2022 年 10 月 20 日　全 12 頁

退職確率を分析する

～数理モデルを用いた退職確率分析～

コンサルティング第三部 コンサルタント　江藤 俊太郎

［要約］

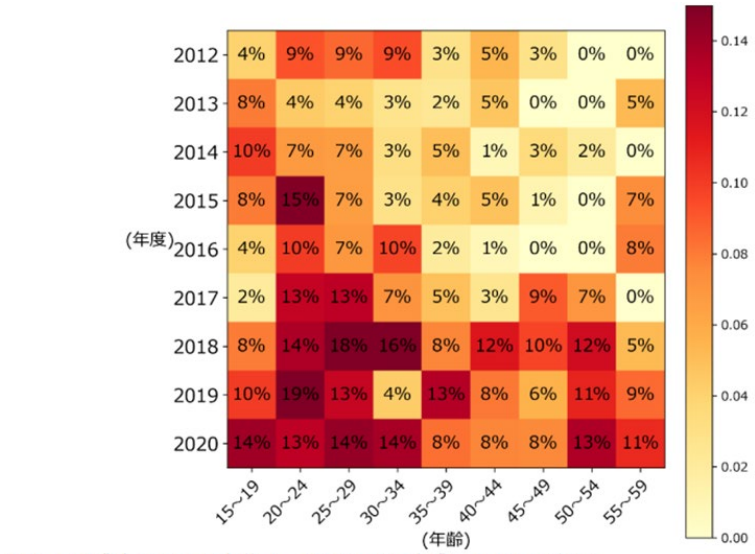
- 人材の維持・保持（リテンション）のためにも、また人的資本経営のためにも、退職の要因を把握することは重要な課題だ。
- そこで本稿では数理モデルを用い、その結果を踏まえて一定期間にどの程度退職者がでるのかという退職確率の分析、さらに予測を行ってみた。
- 採用したのは、エンゲージメント・データと組み合わせて退職確率に影響を与えている要素を明らかにするモデルである。
- 分析のための分析ではなく、データにより、何が大きく退職を決めるファクターなのかを可視化し、より有効な人事施策を取ること、それが最終目標である。

1. 退職確率分析の重要性とその手法

少子高齢化が進む中で優秀な人材の維持・保持（リテンション）は重要な課題である。たとえば次頁図表 1 の退職率[1]ヒートマップを持つこの会社（サンプル）では、25-34 歳の退職傾向が近年顕著に悪化している。しかし、ヒートマップは退職の動向を見極めるのにきわめて有用だが、この情報だけでは、何が退職の原因かは判別しがたい。

[1] 会計計算等、様々な局面で用いられる一般的な退職率は、過去の実績データを用いて算出した年齢(勤続年数)別の退職率のことを指す。ヒートマップはこれを基に作られている。一方、本稿で特にことわりなく退職確率といった場合、一定期間のうちにその従業員が退職する確率のことを指す。

（図表 1）退職率ヒートマップ



（注）色の濃度は 0%を最も薄く、15%以上を最も濃い色として表示

出所：ダミーデータより、大和総研作成

事前に退職の兆候を捉えたいというニーズが経営陣にはあるはずだ。もちろん、企業の人事部門は膨大なデータを持っている。しかし、だからこそ、そこから法則性を導き出していくのは困難にみえる。

そこで本稿では、数理モデルを用い、その結果を踏まえて退職確率の分析、さらに予測モデルの構築を行ってみた。

ところで、退職確率分析では、複数のパラメータから売上高（金額）を予想する重回帰分析などとは異なり、従業員が一定期間後に在職しているか、退職しているか、どちらかしかない状態を対象とする（2 値問題）。このように 2 つの状態を対象とする場合、データ分析で一般によく使われるものに以下のものがある。

① ロジスティック回帰分析[2]

② AI における機械学習（たとえば、タイタニックの乗客名簿の分析が有名である）

しかし、これらの手法を、表面的に用いるだけでは、時間の要素を容易に取り込めないという弱点がある。また、退職の場合、「いつ」という観点は貴重なデータであり、「いつ」「どれくらいの割合で」退職するかどうか知りたいというニーズがあろう。

そこで、豊富な研究蓄積があり、説明もわかりやすい、イベント・ヒストリー・アナリシスの手法を退職確率分析に応用してみる。イベント・ヒストリー・アナリシスとは、医学（疫学・薬学分野）では生存時間分析、工学では故障時間分析、マーケティングでは顧客離反分析（チャー

[2] ロジスティック回帰は機械学習のアルゴリズムとしても使われているが、統計的手法として古典的部類に入るためあえて別記している。

ン分析）として利用されているものである。やや専門的な用語で表現すると、①カプラン・マイヤー推定量（単変量解析）、②Cox 比例ハザードモデル（多変量解析）の退職確率分析への応用可能性を探っていく。

2. 退職確率分析の実際の手順(1)データの準備

採用したのは、エンゲージメントのデータが退職確率に影響を与えているとの前提で構築したモデルである。入力データのイメージは図表 2 のようになる（エンゲージメント調査の結果等を加工して作成したダミーデータ）。在職者 401 人の会社であり、観測期間内（約 3 年間）の退職者は 99 人。ID1 の従業員は、観測開始 959 日目にして、現在も在職中。ID2 の従業員は、343 日目に退職した。

それぞれに紐づいている「業務のやりがい」「報酬に対する満足度」「人間関係の満足度」「将来の企業・業界展望」は、人事指標やエンゲージメント・データとして比較的一般的な項目かと思われる（厚生労働省の「雇用動向調査結果」の離職理由を参考にしている）。またここでは、他に「月平均の時間外労働時間」を労働条件として設けている。データの取得により観測期間が開始されるが、最も長いケースで 1091 日（約 3 年）前である。

サンプルでは、主に 2 段階評価にしているが、5 段階評価でも構わない。連続した数値も分析可能である。また、図表 2 のように、エンゲージメントのデータが、退職者も在職者同様残っていることが前提である。

（図表 2）サンプル・データ

ID	状態 （0＝在職中、 1=既に退職）	観測期間 （日）	業務のやりがい （1＝感じる、 0=感じない）	報酬に対する 満足度 （1＝感じる、 0=感じない）	人間関係の 満足度 （1＝感じる、 0=感じない）	将来の企業 ・業界展望 （1＝明るい、 0=そうでない）	…	時間外労働 （月平均）
1	0	959	1	1	0	0	…	10.1
2	1	343	1	0	0	0	…	46.6
3	0	846	0	1	1	1	…	30.1
4	1	61	1	0	1	0	…	22.0
5	0	455	0	1	0	0	…	22.0
6	0	968	1	0	0	0	…	6.7
7	0	614	0	1	1	0	…	18.8
8	1	133	0	1	1	1	…	30.7
9	0	104	1	0	1	1	…	33.0
10	0	1001	1	1	0	1	…	7.6
⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
498	0	524	1	0	1	1	…	21.8
499	0	879	1	0	1	0	…	4.8
500	0	74	0	0	0	1	…	13.5

出所：ダミーデータより、大和総研作成

図表 2 には盛り込んでいないが、退職を説明する必要データ(説明変数と呼ぶ)には他にも次のようなものが挙げられるだろう。「性別」「年齢」「勤務地」「異動の頻度」「各種パフォーマンス指標[3]」「研修受講歴」「キャリアアップ志向」(資格のあるなし)「信頼できる相談者（メンター）の存在」等。

図表 3 は図表 2 のデータの記述統計量である。たとえば、状態 0 である 401 人の在職者中、「業務のやりがい」については平均値 0.656、「時間外労働」の平均値は 20.9 時間、時間外労働の最小値は 0.1 時間、最大値は 39.8 時間である。

同じく、状態 1 である 99 人の退職者中、「業務のやりがい」については平均値 0.424、「時間外労働」の平均値は 30.3 時間、時間外労働の最小値は 1.6 時間、最大値は 58.0 時間となる。

（図表 3）記述統計量

状態	変　数	n	平　均	不偏分散	標準偏差	最小値	最大値
全　体	観測期間	500	502.660	102636.682	320.370	1.000	1091.000
	業務のやりがい	500	0.610	0.238	0.488	0.000	1.000
	報酬に対する満足度	500	0.546	0.248	0.498	0.000	1.000
	人間関係の満足度	500	0.584	0.243	0.493	0.000	1.000
	将来の企業・業界展望	500	0.636	0.232	0.482	0.000	1.000
	時間外労働	500	22.783	153.700	12.398	0.100	58.000
状態=0	観測期間	401	555.853	98722.216	314.201	1.000	1091.000
在職者	業務のやりがい	401	0.656	0.226	0.476	0.000	1.000
	報酬に対する満足度	401	0.576	0.245	0.495	0.000	1.000
	人間関係の満足度	401	0.584	0.244	0.494	0.000	1.000
	将来の企業・業界展望	401	0.653	0.227	0.476	0.000	1.000
	時間外労働	401	20.917	129.903	11.397	0.100	39.800
状態=1	観測期間	99	287.202	61187.796	247.362	4.000	925.000
退職者	業務のやりがい	99	0.424	0.247	0.497	0.000	1.000
	報酬に対する満足度	99	0.424	0.247	0.497	0.000	1.000
	人間関係の満足度	99	0.586	0.245	0.495	0.000	1.000
	将来の企業・業界展望	99	0.566	0.248	0.498	0.000	1.000
	時間外労働	99	30.341	180.441	13.433	1.600	58.000

出所：図表 2 のデータより、大和総研作成

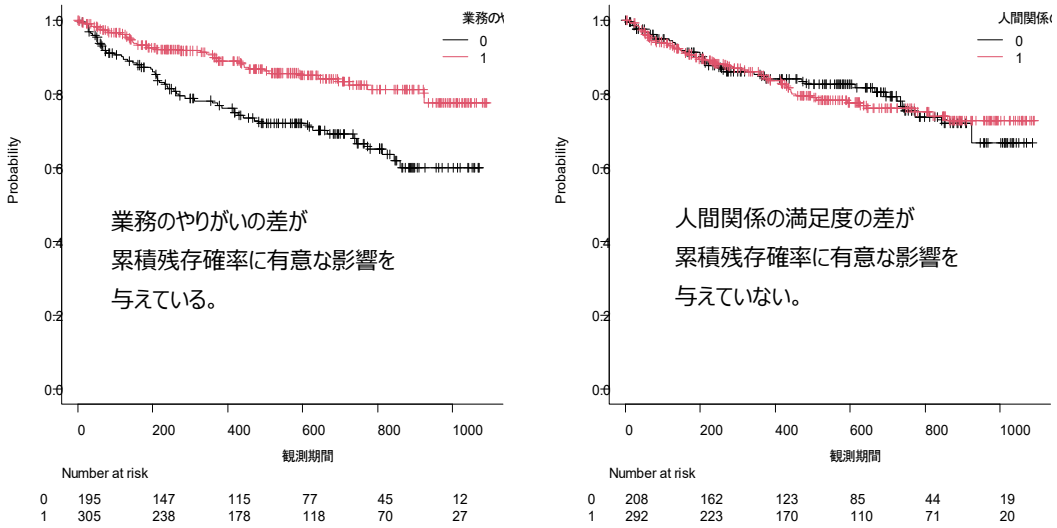
[3] 仕事の質・量、業務知識、創造性等のパフォーマンスを評価する様々な指標

3. 退職確率分析の実際の手順(2)カプラン・マイヤー推定量の検定

さて、このデータを基にまず、カプラン・マイヤー推定量、すなわちカプラン・マイヤー曲線のグラフをみてみる。カプラン・マイヤー曲線は、いわゆる生命表（脱退残存表）と呼ばれる表をグラフで描画したもの。この表では退職者とそれ以外の脱落者（この場合、観測打ち切り）を母数から順々に除いて、残存確率を都度計算していく。グラフは、時間とともに下りていく階段式に描かれる。

横軸を時間、縦軸を累積残存確率（＝1-累積退職確率）で表すことになっている。図表 4 では、「業務のやりがいは退職確率に影響を与えている」(左)「人間関係の満足度は退職確率に影響を与えていない」(右)ことが一目瞭然である。なお、ログランク検定[4]という手法を並行して用いると、差のあるなしを数値で確認できる。

（図表 4）カプラン・マイヤー曲線



出所：図表 2 のデータより、大和総研作成

まずは、このように変数の種類に応じてカプラン・マイヤー曲線をチェックすることがモデル構築の第一段階である。もちろん、カプラン・マイヤー曲線を使わなくとも、プログラム言語 R や Python を使えば、一足飛びにモデル構築に変数選択を行うことは可能だ。ただし、それでもこのグラフを描画する意味はある。

その最も大きなものは、入力値を連続変数として扱ってよいか、カテゴリー化したほうがよい

[4] 2 群の生存時間(この場合、在職時間)に差があるかどうかを検定する手法のこと

かの区別がわかることである。たとえば残業時間 10 時間以下と、30 時間以上がともに退職確率を上げる要因で、20 時間程度では逆に退職確率を下げる要因になる、といった現象もありうる。こうした場合、カプラン・マイヤー曲線での事前の検証が必要になる。

仮に、最終的に適切な退職モデル構築に行きつかなかった場合も、たとえばカプラン・マイヤー曲線は新入社員の 3 年退職確率（離職率）等の分析などに有用である。配属時のチューターの違いや、配属地、SPI[5]などの適性検査の点数などに退職確率が影響されているのではないか、との仮説を立てて検証すると、その結果を統計的に知ることができる。

4. 退職確率分析の実際の手順（3）モデルの構築

次に、いよいよモデルの構築に進む。採用する Cox 比例ハザードモデルは、カプラン・マイヤーが単変量解析であるのに対して、多変量の解析である点に特徴がある。同時に複数の説明変数（＝共変量）をモデルに組み込め、かつ、相対リスク度を数値で示せるメリットもある。イメージとしてはベースとなる退職確率曲線（ベースラインハザードと呼ぶ）があって、ハザード比と呼ばれる係数がパラレルに影響を与えるというモデルである。

このモデルで使う「ハザード関数」は以下の式で表される。専門的に言うと、β が共変量に対応する偏回帰係数、x が共変量を表す。

$$h(t|x) = h_0(t) * \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)$$

Cox 比例ハザードモデルでは、係数 β を指数変換したハザード比$= \exp(\beta_n)$に注目する。R や Python 言語を使えば、この β を簡単に導出することができる。変数選択の結果は、カプラン・マイヤー曲線が示唆した通り「業務のやりがい」「報酬に対する満足度」「時間外労働」が残った。結局、図表 2 のデータを基に計算されたハザード関数の式は以下のようになる。

$$h(t|x) = h_0(t) * \exp(-0.8038 * \text{業務のやりがい} - 0.5008 * \text{報酬に対する満足度} + 0.0609 * \text{時間外労働})$$

この式は以下のように解釈できる。業務のやりがいがある人はない人に比べて退職のリスクを 45%＝exp(-0.8038)に下げる。報酬に対する満足度がある人はない人に比べて退職のリスクを 61%＝exp(-0.5008)に下げる。月の時間外労働が 1 時間増えると 6%=exp(0.0609)退職のリスクを上げる。なお、対数線形性の仮定により、残業が 20 時間増えるとこのモデルの場合退職のリスクは 1.06^20＝3.2 倍増えることになる。

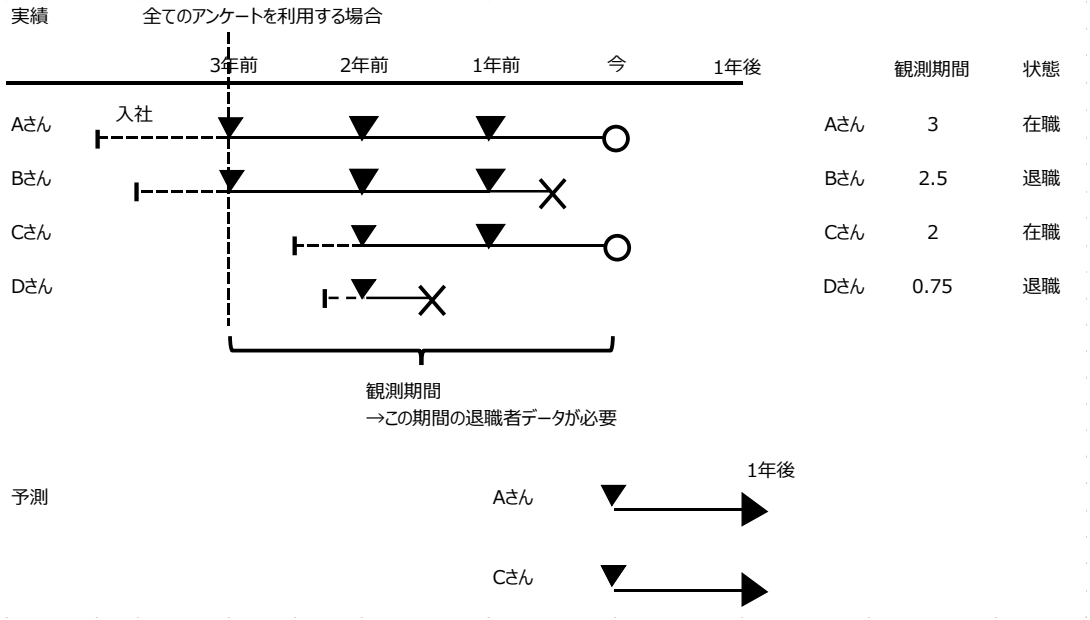
[5] SPI とは、株式会社リクルートマネジメントソリューションズが開発・提供している適性検査のこと

ただし、Cox 比例ハザードモデルを使う場合、「比例ハザード性[6]」が担保されているかどうかに注意しなければならない。たとえばR では、3 つの方法[7]でそれを確認することができる。今回のケースではどの説明変数も比例ハザード性を満たすことが確認された。

また、今回は、各共変量が時間において不変であることも前提としている。これは、観測期間（この場合最長 3 年前）、開始時のエンゲージメントのデータが経時的に変化しない、あるいはその時点でのエンゲージメント・データ取得結果が退職行動を決めるというかなり強い仮定である。

実際には、エンゲージメント・データの取得は繰り返し行うだろう。そして、各種エンゲージメント・データについても月平均の残業時間についても、経時的に変化する可能性が高い。変化した直近の値が最も強く退職行動に結びついていることは十分に考えられる。その場合、図表 5 のような考え方に従って、一人のレコードを複数に分割する処理が必要となる。たとえば B さんは半年前に辞めたほうである。3 年前、2 年前のデータは退職に帰結しなかったが、1 年前のデータは退職に帰結した。こういった区切ったデータを、時間依存共変量を考慮に入れた Cox 比例ハザードモデルとして計算する。

（図表 5）経時的な変化を盛り込む場合



出所：大和総研作成

[6] 共変量の 2 群間において、ハザード比が時間によらず一定であることを指す。
[7] ①log-log プロットを確認する、②cox.zph 関数を用いる、③Schoenfeld 残差を確認する。

長所は各時点での各エンゲージメントの効果を盛り込めること、サンプル数が増えるので説明力が増すことである。短所は前工程の時間がかかるのと、予測がレコードを区切った期間（この場合 3 年→1 年）に限定されることである。

図表 2 のデータに基づく本稿のモデルは、あくまで時間による変化を前提としていないシンプルなモデルであることに留意されたい。

5. モデルのアウトプット

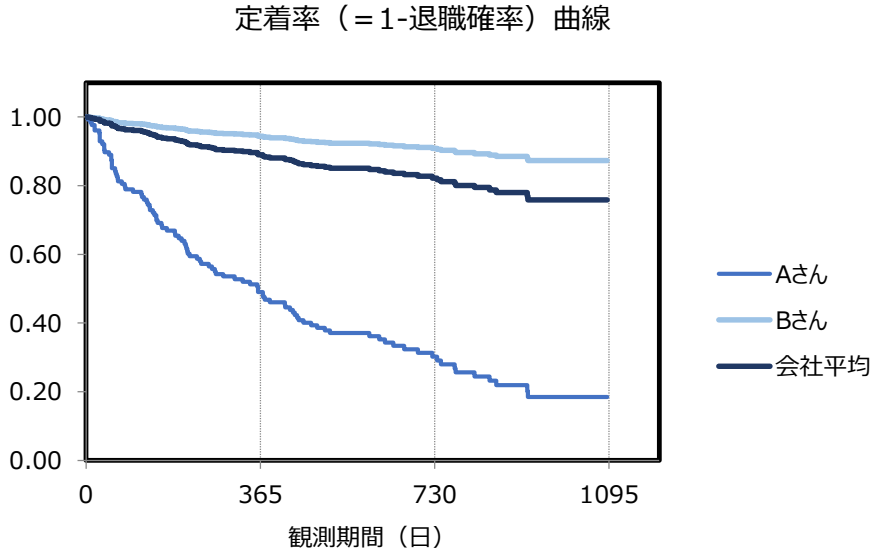
モデル式が導出されれば、以下のようなアウトプットを作成可能である。

① 個人別・部署別シミュレーション

構築された退職確率のモデルを用いることで、シミュレーションが可能になる。たとえば、先に上げたモデルが構築できた場合、A さん、B さんとモデル会社の平均を比較すると、以下のようなシミュレーション結果になる。部署ごとの比較も可能となる。この場合、A さんの 1 年以内に退職する確率は、51.0%、B さんの 1 年以内に退職する確率は 5.6%、会社全体では 11.0%となる。

（図表 6）シミュレーション

	Aさん	Bさん	会社平均
業務のやりがい	0	1	0.610
報酬に対する満足度	0	1	0.546
時間外労働	40	20	22.783

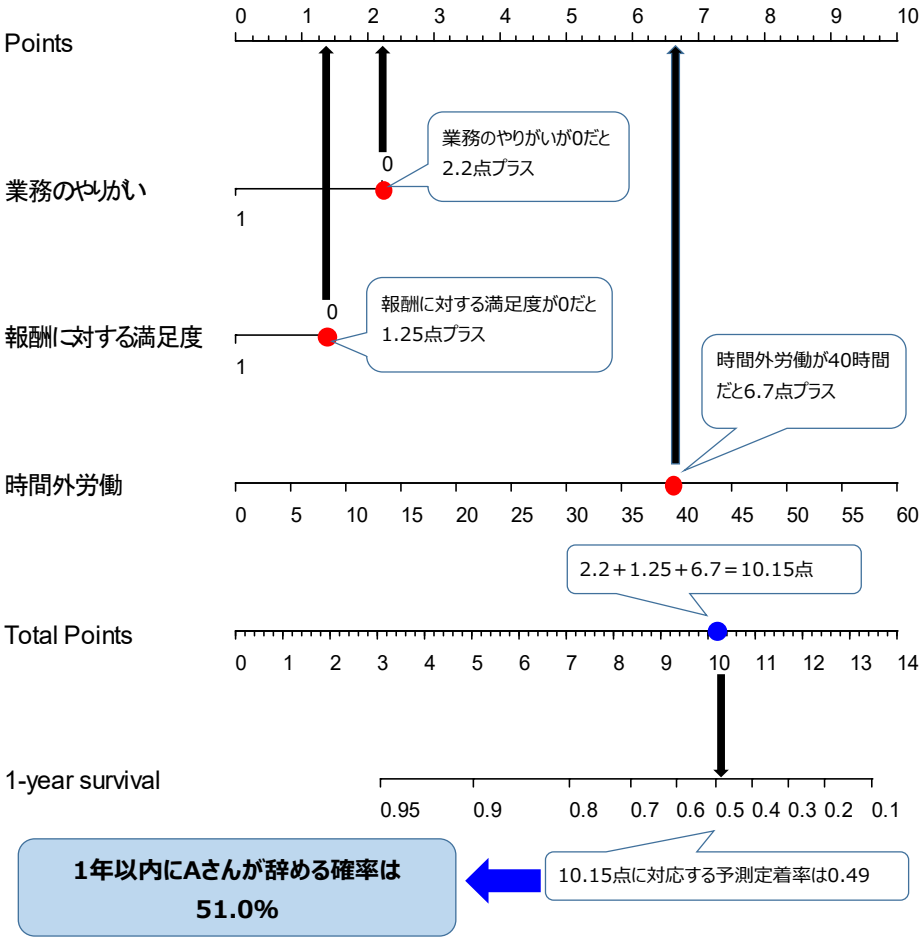


出所：図表 2 のデータより、大和総研作成

② ノモグラム(簡易計算表)による寄与度分析

ノモグラムは、そもそも手元でコンピュータが使えない時代によく使われていた簡易計算表のことであるが、視覚的にどの要素が一番退職に影響を与えているかわかりやすいので、描画してみた。図表 7 では上の 4 行でポイントを計算、下の 2 行で確率を計算する。A さんの例では、「時間外労働」が一番退職確率に大きな影響を与えていることがわかる。

（図表 7）ノモグラム

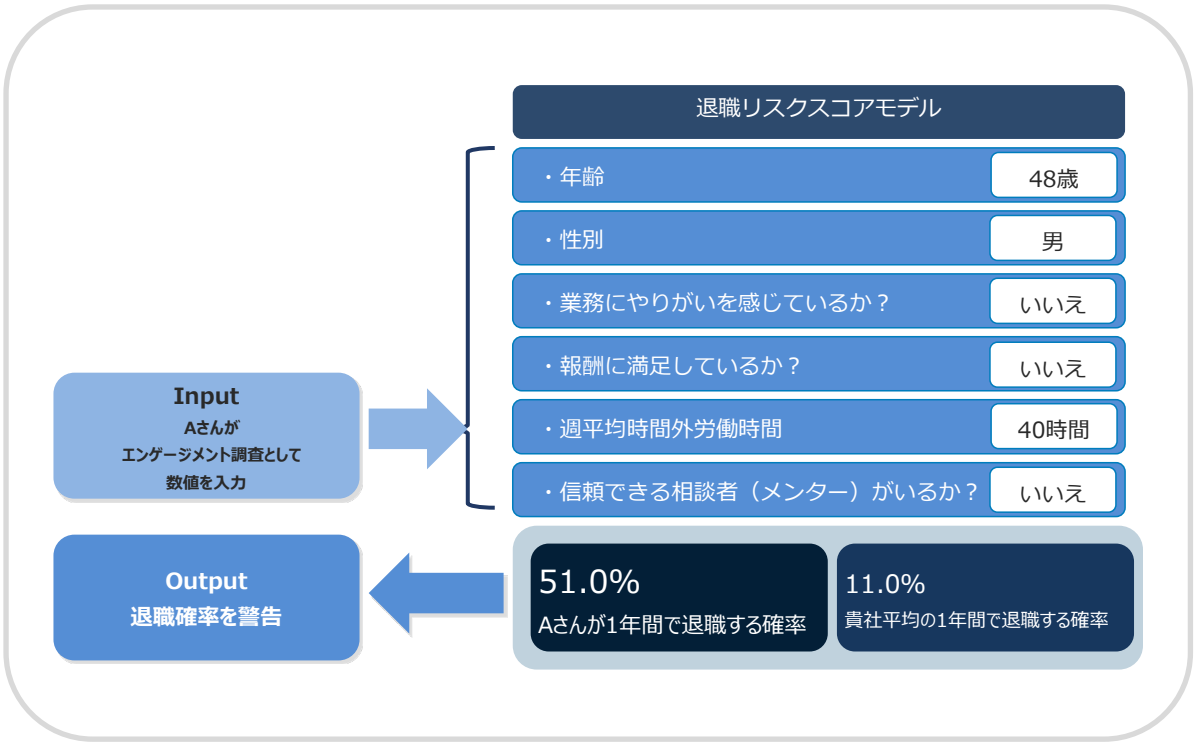


出所：図表 2 のデータより、大和総研作成

③ リスクスコアモデル

米国で開発された、冠動脈疾患の予測に定評のあるフラミンガム・リスクスコア・モデル[8]というものがある。このモデルには、Cox 比例ハザードモデルが使用されている。そこで、同様に退職予測版を構築できないかと考えてみた。なお、この退職リスクスコアモデルは図表 2 のデータより多くの説明変数を盛り込めたと仮定したモデルで構築されていることに注意されたい。（年齢、性別、メンターのありなしは図表 2 のデータには、反映されていない）。

（図表 8）退職リスクスコアモデル



出所：大和総研作成

①～③のようなアウトプットが構築できれば、様々な示唆が得られるはずである。

たとえば、

- 時間外労働時間が退職に与える影響が大きいことがわかった。
- 人間関係のスコアがとりわけある部署で退職確率に影響を与えていることがわかった。
- 社内で相談できる相手（メンター）の存在が重要だとわかった。
- 研修受講歴が、退職確率に影響を与えていることがわかった。

などの結論が導き出せる。当然ながら、単なる分析のための分析ではなく、データにより、何

[8] https://www.mdcalc.com/calc/38/framingham-risk-score-hard-coronary-heart-disease

が大きく退職を決めるファクターなのかを可視化し、企業がより有効な人事施策を取れるようになること、それが最終目標である。

6. 数理モデルの利点と欠点、展望

本稿で採用した Cox 比例ハザードモデルの利点としては、数理モデルなので、どの数値がどの程度退職確率に影響を与えているか、わかりやすく把握可能であることである。欠点としては、比例ハザード性が担保されていない（変数の与える影響が時間によって変動する）場合の適用に工夫がいること、経時的なデータ取得が必須なこと、が挙げられよう。

また、予測するからには、予測モデルとしての精度が重要なポイントである。これにはまず現実の退職者データと比較した各種精度を調べなければならない（これは混同行列[9]を評価したり、ROC 曲線[10]を描くなどしたりして、モデル構築時点で検証することが可能である）。同時に、退職確率の高かった人が実際に辞めたのかどうかを事後的にフォローしていく必要もある。

経時的なデータが入手しづらかったり、AI の導入でより精度の高いモデルが構築できたりするのであれば、実務では AI モデルのほうを採用することも考慮に入れるべきである（大和総研では AI を用いた退職確率分析の研究も進めている）。また、本稿で紹介したイベント・ヒストリー・アナリシスの考え方を AI に応用する試みも現在様々に研究されている。本稿では割愛したが、ランダム・サバイバル・フォレスト[11]等の手法を用いると、単純な Cox 比例ハザードモデルに比べてより精度の高い結果が確認できる。

－以上－

参考文献

- 金 明哲「Ｒによるデータサイエンス」森北出版（2007 年）
- 外山 信夫・辻谷 将明「実践Ｒ統計分析」オーム社（2015 年）
- 中村 剛「医学統計学シリーズ 3　新版　Cox 比例ハザードモデル」朝倉書店（2018 年）
  他

[9] 2 値分類問題で予測された結果をまとめたマトリックスのこと。左上より時計回りに、真陽性（在職と予測され、実際に在職していること）、偽陰性（退職と予測され、在職していること）、真陰性（退職と予測され、退職していること）、偽陽性（在職と予測され、退職していること）の 4 つのクラスで表される。

[10] 「受信者動作特性曲線」の略。陽性、陰性の閾値（しきいち）を動かした場合に、真陽性と偽陽性の割合をプロットして作られるカーブ。曲線が左上に近いほどそのモデルは精度が高いと見なされる。

[11] 機械学習のスタンダードな手法であるランダム・フォレストを生存時間分析に応用したモデル。精度の高さの他にも比例ハザード性の仮定が必要ないなどのメリットがある。