Predictive Modeling of Brazilian Basis Risk: Insights from Time Series and ML Analysis

Introduction:

In financial markets, effective risk management is a foundational element for success and stability. Basis risk, as a continual and distinctive challenge faced by traders and investors, arises from variations between spot market and futures market prices. This introduces a persistent uncertainty, even in the presence of hedging strategies. Addressing this residual risk requires a nuanced understanding of market dynamics and the application of advanced predictive modeling techniques.

This project serves as an exploration of the application of data science in financial risk management. Leveraging time series modeling, our aim is to analyze and predict the basis series, fostering a proactive approach to managing this intricate financial risk.

By utilizing historical data, conducting a comprehensive exploratory data analysis, and deploying advanced time series models and machine learning techniques, this project seeks to unravel the complex dynamics of basis risk. The predictive insights derived from these models hold the potential to refine risk management strategies, equip decision-makers with informed choices, and fortify the overall resilience of financial entities navigating unpredictable markets.

In the subsequent sections, we will delve into the key components, methodologies, and tools that form the foundation of this project. Through addressing challenges, considering external factors, and exploring the potential impact of our findings, our goal is to contribute to the evolving landscape of financial risk management using the lenses of predictive analytics and time series modeling.

Data and Methodology:

Data Source:
The primary focus of this analysis revolves around the Brazilian basis risk, chosen for its liquidity, market activity, and the significant role Brazil plays as the world's largest producer and exporter of coffee. To conduct this study, we utilized the CEPEA-ESALQ Brazilian spot coffee price and ICE coffee futures contract data. The dataset spans from September 02, 1996, to August 08, 2023, providing a robust timeframe for comprehensive analysis.

Data Transformation:
Brazilian spot coffee prices are denominated in Brazilian Reais (BRL) per 60-kilogram bag, while Coffee futures are quoted in US$ cents per pound. To align these two disparate metrics, an initial step involved transforming the Brazilian spot coffee prices into cents per pound. This transformation required considering the USD/BRL exchange rate and converting US$/60-kilogram bags into cents per pound. The conversion coefficient was calculated using the formula:

Conversion Coefficient = $\left(\frac{Brazilian\ Coffee\ Price \times 100}{60} \right) \div\ 2.20462$

This conversion ensured a standardized unit of measurement for both the Brazilian spot coffee price and the Coffee futures.

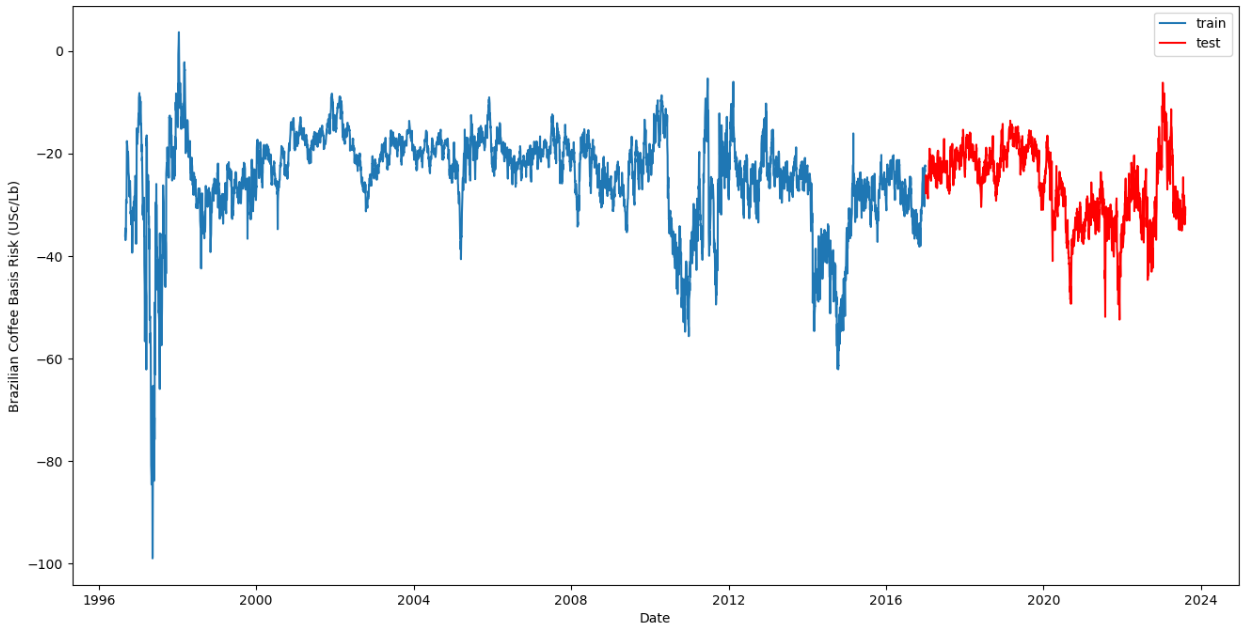
Modeling Approaches:
To comprehend the dynamic and intricate nature of basis risk, we employed three distinct models:

1. ARIMA Model:
   - Autoregressive Integrated Moving Average (ARIMA) is a time series forecasting model designed to discern patterns and trends within sequential data. In our analysis, the ARIMA model's input consisted solely of the Brazilian coffee basis series. This univariate approach aimed to uncover inherent temporal dependencies and structures within the basis risk.

2. XGBoost Model:
   - XGBoost, renowned for its effectiveness in regression tasks, was employed with a multivariate input structure. The input for the XGBoost model included: the Brazilian coffee basis series, the Brazilian spot coffee prices (denominated in Brazilian Reais per 60-kilogram bag), the Brazilian adjusted coffee prices (transformed to US cents per pound), the USD/BRL exchange rate, and the Coffee futures series.
This multivariate input allowed the XGBoost model to consider complex interactions and dependencies among various factors influencing the Brazilian basis risk.

3. Facebook's Prophet Model:
   - Prophet, developed by Facebook, is designed for forecasting time series data with strong seasonal patterns. By employing the Prophet model, we sought to uncover seasonality and capture the underlying patterns in the Brazilian basis risk.

Data Splitting:
Prophet, designed for time series forecasting with strong seasonal patterns, utilized the Brazilian coffee basis series as its sole input. This univariate approach aimed to leverage Prophet's capabilities in capturing seasonality and temporal patterns within the basis risk.



Each model was independently applied to the dataset, allowing for a holistic understanding of the basis risk dynamics from various perspectives. In the subsequent sections, we will delve into the outcomes and insights derived from each model, evaluating their respective performances and contributions to the analysis of Brazilian basis risk.

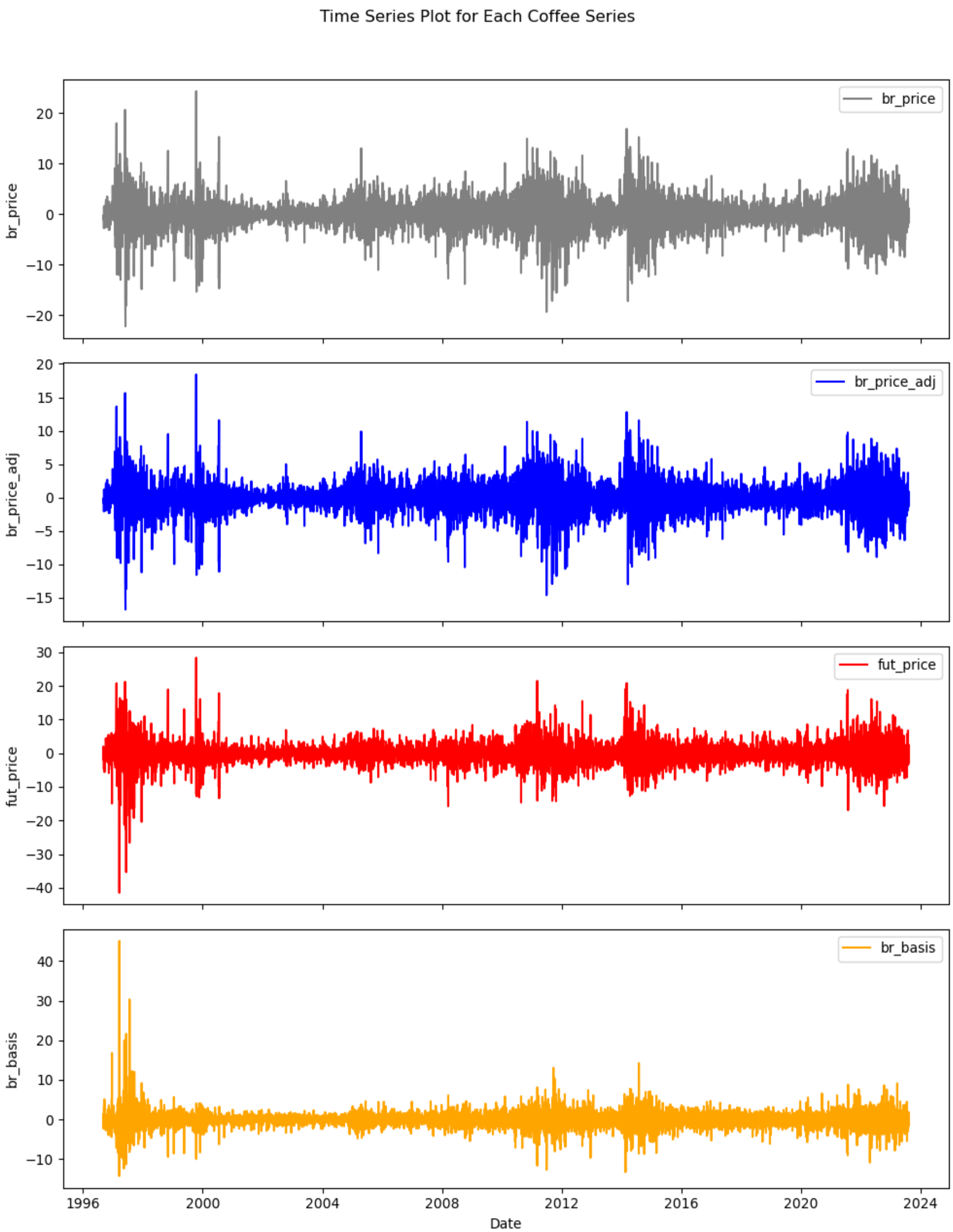
Preprocessing:

Descriptive Statistics:

Initial exploration of the basis series through basic descriptive statistics reveals notable characteristics. The basis series exhibits lower volatility, as indicated by its lowest standard deviation, and demonstrates a narrower range compared to both spot and future market prices. This aligns with the financial concept that basis risk tends to be lower than the inherent risk associated with individual spot or future markets. Additionally, a noteworthy observation is the predominantly negative nature of the Brazilian coffee basis, indicating a prevalent scenario where spot prices are lower than corresponding future prices.

	br_price	br_price_adj	fut_price	br_basis
count	6515.000000	6515.000000	6515.000000	6515.000000
mean	138.433088	104.653779	129.811604	-25.157825
std	59.825587	45.227437	49.486644	9.009344
min	30.920000	23.375155	41.500000	-98.992103
25%	103.840000	78.501813	101.875000	-29.039687
50%	127.380000	96.297775	121.100000	-23.507018
75%	169.150000	127.875401	153.350000	-19.279366
max	349.390000	264.134711	304.900000	3.641727

Stationarity and Correlation:
The preprocessed data includes differentiated series to ensure stationarity, a crucial prerequisite for time series modeling.



Time Series Plot for Each Coffee Series
br_price
20
10
0
−10
−20
br_price_adj
20
15
10
5
0
−5
−10
−15
fut_price
30
20
10
0
−10
−20
−30
−40
br_basis
40
30
20
10
0
−10
1996
2000
2004
2008
2012
2016
2020
2024
Date

The correlation matrix reveals insightful relationships among key variables:

	br_price	br_price_adj	fut_price	br_basis
br_price	1.000000	1.000000	0.778723	-0.132908
br_price_adj	1.000000	1.000000	0.778723	-0.132908
fut_price	0.778723	0.778723	1.000000	-0.725300
br_basis	-0.132908	-0.132908	-0.725300	1.000000

Notably, the coffee future price exhibits a significant positive correlation with the spot Brazilian coffee price (approximately 0.78). However, its correlation with basis risk is negative, as expected (basis = spot - future), albeit at a lower level (approximately -0.73). This discrepancy in correlation levels suggests the influence of additional factors, such as the exchange rate, on basis risk. Thus, the preprocessing stage provides a foundation for in-depth modeling, considering the nuanced relationships within the dataset.

Modeling:

ARIMA Model Selection:

In our pursuit of an optimal ARIMA model, a comprehensive grid search was conducted, exploring various combinations of the autoregressive (p) and moving average (q) parameters. The selection criteria for determining the best model were based on both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

```
Top 5 AIC values:
    p  q           AIC           BIC
35  5  5  21506.443563  21577.932884
34  5  4  21507.746778  21572.737070
15  2  3  21508.285105  21547.279280
29  4  5  21510.197206  21575.187498
9   1  3  21511.051070  21543.546216

Top 5 BIC values:
    p  q           AIC           BIC
8   1  2  21512.571696  21538.567812
9   1  3  21511.051070  21543.546216
14  2  2  21511.546480  21544.041626
15  2  3  21508.285105  21547.279280
20  3  2  21511.611466  21550.605642
```

Despite encountering convergence issues with several of the top AIC models, ARIMA(1,1,3) emerged as the most robust contender, demonstrating convergence and consistently ranking among the top models for both AIC and BIC criteria. The chosen model, ARIMA(1,1,3), will serve as the baseline for further evaluation and comparison with alternative models.

Base Model Comparison: ARIMA(1,1,3) vs. ARIMA(1,1,1)

The choice of a base model plays a pivotal role in predictive modeling. In our analysis of the Brazilian basis risk, we initially selected ARIMA(1,1,3) as the baseline model. However, an alternative contender, ARIMA(1,1,1), exhibits comparable performance while being more parsimonious.

SARIMAX Results			
Dep. Variable:	br_basis	No. Observations:	1604
Model:	ARIMA(1, 1, 3)	Log Likelihood	-3309.003
Date:	Fri, 01 Dec 2023	AIC	6628.007
Time:	08:58:29	BIC	6654.905
Sample:	01-03-2017	HQIC	6637.994
	- 08-08-2023		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.7644	0.065	11.682	0.000	0.636	0.893
ma.L1	-0.8820	0.067	-13.091	0.000	-1.014	-0.750
ma.L2	0.0041	0.028	0.143	0.886	-0.052	0.060
ma.L3	-0.0072	0.027	-0.265	0.791	-0.060	0.046
sigma2	3.6347	0.083	43.573	0.000	3.471	3.798

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	671.30
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	3.64	Skew:	-0.37
Prob(H) (two-sided):	0.00	Kurtosis:	6.08

SARIMAX Results

Dep. Variable:	br_basis	No. Observations:	1604
Model:	ARIMA(1, 1, 1)	Log Likelihood	-3309.027
Date:	Fri, 01 Dec 2023	AIC	6624.055
Time:	08:59:26	BIC	6640.193
Sample:	01-03-2017	HQIC	6630.047
	- 08-08-2023		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.7743	0.037	20.806	0.000	0.701	0.847
ma.L1	-0.8908	0.027	-33.557	0.000	-0.943	-0.839
sigma2	3.6349	0.082	44.480	0.000	3.475	3.795

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	676.45
Prob(Q):	0.96	Prob(JB):	0.00
Heteroskedasticity (H):	3.65	Skew:	-0.37
Prob(H) (two-sided):	0.00	Kurtosis:	6.10

Comparison:
Both models share comparable log likelihoods, AIC, and BIC values. ARIMA(1,1,1) stands out for its parsimony, boasting fewer parameters. Significantly deviating from zero, the parameters in ARIMA(1,1,1) reinforce its statistical robustness. Diagnostic tests reveal that both models' residuals demonstrate a lack of autocorrelation, as indicated by low Ljung-Box Q values. However, the presence of heteroskedasticity in both models suggests variations in residuals' variance over time.

Conclusion:
The ARIMA(1,1,1) model stands out as a more parsimonious alternative, maintaining comparable performance to the ARIMA(1,1,3) model. The choice between these models should consider the trade-off between complexity and predictive accuracy. Further model refinement and comparison will be conducted to enhance our understanding of the basis risk in the Brazilian coffee market.

XGBoost Model:

A comprehensive grid search using GridSearchCV was conducted to optimize the hyperparameters of the XGBoost model. The resulting model exhibits robust performance, effectively capturing the intricate patterns within the data.

The XGBoost model highlights the Brazilian coffee spot price as the most influential feature for predicting the basis, followed closely by the future coffee price and the USD/BRL exchange rate.

Prophet Model:

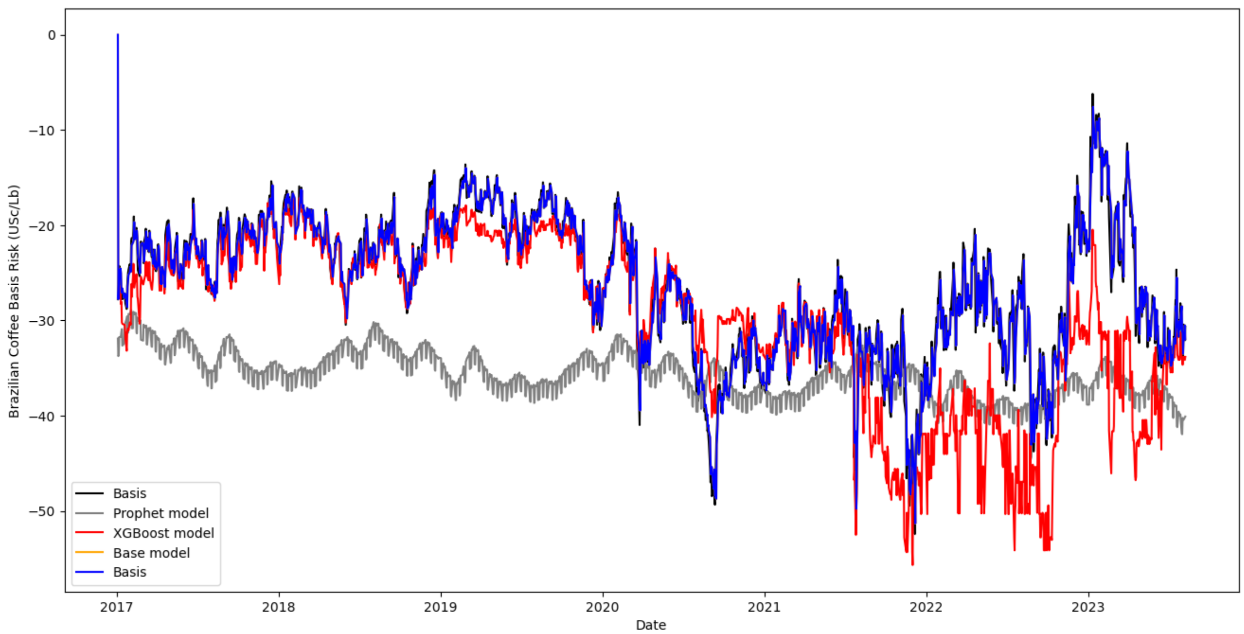
The Prophet model, designed for time series forecasting, adeptly identifies and decomposes the Basis series into distinct trend and seasonality components. This model not only discerns general trends but also recognizes both weekly and yearly seasonality patterns embedded in the data. The ability to capture these nuanced temporal patterns enhances the model's forecasting capabilities for the basis series.

Models Comparison:

	Base model	ARIMA_111	XGBoost	Prophet
MAE	1.419698	1.419602	4.544749	10.098708
RMSE	2.028188	2.028214	7.071423	11.639217

The ARIMA(1,1,1) model demonstrates the best performance with the lowest MAE and RMSE scores, indicating its effectiveness in capturing the underlying patterns in the test data. XGBoost also exhibits competitive scores, showcasing its predictive capabilities. Notably, the Prophet model, designed to handle seasonality, shows higher error metrics compared to the other models, suggesting potential challenges in accurately capturing the complex temporal patterns of the Basis series.

Visual Comparison



The visual comparison of model predictions on the testing data reinforces the efficacy of the ARIMA models, which closely align with the actual data. Both ARIMA(1,1,3) and ARIMA(1,1,1) demonstrate exceptional performance in capturing the nuances of the Basis series.

XGBoost stands out as a robust predictor, especially for short to mid-term data points. Its predictions exhibit a notable alignment with the actual values, reflecting its ability to adapt to the intricacies of the time series data.

Prophet, while capturing broader trends, appears to struggle in accurately predicting the finer details of the Basis series. The model's predictions show deviations from the actual data points, particularly in regions with higher volatility.

Overall, the visual analysis corroborates the numerical comparison, highlighting the strengths and limitations of each model in capturing the dynamics of the Brazilian basis risk.

Conclusion:

In summary, our comprehensive analysis of various time series models for predicting Brazilian basis risk highlights the ARIMA(1,1,1) model as the standout performer. This model exhibits exceptional efficacy in modeling the intricate dynamics of the Brazilian basis risk time series, excelling not only in short-term predictions but also showcasing its competence in forecasting over longer horizons.

While ARIMA(1,1,1) takes the lead, the XGBoost model emerges as a valuable complementary tool, demonstrating particular strength in short to mid-term forecasting. Its ability to provide reliable insights into the near-future dynamics of basis risk enhances its practical utility for market practitioners, hedgers, and investors.

In conclusion, both the ARIMA(1,1,1) and XGBoost models prove to be robust solutions for capturing and predicting the complexities inherent in the Brazilian basis risk. Their accurate predictions serve as valuable inputs for informed decision-making in real-world scenarios, contributing to a more nuanced understanding of market trends and aiding in the development of effective risk management strategies. The versatility and reliability of these models position them as valuable assets for stakeholders navigating the volatile landscape of financial markets.