National College of Ireland

Predicting the Bid-Ask Spread of Equity Options: A Machine Learning Approach Applied to Amazon and AMD

MSc Research Project
FinTech

Pablo José González Pardo
Student ID: x24120511

School of Computing
National College of Ireland

Supervisor: Brian Byrne

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Pablo José González Pardo		
Student ID:	X24120511		
Programme:	MSc FinTech	**Year:**	2024-2025
Module:	MSc Research Project		
Supervisor:	Brian Byrne		
Submission Due Date:	11th August 2025		
Project Title:	Predicting the Bid-Ask Spread of Equity Options: A Machine Learning Approach Applied to AMZN and AMD		
Word Count:	5910	**Page Count:**	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Pablo José González Pardo

Date: 10th of August 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting the Bid-Ask Spread of Equity Options: A Machine Learning Approach Applied to AMZN and AMD

Pablo José González Pardo
x24120511

Abstract

This research investigates the application of machine learning techniques to predict the bid-ask spread of equity options, a key measure of market liquidity and trading costs. While most existing literature focuses on pricing and volatility estimation, this study addresses the topic of direct spread modelling using real world options data from Amazon (AMZN) and AMD, the project develops a full pipeline including feature engineering, exploratory analysis, and the training of both linear and nonlinear models.

Results show that machine learning models significantly outperform traditional linear regressions, especially in capturing nonlinear interactions between variables. Among all models, CatBoost consistently achieves the highest predictive accuracy. Furthermore, explainability tools like SHAP and LIME are used to assess feature importance and enhance model transparency. The findings contribute to both academic understanding and practical applications, providing useful insights for traders, risk analysts, and regulators interested in the behaviour of liquidity in derivatives markets.

1 Introduction

1.1 Background

Options could be defined as financial contracts that give the holder the right to buy or sell a financial instrument at a specific price for a certain period of time, each option has the following key features:

- **Strike price**: The price at which an option can be exercised.
- **Expiration date**: The date at which an option expires.
- **Option premium**: The price at which an option is purchased.
- **Stock price:** The price of the underlying asset.

There are two types of options, **call options**, that give the holder the right, but not the obligation, to buy the underlying asset at the strike price on or before expiration and **put options** that give the holder the right, but not the obligation, to instead sell the underlying asset at the strike price on or before expiration.

Options markets have long suffered from reputational issues due to misuse during past crises however they are now recognized for their essential economic role serving as critical instruments in global financial systems that offer investors the ability to hedge, speculate, and manage risk through derivative contracts. Nowadays, there are several organized options markets in Europe, America, Oceania and Asia, within the United States the largest market is Chicago Board Options Exchange (CBOE) established in 1973, this market started a pivotal moment in options market development with innovations such as standardized strike prices and

expiration dates improved transparency, reduced costs, and increased liquidity. (Vallelado González, 1992)

The core trading session in U.S. option markets runs from 8:30 AM to 3:00 PM, with extended sessions for certain products. The Options Clearing Corporation (OCC) sets daily position limits (number of contracts) and minimum margin requirements (up to 120% for uncovered writers). Regarding the market regulation, U.S. options exchanges operate under SEC oversight and collaborate through various National Market System (NMS) plans, designed to promote cross market transparency, surveillance, and standardization. One key plan is the Consolidated Options Reporting Plan, implemented by OPRA, which processes real-time trade and quote data from all 16 U.S. option exchanges, as required by Regulation NMS Rule 602. The Securities Industry Automation Corporation (SIAC) provides technological infrastructure for data processing. OPRA distributes data to professional and retail subscribers for a fee.
(Andersen *et al.*, 2021)

1.2 Motivation and Research Gap

This research is focused on the bid ask spread defined as the difference between the ask price and the bid price for an asset (options in this project) in the market. Bid price refers to the highest price a buyer is willing to pay for a security at a given time and ask price refers to the lowest price a seller is willing to accept for that same security.
Bid-ask spreads represent a fundamental component of trading costs and liquidity in financial markets. In the context of options, spreads tend to be wider and more volatile than in other asset classes due to factors such as low trading volume, high implied volatility, and the granularity of available contracts. Understanding and forecasting spreads is crucial for investors, market makers, and regulators, as it directly affects execution costs, portfolio management, and perceived market efficiency.

Despite its importance, most academic research has focused on pricing models or volatility estimation, while spread behavior has received relatively limited attention. In addition, existing approaches are mainly focused on parametric models or linear regressions, which are unable to capture complex non-linear interactions between key variables like moneyness, time to expiration, and implied volatility. Moreover, traditional pricing frameworks such as Black-Scholes or Heston treat spreads as exogenous frictions, rather than modeling them directly.

The aim of this research is to address this gap by applying machine learning and neural network models to predict bid-ask spreads of American options and comparing their performance with linear models to demonstrate how the nonlinear relationships affect predictions.

1.3 Research Question

Following the motivation outlined for this study, the key research question can be formulated as:

> To what extent can machine learning models outperform traditional linear models in predicting bid-ask spreads of American options, and which input variables contribute most to model accuracy?

1.4 Objectives

- To identify and evaluate the main features that influence the bid-ask spread in American options.

- Analise nonlinear relationships between features such as time to expiration, moneyness and spread for AMZN calls, puts and AMD calls.
- To apply and compare linear models and non-linear machine learning models for predicting option bid-ask spreads aiming to choose the best performing model.
- To explore the use of explainable AI (SHAP, LIME) to interpret the predictive contribution of each input variable.

1.5 Significance of the Study

The contribution of this research to the bid-ask spread modelling literature are based on the introduction of machine learning techniques capable of capturing non-linear relationships that linear models fail to address. While previous research has explored option pricing and volatility estimation, few studies have focused on spread prediction using real data and modern algorithms. Additionally, this work incorporates interpretability tools to explain the internal logic of predictive models increasing transparency and explainability. By focusing real American options across different issuers and liquidity profiles, this study also provides practical insights for traders, market makers, and regulators seeking to understand or manage liquidity-related costs in fragmented option markets.

1.6 Structure

This paper is structured as follows:
Section 2 reviews the existing literature, combining theoretical perspectives on option pricing and bid-ask spread formation with recent developments in machine learning applied to financial prediction.

Section 3 will describe the research methodology in a quantitative approach justifying each methodological choice.
Section 4 describes the design specification and implementation of the machine learning and neural network techniques.

Section 5 focuses on the implementation phase. It summarizes the final steps taken to operate the methodology

Section 6 presents the evaluation of the models and the key findings of the study.

Finally, **section 7** is a brief reiteration of the initial research questions and summary of the main contributions of the project.

2 Related Work

This section explores the evolution of bid-ask spread modeling, from traditional microstructure-based theories to modern machine learning approaches that aim to predict or explain spreads using high-dimensional market data.

2.1 Classical Theories of Option Pricing

The **Black Scholes model** was the first solution for the price of European options, the model works under several assumptions, such as frictionless markets or constant volatility. Although this model is a standard reference, it failed to account empirical phenomena and the model's assumptions become unrealistic in real markets.(Black and Scholes, 1973)

In response to the limitations of the Black-Scholes model, particularly its inability to capture abrupt movements in the underlying asset price, **Merton** incorporated jump diffusion, allowing for sudden price changes representing an intermediate step between classical and more empirical flexible approaches (Merton, 1976) (Bates, 1995).

Later, **Heston** proposed a stochastic volatility model that captures patterns like volatility skew but still assumes no transaction costs or market frictions (Andersen, 2007) (*Heston Model: Meaning, Overview, Methodology*).

Empirical studies by **Christensen and Prabhala** offers strong empirical support for the hypothesis that implied volatility provides an efficient and unbiased estimate of future realized volatility (RV), reinforcing its value as a predictive input for econometric and ML models (Christensen and Prabhala, 1998).

Other research explores how IV functions vary systematically with moneyness and time to expiry, showing patterns such as volatility smiles (Dumas, Fleming and Whaley, 1996). **Bollen and Whaley** attribute this behavior to order imbalances and constrained arbitrage, directly linking IV to liquidity and market microstructure. (Bollen and Whaley, 2002).

Microstructure models explain spreads as responses to information asymmetry and trading dynamics, **Glosten and Milgrom** show that even in a frictionless market, spreads arise as compensation for trading with informed agents. (Glosten and Milgrom, 1985) Then, **Easley and O'Hara** introduced trade size as an informational signal into spread formation, emphasizing the role of trade aggressiveness (Easley and O'Hara, 1987). This contrasts with Glosten & Milgrom (1985), who treat trades as uniform in size.

Finally, **De Fontnouvelle, Fishe and Harris** demonstrate that inter exchange competition significantly narrows option spreads, highlighting how market structure shapes trading costs (de Fontnouvelle, Fishe and Harris, 2002).

While classical and microstructure models provide theoretical insights, they fall short in capturing the nonlinear, data-rich dynamics of modern markets. This gap is where machine learning offers substantial promise.

2.2 Machine Learning in Option Pricing and Bid Ask Spread Prediction

Hutchinson, Lo and Poggio introduced one of the earliest formal applications of neural networks to the pricing and hedging of derivative securities using market data and avoiding strong theoretical assumptions. However, there are some challenges to address, such as sensitivity to input selection, data requirements, and lack of model interpretability (Hutchinson, Lo and Poggio, 1994). Following the work of Hutchinson, Lo and Poggio, **Gan and Liun** introduced a significant improvement using Residual Neural Networks (ResNet) to overcome gradient vanishing and improve model generalization (Gan and Liu, 2024).

Culkin and Das extend this perspective by training deep feedforward networks to learn the Black-Scholes pricing function from simulated data showing that modern deep learning architectures can approximate complex pricing functions with high precision, and that such models can potentially replicate the behavior of professional traders (Culkin and Das, no date).

Codruț-Florin made a comparative of several machine learning algorithms applied to the pricing of European options and found that tree-based models like XGBoost consistently outperform both traditional models and shallow neural networks in pricing accuracy (Ivașcu, 2021).

In the first paper reviewed, **Sirignano and Cont** applied deep learning to millions of equity order book observations, revealing the potential of neural networks to model market microstructure. (Sirignano and Cont, 2018).

Similarly, **Liou, Liu, and Cheng (2023)** employed CNNs, LSTMs, and GRUs to predict spreads in high-frequency pairs trading, effectively capturing both structural and temporal dependencies (Liou, Liu and Cheng, 2024). In a complementary line, **Avellaneda and Stoikov** developed a high-frequency trading model that links order book dynamics to spread formation via utility-based control, emphasizing the strategic role of liquidity provision (Avellaneda and Stoikov, 2008).

Recent innovations have also introduced physics-inspired architectures, **Hainaut and Casas** employ Physics-Informed Neural Networks (PINNs) to solve the Heston model for option pricing allowing for efficient pricing under stochastic volatility without repeated recalibration. (Hainaut and Casas, 2024).

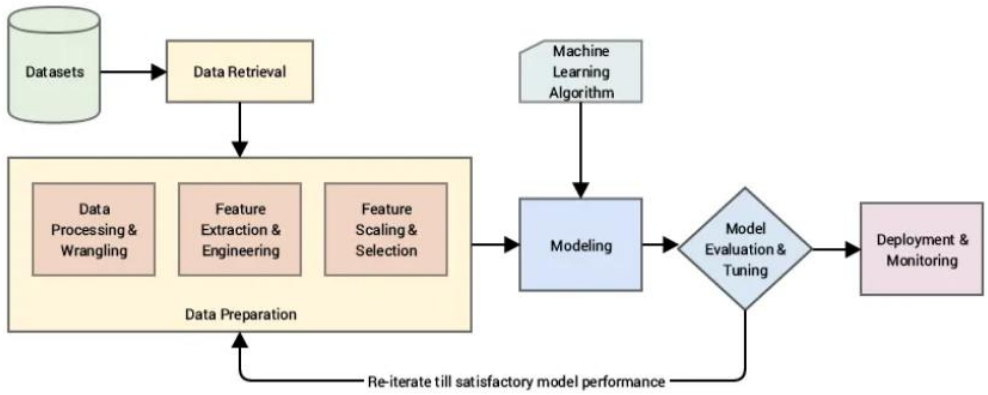
Finally, **John Hull's (2022)** offers a comprehensive overview of how ML techniques are applied to real-world financial problems, including pricing, liquidity, and spread forecasting, reinforcing the idea that ML can detect market signals often missed by classical models (*Hull, J. (2022). Machine Learning in Business and Finance. Wiley.*, no date).

Collectively, these works demonstrate the progression from rigid theoretical structures to flexible, data-driven ML methods, which are better suited for modeling the nonlinearities and frictions inherent in bid-ask spread behavior.

3 Research Methodology

This section outlines the process followed in the research, detailing the data sources, preprocessing steps, feature selection, modeling techniques, and evaluation metrics used to address the research question.

Fig. 1: Research Methodology



Source: (Gunay, 2024)

It is divided into five steps shown in Fig 1:

1. Data Collection: The data used in this study consists of American-style option quotes for the underlying assets AMZN and AMD, collected using the Yahoo Finance API. There are two datasets used in this study that include call and put options. The first dataset contains 13,028 rows of AMZN option data (calls and puts) collected over seven consecutive market days (from June 4, 2025, to June 12, 2025.) and combined into a single dataset after retrieving the data for each individual day. The second dataset contains 1,843 rows of AMD option data collected from one single market day (June 16, 2025). Each dataset contains the following features: bid, ask, strike, lastPrice, volume, impliedVolatility, openInterest, and expirationDate.

2. Data Preprocessing and transformation: After loading the raw data, the preprocessing starts defining two more variables:

-Moneyness, defined as:

Fig 2: Moneyness formula for calls

$$\text{Moneyness} = \frac{S}{K}$$

Where "S" is the underlying stock price and "K" is the strike price. Moneyness indicates whether the option is in- or out-of-the-money, if Moneyness >1 the option is in-the-money and if Moneyness<1 the option is out-the-money.

This definition change for puts where it is defined as follows:

Fig 3: Moneyness formula for puts

$$\text{Moneyness} = \frac{K}{S}$$

Where S is the underlying stock price and K is the strike price. Like for calls, Moneyness >1 the option is in-the-money and Moneyness<1 the option is out-the-money.

-Relative Bid-Ask Spread: This unitless measure represents the percentage difference between ask and bid prices, normalized by the mid-price.

Fig 4: Relative Bid-Ask Spread formula

$$s = \frac{\text{Ask} - \text{Bid}}{(\text{Ask} + \text{Bid})/2}$$

- Tight spreads (0.2%–2%) indicate high liquidity.
- Wide spreads (5%–15%) suggest poor liquidity or high uncertainty.

Then, missing values and outliers such as contracts with bid = 0, volume = 0, or anomalous spreads were removed. Finally, to conclude the preprocessing a few plots, the correlation matrix and descriptive statistics were performed to improve the understanding of the dataset.

4. Modeling and Tuning: This step involves dividing the dataset between features and target, define the train and test dataset, model training and testing and tuning.

In this project, we selected relative bid-ask spread as target feature, we used 80% of the data for training and 20% for testing the models and the following models were implemented and compared:

- Baseline linear model: linear/logistic regression, Ridge or Lasso as an instance of linear model performance.

- Machine Learning models: CatBoost, Random Forest, Gradient Boosting, HistGradienBoosting, XGBoost, KNN Regressor, Decision Tree and Support Vector Regressor.
- **Neural network**: MLP and CNN

Regarding tuning, hyperparameters were tuned using cross-validation for all models optimizing them using Bayesian optimization or Grid search between others.

5. Evaluation Metrics: Involves evaluating the performance of each model predicting the bis-ask spread using MAE, MSE, RMSE and R2 score.
SHAP and LIME were applied to assess model interpretability and feature importance.

4 Design Specification

The implementation of the proposed solution is centered around the analysis and prediction of bid-ask spreads for AMZN and AMD options using machine learning techniques. The objective is to understand how features related to option characteristics, such as moneyness, time to expiration, and implied volatility affect market liquidity as reflected in the spread.
The solution was implemented in Python using google colab and working with the following libraries:

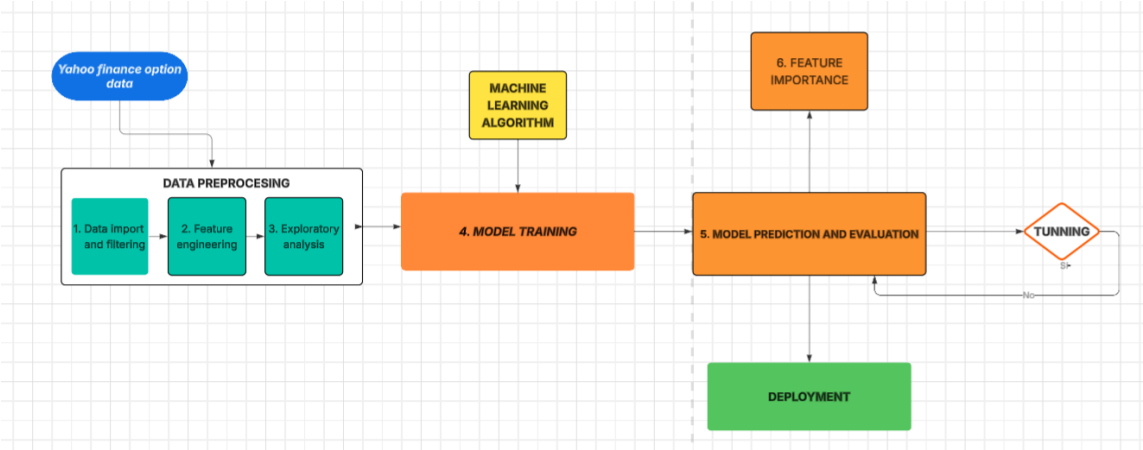
- Data analysis: pandas and numpy
- Data visualization: matplotlib.pyplot, seaborn, plotly.express and matplotlib.patches
- Machine Learning: sklearn.ensemble (Random Forest and Gradient Boosting), sklearn.linear_model, sklearn.svm, sklearn.neighbors, sklearn.tree, sklearn.preprocessing, sklearn.model_selection, sklearn.metrics, GBoost and CatBoost.
- Deep Learning: tensorflow.keras.models, tensorflow.keras.layers, tensorflow.keras.optimizers, tensorflow.keras.callbacks and keras_tuner.
- Feature importance: Shap and Lime
- Optimization: Optuna

4.1 Architecture and Model Development

The modeling architecture consists of a regression-based predictive pipeline aimed at estimating the bid-ask spread using the features mentioned above. A typical supervised learning regression workflow was followed, including:

- **Feature scaling and preprocessing** (if needed).
- **Model training and evaluation** using a suitable algorithm (likely Random Forest, XGBoost, or Linear Regression—exact model to be shown during the viva).
- **Performance assessment** through standard error metrics such as MAE or RMSE (details omitted here due to report length).

Fig 2: Architecture Diagram



Source: Own preparation

The Figure 2 shows the architecture diagram of this project:

1. Data import and filtering: Raw option data is imported from Yahoo finance and filtered to include only the type of options analyzed dropping the observations with unrealistic spreads.
2. Feature engineering: Mid-price is calculated from bid and ask prices, relative spread is computed and used as the prediction target, moneyness is derived from the stock and strike prices and a binary feature (inTheMoney_dummy) is created for ITM classification.
3. Exploratory analysis: The last step of data preprocessing is based on analyse the correlation heatmaps and scatterplots are generated to examine the influence of features like moneyness, implied volatility, and days to expiration on spreads.
4. Model training: Linear , ML models (such as CatBoost, XGBoost, Random Forest, etc) or Neural Networks is trained using selected features with around 80% of the dataset.
5. Prediction and evaluation: The model makes predictions for the spread and is evaluated using MAE, MSE, RMSE and R squared. Then trying to improve models results hyperparameter tunning is applied (if necessary) and tunned models are evaluated with the same metrics.
6. Feature importance: Use feature importance plots, shap and lime to visualize the influence of each feature in the prediction of the model.
7. Deployment: The final model is integrated into a reusable Python pipeline that enables spread prediction for new call option data.

5 Implementation

5.1 Data preprocessing

Using yahoo finance, we obtain 7 days AMZN option data for various strike prices and expiration dates from June 4, 2025, to June 12, 2025, and 1 day AMD option data from June 16, 2025. Each dataset contains the following features:

Variable	Description
Days to expiration (DTE)	Time in days until the option expires
Stock price	Price of the underlying asset
Strike price	The fixed price at which the holder of the option can buy (call) or sell (put) the underlying asset
Last traded price	The most recent transaction price of the option
Bid price	The highest price a buyer is willing to pay

Ask price	The lowest price a seller is willing to accept
Implied volatility	Market expectation of future volatility, typically derived from the Black-Scholes model
Volume	Number of option contracts traded on a given day.
Open interest	Number of outstanding option contracts that have not been settled.
In-the-money dummy	Binary variable: 1 if the option is in-the-money (stock price > strike price), 0 otherwise

In addition to the previous features, we create:

- Moneyness: Defined as:

$$\text{Moneyness} = \frac{S}{K}$$

Where S is the underlying stock price and K is the strike price. If Moneyness>1: In-the-money (ITM) and Moneyness<1: Out-of-the-money (OTM)

- Relative Bid-Ask Spread: Defined as:

$$s = \frac{\text{Ask} - \text{Bid}}{(\text{Ask} + \text{Bid})/2}$$

This unitless measure represents the percentage difference between ask and bid prices, normalized by the mid-price. It is a common liquidity indicator in academic literature.

5.2 AMZN calls implementation

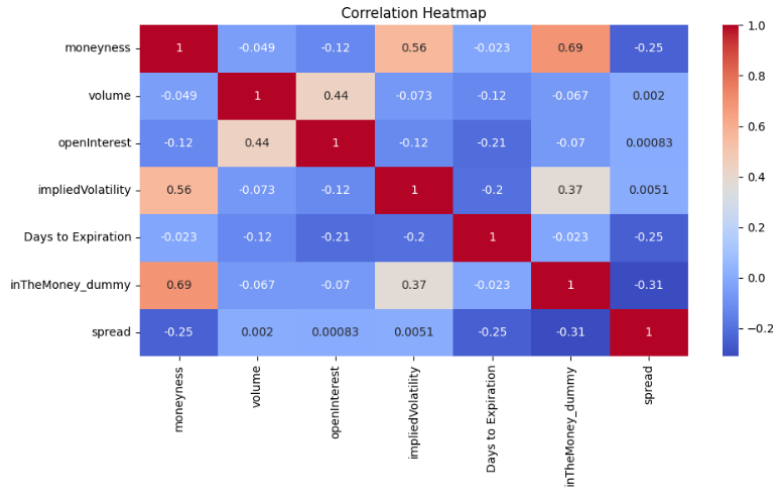
5.2.1 Data cleaning and understanding

After loading the data and create the new features, we start the data cleaning where we convert days to expiration, volume, open interest and implied volatility to numeric and in the money dummy to binary, additionally, we remove entries with null values for key fields and negative or zero spreads. We also drop put values to focus the analysis on calls.

Then, to achieve a better understanding of the data we calculate descriptive statistics, the correlation matrix and a few plots analysing the relationship between implied volatility, days to expiration and moneyness:

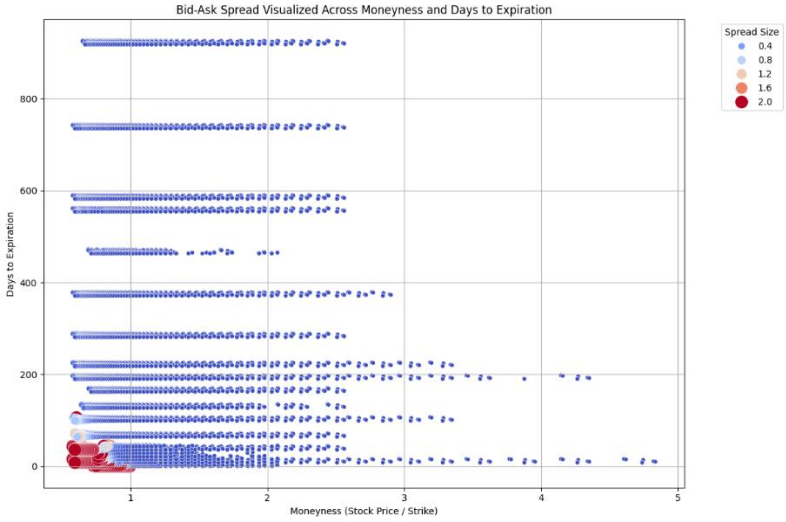
Fig 3: Descriptive statistics and correlation matrix

	moneyness	volume	openInterest	impliedVolatility	Days to Expiration	inTheMoney_dummy	spread
count	6779.000000	6779.000000	6779.000000	6779.000000	6779.000000	6779.000000	6779.000000
mean	1.188435	418.124945	2189.287801	0.535899	258.220829	0.511875	0.146979
std	0.565669	2958.989727	4508.083731	0.454958	264.319444	0.499896	0.425638
min	0.575639	1.000000	0.000000	0.000010	0.000000	0.000000	0.001783
25%	0.791000	2.000000	119.000000	0.331779	38.000000	0.000000	0.011606
50%	1.014195	9.000000	589.000000	0.400824	169.000000	1.000000	0.021053
75%	1.421333	54.000000	2195.500000	0.563115	379.000000	1.000000	0.040268
max	4.835778	80763.000000	50194.000000	7.535157	926.000000	1.000000	2.000000



-Spread statistics show an average transaction cost (14.7) and heavy skew: most options are tight spread, a few very wide (median 2.1%).
-From the correlation matrix we conclude that Volume, Open Interest and Implied Volatility correlation with spread is virtually zero and the feature with the highest correlation with spread is in the money dummy (0.31) followed by moneynesss (0.25) and days to expiration (0.25), this confirms that spread behavior is complex and not well captured by linear correlations alone.

Fig 4: Bid-Ask Spread Visualized Across Moneyness and days to expiration



The key finding of the previous plot are:

•Spread explosions for deep OTM short-dated calls:

The largest red bubbles in the lower-left quadrant (moneyness < 1, low DTE) indicate deep out-of-the-money options with near-term expiry. These are illiquid, rarely traded, and seen as high-risk, leading to very wide spreads to offset hedging and pricing uncertainty.

•Tight spreads around ATM and medium-term options:

Small blue points near moneyness ≈ 1 and DTE between 50–300 days represent actively traded ATM options. These contracts exhibit tight spreads due to strong liquidity and market maker quoting activity.

•Widened spreads for long-dated or extreme moneyness:

Options with extreme moneyness or very long expiration display moderate-to-wide spreads, caused by greater uncertainty, low trading frequency, and limited quoting interest.

This shows a clear relationship between short days to expiration, low moneyness and bigger spreads that we must take into account when predicting the spread.

5.2.2 Modelling

Once we have cleaned the data and improve our understanding of options data, we estimated an OLS regression to see the performance of traditional models with the option relative bid-ask spread prediction using six explanatory variables: moneyness, days to expiration, implied volatility, volume, open interest, and a dummy variable for whether the option is in the money.

Fig 5: Linear Regression AMZN calls

OLS Regression Results

Dep. Variable:	spread	R-squared:	0.185
Model:	OLS	Adj. R-squared:	0.185
Method:	Least Squares	F-statistic:	256.8
Date:	Fri, 01 Aug 2025	Prob (F-statistic):	7.08e-297
Time:	10:02:58	Log-Likelihood:	-3133.3
No. Observations:	6779	AIC:	6281.
Df Residuals:	6772	BIC:	6328.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4549	0.013	34.637	0.000	0.429	0.481
impliedVolatility	0.1233	0.013	9.610	0.000	0.098	0.148
moneyness	-0.1166	0.013	-9.070	0.000	-0.142	-0.091
Days to Expiration	-0.0004	1.87e-05	-21.805	0.000	-0.000	-0.000
volume	-2.13e-06	1.76e-06	-1.209	0.227	-5.58e-06	1.32e-06
openInterest	-6.403e-06	1.18e-06	-5.405	0.000	-8.73e-06	-4.08e-06
inTheMoney_dummy	-0.2254	0.013	-17.458	0.000	-0.251	-0.200

Omnibus:	4252.298	Durbin-Watson:	0.231
Prob(Omnibus):	0.000	Jarque-Bera (JB):	37088.476
Skew:	3.019	Prob(JB):	0.00
Kurtosis:	12.739	Cond. No.	2.07e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.07e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 6 shows:

•**R-squared = 0.185**: The model explains 18.5% of the variance in the spread. These modest results show that the linear regression captures some relationships in the data but leaves substantial variability unexplained.

•**F-statistic = 256.8, p < 0.0001**: The overall model is highly statistically significant, indicating that the independent variables jointly explain a non-negligible portion of the variation in the spread.

•**Implied volatility and ITM status** are the strongest predictors of the bid-ask spread in this model while moneyness and open interest also play meaningful roles. In contrast **Days to expiration** (highly relevant according to correlation matrix and plots) has a coeficient near to zero that make us thing on a nonlinear relationship not captured by the model.

The linear model provides important economic insights, but its performance is limited by linearity and distributional assumptions. It's a strong benchmark and interpretability reference, but for prediction-focused tasks (e.g., in trading systems), machine learning models may offer superior fit and generalization.

Based on that, trying to improve linear results we transition to machine learning models getting the following results:

Fig 6: ML models results

	MAE	MSE	RMSE	R2 Score
CatBoost	0.027126	0.008042	0.089677	0.956834
HistGradientBoosting	0.030211	0.010689	0.103386	0.942628
Random Forest	0.025947	0.010952	0.104652	0.941214
XGBoost	0.027465	0.011172	0.105700	0.940031
Gradient Boosting	0.040272	0.013609	0.116656	0.926954
KNN Regressor	0.033607	0.016972	0.130275	0.908903
Decision Tree	0.032242	0.023452	0.153140	0.874120
Support Vector Regressor	0.085547	0.026462	0.162672	0.857961
Linear Regression	0.224516	0.150899	0.388457	0.190038
Ridge	0.224506	0.150900	0.388458	0.190033
Lasso	0.218473	0.186328	0.431657	-0.000131

CatBoost stands out as the top-performing model achieving the lowest MAE (0.0271), lowest MSE (0.0080), lowest RMSE (0.0897) and with the highest R^2 score at 0.9568, meaning it explains over 95.7% of the variance in the relative spread.
HistGradientBoosting, Random Forest and XGBoost also get high results close to CatBoost being capable of capturing nonlinear relationships, feature interactions, and complex structures often present in option market microstructure.

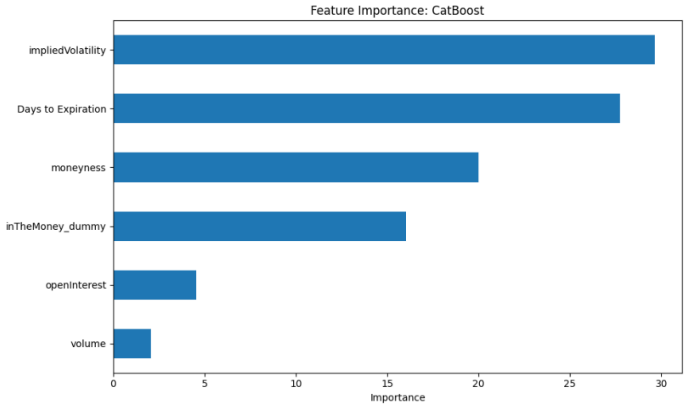
Linear models (Linear Regression, Ridge and Lasso) perform poorly because market microstructure effects are nonlinear and the interactions matter, this is shown in the importance of days to expiration that is not reflected in linear regression. The effect of moneyness may depend on days to expiration or implied volatility as we saw in the previous plots.

The excellent performance of CatBoost, HistGradientBoosting, and XGBoost validates the use of advanced machine learning methods in predicting key pricing variables like the bid-ask spread demonstrating the limitations of traditional linear approaches and support the integration of ML in market microstructure and risk-pricing models.

5.2.3 Feature importance

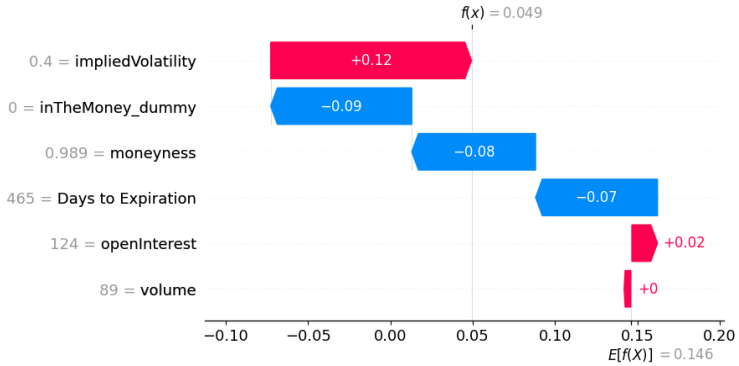
In this section we evaluate the importance of each feature in the best performing model to increase transparency and to know with features have more influence to predict the spread. To do it we use a bar plot, Shap and Lime.

Fig 7: Feature importance CatBoost



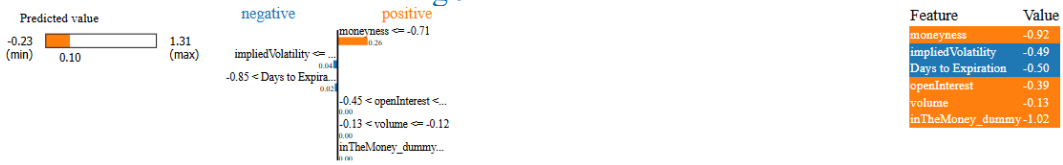
From de bar plot we conclude that the three features with higher importance are IV, Days to expiration and moneyness. It is also essential to perform SHAP values to explain the impact of each feature on individual predictions, offering a granular and model-agnostic explanation.

Fig 8: Shap CatBoost



Shap shows that Higher IV increases the predicted spread (SHAP ≈ +0.12), ITM options significantly reduce the spread (SHAP ≈ -0.09) based on ITM dummy and moneyness and Long-dated options have wider spreads (SHAP ≈ -0.07).

Fig 9: Lime CatBoost



Finally, lime highlights the same three features as most relevant. The CatBoost model predicts a negative spread of -0.24 driven by its significantly out-of-the-money status (Moneyness: -0.92), which is the most impactful factor pushing towards a negative spread, lower implied volatility (-0.49) and fewer days to expiration (-0.50).

The LIME explanation effectively shows that for this specific instance, a combination of these factors, especially its deep out-of-the-money status, leads to the predicted negative spread. The blue color in the final predicted value bar visually confirms the "negative" outcome of the spread.

5.2.4 Hyperparameter tuning

After the implementation of the previous machine learning models, we try to improve their results by applying hyperparameter tuning using Grid or Bayesian, the results are the following:

Fig 10: Hyperparameter tuning metrix

	MAE	MSE	RMSE	R2 Score
CatBoost	0.027126	0.008042	0.089677	0.956834
CatBoost_tuned	0.029647	0.008188	0.090487	0.956051
XGBoost_tuned	0.035351	0.009800	0.098995	0.947397
HistGradientBoosting_tuned	0.029913	0.010415	0.102056	0.944095
HistGradientBoosting	0.030211	0.010689	0.103386	0.942628
Random Forest_tuned	0.026040	0.010736	0.103615	0.942373
Random Forest	0.025947	0.010952	0.104652	0.941214
XGBoost	0.027465	0.011172	0.105700	0.940031
Gradient Boosting	0.040272	0.013609	0.116656	0.926954
KNN Regressor	0.033607	0.016972	0.130275	0.908903
Decision Tree	0.032242	0.023452	0.153140	0.874120
Support Vector Regressor	0.085547	0.026462	0.162672	0.857961
Linear Regression	0.224516	0.150899	0.388457	0.190038
Ridge	0.224506	0.150900	0.388458	0.190033
Lasso	0.218473	0.186328	0.431657	-0.000131

Fig 10 shows a comparation of all models before and after tuning, as we see all models improve their results except CatBoost that remains the best model without tunning, it is important to highlight the high increase of XGBoost after tunning that improve its results from 0.941 to 0.947 being the second best model.

5.2.5 Neural Networks

To complete the modelling we decided to implement two neural networks, Multi-Layer Perceptron (MLP) and Convolutional Neural Network (CNN) with the following results:

Fig 11: All models classification

	MAE	MSE	RMSE	R2 Score
CatBoost	0.027126	0.008042	0.089677	0.956834
CatBoost_tuned	0.029647	0.008188	0.090487	0.956051
XGBoost_tuned	0.035351	0.009800	0.098995	0.947397
HistGradientBoosting_tuned	0.029913	0.010415	0.102056	0.944095
HistGradientBoosting	0.030211	0.010689	0.103386	0.942628
Random Forest_tuned	0.026040	0.010736	0.103615	0.942373
Random Forest	0.025947	0.010952	0.104652	0.941214
XGBoost	0.027465	0.011172	0.105700	0.940031
Gradient Boosting	0.040272	0.013609	0.116656	0.926954
MLP Neural Network	0.040440	0.014755	0.121471	0.920801
1D CNN	0.040927	0.015844	0.125874	0.914955
KNN Regressor	0.033607	0.016972	0.130275	0.908903
Decision Tree	0.032242	0.023452	0.153140	0.874120
Support Vector Regressor	0.085547	0.026462	0.162672	0.857961
Linear Regression	0.224516	0.150899	0.388457	0.190038
Ridge	0.224506	0.150900	0.388458	0.190033
Lasso	0.218473	0.186328	0.431657	-0.000131

Both MLP and CNN got high results with an R2 Score of 0.92 and 0.915 respectively, although these results show a great model performance they don't improve most of the machine learning models.

The last step is to apply tuning to NN trying to improve their results and improve ML models

Fig 12: Neural Networks after tuning

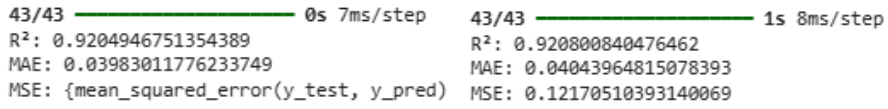


Fig 12 shows the results of MLP (1st image) and CNN (2nd image) after tuning, as we see only CNN improve its results but still not getting the level of ML models such as CatBoost or XGBoost.

5.3 AMZN puts implementation

In the second part, to extend the analysis we see the puts behaviour using the same dataset.

5.3.1 Data cleaning and understanding

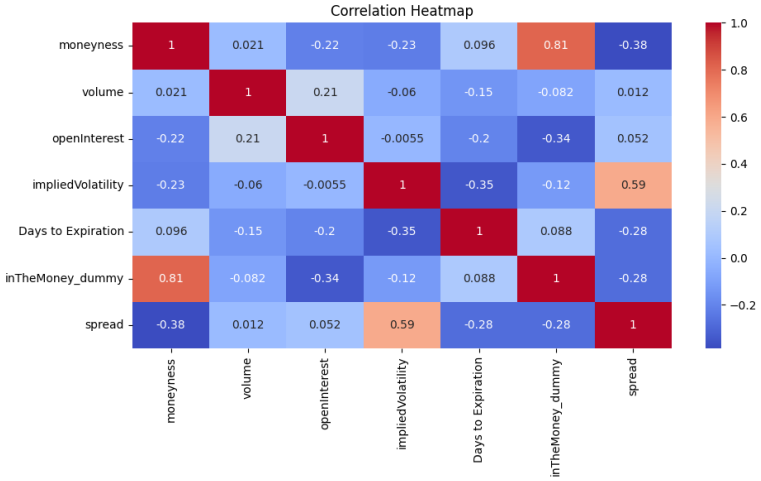
We did the same as for the calls selecting puts only with the difference of changing the moneyness definition as follows:

$$\text{Moneyness} = \frac{K}{S}$$

where S is the underlying stock price and K is the strike price. If Moneyness>1the option is in-the-money (ITM) and if Moneyness<1 the option is out-of-the-money (OTM).

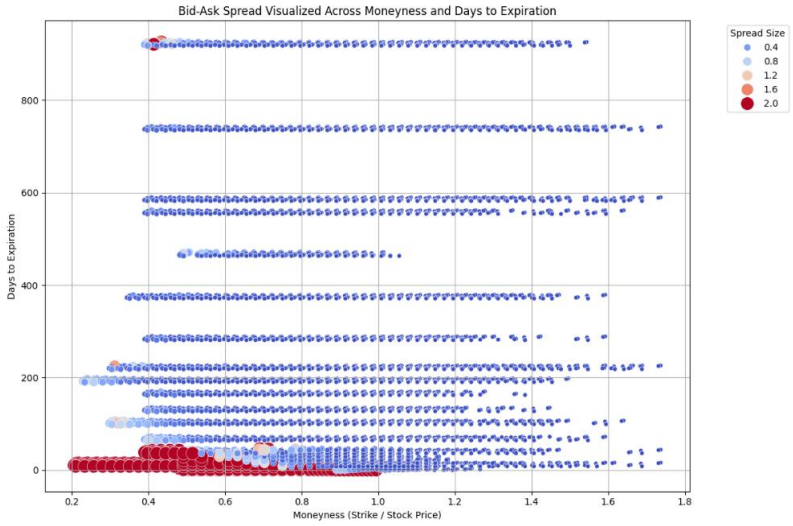
Then we did the correlation matrix where we see a similar behaviour with the difference of a notable increase of implied volatility correlation with spread. This can be explained because for puts, higher implied volatility reflects more uncertainty and risk, leading market makers to widen spreads aligning with market microstructure theory (risk compensation).

Fig 13: Correlation matrix AMZN puts



The days to expiration vs moneyness vs spread plot (Fig 16) shows again that there is an extreme spread concentration for deep OTM Puts with bigger spreads located in the bottom left-hand corner.

Fig 14: Bid-Ask Spread Visualized Across Moneyness and days to expiration



5.3.2 Modelling

We used the same combination of linear, machine learning and Neural Networks models that for calls and then applied hyperparameter tuning getting the following results:

Fig 15: Models results classification for AMZN puts

	MAE	MSE	RMSE	R2 Score
CatBoost_tuned	0.043441	0.018219	0.134977	0.920630
CatBoost	0.044573	0.018331	0.135392	0.920141
HistGradientBoosting	0.047196	0.020346	0.142640	0.911362
Random Forest_tuned	0.041693	0.020393	0.142804	0.911158
Random Forest	0.041870	0.020475	0.143090	0.910801
Gradient Boosting	0.052017	0.020649	0.143699	0.910040
HistGradientBoosting_tuned	0.048256	0.020772	0.144125	0.909507
XGBoost_tuned	0.056218	0.022287	0.149289	0.902906
XGBoost	0.046006	0.022960	0.151527	0.899973
KNN Regressor	0.050237	0.029284	0.171126	0.872424
MLP Neural Network	0.060296	0.030863	0.175678	0.865546
Decision Tree	0.049700	0.032652	0.180698	0.857752
1D CNN	0.070337	0.034650	0.186146	0.849046
Support Vector Regressor	0.089793	0.036197	0.190255	0.842307
Ridge	0.226196	0.141813	0.376581	0.382190
Linear Regression	0.226192	0.141816	0.376585	0.382177
Lasso	0.280242	0.229881	0.479459	-0.001478

Although the linear model performs notably better for puts ($R^2 \approx 0.38$) than for calls ($R^2 \approx 0.18$), the results still indicate a relatively weak explanatory power that suggests that the model is not capturing nonlinear relationships that other nonlinear models.

From ML models, like in call options, CatBoost is the best model with a MAE of 0.043, a MSE of 0.018, a RMSE of 0.135 and a R2 of 0.92 followed by the Random Forest with 0.91 of R2.
The results are lower for puts because they are harder to predict due to:

- Higher skewness in moneyness.
- Wider distribution of spreads, especially for DOTM puts.
- Lower average open interest and volume.

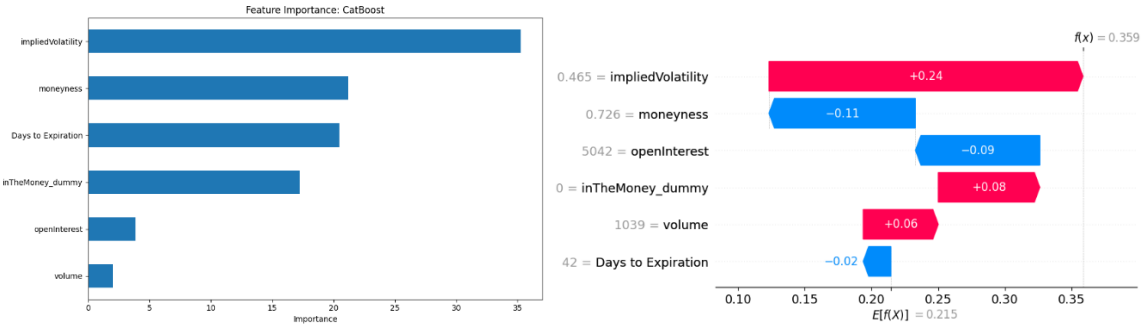
After **hyperparameter tuning** almost all the machine learning models improved and CatBoost improve its results being the best model like for AMZN Calls.

Regarding Neural Networks MLP and CNN were applied and like for AMZN call options, both MLP and CNN improve the results of linear models but achieve lower results that machine learning models such as CatBoost (0.92) or XGBoost (0.918).

5.3.3 Feature importance

We evaluate the importance of each feature in the best performing model (CatBoost) to increase transparency and to know with features have more influence to predict the spread. To do it we use a bar plot, Shap and Lime.

Fig 16: Feature importance, SHAP and LIME CatBoost



Type	Calls	Puts
Feature importance	IV> DTE > moneyness	IV> moneyness > DTE
SHAP	IV increases spread, ITM dummy and moneyness reduce it	IV and moneyness both increase spread, DTE effects are mixed
LIME	Split a low moneyness and low volume: wide spread	High moneyness and high IV: wide spread, DTE and OI help reduce it

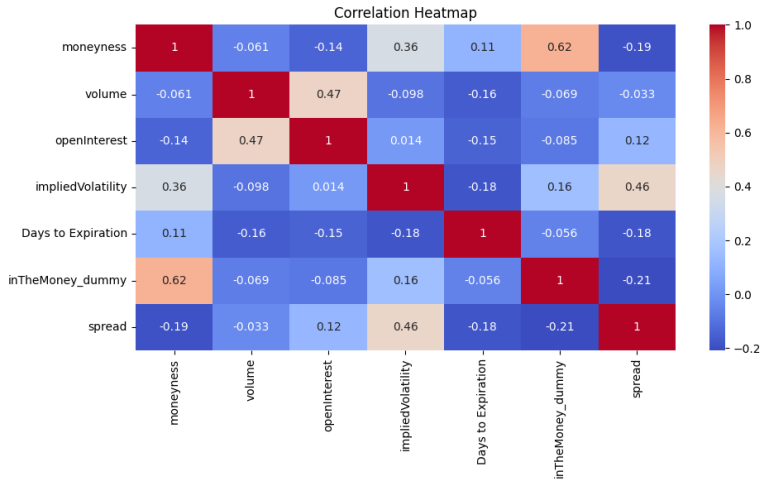
5.4 AMD calls implementation

To conclude, we analyse if there is a patron in common with other companies, we perform the analysis to AMD one day Call options from 16 of June of 2025.
For this analysis we used again yahoo finance to obtain AMD option data for various strike prices and expiration dates from June 16, 2025with the same features of the AMZN dataset used previously.

5.4.1 Data cleaning and understanding

For data cleaning we did the same as for AMZN calls starting with the correlation matrix:

Fig 17: Correlation matrix AMD calls

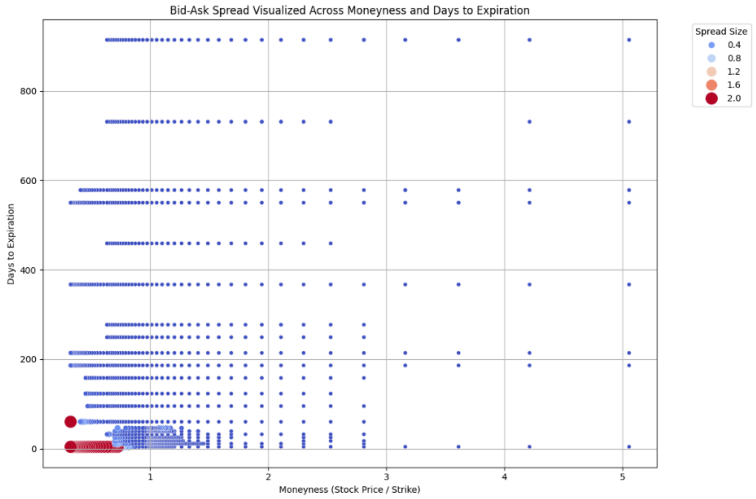


- For AMD, Open Interest and Implied Volatility have higher correlation with spread than AMZN (where is virtually zero)
- The feature with the highest correlation with spread is in the money dummy (0.21) followed by moneynesss (0.19) and days to expiration (0.18)

The correlation matrix for AMD shows a few differences with the AMZN one but the features with higher correlation with spread remain the same.

Then, in the days to expiration vs moneyness vs spread plot we see the same relationship between these three features for AMD and AMZN. Larger spreads are located in the bottom left-hand corner with shorts days to expiration and low moneyness.

Fig 18: Bid-Ask Spread Visualized Across Moneyness and days to expiration



5.4.2 Modelling

We apply the same models for AMD call options showing the following results:

Fig 19: Models results classification for AMZN puts

	MAE	MSE	RMSE	R2 Score
CatBoost_tuned	0.011061	0.000713	0.026708	0.995985
Random Forest_tuned	0.011183	0.000935	0.030582	0.994735
CatBoost	0.011220	0.000967	0.031101	0.994555
Decision Tree	0.011304	0.001060	0.032564	0.994030
Gradient Boosting	0.012465	0.001126	0.033550	0.993663
Random Forest	0.011708	0.001178	0.034321	0.993369
XGBoost_tuned	0.016591	0.001666	0.040811	0.990624
HistGradientBoosting_tuned	0.018700	0.001891	0.043485	0.989355
HistGradientBoosting	0.018491	0.001938	0.044021	0.989091
MLP Neural Network	0.022827	0.002788	0.052805	0.984303
1D CNN	0.026000	0.003500	0.059157	0.980299
KNN Regressor	0.014170	0.003880	0.062286	0.978160
XGBoost	0.018247	0.006234	0.078956	0.964906
Support Vector Regressor	0.059148	0.008878	0.094221	0.950023
Linear Regression	0.137149	0.087970	0.296598	0.504773
Ridge	0.137145	0.088086	0.296794	0.504119
Lasso	0.152911	0.179675	0.423881	-0.011478

Again, the linear model performs better for AMD ($R^2 \approx 0.36$) than for AMZN ($R^2 \approx 0.18$), the results still indicate a relatively weak explanatory power. Again, it's clear that a simple linear framework fails to capture the complexity of option market microstructure, especially for deep OTM contracts and short maturities.

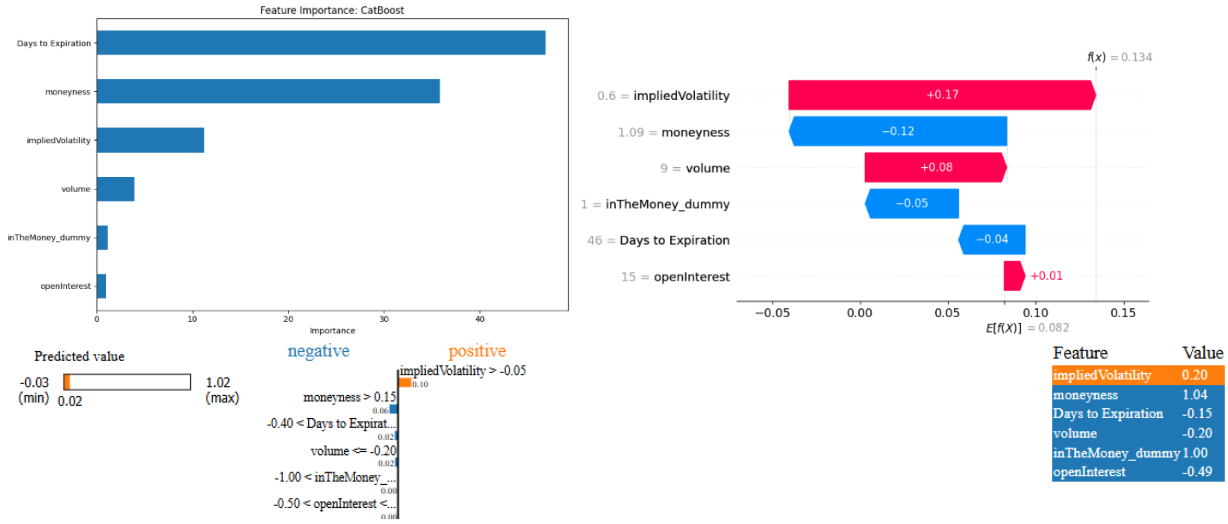
Regarding ML models, CatBoost is the best model with even better results: MAE of 0.0112, an MSE of 0.0009, a RMSE of 0.0311 and a R2 of 0.9945 followed by the Random Forest with 0.9940 of R2. All machine learning models perform better than linear ones and improve their performance after tuning being CatBoost tuned the first in terms of R2 Score.

Like for AMZN both Neural Networks models get high results but don't improve most of the machine learning models.

5.4.3 Feature importance

We evaluate the importance of each feature in the best performing model (CatBoost) to increase transparency and to know with features have more influence to predict the spread. To do it we use a bar plot, Shap and Lime.

Fig 20: Feature importance, SHAP and LIME CatBoost



Implied Volatility is consistently the most influential feature but not always ranked first, days to expiration acts as a regular spread suppressor across all types and moneyness has a dual role: predictive importance and spread-reducing power.

6 Evaluation

The aim of this section is to provide a comprehensive analysis of the results and main findings of the study. It is divided into for subsections: Linear models, machine learning models, feature importance and Neural networks

6.1 Linear models

As we have seen, linear models do not get high results when predicting the spread for all AMZN calls and puts and AMD call due to:

- Market microstructure effects are nonlinear
- The effect of moneyness may depend on days to expiration or implied volatility as we saw in the previous plots.
- Tree-based models like CatBoost and Random Forest automatically capture such interactions.
- Bid-ask spreads often exhibit heavy-tailed distributions.
- Models like CatBoost are robust to outliers and multicollinearity, where linear models often break down.

6.2 Machine Learning models

Fig 21: Machine Learning models results comparation

	MAE	MSE	RMSE	R2 Score
CatBoost	0.027126	0.008042	0.089677	0.956834
CatBoost_tuned	0.029647	0.008188	0.090487	0.956051
XGBoost_tuned	0.035351	0.009800	0.098995	0.947397
HistGradientBoosting_tuned	0.029913	0.010415	0.102056	0.944095
HistGradientBoosting	0.030211	0.010689	0.103386	0.942628
Random Forest_tuned	0.026040	0.010736	0.103615	0.942373
Random Forest	0.025947	0.010952	0.104652	0.941214
XGBoost	0.027465	0.011172	0.105700	0.940031
Gradient Boosting	0.040272	0.013609	0.116656	0.926954
MLP Neural Network	0.040440	0.014755	0.121471	0.920801
1D CNN	0.040927	0.015844	0.125874	0.914955
KNN Regressor	0.033607	0.016972	0.130275	0.908903
Decision Tree	0.032242	0.023452	0.153140	0.874120
Support Vector Regressor	0.085547	0.026462	0.162672	0.857961
Linear Regression	0.224516	0.150899	0.388457	0.190038
Ridge	0.224506	0.150900	0.388458	0.190033
Lasso	0.218473	0.186328	0.431657	-0.000131

	MAE	MSE	RMSE	R2 Score
CatBoost_tuned	0.043441	0.018219	0.134977	0.920630
CatBoost	0.044573	0.018331	0.135392	0.920141
HistGradientBoosting	0.047196	0.020346	0.142640	0.911362
Random Forest_tuned	0.041693	0.020393	0.142804	0.911158
Random Forest	0.041870	0.020475	0.143090	0.910801
Gradient Boosting	0.052017	0.020649	0.143699	0.910040
HistGradientBoosting_tuned	0.048256	0.020772	0.144125	0.909507
XGBoost_tuned	0.056218	0.022287	0.149289	0.902906
XGBoost	0.046006	0.022960	0.151527	0.899973
KNN Regressor	0.050237	0.029284	0.171126	0.872424
MLP Neural Network	0.060296	0.030863	0.175678	0.865546
Decision Tree	0.049700	0.032652	0.180698	0.857752
1D CNN	0.070337	0.034650	0.186146	0.849046
Support Vector Regressor	0.089793	0.036197	0.190255	0.842307
Ridge	0.226196	0.141813	0.376581	0.382190
Linear Regression	0.226192	0.141816	0.376585	0.382177
Lasso	0.280242	0.229881	0.479459	-0.001478

	MAE	MSE	RMSE	R2 Score
CatBoost_tuned	0.011061	0.000713	0.026708	0.995985
Random Forest_tuned	0.011183	0.000935	0.030582	0.994735
CatBoost	0.011220	0.000967	0.031101	0.994555
Decision Tree	0.011304	0.001060	0.032564	0.994030
Gradient Boosting	0.012465	0.001126	0.033550	0.993663
Random Forest	0.011708	0.001178	0.034321	0.993369
XGBoost_tuned	0.016591	0.001666	0.040811	0.990624
HistGradientBoosting_tuned	0.018700	0.001891	0.043485	0.989355
HistGradientBoosting	0.018491	0.001938	0.044021	0.989091
MLP Neural Network	0.022827	0.002788	0.052805	0.984303
1D CNN	0.026000	0.003500	0.059157	0.980299
KNN Regressor	0.014170	0.003880	0.062286	0.978160
XGBoost	0.018247	0.006234	0.078956	0.964906
Support Vector Regressor	0.059148	0.008878	0.094221	0.950023
Linear Regression	0.137149	0.087970	0.296598	0.504773
Ridge	0.137145	0.088086	0.296794	0.504119
Lasso	0.152911	0.179675	0.423881	-0.011478

For the three datasets, Machine Learning models outperform the linear models, in concrete CatBoost got the best results for both AMD and AMZN. Hyperparameter tunning improve the results of almost all the models being the exception the CatBoost for AMZN call which normal model did no improve but got the higher results.

The models achieve better results when predicting the spread of call options, specially for AMD due to the dataset is smaller what means a low number of outliers.

6.3 Feature importance

Data	AMZN calls	AMZN puts	AMD calls
Feature importance	IV > DTE > moneyness	IV > moneyness > DTE	DTE > moneyness > IV
SHAP	High IV pushes spread up; High DTE and ITM status reduce it	High IV and moneyness increase spread, DTE and OI reduce it	High IV pushes spread up, moneyness reduces it and DTE and ITM dummy reduce it
LIME	Split at low moneyness and low volume: wide spread	High moneyness and high IV: wide spread. DTE and OI help reduce it	Implied Volatility is the strongest positive contributor, the strongest negative is moneyness

The main conclusion we obtain about feature importance are that always the same three features (moneyness, days to expiration and implied volatility) are the most important for spread prediction.

6.4 Neural networks

As we have seen before, Multi-Layer Perceptron (MLP) and Convolutional Neural Network (CNN) get high results for both AMZN and AMD outperforming linear models but even after applying hyperparameter tunning they did not improve the results of most of the Machine Learning models.

7 Conclusion and Future Work

7.1 Conclusions

The main conclusions of this project are:

- Variables such as days to expiration and moneyness exhibit non-linear relationships with the bid-ask spread that linear models like logistic regression fail to capture showing that days to expiration is not statistically significant in AMD call options model and giving a coefficient near 0 in AMZN puts and calls models.
- Nonlinear models (Like CatBoost or XGBoost) outperform linear ones (Logistic Regression) in predicting option spreads by capturing complex interactions, U-shaped patterns, and handling outliers effectively.
- For machine learning models the three main features are moneyness, days to expiration and implied volatility being the last one the most important.
- Hyperparameter tuning improves almost all ML models.
- Explainability techniques like SHAP and LIME show that High IV pushes spreads up, high DTE and ITM status reduce spreads and the interaction between variables is complex and non-linear.

- Neural Network such as MLP and CNN improve the linear models results but fail to improve most of machine learning models.
- Days to expiration, moneyness and spread maintain their relationship for AMZN puts and AMD calls as is shown in the plots.
- Puts show more complex interactions between features like IV and moneyness.

7.2 Future work

While the current project demonstrates the feasibility of predicting bid-ask spreads for AMZN and AMD options using machine learning techniques, several areas for future work have been identified:

- The current model is static and cross-sectional. Incorporating temporal features (using LSTM or time-aware models) would allow prediction of how spreads evolve over time, especially around earnings.

- Extend the analysis to AMD puts or other companies to continue confirming the existence of a common patron when predicting spread.

- Testing the model during high-volatility periods (e.g., during earnings calls, FOMC meetings, or geopolitical events) could validate its behavior under extreme market conditions.

References

Andersen, L.B.G. (2007) 'Efficient Simulation of the Heston Stochastic Volatility Model', *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.946405.

Andersen, T. *et al.* (2021) 'A Descriptive Study of High-Frequency Trade and Quote Option Data*', *Journal of Financial Econometrics*, 19(1), pp. 128–177. Available at: https://doi.org/10.1093/jjfinec/nbaa036.

Avellaneda, M. and Stoikov, S. (2008) 'High-frequency trading in a limit order book', *Quantitative Finance*, 8(3), pp. 217–224. Available at: https://doi.org/10.1080/14697680701381228.

Bates, D.S. (1995) 'Testing Option Pricing Models'. Rochester, NY: Social Science Research Network. Available at: https://papers.ssrn.com/abstract=225194 (Accessed: 7 June 2025).

Black, F. and Scholes, M. (1973) 'The Pricing of Options and Corporate Liabilities', *The Journal of Political Economy*, 81(3), pp. 637–654.

Bollen, N.P.B. and Whaley, R. (2002) 'Does Net Buying Pressure Affect the Shape of Implied Volatility Functions?' Rochester, NY: Social Science Research Network. Available at: https://doi.org/10.2139/ssrn.319261.

Christensen, B.J. and Prabhala, N.R. (no date) 'The relation between implied and realized volatility'.

Culkin, R. and Das, S.R. (no date) 'Machine Learning in Finance: The Case of Deep Learning for Option Pricing'.

Dumas, B., Fleming, J. and Whaley, R. (1996) 'Implied Volatility Functions: Empirical Tests'. Rochester, NY: Social Science Research Network. Available at: https://papers.ssrn.com/abstract=7373 (Accessed: 9 June 2025).

Easley, D. and O'Hara, M. (1987) 'Price, trade size, and information in securities markets', *Journal of Financial Economics*, 19(1), pp. 69–90. Available at: https://doi.org/10.1016/0304-405X(87)90029-8.

de Fontnouvelle, P., Fishe, R.P.H. and Harris, J.H. (2002) 'The Behavior of Bid-Ask Spreads and Volume in Options Markets During the Competition for Listings in 1999'. Rochester, NY: Social Science Research Network. Available at: https://doi.org/10.2139/ssrn.363600.

Gan, L. and Liu, W. (2024) 'Option Pricing Based on the Residual Neural Network', *Computational Economics*, 63(4), pp. 1327–1347.

Glosten, L.R. and Milgrom, P.R. (1985) 'Bid, ask and transaction prices in a specialist market with heterogeneously informed traders', *Journal of Financial Economics*, 14(1), pp. 71–100. Available at: https://doi.org/10.1016/0304-405X(85)90044-3.

Gunay, D. (2024) 'Feature Engineering & Data Preprocessing', *Medium*, 30 April. Available at: https://medium.com/@denizgunay/feature-engineering-data-preprocessing-d6bc219b6b93 (Accessed: 31 July 2025).

Hainaut, D. and Casas, A. (2024) 'Option pricing in the Heston model with physics inspired neural networks', *Annals of Finance*, 20(3), pp. 353–376. Available at: https://doi.org/10.1007/s10436-024-00452-7.

Heston Model: Meaning, Overview, Methodology (no date) *Investopedia*. Available at: https://www.investopedia.com/terms/h/heston-model.asp (Accessed: 7 June 2025).

Hull, J. (2022). Machine Learning in Business and Finance. Wiley. (no date).

Hutchinson, J.M., Lo, A.W. and Poggio, T. (1994) 'A Nonparametric Approach to Pricing and Hedging Derivative Securities Via Learning Networks'. Rochester, NY: Social Science Research Network. Available at: https://papers.ssrn.com/abstract=236673 (Accessed: 10 June 2025).

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

Liou, J.-H., Liu, Y.-T. and Cheng, L.-C. (2024) 'Price spread prediction in high-frequency pairs trading using deep learning architectures', *International Review of Financial Analysis*, 96(PB). Available at: https://ideas.repec.org//a/eee/finana/v96y2024ipbs1057521924007257.html (Accessed: 13 June 2025).

Merton, R.C. (1976) 'Option pricing when underlying stock returns are discontinuous', *Journal of Financial Economics*, 3(1), pp. 125–144. Available at: https://doi.org/10.1016/0304-405X(76)90022-2.

Sirignano, J. and Cont, R. (2018) 'Universal features of price formation in financial markets: perspectives from Deep Learning'. arXiv. Available at: https://doi.org/10.48550/arXiv.1803.06917.

Vallelado González, E. (1992) 'Options evolution: the introduction of organized markets in the U.S.A.', *Anales de estudios económicos y empresariales*, (7), pp. 97–110.