



Vertex Explainable AI で 脱ブラックボックス

機械学習モデルをより良く理解するには？

葛木 美紀

Google Cloud カスタマー エンジニア

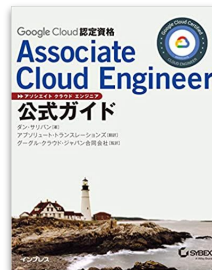
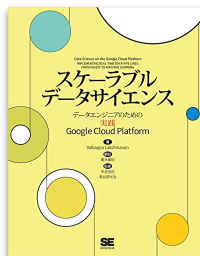
スピーカー自己紹介



Miki Katsuragi
Google Cloud
カスタマー エンジニア
Twitter: myoshimu@

兵庫県出身で二児の母。データベース ベンダーでアナリスト、データ分析基盤の構築や運用を経て、Google で統計や機械学習によるデジタル広告や CM の広告効果測定などの分析業務に従事。現在は Google Cloud でエンジニアとして ML/AI を活用したサービスの開発やデータ分析の提案を担当。
GCPUG 女子会オーガナイザー

著書、レビュー、監修





機械学習における説明性とは

AI はブラックボックス？

企業におけるAI の価値に障害となる要素

わからないことへの不安

79%

出発点がわからない: 63%

間違ったベンダー戦略: 48%

企業での利用が少ない: 40%



さまざまなチームで重要な AI の説明性

知りたい
こと

データサイエンティスト
と ML エンジニア

モデルがうまく
機能しない理由

改善方法

ML システムの
エンドユーザー

予測結果はどれくらい
信頼できるか

予測結果をどのように
使用できるか

監督・
コンプライアンス担当

モデルは安全で
目的に適合しているか

規制への準拠

モデルと予測の背後にある「理由」を明らかに



堅牢で実用的な説明

特徴アトリビューション : モデル全体および特定の予測結果について特徴量の重要性を表示



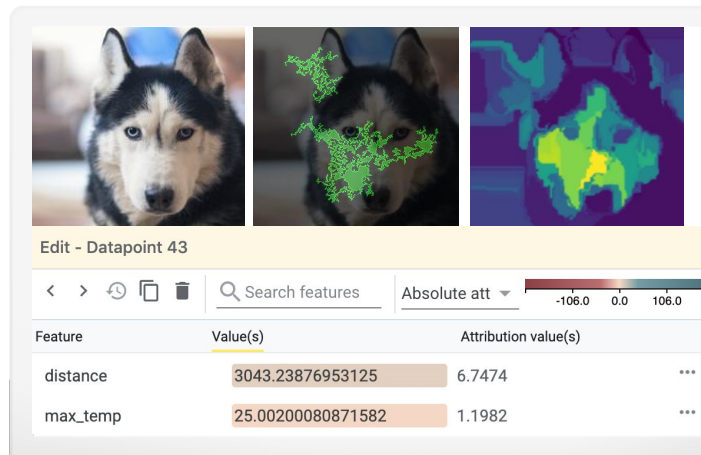
複数のAIプラットフォームサービスに対応

Vertex AI Prediction、AutoML、Workbench



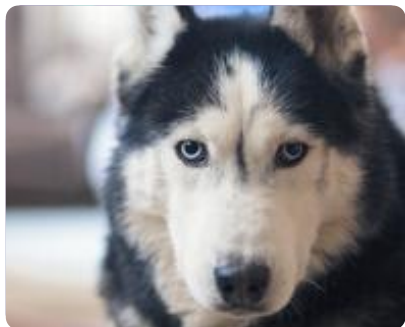
柔軟、高速、スケーラブル

複数の ML フレームワーク、オンラインおよびバッチ処理のユースケースからの表形式、画像、テキストモデルをサポート
フルマネージド、サーバーレスで、高速に処理



さまざまなデータ形式に対応

画像



ML タスク:

画像分類

表形式

Feature name	Feature value
start_hr	18
weekday	1
distance	1395.51
temp	16.168
dew_point	7.83396
wdsp	0
max_temp	20.7239
prcp	0.03
rain_drizzle	0
duration	11

分類 / 回帰

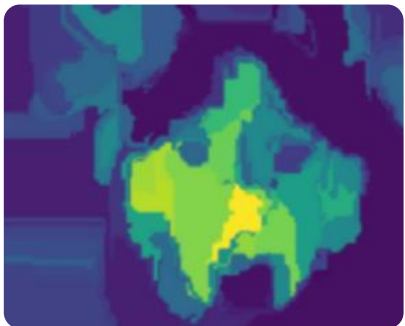
文章

The cake tastes
delicious!

テキスト分類

さまざまなデータ形式に対応

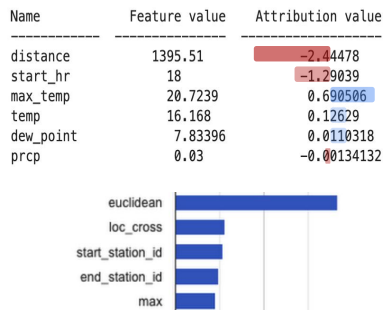
画像



わかること

モデルの分類に最も貢献した
画像ピクセルまたは領域

表形式



各特徴量が、単一の予測結果または
モデル全体にどの程度貢献したか

テキスト

The cake tastes
delicious!

Sentiment score: 0.9

各単語または文節が、
テキスト分類にどの程度貢献したか

Explainable AI の特徴

1

堅牢性

確立された研究に基づく
3つの説明可能性*

- [Sampled Shapley](#)
- [Integrated Gradients](#)
- [XRAI](#)

データサイエンティストと
エンドユーザーにとって
直感的

* 参考 [AI Explainability Whitepaper](#)

2

柔軟さ

複数のモデルタイプを
サポート

- 表形式の分類、回帰
- 画像分類
- テキスト分類

オンラインとバッチ処理に
対応

ML フレームワークに依存しない:
カスタム コンテナとしてデ
プロイされたすべてのモデルと
互換性があります

3

シームレスな連携

XAI 対応製品:

- AutoML Tables
- Vertex Prediction
- Vertex Notebooks

今後の対応予定

- AutoML Vision
- BQML
- Continuous Monitoring
- Others...

4

使いやすさ& スケール

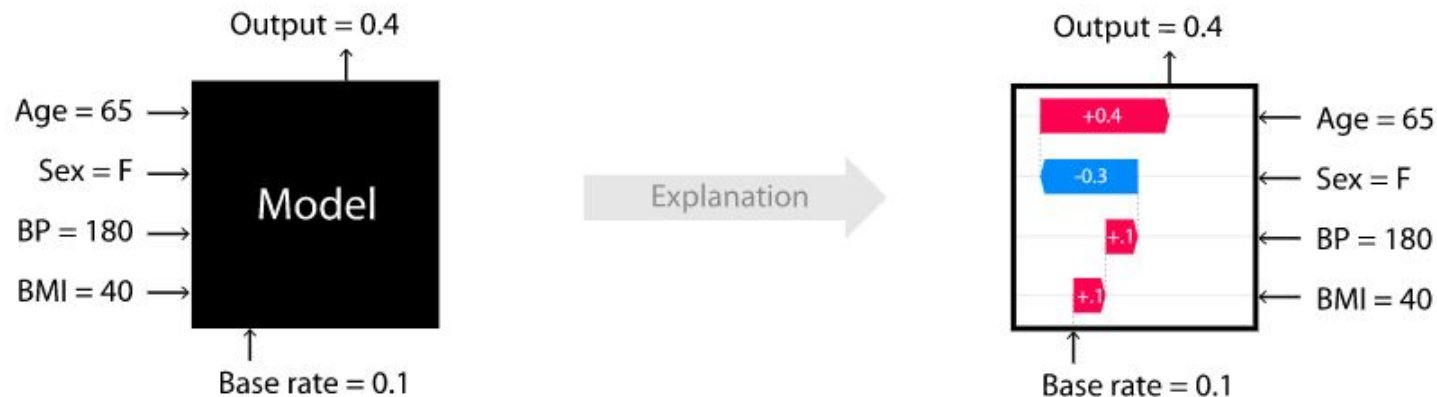
SDK により迅速な
セットアップが可能

マネージド・サーバーレス
サービス

OSS パッケージよりも
大幅に高速な処理性能

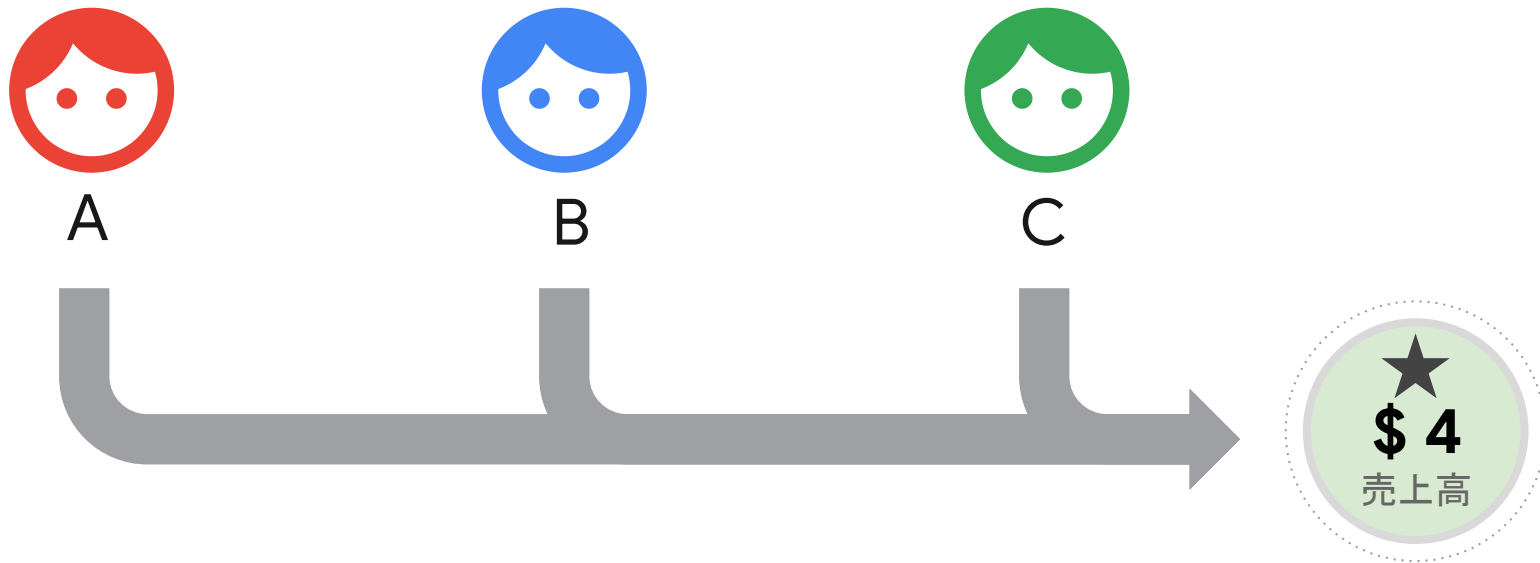
Shapley Value とは

- SHAP (SHapley Additive exPlanations) : 機械学習モデルの出力を説明するためのゲーム理論的アプローチ
- ゲーム理論の古典的なシャープレイ値を利用して、最適な貢献度割り当てを計算

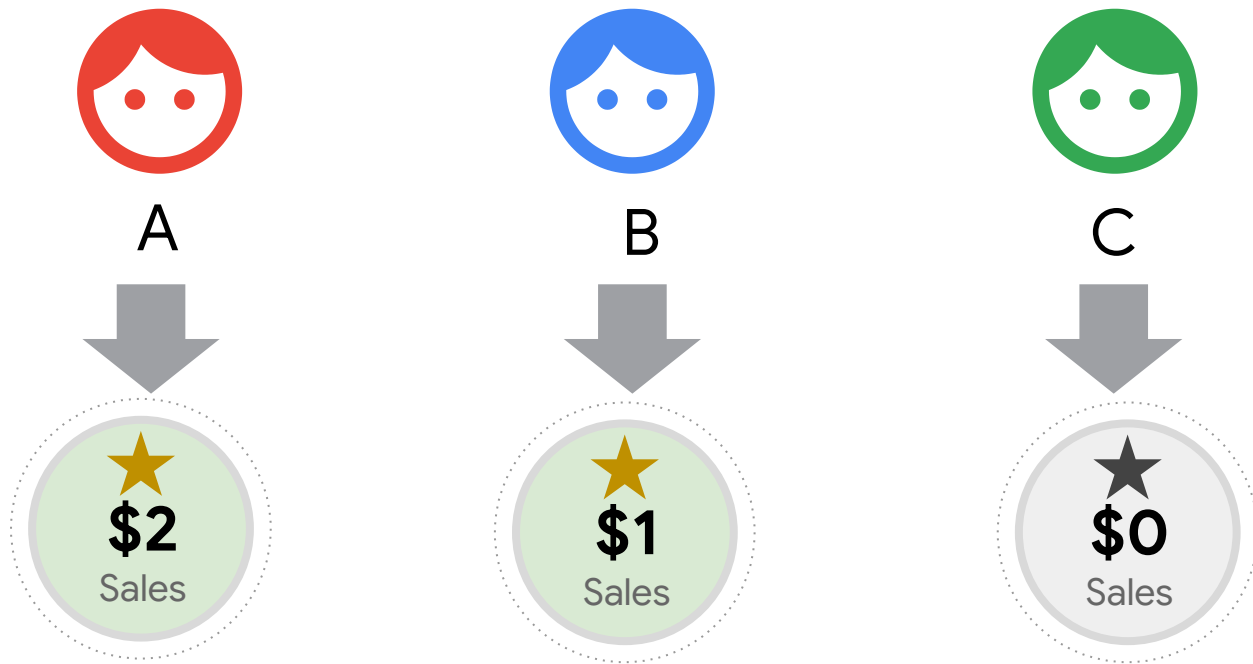


Shapley Value とは

例: 営業チーム全体で \$4 の売上があった場合、各メンバーにどのように報酬を分配するか？



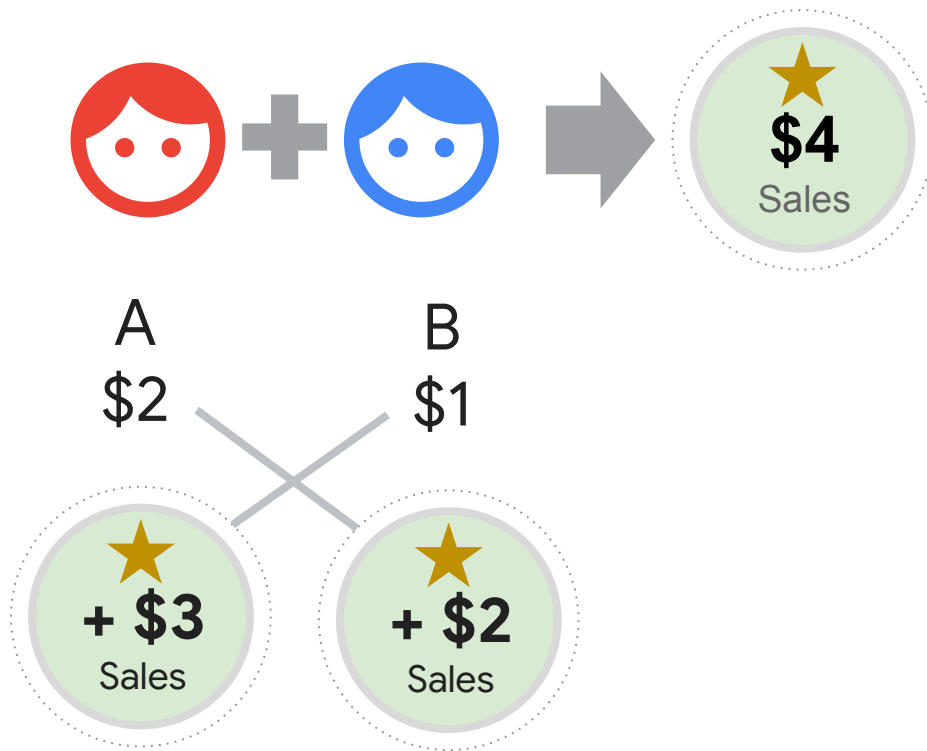
各個人単体のパフォーマンス















チームの一部としての個人パフォーマンス

増分貢献 (Counterfactual Gain)

- A の増分貢献 = $\$4 - \$1 = \$3$
売上合計から B 個人の
達成金額を差し引いた金額
- B の増分貢献: $\$4 - \$2 = \$2$
売上合計から A 単体の
貢献を差し引いた金額



チームの一部としてのパフォーマンスを Matrix に

	A	B	C	AC	BC	BA	ABC
				 	 	 	  
売上合計	\$2	\$1	\$0	\$2	\$1	\$4	\$4

Counterfactual Gain



A



B



C

\$2	-	-	\$2	-	\$3	\$3
-	\$1	-	-	\$1	\$1	\$2
-	-	\$0	\$0	\$0	-	\$0

AC からの増分

Shapley 値の計算方法

出現順序に貢献度をマッピング

	A の貢献度	B の貢献度	C の貢献度	合計
A > B > C	2	2	0	4
A > C > B	2	2	0	4
B > A > C	3	1	0	4
B > C > A	3	1	0	4
C > A > B	2	2	0	4
C > B > A	3	1	0	4
Shapley Value (平均)	2.5	1.5	0	4

Shap vs. Integrated Gradients (統合勾配)

サンプリングされた Shapley

メリット

- 任意の入力で動作可能
- 入力値の貢献度を計算
- **表形式のモデルにフィット**

デメリット

- **計算上より高価**
- **画像モデルには適していない**

統合勾配 + XRAI

メリット

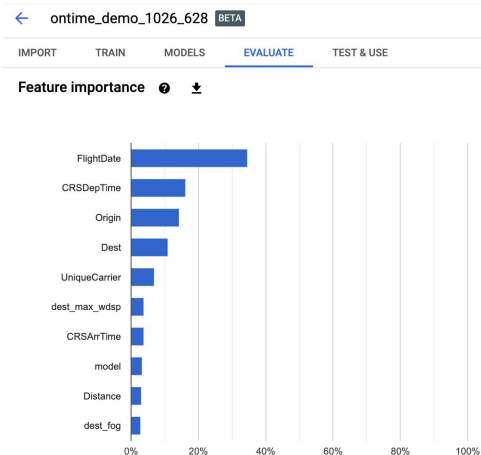
- **計算がより効率的**
- **画像モデルでうまく機能**
- 視覚化のためのいくつかの拡張機能

デメリット

- 勾配爆発 (勾配が無限值になってしまう)
- 微分不可能なモデルでは勾配が計算できない
ので機能しない

表形式データ向け AutoML

各特徴量の重要性を理解する
モデルの全体的な予測力(「グローバル」特徴量の重要性)



特定の特徴量の重要度を把握
(「ローカル」特徴量の重要性)

Feature column name ↑	Data type	Value	Local feature importance ⓘ
CRSArrTime	Numeric	1045	0.007
CRSDepTime	Numeric	720	0.026
Dest	Categorical	RSW	0.038
dest_fog	Categorical	0	0.000
dest_max_wdsp	Numeric	8.9	0.006
Distance	Numeric	979	-0.008
FlightDate	Timestamp	2003-10-17	-0.070



ユースケース

AutoML による白内障手術プロセスの改善

AutoMLを使用して、白内障手術に必要な時間の長さを予測するモデル作成

結果

白内障手術時間の予測を 33%向上

利点

- スタッフと部屋の可用性を最適化し手術をより効率的にスケジューリング
- リソースに制約のある環境でコスト削減

ブログ

: <https://cloud.google.com/blog/topics/customer-s/how-moorfields-is-using-automl-to-enable-clinicians-to-develop-machine-learning-solutions>



参考資料

Getting started



cloud.google.com/explainable-ai

Blog: Introducing Explainable AI



bit.ly/introducing-xai

Docs: AI Platform



bit.ly/xai-docs

Sample code: Explainable AI notebooks



bit.ly/xai-code

Whitepaper



bit.ly/xai-whitepaper

TabNet: Train interpretable-by-design tabular models



bit.ly/tabnet-quickstart

What-If Tool: Analyze models within a notebook



pair-code.github.io/what-if-tool

Thank you.

