データサイエンス

すいすい会

第20回「自動機械学習ツール ForecastFlowのアルゴリズム解説 ～勾配ブースティングを中心に～」

GRI
BIGDATA SCIENCE
& RESEARCH

すいすい会の紹介

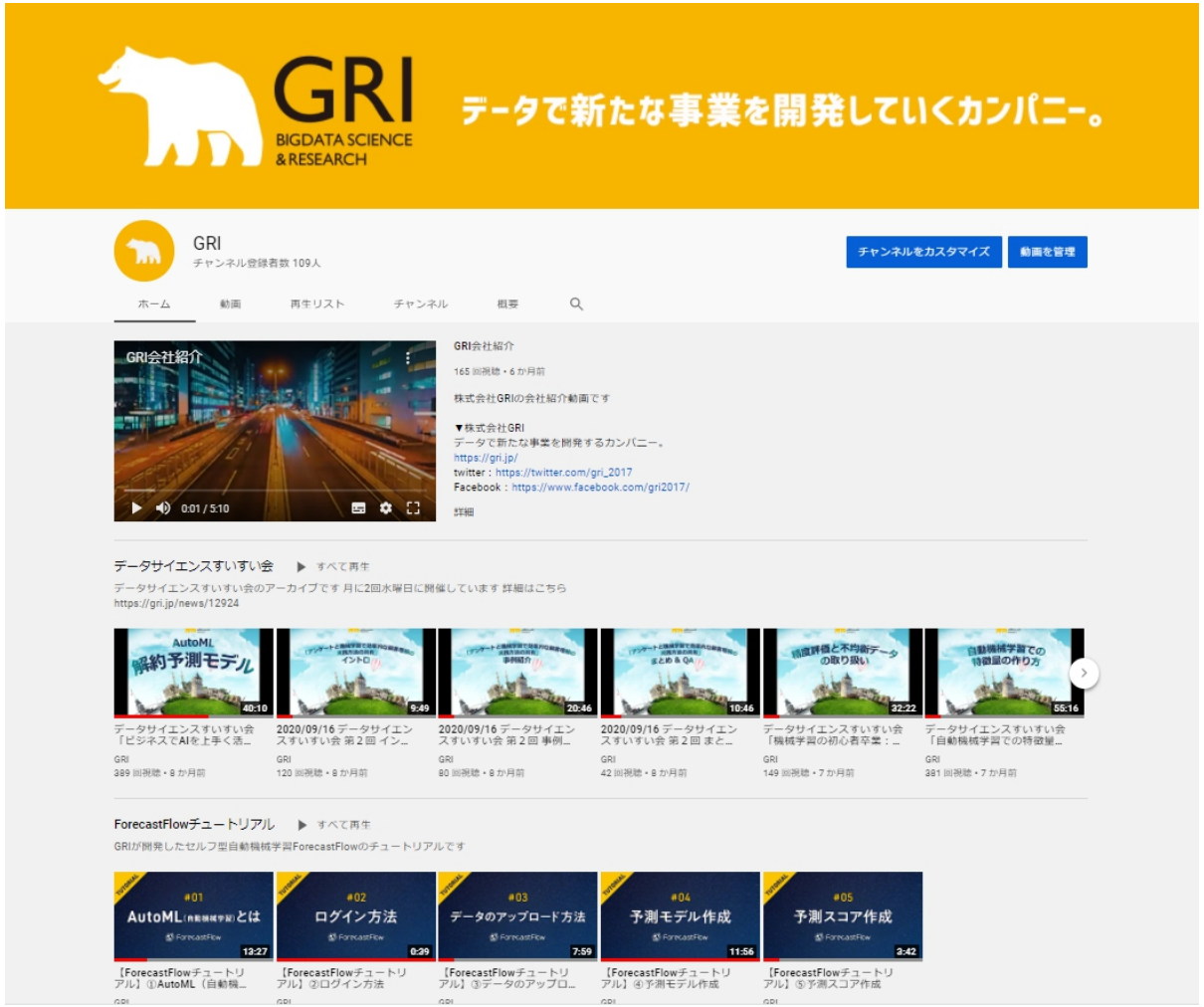
■データサイエンスに関する知見を共有しあい
　みんなで実践的な理解度を高めたい

■資料：　GRIホームページ　https://gri.jp/news/12924

■Slack　ForecastFlowチャンネル　https://join.slack.com/t/forecastflowusers/shared_invite/enQtNTgyMjcxOTg0NzcxLTBkOWEzYWMwNDJmNTkyMDQzYmIxYWU0YWI4ZmU3ZDU0ZTMxNDUwODAxMWFmYmU1YjJiZGI0MjRhYWYyYTNlZTQ

過去のすいすい会の動画（YouTube）

https://www.youtube.com/c/GRIinc



ナビゲータの背景

古幡征史
株式会社GRI　取締役

- 株式会社GRI 取締役
- Ph.D in Computer Science
- GRIにて50以上のAI, BI, 分析基盤構築プロジェクトをリード
- KPMGコンサルティング、University of Southern California、ドワンゴを経て、2016年9月より現職

©GRI Inc.

データサイエンス

すいすい会

第20回「自動機械学習ツール ForecastFlowのアルゴリズム解説 ～勾配ブースティングを中心に～」

GRI
BIGDATA SCIENCE
& RESEARCH

本日の狙い

- 表形式のデータ（ほとんどのビジネスデータ）に対して、機械学習を適用しようと思う際、「初手として勾配ブースティング」を試すべき
  - その理由は何なんだろう？
  - 上手く予測モデルを作るために考慮しておいた方が良い点は？

- 集合知にするため、皆さんの知見をご共有いただきたいです

©GRI Inc.

タイムテーブル

発表タイトル１　「直感的にわかる勾配ブースティング入門」
ゲスト・スピーカー：　株式会社GRI
奥本翔

発表タイトル２　「勾配ブースティングでも変わらない決定木のお気持ち」
ゲスト・スピーカー：　株式会社GRI
岡部泰三

発表タイトル３　「ForecastFlowのリリース情報」
スピーカー：　株式会社GRI
古幡征史

©GRI Inc.

ForecastFlowの
リリース案内

©GRI Inc.

ForecastFlowに関する共有事項

- 自動機械学習ツールForecastFlowの情報源
  - http://forecastflow.jp/

©GRI Inc.

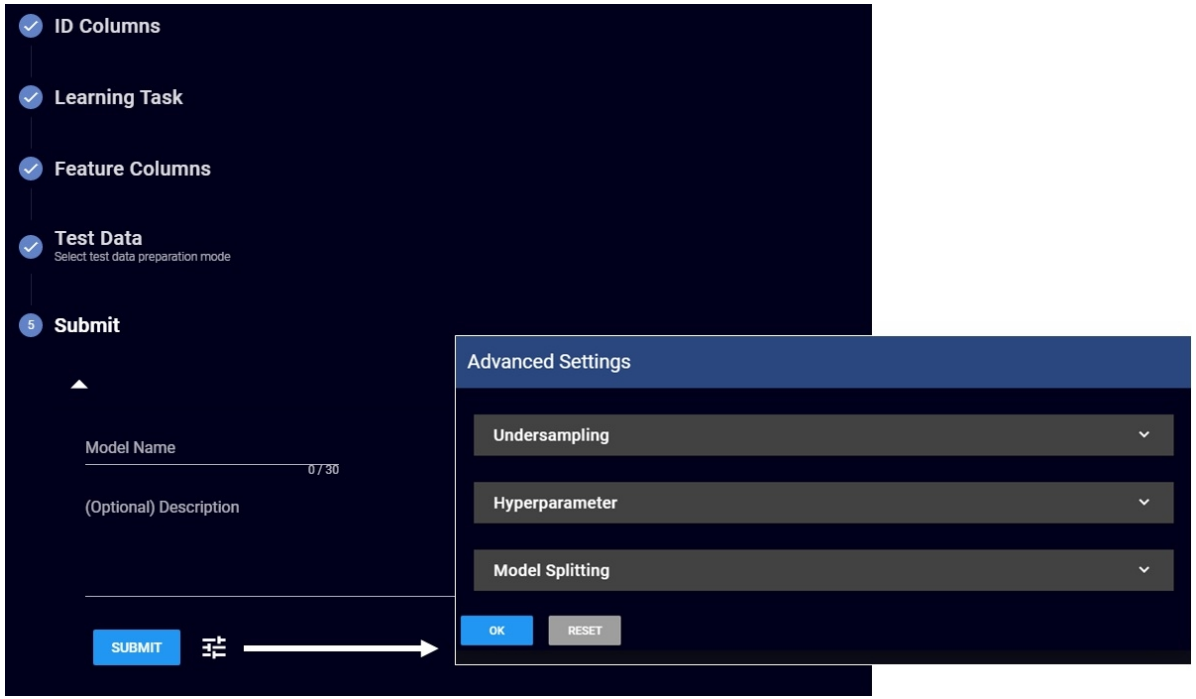
ForecastFlowに関する共有事項

- ForecastFlowのリリース情報（20210705）
- ForecastFlowの既知の問題
- 今後のロードマップ

©GRI Inc.

ForecastFlowのリリース情報（20210705）

- UnderSampling（アンダーサンプリング）：
  - 訓練タスクのモデルタイプがClassification分類（データ上は二値分類）の際、２値の件数が同数になるように、ランダムにアンダーサンプリングされます。
- Hyperparameter（ハイパーパラメタ）：
  - デフォルトで試行回数（Number of Turnings）が20回になっておりますが、試行回数を減らすことで、訓練時間の短縮が見込まれます。なお、試行回数を減らすと予測精度が低下する傾向にあります。
- Model Splitting（モデル分割）：
  - カテゴリ特徴量の値ごとにモデルを自動分割します（"The column to split the model"にてカテゴリ特徴量を選択します）。



アンダーサンプリング機能の注意事項

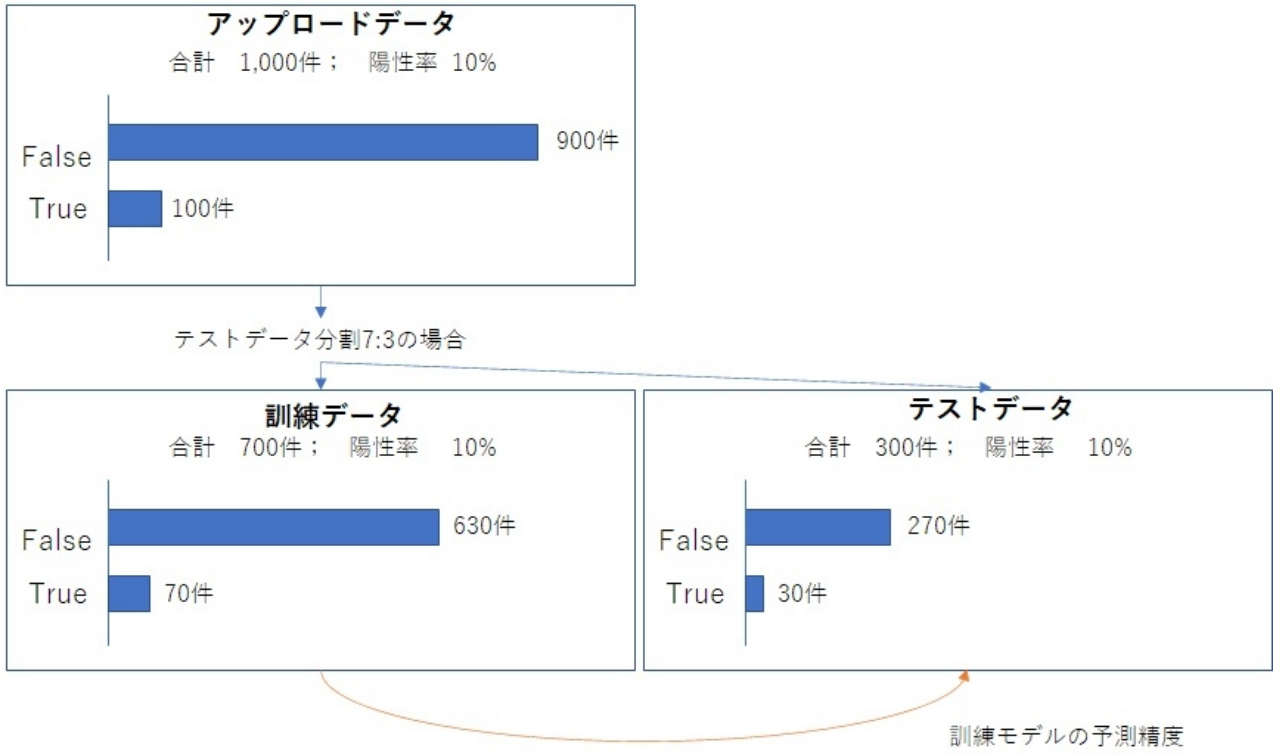
- アンダーサンプリングを利用するシナリオは、不均衡なターゲットデータに対して予測精度を向上させる予測モデルを構築

- 現バージョンにてアンダーサンプリングをした際、実際は予測精度が向上しているにも関わらず、予測精度が減少したように見えるケースが発生します。

©GRI Inc.

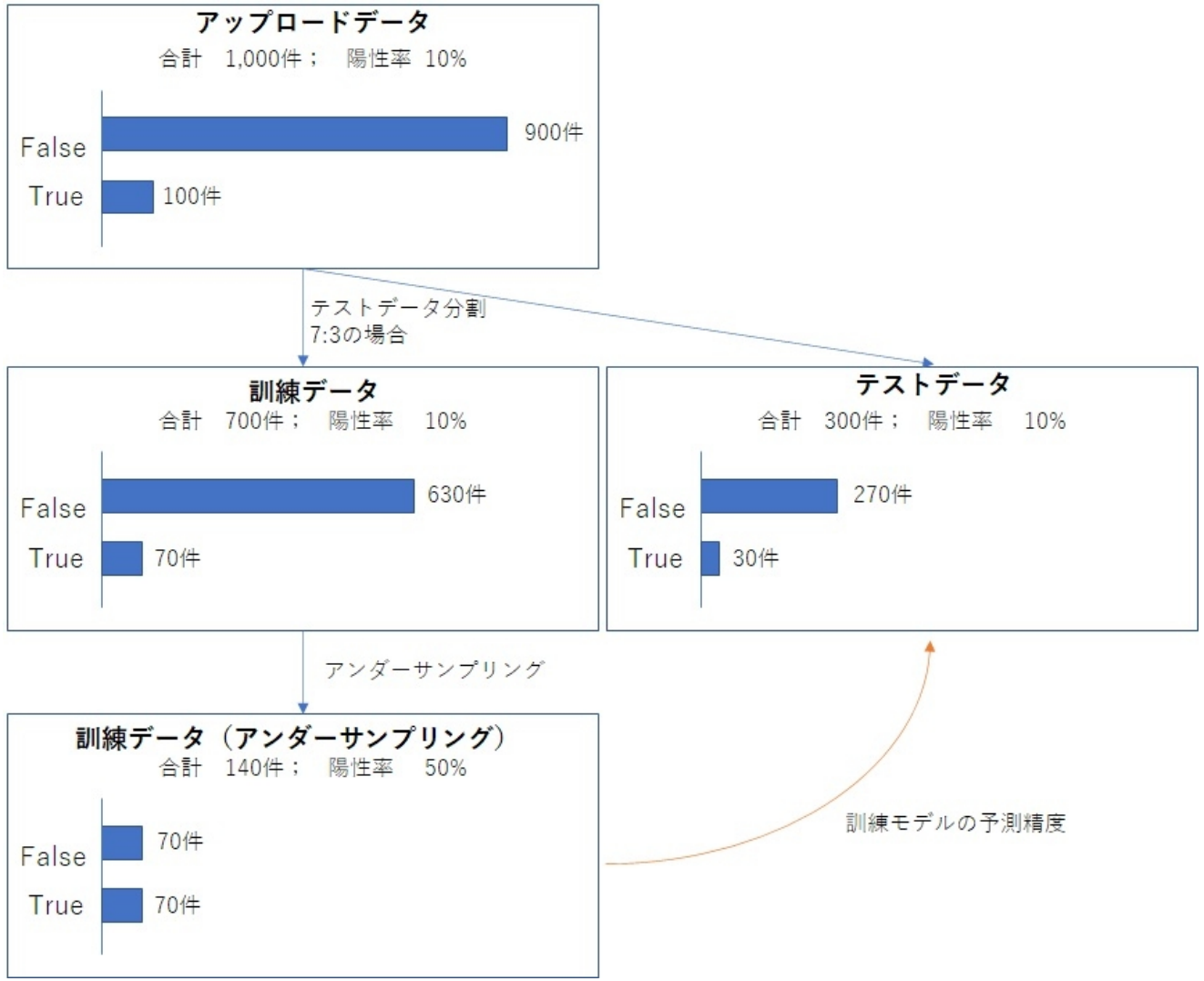
テストデータの分割（再掲）

- 機械学習で予測モデルを構築（訓練）する際、データを訓練データとテストデータに分割します
- 訓練データでモデルを構築して、テストデータで予測精度を評価します



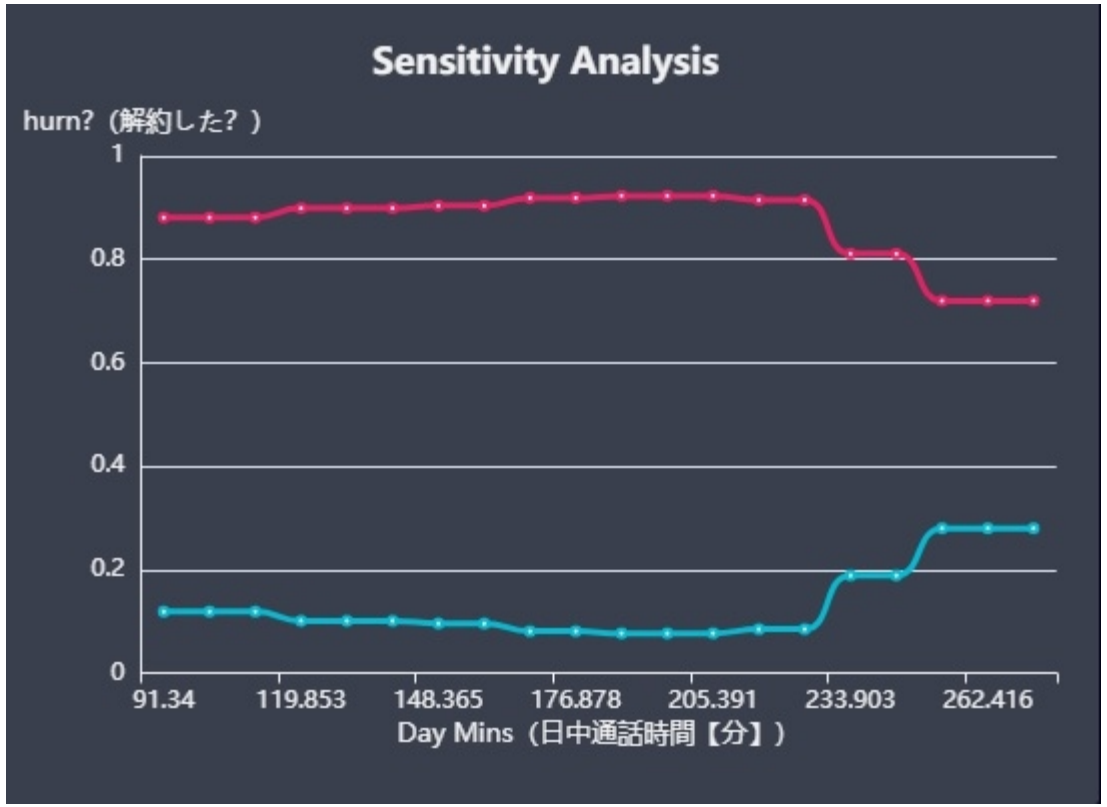
ランダム・アンダー・サンプリングで行っていること

- アンダーサンプリング機能では、訓練データをランダムにサンプリングします。その際、ターゲット（目的変数）のデータ値が同数になるように、ランダムにデータを間引きます。
- アンダーサンプリングされた訓練データで、予測モデルを構築しテストデータで予測します。
- 訓練データ（アンダーサンプリング）とテストデータで分布が異なる
  - 予測スコアの補正をかけている



予測スコアの補正効果による感度分析（Sensitivity Plot）

- 予測スコアの自動補正機能により、かなり現実的な予測スコアが表示されている



©GRI Inc.

予測パフォーマンス・レポートの修正の必要性

- 予測精度を向上させる目的でアンダーサンプリングを行うが、予測スコアが下がって見えてしまうことがある

- 現在のForecastFlowは、予測スコアが0.5になるように調整されている
  - ただし、アンダーサンプリングしたデータで予測モデルを構築した後、予測スコアを補正しているため、先ほどの0.5という閾値を最適化する必要がある（これは、次期バージョンにて修正する予定）

- しばらく、アンダーサンプリングを行った時、予測スコアには最適化が必要なことを、ご理解ください

Performance Report

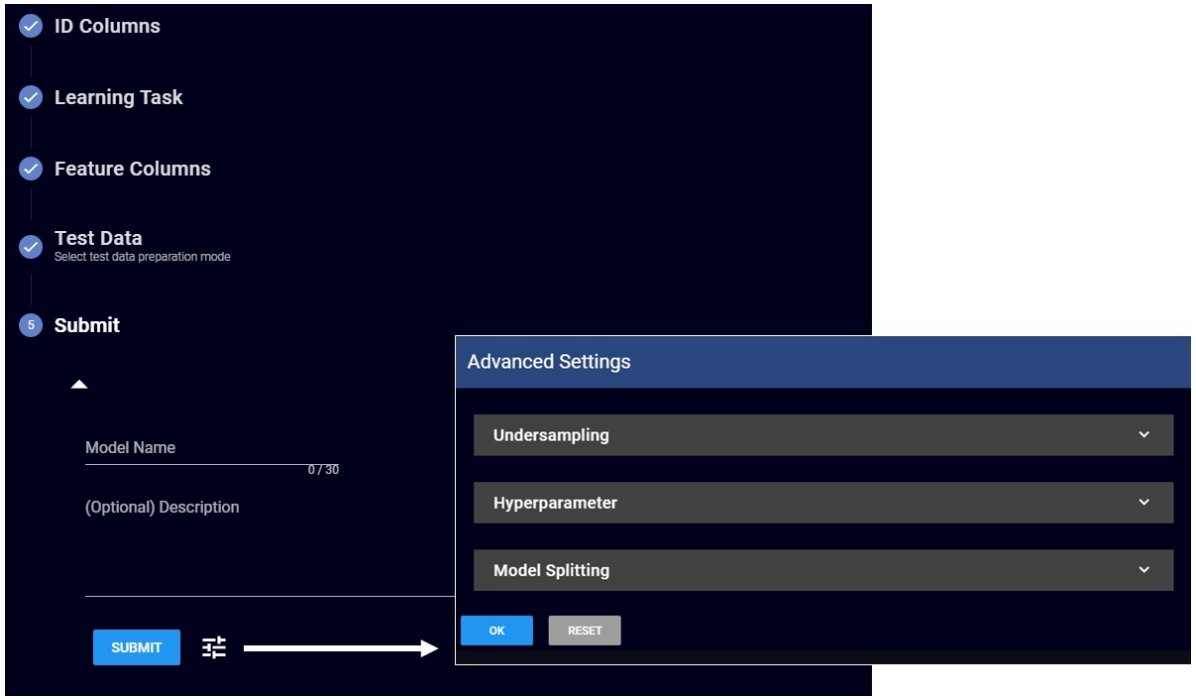
Label Name	F1 Score	Precision Score	Recall Score	Support
False.	0.951	0.932	0.972	855
True.	0.664	0.778	0.579	145

©GRI Inc.

現行リリースの注意点

- UnderSampling（アンダーサンプリング）：
  - 二値分類のみ利用可能

- Model Splitting（モデル分割）とアンダーサンプリングの併用不可:
  - 将来のバージョンアップで対応予定



ForecastFlowのロードマップ

- 20210726週のリリース予定
  - 予測精度（Performance Report）の最適化機能追加
  - チーム機能_v1（チーム内メンバーでのデータ・モデル共有機能）

- 2021年8月以降のリリース予定
  - チーム機能_v2（チーム管理者によるユーザ追加・削除機能）
  - 多言語化（日本語を含む）
  - 予測モデルの解釈画面のダッシュボード化

©GRI Inc.

本日のまとめ

©GRI Inc.

本日のまとめ

- 奥本さんと岡部さんに勾配ブースティングの解説をしました。

- アルゴリズムのお気持ちを考えることにより、モデル改善の参考になればと思います

©GRI Inc.

次回のすいすい会

- 2021年7月28日（水）18:30-19:30

- モダンなデータ分析基盤用ELTツールMatillionの紹介と自動機械学習（AutoML）での使い道
  - 大量のソースシステムからのデータ統合
  - 大規模データの効率的なデータ整形処理
  - 定期ジョブとスケジューリング管理
  - 自動機械学習ツールForecastFlowとの自動連携処理

©GRI Inc.

残っている質問（今後、説明していきたい内容）

- ビジネスの実践例を知りたい
- どのくらい正解データの数があるべきか
- 未来の施策と機械学習の実行タイミングの関係性
- ダウンサンプリングとProbability Calibration
- ForecastFlowのアルゴリズムを知りたい
- 様々な自動機械学習ツールの比較
- 様々なETL/ELTツールの比較
- Feature Store、分析基盤、CDPの説明

©GRI Inc.