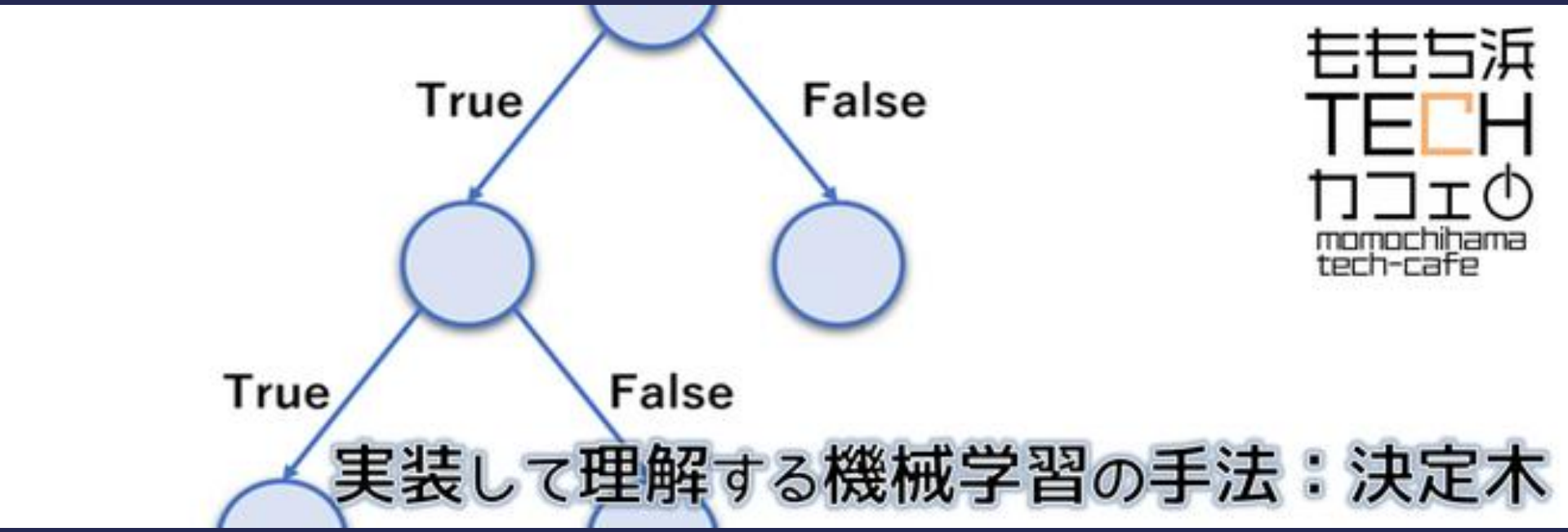


ももち浜TECHカフェ
momochihama tech-cafe

2019.7.16
@福岡システムLSI総合開発センター

今日の目標：動作原理を学ぶ

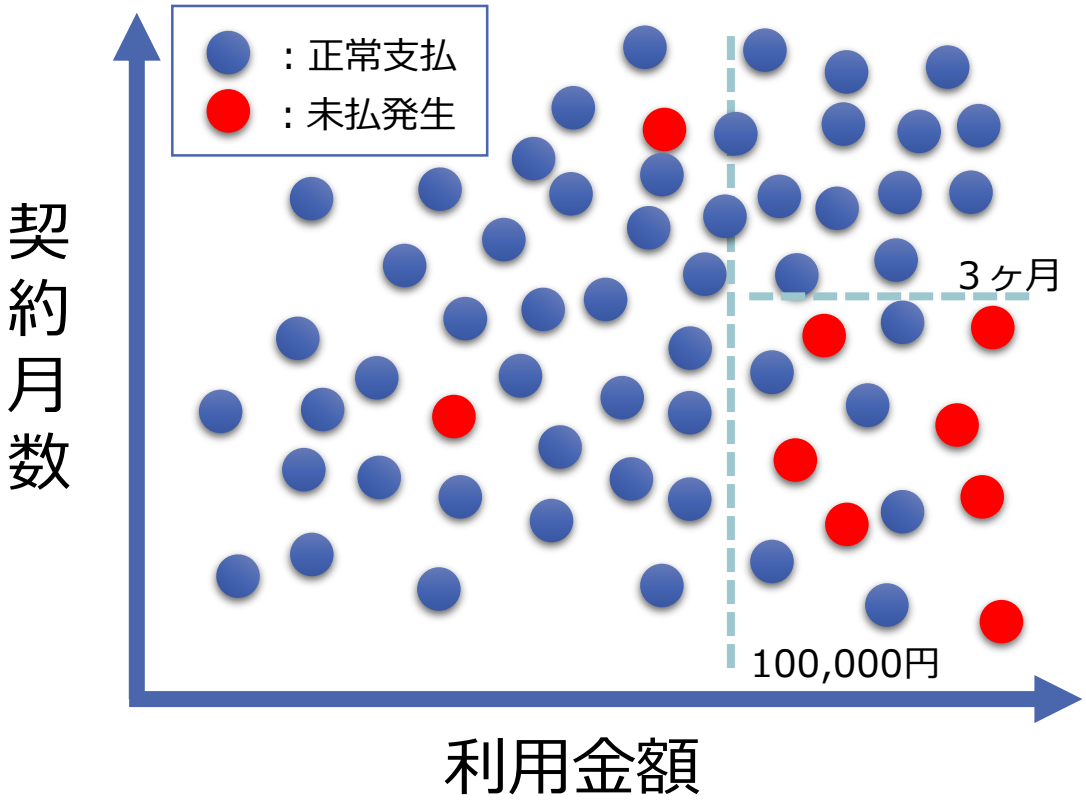
我々は技術者です。技術者にとっての基礎とは何でしょう？
私は「**動作原理を理解する事**」であると考えています。

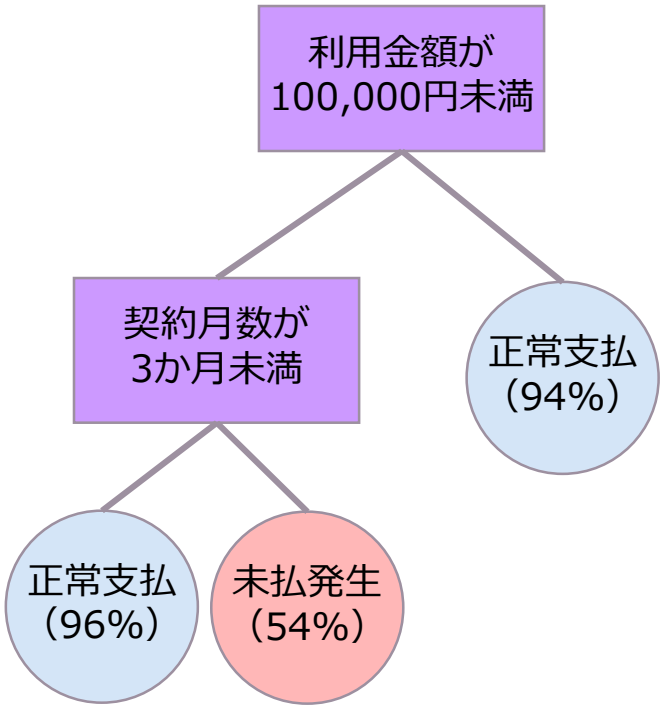
技術者は自分が作り出したものを改修し、さらにいいものを作っていく役目があると思います。このとき、動作原理をしらない部分は改修ができません。

そこで今日は、あえて**１手法の実装を行う**事により、その動作原理の理解の一助にできればと考えております。

決定木とは

目的変数の予測ロジックを、説明変数のロジックツリーの形で表現。
高い解釈可能性をもち、人間に対する説明が必要となる要件が
存在ケースでは利用されることが多いです。

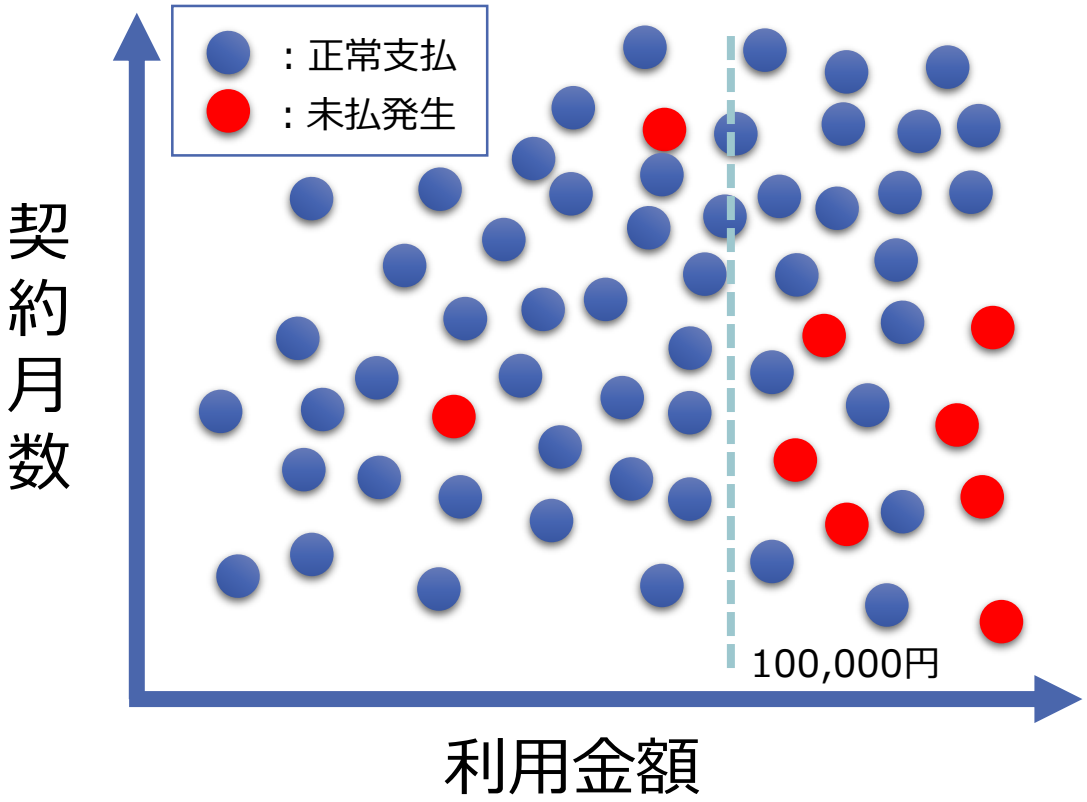


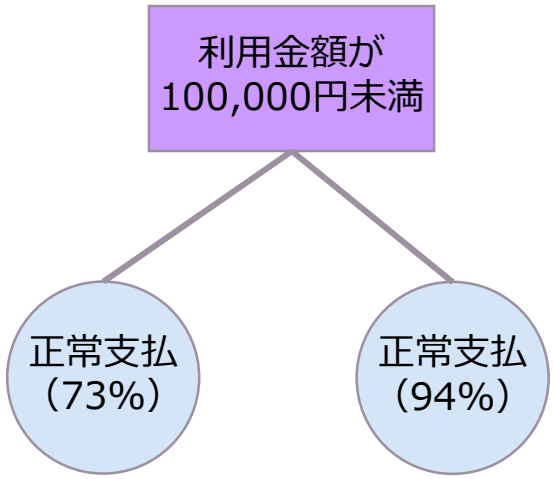


※ 左がFalse、右がTrue

構築手順１：ベストな分割を繰り返して木を育てる

全ての説明変数について、ある閾値をもって集団を２群（以上）に分けベストな分割を探します。分割後のそれぞれの群について、さらに同様の操作を行い、逐次的に木を育てていきます。

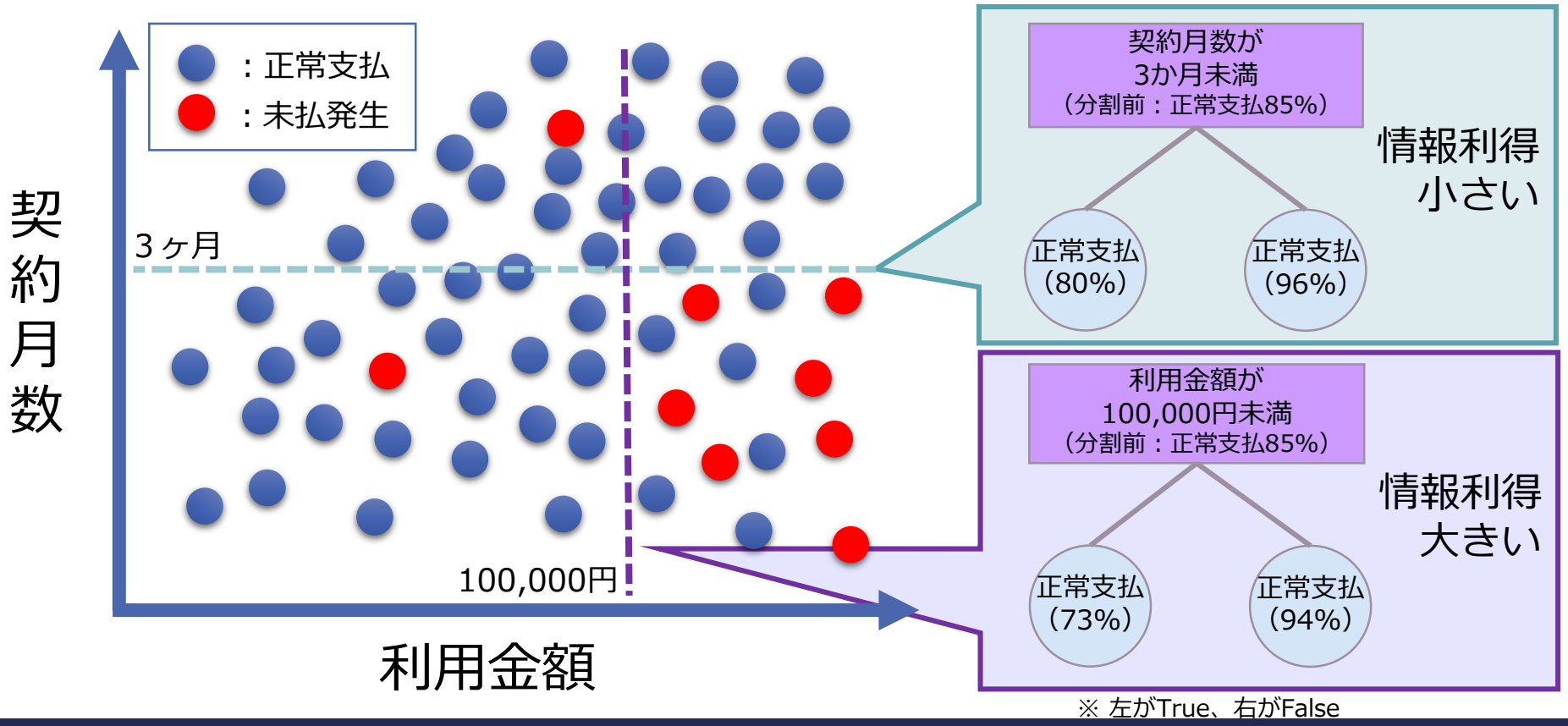




※ 左がFalse、右がTrue

何をもって「ベスト」か？

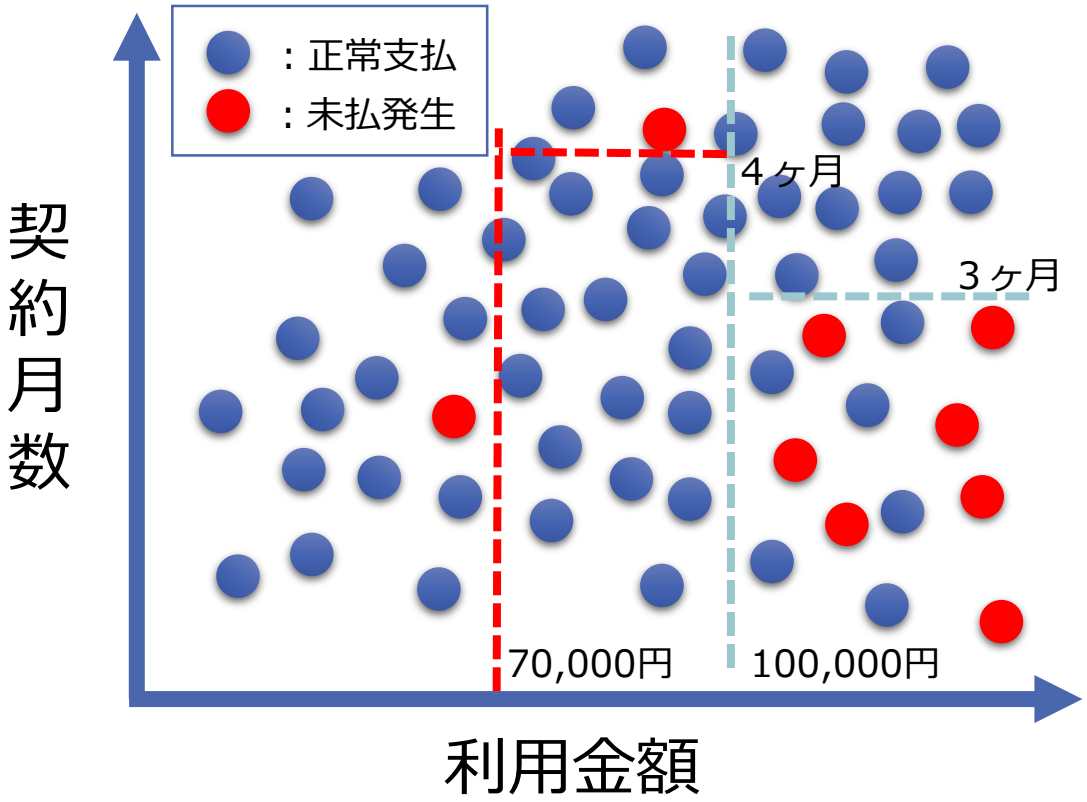
情報利得（定義式は後述）が大きい分割をもって良しとします。
情報利得とは直感的には２群に分けたときに得られる情報量の大きさ。

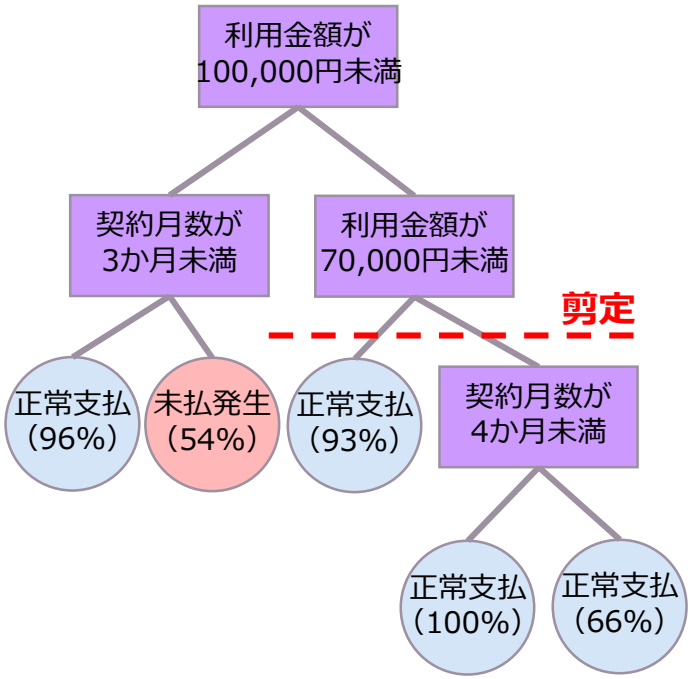


※ 左がTrue、右がFalse

構築手順２：育ち過ぎた木を剪定する

育ち過ぎた木を放置しておくと、単なる外れ値にもフィットしてしまい、
かえって汎化性能を失うことになる（過学習）ので剪定を行います。





※ 左がFalse、右がTrue

決定木に必要な機能

決定木モデルを構築するために必要な手順は…

- ベストな分割を繰り返して木を育てる
- 育ち過ぎた木を剪定する

…である事を説明しました。

これは、そのまま実装に必要なのは…

- fit：モデルを構築する機能
  - build：ベストな分割を繰り返して木を育てる機能
  - prune：育ち過ぎた木を剪定する機能

…である事を意味します。（これ以外に構築した決定木を使って実際に予測を行う機能：predict、木の構造を表示する機能：print が必要になりますが、これらは理解の本質ではないので軽く説明します。）

これから Colaboratory を使って実際に実装を進めます。

最後に

今回の内容を理解（時間の都合上、一緒に実装しなかった部分も再確認）することで、決定木を実装する事は可能になった事と思います。

例えば sklearn ですら以下については実装されていませんが、今回の内容をマスターすれば、これらは全て自力で実装可能です。

- ✓分割の評価をカイ２乗値で実行（CHAIDといいます）したり、Upliftを計算して実行（Uplift Modelingといいます）したりする（あるいは自分で分割を評価する式を考案して実装する）。
- ✓１つのノードから３分割以上の子ノードにわける（本来は可能）。
- ✓カテゴリ変数を説明変数として利用する。

原理をしっていれば、自分で改良する事もできます。

本日のセミナーをキッカケに決定木について、ひとつ深く知ってもらい、機械学習の楽しさを体感していただければ幸甚です。