ChatGPT

J-Quants APIデータを活用したマルチホライズン株価予測データセット提案

提案概要とデータソース一覧

J-Quants公式APIが提供する多様な市場データと、既存プロジェクト「gogooku3」のコードベースで用意された特徴量を組み合わせ、**銘柄×日次のパネルDataFrame**形式でマルチホライズン（1日後/5日後/10日後/20日後）リターン予測に適した特徴量データセットを構築します。以下のデータソース別に、**未来情報を含まない形で**有用な特徴量へ変換する方法を提案します。特徴量はすべて数値型（float）または適切にエンコードされたカテゴリ型に統一し、モデル入力での整合性を保ちます。また、欠損値補完やスケーリングの指針についても言及します。最後に、各特徴量の重要度をHigh/Medium/Lowで評価し、情報係数(IC)やSharpe比改善への期待度を示します。

使用する主なAPIデータ:

- 株価日足OHLCV（終値を中心に調整株価使用） [1] [2]
- 信用取引残高（週次の信用買い残・売り残、および日々公表分） [3]
- 売買内訳データ（現物/信用の新規・返済別の売買代金・出来高） [4] [5]
- 空売り関連データ（業種別空売り比率、空売り残高報告など） [3]
- 投資部門別売買動向（投資主体別の売買代金・先物建玉など） [6]
- 財務・ファンダメンタル（財務諸表、配当情報、決算発表予定日など） [7]
- 銘柄属性（上場市場区分、17業種/33業種コード、規模カテゴリ、信用銘柄区分 等） [8]
- 市場指数（TOPIX、業種指数、スタイル指数など各種指数の日次値） [9] [10]

以下、各データソースから抽出する特徴量とその加工方法を示します。

株価・テクニカル指標に基づく特徴量

株価OHLCV（日足）からは、従来のテクニカル指標を計算します。これらは既存プロジェクトでも実装済みと思われますが、**マルチホライズン予測に有効な安定性のある特徴量**として引き続き活用します [11]。生の価格ではなくリターンやオシレーター系に変換し定常性を確保します [12]。

- **リターン系列**: 1日、5日、20日などの過去リターン（対数または単純）を特徴量化します [11]。例えば「前日リターン」「過去一週（5営業日）リターン」「過去1ヶ月リターン」などです。特に**直近の短期リターン**はモメンタム効果やリバーサル効果を捉え、モデル精度向上に寄与します [13] [14]（実際、前日リターンはモデルで最も重要な特徴量となり、前日下落銘柄ほど翌日リバウンド傾向を予測しました [13]）。マルチホライズンで各ターゲット期間に対応するよう、**過去h日間のリターン**（h=1,5,10,20など）をそれぞれ計算しておきます（**リーケージ対策**: 例えば5日後リターン予測には当日までのリターンのみを使用し、予測期間に重複しないよう適切にシフトします）。

- **移動平均乖離率**: 短期トレンドを捉えるため、5日移動平均や25日移動平均と現在価格の乖離率を算出します [15]。例: `MA5_gap = Close / MA5 - 1`（5日移動平均との差分）等。これにより株価の過熱感や反転シグナルを特徴量化できます。移動平均線からの乖離は**ローリングZスコア**に近く、定常性確保に有効です [12]。既存実装にある場合は重複となりますが、**信用・需給系指標との組み合わせでシナジー**が期待できます（乖離が大きくかつ需給指標も極端な場合など）。

- **ボラティリティ指標**: 過去一定期間の価格変動率（実現ボラ）を特徴量とします [15]。例えば5日、20日間の**ローリング標準偏差**（対数リターンで計算）を取ります。ボラティリティが高い銘柄は予測難易度が高い傾向もあり、モデルにボラ情報を与えることでリスク調整や異常値検知に役立ちます。また**出来高**も活用し、出来高の5日平均対比や出来高急増率などを特徴量化すると、短期的な注目度上昇を捉えられます。これらテクニカル指標は基本的なものですが、**High優先度**で導入すべき標準特徴量です（単体のICは中程度ながら、他特徴量との組み合わせでSharpe向上に不可欠 [14]）。

- **対TOPIX・対業種インデックス残差**: 個別銘柄の動きを市場全体や業種平均から切り離すため、当日または直近の**超過リターン**を特徴量とします。例として「当日リターン － 当日TOPIXリターン」や「過去5日リターン － 過去5日業種指数リターン」等を計算します。これにより市場要因を差し引いた個別要因の強さを定量化できます [16]。「対TOPIX残差」はマクロ要因中立化に役立ち、モデルが純粋なアルファ信号を捉えやすくなるため**優先度High**です（市場全体と無関係に動く銘柄ほど異常値として予測しやすくなり、IC改善が見込まれます）。業種指数に対する超過リターンも同様に算出し、**業種固有のトレンドとの差異**を捉えます。

- **価格モメンタム・リバーサル指標**: 上記リターン系列からさらに、**モメンタム**（中期的持続傾向）や**リバーサル**（短期的反転傾向）のファクターを構築できます。たとえば12週程度の累積リターン（中期モメンタム）と1週間程度のリターン（短期リバーサル）の組み合わせは伝統的ファクターです。J-Quantsのチュートリアルでもモメンタム/リバーサル効果が確認されており [14]、これらは**High優先度**の特徴量です。期待ICは期間によりますが、短期リバーサルは1日〜1週先リターンに有意な正のICを持ち（前日下落銘柄の翌日反発傾向など [13]）、中期モメンタムは1ヶ月先程度まで弱い正のICを持つと考えられます。モデルにはこれらを直接特徴量として与えるか、過去リターンから学習させることも可能ですが、明示的に特徴量化しておくことで学習を安定させます。

信用取引関連の需給特徴量（信用残・信用倍率など）

信用取引データ（証券金融会社発表の週次残高および日々公表分）からは、投資家の信用取引ポジションに基づく**需給指標**を構築します。信用買い・売り残は将来の売買圧力を示唆する重要な手掛かりであり、市場でも注目されています [17] [18]。これらの特徴量は**マーケット中期要因**としてSharpe改善に貢献すると期待されます。

- **信用買い残高・売り残高**: 週次の「銘柄別信用取引残高」API [3] から各銘柄の信用買い残株数（信用買建玉）と信用売り残株数（信用売建玉）を取得し、特徴量とします。絶対値では銘柄ごとに桁が異なるため、**流通株数に対する割合**や**時価総額に対する比率**にスケーリングすると比較可能です。また信用買い残と売り残を組み合わせて**信用倍率（買い残/売り残）**を算出します [19]。信用倍率が高い（買い残が極端に多い）銘柄ほど**将来的な売り圧力が強い**ため上値が重い傾向があり、逆に信用倍率が低い（1倍に近いかそれ以下）銘柄は**将来的な買い戻し圧力＝上昇余地**が大きいと考えられます [20] [21]。実際、「信用買い残は将来の売り圧力、信用売り残は将来の買い圧力」と言われています [18]。このため**信用倍率**やその変化は重要度Highの特徴量です（期待IC中程度〜やや高め。買い残過多銘柄は翌週以降パフォーマンスが低迷しがち、信用取組改善銘柄はアウトパフォーム傾向 [17] [18]）。モデルには信用買い残と売り残を別々に与え、ネットの需給バランスを学習させることも有効です。なお週次データの**公表タイミング**に注意し、公表日以降のサンプルにその週の残高を反映（ラグ導入）して未来情報のリークを防ぎます。

- **信用残高の変化率・増減フロー**: 週次残高の前週比増減を計算し、信用買い残増加率や売り残増加率を特徴量とします。信用買い残が増加している銘柄は投資家の強気姿勢を反映しますが、過熱感も示唆するため短期的には逆指標となる場合があります [17]。一方、信用買い残が減少に転じた銘柄は将来の売り圧力減少で好材料と捉えられることがあります [22]。こうした**信用残高のモメンタム**も有益な特徴量です。変化量は株数ベースより**率**や**対出来高比**にすると、銘柄間で比較しやすくなります。例えば「信用買い残増加率（今週買い残/先週買い残 - 1）」や「信用買い残の週間増加株数 / 発行株

数」を算出します。信用残増減はイベントドリブンな動きも多いため、単独でのICは中程度ですが、他のシグナルと組み合わせると有効（例:　業績悪化で信用買い残が減少＝需給好転だが悪材料伴う場合は慎重に解釈 [23] ）。**優先度Medium**程度ですが、ポートフォリオ全体のSharpe改善に寄与し得ます。

- **日々の信用新規・返済売買推計**: 日次の売買内訳データ（後述）を活用して、**日々の信用建玉残高を推定**することも可能です [24] 。売買内訳には信用取引による新規買い・返済売り、および新規売り・返済買いの出来高が含まれており [25] 、これを累積することで週次発表まで待たず**各日の信用買い残・売り残のおおよその推移**を把握できます [26] 。例えば「当日の信用買い残推定＝前営業日の残高 + 信用新規買い株数 − 信用返済売り株数」といった計算です。この推定残高を特徴量として持たせれば、直近の信用需給の変化をタイムリーに反映できます（週次データを補間する形）。実装がやや複雑なため**優先度Medium**ですが、モデルが需給変化点を捉えやすくなるメリットがあります。

- **信用取引関連その他**: 銘柄ごとの**信用区分**（制度信用銘柄か否か、貸借銘柄か否か）も特徴量になります。J-Quantsの銘柄情報にはMarginCodeがあり、例えば「貸借銘柄（制度信用取引で買建・売建可）」「制度信用買のみ可（貸借不可）」等の区分が取得できます [27] 。信用取引が可能な銘柄ほど個人投資家の投機資金が入りやすく、信用データも存在します。**MarginCode**をカテゴリとしてモデルに与えることで、「信用取引不可銘柄では信用需給特徴量は常にゼロ」という構造を学習させる助けになります（該当しない特徴量の影響を自動で無視しやすくなる）。また信用取引不可の銘柄は需給面で特殊な動きをする可能性があるため、識別する意義があります。これは**優先度Low**（間接的な効果）ですが、特徴量の解釈を助ける補助的カテゴリ情報として追加が望ましいです。

売買内訳データに基づく特徴量（現物vs信用、新規vs返済、空売り比率など）

JPXの売買内訳データ [4] は、個別銘柄の日次出来高・売買代金を**現物/信用別かつ新規/返済別**、さらに空売り区分別に分解した非常に粒度の細かいデータです [28] 。このデータからは、投資家行動の詳細を捉える特徴量を構築できます。特に**信用取引のフロー（新規建て・決済）**や**空売り動向**が直接把握できる点で貴重であり、予測モデルのアルファ源泉として**優先度High**と位置付けます。

- **信用新規買い・信用新規売り比率**:　売買内訳から抽出できる「信用新規買い」「信用新規売り」の出来高を用いて、それぞれ**当日買い注文の何割が信用新規か、当日売り注文の何割が信用新規（=空売り）か**という比率を算出します。 [16] にあるように、**信用買い比率＝信用新規買い / 買合計、空売り比率＝(信用取引以外の空売り + 信用新規売り) / 売合計**と定義できます [29] 。信用買い比率が高い銘柄は「その日の買いの多くが信用取引（レバレッジを掛けた買い）によるもの」であり、個人投資家主体の投機的買いが旺盛であることを示唆します。一方、空売り比率が高い銘柄は「その日の売りの多くが新規の空売り（信用 or 制度外）によるもの」で、弱気な見方・ヘッジ売りが多い状況です [16] 。一般に**信用買い比率が極端に高い局面は過熱感を示し短期的には逆方向の反動が起きやすい**一方 [18] 、**空売り比率が高い局面では将来的な買戻し（ショートカバー）による上昇余地も生まれる** [18] 。したがって、これら比率の**水準（絶対値）**や**異常度（過去との乖離）**を特徴量とすることで、需給の偏りをモデルに織り込めます。具体的には、信用買い比率・空売り比率それぞれの**当日値**に加え、**直近20日平均からのZスコア**や**一年間の分位（percentile）**などを計算し、異常に高い/低い日を検出します。これらは新しい情報源であり**優先度High**、ICも高めが期待できます（例えば信用買い比率が上位5%の日の翌週株価は平均アンダーパフォーム、といった傾向が想定されます）。

- **現物vs信用のネットフロー**:　　売買内訳から**現物買い/売り代金**と**信用買い/売り代金**も取得できます [30] 。これにより「**現物の純買越額**」と「**信用の純買越額**」をそれぞれ日次で算出可能です（買い代金－売り代金）。信用取引の買越が大きい日は個人のレバレッジ買いが流入した日、現物買越が大きい日は機関投資家等の現金買いが入った日と推察できます。特に**信用買越の大きな増加**は短期的に価

格上昇に寄与しますが、その後の反動売り圧力も生みます。一方で**信用売越（信用投資家が売り越し＝ポジション解消）**が進む局面は需給健全化としてポジティブに働く可能性があります。このような**現物・信用別の資金フロー**指標を日次特徴量とすることで、どの主体が株を買っているか/売っているかをモデルに認識させられます。過去の同種指標の分析例では、信用買い越し急増銘柄は短期リターンが低迷しやすいことが示唆されます [17]。従ってこの指標も**優先度High**級で、単体ICはそこそこでも他特徴量と組み合わせてSharpe向上が見込まれます。

- **機関投資家 vs 個人投資家の動向**（推測）： 売買内訳の内訳項目には「信用取引以外の空売り（ShortSellWithoutMargin）」という区分があります [25]。これは**制度信用を使わない空売り**、すなわち機関投資家等が独自に株券を借りて行う空売りを指します [31] [32]。対して「信用新規売り」は個人中心の信用空売りです。二つを比較することで**機関 vs 個人の空売り動向**が推察できます。例えば**“機関投資家による空売り比率”＝信用取引以外の空売り / 売合計**、**“個人信用空売り比率”＝信用新規売り / 売合計**といった指標です。前者が突出して高い場合、機関投資家がその銘柄に強い弱気ポジションを取っていることを示し、将来的な大口の買い戻し余地や、悪材料に対するインサイダー的示唆の可能性もあります。後者が高い場合、個人が空売りに殺到している状況で、こちらも異例と言えます（※多くの銘柄では個人の空売り比率は小さいため、極端な値は何らかの事情を示唆 [32]）。これら細分化した空売り指標は**優先度Medium**です。単体ではノイズも多い可能性がありますが、他の特徴量と組み合わせて解釈することで予測に貢献すると考えます（例えば機関の空売り比率上昇＋業績悪化ニュース＝本格的な下落トレンドの予兆、等）。

- **出来高ファクターとの組み合わせ**: 売買内訳データ由来の指標はいずれも出来高に起因するため、総出来高との比率で正規化されています。必要に応じて**流動性フィルタ**（出来高が極端に少ない日のデータは信用できない等）や**閾値処理**（例えば出来高1万株未満の日は特徴量をゼロクリアする等）も検討します。プロジェクト既存の出来高急増シグナル等があれば、信用/空売りフロー指標とのシナジーを生みます。例えば出来高急増かつ信用買い比率急増であれば「個人投資家が群がった異常値」として強いシグナルとみなす、といったルールの発見につながるでしょう。

空売り・先物ポジション関連の特徴量

空売りデータと先物建玉に関しては、上記売買内訳に含まれるもの以外に、J-Quants APIから取得できる集計データも活用します。マーケット全体や業種ごとの空売り動向、先物ポジションは**市場心理の指標**として機能し、個別銘柄のリターン予測にも間接的に影響を与えます。

- **業種別空売り比率（市場全体）**: `/markets/short_selling` エンドポイントでは市場全体の業種別空売り比率が取得可能です [3]。例えば「プライム市場全銘柄の空売り比率（出来高ベース）40%」のようなデータです [33]。この値自体は個別銘柄に共通の**マーケット指標**となりますが、**日次で変動する市場センチメント**として全銘柄の説明変数に加えることができます。具体的には「本日市場全体で空売りが膨らんでいるかどうか」を示す指標として、全銘柄共通の特徴量列（例: 当日空売り比率(市場)）を追加します。モデルはこれを利用して、日ごとのマーケットモード（リスクオフで全体空売り増など）を学習できます。単独では銘柄選択には効きませんが、**マーケットレジームを捉える要因**として有用です（例えば全体空売り比率急上昇の日は短期リバウンドが起きやすい等の挙動を補足）。**優先度Low**（あれば良い程度）ですが、深層学習モデルでマーケット状態を認識させるために入れておくことを推奨します。

- **空売り残高情報**: 金融庁の規制により、大口空売りポジション（0.2%超）は公表されます。J-Quantsの `/markets/short_selling_positions` では銘柄ごとの空売り残高報告情報が得られる可能性があります [3]。もし取得できる場合、**「空売り残高報告件数」**や**「空売り残高割合（発行株数比）の最大値」**などを特徴量化できます。具体例: ある銘柄に残高報告が出ている＝大口投機筋がショートしている、を示すので、特徴量「大口空売りフラグ（有無）」や「報告残高割合（最大値または合計値）」を作ります。大口の空売り介入は将来の買戻しインパクトも大きく、短期リターンのボラティリティ

要因となります。これも**Low〜Medium優先度**ですが、実装可能なら差別化要因になります（特に報告残高が増加傾向の銘柄は弱気シグナル、高水準から減少に転じたら好材料、等の解釈が可能で、IC改善に寄与する可能性があります）。

- **投資部門別の売買動向**: `/markets/trades_spec`からは投資主体別（海外投資家、個人、信託銀行etc.）の売買動向が取得できます [6]。通常これは**市場全体の週間データ**（例: 先週外国人が現物株を○億買い越し、先物を△枚売り越し）です。個別銘柄には直接紐付きませんが、マクロ指標として日次特徴量に組み込めます。例: 毎週木曜発表の「外国人投資家の先物建玉」が急増/急減した週かどうかをその週の各日に持たせる、など。外国人の大きな買い越しは相場全体上昇への期待材料、売り越しは下落リスク材料となり得ます。このような市場マクロ要因は**Low優先度**ではありますが、モデルのバックグラウンド要因として入れておくと安定性が増します。深層学習モデルでは時系列の文脈としてこうした共通要因を捉えられるため、全銘柄に同じ値が入る特徴量であっても有用です。なお公表頻度・ラグがあるため、利用時はそのタイミング以降に一定期間その値を保持（例えば週次データを週内日次に適用）する形で実装します。

- **先物市場のテクニカル**: `/derivatives/futures`から日経先物やTOPIX先物の建玉や取引情報も取得できます [34]。先物価格や建玉に関する指標（例: 建玉の増減＝投機的資金の流入/流出）は市場先行指標として使われます。これも全銘柄共通となりますが、**先物価格の当日変化率**や**建玉残高の増減率**などを特徴量に加えることができます。例えば先物主導で大きく市場が動いた日は個別銘柄にとってもトレンドフォロー/リバーサルのヒントになります。優先度は低めですが、市場全体の状態をモデルに伝える一手段です。

財務・ファンダメンタル・イベント関連の特徴量

財務諸表データや決算・配当情報からは、中長期のファンダメンタル価値や企業イベントを表す特徴量を構築します。マルチホライズンが最大20営業日（約1ヶ月）程度であるため、ファンダメンタル指標の直接的な効果は限定的かもしれません。しかし**バリュエーションや業績モメンタム**はベースラインのリターンに影響し、中期的なアルファ源泉となります。また決算発表や配当といったイベント周りのフラグは**短期リターンの分布**に大きな変化をもたらすため、モデルにこれら情報を与えておくことは重要です。

- **バリュエーション指標（バリュー因子）**: 財務情報API（/fins/statements, /fins/fs_details など）から取得した直近の財務数値と株価を組み合わせ、**PER, PBR, 配当利回り, EV/EBITDA**等のバリュエーション指標を計算します [35]。例えば**PBR（株価純資産倍率）**は`時価総額 / 純資産`、**ROE**は`当期純利益 / 純資産`などです。J-Quants APIで必要な項目（EPSやBPS、利益、資産など）が取得可能であり、大半は算出可能です [35]。こうした指標は**低PBR・高利回り株は将来的にアウトパフォームしやすい**等の伝統的ファクター効果を持ちます。ただし1ヶ月程度の予測ホライズンでは影響は緩やかであるため、単体ICは低め（長期では有意でも短期ではノイズに近い）。それでも**優先度Medium**として導入を推奨します。理由は：(1) モデルのベースラインとして組み込むことで、極端に割高/割安な銘柄に対する予測のブレを減らせる（例えば割高成長株は恒常的にリターン平均が低めなのでモデルが的外れな強気予測をしにくくなる）、(2) ATFT/GATのようなモデルで**銘柄間の関係性**を学習する際に、バリュエーションは重要な共通要因となり得るためです。具体的な特徴量候補: **予想PER**（または実績PER）, **PBR**, **予想配当利回り**, **利益成長率（今期予想EPS/昨期EPS - 1）**など。予想値を使う場合は適時開示情報やアナリスト予想が必要ですが、利用可能範囲で算出します。

- **クオリティ指標（収益性・成長性）**: ファンダメンタル面では**ROE, ROA, 営業利益率, 売上/利益成長率**など企業の収益力・成長力を示す指標も特徴量とします [35]。これらも短期予測への直接的寄与は大きくないものの、銘柄特性を表す**静的特徴量**としてモデルに組み込む価値があります。Deep Learningモデルではこれらを**銘柄embedding**に学習させることも可能ですが、最初から入力特徴量として与えても良いでしょう。例えば**ROE**が高い（収益性高い）銘柄は下落局面でも下値が堅い傾向があるかもしれず、モデルがそうした挙動を捉える助けになります。優先度はValue因子と同様

Medium程度です。なお財務指標は**四半期ごと**にしか更新されないため、直近決算発表時に値を更新し、それ以外の期間は**前回値を保持**する形で欠損なく扱います（発表翌日から新値を使用しリーク防止）。

- **規模（サイズ）**: 時価総額や発行株式数も重要なファクターです [35]。一般に**小型株はボラティリティが高くリターン分布も肥尾**である一方、**大型株は安定的**という傾向があります。また小型株には流動性制約から来る固有の需給要因があるため、モデルに規模を認識させることは有用です。特徴量として**対数時価総額**（ログマーケットキャップ）や**浮動株比率**などを組み込みます。J-Quantsの銘柄一覧情報から時価総額や発行済み株数が取得できますので、それを用います。優先度はHigh寄りのMedium程度です（IC自体はそれほど高くありませんが、モデルの汎化性能向上に寄与します）。既存プロジェクトで規模を考慮済み（例えば銘柄選定をTOPIX500に限定する等 [36] [37]）であれば重複しますが、特徴量としても保持しておくとモデルが銘柄間比較を適切に行えます。

- **決算発表イベント**:　決算発表は短期株価に大きな変動をもたらすイベントです。J-Quantsの `/fins/announcement` から各銘柄の決算発表予定日を取得できます [38]。これを用いて、**「決算発表前後○日の期間フラグ」**を特徴量とします。例えば決算発表日の前後5営業日は `Earnings_window=1` などとし、それ以外を0とする特徴量です。モデルはこのフラグから「決算直後はリターン分布が通常と異なる（ギャップアップ/ダウンのリスク）」ことを学習できます。実際、サプライズ決算は短期的に急騰・急落を引き起こしやすく、また発表直前は様子見で株価変動が収まる傾向もあります [39]。フラグを与えることで**モデルが自動で不確実性を織り込む**ことが期待できます。決算発表自体をまたぐようなホライズン（10日後や20日後）もあり得るため、その場合モデルはフラグと他のシグナルから適切な予測幅を調整するでしょう。なお結果の良否（サプライズのプラス/マイナス）は事前にはわかりませんが、**オプションとして**決算発表日の**事後リターン**を履歴に組み込んでおき、類似の状況学習に使う手もあります（要注意:　これは将来予測には使えない情報なのでラベル扱いになる）。基本は**決算予定日フラグ**で十分でしょう。この特徴量は**優先度High**です（直接ICを持つものではないものの、モデルの誤差低減に大きく寄与するためSharpe向上効果は高いと考えます）。

- **配当イベント**:　配当も株価に影響します。具体的には**権利落ち日**に配当相当分だけ株価が下落するため、調整が必要です。J-Quantsの `/fins/dividend` から配当予定額や権利確定日が取得できます [38]。**対応策**としては、**株価を配当落ち調整済みの終値で分析する**ことが第一です（J-Quantsの日足APIで調整株価を取得できます [40]）。それでもモデルに認識させたい場合、**権利落ち日フラグ**を特徴量に加えます。権利落ち日は高確率でマイナスリターンとなりますがモデルには「配当による機械的な下落」と伝えることで、誤ったシグナル解釈を防げます。加えて**予想配当利回り**（上述のバリュエーション指標の一つ）も長期要因としてインプットしておくと、配当利回りの高い銘柄群に共通の動きを学習できます。これら配当関係の特徴量は**優先度Medium**です。権利付き最終日は決算ほどボラティリティ増には繋がりませんが、持株のコスト/インカム見通しを左右するため、中期リターンへの影響があります。

- **適時開示（ニュース）情報**: 個別銘柄の突発ニュースやIR（適時開示）は短期株価に大きな影響を与える場合があります。J-Quants　API自体には適時開示の内容までは含まれていませんが、決算以外の重要イベント（増資・株式分割・大型受注・不祥事など）を特徴量に反映できれば予測精度は上がります。簡易的には**「当日適時開示件数」**や**「当日重大開示フラグ」**を入れる方法があります。適時開示の件数データを入手できるなら、当日の開示リリース本数を特徴量にし、本数>0の場合はニュース有りとモデルに伝えます。より進めて、開示のポジティブ/ネガティブを判定する自然言語処理を行い**好材料フラグ/悪材料フラグ**を作ることも可能ですが、こちらは高度であり本プロジェクトのスコープ次第です。**優先度Low**（任意）とします。代替策として、株価が**異常値変動**した場合に事後的にモデルが対処できるよう、ボラティリティ特徴量で吸収させたり、学習時に外れ値データポイントにロバストな手法（例: 損失関数をHuberにする等）を用いることで対応します。

銘柄属性に基づく特徴量（業種・市場区分・スタイルなど）

銘柄の静的属性もモデルに与えるべきです。これは直接的な予測シグナルではありませんが、**銘柄ごとの固有効果**や**他銘柄との関係性**を学習する上で重要な手掛かりとなります [41]。ATFT-GAT-FAN系のモデルでは、銘柄をノードとしたグラフ関係を学習したり、銘柄Embeddingで固有特徴を表現する可能性が高いため、以下のような属性データを特徴量化します。

- **業種コード**: JPXの上場銘柄一覧APIから17業種コードや33業種コードが取得できます [1]。**どの業種に属するか**を示すカテゴリ特徴量を用意します。方法としては**One-hotエンコーディング**（例: 17業種なら17次元のバイナリベクトル）や、もしモデル側でEmbedding可能なら**カテゴリID**を与える形でも良いでしょう。業種によって株価の構造的なクセ（景気敏感かディフェンシブか等）が異なるため、モデルは業種情報から**リターンの共分散構造**を学習できます。また業種別のインデックスや平均値と組み合わせて、**業種内相対価値**を判断することも可能です。業種特徴量自体は**優先度High**です（銘柄embeddingを通じてSharpe向上や汎化性能向上に繋がる [10]）。既存プロジェクトでも業種を特徴量に加えることが示唆されています [41] ので、重複しても必ず含めます。

- **上場市場区分**: プライム市場、スタンダード市場、グロース市場といった**市場区分**をカテゴリ特徴量にします。市場区分は流動性や投資家層の違いを反映しており、例えばプライム市場銘柄は海外機関投資家の影響が大きい、グロース市場銘柄は個人主体でボラティリティが高い、などの傾向があります [42] [43]。実際、**プライム市場銘柄の中央値空売り比率は約40%と高く、スタンダード/グロースは15~20%程度と低い**というデータもあります [44]。このように市場カテゴリで需給構造が異なるため、モデルに区分を教える意義は大きいです。エンコーディングは業種同様にOne-hot化するか、3区分＋「その他（ETF等）」くらいであればバイナリで持たせても良いでしょう。優先度はMediumです。

- **規模カテゴリ・スタイル**: 上記の時価総額を連続値で入れる他に、**サイズカテゴリ**（例: TOPIX Core30，Large70，Mid400 といったインデックス区分 [37]）を特徴量として加えることも考えられます。J-Quants銘柄一覧のScaleCategoryから取得可能です [37]。例えばCore30は超大型株、Mid400は中型株群という分類です。サイズはすでに数値で入れるため冗長かもしれませんが、**スタイルインデックスとの関連**を見るなら有効です。スタイル指数として**JPX日経400**などが典型ですが、J-Quants提供指数にスタイル分類があれば、**その指数への採用フラグ**も特徴量となります（採用されている=質が高い等の示唆があるため）。ただし複雑化する割に効果は限定的かもしれないので、**優先度Low**とします。規模は数値情報で充分捉えられるため、カテゴリとしては無理に追加不要です。

- **銘柄固有ID/Embedding**: 厳密には特徴量ではないですが、ディープラーニングモデルで**銘柄IDに対応するembeddingベクトル**を学習させ、固定ファクターに近い働きをさせることも考えられます。これはモデルアーキテクチャの工夫なので詳細割愛しますが、上記業種・サイズなど静的特徴を加味したい場合にembeddingレイヤーに組み込むアプローチです。いずれにせよ、**銘柄を識別する何らかの特徴**（業種+サイズ+市場の組合せ等）はデータセット上に保持しておきます。

特徴量の前処理方針（欠損値対応・スケーリング・カテゴリ変換）

一連の特徴量を構築した後、**欠損値の補完**や**スケーリング**によってモデル学習に適したデータフレームを整備します。

- **欠損値処理**: 株価由来のテクニカル指標は計算上直近データが無い場合にNaNが生じますが、基本的に**初日など以外は欠損が出ないように設計**します。例えばリターンや移動平均は `fillna(0)` で初期欠損を埋める実装が可能です [45] [46]（初日のリターンは0と置く等）。週次信用残高など**低頻度データ**は、発表されない日は最新値を**前日値で埋めてホールド**します（例: 火曜に更新された信用残を水〜翌月曜まで同値とする） [26]。これにより各営業日でデータフレームを欠損なく埋めます。どうして

も値が存在しない場合（新規上場銘柄の過去指標など）は**中立的な値**で補完します。中立値の例:　リターンなら0、比率類の特徴量なら市場平均もしくは0、カテゴリなら「該当なしカテゴリ」を新設。特に**信用・空売り指標**で「その銘柄は信用取引不可/空売り不可」の場合は、当該特徴量は自然に0となる（取引が無い）ため、それで問題ありません。プロジェクト内でも`fillna(0)`により欠損埋めしている例があります [47]。ただし、0埋めが情報を失わないか注意が必要です。例えばファンダメンタルの増減率で0埋めすると「増減なし」と誤認されるため、その場合は別途**欠損フラグ**を立てる方法もあります。しかし煩雑になるので、今回は「基本的に欠損は前処理で極力発生させない」「発生した場合は0等で埋める」方針とします。

- **スケーリングと正規化**:　モデル収束を助けるため、数値特徴量は**スケーリング（標準化または正規化）**します。具体的には全銘柄・全期間の分布で標準化（平均0・標準偏差1）する方法、または特徴量ごとに5%-95%レンジでクリップしてMin-Max正規化する方法などが考えられます。Deep Learningでは内部でスケーリングを学習できますが、適度なスケールに揃えておくと学習安定性が増します。特に**金額や株数など桁が大きく異なる指標**（例: 時価総額 vs 利益成長率）を同程度のスケール感にします。ほとんどのテクニカル指標や比率は±数値に収まりますが、**歪度の大きい分布**（出来高変化率やPER等）は対数変換やランク変換も検討します。実際、金融データは過分散なため**ビニングやランク化**でロバスト性を上げる手法も有効です [48]。プロジェクトの既存実装で**五分位ビニング**を行っている例があります [48]。これは特徴量をカテゴリに離散化することで外れ値の影響を抑えるものです。同様に、例えば**リターン関連特徴量は中央値基準のZスコアにクリップ**、**PER等は対数を取った上で五分位に分割**といった工夫も可能です。今回は基本的に**標準化(Z-score)**をベースに、必要に応じて上限下限をウィンスorクリップする方針とします。なお**カテゴリ特徴量**（業種や市場区分など）はOne-hotにした後0/1のままで問題ありません（すでにスケール統一済み）。もしバイナリではなく数値IDでembeddingする場合も、embedding層が処理するのでスケーリング不要です。

- **カテゴリ変換**:　前述の通り業種コード・市場区分などは**One-hotエンコーディング**してDataFrameに格納します。例えば業種17区分なら`Industry_1`〜`Industry_17`の列を用意し、属する業種の列だけ1、それ以外0とします。こうすることでカテゴリも数値ベクトルとなり、他の特徴量と型を揃えられます。一部カテゴリには階層関係がありますが（例:　33業種は17業種の細分）、扱いが複雑になるためどちらか一方（代表的には17業種）を採用します。**銘柄コードそのもの**は数値ですが、機械的連番に過ぎないのでそのまま特徴量にはしません（前述のembeddingを用いる場合のみ利用）。**日付**もIndexとして用い、明示的な特徴量にはしませんが、もし曜日や月効果を取り入れるなら日付から**曜日One-hot**や**月エフェクト**を加えることも考えられます（今回は株式データの曜日効果は小さいので省略）。

以上により、最終的に**全銘柄×全営業日のパネルデータ**が完成します。各行はユニークな(銘柄,　日付)で、列には上記で設計した全特徴量（およそ数十〜百数十列程度）が揃います。ターゲット変数として1日後、5日後、10日後、20日後のリターン（対数リターンもしくは単純リターン）を別列で付与します。学習時にはリーケージ防止のためターゲット計算にも調整を入れます（例えば5営業日後リターン計算時に途中の配当落ちがあれば調整済株価で算出）。こうしたデータセットは**深層学習モデルのパイプライン**に直接入力可能であり、時系列モデル（ATFT等）では銘柄ごとの系列データに変換、グラフモデル（GAT等）では業種などで構築した関係グラフと併用する形で利用できます。

既存特徴量との重複・シナジー

gogooku3プロジェクト内ですでに導入済みと思われる特徴量との関係について整理します。先述のテクニカル指標（リターン、移動平均乖離、ボラティリティ等）は**既存のベース特徴量**として重複する可能性が高いです [2]。しかしこれらは引き続き重要な入力であり、本提案の新規特徴量と組み合わせることで**シナジー**を発揮します。例えば、既存で前日リターンやモメンタム特徴量を使っているなら、そこに信用買い比率や空売り比率といった需給系特徴量を加えることで、「**下落しているが信用買いが大量に入った銘柄**」など特殊な状況をモデルが識別できるようになります。これは単一の価格モメンタム特徴量では拾いきれない情報です。

また既存プロジェクトでファンダメンタルデータを一部導入済みであれば（例えばPBRやROEを計算済み）、それらは本提案の**Value/Quality特徴量**と重複します。重複部分は**再利用**し、漏れている指標（例:　配当利回りや成長率など）があれば追加します [35]。ファンダメンタル指標同士は相関が高いため、多く入れすぎると冗長ですが、モデルが重み付けできるので大きな害はありません。むしろ**多面的な財務健全性指標**を与える方が、特定の指標の欠点（例: 会計一時要因でPERが異常値になっている等）を補えます。

プロジェクト内で**ターゲットの計算**や**データラグ処理**について既に配慮されていれば、本提案でも同様の姿勢を取ります。たとえばJPXコンペでは**過去データに2週間の遅延**がある設定でしたが、J-Quantsプレミアムでは当日夜にデータ取得可能なので、最新データを使用します。その意味で、既存コードでの「データ取得遅延処理」が不要になるかもしれません。代わりに信用残高や週次データの**発表遅延**に注意が必要です。この点は上記でラグ導入を提案した通りです。

重複を避けるべき特徴量としては、同じ情報源から計算されるものを二重で入れないことが挙げられます。例えば「5日リターン」は既存にあり、本提案で「対TOPIX5日残差」として**(株価5日リターン - TOPIX5日リターン)**を入れるなら、前者単独の5日リターンはやや冗長になる可能性があります。ただ、モデルが適切に組み合わせを学習するため、**冗長性のある特徴量も許容**します（過度な次元削減はせず、重要度分析であとから絞り込めばよい [13] [49]）。従来特徴量と明確に重複するものは統合・削除も検討しますが、**Sharpe改善に寄与する可能性がある限り極力含める方針**です。特に、既存にない**需給系・イベント系特徴量**はすべて新規追加することで、既存モデルには無かったアルファ源を取り込めます。

特徴量の優先度と期待される効果まとめ

最後に、提案した特徴量群について重要度（優先度）と予想されるモデルパフォーマンスへの効果をまとめます。優先度はHigh/Medium/Lowの3段階で表記し、IC（情報係数）やSharpe比への寄与についてコメントします。

- **High優先度** – モデル予測精度に大きく貢献すると考えられる特徴量
- **テクニカルモメンタム/リバーサル系**（前日リターン、短期リターン、移動平均乖離など） [13] [11]
  理由: 短期の価格動きの癖を直接捉える基本特徴量。特に前日リターンのリバーサル効果は顕著で、モデルの主要な説明変数となりうる [13]。ICは単体で0.02〜0.05程度ながら安定しており、複数組み合わせることでSharpe向上。
- **需給フロー系**（信用買い比率・空売り比率、現物vs信用の資金フロー） [16] [18]
  理由: 他投資家の売買動向という価格以外の情報を提供し、新たなアルファ源泉となる可能性が高い。信用買い比率極端銘柄の短期リターンは逆方向に振れやすいなど、市場で実証的に語られる効果がある [17]。これらはICこそ中程度（0.01〜0.03想定）でも、価格系特徴と独立したシグナルを持ちSharpe改善に有効。
- **信用残高・信用倍率**（信用買い残/売り残水準、増減率） [18] [19]
  理由: 需給の蓄積状態を表し、特に信用倍率は将来の売買圧力バランスを示す先行指標。信用買い残過多はその後の上値抑制要因となり得るため、中〜長めのホライズンで効果を発揮（5日〜20日リターンへの影響）。ICは0に近い〜0.02程度と推測するが、極端値では有意差が出るためモデルの非線形学習と相性が良い。

- **業種・市場区分などカテゴリ情報** [42] [41]
  理由:　特徴量そのもののICは計算困難だが、モデルに銘柄の共通特性を学習させ大局的な判断を助ける。これにより誤検知や過剰適合を防ぐ効果が期待でき、結果的にSharpe向上に寄与。特に業種固有ショックの伝播などをモデルが理解できるようになる点が大きい。

- **Medium優先度** – ある程度予測に貢献し、他特徴量との相乗効果も見込める特徴量

- **ファンダメンタル価値系**（PBR、PER、ROE、成長率、配当利回り） [35]
  理由: 短期アルファとしては弱いが、銘柄の基本的な期待リターン水準を決める要因。割高株 vs 割安株で平均リターンに差があればモデルはそれを加味できる。ICは月次リバランス程度で0.01〜0.02と小さいが、安定した効果。20日先ターゲットには多少効く可能性。
- **イベントフラグ**（決算発表前後、配当権利落ち）
  理由: 直接的な予測というよりリスク管理要素。これを入れることで決算跨ぎの予測誤差が減りSharpe向上に繋がる。ICという観点ではフラグ=1時のリターン分布の特性変化を示すのみで0に近いが、モデルの全体的な精度・安定性を高める効果が大きい。
- **機関/個人別動向**（機関投資家空売り比率、個人信用動向など） [32]
  理由: Highの需給フロー指標をさらに分解したもの。単体ではノイズも多いが、特殊なケースを捉えると有効（例えば機関空売り急増＋株価下落でネガティブシグナル強化など）。IC向上幅は限定的だが、極端値でのモデル判断精度向上に資する。

- **マーケット指標**（市場全体空売り比率、投資主体別売買高、先物動向）
  理由:　　市場全体のムードを表す背景特徴量。個別銘柄選択には直接効かないが、モデルがその日のマーケット環境を認識する助けとなり、間接的に予測精度を上げる。Sharpe改善は小さいものの、全銘柄共通要因のフィルタリングに効果。

- **Low優先度** – 効果が限定的か不確実だが、余力があれば検討する特徴量

- **大口空売り残高情報**（空売りポジション開示）
  理由: データ頻度や網羅性の点で限定的だが、特定銘柄に大口ショートが入っている情報は希少で差別化要素になり得る。ただ頻繁に出るものではなくモデルに与える影響はスポット的。全体IC貢献はごく小さいと見込む。
- **適時開示ニュースフラグ/スコア**
  理由: テキスト解析等が必要で開発コスト大。入れればサプライズ検知に役立つ可能性はあるが、網羅性も課題。労力対効果が低いため優先度は下げた。代替としてボラティリティ等で間接対応。
- **細分カテゴリやスタイル**（33業種細分類、サイズ区分の重複表現など）
  理由: 業種17区分で概ね十分と考えられる部分をさらに細かくする提案。情報量は増えるが次元も増え過学習リスクも高まるため、効果は不明瞭。Deepモデルならembeddingで吸収可能なので無理に増やす必要なし。

以上の提案により、**gogooku3プロジェクトのMLパイプライン**に新たなデータセットを組み込みます。日次で銘柄ごとのテクニカル＋需給＋イベント＋ファンダメンタル特徴量が揃ったDataFrameを生成し、モデルではこれを用いて1日先〜20日先の複数ホライズンリターンを同時予測または個別予測します。新特徴量群は既存の価格系特徴量と補完関係にあり、情報リークを防ぎつつアルファの多様化を実現します。最終的な期待効果として、情報係数のトータルな底上げ（複数特徴量の組み合わせによるIC向上）と、モデルのSharpe比改善（予測精度向上とリスク低減の両面）を達成できると考えます。 [14] [17] 各特徴量の有効性は今後の検証を要しますが、本提案は現在入手可能なJ-Quantsデータを余すところなく機械学習モデルに活用する包括的なアプローチとなります。モデルの解釈可能性向上にもつながるため、ぜひ導入をご検討ください。

参考文献・出典: J-Quants APIリファレンス [6] [9]、J-Quants公式記事 [16] [26]、証券各社の投資情報 [17] [18]、およびJ-Quants株式分析チュートリアル [41] [13] などを参照しました。

---

[1] [3] [6] [7] [9] [34] [38] API仕様 | J-Quants API
https://jpx.gitbook.io/j-quants-ja/api-reference

[2] [10] [11] [12] [13] [14] [15] [35] [36] [37] [41] [45] [46] [47] [48] [49] ゼロから始める株式分析：J-Quants APIを用いた日本株寄り引けロングショート戦略分析｜botter_01
https://note.com/botter_01/n/nc5b1967ea962

4 5 8 16 24 25 26 27 28 29 30 31 32 33 42 43 44 〖J-Quants〗売買内訳データの紹介及びその利用例について #Python - Qiita
https://qiita.com/j_quants/items/ae423b14dcdcf819eb27

17 18 19 20 21 22 23 信用買い残が多い（減少する）とどうなる？どこで見れるか、調べ方も解説
https://kabukiso.com/column/idiom/kaizan_stock_down.html

39 サプライズ決算と株価変動の関係 〜銀行が融資したくなる決算書も
https://www.sin-kaisha.jp/article/settlement/%E3%82%B5%E3%83%97%E3%83%A9%E3%82%A4%E3%82%BA%E6%B1%BA%E7%AE%97%E3%81%A8%E6%A0%AA%E4%BE%A1%E5%A4%89%

40 株価四本値(/prices/daily_quotes) | J-Quants API - GitBook
https://jpx.gitbook.io/j-quants-ja/api-reference/daily_quotes