Option Pricing Strategy

Ayan Goswami

1 Introduction

This document summarizes the findings reported in `strategy_book.pdf`. The original analysis uses option data to evaluate pricing efficiency and develop a trading approach. A K–Nearest Neighbors (KNN) model is employed as a non-parametric estimator of conditional expectations $E(X \mid Y, Z, \ldots)$. The KNN output serves as a prior input in a Bayesian framework to signal long or short trades.

2 Data Loading and Preparation

Option data was loaded from a socket stream leveraging Alpaca's IEX market data (see ./socket), updated every 10 seconds. For the purposes of this investigation, we focused on options with 0 days to expiry (0dte), as they have the largest trading volume and are assumed to self-correct quicker than options with a later expiry. All options between 99% to 101% of the current price were stored in a csv, along with all the corresponding greeks and implied volatility. To read more about how this was calculated see the Appendix. When this investigation was conducted, SPY hovered around 620$, hence for a given timestamp, 14 or 15 calls and puts were stored. This was done to get an even spread of in the money (ITM), at the money (ATM) and out the money (OTM) options.

Moneyness, defined as the relative position of the underlying asset price to the strike price, exhibits a strong relationship with the option premium. For call options, those that are in the money (ITM), i.e., with strike prices lower than the spot price ($K < S$), tend to have higher intrinsic value and thus higher premiums. Conversely, out of the money (OTM) options ($K > S$) carry mostly time value and are priced lower. This nonlinear relationship is especially pronounced in short-dated options, where the time decay is steep and the implied volatility surface varies across moneyness levels. Empirically, our dataset reflects this convex structure, with ATM options ($K \approx S$) typically exhibiting peak implied volatilities and forming the apex of the option price curve.

There is usually a very steep drop-off in 0dte OTM options' prices, mainly due to the fact that they will expire worthless by the end of the day. To enable fair comparison of option prices across different timestamps and mitigate scale

differences between call and put options, we applied a logarithmic normalization procedure to standardize the prices. The transformation is defined as:

$$\text{standardized\_price} = \frac{\log(p+1) - \log(p_{\min}+1)}{\log(p_{\max}+1) - \log(p_{\min}+1)}$$

where p is the latest trade price of the option, and $p_{\min}, p_{\max}$ are the minimum and maximum trade prices for a given option type (call or put) within a specific timestamp.

This transformation is applied separately for call and put options, and performed independently at each timestamp. The rationale is twofold: (1) call and put options have inherently different pricing distributions, and (2) the market conditions change across timestamps, requiring local normalization to preserve intra-timestamp price structure.

Logarithmic scaling was chosen to compress the skewness in price distribution, especially in deep ITM or OTM options where price differences can be exponential. Adding 1 inside the logarithm avoids issues with near-zero prices. The result is a normalized feature `standardized_price` bounded in $[0, 1]$, suitable for downstream tasks such as KNN modeling and residual analysis. The following is a random timestamp slice:

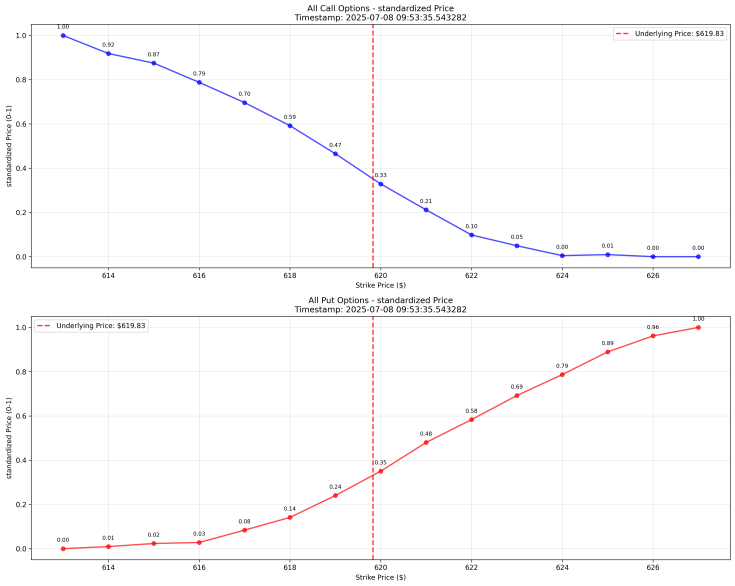

Figure 1: Standardized option prices for a given strike

Option data were loaded from several processed files and split into training and validation sets using a 75%/25% split. Since the θ decay is so high for 0dte options, we focused on trade entry and exit windows between 30 seconds to 5 minutes. This way the α can be measured reliably without accounting for price

decay. Hence, a column was appended to our dataset called `price_diff`, which measured:

$$\text{price\_diff}_x \text{price\_diff}_{x\,\text{price\_diff}_{x\,\text{price\_diff}_x}} = \frac{\frac{\sum_{i=t}^{x+t} price_i}{x}}{price_t}$$

where x is the look-forward period in tens of seconds, and t is the current time period. This will serve as a metric of success and efficacy in our paper.

3 KNN Modeling

The model predicts standardized option prices. A wrapper function scales predictors and returns both root mean squared error (RMSE) and R^2 values. Exhaustive subset selection determined that `delta`, `gamma`, `moneyness`, and `rho` gave the best performance with

$$\begin{aligned} \text{RMSE} &= 0.0269, \\ R^2 &= 0.9946. \end{aligned}$$

Optimal k was found by evaluating a range of values and occurred at $k = 3$. Residual analysis showed near-normal errors; a significance threshold was derived as

$$q = \Phi^{-1}(0.95, \mu_r, \sigma_r) = 0.0146,$$

where r denotes prediction residuals.

4 Hypothesis Testing

Price moves were summarized by

$$price\_diff_x = \frac{\frac{\sum_{i=t}^{t+x} price_i}{x}}{price_t}.$$

Mean returns were compared for options with residuals above q (GEQ), below $-q$ (LEQ), and for all options. The null hypothesis

$$H_0 : \mu_{\text{GEQ}} = \mu_{\text{LEQ}} = \mu_{\text{All}}$$

was rejected for the six-period window with p-values below 0.05. The alternative hypothesis

$$H_1 : \mu_{\text{GEQ}} > \mu_{\text{All}} > \mu_{\text{LEQ}}$$

was supported in several horizons.

5 Trading Simulation

Signals were generated on new data using the KNN predictions. Long trades were entered when residuals exceeded q, and short trades were entered when residuals were less than $-q$. If $price\_diff\_x$ is the average future price ratio, then

$$\alpha_{\text{long}} = (price_t \cdot price\_diff_x) - price_t,$$
$$\alpha_{\text{short}} = price_t - (price_t \cdot price\_diff_x).$$

A test run on 5268 trades produced \$1685 in profit. Comparison against 100 random simulations gave a null threshold of \$445.64, indicating the strategy outperformed chance at the five percent level.

6 Results

Figure 2 shows the average price change by residual group. The high residual group typically exhibits a higher mean price ratio across horizons. Table 1 lists mean future price ratios and p-values for each window.

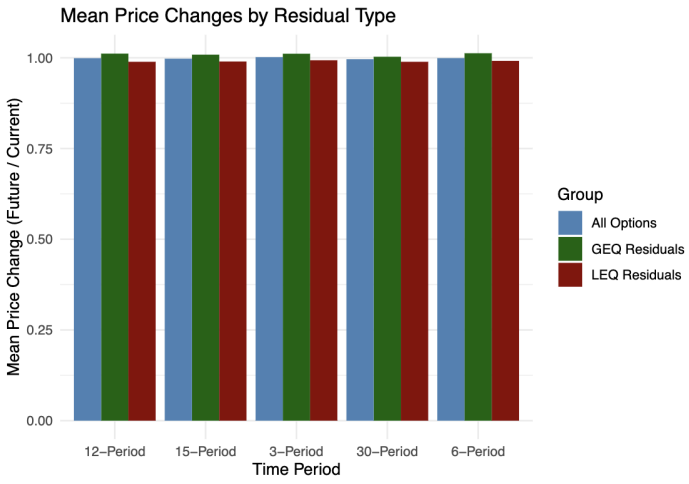

Figure 2: Mean future price ratios by residual group

7 Discussion

The KNN output served as a prior that highlighted relative mispricing. The residual-based signals produced consistent return differences in favor of positive

Period (10s)	Mean (Under)	Mean (Over)	Mean (Overall)	$P(\mu_{\text{GEQ}} = \mu_{\text{All}})$	$P(\mu_{\text{All}} = \mu_{\text{LEQ}})$
3	1.0111	0.9932	1.0019	0.0742	0*
6	1.0125	0.9915	0.9991	0.0265*	0.0006*
12	1.0114	0.9889	0.9989	0.0504	0.0003*
15	1.0086	0.9897	0.9975	0.0818	0.0052*
30	1.0028	0.9889	0.9961	0.2437	0.0405*

Table 1: Mean future price ratios by residual group and associated p-values. * indicates significance at $p = 0.05$.

residuals. While the profits observed in backtesting were modest, the approach performed better than random trading according to the null threshold. Further study could refine feature selection and risk controls.

Moreover, machine learning models such as random forests and boosted trees perform better than a naive Bayes solution such as this one according to Caruana (2006).

8 Conclusion

The KNN model approximates $E(X \mid Y, Z, \ldots)$ and acts as a prior in a Bayesian decision rule. Trades are triggered when observed residuals cross the significance threshold. The strategy generated profits above a random baseline, suggesting predictive value in the modeled relationships.

A Black-Scholes Formulas

A.1 Option Pricing Formulas

The Black-Scholes model for European option pricing is given by the following formulas:

A.1.1 Call Option Price

$$C(S, K, T, r, \sigma) = S \cdot N(d_1) - Ke^{-rT} \cdot N(d_2) \tag{1}$$

A.1.2 Put Option Price

$$P(S, K, T, r, \sigma) = Ke^{-rT} \cdot N(-d_2) - S \cdot N(-d_1) \tag{2}$$

A.1.3 Parameters d_1 and d_2

$$d_1 = \frac{\ln(S/K) + (r + \sigma^2/2)T}{\sigma\sqrt{T}} \tag{3}$$

$$d_2 = d_1 - \sigma\sqrt{T} \tag{4}$$

where:

- S = Current stock price
- K = Strike price
- T = Time to expiration (in years)
- r = Risk-free interest rate
- σ = Volatility of the underlying asset
- $N(\cdot)$ = Cumulative distribution function of the standard normal distribution

A.2 Option Greeks

The Greeks measure the sensitivity of option prices to various factors:

A.2.1 Delta

Measures the rate of change of option price with respect to changes in the underlying asset's price.

Call Option Delta:

$$\Delta_{call} = N(d_1) \tag{5}$$

Put Option Delta:

$$\Delta_{put} = N(d_1) - 1 \tag{6}$$

A.2.2 Gamma

Measures the rate of change of delta with respect to changes in the underlying price.

$$\Gamma = \frac{N'(d_1)}{S\sigma\sqrt{T}} = \frac{e^{-\frac{d_1^2}{2}}}{S\sigma\sqrt{2\pi T}} \tag{7}$$

Gamma is the same for both call and put options.

A.2.3 Theta

Measures the rate of change of option price with respect to the passage of time (time decay).

Call Option Theta:

$$\Theta_{call} = -\frac{SN'(d_1)\sigma}{2\sqrt{T}} - rKe^{-rT}N(d_2) \tag{8}$$

Put Option Theta:

$$\Theta_{put} = -\frac{SN'(d_1)\sigma}{2\sqrt{T}} + rKe^{-rT}N(-d_2) \tag{9}$$

Theta is typically expressed in value per day, dividing by 365.

A.2.4 Vega

Measures the rate of change of option price with respect to changes in volatility.

$$Vega = S\sqrt{T}N'(d_1) \tag{10}$$

Vega is the same for both call and put options and is typically expressed as change per 1% change in volatility.

A.2.5 Rho

Measures the rate of change of option price with respect to changes in the risk-free interest rate.

Call Option Rho:

$$\rho_{call} = KTe^{-rT}N(d_2) \tag{11}$$

Put Option Rho:

$$\rho_{put} = -KTe^{-rT}N(-d_2) \tag{12}$$

Rho is typically expressed as change per 1% change in interest rate.

A.3 Implied Volatility

Implied volatility is the volatility value that, when input into the Black-Scholes formula, yields a theoretical option price equal to the market price. It is typically solved using numerical methods such as the Newton-Raphson method:

$$\sigma_{n+1} = \sigma_n - \frac{BS(S, K, T, r, \sigma_n) - Market\_Price}{Vega} \tag{13}$$

where $BS(\cdot)$ is the Black-Scholes pricing function and iterations continue until convergence.

A.4 Put-Call Parity

For European options on non-dividend paying stocks:

$$C + Ke^{-rT} = P + S \tag{14}$$

This relationship can be used to derive the price of a put option from a call option with the same strike and expiration, or vice versa.

B References

1. Caruana, R.; Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. Proc. 23rd International Conference on Machine Learning.