ニュースメディアでの機械学習活用事例

ー Kaggle コンペ開催を題材にー

石原 祥太郎

株式会社 日本経済新聞社
データサイエンティスト

本発表の概要

日本経済新聞社における機械学習の活用事例の一端を紹介

1. 日本経済新聞社とデータ活用
2. Kaggle コンペの開催
3. 業務における Kaggle の知見の活用

自己紹介：石原 祥太郎

- 大学時代は工学部で情報技術を学びつつ、学生新聞で編集長などを歴任
- 17 年から現職で、データサイエンティストとして記事推薦などデータ活用に取り組む
- 社外活動として Kaggle に取り組み、コンペの優勝や、入門書『 Python ではじめる Kaggle スタートブック』出版を経験
- 20 年に国際ニュースメディア協会「 30 Under 30 Awards 」のアジア太平洋部門の最優秀賞

日本経済新聞社とデータ活用

01

日本経済新聞社

- 創刊 1876 年で 140 年以上の歴史
- 従業員数 3000 人
- 電子版創刊 10 周年
  - 電子版有料会員は 70 万人
  - 900 万人が「日経 ID 」アカウントを保持

老舗企業でありつつも、常に最新技術を取り込みながら成長

- 1972 年：世界初のコンピュータによる一貫新聞製作システム「ANNECS 」を開発
- 1984 年：データベースサービス「日経テレコン」を開始
- 2010 年：日経電子版を創刊
- 2015 年：英国経済紙「フィナンシャル・タイムズ」買収

データ活用の土壌が整っている

- リアルタイムデータ処理基盤「 Atlas 」を開発・OSS公開
  - https://hack.nikkei.com/blog/atlas_opensource_project/
- 記事データや企業情報などを抽出できるAPIも充実
- 多くの人にデータを身近に
  - 自社データを用いた SQL(BigQuery) 文法の学習用教材
  - データ道場
    - https://speakerdeck.com/yosukesuzuki/nikkei-data-driven-20180823

なぜデータ？ ※個人の認識

- 情報技術の発展に伴い、ニュースメディアの在り方が変化
  - 「マスメディア」だけではなく、個々人が発信
  - 一方的な発信だけではなく、読者の反応が見えるように
- データを活用し、読者に適した情報を届ける
  - パーソナライズ・推薦の需要
  - 内製エンジニア組織

Kaggle コンペの開催

02

Kaggle とは

- 機械学習による予測性能を競うコンペティションのプラットフォーム
- Google が 2017 年に Kaggle を買収
- コンペティションだけではなく、ソースコード・データセットの共有や議論の場も提供
- 世界中のデータサイエンティストが集うコミュニティとして認知度が高まり、称号の価値も向上している

u++

Data Scientist at a media company

Tokyo, Tokyo, Japan

Joined 3 years ago · last seen in the past day

https://upura.github.io/

Followers 432

Competitions Master

Home | Competitions (34) | Datasets (2) | Notebooks (73) | Discussion (178) | •••

Edit Profile

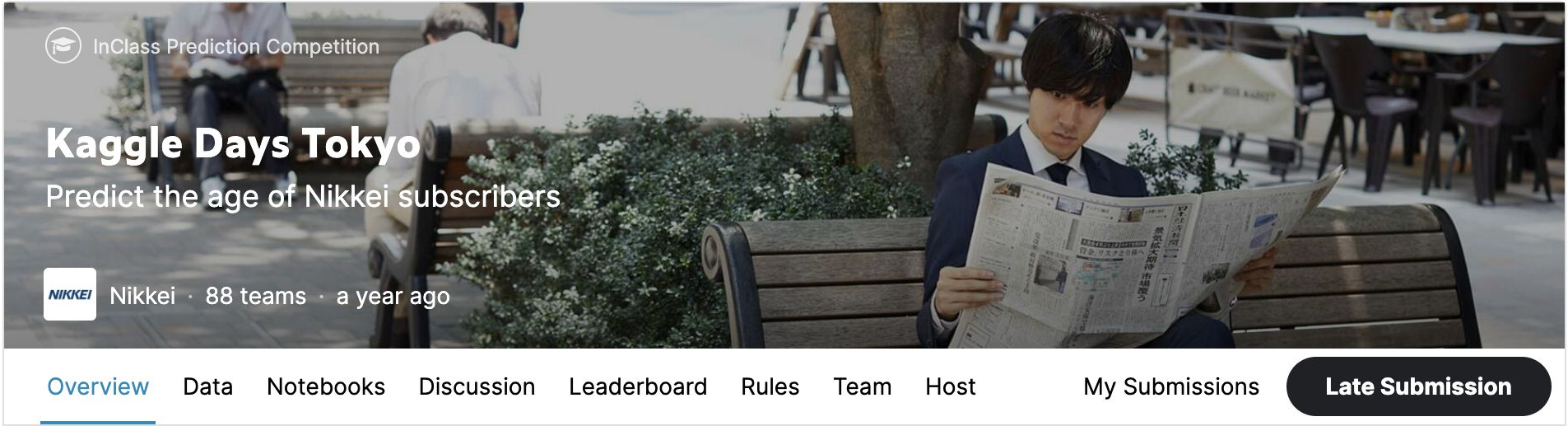
Competitions Master	Datasets Contributor	Notebooks Expert	Discussion Expert
Current Rank 373 of 149,279	Unranked	Rank 140 of 145,018	Rank 130 of 167,444
Highest Rank 229			
1 / 4 / 2	0 / 0 / 0	3 / 3 / 18	16 / 11 / 88
PetFinder.my Ad... · 2 years ago · Top 1% — 1st of 2023	df for visualizati... · 2 years ago — 1 vote	upura-kaggle-tu... · 2 months ago — 478 votes	1st Place Solutio... · 2 years ago — 68 votes
Jigsaw Uninten... · a year ago · Top 2% — 32nd of 3165	submit_files · 9 months ago — 1 vote	python-kaggle-... · 9 months ago — 76 votes	Multi-label com... · 2 months ago — 68 votes
OpenVaccine: C... · 25 days ago · Top 4% — 58th of 1636		MoA: LGBM Ben... · 2 months ago — 68 votes	Real World Probl... · a month ago — 41 votes

https://www.kaggle.com/sishihara

Kaggle コンペの一例

- マレーシアのペットショップの引き取り速度の予測
- COVID-19 のワクチン開発に向けた RNA 分解のモデル開発
- 文書の悪質度の予測
- 動画の「 deep fake 」の検知
- リクルートのレストラン客数の予測
- メルカリの出品価格の予測
- 日本のくずし文字の検出・認識

日経による Kaggle コンペ



2019 年 12 月、日本で開催された「 Kaggle Days Tokyo 」にて

Kaggle Days

- 世界各地で開催されているオフラインイベント
- 普段はオンライン上で競い合っている Kaggler と対面で交流できる貴重な場
- 優れた実績を持つ Kaggler らによるプレゼンテーションや、イベント参加者限定のコンペなどが開催
- ワルシャワ・パリ・サンフランシスコ・ドバイ・北京・東京で開催

Kaggle Days Tokyo

- 12 カ国以上から 465 人もの応募（定数 240 ）
- 2 日目のコンペに 88 チーム 149 人が参加、提出数 1116
- コンペのお題「日経電子版ユーザの年齢推定」
- 入力:
  - 匿名化したユーザの記事閲覧ログ
  - 記事データ
- 出力:
  - ユーザ ID に紐づく年齢

出題の背景

- 属性情報は、ユーザに良質な体験を提供する上で重要
  - 個々人に適した記事の推薦などのパーソナライズ機能の充実を目指している
- 現在は日経 ID に登録する形で、ユーザに属性情報を提供してもらっている
  - 登録の手間
  - 日経 ID を持たないユーザも存在
- 記事閲覧ログや記事データから年齢を予測するモデルが構築できれば、これらの問題に対応できる可能性あり

コンペのルール

- 午前 10 時半に始まり、午後 6 時半までの 8 時間
- 最大 3 人のチーム
- 外部データは日本語の辞書のみを利用可能
- 提出回数はチーム全体で 30 回まで
- コンペ中の 8 時間は、予測用のデータセットのうち 25% のみを用いて計算した暫定のスコアに基づく順位を表示
- 評価指標は、RMSE

最終結果

#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	—	The True Zoo			11.10858	15	1mo
2	—	the overfitting zoo			11.19380	29	1mo
3	▲ 3	Malbin & Seffi			11.26356	30	1mo
4	—	TI			11.28198	18	1mo
5	▲ 2	HILZ			11.29536	18	1mo
6	▲ 2	NARI			11.31640	17	1mo
7	▲ 4	flowlight & higepon			11.32200	26	1mo
8	▲ 7	DSS_Kaggle_Club			11.34872	11	1mo

上位陣の取り組み

- ユーザの記事閲覧ログと記事データから、テーブル形式の特徴量を生成し、機械学習アルゴリズムで予測
- 自然言語処理技術を活用し記事データから効果的な特徴量を抽出していた点が決め手
  - 記事本文の活用（tf-idfなど）
  - 記事の閲覧履歴を系列と見なし Word2Vec など

業務における Kaggle の知見の活用

03

Kaggle コンペの知見

- 開催主としての気付き
  - データの効率的な前処理
  - 自然言語処理の手法を用いた特徴量
  - 想定される性能の見込み
- 国際シンポジウムにも採択
  - “Age Prediction of News Subscribers Using Machine Learning”
    - The Computation + Journalism Symposium 2020
    - https://cj2020.northeastern.edu/research-papers/

Kaggle を通じて得られたこと

- 自分で手を動かした経験
  - 機械学習が得意なこと・苦手なこと
  - データ・手法に対する勘所
  - コードのコピペ
- 自分自身の客観的な指標に
  - 世界基準で自分の取り組みを検証
  - 社内外での認知度に貢献

まとめ

日本経済新聞社における機械学習の活用事例の一端を紹介

1. 日本経済新聞社とデータ活用
2. Kaggle コンペの開催
3. 業務における Kaggle の知見の活用