盛り上がる株式予測コンペの問題設計と勝利者解法考察

AlpacaJapan CPO Tomoya Kitayama

Alpaca

会社概要と発表者紹介

会社概要

会社名	AIpacaJapan株式会社
拠点	東京（本社）、神戸
創業	2013年2月（当時Ikkyo Technologies（株）として設立）
従業員数	57名（うちエンジニアが30名）
事業概要	AIによる市場予測ソリューションを提供するフィンテック企業。 様々なAI技術に加えて、ビッグデータ解析や金融向けデータ・ストレージなどの技術にも強みを持つ

沿革

2013
- 前身親会社AIpacaDB社（米国）が画像認識サービスで創業

2015

2016
- AIpacaDB社がフィンテック領域に事業転換
- 日本では、トレーディング業務をAIによって自動化するアルゴリズムをローンチ
- MUFG主催のフィンテック・アクセラレータ・プログラムにて準グランプリ獲得
- 四元の入社と共に、AIpacaJapan（株）に名称変更

2017
- AIpacaDB社の米国進出(B2C証券)に際し、AIpacaJapanとして国内外B2B AIビジネスに本格参入
- MUFG向けに短期のマーケット予測モデルを本格実装開始

2018
- 日米間の経営とガバナンスの明確化を目的に日本経営陣にてMBOすることで日米分離
- その他様々な予測モデルを提供開始

2020
- MBOから約1年後に、三菱UFJ銀行をリードとするSeries Aラウンドにて6.5億円を増資
- SBIグループ、株式会社ミンカブ・ジ・インフォノイドへのSeries B調達を完了

北山朝也

Chief Product Officer

- 10年間ソニーに在籍し、PlayStationのサポートチームのマネージャーとしてゲームタイトル開発者とPlayStation開発チームの橋渡しを行う、ソニー時代に2008年度IPA未踏スーパークリエイターなども取得
- 2015年よりAIpacaに参画し、金融機関との様々なプロジェクトを企画・実施
- 金融 X AIでビジネスをつくることに日々もがく
- 現在はChief Product OfficerとしてAIpacaのプロダクトを統括

Alpacaの歴史

2013年 - 2014年: 画像認識の受託開発企業としてスタート(任天堂などに技術提供)

2015年 - 2016年: 画像認識モデル作成サービス(Labellio)を京セラに売却、AlpacaAlgoを開発。

2017年: MUFGとコラボを開始。Fintechの道へ入る

2018年 - 2019年: AlpacaJapanがAlpacaUSからMBO。AlpacaUSは米国でAPI証券の道を歩み、AlpacaJapanは日本でAI&Fintechカンパニーとしての成長を目指しBloombergと提携

2020年: PoCメインからASP/プロダクトドリブンを目指し、研究開発を続ける

2021年: J-Quantsの第1, 2回目、アルパカ証券の開業

2022年: J-Quantsの第3回目をKaggleで開催

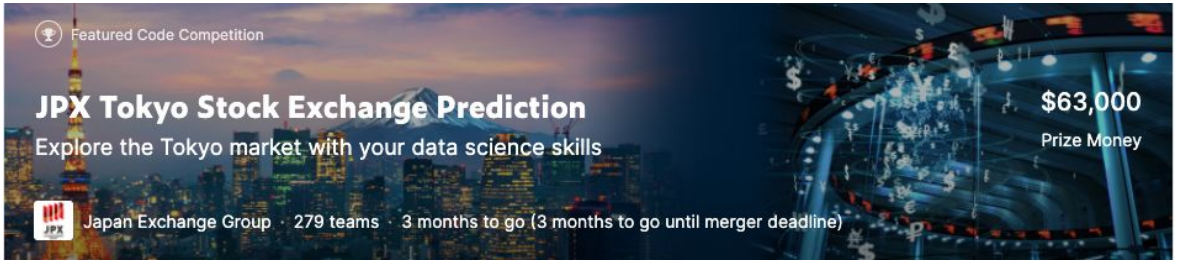
J-Quantsで開催したコンペ

J-QuantsとはJ-Quantsは、投資にまつわるデータ・環境を提供し、個人投資家の皆様によるデータ利活用の可能性を検証するプロジェクトです。個人投資家の皆様が、データを活用して取引できるようになることを目的として、学習コンテンツの提供など各種施策を行います。

これまで開催したコンペ

ファンダメンタル分析
2021/1/29-2021/6/14

ニュース分析
2021/3/19-2021/6/4



Kaggleによる開催
2022/3/19-2022/6/4

株式予測のコンペを開催する難しさ

データサイエンスにおいて、正当な努力がスコアに反映されない問題は非常に嫌われる

- データにノイズが多すぎてパターンを見出すことが出来ない
- 学習用のデータと評価用のデータがあまりにも違いすぎるために学習させたモデルが全く役に立たない
- 突然レジームチェンジが起きてゲームのルールが全く変わってしまう

つまり、株式予測の問題はデータサイエンティストが嫌がる要素をすべて含んでいる

過去にTwo Sigma/Jane StreetなどもKaggleでコンペを主催してきたが、あまり良くない評価

「クソコンペオブザイヤー2021」より抜粋

https://aru47.hatenablog.com/entry/2021/12/12/205905

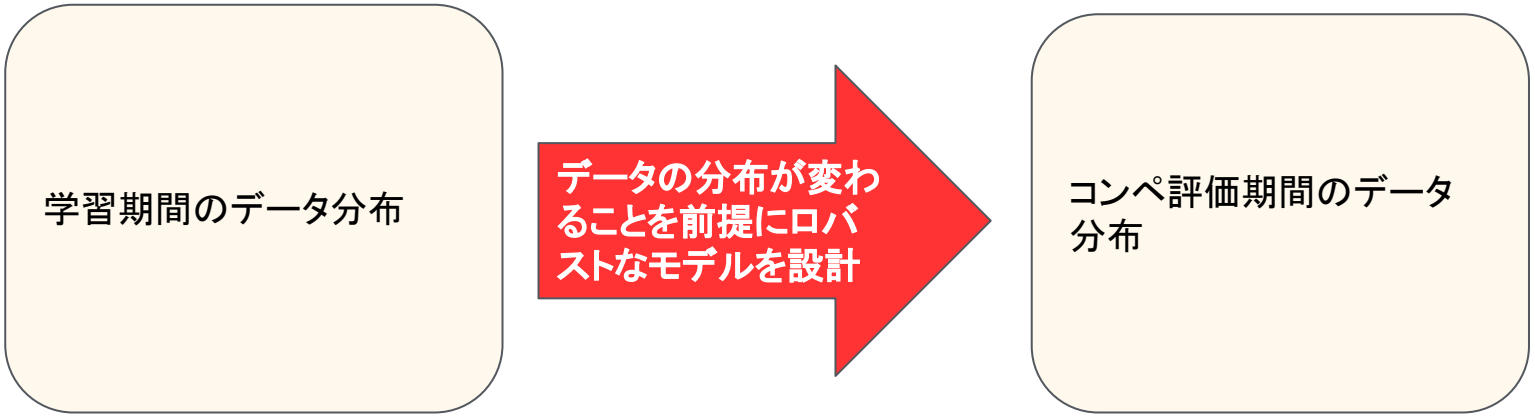
タスク設計が難しすぎる・現在の技術では解くのが困難なため結果運ゲーとなってしまうコンペを指します。例えば6ヶ月後の株式を予測するコンペなどはタスク設計が難しすぎ、運ゲーとなってしまい参加者としてあまり面白みはありません(e.g. Jane Street Market Prediction、Two Sigma: Using News to Predict Stock Movements)。

データサイエンスコミュニティに受け入れられる株式予測コンペについて考える必要がある

問題設計のときに考えなければいけないこと

一般的に金融マーケットの時系列データの2つの異なる期間において、データが同一の分布を永続的に持つことを仮定することはできません。例えば、2020年2月までと2020年3月以降では、新型コロナウイルス感染症等による世界情勢の変化により、マーケットの性質は大きく変化しました。

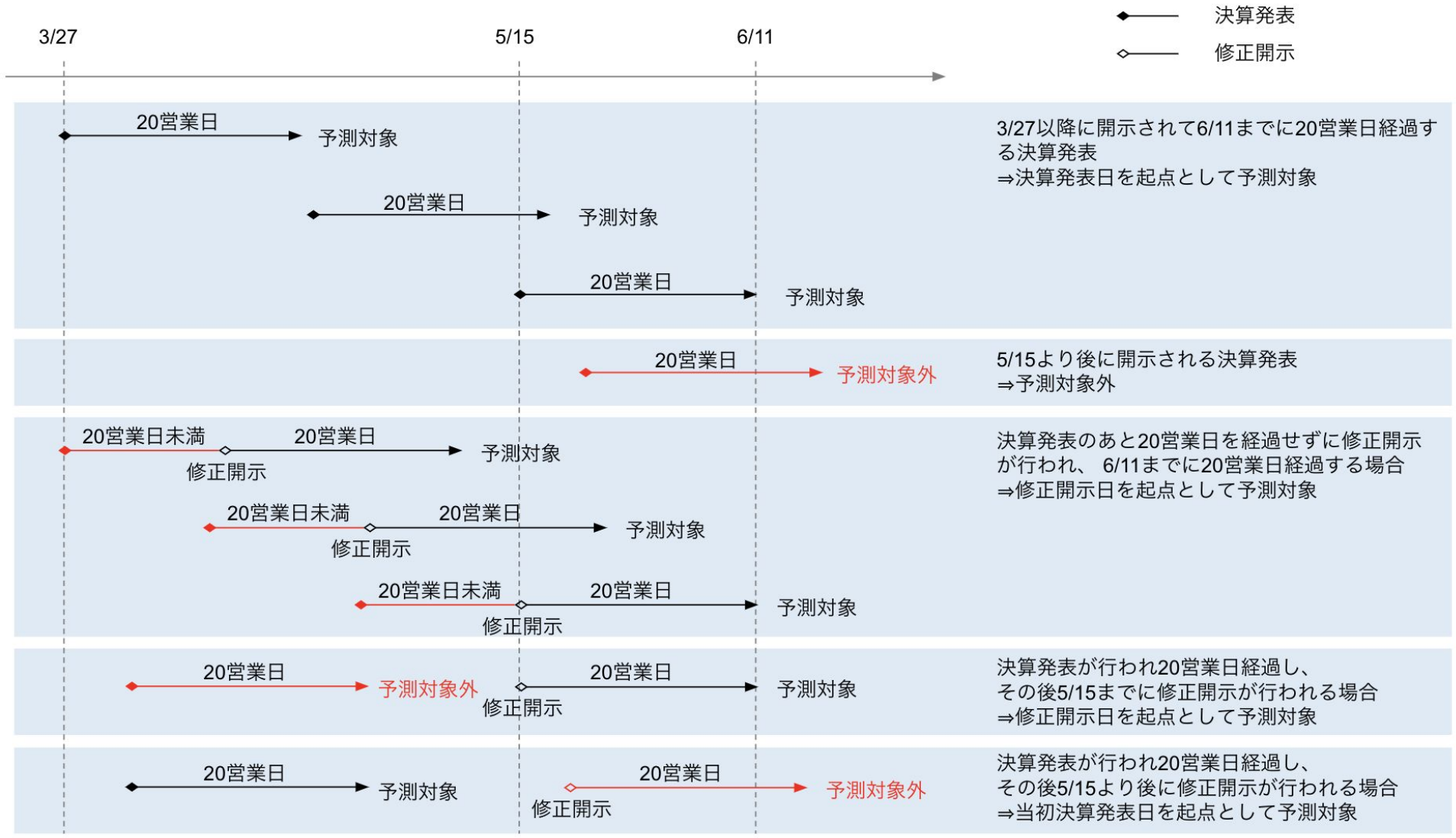
このようにデータ分布の特性が移り変わる金融市場を題材としたコンペの場合、問題設定として、データ分布の変化に依らないロバストなモデルを構築した参加者をコンペの勝利者とするべきと考えました。



この評価を成立させるためには、ロバストな問題を設計する必要がある！

ファンダメンタル分析チャレンジの問題

課題: 決算発表後の株価の20日高値・安値予測



ファンダメンタル分析チャレンジの問題

最高値・最安値への変化率について、それぞれのスピアマンの順位相関係数を合算し評価

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

d：対応するXとYの値の順位の差
n：値のペアの数
X：該当期間の決算日に対して出力されたモデルのスコア
Y：20日営業日以内に発生した最高値への変化率、もしくは最安値への変化率

その上で、それぞれの順位相関係数を以下の通り合算した統合スコアを評価関数とします。

$$score = (\rho_{high} - 1)^2 + (\rho_{low} - 1)^2$$

ρ_{high}：最高値のスピアマン順位相関係数
ρ_{low}：最安値のスピアマン順位相関係数

つまりある一定期間に発表された決算において、発表後に最高値になる可能性の高い決算、最安値になる可能性の高い決算を予測、その予測の **順位**が実際にどの程度似ているかを競います

ファンダメンタル分析チャレンジができるまで

本コンペを設計する上で、JPX様と一緒に過去のすべてのコンペの問題を分析

- 合計12個のコンペが過去にあったがデータサイエンティストの努力が反映されない設計
- データサイエンティストの努力がそのまま反映されるコンペを第一弾に作りたい

実際に20個程度の問題を作成し、有望ないくつかの問題を実際にAlpacaのデータサイエンティストがデータを用意して解いてみて以下の5観点で評価

- 元データの取得しやすさ
- 話題性（面白さ）
- 実取引での有用性
- 金融初学者の参入しやすさ
- 分析しやすさ

データサイエンティストの工夫に比例して、パフォーマンスが向上し、決算分析という王道の問題で、アナリストがよく発表する目標株価と近い概念の「20日以内の高値・安値予測」という問題が完成

実際に問題を解いてみている時のサンプル

特徴量やラベルに対してRMSE、Accuracy、スピアマンの順位相関、ピアソンの相関などを解いてみて、
最もデータサイエンス的な努力に対してスコアが上がりやすく、運要素が極力少ないものを選択

高値予測

month	feature	index	RMSE	accuracy	spearman_corr	corr	R^2 score
1	fundamental+technical	label_high_10	0.092	0.682	0.038	0.100	0.012
	fundamental+technical	label_high_20	0.104	0.698	0.034	0.123	0.016
	fundamental+technical	label_high_5	0.085	0.619	0.034	0.097	0.010
	fundamental_only	label_high_10	0.093	0.678	0.067	0.128	0.017
	fundamental_only	label_high_20	0.105	0.699	0.056	0.136	0.019
	fundamental_only	label_high_5	0.086	0.619	0.069	0.102	0.011
	return_only	label_high_10	0.095	0.670	0.002	0.017	0.001
	return_only	label_high_20	0.110	0.695	0.011	0.023	0.001
	return_only	label_high_5	0.087	0.601	-0.014	0.005	0.001
	technical_only	label_high_10	0.092	0.685	0.048	0.090	0.008
	technical_only	label_high_20	0.106	0.698	0.053	0.106	0.011
	technical_only	label_high_5	0.085	0.619	-0.001	0.036	0.002
2	fundamental+technical	label_high_10	0.096	0.501	0.023	0.144	0.021
	fundamental+technical	label_high_20	0.108	0.504	0.008	0.143	0.020
	fundamental+technical	label_high_5	0.082	0.495	0.048	0.106	0.011
	fundamental_only	label_high_10	0.103	0.510	0.034	0.052	0.003
	fundamental_only	label_high_20	0.120	0.506	0.003	0.043	0.002
	fundamental_only	label_high_5	0.087	0.509	0.054	0.048	0.002
	return_only	label_high_10	0.099	0.503	0.022	0.075	0.007
	return_only	label_high_20	0.114	0.505	0.018	0.108	0.013
	return_only	label_high_5	0.084	0.499	0.014	0.045	0.003
	technical_only	label_high_10	0.097	0.501	0.012	0.132	0.018
	technical_only	label_high_20	0.110	0.504	0.011	0.159	0.025
	technical_only	label_high_5	0.082	0.488	-0.010	0.071	0.006

いろいろな問題に特徴量とモデルを複数用意し、メトリクスのバラツキを確認する

ニュース分析チャレンジの問題

問題

本コンペティションでは、モデルで予測したポートフォリオで得られる利益の総合計をもとに評価します。従って、提出モデルでは、利益の総合計がより高くなるポートフォリオを予測いただきます。利益の総合計は、以下の算出式を用いて計算します。

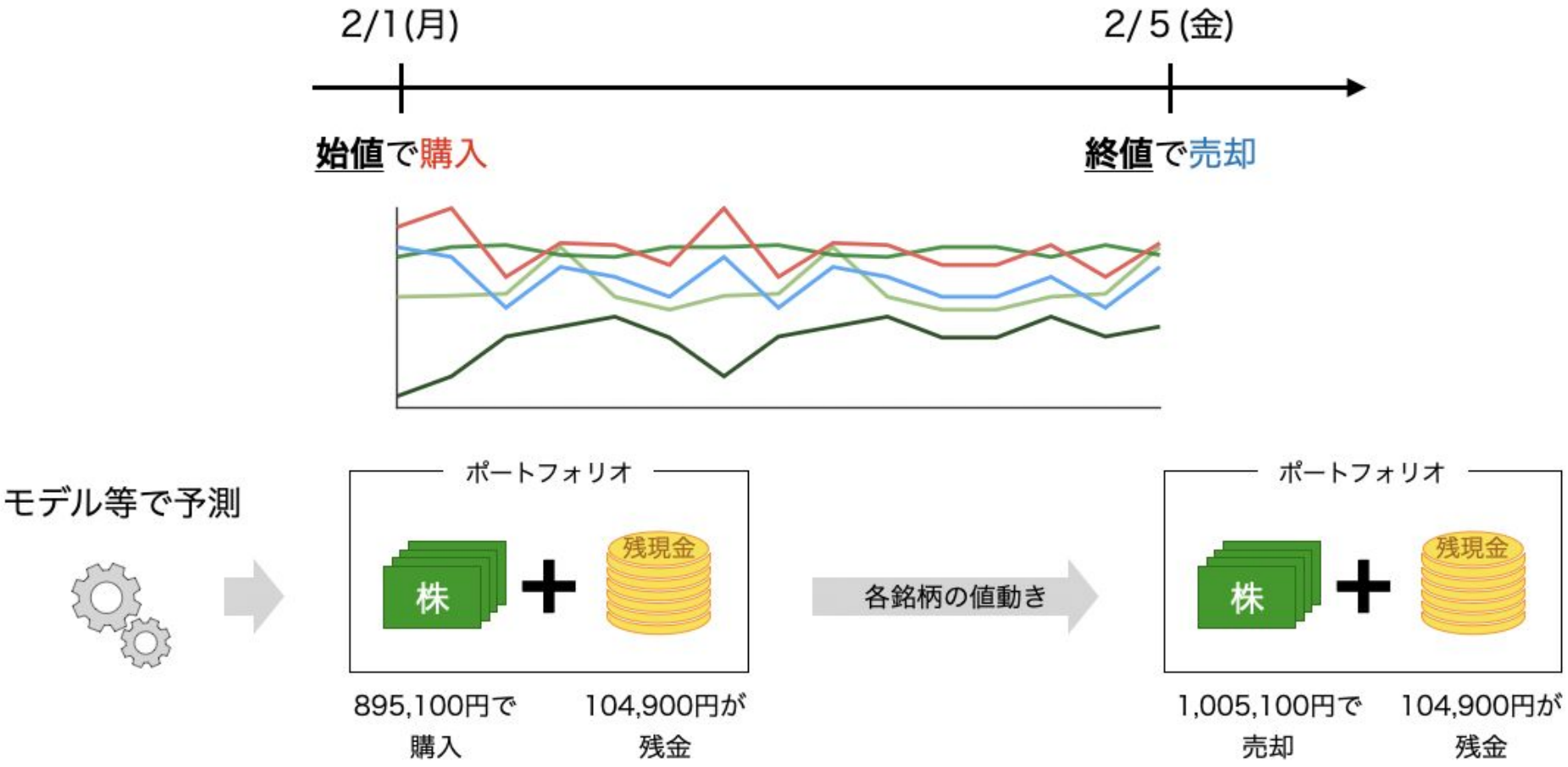
利益の総合計 = Σ [r=1, n] ラウンドrの収益 (Public LB: n=1, Private LB: n=4)
ラウンドrの収益 = ラウンドrの保持現金 + ラウンドrの合計評価額 - 100万円
ラウンドrの保持現金 = ラウンドrの最初営業日において株式の購入に利用せずに残った現金
ラウンドrの合計評価額 = ラウンドrの最終営業日の終値 x 保持する株数

ポイント

実際のトレーディングを想定したコンペ(といっても利食い・損切りなどがないなど制限はあるが)

- 自分の投資戦略に近いものをそのまま利用可能
- ファンダメンタルズ分析チャレンジの成果を活かすことが可能
- ニュースデータのようなオルタナティブデータを活用するアドバンスな手法を試すことが可能

問題の図解



Public期間での運用実績* = **保有している株式の合計評価額** + **保持現金** - **購入原資金**

110,000円 = 1,005,100円 + 104,900円 - 1,000,000円

* 運用実績は株式から得られる配当金等は含まれず、純粋な株価の値動きのみにより決まります

ニュース分析チャレンジができるまで

ユーザーにできるだけ実取引を想定した問題を提供したい

- 実取引をできるだけ想定したコンペ設計
- 運が絡むのはどうしても回避できないが、それでもある程度努力に応じた結果になってほしい

問題設計のポイント

- 売買の試行回数をある程度増やすことでデータサイエンス的な効果が出やすくする
  - さすがに1日単位では短すぎるので週単位とする
  - 1株単位で購入可能とすることで複数種類の株を買いやすくする
- 購入代金を50万円以上と調整可能とすることで下落局面のコントロールもできる
  - フルでお金を投資必須とするよりも幅広いベータコントロールが可能になる
- 運要素があまりにも高い銘柄(株価が数十円で1円の値動きで数%の利益となるなど)は排除
  - 2020年12月末時点で、時価総額が200億を上回っていること

ニュースチャレンジ問題を週単位のポートフォリオ問題として設計

第３回のKaggleの問題設計

1. モデルは毎営業日(t)に、その営業日までの終値(C_t)等のデータを入力データとして、翌営業日の終値(C_t+1)から翌々営業日の終値(C_t+2)までの変化率(r_t+1)の上位200銘柄と下位200銘柄を予測します。
   r_t+1 = (C_t+2 - C_t+1) / (C_t+1)

2. 予測した上位200銘柄に対して1から200位までにそれぞれ2から1の線形のウェイトをそれぞれの変化率に対して掛けたものの総和を計算しこれをS_upとします。

3. 予測した下位200銘柄に対して下位1から200位までにそれぞれ2から1の線形のウェイトをそれぞれの変化率に対して掛けたものの総和を計算しこれをS_downとします。

4. S_upからS_downを引いた結果をR_dayとし「デイリースプレッドリターン」と呼ぶことにします。

5. デイリースプレッドリターンをpublic/private期間中、毎営業日計算し、当該期間の時系列として取得します。デイリースプレッドリターンの時系列の平均/標準偏差をスコアとします。
   a. スコア計算式(xはpublic/private期間の営業日):
      Average(R_day1-dayx) / STD(R_day1-dayx)

6. private期間の最大のスコアを獲得したカグラーが勝利となります。

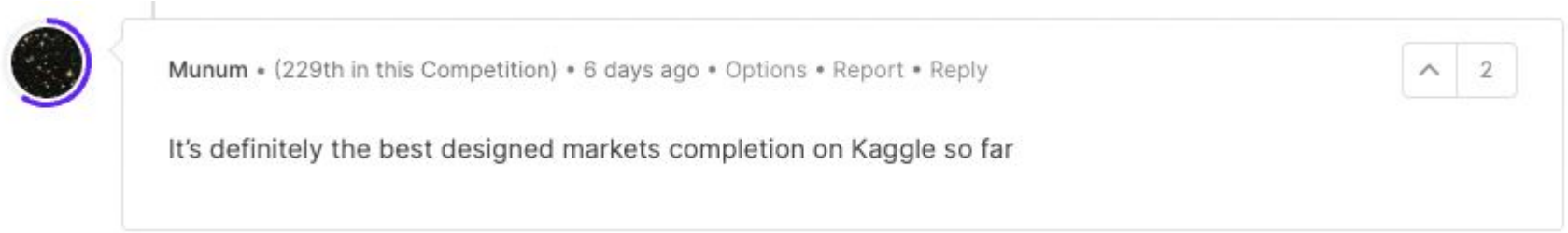
Kaggleの問題設計のポイント

- 個別銘柄の事象にあまり左右されることなくモデルの性能を競い合うことができるように、毎営業日に予測する銘柄数を投資対象銘柄(2000銘柄)に対して**10%以上である上下それぞれ200銘柄の変化率の差分**としています。ただし実務的には機関投資家やファンドの投資対象は50-100銘柄とすることが多く、若干現実の問題設定と乖離がありますが、本コンペティションでは問題の安定性を優先して設計しています。
- デイリースプレッドリターンの計算に際しては、1位により収益率が高い(下位であれば低い)銘柄が配置されるように、**1位から200位に対して1から2の線形のウェイト**を掛けています。
- 「リスクコントロール」も投資にとって重要な要素であるため、競い合うスコアをデイリースプレッドリターンの単純平均や総和ではなく、**デイリースプレッドリターンの時系列の平均/標準偏差**としています。これにより特定の日だけ大勝ちするモデルではなく、データの分布の変化に対応し安定的に高い性能が出るモデルを構築することが必要になります。
- 本コンペでは、**マーケット自体のボラティリティやリスク要素を推測する手がかりとなるオプションデータ等のデータも提供**しています。これらのデータを活用することでより高度なリスクコントロールができるかもしれません。また、**下位200銘柄も予測対象に含めているため、マーケット・ニュートラルな戦略を採用することも可能**(なお、ベータ値をわざと偏らせることでロング側にバイアスを掛ける戦略も可能)で、**ベータ値をコントロールすることでマーケットの値動きに大きく左右されない標準偏差がコントロールされたモデルの構築が可能**です。

Kaggleでの問題設計に対する反応

元ミレニアムの株式運用責任者、元QuantopoanのCIO、現在Numeraiにも出資

Kagglerからも好感触 https://www.kaggle.com/competitions/jpx-tokyo-stock-exchange-prediction/discussion/324143#1784860



上位入賞者の解法はソースコード含めて公開済み

https://www.jpx.co.jp/corporate/news/news-releases/0010/20210813-01.html

こちらのファンダメンタルズ分析チャレンジの解析結果について情報共有致します。

資料概要	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
1位解法	ご発表資料 提出モデルソースコード	提出モデルソースコード
2位解法	ご発表資料 提出モデルソースコード	同左
3位解法	ご発表資料 提出モデルソースコード	提出モデルソースコード
4位解法	提出モデルソースコード	提出モデルソースコード
5位解法	提出モデルソースコード	提出モデルソースコード
6位解法	提出モデルソースコード	提出モデルソースコード
7位解法	提出モデルソースコード	提出モデルソースコード
8位解法	提出モデルソースコード	提出モデルソースコード
9位解法	提出モデルソースコード	提出モデルソースコード
10位解法	提出モデルソースコード	提出モデルソースコード
Web記事賞	発表資料リンク Web記事	—

上位入賞者の解法について モデル作成

利用モデルは以下の通り。モデルが汎化したことの評価方法に工夫が観測されました。

- LGBM 9(リグレッション/ランク学習)
- XGBR 1
- ニューラルネットワークはなし

上位の人ほどCVや評価期間を工夫しており、本問題の本質がよく現れています。

2017-2020までの各年がfoldに対応する4fold-cv。先述の通り、信用できるCV手法が確立できたことが大きく、最終的なprivate LBのスコアはCVのスコアの誤差範囲内に収まっていた。
PublicLBの評価期間は2020年の通年でしたが、Privateの評価期間は3/27～5/15までのおよそ1.5ヶ月であるため、モデリングの評価もこれに準じた期間で行いました。具体的には①2019/12/27～2020/2/15、②2020/3/27～2020/5/15、③2020/6/27～2020/8/15、④2020/9/27～2020/11/15の4つの期間でそれぞれ評価スコアを計算し、それぞれのアベレージを見て汎化性能が改善したかどうかを判断しました。
5FOLDクロスバリデーション。学習時は時系列データとしての考慮は特にしておらず、厳密に言えばリークが発生している可能性はあるが、面倒だったので何も対策は行わなかった。
Time series cross validationを意識して2019年、2020年、2021年をテスト期間とした3つのfoldを作り汎化性能を計測した
汎化性能についても予測ラベルをクロスセクションで正規化する手法が効いたと考えています。実際パブリックにおける精度とプライベートにおける精度はほぼ一致していました。

上位入賞者の解法について 特徴量

ドメイン知識をフル活用(どれが支配的かを改めて選別する)するアプローチ、一般的なトレード指標を使う、差分系列・セクター平均化などのテクニックが観測されました。ドメイン知識では本問題において、レンジの情報が重要であることの指摘が多く、ボラティリティ予測の本質をきちんと知っていると感じました。

ATR（Average True Range）、利益率に関する指標（売上高純利益率など）
PERやPBRなど一般的なトレードで重視される指標
前期からの差分や、前期の予測と今期の値との差分などの時系列方向の特徴量
セクター平均や日経平均の株価、制限値幅による株価の上限・下限
テクニカル: 20,40,60営業日のそれぞれについて、リターン、ボラティリティ、移動平均との乖離率、(最大-最小)/標準偏差、騰落率の平均、騰落率の標準偏差、騰落幅の標準偏差/平均、RSI、MACD
ファンダメンタル: ROE, ROA, EPS, PER, BPS, PBR, 配当性向、四半期ごとの修正回数、前四半期との差分（売上高、営業利益、経常利益、当期純利益、総資産、純資産、一株当たり四半期配当金、一株当たり年間配当金累計）
テクニカル分析の指標を算出できるtaライブラリを用いて，一度全特徴量を用いて予測．その後特徴量間の相関や重要度を観察し，精度に対して悪く働いている特徴量を削除
累計値ではなく各期の伸び率を算出、RSI・MACDなどの株価指標、時価総額・EPS・PER・PBR

上位入賞者のポイント

- **評価関数・手法の設計が最も重要**
  - データサイエンス的な手法でスコアを向上させる方法はたくさん書籍がある
  - 問題はそのスコアが未来に対してもロバストにワークするか
  - 未来に対してのロバストさを検証する方法はたくさんあるが、X-FoldやCPCVが強力な手法として知られており、実際に多くの上位モデルでこれらの派生のテクニックを使っていた
    - CPCV(Combinatorial Purged Cross-Validation)法は時系列データ(特にCVが不安定な場合)に対して安定したvalidationを行うための手法
      - https://zenn.dev/ymd/articles/fd08fb46bc868c などを参照

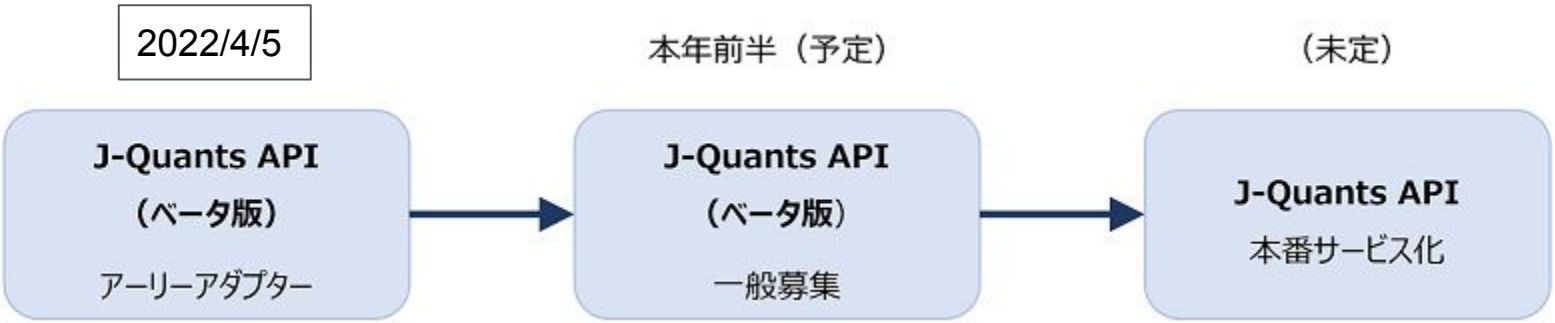
- **ドメイン知識が次に重要**
  - 特徴量設計をしっかり行っていれば、モデル自体はほとんどの場合はあまり大きな差異を生み出さない
  - 特徴量として、差分系列・セクター平均化などのテクニックや、ボラティリティ推定の過去研究はオプションのプライシングでたくさん行われているので、それらの情報が活かされているモデルが強かった
  - ドメイン知識で特徴量を大量に設計して、それらを決定木のモデルで学習される手法がほぼ上位を独占した

J-Quants APIのデモ

本サービスで以下のデータをAPIにより配信する予定です。利用者は自由な切り口で、株価の時系列分析や、上場会社間の比較が可能となります。

- 東証上場会社にかかるヒストリカル株価（2017年以降の四本値・出来高等、株式分割等を調整した調整株価）
- 決算短信情報を整形した財務データ（2017年以降の決算短信サマリー情報）

J-Quants APIは今後β版の一般募集を実施予定です。ぜひ、ご応募ください！



Thank You!