データサイエンス

すいすい会

第15回「自動機械学習Google AutoML TablesとForecastFlowの比較」

GRI
BIGDATA SCIENCE
& RESEARCH

すいすい会の紹介

- データサイエンスの領域に関して知見を共有しあい、みんなで実践的な理解度を高めたい

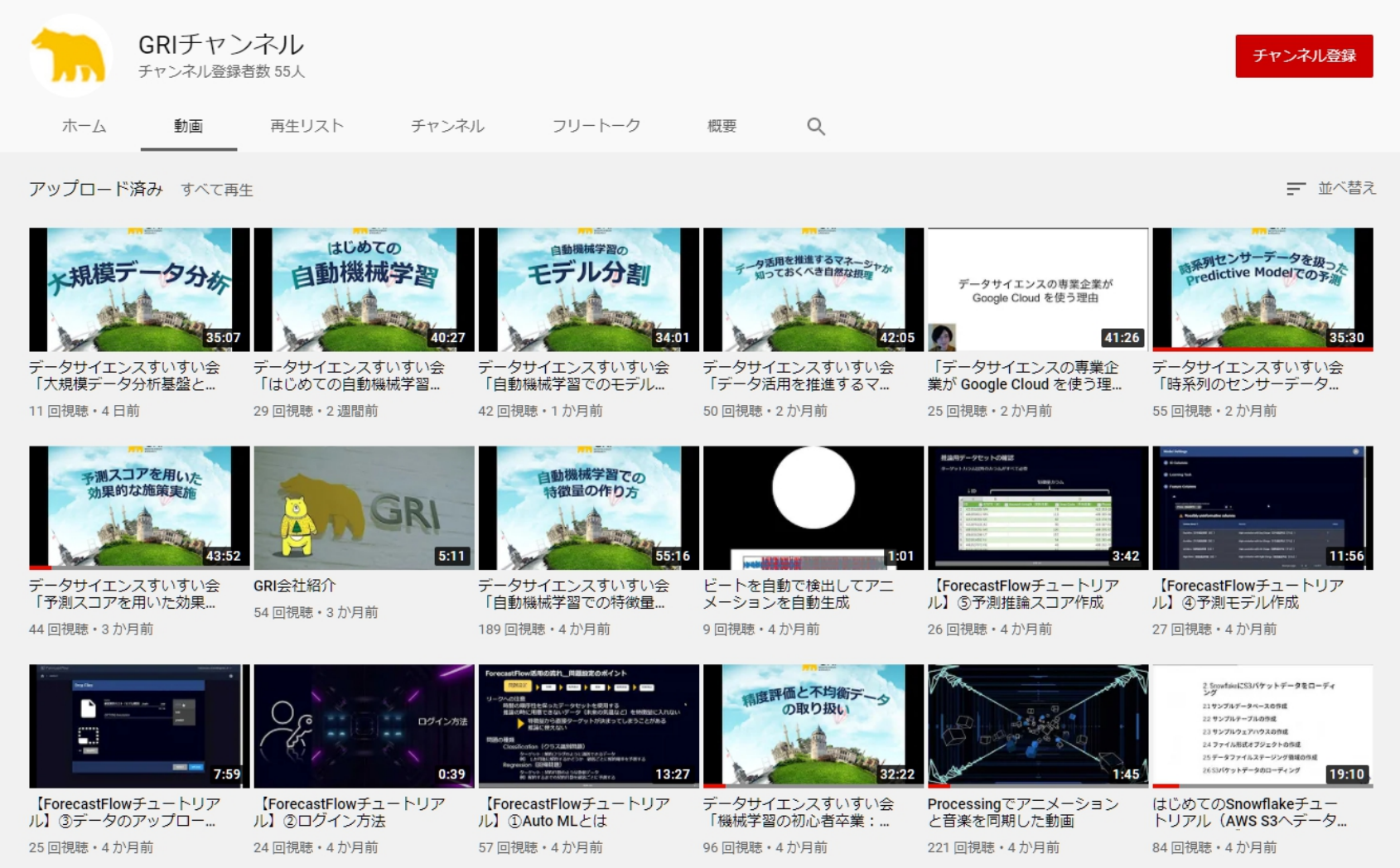
- 自分の知見などを自由に発言してください
  （呼び水のテーマ紹介をします）

- しばらくの期間は、自動機械学習を中心にお話をさせてください

- 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
- YouTube: GRIチャネル　https://www.youtube.com/channel/UCDVGqf-dgczYHMfPzpQ0jNQ

過去のすいすい会の動画（YouTube）

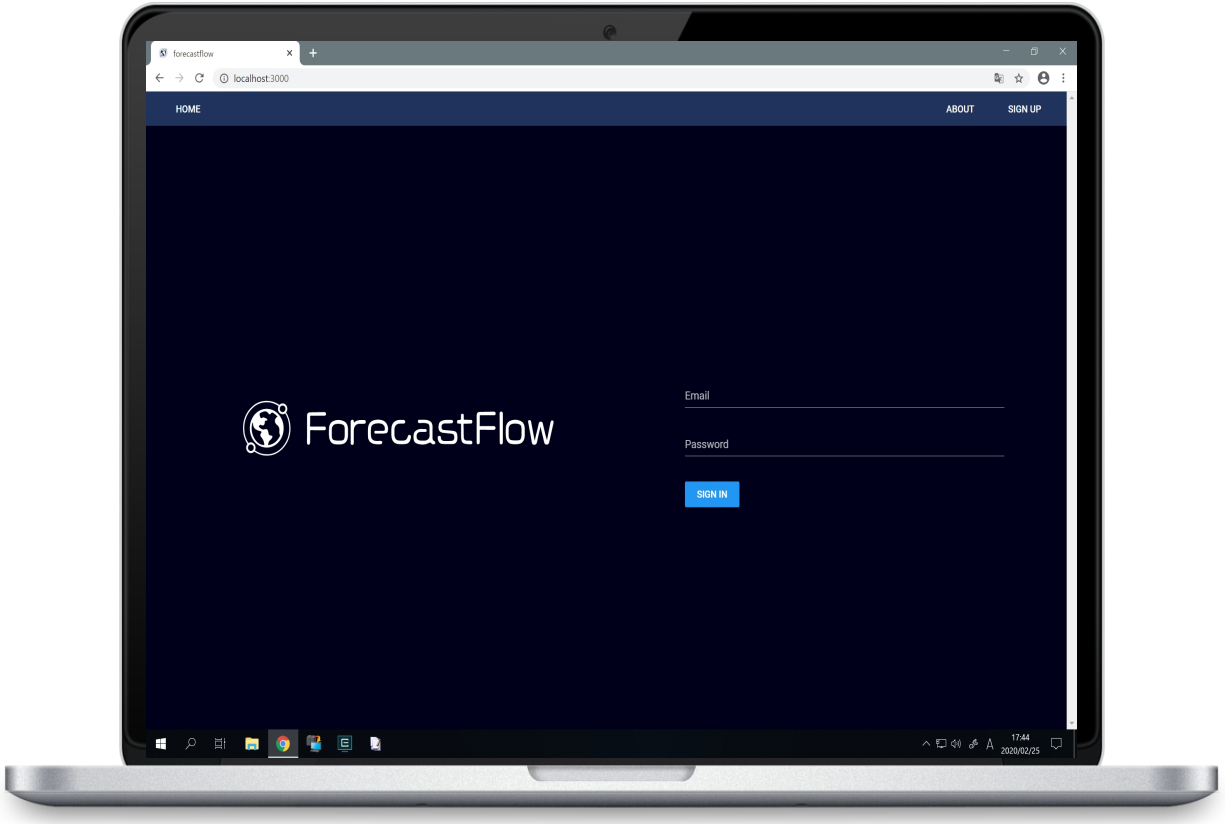
https://www.youtube.com/channel/UCDVGqf-dgczYHMfPzpQ0jNQ/videos



ForecastFlowの簡単説明

入力データに応じて最適な機械学習モデルを自動的に構築する
AutoML（自動機械学習） サービス

- 予測することが得意
- 誰でも予測が「かんたん」にできる



予測活用のベストプラクティスを自動機械学習ツール化

know-how

GRI
BIGDATA SCIENCE
& RESEARCH

ForecastFlow

データサイエンティスト集団

誰でも予測が「かんたん」にできる
（プログラム不要）

Disclaimer

■本発表でGoogle AutoML TablesとForecastFlowを比較します。

■この比較は実機にて行ったものでありますが、特殊なデータの可能性があり、一般性を持った結果ではありません。

■参考程度に見ていただき、ご自身のデータで比較検討することをお薦めいたします。（実データでの比較検証のサポートも可能です）

アジェンダ

- Google AutoML Tablesの基本機能の説明
- 選定基準の設定方法
  - 訓練編
    - 事前的な予測精度
    - 解釈性
    - 訓練速度
    - 試行錯誤のしやすさ
  - 推論編
    - 推論速度
    - 自動化
    - 事後的な予測精度
- 費用
- 連携性

Google Cloud Platform

Google AutoML Tables

- 大規模データの訓練＆推論が高速で処理できる
- 特徴量エンジニアリング等を前提としていない
  - （入力が正しければ、最適な解を出す思想）

デモ

デモのシナリオ～ある米国の電話会社～

■解約が増えていることが悩み

–解約増加の主な原因を探りたい

–原因を探るだけでなく解約を減らす施策も考えたい

–日々更新されるデータを用いて見込み解約顧客を事前に見つけたい

デモのシナリオ～ある米国の電話会社～

■データセットの詳細

- ユーザ数：　3,333
- 特徴量　：　19
- 解約率　：　14.5%(=483/3,333)

デモ

訓練編

事前的な予測精度

■いくつかのデータセットで試したところ、Google AutoML Tablesはトップクラスの予測精度

–F1スコア基準

–ForecastFlowもほぼ同程度の精度

実案件で試したツール

AutoML Tables

ForecastFlow

DataRobot

Prediction One
プレディクション ワン

訓練結果の解釈性

■予測モデルを改善するための手がかりが乏しい

インポート　トレーニング　モデル　評価　テストと使用

モデル

バイナリ分類モデル
trial_2

AUC PR
0.841

AUC ROC	0.883
精度	96.31%
ログ損失	0.157

陽性のクラスである、あまり一般的でないラベルに基づいて指標が生成されます。
精度はスコアしきい値 0.5 に基づいています

モデル ID	TBL5617556638912217088
作成日時	2020/09/14 19:44:59
ターゲット	Churn
特徴量列	19 個含まれます
テスト行	244
最適化の目標	AUC ROC
トレーニング料金	1.396 ノード時間
モデルのハイパーパラメータ	モデル　トライアル
ステータス	未デプロイ

評価全体を見る

バイナリ分類モデル
trial_1

AUC PR
0.844

AUC ROC	0.882
精度	95.9%
ログ損失	0.152

陽性のクラスである、あまり一般的でないラベルに基づいて指標が生成されます。
精度はスコアしきい値 0.5 に基づいています

モデル ID	TBL4538944528156983296
作成日時	2020/09/14 10:47:42
ターゲット	Churn
特徴量列	19 個含まれます
テスト行	244
最適化の目標	AUC ROC
トレーニング料金	0.968 ノード時間
モデルのハイパーパラメータ	モデル　トライアル
ステータス	未デプロイ

評価全体を見る

訓練結果の解釈性

■予測モデルを改善するための手がかりが乏しい

インポート　トレーニング　モデル　評価　テストと使用

モデル
trial_1

バイナリ分類モデル
2020/09/14 10:47:42
トレーニング費用: 0.968 ノード時間

ターゲット	特徴量列	最適化の目標	AUC PR	AUC ROC	精度	ログ損失
Churn	19 個含まれます 244 個のテスト行	AUC ROC	0.844	0.882	95.9%	0.152

指標は、最も一般的でないクラスを陽性のクラスとして使用して生成されます。精度は 0.5 のスコアしきい値に基づいています

→ テスト データセットの予測を BIGQUERY にエクスポート

テスト データセットを BigQuery にエクスポートできるのは最大で 30 日間です

ラベルをフィルタ

False.

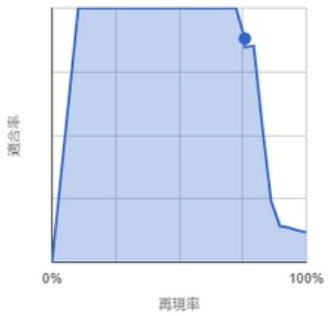
True.

True.

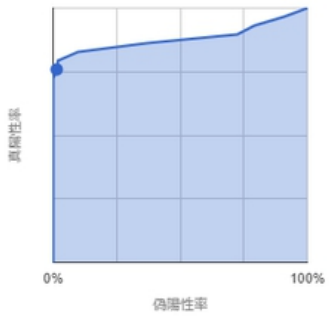
スコアのしきい値 0.50

F1 スコア	0.815
精度	95.9%（234/244）
適合率	88.0%（22/25）
真陽性率（再現率）	75.9%（22/29）
偽陽性率	0.014（3/215）

スコアしきい値は、予測が陽性になるのに必要な最低レベルの信頼度を決定します。モデル評価の詳細



AUC: 0.844　PRC



AUC: 0.882　ROC

訓練結果の解釈性

■予測モデルを改善するための手がかりが乏しい

混同行列

混同行列は、誤った分類が発生している場所（どのクラスが互いに「混同」されているか）を理解するのに役立ちます。各行は予測対象のクラスであり、各列は観測対象のクラスです。表の各セルは、各分類予測が観測対象の各クラスと一致する頻度を示します。

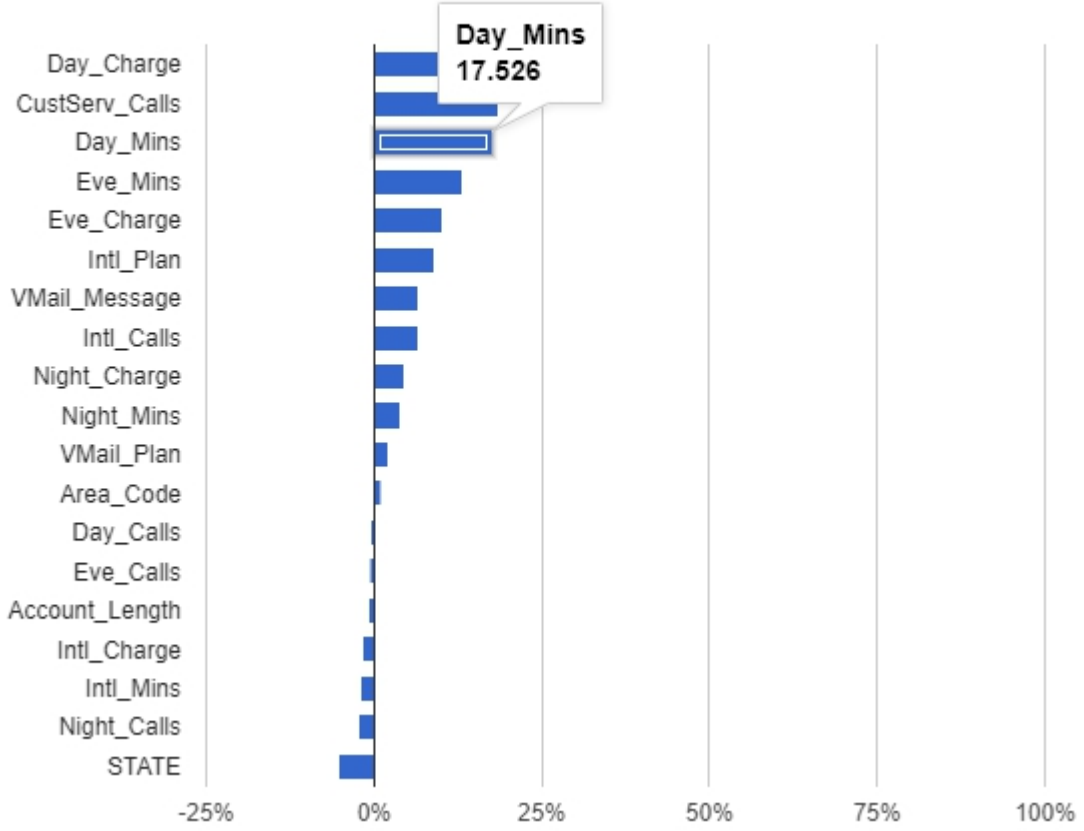
True ラベル: False.
予測ラベル: False.
再現率: 98.6%
212/215 行

正ラベル	False.	True.
False.	99%	1%
True.	24%	76%

訓練結果の解釈性

■予測モデルを改善するための手がかりが乏しい

特徴量の重要度



軸が固定されており、読み取りが困難

訓練速度

■訓練の時間は1~72時間のモードから選択

–長い訓練時間の方が精度が良い傾向

■ForecastFlowで数分で終わるデータセットでも１時間かかる（Google AutoML Tablesはいくつかの不要なアルゴリズムをわざわざ念のため走らせているため）

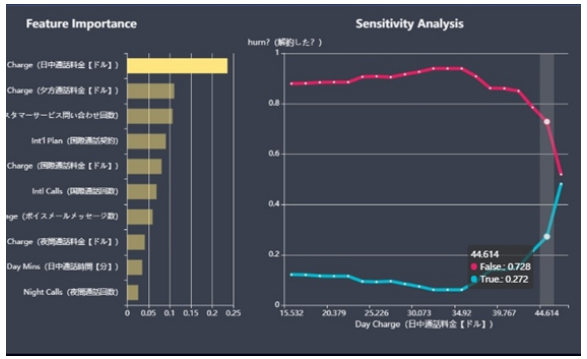
■ただし、ForecastFlowで数十時間かかるデータセットを３時間で終わらせることができる（ForecastFlowは途中打ち切りのような仕掛けがまだない）

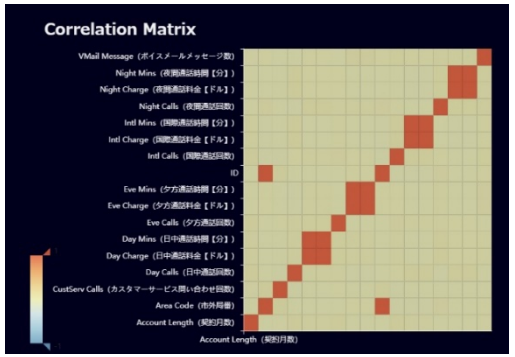
試行錯誤のしやすさ

■モデル改善の手がかりが、本当に何もない

■特徴量間の相関分析、感度分析、モデル分割等、一切ない

ForecastFlowの対話的に訓練結果やデータ理解の画面



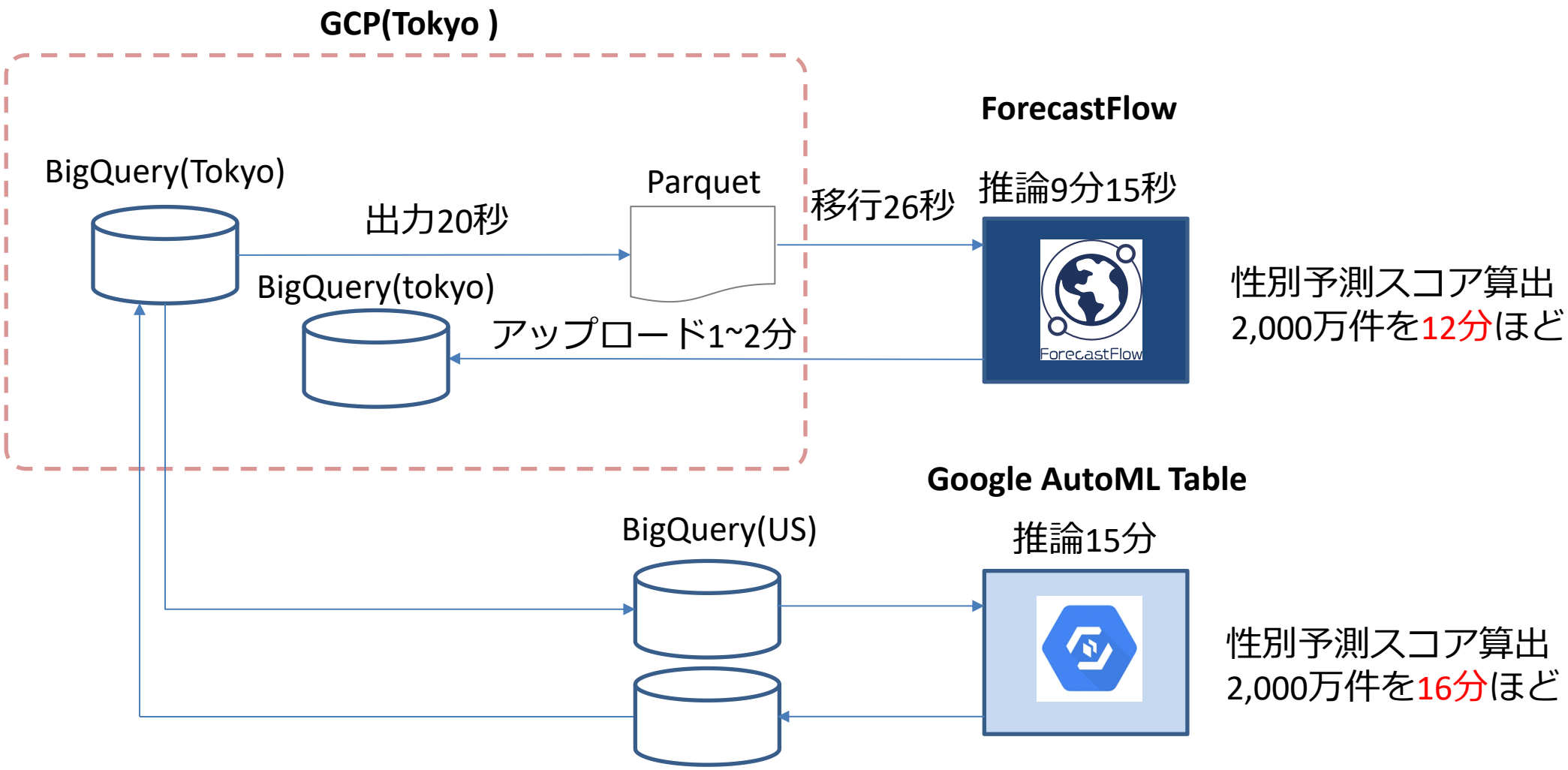


推論編

推論速度は超高速

- 約2,000万件のデータセットを15分で、予測スコアの付与
  - BigQuery連携時
- なお、同じデータをForecastFlowは9分15秒
  - 最近67倍の高速化を実現
  - Parquetファイル連携時

パイプラインでの実行時間（他のツールは、相当時間がかなるデータ量）



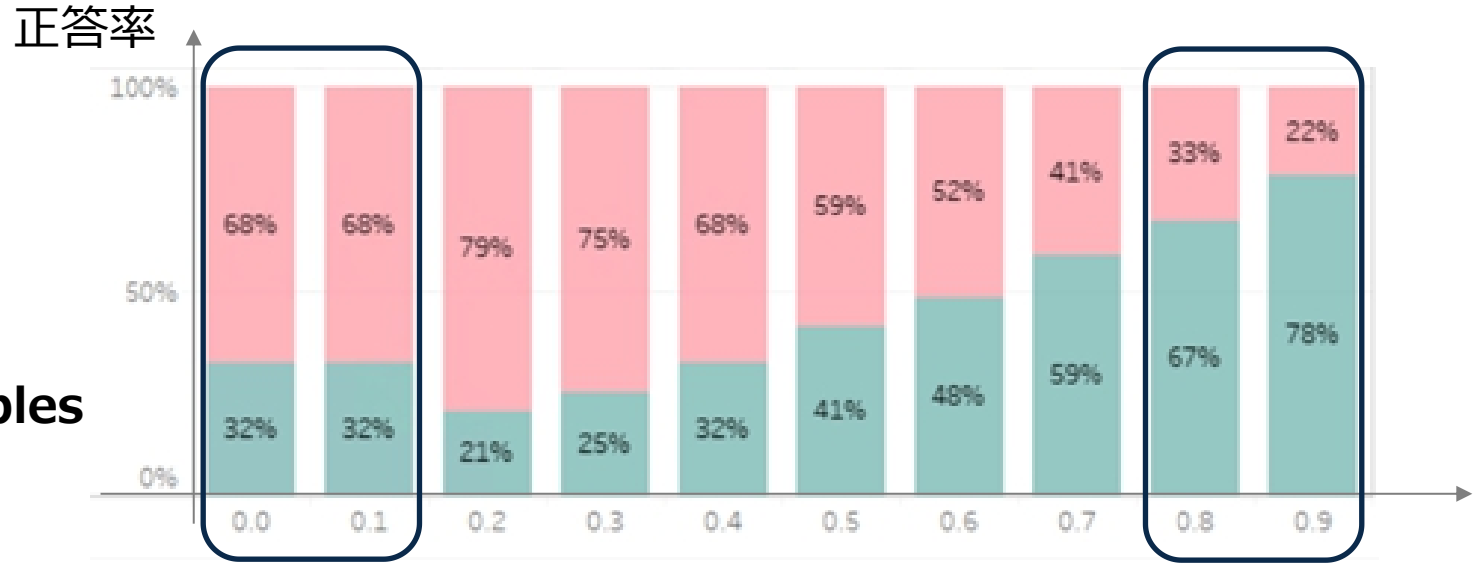
推論の自動化

■スケジュール実行機能

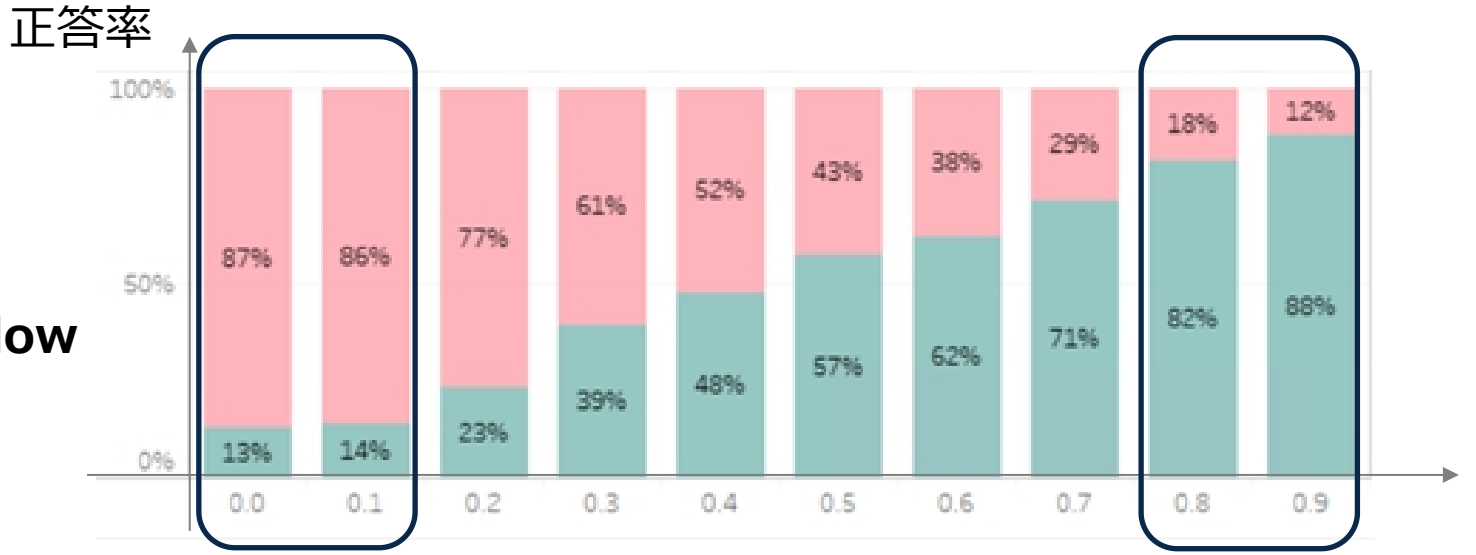
■API経由で自動化を行う

事後的な予測精度（予測スコアと実データの比較）

Google AutoML Tables



ForecastFlow



※Google AutoML Tablesは確度の高い予測スコアが少なかった（両端）

費用

■訓練は時間課金

– 1 訓練当たり、数百円～数万円（大きなデータの場合）のケースが典型的

–試行錯誤をする前提だと、データサイエンティストに精神的に負担

■推論に関しては、ほとんど無料と考えてOK

連携性

■GCPのUSやEUリージョンを前提

■GCS上のCSV、あるいはBigQuery

Google AutoML Tablesに関するまとめ

■試行錯誤を前提にした自動機械学習ではない

■正しいデータセットがあると考えられる場合、高精度な訓練結果を、そこそこの訓練時間で、超高速な推論結果を得られる

■金額は、試行錯誤しないならば安価で済む

■参考資料

–～AutoMLで実践する～ ビジネスユーザーのための機械学習入門シリーズ【第 4 回】AutoML のための ML デザイン

https://cloud.google.com/blog/ja/products/ai-machine-learning/practical-machine-learning-with-automl-series-4

–ただし、AutoML Tablesにない画面（pythonで自分で出力した画面）が含まれているので要注意

次回のすいすい会

- 2021年4月28日（水）12:00-13:00

無料トライアル・アカウントの配布

■ トライアル・アカウントの詳細ページ

- https://forecastflow.org/freetrial/welcome

残っている質問（今後、説明していきたい内容）

- その他の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
- 効果検証
- 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