SIGNATE「BlueCarbonダイナミクスを可視化せよ！」

藤井

中川

西村

森井

目標

回帰予測タスクに挑戦することで機械学習の体系的理解を目指す。

課題設定

沖縄県全域から取り出されたある地点について、緯度経度などのテーブルデータと衛星画像から抽出した情報を基に、海洋の藻の被度（割合）を予測する。

説明変数の内訳

被度文献値（25）
海洋環境値（225）
ランドサット衛星画像データ（64）
2019年に撮影されたセンチネル衛星画像データ（75×3種類）
2000-2020年のランドサット衛星画像データ（50×3種類×20年分）

成果

参加人数1159人中　70位　（提出人数372人）2023/04/28　暫定

本来3ヶ月かけておこなうコンペであったが参加したのが　4月1日からで一ヶ月しかなかった。

賞金も大きく、プロのデータサイエンティストや企業が参加している大会

その中で手探りながらも頑張った結果としては上出来であると自己評価。

課題解決の流れ

情報収集

・直感的な情報の知見

データの前処理

・欠損値の補完
・ノイズの除去
・意味のある特徴量の生成

モデル

・LightGBM
・RandomForest
・kNN（最近傍法）

情報収集

直感的な情報の知見

情報の収集場所と年月に偏りがある。

距離が近く、時間も同じぐらいのところなら被度も同じっぽい

衛星写真から目視できるのであれば Greenの値が大きい

データの前処理

欠損値の補完

衛星写真には欠損が多かった為、前後の年の情報から線形補間、代表値補完を行った。

ノイズの除去

処理として上位 15%、下位 15%を外れ値とし、除去した。

意味のある特徴量の作成

VARIという農業で用いる緑だけを抽出することができる特徴量を作成
海藻には緑だけでなく赤色のものも存在する為、 VARI'として赤だけを抽出することができる特徴量を作成

モデル

LightGBM

Microsoftによって開発された決定木モデルで序盤から終盤まで大活躍
正式名称はLight Gradient Boosting Machine
米国データサイエンスコンペ Kaggleでひっぱりだこの優秀なモデル

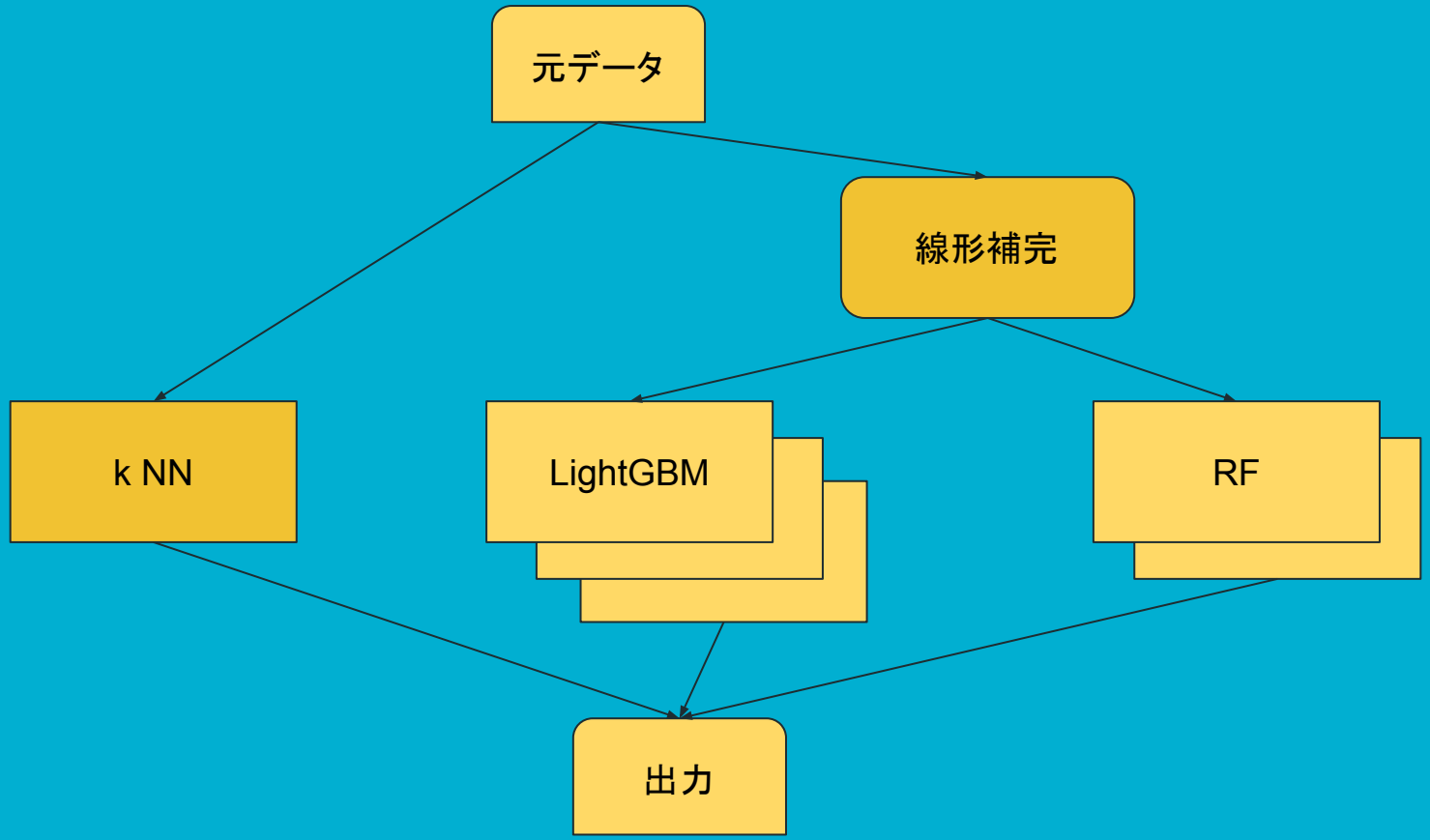
RandomForest

決定木モデル。
ノンパラメトリックなためアンサンブル学習に向いている

kNN（最近傍法）

はじめは敬遠していたが、ものは試しということで採用。
単体でも0.17代のかなり良いスコアを出してくれ、膠着状態からブレイクスルーを生み出した。

予測値を作成するまでの流れ



試したこと

NN（Dence）

言わずとしれた深層学習。
データサイエンスコンペでは深層学習か LightGBMかの2択といわれているそう。
テーブルデータであったため、全結合のニューラルネットワークモデルを使用した。
しかし、今回予測する値が割合であり小さくうまく学習がすすまなかった。
残り時間も少なく、注力できる人がいなかったため断念。

PCA（主成分分析）

処理後も100程度の次元があるため、次元削減のために検討。
導入途中にタイムアップ。

感想

藤井：チーム開発、ルール共有等反省点は多かった。お疲れ様でした。

中川：チーム開発の難しさを痛感しました。

西村：ついていくのが大変でしたが、楽しかったです。

森井：現場の雰囲気を知ることができて楽しかったです。