CF-009

変数の分布に着目した特徴量の追加とアンサンブル学習による予測モデル高精度化の検討
A Study of High Accuracy of Predictive Models by Using Ensemble Learning and the Addition of Features Focusing on the Distribution of Data Variables.

高田 晋太郎† 鯨井 俊宏†
Shintaro Takada Toshihiro Kujirai

1. はじめに

IT や NW 技術、クラウドコンピューティング、モバイル端末、ソーシャルメディア、センサ技術などの発展に伴い、世の中で発生、収集できるデジタルデータは増加し続けている。こうした中、近年ではあらゆる事象をデータとして捉え、AI を適用し種々の業務指標の改善につながる施策の提案等を行う AI ソリューションビジネスが活発である[1]。AI を活用した業務改善施策の立案を行うためには、改善対象とする業務に関して蓄積されたデータをもとにした、事象のモデル化、つまり機械学習による業務指標の予測モデルの生成が必要不可欠である。対象とする業務指標の予測モデルを得ることで、モデルを用いた将来的な指標の予測や、様々な条件下でのシミュレーションを重ねることによって指標を改善するための条件を発見する等、業務指標改善の施策立案を実行することが可能となる。

一般的に機械学習によって予測モデルを得るためには、1) 対象とする事象のデータ取得、2) 事象を適切に表現することが可能な特徴量（説明変数）の設計と 3) 外れ値や欠損値等を含むサンプルに対する適切な対処を経て、学習データの準備を行い、4) 学習データに適した機械学習アルゴリズムの選定と学習時条件の調整（ハイパーパラメータ調整など）、等の工程を要する。学習データを準備するまでの工程 1) ～ 3)においては、データ分析の知識に加え、対象とする業務の性質や特殊性等を理解し、それらを考慮した特徴量の設計が重要であり、これら両方を実行可能な人材が必要となる。また、業務指標改善施策の効果を高めるためには、できるだけ精度の高い予測モデルを得られることが望ましく、4) の工程については、モデル作成者自身の機械学習の知識や、網羅的な学習時条件の探索を可能とする潤沢な計算機リソースなどが求められる。その一方で、AI を用いたソリューションビジネスを様々な業種等に幅広く迅速に展開するには、できるだけ少ないリソースで高精度な予測モデルを生成できることが要件となっており、これを実現する予測モデル学習方式の確立が求められている。

これまでの研究において、2) 特徴量の設計や 4) 機械学習適用時の条件調整等の工程を、機械によって自動的に処理する AutoML が開発されてきた。AutoML では、人間による様々な条件での試行錯誤を計算機側で網羅的に行うことで、最適な条件を見つけ出し人的工数をほとんどかけずに、高精度なモデルを得ることができる手段として、盛んに研究がなされている[2][3][4][5]。その一方で、現実世界における様々な事象において、あらかじめ定められた条件における網羅的な探索のみで、対象の事象を高い精度でモデル化することは難しい。特に 2) 特徴量の設計において、AutoML では定型的な変換処理は自動で行うことができるが、より高い精度のモデルを得るためには、人間による業務特有の性質や特殊性を考慮した設計が重要となる[2]。

本研究では、AI を用いたソリューションビジネスの迅速な展開を実現するため、機械学習による予測モデル生成において自動かつ高い精度を得られる予測モデル学習方式の確立を目的とし、学習データが持つ各説明変数の分布に着目して作成された特徴量を、効果的に予測モデルの精度向上に利用可能なアンサンブル学習方式を提案する。

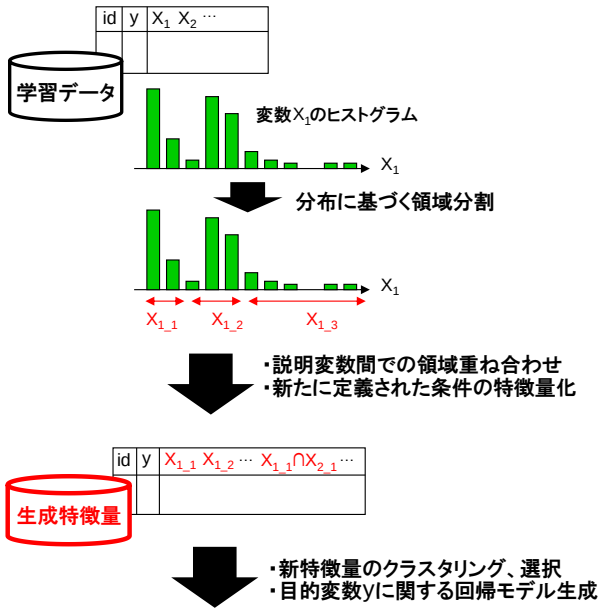
†(株)日立製作所　研究開発グループ　先端 AI イノベーションセンタ Hitachi, Ltd. Research & Development Group, Center for Technology Innovation – Advanced Artificial Intelligence.

2. 従来研究

これまで、機械学習による予測モデル生成において、1 章で述べた各工程の自動化並びに、モデルの高精度化に関する研究が盛んに行われてきた。代表的な取り組みとして、高い精度の予測モデルを得るためのタスクを、特徴量生成や機械学習アルゴリズム選択、ハイパーパラメータ調整等のモデル生成に必要な一連の工程を、各工程における条件組み合わせの最適解を求める CASH 問題(Combined Algorithm Selection and Hyperparameter optimization problem)として捉えた AutoML の研究が挙げられる[2]。AutoML の機能を備えたツールやフレームワークは多数開発、公開がされており使用することが可能である[3][4]。

工程 2) 特徴量設計の自動化に関する取り組みも行われており、最も一般的な設計方法は、事前に定義された処理演算子をもとに特徴量を生成するものである。例えば、単一の変数に対して標準化や、ログスケール変換等規定の変換処理を適用するもの、複数の変数の値を種々の演算子によって組み合わせて新しい特徴量とするもの、また複数の変数値に対する統計量値を新たな特徴量とするものなど、様々なものが存在する[5]。これらの設計方法は事前に特徴量設計として有効であると認知され、AutoML における条件候補として用意されたものである。一方、実際の特徴量設計の工程においては、学習データの特性を見た上で有効な特徴量を設計することで、そのデータにより適した特徴量を生成することができる場合があるが、この工程は人間によるデータ分析と試行錯誤が必要であり、自動化することは非常に難しい。これに対し、工藤らは、目的変数に関連が高い隠れた変数条件を分析するために、各説明変数の値を複数の条件で区切った領域を定義し、かつ他の説明変数の領域との重ね合わせた条件を作りだし、その中から目的変数と相関が高いものを自動で抽出する方法について言及している[6]。本方法の基本的な考え方を図 1 に示す。まず、学習データが持つ各説明変数について、何らかの基準に基づき、その変数が取り得る領域を複数の領域に区切る。本図では例として、その変数における学習データサンプルの分布（ヒストグラム）に基づき、各領域に同数のサンプ

ルが収まるよう 3 個の領域（変数条件）に区切った場合を示している。このような処理を全説明変数において実施し、さらに、説明変数間での各領域を組み合わせた変数条件も作成する。そして、定義された各変数条件を新たな特徴量とみなし、特徴量化を行う。具体的には、各変数条件に各データサンプルが一致するかどうかの二値(0/1)を持つ変数として特徴量を生成する。次に、生成された新特徴量を一つのデータとして見立て、目的変数 y に対する線形回帰モデルの生成を行う（クラスタリング、特徴量選択を含む）。このように生成された回帰モデルの係数を調べることで、どのような変数条件が目的変数と関連が高いのか、分析することが可能となる。本手法は前述のように、目的変数と関連した隠れた変数条件を発見、分析するためには有効な手段となり得るが、新たに生成した特徴量は元の説明変数に比べ情報量が劣化しているため、本方法によって生成される回帰モデル自体の予測精度は高くないと考えられる。



$$y = 0.71X_{1\_1} + \ldots + 0.33X_{1\_1} \cap X_{2\_1} + 0.21$$

図 1 変数の分布に基づく特徴量生成と分析

工程 4) 学習データに適した学習条件の調整においては、一般的なパラメータ最適化に加え、複数の機械学習モデルを組み合わせたアンサンブル学習による予測精度向上がよく知られている[2]。アンサンブル学習は、組み合わせを行うモデル間の多様性に依って、片方のモデルでは考慮できなかった学習の観点を、他方のモデルで補うことで、最終的な予測精度を向上する。特に代表的な方法である Stacking と呼ばれるアンサンブル学習方式は、学習データに対して複数生成された予測モデルの各予測値を、新たな特徴量とみなし、新しい階層での予測モデルを再度学習する手法である。

3. 本研究の目的と方針

前述の通り、本研究では AI を用いたソリューションビジネスの迅速な展開を実現するため、機械学習による予測モデル生成において自動かつ高い精度を得られる予測モデル学習方式の確立を目的とする。2 章で述べたように、AutoML 等の従来研究の取り組みによって、機械学習モデル生成のための多くの工程を自動で実施できる一方で、学習データの特性に沿った特徴量の設計やそれを用いた高精度な予測モデルの生成を自動で行うことは、現状では難しい。

そこで、本研究では[6]における学習データの分布に応じた特徴量の設計方法の考え方をもとに、生成された新たな特徴量を予測精度向上に活用可能な学習方式の開発に焦点をあて、研究に着手した。この理由は下記の通りである

- データの特性に合わせた特徴量設計は、これまで人間が担ってきた領域であり、モデルの高精度化の余地が残されていると考えられる
- 特徴量を新たに追加することで、精度向上の要因が把握しやすい
- 特徴量設計のバリエーションを増やすことに繋がり、AutoML 等の技術との組み合わせが可能である

学習方式を開発するにあたって、まず、予測精度の向上が見込める仮説を立て、それに基づき提案手法を考案した。そして、実データによる予測精度の評価実験を行い仮説の評価、検証を行った。以降は、それぞれについて詳細を述べる。

4. 学習方式の提案

4.1 予測精度向上の仮説

予測精度向上に繋がる仮説として、以下を考えた

仮説 1：生成した特徴量の追加によるモデル精度の向上

図 1 に示した特徴量の設計と分析方法は、一般的な相関分析等による目的変数と関連が高い説明変数の分析と比べ、より局所的な条件において、目的変数と関連が高い事象の有無を発見することが可能となる。このように発見された条件は、通常のデータ分析からでは見つけにくく、同様に一般的な機械学習アルゴリズムにおける学習の過程においても、考慮がなされないような条件となっている場合が考えられる。例えば、決定木モデルの学習過程において、説明変数における分岐条件の決定は、常に目的変数との関係性を基に進められていく。一方、本方法での特徴量生成においては、一旦は目的変数との関係性は考慮せずに、説明変数自体の分布による条件決定が行われる。このような処理によって生成された特徴量は、通常の機械学習アルゴリズムでは考慮しきれなかった条件を表現し得ることが期待され、予測精度向上に寄与すると仮説立てをした。

仮説 2：多様性の更なる付加による精度向上

2 章において述べた通り、アンサンブル学習は複数の予測モデル間で生じる多様性に基づいて、それらを統合することで予測精度を向上する学習方式である。上記、仮説 1 のように、新たに生成した特徴量が、機械学習アルゴリズムで考慮しきれていなかった条件を表現しているような場合に、その条件を持つ学習データとそうでない学習データで、多様性が生じるのではないかと考えた。具体的には、

元の学習データに対して異なる複数のアルゴリズムを用いる一般的なアンサンブル学習よりも、新たな特徴量を含んだ学習データとそうでない元の学習データのそれぞれで生成されたモデルを統合するアンサンブル学習を行うことで、よりモデル間の多様性が高まり、予測精度向上に寄与するのではないかと考えた。

4.2 提案学習方式

前節で述べた各仮説に基づく提案学習方式を図 2 に示す。本提案学習方式は、2 つのステップから構成される。

Step1:追加特徴量生成ステップ

まず、学習対象となる学習データ（元学習データ）に対して図 1 で示した考えに基づいた追加特徴量を生成する。元学習データの各説明変数におけるデータの分布に基づき、複数の領域（変数条件）を定義し、さらに、異なる説明変数間での各領域を組み合わせた変数条件も併せて作成する。作成した各変数条件に各データサンプルが一致するかどうかの二値(0/1)を持つ変数として特徴量を生成し、目的変数との相関に基づき、相関の高い一定数の特徴量を追加特徴量のデータとする。このように、採用する特徴量の数を一定の規則で制限することで、追加特徴量の数が肥大化することを防ぐ。

Step2:アンサンブル学習ステップ

次に、アンサンブル学習時に、学習に用いるデータの段階で多様性を生じさせるために、データ統合を行う。具体的には、元学習データの説明変数に前ステップにて作成した追加特徴量変数を追加した新たな学習データ（元学習データ＋追加特徴量）を作成する。このようにして、得られた新たな学習データは、元学習データと比べて追加された特徴量の分、異なる情報を持っているため、それぞれの学習データから作成される各予測モデルには、より多くの多様性が生じることが予想される。このようにして、得られた元学習データ＋追加特徴量の学習データと、元学習データそれぞれに対し、予測モデルを生成し、作成された各予測モデルをアンサンブル学習によって統合する。

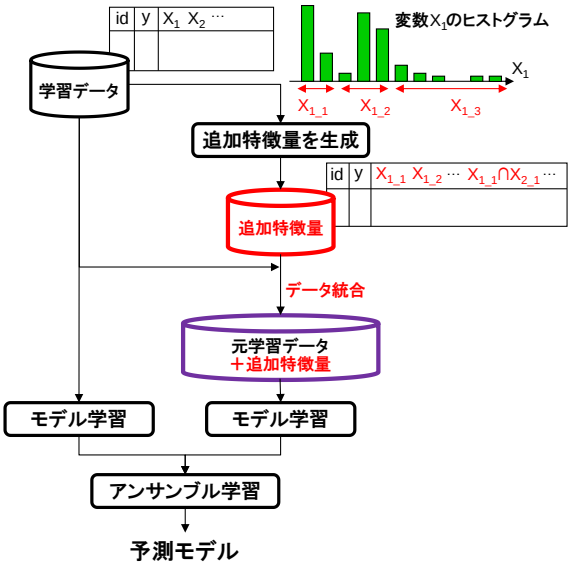


図 2 提案方式の処理

5. 実データによる予測精度評価実験

本章では、実際のデータを用いた提案学習方式の予測精度評価実験について述べる。本実験では、一般的な予測精度による性能評価に加え、前章で立てた二つの仮説に対応する以下の二つの観点について評価を行った。

観点 1：新たに生成した特徴量の有効性評価

前章で述べた特徴量の生成が、予測精度に貢献するかどうかの評価を行った。特に、従来の学習方式で頭打ちとなっていた精度が本特徴量の追加によって向上するか、と実際に予測精度が向上している場合にどのような特徴量が精度向上に寄与しているか、の観点で評価と分析を行った。

観点 2：多様性によるアンサンブル学習時の精度向上評価

学習に用いる学習データの段階から多様性を生じさせ、そこから生成された予測モデルをアンサンブル学習することで、従来のアンサンブル学習と比較し予測精度が向上するかどうか、の観点で評価と分析を行った。

5.1 実験条件

本実験では、現実の事象を対象としたデータに対し、予測精度の評価を行った。これによって、より実践的な変数の分布を持ったデータに対しての有効性を評価可能である。使用したデータは、ある物流倉庫内での作業員による荷物ピッキング作業時間の予測に関するものである。学習データのサンプル数は 14766 個、説明変数は 104 個である。説明変数は例えば「ピッキング対象物の重さ」等、ピッキング作業時間に影響し得る要素のものであり、事前に人間が定義を行ったものである。提案方式における追加特徴量の生成にあたっては、各説明変数の領域を分割後のサンプル数が等しくなるよう 5 分割し、各変数間での各領域の組み合わせは 2 つまでを考慮した。生成した特徴量のうち、目的変数と相関の高い上位 100 個のものを追加特徴量とした。また、生成したモデルの予測精度の評価には、MAE (Mean Absolute Error)と RMSE (Root Mean Squared Error)の二種類の予測誤差の指標を用いた。学習時に用いる機械学習アルゴリズムは Elastic Net (EN) 、Random Forest (RF)、LightGBM (LGBM)の 3 種類を用いた。各アルゴリズムとも学習データの説明変数に対する変数選択効果があり、特徴量が多数追加され変数が多くなってしまうデータにおいても、学習アルゴリズム側での有効な変数の選択を期待できる点と、特に RF と LGBM はデータサイエンティストの間でも広く使用され、高いモデル精度を期待できる点を考慮した。また、アンサンブル学習による複数のモデルの統合には Stacking 法を用い、アルゴリズムには EN を使用した。これは、アンサンブル学習時に複雑なモデルを用いてしまうと、どのような関係性で予測モデルが統合されたのかが、分かりにくくなり、最終的なモデルの解釈性が落ちてしまうのをできるだけ防ぐためである。また、同様の理由でアンサンブル学習の階層は、1 段目に各アルゴリズムでの予測モデル生成を行い、それらの出力(予測値)を統合する 2 段目のモデルを学習する単純な構造を用いた。

5.2 評価 1: 生成特徴量の有効性評価

新たに生成した特徴量が予測精度の観点で、精度向上に寄与するかどうかの評価を行った。比較の対象として、元の学習データに一つの機械学習アルゴリズムを適用した場合と、同様に元の学習データに異なる複数のアルゴリズム

を適用したアンサンブル学習を適用した場合もそれぞれ評価した。本評価結果を表 1 に示す。まず、機械学習アルゴリズム単一のものを用いた場合で、追加特徴量の有無と元学習データに追加特徴量を追加した場合による比較(#1 – 3、#7 – 9 と#10 - 12 の比較)を行った。追加特徴量のみを用いた場合では、RF と LGBM での学習において元学習データを用いた場合に比べ精度が落ちてしまっていることが確認できる(#8, #9)。一方で、追加特徴量を元学習データに追加した場合においては、EN 及び LGBM で両精度が向上、RF はほぼ同等の結果であることが確認された(#10 - 12)。特に単一アルゴリズムで最も精度が良い LGBM を用いた場合において、特徴量の追加によって精度が向上していることを確認できた(MAE が 112.48 から 112.15 に、RMSE が 177.84 から 177.10 に改善)。さらに、アンサンブル学習を行った場合における比較(#4 – 6 と#13 – 15 の比較)をすると、LGBM と EN 及び LGBM と RF を組み合わせた場合において、特徴量の追加によって精度が向上することが確認できた。本評価全体で、従来方法の中で最も高精度であった条件(#6: 元学習データに RF と LGBM でアンサンブル学習)と比較し、元学習データに新たに生成した特徴量を追加したデータを用いることで(#15)、MAE が 112.08 から 118.83 に、RMSE が 176.77 から 176.24 にそれぞれ改善することができた。このことから、特徴量の生成と追加によって、従来の方式から予測精度を向上させる効果があることを確認できたと言える。

表 1　評価 1 評価結果

#	データ	学習方法		MAE	RMSE
1	元学習データ	単一	EN	246.66	677.26
2			RF	113.70	178.12
3			LGBM	112.48	177.84
4		アンサンブル	EN, RF	113.80	178.10
5			EN, LGBM	112.48	177.78
6			RF, LGBM	112.08	176.77
7	追加特徴量のみ	単一	EN	148.60	226.48
8			RF	131.41	203.88
9			LGBM	127.42	198.38
10	元学習データ + 追加特徴量	単一	EN	138.44	235.66
11			RF	113.88	178.24
12			LGBM	112.15	177.10
13		アンサンブル	EN, RF	113.71	178.16
14			EN, LGBM	112.13	177.07
15			RF, LGBM	**118.83**	**176.24**

次に、具体的にどのような追加特徴量が精度向上に寄与しているのかについて調べた。同じ機械学習アルゴリズムにおいて、その特徴量の追加によって精度向上が見られた追加特徴量をピックアップし、それについて詳細に分析した。図 3 に、精度向上に寄与した追加特徴量の中で、特に特徴的であった追加特徴量 X_{Z_1} について、その元となった説明変数 X_Z の分布とともに図示する。上段は、元となった説明変数 X_Z と目的変数 y の散布図を、下段は説明変数 X_Z のヒストグラムをそれぞれ示し、点線で示した説明変数 X_Z の範囲が、追加特徴量 X_{Z_1} で定義される変数条件である。まず、説明変数 X_Z の性質として値が 6000 付近にサンプルが多く分布していることが下段のヒストグラムから分かる。また、目的変数 y との関係性においては、説明変数の値が 5000 手前付近までにおいては、y の値が 0 付近に分布し、5000 付近以降では y 値が 2000 – 3000 付近の大きな値を取った後、ゆるやかに小さくなっていくという特殊な分布であることが分かる。これは、モデル化対象の事象において、本説明変数の値が 5000 付近において何らかのルールや制約等が存在しておりそれを境に目的変数の傾向が変わるという現実世界の事象における特殊性を強く反映したデータであると言える。このような分布を持つ説明変数に対し、予測精度向上に寄与した追加特徴量の変数条件は 0 から 5720 の範囲で定義されるものであった。この条件は前述の通り、この分布の特殊性（5000 付近手前までは y が 0 に近い）を表現するのに適した特徴量であると言え、この変数が精度向上に寄与したことは妥当と考えられる。以上のことから、本方式で追加された特徴量は、前述のように特に現実の事象で見られるような特殊な分布を持つデータに対して、その特殊性をうまく表現できた場合に、予測精度向上に寄与することを期待できる方法であると言える。

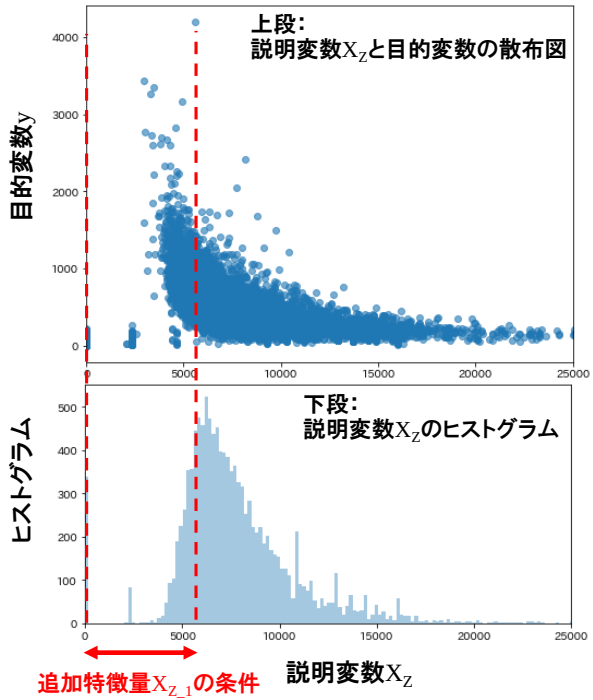


図 3　精度向上に寄与した特徴量の分析

5.3 評価 2：多様性付加による効果の評価

本評価では、学習データレベルでの多様性を付加したアンサンブル学習が、予測精度向上に繋がるかどうかの観点で評価と分析を行った。表 1 で示した結果から、本データにおいては RF と LGBM の組み合わせのアンサンブル学習が最も精度が高かったことが分かる。本評価では、これに該当する#6 と#15 の結果に加え、多様性を付加した条件で学習した結果を比較する。評価結果を表 2 に示す。# 16 と #17 の結果が、片側のデータを元学習データ、他方のデータを元学習データに特徴量を追加したデータとして学習した場合で、それぞれに RF と LGBM のどちらかを適用した場合の結果である。精度結果として、元学習データに RF

を、特徴量追加データに LGBM を適用しアンサンブル学習を行った#16 の条件の精度が最も高かった。しかしながら、学習データを元学習データ＋追加特徴量データを共通して用いてアンサンブル学習を行った#15 の結果と、精度はほぼ同等であった。

表 2　評価 2 評価結果

#	アンサンブル学習条件				MAE	RMSE
	Data 1	Algo.1	Data 2	Algo. 2		
6	元学習データ	RF	元学習データ	LGBM	112.08	176.77
15	元学習データ＋追加特徴量	RF	元学習データ＋追加特徴量	LGBM	**118.83**	**176.24**
16	元学習データ	RF	元学習データ＋追加特徴量	LGBM	**111.82**	**176.22**
17	元学習データ＋追加特徴量	RF	元学習データ	LGBM	112.07	176.75

この結果について、詳細に分析する。#15 と#16 の違いは、RF を適用する学習データの違いであるが、この違いが予測精度には影響していないように考えられる。表 1 の #2 と #11 の比較からも分かるように、RF を用いた学習においては、追加特徴量が予測精度向上に寄与していないことが分かる。これを確かめるため、RF によって生成された予測モデルにおける特徴量変数の重要度を調べた。その結果を図 4 に示す。本図は学習時に用いた学習データの特徴量変数の変数 No.とそれに対応した重要度がグラフ化されている。上段が元の学習データを用いた場合、下段が特徴量を追加した場合の重要度のグラフである。特徴量追加によって追加された特徴量の番号である No.105～No.204 の変数において重要度がほとんど 0 であり、本モデルで考慮された特徴量が上段と下段でほぼ同じであったことが分かる。

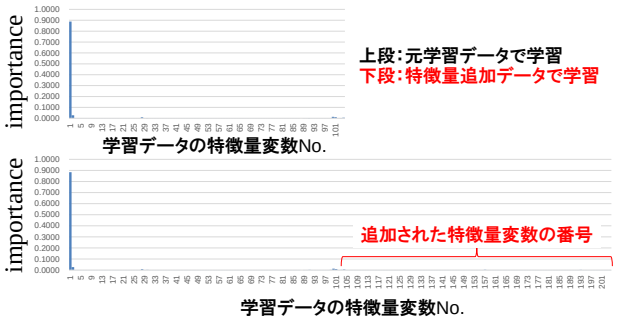


図 4 RF による予測モデルの重要度

一方で、同様に LGBM によって生成された予測モデルの特徴量変数の重要度を図 5 に示す。LGBM の場合、新たに追加された特徴量が僅かながらではあるが、予測モデルで考慮されていることが確認できる。このように新たに考慮された特徴量が予測精度の向上に寄与しているものと考えられる。

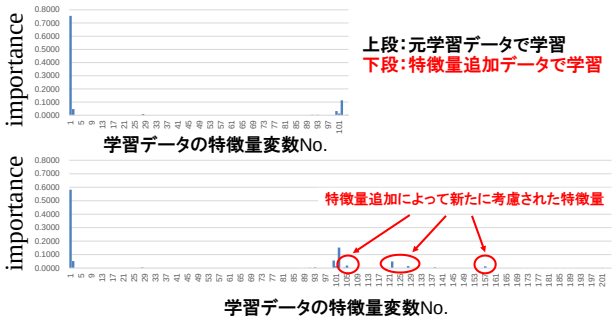


図 5 LGBM による予測モデルの重要度

上記の分析結果から、今回の実験で使用したデータにおいては、RF による学習において特徴量追加データを用いた学習は、実質的には元学習データと同じ予測モデルが生成されているものであることが分かる。このことは、表 2 の#15 における学習は実質的に#16 と同様の学習が行われていると考えることができる。したがって、#15 及び#16 の学習条件によって最も精度が高い予測モデルを得られている要因として、元学習データに対して RF を適用した予測モデルと、特徴量追加データに対して LGBM を追加した予測モデルを組み合わせたことが該当するものと考えられる。つまり、提案学習方式である新たな特徴量の追加と学習データレベルでの多様性の付加を行った学習が、更なる予測精度向上に効果があることを、本評価によって確かめることができたと言える。

5.4 汎用性に関する追加評価と考察

前節までの二つの評価によって、提案学習方式が従来の学習方式に比べ、予測精度の更なる向上を期待できることが確認できた。本節では、提案方式の有効性が他のデータ適用時においても汎用的に期待できるかどうかの追加評価と考察を行った。評価に用いたデータは、上記のデータと同様に物流倉庫内における荷物ピッキング作業時間の予測に関するものであるが、場所が異なる倉庫のものでありデータの性質、分布等は異なる。学習データのサンプル数は 12133 個、特徴量数は 100 個である。追加特徴量の生成方法は同様であり、目的変数と相関の高い上位 100 個のものを追加特徴量とした。

本データにおける予測精度の評価結果を表 3 に示す。まず、ベースラインの精度として、元学習データに単一のアルゴリズム適用した場合では LGBM が最も高い精度が得られた(#20)。同様に、元学習データに一般的なアンサンブル学習を適用した場合は、LGBM と EN の組み合わせが最も高い精度を得た(#27)。これは、本データの場合においては、LGBM と EN によって生成された予測モデル間の多様性が、統合による精度向上に寄与しているものと考えられる。次に、提案方式による効果の確認として、まず、特徴量追加データによる学習結果においては、LGBM、RF を単独で適用した場合は、ともに精度向上効果は見られなかった(#25, #26)。一方で、EN に関しては MAE が 113.33 から 106.51 に改善しており、改善の効果が見られた(#18→#24)。このことから、本データにおいては、LGBM と RF では追加された特徴量が精度向上に寄与しなかったが、EN においては

効果的に寄与している結果となった。次に、元学習データと特徴量追加データの両方を用いた提案方式のアンサンブル学習時においては、LGBM と EN の組み合わせが、全条件の中で最も高い精度が得られることが確認できた(#29, #30)。ここで、#30 においては特徴量追加データに LGBM を適用した条件となっているが、前述の通り、LGBM を用いた学習は追加特徴量が効果的に寄与していないため、実質的に元学習データを使った条件#29 と同等であると言える。このことから、今回データにおける評価においても、#29 の条件である元学習データに LGBM を、特徴量追加データに EN をそれぞれ適用する学習データのレベルで多様性を付加したアンサンブル学習によって、最も高い精度が得られることが分かった。

表 3 別データにおける精度評価結果

#	学習条件				MAE	RMSE
	Data 1	Algo.1	Data 2	Algo. 2		
18	**元学習データ**	EN	-	-	113.33	224.33
19	**元学習データ**	RF	-	-	101.67	148.06
20	**元学習データ**	LGBM	-	-	100.46	147.32
21	**追加特徴量**	EN	-	-	114.52	185.60
22	**追加特徴量**	RF	-	-	110.50	174.89
23	**追加特徴量**	LGBM	-	-	109.17	172.54
24	**元学習データ ＋追加特徴量**	EN	-	-	106.51	231.61
25	**元学習データ ＋追加特徴量**	RF	-	-	101.73	147.96
26	**元学習データ ＋追加特徴量**	LGBM	-	-	100.67	147.21
27	**元学習データ**	LGBM	**元学習データ**	EN	100.00	146.75
28	**元学習データ**	LGBM	**元学習データ**	RF	100.53	148.60
29	**元学習データ**	LGBM	**元学習データ ＋追加特徴量**	EN	**99.55**	**145.48**
30	**元学習データ ＋追加特徴量**	LGBM	**元学習データ ＋追加特徴量**	EN	**99.75**	**145.45**

本評価結果から、別の学習データの場合においても、提案方式の有効性が確認できた。二つのデータに対して提案方式を適用した結果について以下のように考察した。

・新たな特徴量の追加による効果

特徴量の生成と元学習データへの追加によって、予測精度の向上を期待できることが分かった。一方で、データとアルゴリズムの組み合わせによっては、精度向上の効果がない場合も考えられる。効果がある場合は、図 3 で示したように、データの分布の特異性をうまく表現できていて、かつアルゴリズムが学習時にその特徴量を効果的に考慮できている場合である。このような性質は、AutoML 等、網羅的に特徴量の生成とアルゴリズムとの組み合わせを試行するような枠組みと親和性があり、最適な組み合わせを発見することで、より高い確度で予測精度向上が期待できる方式であると言える。

・データレベルでの多様性付加による精度向上効果

新たに生成した特徴量を追加したものと、そうでない元の学習データの二つの学習データを用いて、より多様性を持った予測モデル間でのアンサンブル学習を行うことで、高い予測精度が得られることを確認した。今回の実験で用いた二種類の評価データにおいては、片側を元の学習データ、もう片方を特徴量追加データでそれぞれ予測モデルを生成しアンサンブル学習を行った場合が、最も高い精度が得られる結果となった。一方で、本方式で精度向上が見られるのは、特徴量追加データによって精度向上があった場合に付随して更なる精度向上が見込める手法とも言える。したがって、精度向上が見込めるデータとアルゴリズムの組み合わせ発見が重要となるため、こちらも前述の通り、AutoML 等の枠組みと併せて用いることで、より精度向上効果を見込めると考えられる。

6. おわりに

本研究では、機械学習による予測モデル生成において、自動かつ高い精度を得られる予測モデルの学習方式確立を目的として、学習対象のデータが持つ分布に応じて生成された特徴量と元の学習データとを用いて作成した新たなデータと、元の学習データの２通りの学習データに対して、異なるアルゴリズムを適用しアンサンブル学習を行う学習方式について、提案を行い、実データを用いた評価実験を行った。本評価実験によって、新たに生成した特徴量が予測精度向上に寄与すること、またその特徴量を活用した提案方式のアンサンブル学習が、従来のアンサンブル学習に比べ更なる予測精度向上の効果が見込めることを確認した。本実験における分析と考察においては、学習データと新しく生成した特徴量、及び学習データに適用するアルゴリズムとの組み合わせに依存して、精度向上が期待できるという提案方式の特性は、AutoML 等における最適学習条件を探索する枠組みと親和性が高く、これらを組み合わせることで、より高い効果を期待できることを述べた。

今後は、本提案学習方式を種々のデータに適用し、予測精度向上効果の更なる検証を行うとともに、効果が高いデータの特性や性質などについて明らかにしていくことを予定している。また、新たに生成した特徴量の追加と提案方式のアンサンブル学習を AutoML の枠組みに乗せ、組み合わせ探索のバリエーションを増やすことで、より高い予測精度向上効果をめざす予定である。

参考文献

[1] 難波 康晴, 吉田 順, 徳永 和朗, 原口 拓也, “AI のサービスと基盤業務の高度化への実践的アプローチ”, 日立評論, 2016 年 4 月号, pp.37-40 (2016).
[2] Marc-André Zöller, Marco F. Huber, “Benchmark and Survey of Automated Machine Learning Frameworks,” Journal of Artificial Intelligence Research, Vol.70, pp.409 - 472 (2021)
[3] Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., & Hutter, F., “Practical Automated Machine Learning for the AutoML Challenge 2018,” International Conference on Machine Learning AutoML Workshop, (2018)
[4] R. S., & Moore, J. H., “TPOT : A Tree-based Pipeline Optimization Tool for Automating Machine Learning,” International Conference on Machine Learning AutoML Workshop, pp.66 – 74 (2016)
[5] Sondhi, P., “Feature Construction Methods: A Survey,” Sifaka. Cs. Uiuc. Edu, 69, 70-71 (2009)
[6] F. Kudo, T. Akitomi, N. Moriwaki, “An Artificial Intelligence Computer System for Analysis of Social-Infrastructure Data” IEEE 17th Conference on Business Informatics, pp.85 - 89 (2015).