Stock options pricing via machine learning methods combined with firm characteristics

Panayiotis C. Andreou*†
Chulwoo Han‡
Nan Li†

May 2023

Abstract

This paper proposes machine learning-based option pricing models that incorporate firm characteristics. We employ two semi-parametric models, one that uses machine learning to predict the implied volatility and the other to correct the pricing error of the Black-Scholes model and use 114 firm characteristics as well as option-related variables as the input features. Tested on the stock options in the US market, we find that both models outperform a parametric model even without firm characteristics, and firm characteristics significantly enhance the performance of these models. Idiosyncratic volatility, share price, market equity, illiquidity, and firm age are found to be the most important features.

Keywords: Option pricing; Semi-parametric model; Firm characteristics; Machine learning

*Cyprus University of Technology, Department of Commerce, Finance and Shipping, Archbishop Kyprianou 30, Limassol 3603, Cyprus.
†Durham University Business School, Mill Hill Lane, Durham, DH1 3LB, United Kingdom.
‡Sungkyunkwan University, 25-2 Seonggyungwan-ro, Jongno-gu, Seoul, South Korea.
E-mail addresses: panayiotis.andreou@cut.ac.cy, chulwoo.han@skku.edu, nanli.2@durham.ac.uk

1 Introduction

Options valuation has garnered significant attention in the field of finance, both from theoretical and practical perspectives. Academics are intrigued by the potential for a robust option pricing model to shed light on financial market operations, while market makers aim to use an efficient pricing model to determine prices in the derivatives market. The Black-Scholes (BS) model (Black and Scholes, 1973) is the earliest and most well-known option pricing model. Despite its simplicity, it provides a good estimate of option prices and remains as one of the most important option pricing models. The BS model relies on the assumption that stock returns are normally distributed and have constant volatility over time. However, there is abundant evidence that stock returns have a fat-tailed distribution and exhibit time-varying volatility. To account for the non-normality of stock returns, Corrado and Su (1996) augment the BS model with skewness and kurtosis. To account for the time-varying volatility, Heston (1993) and Hagan, Kumar, Lesniewski, and Woodward (2002) propose stochastic volatility models, which assume the volatility to be a random variable. These models allow a more realistic representation of the underlying asset's volatility and provide more accurate pricing of options. For American options, which face premature exercise and dividend payment issues, tree-based (Cox, Ross, and Rubinstein, 1979; Boyle, 1986) or Monte Carlo simulation-based (Broadie and Glasserman, 1997; Longstaff and Schwartz, 2001; Andersen and Broadie, 2004) methods have been proposed. A closed-form solution can also be obtained for an American option in case of known absolute dividends (Roll, 1977; Geske, 1979; Whaley, 1981) or proportional dividends (Villiger, 2006).

The majority of option pricing research including those mentioned above adopts a parametric method, *i.e.*, they assume a certain distribution for the underlying asset return and derive the fair price of the option under the no-arbitrage condition. While parametric models are preferable as they are established on a solid economic foundation and often allow an analytic option pricing formula, the reality could deviate from their underlying assumptions, and parametric models often struggle to fit the actual option prices observed in the market.

More recently, machine learning-based models have been proposed as an alternative approach. These models have the ability to capture the nonlinear and complex relationships between option prices and their underlying factors, making them more flexible than traditional parametric models. Machine learning-based models can be categorized into two types: non-parametric and semi-parametric models. Non-parametric models are agnostic about the economic theory behind option pricing and predict the option price by learning purely from the data the relationship between the input variables and the option price. One of the earliest studies in this category is work of Hutchinson, Lo, and Poggio (1994), which treats option pricing as a regression problem. They demonstrate that the learning network is superior to traditional parametric methods in option valuation. With advances in machine learning algorithms, researchers have proposed ways to improve the generalization of a neural network, such as Bayesian adjustment, early stopping, and bagging, which allow more robust pricing of options (Gençay and Qi, 2001). Gradojevic, Gençay, and Kukolj (2009)) use modular neural networks to improve prediction performance, while Liang, Zhang, Xiao, and Chen (2009) use a combination of neural networks and support vector regression to reduce pricing errors in traditional option pricing methods such as Monte Carlo simulation, binomial trees, and finite difference methods.

Non-parametric models can fit the data flexibly without making any assumptions about option value. However, the very flexibility can be toxic and expose the models to the risk of overfitting. Semi-parametric models address this risk by combining a parametric model with machine learning. Guided by economic theory, a semi-parametric model can fit the data with a more parsimonious structure. One approach is to employ machine learning to predict unobservable variables such as implied volatility, which are then used as input to a parametric model. Andreou, Charalambous, and Martzoukos (2006) and Andreou, Charalambous, and Martzoukos (2010) predict volatility for the BS model and volatility, skewness, and kurtosis for Corrado and Su (1996)'s model via a neural network and use them as input to the corresponding parametric option pricing model. The other approach

corrects the pricing error of a parametric model using machine learning. Lajbcygier and Connor (1997) employ neural networks to correct the pricing error of the BS model.

An advantage of machine learning is that it can accommodate any features that potentially carry information about the option price. However, previous studies exclusively use only option-related variables, such as the underlying asset price, strike price, time-to-maturity, risk-free rate, and implied volatility. Although the theory states that the option price should be fully described by these variables, the actual price can also be affected by other factors such as the investors' view on the underlying asset. For individual stock options, the actual option price can be higher than the theoretical value if the investors' sentiment toward the underlying firm's future growth is positive. Inspired by this idea, we aim to establish an option pricing model that incorporates additional factors other than the variables used in a parametric model. Without prior knowledge about what other factors could affect the option price, we employ a large set of firm characteristics published in the literature that can potentially carry information about the firm's financial health and future performance, and hence have an impact on the option price. To the best of our knowledge, this research is the first of its kind that incorporates firm characteristics in option pricing.

We employ two semi-parametric methods to incorporate firm characteristics. The first model is based on Andreou et al. (2010)'s generalized parametric function (GPF) model, which employs machine learning for the prediction of implied volatility. The second model is based on Lajbcygier and Connor (1997)'s hybrid (HBD) model, which employs machine learning for pricing error correction. These models are compared against a parametric benchmark model. We choose as the benchmark Dumas, Fleming, and Whaley (1998)'s deterministic volatility function (DVF) model as it is widely used as a benchmark owing to its ability to effectively deal with the volatility smile and its ease of use (Christoffersen and Jacobs, 2004; Berkowitz, 2009; Christoffersen, Heston, and Jacobs, 2009).

We evaluate the models using individual stock options in the US market from 1996 to 2021. As to the firm characteristics, we choose 114 firm characteristics from Jensen, Kelly,

and Pedersen (2021) after removing firm characteristics with many missing values. We first find that the semi-parametric models, GPF and HBD, outperform the benchmark, DVF, even before incorporating firm characteristics. The root mean square errors (RMSEs) of GPF and HBD are respectively 2.447 and 2.527, whereas that of DVF is 5.799. Second, firm characteristics can further improve the performance of GPF and HBD: the RMSEs of the two models are reduced to 2.064 and 2.110, respectively. GPF consistently renders a smaller error than HBD. We conjecture that this is because the range of implied volatility is narrower than the range of option price residual, making a more stable prediction possible. We assess the models with various subsets of options defined by the option type, time-to-maturity, and moneyness, and find that the models perform consistently across these options groups. The usual option features such as time-to-maturity and moneyness are the most important features for pricing options, but firm characteristics are also deemed to play a non-trivial role. Firm characteristics collectively have a feature importance score of 17 out of 100 in HBD. In particular, idiosyncratic volatility, share price, market equity, illiquidity, and firm age turn out to be the most important firm characteristics in option pricing.

The contribution of this paper is twofold. First, we integrate firm characteristics in option pricing and demonstrate that they can enhance option pricing performance. We also identify firm characteristics that are deemed to add the most value. The second contribution is that we evaluate the performance of machine learning-based models for individual stock options. Previous studies employ machine learning mostly to price European index options. However, a machine learning-based model is more suitable for stock options for the following reasons. Stock options are American options and are more challenging to price using a parametric model. Investors' view on the underlying firm's growth can affect their early exercise decision. Firm-specific factors are also likely to be important pricing factors that cannot be captured by a parametric model. Finally, there are a significantly larger amount of stock option data than index option data. A large amount of data is crucial for a machine learning algorithm to generalize without overfitting. We show that the proposed models can

successfully price individual stock options.

This paper is structured as follows. Section 2 describes the option pricing models used in this study, Section 3 details the data and methodology for the empirical analysis, Section 4 presents the empirical results, Section 5 runs robustness tests, and Section 6 concludes.

2 Models

2.1 Parametric model

The BS pricing formula for an option without dividends is given by

$$C^{BS} = SN(d) - Xe^{-rT}N\left(d - \sigma\sqrt{T}\right), \tag{1}$$

$$P^{BS} = -SN(-d) + Xe^{-rT}N\left(-d + \sigma\sqrt{T}\right), \tag{2}$$

$$d = \frac{\ln\left(\frac{S}{X}\right) + rT + \frac{\left(\sigma\sqrt{T}\right)^2}{2}}{\sigma\sqrt{T}}, \tag{3}$$

where C^{BS} and P^{BS} are the price of a call and a put option, respectively, S is the spot price of the underlying stock, X is the strike price of the option, T is the time to maturity, r is the continuously compounded risk-free interest rate, σ is the stock return's volatility, and $N(\cdot)$ denotes the standard normal cumulative distribution. We first use the BS model to derive the implied volatility. To find the implied volatility, we solve an optimization problem that has the following form:

$$Loss_i = \min(Price_{obs,i} - Price_{est,i})^2 \tag{4}$$

, where the $Price_{obs,i}$ is the market price of option i, $Price_{pre,i}$ is the estimated price of option i.

We then use as a benchmark Dumas et al. (1998)'s DVF model, which incorporates

regression-based implied volatility into the BS model. The DVF model aims to capture the curvature of the volatility smile and the term effect by assuming that the implied volatility is a quadratic function of moneyness and maturity:

$$\sigma_{DVF} = \max\left(0.01, a_0 + a_1 K + a_2 K^2 + a_3 T + a_4 T^2 + a_5 KT\right), \tag{5}$$

where $K = S/X$ denotes moneyness.[1] We calibrate the DVF function daily by fitting it to the implied volatility observed in the previous 10 days.

2.2 GPF model

Andreou et al. (2010) propose a semi-parametric option pricing model (GPF model) and show that it exhibits an enhanced pricing performance when applied to index options. The GPF model predicts unobservable option variables (volatility for the BS model and volatility, skewness, and kurtosis for Corrado and Su (1996)'s model) via a neural network and uses them as input to the corresponding parametric option pricing model. An advantage of the GPF model is that it combines the benefits of both parametric and nonparametric methods. The parametric model provides a theoretical framework for option pricing, while the machine learning algorithm generates a more accurate prediction of the input variables. By combining these two approaches, the GPF model can effectively address the limitations of traditional parametric and nonparametric methods in option pricing. The GPF model is flexible and can be adapted to meet the needs of different option pricing scenarios, making it a versatile tool for option pricing research.

For our study, we extend the GPF model in several ways. The original GPF model is a one-step approach, where the machine learning algorithm is trained so that the option pricing error is minimized: the loss function is defined as the mean square error of the option

[1] Dumas et al. (1998) use strike price instead of moneyness. We test both versions and choose moneyness as it gives better results. Andreou, Charalambous, and Martzoukos (2014) also use moneyness.

price. The loss function we optimize in the GPF is shown in Equation 6

$$Loss(GPF) = \min \sum_{i=1}^{N} (\sigma_{obs,i} - \sigma_{pre,i})^2 \tag{6}$$

, where i is the i-th observation, N represents total observations. $\sigma_{obs,i}$ represents the observed implied volatility of option i derived from the market price using the binomial model, and the $\sigma_{pre,i}$ represents the predicted implied volatility of option i.

A downside of this approach is that the objective function becomes highly nonlinear as it involves the parametric option pricing model. We mitigate this issue by training the machine learning algorithm so that the distance between the predicted volatility and the implied volatility is minimized. The implied volatility is derived from the market price using the BS formula. The predicted volatility is then used as input to the BS model to obtain the option price. We also replace the neural network with CatBoost, a gradient boosting method, as gradient boosting methods have shown superior performance in many machine learning competitions.[2] Most importantly, besides the option-related features, we include firm characteristics as additional input features to the machine learning algorithm. As to the option-related features, we use K, K^2, T, T^2, KT, and σ_{imp}, where σ_{imp} is the average of the implied volatilities of all the options with the same underlying asset over the past 10 days. Figure 1 describes the schematic structure of the GPF model.

[Insert Figure 1, here]

2.3 Hybrid model

The hybrid model (HBD model) proposed by Lajbcygier and Connor (1997) also combines the advantages of a parametric option pricing model and machine learning to improve option pricing accuracy. However, it is different from GPF in that it predicts the pricing error of a parametric model via machine learning. HBD first calculates the option price using a

[2] More details of CatBoost can be found in the Appendix B.

parametric model such as the BS model. This model price is then contrasted with the market price to obtain the pricing error (residual). A machine learning algorithm is then employed to predict the residual. The final option price is the sum of the model price and the predicted residual. The loss function we optimize in the HBD is shown in Equation 7

$$Loss(HBD) = \min \sum_{i=1}^{N} (redisual_{obs,i} - residual_{pre,i})^2 \tag{7}$$

, where i is the i-th observation and N represents the total number of observations. $redisual_{obs,i}$ represents the observed pricing residual calculated using the market price of the option minus the parameter model-based option price of option i, and $redisual_{pre,i}$ represents the prediction pricing residual calculated using the market price of the option minus the predicted option price of option i.

We extend the HBD model by incorporating firm characteristics and employing CatBoost instead of a neural network. Figure 2 describes the schematic structure of the HBD model.

[Insert Figure 2, here]

3 Data and Methodology

3.1 Option data

Option data are obtained from OptionMetrics' IvyDB US and cover the options in the US market from 1996.01 to 2021.12. As we need the first two years of the sample to train the machine learning-based models, we set the out-of-sample period to be from 1998.01 to 2021.12. The dataset includes the best bid and ask prices of all options at the close of each trading day, as well as the ticker of the underlying stock, option ID, issue date, expiration date, strike price, volume, and open interest. Following Dumas et al. (1998), we define the option price as the midpoint of the bid and ask prices to reduce the estimation noise of implicit parameters. The underlying stocks' prices are collected from the Securities table in

OptionMetri cs' IvyDB US, and the 3-month Treasury bill rate, which is used as a proxy for the risk-free rate, is obtained from the St. Louis Federal Reserve Economic Data. Following Bakshi, Cao, and Chen (1997) and Andreou et al. (2010), we filter the option data using the following criteria.

1. Options with a trading volume of less than 100 are eliminated as they are deemed to be illiquid and their prices may not represent the actual market price.
2. The time to maturity should be at least six days and no longer than 365 days as options near expiration may induce liquidity-related biases.
3. Options with price quotes less than 0.1 are eliminated.
4. The moneyness of an option should be between 0.8 and 1.2.
5. Options with missing or abnormal implied volatility are eliminated: Some options have abnormal prices that lead to abnormal implied volatilities.

After filtering, the final data set contains 16,050,622 observations. Sample characteristics of the dataset are reported in Table 1 and 2. Table 1 shows that there are more observations of out-of-the-money options than in-the-money options and more observations of call options than put options. The volatility smile is observed in all maturity groups and it is more pronounced in the near-term options. Table 2 reports the number of options per underlying stock year by year. The number of options per stock has increased over time. On average the average number of options increases from 3 in 1996 to 13 in 2021. However, the median number of options remains small at 3 even in 2021, while the maximum number of options increases to 339, which implies that there are only a handful number of stocks with many options and the rest have only a few options associated with them.

[Insert Table 1, here]

[Insert Table 2, here]

3.2 Firm characteristics

Firm characteristics play an important role in understanding the performance of firms and their prospective growth. Since options are derived from the underlying asset, their prices reflect the market's expectation of the firm's future performance, and it is expected that the firm characteristics can provide valuable information for option pricing. Therefore, we aim to explore the impact of firm characteristics on the option price and measure the additional information they provide.

We use 114 firm characteristics that include accounting ratios, momentum features, stock return volatility, and other characteristics related to the firm's operation, growth, risk, and performance. These characteristics are selected from the comprehensive list of firm characteristics in Jensen et al. (2021) after eliminating firm characteristics with more than 20% missing values. We fill the remaining missing values with the cross-sectional median value following Gu, Kelly, and Xiu (2020). These firm characteristics are generated using PyAnomaly, a powerful Python library for firm characteristics generation and asset pricing study.[3] The package provides easy access to various financial data sources and ensures data accuracy and consistency. The list of the firm characteristics with their description can be found in the Appendix. The exact definitions of the firm characteristics can be found in Jensen et al. (2021) or the references therein. In order to avoid forward-looking bias, the firm characteristics in month $t-6$ are matched with the option data in month t.

3.3 Models and evaluation metrics

We compare the out-of-sample option pricing performance of the following five models: the DVF model (DVF), the GPF model without firm characteristics (GPF), the GPF model incorporating firm characteristics (GPF_f), the hybrid model without firm characteristics (HBD), and the hybrid model incorporating firm characteristics (HBD_f).

The volatility function of the DVF model is calibrated daily, whereas the machine

[3] https://pyanomaly.readthedocs.io/en/latest/index.html

learning-based models are trained once a month to reduce the computational cost. The models are trained using the past 24 month data, of which the first 21 months are used as the training set and the remaining three months are used as the validation set. We avoid hyperparameter tuning and use the default hyperparameter values of CatBoost so as to evaluate the models from a conservative perspective.

The models are evaluated using the following error measures: root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). In addition, we employ the model confidence set (MCS) of Hansen, Lunde, and Nason (2011) to compare the models. The purpose of the MCS is to provide a measure of uncertainty in the model by identifying a range of models that are consistent with the data. The MCS is implemented through the following steps.

1. Estimate the parameters of a candidate model using the maximum likelihood or some other estimation technique.
2. Calculate the test statistic for the candidate model, which measures the goodness of fit of the model to the data.
3. Using the test statistic, determine the critical value that corresponds to a certain level of confidence. This critical value can be used to determine the MCS.
4. Evaluate the candidate model by comparing its test statistic to the critical value. If the test statistic is less than the critical value, the model is considered to be a good fit and is included in the MCS.
5. Repeat steps 1-4 for multiple candidate models to form the MCS.

The MCS is similar to a confidence interval, meaning that it contains the best model with $100(1-\alpha)\%$ confidence. As α decreases, the number of models in the MCS increases, just like the size of a confidence interval. The primary output is a set of p-values, with a small p-value indicating that a model is less likely to belong to the MCS. We report the p-values of the models in the empirical analysis.

4 Empirical Analysis

We evaluate the out-of-sample performance of the models using the error metrics defined in the previous section. We first compare the pricing errors of all options and then examine the pricing errors of different subsets defined by option type, moneyness, and maturity to validate the consistency of the pricing performance. We also identify informative firm characteristics by analyzing feature importance.

4.1 Pricing performance

4.1.1 All options

Table 3 reports the pricing performance of the models evaluated using all options. First, comparing the three models that do not incorporate firm characteristics, DVF, GPF, and HBD, we find that GPF performs best, followed by HBD. The RMSEs of GPF, HBD, and DVF are 2.447, 2.527, and 5.799, respectively. The superiority of GPF is also reflected in the other error measures. It is notable that the machine learning-based models can reduce the pricing error almost by half. This result suggests that employing machine learning can effectively reduce pricing error, and the GPF architecture is more effective than that of HBD.

When firm characteristics are incorporated, both GPF and HBD yield smaller pricing errors: the RMSEs of $\text{GPF}_f$and $\text{HBD}_f$are respectively 2.064 and 2.110, which are considerably smaller than those of GPF and HBD. Firm characteristics improve the pricing performance of both GPF and HBD by about 16%, making $\text{GPF}_f$the best performer among all five models, followed by HBD_f. The MCS p-value indicates that $\text{GPF}_f$is the only model that is included in the MCS.

The reason for the superior performance of GPF over HBD appears to be related to the way it integrates machine learning. GPF predicts implied volatility via machine learning while HBD predicts residual. The same error in implied volatility can lead to an error of different magnitudes in option price depending on the moneyness and the maturity of the

option. This added complexity makes the prediction of pricing residual more challenging.

[Insert Table 3, here]

4.1.2 Call vs put options

Table 4 reports the pricing performance of the models for call options (Panel A) and put options (Panel B) separately. It shows that the differences in pricing errors between call options and put options are almost negligible, and $\text{GPF}_f$and HBD_fremain as the best performers for both types of options. For instance, the RMSEs of GPF_f, HBD_f, GPF, and HBD for call options are 2.067, 2.131, 2.442, and 2.520, respectively, while the RMSEs for put options are 2.061, 2.079, 2.455, and 2.538. These models significantly outperform DVF for both option types, whose RMSE is 5.933 for call options and 5.592 for put options. This result suggests that machine learning can enhance the pricing performance for both option types and the performance can be further improved by incorporating firm characteristics. The smaller MAPE of put options can be attributed to the fact that the prices of out-of-the-money put options are higher than the prices of out-of-the-money call options in our sample, as shown in Table 1. Out-of-the-money options are cheaper and bear larger percentage errors. The higher prices of out-of-the-money put options result in smaller overall percentage errors.

[Insert Table 4, here]

4.1.3 Different moneyness options

It is well known that implied volatility differs across moneyness, the phenomenon known as volatility smile or smirk, and moneyness can have a significant impact on option pricing performance. For instance, in-the-money options may require larger price adjustments compared to out-of-the-money options as the former is more likely to be exercised prior to maturity, whereas the latter may be held to maturity. To examine the impact of moneyness on option pricing, we analyze the models' performances for different moneyness options.

Table 5 reports the pricing errors for each moneyness group. GPF$_f$is the best performer for most moneyness groups with the lowest RMSEs: 2.004 for DOTM, 1.943 for OTM, 2.119 for JOTM, 2.588 for ATM, 2.008 for JITM, 1.650 for ITM, 1.615 for DITM. It only ranks second for ATM options, for which HBD$_f$performs best with an RMSE of 2.556. HBD$_f$is the second-best performer in terms of RMSE. It is notable that the benchmark DVF model has a significantly larger error for DOTM options, whereas the proposed models perform consistently across all moneyness groups. Moreover, the performance improvement by firm characteristics is particularly evident when pricing out-of-the-money options. This result suggests that firm characteristics can explain volatility smile to some extent. Given the larger trading volume of out-of-the-money options (as evidenced by support) and the fact that investors commonly use these options as a form of protection due to their lower cost, accurate pricing of these options is of greater significance than that of in-the-money options. Out-of-the-money options are also more difficult to price due to volatility smile. Both GPF$_f$and HBD$_f$exhibit greater performance improvement when pricing these options.

[Insert Table 5, here]

4.1.4 Different maturity options

Table 6 reports the pricing performance of the models for different maturity groups. Again, GPF$_f$performs best in all maturity groups in terms of RMSE, followed by HBD$_f$, GPF, and HBD. The errors tend to increase with time to maturity. The RMSE of DVF increases dramatically from 3.572 (Near-term) to 12.419 (Long-term). In contrast, the proposed models perform consistently throughout the maturity groups and the error increases with maturity only moderately. For instance, the RMSE of GPF$_f$increases from 2.038 (Near-term) to 2.465 (Long-term). Moreover, firm characteristics appear to have a more significant impact on options with longer maturities. For near-term options, the RMSE of GPF is reduced by 11.93% when firm characteristics are incorporated, whereas it is reduced by 20.81% for mid-term options, and by 23.92% for long-term options. Similarly, the RMSE of HBD is

reduced by 14.91% for near-term options, by 17.72% for mid-term options, and by 21.67% for long-term options. Investors who trade long-term options are likely to trade options based on the growth perspective of the underlying firm rather than for speculation. Therefore, firm characteristics are expected to carry more explanatory power for long-term options.

[Insert Table 6, here]

4.1.5 Year-by-year performance

Table 7 reports the MAPEs of each model year by year over the out-of-sample period. We report MAPE instead of RMSE or MAE because both option prices and (absolute) errors increase with time and therefore it is difficult to compare the performance across years with absolute error metrics. As before, $\text{GPF}_f$and HBD_foutperform GPF and HBD, respectively, and DVF performs worst. The proposed models perform particularly well in comparison to DVF in recent years when there are significantly more options in the market. For instance, the MAPEs of $\text{GPF}_f$and DVF in 1998 are 0.160 and 0.306, respectively, whereas those in 2021 are 0.248 and 0.681. The improved performance in recent years can be attributed to the increased volume of the training set. Machine learning algorithms learn patterns from historical data and having a large amount of data for training is critical to avoid overfitting and make more accurate predictions.

The increasing trend of pricing errors corresponds to the growth of trading volume. As more participants engage in the options market and trading activity intensifies, prices can swing more due to more frequent changes of demand and supply dynamics. Increased liquidity also allows for quicker transactions and increased price sensitivity to new information. These factors can contribute to the increasing trend of pricing errors.

[Insert Table 7, here]

4.2 Importance of firm characteristics

CatBoost provides the importance of each input feature as a score that adds up to 100 based on their contribution to prediction. Figures 3 and 4 present the 20 most important features in $\text{GPF}_f$and HBD_f, respectively. The scores in the figures are the average scores of all monthly training results. Although not reported here, we fine that the importance scores are stable over the sample period.

Figure 3 shows that the option-related variables, σ_{imp}, K^2, K, KT, T^2, and T, are the most important features. In particular, σ_{imp} has an average score of 38, which is considerably higher than the other features' scores. This is an expected result as the target variable of the machine learning algorithm in $\text{GPF}_f$is implied volatility. The option-related variables collectively have an importance score of 94 and the rest 6 is distributed across firm characteristics. Among firm characteristics, idiosyncratic volatility (ivol_capm_252d), share price (price), market equity (market_equity), year-1-lagged annual return (seas_1_1an), illiquidity (ami_126d), and firm age (age) are some of the most important features. In particular, idiosyncratic volatility has the highest score of 0.569, which suggests that idiosyncratic volatility carries information about implied volatility. Share price is also identified as an important feature with an importance score of 0.510. This result suggests that moneyness (share price divided by strike price) alone is not sufficient to describe the role of strike price and share price in option value, but they interact in a more complex manner in determining option price.

Figure 4 reveals that the option-related variables are also the most important features in HBD_f. However, their importance scores are lower and firm characteristics play a more important role with an aggregate score of 14 in HBD_f. This is because the machine learning algorithm in HBD_fpredicts the residuals from the BS model. Share price, idiosyncratic volatility, illiquidity, market equity, and firm age (age) are the most important features among firm characteristics. The fact that these firm characteristics are identified as important features in both models implies that they are not picked by chance but do contain information

about the option price.

[Insert Figure 3, here]

[Insert Figure 4, here]

5 Robustness tests

5.1 Training models for each option group

As shown in Table 1, the distribution of the sample is highly imbalanced across different option groups, *e.g.*, there are significantly more near-term options than long-term options. The imbalance in the training set can lead to a biased result as the algorithm may prioritize minimizing the pricing error of near-term options over long-term options. If the relationship between the input and the output is different between these two option groups, training one model for all options will result in a relatively poor performance for long-term options. On the other hand, training a machine algorithm individually for each option group can suffer from a small data problem, which can cause overfitting. The small data problem can be particularly severe in the early years of the sample, where available options are significantly fewer. To test the trade-off between the data imbalance problem and the small data problem, we train the models individually for each option group and compare the results with those from the previous one-model-fits-all case. The results are presented below. Overall, it appears that the relationship between the input features and the option price is not so distinctive across different types of options to require training the models individually for each group, and training one model for all types of options is adequate.

5.1.1 Call vs put options

Table 8 reports the pricing performance of the models that are trained on call and put options individually. The proposed models yield slightly smaller errors when trained individ-

ually. For call options, the RMSEs of GPF$_f$and HBD$_f$are reduced from 2.067 to 2.011 and from 2.131 to 2.075, respectively, and for put options, the RMSEs are reduced from 2.061 to 1.996 and from 2.079 to 2.020. Although the improvements are minor, the result suggests that it is worth training the models separately for each option type.

[Insert Table 8, here]

5.1.2 Different moneyness options

Table 9 reports the pricing performance of the models that are trained on different moneyness groups individually. In some groups, *e.g.*, DOTM and ATM, training the models individually renders better results, but in other groups, the results are mixed. It appears that dividing the sample into many groups results in not enough sample size for each group and the benefit of specifying a model for each group cannot dominate the small data problem.

[Insert Table 9, here]

5.1.3 Different maturity options

Table 10 reports the pricing performance of the models that are trained on different maturity groups individually. The models perform slightly better for near-term and mid-term options, *e.g.*, the RMSEs of GPF$_f$and HBD$_f$are respectively 2.020 and 2.033 for near-term options when trained individually, whereas they are 2.038 and 2.038 when trained using the entire sample. However, the models perform worse for long-term options when trained individually. The poor performance can be attributed to the small sample size of long-term options.

[Insert Table 10, here]

6 Conclusion

This paper proposes machine learning-based option pricing models that incorporate firm characteristics. Individual stock option prices can be affected by the prospects of the underlying firm and our aim is to assess whether firm characteristics can enhance the pricing of stock options. We employ two semi-parametric models, a varient of Andreou et al. (2010)'s generalized parametric function model (GPF) and a varient of Lajbcygier and Connor (1997)'s hybrid model (HBD), and evaluate them using individual stock options in the US market in the period from 1996 to 2021. The results suggest that both GPF and HBD are effective in pricing American options and firm characteristics can significantly improve the performance of these models. Between the two models, GPF consistently performs better. Among the firm characteristics, idiosyncratic volatility, share price, market equity, illiquidity, and firm age are found to be the most important features in predicting option prices. We contribute to the option pricing literature by making the first attempt to incorporate firm characteristics into option pricing and demonstrating its effectiveness.

References

Andersen, L., Broadie, M., 2004. Primal-dual simulation algorithm for pricing multidimensional American options. *Management Science* 50, 1222–1234.

Andreou, P. C., Charalambous, C., Martzoukos, S. H., 2006. Robust artificial neural networks for pricing of European options. *Computational Economics* 27, 329–351.

Andreou, P. C., Charalambous, C., Martzoukos, S. H., 2010. Generalized parameter functions for option pricing. *Journal of Banking & Finance* 34, 633–646.

Andreou, P. C., Charalambous, C., Martzoukos, S. H., 2014. Assessing the performance of symmetric and asymmetric implied volatility functions. *Review of Quantitative Finance and Accounting* 42, 373–397.

Bakshi, G., Cao, C., Chen, Z., 1997. Empirical performance of alternative option pricing models. *The Journal of Finance* 52, 2003–2049.

Berkowitz, J., 2009. On justifications for the ad hoc Black-Scholes method of option pricing. *Studies in Nonlinear Dynamics & Econometrics* 14.

Black, F., Scholes, M., 1973. The pricing of options and corporate liabilities. *Journal of Political Economy* 81, 637–654.

Boyle, P. P., 1986. Option valuation using a tree-jump process. *International Options Journal* 3, 7–12.

Broadie, M., Glasserman, P., 1997. Pricing American-style securities using simulation. *Journal of Economic Dynamics and Control* 21, 1323–1352.

Christoffersen, P., Heston, S., Jacobs, K., 2009. The shape and term structure of the index option smirk: Why multifactor stochastic volatility models work so well. *Management Science* 55, 1914–1932.

Christoffersen, P., Jacobs, K., 2004. The importance of the loss function in option valuation. *Journal of Financial Economics* 72, 291–318.

Corrado, C. J., Su, T., 1996. Skewness and kurtosis in S&P 500 index returns implied by option prices. *Journal of Financial Research* 19, 175–192.

Cox, J. C., Ross, S. A., Rubinstein, M., 1979. Option pricing: A simplified approach. *Journal of Financial Economics* 7, 229–263.

Dumas, B., Fleming, J., Whaley, R. E., 1998. Implied volatility functions: Empirical tests. *The Journal of Finance* 53, 2059–2106.

Gençay, R., Qi, M., 2001. Pricing and hedging derivative securities with neural networks: Bayesian regularization, early stopping, and bagging. *IEEE Transactions on Neural Networks* 12, 726–734.

Geske, R., 1979. The valuation of compound options. *Journal of Financial Economics* 7, 63–81.

Gradojevic, N., Gençay, R., Kukolj, D., 2009. Option pricing with modular neural networks. *IEEE Transactions on Neural Networks* 20, 626–637.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies* 33, 2223–2273.

Hagan, P. S., Kumar, D., Lesniewski, A. S., Woodward, D. E., 2002. Managing smile risk. *The Best of Wilmott* 1, 249–296.

Hansen, P. R., Lunde, A., Nason, J. M., 2011. The model confidence set. *Econometrica* 79, 453–497.

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, 327–343.

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, 851–889.

Jensen, T. I., Kelly, B. T., Pedersen, L. H., 2021. Is there a replication crisis in finance? Tech. rep., National Bureau of Economic Research.

Lajbcygier, P. R., Connor, J. T., 1997. Improved option pricing using artificial neural networks and bootstrap methods. *International Journal of Neural Systems* 8, 457–471.

Liang, X., Zhang, H., Xiao, J., Chen, Y., 2009. Improving option price forecasts with neural networks and support vector regressions. *Neurocomputing* 72, 3055–3065.

Longstaff, F. A., Schwartz, E. S., 2001. Valuing American options by simulation: a simple least-squares approach. *The Review of Financial Studies* 14, 113–147.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A., 2018. CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, pp. 6638–6648.

Roll, R., 1977. An analytical valuation formula for unprotected american call options. *Journal of Financial Economics* 5, 251–258.

Villiger, R., 2006. Valuation of American call options. *Wilmott Magazine* 3, 64–67.

Whaley, R. E., 1981. On the valuation of American call options on stocks with known dividends. *Journal of Financial Economics* 9, 207–211.

Figures

Figure 1. GPF model structure.

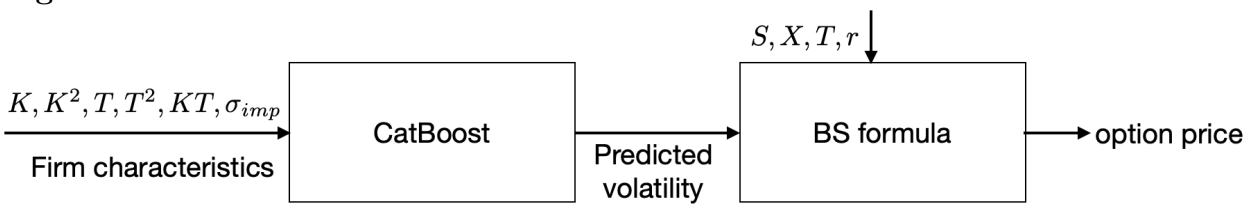


Figure 2. HBD model structure.

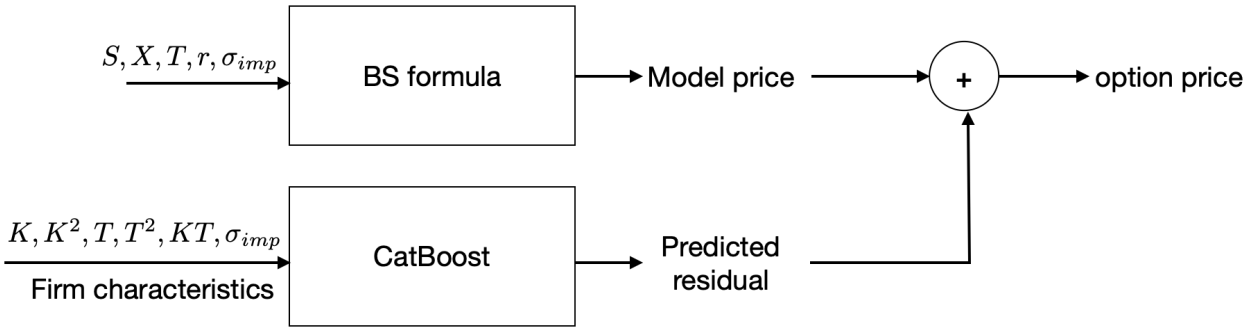


Figure 3. Feature importance score of the 20 most important features in GPF_f.

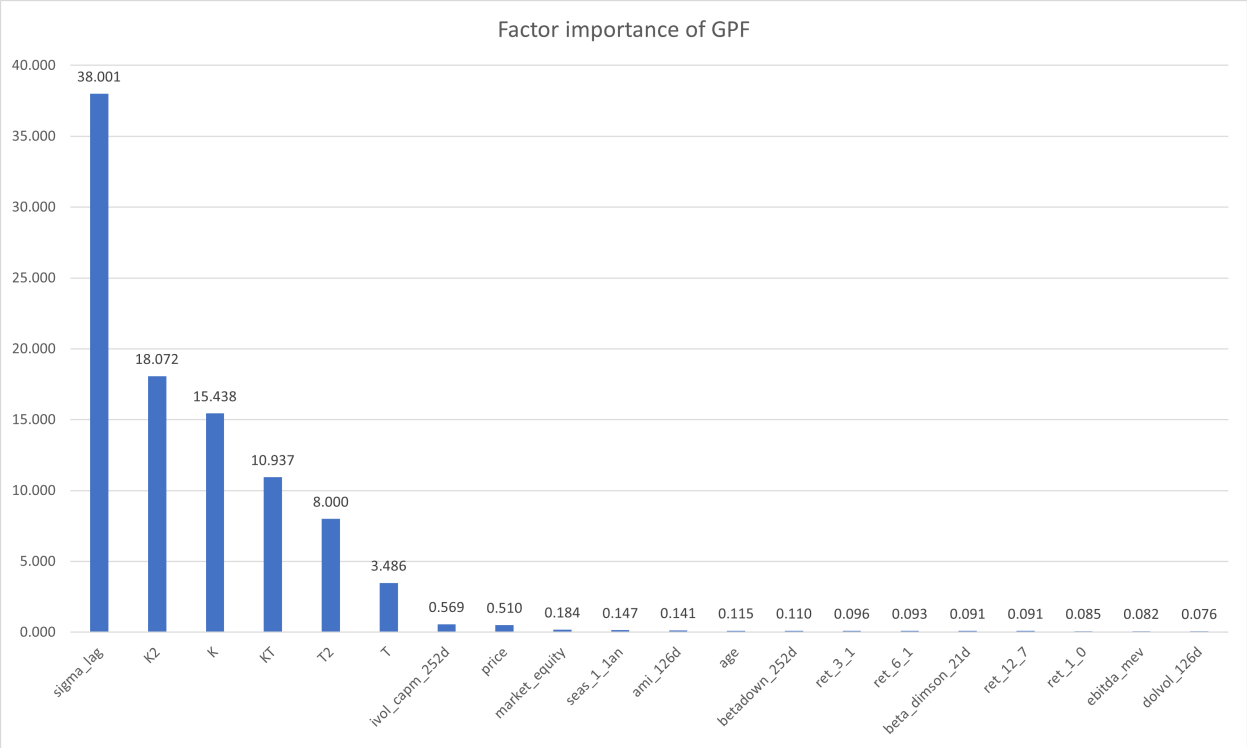
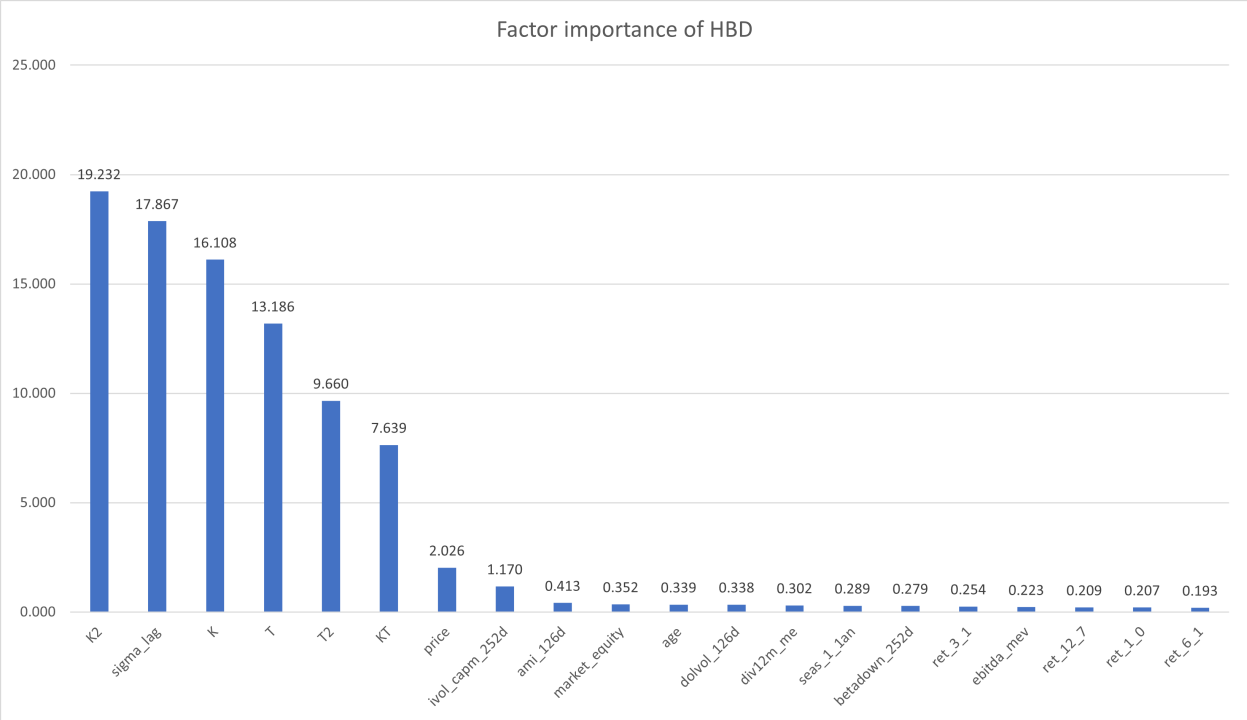


Figure 4. Feature importance score of the 20 most important features in HBD_f.



Tables

Table 1. Characteristics of option data
This table describes the characteristics of the option data. The data is obtained from OptionMetrics' IvyDB US and cover the stock options in the US market from 1996.01 to 2021.12. 'Price', 'Implied volatility', and 'Observations' respectively refer to the average price, average implied volatility, and the number of observations in each subset.

A. Call options

Moneyness	DOTM	OTM	JOTM	ATM	JITM	ITM	DITM
S/X	0.80-0.90	0.90-0.95	0.95-0.99	0.99-1.01	1.01-1.05	1.05-1.10	1.10-1.20
Near-term (6-60 days)							
Price	1.368	1.752	2.578	4.359	5.620	7.488	10.407
Implied volatility	0.857	0.550	0.391	0.346	0.415	0.566	0.813
Observations	909,057	1,245,376	1,722,750	916,423	904,543	457,659	312,472
Mid-term (60-180 days)							
Price	2.388	3.121	4.568	6.798	7.412	8.791	11.118
Implied volatility	0.484	0.361	0.325	0.327	0.353	0.404	0.493
Observations	634,289	571,393	512,475	225,180	273,762	176,100	152,272
Long-term (181-365 days)							
Price	3.979	5.029	6.864	9.321	9.818	11.090	13.623
Implied volatility	0.365	0.315	0.307	0.316	0.332	0.355	0.397
Observations	242,345	167,661	134,502	58,768	74,067	53,679	54,388

B. Put options

Moneyness	DITM	ITM	JITM	ATM	JOTM	OTM	DOTM
S/X	0.80-0.90	0.90-0.95	0.95-0.99	0.99-1.01	1.01-1.05	1.05-1.10	1.10-1.20
Near-term (6-60 days)							
Price	1.510	1.935	2.726	4.242	5.321	7.298	11.360
Implied volatility	0.782	0.543	0.400	0.353	0.426	0.597	0.921
Observations	745,650	894,463	1,175,816	655,465	628,517	289,699	165,035
Mid-term (60-180 days)							
Price	2.700	3.497	4.593	6.084	6.490	7.449	9.980
Implied volatility	0.470	0.387	0.351	0.345	0.371	0.432	0.550
Observations	364,882	288,937	281,504	143,189	180,036	107,844	84,358

Continued on the next page

Long-term (181-365 days)							
Price	4.784	5.782	7.112	8.939	8.863	9.545	12.803
Implied volatility	0.380	0.353	0.341	0.341	0.357	0.385	0.442
Observations	136,102	88,593	78,961	40,505	55,059	39,844	37,614

Table 2. Number of options per stock

This table describes the descriptive statistics of the number of options per stock in each year in the sample period.

Year	No. firms	No. options						
		Mean	Std	Min	25%	50%	75%	Max
1996	142	3	4	1	1	2	3	28
1997	163	4	5	1	1	2	4	38
1998	175	4	6	1	1	2	5	35
1999	185	5	6	1	1	2	5	41
2000	197	5	7	1	1	2	6	42
2001	206	5	6	1	1	2	6	33
2002	206	5	6	1	1	3	6	33
2003	226	5	6	1	1	3	7	31
2004	258	5	6	1	1	3	7	32
2005	282	6	6	1	1	3	8	39
2006	314	7	7	1	1	3	9	51
2007	350	7	8	1	1	4	9	64
2008	344	7	9	1	2	4	9	69
2009	336	7	10	1	2	4	9	69
2010	328	8	10	1	1	4	10	78
2011	335	9	14	1	1	4	11	152
2012	308	9	18	1	2	4	11	251
2013	338	9	17	1	1	4	10	225
2014	356	9	18	1	1	3	9	256
2015	349	8	19	1	1	3	7	246
2016	353	8	16	1	1	3	7	180
2017	396	8	17	1	1	3	7	177
2018	445	9	21	1	1	3	7	212
2019	448	9	21	1	1	3	7	207
2020	497	12	28	1	1	3	9	317
2021	615	13	30	1	1	3	9	339

Table 3. Pricing performance for all options

This table presents the out-of-sample pricing errors of each model for all the options in the sample. The out-of-sample period is from 1998.01 to 2021.12. 'RMSE', 'MAE', and 'MAPE' respectively refer to the root mean square error, the mean absolute error, and the mean absolute percentage error, and 'MCS-p' refers to the p-value of the model confidence set test. The details of each model can be found in Section 3.3.

	GPF_f	HBD_f	GPF	HBD	DVF	Support
RMSE	2.064	2.110	2.447	2.527	5.799	16,050,622
MAE	0.593	0.601	0.671	0.679	1.337	
MAPE	0.219	0.226	0.248	0.237	0.494	
MCS-p	.0001	0.016	0.000	0.000	0.000	

Table 4. Pricing performance for call and put options

This table presents the out-of-sample pricing errors of each model for call and put options. The out-of-sample period is from 1998.01 to 2021.12. 'RMSE', 'MAE', and 'MAPE' respectively refer to the root mean square error, the mean absolute error, and the mean absolute percentage error, and 'MCS-p' refers to the p-value of the model confidence set test. The details of each model can be found in Section 3.3.

	GPF_f	HBD_f	GPF	HBD	DVF	Support
A. Call options						
RMSE	2.067	2.131	2.442	2.520	5.933	9,631,300
MAE	0.591	0.598	0.673	0.683	1.336	
MAPE	0.228	0.235	0.261	0.251	0.517	
MCS-p	1.000	0.008	0.000	0.000	0.000	
B. Put options						
RMSE	2.061	2.079	2.455	2.538	5.592	6,419,322
MAE	0.596	0.607	0.669	0.672	1.339	
MAPE	0.206	0.213	0.227	0.216	0.459	
MCS-p	1.000	0.008	0.000	0.000	0.000	

Table 5. Pricing performance for different moneyness options

This table presents the out-of-sample pricing errors of each model for different moneyness options. The out-of-sample period is from 1998.01 to 2021.12. 'RMSE', 'MAE', and 'MAPE' respectively refer to the root mean square error, the mean absolute error, and the mean absolute percentage error, and 'MCS-p' refers to the p-value of the model confidence set test. The details of each model can be found in Section 3.3.

Model	GPF_f	HBD_f	GPF	HBD	DVF	Support
A. DOTM						
RMSE	2.004	2.108	2.418	2.420	8.483	2,993,209
MAE	0.570	0.593	0.672	0.658	1.811	
MAPE	0.350	0.372	0.439	0.403	1.002	
MCS-p	1.000	0.01	0.000	0.000	0.000	
B. OTM						
RMSE	1.943	2.010	2.345	2.390	3.905	3,207,739
MAE	0.564	0.579	0.656	0.654	1.040	
MAPE	0.300	0.313	0.344	0.325	0.523	
MCS-p	1.000	0.009	0.000	0.000	0.000	
C. JOTM						
RMSE	2.119	2.138	2.544	2.627	5.029	3,855,525
MAE	0.615	0.619	0.703	0.713	1.260	
MAPE	0.230	0.232	0.245	0.241	0.491	
MCS-p	1.000	0.010	0.000	0.000	0.000	
D. ATM						
RMSE	2.588	2.556	2.995	3.234	6.395	2,016,292
MAE	0.745	0.735	0.815	0.859	1.561	
MAPE	0.159	0.157	0.154	0.159	0.354	
MCS-p	1.000	0.007	0.000	0.000	0.000	
E. JITM						
RMSE	2.008	2.048	2.333	2.407	5.197	2,082,314
MAE	0.581	0.586	0.642	0.651	1.224	
MAPE	0.093	0.093	0.094	0.094	0.196	
MCS-p	1.000	0.006	0.000	0.000	0.000	

Continued on the next page

F. ITM						
RMSE	1.650	1.726	1.889	1.933	3.722	1,100,709
MAE	0.490	0.498	0.527	0.534	0.970	
MAPE	0.062	0.062	0.062	0.062	0.124	
MCS-p	1.000	0.006	0.000	0.000	0.000	
G. DITM						
RMSE	1.615	1.743	1.839	1.882	5.520	787,063
MAE	0.476	0.490	0.497	0.508	1.372	
MAPE	0.047	0.048	0.046	0.046	0.122	
MCS-p	1.000	0.009	0.000	0.000	0.000	

Table 6. Pricing performance for different maturity options

This table presents the out-of-sample pricing errors of each model for different maturity options. The out-of-sample period is from 1998.01 to 2021.12. 'RMSE', 'MAE', and 'MAPE' respectively refer to the root mean square error, the mean absolute error, and the mean absolute percentage error, and 'MCS-p' refers to the p-value of the model confidence set test. The details of each model can be found in Section 3.3.

	GPF_f	HBD_f	GPF	HBD	DVF	Support
A. Near-term						
RMSE	2.038	2.038	2.314	2.395	3.572	10,855,219
MAE	0.592	0.593	0.653	0.657	1.095	
MAPE	0.249	0.257	0.282	0.266	0.499	
MCS-p	1.000	0.008	0.000	0.000	0.000	
B. Mid-term						
RMSE	1.987	2.112	2.509	2.567	7.275	3,928,374
MAE	0.548	0.569	0.651	0.660	1.543	
MAPE	0.159	0.164	0.178	0.177	0.467	
MCS-p	1.000	0.009	0.000	0.000	0.000	
C. Long-term						
RMSE	2.465	2.643	3.240	3.374	12.419	1,246,944
MAE	0.732	0.773	0.891	0.920	2.791	
MAPE	0.147	0.154	0.165	0.166	0.526	
MCS-p	1.000	0.013	0.000	0.000	0.000	

Table 7. Pricing performance in each year

This table presents the MAPE of each model year by year. The out-of-sample period is from 1998.01 to 2021.12. 'RMSE', 'MAE', and 'MAPE' respectively refer to the root mean square error, the mean absolute error, and the mean absolute percentage error, and 'MCS-p' refers to the p-value of the model confidence set test. The details of each model can be found in Section 3.3.

Year	GPF_f	HBD_f	GPF	HBD	DVF	Support
1998	0.160	0.160	0.165	0.164	0.306	186,489
1999	0.138	0.143	0.145	0.140	0.268	213,207
2000	0.146	0.143	0.144	0.146	0.267	247,356
2001	0.156	0.158	0.152	0.156	0.357	239,132
2002	0.163	0.170	0.164	0.165	0.406	259,130
2003	0.168	0.166	0.157	0.155	0.580	295,923
2004	0.147	0.146	0.146	0.145	0.592	354,210
2005	0.158	0.158	0.167	0.166	0.407	414,818
2006	0.155	0.157	0.166	0.167	0.320	518,190
2007	0.170	0.167	0.174	0.174	0.352	610,940
2008	0.168	0.176	0.178	0.174	0.354	627,255
2009	0.168	0.177	0.161	0.158	0.472	639,593
2010	0.175	0.181	0.169	0.165	0.798	652,750
2011	0.178	0.186	0.202	0.193	0.481	761,466
2012	0.195	0.203	0.222	0.212	0.461	700,098
2013	0.213	0.221	0.247	0.238	0.554	763,040
2014	0.218	0.216	0.253	0.239	0.514	784,004
2015	0.229	0.239	0.267	0.259	0.398	740,237
2016	0.267	0.272	0.297	0.287	0.469	712,374
2017	0.268	0.272	0.313	0.299	0.619	805,788
2018	0.250	0.258	0.310	0.297	0.412	1,018,956
2019	0.312	0.321	0.368	0.360	0.494	1,070,071
2020	0.278	0.295	0.322	0.298	0.446	1,491,641
2021	0.248	0.259	0.290	0.263	0.681	1,943,954

Table 8. Pricing performance of individual models for call and put options
This table presents the out-of-sample pricing errors of the models that are trained on call and put options individually. The out-of-sample period is from 1998.01 to 2021.12. 'RMSE', 'MAE', and 'MAPE' respectively refer to the root mean square error, the mean absolute error, and the mean absolute percentage error, and 'MCS-p' refers to the p-value of the model confidence set test. The details of each model can be found in Section 3.3.

	GPF_f	HBD_f	GPF	HBD	DVF	Support
A. Call options						
RMSE	2.011	2.075	2.450	2.532	5.933	9,631,300
MAE	0.585	0.591	0.679	0.685	1.336	
MAPE	0.228	0.229	0.258	0.248	0.517	
MCS-p	1.000	0.008	0.000	0.000	0.000	
B. Put options						
RMSE	1.996	2.020	2.465	2.550	5.592	6,419,322
MAE	0.586	0.600	0.670	0.675	1.339	
MAPE	0.200	0.210	0.225	0.216	0.459	
MCS-p	1.000	0.008	0.000	0.000	0.000	

Table 9. Pricing performance of individual models for different moneyness options
This table presents the out-of-sample pricing errors of the models that are trained on different moneyness options individually. The out-of-sample period is from 1998.01 to 2021.12. 'RMSE', 'MAE', and 'MAPE' respectively refer to the root mean square error, the mean absolute error, and the mean absolute percentage error, and 'MCS-p' refers to the p-value of the model confidence set test. The details of each model can be found in Section 3.3.

	GPF_f	HBD_f	GPF	HBD	DVF	Support
A. DOTM						
RMSE	1.995	2.026	2.368	2.437	8.483	2,993,209
MAE	0.568	0.561	0.655	0.658	1.811	
MAPE	0.376	0.365	0.419	0.397	1.002	
MCS-p	1.000	0.163	0.000	0.000	0.000	
B. OTM						
RMSE	1.986	1.988	2.323	2.389	3.905	3,207,739
MAE	0.564	0.557	0.647	0.648	1.040	
MAPE	0.305	0.299	0.334	0.321	0.523	
MCS-p	1.000	0.884	0.000	0.000	0.000	
C. JOTM						
RMSE	2.117	2.182	2.553	2.577	5.029	3,855,525
MAE	0.605	0.605	0.698	0.695	1.260	
MAPE	0.225	0.222	0.239	0.233	0.491	
MCS-p	1.000	0.004	0.000	0.000	0.000	
D. ATM						
RMSE	2.447	2.485	2.978	3.019	6.395	2,016,292
MAE	0.701	0.698	0.812	0.807	1.561	
MAPE	0.148	0.144	0.153	0.149	0.354	
MCS-p	1.000	0.204	0.000	0.000	0.000	
E. JITM						
RMSE	2.023	2.035	2.376	2.437	5.197	2,082,314
MAE	0.570	0.581	0.634	0.641	1.224	
MAPE	0.089	0.090	0.091	0.091	0.196	
MCS-p	1.000	0.801	0.000	0.000	0.000	
F. ITM						

Continued on the next page

RMSE	1.702	1.709	1.940	2.025	3.722	1,100,709
MAE	0.467	0.471	0.512	0.529	0.970	
MAPE	0.057	0.057	0.058	0.059	0.124	
MCS-p	1	0.819	0.000	0.000	0.000	
G. DITM						
RMSE	1.601	1.645	1.897	2.051	5.520	787,063
MAE	0.435	0.446	0.476	0.510	1.372	
MAPE	0.042	0.043	0.042	0.044	0.122	
MCS-p	1.000	0.058	0.000	0.000	0.000	

Table 10. Pricing performance of individual models for different maturity options
This table presents the out-of-sample pricing errors of the models that are trained on different maturity options individually. The out-of-sample period is from 1998.01 to 2021.12. 'RMSE', 'MAE', and 'MAPE' respectively refer to the root mean square error, the mean absolute error, and the mean absolute percentage error, and 'MCS-p' refers to the p-value of the model confidence set test. The details of each model can be found in Section 3.3.

	GPF_f	HBD_f	GPF	HBD	DVF	Support
A. Near-term						
RMSE	2.020	2.033	2.316	2.400	3.572	10,855,219
MAE	0.589	0.589	0.650	0.652	1.095	
MAPE	0.255	0.255	0.279	0.281	0.499	
MCS-p	1.000	0.008	0.000	0.000	0.000	
B. Mid-term						
RMSE	1.966	2.003	2.513	2.560	7.275	3,928,374
MAE	0.546	0.550	0.649	0.649	1.543	
MAPE	0.155	0.160	0.174	0.172	0.467	
MCS-p	1.000	0.009	0.000	0.000	0.000	
C. Long-term						
RMSE	2.643	2.774	3.167	3.359	12.419	1,246,944
MAE	0.753	0.765	0.885	0.925	2.791	
MAPE	0.147	0.148	0.165	0.167	0.526	
MCS-p	1.000	0.073	0.001	0.001	1.000	

Appendices

A. Firm characteristics

Abbreviation	Description	Author	Year	Journal
age	Firm age	Jiang, Lee, and Zhang	2005	RAS
aliq_at	Asset liquidity to book assets	Ortiz-Molina and Phillips	2014	JFQA
ami_126d	Illiquidity	Amihud	2002	JFM
at_be	Book leverage	Fama and French	1992	JF
at_gr1	Asset growth	Cooper, Gulen, and Schill	2008	JF
at_me	Assets-to-market	Fama and French	1992	JF
at_turnover	Capital turnover	Haugen and Baker	1996	JFE
be_gr1a	Chage in common equity	Richardson et al.	2005	JAE
be_me	Book-to-market	Rosenberg, Reid, and Lanstein	1985	JF
beta_dimson_21d	Dimson Beta	Dimson	1979	JFE
betadown_252d	Downside beta	Ang, Chen, and Xing	2006	RFS
bev_mev	Book-to-market enterprise value	Penman, Richardson, and Tuna	2007	JAR
bidaskhl_21d	High-low bid-ask spread	Corwin and Schultz	2012	JF
capx_gr1	CAPEX growth	Xie	2001	AR
cash_at	Cash-to-assets	Palazzo	2012	JFE
chcsho_12m	Net stock issues	Pontiff and Woodgate	2008	JF
coa_gr1a	Change in current operating assets	Richardson et al.	2005	JAE
col_gr1a	Change in current Ooperating liabilities	Richardson et al.	2005	JAE
cop_at	Cash-based operating profitablility	Ball et al.	2016	JFE
cop_atl1	Cash-based operating profits to lagged assets	Ball et al.	2016	JFE
coskew_21d	Coskewness	Harvey and Siddique	2000	JF
cowc_gr1a	Change in net non-cash working capital	Richardson et al.	2005	JAE
dbnetis_at	Net debt finance	Bradshaw, Richardson, and Sloan	2006	JAE
debt_me	Debt to market	Bhandari	1988	JFE
dgp_dsale	Gross margin growth to sales growth	Abarbanell and Bushee	1998	AR
div12m_me	Dividend yield	Litzenberger and Ramaswamy	1979	JF
dolvol_126d	Dollar trading volume	Brennan, Chordia, and Subrahmanyam	1998	JFE
dolvol_var_126d	Volatility of dollar trading volume	Chordia, Subrahmanyam, and Anshuman	2001	JFE

Continued on the next page

dsale_drec	Sales growth to receivable growth	Abarbanell and Bushee	1998	AR
ebit_bev	Return on net operating assets	Soliman	2008	AR
ebit_sale	Profit margin	Soliman	2008	AR
ebitda_mev	Enterprise multiple	Loughran and Wellman	2011	JFQA
emp_gr1	Employment growth	Belo, Lin, and Bazdresch	2014	JPE
eqnetis_at	Net equity finance	Bradshaw, Richardson, and Sloan	2006	JAE
eqnpo_12m	Composite equity issuance	Daniel and Titman	2006	JF
eqnpo_me	Net payout yield	Boudoukh et al.	2007	JF
eqpo_me	Payout yield	Boudoukh et al.	2007	JF
f_score	Piotroski F-score	Piotroski	2000	AR
fcf_me	Cash flow-to-price	Lakonishok, Shleifer, and Vishny	1994	JF
fnl_gr1a	Change in financial liabilities	Richardson et al.	2005	JAE
gp_at	Gross profits-to-assets	Novy-Marx	2013	JFE
gp_atl1	Gross profits-to-lagged assets	Novy-Marx	2013	JFE
inv_gr1a	Inventory change	Thomas and Zhang	2002	RAS
iskew_capm_21d	Idiosyncratic skewness (CAPM)	Bali, Engle, and Murray	2016	BOOK
iskew_ff3_21d	Idiosyncratic skewness (FF3)	Bali, Engle, and Murray	2016	BOOK
iskew_hxz4_21d	Idiosyncratic skewness (q-factor)	Bali, Engle, and Murray	2016	BOOK
ivol_capm_21d	Idiosyncratic volatility (CAPM)	Ang et al.	2006	JF
ivol_capm_252d	Idiosyncratic volatility	Ali, Hwang, and Trombley	2003	JFE
ivol_ff3_21d	Idiosyncratic volatility (FF3)	Ang et al.	2006	JF
ivol_hxz4_21d	Idiosyncratic volatility (q-factor)	Ang et al.	2006	JF
lnoa_gr1a	Change in long-term net operating assets	Fairfield, Whisenant, and Yohn	2003	AR
lti_gr1a	Chagne in long-term investments	Richardson et al.	2005	JAE
market_equity	Market equity	Banz	1981	JFE
mispricing_mgmt	Mispricing factor: Management	Stambaugh and Yuan	2016	RFS
mispricing_perf	Mispricing factor: Performance	Stambaugh and Yuan	2016	RFS
ncoa_gr1a	Change in non-current operating assets	Richardson et al.	2005	JAE
ncol_gr1a	Change in non-current operating liabilities	Richardson et al.	2005	JAE
netdebt_me	Net debt-to-price	Penman, Richardson, and Tuna	2007	JAR
netis_at	Net external finance	Bradshaw, Richardson, and Sloan	2006	JAE
nfna_gr1a	Change in net financial assets	Richardson et al.	2005	JAE
ni_be	Return on equity	Haugen and Baker	1996	JFE
ni_inc8q	Number of consecutive quarters with earnings in...	Barth, Elliott, and Finn	1999	JAR

Continued on the next page

ni_me	Earnings to price	Basu	1983	JFE
niq_at	Quarterly return on assets	Balakrishnan, Bartov, and Faurel	2010	JAE
niq_at_chg1	Change in quarterly return on assets	Balakrishnan, Bartov, and Faurel	2010	JAE
niq_be	Return on equity (quarterly)	Hou, Xue, and Zhang	2015	RFS
niq_be_chg1	Change in quarterly return on equity	Balakrishnan, Bartov, and Faurel	2010	JAE
niq_su	Earnings surprise	Foster, Olsen, and Shevlin	1984	AR
nncoa_gr1a	Change in net non-current operating assets	Richardson et al.	2005	JAE
noa_at	Net operating assets	Hirshleifer et al.	2004	JAE
noa_gr1a	Change in net operating assets	Hirshleifer et al.	2004	JAE
oaccruals_at	Operating accruals	Sloan	1996	AR
oaccruals_ni	Percent operating accruals	Hafzalla, Lundholm, and Van Winkle	2011	AR
ocf_at	Operating cash flow to assets	Bouchard et al.	2019	JF
ocf_at_chg1	Change in operating cash flow to assets	Bouchard et al.	2019	JF
ocf_me	Operating Cash flows to price	Desai, Rajgopal, and Venkatachalam	2004	AR
op_at	Operating profits-to-assets	Ball et al.	2016	JFE
op_atl1	Operating profits-to-lagged assets	Ball et al.	2016	JFE
opex_at	Operating leverage	Novy-Marx	2011	JFE
prc_highprc_252d	52-week high	George and Hwang	2004	JF
price	Share price	Miller and Scholes	1982	JPE
qmj	Quality minus Junk: Composite	Assness, Frazzini, and Pedersen	2018	RAS
qmj_growth	Quality minus Junk: Growth	Assness, Frazzini, and Pedersen	2018	RAS
qmj_prof	Quality minus Junk: Profitability	Assness, Frazzini, and Pedersen	2018	RAS
qmj_safety	Quality minus Junk: Safety	Assness, Frazzini, and Pedersen	2018	RAS
resff3_12_1	12 month residual momentum	Blitz, Huij, and Martens	2011	JEF
resff3_6_1	6 month residual momentum	Blitz, Huij, and Martens	2011	JEF
ret_1_0	Short-term reversal	Jegadeesh	1990	JF
ret_12_1	Momentum (12 month)	Jegadeesh and Titman	1993	JF
ret_12_7	Intermediate momentum (7-12)	Novy-Marx	2012	ROF
ret_3_1	Momentum (3 month)	Jegadeesh and Titman	1993	JF
ret_6_1	Momentum (6 month)	Jegadeesh and Titman	1993	JF
ret_9_1	Momentum (9 month)	Jegadeesh and Titman	1993	JF

Continued on the next page

rmax1_21d	Maximum daily return	Bali, Cakici, and Whitelaw	2011	JFE
rmax5_21d	Highest 5 days of return	Bali, Brown, and Tang	2017	JFE
rmax5_rvol_21d	Highest 5 days of return to volatility	Assness et al.	2020	JFE
rskew_21d	Return skewness	Bali, Engle, and Murray	2016	BOOK
rvol_21d	Return volatility	Ang et al.	2006	JF
sale_bev	Asset turnover	Soliman	2008	AR
sale_gr1	Annual sales growth	Lakonishok, Shleifer, and Vishny	1994	JF
sale_me	Sales to price	Barbee, Mukherji, and Raines	1996	FAJ
saleq_su	Revenue surprise	Jegadeesh and Livnat	2006	JFE
seas_1_1an	Year 1-lagged return, annual	Heston and Sadka	2008	JFE
seas_1_1na	Year 1-lagged return, nonannual	Heston and Sadka	2008	JFE
sti_gr1a	Change in short-term investments	Richardson et al.	2005	JAE
taccruals_at	Total accruals	Richardson et al.	2005	JAE
taccruals_ni	Percent total accruals	Hafzalla, Lundholm, and Van Winkle	2011	AR
tangibility	Tangibility	Hahn and Lee	2009	JF
tax_gr1a	Tax expense surprise	Thomas and Zhang	2011	JAR
turnover_126d	Share turnover	Datar, Naik, and Radcliffe	1998	JFM
turnover_var_126d	Volatility of share turnover	Chordia, Subrahmanyam, and Anshuman	2001	JFE
zero_trades_126d	Zero-trading days (6 months)	Liu	2006	JFE
zero_trades_21d	Zero-trading days (1 month)	Liu	2006	JFE
zero_trades_252d	Zero-trading days (12 months)	Liu	2006	JFE

B. CatBoost

The machine learning algorithm we employ is CatBoost proposed by Prokhorenkova, Gusev, Vorobev, Dorogush, and Gulin (2018). CatBoost is an ordered boosting algorithm based on a gradient boosting decision tree. The CatBoost algorithm starts by initializing a set of decision trees, known as the base predictors. In each iteration of the algorithm, a new decision tree is added to the set of base predictors, taking into account the residuals from the previous iteration. The residuals are computed as the difference between the true target values and the predictions from the current set of decision trees. The algorithm continues to add decision trees until a stopping criterion is met, such as reaching a maximum number of trees or a minimum improvement in the residuals.

CatBoost is advantageous for regression problems due to its ability to handle data with different data types and missing values in a way that does not require preprocessing or feature engineering. Additionally, CatBoost uses an ordered boosting technique, which is known to be effective for regression problems as it allows for quick convergence and reduces overfitting. The use of oblivious decision trees as the basic predictor also contributes to CatBoost's performance as these trees are able to capture complex nonlinear relationships in the data. These features combined together make CatBoost an attractive choice for regression problems, particularly when working with datasets that are challenging to preprocess or engineer. The core concepts used in CatBoost include the following:

- **Gradient boosting:** CatBoost is based on gradient boosting, a technique that uses an ensemble of weak learners (usually decision trees) to create a strong model. Gradient boosting is used to minimize the loss function by updating the model iteratively based on the gradient of the loss.
- **Decision trees:** CatBoost uses decision trees as the base model in its ensemble. Decision trees split the feature space into regions and make predictions based on the observations in those regions.
- **Oblivious trees:** CatBoost uses oblivious trees, a type of decision tree that is not affected

by the order of the features. This helps to reduce the correlation between trees and make the predictions more stable.

- **Gradients and loss function:** CatBoost calculates gradients for each observation based on the current prediction and actual value, and updates the model to minimize the loss function. The loss function measures the difference between the prediction and actual value and is used to determine how much the model should be updated at each iteration.
- **Boosting iterations:** CatBoost iteratively updates the model by adding trees that fit the residuals until the loss function reaches a minimum. The final prediction is made by combining the predictions from all the trees in the ensemble.

The process of CatBoost regression is as follows. Let (X, y) denote the training set, where X is a $N \times K$ matrix of input features for N samples and K features, and y is a N-dimensional vector of target values. The objective is to find a function $F(X)$ that minimizes the cost function $J(y, F(X))$, which is the root mean square error in our task. In each iteration $t,\ t = 1, 2, ..., T$, the algorithm trains a new decision tree $g_t(X)$ to approximate the negative gradient of the loss function with respect to the current model $F_{t-1}(X)$. The model is then updated by the following equation:

$$F_t(X) = F_{t-1}(X) + \eta g_t(X), \quad t > 1, \tag{8}$$

where η is a learning rate. When $t = 1$, $F_1(X)$ is set to be $g_1(X)$.