Market Risk Forecasting: Traditional vs. Machine Learning Approaches

Guillaume Friez

2023-2025

Supervisor : Prof. Guillaume Coqueret

MSc In Finance 2025

em lyon business school

Introduction

Context

- Since 2008 and again during the COVID-19 crisis, financial institutions have experienced major failures in their risk models.
- Traditional Value-at-Risk models failed to anticipate tail events, resulting in regulatory breaches and unexpected capital requirements.
- These repeated breakdowns highlight the urgent need to revisit how we forecast market risk, especially under extreme conditions

Motivation

- Traditional models often fall short during market turmoil
- While ML offers more flexibility, it must also meet regulatory demands for interpretability
- During my internship at Rothschild & Co, I saw firsthand the limits of VaR models under stress, which led me to explore smarter, more robust alternatives

Objective

This thesis aims to compare traditional, ML, and hybrid market risk forecasting models.

We assess them based on :

1. **Accuracy**
2. **Robustness** during crises
3. **Interpretability** (SHAP and LIME)

The goal is to determine if AI-based models can outperform classical approaches without sacrificing explainability.

Research Question & Hypotheses

em lyon business school

Main question

Can ML and hybrid models improve market risk prediction while remaining interpretable and robust in crises ?

Hypotheses:

- **H1a**: ML/hybrid models → higher accuracy
- **H1b**: ML/hybrid models → more robust in stress
- **H1c**: ML/hybrid models → interpretable enough via SHAP/LIME
- **H0**: No significant improvement over traditional methods

Research Design

Assets	ETFs, equities, bonds, cyrptos and commodities
Risk Metrics	VaR, CVaR, volatility and drawdown
Horizons	1-day and 21-day
Dimensions	Accuracy, Robustness and Interpretability

Forecasting Models Compared

Category	Models Implemented	Notes
Traditional	Historical VaR, Parametric VaR, CVaR, GARCH	Basel III compliant, but limited in tail risk and regime change handling
Machine Learning	Random Forest, Gradient Boosting, XGBoost, LSTM	Data-driven, distributional assumptions
Hybrid	GARCH-LSTM	Combines d-short-term volatility from GARCH + LSTM memory for sequence modeling

Input Data and Risk Targets

Data Sources

- Market and macro data from Refinitiv Eikon
- Multiple asset classes : ETFs, Equities, Bonds, Cryptocurrencies, Commodities

Contextual Features (by Asset Class)

- **Commodities:** volume-weighted average price, number of intraday price moves, indicative NAV
- **Cryptocurrencies** → medium and long-term moving averages (50D, 200D) to capture momentum
- **Equities** → liquidity-adjusted price levels, short-term trend indicators, block trading volume
- **ETFs** → indicative net asset value (NAV), institutional block trade counts
- **Fixed Income** → indicative NAV, bid and ask price quotes for spread estimation

Feature Categories

- **Price-based**: returns, volatility (10/30/60d), drawdown
- **Technical Indicators**: RSI, Moving Averages, momentum, turnover…
- **Macroeconomic**: VIX, CPI, 10Y yields, PLI, Fed rates…
- **Engineering Metrics** : skewness, kurtosis, realized variance, liquidity shocks…

Forecast Targets

- **Value-at-Risk** (1-day, 21-day at 95% and 99%)
- **Conditional VaR (CVaR)**
- **Realized Volatility**
- **Maximum drawdown**

Pipeline & Methodology

em lyon business school

Data Collection

- Historical time series from Refinitiv
- Asset-level folders by class (equities, crypto, bonds…)
- Macroeconomic time-aligned dataset

Preprocessing & Feature Engineering

- Cleaning: missing values, outlier filtering, forward-fill for macro
- Rolling window computations: returns, volatilities, drawdowns
- Feature scaling (Min-Max) and standardization
- Dynamic feature injection depending on asset class
- Validation: assets filtered based on data quality (NaN %, min history…)

Model Training

- Time-aware train/test split (last 20% held out)
- Recursive Feature Elimination (RFE)
- RandomizedSearchCV for hyperparameter tuning
- One model per target, asset and horizon (no multi-tasking)
- LSTM inputs reshaped into 3D sequences (window-based)

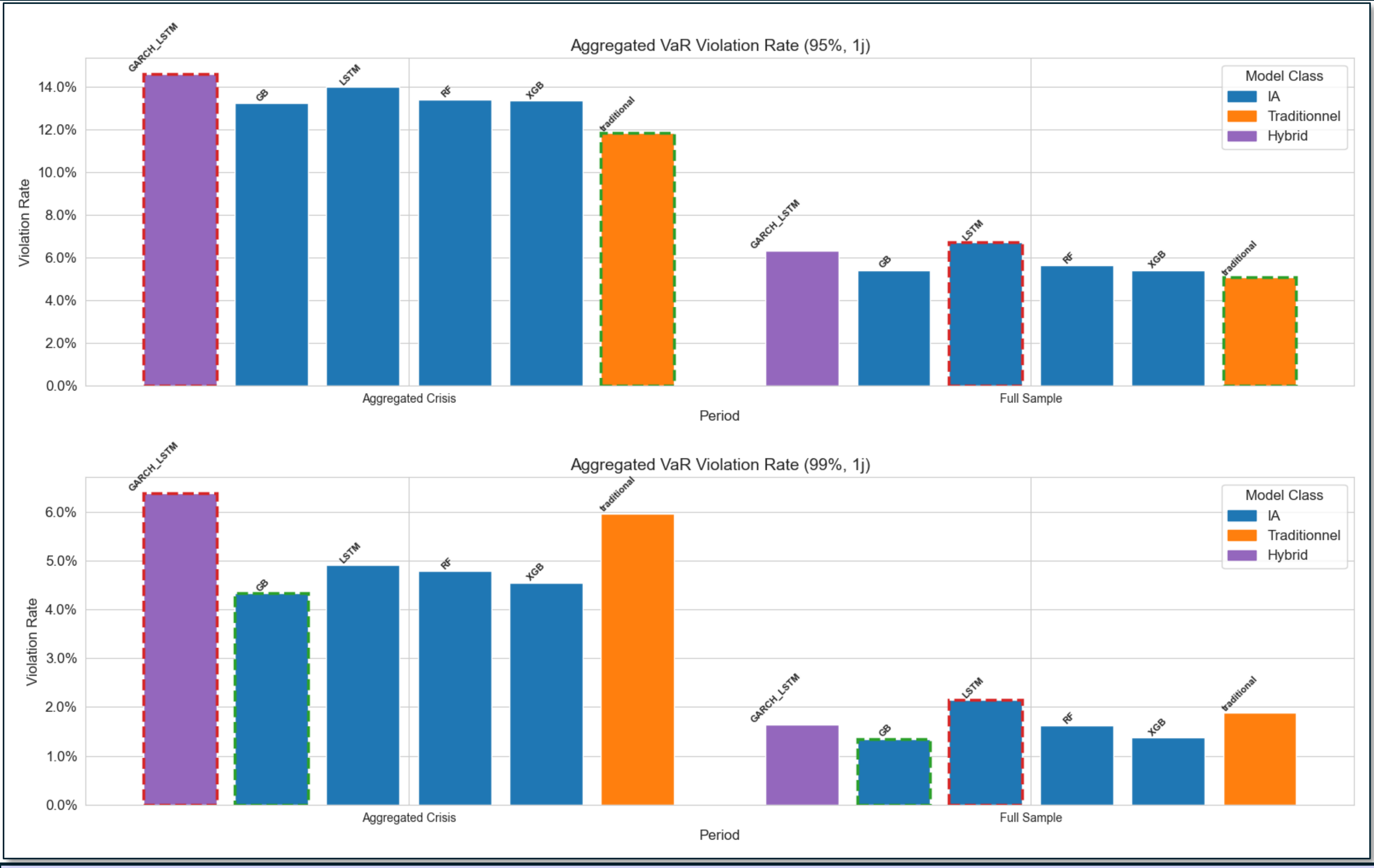
Evaluation & Backtesting

- VaR: Quantile loss, Violation Rate, kupiec Test
- CVaR: Quantile Loss
- Volatility / Drawdown: RMSE, MAE, Adjusted R^2
- Interpretability :
  - SHAP (RF, XGB, GB)
  - LIME (LSTM, GARCH-LSTM)

RESULTS

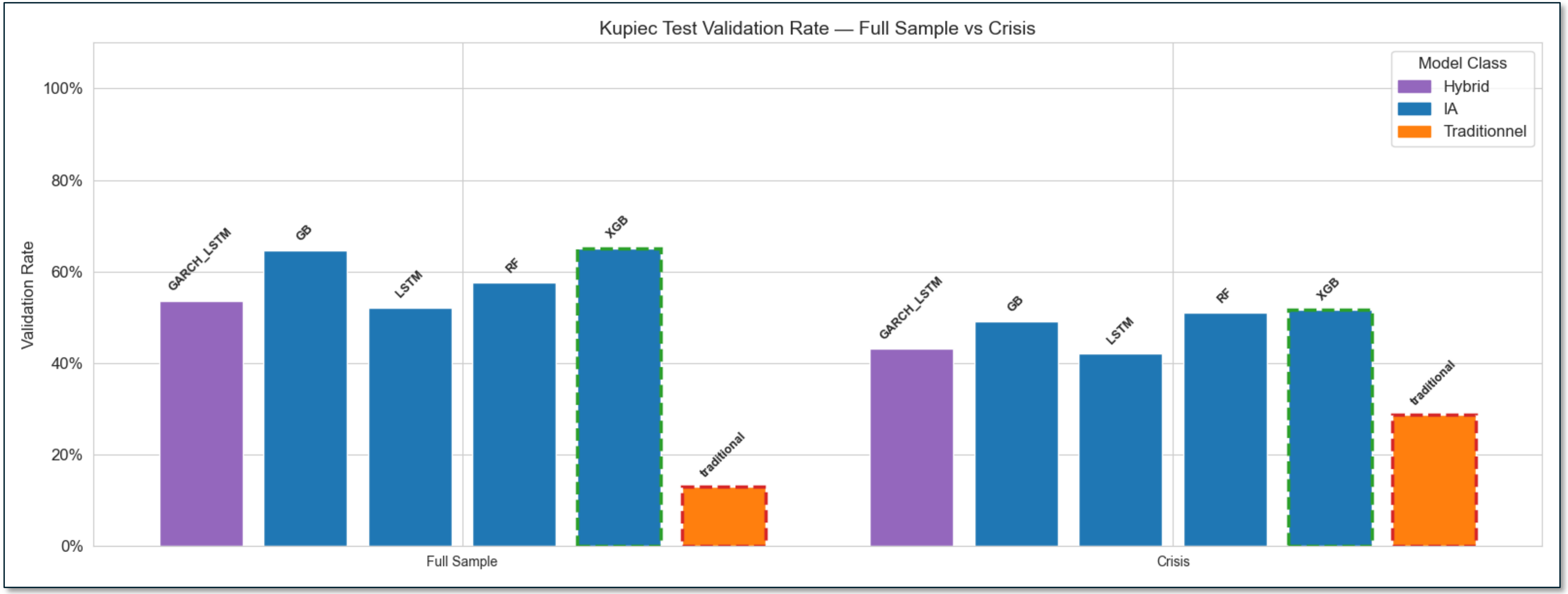
VaR Violation Rate

em lyon business school



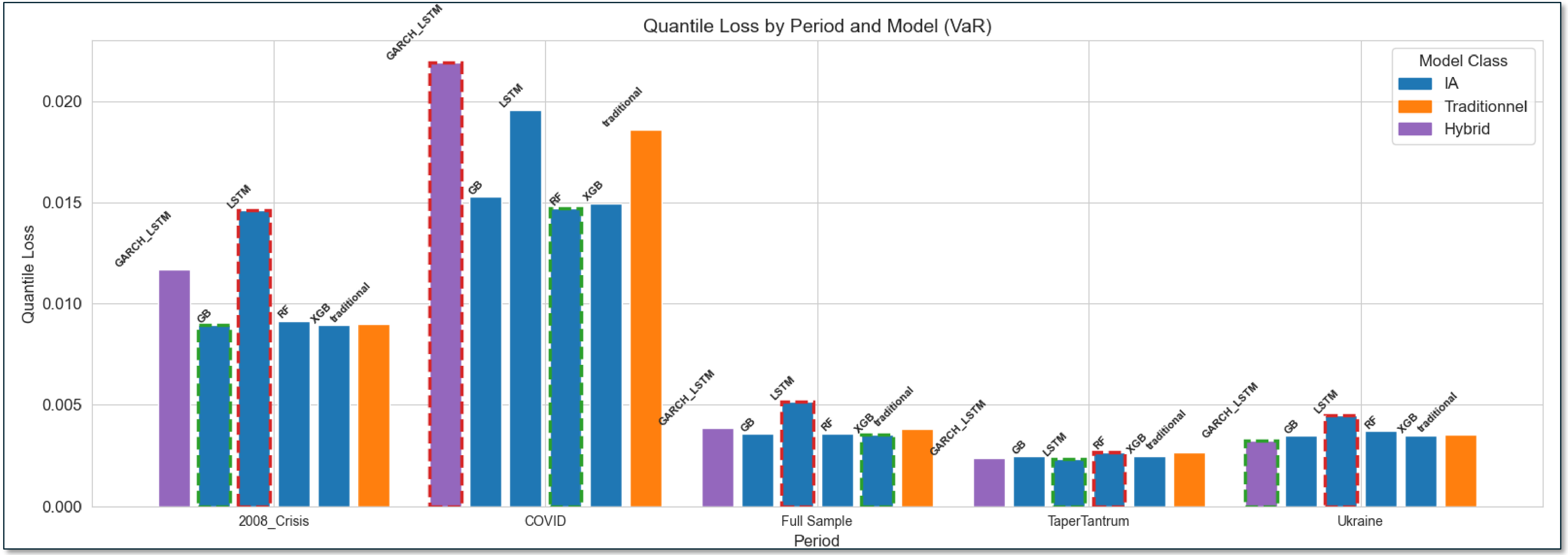
- **Gradient Boosting, XGBoost, and Random Forest** show strong calibration at both 95% and 99%, even under crisis conditions.
- **Traditional models** are too conservative in calm periods but fail during crises, with violation rates rising above 12%.
- **LSTM and GARCH-LSTM** consistently underpredict tail risk, with violation rates reaching 14–15% at 95% and up to 6% at 99%.

Kupiec Backtest Results



XGBoost passes in both full sample and crisis periods.
Random Forest and **Gradient Boosting** also show strong validity.
Traditional models fail consistently, especially under stress.
LSTM and GARCH-LSTM show poor calibration, with clustered breaches.

Quantile Loss (VaR)

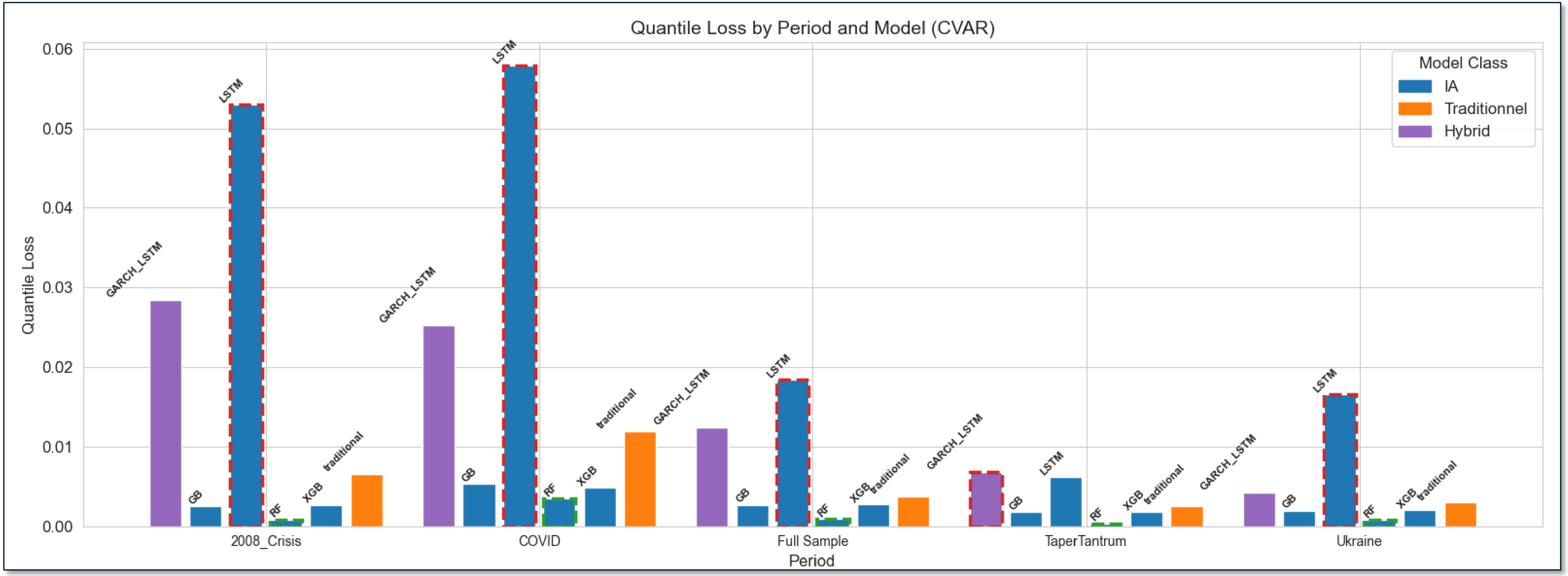


XGBoost, RF, and GB show the **lowest quantile loss**, including during 2008 and COVID.

Traditional models are acceptable overall but struggle under certain conditions of stress.

LSTM and GARCH-LSTM perform poorly in crises, with high loss values and unstable behavior.

CVaR Results



Tree-based models (XGB, RF, GB) provide the **most accurate estimates** of expected tail losses.

Traditional models are less precise, especially in turbulent periods.

LSTM and hybrids again fail in crises, with large CVaR loss and erratic behavior.

Volatility & Drawdown Prediction

MAE

Category	GARCH_LSTM	GB	LSTM	RF	traditional	XGB
Drawdown (Crisis)	0,05	0,04	0,23	0,02	0,06	0,04
Drawdown (Full Sample)	0,03	0,03	0,13	0,01	0,04	0,03
Volatility (Crisis)	0,15	0,15	0,21	0,13	0,20	0,15
Volatility (Full Sample)	0,09	0,11	0,14	0,08	0,15	0,11

RMSE

Category	GARCH_LSTM	GB	LSTM	RF	traditional	XGB
Drawdown (Crisis)	0,07	0,06	0,31	0,03	0,07	0,05
Drawdown (Full Sample)	0,05	0,05	0,18	0,02	0,06	0,05
Volatility (Crisis)	0,20	0,20	0,29	0,18	0,28	0,20
Volatility (Full Sample)	0,13	0,17	0,21	0,13	0,24	0,16

R^2 adjusted

Category	GARCH_LSTM	GB	LSTM	RF	traditional	XGB
Drawdown (Crisis)	0,18	0,77	-290,20	0,88	-2,62	0,77
Drawdown (Full Sample)	0,16	0,78	-197,70	0,88	-0,16	0,78
Volatility (Crisis)	0,12	0,48	-0,25	0,51	-1,81	0,49
Volatility (Full Sample)	0,11	0,49	-0,34	0,52	-0,20	0,49

- **RF and XGB** consistently deliver the **lowest errors** and **highest R^2**, both in normal and crisis regimes.
- **LSTM** shows severe instability, with extreme errors and deeply negative R^2 — especially on drawdown.
- **Traditional models** and **GARCH-LSTM** perform slightly better but still fall short in explaining risk under stress.
- **Yellow cells highlight the second-worst performers,** confirming that only tree-based models are reliably robust.

INTERPRETABILITY (XAI)

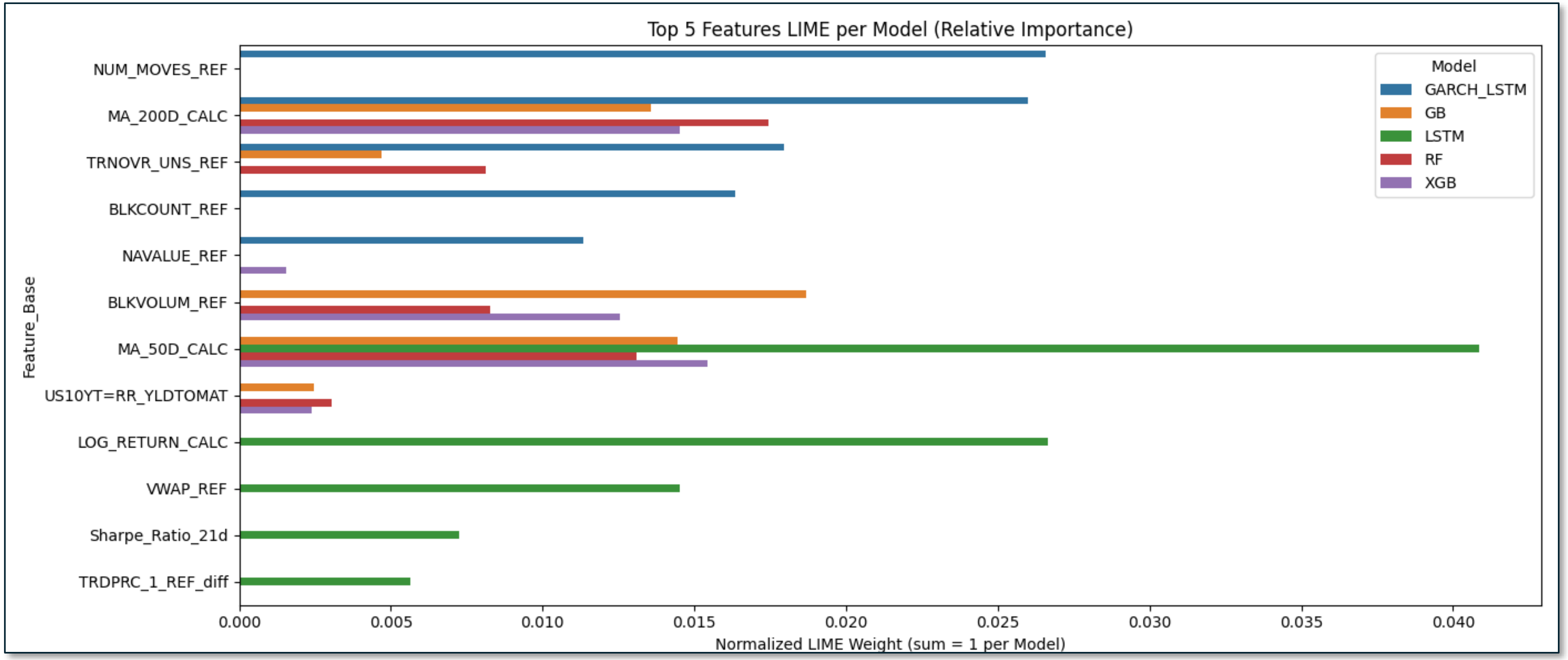
LIME Analysis

MA_50D and **LOG_RETURN** are highly used by LSTM, but lead to fragile performance

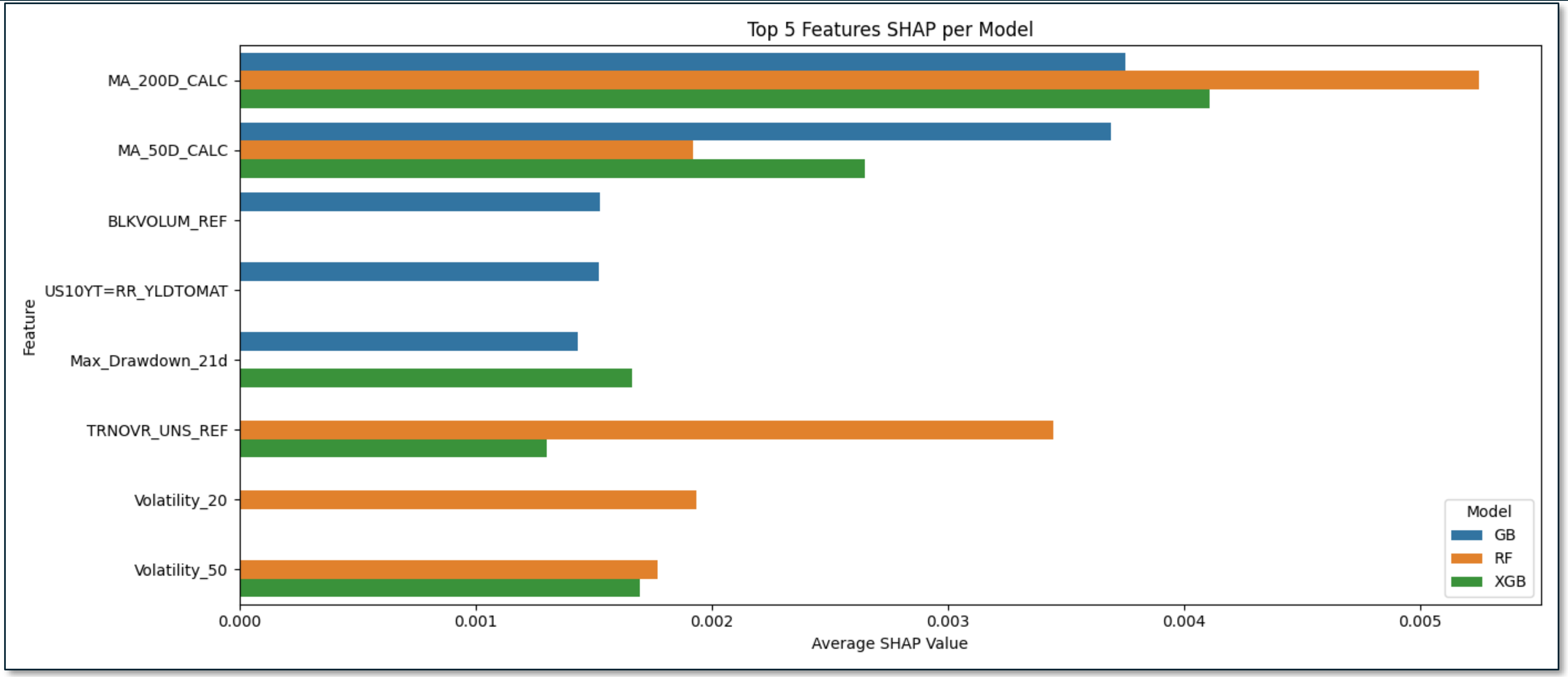
Volume and NAV-based features (e.g., BLKVOLUM, NAVALUE) are shared by tree-based models

US10Y emerges as a strong macro predictor in GB and XGB

→ Tree-based models focus more on **economically relevant and stable signals**



SHAP Analysis

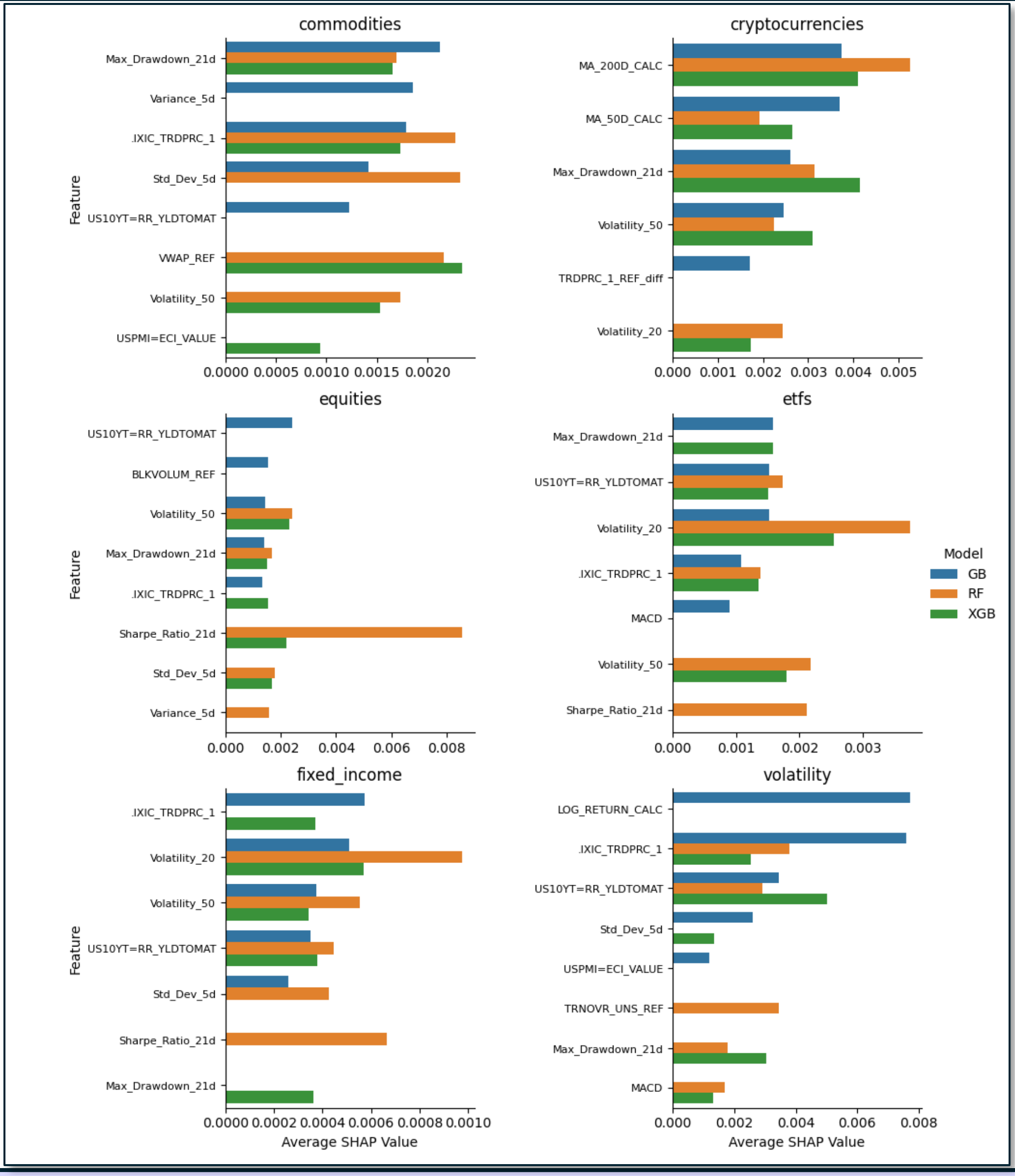


MA_200D and **MA_50D** (trend) are top predictors across models

Volatility & turnover features (Vol_50, TRNOVR) strongly contribute

Macro-input like US10Y appear in GB — intuitive and explainable

SHAP by Asset Class



SHAP attribution varies meaningfully by asset class:

- **Bonds** → macro rates (US10Y), yield spreads
- **Cryptos** → trend and drawdown metrics dominate
- **ETFs** → NAV deviation, liquidity/volume signals

→ The model dynamically learns the relevant drivers depending on asset structure.

Tree-based models not only perform well — they also provide **clear, stable and economically meaningful explanations**. Their ability to adapt to each asset class reinforces their credibility for real-world risk management.

WRAP-UP

Summary Table

Status	Model	Accuracy	Robustness	Interpretability
✓	**XGBoost**	Best Overall (low ViR, Kupiec OK)	Strong in stress, stable R^2	SHAP/LIME Consistent
✓	**Random Forest**	Stable VaR + Kupiec passed	Best on drawdown/volatility metrics	SHAP/LIME interpretable
✓	**Gradient Boosting**	Minor violations, solid shape	Good balance	Globally coherent
✗	**Traditional**	Systematic rejection (Kupiec failed)	Fragile under stress (neg. R^2)	Transparent
⚠	**LSTM**	Severe over-violation & test failure	Catastrophic drawdown & volatility forecasts	Transparent but no SHAP
⚠	**GARCH-LSTM**	Fails on calibration	Slight improvement over LSTM	Transparent but no SHAP

Limitations & Future Work

Limitations

- **LSTM and GARCH-LSTM models underused**
  Due to time and compute constraints, hyperparameter tuning was minimal, limiting their potential.
- **Interpretability on sequential models is partial**
  SHAP/LIME offer only approximate insights on LSTM architectures.
- **Feature selection not fully optimized**
  Top SHAP/LIME variables were not yet reused to retrain models.
- **Vendor bias & data scope**
  The study relied solely on Refinitiv data — results lack external validation.
- **Real-time performance not assessed**
  Models were not tested in live market streaming conditions.

Future Work

- **Re-train models using SHAP-selected features**
  Iterative feature refinement could improve both performance and interpretability.
- **Explore attention-based architectures**
  Improve tuning and architecture for LSTM/GARCH-LSTM
- **Expand to real-time and news-based variables**
  Add macroeconomic and news-based variables
- **Add new models to the comparison**
  Modelize GAN model and apply it to EVT theory
- **Improve hybrid models with variants**
  Apply hybrid models with different types of GARCH models (EGARCH, GJR-GARCH)