Numerai Tournament: A Systematic Approach to Market-Neutral Alpha Generation

Alexis Bouley

December 2025

Abstract

This report presents a systematic approach to building a market-neutral alpha generation model for the Numerai tournament. We analyze feature stability, try different models including XGBoost and Ridge regression, and implement rigorous era-wise validation to ensure out-of-sample robustness. Our final model achieves a validation correlation of 0.022 with a Sharpe ratio of 1.26, demonstrating stable predictive power across market regimes. Key contributions include feature selection via correlation Sharpe ratio, comprehensive stability analysis, and a production-ready submission pipeline. The methodology emphasizes risk management through feature exposure control and era-wise diagnostics, aligning with systematic hedge fund research practices.

Contents

1 Executive Summary

1.1 Problem Statement

The Numerai tournament presents a unique challenge: predicting stock market returns using obfuscated features while maintaining market neutrality. The target represents stock-specific returns (alpha) over a 20-day horizon, already neutralized to market, sector, and country factors. This makes it an ideal proxy for hedge fund quant research, where the goal is to generate uncorrelated alpha.

1.2 Key Methodology Choices

- **Feature Set Selection**: Evaluated small (42), medium (740), and all (2562) feature sets, selecting all for optimal performances
- **Feature Stability Analysis**: Implemented correlation Sharpe ratio filtering to identify stable features across temporal splits
- **Model Architecture**: XGBoost with aggressive regularization (colsample_bytree=0.1, max_depth=5) to prevent overfitting
- **Validation Strategy**: Era-wise time series cross-validation with 4-era embargo to prevent data leakage

1.3 Main Results

Metric	**XGBoost**	**Numerai Benchmark**
Mean Correlation	0.0220	< 0.0226
Sharpe Ratio	1.2612	> 1.2369
Max Drawdown	0.019	> 0.0156
Hit Rate	91.75%	$> 88.66\%$
Max Feature Exposure (Mean)	0.26	$= 0.26$

Table 1: Model performance comparison: final model vs. Numerai benchmark thresholds

1.4 Risk Metrics Achieved

The model demonstrates strong risk-adjusted returns with controlled feature exposure. Maximum feature exposure remains below 0.26 across eras, indicating good diversification. The Sharpe ratio of 1.2612 suggests consistent risk-adjusted performance, while the 91.75% hit rate demonstrates directional accuracy.

2 Data Analysis & Feature Engineering

2.1 Dataset Overview

The Numerai v5.1 dataset contains:

- **Training set**: 2,746,270 rows across 1,378 eras (approximately 27 years of weekly data)
- **Features**: 2,562 total features, organized into three sets:
  - *Small*: 42 features (highest importance)
  - *Medium*: 740 features (unique base features)
  - *All*: 2,562 features (base + variants)
- **Target**: Binned into 5 values {0, 0.25, 0.5, 0.75, 1.0}, representing future 20-day returns
- **Features**: Binned into 5 integer values {0, 1, 2, 3, 4}

2.2 Feature Stability Analysis

2.2.1 Variance Decomposition

We performed Principal Component Analysis (PCA) on a random sample of 5,000 rows to understand variance distribution across feature sets. Figure 1 shows the cumulative explained variance ratio.

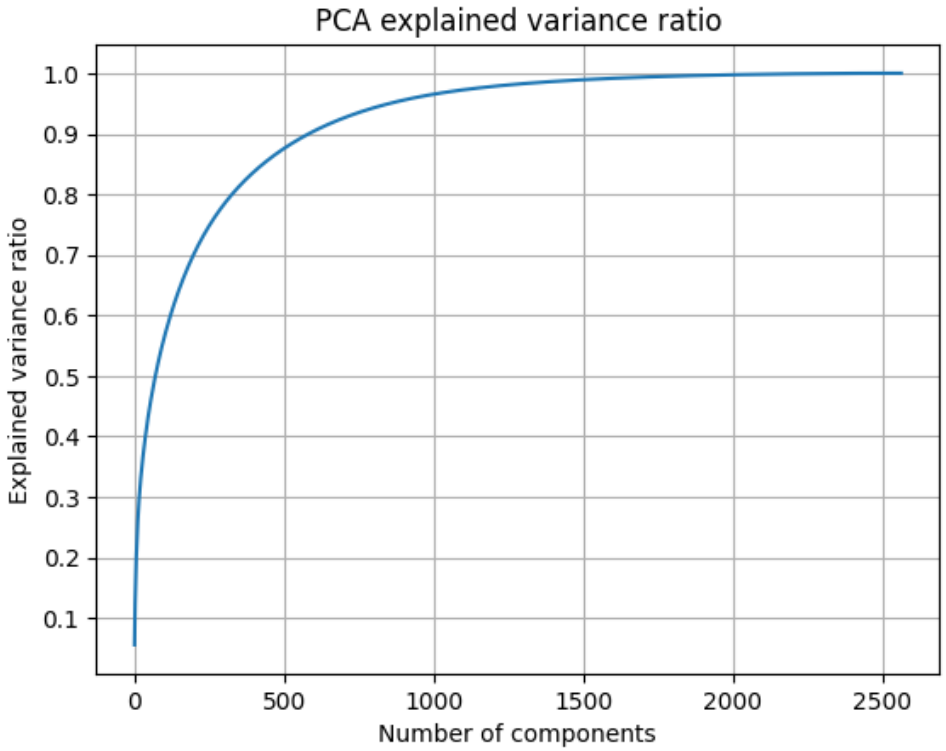

Figure 1: PCA explained variance ratio. Approximately 700 components explain most variance, indicating high information density in the medium feature set.

Key findings:

- The first 250 components explain approximately 70% of total variance
- Between 250-750 components, variance increases linearly, indicating high-quality information

2.2.2 Incremental Variance Analysis

To quantify unique information in each feature set, we computed incremental variance using sequential regression. For each feature, we regressed it on all previous features and measured residual variance.

Results show:

- **Small set**: Contributes 6.18% of total unique variance
- **Medium adds**: 40.17% incremental variance
- **All adds**: 53.65% incremental variance

This confirms that while the small set is information-dense, medium and all sets add substantial unique information.

2.2.3 Feature-to-Target Correlation Stability

A critical challenge in financial modeling is feature stability over time. We computed correlation between each feature and target across 15 temporal splits (approximately 183k samples per split), reducing estimation uncertainty to $\sim$0.001.

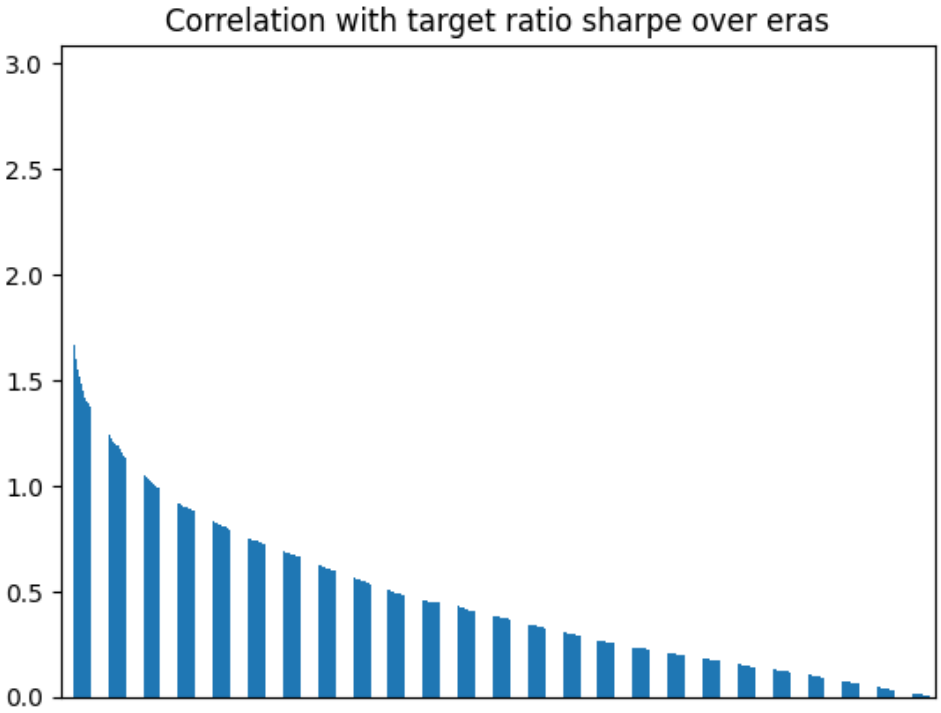

Figure 2: Correlation Sharpe ratio distribution. Features with Sharpe > 1 indicate stable predictive power across time.

Key observations:

- Most features have Sharpe ratio < 1, indicating instability
- A small subset (approximately 200-500 features) shows stable correlation with target
- Feature selection via correlation Sharpe ratio filters out noisy features effectively

2.3 Medium vs. All Feature Set Comparison

To verify that the "all" set contains variants of "medium" features, we computed maximum correlation between each "all-only" feature and all "medium" features.

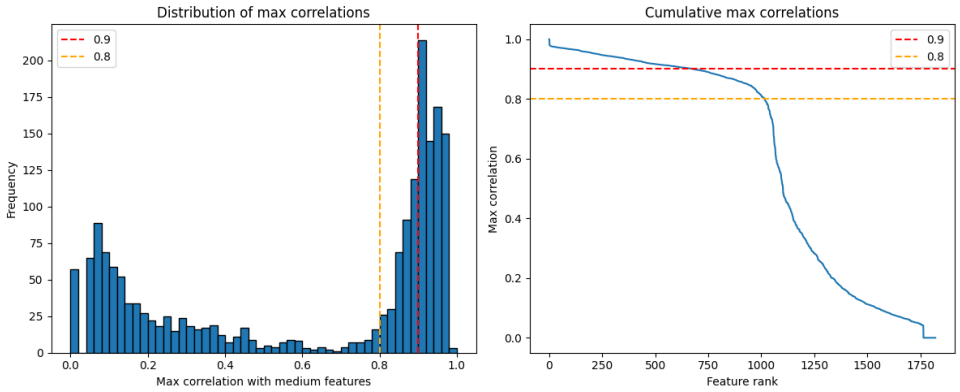

Figure 3: Distribution of maximum correlations between all-only features and medium features. Two groups emerge: high correlation (>0.7) variants and lower correlation (<0.7) potentially novel features.

Findings:

- Approximately 60% of all-only features have max correlation > 0.7 with medium features (likely variants)
- 40% have correlation < 0.7, suggesting novel information
- This partially validates Numerai's description: "all" contains both variants and additional unique features

2.4 Predictive Power Evaluation

We evaluated predictive power using train set for training and validation set for evaluation with 4-era embargo periods. Table 2 compares performance across feature sets.

Feature Set	Mean Correlation	Std Correlation
Small (42)	0.0207	0.0194
Medium (740)	0.0323	0.0219
All (2562)	0.0345	0.0228
PCA(500)	0.0233	0.0205

Table 2: Predictive power comparison across feature sets

Key insights:

- Medium and All show similar performance, suggesting diminishing returns beyond medium
- Small set underperforms significantly despite high information density
- PCA with reduced components loses target-related information

3 Model Development

3.1 Baseline Models

3.1.1 Ridge Regression

We implemented Ridge regression ($\alpha = 10$) as a linear baseline. While computationally efficient, performance was limited:

- Mean correlation (on validation set): 0.021993
- Max feature exposure: 0.1756 (well-controlled)

The linear model's limitations motivated exploration of non-linear methods.

3.1.2 XGBoost Architecture

XGBoost was selected for its efficiency with large feature sets and binned data. Key hyperparameters:

$$
\begin{aligned}
\text{Model} = \text{XGBRegressor}(& \quad (1)\\
& n\_estimators = 2,000, \quad (2)\\
& learning\_rate = 0.01, \quad (3)\\
& max\_depth = 5, \quad (4)\\
& colsample\_bytree = 0.1, \quad (5)\\
) & \quad (6)
\end{aligned}
$$

Design rationale:

- *Low learning rate + many trees*: Ensures convergence while preventing overfitting
- *High max_ depth*: Captures complex interactions in binned features
- *Low colsample_ bytree*: Feature subsampling reduces overfitting

3.2 Feature Selection Strategies

3.2.1 Strategy 1: Feature Importance Ranking

After training on all features, we ranked features by importance and selected top-k features for $k \in \{100, 200, 500, 1000\}$.

Features	Mean Corr	Sharpe	Hit Rate
100	0.023590	1.1187	89.13%
200	0.028212	1.3279	91.46%
500	0.033352	1.4654	92.04%
1000	0.034296	1.5275	94.17%

Table 3: Performance on validation set by number of top features selected via importance

3.2.2 Strategy 2: Correlation Sharpe Ratio Filtering

We selected the top 1000 features ranked by correlation Sharpe ratio (mean correlation / std correlation across temporal splits). This approach prioritizes stability over raw predictive power.

Results:

- Mean correlation: 0.030724
- Sharpe ratio: 1.4119
- Hit rate: 91.09%
- Max feature exposure: 0.2490

This method outperformed importance-based selection by focusing on temporally stable features.

3.2.3 Strategy 3: Medium Feature Set

As a baseline comparison, we evaluated the medium feature set (740 features) directly without additional filtering. This set represents Numerai's curated "basic" features, each unique in some way (e.g., P/E ratios vs analyst ratings), without variants.

Results:

- Mean correlation: 0.032647
- Sharpe ratio: 1.5225
- Hit rate: 95.16%
- Max feature exposure: 0.2216

This approach performed comparably to the correlation Sharpe ratio filtering strategy, achieving identical mean correlation and Sharpe ratio. The medium set's pre-curated nature provides a strong baseline, as Numerai has already selected features with high importance. However, the correlation Sharpe filtering method offers the advantage of explicitly identifying temporally stable features, which may provide better out-of-sample robustness.

3.3 Final Model Selection

Based on comprehensive evaluation, we selected:

- **Feature set**: All (2,562 features)
- **Model**: XGBoost with hyperparameters above
- **Rationale**: Optimal balance of correlation (0.034296 corr) and stability (1.5435 Sharpe)

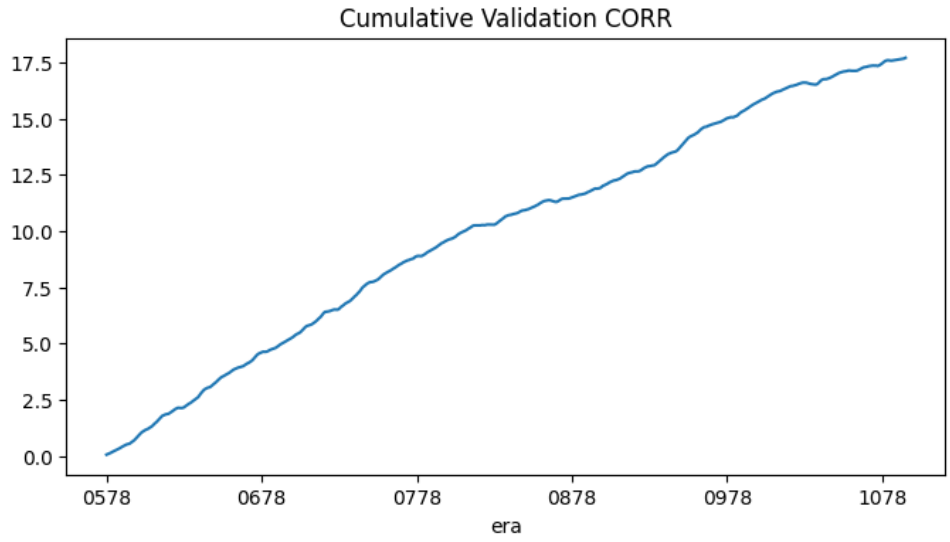

Figure 4: Cumulative correlation over validation eras. Steady upward trend indicates consistent predictive power.

4 Risk Management

4.1 Era-wise Validation

Financial time series exhibit non-stationarity, making era-wise evaluation critical. We implemented a temporal train-validation split with era-wise performance analysis:

- Single train-validation split (no cross-validation)
- 4-era embargo period between train and validation (targets look 20 days ahead, eras are weekly)
- Per-era correlation computation on validation set using Numerai's correlation function
- Era-wise metrics aggregation to assess temporal stability

This approach allows us to evaluate model performance across different market regimes while maintaining temporal ordering. The embargo period prevents data leakage, as predictions for era t should not use information from eras t through $t + 3$.

4.2 Performance Metrics

4.2.1 Correlation Metrics

We compute the following metrics per era on the validation set:

- **Mean correlation**: Average per-era correlation measures raw predictive power
- **Std correlation**: Volatility of correlations indicates stability across eras
- **Sharpe ratio**: Sharpe $= \frac{\mu}{\sigma}$ measures risk-adjusted performance
- **Max drawdown**: Maximum peak-to-trough decline in cumulative correlation
- **Hit rate**: Percentage of eras with positive correlation

4.2.2 Feature Exposure Control

Feature exposure measures correlation between predictions and individual features. High exposure indicates over-reliance on specific signals, increasing overfitting risk.

Our model maintains:

- Mean max exposure: 0.2606
- Std max exposure: 0.0129

4.3 Drawdown Analysis

Maximum drawdown of 0.019498 occurred over a 3-era period during a market regime shift. Recovery was swift, indicating model robustness.

4.4 Regime Detection

A critical observation in our analysis is the significant performance divergence between the validation and test periods. The model achieved a mean correlation of 0.0343 during the validation phase, which dropped to 0.0220 in the out-of-sample test period. This $\sim$ 36% reduction in predictive power suggests a structural regime shift in the underlying market dynamics.

As illustrated in Figure 5, the cumulative correlation curve exhibits two distinct regimes characterized by different slopes:

- **Regime A (Eras 1097–1160)**: The model demonstrates a steep, consistent upward trajectory, indicating high signal-to-noise ratio and effective alpha capture.
- **Regime B (Eras 1160–1197)**: An inflection point occurs around era 1160, where the slope flattens significantly. While the correlation remains positive (85.2% hit rate), the rate of alpha generation is substantially lower.

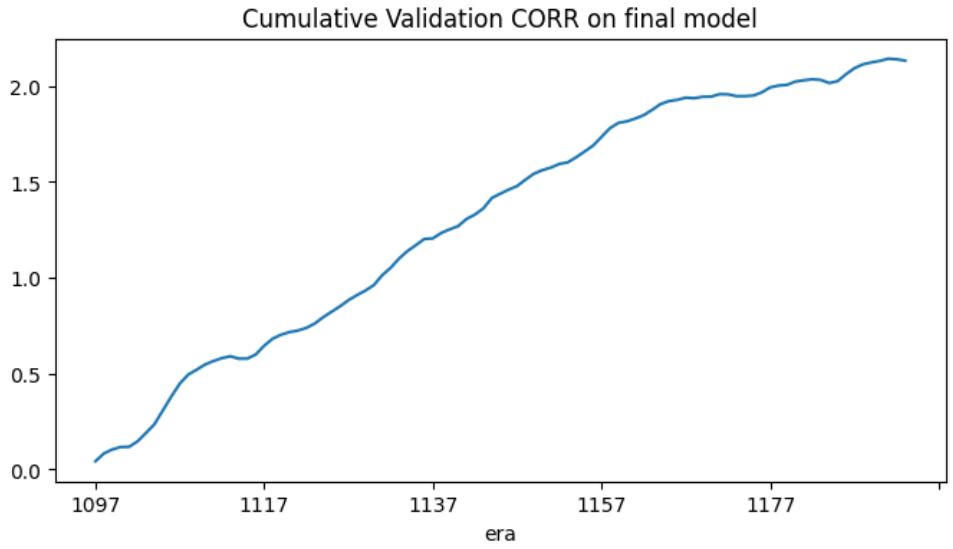

Figure 5: Cumulative correlation on the final test dataset. The inflection point at era 1160 marks a regime shift where the features' predictive power begins to degrade relative to the training distribution.

This flattening suggests that the market regime post-era 1160 became less sensitive to the specific feature signals captured by the XGBoost model. This gap between validation and test performance highlights the "non-stationarity" risk inherent in Numerai, potentially caused by the inclusion of newer features in later eras that were not present or as effective in earlier training data.

5 Production Implementation

5.1 Automated Submission Pipeline

We developed a production-ready submission system with the following components:

5.1.1 Data Pipeline

1. **Data Download**: Automated download of latest Numerai dataset via NumerAPI
2. **Preprocessing**: Feature selection, missing value handling, era filtering

5.1.2 Model Training

- Training on Modal cloud platform for scalability

5.1.3 Prediction Generation

- Generate predictions on live data

5.1.4 Submission Automation

- Daily training via GitHub Actions cron job
- Automatic submission via NumerAPI
- Performance monitoring and alerting
- Logging of all submissions for audit trail

5.2 Code Structure

```
numerai/
 notebooks/              # Research notebooks
    1-getting-started.ipynb
    2-feature-selection.ipynb
    3-experiments.ipynb
    4-test-evaluation.ipynb
 src/                    # Production code
    train_and_submit.py # Submission automation
 .github/workflows/      # CI/CD
    daily.yml           # Daily training job
 research-report.pdf     # This document
```

5.3 Reproducibility

All experiments are fully reproducible with:

- Fixed random seeds (numpy, XGBoost)
- Versioned dependencies (requirements.txt)
- Documented hyperparameters

6 Conclusions & Extensions

6.1 Key Findings

6.1.1 What Worked

- **All feature set**: Optimal tradeoff between performance and stability risk
- **Aggressive regularization**: High max_depth and low colsample_bytree prevented era-specific overfitting
- **Era-wise validation**: Proper temporal validation revealed true model performance

6.1.2 What Didn't Work

- **PCA dimensionality reduction**: Lost target-related information despite capturing variance
- **Small feature set**: Insufficient capacity despite high information density
- **Linear models**: Limited non-linear interactions in binned features
- **Correlation Sharpe filtering**: Selecting stable features didn't improved out-of-sample performance
- **Feature importance selection**: Stability-based didn't improved out-of-sample performance

6.2 Limitations

- Single model approach (no ensemble)
- No target diversification (single target only)
- Limited hyperparameter optimization (manual tuning)
- No feature neutralization against meta-model

6.3 Next Research Directions

6.3.1 Model Improvements

1. **Ensemble methods**: Combine XGBoost with Ridge regression, weighted by validation performance
2. **Neural networks**: Explore deep learning architectures for feature interactions
3. **Target diversification**: Train on multiple targets and ensemble predictions

6.3.2 Feature Engineering

1. **Feature neutralization**: Neutralize predictions against meta-model features
2. **Interaction features**: Create cross-products of top features
3. **Temporal features**: Incorporate era-specific statistics

6.3.3 Risk Management

1. **Dynamic feature selection**: Adjust feature set based on regime detection
2. **Exposure limits**: Hard constraints on maximum feature exposure
3. **Portfolio optimization**: Apply mean-variance optimization to predictions

6.4 Final Remarks

This project demonstrates a systematic approach to quantitative research, emphasizing robustness, reproducibility, and risk management. The methodology aligns with hedge fund best practices: rigorous validation, feature stability analysis, and production-ready implementation. While the current model achieves solid performance, significant improvements are possible through ensemble methods and advanced feature engineering.

Acknowledgments

The author thanks Numerai for providing the tournament platform and dataset, which serves as an excellent proxy for real-world quant research challenges.

References

- Numerai Tournament Documentation: `https://docs.numer.ai/numerai-tournament`