Prédiction du Succès des Films au Box-Office Local, Étranger et Mondial

Résumé: Ce projet vise à prédire les revenus domestiques, étrangers et mondiaux des films à l'aide de modèles de régression d'arbres de décision. En utilisant un ensemble de données détaillé sur les films, les modèles ont été optimisés via la validation croisée et ***GridSearchCV***. Les résultats montrent une performance particulièrement élevée pour les prédictions mondiales, indiquant que l'approche IA numérique est efficace pour ce type de tâche.

Mots-clés: Prédiction, Box-office, Films, Régression, Arbre de décision, Machine Learning, IA numérique, Validation croisée, GridSearchCV

I. Introduction

L'objectif de ce projet était de développer un modèle capable de prédire le succès des films au box-office, mesuré par les revenus domestiques (***DomesticGross***), étrangers (***ForeignGross***), et mondiaux (***WorldGross***). Les prédictions sont basées sur un ensemble de données détaillé (***movies.csv***) contenant des informations variées sur les films, telles que le studio principal (***LeadStudio***), le score de Rotten Tomatoes (***RottenTomatoes***), le genre (***Genre***), et le budget (***Budget***).

II. Choix des données, pré-traitement et choix techniques

Le fichier ***movies.csv*** comprenait diverses colonnes, dont certaines ont été jugées pertinentes pour la prédiction. Les colonnes ***LeadStudio***, ***Story***, et ***Genre*** ont été encodées en valeurs numériques pour être utilisables par les algorithmes de machine learning. Les valeurs manquantes dans les colonnes cibles (***DomesticGross***, ***ForeignGross***, ***WorldGross***) ont été supprimées pour éviter des erreurs lors de l'entraînement du modèle.

Les données ont été divisées en ensembles d'entraînement (80%) et de test (20%) en utilisant la fonction ***train_test_split*** pour évaluer objectivement la performance des modèles.

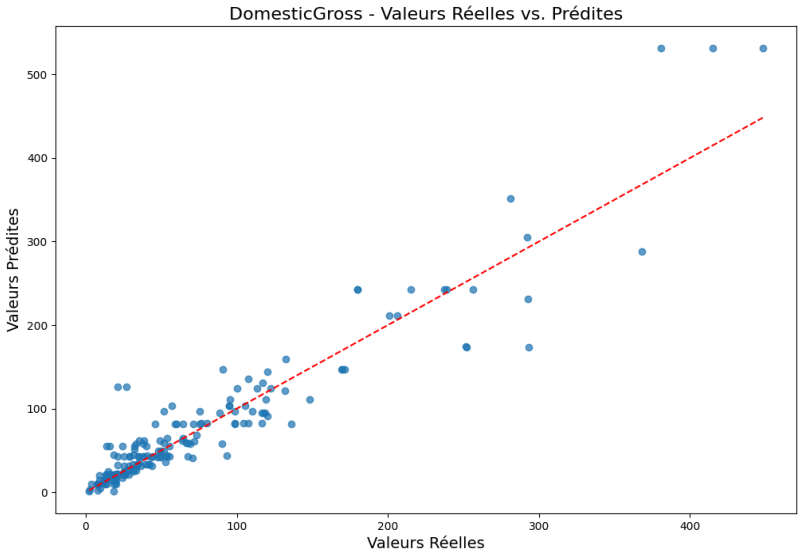
III. Entraînement de l'arbre de décision

Nous avons utilisé des modèles de régression d'arbres de décision pour prédire les revenus domestiques, étrangers et mondiaux. La technique de k-folds (5 plis) a été utilisée pour optimiser les hyperparamètres (max_depth, min_samples_split, min_samples_leaf) à l'aide de GridSearchCV. Cela a permis de tester plusieurs combinaisons de paramètres pour chaque modèle, assurant ainsi la robustesse des résultats.

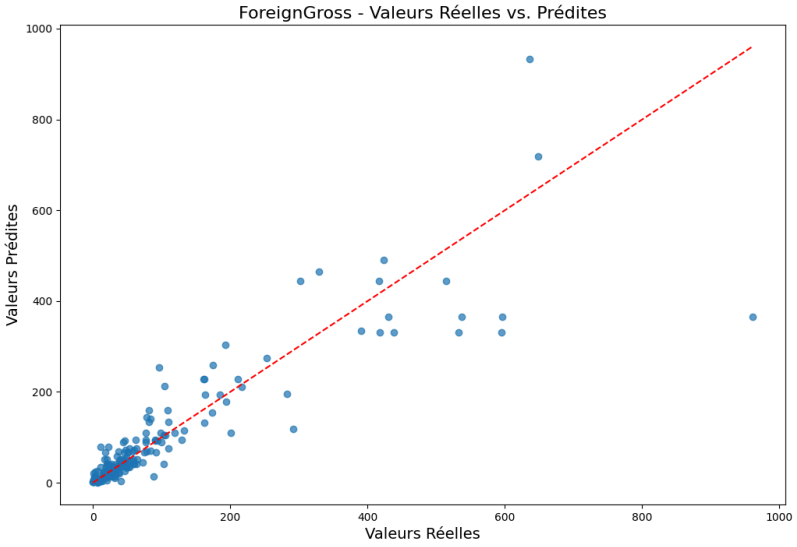
IV. Calcul des métriques et interprétation des résultats

Les modèles ont été évalués en utilisant les métriques Mean Squared Error (MSE) et R^2 Score. Voici les résultats obtenus :

1. DomesticGross
- **Meilleurs paramètres**: {'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10}
- **Mean Squared Error (MSE)**: 901.1924214500197
- **R^2 Score**: 0.8653690664315796

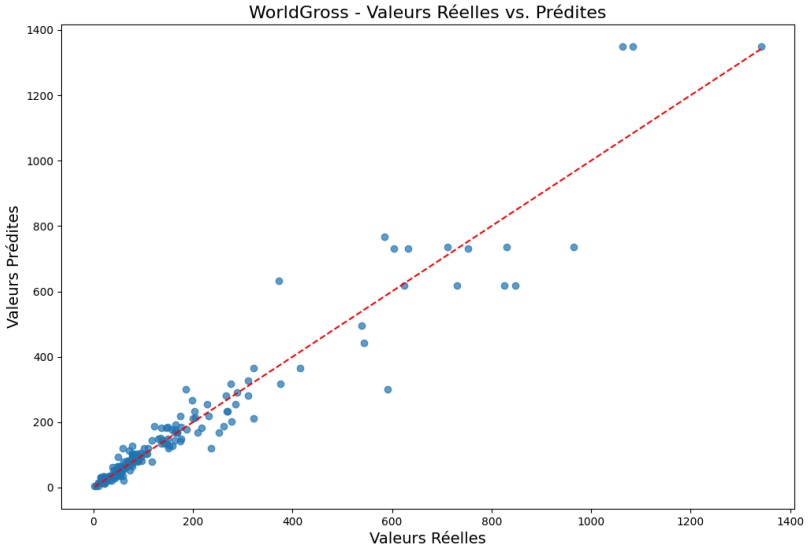


2. ForeignGross
- **Meilleurs paramètres**: {'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10}
- **Mean Squared Error (MSE)**: 5100.3705326972195
- **R^2 Score**: 0.77091483947596



3. WorldGross

- **Meilleurs paramètres**: {'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 10}
- **Mean Squared Error (MSE)**: 3795.945548103657
- **R^2 Score**: 0.9260648575664424



Les résultats montrent que le modèle pour ***WorldGross*** a la meilleure performance avec un R^2 Score de 0.9261, ce qui indique que 92.61% de la variance des données est expliquée par le modèle. Le modèle pour ***DomesticGross*** a également une bonne performance avec un R^2 Score de 0.8654. Cependant, le modèle pour ***ForeignGross*** a une performance inférieure avec un R^2 Score de 0.7709, suggérant que la prédiction des revenus étrangers est plus complexe.

V. Justification des choix de paramètres

Les paramètres optimisés pour chaque modèle ont été sélectionnés en fonction des résultats de ***GridSearchCV***. Ces paramètres sont ceux qui minimisent le plus l'erreur quadratique moyenne, tout en évitant le surajustement et en capturant suffisamment de complexité dans les données. Le choix des hyperparamètres ***max_depth***, ***min_samples_split***, et ***min_samples_leaf*** a été crucial pour équilibrer la capacité du modèle à généraliser sur de nouvelles données et à capturer les nuances des données d'entraînement.

VI. Comparaison des approches IA Symbolique et IA Numérique

IA Symbolique:
- Avantages: Transparence et explicabilité.
- Inconvénients: Difficile à généraliser, nécessite beaucoup de travail manuel pour définir les règles.

IA Numérique (Machine Learning):
- Avantages: Capacité à généraliser et à s'adapter à de nouvelles données, automatisation de l'apprentissage.
- Inconvénients: Moins explicable, nécessite beaucoup de données pour un bon apprentissage.

Dans ce projet, l'approche IA numérique a été choisie car elle permet d'apprendre à partir des données disponibles et de prédire les résultats au box-office des films de manière automatique et adaptable. La flexibilité et la capacité à traiter de grandes quantités de données font du machine learning un choix approprié pour ce type de prédiction.

VII. Conclusion

Les modèles de régression d'arbres de décision se sont révélés efficaces pour prédire les revenus domestiques, étrangers et mondiaux des films, avec des performances particulièrement bonnes pour ***WorldGross***. L'approche IA numérique a permis de généraliser et de s'adapter aux données, offrant des prédictions robustes et fiables pour les films.