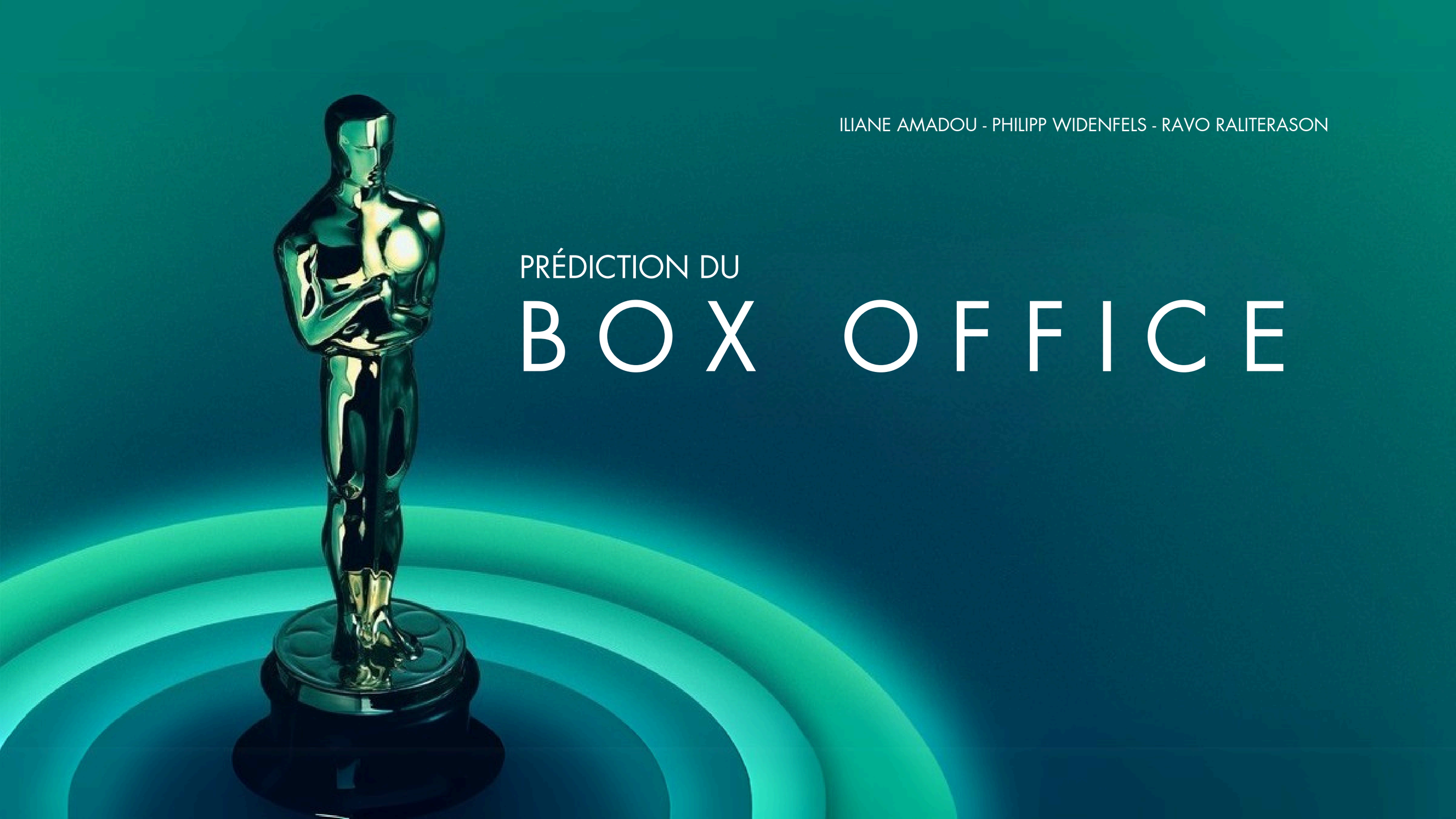
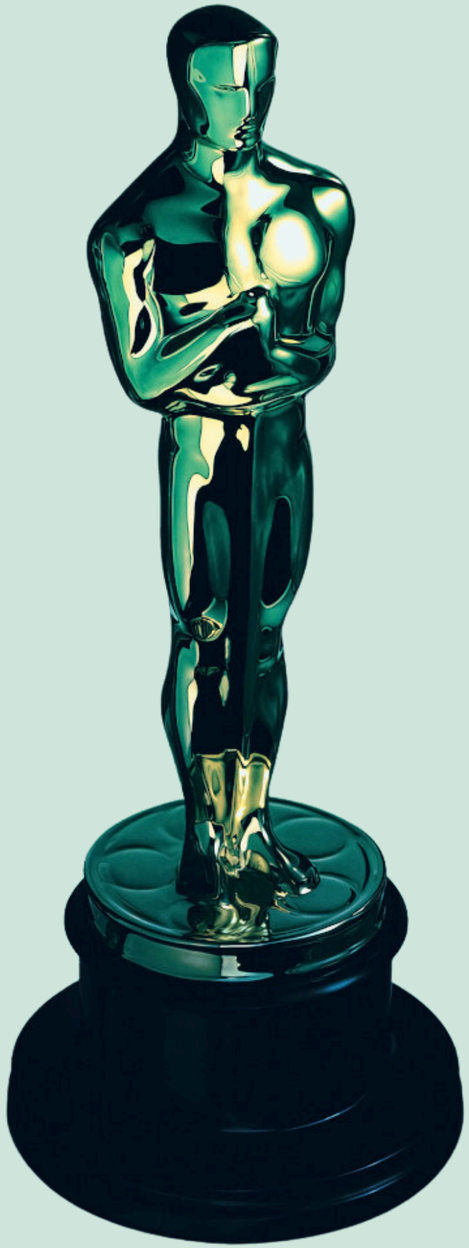


ILIANE AMADOU - PHILIPP WIDENFELS - RAVO RALITERASON

PRÉDICTION DU BOX OFFICE

SOMMAIRE



1 INTRODUCTION

CONTEXTE

L'industrie cinématographique représente un secteur économique majeur où les investissements peuvent atteindre plusieurs centaines de millions de dollars pour un seul film. Dans ce contexte à fort enjeu financier, la capacité à prédire le succès commercial d'un film avant même sa production constitue un avantage stratégique considérable pour les studios et les investisseurs.

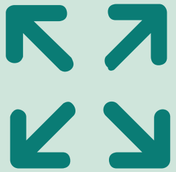
OBJECTIF

Le projet y_BoxOffice vise à développer un modèle de prédiction des revenus au box-office basé exclusivement sur les métadonnées disponibles avant la production d'un film.

MÉTHODOLOGIE

Contrairement à d'autres approches qui s'appuient sur des données post-production comme les votes du public ou les critiques, notre modèle se concentre uniquement sur les informations accessibles au stade initial du projet cinématographique : le genre, le budget prévisionnel, les sociétés de production impliquées, la période de sortie envisagée, et d'autres caractéristiques similaires.

2 PRÉSENTATION DU DATASET



TAILLE

Un dataset conséquent et complet.

24 features
+1.2 M de films

SOURCE

Une base de données en ligne collaborative dédiée aux films et aux personnes impliquées dans l'industrie du cinéma.

TMDb “The Movie Database” (Kaggle)

IDENTITÉ

Toutes les colonnes qui définissent ce qu'est le film et son contenu.

id, imdb_id, title, original_title, original_language, tagline, overview, adult, poster_path, backdrop_path

PRODUCTION

Les colonnes liées au processus de création et aux caractéristiques de production.

budget, production_companies, production_countries, spoken_languages, runtime, genres, keywords

SORTIE

Colonnes concernant la mise sur le marché du film.

status, release_date, homepage

PERFORMANCE

Les métriques liées au succès commercial et critique.

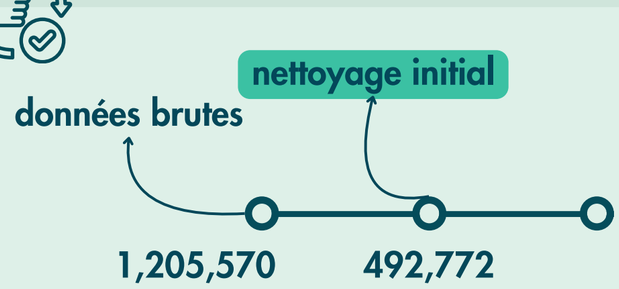
revenue (Y), popularity, vote_averag e, vote_count

3 PRÉ-TRAITEMENT

NETTOYAGE INITIAL

1. **Suppression des colonnes jugées non pertinentes pour notre objectif (prédiction de revenus)**
2. **Application de filtres initiaux pour enlever les données aberrantes ou incomplètes**

Suivi du nombre de lignes restantes



1

Suppression des features non pertinentes

'backdrop_path', 'poster_path', , 'home_pageadult', 'id', 'imdb_id'

Suppression des features non disponibles avant la sortie du film

'vote_average', 'vote_count', 'popularity'

2

Gestion des données aberrantes ou incomplètes

Revenu max raisonnable

Compagnies de production non nulles

Date de sortie valide et avant 2024

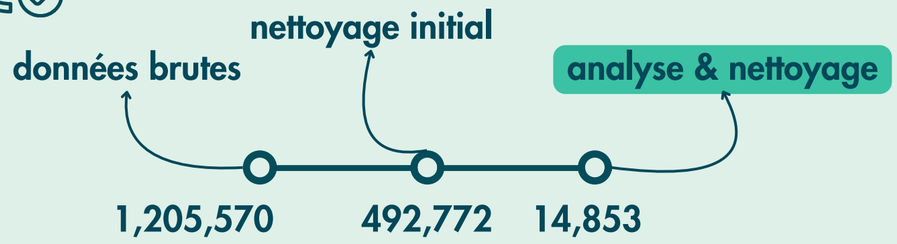
3

PRÉ-TRAITEMENT

ANALYSE ET NETTOYAGE

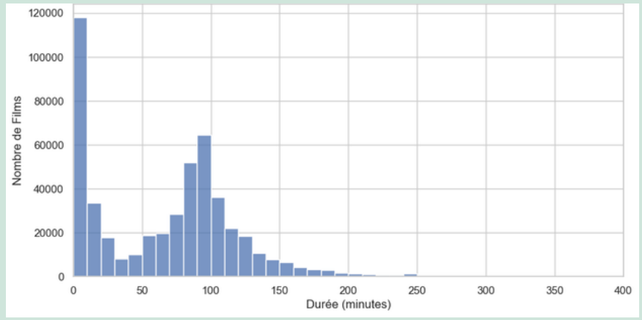
1. Examen de la distribution de la durée des films et des revenus
2. Exclusion des valeurs extrêmes ou non pertinentes (ex: courts métrages, films sans revenus significatifs).

Suivi du nombre de lignes restantes

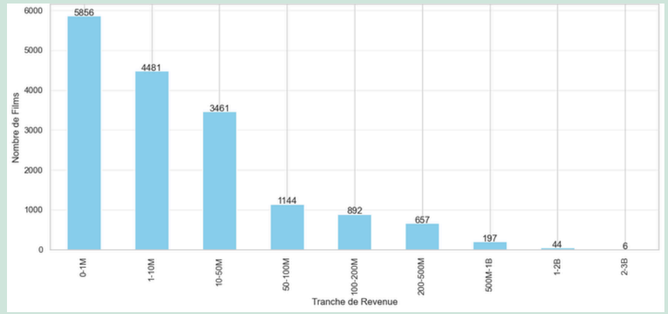


1

Distribution de la durée



Distribution du revenu



2

Filtrage

Durée entre 30 et 280 min
Revenu supérieur à 1000$

3

PRÉ-TRAITEMENT

FEATURE ENGINEERING

1. Text to Numbers (Embeddings): Conversion de `title`, `tagline`, & `overview` en vecteurs numériques riches (Sentence Transformers) pour capturer le sens sémantique.

2. Nettoyage catégorique:
    * Listes structurées: Analyse des `genres` et des `mots clés` séparés par des virgules en listes utilisables.
    * Encodage uni-chaud/multi-chaud:** Transformation des catégories comme `langue_originale`, `genres`, et `mots-clés` en caractéristiques binaires (0/1), ce qui les rend lisibles par la machine.

3. Caractéristiques numériques et de date: Utilisation d'entrées directes comme `budget`, `runtime`, et extraction de `release_year` et `release_month` pour les modèles temporels.

4. Mise à l'échelle des caractéristiques: Normalisation des caractéristiques numériques (y compris les embeddings) pour assurer une contribution équitable au modèle.

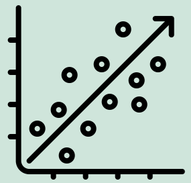
5. Cohérence des colonnes: Veille à ce que toutes les caractéristiques soient parfaitement alignées entre l'entraînement et la prédiction.

4 MÉTHODES D'APPRENTISSAGE

MÉTHODE 1

Régression linéaire

Modèle statistique simple qui cherche à ajuster une droite données pour prédire une variable cible à partir de variables explicatives.



1

AVANTAGES

- Simple et rapide à entraîner.
- Facilement interprétable (coefficients des variables).
- Utile pour comprendre les relations linéaires entre variables.

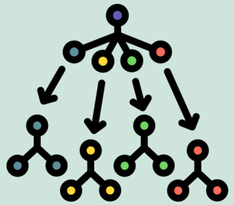
INCONVÉNIENTS

- Suppose une relation linéaire entre les variables, ce qui limite sa flexibilité.
- Sensible aux valeurs aberrantes (outliers).
- Moins performant sur des problèmes complexes/non-linéaires.

MÉTHODE 2

Random forest

Ensemble d'arbres de décision construits sur des sous-échantillons aléatoires des données. Les prédictions sont faites par moyennage



2

AVANTAGES

- Gère bien les relations non linéaires et les interactions complexes.
- Moins sensible aux outliers.
- Réduit le risque de surapprentissage par agrégation.

INCONVÉNIENTS

- Moins interprétable qu'un modèle linéaire.
- Plus lent à entraîner et à prédire (surtout avec beaucoup d'arbres).
- Peut surajuster si le nombre d'arbres est trop petit ou mal paramétré.

5 RÉSULTATS EXPÉRIMENTAUX

MÉTHODE 1

Régression linéaire

R2 score (log): 0.4427

MÉTHODE 2

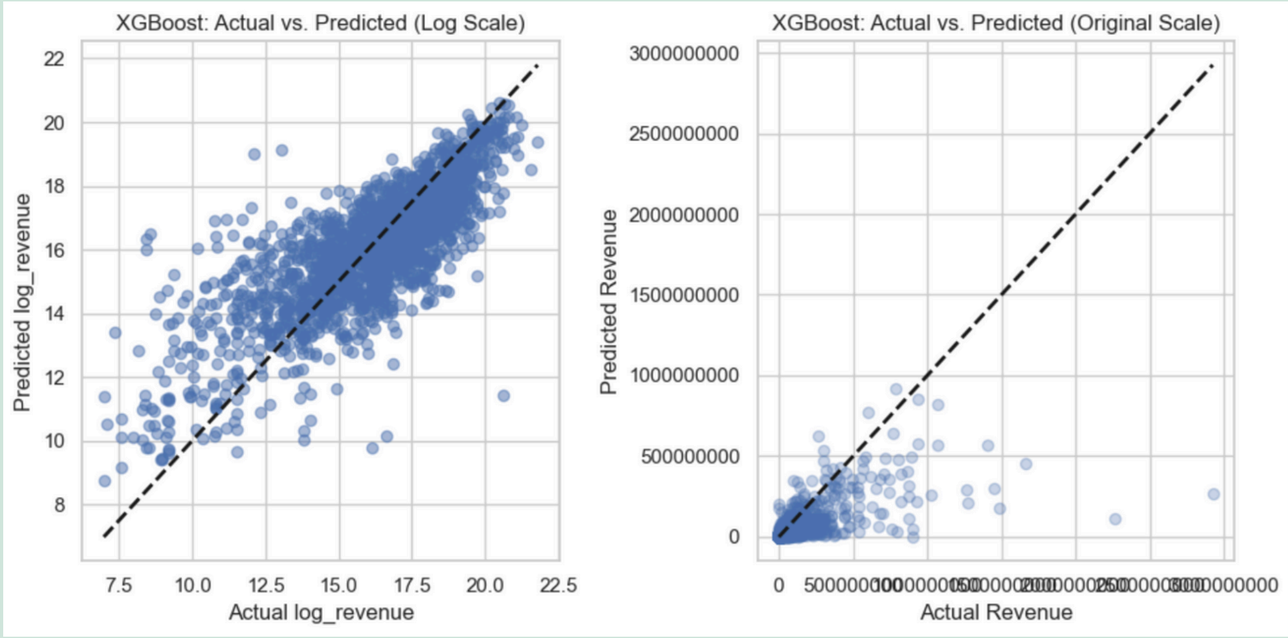
Random forest

R2 score (log): 0.6162

MÉTHODE 3

XGBoost

R2 score (log): 0.6265



6

DÉMO