データサイエンスすいすい会

「特徴量エンジニアリング」

GRI

ForecastFlow

ナビゲータ

- 古幡征史(PhD)
- 株式会社GRI 取締役
- 経歴
  - 筑波大学　　第三学群社会工学
  - 筑波大学大学院　　経営・政策科学研究科修了(MS)
  - University of Western Sydney, PhD in Computer Science（コンピューターサイエンス博士号）
  - Universite Toulouse 1 Capitole, Doctorat en Informatique（コンピューターサイエンス博士号）

  - 丸文株式会社/KPMGコンサルティング/北陸先端科学技術大学院大学/University of Southern California/ドワンゴ
- 分野
  - データ分析＆シミュレーション
  - データ分析基盤構築
  - AI開発
  - BI構築
  - マーケティング、サプライチェーン、物流管理、配送管理、営業、経営、人事管理、財務管理、オークション、制度設計
  - 東京ガス、読売新聞、ソニーネットワークコミュニケーションズ、ドワンゴ、アミューズ、ジブラルタ生命保険、野村総研、東京証券取引所、スパークス、米国運輸省、米国海軍研究所、NASAを含む50社以上

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）

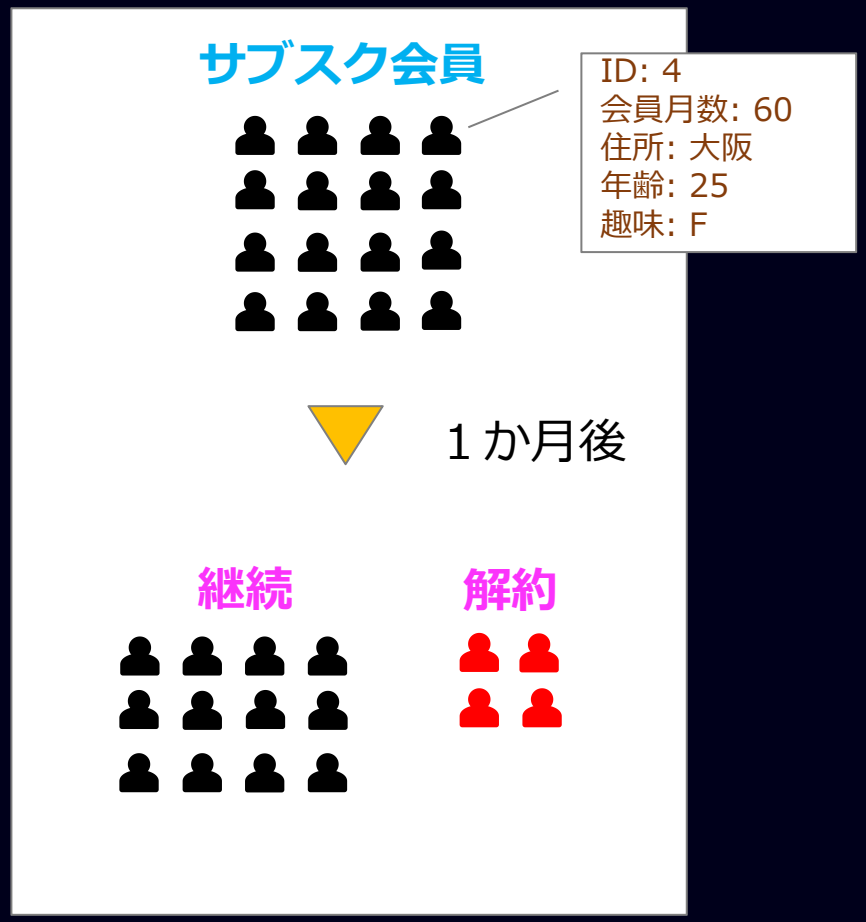
- しばらくの期間は、自動機械学習を中心にお話をさせてください

- 資料：　GRIホームページ
- Slack　ForecastFlowチャンネル
- YouTube: GRIチャネル

©GRI Inc.

表形式の自動機械学習（AutoML）とは

■ 一枚表さえ用意したら、自動で機械学習をさせることができるシステム



機械学習：　ターゲットを特徴量だけで予測すること

ID	ターゲット	特徴量１	特徴量２	特徴量３	特徴量４
1	Yes	120	東京	18	A
2	No	80	千葉	18	B
3	Yes	95	神奈川	20	B
4	Yes	60	大阪	25	F
5	Yes	86	東京	26	G
6	No	92	埼玉	22	C
7	No	105	兵庫	15	C
8	Yes	72	埼玉	19	D
9	Yes	88	東京	28	A

©GRI Inc.

本日の話題：　特徴量エンジニアリング

- 特徴量エンジニアリングの基本形
- 特徴量のアンチパターン
- ForecastFlowで不要な特徴量エンジニアリングのリスト
- Tableauと自動機械学習での特徴量エンジニアリングの切り分け
- 特徴量エンジニアリングのおすすめツール

©GRI Inc.

特徴量エンジニアリングの重要性

- 機械学習の結果から施策へ導くことのできる範囲は、準備した特徴量が予測モデルにおいて効いている場合

- どのような特徴量を準備しておくか？これが機械学習を活用するためのカギ
  - 特徴量エンジニアリング

©GRI Inc.

特徴量エンジニアリングの万能な方法はない

- No Free Lunch Theorem（ノー・フリー・ランチ定理）が、「あらゆる状況に万能な最適化アルゴリズムが存在しない」というように、最適な特徴量を自動生成するのは困難

- 機械的に特徴量を生成することはできるが、施策に役立つ特徴量作りという点においては、現場の知見が不可欠

- よって、現場の人が自動機械学習のプロジェクトに関わるのが最適だと考えている

- 特定用途向けのテンプレートがあるべきだが、GRIから提供できるように努めている（データを見たりヒアリングすれば作れるが、千差万別な現状）

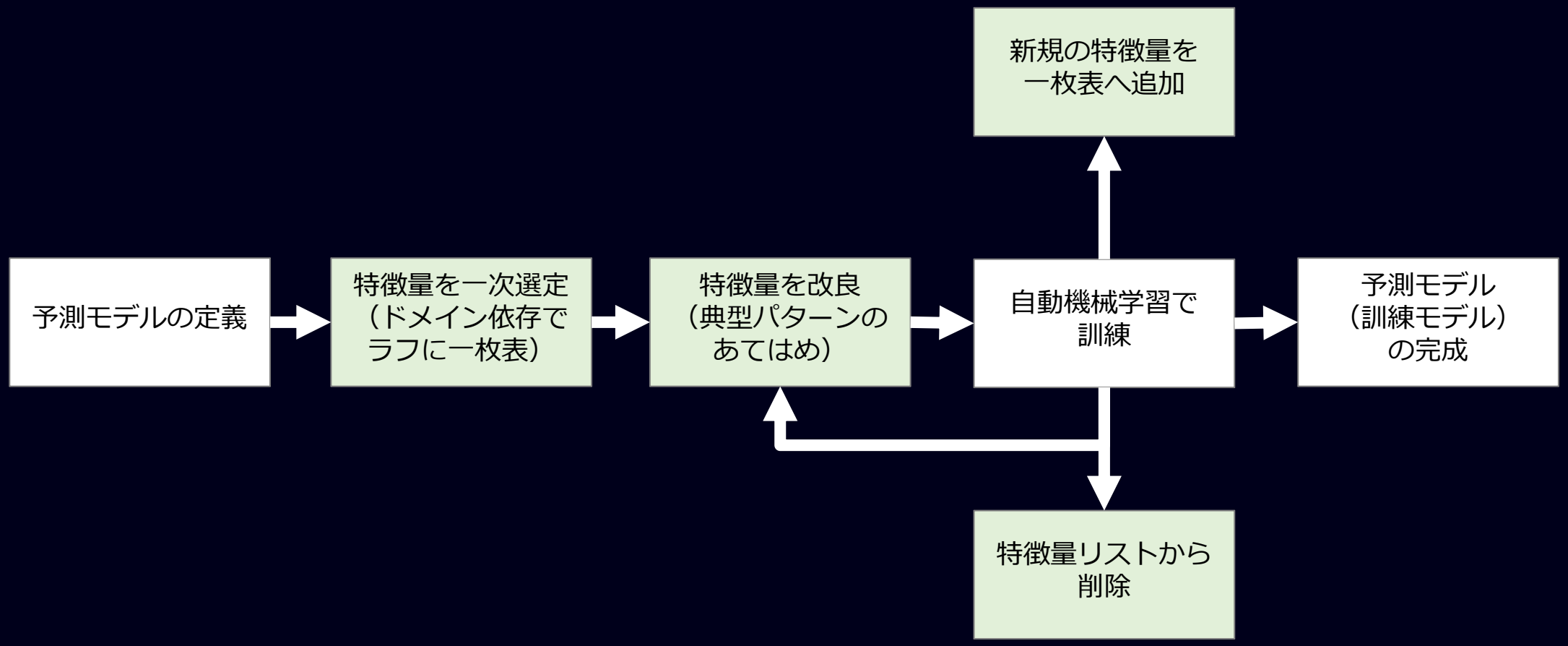
©GRI Inc.

現場の人が使う前提で特徴量エンジニアリングを考える

- 典型例を知ることで、自分たちの現場に合った、特徴量作りに役立つと考えている

©GRI Inc.

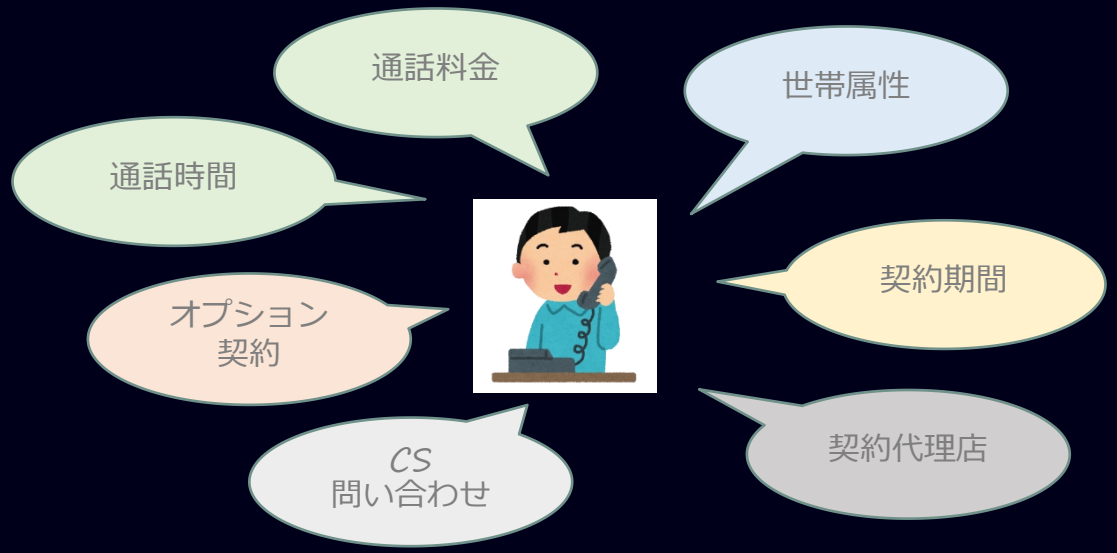
自動機械学習の特徴量エンジニアリングの典型的な流れ



©GRI Inc.

【ユーザに関するドメイン知識】特徴量エンジニアリングの考え方①

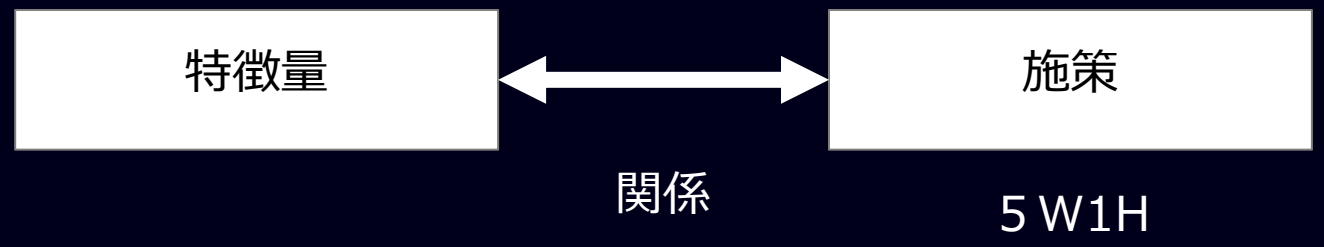
- 特徴量エンジニアリングでは、例えば顧客になりきり、顧客の気持ちを想像しながら、「ユーザはこんな気持ちになるはずだから、こんな人はこんな行動をするはずだから、この特徴量を入れよう」と進めていきます
- チェックする基準
  - ユーザ属性：　性別、年齢
  - ユーザ行動：　利用料、アクセス数



継続／解約を上手く説明できる要因は何だろうか？

【施策に関するドメイン知識】特徴量エンジニアリングの考え方②

- ユーザ属性／行動を説明する特徴量だけでは、施策を考えるには物足りない
- 作った特徴量一覧に対して、施策に関係させられるか？を振り返る
  - 関係部門を頭に入れ、この特徴量に関係する部門で、ユーザにアプローチできることは何か？
  - 当然効くと思われる特徴量が効いていない場合、そこに関連する施策を停止する決断ができるか？



©GRI Inc.

【典型パターン】特徴量エンジニアリングの考え方③

- 典型パターンでは、多くのドメインであてはまる共通した作法

特徴量の表現方法

- 特徴量は大きく分けて２種類
  - カテゴリ（Categorical Feature）
  - 数値（Numerical Feature）

カテゴリ特徴量の例	カテゴリ値の例
性別	男性, 女性
年齢層	10代, 20代, 30代
購入経験	あり, なし

数値特徴量の例	数値の例
推定男性確率	0.94
年齢	12, 46, 58
購入経験フラグ	0, 1
購入金額	5000, 12000

©GRI Inc.

カテゴリ・データに対する操作① <変換>

- カテゴリの種類が大量だと、結果を把握するのが困難になる　→　PDPの例
- 間違っている場合、古い情報
- その他を効果的に使う
  - 集約操作でまとめる
  - 例）メールアドレスのドメイン名

メアドドメイン（元データ）
hotmail.com
saftbank.jp
gri.jp

メアドドメイン（変換後）	変換内容の説明
outlook.com	買収されたので同一サービスに
softbank.jp	スペルを間違えている
その他（企業）	ある括りに集約しないと、あまり出現しないデータになる

カテゴリ・データに対する操作② <カウント>

- カテゴリの出現有無（経験が重要であり、回数に依存しない場合）
  - 購買経験
  - コールセンター問い合わせ経験
  - 過去入会経験
- カテゴリの出現回数をカウント（回数に意味があると思われる場合）
  - 購買回数
  - コールセンター問合せ回数
  - 過去入会回数

ID	POS#
001	P1234
001	P1234
003	P1289

▶

ID	購買経験	購買回数
001	あり	2
002	なし	0
003	あり	1

©GRI Inc.

カテゴリ・データに対する操作③ <相互作用>

- 複数以上のカテゴリ間の相互作用を特徴量にする（斜め方向の交互作用に弱い機械学習アルゴリズムが多い）
  - 例）
    - 性別 x 年代
    - 支払手段 x キャリア（手数料の違いやnotification方法の違いなど）
    - 支払手段 x 購入経路（代理店／販社／商社／Rep）
    - 地域も多くの要素と相互作用

ID	性別（特徴量）	年齢（特徴量）
001	男性	14
002	女性	21
003	男性	33

▶

性別 x 年齢層（特徴量）
男性（10代）
女性（20代）
男性（30代）

©GRI Inc.

数値データに対する操作① ＜基礎統計の利用＞

- ある期間の最大購入金額、最小購入金額、平均購入金額
- 負の購入金額（返品）の回数
- 0の購入金額（クーポン利用）の回数

ID	月間購入金額（平均）	月間購入金額（最大）	月間購入金額（Nケ月前）
001	1,432	1,600	2
002	3,540	5,200	4
003	4,200	4,200	1

©GRI Inc.

時間データに対する操作① <日時→期間>

■ ●●時刻は期間に変換
- 疎なデータになる
- 推論で役に立たない

元データ	元データの例
入会日時	2020-08-30 14:00:00
購入日時	2020-08-30 14:00:00
入会日	2020-08-30
購入日	2020-02-04

特徴量（変換後）	変換データの例	内容説明
会員日数	2	20/9/1 午前0時を起点にした日数
販売開始からの経過（分）	120	20/8/30 午後12時を起点にした時間（分）
入会日の月末までの日数	1	20/8/31を起点にした日数
節分までの日数	-1	20/2/3を起点にした日数

©GRI Inc.

時間データに対する操作② <日時→カテゴリ、数値>

- 日時はカテゴリや数値に変換
  - 疎なデータになる
  - 推論で役に立たない
  - ForecastFlowのAPIで曜日、六曜取得を対応

入会日時
2020-08-30 14:00:00

▶

入会日（季節）	入会日（曜日）	入会日（休日フラグ）	入会日（六曜フラグ）
夏	日曜	休日	赤口

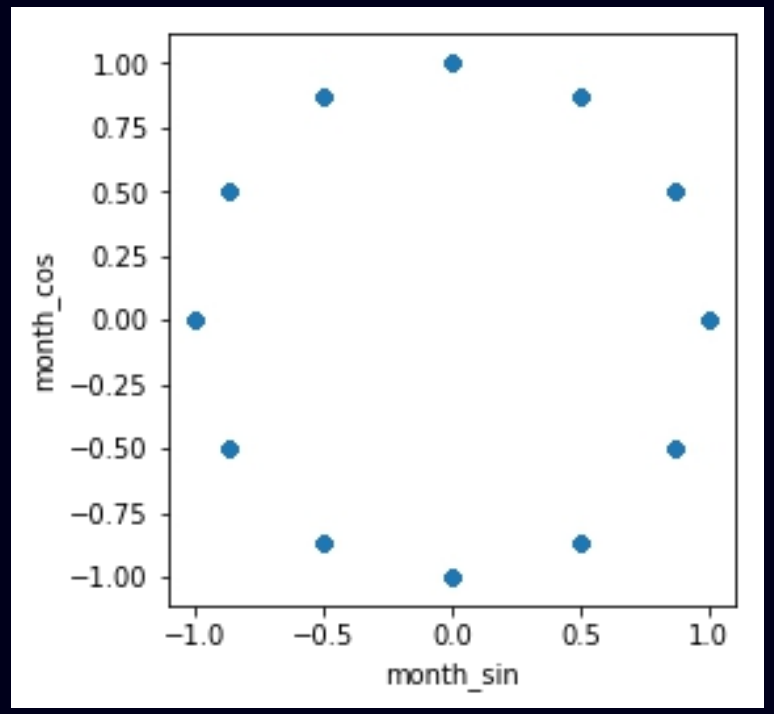
入会年度	入会日（Deci）
2020	2020.66

©GRI Inc.

時間データに対する操作② <日時→数値>

■循環的に表現したい場合（初手はカテゴリ表現で、その後で良い）

- 月、日、時間、分などを三角関数で表現
  - Sin
  - Cos
- 座標で近さを縛れる



https://qiita.com/shimopino/items/4ef78aa589e43f315113

https://datascience.stackexchange.com/questions/5990/what-is-a-good-way-to-transform-cyclic-ordinal-attributes

時間データに対する操作③ <トレンドと対象期間フィルタ>

- 直近１週間の合計購買金額
- 直近１か月の合計購買金額
- 先月と先々月の合計購買金額の差

- 難しい点
  - ユーザによって新規になった期間が異なり、同一の購買期間でない場合
  - モデル分割すべきか検討

ID	合計購買金額（N月）	合計購買金額（N-1月）	合計購買金額の差（N月 － N-1月）
001	1,432	1,600	-168
002	3,540	5,200	-1660
003	4,200	4,200	0

©GRI Inc.

注意：　時間に関して「リーク」を起こしやすい

- 会員の継続予測をする上で、有料コンテンツの最終アクセス日を無加工で利用すると、リーク（そのデータを使うと答えそのものになってしまう）

空間データに対する操作＜変換＞①

- 座標から都道府県＆市区町村へ（細かすぎると解釈できない。推論とバランス）
  - 都道府県：　47
  - 市区町村：　1,741
- 郵便番号から都道府県＆市区町村へ（細かすぎると解釈できない。推論とバランス）
- ForecastFlowのAPIで変換可能（？）

©GRI Inc.

空間データに対する操作＜エンリッチ＞②

- 座標からランドマークへの距離
  - 自宅から最寄り店舗
  - 自宅から最寄り駅
  - 自宅から競合店までの距離の差分
- 座標からランドマークへの最短所要時間
- 座標から都道府県や市区町村のeStat等の調査データ（人口）を付与
- 徒歩５分以内のコンビニ件数
- 店舗名から店舗種別（ロードサイド、駅近、デパート内、ショッピングセンター内）

- Python等を用いる
- 様々なAPIを叩く

©GRI Inc.

時間 x 空間データに対する操作 <エンリッチ> ①

- 日時と場所が特定できると気象データが使える（ForecastFlowのAPI）
  - 天気（表記ゆれ）
  - 気温（平均、最高、最低）
  - 降水量
  - 日照時間
- 場所に関する指定方法
  - 県名（漢字）
  - 都道府県コード
  - 郵便番号

©GRI Inc.

特徴量のアンチパターン

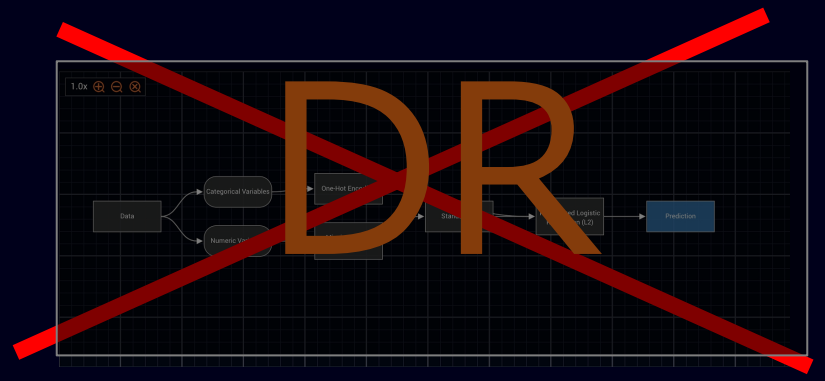
- 日時のデータをそのまま入れてしまう（疎であり、推論時に困難）
  - 入会日時
- 大量の種類のカテゴリ
- ほぼNULLの特徴量
- リークしている
  - 予測の判定時刻以降の最終アクセス日時
- 相関が強い特徴量
  - ForecastFlowは比較的に頑健だが、相関が高いものはドロップした方が無難

©GRI Inc.

ForecastFlowで不要な操作（実務を考えて）

- 下記の操作は、自動機械学習によっては強いられる操作、あるいは精度向上に必要な作業

- 欠損値データのある行の削除作業
- カテゴリの特徴量
  - One-hot encoding :　カテゴリフラグの横持ち
  - Label encoding :　カテゴリを数値への変換
- 数値の特徴量
  - Scaling : 対数や標準偏差、平方根などの変換
  - Regularization: 正則化

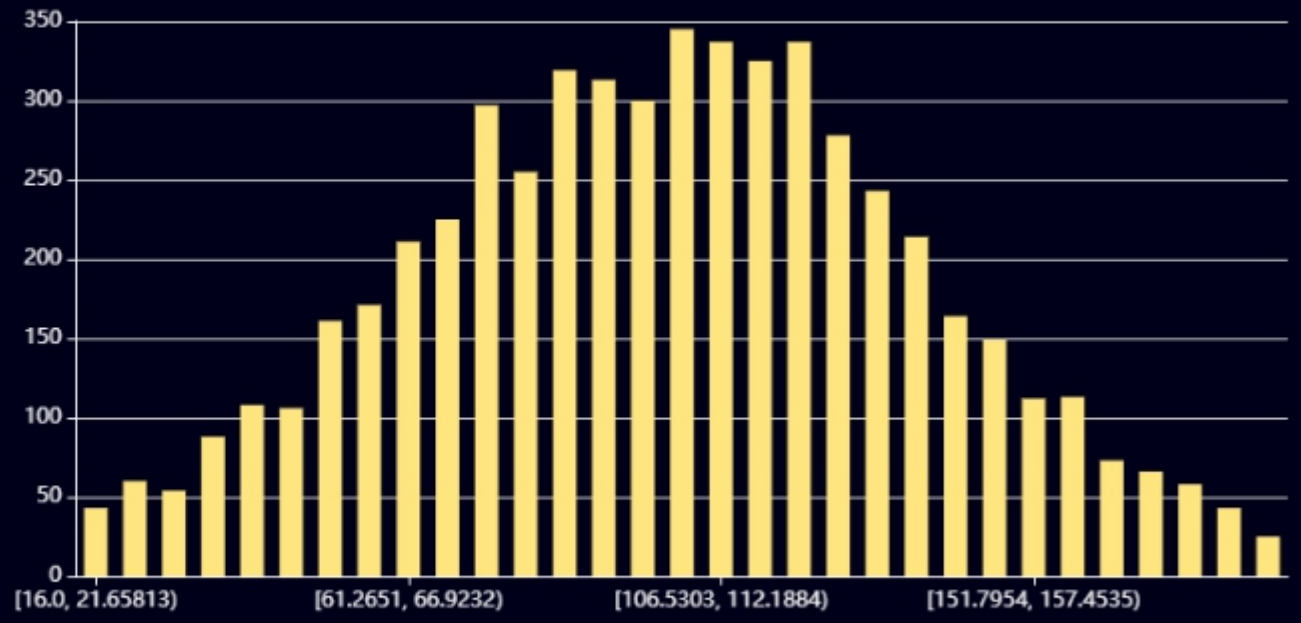


ForecastFlowを使って不要な特徴量を一枚表から削除

■特徴量の分布を確認

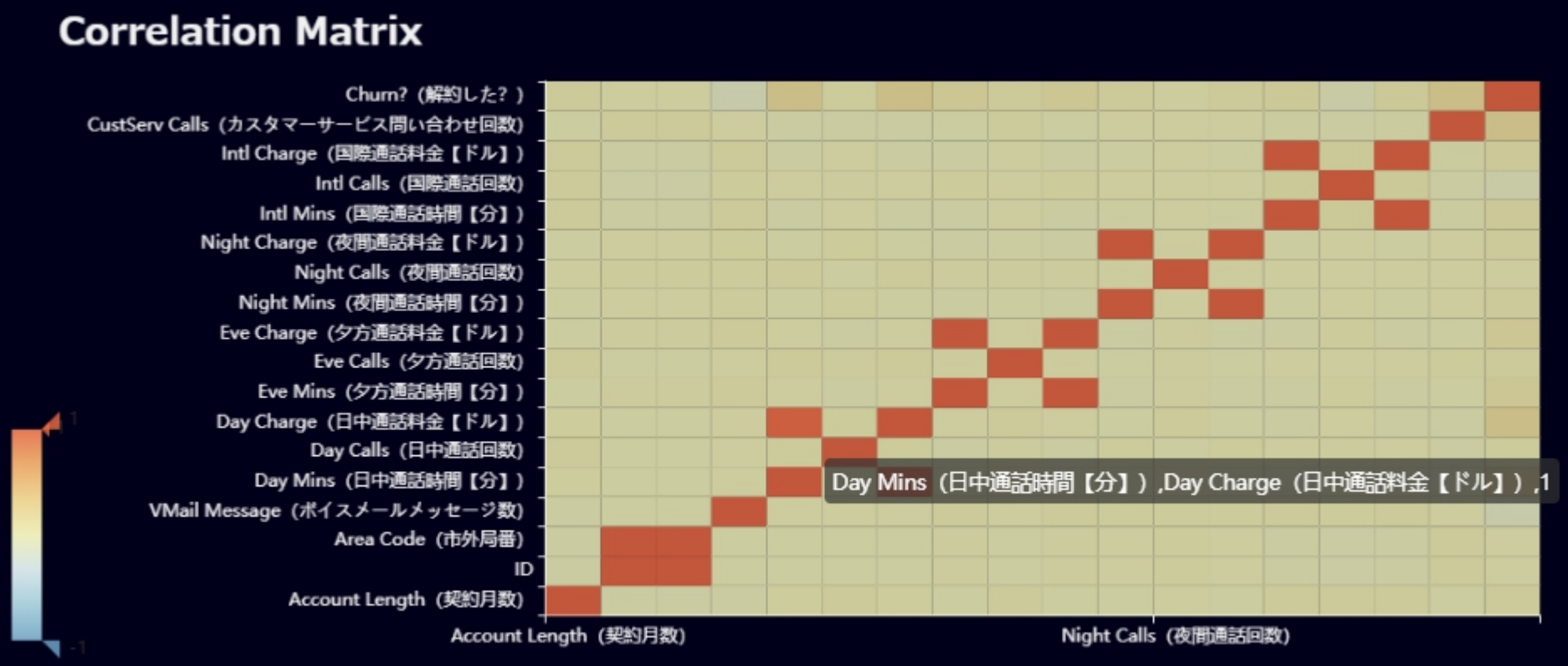
Column
Account Length

Type:	NUM
Is Unique:	false
P-Unique:	44.5%
P-Missing:	0%
Mean:	101.7
STD:	38.3
Min:	1
Max:	243
Plot Range:	[16, 191.4084]
Outliers:	107



ForecastFlowを使って不要な特徴量を一枚表から削除

■数値の特徴量間の相関を確認



ForecastFlowを使って不要な特徴量を一枚表から削除

■不要カラムを一枚表から削除する機能

3 Feature Columns

Select columns NOT including features

Possibly uninformative columns

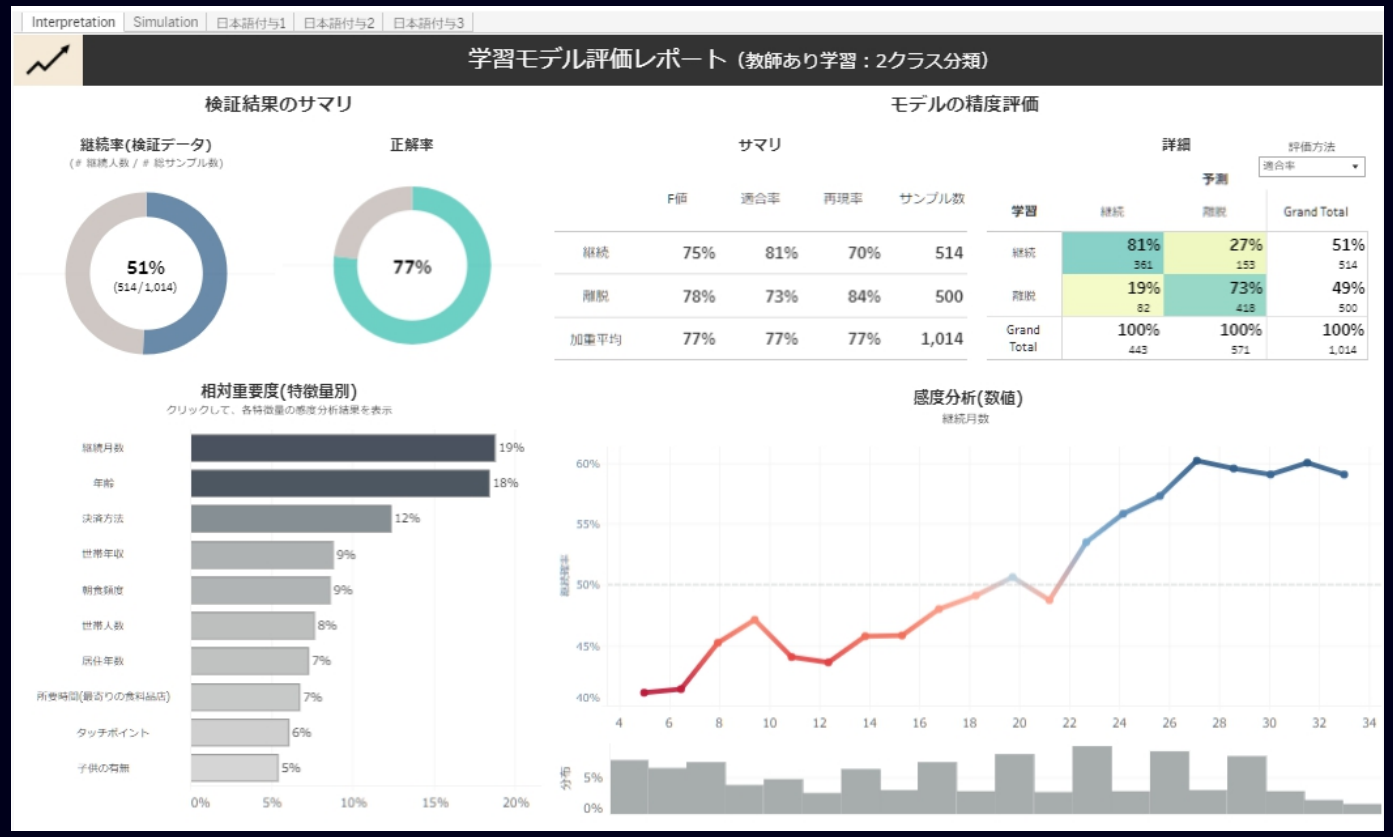
Column Name	Reason	Value
Day Mins（日中通話時間【分】）	High correlation with Day Charge（日中通話料金【ドル】）	1
Eve Mins（夕方通話時間【分】）	High correlation with Eve Charge（夕方通話料金【ドル】）	1
Intl Mins（国際通話時間【分】）	High correlation with Intl Charge（国際通話料金【ドル】）	1
Night Mins（夜間通話時間【分】）	High correlation with Night Charge（夜間通話料金【ドル】）	1

Rows per page: 5 1-4 of 4

©GRI Inc.

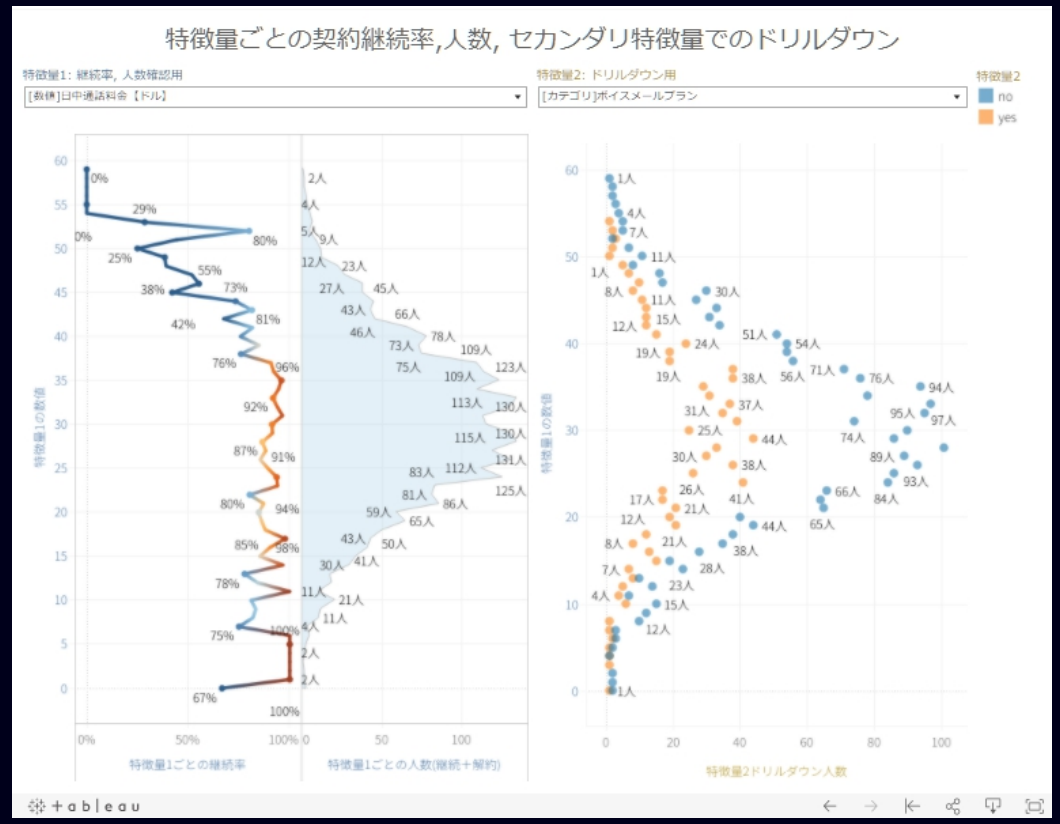
特徴量エンジニアリングでTableauとの違い

ターゲットと多くの特徴量の関係



https://public.tableau.com/profile/griinc6648#!/vizhome/MachineLearningInterpretation_JA/Interpretation

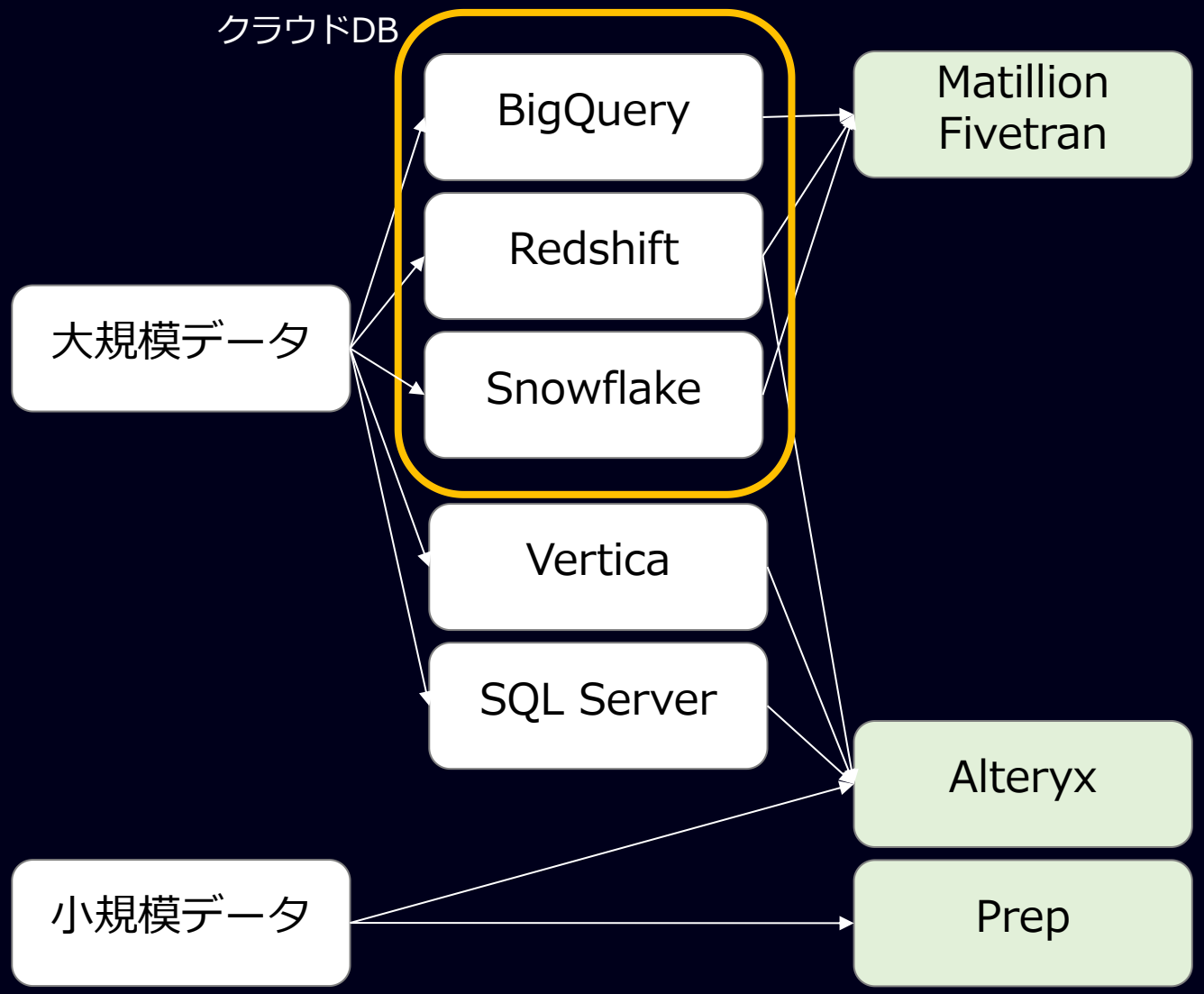
ターゲットと１つの特徴量の関係



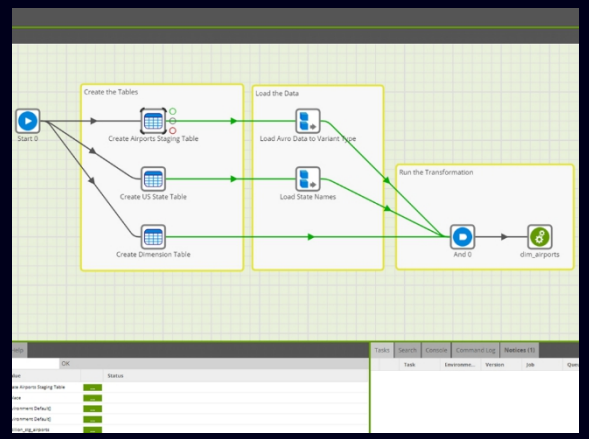
https://public.tableau.com/profile/griinc6648#!/vizhome/CustomerChurnTelephone_/sheet6

- 現場でよく知られている数値
- ドリルダウンして、背後にある関係性を確認（相互作用の特徴量を作るべきか？）

データ別のおすすめETL/ELTツール

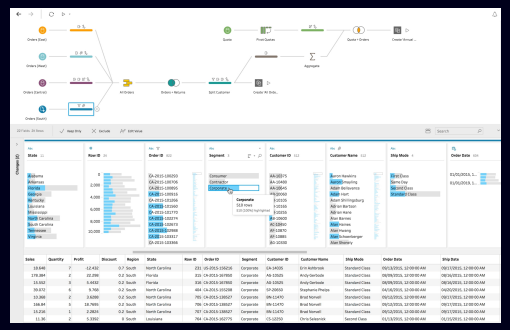


- 高速処理が基本
- 整形処理機能は最小限
- 接続先が分散していてもクラウドに集約しやすい



- 洗練されている
- In-DBで高速処理
- マクロ等
- 高コスト
- 接続先が分散されているとETLに高スペック必要

- Tableauにバンドルで低コスト
- 初心者にわかりやすい
- 大規模データの高速処理は限界



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
- 推論結果の使い方
- 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
- 特徴量エンジニアリング（時系列編）

©GRI Inc.

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
- 推論結果の使い方
- 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
- 特徴量エンジニアリング（時系列編）

©GRI Inc.

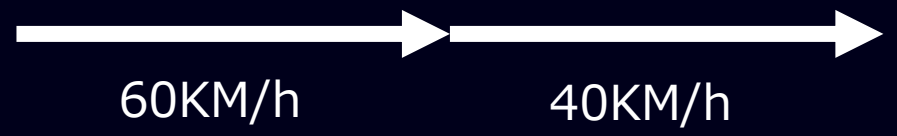
前回のF値の計算（修正）　※率同士の平均

F値：　適合率(precision）と再現率(recall)の平均（調和）

$$= \frac{2}{\text{Precision}^{-1} + \text{Recall}^{-1}}$$

※F値も率と率の平均なので、速度の平均と同じように逆数の平均で計算

■平均速度の計算方法の例



時速60KM
で2時間
= 120KM / 2h

時速40KM
で3時間
= 120KM / 3h

平均時速：　48KM/h　$= \frac{2}{(120/2)^{-1} + (120/3)^{-1}}$

©GRI Inc.

過去のすいすい会の動画

- 第１回「ビジネスでAIを活用するための問題設定」
  - https://www.youtube.com/watch?v=ppkgYD69NDU

- 第２回「アンケートと機械学習」
  - https://www.youtube.com/watch?v=-s1PcLQUBNI
  - https://www.youtube.com/watch?v=ylVT3OyayEk
  - https://www.youtube.com/watch?v=sFVabtfNvIk

- 第３回「分類問題の精度評価手法と不均衡データ」
  - https://www.youtube.com/watch?v=q6WJDTOgotA

©GRI Inc.

GRI出展：　Japan IT Week秋 AI・業務自動化展

- 10/28（水）～30（金）@幕張メッセ
  - すいすい会@AI・業務自動化展（10/28（水）12:00-13:00）
- 自社開発のセルフ型機械学習ツール「ForecastFlow」をぜひ体感して下さい！
- http://forecastflow.jp/event/japanitweek2020/

- ご来場特典として
  - お持ちいただいたデータを無料で解析！ハンズオン形式でレクチャー（要予約）
  - セルフ型機械学習ForecastFlowの1ヶ月無料体験
  - 機械学習超入門講座、講義動画を無料提供
  - GRI オリジナルグッズを進呈

©GRI Inc.