limited in handling temporal dynamics. Future studies may explore time-series models (e.g., RNNs, Transformers) or employ rolling-window training schemes to enable the model to adapt dynamically to changes in market structure and improve cross-period predictive capability [14].

Finally, this paper does not link the mispricing identification results to live trading signals for joint testing. Future research can extend the framework to quantitative strategy design, exploring the operability and return performance of mispricing signals in actual investment, thereby enhancing the practical value of the study.

References

[1] George M. Constantinides, Jens Carsten Jackwerth, and Stylianos Perrakis. (2009) Mispricing of S&P 500 index options. The Review of Financial Studies 22.3: 1247-1277.

[2] Gu, Shihao, Bryan Kelly, and Dacheng Xiu. (2020) Empirical asset pricing via machine learning. The Review of Financial Studies 33.5: 2223-2273.

[3] Chen, T. (2016) XGBoost: A scalable tree boosting system. Cornell University: 785–794.

[4] French, Dan W., and Linda J. Martin. (1988) The measurement of option mispricing. Journal of Banking & Finance 12.4: 537-550.0.

[5] Black, Fischer, and Myron Scholes. (1973) The pricing of options and corporate liabilities. Journal of political economy 81.3: 637-654.

[6] Merton, Robert C. (1976) Option pricing when underlying stock returns are discontinuous. Journal of financial economics 3.1-2: 125-144.

[7] Bollerslev, Tim. (1986) Generalized autoregressive conditional heteroskedasticity. Journal of econometrics 31.3: 307-327.

[8] Bakshi G., Panayotov G., (2008). A framework for studying option mispricing: a general test and empirical evidence. In: AFA 2009 San Francisco Meetings. San Francisco.

[9] Dumas, Bernard, Jeff Fleming, and Robert E. Whaley. (1998) Implied volatility functions: Empirical tests. The Journal of Finance 53.6: 2059-2106.

[10] Bakshi, Gurdip, Charles Cao, and Zhiwu Chen. (1997) Empirical performance of alternative option pricing models. The Journal of finance 52.5: 2003-2049.

[11] Kelly, Bryan, and Seth Pruitt. (2013) Market expectations in the cross-section of present values. The Journal of Finance 68.5: 1721-1756.

[12] Yan, S. (2011). Jump risk, stock returns, and slope of implied volatility smile. Journal of Financial Economics, 99(1), 216-233.

[13] Gençay, Ramazan, and Aslihan Salih. (2003) Degree of mispricing with the Black-Scholes model and nonparametric cures. Annals of Economics and Finance 4: 73-102.

[14] Tashman, Leonard J. (2000) Out-of-sample tests of forecasting accuracy: an analysis and review. International journal of forecasting 16.4: 437-450.