

生産管理 最終レポート

中央大学理工学部
ビジネスデータサイエンス学科
23D7104001I 高木悠人

1. 課題概要

データセット中の末尾 500 レコードのデータの OV 値について予測を行うことを目的とする。ただし、

- 可能な限り少ない説明変数の数で
- 高い予測精度を達成する

ことを条件とする。「可能な限り少ない説明変数の数」については VM で予測するにあたって監視するパラメータを少なくしたいという目的がある。予測精度については RMSE を評価指標とする。

予測モデル構築にあたっては、動的にパラメータを更新しても良いし、固定モデルで予測を行っても良い。固定モデルで予測を行う場合は、末尾 500 ロットの 1 レコード目の `process_end_time` より前の `final_mes_time` で予測モデルを構築すること。

モデル自体については特に制約は設けない。クラスタ分析でグループ分けして予測モデルを組んでも良いし、複数のモデルを組み合わせても良い、またデータの交互作用などをとっても良い。

2. 実行環境

本課題では、以下の環境で実行・分析した。

分析開始日:	2026 年 1 月 10 日
PC (notebook):	Apple macbook air m4
Python 環境:	venv 仮装環境(ローカルでの実行)
Python バージョン:	python-3.12.11
各種パッケージバージョン:	requirements.txt リンクは付録の項に示すこととした
セッション管理:	Python Package(SessionSmith==2.0.0)を利用した
コードバージョン管理:	GitHub(リンクは付録の項に示すこととした)

3. 分析手順

本分析は以下の手順で実施した。

1. データの概要の確認

はじめに、本分析で利用するデータの概要、基本統計量等を確認する。そして、データを訓練データとテストデータに分割する。データ分割の基準は、課題の通り以下のように定義する。

訓練データ:

テストデータに該当しない 1776 行のうち、テストデータの最初の 1 レコードの "process_end_time" より前である 1155 レコード

テストデータ:

末尾 500 レコード

2. データの可視化

欠損等を確認し、存在するのであれば適する形で補完する。そして、訓練データとテストデータの分布の違いや時系列プロット等を実施し、分析方法を検討する。

3. 機械学習自動化ライブラリ (h2o) を用いた分析手法の検討

機械学習の変数選択とモデル選択を自動化する h2o を使い、ある程度分析方法を検討する。

4. モデリング

データの可視化と h2o のモデリング結果を踏まえ、モデリング手法を検討、実装する。

4. データの概要の確認

はじめにデータの概要を確認した。本分析では、時系列データ 2 つと 83 のセンサデータ、1 つの品質データを持つ表形式データを対象とした。各列の詳細を以下に示す。

1. OV

ウェハー上の不良品数を意味しており、値が小さいほど良品と言える。本分析では、不良品数を予測するため、目的変数として定義する。

2. process_end_time

作業が完了した時間を意味する。

3. final_mes_time

作業完了後に検査し完了した時刻を意味する。本分析目的は、検査後に予測モデルの改善を実施しそれ以降の予測精度を向上させることと品質悪化の原因を探ることにある。

4. X{n} ※n: 1~83

センサーデータの計測値を意味する。

次に、各データの基本統計量を確認した。基本統計量の数値については、載せきれないため GitHub の notebook に掲載することとした。

5. データの可視化

次にデータの可視化を行なった。はじめに、目的変数 OV における訓練データとテストデータの分布を以下のように可視化した。

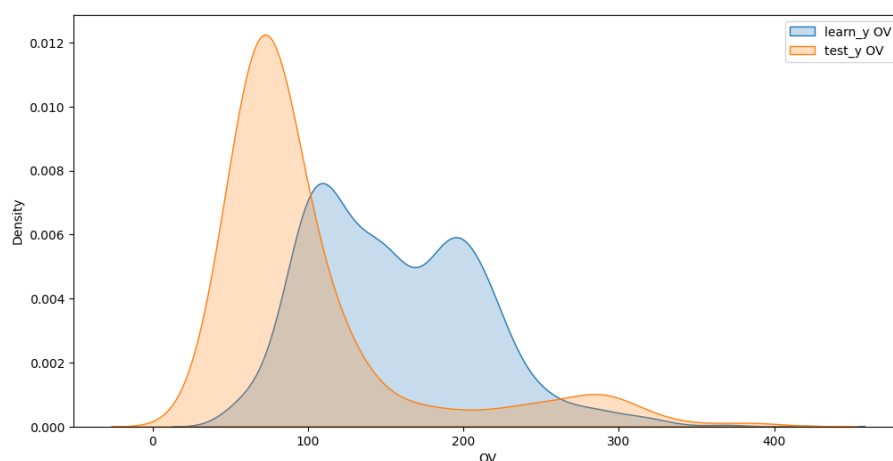


図 1. 目的変数 OV の訓練・テストデータの分布

上記の図をみると、訓練データとテストデータにおける目的変数 OV の分布には明確な差異が存在することが分かる。訓練データは比較的広い範囲に分布し、高い値側にも山が見られる一方で、テストデータは低い値に分布が集中している一山の分布とみなせる。このことから、訓練データとテストデータの間で分布の偏りが生じており、モデルの汎化性能に影響を与える可能性が示唆される。

次に、製造日と製造数の分布について可視化した。

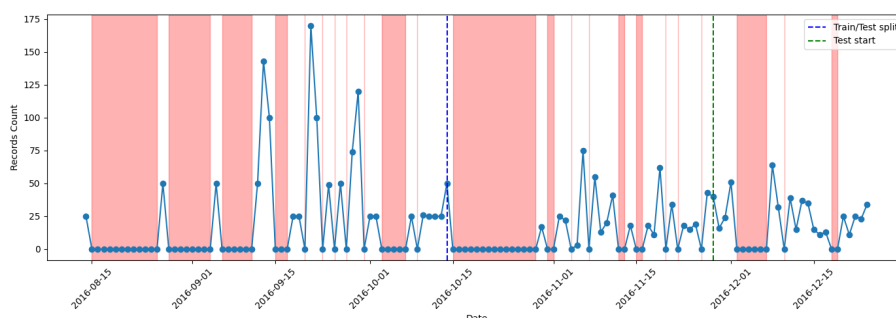


図2. 製造日と製造数の推移

※1 縦青線が訓練データ末尾/ 縦赤線がテストデータ開始

※2 色がついている部分は製造していない期間を意味する

上記の図をみると、製造日ごとの製造数は時間の経過に伴って大きな変動を示しているものの、本分析の目的である品質予測の観点では、稼働日・非稼働日の区別自体が直接的な説明要因とはならないと考えられる。したがって、色付きで示されている製造していない期間はデータの欠損や異常として扱う必要はなく、あくまで観測されている製造日の系列に着目すれば十分であると考ええる。

一方で、時系列全体をみると、訓練データ期間とテストデータ期間とで製造数の水準やばらつきに違いが見られる。このことは、製造条件やプロセス環境が時間とともに変化している可能性を示唆しており、品質に影響を与える潜在的な要因がテスト期間では異なる分布を持っている可能性がある。品質予測モデルにおいては、このような分布の変化が予測誤差の増大につながる恐れがあるため、学習データがテストデータを十分に代表しているかを確認する必要があると考える。

以上より、本データを用いたモデリングでは、日付そのものを強い説明変数として用いるのではなく、製造数や各工程のプロセス変数といった品質に直接関連する特徴量を中心にモデルを構築することが妥当である。また、時間の経過による緩やかな傾向変化を捉えるために、製造日を補助的な変数として扱う、あるいは期間ごとにモデル性能を検証することで、品質予測としての頑健性を確保することが重要であると考えられる。

上記を踏まえて、各変数においても訓練データとテストデータで分布が異なっているかを確認することとした。可視化した結果を以下に示す。

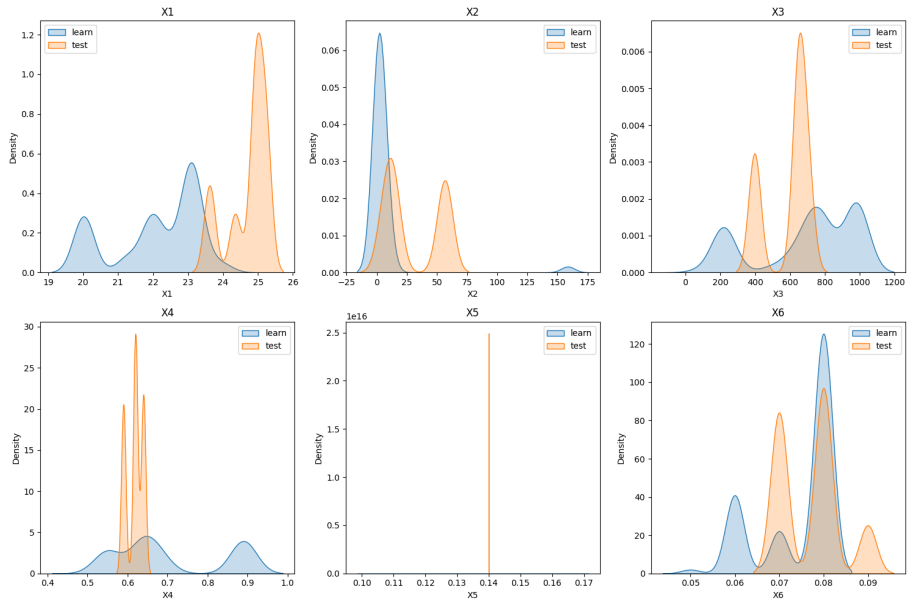


図 3. X1~X6 の訓練・テストデータの分布推移

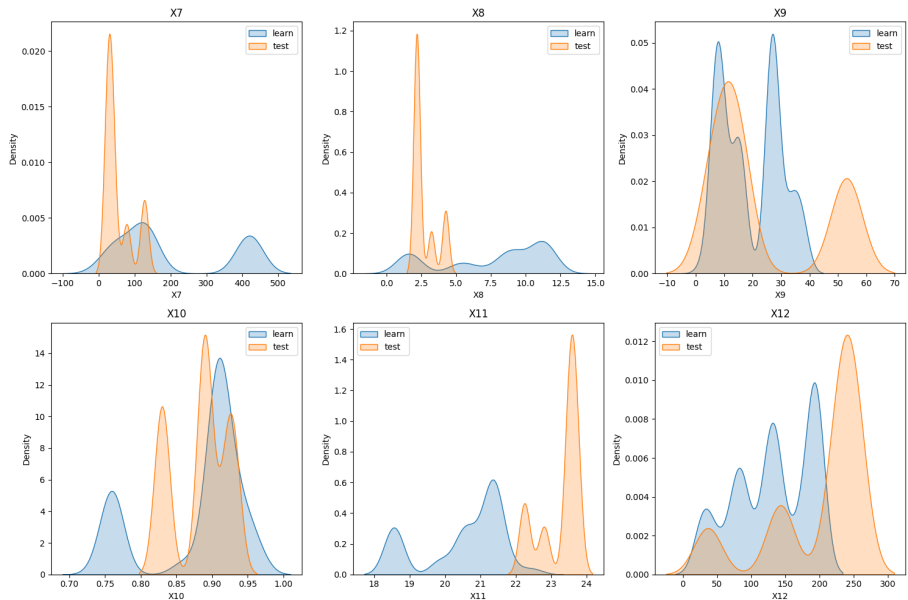


図 4. X7~X12 の訓練・テストデータの分布推移

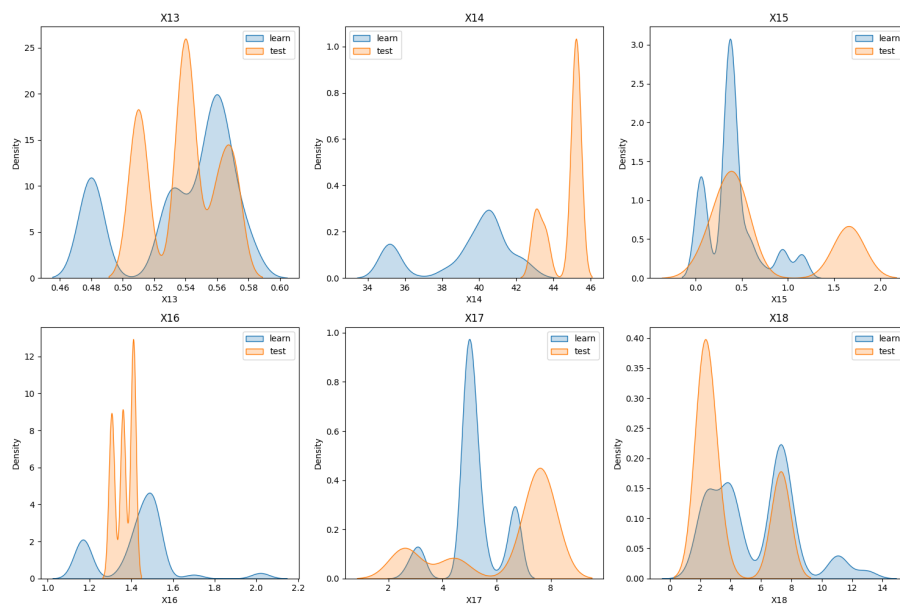


図 5. X13~X18 の訓練・テストデータの分布推移

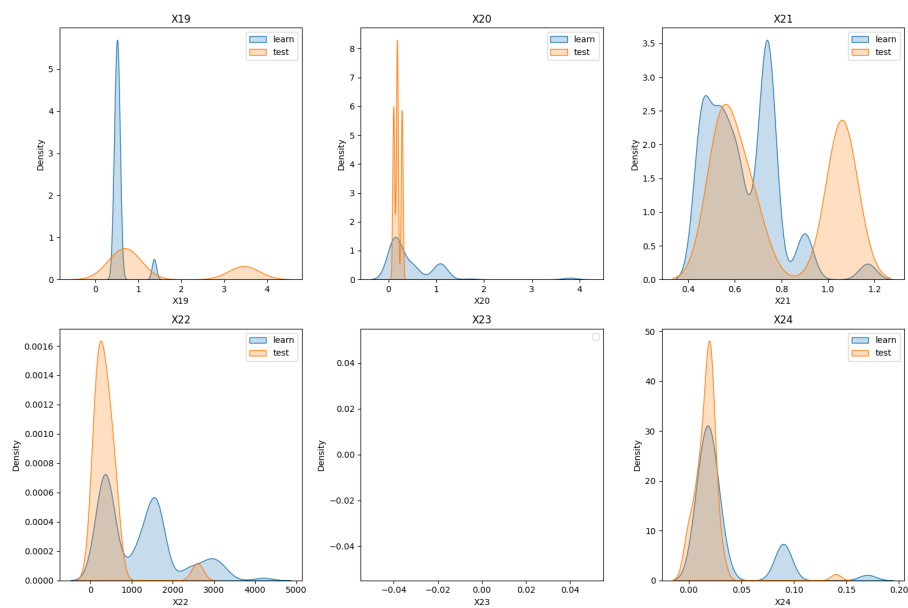


図 6. X19~X24 の訓練・テストデータの分布推移

※X23 は訓練・テストデータとも、全てであった

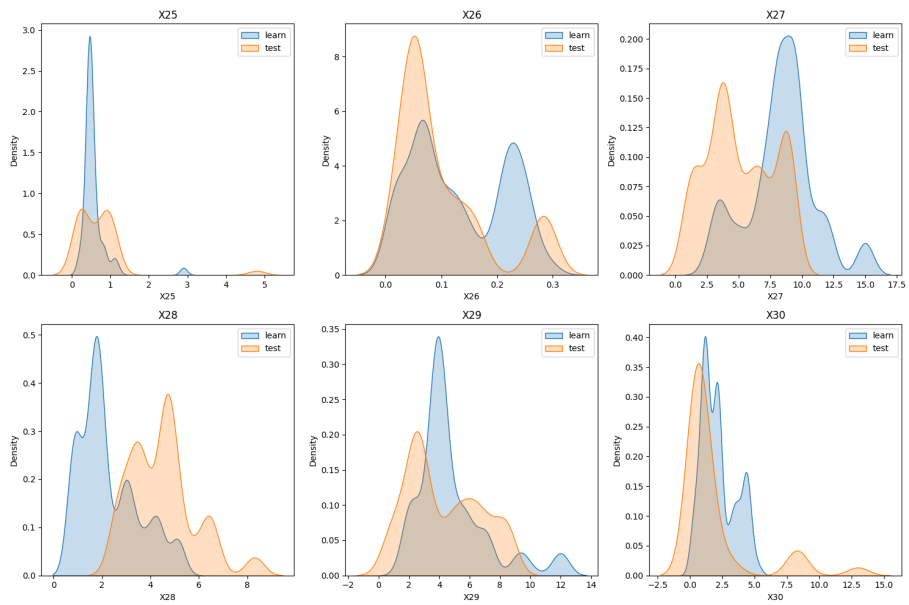


図 7. X25~X30 の訓練・テストデータの分布推移

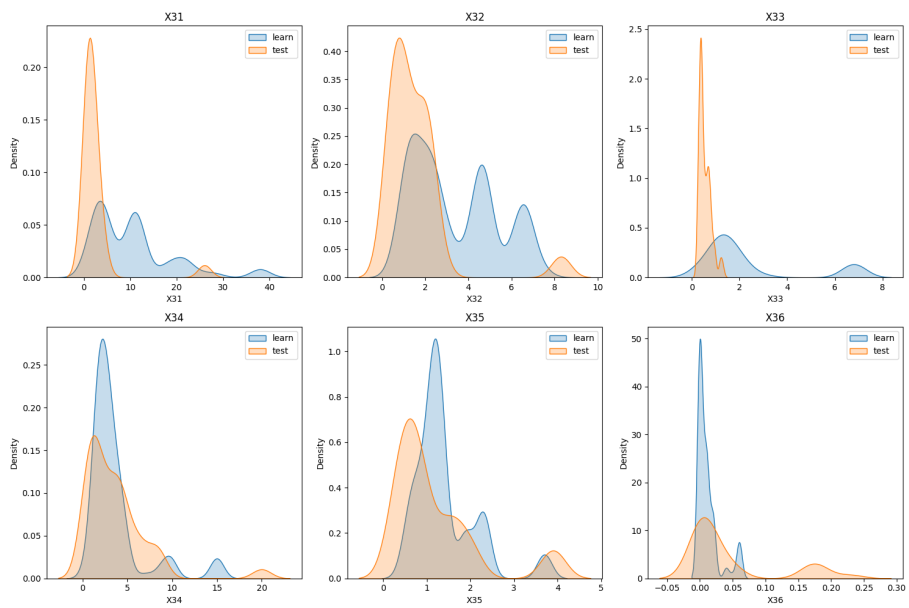


図 8. X31~X36 の訓練・テストデータの分布推移

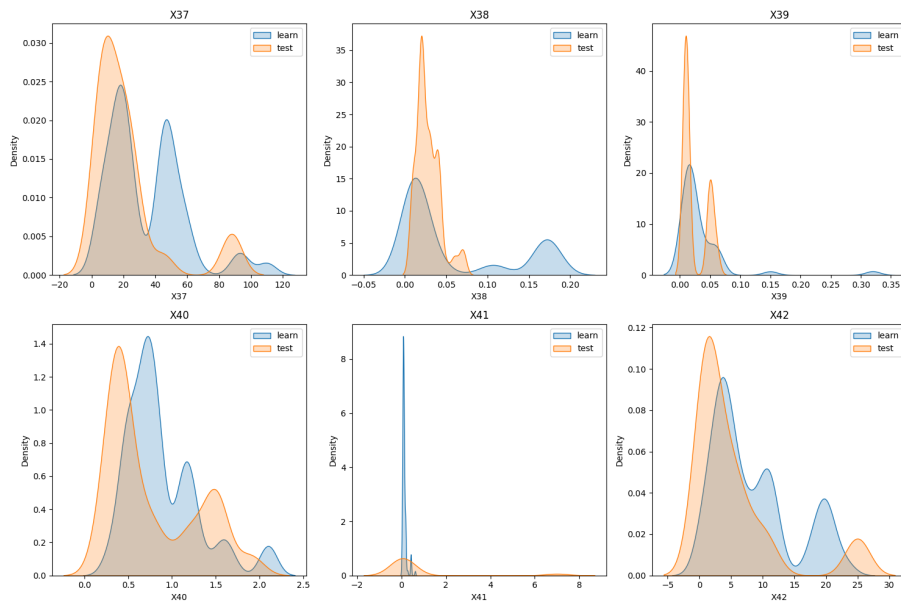


図 9. X37~X42 の訓練・テストデータの分布推移

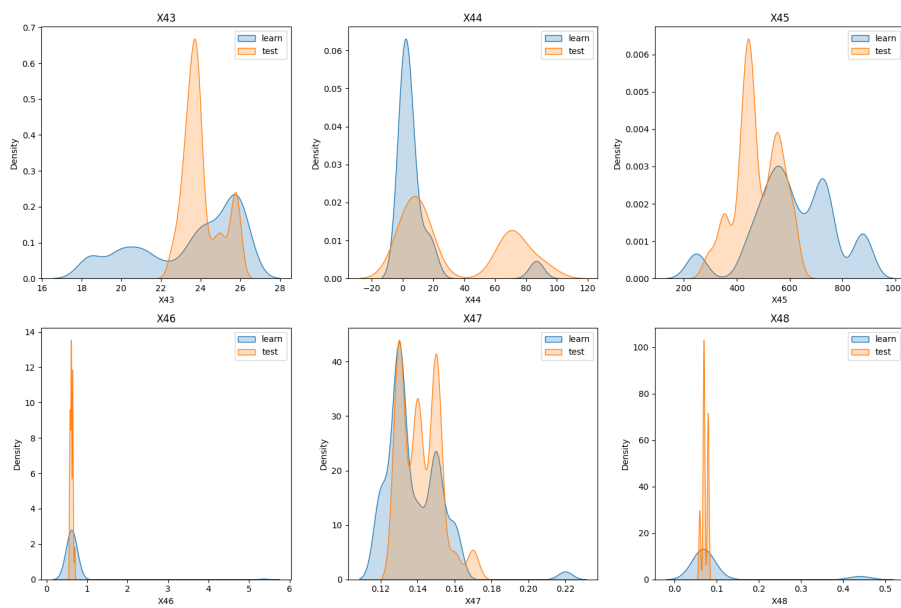


図 10. X43~X48 の訓練・テストデータの分布推移

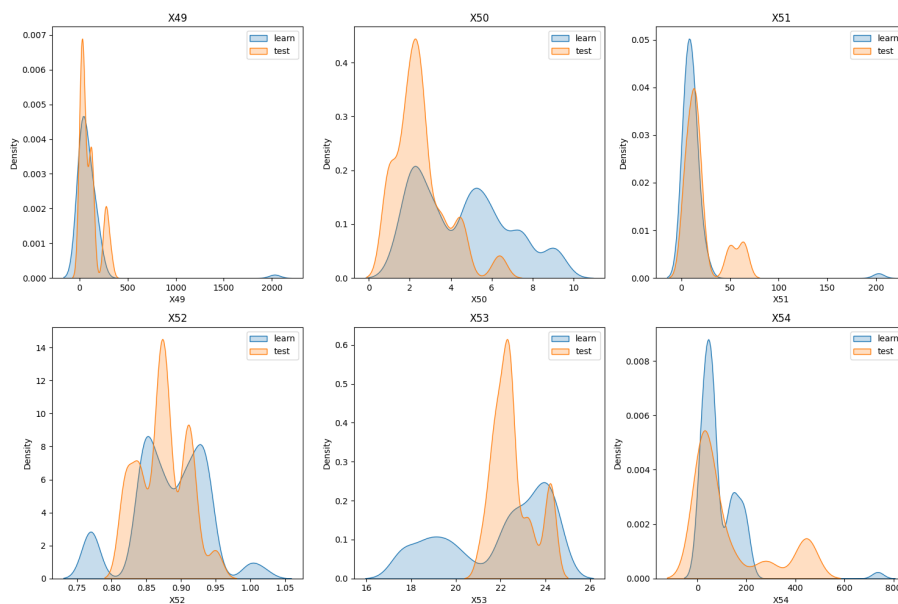


図 11. X49~X54 の訓練・テストデータの分布推移

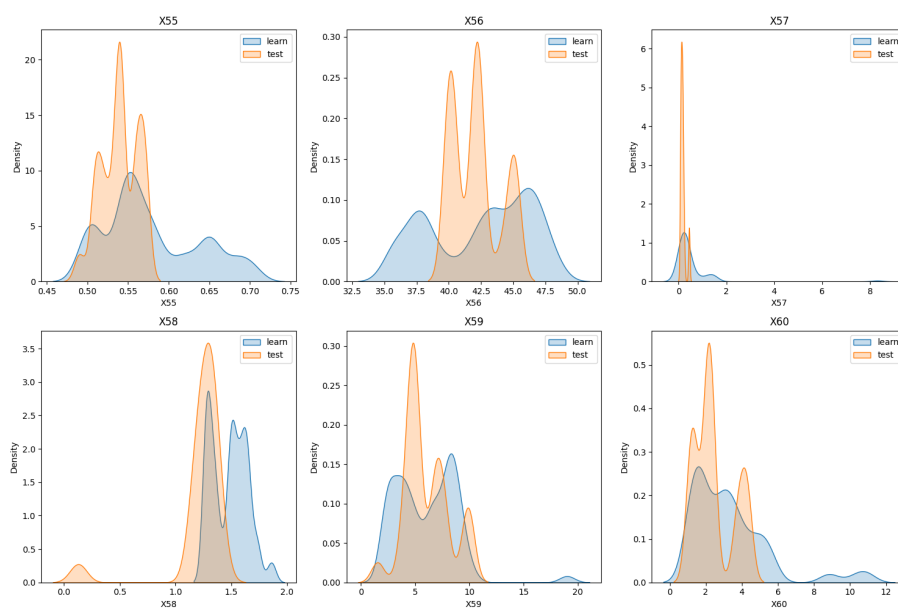


図 12. X55~X60 の訓練・テストデータの分布推移

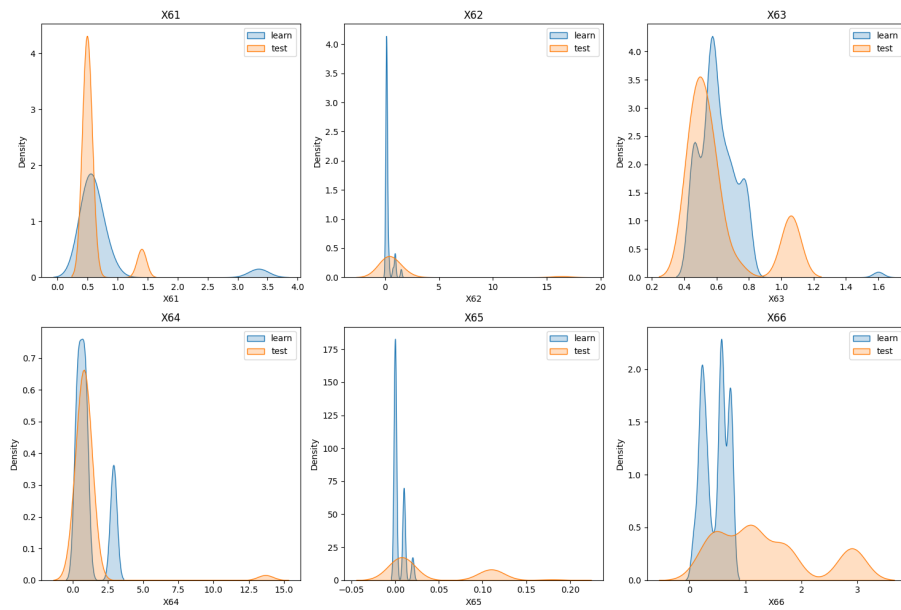


図 13. X61~X66 の訓練・テストデータの分布推移

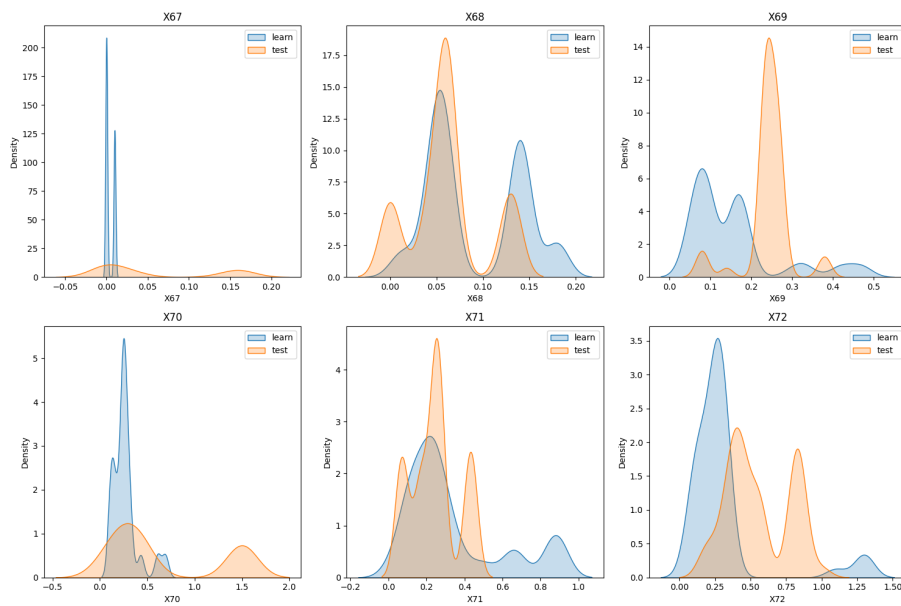


図 14. X67~X72 の訓練・テストデータの分布推移

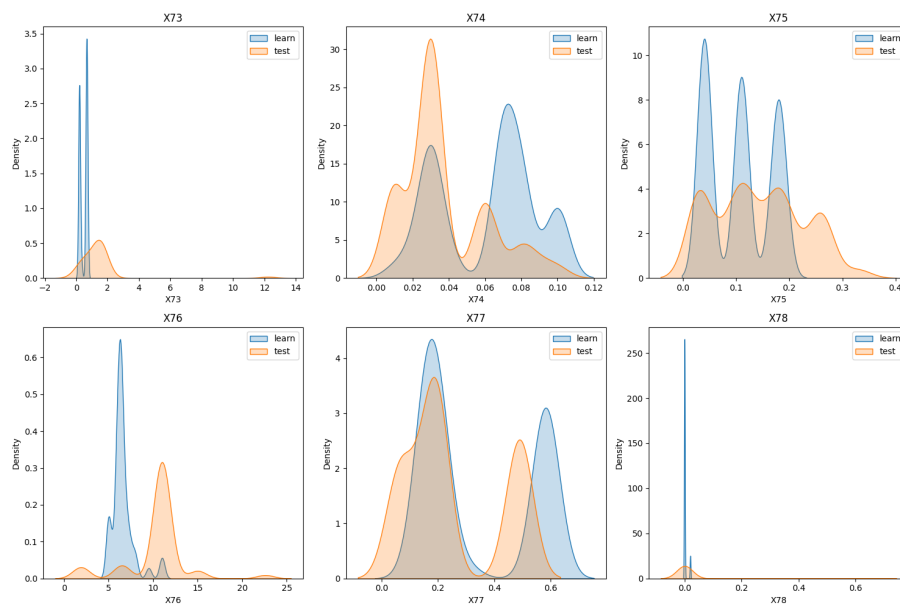


図 15. X73~X78 の訓練・テストデータの分布推移

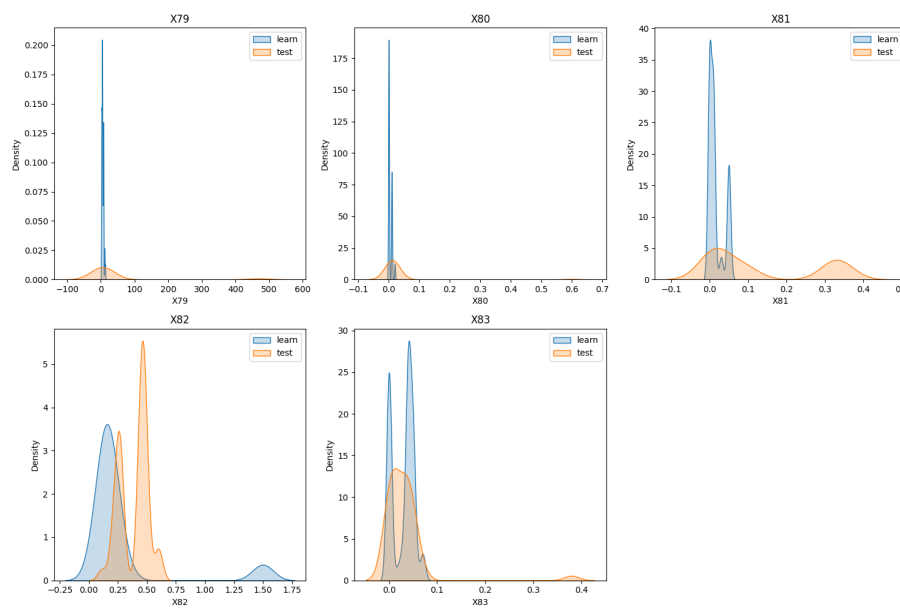


図 16. X79~X83 の訓練・テストデータの分布推移

以上の可視化結果を総合的に確認すると、X1～X83 に含まれる多数の説明変数において、訓練データとテストデータの分布が一致していないことが明らかとなった。この分布差は一部の変数に限定されたものではなく、平均値や中央値といった分布の中心のシフト、分散の増減、分布形状の変化（単峰性から多峰性への変化、あるいはその逆）、さらには外れ値や長い裾の出現といった形で、広範に観測されている。

まず、本分析におけるテストデータは、訓練データの単純な部分集合ではなく、異なる分布特性を持つデータが含まれていると考えられる。品質予測モデルは、訓練データにおいて「説明変数と目的変数 OV の関係性」を学習するため、テストデータが異なる入力分布を持つ場合、モデルは学習時に十分に観測していない領域での予測であり外挿的な予測を行うことになる。このような状況では、モデルの汎化性能が低下し、特に RMSE では、少数の大きな予測誤差が全体の評価値を大きく悪化させる可能性がある。

また、多くの変数において、訓練データとテストデータで分布の広がりが異なっている点も確認された。訓練データでは比較的広い範囲にばらついていた変数が、テストデータでは特定の値域に集中しているケースや、その逆にテストデータのみで裾の長い分布を示すケースが存在する。このような分散の違いは、モデルにおける変数の寄与度や重み付けに影響を与え、訓練時には有効であった説明変数が、テストデータでは十分な情報を持たない、あるいは過度に影響を持つといった不安定な挙動を引き起こす要因となり得る。

さらに、分布形状に着目すると、訓練データとテストデータでピーク数が異なる変数も多く見られた。これは、同一の工程・センサーデータであっても、期間によって複数の状態が異なる割合で混在している可能性を示唆している。品質予測の観点では、どの状態がどの程度出現するかが変化すること自体が予測難易度を高める要因となるため、このような分布構造の変化をモデリング時に考慮する必要があると考える。

一方で、すべての説明変数が同程度に不安定であるわけではなく、X18 や X41、X44、X79 などのように訓練・テスト間で分布の重なりが比較的大きく、中心や形状が概ね一致している変数も一定数存在する。これらの変数は、期間をまたいでも品質に関する情報を比較的一貫して保持していると考えられ、モデルの汎化性能を支える基盤となる可能性がある。しかし、全体として見ると、分布が変化している変数の影響を無視することはできず、説明変数全体としては同一分布仮定が必ずしも成立していない状況にあると判断できる。

以上のことから、本データセットを用いた品質予測では、単に高性能なモデルを選択するだけでなく、分布差の存在を前提としたモデリング戦略が重要となる。具体的には、説明変数の数を可能な限り少なくするという課題条件を踏まえつつ、訓練・テスト間で分布が比較的安定している変数を優先的に採用すること、また外れ値や分布の歪みに対してロバストなモデルや正則化手法を用いることが有効であると考えられ

る。このような設計により、分布シフトの影響を抑えつつ、実運用を想定した品質予測モデルの構築が可能になると考えられる。

次に、説明変数間の相関をヒートマップにして可視化した。可視化した図を以下に示す。

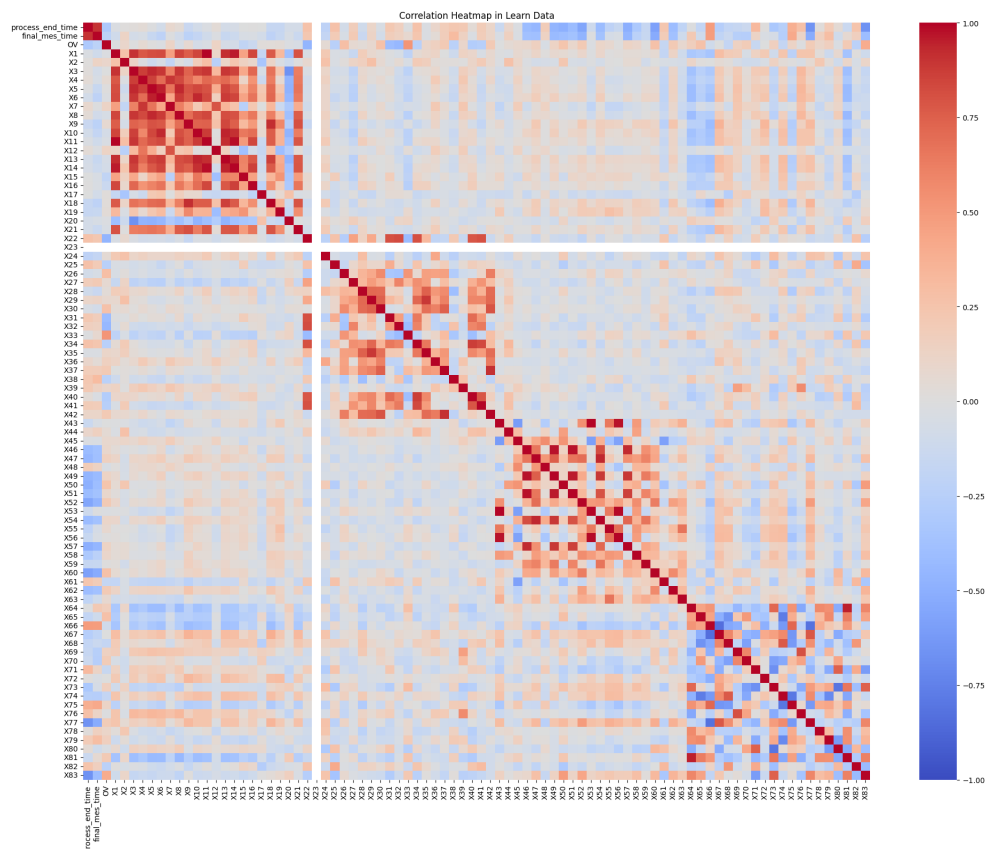


図 17. 相関ヒートマップ

上記をみると、説明変数間には明確な相関構造が存在しており、X1～X83 は互いに独立な特徴量の集合ではないことが分かる。特に、ヒートマップ上で赤色、青点が集中しているブロック状の領域が複数確認でき、これは特定のセンサ群同士が強い相関を持っていることを示している。このような相関構造は、同一工程内で計測されたセンサや、物理的・化学的に密接に関連する指標が含まれている可能性を示唆しており、データが工程構造を反映した形で取得されていることを裏付けている。

一方で、相関の強いブロック間の相互相関は比較的弱く、ブロック同士はある程度独立した情報を持っていると解釈できる。このことから、本データセットは「完全に冗長なデータ」ではなく、いくつかの相関の強い変数群（クラスター）が組み合わさって構成されていると考えられる。ただし、同一ブロック内では相関係数が高い変数が多

数存在するため、これらをそのまま全てモデルに投入すると、多重共線性の影響によりモデルが不安定になる可能性が高い。

また、目的変数である OV と各説明変数との相関を見ると、単一の変数が極端に高い相関を示しているわけではなく、OV は複数のセンサ情報の組み合わせによって決定されていることが示唆される。これは、品質（不良品数）が単一工程や単一センサに強く依存するのではなく、複数工程・複数条件の累積的な影響を受けているという製造プロセス上の直感とも整合的である。そのため、単純な一変量モデルでは十分な予測精度を得ることが難しく、複数の説明変数を適切に組み合わせる必要があると考えられる。

以上を踏まえると、本データを用いたモデリングにおいては、相関の強い変数群から代表的な特徴量を選択する、あるいは正則化や木系モデルのように多重共線性の影響を受けにくい手法を採用することが重要であると考えられる。また、説明変数の数を可能な限り少なくするという課題条件とも整合的に、相関構造を活用して情報の冗長性を削減することは、モデルの安定性向上と解釈性の確保の両面で有効であると考えられる。

次に、OV との変数の相関を相関の強い説明変数ブロックごとにどのような傾向があるのか可視化した。相関を表す棒グラフを以下に示す。

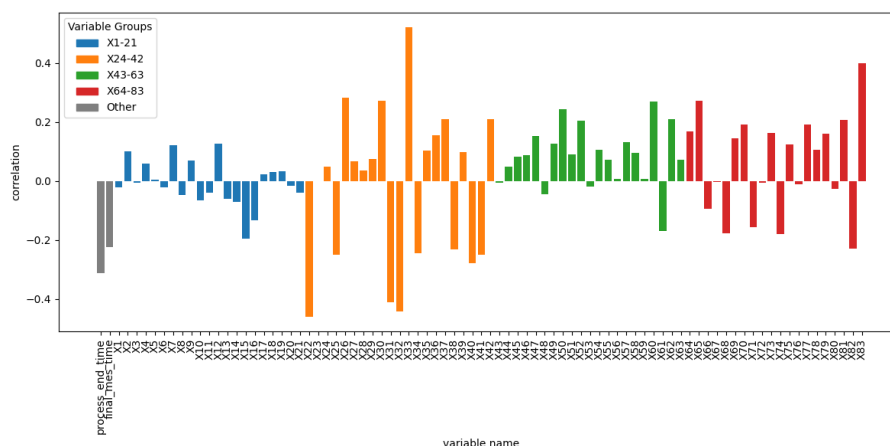


図 18. OV と各説明変数間の相関推移

上記をみると、目的変数である OV と各説明変数との相関は一様ではなく、変数群ごとに異なる傾向を示していることが分かる。まず、X1～X21 に属する変数群では、全体として相関係数の絶対値が比較的小さく、正負いずれの方向についても弱い相関に留まっている変数が多い。このことから、これらの変数は単独では OV を強く説明する情報を持たない可能性が高く、品質予測においては補助的な役割を果たす特徴量群であると考えられる。

一方で、X24～X42 の変数群では、正負いずれの方向においても比較的大きな相関を示す変数が複数確認できる。特に、負の相関が顕著な変数と正の相関が顕著な変数が混在しており、このブロックが OV の増減と密接に関係していることが示唆される。このような傾向は、当該センサーが多く計測しているセクションにおける条件変化が不良品数に直接的な影響を及ぼしている可能性を示しており、品質劣化の主要因がこの変数群に含まれている可能性が高いと考えられる。

また、X43～X63 の変数群では、全体として正の相関を示す変数が多く、OV の増加に伴ってこれらのセンサ値も増加する傾向が見られる。一部には負の相関を示す変数も存在するものの、ブロック全体としては比較的一貫した方向性を持っている点が特徴的である。このことから、この変数群は品質悪化を表す「状態指標」として機能している可能性があり、OV の変動を捉える上で有用な情報を提供していると考えられる。

さらに、X64～X83 の変数群では、相関係数の絶対値が大きい変数が複数存在し、特に正の相関が強く表れている点が確認できる。このブロックは、OV と強い関係を持つ変数が集中している領域であり、品質予測モデルにおいて中核的な役割を果たす特徴量群である可能性が高い。一方で、同一ブロック内において相関の向きや大きさにばらつきがあることから、これらの変数を全て同時に使用すると多重共線性の影響を受けやすく、モデルの不安定化を招く恐れがある点には注意が必要である。

以上を総合すると、OV は特定の単一変数によって強く規定されているのではなく、相関構造を持つ複数の変数群の影響を同時に受けていると考えられる。この結果は、前節で確認した相関ヒートマップにおけるブロック構造とも整合的であり、製造プロセスが複数の工程・状態の組み合わせによって品質に影響を与えていると考える。

したがって、本課題の条件である「可能な限り少ない説明変数の数で高い予測精度を達成する」という目的を達成するためには、単純に相関の大きい変数を機械的に選択するのではなく、相関ブロックごとに代表的な変数を選択する、あるいは正則化手法や木系モデルを用いて冗長な情報を抑制しつつ重要な情報を抽出する戦略が有効であると考えられる。

特に、相関が強い変数群から少数の代表変数を抽出することで、モデルの解釈性を高めると同時に、監視すべきパラメータ数を削減するという実運用上の要請にも応えることができる。このような観点から、次節以降のモデリングでは、相関構造と分布安定性の両方を考慮した特徴量選択を行い、品質予測モデルの構築を進めることが妥当であると判断した。

6. 機械学習自動化ライブラリを用いた分析手法の検討

機械学習自動化ライブラリ(h2o)を用いて GBM モデルを構築した結果を以下に示す。

表 1. 変数重要度の上位 10 変数

	variable	relative_importance	scaled_importance	percentage
0	X34	2.233030e+06	1.000000	0.157967
1	X32	1.412960e+06	0.632754	0.099954
2	X31	1.180889e+06	0.528828	0.083537
3	X33	7.105086e+05	0.318181	0.050262
4	X25	7.099123e+05	0.317914	0.050220
5	X37	6.934696e+05	0.310551	0.049057
6	X40	6.618311e+05	0.296382	0.046819
7	X72	5.071781e+05	0.227125	0.035878
8	X22	5.014171e+05	0.224546	0.035471
9	X41	4.773238e+05	0.213756	0.033766

MSE: 5812.25142039866

RMSE: 76.23812314320612

MAE: 67.92062194142275

RMSLE: 0.656052249504877

Mean Residual Deviance: 5812.25142039866

gbm

GBM_grid_1_AutoML_2_20260122_202917_model_9

図 19. h2o のモデリング結果

本結果から、目的変数 OV の予測においては、全説明変数が一様に寄与しているわけではなく、特定の少数変数が予測性能に大きく寄与していることが明らかとなった。

最も高い重要度を示したのは X34 であり、寄与率は約 16% に達している。これは、モデルが行った分割の中で、X34 が最も頻繁かつ効果的に用いられたことを意味しており、当該センサが品質変動を捉える上で極めて重要な情報を含んでいる可能性を示唆している。続いて X32、X31、X33、X25 といった変数が高い重要度を示しており、上位数変数が予測において中核的な役割を果たしていることが分かる。

注目すべき点として、これら上位変数の多くは、前節で確認した OV との相関が比較的大きい変数群 (X24~X42、X64~X83) に属している。これは、単純な相関分析によって示唆された「品質変動に影響を与える工程セクション」が、GBM という非

線形モデルにおいても同様に重要視されていることを意味しており、可視化分析と機械学習モデルの結果が整合的であることを示している。

一方で、相関係数が比較的大きかったすべての変数が高い重要度を示しているわけではない点も重要である。GBM は木構造に基づくモデルであり、相関の強い変数群が存在する場合、それらの中から代表的な変数が優先的に分割に用いられる傾向がある。このため、同一ブロック内に属する他の変数は、情報としては有用であっても、モデル内での重要度が相対的に低く評価される可能性がある。これは、前節で指摘した多重共線性の問題を、GBM が内部的に緩和しつつ処理している結果と解釈できる。

また、本課題の条件である「可能な限り少ない説明変数の数で高い予測精度を達成する」という観点から見ると、本結果は非常に示唆的である。83 個存在するセンサ変数のうち、GBM モデルが特に重要と判断した変数はごく一部に集中しており、品質予測に必要な情報は限られたセンサから得られている可能性が高い。これは、実運用において監視すべきパラメータ数を削減したいという VM の目的とも整合的であり、h2o を用いた変数重要度分析が、単なる予測精度向上だけでなく、監視設計の合理化にも寄与し得ることを示している。

ただし、ここで得られた変数重要度は、あくまで GBM モデル内部における相対的な指標であり、各説明変数が OV に対して因果的にどの程度影響を及ぼしているかを直接示すものではない。従って、本分析結果を踏まえ変数選択とモデル選択を継続することとした。

7. 説明変数のクラスタリングと複数モデルによる予測

本分析で利用するデータは、ヒストグラムからわかる通り単峰的ではないため非線形なモデルを適用するか、複数のセグメントに分割し、モデルを使い分けることが理想的と言える。そこで、説明変数を用いてクラスタリングしその結果を元にモデルを使い分けることで、OV の予測精度を挙げられるよう試みた。

1. モデル選択の前処理: 説明変数クラスタリング

はじめに、説明変数を用いてクラスタリングした。本分析では説明変数間に相関の強いグループが 4 つ存在していた。それぞれの説明変数グループを以下に示す。

表 2. 説明変数グループ

グループ	該当説明変数
A	X1~X21
B	X22~X42
C	X43~X63
D	X64~X83

上記のグループごとに、k-means 法でクラスタリングし、そのグループに属するか one-hot-encode した結果間の相関ヒートマップを作成した結果を以下に示す。

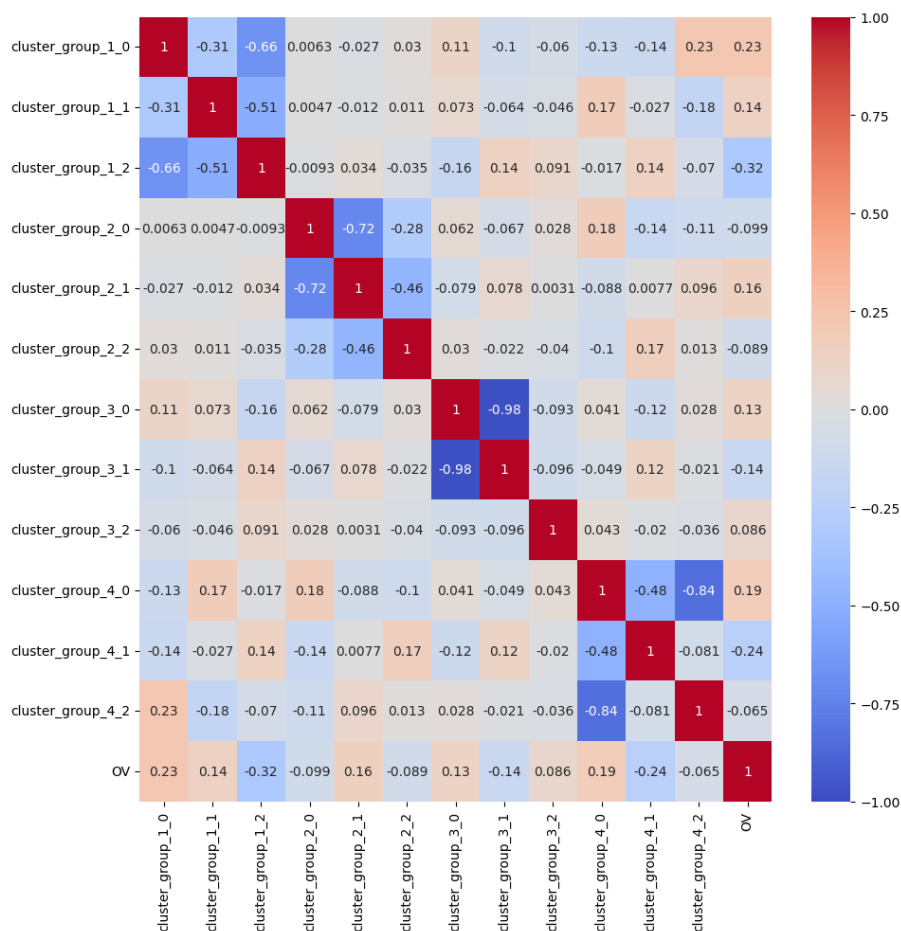


図 20. group ごとにクラスタリングした際の相関ヒートマップ

※cluster_group_{説明変数グループ}_{クラスタリング No.}

上記を見ると、クラスごとに OV との相関があり、特に`cluster_group_1_2` と OV の相関の絶対値が高いと言える。そこで、どのレコードが group_1_2 に該当しているのか確認するため、以下の図を作成した。

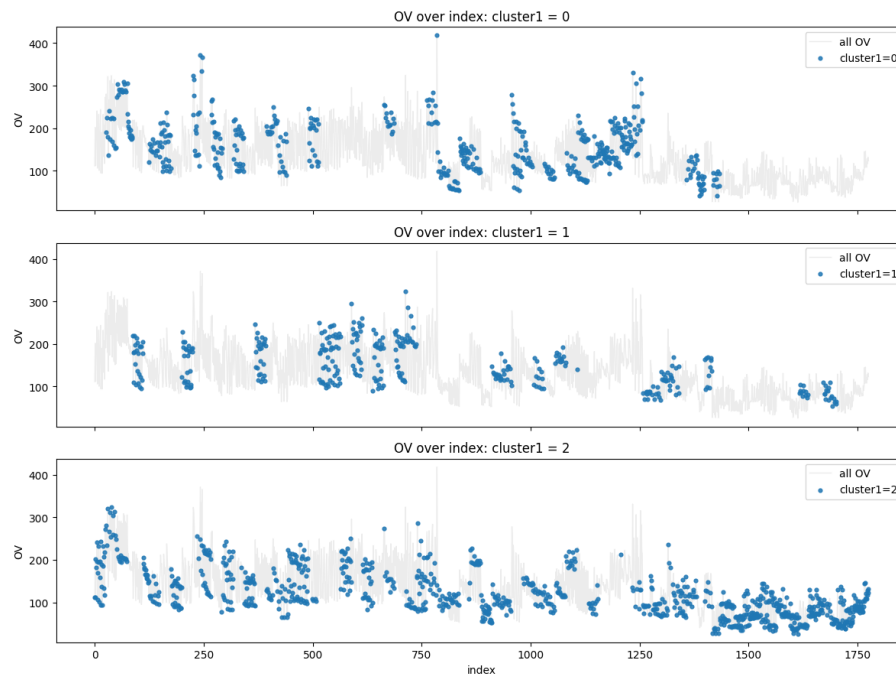


図 21. group_1 の分布

上記の図は、group1 に属するデータについて、クラスタリングによって得られた 3 つのクラス (cluster1 = 0, 1, 2) ごとに、インデックス順（製造順）で目的変数 OV の推移を可視化したものである。各図では、全体の OV 推移を背景として薄く描画し、その上に該当クラスに属するデータ点を強調表示している。

まず全体的な傾向として、OV の値はインデックスに沿って緩やかな低下傾向を示しており、製造期間の後半に向かうにつれて不良品数が減少している様子が確認できる。このことから、本データは時間的に非定常な構造を持っており、製造条件や工程管理状態が期間とともに変化している可能性が示唆される。このような非定常性の存在は、時系列予測や品質予測モデルにおいて重要な前提条件となる。

次に、各クラスにおける点の分布構造に注目すると、いずれのクラスにおいても、データ点がインデックス方向に一樣に分布しているのではなく、特定のインデックス範囲に集中して出現していることが分かる。特に cluster1 = 1 の図の左側では、点が縦方向に並ぶ縦縞状の分布が顕著に観察される。この縦縞構造は、同一または近接したインデックスにおいて複数のロットが連続して製造され、それらが同一クラスに分類されていることを意味している。

この現象は統計的には、クラスタ割当てが独立同分布ではなく、時系列的な自己相関を持っていることを示唆している。よって、あるロットが特定のクラスに属する場合、その前後のロットも同一クラスに属する確率が高く、製造工程の状態が短時間では大きく変化しないことを反映していると解釈できる。これは、同一時間

帯に製造されたロットが類似した説明変数（センサ値）の値を持つという製造プロセス上の直感とも整合的である。

一方で、クラス間での OV 分布に着目すると、 $\text{cluster1} = 0, 1, 2$ のいずれにおいても、OV の取り得る値域やばらつきの大きさは概ね類似しており、クラス間で明確な分布の分離は確認できない。例えば、いずれのクラスにおいても高い OV 値と低い OV 値が混在しており、特定のクラスが一貫して高不良または低不良を示すわけではない。このことは、クラスタリングが主として説明変数の類似性に基づいて行われており、目的変数である OV を直接的に分離する構造を持っていないことを意味している。

統計的に見れば、本クラスタリングは「説明変数空間における局所的な状態の違い」を捉えることには成功しているものの、「OV の条件付き分布がクラスごとに有意に異なる」というレベルには至っていないと評価できる。したがって、各クラスを切り替え条件として異なる OV 予測モデルを構築した場合、モデル間で予測対象となる OV の分布が大きく異ならないため、切り替えによる性能向上効果は限定的であると考えられる。

以上より、 group1 におけるクラスタリング結果は、製造工程の時間的連続性や説明変数の局所的な類似性を反映した構造を持つ一方で、OV の分布を明確に分離する指標としては十分ではない。このため、本クラスタを OV 予測モデルのスイッチング条件として直接利用することは適切ではなく、むしろ「工程状態の遷移や安定区間を把握するための補助的な情報」として位置付けることが妥当であると評価する。

2. モデル選択の前処理: ロットごとの基本統計量から IsolationForest でスパイク検知前項で得られた示唆を元に、ロットごとに特徴を踏まえることで、適したモデル選択を実施できると考え、ロットごとに特性を予測するモデルを構築することとした。はじめに、ロットごとの OV 分布を以下に示す。

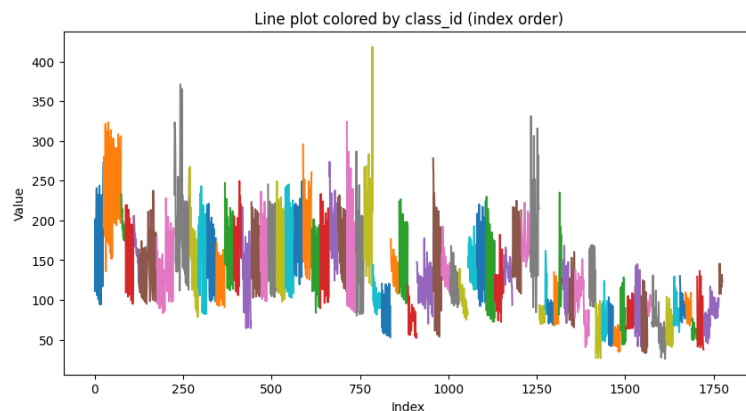


図 22. ロットごとに色分けした OV 推移

上記の図は、各ロットにおける目的変数 OV の推移をインデックス順に可視化し、クラスごとに色分けしたものである。この図から明らかなように、OV の推移には大きく二種類の挙動が存在する。すなわち、短いインデックス区間において OV が急激に増減し、縦方向に大きく伸びる「スパイク状」の挙動を示すロットと、比較的緩やかな変動を保ちながら推移するロットである。

このような挙動の違いは、単なるノイズの大小として片付けられるものではなく、確率過程としての性質がロットごとに異なる可能性を示唆している。具体的には、なだらかに推移するロットでは、OV の変動が比較的安定した分布（分散が小さい条件付き分布）に従っている一方で、スパイクを伴うロットでは、突発的な外乱や工程状態の変化により、分散が一時的に大きくなる非定常な挙動が生じていると解釈できる。

統計学的観点から見ると、このようなデータ構造は「全ロットが同一の確率分布から生成されている」という同分布仮定を満たしていない可能性が高い。すなわち、OV は単一の平均・分散を持つ分布に従うのではなく、ロットの状態に応じて分散構造や分布形状が切り替わる混合分布として生成されていると考える方が自然である。この場合、全データを一括して学習した単一モデルは、スパイクを伴うロットと安定したロットの双方を同時に説明しようとするため、結果としてどちらに対しても十分に適合しない「平均的なモデル」になりやすい。

また、スパイクの存在は、目的変数の分布に裾の重い構造をもたらす要因となる。RMSE のような二乗誤差に基づく評価指標では、少数の大きな誤差が全体の評価値を支配するため、スパイクが発生するロットを適切に捉えられないモデルは、全体として著しく性能が低下する恐れがある。一方で、スパイクを過度に重視したモデルは、安定したロットに対して過剰に変動の大きい予測を行い、こちらでも誤差が増大する可能性がある。

このような構造を踏まえると、ロットごとの推移形状に応じて予測モデルを切り替える、あるいはロットをいくつかの状態に分類した上で、それぞれに適したモデルを適用するという戦略は、統計的に妥当なアプローチであると考えられる。すなわち、スパイクが頻発するロットに対しては、外れ値や急変に対してロバストなモデルや、短期的な変動を重視したモデルを用い、なだらかに推移するロットに対しては、分散の小さい前提のもとで安定的な関係性を学習するモデルを用いることで、それぞれの条件付き分布により適合した予測が可能となる。

さらに、ロットごとの挙動の違いを事前に識別することができれば、予測モデルの精度向上だけでなく、製造工程の監視や異常検知の観点からも有用な情報が得られる。例えば、通常はなだらかな推移を示すロットが突如スパイク的挙動に移行した場合、それ自体を工程異常の兆候として捉えることができる。このように、推移

形状に基づくロット分類は、単なる予測精度向上に留まらず、品質管理上の意思決定支援にも寄与し得る。

以上より、本データにおける OV の推移は、単一の確率モデルで一様に扱うことが困難な異質性を内包しており、ロットごとの推移特性を考慮したモデリング戦略が必要であると結論付けられる。本分析では、この問題意識に基づき、スパイク的挙動を示すロットと安定的に推移するロットを区別した上で、それぞれに適した予測モデルを構築することにより、単一モデルによる予測と比較して、より高い予測精度を達成できる可能性を検討することとした、そこで、スパイクに該当するロットを OV の分散が高いロットとして解釈できると推測し、分散 50 以上を対象に抽出した結果を以下に示す。

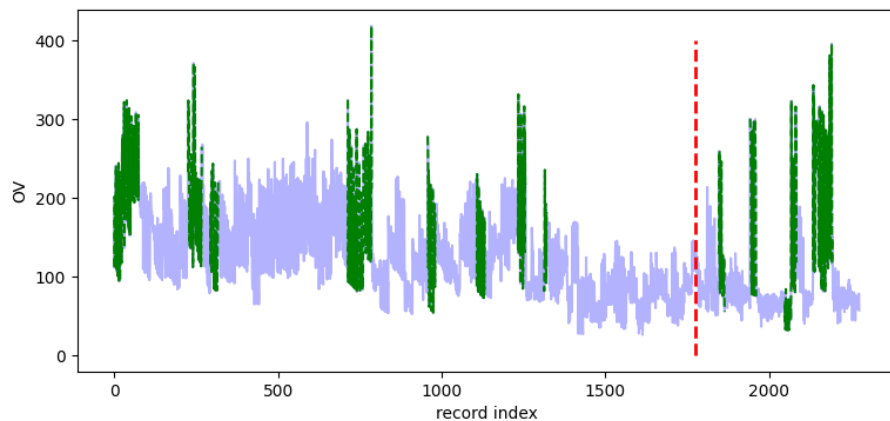


図 23. OV の分散 50 以上のロット

上の図は、ロットごとの OV の推移に着目し、ロット内における OV の分散が 50 を超えるものを抽出して可視化した結果を示している。背景として全体の OV 推移を薄色で示し、その上に分散の大きいロットを強調表示することで、OV の変動構造を時系列的に把握できるようにしている。

本図から、OV の推移には一様な変動構造は存在せず、短いインデックス区間で急激な増減を示す「スパイク的挙動」を持つロットと、比較的滑らかに推移するロットが混在していることが確認できる。このことから、ロットごとに OV の分散構造が大きく異なっており、OV が単一の分布から生成されているという同分散・同分布の仮定は成立していないと考えられる。すなわち、本データは安定状態と不安定状態が混在した混合的な生成構造を持つと解釈できる。

ここで用いたロット内分散 50 という基準は、通常状態における OV の変動幅と比較して明確に変動が拡大する境界として設定したものであり、安定ロットとスパイクロットを識別する実用的な指標として機能している。また、スパイクロット

はランダムに出現するのではなく、特定のインデックス区間に集中しており、製造工程の状態が時間とともに変化している可能性を示唆している。

このような構造を持つデータに対して、全ロットを一括して単一の予測モデルで学習する場合、安定状態とスパイク状態の双方を同時に説明する必要が生じ、結果として予測精度が低下する恐れがある。特に RMSE のような二乗誤差指標では、少数のスパイクが評価値を支配しやすい。

以上より、ロット内分散に基づいてスパイク的挙動を示すロットを識別し、安定ロットとは異なる分布として扱うことは統計学的に妥当であると考えられる。

そして、このスパイクロットの抽出は、目的変数 OV を用いた事後的評価であるため、説明変数のみに基づく事前検知手法として Isolation Forest を用いた異常検知を行った。OV の分散が小さいロットを平常ロットと定義し、その説明変数分布を基準として学習したモデルを全ロットに適用した。

以下の図は、Isolation Forest により異常と判定されたロットを時系列上に示したものである。

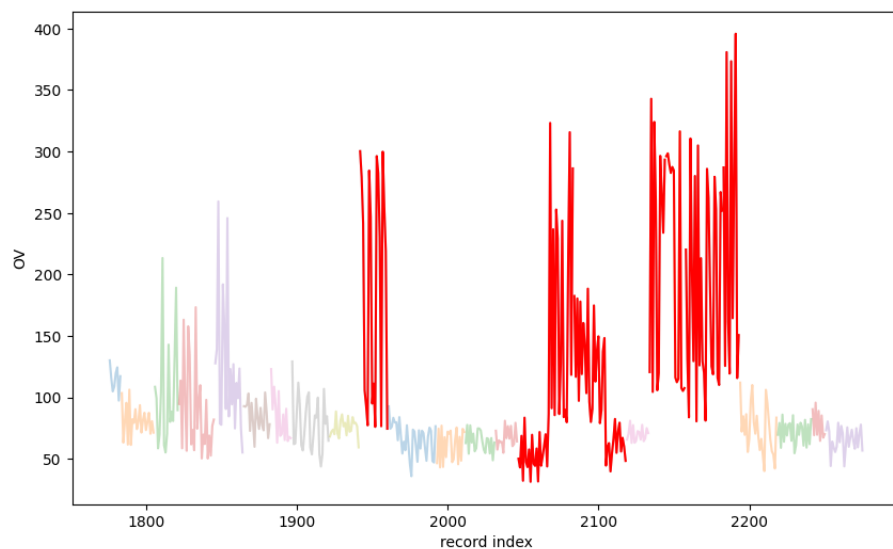


図 24. スパイク検知 Isolation Forest の予測結果(テストデータ)

※説明変数は、["X36", "X27", "X30", "X41"]

上記を見ると、OV の分散を利用して抽出されたスパイクロットと出現タイミングがおおむね一致しており、説明変数のみからスパイク的状态を一定程度再現できていることが確認できる。一方で一致しない区間も存在し、説明変数では捉えきれない要因の影響も示唆される。

以上の結果から、Isolation Forest を用いた説明変数ベースの異常検知は、スパイクロットの事前検知手法として有効な可能性を持つと評価できる。

3. OV 予測モデルの構築

前節までの分析により、本データセットにおける目的変数 OV は、説明変数との関係性が単純な線形構造ではなく、さらに時間経過や工程状態の変化に伴って分布特性が変化する非定常な性質を持つことが示唆された。また、ロットごとの OV 推移を観察すると、なだらかな変動を示す区間と、突発的に大きな変動（スパイク）を示す区間が混在しており、すべてのデータを同一の確率構造として扱うことは適切でないと判断される。このようなデータ特性を踏まえ、本分析では「スパイク的挙動を示すロット」と「比較的安定した挙動を示すロット」を区別し、それぞれに適した予測モデルを構築する二段階構成の予測フレームワークを採用した。すなわち、まず説明変数のみを用いてスパイク発生の可能性を事前に判定し、その結果に基づいて OV 予測モデルを切り替える構成とした。

1. モデルの全体像

(ア) スパイク検知モデル

ロットごとの OV 分散を用いてスパイクロットを除いたフラットロットのみを学習データとした Isolation Forest で、スパイクロットを検出するモデルを作成した(前項)。

(イ) OV 予測モデル 1(スパイクロット用)

スパイク発生が予測されたロットに対して適用する、スパイクを重視した回帰モデルを作成した。

(ウ) OV 予測モデル 2(フラットロット用)

スパイクが発生していないと判定されたロットに対して適用する回帰モデルを作成した。

2. OV 予測モデルに利用する説明変数の選択

OV 予測モデルでは、前節までの相関分析および h2o による変数重要度分析の結果を踏まえ、説明変数を必要最小限に絞り込んだ。特に、以下の観点を重視した。

- ・ 訓練データとテストデータ間で分布が比較的安定していること
- ・ OV との相関が比較的大きく、かつ他変数との冗長性が高すぎないこと
- ・ 実運用において監視対象として現実的な変数数であること

その結果、OV 予測モデルでは主に 'X15', 'X30', 'X43', 'X68', 'X2', 'X24', 'X33', 'X37' など、GBM において重要度が高く、相関構造上も中核を成す変数群を中心に構成した。

3. モデルの学習方法

OV 予測モデルの学習は、課題条件に従い以下のデータ分割に基づいて実施した。

表 3. データ分割

訓練データ	テストデータに該当しないデータのうち、テストデータ先頭レコードの process_end_time より前の final_mes_time を持つレコード
テストデータ	末尾 500 レコード

また、スパイクロットにおいては OV の分散が大きく、通常の最小二乗誤差に基づく学習ではスパイクが過小評価される恐れがある。そのため、スパイクロット用モデルでは 高 OV 領域の誤差を相対的に重視する重み付き学習を導入し、品質悪化時の予測性能を高める設計とした。

4. 評価表新と評価指標

モデルの評価指標には、課題条件に従い RMSE (Root Mean Squared Error) を用いた。特に、本分析では以下の 2 点を重視して評価を行った。

- ・ 全テストデータに対する RMSE
- ・ スパイク区間を含むロットにおける RMSE
- ・ フラット区間を含むロットにおける RMSE

これは、全体の平均的な性能だけでなく、実運用上重要となる「品質が悪化する局面での予測性能」を適切に評価するためである。

5. 結果

スパイク区間には Random Forest Regressor、フラット区間には StandardScaler と LightGBM Regressor を組み合わせたパイプラインを用いた。結果、最良の説明変数組合せは以下の通りであった。

表 4. 利用した説明変数

スパイク区間	X15, X30, X43, X68
フラット区間	X2, X24, X33, X37

このときの予測精度（RMSE）は以下の通りである。

表 5. 予測精度

全体	44.561
スパイク区間	72.299
フラット区間	24.947

予測結果と真値を以下の図に示す。

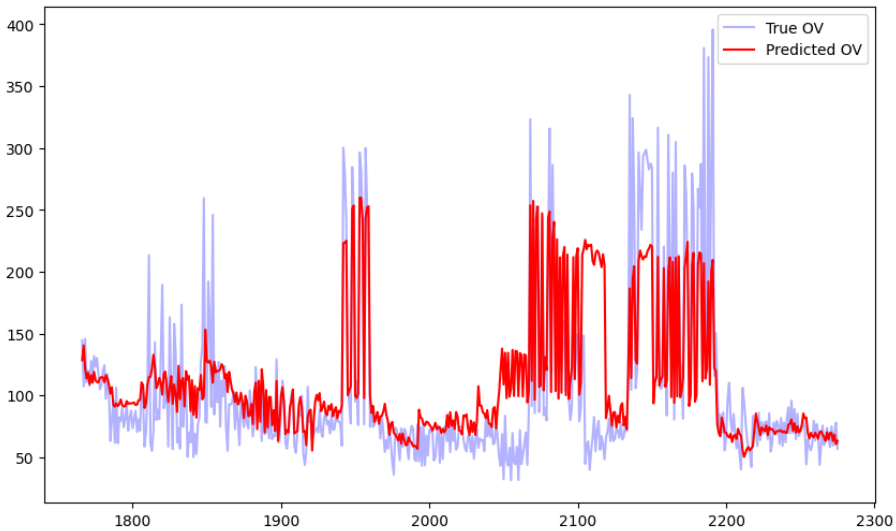


図 25. 予測結果

図に示す通り、フラット区間においては予測値（赤線）が真値（青線）に良好に追従しており、OV の水準変化や緩やかなトレンドを安定的に捉えられていることが確認できる。一方で、スパイク区間ではピーク値を完全には再現できていないものの、スパイク発生タイミング自体は概ね捉えられており、単一モデルによる予測と比較して極端な過小評価は抑制されていると評価できる。

8. 全体の考察

本課題では、製造プロセスにおける品質指標 OV を対象として、相関分析や変数重要度分析を通じてデータの構造を把握した上で、予測モデルの構築を行った。分析の結果、OV は特定の単一変数によって決定されるものではなく、複数工程・複数変数の組み合わせによって影響を受ける指標であり、さらに時間経過や工程状態の変化に伴って分布特性が変化する非定常な性質を持つことが明らかとなった。また、ロットごとの OV 推移を観察すると、比較的安定した挙動を示す区間と、突発的に大きな変動を示すスパイク区間が混在しており、全データを同一の確率構造として扱うことは適切でないと判断された。

これらのデータ特性を踏まえ、本分析ではスパイク検知と予測モデルの切り替えを組み合わせた二段階構成のフレームワークを採用した。その結果、特にフラット区間においては予測値が真値に良好に追従し、RMSE も 24.947 と比較的低い値を達成した。このことから、OV が安定している状態においては、説明変数と目的変数の関係が比較的安定しており、機械学習モデルによる予測が有効であることが示唆される。一方で、スパイク区間ではピーク値を完全に再現することは困難であったものの、スパイク発生のタイミング自体は概ね捉えられており、単一モデルによる予測と比較して極端な過小評価は抑制されている。

スパイク区間の予測精度が相対的に低くなった要因としては、スパイクが外生的要因や未観測変数の影響を強く受ける可能性が高いことに加え、OV の分散が大きく外れ値的挙動を含むため、RMSE が大きくなりやすいという評価指標の特性も影響していると考えられる。この点を踏まえると、本分析におけるスパイク区間用モデルは、スパイクの大きさを精密に予測するというよりも、品質悪化が生じる兆候を事前に捉えることを重視した設計として妥当であったと言える。また、最終的に採用された説明変数は 8 変数と比較的少数であり、実用的な予測精度を達成できた。

以上より、本課題で構築した予測モデルは、非定常性や異質性を持つ品質データに対して、単一モデルではなく状態に応じてモデルを切り替えるという考え方の有効性を示したものであり、製造プロセスにおける品質監視や予兆検知の観点からも有用であると結論付けられる。一方で、スパイク区間におけるピーク値予測精度の向上や、外生情報を含めたモデル拡張、残渣学習したモデルの構築、テストデータ一括予測ではなく逐次学習と予測による精度向上といった課題も残されており、これらは今後の検討課題とする。

9. 参考文献

1. Peter Bruce、Andrew Bruce、 Peter Gedeck , 『データサイエンスのための統計学 入門 第2版』 , Oreilly Japan
2. 門脇 大輔, 『Kaggle で勝つデータ分析の技術』 ,2019.
3. 井手剛, 『入門機械学習による異常検知: R による実践ガイド』 , 2015.

10. 付録

今回の分析で利用した分析ファイルを GitHub リンクとして以下に示す。