Rist Meetup 2024

Kaggle駆動ソリューション / プロダクト開発

2024年10月12日

株式会社Rist
石 圭一郎 / @ishikei

Rist

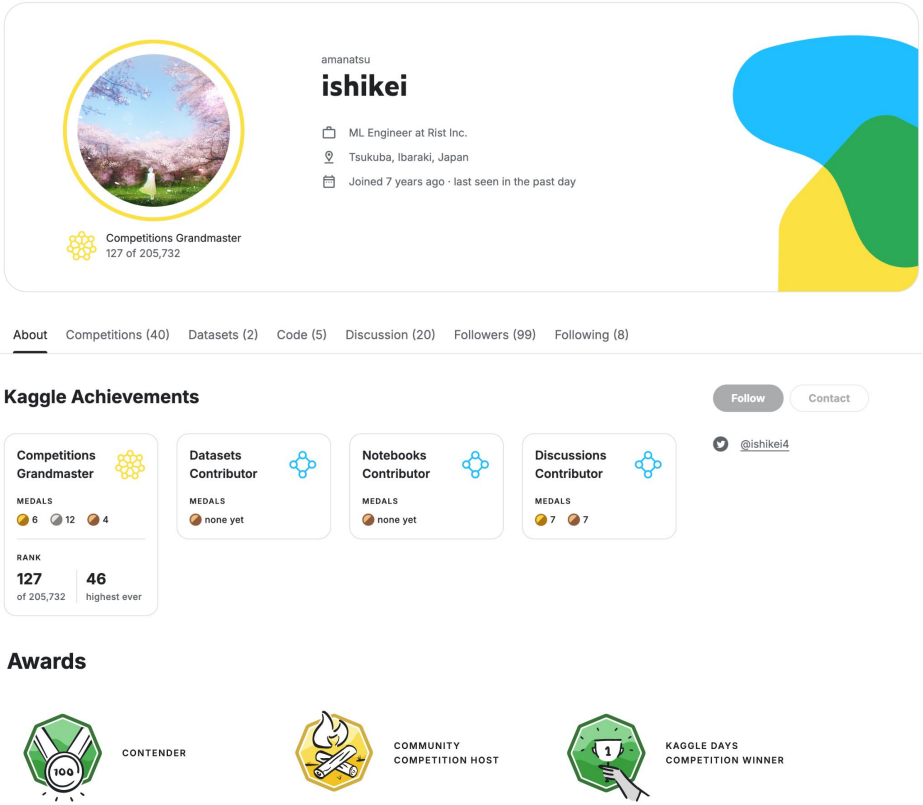
目次

1. 自己紹介
2. Rist Kaggleチーム
3. Kagglerの問題解決能力を業務でもフルに活かし切りたい！（抽象 / しくみの話）
4. おまけ：「Kaggleテクは業務の役にたつ」（具体 / テクニックの話）

石 圭一郎 / ishikei

Competition Grandmaster
株式会社Rist：ビジョンソリューション部

- 主に画像認識まわりのお仕事してます
- PAPA Kaggler：娘(4歳)
  - 課題：幼稚園でのパパ友作り
- 趣味：映画鑑賞 (最近はホラー系が好き)
  - 最近みた中でおすすめ：カルト (白石晃士)
- 最近の推しコンペ
  - Benetech(2023) / RSNAシリーズ
    - 総合力 or 発想力勝負のコンペが好き



沿革

- 2019/12：Rist Kaggleチーム発足 (onoderaさんアドバイザー就任)
- **2024/10現在、GM9名 / Master3名 が在籍 (Rist従業員数67名)**

社員のうち
5.6人に1人がMaster以上

(Kaggleチームとしての) 主な活動内容

- **Kaggle workshop (隔週)**
  - コンペのまとめ、論文まとめ、案件紹介など…
  - たまにオフラインでも開催
- **業務時間内でのコンペ参加**
  - 業務内のKaggle比率：30% ~ 50%
  - 計算リソース補助
    - 個人KagglePC (ishikeiの場合: RTX4090x2)
    - Kaggleチーム優先A100クラスタ

2019/12 Kaggleチーム1人目

2024年度 キックオフMTG

Kagglerがかかわる主な業務 (第2回 関東Kaggler会 スポンサーLT資料より)

受託・Product開発それぞれで<u>Kagglerが活躍しています</u>

今日はここの「仕組み」の話

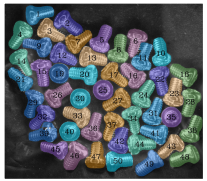
京セラロボティクスサービス
関わったKaggler: tascj, ishikei, takuoko, smly, fam_taro

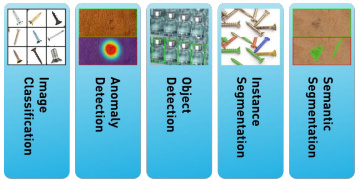
受託データ分析・モデル開発事業

- かなり直接的にKaggle的知見が活かされる領域
- 課題設計やモデル開発の精度、およびそのサイクルを回す速度
- 「どのKagglerがどんなコンペに出てる (た) か」を常に共有
  → 可能な限り人材を最適配置して案件にいどむ体制を構築

Deep Counter

Deep Inspection PRODUCT





Product系事業

- 主にProductの内部アルゴリズムはKagglerが関わることが多い
- アプリケーションとの繋ぎこみの部分は親会社のリソースも活用

Kagglerの問題解決能力を業務でもフルに活かすためには？

Kagglerと一括りにしても中には色々なタイプがいる

- 得意なデータ形式：画像 / テーブル / NLP / 信号データ...
- 得意なタスク：例えば画像ならClassification / Detection / Segmentation / Instance Segmentation...
- 得意な取り組み方：
  - ドメイン知識ガッツリ取り入れながら特化モデル作る
  - モデルの技術的な課題を特定してそれを解決する方法を研究する...

▶ **Kagglerと案件とのミスマッチをなくしたい！**

Kaggleを案件で活かすために社内で実施していること

- Kaggle実績シートの作成 / 共有
- 社内でKaggleの話題に (カジュアルに) 触れる機会を多く作る

Kaggle実績シートの作成 / 共有

- もともと社内Kaggleランク管理用に各Kagglerのコンペ実績をまとめたシートは存在していた
- ただ、それらは単純にランク管理のためであり、コンペ詳細などはまとまってなかった

▼

上記を参考に2種類のシートを作成

社内共有用Kaggle実績まとめ

- コンペの内容、ドメインについてスプレッドシートにまとめる
- その他、各メンバーが得意なタスクや、ドメイン知識のある領域についても記載
- マーケやPMチームと共有し、案件アサイン時の参考とする

お客様への紹介用Kaggle実績まとめ

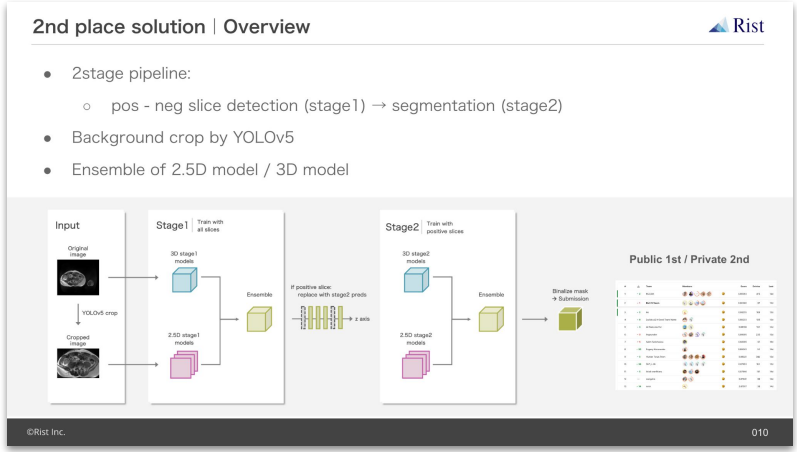
- 1コンペ/ページ程度でスライドにまとめる
- 課題の概要、ドメイン、どのような実績を残したかを記載
- 商談の際の話のタネになったり、提案に説得力がでる

社内でKaggleの話題に (カジュアルに) 触れる機会を多く作る

- 直近であったKaggleコンペについては社内slack上や 朝会などで雑談ベースで話したり
- Kaggle workshop (Kaggleチームの隔週MTG) は、Kaggler以外も参加可能にしている
- ただし、そこで用いる資料などはKaggler向けを前提として作る（発表のハードルを下げる & 負荷削減）
    - 例えば画像コンペの話なら、逐一「ConvNeXtとは」「Pseudo-labelとは」みたいな説明は設けない

日付	担当１	お題１	URL1	担当２	お題２	URL2
2024年						
1/19	Liu	UBC-OCEAN		小野寺	RNAコンペ	https://docs.goo
2/16	Liu	SenNet+HOAコンペ振り返り	https://www.kag			
3/29	fujimoto	IMC'24 紹介		小野寺	GTC2024	
4/19	ishi,tsuji,nikaido	HMSまとめ	https://docs.goo	fujimoto	Image Matching のサーベイ紹介	https://docs.goo
5/10	全員	kaggleチームキックオフ@京都	https://docs.goo			
5/24	chen	introduction to all-BF16 training	https://docs.goo			
6/7	onodera	SqueezeFormer vs Conformer				
6/21	kambe, liu	SSII報告	https://docs.goo	竹ノ内	鳥コンペ	2024062…
7/5	nikaido	文書画像理解タスクまとめ		Chen	Automated Essay Scoring 2.0	https://docs.goo
7/19	takoi	LEAP解法共有	https://docs.goo			
8/23	liu, chen	LMSYS解法共有	https://docs.goo	竹ノ内	MOT	2024082…
8/30	yamaguchi	2023年に金メダルを獲得した7個のコン	https://docs.goo			
9/13	ozaki	MOT2	2024091…			
9/27	kambe	LLM 20 Questionsコンペ振り返り	2024092…			
10/11	fujimoto	論文紹介: PEFT for Large Models: A Comprehensive Survey				

今年のKaggle workshop一覧



例：UWMコンペ解説 (2022)

Q：Kaggler向けの資料だけで本当に大丈夫？？

- 詳しいモデルの詳細が分からなくても、どんなドメイン？ データの性質は？ 何がポイントだった？
  といった情報が伝われば、十分案件受領時やアサイン時の参考になる
- 非エンジニアのデータ分析に対する解像度が上がると、その分案件の成功確率も上がる
- 例：(エンジニアが同席しないような) 先方との事前打ち合わせ段階で、データリークや性質に
  気をつけた話ができるようになる
  - 結果的に双方利益がある場合が多い

結果どうなった？

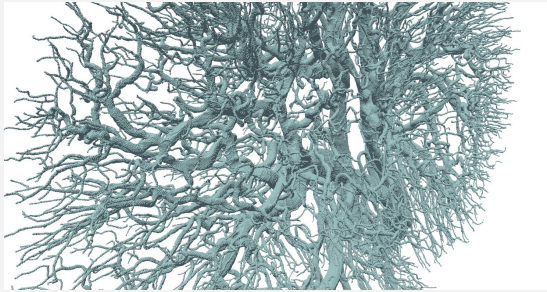
- 案件アサイン時のミスマッチ低減（体感）
- 副次的な効果として、ある案件の相談がきた時に「それに似たタスクxxxさんがやってましたよ」がパッと出てくるようになった

役に立った事例

- 案件のドメイン / タスク：**ライフサイエンス分野 / 3D (Instance) Segmentationタスク**
- **SenNet**コンペという血管の3D Segmentationタスクで直近で実績が作れていた
- 社内でKaggle実績を共有してたため、案件受領時に上記のメンバーをスムーズにアサインできた
- → 先方の予想を上回る納期で高精度なモデルを納品でき、お客様からの評価にもつながった

SenNetコンペ

- HiP-CTにより撮影されたヒト腎臓の画像から血管部分をSegmentationする課題
- 3D rotateによるAugmentation, 2.5Dモデルなどがポイントだった



おまけ：「Kaggleテクは業務の役にたつ」

Kaggleで効いた考え方 / テクニックは案件での打率も比較的高い

- Backbone選択 / Multi-head, Aux Loss / 2.5D / Multi-stage / Pseudo-label...
- 様々なコンペ設計に触れるなかで、良い / 悪い課題設定やデータの性質が
  身に染みてわかる

  この痛みを伴った (?) 経験が大事
- その中の一部をピックアップしてTips的に紹介

Q：案件でKaggleみたいに精度追い求めて意味あるの？

- 案件によってはもちろんある。
- 例えば**製造業では（特にRecallの）精度要求がかなり厳しかったり**する。そういう場面では、データ面での改善 (クリーニング、追加データ収集)と並行して、**モデル側もかなり追い込んでチューニングしていくことは良くある**
- 現状ヒトが目視確認している作業などは、意外と速度要求自体は厳しくない (こともある)
- 「精度勝負では他社に負けません！」ということ自体が売りになったりもする
- 体感としては、
    - 精度60% (or Null) → 80%↑にしたい！　/　精度90% → 99%↑にしたい！
- のどちらもあり得る

適切な精度検証スキームに気をつける

- Kaggleでは信頼できるCV構築が何よりも大事
  - train-test split (場合によってはデータセットシフト含む) / データLeak...
- KaggleではLBを構築するデータへの考察も大事
  - 例えばお客様への説明では、Public test = Validation / Private test = (運用で得られる) Testデータ、と話すこともある
- shakeしない解法には、案件でも使える (一定程度) 運用に耐えるための工夫が隠れていることも多い

適切な精度検証の重要性がわかっていれば、案件**データの受領前段階でリスクを回避**することができる

- 例えば機器の点検業務を行うAI開発で、同じ場所 (例：同じ工場内) に存在するデータをTrain - Test とするのではなく、できれば事業所単位でsplitしてTestに用いる
    - ↑の重要性を理解してもらい、時にはお客様に協力していただき追加データ収集することも（もちろんそれが出来る関係性の構築はかなり大事）

モデル選択

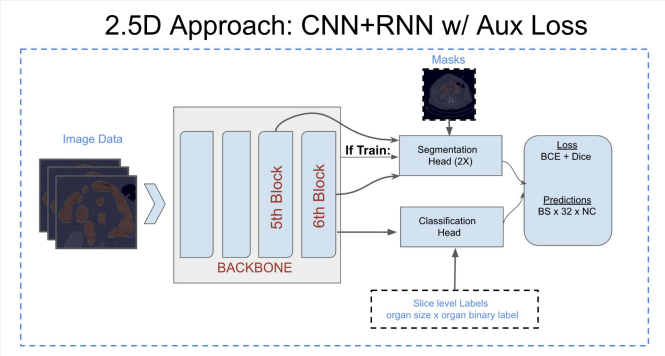
- 例えば画像系モデルのBackboneなどは、Kaggleでの淘汰がかなり進んでいるため、案件でもそこから選択することで不要な試行錯誤を減らせる
    - EfficientNet / ConvNeXt / Swin / (最近だと) MaxViT

Aux Loss

- 本来予測したいTargetとは別のLabel / Mask情報などを同時に学習する
  - 関連するタスクを同時に学習することで、Backboneのより効率的な学習を促す
  - 特に画像分類におけるMask情報の同時予測は案件でも効果を感じやすい (体感)
- 入力としてLabel / Maskを利用する場合とは違い、予測 (運用) 時には上記情報は不要 (+forwardパスも通さなくて良い) ため、運用でも使いやすい手法

RSNA2023コンペ

- Multi-slice CT画像から外傷の程度を予測する課題
- 2stageかつ2.5Dモデルによるアプローチが主流
- 臓器部分を示すSegmentation maskをAux Lossとして学習することで精度向上した (1st place)



RSNA2023: 1st place solution

モデル構造：2.5D

- Multi-slice CT / MRIや動画データ など、XY面内だけでなく、その前後に繋がりのあるデータ
- このようなデータには2.5Dモデルと呼ばれるモデル構造が有効
- slice / シーケンスデータを個別にBackboneに通して特徴抽出した後、LSTMやConv, MLPなどでそれらの情報を統合する
- Kaggleでも近年流行りの手法で、同じようなタスクではもちろん案件でも効く
- データの性質によってモデルの前半で混ぜるか (2D inputのchannel方向に積む)、モデルの後半で混ぜるか (2.5D：各sliceを個別にEncoderに通して後で集約)を変えると良さそう（体感）
- データの性質：sliceごとの類似性や、入力したいslice長など

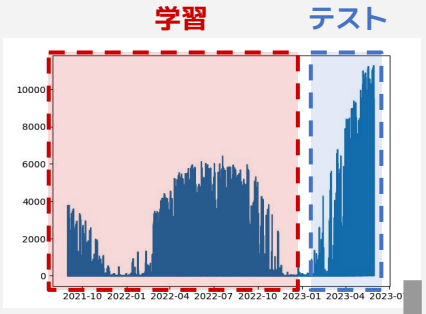
社内でKaggleが役に立った経験を聞いてみました！

モデルのロバスト性向上

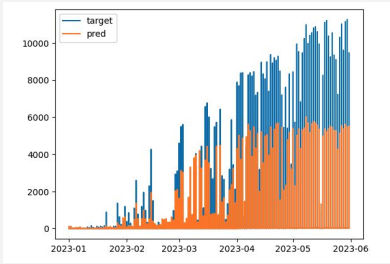
- 例：未来の上昇トレンドにも対応可能なモデルの構築（Takoiさん：Enefitコンペからの知見）
  - GBDTで予測する際に、ラベルを変換してN日前からの差分を予測対象とする

Enefitコンペ

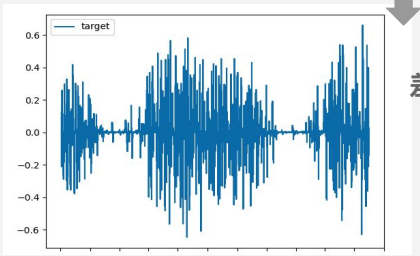
- ある地域におけるエネルギーの生産 / 消費量を予測する課題
- Time-series APIを利用して、モデル提出後に実際の未来のデータで評価
- → モデルのロバスト性が非常に重要



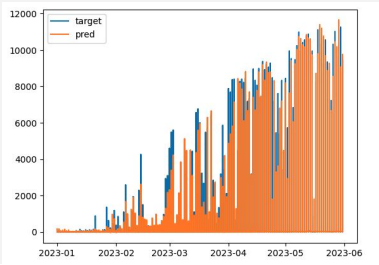
そのまま予測すると…



Target変換



差分予測でロバスト性向上



ライセンスについて

- 最近のコンペはライセンスに ~~うるさい~~ 厳密なことも多い
- そのようなコンペに多くでることで、案件時のライセンス問題への嗅覚もするどくなる (気がする)
- 基本的にはMIT / Apache-2.0のようなライセンスをKaggleでも案件でも使用するようにしている
- ライセンスに気をつかっているモデルライブラリーを使用するのも一つの手
    - 例えばDetection / Instance Segmentationだと**MMDetection**などは良き (Apache-2.0)
        - YOLOv5などのライセンス的に微妙なモデルはしっかり退避させている
            - (MMYOLOとして別ライブラリーになっている)

さいごに

- Kaggleを業務に役立てるには
  - 実力をフルに発揮できるような環境作り！
  - 周りのチームとのコミュニケーションやデータ分析リテラシーの底上げ！
- 技術が洗練される場としてのKaggle
  - Kaggleで有効な手法はどんどん業務でも使っていこう！

- Kaggleを業務に役立てるには
  - 実力をフルに発揮できるような環境作り！
  - 周りのチームとのコミュニケーションやデータ分析リテラシーの底上げ！
- 技術が洗練される場としてのKaggle
  - Kaggleで有効な手法はどんどん業務でも使っていこう！

Kaggleは業務の役にたつ！

以上