SASによる新しい大規模統計学入門（１）

- Large scale statistical forecasting using by SAS -

小野　潔　松澤　一徳

Kiyoshi Ono,　Kazunori Matsuzawa

株式会社インテック 金融ソリューション開発センター

Financial Solutions Development Center, Financial Solutions Service Division, INTEC Inc

株式会社インテック（INTEC Inc.）

設立　　1964年1月11日
資本金　208億30百万円
代表者　代表取締役社長　日下　茂樹
東証一部上場　（証券コード：3626　ITホールディングス株式会社　TIS株会社と当社の共同持株会社）
従業員数 3,666名　グループ会社60社　　グループ従業員　19,472名（2014/9/30現在）
事業内容 技術研究、ICTコンサルティング、ソフトウェア開発
　　システム・インテグレーション、ネットワークサービス、アウトソーシングサービス

本日の目的

本チュートリアルでは**SAS 9.4のリリース**に際し、私たちデータ分析者が新たに享受できるメリット‥**より容易にビッグデータの分析やDeep Learningの実装が可能になった**ことについての概要についてご紹介するものです。

前半では機械学習のこれまでの歩みを振り返った上で、その基本的な手法である**決定木分析**、その応用分野である**ランダムフォレスト**、高度に複雑化する処理の高速化を可能にする**SAS High Performance Analytics**についての概要を説明します。

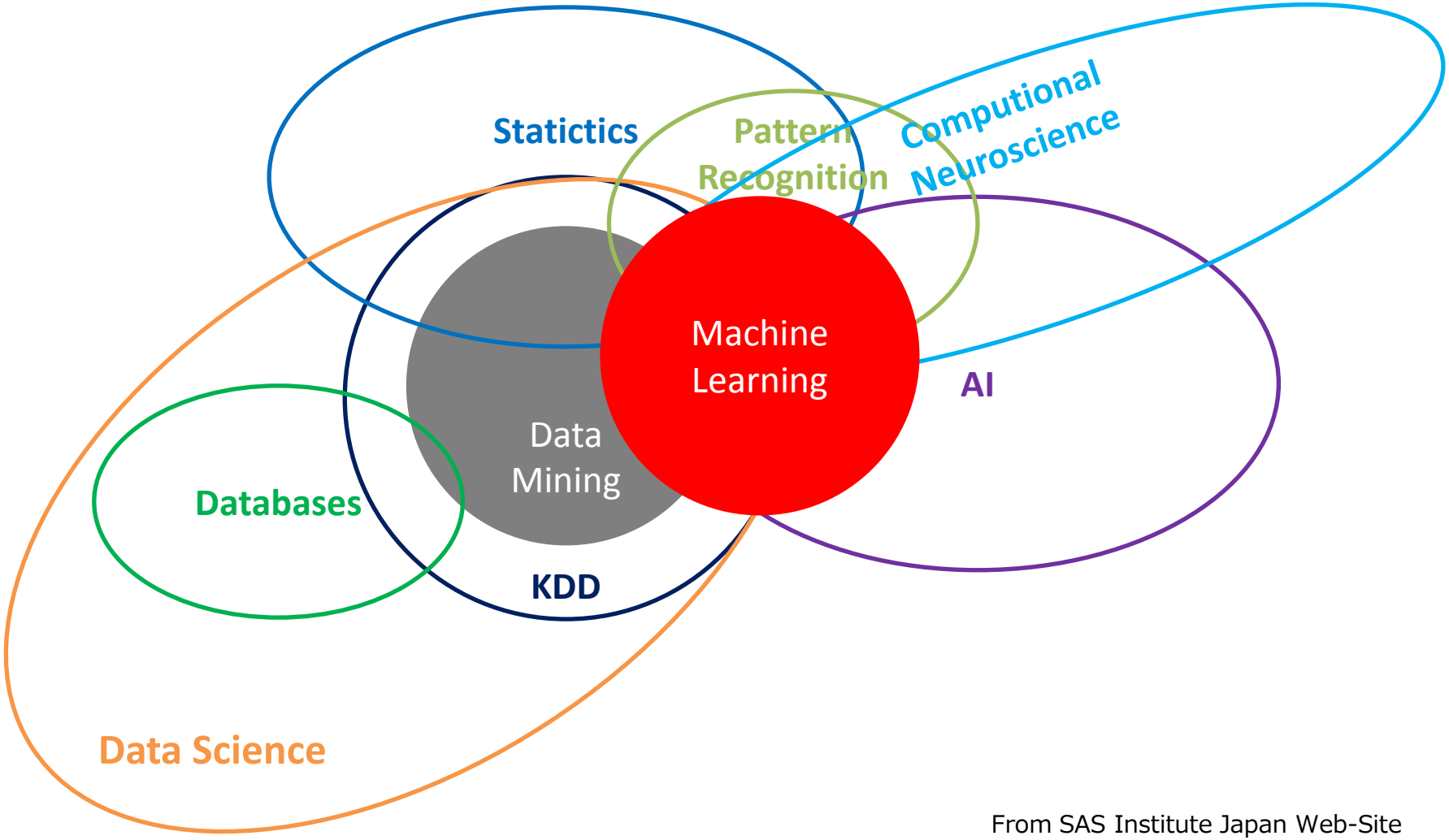
後半では機械学習の最新テクノロジーである**Deep Learning**のご説明と、それを**具体的にSASで実装する手法**についてご紹介します。

尚、今回ご紹介した内容はSAS様のほうで**本日14:00よりD会場にて**深堀した内容をご紹介しています。そちらも合わせてご受講いただけますと幸いです。

- SASによるインメモリ分散並列処理－レコメンドプロシジャ入門
- SAS High-Performance Analyticsによるビッグデータ解析

機械学習とは？

機械学習の領域



From SAS Institute Japan Web-Site

機械学習の種類

Machine Learning

教師あり学習

目標変数あり
（未知データ予測目的）

（例）多変量解析、ロジスティク回帰判別分析、k近傍法、決定木、ランダムフォレスト、　サポートベクターマシン（SVM)、ニューラルネットワーク‥

教師なし学習

目標変数なし
（未知データ規則性発見目的）

（例）クラスタリング、 階層クラスタリング、アソシエーションルール（A priori）、One-class　SVM、自己組織化マップ（SOM）、PCA

半教師あり学習

教師あり学習と
教師なし学習の併用

（例）オートエンコーダー、TSVN、EM

機械学習の歴史

1946年　IBMによる世界初のコンピュータENIACの開発

1956～1960年代 第１次AIブーム　「探索・推論問題の解法」

・IPLによるプリンキピア・マテマティカの定理証明（1957）
・ニューラルネットワーク（1963）
・遺伝子アルゴリズム（1958）

1956年　ダートマスワークショップにて初めて人工知能（Artificial Intelligence）が言葉が出現

1964～1966年　ジョセフ・ハイゼンバウムの対話システムELIZA（人工無能）登場（チューリング・テスト合格？）

1970年代　AIの冬の時代到来（機械翻訳絶望、現実問題がとけず）

1973～1976年　スタンフォード大学にてエキスパートシステム（MYCIN（マイシン））の開発（知識工学の幕開け）

1980年代 第2次AIブーム 「知識工学の時代（自然言語・画像・音声理解）」

1982～1994年 通産省が第5世代コンピュータプロジェクト（1981年）に570億円を投資

1990年～　再びAIの冬の時代到来　知識（ルール）獲得の失敗

1990年代　オントロジーの発達（概念集合の体系化）

2003年～　IBM "Practical intelligent Question Answering Technology プロジェクト"
論理形式分析と機械翻訳を結合するも、成果あげられず

2010年～ 第3次AIブーム 「自己学習、表現の時代」

2012年　トロント大学 DeepLearnningによる物体認識コンペティションにて圧勝

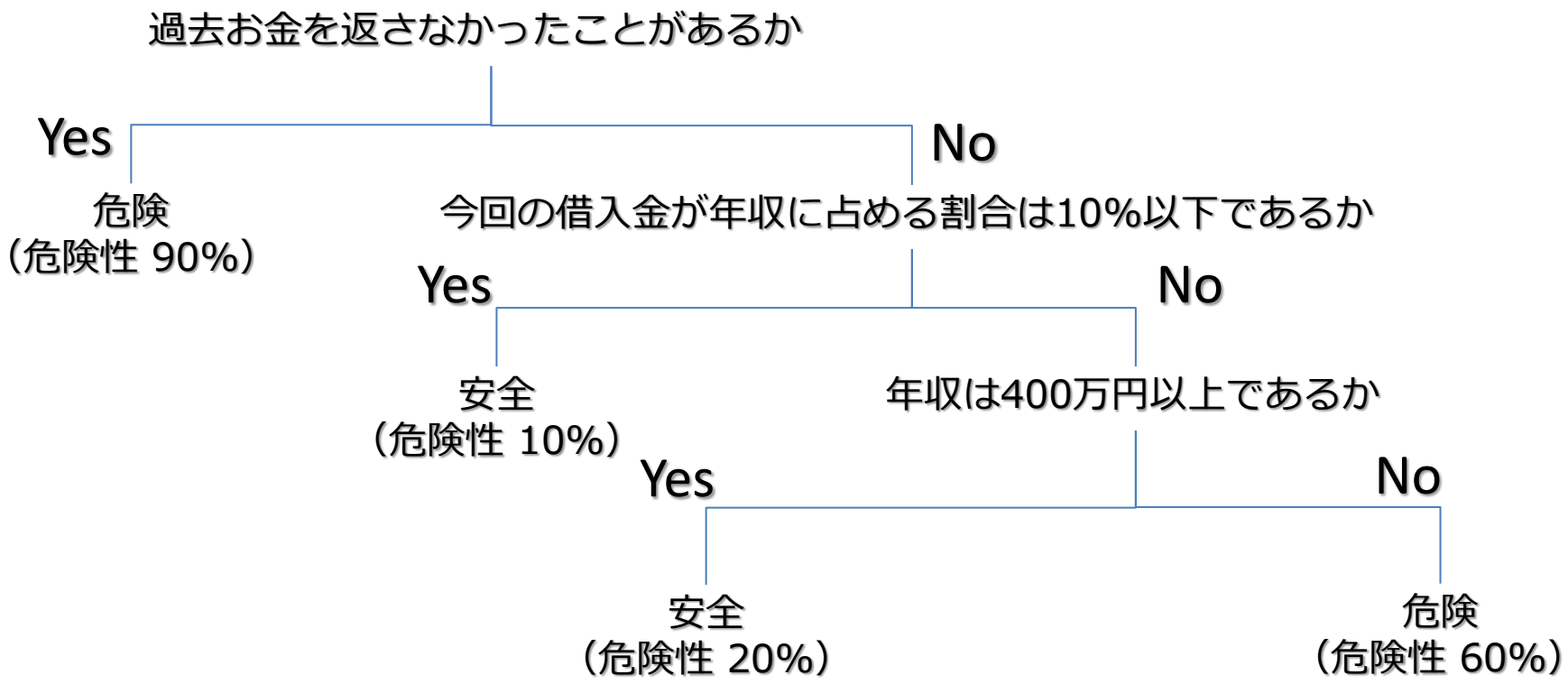
2011年　IBMの質問応答システムWatsonが米国クイズ番組「Jeopardyジェパディ）」でクイズ王に勝利

2045年頃 AIが人類を越す年　‥？（by レイ・カーツワイル）

典型的手法

- 決定木分析 -

決定木分析とは



IF=THENルールによりすべてのデータについて予測値を分類する手法

SASコード

```
PROC HPSPLIT DATA=l.SOURCE MAXDEPTH=7 MAXBRANCH=2;

   TARGET A_GOAL;                                  ①目的変数を指定（2値）
   INPUT VAR1 VAR2 ‥ VAR3_FL1/ LEVEL = int;        ②説明変数を指定
   CRITERION entropy;                              ③分割基準
   PRUNE misc / N<=15;                             ④剪定基準
   PARTITION FRACTION(Validate=0.2);               ⑤バリデーションデータ比率
   RULES FILE='hpsplit-rules.txt';                 ⑥ルール出力先
   SCORE OUT=scored;                               ⑦スコア出力先

RUN;
```

出力結果（IF-Thenルール）

```
*-------------------------------------------------------------*
NODE = 2
*-------------------------------------------------------------*
MISSING(外部信用格付) OR (外部信用格付 IS ONE OF 1, 2)
    PREDICTED VALUE IS 0
    PREDICTED 1 = 0.06343( 136/2144)
    PREDICTED 0 = 0.9366( 2008/2144)

*-------------------------------------------------------------*
NODE = 3
*-------------------------------------------------------------*
(年間返済比率 <19.99996)
AND (外部信用格付け IS 3)
    PREDICTED VALUE IS 1
    PREDICTED 1 = 0.5398( 414/767)
    PREDICTED 0 = 0.4602( 353/767)

. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . .
```

←ノードID

← 条件式

← 本ノードが1（危険）となる予想確率
← 本ノードが0（正常）となる予想確率

←ノードID

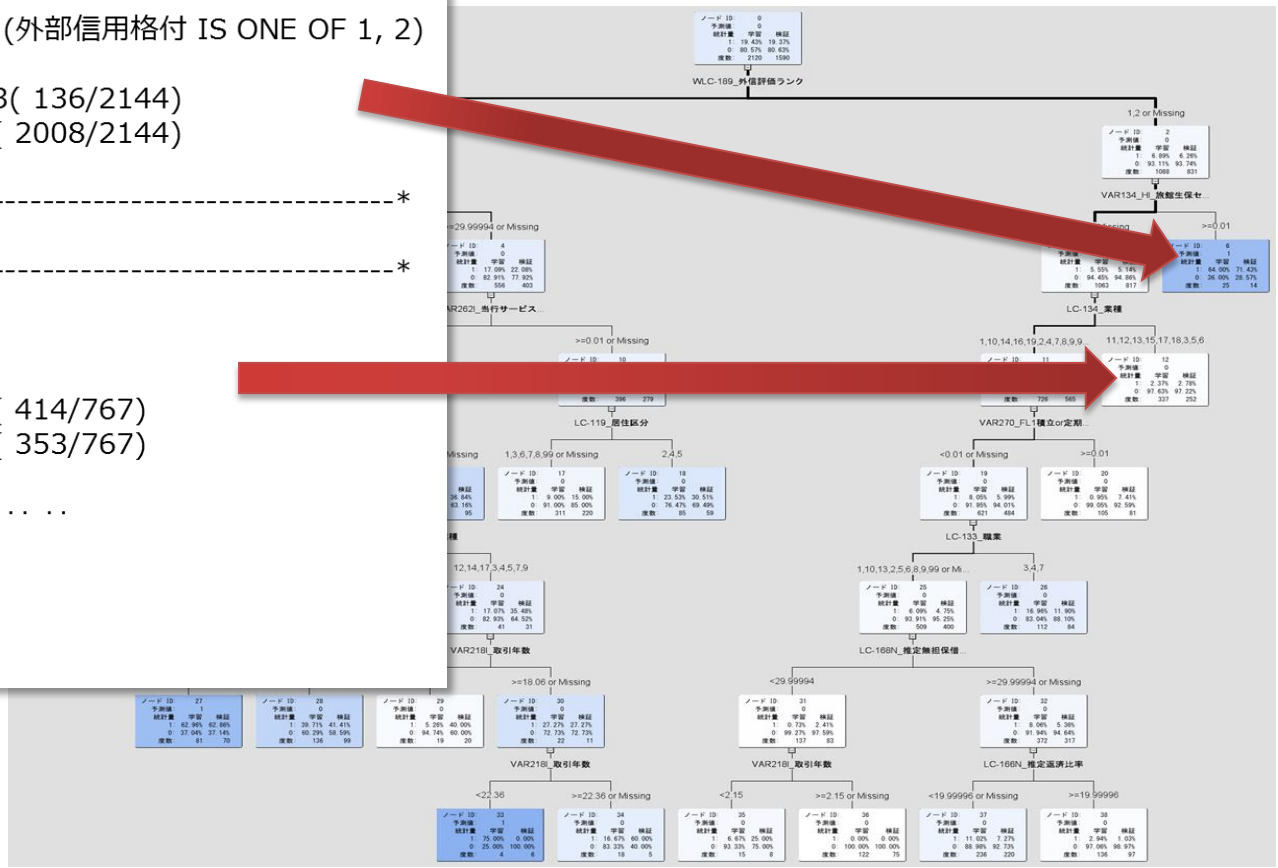
← 条件式

← 本ノードが1（危険）となる予想確率
← 本ノードが0（正常）となる予想確率

全体の意味の読み取り難しい・・・・

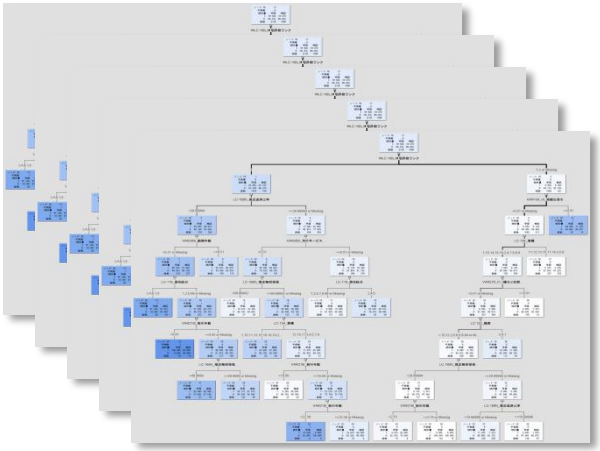
SAS Enterprise Minerによるビジュアル化

```
*-----------------------------------------------------------*
NODE = 2
*-----------------------------------------------------------*
MISSING(外部信用格付) OR (外部信用格付 IS ONE OF 1, 2)
    PREDICTED VALUE IS 0
    PREDICTED 1 = 0.06343( 136/2144)
    PREDICTED 0 = 0.9366( 2008/2144)

*-----------------------------------------------------------*
NODE = 3
*-----------------------------------------------------------*
(年間返済比率 <19.99996)
AND (外部信用格付け IS 3)
    PREDICTED VALUE IS 1
    PREDICTED 1 = 0.5398( 414/767)
    PREDICTED 0 = 0.4602( 353/767)

. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
```



ランダムフォレストの考え方

X個の説明変数からランダムでN個の変数を選択し
決定木を生成（これを任意の数繰り返す）



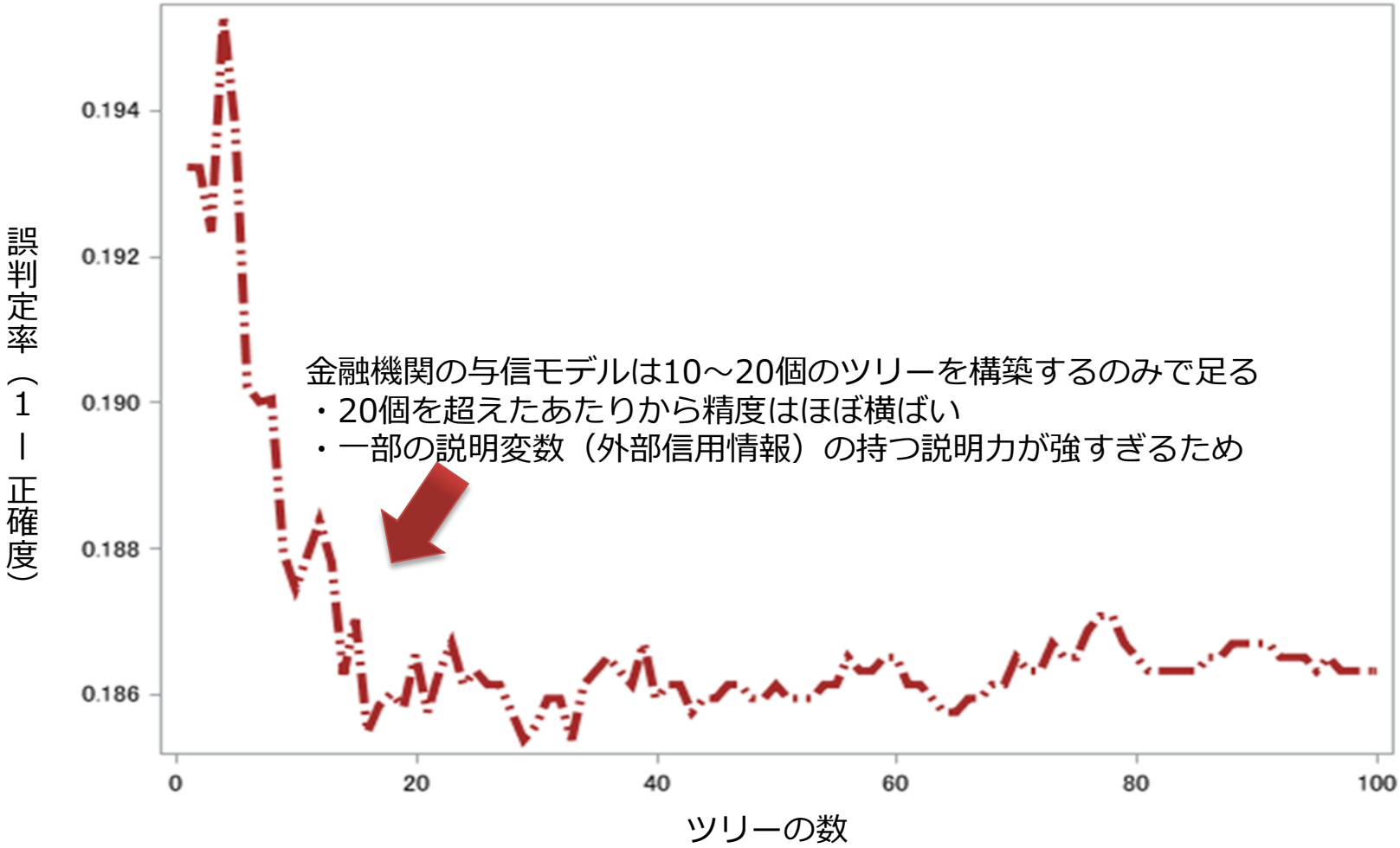
複数の決定木

各決定木の算出スコアの平均値等から
最終的なモデルのスコアを算出

SASコード

```
proc  hpforest data=l.SOURCE
        maxtrees=100                          ①作成する決定木の数
        vars_to_try =8                        ②説明変数の数
        trainfraction=0.6;                    ③トレーニングデータ比率
  target A_GOAL/level=binary;                 ④目的変数
  input VAR1 VAR2 …/level=nominal;            ⑤説明変数
  ods output FitStatistics = fitstats_vars8(rename=(Miscoob=VarsToTry8));
run;
```

ランダムフォレストの効果

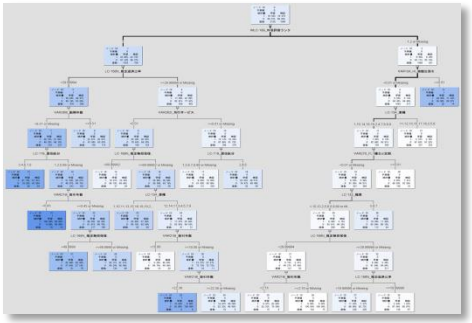


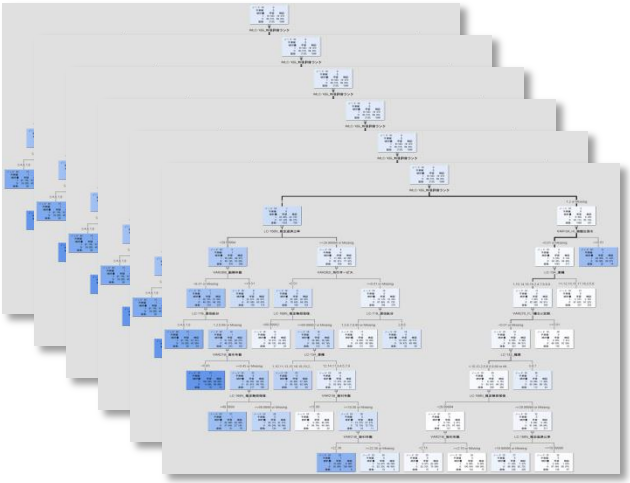
金融機関の与信モデルは10～20個のツリーを構築するのみで足る
・20個を超えたあたりから精度はほぼ横ばい
・一部の説明変数（外部信用情報）の持つ説明力が強すぎるため

SAS High Performance Analysticsのご紹介

- 複雑化する分析手法への対応 -

High Performance Analyticsの必要性





×200

分析手法の高度化に伴い、データ分析にかかる時間は長期化する一方‥ ‥ ‥ ‥ ‥ ‥

一般的な解決策

・並列分散処理　（マルチコアCPUに対し処理を分散・並列化させる）
・インメモリ技術（HDDに比べ圧倒的なI/O速度を有するメモリ内で分析処理を完結）

→ これらの技術を容易に扱えることを実現したのがSAS High Performance Analytics！！

High Performance Analytics の特徴

従来コード（非線形最小二乗法）

```
Proc nlin data=Enzyme method=marquardt;
  parameters theta1=155 theta2=0 to 0.07 by 0.01;
  model Velocity = theta1*Concentration / (theta2 + Concentration);
Run;
```

High Performance Analyticsのコード

```
Proc hpnlmod data=enzyme;
  parameters theta1=155 theta2=0 to 0.07 by 0.01;
  model Velocity ~ residual(theta1*Concentration / (theta2 + Concentration));
Run;
```

処理内容に大きな違いがない（ストレスなく高速分析を利用できる）

High Performance Analytics 対応製品

対応製品	SASプロシジャ	概要
共通	HPDS2	次世代データステップ
	HPDMDB	要約統計量
	HPSAMPLE	サンプリング
	HPCORR	相関分析
	HPIMPUTE	欠損値補完
	HPBIN	変数のビン化
	HPSUMMARY	記述統計

対応製品	SASプロシジャ	概要
SAS® High-Performance Statistics（対応製品：SAS/STAT）	HPCANDISC	正準判別分析
	HPFMM	一般化線形モデル（混合モデル）
	HPGENSELECT	一般化線形モデル
	HPLMIXED	線形混合モデル
	HPLOGISTIC	ロジスティック回帰、変数選択
	HPNLMOD	非線形回帰、最尤法
	HPPLS	PLS回帰・主成分回帰・縮小ランク回帰
	HPPRINCOMP	主成分分析
	HPQUANTSELECT	分位点回帰モデルにおける変数選択
	HPREG	線形回帰、変数選択
	HPSPLIT	決定木モデル

対応製品	SASプロシジャ	概要
SAS® High-Performance Data Mining（対応製品：SAS Enterprise Miner）	HP4SCORE	HPFORESTモデルに基づくスコアリング
	HPBNET	ベイジアンネットワーク
	HPCLUS	クラスタリング
	HPDECIDE	意思決定の情報から最適ディシジョン選択
	HPFOREST	決定木ベースの予測モデリング
	HPNEURAL	ニューラルネットワーク
	HPREDUCE	変数の選択/縮約、相関や共分散の算出
	HPSVM	サポートベクターマシン
	HPTSDR	時系列データの次元削減
SAS® High-Performance Text Mining（対応製品：SAS Text Miner）	HPTMINE	テキストマイニングモデル
	HPTMSCORE	テキストデータスコアリング

対応製品	SASプロシジャ	概要
SAS® High-Performance Econometrics（対応製品：SAS/ETS）	HPCDM	複合分布モデル(試用版)
	HPCOPULA	コピュラによる多変量分布の依存関係のシミュレーション
	HPCOUNTREG	計数データ(カウントデータ)に対する回帰
	HPPANEL	パネルデータモデル
	HPQLIM	制限従属変数モデル
	HPSEVERITY	リスク規模分布の推定
SAS® High-Performance Forecasting（対応製品：SAS Forecast Server）	HPFORECAST	時系列予測モデル

対応製品	SASプロシジャ	概要
SAS® High-Performance Optimization（対応製品：SAS/OR）	HPLSO	遺伝的アルゴリズムをベースとした最適化法
	OPTMILP	混合整数線形計画法
	OPTLP	線形計画法
	OPTMODEL	様々な最適化問題に包括的に対応した機能
(Social Network Analysis Server) SAS® Customer Link Analytics SAS® Fraud Framework	OPTGRAPH	グラフ理論、ネットワーク解析

From SAS Institute Japan Web-Site

ご清聴ありがとうございました