Ground Truth

Project Plan

Team: Connor Garrett Watson , Ronald Stuart Holland , Francisco Munoz , Mark A Gibbons
Date: Sep 19, 2025

WHY

Problem Statement

Financial markets for agricultural commodities lack accessible, transparent AI-driven forecasting tools that integrate diverse data sources. Current solutions either require extensive coding expertise or provide opaque "black box" signals. Retail traders and smaller commodity producers need data-driven forecasts that combine market data, fundamentals, logistics, climate, and sentiment analysis into actionable insights.

Assumptions

1. **Information Edge**: Systematic processing of multi-modal public data (market, climate, logistics, sentiment) provides predictive value beyond traditional price-only models
2. **Signal Fusion**: Combining structured (LSEG market data, USDA fundamentals) with unstructured data (GDELT news sentiment) improves forecast accuracy
3. **Modern Architecture Advantage**: State-of-the-art time series models (TimesFM/TimeCopilot) outperform classical approaches for multi-variate commodity forecasting

Impact and Market Opportunity

Market Size:

- TAM: $51.4B algorithmic trading market (2024)
- SAM: $7.18B projected retail algorithmic trading market by 2030 (12.7% CAGR)
- SOM: Commodity-focused segment representing a "blue ocean" opportunity, largely uncontested compared to equity-focused competitors

Impact Quantification: Performance measured through backtested Information Ratio (target >0.5), Sharpe Ratio, and Maximum Drawdown on historical coffee and sugar futures data.

Target Customer

- **Primary**: Sophisticated retail traders ("prosumers") seeking transparent, AI-driven commodity signals

- **Secondary**: Small-to-medium commodity producers needing market insights for hedging decisions
- **Use Case**: Hourly forecasts for coffee and sugar futures with explainable feature attributions

WHAT

MVP (POC Scope - 12 weeks)

A fully automated daily forecasting system for coffee (KC) and sugar (SB) futures that:

- Ingests multi-modal data via AWS infrastructure
- Generates calibrated hourly price forecasts (min, max, close) at a 48 hour horizon
- Provides feature importance explanations

Key Features

1. **Automated Data Pipeline**: Daily ingestion of LSEG market data, USDA fundamentals, CHIRPS/ERA5 climate data, MarineTraffic logistics, and GDELT sentiment
2. **Advanced Forecasting**: TimesFM/TimeCopilot models incorporating both structured and unstructured features
3. **Risk Governance**: Position limits, turnover controls, and confidence-based sizing
4. **Explainability**: SHAP values and feature attribution for transparency

Value Proposition

Unlike QuantConnect (requires coding) or Danelfin (black-box equities), we provide transparent, multi-modal commodity forecasts combining market, fundamental, climate, and sentiment data with clear explanations of driving factors.

HOW

Data Strategy

Using **Option C - Enterprise (LSEG/Refinitiv)** via UCB license, supplemented with:

- **Market**: Continuous futures with roll adjustments (KC, SB)
- **Fundamentals**: USDA PSD, CONAB, ICO/ISO reports
- **Climate**: CHIRPS precipitation, ERA5 temperature, MODIS NDVI
- **Logistics**: MarineTraffic port activity, Cecafé export data
- **Sentiment**: GDELT news analysis with FinBERT

Technical Approach

Three-Agent Architecture:

1. **Agent S (Researcher)**: Data curation, feature engineering, sentiment analysis
2. **Agent T (Forecaster)**: TimesFM/TimeCopilot model training and prediction
3. **Agent R (Risk Governor)**: Trading rules, position management, signal generation

Infrastructure: AWS-based with Lambda functions, S3 storage, EventBridge scheduling, API Gateway serving

Project Management

Member	Primary Role	Secondary Role
Francisco	Data Engineering Lead	ML Engineering Support
Connor	Time Series Models Lead	Project Management
Tony	NLP/Sentiment Lead	Finance Domain Expert
Stuart	System Architecture Lead	Data Engineering Support

12-Week Timeline

Weeks 1-3: Data Pipeline Setup

- Establish AWS infrastructure and LSEG connectivity
- Implement Agent S data ingestion for all sources
- Create continuous futures series with proper roll methodology

Weeks 4-6: Feature Engineering

- Process climate data with geospatial joins
- Implement FinBERT sentiment extraction from GDELT
- Build feature store with 2+ years historical data

Weeks 7-9: Model Development

- Implement TimesFM/TimeCopilot forecasting models
- Integrate structured and unstructured features
- Calibrate probabilities and validate predictions

Weeks 10-11: Backtesting & Evaluation

- Walk-forward validation on historical data
- Calculate performance metrics (IR, Sharpe, MDD)
- Implement Agent R risk governance rules

Week 12: Deployment & Documentation

- Deploy API via API Gateway + Lambda
- Create QuickSight dashboard
- Complete documentation and handover

Success Metrics

- Information Ratio > 0.5 vs buy-and-hold benchmark
  - IR = (Portfolio Return - Benchmark Return) / Tracking Error
  - Tracking Error = standard deviation of excess returns
- AUC > 0.65 for directional predictions
- Brier Score < 0.20 (well-calibrated probabilities)
- System Availability > 99%

Risk Mitigation

- **Data Quality**: Automated validation checks, redundant sources
- **Model Overfitting**: Rolling window validation, regularization
- **Infrastructure**: CloudWatch monitoring, failover procedures
- **Compliance**: Clear disclaimers, no PII collection

This focused plan delivers a working POC for coffee and sugar forecasting within the 12-week timeframe while laying groundwork for future expansion to the full commodity universe vision.