Kairos ML Alpha v2 - Migration Summary

Date: January 1, 2026

What Was Accomplished

1. Fixed feat_fundamental Coverage

- **Problem:** Quarterly data wasn't forward-filled to daily dates (1.6% coverage)
- **Solution:** Created `rebuild_feat_fundamental.py` that forward-fills SF1 quarterly data
- **Result:** 100% coverage on daily dates
- **Script added to pipeline:** Phase 3 in `run_pipeline.py`

2. Trained XGBoost ML v2 on Raw Features

- **Problem:** ML v1 trained on pre-blended composites with baked-in sign assumptions failed
- **Solution:** Train on 23 raw features, let ML learn optimal directions

Features used (23 total):

```
Fundamental: earnings_yield, fcf_yield, roa, book_to_market, operating_margin, roe
Volatility:  vol_21, vol_63, vol_blend
Beta:        beta_21d, beta_63d, beta_252d, resid_vol_63d
Price:       hl_ratio, range_pct, ret_21d, ret_5d
Momentum:    mom_1m, mom_3m, mom_6m, mom_12m, mom_12_1, reversal_1m
```

3. Results

Metric	v1 (Failed)	v8 Baseline	ML v2
Mean IC (CV)	+0.005	+0.021	**+0.035**
Backtest Sharpe	-	1.14	**1.20**
Annual Return	-	22.8%	**30.0%**

Key Files

File	Purpose
`scripts/features/rebuild_feat_fundamental.py`	Forward-fill quarterly fundamentals (in pipeline)
`scripts/ml/train_xgb_alpha_v2.py`	XGBoost training on raw features
`scripts/ml/outputs/model_classification_v2.json`	Trained classifier model
`scripts/ml/outputs/model_regression_v2.json`	Trained regression model
`scripts/ml/outputs/feature_medians_v2.json`	Feature medians for imputation

Database Tables

Table	Description
`feat_alpha_ml_xgb_v2`	ML predictions (ticker, date, alpha_ml_v2_reg, alpha_ml_v2_clf)
`feat_matrix_v2`	Now includes `alpha_ml_v2_clf` column

Optimal Production Configuration

```python
```

```
CONFIG = {
    "top_n": 75,
    "target_vol": 0.25,          # 25% (was 20%)
    "alpha_column": "alpha_ml_v2_clf",  # ML signal (was v8)
    "min_adv": 2_000_000,
    "max_position_pct": 0.03,
    "max_sector_mult": 2.0,
    "lambda_tc": 0.5,
    "max_turnover": 0.30,
    "vol_column": "vol_blend",
    "adv_column": "adv_20",
}
```

Backtest Results (2015-2025)

Parameter Set	Annual Return	Sharpe	Max DD
75 pos, 5d, 20% vol	24.0%	1.20	-28.3%
75 pos, 5d, 25% vol	**30.0%**	**1.20**	-34.6%
50 pos, 5d, 25% vol	28.4%	1.14	-31.7%

Winner: 75 positions, 5-day rebalance, 25% vol target

SHAP Feature Importance (Top 10)

1. vol_63 (63-day volatility)
2. vol_blend
3. mom_3m (3-month momentum)
4. earnings_yield
5. mom_1m
6. beta_21d
7. fcf_yield

8. beta_63d
9. roa
10. ret_5d

Known Issues

ML Predictions NOT in Pipeline Yet

IMPORTANT: After running the full pipeline (`run_pipeline.py`), you must manually:

1. Regenerate ML predictions (pipeline rebuilds `feat_matrix_v2` fresh without ML columns)
2. Backfill recent dates that lack forward returns

After each pipeline run, execute:

```bash
# Retrain or just regenerate predictions
python scripts/ml/train_xgb_alpha_v2.py --db data/kairos.duckdb --skip-cv

# Then update feat_matrix_v2 with ML column (run the backfill script)
```

TODO for next session: Add ML prediction step to pipeline Phase 5 or 6.

Recent Dates Missing Predictions

The ML training script filters on `ret_5d_f IS NOT NULL`, so the last 5 trading days won't have predictions automatically. Run backfill script after each pipeline run.

Next Steps (for future sessions)

1. **Longer prediction horizons** - Train on ret_10d_f or ret_20d_f
2. **Add more features** - insider_composite_z, institutional ownership
3. **Ensemble** - Blend ML v2 + v8 for robustness
4. **Regime conditioning** - Different models for different market regimes
5. **Integrate ML prediction into pipeline** - Auto-generate predictions for all dates

Commands Reference

Run full pipeline:

```bash
python run_pipeline.py --db data/kairos.duckdb
```

Train ML v2:

```bash
python scripts/ml/train_xgb_alpha_v2.py --db data/kairos.duckdb
```

Backtest ML v2:

```bash
python scripts/backtesting/research/backtest_academic_strategy_risk4.py \
  --db data/kairos.duckdb \
  --alpha-column alpha_ml_v2_clf \
  --target-column ret_5d_f \
  --top-n 75 --rebalance-every 5 --target-vol 0.25 \
  --start-date 2015-01-01 --end-date 2025-12-31
```

Generate production rebalance:

```bash
python scripts/production/generate_rebalance.py --db data/kairos.duckdb --date 2025-12-31
```