ソニーグループ合同
データ分析コンペティション
1位解法

アジェンダ

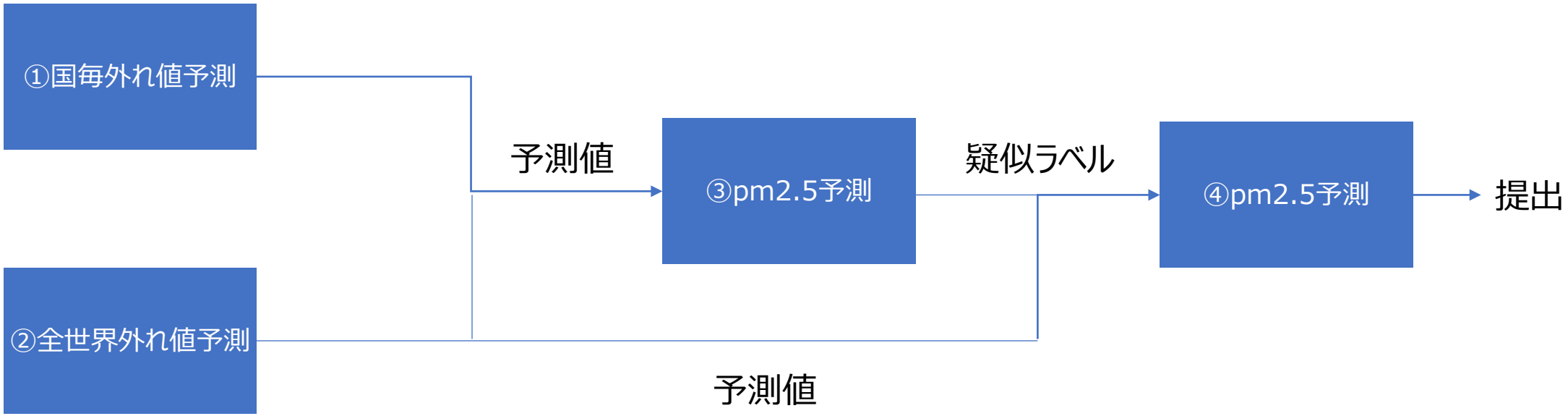
1. 自己紹介
2. モデル構成
3. 使用特徴量
    - 空間
    - 時間
    - 大気
    - 外れ値
    - 外部データ
4. 感想

１．自己紹介

- ユーザー名：pokapokalemon
- 都内のユーザー系IT企業でデータ活用業務に従事
- コンペ参加は4回目

２．モデル構成

4種のlightgbmモデルを使用



３．作成特徴量（空間）

- Kmeansによるクラスタリング、クラスタ値の代入
- 首都との距離、国内最大pm25値の観測点との距離
- 国重心からの距離、距離を国土面積で割り算

３．作成特徴量（時間）

- 目的変数 lag特徴量（shift）
  ⇒（国毎）年月、週、曜日・（クラスタ毎）年月

- co_mid lag特徴量（diff）

- コロナ政府対応指数のlag特徴量（shift、diff、rolling）

- （国毎）年月日、週、曜日などのターゲットエンコーディング

3．作成特徴量（時間）

計測頻度により1lagの意味合いが変わってしまう

計測頻度**高**

日付	pm2.5
2019/1/1	50
2019/1/2	60

1行shift

日付	pm2.5_shift1
2019/1/1	nan
2019/1/2	50

計測頻度**低**

日付	pm2.5
2019/1/1	40
2019/3/1	30

1行shift

日付	pm2.5_shift1
2019/1/1	nan
2019/3/1	40

３．作成特徴量（大気）

- mid max min特徴量を合計、合計値の引き算
- co_mid+no2_midの作成
- 飽和水蒸気圧の算出
- 各地点の天候特徴量（外部APIを使用）

３．作成特徴量（大気）

天候データは下記項目の数日～1か月のlag特徴量を作成

日付	積雪量(cm)	日照時間	紫外線の強さ	体感気温	雲割合	雨量(mm)	視認性	風向	地点
2019/1/1	0	11.6	7	35	46	1.4	10	280	12.46113,130.84184
2019/1/2	0	10.3	7	36	40	4.6	10	281	12.46113,130.84184
2019/1/3	0	11.6	7	37	48	1.4	10	284	12.46113,130.84184

３．作成特徴量（外れ値）

- 国毎・全データ外れ値予測モデルの構築
- 国毎外れ値発生日フラグの作成
- 年末年始、クリスマス、旧正月等イベントフラグの作成

３．作成特徴量（外部データ）

No.	カテゴリ	内容
1	天候・大気	各観測地点の天候
2	天候・大気	ケッペンの気候区分
3	天候・大気	co2データ
4	国土情報	国土の広さ、人口密度、人口
5	国土情報	国毎中心位置
6	国土情報	首都の位置、首都人口
7	国土情報	各都市の人口
8	国土情報	国毎の宗教人口
9	エネルギー	電気使用量
10	エネルギー	クリーンエネルギー率
11	コロナ	自動車売上
12	コロナ	公園や職場への移動指数
13	コロナ	コロナへの政府対応指数
14	コロナ	コロナ感染者数

４．感想

今回、参加1か月で約350回の精度検証を行っていました。
これだけ夢中になれるコンペを主催いただいた皆様ありがとうございました！

ご清聴ありがとうございました！