

オイル温度予測モデルの構築

野村 龍王



目次

(1)背景: オイル温度予測の重要性や課題

(2)データの分析結果: EDAの結果と課題の抽出

(3)技術概要: 使用したモデルと特徴量エンジニアリングの手法

(4)評価指標: モデル評価に使用した指標とその結果

(5)検証内容: 改善策の背景にある仮説

(6)検証結果1: 仮説1の検証結果と考察

(7)検証結果2: 仮説2の検証結果と考察

(8)まとめ: 結果のまとめと今後の展望

(1)背景： オイル温度予測の重要性や課題

<重要性>

- ・ 機械の性能と寿命を最適化し、故障や損傷のリスクを軽減する。
- ・ エネルギー効率を向上させ、運転コストを削減する。

<課題>

- ・ 有用負荷と無用負荷の複雑な相互作用を適切にモデル化する必要性
- ・ 負荷レベルの変動に対する温度変化の時間遅れを正確に捉えること

(2) データ分析の結果：EDAの結果と課題の抽出

基本統計量

	HUFL	HULL	MUHL	MULL	LUFL	LULL	OT
平均	7.375	2.242	4.300	0.882	3.066	0.857	13.325
標準偏差	7.068	2.042	6.827	1.809	1.165	0.600	8.567
最小値	-22.706	-4.756	-25.088	-5.934	-1.188	-1.371	-4.080
25%	5.827	0.737	3.296	-0.284	2.315	0.670	6.964
50%	8.774	2.210	5.970	0.959	2.833	0.975	11.396
75%	11.788	3.684	8.635	2.203	3.625	1.218	18.079
最大値	23.644	10.114	17.341	7.747	8.498	3.046	46.007

<基本統計量から考えられる課題>

1. データの不均衡：HUFLとMUFLの標準偏差が大きい

これにより、モデルが特定の変数に過度に影響を受ける可能性が考えられる

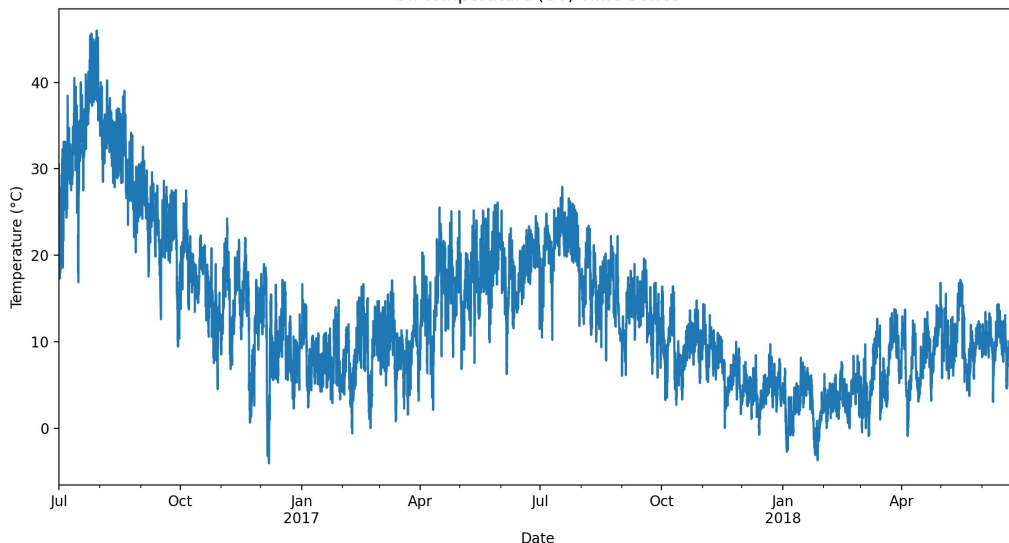
2. 外れ値の存在：最小値と最大値が四分位範囲から大きく離れている変数がある

これらの外れ値が予測精度に悪影響を与える可能性が考えられるため適切な前処理が必要

3. ラグ効果の考慮：負荷の変化がオイル温度に影響を与えるまでの時間差（ラグ）が存在する可能性がある

(2) データ分析の結果： EDAの結果と課題の抽出

Oil Temperature (OT) Time Series



左上図：油温の時系列グラフ

左下図：月別平均油温グラフ

<グラフから考えられる課題>

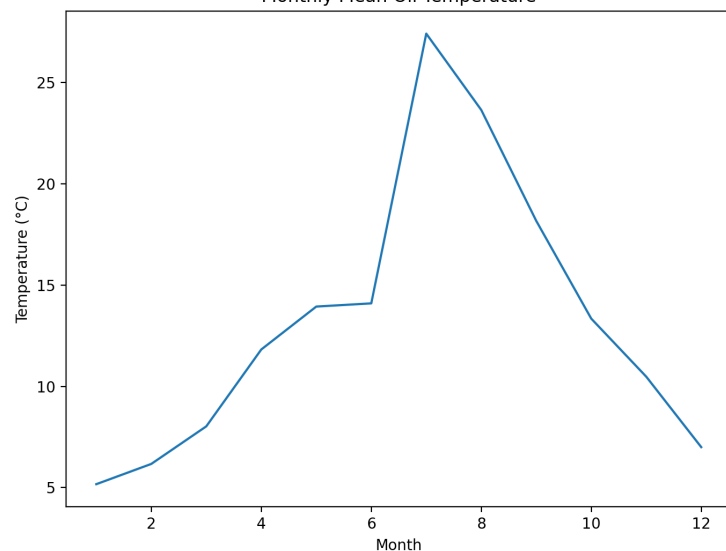
(1)強い季節性：月別平均オイル温度のグラフから、明確な季節パターンが見られる
夏季と冬季の温度の違いを予測モデルに組み込む必要がある

(2)短期的な変動： 日々のオイル温度に大きな変動が見られる

(3)非線形性： 月別平均温度グラフでは、温度上昇と下降のパターンが非対称

(4)異常値と外れ値： 時系列グラフには、通常のパターンから大きく外れた値がいくつか見られる
これらが真の異常値なのか、それとも測定エラーなのかを判断し、適切に処理する必要がある

Monthly Mean Oil Temperature



(3)技術概要： 使用したモデルと特徴量エンジニアリングの手法

使用したモデル

Linear Regression

理由

シンプルかつ解釈が容易で、変数間の直線的な関係を捉えるのに有効であるから。油温と各データの関係が線形である場合、非常に高速かつ低コストで予測を行うことが可能であるから。

Random Forest

理由

個々のロードデータが異なる影響を与える場合でも、その複雑さを扱うことができ、ノイズに対する強靱性を持っているから。

Gradient Boosting

理由

非線形性や複雑な相互作用をモデル化するのに適しており、油温に影響を与える負荷データとの複雑な関係を捉えることが可能であるから。

XGBoost

理由

高速な学習プロセスと優れた予測性能を持ち、欠損値の扱いにも強いから。

LightGBM

理由

大規模データや高次元データに対して高速に処理できるのが強みで、非線形な関係や相互作用を捉える能力が高く、時系列データでも効果的に予測が可能であるから。

ARIMA

理由

時系列データの季節性やトレンドを捉えることが可能であるから。

(3)技術概要: 使用したモデルと特徴量エンジニアリングの手法

1. 時間ベースの特徴量の追加

hour, day, month, year, weekday

```
df['hour'] = df.index.hour  
df['day'] = df.index.day  
df['month'] = df.index.month  
df['year'] = df.index.year  
df['weekday'] = df.index.dayofweek
```

2. ラグ特徴量の追加

HUFL_lag1, HUFL_lag24など

```
for col in ['HUFL', 'HULL', 'MUFL', 'MULL', 'LUFL', 'LULL', 'OT']:  
    df[f'{col}_lag1'] = df[col].shift(1)  
    df[f'{col}_lag24'] = df[col].shift(24)
```

これらの特徴量を追加した理由

時間ベースの特徴量の追加: 油温は日周期や月周期に依存する可能性があるため、これらの特徴量が重要であると考えたから

ラグ特徴量の追加: 油温の予測には過去の値が大きな影響を与えるため、過去の値を特徴量として追加することで、時系列の依存関係を捉えることが可能になると考えたから

(4) 評価指標：モデル評価に使用した指標とその結果

使用した指標

R-squared（決定係数）

理由：モデルがどれだけの情報を説明できるかを一目で理解できるため、回帰モデルの性能を直感的に評価するのに便利であるから

Mean Squared Error (MSE)

理由：大きな誤差を強調し、モデルが予測に対してどれだけの精度を持っているかを示す。
二乗和を用いるため、外れ値に敏感

Root Mean Squared Error (RMSE)

理由：MSEよりも直感的で解釈しやすい指標であり、予測の誤差の平均的な大きさを示すから

Mean Absolute Error (MAE)

理由：外れ値の影響を最小限に抑えた評価ができるため、誤差の平均的な大きさを示す指標となるから

(4) 評価指標：モデル評価に使用した指標とその結果

訓練データ

	Linear Regression	Random Forest	Gradient Boosting	XGBoost	LightGBM	ARIMA
R-squared	0.98661	0.99802	0.98925	0.99746	0.99302	0.98574
MSE	0.014284	0.0021022	0.011553	0.0027704	0.0075946	0.015187
RMSE	0.11929	0.045767	0.10745	0.050657	0.08672	0.12301
MAE	0.081975	0.030819	0.074085	0.036702	0.060616	0.084179

テストデータ

	Linear Regression	Random Forest	Gradient Boosting	XGBoost	LightGBM	ARIMA
R-squared	0.95044	0.91041	0.81925	0.78932	0.84529	-1.2232
MSE	0.012011	0.21422	0.04557	0.051523	0.038409	0.44355
RMSE	0.10507	0.13562	0.16758	0.18869	0.16046	0.65868
MAE	0.077044	0.097421	0.12373	0.13718	0.11396	0.55957

<結果から見えること>

- ・テストデータの各モデルを比較すると、Linear Regressionはかなり良いスコアであることから、線形回帰モデルが適している可能性が考えられる。

- ・それ以外のモデルでは訓練データとテストデータのスコアにかかりさがあることから、過学習している可能性が考えられる。

- ・また、ARIMAについてはテストデータのR-squaredの結果が負の値であることから、ARIMAは多変量モデルには適していない可能性がある。

(5)検証内容：改善策の背景にある仮説

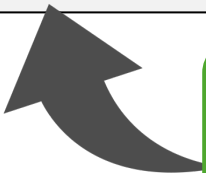
<結果から考えられる改善策の仮説1>

特徴量の数が多いことで過学習を起こしている可能性が考えられる
では、特徴量の中でも重要度の高いものを厳選して訓練を行うことで過学習を防げて、かつ
より精度を高めることができるのではと考察

追加

3. 特徴量選択

SelectKBestとRandom Forestを使用して、最も予測性能に寄与する特徴量を選定
初期段階で SelectKBestを用いて、30個の特徴量を選択し、さらにRandom Forestを使って特徴量
の重要度を評価し、最終的に上位5つの特徴量を使用



重要なものだけを選択することで、より精度
の高い予測が可能になり、かつ計算コストを
削減できるから

(5)検証内容：改善策の背景にある仮説

<結果から考えられる改善策の仮説2>

ARIMAは多変量のデータの予測に適していないに可能性が考えられる
別のモデルの方に変更したほうが良いのではと考察
ARIMAを改良したSARIMAXに変更

変更

先ほどの結果から、Linear Regression(線形回帰)が良いスコアであることから、今回のデータは**線形**の関係性である**可能性が高い**ことや、**季節性の考慮も優れている**ことからSARIMAXを選定

ARIMA

理由
時系列データの季節性やトレンドを捉えることが可能であるから。

SARIMAX

理由
ARIMAに比べ、季節性やトレンドを捉えることに優れており、他の外部要因が油温(OT)に与える影響をモデルに組み込むことが可能であるから。

(6) 検証結果1: 仮説1の検証結果と考察

訓練データ

	Linear Regression	Random Forest	Gradient Boosting	XGBoost	LightGBM	ARIMA
R-squared	0.98603	0.9978	0.98779	0.99506	0.99023	0.98574
MSE	0.014899	0.00234	0.013078	0.0053776	0.01053	0.015187
RMSE	0.12185	0.048285	0.11429	0.072052	0.10253	0.12301
MAE	0.083842	0.033169	0.079702	0.050881	0.071692	0.084179

テストデータ

	Linear Regression	Random Forest	Gradient Boosting	XGBoost	LightGBM	ARIMA
R-squared	0.95731	0.92312	0.92633	0.88086	0.89357	-1.2217
MSE	0.010017	0.018195	0.017646	0.028816	0.026158	0.4434
RMSE	0.097881	0.12733	0.12287	0.14902	0.14072	0.65857
MAE	0.068241	0.09064	0.084775	0.10276	0.096003	0.55939

<考察>

- ・依然としてLinear Regressionが一番良い結果を出していることからやはり線形の関係性が高い可能性がある。

- ・それ以外のモデルのスコアは、かなり良くなったが、まだ訓練データのスコアと比較すると差が見られることから未だ過学習が起きている可能性がある。

- ・ARIMAのスコアはほとんど変わっていないことから、別の時系列予測モデルに変更することが望ましい可能性がある。

(7) 検証結果2: 仮説2の検証結果と考察

訓練データ

	Linear Regression	Random Forest	Gradient Boosting	XGBoost	LightGBM	SARIMAX
R-squared	0.98603	0.9978	0.98779	0.99506	0.99023	0.982379
MSE	0.014899	0.00234	0.013078	0.0053776	0.01053	0.017621
RMSE	0.12185	0.048285	0.11429	0.072052	0.10253	0.131654
MAE	0.083842	0.033169	0.079702	0.050881	0.071692	0.093106

テストデータ

	Linear Regression	Random Forest	Gradient Boosting	XGBoost	LightGBM	SARIMAX
R-squared	0.95731	0.92312	0.92633	0.88086	0.89357	0.982379
MSE	0.010017	0.018195	0.017646	0.028816	0.026158	0.022776
RMSE	0.097881	0.12733	0.12287	0.14902	0.14072	0.147930
MAE	0.068241	0.09064	0.084775	0.10276	0.096003	0.119335

<考察>

- ・テストデータのR-squaredのスコアはLinear Regressionと比べてSARIMAXの方が高いことから、モデルの説明力が優れていることが分かる

- ・MSE、RMSE、MAEはLinear Regressionの方が低いことから、予測の精度はLinear Regressionの方が高いことがわかる

- ・データの説明力を重視するならSARIMAX、予測の精度を重視するならLinear Regressionが適していると考えられる

(8)まとめ:結果のまとめと今後の展望

<結果>

- ・時間ベースの特徴量の追加とラグ特徴量を追加し、これらを重要度で厳選したもので学習を行うことでより精度の高い予測を行うことが可能になることがわかった
- ・線形の関係性が見られることや、外部要因が予測に影響を与えることからSARIMAXは予測における説明力が高い傾向が見られたが、精度の面では線形回帰であるLinear Regressionが高くなることがわかった

<今後の展望>

【ハイブリッドモデルの開発】

SARIMAXの時系列特性の捕捉能力と、Linear Regressionの予測精度の高さを組み合わせたハイブリッドモデルの開発を行うことで、より精度の高い予測モデルを作成できると考えられる。

【特徴量エンジニアリングの改善】

ドメイン知識を活用して、予測に有用な外部データ（天候データ、経済指標など）を導入することで、より精度をあげられる可能性がある