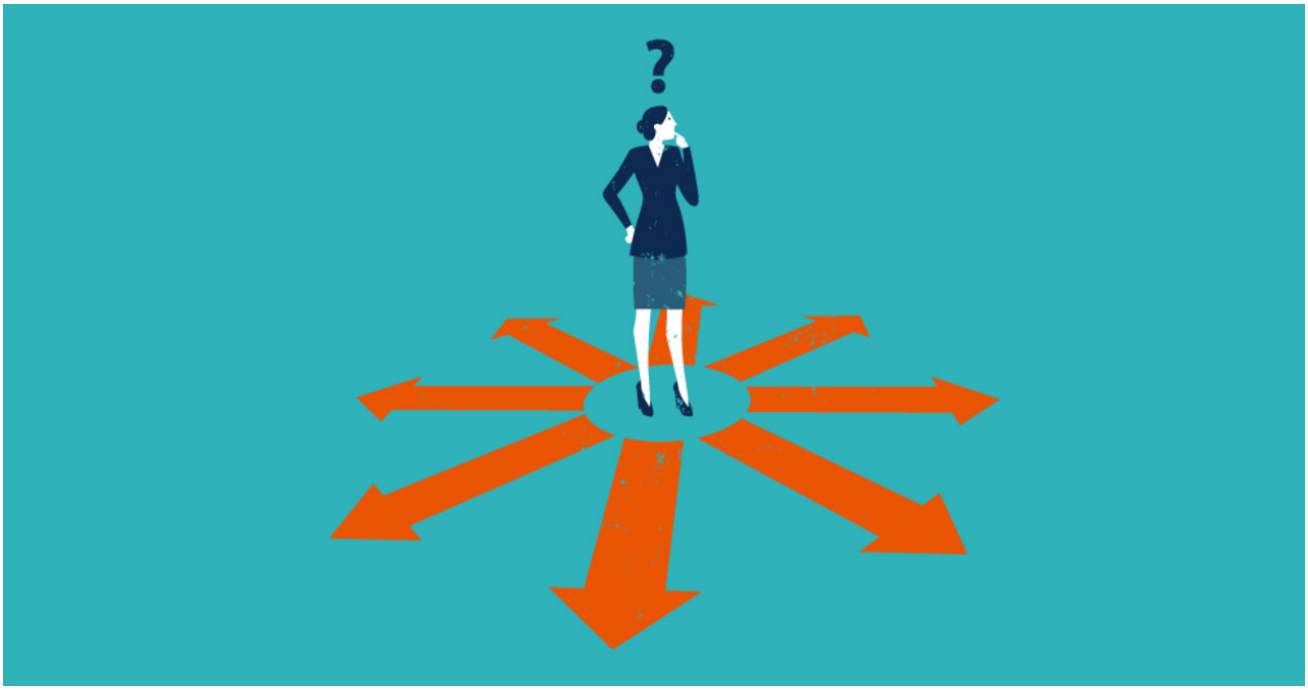
データ活用のための『リサーチデザイン』の考え方

DataVehicle

リサーチデザインの２つの要素



「何に活かしていいかわからない」問題

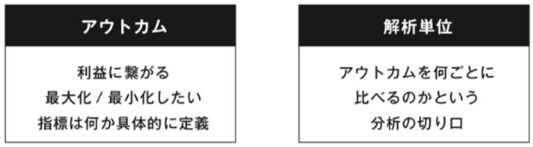
前章では、業務処理のために蓄積されたデータを、活用可能な状態に加工するためにはどのようにすればよいか、ということを中心に説明しました。そこでは、すべての項目データを完璧にいつでも活用可能な状態にすべきというわけではなく、「何にどう活用するか」という目的によって、最優先で使うべき項目もあれば、ほとんど不要な項目もあると考えられます。

冒頭で述べたように、データ分析を行なうにせよ、AI を開発するにせよ、「何にどう活用すべきか」という点については、統計学や機械学習の専門書にあまり展開されていません。「経験やセンスが大事だ」という人がいれば、「他社の先行事例を調べよう」という人もいます。しかし、経験やセンスを有していない人が、先行事例の存在しない領域でどうやってこの判断を下せばよいでしょうか?

幸い私たちはこうした相談に応えることができますが、それは素晴らしいセンスを持ち合わせているからでも、どこかの会社の事例をこっそり漏らしているからでもありません。多くの統計学や機械学習の本に「何にどう活用すべきか」という考え方が書かれていないといいまし たが、実はこの問題を教えてくれるのは別の分野の教科書です。その分野は「リサーチデザイン」と呼ばれ、もともとは研究者が、良質の研究アイデアを生み出し、研究計画を立てて実行し、よい論文を書けるような考え方で、アメリカの大学院などで教えられています。私は研究者になるべくそうした勉強をしていましたが、日本では、あまり体系的に教えられていないこの知恵は、ビジネスマンがデータ活用を考える上で、とても役に立ちます。

データ活用においてもっとも大事な
リサーチデザインの２要素

研究者として独り立ちするには、そもそも「科学とは何か」「知識とは何か」というところから論じるような、難しい本一冊分以上の勉強をする必要があります。しかし、データを活用したいビジネスマンが考えなければならないのは、データ分析であれば、「何を最大化/最小化したいのか」と「それを何毎に比べるのか」という二点だけです。私たちは前者を「アウトカム」と呼び、後者を「解析単位」と呼んでいます。(図表２-１)



図表2-1　リサーチデザインの2つのポイント

リサーチデザインの考え方では、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」ために行われるとされます。もちろんここで、この「違い」について考えることもできるわけですが、それがどれほどの意味を持つかはまた別の話でしょう。 たとえば顧客データを分析した結果、「女性は男性と比べて平均年齢が低い傾向にある」という分析結果を得て喜ぶ人はごくまれでしょう。一方で、「女性は男性と比べて客単価が高い傾向にある」という結果が得られたらどうでしょうか？１人の新規顧客を獲得した場合のメリットが、男性より女性の方が大きいのであれば、女性がよく見る広告媒体を使うとか、女性の人通りが多いエリアに出店するといったことにより、効率的に売上を増やすことができるかもしれません。

この２つの話のどこが違うかといえば、前者は性別も年齢も、それ自体を目的にしているものではありません。一方、後者における「客単価」は、その向上を目的にした業務が多くの会社内に存在しています。つまり、データ分析では業務上の「目的」になりうる成果を分析することが望まれます。この成果すなわち「どんなよいインパクトがあるのか」ということを目的とする指標のことを研究者たちはアウトカムと呼びます。

では、ビジネスにとって究極的な目的とは何でしょうか？それはたとえば「長期的に持続可能な発展を遂げる」ということが挙げられます。そのためには安定して利益をあげる、つまり売上を伸ばして、不要なコストを抑えることが求められます。マーケティング部門は売上を伸ばすこと、調達や購買の担当者はコストを減らすことが仕事です。短期的にはマーケティング担当者は「自社のブランドイメージを上げたい」と考えるかもしれませんが、これが「究極的な目的」である「持続可能な発展」に対して間接的に影響を与えることがよくあります。

ブランドイメージをよくすることによって、余計な販促費をかけたり値引きしたりせずとも、自然と高単価で売れていく、というメリットもあるかもしれません。しかし、ブランドイメージはあまりよくないのに売れている商品とか、儲かっている企業もないわけではありません。あるいは「よい」というブランドイメージが、ルイ・ヴィトンのような高級感や、アップルのような先進性ではない市場もあります。「元気で親しみやすい」というイメージが有効に働く商材においては、下手にブランディングに力を入れてしまうとむしろ売上が下がる可能性さえあります。

「目的と手段を混同しない」というのはリサーチデザインにおいて大事な考え方です。ブランドイメージの向上が、売上や利益率の改善という目的に対する手段であるならば、アウトカムは売上や利益率とすべきです。まずは「よく売れる商品とそうでない商品の違いは何か」という問いを立て

て、その答えの１つとしてブランドイメージがあるという方が適切な分析を導 きます。

同じ売上というアウトカムでも、「よく売れる商品とそうでない商品の違いは何か」「よく買ってくれる顧客とそうでない顧客の違いは何か」「よく売る従業員とそうでない従業員の違いは何か」というようにさまざまな切り口が考えられます。この「商品」、「顧客」、「従業員」というのが前述の「解析単位」であり、「活用のためのデータ」を加工する際には、必ずこの解析単位１つにつき１行ずつ、という形式のデータにしなければなりません。

業務を知らないデータサイエンティストの失敗例

このように、「究極的な目標」に近づけるような、よいアウトカムと解析単位を設定できれば望ましいのですが、上手く行かないこともしばしば起こります。私たちへの相談で一番多いのは「データはあるけれど、どうしていいかわからない」というものですが、その次の相談内容は「外部あるいは最近採用したデータサイエンティストが、何回指摘してもナンセンスな分析結果を出してくるのだが・・・」というものです。彼らの多くは統計学や機械学習を勉強してきた優秀な人なのでしょうが、そうした結果にたどり着いてしまう状況の多くは、手法以前にアウトカムや解析単位の設定が上手くいっていないことにあります。

たとえば高級デパートに出店するアパレル企業で「売上の高い顧客とそうでない顧客の違いを見つける」という分析を行った結果、「セール期間中にまとめ買いをする顧客の売上が高い」という結果が得られたとしましょう。あまりアパレル業界に明るくない分析者は、「とにかく売上が高ければよいのだろう」と考えたのでしょうが、この企業としてはまったく望んでいない分析結果です。

雑誌に広告を掲載し、高級デパートに出店するような会社にとって、「セール時にしか商品が売れない」というのは、大げさに言うとそのブランドの死を意味します。実際、大幅に値引いた商品単価では、広告費を賄えるほどの利益は出ません。春物、冬物といったそれぞれのシーズンの前半に、定価で買ってくれる人こそが彼らにとっての優良顧客ということになります。

とすれば、この会社はどのようなアウトカムを設定すべきでしょうか？セール期間に商品の値段が変動するため計算は複雑になりますが、「顧客」という解析単位を使うにしても、売上ではなく粗利を合計した「粗利総額」の方がよいアウトカムではないでしょうか。またこのような業界では「プロパー消化率」といった指標が用いられることもあります。これは生産したり、仕入れたりした商品数のうち、どれほどが定価販売かというものですが、「プロパー消化率の高い商品と低い商品の違いは何か」という分析をしてみてもよいのかもしれません。

このように、データ分析では適切なアウトカムと解析単位を設定できれば「どこから手をつけていいかわからない」という問題も、「出てきた結果がナンセンス」という問題も回避することが可能です。後で詳述しますが、機械学習で予測したり、人間の認知活動を自動化する AI を作ったりする場合にも、同様のことが言えます。

次項以降ではこのアウトカムと解析単位を、よりよいものにするコツについて考えていきましょう。

アウトカムを設定するコツ

よいアウトカムとは何か

アウトカムと解析単位という２点が適切に定まれば、「どこから手をつけていいかわからない」という問題も、「出てきた結果がナンセンス」という問題も回避することができます。これがデータを使って「よいリサーチクエスチョンを考えることができた」という状態です。

とはいえ、慣れないうちはそれが「よいアウトカム」なのかどうか、自信を持って判断するのは難しいかもしれません。そこで次に、よいアウトカムを定めるためのいくつかの考え方について学んでいきましょう。

ビジネスにおけるよいアウトカムとは次の３つを満たすものということができます。

①自社の長期的な利益に確実に寄与する
②そのアウトカムによるマネジメントでズルがしにくい
③そのアウトカムの値が同じならどんな状況でもうれしさは同じ

これらは「３つ」とはいいながら、同じことを、違う言葉で表現したものになります。それぞれについて見てみましょう。

長期的な利益に確実に寄与するか

とりあえず長期的かどうかはさておき、利益に関係するものがよいアウトカムだと知っていれば、膨大なデータの「どのあたりに注目したらよいのか」という迷いは晴れます。データの中からアウトカムの目星をつけるために最初におこなうべきことは、「何円」「何ドル」といった単位で入力されているデータがあるかないかのチェックです。

ビジネスの中で、このような通貨の単位で入力

されているデータというのは、基本的に売上かコストに関係していることがほとんどでしょう。この両者の差が利益である、と考えられるので、このあたりから「利益が大きいか小さいか」「売上が高いか低いか」「コストが高いか低いか」といった違いを見つけられる可能性があります。

マーケティングや営業系の領域であれば、粗利か売上か、という形でこのように確認するだけでいったん最初の目星がつきます。製造や物流の領域ではそこまで話が単純ではありません。この領域では、自分たちで直接的に売上に関わらず、使えるデータの中にお金に関わる項目が含まれていないこともしばしばです。

こうした場合におすすめなのは「日常的に存在する大きなコスト要因」が何かと考える方法です。たとえば材料や仕掛かりを破棄することになってしまうとか、設備が故障し手待ちが生じてしまうことはないでしょうか？あるいは、入荷や出荷の物流ミスで機会損失になってしまってはいないでしょうか？また、売上に対して過剰に製造してしまい、倉庫のコストが余計にかかったり、キャッシュフローが圧迫されたりはしないでしょうか？

「原材料を破棄することになる日とそうでない日」の違いが正確にわかれば、オペレーションを一部変更して、対策したり仕入れを絞ったりすることができます。

大まかな目安ですが、製造でも物流でも、データがきちんと活用されてこなかった分野に挑戦すると、そこから比較的短期間で数%程度のコストダウン方法が見つかるのはそれほど珍しいことではありません。つまり「あまり注意していなかったけれども、原材料の破棄で年間およそ数十億円のコストがかかっている」という状況なら数千万円のコスト削減方法が見つかる可能性があるということです。

ただし、いくらコスト自体が大きくても「日常的に存在する」ものでなければ、データによっては解決しにくい問題となります。生産拠点が災害に見舞われ、工場が倒壊したり、水没したりした場合の被害は甚大ですが、数年〜数十年に一回起こるかどうかという状況の防止や予測には、残念ながらデータはそれほど力を発揮できません。

また日本企業ではすでにいきわたり、成果が出にくいアウトカムとして、「ある程度枯れた製品」の「歩留まりをあげる」というものがあります。伝統的に統計的品質管理という考え方が根付く日本の生産現場では、戦後からこうしたテーマでの改善活動がやられてきました。それゆえに、IoT を応用して製造の現場でデータ活用しよう、という話になったときに多くの人がまず品質の向上を思いつきます。

「ある程度枯れた製品」の歩留まりがどの程度かといえば、その単位は ppm(parts per million) すなわち、100 万個中に不良品がいくつかあるかないかという水準に達していることも珍しくありません。月産 100 万個の製品の不良率が 4 ppm であったとして、それを半減させるアイデアをデータから見つけることはたいへんな困難ですが、達成したときに得られる成果は「月に 2 個だけ不良品が減る」というようなものです。社会的責任はあるにせよ、データ分析のためにかけるコストを上回るほどのビジネスメリットは得にくいかもしれません。

そのアウトカムでズルはできるか

前回示したように、まずは利益に直結するか、と

いうことを考えて通貨の単位で記録されているデータ、日常的に大きなコストがかかっているデータに着目すればアウトカムの目星がつけられます。

ただ「長期的に確実に」ついては難しいところです。それがわかれば苦労しないよと思われる方もいるかもしれませんので、この部分については「ズルができるか」という別の角度からチェックしてみましょう。

システム開発やコンサルティングを請け負う会社の営業活動についてのアウトカムを考えてみましょう。一般的に、こうした会社の営業スタッフは、担当した案件でいくらの売上をあげるか、ということで評価されることが多いですが、これは長期的な利益に繋がる適切なアウトカムでしょうか？

そうとは言い切れないことが、この「ズルができるか」というところからの視点でわかります。なぜなら、営業スタッフが売上だけで評価されるなら、「不当なダンピングを行って売上をあげる」というズルが成立してしまうからです。

たとえば、エンジニアの人件費、必要なハードウェアやソフトウェアなどの調達込みではどう考えても２千万円のコストがかかる、という案件があったとしましょう。誠実な営業スタッフは、諸々のリスクや会社の利益を考え、最低でも３千万円の金額で提案を行ないます。

一方でズルい営業スタッフは、これを１千万円で提案してしまうかもしれません。提案された側から見れば、安い分魅力的に見えるでしょう。こうしてズルい営業スタッフは１千万円の売上をあげたことになりますが、必要なリソースの半分しかもらえないのだから、会社としては赤字プロジェクトになってしまいます。

これでは長期的な利益に繋がるはずもありません。プロジェクト完了ごとにきちっと会計をしめて、どれだけの黒字を生み出せたのかをアウトカムとした方がよりズルしにくいものになるでしょう。

こうしたズルは売上側だけではなく、製造や物流側におけるオペレーションによっても起こりえます。「公平な第三者によって評価される品質」や「 正直に報告されるトラブルの件数」を アウトカムとして管理できれば生産性は向上するかもしれません。

しかし、自前で報告することが前提とされて「品質」をあげようとすれば、データを改竄して過大によく報告したり、手心を加えてもらえる人に頼んだりすることで簡単に達成できます。トラブルの報告件数を小さくするなら、バレない範囲のものをすべて隠したり、上がった報告を誰かが握りつぶすだけでも簡単に達成できます。

このようにどのようなズルが考えられて、それに対し、どうにすればもっとズルしにくくなるか、と考えることが、アウトカムのブラッシュアップに繋がります。

同じ値ならうれしさは同じか

また別の表現をすると、よいアウトカムとは「どれだけ極端に状況が違っても、得られた値が同じならうれしさも同じ」というものになります。

コストの余裕を持った見積もりで１千万 円を売上げるのと、不当なダンピングで同じ１千万円を売上げるのとでは、うれしさが異なります。この「見積もりの妥当さ」を加味して、売上を評

価するためには、最終的にはプロジェクト終了ごとに振り返って、その利益をみればいい、といえます。

また、トータル５千万円のコストがかかるプロジェクトで６千万円の売上をあげても、 ９千万円かかるプロジェクトで１億円の売上をあげても、１千万円の黒字ならうれしさは同じだろうか？と考えてみてもよいわけです。

あるいは、内部エンジニアの数が限られていて、売上の規模が大きくなればよりおおきな粗利が必要というような状況であれば、エンジニア１人月あたりの粗利額、といったアウトカムを設定した方がよいかもしれません。

人数、期間、値引き、取引の特性など、「状況の違い」としてさまざまな条件が考えられますので可能な限りたくさん、極端な状況を想定してみて、「その場合でもやはりうれしさは同じだろうか？」と考えてみると、アウトカムの思わぬ問題に気づくことがあります。

また、この「うれしさ」という点は、アウトカムの大事なポイントで、私たちが相談される際にも「このデータの中の何がどうなればうれしいですか?」と聞き返しています。

お客さんがもっとたくさん買ってくれたらうれしい、従業員がもっとたくさん売ってくれたらうれしい、設備がすぐに壊れなければうれしい、などその答えには、さまざまな可能性があると思いますが、そうした「うれしさ」をデータに基づいて、具体的に定義できるかが、アウトカムを考えるという行為です。

解析単位を決めるための４つのルール

解析単位を選ぶコツ

前節ではアウトカムを定義するためのコツを説明しましたが、ここでは解析単位について考えます。第１章のデータ整備のところで「活用のためのデータを何に対して１行ずつにするか」を述べましたが、これが「解析単位は何にするか」と考えることと同じことです。

データ整備の説明の時にはどのような切り口にしたいか考えましょう、と書きましたが、実はこれにも明確なルールが存在しています。第１章で述べた特徴と重複もありますが、「解析単位」を決める時には次の４つに気をつけましょう。

① 最終的に最低でも数十行以上になるものであること
② 数十行ごとの違いは「自明」ではないこと
③ 今あるデータから可能な限りたくさんの特徴が考えられること
④ 最終的に何かを「変える」ことができること

ではそれぞれについて確認していきましょう

なぜ数十以上の解析単位が必要か

１つめについては、すでに述べた、活用のためのデータの特徴というところに由来しています。データ分析にせよ、AI に利用するにせよ、最低限数十行はなければ役に立たない、という話です。たとえば東京と大阪の２カ所にしか営業所が存在していない企業において、「業績がよい営業所とよくない営業所の違いはどこにあるか」と考えることはできません。もちろん東京の方がよいとか、大阪の方がよいことはわかりますが、両者の間に存在するさまざまな違いのうち、「何がそこに関係しているか」「何のデータから判断できるか」をデータから判断することはできないからです。これが、数十の事業所があれば「高業績の事業所ばかりに共通した条件がある」

とか「ある値が高ければ高いほど業績は高い傾向にある」という特徴をデータの中から見つけることができます。それは「たまたま共通しているだけ」という話なのか、「たまたまだけではこんな共通点は存在しない」という話なのか、データに基づいて判断できる、というのが現代的な統計学によるデータ分析や、統計的機械学習と呼ばれるものの考え方です。

この考え方がわかっていれば、「性別」という解析単位があり得ないことも判断できるでしょう。この場合も、２つしか事業所がない場合と同様に、男性と女性のどちらが優良顧客そうか、ということはわかっても、データから、「その間のどのような違いによってこの差が生まれているのか」ということを判断できないからです。

解析単位の自明な違いとは

年齢、という解析単位はどうでしょうか？年齢なら 10 代の若者から 70 代以上の高齢者まで、１歳刻みに集計していくと、数十行以上のデータは得られます。このようなデータにおいても「18 歳の顧客よりも 65 歳の顧客の方がたくさん購買してくれる」といった集計は行えますが、これは意味を持つでしょうか？

それは、ここまでの説明で述べたように「たくさん購買してくれる顧客の年齢と、そうでない顧客の年齢の間にどのような違いがあるか？」というような分析に意味があるか、という話ですが、その答えは No でしょう。購買する日時や商品の傾向という、それ以外のデータで 「年齢の間の違い」を表現するまでもなく、18 歳と 65 歳の間には明らかな違いが存在しています。データの中にはそんな項目は含まれていませんが、たとえば 18 歳の方が平均的には独身率が高く、学生の割合が大きく、体力や食欲があることでしょう。逆に 65 歳の方は金融資産をたくさん持っている、ということがいえるかもしれません。

このような状況で、他のデータからさまざまな説明変数や特徴量を考えて「年齢ごとに異なる購買金額の違いを説明したり、予測したりするものを探す」というのはナンセンスです。65 歳の顧客は 18 歳の若者よりもよくお酒を買っているかもしれません。しかし、「お酒を買う頻度の多さ」という違い以前の問題として、18 歳と 65 歳は「自明な違い」を持っています。果たして「お酒を買う頻度の多さ」が優良顧客度合いに関係しているのか、そもそもの「年齢が高いかどうか」という違いが関係しているのか、データからではどうやっても判断できません。

これは顧客の属性についてだけではなく、たとえば「商品ジャンル」といったものについても同様のことがいえます。少し大規模な小売店になると、扱っている商品ジャンルも野菜や精肉、加工食品、洗剤、下着、食器など、数十以上にわたるでしょう。しかし「よく売れる商品ジャンルとそうでない商品ジャンルの違いは何か？」と考えることもやはり意味はありません。お総菜の販売金額は多いが、文房具の販売金額は少ないという状況で、前者が「購買者の平均年齢が高い」とわかったとしても、これが販売金額の差と関係しているのか、そもそものお総菜と文房具の間の「自明な違い」が関係しているのか、これもやはり判断がつかないからです。

年齢は「顧客の一つの属性」であり、商品ジャンルは「商品の一つの属性」であると考えることもできます。このような属性の一つを解析単位にするのではなく、あくまで「顧客」あるいは「商品」を解析単位として選ぶのがよいでしょう。「この顧客とあの顧客の間にどのような違いがあるか」「この商品とその商品の間にどのような

違いがあるか」というのはいくらでも、 自明でない違いが考えられることができます。

今あるデータから考えられる特徴の種類

データ上で「自明ではない違いがある」ということは「理論上考えられる」というだけでは不十分です。顧客なら顧客、商品なら商品で、理論上さまざまな違いが考えられるでしょうが、 そうした違いを表現するようなデータは入手可能でしょうか？

たとえば顧客については単純な ID POS のデータでも、登録された属性や、過去のよく買い物に来る時間帯や曜日、といった情報を「顧客の特徴」として考えることができますが、 もし商品を解析単位にしようとすれば、その商品が「何のジャンルの商品か」とかいう特徴がわかっていなければ、分析結果や予測精度は不十分なものになってしまうでしょう。「ジャンルかはわからないけど、深夜に売れがちなものがよく売れている」とか「ジャンルかはわからないけど、女性に支持されている商品がよく売れている」というのは、なんとも気持ち悪い分析結果ですし、それがお酒なのか、お菓子なのかといった商品ジャンルの情報をつけた方が予測精度も高くなるはずです。

またサービス業において、顧客の購買履歴の中に「接客した従業員の ID」データが含まれることがしばしばあります。レジを打つ際に、胸につけた ID カードのバーコードを読み取るシーンを目にした方もいるでしょう。販売スタッフが数十人以上いる会社であれば、この ID のデータを使って「たくさん販売している従業員とそうでない従業員の違いはどこにある か？」というアウトカムと解析単位を定めて、従業員 1 人 1 行ずつの「活用のためのデータ」へ加工することも可能でしょう。従業員についても、性別や年齢、勤続年数、持っている資格や過去に受けた教育、心理的な特性など、さまざまな違いが考えられるはずです。

しかし、こうしたデータが入手できないのであれば、この考え方は絵に描いた餅です。理論上そうであったとしても、紙の履歴書がオフィスのどこかに保存されているだけ、という状況であれば、データを入力して、レジのバーコードと同じ ID を入力して、といった手間をかける必要があります。

それだけの価値があるのなら行なうべきですが、それほど明確な意義があるわけでもない、というのであれば、「現時点のデータにおいて特徴を表現できる項目がよくそろっている」という解析単位から着手してみるとよいでしょう。データ整備のところでも述べたように、「できそうなところからやってみる」というのが基本です。

解析単位に対して「変える」アクションとは
解析単位を決めるために注意する最後の項目は、最終的に出てきた分析結果に基づいて、アクションを取れるのかどうか、「変えられるか」ということです。分析結果が出た後で、基本的には何かを「変える」アクションをしなければデータ分析の結果は活かせません。変えるものとは「解析単位の状態」や「リソースの配分の仕方」です。たとえば顧客を解析単位とした分析から、「精肉を買っている人が優良顧客である可能性が高い」とか、「土曜日の午前中に来店している人が優良顧客である可能性が高い」といった結果が得られたとしましょう。

この場合、「解析単位の状態を変えるアクション」とは、今いる顧客に対して「精肉を試しに買わせてみる」とか、「試しに土日の午前中に来店させ

てみる」ということに該当します。具体的には、精肉の値引きを行なったり、魅力的な商品を充実させたり、一部の顧客に優待クーポンを発行したり、あるいは土日の午前中にタイムセールを行なうことなどです。こうした誘導を行なうことで、これまでその商品を、あるいはその時間帯に買ったことのない人を「買ったことのある人」に変えることができるかもしれません。その結果、その人の購買パターンが変化し、優良顧客になるのではないか、というアクションが考えられます。

あるいは、同じ分析結果に対して「リソースの配分を変えるアクション」も考えられます。「精肉を買うかどうか」で優良顧客の度合いが変わるというより「精肉を買う家庭かどうか」で優良顧客かどうかが異なる可能性も考えられるかもしれません。このことは食べ盛りの男の子がいる家庭が自社にとっての優良顧客であるという状況を示しており、ほとんど精肉を買わないような家庭に「肉を買わせる」という変化をもたらしても優良顧客度合いに変化はない、と考えることもできます。このような場合、「食べ盛りの男の子がいるような家庭」に向けて重点的にその対象となるメディアに広告を展開することでマーケティングが効率化される可能性があります。具体的には、地域のスポーツ施設に広告を出したり、精肉売り場で男の子に受けがよいメニューのレシピを配布したり、部活帰りの学生に、チラシを持ち帰らせたり、といったことが考えられます。

このような展開が解析単位の状態か、あるいはそれに基づくリソース配分を「変える」というアクションの考え方です。解析単位を設定するにあたっては、そうした変化をもたらすことができるかもよく考えておきましょう。

前述した販売スタッフを解析単位とした分析で、たとえば「特定の資格を持っていると販売成績が高い」といった結果が出たとしましょう。この場合、「状態を変える」方向では、既存のスタッフに対してその資格を取るよう補助を出したり、勤務内での資格取得の勉強を奨励したりするアクションが考えられます。「リソース配分を変える」方向では、その資格が取れる学校に重点的に求人を出す、あるいは有資格者により高い報酬を約束するなどの方法が考えられます。このようなアクションは人事や営業管理の責任者、あるいは経営責任者が了承すれば実行できますが、果たしてそれは現実的でしょうか？こうした権限が自分になく、また責任者も分析の結果から新しいことを試したがるようでなければ、「変える」アクションが可能な、別の解析単位にした方がよいかもしれません。

分析前にこのような視点を持つことができれば、「分析したのに何の役にも立たなかった」というリスクが回避できます。

仮説ではなく解析単位を考える理由

以上が解析単位の考え方ですが、こうしたやり方に違和感を持たれる方もいるかもしれません。BI ツールやエクセルなどで「男女間で購買金額に差があるか」とか「年代によって購買金額に差があるか」といったグラフを描くことが分析だと思っている方は、なぜ「解析単位」や「アウトカム」という概念を導入するのか、と疑問を持たれてしまうかもしれません。

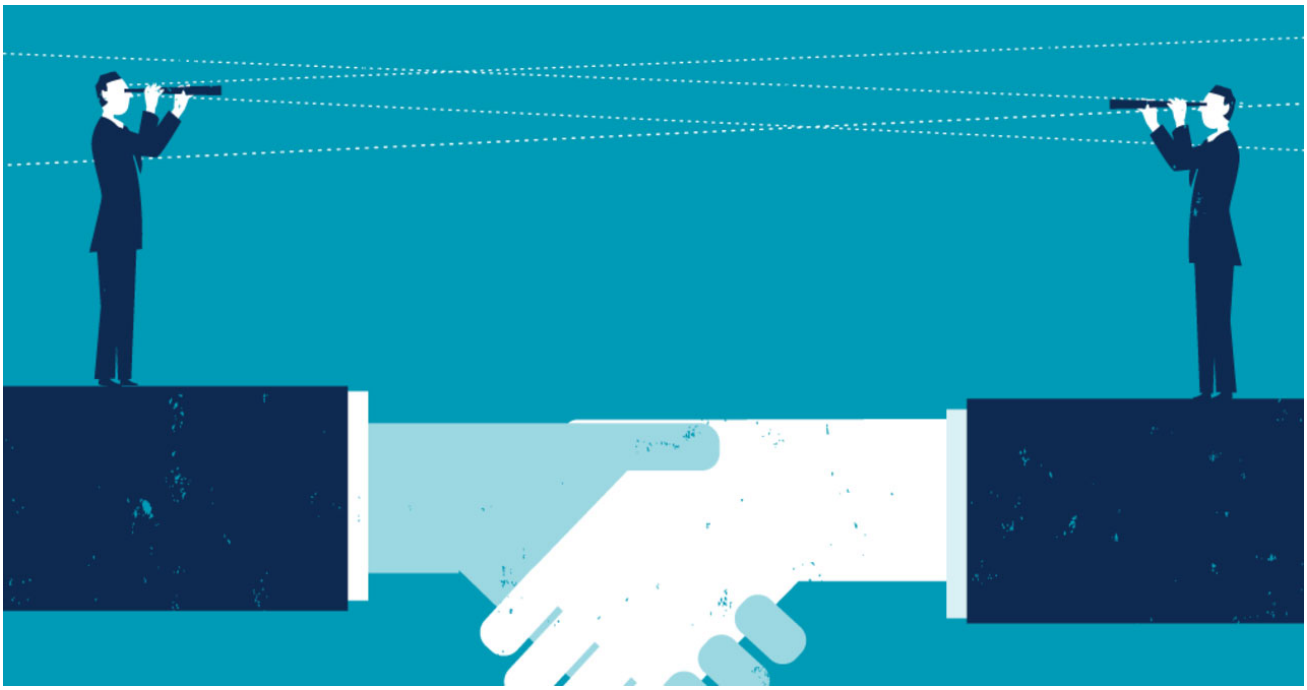
「男女間で購買金額に差がある」「年代によって購買金額に差がある」といった設定は主語と述語を含む仮説で、Yes か No かで答えられるものです。それも、かなりベタな仮説です。私たちには高額な予算をかけて BI ツールを導入し、どん

な項目でも見える化できるようになったものの、「いったい何を見ていいかわからない」という相談もしばしば寄せられます。相談者の多くはせっかくの BI ツールを目の前にして、男女別とか、年代別とか、月別とかの指標で売上を確認するぐらいのアイデアしか浮かばないのだそうです。「仮説を考えるセンスで困る」という状況なのでしょう。

解析単位とアウトカムを定める考えでは「Yes か No かで答えられない問い」を立てます。たとえば解析単位が顧客で、アウトカムが購買金額なら「購買金額の多い顧客とそうでない顧客の間にはどのような違いがあるのか」という問いを考えることになりますが、これは Yes か No かではなく、いくらでも答え方があります。

その解析単位同士の違いは、データから網羅的にあげることができるでしょう。性別、年代といった顧客マスターに含まれる情報はもちろん、顧客と紐付く購買履歴から、「過去の購買に占める日曜日の割合」というような、さまざまな顧客の特徴を考えることができるということを前章のデータ整備ですでに述べました。このように考えられ得るあらゆる特徴の中から、 何がどれくらい関係しているかを探すことがデータ分析の仕事です。それによって、よい仮説が思いつけない人でも、データからアウトカムつまり「利益につながる」重要な要因が何か、を発見することができます。解析単位とアウトカムを採用する大きな理由は、「的確な問い」 を考えるためなのです。

何がその違いと関係しているのか
～基本的なデータ分析の読み方



効率的な「違いの見つけ方」

アウトカムと解析単位が決まり、データから考え得る限りさまざまな解析単位ごとの説明変数を加工することができたら、いよいよ分析に入りましょう。

再述しますが、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」のような、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。

基本的な考え方としては、網羅的に考えられた説明変数のそれぞれと、アウトカムとの間に関連があるのかないのか、あるとすればどの程度の関連があるのかということを探していくことになります。ただし、BI ツールやエクセルで片端から「説明変数を横軸に」「アウトカムを縦軸に」と集計すればよいというものではありません。このような考え方には大きく分けて３つの問題が存在しています。

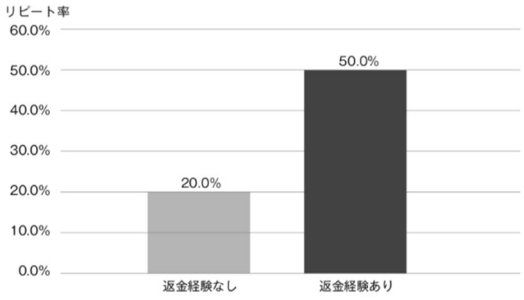
①その関連性が「たまたまの差」なのかどうか判断できない
②複数の説明変数が絡んでくるような関連性について判断できない
③単純にとんでもない作業量がかかる

こうした問題を解決する便利な統計手法もすでに発明されているので最終的には心配する必要はありませんが、BI ツールでの単純な集計で満足してしまわないよう、それぞれ見ていきましょう。

その差は「たまたまの差」なのか

たとえばあなたの会社で既存顧客の離反率に対して悩んでいたとしましょう。毎月、全体のうち約%もの顧客が離反してしまうため、新規顧客を獲得するコストはその後の売上によっても効率よくペイすることができません。そんな状況で、「離反してしまう顧客とそうでない顧客の違いはどこにあるか」と分析するのはとても合理的です。

そこであなたは BI ツールを駆使し、さまざまな説明変数を横軸にした膨大な数のグラフを描画したところ、次に示すように「返金処理の経験がある顧客の離反率が高い」という結果が得られました。(図表 2-2)



図表2-2　BIツールから得られたグラフの例

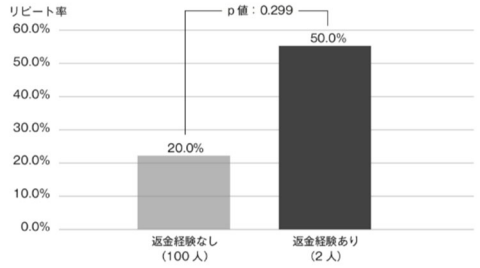
過去 1 か月間のデータを使って、それ以外の顧客では 20%の離反率であるのに対し、返金処理の経験がある顧客では 50%もの離反率が示されています。ここから、「返金処理の対応に何か不満を持たれているのではないか」とか、「そもそも返金を頼みたくなるような商品とは何だったのか」という侃々諤々の議論もできますが、その前にするべきことがあります。

それは、「たまたまの差」なのではないかと考えることです。これが、全体で 102 人のデータで、そのうち返金処理の経験者が 1 人しかいなかったとすればどうでしょうか？ 2 人中 1 人がたまたま離反していただけなら、「たまたま」という気もします。返金経験者が 2 人しかいないのならば、得られる割合は 0 %、50%、100%のいずれかの値にしかなりません。20%ちょうどという値にならない以上、たまたま 2 人とも離反してなければ「なぜか離反率が低い！」というグラフが見えていたでしょう。

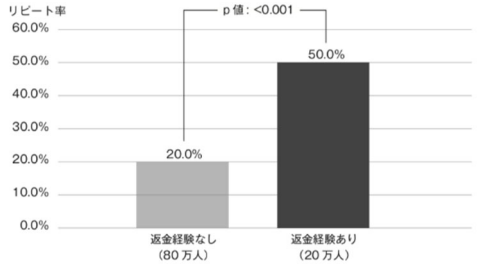
しかし、100 万人の顧客のうち 20 万人が返金処理をしていた、という状況ではどうでしょうか？つまり、返金処理を経験していない 80 万人中の 20%で 16 万人が離反し、経験している 20 万人中その半分の 10 万人が離反している、という状況です。さすがに何万人もの顧客が「たまたま」離反しているとは考えにくいはずです。

このように、単純に割合や平均値を集計して比較しただけでは、「たまたまの差」かどうかという判断ができません。統計学では「p 値」という「本来まったく差がなかった状態で、このような(あるいはそれ以上に差がつく)分析結果が得られる確率はどれほどあるのか」を考えます。この確率が低くければ「たまたまとは言い難い」と考えられますし、高ければ「たまたまという可能性が捨てきれないからあまり気にしない方がいい」と考えられます。

実際に両者のケースについて「p 値」を計算してみると次のようになります。(図表 2-3、2-4)



図表2-3　解析単位が少ない場合の p 値の例



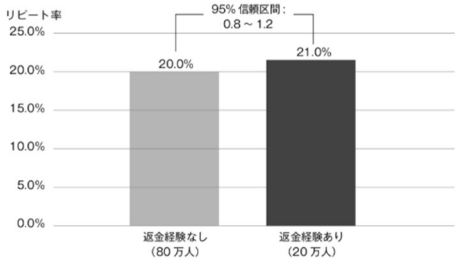
図表2-4　解析単位が多い場合の p 値の例

つまり、102 人中 2 人の返金経験者だけで 50%という高い離反率が出た、という程度の差はまったくの偶然でも 0.299 という高い確率で得られるということです。「3 割バッターがヒットを打った」といわれても誰も驚かないのと同様に、この値がまったくの「たまたま」で得られたといわれても否定することはできません。一方で、100 万人中 20 万人の返金経験者がこれほど高い離反率を示すという結果が「たまたま」得られる確率は 0.1%もありません。そのような奇跡が起こっただけです、という説明にはムリがあります。

統計学では、両者の間に差があるのかどうか、という仮説を検証するために p 値を計算する仮説検定を行ないます。なお、慣例的には 5 %未満すなわち「20 回に 1 回も得られない」ような結果が出れば、その差は「たまたま」とは考えにくいクリアなものだと考えます。

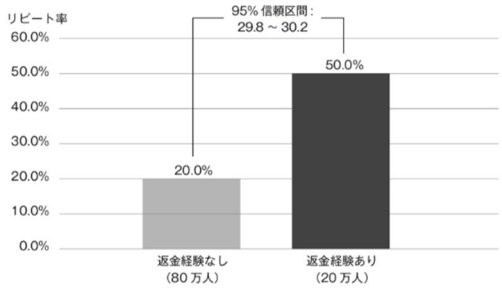
ただし、数十万人同士で比べれば、かなり小さな差であっても「たまたま」とはいい難いという結果が得られることもあります。たとえば同じように 100 万人中 20 万人の返金経験者のうち、ちょうど 21%(4.2 万人)が離反していて、残り 80 万人の非経験者では 20%(16 万人)が離反していたとしましょう。わずか 1 %ポイントの差しかないため、実用上はあまり気にならないかもしれません。しかし、仮に両グループの間にまったく差がなかったのだとすれば「たまたま返金経験者にこれほど離反者が偏る」確率はとても低く、p 値はやはり 0.1%未満です。

したがって p 値で「たまたまの差かどうか」を判断するだけではなく、「たまたまのバラツキなどを考慮した上で、だいたいどの程度の差なのか」を確認した方がよい場合もあります。「95%信頼区間」という指標を使えば、次図のように、いくつからいくつまでの値だと考えるのが妥当かという目安を示せます。今回の状況であればこの 1 %ポイントの差はたまたまのバラツキなどを考慮すると「おおむね 0.8〜1.2 ポイントと考えるのが妥当」と判断できます。(図表 2-5)



図表2-5　解析単位は多いが「わずかな差」の95%信頼区間

なお、同様に先ほどの「20 万人の返金経験者のうち 10 万人が離反」といった状況に対する 95%信頼区間を示すと、「返金経験者の方が 29.8〜30.2 ポイントも離反率が高い」という結果になります。(図表 2-6)



図表2-6　先ほどの例における95%信頼区間

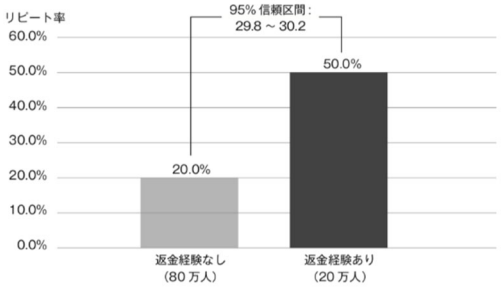
「たまたまの差」であれば、そこからのアクションはムダになってしまうかもしれませんが、統計学の見方を身につければ、そうしたリスクは回避できます。また第一章で「ごくわずかの解析単位しか該当しない」分け方は避けるべきだと述べましたが、それは、多少アウトカムに差がつこうとも、「p 値が大きくてたまたまという可能性が捨てきれない」という結果にしかならないためであることが、改めて理解してもらえたのではないでしょうか。

他の説明変数が絡んだ関係

「p 値」や「95%信頼区間」といったデータの見方を理解できれば「たまたまの差」に惑わされることはなくなります。それでも BI ツールなどで描くグラフの多くは、１つの説明変数と１つのアウトカムの間の関係を２次元的にしか把握できません。立体的なグラフィックを使ったり、色分けをしたり、さまざまな手を尽くして複数の説明変数をグラフ上に同時に表示できる機能を持った BI ツールもありますが、今度は「ごちゃごちゃしてよくわからない」ものになりがちです。

また、「過去にクーポンつきのダイレクトメール(DM)を受け取ったことのある顧客はその後のリピート率が高い」というグラフが得られたとしても、即座に「もっとたくさんクーポンを送ってリピート率をあげよう！」というアクションを起こすべきとはなりません。なぜなら「特定の属性の顧客によくクーポンを送っている」という実態があるのだとすれば、その条件も考慮しなければならないからです。

たとえば、「首都圏在住の顧客には最近高い割合でクーポンを送っている」という事情があれば、顧客全体で「クーポンの有無別にリピート率を比較する」だけでなく、顧客全体を首都圏に住んでいるかいないかも分けた上で同様の集計を行なった結果を見ると、次のようなケースも考えられます。(図表 2-7)



図表2-6　先ほどの例における95%信頼区間

つまり、全体ではクーポンを受け取っていない人のリピート率は 18.0%で、受け取っている場合は 28.9%とこちらの方が高い結果になりますが、首都圏に住んでいればクーポンの有無にかかわらずどちらのリピート率も 40.0%です。首都圏以外の地域に住む人もクーポンを受け取っていようがいまいが、同じく 15.0%というリピート率を示しています。このような状況では「クーポンを送ればリピート率をあげられる！」とは考えにくいでしょう。単に、全体で見た「クーポンを受け取った顧客」のグループには相対的にリピート率の高い首都圏の顧客が含まれていただけ、と解釈する方が妥当です。BI ツールのグラフだ

けを見て、焦ってクーポンを送っても、おそらくリピート率の向上は望めなかったでしょう。

このように、「何らかの説明変数の違いによって、アウトカムにどれだけ差があるか」を考えるためには、興味のある説明変数以外の条件をできるだけそろえなければいけません。この場合、首都圏に住んでいるかどうかで全体のデータをグループ分けしてから分析しましたが、このようなやり方を「サブグループ解析」と呼びます。

ただ、本書で考えてきたような、「活用のためのデータ」を用意すれば、ちょっとした業務のデータから、数百個以上もの説明変数を挙げることができます。これらの１つ１つを「サブグループ解析」して、確認するというのは現実的ではありません。数百項目を、ざっくり 10 個程度のサブグループに分けたとしても、数千枚のグラフを確認して、今回の例のように首都圏在住かどうか、という条件のような「サブグループ分けした時に差が消えるようなグラフはないか」と考えるのは不可能といっても過言ではありません。

サブグループ解析にかかる手間の
膨大さを避ける多変量解析

「多変量解析」と呼ばれる統計学の手法を使えば、このようなサブグループ解析の手間を解決することができます。具体的には、アウトカムが売上などの数値であれば重回帰分析、アウトカムが「リピートするか離反するか」といったように２つに分かれるようなものであればロジスティック回帰分析といったものを用いることができます。

これらの手法は「横軸に１つの説明変数を考えて、縦軸にアウトカムを考えて」といった１対１の関係ではなく、「複数の説明変数とアウトカムの関係を一気に説明する」というものです。それぞれの説明変数とアウトカムの関係性は、「それ以外の説明変数の条件が一定だったとして」という条件の上で示されます。先ほどの、「首都圏に居住しているかどうか」と、「クーポンを受け取ったかどうか」という２つの説明変数を同時に使って、リピート率との関係を分析すると結果は次のようになります。(図表 2-8)

説明変数	オッズ比	95%信頼区間			p値
首都圏在住	3.78	2.58	–	5.54	<0.001
クーポンあり	1.00	0.65	–	1.53	1.000

図表2-8　ロジスティック回帰分析の結果

ロジスティック回帰分析の結果は「オッズ比」という指標で示されます。これが 1 より大きいということは「該当しやすい」、小さいということは逆に「該当しにくい」ということを示します。たとえば「首都圏に在住している」という説明変数の状態を取る場合、１より大きい 3.78 という値なので、首都圏在住の顧客はサブグループ解析で確認した際と同様に、「リピートしやすい」傾向にあるといえるでしょう。また、このオッズ比においても 95%信頼区間と p 値を考えることができます。分析する顧客の人数が少なければ、データのばらつきによって「たまたま差がつく」すなわち「たまたま１より大きなオッズ比が計算される」という可能性もなくはありませんが、95%信頼区間の値を見れば大まかにこのオッズ比は 2.58~5.54 という間にあると考えればよさそうです。したがって p 値をみると、本来であれば首都圏在住者かどうかとリピート率に差がなかった場合に、たまたまこれほど(あるいはもっと強い)関連性が見られる確率は 0.1%もないということがわかります。

しかし「首都圏在住かどうか」という条件をそろえた上で、「クーポンの有無がリピート率に関係

するか」という結果をみると、そのオッズ比はちょうど 1 です。これはクーポンを受け取っていようがいまいが、リピート率が高いわけでも低いわけでもない、という結果を示しています。オッズ比が 1 というのは関連性の強さでいえば「全然関連していない」というもっとも弱い結果ですので、p 値も 1.000 と、確率としてもっとも大きな値になります。

本冊子は「統計学の教科書」ではありませんので、ロジスティック回帰の内容に興味のある方は拙著『統計学が最強の学問である 実践編』などを参照いただければ幸いですが、こうしたロジスティック回帰によって、サブグループ解析と同じように「首都圏在住者のリピート率は高い」「首都圏在住かどうかの条件をそろえるとクーポンの有無はまったく関係ない」という結果が得られることが理解いただけると思います。

ここから後は、説明変数が数百個になろうが、重回帰分析やロジスティック回帰分析といった手法を使えば大丈夫です。あたかもそれらすべての説明変数でサブグループ解析をやったかのように、「すべての説明変数の条件を互いにそろえ合った上で、それぞれの説明変数がどれほどアウトカムに関連しているか」を考えることができるでしょう。

ただし、考えた説明変数すべてを用いるべきかというとそういうわけではありません。今回の例の「クーポンを受け取ったかどうか」というように、まったくアウトカムと関係ない説明変数を考慮する必要はないでしょう。今回のように「まったく差がつかない」という場合ならまだしも、「まったく差がつかないはずなのに、たまたまちょっとだけ差がついてしまった」という状況が困りものです。「たまたまついた差」の影響を考慮して他の説明変数とアウトカムの関係性を考えても、それはあまり今後の役に立たないのです。次に同じデータを同じように分析したとして、この「たまたまついた差」は、もっと大きくなったり、小さくなったりと変化してしまうからです。

統計学はもちろんこうした事態に対処するための手法も発明しています。それはスパースモデリングという考え方です。この手法を使えば、同じロジスティック回帰分析や重回帰分析といった手法を使うにせよ、説明変数の候補がたくさんある中から、互いに条件をそろえる必要がある、アウトカムと明確な関係を持つような説明変数だけを使う組み合わせを自動的に考えることができます。このスパースモデリングの考え方は AI のための機械学習技術にも応用され、ディープラーニングを用いる中で、必要な項目だけを自動的に取捨選択することで、計算量を抑えつつ精度を向上させるという研究が存在しています。

実は本冊子が、業務のためのデータから可能な限り多くの説明変数や特徴量の候補を考えよう、という立場なのは、スパースモデリングの存在によるところが大きいです。「仮説を考える」という頭の使い方によって人間の可能性を狭めるよりも、可能性を広げることに人間の頭を使った方が、人間とコンピュータとの作業分担としては賢明です。コンピュータはまだ「可能性を広げる」ことは得意ではありませんが、スパースモデリングを使えば「必要なものを自動的に取捨選択する」という作業は高速かつ正確に行なうことができます。

このようにして得られた分析結果をもとに、「解析単位の状態を変える」または「リソースの配分を変える」というアクションが取れれば、ビジネスの中で大きな価値を生むことになるでしょう。

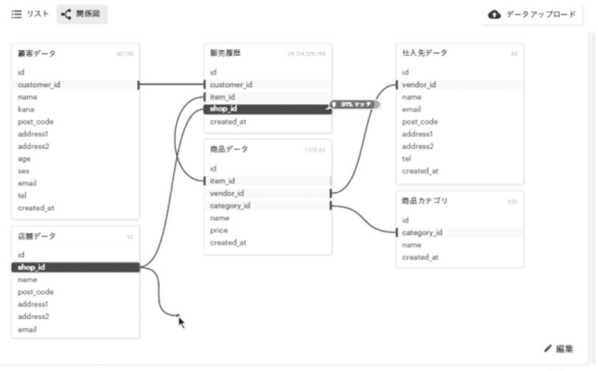
dataDiver を使った洞察を探す分析

ここまで「どのように分析の方針を立てるか」というリサーチデザインの考え方を説明しました。統計学や機械学習の手法に精通しても、この部分でうまく行かないと説得力のある洞察は得られません。

一方で、ここまでのことが理解されれば、データ分析自体はかなりのところまで自動化できる、というのが私たちの考えです。この考え方を実現するために作ったのが「dataDiver」（データ・ダイバー）という製品です。

dataDiver は BI ツールとも違い、AI というわけでもありません。創業以来数年、dataDiver の適切なカテゴリーが世間に存在していなかったのですが、最近のガートナーによれば「拡張アナリティクス」というものになるようです。彼らのいう拡張アナリティクスとは、データを準備し、洞察を獲得するプロセスを自動化し、企業内の意志決定や行動を最適化するものであり、さまざまな状況で専門的なデータサイエンティストを不要にするそうです。

弊社の dataDiver はまさにそうしたコンセプトで作られており、他の分析ツールにはない独自の機能として「関係図」という概念が存在しています。



画面上でデータの項目をドラッグすると「どのキーとどのキーが対応しているか」を簡単に設定することができます。このキーの対応関係についての情報と、データの中身が文字か、数値か、

日付時刻か、という情報を読み取ることで、「顧客ごと」「店舗ごと」「商品ごと」などのさまざまな解析単位ごとに１行ずつのデータを自動的に生成し、その中には私たちの引き出しにある、あらゆる形式で加工された「一言で言って中身がわかる範囲の説明変数」が含まれます。

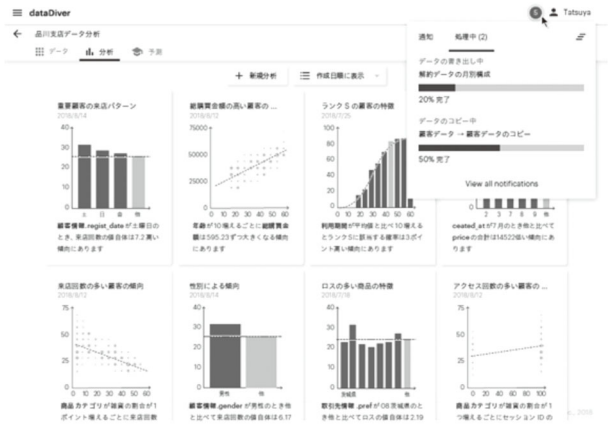
このようにデータの加工が自動化されているため、「解析単位とアウトカムを指定する」だけで、dataDiver における分析操作が実行できます。顧客ごとに過去の購買金額を合算したライフタイムバリューを左右する説明変数を探したければ、まず「顧客ＩＤごとに」と指定し、そこから紐付く「購買金額（購買履歴の中の price という項目）」の「合計」が「少ないことが課題」と入力すれば、すぐにデータの自動加工と分析がスタートします。この「合計」や「件数」といった集計方法も、データ間の関係性から自動的に意味のある選択肢が提示されます。

かなり大規模なデータであっても数分程度の時間で、加工した説明変数の中から最適な組み合わせを探索し、何がどれくらい効いているのか、という結果を日本語で教えてくれます。もちろん機械的な基準で「もっとも説明力の高い説明変数の組み合わせ」を探索していますので、「当たり前すぎる」分析結果が得られてしまうかもしれません。このような場合、右側の「除外」ボタンを押して再計算をすれば、そうした説明変数がデータに含まれていなかったものとして探索をやり直します。

データサイエンティストが社内にいたとしても、このような見直しにはせいぜい翌週、遅ければ「来月の定例会議までに」というスピード感の時間がかかってしまいますが、dataDiver を使えば、数分程度で見直せますので、会議の最中にリアルタイムで分析することも不可能ではありません。

こうした試行錯誤を繰り返せば、最終的に納得感のある興味深い洞察が得ることができます。そこからは、「変える」「ずらす」という方向でどのようなアクションが考えられるかを議論しましょう。

もちろんこれらの結果について視覚的に「どういう状態になっているか」というグラフを確認でき、分析結果を社内で共有するのも簡単です。クラウドあるいは社内のオンプレミス環境であっても、ネットワークに接続されている利点を活かし、ブラウザさえあれば分析結果をすぐに見ることができます。また会議資料用に、分析結果を PowerPoint や Excel 形式で書き出すこともできます。



このように dataDiver は、データ準備や分析、視覚化のかなりの部分を自動化するだけでなく、さらに会議資料作成といったコミュニケーションの部分まで強力にサポートします。ぜひこちらの成果を活かして、リサーチデザインとそこからのアクションという本当に価値を生む仕事に集中していただきたいと思います。