１Ｅ０４

日本における機械学習の人材育成の課題と分業化の提案

山本　雄介（北陸先端大）、内平　直志（北陸先端大）

1. はじめに

日本において機械学習やAIが注目を集め、企業におけるDigital Transformation（DX）の重要性が言われている。日本では特に機械学習やAIを担う人材不足が問題で、企業内研修、民間、大学による教育が盛んになってきている。ただしこれらの人材育成プログラムではクラスタリング、深層学習など問題を解く手法の教育が中心となっており、実際の機械学習プロダクトを構築する上で必要不可欠な、その他の工程を担う人材を育成する視点が欠けている。本発表ではデータ環境構築の専門家であるデータエンジニアと、ビジネス課題から解くべき機械学習の問題を定義し開発者と意思疎通をする分析トランスレーターの重要性を提案する。

2. AI・機械学習に必要な人材

日本ではAIを担う人材不足が言われている。例えばみずほ情報総研株式会社[1]では、2018年時点で3.4万人、2030年には14.5万人不足するとの試算をしている。

ただし、求められるAI人材というのは、どういう専門性を持つ人材なのか、どのような種類の人材なのか、というはっきりと合意されたものはない。上記レポートでは「AI人材に関する明確な定義はない」[1]とAI人材の定義がないことを認めている。ダベンポート[2]は「組織の分析力を支える要素」として、Data、Enterprise、Leadership、Target、Analystの５つの要素を挙げ、「DELTA」モデルを提唱しているが、「データなしに分析をする事はできない」とデータの重要性は議論しているものの、誰がどうやるのかということには触れていない。

機械学習のモデルを作る機械学習エンジニアやデータサイエンティストが必要であるという認識は広まっている。ダベンポート[3]は必要とされる人材として「データサイエンティスト」と説明し、データサイエンティストに求められる知識、能力として、プログラミング、ビッグデータ技術、科学的根拠に基づく意思決定、コミュニケーション能力、意思決定のスコープを設定する能力、統計分析手法、機械学習に関する知識、画像など非構造化データの解析能力、データ分析をどこに当てはめるのが理想か察知する能力、などをあげている。

このように、求められるAI人材は曖昧でありつつも、機械学習のモデルを作る機械学習エンジニアやデータ分析を行うデータサイエンティストが必要という認識は広がっている。また、３章で後述するように、大学や民間機関によるAI教育はモデル手法を学ぶものが中心になっている。

しかし筆者の実務上の経験から、機械学習やAIを生かす組織を作るには、それだけでは不十分と感じる。もっと別の役割を持った人材が必要なのではないか、特にまだ認識が広まっていないデータエンジニアと分析トランスレーターが重要なのではないか、と本章では考察する。

2-1: AI・機械学習に必要な人材：データエンジニア

データエンジニアとは、分析環境構築、データ保存のためのテーブルのデザイン、データを加工し分析できるようにする一連のプロセスを構築、自動化、モニタリングの仕組みを整備する役割を持っている。

Sculleyらによれば、機械学習システムにおいてモデルの学習や予測に費やされる割合というのはごくわずかに過ぎない[4]。図2-1の通り、データの収集、特徴量の抽出、データの検証、分析ツールの準備、プロセス管理のツール、機器の管理、サーバーなどのインフラの整備、モデルのモニタリングなど、現実の機械学習のシステムでは機械学習モデル以外に多くの要素が複雑に絡み合っており[22]、モデルだけ作れば機械学習のシステムが完成するということはまずない。

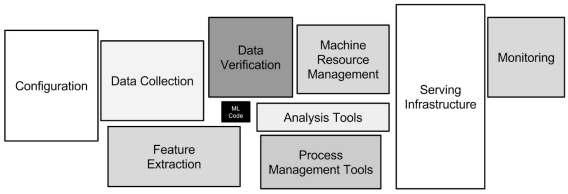

図2-1: 中心黒塗りになっている小さな四角が”ML Code”と、機械学習システムの中のほんの一部を占めているに過ぎない[4]。

また、Rogati[5]はAIや機械学習を導入しようとする企業や組織でよくあるケースとして、「最も基本的なデータサイエンスのアルゴリズムやオペレーションを実装し、その

利点を享受するためのインフラを整備していない」と指摘する。図2-2のように第一段階として「機器、ログ、センサー、外部データ、ユーザーによって生み出されたコンテンツ」などのデータを収集する段階がある。そして第二段階として、「安定したデータの流れ、インフラ、パイプライン、データの抽出・加工・格納、構造化、非構造化データの保存」といったデータを移動・格納する段階がある。これら第一、第二段階はデータエンジニアがいなければ効率的にできず、機械学習モデルを作るスキルとは全く異なるスキルを要する。MLモデル構築や深層学習をする”Learn/Optimize”という段階は第５段階以降である。

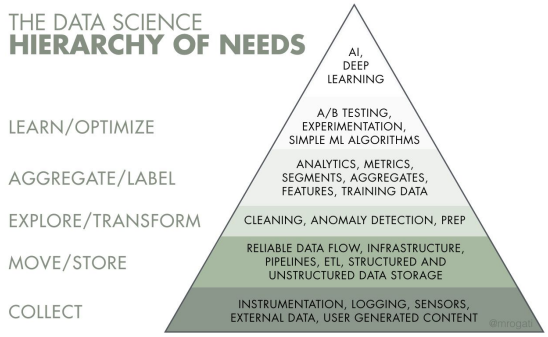

図2-2: AI活用に至るまでの段階の図例[23]

このように、機械学習やAIのアルゴリズムを開発したり、モデルの構築をする以前に、データ環境を構築するというプロセスが必須であり、データエンジニアがいなければ効率的にこのプロセスが完了できないことがわかる。

2-2: AI・機械学習に必要な人材：分析トランスレーター

Henkeら[6]によれば、分析トランスレーターの役割は、データサイエンティストや機械学習の専門家の持つ専門知識を、マーケティング、サプライチェーン、製造、リスク管理などのビジネス側で持っている知識、経験、課題を結びつけることである。結果的に、機械学習などの分析手法から得られた知見を会社・組織の運営に活かすという役割を持つ。

例えば不正注文を検知する機械学習モデルを実装するようなケースでの分析トランスレーターの役割を見てみる。まずは不正注文が問題になっている、という事実を見つけるところから始める。ユーザーから頼んでいないのにクレジットカードに請求があったとか、人気の商品が買い占められてしまって欲しいのに買えない、というような苦情があるという話を、企業の販売部門などから聞きつけるところから始まる。

そして、不正注文はおそらく注文データを使えば機械学習を使って自動的に発見ができるのではないか、というところまで推測を立て、具体的にいくらぐらいの被害が出ているのかを大まかに見積もる。解決できれば年間いくらのコスト削減につながり、大きな機会があると分かれば、データサイエンティストなど機械学習の専門家に異常検知、データ分類に関わるデータ分析課題があると話を通す。

実際にモデルが出来上がれば、モデルの精度をビジネス側にわかる形で伝える。モデルを評価するArea Under Curve（AUC）などの指標はビジネス側の意思決定には役立たない。そこで分析トランスレーターが、開発したモデルを適用すれば、一月あたり何個の不正注文を発見でき、いくらのコスト削減が見込め、かつ間違ってユーザーの正しい注文をキャンセルしてしまう回数は月にたかだか何件である、というような説明をし、モデルを実際に使うのかどうか、意思決定をする手助けをする。

このように分析トランスレーターは自分でモデルを構築するということはない。にも関わらず、分析トランスレーターの役割がなければそもそもの課題を見つけてくることができず、課題を機械学習の問題に置き換えることもできず、モデル構築後にビジネス側へインパクトをわかりやすい言葉で説明もできず、機械学習モデルを導入するという意思決定までつながらない。

2-3: AI・機械学習に必要な人材の分業

このように、モデル構築を行うデータサイエンティストや機械学習エンジニアとは別に、データ環境を構築するデータエンジニアと、問題の定義、要件の整理・説明を行う分析トランスレータの３つの役割が、機械学習やAIの製品開発には必要であることがわかる。

また、これらの役割は関係するところもあるものの、全く別の専門性を要し、それぞれを別々の人材が担当する分業の発想が必要である。

3. 日米におけるAI・機械学習人材育成の比較分析

ところが、日本では大学、民間教育ともに、モデル構築手法に教育が偏っており、データエンジニア、分析トランスレーターを養成しようという試みがない。

データサイエンティスト協会のスキルシート[7]でも、データサイエンス、データエンジニアリング、ビジネスと３つのスキルを提言しているが、分業を示唆していない。異なる専門性を持ったデータサイエンティストと呼ばれる人材が必要との印象を受ける。

アメリカにおいても、大学では日本同様モデル構築手法に教育が偏っているが、GAFAなどAIを活用する企業からの要望に応える形でできたMassive Online Open Course (MOOC)では、データエンジニア、分析トランスレーター(PM)を養成するプログラムが別に存在する

3-2 日本におけるAI教育

日本においては民間のブートキャンプのような短期教育機関においても、大学のような公的教育機関においても、機械学習のモデル構築といった手法の教育に偏っている。データの加工に関する教育はあるものの、データをどう言った形で保存するのか、データをどうやってモデルに渡すのか、データを加工する環境をどう構築するのか、といったデータに関する教育はほぼない。

以下では日本の公的機関と民間機関におけるAI教育を分析する。

3-2-1 滋賀大学データサイエンス学部

滋賀大学データサイエンス学部のカリキュラム[8][9]は専門教育科目として表3-1のようになっている。

この中では、データエンジニアリング科目が６科目と極端に少ない。最多科目数を持つデータサイエンス専門科目は、テキストマイニング、多変量解析、機械学習、時系列解析といったデータはある前提でモデルを作るという科目になっており、データを加工する、環境を構築する、分析しやすい構造にデータベースを構築する、という教育はあまりされていないことがわかる。

科目グループ	科目数
データエンジニアリング科目	6
データアナリシス系科目	9
データ解析科目	2
データサイエンス専門科目	44
価値創造基礎科目	22
価値創造応用科目	22
データ駆動型PBL演習科目	6

表3-1: 滋賀大学データサイエンス学部における科目群ごとの科目数

3-2-2 北海道大学数理・データサイエンス教育プログラム

北海道大学でも2019年より学部プログラムを開始している。カリキュラム[10][11]を見ると、一般科目として、情報学、統計学、線形代数、微積分があり、さらに基礎科目で重回帰分析、画像解析などの数理分野、生物系の統計力学など生命分野の科目、行動計量学、教育、心理学などの社会分野の科目がある。このうち、データエンジニアリングに関わっていそうな科目は「データベースとWebインテリジェンス」の１科目のみである。

3-2-3 民間企業によるAI教育

大学だけでなく、民間の会社が提供するAI教育プログラムもモデル構築に集中している。

株式会社キカガクが提供する「自走できるAI人材になるための６ヶ月長期コース」[12]は事前準備として微分、単回帰分析、Python入門、線形代数などの基礎科目、1-4週目は機械学習、深層学習、Webスクレイピング、アプリ開発、5-8週目は環境構築、画像処理、時系列・自然言語処理、アプリ開発、9週目以降はアプリ開発となっている。Dockerの基礎、サーバ連携といったデータエンジニアリングに関わるような科目があるが、ほとんどはモデル開発のトレーニングとなっている。

エッジテクノロジー株式会社が提供するAI JobColle[13]では機械学習講座、統計+R講座、ディープラーニング講座、ケーススタディ実演講座といった同じようにモデルを作ることを学ぶ講座が中心になっている。

このように、日本の公的機関におけるAI教育は、モデル構築といった手法に関わる教育に特化していることがわかる。また、民間のAI教育においても、一部データエンジニアリングに関わる講座が含まれているものの、やはり中心はモデル構築といった手法に関わる教育となっていることがわかる。

3-3 米国におけるAI教育

米国におけるAI教育も、大学では日本同様モデル構築など手法に関わる教育が中心になっている。一方MOOCでの民間が提供する教育では、企業がカリキュラム開発に携わっていることもあり、より実践的で機械学習、AI開発の周辺分野の教育も提供されている。

3-3-1 University of Denver, Data Science

University of Denverの提供するデータサイエンスコース[14]では、必須科目15科目のうちデータエンジニアに関わるところはDatabase Organization & Management I、Parallel and Distributed Computingの２科目に限られる。残りはプログラミング、数学、確率統計といった基礎科目の6科目と、機械学習やデータ可視化などデータ分析の実務的な科目の7科目がある。

データ分析や機械学習の実践的なモデル作りが７科目、そのための基礎科目が６科目あるのに対し、データベース管理、並列処理のデータエンジニアリング系の科目が２つしかなく、モデル作りや統計手法に科目が偏っているのがわかる。

3-3-2 UC Barkley, The Online Master of Information and Data Science

UC BarkleyのThe Online Master of Information and Data Science[15]においても、カリキュラムは機械学習、統計分析が中心になっている。

カリキュラムには、”The core curriculum focuses on the following key skills”として、Research Design、Data Cleansing、Data Engineering、Data Mining and Exploring、Data Visualization、Information Ethics and Privacy、Statistical Analysis、Machine Learningの８つを挙げている。Data Engineeringがその１つに入っていることから、なんらかのデータエンジニアリング関係の科目が含まれていると思われるが、データ前処理、データマイニング、機械学習、データ可視化、統計分析といったデータが備わっていてそれをどうモデル化するか、というスキルが多い。データエンジニアを養成する、という構成にはなっていない。

3-4 米国民間企業（MOOC）によるAI教育

米国の教育で特色があるのが、より企業が求める人材ニーズに沿った実務的なコースを提供していると思われる、MOOCの教育内容である。Udacity、Coursera、UdemyといったMOOCが米国にはいくつか存在するが、いずれにおいてもデータエンジニアに特化したコースや、PM向けのコースが存在する。例えばUdacityにおいてはGoogleと一緒に開発しているコースもある[17]など、企業と一緒にコースを開発しているMOOCも存在し、よりAIを推進している企業が求めているスキルが反映されていると考えられる。

3-4-1 Udacity Data Engineering nano-degree

Data Modeling (create relational and NoSQL data models), Cloud Data Warehouses, Spark and Data Lakes, Data Pipelines with Airflow, など、単なるデータ加工だけではなく、構造データ、非構造データを保存するデータベースをどのように構築するか、それらデータベースをクラウド上でどう作るか、Sparkといった分散処理、Data Lakeというデータ保存方法、データ加工のプロセス造りとその自動化に関する項目など、これ自体として専門のコースになっており、データサイエンティストが同時にやるようなボリュームではないことがわかる[17]。

UdacityではAI Product Manager[18]というPM向けのコースも存在する。機械学習モデルのトレーニング、評価方法といった機械学習の基礎自体も学ぶと同時に、AI製品のビジネス的価値をどう評価するか、製品が市場で受け入れられ、拡大するために何をすべきか、ということを学ぶ。PMと名前がついている通り、機械学習の開発を統括するのは、データサイエンティストや機械学習のエンジニアとは別の人が行う、という発想がある。

3-4-2 Udemy

Udemy[19]でもBig Data Engineerというカテゴリーで14のコース（2020年9月現在）が提供されている。主な内容は、構造化データを扱うMySQLなどのデータベース、非構造化データを扱うNoSQL、Hadoop、MongoDBといった技術、ビッグデータを並列処理で高速に加工するSparkといった技術が学べる構成になっており、これだけでデータの専門家をトレーニングする、という内容になっている。また、”Practical Project Management for Machine Learning Projects”[20]というPM向けのコースが存在する。ビッグデータ、AI、IoTを活用するプロジェクトが増えてきており、これらのプロジェクトを進行する必要性が出てきたと説明している（Udemy 2020）。

3-5 日米におけるAI・機械学習人材育成のまとめ

このように民間企業が提供するMOOCでは、日本でよく見られるようなデータサイエンス、機械学習のトレーニングも提供されているものの、よりデータに特化したようなコースや、機械学習をソフトウェアなどのサービスにどう活用し、どのように開発を進めていくのか、というようなPM向けのコースが存在する。

いずれにおいても、データエンジニアリングと機械学習プロジェクトのプロジェクトマネジメントには、これまでとは違う別のスキルが必要、という発想があると思われる。これらのMOOCは個人が自分のキャリア開発や転職支援のために自費を払って参加したり、企業が自社社員の教育のために投資するためのコースとなっているため、より企業が求める人材のスキルが反映されると考えられ、データエンジニア、PM for Machine Learningとデータサイエンスや機械学習とは別のコースになっている、という点が米国におけるAI製品、機械学習を生かしたサービス、製品開発の成功の一端ではないかと推察される。

4. 考察と提言

以上、AI・機械学習を生かす組織にはモデルを構築する機械学習エンジニアやデータサイエンティストだけでなく、データ環境の整備をするデータエンジニアと、問題の発見や定義を行う分析トランスレーターが重要であることを考察した。また、日本では人材の育成がモデル構築ができる人材に偏りがあることを、日米の大学、民間教育を比較し分析した。

日本では人材不足が叫ばれているものの、これまで日本の教育機関において、民間、公的機関ともにデータ分析や機械学習に用いられる手法のトレーニングにばかり重点がおかれていることを、日米の教育機関のカリキュラムを比較し、検証した。機械学習やAI開発が進んでいるアメリカにおいては、データ環境を整備するデータエンジニア、ビジネス課題から解くべき課題を定義する分析トランスレー

ターの役割が重要視されつつあり、特にアメリカのMOOCで提供される教育にはそのような分業化が進んでいることが観察された。

また、実際の機械学習を組み入れたシステムやサービスにおいて、実例（不正注文検知）を用いてモデル作りの要素が問題解決の一部でしかないことを確認した。
そもそもどのような課題を解くべきなのか、ビジネス側と開発側の間に立って機械学習の問題に定義し直す役割を持つ分析トランスレーターと、機械学習モデルを開発する大前提として、データ環境を整備したり、モデル開発をしやすい環境を整えるデータエンジニアの重要性を詳細に検証した。

本論考では、実際の企業や組織における事例やインタビュー調査等はできていない。今後の課題として、機械学習やAIを活用できているいくつかの企業、組織において、どのような役割を持った人材がいるのか、それぞれが機械学習の活用にどのように貢献しているのか、分析トランスレーターとデータエンジニアの役割を持つ人材はいるか、というより具体的な調査が必要と思われる。

参考資料

1. みずほ情報総研株式会社, 2019, 「平成 30 年度我が国におけるデータ駆動型社会に係る基盤整備（IT 人材等育成支援のための調査分析事業）－ IT 人材需給に関する調査 －調査報告書」, (2020/06/30取得, https://www.meti.go.jp/policy/it_policy/jinzai/houkokusyo.pdf)
2. トーマス・H・ダベンポート, ジェーン・G・ハリス, ロバート・モリソン, 2011, 「分析力を駆使する企業発展の五段階」日経BP
3. トーマス・H・ダベンポート, 2014, 「データ・アナリティクス3.0 ビッグデータ超先進企業の挑戦」日経BP
4. Sculley etal., 2015, “Hidden Technical Debt in Machine Learning Systems”, (2020/08/20取得, https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf)
5. Monica Rogati, 2017, “The AI Hierarchy of Needs”, (2020/06/30取得, https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007)
6. Nicolaus Henke , Jordan Levine and Paul McInerney, 2018, “You Don’t Have to Be a Data Scientist to Fill This Must-Have Analytics Role”, Harvard Business Review, (2020/06/30取得, https://hbr.org/2018/02/you-dont-have-to-be-a-data-scientist-to-fill-this-must-have-analytics-role)
7. 一般社団法人データサイエンティスト協会、2020 「データサイエンティストスキルチェックリスト ver3.01」
8. 滋賀大学データサイエンス学部カリキュラム, (2020/06/30取得, https://www.ds.shiga-u.ac.jp/about/ds/curriculum/)
9. 滋賀大学データサイエンス学部授業科目ナンバリング一覧表, (2020/06/30 取得, https://www.ds.shiga-u.ac.jp/ds_ms_2018/wp-content/uploads/2019/08/H30データサイエンス学部のナンバリングについて（HP用）.pdf)
10. 北海道大学数理・データサイエンスセンターカリキュラム, (2020/06/30取得, https://www.mdsc.hokudai.ac.jp/curriculum/)
11. 北海道大学数理・データサイエンスセンター専門教育科目一覧, (2020/06/30取得, https://www.mdsc.hokudai.ac.jp/mc-major/)
12. 株式会社キカガク, (2020/06/30 取得, https://www.kikagaku.co.jp/seminars/longterm/)
13. AI JobColle, (2020/06/30取得, https://www.aijobcolle.com/)
14. University of Denver, Data Science, (2020/06/30取得, https://ritchieonline.du.edu/data-science/curriculum/course-descriptions/)
15. UC Barkley, The Online Master of Information and Data Science, (2020/06/30取得, https://datascience.berkeley.edu/)
16. The Economist, 2017, “Established education providers v new contenders”, Jan 12th 2017 edition,
17. Udacity, Data Engineering nano-degree, (2020/06/30取得, https://www.udacity.com/course/data-engineer-nanodegree--nd027)
18. Udacity, AI Product Manager, (2020/09/05取得, https://www.udacity.com/course/ai-product-manager-nanodegree--nd088)
19. Udemy, (2020/09/05取得, https://www.udemy.com/user/big-data-testing/)
20. Udemy, Practical project management for machine learning projects, (2020/09/05取得, https://www.udemy.com/course/practical-project-management-for-machine-learning-projects/)