Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
FINANCE

MASTER´S FINAL WORK
DISSERTATION

CONFORMAL PREDICTION OF OPTION PRICES WITH MACHINE LEARNING

BEATRIZ PINHEIRO TOMAZ DUARTE LEITE

JUNE - 2025

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
FINANCE

MASTER´S FINAL WORK
DISSERTATION

CONFORMAL PREDICTION OF OPTION PRICES WITH MACHINE LEARNING

BEATRIZ PINHEIRO TOMAZ DUARTE LEITE

SUPERVISION:
RAQUEL M. GASPAR
JOÃO A. BASTOS

JUNE - 2025

To my family, for giving me
the strength to face this
journey and the conscience
to appreciate the privilege it
is to walk this path.

ACKNOWLEDGMENTS

This dissertation marks the end of a two-year academic journey and represents the culmination of six months of hard work. Even though this is a personal achievement, it would not have been possible without the help of those close to me. Therefore, I would like to express my gratitude:

To my parents, Maria and Alberto, for investing in my education and always supporting me.

To my grandmother, Edite, for the affection she gives me and to the memory of my grandfather, Alberto, who always believed in my potential.

To my uncles, Roberto and Sérgio, for their guidance and for being a source of inspiration and an example to follow.

To my boyfriend, Miguel, whose support gave me the strength to always keep going, and for making this journey a lot more fun.

To my friends, who heard me saying “I can’t, I have to work on my thesis” a lot in the past months – this will be compensated in the future.

And to my supervisors, Professor Raquel and Professor João, for their constant guidance and advice throughout the stages of this process.

ABSTRACT

The acknowledgement and quantification of uncertainty in option pricing represents a critical challenge in the field of finance that has been largely underexplored. The present study proposes to quantify the uncertainty of option pricing by using conformal prediction, thus aiming to fill a gap in existent literature. Conformal prediction is a technique for constructing prediction intervals with valid coverage in finite samples without making distributional assumptions. Using a large dataset of call and put options on the S&P500, we conduct an empirical study to evaluate the performance of conformal prediction intervals for gradient boosting machines. The empirical results indicate that the prediction intervals reach an empirical coverage equal to the nominal target, which is not observed in non-conformal methodologies. Furthermore, we observe systematic variations in the width of the intervals across option characteristics. Notably, out-of-the-money options and options with a short time-to-maturity have relatively wider prediction intervals, suggesting higher pricing uncertainty. We also observe that short-term put options have wider intervals than short-term call options, due to their inherent payoff differences. Overall, the findings validate the use of conformal prediction in the field of option pricing and highlight its practical value for financial decision-making under uncertainty.

KEYWORDS: Conformal Prediction; Option Pricing; Machine Learning; Quantile Regression; LightGBM.

JEL CODES: C13; C15; C31; G13; G17.

RESUMO

O reconhecimento e a quantificação da incerteza na determinação do preço das opções representam um desafio critico na área das finanças que tem sido largamente subexplorado. O presente estudo propõe-se quantificar a incerteza na determinação do preço das opções recorrendo à metodologia de previsão conformalizada, procurando assim colmatar uma lacuna existente na literatura. A previsão conformalizada é uma técnica que permite construir intervalos de previsão com cobertura válida em amostras finitas, sem fazer suposições sobre a distribuição da amostra. Recorrendo a uma vasta base de dados de opções no S&P500, realizamos um estudo empírico para avaliar o desempenho de intervalos de previsão conformalizados para *gradient boosting machines*. Os resultados empíricos indicam que os intervalos atingem uma cobertura empírica igual à cobertura nominal, o que não é assegurado em metodologias não conformes. Além disso, observamos variações sistemáticas na largura dos intervalos consoante as características das opções. Nomeadamente, as opções out-of-the-money e as opções com um curto período até à maturidade têm intervalos de previsão relativamente mais largos, o que sugere uma maior incerteza na determinação do preço. Observamos também que as *calls* com curto período até à maturidade têm intervalos mais largos do que as *puts* com curto período até à maturidade, devido às diferenças inerentes de retorno. Em suma, os resultados validam a utilização da previsão conformalizada no âmbito da previsão de preços de opções financeiras e salientam o seu valor prático para a tomada de decisões financeiras em condições de incerteza.

PALAVRAS-CHAVE: Previsão Conformalizada; Precificação de opções; Machine Learning; Regressão de Quantis; LightGBM.

CÓDIGOS JEL: C13; C15; C31; G13; G17.

GLOSSARY

ATM – At-the-money.

CQR – Conformal Quantile Regression.

EFB – Exclusive Feature Bundling.

GOSS – Gradient-based One-side Sampling.

IV – Implied Volatility.

ITM – In-the-money.

LightGBM – Light Gradient Boosting Machine.

LT – Long Term.

ML – Machine Learning.

MT – Medium Term.

NQR – Non-conformal Quantile Regression.

OTM – Out-of-the-money.

ST – Short Term.

XGBoost – Extreme Gradient Boosting Machine.

TABLE OF CONTENTS

List of Figures

List of Tables

1. INTRODUCTION

In financial markets, the use of derivatives broadly increased in recent years. An option is defined as a contract between two parties that gives the buyer the right, but not the obligation, to buy or sell a particular asset in the future at a certain predefined price (Bouzoubaa and Osseiran, 2010).

The accurate pricing of options represents a fundamental element of risk management and trading strategies. Traditionally, option pricing has been based on models such as the Black-Scholes model (Black & Scholes, 1973). However, the model relies on several assumptions that do not align with real-world market conditions, including constant volatility, interest rates and dividend yields, as well as the log-normally distributed stock prices and the no-arbitrage assumption. Considering the above, the emergence of machine learning presented an opportunity to explore a novel approach to options pricing. The seminal contributions of Hutchinson et al. (1994) and Malliaris & Salchenberger (1993) established the foundations for the development of more sophisticated machine learning option pricing in contemporary research. Neural Networks are the most frequently employed algorithms for predicting option prices (e.g. Fang & George, 2017; Gan & Liu, 2024; Gaspar et al., 2020; Umeorah et al., 2023). Other machine learning algorithms, such as tree-based ensembles, are also widely used (e.g. Ech-Chafiq et al., 2023; Shubham et al., 2023).

Even though these algorithms yield good results, often outperforming the accuracy of traditional pricing models, the element of uncertainty associated with option price predictions has been predominantly disregarded in literature. From the risk management point of view, it seems relevant to be able to quantify that uncertainty, particularly in high-stakes decisions such as option trading. This can be achieved through prediction intervals, using upper and lower bounds that encompass the response variable with high probability. According to Romano et al. (2019), the ideal prediction intervals should meet two criteria: first, given a significance level α, the prediction interval should provide a coverage of $1 - \alpha$ in finite samples without making strong distributional assumptions. Second, the intervals should be as narrow as possible, to ensure that the predictions are informative.

We employ a methodology of conformal prediction, that ensures that the empirical coverage level of the intervals closely aligns with the nominal target. To assess the level of uncertainty of the predictions, we can observe how the relative width of the intervals varies for different regions of the regressor space. A narrower interval suggests a more confident prediction, keeping the empirical coverage level at the nominal target.

Moreover, we can analyse how the option characteristics influence uncertainty, namely the moneyness and time to maturity of the options. The objective of this dissertation is, therefore, to quantify the uncertainty of option pricing by using conformal prediction, thus aiming to fill a gap in existent literature. This study is one of the first to use conformal prediction in the field of option pricing, building upon the work of Bastos (2024).

This dissertation is structured as follows. The next section introduces the state-of-the-art in constructing prediction intervals as a measure of option price uncertainty, followed by a discussion on different machine learning algorithms that can be employed for that use. In Section 3, the conformal prediction framework is explained in detail, as well as the predictive model chosen for that purpose. Section 4 explains the data selection process and its descriptive statistics. In Section 5, we present and discuss the results. Finally, Section 6 summarizes the key findings from the study and suggests further developments.

2. LITERATURE REVIEW

2.1. Prediction Intervals

The issue of quantifying uncertainty in option price predictions can be addressed using prediction intervals. Various approaches have been proposed and developed for the construction of prediction intervals in the field of statistics, which can then be applied to several areas.

A straightforward approach to construct prediction intervals is to adopt the split conformal prediction method, introduced by Papadopoulos et al. (2002). This method involves splitting the data into training, calibration and test sets. The training set is used to train the model, while the calibration set is used to compute "conformity scores". The test data is used to create and evaluate the quality of the intervals. A regression model is trained on the test data and for each observation the conformity scores are computed,

using the absolute residuals. However, this method has a major limitation: it produces intervals with a fixed length, which does not accurately reflect the varying widths of intervals observed empirically for option prices, as noted by Bastos (2024).

To address the aforementioned limitation, Papadopoulos et al. (2011, 2008) proposed a locally adaptive approach to make conformal prediction adaptive to heteroskedasticity. This method can generate non-constant and narrower intervals by adapting to local properties of the data. However, Romano et al. (2019) identified several limitations to this approach, namely the inflation of the prediction intervals when the data is homoscedastic and the underestimation of the prediction error, thus resulting in some loss of adaptivity of the intervals.

Considering the limitations of these methods, we use the method of conformal quantile prediction in this study, as proposed by Romano et al. (2019). This method also requires splitting the data into training and calibration sets. After defining a coverage level $1 - \alpha$, we fit two quantile regression models on the training set, which can be any regression model. Next, we compute the conformity scores that quantify the magnitude of the error made by the plug-in prediction interval. Afterwards, we compute the quantile of the empirical distribution of the conformity scores. Finally, the output should be a conformalized prediction interval for the option price.

2.2. Machine Learning Model for Quantiles

The method of conformal quantile regression for prediction intervals requires training a regression model to predict quantiles. Ivașcu (2021) compares the performance of several machine learning models for option pricing, concluding that tree-based ensembles, such as gradient boosting machines and random forests, outperform neural networks and support vector machines in the pricing errors. In this study, we use a tree-based algorithm as the quantile regression model, more specifically a modified version of a gradient boosting machine.

A decision tree resorts to recursive binary splitting of the predictor space to generate a prediction. We start with a root node containing all the observations and then recursively split the data into subsets that are finer and finer as the trees progress. The process continues until the stopping criterion is met. The prediction is given by the mean value of the training observations in the leaf node. This algorithm is very efficient and

simple to understand, but it is also prone to overfit new data. To overcome that limitation, Ho (1995) introduced an ensemble method by building multiple trees in different subspaces of the feature space, thus increasing the generalization accuracy. Friedman (2001) introduced a gradient boosting machine, an algorithm of sequential trees in which each tree learns and enhances its performance based on the error residuals of the preceding tree. The extreme gradient boosting, or XGBoost, (Chen and Guestrin, 2016) is one of the most popular algorithms that implement gradient boosting. Its popularity comes from the adaptability of the model. Unlike the traditional gradient boosting machine, this model produces trees in parallel rather than sequentially. The Light Gradient Boosting Machine, or LightGBM (Ke et al., 2017) implements gradient boosting but focuses on being computationally efficient. This algorithm grows trees sequentially. Besides its speed, the algorithm also has other benefits: it performs well when trained with large datasets, even reaching a faster training time than the XGBoost, and it has a low memory consumption, as it converts continuous values to discrete bins (Mienye and Sun, 2022). On the other hand, the algorithm can lead to overfitting if it is trained in a small dataset, as the trees could be too complex.

3. METHODOLOGY

3.1. Conformal Quantile Regression

The following methodology is based on the work of Romano et al. (2019). Let Y represent an option price and X represent the corresponding features. Given a training set of n option prices $\{Y_i\}_{i=1}^{n}$ and corresponding features $\{X_i\}_{i=1}^{n}$, we train a machine learning model. Our aim is to compute a prediction interval $C(X_{n+1}) \subseteq \mathbb{R}$ for an unknown option price given its known features. For a certain desired nominal coverage $1-\alpha$, we have

$$P\{Y_i \in C(X_{n+1})\} \geq 1-\alpha\,, \tag{1}$$

where we expect the prediction interval to have a $1-\alpha$ probability of containing the option price. The samples from the dataset should be exchangeable and drawn from a joint distribution – for instance, if the observations are independent and identically distributed, the assumption holds.

We begin by splitting the initial training data into two disjoint subsets: an actual training set $\{(X_i, Y_i) : i \in \mathrm{I}_1\}$ and a calibration set $\{(X_i, Y_i) : i \in \mathrm{I}_2\}$ to obtain conformity scores. We should also define the miscoverage level $\alpha \in [0,1]$. The conditional distribution function of option prices Y given the set of features X is defined as

$$F(y|X = x) = P(Y \leq y|X = x), \tag{2}$$

and the conditional quantile function is

$$q_\alpha(X) = inf\{Y \in \mathbb{R} : F(Y|X = x) \geq \alpha\}. \tag{3}$$

The goal of the conditional quantile regression is to estimate quantiles, namely the median, of Y conditional on X. By fitting two conditional quantile models $q_{\frac{\alpha}{2}}(X)$ and $q_{1-\frac{\alpha}{2}}(X)$ on the training set, we obtain a conditional prediction interval for Y_{n+1}:

$$C(X_{n+1}) = \left[q_{\frac{\alpha}{2}}(X_{n+1}), q_{1-\frac{\alpha}{2}}(X_{n+1})\right]. \tag{4}$$

The methodology allows any regression model to be used for quantiles. We use a gradient boosting machine as the regression quantile model, specifically a Light Gradient Boosting Machine, as explained in detail in Section 3.2.

Next, we compute the conformity scores that quantify the error of the prediction interval. The conformity scores are computed for each observation within the calibration set and are given by:

$$E_i = \max\left[\hat{q}_{\frac{\alpha}{2}}(X_i) - Y_i, Y_i - \hat{q}_{1-\frac{\alpha}{2}}(X_i)\right], \forall\, i \in \mathrm{I}_2. \tag{5}$$

If our prediction of Y_i is below the lower end of the prediction interval $Y_i < \hat{q}_{\frac{\alpha}{2}}(X_i)$, then the magnitude of the error is given by $E_i = \hat{q}_{\frac{\alpha}{2}}(X_i) - Y_i$. Similarly, if the prediction of Y_i is above the upper end of the prediction interval $Y_i > \hat{q}_{1-\frac{\alpha}{2}}(X_i)$, and the magnitude of the error is given by $E_i = Y_i - \hat{q}_{1-\frac{\alpha}{2}}(X_i)$. At last, if the prediction of Y_i belongs within the prediction interval $\hat{q}_{\frac{\alpha}{2}}(X_i) < Y_i < \hat{q}_{\frac{\alpha}{2}}(X_i)$, then the magnitude of the error corresponds to the largest of the non-positive differences. We can illustrate the above with a set of examples.

We can take an interval of [0.006; 0.01]. Starting with the case of undercoverage, let us consider $Y_i = 0.004$. Since the prediction is below the lower end of the interval $\hat{q}_{\frac{\alpha}{2}}(X_i) = 0.006$, the conformity score is given by max [0.006-0.004; 0.004-0.01] = max [0.002; -0.006] = 0.002. For the case of overcoverage, we can consider $Y_i = 0.012$. The prediction is above the upper end of the interval $\hat{q}_{1-\frac{\alpha}{2}}(X_i) = 0.01$ and the conformity score is given by max [0.006-0.012; 0.012-0.01] = max [-0.006; 0.002] = 0.002. Finally, if the prediction is within the interval $Y_i = 0.008$, then the conformity score is always non-positive max [0.006-0.008; 0.008-0.01] = max [-0.002; -0.002] = -0.002.

We then compute a quantile of the empirical distribution of the conformity scores, which conformalizes the plug-in prediction interval:

$$Q_{1-\alpha}(\mathrm{I}_2) = \frac{(1-\alpha)(1+\mathrm{I}_2)}{\mathrm{I}_2}\text{-th empirical quantile of}\{E_i : i \in \mathrm{I}_2\}. \quad (6)$$

Finally, our output of this methodology is a conformalized prediction interval for Y_{n+1}:

$$C(X_{n+1}) = \left[\hat{q}_{\frac{\alpha}{2}}(X_{n+1}) - Q_{1-\alpha}(\mathrm{I}_2), \hat{q}_{1-\frac{\alpha}{2}}(X_{n+1}) + Q_{1-\alpha}(\mathrm{I}_2)\right]. \quad (7)$$

The methodology of conformal quantile regression ensures valid coverage in finite samples, without making distributional assumptions (Romano, 2019). This means that, for a given prediction interval, the empirical coverage level is very close to the nominal coverage level, i.e., if we want to obtain prediction intervals with 90% coverage, we should take $\alpha = 0.1$, fit the conditional quantile function at the 5% and 95% levels, obtain the corresponding intervals, and the empirical coverage level we obtain from it is very close to the 90% level we wanted. The methodology also ensures statistical efficiency of quantile regression and can take any algorithm for quantile regression. Finally, the prediction intervals are adaptive to heteroscedasticity, since they are calibrated using conditional quantile regression instead of conditional mean regression.

3.2. Light Gradient Boosting Machine

In this section, the LightGBM algorithm is introduced in detail. The algorithm implements gradient boosting and is designed to be highly computationally efficient, making it very fast. It can be used for classification, ranking and regression. LightGBM

produces trees sequentially, as seen in Figure 1, unlike algorithms such as the XGBoost, that produce trees in parallel, as per Figure 2. This ensures that the algorithm is effective when processing large-scale data and features. This algorithm contains two novel techniques: gradient-based one-side sampling (GOSS), which is used to deal with a large number of observations in the dataset, and the exclusive feature bundling (EFB), used to deal with a large number of features (Ke et al., 2017). GOSS is a subsample technique that is used in the training dataset for building the base trees in the ensemble. The technique increases the importance of the observations with higher uncertainty in their classifications, which are the ones with a higher gradient, as these should contribute more to the information gain (Bentéjac et al., 2021). The EFB technique combines sparse features into a single feature when those features do not have non-zero values simultaneously, i.e., for mutually exclusive features (Bentéjac et al., 2021).

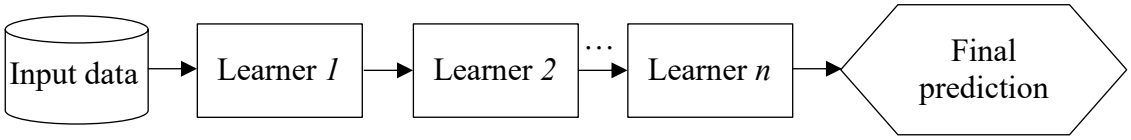

FIGURE 1 – Block diagram of sequential ensemble learning.

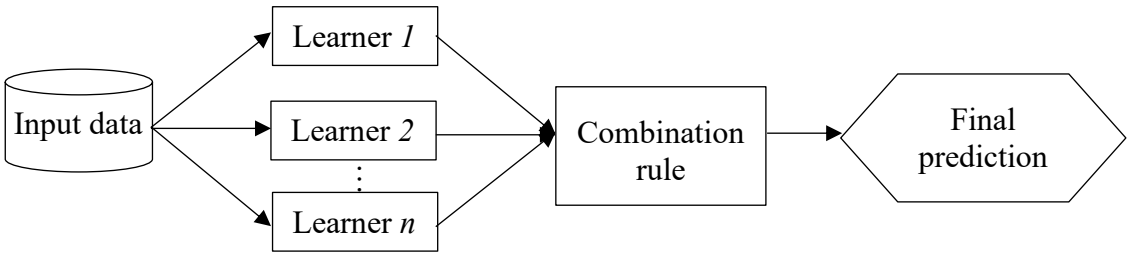

FIGURE 2 – Block diagram of parallel ensemble learning.

The prediction $\hat{Y}$ of the algorithm is obtained by summing the predictions of all the decision trees $\{f_k(X)\}_{k=1}^{K}$:

$$\hat{Y} = \sum_{k=1}^{K} f_k(X). \tag{8}$$

The initial tree $f_1(X)$ is a regular decision tree trained on the training dataset. The subsequent decision trees incrementally added to the committee, however, are trained on

the errors produced by the trees coming before. The objective of this process is to rectify the errors made by the previous decision trees. For each iteration, the tree to be added to the committee is the one that minimizes the regularized loss function:

$$\sum_{i=1}^{n} L(Y_i,\ \hat{Y}_i^{(k-1)} + f_k(X_i)) + \gamma T + \frac{1}{2}\lambda\|w_k\|^2, \tag{9}$$

where the last two terms are regularization terms. These terms serve to penalize complex trees, thus preventing the committee from overfitting the training data (Bastos, 2024).

As our objective is to predict quantiles, we use a pinball loss function as the loss function. The pinball loss function is a metric to assess the accuracy of quantile predictions (Romano, 2019):

$$L_\alpha(y,z) = \begin{cases} (y-z)\alpha, & if\ y \geq z \\ (z-y)(1-\alpha), & otherwise \end{cases} \tag{10}$$

where y represents the option price Y_i, z represents the quantile forecast $\hat{Y}_i^{(k-1)} + f_k(X_i)$ and α represents the target quantile. From Eq. (10) we can see that the pinball loss function is always positive. The more accurate quantile predictions have the lower pinball loss functions.

Since the accuracy of the predictions is highly influenced by the hyperparameters, we need to determine the optimal number and range of variation of the hyperparameters of the model. We use the hyperparameters' search space proposed by Bastos (2024) and perform a grid-search analysis to find the optimal values. The hyperparameters included in the grid-search, presented in Table I, are the following:

- Number of trees in the ensemble: this refers to the total number of decision trees in the ensemble. Although the first trees added usually offer significant improvements in out-of-sample accuracy, the marginal benefit typically diminishes as more trees are added, reflecting the phenomenon of diminishing returns.
- Maximum number of leaves per tree: this parameter limits the number of terminal nodes in a tree to help control model complexity. Overly complex models may lead to overfitting, which happens when a model follows the noise too closely instead of the underlying pattern between the target Y and the inputs X. An

overfitted model will fit well the training data but will undermine the out-of-sample accuracy.

- Maximum tree depth: this parameter imposes a limit on the depth of each decision tree, also to control model complexity and help mitigate the risk of overfitting.
- Learning rate: the learning rate controls the step size in the gradient descent algorithm. If this parameter is too small, the convergence may be slow. On the other hand, if it is too large, the gradient descent might fail to converge.

TABLE I

HYPERPARAMETERS SEARCH SPACE FOR THE LIGHT GRADIENT BOOSTING MACHINE

Hyperparameter	Search space
Number of trees in the ensemble	{100, 500, 1000, 2500, 5000}
Maximum number of leaves per tree	{32, 64, 128, 256}
Maximum tree depth	{8, 16, 32, 64}
Learning rate	{0.005, 0.01, 0.05, 0.1, 0.5}

Once we perform the grid-search, we determine the optimal values for the hyperparameters, thus obtaining models with the lowest mean absolute error on the validation data. After the models are trained, we can calculate prediction intervals for the test data. In all computations, we use a random set with 20% of the observations as validation data.

4. DATA

4.1. Description, Treatment and Statistics

We use data extracted from the Ivy DB US database by OptionMetrics. The dataset comprises European call and put options on the S&P 500 index, traded from January 2018 to December 2022. For each option, we have the date of the observed option price, a call-put flag to distinguish calls and puts, the strike price, volume of option contracts, implied volatility, days to expiration, option price, spot price, dividend yield and risk-free rate.

In order to guarantee the quality of the data, we apply liquidity filters. Observations with volume lower than 20 contracts are excluded from the analysis (Gaspar et al., 2020), as illiquid options might increase noise in the model due to their potential stale prices and wide bid-ask spreads. Observations with price below 0.05 USD are excluded, to eliminate very cheap options which are usually deep-out-of-the-money and are unlikely to be exercised. We also exclude observations with maturity of fewer than 10 trading days and higher than 1 year (Bastos, 2024) to ensure more stability in the model. Options very close to expiration date can have big variations in price due to time decay and gamma effects, while options with a long time until maturity may be less liquid. We also cap volatility at 0.8. Additionally, we exclude deep-in-the-money and deep-out-of-the-money contracts from the analysis ($0.5 < S/K < 1.5$) (Bastos, 2024), as these options behave differently than near-the-money options, since they are less sensitive to market data.

After all the filters are applied, the total sample size is 1,831,515 option contracts, with 704.915 call contracts and 1,126,600 put contracts. Although the sample size significantly decreased after the filters were applied, it is still much larger than what was found in the literature – for instance, Bastos (2024) used 120,180 empirical observations.

Figure 3 illustrates the option prices against moneyness. Call options have a price near zero when they are deep out-of-the-money, i.e., with moneyness significantly below 1, as it is unlikely that they will be exercised. As moneyness increases, the prices increase. For put options, we see they have a price near zero when they are deep out-of-the-money, i.e., with moneyness significantly above 1, and their price increases when they are in-the-money. We can also observe that put option prices can go much higher than call option prices, for a few observations. This could be attributed to the fact that investors are more willing to seek downside protection, and tendency to perceive a large downward movement in an asset as more probable than a large upward movement (Bouzoubaa and Osseiran, 2010). Given that put options can be used as a hedging instrument against a downward movement in an asset price, they confer the downside protection that the investors seek, thus putting upward pressure in the price of put options.

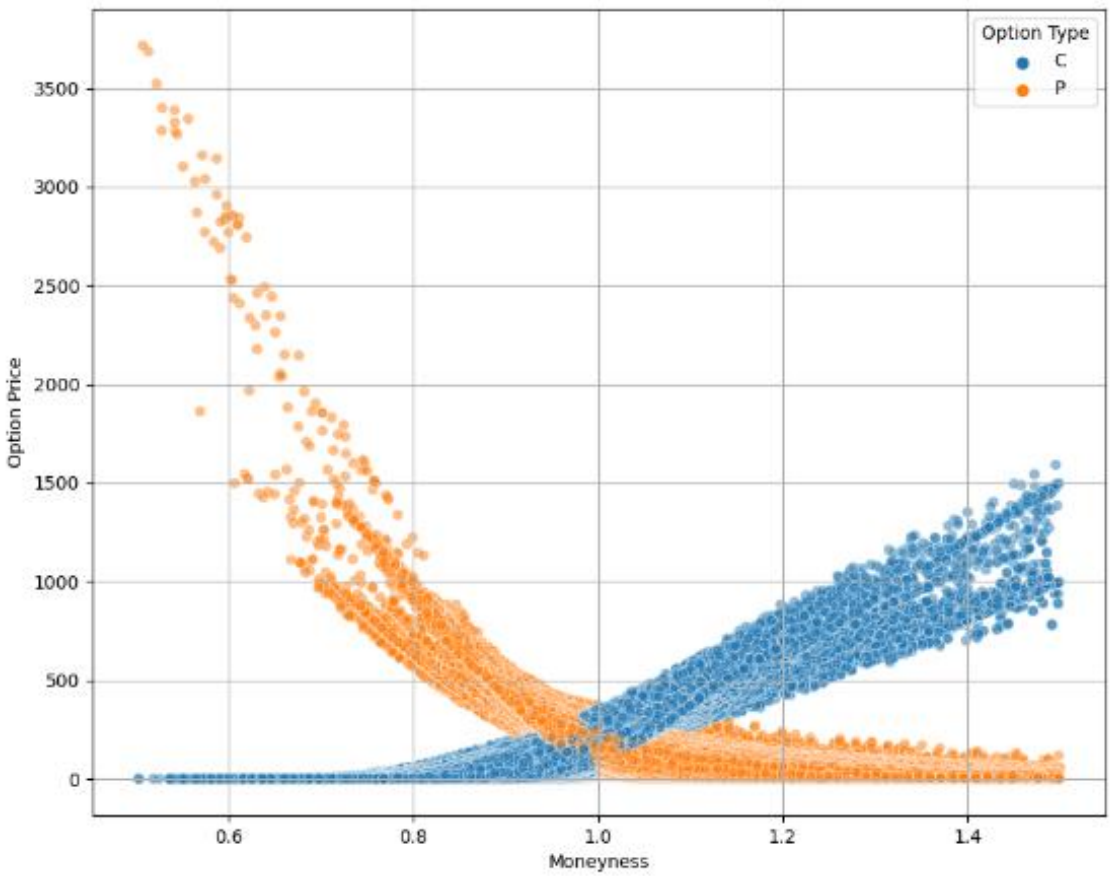

FIGURE 3 – Option prices against moneyness

Figure 4 and Figure 5 represent, respectively, a call and put option price surface, plotted against moneyness and maturity. We can observe the same pattern as described in the previous figure regarding moneyness. We do not observe a significant change in option prices against maturity as we are dealing with a relatively short period, ranging from 10 trading days to one year, which excludes very short-term options and very long-term options from the analysis.

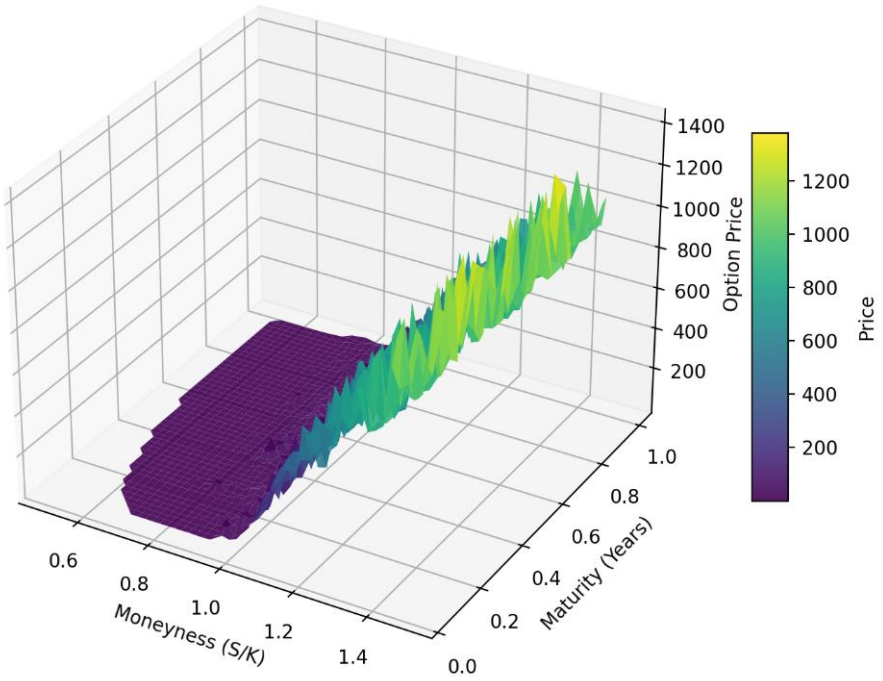

FIGURE 4 – Call option prices against moneyness and maturity.

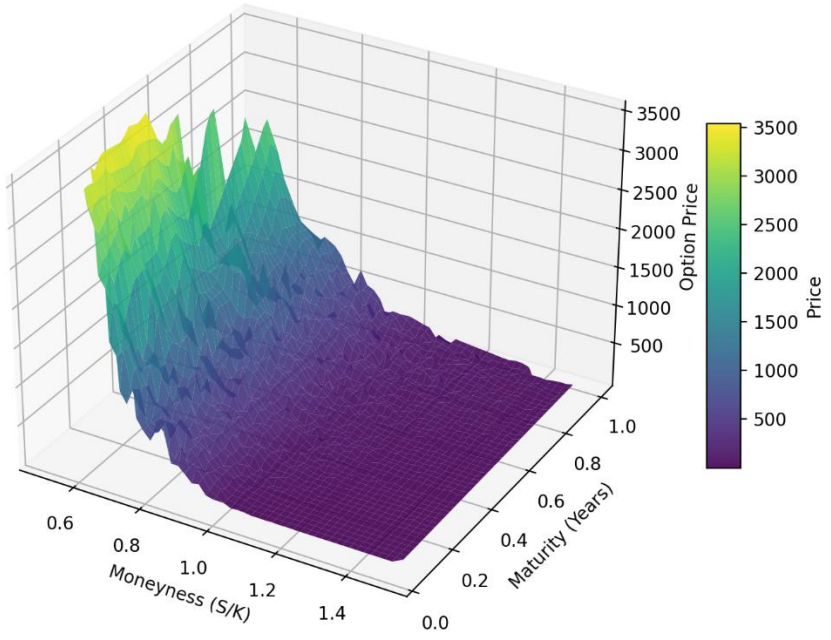

FIGURE 5 – Put option prices against moneyness and maturity.

Table II presents the descriptive statistics of the input variables and response variable, split between call options and put options. The median option price scaled by the strike is 0.008 for both call and put options. The median moneyness for call options is 0.973, slightly lower than the median moneyness for put options of 1.067, although both are near ATM. At least 75% of the call option prices are below one, and at least 75% of the put option prices are above one, indicating that these options are typically out-of-the-money. The median implied volatility for put options is 0.246, higher than the median for call options of 0.163. This is consistent with the expected from volatility skew theory – if an asset or index price drops, volatility tends to increase. Since puts pay on the downside and investors tend to look for downside protection, the price of OTM put options tends to be higher than price of OTM call options, ceteris paribus. This also reflects on the implied volatility of OTM puts tending to be higher than the OTM calls (Bouzoubaa and Osseiran, 2010). The median maturity is a little over a month for both call and puts. An interesting result to observe is the high skewness of the option prices in percentage of the strike, especially for call options, which implies heavy right tails. This highlights that the distribution is asymmetric, which is not a problem in quantile regression. Since the methodology models a conditional quantile instead of the conditional mean of the target variable, we do not have to assume normality of the distribution.

TABLE II

STATISTICS ON INPUT VARIABLES AND RESPONSE VARIABLE

Call options						
	Option price scaled by the Strike	Moneyness	Implied volatility	Maturity	Risk-free rate	Dividend yield
Mean	0.017	0.965	0.174	0.207	0.014	0.014
Std dev	0.031	0.060	0.073	0.192	0.011	0.004
Min	0.000	0.503	0.049	0.044	0.001	0.000
25%	0.002	0.941	0.120	0.079	0.002	0.012
50%	0.008	0.973	0.163	0.127	0.015	0.015
75%	0.021	0.996	0.211	0.262	0.023	0.017
Max	0.521	1.500	0.799	0.996	0.049	0.021
Median	0.008	0.973	0.163	0.127	0.015	0.015
Skewness	7.034	0.217	1.853	1.880	0.436	-1.352
Kurtosis	74.958	10.607	7.053	3.288	-0.850	2.069
Put options						
	Option price scaled by the Strike	Moneyness	Implied volatility	Maturity	Risk-free rate	Dividend yield
Mean	0.014	1.105	0.263	0.211	0.014	0.015
Std dev	0.019	0.120	0.109	0.189	0.011	0.004
Min	0.000	0.507	0.038	0.044	0.001	0.000
25%	0.003	1.019	0.185	0.083	0.002	0.012
50%	0.008	1.067	0.246	0.139	0.016	0.016
75%	0.019	1.161	0.316	0.274	0.023	0.017
Max	0.489	1.500	0.800	0.996	0.049	0.021
Median	0.008	1.067	0.246	0.139	0.016	0.016
Skewness	4.179	1.165	1.130	1.800	0.349	-1.331
Kurtosis	34.378	0.948	1.866	3.015	-0.871	2.420

Figure 6 displays the histograms for each input variable, and Figure 7 displays the associated box plots. Implied volatility is right-skewed and has several outliers in the upper end of the boxplot – meaning the values that fall over the 75th percentile. Moneyness is centred around ATM options. The maturity is heavily right skewed, centred around a month-long maturity.

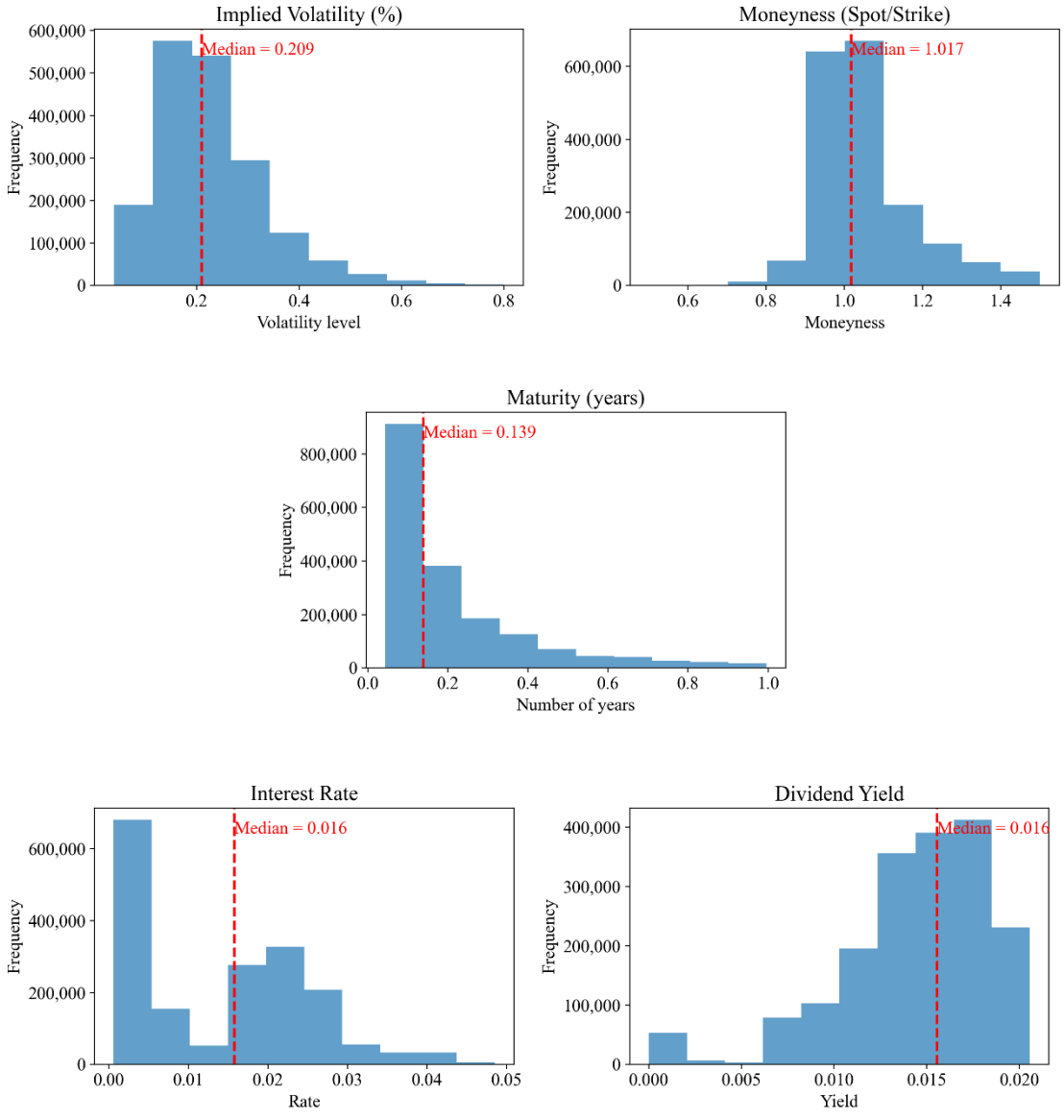

FIGURE 6 – Input variables' histograms.

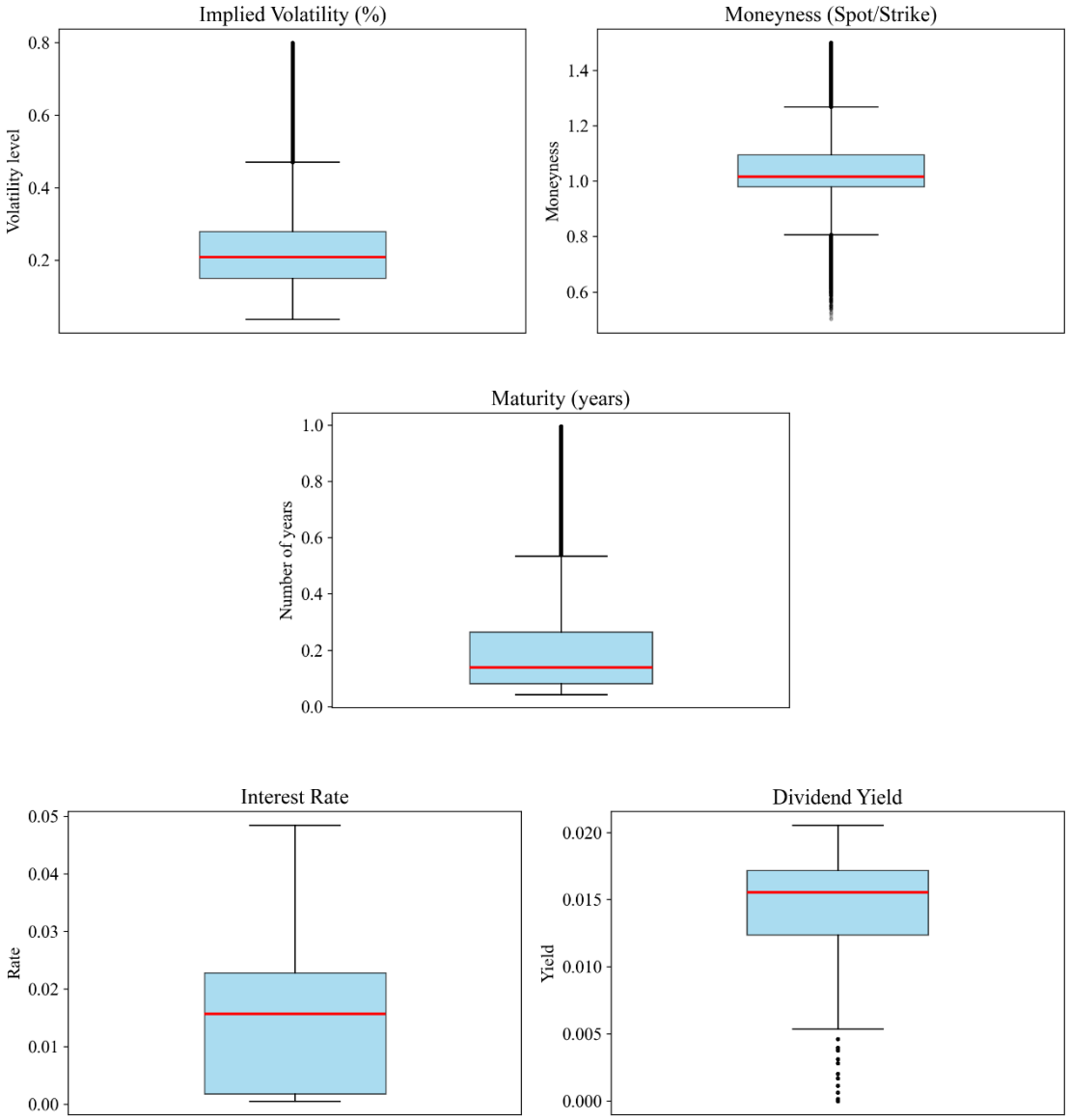

FIGURE 7 – Input variables' boxplots.

4.2. Training, calibration and test datasets

All the models perform a random split of 80/20 between the training and test sets, and the CQR employed one more split of 80/20 within the training set to split the data in training set and calibration set, respectively. The training dataset is a set of observations that are used to train the models with pinball loss functions. The calibration dataset is

used to predict quantiles and compute conformity scores. Finally, the test dataset allows us to make an unbiased evaluation of the conformal prediction intervals.

To assess whether the training and test datasets have a similar distribution, we employ the Kolmorogov-Smirnov test feature-by-feature. The findings presented in Table III lead us to conclude that the distributions are likely the same, as evidenced by the p-value higher than 0.05 for every variable and by the low K-S statistic, which measures the largest difference between the empirical distribution function of both datasets.

TABLE III

KOLMOGOROV-SMIRNOV BY FEATURE

	Implied volatility	Moneyness	Maturity	Risk-free rate	Dividend yield
p-value	0.9530	0.8092	0.3600	0.5159	0.8833
statistic	0.0010	0.0012	0.0017	0.0015	0.0011

For small Kolmogorov-Smirnov statistics (*p*-value higher than 0.05), we do not reject the null hypothesis of the datasets being drawn from the same distribution

To mitigate any issue of sampling variability, the calculations are performed 50 times using different random splits of the data into training, calibration and test datasets. The results presented come from an average of the 50 trials, thus increasing reliability.

The models use an array of features to compute the target variable. For each option, the target variable is the option price in percentage of the strike (price/K). The scale is normalized when we divide the option price by the strike price, thus ensuring a more stable target variable. After the prediction, we rescale by multiplying the result by the strike to have the option price. The model features are the moneyness (S/K), maturity, implied volatility, interest rate and dividend yield.

An option is said to be at-the-money (ATM) when the strike price is equal to the current spot price of the underlying asset (Bouzoubaa and Osseiran, 2010). Here, we considered as ATM the options with less than two percent deviation from the spot price, so $0.98 < S/K < 1.02$. A call option is in-the-money (ITM) when the strike price is below the current price of the underlying asset, and out-of-the-money (OTM) when the strike price is above the spot price. Therefore, above the 1.02 moneyness threshold we have the

ITM calls and below the 0.98 threshold are the OTM calls. For put options, the opposite happens – a put option is ITM when the strike price is above the spot price and OTM when the strike price is below the current trading price. So, above the 1.02 moneyness threshold we have the OTM puts and below the 0.98 threshold are the ITM puts. Naturally, at the maturity date, only ITM options will be exercised, as the remaining states of moneyness at maturity imply that the options are worthless.

For maturity, we considered short-term (ST) to be below 30 trading days, long-term (LT) when the maturity is above 60 days and medium-term (MT) for the ones falling in between. Table IV displays the distribution of moneyness and maturity for call and put options in the dataset. We should note the bias in the distribution – ATM, OTM and short-term options are over represented in the dataset, which is expected since these are the most traded options in the markets, but might still influence the precision of the predictions for ITM and long-term options (Gaspar et al., 2020).

TABLE IV

SAMPLE MONEYNESS AND MATURITY.

	Moneyness			Maturity		
	ITM	ATM	OTM	ST	MT	LT
Call options	7.49%	35.19%	57.32%	46.92%	27.22%	25.86%
Put options	4.51%	21.2%	74.29%	44.37%	28.49%	27.14%

5. RESULTS

In this section, we present the results of the conformal quantile regression models for predicting option prices and compare them with a non-conformal quantile regression model. We use empirical coverage and relative width of prediction intervals as performance metrics. In Section 5.1 we perform an empirical coverage analysis and in Section 5.2 we analyse the conditional coverage of the model.

Figure 8 displays the conformal prediction intervals plotted against moneyness. Figure 9 shows the conformal prediction intervals plotted against time-to-maturity. In the upper plots are represented the call options and in the lower plots are represented the put options. The nominal coverage of the intervals is set to 0.9, meaning that, if the conformal

intervals are valid, approximately 90% of the actual option prices should fall within the predicted intervals. The bars in grey represent the instances where the prediction interval failed to cover the actual value. The dots represent actual option prices missed by the intervals. The variations in the width of the intervals across moneyness reflect the model's ability to adapt to uncertainty. Wider intervals are observed in regions where the model has less confidence in the predictions, while narrower intervals correspond to more confident predictions.

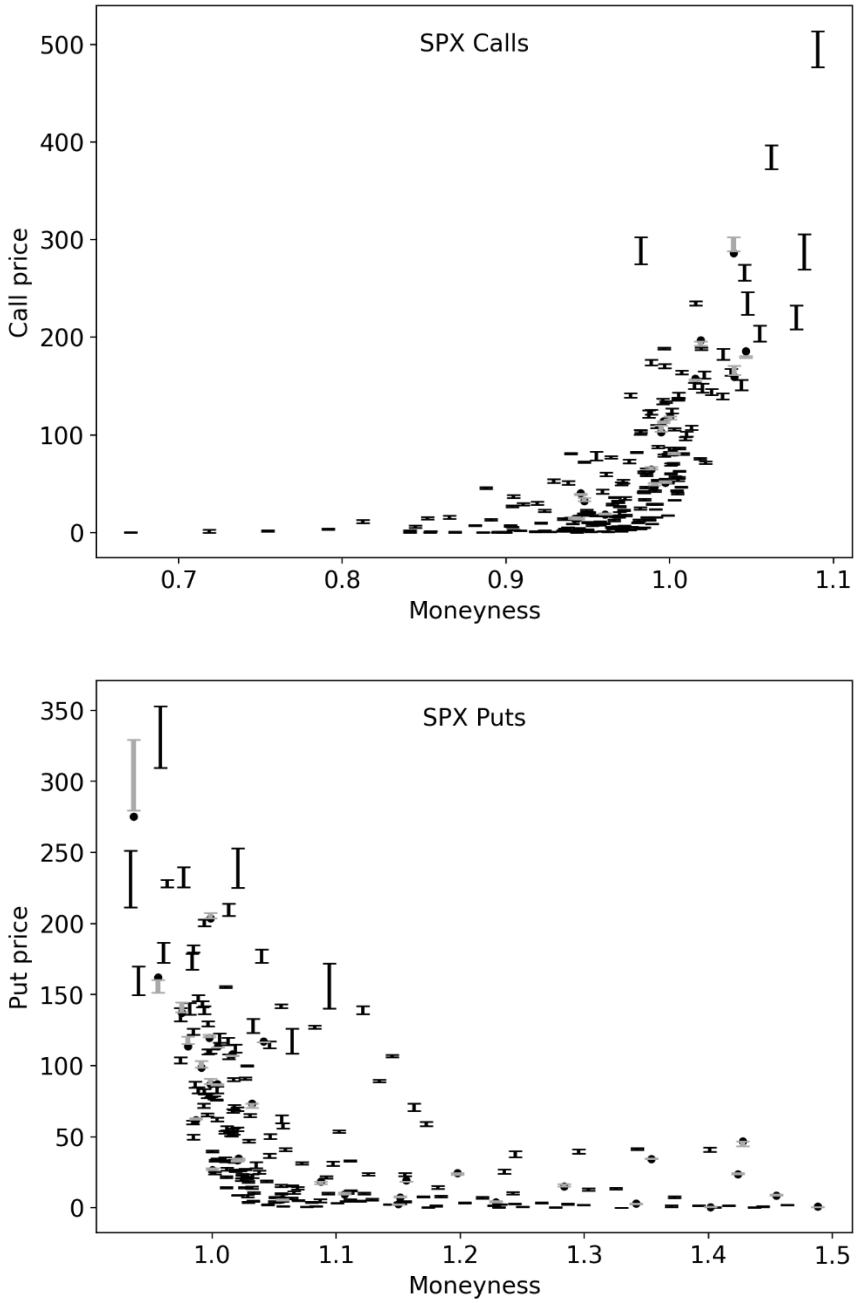

FIGURE 8 – Conformal prediction intervals against option moneyness (S/K).

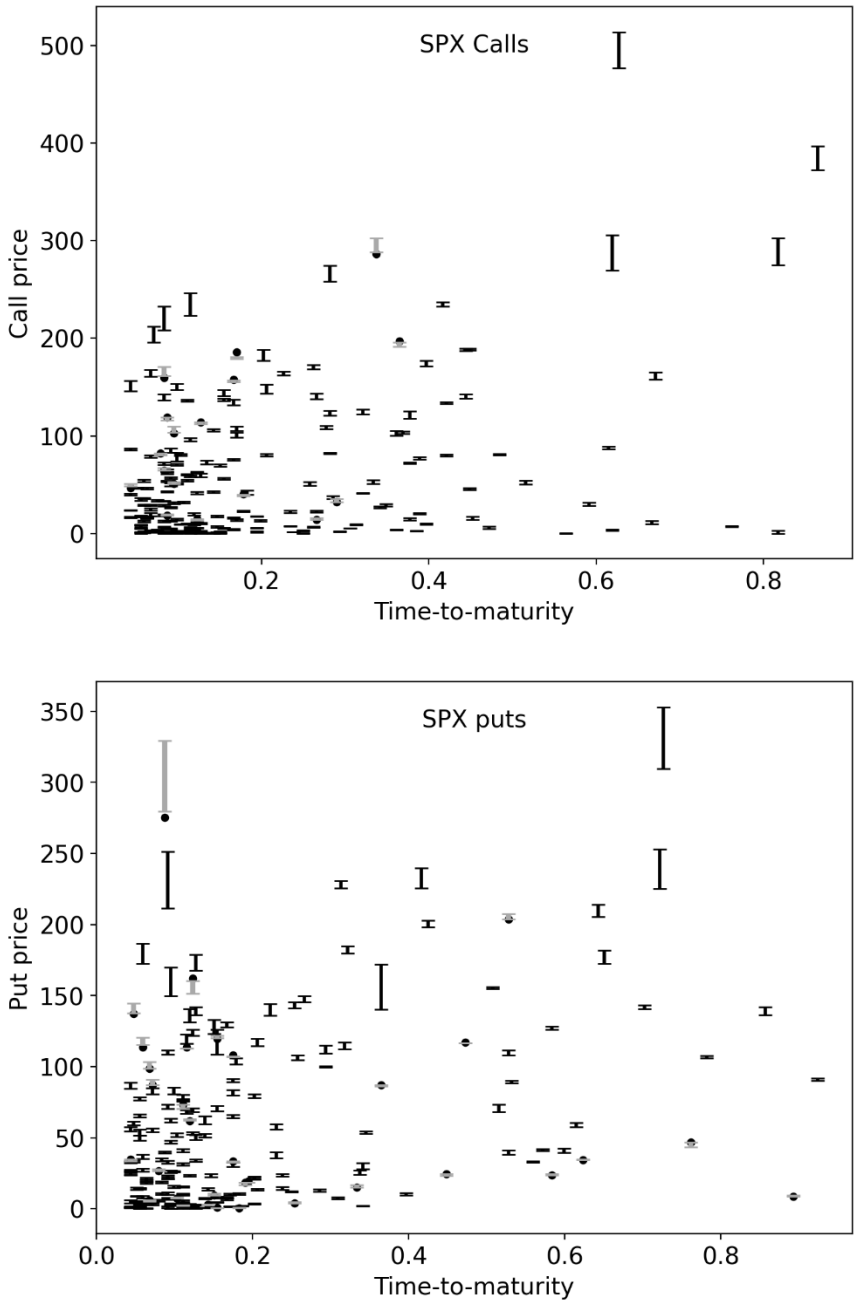

FIGURE 9 – Conformal prediction intervals against time-to-maturity.

5.1. Empirical coverage

To assess the validity of the coverage guarantees given by the CQR model, we resort to two important metrics – empirical coverage and relative width. Empirical coverage is calculated as the proportion of observed option prices that lie within the prediction intervals. The median relative width of the intervals refers to the median of

their size. The wider the intervals are, the more easily we can expect the actual observation value to fall within the interval. However, for an interval to be informative, it should be as narrow as possible. Conversely, for narrower intervals, it is more challenging for the intervals to encompass the true observed value. This reveals a trade-off relationship between the two metrics – if we wish to increase the empirical coverage of a prediction interval, we will increase the size of the prediction interval, all else equal.

Table V presents the empirical coverage and median relative width of prediction intervals, using both conformal and non-conformal methodologies. CQR refers to the intervals obtained from the conformal machine learning models for quantiles, while NQR refers to the intervals using plain machine learning models for quantiles. The CQR intervals reach an empirical coverage level very close to the nominal coverage level of 0.9, thus guaranteeing the finite-sample coverage. In contrast, the NQR presents empirical coverage levels below the nominal target, meaning that the model undercovers the actual option prices. The values reported throughout the analysis result from an average of 50 random split of the data into training, calibration and test sets.

The median relative width of the NQR intervals is lower than the CQR intervals. This is expected, since non-conformal models do not provide a finite-sample coverage guarantee and therefore tend to underestimate the uncertainty. The wider CQR intervals reflect the ability to capture the uncertainty inherent to estimating option prices. It is also worth noting that the CQR model is trained on less data than the NQR model, given an additional split is performed, to hold 20% of the observations from the training dataset to serve as the calibration dataset.

TABLE V

EMPIRICAL COVERAGE AND MEDIAN RELATIVE WIDHT OF THE PREDICTION INTERVALS FROM CQR AND NQR

	Empirical coverage		Relative width	
	NQR	CQR	NQR	CQR
Calls	0.8096	0.9000	5.38%	6.94%
Puts	0.8134	0.9000	7.28%	8.35%

5.2. Conditional coverage

The empirical coverage analysis satisfies the marginal coverage guarantee expressed in Eq. (1). However, this guarantee does not automatically hold for a conditional coverage analysis. To obtain valid conditional coverage across different regions of the regressor space, i.e., to guarantee that the coverage of the intervals is still the same as the nominal coverage of 90% even in subsets of the data, it would be necessary to train individual models on each subset. Table IV suggests that for certain subsets of the data, such as ITM and long-term options, the data could be limited to effectively train separate models for each region. Nevertheless, it remains an interesting analysis to explore the behaviour of the conditional coverages in subsets of the data when the model is trained on the full dataset.

Table VI displays the conditional coverages and median relative widths for different levels of time to maturity. We can observe that the empirical coverage of the prediction intervals increases for shorter maturities, closely aligning with the nominal level for medium-term maturities. However, long-term options tend to undercover, while short-term options tend to overcover. The relative width of the prediction intervals is lower for longer maturities, suggesting that the model is less confident in predicting option prices for shorter term options. The results are consistent with the expected, given that option prices can vary significantly near the strike price as they get closer to maturity. Finally, it is worth comparing the results for the intervals of call option prices and put option prices. While they follow the same pattern, it is visible that the median relative width of short-term options is significantly higher for puts when compared to calls. This could be due to the higher implied volatility in puts than in calls, as we observed in Table III, which is linked to the different implied volatility skew. Put options generate positive payoffs when asset prices decrease, namely during market crashes, which are frequently accompanied by increases in volatility. The market tends to perceive negative trends in asset prices as more probable than positive ones, meaning that after a decrease in prices, the investors tend to anticipate further decreases. (Bouzoubaa and Osseiran, 2010). This perception results in the downside of an asset having a higher implied volatility associated with it than the upside. The higher implied volatility reflects the higher uncertainty, which results in wider prediction intervals for put options.

TABLE VI

CONDITIONAL COVERAGE AND MEDIAN RELATIVE WIDHT OF CQR FOR DIFFERENT LEVELS OF TIME TO MATURITY

	Empirical coverage			Relative width		
	ST	MT	LT	ST	MT	LT
Calls	0.9262	0.9112	0.8441	7.91%	7.20%	5.57%
Puts	0.9270	0.9082	0.8502	11.31%	8.53%	5.49%

ST: $T < 30$, MT: $30 \leq T < 60$, and LT: $T \geq 60$, with T = maturity in days

Table VII shows the conditional coverages and median relative widths for different levels of moneyness. The empirical coverage of the prediction intervals decreases with moneyness, as we observe higher empirical coverages for OTM options. ITM options tend to undercover actual prices, even more than observed for options with higher time to maturity. However, it is worth noting that they are closer to the nominal target than the results provided by the NQR models. The relative width of the prediction intervals tends to decrease with moneyness, which suggests that the model is less confident in predicting option prices for OTM options. This finding is in line with the expected, since OTM options have low intrinsic value, and are prone to lose value quickly as they approach maturity. ATM options present the narrowest intervals. This could be due to the high concentration of options near-the-money in our dataset, which is an accurate reflection of the same concentration seen in the market. Given the high number of observations in this small range of moneyness, the model can generate more confident predictions.

TABLE VII

CONDITIONAL COVERAGE AND MEDIAN RELATIVE WIDHT OF CQR FOR DIFFERENT LEVELS OF MONEYNESS

	Empirical coverage			Relative width		
	OTM	ATM	ITM	OTM	ATM	ITM
Calls	0.9171	0.8911	0.8108	12.93%	3.79%	5.86%
Puts	0.9086	0.8850	0.8287	10.85%	4.22%	8.81%

OTM call and ITM put: $M < 0.98$, ATM: $0.98 \leq M < 1.02$, ITM call and OTM put: $M \geq 1.02$, with M = moneyness

6. CONCLUSIONS

This study addresses the problem of quantifying uncertainty in option pricing using conformal quantile prediction. We conduct an empirical experiment using call and put options on the S&P500, traded between January 2018 to December 2022, with a total of 1.8 million observations. We use conformal quantile regression to generate statistically rigorous prediction intervals for machine learning models that predict option prices. The predictive model employed is a modified version of a gradient boosting machine.

The results confirm the empirical coverage guarantee provided by the conformalized model, contrasting with the non-conformal model that undercovers significantly the actual option prices. Furthermore, we uncover several findings regarding the uncertainty in option pricing. We obtained wider prediction intervals for options closer to maturity, suggesting a less confident prediction by the model. Indeed, option prices can vary significantly near the strike price as they get closer to expiration. We also observed wider intervals for out-of-the-money options, as these possess low intrinsic value. Finally, we uncover a novel difference between the intervals for call and put options with short time-to-maturity, with the latest presenting significantly wider intervals than the former, thereby offering an interesting insight into the asymmetries of market expectations and model uncertainty.

This study considered a much larger number of observations than those found in similar literature, with over 1.8 million observations. Additionally, we analysed both call and put options. Future research could build upon these findings, as it could be meaningful to analyse the results with separate models for different regions of the dataset, to allow a conditional coverage guarantee. This framework could also be used with a different predictive machine learning model, or for price predictions of a different type of financial asset.

REFERENCES

Bastos, J.A. (2024). Conformal prediction of option prices. *Expert Systems with Applications, 245*, 123087. https://doi.org/10.1016/j.eswa.2023.123087

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, *81*(3), 637–657. Scopus. https://doi.org/10.1086/260062

Bouzoubaa, M., & Osseiran, A. (2010). *Exotic Options and Hybrids—A Guide to Structuring, Pricing and Trading*. John Wiley & Sons, Ltd.

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. *13-17-August-2016*, 785–794. Scopus. https://doi.org/10.1145/2939672.2939785

Ech-Chafiq, Z.E.F., Labordere, P.H., & Lelong, J. (2023). Pricing Bermudan Options Using Regression Trees/Random Forests. *SIAM Journal on Financial Mathematics*, *14*(4), 1113–1139. Scopus. https://doi.org/10.1137/21M1460648

Fang, Z., & George, K. M. (2017). *Application of Machine Learning: An Analysis of Asian Options Pricing Using Neural Network*. 142–149. Scopus. https://doi.org/10.1109/ICEBE.2017.30

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gan, L., & Liu, W.-H. (2024). Option Pricing Based on the Residual Neural Network. *Computational Economics*, *63*(4), 1327–1347. Scopus. https://doi.org/10.1007/s10614-023-10413-3

Gaspar, R.M., Lopes, S.D., & Sequeira, B. (2020). Neural Network Pricing of American Put Options. *Risks*, *8*(3), Artigo 3. https://doi.org/10.3390/risks8030073

Ho, T.K., 1995. Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994

Hutchinson, J.M., Lo, A.W., & Poggio, T. (1994). A Nonparametric Approach to Pricing and Hedging Derivative Securities Via Learning Networks. *The Journal of Finance*, *49*(3), 851–889. https://doi.org/10.1111/j.1540-6261.1994.tb00081.x

Ivașcu, C.-F., (2021). Option pricing using Machine Learning. *Expert Systems with Applications, 163*, 113799. https://doi.org/10.1016/j.eswa.2020.113799

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, *30*. https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Malliaris, M., & Salchenberger, L. (1993). A Neural Network Model for Estimating Option Prices. *Applied Intelligence*. https://epublications.marquette.edu/mgmt_fac/287.

Mienye, I.D., & Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*, *10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Papadopoulos, H., & Vovk, V. (2008). *Normalized nonconformity measures for regression Conformal Prediction*. 64–69.

Papadopoulos, H., Proedrou, K., Vovk, V., & Gammerman, A. (2002). Inductive Confidence Machines for Regression. Em T. Elomaa, H. Mannila, & H. Toivonen (Eds.), *Machine Learning: ECML 2002* (pp. 345–356). Springer. https://doi.org/10.1007/3-540-36755-1_29

Papadopoulos, H., Vovk, V., & Gammerman, A. (2011). Regression Conformal Prediction with Nearest Neighbours. *Journal of Artificial Intelligence Research*, *40*, 815–840. https://doi.org/10.1613/jair.3198

Romano, Y., Patterson, E., & Candes, E. (2019). Conformalized Quantile Regression. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html.

Shubham, K., Tiwari, V., & Patel, K. S. (2023). Predictive Learning Methods to Price European Options Using Ensemble Model and Multi-asset Data. *International Journal on Artificial Intelligence Tools*, *32*(07), 2350034. https://doi.org/10.1142/S0218213023500343

Umeorah, N., Mashele, P., Agbaeze, O., & Mba, J. C. (2023). Barrier Options and Greeks: Modeling with Neural Networks. *Axioms*, *12*(4). Scopus. https://doi.org/10.3390/axioms12040384

DISCLAIMER

I declare that this dissertation was completed with strict adherence to the academic integrity policies and guidelines set forth by ISEG, Universidade de Lisboa. The work presented herein is the result of my own research, analysis and writing, unless otherwise cited. In the interest of transparency, I provide the following disclosure regarding the use of artificial intelligence (AI) tools in the creation of this dissertation:

I disclose that AI tools were employed during the development of this thesis as follows:

- Generative AI tools were consulted to help debug and improve the code behind the models used. However, all writing, synthesis and critical analysis are my own work.

Nonetheless, I have ensured that the use of AI tools did not compromise the originality and integrity of my work. All sources of information have been appropriately cited in accordance with academic standards. The ethical use of AI in research and writing has been a guiding principle throughout the preparation of this dissertation.

I understand the importance of maintaining academic integrity and take full responsibility for the content and originality of this work.

Beatriz Leite, 30-06-2025.