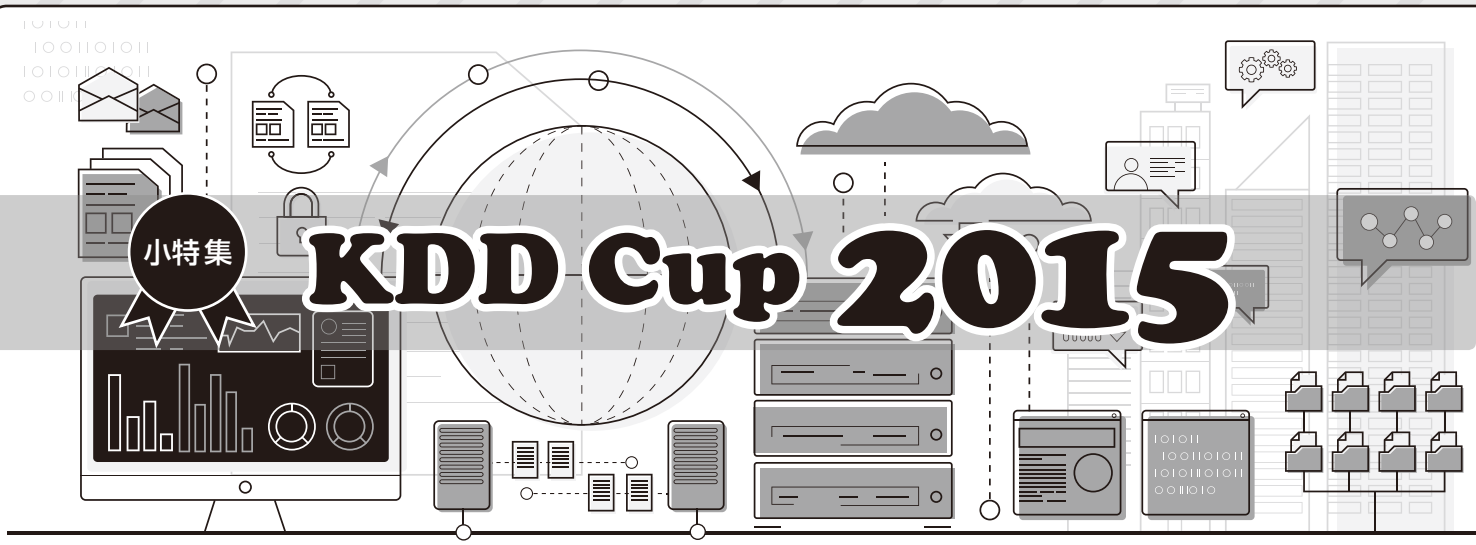


## 編集にあたって

油井　誠（トレジャーデータ（株））

KDD Cup（Knowledge Discovery and Data Mining Cup）☆1 は ACM（Association for Computing Machinery）の分科会である SIGKDD（Special Interest Group on Knowledge Discovery and Data Mining）が毎年開催する知識発見とデータマイニングに関する競技会である．

1997 年に第 1 回が開催されて以降，毎年異なるさまざまな課題に対して，大学や企業の研究者を始め，企業の第一線で活躍しているデータ分析のプロフェッショナル，はたまた世にその名をとどろかせたいと思っているベンチャー企業の技術者や個人まで多くの参加者がデータ分析技能を競い合っている．企業からの注目度も高く，例年，日本企業からもデータ分析のプロフェッショナルが自己研鑽や会社の技術力のアピールを目的とした参加がある．毎年，時代を反映した課題とスポンサー企業から現実のデータが提供されるのが特徴である．

第 19 回目の開催となる KDD Cup 2015 には過去最多の 821 チーム（1,263 人）が参加する中，日本人の参加する 4 チームが入賞を果たした．

本特集では，始めに本大会の課題，ルールおよび予測結果の評価方法を解説し，次の記事から上位に入賞した 4 チームの本大会での取り組み内容と各チームが用いた特徴的な分析手法を解説する．

1 位の Intercontinental Ensemble チームは優勝の要となったスタッキングと呼ばれるメタ特徴量の生成手法と国際混成チームでアンサンブルモデルを作り上げる上で役立ったチームガイドラインを紹介する．次に，2 位の企業合同参加の FEG&NSSOL@Veraci チームが特に注力した特徴量設計とチーム戦でのモチベーション維持方法，役割分担について述べる．6 位の KDDILABS&Keiku チームはユーザ間の類似度や ARIMA と呼ばれる時系列予測モデルを特徴量に利用したことを報告する．最後に，10 位に学生参加の kyazuki&DT@Keio Univ. Ohmori Lab チームが特徴量設計とモデル構築手法に加え，学生の視点からデータ分析コンペティションに参加する意義を述べる．

KDD Cup 2015 では，データ分析コンペサイト「Kaggle」☆2 の上位ランク保持者がタッグを組み，各個人が生成した予測結果や予測モデルを組み合わせることで優勝を果たしたが，ほかの上位入賞者も複数人の学習モデルを統合している点が興味深い．いわゆる「**三人寄れば文殊の知恵**」が実践されていると言える．しばしばデータ分析コンペでは，「**過学習することが勝つ秘訣だ**」と皮肉をこめて語られるが，一方で過学習を避けるアンサンブル学習などの取り組みがいかに重要か分かる．

本特集で入賞者が利用したアンサンブル学習，Gradient Boosting Decision Tree（決定木を利用した勾配ブーティング），スタッキング，各種特徴量設計手法やライブラリ情報は機械学習を現実問題に適応

☆1 http://kdd.org/kdd-cup

☆2 https://kaggle.com/

する上で役立つヒントとなる．入賞者が惜しみなく各種データ分析技法を解説する本特集記事が，データ分析や機械学習をビジネスや現実の問題に導入しようとする読者の役に立つことを願ってやまない．

## KDD Cup 2015 について

今年度（KDD Cup 2015）の課題は，**「オンライン学習講座（Massive Open Online Course（MOOC））の受講者の離脱確率を予測する」**というものである．データセットには，中国の e ラーニングサイト「XuetangX」から実受講ログデータが提供された．競技参加者は MOOC の履修ごとの脱落確率を予測し，その精度を競う．

### 提供されたデータと Dropout の定義

e ラーニングサイトの各講座それぞれの開講後 30 日分の受講ログデータが提供された．提供データの構成は次のとおりである．

- 講座数：39 講座
- ユーザ数：112,448 人
- 履修数：200,905（うち，訓練用が 120,543，提出用が 80,362）

ログデータは Enrollment（講座ごとのユーザの履修）単位で生成されていて，開講後の 31 日目から 40 日目の 10 日間で履修ごとに受講によるアクセスが発生しない（これを Dropout と呼称する）確率を予測する．**図 -1** に Dropout および Enrollment の概念を掲示する．

### 予測結果の提出までの流れ

競技参加者は，まず訓練用のデータから脱落予測のモデルを組み立て，その予測モデルを利用して評価用のデータから脱落確率を導く．そして，履修ごとの脱落予測確率を大会のサイトに提出する．各参加チームには 1 日あたり 5 回までの予測結果の提出が許されている．

開催期間中は，検証データの一部を用いて予測の精度に応じた暫定スコアが発表される．参加者は暫定スコアを参考にして予測精度の改善を行う．ただし，最終的な正式スコアの算出では別の検証用データセットが利用されるため，暫定スコアに過剰適合するのではなく，汎化能力（未知のデータに対する予測できる能力）の高い学習モデルを構築することが求められる．

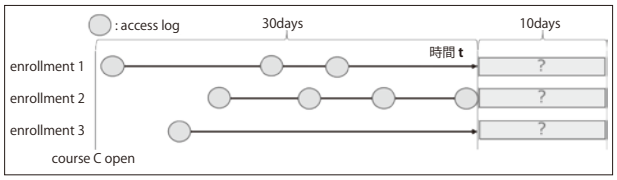

図 -1　Dropout 予測の概念

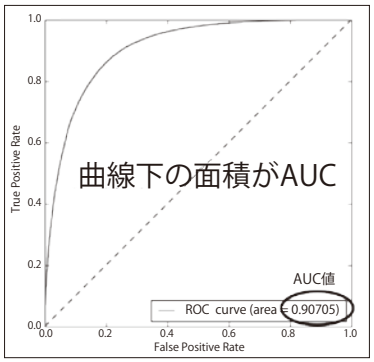

図 -2　Area Under Curve

### 評価方式

予測結果には履修ごとの Dropout のありなしではではなく，履修ごとの解約確率を提出する．そして，提出データは Area Under Curve（AUC）と呼ばれる指標で評価される．

AUC は，**図 -2** のように，横軸に偽陽性（False Positive）の割合，縦軸に真陽性（True Positive）の割合をとった Receiver Operating Characteristic（ROC）曲線下の面積で分類器の性能を示す．直感的には，二値分類で真の値が正のものと負のものをランダムに 1 つずつ選んだときに，正の値の予測解約率が負の値の予測解約率以上になる確率を持って分類器の性能を示すものである．偽陽性を低くすると真陽性も低くなりがちであり，偽陽性を高くすると真陽性も高くなる．AUC は 0 から 1 までの値をとり，完全な分類が可能なときの面積は 1 で，ランダムな分類の場合は 0.5 になる．

本大会では，上位入賞者は 0.90 以上であり，優勝した Intercontinental Ensemble チームの AUC は 0.9091817339587759 で 2 位のチームとの差はわずか 0.00031856399 である．なお，スコアの差は軽微であるが，ビジネスではこの軽微の差が大きな収益の差を生むことがあることに留意されたい．

（2015 年 10 月 31 日）