

Study of Implied Volatility Surface with Machine Learning Techniques

Sérgio da Cunha Sobral

Master in Financial Mathematics

Supervisors:

PhD Diana Elisabeta Aldea Mendes, Associate Professor,
ISCTE-IUL

October, 2025

Finance Department

Mathematics Department

Study of Implied Volatility Surface with Machine Learning Techniques

Sérgio da Cunha Sobral

Master in Financial Mathematics

Supervisors:

PhD Diana Elisabeta Aldea Mendes, Associate Professor,
ISCTE-IUL

October, 2025

*“Que sempre que um homem sonha, o mundo pula e avança como bola colorida entre
as mãos de uma criança.”*

António Gedeão

A todos aqueles que me permitem sonhar mais alto.

Acknowledgements

This dissertation is the result of the dedication, guidance, and support of many people, whose contributions I gratefully acknowledge.

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Doutora Diana Aldea Mendes. From the moment she accepted this challenge, her support has been constant and invaluable. I am deeply thankful not only for the knowledge, guidance, and encouragement she provided, but also for making this journey so much more enjoyable through the remarkable person she is.

I am also thankful to Professor Doutor João Pedro Nunes and Professor Doutor João Pedro Ruas for kindly providing access to the SPX options dataset, which was fundamental for this research.

To my colleagues and friends, Afonso and José, I am grateful for having found such genuine friendship as a side effect of this adventure. And to Cláudio, whose friendship goes back to our undergraduate years and has since grown stronger, it brings me great joy to close yet another chapter alongside you. I will always carry the memories we built together — the four of us — from the toughest moments to the dinners filled with laughter and good energy.

To my family, a thank you that words can never fully express. To my grandparents, my father Ilídio, my mother Fernanda, and my sister Letícia — your unconditional love and unwavering trust in my choices are my greatest strength. Taking bold steps feels easier when one knows there is solid ground of love and support to fall back on.

Finally, a special thank you to my girlfriend, Sara, for encouraging me to take the first step and for your unconditional support throughout this journey. The path was not always an easy one, but having you by my side — with your patience, kindness, and love — made me never lose sight of where I was heading. We made it together.

Abstract

Modelling the implied volatility surface (IVS) is fundamental for option pricing and for understanding the structure of volatility in financial markets. This work analyzed different approaches to IVS modelling, from classical models to machine learning techniques, including hybrid models that combine both paradigms.

Daily data on SPX options provided by the CBOE were used, covering the period from 2009 to 2019. The main explanatory variables were log-moneyness and time to maturity, complemented, in the machine learning models, by the risk-free rate and a binary option-type indicator, to assess the additional contribution of these variables.

All models followed the same training, tuning, and test process, ensuring comparability. Performance was evaluated using error metrics, such as R^2 and RMSE, complemented by residual analyses. The classical models — linear regression (OLS) and a fifth-degree polynomial model — highlighted the nonlinear nature of the IVS, whereas the cubic B-spline model achieved significant gains, though some residual structure remained unexplained.

The machine learning techniques (Random Forest and XGBoost) showed substantial improvements in the metrics and almost eliminated residual structure, demonstrating a superior ability to adapt to complex patterns. The hybrid models, which apply machine learning to the spline model residuals, achieved results equivalent to or slightly better than those of the standalone machine learning models while preserving an interpretable foundation.

This outcome confirms that integrating classical and machine learning models provides an effective balance between structure and flexibility, offering a robust framework for IVS modelling.

Keywords: Implied Volatility Surface, Machine Learning Techniques, Hybrid Models, SPX Options

JEL Classification: C45, C53

Resumo

A modelização da superfície de volatilidade implícita (IVS) é fundamental para a precificação de opções e para compreender a estrutura da volatilidade nos mercados financeiros. Este trabalho analisou diferentes abordagens para a modelização da IVS, desde modelos clássicos até técnicas de *Machine Learning*, incluindo modelos híbridos que combinam ambos os paradigmas.

Foram utilizados dados diários de opções SPX disponibilizados pela CBOE, abrangendo o período de 2009 a 2019. As variáveis explicativas principais foram o *log-moneyness* e o tempo até à maturidade, complementadas, nos modelos de *Machine Learning*, pela taxa de juro livre de risco e por um indicador binário de tipo de opção, de forma a avaliar o contributo dessas variáveis adicionais.

Todos os modelos seguiram o mesmo processo de treino, afinação e teste, garantindo comparabilidade. O desempenho foi avaliado através de métricas de erro como o R^2 e o RMSE, complementadas por análises de resíduos. Os modelos clássicos, nomeadamente a regressão linear (OLS) e o modelo polinomial de quinto grau, evidenciaram a natureza não linear da IVS, enquanto o modelo de *Splines B-Cúbicos* trouxe ganhos significativos de flexibilidade, ainda que mantendo alguma estrutura residual por explicar.

As técnicas de *Machine Learning* (*Random Forest* e *XGBoost*) apresentaram melhorias substanciais nas métricas e eliminaram praticamente toda a estrutura dos resíduos, demonstrando uma capacidade superior de adaptação a padrões complexos. Por sua vez, os modelos híbridos, baseados na aplicação de *Machine Learning* sobre os resíduos do modelo de *Splines*, alcançaram resultados equivalentes ou ligeiramente superiores, preservando uma base interpretável.

Este resultado confirma que a integração entre modelos clássicos e *Machine Learning* oferece um equilíbrio eficaz entre estrutura e flexibilidade, constituindo um quadro robusto para a modelização da IVS.

Palavras-chave: Superfície de Volatilidade Implícita, Técnicas de *Machine Learning*, Modelos Híbridos, Opções SPX

Classificação JEL: C45, C53

Table of Contents

Acknowledgements.....	I
Abstract.....	III
Resumo	V
Table of Contents	VII
List of Tables.....	IX
List of Figures	XI
Chapter 1 Introduction	1
Chapter 2 Literature Review	5
Chapter 3 Data and Sample Preparation	9
3.1 Market and Data Overview	9
3.2 Variable Construction and Features	10
3.3 Construction of the Modelling Dataset	12
3.3.1 Liquidity and Maturity Filters	12
3.3.2 Risk-Free Rate Alignment.....	13
3.3.3 Theoretical IV and NaN Removal.....	14
3.3.4 Final Sample Summary	16
3.4 Descriptive Statistics of the Final Sample	17
3.5 Sample Splitting and Embargo Design	20
Chapter 4 Methodology	21
4.1 Overview of the Modelling Framework	21
4.2 Classical Models	21
4.2.1 Ordinary Least Squares (OLS).....	22
4.2.2 Polynomial Regression.....	23
4.2.3 Cubic B-Splines	24
4.3 Machine Learning Models	26
4.3.1 Random Forest	26

4.3.2	Extreme Gradient Boosting.....	28
4.4	Hybrid Models	30
4.5	Evaluation Framework and Performance Metrics	31
Chapter 5 Results and Discussion.....		33
5.1	Overview of the Evaluation Framework.....	33
5.2	Results from Classical Models.....	33
5.2.1	Ordinary Least Squares	33
5.2.2	Polynomial Regression.....	35
5.2.3	Cubic B-Splines	37
5.3	Results of Machine Learning Models	39
5.3.1	Random Forest	39
5.3.2	XGBoost.....	42
5.4	Results of Hybrid Models	45
5.4.1	Hybrid Random Forest.....	45
5.4.2	Hybrid XGBoost	48
5.5	Comparative Analysis and Key Insights	51
Chapter 6 Conclusions and Recommendations.....		53
References.....		- 57 -
Appendices.....		- 60 -

List of Tables

Table 3.1 - Description of the raw variables in the SPX option dataset.	10
Table 3.2 - Impact of Liquidity and Maturity Filters on Sample Size.	13
Table 3.3 - Observations Retained After Risk-Free Rate Alignment.	14
Table 3.4 - Observations Retained After Theoretical IV Quality Filter.	16
Table 3.5 - Evolution of Dataset Size Across Filtering Steps.	16
Table 3.6 - Summary statistics of theoretical implied volatility (final modelling sample).	17
Table 3.7 - Summary of Data Splitting and Embargo Design.	20
Table 4.1 - Hyperparameters of the Random Forest model.	27
Table 4.2 - Main hyperparameters of the XGBoost model and their functional roles.	29
Table 4.3 - Evaluation metrics and their interpretation.	32
Table 5.1 - OLS Coefficients, Standardized Betas and Relative Weights.	33
Table 5.2 - Performance metrics of the OLS model.	34
Table 5.3 - Tuning results for polynomial regression models with different polynomial degrees.	35
Table 5.4 - Top five polynomial terms and respective normalized coefficients ($d = 5$, $\alpha^* = 0.001$).	35
Table 5.5 - Performance metrics of the polynomial regression model.	36
Table 5.6 - Tuning results for cubic B-spline models with different knot configurations in moneyness (m) and maturity (T).	37
Table 5.7 – Top five Cubic B-Splines terms and respective normalized coefficients ($d = 3$, $\alpha^* = 0.01$).	37
Table 5.8 - Performance metrics of the Splines model.	38
Table 5.9 - Predictor relevance analysis for the Random Forest model.	39
Table 5.10 - Hyperparameter grid used for Random Forest tuning and the selected configuration based on 2-fold cross-validation (60 fits in total).	40
Table 5.11 - Performance metrics of the Random Forest model.	40
Table 5.12 - Variable importance measures for the Random Forest model.	41
Table 5.13 - Predictor relevance analysis for the XGBoost model.	42
Table 5.14 - XGBoost hyperparameter search ranges and selected values.	43
Table 5.15 - Performance metrics of the XGBoost model.	43
Table 5.16 - Variable importance measures for the XGBoost model.	44

Table 5.17 - Hyperparameter grid used for Hybrid Random Forest tuning and the selected configuration based on 2-fold cross-validation (96 fits in total).....	45
Table 5.18 - Performance metrics of the Hybrid Random Forest model.	46
Table 5.19 - Variable importance measures for the Hybrid Random Forest model.....	47
Table 5.20 – Hybrid XGBoost hyperparameter search ranges and selected values.....	48
Table 5.21 - Performance metrics of the Hybrid XGBoost model.	49
Table 5.22 - Variable importance measures for the Hybrid XGBoost model.	50
Table 5.23 - Comparative performance of all models on the test sample.....	51

List of Figures

Figure 3.1 - Monthly Number of SPX Option Records (2009–2019).	9
Figure 3.2 - Distribution of option time to maturity (T) in the raw SPX dataset.....	11
Figure 3.3 - Percentage of Missing Implied Volatilities by Year.	14
Figure 3.4 - Density of log-moneyness for valid and NaN implied volatilities.....	15
Figure 3.5 - Percentage of Outliers ($IV > 2$) by Year.	15
Figure 3.6 - Implied Volatility Surface (3D, Median %).	18
Figure 3.7 - Total Variance Surface (3D, Median).....	19
Figure 5.1 - Predicted implied volatility surface – OLS model (test sample).	34
Figure 5.2 - Predicted implied volatility surface – Polynomial Regression (test sample).....	36
Figure 5.3 - Predicted implied volatility surface – Splines Model (test sample).....	38
Figure 5.4 - Predicted implied volatility surface – Random Forest Model (test sample).....	41
Figure 5.5 - Predicted implied volatility surface – XGBoost Model (test sample).	44
Figure 5.6 - Predicted implied volatility surface – Hybrid Random Forest Model (test sample).	47
Figure 5.7 - Predicted implied volatility surface – Hybrid XGBoost Model (test sample).....	50

Chapter 1

Introduction

In modern financial markets, options are particularly relevant due to their flexibility in constructing asymmetric payoff profiles and their effectiveness in hedging exposure to market fluctuations. Financial institutions, portfolio managers, and individual investors increasingly rely on options to implement tactical views, protect portfolios or exploit arbitrage opportunities. As a result of this increasing demand, the options market has experienced exponential growth since the establishment of the Chicago Board Options Exchange (CBOE) in 1973. According to data from the Options Clearing Corporation (OCC), the volume of equity option trading increased from 1.12 million contracts in 1973 to over 3.69 billion in 2017, and further to 6.52 billion contracts in 2024 (Li, 2021; *OCC*, 2024). Alongside their expanding role in financial markets, options are also used to extract forward-looking measures of uncertainty, as their prices reflect expectations about future asset movements. In this context, implied volatility (IV) stands out as a central concept, as a key variable for estimating the level of uncertainty anticipated by the market over the life of the option (Liu et al., 2021).

The concept of implied volatility emerged as a practical outcome of the Black–Scholes–Merton (1973) option pricing model. In the years following its publication, the model became widely used in financial markets to determine fair values for options and, since volatility was the only input that could not be directly observed in the market, practitioners began inferring it from actual option prices, reversing the model. One of the earliest documented instances of this approach can be found in Latané and Rendleman (1976), who extracted implied standard deviations from option prices. This inversion process led to the widespread use of implied volatility as a proxy for the market’s expectations.

However, 1987 stock market crash marked a turning point in the equity options market. In the years preceding the event, the existence of volatility smiles had already been documented, but implied volatility curves for index options remained relatively flat, so the phenomenon was not considered particularly relevant. After the crash, markets adjusted, and it became increasingly clear that the Black–Scholes assumption of constant volatility could not hold in practice. Subsequent empirical studies, such as Dumas et al. (1998) and Peña et al. (1999), documented systematic deviations from constant volatility across strikes and maturities in different option markets. By examining implied volatility as a dynamic function over strike and maturity, these and other empirical findings paved the way for the development of the implied volatility surface (IVS) concept. This marked a paradigm shift where the question was no

longer whether volatility was constant, but rather whether the IVS could be understood as a structured and predictable surface.

Early models of the IVS were based on empirical techniques such as polynomial regressions or spline interpolation. These approaches aimed to smooth observed market data but lacked theoretical foundations. As a result, they often produced unrealistic shapes, failed to ensure arbitrage-free pricing and struggled to capture the dynamic structure of the implied volatility surface, particularly in regions with sparse data (Dumas et al., 1996).

To overcome the limitations of empirical models, a new class of stochastic volatility models emerged in the 1990s and early 2000s. These models describe volatility itself as a random process, allowing the implied surface to evolve dynamically over time. A well-known example is the Heston model (1993), where volatility follows its own stochastic equation. Although more theoretically grounded, these models still do not fully solve the problem of modeling the IVS. Even with the significant progress made by structural models, studies such as Fan et al. (2022) show that classical approaches, including widely used models such as Heston and SABR, still fail to accurately capture the shape of the IVS across different market regimes. These findings reinforce the need to explore more flexible, data-driven models.

The idea of treating the IVS as a structured, low-dimensional object, whose evolution can be directly analyzed and modeled, was studied in the work of Cont and Da Fonseca (2002). Using principal component analysis (PCA) on a time series of implied volatility surfaces, they showed that most of the variation in implied volatilities can be explained by just three main components: level, slope, and curvature. This decomposition revealed that, despite its apparent complexity, the IVS evolves in a stable and predictable manner over time. Their findings laid the groundwork for factor-based and data-driven approaches to volatility modeling, suggesting that statistical and computational tools could effectively capture both the structure and dynamics of the IV surface.

Building on this perspective, machine learning (ML) has emerged as a powerful and flexible alternative for modeling the implied volatility surface. Unlike traditional models, ML approaches do not rely on rigid functional forms and can capture complex, nonlinear relationships directly from the data. Methods such as Random Forests, XGBoost and Neural Networks can efficiently handle high-dimensional inputs, model intricate interactions and generalize well across different market conditions. This makes them particularly suitable for modeling the IVS, where subtle patterns and regime shifts are common. Several recent studies support the growing use of ML in this context: van Lent (2023) work shows that ML models outperform classical benchmarks in capturing structural features of the IVS; Liu (2024) report

superior performance of Random Forests over linear models in American option pricing; and Maingo et al. (2025) successfully apply hybrid XGBoost models to forecast realized volatility. Furthermore, Dufays et al. (2025) propose an integrated framework for real-time IVS modeling using hybrid ML approaches, highlighting their practical relevance in both academic and industry settings. Together, these contributions reinforce the growing recognition of machine learning as a flexible and powerful tool for volatility modeling.

Motivated by this evidence, this dissertation aims to explore the modeling of the implied volatility surface using machine learning techniques, by evaluating their performance against classical benchmarks in terms of accuracy, generalization, and structural consistency. To achieve this objective, the analysis is structured into three main stages. First, a set of classical regression-based models will be developed to establish a reference benchmark. Linear and polynomial regressions will serve as an initial exploratory step, providing a simple yet informative baseline, while the spline-based model will introduce greater flexibility within the classical framework and allow for smoother variations across strikes and maturities. In the second stage, modern machine learning techniques will be applied for the direct modeling of the implied volatility surface, focusing on Random Forests and XGBoost, two algorithms widely used in financial modeling for their robustness, interpretability, and strong empirical performance. Finally, to assess the versatility of these methods, a hybrid framework will be proposed in which machine learning is employed not to model the IVS directly, but to learn from the residuals of the spline model. This final stage will test whether ML can effectively extract the remaining structure unexplained by classical models, demonstrating its potential to complement traditional approaches and enhance overall modeling accuracy. In addition to the empirical investigation, this dissertation develops a reproducible and modular framework for modeling the implied volatility surface, characterized by structured data handling, consistent validation procedures, and transparent model implementation.

The remainder of this dissertation is organized as follows. Chapter 2 reviews the main contributions in the literature on implied volatility and volatility surface modeling, with particular emphasis on machine learning approaches, especially Random Forest and XGBoost. Chapter 3 describes the data sources, variable construction, and preprocessing steps used to build the modeling dataset. Chapter 4 outlines the methodological framework, detailing the classical, machine learning, and hybrid models, as well as the evaluation metrics and validation procedures. Chapter 5 presents and discusses empirical results, comparing the performance of the different models and highlighting the main findings. Finally, Chapter 6 concludes the dissertation by summarizing the key results and providing recommendations for future research.

Chapter 2

Literature Review

The introductory chapter outlined the background of implied volatility modelling, from early empirical observations to the structured interpretation proposed by Cont and Da Fonseca (2002), and the subsequent methodological developments that have addressed it. That discussion showed how the implied volatility surface (IVS) evolved from an empirical irregularity to a systematic, measurable pattern over time. Building on that foundation, this chapter reviews key methodological contributions in the modelling of the IVS, organized into three main directions: classical econometric models, machine learning methods, and hybrid frameworks that integrate both.

Ordinary Least Squares (OLS) regression provides a natural starting point for modelling the IVS. In the standard econometric formulation (Greene, 2018), OLS provides a transparent and interpretable framework for estimating the relationship between implied volatility, log-moneyness, and time to maturity. However, due to its simplicity, this linear specification fails to capture the curvature and asymmetries typically observed in option markets, leading to systematic biases across moneyness levels and maturities.

A natural extension to address the inherent nonlinear patterns is the use of polynomial regressions models. Dumas et al. (1998) introduced a quadratic specification of implied volatility as a function of strike and maturity — the deterministic volatility function (DVF) — showing that such models can reproduce the smile and skew patterns typically observed in option markets. However, their empirical results revealed poor temporal stability, with parameters varying considerably across periods and hedging performance deteriorating out of sample. This instability motivated subsequent studies to explore regularized polynomial regressions (Hastie et al., 2009), where ridge penalization helps control multicollinearity and reduce variance among highly correlated regressors such as strike and maturity transformations.

An alternative to global polynomial regressions is spline-based methods, which impose local smoothness while avoiding a rigid global functional form. Fengler (2005) formalized this approach using tensor-product B-splines to jointly model the surface across the strike and maturity dimensions. The method allows the curvature of the IVS to vary locally while maintaining differentiability and smooth transitions across the strike–maturity domain. In subsequent work, Fengler et al. (2009) extended this framework into a semiparametric factor model, showing that a small number of smooth latent factors can explain most of the temporal

variation in the surface. Empirical evidence from these studies indicates that spline specifications achieve higher goodness of fit and greater temporal stability than polynomial models, making them a strong benchmark for comparison with the hybrid and machine learning approaches discussed later in this dissertation.

Machine learning methods, particularly ensemble tree-based algorithms such as Random Forest and XGBoost, have gained prominence in financial applications. These models combine predictive flexibility with robustness to high-dimensional and noisy data, providing an effective alternative to classical econometric approaches for modelling the IVS.

Among these, Random Forests (Breiman, 2001), represent one of the most influential ensemble learning methods for regression and classification problems. Building on earlier concepts of bootstrap aggregation (Breiman, 1996) and the random subspace method (Ho, 1998), the algorithm constructs an ensemble of decision trees trained on random subsets of both observations and features. The final prediction results from averaging across trees, reducing variance and mitigating overfitting while capturing nonlinear and interaction effects beyond the reach of linear models.

The main hyperparameters, such as the number of trees, maximum depth, and minimum samples per split, govern the balance between bias and variance and determine model stability. As discussed by Probst et al. (2019), increasing the number of trees generally enhances predictive robustness, while controlling tree depth prevents excessive noise fitting. Randomized feature selection at each split reduces inter-tree correlation, improving ensemble performance relative to single-tree models. The importance of efficient hyperparameter optimization has also been highlighted by Bergstra and Bengio (2012), who showed that random search can outperform exhaustive grid search in high-dimensional parameter spaces.

In financial applications, Random Forests have shown solid predictive accuracy and robustness. Luong et al. (2018) applied the method to forecast realized volatility, obtaining lower forecast errors compared with traditional econometric benchmarks. Olsen et al. (2024) extended this analysis to currency option markets, finding that Random Forests delivered competitive results, although not consistently outperforming deep learning architectures such as Long Short-Term Memory (LSTM) networks. Overall, these studies confirm the method's effectiveness for modelling volatility-related phenomena and motivate its application to the implied volatility surface, where complex nonlinear relationships between strike, maturity, and market conditions must be captured effectively.

Gradient boosting methods extend the ensemble logic of Random Forests by sequentially combining multiple weak learners, each correcting the residual errors from the

previous iteration. This iterative structure allows the model to minimize a chosen loss function through successive refinements. The theoretical foundation of this approach was formalized by Friedman (2001), who introduced the Gradient Boosting Machine as a flexible framework for optimizing additive models. Unlike Random Forests, which focus on variance reduction through randomization and averaging, gradient boosting primarily targets bias reduction by constructing trees that progressively correct the mistakes of earlier ones.

XGBoost, introduced by Chen and Guestrin (2016), represents one of the most efficient and scalable implementations of gradient boosting. The algorithm includes several key innovations: second-order gradient optimization for faster convergence, L1 and L2 regularization to prevent overfitting, and parallel processing to enhance computational performance. It also supports advanced features such as early stopping, tree pruning, and automatic handling of missing values, making it particularly suitable for large-scale financial datasets. The main hyperparameters — learning rate, maximum depth, number of boosting rounds, and regularization terms — control the trade-off between model complexity and generalization. In practice, smaller learning rates combined with a larger number of boosting rounds often yield more stable generalization performance.

Building on these algorithmic advances, empirical studies have increasingly adopted XGBoost for financial prediction tasks. The method has been successfully applied to asset return forecasting, volatility estimation, and option pricing. Its ability to capture nonlinear dependencies and interactions has made it a benchmark algorithm in modern empirical finance. Studies using XGBoost to predict implied or realized volatility generally report improvements in predictive accuracy compared with both classical econometric models and simpler machine learning approaches. For instance, Teller et al. (2022) show that XGBoost significantly outperforms both Heterogeneous Autoregressive (HAR) and LSTM models in short to long-term realized volatility forecasting. These characteristics make it a strong candidate for modeling the IVS, where predictive precision and computational efficiency are equally critical.

Finally, hybrid frameworks combine the structural interpretability of classical econometric models with the flexibility of machine learning techniques. The core idea is to use a statistical model to capture the main functional relationships in the data and then apply a machine learning algorithm to model the residual component. This two-step strategy allows the model to retain theoretical consistency while correcting for nonlinearities and omitted-variable effects that are difficult to capture parametrically.

Such approaches have been widely explored in econometrics and forecasting. Zhang (2003) demonstrated that hybrid models combining autoregressive structures with neural

networks could significantly improve forecasting accuracy. Kristjanpoller and Minutolo (2018) further advanced this concept by integrating GARCH, neural networks, and technical indicators into a hybrid volatility forecasting framework, showing clear accuracy gains over standalone models. These findings suggest that hybrid models can provide a balanced compromise between interpretability and flexibility, making them particularly relevant for IVS modelling, where preserving structure while improving predictive accuracy is essential.

The literature on IVS modelling shows a clear progression from classical econometric methods to modern machine learning and hybrid frameworks. Classical models such as linear, polynomial, and spline regressions established the foundations of structure and interpretability. Machine learning approaches, particularly ensemble tree-based algorithms, introduced flexibility and improved predictive performance by capturing nonlinear and high-dimensional relationships. Hybrid frameworks then emerged as a natural extension, combining statistical structure with machine learning applied to residuals to achieve a balance between theoretical consistency and accuracy. Together, these developments provide the conceptual foundation for the modelling strategy developed in this dissertation.

Chapter 3

Data and Sample Preparation

3.1 Market and Data Overview

The empirical analysis in this dissertation is based on standardized options written on the S&P 500 Index (SPX), traded on the Chicago Board Options Exchange (CBOE). The CBOE is the largest and most liquid options exchange in the United States, providing reliable and detailed quote data for a wide variety of contracts. SPX options are European-style, cash-settled, and quoted in index points, making them particularly suitable for implied volatility surface modelling due to their high trading volume and consistent market depth across a broad range of strikes and maturities.

The dataset used in this study includes daily data from January 2, 2009, to April 30, 2019, covering over a decade of trading activity and diverse market conditions, including the post-2008 recovery, periods of quantitative easing, and shifts in volatility and interest-rate regimes. Each observation corresponds to a single SPX option contract on a given trading day. The dataset contains both end-of-day and intraday quote information, with the latter captured at 15:45 (Eastern Time). Following CBOE's guidelines for quantitative research, this study relies on the 15:45 snapshot, as it better reflects real-time market liquidity and mitigates potential distortions from illiquid closing trades. The dataset includes the quote date, expiration date, option type (call or put), strike price, bid and ask prices for both the option and the underlying, as well as bid/ask sizes.

The dataset is organized at a daily frequency, with each row representing a specific option contract, uniquely defined by its type, strike, and maturity, on a given trading date. In total, the raw dataset contains approximately 15.1 million option-day observations, equally split between call and put contracts. Figure 3.1 shows the evolution of the number of available options, before any filtering, highlighting the significant growth in traded volume throughout the decade, driven by increased market participation and availability of strikes.

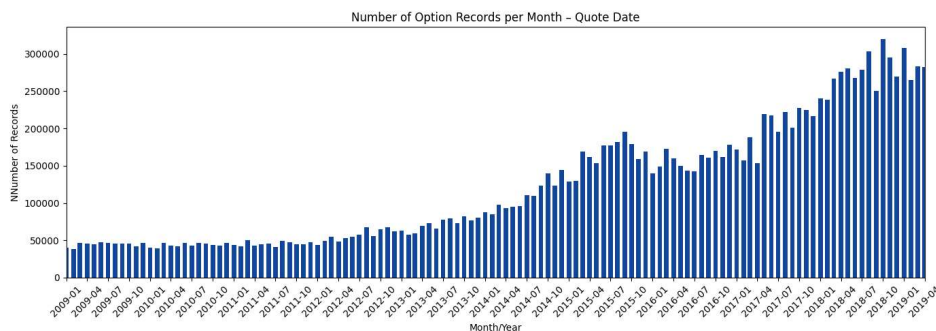


Figure 3.1 - Monthly Number of SPX Option Records (2009–2019).

3.2 Variable Construction and Features

The modelling process relies on both raw variables extracted from the dataset and additional features obtained through analytical transformations and external data integration. Table 3.1 summarizes the raw variables, whereas the constructed ones are presented later in this section.

Table 3.1 - Description of the raw variables in the SPX option dataset.

Variable	Description
quote_date	Date of observation
expiration	Contract expiration date
strike	Option strike price
option_type	Option type (Call or Put); Transform into binary form (type)
bid_1545, ask_1545	Bid/ask prices of the option at 15:45
underlying_bid_1545, underlying_ask_1545	Bid/ask of the underlying asset at 15:45
bid_size_1545, ask_size_1545	Bid/ask sizes at 15:45

The main constructed variables are the time to maturity, the mid-prices of the option and the underlying, and the log-moneyness. The time to maturity, T , is measured as the number of calendar days between the expiration and quote dates, scaled using the ACT/365 convention:

$$T = \frac{\text{expiration date} - \text{quote date}}{365} \quad (3.1)$$

In the case of the option price and the underlying price, both are computed as the midpoint between their respective bid and ask quotes at 15:45. Using the mid-quote provides a consistent estimate of the fair transaction value, reducing bid–ask bounce effects and avoiding the directional bias inherent in relying solely on either bid or ask quotes. This approach is in line with standard practices in empirical option pricing research (e.g., Dumas et al. (1996)) and reflects the methodology recommended by the CBOE for quantitative analysis. Taking this into consideration, the option’s mid-price is calculated as follows:

$$C_{\text{mid}} = \frac{\text{bid}_{1545} + \text{ask}_{1545}}{2} \quad (3.2)$$

The price of the underlying is computed in the same way:

$$S = \frac{\text{underlying_bid}_{1545} + \text{underlying_ask}_{1545}}{2} \quad (3.3)$$

The log-moneyness variable, m , one of the key explanatory factors for the implied volatility surface, is defined as:

$$m = \ln\left(\frac{K}{S}\right), \quad (3.4)$$

where K denotes the strike price and S the underlying's spot price.

Additionally, a risk-free interest rate variable was incorporated into the dataset using daily U.S. Treasury yields obtained from the Federal Reserve Economic Data (FRED). The selected maturities correspond to 1-month, 3-month, 6-month, and 1-year tenors, which together represent the short end of the yield curve most relevant for the options in the sample. This choice follows from the empirical maturity distribution presented in Figure 3.2, which shows that most SPX options have maturities shorter than one year.

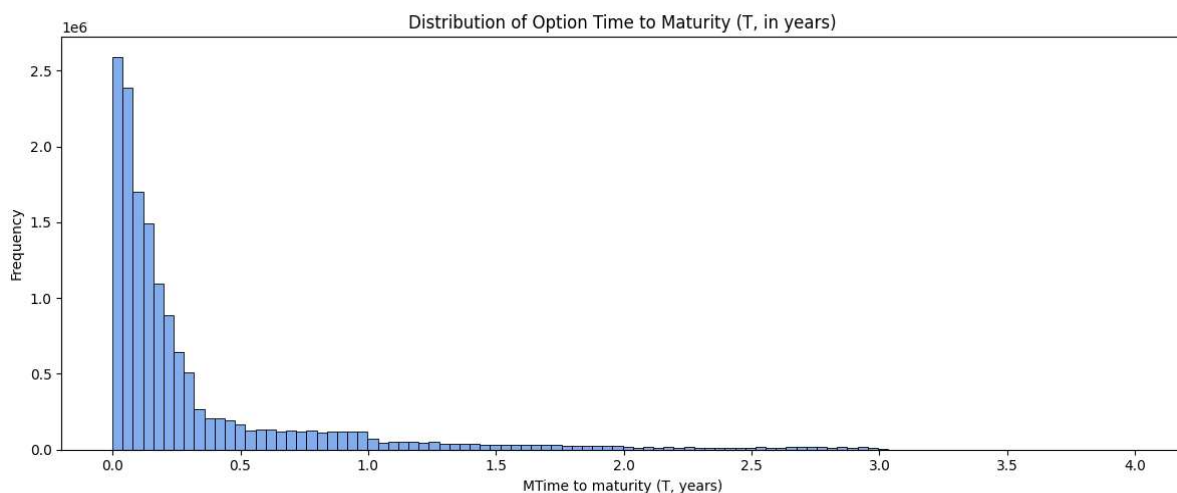


Figure 3.2 - Distribution of option time to maturity (T) in the raw SPX dataset. The figure shows the strong concentration of short-term contracts, with most options expiring within a few months.

The inclusion of the risk-free rate serves two main purposes. First, it is a required input for performing the implied volatility inversion under the Black-Scholes framework, which relies on the option's market price, strike, time to maturity, and risk-free rate to determine the corresponding implied volatility. Second, it was considered as a potential explanatory variable in the machine learning models developed later in this dissertation, to examine whether variations in the risk-free rate could help explain the dynamics of the implied volatility surface.

For each option, a risk-free rate was assigned based on its time to maturity, T . Options with maturities shorter than one month were assigned the 1-month rate (DGS1MO), while those

with one-year maturity were assigned the 1-year rate (DGS1). For intermediate maturities, the applicable rate was obtained through linear interpolation between the nearest available tenors. For instance, maturities between 1M and 3M used interpolation between DGS1MO and DGS3MO; those between 3M and 6M used interpolation between DGS3MO and DGS6MO; and those between 6M and 1Y used interpolation between DGS6MO and DGS1. This procedure yields a maturity-adjusted risk-free rate consistent with each option's time to expiration and is described in greater detail in the following section, which outlines the construction of the final dataset.

3.3 Construction of the Modelling Dataset

The final modelling dataset was obtained after a series of data preparation steps designed to ensure the quality, consistency and relevance of the observations used for model training and evaluation. These steps included applying liquidity and maturity filters, aligning each option with its corresponding risk-free rate, and computing the theoretical implied volatility by inverting the Black-Scholes model. Each of these components is described in the following subsections.

3.3.1 Liquidity and Maturity Filters

The first step in constructing the modelling dataset consisted of retaining only observations that were both liquid and representative of maturities relevant to the implied volatility surface.

To ensure adequate liquidity, only option quotes with strictly positive bid and ask sizes were retained, that is, options satisfying $bid_size_{1545} > 0$ and $ask_size_{1545} > 0$. This filter removes stale or non-tradable quotes, which often come from strikes or maturities with little or no trading activity. Keeping only quotes with active bid and ask sizes helps ensure that mid-prices and implied volatilities reflect genuine market conditions rather than outdated or indicative quotes.

Regarding maturity time, the sample was limited to options with maturities between 7 and 365 calendar days ($7 \leq T_days \leq 365$). Very short-term options (less than a week to expiration) often exhibit unstable, noisy implied volatilities, while very long-term options, commonly referred to as LEAPS (Long-term Equity AnticiPation Securities), tend to be less liquid and less relevant for modeling the short-to medium-term structure of the implied volatility surface.

Table 3.2 reports the number of observations retained after each filtering step. The liquidity filter eliminated 1.33 million observations (−8.75%), while the maturity filter removed an additional 2.06 million (−14.87%). In total, 3.38 million records (−22.32%) were excluded, leaving approximately 11.76 million option-day observations for subsequent filtering and analysis.

Table 3.2 - Impact of Liquidity and Maturity Filters on Sample Size.

Step	Remaining Observations	Δ (absolute)	Δ (%)
Original sample	15,144,974	—	—
After liquidity filter	13,819,247	−1,325,727	−8.75%
After maturity filter	11,763,998	−2,055,249	−14.87%
Total reduction	—	−3,380,976	−22.32%

3.3.2 Risk-Free Rate Alignment

The next step in constructing the modelling dataset involved ensuring that every option observation was associated with a valid risk-free rate. As described in Section 3.2, each option was matched to a corresponding U.S. Treasury yield based on its time to maturity using short-term tenors obtained from the FRED database. This alignment guarantees consistency between option and interest rate data, which is essential for the correct computation of theoretical implied volatilities.

Since not all U.S. federal holidays correspond to full market closures for Treasury securities, there are days officially marked as holidays for the equity markets during which the Treasury market remained open and official yields were published by the Federal Reserve. Consequently, excluding all federal holidays would incorrectly remove valid observations from the dataset. To address this, all dates with available Treasury yields in the FRED series were retained, even if they coincided with stock market holidays, whereas dates without published yields, corresponding to full Treasury market closures, were removed. The excluded days were manually verified to confirm their origin. All missing values corresponded to well-documented events in which the U.S. financial markets were fully closed, such as Good Friday, national days of mourning, or exceptional events, like Hurricane Sandy.

Table 3.3 summarizes the effect of this step. Only 84,148 observations (−0.72%) were removed, confirming that the alignment process preserved virtually the entire dataset while ensuring that each remaining record had a valid and consistent risk-free rate.

Table 3.3 - Observations Retained After Risk-Free Rate Alignment.

Step	Number of Observations	Δ (absolute)	Δ (%)
After liquidity and maturity filters	11,763,998	—	—
After risk-free rate alignment	11,679,850	-84,148	-0.72%

3.3.3 Theoretical IV and NaN Removal

The original dataset did not provide theoretical implied volatility values; therefore, it was necessary to compute the IV for all contracts to serve as the target variable in the modelling process. Implied volatilities were obtained by numerically inverting the Black–Scholes formula using the `implied_volatility` function from the open-source *py_vollib* Python library. This solver estimates the volatility parameter that equates the theoretical Black–Scholes price to the observed market mid-price, based on each option’s mid-price, strike, time to maturity, underlying mid-price, risk-free rate, and option type (call or put). The *py_vollib* implementation relies on the Newton–Raphson iterative method, which is computationally efficient but can be numerically unstable for deep in-the-money (ITM) and deep out-of-the-money (OTM) options.

After the inversion, descriptive statistics were computed to assess the quality of the resulting implied volatilities. Out of 11.68 million observations, approximately 2.82 million (24.13%) failed to converge and returned NaN values, while no negative volatilities were detected. The proportion of missing values remained relatively stable over time, ranging from about 18% in 2018 to 29% in 2013, with most years falling between 22% and 26%. No clear upward or downward trend was observed over the sample period (Figure 3.3).

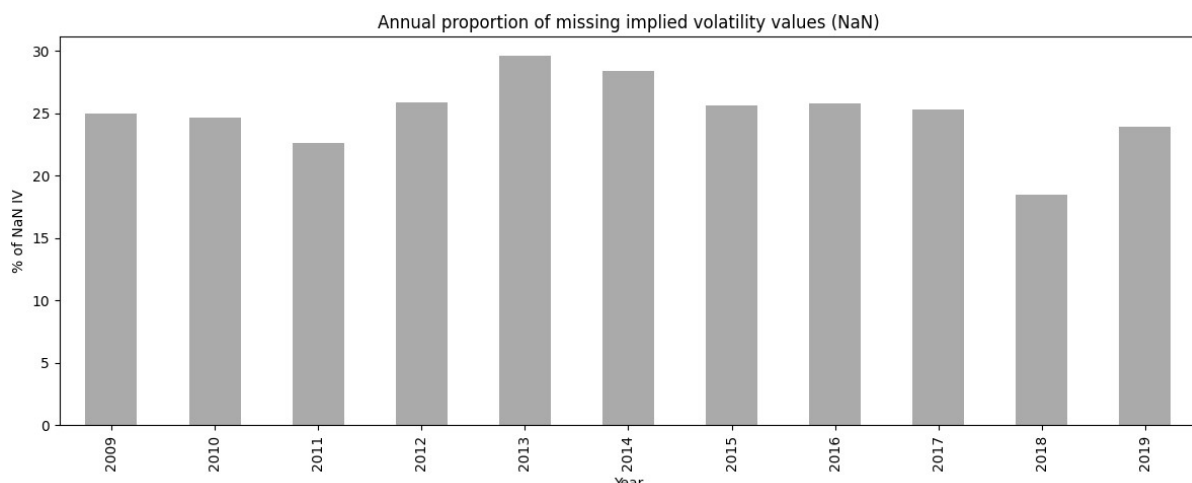


Figure 3.3 - Percentage of Missing Implied Volatilities by Year.

The proportion fluctuates between 18% and 29%, showing no persistent trend.

When comparing the distribution of log-moneyness for options with valid and missing implied volatilities (Figure 3.4), we see that valid observations are concentrated around the at-the-money region ($m \approx 0$), whereas missing values are predominantly associated with deep out-of-the-money puts and deep in-the-money calls. The average log-moneyness among missing cases (-0.3873) contrasts with -0.0856 for valid ones, confirming that numerical convergence failures primarily occur in the tails of the moneyness distribution.

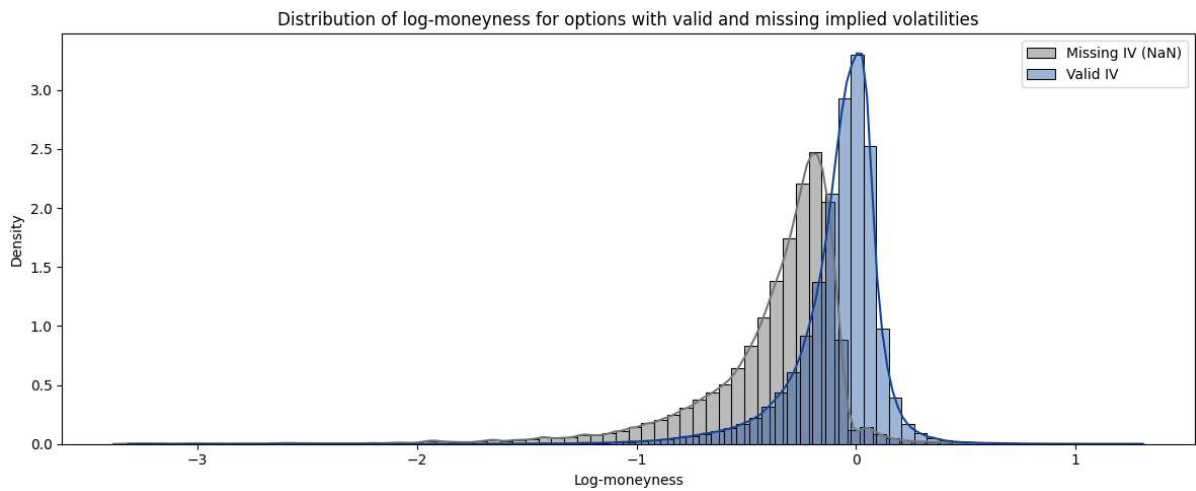


Figure 3.4 - Density of log-moneyness for valid and NaN implied volatilities. Most valid IVs cluster near the at-the-money region ($m \approx 0$), whereas missing values concentrate in deep OTM puts and ITM calls.

Additionally, a small number of extreme outliers were identified above 2 in volatility units, representing only 978 observations (0.0084% of the total sample) and with no concentration in any specific period (Figure 3.5).

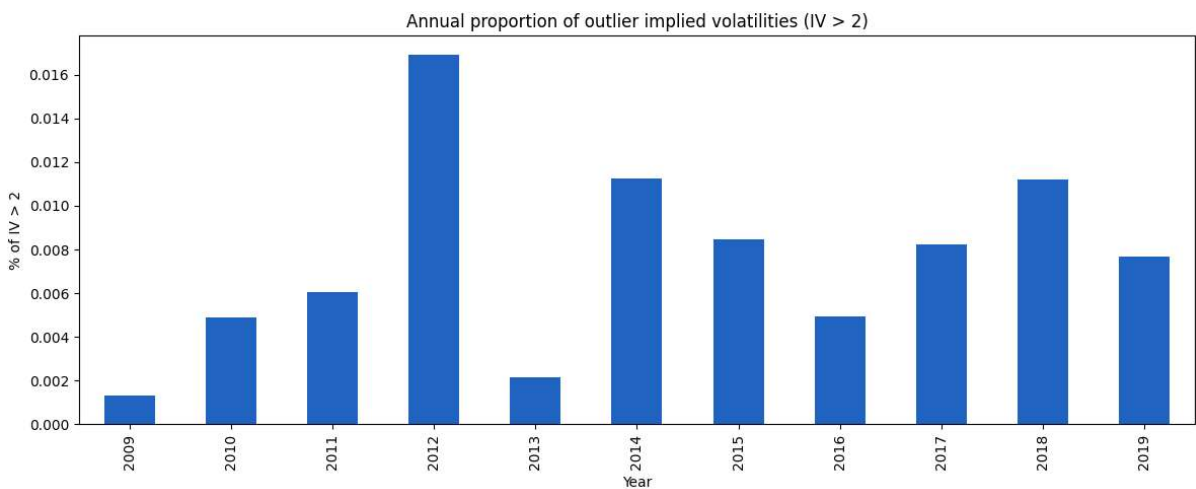


Figure 3.5 - Percentage of Outliers ($IV > 2$) by Year.

As these missing values stem from inherent numerical limitations rather than data errors, no imputation was attempted. All observations with missing, null, or unrealistic implied volatilities ($IV > 2$) were removed from the modelling dataset. This step excluded approximately 2,82 million records (-24%), primarily affecting deep OTM or ITM options (Table 3.4). The resulting dataset therefore preserves high quality and representativeness for the subsequent modelling phase.

Table 3.4 - Observations Retained After Theoretical IV Quality Filter.

Step	Number of Observations	Δ (absolute)	Δ (%)
After previous filters	11,679,850	—	—
After theoretical IV quality filter	8,860,820	-2,819,030	-24.14%

3.3.4 Final Sample Summary

The final dataset contains 8,860,820 observations. Table 3.5 summarizes the progressive reduction in sample size across the various cleaning and filtering stages. Starting from the original dataset of 15.14 million option-day observations, the liquidity and maturity filters removed roughly 22% of records, while the risk-free rate alignment eliminated less than 1%. The final quality filter, applied to the theoretical implied volatility, excluded approximately 2.82 million additional records (-24.14%). Overall, 6.28 million observations were removed, resulting in a 41.49% reduction relative to the original dataset.

Table 3.5 - Evolution of Dataset Size Across Filtering Steps.

Step	Number of Observations	Δ (absolute)	Δ (%)
Original dataset	15,144,974	—	—
Liquidity filter	13,819,247	-1,325,727	-8.75%
Maturity filter	11,763,998	-2,055,249	-14.87%
Risk-free rate alignment	11,679,850	-84,148	-0.72%
Theoretical IV quality filter	8,860,820	-2,819,030	-24.14%
Total reduction	—	-6,284,154	-41.49%

The proportion of calls and puts was nearly balanced before applying the quality filter on theoretical implied volatility (Calls 50.83%, Puts 49.17%). After this step, the final modelling sample comprised 35.94% calls (3,184,554 contracts) and 64.06% puts (5,676,266 contracts).

3.4 Descriptive Statistics of the Final Sample

Before proceeding with model estimation, it is essential to describe the statistical properties of the final dataset, particularly the behavior of the theoretical implied volatility after cleaning and filtering. Understanding its empirical distribution provides a valuable reference for interpreting the magnitude of the evaluation metrics presented later in the methodology and results chapters.

The summary statistics for the theoretical implied volatility are reported in Table 3.6. The variable displays a mean of 0.2186, a standard deviation of 0.1238, and interquartile values of 0.1321 (25th percentile), 0.1857 (median), and 0.2739 (75th percentile). The observed minimum is 0.0004 and the maximum 1.9987, as values above 2 were excluded to remove unrealistic outliers. These figures indicate that most implied volatilities in the sample lie between 10% and 30%, consistent with the historical range typically observed in SPX options.

Table 3.6 - Summary *statistics of theoretical implied volatility (final modelling sample)*.

Statistic	Value
Count	8,860,820
Mean	0.2186
Standard deviation	0.1238
Minimum	0.0004
25th percentile	0.1321
Median (50th percentile)	0.1857
75th percentile	0.2739
Maximum	1.9987

To visualize the structure of the implied volatility surface and confirm its expected empirical features, a two-dimensional grid was constructed using log-moneyness (m) and time to maturity (T) as the main dimensions. The grid, trimmed between the 1st and 99th percentiles and divided into approximately 50×40 bins, was aggregated by the median implied volatility in each cell. Two representations were analyzed: the implied volatility itself, expressed in percentage terms for interpretability, and the total variance ($w = \sigma^2 T$), which provides a smoother and temporally consistent view of the volatility structure.

Figure 3.6 shows the empirical IVS obtained from the final dataset. Implied volatility reaches its minimum near the at-the-money level ($m \approx 0$) and increases on both sides, forming a clear smile. This pattern reflects the sensitivity of implied volatility to strike variations around the ATM region and is consistent with the behavior typically observed in SPX options.

Looking more closely at the smile, the surface shows a clear asymmetry between the two wings. Implied volatility increases sharply for negative log-moneyness values ($m < 0$), corresponding to deep out-of-the-money (OTM) puts and in-the-money (ITM) calls. This left-hand steepness reflects the negative volatility skew typical of equity index options, driven by the demand for downside protection and the risk premium embedded in OTM puts. On the opposite side ($m > 0$), representing OTM calls and ITM puts, volatility rises more gradually, producing a flatter right wing. This asymmetry confirms the persistent left-skew pattern widely documented in SPX options.

The surface also changes with maturity. Short-term options display higher implied volatilities and stronger skewness, especially on the put side. As maturity increases, volatility levels decrease and the smile flattens, showing less variation across moneyness. This behavior follows the typical maturity-volatility relationship observed in index options.

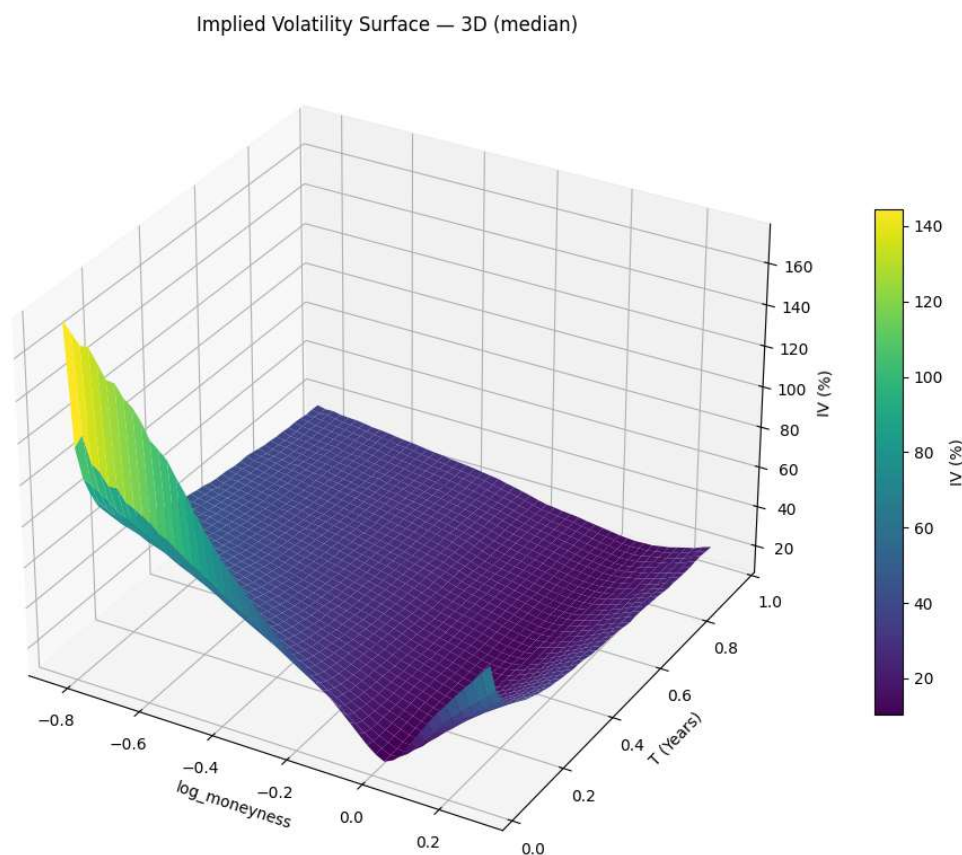


*Figure 3.6 - Implied Volatility Surface (3D, Median %).
The surface shows the typical left-skew pattern of SPX options*

Turning to total variance, Figure 3.7 presents the surface computed from the final dataset. For all moneyness levels, total variance increases with maturity, confirming the expected monotonic behavior that reflects the accumulation of variance over time. The growth of $w(T)$ is approximately linear, but the slope varies with moneyness: the left wing (OTM puts) grows faster with T , while the right wing (OTM calls) increases more gradually. The asymmetry across moneyness remains visible, with higher total variance on the put side and lower levels on the call side.

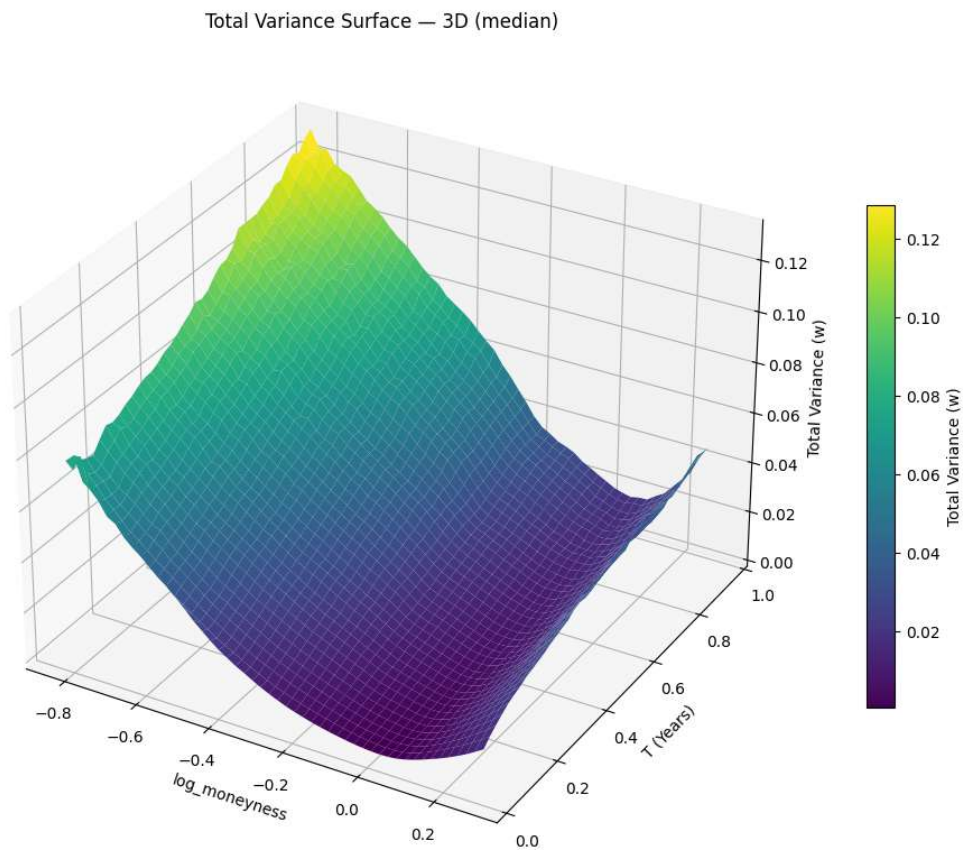


Figure 3.7 - Total Variance Surface (3D, Median).

Overall, the final sample exhibits a realistic and internally consistent implied volatility structure, characterized by a pronounced left-skew, higher short-term volatility, and monotonic growth of total variance with maturity. These properties confirm both the integrity of the data and the adequacy of the preprocessing steps, ensuring a reliable foundation for the modelling phase that follows.

3.5 Sample Splitting and Embargo Design

To ensure a realistic and robust evaluation of the models, the final dataset was split into training, tuning (validation), and test sets in a chronological order. Because the number of listed SPX options increases substantially over time, a split based on quote dates would have resulted in an unbalanced division, with the most recent years dominating the test set. To avoid this distortion, the separation was instead performed by the number of observations (rows), which preserves consistent statistical proportions across subsets despite variations in daily trading volume.

The dataset was first divided into training and test sets with approximately an 85/15 ratio. The training subset covers the period from January 2, 2009, to September 21, 2018, while the test subset begins on October 1, 2018, aligned with the start of the fourth quarter of 2018, and extends until April 30, 2019. Within the training portion, a secondary split of approximately 80/20 was applied to create the inner training and inner tuning samples. The inner training subset spans from January 2, 2009, to December 26, 2017, and the inner tuning subset from January 2, 2018, to September 21, 2018, ensuring that the tuning phase uses the most recent segment of the training data while preserving a broad historical base for model estimation.

Note that embargo windows were applied to these splits to prevent temporal leakage. Between the training and tuning samples, an internal embargo of three business days was used at the end of 2017, while between the complete training and test samples, an external embargo of five business days was applied before the start of October 2018. These embargo periods help preserve temporal independence across modelling stages and reduce the risk of information leakage between adjacent time segments. The resulting samples and their characteristics are summarized in Table 3.7.

Table 3.7 - Summary of Data Splitting and Embargo Design.

Sample	Start Date	End Date	Observations	Days	% of total	% of train
Full Training	2009-01-02	2018-09-21	7,505,238	2,427	84.70%	—
Test	2018-10-01	2019-04-30	1,292,514	142	14.59%	—
Inner Training	2009-01-02	2017-12-26	5,970,843	2,244	67.38%	79.56%
Inner Tuning	2018-01-02	2018-09-21	1,534,395	183	17.32%	20.44%
Embargo periods	—	—	—	—	0.71%	—

This configuration ensures a comprehensive historical base for model training (2009–2017), a recent segment for hyperparameter tuning (2018 Q1 - Q3), and a well-isolated test window (Q4-2018 to April 2019) for out-of-time evaluation.

Chapter 4

Methodology

4.1 Overview of the Modelling Framework

Having established the dataset and feature construction in the previous chapter, the next step is to define the methodological framework for modelling the implied volatility surface. The modelling design follows a progressive structure, starting from simple parametric formulations and moving towards increasingly flexible data-driven methods. Each model is developed under a unified empirical framework, with standard data partitions, evaluation metrics, and validation logic, ensuring that differences in performance reflect methodological characteristics rather than data inconsistencies. This structure provides a coherent basis for analyzing how varying levels of model complexity contribute to explaining the shape and dynamics of the IVS.

For clarity of interpretation, the approaches are organized into three groups. The classical models serve as interpretable baselines and include a simple linear regression model based on ordinary least squares method (hereafter denoted as the OLS model, following the notation used in Python libraries), a Polynomial Regression, and a Cubic B-Spline. The machine learning models, Random Forest (RF) and XGBoost (XGB), relax functional assumptions and learn nonlinear patterns directly from the data. Finally, the hybrid models combine both paradigms by fitting ML algorithms to the residuals of the spline model, allowing the ML component to capture remaining nonlinear structures unexplained by the classical specification.

All models share a consistent design to ensure comparability. The core features are the logarithmic moneyness (m) and time to maturity (T), while the ML and hybrid frameworks also include the risk-free rate (r) and an option type indicator ($type$) to test for additional discriminant power and nonlinear interactions. The evaluation framework, detailed later in this chapter, is applied consistently across all models to facilitate fair and transparent comparison.

4.2 Classical Models

The first group of models corresponds to classical regression-based approaches, which rely on explicit functional relationships to describe how implied volatility varies with moneyness and time to maturity. These models provide interpretable baselines and help illustrate the limitations of traditional formulations when applied to the highly nonlinear structure of the IVS.

Within this group, the Ordinary Least Squares (OLS) and Polynomial Regression models are fully parametric, assuming a fixed global functional form whose parameters are directly estimated from the data. These two models are intentionally simple and mainly illustrative, showing how rigid global structures fail to reproduce the complex curvature typically observed in the IVS. The Cubic B-Spline model extends this classical framework by introducing local flexibility and smoothness through spline basis functions. This semi-parametric approach provides a more adaptable representation of the IVS and serves as the benchmark against which the performance of the machine learning models is later assessed.

4.2.1 Ordinary Least Squares (OLS)

The Ordinary Least Squares (OLS) model represents the simplest specification within the classical family, following the standard linear regression econometric formulation (Greene, 2018). It assumes a global linear relationship between implied volatility and its explanatory variables, providing a transparent and interpretable baseline for comparison with more flexible modelling approaches. Formally, the model is expressed as:

$$IV_i^{OLS} = \beta_0 + \beta_1 m_i + \beta_2 T_i + \varepsilon_i, \quad (4.1)$$

where IV_i^{OLS} denotes the fitted implied volatility for observation i , $m_i = \log(K_i/S_i)$ represents the logarithmic moneyness, T_i is the time to maturity (in years), and ε_i is assumed to be Gaussian white noise, with zero mean and constant variance. The parameters β_0 , β_1 , and β_2 are obtained by minimizing the sum of squared residuals (that is, OLS):

$$\hat{\beta} = \arg \min_{\beta} \sum_i (IV_i - \mathbf{x}_i^T \beta)^2, \quad (4.2)$$

where $\mathbf{x}_i^T = [1, \log(K_i/S_i), T_i]$ denotes the vector of explanatory variables including the intercept term. The model is estimated using the *scikit-learn* implementation of linear regression, which computes the closed-form least-squares solution. Since this specification does not include any hyperparameters, no tuning or cross-validation was required prior to fitting. The resulting fitted values define the predicted implied volatility surface (IVS^{OLS}).

Although OLS is simple and easy to interpret, its linear functional form limits its ability to capture the nonlinear patterns commonly observed in IVS, such as smiles, skewness or local curvature. It is therefore used mainly as a baseline model, providing a reference for the more flexible approaches introduced next.

4.2.2 Polynomial Regression

Polynomial regression is a natural extension of the OLS framework, allowing for nonlinear relationships between implied volatility and its explanatory variables via polynomial basis expansion. It remains a parametric model but introduces additional flexibility by including higher-order and interaction terms between the main inputs of the surface. This formulation has been previously tested in the context of IVS modelling by Dumas et al. (1996), and is consistent with modern statistical treatments of polynomial and regularized regression (Hastie et al., 2009).

The general polynomial model of degree d is defined as:

$$IV_i^{Poly} = \beta_0 + \sum_{a+b \leq d} \beta_{ab} m_i^a T_i^b + \varepsilon_i, \quad (4.3)$$

where IV_i^{Poly} denotes the fitted implied volatility for observation i , m_i denotes the logarithmic moneyness, T_i is the time to maturity (in years), and ε_i is an error term. The inclusion of cross terms (a, b) allows the model to capture joint nonlinearities between moneyness and maturity.

As the number of polynomial terms grows rapidly with the degree d , the model becomes increasingly prone to overfitting. To address this issue, ridge regularization (L2 penalty) is applied during the estimation of the regression coefficients, by minimizing the following penalized objective function:

$$\hat{\beta} = \arg \min_{\beta} \left[\sum_i (IV_i - x_i^T \beta)^2 + \alpha \sum_j \beta_j^2 \right], \quad (4.4)$$

where $\alpha > 0$ controls the strength of regularization. The optimal value of α is determined through a ridge regression algorithm with cross-validation (RidgeCV), which evaluates multiple α values, logarithmically spaced between 10^{-6} and 10^3 , and automatically selects the one that minimizes the validation error. Higher values of α impose a stronger penalty on significant coefficients, shrinking them toward zero and producing smoother and more stable fitted surfaces. This regularization reduces variance at the cost of a small bias, improving the model's generalization to unseen data.

After selecting the optimal regularization strength α^* for each polynomial degree d , the procedure compares results across all tested degrees to identify the most appropriate specification. To ensure parsimony, a practical selection criterion was defined in which, among all polynomial degrees producing an RMSE within 0.5% of the minimum observed, the

simplest specification corresponding to the lowest degree was chosen. This rule balances model flexibility and interpretability, avoiding unnecessary complexity for marginal gains in fit quality.

The final polynomial regression model is re-estimated on the full training set using the optimal degree d^* and regularization parameter α^* , producing the approximated implied volatility surface IVS^{Poly} . This approach offers greater flexibility than OLS by capturing smooth curvature and moderate nonlinearity while remaining interpretable and computationally efficient, positioning it as an intermediate step between the linear OLS model and the more adaptive spline and machine learning approaches presented later.

4.2.3 Cubic B-Splines

Cubic B-splines provide a flexible and locally smooth representation of the implied volatility surface, allowing curvature to vary across both moneyness and maturity without imposing a rigid global functional form. Unlike the polynomial model, which applies a single equation over the entire domain, splines partition the feature space into smaller segments joined smoothly at predefined knots, producing a piecewise polynomial approximation that adapts to local patterns in the data. The capability of this semiparametric model to capture the structure of the IVS was illustrated by Fengler (2005) and Fengler et al. (2009).

The modelling procedure starts by defining the spline bases over logarithmic moneyness, m_i , and maturity, T_i , which serve as the building blocks of the IVS. Before generating the spline bases, both m_i and T_i are standardized (z-score) to enhance numerical stability and ensure consistent knot placement. The model uses cubic splines ($degree = 3$) with quantile-based knot placement, which allocates more flexibility to regions with denser observations. If quantile placement is not supported, uniform spacing is used as a fallback. Linear extrapolation is applied near the boundaries to maintain stable behavior outside the main data range.

Each dimension is then represented by a one-dimensional cubic B-spline basis:

$$B_m(m_i) = (B_{m,1}(m_i), \dots, B_{m,p}(m_i)), \quad (4.5)$$

$$B_T(T_i) = (B_{T,1}(T_i), \dots, B_{T,q}(T_i)). \quad (4.6)$$

The implied volatility at observation i is then modelled as a tensor-product of these bases, where θ_{rs} are the estimated coefficients and ε_i is a zero-mean error term:

$$IV_i^{Spline} = \sum_{r=1}^p \sum_{s=1}^q \theta_{rs} B_{m,r}(m_i) B_{T,s}(T_i) + \varepsilon_i \quad (4.7)$$

This formulation allows the model to capture smooth local variations in the surface while ensuring continuity and differentiability up to the second order. After constructing the tensor-product basis $z_i = B_m(m_i) \otimes B_T(T_i)$, the coefficients θ are estimated using ridge regularization to control overfitting. The optimization problem is given by:

$$\hat{\theta} = \arg \min_{\theta} \left[\sum_i (IV_i - z_i^T \theta)^2 + \alpha \sum_j \theta_j^2 \right], \quad (4.8)$$

where $\alpha > 0$ determines the strength of the L2 penalty. Higher values of α smooth the surface by shrinking the spline coefficients toward zero. The optimal value of α is selected automatically using a ridge regression algorithm with cross-validation (RidgeCV), testing values logarithmically spaced between 10^{-2} and 10^2 .

The training process follows the same nested structure described previously. For each combination of knots in the moneyness and maturity dimensions ($n_{knots,m}, n_{knots,T}$) and for each tested regularization level α , the model is first fitted on the Inner Training sample to estimate the coefficients θ . The resulting specification is then applied to the Inner Tuning sample, where its predictive accuracy is evaluated through the Root Mean Squared Error (RMSE). This separation ensures that parameter estimation and model selection are performed on independent data, preventing overfitting to the training set. To ensure parsimony, a practical selection criterion was adopted: among all configurations yielding an RMSE within 0.5% of the minimum observed, the simplest specification, corresponding to the lowest total number of knots, was chosen. The selected configuration is then re-estimated on the Full Training sample and finally evaluated on the Test sample. The resulting fitted values define the approximated implied volatility surface, denoted as IVS^{Spline} .

The cubic B-spline model serves as the main classical benchmark, providing a flexible yet interpretable representation of the implied volatility surface. It captures local curvature effectively while maintaining smoothness, offering a realistic approximation of the IV structure. Although its flexibility depends on the number and placement of knots, it remains a solid baseline for comparison with the machine learning and hybrid models that follow.

4.3 Machine Learning Models

Machine learning methods provide a flexible, data-driven alternative to classical parametric approaches, allowing complex nonlinear patterns to be learned directly from the data rather than imposed through predefined functional forms. In the context of implied volatility surface modelling, these algorithms can adapt to intricate relationships between moneyness, maturity, and other explanatory variables without requiring explicit structural assumptions.

This section focuses on tree-based ensemble methods, namely Random Forest (RF) and Extreme Gradient Boosting (XGBoost), which combine predictive accuracy with interpretability and robustness to overfitting. Both models are trained under the same empirical framework used in the classical approaches, enabling a consistent comparison of their ability to capture the nonlinear dynamics of implied volatility.

4.3.1 Random Forest

Random Forest (RF) is a non-parametric ensemble learning method that combines multiple decision trees to produce accurate and stable predictions. The algorithm was formally introduced by Breiman (2001), extending his earlier bagging framework (Breiman, 1996) and the random subspace method proposed by Ho (1998). By integrating bootstrap aggregation with random feature selection at each split, RF builds an ensemble of decorrelated trees whose averaged predictions achieve high accuracy and robustness to overfitting. Owing to these properties, it has become one of the most established algorithms for nonlinear regression and forecasting tasks, including applications in finance where relationships between variables are highly nonlinear.

Let x_i denote the vector of predictors for option i . The Random Forest prediction for implied volatility is given by the average output of all B trees in the ensemble:

$$IV_i^{RF} = \frac{1}{B} \sum_{b=1}^B f_b(x_i), \quad (4.9)$$

where $f_b(x_i)$ is the prediction from the b -th decision tree. By averaging across weakly correlated estimators, the model reduces variance and improves generalization. Model performance during training and tuning was assessed using the coefficient of determination (R^2), which measures the proportion of variance in implied volatility explained by the model and reflects its overall explanatory power.

Given the flexibility of tree-based methods, the set of predictors considered the two variables already used in the classical models, the log-moneyness and time to maturity, but also two additional variables, the risk-free rate (r) and a binary option type indicator ($type$). A preliminary validation compared alternative feature sets on the inner-training and inner-tuning samples, using RMSE and R^2 to assess the marginal contribution of the additional predictors. This pre-calibration step helped determine whether their inclusion improved the model's discriminative power. In addition, a feature-importance analysis was conducted on the final model specification to quantify the relative contribution of each selected variable. This provides an interpretable ranking of predictors according to their influence on the fitted volatility surface.

Once the final set of predictors was defined, the modeling stage was executed using a balanced sampling procedure to mitigate the uneven distribution of observations across moneyness–maturity regions. This ensured that the Random Forest learned a representative structure of the implied volatility surface rather than being biased toward denser areas of the data. The model's behavior is governed by several hyperparameters that control tree depth, smoothness, and feature selection. Following Probst et al. (2019), the tuning focused on the factors with the most significant influence on predictive performance and model stability. The hyperparameters used in the final implementation are summarized in Table 4.1. All models were trained with fixed random seeds to ensure reproducibility, using bootstrap sampling and full CPU parallelization.

Table 4.1 - Hyperparameters of the Random Forest model.

Hyperparameter	Description
n_estimators	Number of trees in the ensemble; higher values reduce variance but increase computation time.
max_depth	Maximum tree depth; limits complexity and overfitting.
min_samples_leaf	Minimum samples per leaf; larger values produce smoother predictions.
min_samples_split	Minimum samples required to split a node; regulates partition granularity.
max_features	Fraction of features randomly selected at each split; balances bias and variance.
max_samples	Fraction of training data used per tree (with replacement); lowers correlation among trees.

Hyperparameters were optimized using *RandomizedSearchCV* (Bergstra & Bengio, 2012) with two-fold cross-validation on the inner training subset. This stochastic approach offers an efficient alternative to exhaustive grid search while still exploring a broad range of parameter combinations. For each configuration, the model's performance was subsequently

evaluated on the inner tuning sample to validate its generalization ability. Parameter ranges were iteratively refined through successive runs and exploratory diagnostics to focus the search on the most informative regions of the parameter space.

The configuration achieving the best combination of cross-validated and out-of-sample (R^2) performance was retrained on the complete training set and finally evaluated on the temporally independent test sample. After several tuning rounds, no further improvements were found, indicating that the model had stabilized. This final version was used as the benchmark Random Forest model for the following analyses.

4.3.2 Extreme Gradient Boosting

XGBoost (Extreme Gradient Boosting) is a tree-based ensemble algorithm that extends the gradient boosting framework to achieve high predictive accuracy and computational efficiency. Unlike Random Forests, which average predictions from independently grown trees, XGBoost builds trees sequentially, each one correcting the residual errors of the previous ensemble. The method was first formalized by Friedman (2001) through the concept of gradient boosting and later optimized by Chen and Guestrin (2016) into the highly efficient XGBoost implementation. Its combination of gradient-based optimization, regularization, and subsampling makes it both flexible and robust against overfitting.

Let x_i denote the vector of predictors for observation i . The predicted implied volatility is obtained as the additive contribution of M sequential trees, each trained to correct the residual errors of the previous ensemble:

$$IV_i^{XGB} = \sum_{m=1}^M \eta f_m(x_i), \quad (4.10)$$

where η is the learning rate and $f_m(x_i)$ denotes the prediction of the m -th tree. Each tree is trained to minimize a regularized objective function of the form:

$$\mathcal{L}^{(m)} = \sum_i l(y_i, \hat{y}_i^{(m-1)} + f_m(x_i)) + \Omega(f_m), \quad (4.11)$$

where $l(\cdot)$ is the squared-error loss and $\Omega(f_m)$ penalizes model complexity through the regularization parameters γ and λ :

$$\Omega(f_m) = \gamma J + \frac{1}{2} \lambda \sum_j w_j^2 \quad (4.12)$$

As in the Random Forest model (Section 4.3.1), the risk-free rate and the option type indicator (type) were briefly tested as additional predictors, alongside the variables already used in the classical models, the log-moneyness and time to maturity. A short pre-calibration assessed their discriminative relevance, and the final model's feature-importance analysis quantified the relative contribution of each variable to the fitted volatility surface.

Also, as in the Random Forest model, a balanced sampling procedure was applied to the training data to mitigate uneven observation densities across moneyness-maturity regions. In the case of XGBoost, the model's hyperparameters, detailed in Table 4.2, control tree depth, learning rate, and regularization strength. The objective function and evaluation metric were kept consistent throughout the training process, while GPU-accelerated tree construction enables efficient parallel training without altering the model's statistical properties.

Table 4.2 - Main hyperparameters of the XGBoost model and their functional roles.

Hyperparameter	Description
n_estimators	Number of boosting rounds (trees).
learning_rate	Step size controlling each tree's contribution.
max_depth	Maximum tree depth; controls model complexity.
min_child_weight	Minimum sum of instance weights per leaf; higher values reduce overfitting.
gamma	Minimum loss reduction required to make a split.
subsample	Fraction of samples used per tree; adds stochasticity.
colsample_bytree	Fraction of features used per tree; analogous to feature subsampling in RF.
reg_lambda	L2 regularization term on leaf weights.
reg_alpha	L1 regularization term on leaf weights.

Hyperparameters were optimized through a random search over predefined ranges, with each configuration trained on a balanced subset of the inner training data and validated on the inner tuning sample using the root mean squared error (RMSE). Successive tuning iterations progressively narrowed the parameter ranges around high-performing regions. The configuration achieving the lowest validation RMSE was retrained on the complete training set and evaluated on the temporally independent test sample.

After multiple iterations, performance gains stabilized, and the resulting specification was adopted as the final XGBoost model.

4.4 Hybrid Models

Hybrid models aim to combine the structural smoothness of classical regressions with the nonlinear flexibility of machine learning algorithms. The main idea is to decompose the prediction of implied volatility into two complementary components: a structured baseline provided by a parametric model and a data-driven correction learned from its residuals. This approach allows machine learning methods to focus on the patterns that remain unexplained by the base model, potentially improving predictive performance without sacrificing interpretability.

Formally, let IV_i denote the observed implied volatility and IV_i^{Base} the fitted value from a classical model. The residual component is defined as:

$$\varepsilon_i = IV_i - IV_i^{Base}, \quad (4.13)$$

and the hybrid prediction can be expressed as:

$$IV_i^{Hybrid} = IV_i^{Base} + \hat{\varepsilon}_i^{ML}, \quad (4.14)$$

where $\hat{\varepsilon}_i^{ML}$ represents the residual correction estimated by a machine learning model. The base model captures the global structure of the implied volatility surface, while the ML component provides localized nonlinear adjustments.

In this dissertation, the spline model was adopted as the parametric baseline (IV^{Base}), given its flexibility and empirical fit to the implied volatility surface. Accordingly, the residuals used as the target variable in the hybrid models were computed as:

$$\varepsilon_i^{Splines} = IV_i - IV_i^{Splines}. \quad (4.15)$$

Two hybrid extensions were then developed by modeling these spline residuals ($\varepsilon_i^{Splines}$) using the Random Forest and XGBoost algorithms, respectively:

$$IV_i^{Hybrid_RF} = IV_i^{Splines} + \hat{\varepsilon}_i^{RF}, \quad (4.16)$$

$$IV_i^{Hybrid_XGB} = IV_i^{Splines} + \hat{\varepsilon}_i^{XGB}. \quad (4.17)$$

Both residual models were trained using the same predictor set employed in the respective ML approach, to ensure methodological consistency and facilitate comparison. Their configuration, tuning, and validation procedures followed precisely those described in Sections 4.3.1 and 4.3.2, differing only in the target variable, which here corresponds to the residuals from the spline model. This setup enables a direct evaluation of whether residual learning contributes incremental predictive power beyond both the classical and direct ML models.

4.5 Evaluation Framework and Performance Metrics

Model performance is evaluated through both in-sample (train sample) and out-of-sample (test sample) assessments. The first measures the model's ability to fit the data used for estimation, while the second tests its generalization capacity on unseen data. This dual setup distinguishes models that simply replicate historical patterns from those that capture the underlying structure of the implied volatility surface.

The performance measures applied consistently across all models are based on error statistics and explanatory power indicators. The direct evaluation metrics considered are the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Median Absolute Error (MedAE), Coefficient of Determination (R^2) and Bias, which capture prediction accuracy, explanatory strength, and systematic bias. Complementarily, two derived diagnostic ratios, RMSE/MAE and MAE/MedAE, provide additional information on the dispersion and asymmetry of the residuals. The corresponding formulas are presented in Appendix A (Formulas A.1–A.7).

The interpretation of the evaluation metrics is anchored in both the statistical properties of the modelling sample (Table 3.6) and empirical benchmarks reported in the literature. Error-based metrics such as RMSE, MAE, and MedAE are evaluated relative to the sample's standard deviation of implied volatility ($\sigma = 0.1238$). In this context, values in the range of 0.035–0.045, corresponding to approximately 30% of standard deviation, are regarded as representative benchmarks of good model performance for implied volatility surface applications. The Bias is evaluated relative to the mean implied volatility (mean=0.2186), where magnitudes below ± 0.02 , corresponding to about 10% of the average IV level, are considered negligible and indicate the absence of systematic directional errors. For the Coefficient of Determination, R^2 values in the 0.80–0.90 range are typically associated with strong explanatory performance in high-dimensional IVS modelling. The diagnostic ratios RMSE/MAE and MAE/MedAE are

interpreted using the statistical reference values described in Appendix A, where ratios around 1.25 and 1.18, respectively, correspond to normally distributed and well-behaved residuals.

Table 4.3 below summarizes the metrics and their interpretation.

Table 4.3 - Evaluation metrics and their interpretation.

Metric	Interpretation
RMSE	Measures the average magnitude of squared errors, giving higher weight to large deviations. In IVS modelling, RMSE values around 0.035–0.045 ($\approx 30\%$ of σ_{IV}) are typically regarded as indicative of good model performance.
MAE	Mean absolute deviation, providing a robust measure of average prediction error. Values in the 0.035–0.045 range are also considered representative benchmarks for accurate predictions.
MedAE	Median absolute deviation, representing the typical prediction error and reducing the influence of outliers. Reference values are of similar magnitude to those of MAE (≈ 0.035 –0.045).
R ²	Proportion of the variance in observed implied volatility explained by the model. Values between 0.80 and 0.90 are generally associated with strong explanatory performance.
Bias	Average signed deviation, indicating whether the model systematically over- or underestimates implied volatility. Magnitudes below ± 0.02 ($\approx 10\%$ of the mean IV) are typically considered negligible.
$RMSE/MAE$	Measures the relative weight of large errors compared to the average absolute error. Ratios around 1.25 are consistent with normally distributed residuals.
$MAE/MedAE$	Compares the mean and median absolute errors, reflecting asymmetry and robustness. Ratios close to 1.18 indicate regular residual behavior.

For direct comparison between any $Model_1$ and $Model_2$, we consider the percentual relative improvement ($RI_{(Model_1; Model_2)}$), defined as:

$$RI_{(Model_1; Model_2)} = \frac{RMSE_{Model_1} - RMSE_{Model_2}}{RMSE_{Model_1}} \times 100 \quad (4.18)$$

Finally, the numerical evaluation is complemented by graphical diagnostics that validate model behavior across the surface. These include heatmaps of absolute errors, three-dimensional IVS plots (observed vs predicted), histograms of residuals, residuals plotted against log-moneyness and maturity and observed versus predicted IV scatterplots. Such visual analyses help confirm the spatial coherence of predictions and detect localized patterns of systematic error.

Chapter 5

Results and Discussion

5.1 Overview of the Evaluation Framework

This chapter presents and discusses the empirical results obtained across all modeling approaches. The analysis is structured in three main blocks, following the methodological framework introduced in Chapter 4: classical models, machine learning models, and hybrid methods. The evaluation procedure follows the methodology described in Section 4.5, including the error metrics and graphical diagnostics used to assess model performance and residual behavior. All results are reported in a structured manner to ensure transparency and comparability, with complementary analysis outputs included in the Appendix for reference.

The overall framework ensures a consistent and reproducible evaluation pipeline, allowing for a fair and meaningful comparison across models and modelling paradigms.

5.2 Results from Classical Models

This section presents the empirical results for the classical modeling approaches described in Section 4.2. The analysis covers the performance of OLS, polynomial, and spline regressions in fitting the implied volatility surface, using the same data partitions, evaluation metrics, and diagnostic tools as established in the methodological framework.

5.2.1 Ordinary Least Squares

The Ordinary Least Squares (OLS) model follows the specification described in Section 4.2.1. As this model does not involve hyperparameter tuning, the regression coefficients were estimated directly from the training sample. The resulting coefficients, standardized betas, and relative weights are reported in Table 5.1. The results indicate that log-moneyness is by far the dominant explanatory variable, accounting for approximately 91% of the model's explanatory power. The time to maturity (T) plays a secondary role, with a modest yet consistent negative impact on implied volatility. This outcome is consistent with the typical shape of the IVS, where variations across moneyness are more pronounced than across maturity.

Table 5.1 - OLS Coefficients, Standardized Betas and Relative Weights.

Variable	Coefficient (β_i)	Standardized β_i^*	Relative Weight (%)
Intercept	0.1924	—	—
log_moneyness	-0.3735	-0.6259	91.04%
T (time to maturity)	-0.0333	-0.0616	8.96%

Considering the model's performance metrics presented in Table 5.2, the OLS regression achieves a R^2 value of 38.4% and 41.4% for the training and test samples, respectively, confirming its inability to capture the nonlinear structure of the IVS. The RMSE obtained in the test sample (0.0892) represents roughly 72% of the overall sample volatility ($\sigma = 0.1238$), highlighting the model's limited explanatory power and high residual variance relative to the intrinsic variability of implied volatilities.

Table 5.2 - Performance metrics of the OLS model.

Sample	RMSE	MAE	MedAE	$RMSE/MAE$	$MAE/MedAE$	R^2 (%)	Bias
Train	0.0981	0.0655	0.0507	1.498	1.292	38.36	0.0000
Test	0.0892	0.0571	0.0392	1.562	1.456	41.39	-0.0140

The high RMSE/MAE and MAE/MedAE ratios further suggest the presence of heavy tails and asymmetry in the error distribution. This is confirmed by the complementary graphical diagnostics (Appendix B, Figure B.1 – Figure B.5), which show systematic residual patterns consistent with unmodeled nonlinear effects.

Overall, the OLS model primarily serves as a baseline benchmark, illustrating the limitations of a purely linear specification. This is well evident by the fitted implied volatility surface (Figure 5.1), which is nearly planar, diverging markedly from the theoretical shape of the surface shown in Figure 3.6.

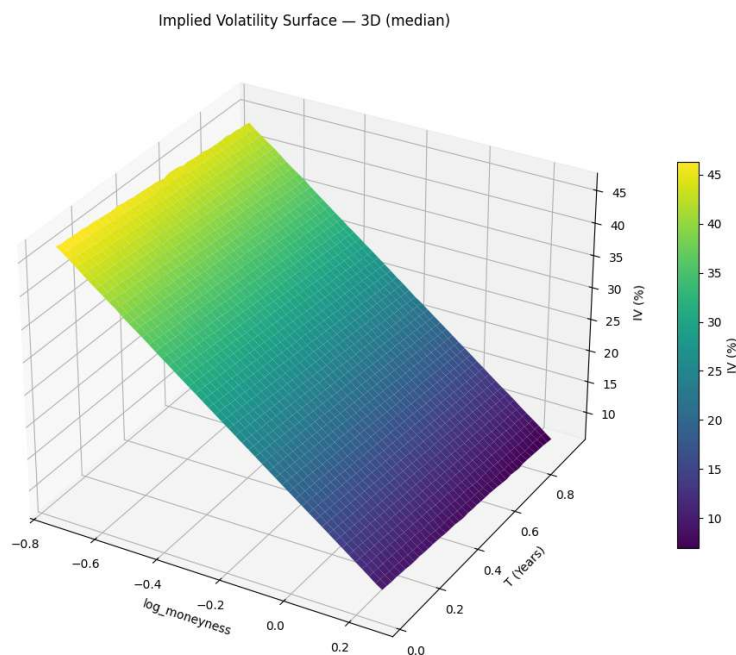


Figure 5.1 - Predicted implied volatility surface – OLS model (test sample).

5.2.2 Polynomial Regression

The polynomial regression model follows the specification described in Section 4.2.2. The analysis began with a systematic tuning procedure that tested combinations of polynomial degrees (d) and regularization parameters (α). The corresponding results are presented in Table 5.3. The selected specification, highlighted in bold, corresponds to a degree 5 polynomial with $\alpha^* = 0.001$. Beyond this point, additional complexity provided no meaningful gain and, for higher degrees, led to overfitting. The stable bias and smooth increase in explanatory power across degrees indicate that the regularization term effectively constrained coefficient magnitudes, ensuring numerical stability and preventing divergence as polynomial order increased.

Table 5.3 - Tuning results for polynomial regression models with different polynomial degrees. The selected specification (in bold) corresponds to degree 5 polynomial, $\alpha^ = 0.001$, determined by the 0.5% improvement and simplicity criterion.*

Degree	RMSE	R ² (%)	Bias	α	Features
1	0.0902	48.20	0.0112	10.000000	2
2	0.0789	60.36	0.0090	10.000000	5
3	0.0684	70.25	0.0076	1.000000	9
4	0.0613	76.07	0.0074	0.100000	14
5	0.0583	78.37	0.0080	0.001000	20
6	0.0593	77.64	0.0071	0.000100	27
7	0.4301	-1078.05	-0.0265	0.000001	35

After selecting the optimal degree and regularization parameter, the model was retrained on the whole training sample. The complete set of normalized coefficients is reported in Appendix C (Table C.1), while Table 5.4 presents the top five polynomial terms ranked by absolute normalized coefficient magnitude.

Table 5.4 - Top five polynomial terms and respective normalized coefficients ($d = 5$, $\alpha^ = 0.001$).*

Term	Normalized Coefficient (β)	Weight (%)
T^3	-26.2691	17.32
T^4	25.2484	16.65
$\log_moneyness \times T^3$	16.2888	10.74
$\log_moneyness \times T^2$	-13.7067	9.04
T^2	12.1682	8.02

The dominant terms are the higher-order maturity components T^3 and T^4 , which together explain more than one-third of the model's total weight. These nonlinear terms make time to maturity much more relevant than in the OLS model, where it accounted for only about

10% of the variance, compared with 90% from log-moneyness. The inclusion of interactions such as $\log_moneyness \times T^3$ and $\log_moneyness \times T^2$ further confirms that moneyness and maturity jointly shape the curvature of the implied volatility surface, consistent with the smile and term-structure patterns observed in SPX options.

Looking now at the model's performance metrics (Table 5.5), both in-sample and out-of-sample achieve an R^2 already within an acceptable range, showing a substantial improvement compared with the OLS model. However, the RMSE remains relatively high when compared with the theoretical sample volatility (approximately 48%), indicating that a significant portion of unexplained variance persists. The residual ratios are closer to the theoretical values, suggesting less skewness and thinner tails in the error distribution.

Table 5.5 - Performance metrics of the polynomial regression model.

Sample	RMSE	MAE	MedAE	RMSE/MAE	MAE/MedAE	R^2 (%)	Bias
Train	0.0632	0.0445	0.0343	1.420	1.296	74.42	0.0000
Test	0.0594	0.0420	0.0326	1.414	1.289	74.01	-0.0122

From a graphical standpoint, the results are also satisfactory. The fitted IVS (Figure 5.2) shape is closer to the theoretical one, while residuals appear narrower and more symmetric. Still, some degree of underestimation persists for short maturities and extreme moneyness regions, as illustrated in Appendix B (Figure B.6 – Figure B.10).

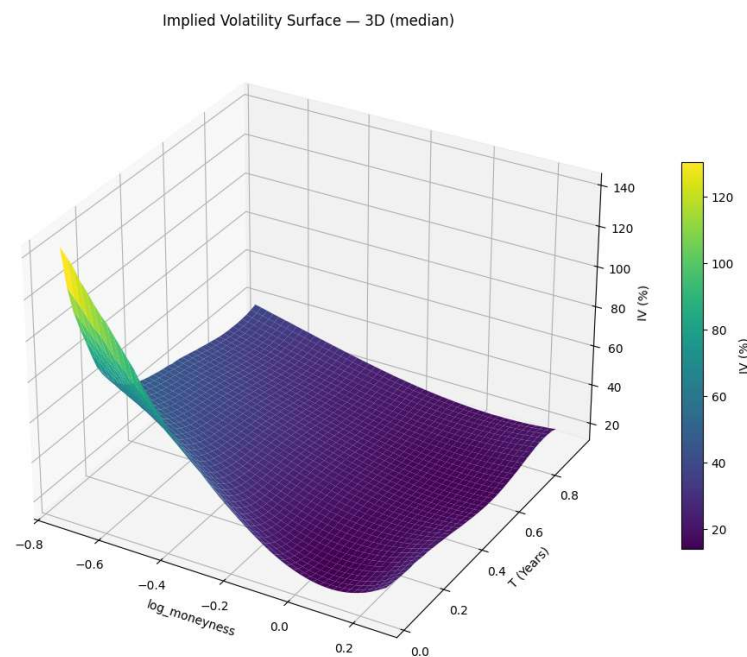


Figure 5.2 - Predicted implied volatility surface – Polynomial Regression (test sample).

5.2.3 Cubic B-Splines

The spline regression model follows the specification described in Section 4.2.3. The first step involved a tuning procedure that tested several knot configurations across moneyness (m) and maturity (T). The results showed clear performance gains up to a certain level, with knots $(m, T) = (12, 4)$ and $\alpha^* = 0.01$, after which additional knots produced negligible improvements. The final configuration, summarized in Table 5.6, includes 65 features and represents the optimal trade-off between flexibility and efficiency.

Table 5.6 - Tuning results for cubic B-spline models with different knot configurations in moneyness (m) and maturity (T).

Selected configuration (0.5% improvement rule + simplicity): knots(m, T) = (12, 4), α^ (RidgeCV) = 0.01, 65 features expanded.*

n_knots_m	n_knots_t	RMSE	MAE	R ² (%)	Bias	Features
10	3	0.0386	0.0280	90.53	0.0057	44
10	4	0.0380	0.0277	90.79	0.0054	55
10	5	0.0379	0.0276	90.86	0.0054	66
12	3	0.0385	0.0280	90.54	0.0057	52
12	4	0.0380	0.0277	90.82	0.0054	65
12	5	0.0378	0.0276	90.88	0.0054	78
14	3	0.0385	0.0280	90.55	0.0057	60
14	4	0.0380	0.0277	90.82	0.0054	75
14	5	0.0378	0.0276	90.89	0.0054	90

After selecting the optimal configuration, the model was retrained on the full training sample. The complete set of normalized spline coefficients is reported in Appendix D (Table D.1), while Table 5.7 presents the top five terms ranked by absolute normalized coefficient.

Table 5.7 – Top five Cubic B-Splines terms and respective normalized coefficients ($d = 3$, $\alpha^ = 0.01$).*

Term	Normalized Coefficient (β)	Weight (%)
Bm2×BT0	+92.4555	16.82
Bm1×BT0	-65.2275	11.86
Bm1×BT1	+61.0031	11.10
Bm0×BT3	+43.9439	7.99
Bm11×BT0	+29.2769	5.32

The largest weights concentrate at short maturities (BT_0 and BT_1), showing that most curvature is captured at the front end of T . Along moneyness, the strongest basis functions are around the at-the-money region (Bm_{-2} to Bm_2) with mixed signs, consistent with a smile turning point around ATM. Non-negligible weights on wing bases (Bm_{11} , Bm_{12}) and smaller

contributions at longer maturities (BT_3-BT_4) indicate that the model captures both local ATM curvature and wing steepness while preserving smoothness across maturities.

The flexibility revealed by the spline basis functions is also reflected in the model's performance metrics (Table 5.8). The R^2 of 86.3% on the test sample is a substantial improvement over the polynomial model and the corresponding RMSE of 0.0431 represents roughly 35% of the sample volatility, indicating an acceptable level of residual variability.

Table 5.8 - Performance metrics of the Splines model.

Sample	RMSE	MAE	MedAE	RMSE/MAE	MAE/MedAE	R^2 (%)	Bias
Train	0.0459	0.0327	0.0250	1.402	1.310	86.51	0.0000
Test	0.0431	0.0324	0.0249	1.331	1.298	86.33	-0.0103

The residual ratios are similar to those of the polynomial model, indicating comparable residual symmetry and moderate tail thickness, however, the complementary graphics (Figure B.11 - Figure B.15) reveal clear improvement. The residuals are more tightly distributed around zero, with reduced heteroskedasticity and no visible structural patterns, and although slight underestimation remains at extreme moneyness and very short maturities, the overall residuals appear more homogeneous and random.

The resulting IVS (Figure 5.3) displays smoother curvature and a closer match to the theoretical shape, especially near the at-the-money region and for short maturities.

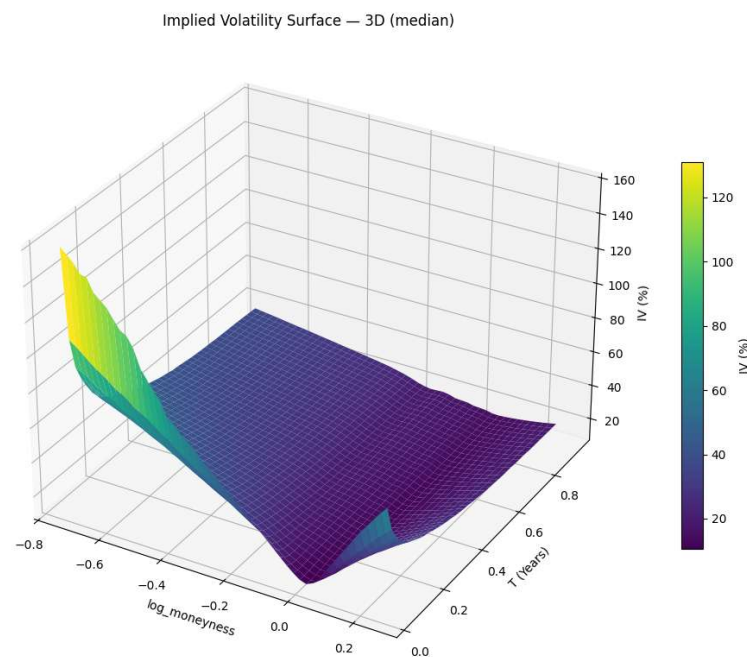


Figure 5.3 - Predicted implied volatility surface – Splines Model (test sample).

5.3 Results of Machine Learning Models

This section presents the empirical results for the machine learning models described in Section 4.3. The analysis includes the Random Forest and XGBoost regressors, designed to capture nonlinear relationships between implied volatility, moneyness and maturity, and to potentially incorporate additional predictors such as the risk-free rate or option type.

These ensemble-based methods aim to enhance predictive accuracy and generalization beyond the capabilities of the classical models.

5.3.1 Random Forest

The Random Forest model follows the specification described in Section 4.3.1. The first step involved a predictor relevance analysis to evaluate the contribution of the additional predictors, the risk-free rate (r), and the option type indicator ($type$). The results, presented in Table 5.9, show that including the risk-free rate led to a slight deterioration in performance, whereas the option type yielded a clear improvement. Based on these findings, a final feature configuration (m , T , $type$) was tested, confirming that the option type indicator provides additional explanatory power ($\Delta R^2 = +1.74$ and $\Delta RMSE = -0.0038$).

Table 5.9 - Predictor relevance analysis for the Random Forest model. Each row compares two feature sets, showing the change in RMSE and R^2 when adding a new predictor. The final feature configuration adopted was ($\log$ moneyness, T , $type$).

Model Comparison	RMSE (1)	RMSE (2)	$\Delta RMSE$	R^2 (1)	R^2 (2)	ΔR^2
(A) m , T , $r \rightarrow + type$	0.04494	0.04216	-0.00278	87.14	88.68	+1.54
(B) m , T , $type \rightarrow + r$	0.03444	0.04216	+0.00773	92.45	88.68	-3.77
(C) m , $T \rightarrow + type$	0.03819	0.03444	-0.00376	90.71	92.45	+1.74

Having defined the final set of predictors, the next step consisted of the hyperparameter tuning phase. Following the respective methodology, several rounds of execution were performed to iteratively refine the parameter ranges, converging on a stable, computationally efficient configuration.

Table 5.10 presents the final grid used and the selected hyperparameter values obtained from 2-fold cross-validation (60 fits in total). The chosen configuration favored a moderate tree depth ($max_depth = 14$) and a relatively large ensemble size ($n_estimators = 500$), ensuring stable and smooth predictions. The constraints on node size ($min_samples_split = 200$; $min_samples_leaf = 50$) imposed a coarse partitioning structure, effectively preventing overfitting and limiting unnecessary splits. The sampling parameters ($max_samples = 0.6$; $max_features = 0.8$) introduced additional randomness in the selection of observations and

predictors. Given that only three explanatory variables were used, $max_features = 0.8$ means that each split considered two predictors at random, which provided limited but useful diversity among trees. The relatively small subsample fraction ($max_samples = 0.6$) further supported generalization by training each tree on partially different subsets of data.

Table 5.10 - Hyperparameter grid used for Random Forest tuning and the selected configuration based on 2-fold cross-validation (60 fits in total).

Hyperparameter	Range Tested	Selected Value
Number of trees ($n_estimators$)	[500, 600]	500
Maximum tree depth (max_depth)	[14, 16]	14
Minimum samples per leaf ($min_samples_leaf$)	[50, 100]	50
Minimum samples per split ($min_samples_split$)	[200, 300, 400]	200
Maximum features ($max_features$)	[0.8, 1.0]	0.8
Maximum samples ($max_samples$)	[0.6, 0.8]	0.6

Overall, this configuration represents a well-balanced compromise between flexibility and stability, enabling the Random Forest model to capture nonlinear behavior in the implied volatility surface while maintaining consistent performance across validation folds. This is confirmed by the results reported in Table 5.11, where the model achieves an R^2 of 89.2% on the test sample, an RMSE of 0.0384, and a small negative bias (−0.0114). The RMSE corresponds to roughly 31% of the sample volatility, which reflects a strong explanatory capacity relative to the intrinsic variability of implied volatilities.

Table 5.11 - Performance metrics of the Random Forest model.

Sample	RMSE	MAE	MedAE	RMSE/MAE	MAE/MedAE	R^2 (%)	Bias
Train	0.0421	0.0288	0.0214	1.462	1.346	88.65	0.0000
Test	0.0384	0.0275	0.0193	1.395	1.422	89.16	-0.0114

Although the RMSE/MAE and MAE/MedAE ratios are slightly higher than those of the spline model, this is expected given the overall improvement in fit quality. As model accuracy increases, reflected in smaller RMSE, MAE and MedAE values, residuals become smaller on average but relatively more sensitive to extreme observations. This leads to marginally higher ratio values without degrading performance. The complementary graphical diagnostics (Figure B.16–Figure B.20) reinforce these findings: the observed-versus-predicted plot shows a dense alignment along the diagonal, the residual histogram is centered and nearly symmetric, and the residuals plotted against both maturity and log-moneyness reveal no pronounced structural patterns. Only mild underestimation persists at very short maturities and

in the deep wings, suggesting that most nonlinearities are successfully captured by the ensemble.

Regarding feature relevance, Table 5.12 summarizes both impurity-based and permutation importance measures. The results confirm that log-moneyness is by far the most influential variable, followed by time to maturity, while the option type plays a secondary but non-negligible role. This ranking aligns with the earlier predictor relevance analysis, where including the option type variable yielded the largest marginal improvement in explanatory power.

Table 5.12 - Variable importance measures for the Random Forest model.

Feature	Impurity Importance	Permutation Importance	Permutation Std
log_moneyness	0.7647	2.6538	0.0030
T	0.1510	0.8022	0.0014
option type	0.0843	0.0673	0.0002

Overall, the fitted implied volatility surface obtained from the Random Forest model (Figure 5.4) reproduces the expected smile pattern and smooth term-structure evolution, closely resembling the theoretical shape while retaining local flexibility and numerical stability.

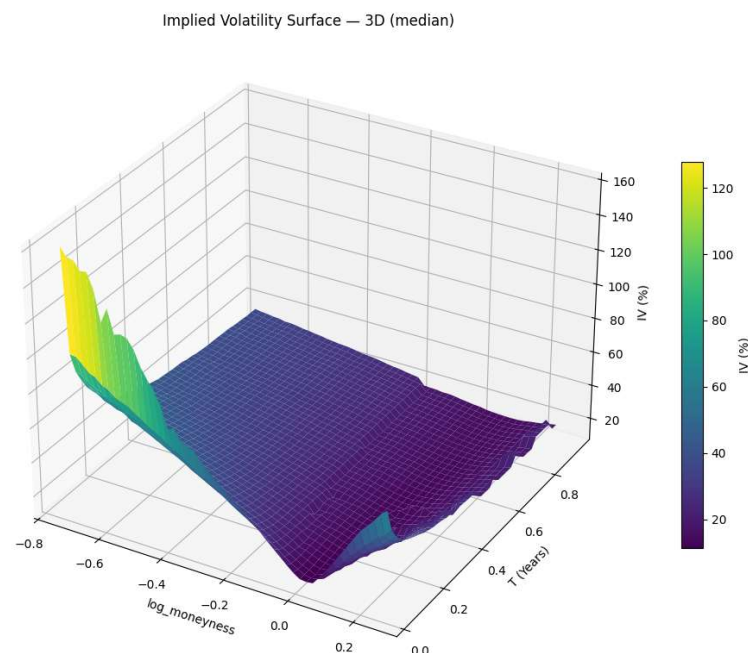


Figure 5.4 - Predicted implied volatility surface – Random Forest Model (test sample).

5.3.2 XGBoost

The XGBoost model follows the specification described in Section 4.3.2. As with Random Forest, the first step was a predictor relevance analysis to assess the contribution of the additional variables, and the results (Table 5.13) led to the same conclusion. Adding the risk-free rate slightly worsened performance, whereas the option type indicator produced a clear improvement.

The model final feature set (m, T, type) shows a markedly stronger incremental value from the inclusion of the option type variable, with $\Delta R^2 = +17.48$ and $\Delta RMSE = -0.0280$ confirming substantial additional explanatory power relative to specifications without this indicator.

Table 5.13 - Predictor relevance analysis for the XGBoost model.

Model Comparison	RMSE (1)	RMSE (2)	$\Delta RMSE$	R^2 (1)	R^2 (2)	ΔR^2
LM, T, r $\rightarrow + \text{type}$	0.04577	0.04502	-0.00075	86.66	87.10	+0.43
LM, T, type $\rightarrow + \text{r}$	0.03518	0.04502	+0.00984	92.12	87.10	-5.02
LM, T $\rightarrow + \text{type}$	0.06311	0.03518	-0.02793	74.64	92.12	+17.48

Having defined the final set of predictors, the next step consisted of the hyperparameter tuning phase. Table 5.14 presents the final grid used, efficiently refined through several rounds of executions, and the selected hyperparameter values obtained from cross-validation.

The optimal configuration combines a moderate learning rate (0.0347) with a relatively small number of boosting rounds ($n_estimators = 1600$), indicating a stable convergence process where each tree contributes incrementally to improving the model fit. The selected tree depth ($max_depth = 5$) and minimum child weight (5) balance flexibility and regularization, allowing the model to capture relevant nonlinearities without overfitting, and the parameters $gamma$, $subsample$ and $colsample_bytree$ introduce control over complexity and randomness. The low $gamma$ value (0.0563) enables frequent but meaningful splits, while the sampling fractions ($subsample = 0.7050$, $colsample_bytree = 0.6946$) ensure adequate stochasticity across trees, improving generalization. Finally, the regularization terms ($reg_lambda = 4.3835$, $reg_alpha = 0.8137$) apply moderate L2 and relatively strong L1 penalization on leaf weights, contributing to numerical stability and model sparsity.

Overall, the selected configuration yields a well-regularized, efficiently trained ensemble, capable of capturing subtle nonlinear patterns in the IVS while maintaining robust performance and computational efficiency.

Table 5.14 - XGBoost hyperparameter search ranges and selected values.
U(a,b) denotes a continuous uniform search interval.

Parameter	Search Range	Selected Value
n_estimators	{1600, 2000, 2400, 3000}	1600
learning_rate	U(0.020, 0.040)	0.0347
max_depth	{3, 4, 5}	5
min_child_weight	{5, 7, 9, 11}	5
gamma	U(0.05, 0.30)	0.0563
subsample	U(0.65, 0.85)	0.7050
colsample_bytree	U(0.65, 0.85)	0.6946
reg_lambda	U(1.0, 6.0)	4.3835
reg_alpha	U(0.10, 0.90)	0.8137

Looking at the performance metrics in Table 5.15 we can see that the XGBoost model achieves an R^2 of 87.5% on the test sample, with an RMSE of 0.0412 and a negligible bias (–0.0158). Although these values are slightly below those obtained by the Random Forest model, they still indicate strong predictive accuracy and discriminative capacity. The relatively close results between training and test samples confirm the model’s robustness and absence of overfitting, consistent with the regularization mechanisms and sampling constraints introduced during tuning.

Table 5.15 - Performance metrics of the XGBoost model.

Sample	RMSE	MAE	MedAE	RMSE/MAE	MAE/MedAE	R^2 (%)	Bias
Train	0.0436	0.0299	0.0222	1.461	1.348	87.79	0.0000
Test	0.0412	0.0293	0.0199	1.404	1.474	87.51	-0.0158

As in the Random Forest model, the residual ratios remain moderately above the theoretical benchmarks, reflecting the same pattern of concentrated yet occasionally heavy-tailed errors. This behavior is consistent with the higher precision of boosted trees, which tend to capture most of the variance while leaving only small residual deviations concentrated in extreme observations. The complementary graphical diagnostics (Figure B.21 – Figure B.25) confirm this result, showing tightly clustered points around the diagonal, a nearly symmetric residual distribution, and limited structure along maturity or moneyness dimensions. Overall, the residuals are well-behaved and display a level of homogeneity comparable to that of the Random Forest model, confirming that the boosting procedure achieves similar explanatory power with a slightly different bias–variance trade-off.

Finally, the last step of the modelling procedure consisted of analyzing the importance of the features used in the XGBoost model. As anticipated in the predictor relevance analysis,

the option type indicator (type) was expected to play a more significant role here than in the Random Forest model. The results summarized in Table 5.16, show that log-moneyness remains the dominant driver of the implied volatility surface, followed by time to maturity (T) and the option type indicator. However, compared with the RF model, the latter assumes indeed a much larger relative importance, especially in gain terms, suggesting that XGBoost exploits interactions between moneyness and option type more effectively.

Table 5.16 - Variable importance measures for the XGBoost model.

Gain importance measures each variable's contribution to loss reduction, while permutation importance captures the rise in prediction error when that variable is shuffled.

Feature	Gain Importance	Permutation Importance	Permutation Std
log_moneyness	174.2503	2.9115	0.0233
T	40.9402	0.7413	0.0040
option type	84.3403	0.2074	0.0017

These results indicate that while moneyness remains the dominant driver of variance, XGBoost assigns greater weight to the interaction between maturity and option type, capturing local asymmetries between calls and puts. The fitted implied volatility surface (Figure 5.5) is smooth and realistic, closely matching the empirical patterns. Overall, the XGBoost model achieves performance comparable to the Random Forest, confirming its robustness and effectiveness in modelling the IVS.

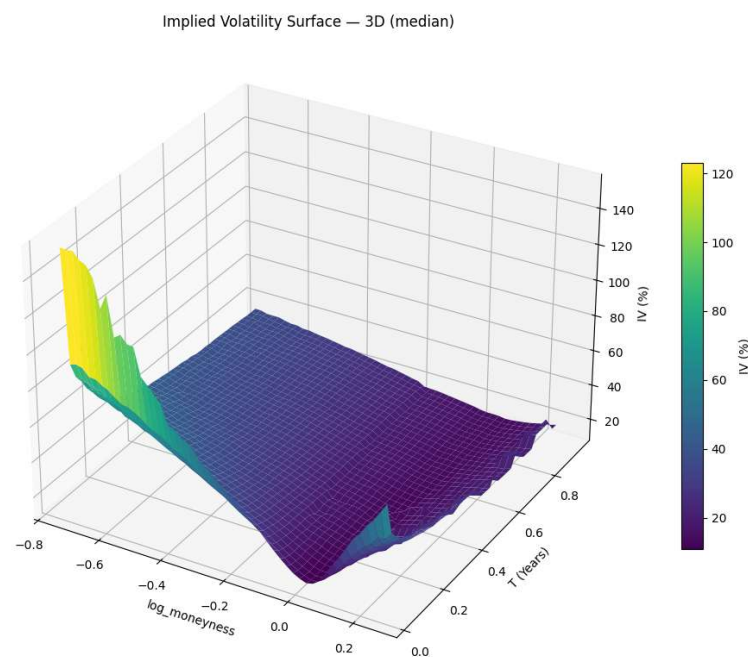


Figure 5.5 - Predicted implied volatility surface – XGBoost Model (test sample).

5.4 Results of Hybrid Models

This section presents the empirical results for the hybrid machine learning models described in Section 4.4. Two hybrid specifications were developed by combining the cubic B-spline regression with a subsequent machine learning model applied to its residuals, using Random Forest and XGBoost, respectively. The methodological structure of each hybrid model follows the same settings and tuning principles of its standalone counterparts, ensuring full comparability across approaches.

The main objective of these hybrid models is to improve predictive performance without sacrificing interpretability, leveraging the spline model's smooth structural fit and the machine learning component's ability to capture residual nonlinearities.

5.4.1 Hybrid Random Forest

The first hybrid model combines the cubic B-spline regression with a subsequent Random Forest applied to its residuals. The process begins by estimating residuals from the spline model, whose results, presented in Section 5.2.3, represent the component of implied volatility unexplained by the spline fit. The second stage applies the same Random Forest methodology previously used in Section 4.3.1, suitably adjusted to predict this residual component. To ensure methodological consistency, the same set of predictors was used.

Passing to the tuning phase, Table 5.17 summarizes the hyperparameter grid tested and the final configuration selected via cross-validation. The optimal model uses 1,200 trees with a maximum depth of 12, indicating a preference for moderately complex, shallow trees relative to the standalone Random Forest. The chosen constraints on node size ($min_samples_split = 150$; $min_samples_leaf = 100$) further limit unnecessary partitions, ensuring smoother predictions over the residual surface. Sampling parameters ($max_samples = 0.6$; $max_features = 0.8$) maintain the same degree of randomness and diversity across trees as in the previous models, promoting generalization and reducing variance.

Table 5.17 - Hyperparameter grid used for Hybrid Random Forest tuning and the selected configuration based on 2-fold cross-validation (96 fits in total).

Hyperparameter	Range Tested	Selected Value
Number of trees ($n_estimators$)	{800, 1000, 1200}	1200
Maximum tree depth (max_depth)	{12, 14, 16}	12
Minimum samples per leaf ($min_samples_leaf$)	{30, 50, 100}	100
Minimum samples per split ($min_samples_split$)	{100, 150, 300}	150
Maximum features ($max_features$)	{"sqrt", 0.8}	0.8
Maximum samples ($max_samples$)	{0.6, 0.8, 0.9}	0.6

Compared with the standalone Random Forest, the tuned configuration reflects the different nature of the prediction task. Here, the target variable corresponds to the spline residuals, which exhibit lower variance and weaker structural patterns, so the model naturally converges toward a simpler configuration, with fewer effective splits and stronger regularization, while preserving sufficient flexibility to capture the remaining nonlinear components of the IVS.

With the hyperparameters defined, the residuals were modelled both in-sample and out-of-sample, following the procedure described in Section 4.4. The final implied volatility surface ($IV_i^{Hybrid_RF}$) was then reconstructed using Equation (4.17), combining the spline baseline with the Random Forest-based residual predictions. The resulting performance metrics are reported in Table 5.18, showing a R^2 of 89.2% on the test sample and an RMSE of 0.0383, corresponding to roughly 31% of the sample volatility. These results indicate that the hybrid model effectively improves upon the spline-only baseline while preserving the smoothness and interpretability of the original structure.

Table 5.18 - Performance metrics of the Hybrid Random Forest model.

Sample	RMSE	MAE	MedAE	RMSE/MAE	MAE/MedAE	R² (%)	Bias
Train	0.0421	0.0288	0.0213	1.462	1.349	88.64	0.0000
Test	0.0383	0.0275	0.0192	1.395	1.428	89.18	-0.0113

Compared with the standalone Random Forest model, the hybrid version delivers virtually identical performance across all error metrics, confirming that the spline pre-fitting step does not degrade predictive accuracy. In fact, the slightly higher R^2 and marginally lower RMSE suggest that the hybrid structure provides a slight gain in stability and smoothness, particularly by isolating residual nonlinearities while preserving the global shape captured by the spline.

The residual ratios are also consistent with those of the Random Forest model, reflecting the same pattern of compact, near-symmetric residuals with limited heavy tails. This similarity is equally observed in the graphical diagnostics (Figure B.26 – Figure B.30), where residual plots exhibit patterns almost indistinguishable from those of the standalone model.

Overall, the hybrid approach successfully integrates the interpretability of classical models with the predictive strength of ensemble methods, yielding a well-balanced and robust IVS representation.

In terms of feature importance, the results in Table 5.19 show a more balanced contribution across predictors than in the standalone Random Forest model. Log-moneyness

remains the most influential variable, but its dominance is reduced (impurity importance of 0.46 versus 0.76 previously), while time to maturity gains relative importance, suggesting that most of the residual variability comes from local effects along the maturity dimension. The option type indicator retains a minor yet non-negligible contribution, correcting small asymmetries between calls and puts not captured by the spline baseline. This more even distribution of importance confirms that the hybrid Random Forest primarily refines the spline fit by modelling subtle nonlinear interactions rather than reconstructing the global structure of the implied volatility surface.

Table 5.19 - Variable importance measures for the Hybrid Random Forest model.

Feature	Impurity Importance	Permutation Importance	Permutation Std
log_moneyness	0.462	0.437	0.0012
T	0.398	0.322	0.0004
option type	0.140	0.081	0.0004

Overall, the strong predictive performance achieved by the standalone machine learning models can be effectively reproduced through this hybrid approach. By modelling only the residual component, the Random Forest successfully refines the spline baseline while preserving its structural interpretability. The resulting implied volatility surface is presented below, in Figure 5.6.

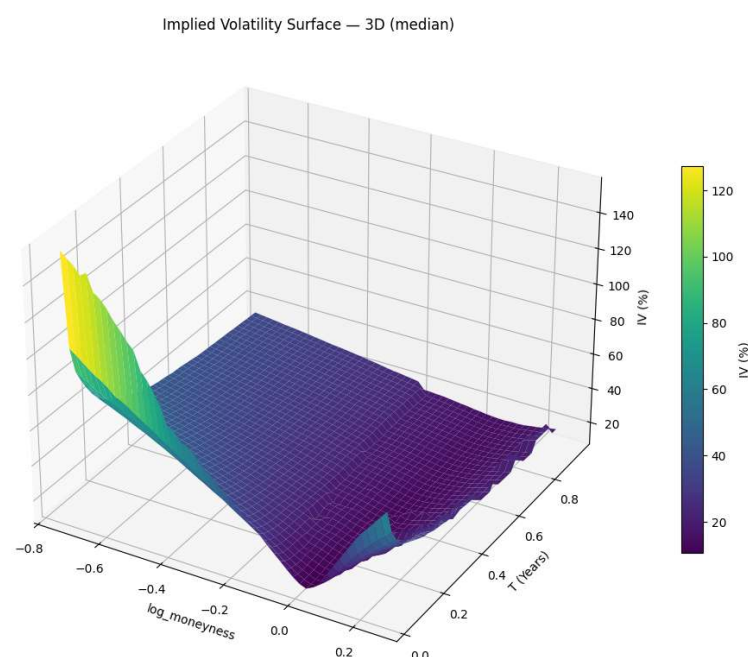


Figure 5.6 - Predicted implied volatility surface – Hybrid Random Forest Model (test sample).

5.4.2 Hybrid XGBoost

The second hybrid model combines the same cubic B-spline regression with a subsequent XGBoost applied to its residuals. It also uses the residuals obtained from the spline model presented in Section 5.2.3, applying the XGBoost methodology previously described in Section 4.3.2. To ensure methodological consistency and comparability with previous models, the same set of predictors was used.

Regarding the tuning phase, Table 5.20 presents the grid of hyperparameters tested and the final configuration selected through cross-validation. The optimal specification combines a large number of boosting rounds ($n_estimators = 5000$) with a low learning rate (0.0228), suggesting a slow, stable convergence process. The model adopts moderate tree depth ($max_depth = 5$) and a high minimum child weight (11), both contributing to strong regularisation and preventing overfitting to small residual fluctuations. The gamma value (0.2773) introduces a conservative split threshold, while the sampling parameters ($subsample = 0.7104$; $colsample_bytree = 0.7317$) maintain sufficient stochasticity to improve generalisation. Finally, the regularisation terms ($reg_lambda = 5.7313$; $reg_alpha = 0.3435$) impose additional L2 and L1 penalties, ensuring numerical stability and reducing model variance.

Table 5.20 – Hybrid XGBoost hyperparameter search ranges and selected values.
U(a,b) denotes a continuous uniform search interval.

Parameter	Search Range	Selected Value
n_estimators	{4000, 5000, 6000}	5000
learning_rate	U(0.020, 0.040)	0.0228
max_depth	{3, 4, 5}	5
min_child_weight	{5, 7, 9, 11}	11
gamma	U(0.05, 0.30)	0.2773
subsample	U(0.65, 0.85)	0.7104
colsample_bytree	U(0.65, 0.85)	0.7317
reg_lambda	U(1.0, 6.0)	5.7313
reg_alpha	U(0.10, 0.90)	0.3435

Although this tuning was performed on a different target variable (spline residuals instead of implied volatility), the overall configuration remains broadly consistent with that of the standalone XGBoost model. The main differences, particularly the higher *min_child_weight*, larger *gamma*, and stronger regularization, reflect the lower variance and smoother structure of the residuals, which require a more constrained, conservative model. Overall, the tuned parameters confirm that the hybrid specification favors robustness and

controlled flexibility, adapting effectively to the residual learning task without loss of generalization.

After the tuning step, the residuals were modelled on both the training and test samples, and the final implied volatility surface ($IV_i^{Hybrid_XGB}$) was obtained using Equation (4.18), combining the spline baseline with the XGBoost-based residual predictions. The corresponding performance evaluation results are presented in Table 5.21. The model achieves a R^2 of 88.9% and an RMSE of 0.0388 on the test sample, representing approximately 31% of the sample volatility. The bias remains negligible (-0.0110), confirming that the hybrid model provides accurate and well-centered predictions across samples.

Table 5.21 - Performance metrics of the Hybrid XGBoost model.

Sample	RMSE	MAE	MedAE	RMSE/MAE	MAE/MedAE	R^2 (%)	Bias
Train	0.0426	0.0293	0.0220	1.454	1.335	88.34	0.0000
Test	0.0388	0.0280	0.0199	1.385	1.409	88.92	-0.0110

Compared with both the standalone XGBoost and the Hybrid Random Forest models, this specification delivers very similar performance, confirming that the hybrid structure preserves the predictive capacity of the underlying ML algorithm. The small improvement in stability and bias reduction, together with nearly identical RMSE and R^2 values, suggests that the spline pre-fitting successfully removes the large-scale deterministic patterns of the IVS, allowing the boosting algorithm to focus solely on local nonlinearities. The residual ratios ($RMSE/MAE = 1.39$; $MAE/MedAE = 1.41$) remain close to theoretical benchmarks and consistent with the previous models, indicating compact and symmetric residuals with limited heavy tails.

The graphical diagnostics (Appendix B, Figure B.31 – Figure B.35) corroborate these quantitative results. The observed-versus-predicted plot shows a dense, well-aligned pattern along the diagonal, indicating high fit accuracy across the sample. The residual histogram is centered and approximately symmetric, while the residual plots against both log-moneyness and maturity display no significant structural patterns or heteroskedasticity. Minor underestimation persists at very short maturities and deep out-of-the-money regions, consistent with the behavior observed in the Hybrid Random Forest model. Overall, the graphical evidence confirms that the hybrid XGBoost specification achieves a similarly smooth and well-behaved fit, capturing the main nonlinearities of the implied volatility surface without introducing instability or overfitting.

At last, looking at the feature importance results reported in Table 5.22, the dominant role of log-moneyness is evident, remaining by far the most influential variable in explaining the residual structure. Time to maturity provides a secondary yet meaningful contribution, capturing local temporal effects not fully accounted for by the spline baseline, and the option type indicator shows only marginal influence, suggesting that most asymmetries between calls and puts were already absorbed during the initial spline fitting.

Table 5.22 - Variable importance measures for the Hybrid XGBoost model.

Feature	Gain Importance	Permutation Importance	Permutation Std
log_moneyness	82.9970	0.3694	0.0014
T	4.6324	0.2109	0.0019
option type	1.8668	0.0261	0.0003

Overall, the importance pattern closely mirrors that of the Hybrid Random Forest, indicating that both hybrid models capture similar residual dynamics with comparable explanatory balance. In contrast, the standalone XGBoost displayed a more balanced weighting across predictors, reflecting its direct modelling of the full implied volatility surface rather than the residual component alone.

The resulting implied volatility surface is presented below, in Figure 5.7.

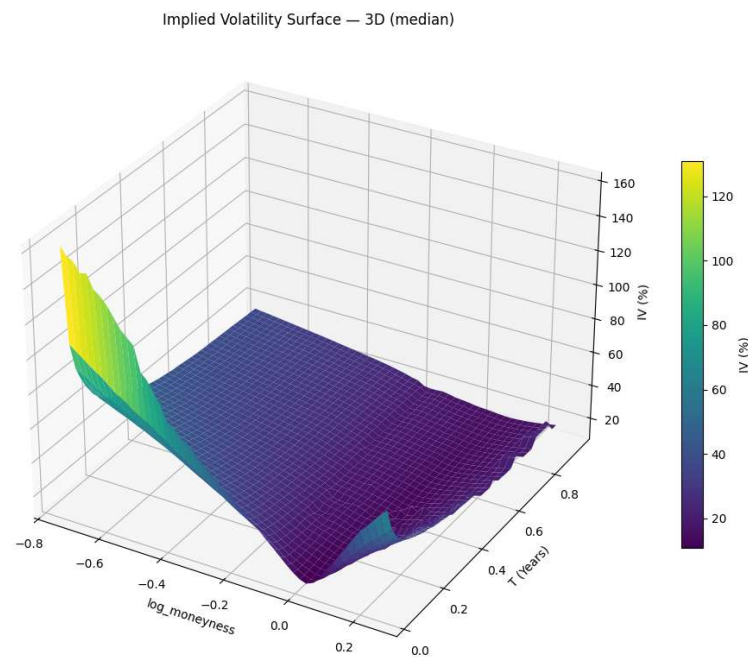


Figure 5.7 - Predicted implied volatility surface – Hybrid XGBoost Model (test sample).

5.5 Comparative Analysis and Key Insights

This section summarizes and compares the empirical results obtained across all modelling approaches, consolidating the quantitative metrics and graphical analyses presented throughout this chapter. The comparison follows the methodological hierarchy established from classical models to machine learning and hybrid methods, allowing a consistent assessment of how each one contributes to the modelling of the implied volatility surface.

Table 5.23 presents the comparative results for the test sample, in terms of coefficient of determination (R^2), bias, root mean squared error (RMSE), the ratio of RMSE to the sample volatility, and the relative improvement (RI) defined based on formula 4.18.

Table 5.23 - Comparative performance of all models on the test sample.

The table reports the coefficient of determination (R^2), bias, root mean squared error (RMSE), the ratio of RMSE to the sample volatility ($\sigma = 0.1238$), and the relative improvement (RI) with respect to the benchmark model in each comparison.

Model	R^2 (%)	Bias	RMSE	$RMSE/\sigma$	Relative Improvement (RI)
OLS	41.39	-0.0140	0.0892	72.03%	–
Polynomial	74.01	-0.0122	0.0594	47.96%	$RI_{(OLS; Poly)} = +33.41\%$
Cubic B-Splines	86.33	-0.0103	0.0431	34.81%	$RI_{(Poly; Splines)} = +27.44\%$
Random Forest	89.16	-0.0114	0.0384	31.03%	$RI_{(Splines; RF)} = +10.90\%$
XGBoost	87.51	-0.0158	0.0412	33.28%	$RI_{(Splines; XGBoost)} = +4.41\%$
Hybrid – RF	89.18	-0.0113	0.0383	30.93%	$RI_{(Splines; H_{RF})} = +0.26\%$
Hybrid - XGBoost	88.92	-0.0110	0.0388	31.34%	$RI_{(Splines; H_{XGBoost})} = +5.83\%$

The progression from the linear OLS model to the fifth-degree polynomial regression marks the largest single improvement, with the RMSE decreasing by about 33% and R^2 rising from 41% to 74%. This confirms that introducing nonlinear terms for moneyness and maturity allows the model to capture much of the curvature and slope structure of the IVS that a linear specification cannot.

Moving from polynomial to cubic B-splines brings a further 27% reduction in RMSE and an increase in R^2 to 86.3%. The spline framework offers a smoother and more flexible representation, reproducing the smile and term-structure effects with high stability and minimal bias.

The transition to machine learning models consolidates the performance gains. The Random Forest reaches an RMSE of 0.0384 (around 31% of the sample volatility) and a R^2 of 89.2%, representing an additional 11% improvement relative to the spline model. The XGBoost

model achieves similar accuracy, with slightly higher RMSE but comparable explanatory power, showing that both ensemble approaches effectively capture nonlinearities and complex interactions.

The hybrid models, combining the spline baseline with residual learning, achieve almost identical results to their machine learning counterparts. The Hybrid Random Forest and Hybrid XGBoost models deliver R^2 values of 89.2% and 88.9%, respectively, and RMSEs close to 0.038, confirming that their approach matches ML accuracy while preserving the interpretability of the spline foundation. The marginal differences in performance indicate that once the main structure is captured by the spline, the residual learning stage primarily refines local irregularities rather than altering the global fit.

Overall, the quantitative results reveal a clear progression in explanatory power and precision, culminating in a stable goodness of fit around $R^2 \approx 89\%$. Each methodological refinement, from polynomial flexibility to spline smoothness and ensemble learning, adds measurable value, with the hybrid framework demonstrating that such gains can be achieved without compromising interpretability.

The graphical analyses support these quantitative findings. The fitted implied volatility surfaces evolve from a nearly planar shape in the OLS model to the realistic curvature seen in the spline and ensemble models. Residuals become progressively narrower, more symmetric, and less heteroskedastic, showing that later models capture the key nonlinear and local dynamics of moneyness and maturity. The hybrid variants produce residual plots almost indistinguishable from the pure machine learning versions, reflecting consistent behavior and strong generalization across samples.

Taken together, the results illustrate a gradual trade-off between model transparency and flexibility. Classical regression models provide clear interpretability but limited accuracy, while machine learning models achieve high precision at the expense of clarity. The hybrid approach offers a practical middle ground, retaining the spline model's interpretability while reaching the predictive performance of ensemble learning.

Chapter 6

Conclusions and Recommendations

This dissertation set out to explore how different modeling approaches can represent the structure of the implied volatility surface (IVS) of the S&P 500. The goal was not only to identify which techniques perform best, but also to understand how far one can go in improving predictive accuracy without losing interpretability or methodological coherence. This work was built on a simple, yet essential idea: although the IVS may appear complex and irregular, its shape is driven by underlying relationships that can be analyzed and captured. The challenge lies in finding a model that can describe this behavior while preserving an interpretable, coherent representation.

The study was organized around three main categories of models. The first group, built upon traditional statistical formulations (OLS, polynomial regression, and cubic B-splines), provided a controlled starting point for understanding the functional patterns of the IVS. The second group, consisting of machine learning models such as Random Forest and XGBoost, represented a new paradigm in which the functional form is not predefined but learned directly from the data, allowing nonlinear interactions to emerge naturally. Finally, the hybrid models combined both perspectives, using machine learning not as a replacement but as a complement to strengthen the explanatory power of a classical base such as the spline model. All models were developed under a common pipeline, ensuring comparability across training, tuning, and evaluation stages.

The results showed a clear and consistent progression in performance as model flexibility increased. The evolution from OLS to polynomial regression mainly served to illustrate the complexity of the IVS. The linear model acted almost as a test of the obvious, confirming that the surface is highly nonlinear and cannot be represented by a simple plane. Moving to a polynomial specification revealed that a reasonable fit could only be achieved by extending the degree up to five, which already signals the irregular and intricate nature of the IVS. At the same time, the polynomial model also revealed its own limitations, as adding additional degrees no longer improved the fit and even degraded it in some regions. This confirmed that such rigid global structures cannot fully capture the heterogeneous patterns of implied volatility.

Seeking a more flexible approach within the family of classical models, the next step was to model the surface using cubic B-splines. This specification led to a marked improvement in performance, confirming that global polynomial structures were too rigid to reproduce the

irregularities and local variations that characterize the IVS. Spline-based formulations provided a more adaptable representation, with a fitted surface already displaying shapes consistent with theoretical expectations. Nevertheless, a closer examination of the residuals revealed patterns that were not merely noise, but signs of remaining structure not captured by the model. This observation motivated the exploration of more flexible techniques, leading to the use of machine learning models as the next step in the analysis and the focus of this dissertation.

This stage marked the transition to the second group of models, which explored machine learning techniques. In this case, the two algorithms should not be seen as sequential developments but as distinct approaches within the machine learning framework. Random Forest relies on bagging, aggregating multiple trees trained on random subsets of data to reduce variance, while XGBoost follows a boosting strategy, where trees are built sequentially to correct the residuals of previous trees. Both achieved lower errors and produced residuals that were more concentrated and free from discernible patterns. The Random Forest achieved slightly better performance metrics, while XGBoost distributed explanatory power more evenly across predictors, suggesting stronger adaptability across different forecasting conditions.

Finally, the hybrid approach was explored by modeling the residuals of the spline model using each of the machine learning techniques from the previous group, following the same setup to ensure full comparability. The results were equally strong, even showing a slight improvement in some of the evaluation metrics. This finding is conceptually relevant, as it demonstrates that high predictive accuracy can be achieved while maintaining a classical and interpretable foundation. In essence, the flexibility of machine learning can serve as a complement rather than a replacement, refining the residual structure of traditional models without losing their analytical transparency.

A natural next step is to consider opportunities to strengthen the study. It is important to reflect on how the construction of the modelling table was approached, since it followed deliberately conservative choices. The cleaning rules focused on stable and information-rich regions of the IVS, which helped ensure robust estimates but also reduced the complexity of the task. An obvious extension would be to widen the scope towards more demanding areas of the surface, including very deep in or out of the money options, as well as shorter and longer maturities. Alternatively, recognizing that these regions often display distinct structural behaviors, dedicated models could be developed for those sections and later combined with the current framework, so that the entire surface benefits from specialized treatment where it is most needed.

There is also room to improve the way implied volatilities are obtained. Ideally, the dataset would include theoretical IVs computed with high-precision institutional methods. In the absence of such data, this work relied on *py_vollib*, a solid and widely used library but not an industrial standard, meaning that small numerical deviations are to be expected. Future research could complement or cross-validate these calculations using alternative procedures or data sources, reducing noise measurement and improving input reliability. Such an extension could also prevent the exclusion of a considerable number of observations that were removed due to the inability to compute valid IV values.

Beyond the aspects related to data and the construction of the modelling base, it is also worth considering methodological enhancements that could serve as natural extensions of the present work. One particularly relevant direction concerns the limited set of explanatory variables used throughout the analysis. A possible improvement would be to incorporate exogenous factors that reflect broader market or macroeconomic conditions. Indicators such as the Economic Policy Uncertainty (EPU) index or the VIX could enrich the explanatory framework, allowing the models, particularly the machine learning ones, to better exploit their flexibility and potentially improve both fit and interpretability.

Another opportunity lies in optimizing hyperparameters. Future research could adopt more systematic search procedures, such as Bayesian optimization or extended grid search, covering wider parameter ranges and testing a larger number of configurations. This could translate into small but meaningful gains in predictive accuracy and model stability.

It would also be valuable to test the proposed pipeline under different structural regimes, applying the same framework to more recent and volatile periods, such as during the COVID-19 crisis. Doing so would provide a stronger assessment of model robustness and its capacity to adapt to abrupt changes in market dynamics. Being even more ambitious, one could challenge the cross-sectional nature of the modelling strategy itself. In the current design, each trading day was treated as an independent surface, allowing the analysis to focus on the shape of the IVS at each moment but without capturing its temporal evolution. A natural extension would be to approach the problem from a time-series perspective, studying how the parameters or structural components of the IVS evolve over time and whether those dynamics can be systematically modelled.

Finally, several alternative, somewhat divergent paths remain aligned with the central concept of this dissertation and would be valuable to explore in future research. The most natural direction would be to investigate different modelling frameworks, either by replacing the spline specification with other interpretable bases for hybrid models or by testing

complementary machine learning architectures. For instance, stochastic volatility models such as Heston or SABR could be used as functional cores, serving as the classical component of a new hybrid setup where machine learning enhances the residual structure. On the other hand, more sophisticated ML architectures, including neural networks, LSTM or transformer models, could capture non-linear and temporal dependencies that go beyond the scope of the current design. Another possible direction, conceptually aligned with hybrid philosophy but more application-oriented, would be to extend the use of the modelled volatility surfaces to other financial contexts. For instance, employing the estimated IV surfaces in option pricing, hedging, or stress-testing frameworks could provide a natural way to evaluate how the empirical performance of these models translates into practical relevance. Rather than comparing their accuracy against theoretical benchmarks, such an approach would assess their effectiveness in real-world market applications, linking statistical performance with tangible decision-making value.

Overall, the results of this dissertation show that the implied volatility surface is far from random. It has a consistent, learnable structure that can be effectively modelled when classical statistical reasoning is combined with data-driven flexibility. Hybrid approaches seem to offer a promising balance between interpretability and accuracy. In this sense, the work illustrates that the future of volatility modelling may not lie in choosing between econometrics and machine learning, but in finding ways to integrate them, combining the structure and intuition of the former with the adaptability and power of the latter.

References

- 01-03—OCC Annual 2024 and December 2024 Volume. (2024).
<https://www.theocc.com/newsroom/views/2025/01-03-occ-annual-2024-and-december-2024-volume>
- Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13(10), 281–305.
- Black, F., & Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81(3), 637–654.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.
<https://doi.org/10.1007/BF00058655>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
<https://doi.org/10.1023/A:1010933404324>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
<https://doi.org/10.1145/2939672.2939785>
- Cont, R., & Da Fonseca, J. (2002). Dynamics of implied volatility surfaces. *Quantitative finance*, 2(1), 45.
- Dufays, A., Jacobs, K., & Rombouts, J. (2025). *A Framework for Real-Time Modeling and Forecasting of Large Unbalanced Option Implied Volatility Surfaces* (SSRN Scholarly Paper No. 5275880). Social Science Research Network. <https://doi.org/10.2139/ssrn.5275880>
- Dumas, B., Fleming, J., & Whaley, R. E. (1998). Implied Volatility Functions: Empirical Tests. *The Journal of Finance*, 53(6), 2059–2106.
- Fan, Q., & Feng, S. (2022). An empirical study on the characterization of implied volatility and pricing in the Chinese option market. *Finance Research Letters*, 49, 103160.
<https://doi.org/10.1016/j.frl.2022.103160>
- Fengler, M. R. (2005). *Semiparametric Modeling of Implied Volatility*. Springer-Verlag.
<https://doi.org/10.1007/3-540-30591-2>

- Fengler, M. R. (2009). Arbitrage-free smoothing of the implied volatility surface. *Quantitative Finance*, 9(4), 417–428. <https://doi.org/10.1080/14697680802595585>
- 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>
- Greene, W. H. (2018). *Econometric Analysis*. Pearson.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. <https://doi.org/10.1007/978-0-387-84858-7>
- Heston, S. L. (1993). A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *The Review of Financial Studies*, 6(2), 327–343.
- Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 832–844. <https://doi.org/10.1109/34.709601>
- Kristjanpoller, W., & Minutolo, M. C. (2018). A hybrid volatility forecasting framework integrating GARCH, artificial neural network, technical analysis and principal components analysis. *Expert Systems with Applications*, 109, 1–11. <https://doi.org/10.1016/j.eswa.2018.05.011>
- Latané, H. A., & Rendleman, R. J. (1976). STANDARD DEVIATIONS OF STOCK PRICE RATIOS IMPLIED IN OPTION PRICES. *The Journal of Finance*, 31(2), 369–381. <https://doi.org/10.1111/j.1540-6261.1976.tb01892.x>
- Lent, M. A. Y. van. (2023). *Predicting option implied volatility features using machine learning models: A comparative study with traditional implied volatility models*. <https://thesis.eur.nl/pub/67130>
- Li, K. (2021). The effect of option trading. *Financial Innovation*, 7(1), 65. <https://doi.org/10.1186/s40854-021-00279-5>
- Liu, C. (2024). (PDF) Implied Volatility Forecasting for American Options Based on Random Forest Regressor, Linear Regression Model. *ResearchGate*. https://www.researchgate.net/publication/380931219_Implied_Volatility_Forecasting_for_American_Options_Based_on_Random_Forest_Regressor_Linear_Regression_Model

- Liu, D., Liang, Y., Zhang, L., Lung, P., & Ullah, R. (2021). Implied volatility forecast and option trading strategy. *International Review of Economics & Finance*, 71, 943–954. <https://doi.org/10.1016/j.iref.2020.10.023>
- Luong, C., & Dokuchaev, N. (2018). Forecasting of Realised Volatility with the Random Forests Algorithm. *Journal of Risk and Financial Management*, 11(4), 61. <https://doi.org/10.3390/jrfm11040061>
- Maingo, I., Ravele, T., & Sigauke, C. (2025). A Fusion of Statistical and Machine Learning Methods: GARCH-XGBoost for Improved Volatility Modelling of the JSE Top40 Index. *International Journal of Financial Studies*, 13(3), 155. <https://doi.org/10.3390/ijfs13030155>
- Olsen, A., Djupskås, G., de Lange, P. E., & Rissstad, M. (2024). Forecasting implied volatilities of currency options with machine learning techniques and econometrics models. *International Journal of Data Science and Analytics*, 20(2), 1329–1347. <https://doi.org/10.1007/s41060-024-00528-7>
- Peña, I., Rubio, G., & Serna, G. (1999). Why do we smile? On the determinants of the implied volatility function. *Journal of Banking & Finance*, 23(8), 1151–1179. [https://doi.org/10.1016/S0378-4266\(98\)00134-4](https://doi.org/10.1016/S0378-4266(98)00134-4)
- Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, 9(3), e1301. <https://doi.org/10.1002/widm.1301>
- Teller, A., Pigorsch, U., & Pigorsch, C. (2022). *Short- to Long-Term Realized Volatility Forecasting using Extreme Gradient Boosting* (SSRN Scholarly Paper No. 4267541). Social Science Research Network. <https://doi.org/10.2139/ssrn.4267541>
- Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159–175. [https://doi.org/10.1016/S0925-2312\(01\)00702-](https://doi.org/10.1016/S0925-2312(01)00702-)

Appendices

Appendix A - Evaluation Metrics

The evaluation metrics presented in Table 4.3 of Chapter 4.5 are formally defined below. These measures quantify model performance in terms of error magnitude, explanatory strength, and systematic bias. In the context of this appendix, y_i and $\hat{y}_i$ denote the observed and predicted implied volatilities, respectively, while $\bar{y}$ represents their sample mean.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (A.1)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (A.2)$$

$$MedAE = \text{median}(|\hat{y}_i - y_i|) \quad (A.3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (A.4)$$

$$Bias = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i) \quad (A.5)$$

The performance evaluation also considered two derived error ratios, the RMSE/MAE and MAE/MedAE, which provide additional information about the distributional properties of model residuals. Both ratios have theoretical foundations derived under the assumption of normally distributed, zero-mean residuals, and their reference values are supported by empirical evidence in literature.

The first ratio compares the root mean squared error (RMSE) with the mean absolute error (MAE):

$$\frac{RMSE}{MAE} = \frac{\sqrt{E[\varepsilon^2]}}{E[|\varepsilon|]}, \quad (A.6)$$

where $\varepsilon_i = \hat{y}_i - y_i$ denotes the model errors.

If the residuals follow a normal distribution, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, the two expectations admit analytical expressions:

$$RMSE = \sqrt{E[\varepsilon^2]} = \sigma, \text{ and } MAE = E[|\varepsilon|] = \sigma \sqrt{\frac{2}{\pi}}.$$

The theoretical ratio is therefore $\frac{RMSE}{MAE} = \sqrt{\frac{\pi}{2}} \approx 1.253$.

This value corresponds to the case of normally distributed errors with moderate tails. The ratio increases when large errors have a stronger impact, for example when the distribution of residuals is heavy-tailed or affected by outliers, and decreases when errors are more homogeneous. In this sense, $RMSE/MAE$ can be interpreted as a practical proxy for the kurtosis of the error distribution: higher values reflect fatter tails, while ratios close to unity suggest smoother or underfitted models.

The second derived ratio, $MAE/MedAE$, compares the mean and median of the absolute errors and serves as an indicator of asymmetry and robustness of the error distribution:

$$\frac{MAE}{MedAE} = \frac{E[|\varepsilon|]}{\text{median}(|\varepsilon|)}. \quad (\text{A.7})$$

Under the same normality assumption, the expected mean absolute error is $MAE = \sigma\sqrt{2/\pi} \approx 0.7979\sigma$, and the median absolute error is $MedAE = 0.6745\sigma$, since 0.6745 corresponds to the 75th percentile of the absolute value of a standard normal variable. The resulting theoretical value is $\frac{MAE}{MedAE} = \frac{0.7979\sigma}{0.674\sigma} \approx 1.183$.

This reference value is valid for symmetric and light-tailed error distributions. When the distribution becomes more asymmetric or exhibits heavier tails, the mean absolute error increases relative to the median, raising the ratio above the theoretical level. Empirical evidence suggests that for near-normal residuals, the $MAE/MedAE$ ratio typically falls between 1.2 and 1.3, while values exceeding 1.4 indicate a stronger influence of outliers or pronounced tail behavior. Conversely, ratios close to 1.0–1.1 are characteristic of very concentrated residuals, potentially resulting from over-smoothed or underfitted models.

Together, these two ratios provide complementary diagnostic information: $RMSE/MAE$ reflects the tail weight of the error distribution, while $MAE/MedAE$ captures its asymmetry and robustness.

Appendix B – Complementary Graphical Analysis

This annex presents graphical results for each model in Chapter 5, including the fitted total variance surface, observed versus predicted implied volatility and residual analyses (histogram, versus log-moneyness and versus maturity). These visualizations complement the quantitative metrics, offering an intuitive view of model performance and residual behavior.

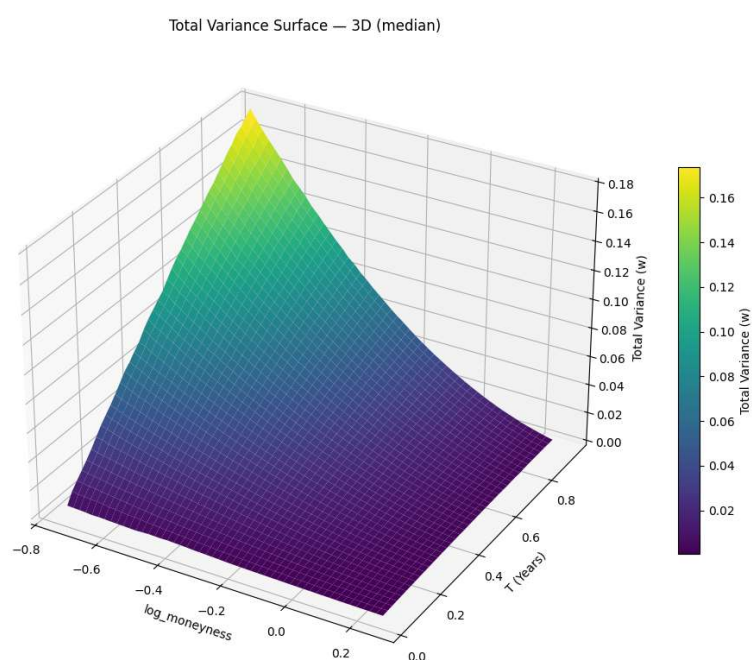


Figure B.1 - Total variance surface (3D view) — OLS Model.

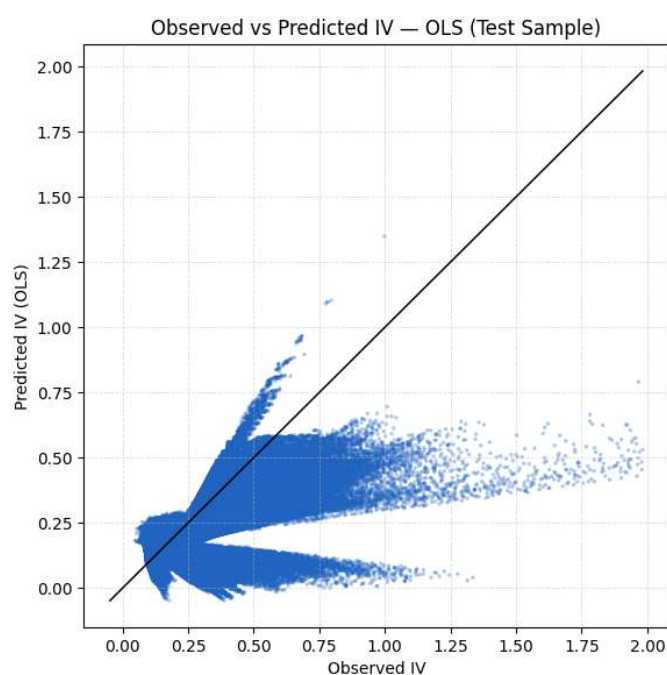


Figure B.2 - Observed vs predicted implied volatility — OLS Model.

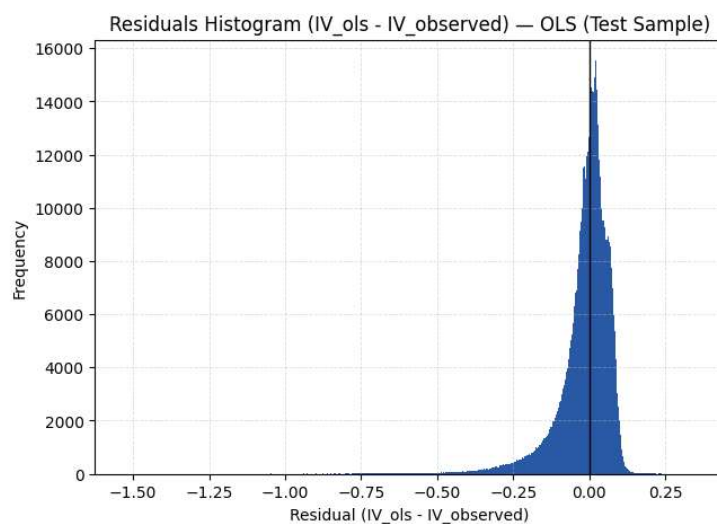


Figure B.3 - Residuals histogram — OLS Model.

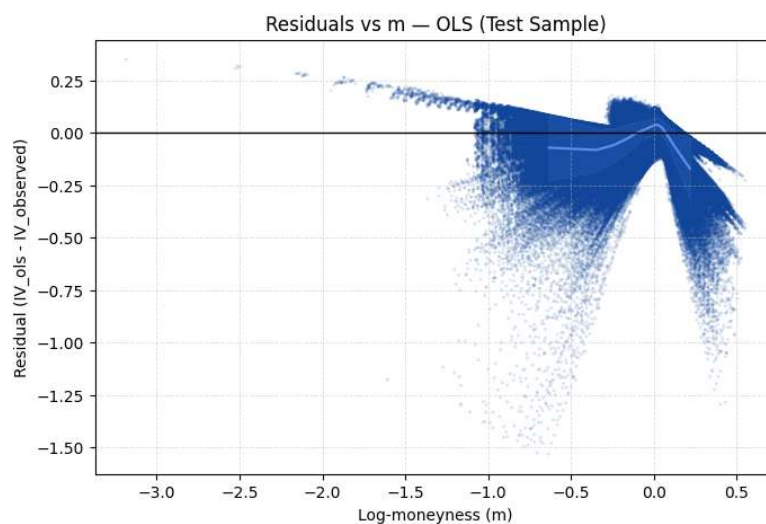


Figure B.4 - Residuals vs log-moneyness – OLS Model.

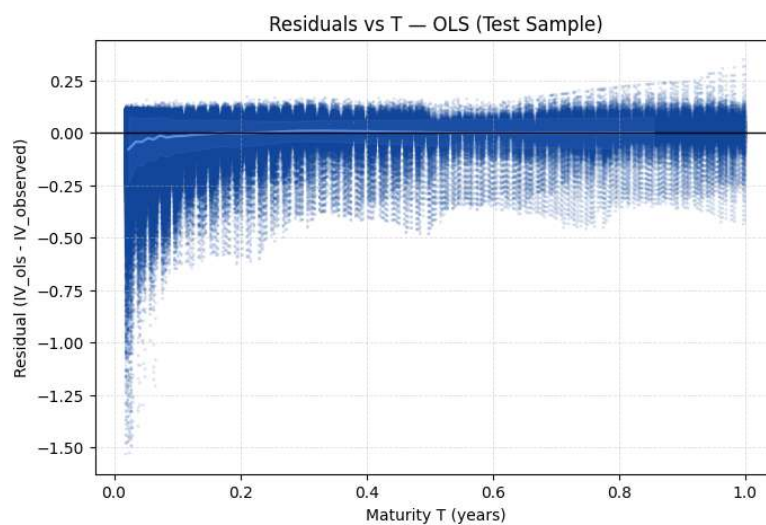


Figure B.5 - Residuals vs maturity (T) — OLS Model.

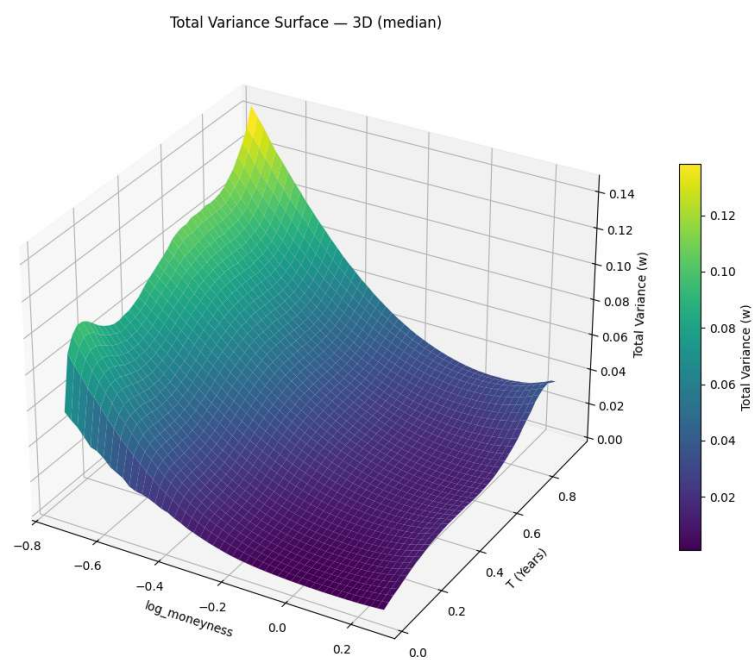


Figure B.6 - Total variance surface (3D view) — Polynomial Model.

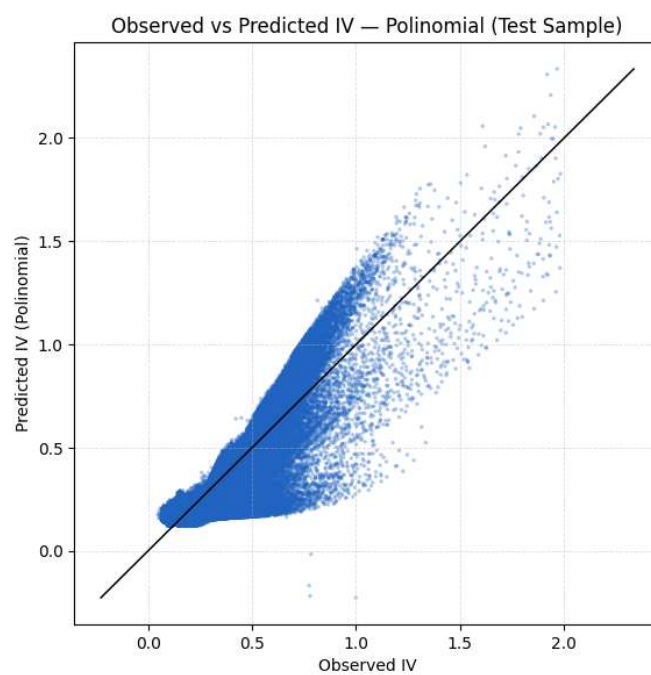


Figure B.7 - Observed vs predicted implied volatility — Polynomial Model.

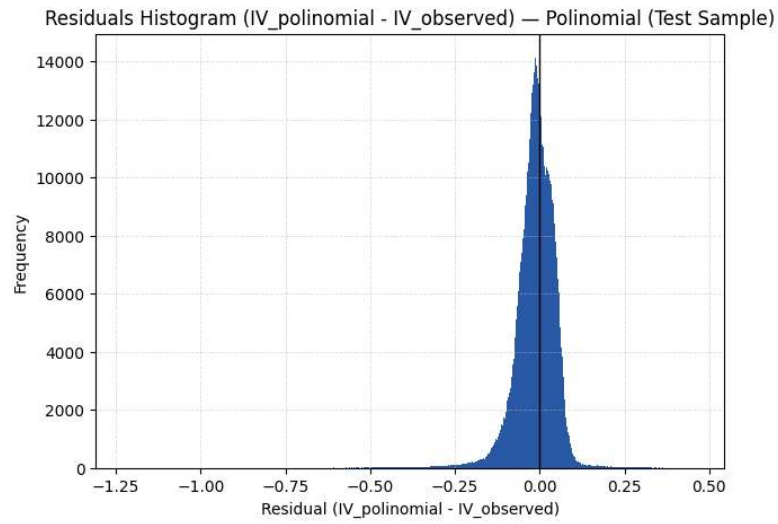


Figure B.8 - Residuals histogram — Polynomial Model.

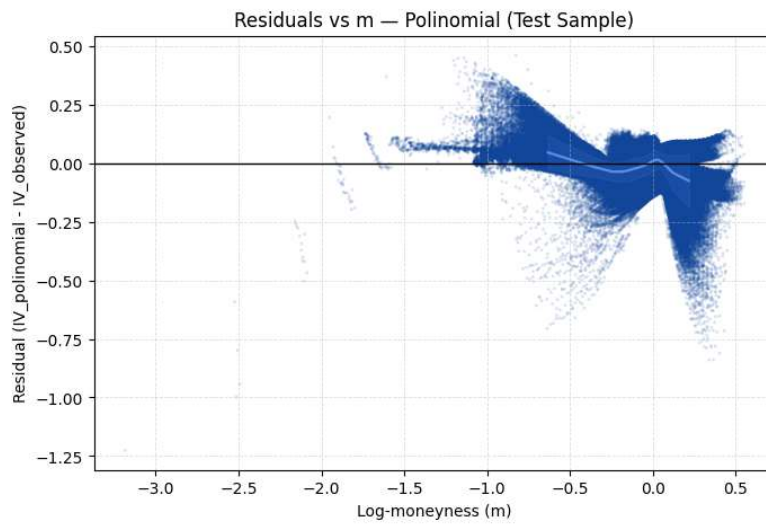


Figure B.9 - Residuals vs log-moneyness — Polynomial Model.

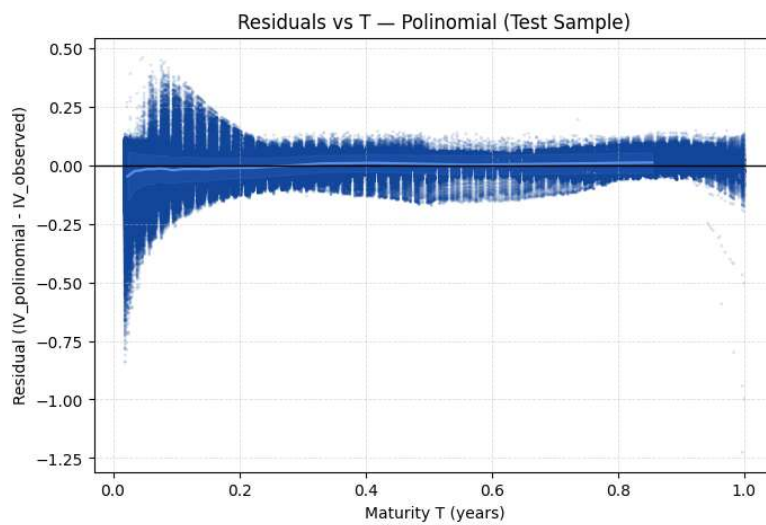


Figure B.10 - Residuals vs maturity (T) — Polynomial Model.

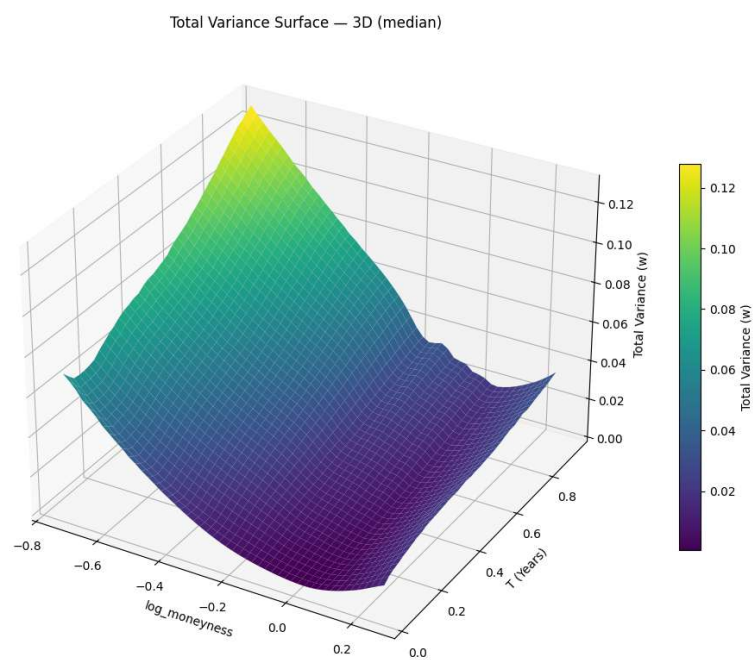


Figure B.11 - Total variance surface (3D view) — Splines Model.

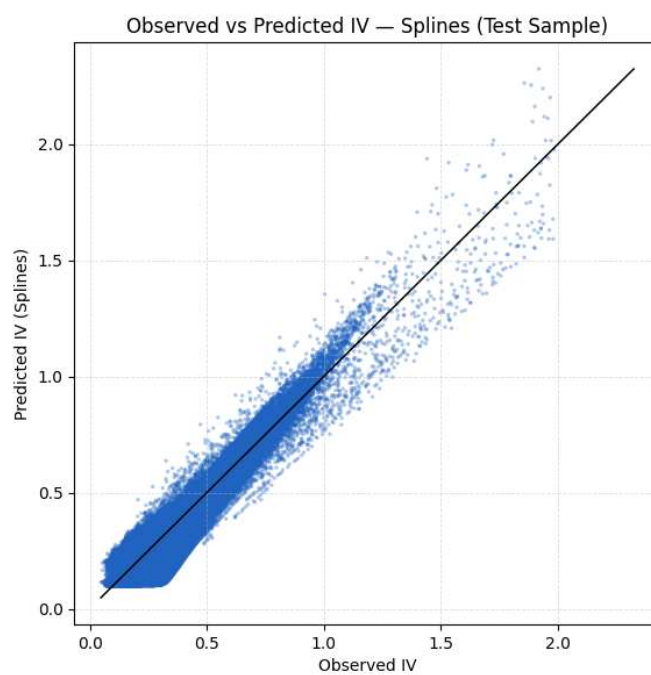


Figure B.12 - Observed vs predicted implied volatility — Splines Model.

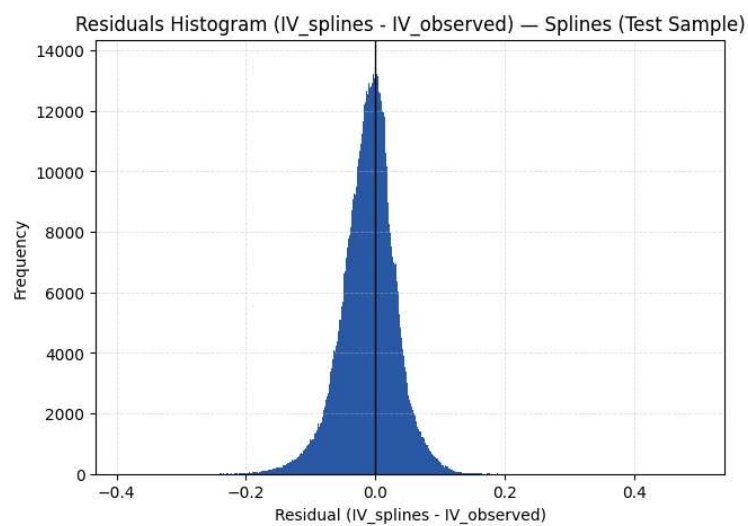


Figure B.13 - Residuals histogram — Splines Model.

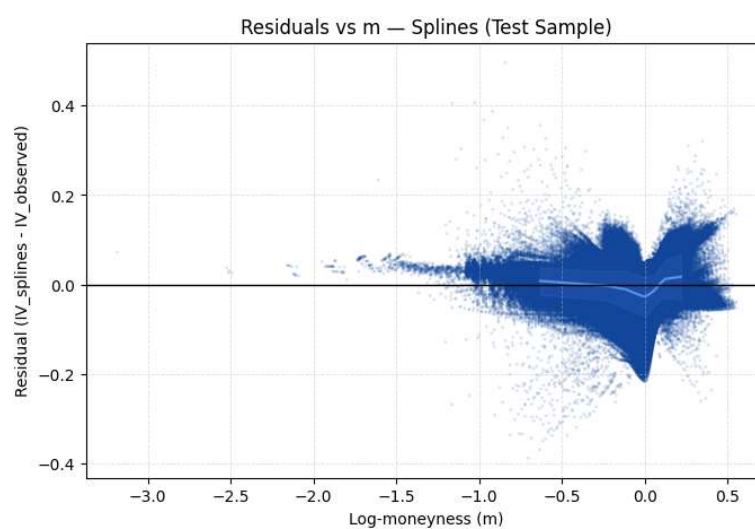


Figure B.14 - Residuals vs log-moneyness – Splines Model.

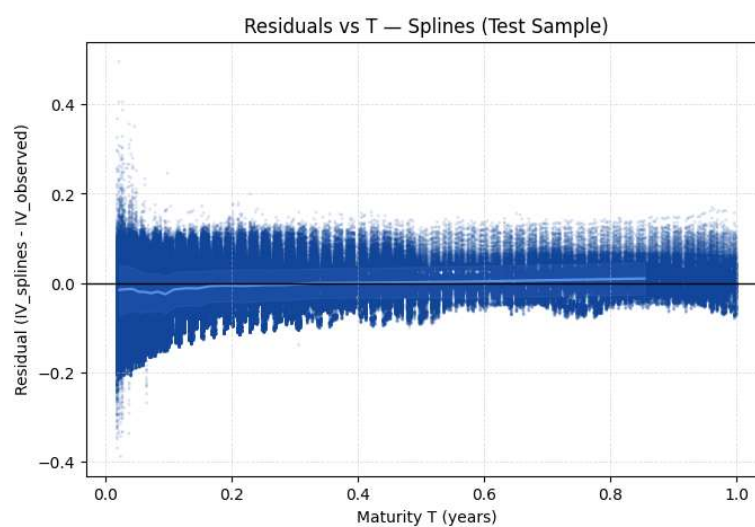


Figure B.15 - Residuals vs maturity (T) — Splines Model.

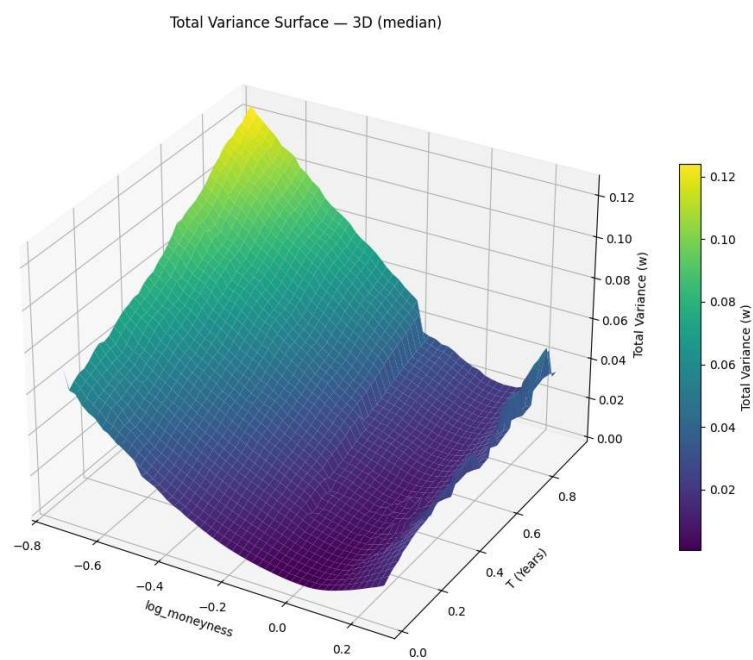


Figure B.16 - Total variance surface (3D view) — Random Forest Model.

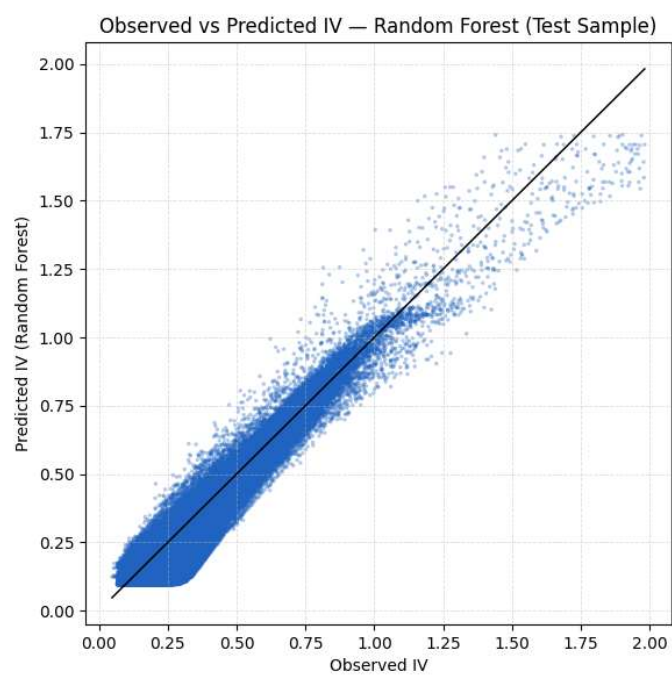


Figure B.17 - Observed vs predicted implied volatility — Random Forest Model.

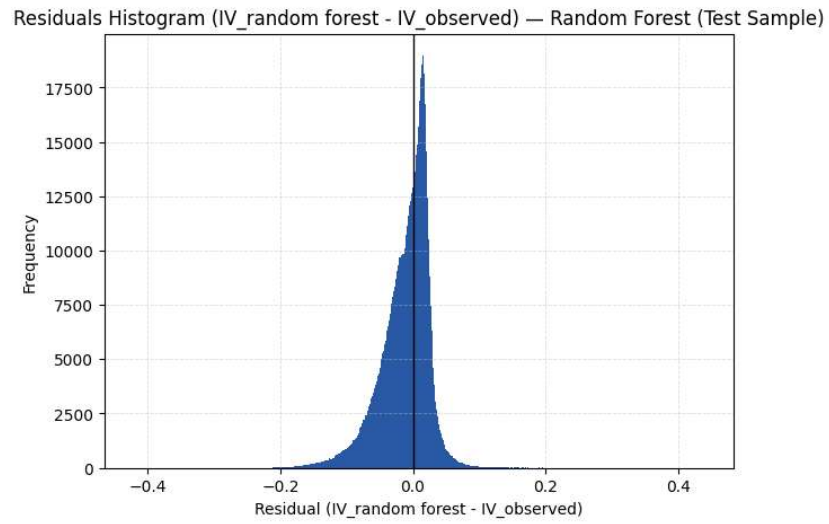


Figure B.18 - Residuals histogram — Random Forest Model.

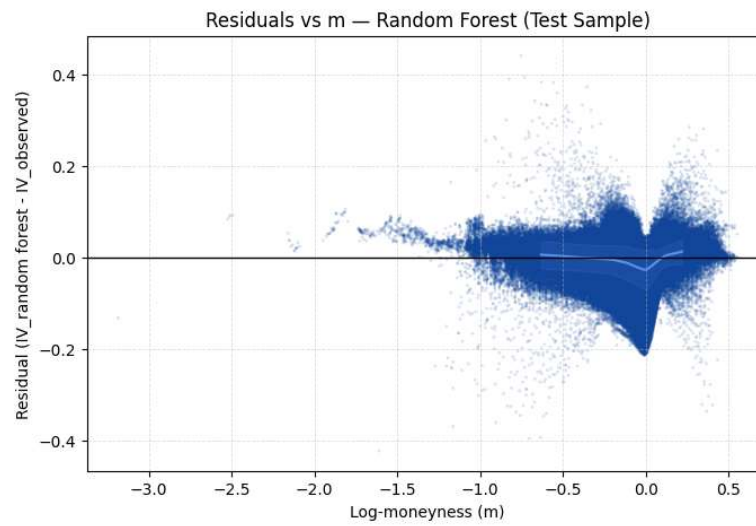


Figure B.19 - Residuals vs log-moneyness — Random Forest Model.

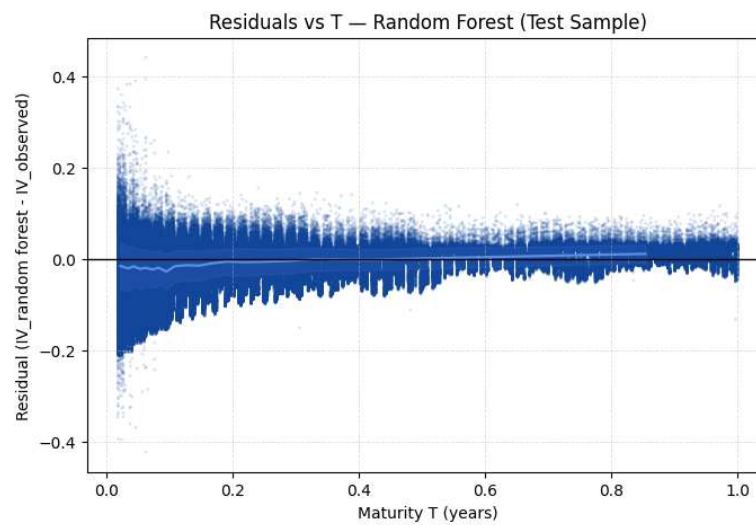


Figure B.20 - Residuals vs maturity (T) — Random Forest Model.

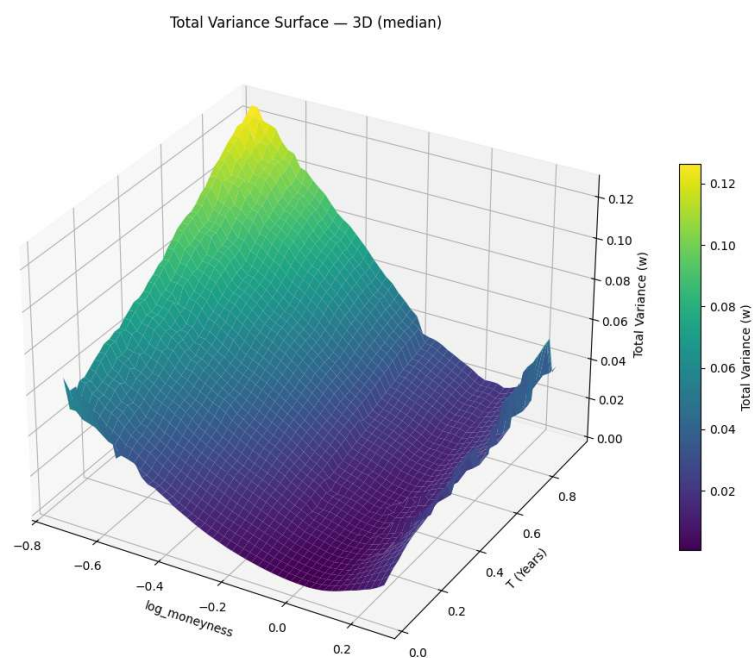


Figure B.21 - Total variance surface (3D view) — XGBoost Model.

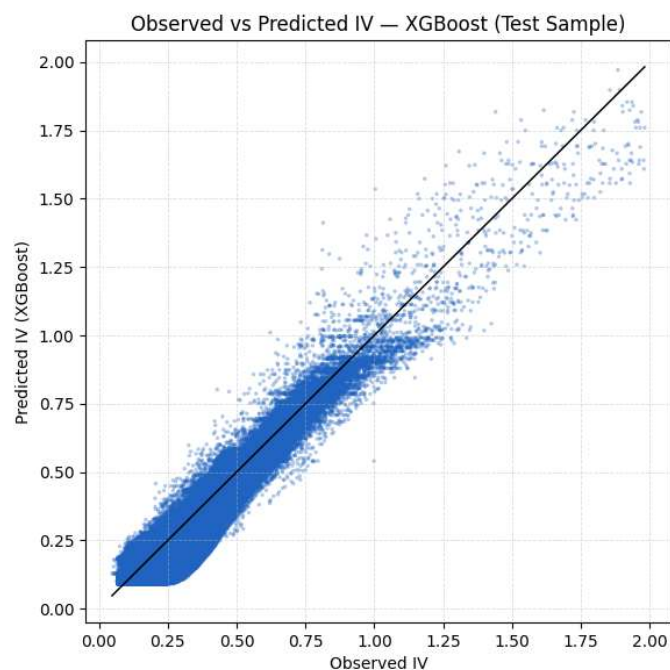


Figure B.22 - Observed vs predicted implied volatility — XGBoost Model.

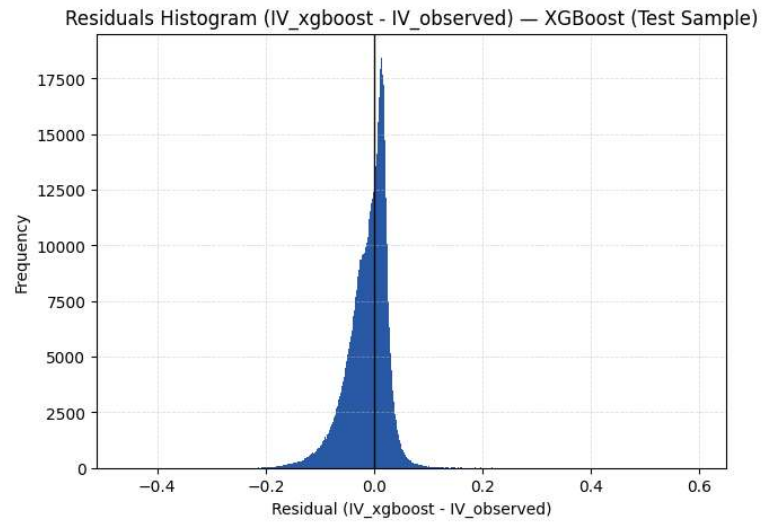


Figure B.23 - Residuals histogram — XGBoost Model.

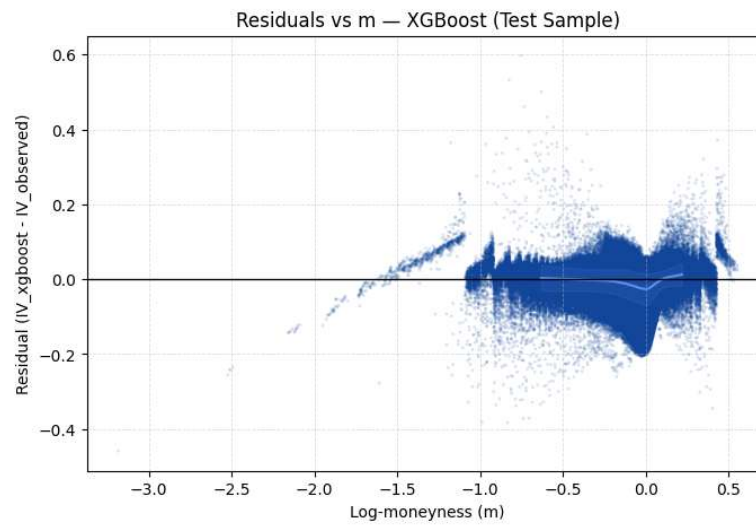


Figure B.24 - Residuals vs log-moneyness — XGBoost Model.

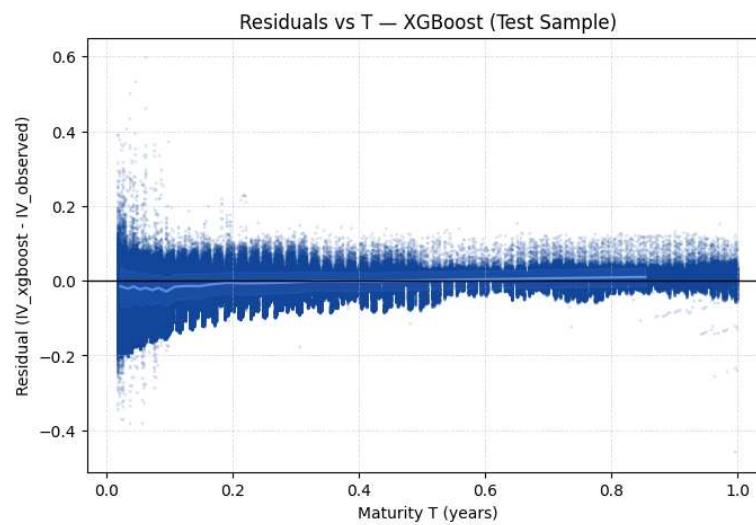


Figure B.25 - Residuals vs maturity (T) — XGBoost Model.

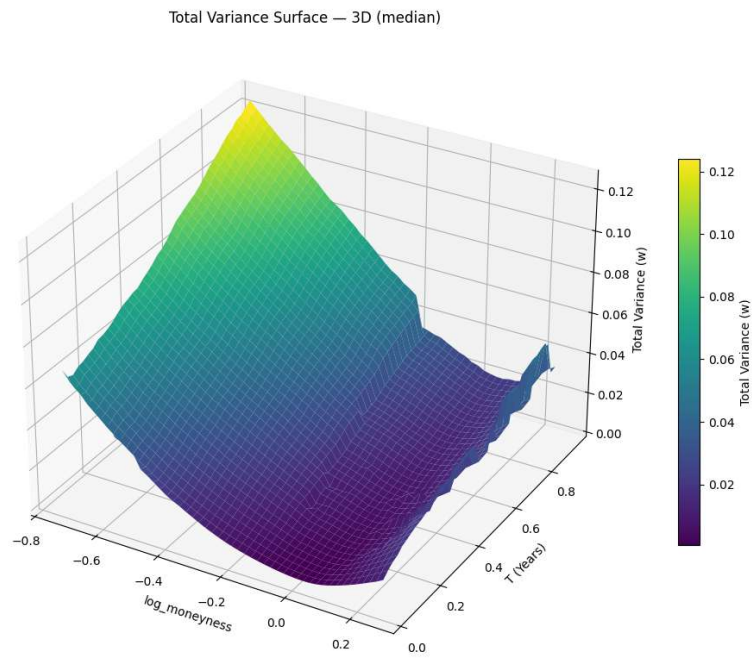


Figure B.26 - Total variance surface (3D view) — Hybrid Random Forest Model.

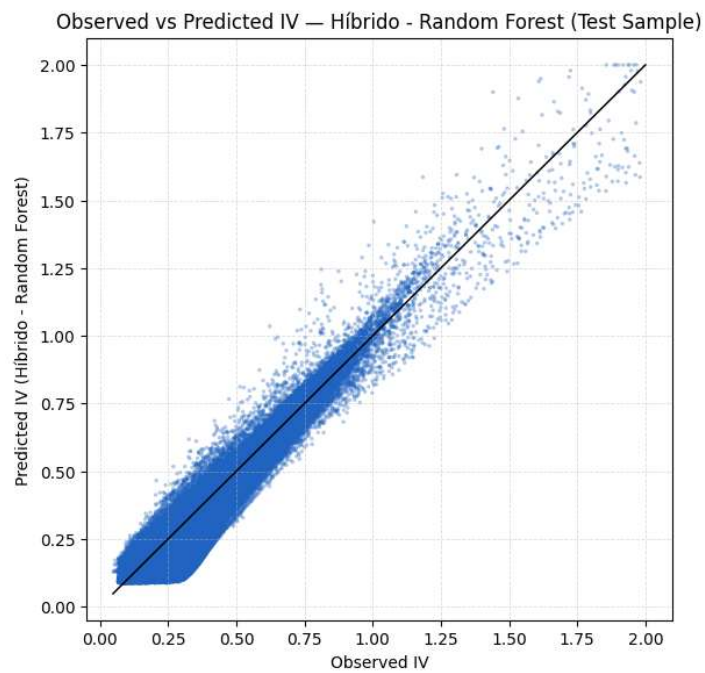


Figure B.27 - Observed vs predicted implied volatility — Hybrid Random Forest Model.

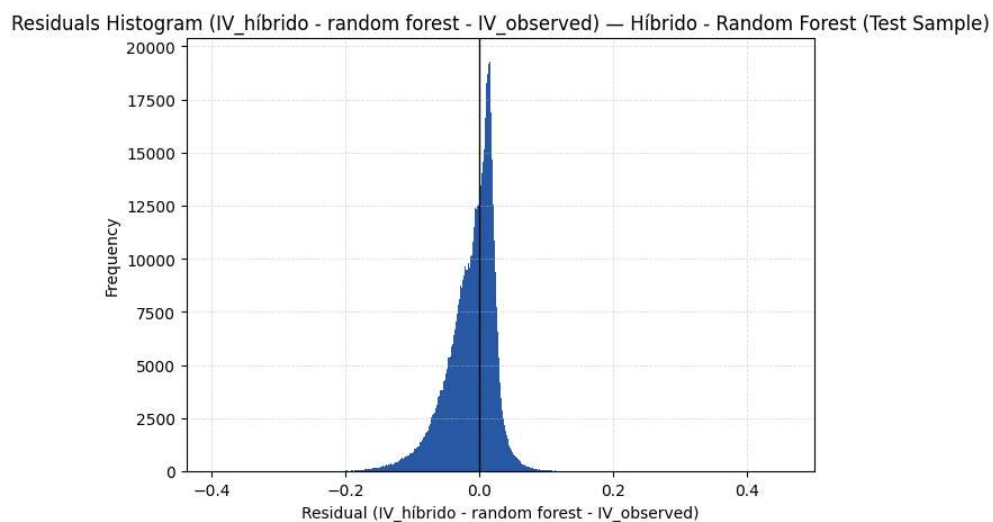


Figure B.28 - Residuals histogram — Hybrid Random Forest Model.

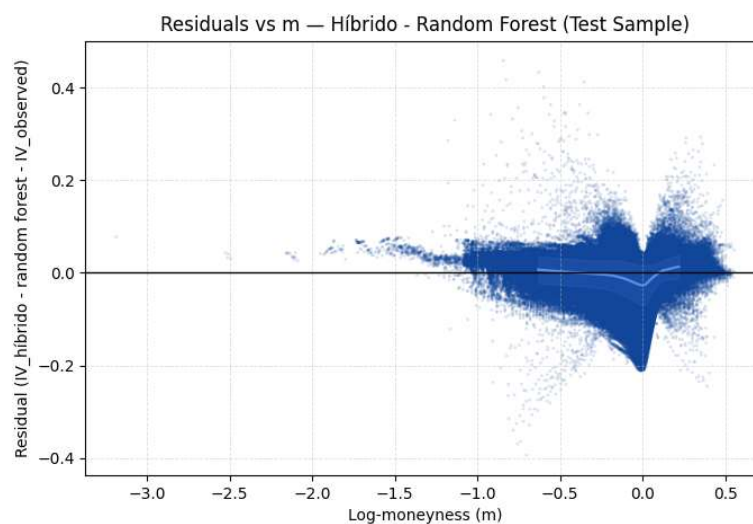


Figure B.29 - Residuals vs log-moneyness — Hybrid Random Forest Model.

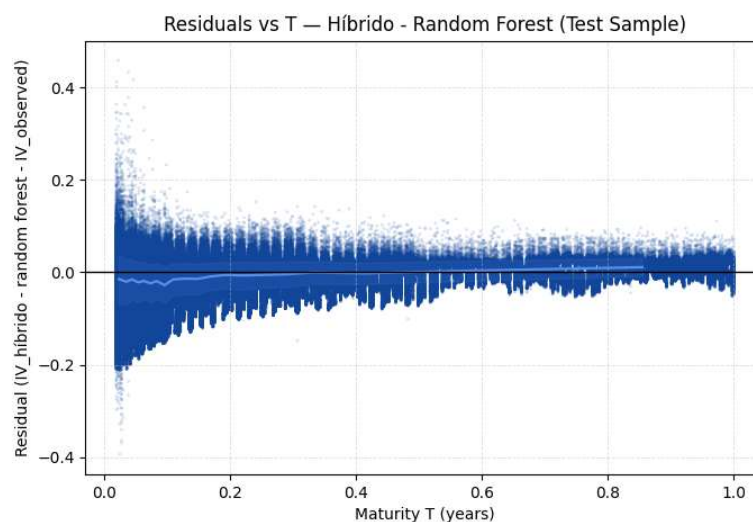


Figure B.30 - Residuals vs maturity (T) — Hybrid Random Forest Model.

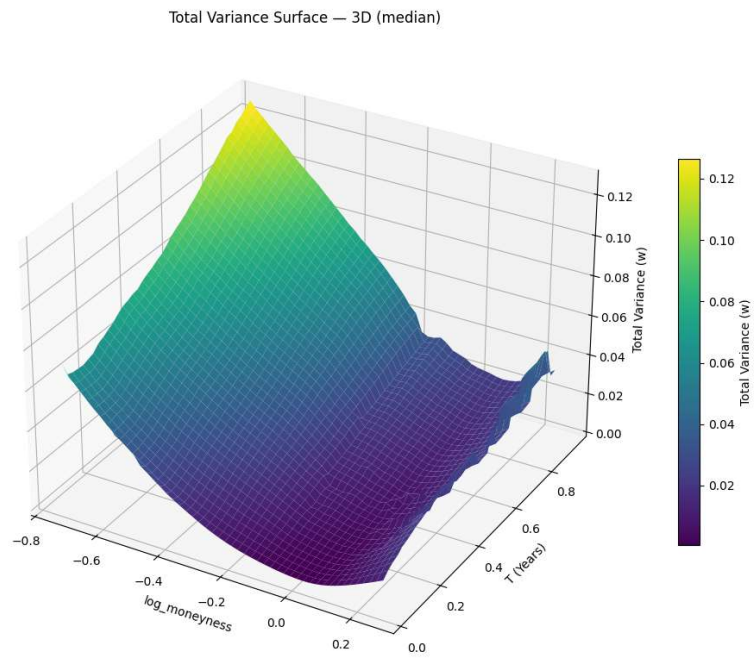


Figure B.31 - Total variance surface (3D view) — Hybrid XGBoost Model.

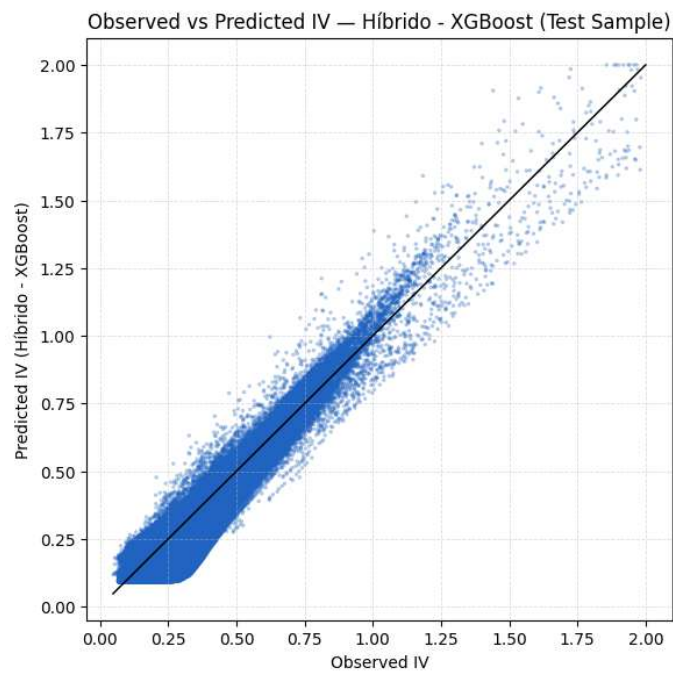


Figure B.32 - Observed vs predicted implied volatility — Hybrid XGBoost Model.

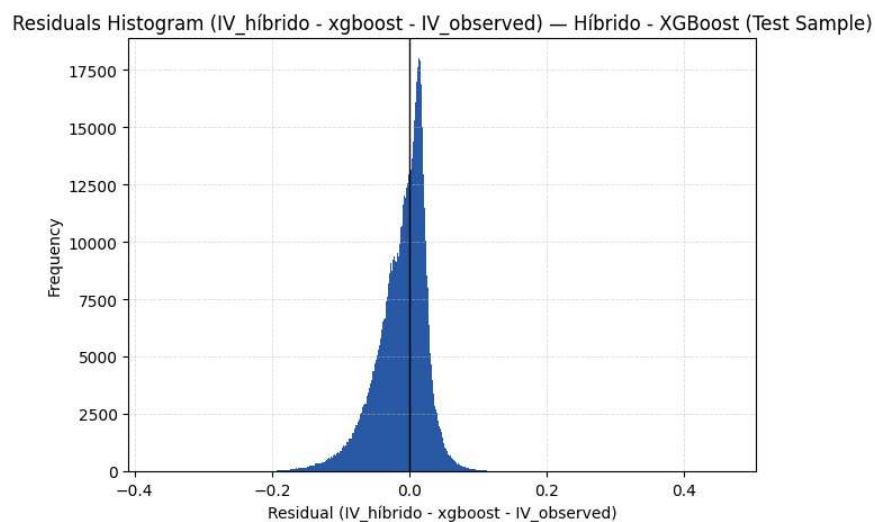


Figure B.33 - Residuals histogram — Hybrid XGBoost Model.

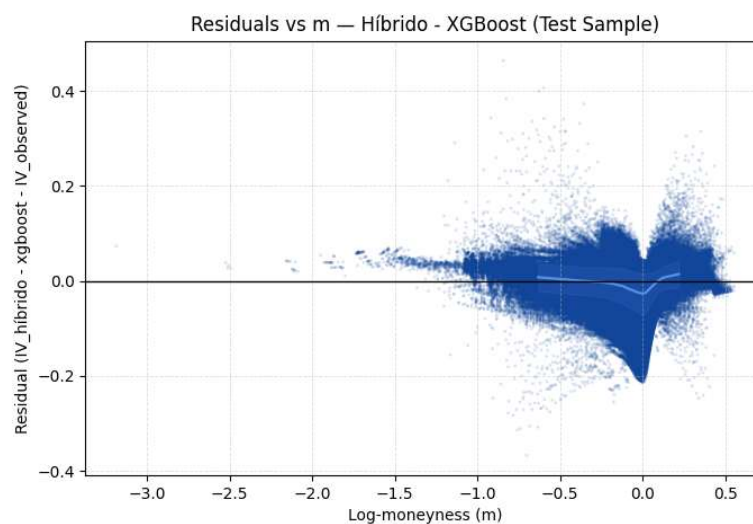


Figure B.34 - Residuals vs log-moneyness — Hybrid XGBoost Model.

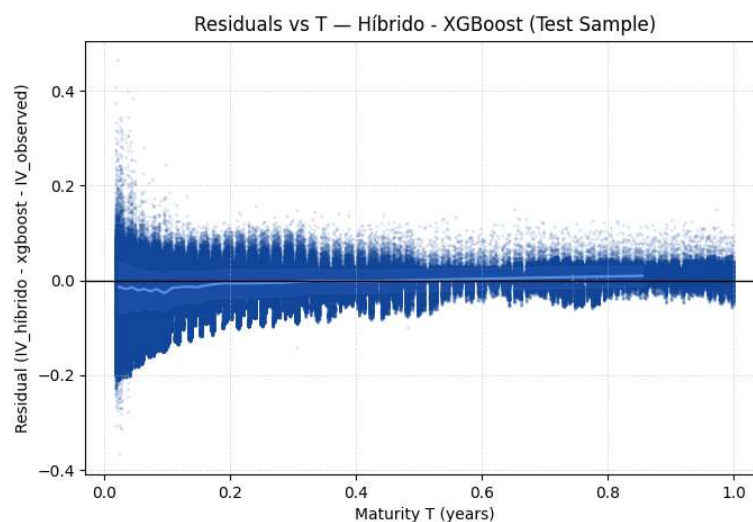


Figure B.35 - Residuals vs maturity (T) — Hybrid XGBoost Model.

Appendix C - Polynomial Regression Terms and Respective Coefficients

This appendix reports the complete set of polynomial terms and their corresponding normalized coefficients estimated for the fifth-degree polynomial regression model with RidgeCV ($\alpha^* = 0.001$).

The coefficients are ranked by absolute magnitude to facilitate interpretation of the relative importance of each term in shaping the implied volatility surface.

Table C.1 - Normalized coefficients of the fifth-degree polynomial regression model estimated with RidgeCV ($\alpha^ = 0.001$).*

The table reports the 20 polynomial terms ranked by absolute normalized coefficient.

Term	Normalized Coefficient (β)	Weight (%)
T^3	-26.2691	17.32
T^4	25.2484	16.65
$\log_moneyness \times T^3$	16.2888	10.74
$\log_moneyness \times T^2$	-13.7067	9.04
T^2	12.1682	8.02
T^5	-8.9591	5.91
$\log_moneyness^2 \times T^2$	8.2490	5.44
$\log_moneyness^2 \times T$	-8.1474	5.37
$\log_moneyness \times T^4$	-6.7261	4.44
$\log_moneyness \times T$	4.8812	3.22
$\log_moneyness^3 \times T$	-4.1971	2.77
$\log_moneyness^2$	3.8955	2.57
$\log_moneyness^3$	3.0932	2.04
$\log_moneyness^2 \times T^3$	-2.9313	1.93
T	-2.2056	1.45
$\log_moneyness^3 \times T^2$	1.9409	1.28
$\log_moneyness$	-1.0613	0.70
$\log_moneyness^4 \times T$	-0.7255	0.48
$\log_moneyness^4$	0.7241	0.48
$\log_moneyness^5$	-0.2119	0.14

Appendix D – Cubic B-Splines Terms and Respective Coefficients

This appendix reports the complete set cubic b-splines terms and their corresponding normalized coefficients estimated (RidgeCV ($\alpha^* = 0.01$)). The 65 coefficients are ranked by absolute magnitude to facilitate interpretation of the relative importance of each term in shaping the implied volatility surface.

Table D.1 - Normalized coefficients of the cubic B-spline regression model (degree = 3, $\alpha^ = 0.01$). The table lists the 65 spline basis terms ranked by absolute normalized coefficient, with Bm and BT denoting basis functions in moneyness and maturity, respectively.*

Term	Normalized Coefficient (β)	Weight (%)	Term	Normalized Coefficient (β)	Weight (%)
Bm2×BT0	+92.4555	16.82	Bm5×BT4	+1.5530	0.28
Bm1×BT0	-65.2275	11.86	Bm5×BT1	+1.4722	0.27
Bm1×BT1	+61.0031	11.10	Bm3×BT3	+1.4317	0.26
Bm0×BT3	+43.9439	7.99	Bm11×BT4	+1.3770	0.25
Bm11×BT0	+29.2769	5.32	Bm11×BT3	+1.3512	0.25
Bm0×BT4	+27.0997	4.93	Bm7×BT4	+1.2299	0.22
Bm1×BT2	+24.5392	4.46	Bm4×BT4	+1.2224	0.22
Bm12×BT1	+19.2675	3.50	Bm5×BT2	+1.0752	0.20
Bm12×BT0	+16.1635	2.94	Bm9×BT0	+1.0653	0.19
Bm2×BT1	+15.2823	2.78	Bm8×BT0	+1.0365	0.19
Bm3×BT0	+13.3240	2.42	Bm6×BT4	+1.0320	0.19
Bm12×BT2	+12.4981	2.27	Bm10×BT4	+1.0248	0.19
Bm12×BT3	+9.5883	1.74	Bm9×BT4	+1.0193	0.19
Bm12×BT4	+9.4923	1.73	Bm8×BT4	+0.9944	0.18
Bm11×BT1	+8.9808	1.63	Bm6×BT3	+0.9897	0.18
Bm1×BT3	+8.3483	1.52	Bm5×BT3	+0.9162	0.17
Bm4×BT0	+7.6284	1.39	Bm9×BT3	+0.8981	0.16
Bm2×BT2	+7.4464	1.35	Bm6×BT2	+0.8913	0.16
Bm1×BT4	+5.7481	1.05	Bm7×BT3	+0.8893	0.16
Bm2×BT4	+5.2223	0.95	Bm4×BT3	+0.8644	0.16
Bm11×BT2	+5.1345	0.93	Bm6×BT1	+0.8611	0.16
Bm5×BT0	+5.0046	0.91	Bm7×BT2	+0.7636	0.14
Bm10×BT0	+4.5443	0.83	Bm8×BT3	+0.7631	0.14
Bm6×BT0	+4.2462	0.77	Bm0×BT2	+0.7557	0.14
Bm2×BT3	+2.8648	0.52	Bm8×BT2	+0.5721	0.10
Bm3×BT1	+2.5760	0.47	Bm10×BT1	+0.5620	0.10
Bm7×BT0	+2.2951	0.42	Bm7×BT1	+0.5598	0.10
Bm0×BT0	-2.2837	0.42	Bm10×BT2	+0.4313	0.08
Bm0×BT1	-2.1350	0.39	Bm9×BT2	+0.4178	0.08
Bm3×BT2	+2.0659	0.38	Bm10×BT3	+0.3583	0.07
Bm4×BT1	+1.9192	0.35	Bm8×BT1	+0.3212	0.06
Bm3×BT4	+1.7998	0.33	Bm9×BT1	+0.1383	0.03
Bm4×BT2	+1.5631	0.28			