Kaggle Prediction Competition

TMDB Box Office Prediction

Haoyue YUAN
https://github.com/aqwpm/TMDB_Prediction

Table de matières

Introduction

Dans un monde où l'industrie cinématographique a généré environ 41,7 milliards de dollars en 2018, comprendre les facteurs qui influencent les revenus au box-office est devenu crucial. Pour les cinéastes, les distributeurs, les investisseurs et les annonceurs, savoir ce qui fait ou défait un succès financier est vital. Alors, est-ce le réalisateur qui compte le plus, le budget de production ou peut-être même les mots-clés du scénario ? Cette étude vise à répondre à ces questions en utilisant des méthodes de machine learning pour prédire les revenus au box-office des films.

Objectif du Travail

L'objectif de ce travail est de construire un modèle prédictif capable de prédire les revenus mondiaux au box-office d'un film en utilisant un ensemble de métadonnées associées. Nous visons non seulement à réaliser un modèle performant mais aussi à comprendre les variables qui jouent un rôle significatif dans la détermination des revenus au box-office.

Nous aborderons ce sujet en suivant une méthode systématique :

- Analyse exploratoire des données (EDA) et prétraitement de l'ensemble de données pour en comprendre la structure, les relations et les anomalies.
- Ingénierie des caractéristiques: Utilisation des caractéristiques existantes pour en extraire ou combiner de nouvelles informations qui pourraient être utiles pour la prédiction.
- Suivi de la performance et arrêt anticipé: Ceci afin d'éviter les problèmes de surajustement du modèle. Nous examinerons également l'importance des variables pour vérifier si les variables en cours d'observation sont logiques.
- Réglage des paramètres des modèles et la modélisation: Pour optimiser leurs performances.
- Comparaison de différents modèles: Utilisation de métriques telles que le RMSE pour évaluer et comparer les performances de différents modèles.

À la fin, bien que nous ne puissions pas vérifier nos prédictions avec les vrais revenus de l'ensemble de données de test, nous fournirons néanmoins des prédictions basées sur les modèles que nous aurons développés.

Présentation du jeu de données

Source et Méthodologie de Collecte

Le jeu de données a été fourni dans le cadre d'une compétition Kaggle antérieure. Il est tiré de **The Movie Database (TMDB)** et rassemble des métadonnées sur un total de plus de 7 000 films. Pour cette étude, l'accent sera mis sur l'ensemble de données d'entraînement, qui contient des métadonnées sur un sous-ensemble de 3000 films.

Caractéristiques Principales des Variables

Le jeu de données est riche en caractéristiques qui peuvent être explorées et utilisées pour la prédiction. Voici une explication des attributs clés :

- **ID**: Identifiant unique de chaque film, sous forme d'entier.
- **Belongs_to_collection**: Informations sur la collection du film, si applicable, au format JSON.
- **Budget**: Budget du film en dollars. Une valeur de 0 indique que le budget est inconnu.
- **Genres**: Genres auxquels le film appartient, présentés au format JSON.
- **Homepage**: URL de la page d'accueil officielle du film.
- **Imdb_id**: Identifiant IMDB du film, permettant d'accéder à sa page IMDB.
- **Original_language**: Code à deux chiffres de la langue originale du film.
- **Original_title**: Titre original du film.
- **Overview**: Brève description du film.
- **Popularity**: Popularité du film, sous forme de nombre à virgule flottante.
- **Poster_path**: Chemin d'accès à l'affiche du film.
- **Production_companies**: Noms et identifiants TMDB des sociétés de production, au format JSON.
- **Production_countries**: Codes à deux chiffres et noms complets des pays de production, au format JSON.
- **Release_date**: Date de sortie du film, au format mm/jj/aa.
- **Runtime**: Durée totale du film, en minutes.
- **Spoken_languages**: Langues parlées dans le film, au format JSON.
- **Status**: Statut du film, indiquant s'il est sorti ou s'il est encore à l'état de rumeur.
- **Tagline**: Slogan du film.
- **Title**: Titre anglais du film.
- **Keywords**: Mots-clés associés au film, au format JSON.
- **Cast**: Distribution du film, au format JSON.
- **Crew**: Équipe du film, y compris les réalisateurs, les scénaristes et autres membres de l'équipe, au format JSON.
- **Revenue**: Revenus totaux générés par le film, en dollars.

Cette riche variété de caractéristiques offre une opportunité unique pour une analyse détaillée et la construction d'un modèle de prédiction robuste.

Exploratory Data Analysis (EDA)

Exploration initiale du jeu de données

Pour commencer notre analyse, nous avons chargé le jeu de données de formation train_df qui contient 2100 entrées et 23 colonnes. Ces colonnes varient en termes de types de données : on y trouve des entiers (int64), des flottants (float64) et des chaînes de caractères (object).

- id, budget, revenue sont des nombres entiers
- popularity, runtime sont des nombres à virgule flottante
- Le reste sont des types objet, qui pourraient être des chaînes de caractères ou des collections comme des listes.

Valeurs manquantes

Pour identifier les colonnes qui contiennent des valeurs manquantes, nous avons calculé et visualisé le nombre de valeurs **NA** par colonne pour le dataset d'entraînement (figure 1 à gauche) et pour le dataset du test (figure 1 à droite).

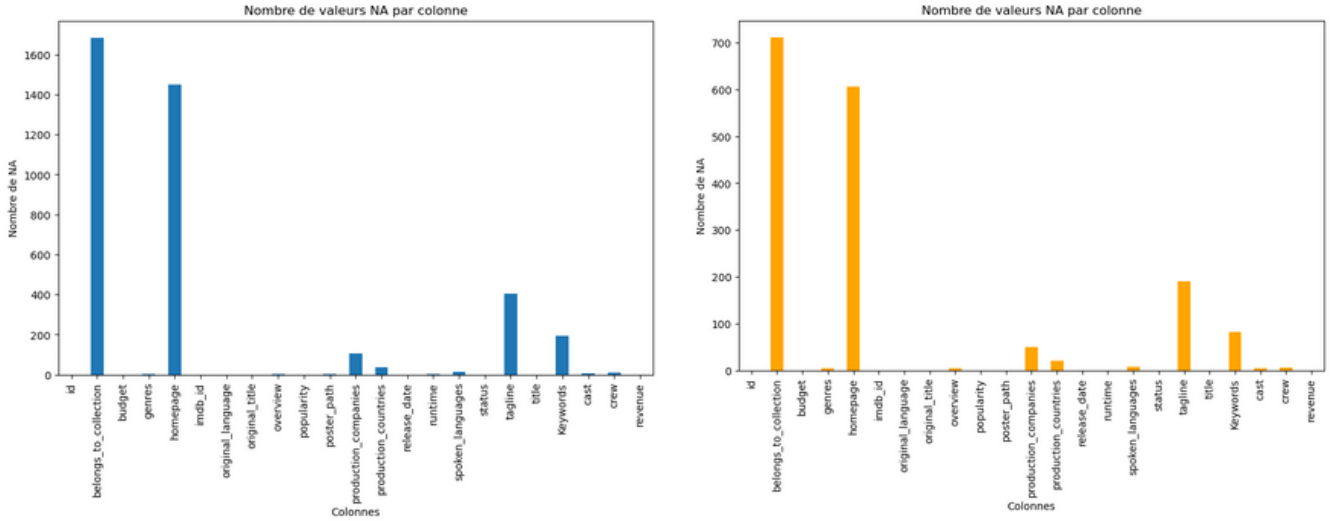

Figure 1

Nous constatons que les colonnes 'belongs_to_collection' , 'homepage' , 'homepage' ont relativement plus des valeurs manquante.

Correction des anomalies

Nous avons identifié et corrigé des valeurs aberrantes dans les colonnes **revenue** et **budget** à l'aide d'informations externes*.

Traitement des dates

Nous avons converti les dates de sortie des films au format datetime et extrait plusieurs nouvelles caractéristiques telles que release_year, release_month, release_day et release_dow (jour de la semaine).

Exploratory Data Analysis (EDA)

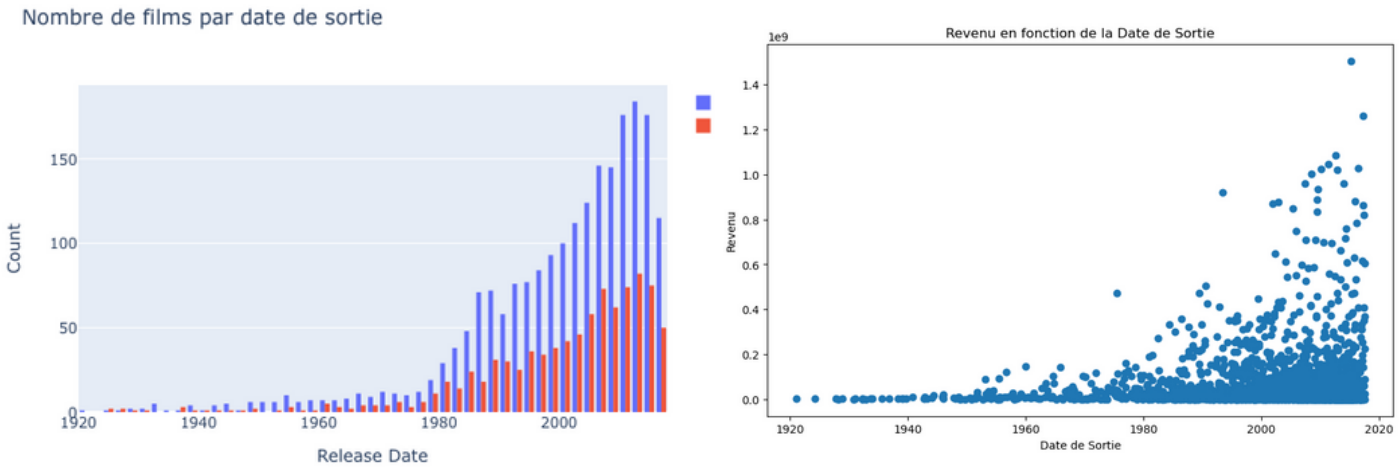

Figure 2

À partir de ce graphique interactif, il est clair que la majorité des films de la base de données ont été sortis entre les années 2000 et 2020. Cela soulève la question du biais d'échantillonnage dans la base de données. Étant donné le faible nombre de films sortis dans les années 1990, il est difficile d'établir une comparaison significative entre les films de cette décennie et ceux sortis entre 2000 et 2020.

En outre, il est intéressant de noter que la tendance du revenu généré par les films suit celle du nombre de films sortis au fil du temps. Ce constat est assez logique : plus le nombre de films produits augmente, plus les revenus générés sont susceptibles d'être élevés.

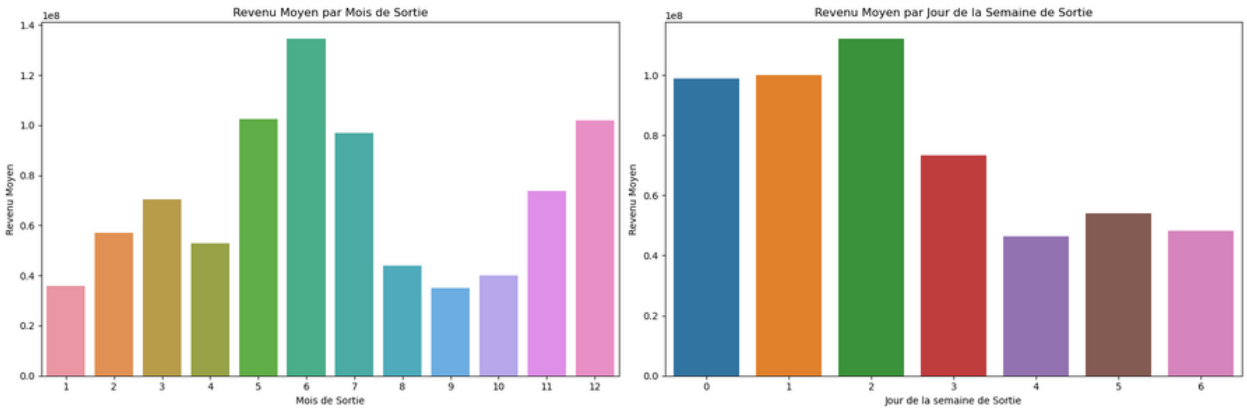

Figure 3

L'analyse des données révèle un pic dans le revenu moyen généré par les films durant les mois de juin, juillet et décembre. Cette tendance s'explique logiquement par le fait que ces périodes correspondent généralement aux vacances, pendant lesquelles les gens sont plus enclins à aller au cinéma. De plus, nous observons des pics de revenus moyens à certains jours de la semaine. Bien que notre jeu de données ne spécifie pas quels jours de la semaine ces pics surviennent, il est intuitif de supposer qu'il s'agit des week-ends, périodes où les gens ont généralement plus de temps libre pour des activités de loisir comme le cinéma.

Exploratory Data Analysis (EDA)

Relation entre les caractéristiques (de type numérique) et le revenu

Pour mieux comprendre comment différentes variables affectent le revenu, nous avons tracé des graphiques de dispersion pour les paires **Budget vs Revenue**, **Popularity vs Revenue** et **Runtime vs Revenue**.

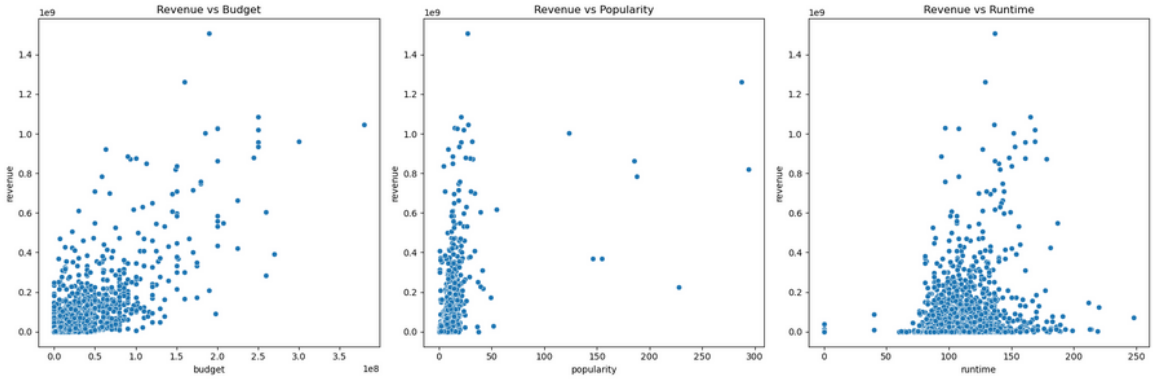

Figure 4

Selon le graphique de dispersion, il est manifeste qu'une corrélation positive existe entre le budget alloué à un film et les revenus qu'il génère. Cela suggère que des investissements plus importants dans la production d'un film sont généralement associés à des rendements financiers plus élevés. En revanche, le graphique ne montre pas de corrélation notable entre la popularité d'un film et ses revenus, ni entre sa durée et les recettes qu'il engendre. Ces observations remettent en question l'idée selon laquelle un film plus populaire ou plus long serait nécessairement plus lucratif.

Matrice de corrélation

Enfin, pour avoir un aperçu de la relation linéaire entre les différentes caractéristiques numériques, nous avons calculé la matrice de corrélation.

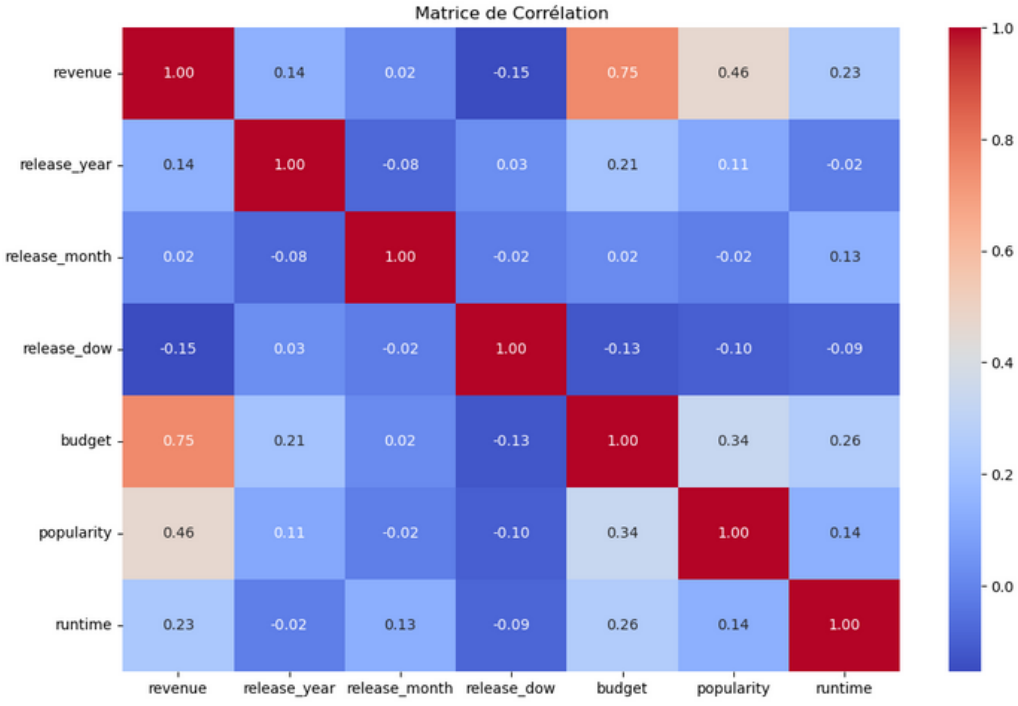

Figure 5

Les résultats confirment nos observations antérieures : il existe une forte corrélation entre le budget et les revenus, corroborant l'idée que des investissements plus élevés dans la production sont souvent associés à des rendements financiers plus importants. Par ailleurs, une corrélation modérée est également observable entre les revenus et la popularité du film, encore moins modérée pour la durée du film.

Ce travail préliminaire d'analyse exploratoire nous donne une meilleure compréhension du jeu de données, ce qui sera crucial pour la modélisation qui suivra.

Data Pre-processing

Traitement des Variables Non-Numériques

- **Transformation des Données JSON en Dictionnaires**

Avant de plonger plus profondément dans les analyses, il est crucial de traiter les variables qui ne sont pas de type numérique. Ce traitement est nécessaire car les modèles statistiques et d'apprentissage automatique fonctionnent généralement mieux avec des données numériques. En fait les données JSON sont souvent des collections d'objets qui peuvent être difficiles à analyser sans un prétraitement approprié donc le format JSON non-structuré dans certaines colonnes doit être transformé en une forme plus manipulable pour les analyses ultérieures.

Nous avons identifié les colonnes suivantes comme contenant potentiellement des données JSON : ['**belongs_to_collection', 'genres', 'production_companies', 'production_countries', 'spoken_languages', 'Keywords', 'cast', 'crew'**], et nous allons transformer ces données en dictionnaire.

Une fois ce prétraitement terminé, nous examinerons les relations entre ces variables nouvellement transformées et le revenu du film, afin de déceler des schémas ou des facteurs qui pourraient influencer les performances financières d'un film.

- **Conversion des Dictionnaires en Colonnes Individuelles**

Après avoir converti les chaînes de caractères en dictionnaires pour certaines variables, l'étape suivante de notre prétraitement est de décomposer ces dictionnaires en des colonnes individuelles. Ce processus facilite grandement la modélisation subséquente car il convertit les données catégorielles complexes en un format numérique qui peut être directement utilisé dans des modèles statistiques ou d'apprentissage automatique.

Bien que cette technique offre plusieurs avantages pour la modélisation, elle présente également des inconvénients importants. Le plus notable est sans doute l'augmentation de la dimensionnalité de l'ensemble de données. En créant une colonne distincte pour chaque valeur unique de chaque variable catégorielle, le nombre total de colonnes (features) peut augmenter considérablement.

Analyse des Relations avec le Revenu

- **Belong_to collection**

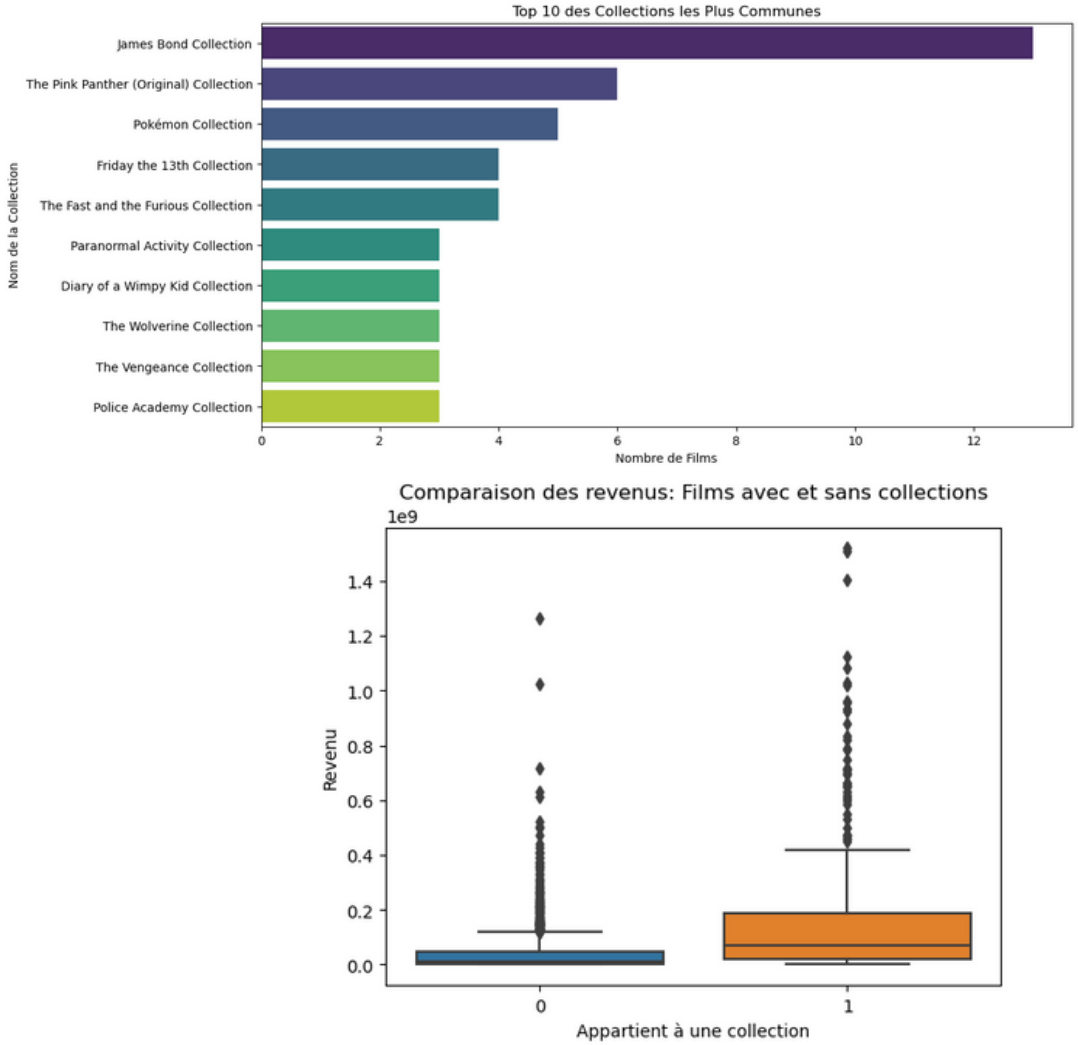

Figure 6

Il est évident que les films qui sont avec collections, ont les revenus plus hautes par rapport aux ceux qui sans collections.

- **production_countries**

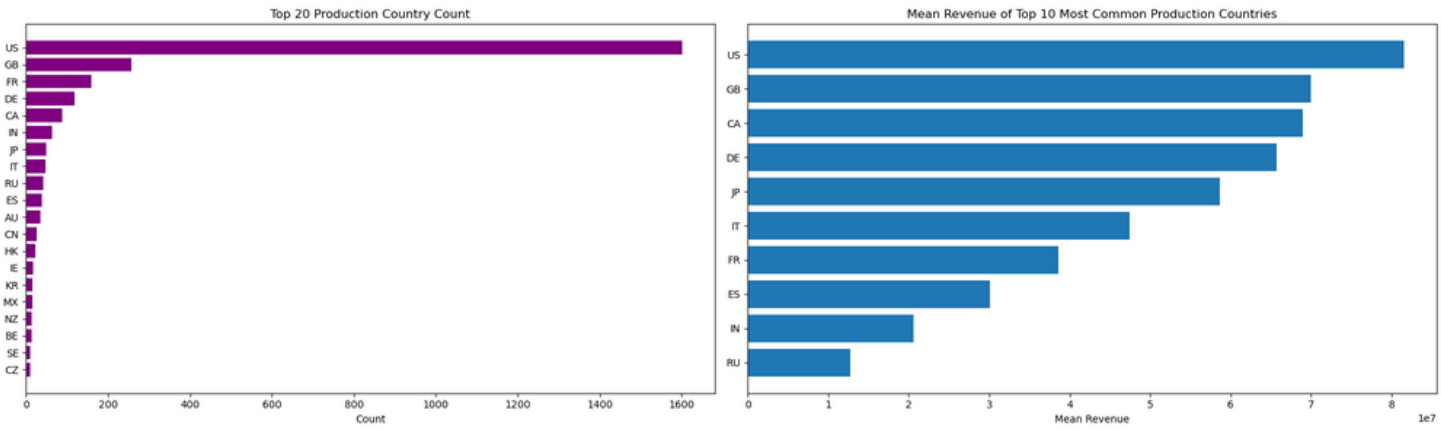

Figure 7

Assez logique comme il y a plus de film fabriqués par américains donc reçoivent le plus gros succès.

- **Keywords**

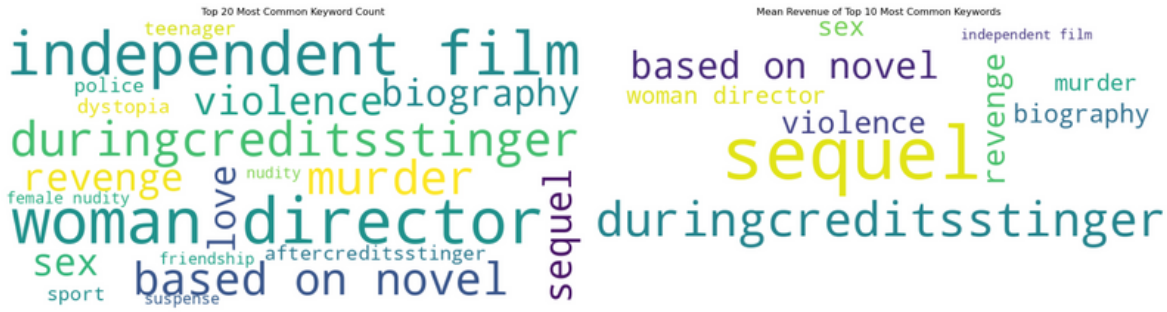

Figure 8

Nous pouvons également voir certains mot-clé ont souvent plus grand succès, nous observons selon le wordcloud les mots comme '**sequel', 'violence'**, etc.

Nous allons répéter ce technique pour toutes les autres variables mentionnées précédemment.

Choix et entraînement des modèles

Après avoir prétraité nos données, nous passons à une étape importante de notre analyse : le choix et l'entraînement des modèles de machine learning. Cette partie décrit les différentes méthodes et approches adoptées pour la modélisation.

- **Analyse de la Distribution des Variables d'Intérêts**

Avant de plonger dans la modélisation, il est essentiel d'analyser la distribution de nos variables cibles et d'autres caractéristiques importantes.

Distribution du Revenu

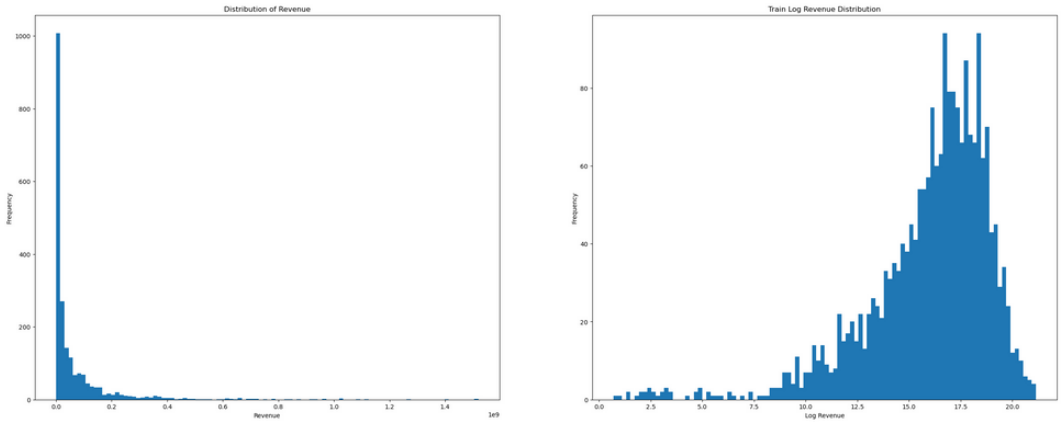

Figure 9

Le coefficient d'asymétrie (Skewness) du revenu est de 4,8, ce qui indique une distribution fortement asymétrique.

Distribution des Variables Biaisées : Budget et Popularité

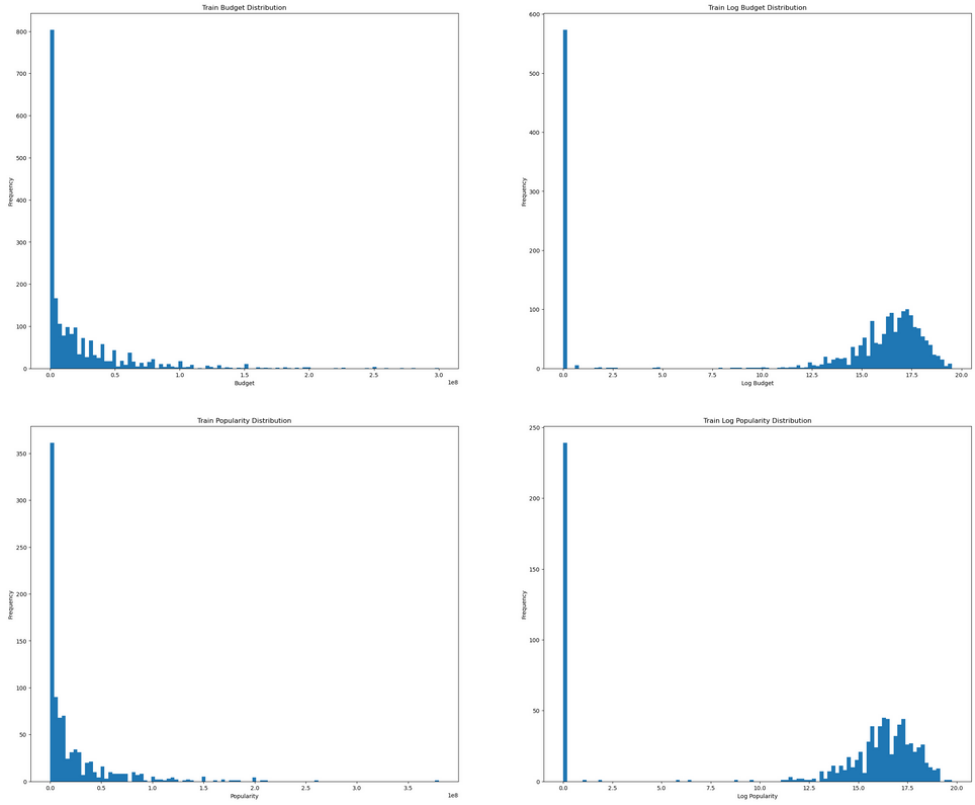

Figure 10

Nous avons également étudié la distribution des variables budget et popularité, et nous avons trouvé que ces variables sont également asymétriques, avec des coefficients d'asymétrie respectifs de 3,0 et 15,0 pour le jeu d'entraînement et de 3,3 et 10,6 pour le jeu de test.

- **Transformation Logarithmique**

Pour rendre nos variables plus "normales", nous avons appliqué une transformation logarithmique. La transformation logarithmique est bénéfique car elle permet de réduire l'effet des valeurs aberrantes et de rendre la distribution plus proche d'une distribution normale.

Nous avons ensuite également **remplacé des valeurs NaN et Inf par zéro** et **créé diverses nouvelles caractéristiques** qui pourraient être informatives pour nos modèles puis nous avons **supprimé les colonnes** qui ne sont pas nécessaires pour la modélisation. Nous avons enfin **séparé nos données en ensembles d'entraînement et de validation** pour évaluer la performance de nos modèles futurs.

Avec ces préparatifs en place, nous sommes maintenant prêts à passer à l'étape suivante, qui consiste à sélectionner et entraîner nos modèles de machine learning.

Choix du modèle : LightGBM

LightGBM (Light Gradient Boosting Machine) est une bibliothèque pour l'entraînement de modèles basés sur des arbres de décision boostés. Le terme "light" souligne que cette bibliothèque est extrêmement rapide et consomme moins de ressources par rapport à d'autres implémentations d'arbres boostés. Il est particulièrement efficace pour les problèmes de grande dimension et de grande envergure.

Dans le contexte de notre objectif, qui est la prédiction du revenu des films, LightGBM présente plusieurs avantages. La nature complexe et à grande échelle de notre ensemble de données, avec de nombreuses caractéristiques de types différents (numériques, catégorielles, texte, etc.), fait de LightGBM un choix adapté.

Choix du modèle : XGBoost

XGBoost (eXtreme Gradient Boosting) est également une bibliothèque très efficace pour l'entraînement de modèles basés sur des arbres de décision boostés. Il est bien connu pour sa robustesse et sa capacité à produire des modèles performants. Il offre une multitude d'options de régularisation pour prévenir le surajustement, ainsi que plusieurs fonctionnalités pour le réglage des hyperparamètres, la gestion des valeurs manquantes et l'optimisation de l'efficacité computationnelle. Il est souvent utilisé dans des solutions gagnantes de compétitions de science des données et est largement apprécié pour sa flexibilité et sa polyvalence.

le choix de XGBoost pour nous repose sur sa robustesse, sa précision, et sa flexibilité, ce qui en fait un outil très approprié pour la prédiction du revenu des films dans le cadre de notre étude.

- **Entraînement du modèle**

Pour entraîner notre modèle, nous avons divisé notre ensemble de données en deux parties : un ensemble d'entraînement et un ensemble de validation. L'ensemble d'entraînement est utilisé pour ajuster les paramètres du modèle, tandis que l'ensemble de validation nous aide à évaluer les performances du modèle pendant la phase d'entraînement. Cette approche nous permet de vérifier comment le modèle performe sur des données qu'il n'a jamais vues, offrant ainsi une évaluation plus réaliste de sa performance.

- **Stratégie d'Arrêt Anticipé**

Pour éviter le surajustement, c'est-à-dire une situation où notre modèle apprend "par cœur" les données d'entraînement au détriment de sa performance sur de nouvelles données, nous avons employé une technique appelée "arrêt anticipé". Cette technique consiste à arrêter l'entraînement si le modèle ne s'améliore pas pendant un certain nombre d'itérations consécutives sur l'ensemble de validation. Dans notre cas, nous avons défini ce nombre à 200 itérations. Cela signifie que si le modèle ne parvient pas à améliorer sa performance sur l'ensemble de validation pendant 200 itérations, l'entraînement s'arrête prématurément.

- **RMSE (l'Erreur Quadratique Moyenne Racine)**

RMSE est généralement associée à l'évaluation du modèle. Nous avons utilisé RMSE pendant la phase d'entraînement qui nous permet d'obtenir un retour instantané sur la qualité du modèle. Si le RMSE sur l'ensemble de validation commence à augmenter ou cesse de diminuer, cela peut être un signe que le modèle est en train de surajuster les données d'entraînement.

Résultats

Poids des caractéristiques du modèle

- **LightGBM**

Pour le modèle LightGBM, l'entraînement s'est arrêté à la meilleure itération, avec un RMSE de 2.1492 sur l'ensemble de validation. Les caractéristiques les plus importantes pour ce modèle sont **_year_to_log_budget** (0.2979), **_budget_year_ratio** (0.1000), et **release_year** (0.0682), entre autres.

- **XGBoost**

Le modèle XGBoost a obtenu un RMSE de 2.04028 sur l'ensemble de validation. Les caractéristiques les plus influentes étaient **_budget_year_ratio** (0.1900), **_year_to_log_popularity** (0.0345), et **log_budget** (0.0276).

Comparaison des modèles

	Id	PredictedRevenue_model_lgb	PredictedRevenue_model2_xgb	revenue
0	2152	23479512	18428486	13670688
1	2170	3554525	6337898	23800000
2	193	1201419	650157	37915971
3	890	760124	510811	13102295
4	2579	1788123	3757451	2200000

Figure 11

RMSE

- LightGBM : 75,652,777.17
- XGBoost : 74,074,840.48

Histogramme des erreurs de prédiction

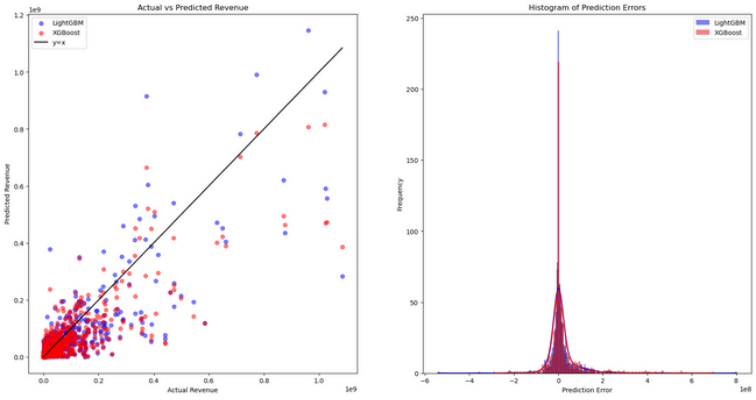

Figure 12

Matrice de corrélation.

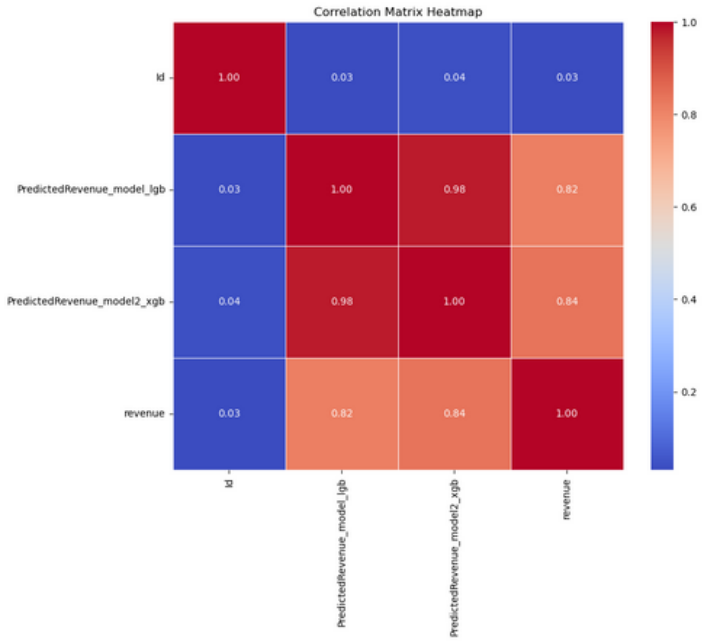

Figure 13.

Residus

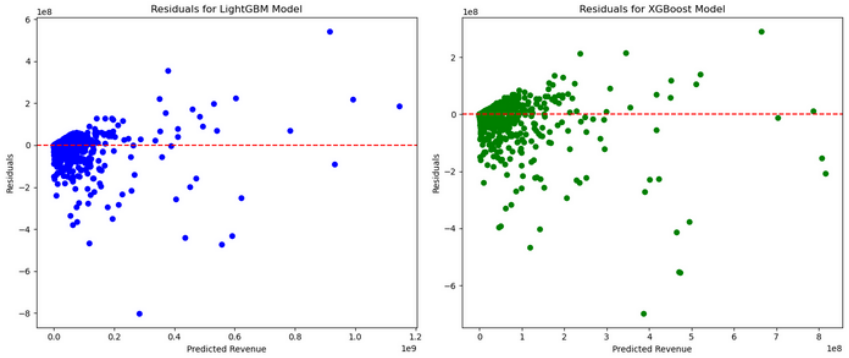

Figure 14

Test statistique

Le test t de Student pour les erreurs résiduelles des deux modèles a donné une valeur de t-statistique de 2.959 et une valeur p de 0.0032.

Interpretation

Il est à noter que **_budget_year_ratio** est une caractéristique clé pour les deux modèles, suggérant que la relation entre le budget et l'année de sortie a une influence significative sur le revenu des films.

Le modèle XGBoost a un RMSE légèrement plus faible que le modèle LightGBM, indiquant qu'il pourrait être légèrement plus précis dans ce cas.

Les distributions des erreurs pour les deux modèles semblent similaires, mais des analyses plus détaillées sont nécessaires pour tirer des conclusions plus fermes.

La matrice de corrélation offre un aperçu des relations linéaires entre les différentes caractéristiques. Elle peut être utile pour identifier les caractéristiques qui pourraient être redondantes ou hautement corrélées.

Le graphe des résidus montre comment les erreurs sont réparties autour de la ligne zéro. Une distribution aléatoire des points autour de la ligne zéro est généralement un bon signe, indiquant que le modèle est bien ajusté.

Avec un seuil de signification de 0.05, la valeur p inférieure à ce seuil indique qu'il y a une différence statistiquement significative entre les performances des deux modèles.

Conclusion

Nos analyses ont démontré que les deux modèles, LightGBM et XGBoost, sont tous deux performants et affichent des erreurs assez similaires. Cependant, il est important de noter que, même si les performances sont proches, le modèle LightGBM s'est révélé légèrement plus précis en termes de RMSE. Cette observation est statistiquement significative, comme l'a confirmé le test t de Student.

Globalement, il est encourageant de voir que les deux modèles ont réussi à fournir de bonnes prédictions sur un ensemble de données complexe et à grande échelle. Cela valide la robustesse et la fiabilité des deux techniques de boosting d'arbres pour les tâches de prédiction en général, et pour la prédiction du revenu des films en particulier.

Néanmoins, il convient de mentionner que des opportunités d'amélioration existent. Par exemple, l'introduction de nouvelles caractéristiques, telles que les données sur le marketing, les avis des critiques ou même des données temporelles plus granulaires, pourrait potentiellement améliorer les performances des modèles. De plus, l'exploration d'autres algorithmes et techniques de modélisation pourrait également offrir des voies d'amélioration.

En résumé, bien que nos modèles soient performants, le domaine de la prédiction du revenu des films reste un champ ouvert à de nouvelles recherches et améliorations.