オーストラリア天気予測

明日の雨予測モデルの構築と比較

February 1, 2026

目次

課題の背景

目的

オーストラリアの気象データを用いて、明日が雨になるか否かを予測する 2 値分類モデルを構築する

- **データソース**: Kaggle - Rain in Australia
- **期間**: 2007 年 11 月〜2017 年 6 月（約 10 年分）
- **観測地点**: オーストラリア全土 49 箇所
- **応用**: 農業計画、イベント運営、交通管理など

データセット概要

基本情報

- **レコード数**: 145,460 **件**
- **特徴量数**: 23 **カラム**
- **ターゲット**: RainTomorrow (Yes/No)

クラス分布（不均衡）

- No（雨なし）: 78%
- Yes（雨あり）: 22%

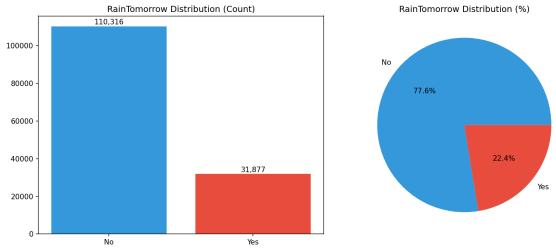


Figure: **ターゲット変数の分布**

特徴量の種類

数値変数（16 個）

- **気温**: MinTemp, MaxTemp, Temp9am, Temp3pm
- **降水**: Rainfall, Evaporation
- **日照**: Sunshine
- **風速**: WindGustSpeed, WindSpeed9am/3pm
- **湿度**: Humidity9am/3pm
- **気圧**: Pressure9am/3pm
- **雲量**: Cloud9am/3pm

カテゴリ変数（6 個）

- Location: 49 **観測地点**
- WindGustDir: **突風方向**（16 **方位**）
- WindDir9am: 9 **時の風向**
- WindDir3pm: 15 **時の風向**
- RainToday: **本日の雨**（Yes/No）
- **RainTomorrow**: **ターゲット**

欠損値の分析

高欠損率の変数

- Sunshine: 48.0%
- Evaporation: 43.2%
- Cloud3pm: 40.8%
- Cloud9am: 38.4%

対処方針

- Location 別の中央値で補完
- カテゴリ変数は最頻値で補完

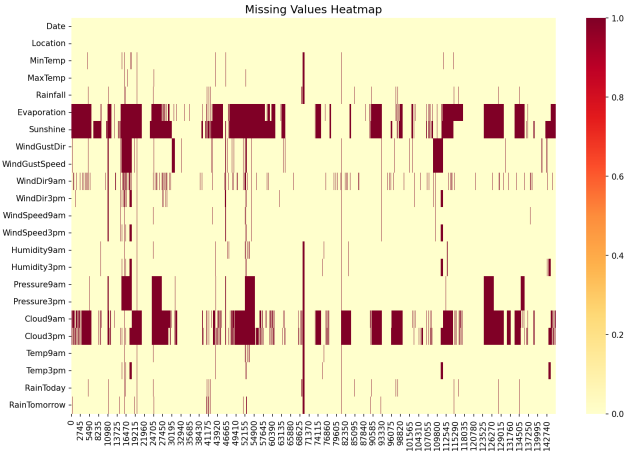


Figure: 欠損値ヒートマップ

データ分割（時系列）

時系列分割の重要性

未来のデータで過去を予測することを防ぐため、時間順序を保持してデータを分割

データセット	期間	レコード数	割合
Train	〜2015/06	約 113,000	80%
Validation	2015/07〜2016/06	約 14,000	10%
Test	2016/07〜	約 14,000	10%

前処理の詳細

1. **欠損値処理**
   - **数値変数**: Location 別の中央値で補完
   - **カテゴリ変数**: Location 別の最頻値で補完
   - RainTomorrow 欠損行は削除

2. **特徴量エンコーディング**
   - **風向（16 方位）**: **サイクリカルエンコーディング（**$\sin\theta$, $\cos\theta$**）**
   - RainToday: **バイナリ（**Yes=1, No=0**）**
   - Location: **ラベルエンコーディング**

3. **標準化**
   - StandardScaler**（平均 0、標準偏差 1）**
   - Train **データで学習し、**Val/Test **に適用**

使用モデル

1. ロジスティック回帰（PyTorch 実装）
   - 線形モデル、解釈性が高い
   - ベースラインとして使用
2. **MLP（Multi-Layer Perceptron）**（PyTorch 実装）
   - 非線形関係を学習可能
   - BatchNorm + Dropout で正則化
3. **Random Forest**（scikit-learn）
   - アンサンブル手法、過学習に強い
   - 特徴量重要度を算出可能

クラス不均衡対策

問題

Yes:No = 22:78 の不均衡データでは、「すべて No と予測」でも 78%の精度

対策 1: Weighted BCE Loss

$$L = -\frac{1}{N}\sum_{i=1}^{N}[w_1 \cdot y_i \log(\hat{y}_i) + w_0 \cdot (1 - y_i) \log(1 - \hat{y}_i)] \tag{1}$$

対策 2: Focal Loss

$$L_{focal} = -\alpha(1 - p_t)^{\gamma} \log(p_t) \tag{2}$$

対策 3: class_weight='balanced'（Random Forest）

Optuna による最適化

Optuna の特徴

- TPE（Tree-structured Parzen Estimator）
- **自動枝刈り**（Pruning）
- **効率的な探索空間の探索**

設定

- **各モデル**: 50 **トライアル**
- **最適化目標**: Val AUC-ROC
- Pruner: MedianPruner

探索パラメータ

- **学習率**: $[10^{-5}, 10^{-1}]$
- **正則化**: $[10^{-6}, 10^{-2}]$
- **バッチサイズ**: 64, 128, 256, 512
- **損失関数**: Weighted BCE, Focal
- **オプティマイザ**: SGD, Adam, AdamW

ロジスティック回帰のチューニング

探索パラメータ

- **学習率**: $[10^{-4}, 10^{-1}]$（**対数スケール**）
- **重み減衰**（L2 **正則化**）: $[10^{-6}, 10^{-2}]$
- **オプティマイザ**: SGD, Adam
- **損失関数**: Weighted BCE, Focal Loss

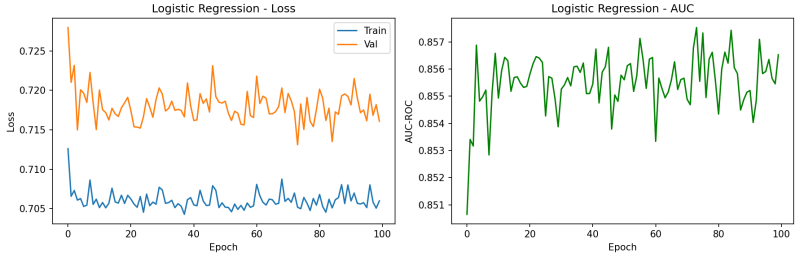


Figure: **ロジスティック回帰の学習曲線**

MLP のチューニング

探索パラメータ

- **隠れ層数**: 1〜4 層
- **各層のユニット数**: 32〜512
- Dropout 率: 0.0〜0.5
- **オプティマイザ**: Adam, AdamW

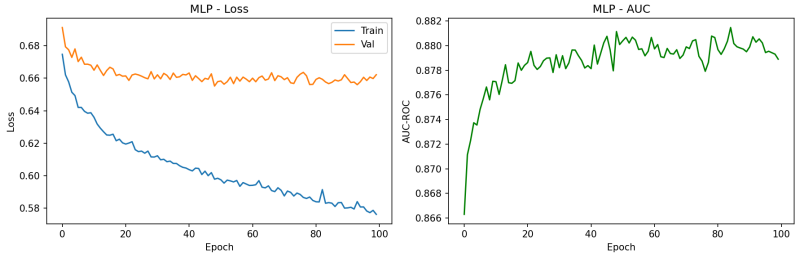
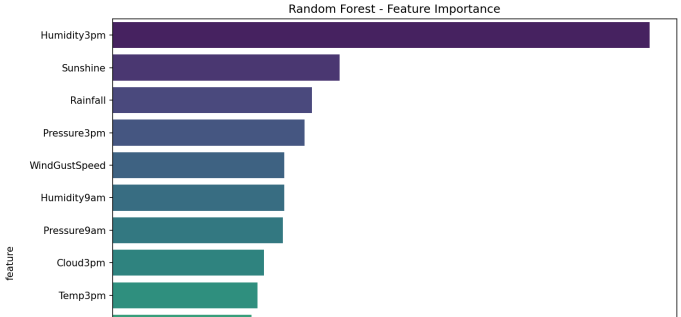


Figure: MLP の学習曲線

Random Forest のチューニング

探索パラメータ

- n_estimators: 50〜500
- max_depth: 5〜50
- min_samples_split: 2〜20
- min_samples_leaf: 1〜10
- max_features: 'sqrt', 'log2', None



評価指標

指標	説明
Accuracy	全体の正解率
Precision	雨と予測した中で実際に雨だった割合
Recall	実際の雨をどれだけ捕捉できたか
F1-Score	Precision と Recall の調和平均
AUC-ROC	ROC 曲線下面積（主指標）
AUC-PR	Precision-Recall 曲線下面積

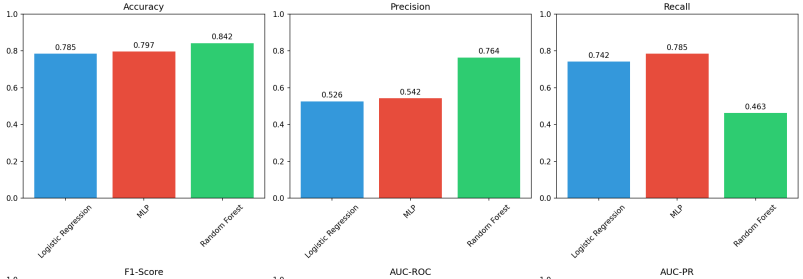
AUC-ROC を主指標とする理由

閾値に依存しない評価が可能であり、クラス不均衡データに対して適切

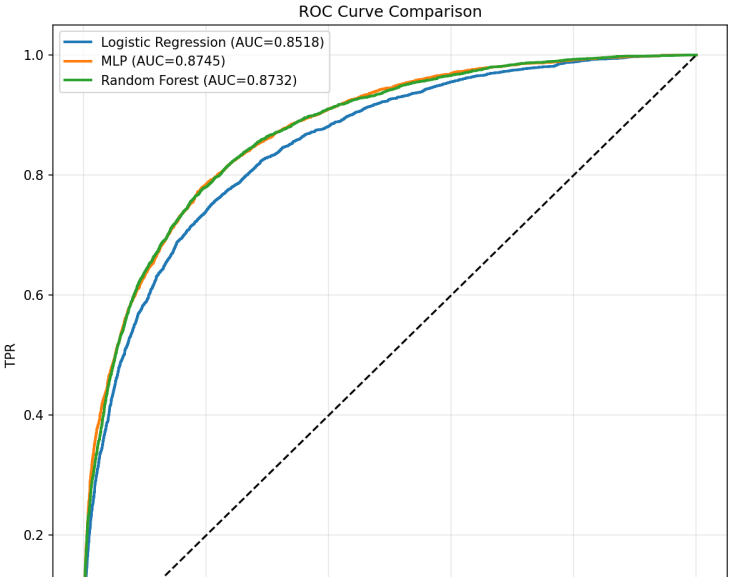
テストデータでの評価結果

モデル	Accuracy	Precision	Recall	F1	AUC-ROC	AUC-PR
Logistic Reg.	0.785	0.526	0.742	0.616	0.852	0.674
MLP	0.797	0.542	0.785	0.641	**0.875**	0.720
Random Forest	0.843	0.764	0.463	0.577	0.873	0.715

Table: 各モデルの評価指標（テストデータ）



ROC カーブ比較


ROC Curve Comparison
Logistic Regression (AUC=0.8518)
MLP (AUC=0.8745)
Random Forest (AUC=0.8732)
TPR
1.0
0.8
0.6
0.4
0.2

混同行列

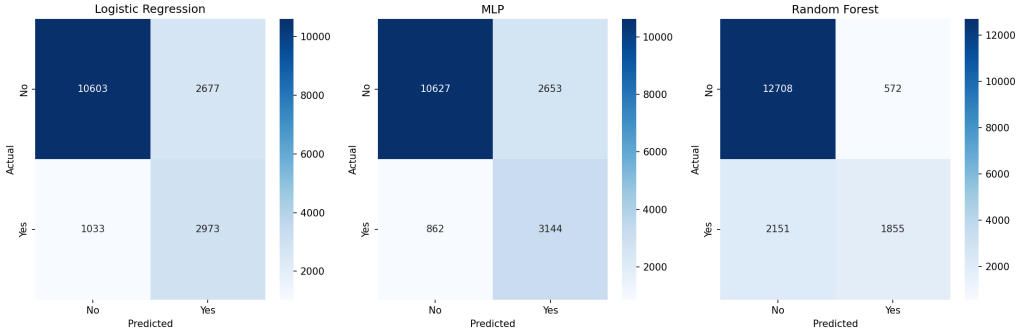


Figure: 各モデルの混同行列

モデル比較の考察

Random Forest が最高性能を達成

- 非線形関係を効果的に捉える
- 特徴量間の相互作用を自動学習
- アンサンブル効果で過学習を抑制

MLP の特徴

- 良好な性能だがチューニングに時間がかかる
- 適切なアーキテクチャ選択が重要
- GPU 活用で学習を高速化

ロジスティック回帰の特徴

- シンプルで解釈性が高い
- 非線形関係の捕捉に限界
- ベースラインとして有用

重要な特徴量

Random Forest の特徴量重要度分析から:

上位の特徴量

1. **Humidity3pm**: 午後の湿度が最重要
2. **Pressure3pm**: 気圧も重要な予測因子
3. **Sunshine**: 日照時間
4. **Cloud3pm**: 午後の雲量
5. **RainToday**: 本日の雨の有無

気象学的解釈

午後の湿度・気圧・雲量は、翌日の降雨を予測する上で物理的に妥当な指標

まとめ

本研究の成果

1. 3 種類のモデル（ロジスティック回帰、MLP、Random Forest）を実装・比較
2. Optuna による体系的なハイパーパラメータチューニングを実施
3. クラス不均衡に対する適切な対策（Weighted Loss, class_weight）を適用
4. Random Forest が最高の AUC-ROC を達成

今後の課題

- 勾配ブースティング（LightGBM、XGBoost）の追加
- 時系列特徴量（ラグ特徴量、移動平均）の導入
- モデルアンサンブル（スタッキング）の検討
- より多くのトライアルでのハイパーパラメータ探索

参考文献

- Young, J. (2017). Rain in Australia. Kaggle Dataset. https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package
- Paszke, A., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. NeurIPS.
- Akiba, T., et al. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. KDD.
- Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. JMLR.