Appendix: Additional Results for KOSPI Excess Return Prediction

20203955 Park WonKyu, 20214677 Lee YuJung, 20201799 Jung SeungHwan

1. Intro

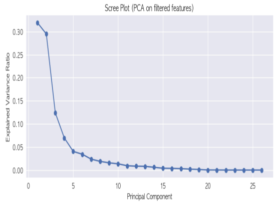
This study extends the S&P 500 excess-return prediction framework to the Korean equity market (KOSPI) to evaluate whether short-horizon predictability and factor structures generalize across markets. The S&P 500 results showed extremely weak but non-zero signals in trend, volatility, and macro features. However, KOSPI differs markedly in microstructure—featuring heavier retail participation, higher short-term volatility, and stronger exposure to global flows—raising the likelihood of even greater unpredictability. Accordingly, this Cross-Market Extension is not merely a dataset substitution, but an empirical assessment of how market structure and information diffusion shape the performance of ML-based forecasting models.

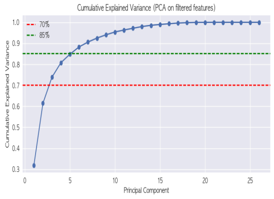
2. Data Description

The KOSPI forecasting dataset was constructed by combining market data obtained from Yahoo Finance with a domestic news sentiment index derived from Korean economic news (Bank of Korea ECOS). Using these collected variables, we generated a wide range of derived features to capture market structure. Detailed descriptions of all variables and engineered features can be found in the dataset card. Most technical indicators were already included and used as provided; we only added one long-term feature, the 5-year rolling mean.

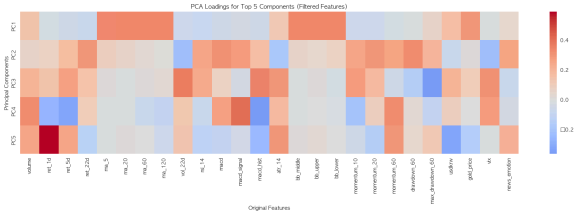
3. PCA Dimension Reduction

The modeling phase applied the key insight obtained from the S&P 500 analysis—namely, that severe multicollinearity and noise dominate financial time-series features, and PCA-based dimensionality reduction is effective under such conditions. The Cross-Market Extension trained PCA exclusively within each training fold of a TimeSeriesSplit, applying only the transform step to validation and test segments. This leakage-free PCA design prevents the model from implicitly accessing future covariance structures, a particularly critical requirement in financial modeling.

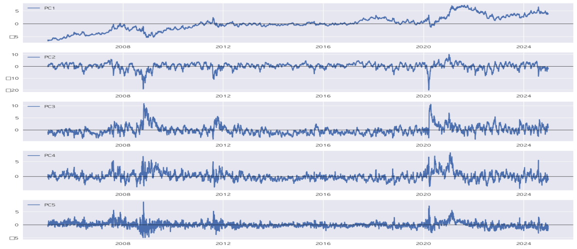




The results showed that the top five principal components explained roughly 85% of total variance.



The first component behaved like a broad market-level beta or long-term trend factor. The second captured volatility-regime dynamics, responding sharply during periods of short-term market stress. The third and fourth components represented momentum–reversal structures and MACD-related short-term trend signals. Notably, the fifth component showed elevated loadings for the news sentiment index and the exchange rate, reflecting Korea-specific psychological and external-shock channels not present in the S&P 500 environment.

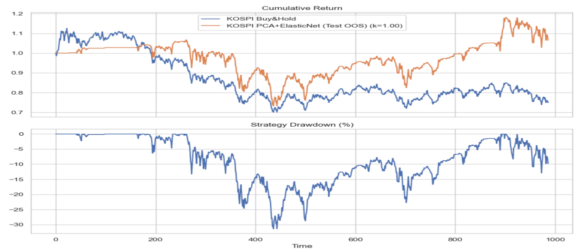


4. Modeling Framework

ElasticNet regression was applied to the PCA factors to forecast next-day excess returns. Its combined L1–L2 regularization offers stable coefficient estimation and removes irrelevant components, making it effective for noisy, high-dimensional financial data. In contrast to the S&P 500 results, nonlinear tree models such as LightGBM and XGBoost failed to extract meaningful signals in KOSPI, while ElasticNet remained the most reliable. This indicates that nonlinear models overfit more easily when short-horizon signals are extremely weak, consistent with KOSPI's higher noise-to-signal ratio.

Model	RMSE _mean	RMSE _std	Corr _mean	Corr _std
ElasticNet_raw	0.0096	0.0023	0.0638	0.0639
ElasticNet_PCA	0.0096	0.0023	0.0670	0.0483
LightGBM	0.0116	0.0040	0.0785	0.0724
XGBoost	0.0115	0.0033	0.0679	0.0488

The predicted signal was used to construct a dynamic asset-allocation strategy. Portfolio weights were designed to scale linearly with the standardized predictive signal and were clipped between 0 and 2, in accordance with the assignment guidelines. To satisfy the core project constraint—strategy volatility ≤ 1.2× benchmark volatility—the initial position series was proportionally rescaled. This adjustment effectively controlled exposure during volatility-regime shifts and resulted in a strategy resembling market-timing behavior through time-varying beta adjustments.



The fully out-of-sample backtest showed clear differences from the benchmark. While Buy-and-Hold generated a negative annualized return, the PCA+ElasticNet strategy achieved a positive 3.46% return with a Sharpe-variant of 0.18, indicating that the model captured a weak yet non-zero signal in the Korean market. Annualized volatility was 19.2%, exactly satisfying the 1.2× volatility cap, and the ~—31% maximum drawdown was comparable to the market's, as the strategy adjusts exposure rather than performing security selection.

These results offer several implications. First, the presence of small but detectable predictability in Korean excess returns serves as a mild challenge to the strict weak-form EMH. Second, the PCA–ElasticNet combination again functions effectively as a factor-extraction and noise-reduction method, suggesting cross-market robustness. Third, the inclusion of the news sentiment index—absent in the S&P 500 data—demonstrated its influence on short-term volatility and factor loadings, highlighting the heightened role of psychological forces in KOSPI. Finally, the study shows how ML-based forecasting adapts to different market microstructures while still meeting practical constraints such as volatility limits.

5. Conclusion

In summary, this cross-market extension shows that a forecasting strategy developed for the S&P 500 can partially transfer to the Korean market. Predictability is weaker but non-zero; sentiment and macro shocks play a larger role; and nonlinear models lose effectiveness in higher-noise environments. These findings help explain how structural differences across markets affect the generalizability of ML-based prediction frameworks. Future work may extend the analysis to NASDAQ, cryptocurrencies such as BTC, or multi-asset settings to evaluate predictability across diverse financial environments.