The background of the slide features a light gray, semi-transparent image of two film strips. One strip is on the left, curving upwards, and the other is on the right, curving downwards. They overlap in the center, creating a sense of depth and movement.

Customized Product Proposal

-

Recommender System for Movies

Jérémy Lecourt / DataScientest
Projet d'étude en Data Science

Contexte du projet

Les systèmes de recommandation sont omniprésents sur la toile :

- Intérêt économique (chiffre d'affaire)
- Intérêt métier (satisfaction client / personnalisation)
- Intérêt technique (web app / expérience utilisateur)
- Intérêt scientifique (machine learning avancé)

⇒ Application aux recommandations de Film dans le cadre de ce projet

Objectifs du projet

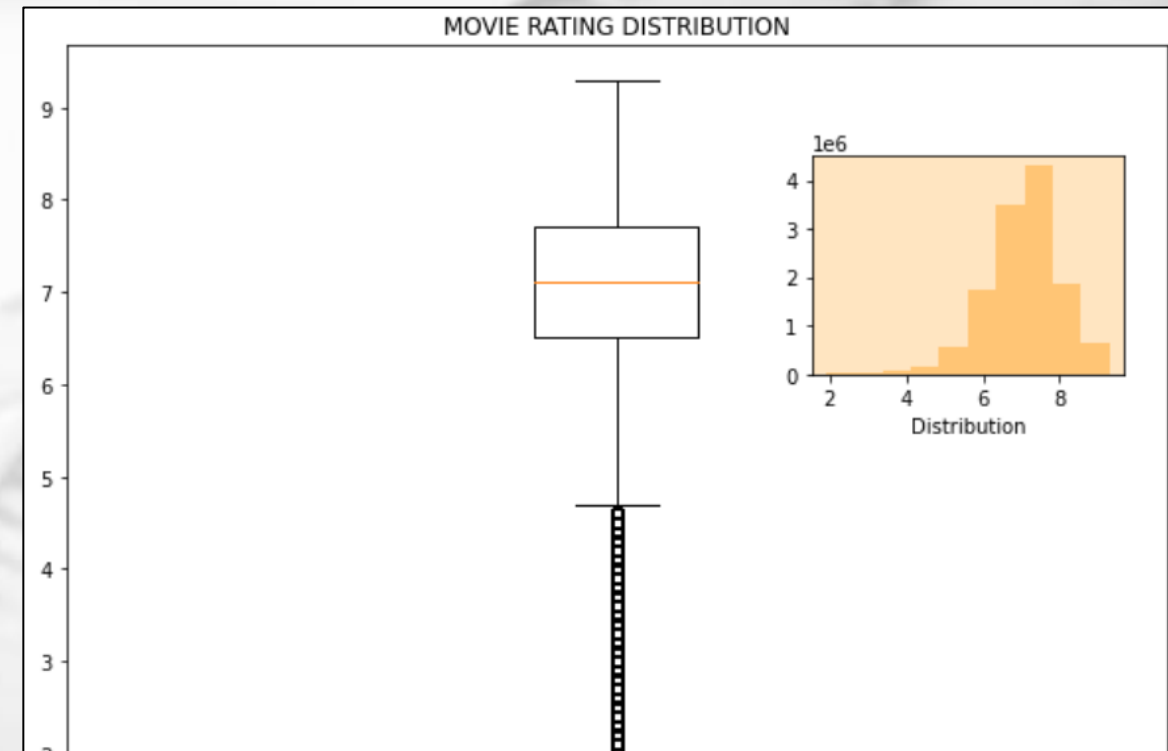
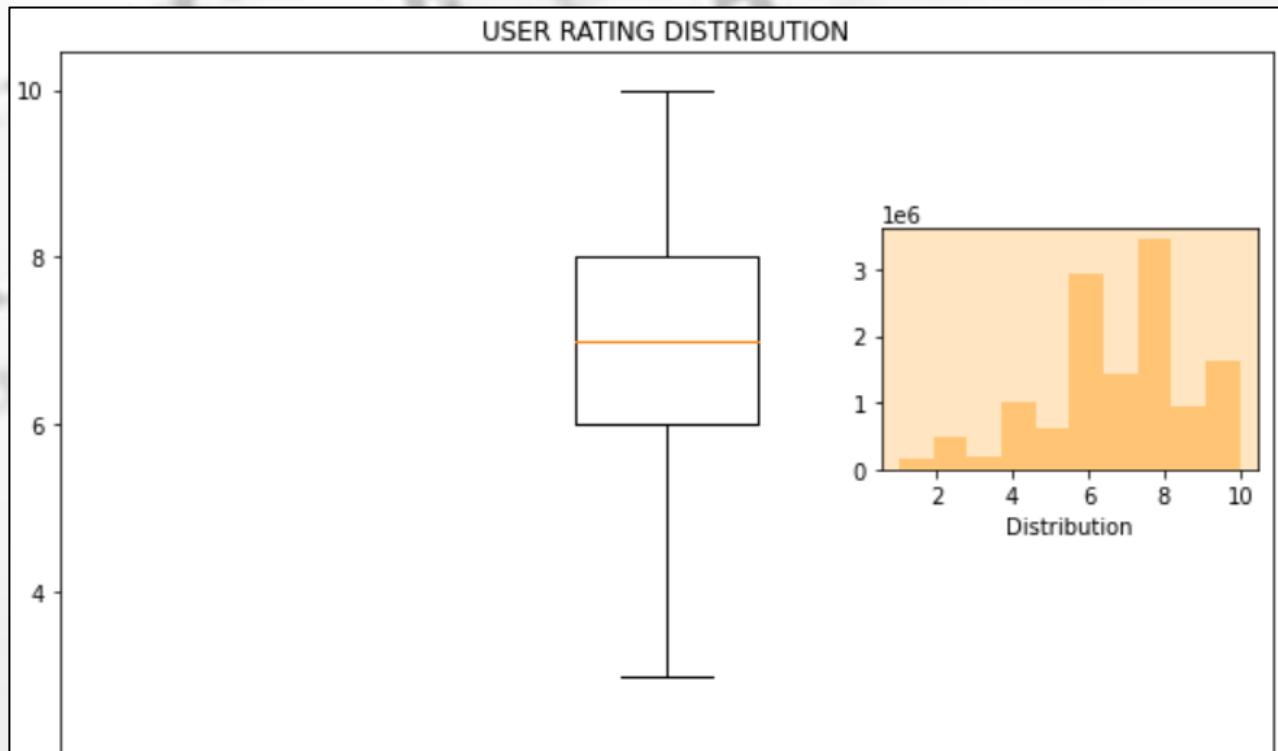
- Mettre en application toutes les étapes d'un projet de machine learning, jusqu'à la mise en production d'une application web conviviale pour établir des propositions de Films personnalisées
 - Web Scraping / Collecte de Données
 - Data Visualisation / Feature Engineering
 - Algorithme de Machine Learning avancé / Optimisation
 - Maquette d'une Application Web
- Améliorer l'approche des demandes personnalisées en intégrant le contexte ou le comportement perçu lors de la notation utilisateur

Collecte des données

- Notes utilisateur et caractéristiques des films récoltées sur IMDB.com
 - Plus de 10 millions de notations, entre 1995 et 2014
 - Films et utilisateurs internationaux (version EN)
 - Commentaires non fournis dans les « Datasets » disponibles en ligne
- Travail réalisé sur les « Datasets » disponibles
 - Croisement des différentes données récoltées via un identifiant unique de Film
 - Nettoyage et reformatage de certaines variables
 - Décomposition des principaux genres associés à chaque Film
 - Filtre temporel tout en gardant un volume de données honorable (env. 100K ratings)

Visualisation : Distribution des notes

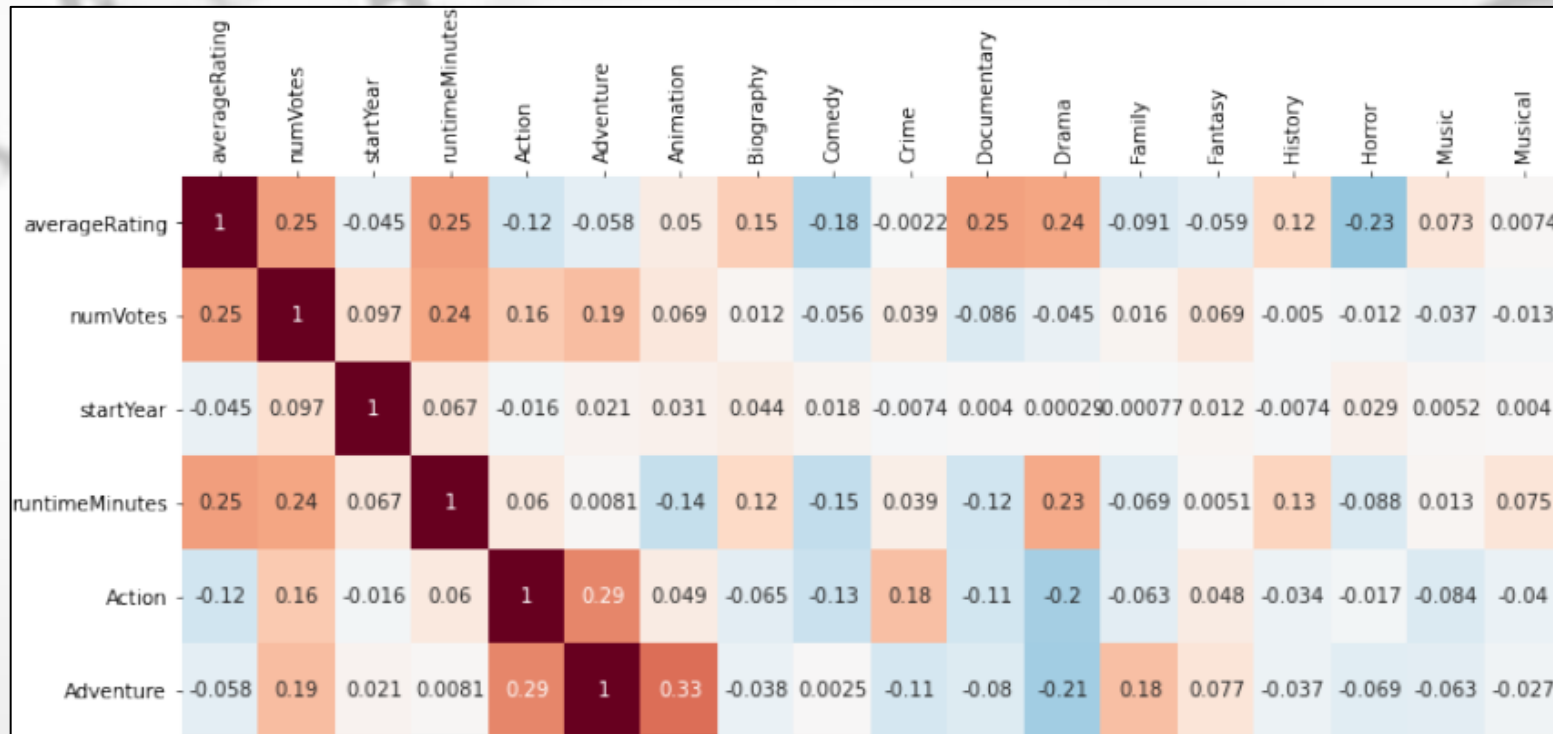
- Distribution des notes moyennes utilisateurs et films :



⇒ Distribution « gaussienne » pour les Films, plus éparpillées pour les Clients, moyenne autour de 7/10

Visualisation : Corrélations des variables

- Corrélations entre la variable cible et la « metadata » sur les Films :



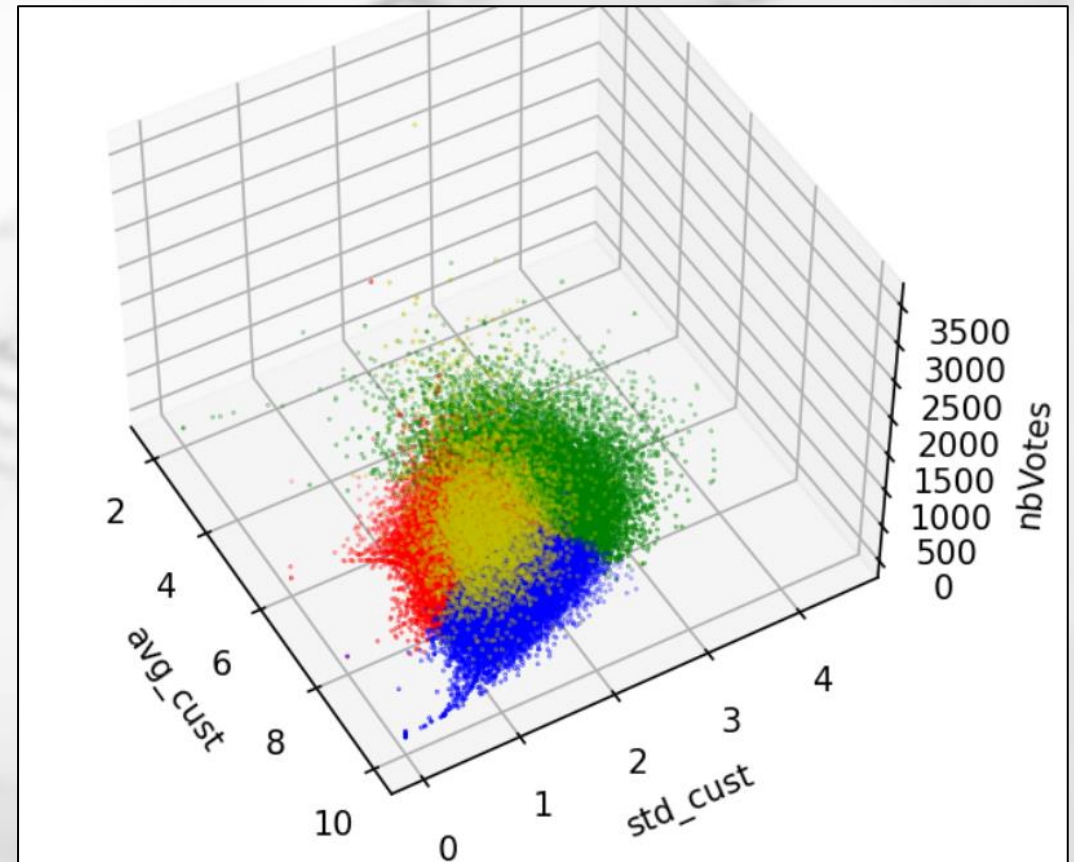
⇒ Corrélations faibles avec la variable cible (notes), approche « Content-Based » écartée

Visualisation : Clustering client

- Feature engineering et Visualisation des clusters clients identifiés :

- « Rouge » : cluster client « classique »
- « Bleu » : cluster client « notes élevées »
- « Vert » : cluster client « notes éparpillées »
- « Jaune » : cluster « Rouge étendu », avec plus de votes effectués (cluster minoritaire)

⇒ Impact du Clustering Client sur la performance des algorithmes Recommender System à venir ?



Modélisation Recommender System

- Différentes approches algorithmiques, via la librairie « Surprise » :
 - NormalPredictor (Distribution Normale Centrée + Maximum Likelihood Estimation)
 - Item-Based et User-Based : KNNBasic et KNNWithMeans (recalibrage inter-client)
 - Matrix Factorisation : SVD (explicite) et SVDpp (explicite et implicite)
- Comparaison des résultats selon les approches :
 - Evaluation via l'écart quadratique moyen (RMSE)

NormalPredictor	KNNBasic	KNNWithMeans	SVD	SVDpp
1,37	0,86	0,84	0,81	0,80

Optimisation Recommender System

Les performances ont été améliorées itération après itération :

➤ Via le choix de l'algorithme

La RMSE est passée de **1,37** à **0,80**, depuis le NormalPredictor jusqu'au SVD classique

➤ Via l'amélioration de la dispersion du dataset : « sparcity »

La RMSE est passée de **0,85** à **0,80**, avec une dispersion qui a évoluée de 99% à 92,3%

➤ Via le choix des paramètres optimaux sur l'algorithme et le dataset retenus

La RMSE est passée de **0,80** à **0,78**, au bout de 10h d'entraînement/évaluation (1800 combinaisons).

Améliorations : contexte / comportement

- Approches entrevues pendant le projet
 - « Webscraping » pour étudier les commentaires associés aux notes
 - Impact du clustering réalisé, notamment pour le cluster « client éparpillé »
- Ouverture à d'autres approches
 - Analyse de contexte via les commentaires (ajustement/filtrage notation)
 - Analyse de sentiment via les commentaires (cohérence notation et commentaire)
 - Ajout des données de navigation / comportement des clients (approches hybrides)

Conclusion

- Du point de vue de la Formation DataScientest
 - Projet complet : mise en pratique de plusieurs blocs du cursus, et librairie hors cursus
 - Problématique « grandeur nature » (complexité de récolte de la donnée appropriée)
 - Ouvertures et améliorations possibles pour de futurs projets
- D'un point de vue personnel
 - Lien avec une partie de mon expérience professionnelle (vente omnicanale)
 - Satisfait des connaissances acquises lors de cette reconversion (variées et pointues)
 - Remerciement pour l'adaptabilité du cursus, en parallèle de mon activité