

津田塾大学総合政策学部総合政策学科 | データ政策科学

# イントロダクション

松本 崇斗(Takato Matsumoto)  
takato.matsumoto0114@gmail.com

## 質問について



- sil.doで受け付けます  
<https://www.sli.do/>
- Event code  
# SA202
- ミーティングツールで質問が可能であればそちらでも大丈夫です

# コンテンツ

1. Kaggle (Signate)について
2. Decision Tree (決定木)

# コンテンツ

1. Kaggle (Signate)について
2. Decision Tree (決定木)

# Kaggle (Signate)について

- Q. データ分析をどうやって学んだら良いかわからない
- Kaggleで実践をしながらデータ分析に必要なスキルを学習してみるのをお勧めします！
  - 今回はKaggleの日本版であるSignateを使ってみます
  - <https://signate.jp/>

# Kaggle (Signate)について

## ■ Kaggle

- データ分析コンペティションプラットフォーム
- <https://www.kaggle.com/>
- <https://ja.wikipedia.org/wiki/Kaggle>



スポンサー



データ分析の問題  
多くが予測モデルの作成



競技者  
予測モデルを作成



予測結果を提出

予測結果の採点

# Kaggle (Signate)について

## ■ Kaggleで学べること

- 数学
- 統計
- 機械学習の理論
- コーディング
- データの取り扱いなど

## ■ Kaggleで学びにくいこと

- データ収集の設計
- ビジネス理解

# Kaggle (Signate)について

- Kaggle (Signate)では,データサイエンティストが行う業務に必要な知識の多くを学ぶことができる
- コンペティションに参加しながら,Decision Tree (決定木)分析を学んでみましょう



# Kaggle (Signate)について | 国勢調査からの収入予測

## ■ 国勢調査からの収入予測

- <https://signate.jp/competitions/107/data>
- アメリカの国勢調査から収入が\$50Kを超えるかどうかを予測するタスク
- 2分類のクラス分類

## ■ `data_studies_policy/chapter_2/notebook/eda.ipynb` を開いてください

# コンテンツ

1. Kaggle (Signate)について
2. Decision Tree (決定木)

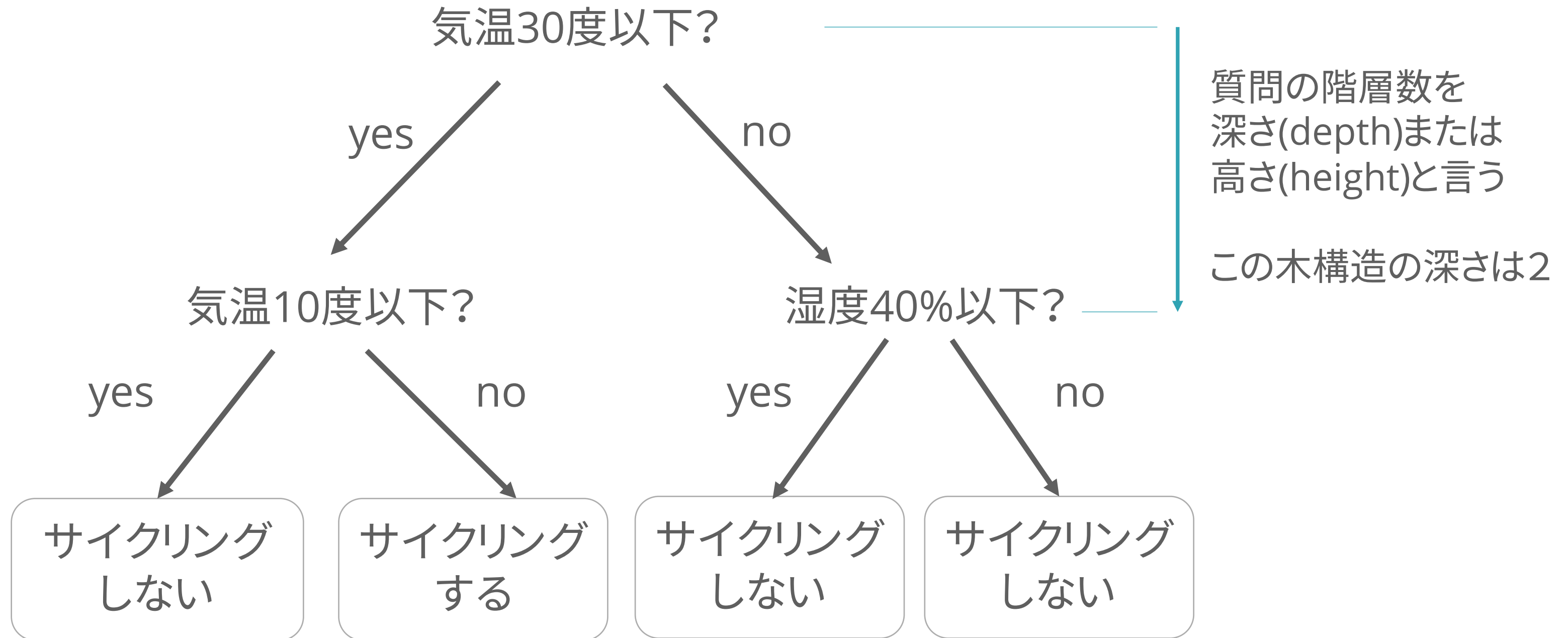
# Decision Tree (決定木)

## ■木構造を用いた予測モデル

- 解釈可能性が高い
  - なぜそのような結果になったかが分かりやすい
  - 説明責任が高いケースに用いられる
- 分類または回帰モデルがある

# Decision Tree (決定木)

## ■ 木構造とは



# Decision Tree (決定木)

## ■ 解釈可能性が高いためビジネスでよく利用される

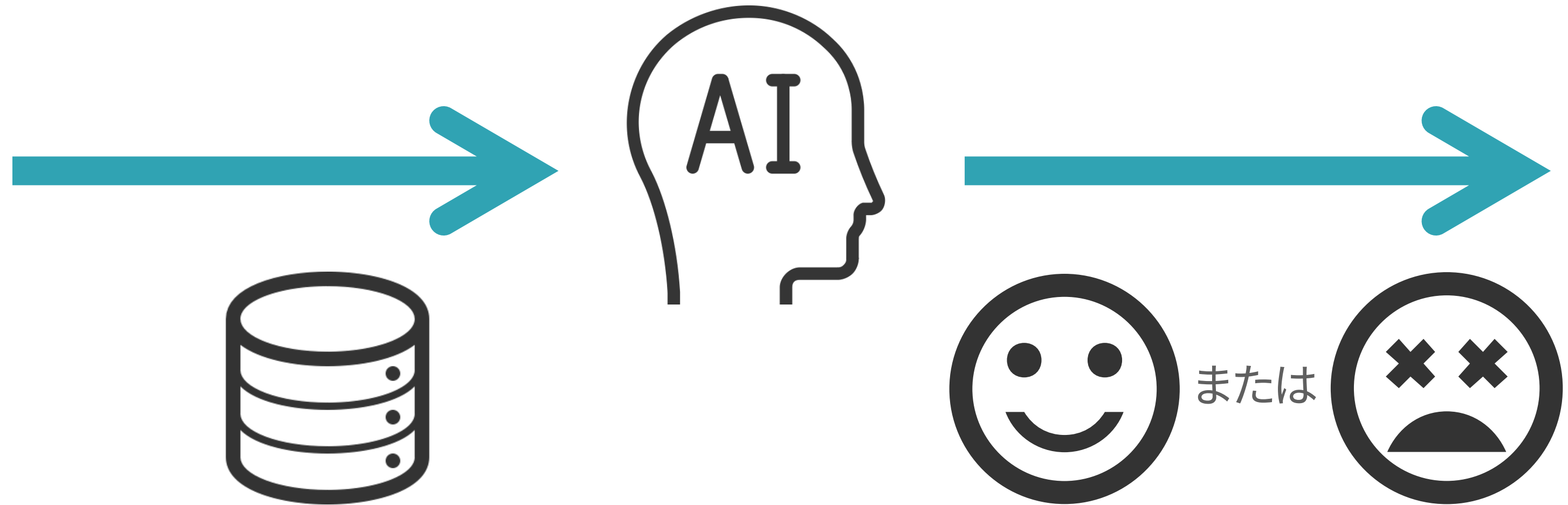
- 要因の把握ができる
  - 顧客の嗜好, 購入動機の実分析 (マーケティング)
  - 顧客の信用度の予測 (金融機関など)

## ■ 解釈可能性

- 予測モデルが, 予測を行うプロセスの理解のしやすさのこと
- Deep Learningなどの手法は, ブラックボックスであるため解釈可能性が低い

# Decision Tree (決定木) | 金融機関の例

予測モデル



顧客データ

収入, 年齢, 資産, 取引履歴

顧客の貸し倒れリスク

お金を貸しても大丈夫か否か

# Decision Tree (決定木) | 金融機関の例

解釈可能性が低い予測モデルを使った場合



# Decision Tree（決定木） | 金融機関の例

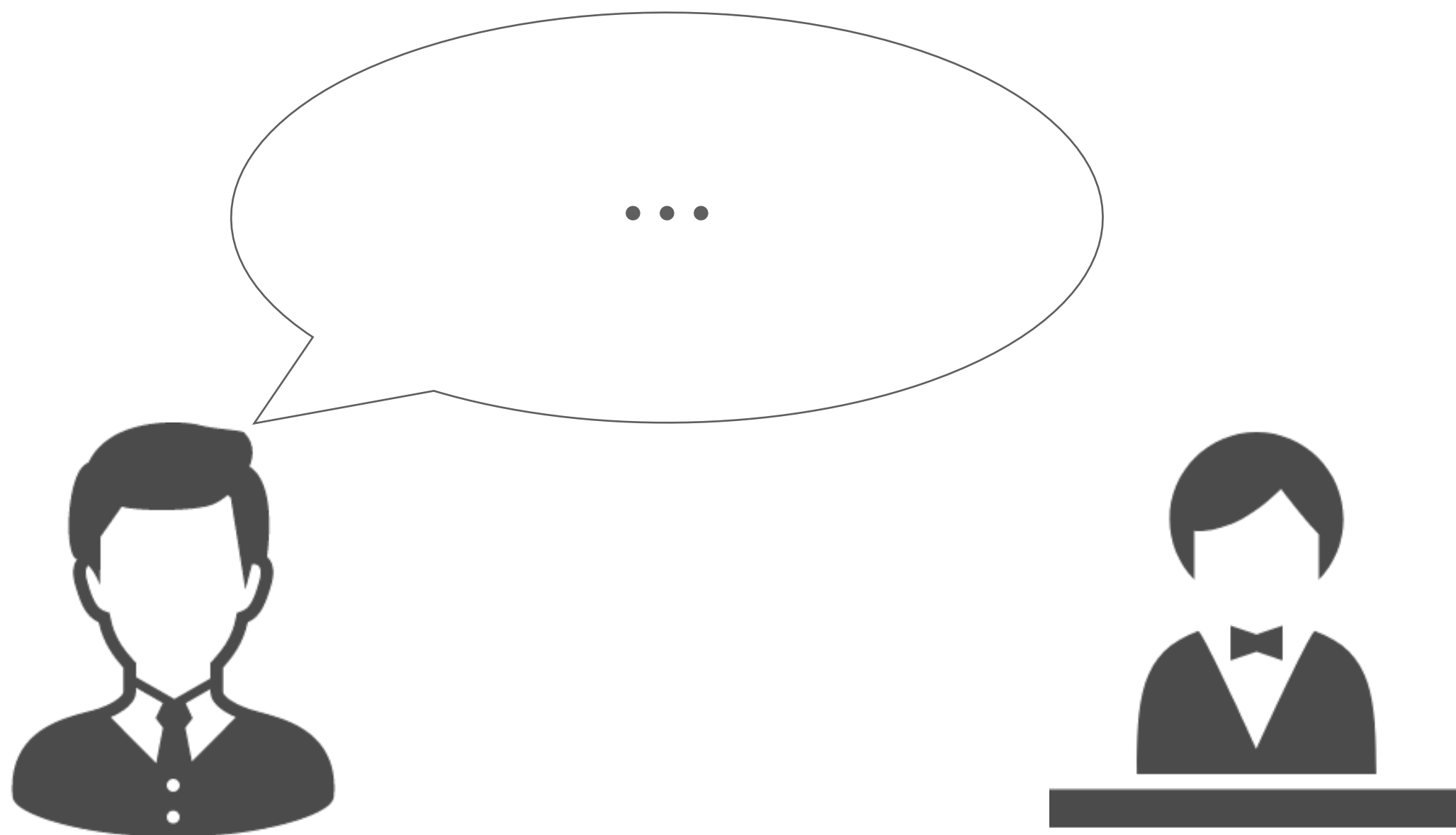
解釈可能性が低い予測モデルを使った場合





# Decision Tree（決定木） | 金融機関の例

解釈可能性が低い予測モデルを使った場合



# Decision Tree (決定木)

- なぜそうなったかといった説明責任が高いケースでは  
解釈可能性が高いモデルを使用することが推奨される
- 解釈可能性が高いモデルを使うことで, どう言った理由で  
顧客にお金を貸すことができないかを説明できる
  - 収入が, 希望の貸金に比べ低い
  - 過去に貸し倒れたことがある, など
- 目的変数に影響のある説明変数がわかるため  
原因の調査や改善策の立案に役立つ

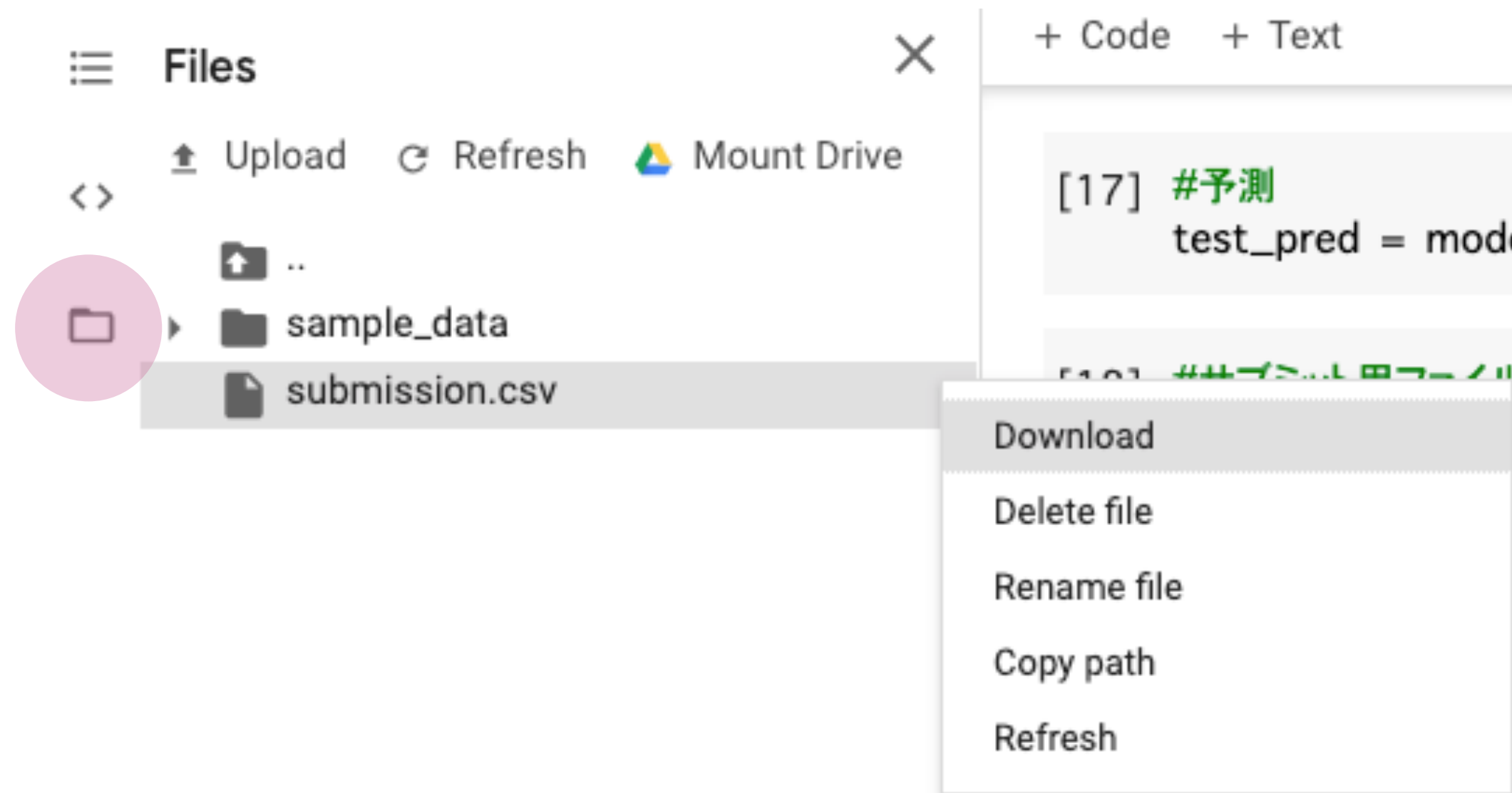
# Decision Tree (決定木)

- `data_studies_policy/chapter_2/notebook/modeling_dtr_1.ipynb`  
を開いてください

# Signateへ予測結果の提出

## ■ 実際にSignateに予測結果を投稿してみましょう

- data\_studies\_policy/chapter\_2/notebook/modeling\_dtr\_2.ipynbを実行します
- フォルダマークをクリックしてsubmission.csvをダウンロードしてください



# Signateへ予測結果の提出

## ■ 以下のページにアクセスして、投稿ボタンをクリック

- <https://signate.jp/competitions/107/data>



# Signateへ予測結果の提出

## ■ ファイルを選択で先ほどのsubmission.csvを選択し投稿できます

投稿

---

投稿ファイル※ (投稿可能なファイルの拡張子は csv です)

ファイルを選択

ファイル未選択

---

メモ (400字以内)

> 投稿

\*1日10回まで投稿が可能です。  
投稿した結果は登録されたメールに送信されます。

---

キャンセル

# Signateへ予測結果の提出

- 投稿すると自分のスコアとランキングが表示されます
- モデルの説明変数を工夫してハイスコアを目指してみてください

274	<a href="#">hakatamentaiko</a>		0.8405503	7
275	<a href="#">Kob</a>		0.8392605	3
276	<a href="#">en1968</a>		0.8385849	6
277	<a href="#">Takato</a>		0.8357595	3
278	<a href="#">dev-ict</a>		0.8347767	5
279	<a href="#">k.kubo</a>		0.8347767	2
280	<a href="#">k.yaguchi</a>		0.8347767	2

- データを用いて、所得格差を是正する政策を立案してください
  - 何が所得格差の問題となっているか調査
  - A4レポート一枚