# データ解析コンペティションへの挑戦

中田　和秀

ここ数年，われわれの研究室はデータ解析コンペティションに参加している．このコンペティションへの参加を通じて感じたこと，また，充実したデータ解析をグループで進めていくコツについて，われわれの経験をもとに紹介する．

キーワード：データ解析コンペティション，人材育成，機械学習

## 1. データ解析コンペティションとは

ここ数年，研究室で有志が集い「データ解析コンペティション」(以下「コンペ」という）に参加してきた．このコンペは経営科学系研究部会連合協議会が主催しており，共通の実データをもとに参加者が分析を競うものである．本誌 2 月号は，このデータ解析コンペティションの特集であり，ご覧になった方も多いだろう．ここ 3 年間の参加状況を次に示した（なお，2016 年と 2015 年は 2 つの提供データに対し独立してコンペが行われており，その合計数である)．

2017 年：93 チーム，約 550 人

2016 年：118 チーム，約 700 人

2015 年：104 チーム，約 600 人

上記より参加者も多く盛況であることが見て取れる．

近年，Kaggle，KDD Cup，Deep Analytics，マーケティング分析コンテストなど，国内外で実データを分析するコンペティションは盛んに催されている．それらの中で，本コンペの特徴として次の点が挙げられる．一つ目の特徴は，1994 年から毎年開催されており，今年で 25 年目という大変に歴史のあるコンペであることである．開催当初は現在のようにデータマイニングやデータサイエンスといった言葉の認知度は高くなく，そのような時代から毎年開催されてきたことは特筆に値する．その継続性が参加者数，分析の質，注目度などに繋がっていると感じる．二つ目の特徴は学会系のコンペであり，学生や大学教員の参加が多いということである（参加者のおよそ 7 割は学生だという)．そのため，分析には実用性のみならず学術的な新規性も問われる傾向にあり，その最終形が本誌特集に掲載される査読付き論文となる．また，発表会場では学生の教育の場という温かい雰囲気もいくぶん感じられる (学生教育という点における運営者側の意図が [1] で説明されている)．三つ目の特徴は，「画像から不良品を判別する」や「入院の期間を予測する」といった明確な目的が与えられ，その精度を数字の大小で競うのではなく，データが与えられるだけで，それをどのような目的で分析し，結果をどのようにビジネスに役に立つかも含めて考えることが求められていることである．よって，データ解析の知識・技術だけでなく，ビジネス視点でのアイデアも問われる．また，Kaggle のようにリアルタイムで成績が更新され順位がわかるような仕組みはなく，最後にプレゼンテーションを行い，それを審査員が評価する．なお，約 100 チームを一斉に集めて発表会を行うことはできないため，まず各研究部会に分かれて発表会を開催し，そこで選ばれた 15 チーム程度が最終の成果報告会で発表を行うという，2 段階の審査となっている．例年のスケジュールは大まかに次のとおりである．

8 月頃　発会式，参加申し込み

9 月頃　データ配布

11 月頃　研究部会で中間発表

翌 2 月頃　各研究部会で発表会

翌 3 月頃　最終の成果報告会

翌 7 月頃　本誌特集号へ論文投稿

本稿が掲載されて 3 カ月後くらいに，次回のコンペティションが開催されるはずである．

## 2. 学生が参加する意義

研究室として何度かこのコンペに参加して強く感じるのは，学生たちにとってこのコンペがとても有意義な体験となっていることである．まず一つに，このコンペに参加することが，実データを分析する貴重な機会となっていることが挙げられる．実データの分析は，

なかた　かずひで
東京工業大学工学院
〒 152–8552　東京都目黒区大岡山 2–12–1
nakata.k.ac@m.titech.ac.jp

 Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited. 

ある程度整えられているデータの分析とは違う，実データならではの苦労も多いのだが，その困難を乗り越えて実務で役立つという目に見える成果が出せる喜びがある．大学周辺ではなかなか触れる機会が訪れない実データを使った分析は貴重な経験となる．次に，コンペの参加によって，学生たちが著しく成長できることが挙げられる．半年近くにわたる取り組みは大変ではあるが，大変だからこそ学生たちの知識・技術・問題発見力・分析力・ディスカッション能力・プレゼンテーション能力などが大きく向上する．また，仲間と共同で最後までやり遂げるという責任感も身につく．3 月にコンペを終えた学生たちが，半年前と比べ見違えるように頼もしくなることも多く，その教育的価値は計り知れない．さらに近年では，このコンペでの実データ分析経験を活かして，卒業後にデータアナリストやデータサイエンティストになる学生も増えてきた．すなわち，コンペの参加によって将来の仕事の選択肢が広がる可能性もある（就活でも，コンペ参加の話はウケがよいと聞く）．そういえば，卒業後は主戦場を Kaggle に移し，Kaggle Master になった学生もいた．なお，事前に申請すれば，コンペでの分析結果を卒論や修論として発表することも認められており，そちらでも実利があるかもしれない．

本稿を読まれている学生で少しでも興味をもたれたならば，ぜひとも参加をお薦めしたい．とはいっても，データ分析の知識も技術も乏しい状況で，いきなりコンペに参加して戦えるのかと不安に思われる方もいるかもしれない．実は，筆者も最初はそう思っていたのだが，意外と健闘できるというのが現在の感想である．その理由を述べる前に，このコンペに参加するチームの傾向をみてみよう．参加チームは，分析者の所属によって大きく次の三つに分類できる．

・学生主体のチーム
・大学教員主体のチーム
・社会人主体のチーム

われわれのチームは学生主体のチームである．大学教員や社会人が主体のチームは，知識や技術があると思われるが，当然仕事が優先であり，平日夜や土日に分析を行わざるを得ない（それも残業や家族サービスでつぶれることが多いと聞く）．よって，あまり分析に時間をかけることができない．一方，学生主体のチームは，多くの時間を費やして分析を進めることができる．後述するように，データ分析は時間をかけないとよい結果を出すのは難しいため，これが学生チームの最大の武器であり，健闘できる要因となる．逆にいうと，知識や技術で劣っている学生主体のチームが，時間もかけないのでは勝つことは難しい．

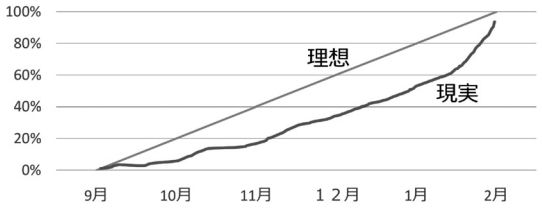


**図 1**　完成度の推移

## 3.　データ分析の進め方

ここからは，初めてデータ解析コンペティションに参加する学生が，充実した分析を行うためのポイントを説明したい．われわれが得た知見がデータ分析の初心者に少しでも参考になれば幸いである．

### 3.1　ゼミ

学生チームが充実した分析を行うためには時間をかけることが重要であることをすでに述べた．しかしながら，「結果が出たら集まって議論をしよう」というような方針では，人間の特性として先延ばしになりがちである．あっという間に発表間近になって焦り，もっと早く始めておけばよかったと後悔することになる．そのような事態を避けるため，データを受け取った直後から，毎週時間を決めてゼミを行うことをお薦めする．このゼミで進捗状況の確認とそれに関するディスカッションを行い，今後の方向性を決めることによって，当初から時間をかけて着実に分析を進めることができる．

また，グループで分析を行うと，参加者の中で「やる気」に差が出てくることがある．分析をサボる人が出ると，それまで頑張ってきた人のやる気を削ぎ，全体のパフォーマンスが落ちることが多い．チーム全体の士気を落とさないことも肝心で，われわれのチームでは（原則として）2 週間に一度以上の頻度で各自が得た成果を報告することにしてきた．

そのようなやり方によって，当初から時間をかけて分析を進めているのであるが，やはり最後は慌てて，発表前 2〜3 週間の頑張りでなんとか完成させることにはなる．われわれの経験をもとに，経過時間と完成度の関係を表すイメージ図を図 1 で示した．理想どおりにはいかないものの，上記の方針によって発表 2〜3 週間前まで 70％程度は進んでいるため，最後の踏ん張りでなんとか 90％以上の完成度にもっていくことができる．これまでさまざまなチームの発表を聞いてき


Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.

たが，分析のアイデアは面白いのだが，それが昇華しきれておらず，もう一段深く分析できたら素晴らしいものになったのに，と感じることも多い．これは，当初から多くの時間をかけることによって分析を完遂させることが重要だということを示唆している．

### 3.2 データの前処理

実データというのはそのままでは分析には使えず，データをクレンジングするさまざまな処理が必要である．たとえば，欠損値に対するデータの補完や該当レコード・フィールドの除去，外れ値や異常値に対する除去や修正が挙げられる．このような作業を自動で行う手法も提案されているが，大抵うまく働かない．よって，データを丹念に調べ，人間の知識・経験などをもとに地道に手動で行う必要がある．この前処理は専門家が行っても時間を費やすことが多く，実務家が記した著書 [2] でも次のような記述がある．

> 筆者の経験上，データ分析では前処理プロセスが最も時間を要する場合が多く「データ分析作業全体の 8 割を占める」という説までであります．

この前処理は大変重要であり正しいデータを作っておかないと，いくら分析してもよい結果は得られない (garbage in, garbage out)．その事実を端的に表現している [3] の一節を紹介する．

> 良きデータこそが良きデータ分析の要です．その後の分析手法の適用がデータ解析の成功に与える比重は，データの良さに対して微々たるものです．

われわれもこの前処理の大変さと大切さは痛感しており，上記の一節に深く納得をする．しかしながら，この前処理の努力は，論文査読ではほとんど評価されないという悲しさがある．それはともかく，前述したデータ分析には時間がかかるというのは，この理由が大きく，特に学生主体チームは時間をかけて丹念にデータを追い，適切に前処理を行うことが肝要である．

### 3.3 分析

このコンペでは，分析の「目的」と「方法」両方を自分たちで設定する．このためには，問題発見力と問題解決力が共に必要となる．うまく問題発見をするには，平凡な意見であるが，普段から幅広い興味をもつことと，チーム内でディスカッションを重ねることが大事だろうと思う．分析の方向が定まれば，現在ではデータ処理にはさまざまなフリーのツールが使えるため，プログラミングの技術が低くてもある程度分析を進めることができる．われわれのチームが利用している次の二つの無料ソフトウェアは，どちらも高度な機械学習の手法を手軽に利用できるものである．

**Python** 汎用スクリプト言語で，Scikit-learn, Pandas, Chainer などの機械学習パッケージが利用できる

**R** 統計処理用フリーソフトで，データ分析のライブラリが充実している

また，企業のご好意で商用ツールが提供されることもある．もちろん，ツールが高性能といえども，データを入力すればすぐに素晴らしい結果が出力されるというわけではない．分析に使う特徴量を巧みに設計することによって，分析のパフォーマンスが向上することも多い．また，複雑な分析モデルでよい結果を得るには膨大なデータ数が必要なため，単純な分析モデルを用いるほうがよい結果となることも多い．すなわち，闇雲に複雑な手法を使うのも考えものであり，いくつかの手法を試行錯誤することが必要である．

なお，分析後はできる限り分析手法の妥当性の検証は行ったほうがよい．それを行うことによって格段に説得力が増す．ただし，実務の現場では A/B テストを行うことによって，簡単に検証ができることもあるが，コンペでは正確な検証が困難であることが多い．この点はわれわれも毎回頭を悩ますところである．

### 3.4 プレゼンテーション

データ解析コンペティションでは，最後に分析結果を発表をして審査を受けることになる．運営側から公表されている審査の観点は以下のようになっている．

**学術的新規性**

- 有効・分析モデルの新規性
- 新たな消費者行動モデル
- 新たな統計モデル　　など

**ビジネス視点での有効性**

- マーケティングアクションの提示
- 新たなビジネスモデルの創生　　など

**プレゼンテーション**

- 説得力
- 適切な質疑応答　　など

これまで何度か審査される側を経験してきたが，上記の視点に沿ってしっかりと公平に審査が行われているように感じている．よって，これらの項目を意識して 15 分程度のプレゼンテーションを行うとよい．審査員は経営科学系研究部会連合協議会を構成する研究部会の代表者，実務家，データ提供元など 5〜10 名である．経営科学系研究部会連合協議会は次のようにさまざまな学会や企業から構成されている．

 Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited. 

**図 2**　発表会場の様子

**図 3**　表彰式後の記念撮影

・日本オペレーションズ・リサーチ学会
　データサイエンスとマーケティング分析研究部会
・日本マーケティング・サイエンス学会
　ID 付 POS データ活用研究部会
　消費者・市場反応の科学的研究部会
　消費者行動の学際的研究部会
　市場予測のための消費者行動分析研究部会
・日本計算機統計学会
　データ解析スタディーグループ
・日本データベース学会
　ビジネスインテリジェンス研究グループ
・ACM SIGMOD 日本支部
・日本経営工学会 経営情報部門
・株式会社 NTT データ技術革新統括本部技術開発本部
・株式会社産業科学研究開発センター

すなわち，審査員の大半は OR を専門としていない人であることに留意する必要がある．専門外の人にも十分伝わるよう，シンプルで明快なストーリーを作り，視覚的な理解も利用した，わかりやすい発表が必要になる．

発表風景を撮った写真を図 2 に掲載した．学生にとって，このような大きな会場で発表することは慣れていないため緊張したかもしれないが，それもよい経験となっただろうと思う． 図 3 は，表彰式後に記念撮影した写真である．学生たちはすべてをやり遂げたという，表情をしており，分析の充実ぶりがうかがえる．

## 4. おわりに

参加する側にとって，データ解析コンペティションは実データに触れながら自らが成長できるという素晴らしい場である．一方，大変な価値を秘めているデータを公開することに対し，提供企業側に抵抗感があることも否めなく，毎年コンペ用の実データを用意する運営側の苦労は想像に難くない．また，100 チーム，500 人以上の参加者の管理も並大抵のことではなく，滞りなくコンペティションを終えるためには，大変な労力を伴っていると思われる．そのような中，毎年データ解析コンペティションを開催していただいていることに対し，生田目先生をはじめとする関係者の皆様にはこの場をお借りし深く感謝したい．

### 参考文献

[1] 生田目崇，“OR におけるマーケティング教育と研究―「データ解析コンペティション」を通して―，”オペレーションズ・リサーチ：経営と科学，**61**(11), pp. 774–777, 2016.
[2] 中川慶一郎，小林佑輔（編），『データサイエンティストの基礎知識―挑戦する IT エンジニアのために―』，リックテレコム，2014.
[3] あんちべ，『データ解析の実務プロセス入門』，森北出版，2015.


*Copyright © by ORSJ. Unauthorized reproduction of this article is prohibited.*