Compte rendu d'analyse : Modélisation et évaluation

1. Introduction

Contexte et problématique

L'objectif du projet est de construire un modèle de machine learning capable de prédire une variable cible à partir d'un ensemble de caractéristiques issues d'un jeu de données tabulaire. Le dataset utilisé (format CSV) contient **8 variables explicatives** et une variable cible numérique. La problématique consiste à déterminer si un modèle de régression peut fournir des prédictions fiables, et à analyser ses performances et limitations.

Objectifs

- Explorer et préparer les données.
- Sélectionner et justifier une méthode de modélisation.
- Évaluer quantitativement les performances du modèle à l'aide de métriques adaptées (MSE, RMSE, R^2).
- Identifier les erreurs du modèle et proposer des pistes d'amélioration.

---

2. Méthodologie

2.1 Nettoyage et préparation des données

- **Vérification des valeurs manquantes** : aucune valeur manquante observée, donc aucune imputation nécessaire.
- **Séparation entraînement / test** :
  - 80% pour l'entraînement
  - 20% pour le test
  - `random_state = 42` pour assurer la reproductibilité
- **Absence de normalisation** : acceptable pour les algorithmes d'arbres (Random Forest), car ils n'y sont pas sensibles.

2.2 Choix du modèle

Le modèle utilisé est un **Random Forest Regressor**. Justifications :

- Robuste aux données bruitées.
- Capable de capturer des relations non linéaires.
- Peu sensible au scaling et au choix des variables.
- Souvent performant sans réglage initial complexe.

2.3 Justification des métriques

- **MSE (Mean Squared Error)** : quantifie l'erreur quadratique moyenne.
- **RMSE (Root Mean Squared Error)** : interprétable dans l'unité de la cible.
- **R^2 Score** : proportion de variance expliquée par le modèle.

---

3. Résultats & Discussion

3.1 Performances obtenues

D'après les résultats du notebook :

- **MSE test :** ≈ **719 713.33**
- **RMSE test :** √MSE ≈ **848.36**
- **R^2 Score : 0.48**

3.2 Interprétation

- Le modèle explique **environ 48%** de la variance de la variable cible.
- Une **erreur moyenne (RMSE) relativement élevée** indique des prédictions encore éloignées des valeurs réelles.
- Bien que le modèle capture une partie du signal, il existe un **important résidu de variance non expliquée**.

3.3 Analyse des erreurs

Aucune matrice de confusion n'est incluse (normal, car il s'agit d'un **problème de régression**). Les erreurs peuvent provenir de :

- Relations non capturées par les arbres malgré leur flexibilité.
- Variables explicatives insuffisantes ou manquantes.
- Données naturellement bruitées.

- Hyperparamètres par défaut non optimisés.

Une analyse plus fine des résidus (non fournie dans le notebook) serait utile pour vérifier :

- Hétéroscédasticité
- Distribution des erreurs
- Patterns résiduels indiquant un manque structurel du modèle

---

4. Conclusion

Limites du modèle

- Le modèle atteint des performances moyennes ($R^2 \approx 0.48$).
- Erreurs importantes (RMSE > 800), signifiant que les prédictions sont encore imprécises.
- Absence d'optimisation des hyperparamètres.
- Possible insuffisance des variables explicatives pour capturer la complexité du phénomène.

Pistes d'amélioration

1. **Optimisation du Random Forest**
    - Ajuster `n_estimators`, `max_depth`, `min_samples_split`, etc.
    - Utiliser GridSearchCV / RandomizedSearchCV.
2. **Tester d'autres algorithmes**
    - Gradient Boosting Regressor
    - XGBoost, LightGBM
    - Régression linéaire ou régularisée (Ridge, Lasso)
3. **Feature engineering**
    - Création de nouvelles variables
    - Transformation logarithmique si la cible est très dispersée
    - Analyse de corrélations
4. **Visualisation des résidus**
    - Détecter des patterns non capturés
    - Vérifier la normalité, variance non constante, etc.
5. **Augmentation du dataset**
    - Obtenir plus de données pour réduire le sur-ajustement.