# 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 髙木悠人

---

## 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を用いた分析手法の検討

4. モデリング
   データの可視化を踏まえ

## 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。

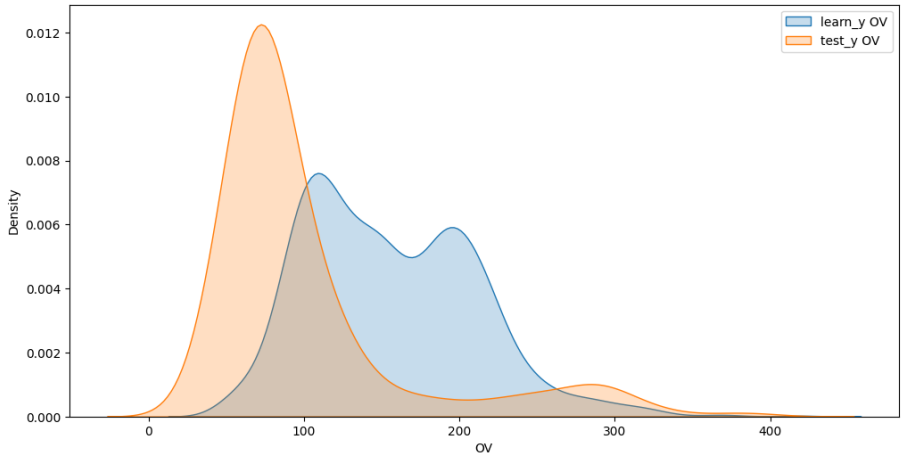


**図. 目的変数 OV の訓練・テストデータの分布**

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。

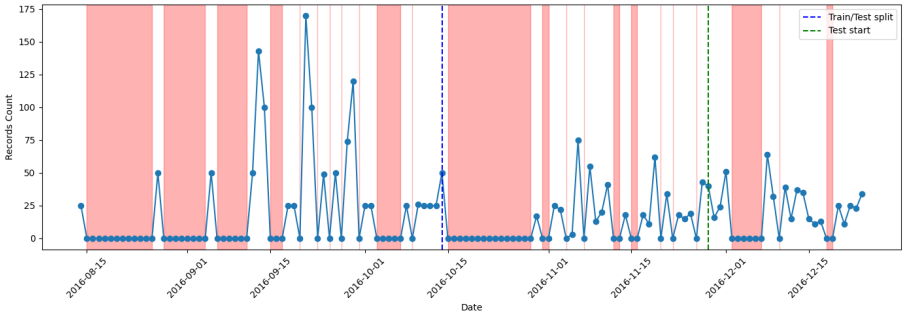


**図. 製造日と製造数の推移**

※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。

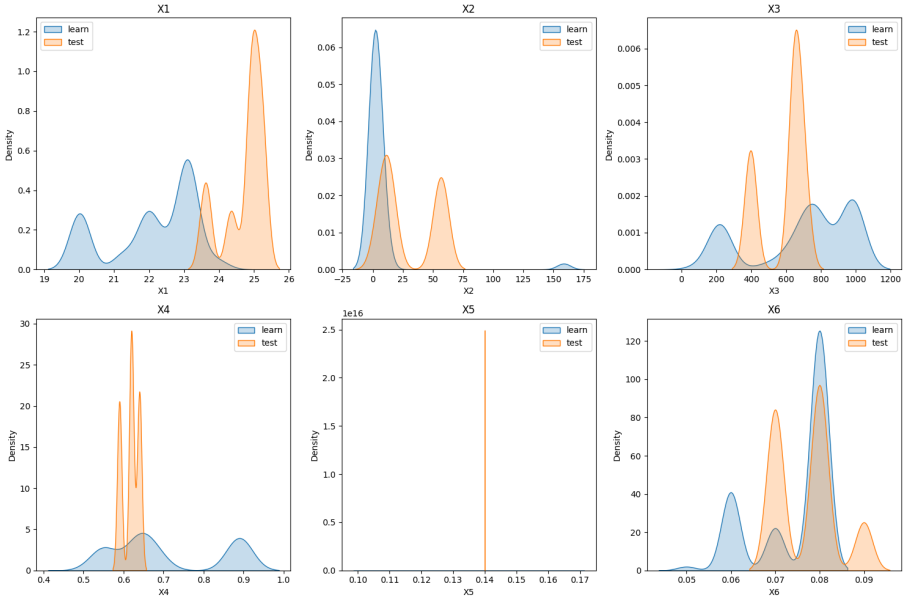


図. X1~X6 の訓練・テストデータの分布推移

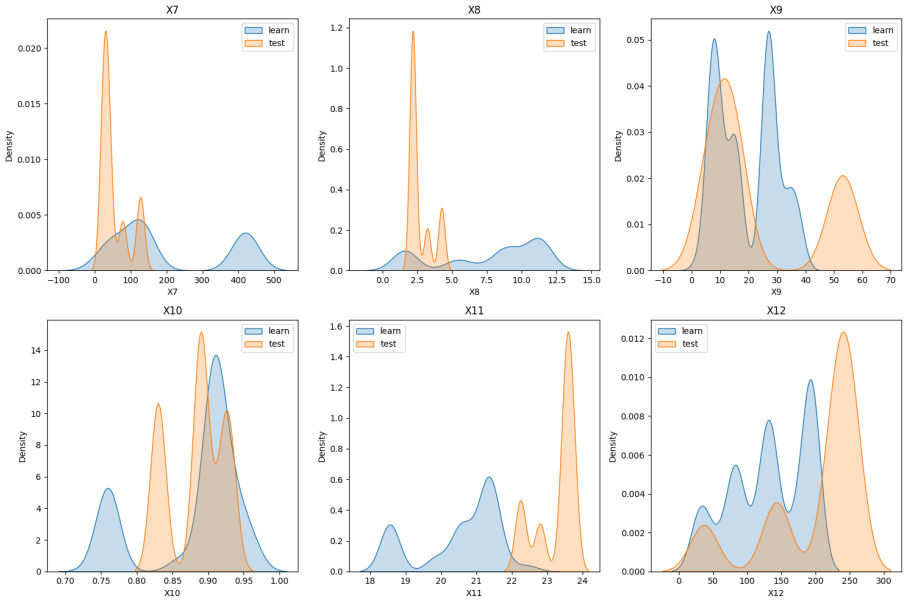


図. X7~X12 の訓練・テストデータの分布推移

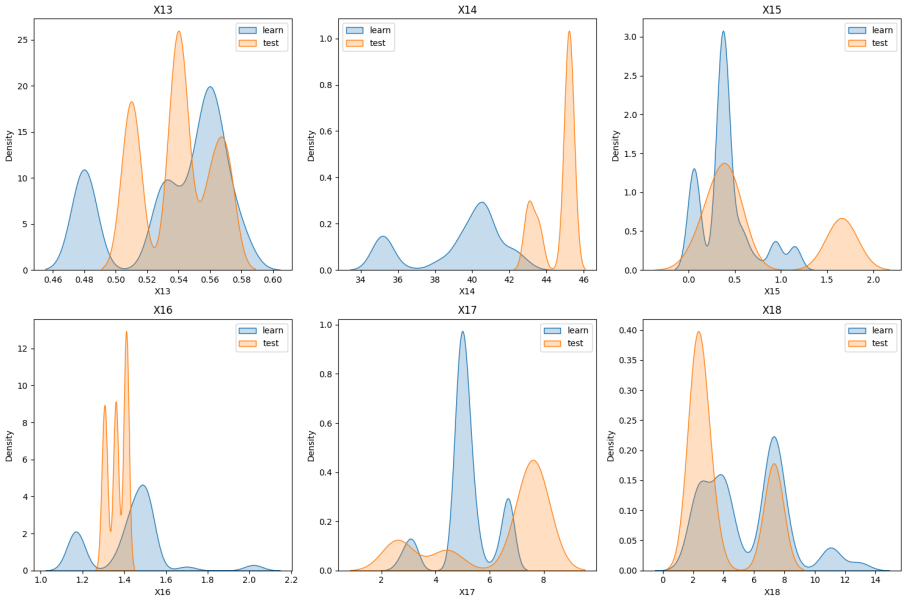


図. X13~X18 の訓練・テストデータの分布推移

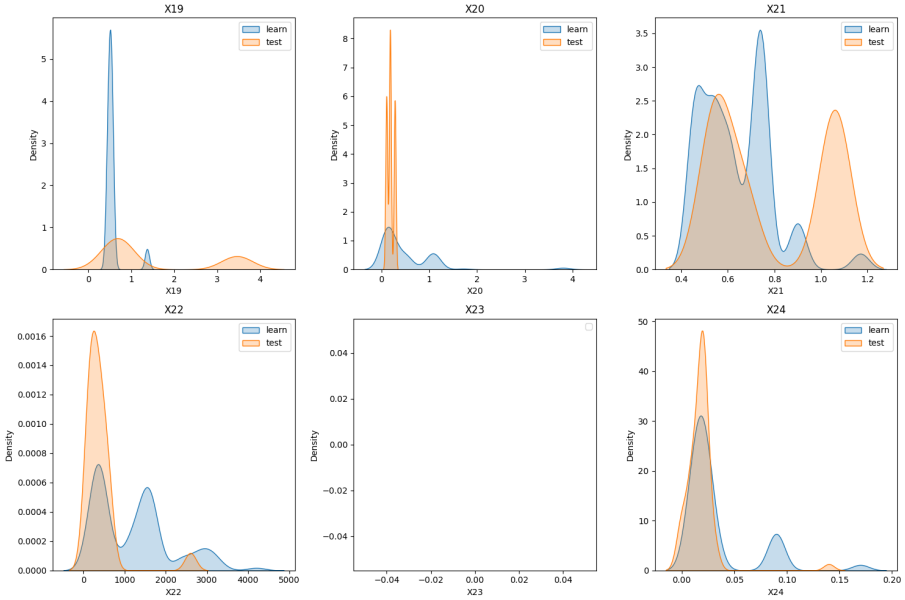


図. X19~X24 の訓練・テストデータの分布推移

※X23 は訓練・テストデータとも、全てであった

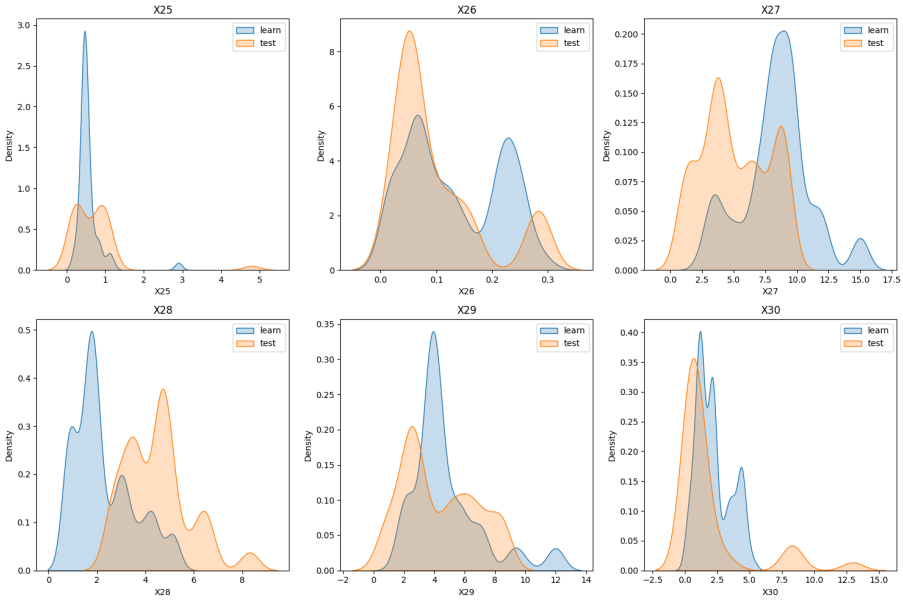


図. X25~X30 の訓練・テストデータの分布推移

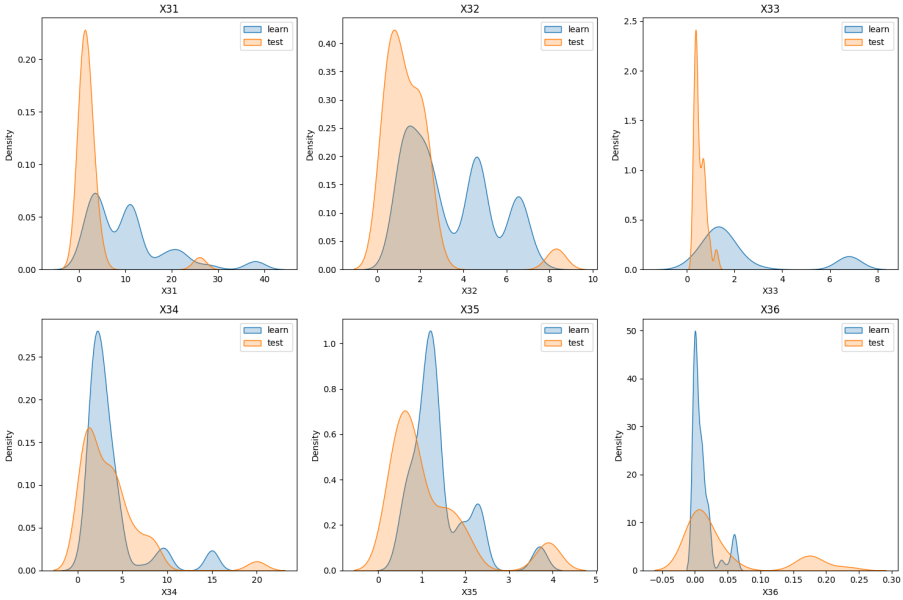


図. X31~X36 の訓練・テストデータの分布推移

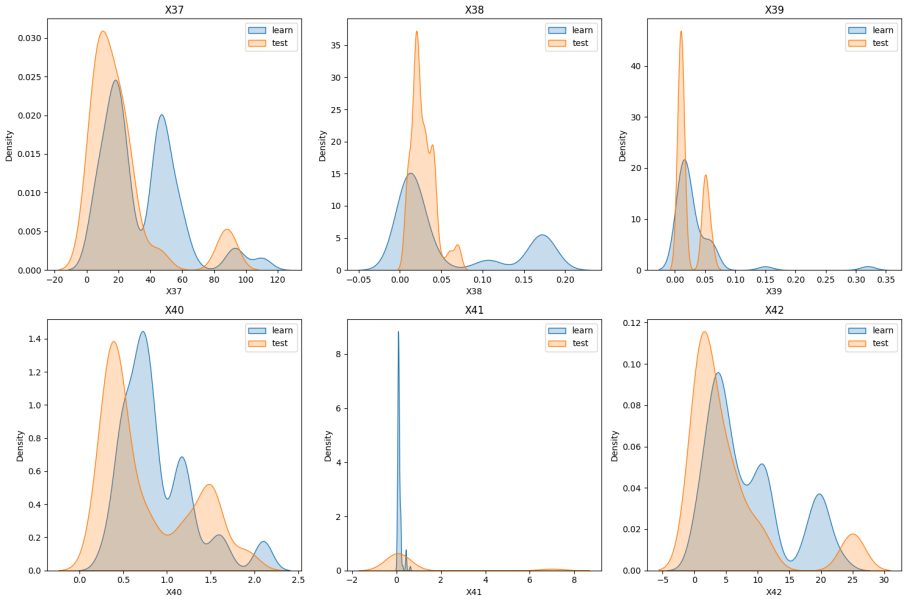


図. X37~X42 の訓練・テストデータの分布推移

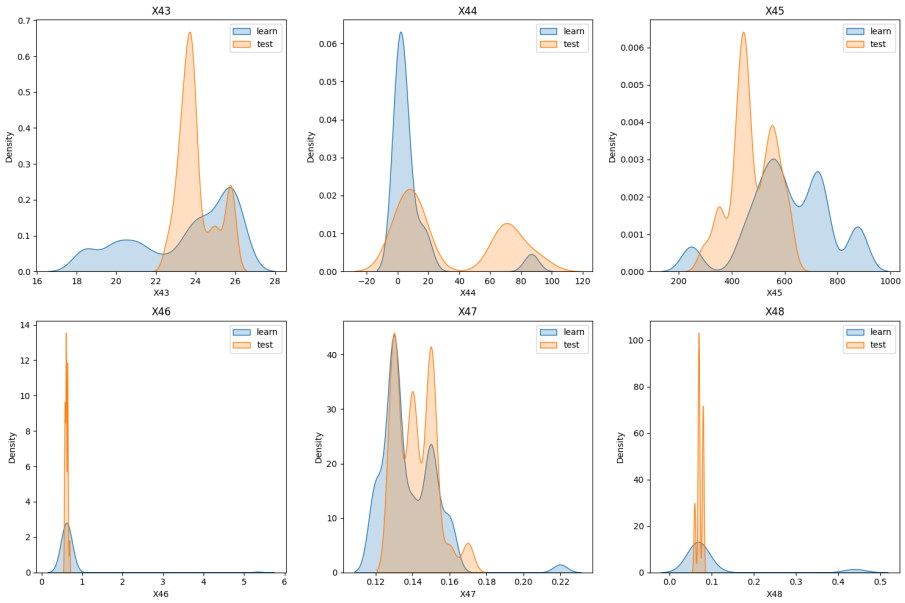


図. X43~X48 の訓練・テストデータの分布推移

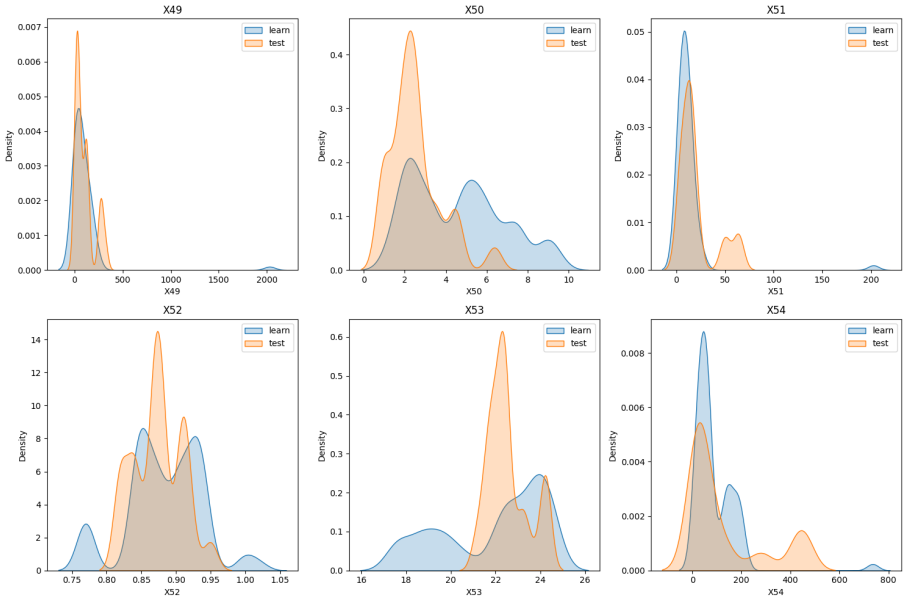


図. X49~X54 の訓練・テストデータの分布推移

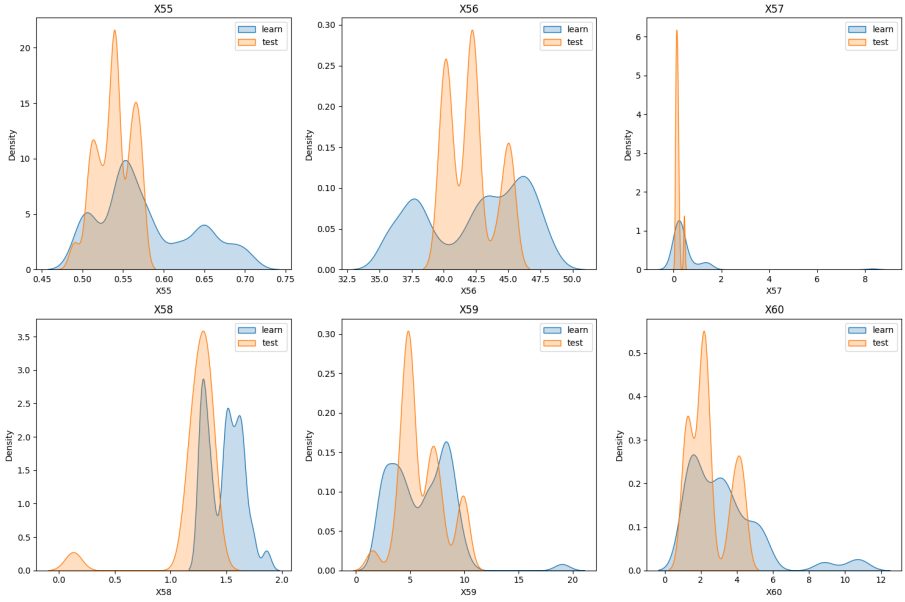


図. X55~X60 の訓練・テストデータの分布推移

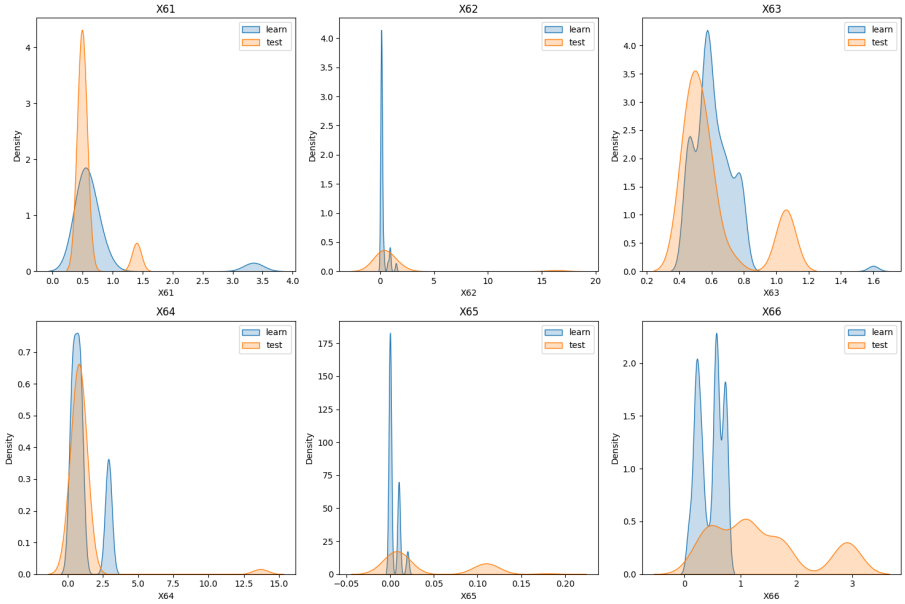


図. X61~X66 の訓練・テストデータの分布推移

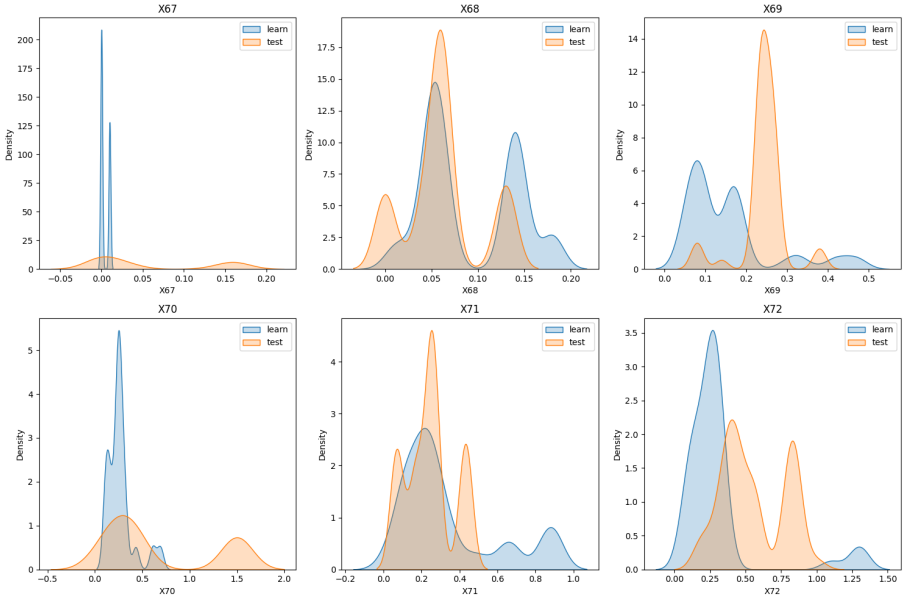


図. X67~X72 の訓練・テストデータの分布推移

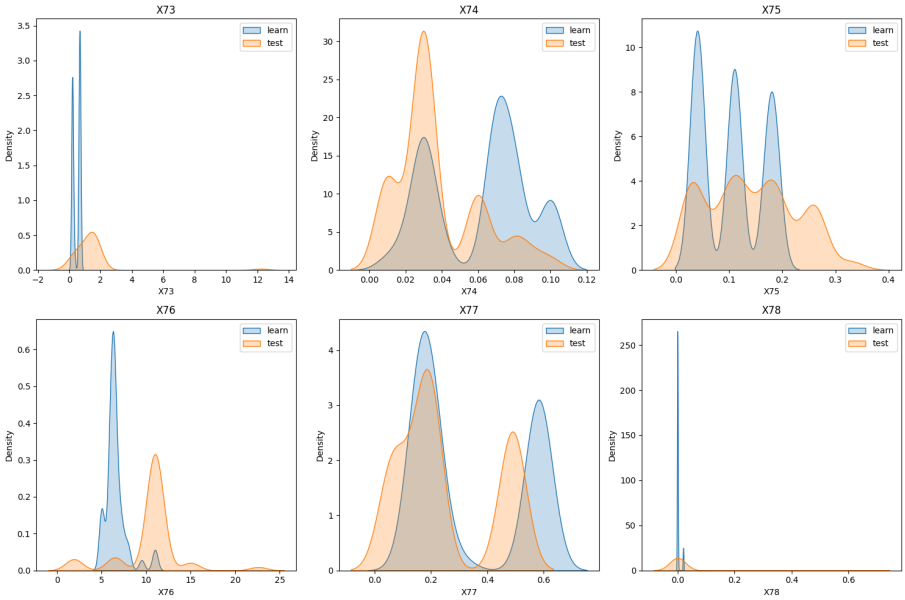


**図. X73~X78 の訓練・テストデータの分布推移**

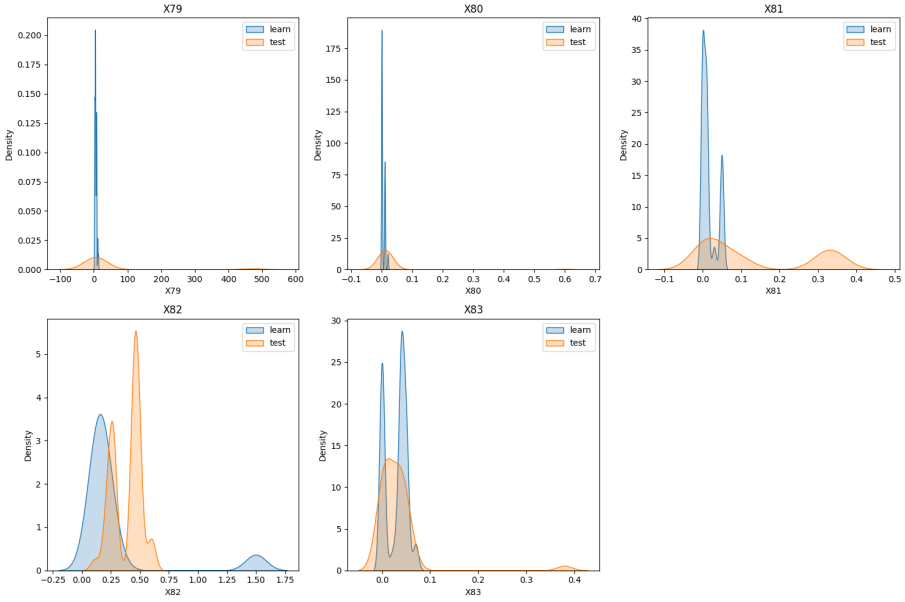


**図. X79~X83 の訓練・テストデータの分布推移**

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。

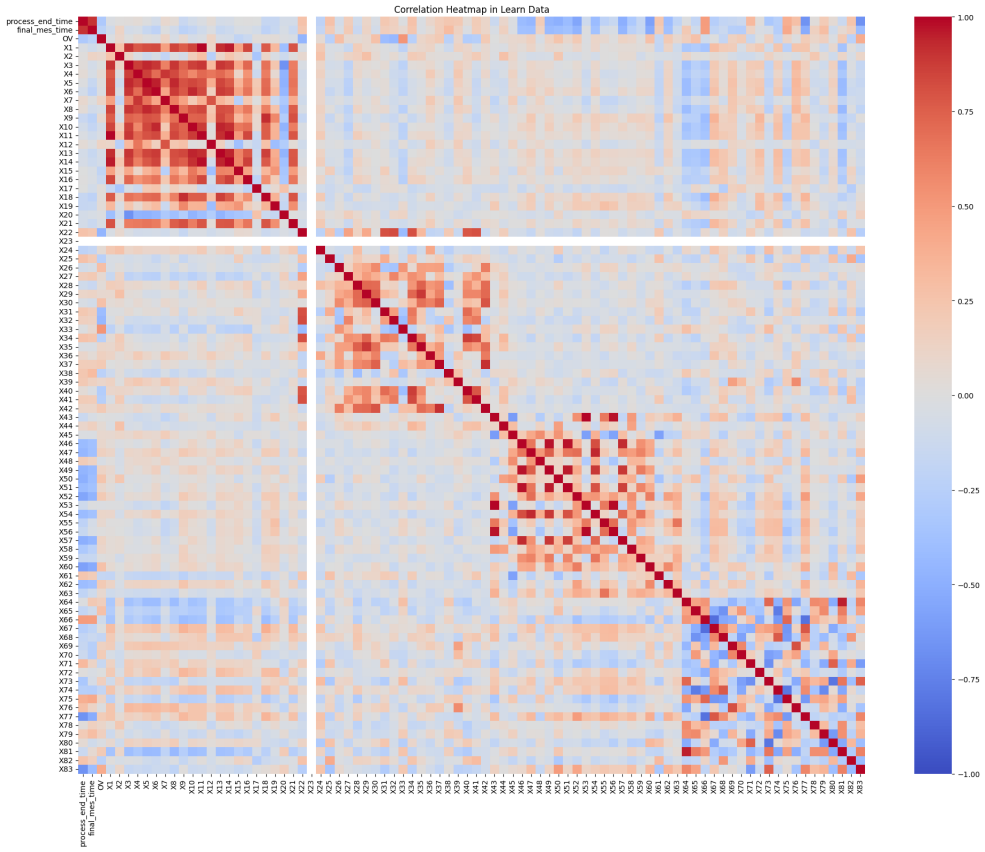


**図. 相関ヒートマップ**

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正の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）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。

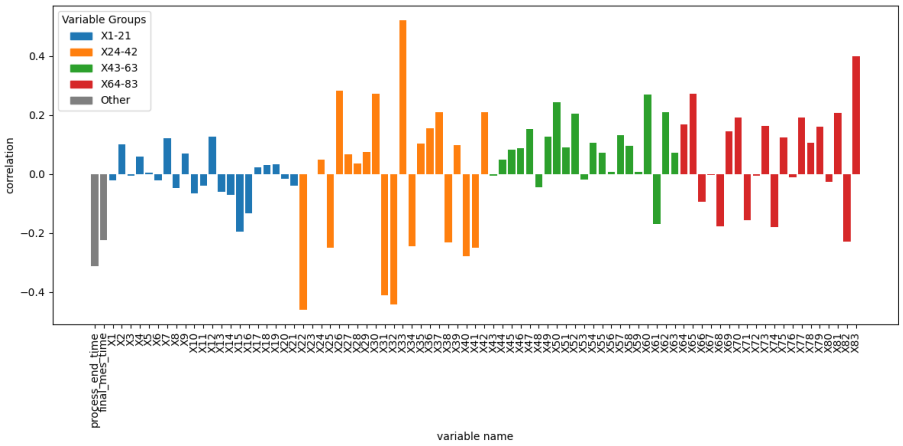


**図. OV と各説明変数間の相関推移**

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

## 7. 参考文献

## 8. 付録