2022 年 6 月
ReCent
Actuarial News

従来の方法では見つけられなかった
インサイトを生み出す。

hr | bluebox：機械学習に関する実用的なインサイト

先進的なデータ分析法が、将来のビジネスの進め方を方向づけると、業界では広く理解されています。しかし、（再）保険業者の多くは、データ分析の専門知識の恩恵にあずかる方法について、その実例を持ちません。ハノーバー・リーの hr | bluebox サービスは、熟練したデータサイエンティストが機械学習アルゴリズムを応用し、それによってクライアントのポートフォリオの早期解約要因を検出します。

何をしているのか？

この無料のサービスにより、当社は、顕著に早期解約の挙動が見られるポートフォリオ・セグメントを特定し、その特徴を明らかにすることを目指しています。それにより、解約の可能性が高い保険契約者や解約の可能性が低い保険契約者をポートフォリオの中から特定し、将来そうした保険契約者を判別するためのシンプルなルールを明確にすることができます。

例えば、「A、C または F 地域に居住し、職業コードが 1、2、3、4 または 5 の保険契約者の場合、平均の早期解約率は 83%と想定する」といったルールが考えられます。今後、クライアントとともに、こうしたインサイトを活用し、早期解約率を下げていく予定です。ビジネスへの活用法としては、例えば、解約の可能性が低い保険契約者に焦点を合わせて、よりカスタマイズされたマーケティングキャンペーンを行うことなどが考えられます。こうすれば、利益をもたらす前に解約してしまうような保険契約者に費やす新契約費が減ります。

「早期」解約の定義は、クライアントによって異なります。多くの場合、経済的ダメージが最も大きい時期によって決まり、例えば最初の 6 カ月以内に解約、最初の 1 年以内に解約などを「早期」と定義します。

機械学習を活用する理由

機械学習は、典型的なデータセットに適用の可能性があるルールが多数に及ぶ場合に、とりわけ役立ちます。

この点を分かりやすく説明するために、たった一つのルールに対して取りうる組み合わせの数について、少し考えてみましょう。説明変数が、例えば、保険契約者の職業コードと居住地域というように二つあり、それぞれの変数が 10 の異なる値（職業コードは A～J、地域は 1～10）を取りうる場合、1,046,528 通りもの非自明なルールの定義の仕方があります。

多くのデータセットで、含まれる説明変数が 100 に達することも珍しくないことを踏まえると、考えうる全てのルールを手作業で評価するのは不可能です。一方、「決定木（Classification and Regression Trees；CART）」アルゴリズムは、高次元のデータを効率的に処理することができます。加えて、CART は、そもそもの仕組みから、ポートフォリオの中から解約の可能性が低い保険契約者と解約の可能性が高い保険契約者を特定することに適しているのです。その方法は、無数のルールを系統的に構築・評価し、そこから、少数の本質的なルールに絞り込んでいくというものです。

これらのルールは、個々のポートフォリオで分析に利用できる説明変数に基づいており、したがって、利用可能な説明変数によって制約されます。そのため、結果として得られたルールや対応するセグメントを慎重に評価することによって、ビジネスにおいて実行に移すことができるインサイトを生み出します。

CART のしくみ

CART アルゴリズムは、解約挙動という点で、ポートフォリオ・セグメントの「純度」を最大化することを目標としています。例えば、セグメント***A*では、**全ての保険契約が解約された、あるいは、全ての保険契約が解約されていない場合、その純度は 100%とされ、不純度の値は$I(\boldsymbol{A}) = 0$となります。逆に、保険契約のちょうど半分が解約されたセグメントの不純度は、100%となります。不純度の指標として用いられるのはジニ不純度で、次のように定義されます。

$$I(\boldsymbol{A}) = 2 \cdot p \cdot (1 - p),$$

このとき、pは、セグメント$\boldsymbol{A}$の保険契約の解約率を表わ

hannover re®

しています。CART アルゴリズムの目標は、全てのセグメントの不純度の加重和が最小となるように、ポートフォリオをセグメント分けすることです。

CART は、以下のように、分割をグリーディ（貪欲）法により、段階的に進めます。最初にポートフォリオを二つのセグメントに分割します。この時のセグメント分けのルールは、「保険契約者が地域 1 または 3 に居住している」「保険契約者が地域 1 または 3 に居住していない」などのシンプルなルールです。

これを図にすると（上下逆向きの）木の形になります。このとき、最上部の「ルート（根）」$\boldsymbol{A}_0$はポートフォリオ全体を、「ブランチ（枝）」の$\boldsymbol{A}_L$（左側）と$\boldsymbol{A}_R$（右側）は最初の分割によって得られた二つのセグメントを表わしています。セグメント分割の条件は、考えうる全ての選択肢（地域 1 に居住している／していない、地域 2 に居住している／していない、地域 1 または地域 2 に居住している／していない、職業クラスが A である／ない、など）の中から、不純度が最も減少するものを選びます。不純度は、次式によって算出します。

$$\Delta I = n_0 \cdot I(\boldsymbol{A}_0) - [\, n_L \cdot I(\boldsymbol{A}_L) + n_R \cdot I(\boldsymbol{A}_R) \,]$$

このとき、$I(\boldsymbol{A}_0)$、$I(\boldsymbol{A}_L)$、$I(\boldsymbol{A}_R)$は、$\boldsymbol{A}_0$、$\boldsymbol{A}_L$、$\boldsymbol{A}_R$のそれぞれの不純度を表わしています。また、不純度は、サンプル数n_0、n_L、n_Rによって、それぞれ重みづけを行っています。

最初の分割が確定したら、CART アルゴリズムは、ポートフォリオの分割を繰り返し、さらに小さなセグメントに分けていきます。それとともに、木は、どんどん枝が増えて大きくなっていきます。分割するたびに、アルゴリズムが、既存のルールに追加することができる、合理的に考えうる全ての条件（例えば「保険契約者が地域 1 または 3 に居住しており、**かつ**、保険金額の合計が 3,000 ユーロより大きい」など）を試し、再度、不純度が最も減少する条件を選びます。

分かりやすくするため、10 件の保険契約からなるシンプルなデータセット例（図 1 参照）で考えてみましょう。それぞれの保険契約は、二つの説明変数、すなわち、地域（丸またはひし形）と職業クラス（濃い青色または薄い青色）によって特徴づけられます。また、解約された保険契約には、赤色のバツ印が付けられています。このとき、ポートフォリオ全体（木の最上部のルート）のジニ不純度は、次式で計算できます。

$$I(\boldsymbol{A}_{\text{portfolio}}) = 2 \cdot p \cdot (1-p) = 2 \cdot 0.5 \cdot 0.5 = 0.5.$$

CART アルゴリズムは、最初の分割時に、二つの選択肢を検討します。地域 1 と地域 2 で分割するか、職業クラス A と職業クラス B で分割するかの二つです。前者の選択肢を選んだ場合、不純度の減少は次式の通りです。

$$\begin{aligned}\Delta I^{\text{region}} &= n_{\text{portfolio}} \cdot I(\boldsymbol{A}_{\text{portfolio}}) \\ &\quad - [\, n_1 \cdot I(\boldsymbol{A}_1) + n_2 \cdot I(\boldsymbol{A}_2) \,] \\ &= 10 \cdot 0.5 - [6 \cdot 2 \cdot 1/3 \cdot 2/3 + 4 \cdot 2 \cdot 3/4 \cdot 1/4] \\ &= 0.83\end{aligned}$$

それに対し、後者の選択肢を選んだ場合、純度の減少は次式の通りわずかしかありません。

$$\begin{aligned}\Delta I^{\text{occ}} = n_{\text{portfolio}} \cdot I(\boldsymbol{A}_{\text{portfolio}}) &- [\, n_{\text{A}} \cdot I(\boldsymbol{A}_{\text{A}}) + n_{\text{B}} \cdot I(\boldsymbol{A}_{\text{B}}) \,] \\ &= 10 \cdot 0.5 \\ &\quad - [5 \cdot 2 \cdot 3/5 \cdot 2/5 + 5 \cdot 2 \cdot 2/5 \cdot 3/5] \\ &= 0.20.\end{aligned}$$

したがって、最初は、地域によって分割することになります。続いて、アルゴリズムは、地域 1 の保険契約について 2 回目の分割を行う意味があるかどうかを、不純度の減少という点から評価します。不純度の減少がゼロであることが明らかになったため、木の左側については、さらなる分割は行われません。一方、地域 2 の保険契約については、職業クラス A と B に分割すると、不純度の減少は 0.52 となり、ゼロではありません。

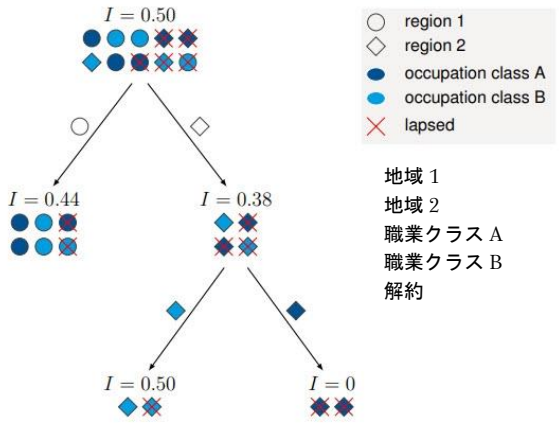

図 1：分類木の例

完全自動化ワークフローの神話

この機会に、機械学習に関するよくある誤解を取り上げたいと思います。分析では、ほぼあるいは完全に自動的に結果が生成されるため、分析に当たり、手作業はほとんどあるいは全く必要がないという誤解です。しかし残念ながら、これは、多くの場合、真実からかけ離れています。

当社における典型的なプロジェクト・サイクルは、準備と、クライアントから提供を受けた生データのクリーニングから始まります。これには、品質チェックや欠損値の検出なども含まれます。

このプロセスを支援するパッケージを開発しましたが、それでもなお、結果の品質を保証するために、データ作成の多くのステップは、手作業で行う必要があります。データチェックだけでなく、分析に用いる新たな変数の構築も行います。これらの新たな変数は、データセットから入手できる情報に基づいているものもありますが、外部のデータソース（郵便番号に基づく社会経済的ステータスなど）を利用して強化しているものもあります。

通常、このプロセスには、プロジェクト・サイクルの大きな部分を費やす必要があり、軽視すべきものではありません。しかし、このようなクリーニングと強化が行われた後のデータセットを受け取るだけでも、多くのクライアントにとって、大きなサポートとなります。

データを正しいフォーマットに収めたら、モデルを当てはめます。このプロセスは概ね自動化されているため、それほど多くの時間を要することはありません。ぴったり当てはまるモデルが得られたら、その後は、二つのステップが残されているだけです。

- 第一は、特定されたセグメントの品質をチェックし、将来その結果を確実に活用できるようにすることです。
- 第二は、分析結果をビジネスの視点で解釈し、実行に移すことができる具体的な対策を立てることです。

これら二つのステップにより、当社の hr | bluebox ソリューションは、市場で提供されている他のサービスと一線を画しています。

hr | bluebox がもたらす付加価値：信頼性の高い解約予測

当社は、結果の品質を様々な角度から保証しています。当社の主たる目標は、特定したセグメントやルールの信頼性を可能な限り高めることです。

信頼性に対する脅威の一つに、偶然の発見があります。解約と説明変数の間に系統的な関係がなくても、CART アルゴリズムは、ただ単なる偶然によって解約率が平均よりわずかに高いまたは低いポートフォリオ・セグメントを検出してしまう可能性があるのです。これは、過学習（overfitting）の典型的な事例です。そのため、検出したセグメントが、単なるノイズではないことを保証する必要があります。

そのために用いるツールの一つに、いわゆるファンネルプロット（図 2 参照）があります。ファンネルプロットに示した番号付きの点は、分析によって特定されたポートフォリオ・セグメントを表わしており、x 軸はポートフォリオの中でそのセグメントが占める割合、y 軸は解約率を示しています。

これらの点を、ファンネル（灰色の部分）と比較します。ファンネルは、解約が完全にランダムである場合、すなわち、いずれの説明変数とも無関係である場合に想定されるポートフォリオに占めるセグメントの割合と解約率の関係の範囲を示しています。

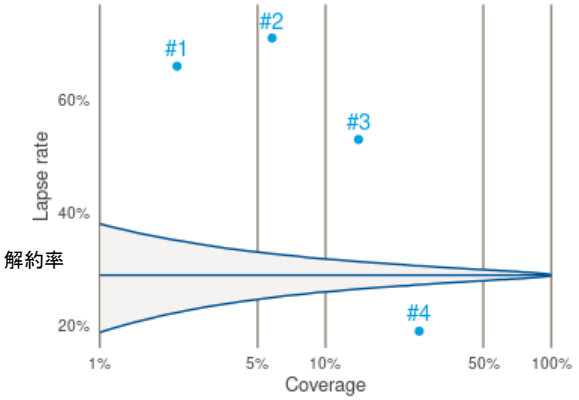

図 2：ファンネルプロット

このプロットは、人工のデータセットに基づくもので、前記の例のデータに基づくものではないことにご留意ください。ポートフォリオに占める割合が小さければ小さいほど、解約率の範囲が広がります。全くの偶然によって左右される可能性が高くなるからです。

より重要なのは、ファンネルの範囲を示す境界線と選定された点とが遠くかけ離れていることです。点#1（ポート

フォリオ・セグメント#1）について分析すると、ポートフォリオに占める割合が 2.2%の場合、ファンネルの範囲は 27.5%～40.6%であることが分かります。ファンネルの境界線と選定されたセグメントの解約率（セグメント#1 の場合は 76%）との距離が、信頼性を示す重要な指標となります。この場合、点#1 とファンネルの境界線との隔たりが大きいことから、このセグメントは偶然の発見ではないことの確からしさが比較的高いと言えます。

セグメントの検証が終わったら、ある変数によるものとして観測された影響が、実のところは別の変数の影響によるものなのかどうかをチェックします。そのために、データセットから最も重要な変数を除外するという作業を繰り返して、複数のモデルを比較します。その結果、特定されていた変数とは別の変数によって同程度の大きさの影響が生じていることが分かった場合には、この点についてクライアントと話し合い、明らかにします。

場合によっては、ルールの定義に用いた変数を、ビジネスの視点から見てより実行に移しやすい対策を立てるために、別の変数に置き換えることも考えられます。しかし、このような変更が何をもたらすかを目に見えるようにするためには、クライアントが決定したルールの変更の影響をインタラクティブに視覚化する必要があります。

ルールのインタラクティブな調整：データの視覚化

ルール変更の説明のために、当社では、分析によって得られたセグメントの意味をインタラクティブに視覚化できるダッシュボードを開発しました。

加えて、それぞれのセグメントに関して、条件を変更したり、特定の変数を除外／変更したり、新たなルールを追加したりすることもできます。そうした調整を行うと、ダッシュボードは、調整後の解約率や調整後のセグメントがポートフォリオに占める割合を表示します。このツールは、実施した分析に対する相互理解を深めるとともに、特定されたルールに基づいて、実行しやすく、最適な対策を立案する上でも役立ちます。

まとめ

機械学習を利用すれば、分析が円滑に行えるだけでなく、早期解約の複雑な影響要因を検出するためにも役立ちます。とはいえ、そのワークフローは、決して完全に自動化できるものではありません。また、分析結果を現実の実務に適用するには、モデルに当てはめるだけでなく、様々な調整が必要になります。しかし、当社のツールや戦略を利用すれば、従来の方法では見つけられなかったインサイトを生み出すことができます。

作成者

Lukas Herrmann
L&H - Data Analytics
Tel. +49 511 5604-2630
lukas.herrmann@hannover-re.com

LinkedIn でフォローすると、Life & Health の最新情報が得られます。