Machine Learning-Based Predictive Allocation for S&P 500 Excess Returns: A Time-Series Approach Under the Efficient Market Hypothesis

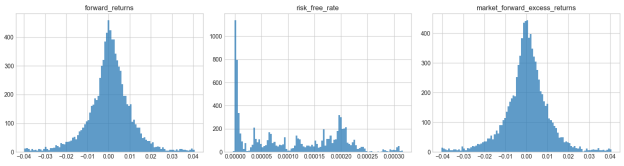
20203955 Park WonKyu, 20214677 Lee YuJung, 20201799 Jung SeungHwan

Intro

This project uses the Hull Tactical Market Prediction dataset to forecast next-day S&P 500 excess returns and build a portfolio with exposure between 0 and 2 under a 120% volatility cap. Because financial returns have near-zero means, high noise, and weak autocorrelation, the task extends beyond standard regression: it evaluates whether any economically meaningful signal exists within strict risk limits. Although the Efficient Market Hypothesis implies minimal predictability, patterns such as volatility clustering and regime shifts suggest that limited structure may persist. Thus, the project investigates how much signal machine learning can extract from noisy market data and how this translates into risk-adjusted performance.

Data Description

The dataset contains 9,021 daily observations and 98 features describing S&P 500 market conditions and related economic factors. The target variable, market_forward_excess_returns, is a processed excess-return series with mean near zero and volatility around 1%, consistent with typical equity-index behavior. Features span macro variables (E), momentum signals (M), volatility measures (V), sentiment indicators (S), and rate-related factors (I). Many variables exhibit multicollinearity and occasional missing values, reflecting structural shocks rather than simple data errors.



Efficient Market Hypothesis

The prediction task directly challenges the weak-form EMH, which asserts that past prices and returns contain no exploitable information about future excess returns. In line with EMH, daily S&P 500 excess returns appear almost random, with a very low signal-to-noise ratio. At the same time, documented phenomena such as volatility clustering, macro shocks, and regime shifts suggest that state variables—rather than return direction itself—may be partially predictable. The project is thus framed as an empirical test of whether machine learning can exploit this limited predictability without violating the spirit of EMH.

EDA and Stylized Factors

Analysis showed that missing values were concentrated in specific macro, volatility, and momentum variables (e.g., E7, V10, S3, M1, M13). Rather than viewing these as simple data-quality issues, they reflect structural properties of financial markets, where certain indicators are reported irregularly or become unavailable during rapid market shifts. Instead of discarding these patterns, we used them to encode “macro shock intervals” by converting them into rolling z-scores and shock indicators. Likewise, outliers were treated not as noise but as signals of extreme returns or macro disruptions, forming the basis for regime-oriented feature construction.

To assess multicollinearity, we computed VIF values for major features. Extremely high VIFs were observed for market_forward_excess_returns and forward_returns (≈7,800), with additional P-series (P10, P11), V-series (V9, V10), and select E-series variables also reaching into the hundreds. This reflects the nature of a single-asset dataset (S&P 500), where multiple features load onto only a few latent factors. These results shaped our modeling strategy: for linear models, PCA followed by ElasticNet was appropriate, while non-linear tree-based models (LightGBM, XGBoost) were

suitable due to their ability to handle interactions and perform implicit dimensionality reduction.

The EDA confirmed four stylized facts:
(1) daily excess returns are nearly directionally unpredictable,
(2) returns show fat-tailed behavior with recurrent extreme events,
(3) volatility varies over time and clusters, and
(4) strong correlations indicate a pronounced latent factor structure.

Feature Engineering

Because financial time series contain strong noise and regime-dependent behavior, raw data alone is insufficient for capturing the subtle signals required to predict excess returns. To address this, we engineered features across five categories.

Lag-based features. Although directional prediction is notoriously difficult, short-term shocks and residual effects may persist during regime transitions. To capture dynamics across multiple horizons, we generated lag values for 1, 2, 5, 10, 21, and 63 days.

Rolling-window statistics. Since volatility clustering is a dominant property of financial markets, we computed rolling means, standard deviations, min/max values, and z-scores over 5, 10, 21, and 63-day windows. These serve as a proxy for the market's signal-to-noise ratio and help detect structural shifts.

Explicit volatility-regime indicators. Volatility is central to market state identification. We compared 21-day vs. 63-day volatility to flag high-volatility periods and defined crisis regimes when the 21-day volatility was above the 90th percentile. We also included a volatility slope (vol_slope) as a leading indicator of rising or falling market risk.

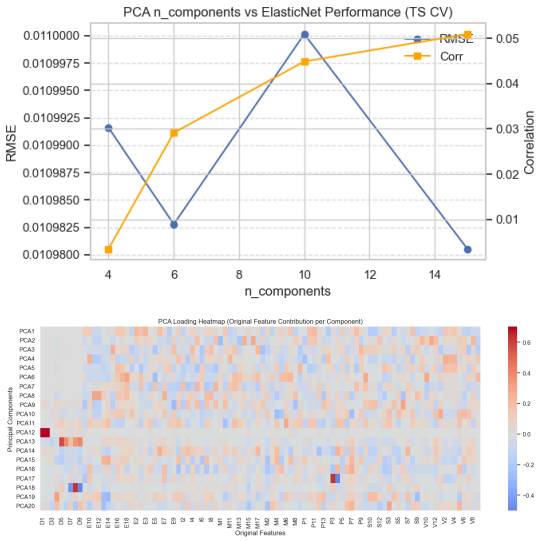
Macro shock and macro regime variables. Since macroeconomic indicators update at different frequencies, deviations from baseline were normalized using rolling z-scores. Absolute z-scores above 2 were interpreted as macro "shocks." When multiple shocks co-occurred, aggregated indicators such as macro_shock_sum and macro_crisis captured broader systemic stress.

Interaction features. Momentum–volatility interactions (M × V) were included to represent well-known factor relationships. We also constructed macro spreads and interest rate spreads (e.g., E2–E11, E7–E12), which contain structural information about yield curve changes and economic cycle transitions. Overall, this feature engineering process encodes market structure into a high-dimensional factor representation, enabling the model to perform **state recognition** rather than relying solely on directional predictions.

Modeling Strategy

We employed a hybrid modeling strategy utilizing both linear and non-linear approaches to address the specific characteristics of financial time series.

Linear Approach (ElasticNet with PCA): To mitigate the severe multicollinearity inherent in the data, we adopted ElasticNet, which applies both L1 and L2 penalties for stable estimation. Crucially, the input data first underwent standardization and dimensionality reduction via PCA. We selected the top 15 principal components, explaining approximately 70–75% of the total variance, to effectively compress the latent factor structure before feeding it into the model.



Non-linear Approach (LightGBM & XGBoost): We utilized LightGBM and XGBoost to capture complex non-linear relationships and interactions. These tree-based gradient boosting models are highly efficient at modeling intricate financial factor structures. Unlike the linear approach, they were trained directly on raw high-dimensional features

generated during the Feature Engineering phase, leveraging their inherent capability to handle unscaled inputs.

Blending Strategy: The core of our strategy lies in blending disparate signals to secure predictive stability. In environments with extremely low signal-to-noise ratios, no single model demonstrates consistent dominance. Therefore, averaging weak predictive signals proved effective in reducing variance. Our final model combines these linear and non-linear outputs to achieve a robust performance that exceeds that of any individual model.

Time-Series Cross Validation

Preventing data leakage is paramount in financial modeling. Standard k-fold cross-validation relies on random splitting, which ignores chronological order and introduces severe look-ahead bias. To mitigate this, we employed a walk-forward validation strategy based on TimeSeriesSplit.

Implementation: In each fold, models were trained on historical data up to a specific cutoff and validated on the immediate subsequent period. We enforced strict separation in preprocessing: for ElasticNet, standardization and PCA were fitted exclusively on the training interval of each fold to ensure no future information leaked into the transformation. Tree-based models (LightGBM, XGBoost) followed the same rigorous temporal segregation.

Significance: This framework faithfully replicates the Out-of-Sample (OOS) environment of real-world backtesting. The validation results derived from this process serve as a robust, logical basis for evaluating model quality and determining the optimal weights for the final ensemble blending.

Model Comparison and Blending

TS-CV results showed that the Constant-Mean Benchmark achieved an RMSE of ~0.0108, while all individual models underperformed it: ElasticNet ~0.0111, LightGBM ~0.0122, and XGBoost ~0.0124—consistent with EMH expectations.

Model	RMSE_mean	RMSE_std	Corr_mean	Corr_std
Constant-Mean Benchmark	0.0108	0.0027	NaN	NaN
ElasticNet	0.0112	0.0028	0.0349	0.0565
LightGBM	0.0122	0.0025	0.0297	0.0479
XGBoost	0.0124	0.0025	0.039	0.0504

However, because model errors were partially uncorrelated, blending offered the potential to reduce idiosyncratic noise. A grid search over OOF predictions identified optimal ensemble weights.

Minimizing RMSE yielded ElasticNet 0.95, LightGBM 0.05, and XGBoost 0.00. Correlation-based optimization converged even more strongly toward ElasticNet alone. These results indicate that non-linear models extracted little meaningful signal, whereas ElasticNet reliably captured weak linear structure via PCA-transformed factors. LightGBM contributed marginally, and XGBoost added no value.

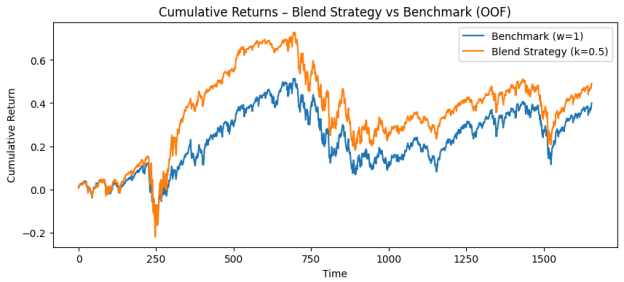
Thus, the final model was a simple ensemble of 95% ElasticNet and 5% LightGBM, a choice supported by TS-CV evidence and designed to avoid overfitting.

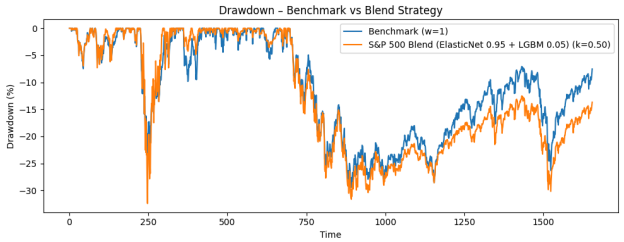
Strategy Evaluation under Volatility Constraint

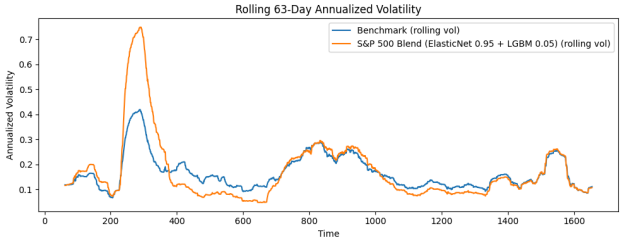
To satisfy project constraints (weights [0, 2], volatility ratio 1.2), we employed a standardized weighting scheme $1 + k \cdot z$ with dynamic rescaling. Grid search optimization identified a conservative k=0.5 as optimal for the Modified Sharpe ratio2, confirming that aggressive weighting amplifies noise in low-signal regimes.

```
Blend Strategy (ElasticNet 0.95 + LGBM 0.05)
Best k: 0.5
mean_return_bench: 0.00026529112000813306
mean_return_strat: 0.00033078310976902305
vol_bench: 0.17661231447784878
vol_strat: 0.21193477643068365
vol_ratio: 1.1999999878675793
sharpe_bench: 0.3785316807080875
sharpe_strat: 0.39331599476954743
final_cumret_bench: 0.3997951143097216
final_cumret_strat: 0.4913627553218616
```

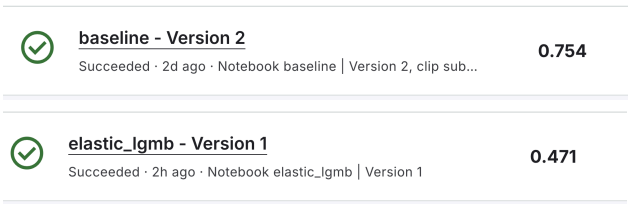
Summarizing the final performance, the strategy recorded a mean excess return of approximately 0.000330, compared to the benchmark's 0.000265. The Sharpe ratio showed a marginal improvement, rising from approximately 0.378 for the benchmark to 0.393 for the strategy. Cumulative returns also improved, increasing from 0.399 to 0.491. The volatility ratio stood at 1.20, converging exactly at the constraint limit. While this margin of improvement is minimal and perhaps insufficient for practical high-frequency trading, it demonstrates that machine learning can capture subtle structural signals within the boundaries permitted by the EMH.







Kaggle Performance Analysis and Discussion



The Public LB underperformance (0.47 vs 0.7) stems from a structural conflict, not model deficiency. Our real-time pipeline pre-loads a 500-day history buffer from the training set. Since the Public LB duplicates this exact training segment, merging the inputs creates a data duplication artifact, severely distorting rolling features. This score drop paradoxically validates that the history buffer is active. We anticipate nominal performance on the Private LB, where inputs are unique future data, eliminating the duplication issue.

Limitations and Future Work

This study has several limitations. Excluding deep learning models such as LSTMs, GRUs, and Transformers reduced our ability to capture long-term dependencies and nonlinear memory effects. Additionally, while PCA helped simplify the feature space, it weakened the interpretability of individual macroeconomic variables.

Future work could incorporate attention-based architectures or alternative datasets—including market microstructure signals and news sentiment—to expand informational predictability. Improvements in rolling-window Time-Series Cross-Validation and meta-learning approaches that directly optimize Sharpe-based metrics also represent promising directions.

Conclusion

This project approached S&P 500 excess-return prediction by emphasizing structural market features—such as volatility regimes and macro shocks—rather than raw directional movements. Consistent with the EMH, no model clearly surpassed the Constant-Mean Benchmark; however, Time-Series Cross-Validation showed that a blended model (95% ElasticNet, 5% LightGBM) produced the most stable signals, indicating that linear models often outperform complex non-linear ones in high-noise settings.

The final strategy achieved a small but measurable Sharpe ratio improvement under volatility constraints. This reinforces that short-term return direction is nearly unpredictable, yet machine learning can still extract limited structural signals within the bounds of market efficiency.