2023

DEVOIR D'ÉQUIPE

TECH30724 - INTRODUCTION À PYTHON

AMIRA HAMDI - 11305187
CYRIELLE BOCQUET - 11315583
LUIS ESTUARDO VELA - 11249612
MARIE EMMANUEL CÉLESTIN - 11145012
TING-WEI CHUANG - 11315273
TRACY COTÉ-TURGEON - 11223862

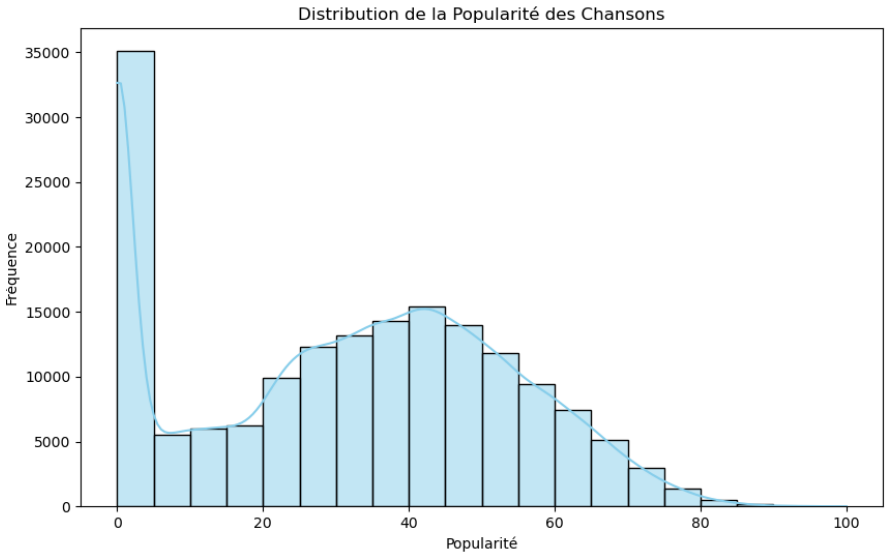
Table des matières

1. Présentation de graphiques autour de la variable Popularité et nos observations

Voici comment la variable est définie dans la description des données :

Popularité : la popularité d'un titre est une valeur comprise entre 0 et 100, 100 étant la valeur la plus populaire. La popularité est calculée par Spotify et est basée, en grande partie, sur le nombre total de lectures du titre et sur leur caractère récent.

Graphique 1 : Distribution de la variable popularité



On constate tout d'abord grâce au graphique qu'il y a environ 35 000 morceaux sur les 170 653, soit 21% des morceaux avec une note de popularité entre 0 et 4. On remarque aussi qu'il n'y a pas beaucoup de morceaux avec une note de popularité très élevée c.-à.-d. entre 80 et 100. On peut dire que les morceaux très populaires sont rares. Autrement dit, on peut penser que le max = 0 et le min = 100.

La note de popularité maximale, en omettant la note de 0, se situe entre 40 et 45. La majorité des morceaux ont une note entre 40 et 45.

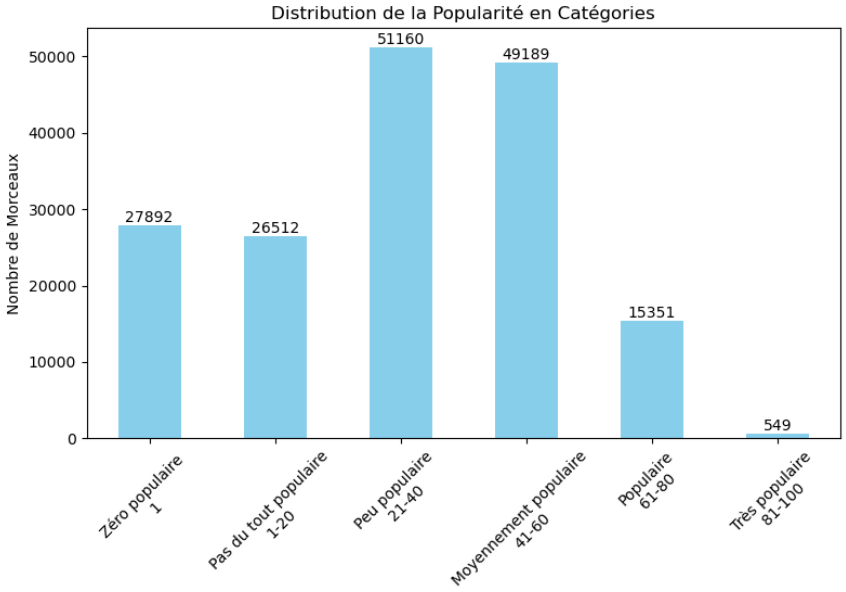
Graphique 2 : Distribution de la popularité par catégories

Avec les résultats du premier graphique, on pense qu'il serait pertinent de créer des catégories de popularité pour pouvoir avoir une meilleure visualisation de la distribution de la variable.

Nous avons créé 6 catégories :

- Zéro populaire qui regroupe la note de 0.
- Pas du tout populaire qui regroupe les notes allant de 1 à 20.
- Peu populaire qui regroupe les notes allant de 21 à 40.
- Moyennement populaire qui regroupe les notes allant de 41 à 60.
- Populaire qui regroupe les notes allant de 61 à 80.
- Très populaire qui regroupe les notes allant de 81 à 100.

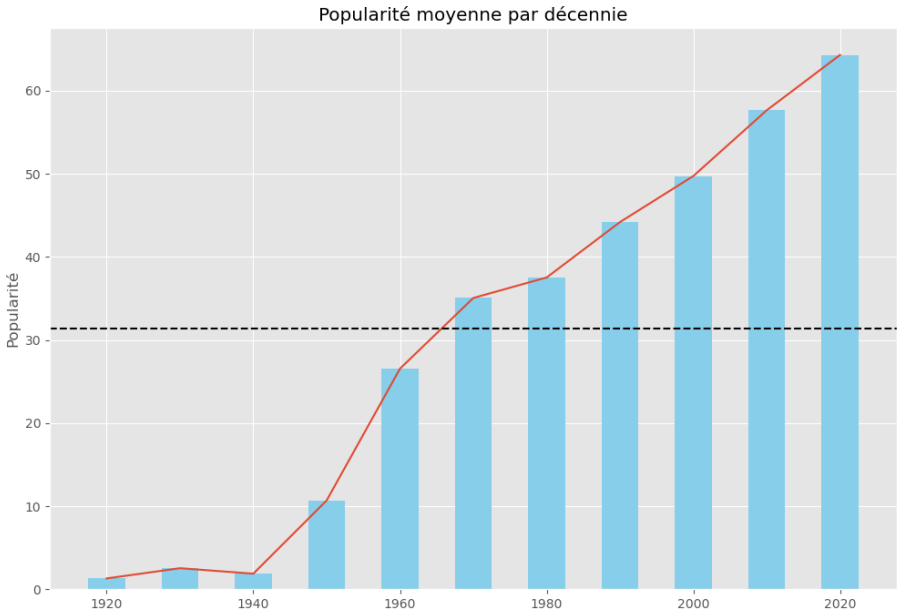
On obtient le graphique suivant :



Ici, on observe 2 catégories qui se démarquent : La première qui contient le plus de morceaux est la catégorie peu populaire (30% des morceaux) suivie de près de la catégorie moyennement populaire (29% des morceaux).

On confirme que la catégorie très populaire contient très peu de morceaux (0.3% des morceaux).

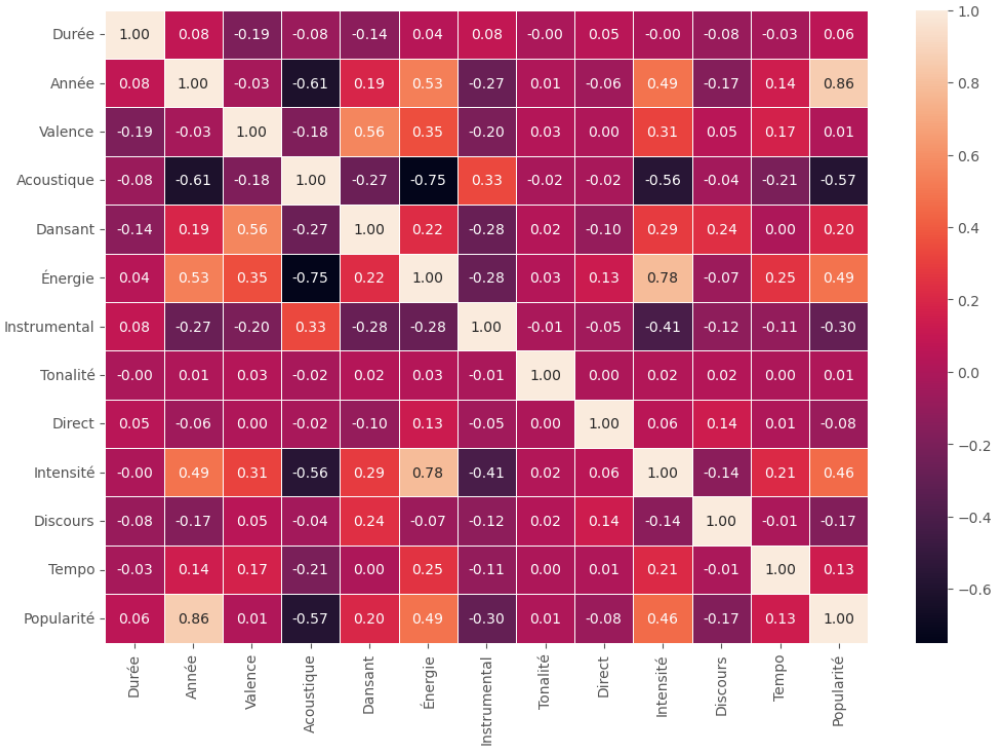
Graphique 3 : Évolution de la popularité moyenne par décennie



On observe que la popularité moyenne des morceaux était très faible (en dessous de la note de 10) jusqu'à la fin des années 50 où elle augmente pour atteindre la note de 20. À partir de là, on remarque une progression positive à travers les décennies avec une augmentation plus ou moins rapide selon les décennies. À noter ici que la décennie 2020 ne contient en réalité que l'année 2020 puisque nos données vont jusqu'à cette année-là. La popularité moyenne des morceaux de l'année 2020 est autour de 65. On observe également avec la ligne en pointillé que la moyenne de la popularité est d'environ 31.

Autrement dit, plus un morceau est récent et plus sa note de popularité augmente et c'est ce qu'on observe avec notre graphique. On peut penser aussi que Spotify étant une plateforme de streaming musical en ligne avec un public cible de jeunes adultes, les morceaux les plus écoutés ont une faible probabilité d'être très anciens ce qui correspond à ce qu'on voit dans notre graphique entre les années 20 à 50.

Graphique 4 : Analyse des corrélations de la variable popularité



Grace à cette matrice, on confirme notre observation, la variable Année est fortement corrélée positivement à la popularité (0.86), c.-à-d. plus l'année augmente, plus la popularité augmente.

On observe que les variables énergie et intensité ont une corrélation positive modérée avec la variable popularité avec un coefficient de corrélation respectif de 0.49 et 0.46. On note qu'il existe une forte corrélation positive entre les variables énergie et intensité c.-à-d. plus l'intensité sonore globale d'un morceau augmente, plus l'énergie du morceau augmente.

La variable acoustique a une corrélation négative modérée avec la variable popularité c.-à-d. plus le morceau est acoustique, plus la variable popularité diminue. Même chose pour la variable instrumental c.-à-d. plus il est probable que le morceau ne contienne pas de contenu vocal et plus la variable popularité diminue.

On note que la variable acoustique a une forte corrélation négative avec la variable énergie et un corrélation négative modérée avec les variables année et intensité. Plus le morceau est acoustique, plus la variable énergie, année et intensité diminuent.

2. Modèle prédictif de la variable « Popularité »

Dans cette section, nous allons élaborer un modèle prédictif pour estimer la popularité des morceaux de l'ensemble de données Spotify. La popularité est une variable importante pour évaluer le succès d'une chanson, et nous chercherons à développer un modèle précis pour la prédire.

Voici un extrait notre jeu de données :

	ID	Artistes	Titre	Durée	Année	Valence	Acoustique	Dansant	Énergie	Explicite	Instrumental	Tonalité	Direct	Intensité	Mode	Discours	Tempo	Popularité
0	4BJqT0PrAfrxzMOxytFOIz	['Sergei Rachmaninoff', 'James Levine', 'Berli...	Piano Concerto No. 3 in D Minor, Op. 30: III. ...	831667	1921	0.0594	0.98200	0.279	0.211	Non	0.878000	10.0	0.6650	-20.096	Majeur	0.0366	80.954	4
1	7xPhfUan2yNtyFG0cUWkt8	['Dennis Day']	Clancy Lowered the Boom	180533	1921	0.9630	0.73200	0.819	0.341	Non	0.000000	7.0	0.1600	NaN	Majeur	0.4150	60.936	5
2	1o6I8BglA6ylDMrIELygv1	['KHP Kridhamardawa Karaton Ngayogyakarta Hadi...	Gati Bali	500062	1921	0.0394	0.96100	0.328	0.166	Non	0.913000	3.0	0.1010	-14.850	Majeur	0.0339	110.339	5
3	3ftBPsC5vPBKxYSee08FDH	['Frank Parker']	Danny Boy	210000	1921	0.1650	0.96700	0.275	0.309	Non	0.000028	5.0	0.3810	-9.316	Majeur	0.0354	100.109	3

Prétraitement des données

A. Nettoyage des données

Avant de construire le modèle, nous avons effectué des étapes de prétraitement des données. Cela comprend le traitement des valeurs manquantes, la conversion des variables catégorielles en variables numériques (par exemple, le mode, l'explicite), et la division des données en ensembles d'entraînement et de test.

Identifions les colonnes catégorielles qui comportent beaucoup de valeurs uniques car nous allons les exclure:

ID	170653
Artistes	34088
Titre	133633

Nous avons aussi exclu l'année car l'inclusion de celle-ci pourrait améliorer la performance du modèle en capturant cette tendance linéaire de la popularité par rapport à la date de sortie, nous avons choisi de la retirer pour éviter de biaiser les données. En effet, comme nous pouvons le voir dans le graphique 3, il y a une forte corrélation entre l'année et la popularité. Cette corrélation linéaire pourrait introduire un biais dans le modèle en surestimant l'importance de l'année de sortie au détriment d'autres caractéristiques des titres qui pourraient être plus significatives pour prédire la popularité. L'année de sortie d'un titre et la décennie n'a pas été considérée comme une caractéristique déterminante dans ce calcul de popularité, raison pour laquelle elle a été retirée de notre ensemble de données pour la modélisation. De plus, nous avons aussi pris en compte si nous la considérions comme une variable catégorielle, nous l'aurions évincé car elle compterait 100 valeurs uniques.

B. Division des données

Il reste deux variables catégorielles binaires : Explicite et Mode.

Voici les variables numériques : Durée, Valence, Acoustique, Dansant, Énergie, Instrumental, Tonalité, Direct, Intensité, Discours et Tempo.

Nous allons maintenant créer deux jeux de données, l'un contenant seulement notre variable cible c'est-à-dire, la popularité qu'on nomme « y » et un autre qui contient toutes les autres colonnes à l'exception des colonnes ID, Artistes, Titre et Année qu'on nomme « X ».

Une fois cette étape réalisée, nous identifions les colonnes numériques et catégorielles puis nous devons séparer notre ensemble de données en trois parties car nous avons besoin de trois étapes : l'entraînement, la validation et le test. On conserve 60% des données pour l'entraînement, 20% des données pour l'ensemble de validation et 20% pour l'ensemble de test.

On obtient les jeux de données suivants :

- X_train: (102391, 13) pour l'entraînement.
- X_val: (34131, 13) pour la validation.
- X_test: (34131, 13) pour le test.

La prochaine étape est d'imputer les valeurs manquantes, pour cela, on doit d'abord séparer les variables numériques et catégorielles dans chacun des échantillons puis conserver en mémoire le nom des index de chaque ligne ce qui nous permettra de regrouper les jeux de données plus facilement.

Pour les variables numériques, on va identifier la médiane de chaque colonne numérique de l'échantillon d'entraînement puis on va remplacer les données manquantes par la médiane et l'appliquer à tous les échantillons.

Pour les variables catégorielles, on va identifier le mode de chaque colonne catégorielle de l'échantillon d'entraînement puis on va remplacer les données manquantes par le mode et l'appliquer à tous les échantillons.

En effectuant ces étapes nous obtenons à la fin des arrays qu'il faut qu'on transforme en dataframe.

On doit par la suite encoder les variables catégorielles en nombre entier pour que l'information puisse être utilisée dans notre modèle.

Finalement, on va centraliser et normaliser nos données car on n'a pas la même échelle pour toutes nos variables comme la durée qui est milliseconde et valence qui est un score entre 0 et 1.

Voici notre jeu de données de test par exemple après normalisation:

	encoder__Explicite_Oui	encoder__Mode_Mineur	remainder__Durée	remainder__Valence	remainder__Acoustique	remainder__Dansant
69125	0.0	0.0	-0.003166	-0.180288	-0.145939	0.388664
19786	1.0	1.0	-0.053890	0.064904	0.000000	1.145749
130015	0.0	0.0	-0.534977	0.788462	-0.562563	0.303644
92314	1.0	0.0	0.049146	-0.204327	-0.628807	0.797571
101456	0.0	0.0	-0.149416	0.000000	-0.058376	-0.267206
...	...	...	...	...	...	...
110021	0.0	1.0	-0.063832	-0.978365	0.590102	-0.469636
106959	0.0	0.0	-0.274183	-0.853365	0.437817	0.761134
89694	0.0	0.0	-0.185573	0.762019	-0.619289	0.773279
32335	0.0	0.0	0.115982	0.935096	-0.529188	0.712551
123602	0.0	0.0	1.880993	-1.180529	0.269036	0.773279

remainder__Énergie	remainder__Instrumental	remainder__Tonalité	remainder__Direct	remainder__Intensité	remainder__Discours	remainder__Tempo
-0.205950	-0.002285	0.000000	1.111801	-0.438282	0.000000	-0.150753
0.340961	-0.002317	0.166667	-0.105590	0.311860	7.813333	0.434373
0.897025	-0.002317	0.166667	0.012422	0.544511	-0.056000	0.318800
0.274600	-0.002317	-0.666667	-0.099379	0.662969	3.840000	0.928834
0.016018	0.018406	0.000000	0.565217	0.214733	-0.506667	-0.653564
...	...	...	...	...	...	...
-0.782609	0.011498	0.000000	-0.173913	-0.690415	-0.309333	-0.296501
-0.398169	-0.002158	-0.500000	-0.466460	0.133959	-0.269333	0.260752
0.835240	-0.002317	-0.166667	2.223602	0.875569	-0.381333	0.121978
0.704805	0.013836	0.666667	-0.043478	0.078783	-0.213333	0.215277
-0.384439	0.000670	-0.166667	0.645963	-0.198521	-0.341333	0.192812

Choix et entrainement des modèles

Pour construire le modèle, nous avons sélectionné un ensemble de caractéristiques pertinentes qui pourraient exercer une influence sur la popularité. Ces caractéristiques incluent la valence, l'énergie, la tonalité, la durée, etc. Ces caractéristiques servent à l'entraînement des 3 modèles de régression suivante:

i. Régression linéaire
ii. Régression Ridge
iii. Régression Lasso

Les modèles de régression linéaire, Ridge et Lasso sont plus simples et plus faciles à interpréter que les modèles plus complexes.

De plus, ils sont souvent utilisés pour réduire le surajustement dans les données et permettent en général d'améliorer les performances du modèle. Les modèles de régression linéaire fonctionnent bien même avec un petit nombre d'observations. Aussi, ils ont généralement des temps d'entrainement plus courts par rapport à d'autre modèles.

La régression Ridge permet de contrôler la multicolinéarité c.-à-d. qu'elle va identifier les variables explicatives qui sont fortement corrélées entre elles et réduire leur influence. On a constaté qu'il y avait des fortes corrélations négatives entre énergie et acoustique et une forte corrélation positive entre valence et dansant ainsi qu'entre énergie et intensité.

Ce modèle de régression linéaire régularisée ajoute une pénalité L2 aux coefficients du modèle, ce qui aide à réduire le surajustement en contrôlant la taille des coefficients. Il est plus flexible que la régression linéaire standard, ce qui lui permet de mieux gérer les cas où il y a un grand nombre de caractéristiques potentiellement corrélées entre elles. Cependant, il ne fait pas de sélection automatique de variables.

La régression Lasso a la capacité d'éliminer les variables moins importantes en faisant une sélection automatique et en conservant uniquement les variables les plus importantes. Cette technique peut s'avérer plus adéquate lorsqu'on a des valeurs extrêmes. Dans notre cas, ce modèle nous permettrait de faire la sélection des variables importantes, sans nécessité d'une intervention humaine.

À la différence de la Ridge, la régression Lasso utilise une pénalité L1, ce qui conduit à la réduction des coefficients à zéro pour certaines variables moins importantes, agissant ainsi comme un outil de sélection automatique de variables. Cela permet de simplifier le modèle en ne conservant que les caractéristiques les plus importantes. Cependant, il peut être instable lorsqu'il y a une corrélation forte entre les caractéristiques.

La régression linéaire : il s'agit d'un modèle de base qui suppose une relation linéaire entre les caractéristiques et la variable cible, la popularité dans notre cas. C'est un modèle simple à interpréter et à comprendre, mais il peut être limité pour capturer des relations non linéaires dans les données, ce qui peut entraîner une baisse de performance dans des scénarios complexes.

Nous avons examiné trois modèles de régression distincts pour estimer la popularité des morceaux. Chacun de ces modèles possède ses propres caractéristiques et avantages spécifiques.

Chacun de ces modèles offre un compromis entre la simplicité, la flexibilité et la capacité à gérer les problèmes de surajustement. Le choix du modèle dépend de la complexité des données, de la taille de l'ensemble de données et de la priorité accordée à l'interprétabilité par rapport à la performance prédictive.

Évaluation des modèles

Nous avons évalué la performance de chaque modèle sur l'ensemble de test en utilisant des métriques telles que le coefficient de détermination (R^2), l'erreur absolue moyenne (MAE) et l'erreur quadratique moyenne (RMSE). Ces métriques nous ont permis de comprendre comment chaque modèle se comporte pour prédire la popularité des morceaux.

Modèle	R^2	MAE	RMSE
Régression linéaire	0.44	13.17	267.89
Régression Ridge	0.29	15.19	479.36
Régression Lasso	-1.96	18.58	479.38

Sélection du meilleur modèle

En fonction des performances évaluées, nous avons sélectionné le modèle linéaire car il présente les meilleures performances. Bien que les modèles Ridge et Lasso affichent des valeurs similaires pour MAE et RMSE, la régression linéaire est légèrement supérieure. Nous avons également interprété ce modèle pour comprendre quelles caractéristiques influencent le plus la popularité des morceaux. De plus, nous avons exploré différentes valeurs pour le paramètre alpha dans les modèles Ridge et Lasso, pour démontrer leur sensibilité aux variations des paramètres.

Interprétation du modèle

La régression linéaire a obtenu le R2 le plus élevé parmi les modèles testés avec une valeur 0.44. Cela indique que 44% de la variance de la popularité peut être expliquée par les caractéristiques prises en compte dans le modèle. De plus, la régression linéaire affiche également les valeurs les plus faibles pour les métriques MAE et RMSE, avec respectivement 13.17 et 267.89.

Ces résultats suggèrent que la régression linéaire parvient à prédire la popularité avec une précision relativement plus élevée par rapport aux modèles de régressions Ridge et Lasso.

Comparativement, les modèles de régression Ridge et Lasso ont des performances inférieures en termes de R2, MAE et RMSE.

Ces résultats confirment que la régression linéaire est le choix optimal parmi ces modèles pour prédire la popularité des morceaux, en utilisant les caractéristiques considérées dans cette étude.

Conclusion

En conclusion, le modèle sélectionné, à savoir la régression linéaire, a démontré de bonnes performances pour prédire la popularité des morceaux. L'analyse des caractéristiques a mis en évidence l'impact de certaines variables sur la popularité, telle que la valence et l'énergie sur la popularité des chansons, offrant des insights sur les facteurs qui contribuent à la popularité d'une chanson, aidant potentiellement les artistes, producteurs et plateformes musicales à optimiser la création et la promotion de morceaux plus attrayants pour le public.