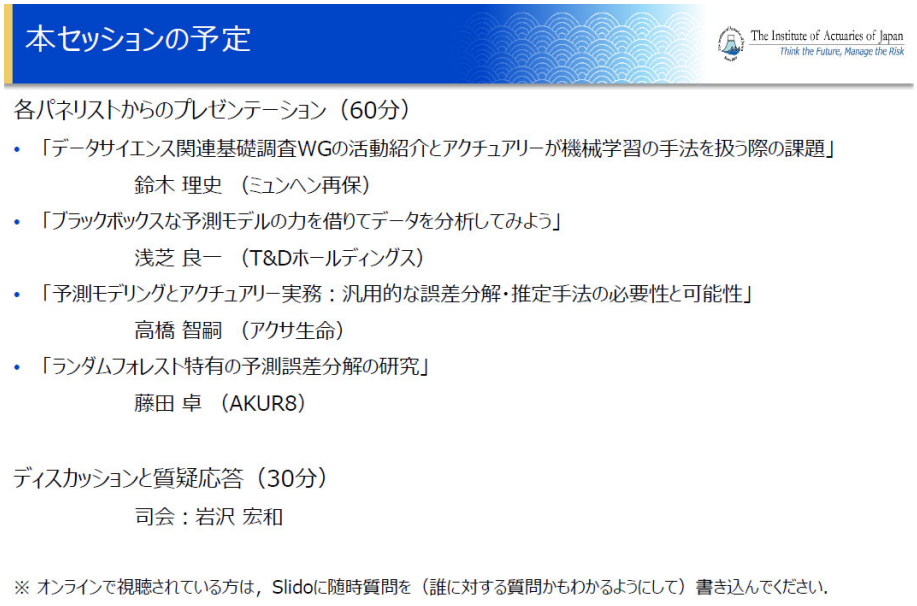
アクチュアリー×データサイエンス　どう使う？　どう学ぶ？
〈データサイエンス基礎調査 WG〉

早稲田大学大学院（司会）　岩沢 宏和 君
T&D ホールディングス　浅芝 良一 君
ミュンヘン再保　鈴木 理史 君
アクサ生命　高橋 智嗣 君
AKUR8　藤田 卓 君

岩沢　データサイエンス関連基礎調査ワーキングのメンバーによるパネルディスカッション、「アクチュアリー×データサイエンス　どう使う？　どう学ぶ？」を開始いたします。本セッションの司会を担当いたします、岩沢です。どうぞよろしくお願いいたします。



本セッションで予定している内容は、スライドをご覧ください。

最初に 60 分ほど、４人の方々にプレゼンテーションしていただいて、そのあとに 30 分ほど、同じ４人の方にディスカッションしていただく予定をしております。これまで 4 人の方が準備されてきた内容を拝見していますが、盛りだくさんで密度が濃いです。時間はタイトです。ですから、私はほとんど話しません。また、最後にご質問を受ける時間はありますけれども、十分取れないかもしれません。ご了解いただければと思います。

では、早速プレゼンに入りたいと思います。最初はミュンヘン再保険の鈴木さんから、「データサイエンス関連基礎調査ＷＧの活動紹介とアクチュアリーが機械学習の手法を扱う際の課題」ということで、お願いいたします。

１．データサイエンス関連基礎調査WGについて

- 2019年、会員のデータサイエンスに関する知識とスキル向上を目的とし、**データサイエンス関連基礎調査WG**が設立された。
- ＷＧは以下のチームに分かれて活動

■ ＩＭＬチーム　：　機械学習モデルの汎用的な解釈手法を研究
■ リスク研究チームＡ班　：　汎用的に使用可能な予測誤差分解の手法を研究
■ リスク研究チームＢ班　：　ランダムフォレストによる予測誤差推定・分解の研究
■ 試験問題研究チーム　：　SOAのExam PA(Predictive Analytics)の試験問題の翻訳を会員へ共有する活動を実施
■ 情報共有チーム　：　機械学習に関する情報を会員へ共有する活動を実施

3

鈴木　最初のプレゼンテーションを担当します、ミュンヘン再保険会社の鈴木理史と申します。どうぞよろしくお願いいたします。私からは、本セッションのイントロダクションとして、我々データサイエンス関連基礎調査ＷＧとは何か、何をしてきたか、何をしているかということと、アクチュアリーが機械学習の手法を扱う際の課題は何かについて、お話をさせていただきます。

我々データサイエンス関連基礎調査ＷＧは、2019 年に本会会員のデータサイエンスに関する知識とスキル向上を目的として設立されました。3 ページに示している 4 つのチームに分かれて活動しております。

１．データサイエンス関連基礎調査WGについて

実施時期	内容	アクチュアリージャーナル
2017年7月	**第３回例会　（会員報告）** 「将来を見据えたアクチュアリー教育像　～ＩＡＡシラバス改定の議論～」	101号
2017年11月	**データサイエンスに関する会員アンケートを実施**	104号
2018年8月	**「若手ディスカッション」を組成（のちのWGの原型）**	
2018年12月	**第７回例会　（休日シンポジウム）**　･･･早稲田大学と共催 ■特別講演　「会計に関するデータサイエンス」 ■「データサイエンスの技術」というテーマで若手ディスカッションから４つのプレゼンとパネルディスカッションを実施	106号
2019年9月	**データサイエンス関連基礎調査WG 設置**	
2020年3月	**予測モデリングにおける誤差評価に関する研究報告** データサイエンス関連基礎調査WG　リスク研究チーム	110号
2020年7月	**SOAのExam PA: Predictive Analyticsの翻訳** （2018年12月、2019年6月のSOA試験） データサイエンス関連基礎調査WG　試験問題研究チーム	112号
2020年9月	**報告　Rを用いたデータの可視化技術 解説書** データサイエンス関連基礎調査WG　可視化研究チーム	112号
2020年12月	**第4回例会** 「予測モデリングにおける誤差評価に関する研究」 ※アクチュアリージャーナル110号の解説	

4

4 ページでは、ワーキンググループが設立される前後のアクチュアリー会のデータサイエンスに関する活動をまとめております。2017 年にＩＡＡ（国際アクチュアリー会）のシラバスが改定されました。そこで、「データとシステム」という項目が新しく追加されております。これを踏まえて日本アクチュアリー会で

は、データサイエンスに関する会員アンケートを実施し、現状を調査。「若手ディスカッション」という後のワーキンググループの原型になる活動を開始いたしました。その後、2019 年に正式なワーキンググループとして立ち上げ、翌年には『アクチュアリージャーナル』に三つの記事を投稿しております。

1．データサイエンス関連基礎調査WGについて

実施時期	内容	アクチュアリージャーナル
2022年6月	**Actuarial Colloquia 2022（オンライン）** 論文"A New Framework of Prediction Error Decomposition for the Machine Learning Era"を発表 データサイエンス関連基礎調査WG リスク研究チームA班	
2022年9月	**Convention A （オンライン）** WGの研究成果の発表を含むデータサイエンスに関するセッションを実施 日本アクチュアリー会事務局、ASTIN関連研、データサイエンス関連基礎調査WG	
2022年12月～2023年3月	**ムーンライトセミナー** Interpretable Machine Learningというテーマで実施 データサイエンス関連基礎調査WG IMLチーム	
2023年5月	**ICA2023（シドニー）** 論文"Random Forest Model with Prediction Error Decomposition Function"を発表 データサイエンス関連基礎調査WG リスク研究チームB班	
2023年6月	**SOAのExam PA: Predictive Analyticsの翻訳** （2019年12月、2020年6月のSOA試験） データサイエンス関連基礎調査WG 試験問題研究チーム	123号
2023年9月	**論文 予測モデリングとアクチュアリー実務：汎用的な誤差分解・推定手法の必要性と可能性** データサイエンス関連基礎調査WG リスク研究チームA班	124号

▶ 5 (*)赤枠で囲んだものが本日の各チームからの発表に特に関係の深いもの。

5 ページは、直近に行ったワーキンググループの活動を示しております。特に、赤枠で囲まれているＩＭＬチームによるムーンライトセミナー、リスク研究チームＢ班によるＩＣＡ2023 でのシドニーでの論文発表、そして、直近の『アクチュアリージャーナル』124 号でのリスク研究チームＡ班における論文の掲載の３つは、本日のプレゼンテーションのベースとなっております。

我々の活動は、単にデータサイエンスを使ってみた、やってみたというような簡単なものではございません。アクチュアリーが、我々の専門領域の中でデータサイエンスをどのように使っていくべきか、使うにあたってどのようなことが課題なのかという問題意識を踏まえて、アクチュアリーらしい機械学習手法の使い方を念頭に置いて活動しております。6 ページより、その問題意識の一端をご説明させていただきます。

2．統計解析におけるパラダイムシフト

新しい統計解析（多くの場合機械学習も含まれる）において、アクチュアリーが慣れ親しんだ旧来の統計解析から大きなパラダイムシフトが起きている。

旧来の統計解析	新しい統計解析
モデルの正しさを求める。	**予測の正しさ**を求める。
モデルが正しいことが仮定ないし目標。	モデルが正しいことは仮定もされないし目指されもしない。
どのデータ（特徴量）を説明変数として使うかは、統計解析をはじめる段階ですでに決定している。	どれを使ったらよいかわからないデータ（特徴量）がたくさんあり、統計解析をはじめる段階では説明変数は特定されていない。

このパラダイムシフトの中で、予測精度の高い予測モデルを得たが、

アクチュアリーの実務に必要な「信頼性の評価」に課題

⇨ 本WGの活動で実現を目指している

▶ 6 (*)2017年度　第8回例会　第1部「プレディクティブ・モデリングの概要と最新動向」（岩沢宏和氏）を参照（アクチュアリージャーナル104号に講演録が掲載）

6 ページは、本日のセッションのモデレーターであります岩沢宏和先生が、2017 年にアクチュアリー会の例会でプレゼンテーションされたものをまとめたものです。機械学習も含まれる新しい統計解析は、単に古いモデルより良いモデルになった、あるいは高度な数学を使うようになったというような単純な進化ではなくて、特に目的の中で一種のパラダイムシフトが起きているということを念頭に置く必要があると考えています。

旧来の統計解析の手法では、現実世界を適切に反映した正しいモデルを得ることを目的としております。一方で新しい統計解析では、将来の予測値を得ることを目的としています。その目的の中で、理論的な背景も大きく異なります。旧来の統計解析では、現実を正しく表したモデルを得ることを仮定ないしは目的とします。後でもご説明いたしますが、皆さんもおなじみの検定や区間推定などは、モデルが正しいことを前提としたものになります。しかし、新しい統計解析の中では、そのモデルが現実を表した正しいモデルであることは、仮定もしていないし、目指されてもいない。あくまでも将来の予測の正しさを求めています。

また、説明変数の候補に対して想定している状況も大きく異なります。旧来の統計解析では、どのデータを説明変数として使うかは、既に決定されている場合が多いです。一方で、新しい統計解析で想定される状況は、統計解析を始める段階では、説明変数は特定されていない、すなわちどれを使えばいいかは分からないけれども、データはたくさんあるというような状況です。ビッグデータと呼ばれるような大容量のデータを扱う昨今の状況にふさわしい変化だと思いますけれども、このような状況の違いがあります。これらのパラダイムシフトの中で、予測精度の高いモデルを得たのですが、捨て置かれていることがあります。それは、アクチュアリーの実務に必要な「信頼性の評価」です。これを解決することを目指して、本ワーキンググループでは活動しております。

2．旧来の統計解析の手法と機械学習の手法の違い

これらの違いによりアクチュアリー実務に応用する上で課題。（後述）

	旧来の統計解析の手法	機械学習の手法
確率分布の仮定	モデルの中に目的変数が従う確率分布が組み込まれている。 例：線形モデルでは正規分布	モデルの中で目的変数に確率分布を仮定しているわけではない。
変数選択	モデルを当てはめる段階で変数をうまく選ぶことが重要。	モデル構築の前に必ずしも変数を厳選する必要はなく、モデル構築の中で自動的に変数が選択される手法も多い。
ハイパーパラメータの有無	モデルを微調整するために任意に指定できる要素はほとんどない。	モデルを微調整するために任意に指定できるパラメータ（ハイパーパラメータ）があり、自動的な試行の繰り返しなどによってチューニングが行える。
アウトプット	不確実性を踏まえた評価。（区間推定、有意水準）	精度に重きを置いた予測。

▶ 7

7 ページをご覧ください。このページでは、先ほどご説明した目的や仮定の違いに裏づけられた、細かな手法の違いをご説明いたします。表形式にまとめていますが、後ほど具体例で説明いたしますので、現時点でピンとこなくてもご安心ください。一番大きな違いは、確率分布の仮定です。旧来の統計解析の中では、モデルの中に目的変数が従う確率分布が組み込まれております。例えば、線形モデルでは正規分布が仮定されております。一方で、機械学習の手法の中では、モデルの中で目的変数に確率分布を仮定しているわけではありません。この違いは、後に不確実性の評価に大きく関わってきます。

また、先ほどお話ししました説明変数の選択について、旧来の統計解析では、ここをうまく選択することが極めて重要でした。一方で機械学習の手法では、モデル構築の際に、とにかく入れられるだけのデータをありったけ説明変数の候補として突っ込んで、モデルの構築の中でアルゴリズムによって自動的に選択されるという手法が多くなっております。これも後でお話しいたします。

また、特徴的なものは、ハイパーパラメータです。ハイパーパラメータとは何なのかは、このあとにまたご説明いたしますが、旧来の統計解析では、モデルを微調整するために任意に指定できる要素はほとんどありませんでしたが、機械学習の手法では、ハイパーパラメータというチューニングできるパラメータがあります。

また、最後にアウトプットですが、これがアクチュアリー実務を考えた上で一番大きな違いです。旧来の統計解析では、区間推定や有意水準などの不確実性を踏まえた評価をアウトプットとして出しましたが、機械学習においては、精度に重きを置いた予測を点として出すという違いがございます。

３．代表的な機械学習の手法（赤字は本WGの発表で使用）

- 回帰系
  - Ridge回帰
  - Lasso
  - サポートベクトル回帰
- 決定木系
  - 回帰木
  - ランダムフォレスト
  - XGBoost
  - LightGBM
- ニューラルネットワーク系
  - パーセプトロン
  - 3層ニューラルネットワーク
  - 畳み込みニューラルネットワーク
  - 再帰型ニューラルネットワーク

本プレゼンテーションでは
この２つの手法の概要および
の旧来の統計解析手法との違いを説明。

8 (*) 教師あり学習の回帰のみを掲載している。

８ページからは、今までお話しした違いについて、具体例を用いて説明いたします。本日のプレゼンテーションでは、赤字で表示した機械学習手法が登場しますが、このうち私のプレゼンテーションでは、サポートベクトル回帰とランダムフォレストについて説明しまして、旧来の統計解析手法との違いをご説明させていただきます。

３．代表的な機械学習の手法　サポートベクトル回帰

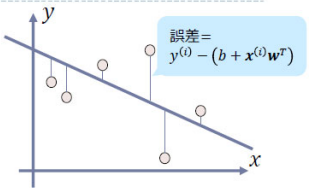
線形回帰

誤差の２乗を最小化するb,wを求める。

切片　回帰係数

$$\sum_{i=0}^{n}\left(y^{(i)}-\left(b+\boldsymbol{x}^{(i)}\boldsymbol{w}^{T}\right)\right)^{2}$$

誤差の2乗



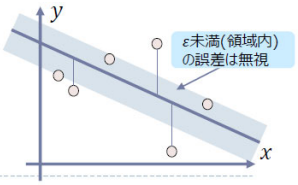
サポートベクトル回帰

下記の式（損失関数）を最小化するb,wを決定する。その際、
誤差の絶対値がε以下の時の誤差を無視（ノイズの影響を受けづらくなる）
正則化項により特徴量の重みwが大きくなりすぎないようにする（過学習を防ぐ）

誤差と正則化項のバランス　無視する誤差の閾値

$$C\sum_{i=0}^{n}\max\left\{0,\left|y^{(i)}-\left(b+\boldsymbol{x}^{(i)}\boldsymbol{w}^{T}\right)\right|-\varepsilon\right\}+\frac{1}{2}\|\boldsymbol{w}\|^{2}$$

ε許容誤差　正則化項



9

まず、サポートベクトル回帰です。サポートベクトル回帰の説明のために、まずは線形回帰をベースに説明いたします。アクチュアリー試験の数学の中で皆さんも勉強された、もしくは今でも勉強していると思いますが、誤差の２乗を最小化するｂとｗを求めるものになります。なお、機械学習手法で使われる記号に合わせるためにｂとｗにしていますけれども、アクチュアリー試験ではαとβと呼んでいたものだと思っていただければと思います。

一方でサポートベクトル回帰は、9 ページの下の方に書いてある式を最小化する b と w を決定します。この式を「損失関数」といいます。損失関数は、機械学習においてパラメータを求める際に最小化する関数の一般的な名称で、このあとも何度も出てきますので、ご留意ください。

この損失関数の式の意味をご説明しますと、C が掛かっている部分は、誤差が ε 以下のときにその誤差を無視するということを意味します。この図で言うと、グレーの部分の誤差は無視して、それより離れている部分だけを評価します。また、「正則化項」というものがあります。これは、特徴量の重み、皆様が慣れている言葉では回帰係数の β の大きさを、罰則項として入れているものです。この意味するところは、特定の回帰係数が大きくなりづらくなって、過学習、すなわちオーバーフィッティングが防がれるという効果があります。

３．代表的な機械学習の手法　サポートベクトル回帰

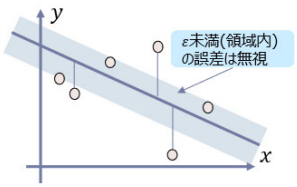
εを(無視する誤差の幅)を大きくすればノイズの影響は減らせるが、
少ないサンプルの誤差から回帰係数wが決定されることになる。
Cを大きく(正則化項の影響を小さく)すれば、
訓練データにはフィットするが過学習は起きやすくなる。

⇨ εやCは**ハイパーパラメータ**と呼ばれ、モデルに外部から指定する値。
上記のバランスを取りながら予測精度が高まるようにチューニングする。

誤差と正則化項のバランス　　無視する誤差の閾値

$$C\sum_{i=0}^{n}\max\{0,|y^{(i)}-(b+\boldsymbol{x}^{(i)}\boldsymbol{w}^{T})|-\varepsilon\}+\frac{1}{2}\|\boldsymbol{w}\|^{2}$$

ε許容誤差　　正則化項

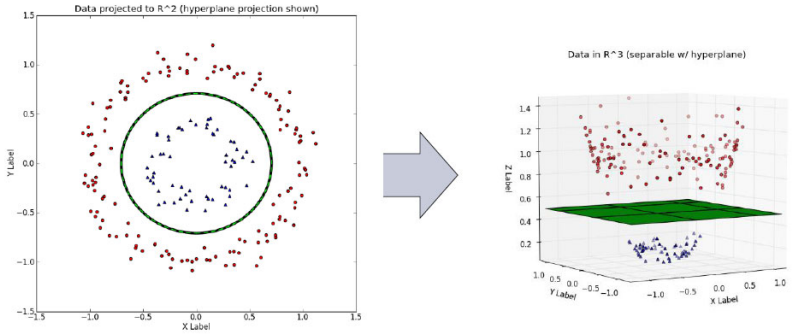


10

この ε と C をどのように設定するかで、何となく結果が変わるということは、皆さんも想像できると思います。これをどのように決めるかですが、適当に決めます。いやいや、「適当に決める」と言うと語弊があるので、訂正いたしますが、モデルの予測精度が高まるように、試行錯誤を繰り返してチューニングするというものです。このようなパラメータを、ハイパーパラメータといいます。一意に決めるようなパラメータではなくて、予測精度が上がることを目指して、試行錯誤の中でチューニングしていく。客観性や再現可能性を重視するアクチュアリーの実務からすると、少し気持ち悪いところではありますが、予測精度を高めることを目的としているので、このようなことも許されるという特徴があるのが機械学習になります。

３．代表的な機械学習の手法　サポートベクトル回帰

非線形なモデルが構築できるように、サポートベクトル回帰はカーネル関数を用いた写像（非線形な変数への変数変換）と一緒に用いられることが多い。
計算が困難な場合が多い写像後の関数をそのまま計算するのを避け、比較的計算しやすいその内積（カーネル）のみを計算する手法をカーネルトリックと呼ぶ。
（カーネルについてはAppendix参照）



11 画像引用元：https://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html

11 ページでは、カーネルの話をします。少し難しい話で、本日のプレゼンテーションの本論ではないので、何となく聞いていただければと思いますが、今までお話ししたサポートベクトル回帰は、所詮、線形回帰ですので、実際のところ、それほど予測精度が良くありません。ですから、非線形なモデルが構築できるように、カーネル関数を用いた非線形な変数への変数変換と一緒に用いられることが多くなっています。

カーネルの詳細については割愛させていただきますけれども、変数変換のイメージをざっくりとご説明いたします。例えば下の図は、回帰ではなく、サポートベクトルマシンという分類の図ですが、左側の円のように線を引きたいときに、この円のような線を引くアルゴリズムはなかなか複雑になりますが、これをＺ軸方向に引き延ばして、右側のように線形の平面で切ろうというものになります。実はカーネル自体もハイパーパラメータで、機械学習の手法は、かなり柔軟性を持ちながらモデルを構築できるということは、皆さんも何となくお分かりいただけたと思います。

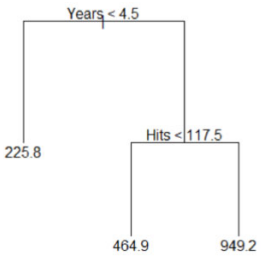
3．代表的な機械学習の手法　決定木

決定木とは、樹形図（木）を用いて目的変数を予測する手法。
回帰と分類の双方に使用可能。回帰を行う場合、回帰木と呼ばれる。
樹形図をたどればよいので、**結果の解釈が非常に容易。**

例：メジャーリーグ打者の成績と年収のHitters データ
（統計ソフトRのISLRパッケージの組込データ）

目的変数：1987 年の年棒（千ドル）（Salary）
特徴量1：メジャーリーグでのプレー年数（Year）
特徴量2：1986年のヒット本数（Hits）

[右の回帰木の読み方]
- Years が4.5 より小さければ225.8（千ドル）と予測
- Years が4.5 より大きければ、Hits の値を見て、
- Hits が117.5 より小さければ464.9 （千ドル）、
- Hits が117.5 より大きければ949.2 （千ドル）と予測



解釈は容易だが、**予測精度はあまり高くならない**ことが多い。

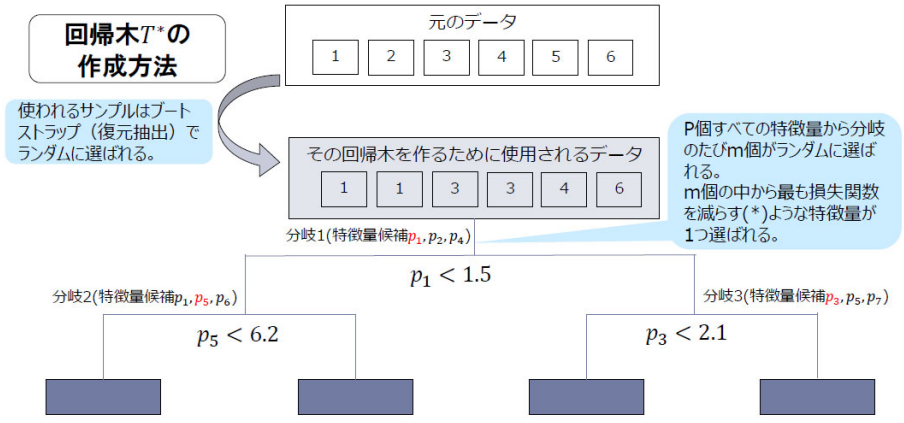
12

続いて、12 ページからランダムフォレストをご説明いたします。ランダムフォレストの説明の前に、ベースとなる決定木をご説明いたします。決定木とは、樹形図を用いて目的変数を予測する手法で、回帰と分類の両方に使用可能です。回帰を行う場合は「回帰木」と呼ばれます。樹形図をたどればよいので、結果の解釈は非常に容易です。

例として、メジャーリーグの打者の成績と年俸のデータを、回帰木を示しております。このデータからご説明しますと、メジャーリーグのプレー年数が 4.5 年より低ければ、年俸は 225.8 千ドル。4.5 年より長くヒット数が 117.5 より低ければ年俸は 464.9 千ドル、高ければ年俸は 949.2 千ドルということで、予測のパターンとしては三つしかございません。一見、あまり精度が良くないモデルであることは、お分かりいただけると思います。

3．代表的な機械学習の手法　ランダムフォレスト

ランダムフォレストは、回帰木を**ランダムに**多数（B個）生成し、
平均をとることで予測する手法。

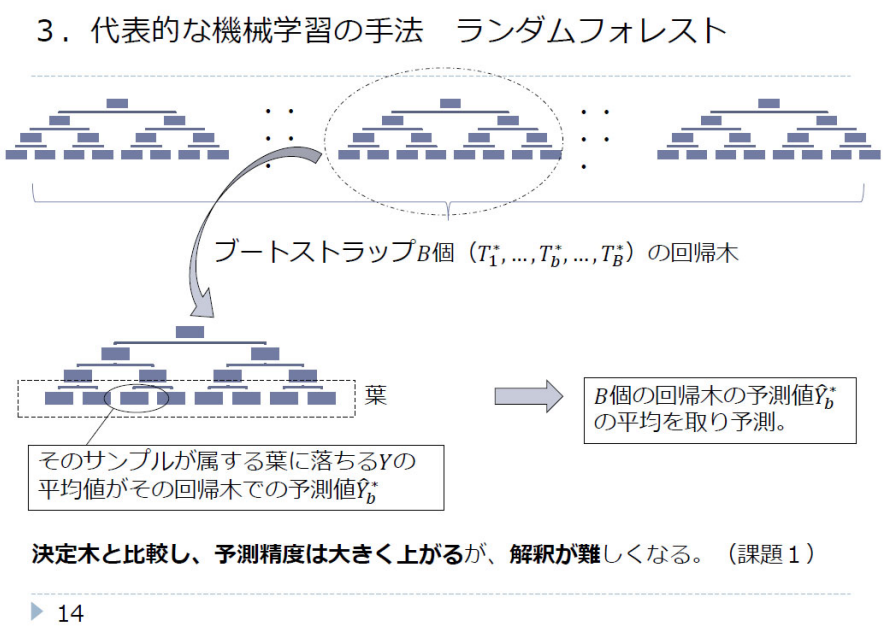


13 (*)分割後の領域を$R1$と$R2$、Y_i^*をブートストラップサンプル、$\bar{Y}_{R1}^*$と$\bar{Y}_{R2}^*$をその領域内の平均として
$\sum_{i \in R_1}(Y_i^* - \bar{Y}_{R1}^*)^2 + \sum_{i \in R_2}(Y_i^* - \bar{Y}_{R2}^*)^2$ を最小化するように分岐するアルゴリズムが使われることが多い。

ランダムフォレストは、この回帰木をランダムに多数（Ｂ個）生成し、平均を取る手法です。ランダムということがポイントで、一つ一つの回帰木の作成に使用するサンプルもランダム、各分岐で説明変数の候補となる特徴量の選択もランダムになります。

以下の例では、サンプルが１から６までの６個あるとします。この中から、ブートストラップ法と呼ばれる、同じサンプルを何度も取る復元抽出でサンプルを選びます。同じサンプルを何度も取れるので、１が２回、３が２回、また、２と５が使われていなかったりします。このように得られたサンプルを用いて回帰木を作ります。

最初の分岐では、特徴量の候補の数はＰ個が全部だとして、その中からｍ個だけランダムに選ばれます。この例ではp_1、p_2、p_4の三つです。この三つの中から、最も損失関数を減らす特徴量、この例ではp_1を使って分岐をします。この特徴量の候補が、分岐のたびに変わります。左側の分岐では、p_1、p_5、p_6が候補でp_5が使われ、右側の分岐ではp_3、p_5、p_7が候補でp_3が使われるというように、分岐のたびにランダムに特徴量の候補が選ばれます。このような感じで回帰木をたくさん作っていくものが、ランダムフォレストという手法です。

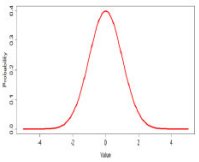


それぞれの回帰木の予測値は、最後に落ちたところのＹの平均値です。これをＢ個、たくさん作って平均します。回帰木を作るときに特徴量やサンプルがランダムに選ばれているので、その中で非線形の効果や交互作用などがうまい具合に調整されて、いい感じの予測値になる。ざっくり言うとランダムフォレストはそのような感じの手法でして、予測精度がそれなりに高い手法として知られております。

ただ、ランダムフォレストは決定木とは異なり、どの特徴量が作用しているのかを解釈することが、非常に難しくなります。一つ一つの木を見ていこうにも限界があり、木の数Ｂ個というのはケース・バイ・ケースではありますが 500 個や 1,000 個などありまして、これを一つずつ見ていくのは現実的ではありません。確かに予測精度は高いかもしれませんが、解釈という意味では非常に難しい例として、ランダムフォレストは分かりやすいと思っております。これが一つ目の課題です。

４．アクチュアリーが機械学習の手法を扱う上での課題

旧来の統計解析の手法では、モデルの中で目的変数が従う確率分布が仮定。
⇒モデルが正しいとすれば予測誤差の定量化の議論が可能。(*)
機械学習の手法は、確率分布を仮定した手法ではない。
⇒予測誤差の定量化の議論は難しい。（課題２）

線形モデル	サポート ベクトル回帰	ランダム フォレスト
$Y_i \sim N(\mu_i, \sigma^2)$ $\mu_i = b + \boldsymbol{x^{(i)} w^T}$ を仮定したモデル。 	予測精度の向上を目的として、以下の損失関数を最小化する回帰式を求めている。 $C\sum_{i=0}^{n}\max\{0, \lvert y^{(i)} - (b + \boldsymbol{x^{(i)} w^T})\rvert - \varepsilon\} + \frac{1}{2}\lVert \boldsymbol{w} \rVert^2$ モデルの中でY_iに対して確率分布が仮定されているわけではない。	１つ１つの回帰木は、損失関数を減少させるアルゴリズムで分岐させた結果得られるもので、その中に確率分布の仮定はない。 

15 (*) しかし、この方法で得られる誤差はすべての誤差を考慮できていない。リスク研究チームA班の発表参照。

続いて、誤差の定量化についての課題をお話しします。冒頭でお話ししたとおり、旧来の統計解析の手法の中では、モデルの中で目的変数が従う確率分布が仮定されています。ですから、モデルが正しいと仮定すれば、予想誤差の議論が可能です。

例えば線形モデルでは、仮定した正規分布に沿った不確実性の評価が可能です。一方、サポートベクトル回帰は、見た目は回帰の式をしていますけれども、このＹに正規分布などが仮定されているわけではありません。あくまで、この損失関数を最小化すると予測精度のいい回帰式が得られるという考えのもとに作られていますので、この得られた予測値の確率分布がどのようになっているかを議論するのは簡単ではありません。ランダムフォレストは、もっと分かりやすいと思います。ランダムフォレストは、先ほどの話の中で確率分布に関する話は出てこなくて、一つ一つの回帰木の決定の中で損失関数を減らす分岐をして、その結果、得られる回帰木をたくさん作って平均しているという手法ですので、この中に確率分布の議論はありません。

４．アクチュアリーが機械学習の手法を扱う上での課題

課題１：予測精度が高い手法ほど解釈が難しくなる傾向
課題２：予測誤差の定量化の議論は難しい

16

まとめますと、１つめの課題は予測精度が高い手法ほど解釈が難しい傾向があること、２つ目の課題は機械学習の手法は誤差の定量化は難しいということ、この二つが、アクチュアリーが機械学習の手法を扱う上での主な課題として認識しております。本日の３名のプレゼンテーションは、これらの課題に対する一つの我々の考えをお示しするものになります。詳細はほかの３名からご説明いただきますので、割愛いたします。ここで浅芝さんにバトンタッチしたいと思います。

５．本日のこの後のプレゼンテーションについて

- ＩＭＬチーム（浅芝　良一さん）
  PFI(Permutation Feature Importance), ICE(Individual Conditional Expectation), SHAP(SHapley Additive explanation)などのモデル解釈に有益な指標およびそのモデリングプログラムの実例を紹介。

- リスク研究チームA班（高橋　智嗣さん）
  機械学習の手法により得られた予測値の予測誤差を推定し、
  さらにその要因をプロセス誤差、パラメータ誤差、モデル誤差に分解する
  汎用的な予測誤差分解手法を紹介（本WG独自の研究成果）

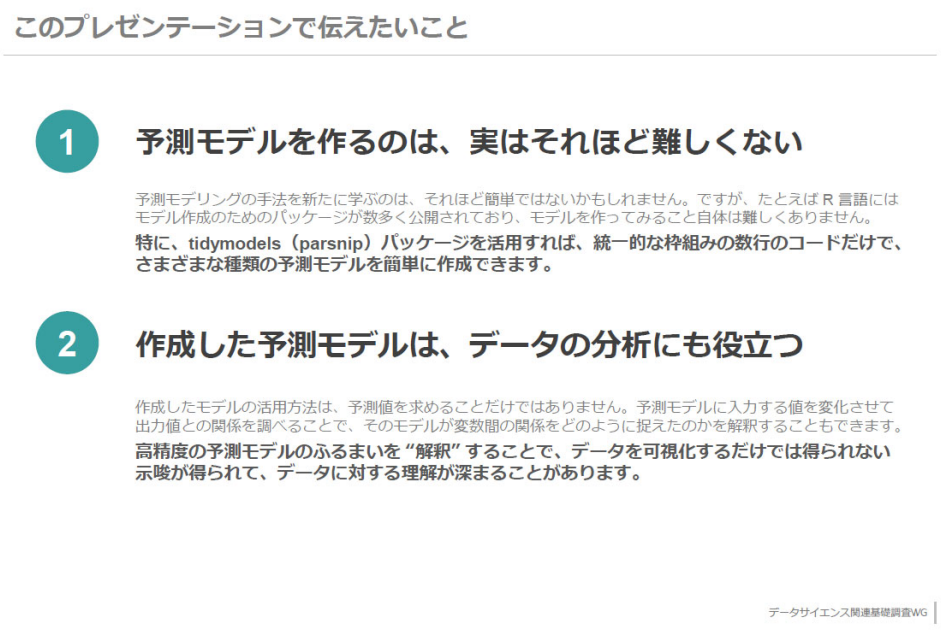
- リスク研究チームB班（藤田　卓さん）
  リスク研究チームA班の研究している汎用的な予測誤差分解手法を改善する
  ランダムフォレストに特化した手法を紹介（本WG独自の研究成果）

17

岩沢　鈴木さん、ありがとうございました。私の代わりに全体の案内をしていただいたので、スムーズに入っていけると思います。次は、Ｔ＆Ｄホールディングスの浅芝さんで、タイトルは「ブラックボックス

な予測モデルの力を借りてデータを分析してみよう」です。それでは、お願いいたします。

浅芝　ありがとうございます。ここからは、ＩＭＬチームの浅芝がご報告いたします。よろしくお願いします。



このプレゼンテーションを通じて私がお伝えしたいことは、二つあります。一つめは、予測モデルを作ることは、実はそれほど難しくないということです。今回はＲ言語を使用しますが、Ｒでは、いろいろな種類の予測モデルを作成するためのパッケージが既に公開されています。パッケージの使い方を一つ一つ調べていくことは大変ですが、tidyverse や tidymodels などのパッケージを活用すると、統一的で、しかもシンプルなコードで、さまざまな種類の予測モデルを作ることができます。

二つめは、作成した予測モデルは、データの分析にも役立つということです。さまざまな予測モデリング手法の中には、予測における入力と出力の関係が分かりにくい一方で、作成者が特に工夫をしなくても高い予測精度を発揮するものがあります。そのような予測モデルを「解釈」する手法を使うと、データを可視化するだけでは分からなかったような示唆が得られて、その後のデータ分析に生かせることがあります。ということで、さっそく、予測モデルを作成してみましょう。

・・・ということで、さっそくやってみよう

- **insurance データセット**[*1]
  - ある医療保険制度における、被保険者1,338人の属性と医療費に関するデータです。

変数名	説明	値					
age	年齢	19	18	28	33	32	・・・
bmi	BMI	27.9	33.8	33.0	22.7	28.9	・・・
children	子供の数	0	1	3	0	0	・・・
smoker	喫煙（1：あり、0：なし）	1	0	0	0	0	・・・
gender	性別（1：女性、0：男性）	1	0	0	0	0	・・・
charge	医療費	16,885	1,726	4,449	21,984	3,867	・・・

  - 今回は、**被保険者の属性をもとに医療費の値を予測する回帰モデル**を作成します。

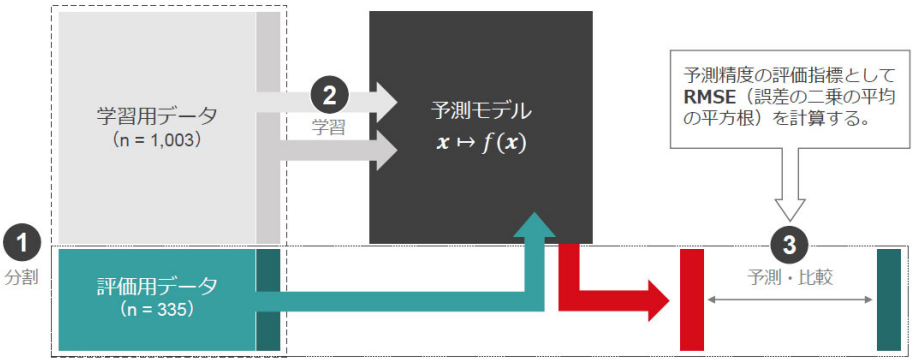
(*1) この発表では、liverパッケージに収録されている同名のデータセットを利用します。同じデータを https://www.kaggle.com/datasets/mirichoi0218/insurance などでも入手できますが、一部の変数名が上の表とは異なるので注意してください。なお、変数 region を削除し、smoker と gender を数値に変換しています。

データサイエンス関連基礎調査WG | 3

まず、今回使用するデータについて、簡単にご説明いたします。今回は、insurance データセットを使用します。これは、ある架空の医療保険制度における被保険者 1,338 人の属性と医療費に関するデータです。被保険者の属性としては、年齢、ＢＭＩ、子供の数、喫煙状況、性別が与えられています。今回の発表では、被保険者の属性から、その医療費の値を予測する回帰モデルを作成します。なお、発表資料と一緒にＲのコードを掲載（本記事の末尾に再掲）しておりますので、お手元にＲの実行環境がある方は、ぜひ実行してみてください。

予測モデルの作成から評価まで

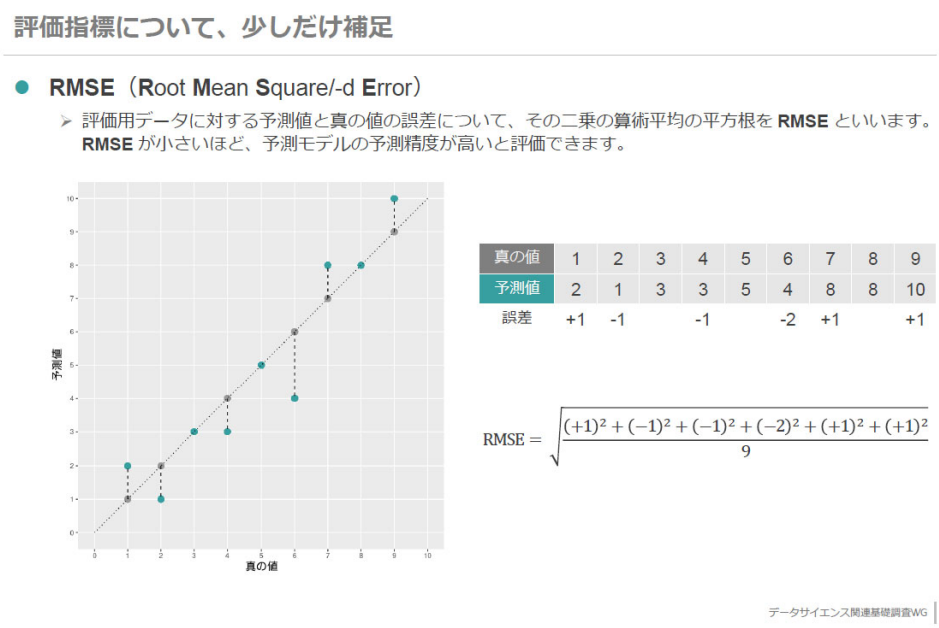
1 データを**学習用データ**と**評価用データ**に分割します。
2 **学習用データ**のみを用いて予測モデルを作成します。
3 **評価用データ**の医療費を予測して真の値と比べます。



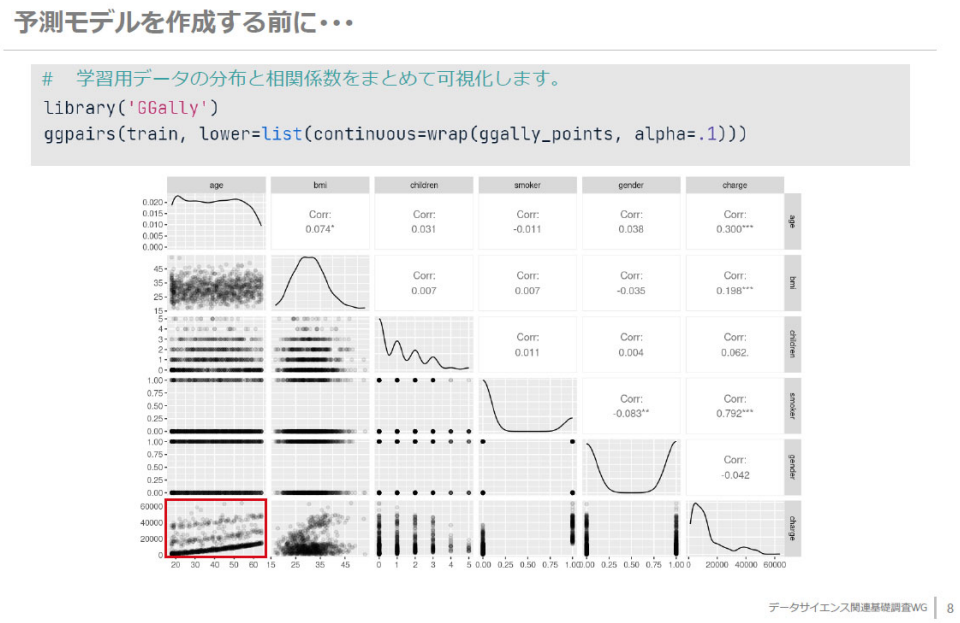
データサイエンス関連基礎調査WG | 5

ここで、予測モデルを作成してから評価するまでの流れを、三つのステップに分けて簡単にご説明しま

す。まず、使用するデータを、予測モデルの作成に使う「学習用データ」と、予測モデルの評価に使う「評価用データ」に分割します。次に、学習用データの情報だけを使って予測モデルを作成します。最後に、作成した予測モデルに評価用データの属性を入力することで、評価用データに対する医療費の値を予測し、これを正解である真の値と比較することで、予測モデルを評価します。



予測精度の評価についても、簡単に補足いたします。今回は、ＲＭＳＥという指標を用いて予測モデルを評価します。これは Root Mean Square/-d Error の略で、真の値と予測値の各ペアについてその差を２乗し、その平均を取り、さらにその平方根を取った値のことです。ＲＭＳＥが小さいほど予測が正確であるといえ、予測が完璧であれば、ＲＭＳＥの値は０になります。



なお、学習用データの分布をまとめてプロットすると、図のようになります。左下は、被保険者の年齢

に対して医療費をプロットしたものですが、横軸に対して３本の線を描いているというような特徴が見られます。

パッケージを活用してモデルを作ろう

- tidymodels の **parsnip パッケージ**
  - **parsnip パッケージ**を使えば、統一的なコードによって、多彩な種類の予測モデルを作成できます。

```
model = linear_reg() %>%          ❶ タイプ（予測モデルの種類）を指定
    set_engine('lm') %>%          ❷ エンジン（パッケージ名など）を指定
    set_mode('regression') %>%    ❸ モード（回帰や分類などの種別）を指定
    fit(charge~., train)          ❹ 回帰式*1と学習用データを渡して学習
```

  - 今回は、下表の予測モデルを作成します。

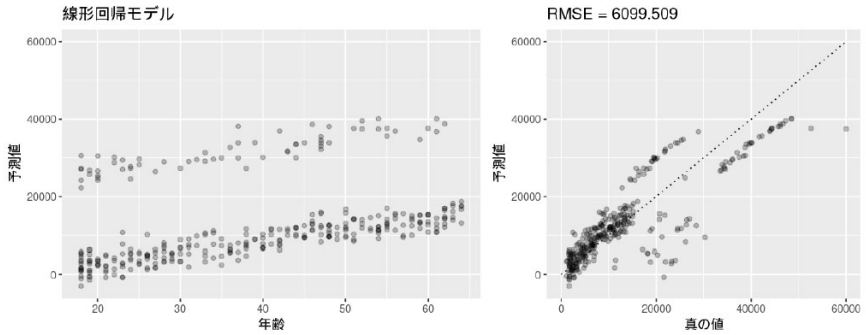
予測モデル	タイプ	エンジン	モード
線形回帰モデル	linear_reg	lm*2	regression*2
サポートベクトル回帰モデル－線形カーネル	svm_linear	kernlab	regression
サポートベクトル回帰モデル－RBFカーネル	svm_rbf	kernlab*2	regression
ランダムフォレストモデル	rand_forest	ranger*2	regression
勾配ブースティングモデル－XGBoost	boost_tree	xgboost*2	regression
ニューラルネットワークモデル	mlp	nnet*2	regression

(*1) Rの多くのパッケージでは、回帰式（formula）を"目的変数~説明変数からなる式"という記法で指定します。なお、ドット（.）はデータに含まれるすべての変数の一次の項を表します。
(*2) デフォルト値のため、コード上は指定を省略することができます。

さて、今回は、tidymodels パッケージの一部にもなっている parsnip パッケージを使って予測モデルを作成していきます。parsnip パッケージでは、まず作成した予測モデルのタイプを指定し、次に、そのモデルの作成に使用するエンジンを指定し、さらに、「回帰」や「分類」などのタスクの種類を表すモードを指定し、最後に回帰式と学習用データを入力することで、モデルを作成することができます。Ｒでは、予測モデルを構築するためのパッケージが数多く公開されていますが、この parsnip パッケージを使うと、いろいろなパッケージの使い方をいちいち調べたり、覚えたりしなくても、シンプルで統一されたコードの書き方を覚えるだけで、多彩な種類の予測モデルを作成することができます。今回は、下の表に記載した六つの設定で予測モデルを作成していきます。

まずは線形回帰モデル（重回帰分析）から

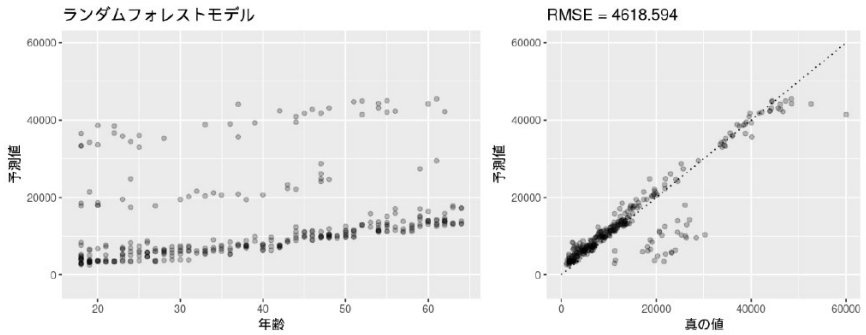
```
#　線形回帰モデルを作成します。作成した予測モデルはリストに格納します。
models = list()
models[['lm']] = linear_reg() %>% fit(charge~., train)
models[['lm']]$title = '線形回帰モデル'
evaluate_model(models[['lm']], test)
```



まず、線形回帰モデルを作ってみましょう。このスライドのコードの２行めで、モデルのタイプを linear_reg と指定し、回帰式と学習データを入力することで予測モデルを作成しています。なお、エンジンとモードについては、linear_reg 関数の規定値をそのまま使用するため、コードでの指定を省略しています。下段のプロットは、それぞれ評価用データにおける年齢と真の医療費を横軸に取って、予測値を縦軸に取ってプロットしたものです。この予測モデルのＲＭＳＥは少し大きく、6,099 となりました。

決定木系のモデルも作ってみよう

```
#　ランダムフォレストモデルを作成します。
models[['rf']] = rand_forest() %>%
    set_mode('regression') %>% fit(charge~., train)
models[['rf']]$title = 'ランダムフォレストモデル'
evaluate_model(models[['rf']], test)
```



次に、ランダムフォレストモデルを作成してみましょう。モデルのタイプを rand_forest と指定し、モードを regression、すなわち回帰と指定することで、ランダムフォレスト回帰モデルを作成できます。線形回帰モデルに比べて予測精度が大きく改善し、ＲＭＳＥは 4,618 まで低下しました。ここでは省略しますが、他の予測モデルも、同じようなコードで作成することができます。

予測モデルを作るのは、実はそれほど難しくない

- **作成した予測モデルの一覧**
  - ここまでに見たように、R のパッケージ（特に tidyverse や tidymodels に含まれるもの）を活用すれば、下の一覧に含まれるような予測モデルを、とてもシンプルなコードで作成することができます。

予測モデル	RMSE
線形回帰モデル	6099.509
サポートベクトル回帰モデル – 線形カーネル	6662.996
サポートベクトル回帰モデル – RBFカーネル	4731.479
ランダムフォレストモデル	4618.594
勾配ブースティングモデル – XGBoost	4783.858
ニューラルネットワークモデル	4545.479

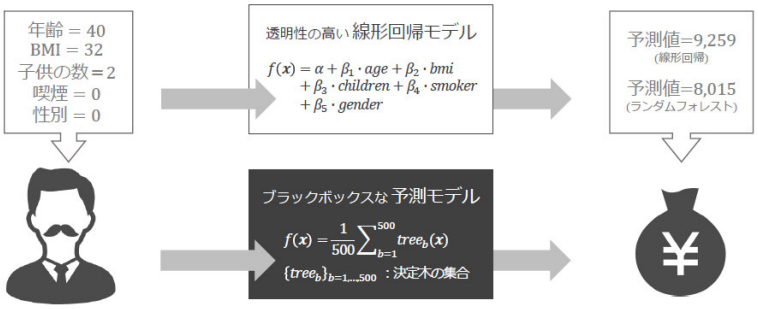
  - **予測モデルを作るのは、**少なくとも作ってみるだけなら･･･**それほど難しくありません！**

データサイエンス関連基礎調査WG | 17

資料の中で作成している予測モデルの種類と、評価用データに対するＲＭＳＥの値を、スライドの表にまとめています。ここまでに見たように、Ｒでは、便利なパッケージを活用することで、本当にいろいろな予測モデルを、シンプルで統一的なコードで作成することができます。予測モデルを作ることは、少なくとも作ってみるだけであれば、それほど難しくはないということをお伝えできたのではないかと思います。

予測モデルをデータ分析に活用する

- **ブラックボックスモデル**
  - 入力された変数の値から予測値を計算するまでの過程が複雑で、把握が困難であるような予測モデルを、**ブラックボックスな予測モデル**と呼びます。予測モデルのブラックボックス化は、予測の計算ステップが多すぎたり、入力された変数を変換する先の次元が大きすぎたりするために起こります。

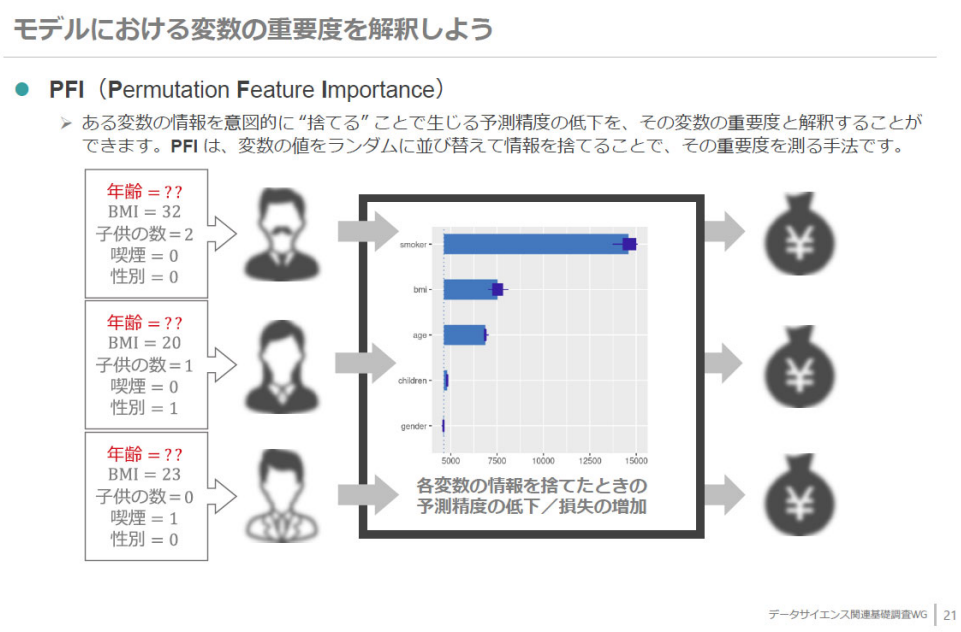


データサイエンス関連基礎調査WG | 19

ここからは、作成したモデルを解釈する手法を、その実例とともに紹介いたします。ここまでに作成した予測モデルのいくつかは、入力された変数の値から予測値を計算するまでの過程が複雑で、人間の能力では、その過程を把握することが非常に難しいようなものになっています。たとえば、今回作成したラン

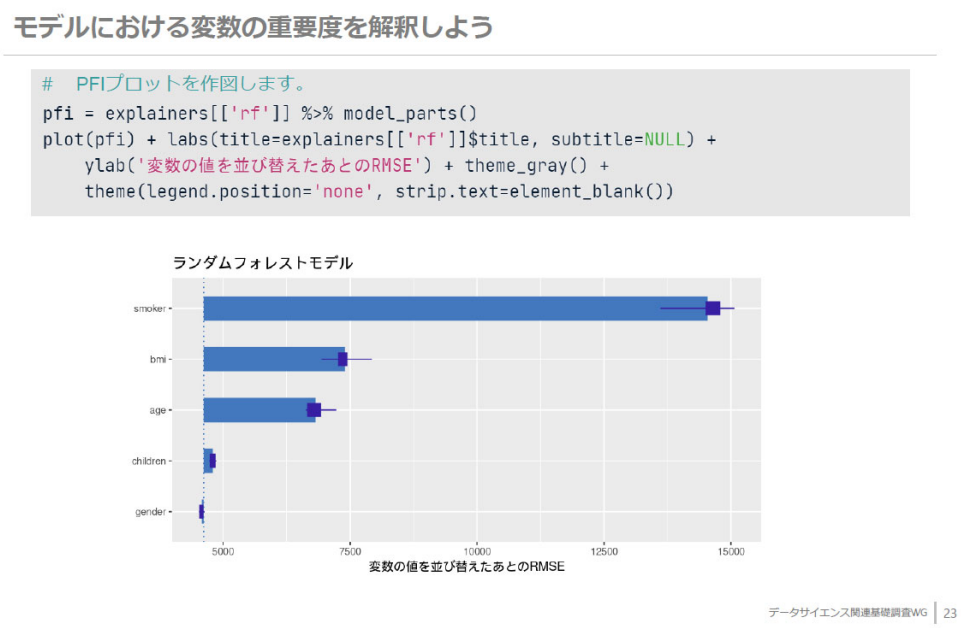
ダムフォレストモデルでは、1 人の被保険者に対する医療費の予測値を得るために、500 本の異なる決定木をそれぞれたどってみる必要があります。そのような計算を人間の頭の中で実行することは、全く現実的ではありません。このように、入力と出力の関係を把握することが難しい予測モデルのことを、「ブラックボックスモデル」と呼ぶことがあります。

さて、ブラックボックスモデルであっても、コンピューターの計算力を使えば、予測値を得ること自体は非常に簡単です。そこで、たとえば入力値を少しずつ変更してみて、出力される予測値と入力値との関係を調べることで、予測モデルの関数的なふるまいのようなものを知ることができます。今回は、予測精度が高かったランダムフォレストモデルのふるまいを解釈して、線形回帰モデルを改良するヒントとして活用できないか、考えてみましょう。



具体的なモデル解釈の手法として、最初にＰＦＩをご紹介します。ＰＦＩは Permutation Feature Importance の略で、モデルが高い予測精度を達成する上で、どの変数が重要であったのかを解釈する手法です。ＰＦＩのアイデアはシンプルです。データから予測値を算出するときに、ある変数の情報を意図的に捨てると、予測精度が当然悪化します。ＰＦＩでは、この予測精度の悪化を、その変数の重要度と解釈します。

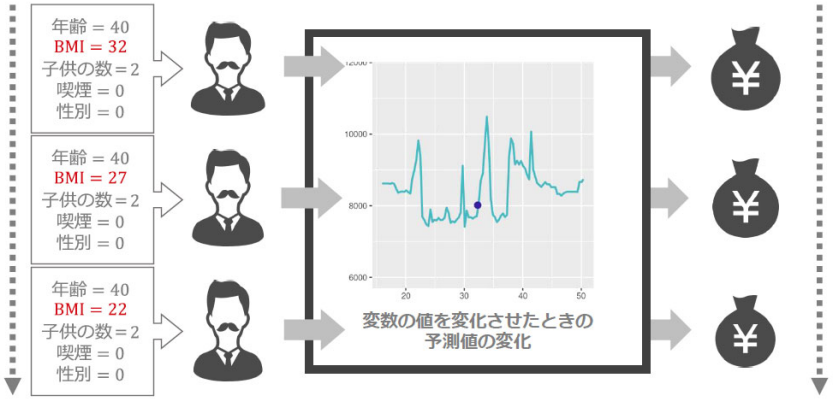
変数の情報を捨てるということは、その情報を使わずに予測を行うということですが、ＰＦＩでは、評価用データの中で対象となる変数の値をランダムにシャッフルしてしまいます。そうすることで、その変数の情報を予測において使えないという状態を実現します。つまり、各変数の値をデータの中でシャッフルしてから予測をすれば、当然、予測の精度は悪化し、ＲＭＳＥの値が増加してしまいますが、この増加幅をシャッフルした変数の重要度と見なすということです。



こちらのスライドの図は、今回作成したランダムフォレストモデルのＰＦＩをプロットしたものです。これを見ると、被保険者の属性のうち、喫煙状況・ＢＭＩ・年齢の三つが重要で、特に喫煙に関する情報を失うと、ＲＭＳＥが元の 4,618 から、およそ１万 4,600 まで大きく増加してしまうことが分かります。一方で、子供の数や性別は、たとえその情報を捨てたとしても、モデルの予測精度にほとんど影響しないようです。

変数と予測値の関係を解釈しよう

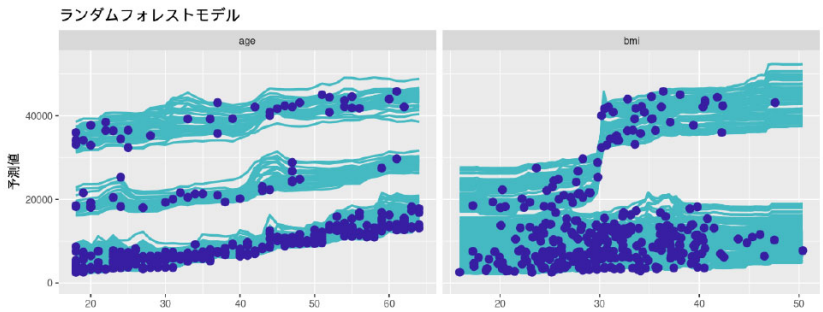
- **ICE（Individual Conditional Expectation）**
  - 個別の予測について、ある特定の変数の値を変化させてみて、それに伴って予測値がどのように変化するかを調べることで得られるプロットを **ICE** といいます。



次に紹介するのは、ＩＣＥという手法です。これは Individual Conditional Expectation の略で、個別の予測について、ある変数の値だけを少しずつ変化させていくときに、予測値がどのように変化していくかを調べることで、その変数に関する一変数関数を作っていく手法です。

変数と予測値の関係を解釈しよう

```
#  ICEプロットを作図します。
ice = explainers[['rf']] %>% predict_profile(test)
plot(ice, variables=c('age', 'bmi')) +
    labs(title=explainers[['rf']]$title, subtitle=NULL, y='予測値') +
    ylim(0, 53000) + theme_gray()
```

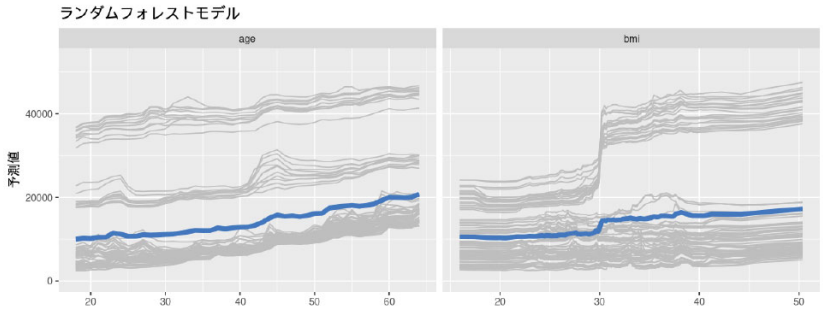


こちらの図は、評価用データの全ての予測についてＩＣＥを計算し、まとめてプロットしたものです。ＩＣＥでは、この図のように、複数の予測値に対して計算した関数を、ひとまとめにプロットするということもよく行われます。左下のプロットは、横軸に被保険者の年齢を取ったものです。このプロットからは、おおむねどの被保険者についても、仮にその人がもっと高齢であれば、医療費はほぼ単調に増加していただろうということが読み取れます。

また、右下のプロットは、横軸に被保険者のＢＭＩを取ったものです。このプロットからは、もしＢＭＩがもっと高ければ、医療費が非常に大きくなっていただろうというグループと、もしＢＭＩがもっと高かったとしても、医療費はそれほど変わらなかっただろうという二つのグループが、データの中に混在していることが分かります。このような場合、ＢＭＩと交互作用を持っている他の変数が存在することが強く示唆されます。

変数と予測値の関係を解釈しよう

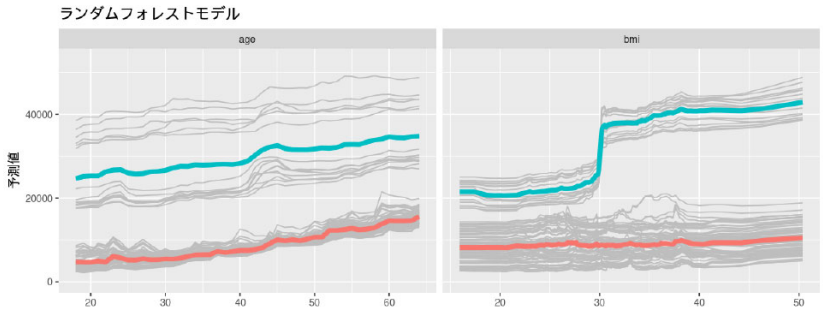
```
#　PDP（Partial Dependence Plot）を作図します。
pdp = explainers[['rf']] %>% model_profile()
plot(pdp, variables=c('age', 'bmi'), geom='profiles') +
    labs(title=explainers[['rf']]$title, subtitle=NULL, y='予測値') +
    ylim(0, 53000) + theme_gray() + theme(legend.position='none')
```



こちらの図は、評価用データのＩＣＥの平均値をプロットしたものです。このような図は、Partial Dependence Plot、省略してＰＤＰと呼ばれます。ＰＤＰは、ある変数の値を変えたときに、その変数が予測値に与える平均的な影響として解釈することができます。

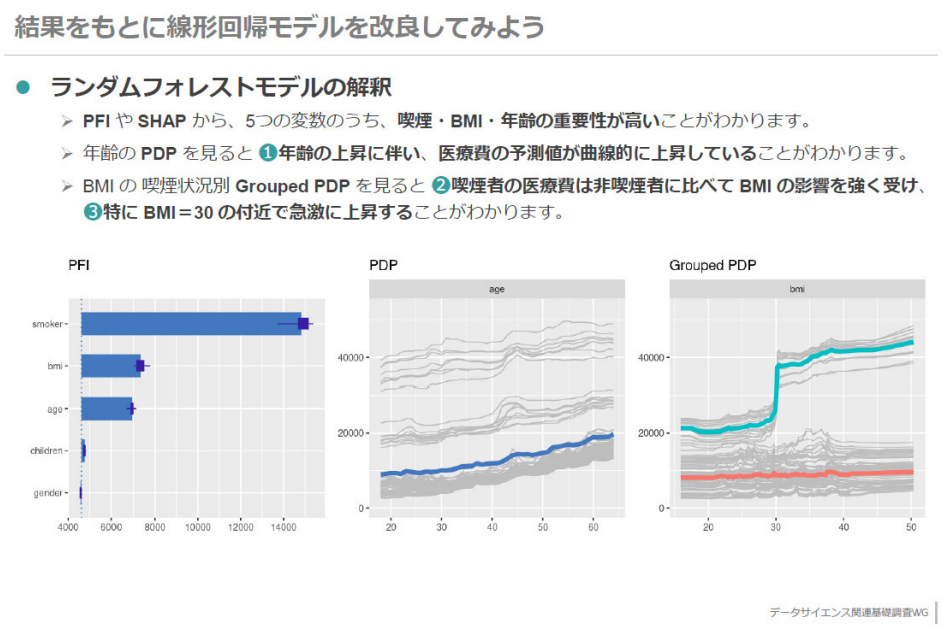
変数と予測値の関係を解釈しよう

```
#　喫煙の有無でグループ化したPDP（Grouped PDP）を作図します。
gpd = explainers[['rf']] %>% model_profile(groups='smoker')
plot(gpd, variables=c('age', 'bmi'), geom='profiles') +
    labs(title=explainers[['rf']]$title, subtitle=NULL, y='予測値') +
    ylim(0, 53000) + theme_gray() + theme(legend.position='none')
```

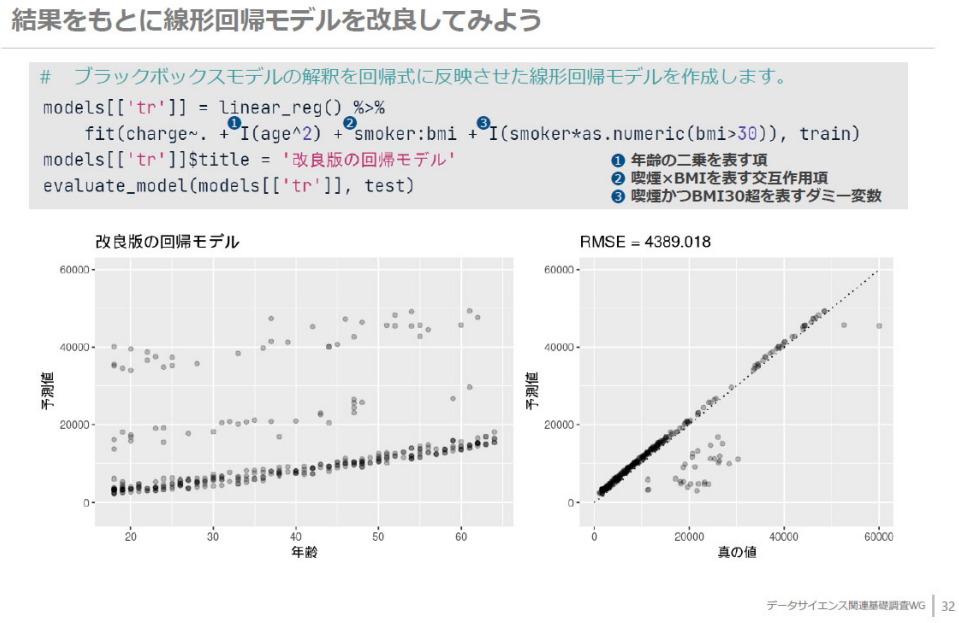


こちらの図は、被保険者の喫煙の有無で評価用データをグルーピングして、そのグループ別にＰＤＰを

描いたものです。緑の線は喫煙者、オレンジの線は非喫煙者における予測値の変化を表しています。



さて、それでは、ランダムフォレストモデルを解釈した結果を用いて、線形回帰モデルを改良できないかを考えてみましょう。ＰＦＩからは、五つの属性のうち、喫煙・ＢＭＩ・年齢の重要性が高いことが分かります。また、年齢のＰＤＰを見ると、年齢の上昇に伴って医療費の予測値が、直線というよりは、やや曲線的に上昇していることが分かります。さらに、ＢＭＩの喫煙状況別ＰＤＰを見ると、喫煙者の医療費は非喫煙者に比べてＢＭＩの影響を強く受けており、特にＢＭＩが 30 を超えるあたりで急激に上昇することが分かります。これらの知見を、線形回帰モデルの回帰式に反映させてみましょう。



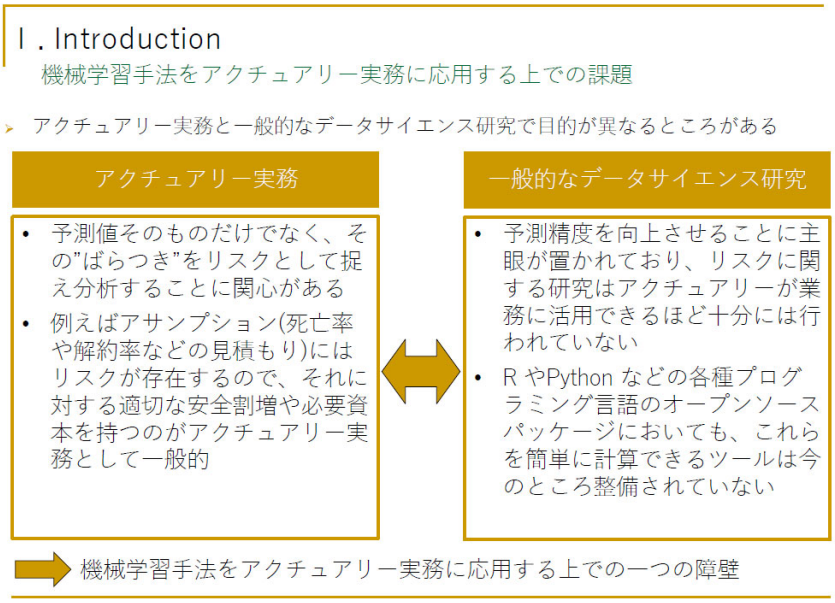
Ｒでは、予測モデルを作成するときに、回帰式を指定することができます。ここでは、モデルの回帰式に対して、スライドに記載の①から③までの三つの項を追加してみました。①の項は、年齢の２乗を表しています。②の項は、喫煙状況とＢＭＩの交互作用を表す項です。③の項は、喫煙者でかつＢＭＩが 30 を

超えるときにだけ１という値を取り、他のときには０という値を取るダミー変数です。これは、ＢＭＩの非連続的な影響を表現したものです。この回帰式で作成した回帰モデルは、ほとんどの被保険者についてかなり正確な予測を出力しており、ＲＭＳＥの値も 4,389 まで改善しました。

最後にもう一度、今回作成した予測モデルと、そのＲＭＳＥを一覧でまとめておきます。今回は、ランダムフォレストモデルの解釈から得た知見を用いて線形回帰モデルを改良することで、シンプルでありながら、かつ、ブラックボックスモデルを上回る精度の回帰モデルを作成することができました。正直なところ、もっと複雑な現実のデータを分析する場面では、今回のような予測精度の逆転が起きることは、まずないだろうと思います。ただ、少なくとも、ブラックボックスな予測モデルは、実務で使う予測モデルとしては採用できないという場合でも、もっと透明性の高い線形回帰のようなモデルを改善するためのヒントを与えてくれるかもしれません。

以上で、私からの報告を終わります。ありがとうございました。

岩沢　浅芝さん、ありがとうございました。素晴らしいですね。皆さんも使う気になったのではないかと思います。次は、アクサ生命の高橋さんで、タイトルは「予測モデリングとアクチュアリー実務：汎用的な誤差分解・推定手法の必要性と可能性」です。では、お願いいたします。



高橋　アクサ生命の高橋です。よろしくお願いいたします。まず、イントロダクションとして、機械学習手法をアクチュアリー実務に応用する上での課題についてご説明いたします。

アクチュアリー実務と一般的なデータサイエンス研究では、目的が異なるところがあります。アクチュアリー実務では、予測値そのものだけではなく、そのばらつきをリスクとして捉えて、分析することに関心があります。例えばアサンプションの場合には、アサンプションの見積もりには必ずリスクが存在するので、それに対する適切な安全割増や必要資本を持つことが、実務として一般的です。一方、一般的なデータサイエンス研究では、予測精度を向上させることがメインのターゲットでありリスクに関する研究は、アクチュアリーが業務に活用できるほど十分には行われていないと認識しております。以上のように、両者にギャップがありますので、機械学習手法をアクチュアリー実務に応用する上で、一つの障壁になっていると考えております。

Ⅰ. Introduction
アクチュアリー実務における誤差（リスク）の捉え方

- アクチュアリーはかつてより予測誤差（予測値と実績値の差）を以下の3つの誤差（リスク）に分解し捉えてきた。
- ASOP43による定義は以下のとおり。（以下、発表者にて日本語訳）

Model Error
- 手法が状況に応じて適切でない、あるいはモデルが相応しいものではない不確実性

Parameter Error
- 手法やモデルで使用されるパラメータが相応しいものではない不確実性

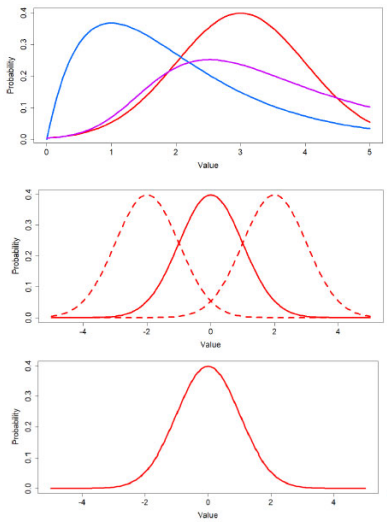
Process Error
- パラメータが分かっている場合であっても、内在する不確実性によって偶然的に変動する不確実性

※ASOP43ではErrorではなくRiskと呼んでいるが、本研究では機械学習手法等を用いた予測における予測誤差の分解を想定していることから、Errorと読み替えている。

2

では、アクチュアリーが、誤差やリスクなどをどのように捉えて分析してきたか、一つのフレームワークとして、誤差やリスクなどを、モデル誤差とパラメータ誤差、プロセス誤差の三つの要素に分解して捉えてきました。この三つがそれぞれ何を意味しているか、このスライドに定義を記載していますが、文字では分かりにくいと思いますので、次のスライドでグラフを使ってイメージで説明したいと思います。

Ⅰ. Introduction
各誤差に関する一般的な解釈



モデル誤差…
モデルが適切でないことに起因

パラメータ誤差…
モデルは適切であるとして、パラメータが適切でないことに起因

プロセス誤差…
モデルとパラメータは適切であるとして、内在する不確実性に起因

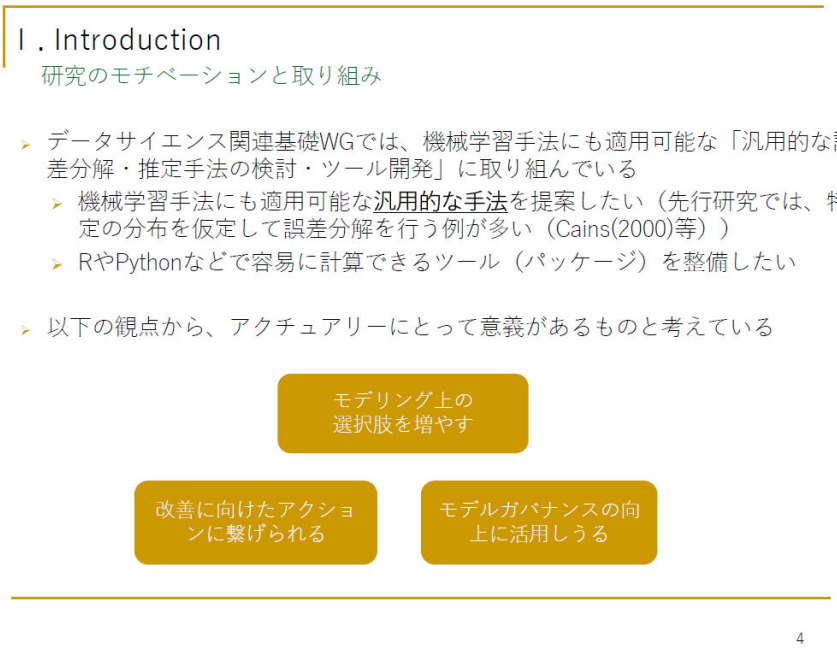
3

まず、モデル誤差は、「モデルが適切でないことに起因する誤差」です。例えば、一番上のグラフ上の赤い実線が真のモデルだとします。分析上、青い実線を使っている場合は、そもそも想定しているモデルの関数形が違いますので、誤差が生じます。そのようなものをモデル誤差と呼びます。

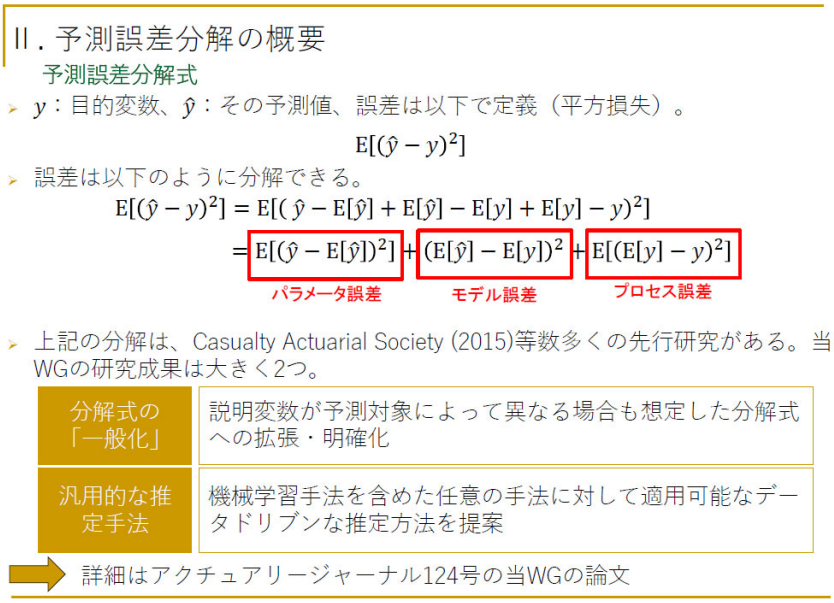
次にパラメータ誤差は、「モデルは適切ではあるけれども、パラメータが適切でないことに起因する誤差」です。グラフで見ると、赤い破線のようなグラフを分析上使った場合に、関数形は合っているのですが、中心の位置を示すパラメータの推定が少しずれておりまして、それによって誤差が生じます。そのような

ものをパラメータ誤差と呼びます。

最後にプロセス誤差ですが、「モデルとパラメータは適切であるけれども、なお内在する不確実性に起因する誤差」です。グラフのように、真の確率分布、真のモデルを完全に知っていたとしても、分析対象としている事象自体が確率的な事象ですので、必ず推定に誤差が生じます。そのようなものをプロセス誤差と呼びます。



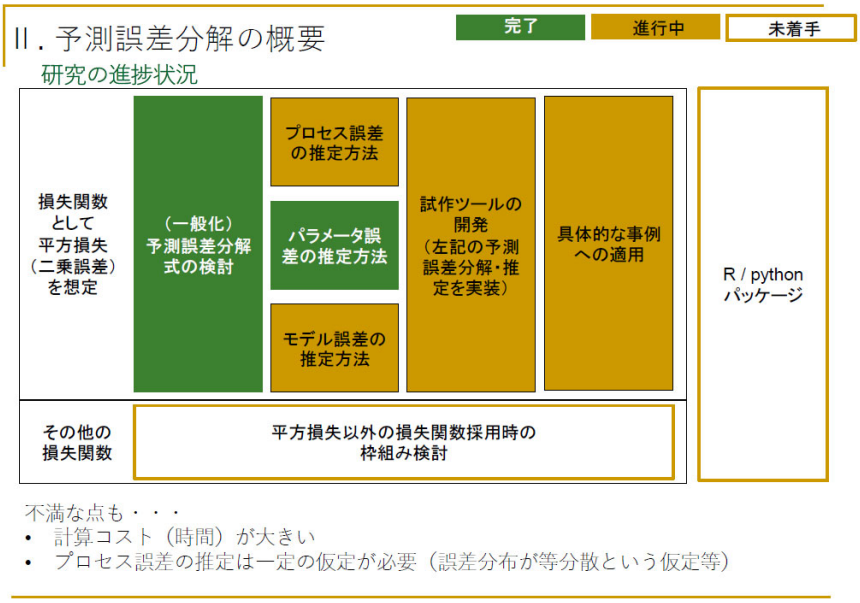
以上、アクチュアリー実務と一般的なデータサイエンス研究のギャップおよび、アクチュアリーが誤差やリスクなどをどのように分析してきたかを紹介しました。ワーキンググループでは、それらを融合させるべく、機械学習手法に適用可能な汎用的な誤差分解、推定手法の検討に取り組んでおります。つまり、やりたいこととしては、どんな機械学習手法を持ってきたとしても、先ほど言った三つの要素に誤差を分解可能な汎用的な手法を提案したいと考えております。それの実現によって、アクチュアリーにとってもいろいろな意義があります。ここに三つ挙げていますが、この中身については、後ほどケーススタディとともにご説明いたします。



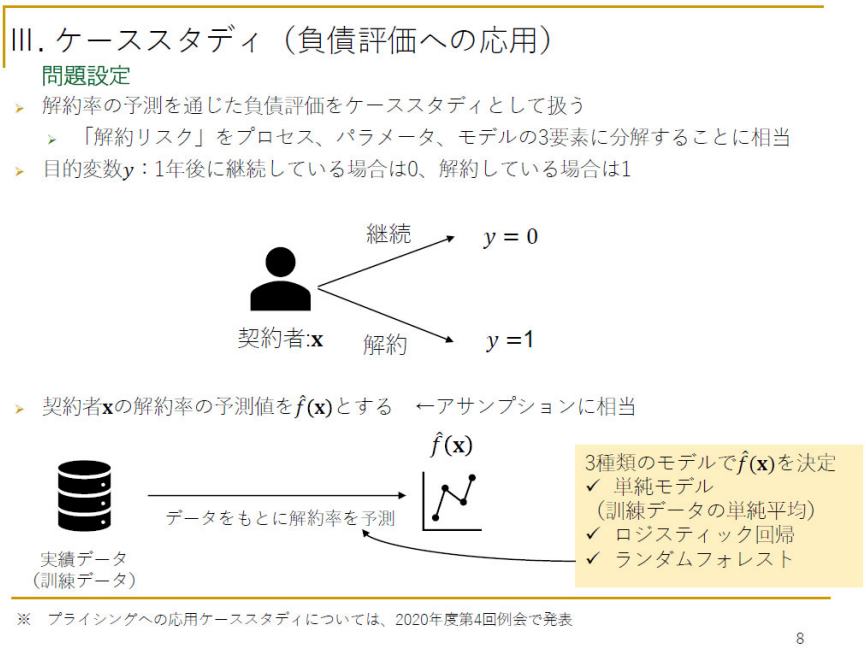
このスライドでは、誤差分解をどのようにやるのか、簡単に１枚で説明したいと思います。まず、記号の定義として、y を目的変数、y ハットをその予測値とします。誤差は、スライドの一番上に書いてある式の平方損失といわれているもので定義します。よくあるテクニックだと思いますが、中間変数を挟んで展開すると三つの項が出てきて、赤く囲っているそれぞれの項を、パラメータ誤差、モデル誤差、プロセス誤差と呼んでいます。

ここでは一つだけ例をとって、真ん中の項がなぜモデル誤差と呼ばれるかを簡単に説明します。真ん中の項の y ハットの平均は、モデルから出力される値の平均値を示しています。一方、y の平均は、モデルは関係なく、実際の観測から得られる値の平均値を表しており、その差分はモデルと実際の事象のバイアスを測っているような項であり、従ってモデル誤差と解釈することができます。

ここで説明した誤差分解は多くの先行研究がありますが、我々の成果としては、分解式の「一般化」と、汎用的な推定手法の提案という二つになります。分解式の一般化については、今日の発表の中では、上の式と同じものと思っていただいて大丈夫です。汎用的な推定手法については、どのような機械学習手法を持ってきたとしても、上の三つの項を具体的にデータドリブンで評価できる手法を提案しております。手法の中身は、今日は時間がなくて説明できないので、ご興味があれば、『アクチュアリージャーナル 124 号』をご参照いただければと思います。

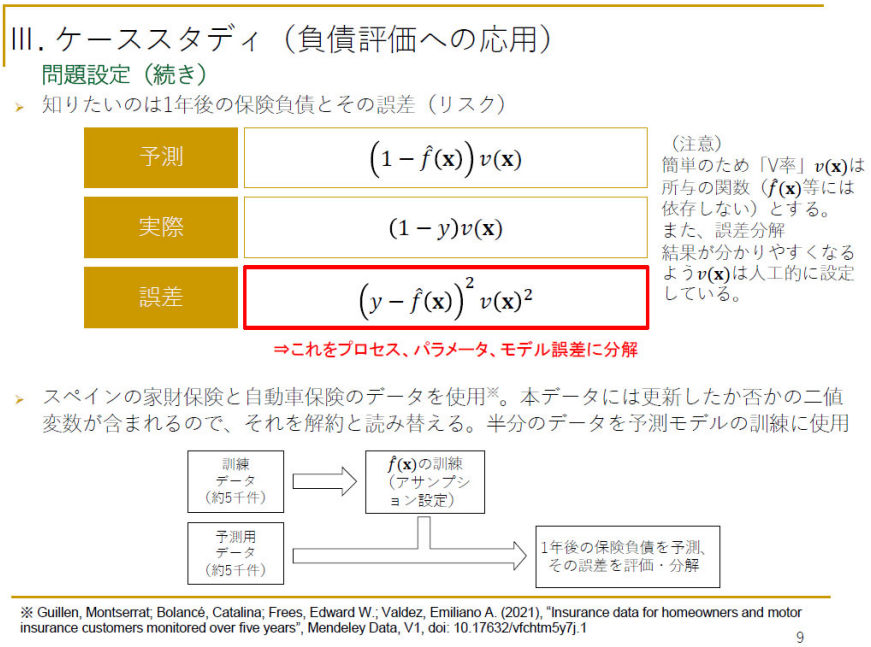


このスライドは研究の進捗状況を示しています。今日はあまり詳しくは説明しませんが、一言で言うと、ある程度の形はできているが、まだまだチャレンジする課題が多いという状況です。例えば、一番下に書いてあるとおり、プロセス誤差の推定については、一定の強めの仮定が必要という状況です。このような課題もあるので、今後改善していく必要があると思っております。



ここまで誤差分解の概略を説明してきましたが、ここからは、ケーススタディを使って、誤差分解を行うとどのようなアウトプットが得られるか、どのようなうれしいことが起きるかというところを、詳しく説明していきたいと思います。

ケーススタディとしては、解約率の予測を通じた負債評価を考えます。目的変数ｙは、１年後にある契約が継続している場合を０、解約している場合は１という変数と定義します。ある契約者の解約率の予測値を、ｆハット(x)という関数で置きます。これは、実績データ（または訓練データ）を基に決定します。これは、まさに一般的なアクチュアリーの実務で行われているアサンプションの決定と、全く同じ流れだと思います。スライドの右下に書いているとおり、ｆハット(x)を決めるために、このケーススタディでは単純モデルとロジスティック回帰、ランダムフォレストの三つの方法を使用します。



先ほど解約率に焦点を当てて話をしましたが、我々が最終的に評価したいのは保険負債だと思います。そこで、１年後の保険負債の予測と実績は、どのように書けるのかということを考えてみます。まず予測については、ｆハット(x)が解約率のアサンプションだったので、一番上のような式で書けます。式中に登場するｖ(x)は１件当たりのＶ率を示すもので、ここでは簡単のために所与の関数としております。保険負債の実際の値は、ｙが０、１の変数だったことを思い出すと、上から二つ目の式のように書けます。誤差の定義から、それの差分を取って２乗したもの、赤枠で囲った所が、今回考えたい誤差になります。ここをプロセス、パラメータ、モデルの三つの要素に分解するということが、主要な目標です。また、このケーススタディデータでは、とあるオープンデータを使用します。そのデータの概要は次のスライドで説明します。

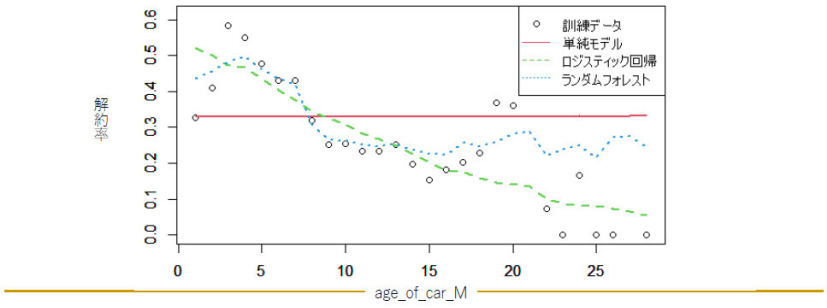
Ⅲ. ケーススタディ（負債評価への応用）

データと予測モデルの説明

- データの主な変数

gender	男性:1、女性:0	car_power_M	車両のパワー
age_client	顧客の年齢	retention	更新:1、非更新:0
age_of_car_M	車両を購入してからの年数		

- 解約率のグラフ：　訓練データとモデル（$\hat{f}(\mathbf{x})$）予測値、変数＝age_of_car_M）



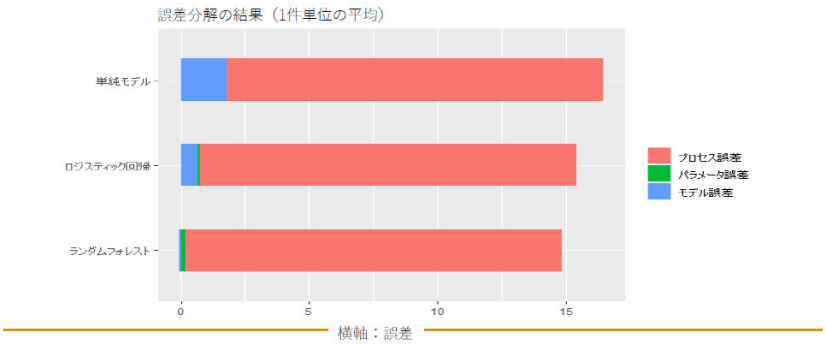
10

データとモデルのイメージをつかんでいただくために、一つのグラフを紹介しております。データの中に、age_of_car_M という、車両を購入してからの年数という変数があります。それを横軸に取って、縦軸は解約率を示しております。この白い丸が、いわゆるデータから作った解約率、粗解約率に相当するものを示しております。赤い直線が、単純モデルで予測した解約率です。単純平均で作っているので一直線になります。緑がロジスティック回帰で予測した解約率で、直線のような感じで落ちていくものです。最後に青い点線がランダムフォレストのもので、表現力が高いモデルですので、解約率が一旦上がって、下がるというような動きも捉えられています。

Ⅲ. ケーススタディ（負債評価への応用）

誤差分解の結果（1件単位）

- どのモデルでもプロセス誤差が大宗
  - 解約という不確実性が強い事象を扱っているためこの結果は自然
- 実際は、1件単位の誤差よりもポートフォリオ単位での誤差に興味がある
  ⇒次ページで、予測対象の保険負債の合計をとったベースで誤差分解を実施

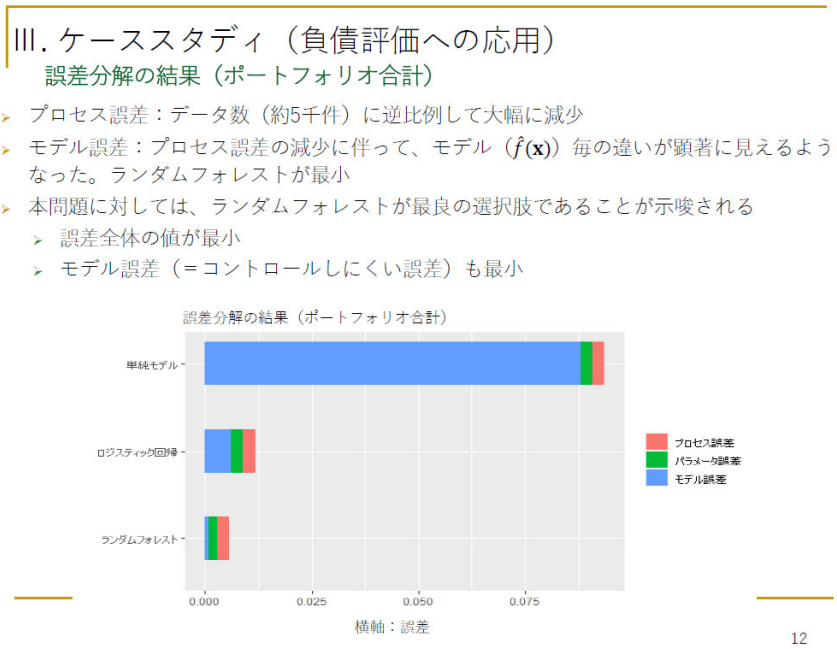


11

前置きが長かったのですけれども、ここでケーススタディの誤差分解の結果を示しております。グラフの見方ですけれども、横軸が誤差で、赤と緑と青で三つの誤差を示しております。縦にモデルを三つ並べ

ています。まず、この誤差分解結果が得られたことが、一つの成果です。例えば、単純モデルをアサンプション作成で使っていたとして、誤差の全体額が横幅全体と見積もれて、そのうちプロセス誤差が後方全てを占めているという知見が、まずここで得られる。これがこのグラフが示している成果です。

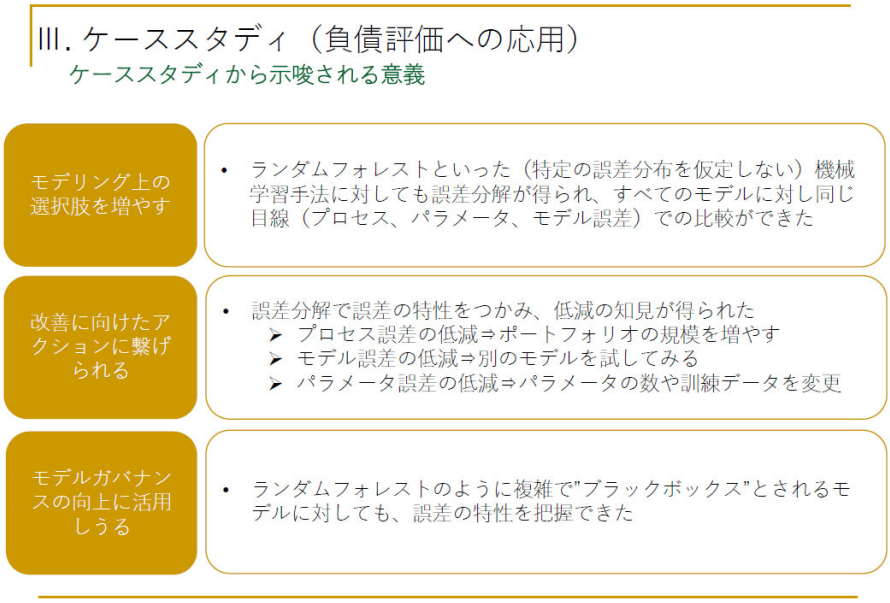
グラフを見ると、どのモデルでも似た結果が得られて、どれもプロセス誤差が大きい状態となっておりますが、これは自然だと思います。今は、契約１件単位の誤差を分解していますけれども、１件の解約を当てるのは、非常に不確実性が強い問題です。例えるならば、コインを投げて裏表を当てるような問題に近いと思います。そのような問題では、プロセス誤差が大きくなることは自然です。



アクチュアリー実務としては、１件単位の誤差を考えるよりは、ポートフォリオを組んで大数の法則を効かせた誤差を考えることが普通だと思いますので、このページで、約 5,000 件のポートフォリオを作って、それに対して誤差を分解した結果を示しています。例えるならば、5,000 枚のコインを投げて、表の出る比率を求めるような問題ですので、プロセス誤差は、ルートｎに比例して非常に小さくなっていることがグラフから見て取れます。

モデルごとの違いも、はっきりと見えるようになりました。仮に単純モデルを分析上使っていたとすると、誤差は大きいです。その誤差の内訳は何かというと、モデル誤差がほぼ全てを占めているということが分かります。これは、最初にグラフを見ていただいたとおり、わざとそのようにしているのですけれども、単純モデルは非常にフィッティングが悪いことから、モデル誤差が大きくなることが納得できると思います。

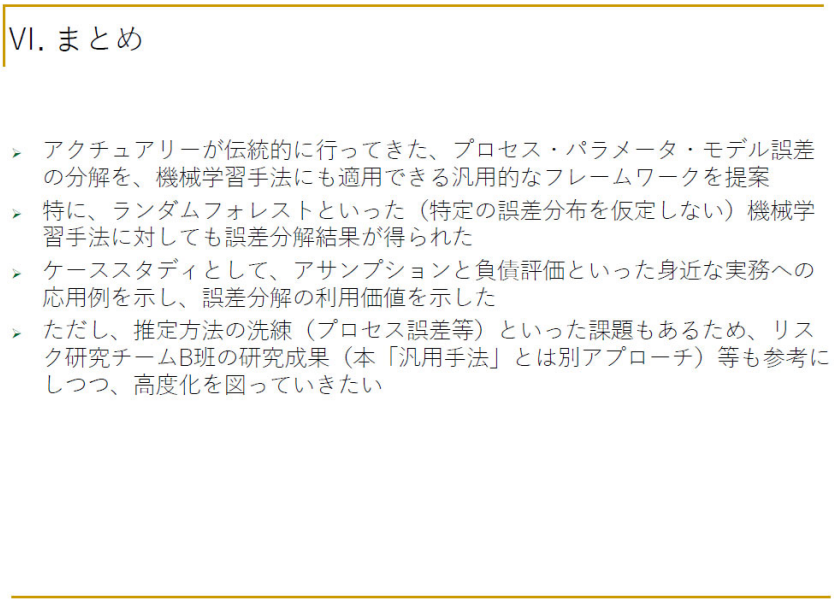
一方で、一番下のランダムフォレストを見てみますと、誤差の全体額も小さいですし、モデル誤差も非常に小さくなっております。ランダムフォレストは、先ほどお見せしたように、表現力が高いモデルで、フィッティングがよかったことの結果だと思います。モデルごとに並べた分析を見ることによって、ランダムフォレストは誤差が小さく、モデル誤差も小さいので、この問題には使いやすいモデルだという知見も得られます。



結果を駆け足で説明してしまいましたが、その結果から得られる意義を、簡単に１枚でまとめております。まず、モデリング上の選択肢が増えるという意義があると思います。先ほど結果を何回も見ていただきましたけれども、特定のモデルに制限されず、三つのモデルの誤差の内訳を見ることができました。横並びでモデルの誤差に関する特性を比較できたので、モデリングに活用できると思います。

さらに、改善に向けたアクションにつなげられるという意義もあると思います。誤差分解をして誤差の特性をつかむことによって、誤差を減らすために、どのようなアクションを執ればいいかということが見えてきます。先ほど実際にお見せしたとおり、プロセス誤差が非常に大きければ、ポートフォリオの規模を増やすことが有効です。モデル誤差が大きいという結果が出れば、先ほど単純モデルとランダムフォレストを比べたように、別のモデルを使ってみて、どのようになるかを試すことができると思います。

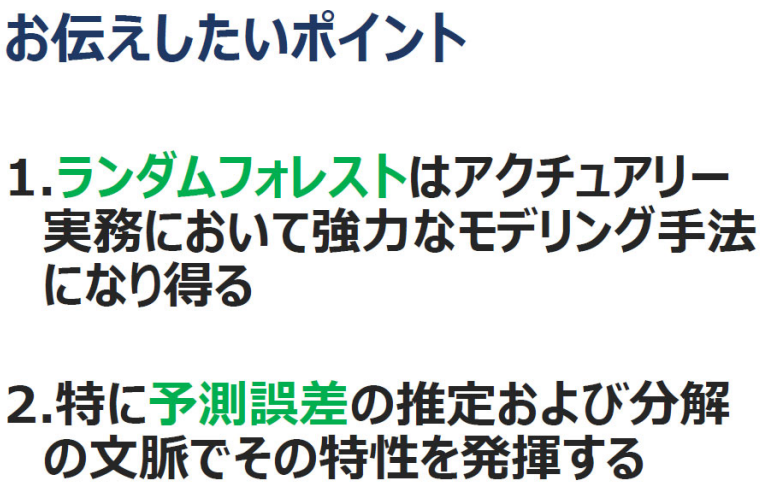
また、モデルガバナンスの向上に活用しうるという意義もあります。ランダムフォレストのように複雑で、一般的にブラックボックスとされるようなモデルに対しても、誤差の特性などを把握できたので、一定のモデルの透明性や弱点の把握などにつながると思います。



最後にまとめです。アクチュアリーが伝統的に行ってきたプロセス、パラメータ、モデル誤差の分解を、機械学習手法にも適用できる汎用的なフレームワークを提案いたしました。身近な負債評価とアサンプションというケーススタディを示して、利用価値も示せたと思います。これが使えそうだと共感していただければ、大変うれしく思います。

ただし、途中でお話したように、プロセス誤差の推定方法の洗練など、課題もたくさん残っています。次に発表のリスク研究チームＢ班が、我々の汎用手法とは別のアプローチで誤差分解に取り組んでいますが、そちらの手法ではこのような問題を解決できている部分もあります。他のチームの研究成果も参考にしながら、手法を高めていければと考えております。私からは以上です。ありがとうございました。

岩沢　高橋さん、ありがとうございました。４人目のプレゼンターは AKUR8 の藤田さんで、タイトルは「ランダムフォレスト特有の予測誤差分解の研究」です。それでは、お願いいたします。

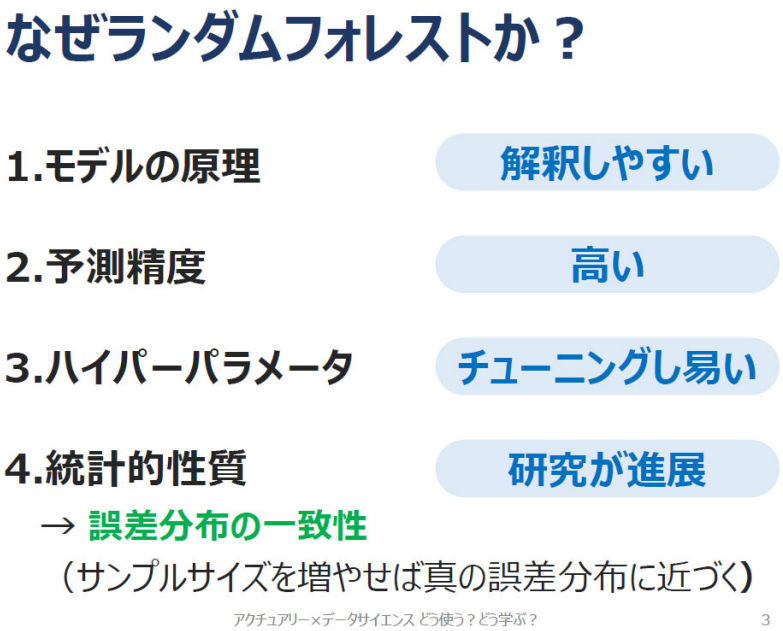


藤田　ありがとうございます。AKUR8 の藤田と申します。よろしくお願いします。「ランダムフォレスト特有の予測誤差分解の研究」という私のパートでは、お伝えしたいポイントが二つあります。

まず一つ目が、ランダムフォレストは、アクチュアリー実務において強力なモデリング手法となりうるということです。

二つ目は、特に予測誤差の推定および分解の文脈で、その特性を発揮することです。この予測誤差に関しては、先ほどの高橋さんのパートで一部課題があるとおっしゃっていましたが、そちらを担保できるような形でランダムフォレストを活用できることをお伝えしたいと思います。

本セッションでは今まで、IML、予測誤差分解と紹介がありましたが、決してそれらと本研究が独立しているわけではなく、密接に関連していると考えています。ＩＭＬに関連させると、ランダムフォレストは機械学習モデルの中でもモデルの原理がわかりやすく、解釈可能性を担保するための技術も多いです。そして予測誤差分解に関しては、ランダムフォレストのユニークな特性が発揮されます。



なぜランダムフォレストなのかということで、改めて簡単にまとめてみました。次の四つの理由があると考えます。まず、複数の決定木を集約するというベースの考え方が単純であるため、モデルの原理が解釈しやすいこと。二つ目に、予測精度が高いことが知られており、三つ目は、ハイパーパラメータのチューニングが他のモデルに比べて容易であること。ここで他のモデルと言っているのは、今回はご紹介を詳しくできていませんが、ニューラルネットワークやＸＧBoost などのモデルを念頭に置いていただければと思います。四つ目として、ランダムフォレストの統計的性質に関する研究が、最近大きく進展してきています。特に誤差分布の一致性に関する研究が、予測誤差の推定に非常に有用であると考えています。誤差分布の一致性とは何かというと、サンプルサイズを増やせば真の誤差分布に近づくというもので、予測誤差の文脈では非常にうれしい性質となっています。つまり、不偏性のようなものではなくて、単にデータのサンプルサイズを増やせば良いということです。

なぜGLMではないのか？

GLMはアクチュアリーにとって馴染みのあるモデル

- 予測精度はランダムフォレストに劣るものの
- 広く知られ原理的な解釈はしやすく
- モデリングもしやすい、一方で

誤差分布の一致性が必ずしも保証されない

（サンプルサイズを増やしても真の誤差分布に近づく保証なし）

ここで、ふと疑問を持たれる方もいらっしゃるかもしれませんが、なぜＧＬＭではないのでしょうか。他の研究でも、この年次大会でも、ＧＬＭという単語を見た方もおられると思います。ＧＬＭは今でもアクチュアリーの間で非常になじみがあって、例えばプライシングやリザーブなど、さまざまなプロジェクトに使われているかと思います。予測精度はランダムフォレストに劣るものの、広く知られて原理的な解釈はしやすく、モデリングもしやすいです。一方で、何が問題かというと、予測誤差の推定という文脈では、誤差分布の一致性が必ずしも保証されないということがあります。アクチュアリーはリスクを扱うので、これは重大な部分だと考えています。つまり、サンプルサイズを増やしても真の誤差分布に近づく保証がないということです。我々は、ＧＬＭと同じぐらい実務で活用できる機械学習のモデルを提唱できればと考えています。

ランダムフォレストへの注目度は？

アクチュアリー界隈では低いかもしれない

- ニューラルネットワークと勾配ブースティングは高い
- データ分析コンペの影響？
- CART（決定木）はやや高い

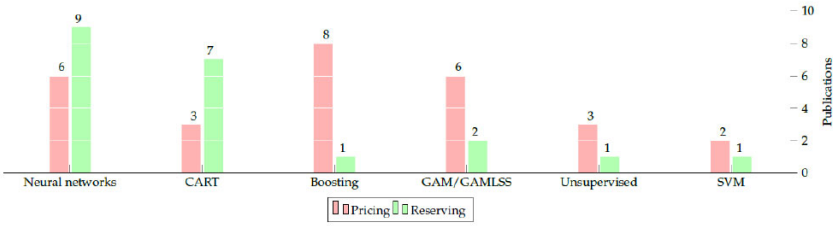
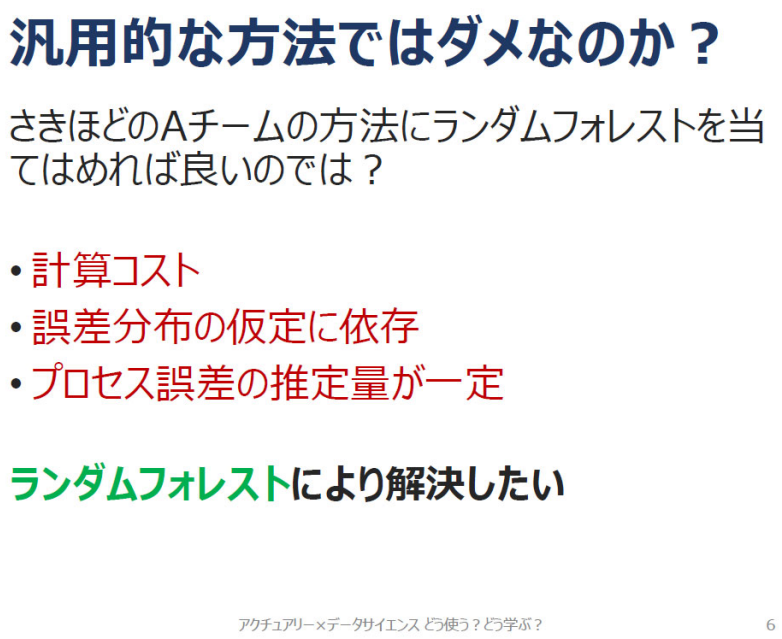


Figure 3. Number of publications by model.

実際のランダムフォレストの注目度は、一体どうでしょうか。こちらにグラフを用意していますが、損害保険分野での数理業務におけるモデルの利活用について、2015 年から 2020 年 8 月までの 77 件の論文を調査した結果です。横軸に各モデルが、ニューラルネットワークや決定木、Boosting などと書いてあって、縦軸に、赤がプライシング業務、緑がリザービング業務で、どれだけ使われているかがカウントされています。ニューラルネットワークと勾配 Boosting が、比較的多いことが分かると思います。これは推察ですが、Kaggle 等のデータ分析コンペティションの使用実績の影響が、一つの理由として挙げられるかもしれません。

さて、この中で、ランダムフォレストの名前がそもそも出てきていません。ランダムフォレストの注目度はそこまでないように見て取れますが、CART すなわち決定木も多く、そのアンサンブルモデルであるランダムフォレストの発展の余地は十分にあると考えていまして、本研究の活動が、これに向けたよい皮切りになればと思っています。



さて、「ランダムフォレスト特有の」と申し上げましたけれども、汎用的な方法ではだめなのでしょうか。先ほど高橋さんからご説明いただいたパートです。予測誤差推定について、先ほどの発表内容にもあったとおり、計算上のコストや誤差分布の仮定への依存性、プロセス誤差の推定量がコンスタントなどの課題が残っていて、これらに対してランダムフォレストの特性が生かされて、解決できると考えています。

予測誤差の文脈で発揮する特性

ノンパラメトリック性

- 特定の分布を想定しないため、データに応じて柔軟にモデリングできる

多数の独立した回帰木による恩恵

- 学習過程で副次的に得られた IB と OOB が活用でき、計算コスト削減につながる

では、ランダムフォレストは、予測誤差評価でどのようなところが発揮できるのかということで、特性を挙げてみました。１点目のポイントは、ノンパラメトリック性です。ランダムフォレストはある種のノンパラメトリックモデルであるため、柔軟にデータに当てはめることができると考えています。２点目としては、学習の過程でＩＢ、ＯＯＢといった副産物を生成するので、それらを活用することで計算コストの削減につながると考えています。

In-Bag と Out-Of-Bag

IB (In-Bag)

- ある回帰木のブートストラップ標本に使用されたサンプル

OOB (Out-Of-Bag)

- ある回帰木のブートストラップ標本に使用されなかったサンプル
  - 学習データに占める OOB は
    $$\left(\frac{n-1}{n}\right)^n \approx \exp(-1) \approx 36.8\%$$
  - OOB予測により汎化性能を評価することができる

ランダムフォレストには、In-Bag と、Out-Of-Bag というユニークな概念があります。あるサンプルが得られない確率は、ここにあるような数式で評価でき、学習サンプルから復元抽出を行う上で、36.8%はブートストラップ標本として選択されません。従って、学習サンプルは、ブートストラップ標本に使われるものと使われないものの２種に分かれます。それぞれを、ＩＢとＯＯＢと呼びます。各学習サンプルにつ

いてＯＯＢの木を集めることで、その対象となるサンプルについて予測を行うことができるのですが、これを「ＯＯＢ予測」と呼び、テストデータに対する汎化性能を測ることができます。

- ここまでがモチベーションとランダムフォレストの特性の紹介
- 以降、本研究のメインテーマである、ランダムフォレストを活用した予測誤差分解方法を紹介

ここまでが、本研究のモチベーションと、ランダムフォレストの予測誤差分解における特性の紹介でした。以降は、本研究のメインテーマである、ランダムフォレストを活用した予測誤差分解方法をご紹介したいと思います。

学習データの分類

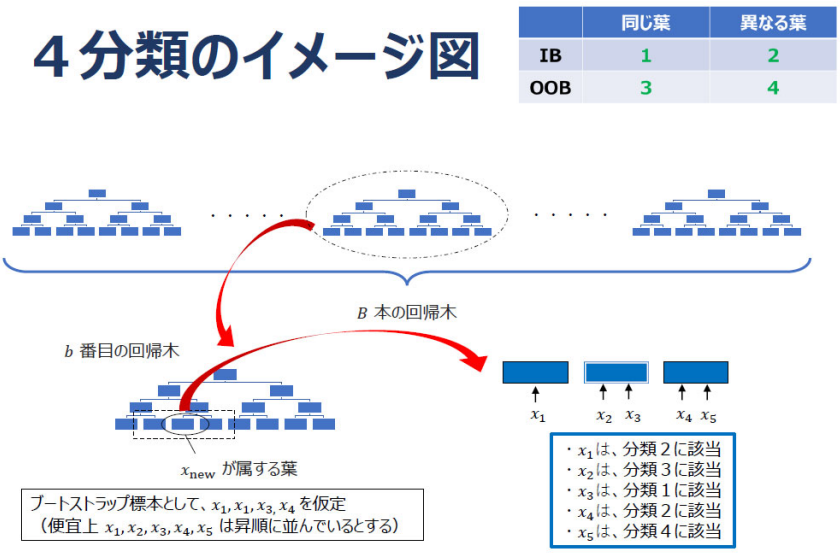
新たな観測対象 $(x_{\mathrm{new}}, y_{\mathrm{new}})$ が与えられたとき、各回帰木に対して各学習サンプルは、次のように分類できる

- IB か OOB か
- $(x_{\mathrm{new}}, y_{\mathrm{new}})$ と同じ葉に属するか異なる葉に属するか

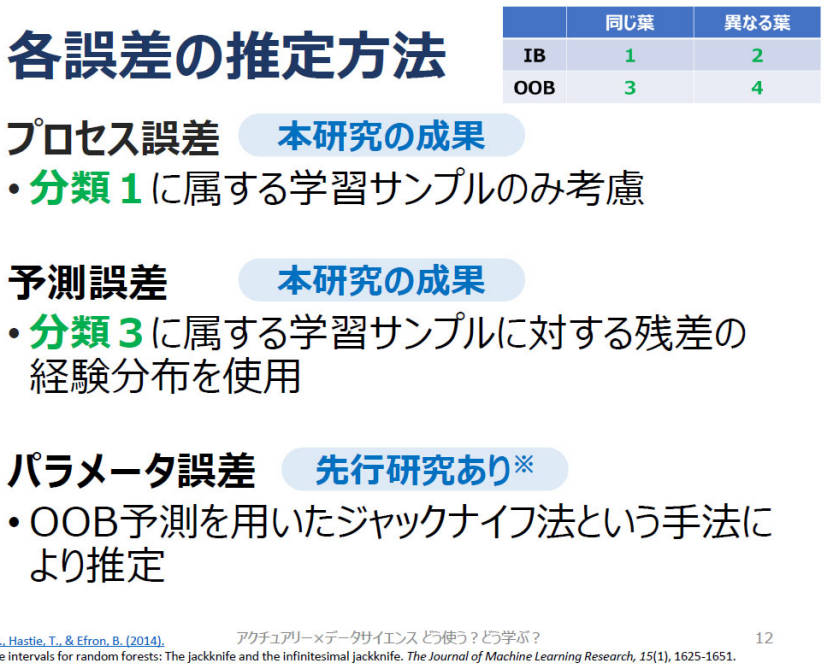
	同じ葉	異なる葉
IB	1	2
OOB	3	4

さて、学習データの分類を考えてみます。新しい観測対象(xnew、ynew)が与えられたときに、ランダムフォレストを構成する各回帰木に対して、各学習サンプルは次のように分類することができます。まず、ＩＢか、ＯＯＢかというパターンです。そして、(xnew、ynew)と同じ葉に属するか、異なる葉に属するかというところです。これによって２×２で４パターンに分かれます。ここで番号を１、２、３、４とつけ

ておきます。



４分類のイメージ図ですが、Ｂ本の回帰木があって、その中からｂ番目の回帰木に着目し、そのうちの一つの葉に xnew が属する葉があったとします。仮にブートストラップ標本として x1、x1、x3、x4 を仮定したときに、この丸で囲んである所には x2 と x3 が入っていて、その左隣には x1、右隣には x4、x5 が入ったとします。そのときにどのように考えるかというと、x1 は、ＩＢですが、実際の xnew が入っている葉とは違う所なので、分類２に該当。x2 は、ＯＯＢで異なる葉なので分類３番。x3 は、ブートストラップ標本に使われていて、xnew と同じ葉に属しているので、分類１に該当するという具合になります。

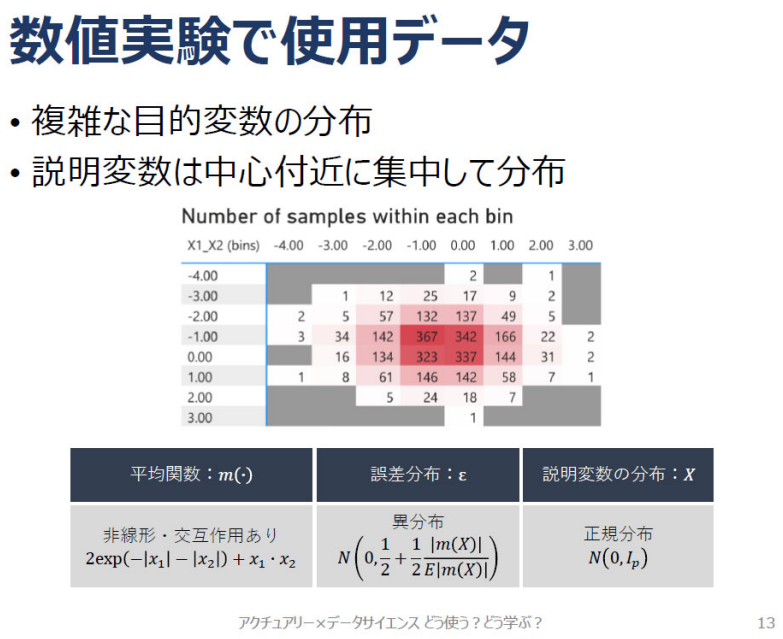


ここで、各誤差の推定方法です。ここでは数式は一切割愛して、どのような感じでそれぞれの誤差を求めていくのか、ご紹介したいと思います。

プロセス誤差については、簡単に言うと、分類１に属する学習サンプルのみを考慮します。分類１は、ブートストラップ標本に使われていて、新しい観測対象と同じ葉に属するものなので、本来であれば、ランダムフォレストで予測を行うために使うサンプルのみを使って、プロセス誤差を評価できます。

予測誤差は全体の誤差ですけれども、こちらは分類３ですね。先ほどＯＯＢ予測のときに見たように、ブートストラップ標本に使われなかったものについて、それを集約して学習サンプルに対する残差の経験分布を使用することで、予測誤差を定量化することができます。

そして、パラメータ誤差ですが、こちらは、ＯＯＢ予測を用いたジャックナイフ法という手法により推定することができます。ジャックナイフ法は、こちらでは割愛しますけれども、先行研究を参考にしています。プロセス誤差と予測誤差については本研究の成果になっています。



省略しましたけれども、実際に定式化した推定量について、定量的にどれだけ正確なものなのかということを、数値実験を用いて確認しました。ここでは人工データと実データを用いて、定量的な検証を行いました。そのうちの一部をざっくりとご紹介したいと思いますが、概ねうまくいっていることを確認しています。

人工データの仮定を下にまとめていますが、複雑な場合を想定して、それでもある許容範囲内でうまくいっているかどうかを確認しています。例えば、これは真の関数になっていますが、交互作用が入っていたり、誤差分布については、正規分布の分散の項が特徴量に応じて異なっていたりという具合です。

この図は、どのような感じでサンプルが分布しているのかを示しています。二つの特徴量Ｘ1、Ｘ2というものがあって、縦軸・横軸にプロットしたときのサンプル数のヒートマップを表しています。サンプル数は中心部分が多く、端に行くほど少なくなっていることが分かります。つまり、端に行くほどデータ数が少なくなっているので、予測も難しくなっていくと予想されます。

目的変数の説明変数ごとの比較

目的変数において、ランダムフォレストの予測値（左）と真の値（右）を比較（近ければ近いほど良い）

- 中心付近はモデルの予測は正確
- 縁における大きな変動は捉えきれていない

Y_estimated

X1_X2 (bins)	-4.00	-3.00	-2.00	-1.00	0.00	1.00	2.00	3.00
-4.00					-0.68		-3.58	
-3.00		2.88	2.44	1.08	-0.70	-2.89	-3.93	
-2.00	0.95	2.68	1.90	0.96	-0.10	-1.90	-2.43	
-1.00	0.88	1.12	1.01	0.99	0.72	-0.33	-1.06	-2.69
0.00		-0.37	-0.24	0.58	1.08	0.95	0.74	-1.45
1.00	-5.28	-2.99	-1.90	-0.30	1.03	1.86	2.36	0.52
2.00			-2.62	-0.81	1.22	2.77		
3.00					1.74			

Y_true

X1_X2 (bins)	-4.00	-3.00	-2.00	-1.00	0.00	1.00	2.00	3.00
-4.00					-2.25		-8.19	
-3.00		6.90	3.55	1.08	-1.11	-3.20	-4.66	
-2.00	6.42	3.58	2.06	0.94	-0.42	-1.99	-4.61	
-1.00	0.31	1.04	0.97	1.11	0.65	-0.30	-1.24	-2.34
0.00		-0.87	-0.23	0.59	1.07	0.97	1.30	0.69
1.00	-5.89	-2.69	-1.87	-0.38	0.93	2.00	2.27	4.67
2.00			-3.65	-0.91	1.60	2.28		
3.00					2.16			

まず、目的変数の説明変数ごとの比較ですけれども、ベースとなるランダムフォレストによる推定結果を表しています。こちらは、まだ予測誤差は関係ありません。右のグラフが真の値で、左の値が推定値になっています。両者は数値が近ければ近いほどいいのですが、中心部分では、よく見ると推定値はほぼ正確ですけれども、端の方では、やはりサンプル数が少ないので、大きな変化を捉えることができなかったようです。

予測誤差の説明変数ごとの比較

予測誤差において、推定値（左）と真の値のシミュレーション値（右）を比較（近ければ近いほど良い）

- サンプル数が多い中心付近では推定が正確
- サンプル数が少ない縁では相対的に乖離がある

Prediction_error_estimated

X1_X2 (bins)	-4.00	-3.00	-2.00	-1.00	0.00	1.00	2.00	3.00
-4.00					1.04		2.17	
-3.00		1.86	1.81	1.30	1.03	1.41	2.42	
-2.00	1.78	1.55	1.44	1.09	0.88	1.18	1.71	
-1.00	1.07	1.07	1.11	1.11	0.87	0.84	1.33	4.63
0.00		1.03	0.95	0.88	1.01	1.04	1.63	4.40
1.00	4.19	2.62	1.59	0.84	1.11	2.18	3.06	5.91
2.00			1.96	0.91	1.32	2.63		
3.00					1.31			

Prediction_error_MC

X1_X2 (bins)	-4.00	-3.00	-2.00	-1.00	0.00	1.00	2.00	3.00
-4.00					2.51		17.32	
-3.00		6.64	2.92	1.29	1.33	2.82	4.87	
-2.00	8.12	3.18	1.69	1.07	0.82	1.60	2.25	
-1.00	2.07	1.17	1.06	1.09	0.90	0.80	1.19	2.35
0.00		1.13	0.81	0.89	1.09	1.04	1.39	1.70
1.00	8.80	2.58	1.55	0.80	1.06	1.76	2.52	11.09
2.00			2.93	1.17	1.35	4.17		
3.00					2.43			

そして、メインの予測誤差の部分ですけれども、種々の誤差のうち、予測誤差から提案手法の精度を評価したものになっています。こちらも二つの図があって、右が真の値をシミュレーションした値、左は推定値となっています。こちらも数値が近ければ近いほどいいものですが、中心部では正確な推定が比較的行われていることが分かるものの、サンプル数が少ない端の方では、相対的に乖離があることが分かって

います。

パラメータ誤差の比較

パラメータ誤差において、極小ジャックナイフ法という手法による推定値（左）と真の値のシミュレーション値（右）を比較（近ければ近いほど良い）

- サンプル数が多い中心付近では推定が正確
- サンプル数が少ない縁では相対的に乖離がある

Parameter_error_estimated

X1_X2 (bins)	-4.00	-3.00	-2.00	-1.00	0.00	1.00	2.00	3.00
-4.00					0.66		0.01	
-3.00		0.01	0.25	0.21	0.14	0.14	0.01	
-2.00	8.15	0.18	0.11	0.04	0.05	0.06	0.21	
-1.00	0.41	0.05	0.07	0.06	0.04	0.06	0.24	5.19
0.00		0.01	0.05	0.05	0.05	0.05	0.47	5.10
1.00	2.14	0.15	0.12	0.04	0.07	0.23	0.47	7.04
2.00			0.11	0.03	0.27	0.62		
3.00					0.01			

Overall average: 0.0871

Parameter_error_MC

X1_X2 (bins)	-4.00	-3.00	-2.00	-1.00	0.00	1.00	2.00	3.00
-4.00					0.42		0.62	
-3.00		0.42	0.27	0.11	0.15	0.28	0.42	
-2.00	1.16	0.36	0.12	0.05	0.06	0.15	0.21	
-1.00	0.72	0.11	0.05	0.04	0.04	0.06	0.12	0.50
0.00		0.11	0.06	0.04	0.04	0.05	0.15	0.50
1.00	1.04	0.28	0.15	0.06	0.05	0.11	0.22	1.79
2.00			0.31	0.12	0.13	0.33		
3.00					0.59			

Overall average :0.0676

分解した要素のうち、パラメータ誤差についても同じような現象が起きていて、中心部分では正確ですが、サンプル数が少ない縁では、相対的に乖離があります。

プロセス誤差の比較

プロセス誤差において、補正項を加えた推定値（左）と真の値のシミュレーション値（右）を比較（近ければ近いほど良い）

- サンプル数が多い中心付近でもやや負のバイアスあり
- サンプル数が少ない縁では相対的に乖離がある

Process_error_estimated

X1_X2 (bins)	-4.00	-3.00	-2.00	-1.00	0.00	1.00	2.00	3.00
-4.00					3.24		1.93	
-3.00		1.40	2.44	1.69	1.31	1.15	1.14	
-2.00	13.02	2.20	1.30	1.01	0.78	0.89	1.49	
-1.00	2.22	0.97	1.01	1.04	0.77	0.63	1.10	5.92
0.00		0.77	0.79	0.76	0.94	0.85	1.59	6.69
1.00	6.74	1.28	0.99	0.62	0.88	1.55	2.37	11.17
2.00			0.89	0.65	1.23	2.86		
3.00					1.65			

Overall average: 0.94364

Process_error_MC

X1_X2 (bins)	-4.00	-3.00	-2.00	-1.00	0.00	1.00	2.00	3.00
-4.00					1.67		4.21	
-3.00		3.31	2.15	1.14	1.10	2.20	2.84	
-2.00	2.95	2.29	1.55	1.01	0.76	1.43	1.87	
-1.00	1.07	1.04	1.01	1.04	0.85	0.74	1.00	1.19
0.00		0.97	0.74	0.84	1.05	0.98	1.20	1.05
1.00	3.20	2.13	1.40	0.74	1.00	1.60	2.07	3.05
2.00			2.21	0.99	1.17	2.48		
3.00					1.62			

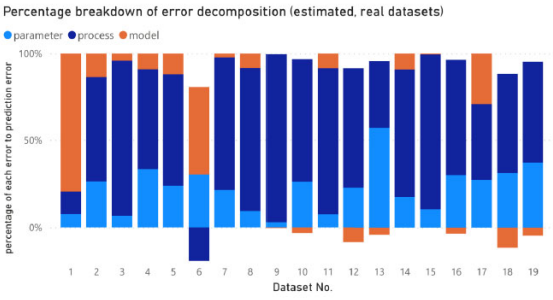
Overall average : 0.99853

プロセス誤差については、他のものに比べて課題がありそうなのですが、中心でも若干の乖離があります。うまくいっていないのではないかと思われるかもしれませんが、かなり複雑な人工データを仮定しているので、それにしては、許容的なものになっているのではないかと考えています。

実データにおける予測誤差分解

複数の実データに対する各誤差の全体に対する比率

- 多くの場合、プロセス誤差50%以上、モデル誤差20%以下
- 改善すべき誤り（負のプロセス誤差、モデル誤差）



また、もう一つ、こちらはインターネット上で入手できる実データを用いて、提案手法を適用して予測誤差分解した結果です。全部で 19 個のデータを使っています。幾つかの例外はあるものの、予測誤差のほとんどは、紺色で表したプロセス誤差によるものであり、オレンジ色で示したモデル誤差によるものはほとんどないことから、推定上はうまく分解ができていると考えています。しかし、プロセス誤差やモデル誤差が誤ってマイナスとなっているケースも幾つかあったり、実際のデータに適用した場合に提案手法の推定がどの程度正確であったかは、疑問もやや残っているケースもあります。

実験部分を要約すると、課題はあるけれども、提案手法はうまく機能していることが観察できたということをお伝えしたかった次第です。

提案手法の考察

- 追加の反復計算が不要なため、計算コストの低減が期待される
- 誤差分布の仮定がなく、学習データに応じた誤差分布が構築可能。またモデル誤差が小さくなることが期待される
- プロセス誤差の推定を新たな観測対象ごとに可能

以上は、数値実験で一定程度確認できた

最後に、提案手法の考察です。計算コストについて定量的にはお示ししていませんが、追加の反復計算が不要なので、計算コストの低減が期待されます。また、誤差分布の仮定がないので、学習データに応じた誤差分布の構築が可能となっています。プロセス誤差の推定を新たな観測対象ごとに可能であり、先ほどの汎用的な手法の課題をクリアしていることも分かっています。これらのほとんどは、数値実験で一定程度確認できています。

今までをまとめますと、ランダムフォレストを用いた予測誤差分解手法を示して、ランダムフォレストがアクチュアリーにとって強力な手法になること、特に予測誤差分解の文脈でその特性が発揮されることを、定性的に、定量的に、簡単にですが、お示ししました。ご清聴ありがとうございました。

岩沢　藤田さん、ありがとうございました。他のパネリストの皆さんも、ありがとうございました。前半は、主に「どのように使うか」について語っていただきましたが、ここからは、主に「どのように学ぶか」について語ってもらいます。プレゼンをしていただいた４人の方に、そのままディスカッションに参加していただきます。どうぞよろしくお願いいたします。

ディスカッション

The Institute of Actuaries of Japan
Think the Future, Manage the Risk

話題１：データサイエンスの面での自己紹介を兼ねて，データサイエンスをいつからどれくらい学んでいるかを大まかに教えてください.

※ 学習の仕方に関する経験を語ってもらえる機会はこのあともあると思うので，この話題は簡潔にお願いします☺

まずは話題１ですが、データサイエンスの面での自己紹介を兼ねて、データサイエンスをいつから、どれくらい学んでいるかを、大まかに教えていただきたいと思います。では、鈴木さんからお願いできますか。

鈴木　私は、大学時代は工学部で、統計学の研究室に所属しておりまして、実験計画法という分野で卒業研究をしました。データサイエンスを勉強しだしたのはアクチュアリー会の正会員になってからで、専らアクチュアリー会の委員会活動を通じて学んできたという感じです。

最初はデータサイエンスを学ぼうと思って始めたのではなくて、データサイエンス・ワーキンググループより前にＡＳＴＩＮ関連研究会に所属しておりまして、そこでＣＡＳの Predictive Modeling の教科書の翻訳を担当いたしました。教科書の中身としては、機械学習より少し前といいますか、ＧＬＭやＥＤＡ

などの内容が基本でしたけれども、アクチュアリー試験の基礎科目の数学を超えて、より高度な数学を勉強しようというモチベーションがその活動を通じて高まって、今もこのワーキンググループで活動しております。

岩沢　ありがとうございます。では、浅芝さん、お願いします。

浅芝　私は、大学時代は、科学史・科学哲学という分野を専攻しており、当時、データサイエンスや機械学習を体系的に学んだことはありませんでした。データサイエンスの勉強を本格的に始めるきっかけになったのは、2021 年度に受講したデータサイエンス専門講座です。また、講座を受けるかたわらで、kaggle というウェブプラットフォームで開催されているデータサイエンスコンペティションに、少しだけ挑戦してみました。

岩沢　ありがとうございます。会場の方からすると、パネリストのみなさんが座っている並び順と違ってしまいますが、次に、高橋さん、お願いいたします。

高橋　私は、５、６年前ぐらいからデータサイエンスを学び始めました。学生の頃は物理を専攻していましたが、数値計算を使ってシミュレーションをやるような研究を行っていましたので、プログラムを書いて分析することに抵抗はありませんでした。当時、データサイエンスという言葉が世間でも非常に流行っていて、面白そうだな、自分の得意分野とマッチしそうだなという感じで興味を持って、勉強し始めたというのがきっかけです。ただ、ふだんの業務で専門的にデータサイエンスを活用しているわけではないです。

　このワーキンググループには 2020 年から参加させていただいておりまして、暇な時間を見つけて研究活動を行っているという状況です。

岩沢　ありがとうございます。それでは、藤田さん、お願いします。

藤田　はい。私は元々、学士・修士ともに機械学習関連の研究をしていて、ある意味では、データサイエンスに関わっていたといえば関わっていたと思います。新卒で会社に入社してから、５年ほどブランクがありましたけれども、2016 年、2017 年あたりから、アクチュアリー会のＡＳＴＩＮ関連研究会やデータサイエンス・ワーキンググループに所属するようになって、本業のかたわら、データサイエンスのプロジェクトに関与することになりました。

　このような対外活動の成果を実際の本業に生かしたり、業務に還元する機会も増えていって、今はインシュアテック企業で働きながら、業務で主にデータサイエンス系のプロジェクトに従事しているという感じです。

岩沢　ありがとうございます。お話しいただいたところからすると、藤田さんが、一番データサイエンスを仕事に使っているという感じですね。

話題2：データサイエンスを学ぼうと考えた理由や，その後モチベーションを保てている理由などを簡潔に教えてください．

※ たとえば，業務との関係や学んだことが役に立っているかといった観点で答えてもらってもよいと思います．「簡潔に」としましたが，この話題はたくさん語りたい，という方はたくさん語ってもらってもかまいません．以下の話題も同様です．

岩沢　それでは、話題２として、データサイエンスを学ぼうと考えた理由や、今の自己紹介と重なる部分はあるかもしれませんが、その後モチベーションを保てている理由などを、簡潔に教えていただければと思います。では、鈴木さんから、お願いします。

鈴木　先ほど藤田さんは、インシュアテック企業で専門的にデータサイエンスをされているということでしたが、私は、恥ずかしながら、業務の中では一切データサイエンスに関わる仕事はしておりません。専ら委員会活動で勉強しているという感じです。

モチベーションを保てている理由ですが、私個人としては、データサイエンスのスキルが、将来、アクチュアリーの必須スキルになることを確信しているからです。私の想像ですけれども、現在で言うところの責任準備金の経済価値評価などに近いような位置づけになるのではないかと思っています。例えば、今、生保のアクチュアリーの方で、保証とオプションの時間価値とは何か、どのように計算するかを知らない人はあまりいないと思いますが、20 年前はあまり知れ渡っていなかった。そのような感じで、データサイエンスについても、20 年後や 30 年後に私がもっと実務の中で高い立場になったときに、今まで勉強してきたことが必ず役に立つと確信していまして、それでデータサイエンスを勉強するモチベーションを保っております。

岩沢　ありがとうございます。では、浅芝さん、よろしいですか。

浅芝　私がデータサイエンスに関心を持ったのは、過去の年次大会で、特に強化学習を応用したものでしたが、機械学習関係の発表が、朝一番にもかかわらず多くの聴講者を集めているのに驚かされたことでした。学習に対するモチベーションの観点では、単純に楽しいから勉強が続いているということももちろんありますが、アクチュアリー会のデータサイエンス関連基礎調査ワーキンググループや、所属会社の有志による勉強会にも参加しており、他の方と一緒に学ぶことでモチベーションを維持しやすい環境ができていると感じています。

データサイエンスを実際の業務に役立てる方法については、所属会社での勉強会でも、ちょうどアイデアが出ているところです。たとえば、発生率や解約率のアサンプションを設定するときに、「まずはブラックボックスモデルでもいいから精度の高い予測モデルをとりあえず作ってみて、それを解釈する」というアプローチが、ベスト・エスティメイト・アサンプションを群団で分けるときの切り口や水準を決めるときの参考になることもあるのではないかと思っています。

岩沢　ありがとうございます。それでは、高橋さん、お願いします。

高橋　きっかけは先ほど話してしまったので、主にモチベーション面でお話しさせていただこうと思います。業務で専門的に活用しているわけではないと言いましたが、学んだことが役に立っていると感じることはたくさんありまして、そのようなところがモチベーションになっていると感じております。

例えば、私はずっと生命保険分野を経験していますが、アサンプションを作成するときに、ＧＬＭやデータサイエンスのフレームワークを試してみて、アサンプションの精度が向上するか、透明性や客観性は向上するかというような検討をしたりしたことがありました。もう少し機械学習寄りの話では、先ほどの浅芝さんのお話に近いのですが、ランダムフォレストから得られるＰＦＩという特徴量重要度を見て、新しい知見が得られないか、人間の判断を少なくできないかなど、実験的な要素が強いようなことを考える機会もありました。このように非常に学んだことが役に立っていると感じていて、そこはモチベーションになっています。

岩沢　ありがとうございます。では、藤田さん、お願いします。

藤田　データサイエンスを学ぼうと考えた理由は、元々データサイエンスなどの理論的なこと、分析的なことが自分の興味の対象だということもありますし、単純に面白いからです。また、時代の潮流として、やっておいた方がいいのかなということもあります。

それらが正直なところですが、その後モチベーションを保てている理由としては、今までの発表を聴いて皆さんも感じていただいていると信じていますが、実際に新たな視点や分析の幅が広がることが多いと思うためです。データサイエンスは、単に「この新しいモデルを当てはめてみよう」というモデリングの手法だけではなくて、分析プロセス全体を対象とするものなので、応用面は多岐に渡り、参考になると思います。

私の個人的な話では、前職の話ですけれども、別の分析系のチームがあって、そこでデータサイエンス関連のプロジェクトを通して、新たなプロジェクトを立ち上げてチーム間のコラボレーションが生まれたり、新たな付加価値も生まれたりして、非常によかったと思っています。

話題3：データサイエンスのおすすめの学習方法を教えてください.

※ どういう本やどういう内容を学ぶのがおすすめかといった観点や，（話題1で語り切れなかった）「自分はどう学んできたか」という話や，コンピュータ環境の整え方に関することなど，いろいろな切り口があると思うので，ごく簡潔に，しかし，2回ずつくらい答えてもらいましょうか．これが中心的な話題なので，3回以上答える方がいてもかまいません.

岩沢　ありがとうございます。いろいろと聞きたいところですけれども、まだ話題がたくさんあるので、次の話題３に移ります。これがメインの話題です。どのように学ぶかというテーマでして、パネリストの皆さんは、勉強を始めてからの長さもさまざまなのですが、それぞれの経験を通して、データサイエンスのお勧めの学習方法をお話しいただければと思います。これは、学習年数としては浅芝さんが一番短いので、まずは浅芝さんにお願いします。

浅芝　はい。私は、アクチュアリー会で開講していて、岩沢先生が講師をされている「データサイエンス専門講座」に参加してみることをお勧めしたいと思っております。以上です。

岩沢　ありがたいですね。その講座の案内はスライドの最後にもあります。専門講座の受講経験者ということでは、ここに高橋さんもいらっしゃるので、追加でコメントをいただければと思います。

高橋　はい。私は 2019 年に参加させていただきましたが、ある程度知識を持ったような状態で参加しました。ただ、アクチュアリー向けの講座ということで、改めて学んだ、気づいたことは非常に多かったと思っております。講師は専門家の方ですので、講義で生の声を聞いて、本では得られないような気づきも非常に多かったと思っております。ですから、これから学ぼうとする人も、ある程度学んだことがある人も、どちらにもお勧めできると思っております。

岩沢　ありがとうございます。それでは、鈴木さん、お願いできますか。

鈴木　私からは、ともかく実装してみることをお勧めします。R や Python のパッケージで、代表的なデータサイエンスの手法は簡単に実装できます。しかも、R や Python は無料です。インターネットを探せばプログラムコードはたくさん見つかりますし、てまえみそですけれども、我々ワーキンググループのジャーナルの報告書にはたくさんプログラムコードが載っていますので、ぜひ使ってください。

とりあえず実装してみて、パラメータを変えると、当然結果は変わります。その後、そのパラメータがどのような意味を持っているのだろうということを後から勉強するという、このようなやり方が、一番手っ取り早くアルゴリズムを理解できるのではないかと思っております。

岩沢　具体的なアドバイスをありがとうございます。それでは、藤田さん、お願いいたします。

藤田　自分の経験を踏まえて、どのような学習方法がよかったか振り返ってみると、三つほどあると思いました。まず、理論的なモデルの数式などを学んだり、どのような仕組みになっているのかを学ぶ座学も重要だと思いますが、それと同じくらい、自分の手を動かして学ぶ、ハンズオンが大事だと思いました。では、どこでハンズオンするかというと、アクチュアリー会の講座もありますし、他にもインターネットで探せば、無料ないし安価で参加することができるセミナーや勉強会が、オンライン、オフラインにかかわらず、たくさんあるので、積極的に参加することがいいと思いました。

また、実務との関連がイメージできた方が、学ぶモチベーションをキープできると思います。データサイエンスは、本当に範囲が広いです。モデル一つを取っても学習するのに時間がかかると思うので、実務との関連がイメージできるなど、モチベーションにどのようにつなげるかというところが、学習を維持するポイントになると思います。先ほどはセミナーや勉強会と言いましたが、具体的なケーススタディや事例を紹介するセミナーもたくさんあります。そのようなことを聴講できる場に参加することも重要だと思います。個人的には、海外のカンファレンスで多いイメージですが、日本でももちろんあると思います。

それから、そのようなセミナーや勉強会、ハンズオンを通して学んだ内容を、最後は自分が持っている会社のデータや、それが難しければ自分が拾える有名なデータでもいいと思いますが、それに落とし込んでみて、学んだことと全く同じことをしてみるだけでもいいと思います。データによって全く結果が違ったり、結果が違うことでより理解が深まる部分もあると思うので、お勧めしたいと思います。

岩沢　ありがとうございます。他にも何か、では、浅芝さん、お願いします。

浅芝　月並みですが、関心のある本を１冊手に取って勉強するのが、非常にいいのではないかと思います。講座やセミナーと違って、学習内容を自分で主体的に選べる点が一番のメリットです。本日の会場から東京駅を挟んで反対側にある「丸の内オアゾ」に丸善書店があります。そこの３階には、データサイエンス関係の技術書が豊富にそろっているので、本日の年次大会にご参加された方は、帰りに立ち寄ってみると、きっと興味を惹く一冊が見つかるのではないかと思います。

岩沢　ありがとうございます。高橋さん、お願いします。

高橋　我々のワーキンググループでは、ＳＯＡのプレディクティブ・アナリティクスという試験問題を和訳して、会員に共有するという活動を行っています。私はその活動に直接関わっているわけではないので、個人的な感想になってしまいますが、模擬データが試験問題で与えられて、自分でちょっとしたコードを書きながら、データの加工から分析までをやるという問題になっていて、一連の流れがトレーニングできるので、そのような能力を高めたい人には非常にお勧めの題材だと思いました。

何より自分が役に立ったと感じるものが、その試験が分析して終わりではなくて、「非専門家を想定してレポーティングしなさい」という内容も含まれていて、まさにそのような能力は、アクチュアリーが実務でデータサイエンスを使う上では本当に必要だと思いますので、いいトレーニングの題材だと思いました。以上です。

岩沢　ありがとうございます。大変参考になったと思います。

ディスカッション

The Institute of Actuaries of Japan
Think the Future, Manage the Risk

話題4：少し大所高所の見地に立ったとき，アクチュアリーは，どのような内容のデータサイエンスを，どのように学んでいったらよいと思いますか.

※ 自分のことは棚に上げてもよいし，また，自分の夢のようなものを語ってもらってもかまいません.

岩沢　話題４に移りたいと思います。少し大所高所の見地に立ったとき、アクチュアリーはどのような内容のデータサイエンスを、どのように学んでいったらよいと思うかということで、自分のことを棚に上げてもいいので、ご意見を言っていただければと思います。では、鈴木さん、お願いします。

鈴木　今日は解釈の可能性というような話もありましたけれども、アクチュアリーの仕事は、とにかく当たればいいというものではないと思うので、過度に予測精度を追い求めて、解釈できない複雑なモデルを作るようなスキルは要らないと思っております。

一方で、我々の伝統的に使っている手法では到底太刀打ちできないような、予測精度の高いモデルがあることも事実です。私としては、我々アクチュアリーが、今日出てきたような基礎的なデータサイエンスの手法を学んで、みんなが当たり前のように話せるようになれば、もっと発展していくのではないかと思っております。

岩沢　ありがとうございます。では、藤田さん、よろしいですか。

藤田　完全に自分のことを棚に上げているかもしれませんが、データサイエンスは、今では業界ごとにそれぞれの考え方や視点などが確立されつつあると思っています。他の業界のデータサイエンスがどのような状況か、きちんと詳しく知っているわけではないですが、アクチュアリーにとっても、アクチュアリーにとってのデータサイエンスというものがあると思います。それを学ぶことが理想的だと思いますが、で

は、どうするかということで、それを深掘りして会員の皆様に発信していくことが、我々が所属するデータサイエンス・ワーキンググループの活動使命の一つだと考えています。
　本セッションでも紹介があったような解釈可能性や予測誤差の話は、まさにアクチュアリーにとってのデータサイエンスそのものに該当すると思うので、宣伝ではありませんが、まずはデータサイエンス・ワーキンググループの、発行しているジャーナル等を読んでいただくことが、いい第一歩となるのではないかと思います。

岩沢　ありがとうございます。では、高橋さん、お願いします。

高橋　周りの人に話を聞いてみると、データサイエンスは役に立ちそうだけれども、具体的に業務にどのように役に立てればいいか、想像できないという人が多いようなイメージを持っていて、私自身も持っているところもあるかもしれません。
　今日のセッションでもたくさん話題が出てきましたが、データサイエンスは非常にアクチュアリーにとっても役に立つものだと思っておりまして、アクチュアリー会としても、そのような共通認識を高めていければと、偉そうなことを言ってしまいますけれども、考えております。少し回答とはずれてしまうかもしれませんが、どんどん学んでみて、それを活用してみて、いろいろと発見してみるということが、学ぶべきことだと私は考えております。

岩沢　ありがとうございます。では、浅芝さん、お願いします。

浅芝　何を学ぶべきかという問いに対しては、「分かりません」と答えさせてください。そもそも「大所高所の見地からの見え方」が、私の実力では全く見えてこないという事情もあるのですけれども、それ以上に、データサイエンスの領域は本当に日進月歩だと思います。書店のコーナーを眺めると、最近はＬＬＭ、生成ＡＩ、チャットＧＰＴなどの書籍が平置きされていますが、日々新しい書籍と入れ替わっているという状態で、データサイエンスでできることの限界がどんどん押し広げられているのを実感するところです。ですから、ここは思い切って、あるいは割り切って、各自が関心を持ったテーマを学んでみるということが、今後何がアクチュアリーの世界で活用できるかがだんだん分かってきたときに、誰かがそれをできる、という状況ができやすく、好ましいのではないかと思っております。

岩沢　ありがとうございます。

話題5：データサイエンスを学ぶことに不安をもっているアクチュアリーを思い浮かべて，アドバイスやヒントとなりそうなことを教えてください．

※ これまでの話題の中では触れられなかったご自身や周りの方の経験などについて紹介してもらってもよいかもしれません．あるいは，逆に，これまで触れたことのうち，この話題の観点で再度強調すべきことを簡潔に述べ直すなり敷衍するなりしてもらってもかまいません．

岩沢　ちょうど話題５につながりそうなお話でした。今日会場に来ていただいている方、あるいはオンラインで参加されている方も、データサイエンスに興味を持ってこのセッションに来られたと思いますけれども、データサイエンスを学ぶことには不安を持っているアクチュアリーは非常に多いのではないかと思います。そのような方々を思い浮かべて、アドバイスやヒントなど、自由に語っていただければと思います。高橋さん、よろしいですか。

高橋　この問いを初めて見て「不安って何だろう」と思ったときに、モチベーション面で不安を抱いている方を初めに想像しました。データサイエンスに興味があるという方は非常に多いと想像していますが、自分の興味以外の理由で学ぶことの理由を見いだせない。「学んでもどうなるんだろう」、「役に立つのかな」と不安に思っている方は、ぜひ今日のセッションの話題を参考にしていただければと思います。

私自身も、最初に言いましたけれども、データサイエンスに非常に詳しいわけではなくて、最初は趣味で勉強したような感じなので、そのような人でも生かせる課題は実務に転がっていると思います。

岩沢　ありがとうございます。では、鈴木さん、よろしいですか。

鈴木　機械学習の専門書をオアゾの丸善に行って手に取ると、皆さん、「うっ」となると思います、数式が難しくて。ただ、実はあまり恐れることはないと思います。確かに数学としてきちんと理解して、その式をきちんと解こうと思うと、最適化問題などが出てきたりして、非常に難しいのですけれども、アルゴリズムとして何をやっているかをざっくり理解するだけであれば、インターネットなどで調べると視覚的な説明をしてくれる方がいます。

RやPythonのパッケージの中で最適化の部分などはやってくれるので、実際に我々は、最適化計算をする必要はありません。ですから、基礎的なアルゴリズムをざっくり理解するぐらいの気持ちで、それほど数学を恐れずに臨むのが良いと思います。

岩沢　ありがとうございます。では、浅芝さんから、ぜひお願いします。

浅芝　個人的な経験から申しますと、たとえば、文系出身でプログラミング特有の考え方に慣れていなかったりすると、機械学習の勉強をせっかく始めてみても、壁にぶち当たることがあると思います。具体的には、教科書にあるプログラミングのコードなどを書き写したあと、自分で少し工夫して変えてみたときに、エラーが出たり、全く実行が終わらなかったりすることがよくあると思います。そのときに、Pythonなどの言語で「データ構造」、「アルゴリズム」などのキーワードで本を探して、１冊勉強してみると、基礎体力がついて、壁を乗り越えていけるようになるのではないかと、実感として思います。

岩沢　説得力がありますね。ありがとうございます。では、締めというか、藤田さん、お願いします。

藤田　恐縮です。他の皆さんと重なるところも多いと思いますが、データサイエンスの勉強を始める上では、参考資料やセミナーの機会はたくさんありますし、環境についても今ではクラウドベースのツールなどを簡単に使うこともできるので、環境のセットアップの敷居も低くなっているので、そこは心配ないと思います。

　より重要なことは、繰り返しになりますが、アクチュアリーにとってのデータサイエンスとはどのようなものなのかということを常に念頭に置いて、業務に生かすためにはどうすればいいのかという視点を培うことが大事だと思います。それがモチベーションアップにもつながりますし、不安をうまい具合に払拭できるものだと思います。データサイエンスを学ぶことへの不安には、やはり目的意識を明確化することがいいと思います。そして、データサイエンス・ワーキンググループの活動報告の成果が、ヒントを与えてくれると信じています。

岩沢　ありがとうございました。それでは、時間に余裕があったとき用の話題も用意していたのですが、それは飛ばしまして、質疑をしようと思います。オンライン参加者からは質問が来ていないですね。会場からご質問があれば、お受けしたいと思いますが、いかがでしょうか。

勝野　住友生命の勝野です。本日は、どうもありがとうございました。藤田さんに質問です。先ほど、機械学習を使って価値創造をしたという話をされたと思いますけれども、どのような価値創造をされたか、教えていただければうれしいです。

藤田　あまり細かくは話せませんが、自然災害モデルにおける例を挙げます。自然災害モデルでは、従来は物理的なモデルによるアプローチが主流です。例えば地震であれば、地震の発生過程において、プレートテクトニクスという概念があって、プレート同士でひずみが生じて地震波が発生して、例えば日本に地震波が到達するまでにある程度減衰して、その上に建っている建物が揺れて、建物の構造ごとに、木造は壊れやすい、鉄骨造だと頑健ということを踏まえながら、最終的に保険金の支払いがいくらになる、という感じで、演繹的なプロセスが組み込まれていることが一般的です。

　近年は、例えば東日本大震災が起きて、今までの経験則が合わない、どんどん新しい事象が発生する中で、別なアプローチが必要ではないかという話は、アカデミックの世界でも多少はあったようです。そう

いった背景で、保険という観点から何かできないかということで始まりました。そのような意味で、結構新しい内容に取り組めたということと、その結果を基にアカデミックなところとコラボレーションする機会もあったりして、それが収益につながったかというと、そのような意味ではないかもしれませんが、新たな機会は生めたと思います。

勝野　ありがとうございました。

岩沢　他はいかがですか。

横尾　貴重なお話をありがとうございました。明治安田の横尾と申します。皆さんに伺いたいのですが、今回の機会学習やデータサイエンスのお話を伺って、思ったところがあります。内容としては非常に高度な話をされていると思うので、実際にこれをアクチュアリー業務に活用する形になった際に、レポーティングということがＳＯＡの試験の話で出てきたと思いますが、アクチュアリーの方だけでなく、会社の中の方に成果や分析結果を伝えることが難しいと個人的には考えています。

具体的に、妄想などという話ではなくて、以前に別のところで、機械学習ではないですけれども、データサイエンスの講座のようなものを受けたことがありまして、その中でも、分析結果を周囲に報告する際には、使っても回帰分析ぐらいのモデルでないと周囲には伝わらないのではないかということで、モデルの制約のようなところも現実を突きつけられたこともあります。そこで、より高度な内容を業務一般に落とし込むための方策ではないですが、どのようなところに注意すればよりレベルアップできると考えていらっしゃるか、伺いたいと思います。

岩沢　ありがとうございます。時間の関係で全員からは聞けないので、どなたか１人か２人、短く言えますか。高橋さんは、先ほど関係あることをおっしゃっていましたね。

高橋　ありがとうございます。レポーティングということで、非常に難しい問題かと思いますが、今までもアクチュアリーは、結構難しいことを経営層などにレポーティングしていると思います。モデルの詳細な技術的なところは、分析家の中で共有すればいいと思いますが、そこから出てくるメッセージは、テクニカルなところを離れて、会社の経営やアクションにつながるという、アウトプットとしては得られると思います。

その得られる過程を説明するために、機械学習に踏み込まなければいけないのかもしれませんけれども、そのようなアウトプットを大事にレポーティングしていくことが必要だと思います。私も全くできていないので、これから勉強だと思っております。

岩沢　ありがとうございます。お時間の関係で、他のご質問も、休み時間等に聞いていただければと思います。

データサイエンス専門講座のご紹介

参考

- **講師陣**

前編講師：野村 俊一 早稲田大学大学院 准教授

早稲田大学で統計科学を専門として、データサイエンス関連の授業も担当
当会2017年度優秀論文賞受賞（統計数理研究所在籍時）
2021年度早稲田大学ティーチングアワード受賞（受賞科目はデータサイエンスⅡ）

後編講師：岩沢 宏和 早稲田大学大学院 客員教授

当会基礎講座・追加演習講座（損保数理）講師
2021年Hachemeister賞（※）受賞（内容はデータサイエンス）
※ 損保分野の世界最高峰の賞

- **充実した講義内容**

初学者(若手からベテランまで)向けの体系的なカリキュラムで一通りの知識を習得可能
独学で学んだ方の学び直しにも活用

- **個人でも負担しやすい価格設定**

受講料は５万円。公益社団法人として会員に提供している講座のため、受講料が安い

申込受付中！！
11 月 7日（火）正午 受付締切
kouza@actuaries.jpへご連絡下さい！

データサイエンス専門講座のご紹介(内容・日程(注1))

The Institute of Actuaries of Japan
Think the Future, Manage the Risk

科目	講師	内容	日程
前編 （講義中心）	野村俊一 早稲田大学大学院 准教授	(第１回) Ｒの導入と簡単な回帰モデル (第２回) 線形回帰モデル (第３回) 主成分分析・クラスタリング (第４回) 決定木 (第５回) 一般化線形モデル１ (第６回) 一般化線形モデル２ (第７回) 時系列解析	11/28～1/23 (原則火曜日(注2))
後編 （実習中心）	岩沢宏和 早稲田大学大学院 客員教授	(第８回) 予測モデリングの基本手順 (第９回) 探索的データ解析（ＥＤＡ） (第10回) 予測モデリング用のモデル例 (第11回) モデルの選択・評価の方法 (第12回) 回帰問題での実践（ＧＷ） (第13回) 分類問題での実践（ＧＷ）	2/3～3/16 （原則土曜日(注3)）

（注１）詳細は「2023年度アクチュアリー講座 専門講座（データサイエンス）実施要領」をご覧ください。
（注２）講義時間は1.5時間。各回の講義の前に、事前オンデマンド配信を視聴いただきます。
（注３）講義時間は3時間

最後に一つ案内があります。パネリストのうちのお二人も入り口としてお使いになったデータサイエンスの専門講座の案内です。内容については、時間もないので、スライドを２枚用意していて、２枚目に詳しく出ておりますので、ぜひごらんください。ここで伝えなければいけないことは、締切間近だということで、１週間を切っています。講義時間は全部で 39 時間だと思います。それで５万円と、考えられないほど格安ですので、内容もよくごらんいただいた上で、ぜひご参加いただければと思います。

それでは、やや時間も延長ぎみでしたが、これで終わりとしますので、改めましてパネリストの皆様、ありがとうございました。また、会場やオンラインで参加の方々も、ありがとうございました。本セッションはもう時間切れですけれども、休憩時間にまた質問していただければと思います。これで本パネルディスカッションは終了いたします。どうもありがとうございました。このあと大会委員より事務連絡がありますので、そのまましばらくお待ちください。どうもありがとうございました。

（参考資料）

ブラックボックスな予測モデルの力を借りてデータを分析してみよう - 実行用 R スクリプト

予測モデルを作るのは、実はそれほど難しくない

```
# 必要なパッケージのインストール状況を確認し、見つからない場合はインストールします。
pkgs = c('tidyverse', 'tidymodels', 'GGally', 'patchwork', 'gridExtra', 'liver',
         'ranger', 'xgboost', 'kernlab', 'nnet', 'DALEX', 'shapviz', 'fastshap')
for (pkg in pkgs) {
    if (!require(pkg, quietly=TRUE, character.only=TRUE)) install.packages(pkg)
}
```

```
# データを読み込み、変数のうち gender と smoker を数値化、region を削除します。
library('tidyverse')
library('liver')
data(insurance)
df = insurance %>%
    mutate(gender=as.numeric(gender=='female'),
           smoker=as.numeric(smoker=='yes')) %>%
    select(-region)
df %>% str()
```

```
# データを学習用データと評価用データに分割し、分割後のデータ件数を確認します。
library('tidymodels')
set.seed(2023)
split_data = initial_split(df, prop=3/4)
train = training(split_data)
test = testing(split_data)
cat(nrow(train), nrow(test))
```

```
# 学習用データの分布と相関係数をまとめて可視化します。
```

```
library('GGally')
ggpairs(train, lower=list(continuous=wrap(ggally_points, alpha=.1)))


# 予測モデルに対して、予測値のプロットと評価を表示する関数を定義します。
library('patchwork')
evaluate_model = function(model, new_data){
    aug = augment(model, new_data=new_data)
    loss = rmse(aug, charge, .pred)$.estimate %>% round(3)
    p = ggplot() + ylim(-3000, 60000) + ylab('予測値')
    p1 = p + geom_point(aes(x=age, y=.pred), data=aug, alpha=.25) +
        labs(title=model$title, x='年齢')
    p2 = p + geom_point(aes(x=charge, y=.pred), data=aug, alpha=.25) +
        geom_line(aes(x=c(0, 60000), y=c(0, 60000)), lty=3) +
        labs(title=paste('RMSE =', loss), x='真の値')
    p1 + p2 + plot_layout(widths=c(4, 3))
    }


# 線形回帰モデルを作成します。作成した予測モデルはリストに格納します。
models = list()
models[['lm']] = linear_reg() %>% fit(charge~., train)
models[['lm']]$title = '線形回帰モデル'
evaluate_model(models[['lm']], test)


# サポートベクトル回帰モデル（線形カーネル）を作成します。
models[['svl']] = svm_linear() %>% set_engine('kernlab') %>%
    set_mode('regression') %>% fit(charge~., train)
models[['svl']]$title = 'サポートベクトル回帰モデル - 線形カーネル'
evaluate_model(models[['svl']], test)


# サポートベクトル回帰モデル（RBF カーネル）を作成します。
models[['svr']] = svm_rbf() %>%
    set_mode('regression') %>% fit(charge~., train)
models[['svr']]$title = 'サポートベクトル回帰モデル - RBF カーネル'
evaluate_model(models[['svr']], test)
```

```
# ランダムフォレストモデルを作成します。
models[['rf']] = rand_forest() %>%
    set_mode('regression') %>% fit(charge~., train)
models[['rf']]$title = 'ランダムフォレストモデル'
evaluate_model(models[['rf']], test)


# 勾配ブースティングモデル (XGBoost) を作成します。
models[['gbm']] = boost_tree() %>%
    set_mode('regression') %>% fit(charge~., train)
models[['gbm']]$title = '勾配ブースティングモデル - XGBoost'
evaluate_model(models[['gbm']], test)


# ニューラルネットワークモデルを作成します。
models[['nn']] = mlp(penalty=10, epochs=600) %>%
    set_mode('regression') %>% fit(charge~., train)
models[['nn']]$title = 'ニューラルネットワークモデル'
evaluate_model(models[['nn']], test)


## 作成した予測モデルは、データの分析にも役立つ


# DALEX パッケージで予測モデルを解釈するための前準備を行います。
library('shapviz')
library('DALEX')
explainers = list()
for (name in names(models)){
    explainers[[name]] = DALEX::explain(model=models[[name]],
                                        data=select(test, -charge),
                                        y=test$charge)
    explainers[[name]]$title = models[[name]]$title
}
```

```
# PFI プロットを作図します。
pfi = explainers[['rf']] %>% model_parts()
plot(pfi) + labs(title=explainers[['rf']]$title, subtitle=NULL) +
    ylab('変数の値を並び替えたあとの RMSE') + theme_gray() +
    theme(legend.position='none', strip.text=element_blank())


# ICE プロットを作図します。
ice = explainers[['rf']] %>% predict_profile(test)
plot(ice, variables=c('age', 'bmi')) +
    labs(title=explainers[['rf']]$title, subtitle=NULL, y='予測値') +
    ylim(0, 53000) + theme_gray()


# PDP (Partial Dependence Plot) を作図します。
pdp = explainers[['rf']] %>% model_profile()
plot(pdp, variables=c('age', 'bmi'), geom='profiles') +
    labs(title=explainers[['rf']]$title, subtitle=NULL, y='予測値') +
    ylim(0, 53000) + theme_gray() + theme(legend.position='none')


# 喫煙の有無でグループ化した PDP (Grouped PDP) を作図します。
gpd = explainers[['rf']] %>% model_profile(groups='smoker')
plot(gpd, variables=c('age', 'bmi'), geom='profiles') +
    labs(title=explainers[['rf']]$title, subtitle=NULL, y='予測値') +
    ylim(0, 53000) + theme_gray() + theme(legend.position='none')


# 1 番目のインスタンスに関する SHAP ウォーターフォール図を作図します。
shap1 = explainers[['rf']] %>% predict_parts(test[1,], type='shap')
sv_waterfall(shapviz(shap1)) +
    ggtitle(paste(explainers[['rf']]$title, '- 1 番目のインスタンス')) +
    xlab('予測値') + theme_gray()


# 10 番目のインスタンスに関する SHAP ウォーターフォール図を作図します。
shap2 = explainers[['rf']] %>% predict_parts(test[10,], type='shap')
sv_waterfall(shapviz(shap2)) +
```

```
    ggtitle(paste(explainers[['rf']]$title, '- 10番目のインスタンス')) +
    xlab('予測値') + theme_gray()


# ブラックボックスモデルの解釈を回帰式に反映させた線形回帰モデルを作成します。
models[['tr']] = linear_reg() %>%
    fit(charge~. + I(age^2) + smoker:bmi + I(smoker*as.numeric(bmi>30)), train)
models[['tr']]$title = '改良版の回帰モデル'
evaluate_model(models[['tr']], test)


## Appendix


# 評価用データ全体に対するSHAP Importanceを計算します。
library('fastshap')
pred_parsnip = function(object, newdata){
    predict(object, new_data=newdata)$.pred
    }
shap = fastshap::explain(object=models[['rf']],
                         X=select(train, -charge),
                         newdata=select(test, -charge),
                         pred_wrapper=pred_parsnip,
                         nsim=20)
sv = shapviz(shap, X=test)
p = sv_importance(sv, 'bee') + ggtitle(models[['rf']]$title)
q = sv_importance(sv) + ggtitle('SHAP Importance')
p + q


# BMIに関する各予測モデルのGrouped PDPをまとめて作図します。
library('gridExtra')
plots = list()
for (name in names(explainers)){
    gpd = explainers[[name]] %>% model_profile(groups='smoker')
    p = plot(gpd, variables=c('bmi'), geom='profiles') +
        labs(title=explainers[[name]]$title, subtitle=NULL, y=NULL) +
        ylim(-3000, 60000)
```

```
    plots[[name]] = p + theme_gray() + theme(legend.position='none')
}
grid.arrange(grobs=plots, ncol=3)
```