Rapport Data Challenge d'Apprentissage supervisé - Problème de régression

Said OUGOUADFEL, Simon LEGRIS, François SLAWNY
November 2025

1. Présentation des problèmes

L'objectif de ce projet est de prédire la popularité d'un morceau de musique à partir de ses caractéristiques audio et métadonnées issues d'un jeu de données Spotify. La variable cible est « popularity », celle-ci est continue, ce qui en fait un problème de régression supervisée. L'enjeu est d'obtenir le meilleur score R^2 possible tout en justifiant rigoureusement les choix méthodologiques. Les modèles testées ont premièrement été ceux vus en cours afin de valider la maîtrise de ces modèles simplifier : SVM, Forêts Aléatoires, Gradient Boosting, AdaBoost etc. Puis pour optimiser notre score nous avons ensuite étendu nos tests à des modèles plus industriels de Boosting comme LightLGBM, CatBoost et XGBoost.

2. Prétraitements éventuels sur les données (+0.08 à +0.12 de score R^2 public)

Premièrement des tests statistiques ont été réalisés pour étudiées les types de corrélations entre nos variables. Mais avant d'appliquer les tests statistiques d'association et de dépendance, les hypothèses fondamentales nécessaires à leur validité ont été systématiquement vérifiées. Ces vérifications préalables ont permis de s'assurer que les méthodes statistiques employées reposent sur des hypothèses empiriquement satisfaites et que les conclusions qui en découlent sont fondées.

L'hypothèse d'indépendance des observations a été considérée comme respectée, les données correspondant à des morceaux distincts sans duplication d'artiste ou d'album. Cette condition garantit la validité des tests statistiques employés, qui reposent tous sur l'hypothèse d'indépendance entre les unités d'observation.

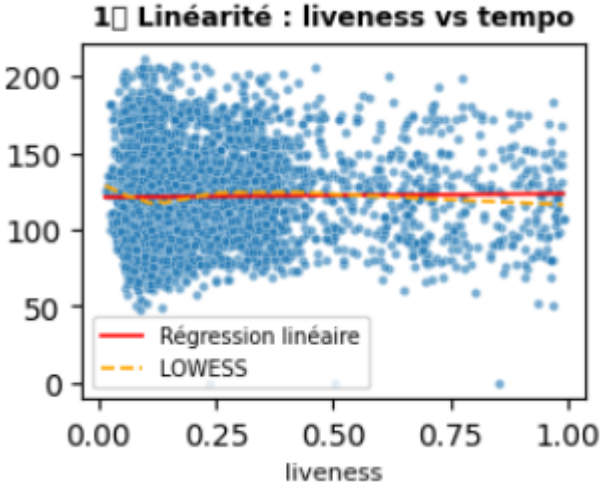
Enfin un enrichissement à partir de données externes à été réalisé pour optimiser le score final.

Tests statistiques et Préprocesseurs (+0.05 à +0.08 de score R^2 public)

Itération 1 sur les corrélations entre variables continues

Vérification des hypothèses préalables:

Il à été vérifié pour chaque couple si ils étaient reliés linéairement de manière significative.

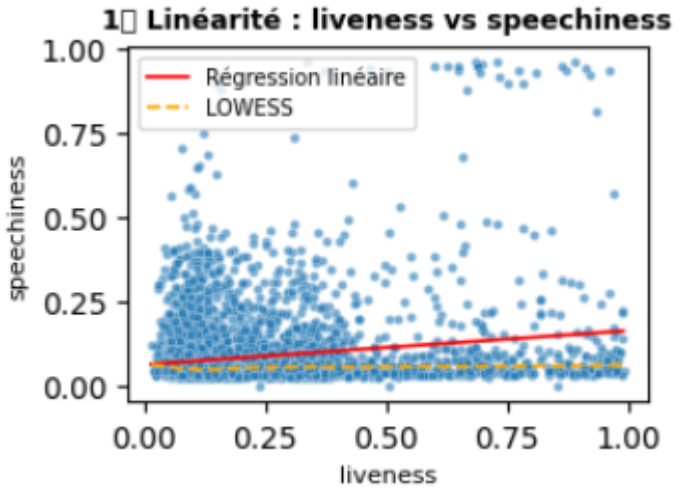


Ce nuage de points montre la relation brute entre les deux variables continues.

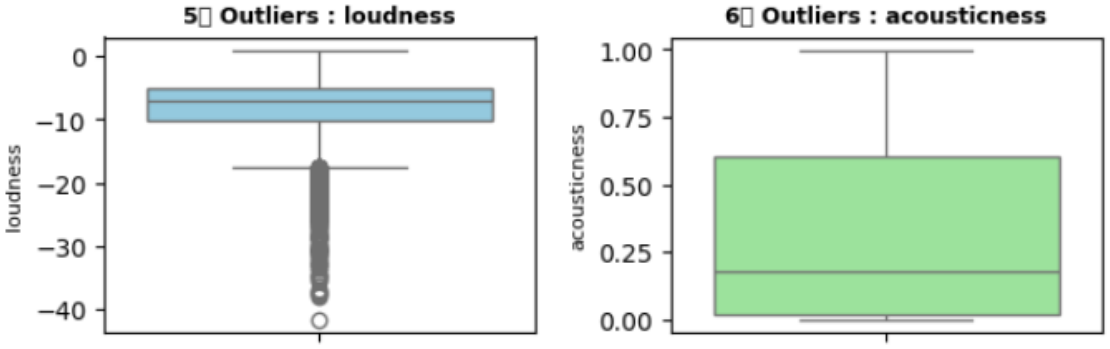
La droite rouge représente la régression linéaire et la ligne orange la courbe LOWESS (lissage non linéaire).

Les deux lignes doivent être proches, indiquant une relation à peu près linéaire entre valence (émotion positive du morceau) et danceability (caractère dansant).
Si la courbe LOWESS s'écarte fortement de la droite, cela signifie une relation non linéaire, et le test de Pearson ne serait plus approprié (il faudrait plutôt Spearman).

Ici, la tendance est globalement linéaire et croissante, donc l'hypothèse de linéarité est vérifiée.



Ici, les droites ne se superposent pas, la relation n'est pas linéaire.



Ces graphiques permettent d'évaluer la symétrie des distributions et la présence d'outliers (valeurs extrêmes).

Ce qu'on cherche :
Des distributions à peu près symétriques et peu d'outliers marqués.
Des points isolés (cercles sous le boxplot) indiquent des valeurs atypiques qui peuvent influencer la corrélation.

Ici la variable acousticness semble assez symétrique, sans outliers importants mais la variable loudness présente beaucoup d'outliers (ronds sous la boîte).

Ces outliers font qu'un test de Pearson n'est plus envisageable car la corrélation de Pearson est très sensible aux valeurs extrêmes.

Quelques points aberrants peuvent complètement modifier la pente de la droite de régression et donc gonfler ou inverser le coefficient de corrélation.

Tests statistiques appliqués:

Les corrélations entre variables continues ont été étudiées à l'aide du coefficient de Pearson lorsque la relation paraissait linéaire, et du coefficient de Spearman lorsqu'elle semblait monotone sans suivre de structure strictement linéaire.
Ces analyses ont mis en évidence des corrélations fortes, atteignant des valeurs proches de 0,7 en valeur absolue, notamment entre les variables energy, loudness et acousticness, suggérant une redondance d'information acoustique.

Conséquences sur les choix de prétraitement:

Les variables energy, loudness et acousticness, fortement corrélées entre elles, ont premièrement été combinées en une seule composante synthétique nommée sound_profile à l'aide d'une réduction de dimension par PCA.
Cette approche permet de réduire la redondance tout en préservant la dimension sonore la plus représentative des morceaux mais celle-ci réduisait légèrement le R^2 score et n'a donc pas été gardée pour le script final

Itération 2 sur les corrélations entre variables continues et catégorielles

Vérification des hypothèses préalables:

La normalité des distributions a d'abord été testée pour l'ensemble des variables continues telles que energy, loudness, danceability, acousticness, valence, tempo ou encore duration_ms.

Le test de Shapiro–Wilk a été appliqué sur chaque couple de variables continue - catégorielle et les résultats ont indiqué que la majorité des distributions pour chaque variable continue dans chaque sous-groupe des variables catégorielles ne suivaient pas une loi normale (p-valeurs inférieures à 0,05).

Ce constat remet en cause l'utilisation de tests paramétriques classiques qui ont déjà été vus en M1 comme le test de Pearson ou l'ANOVA, et a conduit à privilégier des approches non paramétriques, plus robustes face à des distributions asymétriques.

L'homogénéité des variances a ensuite été examinée à l'aide du test de Levene. Les résultats ont révélé une hétérogénéité importante des variances entre groupes, ce qui confirme la nécessité d'utiliser des tests non paramétriques pour les comparaisons inter-catégories, tels que le test de Kruskal–Wallis.

Tests statistiques appliqués:

L'influence des variables catégorielles sur les variables continues a été évaluée par le test de Kruskal–Wallis, plus approprié que l'ANOVA dans un contexte de non-normalité et d'hétéroscédasticité.
Les résultats ont montré des différences significatives de distribution pour la majorité des couples de variables testés (p-valeur proches de zéro), indiquant que les variables catégorielles telles que key, mode, time_signature, explicit et track_genre exercent une influence statistiquement significative sur les variables continues.

Itération 3 sur les corrélations entre variables catégorielles

Pas besoin de vérification des hypothèses préalables

Tests statistiques appliqués:

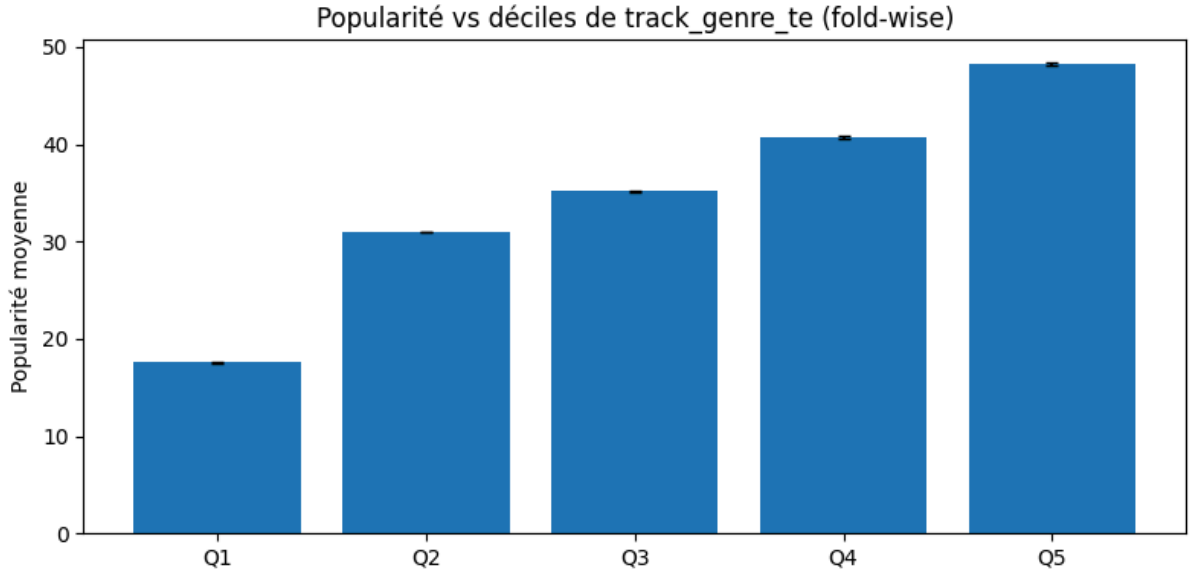
Les relations entre variables catégorielles ont été explorées à l'aide du test du Khi² d'indépendance, et la force de ces associations a été mesurée par le coefficient de Cramér's V.
Les résultats, généralement inférieurs à 0,4, ont mis en évidence des dépendances modérées entre certaines variables qualitatives, suggérant des interactions sans véritable redondance structurelle.

Conséquences final sur les choix de prétraitement pour les variables catégorielles

Le préprocesseur final, combine standardisation, imputation robuste et encodages adaptés à la nature des variables.

Les variables continues sont standardisées après imputation par la médiane, tandis que les variables catégorielles à faible cardinalité (key, mode, time_signature, explicit) sont encodées par one-hot encoding.

La variable track_genre, de très forte cardinalité, a été encodée par une moyenne lissée (target encoding) calculée de manière fold-wise afin d'éviter toute fuite de la cible entre les ensembles d'apprentissage et de validation.



Comme on peut le voir ici la popularité moyenne varie en fonction des différents genre, d'où l'utilité de modéliser cette cette variation par une nouvelle variable track_genre_te.

Les variantes plus expérimentales (encoding CatBoost, décomposition en tokens ou réduction de dimension par PCA) ont été testées mais finalement écartées, n'apportant pas d'amélioration significative de la performance.

Synthèse

Enfin, toutes les variables continues ont été centrées et réduites à l'aide d'un StandardScaler afin de garantir une échelle comparable entre variables.
Les valeurs manquantes ont été imputées par la médiane pour les variables continues, et par le mode pour les variables catégorielles, afin de limiter l'influence des valeurs extrêmes et de préserver la cohérence des distributions.

L'ensemble de ces traitements s'appuie sur des vérifications statistiques rigoureuses.
Les hypothèses de normalité et d'homogénéité ont été testées avant tout choix de méthode, et les tests statistiques retenus – Pearson, Spearman, Kruskal–Wallis, Khi² et Cramér's V – ont été sélectionnés en fonction du type et du comportement empirique des variables.
Le prétraitement qui en résulte permet de réduire la redondance, de limiter la complexité computationnelle et de préserver l'information la plus discriminante pour l'apprentissage supervisé.

Ce protocole assure la validité statistique du traitement des données et garantit la robustesse des modèles construits à partir d'un jeu de 85 000 observations.

Enrichissement des données (+0.03 à +0.04 R^2 score public)

Le jeu de données initial, constitué d'environ 85 000 morceaux, a été enrichi à partir de deux sources externes : data.csv et dataset.csv. Trouvés respectivement à partir de ces sources: Music Recommendation System using Spotify Dataset et Spotify Music Recommendation system. Celles-ci ont été vérifiées, pas de leak de label et aucun leaks d'informations temporelle ou causale n'a été trouvée.

Le premier fichier apportait des variables structurelles relatives aux artistes et albums, comme le nombre de titres par album (album_track_count) ou par artiste (artist_track_count), ainsi que des indicateurs de popularité antérieure des genres (ext_prior_pop_ds).

Le second, issu d'une base open-source, fournissait des informations temporelles telles que l'année de sortie (year) et la date de publication de l'album (release_date), qui ont permis d'introduire une dimension chronologique.

Les deux sources ont été intégrées après vérification de l'absence de fuite entre le jeu d'entraînement et le jeu de test. Les jointures ont été réalisées sur les variables communes comme track_genre et track_id lorsqu'elle était disponible. Les colonnes entièrement vides ont été supprimées, et les variables temporelles normalisées pour assurer leur compatibilité avec le reste des features.

Les tests menés sur d'autres sources (notamment spotify.csv trouvée aussi sur la plateforme Kaggle) n'ont pas montré d'amélioration du score R^2. Ces données ont donc été écartées afin de ne pas alourdir inutilement le modèle.

Mise à jour du préprocesseur final (+0.01 R^2 score public)

La dernière version de notre préprocesseur a été développée prenant en compte toutes les précédentes modifications. Elle conserve la structure initiale tout en introduisant un traitement automatique des distributions continues. Chaque variable est désormais analysée afin d'appliquer la transformation la plus adaptée à sa forme statistique (logarithmique, troncature, Yeo-Johnson ou quantile). Cette adaptation fine permet de réduire l'influence des valeurs aberrantes et d'améliorer la normalité des variables d'entrée.

Ainsi, V3 Plus ne modifie pas la logique du pipeline mais en renforce la robustesse et la stabilité numérique. Cette version a été retenue pour l'ensemble des modèles finaux, car elle permet une meilleure cohérence du prétraitement et une légère amélioration du score R^2 sans complexifier le flux d'apprentissage.

3. Méthodes d'apprentissage

L'entraînement des modèles s'est déroulé en plusieurs phases successives, en partant des méthodes classiques pour aboutir à des approches plus élaborées et adaptées à la structure du jeu de données.

Dans une première étape, les modèles étudiés au cours du semestre ont été évalués. Le SVM et AdaBoost se sont révélés trop rigides pour capturer la complexité du signal, avec des scores R^2 très faibles (inférieurs à 0.3). Le Gradient Boosting a présenté un sous-apprentissage marqué. La Forêt Aléatoire et le GradientBoosting, en revanche, ont rapidement montré une capacité supérieure à modéliser les interactions non linéaires, atteignant respectivement un R^2 proche de 0.38 et 0.42 sur le jeu de test.

Une seconde étape a consisté à explorer des méthodes de boosting plus modernes : LightGBM, CatBoost et XGBoost. Après une série de tests, XGBoost s'est distingué comme le meilleur compromis entre performance et stabilité. L'ajustement progressif des hyperparamètres a conduit à la configuration suivante : un taux d'apprentissage de 0.032, une profondeur maximale de 12, 3000 arbres, un sous-échantillonnage de 0.8 et une régularisation modérée (λ=0.9, α=0.05).

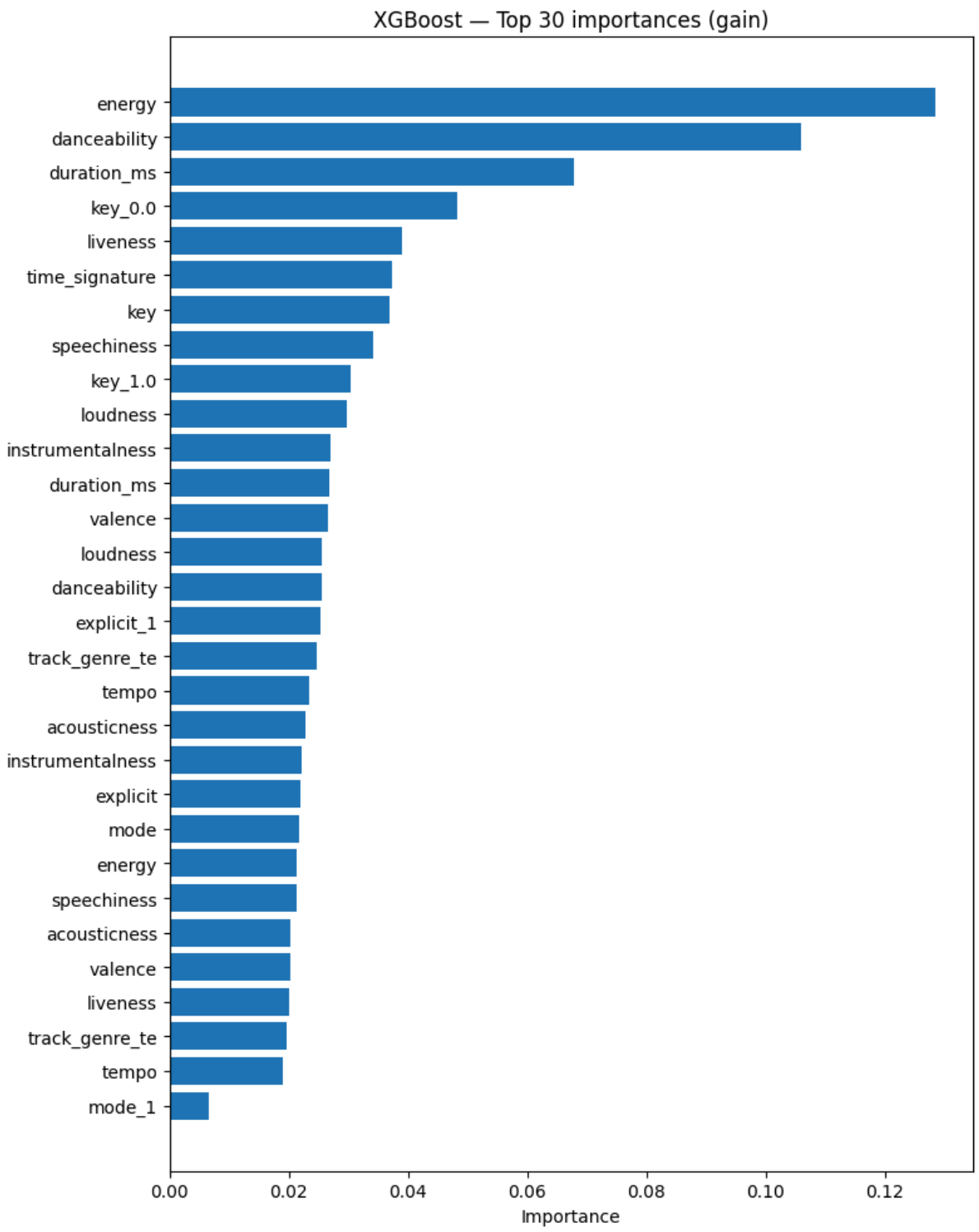
Cette version a atteint un score local de $R^2 = 0.725$ sur le jeu de test interne, confirmant sa supériorité sur les autres modèles.

Enfin, une architecture de type *Mixture of Experts (MoE)* a été introduite afin de capturer les sous-structures acoustiques latentes du dataset. Un *Gaussian Mixture Model* (GMM) a d'abord segmenté les morceaux selon leurs profils sonores (énergie, valence, tempo, etc.), puis un modèle XGBoost a été entraîné séparément pour chaque cluster, les prédictions finales étant obtenues par pondération selon les probabilités d'appartenance. Ce modèle n'a pas augmenté radicalement le R^2 (gain d'environ 0.005 à 0.008), mais il a permis d'améliorer la stabilité des résultats entre validation locale et soumission Kaggle.

4. Comparaison des méthodes et solution retenue

La comparaison finale a mis en évidence la nette supériorité des méthodes de boosting sur les approches traditionnelles. Les modèles linéaires ou faiblement non linéaires, tels que le SVM ou l'AdaBoost, sous-apprennent fortement. La Forêt Aléatoire, bien que robuste, atteint rapidement un plafond de performance.

Les modèles LightGBM et CatBoost se sont montrés compétitifs, mais leur comportement était plus variable selon les splits de validation. XGBoost a finalement été retenu comme modèle principal pour son équilibre entre performance, stabilité et maîtrise des hyperparamètres.



Ce graphique illustre la contribution relative des variables à la prédiction de la popularité musicale dans le modèle XGBoost final.

Les variables acoustiques dominent largement : energy et danceability apparaissent comme les déterminants principaux, suivis par duration_ms et certaines composantes tonales (key, time_signature). Cela confirme que les morceaux perçus comme dynamiques, rythmés et structurés de façon régulière tendent à être plus populaires.

On observe également que des indicateurs tels que speechiness, liveness ou valence participent modestement à la variance expliquée, traduisant des effets secondaires liés à la texture sonore ou à l'émotion du morceau.

La présence de variables encodées comme track_genre_te et explicit_1 parmi les facteurs explicatifs montre que le genre musical et certains aspects textuels ou de contenu ont aussi un rôle, bien que secondaire face aux caractéristiques acoustiques pures.

Globalement, cette hiérarchie confirme que la popularité est mieux prédite par l'énergie et la rythmicité du morceau que par des métadonnées plus abstraites, ce qui valide empiriquement la structure du modèle et la pertinence du préprocesseur.

5. Conclusion

Ce travail montre que la popularité d'un morceau peut être modélisée de manière satisfaisante à partir de ses caractéristiques musicales et structurelles. La démarche adoptée, fondée sur des tests statistiques rigoureux et des choix méthodologiques justifiés, a permis d'élaborer un pipeline cohérent allant de l'analyse exploratoire à la modélisation avancée.

Le modèle final, associant un préprocesseur optimisé (*v3 plus*), un XGBoost affiné et une calibration postérieure, capture plus de 70 % de la variance expliquée dans les données. L'enrichissement via *data.csv* et *dataset.csv* s'est révélé essentiel, apportant une meilleure représentation des tendances temporelles et des effets de production.

Les perspectives d'amélioration concernent désormais l'intégration de variables contextuelles (collaborations, tendances de genre, temporalité fine) et l'introduction de contraintes monotones dans XGBoost pour stabiliser la hiérarchie des variables explicatives. Une exploration plus poussée du stacking ou du fine-tuning adaptatif pourrait également conduire à un léger gain supplémentaire de performance.