



J-Quants

- ・ファンダメンタルズ分析チャレンジ
- ・ニュース分析チャレンジ

解法について

2021年7月19日
UKI@blog_uki

自己紹介

<現職>

- ・システム開発、AI開発、金融ストラテジー開発
- ・自社資金運用 ※非金融業

<バックグラウンド>

- ・元はハードウェアエンジニア（車載機器設計）
- ・2014年から現職、2015年頃から業務へ機械学習を導入

<研究内容>

- ・ファイナンスデータの分析および収益化のためのFrameworkの研究
- ・データサイエンティストではありませんのでご留意願います。
（最新の機械学習手法やKaggleで使うようなテクニックはワカナイ）

所見：株×データサイエンスについて

- ・一般に手に入るデータ（構造化データ）で十分収益化が可能
- ・株×データサイエンスの2大問題

(1)低S/N比

- ・モデリングで対処可能
- ・実売買で制約を受ける

(2)非定常性

- ・市場は常に変化する
- ・よって可読性が不可欠

現時点においてファイナンス（マーケット予測）における機械学習の役割は複雑なモデリングではなく市場探索のためのツール。

ファンダメンタルズ 分析チャレンジ

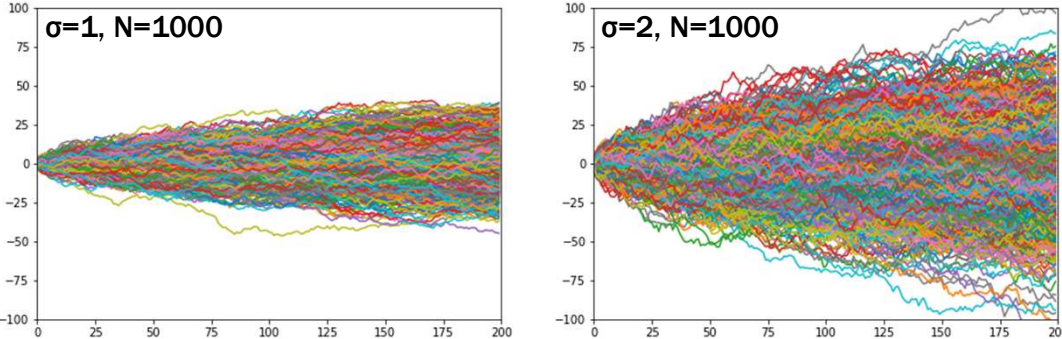
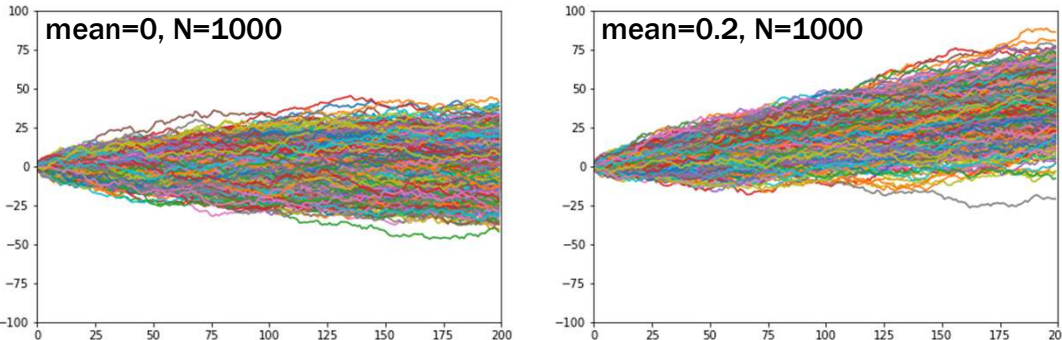
体系的なアプローチ

参考文献：

執行戦略と取引コストに関する研究の進展（杉原, 2012）

High Frequency Trading in a Limit Order Book（Avellaneda, Stoikov, 2008）

高値／安値を当てる問題はランダムウォークがそのベースとなる。

基 本 的 特 性	説 明	予測可能性
<p>①ボラティリティと高値／安値の関係</p> <div data-bbox="165 619 1223 959"></div>	<p>ボラが大きいと 高値：大（正相関） 安値：小（負相関）</p>	<p>比較的容易 自己相関あり</p>
<p>②トレンド（方向予測）と高値／安値の関係</p> <div data-bbox="165 1086 1223 1426"></div>	<p>トレンド発生すると 高値：大（正相関） 安値：大（正相関）</p>	<p>本来難しい 今回に限り 比較的容易 決算という カタリストの存在</p>

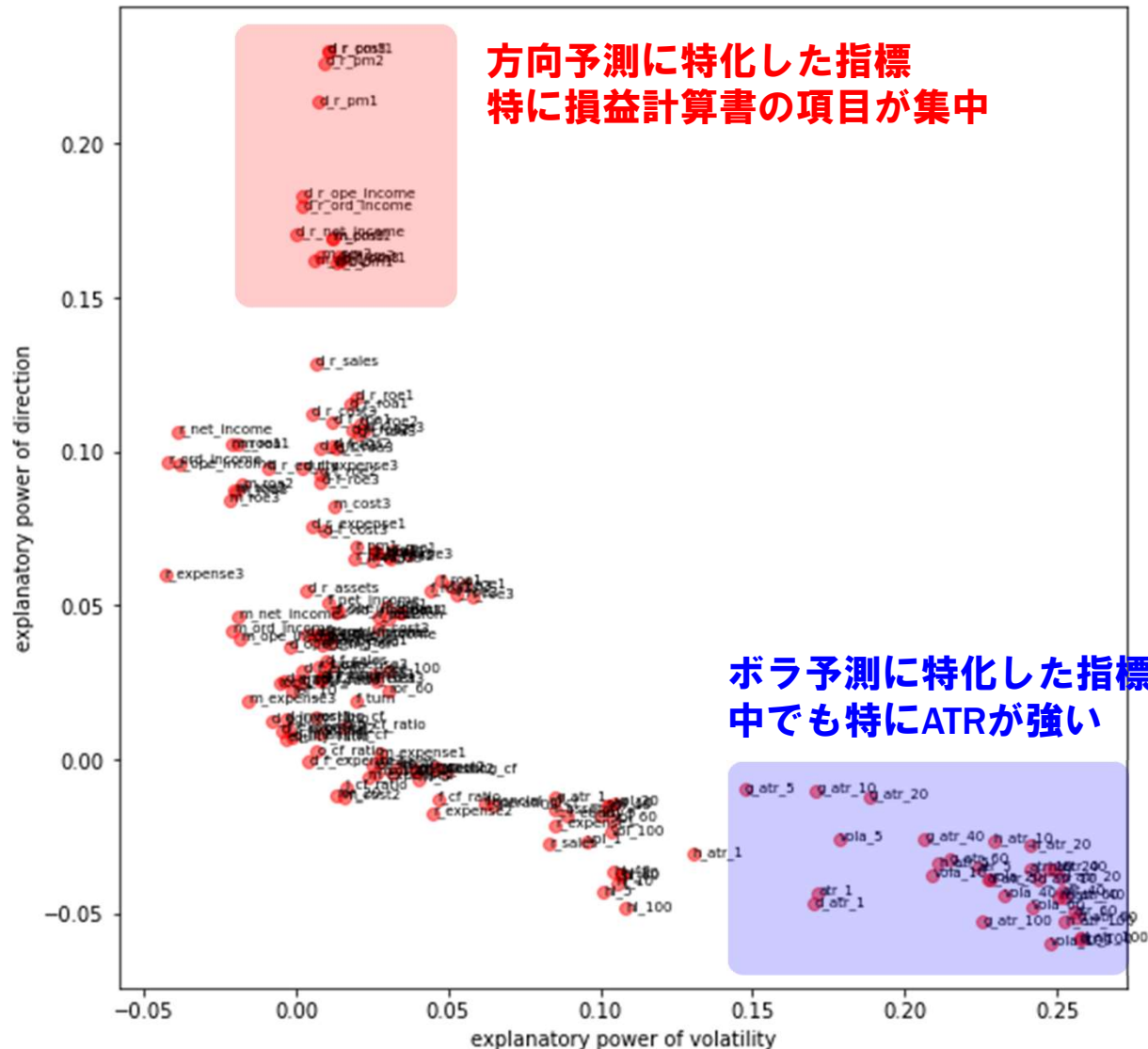
ランダムウォークにおける上記特性がモデルのベースとなる

特徴量一覧

情報ソース		名 称	記 号	備 考	総 数	
価格データ		終値	close		1	63
		騰落率	ror_n	n=1,5,10,20,40,60,100	7	
		売買代金平均	vol_n	出来高×価格 n=1,5,10,20,40,60,100	7	
		売買代金乖離率	d_vol	vol_1/vol_20	1	
		ATR(Average True Range)	atr_n	n=1,5,10,20,40,60,100	7	
		ATR乖離率	d_atr	atr_1/atr_20	1	
		ATR(ギャップ部)	g_atr_n	n=1,5,10,20,40,60,100	7	
		ATR(胴体部)	d_atr_n	n=1,5,10,20,40,60,100	7	
		ATR(ヒゲ部)	h_atr_n	n=1,5,10,20,40,60,100	7	
		ボラティリティ(標準偏差)	vola_n	n=5,10,20,40,60,100	6	
		高値安値バンド	hl_n	n=5,10,20,40,60,100	6	
		流動性	mi_n	値幅÷売買代金で模擬 n=5,10,20,40,60,100	6	
財務データ	ダミー	本決算フラグ、訂正フラグ	annual, revision		2	119
	raw 特徴量	売上高/営業利益/経常利益/純利益	*_sales, *_ope_income *_ord_income, *_net_income	*=r, f (r:result, f:forecast)	8	
		各コスト(上の指標の差分)	*_expense1~3	*=r, f (r:result, f:forecast)	6	
		総資産、純資産	r_assets, r_equity		2	
		キャッシュフロー(営業/財務/投資)	*_cf	*=operating, financial, investing	3	
	ratio 特徴量	売上高利益率(純利益/経常/営業)	*_pm1~3	*=r, f (r:result, f:forecast)	6	
		ROE(純利益/経常/営業)	*_roe1~3	*=r, f (r:result, f:forecast)	6	
		ROA(純利益/経常/営業)	*_roa1~3	*=r, f (r:result, f:forecast)	6	
		売上高コスト率	*_cost1~3	*=r, f (r:result, f:forecast)	6	
		売上高回転率	*_turn	*=r, f (r:result, f:forecast)	2	
		財務健全性	equity_ratio		1	
		総資本CF比率	*_cf_ratio	*=o, f, i	3	
	diff 特徴量	raw特徴量の前期差分	冠詞にd_を加える		19	
		ratio特徴量の前期差分	冠詞にd_を加える		30	
		raw特徴量とratio特徴量のうち、 実績と予想の差分が取れるもの	冠詞にm_を加える	売上高回転率は除く	19	

価格系63、財務系119、計182の特徴量を作り網羅的に探索した

特徴量マップ



＜考え方＞

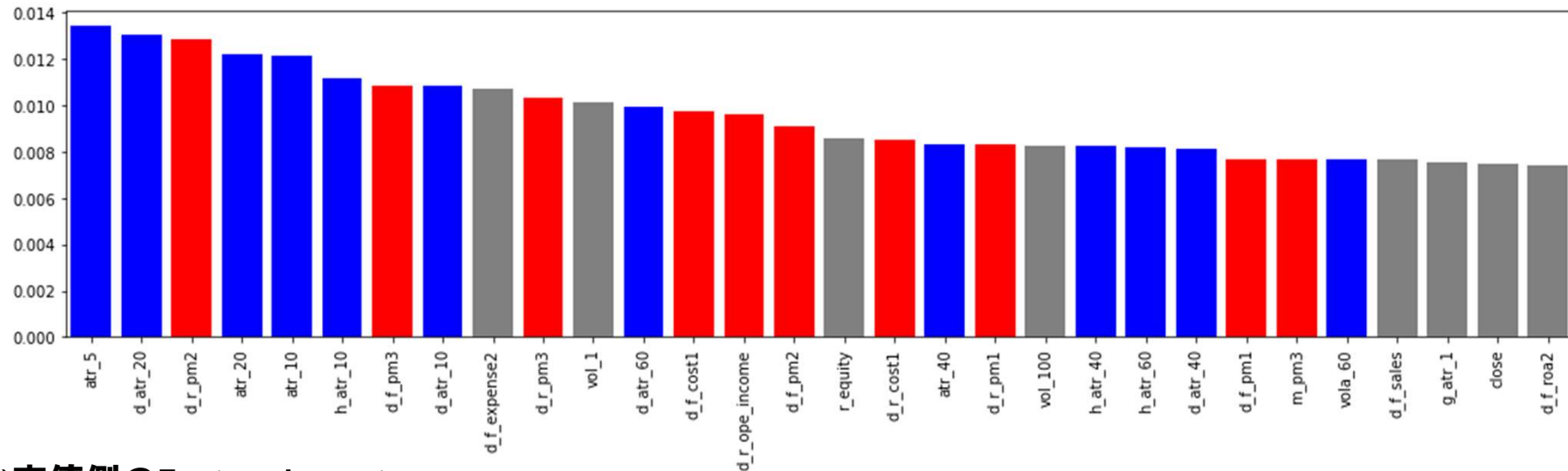
- ボラ予測に対する特徴量は、
高値側で**正**相関、安値側で**負**相関
- 方向予測に対する特徴量は、
高値側で**正**相関、安値側で**正**相関
- ボラ予測の評価指標として
(高値側Rank相関 - 安値側Rank相関)/2
- 方向予測の評価指標として
(高値側Rank相関 + 安値側Rank相関)/2
を計算してマッピング。

特徴量によって得意分野が偏り且つそれぞれの予測精度が非常に高い (IC>0.2)
ボラ予測か方向予測か考えながら特徴量を生成する。

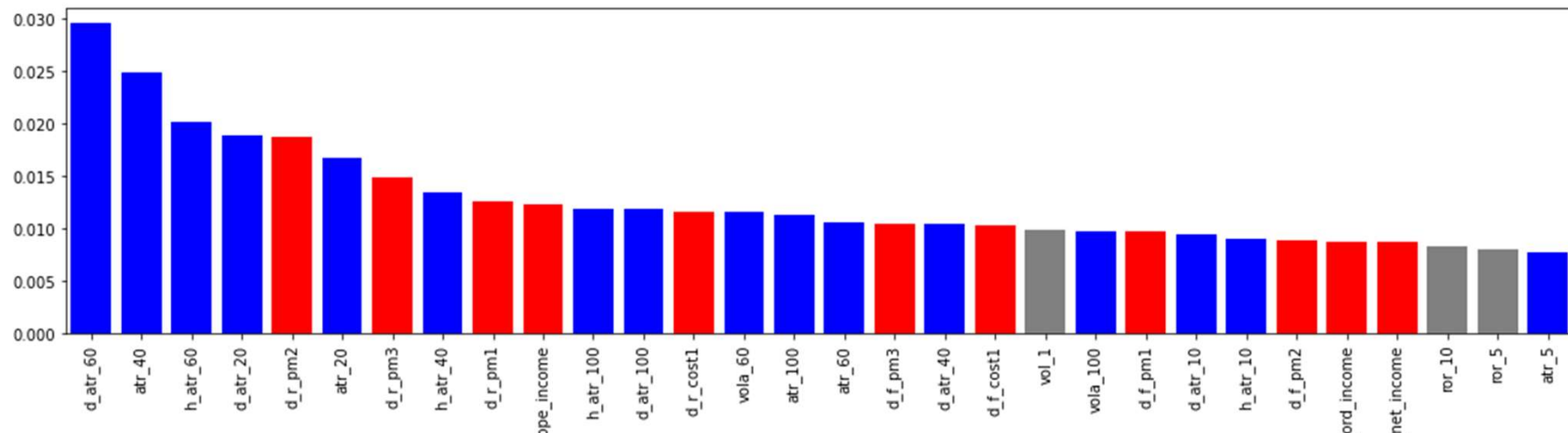
特徴量の重要度（Feature Importances）

■方向予測に特化（前頁の赤枠）
■ボラ予測に特化（前頁の青枠）

(1)高値側のFeature Importances



(2)安値側のFeature Importances



ATR（ボラ予測）とProfit Margin（方向予測）が上位を占める。
ボラ予測と方向予測の両輪が必要。

構築したモデル

<学習器>

XGBRegressor(max_depth=6, learning_rate=0.01, n_estimators=3000, colsample_bytree=0.1)

※高値側、安値側でそれぞれ学習

<特徴量>

182特徴量（価格系63、財務指標系119）

<モデル構築上の留意点>

予測対象が非常にロバスト（ボラの自己相関＋カタリストによる方向予測）
さらにPrivateは期間が限定されており、且つ相対的な予測である。
よって、

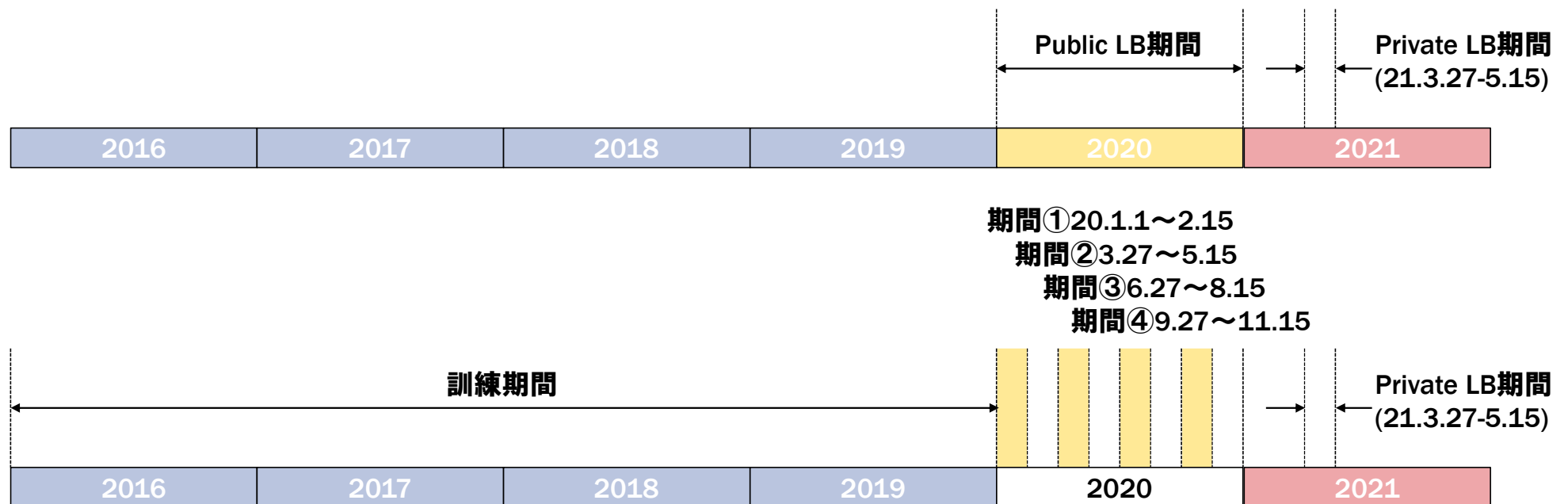
- ・ CV不要 : 与えられたサンプルを全て同時に活用する。
- ・ 前処理不要 : 相対予測なのでカット水準で間違いが発生しにくい。
※前処理する場合はリークに細心の注意が必要
- ・ 正則化不要 : 劣化しても十分なパフォーマンスが残る。

シンプルなモデリングでワークするが、評価期間の選定に落とし穴あり（次頁）

モデル評価上の留意点

Public期間での評価ではなく、独自の評価が必要。

- そもそもPrivate LB期間はPublicよりも短い。
- Public LB期間の2020年はコロナで安値側ターゲット分布に異常が発生



- 評価は1年間ではなくPrivateと同等の日数×4期間の平均で行った。
- 評価期間の選定がモデリング成否を分けた感触。

ファンダメンタルズ分析チャレンジ 知見まとめ

決算発表後の高値・安値の相対順位は、かなりの確度で予測可能。

- ①値動きの方向予測はProfit Marginで殆ど説明可能
 - ・各期において売上高経常利益率[%]を計算する
 - ・前期からの変化分が重要
- ②ボラティリティ予測はATRが適している
 - ・体系的には標準偏差が用いられるが、実務的にはATRがよい

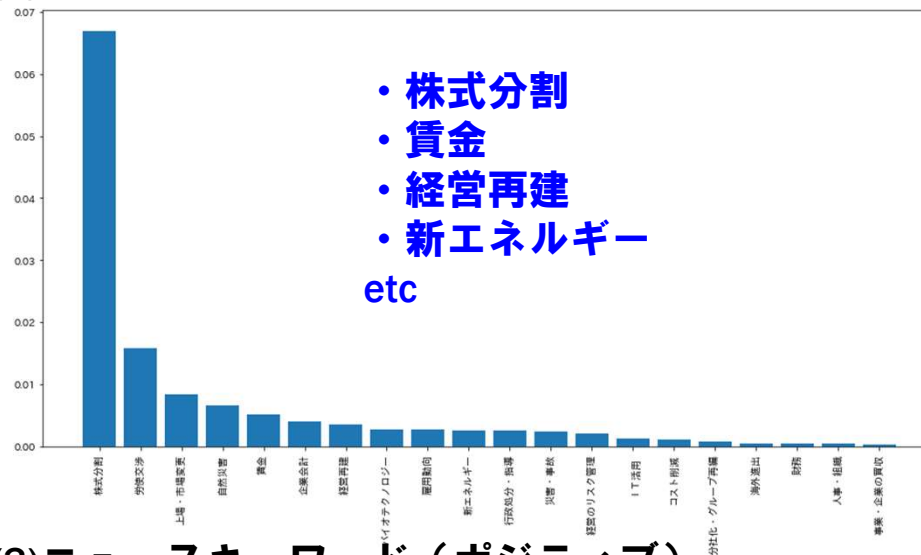
ニュース 分析チャレンジ

<条件>

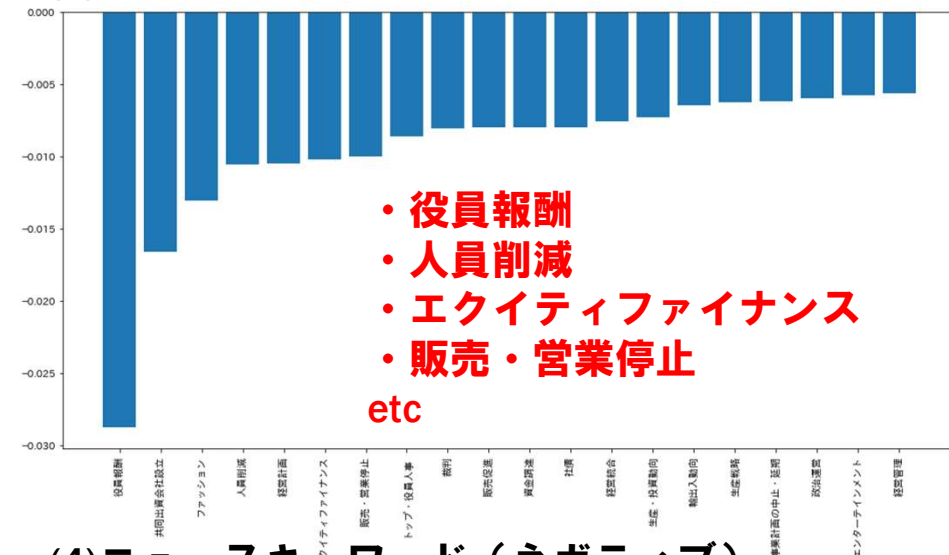
- ・ニュースデータは2020年のみ
- ・複数銘柄の含まれたニュースは除外
- ・サンプル数50以上のワードのみ抽出
- ・5日RT（マーケット寄与分を控除）

ニュースのテーマ／キーワードとリターンの関係性

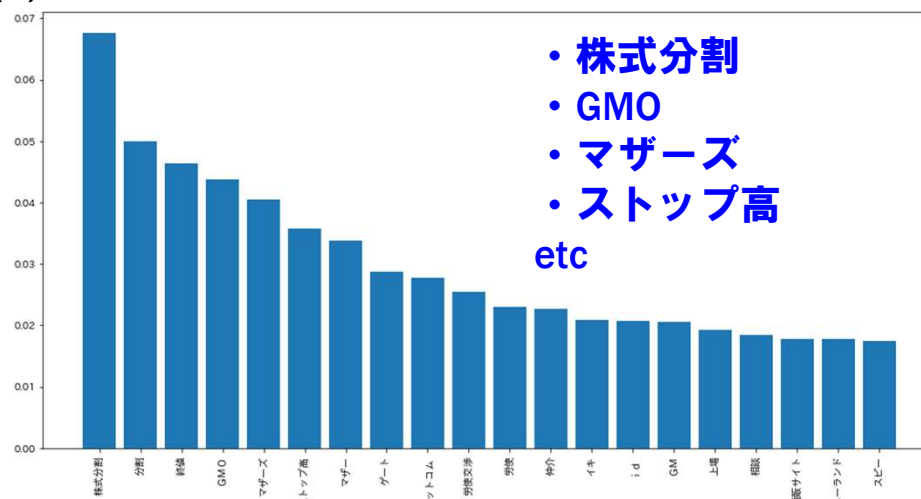
(1) ニューステーマ（ポジティブ）



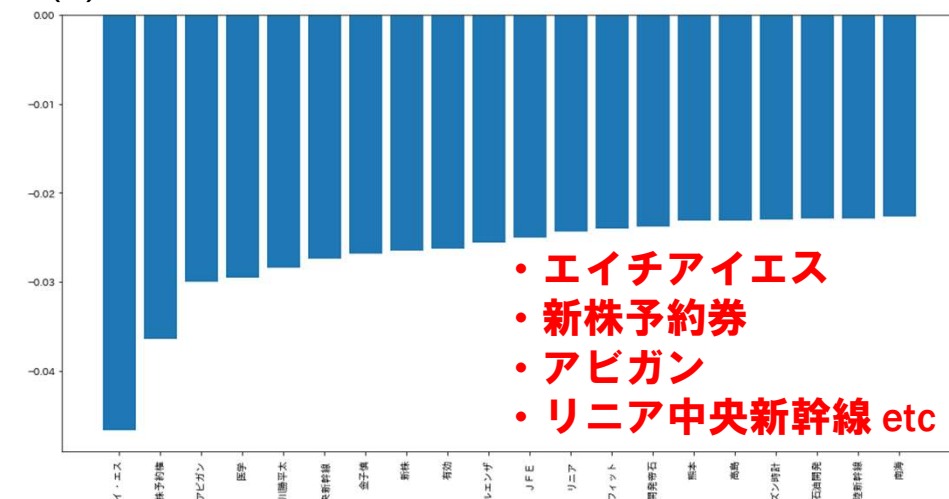
(2) ニューステーマ（ネガティブ）



(3) ニュースキーワード（ポジティブ）



(4) ニュースキーワード（ネガティブ）



テーマのほうが既知のイメージと合致しやすい（キーワードは局所的）。
サンプル数や季節性の問題があり、今回はモデルへの採用を見送った。

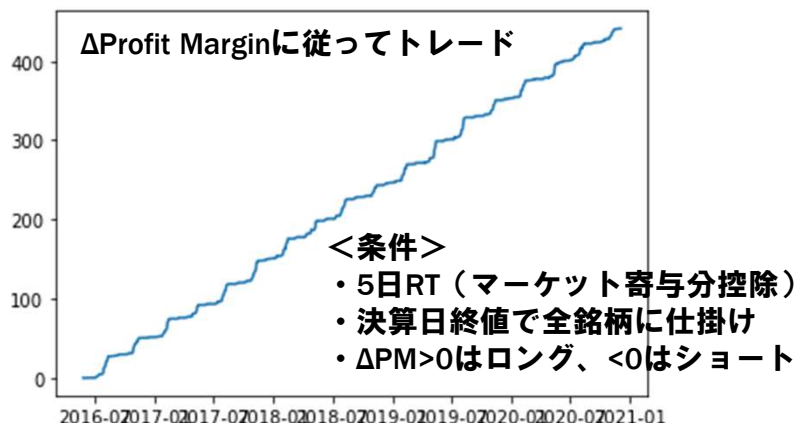
ポートフォリオ構築手法 選定の経緯

手 法	説 明	コンペ情報ソース		
		価格	財務	Text
テキストベース (ニュース)	先行研究で優れた事例あり 参考文献：Predicting Returns with Text Data (Ke, Kelly, Xiu, 2019) 今回はサンプルや季節性の事情で見送り	○		○
決算ベース (ファンダモデル流用)	モデル構築済みでコンペ主旨に沿う ボトルネックあり (次頁)	○	○	
クオンツ	機械学習の先行事例多数 参考文献：Deep Factor Model (Nakagawa, Uchida, Aoshima, 2018) 銘柄分散が必要で利幅が小さめ	○	○	
異常検知	ボラ拡大兆候の検出、方向予測は不向き 参考文献：VPINを用いた短期的な市場変動予測 (脇屋, 大屋, 2016)	○		
テクニカル	<ul style="list-style-type: none"> • 古典的手法 (モメンタム・リバーサル) • 機械学習 (時系列クラスタリング) 参考文献：Stock Price Prediction with Fluctuation Patterns Using IDTW and kNN (Nakagawa, Imamura, Yoshida, 2018) <ul style="list-style-type: none"> • 深層学習 (DNN for Candlesticks) 参考文献：Using DNN and Candlestick Chart Representation to Predict Stock Market (Kusuma, Ho, et al. 2019)	○		

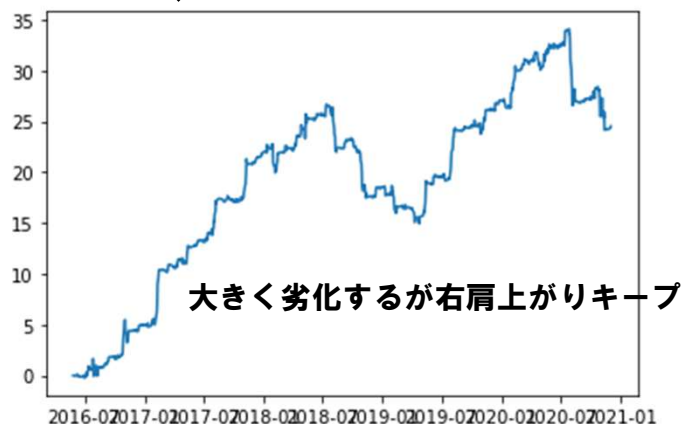
ニュースベースの代わりにファンダ分析チャレンジのモデルを流用した

ファンダ分析モデル流用時のボトルネックと対応結果

(1)翌日仕掛け時の収益性低下



- ↓
- 決算発表は15:00以降に集中
= 当日終値では建てられない
 - 翌日仕掛けにするとどうなるか？



一定のアルファは残存すると判断

(2) 決算銘柄数の偏り

2021年 5月

日	月	火	水	木	金	土
						01
02	03	04	05	06 50件	07 189件	08
09	10 100件	11 255件	12 329件	13 588件	14 1009件	15
16	17 64件	18 11件	19 8件	20 21件	21 2件	22
23	24 1件	25 2件	26 2件	27 1件	28 5件	29
30	31 7件					

決算スケジュール（国内株式） 出所：SBI証券

- ROUND1~2でスタートダッシュする戦略だと割り切った
- 予測日を起点に100銘柄以上が含まれる日までルックバック（情報の新鮮さ）
- PFは高値Pred上位5銘柄×20万で構築

結果として最も決算の多いROUND2利益が決め手となり上位入賞できた。

総括

コンペティションに参加した感想

<ファンダメンタルズ分析チャレンジ>

- ・ロバストな課題を選定し、チュートリアルを通じて初学者にも決算データ分析のノウハウを習得させる

<ニュース分析チャレンジ>

- ・ロングのみだがキャッシュ比率に柔軟的な制約を設けることで、相場環境まで意識した実践的PF構築を促す

- ・二段構えのコンペ構成が非常に秀逸だと感じました。
- ・またチュートリアルの内容の深さとボリュームには感銘を受けました。

→総じてとても有用なコンペであったと考えています。

Further Work

本研究の内容を用いて安定して利益を上げることができるのか？

答えは「このままでは難しい」。価格への織り込みが瞬間的で利幅が小さい。

<方針①>

好決算だが反応の甘い銘柄を狙うための
フレームワーク（株価がジャンプしないが
確実にトレンドシフトする銘柄）

事例：大和証券AIセレクト株式銘柄



<方針②>

来季の決算を予測する、そして仕込む。

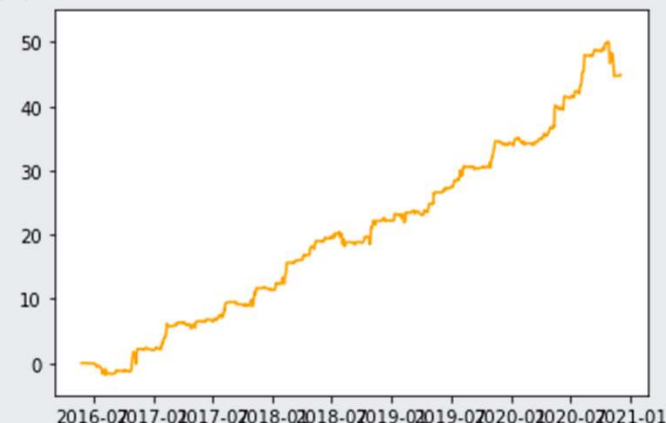
- 例えば利益率は負の自己相関を持つ
- 収益構造（人件費、広告費）の変化が
来期決算に及ぼす影響分析

<検証結果>

(1)改善前



(2)改善後



更なるデータ分析によりパフォーマンス向上できる可能性

個人投資家によるデータ分析の意義

- ・「データを使って豊かになる」。これは時代の潮流として当然。
- ・株式投資の収益源の探索は深化を続け、情報戦は激化の一途を辿る
 - ※一例として、データドリブンな投資としてロボアドの存在感が増すがこれらのロボアドのトレンドも個別銘柄へとシフトしつつある（ダイワ、カブコム、ウェルスウィング、アルパカロボ）
- ・資産形成は国民への社会的要請になりつつあり、「データ×資産形成」というジャンルを切り開く機運は訪れている
- ・個人投資家はデータドリブンな投資の理解を深め活用するとともに自己運用できればなお嬉しい

今の時代の投資にはデータ分析は不可欠、そしてそれを実現するためにAPIが必要

ご清聴ありがとうございました

構築モデル

