35137 - Machine Learning in Finance
Group 14: Jack Gordon, Kathryn Wason, Christian Bohren
Updated project proposal & Workplan

UPDATED: Are “smarter” models actually better at predicting FX moves around Fed meetings?

What we aim to explore

Foreign exchange (FX) markets move fast on Fed days (i.e., when the US Federal Reserve announces changes to the Federal Funds Rate). Traders react these days to three things, including the rate change itself, potential implications on future rates, and the Fed’s tone.

We want to use analytical techniques to evaluate, on Federal Open Market Committee (FOMC) decision days, how much more accurate we can get at predicting USD FX moves by increasing complexity in models. Would moving up a “model ladder” (below) improve performance?

1. **Simple baselines** (e.g., predict 0; random walk)
2. **Theory-style baselines** (e.g., interest rate differentials / parity ideas)
3. **Supervised ML on structured data** (e.g., implications for future rate curve moves, related cross-asset features from other market products)
4. **Add text signals** from Fed communication (LLM rubric scoring first; embeddings later) - *meeting statement for first window, press conference for second window (below).*

Scope and design

Currency pairs: We plan to focus on 4 key currency pairs versus the US Dollar (USD) - the Euro (EUR), the Japanese Yen (JPY), the British Pound (GBP), and the Canadian Dollar (CAD). These currencies are less “managed” by their central government and collectively account for ~48.4% of daily FX turnover in April 2025, per the BIS. These currencies are together very liquid, are floating, and have significant USD exposure.

Window for prediction: We would consider two intraday windows, aligned to how markets trade to account for the Fed:

- Statement reaction: 2:00-2:30p ET - capturing the immediate response to the decision and formal Fed statement.
- Post-statement “digestion”: 2:30-4:00p ET - captures how the market broadly adapts to the decision and (when applicable) the associated press conference from the Fed Chair.

Scope of data: We focus on the FOMC meetings since 2021 (N=41) given our limits on accessing data efficiently without access to an API.

Data and modeling approach

Data: This requires us to use a mix of structured and unstructured data; structured data is primarily found from Bloomberg (using Booth’s Bloomberg Terminals) and online sites.

35137 - Machine Learning in Finance
Group 14: Jack Gordon, Kathryn Wason, Christian Bohren
Updated project proposal & Workplan

- Intraday FX: This is our target variable, using the USDEUR, USDJPY, USDGBP, and USDCAD intraday prices at a 5-minute frequency. Positive values mean that the USD appreciates.
- Rates context: We add incremental structured features to capture macroeconomic variables, including US Treasury yields (e.g., 2Y, 10Y in window) and risk proxies (e.g., S&P 500 moves, VIX).
- Fed communications text: Statement text coming out of each meeting and press conference transcript text, when available.

Modeling approach: In line with the prior ladder, we could consider the following categories of models to see whether extra complexity drives value:

- Must-beat: Zero-mean and historical average baselines.
- Theoretical baselines: Utilizing differential / forward-premium predictors based on theory for international economic policy to see whether basic relationships are predictive.
- Supervised ML: Using models appropriate for small samples (given few Fed meetings), e.g., OLS and Ridge / LASSO / Elastic Net. We can also consider ways to add nonlinearities to predict values.
- Textual signals: We can add additional features focused on the text on top of supervised ML, including a rubric-based scoring run by an LLM (e.g., hawkish v. dovish, inflation v. labor emphasis) and an embeddings approach.

End goals

At the end of this, we would hope to be able to determine:

- Errors: RMSE / MAE for each currency pair and model.
- Improvement: Out-of-sample improvement versus the baseline, using 20% of data for testing versus 80% on training (potentially considering cross-validation).

In addition, we may do a basic view of trading costs to see whether any improvements could have practical value, though we are focused primarily on potential incremental signal.

35137 - Machine Learning in Finance
Group 14: Jack Gordon, Kathryn Wason, Christian Bohren
Updated project proposal & Workplan

NEW: Actions to complete

0) Project scaffolding + reproducibility (do first)

0.1 Repo + data layout

Deliverable: a repo structure that makes the pipeline rerunnable.

Suggested tree:

- `data_raw/`
  - `fomc_metadata.csv`
  - `policy_rates_daily.csv`
  - `intraday_fx/{pair}.csv`
  - `intraday_rates/{2y,10y}.csv`
  - `intraday_index/{spx,vix}.csv`
  - `statements/` (if collected)
  - `transcripts_pdf/`
- `data_clean/` (standardized schemas, cleaned)
- `features/` (model-ready tables)
- `notebooks/` (EDA + sanity checks)
- `src/`
  - `ingest.py`, `clean.py`, `features_structured.py`, `features_text.py`, `models.py`, `eval.py`, `plots.py`
- `outputs/`
  - `tables/`, `figures/`, `model_artifacts/`
- `report/` (final writeup + slides)

0.2 Environment + configs

Deliverables

- `requirements.txt` (or `environment.yml`)
- `config.yaml` containing:
  - window definitions (2:00–2:30; 2:30–4:00 ET)
    ML in Finance - Final Project -…
  - list of pairs/assets
  - feature toggles (structured-only vs +text)
  - CV scheme settings (grouped by meeting, etc.)

35137 - Machine Learning in Finance
Group 14: Jack Gordon, Kathryn Wason, Christian Bohren
Updated project proposal & Workplan

---

1) Data ingestion + cleaning (standardize every raw source)

1.1 FOMC metadata (meeting-level)

Inputs: dates, announcement time, target range bounds, votes for/against (and any flags).
Processing steps

- Standardize `announcement_datetime_et` (timezone-aware), `meeting_id` (e.g., `YYYY-MM-DD`), and numeric fields.
- Create derived fields:
    - `target_mid = (lower+upper)/2`
    - `change_mid` vs prior meeting
    - `is_hike/is_cut/is_hold`
    - `dissent_count = votes_against` (or separate if you have details)

Output: `data_clean/fomc_metadata.parquet`

1.2 Daily policy rates (Fed/ECB/BOJ/BOC/BOE)

Processing steps

- Standardize to one row per day per central bank: `date`, `cb`, `policy_rate`.
- Forward-fill only where economically correct (policy rates are step functions).
- Create rate differentials used for theory features:
    - `diff_fed_ecb`, `diff_fed_boj`, `diff_fed_boe`, `diff_fed_boc`
- For each FOMC date, create:
    - same-day differential (as-of close prior day or as-of morning—pick one convention and lock it)

Output: `data_clean/policy_rates_daily.parquet`

1.3 Intraday bars (FX, SPX/VIX, UST 2Y/10Y)

You'll want a *single canonical intraday schema* across all assets:

Canonical schema

- `timestamp_et` (end time of bar, ET, tz-aware)

- `asset` (e.g., `USDJPY`, `SPX`, `UST2Y`)
- `open`, `high`, `low`, `close` (numeric)
- `bar_minutes` (=5)
- `meeting_id` (derived from date)
- `window` (Statement vs Digestion; computed from timestamp)

Cleaning steps (common gotchas)

- Deduplicate timestamps; enforce 5-min grid.
- Handle missing bars:
  - flag missingness
  - decide: drop affected meetings for that asset/window *or* impute conservatively (usually better to drop for small-N integrity).
- Confirm unit conventions:
  - yields in percent vs decimal
  - VIX level vs percent (usually level)
- **FX direction convention:** you stated “positive means USD appreciates.” Enforce this mechanically:
  - compute `fx_return = log(close/open)` on a USD-strength basis
  - if any series is quoted as USD-per-FX (the “wrong” direction), invert prices before return calc.

Outputs

- `data_clean/intraday_bars.parquet` (stacked long table)
- `data_clean/qc_report.csv` with per-meeting missing bars counts by asset

---

2) Construct targets (your y’s) and the modeling panel

Your unit of observation should be **(meeting × pair × window)**. That gives up to:

- 41 meetings × 4 pairs × 2 windows = 328 rows (before missingness).

2.1 Define targets per row

For each `meeting_id`, `pair`, `window`:

Core targets

- `y_return`: log return over the window

- `y_abs_return`: absolute return (magnitude)
- `y_direction`: sign(y_return) as a classification target

Optional "microstructure" outcomes for analysis

- `range = log(high/low)` aggregated over the window
- realized vol proxy: sum of squared 5-min returns in the window

Output: `features/targets.parquet`

2.2 Merge into one panel

Join keys

- `meeting_id` + `pair` + `window`

Output: `features/panel_base.parquet` (targets + identifiers + meeting metadata)

---

3) Structured feature engineering (rungs 1–3 of ladder)

This is where you build a **feature dictionary** and keep it disciplined (small sample).

3.1 Must-beat baseline features (for benchmarking only)

- none needed; baselines operate directly on y

3.2 Theory-style features (rate differentials / parity ideas)

Per meeting × pair (copied to both windows unless you choose otherwise):

- `policy_diff` (Fed minus foreign CB)
  ML in Finance - Final Project -…
- `policy_diff_change_1m` (or change since last meeting)
- `fed_policy_level`, `foreign_policy_level`
- optional: "carry proxy" = policy_diff (same thing, but you may label it clearly)

Note: With only intraday FOMC-day FX, you probably *won't* implement full UIP/Fama regression cleanly; instead, treat these as "theory-motivated predictors."

3.3 Cross-asset reaction features (structured ML rung)

For each meeting, compute by window:

- UST: `ust2y_ret`, `ust10y_ret`, `slope_change` (2y–10y)
- SPX: `spx_ret`
- VIX: `vix_ret` (or Δlevel)

Then decide the information set per task:

Task A (Statement window model, 2:00–2:30):

- Use meeting metadata + policy differentials + (optionally) pre-2:00 context only.
- If you include cross-asset moves *from the same window*, be explicit in report that this is *contemporaneous explanatory modeling*, not a tradable forecast.

Task B (Digestion window model, 2:30–4:00):

- Use **Statement window** cross-asset features (2:00–2:30) as predictors for 2:30–4:00.
    - This is the cleanest "nowcast" setup and is usually defensible.

3.4 Meeting metadata features

- `change_mid` (bps)
- `is_hike/is_cut/is_hold`
- `dissent_count`
- `meeting_number_since_2021` (trend proxy)
- "cycle regime" proxy (optional): e.g., `fed_policy_level` bucketed

3.5 Feature QA

Deliverables

- `features/structured_features.parquet`
- `outputs/tables/feature_missingness_by_asset.csv`
- Correlation heatmaps (structured) and variance checks to catch near-constants

---

4) Text feature engineering (rung 4)

You have two text sources aligned to the two windows in your plan: **statement text for 2:00–2:30; press conference transcript for 2:30–4:00**

ML in Finance - Final Project -…

35137 - Machine Learning in Finance
Group 14: Jack Gordon, Kathryn Wason, Christian Bohren
Updated project proposal & Workplan

4.1 Text ingestion

Transcripts (PDFs)

- Extract text per meeting into `data_clean/transcripts_text/{meeting_id}.txt`
- Clean:
  - remove headers/footers/page numbers
  - normalize whitespace
  - optional: keep speaker tags if present (Powell vs reporter)

Statements

- If you have statement text already, same treatment into `data_clean/statements_text/`
- If not: either (a) drop from text scope for Task A, or (b) collect quickly from Fed archive and document the source.

4.2 Rubric scoring via LLM (fast + interpretable)

Define a fixed scoring schema; for each document, output structured JSON fields like:

- `hawk_dove_score` (e.g., -2..+2)
- `inflation_focus` (0..1)
- `labor_focus` (0..1)
- `growth_recession_risk` (0..1)
- `uncertainty` (0..1)
- `forward_guidance_strength` (0..1)
- `balance_sheet_mentions` (count)
- `risk_management_language` (0..1)

Critical controls

- Use the *same prompt* for all meetings.
- Save the prompt, model/version, and raw JSON outputs for auditability.
- Run simple inter-measure checks (e.g., does hawkish score correlate with 2Y move directionally?)

Outputs

- `features/text_rubric_scores.parquet`
- `outputs/tables/text_score_summary.csv`

4.3 Embeddings (optional / stretch)

- Compute embeddings per document (statement and/or transcript)
- Reduce dimensionality:
    - PCA to 5–20 components **max** (small-N)
- Treat these as additional features under strong regularization.

Outputs

- `features/text_embedding_pcs.parquet`
- Explained-variance plot for the PCA components

---

5) Modeling suite (the full ladder)

You should run **separate model tracks** for:

- Window A: Statement reaction (2:00–2:30)
- Window B: Digestion (2:30–4:00)

…and for each track, run both:

- **Regression** (predict `y_return`)
- **Classification** (predict `y_direction`)

5.1 Data splitting (avoid leakage)

Given event clustering, split by **meeting_id** (not by row):

- any given meeting's rows (all pairs) must be entirely in train or test.

Recommended evaluation:

- primary: **time-ordered splits** (rolling/expanding)
- secondary: leave-one-meeting-out (robustness)
- hyperparameter tuning: nested CV, grouped by meeting

5.2 Models to include (wide but realistic for small sample)

Rung 1 — Must-beat baselines

- Predict 0
- Predict historical mean by pair (train set only)
- Random-walk sign (direction = last meeting for that pair/window)

Rung 2 — Theory baselines

- OLS with only policy differential features (+ intercept, pair fixed effects)
- Simple “carry sign” classifier: sign(policy_diff)

Rung 3 — Structured ML (small-sample friendly)

- OLS (with pair FE)
- Ridge / LASSO / Elastic Net
- Huber regression (robust to outliers)
- (Optional) Gradient-boosted trees *with strict constraints*:
  - shallow depth, strong regularization, and meeting-group CV

Rung 4 — Add text

- Repeat ridge/elastic net adding rubric scores
- Repeat ridge adding PCA embedding components (if used)
- Ablation: structured-only vs structured+text

5.3 Model outputs you will need for the report

For each window × pair (and pooled):

- performance table: MAE, RMSE, directional accuracy
  ML in Finance - Final Project -...
- “incremental ladder” plot: baseline → theory → structured → +text (ΔRMSE / ΔMAE)
- coefficient tables for linear models (and stability across splits)
- feature importance summaries:
  - standardized coefficients (linear)
  - permutation importance (model-agnostic; safer than SHAP with tiny N)

Deliverables

- `outputs/tables/model_performance.csv`
- `outputs/figures/ladder_improvement.png`
- `outputs/figures/residual_plots.png`
- `outputs/tables/top_features_by_model.csv`

---

6) Diagnostics, robustness, and “strategy check” outputs

6.1 Diagnostics

- residual distribution by window (fat tails are expected)
- performance by regime:
  - hiking cycle vs cutting/hold periods (use fed policy level or meeting epoch)
- sensitivity to outliers (winsorize y? report both)

6.2 Robustness checks (pick 2–3)

- direction-only vs regression consistency
- excluding the largest 1–2 moves (stress test)
- pooled vs per-pair models
- using only 2Y (drop 10Y) or only SPX/VIX to test redundancy

6.3 Basic “strategy check” (lightweight)

Not a full backtest—just an economic sanity check:

- Take the sign prediction and apply to a notional $1 exposure
- Assume a conservative transaction cost per trade (document assumption)
- Report average net per-trade and hit-rate

Deliverables

- `outputs/tables/strategy_check.csv`
- a single slide-ready chart: cumulative P&L over meetings (ordered by time)