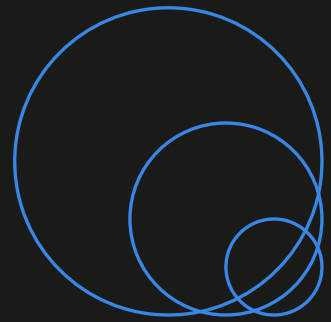
# Ex.2 演習の進め方 ―Volveフィールドデータで代理モデルを構築する―

実践 貯留層工学 データ駆動型シミュレータ編 第2回 演習パワポ資料



# 本日のアジェンダ (1/2)

- 演習(Ex.2)の目的とゴール
- 使用データ: Equinor Volveフィールドデータ(第1回からの接続)
- 入力・出力変数の設計と前処理
- RSM・回帰モデルの構築演習
- ランダムフォレストモデルの構築演習
- ニューラルネットワークモデルの構築演習(任意課題)

# 本日のアジェンダ (2/2)

- 精度検証(ホールドアウト・k-fold交差検証)
- 検証で分かったこと ―時系列データにおける分割の落とし穴

パートF

# Ex.2 演習の進め方

› 本資料のねらい: 第2回演習(Ex.2)を、調査部門/参考資料/Volveの実データと第1回演習の成果に沿って具体的に進められるようにする

› 対象データ: Equinor Volveフィールドデータのうち、油田全体の月次生産量・圧力データ(Volve_Production_Data_Processed.csv、112行)

› 第1回Ex.1で得た気づき(欠損の構造・0値混入・日付表記の落とし穴)を踏まえ、パートB〜Eの内容を代理モデル構築まで実践する

# F1: 演習(Ex.2)の目的

■ 目的1: 第1回で読み込んだVolveの月次生産量・圧力データから、代理モデル構築用の入力・出力データを整理する

■ 目的2: パートB(RSM・回帰)・パートC(ランダムフォレスト、任意でニューラルネットワーク)の手法で、貯留層圧力を予測する代理モデルを構築する

■ 目的3: ホールドアウト法・k-fold交差検証(パートE)により精度を検証し、手法間で比較する

調査部門/参考資料/Volve/README.md；調査部門/成果物『データ駆動型シミュレータ編_第1回_演習notebook.ipynb』

# 本日の演習(Ex.2)の流れ

# F2: 使用データセットの確定(第1回からの継続)

■ 第1回詳細版F2ではVolve/SPE9/SPE10を候補として挙げていたが、社内格納済みで演習に使えるVolveデータを本演習の確定データとする

■ 使用ファイル: Volve_Production_Data_Processed.csv(2007年9月～2016年12月・月次112行、貯留層圧力p・累積生産量Np/Gp/Wp・累積圧入量Gi/Wiを収録)

■ 第1回で作成した坑井別特徴量テーブル(well_feature_table.csv、7坑井)は行数(サンプル数)が少なすぎるため参考情報にとどめ、本演習の主教材は月次112行の油田全体データとする

調査部門/参考資料/Volve/README.md

# F3: 入力変数・出力変数の設計

■ 出力変数(予測対象): p (psia) ―貯留層圧力。マテリアルバランス的には累積生産・圧入量から決まる応答量

■ 入力変数(候補): 経過日数(elapsed_days)・累積油量Np(STB)・累積ガス量Gp(SCF)・累積水量Wp(STB)・累積ガス圧入量Gi(SCF)・累積水圧入量Wi(STB)

■ Volve_input_to_mbal.csvは同種の変数をマテリアルバランス入力形式で保持しており、本演習の入力設計と対応関係にある

調査部門/参考資料/Volve/README.md

# 入力変数と出力変数

| 入力変数(X) | 出力変数(y) |
| --- | --- |
| 経過日数 | p(貯留層圧力, psia) |
| Np(累積油量) | |
| Gp(累積ガス量) | |
| Wp(累積水量) | |
| Gi・Wi(累積圧入量) | |

# F4: データ準備の具体的手順

■ pandas.read_csvで読み込み、Date列を第1回同様pd.to_datetime(dayfirst=True)で変換する(本ファイルもDD/MM/YYYY表記のため)

■ データ開始日からの経過日数(elapsed_days)を新たな入力変数として追加する

■ sklearn.preprocessing.StandardScalerで入力変数を標準化し、train_test_splitで訓練/テストデータに分割する(パートE1参照)

# F5: RSM/回帰モデルの構築演習(検証結果)

■ sklearn.preprocessing.PolynomialFeatures(degree=2)+LinearRegressionで2次の応答曲面モデルを構築する

■ ホールドアウト検証(訓練89件/テスト23件): 線形回帰はRMSE 89.7psia・R2 0.831、2次多項式回帰はRMSE 75.5psia・R2 0.881に改善

■ 入力変数のうちNp・Gpは相関係数0.9998とほぼ完全に共線関係にあり、パートB6の多重共線性対策(標準化・変数整理)の実例として使える

調査部門/成果物『データ駆動型シミュレータ編_第2回_演習notebook.ipynb』(検証記録)

# F5補足: 回帰式(線形回帰)

■ 実測単位のまま(標準化なし)で線形回帰を学習すると、次の式が得られる(RMSE 90.4psia・R2 0.829、標準化ありの結果RMSE 89.7psia・R2 0.831とほぼ同一)

■ p (psia) = 4787.51 − 1.83e-08×経過日数 − 5.10e-05×Np − 3.38e-07×Gp − 2.04e-04×Wp + 0×Gi + 2.14e-04×Wi

■ 係数の符号が物理的直感と一致する点に注目: 生産量Np・Gp・Wpの係数は負(生産で圧力が下がる)、圧入量Wiの係数は正(圧入で圧力が保たれる)。Gi(ガス圧入)は全期間ゼロのため係数も0

調査部門/成果物『データ駆動型シミュレータ編_第2回_演習notebook.ipynb』(検証記録)

# F5補足: 回帰式(2次多項式回帰・上位項)

| 項(標準化後の入力変数の積) | 係数 |
|---|---|
| 切片 b0 | 4933.208 |
| Wp × Wi | -515784.5 |
| Gp × Wi | -375033.1 |
| $Wi^2$ | 340464.4 |
| 経過日数 × Wi | 241879.7 |
| 経過日数 × Np | -209449.8 |
| $Wp^2$ | 191461.6 |
| Np × Wp | 167306.4 |
| 経過日数 × Wp | -157251.6 |

調査部門/成果物『データ駆動型シミュレータ編_第2回_演習notebook.ipynb』(検証記録)。全27項(定数項含む)のうち係数の絶対値が大きい上位8項。RMSE=75.51psia・R2=0.881

# F5補足: 多項式回帰の式は『解釈』には使えない

■ 2次多項式回帰は精度こそ高い(R2 0.881)が、上位項の係数が数十万〜数十万のオーダーで大きく振れており、線形回帰(数式がそのまま物理的に解釈できる)とは対照的

■ 原因はF5で確認した入力変数間の多重共線性(Np-Gpの相関0.9998等)。標準化後の交互作用項同士も強く相関するため、個々の係数が不安定になり、式の値だけでは『どの変数が効いているか』を読み取れない

■ 『精度は出せるが式の解釈はできない』というRSM/多項式回帰の限界を実データで確認できた。変数の効き方を知りたい場合はランダムフォレストのfeature_importances_(F6)の方が信頼できる

# F6: ランダムフォレストモデルの構築演習(検証結果)

■ sklearn.ensemble.RandomForestRegressor(n_estimators=300)で学習。ホールドアウト検証でRMSE 49.3psia・R2 0.949、OOBスコア0.91とRSMより高精度

■ feature_importances_では経過日数(0.23)・Np(0.22)・Gp(0.19)・Wi(0.18)・Wp(0.18)が拮抗する一方、Gi(累積ガス圧入量)は重要度0.00

■ Volveは水攻(water injection)主体でガス圧入(Gi)は全期間ゼロであり、実データ確認で重要度がゼロになる理由も特定できる ―データに現れない変数は重要度も出ない、という実例

調査部門/成果物『データ駆動型シミュレータ編_第2回_演習notebook.ipynb』(検証記録)

# F6補足: 予測値 vs 実測値(ランダムフォレスト、テスト23件) (1/2)

| 日付 | 実測p(psia) | 予測p(psia) | 誤差(予測-実測) |
|---|---|---|---|
| 2007-09-01 | 4780.59 | 4772.64 | -7.95 |
| 2008-01-01 | 4780.59 | 4765.57 | -15.02 |
| 2008-07-01 | 4362.30 | 4354.78 | -7.52 |
| 2008-08-01 | 4309.95 | 4284.65 | -25.30 |
| 2008-09-01 | 4231.05 | 4155.71 | -75.34 |
| 2009-03-01 | 4786.10 | 4608.08 | -178.02 |
| 2009-11-01 | 4657.89 | 4626.86 | -31.03 |
| 2010-04-01 | 4587.40 | 4676.65 | 89.25 |
| 2010-09-01 | 4881.97 | 4885.35 | 3.38 |
| 2011-01-01 | 4886.18 | 4866.64 | -19.54 |
| 2011-03-01 | 4896.76 | 4877.81 | -18.95 |
| 2011-05-01 | 4889.22 | 4890.64 | 1.42 |
| 2011-08-01 | 4910.69 | 4893.33 | -17.36 |
| 2012-02-01 | 4907.06 | 4891.40 | -15.66 |

# F6補足: 予測値 vs 実測値(ランダムフォレスト、テスト23件) (2/2)

| 日付 | 実測p(psia) | 予測p(psia) | 誤差(予測-実測) |
|---|---|---|---|
| 2012-04-01 | 4877.04 | 4862.64 | -14.40 |
| 2013-02-01 | 4846.15 | 4854.09 | 7.94 |
| 2013-05-01 | 4893.72 | 4920.03 | 26.31 |
| 2013-06-01 | 4904.74 | 4934.73 | 29.99 |
| 2013-10-01 | 4924.32 | 4931.06 | 6.74 |
| 2014-05-01 | 4881.39 | 4943.09 | 61.70 |
| 2015-04-01 | 5052.10 | 5039.80 | -12.30 |
| 2015-10-01 | 4985.67 | 5000.02 | 14.35 |
| 2016-04-01 | 5070.37 | 5106.49 | 36.12 |

調査部門/成果物『データ駆動型シミュレータ編_第2回_演習notebook.ipynb』(検証記録)。RMSE=49.32psia, MAE=31.11psia, R2=0.949, OOB=0.910

# F6補足: 特徴量重要度(ランダムフォレスト)

| 入力変数 | 重要度 |
|---|---|
| 経過日数(elapsed_days) | 0.231 |
| Np(累積油量, STB) | 0.216 |
| Gp(累積ガス量, SCF) | 0.191 |
| Wi(累積水圧入量, STB) | 0.182 |
| Wp(累積水量, STB) | 0.180 |
| Gi(累積ガス圧入量, SCF) | 0.000 |

調査部門/成果物『データ駆動型シミュレータ編_第2回_演習notebook.ipynb』(検証記録)。GiはVolveで全期間ゼロ(ガス圧入なし)のため重要度0

# F6補足: 5-fold交差検証の結果(ランダムフォレスト)

| Fold | R2 |
|---|---|
| Fold 1 | 0.950 |
| Fold 2 | 0.901 |
| Fold 3 | 0.768 |
| Fold 4 | 0.897 |
| Fold 5 | 0.972 |
| 平均(標準偏差) | 0.897 (0.071) |

調査部門/成果物『データ駆動型シミュレータ編_第2回_演習notebook.ipynb』(検証記録)。ホールドアウト単独のR2(0.949)より低いfoldがあり、単一分割に頼らない確認の重要性を示す

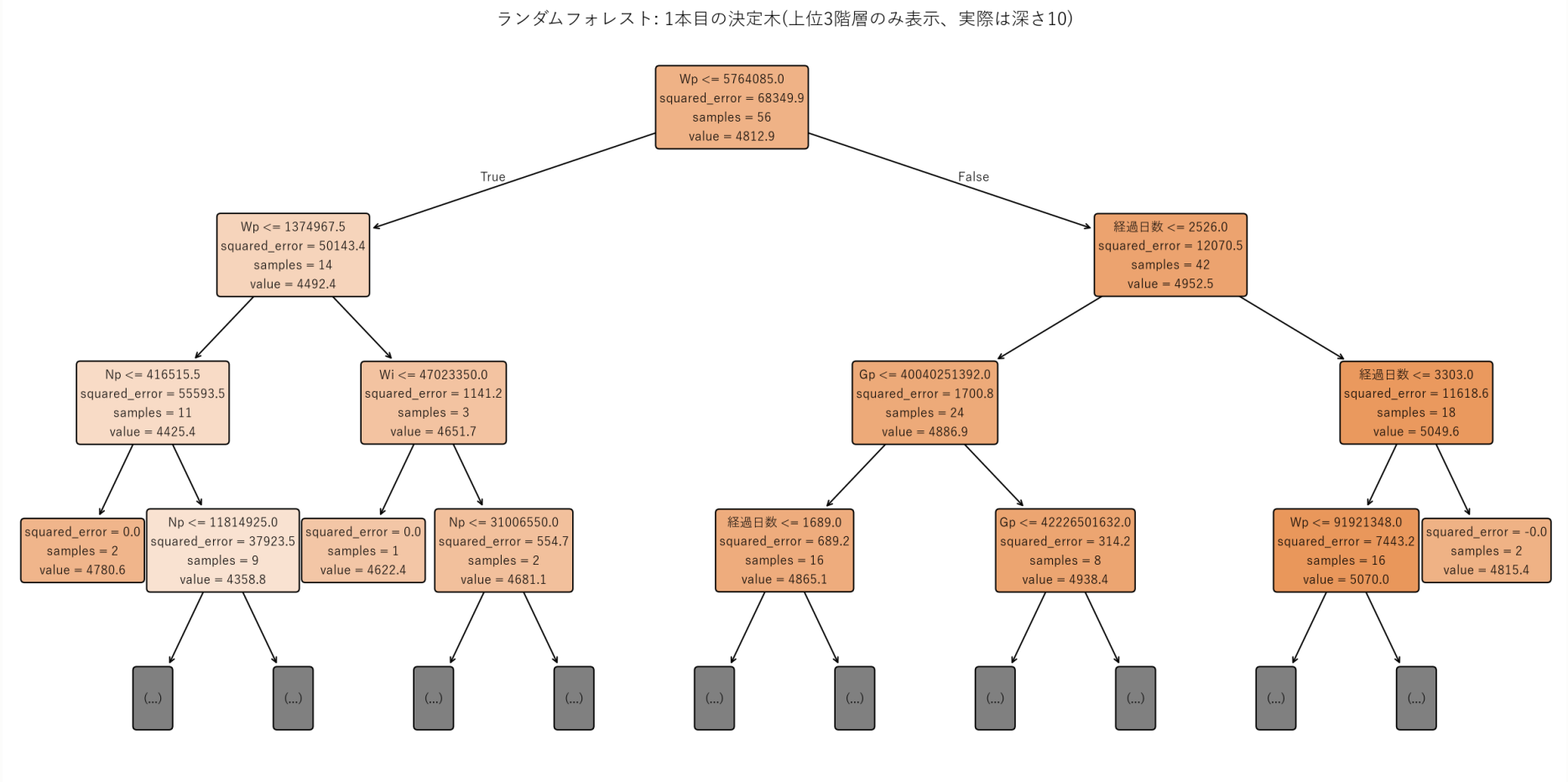
# F6補足: ランダムフォレストの中身を覗いてみる

■ ランダムフォレストは決定木を多数(本演習ではn_estimators=300)組み合わせたアンサンブル学習であり、最終予測は300本の木の予測の平均で決まる(パートC2)

■ 1本1本の木は『ある条件で分岐 → さらに分岐 → 葉に到達したらその葉に属する訓練データの平均値を予測値とする』という単純なルールの繰り返しで構成される

■ 次スライドでは、実際に学習させた1本目の木を可視化し、どの変数・しきい値で分岐しているかを確認する

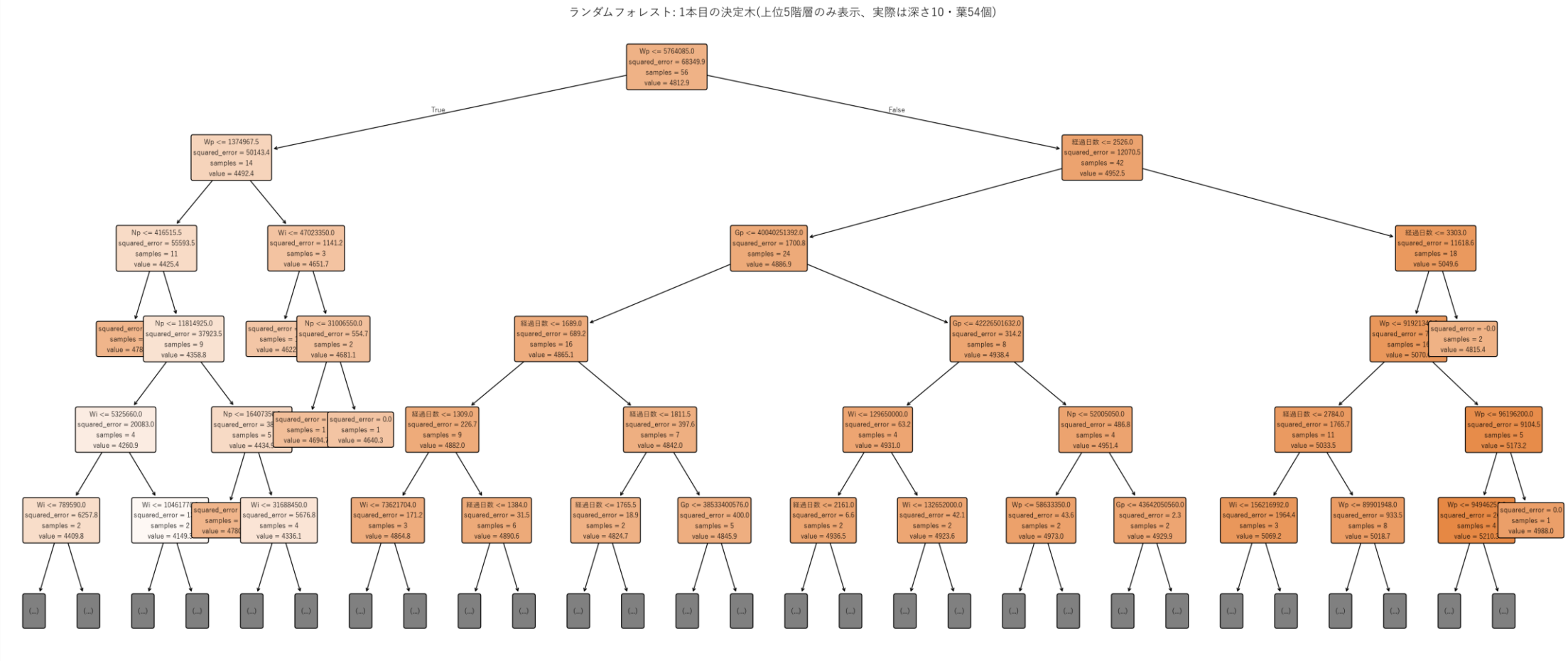
# F6補足: 決定木1本の中身(概要)




ランダムフォレスト(300本)のうち実際の1本目の木。上位3階層のみ表示(実際は深さ10・葉54個)

調査部門/成果物『データ駆動型シミュレータ編_第2回_演習notebook.ipynb』(検証記録)。sklearn.tree.plot_treeで生成

# F6補足: 決定木1本の中身(詳細)




同じ木を上位5階層まで表示。分岐を重ねるほど条件が細分化され、末端の葉ほど少数サンプルに基づく予測になる

調査部門/成果物『データ駆動型シミュレータ編_第2回_演習notebook.ipynb』(検証記録)。sklearn.tree.plot_treeで生成

# F7: ニューラルネットワークモデルの構築演習(任意課題・検証結果)

■ 時間・習熟度に余裕のある受講者向けに、sklearn.neural_network.MLPRegressor(隠れ層32・16ノード)で構築する任意課題とする

■ ホールドアウト検証ではRMSE 137.2psia・R2 0.606と、RSM(R2 0.881)・ランダムフォレスト(R2 0.949)を下回る結果になった

■ 訓練データが89件と少なく、ニューラルネットワークはデータ量を要するというパートC10の指摘を実データで裏付ける結果になった。第4回のPINN等、事前知識(物理制約)を組み込む手法の必要性を予告する材料にもなる

調査部門/成果物『データ駆動型シミュレータ編_第2回_演習notebook.ipynb』(検証記録)

# F8: 精度検証演習(ホールドアウト・k-fold比較)

■ 3手法(RSM/多項式回帰・ランダムフォレスト・NN)についてホールドアウト法でRMSE・MAE・R2を算出し、一覧表で比較する

■ ランダムフォレストについて5-fold交差検証を実施したところ、R2は0.768〜0.972(平均0.897)とfoldにより変動し、単一のホールドアウト結果(R2 0.949)だけに頼らない精度確認の重要性を確認できる

■ 手法間の比較は必ず同じテストデータ・同じ評価指標で揃える(パートE4)

調査部門/成果物『データ駆動型シミュレータ編_第2回_演習notebook.ipynb』(検証記録)

# 分割方法によってモデル評価が一変する

| ランダム80/20分割 | | 時系列分割(2015年2月以降をテスト) |
|---|---|---|
| RF: R2 0.949 | 同じデータ・同じモデルでも | RF: R2 -0.42 |
| 多項式(2次): R2 0.881 | | 多項式(2次): R2 -12.15 |
| 精度は良好 | | 精度が破綻(外挿) |

# F9: 検証で分かったこと ―時系列データの分割方法(重要な気づき)

■ ランダムな80/20分割ではRF R2 0.949だったが、時系列に沿って前半80%を訓練・後半20%(2015年2月～2016年12月)をテストにすると、RF R2は-0.42、多項式回帰はR2 -12.15まで悪化する

■ 原因はテスト期間の圧力(最大5273.87psia)が訓練期間の圧力範囲(最大5114.90psia)を超えており、モデルが学習データ範囲外への外挿を強いられるため(パートE5・B10の『外挿領域での精度低下』の実例)

■ ランダム分割による検証結果だけを鵜呑みにせず、時間的な構造を持つデータでは時系列分割での検証も併用する必要がある、という実務上重要な教訓

調査部門/成果物『データ駆動型シミュレータ編_第2回_演習notebook.ipynb』(検証記録)

# F10: 結果比較・可視化

- 予測値 vs 実測値の散布図(対角線からのズレで精度を視覚的に確認、手法ごとに作成)
- 手法別のRMSEを棒グラフで比較する(ランダムフォレストが最も精度が高いことを可視化)
- 時系列分割のテスト区間では、実測値と予測値の推移を重ねて描き、外挿による乖離を視覚的に確認する

調査部門/成果物『データ駆動型シミュレータ編_第2回_演習notebook.ipynb』(検証記録)

# まとめ

■ 本演習(Ex.2)では、Volveの月次生産量・圧力データ(112行)を用いて、RSM(多項式回帰)・ランダムフォレスト・(任意)ニューラルネットワークの3手法で貯留層圧力の代理モデルを構築する

■ ホールドアウト検証ではランダムフォレストが最も高精度(R2 0.949)、次いでRSM(R2 0.881)、ニューラルネットワークはデータ量不足で精度が伸びない(R2 0.606)という結果を実データで確認した

■ 本資料の内容は実際にVolveデータで検証済み(検証記録: 調査部門/成果物/データ駆動型シミュレータ編_第2回_演習notebook.ipynb)。ランダム分割と時系列分割で精度が大きく異なるという重要な気づきを教材に反映した

# 参考資料一覧(出典) (1/2)


■ Equinor, 'Volve field data set' https://www.equinor.com/energy/volve-data-sharing

■ energistics.org, "Equinor's Volve Field Test Data" https://energistics.org/equinors-volve-field-test-data

■ yohanesnuwara/pyreservoir(データ再配布元) https://github.com/yohanesnuwara/pyreservoir/tree/master/data/volve

■ 調査部門/参考資料/Volve/README.md(格納データの詳細・出典・ライセンス注意)

■ 調査部門/成果物『データ駆動型シミュレータ編_第1回_演習notebook.ipynb』(第1回の検証記録・本演習の前提)

■ 調査部門/成果物『データ駆動型シミュレータ編_第2回_演習notebook.ipynb』(本資料の内容を実データで検証した記録)

■ scikit-learn公式ドキュメント https://scikit-learn.org/stable/

■ scikit-learn, 'Cross-validation: evaluating estimator performance' https://scikit-learn.org/stable/modules/cross_validation.html

# 参考資料一覧(出典) (2/2)

■ scikit-learn, 'Tuning the hyper-parameters of an estimator' https://scikit-learn.org/stable/modules/grid_search.html