

【AIスコアの概要】

AIスコアとは

決算や業績予想修正の発表が株価に与える影響を評価し、株価の変化率を測る参考指標です。

例えば、ある決算発表に対してスコアが+1.00であれば、過去データに基づいて「この決算内容であれば、平均的に株価が1%上昇する」と評価されたことになります。

AIスコアが対象とするのは決算発表当日（発表が大引け後の場合は翌日）の株価変化率、およびその1日後、5日後の変化率についてもスコアを算出します。

【入力データ例】

以下は、AIスコアの算出にあたって入力するデータの一例です。これは通期の決算発表で、発表された実績値の一部を以下に記載します（実際のデータは非常に多くの項目を含むため、一部を抽出・整形して表示しています）：

	売上高	営業利益	経常利益	当期利益
発表実績	36,997	4,968	4,787	2,687

上記の数値を評価するため、入力データには「前期の実績値」「直近の予想値」「直近のQUICKコンセンサスの値」なども含まれます：

	売上高	営業利益	経常利益	当期利益
前期実績	46,346	5,940	5,452	3,408
直近予想	50,612	5,979	5,579	3,515
直近コンセンサス	38,100	4,600	4,350	2,750

これらのデータをもとに、プログラム内部では増減率を算出します：

	売上高	営業利益	経常利益	当期利益
対前期実績	-20.2%	-16.4%	-12.2%	-21.2%
対直近予想	-19.1%	-5.4%	-2.2%	-15.5%
対直近コンセンサス	-2.9%	8.0%	10.0%	-2.3%

この例では、いずれの数値も前期実績や直近の予想に対しては低い結果となっています。ただし、直近のコンセンサスと比較した場合、営業利益と経常利益についてはプラス評価となっています。

AIスコア算出プログラムはこの値をもとに、「過去のデータに基づくと、このようなパターンでは株価がどう動いたか」という観点からスコアを算出します。このとき、将来の株価には決算数値のほかに直近の株価の動きも影響しますので、入力データには移動平均乖離率も含めております：

	5日	25日	75日
移動平均乖離率	2.12%	8.79%	13.54%

これらの値をもとに、以下のような予測値が出力されます：

	当日	1営業日後	5営業日後
株価変化率のスコア	3.2%	0.6%	-0.5%

これは決算発表の当日（大引け後の発表の場合は翌日）の株価は3.2%上昇する一方、1営業日後、5営業日後の変化率は小さく影響はほとんど残っていない、という評価を示しています。

上記は「平均的な」スコアであり、これに加えて「区間値」も出力します：

	当日	1営業日後	5営業日後
上限（95%点）	12.7%	6.5%	4.3%
下限（5%点）	-6.4%	-5.3%	-5.2%

これは90%区間と呼ばれるもので、「この区間に値の90%が入る」ものとして利用することができます。当日の例でいえば、6%近く下落するおそれも十分あるものの、より上昇幅が大きく12%以上上昇する可能性もあることが読み取れます。

かなり大きな幅をもっていますが、1営業日後、5営業日後ではこの状態は落ち着き、ほぼ上下に対称な区間が出力されています。

【スコア算出メカニズム】

算出には機械学習および統計学の手法を用います。まず、蓄積された過去の決算発表と株価のデータから、決算発表と将来の株価の関係を「学習」します。

「学習」とは機械学習の用語で、ある入力に対してどのような出力が得られるか、という関係を再現することを表します。この関係は「モデル」と呼ばれる仕組みで表現されます。AIスコアにおいては、入力として決算発表データを「モデル」に与えることで、将来の株価変化率を出力するように処理を行います。

学習したモデルを用いることで、発表された決算に対して「従来の類似決算のパターンに基づけば、将来の株価がどのように動くか」という評価を行うことができます。

【データの例】

モデルの学習にあたっては、蓄積された過去データを使用します。発表された実績や予想をQUICKコンセンサスや直近予想、直近実績などと比較して変化率を算出します。

変化率は以下で定義されます：

$$\text{変化率} = \frac{\text{発表値} - \text{比較対象}}{|\text{比較対象}|}$$

以下は上記の変化率について、比較対象が0の場合、または存在しない場合なども含めてカテゴリ分けした例です。当期利益の発表実績を対直近予想で評価したもので、平均的には利益が高いほど株価が上昇する傾向にあります。

カテゴリ	当期利益 実績評価 対直近予想	株価変化率平均	件数
A	変化率 $\geq 10\%$	0.16%	5,090
B	$3\% \leq \text{変化率} < 10\%$	-0.30%	3,602
C	$-3\% < \text{変化率} < 3\%$	-0.38%	8,969
D	$-10\% < \text{変化率} \leq -3\%$	-0.70%	1,809
E	変化率 $\leq -10\%$	-1.15%	3,112
-	直近予想なし	-1.32%	328

同様のことを他の発表値についても行うことができます。例えば経常利益の場合、以下のようになります：

カテゴリ	経常利益 実績評価 対直近予想	株価変化率平均	件数
A	変化率 $\geq 10\%$	0.05%	4,863
B	$3\% \leq \text{変化率} < 10\%$	-0.35%	4,045
C	$-3\% < \text{変化率} < 3\%$	-0.38%	9,663
D	$-10\% < \text{変化率} \leq -3\%$	-0.72%	1,853
E	変化率 $\leq -10\%$	-1.06%	2,138
-	直近予想なし	-1.34%	348

同じ「変化率 $\geq 10\%$ 」であれば、経常利益より当期利益のほうが株価の上昇率が高いといえそうです（件数も影響しますが、ここではそれほど変わりません）。

売上高の場合は以下のようになります：

カテゴリ	売上高 実績評価 対直近予想	株価変化率平均	件数
A	変化率 $\geq 3\%$	-0.53%	2,004
B	$1\% \leq \text{変化率} < 3\%$	-0.35%	3,167
C	$-1\% < \text{変化率} < 1\%$	-0.31%	12,493

D	-3% < 変化率 ≤ -1%	-0.35%	2,457
E	変化率 ≤ -3%	-0.67%	2,308
-	直近予想なし	-1.06%	481

売上高は当期利益に比べると直近予想からの乖離が小さいため、カテゴリ分けの閾値も異なっています。かつ、予想との乖離が±1%となるCのときがもっとも株価変化率は高く、利益のように「高いほどよい」という傾向は見られません。

一方で売上高をQUICKコンセンサスと比較した場合は様子が変わってきます：

カテゴリ	売上高 実績評価 対直近コンセンサス	株価変化率平均	件数
A	変化率 ≥ 3%	1.33%	338
B	1% ≤ 変化率 < 3%	0.80%	988
C	-1% < 変化率 < 1%	0.25%	2,739
D	-3% < 変化率 ≤ -1%	-0.65%	857
E	変化率 ≤ -3%	-1.15%	339
-	直近コンセンサスなし	-0.56%	17,649

QUICKコンセンサスは対象銘柄が限られ、またアナリスト予想が2社以上の値が存在するときしか利用しません。これは1社予想ではコンセンサスとしての精度が大きく落ちると考えられるためです。件数は少ないものの、コンセンサスからプラス方向に乖離が大きいほど株価が上昇する傾向がみられます。

ここまで見たなかでは、売上高を直近コンセンサスと比較したときが最も株価との連動性が高いといえそうです。一方でコンセンサスは利用できない銘柄もあることから、全体でみたときに適用しづらい側面もあります。

【モデルについて】

上記のとおり、各発表数値をさまざまな指標と比較することで、決算発表後の株価変化率と関連のあるデータ項目を作成することができます。

このとき、各項目には

- 株価変化率と関連の強いもの、弱いものがある
- 関連は強いものの、一部の銘柄にしか適用できないものもある
- 互いに似た値をとるものもあれば、異なる側面を評価しているものもある

といった特徴があります。したがって、複数の項目を組み合わせることに意味があります。

単一の項目であれば前出のように一覧表の形で評価することができます。ただ、複数の項目の場合、組み合わせ数が増えすぎてしまい、組み合わせあたりのデータ件数が少なくなってしまう。数十件あればともかく、2件、3件といったデータを元に推論を行うことは誤差が大きくなり危険です。

その場合、モデルを用いることで、このような問題を解決することができます。いくつかの手法がありますが、基本的に以下のような特徴を持ちます：

- 必要な項目のみを使用する
- 項目によって予測への寄与度(重要度)が異なる
- 複数の項目を組み合わせても誤差を抑える仕組みがある

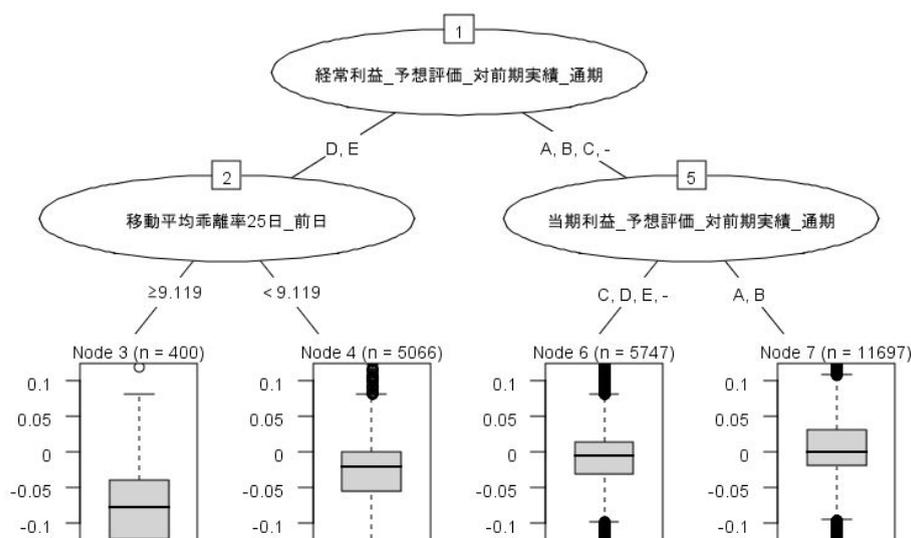
以下、代表的なモデルについて紹介します。

記法として、予測に用いる各データ項目をX、予測対象となる値(株価変化率など)をYとします。

①決定木

決定木は入力Xの値による「場合分け」を順に繰り返す手法です。このとき、もっとも分割による効果が高い(Yの値が大きいものと小さいものに分かれやすい)項目から順に場合分けを行います。代表的な手法であるCARTアルゴリズム(およびその類似手法)では、各項目でもっとも効率のよい二分割を行います。

以下はシンプルな決定木の例です(実際はもっと多くの層が生成されます):



ここでは合計4つのパターンにデータを分割しています。もっとも株価が上昇する組み合わせは「経常利益の予想が前期実績に比べてあまり低くなく」、かつ「当期利益の予想が前期実績よりも高い(AまたはB)」というパターンです。

逆に株価が下落する組み合わせは「経常利益の予想が前期実績より低く(DまたはE)」、かつ「前日株価の移動平均乖離率が約9.1%以上」というものです。事前期待が非常に高い反面、今後の予想が振るわなかったケースといえます。

決定木はこの通りシンプルな仕組みであることから、わかりやすく比較的ノイズに強い(極端な値の影響を受けにくい)という利点があります。一方でデータを分割する特性から、予測精度を担保するためには比較的少数の項目しか利用できないといったケースもあります。

②線形回帰

線形回帰は数式として表されるモデルで、以下のように「項目ごとに係数をかけ、それを足し合わせる」という形が基本となります:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

ここでカテゴリの場合、データ項目ひとつに対して複数の変数が対応することに注意が必要です。例えば「売上高 実績評価 対直近予想」というカテゴリ項目ひとつに対して、以下のような対応となります：

カテゴリ	X1	X2	X3	X4	X5
A	1	0	0	0	0
B	0	1	0	0	0
C	0	0	1	0	0
D	0	0	0	1	0
E	0	0	0	0	1
-	0	0	0	0	0

カテゴリは計6種類あり、「比較対象なし、または発表なし」を表す「-」を基準として計5つの変数が生成されます。したがって、モデルに含まれるカテゴリ数の5倍の変数が存在します。

この項目だけを使ってモデルを作ると、5つの変数を含む以下のような数式となります：

$$Y = \beta_0 + \beta_A X_A + \beta_B X_B + \beta_C X_C + \beta_D X_D + \beta_E X_E$$

得られた係数の値は以下の通りです：

	係数	パーセント表記
β_0	-0.01061	-1.06%
β_A	0.005261	0.53%
β_B	0.007125	0.71%
β_C	0.007524	0.75%
β_D	0.007061	0.71%
β_E	0.003942	0.39%

カテゴリが「-」のときは

$$Y = \beta_0 + \beta_A 0 + \beta_B 0 + \beta_C 0 + \beta_D 0 + \beta_E 0$$

となるので、予測値は定数項(β_0)のみで -1.06% となります。

カテゴリが「B」であれば、

$$Y = \beta_0 + \beta_A 0 + \beta_B 1 + \beta_C 0 + \beta_D 0 + \beta_E 0$$

となり、予測値は $-1.06\% + 0.71\% = -0.35\%$ となります。

複数の項目を含む場合、数式はより長いものとなります。あえて書けば以下ようになります：

$$Y = \beta_0 + \left(\beta_A^{(1)} X_A^{(1)} + \dots + \beta_E^{(1)} X_E^{(1)} \right) + \left(\beta_A^{(2)} X_A^{(2)} + \dots + \beta_E^{(2)} X_E^{(2)} \right) + \dots$$

回帰分析の場合、単純な組み合わせと比べればより多くの項目を同時に予測に用いることができます。ただし際限なく項目を増やすと予測精度はかえって低下するため、重要度の高い項目のみを適切な個数だけモデルに含めるといった工夫が必要です。

代表的な方法がステップワイズ法で、決定木と同様に「もっとも効果の高い項目からひとつずつ」モデルに追加し、「これ以上は予測精度が向上しない」と判断したところで止める、といった手順を踏みます（逆にひとつずつ減らす方法もあります）。ほかにもElasticNet法など、類似の概念でさまざまな手法が存在します。

例として、重要度の高い順にモデルに項目を追加した結果、以下のような項目の組み合わせが得られました：

経常利益_予想評価_対前期実績_通期
当期利益_予想評価_対直近コンセンサス_通期
移動平均乖離率75日_前日
経常利益_実績評価_対直近予想
当期利益_予想評価_対前期実績_通期
経常利益_実績評価_対前期実績
営業利益_予想評価_対前期実績_半期
営業利益_実績評価_対直近コンセンサス
当期利益_実績評価_対直近予想
売上高_予想評価_対前期実績_通期
営業利益_予想評価_対直近コンセンサス_通期
移動平均乖離率5日_前日
営業利益_予想評価_対前期実績_通期
当期利益_実績評価_対前期実績
売上高_実績評価_対前期実績
営業利益_実績評価_対直近予想
発表実績_総資産経常利益率(ROA)
発表実績_株主資本当期利益率(ROE)
売上高_実績評価_対直近コンセンサス
ROA_対前期差分
ROE_対前期差分

発表実績_売上高営業利益率
売上高営業利益率_対前期差分
当期利益_予想評価_対前期実績_半期

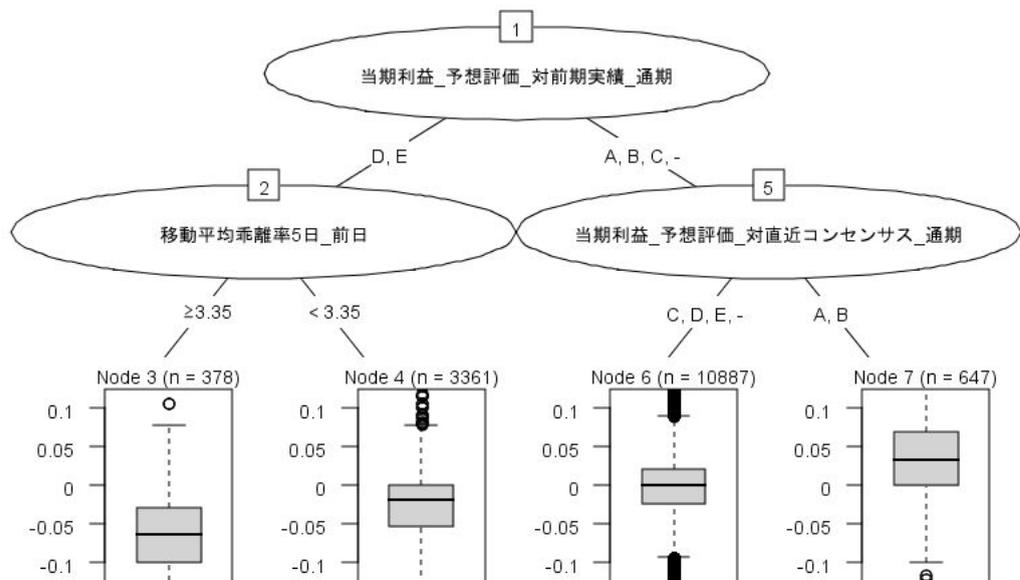
このように多くの項目を組み合わせてつづ、かつ不要なものや重複する情報を省略することで精度のよい予測が実現されます。

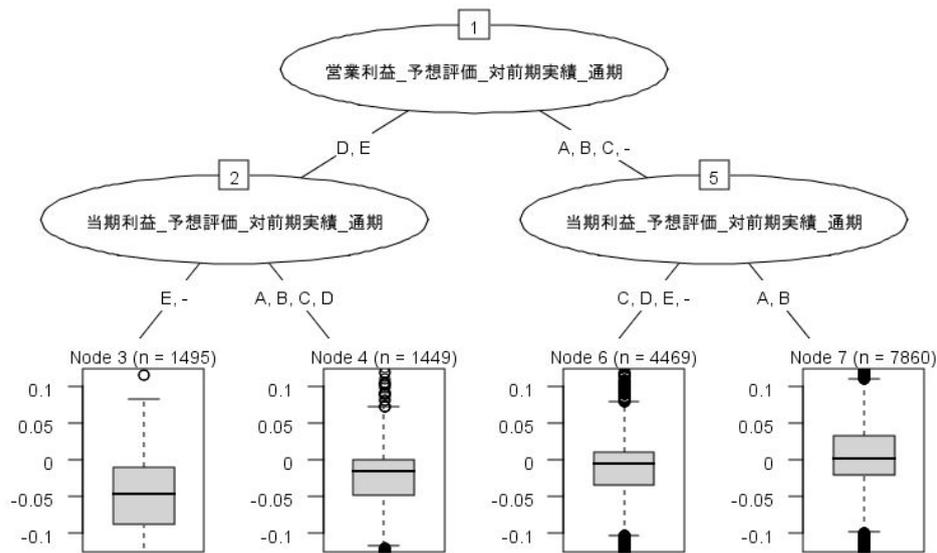
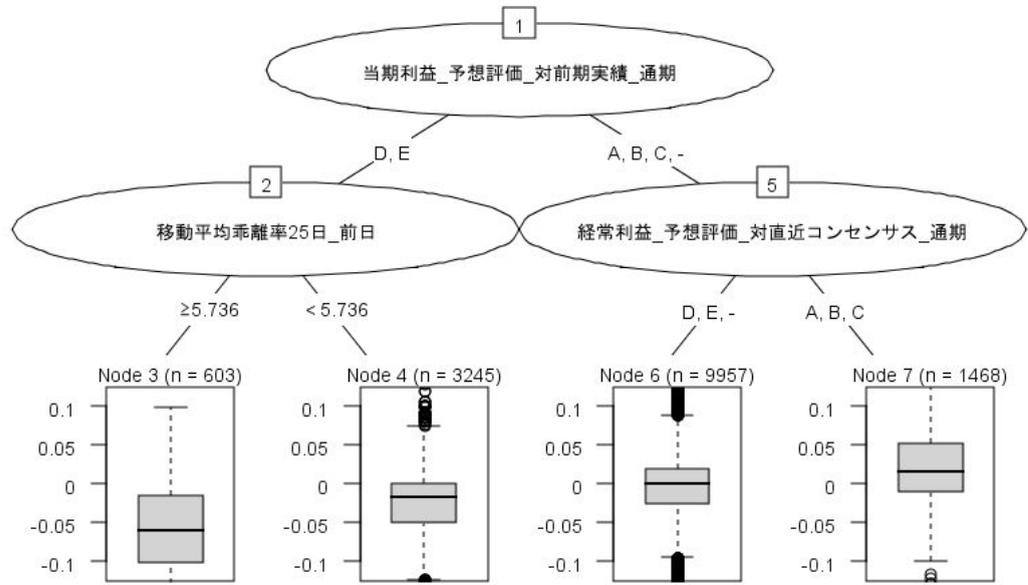
③ランダムフォレスト

ランダムフォレストは決定木にもとづくアルゴリズムですが、以下の特徴を持ちます：

- 元のデータからレコード、項目ともにランダムにサンプリングし、複製データを作成する
- 複製データから決定木を作成する。これを多数回(例：500回)繰り返す

ここまで見てきた例では、以下のように「似ているが異なる」決定木モデルが多数生成されます：





スコアを算出する際には、まず得られた多数の決定木それぞれでスコアを算出します。決定木の数だけ得られたスコア(例:500個)を平均することで最終的なスコアが得られます。

それぞれの決定木が少しずつ異なるため、決定木に比べて多くの項目が利用されます。かつ、多くのスコアを平均することでノイズを抑える効果もあると考えられます。

他にも多くの手法が存在しますが、いずれも上記でみたような「精度の高い予測をするため、重要な項目だけを適切に組み合わせる」というメカニズムを持っています。

複数のモデルを作成する場合、通常もっとも精度の高いものを選びます。これをモデル選択とよび、AIスコアにおいては以下の点を重視して、現時点で精度が高いものを選んでいきます。今後も定期的にモデルの作成やチューニングなど、見直しを行う予定です。

- 平均的な精度。スコアが実際の株価変化率とどの程度離れているか。予測誤差を二乗して平均した「平均二乗誤差」などを用いて評価します。
- 予測の幅。「この区間に実績値の90%が入る」といった区間をモデルから生成したとき、実際に90%に近い実績値がその区間に入るか

以上

・本情報に関する著作権を含む一切の権利は、株式会社QUICK（「QUICK」）またはその提供元（「情報源」）に帰属します。

・本情報は信頼できると判断される情報をもとにQUICKが提供したものです。その正確性、完全性を保証するものではありません。本情報の表示、更新は、システム上の理由（保守、障害復旧、サービス改変など）によって、遅延、中断することがあります。本情報によって生じたいかなる損害についても、QUICKおよび情報源は、一切責任を負いません。

・本情報は投資判断の参考としての提供を目的としているものであり、投資勧誘を目的にしたものではありません。記載内容は提供日時点のものであり、将来予告なしに変更されることがあります。

・本情報は過去の運用実績ならびに過去の運用実績に基づき計算されたものであり、将来における運用成果をお約束・保証するものではありません。また、手数料、税金等のコストを考慮していません。

・本情報は、閲覧者ご自身のためにのみご利用いただくものとし、第三者への提供は禁止します。また、本情報の内容について、蓄積・編集・加工・転用・複製等を禁止します。
