Système de Recommandation d'Activités : Démarche, Modèles et Évaluation

June 9, 2025

1. Introduction

Ce document présente l'ensemble de la démarche mise en œuvre pour créer un système de recommandation d'activités personnalisées. En l'absence de données réelles, une stratégie de simulation de données a été adoptée pour modéliser des comportements utilisateurs plausibles. Cette approche a permis de créer un environnement de test complet pour l'entraînement et l'évaluation de modèles de recommandation.

2. Construction du jeu de données

2.1 Personas créés

Afin de simuler des utilisateurs aux profils variés, cinq personas typiques ont été définis :

- **Léo** (29 ans, Lyon) : passionné de sport, pratique le crossfit, l'escalade, la course à pied. Fréquence : 4 fois/semaine.
- **Clara** (35 ans, Paris) : amatrice de culture, aime les musées, les expositions, les conférences. Fréquence : 2-3 fois/mois.
- **Jeanne** (42 ans, Nantes) : créative, passionnée d'activités manuelles comme la poterie, la couture, la peinture. Fréquence : hebdomadaire.
- **Mehdi** (38 ans, Montpellier) : chercheur de bien-être, pratique le yoga, le spa, la méditation pour gérer son stress. Fréquence : 1-2 fois/semaine.
- **Lucas** (27 ans, Marseille) : amateur de sensations fortes, adepte de simulateurs, d'escape games extrêmes et d'activités à adrénaline. Fréquence : occasionnelle.

2.2 Tables générées

1. **activites.csv** : 100 activités classées selon les catégories Sport, Bien-être, Culture, Manuels, Extrêmes, localisées à Paris avec likes et rating.
2. **utilisateurs.csv** : 10 000 utilisateurs simulés, rattachés à un persona-type.
3. **historique_utilisateur.csv** : liste des activités passées (1 à 5 par utilisateur).
4. **interactions_utilisateur.csv** : données enrichies avec date et note explicite donnée à chaque activité.
5. **vendors.csv** : prestataires d'activités, associés par ID aux activités.

3. Modèle de Recommandation

3.1 Objectif

Prédire la note qu'un utilisateur donnerait à une activité (`user_rating`) afin de recommander les activités les plus susceptibles d'être appréciées.

3.2 Données utilisées pour l'entraînement

- `user_id`, `activity_id` : identifiants
- `category` : type d'activité (encodé)
- `vendor_id` : prestataire (encodé)
- `likes` : popularité sociale
- `rating` : moyenne des avis

Nota : la variable `persona_category` *a été volontairement exclue pour favoriser un modèle plus neutre et généralisable.*

4. État de l'art : Algorithmes de Recommandation

4.1 Modèles testés

- **Random Forest Regressor** : algorithme à base d'arbres de décision. Il fonctionne bien sans normalisation mais peut sur-apprendre si les données sont très corrélées.
- **LightGBM** : modèle gradient boosting très rapide et efficace, adapté aux grands volumes de données tabulaires.
- **XGBoost** : similaire à LightGBM, performant mais plus lent. Référence dans les compétitions Kaggle.
- **Linear Regression** : baseline simple, utile pour avoir une mesure de référence.

4.2 Alternatives possibles

- **Collaboratif pur** : apprentissage basé uniquement sur les similarités entre utilisateurs ou items (*ex: SVD, kNN*).
- **Filtrage implicite** : intégration de clics, vues ou actions non explicites (*ex: LightFM*).
- **Systèmes hybrides** : combinaison de données de contenu et d'interaction.

5. Evaluation des Modèles

5.1 Métriche : RMSE

RMSE (Root Mean Squared Error) est une métrice classique pour évaluer la précision des modèles de régression. Elle mesure l'écart quadratique moyen entre la note prédite et la note rélle donnée par l'utilisateur. Une RMSE proche de 0 indique une bonne précision.

5.2 Résultats obtenus (80% entrâinement / 20% test)

Modèle	RMSE
Random Forest	0.5646
LightGBM	**0.4815**
XGBoost	0.4966
Linear Regression	0.4932

5.3 Analyse

LightGBM s'impose comme le meilleur compromis entre performance et généralisation. Il capture mieux les interactions complexes entre utilisateur et activité sans sur-apprentissage.

6. Recommandation finale

Pour chaque utilisateur, toutes les activités sont prédites avec une note probable. Les 5 activités ayant la note prédite la plus élevée sont recommandées.

7. Perspectives d'Amélioration

- Ajouter des variables historiques (moyenne de notes utilisateur, fréquence des activités)
- Intégrer la temporalité (effet de saison, récence)
- Tester des modèles collaboratifs comme *Surprise.SVD*