データサイエンス

すいすい会

第13回「自動機械学習での特徴量エンジニアリングのテクニカル解説」

GRI
BIGDATA SCIENCE
& RESEARCH

本日のアジェンダ

- 特徴量エンジニアリングのテクニカル解説

時刻データの扱い（時系列予測ではなく）
•時刻ではなく、リードタイムや期間に変換

欠損値の扱い
•データに欠損値がある時のシチュエーション
•データに欠損値がある時のForecastFlowの挙動（訓練時、推論時）
•データに欠損値がある時のデータ準備の手順

カテゴリデータの扱い
•カテゴリデータの持ち方（ワンホット・エンコーディングとの比較）
•カテゴリデータを括る
•数値をカテゴリに変換すべきか

©GRI Inc.

一枚表を自動機械学習に入れる

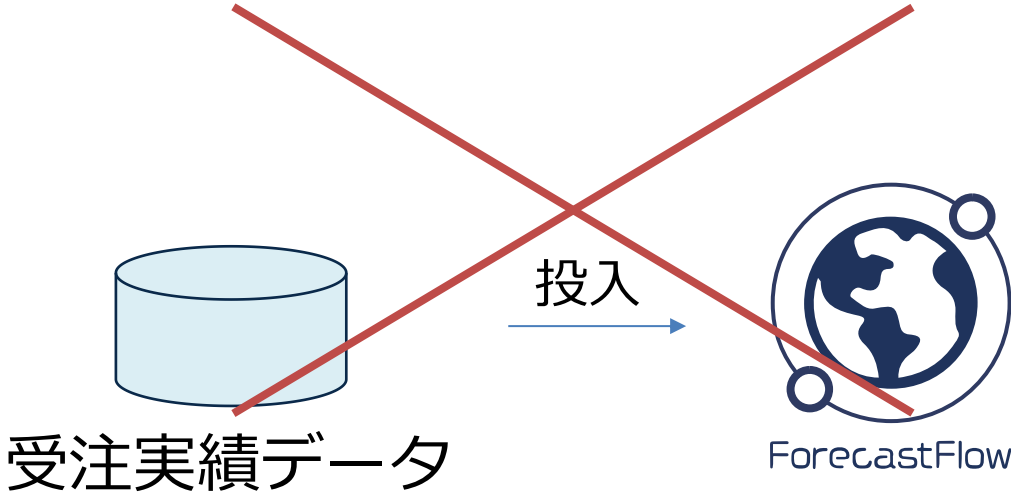
顧客ID　正解ラベル　特徴量

	A	B	C	D	E
1	ID	Churn?（解約した？）	STATE（州）	Account Length（契約月数）	Area Code（市外局
2	4153824657	False.	KS	128	415
3	4153717191	False.	OH	107	415
4	4153581921	False.	NJ	137	415
5	4083759999	False.	OH	84	408
6	4153306626	False.	OK	75	415
7	5103918027	False.	AL	118	510
8	5103559993	False.	MA	121	510

カテゴリ　数値　数値

よくある勘違い①

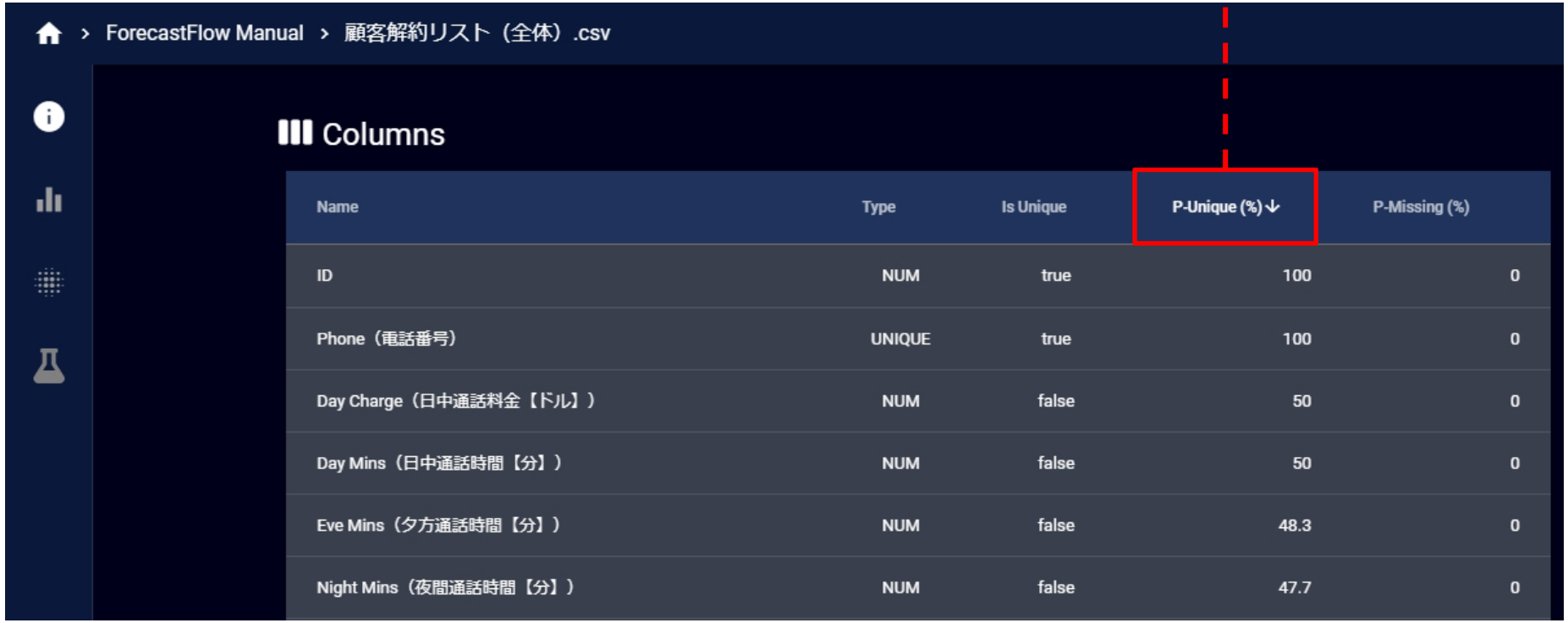
- データベースに入っているデータを入れれば、自動機械学習がよしなにやってくれる
  - 解釈しやすい特徴量を効率的に作る技術はまだない（どちらかを犠牲にする必要がある）



よくある勘違い②

- ユニーク性の高いデータを一枚表に準備してしまう
  - 日時に関するデータ
    - 生年月日
    - 初回受注日時
    - 最終受注日時

データ確認画面のP-Uniqueにて確認
ユニーク性の高さを確認



Name	Type	Is Unique	P-Unique (%) ↓	P-Missing (%)
ID	NUM	true	100	0
Phone（電話番号）	UNIQUE	true	100	0
Day Charge（日中通話料金【ドル】）	NUM	false	50	0
Day Mins（日中通話時間【分】）	NUM	false	50	0
Eve Mins（夕方通話時間【分】）	NUM	false	48.3	0
Night Mins（夜間通話時間【分】）	NUM	false	47.7	0

©GRI Inc.

ユニーク性の解消方法（生年月日）

- ４種類の回避方法

データ型

年齢
（数値）

年齢層
（カテゴリ）

X

基準日の設定

更新日基準
（受注日などを基準日にする）

学制年度基準
（４月２日を基準日にする）

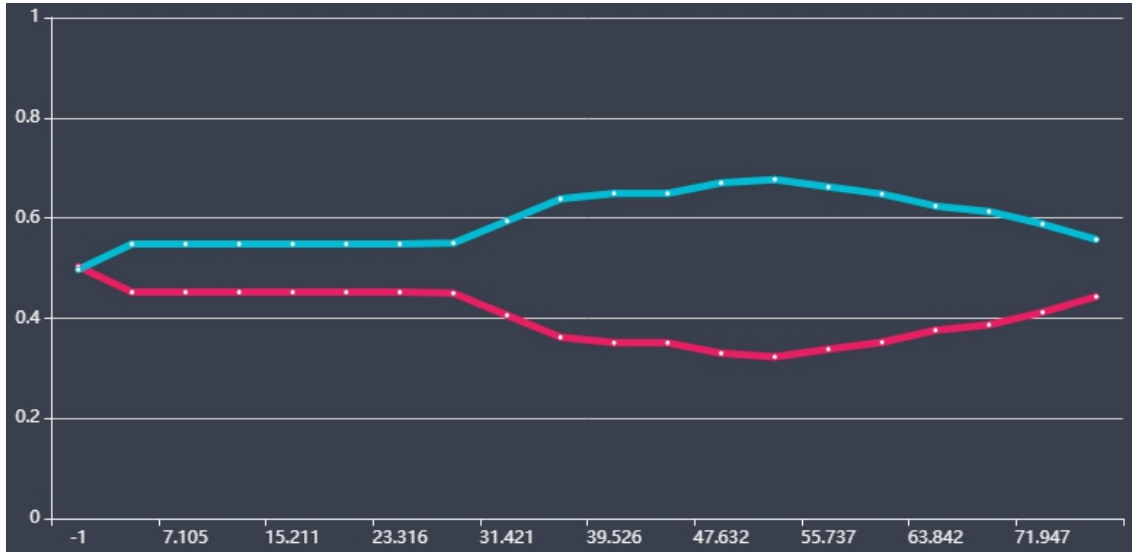
年齢　　= Datediff（Year, 生年月日, 基準日）
年齢層　= IF 年齢 …

©GRI Inc.

ユニーク性の解消方法（生年月日）

- 年齢（数値）と年齢層（カテゴリ）の違い
- 年齢（数値）の方が予測精度が高くなる傾向、ただしカテゴリの方が解釈しやすいことが多い（年齢層の区切り方）

予測確率



年齢（数値）

年齢層（カテゴリ）

予測確率

©GRI Inc.

ユニーク性の解消方法（初回受注日時など）

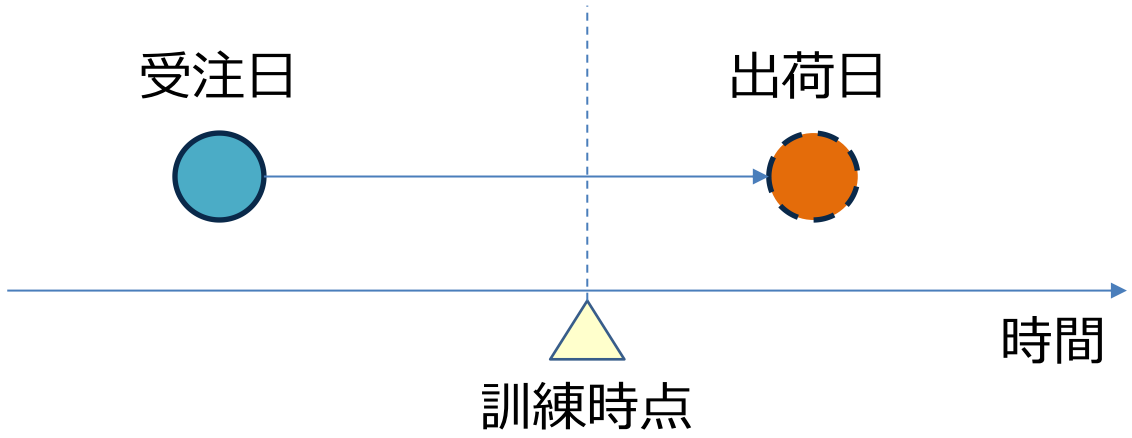
- 何を見たいかによって、変換方法が変わる

- サービスの古参度合いを表現（初期、２～３年目、それ以降）
- 開始日に依存しない月額課金でのお得感を表現（初回受注日時から月末日までの日数）
- ユーザの待ち時間を表現（受注日から出荷日までのリードタイム日数）

©GRI Inc.

訓練時点での不完備データ（出荷日が決まっていない場合など）

- ユーザの待ち時間を表現（受注日から出荷日までのリードタイム日数）
- 特に出荷が決まっていない時、欠損値（NULL）にすべきか？
- 出荷ミスとの違いを、どのように明示すべきか？



©GRI Inc.

ForecastFlowで特徴量が欠損値（NULLの値）の時の挙動

- ４種類のパターン

データ型

数値型

カテゴリ型

X

訓練 or 推論

訓練時の欠損

推論時の欠損

©GRI Inc.

数値型の欠損値（Null）の扱い【訓練時】

- ForecastFlow内部では、一度Nullを無視してモデル構築を行い、より精度が向上する方にNullを割り当てる

- 明示的にNullのデータの影響を知りたい場合、最小値と最大値の範囲外の値に置き換えたデータを準備する（ただし、あまりにも離れた値にすると、感度プロット（Sensitivity Plot）で見づらいので注意

©GRI Inc.

数値型の欠損値（Null）の扱い【推論時】

- 訓練時点で欠損値がなかったにも関わらず、推論時点で欠損値が出てきた場合、0として扱った上で予測スコアを算出する

©GRI Inc.

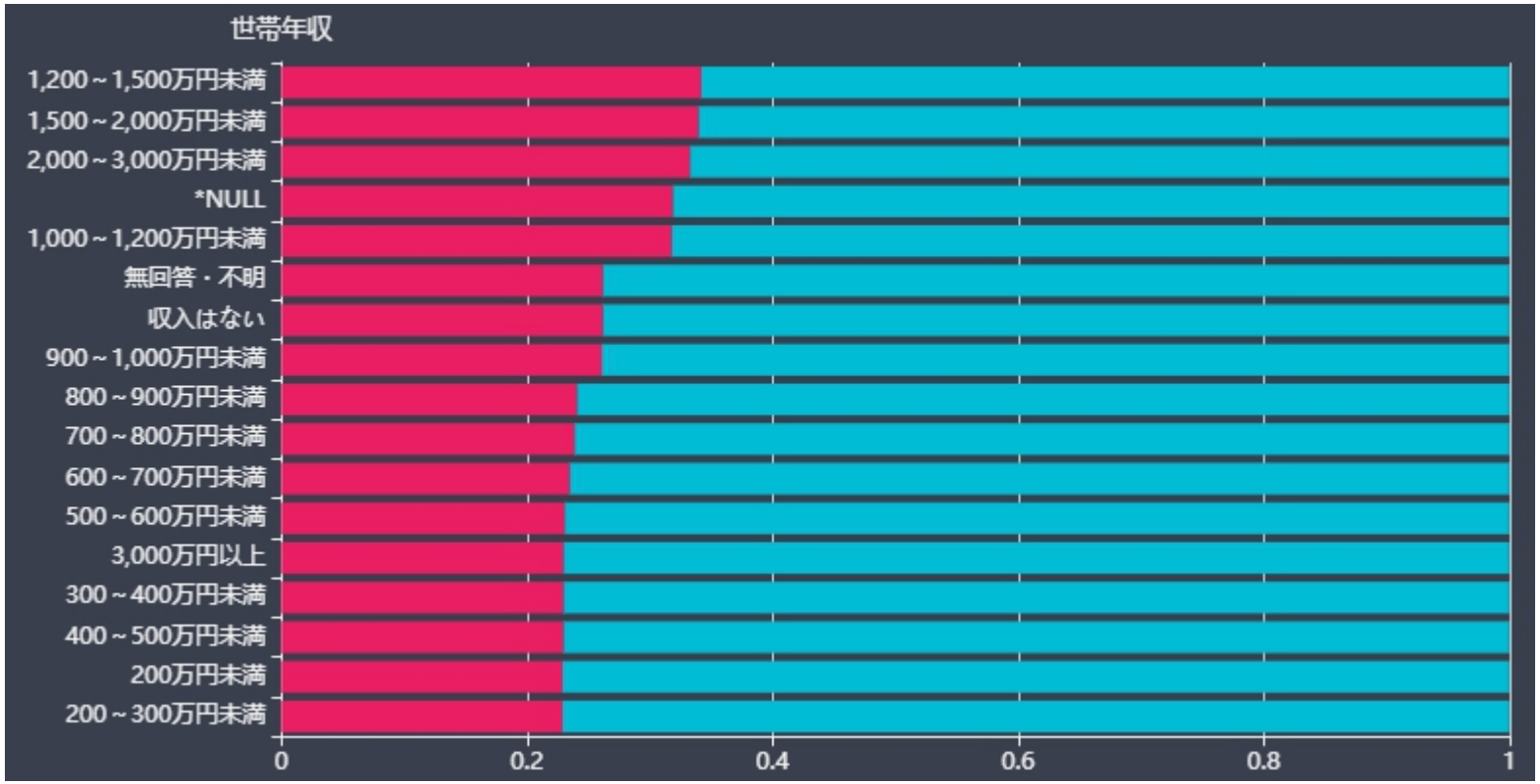
カテゴリ型の欠損値（Null）の扱い【訓練時、推論時】

- カテゴリ型の欠損値（Null）の場合、訓練でも推論では「*NULL」として扱われる

データの中の欠損値（ForecastFlowで「*NULL」の設定）

アンケート収集システムで検知された欠損値

アンケートで「収入はない」と回答



欠損値の入り方の典型例

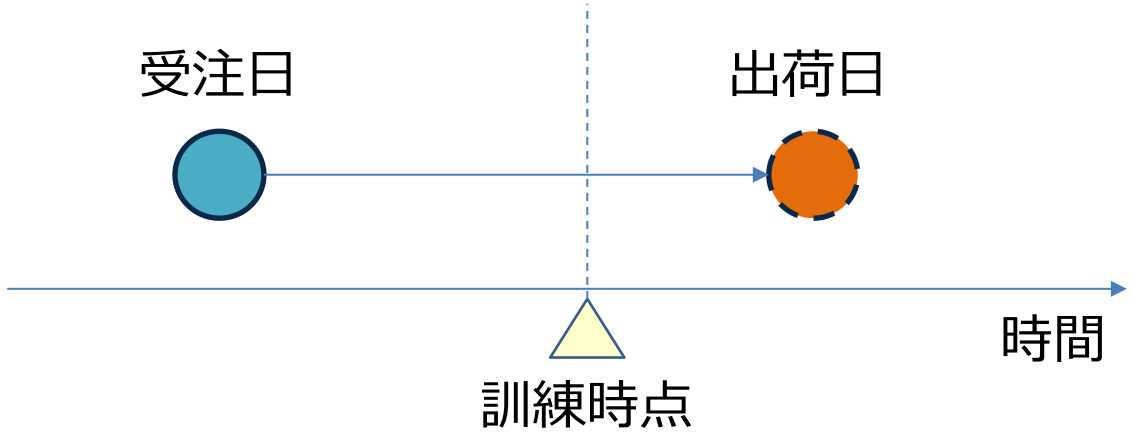
- システムの変更（過去には収集していなかった）
- ユーザが回答したくない項目（年齢などのプライバシー配慮）
- プロセス上の課題（出荷日など）
- ユーザ群自体が違う
- 機器の故障

©GRI Inc.

訓練時点での不完備データ（出荷日が決まっていない場合など）

- ユーザの待ち時間を表現（受注日から出荷日までのリードタイム日数）
- 特に出荷が決まっていない時、欠損値（NULL）にすべきか？
- 出荷ミスとの違いを、どのように明示すべきか？

- 方針１：　Nullのままにする
- 方針２：　平均的な出荷リードタイム日数を埋める（出荷ミスではない場合）
- 方針３：　-1 or 100などの値で埋める（出荷ミスの場合で単一値と集約値）
- 方針４：　行自体を削除する

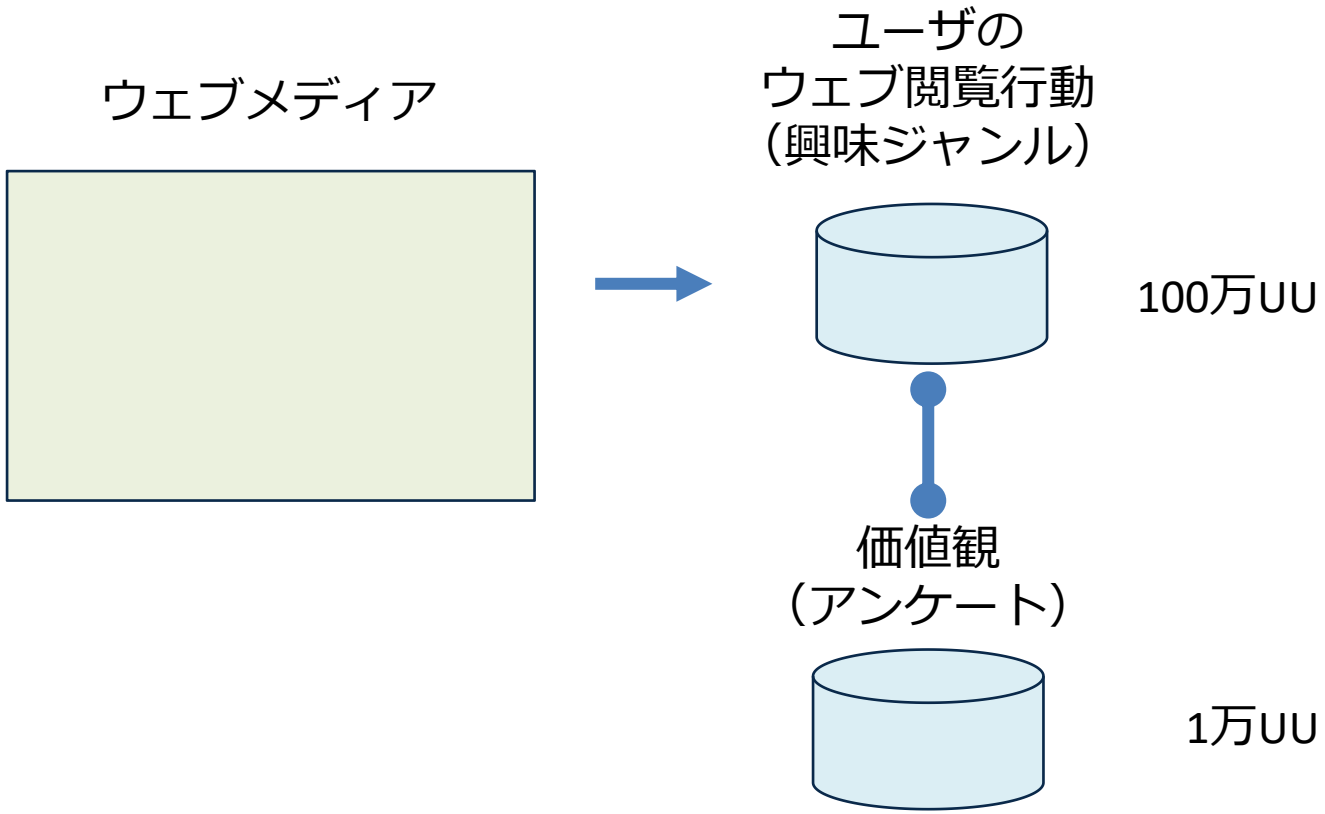


©GRI Inc.

モデル分割と不完備データ

- 欠損値の入り方がユーザの状況によって大きく変わる時、モデル自体を分割した方が良い時がある

メディア企業でのZero or 1st partyデータでのユーザ属性エンリッチ予測



◆モデルの分割
- 初回来訪ユーザ
- ライトユーザ
- ヘビーユーザ

©GRI Inc.

カテゴリデータの扱い

- カテゴリデータは１列で扱うべきか、One-hotエンコーディング（横持ち）すべきか？

１列のカテゴリ

ID	学歴
A001	中学卒
A002	高校卒
A003	大学卒
A004	大学院卒

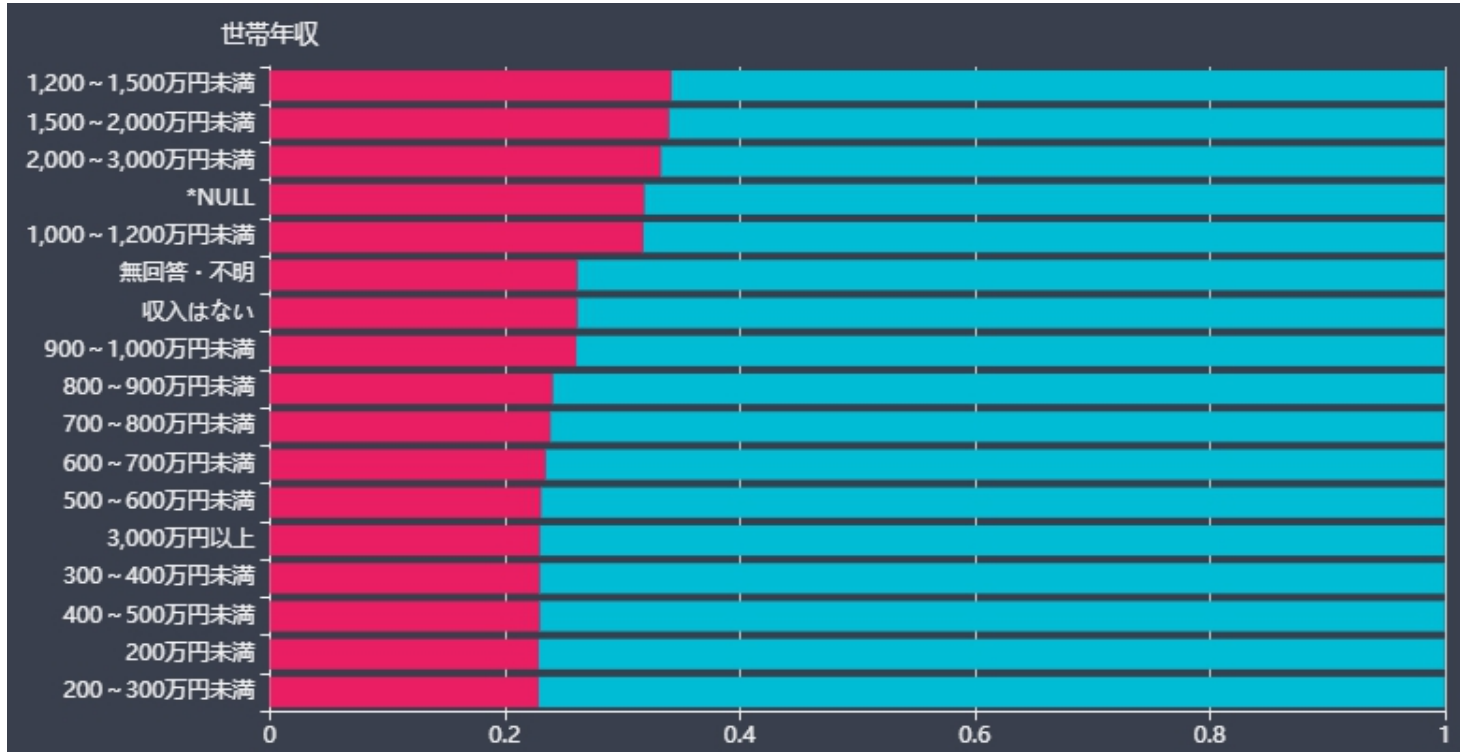
One-hot エンコーディング

ID	中学卒	高校卒	大学卒	大学院卒
A001	1	0	0	0
A002	0	1	0	0
A003	0	0	1	0
A004	0	0	0	1

©GRI Inc.

カテゴリデータの扱い

- どのような問題を扱っているのか？
- カテゴリの数の可能な部分集合のうち、予測精度が高くなる部分集合を決める
- 2^(k-1)-1の組み合わせ：　年収の例だと65,535個の組み合わせ

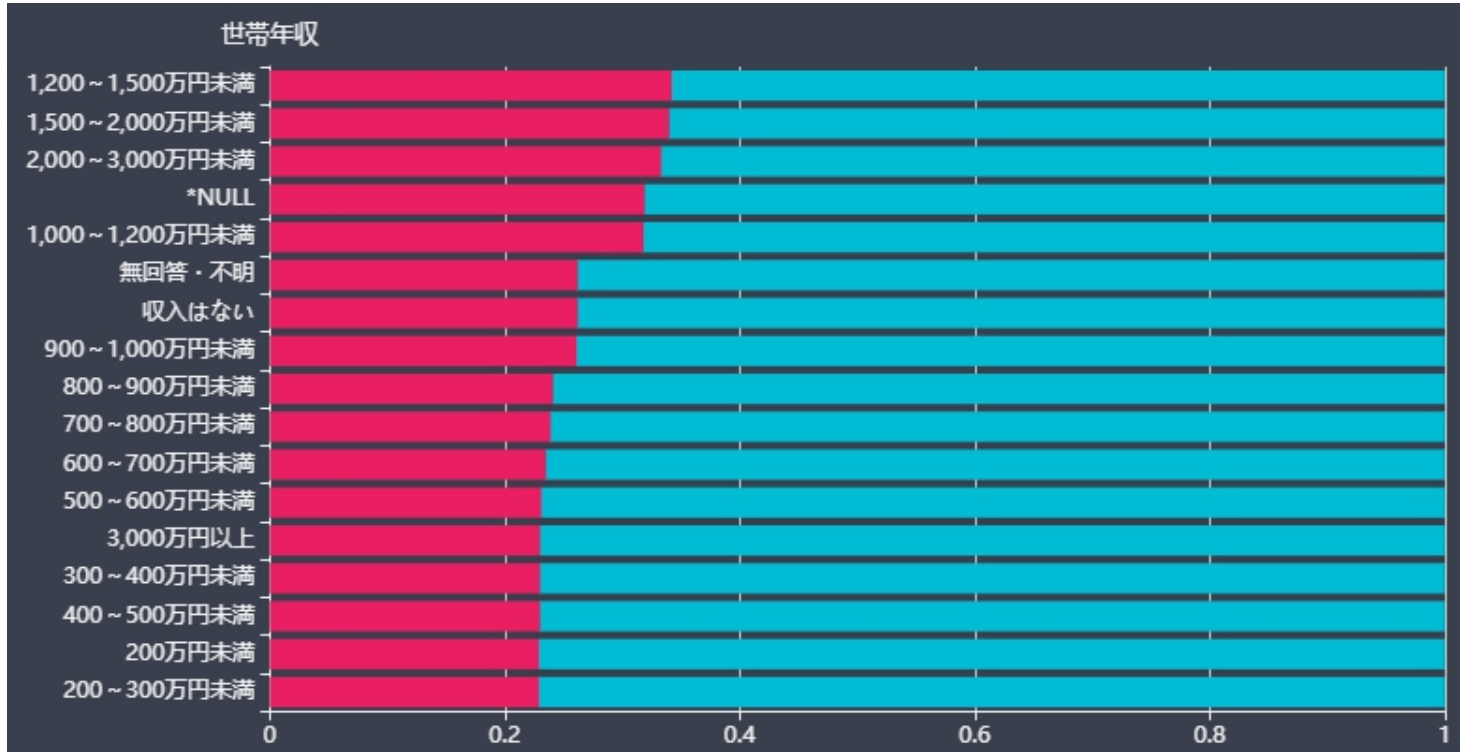


©GRI Inc.

カテゴリデータの扱い

- １列で扱った方が良い
  - Sensitivity Plot（感度分析）での解釈性が良い
  - Importance Plot（重要特徴量）のランクも少なくて済む
  - １列のデータの種類が多い時、予測精度が良い傾向のアルゴリズム

Fisher(1958) On Grouping for Maximum Homogeneity

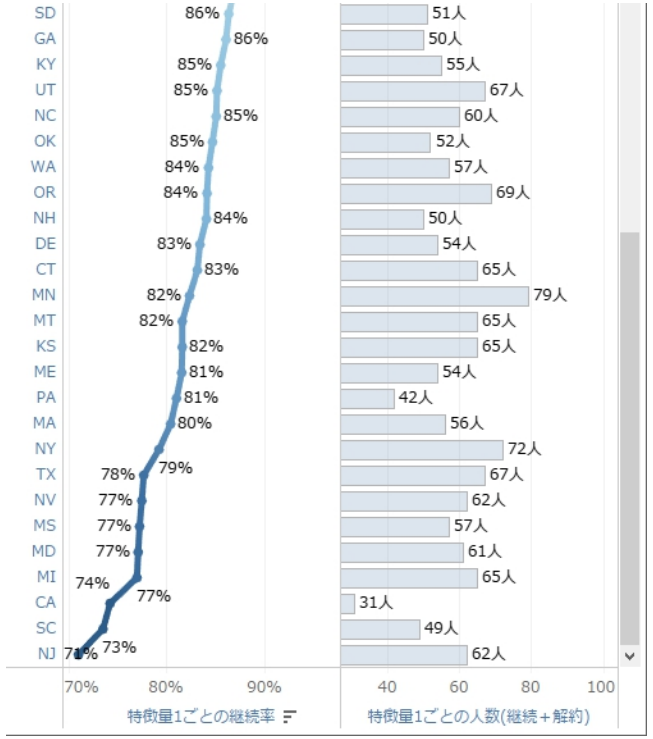


カテゴリデータの扱い

- １列で扱った方が良い
  - １列のデータの種類が多い時、予測精度が良い傾向のアルゴリズム

Fisher(1958) On Grouping for Maximum Homogeneity

  - ただし、該当数の少ないカテゴリラベルは、似ているもの同士でくくる



カテゴリデータの扱い

- カテゴリから数値にした方が良い時
  - カテゴリが順序性を持つとき、数値で持った方が精度が高くなる可能性がある
    - 悪い　→　1
    - ふつう　→　2
    - 良い　→　3

- カテゴリデータの最大の種類数：　2,147,483,647

©GRI Inc.

まとめ

■時刻データの扱い（時系列予測ではなく）

- 時刻ではなく、リードタイムや期間に変換

■欠損値の扱い

- データに欠損値がある時のシチュエーション
- データに欠損値がある時のForecastFlowの挙動（訓練時、推論時）
- データに欠損値がある時のデータ準備の手順

■カテゴリデータの扱い

- カテゴリデータの持ち方（ワンホット・エンコーディングとの比較）
- カテゴリデータを括る
- 数値をカテゴリに変換すべきか

©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
- 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

©GRI Inc.

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

©GRI Inc.