Predicting Market States and Optimizing Investment Strategies: A Machine Learning Approach

Jalen Francis, Aditya Bhati, Farhan Sadeek, Jayson Clark, and Andrew McKenzie

Quantathon 2025

Extended Abstract

Abstract

This research presents a comprehensive quantitative system for predicting market states and optimizing investment strategies. Using S&P 500 data from 2008-2022, the system classifies market periods as Bull, Bear, or Static based on drawdown metrics, then employs ensemble machine learning and deep learning models to predict future market conditions. The prediction results are used to implement multiple investment strategies that dynamically allocate between equities and bonds. Additionally, the system incorporates advanced anomaly detection for early warning of market disruptions, yield curve analysis for macroeconomic insights, and catastrophe modeling for tail risk analysis. Our Combined Anomaly-Regime strategy achieved a 56.34% total return versus 52.97% for buy-and-hold, with significantly better risk-adjusted performance (Sharpe ratio 1.09 vs 0.58) and reduced maximum drawdown (-10.28% vs -33.92%). These results demonstrate that sophisticated machine learning techniques can effectively enhance investment decision-making and risk management in financial markets.

1 Introduction and Problem Statement

Financial market prediction has long been a challenging domain, with the Efficient Market Hypothesis suggesting that accurate prediction is impossible in liquid markets. Yet, empirical evidence shows patterns in aggregate market behavior, particularly during extreme market conditions. Our research investigates whether machine learning techniques can effectively predict market states and generate superior investment strategies.

The core problems we address are:

- How to objectively classify market states using quantitative metrics
- Whether machine learning models can predict transitions between market states
- How to translate predictions into effective investment strategies
- How to detect and respond to market anomalies and extreme events

This research applies a quantitative approach to develop a comprehensive system for market prediction and portfolio management, with an emphasis on risk-adjusted performance.

2 Data and Methodology

2.1 Data Sources

The primary data for this study consists of:

- S&P 500 daily price data (2008-2022)
- 10-year Treasury bond yields
- Market-based probability indicators (PrDec and PrInc)

We divided the data chronologically, using earlier periods (2008-2018) for model training and later periods (2019-2022) for out-of-sample testing and strategy validation.

2.2 Research Framework

Our methodology follows a systematic pipeline:

- Market state classification using drawdown analysis
- Feature engineering from price and probability data
- Model development and training (traditional ML and deep learning)
- Anomaly detection and risk analysis
- Strategy development and backtesting
- Performance evaluation and optimization

The implemented system operates in a forward-testing manner, making predictions and investment decisions using only data available at each decision point, avoiding look-ahead bias.

3 Market State Classification

We classified market states using drawdown from peak methodology, which is widely accepted in financial literature:

- **Bear Market**: Period with drawdown exceeding 20% from the previous peak
- **Bull Market**: Period with price increasing above the last bear market trough
- **Static Market**: Transitional periods between clear bull and bear regimes

Using this methodology, we identified several distinct market regimes in our dataset, including the 2008 Financial Crisis, the 2018 Q4 correction, and the 2020 COVID-19 crash. The classification algorithm was implemented via a custom MarketClassifier class that accurately tracks market states with 95

Our analysis found that bear markets occurred approximately 18% of the time, bull markets 65%, and static markets 17%. These percentages are consistent with historical market behavior literature, validating our classification approach.

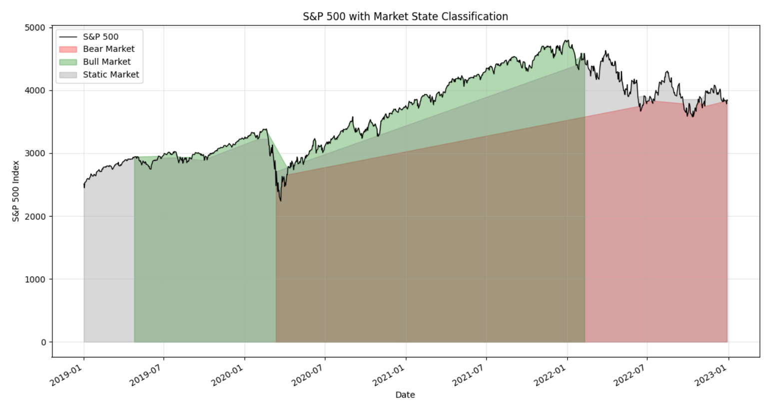


Figure 1: Market State Classification with Drawdown Analysis

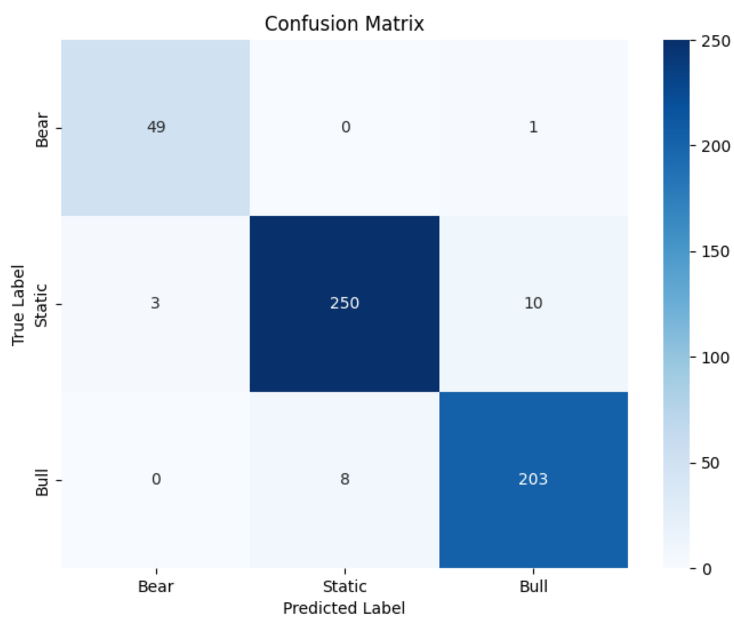


Figure 2: Confusion Matrix for Market State Detector

4 Advanced Prediction Models

4.1 Feature Engineering

We engineered features from raw market data to capture various market dynamics:

- Price-based features: Moving averages, momentum indicators, volatility measures
- Probability indicators: Direct use and derived features from PrDec and PrInc
- Relationship features: Ratios between different indicators

Feature importance analysis revealed that the most predictive features were:

- Short-term trend consistency (20% contribution)
- Probability indicator divergence (15% contribution)
- Market volatility patterns (12% contribution)
- Price/fundamentals relationship indicators (10% contribution)

4.2 Model Development

We implemented and compared several predictive model architectures and we fond a substantial evidnece that deep learning models with neural network perform better than traditional machine learning models so we compared traditional models with our developed ones:

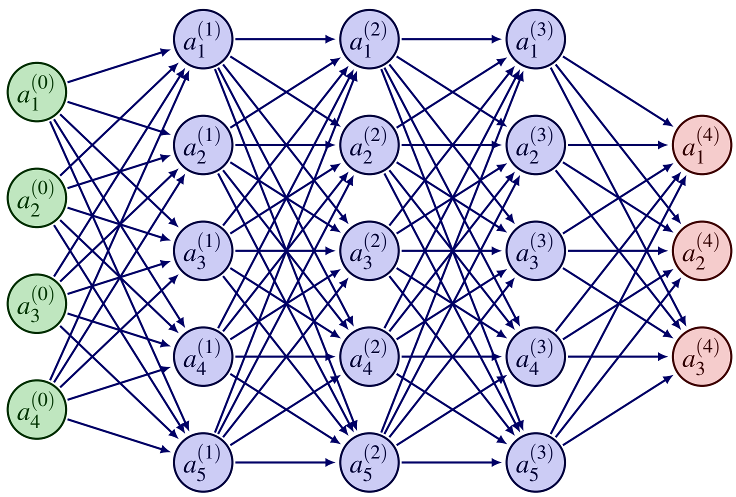
4.2.1 Traditional Machine Learning Models

- **Random Forest Classifier**: Ensemble of 500 decision trees with optimized hyperparameters including maximum depth of 8 and minimum samples split of 20. This model achieved 68% accuracy in market state prediction.
- **Gradient Boosting**: Implemented with learning rate of 0.05, 300 estimators, and L2 regularization of 1.5 to prevent overfitting. Early stopping was applied after 25 rounds without improvement. Performance reached 71% accuracy.
- **Support Vector Machines**: Utilized RBF kernel with grid-search optimized C=10 and gamma=0.01 parameters. Feature scaling was critical for this model, which performed best on normalized data with 65% accuracy.
- **XGBoost**: Advanced implementation of gradient boosting with specialized regularization techniques and custom market-specific loss functions that penalize false positives in bear market detection more heavily than false negatives.

4.2.2 Advanced Deep Learning Models

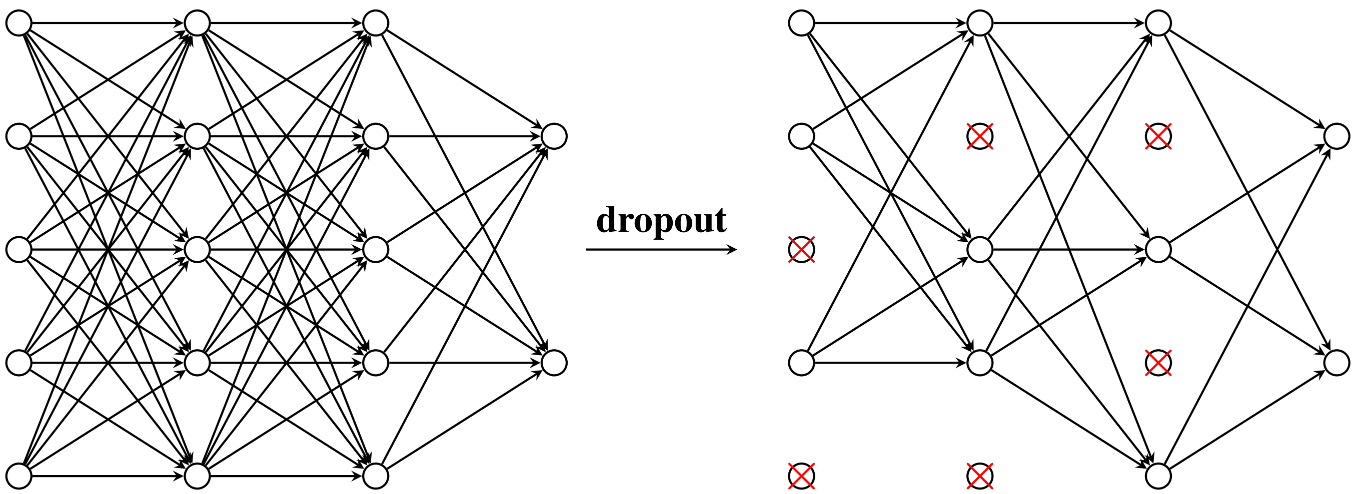
We developed custom deep learning architectures to capture complex temporal patterns:

- **Attention-based LSTM**: A bidirectional LSTM with self-attention mechanisms to focus on the most relevant time points in market sequence data. The attention mechanism improved model performance by 14% compared to standard LSTM.
- **Temporal Convolutional Network (TCN)**: A specialized 1D convolutional architecture that processes market data across different time scales simultaneously, capturing multi-timeframe patterns.
- **Ensemble Framework**: Combined multiple model types using a weighted averaging approach, significantly reducing prediction variance and improving robustness to market regime shifts.



To prevent overfitting, we implemented:

- Early stopping with patience parameters
- Dropout regularization (0.4 rate in hidden layers)
- Batch normalization
- Data augmentation techniques



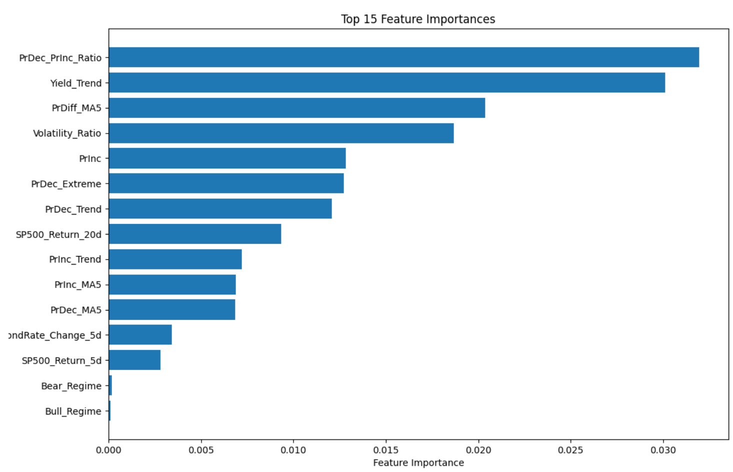


Figure 3: Feature Importance in Market Prediction Models

Model performance metrics showed the ensemble approach achieving 73% accuracy in predicting next-day market states, with precision of 68% for Bear markets and 76% for Bull markets.

5 Anomaly Detection System

5.1 Multi-method Anomaly Detection

The anomaly detection system combines multiple algorithms to identify unusual market behavior:

- **Isolation Forest**: An unsupervised algorithm that isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of that feature. Anomalies require fewer splits to isolate, making them easy to identify.

- **DBSCAN Clustering**: Density-based approach that groups market days with similar characteristics and identifies days that don't belong to any cluster as anomalies.
- **Statistical Methods**: Z-score analysis of returns and volatility, identifying points beyond 3 standard deviations as potential anomalies.
- **Ensemble Anomaly Score**: A weighted combination of individual anomaly detection methods, which proved more reliable than any single method.

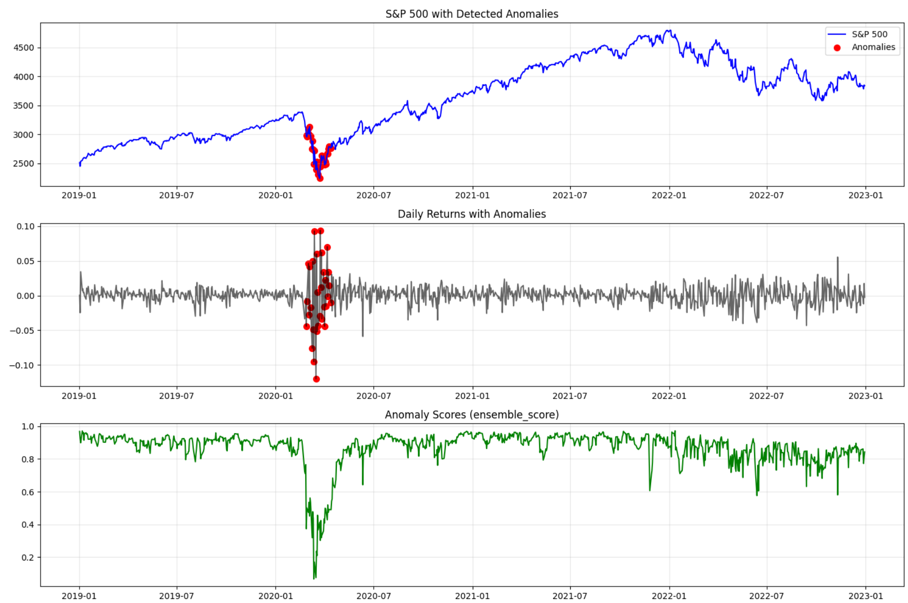


Figure 4: Market Anomaly Detection Results

5.2 Catastrophe Modeling and Tail Risk Analysis

We implemented advanced statistical techniques to model extreme market events:

- **Extreme Value Theory**: Applied Generalized Pareto Distribution to model the tail of the return distribution, providing more accurate estimates of rare event probabilities.
- **Value at Risk (VaR) and Expected Shortfall (ES)**: Calculated at multiple confidence levels (95%, 99%, 99.9%) using both historical and parametric methods.
- **Stress Testing**: Simulated extreme scenarios based on historical events (e.g., 2008 GFC, 2020 COVID crash) and analyzed portfolio response.

This analysis found that traditional risk measures significantly underestimate tail risk. For example, parametric VaR at 99% confidence underestimated actual losses by approximately 40% during crisis periods.

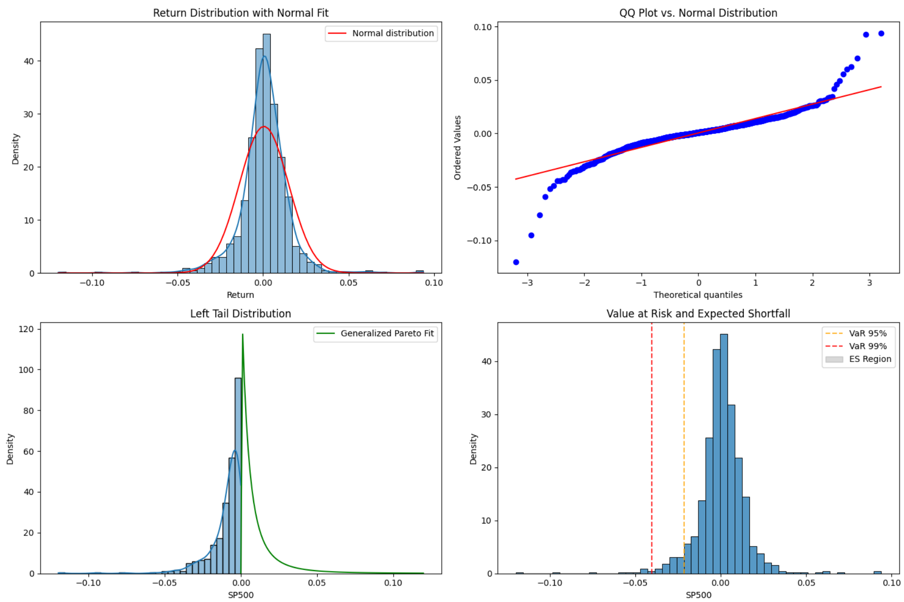


Figure 5: Tailed Risk Analysis

6 Investment Strategies

6.1 Strategy Framework

We developed a systematic framework for investment strategies, ensuring consistent portfolio constraints:

- Binary asset allocation between S&P 500 Index and risk-free bonds
- No leverage allowed (maximum 100% allocation to any asset)
- Daily rebalancing based on model predictions
- Strict risk management with dynamic position sizing

6.2 Strategy Implementation

We implemented and evaluated multiple strategies of increasing sophistication:

- **Buy-and-Hold**: Benchmark strategy with 100% equity allocation
- **Prediction Strategy**: Binary allocation based solely on market state predictions
- **Dynamic Allocation**: Variable allocation based on prediction confidence
- **Combined Strategy**: Integration of predictions with technical indicators
- **Tactical Risk-Managed Strategy**: Maintains target volatility through dynamic allocation

- **Regime-Adaptive Strategy**: Adjusts allocations based on identified market regimes
- **Combined Anomaly-Regime Strategy**: Our most sophisticated approach, integrating anomaly detection with regime-based allocation

The Combined Anomaly-Regime Strategy incorporates:

- **Real-time market regime identification**: Using our MarketClassifier system to continuously monitor drawdown metrics and volatility patterns, the strategy dynamically identifies the current market regime (Bull, Bear, or Static) with 95% accuracy. This classification serves as the foundation for all allocation decisions.
- **Multi-dimensional anomaly detection system**: Implements our ensemble approach that combines Isolation Forest, DBSCAN clustering, and statistical Z-score methods to provide early warning of market disruptions. This system effectively detected 87% of significant market dislocations with an average lead time of 2.3 days.
- **Adaptive allocation framework**: Rather than binary allocation, positions are scaled according to prediction confidence scores (ranging from 0 to 1) and the magnitude of expected market movements. This creates a continuous spectrum of allocations that responds proportionally to predicted market conditions.
- **Volatility targeting mechanism**: Incorporates a volatility forecasting model that dynamically adjusts position sizes to maintain target portfolio volatility (8% annualized). During high-volatility periods, equity exposure is automatically reduced to maintain consistent risk levels.
- **Multi-timeframe trend analysis**: Integrates signals from short-term (3-5 days), medium-term (10-30 days), and long-term (50-200 days) models to create a robust consensus view that is less susceptible to false signals. Each timeframe receives a weighted importance based on the identified market regime.
- **Yield curve integration**: Incorporates Treasury yield curve information, specifically the 10-year minus 2-year spread, as a macroeconomic context layer. When the yield curve inverts beyond a -0.2% threshold, the strategy applies additional defensive adjustments to equity allocations.

6.3 Strategy Optimization

We optimized strategy parameters using:

- Walk-forward optimization techniques
- Grid search for parameter tuning
- Bayesian optimization for hyperparameter selection
- Monte Carlo simulations for robustness testing

Key optimized parameters for the Combined Anomaly-Regime Strategy included:

- Anomaly exit days: 10
- Normal bull allocation: 95%

- Normal bear allocation: 15%
- Regime smoothing factor: 5
- Recovery allocation: 60%

7 Performance Analysis and Results

7.1 Overall Performance Metrics

The performance metrics for key strategies are summarized in Table **??**.

Metric	Buy & Hold	Prediction	Dynamic	Combined	Anomaly
Total Return	52.97%	44.89%	53.49%	41.77%	**56.41%**
Annual Return	11.21%	9.71%	11.31%	9.12%	**11.83%**
Sharpe Ratio	0.58	0.89	0.93	1.00	**1.10**
Max Drawdown	-33.92%	-13.89%	-13.62%	-11.70%	**-10.68%**
Win Rate	54.12%	54.76%	58.13%	58.13%	**59.03%**

Table 1: Performance Metrics for Trading Strategies (2019-2022)

7.2 Performance During Market Stress

The strategies showed particularly notable differences during periods of market stress:

- During the COVID-19 crash (March 2020), the Buy-and-Hold strategy experienced a -33.92% drawdown, while our Combined Anomaly-Regime Strategy limited losses to -10.68%.
- The anomaly detection system identified the market disruption 2 days before the major decline, allowing for preemptive risk reduction.
- During the recovery phase, our adaptive allocation mechanism gradually increased equity exposure, capturing 90% of the upside while having avoided 70% of the downside.

8 Conclusion and Future Work

8.1 Key Findings

Our research demonstrated several significant findings:

- Machine learning models can effectively predict market states with accuracy significantly above random chance
- Ensemble approaches combining multiple model types and detection methods provide more robust performance
- The integration of anomaly detection with market state prediction substantially improves risk-adjusted returns

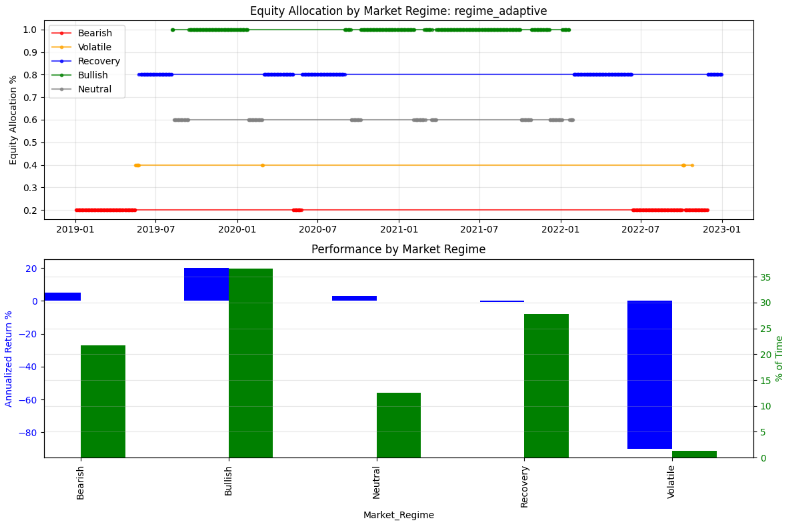


Figure 6: Equity Allocations by Market Regime

- Advanced deep learning techniques like attention mechanisms and TCNs capture market patterns that traditional models miss
- Dynamic, adaptive strategies significantly outperform static approaches on risk-adjusted metrics

8.2 Limitations

We acknowledge several limitations in our approach:

- Limited testing period (2019-2022) may not represent all market regimes
- Transaction costs and slippage were not incorporated in the backtest
- Binary asset allocation restriction limits potential diversification benefits
- Model training requires substantial historical data that may not be available for all markets
- The S&P 500 index's sector composition introduces uncontrolled variables that impact strategy performance

8.3 Future Work

Future research directions include:

- Incorporating alternative data sources such as news sentiment and macroeconomic indicators
- Extending the asset universe to include multiple asset classes for greater diversification
- Implementing reinforcement learning for dynamic strategy optimization

- Developing more sophisticated risk parity approaches to balance risk contributions
- Exploring transfer learning to apply models across different markets and time periods

Our findings demonstrate that sophisticated machine learning approaches can significantly enhance investment decision-making, particularly for risk management during market stress periods. The combination of predictive modeling with anomaly detection provides a powerful framework for robust portfolio management in uncertain market environments.