



機械学習コンテスト

—データサイエンティスト競争（共創）の場—

藤若雅也（NEC 情報・ナレッジ研究所）

広がりを見せる機械学習コンテスト

❖ 機械学習コンテストとは？

企業が保有するさまざまなデータを分析して、売上向上や業務効率改善に活用するビッグデータ分析が注目を集めている。ビッグデータ分析はデータを保有する企業が自ら分析する場合もあれば、データ分析を専門に行う企業と契約して行う場合などその形態はさまざまだが、ここ数年で広がりつつあるのが誰でも参加可能な機械学習コンテストという形だ。機械学習コンテストはデータを保有する企業（スポンサー）が課題とデータを公開して参加者を募り、課題の出来を参加者に競わせるというもので、コンテストによっては賞金が1億円を超えるものもあり、優秀な成績を収めた参加者が賞金を獲得することができる。

有名な機械学習コンテストとしては、2006年に米国動画ストーリーミング大手のNetflixが自社のレコメンデーション機能（Cinematch）の精度を強化する目的で開催した賞金100万ドルのコンテスト「Netflix Prize」である。2011年までを期限とし

て、Cinematchより10%以上の精度向上を達成したチームが現れた時点でコンテスト終了となるもので、最終的に世界186カ国4万チーム以上が参加した。コンテスト開始後2週間でCinematchより優れたアルゴリズムが登場し、徐々に精度を上げていき、遂に2009年にAT&TやYahoo!の研究者などから構成される合同チームが条件を満たすアルゴリズムを開発し、賞金を獲得した。

❖ コンテスト開催の場を提供する企業

Netflixは自社でコンテストを開催したが、近年、このようなコンテストの場をプラットフォームとして提供する企業が登場してきており、FacebookやMicrosoftといったトップ企業もその場を活用して自身が持つ課題の解決やリクルーティング等に役立てている。このようなコンテストの場を提供する企業としてはTopcoder、Kaggle、CrowdANALYTIX等が有名である（表-1）。Topcoderは元々プログラミングのスキルを競う競技プログラミングの場として有名であり、2013年の米アピリオによる買収以降、機械学習のコンテストにも力を入れてきている。Kaggleは機械学習に特化してコンテストを提供する企業であり、近年急速にユーザ数を伸ばしている。

ユーザ数はクラウドソーシングも手掛けるTopcoderが最も多いが、機械学習コンテストの場としてはKaggleが世界最大である。本稿の後半では、このKaggleにフォーカスをあてて具体的な参加方法や仕組みなどを紹介する。

企業名	概要	ユーザ数 (2015/3時点)	設立年
Topcoder	競技プログラミングを始め、クラウドソーシング、コンテスト等を手がける。現在はグラフィックデザイン、ソフトウェア開発、データ分析（機械学習）の3つのカテゴリでコンテストを開催している。	760,732	2001
Kaggle	データ分析に特化したコンテストを開催し、近年急速にユーザ数を伸ばしている。世界最大規模のデータサイエンティストコミュニティを持つ。	272,304	2010
CrowdANALYTIX	インドのスタートアップ。Kaggle同様データ分析コンテストに特化している。	7,578	2012

表-1 コンテスト運営企業

❖ 広がりを見せるわけ

このような機械学習コンテストが広がりを見せている背景として、機械学習ツールの充実化とコンピュータリソースの低廉化が挙げられる。ツールとしては、統計分析を行うための言語である R や Python の統計分析用のモジュールである scikit-learn などがあり、オープンであるため誰でも手軽に利用することができる。また、Apache Hadoop や Apache Spark などのビッグデータ分析基盤も登場し、大規模なデータでも比較的簡単に分析できるようになってきた。

コンピュータリソースについても、今や高価なサーバを購入せずとも、Google Compute Engine や Amazon EC2 といった IaaS が安価に利用できる。たとえば、Amazon EC2 では 1 仮想 CPU、3.75GiB メモリの仮想マシンが 1 時間あたり約 1 円で利用できてしまう（2015 年 3 月時点のスポットインスタンス価格）。データ分析では、一時的にリソースが必要になる場合が多く、リソースを時間単位で購入できる IaaS とはとても相性が良い。

❖ コンテスト参加のメリットと注意点

機械学習コンテストはスポンサー企業、コンテスト参加者それぞれにメリットがある。スポンサー企業のメリットとしては、自身が保有する課題の解決を短期かつ安価に実現できる点にある。また、その品質も高く、たとえば Kaggle では大学の教授やトップ企業のエンジニア等超一流の人材が参加している。気を付ける点としては、企業は自身が保有するデータを公開するため、他社に利用されたり、戦略を悟られる可能性もある。また、データ公開にあたってはプライバシーの考慮も必要である。前述した Netflix では、「Netflix Prize」の続編となるコンテストを開催しようとしたが、米国連邦取引委員会（FTC）がプライバシーの観点で懸念を示し、最終的に開催を断念したという経緯がある。ただ、Kaggle 等では、データを匿名化するなどのサポートも行っているため、そのあたりのハードルは下がってきている。

コンテスト参加者のメリットもいろいろある。機

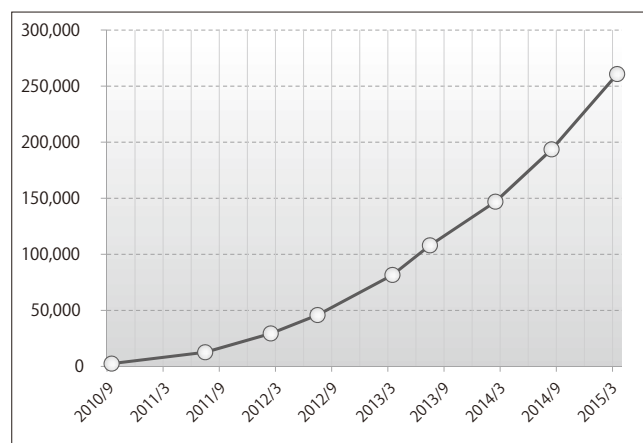


図-1 Kaggle ユーザ数

械学習を勉強したい参加者にとってはリアルなデータを使えるというのは非常にありがたい。また、コンテストで優秀な成績を収めることで自身の能力を企業にアピールすることができ、就職や転職などにも役立つ。実際 Topcoder はリクルートセンターとしての側面が強く、Google、Yahoo! を始めとした多くのトップ企業がリクルートの場として利用している。また、Kaggle でも Facebook がリクルート用コンテストを開催し、何人もの参加者が Facebook に採用されたという実績がある。Kaggle としても学生の囲い込みにも力を入れていて、Kaggle-In-Class という授業用にコンテストを開催できる仕組みを無償で提供している。

Kaggle の概要

❖ 世界最大の機械学習コンテストの場

Kaggle は 2010 年にオーストラリア（メルボルン）に設立された企業で、2011 年に拠点をサンフランシスコに移している。現在 100 以上の国、200 以上の大学からデータサイエンティストが参加している。Kaggle 自体はデータや課題を提供するスポンサー企業とコンテスト参加者を仲介する役割を担っている。従業員は 12 人と少ないが、近年そのユーザ数を急速に伸ばしていて、現在 27 万人を超えるユーザを抱えている（図-1）。Kaggle 自身、その従業員の多くは Kaggle のコンテストからリクルートしている。

コンテスト名	スポンサー	コンテスト概要	賞金	開催期間
National Data Science Bowl	Booz Allen Hamilton	海中カメラによって撮影されたプランクトン画像イメージを約 120 のクラスに分類し、その分類精度を競う。3 万枚の学習用画像イメージと 13 万枚のテスト用画像イメージが用意されている。	\$175,000	2014/12/15 ～ 2015/3/16
Driver Telematics Analysis	AXA	約 2,500 名のドライバの自動車走行データ (GPS) からドライバの運転の癖を抽出し、ドライバの識別精度を競う。	\$30,000	2014/12/15 ～ 2015/3/16
March Machine Learning Mania 2015	HP	過去 30 年の NCCA (全米大学体育協会) のバスケットボールの試合データを用いて、2015 年のトーナメントの勝敗予測を行う。	\$15,000	2015/2/2 ～ 2015/3/14

表-2 開催中コンテスト一覧 (一部)

❖ コンテストが開催されるまで

コンテストを開催したいスポンサー企業は、Kaggle と契約し、コンテスト開催費用と賞金を支払う。Kaggle は、コンテスト開催に必要なプラットフォームを提供するとともに提供データの一貫性チェックや評価基準の選定等のコンサルティングを行う。プラットフォームは自動ランキング機能や参加者同士の情報交換の仕組み(フォーラム)を提供する。契約後、Kaggle のスタッフとともに数週間のデータ準備などを行い、コンテスト開催となる。

❖ どのようなコンテストがあるか

Kaggle のコンテストのカテゴリとしては、大きく商業課題 (Featured)、研究課題 (Research)、お遊び (Playground)、練習用 (Getting Started) に分類される。商業課題や研究課題の多くには賞金がついている。研究課題では賞品として ICML や ICDM といった機会学習分野のトップ学会への招待やジャーナルへの掲載などがつく場合もある。

最も有名なコンテストの 1 つが 2011 年から 2 年に渡って開催された賞金 300 万ドルの「Heritage Health Prize」だ。病院患者の過去のデータから次の年のその患者の入院日数を予測するというもので、スポンサー企業 (Heritage Provider Network) が提示した予測精度を満たした 2013 年の優勝者が 300 万ドルを獲得することができる。残念ながらこのコンテストの優勝者はその予測精度を満たすことができなかったため 300 万ドルは獲得できなかったが、

優勝賞金として 50 万ドルを獲得した。

現在 (2015 年 3 月時点) 開催されているコンテストの一部を表-2 に示す。「National Data Science Bowl」は、海中カメラのプランクトンの画像データを学習して、プランクトンを識別できるようにする、というもので優勝したアルゴリズムは今後の海洋研究に使われることになる。「Driver Telematics Analysis」は、自動車保険を取り扱っている AXA がスポンサーになっていて、コンテストの内容としては、約 2,500 名の自動車のドライバの走行データを学習して、ドライバを識別するアルゴリズムを開発するというものだ。近年、欧米では自動車に自動車事故通報システムを搭載する動きがあり、これと連動して、保険業界でも自動車に搭載されたシステムの情報を活用した保険「テレマティクス保険」を普及させようという動きが出てきている。そのような動きに先駆けて技術を先行開発しようという狙いがうかがえる。

❖ コンテストの賞金

これまで Kaggle で開催されたコンテストは約 150 あり、その賞金別件数は図-2 のようになる。Kaggle は Topcoder と比べ、高額な賞金が出るコンテストが多く、10 万ドルを超えるものが過去 10 回も存在する。

Kaggle への参加方法

ここからは、Kaggle の具体的な参加方法、分析

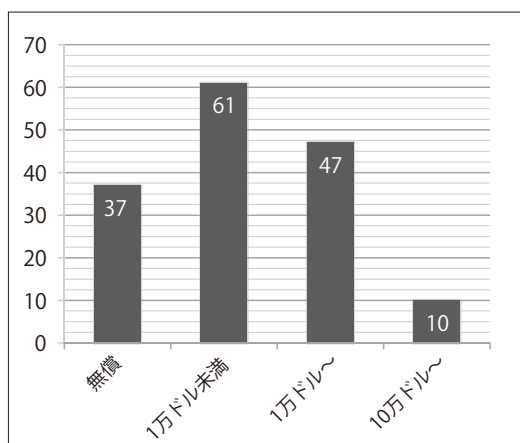


図-2 賞金額別コンテスト数

する際のポイントや注意点などを解説する。

Step1：参加者登録する

Kaggle のコンテストに参加するためには、まず、ユーザ登録が必要である。Google, Yahoo! や Facebook のアカウントを持っていれば簡単に登録できる。

Step2：コンテストを選ぶ

Kaggle では大体 10 以上のコンテストが常時開催されている。一部のコンテストを除き基本的にどのコンテストにも参加可能だが、賞金額が大きいコンテストや、分析が行いやすいコンテストは人気があり、上位に食い込むのも難しくなる。また、開催期間が短いコンテストはあまり人気がなく、時間がとれる人にとっては狙い目かもしれない。

今回は「Driver Telematics Analysis」を詳細に見てみることにする。

Step3：データをダウンロードする

データはコンテストの専用ページからダウンロードできる。圧縮ファイルで数 GB を超えるコンテストも多いので、自身の分析環境のディスク、メモリ容量をチェックしておくことが必要だ。

「Driver Telematics Analysis」の場合、データは圧縮済みファイルで 1.44GB あり、展開後は 5GB を超えるサイズとなる。約 2,500 名のドライバの 200 回の走行データがそれぞれ csv ファイルで用意されている。各 csv ファイルには 1 秒ごとの GPS 位置情報が記載されていて、数分～数十分の走行データとなっている。データはプライバシー保護のため、相

対的な GPS 位置に変換されている。

Step4：データを分析する

データの分析は各自どのように行っても構わないが、コンテストによっては「賞金を受け取るためにはオープンソースを利用しなければならない」などのルールが設けられている場合が多く、ルールは一通り目を通しておく必要がある。また、コンテストによっては、分析方法に関するチュートリアルが用意されている。分析した結果は、決められたフォーマットに従う必要があり、データダウンロードページにサンプル提出ファイルが用意されている。

「Driver Telematics Analysis」の分析では、各ドライバの 200 の走行データのうちいくつかに偽物（ほかの人が運転した走行データ）が紛れていて、それらを特定することが目的になっている。答えは公開されておらず、投稿して初めて結果が分かる。ドライバを特定するには、ドライバの運転の癖を走行データから抽出する必要がある。

Step5：フォーラムに目を通す(オプション)

各コンテストには、フォーラムと呼ばれる参加者同士の情報交換ページが用意され、参加者はその場で自由に情報のやりとりをすることができる。多くのコンテストでは、プログラムや提供データに関する他チームとのプライベートな情報交換は禁止されているものの、フォーラムのようにパブリックな場所での情報交換は許可されている。

「Driver Telematics Analysis」のフォーラムでは、コンテストの内容に近い論文の情報（スタンフォード大学のドライバ認識に関するもの等）や自分が試したアルゴリズムの結果や、データにおかしなものがあるとか、さまざまな内容が議論されている。この分野に詳しくない人でもフォーラムを一通り眺めるだけで、かなりの情報を得ることができる。

Step6：チームを作る(オプション)

コンテストは個人で参加するか、チームで参加するかを選択することができ、途中でチームに参加することもできるし、他のチームとマージすることもできる。チームで参加する場合、チーム内で情報共

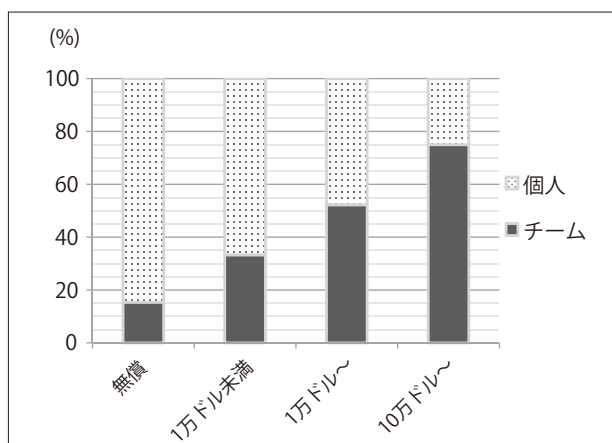


図-3 優勝者の内訳

有や議論が行える点がメリットとしてある。特に賞金額の高いコンテストでは、チームを作ることによって上位に食い込める可能性が高くなる。Kaggleの過去のコンテストの優勝者の内訳を見てみると(図-3)、賞金額の高いコンテストほどチームでの優勝が多いことが分かる。賞金額の高いコンテストは、参加者も多く一般にレベルが高い。チームを作りお互いのアルゴリズムをチーム内で公開して議論・改善することで、より優れたアルゴリズムを開発することができる。前述のフォーラムを使ってチームメンバーを募集している参加者も見かけられる。

Step7: 分析結果を投稿する

投稿ページ(図-4)で分析結果のファイルをドラッグ&ドロップするだけで投稿が完了する。投稿後数秒でLeaderboardというランキングページに自分の順位が掲載され、前回の投稿からどのくらいランクアップしたかなども表示される。多くのコンテストでは投稿回数に"1日5回まで"といった制限が設けられている。

注意すべきは、コンテスト開催中のLeaderboardランキングは一部のデータの評価結果に基づく暫定版となっており、コンテスト開催中は正確な順位は分からない点である。これは評価に使ったデータに特化した学習をしてしまい未知のデータに対する精度が悪くなる、いわゆる過学習の防止を目的としたもので、最終的なランキングはコンテスト期間終了後に残りのデータの評価結果を反映させたもので決定される。

コンテスト期間終了後もフォーラムでは盛り上が

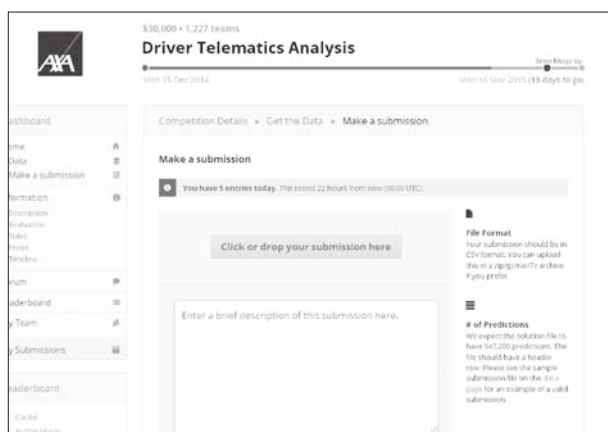


図-4 分析結果投稿ページ

<http://www.kaggle.com/c/axa-driver-telematics-analysis/submissions/attach>

りを見せる。ランキング上位の参加者が自身のソースコードのレポジトリを公開することもあり、その内容についての議論が活発に行われている。Kaggleではコンテスト終了後も投稿は行えるようになっており、公開されたソースコードを実際に試すことで最先端の技術を身をもって体験することができる。

競争と共創

機械学習コンテストというと参加者同士が単純に結果を競い合うというイメージが強いが、実際はフォーラムやチーム形成などを通して活発な情報交換が行われていて、お互いが切磋琢磨できる環境になっている。力試しをしたいと考えている人はもちろんのこと、これから機械学習を始めたい、リアルなデータで実践力をつけたい、と考えている人は一度挑戦してみてもはどうだろうか？

このような機械学習コンテストの普及に伴い、今後コンテストにおける成績がデータサイエンティストの能力を示す1つの指標として定着する日が来るかもしれない。

(2015年3月2日受付)

■ 藤若雅也 m-fujiwaka@cd.jp.nec.com

2007年日本電気(株)入社。携帯端末の開発やデータセンター、キャリアシステムのキャパシティ分析、性能分析などの研究に従事。