

Comprendre son modèle de Machine Learning grâce à l'Explicabilité

Manon MARTIN



Crédit Agricole Titres

- Spécialiste des traitements financiers
- Gère les plateformes pour les caisses régionales
- \approx 1000 employé.es sur 3 sites en France

Contexte

- IA Act
- Modèles à améliorer
 - Cas sur les transferts PEA

Explicabilité



Favorise la compréhension des
prédictions des modèles



Favorise une adoption plus confiante
et éthique de l'intelligence artificielle

Modèles interprétables

Modèles intrinsèquement compréhensibles

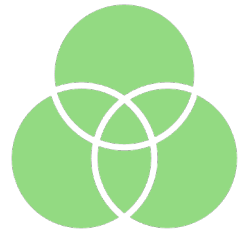
Régression linéaire

Arbre de décision



Performances moindres

Méthodes model-agnostic ou model-specific



Méthodes model-agnostic

Travaille sur les paires d'entrée/sortie



Méthodes model-specific

Travaille avec les paramètres internes du modèle

Méthodes globales ou locales



Méthodes globales

Vue d'ensemble du modèle



Méthodes locales

Cible une instance en particulier

Dataset pour le TP – German Credit Risk

- Age (numérique) : L'âge des individus en années
- Sexe (texte) : Le sexe des individus
- Job (numérique) : Niveau de compétence et type de résidence des individus
- Housing (texte) : Type de logement des individus
- Saving accounts (texte) : Niveau des comptes d'épargne
- Checking account (numérique, en DM) : Montant des comptes courants, exprimé en Deutsche Mark (DM)
- Credit amount (numérique, en DM) : Montant du crédit, exprimé en Deutsche Mark (
- Duration (numérique, en mois) : Durée du crédit en mois
- Purpose (texte) : Objet du crédit
- Risk (texte) : cible (bad or good)

SHAP (Shapley Additive exPlanations)

Utilise les valeurs de Shapley



Calcule la contribution marginale de chaque caractéristique à chaque prédiction

Théorie des jeux – Valeurs de Shapley

First	Second	Third
\$10,000	\$7,500	\$5,000



\$10,000

Théorie des jeux – Valeurs de Shapley

First	Second	Third
\$10,000	\$7,500	\$5,000



Joueur 1



Joueur 2



$C^{12} = \$10,000$

C^{12} est la valeur de la coalition entre le joueur 1 et le joueur 2

Théorie des jeux – Valeurs de Shapley

First	Second	Third
\$10,000	\$7,500	\$5,000



Joueur 1



Joueur 2

Valeurs de coalition :

$$C_{12} = 10\,000$$

$$C_1 = 7\,500$$

$$C_2 = 5\,000$$

$$C_0 = 0$$

Théorie des jeux – Valeurs de Shapley

Contribution marginale du Joueur 1 :

$$C_{12} - C_2 = 5\,000$$

$$C_1 - C_0 = 7\,500$$

$$(5\,000 + 7\,500) / 2 = \$6\,250$$

Valeurs de coalition :

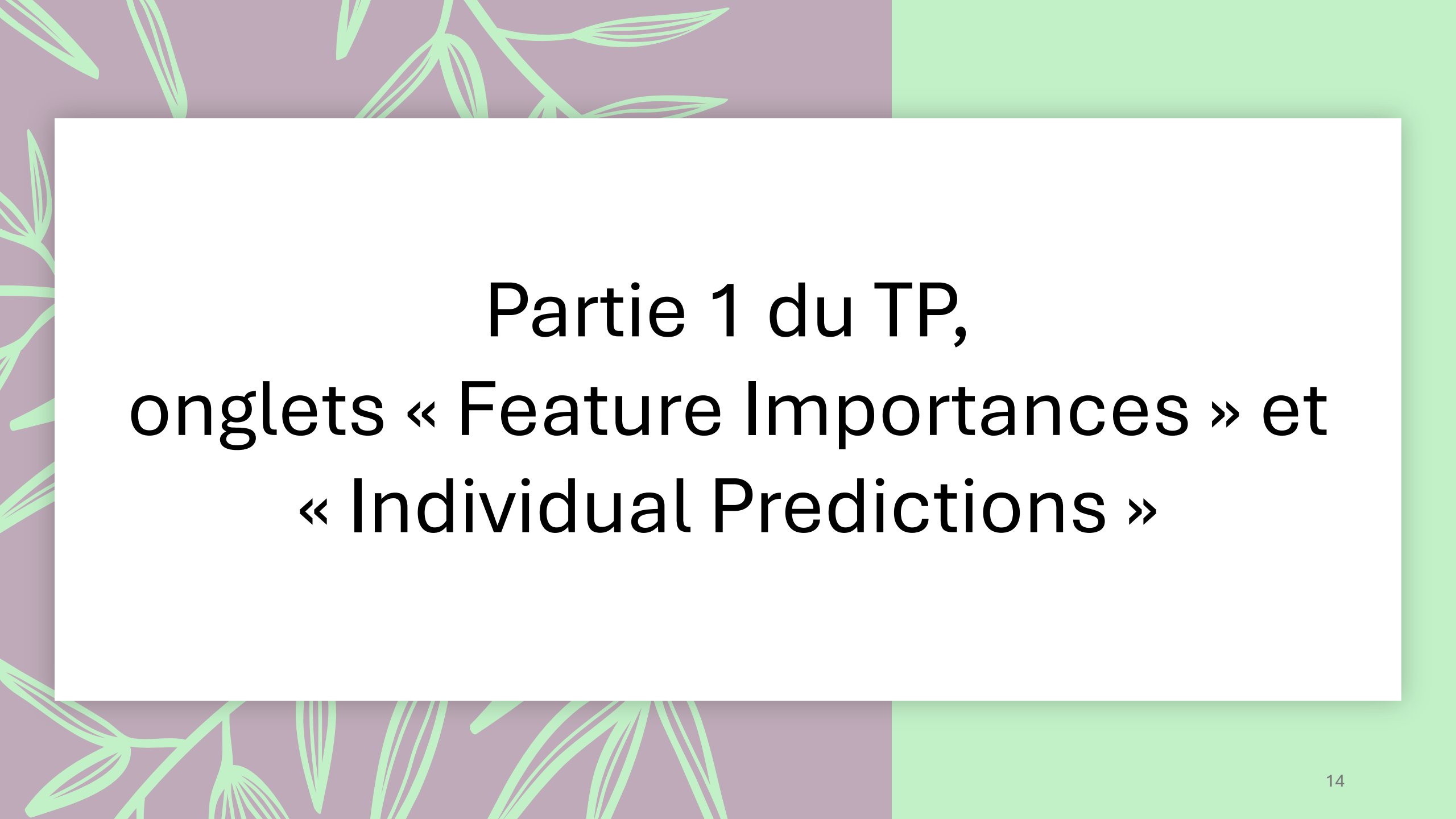
$$C_{12} = 10\,000 \mid C_1 = 7\,500 \mid C_2 = 5\,000 \mid C_0 = 0$$

Contribution marginale du Joueur 2 :

$$C_{12} - C_1 = 2\,500$$

$$C_2 - C_0 = 5\,000$$

$$(2\,500 + 5\,000) / 2 = \$3\,750$$



Partie 1 du TP, onglets « Feature Importances » et « Individual Predictions »

Explications Contrefactuelles

- Scénarios modifiés de l'instance originale qui changent la prédiction
- Comprendre les seuils de décision du modèle
- Identifier les caractéristiques qui influencent les prédictions

Partie 1 du TP, onglet « What If... »

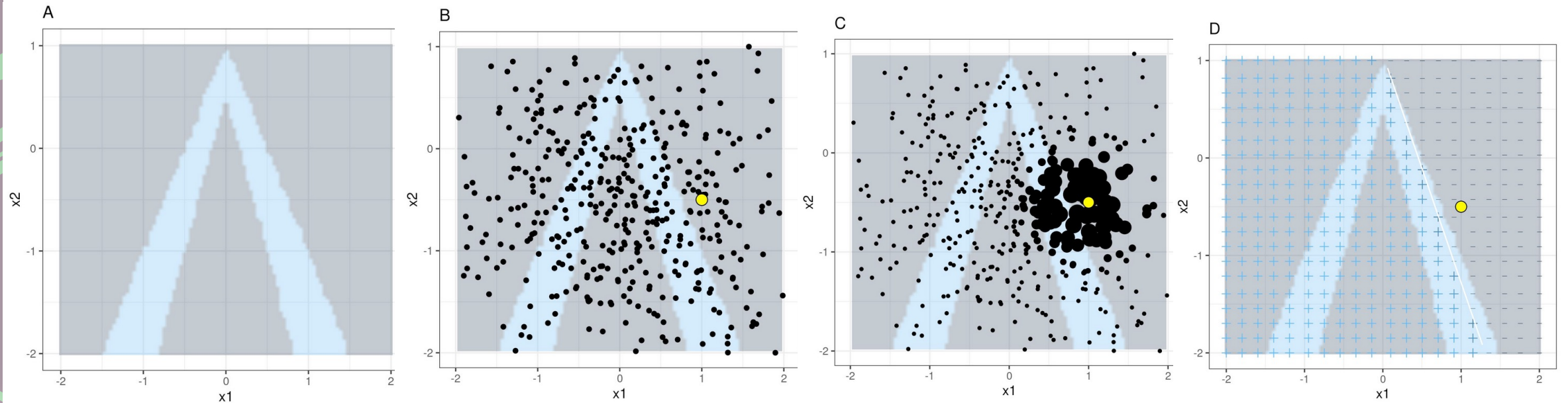
LIME (Local Interpretable Model-agnostic Explanations)

Génère des explications locales en perturbant le voisinage de la prédiction à expliquer



Ces exemples perturbés sont ensuite utilisés pour entraîner un modèle interprétable

LIME (Local Interpretable Model-agnostic Explanations)



LIME sur des images

- Segmentation de l'image en superpixels
- Création de plusieurs versions perturbées de l'image
- Entraînement d'un modèle simple sur les images perturbées
- Utilisation des caractéristiques de présence/absence de chaque superpixel
- Les poids du modèle simple indiquent l'importance de chaque superpixel pour la prédiction

Partie 2, 3 et 4 du TP

**Merci pour votre
participation !**