Jedha

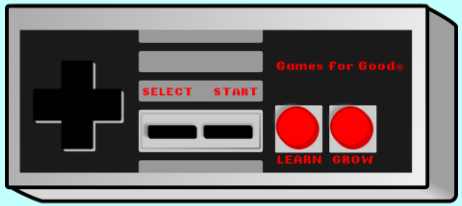
Data Science Bootcamp

Marché du Jeu Vidéo

Etude de l'impact des systèmes de classification et de notation sur les ventes mondiales de jeux vidéo depuis 1980.

Maxime LARCHER

Agenda



- Contexte de l’étude
- Nettoyage du dataset et Exploratory Data Analysis (EDA)
- Modèle de prédiction
- Résultats et conclusion

Contexte de l’étude

- Vous êtes Data Analyst d’un grand studio de jeu vidéo.

- Vous êtes contacté par le Head of Production du studio qui cherche à définir sa stratégie de référencement et de Marketing pour le portfolio de jeux de l’année à suivre.

- Pour orienter ses ressources, ce dernier vous demande de fournir une étude permettant d’estimer l’impact des systèmes de notation (utilisateur et presse) et de classification (rating par âge type ESRB) sur les ventes globales de jeux.

- Il vous fournit une base de données brutes, retraçant ces données de 1980 à fin 2016.

Nettoyage du modèle et EDA

Nettoyage du dataset et EDA

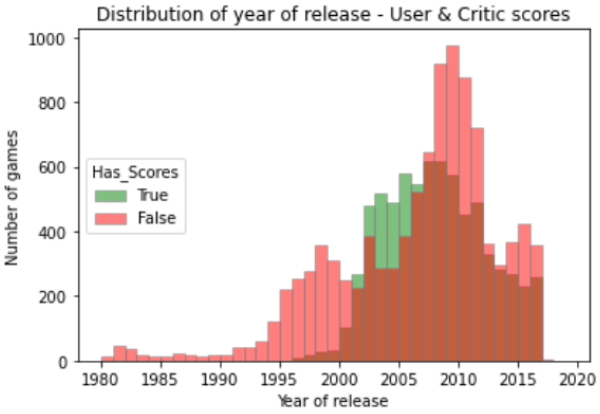
Découverte des données et explication

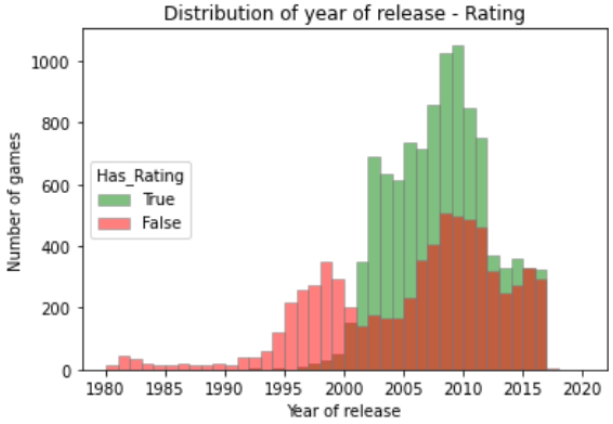
- **Nom** et **genre** (*sport, action,...*) du jeu vidéo
- **Studio éditeur, console** et **année de sortie** du jeu vidéo
- **Ventes mondiales en million d'exemplaires vendus** (et ventilation entre les régions Europe / Amérique du Nord /Japon / Autres régions)
- **Notes attribuées par la presse** (critiques professionnels) **et utilisateurs** sur les médias de référence (blog, web magazine, plateformes de notation...)
- **Nombre de votants Presse** et **Utilisateurs**
- **Classification** d'après le système de référence **European Systemic Risk Board** (ESRB)

Nettoyage du dataset et EDA

- **Première observation** : le dataset comporte de nombreuses valeurs manquantes
  - Sur un volume faible pour les informations relatives à l'année de sortie ou l'éditeur
  - Sur un volume très important pour les informations relatives à la notation, au nombre de votants et à la classification (*rating*).
    Ces données constituent le **cœur de notre étude** et sont difficiles à contourner par une stratégie d'imputation : on choisira donc de supprimer ces entrées et de recentrer le périmètre de l'étude.

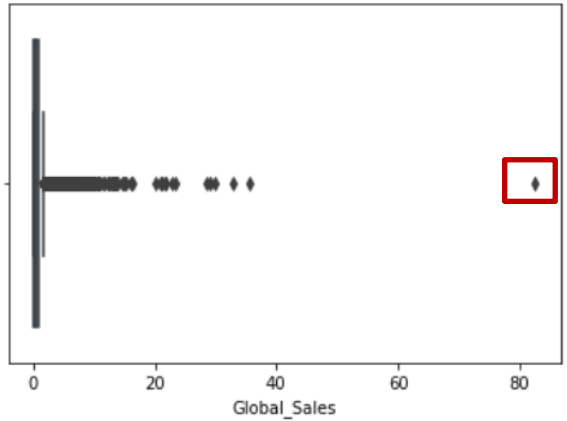
#	Column	Non-Null Count	Dtype
0	Name	17416 non-null	object
1	Platform	17416 non-null	object
2	Year_of_Release	17408 non-null	float64
3	Genre	17416 non-null	object
4	Publisher	17415 non-null	object
5	NA_Sales	17416 non-null	float64
6	EU_Sales	17416 non-null	float64
7	JP_Sales	17416 non-null	float64
8	Other_Sales	17416 non-null	float64
9	Global_Sales	17416 non-null	float64
10	Critic_Score	8336 non-null	float64
11	Critic_Count	8336 non-null	float64
12	User_Score	7798 non-null	float64
13	User_Count	7798 non-null	float64
14	Rating	10252 non-null	object

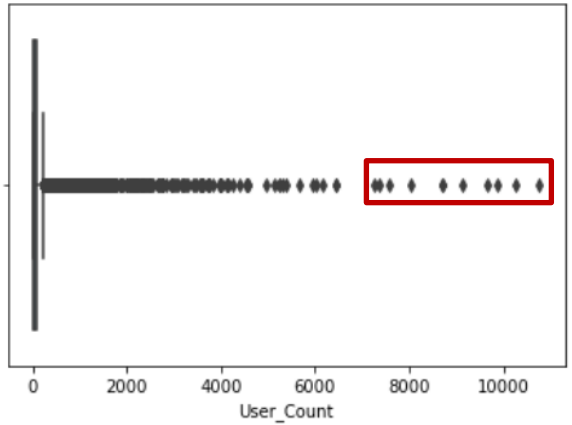




Nettoyage du dataset et EDA

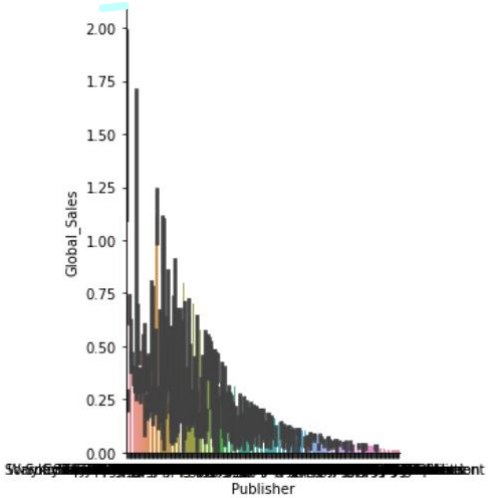
- **Seconde observation** : le Dataset fait apparaitre de sérieuses données aberrantes (outliers)
  - Ces données aberrantes concernent les variables Ventes Globales/EU/NA ainsi que le compte de votants Utilisateur.
  - Ces données peuvent avoir du sens, notamment pour l'étude des **bestsellers** ou des phénomène de buzz, mais on décidera de s'en passer dans le cadre de ce premier niveau d'étude.



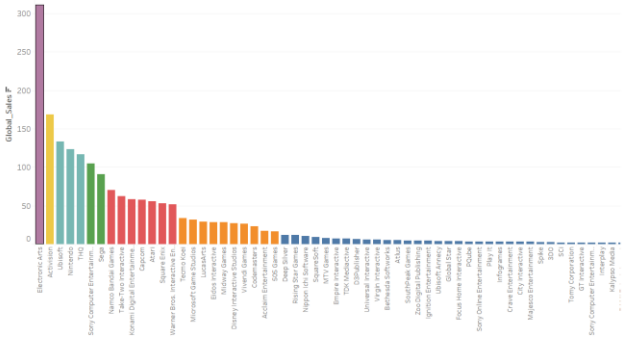


Nettoyage du dataset et EDA

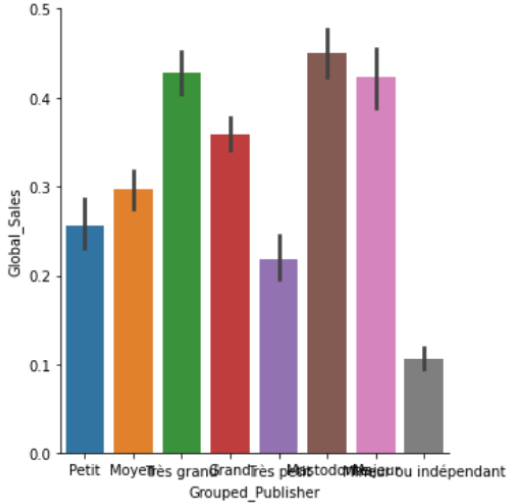
- **Troisième observation** : la variable éditeur comporte pas moins de 260 valeurs uniques
  - **Risque identifié** : niveau de détail/ventilation menant à une perte de précision (volume non-significatif par éditeur, forte variance selon le succès des sorties)
  - **Solution proposée** : catégorisation des éditeurs par taille de marché (majeur, indépendant…)



Répartition des ventes globales par éditeur, pré-traitement



Groupement des éditeurs par taille via Tableau Desktop



Aperçu de la répartition des ventes globales par éditeur, post-traitement

Nettoyage du dataset et EDA

Autres corrections mineures

```
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Name             17416 non-null  object
 1   Platform         17416 non-null  object
 2   Year_of_Release  17408 non-null  float64
 3   Genre            17416 non-null  object
 4   Publisher        17415 non-null  object
 5   NA_Sales         17416 non-null  float64
 6   EU_Sales         17416 non-null  float64
 7   JP_Sales         17416 non-null  float64
 8   Other_Sales      17416 non-null  float64
 9   Global_Sales     17416 non-null  float64
 10  Critic_Score     8336 non-null   float64
 11  Critic_Count     8336 non-null   float64
 12  User_Score       7798 non-null   float64
 13  User_Count       7798 non-null   float64
 14  Rating           10252 non-null  object
```

Correction de la donnée temporelle

Critic_Score	Critic_Count	User_Score	User_Count
76.0	51.0	8.0	324.0

Harmonisation des bases de notation

Nettoyage du dataset et EDA

Conclusion du Data Cleaning

- Un nombre non-négligeable de données manquantes dans le contexte de notre étude qui a amené une coupe nette dans le volume d'information à disposition.
- L'étude de stratégies de contournement différentes pourrait constituer une piste intéressante de complétion du modèle (étude dédiée aux jeux sans notation/rating ?)

- Des Outliers clairement identifiés sur les ventes et nombres de votants, écartées dans le cadre de ce premier niveau d'analyse mais dont l'étude dédiée pourrait constituer un nouvel axe de lecture (bestsellers, effets de buzz)

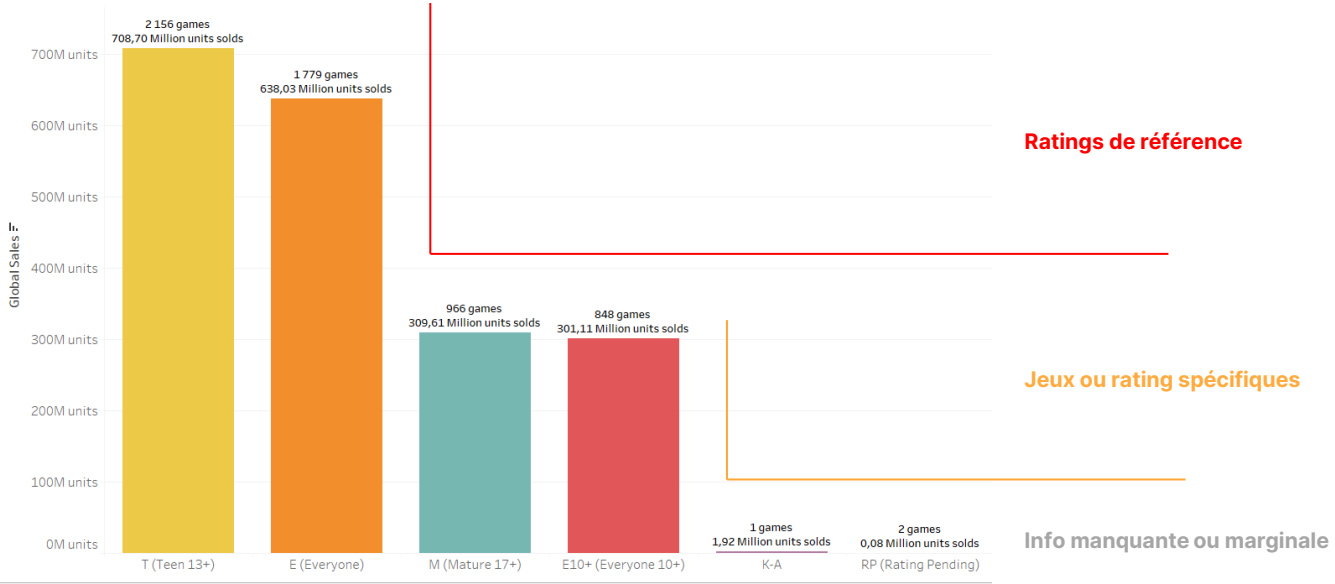
- Harmonisation des catégories et des bases de notation

```
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Name             5752 non-null   object
 1   Platform         5752 non-null   object
 2   Genre            5752 non-null   object
 3   Publisher        5752 non-null   object
 4   NA_Sales         5752 non-null   float64
 5   EU_Sales         5752 non-null   float64
 6   JP_Sales         5752 non-null   float64
 7   Other_Sales      5752 non-null   float64
 8   Global_Sales     5752 non-null   float64
 9   Critic_Score     5752 non-null   float64
 10  Critic_Count     5752 non-null   float64
 11  User_Score       5752 non-null   float64
 12  User_Count       5752 non-null   float64
 13  Rating           5752 non-null   object
 14  Age              5752 non-null   float64
 15  Grouped_Publisher  5752 non-null   object
dtypes: float64(10), object(6)
memory usage: 923.9+ KB
```

Nettoyage du dataset et EDA

EDA : Rating et ventes globales de jeu

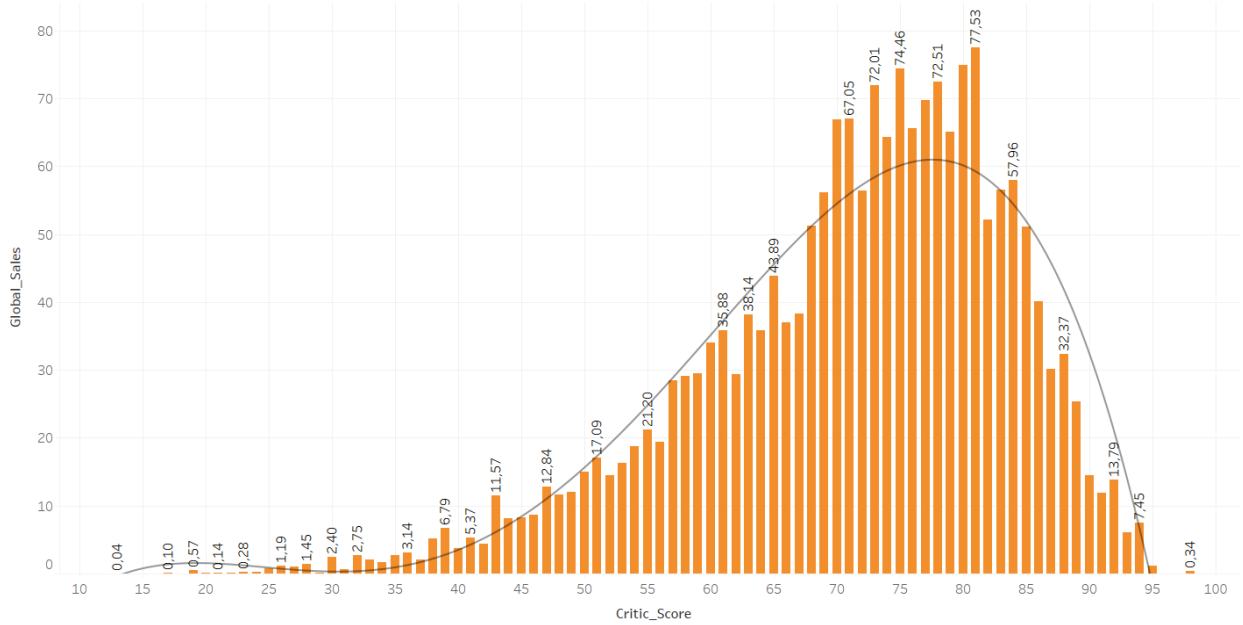
- Trois groupes semblent se dessiner en terme de Rating et d'impact sur les ventes globales



Nettoyage du dataset et EDA

EDA : Ventes globales par notation Presse

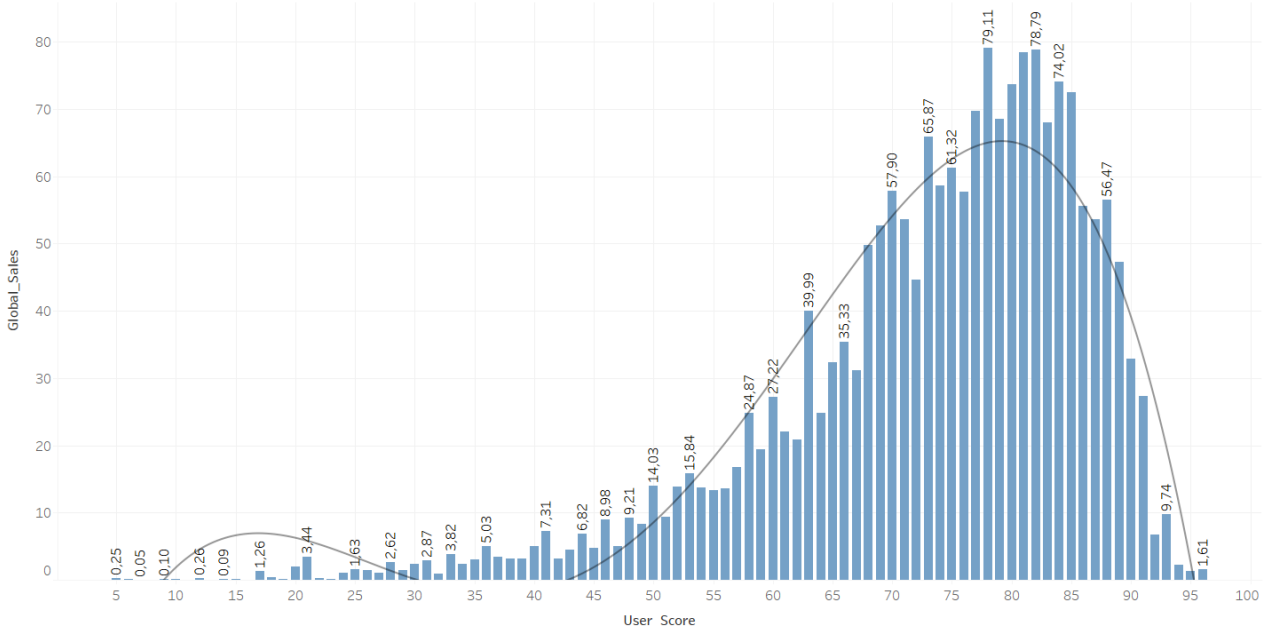
- Une tendance se dessine, centrée autour des notations 70-85



Nettoyage du dataset et EDA

EDA : Ventes globales par notation Utilisateurs

- Une tendance se dessine, centrée autour des notations 70-85



Modèle de prédiction

Modèle de prédiction

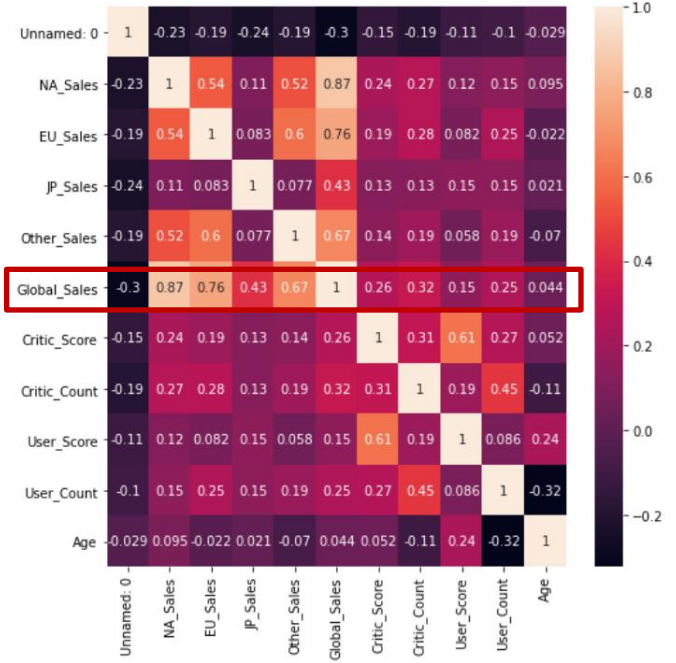
Hypothèses retenues pour la conception du modèle de prédiction

- Variable cible : Ventes globales

- Variables explicatives :
  * Age du jeu
  * Editeurs catégorisés
  * Console et genre du jeu vidéo
  * Notes utilisateurs et nombre de votants
  * Notes Presse et nombre de votants
  * Rating du jeu vidéo

  /! On ne retiendra pas les variables NA/EU/JP/Other Sales, étant une décomposition directe de la target

- Modèle de prédiction utilisé : Régression linaire multiple

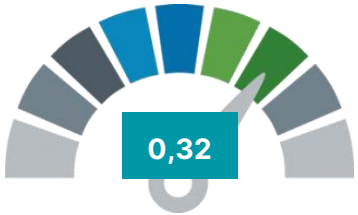


Résultats et conclusion

Résultats et conclusion

Performance du modèle

Coefficient de détermination sur le jeu d'entraînement



Coefficient de détermination sur le jeu de test

En l'état, le modèle de prediction et les features sélectionnés permettent d'expliquer 34% de la variation de la cible Ventes Globales

Est-ce la conclusion finale de l'étude (correlation non-évidente entre la notation/classification et les ventes) ? Aller plus loin ?

Pistes d'exploration et d'amélioration

Enrichissement du dataset, ajout de features explicatives

Gestion des outliers : affiner la détection des outliers ou utiliser un modèle de preprocessing adapté pour la gestion des valeurs extrêmes (RobustScaler)

Feature engineering : travail autour du volume de NULL écarté

Travail autour du régresseur : exploration de nouveaux modèles de prediction (XGBRegressor) et optimisation des hyperparamètres

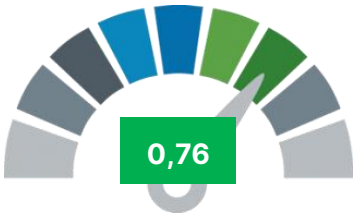
Résultats et conclusion

Exploration autour du régresseur XGBoost

Le **Boosting de Gradient** est un algorithme d'apprentissage supervisé dont le principe et de combiner les résultats d'un ensemble de modèles plus simple et plus faibles afin de fournir une meilleure prédiction.

Ce modèle de prédiction est l'un des plus **populaires** et **puissants** de la sphère actuelle de Data Science.

Performance du modèle



Coefficient de détermination sur le jeu d'entraînement

Coefficient de détermination sur le jeu de test

Le choix du régresseur a résulté en un fort impact sur la performance générale du modèle de prédiction : ce choix constituait bien une piste d'amélioration pertinente de notre travail de prédiction.

Jedha

Merci,

à bientôt !