ヤマトグループ
5Daysコンペインターン

Team C:
森 雅也(Masaya Mori)
石橋 舜(Shun Ishibashi)
鈴木 晴勝(Harumasa Suzuki)

発表内容

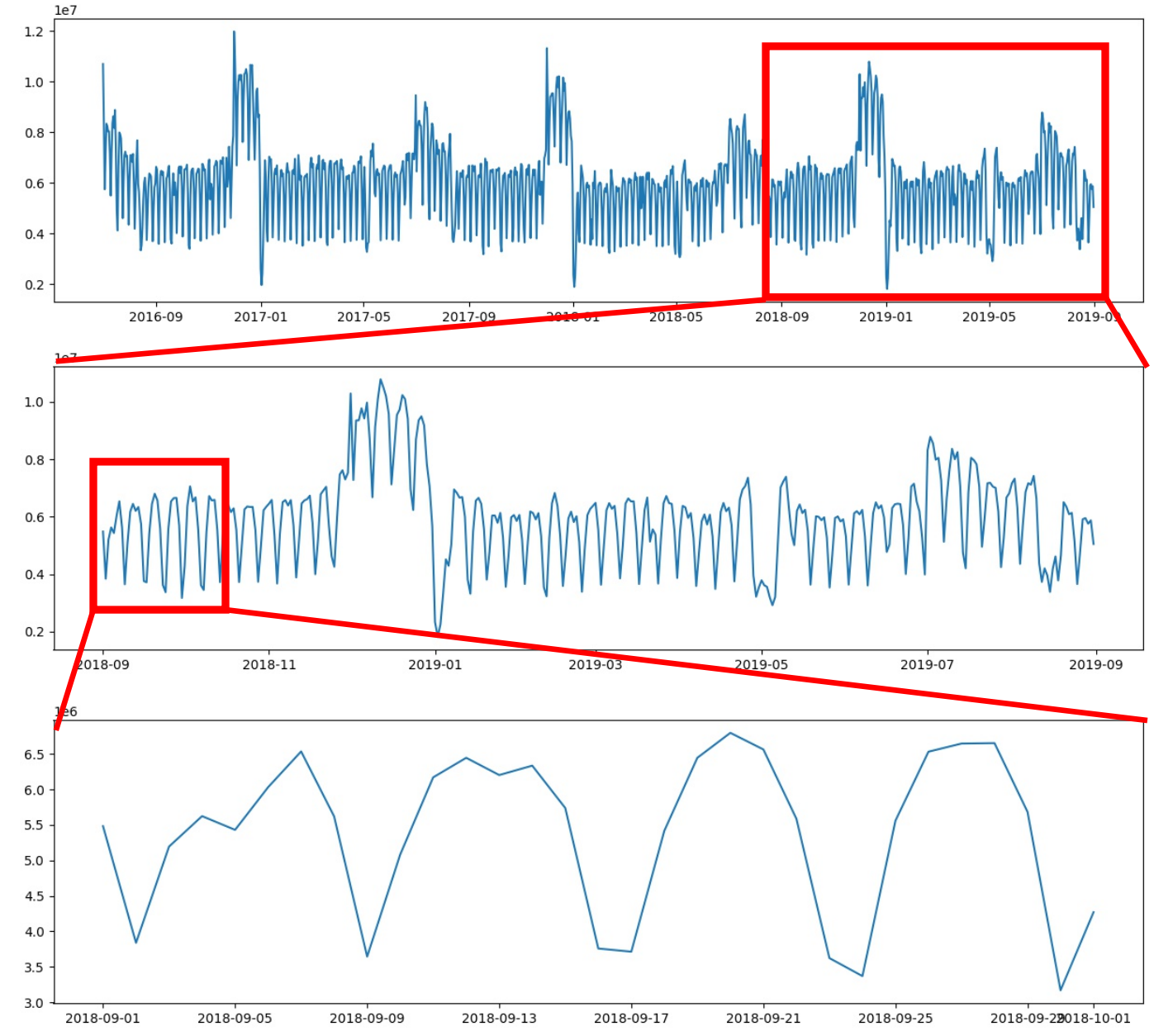
1. データ分析
   → 波形の可視化，自己相関の確認

2. 特徴量エンジニアリング
   → ラグ特徴量，ラグ間の分布の統計量，月・曜日情報の加工

3. モデルの選択
   → 多項式，Ridge回帰，K近傍回帰，ランダムフォレスト，LightGBM

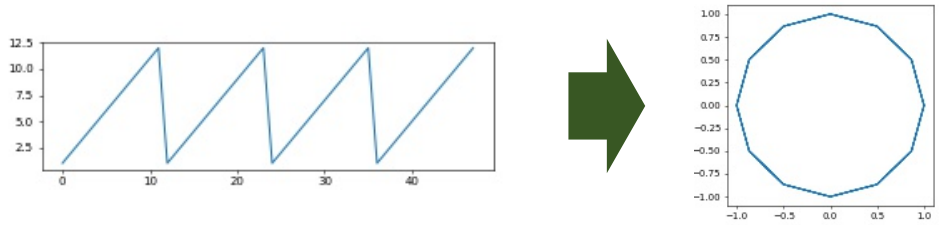
4. モデルの設計
   → アンサンブル学習(スタッキング)

データ分析/特徴量エンジニアリング



- 全ての期間を通して右肩上がり・下がりなどのトレンドは無さそう，値が大きい
  →対数系列

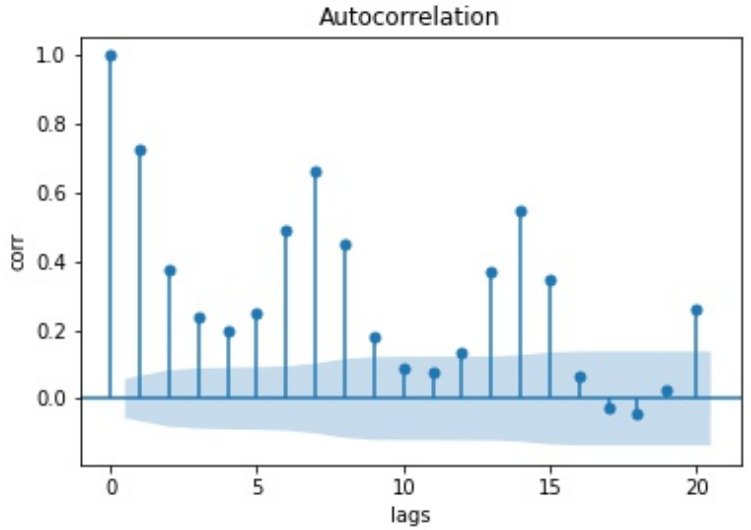
- 月ごとの配送量に違い，特に7月と12月
  → 月を三角関数に変換し追加



- 月～土が配送量多め，日・祝日が配送量少なめ
  → 日・祝日情報の追加
- 曜日により配送量にばらつき
  →曜日を三角関数に変換し追加
- 1週間ごとに周期性？ → 自己相関

データ分析/特徴量エンジニアリング

自己相関：周期性がわかる



- 6日周期，7日周期，8日周期，13日周期，14日周期，15日周期で強い相関
  → ラグ特徴量に変換し入力へ
- 周期間の分布の統計量が知りたい
  → 周期間の平均・分散・尖度・歪度を入力

まとめると

1. 対数系列のラグ特徴量(6,7,8,13,14,15)
2. 周期間の平均・分散・尖度・歪度
3. 日曜・祝日情報
4. 月のsin値とcos値
5. 曜日のsin値とcos値

モデルの選択

マクリダキスら[1]によると,

時系列予測は単純なモデルの方が予測精度が高い

(統計モデル > 機械学習モデル > 深層学習モデル)

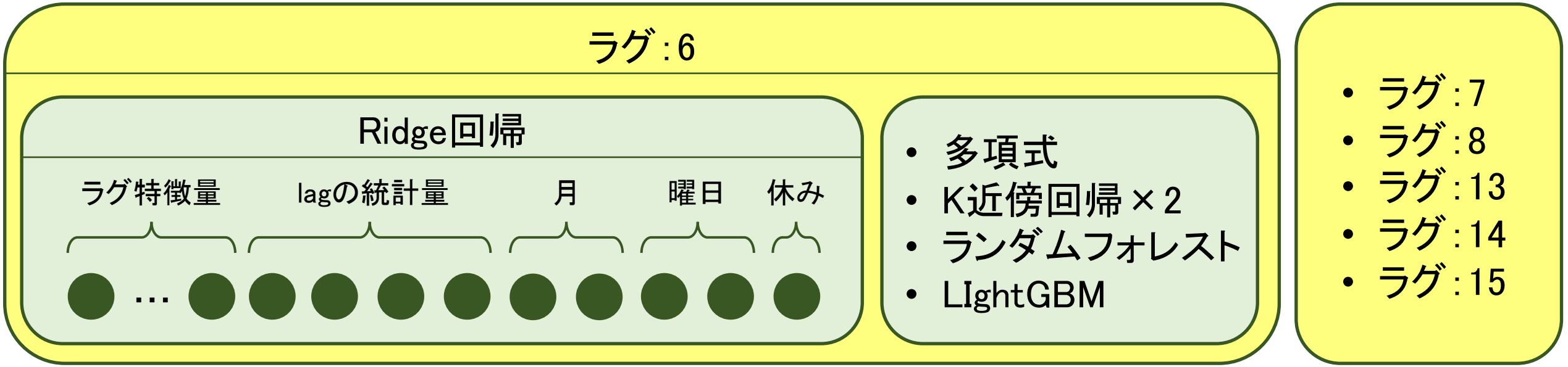
- VAR
- ARIMA
- SARIMAX
- GARCH
- 状態空間モデルなど…

問題1:簡単に使えない
→ 時間的に厳しい
問題2:検定が多数(単位根検定, 残差の正規性…)
→ そもそも使えない可能性
問題3:内容が難しい
→ 1週間で理解は厳しい

機械学習モデルでアプローチ！！！

[1] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS one*, *13*(3), e0194889.

モデルの設計



モデル数：6 × ラグ数：6 = 36個の予測値

スタッキング（線形回帰）

<u>最終的な予測値</u>

まとめ

特徴量：

1. ラグ特徴量(6,7,8,13,14,15)
2. 周期間の平均・分散・尖度・歪度
3. 日曜・祝日情報
4. 月のsin値とcos値
5. 曜日のsin値とcos値

機械学習モデル：

1. 多項式
2. Ridge回帰
3. K近傍回帰×2
4. ランダムフォレスト
5. LightGBM
6. 線形回帰(スタッキング)