Note Explicative sur le Projet de Recherche

1. Contexte et Question de Recherche

Dans le cadre de ce projet, nous avons choisi de faire une étude sur des données concernant le cinéma. Nous avons donc choisi un jeu de donné IMDB, puis nous nous sommes naturellement intéressé aux notes des films. On s'est donc posé la question :

" Quels éléments pour influencer la note d'un film"

Cette problématique peut être l'objet par exemple de questionnement d'un producteur qui souhaite produire un film à succès. Il faut cependant nuancer, le fait qu'un film ait une bonne note ne signifie pas grande recette.

Nous avons donc décidé de faire une conception et réalisation d'un outil permettant de prédire la note d'un film en fonction de 3 paramètres (durée, genre, et budget). Pour ce faire, il nous faut bien choisir nos analyses statistiques ainsi que notre modèle de prédiction.

2. Choix du modèle et justification

Notre objectif étant de pouvoir prédire la note d'un film en fonction de 3 paramètres. La variable qu'on cherche (la note) est donc dépendante des autres variables. De plus, les variables n'agissent pas de manière indépendante. Par exemple, les films d'animation sont souvent plus courts et moins coûteux que des blockbusters.

Le meilleur choix pour notre modèle est donc une régression linéaire pour prendre en compte tous les paramètres.

3. Méthodologie et traitement des données

Tout d'abord nous avons récupéré un fichier .CSV de la base de données du site IMDB, contenant un jeu de données de films. Chaque ligne du fichier contient un film et ses caractéristiques. Malheureusement, le jeu de données ne contenait pas le budget de chaque film. Il nous fallait donc trouver un moyen d'avoir cette valeur car il nous paraissait évident qu'elle était importante pour la prédiction.

Nous avons donc récupéré cette valeur grâce à une API. Après avoir concaténé les budgets avec le jeu de donnée, notre fichier était prêt à être nettoyé.

Pour pouvoir utiliser le jeu pour notre modèle il nous fallait nettoyer le jeu pour supprimer les colonnes inutiles, renommer les colonnes nécessaire, suppression des lignes incomplète et la suppression des lignes avec un nombre de vote trop faible.

Nous avons ensuite transformé la colonne textuelle des genres en une série de colonnes numériques binaires valant 0 ou 1, car l'algorithme de régression ne peut traiter que des chiffres. Nous avons ensuite combiné ces indicateurs de genre avec le budget et la durée des films pour constituer nos variables explicatives, tout en définissant la note moyenne comme la cible à prédire. Enfin, ce jeu de données nettoyé a été divisé aléatoirement en deux parties : 80 % pour entraîner le modèle à calculer les corrélations, et 20 % réservés pour tester sa fiabilité sur des films inconnus.

Le script nous permet d’accéder au modèle avec une interface web. Ce qui nous permet de prédire la note d’un film en fonction des paramètres choisis.

4. Résultats et interprétation

Nos résultats indiquent que le modèle explique environ 21 % de la note finale (R2=0.21) ce qui confirme que le succès d'un film repose avant tout sur sa qualité artistique (scénario, jeu d'acteur) plutôt que sur sa fiche technique. Cependant cela peut aussi être lié au fait que notre jeu de données ne contiendrait pas suffisamment de données pour entraîner le modèle. Si l'analyse a prouvé, contre toute attente, que le budget n'a aucune influence sur la qualité perçue, elle a en revanche mis en évidence que les films plus longs et certains genres spécifiques (comme l'Animation ou l'Histoire) partent avec un avantage significatif.

5. Limites et pistes d’amélioration

La principale limite réside dans notre coefficient de détermination R2 de 21%. Cela signifie que près de 80% de la note d'un film échappe à notre algorithme. Cette ombre s'explique par l'absence de variables clés ou le manque de lignes dans notre jeu de données. En effet, notre régression ne prend pas en compte la notoriété des acteurs, la réputation du réalisateur ni la qualité du scénario qui sont pourtant des moteurs essentiels de la réception critique.