# Prédire et Maîtriser l'Attrition des Talents

Cette présentation détaille un projet de Machine Learning visant à **identifier et prédire les risques de démission** (Attrition) au sein de l'entreprise. Notre objectif est de transformer les données en informations exploitables pour l'équipe RH, permettant des interventions proactives et ciblées pour la rétention des talents.

Audience: Jury de direction et équipe RH. Accent mis sur l'interprétabilité et la valeur opérationnelle du modèle.

Coulibaly Cheick Oumar

CHAPITRE 1 : CONTEXTE & OBJECTIFS

# L'Impératif de l'Interprétabilité

## Problématique Opérationnelle

L'attrition des employés représente un coût significatif, non seulement en termes de recrutement et de formation, mais aussi en perte de savoir-faire et de productivité. Nous devons passer d'une réaction à la démission à une **anticipation basée sur des données**.

- Identifier les employés présentant un risque élevé de départ.
- Comprendre les leviers sous-jacents pour des actions RH efficaces.
- Prioriser les ressources limitées de l'équipe de rétention.

## Objectif Technique Clé

> Construire un modèle de classification binaire transparent et interprétable dont les décisions peuvent être expliquées aux décideurs et aux managers.

CHAPITRE 2 : PRÉPARATION DES DONNÉES

# Gérer la Complexité des Données RH

## Colinéarité des Features

Observation d'une forte dépendance entre certaines variables (ex: Revenu Mensuel et Niveau Hiérarchique).

**Stratégie retenue :** Conservation de toutes les variables pour maximiser l'information, en acceptant que cela puisse complexifier légèrement l'interprétation initiale.

## Déséquilibre de Classes

Le déséquilibre est majeur : 84% des employés n'ont pas démissionné (Classe Non) contre seulement 16% qui l'ont fait (Classe Oui).

**Stratégie retenue :** Séparation stratifiée des ensembles d'entraînement/test et application de `class_weight='balanced'` pour donner plus de poids aux exemples d'attrition.

## Encodage des Variables

Les variables catégorielles ont été transformées pour être utilisables par le modèle.

**Méthode :** Utilisation de l'encodage Target/Label, plus efficace pour les modèles linéaires que le One-Hot Encoding sur un grand nombre de catégories.

CHAPITRE 3 : MODÉLISATION ET PERFORMANCE

# Choix du Modèle : Simplicité et Interprétabilité

## Justification de la Régression Logistique

Malgré l'existence de modèles plus performants (arbres, réseaux de neurones), la **Régression Logistique Optimisée** a été préférée.

- Facilité d'interprétation des coefficients (risques positifs/négatifs).
- Relation claire entre les facteurs et la probabilité de démission.
- Alignement parfait avec notre objectif de modèle transparent.


---

## Métrique de Sélection : PR AUC

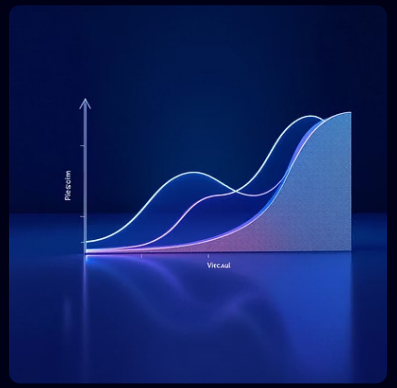
Pour un problème fortement déséquilibré comme celui-ci, l'Accuracy (précision globale) et même le ROC AUC sont trompeurs. Le **PR AUC (Precision-Recall Area Under Curve)** est la métrique la plus fiable et la plus honnête pour évaluer notre capacité à détecter correctement la classe minoritaire (la démission).

# Performance Benchmark et Diagnostic

## PR AUC Optimal : 0,6054

Ce score représente la meilleure performance stable et reproductible que le modèle de Régression Logistique a pu atteindre.

L'analyse des courbes d'apprentissage (Learning Curves) montre que le modèle linéaire a atteint un plateau de performance stable, ce qui indique qu'il capture efficacement l'essentiel des relations présentes dans les données.Ce léger sous-apprentissage est en réalité un compromis positif, car il garantit une interprétabilité maximale tout en offrant une performance solide, tout en évitant le surapprentissage (overfitting) sur le jeu de données.

CHAPITRE 4 : DIAGNOSTIC OPÉRATIONNEL

# Le Compromis entre Détection et Fiabilité

Le seuil de décision du modèle a été choisi pour **maximiser le F1-Score** sur la courbe Précision/Rappel, offrant le meilleur équilibre entre la capacité de détection et la fiabilité des alertes.

## Rappel (Recall) : 63,8%

64% des démissions réelles (Vrais Positifs) sont correctement identifiées par le modèle. C'est notre capacité à détecter le risque.

## Précision (Precision) : 42,9%

Seulement 43% des alertes sont de véritables risques de démission. Cela signifie que 57% des alertes sont des Faux Positifs (Fausses Alarmes).

# Matrice de Confusion au Seuil Optimisé

| | Prédit Non-Départ | Prédit Départ |
|---|---|---|
| **Réel Non-Départ** | VN = 207 (Vrais Négatifs) | FP = 40 (Faux Positifs) |
| **Réel Départ (Attrition)** | FN = 17 (Faux Négatifs) | VP = 30 (Vrais Positifs) |

**Conclusion Opérationnelle :** Le modèle est un outil de couverture des risques performant (détectant la majorité des démissions), mais il génère un nombre significatif de fausses alertes (Faux Positifs). Il doit donc être utilisé pour la **priorisation** des entretiens de rétention, et non comme un diagnostic définitif.

CHAPITRE 5 : INTERPRÉTABILITÉ ET CAUSES

# Facteurs Clés de l'Attrition (Analyse SHAP)

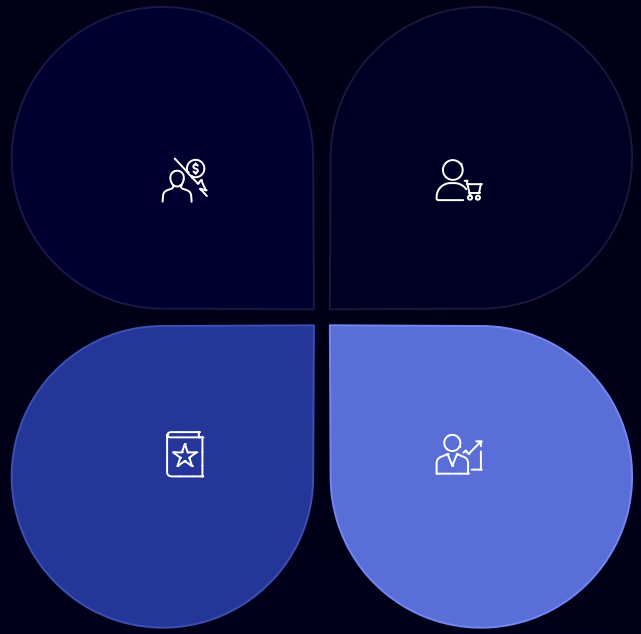
L'analyse des **Valeurs SHAP** (via LinearExplainer, compatible avec la Régression Logistique) permet de décomposer l'influence de chaque variable sur la prédiction de risque de démission, fournissant une explication claire et causale.

### Bas Revenu Mensuel

Le risque le plus élevé. Les salaires faibles augmentent la probabilité de recherche d'opportunités externes.

### Courte Ancienneté au Poste

Les employés récemment arrivés ou promus sont plus susceptibles de repartir s'ils ne trouvent pas rapidement satisfaction.



### Haute Expérience Totale

Les vétérans de l'entreprise tendent à être plus stables, valorisant la continuité et les acquis (avantages, retraite, etc.).

### Haut Niveau Hiérarchique

Facteur de stabilité. Souvent associé à une rémunération et des responsabilités plus satisfaisantes.

CHAPITRE 6 : RECOMMANDATIONS FINALES

# Actions RH Prioritaires Basées sur les Données

1

## Revue Salariale Ciblée

Pour contrer le facteur "Bas Revenu", organiser des revues salariales spécifiques et rapides pour les employés identifiés comme étant à haut risque par le modèle.

2

## Points de Contrôle Carrière

Pour les employés avec une courte ancienneté au poste (<18 mois), instaurer des entretiens de suivi pour vérifier l'adéquation du rôle et prévenir la stagnation.

3

## Priorisation des Entretiens

Utiliser le modèle pour **classer les risques**. Concentrer les efforts d'entretien de rétention sur le top 10-15% des employés avec le plus haut score de risque.

Coulibaly Cheick Oumar

# Le Modèle : Un Outil d'Aide à la Décision

## Considérations Éthiques et Opérationnelles

Il est crucial de se souvenir du taux de Faux Positifs (≈57%). Le score de risque est une **alerte**, non une vérité absolue.

- Le modèle ne doit pas être utilisé pour prendre des décisions individuelles automatiques (licenciement, promotion).
- Le modèle sert uniquement à prioriser les employés qui nécessitent une conversation **humaine** de rétention.
- Les facteurs identifiés (Revenu, Ancienneté) offrent des pistes claires pour les thèmes abordés lors de ces entretiens.

**Planifier une Démo du Modèle**

Coulibaly Cheick Oumar