【研究ノート】

# ChatGPT のデータ解析系授業への影響について

三宅茂樹

## Impact of ChatGPT on Data Analysis Courses

Shigeki Miyake

**要旨：**この研究ノートは、ChatGPT をデータ解析系授業に導入することの影響に焦点を当てている。主に、テーブルデータと画像データの演習問題に ChatGPT を使用し、その効果と限界を探求している。テーブルデータの問題では、ChatGPT がデータ整形からモデル評価までのプロセスを効率化し、平均的な精度のモデルを容易に構築できることが示されている。しかし、精度を向上させるには、データ解析の基本的な知識が不可欠である。画像データの問題では、ChatGPT の実行環境の制限によるボトルネックと、深層学習の基本的な知識の重要性が明らかになっている。全体として、ChatGPT は教育プロセスを効率化する可能性があるが、短期的には従来の教育方法の重要性は変わらない。一方、中長期的には授業の内容やアプローチの再評価が必要である。

### 1. 背景

#### 1.1 ChatGPT について

ChatGPT は OpenAI によって開発された先進的な自然言語処理モデルである。人間のように流暢かつ自然な言語でのコミュニケーションが可能で、膨大なテキストデータから学習を行っている。このシステムは GPT（Generative Pre-trained Transformer）技術に基づいており、大規模なデータセットからパターンを学習し、新しいテキスト生成に応用する。ChatGPT は質問への回答、テキスト生成、データ解析の課題解決など、幅広い用途に利用されている。特に、コンピュータプログラミングやデータサイエンス分野での応用が注目され、学生や専門家にとって重要なリソースとなっている。このモデルの主な特徴は柔軟性と広範な応用可能性にあり、日常のコミュニケーション、学問的な探求、技術的な問題解決において新たな可能性を提供している。

#### 1.2 データ解析系授業への影響について

従来の授業設計では、最新の技術進化、特に ChatGPT のような高度な自然言語処理モデルの使用を想定していない。現代の技術環境において、このようなツールの導入は、学生が直面する実際の課題と授業内容の間にギャップを生じさせる。実際に ChatGPT を授業に用いた際の実践的報告が出始めているが、これらは主にプログラミングや語学教育に

関連している [1][2][3]。データサイエンス系授業での具体的な実践例の報告はまだ少ない。[4] はデータ解析系授業に ChatGPT を導入した実践例といえるが、主に学生アンケート結果の集計について記述していて、導入の影響分析、対処法・改善案などの提案には至っていない。

ChatGPT には外部からデータを与えることでそのデータを解析する機能 [1] があり、コーディングまで行うことが可能であるが、これは従来の授業設計の想定外である。そのため、授業内容や演習問題の調整に関しても、ChatGPT のようなツールを適切に組み込む方法については明確なガイドラインが存在しない。ChatGPT を授業に導入した場合のメリットやデメリットが未だ明確ではないが、教育者は授業の目的と手段を再評価し、どのような影響があり得るかを明らかにしながら、学生にとって最も有益な学習体験を提供するための新たなアプローチを模索する必要がある。

### 1.3 本研究ノートの役割

本研究ノートの主な役割は、ChatGPT を用いて典型的な演習問題に対して学生の立場でアプローチを行い、その過程で得られた知見を通じて、ChatGPT の可能性と限界を明確にすることにある。具体的には、ChatGPT がデータ解析の分野でどのように利用できるか、その機能と効果を詳細に分析する。また、学生が ChatGPT を使用する際のメリットとデメリットを検証し、教育的観点からその意義を探求する。さらに、授業への影響と、ChatGPT の導入に伴う教育方法の適応や改善策を提案する。なお本稿で使用した ChatGPT のバージョンは ChatGPT4 である。ChatGPT3.5 は無料で利用できるため、現時点で学生が最も使用しているバージョンであると予想できるが、大規模言語モデルの性能は時間が経てば各段に上がっていくため、本稿での調査や考察の急激な陳腐化を防ぐために敢えて 3.5 ではなく 4 を使用したことを注意しておく。

この研究ノートは、AI と教育の融合が進む現代において、教育者と学生の両方にとって重要な指針を提供することを目的としている。ChatGPT の持つ技術的な特性と教育への応用を深く理解することで、今後のデータ解析教育のあり方についての議論を深め、より効果的な学習方法の開発に寄与することを期待している。

## 2. ChatGPT を使った演習問題へのアプローチ

第 2 章では、ChatGPT のデータ解析系授業への影響を探るために、具体的な演習問題に焦点を当てる。特に ChatGPT を用いて典型的なデータ解析の演習問題を学生目線でのアプローチを行う。選択した問題は、データ解析コンペティションサイトである SIGNATE[5] の練習問題から 2 題を取り上げた。SIGNATE はデータ解析のコンペを提供するサイトであり、その練習問題も実践的なデータ解析スキルの習得に役立つものが多い。本章では、これらの問題に対してどのように ChatGPT を使ってアプローチしたのかを記述する。特に、従来のアプローチと ChatGPT を用いたアプローチとの比較に重点を置き、その相違点や効果を分析する。この分析を通じて、ChatGPT がデータ解析の学習プロセスにどのように影響を与え、従来の教育方法にどのような新たな視点をもたらすのかを探求する。さらに、ChatGPT の使用が演習問題の解決方法にどのような変化をもたらすかについても詳

しく調査し、データ解析教育への具体的な応用可能性について考察する。

## 2.1 自動車の走行距離予測 [6]

### 2.1.1 問題概要

SIGNATE の練習問題「自動車の走行距離予測」は、典型的なテーブルデータを用いた問題であり、自動車のさまざまな属性からその燃費性能（1 ガロンあたりの走行距離）を予測することが目的である。この問題では、車種ごとの重量、排気量、馬力などの物理的特性に加え、年式や製造国などのデータが提供される。参加者はこれらの情報を基に、燃費性能を正確に予測するモデルを構築する。特に、本問題はデータ解析の基本的なスキルを試すのに適しており、回帰分析など機械学習アルゴリズムの理解と応用が中心となる。さらに重要な点は、データの処理や分析がすべて ChatGPT の実行環境上で行えることである。これにより、従来のデータ解析方法と比較して、ChatGPT を用いたアプローチがどのように異なるかを評価する絶好の機会を提供する。

### 2.1.2 ChatGPT を使ったアプローチ

まず、提供されたテーブルデータ（csv 形式）を ChatGPT に pandas dataframe として読み込ませ、以後データ解析の標準手順（データ整形、探索的データ分析（EDA）、モデル選択・学習、評価）に従った。

データ整形の段階では、欠損値の処理方法を ChatGPT に提案してもらい、その提案に基づいて欠損箇所を埋めた。次に、EDA では ChatGPT が目的変数と説明変数との相関をグラフ化し、データの関係性を明確に示した。この視覚化はデータの理解を深めるのに役立つが、従来はコードを手書きしながら行うのでコードのバグ取りもしながらデータの理解も進めなければならなかった。こういった状況はデータ解析を始めたばかりの者にとってはそれなりに難関であった。ChatGPT を使えばデータの理解に集中できるのが大きなメリットである。

EDA の結果を踏まえて、ChatGPT は回帰モデルを提案し、そのモデルで学習と評価を行った。また、ChatGPT はテストデータに対する予測実行から、予測結果を SIGNATE の投稿形式で保存するまで自然言語で指示するだけで完結するのも初心者には助かる。得られた結果は順位的に中程度またはやや上位であり、ベースラインとしては十分な結果である。これまでのプロセスには 10 分余りしかかからなかった。

ここから重要な点は、初回サイクルで得られたベースラインの結果からモデルをどのように改良していくか、である。この段階では利用者のデータ解析に関する知識（特徴エンジニアリング、学習モデルの選択、ハイパーパラメータのチューニングなど）が不可欠となる。ChatGPT は改良の方針を提案してくれるが、どの提案を選択するかは、利用者の経験と知識に依存する。

今回は学習モデルとして SVM を採用し、正則化係数とカーネルについてハイパーパラメータのチューニングを行った。ハイパーパラメータのチューニングにはこちらから指示しなくても scikit-learn の Grid SearchCV を使用していた。ハイパーパラメータのチューニングが完了した時点で投稿した結果は上位の 1 割程度であった。ここまでにかかった時間

はほぼ 1 時間であり、大変効率的にデータ解析のサイクルを回せていることに気づく。

このように、ChatGPT はデータ解析のプロセスを大幅に効率化するが、最終的な判断や精度の改良には人間側の専門知識が重要な役割を果たすことが明らかとなった。

### 2.2 画像ラベリング（20 種類）[7]

#### 2.2.1 問題概要

SIGNATE の練習問題「画像ラベリング（20 種類）」は、前述の「自動車の走行距離予測」とは異なり、画像データを対象とする深層学習モデルの構築を目指す問題である。この課題では、動物、乗り物、自然の風景など 20 種類の異なるカテゴリーに属する画像が提供され、参加者はこれらを適切に分類するモデルを開発する。この問題は、テーブルデータを扱う前問題と異なり、コンピュータビジョンと画像処理の基礎知識が中心となり、特に畳み込みニューラルネットワーク（CNN）などの深層学習アルゴリズムの理解と応用が重要である。

この深層学習の問題は学習時に高い計算能力を要求し、特に GPU の使用が必須となる点が特筆すべきところである。ChatGPT 内の実行環境では GPU を使用できないため、Google Colaboratory などの別の実行環境を用意し、ChatGPT が提供するコードを実行する必要がある。提供された画像データセットを用いてモデルをトレーニングし、新たな画像に対する分類精度を高めることが求められる。この問題は、実際の画像分類の課題に近く、深層学習技術の応用能力を測る機会を提供する。テーブルデータと画像データ、それぞれの問題において必要とされるデータ解析のアプローチの違いを明確に示している。

#### 2.2.2 ChatGPT を使ったアプローチ

「画像ラベリング（20 種類）」問題への ChatGPT を使ったアプローチでは、まず訓練データをアップロードしたディレクトリを指定し、そこからデータを読み込むように指示した。しかし、最初に実行したコードでは画像データのサイズが大きすぎてメモリがクラッシュする問題に直面した。これに対処するため、画像データをバッチごとにメモリに読み込むよう ChatGPT に指示したところ、ジェネレータを用いたコードが回答された。このジェネレータを使用するために、訓練データが保存されているディレクトリ内にラベルに対応するサブディレクトリを作成する必要があったが ChatGPT の回答に従ってこれを容易に実施できた。

学習モデルに関しては、最初から CNN を構築するのではなく、既存のモデルをファインチューニングする方針を取った。ちなみに ChatGPT の提案は VGG16 であった。モデルの学習後、テストデータに対する予測を行い、SIGNATE の投稿形式で保存するプロセスも前の問題と同様に実施した。

このようにベースラインの学習モデルまでは前問同様に容易に構築可能である。前問同様であるが本質的な問題はベースラインからどのようにモデルを改良していくかにある。今回は、data augmentation（データの水増し）を行うように指示した。正確に述べると、ベースライン構築のコードはすでに data augmentation を行う設定だったが初回のサイクルではオーバーフィッティングを起こすことが重要なので敢えてこれを行わなかった。

さらに、オーバーフィッティングを緩和する方法に関しても ChatGPT より複数の回答

を得た。しかし、授業における演習を想定しているため、時間のかかるハイパーパラメータのチューニングまでは行わなかった。

結局、ChatGPT はデータの準備からモデルの評価に至るまでのプロセスを効率化する一方で、ベースラインからさらに精度を改善するためには深層学習の基本的な理解が依然として重要であることが理解できた。

## 3. 考察と結論

第 2 章で実施した SIGNATE 練習問題へのアプローチを通じて得られた重要な知見は、ChatGPT を活用することでデータ解析のプロセスが効率化される一方で、より高い精度を求める場合には従来同様専門的な知識が必須であることである。テーブルデータの解析では、ChatGPT の実行環境上でデータ整形から学習モデルの評価までの一連のプロセスを迅速に完了することが可能であった。これにより、従来の手法に比べてデータ解析のサイクルを格段に速く進めることができる。従ってグループワークの形式で問題に取り組む際には、作業単位での役割分担の意義が薄れ、代わりにデータ理解やアイデアの交換がより重要な役割を果たすようになると考えられる。

一方で、GPU を要する深層学習の問題においては、ChatGPT の実行環境だけでは不十分であり、Google Colaboratory などの外部環境との連携が必要となる。従って、学習時間が作業のボトルネックとなる点は従来と変わらない。テーブルデータの問題と同様に、コードやアルゴリズムに詳しくない場合でも平均レベルのアウトプットを容易に得ることができるが、アウトプットの精度をさらに高めるためには深層学習の深い理解が必要となる。

これらの知見は、ChatGPT の利用がデータ解析の教育や実務に与える影響を考える上で重要である。一方で効率化が進み、もう一方で高度な成果を得るためには、従来と同様に深い専門知識が求められる。このバランスを理解し、適切に対応することが、今後のデータ解析教育や実務の発展において鍵となる。

### 3.1 メリットとデメリット

学生が ChatGPT を活用して演習問題に取り組むことには、明確なメリットとデメリットが存在する。メリットとしては、まずコーディングに関する手間が軽減され、分析そのものに集中できる点が挙げられる。特にテーブルデータの解析など、ChatGPT の実行環境内で完結できるケースでは、データ解析のサイクルを大幅に速めることが可能となる。これにより、データの深い理解や学習モデルの改良に向けたアイデアの深堀りにより多くの時間を割くことができる。

しかし、一方でデメリットも存在する。ChatGPT を用いることで、自然言語で指示を出すだけで適切なコードが返ってくるため、学生がコーディングスキルを習得する機会が減少する可能性がある。これは、コードを記述する技術やプログラミングの論理的思考を養う過程が疎かになるということを意味する。プログラミングスキルはデータ解析だけでなく、様々な分野での問題解決や新たなアイデアの実現に不可欠であるため、このスキルの習得は重要である。

### 3.2 授業への影響

ChatGPT の利用がデータ解析の教育に与える影響を考えると、コードやアルゴリズムの理解が依然として重要であることが明確になった。本節では今後 1 ～ 2 年スパンでの短期的な授業への影響と、それよりも長いスパンでの中長期的な授業への影響について考える。

#### 3.2.1 短期的な影響

プログラミングの基礎知識やデータ解析の論理的な思考プロセスは、ChatGPT を用いても変わらず重要な要素である。従って、授業の大枠に関しては、従来のアプローチを維持する必要がある。ただし、演習問題に ChatGPT を取り入れることで解析のサイクルを数倍速めることが可能となる。これにより、学生はより迅速にそれなりの結果を得られるようになる。この変化は、授業における焦点を学習モデルの初期構築から、より高度な改良や最適化のプロセスへとシフトさせる。教師は、この新しい動向に対応するために、学生がモデルを深く理解し、効果的に改良する方法について指導する準備を整える必要がある。

総じて、ChatGPT の導入は短期的には教育の方法を根本から変えるものではないが、教師と学生の両方に新たな機会と課題を提供する。このツールを効果的に活用し、同時にデータ解析の基本的なスキルと理解を維持・深化させることが、今後のデータ解析教育の発展に不可欠である。

#### 3.2.2 中長期的な影響

中長期的な観点から ChatGPT の導入が教育に与える影響を考えると、その効果はデータ解析の授業に限定されない。ChatGPT を活用することで、対話形式の自学自習が可能となり、学習者はデータ解析をはじめとする多様な分野で知識を深めることができるようになる [2)]。これにより、自律的な学習スタイルが促進される。しかし、ChatGPT は現時点ではハルシネーションの問題を抱えており、完全に信頼できるチューターや情報源として機能するには至っていない。ただし、技術の進歩によりこの問題は遠くない将来には大幅に緩和される可能性がある。

近い将来、学生は ChatGPT を使って高度な概念も自学自習できるようになり、かつ、ChatGPT の解析手法もベースライン以上に進むことができるレベルに到達すると予測される。その結果、大学の授業形態自体にも変化が生じる可能性がある。基本的な概念の習得は学生自身の自学自習に委ねられ、大学の授業では概念の習得の確認や応用的な問題の演習に焦点が置かれるようになると考えられる。これは、反転授業や分野をまたいだプロジェクトベースの演習が主流となることを意味する。

中長期的な影響の文脈において、データ解析系授業への具体的な影響は以下のように考えられる。まず、反転授業においては、学生が ChatGPT を活用して教師が与えた問題を解析し、その結果をグループワークで討議する形式が一般的となると予想される。ChatGPT のアプローチは単一ではないと予想される [3)] ので複数の結果を持ち寄ることで、学生はベースラインからどのような工夫を加えることで精度を向上させることができたのかを深く理解し、共有する機会を得る。

プロジェクトベースの演習については、学生が自らテーマを探求し、それについての解

析を行うという基本的な枠組みは現在と変わらない。ただし、ChatGPT との対話を通じて、全ての参加学生がベースライン以上の精度に至ることが可能となる。これにより、各学生のプロジェクトへの関わりが従来よりも深まり、より充実した学習経験を提供することになる。

中長期的には、ChatGPT の利用が教育全体のあり方に影響を及ぼす可能性がある。データ解析系授業においても、基本的な概念の習得は学生自身の自学自習によって行われ、大学の授業ではその確認や応用的な問題の演習に重点を置く形式が主流となる見込みである。このように、ChatGPT は学習方法や授業形式に変革をもたらし、学生が体系的な知識と批判的思考法を効果的に身に着けるための新たな手段を提供することになると考えられる。

**注**

1) 当初 ”Code Interpreter” という名称で導入されたが、その後 ”Advanced Data Analysis” に改名され、現時点では特定の名前はあてがわれていないようである。
2) プログラミング言語の自学自習の例は既に教科書レベルで紹介されている [8]。
3) 現時点でも例えばベースラインからの向上案として特徴エンジニアリングやハイパーパラメータのチューニングなど複数の案が提示される。

**参考文献**

[1] 倉光君郎 ,「ChatGPT と Colab の連携：ChatGPT を活用したプログラミング演習の実例」, 教育機関 DX シンポジウム（2023 年 5 月 12 日）, available at https://edx.nii.ac.jp/lecture/20230512-06.
[2] 柳瀬陽介 ,「大学英語教育における ChatGPT 活用型授業実践：英語教師が認識する生成系 AI 活用の可能性と限界」, 教育機関 DX シンポジウム（2023 年 10 月 13 日）, available at https://edx.nii.ac.jp/lecture/20231013-05.
[3] 山田優 ,「大規模言語モデル ChatGPT を活用した翻訳トレーニングと外国語教育」, 教育機関 DX シンポジウム（2023 年 11 月 13 日）, available at https://edx.nii.ac.jp/lecture/20231113-06.
[4] Y. Zheng, “ChatGPT for teaching and learning: an experience from data science education,” available at arXiv:2307.1665v1 [cs.CY], 2023.
[5] https://signate.jp/
[6] 自動車の走行距離予測 , available at https://signate.jp/competitions/121.
[7] 画像ラベリング（20 種類）, available at https://signate.jp/competitions/108.
[8] 小野哲 ,「ソフトウェア開発に ChatGPT は使えるのか？ー設計からコーディングまで AI の限界を探る」, 技術評論社 , 2023.