Open-Data Based India VIX Forecasting and Regime-Aware Equity Timing

DES646 Course Project Report

Team Members

SATYAM SINGH (220985)
SHASHWAT GAUTAM (221005)
VANSH MINA (221167)
HASHVITH (221158)

November 9, 2025

Abstract

This report documents an end-to-end, open-data driven framework for forecasting India VIX and converting those forecasts into regime-aware risk-on / risk-off signals for the Indian equity market. Using only publicly available data, we (i) construct a robust data pipeline from minute-level quotes to weekly aggregates, (ii) engineer economically meaningful volatility and regime features, (iii) train complementary linear and non-linear models (Elastic Net and gradient boosting) to predict next-week VIX returns, and (iv) combine these models into an interpretable, regime-aware blended signal that can guide allocation, hedging, or overlay strategies. The design philosophy follows an "open, explainable, retail-friendly" approach: every transformation is transparent, reproducible, and aligned with realistic implementation constraints.

Contents

1 Introduction

Volatility indices condense the market's expectation of short-term risk into a single, tradable number. For India, the India VIX—derived from NIFTY 50 option prices—serves as a forward-looking indicator of expected equity market volatility. Timely and interpretable signals about the likely direction or regime of India VIX are valuable for:

- Dynamic equity allocation (risk-on vs. risk-off),
- Hedging decisions using index derivatives,
- Understanding stress build-up and regime shifts.

In parallel, recent work on open-data equity-timing for the Indian market highlights that reasonably strong timing and risk signals can be built entirely from public sources: index levels, factor indices, mutual fund flows, macro calendars, and implied volatility. This project instantiates that philosophy in a focused way: we design a complete, well-engineered India VIX forecasting module, suitable as a plug-in to a larger open-data timing framework.

Our goals are: (1) to treat modeling as a careful engineering pipeline rather than a black box, (2) to emphasize interpretability alongside accuracy, and (3) to respect realistic constraints (no look-ahead leakage, transparent features, and simple deployment logic).

2 Objectives

The specific objectives of this project are:

(O1) Build a clean pipeline from raw, high-frequency India VIX data to weekly bars.
(O2) Engineer a compact yet expressive feature set capturing level, momentum, realized volatility, volatility-of-volatility, and drawdown states.
(O3) Predict next-week India VIX returns using two complementary models: a regularized linear model (Elastic Net) and a tree-based non-linear model.
(O4) Design a regime-aware blending scheme to convert model outputs into risk-on / risk-off signals.
(O5) Present the entire framework in an explainable and reproducible manner, suitable for teaching and for open-data retail use.

3 Data Description and Design Principles

3.1 Data Sources

- **India VIX**: Historical minute-level or high-frequency data with timestamped quotes, from which we derive daily and weekly series.
- **Index-Level References (Optional)**: NIFTY 50 levels or basic index returns to contextualize volatility regimes (used only when available from public sources).

All data sources are: (i) publicly available or reproducible using official exchange data, (ii) free from proprietary vendor-only fields, and (iii) aligned with the open-data philosophy.

3.2 Open-Data and No-Leakage Design

Key principles:

- **Chronological Discipline**: Every feature at week w only uses information available up to the end of week w.
- **Reproducibility**: All transformations are simple, documented, and can be rebuilt from raw CSV-style data.

- **Explainability**: Feature definitions are economically interpretable (e.g., "1-month VIX momentum", "realized vol", "stress regime").

4 Methodology

4.1 Step 1: Minute-Level Cleaning

Let P_t denote the India VIX level at minute t.

(1.a) Parse timestamps and sort records strictly by time.
(1.b) Standardize column names (e.g., `open`, `high`, `low`, `close`).
(1.c) Compute minute returns:

$$r_t^{\text{min}} = \frac{P_t - P_{t-1}}{P_{t-1}},$$

handling missing values and obvious bad ticks via simple filters.

This produces a consistent intraday series suitable for realized volatility estimation.

4.2 Step 2: Daily Aggregation

For each trading day d, define:

$$\begin{aligned} O_d &= \text{first } P_t \text{ of day } d, \\ H_d &= \max_{t \in d} P_t, \\ L_d &= \min_{t \in d} P_t, \\ C_d &= \text{last } P_t \text{ of day } d. \end{aligned}$$

Realized intraday volatility for day d is approximated as:

$$\sigma_d^{\text{RV}} = \sqrt{\sum_{t \in d} (r_t^{\text{min}} - \bar{r}_d)^2},$$

where $\bar{r}_d$ is the mean of minute returns on day d.

Additional daily features:

$$\text{range\_abs}_d = H_d - L_d, \quad \text{range\_pct}_d = \frac{H_d - L_d}{C_d}.$$

Days with incomplete or unreliable data are dropped.

4.3 Step 3: Weekly OHLC and Volatility Features

We aggregate to weekly frequency using a Friday-close (or last trading day) convention.

For week w:

$$\begin{aligned} O_w &= \text{first } O_d \text{ of week } w, \\ H_w &= \max_{d \in w} H_d, \\ L_w &= \min_{d \in w} L_d, \\ C_w &= \text{last } C_d \text{ of week } w. \end{aligned}$$

We also compute:

- Mean realized volatility:

$$\sigma_w^{\text{RV}} = \frac{1}{|w|} \sum_{d \in w} \sigma_d^{\text{RV}}.$$

- Average daily range:

$$\overline{\text{range\_pct}}_w = \frac{1}{|w|} \sum_{d \in w} \text{range\_pct}_d.$$

Weeks with insufficient daily coverage are removed.

4.4 Step 4: Weekly Returns and Momentum

The one-week VIX return is:

$$r_{w,1}^{\text{VIX}} = \frac{C_w - C_{w-1}}{C_{w-1}}.$$

We define multi-horizon momentum features:

$$\text{mom}_k(w) = \frac{C_w - C_{w-k}}{C_{w-k}}, \quad k \in \{2, 4, 8, 12\}.$$

These capture short- to medium-term trends in implied volatility.

4.5 Step 5: Volatility-of-Volatility

To capture instability in volatility itself, we compute rolling standard deviations of weekly returns:

$$\text{vol}_k(w) = \text{Std}\left(r_{w',1}^{\text{VIX}}\right)_{w'=w-k+1}^{w}, \quad k \in \{4, 8, 12\}.$$

High values of $\text{vol}_k(w)$ indicate unstable or shock-prone regimes.

4.6 Step 6: Drawdowns and Regime Buckets

We define a 52-week rolling peak in VIX:

$$C_{52}^{\max}(w) = \max_{w-51 \le j \le w} C_j,$$

and the corresponding drawdown in volatility space:

$$\text{DD}_{52}(w) = \frac{C_w - C_{52}^{\max}(w)}{C_{52}^{\max}(w)}.$$

Using the empirical distribution of C_w (or $r_{w,1}^{\text{VIX}}$), we create discrete regimes:

- Regime 0: Very Low VIX (compressed risk premia),
- Regime 1: Normal Range,
- Regime 2: Elevated,
- Regime 3: High / Stress.

These regime labels are used both as features and to control model blending.

4.7 Step 7: Target Construction

Our prediction target is the next-week VIX return:

$$y_w = \frac{C_{w+1} - C_w}{C_w}.$$

For each week w, we construct a feature vector X_w using only information up to week w, and pair it with target y_w. Rows with any missing rolling-value features are dropped.

5 Modeling Framework

5.1 Train–Test Split

To respect time ordering, we use a chronological split:

- Oldest $\approx 70\%$ of observations: training set.
- Most recent $\approx 30\%$: test set.

No shuffling is performed. Any scaling or transformation is fitted on the training set and applied to the test set.

5.2 Model 1: Elastic Net Regression

Elastic Net serves as an interpretable baseline. Let X be the feature matrix and y the target vector. Elastic Net solves:

$$\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \left(\alpha \|\beta\|_1 + \frac{1-\alpha}{2} \|\beta\|_2^2 \right) \right\},$$

where:

- λ controls the overall strength of regularization,
- α interpolates between LASSO ($\alpha = 1$) and Ridge ($\alpha = 0$).

Hyperparameters (λ, α) are selected via cross-validation on the training period. The resulting coefficients highlight which features (momentum, vol-of-vol, drawdowns, regimes) linearly drive next-week VIX moves.

5.3 Model 2: Gradient Boosting / LightGBM

To capture non-linearities and interactions, we use a gradient boosted decision tree model (e.g., LightGBM) on the same feature set.

Key settings (conceptually):

- Depth and number of trees constrained to avoid overfitting.
- Learning rate set conservatively.
- No leakage: trees see only X_w at week w for predicting y_w.

This model can learn effects such as:

- Momentum behaving differently in low vs. high VIX regimes.
- Thresholds in vol-of-vol beyond which future spikes are more likely.

Feature importance and partial dependence plots are used to explain these learned relationships in intuitive terms.

5.4 Regime-Aware Blended Signal

Let $\hat{y}_w^{(E)}$ be Elastic Net predictions and $\hat{y}_w^{(G)}$ be gradient boosting predictions.
We define weights as functions of the current regime R_w:

$$\omega_E(R_w),\ \omega_G(R_w) \geq 0, \quad \omega_E(R_w) + \omega_G(R_w) = 1.$$

A simple scheme:

- In Regimes 0–1 (low/normal): put more weight on the non-linear model.
- In Regimes 2–3 (elevated/stress): put more weight on Elastic Net for stability.

The blended forecast is:

$$\hat{y}_w^{\text{blend}} = \omega_E(R_w)\,\hat{y}_w^{(E)} + \omega_G(R_w)\,\hat{y}_w^{(G)}.$$

We convert this into a directional signal:

$$S_w = \begin{cases} +1 & \text{if } \hat{y}_w^{\text{blend}} > \theta, \\ -1 & \text{if } \hat{y}_w^{\text{blend}} < -\theta, \\ 0 & \text{otherwise,} \end{cases}$$

for a small threshold $\theta > 0$ to ignore noise.
Interpretation:

- $S_w = +1$: VIX expected to rise $\Rightarrow$ Risk-Off bias.
- $S_w = -1$: VIX expected to fall $\Rightarrow$ Risk-On bias.
- $S_w = 0$: Uncertain / Neutral.

6 Evaluation

6.1 Forecast Metrics

On the test set, we evaluate:

- Root Mean Squared Error:
$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_w (y_w - \hat{y}_w)^2}.$$
- Mean Absolute Error (MAE).
- Directional Accuracy:
$$\Pr\left[\text{sign}(\hat{y}_w) = \text{sign}(y_w)\right].$$
- Pearson correlation between y_w and $\hat{y}_w$.

We compare: (i) Elastic Net, (ii) Gradient Boosting, and (iii) the blended model.

6.2 Illustrative Strategy Backtest

To connect forecasts with economic meaning, we consider a stylized strategy:

- Go long a VIX-proxy when $S_w = +1$,
- Go short (or underweight hedges) when $S_w = -1$,
- Stay flat when $S_w = 0$.

Weekly strategy returns:

$$r_w^{\text{strat}} = S_w \cdot y_w.$$

From $\{r_w^{\text{strat}}\}$ we derive:

- Cumulative return curve,
- Annualized volatility,
- Sharpe ratio (with a simple risk-free proxy),
- Maximum drawdown,
- Hit-rate (fraction of profitable weeks).

This backtest is presented only as an illustration. Any unusually strong performance is treated with caution and subjected to sanity checks such as rolling-window evaluation and hyperparameter robustness.

7 Explainability, UX, and Integration

7.1 Explainable Outputs

To make the system accessible to non-expert users:

- Provide a per-week "signal card" listing: top features influencing the forecast, current VIX regime, and the sign of S_w.
- Use simple labels: *Calm*, *Normal*, *Watch*, *Stress* for regimes.
- Visualize: time-series of VIX, predicted vs. realized returns, and regime shading.

7.2 Integration into a Multi-Signal Open-Data Stack

The VIX module can plug into a broader equity timing framework by:

- Conditioning factor/sector rotation on VIX regime (e.g., de-leveraging in stress).
- Linking with mutual fund flow-based signals and index momentum for a combined ranking model.
- Using the blended VIX forecasts as a risk-scaling input to control exposure.

This modular structure ensures that improvements in one block (e.g., better open-data flow signals) do not break the transparency of the whole system.

8 Limitations and Future Work

8.1 Limitations

- Reliance on continuous, good-quality high-frequency data.
- Model stability under structural breaks (regulatory changes, liquidity shifts).
- Risk of overfitting if too many features or overly complex trees are used.
- Focus on a single risk indicator; real systems should combine multiple signals.

8.2 Future Extensions

- Introduce probabilistic forecasts (e.g., quantile regression, conformal prediction) for VIX moves.
- Combine VIX-based regimes with open-data signals from factor indices, sector returns, and mutual fund flows.
- Build an interactive dashboard (Python/JS) for educational visualization.
- Perform extensive robustness checks: walk-forward validation, alternative regime definitions, and cost-adjusted backtests.

9 Conclusion

We presented a fully specified, open-data based framework for forecasting India VIX and transforming these forecasts into regime-aware, interpretable signals for Indian equity timing. The pipeline—from minute-level data engineering to model blending and signal interpretation—is transparent and reproducible, making it suitable both as a DES646 course project and as a building block for real-world, retail-friendly decision tools.

Note: All modeling steps are designed to be implementable directly from publicly available data sources and standard Python notebooks, without proprietary data or black-box dependencies.