Prédiction de l'inflation future par apprentissage supervisé

Approche par classification binaire

Rakotoarimanana Joy Sandra Hasina

M1 MAFI

Table des matières

1 Introduction

L'inflation correspond à l'augmentation durable du niveau général des prix et constitue un indicateur central de la stabilité macroéconomique. Une inflation élevée ou volatile peut avoir des effets négatifs sur le pouvoir d'achat, l'investissement et la croissance.

Au-delà de l'observation a posteriori, l'enjeu principal réside dans la capacité à **anticiper l'évolution future de l'inflation**. L'objectif de ce travail est de construire un modèle de classification permettant de prédire si l'inflation de l'année suivante dépassera un seuil critique de 5 %.

Ce seuil est couramment utilisé en macroéconomie pour caractériser une situation inflationniste préoccupante.

2 Présentation des données

Le jeu de données utilisé est le jeu *Global Economic Indicators (2010–2025)*, disponible sur la plateforme Kaggle. Il regroupe, pour un ensemble de pays et sur plusieurs années, des indicateurs macroéconomiques clés tels que :

— l'inflation (CPI),
— la croissance du PIB,
— le taux de chômage,
— les indicateurs budgétaires et financiers.

Les données sont structurées sous forme de panel pays–année, ce qui permet d'exploiter la dimension temporelle pour la construction de la variable cible.

3 Préparation des données

3.1 Nettoyage et normalisation

Les noms de variables ont été standardisés afin de respecter les conventions usuelles en modélisation :

— utilisation de minuscules,
— format *snake_case*,
— suppression des caractères spéciaux.

Les valeurs manquantes des variables explicatives ont été imputées par la médiane, afin de limiter l'influence des valeurs extrêmes.

3.2 Construction de la variable cible

La variable cible correspond à l'inflation CPI de l'année $t+1$ pour un même pays. Formellement, pour un pays i et une année t :

$$y_{i,t} = \begin{cases} 1 & \text{si Inflation}_{i,t+1} > 5\% \\ 0 & \text{sinon} \end{cases}$$

Cette transformation permet de formuler le problème comme une tâche de **classification binaire supervisée**.

Les observations pour lesquelles l'inflation future n'est pas disponible ont été exclues du jeu de données.

4 Méthodologie

4.1 Séparation des données

Les données ont été divisées en trois sous-ensembles :
— 60 % pour l'entraînement,
— 20 % pour la validation,
— 20 % pour le test.

La séparation a été réalisée de manière stratifiée afin de préserver la proportion des classes dans chaque sous-échantillon.

4.2 Modèles évalués

Plusieurs algorithmes de classification ont été comparés :
— Régression logistique,
— KNN ou k plus proches voisins,
— Arbre de décision,
— Random Forest,
— Gradient Boosting,
— Support Vector Machine,
— XGBoost,
— Naive Bayes.

La métrique principale retenue pour la comparaison est l'**AUC ROC**, car elle mesure la capacité de discrimination du modèle indépendamment du seuil de décision.

5 Optimisation des modèles

Les trois modèles présentant les meilleures performances sur l'ensemble de validation ont été retenus pour une optimisation des hyperparamètres par *Grid Search*.

L'optimisation a été réalisée uniquement sur l'ensemble d'entraînement, afin d'éviter toute fuite d'information.

6 Résultats

6.1 Comparaison des modèles

Les performances des différents modèles ont été évaluées sur l'ensemble de validation à l'aide de plusieurs métriques : Accuracy, Precision, Recall, F1-score et AUC ROC.

Le tableau 1 présente les résultats obtenus.

TABLE 1 – Résultats des modèles évalués

Modèle	Accuracy	Precision	Recall	F1-score	AUC
XGBoost	0.834	0.755	0.741	0.748	0.909
Random Forest	0.830	0.761	0.711	0.735	0.905
Gradient Boosting	0.838	0.771	0.729	0.749	0.902
Decision Tree	0.758	0.632	0.651	0.641	0.731
KNN	0.609	0.356	0.223	0.274	0.558
Naive Bayes	0.383	0.344	0.952	0.506	0.538
Logistic Regression	0.669	0.000	0.000	0.000	0.523
SVM	0.669	0.000	0.000	0.000	0.466

6.2 Sélection des trois meilleurs modèles

Sur la base de la métrique AUC ROC, trois modèles se distinguent nettement :
— XGBoost,
— Random Forest,
— Gradient Boosting.

Ces modèles présentent une capacité de discrimination significativement supérieure aux autres approches testées, en particulier par rapport aux modèles linéaires et aux méthodes basées sur la distance.

La figure 1 présente les courbes ROC des trois meilleurs modèles évalués sur l'ensemble de test.

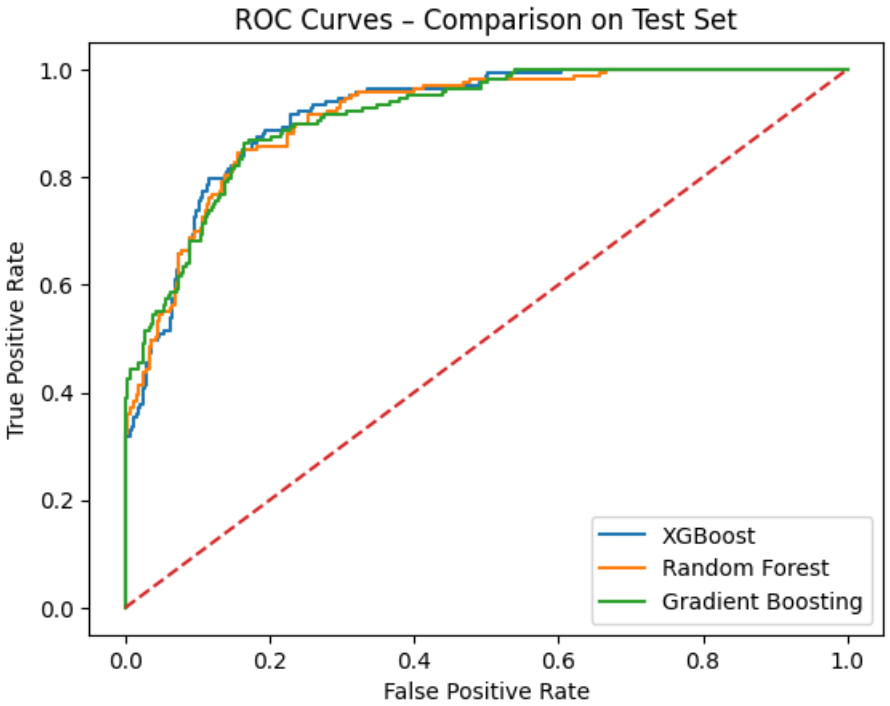

FIGURE 1 – Courbes ROC des trois meilleurs modèles sur l'ensemble de test

6.3 Sélection du modèle final

Parmi les trois modèles optimisés, le modèle **XGBoost** obtient la meilleure performance globale sur l'ensemble de test.

Ses métriques finales sont les suivantes :

— Accuracy : 0.849
— Precision : 0.786
— Recall : 0.749
— F1-score : 0.767
— AUC ROC : 0.918

La figure 2 présente les courbes du modèle **XGBoost**.

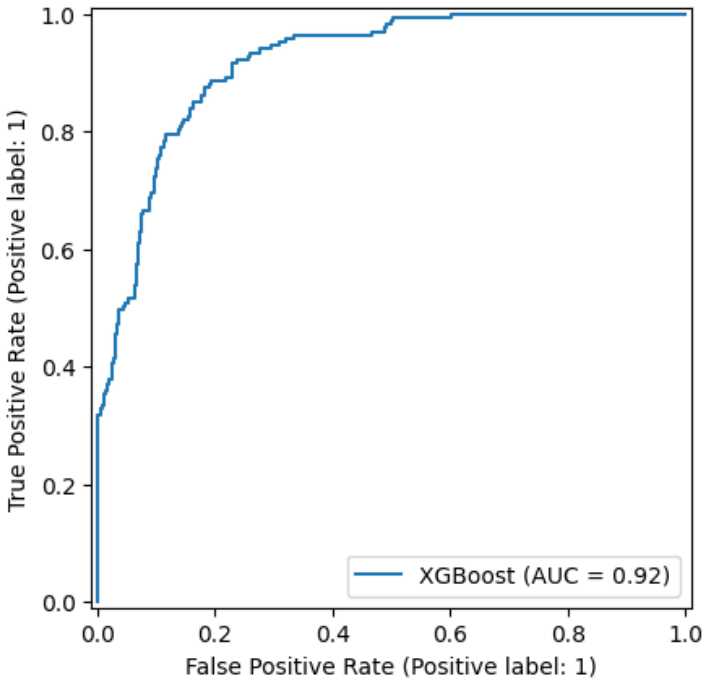

FIGURE 2 – Courbe AUC ROC - XGBoost

Ces résultats indiquent une bonne capacité du modèle à détecter les situations d'inflation élevée tout en maintenant un compromis satisfaisant entre précision et rappel.

6.4 Analyse de la matrice de confusion

La figure 3 présente la matrice de confusion associée au modèle XGBoost sur l'ensemble de test.

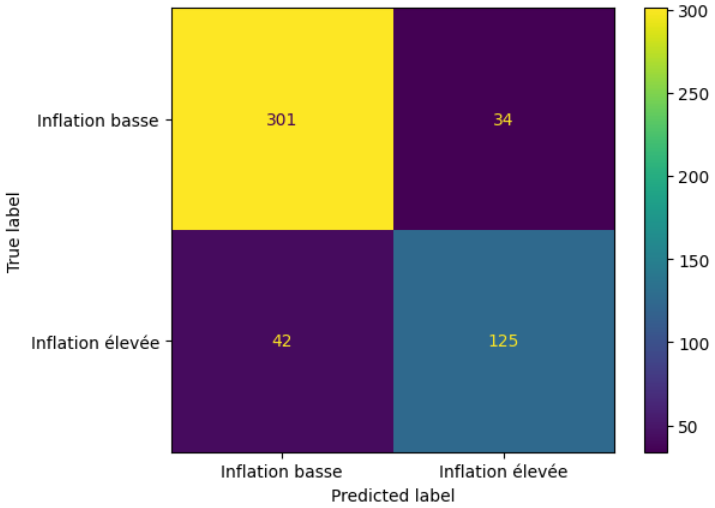

FIGURE 3 – Matrice de confusion du modèle XGBoost

6.5 Interprétation des résultats

La matrice de confusion montre que le modèle XGBoost parvient à identifier correctement une large proportion des situations d'inflation élevée, tout en limitant le nombre de faux positifs.

La valeur élevée de l'AUC ROC confirme la capacité du modèle à discriminer efficacement les observations à risque inflationniste, indépendamment du seuil de décision retenu.

7 Limites et perspectives

Plusieurs limites doivent être soulignées :

— l'absence de discrimination par pays,
— l'absence d'une séparation strictement temporelle,
— l'imputation simple des valeurs manquantes,
— l'absence de calibration des probabilités.

Des extensions possibles incluent l'intégration de méthodes de validation temporelle, la calibration des probabilités et l'ajout de variables retardées.

8 Conclusion

Ce travail montre qu'il est possible d'anticiper les périodes de forte inflation à partir d'indicateurs macroéconomiques observés, en formulant le problème comme une tâche de classification binaire.

Les résultats obtenus soulignent l'intérêt des méthodes d'ensemble, notamment XGBoost et Random Forest, pour la prédiction de l'inflation future.