機械学習による
予測と最適化

DataVehicle

洞察・予測・最適化
～AI 開発で同様のところと違うところ

AI 開発ではどうなのか

データ分析のような定量的な研究は、「何かと何かの違いを生んでいる原因がどこにあるかを考える」ために行なわれます。したがって「購買金額の高い顧客とそうでない顧客の違い」のように、アウトカムがよい状態の解析単位と、そうでない解析単位の違いはどこにあるのか、と考えればよいわけです。アウトカムとは可能な限り確実に長期的な利益につながる指標であり、優先して考慮すべき解析単位の特徴についても学びました。このような分析を行なえば、「解析単位の状態を変える」か「リソース配分を変える」かというアクションにしたがって利益を向上できるかもしれません。

これが分析による「データの活用」ですが、読者の中にはデータ分析ではなく、「AI への活用」についての期待で本書を手に取られている方もいらっしゃるでしょう。

一般的な呼び方は異なりますが、「このデータに対してどのような AI を作るか」ということを考える場合にも「アウトカム」と「解析単位」という考え方は役立ちます。AI の領域においてアウトカムは「目的変数」あるいは「外的基準」と呼ばれることが多く、解析単位にあたる表現はあまり一般的ではありませんが、「対象」と呼んでいます。

つまり、データ分析では「アウトカムがよい状態の解析単位とそうでない解析単位の違いは何か？」を考えましたが、AI 開発においては「対象ごとに目的変数を正確に予測し、最適な選択肢を提示する」というように考えていきます。

また、データ分析と AI の間ぐらいの位置に「最適な選択肢を提示する」ということを抜きにして、「条件を入れれば目的変数を正確に予測し出力する」という、統計学や機械学習を応用した「予測モデル」という考え方が存在しています。

データ分析の場合と同様、予測モデルや AI の開発においてもこれら 2 点(目的変数と対象)を整理することが意義ある目的設定となり、「活用のためのデータ」として、解析単位や対象ごとにデータを整理することも意義があります。また、対象ごとの特徴量を網羅的に考えれば高い精度での動作につながる点も共通しています。

データ分析と AI 開発で異なる「活かし方」

一方で、根本的な違いはその活かし方にあります。データ分析では過去のデータからもっともよく当てはまる説明変数とアウトカムの間の関係性を導き出します。そこから、何がどれくらい関係しているのか、ということがわかれば、前述の通り「解析単位の状態やリソース配分を変える」というアクションにしたがってアウトカムをよりよく変えていくことができます。データ分析の手法に「統計的因果推論」と呼ばれるものがありますが、その名の通り「因果関係」を適切に「推論」することができれば、その原因を改善してよりよい結果を得ることができるでしょう。たとえば現代において「喫煙すると死亡率が上昇する」という知見も統計的因果推論にしたがって得られた成果で、政府が喫煙対策に取り組めば死亡率が下がる、というアクションが考えられます。

予測モデルを作る際には、このようなデータ分析と同じ手法を使うこともあります。「もっともよく当てはまる関係性」を使えば、このまま同じ状態が続く限り正確に未来を予測できるからです。同じ手法を使っていても、その活用方法は異なります。先ほどの喫煙と死亡の例でいえば、喫煙への対策ではなく「このままの喫煙率が続くとすれば 10 年後に何人が亡くなるのか」といった予測に使うわけです。この予測がうまく行けば、年後に病院のベッドがどれだけ必要なのか、どれぐらいの量の薬を用意すればいいか、ということがわかります。そのことでムダなリソースを省いたり、不足して困る事態を避けたり、ということができるでしょう。これはビジネスの場面においても同様です。適切な予測モデルをつくれば、生産量や在庫、コールセンターの人員数などのリソースをムダにせず、不足もしない最適な量を考えることができるでしょう。

さらにいえば、予測モデルを作る手法では、データ分析よりはるかに複雑で、中身がブラックボックスとなるようなものも許容される、という側面もあります。データ分析では最終的に「何がどれくらい関係しているか」と人間が理解しなければなりませんが、予測モデルではとにかく予測が正確に当たりさえすればよいわけです。多少中身が複雑でブラックボックスになっていても、予測値と実際の値の間のズレが小さくなっていればよいと考えます。そのため、「データからもっともあてはまりのよいものを考える」という数学的な側面はデータ分析と同じであっても、より複雑で、統計的機械学習と呼ばれる手法が予測モデルではよく使われます。

最近の AI の発展は、こうした統計的機械学習にしたがって支えられていますが、AI すなわち人工知能というからには目的変数の予測値を正確に算出するだけでなく、有限の選択肢のうちもっとも「よい」ものを提示するところまでが含まれます。(ただし、単なる機械学習による予測モデルのことを AI として売っている会社も見かけます。)

インターネットにアップされた 1 枚ずつの画像を

対象に、統計的機械学習にしたがってどのような単語のラベルがつけられる確率が高いか、という目的変数を正確に予測する AI が、近年の AI 研究のブレイクスルーとなりました。これによって、画像に何が写っていると判別すればよいか、という最適な選択肢を提示できるわけです。こうした技術を応用すれば、１枚ずつのレントゲン画像から、さまざまな病気に罹っている確率という目的変数を正確に予測することもできます。そこから、何の病気への警戒を表示するか、というもっとも適切な選択肢を提示するわけです。他にも「ロゴデザイン」や、「文章の翻訳」といったさまざまなものについて最適な選択肢を提示する AI が開発されています。

予測モデルや AI の価値を分けるところ

このように十分なデータをうまく使いこなせば、とても便利な道具を作れる機械学習の技術ですが、データ分析でリサーチデザインという考え方が欠如していたときと同じような問題が存在しています。リサーチデザインという考え方を知らない人が、どのような目的で何のデータをどう分析してよいかわからない、という状態に陥るという話を前章ではしました。あるいは、リサーチデザインの考え方がわかっていなかったために、せっかくのデータからナンセンスな結果しか見つけられなかった、という人もいます。

AI を開発する場合も同様で、この世にはナンセンスな AI 製品がいくつもあります。私たちのところには近年こうしたものについての相談も寄せられていて「せっかく大学と協力して作った AI 製品がすこしも売れないのだがどうすればよいか」といったことも数多くあります。予測モデルについても「予測自体はうまく行っているが、コスト削減につながるわけではない」といわれることもあります。また、「開発の時点ではうまく機能したように見えたが、実際に運用するととんちんかんな結果が出てくる」といわれることもあります。

この数年で、ディープラーニングをはじめとする統計的機械学習の手法は大きく進歩し、使いやすい道具や、わかりやすい資料もたくさん生まれました。これらを活かせば何かできそう、と思うのはとても夢のある考え方で、実際にうまく活かせた人たちもたくさんいます。しかし、手法の中身やそれを実装するためのプログラムの書き方、という知恵はどんどん普及しているのに対し、実際にそれをどう活かすか、という知恵はあまり知られていません。

データ分析におけるリサーチデザインという考え方を持っている状態で「価値を生まない AI や予測モデル」の相談を多数もらった私たちは、これらに対してもリサーチデザインの考え方を応用すれば、「どのような課題に取り組むべきか」を判断できることに気づきました。

また、機械学習手法の中身について説明している資料であっても、予測モデルや AI の精度が「今後も有効に機能するようなものなのか」をどう検証するかという点についてはあまり詳しくは書かれていません。この点については統計学を用いた場合とまったく同様に、クロスバリデーション法という考え方を使えば、機械学習の精度もきちんと評価できます。こちらについても本章の後半で詳しく説明しましょう。

私たちの考えでは、多くの企業やビジネスマンにとって、機械学習技術から価値を生めるかどうかを分けるのは、手法自体の理解やそれを実装する技術力ではありません。すでに述べたようにディープラーニングであっても「便利な道具と資料が普及してきた」ために、実装自体はとても簡単になってきています。具体的にいうと、画像を認識

して最適な選択肢を選ぶといった「ちょっとしたAI」なら、「活用のためのデータ」が整備された状態から、数十行ほどのプログラムを書くだけで実装できます。

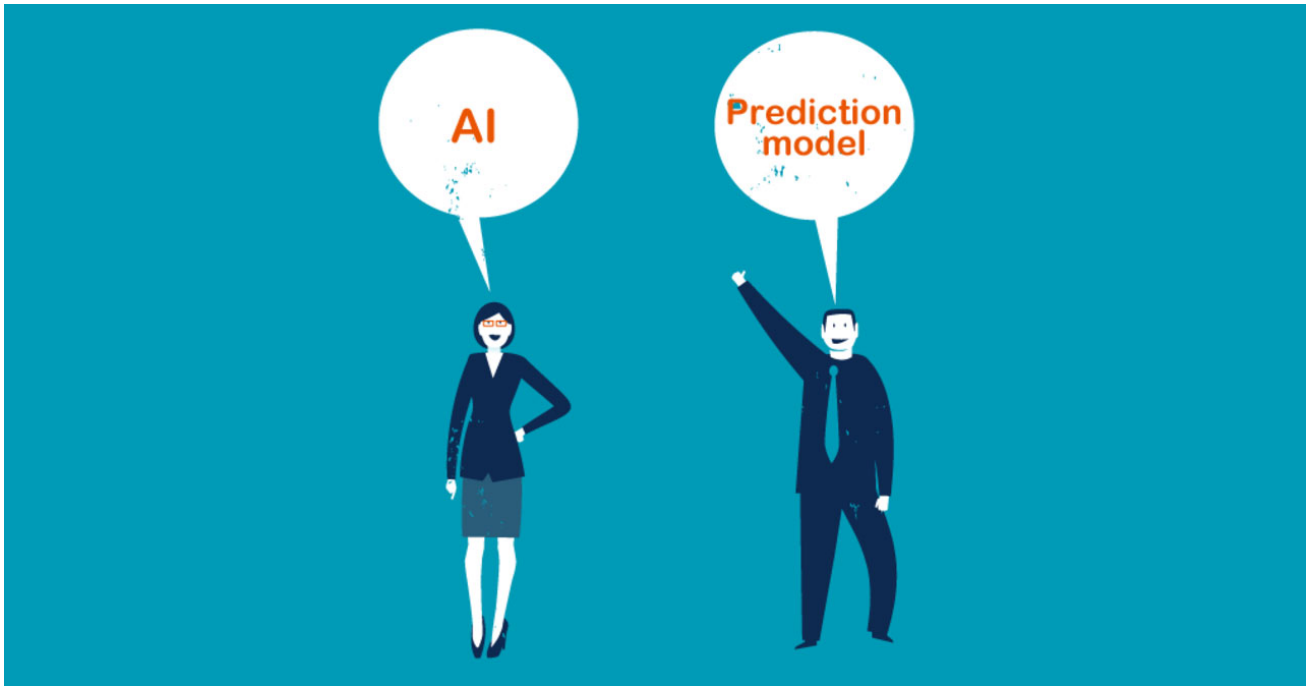
Google の TensorFlow は、無料で商用利用できるソフトウェアライブラリですが、彼らはこの中で「直感的な短いコードを書くだけで、ディープラーニングが実装できる」という機能を積極的に開発しているぐらいです。それ故か、IT 系の専門媒体の編集長に聞いたところ、AI の開発を外注できる企業は、大手から中小まで合わせるとすでに国内で 100 社を超えているそうです。大学の工学部サイトを調べれば、AI 開発の共同研究を受け付けている先生方を見つけられるかもしれません。皆さん自身が機械学習技術を知らなくても、必要な予算を確保して彼らに対して適切な依頼をすれば、AI 製品を作り上げることが可能です。

その一方で、「誰もお金を払う人がいない」AI が、この瞬間にも作り続けられています。納品時に報告されたよりも低い精度でしか、運用上機能しないというケースも起こります。

したがって、リサーチデザインに基づく「課題の設定」が適切かという点と、クロスバリデーション法にしたがってきちんと性能を評価できるかという点をクリアできる側面の方が、多くの人にとって「機械学習技術を活かす」上で重要になります。前者ができていればマネタイズに失敗するリスクは下げられ、AI 開発の外注も明確にオーダーすることができます。また、後者ができていれば機械学習手法の中身はわからなくても「それが今後機能するかどうか」を正確に判断できます。

極端な話、予測モデルや AI の性能を評価するにあたり、依頼する会社の規模も、技術者の年齢や性別、学歴すら考慮する必要はありません。いくつかの課題に対し、機械学習手法を適用させて、クロスバリデーション法に基づき、どの課題にも安定して高い性能を示す人がいれば、確実に「腕利きの機械学習技術者」と判断してよいでしょう。本章ではここから、こうした「機械学習技術を活かす方法」について詳しく説明していきます。

予測モデルと AI の使い分け



正確な予測が価値を生むとき

まずリサーチデザインを応用して、予測モデルや AI 開発の課題設定について考えてみましょう。予測モデルとは統計手法や機械学習手法を使って「とにかく正確に何かの値を予測してその結果を出力するもの」と述べました。一方 AI については「予測に基づき、最適な選択肢を提示する」ものと述べました。

いずれにしても、対象ごとの目的変数を正確に予測することは価値につながります。株や不動産、原油などの価格が今後どうなるか、人間より正確に予測できるようになった人は大金持ちになれるでしょう。売上やキャッシュなどの経営指標が今後どうなるか正確に予測できる人は、会社のリソースを限界まで使って、効率的な経営をすることができます。スポーツチームはドラフトやフリーエージェントから獲得可能な選手のうち、長期的に見て、誰がどれぐらい活躍できるかをうまく予測できると、年俸を抑えた上で勝利を重ねることができます。大災害が起こるかどうかを正確に予測すれば、被害を抑えて人命が救われます。盤面上のどの場所に石や駒を置けば勝率が高いかを、誰よりも正確に予測すれば囲碁や将棋でも勝つことができます。

これらは「予測の正確さ」、つまり人間が直感的に行なうよりも正確な予測ができるという「予測精度の改善価値」が重要になるパターンです。一方で、人間より精度の低い予測であっても、予測モデルや AI が価値を生むパターンはあります。それは「手間の省力化」という側面です。

省力化というもう 1 つの価値

Amazon のページを開くと、そこに自分におすすめの商品が表示されます。この仕組みの背後でも機械学習手法が応用されています。何の商品を表示すべきかという「最適な選択肢」を提示していると考えればこれも一種のAIと考えることができます。すなわち、書店員や図書館の司書、あるいは本好きの友人に対して「こういう本が好きで何か他にも読んでみたいんですけど、何かおすすめはありますか?」と聞いて教えてもらうという人間の行為を、人工的に代替したものだと考えてもそう間違いはないでしょう。

この Amazon のおすすめが人間より正確かというと、必ずしもそうではありません。自分と相性のよい書店員が働くお店では、いつも行くコーナーに Amazon では見つからなかった本との出会いが存在します。

一方、Amazon は執拗に私に対して、拙著の『統計学が最強の学問である』を推薦してきます。確かに私は、統計学や機械学習の手法や応用例について書かれた本を大量に買っているので、データ上「あれだけ売れた類書をまだ買っていない」と判断されるのはよくわかりますが。

しかし、よい書店員や司書、友人の時間は無限ではありません。世界中の Amazon の全ユーザーに対して、いつでもどこでも 24 時間電話一本ですぐに「こういう本が好きなんだけどおすすめはありますか？」と質問できるコンシェルジュサービスを提供しようとすれば、おそらくあの Amazon ですらすぐに吹き飛ぶほどのコストがかかってしまうことでしょう。「人間がよく考えた結果」より多少精度が低かろうと、一度きちんと仕組みを作ってしまえば、ごくわずかなコストで自動的に人間の仕事を代替できるというのは確かにビジネス的な魅力です。

２つの価値のトレードオフ

データ分析においては、とにかくアウトカムが長期的な利益に直結するものでなければならない、と述べましたが、予測モデルと AI の価値はこのような２つの価値の掛け合わせで決まってきます。また両者の間にはある程度のトレードオフが存在しています。

予測精度の改善自体が大きな価値を生むのは、たった 1 つの予測に基づいて、大金が動くようなもの、大量の人命がかかっているようなものごとです。このような仕事は人類全体で見ればごくわずかな一部の専門家が行なうものであり、また専門家であったとしても仕事時間のすべてを予測だけに費やしているとは限りません。したがってこうした予測を AI で自動化したとしても、省力化のメリットは大きくないかもしれません。極端な話、「大地震を正確に予測する」というのは日本人にとって大きな価値になりますが、地震を研究する学者が自分たちの省力化になるからとお金を払ってこの AI を買ってくれることはないでしょう。

「予測精度の改善価値」があまりに高すぎるような課題では、「人間よりもかなり精度が下がるけど省力化につながるからよい」といった判断はしにくくなります。日本において、救急車を呼ぶべきかどうか電話で相談する窓口を設けている自治体は少なくありません。この窓口の業務を AI 化するには、人命に関わる「予測精度の改善価値」が大きすぎるために、人間と同等以上の予測精度が得られないのであればその導入に対して慎重にならざるを得ないでしょう。

逆にいえば、「予測精度の改善価値」がそれほど高くない課題であれば、精度が下がったとしても、省力化ができるというだけで価値になります。先ほどの「何の本をおすすめすれば相手は喜んで購

入するか」という確率を、Amazon のレコメンドエンジンは熟練した書店員ほどには正確に予測できない、という話を思い出しましょう。このような課題は多少精度が下がったとしても大きな問題になることはありません。社会の中には、専門家として、正確な予測をしなければならない課題だけでなく、日常的に「多少精度が下がっても問題にはならないが、社会全体では大量の時間や手間がかかる課題」というものがたくさん存在しています。社会全体に存在する「大量の時間や手間」のことを私たちは「課題の総負荷量」と呼んでいます。

予測モデルか AI か

実は統計学や機械学習を、主に「予測精度の改善価値」のために使うか、「課題の総負荷量」の低減のために使うか、という目的に応じて、同じアルゴリズムでも予測モデルとして使うか、AI として使うかという使い分けを考えることができます。

正確な予測値自体が価値を持ち、それが得られれば十分である、というのであれば、予測モデルを作ればよいでしょう。こうした課題は、「自動化」されることが必ずしも望ましいものではありません。

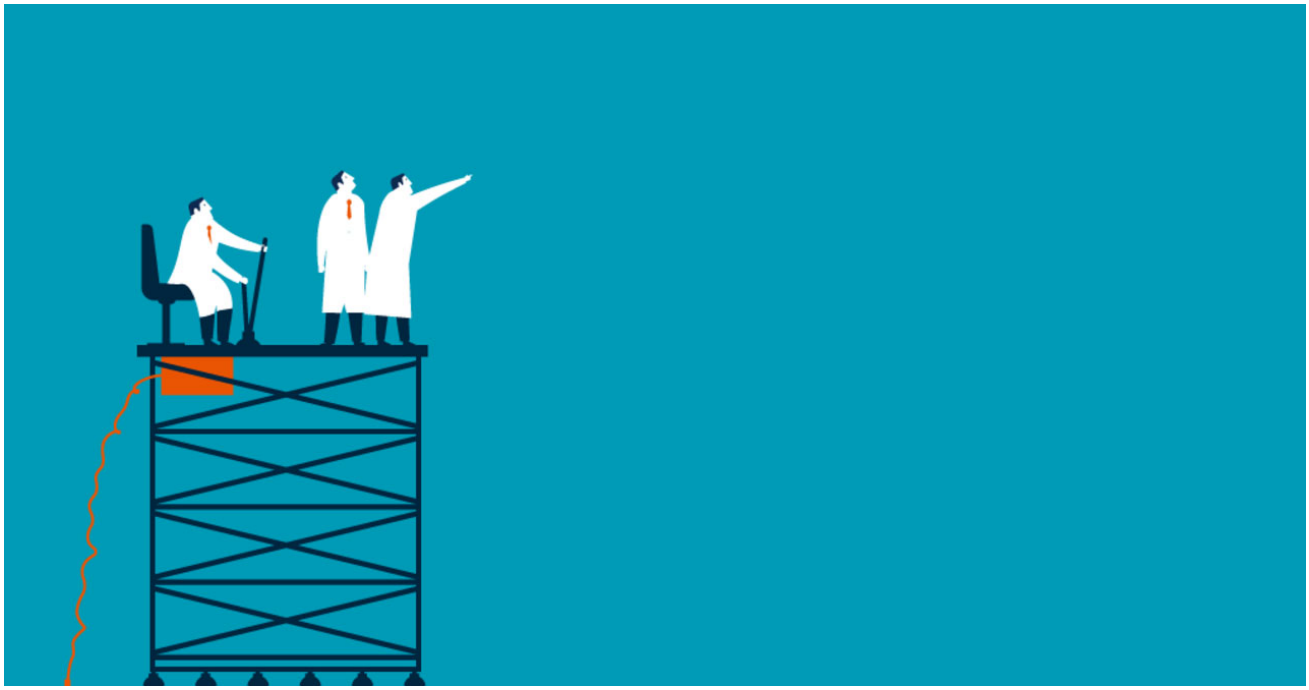
製造業の中で、特定の製品の需要を正確に予測し、生産計画を立てるという課題について考えてみましょう。需要を正確に予測することで在庫やキャッシュフローのムダを省くことができますので、予測精度の改善価値があることは間違いありません。課題の総負荷量という点ではどうでしょうか？専門スタッフでも、毎日、需要予測だけを考えているわけではないので、自動化して省ける人件費などたかがしれています。さらに自動化することで「たまたま機械学習が大きく予測を外す」という状況になった時のリスクをコントロールしたり、責任を取ったりすることができなくなってしまいます。とすれば、正確でかつ専門スタッフにとり、使いやすい予測モデルを作り、管理できるようなツールの方が AI よりも役に立つはずです。

同じ製造業の仕事であっても、原材料を仕分ける、製品を指示通りに箱や袋に詰める、シールやタグを貼る、そのミスをチェックするという作業について、大量のアルバイトの人海戦術で乗り切っている工場もまだまだたくさんあります。こうした作業の精度は１、２％程度、人間が行うよりも低かったとしても、大した問題にはなりませんが、課題の総負荷量は世界レベルで見れば大きなものになるでしょう。つまり、もし「人間と比べてそう悪くない」という精度でこうした課題を自動化できるのだとすれば、それは大きな価値を生むということです。

あるいは総負荷量の高い課題は、オフィスや工場といった仕事の場だけに存在しているわけではありません。家庭でも、夕食の献立を考えて買い物をしたり、ゴミを分別したりといった、「多少精度が悪くてもそれほど問題にならないが、世界全体では大きな時間がかかっていること」はたくさんあります。こうした課題を解決する AI 製品の開発に成功すれば、大きな付加価値を持つ商品として世界中でヒットするかもしれません。

皆さんにとって、あるいは皆さんの顧客にとって懸念となるのはどちらのタイプの課題でしょうか？これらの使い道を間違えないように是非よく考えてみましょう。

何をデータで予測させるべきか



何の予測モデルを作るべきか

予測モデルと AI の使い分けが理解できたら、それぞれについて「何を予測させるべきか」を考えてみましょう。予測モデルの価値とは(人間以上に)予測精度の改善できることそれ自体だと述べました。とするならば、何を予測させるかを決めることというのが予測モデルについてのリサーチデザインに該当することになります。

つまり、失敗した予測モデルの多くは「予測できてもしょうがないこと」を正確に予測しようとしていると表現することもできるでしょう。その中には「正確な予測に基づく有効な対処が存在しないこと」も含まれます。

顧客の初回購買の情報から「その後、顧客が優良顧客となり得るかどうか」を正確に予測できるようになったとします。しかし、これが「へ〜この人、優良顧客になりそうなんだ」とか「へ〜この人もう来ない可能性高いんだ」と一喜一憂するだけなら意味はありません。あるいはこうした人たちにクーポンを送りつける、というのも意味があるかはわかりません。機械学習における予測値は「このままの状態が続くとして」という仮定のもとで得られたものです。何もしなくても優良顧客として大量に購買してくれた人に対してムダに 10%オフの割引をオファーすることになるのか、それともクーポンをきっかけにして、さらに値引き分以上にたくさん購買して利益をもたらしてくれるのか、予測モデルは何も答えてくれないわけです。

このような状況で、「クーポンを送るというアクションありき」であれば、予測すべきはむしろ「クーポンを送った場合と送らなかった場合でどの程度その後もたらしてくれる粗利に差がつくか」ということです。クーポンの送付にかかるコスト以上に大きな粗利が見込める顧客を、正確に見分ける

ことができれば大きな利益につながるでしょう。しかしこれはディープラーニングを含め、一般的な機械学習手法では予測できない値です。本書の最後で詳しく述べますが、こうした因果関係を正確に知るためにはA/Bテストや、統計的因果推論といった手法が必要になってきます。

少し話がそれましたが、人はしばしばこうした「有効な対処が存在しないこと」に関する予測モデルを作ってしまいます。このような状況を避けるためにどう考えればよいのでしょうか？重要なポイントは2つあります。

①予測精度が感覚的に今の1.05倍になればいくら儲かるか
②「今の状態が続くとして」という仮定はどの程度成り立つか

予測精度の不完全さ

「予測精度の改善価値」という考え方に立ち返ってみましょう。当然予測精度には限界があり、100%その後に得られる値と一致する、という以上の予測は存在しません。為替なり、株価なり、1時間後の指標を100%当てることができれば、その人はあっという間に大富豪になることでしょう。

神ならぬ私たちは、さすがに完璧な予測はできません。それは高度な統計解析や機械学習を使った状態でも例外ではなく、アルゴリズムがどうこうというより「予測するのに十分なデータが存在していない」という点にあります。ある液晶テレビがどれぐらい売れるかどうか、という予測について考えてみましょう。

基本的に会社が作る製品には「規格」や「仕様」という概念があり、色や機能、使われる部品や接続できる端子といったさまざまなデータが存在しています。これらのスペックがまったく同じ製品を出したとしても、ちょっとしたデザインのディティールが違うというだけでその売上は大きく変わるかもしれません。また、ほとんど同じデザインで同じスペックの商品を販売しても、それを売っている企業のブランドイメージや用いる広告によっても売り上げは異なることでしょう。さらには、たまたまシェアの大きい小売店のバイヤーと仲のよい営業スタッフがいるとか、たまたま影響力のあるインフルエンサーが手にする機会があって、「これいいよ」とSNSで発言しただけでも売り上げは大きく変わります。

このような状況で製品スペックのデータだけを使っていくら高度なアルゴリズムを駆使しても、売上を正確に予測することはできません。アルゴリズムを改善するよりも「商品や販売する企業のブランドイメージ」「デザインや広告に対するイメージ」「小売店のバイヤーからの商品評価や取り扱いの意向」といったさまざまなデータを入手した方が正確な予測につながるはずです。ただ、そこまでやってもなお、マスメディアやSNS上、さらにいえばローカルなコミュニティの中でインフルエンサーとなるような人が「たまたま手に取るか」というデータは収集することもコントロールすることもそうそうできるものではありません。

「完全な予測」ができないからといって予測を行うことがムダ、ということはありません。具体的に何台ということはわからなくても、液晶テレビならインチのものとインチのもので、それぞれ何台ぐらい売れそうか、くらいは多くの家電メーカーが過去の実績を把握しているはずです。その感覚に基づいて需要を予測し、生産計画を立て、うまく経営が回っていることでしょう。

この状況で、データに基づく予測モデルを緻密に作る必要があるのでしょうか?その判断の基準と

して「予測精度があと 1.05 倍になったときに」というシミュレーションを私たちはお勧めしています。

1.05 倍の予測精度の向上価値

なぜ「感覚的に 1.05 倍」か、という点について説明しておきましょう。おおむね現在行っている経験や勘による意思決定から、短期間でデータを活用して改善できる予測の精度についての私たちの経験的な改善余地ですが、ただの経験則というだけではありません。MIT スローン経営大学院の Brynjolfsson らの論文によれば、「データに基づいて意思決定する企業は 5 ～ 6 %生産性が高い」という結果が示されています。私たちの経験則と、経営学者の分析結果が一致するのであれば、一つの目安としては十分に参考になるでしょう。人間の直感の倍もの精度で未来を予測しようとするとたいへん難しいですが、1.05 倍程度ならけっこう何とかなるものです。

この「1.05 倍の精度」とはどういう状況でしょうか？予測しなければいけない値が、商品の出荷数のように数字で表わされるものだったとき、経験的な判断により 1 万個の製品を生産したが、実際には 8 千個しか売れず、 2 千個の商品が余ってしまったので、それらを売り切るために余計な販促費をかけ、最悪捨てることになる、というのが現状だったとしましょう。このような場合、短期的な成果として、データをうまく使えばこの余ってしまう商品を 5 %ほど削減して、1900 個ほどにすることはそう難しくありません。

その結果、この会社はいくら儲かるでしょうか？予測の失敗がすなわち商品の破棄につながるなら、その差である 100 個分の製造原価が短期的に得られる予測精度の価値になります。1 つの商品に 10 万円ほどの製造原価がかかるなら、この「1.05 倍の精度」の予測モデルは 1000 万円ほどの利益に貢献したことになります。あるいは販促費などを使うにしても「捨てるよりは多少お金が残る」というのならばもう少しこの金額は控えめなものになるかもしれません。

あるいは定性的に「該当するか否か」という状態を予測する状況についても考えてみましょう。製品の製造に必要な機械が「そろそろ故障しそう」という状況について、年間 100 回故障するうちの回については経験と勘で予め対策が取れて問題になっていなかったとします。

この予測精度を 1.05 倍にできるということはすなわち、100 回中の回についてこのような対策が取れるというイメージです。この年間 4 回分の対策により、工場が止まってしまう機会損失や人件費のロスがどの程度の金額になるのか、というのがこの予測モデルの価値になります。

このように考えた上で特に金銭的なメリットが感じられないのであれば、それは正確な予測モデルが求められていない状況なのでしょう。第 2 章で述べたように、すでに「100 万個中数個程度の不具合」といったレベルになった品質に対して、データ分析がそれほど大きな貢献をできないのと同じような話です。勘と経験だけですでにある程度正確に予測できていることに対して予測モデルを使っても大きな価値は生まれません。また、それほどうまく予測できていない状況で、たいしたコストが生じていないような課題に対しても、やはり予測モデルはあまり価値を生みません。

データを使って予測モデルをきちんと作った方がよい課題とは、「現在の予測がいいかげんでムダの大きい領域」だということができるでしょう。生鮮食料品の廃棄や値引きする金額が年間百億円にものぼる、というスーパーマーケットチェーンで

あれば「1.05 倍の精度」の価値は 5 億円ほどになります。自社あるいはクライアントの企業の中で、どこにこうした「予測がいい加減であるが故のムダ」があるかを考えることが、価値を生む予測モデルの開発につながるわけです。

「今の状態」はどの程度続くか

ただし、予測精度のいい加減さによるロスが大きいからといって、必ずしも予測モデルが機能するわけではありません。それが 2 つめのポイントである、「今の状態が続くとして」という仮定はどの程度成り立つのか、という考え方です。

基本的に機械学習とは「今データが得られた状態が続くとして」という仮定のもとで予測を行う方法です。生鮮食料品がいつ、どれくらい売れるか、ということを過去 1 年間のデータから正確に予測できるような予測モデルを作ったとして、この「過去 1 年間」の状況から、データに含まれていない変化が生じてしまえば予測は大きく外れることもあります。

スーパーマーケットの近くに大きなタワーマンションが立てば、それだけでたくさんの人が買い物に来て、生鮮食料品だろうがなんだろうが、たくさんのものが売れることになるでしょう。また、同じ人たちがずっと店舗の周辺に住み続けていたとしても、加齢と共に食べられる量が少なくなることもあります。「魚が体によい」とか「バターは体に悪い」といったような研究成果が広く知られるようになると、それ以降食品の売り上げが急に上がったり、下がったり、ということも当然考えられます。どこかの地域で大地震が起こったために「ミネラルウォーターをたくさん買い込んで置いておこう」と思う人だっています。これらが「今の状態」が続かなくなるという例です。

したがって、予測モデルとは一度作ればそれで終わり、というものではなく継続的なメンテナンスが必要なものです。また、時として「今の状態」が崩れたために大きく外すリスクもあります。これはもちろん人間が経験と勘で予測をしていたとしても避けられないリスクですので、データを使った場合に限った問題ではありません。しかし、「今の状態」がどれだけ安定的に続くのかを考えておくことで、今から作る予測モデルがどの程度の期間「予測精度の改善価値」を安定的に生み続けるのかと見積もることができます。

そう考えると、商品の需要予測というテーマでも、生鮮食料品よりアパレルやエンターテイメントという「流行りすたり」がある領域では、「今の状態」が維持されにくいかもしれません。あるいはテクノロジー面でのイノベーションのスピードが速い領域についても、突然ディスラプティブな製品が生み出されて、それまでよく機能していた予測モデルが突然役に立たなくなるかもしれません。現代の日本で楽曲の販売数についての予測モデルを作ろうとすれば、その予測値は「ジャニーズか」「秋元康プロデュースのアイドルか」といった条件によって大きく左右されるはずです。ほんの数年後にはそうした文化が「イケてない」と捉えられ、ほとんど売れなくなることもあり得ないとはいえません。また、iPhone の発売前のデータを使って、携帯電話のスペックから販売台数を予測していれば、ワンセグかどうか、Felica で決済ができるかという条件が重要とされていたのではないでしょうか。iPhone の登場で日本国内の携帯電話市場は大きく変化し、それ以前に作った予測モデルはほとんど機能しなくなりました。

皆さんが何かの予測モデルを作ろうとするのであれば、以上のような 2 点について吟味してみることをおすすめします。

AI 開発における適切な課題設定とは

省力化に着目した AI の課題設定

「予測精度の改善価値」にフォーカスする予測モデルについては「現状の 1.05 倍の精度で予測してどの程度のコスト削減効果があるか」と考えればよいことを学びました。また「データを収集したような状態」がどれだけ安定的に続くかというのも重要なポイントです。

では、もう一つの機械学習の使い方である、(多少人間より精度が低くてもよいので)自動化/省力化することを価値とする AI についてはどのように課題設定をしたらよいでしょうか？AI すなわち人工知能が「人間の認知活動を機械に代替させるもの」である以上、アイデア出しのために次のような穴埋め問題を考えてみましょう。

○○(人間の行う認知活動)とは
××の有限な選択肢/組み合わせのうち
最適なものを選んでいるに過ぎない。

○○のところに会社の中での業務を入れれば、それは BtoB 向けの AI 製品/サービスあるいは社内用の AI ツールになります。○○のところに家事や趣味のサポートといったものを入れれば、BtoC 向けの AI 製品/サービスとなります。

ロゴデザインを行う AI を提供する会社はいくつもありますが、それは「ロゴデザインとは、フォント、アイコンおよびそれらのレイアウトの組み合わせのうち最適なものを選んでいる」だけだからと考えるから実現できています。

AI が相性のよい相手(異性や求職者)を紹介してくれる、というサービスもいくつもありますが、これも「紹介とは登録される人間のうち最適なものを選んでいる」だけだからです。「AI が夕飯の献立を考えてくれる」というのも「献立を考えるとは

登録されたレシピの組み合わせのうち最適なものを選んでいる」からできることです。

もちろん厳密にいえば、一流のデザイナーはロゴに使うフォント自体に手を加えることがよくあります。同じ人を紹介するにしても「その人のどういう側面をどう説明するか」によって魅力が伝わるかどうかが変わります。献立についても、相手によって既存のレシピから少し材料や調理方法をアレンジした方がよい、という状況もあるでしょう。

しかし「人間より多少精度が下がっても省力化に価値がある」のだとすれば、こうした割り切りこそが現実的な AI 開発の第一歩となります。次にこの「最適な」の基準として、どのようなデータを使うかというところだけを言語化できれば、少なくとも機械学習の技術者とコミュニケーションを始めることができます。「それがユーザーに採用されるか」という確率かもしれませんし、5 段階評価で満足度を聞いてもよいでしょう。あるいは人間が行った判断とどの程度一致するか、という考え方もできます。

AI の価値を判断する基準

総負荷量	世の中でどれだけ人を煩わせているか
安定性	「データの得られた状態」はどの程度続きそうか
有効性	選択肢のうち問題とならないものの割合
責任性	選択を誤った場合のリスクと責任の所在
感情性	人間が行なうこと自体の「うれしさ」

図表3-1　ＡＩの価値を判断する基準

AI に解決させる課題の総負荷量とは

AI で省力化するからには、その課題の総負荷量が大事であるという話をしました。総負荷量とは具体的にいうと、「世界中でどれだけの人が、何時間くらい煩わされているか」ということです。ごくわずかな専門家が、年間数時間程度、会議で話し合ったことがある、というようなテーマであれば、それを代替してくれる AI を作ったとしても大きなお金は生まれません。

また、数多くの人間が長時間頭を使っていたとしても、自ら望んで楽しい思いをしていることであれば、やはり「AI に任せて、やらなくていいから」といわれても、それにお金を払うことはありません。囲碁や将棋を楽しむ人たちに、「AI に任せればいいから明日から囲碁や将棋のことを考えなくてもいいですよ」と呼びかけても、誰もわざわざお金を払って AI を使うことはありません。

BtoB 向けの AI としては、大会社の中にほぼ毎日のように存在して、多くの人が一度くらいは経験して、誰もその仕事に愛着を感じるわけでもない、というような業務があればそれは理想的なものでしょう。

ほとんどすべての小売店にはレジを打つ仕事や、商品を棚に並べる仕事が存在しています。多くのオフィスでは、経費精算の申請や確認、飲み会をアレンジするとか、ゴミを分別して捨てるとか、社内外の誰かにお願いしたタスクが放置されていないかリマインドするといった仕事があります。

こうしたビジネスの現場で使う AI を考える際には、「自分たちの会社の中でどの程度の人数・時間がかかっているか」と考えるのも 1 つの方法ですが、それだと圧倒的に総負荷量に制限がかかってしまうことになりがちです。たとえばある企業で働く千人の作業を省力化する AI を作る手間と、同じ業界で同じように働く 1 千万人の作業を省力化する AI を作る手間を比べた際に、後者に 1 万倍の手間がかかるということはありません。高々数倍というのがいいところでしょう。少なくとも後者の方が 1000 倍以上高い ROI を示すことになりま

す。

つまり、AI を活用しようというのであれば「自社の中での効率化のために」などというみみっちいことを考えていては損です。日本国内の同業者内では使ってもらえるとか、可能であればより広範に「同じようなすべてのオフィス」「同じようなすべての小売店舗」「同じようなすべての工場」といった範囲でより大きな総負荷量を解決する製品化を目指せないかを検討してみましょう。

これは BtoC 向けに製品やサービスを考える上でも、同じことがいえます。ニッチな趣味の一瞬を切り取ったものなどではなく、多くの人が毎日、毎週のようにそこそこの時間「面倒くさい」と感じるような課題こそ、AI で解決すれば大きな市場が見込まれます。極端な例を挙げれば、掃除や洗濯、食器洗いといった家事を AI で今より大きく省力化することに成功した人はきっと大富豪になれることでしょう。あるいは「消費者自身がやっていること」でなく、「消費者に対して人間が提供しているサービス」のうち大きな市場を狙うという考え方もあります。家電製品を買うときの相談に乗ってくれたり、要望を伝えれば好みのワインを選んでくれたり、英会話を教えてくれたり、というサービスのお世話になった人は数多くいることでしょう。こうしたサービスも AI で代替することができれば多くの価値につながります。

総負荷量とトレードオフになる安定性と有効性

このように総負荷量という考え方を持って「どれだけの煩わしさを解決できるか」「もっと範囲を広げて考えられないか」というのは AI の課題設定を行う上で最初に検討すべき重要な視点です。しかし総負荷量が大きくなるようにとにかく範囲を広げさえすればよい、というわけでもないのが難しいところです。

なぜなら課題の範囲を広げれば広げるほど「安定性」が減少する傾向にあります。予測モデルの時にも言及しましたが、機械学習とは「このデータが得られた状態がずっと続くとして・・・」という仮定を置いた上で予測を行うものです。この仮定が崩れるのなら予測が正確に当たるかは保証されず、AI が出してくる「最適な」はずの選択がナンセンスなものになる可能性があります。

実は、囲碁や将棋の AI が成功したのには、この「安定性の高さ」が関係しています。大不況が訪れようが、近くで災害が起ころうが、革命的なテクノロジーが発見されようが、囲碁や将棋のルールが変わることはありません。これが「安定」ということです。

しかし予測モデルのところで、アパレルやエンターテイメントに関する予測については「流行りすたりの影響で予測モデルが機能しなくなることもある」と述べました。これは AI についても同様で、何のメンテナンスもしなければ、「AI が服装をコーディネイトしてくれる」とか「AI が BGM を選曲してくれる」といった仕組みはいつの間にか「微妙に古臭い」というものになってしまうかもしれません。そのような事態を防ぐためには継続的にコストをかける必要があります。

課題の範囲を広げすぎると、安定性以外に有効性の問題が生じることがあります。私たちが AI について、有効性が低いと表現するとき、それは選択肢の組み合わせが膨大すぎて、その多くがまったく意味をなさないとか、害になってしまうという状態を指します。

すでに述べたように「ロゴデザインをする」という AI サービスを提供している会社は複数あります。広告関係の仕事でも、「コピーを書く」という

AI はすでに電通や博報堂などにおいて、ある程度の実績が出ています。これらは、「単語数個程度」や「フォント・アイコン・配置」といった有限な要素の組み合わせの形で表現できるものです。また、最低限の作法さえおさえれば「まったく意味をなさない」とか「害悪になる」ということもそれほどありません。

これが「15 秒のコマーシャルフィルムを制作する」となるとそのリソースの組み合わせは膨大な数になります。単純なデジタルデータとして考えても、１画面の画素数×フレームレート×15 秒分というとんでもない画素の組み合わせになり、それで組み合わされたもののほとんどはノイズのようなものにしかなりません。また、要素を還元して、絵コンテであるとか、その背景の舞台であるとか、セリフ、登場するキャラクター、全体的な色調といったものを考えたとすれば、やはりその組み合わせはとんでもない膨大なものになってしまいます。その多くは「意味不明な場所で意味不明なキャラクターが意味不明なことを話す」という二日酔いの時に見る夢のような内容になってしまうことでしょう。

「15 秒のコマーシャルフィルム」でさえこのような状態なのに、これを「この世のすべての動画コンテンツ」の範囲に広げてしまえば、総負荷量こそ大きくなるかもしれませんが、AI の開発の手間やコストは数倍どころでは済まないはずです。

同様に、「ゴミ捨て」という単純なタスクでも、この世のすべての範囲に拡張しようとすれば、鉛やカドミウムなどの重金属を含む化学系廃棄物や、病原菌に汚染されている可能性のある医療廃棄物なども含んでしまいます。世の中には私たちが今まで人生で一度も見たことのないような「ゴミ」がいくらでもありますが、それらをすべて正確かつ有害とならないように分別するのはたいへんな困難を伴います。

これを「ホワイトカラーのオフィスで出るゴミの分別」と限定すればどうでしょうか？こうしたオフィスで出るゴミのほとんどは燃えるもので、それ以外にはせいぜい食べ物や飲み物の空き容器ぐらいのものです。重金属や病原菌の心配をする必要はほとんどありませんし、家庭と比べても、スプレー缶や包丁といった危険で捨てにくいものが出る可能性はまれにしかありません。しかしそうした絞り込みを行ってなお「オフィスでのごみの分別」タスクはそれなりに大きな総負荷量を持っています。安定性という観点でいっても、オフィスでゴミを分別するタスクのコツが、アパレルや音楽の流行ほど急激に変化することはありません。

このように、いったん総負荷量が大きくなるよう範囲を広げたあと、逆に「総負荷量はかなり大きいまま、安定性および有効性を大きく向上させるようにできないか」という絞り込みを考えるのも大事なコツです。

責任性や感情性がある課題は
人間と組み合わせたシステムを

あと 2 点、責任性と感情性という考え方についてもチェックしてみましょう。囲碁や将棋に勝つか負けるかというのは、本人はともかく社会的にそれほど大きな問題にはなりません。これが課題の「責任性が低い」ということです。しかし、誰かの生命や名誉、自由といった、お金だけでは済まないようなものを対象とする AI についてはその責任について注意が必要です。

予測精度の改善価値と総負荷量のトレードオフのところで言及した「救急車を呼ぶべきかどうか」という課題も責任性が高い課題の一つです。司法や防犯といった領域で、判断を間違えた場合に冤

罪や風評被害を生んでしまうようなAIもそうでしょう。以前 Amazon が、人材採用の判断を行なうAI を「差別を助長する」という危険性から廃止しましたが、こうした判断も責任性の問題と考えられるかもしれません。

もう一方の感情性とは、「人間がやるからこそうれしい」ところです。温泉宿に行った際、女将さんとの会話はそれほどバリエーションがあるわけではありません。また、出張先で見つけたお店のバーテンダーの会話もそれほどバリエーションがあるわけではありません。したがって、カメラから読み取った自分の画像から「お疲れたまってらっしゃいますか？」「今日はこの近くでお仕事ですか？」といった会話内容のうち、どのようなものを提示すべきかということをAIに考えさせることはできます。しかし、AI が代わりにこうした会話をやってくれたからといってちっともうれしくありません。

このように責任性と感情性が高い課題では、AI 単体ですべてを解決せずに、責任を取ったり感情を表現できたりする人間と組み合わせた仕組みの設計がおすすめです。AI と人間をうまく組み合わせれば、高精度で効率的に業務を果たすことができることでしょう。つまり、責任性が高い課題であれば AI が自動的に何かを行うのではなく、専門家に正確な予測値や判断のために必要な情報などをわかりやすいインターフェースで伝えるという方法があります。あるいは感情性でいえば、顧客にはわからないように接客スタッフの手元の端末やイヤホンなどに「相手はこの話をすると喜ぶ」という情報を表示してみてもよいかもしれません。人間の記憶力や判断スピードをAIが補ってくれれば、省力化しながら満足度の高い接客を実現できる可能性もあります。

これだけのことがわかっていれば「何に AI を使っていいかわからない」とか「AI を作ってみたけどマネタイズがうまくいかない」といった事態はだいぶ避けやすくなるはずです。皆さんも是非ご自身のアイデアと機械学習技術を活かして、世界から「負荷」を減らしてみてください。

ズルのできない予測精度の検証方法

「精度 100%の予測」といわれたら、まず疑うべきこと

適切な課題を設定できれば、予測モデルも AI も開発できるようになります。ここまでの具体化ができていれば、統計解析や機械学習の専門家でなくても、日本国内だけで 100 社以上あるという機械学習の開発を請け負う会社に依頼すれば、少なくとも何のアウトプットも出てこないということはありません。

最後の問題は、社内の技術者であれ社外の技術者であれ、あるいは自分自身でも、どのぐらいの精度で予測し、どの程度最適な選択肢を提示できるかを確認することです。これらが「現在の経験と勘による判断よりも高性能」、または「人間がやった場合と比べてそう遜色ない」という水準であれば、課題設定が間違っていない限り、大きな価値を生むことでしょう。今後「このままの状態が続く」という仮定が崩れるような変化に対応するため、定期的なメンテナンスが必要になる可能性はありますが、試験的に導入してみる価値は間違いなくあります。

一方で、予測モデルや AI の精度が、人間と比べてあまりに劣るようであれば開発作業をやり直さなければなりません。予測に必要なデータが他に考えられないか検討し、収集して、アルゴリズムや細かいオプションなどを変えてみて、精度の出る方法がないか検討しましょう場合によっては、実際の予測結果を見てみることで、初めて自分たちが無謀な挑戦をしようとしていたことに気づくこともあります。

性能評価の結果如何で、その後取るべき方向は変わってくるわけですが、問題はこうした開発作業を依頼した技術者が本人の誠実さはさておき、少なくとも「予測精度を高く見せた方が得」という

立場にいることです。本人の無知によるものか、意図的かはわかりませんが、納品された予測モデルや AI が「開発時に示されたような性能で動作していない」ということがしばしば起こります。それは、たいていの場合、予測モデルや AI 開発に使ったデータのみに「過剰に」フィットさせて、見かけの上では高い精度を示したものの、むしろその後の新しいデータに対しての精度が下がってしまう、という現象が起こってしまっているのです。

過学習(オーバーフィッティング)の例

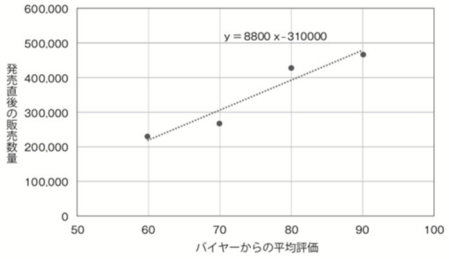
このような現象のことを過学習とか過適合、あるいはオーバーフィッティングと呼びます。単純な例として、次表に示すように、メーカーが新製品の発売前にバイヤーを集め、100 点満点で製品を評価してもらった結果の平均値から、新製品が発売後 1 カ月間でどれほど売れるかと予測する状況を考えてみましょう。

商品	バイヤー評価（平均値）	発売直後の販売数
A	60	230,000
B	70	270,000
C	80	430,000
D	90	470,000

表3-2　バイヤー評価と発売後の販売数量（表）

人間が見ればグラフを見ただけで「何となく右肩上がりだな」というトレンドが見えますし、その直感に基づいて試しに単回帰分析という、直線的に横軸と縦軸の一番あてはまりのよい関係を考える統計手法を使ってみた結果もあわせて示します。(図表 3-2)

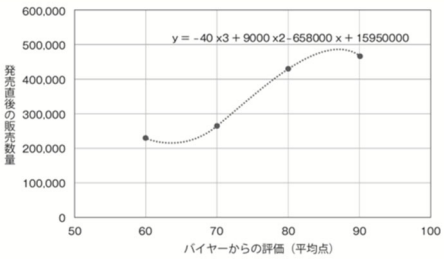
慣例的に数学では横軸を x、縦軸を y で示しますので、今回の場合は x がバイヤーからの評価の平均点、y が発売後の販売数になります。したがってこのトレンドを示す直線の式が y=8,800x-310,000 ということは、バイヤーからの評価が 1 点上がるごとに 8,800 個ずつ発売後の販売数が増える、という傾向を示しています。また、この式に基づくとバイヤーからの評価が平均 60 点なら 8,800×60-310,000 という計算から 218,000 個ほど売れるはずだ、と予測できます。同様に、70 点なら 306,000 個、80 点なら 394,000 個で 90 点なら 482,000 個ほど売れるはずだということになります。



図表3-3　バイヤーからの評価と発売後の販売数量

このような数少ないデータでもこれほどクリアなトレンドがあれば、この「1 点あがるごとに 8,800 個ずつ売れる」という結果は「たまたまのバラつきだけで出るようなものではない」と統計的に判断されるような p 値が得られます。(図表 3-3)

ただし一方で、この予測は「完全に正確なもの」というわけではありません。バイヤーの評価が平均 70 点だったものは 270,000 個ほどしか売れていませんので、実際の値は予測値より 3 万個以上少ない、ということになります。同様に平均 80 点だったものは 430,000 個も売れていますので実際の値は予測値よりも 3 万個以上も多くなっています。

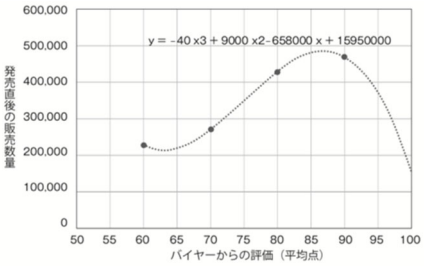


図表3-4　過学習をしてしまった例

付き合いの長いバイヤーの評価とはいっても「たまたま市場を読み違える」とか「たまたま好みが

市場とズレる」ということもあるでしょう。それ以外にも、たまたま広告がウケたとか、どこかのメディアで大きく取り上げられたといった事情もあるでしょうから、バイヤーの評価だけで完璧に予測できるとは限りません。この程度の予測値と実測値のズレが生じることは十分にあり得ます。(図表 3-4)

しかし、いたずらに出てきた予測精度を問題視して、とにかく予測値と実際の値とのズレがほぼゼロになるようにしようとすると、過学習の問題が生じてしまうことが多々あります。直線ではなく曲線状になるように、高校で習う 3 次関数で予測するとまったくズレずに完全な当てはめができるようになります。



図表3-5　過学習による不適切な予測

しかしこの「ズレの小ささ」は健全な状態ではありません。新たに新商品を開発したところ、調査に参加したバイヤー全員が大満足でまさかの平均 100 点、という素晴らしい商品が生まれたとします。このとき、先ほどの直線の式からは「発売後 57 万個売れる」という予測値が得られます。一方でこの 3 次関数は 90 点を超えた時点ですでにピークを越えて販売数量が落ち始める、動きを示しています。実際にこの 3 次関数に「x が 100」という値を用いると、なぜか平均 100 点というこれまでにない素晴らしい商品であるはずなのに、たった 15 万個しか売れない、という考えにくい予測値が得られてしまいます。(図表 3-5)

これらは、わかりやすいように 1 つの説明変数から 1 つの目的変数を予測しようとした例なので、グラフを見れば誰でもその間違いに気づくことができます。しかし複数の説明変数を用いた統計解析や、機械学習ではこのように視覚化することができませんし、特に機械学習では中身がブラックボックスとなるような、「解釈はしにくいけどとにかく予測値と実際の値のズレが小さくなるように」という計算を行うため、過学習を起こしているかどうかについてさらに確認が困難です。

過学習を見抜くためのフェアな評価方法

過学習がなぜ問題になるかというと、今あるデータに対してもっとも予測値と実際の値のズレが小さくなるように計算した状態で、その計算に使ったデータにおける「予測値と実際の値のズレ」を評価しようとしていたからでした。本当に知りたいことは「いまあるデータにおけるズレ」ではなく、新たに開発された商品のように「次に得られるデータ」において、できるだけ正確に予測できる状態であるはずです。このギャップが、過学習の生じている予測モデルや AI において「実際に使い始めたときに思ったような精度でない」という残念な結果の背後にあります。

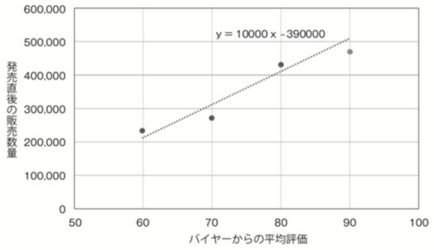
そのため、「過学習」を検証するには、とりあえず実際に動かし始めて、ある程度何度か「新しいデータを予測して実際の値とのズレを確認してみる」というのも 1 つの方法です。しかしその時点で、予測精度が不十分だったことに気づいてまた改めて開発をやり直すのでは時間がかかってしまいます。

そこで行うのがクロスバリデーション(交差検証)法と呼ばれる評価方法です。クロスバリデーション法は今扱えるデータを、予測モデル(あるいは AI)を作る「トレーニングデータ」と、その精度を検証するためのテストデータに分けて、「新しくデ

ータが来た時にどの程度の性能になるか」を検証します。予測モデルを訓練するから「トレーニング」、その精度を試すから「テスト」というわけです。なお、基本的にはトレーニングデータの方がたくさん必要なので、今回の例のように使えるデータの件数が少ない場合には「１個だけがテストデータで残りすべてがトレーニングデータ」とするのが一般的です。

試しに、さきほどの４つの製品について直線的な関係を考える単回帰分析を行うとして、その予測精度をクロスバリデーション法で評価してみましょう。バイヤーからの評価の平均点が 60〜80 点の商品だけであてはまる直線を考えて、その精度を残った 90 点の商品で見ると次のようになります。

３つのトレーニングデータから得られた「もっとも当てはまりのよい直線」は y=10,000x-390,000 というものだったので、この x の値に 90 という値を入れると 51 万個という予測値が得られ、実際の値は予測値よりも 40,000 個も少なくなっています。(図表 3-6)



図表3-6　クロスバリデーション法の考え方

３つのトレーニングデータから得られた「もっとも当てはまりのよい直線」は y=10,000x-390,000 というものだったので、この x の値に 90 という値を入れると 51 万個という予測値が得られ、実際の値は予測値よりも 40,000 個も少なくなっています。(図表 3-6)

ただしクロスバリデーション法はこれで終わりではありません。この縦に点の商品で予測し、90 点の商品で精度を検証する、という評価が「たまたまもっともうまくいったケース」だからかもしれないからです。

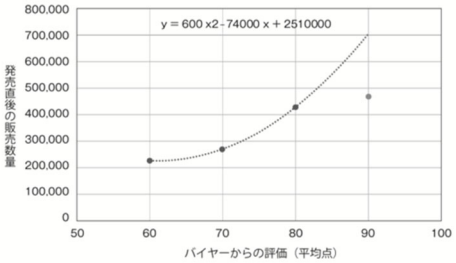
したがって、それ以外にも「80 点の商品をテストデータにして残りをトレーニングデータにする場合」「70 点の商品をテストデータにして残りをトレーニングデータにする場合」「60 点の商品をテストデータにして残りをトレーニングデータにする場合」という、考えられる４つすべての状況について同様の評価を行って、その結果をまとめると次のようになります。(図表 3-7)

テストデータの商品	テストデータの バイヤー評価（平均）	実際の値は 予想値より・・・	その他絶対値
A	60	40000	40000
B	70	-51428.6	51428.6
C	80	51428.6	51428.6
D	90	-40000	40000
ズレの絶対値の平均(平均絶対誤差:MAE)			45714.3

図表3-7　クロスバリデーション法による評価のまとめ

最初に見たバイヤーからの評価が 60〜80 点の商品をトレーニングデータとした予測モデルを、90 点の商品をテストデータとした場合には、実際の値が予測値よりも 40,000 個も少ない、という結果でした。それ以外の組み合わせについても当然こうしたズレがあるわけですが、最終的にはこれらをまとめて「ズレの絶対値を平均したもの」である平均絶対誤差(MAE:MeanAbsoluteError)を用いて、よく予測モデルの性能は評価されます。今回でいえば、プラス側に外すにせよ、マイナス側に外すにせよ、平均すればこのようなバイヤーからの評価の平均値を使った直線的な販売数量の予測を今後用いた場合、「平均して 4.6 万個ほどは前後する可能性がある」と考えて運用しなければいけません。この水準が、現行の「勘と経験に基づく予測」と比べてどの程度、適切なものかも検討する必要があるでしょう。

なお、同様にトレーニングデータとテストデータを分けた場合、「トレーニングデータにおいて予測値と実際の値のズレが 0 になる曲線を当てはめる」という方針は、次に示すようにヒドい検証結果となることがすぐにわかります。(図表 3-8)



図表3-8　クロスバリデーション法で見つける過学習の問題

このように「1 件のデータだけをテストデータとし、それ以外をトレーニングデータとして予測を行った結果どの程度ズレるかを計算する」ということを、すべてのデータが 1 度ずつテストデータとなるよう繰り返した上で、平均絶対誤差(MAE)などの形でまとめて評価する、というのが leave-one-out クロスバリデーション法です。leave-one-out とは「1 つだけ外して」という意味です。

これ以外にも「全体の 10 分の 1 をテストデータとしてそれ以外をトレーニングデータに」というやり方で、やはりすべてのデータが 1 度ずつテストデータになるよう 10 回繰り返す場合は「10 個に折りたたむ」という意味で 10-fold クロスバリデーション法と呼ばれます。

また今回は、何個売れるかという数値を予測しようとしたため、平均絶対誤差などで「いくつ外しうるか」を考えましたが、「顧客が解約するかどうか」という状態についての予測モデルもありますし、省力化のための AI の多くも「選ばれるか否か」という状態を予測します。このような場合はすべてのデータの中で、予測値と実測値がどの程度一致するかという的中率を用いて評価することになりますが、やはりクロスバリデーション法を使えば正確に評価できることに変わりはありません。

運用上注意すべき「どちらの外し方が怖いか」

このようにクロスバリデーション法を用いれば、社内や社外の技術者、あるいは自分自身で作った予測モデルや AI が、今後どの程度の精度で機能しうるかを正確に判断することができます。「データの得られた状態がこのまま続くとして」という仮定が成り立つ限り、基本的にはクロスバリデーション法から考えられる精度の範囲で予測は成立すると考えられるでしょう。

研究者ならば「これだけの精度で予測することができた」と論文に書けばよいですが、ビジネス上の判断としてはそうはいきません。ビジネスでは「平均的にいくつ外すか」「平均的にどれほど的中するか」という精度だけでなく、「どちら側に外す方が怖いか」という問題を考えなければならないからです。

先ほど商品の販売数量を予測すると、プラスマイナスそれぞれに 4 ～ 5 万個ほど外す可能性がある、という検証結果が得られました。数学的にはプラス 5 万もマイナス 5 万も同じ絶対値ですが、ビジネス上これらは同じ意味を持つでしょうか？実際の値が予測値より 5 万個多い、というとき、その商品は倉庫に保管され、消費期限が切れる前に売り切るために、営業さんが走り回ったり、追加の販促費が必要になったりするかもしれません。一方で、実際の値が予測値より 5 万個少なければ、多くの店舗で品切れが続出し、問い合わせ対応に追われ、機会損失が悔やまれるかもしれません。

このどちらが怖いかは商品自体の消費期限や、取引先との関係、自社のブランドイメージによっても大きく左右されます。同じアパレル業界でも、流行の変化が速いレディースものであれば余らせ

ることの方が怖いかもしれませんし、男性向けのビジネスウェアであれば次年度に売り切ることもできるためむしろ機会損失の方が怖いかもしれません。

したがって予測モデルが上がり、クロスバリデーション法による検証ができたら、その予測モデルに基づいて「怖い方」の可能性をある程度防ぐようにした場合、どの程度「怖い方」の問題が解決して、どの程度そのためのコストが許容されるかをシミュレーションしましょう。商品を余らせることが怖いのであれば「常に予測値より5万個少なめに生産する」という方針を取ればおおむねその心配は避けられます。一方でその商品については、最悪万個分の機会損失になるかもしれません。この分の需要を他の商品で代替できたり、品薄感がブランドイメージを向上させたりするのであれば、それも一つの戦略です。

一方で機会損失や欠品が怖いのであれば、常に予測値よりも5万個多く生産しておけばおおむねその心配は避けられます。場合により、最悪10万個ほど余らせてしまうことがあるかもしれませんが、その時点でいったん追加生産を止めて、いずれ売り切ることができるのであればそうした方法も合理的です。また、その中間あたりで「5万個多く生産して最悪10万個余るのは嫌だけど、3万個多く生産して8万個余る程度なら許容できる」という判断もあるかもしれません。

このように「この予測モデルでこれぐらいのバッファを持っていたら実際にどのようなことが起こり得るか」と考えるわけです。また、今回の例のデータでは極端に予測が外れるものはありませんが、プラス側にせよマイナス側にせよ「極端に外すケース」が、全データの中にどれぐらいの割合で含まれていて、その場合どの程度の経営上のデメリットがあるかを考えておきましょう。おそらく勘と経験で予測していた場合にも需要を読み誤るトラブルは何度もあったはずですが、そうした場合と比べて新たな予測モデルの性能は優れているといえるでしょうか？それとも「極端に外しそう」という部分だけは人間の勘で注意した方がよいでしょうか？

この「どちらに外す方が怖いか」という考えは、ある状態に該当するかしないかという目的変数を予測することが多いAIについても同様に応用することができます。「特定の病気かを診断するAI」というものがあったとして、精密検査に回すかどうかの判断で使うAIなら病気の人を見逃す方が問題になります。一方で、手術でメスを入れるかどうかという判断で使うような場合であれば、病気じゃない人を間違って病気だと誤診してしまう方が問題になります。誤って見逃す方が問題なのか、誤って混入させてしまう方が問題なのかというのは、同じ課題にフォーカスしていても、ユーザー層や使用目的によって、しばしば異なるわけです。このような場合についても、どちらの間違いを重視して、どのように運用した場合にどの程度「重視しない側の間違い」が起こりうるかシミュレーションしておくとよいでしょう。

以上のようなことが理解できれば、統計学を使っても機械学習を使っても、データに基づく予測や省力化を皆さんのビジネスの役に立てることができるはずです。

dataDiver を使った予測とその精度検証

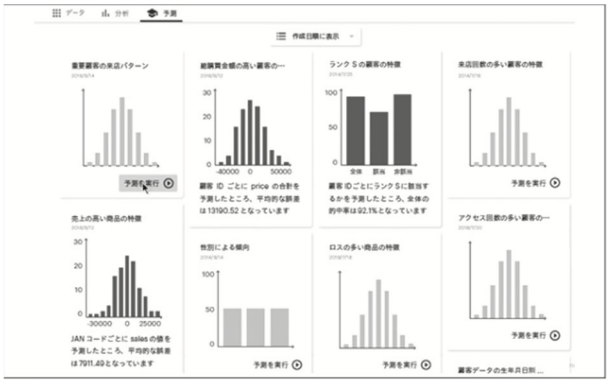
過学習を見抜くためのフェアな評価方法

「dataDiver」(データ・ダイバー)はもともと「何をしたら儲かりそうか」というビジネスの洞察をするために作られた製品ですが、その分析結果を活かすと「経験と勘」よりも高い精度の予測モデルを作ることが簡単にできます。そこで ver2.0 にメジャーバージョンアップするのに合わせて、予測モデルに関する機能を充実させました。

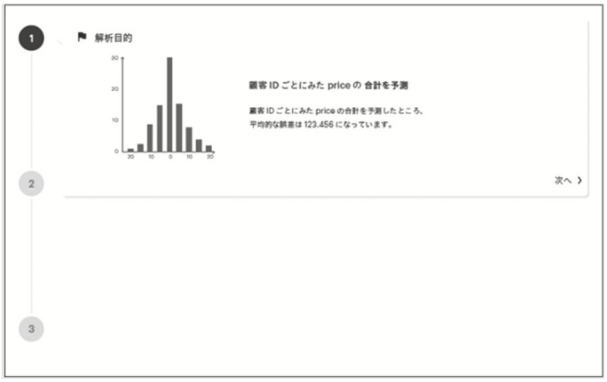
クロスバリデーション法についての説明を本章では丁寧にしましたが、実際にこれを行なおうとすると、基本的には R や Python のプログラムを書かなければいけません。外注先から予測モデルが納品される場合には「クロスバリデーション法による評価の結果を納品すること」という条件をつければよいわけですが、dataDiver を使って自分で予測を行なおうとした場合、当然予測値の算出やクロスバリデーションも自分で行なわなければなりません。そうした操作も簡単にしたことで、予測用としてもずいぶん使いやすくなりました。

コラム②で実施した作業は「分析メニュー」から行ない、結果を保存することができます。その状態で「予測メニュー」に切替えると、保存された分析結果をもとに予測を行なうことができるわけです。

予測メニュー側から過去に行なった分析結果を開くと、まずアウトカムが数値か、何かの分類に該当するか、といった条件に合わせて、10-fold クロスバリデーション法が実行され、公正にその予測精度を検証し、どの程度の平均的な誤差(MAE)や的中率だったのかという検証結果が表示されます。(dd 3-1)



この精度が満足いくものであれば、実際に「このような条件であれば予測値はいくつになるのか」とシミュレーションに使うこともできます。分析操作が終わった時点で「説明変数に対応してアウトカムがどれだけ変化するか」という係数はわかっていますが、「説明変数がいくつのとき」という条件を画面上から入力していけば、そこから予測値が算出され、その結果の複数保存を最後に一括ダウンロードできる、というわけです。また、条件をひとつ 1 つ入力するのではなく、「分析に使ったデータを差し替える形で一気に予測値を算出する」ということもできます。(dd 3-2)



これだけでも結構な精度で予測できますが、dataDiver は基本的に洞察をわかりやすくするためのツールですので、「説明変数とアウトカムの関係が一言で表現されてわかりやすいように」というような予測モデルしか作りません。したがってさまざまな説明変数が複雑に絡み合うことでアウトカムが上下する、というような関係がデータの背後に存在している状況では、そうした仕組みを含む複雑な予測モデルの方が高い精度で予測できるかもしれません。(dd 3-3)

そのため、予測メニューでは分析用に加工したデータをダウンロードできるようになっています。ここまで多様な説明変数を含む「活用のためのデータ」が仕上がっていれば、機械学習の専門家であればすぐにでも高精度な予測モデルを作り始めることができるでしょう。

彼らに依頼する際に「きちんとクロスバリデーション法で評価すること」という要請を忘れなければ、データ準備を含めたプロジェクトよりはるかに小さいコストで、正確な予測モデルを手に入れることができるはずです。