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

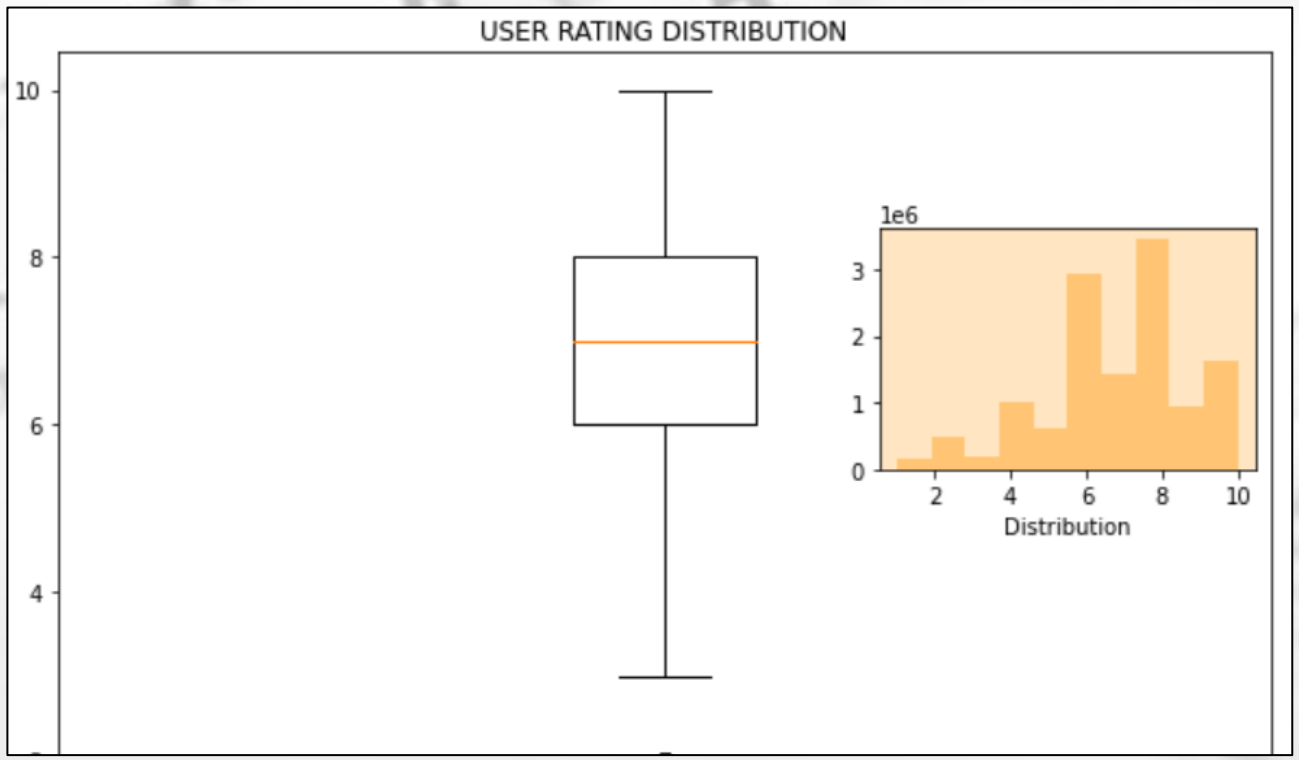
- Améliorer l'approche des demandes personnalisées en intégrant le contexte ou le comportement perçu lors de la notation utilisateur

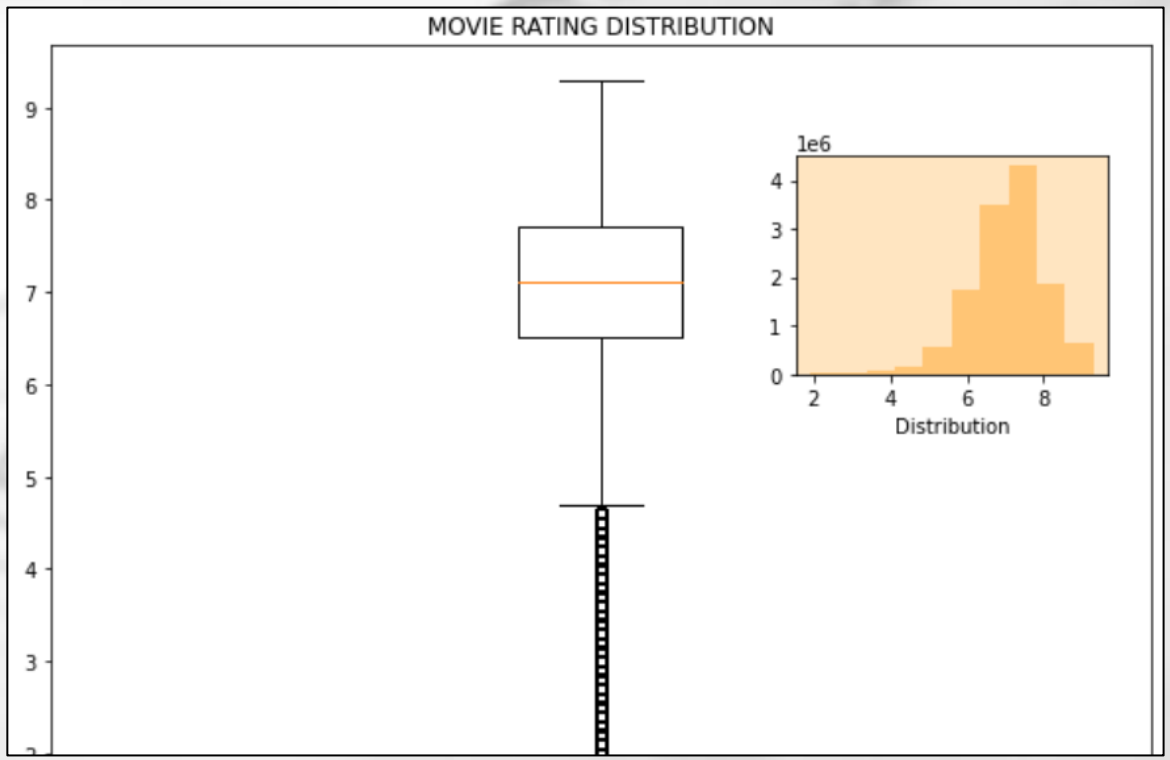
Collecte des données

- 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)

Data Visualisation : Distribution des notes

- Distribution des notes moyennes utilisateurs et films :

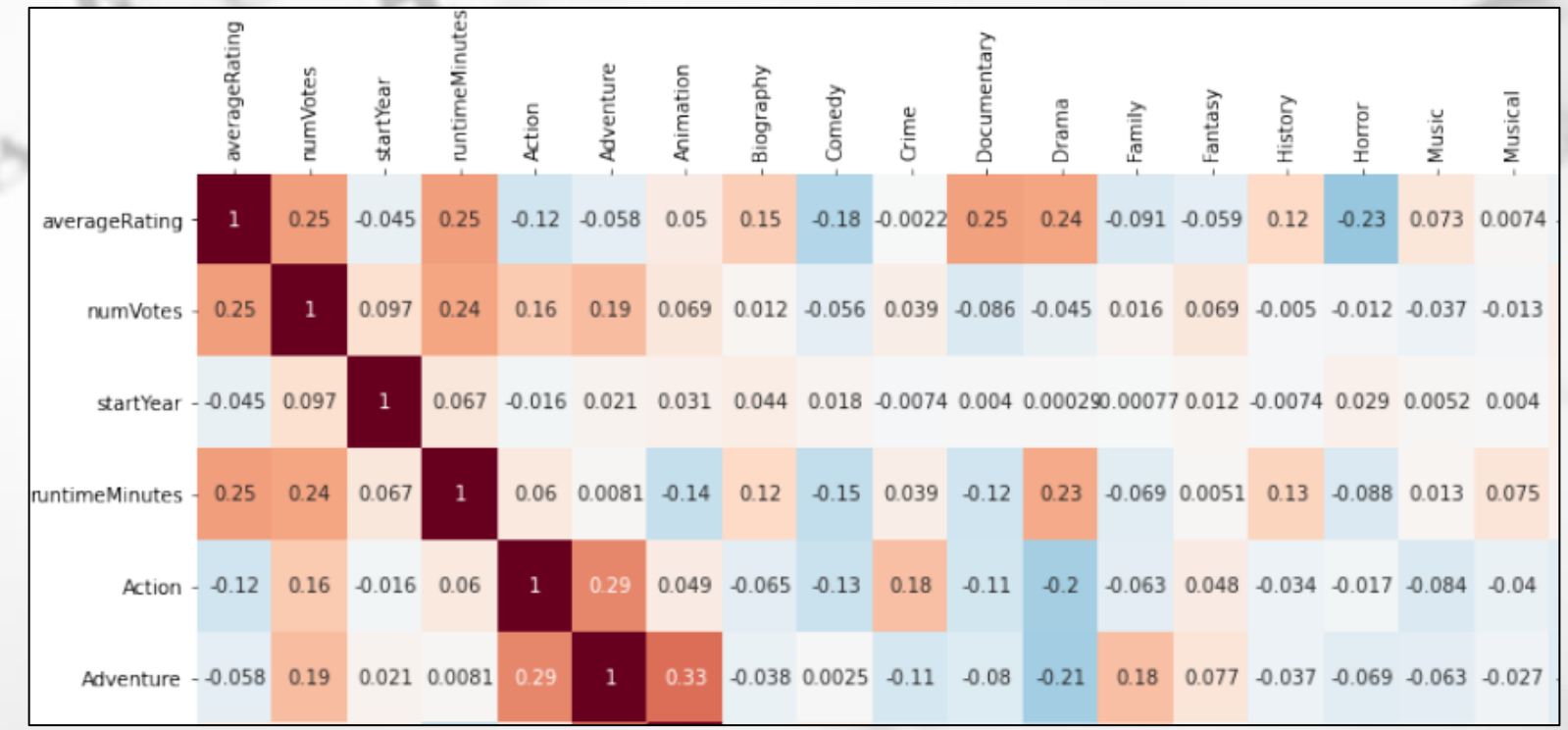




⇒Distribution « gaussienne » pour les Films, plus éparpillées pour les Clients, moyenne autour de 7/10

Data Visualisation : Corrélations des variables

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



	averageRating	numVotes	startYear	runtimeMinutes	Action	Adventure	Animation	Biography	Comedy	Crime	Documentary	Drama	Family	Fantasy	History	Horror	Music	Musical
averageRating	1	0.25	-0.045	0.25	-0.12	-0.058	0.05	0.15	-0.18	-0.0022	0.25	0.24	-0.091	-0.059	0.12	-0.23	0.073	0.0074
numVotes	0.25	1	0.097	0.24	0.16	0.19	0.069	0.012	-0.056	0.039	-0.086	-0.045	0.016	0.069	-0.005	-0.012	-0.037	-0.013
startYear	-0.045	0.097	1	0.067	-0.016	0.021	0.031	0.044	0.018	-0.0074	0.004	0.00029	0.00077	0.012	-0.0074	0.029	0.0052	0.004
runtimeMinutes	0.25	0.24	0.067	1	0.06	0.0081	-0.14	0.12	-0.15	0.039	-0.12	0.23	-0.069	0.0051	0.13	-0.088	0.013	0.075
Action	-0.12	0.16	-0.016	0.06	1	0.29	0.049	-0.065	-0.13	0.18	-0.11	-0.2	-0.063	0.048	-0.034	-0.017	-0.084	-0.04
Adventure	-0.058	0.19	0.021	0.0081	0.29	1	0.33	-0.038	0.0025	-0.11	-0.08	-0.21	0.18	0.077	-0.037	-0.069	-0.063	-0.027

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

Data Visualisation : Clustering client

- Feature engineering et Visualisation des clusters clients identifiés :

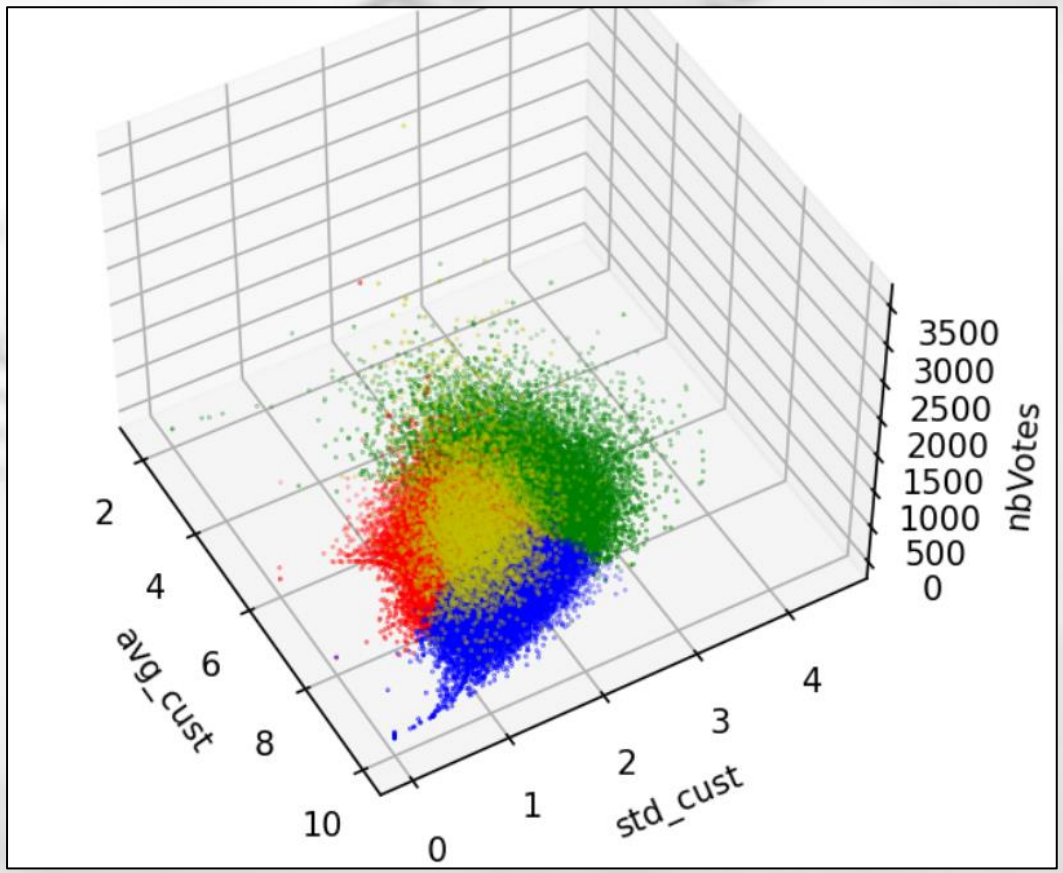
➢ « Rouge » : cluster client « classique »

➢ « 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 Likehood 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).

Amélioration via le « contexte de notation»

- 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)

Intégration dans une application Web

- Proposition de Films personnalisées : « TOP 100 » + « TOP 3 / GENRE »

Veuillez sélectionner le genre souhaité :

Choisir le genre de votre film à regarder:

Comedy

	movieId	originalTitle	genres_y	rating	startYear	runtime
5	588	Aladdin	Adventure,Animation,Co…	4.4812	1992	
8	3114	Toy Story 2	Adventure,Animation,Co…	4.3648	1999	
9	50872	Ratatouille	Adventure,Animation,Co…	4.2547	2007	

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