機械学習コンペティション・プラットフォーム「Kaggle」へのチャレンジ

AIブームの中、機械学習モデルの予測精度を競うハッカソンが増えてきています。その中でも特に注目を集めているのが、世界最大級の機械学習コンペティション・プラットフォームである「Kaggle」です。Kaggleは、提示された課題に対して、企業や研究機関などが投稿したデータを用いて、世界中のデータ・サイエンティストや研究者がその最適な予測モデルを構築し競い合う場です。筆者は2018年8月に開催されたコンペティションにおいて、世界中から集まったメンバーと協力し、当時過去最大の7,000を超える参加チームの中で2位に入賞しました。本コラムでは、筆者が考えるKaggleの魅力やコンペティションにおける工夫などを紹介します。

1. Kaggleに参加したきっかけとその魅力

筆者がKaggleに参加し始めたのは今から約3年前です。もともと大学で数学を専攻していたこともあり、現実課題を数理科学を用いて解決することに興味を持っていました。そこで、データ分析や機械学習を使った実践的な経験を積み、そのアウトプットを評価できる最適な場であるKaggleに参加し始めました。参加した最初の頃は良い結果が得られませんでしたが、徐々にコツをつかみ、今までに10回入賞することができました。

Kaggleは、企業や研究機関などが主催者となりコンペティションを開催する形式を採っています。主催者から出される課題は、PCのマルウェア感染予測、レントゲン画像を用いた肺炎の自動検出などさまざまです。参加者は課題を解決するための最適な予測モデルを構築し、モデルはWeb上で即時に採点されてスコアがランク付けされます。上位に入るとメダルが獲得でき、中でも上位の金メダルを獲得すると、トッププレイヤーの証である“Kaggle Master”の称号が与えられます。筆者はKaggle Masterを保有しています。さらに複数の金メダルを獲得すると、“Grand Master”の称号が与えられます。

筆者にとってKaggleの最大の魅力は、世界中のデータ・サイエンティストと競い合えることです。Kaggleにはユーザーランキングが存在しており、常に向上心が掻き立てられます。また、コンペティション開催中はディスカッションというページが開設され、参加者同士がアイデアや最新の研究成果を試した結果などを共有し合い、知見を深めることができます。

2. Kaggle入賞のコツ

Kaggleのランク上位のレベルになると、モデルの精度を向上させるための常套手法を皆使用しているため、スコアに差がつきにくいことが多くあります。そのため、課題に関する情報をよく読み、自分の構築したモデルは最適化できているか、提供されたデータで必要な情報は十分そろっているかなどを見直すことが重要になります。そして、評価指標に合わせてモデルをチューニングする、必要なデータを自ら作り出すなど、より細かい部分で差をつけられるかが勝負の分かれ目になります。

ここで、筆者のチームが2位に入賞したコンペティションで工夫した点について、具体例を交えて紹介します。課題は「クレジット会社のローンの支払いリスクを予測する」というもので、簡単に言うと、顧客がローンを滞りなく返済できるかを予測するというものです。提供されたデータは、顧客の性別や借入金額、月々の返済金額といったローン申請に関わるデータ、過去のローンに関する履歴データ、外部機関による信用度スコアなど、複数のテーブルと合計数百を超える列が存在して

いました。それぞれのテーブルの関係性も複雑で、その中からいかに重要な情報を抽出できるかが鍵でした。

筆者は、Kaggleでのタスクを**図1**のように考えています。データの探索的解析と特徴量抽出において、筆者のチームは、「①複数存在するテーブルをどのように活用するか」「②ローン返済におけるリスクを特定するための重要な情報が不足していないか」という2点に着目しました。

【着目点①】複数存在するテーブルをどのように活用するか

提供された過去のローン履歴データについて、当初は過去の平均借入金額などを算出して使用していましたが、それだけではなく、さらに過去の情報だけから現在のローンに関するリスクを予測するモデルを構築し、その予測スコアも使用することにしました。その結果、現在のローンの情報は一切考慮せずに、過去の状況から現在顧客がどれぐらいのリスクを持っているかといった補助的な情報を組み込むことができ、スコアを向上させることができました。

【着目点②】重要な情報が不足していないか

提供されたデータの中に、ローンの支払いリスクを予測するための利率の情報がないことに気付きました。さらに分析を進めていくと、参考データとして提供された過去のローン履歴データの中には、元本、毎回の返済額、返済期間の情報があり利率を求められるものの、現在のローン申請データには返済期間の情報が不足していて利率を求められませんでした。そこで、過去のローン履歴データを教師データとしてモデルを構築し、現在のローンの返済期間を予測することにしました。その予測返済期間を用いて利率を算出し、それをモデルに組み入れることでスコアを向上させることに成功しました。

このように、わずかな精度を競うことが多いため、一見すると汎用的な予測モデルの構築技術とはかけ離れているように思うかもしれません。しかし、わずかな数値を向上させるにはモデルの基礎から応用まで、より高度な理解が必要です。モデルはわずかな変更を加えるだけで精度が変化しますが、その変化が偶然なのか、それとも汎用的なのかを見極めることが重要です。どの入力データが効果的か、モデルがどのように入力データから学習しているかといった理解やそれを実装する技術力があるからこそ、細かいレベルで精度を向上させることができるのです。

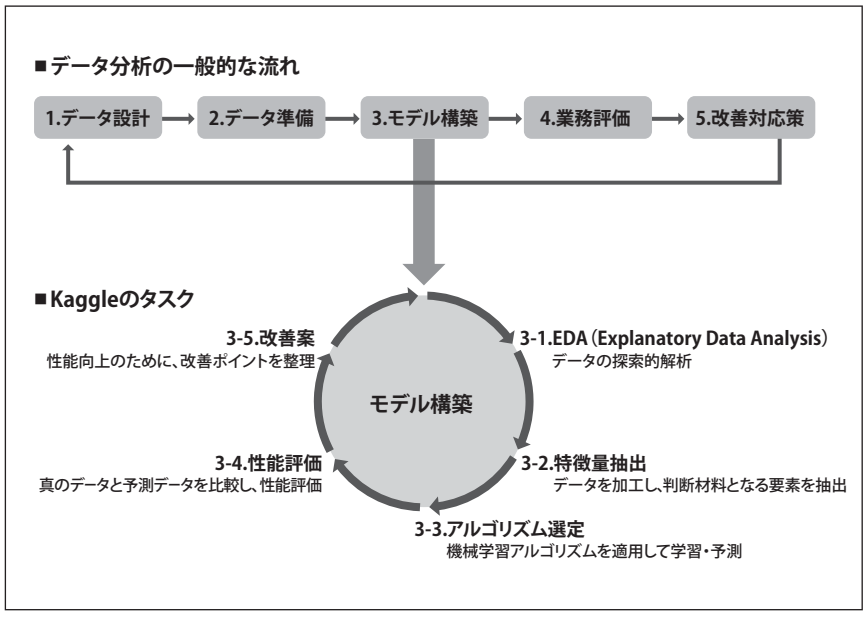

図1. データ分析におけるKaggleのタスク

3. 今後のチャレンジ

今後は、世界トップクラスの専門家としてのKaggle Grand Masterを目指したいと思います。また、Kaggleだけではなく、さらにスキルを向上させて、世界で著名な機械学習・AIの学会で論文発表を行うことで、最先端の知見を応用するだけでなく、自らも新たな知見を生み出すことにチャレンジしていきたいです。

そして、自身の専門性の向上だけでなく、仲間やお客様を巻き込み、協力していくことがとても重要だと考えています。Kaggleで経験してきたように、自分一人では困難なことも、チームの仲間と協力することでより多くの知見が集まり、乗り越えることができます。世界中にあるIBMの研究所やお客様とも協力し、最大限の力を発揮することで、テクノロジーや社会の発展に貢献していきたいと思っています。

日本アイ・ビー・エム株式会社
東京基礎研究所
先進保険ソリューション
研究員

岩森 俊哉
Toshiya Iwamori

2014年日本IBM入社。入社後、金融のデリバリーやデータ・サイエンティストとして業務分析を経験。2018年8月より東京基礎研究所に異動し、現在は医療分野における機械学習の研究に携わっている。