# ProbSpace 花粉飛散量予測 3位解法

2023年2月3日 maruyama

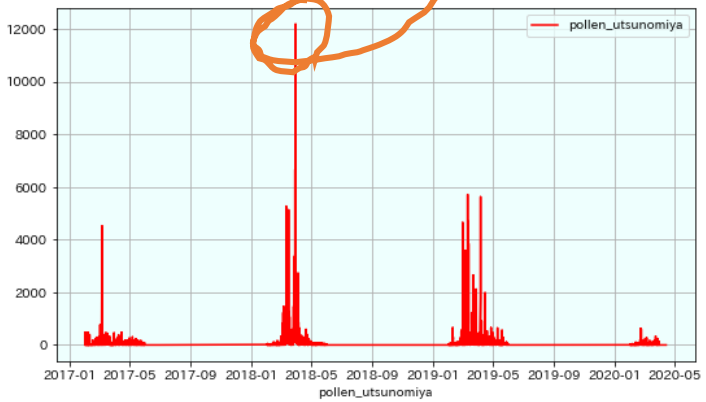
# 目次

- 分析方針を立てるまで
  - トピックを読んで要点を把握する
  - 要点から仮説を立てる
  - 仮説をもとに分析方針を立てる
  - ちゃんとCVしよう…… (反省)
- 解法
  - 解法の概要
  - 解法の詳細｜特徴抽出
  - 解法の詳細｜学習
  - 解法の詳細｜アンサンブル
  - 解法の詳細｜後処理

# トピックを読んで要点を把握する

参加時点でコンペ開始から3か月が経過しており
議論が進んでいたため、トピックを参考に分析方針を立てていった。

- 花粉飛散量がバーストする時刻がある。
  - “EDA” (@kotrying)
- 乱数シードを変えるだけで大幅に予測精度が変わる。
  - “幸運なseed値？” (@kotrying)
- 2020年は例年より花粉飛散量が少ない。
  - “Model (LightGBM Base)” (@kotrying)
  - “Targetの補正に関して” (@uchs)




宇都宮の花粉飛散量の推移 (kotrying氏のEDAから引用)

# 要点から仮説を立てる

- 花粉飛散量がバーストする時刻がある。
  → バースト時刻がMAEを支配しているはず。
- 乱数シードを変えるだけで大幅に予測精度が変わる。
  - “乱数シードを変えるだけで”
    → 情報が不足している。
  - “大幅に”
    → バースト時刻の予測を当てたり外している。
  - “予測精度が変わる”
    → バーストする時刻は当てられているが、花粉飛散量を当てたり外したりしている。
- 2020年は例年より花粉飛散量が少ない。
  →（素直に「少ないんだなあ」と受け止める）

# 仮説をもとに分析方針を立てる

- バースト時刻の予測がMAEを支配しているはず。
  → バーストの予測に注力する。
- バーストする時刻は当てられそうだが、
  その時刻の花粉飛散量は情報不足により予測できなさそう。
  → バーストする時刻を当てるモデルの開発に専念し、
  花粉飛散量の予測自体はPublic LBを参考に手動で調整する。
- 2020年は例年より花粉飛散量が少ない。
  → Public LBを参考に予測結果を低めに手動補正する。

# ちゃんとCVしよう……(反省)

Public LBだけを当てに行き過ぎてPrivate LBで撃沈……

Public LB

| 順位 | チーム名 | メンバー | 最新 | 提出回数 | ベストスコア |
|---|---|---|---|---|---|
| 1 | - | maruyama | 21日前 | 17 | 10.08458 |
| 2 | - | cczouk | 22日前 | 25 | 10.68955 |
| 3 | - | mobi_morita | 21日前 | 64 | 10.80668 |
| 4 | PBDSC | panpanpanda<br>T.T | 21日前 | 94 | 11.40796 |
| 5 | AI-FOX | Yosemite<br>MOK<br>te2T<br>moto1963<br>saru_da_mon | 21日前 | 23 | 11.66667 |

Private LB

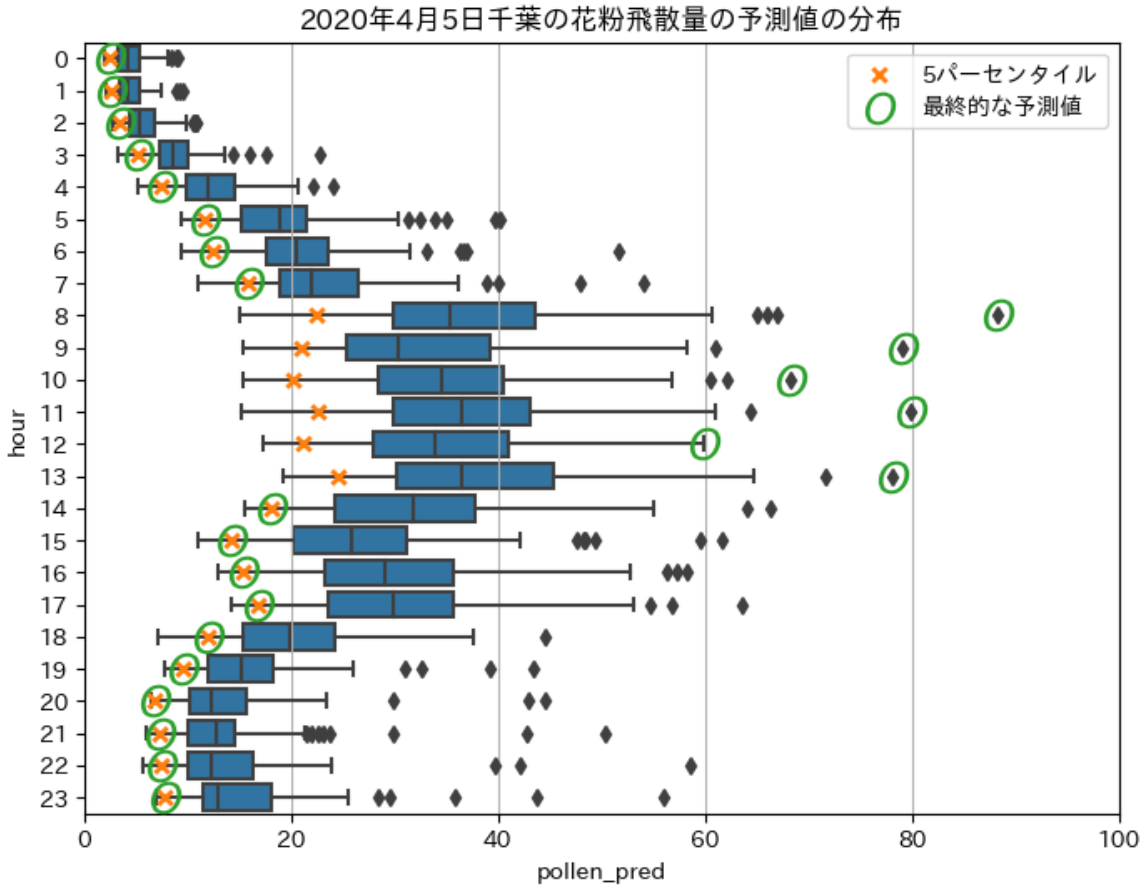
| 順位 | チーム名 | メンバー | 最新 | 提出回数 | ベストスコア |
|---|---|---|---|---|---|
| 1 | - | cczouk | 22日前 | 25 | 7.86103 |
| 2 | AI-FOX | Yosemite<br>MOK<br>te2T<br>moto1963<br>saru_da_mon | 21日前 | 23 | 8.17224 |
| 3 | - | maruyama | 21日前 | 17 | 8.19703 |
| 4 | - | kif | 22日前 | 60 | 8.24164 |
| 5 | - | uchs | 21日前 | 44 | 8.42503 |

# 目次

- 分析方針を立てるまで
  - トピックを読んで要点を把握する
  - 要点から仮説を立てる
  - 仮説をもとに分析方針を立てる
  - ちゃんとCVしよう…… (反省)
- 解法
  - 解法の概要
  - 解法の詳細｜特徴抽出
  - 解法の詳細｜学習
  - 解法の詳細｜アンサンブル
  - 解法の詳細｜後処理

# 解法の概要

- ブートストラップ法で
  予測モデルを大量に作り、
  予測分布を出す。
- 不確実性の高い時刻を
  バースト時刻とみなす。
  (バースト時刻の予測)
- バースト時刻は予測分布の
  上の方の値を予測値とする。
  (バースト時刻の花粉飛散量の予測)
- 非バースト時刻は予測分布の
  下の方の値を予測値とする。
  (2020年の花粉飛散量が
  例年より少ない問題への対処)



# 解法の詳細｜特徴抽出

- 降水量・気温・風速
- 降水量・気温・風速の指数移動平均 (半減期：1時間・1日・1週間)
- 2週間前の花粉飛散量の指数移動平均 (半減期：1週間)
- 年・月・時間
- 拠点

※予測対象拠点の天候情報のみ使用。

# 解法の詳細｜学習

- モデル
  - LightGBM
  - 拠点別に予測モデルを作ることはせず、
    説明変数に拠点を入れた共通の予測モデルを1つだけ作成。
- 損失関数
  - RMSLE
- ハイパーパラメーター
  - Optunaの LightGBMTunerCV で最適化。

# 解法の詳細｜アンサンブル

- 学習データを2週間ごとのグループに分割し、
  グループ単位でランダムにサンプリングしてデータセットを100個作り、
  それらのデータセットを用いて予測器を100個作成。
- 100個の予測値の5パーセンタイルを最終的な予測値として出力。
- ただし、花粉飛散量がバーストしていると思われる時刻については、
  以下の値を最終的な予測値として出力。
  - 千葉で5パーセンタイルが20を超えている時刻
    - 100個の予測値の最大値
  - 千葉で5パーセンタイルが27を超えている時刻
    - 2019年以前の4月第1週～第2週の花粉飛散量の99パーセンタイル

※補正の対象地域や対象時刻はPublic LBのスコアを参考に選定。

# 解法の詳細｜後処理

- 予測値を4の倍数に丸める。
- 負の予測値を0に置換。

# 予測精度の推移

| 提出内容 | Public LB | Private LB |
|---|---|---|
| 予測結果の不確実性が高い問題への対処<br>(バギング) | 13.00000 | 9.65923 |
| 2020年の花粉飛散量が例年より少ない問題への対処<br>(平均値から5パーセンタイルへの変更) | 12.10448 | 8.85130 |
| 花粉飛散量バースト補正<br>(5パーセンタイルが20以上・千葉のみ) | 11.03980 | 8.54399 |
| 花粉飛散量バースト補正<br>(5パーセンタイルが20以上・府中のみ) | 12.10448 | 8.85130 |
| 花粉飛散量バースト補正<br>(5パーセンタイルが20以上・宇都宮のみ) | 12.81095 | 9.40397 |
| 花粉飛散量バースト補正<br>(5パーセンタイルが27以上・千葉のみ) | 10.08458 | 8.19703 |

# まとめ

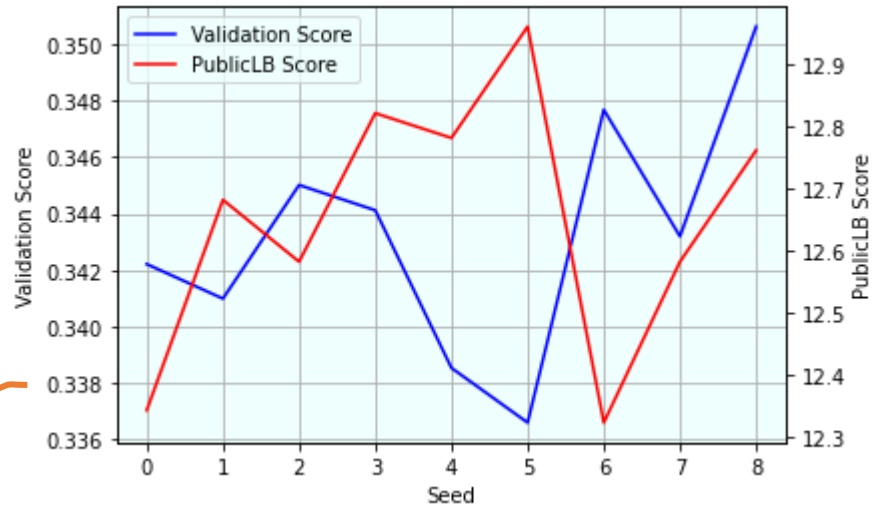
花粉飛散量予測の以下の問題点を予測分布の推定により解決し、
高精度な花粉飛散量予測 (Public1位、Private3位) を実現した。

- 与えられた天候データだけでは情報が不足している。
  → 予測の不確実性を考慮するため、ブートストラップ法で予測分布を推定。
- 花粉飛散量がバーストする。
  → 予測分布をもとにバースト時刻を特定し、予測値を上方修正。
- 2020年の花粉飛散量が例年より少ない。
  → 予測分布をもとに予測値を下方修正。

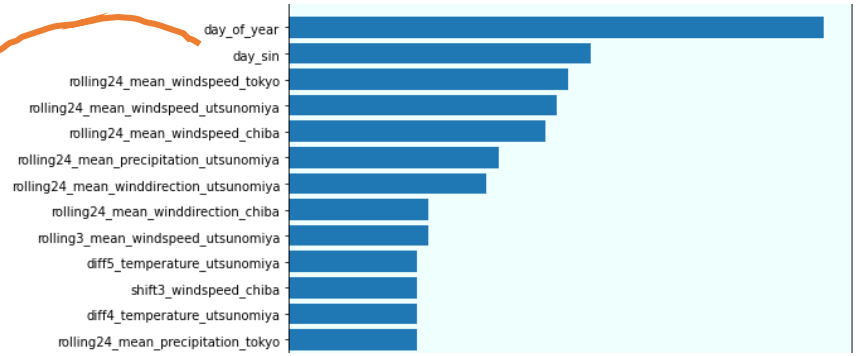
# 所感

CVの設計が難しかった (あまりCVしてないけど……)。

- CVとLBが連動しない。
  - 花粉飛散量がバーストするから。
  - → (対策思いつかず……)
- CVが良くなりすぎる。
  - 時系列データのためi.i.d.が成り立たないから。
  - 対策なしCVで特徴選択すると、
    日時を丸暗記する系の特徴量が選ばれがち。
  - → 2週間ごとのグループに切ってGroupKFold
- CVとLBのスケールが合わない。
  - 2020年の花粉飛散量が例年より少ないから。
  - → 各年度の花粉飛散量のスケールを2020年に合わせてからCV (uchs氏の5位解法)




CV V.S. LB (kotrying氏の「幸運なseed値？」から引用)




特徴量重要度 (kotrying氏の「Model (LightGBM Base)」から引用)