EXECUTIVE SUMMARY

Team Number: Team 5

Team Members: Anabella Trias, Anthony Sierra

Date: December 12, 2025

Subject: Gold vs. S&P 500 Outperformance Prediction Using Machine Learning

---

Problem Statement:

Investors face persistent challenges when allocating capital between safe-haven assets like gold and growth-oriented equities such as the S&P 500. Traditional allocation strategies rely heavily on qualitative judgment and lagging indicators, offering limited support for timing allocation shifts. This project develops a machine learning–based classification framework to predict whether gold or equities are likely to outperform over a 90-day horizon, providing a systematic, data-driven decision support tool.

Methodology:

Data: Integrated 12 sources (Investing.com and FRED) spanning 2006-2024, creating 4,458 daily observations with 40+ engineered features: returns (5/20/60-day), moving averages, volatility metrics, macroeconomic indicators (CPI, unemployment, Fed Funds, M2, yields), and derived features (gold-silver ratio, real interest rates).

Model development occurred in two stages:

1. **Baseline modeling using Logistic Regression, Random Forest, and XGBoost for interpretability and error analysis.**
2. **PyCaret was used as an exploratory benchmarking tool to evaluate 15 classification algorithms using stratified 10-fold cross-validation applied within the training window only. While this approach ensures consistent preprocessing and fair comparison across models, it does not fully preserve temporal ordering and therefore may inflate performance estimates for time-dependent financial data.**
3. **Model Comparison and Selection:**

A range of models—including Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and Extra Trees—were evaluated using cross-validated performance metrics. Results indicate a clear performance gap between linear and ensemble-based methods.

Table 1: Model Performance Comparison (Cross-Validated)

Rank	Model	CV AUC	CV Accuracy	Rationale
1	Extra Trees	0.9897	95.08%	Best AUC, strongest nonlinear capture
2	Random Forest	0.9893	95.02%	Comparable but slightly weaker
3	LightGBM	0.9881	94.97%	Faster, marginally lower accuracy
4	XGBoost	0.9880	95.02%	Strong but not superior
5	Logistic Regression	0.8210	74.35%	Interpretable baseline

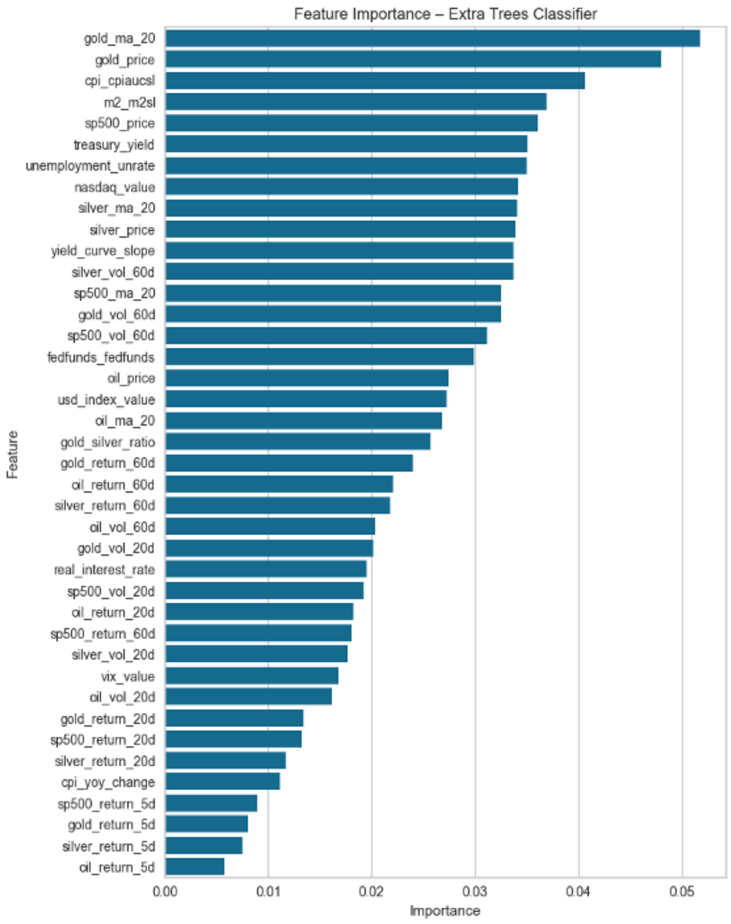
Table 1 presents cross-validated model performance under a stratified sampling framework. While ensemble-based methods such as Extra Trees exhibit extremely high accuracy and AUC scores, these results should be interpreted as upper-bound estimates rather than deployable performance. Financial time series exhibit strong temporal dependence, and models evaluated without walk-forward validation often overstate predictive power.

<u>Major Findings:</u>

The following findings summarize both exploratory ensemble benchmarking results and realistic time-aware baseline performance.

1. **Benchmark Performance (Exploratory):**
   Extra Trees achieved the highest cross-validated ROC AUC and accuracy under stratified evaluation, indicating strong capacity to capture nonlinear relationships among macroeconomic variables. However, when evaluated under time-aware splits, performance was substantially lower and more consistent with baseline models.

2. **Feature Importance (Figure 1):**
   The strongest predictors were:
   - **Gold 20-day moving average**
   - **Gold and equity price levels**
   - **CPI inflation and M2 money supply**
   - **Yield curve slope and real interest rates**



As shown in Figure 1, these variables consistently ranked highest in importance, suggesting that macro-monetary conditions drive many of the model's signals

3. **Economic Patterns:**

   Gold outperformed 42.8% of observations (stocks 57.2%). Gold dominated during crises (2008, 2020); stocks during expansions (2013-2019).

4. **Threshold Optimization (Baseline Model):**

   Logistic Regression threshold tuning (0.4 vs. 0.5) improved recall to 88%, demonstrating how probability thresholds can be aligned with business objectives (minimizing missed gold rallies). These findings highlight that decision thresholds are as important as model selection itself, particularly in financial applications where missed defensive opportunities carry asymmetric risk.

5. **Error Behavior (Baseline Models): False negatives correspond to missed gold outperformance windows (opportunity cost), while false positives represent**

conservative allocations where stocks performed better. Errors were most common during volatile or rapidly shifting macro regimes, reinforcing the need to combine model signals with market context.

Recommendations:

Investment & Portfolio Use:

- **Use model predictions only when confidence is high, avoiding action on weak or ambiguous signals.**
- **Treat outputs as decision support, not automated investment instructions.**
- **Combine model predictions with macroeconomic analysis, market context, and professional judgment before making allocation changes.**
- **Recognize that realistic out-of-sample performance is expected to fall in the 65–75% range, consistent with time-aware baseline models, rather than cross-validated ensemble estimates.**

Monitor key economic and market indicators highlighted by the model to understand why a signal is generated.

Technical & Operational:

- **Use the deployed API to generate on-demand or scheduled predictions as market conditions evolve.**
- **Apply batch predictions to evaluate multiple scenarios and stress-test outcomes under different macroeconomic assumptions.**
- **Retrain and reassess the model periodically to ensure performance remains stable as market regimes change.**
- **Deployment: A FastAPI backend and Streamlit frontend support both individual and batch predictions.**

Analytical Overview:

Extra Trees outperform by averaging fully randomized decision trees, enabling it to capture nonlinear interactions while reducing overfitting. Model validation followed a two-stage approach: stratified cross-validation was used for exploratory benchmarking and model comparison, while a strict temporal holdout was used to evaluate baseline models under realistic deployment conditions. The final pipeline is reproducible, versioned, and deployable, supporting both research and applied investment workflows.

Limitations:

1. **Stratified cross-validation does not preserve time order, leading to optimistic performance estimates for ensemble models; walk-forward validation is required for realistic deployment**
2. **Training on 2006-2017 may miss future regimes**
3. **Features exclude sentiment/geopolitical data.**

Future Work: Walk-forward validation, sentiment integration, ensemble stacking (Extra Trees + XGBoost), transaction cost modeling, automated retraining with drift detection.

Impact: This project transforms qualitative asset allocation decisions into a quantitative, explainable framework. While exploratory ensemble models demonstrate strong theoretical potential, realistic time-aware performance suggests achievable accuracy closer to 65–75%. Even at this level, the system provides meaningful value by reducing subjective bias, improving consistency, and supporting disciplined allocation decisions during periods of macroeconomic uncertainty