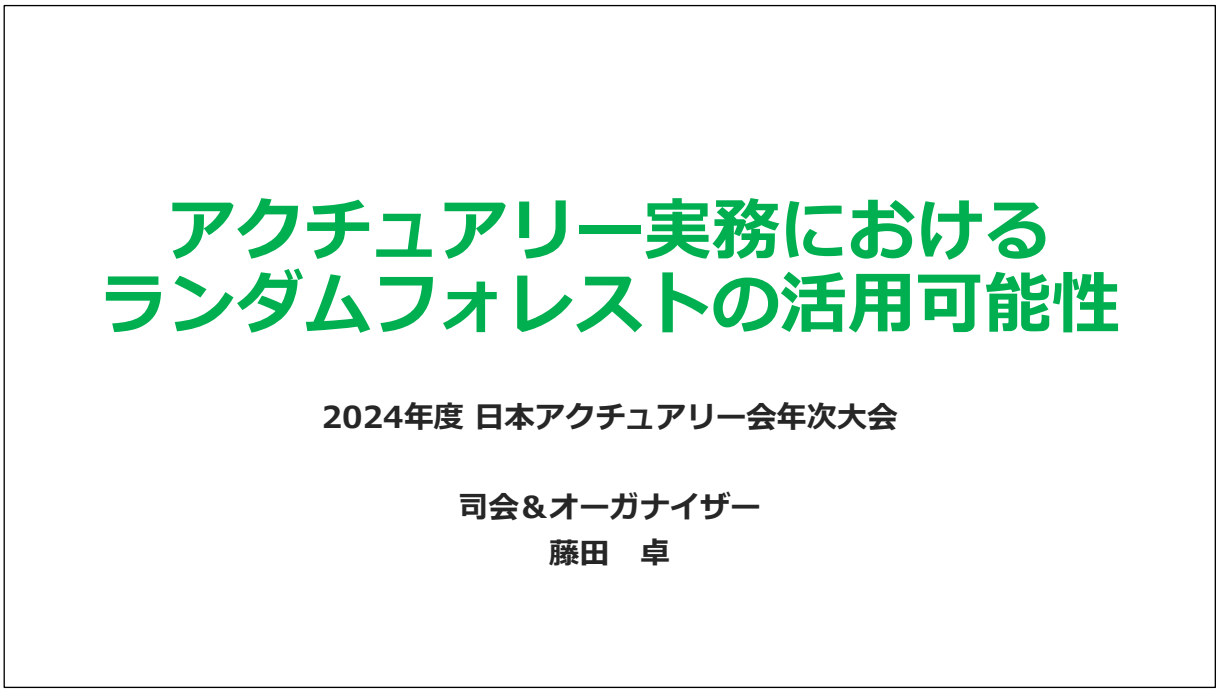
アクチュアリー実務におけるランダムフォレストの活用可能性
＜ASTIN 関連研究会＞

AKUR8　藤田　卓　君
三井住友信託　小田　直人　君
松江　康紘　君
田中　豊人　君



【藤田】　皆さん、こんにちは。セッション A-4「アクチュアリー実務におけるランダムフォレストの活用可能性」にお越しいただき、まことにありがとうございます。

本セッションのオーガナイザーを務める、藤田卓と申します。

アジェンダ

1. イントロダクション
2. パネリストによるプレゼンテーション（20分*3）
3. Q&Aセッション（10分）
4. パネルディスカッション（15分）

2

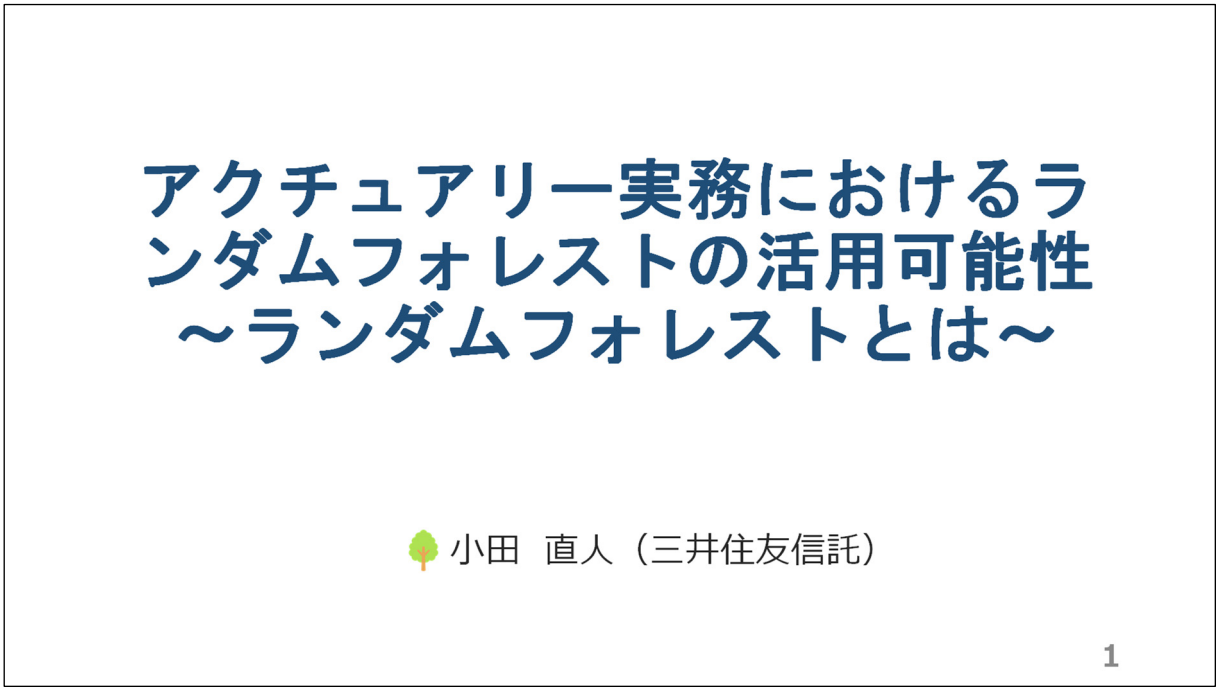
本セッションは、ASTIN 関連研究会主催のセッションとなります。時間は 90 分で、前半、3 名のパネリストをお迎えして、ランダムフォレストにおける各テーマのプレゼンテーションを行っていただきます。そちらを大体 60 分ほど予定していまして、その後に質疑応答の時間を設けております。その後、パネルディスカッションという形式で、15 分程度、ランダムフォレストの実務への活用可能性についてディスカッションを行います。

プレゼンテーション

3

まず、小田直人さんから「ランダムフォレストとは」という題名で、ランダムフォレストの基礎的なところについてお話しいただきます。その後、松江康紘さんから、「Random Planted Forests」という、ランダムフォレストを解釈可能性の観点から拡張を行ったものについてご紹介いただきます。最後に、田中豊人さんから、生存時間分析とランダム・サバイバル・フォレストのご紹介という題名で、ランダムフォレストを活用した生存時間分析について発表いただきます。

それでは、早速まず小田さんから、プレゼンテーションをお願いできますでしょうか。



【小田】　ご紹介にあずかりました、三井住友信託の小田と申します。私からは、ランダムフォレストの基

本的なところにつきまして、15 分程度お時間をいただきまして説明させていただきたいと思います。

本日の説明内容

1. 「ランダムフォレスト」とはどのようなものか
2. ランダムフォレストとアクチュアリーとの親和性
3. ランダムフォレストと他の手法との比較
4. ランダムフォレストのアクチュアリー実務への応用可能性

2

本日の説明内容ですけれども、四つほどご用意しております。一つ目は「ランダムフォレストはどのようなものか」ということについての簡単な仕組みを説明いたします。二つ目としまして「ランダムフォレストとアクチュアリーとの親和性」、三つ目としまして「ランダムフォレストと他の手法との比較」を簡単に説明いたします。四つ目としまして「ランダムフォレストのアクチュアリー実務への応用可能性」を説明いたします。実務の応用については、まだまだこれからというところなのですけれども、幾つか可能性をお示ししたいと思います。

本日の説明内容

1. 「ランダムフォレスト」とはどのようなものか

2. ランダムフォレストとアクチュアリーとの親和性
3. ランダムフォレストと他の手法との比較
4. ランダムフォレストのアクチュアリー実務への応用可能性

3

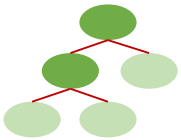
それでは、一つ目のテーマです。「ランダムフォレストについて」ということでございます。

1．「ランダムフォレスト」とはどのようなものか

「ランダムフォレスト」とは、・・・

・数多くある機械学習手法の中の一つ

・機械学習手法の中での分類としては、「決定木」（ツリー）
を元にした手法の一つに分類される



4

ランダムフォレストは、数ある機械学習のうちの一つとなっています。その機械学習の中でもツリー系と言われているものの一つということになっています。

1．「ランダムフォレスト」とはどのようなものか

ランダムフォレストの仕組みは、以下のとおり。

1．フォレスト
- １つの決定木だけから推定するのでは推定値に偏りが出るので、多数の決定木を作成してその集合体として「フォレスト」を作り、その平均値を取ることで精度を上げる。

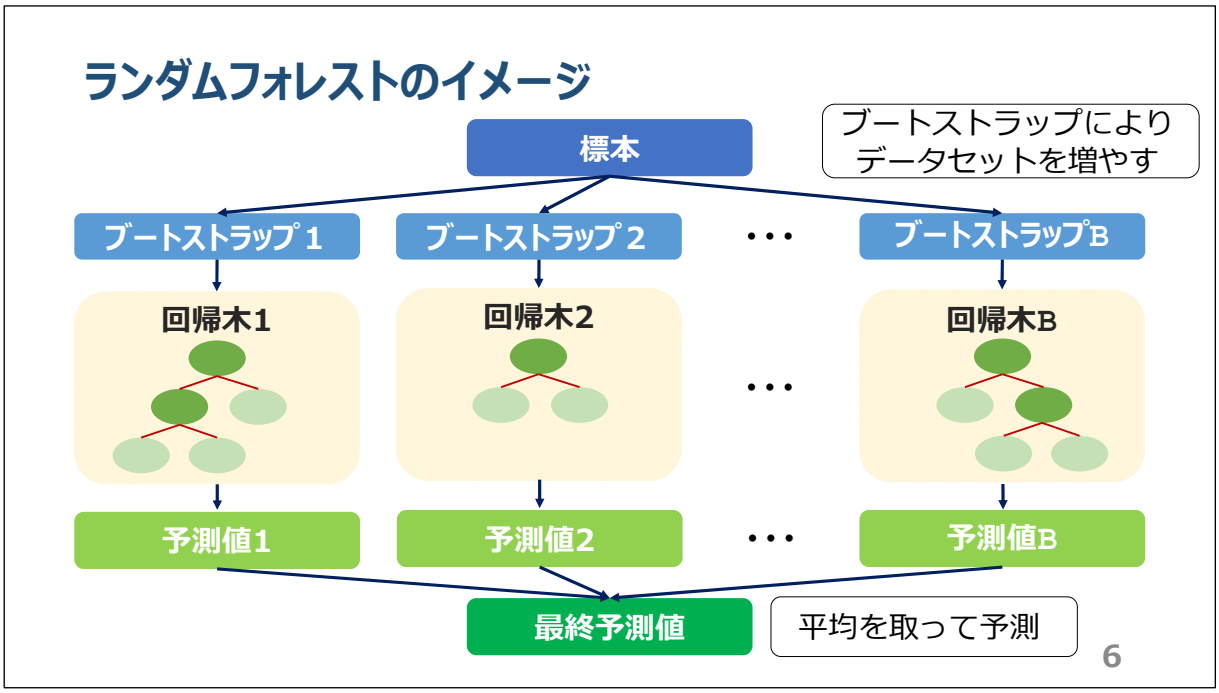
2．ランダム（標本のランダム性と特徴量選択のランダム性）
- 「ブートストラップ」という手法（元の標本からランダムにデータを復元抽出して新たな標本を作る）を使って、ランダムに標本を作成
- 分岐に使われる特徴量のランダムな選択：各決定木において、予め指定された個数の特徴量が、分枝のたびにランダムに選ばれ、最良のものが分岐に使われる

5

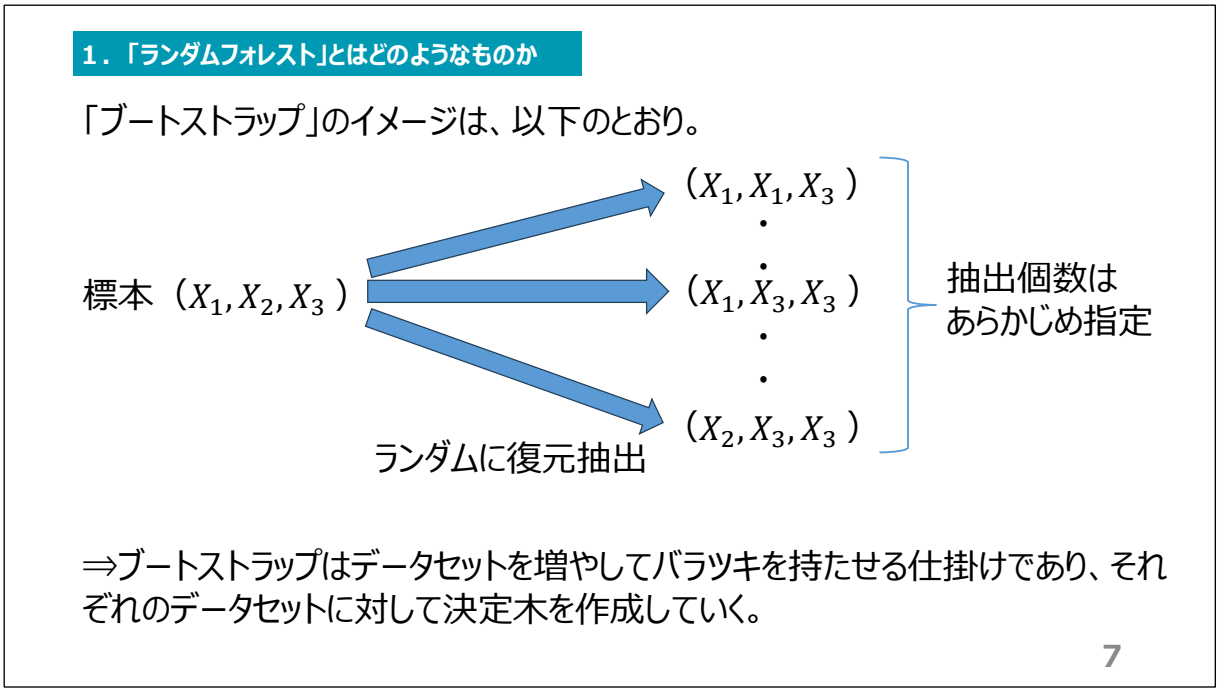
ランダムフォレストの特徴ということなのですけれども、この名のとおり「ランダム」と「フォレスト」、この二つの特徴があるということでございます。

まず「フォレスト」の方です。「フォレスト」は「森」という意味なのですけれども、多くの木を作って、そこから森を作って、森の中で平均を取ることで精度を上げていくという方法となります。一つの木だけから推定すると分散が大きくなるというところなのですけれども、多くの木を作って平均を取るということで分散を小さくすることが特徴になります。

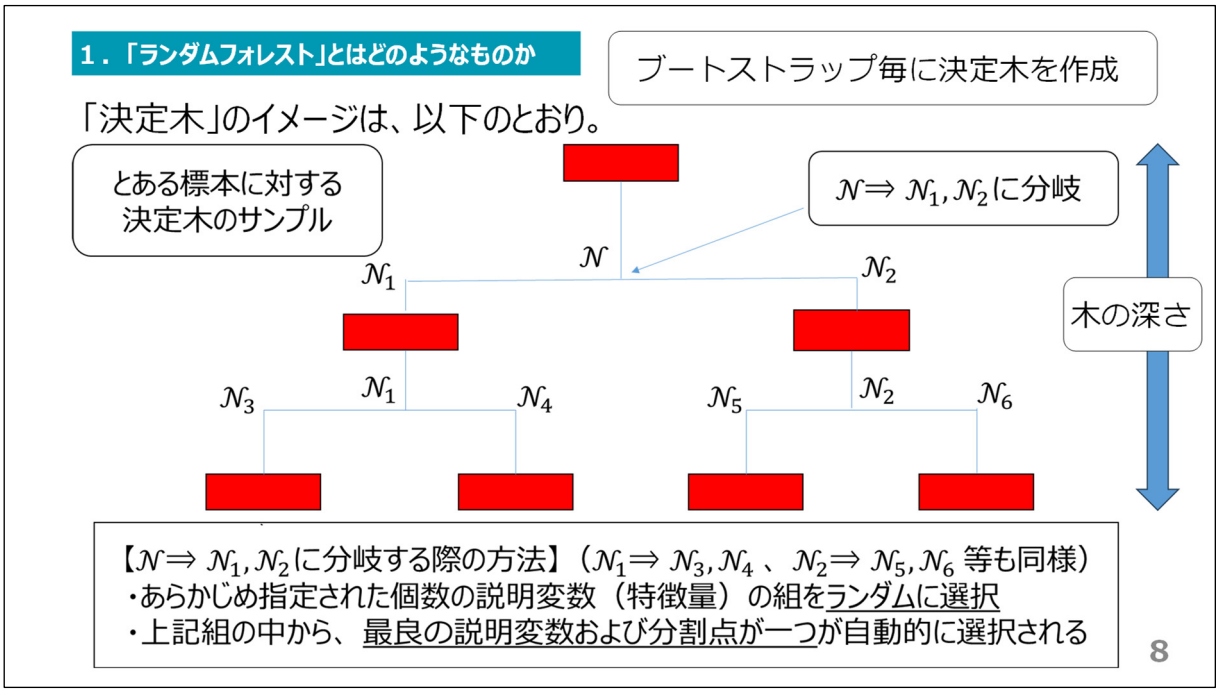
二つ目は「ランダム」です。「ランダム」というときに二つのランダム性があります。一つ目は標本のランダム性、二つ目は特徴量選択のランダム性になります。一つ目は、標本のランダム性なのですけれども、多くの木を作るにあたってブートストラップという手法を使うところでランダム性が使われます。これは後でお話しいたします。二つ目のランダム性は、木の枝を作る際のランダム性ということになりまして、これも後ほどお話しいたします。二つのランダム性があるということでございます。



ランダムフォレストのイメージです。このような形で標本がありまして、その標本に対してブートストラップでたくさん標本を作りまして、それを基に木をたくさん作って、それで、最後、それぞれ予測値を作って、その平均を取って最終予測をするということで、このようなイメージになります。

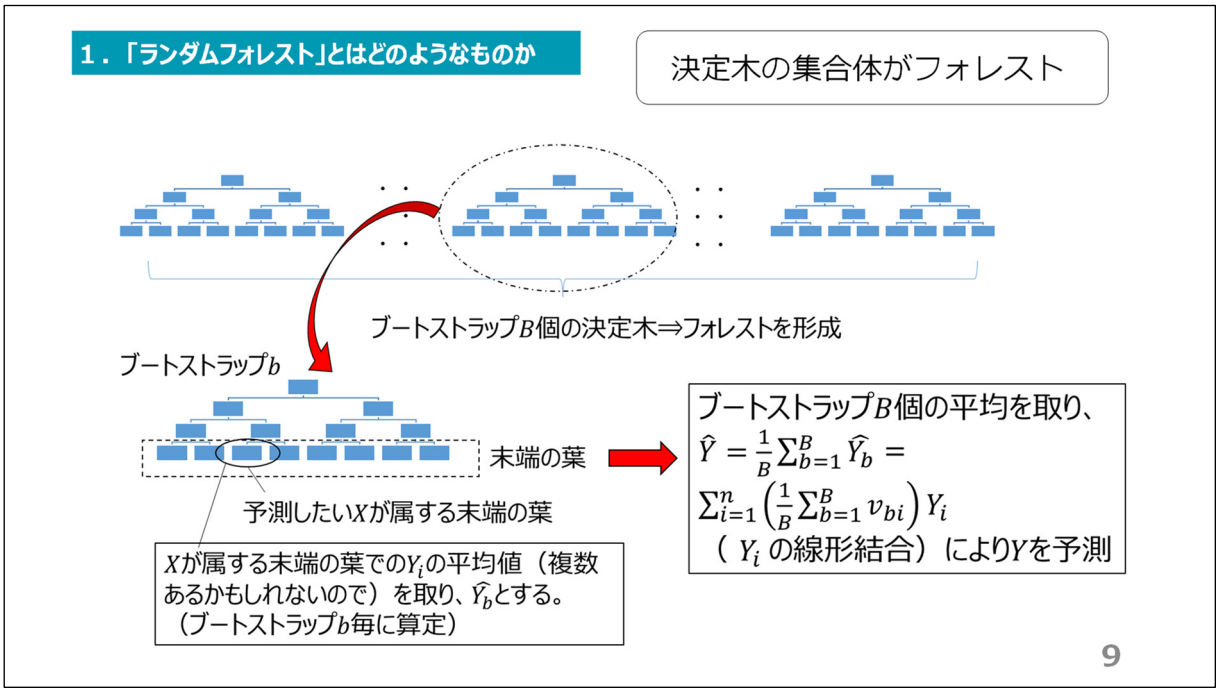


「ブートストラップ」という言葉を連呼してきたのですけれども、ここでは「ブートストラップ」のイメージを示しています。三つだけ標本があるという前提、実際にはもっと、多く標本はあるのですけれども、ここでは三つの標本に対してブートストラップをイメージしているのですけれども、ランダムに復元抽出をするということでデータセットをたくさん作るということになります。復元抽出なので、同じ標本が多く採用されます。今回のサンプルでは X_1 が二つ出てくることになるのですけれども、このような形で多くのデータセットを作るということになります。

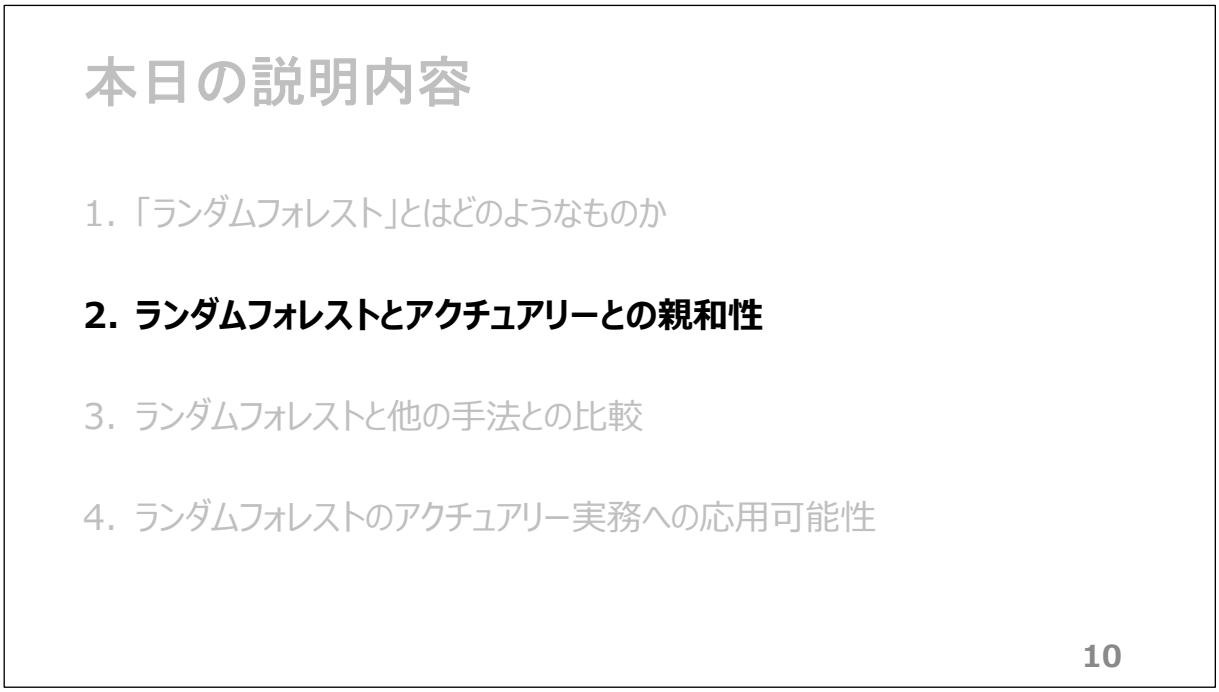


多くのデータセットを作って、それぞれに対して木を作っていくということなのですけれども、その木をどう作るかというところのイメージを示しています。木を作るには枝を作っていく必要があります。この四角囲いの一つ目のところに書いてありますけれども、木の枝を作る際に、全部の変数の中から分岐する際の変数を一つ選ぶというわけではなく、あらかじめ指定された個数の変数の組をランダムに選択して、それを基に分岐していく。このように制限を加えるということが特徴になります。

もう一つ、木の深さについても制限を加えることになっています。これだけではないですけれども、主な特徴はこの二つです。このように制限をいろいろ加えることで、似たような木を多く作らないようにするということになりますし、学習しすぎないようにするといった工夫がされているということになります。



木を作った後、どうするのかということです。木を作った後は、それぞれフォレストを作って、そのフォレストについて予測値を使って平均を取って最終予測をする、このような流れ図になっているというところでございます。



以上がランダムフォレストの簡単な説明でしたが、二つ目のテーマとしまして「アクチュアリーとの親和性」を見ていきたいと思います。

2．ランダムフォレストとアクチュアリーとの親和性

アクチュアリーが業務を行うのにあると良いものは、・・・
（他にもいろいろあるとは思いますが・・・）

・理論的にしっかりした統計的性質をバックに算出すること

・算出結果が解釈しやすく説明しやすいこと

・予測精度が高いモデルを使用していること

11

アクチュアリーの業務と言っても一口に言えないのですけれども、ここに示してあるような三つの特徴があると良いということかと思っています。

一つ目としては「統計的性質がきちんとバックにある」ということでございまして、よく分からないけれども当たるといったことではないというところです。二つ目としては「解釈しやすい、説明しやすい」というところでございます。このような性質は、社内外の説明には大事なところになってきます。それから三つ目としては「予測精度が高い」というところです。せっかく推定してもそれほど当たらないということでは、不都合というところになるかと思います。

ランダムフォレストが、これらの性質を満たしているかどうかというところなのですが、それを一つ一つ見ていきたいと思います。

2．ランダムフォレストとアクチュアリーとの親和性

「理論的にしっかりした統計的性質をバックに算出すること」について

・アクチュアリー業務の理論的な根拠には、「統計学」がある。例えば、点で推定するだけであれば統計学が無くても困らないかもしれないが、区間で推定するには何らかの統計学が必要である等々。

・機械学習の一つであるランダムフォレストには一見統計的性質が無さそうだが、ランダムフォレストの特徴を活かして「誤差分布の近似」ができる。

・「誤差分布の近似」を使って、一定の条件の元で、予測値の一致性（データを増やせば真のものに近づく）が示される。

・こうした統計的性質があることは、使用する上での安心感につながる。

12

一つ目、統計的性質のところです。

ここでは詳しく話さないのですけれども、独自の特徴として、二つ目のところに記載しているのですけれども、「誤差分布の近似がランダムフォレストではできます」というところがあります。これができるとリスク量を測るために有効な手段を得ることができるというところになりまして、このような「統計的性質を十分持っているとランダムフォレストは言える」と考えられると思っています。

2．ランダムフォレストとアクチュアリーとの親和性

「算出結果が解釈しやすく説明しやすいこと」について

・ランダムフォレストに限らず、機械学習モデルの各特徴量の重要度合いや、予測結果を解釈するための技術の研究が進んでいる（Interpretable Machine Learning）

・ランダムフォレストの場合は、特徴量重要度と呼ばれる指標を出力する機能が、種々のプログラミング言語のパッケージに内蔵されていることが多い

・ Interpretable Machine Learningの進展により、従来解釈しにくいとされてきた機械学習モデルの解釈可能性が向上

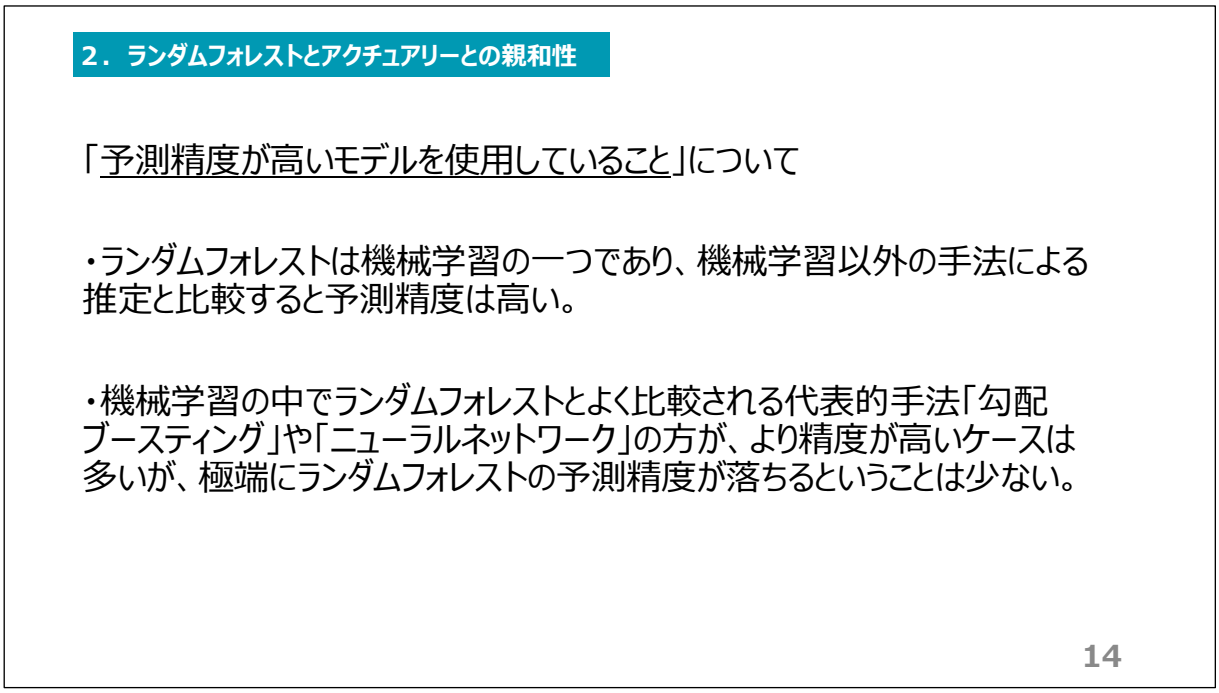
・ただ、機械学習ではあるので、限界はある。

13

それから二つ目のところですけれども、「解釈しやすく説明しやすい」というところです。

ランダムフォレストも機械学習の一つなので、やはり限界はあるのですけれども、一つ目、二つ目、三つ目に記載しておりますように、「Interpretable Machine Learning」といったところが発展してきていまして、

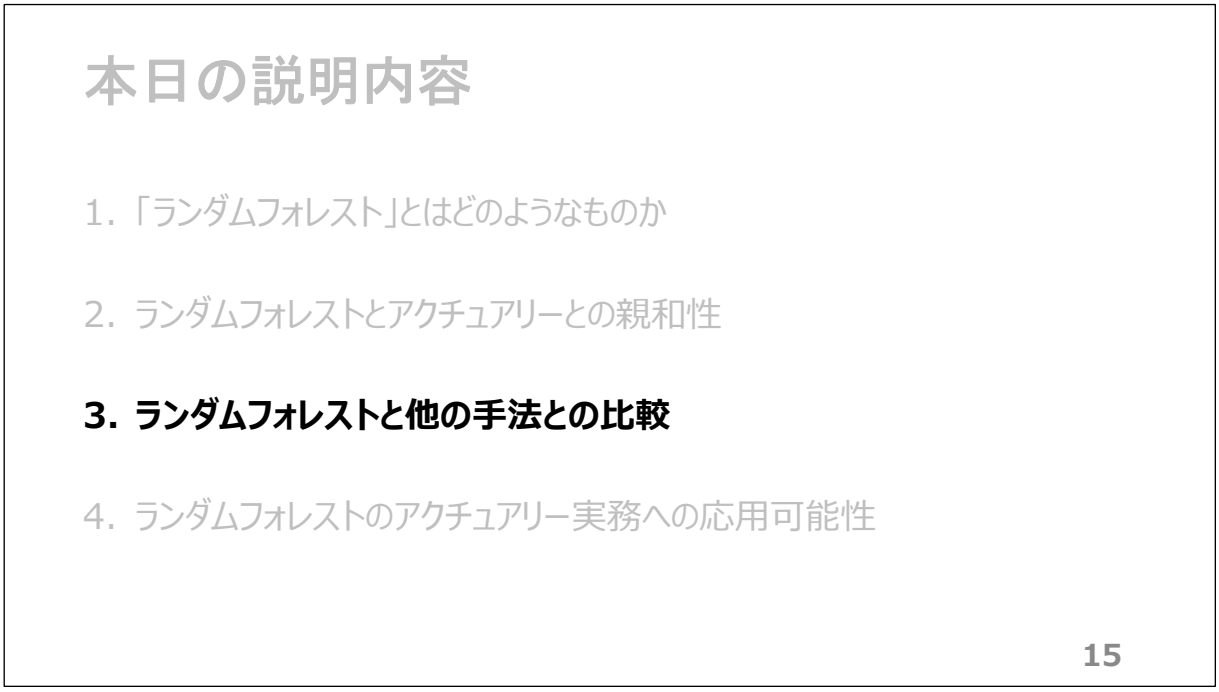
ランダムフォレストに対応したいろいろな準備がされているというところです。これも満たされているのではないかというところです。



三つ目として「予測精度」のところです。

後で出てきますが、他の機械学習と比べて非常に高いというわけでは必ずしもないのですけれども、機械学習以外と比べると高いと言えると思いますし、それなりに予測精度は高いと言えると思います。

ということで、ランダムフォレストはアクチュアリーと親和性があると言っていいのではないかと思っています。



三つ目のテーマ、「他の手法との比較」を説明いたします。

3．ランダムフォレストと他の手法との比較

アクチュアリーが伝統的に行ってきた機械学習以外の手法もしくは機械学習の他の手法とランダムフォレストの手法との比較を行う。

比較にあたっては、主に以下の観点で行う。

・予測精度
・統計的性質のバックグラウンド
・結果の解釈可能性
・取り扱いやすさ
・結果の安定性、頑健性
・外挿（＊）が可能かどうか

（＊）与えられたデータの範囲外で推定を行うこと

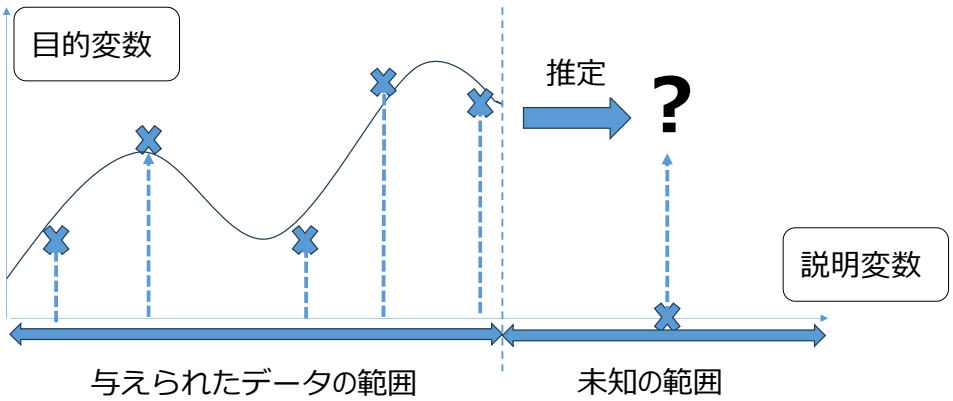
16

他の手法との比較にあたって、観点を六つほど挙げています。最初の三つは先ほど来見てきたものなのですけれども、四つ目の「取り扱いやすさ」、五つ目の「安定性」というものを加えております。また、六つ目として「外挿ができるかどうか」を加えています。

3．ランダムフォレストと他の手法との比較

外挿とは、与えられたデータの範囲外で推定を行うことを指す。

例えば、以下の図で、与えられたデータの範囲から外れた値に対して推定を行う場合を指す。



17

この「外挿」は何かと言いますと、イメージ図を書いています。与えられたデータの範囲の外で推定する、これができるかどうかということになります。この図で言いますと、左半分が与えられたデータの範囲なのですけれども、そこから右半分を推定することができるかというところになります。

3．ランダムフォレストと他の手法との比較

ランダムフォレストは以下のとおりと考えられる（黄色は前記２で記載の性質）

性質	ランダムフォレスト
予測精度	比較的高い
統計的性質のバックグラウンド	あり
結果の解釈可能性	大きいが限界あり
取り扱いやすさ	ハイパーパラメーターの数はやや少なく扱いやすい
結果の安定性、頑健性	比較的安定的で頑健
外挿が可能かどうか	難しい

18

ランダムフォレストは、まずどうなのかと言うと、先ほど見てきたような形で最初の三つはあります。四つ目の「取り扱いやすさ」、これも、ハイパーパラメータの数がやや少ないというところで、取り扱いやすいと言えるかと思います。五つ目の「安定性」もあります。六つ目の「外挿」は難しいのですけれども、それ以外は満たされているというところになります。

3．ランダムフォレストと他の手法との比較

GLM（一般化線形モデル）については以下のとおりと考えられる。

性質	GLM
予測精度	（機械学習と比べて）高くない
統計的性質のバックグラウンド	あり。ただし、「一致性」なし。
結果の解釈可能性	大きい
取り扱いやすさ	扱いやすい
結果の安定性、頑健性	比較的安定的で頑健
外挿が可能かどうか	可能

19

一方で、GLM（一般化線形モデル）。これにつきましては、この表のとおりということになります。二つ目から六つ目まで優れた性質を GLM は持っているのですけれども、一つ目の予測精度のところ、これについては、ランダムフォレストと比較してそれほど高くないというところです。

3．ランダムフォレストと他の手法との比較

勾配ブースティングやニューラルネットワークについては以下のとおりと考えられる。

性質	勾配ブースティング	ニューラルネットワーク
予測精度	高い	高い
統計的性質のバックグラウンド	少ない	少ない
結果の解釈可能性	小さい	小さい
取り扱いやすさ	ハイパーパラメーターの数が多く、ランダムフォレストと比較して取り扱いにくい	ハイパーパラメーターの数が多く、ランダムフォレストと比較して取り扱いにくい
結果の安定性、頑健性	ハイパーパラメーターの数が多く、調整も難しく、調整次第によっては過学習が起きてしまう。	データが少ないと不安定。また、勾配消失等、学習が停滞する可能性があり必ずしも安定的とはいえない
外挿が可能かどうか	難しい	可能

20

一方で、他の機械学習、一般的と言われている勾配ブースティングやニューラルネットワーク、これと比べるとどうかということです。予測精度はフォレストより高いと言えるのだけれども、二つ目から五つ目のところ、これについてはランダムフォレストより劣るのではないかと思われます。

本日の説明内容

1．「ランダムフォレスト」とはどのようなものか

2．ランダムフォレストとアクチュアリーとの親和性

3．ランダムフォレストと他の手法との比較

4．ランダムフォレストのアクチュアリー実務への応用可能性

21

最後、「実務への応用可能性」について説明いたします。

４．ランダムフォレストのアクチュアリー実務への応用可能性

機械学習の実務への応用に際しては、一般的に、以下のような課題がある。

・データの量は十分あるか。

・データの質は十分高いか。

・予測値は正確でも、どのぐらいずれるかといった誤差の分布が不明であるため、資本を積む等の目的に適用しにくい。（ニューラルネットワーク等）

・算定結果が入力値に大きく変動し不安定である傾向があり、怖くて使えない。（ニューラルネットワーク等）

⇒

・１つ目、２つ目については、不可避の課題であり、実務への応用の際には注意しながら対応。

・ランダムフォレストについては、３つ目、４つ目の課題は問題ない。

22

「ランダムフォレストは意外と使えるのではないか」と思っていただけたかもしれないのですけれども、実務への応用可能性はどうなのかというところを見ていきます。一つ目、二つ目のデータの質・量について、機械学習の実務への応用に関しては課題があります。やはり、質・量がないといけません。これは、ランダムフォレストに限らず、そうなのですけれども。一方、三つ目の「統計的性質」や四つ目の「安定性」、これも大事です。これはランダムフォレストにはあるかと思います。

４．ランダムフォレストのアクチュアリー実務への応用可能性

ここでは、２つほど応用を紹介する。１つ目の応用は「予測誤差分解」。

【応用１】予測誤差分解

・プライシングやリザービングにおいて、ランダムフォレストを使用して予測モデルを構築することが考えられる。

・例えば、自動車保険であれば、年齢や車両の種類、地域等のデータからクレームの損害規模そのものを予測するだけでなく、予測誤差を推定し、その予測誤差分解を行うことが考えられる。

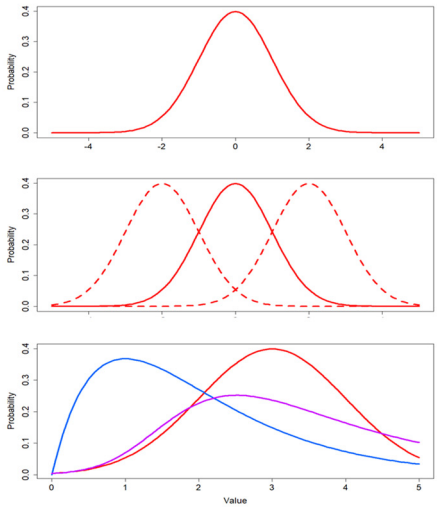
・予測誤差分解は、予測誤差をプロセス誤差（誤差分布に従う誤差であり避けられない誤差）、パラメータ誤差（パラメータ選択で生じる不確実性）、モデル誤差（モデルが正しくないことによる誤差）に分解するものであるが、ランダムフォレストを使用することでうまく分解できる。

23

このようなことを踏まえた上で、具体的な応用を二つ挙げるのですが、まず一つ目です。「予測誤差分解」というものとなります。時間の関係で詳細はお話できないのですけれども、ランダムフォレストの特徴を使って、予測誤差の推定に関して予測誤差の分解を行うことができるということがございます。

４．ランダムフォレストのアクチュアリー実務への応用可能性

予測誤差分解のイメージ図



＜プロセス誤差＞
モデル、パラメーターが正しいとしても消えない不確実性

＜パラメーター誤差＞
モデルが正しいとしてもパラメーター選択で生じる不確実性

＜モデル誤差＞
モデルが正しくないことによって生じるもの

24

「予測誤差の分解は何ですか」と言いますと、このスライドにありますように、プロセス誤差、パラメータ誤差、モデル誤差、このような三つの誤差に分解するということで、このように分解することで予測誤差をよりよく理解していこうということとなります。

４．ランダムフォレストのアクチュアリー実務への応用可能性

＜予測誤差推定を行うことの効用＞

○予測誤差推定を行うことで、プライシングについては、より安全な設定や、逆に、許容できる範囲で、より保険購入者に有利な設定を行うことができる。また、リザービングについては、予測誤差を考慮することが、将来の保険金支払いに備えて資本を積む際の目安になる。

25

予測誤差の推定のメリット、はスライドのとおりですが、これも時間の関係で詳細は割愛いたします。

４．ランダムフォレストのアクチュアリー実務への応用可能性

＜予測誤差を分解することの効用＞

〇プロセス誤差はデータに内在する不可避の誤差と考えられ、これを小さくしたり制御したりすることは難しいものと考えられる一方、パラメータ誤差は、データの有限性に起因する誤差であり、制御可能と考えられる。したがって、予測誤差分解を行うことにより、以下の効用があるものと考えられる。

・プロセス誤差を把握することにより、予測誤差を小さくしうる限界がどの程度かがわかる。

・パラメータ誤差を把握することにより、予測誤差の中で制御可能な部分がどの程度の割合を占めるかがわかり、どの程度データを集めるか等に役立てられる。

26

次のスライドで、予測誤差分解のメリットを記載しています。これも詳細は割愛いたします。予測誤差分解することでの色々なメリットを記載しております。

４．ランダムフォレストのアクチュアリー実務への応用可能性

２つ目の応用は、以下のとおり。

【応用２】ランダムフォレストを使用して検証を行う

・行政あて申請や商品への採用は、法令の縛りや現行の事務ルールとの関係性もあるため、なかなか難しい。

・一方で、正式計算前の検証として当たりをつける、もしくは、正式計算の検証を行うためにランダムフォレストを使用することが考えられる。

・また、既存のモデルのパラメーター等を推定するためにランダムフォレストを使用することも考えられる。

27

駆け足ですけれども、二つ目の応用については、ランダムフォレストを使用して検証を行うということで、具体的なものではなくて恐縮なのですけれども、このような観点があります。行政あての申請や商品化することを、いきなりやることは、なかなか難しいという面があるのですけれども、正式計算前の検証をしたり、正式計算の検証を行ったり、もしくはパラメータを推定したり、そういったもののためにランダムフォレストを使用することが考えられるということです。海外の文献でも、そのような方向性がございます。

４．ランダムフォレストのアクチュアリー実務への応用可能性

また、ランダムフォレストの応用というよりも、ランダムフォレストのバリエーションということになるが、以下のようなものがあり、この後、松江さん、田中さんから説明予定。

〇ランダムプランテッドフォレスト（RPF）
⇒ランダムフォレストに対して、説明可能性を高め、予測精度を向上させたもの
〇生存時間分析（ランダム・サバイバル・フォレスト）
⇒時間経過に伴う特定のイベント（病気、死亡、機械の故障等）の発生を分析する方法だが、ランダムフォレストを活用したものを一部使用

28

ということで、私からの説明は以上になるのですけれども、この後、ランダムフォレストをルーツに持つランダムプランテッドフォレスト、それからランダム・サバイバル・フォレストについて、それぞれ、松江さんと田中さんから説明していただきます。

ご清聴、ありがとうございました。

29

ご清聴、ありがとうございました。

【藤田】　小田さん、ありがとうございました。

それでは、引き続き、松江さんのプレゼンテーションに移りたいと思います。松江さん、ご準備ができ次第、お願いします。

アクチュアリー実務におけるランダムフォレストの活用可能性 ～Random Planted Forests～

解釈性と精度の両立

松江　康紘

1

【松江】 ご紹介にあずかりました、松江と申します。ランダムフォレスト RF の改善版である Random Planted Forests、ここでは「RPF」と呼びますが、そちらについて、ご説明させていただきます。

Agenda

1. ランダムフォレストを実務に活用する際の課題
2. Random Planted Forests(RPF)の概要
3. RPFの性能評価
4. RPFの実データ適用

2

章立てです。まずランダムフォレストの課題がございまして、それを説明した上で、Random Planted Forests（RPF）がどのように解決しているかをご説明し、人工データ、実データを用いた RPF の評価を行います。

サマリー

- ランダムフォレスト(RF)を実務活用するにあたり、解釈性・精度の両面で課題が残っている。
- RFに両面から改善を施したRandom Planted Forest(RPF)を紹介。
  低い次数の**交互作用の組み合わせごとに木を構築**し、それぞれ深い分割を許容することで、
  - モデルの予測値を、少数の変数で決まる**シンプルな関数の和として表現**できる。
  - RFが捉えることが難しい**線形関係なども、小さいバイアスで捉えられる**。
- 人工データによる精度評価で他アルゴリズムと比較。
  - **スパース（高バリアンス）な設定で高い精度**を発揮した。
  - 密（高バイアス）な設定では勾配ブースティングに及ばず。
- 自動車保険の実データでも精度評価を実施。
  - RPFはGBMに並ぶ高い精度となった。
  - 交互作用の次数を上げても精度の改善は見られず。
  - 推定された主効果はいずれも納得性の高い解釈が可能だが、**交互作用には軽微なもの、ノイズと思われるもの**もあり、実用上はある程度**絞り込む必要がある**と考えられる。

3

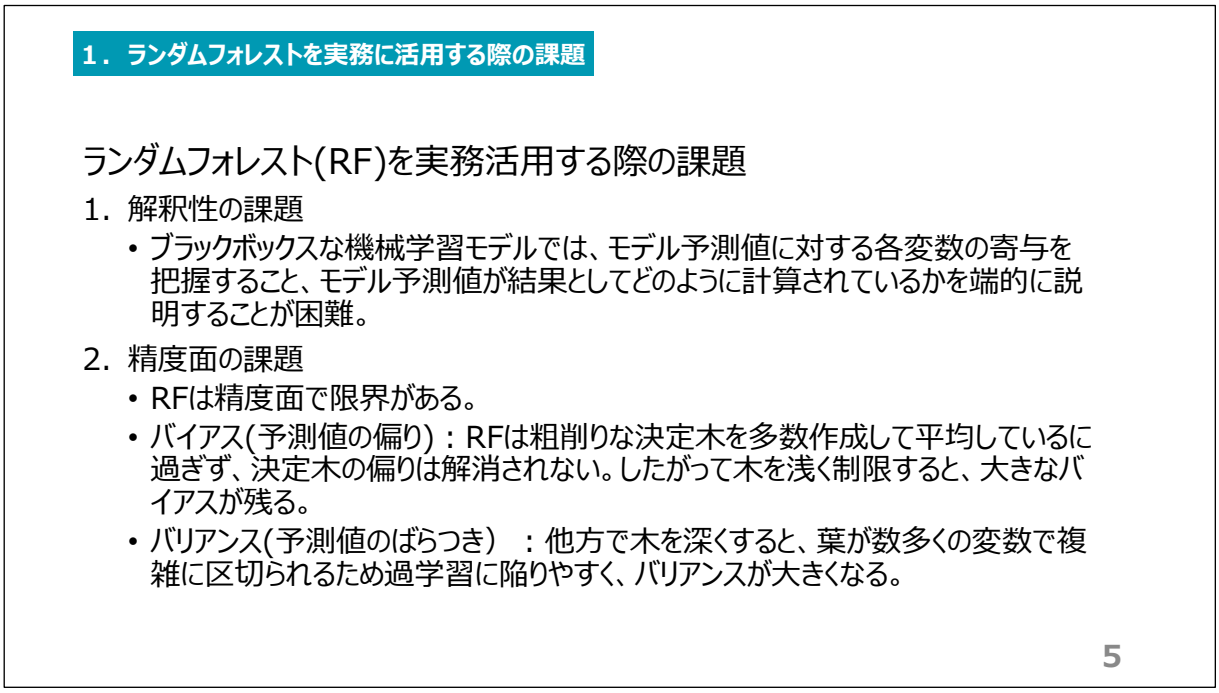
サマリーは、ご参考ですが、こちらのとおりとなります。

Agenda

1. ランダムフォレストを実務に活用する際の課題
2. Random Planted Forests(RPF)の概要
3. RPFの性能評価
4. RPFの実データ適用

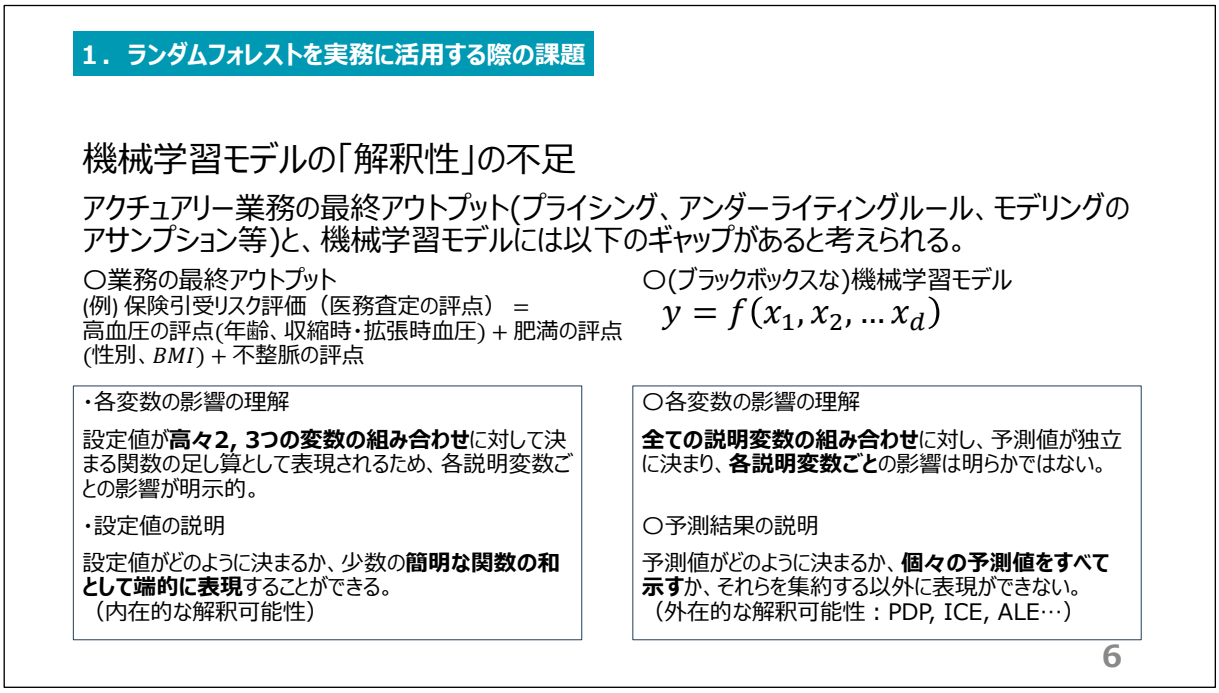
4

ランダムフォレストを実務に活用する際の課題につきまして、まず申し上げます。



二つありまして、一つ目が解釈性の課題がございます。先ほど、解釈しやすいとは申し上げたものの、ランダムフォレストに限らずブラックボックスな機械学習モデルは、各変数の寄与、つまりモデルの予測値に対して各変数がどう寄与しているのか、また予測値がどのように計算されているか、ブラックボックスなので、端的に説明する、分かりやすく説明することは、困難です。

精度面でも課題がありまして、ランダムフォレストは、荒削りでバイアス（予測値の偏り）が残る決定木を、ランダムにたくさん作成して平均をとるバギングという手法を使っているわけですけれども、この方法だと、決定木一つ一つのバイアスは解消されないといった課題がございます。これを無理に解消しようとして、木を複雑化、例えば木の深さを深くすると、決定木は同時に過学習しやすいアルゴリズムとして知られておりまして、今度はバリアンスが大きくなってしまう。バリアンスというのは予測値のばらつきで、データのノイズも学習しまい、汎化性能が落ちるのです。

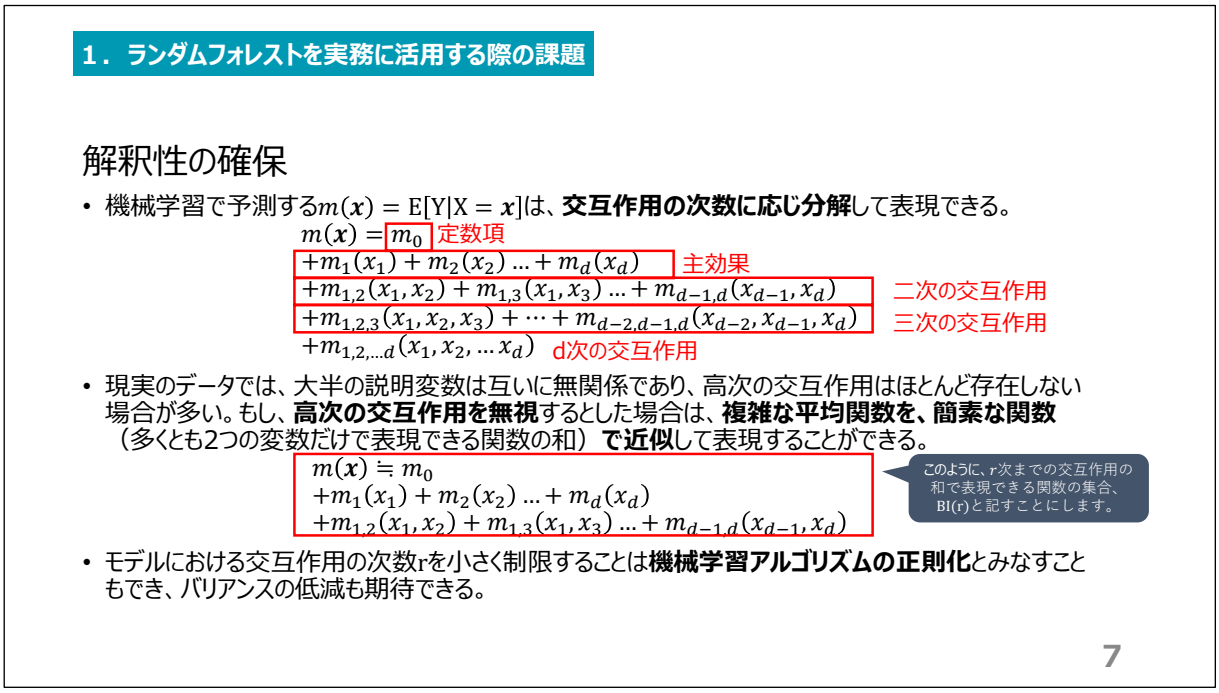


では、一つ目の解釈性につきまして、もっと詳しくご説明します。アクチュアリー業務の最終的なアウトプット、いろいろなケースがあると思うのですけれども、それと機械学習モデルの間にはギャップがあると、私は考えております。私は、医務査定の高度化に向けた分析経験がございますので、そちらの例を取ってご説明します。

アンダーライティングで、ある契約のリスク評価を決めるときは、一般的に、例えば高血圧や肥満など、リスク因子ごとのリスク評価の足し算としています。それぞれのリスク因子のリスク評価としては、健康診断項目や告知項目など限られた項目を使ってシンプルなルールで決まります。従って、例えば、高い血圧がどのようにリスク評価を高めているのかというようなところは明示的に見て取れますし、最終的なリスク評価の点数も、どのように計算されているか一つ一つの要素に分解してシンプルに説明することができます。

他方、ブラックボックスな機械学習モデルでとても厄介なことは、右にあるとおりですけれども、全ての説明変数の組み合わせを与えて初めて予測値が決まります。ですから、ある変数がどのように影響しているかは明らかではありません。あと、予測値の決まり方も、当然ながら、ブラックボックスなので、その過程、中身は分からないわけです。従って、全ての予測値、つまりテストデータに対する全ての個々の予測値を示すか、あるいは、それらを集約して表現する以外に説明の仕方がないのです。

巷には IML、つまり解釈可能な機械学習の手法として様々なものがございますが、こちらも、やはり全ての予測値を集約して表現するものにすぎません。従って、ブラックボックスな機械学習モデルを業務のアウトプットの直接の参考にして、例えば、医務査定のルールをリアルワールドの実績に基づいて改善することはなかなか難しいのではないかと考えられます。



では、この解釈性の課題に解決を与えるために、関数分解という概念を導入いたします。上の段です。こちら、目的変数 Y を X で表すモデルの関数が m(x) です。こちらは、一般に上段のとおり分解することができまして、まず定数項、そして主効果と呼ばれるところ、こちらは変数 x_1 のみで決まる関数 m_1 から、x_d のみで決まる m_d までです。こちら 1 変数だけで決まるので、メインの主効果と呼びます。

次は、x_1、x_2 二つ与えて初めて決まる $m_{1,2}$。こちら、2 次の交互作用と呼びます。さらに 3 次から d 次まで、このように分解できて、中心化という、全データポイントで平均して 0 になるというような制約をつければ、この分解は一意となっております。

以上のとおり申し上げたのですけれども、現実のデータだと、何十次元の交互作用のようなものは、あまりないと考えられます。実際のデータに触れた方だとお気づきかもしれないのですけれども、ある変数と関係している変数は、例えばデータセットで 100 変数あったとしても、ごく一部です。交互作用を持つものも、ごく一部です。もし高次の交互作用を無視できると仮定した場合、この場合は、例として 3 次以上が無視できるほど小さいと仮定しますと、その m(x) の関数は、このような主効果と 2 次の交互作用のみで書くことができます。

真のモデル m(x) を、この右辺のシンプルな形の機械学習モデルで表すことができれば、先ほどの業務アウトプットと同じぐらいシンプルだと言えます。また、モデルの構造をシンプルにするということは、ある意味、正則化と見なすことができまして、高次の交互作用を無理に学習しようとしないから、過学習やバリアンスの抑止も期待することができます。

1．ランダムフォレストを実務に活用する際の課題

（参考）他の機械学習モデルの交互作用次数

アクチュアリーによる採用実績の多いGLM、GAMは交互作用次数が低いが、木構造・ニューラルネットワークなどのいわゆる「ブラックボックスモデル」には、交互作用次数の高いものが多い。

- 線形モデル：$y = a + b_1x_1 + b_2x_2 + \cdots + b_dx_d$
- 加法モデル：$y = a + f_1(x_1) + f_2(x_2) + \cdots + f_d(x_d)$

BI(1)

- 二次加法モデル：$y = a + f_1(x_1) + f_2(x_2) + \cdots + f_d(x_d) + f_{1,2}(x_1, x_2) \ldots + f_{d-1,d}(x_{d-1}, x_d)$

BI(2)

- 勾配ブースティング(木の深さ:r)
- ランダムフォレスト(木の深さ:r)

BI(r)

- ニューラルネットワーク

BI(d)

8

参考までに、こちらの交互作用の次数という観点から他のモデルも分類いたしますと、アクチュアリーによく採用されている GLM や GAM については 1 次元、2 次の交互作用を入れても、せいぜい 2 次元なのに対して、ブラックボックスと呼ばれるような木構造モデルは、木の深さそのものが交互作用の次元となってしまい、深くすると当然複雑になります。あとニューラルネットに至っては、密な層があると全部の変数の交互作用が発生してしまい、かなり複雑なモデルと言えます。

以上より、アクチュアリー業務における活用可能性を高めるためにも、交互作用の次数を下げるといったアプローチは、有効ではないでしょうか。

1．ランダムフォレストを実務に活用する際の課題

ランダムフォレストの精度面における限界

RFは、バイアスとバリアンスのトレードオフ関係において、次のジレンマに直面する。

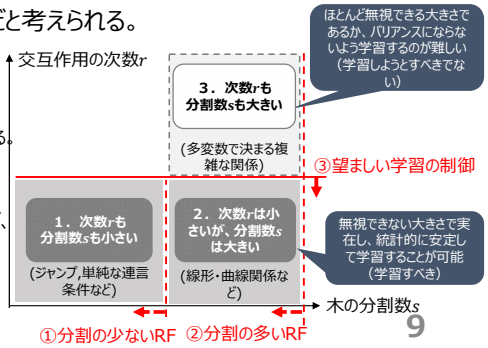
- 解釈性を確保するためにランダムフォレストで交互作用の次数を制限しようとすると、決定木の分割数を少なくする必要があるが、これでは**モデルが粗削りになりすぎて**しまい、バイアスが大きくなってしまう。
- 分割を多く許容すると、**複雑な高次交互作用まで学習しようとしてしまい**、バリアンスの原因となる。

⇒これはRFが、主に**木の分割数sで学習を制御**しているためだと考えられる。

このジレンマは右図の通り図解できる。

- ①正則化の強い(分割の少ない)RFは、右図における「2」を捉えられず、バイアスが残る。
- 他方で、②正則化の弱い(分割の多い)RFは、「2」を捉えることができても、「3」も捉えようとしてバリアンスが大きくなってしまう。

⇒③のとおり、**sではなく交互作用の次数rに上限を設けて学習を制御**すれば、「3」を捉えようとすることによるバリアンスを防止しつつ、「2」を捉えないことによるバイアスも防ぐことができる。



9

次に、「精度面における限界」についてご説明します。最初に申し上げたとおり、ランダムフォレストは、

決定木をバギングしているので、バイアスが残ってしまう。そして、それを解消しようとして木を深くするとバリアンスの原因となってしまう。こちらを詳細に表したものが、右下の図表になっています。こちらは、目的変数 Y と X の関係全体を表していて、それを横軸、関係を捉えるためにどれだけ木構造アルゴリズムにて分割が必要かといった切り口と、縦軸、交互作用の次数で分類しております。

まず、一番右下の 1 番、交互作用の次数が低くて少しの分割で捉えられるような関係としては、一変数によるジャンプがあります。ある閾値で、いきなり Y が少し上がっているというような、一番シンプルな関係が 1 番です。右下の 2 番目は、交互作用の次数は小さい、つまりほとんど変数が登場しないけれども、たくさん分割しないと捉えられないような関係です。例えば線形関係については、階段関数で近似するわけですが、このようにたくさん分割しないと精緻に近似できないものが 2 番です。そして 3 番が、交互作用の次数が高く、分割数もたくさん分割しないと捉えられないもので、こちらは、多変数で決まる高次の交互作用となっています。こちら、先ほど申し上げたとおり、現実にはほとんど存在しないし、学習しようとするとかなり過学習してしまうといったところから、学習しないことが望ましいと言えます。

先ほどのランダムフォレストの話に戻りますと、木の浅いランダムフォレストは、1 番の赤の点線の左側しか学習しないアルゴリズムで、2 番の線形関係を捉えられず、それがバイアスとして残ります。他方、木を深くしてしたものについては、2 番は捉えられるのですけれども、3 番の高次の交互作用も無理に学習しようとしてしまって、バリアンスが生じる原因となって、いずれにせよ精度が不満足なものになります。最善の方法としては、分割数ではなくて交互作用の次数に上限を設けることで、1 と 2 だけを学習して、3 番は学習しないといったことが可能になります。

これまで申し上げた解釈性や精度の課題の解決策を取り入れているものが、これからご紹介する Random Planted Forests になります。

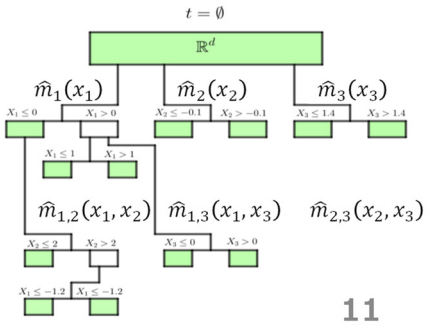
2．Random Planted Forestsの概要

Random Planted Forestsのモデル構造

- Radom Planted **Tree**では、関数分解の形で、**各変数の組tごとに、tによる交互作用に対応する木$\hat{m}_t(x)$**を構築する。
- 主に、ハイパーパラメータ：**交互作用の最大次数r**で学習を制御する。
- 各々の木は、変数の組tに含まれる変数にて、数多く分岐することができる。

○例：次元 $d = 3$、$r = 2$の場合

$$\hat{m}(x) = \hat{m}_1(x_1) + \hat{m}_2(x_2) + \hat{m}_3(x_3) + \hat{m}_{1,2}(x_1, x_2) + \hat{m}_{1,3}(x_1, x_3) + \hat{m}_{2,3}(x_2, x_3)$$



11

モデルの構造としましては、左下にあるとおり、予測値が主効果の項と 2 次の交互作用の項に分かれて表現されていて、右側を見てみますと、それぞれに対して木が作られているといったところが見て取れます。

2．Random Planted Forestsの概要

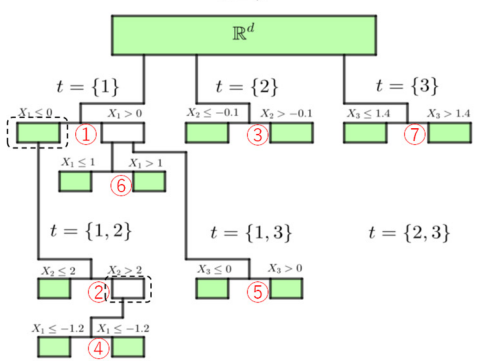
（ご参考）Random Planted Treeの分割方法

- 根：R^dから開始して葉を逐次分割・追加し、残差(次頁説明)が最も減少する、葉・変数・閾値で分割していく。
- 分割後の葉は、それらが属する組み合わせtの木に追加
- 分割には二種類のパターンがある。
  - 既存変数による再分割
    分割前の葉を規定する変数の組tのうち一つで再び分割する場合、分割前の葉は削除し(白)、分割後の葉で置き換える。
  - 新規変数による一つ高次の分割
    分割前の葉を規定する変数の組tと異なる変数で新たに分割する場合、分割前の葉は残し(黄)、分割後の葉を新規追加。

例②：$t = \{X_1\}$の木に属する葉「$X_1 \leqq 0$」を$X_2 = 2$で分割
⇒$t = \{X_1\}$に属さない新たな変数X_2による分割。上の葉は残し、新たな葉「$X_1 \leqq 0$かつ$X_2 \leqq 2$」「$X_1 \leqq 0$かつ$X_2 > 2$」を$t = \{X_1, X_2\}$に追加。

例④：$t = \{X_1, X_2\}$の木に属する葉「$X_1 \leqq 0$かつ$X_2 > 2$」を$X_2 = -1.2$で
⇒$t = \{X_1, X_2\}$にすでに属する変数X_2による分割。上の葉は削除し、新たな葉「$X_1 \leqq -1.2$かつ$X_2 \leqq 2$」「$-1.2 < X_1 \leqq 0$かつ$X_2 > 2$」を$t = \{X_1, X_2\}$に追加に追加。

○例：変数X_1, X_2, X_3、r = 2



12

では、これらをどのように分割していくのかというところですけれども、ほとんどランダムフォレストと同じになりまして、残差、つまりモデルの予測値と Y の差が一番減るような分割を選んでまいります。

ただ、1 点違うところがあって、例えば、この $x_1 \leq 0$①の左側にある葉っぱを $x_2 > 2$②の二つの葉っぱに分ける分割を見てまいりますと、分割前の葉っぱは x_1 だけで特徴づけられるので、t={1}、つまり x_1 のみの主効果の木に入っているのですけれども、分割後は x_1 と x_2 両方の変数で特徴づけられるわけですから、x_1、x_2 の交互作用の木に入るべきです。

従って、木を新設しまして、その木に二つの分割後の葉っぱを追加します。他方で、分割前の x_1 の主効果の木はそのまま残します。このように、交互作用や主効果ごとに木を管理しますので、ここはランダムフォレストと少し違ったところです。

2．Random Planted Forestsの概要

（ご参考）Random Planted Treeの予測値

変数の組tに対する木の葉lごとの値$m_{t,l}$があり、新たな入力xに対し、xが属するすべての葉の$m_{t,l}$の合計が予測値となる。

$$\hat{m}(\boldsymbol{x}) = \sum_{t \in T_r} \hat{m}_t(\boldsymbol{x}) \ , \hat{m}_t(\boldsymbol{x}) = \sum_{l=1}^{L_t} m_{t,l} \, \mathbf{1}\left(\boldsymbol{x} \in \mathcal{I}_{t,l}\right)$$

○$m_{t,l}$ は以下の通り分割ごとに、帰納的に計算される。

- 分割s回目で葉l: $m_{t,l}$ が変数kで、葉l',$l'+1$に分割されるとする。
- 分割$s-1$回目時点でY^iとモデル予測値の間に残る残差R_i^{s-1}を、それぞれの葉$\mathcal{I}^{l'}$,$\mathcal{I}^{l'+1}$の中で平均した値を$g_{\mathcal{I}^{l'}}, g_{\mathcal{I}^{l'+1}}$とおく。
- 分割後二つの葉l',$l'+1$の$m_{t',l'}, m_{t',l'+1}$は以下の通り設定する。
  - $k \in t$(既存変数による再分割)の場合は、$m_{t,l} + g_{\mathcal{I}^{l'}}, m_{t,l} + g_{\mathcal{I}^{l'+1}}$
  - $k \notin t$(新規変数による高次の分割)の場合は、$g_{\mathcal{I}^{l'}}, g_{\mathcal{I}^{l'+1}}$
- 最後に残差の更新を行う：$R_i^s \coloneqq R_i^{s-1} - \mathbf{1}\left(\boldsymbol{x} \in \mathcal{I}^{l'}\right) g_{\mathcal{I}^{l'}} - \mathbf{1}\left(\boldsymbol{x} \in \mathcal{I}^{l'+1}\right) g_{\mathcal{I}^{l'+1}}$

○葉l, t,変数kおよび、分割の閾値cは、$\sum_{i=1}^{n} (R_i^s)^2$が最小となるものが探索・選択される。

○分割sは、ハイパーパラメータ:n_splitsの回数で止める。

13

値の決め方としましても、基本、ランダムフォレストと同じように、葉っぱを分割することによる残差をどんどん更新していきます。ただ、木を新設する際は、葉っぱの値の更新分を新しい木の葉っぱの値として設定して、元の分割する前の葉っぱ、

2．Random Planted Forestsの概要

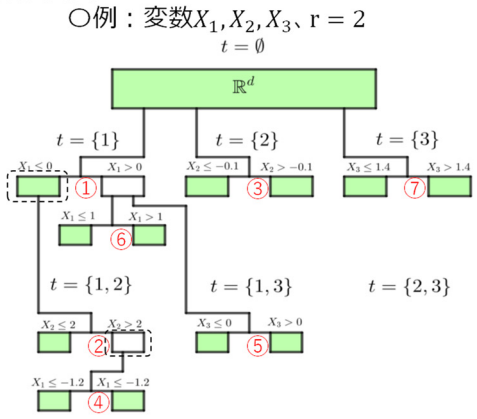
（ご参考）Random Planted Treeの分割方法

- 根：R^dから開始して葉を逐次分割・追加し、残差(次頁説明)が最も減少する、葉・変数・閾値で分割していく。
- 分割後の葉は、それらが属する組み合わせtの木に追加
- 分割には二種類のパターンがある。
  - 既存変数による再分割
    分割前の葉を規定する変数の組tのうち一つで再び分割する場合、分割前の葉は削除し(白)、分割後の葉で置き換える。
  - 新規変数による一つ高次の分割
    分割前の葉を規定する変数の組tと異なる変数で新たに分割する場合、分割前の葉は残し(黄)、分割後の葉を新規追加。

例②：$t = \{X_1\}$の木に属する葉「$X_1 \leqq 0$」を$X_2 = 2$で分割
⇒$t = \{X_1\}$に属さない新たな変数X_2による分割。上の葉は残し、新たな葉「$X_1 \leqq 0$かつ$X_2 \leqq 2$」「$X_1 \leqq 0$かつ$X_2 > 2$」を$t = \{X_1, X_2\}$に追加。

例④：$t = \{X_1, X_2\}$の木に属する葉「$X_1 \leqq 0$かつ$X_2 > 2$」を$X_2 = -1.2$で
⇒$t = \{X_1, X_2\}$にすでに属する変数X_2による分割。上の葉は削除し、新たな葉「$X_1 \leqq -1.2$かつ$X_2 \leqq 2$」「$-1.2 < X_1 \leqq 0$かつ$X_2 > 2$」を$t = \{X_1, X_2\}$に追加に追加。

○例：変数X_1, X_2, X_3、r = 2



12

先ほどで言うと $x_1 \leqq 0$ の葉っぱの値は、そのままにするといった点で少し違ってまいります。

説明し忘れたのですけれども、「交互作用の次元で制限をする」と申し上げましたけれども、分割数の上限

も設定しております。それは、各々の小さい木ごとではなくて、全体で何十回までという形で制限をしますので、例えば X_1 による強い線形関係がある場合は、この X_1 のこちらの木だけをどんどん成長させて、たくさん分割することができる。従って、そのような強い線形関係も緻密に捉えることが可能です。

2．Random Planted Forestsの概要

Random Planted **Forests**の構築

RF同様、学習データのブートストラップに加え、各分岐で用いる**変数の組*t***をランダムに絞り込むことで、独立なRandom Planted Treeを多数構築したうえで平均を取る。

〇関連するハイパーパラメータ

$ntrees \in \mathbb{N}$: Random Planted Treeを構築する本数

$t\_try \in (0,1]$: 分割を行う変数の組tの全ての候補のうち、t_tryの割合の候補に絞り込んで分割を探索する。

$split\_try \in \mathbb{N}$:変数kで分割を試みる閾値cの候補の数

14

これまでは、Random Planted Tree についてのご説明でした。それをバギングして Forests にする方法もほとんどランダムフォレストと同様でして、データのブートストラップに加えまして、分割する交互作用変数の組 t を一定割合、ランダムに絞り込むことによって、ランダム性を保つといった処理を行っています。

2．Random Planted Forestsの概要

（ご参考）Random Planted Forestsの収束性

$Y = m(X) + \varepsilon, E[\varepsilon] = 0, \varepsilon \perp X$とする。

- 特定の正則条件のもと、サンプル数$n \to \infty$でRandom Planted TreeおよびRPFの推定値$\hat{m}(X)$は、真の関数$m(X)$に収束する。
  - Treeは、次のオーダーでL^2収束：$O\left(h_n + \sqrt{\frac{1}{nh_n{}^r}}\right)$ （h_n:葉の幅）
  - Forestは、次のオーダーでL^1収束：$O\left(h_n{}^2 + \frac{h_n}{\sqrt{nh_n^{2r}}} + \sqrt{\frac{1}{nh_n{}^r}}\right)$
- 葉の幅が細かくなるスピードが十分に遅いと($nh_n^{2r} \to \infty$)Forestのほうが収束速度が速い。

15

では、最後に、収束性、理論的性質です。ランダムフォレストにつきましても、先ほど小田さんがご説明したとおり、一致性が知られております。RPF につきましては、Tree と Forest それぞれについて L^2、L^1 収束

が知られています。こちら、単なる一致性の確率収束よりも強い収束といったところで、割といい性質を持っています。

Agenda

1. ランダムフォレストを実務に活用する際の課題
2. Random Planted Forests(RPF)の概要
3. RPFの性能評価
4. RPFの実データ適用

16

次に、RPF の人工データを用いた性能評価を行っています。

3．RPFの性能評価

使用したデータパターン、比較対象のアルゴリズム

RPFの論文では、以下通り人工データを生成したうえで、RPFと他の機械学習アルゴリズムで構築したモデルの精度を比較している。

$$Y_i^s = m(X_i^s) + \varepsilon_i^s, \qquad i = 1, \ldots, 500, \qquad s = 1, \ldots, 100$$

○生成データのパターン

sparse：特定変数以外はノイズ変数

Description	Meaning
additive+sparse	$m(x) = m_1(x_1) + m_2(x_2)$
hierarchical interaction +sparse	$m(x) = m_1(x_1) + m_2(x_2) + m_3(x_3) + m_{1,2}(x_1, x_2) + m_{2,3}(x_2, x_3)$
pure interaction+sparse	$m(x) = m_{1,2}(x_1, x_2) + m_{2,3}(x_2, x_3)$
additive+dense	$m(x) = m_1(x_1) + \cdots + m_d(x_d)$
hierarchical interaction+dense	$m(x) = \sum_{k=1}^{d} m_k(x_k) + \sum_{k=1}^{d-1} m_{k,k+1}(x_k, x_{k+1})$
pure interaction+dense	$m(x) = \sum_{k=1}^{d-1} m_{k,k+1}(x_k, x_{k+1})$
smooth	$m_k(x_k) = (-1)^k 2\sin(\pi x_k)$ $m_{k,k+1}(x_k, x_{k+1}) = m_k(x_k x_{k+1})$
jump	$m_k(x_k) = \begin{cases} (-1)^k 2\sin(\pi x_k) - 2 & \text{for } x \geq 0, \\ (-1)^k 2\sin(\pi x_k) + 2 & \text{for } x < 0 \end{cases}$ $m_{k,k+1}(x_k, x_{k+1}) = m_k(x_k x_{k+1})$

dense：全ての変数が等しく寄与

真のモデルはいずれもBI(1)あるいはBI(2)以内に収まる

○比較対象のアルゴリズム

Description	short
A gradient boosting variant	xgboost
rpf	rpf
Random forest	rf
Smooth backfitting[1]	sbf
Generalized additive models[2]	gam
Bayesian additive regression trees	BART
multivariate additive regression splines	MARS
1 nearest neighbours	1-NN
Sample average	mean

○チューニングするハイパーパラメータの範囲

Method	Parameter range
xgboost	max.depth = 1 (if depth = 1), = 2,3,4, eta = 0.005, 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, nrounds = 100, 300, 600, 1000, 3000, 5000, 7000.
rpf	max.interaction = 1 (if max.interaction = 1), = 2 (if max.interaction=2), = 1, 2, 3, 4, 30, t.try = 0.25, 0.5, 0.75, nsplits = 10, 15, 20, 25, 30, 40, 50, 60, 80, 100, 120, 200 , split.try = 2, 5, 10, 20.
rf	m.try = $[d/4]$, $[d/2]$, $[3d/4]$, $[7d/8]$, $[d]$, min.node.size = 5, ntrees = 500, replace = TRUE,FALSE.
sbf	bandwidth = 0.1, 0.2, 0.3.
gam	select =TRUE, method = 'REM' (in sparse models), default settings (in dense models).
BART	power(β) = 1, 2, 3, ntree = 50,100,150,200,250,300, sparsity parameter = 0.6,0.75,0.9.
MARS	degree = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, penalty = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

17

こちらは、原論文の結果をそのまま紹介する形にはなってしまうのですけれども、サンプル数が 500 と、かなり少ないです。ノイズが正規ノイズで回帰問題という、とてもクラシカルな設定です。生成データのパターンとしては、Sparse と Dense があります。Sparse が、主要な一つか二つの変数だけが効いていて、他の説明変数は全部ノイズというパターン。Dense が、全ての説明変数が等しく寄与しているパターンです。その他は、ジャンプがあったりなかったり、交互作用があったりなかったり、というようなパターンで大量に作

っています。

比較対象のアルゴリズムは XGBoost（GBM）や GAM、ランダムフォレストなど、主要なものと比較しております。

3．RPFの性能評価

Sparseなパターンにおける精度比較

バリアンスが課題となるため、計算コストが膨大なBARTを除き、RFやGBMをしのぎ、RPFがトップクラスの精度となった。

なお、真のモデルではr = 2 であるにも関わらず、RPFをr = 2に制限すると、rを制限しないRPFに多少劣る結果になっている。これは、前者では二次の交互作用として抽出できなかったバイアスが、後者でより高次の交互作用として抽出されたからだと考えられる。

Description	Meaning
additive+sparse	$m(x) = m_1(x_1) + m_2(x_2)$
hierarchical interaction +sparse	$m(x) = m_1(x_1) + m_2(x_2) + m_3(x_3) + m_{1,2}(x_1, x_2) + m_{2,3}(x_2, x_3)$
pure interaction+sparse	$m(x) = m_{1,2}(x_1, x_2) + m_{2,3}(x_2, x_3)$
additive+dense	$m(x) = m_1(x_1) + \cdots + m_d(x_d)$
hierarchical interaction+dense	$m(x) = \sum_{k=1}^{d} m_k(x_k) + \sum_{k=1}^{d-1} m_{k,k+1}(x_k, x_{k+1})$
pure interaction+dense	$m(x) = \sum_{k=1}^{d-1} m_{k,k+1}(x_k, x_{k+1})$
smooth	$m_k(x_k) = (-1)^k 2\sin(\pi x_k)$ $m_{k,k+1}(x_k, x_{k+1}) = m_k(x_k x_{k+1})$
jump	$m_k(x_k) = \begin{cases} (-1)^k 2\sin(\pi x_k) - 2 & \text{for } x \geq 0, \\ (-1)^k 2\sin(\pi x_k) + 2 & \text{for } x < 0 \end{cases}$ $m_{k,k+1}(x_k, x_{k+1}) = m_k(x_k x_{k+1})$

○1・2次交互作用、ノイズ変数あり、ジャンプあり

Method	dim=4	dim=10	dim=30	BI($r = 0/1/2$
xgboost(depth=1)	2.974 (0.112)	3.046 (0.12)	3.098 (0.223)	1
xgboost	1.02 (0.152)	1.28 (0.16)	1.418 (0.156)	×
xgboost-CV	1.049 (0.125)	1.279 (0.157)	1.475 (0.185)	×
rpf (max_interaction=1)	2.941 (0.117)	2.942 (0.123)	2.913 (0.197)	1
rpf (max_interaction=2)	0.767 (0.096)	1.082 (0.139)	1.34 (0.132)	2
rpf	0.745 (0.089)	1.093 (0.142)	1.307 (0.113)	×
rpf-CV	0.769 (0.101)	1.167 (0.152)	1.404 (0.14)	×
rf	0.914 (0.091)	1.237 (0.121)	1.415 (0.152)	×
sbf	2.791 (0.098)	2.926 (0.12)	3.756 (0.284)	1
gam	2.782 (0.085)	2.728 (0.105)	2.793 (0.208)	1
BART	0.611 (0.078)	0.644 (0.106)	0.67 (0.094)	×
BART-CV	0.661 (0.111)	0.772 (0.173)	0.791 (0.133)	×
MARS	2.306 (0.17)	2.325 (0.145)	3.374 (2.716)	×
1-NN	4.559 (0.409)	8.883 (0.692)	13.434 (0.674)	×
average	8.721 (0.334)	8.449 (0.229)	8.638 (0.412)	0

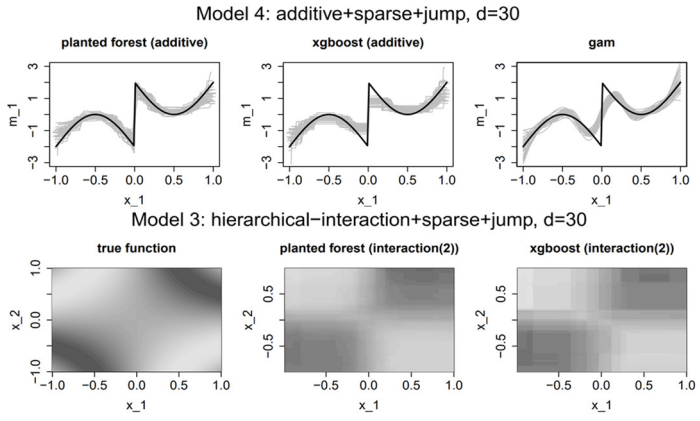
18

結果ですけれども、RPF は、この Sparse なパターンで、かなりいい精度を発揮しています。このような場合ではノイズがたくさん発生してしまうので、バリアンスをいかに抑えるかといったところが課題です。ですから、勾配ブースティングはモデルの残差をどんどん埋めていくようなアプローチなのですけれども、バイアスをどん欲に減らしていく方法なので、バリアンスの対処が結構苦手なのです。ということもあって、RPF が GBM をしのぎ、トップクラスの精度となっております。

3．RPFの性能評価

Sparseなパターンにおける学習モデル構造（解釈）の比較

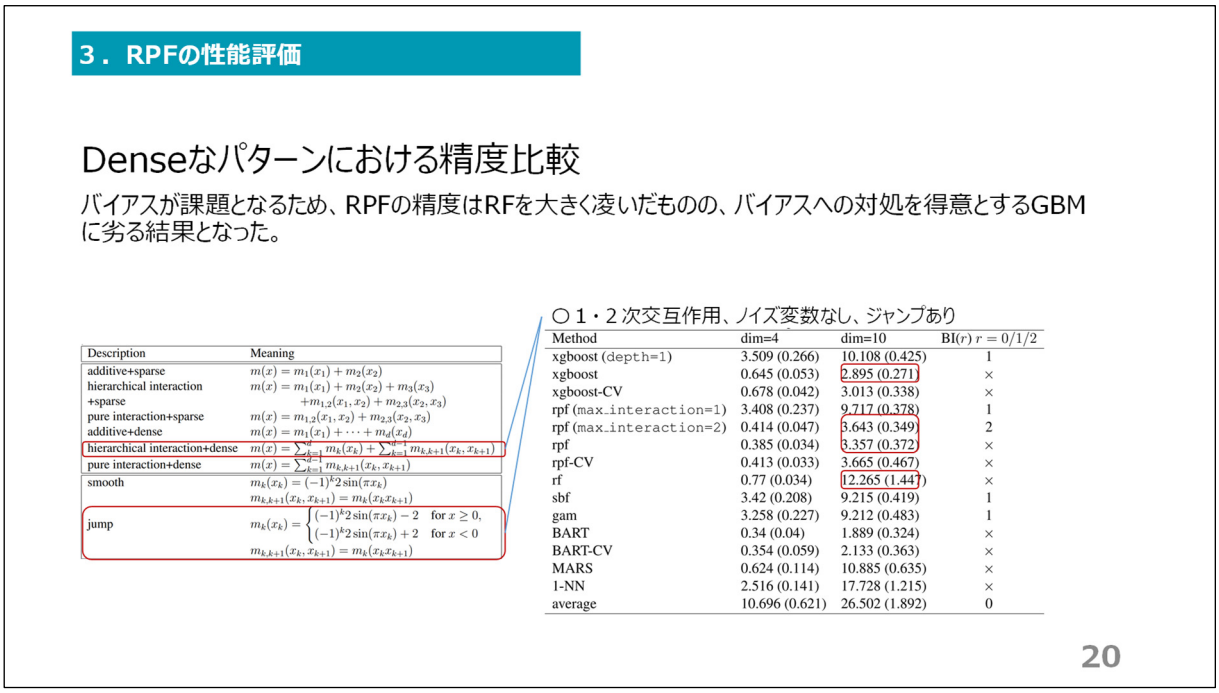
RPFは、GBMやGAMと比較して、主効果・交互作用を細かく把握できている。



19

精度だけではなくて、関数を近似できているかも見ます。上段が主効果です。黒い実線が真の主効果の関数で、この灰色の線が各モデルによる近似となっております。一番上段の左が RPF で、真ん中が GBM。GBM が x_1=0 でジャンプが取れておらず、右の GAM については、スプライン関数は連続なのでジャンプが取れていないです。GBM がうまくいっていない理由としては、都度、浅い決定木を作って残差から差し引くというような処理を何度も繰り返すのですけれども、x_1=0 付近のかなりローカルな関係を、浅い決定木で捉えることは結構難しかったようです。他方、RPF の左については x_1 の木を、葉っぱをどんどん細分化することはできるので、このようなローカルな関数の変化についても緻密に捉えることができたと考えられます。

下は 2 次元ですけれども、下段の左が真の関数、真ん中が RPF で、右が GBM です。やはり RPF の方が、若干誤差はあるのですが、滑らかに把握できていることが見て取れます。



Description	Meaning
additive+sparse	$m(x) = m_1(x_1) + m_2(x_2)$
hierarchical interaction +sparse	$m(x) = m_1(x_1) + m_2(x_2) + m_3(x_3) + m_{1,2}(x_1, x_2) + m_{2,3}(x_2, x_3)$
pure interaction+sparse	$m(x) = m_{1,2}(x_1, x_2) + m_{2,3}(x_2, x_3)$
additive+dense	$m(x) = m_1(x_1) + \cdots + m_d(x_d)$
hierarchical interaction+dense	$m(x) = \sum_{k=1}^{d} m_k(x_k) + \sum_{k=1}^{d-1} m_{k,k+1}(x_k, x_{k+1})$
pure interaction+dense	$m(x) = \sum_{k=1}^{d-1} m_{k,k+1}(x_k, x_{k+1})$
smooth	$m_k(x_k) = (-1)^k 2\sin(\pi x_k)$ $m_{k,k+1}(x_k, x_{k+1}) = m_k(x_k x_{k+1})$
jump	$m_k(x_k) = \begin{cases} (-1)^k 2\sin(\pi x_k) - 2 & \text{for } x \geq 0, \\ (-1)^k 2\sin(\pi x_k) + 2 & \text{for } x < 0 \end{cases}$ $m_{k,k+1}(x_k, x_{k+1}) = m_k(x_k x_{k+1})$

Method	dim=4	dim=10	BI(r) $r = 0/1/2$
xgboost (depth=1)	3.509 (0.266)	10.108 (0.425)	1
xgboost	0.645 (0.053)	2.895 (0.271)	×
xgboost-CV	0.678 (0.042)	3.013 (0.338)	×
rpf (max_interaction=1)	3.408 (0.237)	9.717 (0.378)	1
rpf (max_interaction=2)	0.414 (0.047)	3.643 (0.349)	2
rpf	0.385 (0.034)	3.357 (0.372)	×
rpf-CV	0.413 (0.033)	3.665 (0.467)	×
rf	0.77 (0.034)	12.265 (1.447)	×
sbf	3.42 (0.208)	9.215 (0.419)	1
gam	3.258 (0.227)	9.212 (0.483)	1
BART	0.34 (0.04)	1.889 (0.324)	×
BART-CV	0.354 (0.059)	2.133 (0.363)	×
MARS	0.624 (0.114)	10.885 (0.635)	×
1-NN	2.516 (0.141)	17.728 (1.215)	×
average	10.696 (0.621)	26.502 (1.892)	0

Dense のパターンだと、こちらは、全部の変数の木を取らないといけないので、取りこぼしによるバイアスが課題になるわけです。そこでは、GBM が一番良かったのですけれども、RPF もランダムフォレストと比べれば随分精度がましでした。ランダムフォレストは、EDA のようなベンチマークで使われることが多いと思うのですけれども、バイアスがとても厄介な場合だとなかなか信頼できないことが改めて確認でき、RPF がその用途に適しているのではないかと考えられます。

Agenda

1. ランダムフォレストを実務に活用する際の課題
2. Random Planted Forests(RPF)の概要
3. RPFの性能評価
4. RPFの実データ適用

21

最後、実データに適用いたします。

4．RPFの実データ適用

実データを用いたRPFの評価の目的

- RPFの論文では**交互作用が2次まで**($\mathbf{r = 2}$)の人工データでRPFの性能評価がされているが、アクチュアリーが実務で扱うデータには、より高次の交互作用も含まれると考えられる。
- 当発表では**より高次の交互作用も含む**と考えられる実データを用いて、以下を確認・試行する。
  - RPFと他アルゴリズムの精度比較
  - RPFの交互作用次数r = 2として、精度面で問題はないか
  - RPFによる主効果・交互作用への関数分解

22

人工データでは交互作用が 2 次までとなっておりましたが、実データでは 3 次や 4 次など、少し高次の交互作用も若干あるだろうと、極端に高い次元のものはないにしろ、そのような可能性があります。ですから、実データを用いて、それでも RPF の精度に遜色はないか、また RPF モデルにおける交互作用の最大次数を 2 に制限して精度面で問題がないか、そして、主効果と交互作用を正しく捉えられるか、といったところを確認いたしました。

4．RPFの実データ適用

実データによる数値実験の概要

- 使用データ
  Kaggle:自動車保険の支払請求データ
  - https://www.kaggle.com/datasets/floser/french-motor-claims-datasets-fremtpl2freq
  - エクスポージャー期間ごとの件数データだが、現時点ではパッケージがポアソン分布・オフセット等に対応していないため、エクスポージャ＝1年間（途中契約・解約なし）に母集団を制限し、請求有無（0件 or 1件以上）の分類問題とした。
  - 計算負荷等の理由から、カーディナリティの高い列を削除し、サンプル数も1/10に絞っている。
  - 学習とテストデータに、8:2の割合で分割した。
- パッケージ
  randomPlantedForest(https://github.com/PlantedML/randomPlantedForest)
- データセットの概要

説明変数　　目的変数　レコード・正解ラベルの数

Area	VehPower	VehAge	DrivAge	BonusMalus	VehGas	Density	ClaimNb
<fct>	<int>	<dbl>	<dbl>	<int>	<fct>	<int>	<dbl>
C	7	16	31	72	Regular	165	0
A	8	15	51	50	Diesel	32	0
A	4	7	34	64	Regular	14	0

ClaimNb	n
<dbl>	<int>
0	16098
1	836

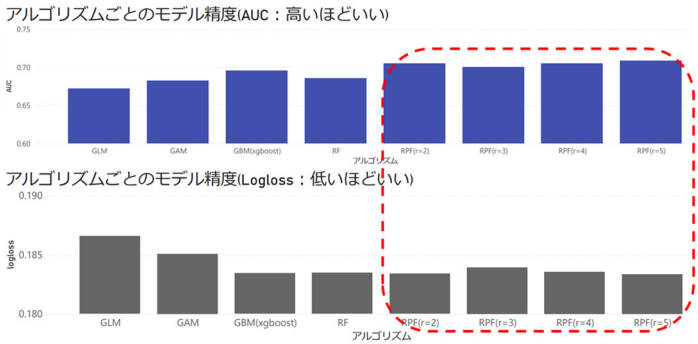
23

データとしては、Kaggle の automobile の自動車保険の支払請求データを使いました。こちらは、RPF のライブラリが開発途上で制約がありまして、カウンティングデータが捉えられないため分類データにして、計算負荷の事情から行や列を一部削除するというようなことを行っておりますので、データセットとしては 7 変数で、レコードが全部で 17,000 件、うち正解ラベルが 800 件ちょっとといった、結構小さいデータとなりました。

4．RPFの実データ適用

予測精度評価

- RPF(r = 2)の精度は、GLMやGAMを上回り、RFやGBMもわずかに上回る結果となった。
- RPFでは、r ≧ 3と次元を高くしても、r = 2と比べて大きな精度の改善は見られなかった。



24

このようにデータが限られるといったこともありまして、RPF の精度、この赤点線枠で囲った一番左が交互作用の次数が 2 の RPF ですけれども、上の AUC はトップクラスで、Logloss（対数損失）が一番低いといった感じで精度が出せております。

赤点線枠で囲ったところが、次元 2、3、4、5 と、右に行くにつれて変わってくるのですけれども、精度の改善はほとんどなかったといったところで、このセッティングだと、交互作用の次元を 2 にしても、ほとんど学習上問題なかったと言えます。3 次、4 次の交互作用もあるにはあるのでしょうけれども、学習できたものもあれば、過学習などでうまく取れなかったものもあると考えられます。

4．RPFの実データ適用

RPFによる関数分解

パッケージ：randomPlantedForestでは、各予測対象$\boldsymbol{x^i}$に対し、関数分解の全ての要素$\hat{m}_t(\boldsymbol{x^i})$の値を出力できる。

r = 2の場合、全サンプルの値$\{\hat{m}_t(\boldsymbol{x^i})|i = 1\sim n\}$を$x_t^i$に対してプロットすることで、各$t$の交互作用を可視化することができ、可視化された曲線およびヒートマップの和として予測値が直感的に表現できる。

$$Y|_{X=\boldsymbol{x}}\sim Be(logit(\hat{m}(\boldsymbol{x})))$$

$$\hat{m}(\boldsymbol{x}) = \hat{m}_0 + \hat{m}_1(x_1) + \hat{m}_2(x_2) + \cdots + \hat{m}_{1,2}(x_1, x_2) + \cdots$$

○出力イメージ

sample_id	BonusMal	DrivAge	…	y_true	y_pred	intercept	effect_BonusMalus	effect_DrivAge	…	effect_BonusMalus:DrivAge	…
1	50	71		0	0.030187	-3.08156	-0.261340706	0.145320256		-0.071304208	
2	80	23		0	0.073768	-3.08156	1.18543606	-0.361946059		-0.355059096	
3	50	70		0	0.038012	-3.08156	-0.261340706	0.145320256		-0.071304208	
4	50	57		0	0.036673	-3.08156	-0.261340706	0.144177544		-0.070161496	
5	80	66		0	0.124212	-3.08156	1.18543606	0.145320256		0.088474347	

25

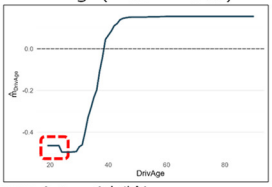
最後に、主効果や交互作用を可視化いたしました。こちらのライブラリでは、新しいインプット x に対して、予測値を各主効果や交互作用の成分に分けて出力することができます。ですから、全てのテストサンプルに対して、このような値を出していって、m_1 や m_2 の主効果、交互作用ごとに可視化しますと、曲線やヒートマップとして表現できます。

4．RPFの実データ適用

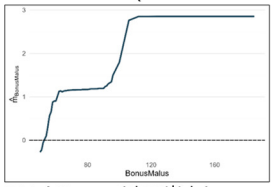
RPFによる関数分解（主効果）

- 推定された主効果は、いずれも納得性の高い解釈が可能である。
- とくに、DrivAgeの主効果では、RPFが年少者における僅かなリスク上昇を捉えることにも成功している。

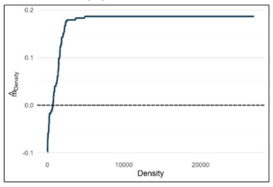
○DrivAge(運転手の年齢)



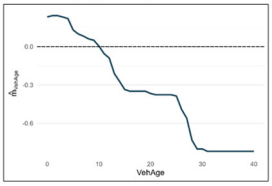
○Bonusmalus(等級割引／割増)



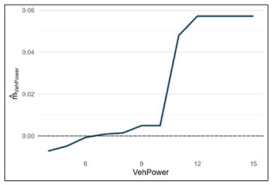
○Density(運転手が住む街の人口密度)



○VehAge(車齢)

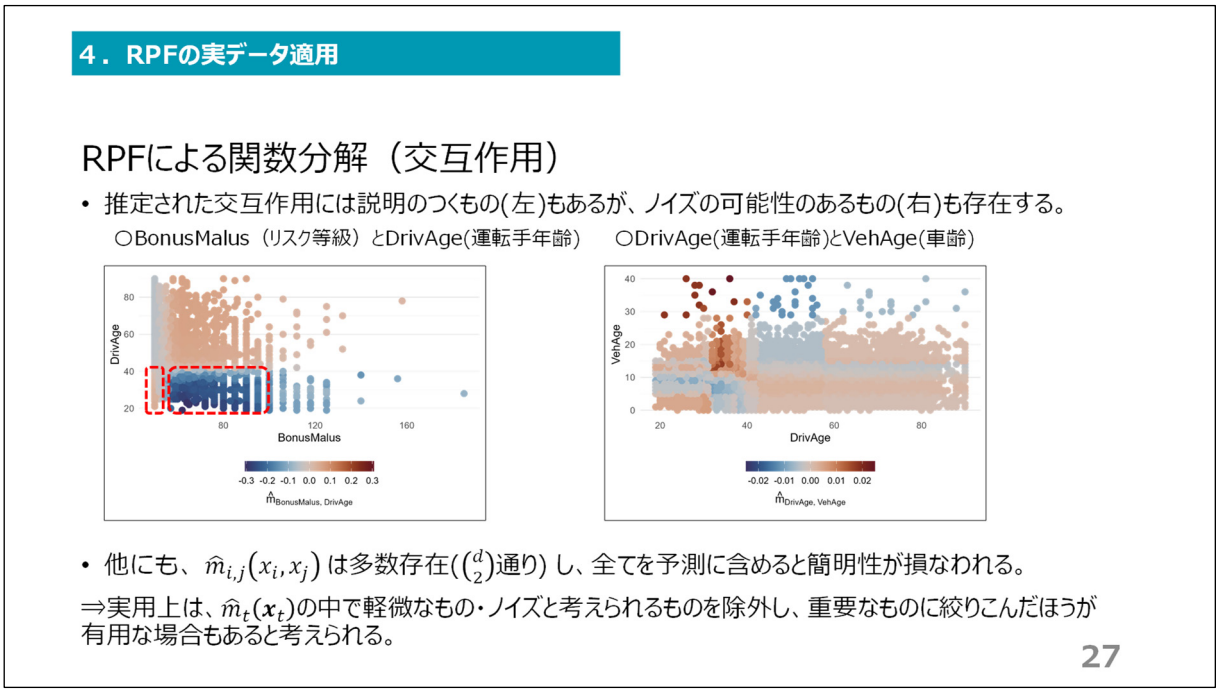


○VehPower(車の出力)



26

まず主効果は、1 変数による影響を見ますと、いずれも納得性の高い解釈が可能です。運転手の年齢や事故に基づく割引／割増、そして、人口密度、車の出力なども高ければ高いほど支払率が高いでしょうと。これは直感的か分からないですけれども、古い車を使う人は、新しいものに目移りせずに保守的な人が多くて、運転も結構安全な運転が多いといったところで、支払率が低くなっています。また、運転手の年齢がかなり若いとリスクが高いといったところが知られていて、若干捉えられてはいるのですけれども、800 件の中に若年のサンプルがあまりなかったのかもしれません。



最後に、交互作用です。こちらは、左側がとても強い交互作用で、リスク等級／リスク割増と運転手の年齢の交互作用で、左側の赤い点々では、若年層でかなり割引が高い人、こちらペーパードライバーですけれども、そのような若年層のペーパードライバーは特にリスクが高いということが、このプラスの交互作用で見て取れます。逆にリスク等級が少し上がると、少し事故をしていても、運転経験があるといったところで、思ったほどリスクが高くならない。

右側は、交互作用としては小さくて、プラマイが凸凹しているところが見られまして、こちらはノイズである可能性もあります。交互作用は、2 次元でも、かなりたくさんありまして、ほとんどが右側のノイズのようなものになっています。これらは、解釈の目的からは、できるだけ除去できればいいと考えられます。

4．RPFの実データ適用

（ご参考）今後の課題：交互作用$\hat{m}_t(\boldsymbol{x}_t)$の絞り込み

- 統計的アプローチ：H統計量
  帰無仮説：「x_iとx_jの間に交互作用はない」が棄却される変数の組tのみを採用する。
  ⇒統計的に有意とは言えずとも、yの予測に対する影響の大きな交互作用を取りこぼしてしまう。
- 機械学習的アプローチ：Lasso回帰
  学習とバリデーション、テストに分け、①学習で*RPF*を用いて$\hat{m}_t(\boldsymbol{x})$を構築、②バリデーションにて$\hat{m}_t(\boldsymbol{x})$を特徴量として、$y$を$L1$正則化とともに線形回帰、③テストにて②のモデルを精度評価する。
  $L1$正則化により、yの予測に寄与しない$\hat{m}_t(\boldsymbol{x})$の係数は0となる。

$$\hat{m}(\boldsymbol{x}) = a + b_1 \cdot \hat{m}_1(x_1) + \cancel{b_2 \cdot \hat{m}_2(x_2)} + b_3 \cdot \hat{m}_3(x_3) + \cancel{b_{1,2}\, \hat{m}_{1,2}(x_1, x_2)} + \cdots$$

係数0　係数0

⇒ 影響の小さい交互作用を取りこぼすうえ、係数＝0となる意味の解釈が難しい。

- RPF学習におけるtの制限
  あるいは、特定の$\hat{m}_t(\boldsymbol{x})$を除外するだけではバイアスの原因となるため、あらかじめ変数の組tで分割しない(交互作用を学習しない)ように制限をする方法も考えられる。（未実装）

28

ただ、どうやって絞り込んで除去するかといったところは今後の課題でして、統計的な交互作用の検定、あるいは、交互作用の値を使って y を機械学習的に予測して、Lasso 回帰（正則化）を行って、要らないものを落とすというアプローチも考えられますけれども、いずれにせよ、そのまま取ってしまうとバイアスの原因になってしまうので、あらかじめ要らない交互作用は木を作らないように学習を制限したいところですが、こちらは実装されておりません。今後のわれわれの課題になると考えております。

サマリー（再掲）

- ランダムフォレスト(RF)を実務活用するにあたり、解釈性・精度の両面で課題が残っている。
- RFに両面から改善を施したRandom Planted Forest(RPF)を紹介。
  低い次数の**交互作用の組み合わせごとに木を構築**し、それぞれ深い分割を許容することで、
  - モデルの予測値を、少数の変数で決まる**シンプルな関数の和として表現**できる。
  - RFが捉えることが難しい**線形関係なども、小さいバイアスで捉えられる**。
- 人工データによる精度評価で他アルゴリズムと比較。
  - **スパース（高バリアンス）な設定で高い精度**を発揮した。
  - 密（高バイアス）な設定では勾配ブースティングに及ばず。
- 自動車保険の実データでも精度評価を実施。
  - RPFはGBMに並ぶ高い精度となった。
  - 交互作用の次数を上げても精度の改善は見られず。
  - 推定された主効果はいずれも納得性の高い解釈が可能だが、**交互作用には軽微なもの、ノイズと思われるもの**もあり、実用上はある程度**絞り込む必要がある**と考えられる。

29

サマリーですけれども、時間がないので省略させていただきます。

ご清聴、ありがとうございました。

30

では、ご清聴どうもありがとうございました。

【藤田】　松江さん、ありがとうございました。

それでは、続きまして、田中さんからのプレゼンテーションに移りたいと思います。

生存時間分析と
ランダム・サバイバル・フォレストの
ご紹介

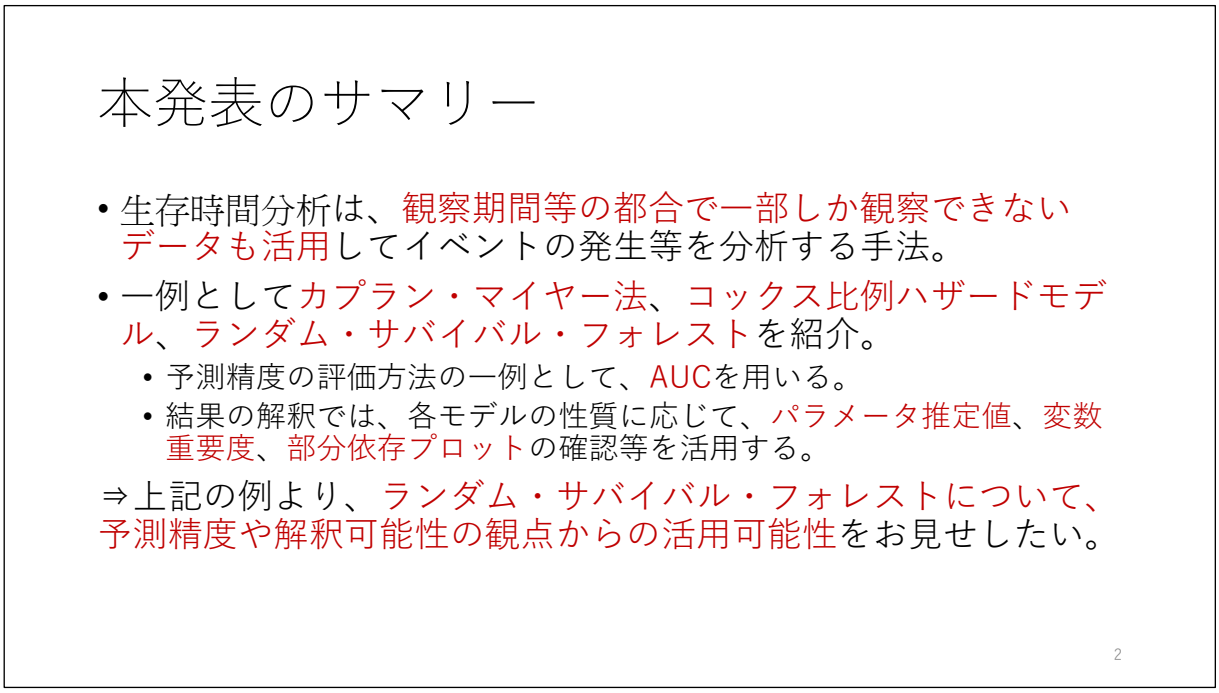
田中　豊人

1

【田中】　ご紹介にあずかりました田中と申します。

「生存時間分析とランダム・サバイバル・フォレストのご紹介」という題名で、発表させていただきます。正直に申し上げると、これからご紹介する内容には、学術的な新規性は特にございませんが、この分野を知

っていただく契機として、「生存時間分析とは何か」、そして「ランダムフォレストの生存時間分析への応用例の一つである、ランダム・サバイバル・フォレストとは何か」の 2 点をお伝えできればと考えておりますので、よろしくお願いします。



最初に、発表内容をお伝えいたします。一つ目は「生存時間分析とは何か」です。生存時間分析とは、観測期間等の都合で一部しか観測できないデータも、どうにか活用することで、イベント、例えば死亡、もしくは何らかの機械の故障などの発生率を分析する手法でございます。二つ目は「ランダムフォレストの生存時間分析への応用」として、まず同分野における典型的な手法である、カプラン・マイヤー法、コックス比例ハザードモデルをご紹介した上で、ランダムフォレストの応用形であるランダム・サバイバル・フォレストをご紹介します。

ここで、一般的にモデルの評価では、予測精度と結果の解釈という 2 つの軸があり、生存時間分析での予測精度の評価指標としては「AUC」を用います。AUC の概要は後ほどご説明します。結果の解釈は、モデルによって方法も変わってまいりますが、モデルのパラメータの推定結果や、変数重要度や部分依存プロット等を確認します。これらをキーワードに、後ほど具体的なご説明をいたします。

今回の主題はランダムフォレストでございますが、本パートでは、冒頭お伝えした通り「生存時間分析とは、そもそも何なのか」、「ランダムフォレストを生存時間解析にどのように活用するのか」の 2 点をお伝えすることで、少しでもご興味を持っていただき、「実務で使ってみたいな」と思って頂くきっかけとなりましたら何よりです。

1. 生存時間分析の概要

- 生存時間分析とは
- 手法例の紹介：　カプラン・マイヤー法
  コックス比例ハザードモデル
  ランダム・サバイバル・フォレスト

3

生存時間分析とは

- 時間経過に伴う特定のイベント（病気、死亡、機械の故障等）の発生を分析する手法で、以下の特徴がある。
  - 打ち切り：観察期間中に特定のイベントが発生しなかった場合等に、当該データ（以降、個体）を「打ち切り」として扱う（次頁詳述）。
  - 生存関数：時間経過とイベントの未発生確率の関係を表す関数。
  - ハザード関数：特定の時点でイベントが発生するリスクを示す関数。
- 本資料では、以下の３つの手法を対象に、手法の概要を説明し、Rを用いた簡単な実装例を示す。
  - カプラン・マイヤー法（KM）
  - コックス比例ハザードモデル（CoxPH）
  - ランダム・サバイバル・フォレスト（RSF）

4

まずは生存時間分析の概要ですが、先ほど申し上げたとおり、時間経過とともに発生する死亡や機械の故障などの発生率を評価するための手法でして、三つのキーワードがあります。

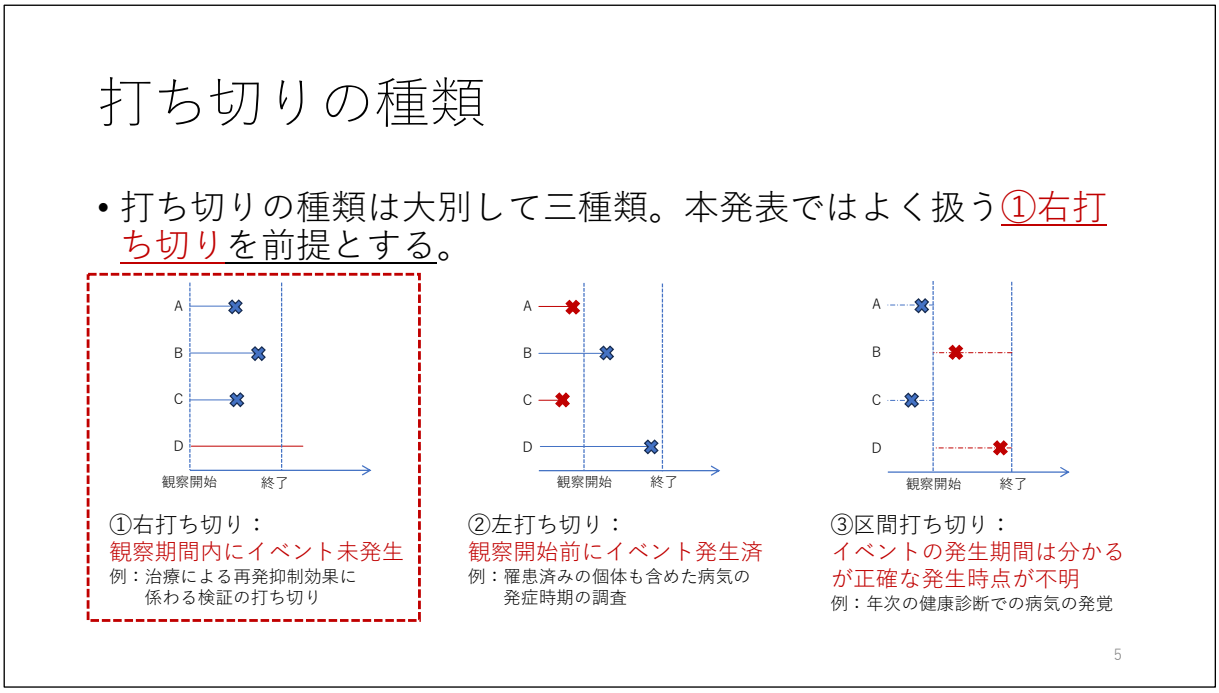
一つ目が、「打ち切り」と呼ばれているものです。当然、観察は有限期間でしかできないので、着目するイベント、例えば病気の再発があったとしても、全期間にわたってイベントを観測しきることは困難です。このような特性を「打ち切り」と呼びますが、このように観測が打ち切られたデータもどうにかしてイベント発生率の推計に活用する、ということが生存時間分析の一つのポイントになります。

２つ目が、生存時間分析で、結局、何を評価したいか。それは、生存関数と呼ばれるものです。生存時間分析は、横軸が経過時間、縦軸がイベントの未発生率として表現されるイメージです。当然、時間の経過とと

もにイベントの未発生率が下がると思うのですが、その経過時間と未発生率の関係を評価することが目的になります。

3 つ目はハザード関数で、これは生存関数の対になるような概念で、各時点の瞬間的なイベント発生率です。なぜ対と表現したかというと、ハザード関数から生存関数を求められますし、生存関数からハザード関数を求められるためです。

これら三点が、生存時間分析を理解する上でのキーワードになります。これらを踏まえて、こちらに記載の三つの手法をご紹介する予定です。



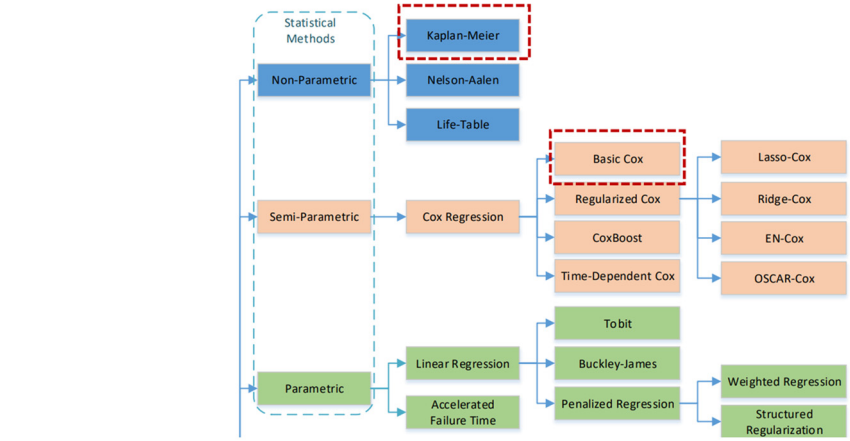
「打ち切り」に関してもう少し詳細を補足します。「打ち切り」は、大きく分けて三種類ございます。一つ目が「右打ち切り」です。「右打ち切り」は、生存時間分析でよく使われており、今回も、「右打ち切り」を前提として各手法をご説明する予定です。例えば、治療による再発抑制効果の有無を確認したい場合を考えます。実際には、観測を開始してから、全期間において、再発有無を確認することは困難で、観察は途中で終了してしまいます。これがまさに時間的な右側の打ち切りであり、「右打ち切り」と呼びます。

「左打ち切り」は、観察開始前にイベントが発生済みのデータが含まれる、例えば、観察開始時に既に罹患している個体のデータも活用して発生確率を推計したい場合です。

最後に「区間打ち切り」です。これは、イベントの発生期間が分かるが正確な発生時点が分からないデータを指します。健康診断をイメージしていただくと分かりやすいです。直前の健康診断で「病気と診断されました」といわれても、「正確に、いつ病気になったのか」は分からず、言えることは、前回と直近の健康診断の間のどこかで病気になった可能性が高い、ということのみです。このようなデータを「区間打ち切り」と呼びます。

このように「打ち切り」は 3 種類あるのですが、先ほど申し上げたとおり、今回は「右打ち切り」を題材にいたします。

生存時間分析手法の全体像 1/2
～統計的モデル～

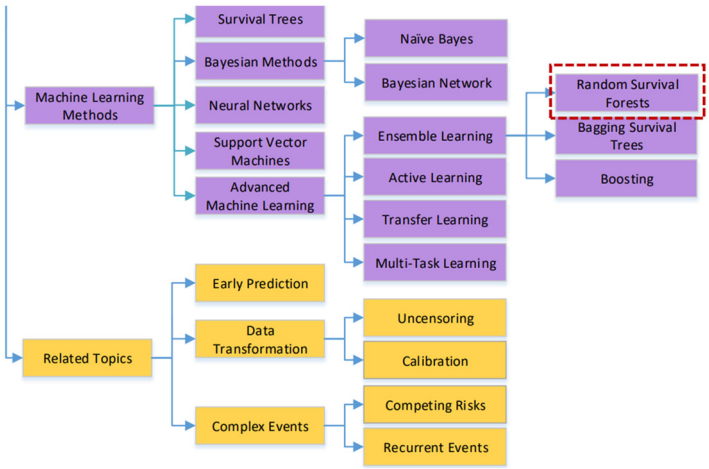


出典：Wang P, Li Y, Reddy CK. 2019. Machine learning for survival analysis: a survey. ACM Computing Surveys 51(6):1-36.

6

生存時間分析の手法は、先ほど触れた 3 つの手法だけでなく、実は沢山ありまして、統計モデルでは、ノンパラメトリック、パラメトリック、セミパラメトリックに分類される各手法があります。

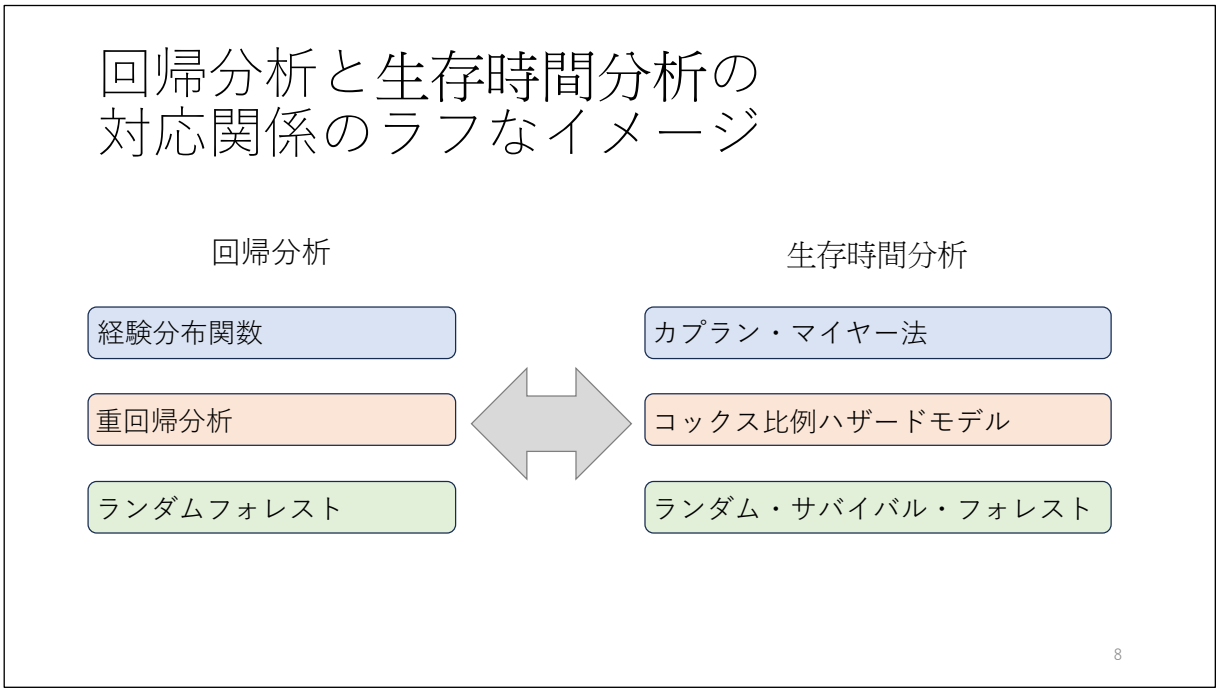
生存時間分析手法の全体像 2/2
～機械学習モデル～



出典：Wang P, Li Y, Reddy CK. 2019. Machine learning for survival analysis: a survey. ACM Computing Surveys 51(6):1-36.

7

機械学習の手法にも、様々なものがあります。一つひとつのモデルをご紹介はしませんが、「今回はほんの 3 つの手法をご紹介するが、実際には様々な手法がある」という点を、ご承知いただければと思います。



生存時間分析の手法のイメージを鮮明化いただくため、回帰分析と生存時間分析における各手法の関係性をご説明します。こちらは大変ラフなイメージでございます。なぜ「ラフ」かというと、この説明は、数学的な根拠に基づいているわけではないためです。あくまで、皆さんの慣れ親しんでいる回帰分析との関係を概略的にお伝えしたいという趣旨で、この対応関係をお見せしております。

回帰分析の経験分布関数はカプラン･マイヤー法に対応して､重回帰分析はコックス比例ハザードモデル、ランダムフォレストはランダム・サバイバル・フォレストに対応します。例えば、回帰分析は、各個人の較差を反映したいとしたら、回帰係数を導入して対応できますが、同じくコックス比例ハザードモデルでも各個人の較差を、回帰係数により反映できます。そのような手法間の類似性に基づき、対応関係をお見せしております。

カプラン・マイヤー法 1/2
～手法の概要～

- 打ち切りの個体（観察期間中にイベントが発生しなかった個体）も含めて生存関数を推定するノンパラメトリックな手法。
- 本手法では、生存関数$S(t)$（時刻tまでイベントが未発生の確率）の推定値$\hat{S}(t)$を以下の通り推定する。

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$$

  - t_i：イベントが発生した時点
  - d_i：t_iでイベントが発生した個体数
  - n_i：時点t_iの直前までイベントが未発生の個体数

カプラン・マイヤー法の原論文：Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457-481.

9

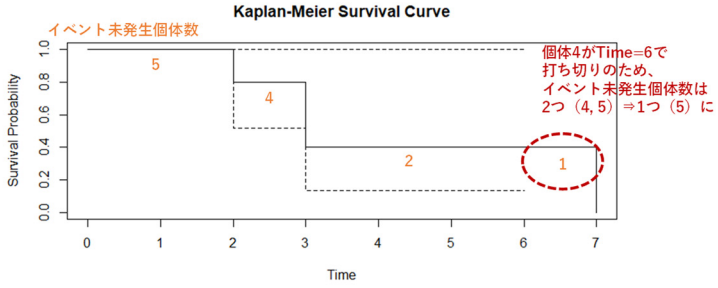
では、それぞれのモデルの概要をお伝えしたいと思います。まずカプラン・マイヤー法です。これは「打ち切りの個体も含めた生存関数を推定するノンパラメトリックな手法」ですが、例をお伝えしたほうが早いと思いますので、次頁をご覧ください。

カプラン・マイヤー法 2/2
～簡易データによる推定イメージ～

- カプラン・マイヤー法による生存関数の推定イメージは以下の通り（R package: 'survival'を使用）。 青：死亡　赤：打ち切り
  - データ：

個体	1	2	3	4	5
イベント発生時点	2	3	3	6	7

  - 生存関数の推定結果：



注：点線は95%信頼区間

10

ここでは、個体が 5 人います。「イベントの発生時点」は死亡した時点だと思ってください。そうすると、一番左側の個体 1 は、時刻 2 で死亡しました。個体 2 と 3 は時刻 3 で同時に死亡しました。そのような形式で並んでいて、個体 4 の人は、打ち切りで、時刻 6 までは生存していたものの、それ以降は死亡したかどうか分かりません。そして個体 5 は時刻 7 で死亡したと仮定します。

冒頭に申し上げましたが、生存時間分析の目的は生存関数の評価、つまり横軸を経過時間で、縦軸を生存

確率とする関数を評価することです。こちらの例では、観察開始時から時刻 2 にかけて、5 人生存していたと仮定しております。つまり、生存確率 100%です。時刻 2 で 1 人亡くなってしまうので、生存確率は 5 分の 4 で、80%になります。そこから、時刻 2 以降でリスクにさらされている 4 人のうち、時刻 3 で更に 2 人亡くなるので、時刻 2 での生存を前提とした条件付き生存確率として、4 分の 2 で 50%が生存確率になります。これに時刻 2 までの生存確率 80%を乗じて、80%×50%で 40%が時刻 3 以降の生存確率となります。

そこから、2 人だけ残るのですが、時刻 6 において、一人離脱するので、観測対象となる個体数は、2 から 1 に減ります。ただ、この離脱によって生存関数は変わらないまま進み、時刻 7 で 1 人亡くなる、つまり一人残った最後の個体が亡くなるので、それは時刻 7 時点の条件付き生存確率が 1 分の 0 で、0 となることを意味し、結果として生存関数上、時刻 7 以降の生存確率は 0 になります。このように、ノンパラメトリックに推定する手法が、カプラン・マイヤー法です。

コックス比例ハザードモデル 1/6 ～手法の概要～

- 複数の説明変数が生存時間に与える影響を評価するセミパラメトリックモデル。本モデルではハザード関数$h(t|X)$を以下の通り推定し、生存関数$S(t)$を推定する。
  - $h(t|X) = h_0(t)\exp(\beta_1 X_1 + \cdots + \beta_p X_p) = h_0(t)\exp(\beta \cdot X)$
    - $h(t|X)$：時点tにおける説明変数Xでの条件付きハザード関数
    - $h_0(t)$：ベースラインハザード関数（全てのXが0の時のハザード率）
    - $\beta_1, \ldots, \beta_p$：各共変量$X_1, \ldots, X_p$の回帰係数
  - $h_0(t)$は$\exp(\cdot)$の部分（ハザード比）とは異なり、具体的な形状を指定せず、ノンパラメトリックに扱う。

  ⇒このため、本モデルはセミパラメトリックと呼ばれる。

コックス比例ハザードモデルの原論文：Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187-220.

11

次に、コックス比例ハザードモデルです。ポイントは、本手法はセミパラメトリックモデルであり、また評価対象は、生存関数ではなく、ハザード関数である点です。先ほど「生存関数とハザード関数は対応関係ある」と申し上げましたが、本手法では、生存関数にかわり、ハザード関数、つまり瞬間的な死亡率を評価した上で、生存確率を評価するアプローチになります。スライド上の h_0 が、ベースとなるハザード関数です。つまり、回帰の説明変数が全部 0 のときのハザード関数に対応します。

本手法名に比例ハザードという表記が含まれますが、これはまさに、エクスポネンシャルの中に回帰係数を導入し、比例的な構造によってリスク較差を表現しているためです。これがコックス比例ハザードモデルという手法名の由来でございます。

例えば体重が増えるほどリスクが増えると仮定する場合、X_1 が体重だとしたら β_1 はプラスのパラメータと仮定して、リスク較差を表現できます。なお、h_0 はベースとなるハザード関数ですが、これは、ノンパラメトリックに推計します。一方、比例ハザードの部分はパラメトリックですので、パラメトリックとノンパラメトリックが混ざっており、本手法はセミパラメトリックな手法と呼ばれます。

コックス比例ハザードモデル 2/6
～（ご参考）ハザード関数と生存関数の関係～

- ハザード関数$h(t)$は、以下の通り、時点tにおいてイベントが発生する瞬間的な発生確率を表す。
  - $h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$
    - T：イベントが発生する時点を表す変数
    - $P(t \leq T < t + \Delta t | T \geq t)$：時点$t$までイベントが発生しておらず、次の微小時間$\Delta t$以内にイベントが発生する確率
  - $H(t) = \int_0^t h(u)\,du$を累積ハザード関数と呼ぶ。
- 上記の$h(t)$に基づき、生存確率$S(t)$は以下の通り計算される。
  - $S(t) = \exp\left(-\int_0^t h(u)\,du\right) = \exp(-H(t))$
    - $\frac{dS(t)}{dt} = \frac{dP(T>t)}{dt} = \frac{d(1-P(T \leq t))}{dt} = -P(T=t) = -S(t) \cdot h(t)$より上式が得られる。

12

ハザード関数を生存関数に変換ができると話しましたが、ここでは、その変換方法について、説明しております。本頁ならびに次の数頁では、こうした手法の詳細や、人工データによる分析例等をお見せしておりますが、本発表では時間の都合で、省略させていただきます。

コックス比例ハザードモデル 3/6
～（ご参考）ハザード関数の推定方法～

- ハザード関数$h(t|\mathrm{X})$の回帰係数βは、以下の部分尤度を最大化する$\hat{\beta}$として求められる。
  - $L(\beta) = \prod_{i \in D} \frac{\exp(\beta \cdot X_i)}{\sum_{j \in R(t_i)} \exp(\beta \cdot X_j)}$
    - D：イベントが発生した個体の集合
    - $R(t_i)$：t_iの直前でイベント未発生の個体の集合
    - X_i：個体iの説明変数ベクトル$X_i = (X_{i1}, \ldots, X_{ip})$

13

コックス比例ハザードモデル 4/6
～（ご参考）ベースラインハザード関数の推定方法～

- ベースラインハザード関数$h_0(t)$は、推定した累積ハザード関数$\hat{H}_0(t)$をtで微分する（但し本例では離散時間t_iを考えるため増分を取る）ことにより以下の通り与えられる。
  - $\hat{H}_0(t) = \sum_{t_i \le t} \frac{\delta_i}{\sum_{j \in R(t_i)} \exp(\hat{\beta} \cdot X_j)}$ ベースラインのリスクを評価するため、分子は$\exp(\hat{\beta} \cdot X_j)$ではなく$\delta_i$
    - δ_i：イベントが発生した場合1、発生していない場合0
    - $R(t_i)$：t_iの直前でイベント未発生の個体の集合
    - $\hat{\beta}$：前頁で推定した回帰係数
  - $h_0(t_i) = \frac{\delta_i}{\sum_{j \in R(t_i)} \exp(\beta \cdot X_j)}$

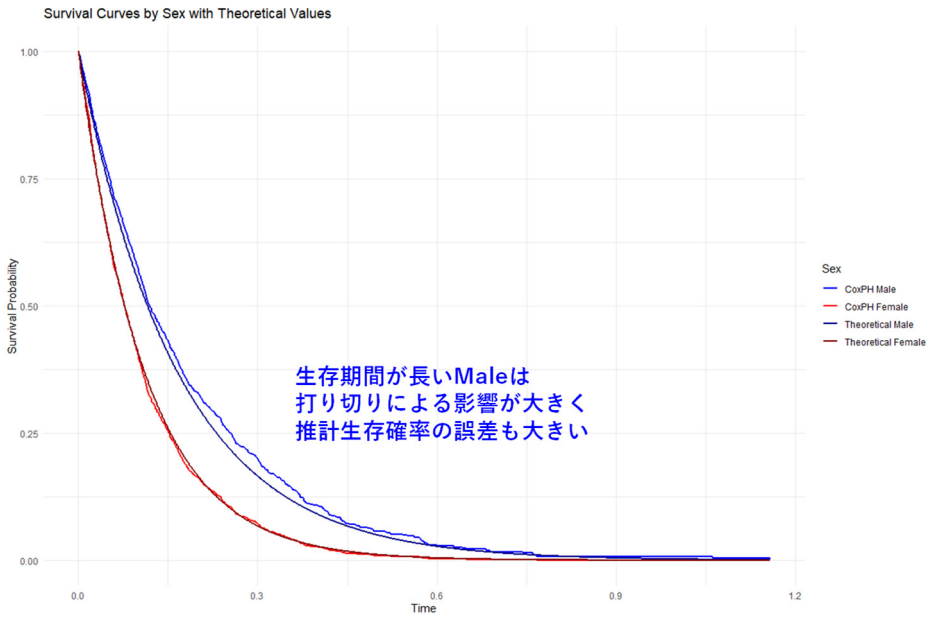
コックス比例ハザードモデル 5/6
～簡易データによる推定イメージ①～

- 人工データを用いたコックス比例ハザードモデルによる生存関数の推定イメージは以下の通り。
  - データ：説明変数（Age、Sex、Weight）、打ち切り時刻、イベント発生時刻に以下を仮定し、1,000個サンプルを生成。
    - $Age \sim N(\mu = 50, \sigma = 10)$
    - $Sex \sim 1: Male(p = 0.5), 2: Female(p = 0.5)$
    - $\text{Weight} \sim N(\mu = 70, \sigma = 10)$
    - 打ち切り時刻 $\sim \exp(0.5)$ 期待値では、時刻2（=1÷0.5）において打ち切りが起こる
    - イベント発生時刻 $\sim \exp(H)$
    - $H(= Hazard) = 0.02 + 3 \cdot (sex - 1) + 0.05 \cdot age + 0.05 \cdot weight$
  - 上記に対して、$h(t|X) = h_0(t)\exp(\beta_1 \cdot Age + \beta_2 \cdot sex + \beta_3 \cdot Weight)$を仮定（即ちデータの生成方法と同じ構造を仮定）してモデルを当てはめる。
  - R package: ‘survival’を使用。

コックス比例ハザードモデル 6/6
～簡易データによる推定イメージ② ～

- Age, Sex, Weightをmedianとして、Sexのみを動かして推計された生存関数をプロットすると次頁の通りとなる。生存関数の理論値（Theoretical）は、 $\exp(-H \cdot time)$によりプロット。

16



17

ランダム・サバイバル・フォレスト 1/4
～手法の概要～

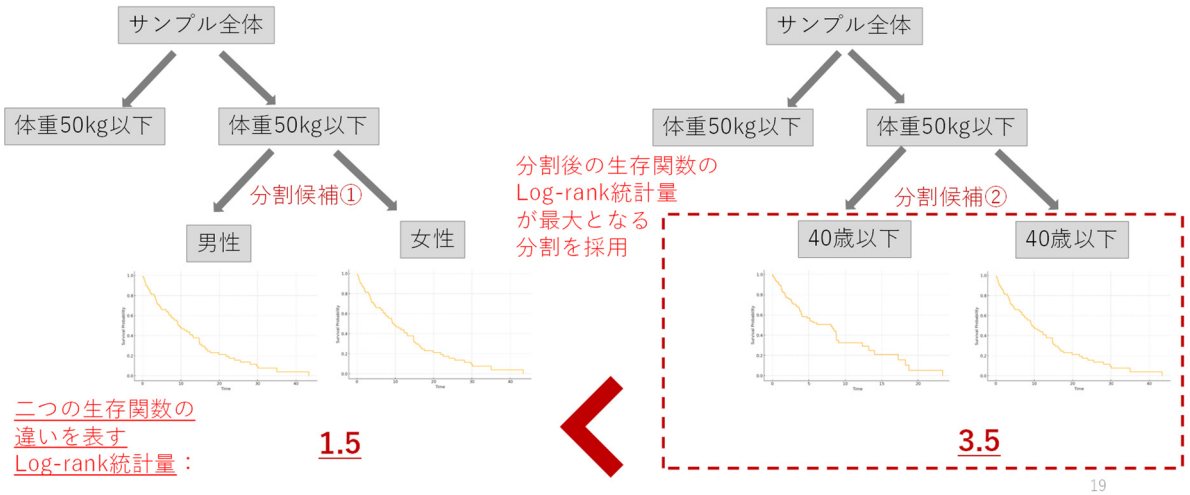
- ランダムフォレストを拡張したノンパラメトリックな手法で生存関数を推定する。アルゴリズムの概要は以下の通り。
  - 元のデータセットからブートストラップサンプルを生成し、それぞれに対して、以下通り決定木の学習を行う。
    - 各決定木の各ノードで、ランダムフォレスト同様、説明変数の全候補から一部をランダムに選び、その中で、最適な分割を与える説明変数により、同分割を実施する（分割のイメージや分割基準は次頁参照）。
  - 各決定木$(b = 1, \dots, B)$より得られた累積ハザード関数$H_b(t)$の単純平均値$H(t) = \frac{1}{B}\Sigma_{b=1}^{B} H_b(t)$から、生存関数$S(t) = \exp(-H(t))$を推定。

ランダム・サバイバル・フォレストの原論文：Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics, 2*(3), 841-860.

18

そして今回メインでご紹介したい、ランダム・サバイバル・フォレストでございます。手法のポイントはランダムフォレストと同様です。ブートストラップサンプリングにより多数のサンプルを発生させて木を構成した結果に基づき平均を取ります。ランダムフォレストとランダム・サバイバル・フォレストの大きな違いは、「どのようにして枝を分けるのか」です。ランダムフォレストでは、目的変数についての予測誤差を二乗誤差で評価し、その二乗誤差を最小化するべく枝を分けていきますが、ランダム・サバイバル・フォレストにおいて、「生存関数の評価ではどのように枝を分けていくのか」について、次頁でご説明します。なお、この枝を分ける操作を分枝と呼びます。

ランダム・サバイバル・フォレスト 2/4
～分割のイメージ（Log-rank統計量による分割）～



19

これはイメージですが、左図・右図のどちらにおいても、最初の分枝において、サンプル全体から体重が

50kg 以下・以上によってグループ分けしています。しかしそこから、「50kg 以下の人を男・女に分けるのですか、それとも 40 歳以上・以下で分けるのですか、どちらで分けましょう？」というところで意見が分かれています。この例を用いて、分割を決定する方法をご説明します。まずは左図をご覧ください。

体重 50kg 以下について、「男・女」で分枝していますが、その分枝後のグループ毎に、生存関数を、カプラン・マイヤー法で評価します。ここで、生存関数の両者の差を定量化する指標として、Log-rank 統計量といわれているものを用います。この統計量が大きいほど、際立って違う生存関数であることが示されます。この統計量を、左図の「男・女」の間、もしくは右図の「40 歳以下・以上」間で比較すると、40 歳以下・以上の間の方が、生存関数の相違が大きくなっています。つまり右図の分枝の方が、生存関数の観点からは、より明確にグループ間の特徴を分けることができている、と言うことができます。

このように Log-rank 統計量に基づき生存時間関数の違いを評価し、その違いを最大化する分割を採用する、というのがランダム・サバイバル・フォレストにおける分枝のポイントになります。

ランダム・サバイバル・フォレスト 3/4 ～（ご参考）Log-rank統計量の詳細～

- ある説明変数によりデータをグループA, Bに分割することを考える。分割点は、以下のLog-rank統計量U（期待イベント数・イベント数の分散は超幾何分布の仮定より導出）が最大となるもの※を選ぶ（なお他の統計量による分割方法も存在する）。

※ Log-rank統計量Uが大きい場合、2つの集団間の生存関数に大きな差異があること（性質の異なる2つの集団への分割）を意味し、同分割の有効性が示される。

- $Log-rank$統計量[注] $= \dfrac{\left(\sum_j d_A(t_j) - E_A(t_j)\right)^2}{V}$
  - $n_i(t_j), d_i(t_j)\ (i = A, B)$ ： 時刻t_jの各グループの総個体数とイベント発生数
  - $E_i(t_j)\ (i = A, B) = \dfrac{n_i(t_j)}{n_A(t_j)+n_B(t_j)}\left(d_A(t_j) + d_B(t_j)\right)$：期待イベント数
  - $V = \sum_j \dfrac{n_1(t_j)n_2(t_j)\left(d_1(t_j)+d_2(t_j)\right)\left(n_1(t_j)+n_2(t_j)-d_1(t_j)-d_2(t_j)\right)}{\left(n_1(t_j)+n_2(t_j)\right)^2\left(n_1(t_j)+n_2(t_j)-1\right)}$：イベント数の分散

注）ここではグループAで統計量を計算しているが、グループBで計算しても値は一致する

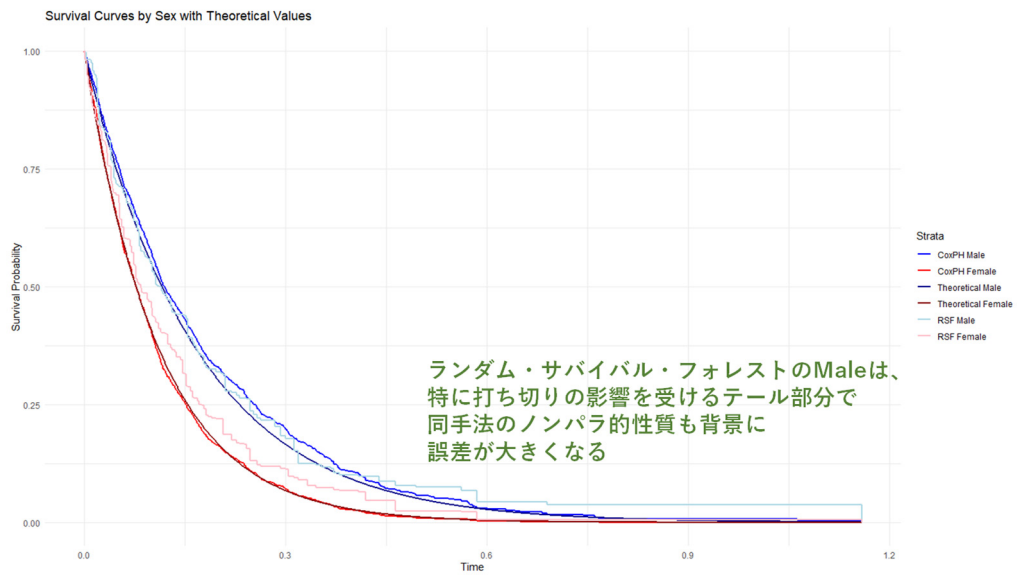
20

Log-rank 統計量の詳細について興味がある方は、こちらのスライドをご覧いただければと思いますが、今回は時間の都合で説明を省略させて頂きます。

ランダム・サバイバル・フォレスト 4/4 ～簡易データによる推定イメージ～

- コックス比例ハザードモデルに適用したデータに対して、 Age, Sex, Weightはmedianとして、生存関数を推定した結果は次頁の通り
  - 分析にはR package: 'randomForestSRC'を使用する。ハイパーパラメータはntree=1,000、その他はdefaultの設定。
- コックス比例ハザードモデル（人工データの生成モデルそのもの）より(当然)精度は劣るが、ランダム・サバイバル・フォレストは、ノンパラメトリックな手法に係わらず、データに仮定した男女較差を一定程度捕捉できている。

21



22

人工データに関する分析も同じく、今回はスキップさせて頂きますが、興味がある方はご覧いただけますと幸いです。

2. 分析例

- 使用データ（乳がん患者の生存確率）
- 分析方法
- 各三手法による分析結果

23

本章では、データを使った分析例をご紹介します。

使用データ 1/2
～概要～

- Kaggle: Breast Cancer (METABRIC)
  - https://www.kaggle.com/datasets/gunesevitan/breast-cancer-metabric
  - 乳がん患者2,509名に対するがん診断後の生存期間・無再発生存期間と関連情報（診断時年齢、手術種類等）のデータ

24

Kaggle から取得した乳がん患者のデータを使用します。約 2,500 人のがん患者に対して、がん診断後の生存期間と共に、各患者に紐づく診断時の年齢や手術の種類などの属性データを含んでおります。

数値実験の使用データ 2/2
～データ加工～

- 今回の分析対象は、再発の有無ではなく、生存状態。
- 元データの全部34カラム注のうち以下の6つを特徴量として選定。
- 生存状態/全生存期間、各特徴量が欠損でない1,390個のデータ(全体の約55%)が分析対象。

項目名	説明	入力内容	欠損割合	モデルへの入力方法
Age at Diagnosis	診断時の患者の年齢	21.9歳～96.3歳	0.4%	数値型
Type of Breast Surgery	乳癌の手術の種類	乳房温存/乳房摘出	22.1%	カテゴリカル
Tumor Stage	腫瘍の進行度を示すステージ	0～4（1刻み）	28.7%	数値型（0~4をそれぞれ整数に変換）
Cellularity	腫瘍の細胞密度を示す指標	Low/Moderate/High	23.6%	数値型（Low:1, Moderate:2, High:3）
Tumor Size	腫瘍のサイズ	1.0~182.0（直径mm？）	5.9%	数値型
Nottingham prognostic index	乳癌の予後を予測するためのスコア	1.0～7.2	8.8%	数値型
Overall Survival Status	生存状態	死亡/生存	21.0%	カテゴリカル
Overall Survival (Months)	全生存期間（診断から死亡 or 最後の追跡までの月数）	0.0~355.2カ月	21.0%	数値型
Relapse Free Status	再発の有無	再発あり/再発なし	0.8%	カテゴリカル
Relapse Free (Months)	再発までの期間（再発がなければ死亡または最後の追跡までの期間）	0.0~384.2カ月	4.8%	数値型

注：カラムの中には今回の分析には有用でないもの（患者ID、性別（データは全て女性）等）も含まれている

25

説明書きが多くて恐縮ですが、生存関数の評価には、生存状態と、全生存期間、つまり観測期間中に実際に生きていた期間の情報を使用します。これらは赤色の項目に対応します。この生存関数が目的変数になります。この生存関数を説明するために用いる指標は緑色の項目に対応します。例えば、診断時の患者の年齢、腫瘍のサイズ等に基づき、生存関数を評価します。赤色と緑色の項目が使用データですが、これらの項目で欠損がある個体は、今回は簡単のため全部除外しております。「データは約 2,500 個ある」と申し上げましたが、この欠損値を持つ個体の除外により、実際に使うデータは 1,390 個となります。

（予測精度）モデルの評価方法

- 全データの75%が学習データ、残り25%はテストデータ。
- 以下の３つの手法について、学習データでモデル化した上で、テストデータから各時点の生存確率を計算し、AUC（Area Under the Curve）により精度を評価。
  - カプラン・マイヤー法(KM)
  - コックス比例ハザードモデル(CoxPH)
  - ランダム・サバイバル・フォレスト(RSF)
- AUCはROC（Receiver Operating Characteristic）曲線の下の面積に対応し、0から1の値を取り、値が大きいほど予測精度が高いことを示す。

注：AUCによるモデル評価のプロセスはKaggleを参照：https://www.kaggle.com/code/gunesevitan/survival-analysis, 2024/10/13 accessed.

26

詳細は後述しますが、モデルの評価として、「AUC」を用います。モデルの予測精度の検証にあたり、全データの 75%を学習データ、25%をテストデータとします。つまり、75%のデータを使ってモデルを構築し、25%

のデータを用いて正しく予測できているかを評価します。「AUC とは何か？」ですが、AUC とは、ROC の下側の面積に対応していて、大まかに申し上げると「AUC が大きいほど予測精度が高いといえる」という指標です。現段階では、何を言っているか分からないと思いますが、ここではポイントとして、「AUC は大きい方がよい」というイメージを持っていただければと思います。

ROC曲線とAUC 1/3 ～偽陽性率と真陽性率～

- ある時点において、各テストデータに対して各モデルで予測した死亡確率[注]を「スコア」として、スコアが或る閾値以上（以下）で死亡（生存）と予測する場合の偽陽性率と真陽性率を計算（陽性＝死亡、陰性＝生存に対応）。
  - 偽陽性率 $= \frac{\#\text{誤って陽性判定}}{\#\text{陰性母集団}} = \frac{\#\text{誤って陽性判定}}{\#\text{誤って陽性判定} + \#\text{正しく陰性判定}}$
  - 真陽性率 $= \frac{\#\text{正しく陽性判定}}{\#\text{陽性母集団}} = \frac{\#\text{正しく陽性判定}}{\#\text{正しく陽性判定} + \#\text{誤って陰性判定}}$
- 上記の閾値を変えて偽陽性率と真陽性率でプロットしたものがROC曲線であり、その下側の面積がAUCに対応。

注：手法によっては、死亡確率の代わりに各サンプルのリスクスコア（例えばコックス比例ハザードモデルでは$\exp(\beta \cdot X)$）を用いる場合がある。

27

AUC を説明する上で重要な指標として、偽陽性率と真陽性率があります。今回の生存時間分析という文脈で説明するため、陽性は死亡、陰性は生存と置き換えて説明します。ここで、推計した死亡率をスコアとして、そのスコアがある閾値以上の場合には死亡と判定する、ということを考えます。偽陽性率は、実際には生存しているのに「死亡」と判定されてしまった人の割合です。そして真陽性率は、死亡した人を正しく「死亡」と判定した人の割合です。閾値を下げてスコアが何であろうと全員死亡と判定してしまえば、真陽性率は上がる一方、偽陽性率も一緒に上がってしまいます。このように、閾値を変えながら、真陽性率と偽陽性率の関係をプロットしたものを ROC 曲線と呼びまして、その曲線の下側の面積を AUC と呼びます。

ROC曲線とAUC 2/3
～ROC曲線の計算例～

- あるモデルの推計結果と３つの閾値（X）による偽陽性率と真陽性率の計算例。
- Xを動かしてROCをプロットする。

がん診断150カ月後		推計死亡率≧Xを死亡と予測した時の推計結果		
実際の状態	推計死亡率	X:30%	X:50%	X:70%
死亡	90%	死亡	死亡	死亡
生存	78%	死亡	死亡	死亡
死亡	65%	死亡	死亡	生存
死亡	56%	死亡	死亡	生存
死亡	52%	死亡	死亡	生存
生存	45%	死亡	生存	生存
生存	36%	死亡	生存	生存
死亡	29%	生存	生存	生存
生存	25%	生存	生存	生存
生存	19%	生存	生存	生存
生存	16%	生存	生存	生存
偽陽性率= $\frac{\#誤って死亡判定}{\#実際の生存者}$		$\frac{3}{6}=50\%$	$\frac{1}{6}=17\%$	$\frac{1}{6}=17\%$
真陽性率= $\frac{\#正しく死亡判定}{\#実際の死亡者}$		$\frac{4}{5}=80\%$	$\frac{4}{5}=80\%$	$\frac{1}{5}=20\%$

閾値が下がると
偽陽性率は上がるが
真陽性率も上がる
トレードオフの関係

28

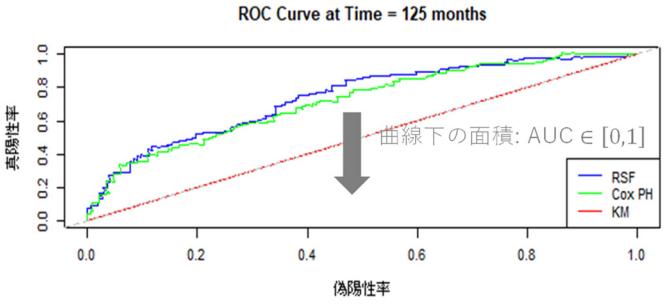
偽陽性率、真陽性率の計算方法に関して、例を用いてご説明します。この例は、がん診断後 150 か月後の状況です。実際の状態として、左側に「死亡」や「生存」とありますが、それに対して右隣の列には、死亡率の推計値が記載されております。これらの推計値は、生存時間分析の何れかのモデルで推計した結果と仮定します。この X が、先ほど申し上げた閾値に対応しまして、例えば一番左側のものは、閾値が 30%を超えたら全員死亡と判断することを意味します。

偽陽性率ですが、実際の生存者数を母集団として、誤って死亡判定してしまった人の割合です。実際の生存者数は左側の中の「生存」となっている人の人数に対応します。その中で間違って死亡と判断してしまった人なので、生存なのだけれども死亡、例えば上から 2 行目の個体が対応します。このような個体を集めて算出した比率が、偽陽性率となります。

真陽性率は、実際の死亡者を母集団として、誤って生存判定してしまった人の割合です。本例では下から 4 行目の個体に対応します。閾値 X が 30%の場合、真陽性率は 80%となります。閾値は、最右列から左側になるにつれて下がっていますが、これにより、どんどん死亡と判断しやすくなるので、真陽性率は上がるものの偽陽性率も上がってしまう、ということになります。

ROC曲線とAUC 3/3
～AUCの計算例～

- 各手法に基づくROCの例（結果の解釈は後述するが、テストデータを用いて、がん診断後125カ月経過後の死亡率に対して適用したもの）。
- この例では最大のAUCを与えるRSF（0.74）が最良。



	AUC		
Time (month)	KM	CoxPH	RSF
125	0.50	0.72	0.74

注）KM：カプラン・マイヤー法（'Survival'）、CoxPH：コックス比例ハザードモデル（'Survival'）、RSF：ランダム・サバイバル・フォレスト（'randomForestSRC'）
括弧内は使用したR Package。ROC曲線・AUCの計算には、R packageの"survivalROC"を使用。

29

前述の両者の関係を、横軸を偽陽性率、縦軸を真陽性率としてグラフで表しております。ここでは、125 か月後の生死の情報に対する推計を例示しております。完璧なモデルでは、真陽性率 100%で、偽陽性率が 0%になるので、グラフの長方形の左上頂点を含む形で、左辺と上辺に沿ったプロットとなります。この場合、その曲線下の面積は 1 になります。これが AUC に対応します。そのため、ACU の最大値は 1 です。ここでランダム・サバイバル・フォレストを RSF、コックス比例ハザードモデルを Cox PH、カプラン・マイヤーを KM として、右表に AUC を掲載しております。「AUC は、曲線の下の面積に対応し、値が大きいほど予測精度が高いといえ、最大値は 1 である」ということを念頭に、青の線を見ていただきますと、本例では、この青い線に対応するモデルが 3 手法の中で最良の手法といえます。つまり、ランダム・サバイバル・フォレストが最良の手法となります。

＜予測精度＞評価結果 1/5
～AUC～

- 前述の通り、1,390個のデータのうち、75%を学習データ、残り25%をテストデータとして、がん診断後経過時間 time=50, 125, 200カ月のそれぞれに関してAUCを計算した結果は以下の通り。
- 何れの時点でもRSFが最良の結果を与えた。
- なおKMは全ての個体に対してリスク較差がない（同じ死亡率を適用する）ため、時点に依らず0.5（完全にランダムな予測と同じ）となる。

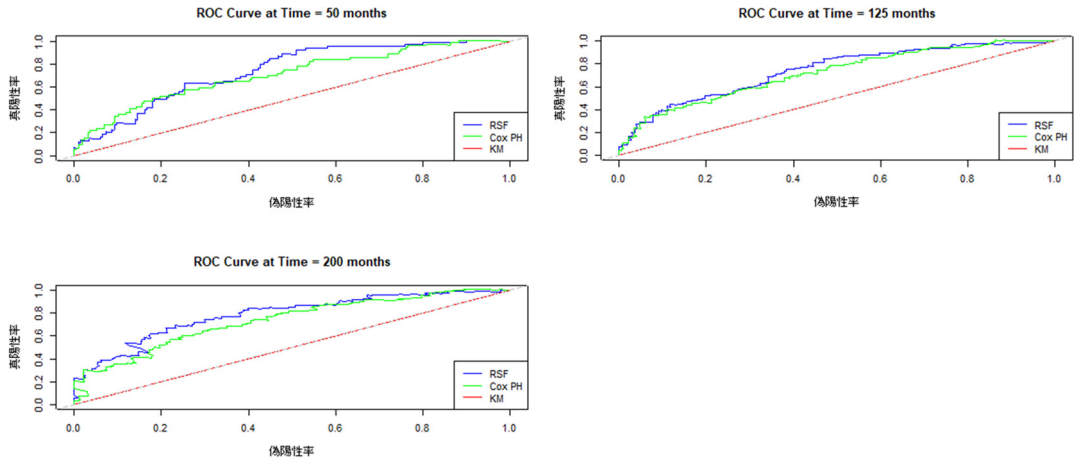
Time (month)	AUC		
	KM	CoxPH	RSF
50	0.5	0.71	**0.74**
125	0.5	0.72	**0.74**
200	0.5	0.74	**0.78**

注）KM：カプラン・マイヤー法（'Survival'）、CoxPH：コックス比例ハザードモデル（'Survival'）、RSF：サバイバル・ランダムフォレスト（'randomForestSRC'）
括弧内は使用したR Package。ROC曲線・AUCの計算には、R packageの"survivalROC"を使用。

30

これまでのご説明を踏まえて、各手法について、偽陽性率と真陽性率を計算して AUC を計算・比較した結果は次のとおりです。経過時間として、50 か月、125 か月、200 か月後の 3 パターンを定めて、それぞれについて、各個体の生死に関する情報が得られますので、そこから AUC を計算します。AUC の計算結果は、この表に記載されております。いずれの場合にも、ランダム・サバイバル・フォレストが最大の AUC を与える結果となりました。

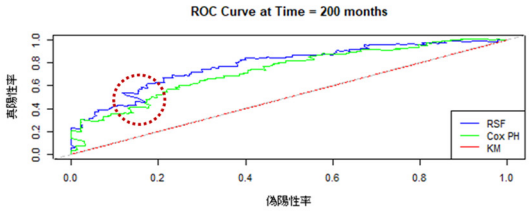
＜予測精度＞評価結果 2/5
～ROC曲線～



注）KM：カプラン・マイヤー法（'Survival'）、CoxPH：コックス比例ハザードモデル（'Survival'）、RSF：サバイバル・ランダムフォレスト（'randomForestSRC'）
括弧内は使用したR Package。ROC曲線・AUCの計算には、R packageの"survivalROC"を使用。

31

＜予測精度＞評価結果 3/5
～（ご参考①）ROC曲線が左上に延びる現象の要因～



＜ROC曲線が左上に延びている理由＞
- 打ち切りデータも活用するべく、KMで推定した母集団に対して、閾値c の真・偽陽性率を下式で推定するため。

真陽性率 $= \dfrac{\{1-\hat{s}_{KM}(t|X>c)\}\cdot\{1-\hat{F}_x(c)\}}{1-\hat{s}_{KM}(t)}$ 時点tのKMでの推定死亡者中のスコアc以上の割合

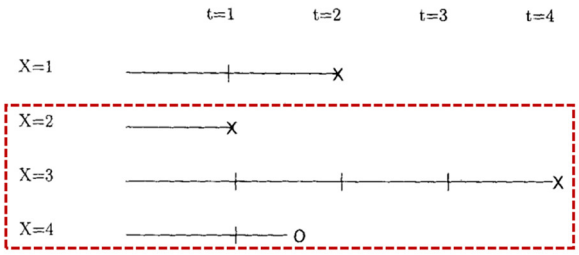
偽陽性率 $= 1-\dfrac{\hat{s}_{KM}(t|X\le c)\cdot\hat{F}_x(c)}{\hat{s}_{KM}(t)}$ 時点tのKMでの推定生存者中のスコアc以上の割合

$\hat{F}_x(c)$: 各手法による推定スコアXの累積分布関数
$\hat{s}_{KM}(t|\cdot)$: KM法による時点tの推計生存確率

単調性（ $\mathbf{P}(X<c'|\text{死亡})\le\mathbf{P}(X<c|\text{死亡}), c'<c$）が保証されない（次頁例参照）

注）ROC曲線の推定手法の詳細は、R package"survivalROC"の以下の原論文を参照。
Heagerty, P.J., Zheng, Y. (2005) Survival Model Predictive Accuracy and ROC Curves Biometrics, 61, 92 – 105.

＜予測精度＞評価結果 4/5
～（ご参考②）ROC曲線が左上に延びる現象例～



×：死亡
O：打ち切り

- $\hat{s}_{KM}(t=2|X>0)=\frac{3}{4}\cdot\frac{1}{2}=\frac{3}{8}$
- $\hat{s}_{KM}(t=2|X>1)=\frac{2}{3}\cdot\frac{1}{1}=\frac{2}{3}$
- $\hat{P}(t=2, 1\ge X>0|\text{死亡}) =$
$\hat{s}_{KM}(t=2|X>0)\cdot\left(1-\hat{F}_X(0)\right)$
$-\hat{s}_{KM}(t=2|X>1)\cdot\left(1-\hat{F}_X(1)\right)$
$=\frac{3}{8}\cdot(1-0)-\frac{2}{3}\cdot\left(1-\frac{1}{4}\right)=-\frac{1}{4}<0$ 負の値になってしまう

単調性（$\mathbf{P}(X<c'|\text{死亡})\le\mathbf{P}(X<c|\text{死亡}), c'<c$）が保証されない

これは「打ち切り個体$X=4$が少なくともt = 1まで生存した」という情報が各時点tの生存確率推計に与える影響が、推定対象であるXのセットに依存するため

注）本例の出典： Heagerty, P.J., Zheng, Y. (2005) Survival Model Predictive Accuracy and ROC Curves Biometrics, 61, 92 – 105.

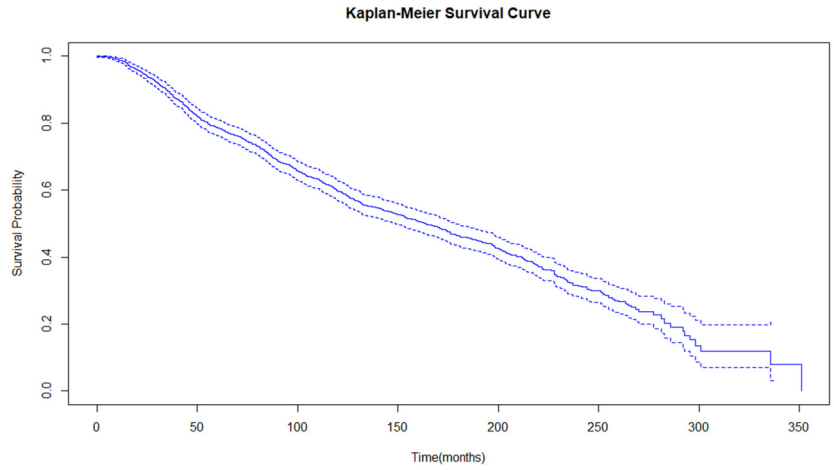
＜予測精度＞評価結果 5/5
～各時点でのAUCプロット～

各時点でAUCが
RSF ＞ **CoxPH**

34

予測精度に関する補足のスライドもいくつかご用意しておりますが、今回は時間の関係でスキップさせていただきます。

＜結果解釈＞全データでのモデル推計結果 1/4
～KM（生存曲線）～

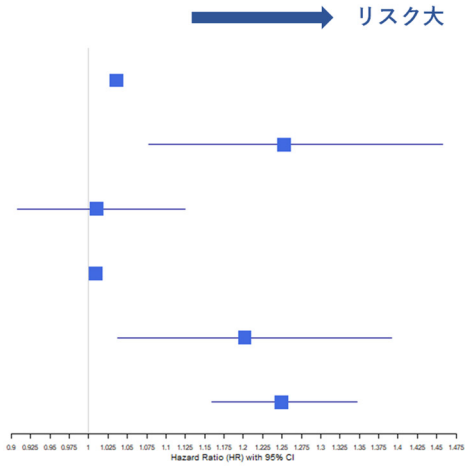


35

次は、結果の解釈です。予測精度の評価として、AUC が利用できる旨をご説明しましたが、結果の解釈としては、カプラン・マイヤー法では、例えば、生存曲線を用いた結果の確認ができます。

＜結果解釈＞全データでのモデル推計結果 2/4 ～CoxPH（パラメータ推定値）～

Variable	Coefficient	HR (95% CI)
Age.at.Diagnosis	0.035	1.036 (1.029 - 1.042)
Type.of.Breast.SurgeryMastectomy	0.225	1.253 (1.077 - 1.457)
Cellularity	0.011	1.011 (0.908 - 1.125)
Tumor.Size	0.008	1.009 (1.004 - 1.013)
Tumor.Stage	0.184	1.202 (1.037 - 1.392)
Nottingham.prognostic.index	0.223	1.250 (1.159 - 1.347)

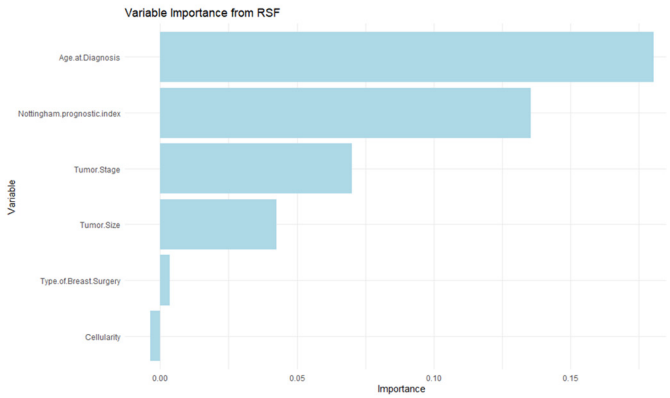


注）プロットにはR packageの'forestplot'を使用。Breast.SurgeryMastectomyは「乳房摘出」に対応（Breastのもう一つのカテゴリは「乳房温存」）。

36

コックス比例ハザードモデルでは、回帰係数の推計結果が確認できます。回帰係数が正値の場合は、各説明変数が増えると死亡リスクが増えることを意味し、回帰係数の絶対値が大きいほど、リスクを増大させる影響が大きくなることを示します。診断年齢や腫瘍のステージに対応する回帰係数は正値となっておりますので、診断年齢の上昇、腫瘍ステージの上昇により、リスクが増えると考えられます。

＜結果解釈＞全データでのモデル推計結果 3/4 ～ RSF（変数重要度）～



変数重要度（Variable Importance）
- 説明変数をシャッフルして予測性能がどれだけ低下するかを測定することで算出する指標。
- 重要な説明変数をシャッフルすると、予測性能が大きく低下し、重要度スコアが高くなる。

ランダム・フォレスト系のアルゴリズムで良く用いられる指標

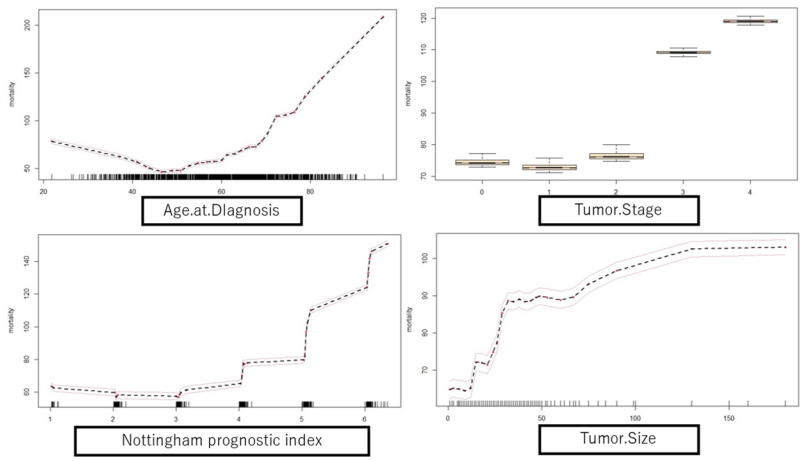
注1）変数重要度の算出には、R packageの'randomForestSRC'を使用。
注2）予測性能を測る指標は、Harrell's C-index（全ペアに関して、どちらが長く生存するかを正しく予測できた割合により計算）が用いられている。詳細は以下を参照：
Ishwaran, H., & Kogalur, U. B. (n.d.). Random Forests for Survival, Regression, and Classification (RF-SRC). Retrieved from https://www.jstatsoft.org/article/view/v063i01, 2024/10/13 accessed.

37

ランダム・サバイバル・フォレストにおける結果の解釈方法として、例えば変数重要度が利用できます。この例では、一番上の項目が患者の年齢を表していますが、患者の年齢以外を固定して、患者の年齢をランダムにシャッフルしたときに、どれだけ予測精度が下がるのかを示しています。予測精度が大きく下がる場合、その変数は予測上重要だと判断できます。この予測精度の低下の度合いに関して、一つひとつの変数に

ついて評価した結果が、こちらの変数重要度となります。

＜結果解釈＞全データでのモデル推計結果 4/4 ～ RSF（部分依存プロット）～



部分依存プロット
（Partial Dependency Plot）
- 特定の変数が
  モデルの予測に与える影響を
  視覚的に理解するためのツール。
- 指定した説明変数のみを
  動かしたときの予測結果の
  平均的な変化を示す。

ランダム・フォレスト系のアルゴリズムに限らず様々なモデルに用いられる。

RSFはCox PHとは異なり、非線形な効果を表現できる。

注）プロットにはR packageの'randomForestSRC'を使用。横軸の黒い線群はデータポイントを視覚的に示したもので、密集している部分は多くのデータが分布していることを表す。
mortality（累積ハザードの合算値）は100を基準として正規化されている（個体間で全ての特徴量が同じ場合はmortalityが100となる）。また点線の両端の赤線は±2σに対応。

38

他には、部分依存プロットというものもあります。これは、説明変数を一つ選んで、例えば、診断の年齢だけを動かして他を全部固定してあげることを考えた場合に、ハザードがどう変わるのかをプロットします。例えば、患者の年齢ですが、左上のプロットを見ていただきますと、年齢が上がるにつれてハザードが上がっていること、他の例としては右上のプロットですと、腫瘍のステージが上がるごとにハザードが上がることが確認でき、こうした各変数とハザードの関係を可視化することができます。

本発表のサマリー（再掲）

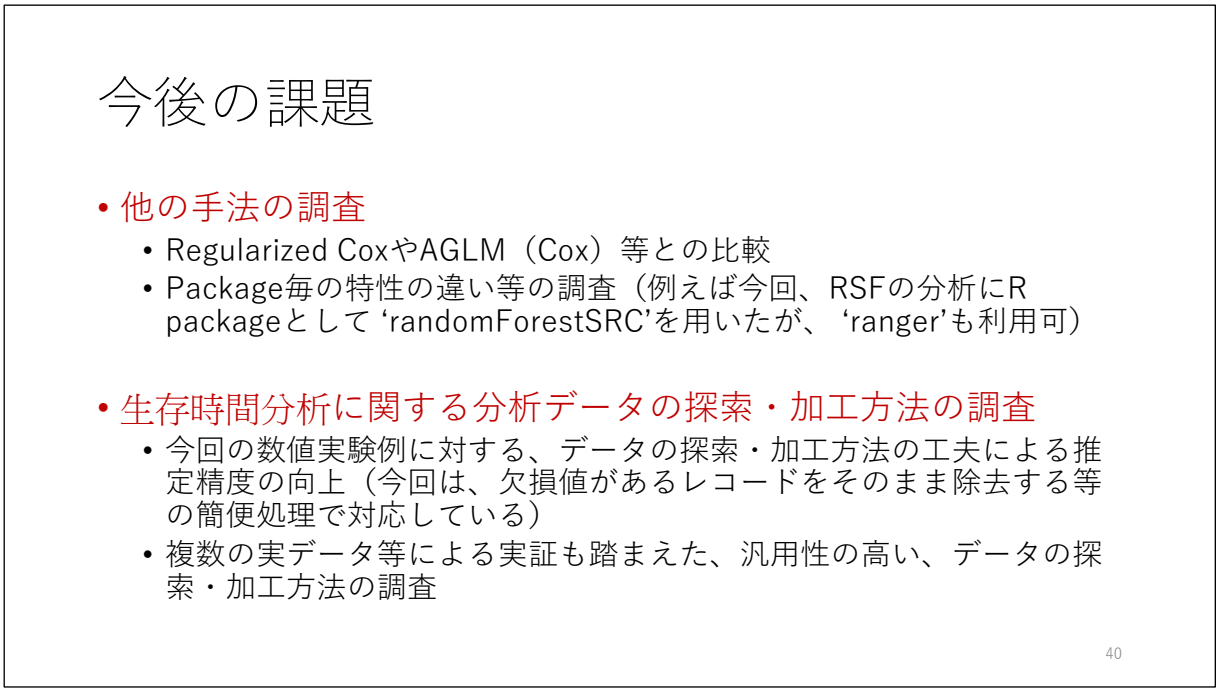
- 生存時間分析は、観察期間等の都合で一部しか観察できないデータも活用してイベントの発生等を分析する手法。
- 一例としてカプラン・マイヤー法、コックス比例ハザードモデル、ランダム・サバイバル・フォレストを紹介。
  - 予測精度の評価方法の一例として、AUCを用いる。
  - 結果の解釈では、各モデルの性質に応じて、パラメータ推定値、変数重要度、部分依存プロットの確認等を活用する。

⇒上記の例より、ランダム・サバイバル・フォレストについて、予測精度や解釈可能性の観点からの活用可能性をお見せしたい。

39

本発表では、「生存時間分析とは、何なのか」、そして「生存時間分析における典型的な手法やランダム・サバイバル・フォレストの概要」、そして、「AUC による予測精度の評価や、パラメータの推定値・変数重要度・部分依存プロットなどによる結果の解釈」等についてご説明しました。今回の発表を通じて、少しでも

生存時間分析やランダムフォレストにご興味を持っていただけましたら何よりでございます。



今後の課題ですが、冒頭にて、生存時間分析には様々な手法がある中で、今回、ほんの一部だけを切り取ってご紹介している旨をお伝えしましたが、今後は、より多くのモデルの比較を進めていくこと、そして、アクチュアリーとしては、やはり、各手法の詳細をきちんと深くまで理解することも重要だと思いますので、パッケージをそのまま使うのではなくて、一つひとつ深掘りしていくことが大切と考えております。

あとは、今回、データを用いた分析では、欠損値をそのまま除いて分析しておりますが、除外してしまったデータも活用できるよう、データ加工の方法を検討する等して、よりよい予測や推定に繋げることも考えられるかと思っています。

私からは以上でございます。ありがとうございました。

Q&Aセッション

皆様からご質問お願いします！

5

【藤田】　田中さん、ありがとうございました。

それでは、続いて質疑応答の時間に移りたいと思います。まずは、会場の参加者の方から、ご質問のある方は挙手にて、どなた宛てなのかも明示していただいた上で、お願いいたします。

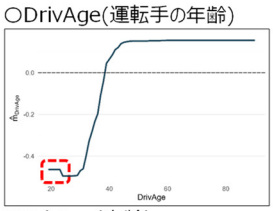
【A】　はい。発表ありがとうございました。松江さんに質問です。

4．RPFの実データ適用

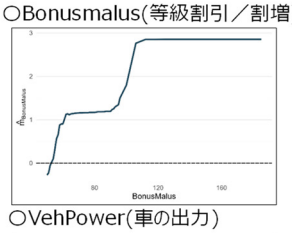
RPFによる関数分解（主効果）

- 推定された主効果は、いずれも納得性の高い解釈が可能である。
- とくに、DrivAgeの主効果では、RPFが年少者における僅かなリスク上昇を捉えることにも成功している。

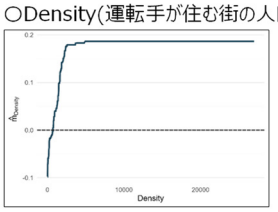
○DrivAge(運転手の年齢)



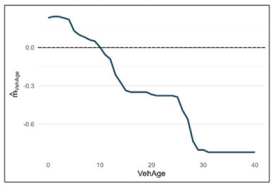
○Bonusmalus(等級割引／割増)



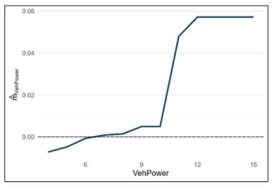
○Density(運転手が住む街の人口密度)



○VehAge(車齢)



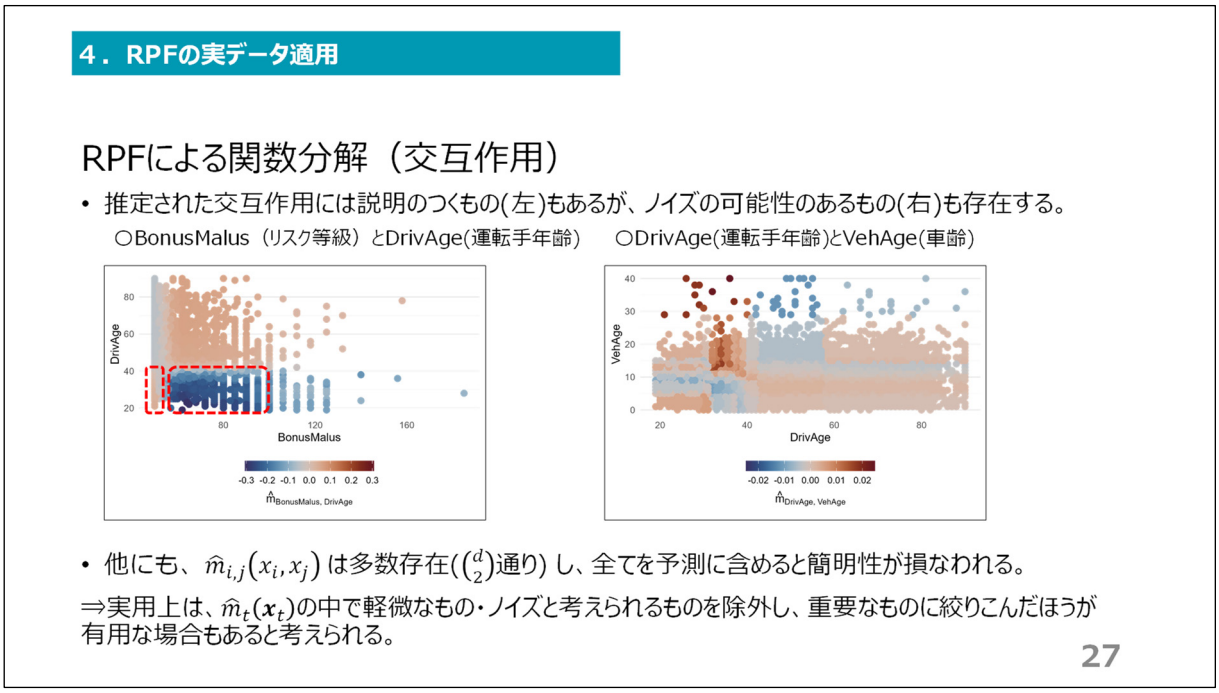
○VehPower(車の出力)



26

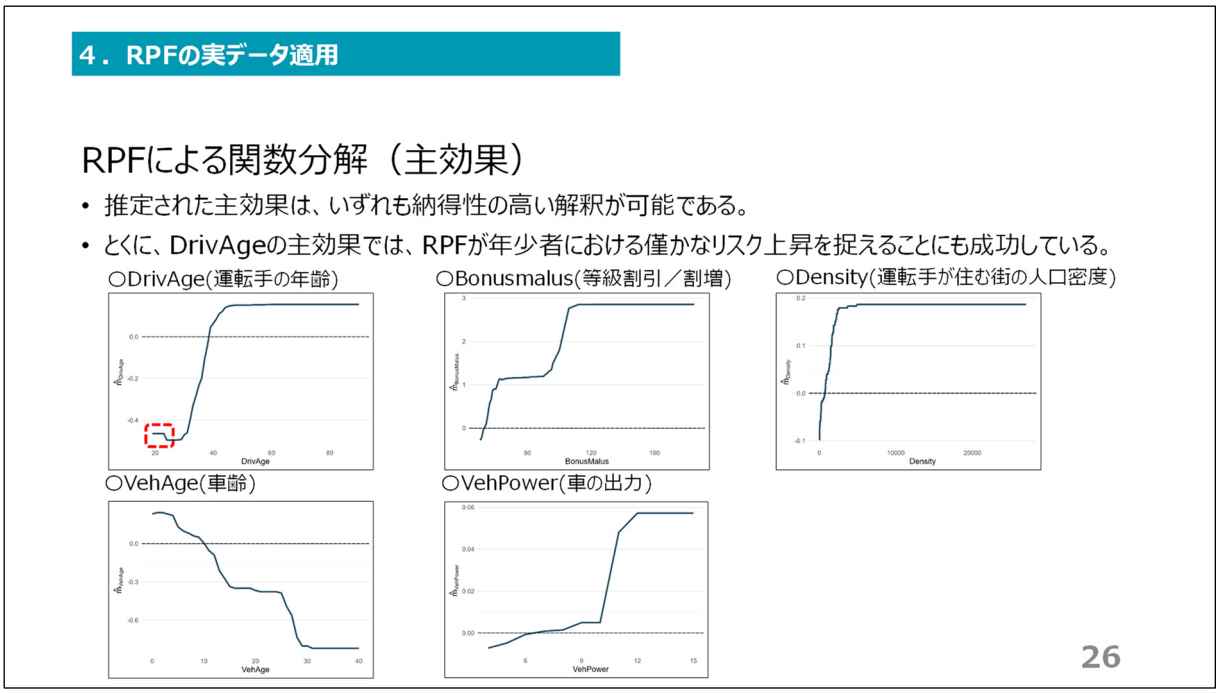
発表の中で、最後、自動車の事故のところで、すみません、私、もしかしたら聞き逃してしまったのかもしれないのですが、「若齢でリスクが上がります」と。それは、とてもよく分かるのです。逆に、高齢のところなのですけれども、例えば 70 歳や 80 歳、90 歳ぐらいでリスクが上がるということだったら、とてもよく

分かるのですけれども、そのグラフだと、40 歳ぐらいでリスクが急上昇しているように見えて、どうなのだろうと思ったので、もし、その点について何か教えていただけることがあれば、教えていただきたいということが 1 点です。

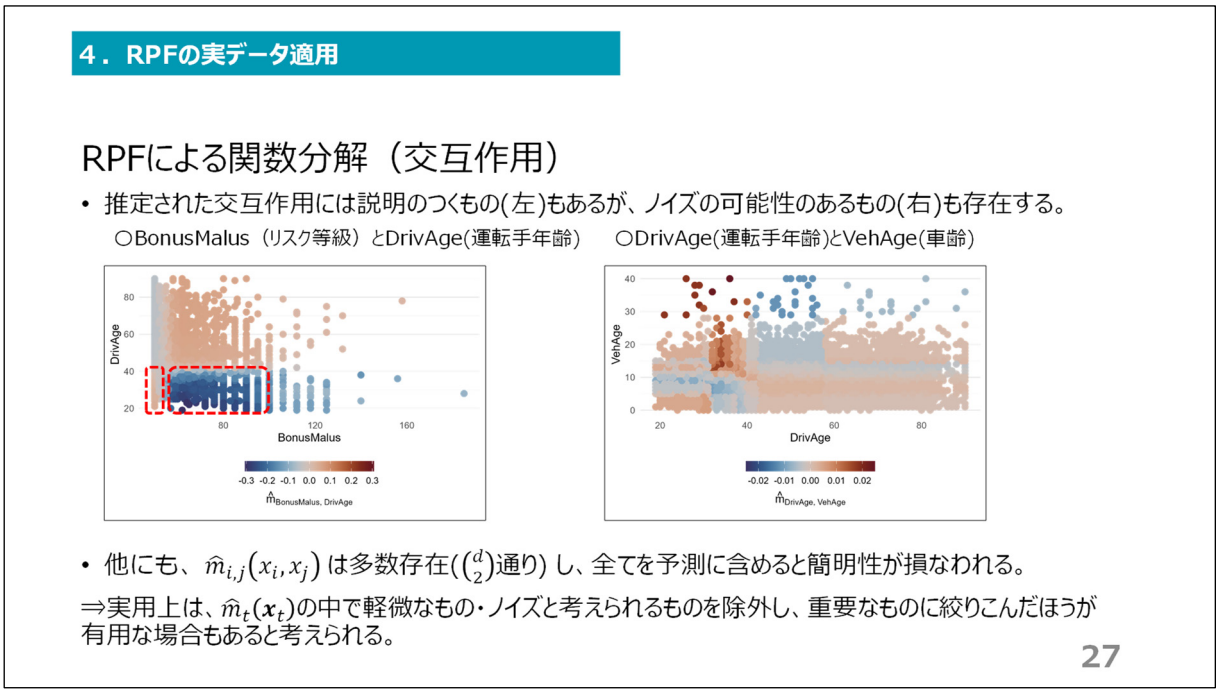


あともう 1 点が、次のページで、左側の「交互作用があります」というところ。いろいろ交互作用があるということ自体は理解したのですけれども、具体的に、どのような年齢とボーナス・マラスで交互作用が起きているのかということを、ご説明いただけるとうれしいというところでございます。

【松江】　分かりました。ありがとうございます。



なぜ平たんかといった、そこまでは深掘りはできておりません。サンプルを結構ダウンサンプリングしてしまっているので、高年齢が十分含まれているかどうか分からないということもありますし、なぜ 40 で止まっているのかは、細かくは見られていないです。確かに、高齢者になると認知機能が低下して事故率が上がるというようなことも考えられるので、おっしゃるとおり、なぜ上がらないのかといったところは、持ち帰りといいますか、確認できればと思っています。すみません。



交互作用、こちらはなかなか見にくいのですけれども、40 歳未満、さらに若ければ若いほど赤いということは見て取れて、ボーナス・マラスも、数値としては 50 などほぼ一番小さい値の辺りで、かなりプラスの交互作用が見られて、少しでもそこから増えるとマイナスとなっているといったところが見て取れます。

【A】　ありがとうございました。

【藤田】　ありがとうございます。他の方、いかがでしょうか。

【B】　すみません、田中さんに質問です。

生存時間解析をアクチュアリーサイエンスに応用する中で一番ぱっと思いつくものは、解約率や生存率・継続率かと思いまして。実際に読んだことはないのですけれども、生存時間解析がランダム・サバイバル・フォレストを含めて使われて、解約率モデルを作っているというものを見たことがあります。一方で、解約は、多分、動的解約のような外部的な金利要因などで時系列どんどん変わっていく、特徴量によって影響を受けるのではないかと思っていて。

質問したいことはコックス比例ハザードモデルのような式で書いてあげるモデルだと、多分、外部の金利など、特徴量がどんどん変わっていっても、生存時間のモデルを変化することはできると思うのですけれども、ランダム・サバイバル・フォレストだと、できなさそうな理解をしたのです。時系列で変わる特徴量も入れることは可能でしょうか。

【田中】　ご質問ありがとうございます。

質問のご趣旨は、「例えば金利などの外部変化を明示的に何かロジックとしてモデルに入れることは、コックス比例ハザードモデルだとできるものの、ランダム・サバイバル・フォレストだと難しいのではないか」ということですか。

【B】　そのとおりです、はい。

【田中】　おっしゃるとおりだと思います。

そこは、多分、ランダム・サバイバル・フォレストだけの話ではなくて、ランダムフォレストにもいえることと考えております。交互作用や非線形な効果は、いずれのモデルでも捉えられると思うのですが、明示的なロジック化は難しい、ということがあると思います。

そのため、これは私個人の見解ですが、例えばランダム・サバイバル・フォレストを、そのまま使うのではなくて、あくまでも線形モデル等をベースとして、それを拡張する際の検討材料として、ランダム・サバイバル・フォレストのような非線形モデルを使ってみて、その中で適宜、変数重要度や部分依存プロット等も参照してみるのが良いかと思っております。

【B】　分かりました。ありがとうございます。

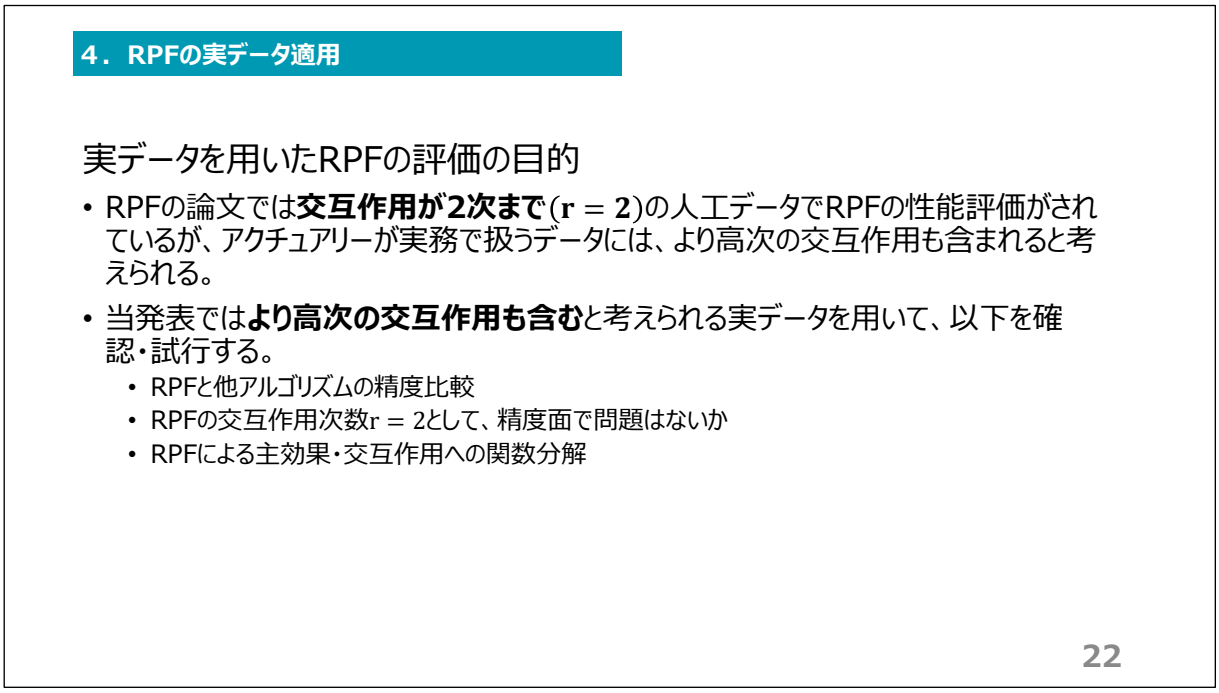
【藤田】　ありがとうございます。他にいかがでしょうか。

それでは、Slido で複数の質問をいただいているので、そちらからピックアップさせていただきます。

そもそものご質問です。「皆様にご質問です。先行研究に対して発表者のご貢献は何でしょうか。具体的に先行研究のご紹介という趣旨なのでしょうか、それとも何か新しいことを提案されているのか知りたいです」というご質問をいただきました。

私からも回答できるのですけれども、では、小田さんからお願いします。

【小田】　私のところは、松江さんや田中さんの説明の準備の側面があるというところです。



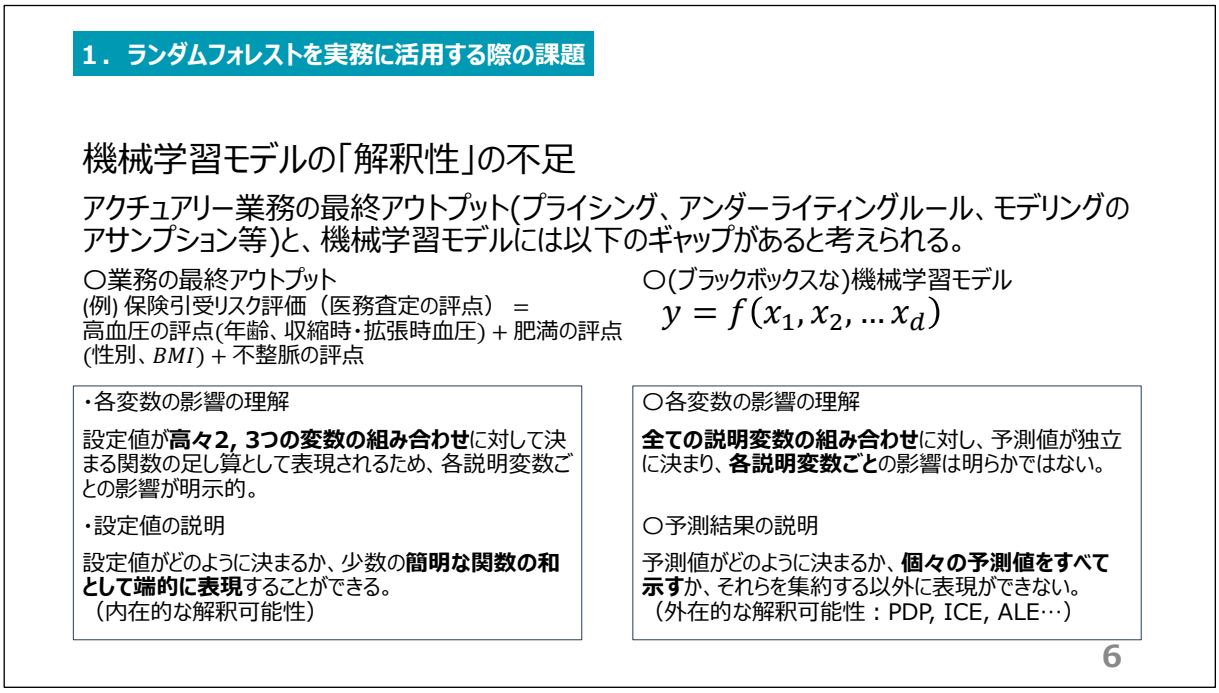
【松江】　松江です。私も、ほとんど先行研究の紹介という形にはなりましたが、実データに当てはめてみて、特に交互作用の次数が 2 とは限らないような場合でどうなのかを確認したことは、成果と言うほどではございませんけれども、プラスアルファのコンテンツだったのではないかと考えております。

【田中】　私のパートは、学術的な新規性はございません。私個人としては、生存時間分析自体をあまり使ったことがなかったため、今回、ランダムフォレストを題材としたセッションを通じて、生存時間分析と共に、ランダム・サバイバル・フォレストの概要に関してご説明することで、皆様にとって、これらの分野にご興味を持っていただく契機となればという趣旨で、発表させていただきました。

【藤田】　ありがとうございます。

このチームは、まさにセッション名どおり、ランダムフォレストをいかにアクチュアリー実務に、活用するかを模索しています。皆さんからおっしゃっていただいたように、先行研究の紹介というところに重きを置いているのですけれども、ランダムフォレストに関する研究は非常に多く、その中からアクチュアリアル・サイエンスに、どの部分が関連しているかというところも、まずピックアップして、今回、ランダムプランテッドフォレストというものと生存時間分析、そちらを選択し、そして、保険データや乳がんに関する、われわれのアクチュアリーの実務とも関連がありそうなデータに適用してみたということを、ご紹介させていただきました。

フロアからは、ご質問、大丈夫ですか。では、続いて Slido で、松江さんにご質問が来ております。「医務査定の評点の分析の経験をお持ちとのことですが、医務査定の評点にランダムプランテッドフォレストを適用することは容易でしょうか。それとも、何か課題と考えられることがあるでしょうか」というご質問をいただいております。



【松江】　ご質問ありがとうございます。私のところで答えます。

医務査定のところといいますと、いろいろなケースが考えられると思っていて、局所的に、例えば血圧の組み合わせによる高血圧のスコアリングを見直すために、RPF を使って、これら変数の交互作用を出してみる、というような使い方はあるのではないかと思うのです。けれども、年齢や性別の扱いをどうするのかということが結構難しくて、別々に変数として入れると、アルゴリズムがそれらばかり取ってしまいますし、オフセットのようなことも、ランダムフォレストだと難しいといったところがあり、適用の仕方は若干迷うところがあるかと思います。

いろいろな疾患のサンプルを全部入れてみて、大局的に見て、高血圧や肥満、どこかの疾患などで、おかしいところはないのか、探索的に使うなど、そのような使い方も、あり得るのではないかと思っております。

補足ですけれども、先ほどのアンダーライティング実務の説明だと、本当にルールベースできっちり決まるかのように言ってしまったのですけれども、項目だけではなくて、査定担当者の判断や、社医の深い判断で決まるようなところもあります。点数がシンプルに決まるというのは、あくまでも原則ベースの話でございまして、決してアンダーライティング自体がシンプルではないというところ、訂正いたします。失礼いたしました。

私からは以上です。

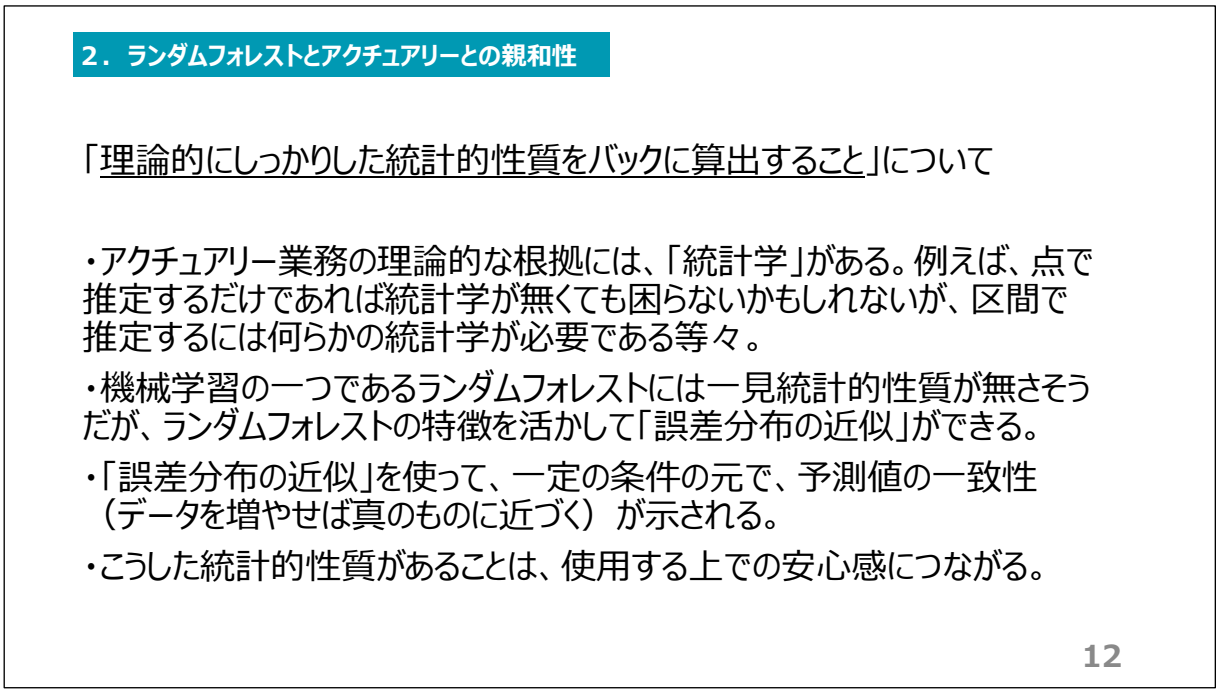
【藤田】　ありがとうございます。実際に、実務にランダムフォレストを活用された経験があるということで、また続くパネルディスカッションでも、そのような点に触れていただければと思います。ありがとうございます。

では、続いて、小田さんへのご質問としていただいております。こちらは、このセッションではランダムフォレストという、スペシフィックにモデルをご紹介しているというところで、やはり他の手法との比較ということにご興味がある方が多いようです。

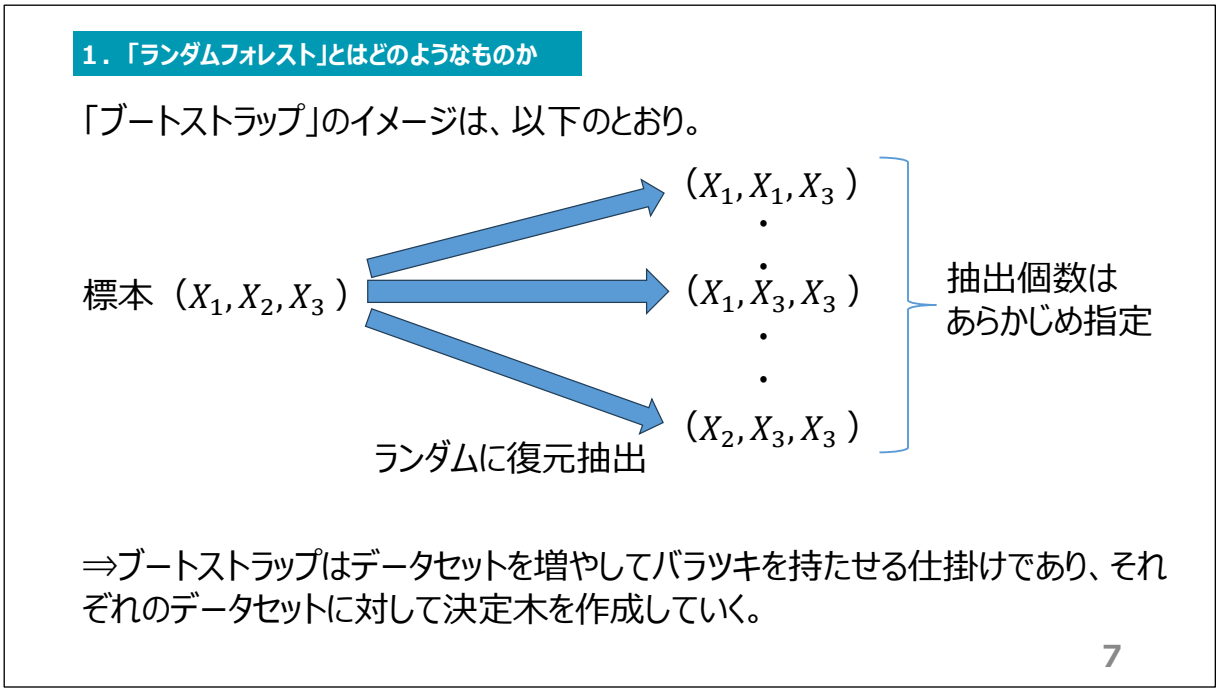
まず、「他の手法との比較において、ランダムフォレストは統計的性質のバックグラウンドがありとしてい

る一方で、勾配ブースティング、GBM は少ないとしているのは、なぜでしょうか」ということが 1 点。「また、結果の解釈の可能性も、そこまでランダムフォレストと GBM では大きな差異はないように感じますが、なぜ GBM の方が小さいとしているのでしょうか」ということです。よろしくお願いいたします。

【小田】　ご質問ありがとうございます。



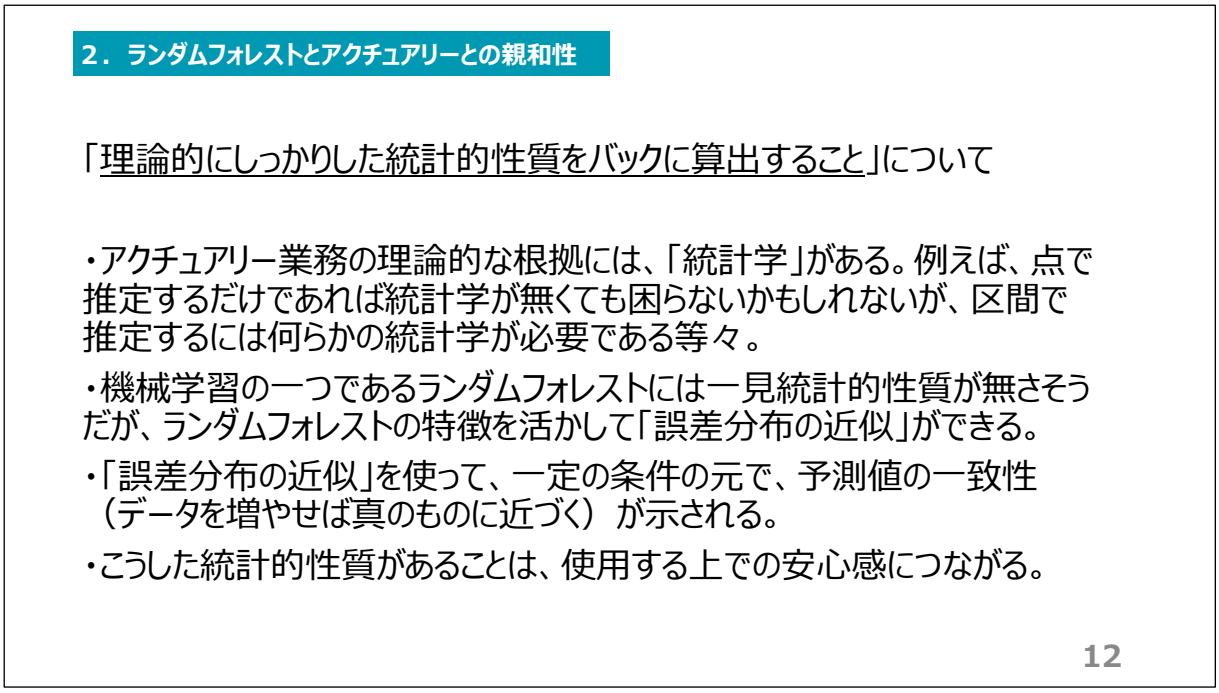
一つ目の「統計的性質」のところは、ここでも少しだけ触れているのですけれども、ランダムフォレストについては、この誤差分布の近似ができるというようになっています。「誤差分布の近似は、そもそも何ですか」ということなのですけれども、



ブートストラップというものを使って、このような形で、たくさんのデータセットを作ります。

例えば一番上のものは、X_1 と X_1 と X_3 なので、X_2 という標本は使われていないのです。使われていない X_2、はアウトオブバックと呼ばれるのですけれども、そのようなものに対して予測ができます。使われていないデータに対して予測をして、それを基に誤差分布を作成できるという仕組みが、ランダムフォレスト特有の性質としてあります。

そうした誤差分布の近似といったものが、GBM やニューラルネットワークのようなところで出来るかどうかは、私は存じ上げません。



更に言うと、誤差分布のところ、一致性ということで収束していきます。そのようなランダムフォレスト特有の統計的性質があるというところが、他との違いだと認識しています。

2．ランダムフォレストとアクチュアリーとの親和性

「算出結果が解釈しやすく説明しやすいこと」について

・ランダムフォレストに限らず、機械学習モデルの各特徴量の重要度合いや、予測結果を解釈するための技術の研究が進んでいる（Interpretable Machine Learning）

・ランダムフォレストの場合は、特徴量重要度と呼ばれる指標を出力する機能が、種々のプログラミング言語のパッケージに内蔵されていることが多い

・ Interpretable Machine Learningの進展により、従来解釈しにくいとされてきた機械学習モデルの解釈可能性が向上

・ただ、機械学習ではあるので、限界はある。

13

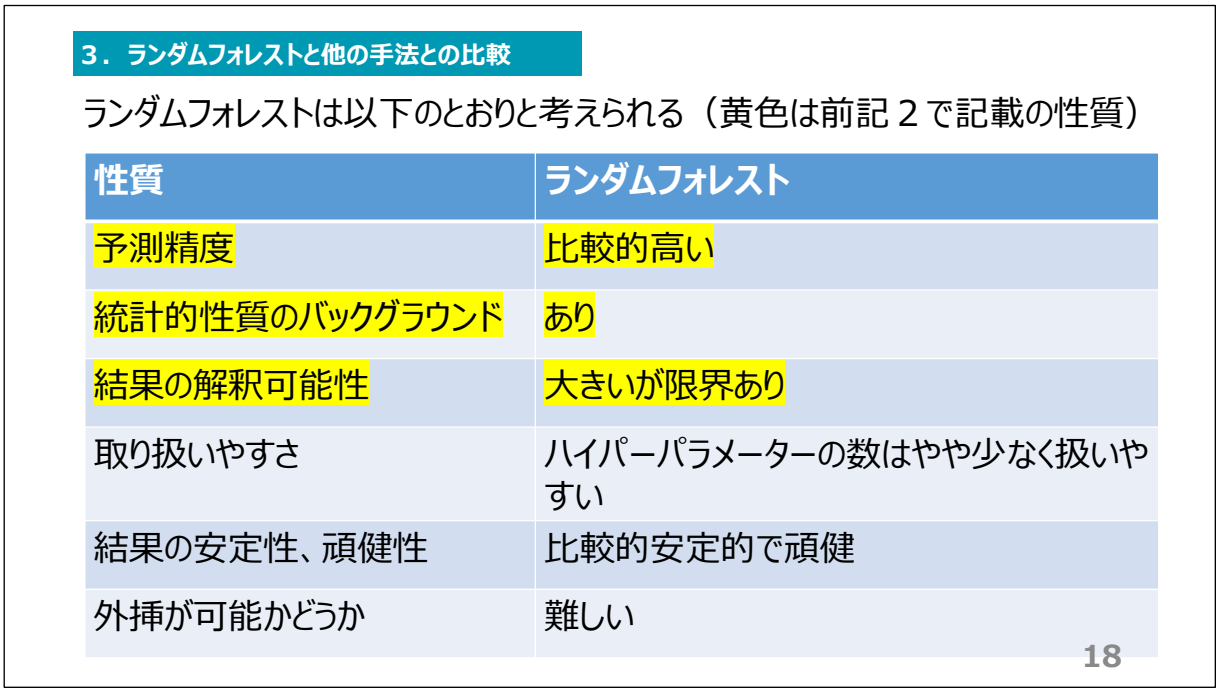
二つ目の「解釈可能性」のところは、ここで書いていますけれども、確かに、Interpretable Machine Learning、これ自体は、別にランダムフォレストだけに適用できるというものではないかと思います。ただ、二つ目にあるような形で、ランダムフォレストは色々なパッケージに内蔵されたり使いやすくなっているというところはあるかと思います。また、これで実際やっているような論文等も多く、そうした説明しやすさというところが、他と比べてあるのではないかと思います。

3．ランダムフォレストと他の手法との比較

勾配ブースティングやニューラルネットワークについては以下のとおりと考えられる。

性質	勾配ブースティング	ニューラルネットワーク
予測精度	高い	高い
統計的性質のバックグラウンド	少ない	少ない
結果の解釈可能性	小さい	小さい
取り扱いやすさ	ハイパーパラメーターの数が多く、ランダムフォレストと比較して取り扱いにくい	ハイパーパラメーターの数が多く、ランダムフォレストと比較して取り扱いにくい
結果の安定性、頑健性	ハイパーパラメーターの数が多く、調整も難しく、調整次第によっては過学習が起きてしまう。	データが少ないと不安定。また、勾配消失等、学習が停滞する可能性があり必ずしも安定的とはいえない
外挿が可能かどうか	難しい	可能

20



ランダムフォレストのところで、「結果の解釈可能性」に「大きいが限界あり」と書いてあるのですが、そのようなところで、GBM などより、いろいろ、そのようなパッケージがあって、準備がされていて、やりやすいが、限界はありますといったところを書いています。

お答えになったかどうかわかりませんが、以上となります。

【藤田】 前半の統計的な性質バックグラウンドというところでも、後半の解釈可能性も、どちらも一部関係していると思うのですけれども、モデルの原理そのものが、相対的に分かりやすいというところもあると思います。GBM もランダムフォレストも複数の決定木を作って予測値を算出していくのですけれども、その作り方にあたって、ランダムフォレストは独立に複数の木を並列的に作る一方で、GBM は一つ前の木に従属させる形で直列的に木を作るというところで、統計的性質のバックグラウンドのところだったり、原理的な解釈可能性が減る要因にもなるのではないかと思いました。

パネルディスカッション

6

まだ質問は残っているのですけれども、次のパネルディスカッションでもできるだけ触れられればと思います。続いて、パネルディスカッションに移りたいと思います。

では、セッション名がアクチュアリー実務におけるランダムフォレストの活用可能性ということなので、お三方の、ランダムフォレストをいろいろ使われて研究されているというご経験も踏まえて、いくつかの質問を、私から、させていただければと思います。

Q.
実際にランダムフォレストを使用してみて、その強みはどのように感じられましたか？

7

まず一つ目です。皆さんのプレゼンテーションの中にもあったと思うのですけれども、改めまして、実際にランダムフォレストを使用されて、どのようなところに強みを感じられたかということを、お聞きしたいです。

では、まず小田さん、いかがでしょうか。

【小田】 私からお話します。

先ほどの私からのお話でも少し触れましたけれども、やはり、一番、私的に大きいことは、比較的簡単に取り扱えるということです。いろいろなチューニングなどを、あまりしなくてよくて、それでいて、まあまあ良い結果が得られるというところ、これが大きいと思っています。

単に精度を高くするということだけであれば、他の手法も、いろいろ優れている面もあるのですけれども、やはりバランスがいいというところです。

あとは、先ほどご質問にもあったのですけれども、統計的性質です。機械学習にもかかわらず誤差分布の近似になるという統計的性質を持っているということは、私も知ったとき、かなり驚いたのですけれども、そのようなところが強みではないかと思っています。

一旦、私からは以上です。

【藤田】 ありがとうございます。

小田さんは、具体的には、どのようなツールを使ってランダムフォレストを実装されているのですか。

【小田】 私は、R を使ってやっていまして。

【藤田】 なるほど。先ほど「使いやすい」とおっしゃったのですけれども、機械学習モデルと言うと、それを作るための、ハイパーパラメータと呼ばれる、ユーザー側が指定する入力があると思いますが、そこで職人技が必要とされるのではないか、ということもあると思うのです。そのようなところも、R では、使いやすいのですか。

【小田】 そうですね、本当に、指定しなければいけないものは、かなり限られるといいますか、そもそもハイパーパラメータの数自体少ないですし、自動的とまでは行かないまでも、かなり容易な形で出来ます。そのような意味で、ランダムフォレストは初心者にも入りやすい手法ではないかと思っています。

【藤田】 ありがとうございます。

では、続いて、まだお時間に余裕があるので、松江さん、何かコメントはありますか。

【松江】 小田さんのご意見と重複するところはあるのですけれども、やはり、データドリブンなところがあります。GLM はフィーチャーエンジニアリングなどを行う上で、どうしても恣意的なところが入ったり、ある程度ドメイン知識がないと、いいものができないといったところがあると思うのです。また、交互作用をあらかじめ把握する必要があるのですけれども。ランダムフォレストは、そのようなことがなく、データドリブンでモデルを作ってくれるといったところが、手軽ではあります。

ただ、こちらはデメリットでもあって、やはり、いいモデルはデータだけだと作れないと思っています。当然、データそのもののバイアスやノイズもありますし、学習する際のバリアンスもあります。ドメインについてのアプリアルな知識と、データから経験的に学べるものを、いかにバランスよく組み合わせるかとい

ったところが重要な中、ランダムフォレストは、経験的なものだけになるので、そこは一つのデメリットになるのではないかと思っています。

そのデメリットに対処するには、例えば、ドメイン知識を入れた GLM とランダムフォレストのブースティングがあります。ニューラルネットのバージョンとして CANN というものがあって、ランダムフォレストでもできないかといったところが考えられます。他には、ハイブリッドモデルがあり、例えばランダムフォレストを一般化したもので、線形回帰モデルの係数などを推定する、しかも、その係数は全範囲で一律ではなくて、局所的な違いも把握するというようなアルゴリズムがありまして、そのように、他の GLM やドメイン知識を入れられるようなシンプルなモデルとの組み合わせでランダムフォレストを使うと、そのような弱みも解消できるのではないかと思っています。

私からは以上です。

【藤田】 ありがとうございます。

そうですよね、データドリブンとドメイン知識ドリブンは、バランスよく組み合わせることが大事だと思いますけれども、ランダムフォレストは、そもそも、ピュアにデータドリブンで見たい場合は、非常に使いやすく、モデルのフィッティングもとてもいいということだと思います。

では、続いて田中さん、いかがですか。

【田中】 生存時間分析の文脈ですが、ランダム・サバイバル・フォレストは、予測精度において従来のモデル対比で良い結果が期待でき、解釈可能性という観点でも助けとなるツールが充実しておりますので、予測精度・解釈可能性の観点からバランスが取れている手法だと考えます。

また手法のコンセプト自体も、いくつも仮想的なサンプルを生成した上で木を構築して平均を取るという、分かりやすいものだと思います。

【藤田】 ありがとうございます。

Q.

ランダムフォレストの実務への具体的な応用について、どのような可能性が考えられますか？

8

Q.

実務に応用する際に、特に注意すべき点や課題は何だと思いますか？

9

それでは、続いてのご質問です。まさに、このセッション名のメインでもある、こちらです。残り二つ用意しているのですけれども、二つの質問は関連しているので、同時にご質問させてください。

まずは、ランダムフォレストの実務への具体的な応用について、どのような可能性が考えられるかということで、松江さんはプレゼンテーションの中で、一部言及していただきましたけれども、改めて皆さんにお伺いしたいです。また、その際に、どのような注意点や課題があるかというところも教えていただけますと幸いです。

まず田中さんから、よろしいですか。順番を変えてみましょう。

【田中】　ありがとうございます。具体的な応用としては、直にモデルを使うことも一つですが、線形モデルの補完的な形で、どのような変数を線形モデルに入れていきましょうかと考えるときに、例えば、非線形な効果や交互作用の効果を入れるための指針として、ランダムフォレストを活用できるのではないか、と考えています。

あともう一つは、先ほどのパートで、小田さんからも言及がありましたが、予測誤差が評価指標ということで、機械学習の手法ですと、どうしても「期待値的な評価が真の値に近ければよい」という意識になりがちですが、アクチュアリーとしては、評価がどれぐらいぶれ得るのかも、興味があると分野と考えております。例えばアクチュアリーとしては、「プロセス誤差やパラメータ誤差はどの程度の水準となっているのか」を確認したいというニーズもあると思います。

例えばパラメータ誤差が大きいのであれば、データを増やす、パラメータ推定の方法を工夫するなど、いろいろ方策があると思います。

そのような誤差分解を考えたときに、実はランダムフォレストは、その手法の特性上、効率的に誤差分解ができるのではないかということで、まさにアクチュアリー会のデータサイエンスワーキンググループで、研究が進められております。ランダムフォレストについて、そのような性質が明らかになってくると、予測精度の向上や、結果の解釈可能性の向上にも繋がりますので、大いにポテンシャルのある手法だと思っています。

【藤田】　ありがとうございます。

後半、予測誤差分解、予測誤差の話に触れていただいたのですけれども、そもそも、このチームは、汎用的な予測誤差分解手法という、データサイエンス関連基礎調査 WG が『アクチュアリージャーナル』で別途発行しているものがあるのですけれども、そちらをランダムフォレストに適用した場合に、ランダムフォレスト特有の良い性質が活用できるのではないかというところで、研究がスタートしたという経緯があります。

アクチュアリーは、いろいろな予測モデルを、活用していくと思うのですけれども、その際に、単なる予測精度だけではなくて、リスク、不確実性に関しても、モデル選択の一つのメトリックとして活用するということで、そのような予測誤差分解の話も非常に重要になるのではないかと思います。

田中さん、ありがとうございます。それでは、続いて松江さん、いかがでしょうか。

【松江】　確かにアクチュアリーの数理モデリングなど、目的が値の正確な推定をすることとなると、他のモデルの方がいいかもしれないのですけれども、例えば、医務査定など実務のルールについては、閾値で区切るのですよね。「血圧がここを超えたらプラス何点にしよう」というように。ランダムフォレストは、差を捉えてセグメント分けをすることが得意と考えています。アルゴリズムとしても、全部同じ値、平均値を取るものが一番悪いモデルで、そこから、セグメントを区切ってセグメントごとの平均値を採用するという感じで、差が最も大きくなるように分割するのがランダムフォレストというアルゴリズムでして、差を捉えるという用途には結構強いアルゴリズムになるのではないかと思っています。

だから、アクチュアリアルな分析でも、例えば、リスク管理のアクチュアリーが解約率の分析をして、事務保全部門に「抑止となる施策を、この契約のセグメントにしてほしい」という提案を行う場合は、ランダムフォレストの分割によって、施策が必要なセグメントを特定する使い方が考えられるのではないかと。逆に、GLM のような連続的なモデルを使っていると、おしなべた線形な効果は捉えられても、ローカルな解約率

が急増するところなどは、なかなか捉えづらいです。かといって、ニューラルネットのような複雑すぎるものになると、相当データ量がないと強みが生かせない。ランダムフォレストは予測精度でパーフェクトというわけではないのですけれども、手軽で、そのわりにいいパフォーマンスを発揮してくれるようなモデルなのではないかと。

もう一つ、先ほど言ったハイブリッドモデルもあるのですけれども、一般化ランダムフォレスト、再びご説明しますけれども、こちら、モーメント条件、つまりXの期待値とXの二乗の期待値など、そのようなものを使った方程式で表現できる統計量が推定できます。普通の機械学習モデルは条件付期待値を推定するアルゴリズムなのですけれども、もっといろいろな統計量、例えば、バリュー・アット・リスクのパーセンタイル点や条件付分散、あとは線形回帰した場合の係数なども、推定することができます。

アクチュアリーのモデリングなどでも、例えば、解約率の景気感応度の係数を設定したり、あとは、健康改善系の商品だったら改善効果の係数を設定したり、そのような係数が、どのような場合でも一律ではなく、様々な変数の値によって異なるだろうと考えられる場合に、一般化ランダムフォレストを使って、「この属性では感応度が少し高くて、この属性の場合は低い」など、そのような分析にも使えたり、活用可能性は広いのではないかと、私は考えています。

以上です。

【藤田】 ありがとうございます。ランダムフォレストの派生形として、おっしゃられたようなモーメント、分散や他の統計量、例えばバリュー・アット・リスクまで予測するモデルがあります。松江さんが、先ほど「医的診査などで使われた」とおっしゃっていたのですけれども、ランダムフォレストを選択された理由の一つには、そのようなところにあるということなのですか。おっしゃっていただいた、使いやすさや、ランダムフォレストの性質やニューラルネットワークとGLMなどの比較に加えて。

【松江】 そうですね、GLMなどを構築して、そこから例えばEDAや主効果を可視化してもあまり意味がなく、ランダムフォレストを構築することで、特に保険引受リスクが低くなるセグメントを特定したりできました。ニューラルネットを構築するとなると、条件体サンプル数が限られていたため全然いいものができませんでした。

【藤田】 ありがとうございます。

では、小田さん、最後にいかがですか。せっかくなので、もし可能であれば、スライドでピックアップできなかった外挿のところ。例えば金利の前提条件の外挿など、多分、アクチュアリーの方は興味があると思うのですけれども。せっかく各モデルの比較のところで外挿というところにも言及していただいたので、もし可能であれば、そこにも触れつつ、何か実務への応用、具体的な可能性、課題にお答えいただけると幸いです。

【小田】 ありがとうございます。

実務への応用可能性のところは、私の説明でもしたのですけれども、やはり、いきなり行政への認可や商品適用ということは、なかなか難しいというところがあって、既存の方法との兼ね合いですね。やはり、そこが大事だということです。既存の方法は、長年使ってきて安心感がありますので。

とはいえ、どのように応用できるかと言うと、例えば、予測モデルに必要なもの、これについてランダムフォレストは使えるということがありますので、リスクの評価、信用リスクの評価といったものに使用できるのではないかと思われます。海外の文献にも、そのようなものが一部あったりしますので、説明可能性さえきちんとできれば、そのようなところも使用できるのではないかと思います。

外挿については触れられずにすみません。

【藤田】 ありがとうございます。

以上、パネルディスカッションとさせていただければと思います。

Slido で、岩沢先生から補足いただいているのですけれども。先ほど、GBM とランダムフォレストの比較のところで、実は、ブースティング木でも OOB という概念はあります。よって、同様な性質を持つ可能性はあると。でも、私が先ほど申し上げたように、「木を直列的に前の木に従属する形で作成するというところで、複雑になり統計的性質の解析は難しいのではないか」ということで補足をいただきました。ありがとうございます。

それでは、お時間になりましたので、このセッションは終了とさせていただきますが、ランダムフォレストの可能性、課題を含めて、いろいろご紹介させていただきました。皆さんも、ランダムフォレストの可能性はご認識いただけたかと思いますが、引き続き、ランダムフォレストが実務に根ざす形を目指して、研究を進めてまいりたいと思います。ご清聴ありがとうございました。

以上、ご清聴ありがとうございました。