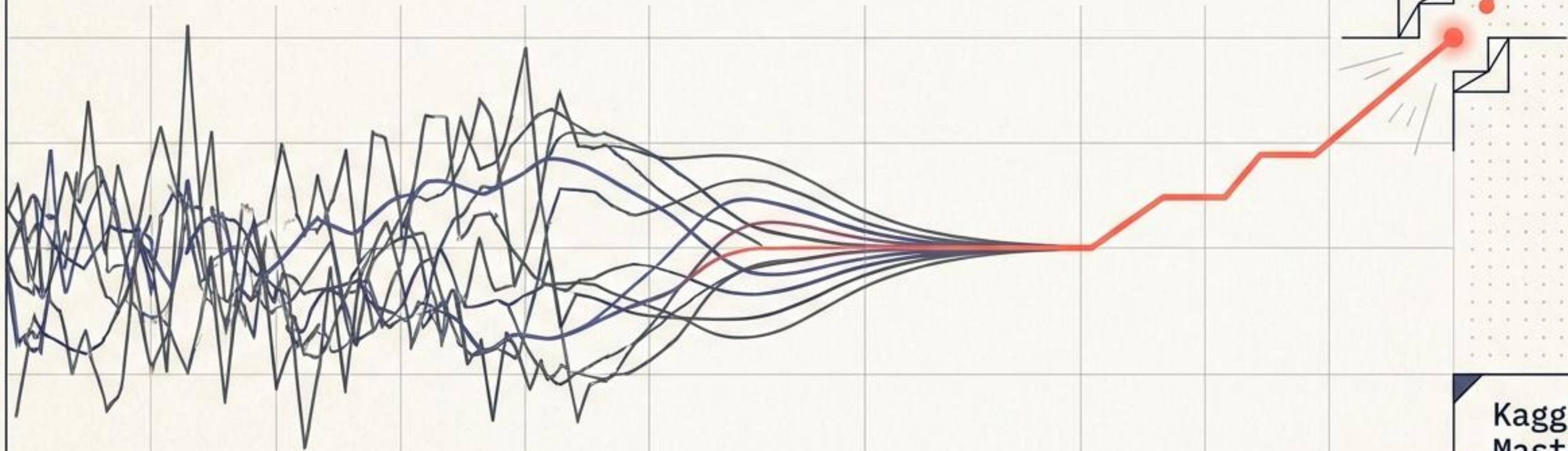


Store Sales Forecasting

AIと共に挑む時系列予測の極意

Google AI Studioを活用した反復的ブレイクスルーの軌跡



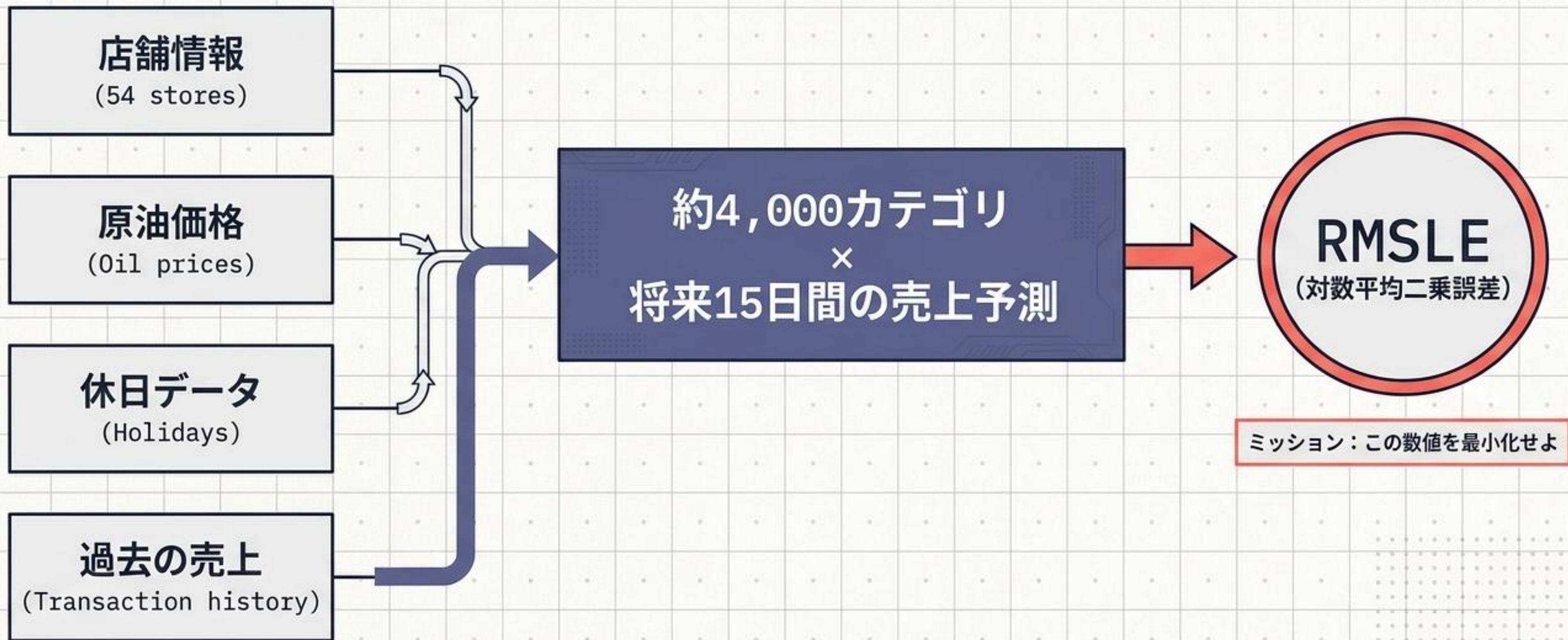
Kaggle
Masterclass



Store Sales - Time Series Forecasting

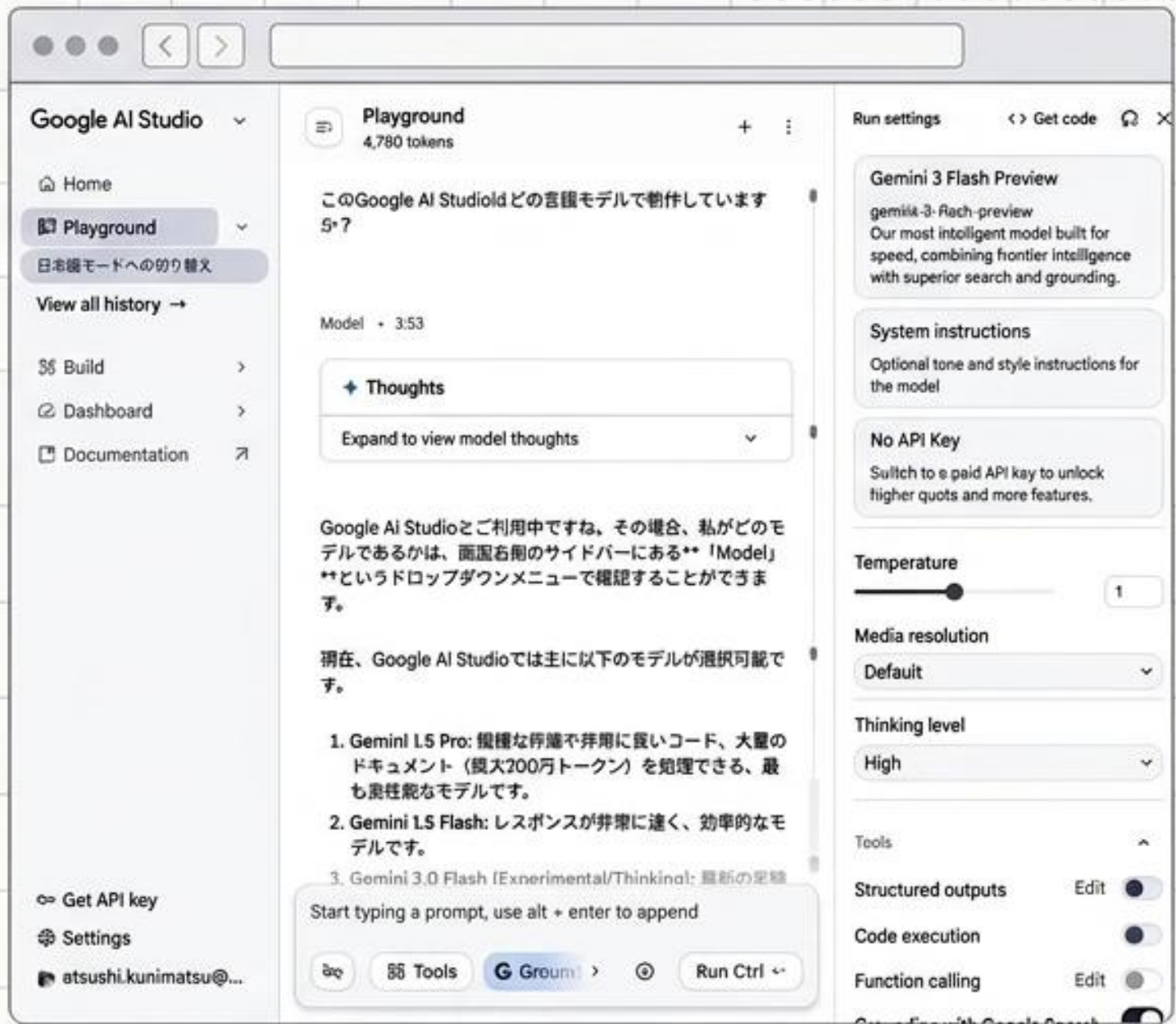
Use machine learning to predict grocery sales

Getting Started · 706 Teams · Ongoing



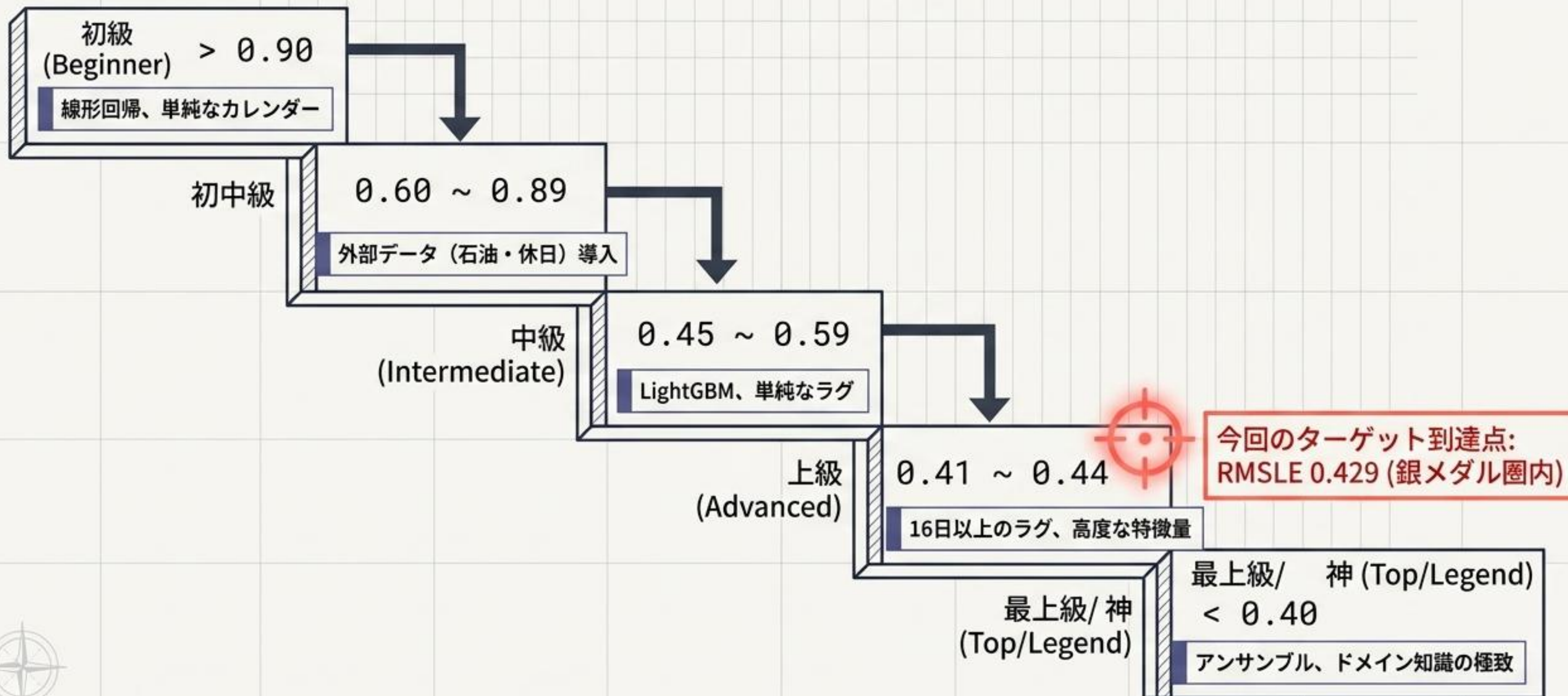
ハイスコアを狙うための武器

	Gemini (無料版)	Google AI Studio
コンテキスト ウィンドウ	制限あり	最大200万トークン (Kaggleの議論や全CSV を一括読み込み可能)
システム プロンプト	汎用・初心者 向け	Kaggle Grandmaster としてのペルソナ 固定が可能
コード実行 (Code Execution)	非対応	バックグラウンドで Pythonを実行し、 重いCSVを直接分析



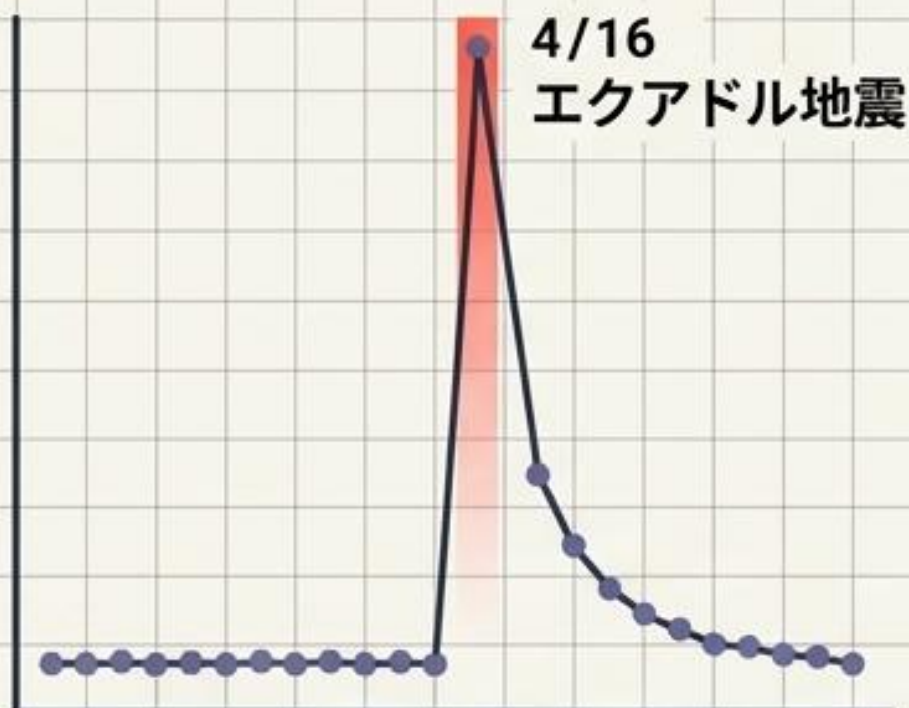
結論: 圧倒的なコンテキストと役割固定が可能な「Google AI Studio」一択。

戦場の見取り図：スコア階層と到達目標



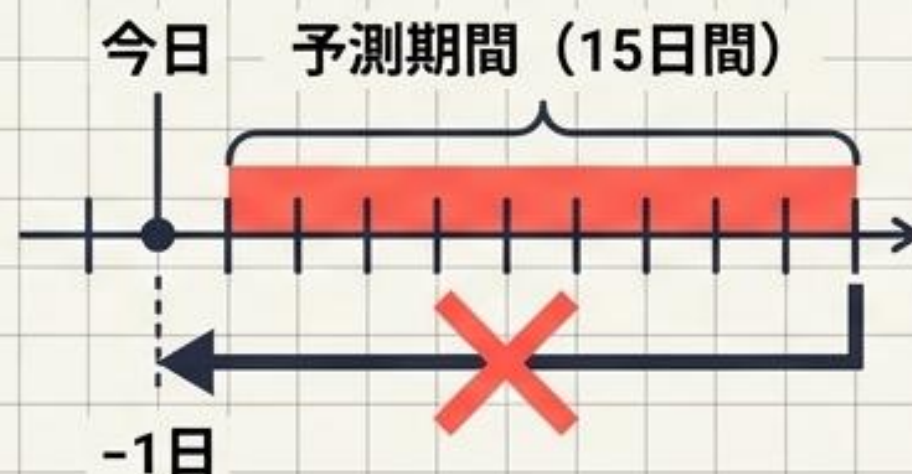
探索的データ分析（EDA）で判明した3つの壁

課題1: 外部ショック



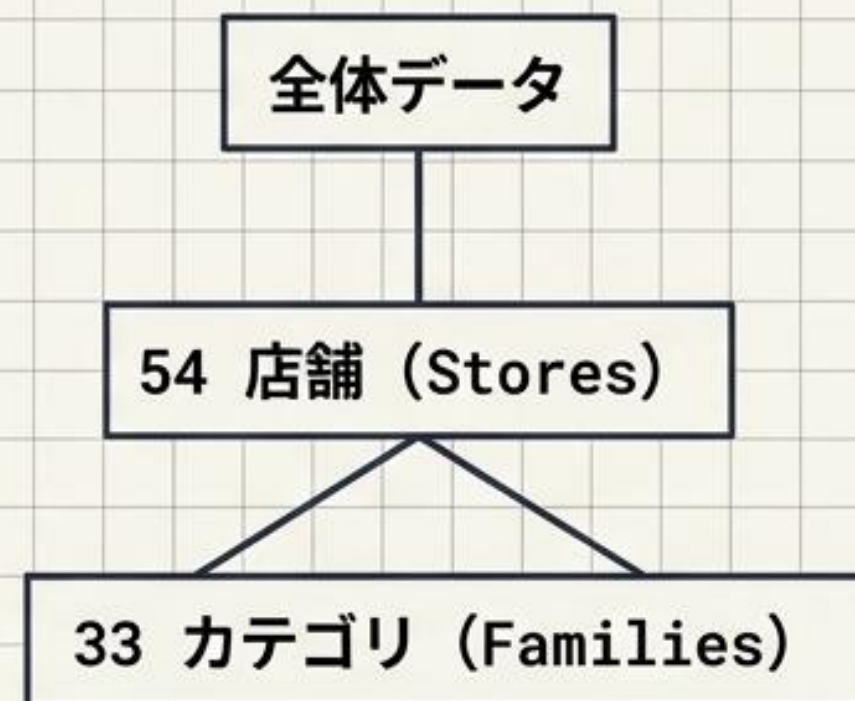
問：外れ値として除外するか、
特需フラグを立てるか？

課題2: ラグの乖離



問：将来の予測に「前日の売上」
は使えない。どう補うか？

課題3: 階層構造



問：1つの巨大モデルで解くか、
個別モデルを作るか？

Phase 1: ベースラインの構築と線形モデルの限界



$$\text{Linear Regression} + \text{カレンダー (曜日・月)} = \text{RMSLE: 0.902}$$



$$\text{原油価格 (Oil)} + \text{休日 (Holiday)} = \text{RMSLE: 0.761}$$

インサイト:

- 0.90から0.76への初期改善。
- しかし、線形回帰には根本的な限界がある。
「時代のトレンド」や、休日とプロモーションが重なるような「非線形な相乗効果」を捉えきれない。



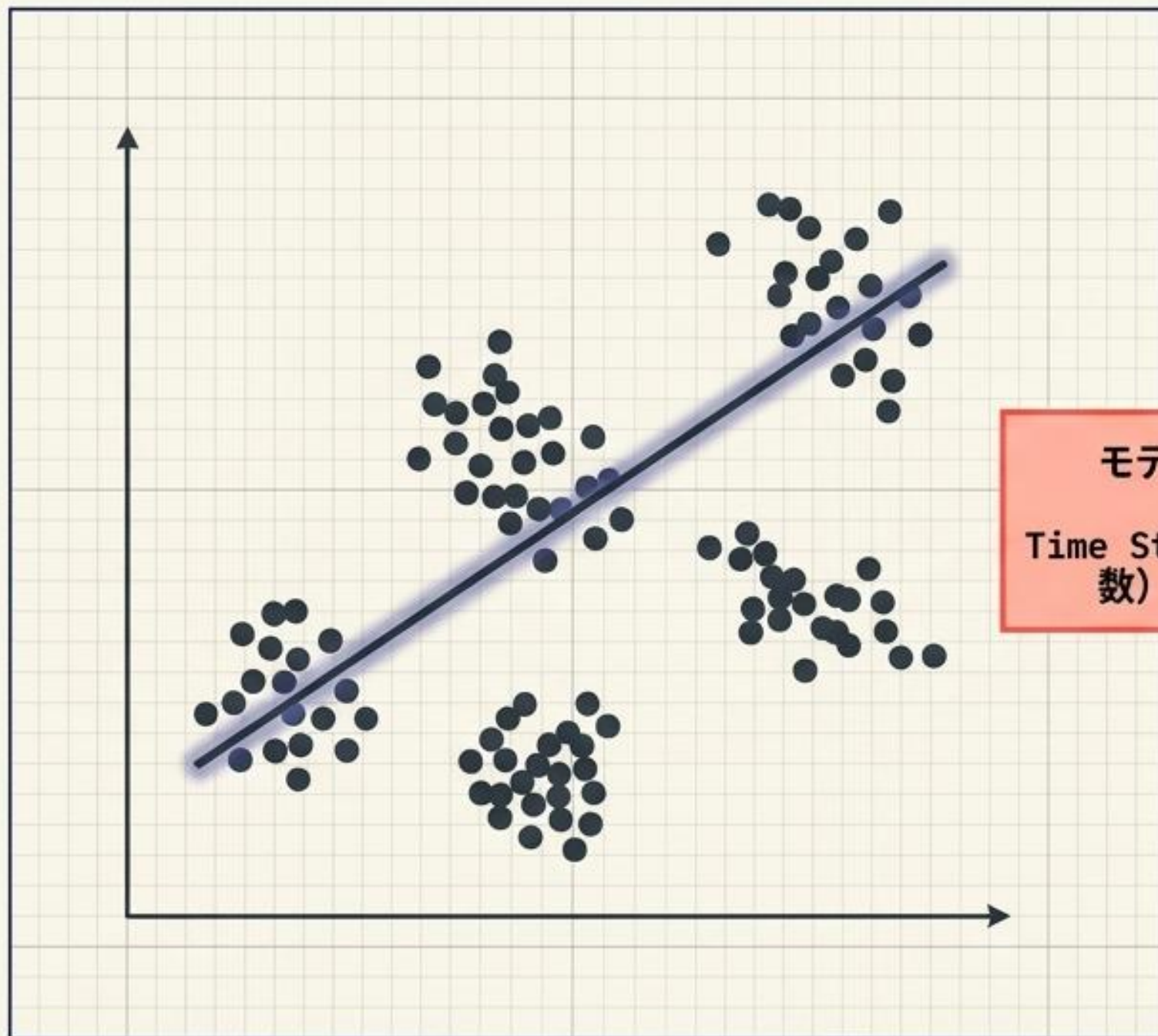
パラダイムシフト：決定木モデル（LightGBM）への移行

Score Badge

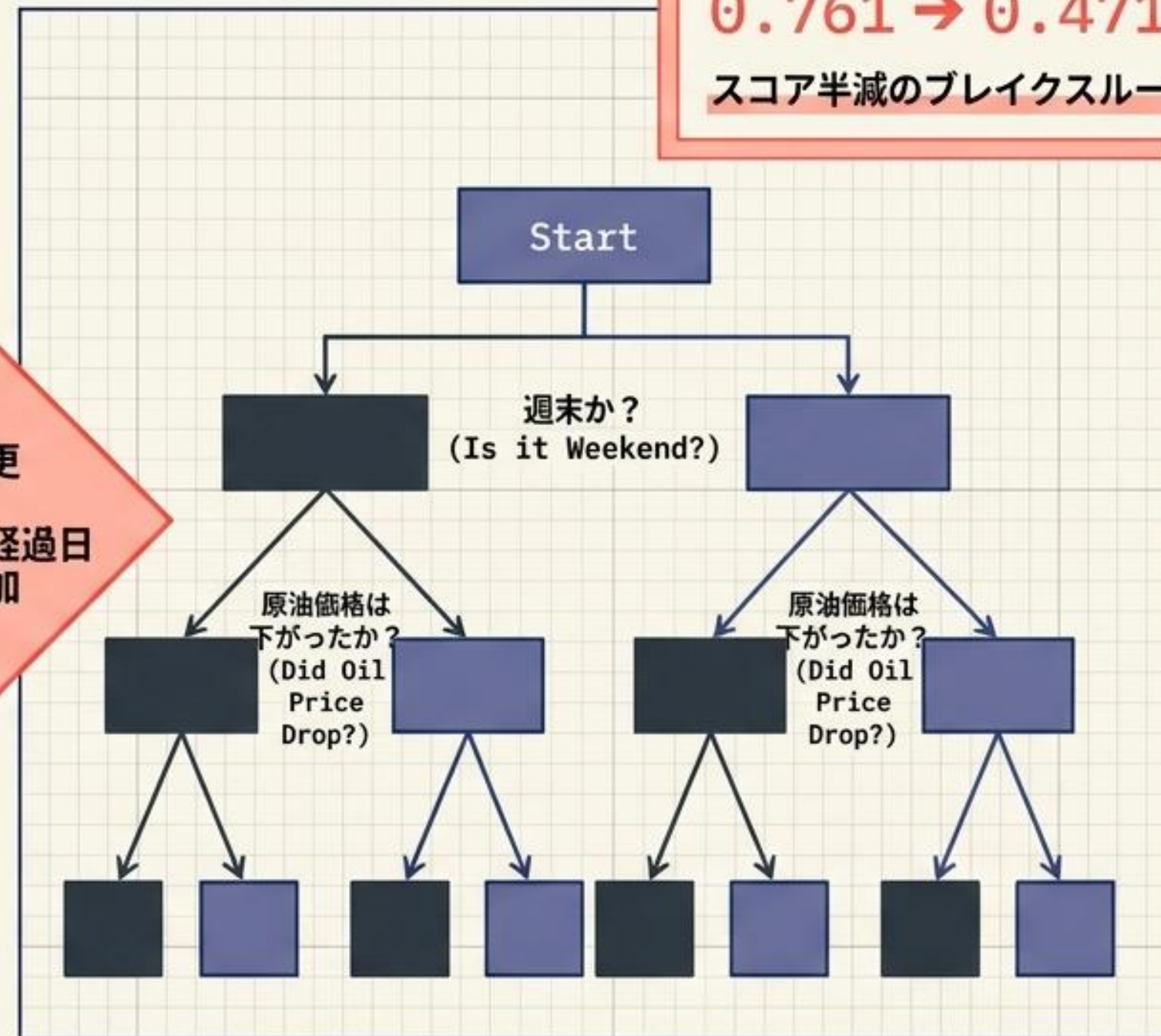
Current Score

0.761 → 0.471

スコア半減のブレイクスルー



モデル変更
+
Time Step (経過日
数) の追加



線形回帰の限界: 複雑な条件分岐ができない

決定木: カテゴリ間の複雑な関係性を自動学習

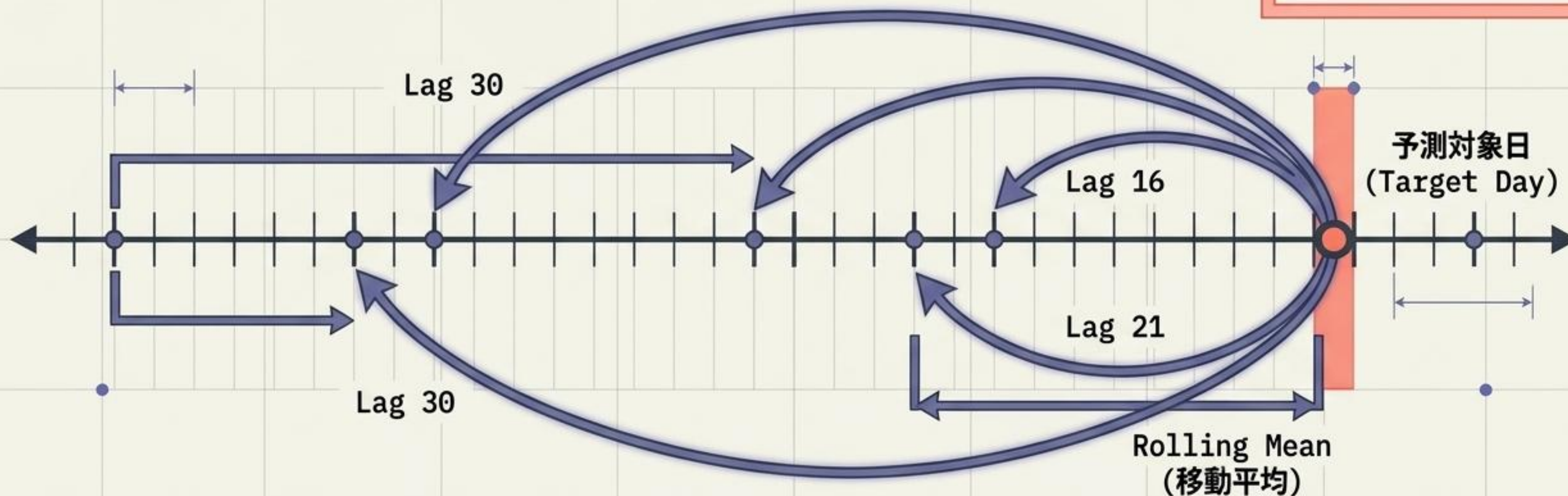
Phase 2: 過去之力（ラグと移動平均の導入）

Score Badge

Current Score

0.471 → 0.461

0.5の壁を完全突破



「過去の売上」が最大の予測因子である。テスト期間(15日間)の未来を予測するためには、データリークを防ぎつつ、16日以上前のデータに依存するラグ設計が必須となる。

アンサンブルの罠 (The Ensemble Stumble)

Score Badge

Current Score

0.461 → 0.464

学びの停滞期 (スコア微増)

LightGBM

店舗ID (1~54) を
『カテゴリラベル』とし
て正しく認識

加重平均
(0.85 : 0.15)

Ridge Regression

店舗IDを『純粋な数値の
大小』 (54は1の54倍)
として誤認

スケールの不一致によるノイズ発生 ⚠



学び: 異なる特性を持つモデルを単純に足し合わせると、スケーリングの不在により逆効果になることがある。

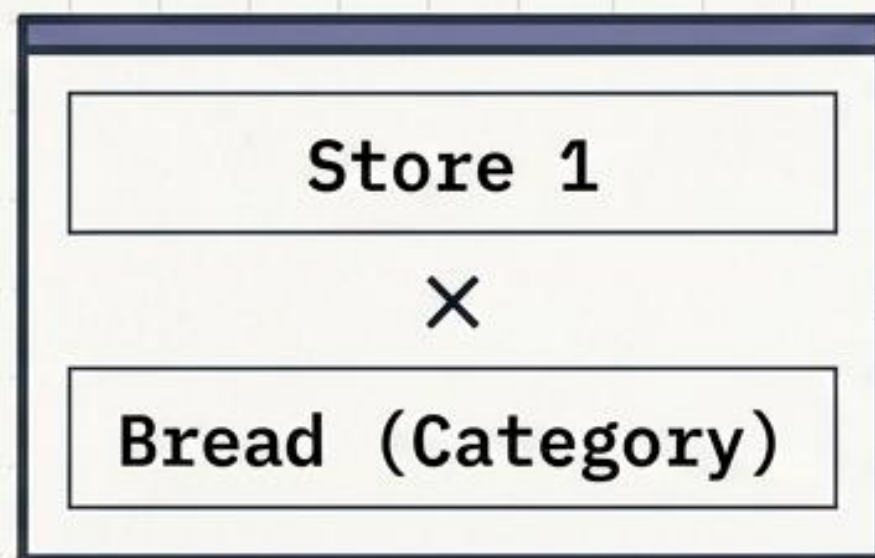
Phase 3: Target Encoding による上級への突破口

Score Badge

Current Score

0.464 → 0.440

上級者の壁への到達

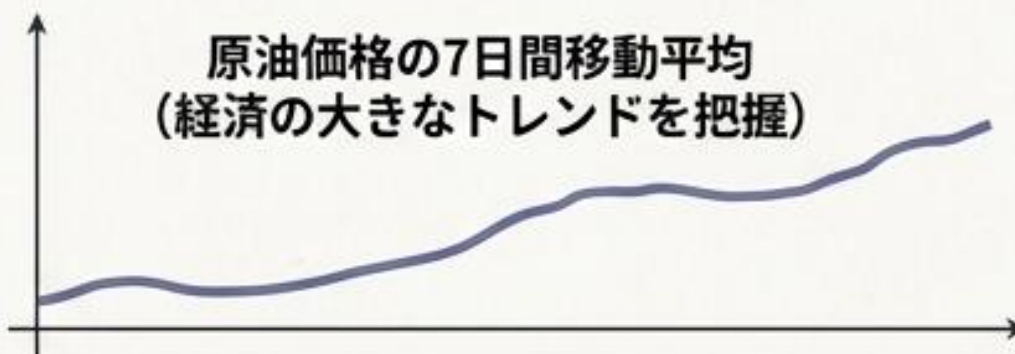


Target Encoding
(学習データのみで計算)



ベースライン数値化:
平均500単位

原油価格の7日間移動平均
(経済の大きなトレンドを把握)



インサイト: 『店舗×商品』という細かいカテゴリの相互作用を、強力なベースライン数値に変換することで学習効率が劇的に向上する。

Phase 4 : 精密射撃 (決定木の弱点を補うチューニング)

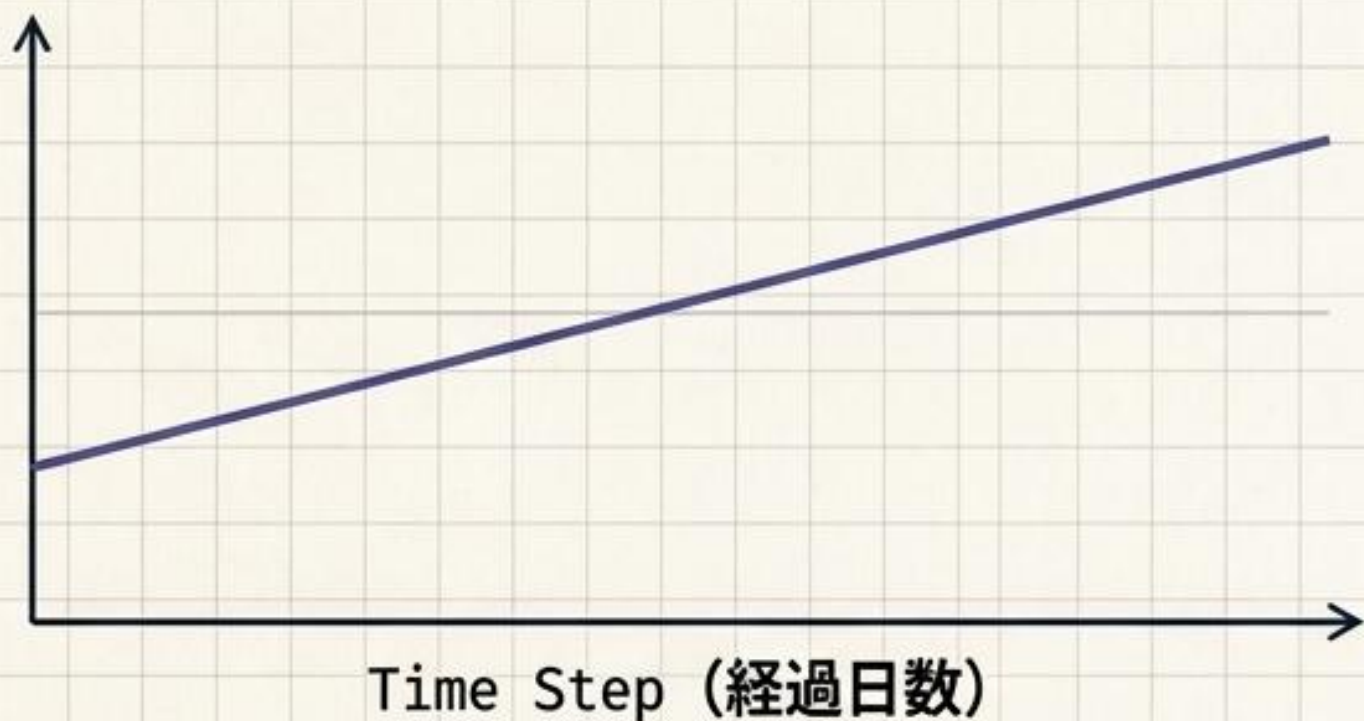
Score Badge

Current Score

0.440

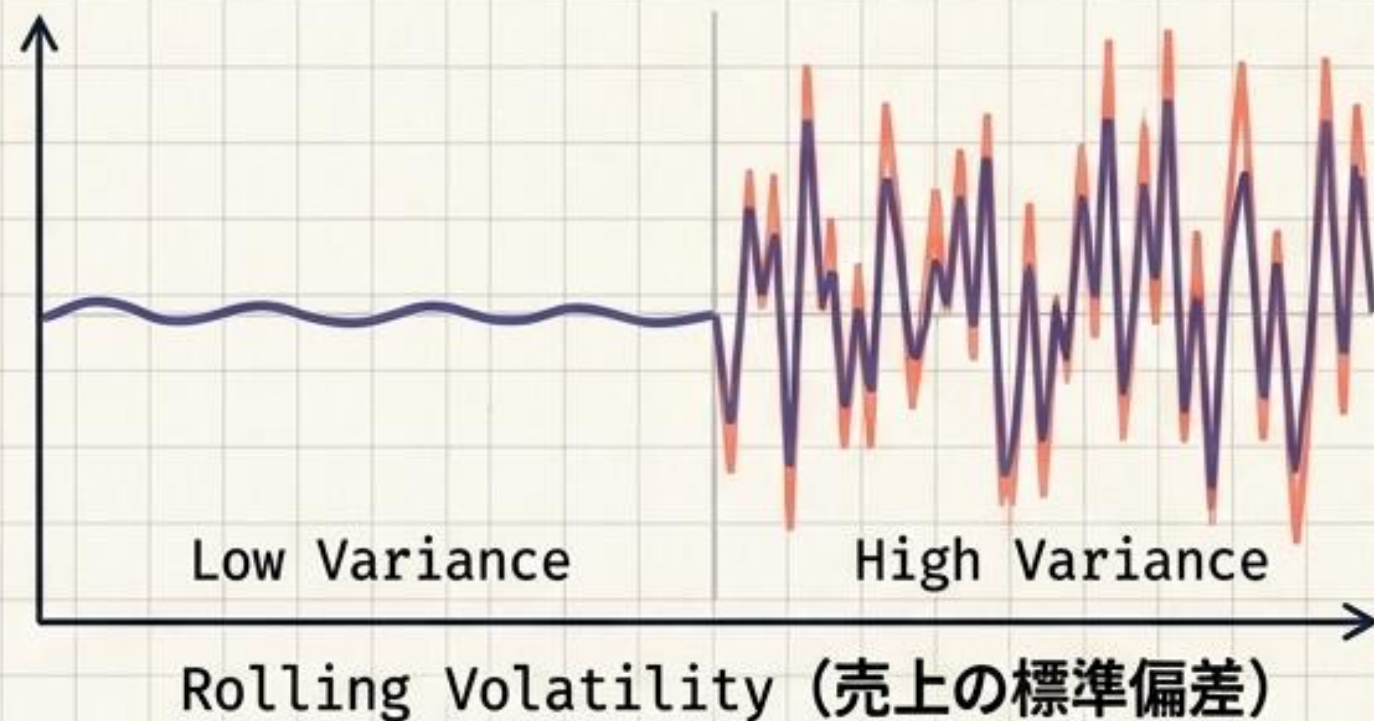
(最終突破への助走)

マクロトレンドの補完



決定木が苦手な『全体的な経済成長による右肩上がりのトレンド』を明示的に教え、将来の過小評価を防ぐ。

ミクロな変動特性の把握



毎日安定して売れる牛乳と、特売で爆発するプロモーション商品の『波の激しさ』を区別させる。

ハイパーパラメータ : `num_leaves=127` でパターン認識を深め、`L1/L2` 正則化で過学習を防止。

Final Breakthrough: 神は細部に宿る (ドメイン知識の極致)

Score Badge

Final Score

0.429

銀メダル圏内達成

今日 (Today)



Lag 364
(昨年と同曜日のトレンド)

364日前 (52週間前)



1月1日の除外:
エクアドルの元旦は全店休業 (売上0)。
8月予測において、この特異日は純粋なノイズ
となるため完全排除。

learning_rate=0.015, feature_fraction=0.6

進化の系譜: 反復的ブレイクスルーの統合マトリックス

Step	Score (RMSLE)	Model	追加された特徴量	得られたインサイト
1	0.902	Linear	Calendar	単純な暦だけでは時系列の複雑さを捉えきれない。
2	0.761	Ridge	Oil + Holiday	経済指標と祝日が売上に直結することを証明。
3	0.471	LightGBM	Time Step	決定木モデルによりカテゴリ間の非線形な関係を学習。
4	0.461	LightGBM	Lags + Mean	過去の売上が最大の予測因子。0.5の壁を突破。
5	0.464	Ensemble	Ridge added	スケールの不一致がノイズを生むという失敗からの学び。
6	0.445	LightGBM	Target Encoding	カテゴリの平均ベースラインが学習効率を劇的に向上。
7	0.440	LightGBM	Volatility	店舗×商品の細かな特性と波の激しさをモデル化。
8	0.429	LightGBM	Lag 364, Jan 1 X	高度なドメイン知識とノイズ除去が最後の壁を破る。

Key Takeaways: AI時代のデータ分析戦略



全体像のコンテキスト共有

単にコードを要求するのではなく、**データスキーマ、コンペのルール、Kaggleの議論**をAI Studioの200万トークンの文脈に全投入する。



仮説に基づいた指示

「**特徴量を作れ**」ではなく、「**経済の遅行性を捉えるため原油価格の7日移動平均を作れ**」と、人間の仮説に基づくエンジニアリングを指示する。



エラーと限界のディスカッション

文法エラーの修正だけでなく、「**線形モデルと決定木のアンサンブルがなぜ失敗したか**」という構造的・数学的な限界についてAIと議論する。

AIは単なるコーダーではなく、最強の壁打ち相手である。