KIHEL HAJAR

Numéro d'étudiant : 24010389 **Classe** : CAC2

Compte rendu d'Analyse : Prétraitement et Étude du Dataset Amazon Sales

Date : 30 Novembre 2025 **Auteur :** KIHEL HAJAR

Sommaire

---

1. Introduction et Définition de la Problématique

Le point de départ de ce projet est l'analyse d'un jeu de données de ventes Amazon dans le but de **prédire la note moyenne d'un produit (`rating`)**. Pour une plateforme de e-commerce, comprendre ce qui influence la satisfaction client est capital. L'objectif a donc été de déterminer si des caractéristiques quantitatives simples (prix, réduction, volumeLe projet d'analyse a débuté par l'étude d'un jeu de données des ventes Amazon. L'ambition principale était d'établir un modèle capable de prédire la note moyenne d'un produit (rating), un indicateur essentiel de la satisfaction client pour toute plateforme de e-commerce. La nature de la variable cible, qui est continue (variant de 0.0 à 5.0), a orienté toute la méthodologie vers un problème de Régression. Avant d'entamer la modélisation, qui visait à déterminer si des caractéristiques quantitatives simples (prix, réduction, volume de revues) étaient suffisantes pour expliquer la qualité perçue des produits, un travail approfondi de nettoyage et de préparation du jeu de données s'est avéré indispensable.

La problématique principale de cette analyse est:

Dans quelle mesure les caractéristiques quantitatives et catégorielles d'un produit Amazon (prix, réduction, volume de revues, et catégorie) permettent-elles de prédire de manière fiable la note moyenne que lui attribuent les clients (rating) ?

2. Préparation et Nettoyage des Données

La première phase a été cruciale, car les colonnes censées être numériques (`discounted_price`, `actual_price`, `discount_percentage`, `rating`, `rating_count`) étaient incorrectement typées en `object` (chaîne de caractères). Ce phénomène était dû à la présence de symboles non numériques comme ' ' (roupie indienne) et '%' (pourcentage). Nous avons donc procédé à la suppression de ces caractères parasites et à la conversion systématique des colonnes en types numériques (float ou Int64), une étape indispensable pour pouvoir effectuer des calculs et entraîner des modèles.

Parallèlement, l'étape de gestion des valeurs manquantes (imputation) a été nécessaire. Les quelques valeurs manquantes dans les colonnes `rating` et `rating_count` ont été remplacées par la **médiane** des valeurs observées (4.1 pour `rating` et 5179.0 pour `rating_count`). Ce choix de la médiane, plutôt que la moyenne, est pertinent pour des distributions très asymétriques, afin de minimiser l'impact potentiel des valeurs extrêmes sur la distribution globale et sur la modélisation ultérieure.

Tâche de Nettoyage	Colonnes Concernées	Méthode Appliquée
Conversion de Type	`discounted_price`, `actual_price`	Suppression de ' ' et conversion en `float`.
Conversion de Type	`discount_percentage`	Suppression de '%' et conversion en `float`.
Imputation des NaN	`rating`, `rating_count`	Remplacement par la médiane respective de la colonne.

---

3. Analyse Exploratoire et Ingénierie des Caractéristiques (EDA & Feature Engineering)

Analyse Exploratoire des Données (EDA)

L'Analyse Exploratoire a confirmé que la variable cible, `rating`, est fortement concentrée dans la plage supérieure (entre 4.0 et 4.5), ce qui est typique des systèmes d'évaluation en ligne. Nous avons également observé que les prix (`discounted_price` et `actual_price`) présentaient une forte asymétrie positive, indiquant que la majorité des produits sont vendus à un prix bas, avec

une longue queue d'articles beaucoup plus chers (outliers). Le volume des revues (`rating_count`) présentait une asymétrie similaire, avec la majorité des produits ayant peu de revues, et quelques articles très populaires.

L'analyse de corrélation, présentée sous forme de matrice, a révélé des corrélations très faibles entre le `rating` et les variables de prix ou de réduction (proches de zéro). Par exemple, la corrélation entre `rating` et `discount_percentage` était légèrement négative (-0.16), suggérant une infime tendance pour les produits fortement réduits à avoir des notes marginalement plus basses, mais sans impact explicatif significatif. Ce constat préliminaire a soulevé des doutes sur le pouvoir prédictif des seules variables numériques.

Ingénierie des Caractéristiques (Feature Engineering)

Pour intégrer des données catégorielles complexes, nous avons simplifié la colonne **`category`** en extrayant la **`primary_category`** (la première catégorie de la hiérarchie). Nous avons ensuite créé une variable binaire nommée **`is_electronics`**, car la catégorie 'Electronics' était la plus représentée dans le jeu de données. Cette simplification a permis de réduire la dimensionnalité des données catégorielles tout en conservant l'information des segments de marché les plus importants.

---

4. Méthodologie de Modélisation et Mise en Place du Pipeline

Préparation Finale

Avant l'entraînement, les étapes suivantes ont été exécutées :

1. **Encodage Catégoriel :** La variable `primary_category` (et toutes ses modalités) a été transformée en colonnes binaires via la technique du **One-Hot Encoding**.
2. **Division des Données :** Le jeu de données a été séparé en un ensemble d'entraînement (80%) pour former les modèles et un ensemble de test (20%) pour évaluer leur capacité de généralisation sur des données non vues.
3. **Mise à l'Échelle :** Les caractéristiques numériques ont été standardisées (à l'aide du **StandardScaler**) sur l'ensemble d'entraînement. Cette étape est essentielle pour que les modèles, notamment ceux basés sur la distance ou l'optimisation par gradient, ne soient pas biaisés par la différence d'échelle entre des variables comme le prix et le nombre de revues.

Sélection des Modèles

Nous avons sélectionné trois architectures de modèles de Régression pour évaluer différentes approches :

- **Régression Linéaire :** Pour établir une base et tester l'existence d'une relation linéaire simple.
- **Forêt Aléatoire (Random Forest) :** Un modèle ensemble basé sur les arbres de décision, réputé pour sa robustesse et sa capacité à gérer les relations non linéaires.
- **Gradient Boosting (GradientBoostingRegressor) :** Un autre modèle ensemble puissant, qui construit les arbres de manière séquentielle, corrigeant les erreurs des arbres précédents.

---

5. Résultats des Modèles et Optimisation

Performance Initiale (Validation Croisée)

Une première évaluation des modèles par **Validation Croisée (5-fold)** sur l'ensemble d'entraînement a confirmé la complexité de la tâche. Les scores R^2 initiaux étaient faibles, indiquant qu'aucun des modèles ne parvenait à expliquer une grande partie de la variance du `rating`. Le RandomForestRegressor a obtenu le meilleur score de base (autour de $R^2 \approx 0.16$), soulignant que la relation, bien que faible, était mieux capturée par une approche non linéaire et arborescente.

Optimisation et Résultats Finaux sur l'Ensemble de Test

Pour maximiser la performance, nous avons procédé à une optimisation des hyperparamètres des deux meilleurs modèles (Random Forest et Gradient Boosting) via **RandomizedSearchCV**. Cette méthode a permis d'explorer efficacement l'espace des hyperparamètres pour trouver la meilleure configuration.

Le modèle qui a finalement fourni la meilleure performance sur l'ensemble de test non vu est le **GradientBoostingRegressor optimisé**.

Modèle	R^2 (Test)	MAE (Test)	MSE (Test)
RandomForestRegressor (Optimisé)	0.1935	0.1744	0.0659
GradientBoostingRegressor (Optimisé)	**0.1968**	0.1802	**0.0656**

Avec un R^2 final de **0.1968**, le Gradient Boosting parvient à expliquer approximativement **20% de la variance** des notes. Ce résultat, bien que le meilleur obtenu, est modeste. L'erreur absolue moyenne (MAE) est d'environ **0.18**, signifiant que, en moyenne, la prédiction est fausse de moins de deux dixièmes de point. Compte tenu de l'étroite plage de variation du `rating` réel (très concentré), cette erreur est techniquement faible, mais le score R^2 révèle que l'information cruciale pour la prédiction manque encore.

6. Conclusion Critique et Perspectives d'Amélioration

Analyse des Limites

Le score R^2 obtenu (inférieur à 0.20) est l'information la plus significative de cette analyse : il indique que **les prix, les réductions et les catégories ne sont que des prédicteurs très mineurs de la note d'un produit**. Ce résultat est cohérent avec l'intuition : la satisfaction client (la note) est principalement liée à la qualité intrinsèque du produit, souvent décrite dans les revues.

La faible variance du `rating` lui-même rend la tâche de régression intrinsèquement difficile. Les produits qui restent visibles sur Amazon ont déjà subi une forte sélection naturelle et ont donc un `rating` déjà élevé, laissant peu de marge pour la prédiction.

Perspectives d'Amélioration (Next Steps)

L'amélioration significative des performances passe inévitablement par l'exploitation des données textuelles. La prochaine étape critique doit être le **Traitement du Langage Naturel (NLP)** sur les colonnes `product_name`, `about_product`, `review_title`, et potentiellement les revues complètes.

- **Analyse de Sentiment :** Extraire une *feature* d'analyse de sentiment (positif/négatif) des titres et contenus de revues.
- **Embedding :** Utiliser des techniques d'embedding (comme Word2Vec ou Tfidf) pour capturer le sens et les thèmes des descriptions de produits.

En intégrant ces nouvelles caractéristiques textuelles dans le pipeline de modélisation, nous nous attendons à ce que le R^2 augmente considérablement, car elles représentent probablement l'information manquante qui explique la satisfaction réelle des utilisateurs.