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

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

なぜこのテーマを選んだか

- 金沢市は年間降水量が多く、雨の日を事前に予測できると日常生活で役立つ
- 気象データを用いた機械学習予測に興味があった
- 2 値分類の実践的な題材として適切だった

日常生活・研究への応用

- 外出・イベントの計画立案
- 農業における作業スケジュール管理
- 交通機関の運行計画への活用

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)
- 期間: 2007 年 11 月〜2017 年 6 月
- 観測地点: オーストラリア全土 49 箇所

クラス分布（不均衡）

- No（雨なし）: 78%
- Yes（雨あり）: 22%

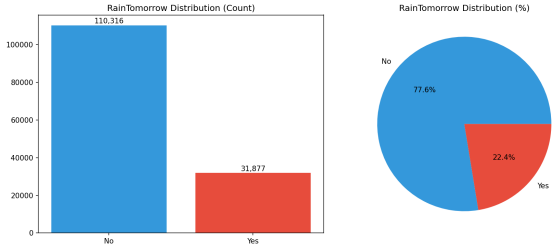


Figure: ターゲット変数の分布

出典: Young, J. (2017). Rain in Australia. Kaggle Dataset.
`https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package`

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

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）
   - 構造: 入力層 → 出力層（1 ユニット）のシンプルな線形モデル
   - 特長: 解釈性が高く、各特徴量の重みから予測への寄与度がわかる
   - 用途: ベースラインモデルとして使用

2. **MLP（Multi-Layer Perceptron）**（PyTorch 実装）
   - 構造: 入力層 → 隠れ層（複数）→ 出力層の多層ニューラルネット
   - 特長: 非線形な関係を学習可能、BatchNorm と Dropout で過学習を抑制
   - 用途: より複雑なパターンの学習

3. **Random Forest**（scikit-learn）
   - 構造: 複数の決定木を並列に学習し、多数決で予測するアンサンブル手法
   - 特長: 過学習に強く、特徴量重要度を算出可能
   - 用途: 非線形関係の効果的な捕捉

クラス不均衡対策

問題点

Yes:No = 22:78 の不均衡データでは、「すべて No と予測」しても 78%の精度になってしまう

実施した対策

1. **ロジスティック回帰・MLP**（PyTorch モデル）
   - 損失関数に重み付けを導入（Weighted BCE Loss）
   - 少数クラス（雨あり）の誤分類に対するペナルティを大きく設定
   - 重み比率: No:Yes = 1:3.5（クラス比率の逆数）
2. **Random Forest**（scikit-learn）
   - class_weight='balanced' パラメータを使用
   - 各クラスのサンプル数に応じて自動的に重みを調整

これらの対策により、少数クラス（雨あり）の検出率を向上させた

ロジスティック回帰のチューニング

探索したハイパーパラメータ

- 学習率: モデルの重み更新の大きさを制御するパラメータ
- 重み減衰（L2 正則化）: 過学習を防ぐための正則化強度
- オプティマイザ: 最適化アルゴリズム（SGD, Adam）
- 損失関数: Weighted BCE Loss

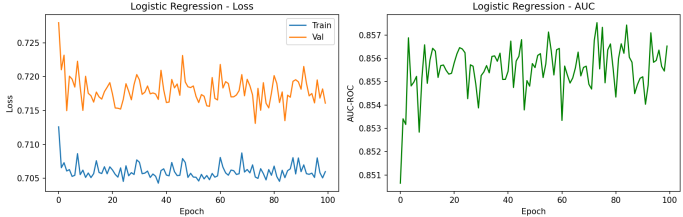


Figure: ロジスティック回帰の学習曲線（エポックごとの損失と Accuracy の推移）

MLP のチューニング

探索したハイパーパラメータ

- 隠れ層数・ユニット数: ネットワークの深さと幅
- Dropout 率: 過学習防止のためのドロップアウト確率
- 学習率・重み減衰: ロジスティック回帰と同様
- オプティマイザ: Adam, AdamW

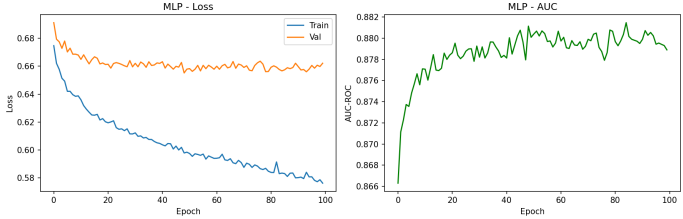


Figure: MLP の学習曲線（エポックごとの損失と Accuracy の推移）

Random Forest のチューニング

探索したハイパーパラメータ

- n_estimators: 決定木の数（多いほど安定するが計算コスト増）
- max_depth: 各決定木の最大深さ（深いほど複雑なパターンを学習）
- min_samples_split: ノード分割に必要な最小サンプル数
- min_samples_leaf: 葉ノードに必要な最小サンプル数
- max_features: 各分割で考慮する特徴量の数

Random Forest は決定木のアンサンブルであり、学習曲線ではなく上記パラメータの組み合わせを探索して最適なモデルを選定した

評価指標

指標	説明
Accuracy	全体の正解率。全予測のうち正しく分類できた割合
Precision	「雨」と予測したもののうち、実際に雨だった割合。誤警報を減らしたい場合に重視
Recall	実際の雨の日のうち、正しく「雨」と予測できた割合。見逃しを減らしたい場合に重視
F1-Score	Precision と Recall の調和平均。両者のバランスを評価

不均衡データでの注意点

Accuracy だけでは正しく評価できない。Precision, Recall, F1-Score を総合的に見ることが重要

テストデータでの評価結果

モデル	Accuracy	Precision	Recall	F1-Score
ロジスティック回帰	0.785	0.526	0.742	0.616
MLP	0.797	0.542	0.785	0.641
Random Forest	**0.843**	**0.764**	0.463	0.577

Table: 各モデルの評価指標（テストデータ）

結果の概要

- Random Forest が Accuracy と Precision で最高
- MLP が Recall と F1-Score で最高
- ロジスティック回帰はベースラインとして妥当な性能

混同行列

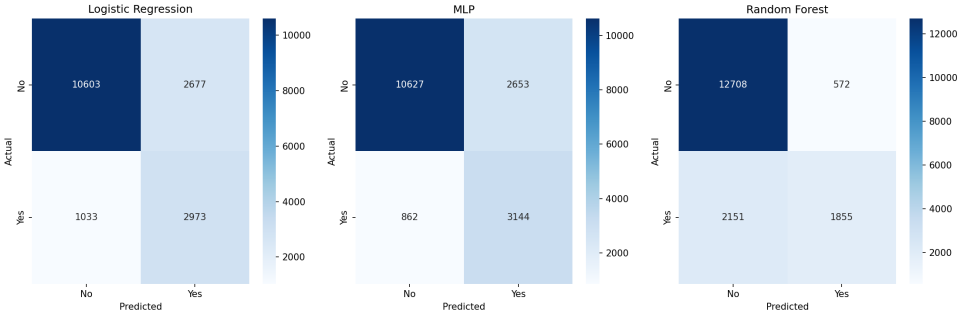


Figure: 各モデルの混同行列

モデル比較の考察

Random Forest の高い Accuracy・Precision の理由

- 決定木のアンサンブルにより、特徴量間の複雑な相互作用を捉えられた
- 気象データは非線形な関係が多く、木構造との相性が良かった
- ただし Recall が低く、雨の日の見逃しが多い

MLP の高い Recall・F1-Score の理由

- 重み付き損失関数により、少数クラス（雨あり）を積極的に検出
- 多層構造により非線形パターンを学習できた
- Precision は低めで、誤警報が多い傾向

ロジスティック回帰の限界

- 線形モデルのため、非線形な関係の捕捉に限界がある
- しかしベースラインとして、他モデルの改善度を測る基準になった

重要な特徴量

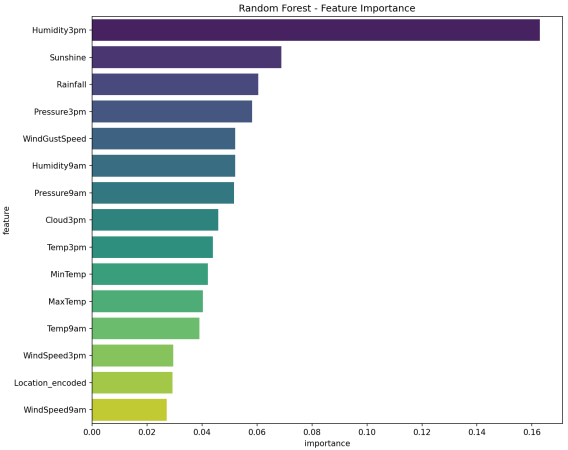
Random Forest の特徴量重要度分析から:

上位の特徴量

1. **Humidity3pm**: 午後の湿度
2. **Pressure3pm**: 午後の気圧
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. 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（重み付き損失関数）を適用
4. Random Forest が Accuracy 0.843、Precision 0.764 で最高性能を達成

今後の課題

- 勾配ブースティング（LightGBM、XGBoost）の追加検討
- 時系列特徴量（ラグ特徴量、移動平均）の導入
- Recall を重視したモデル調整（雨の見逃しを減らす）