10.000 {CODEURS} 2015-2025

DATAGONG

SOUTENANCE DE FIN DE FORMATION EN DATA SCIENCE A DATAGONG

Edith BALLOGOU
&
Marc AZIAN

SOMMAIRE

Introduction

Les tendances politiques sont souvent influencées par des facteurs sociodémographiques tels que l'âge, le niveau d'éducation, le revenu et autres.

Dans le cadre de ce projet, nous allons développer un modèle de classification binaire capable de prédire les résultats de l’élection présidentielle américaine de 2020.

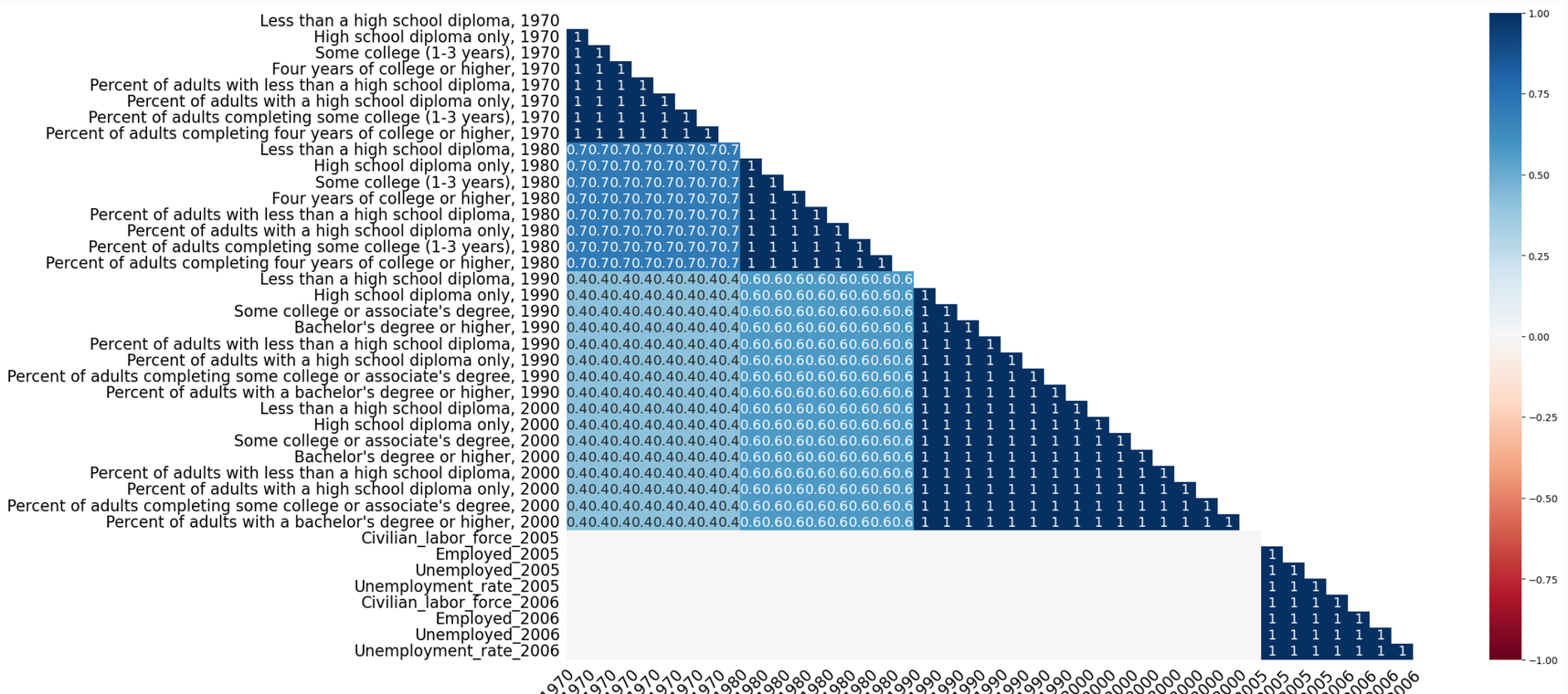
OBJECTIF DU PROJET

L’objectif est d’identifier les facteurs clés influençant le vote afin d’optimiser les stratégies de campagne.

- Éducation : 47 variables et 3283 individus issus des recensements américains depuis 1970
- Population : 165 variables et 3273 individus fournis par l'ERS (Economic Research Service)
- Pauvreté : 34 variables 3193 individus issus du programme SAIPE
- Employabilité : 88 variables et 3275 individus sur le chômage et les revenus
- Résultat 2020 : la base des résultats des élections présidentielles 2020 des USA

PRÉSENTATION DES DONNÉES

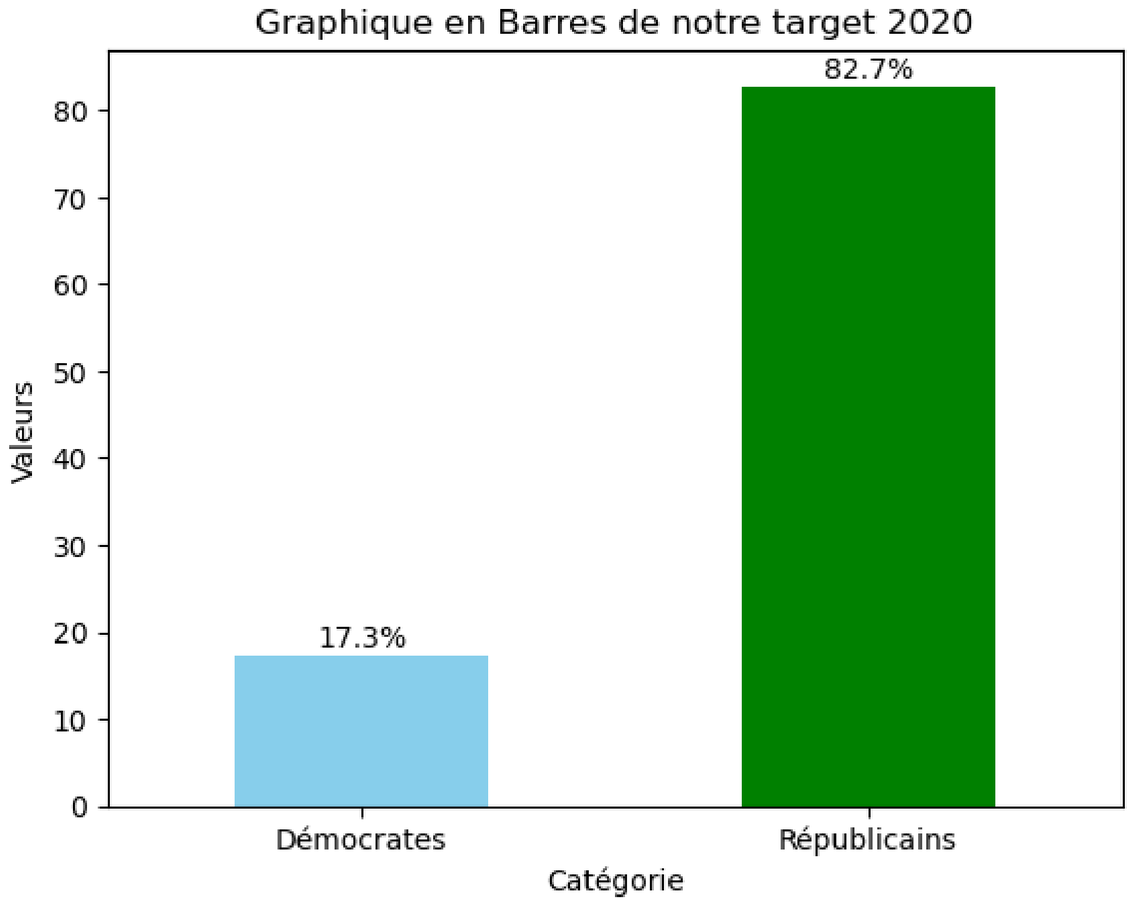
- On peut constater qu'il y a une très forte corrélation entre les valeurs manquantes donc la méthode adaptée pour traiter ses valeurs manquantes est la méthode KNN,
- Détection de valeurs aberrantes et remplacement par les moyennes
- La base finale : 3112 individus et 311 variables



PRÉSENTATION DES DONNÉES

ANALYSE EXPLORATOIRE

QUELQUES RÉSULTATS DE L'ANALYSE EXPLORATOIRE



	target_ 2008	target_ 2012	target_ 2016	target_ 2020
1	71.88	77.95	84.35	82.71
0	28.12	22.05	15.65	17.29

De 2008 à 2020 les républicains ont toujours gagné les élections avec un pourcentage très élevé

MODÉLISATION

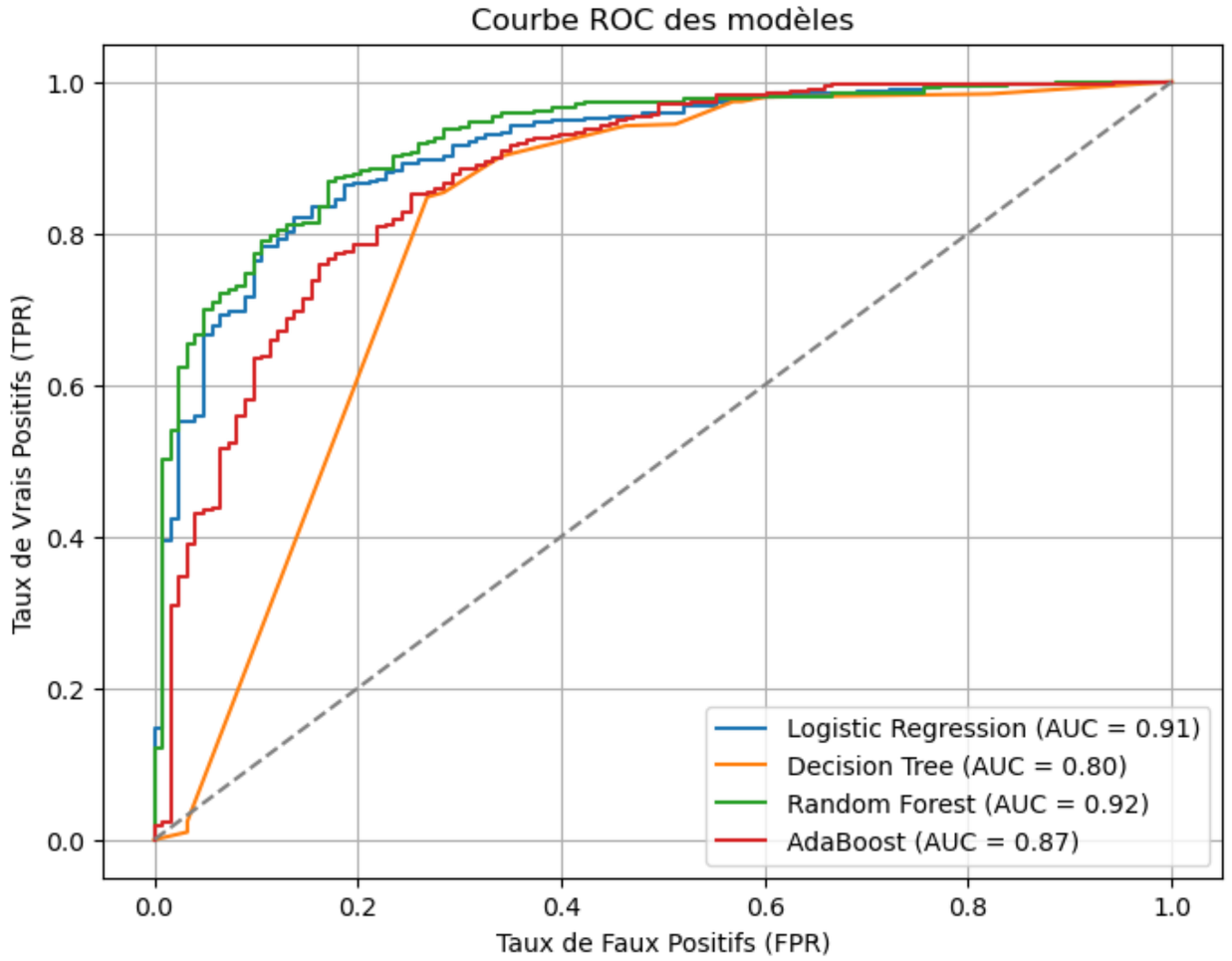
APPROCHE SCIENTIFIQUE

1. Prétraitement des données avec pipeline
2. Modélisation avec plusieurs algorithmes
3. Optimisation des hyperparamètres avec GridSearchCV
4. Évaluation des modèles avec plusieurs métriques
5. Visualisation des performances et des prédictions.

QUELQUES RÉSULTATS ET ORIENTATIONS STRATÉGIQUES ISSUS DE LA MODÉLISATION

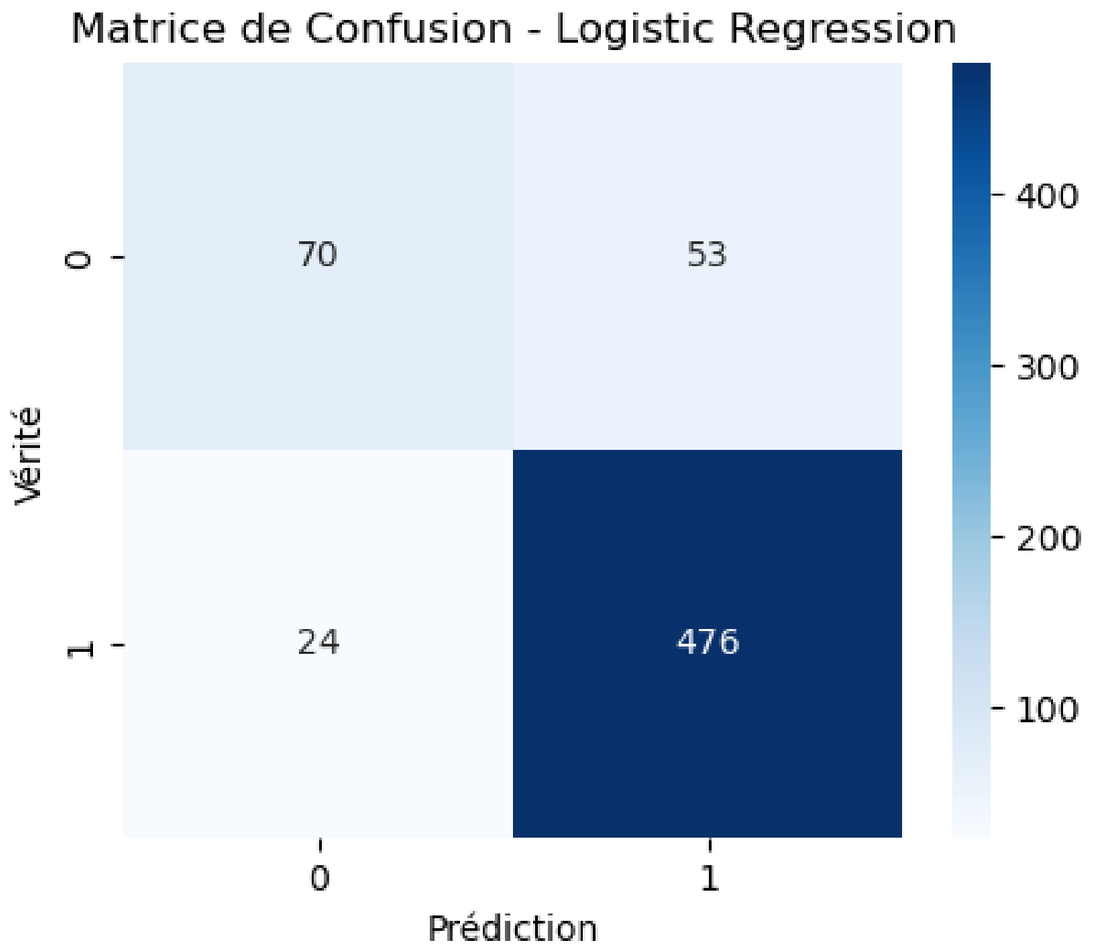
Le modèle forêt aléatoire (random forest) offre la meilleure combinaison de précision et de score F1, ce qui suggère qu'il est le plus performant.

Modèles	F1-score	Accuracy :
Logistic Regression	0.92	0.87
Decision Tree	0.91	0.85
Random Forest	0.93	0.88
AdaBoost	0.92	0.86

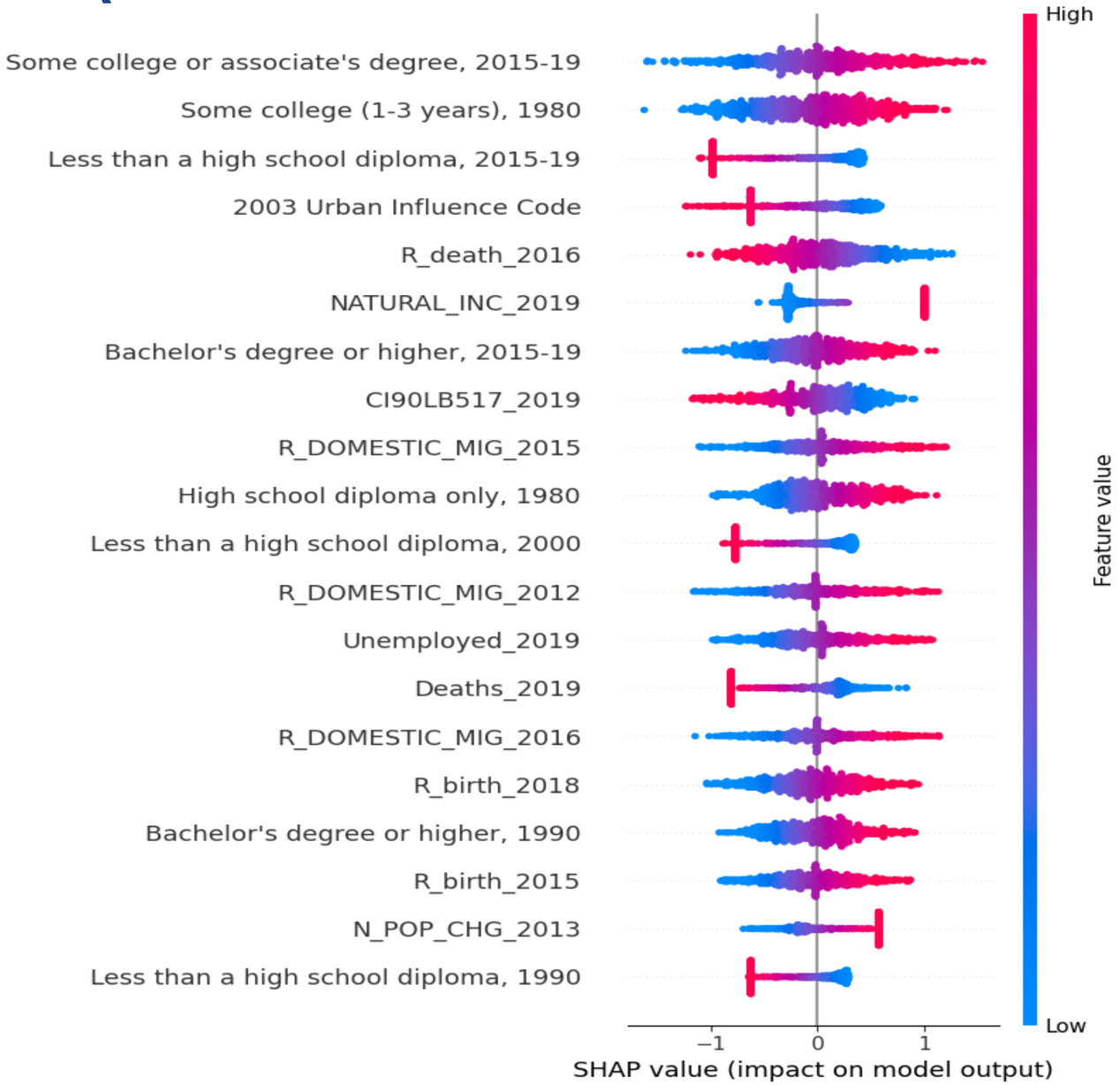


QUELQUES RÉSULTATS ET ORIENTATIONS STRATÉGIQUES ISSUS DE LA MODÉLISATION

- VN (70) : Correct pour le démocrate.
- FP (53) : Erreur, républicain au lieu de démocrate.
- FN (24) : Erreur, démocrate au lieu de républicain.
- VP (476) : Correct pour le républicain.



QUELQUES RÉSULTATS ET ORIENTATIONS STRATÉGIQUES ISSUS DE LA MODÉLISATION



- "Some college or associate's degree, 2015-19" → Une valeur élevée augmente la prédiction.
- "Less than a high school diploma, 2015-19" → Une valeur élevée diminue la prédiction.
- Unemployed_2019 et Deaths_2019 → Une valeur élevée baisse la prédiction.
- R_birth_2015/2018 → Effet plus variable.

QUELQUES RÉSULTATS ET ORIENTATIONS STRATÉGIQUES ISSUS DE LA MODÉLISATION

- Éducation : Mettre en avant des politiques éducatives pour attirer les électeurs instruits.
- Emploi : Proposer des solutions pour réduire le chômage et soutenir l'économie.
- Santé publique : Se concentrer sur la gestion des crises sanitaires pour rassurer l'électorat.
- Approche ciblée : Adapter les messages en fonction des enjeux urbains et ruraux.

CONCLUSION

Random Forest est le modèle le plus précis pour prédire les résultats des élections de 2020 (accuracy de 88,28 %), tandis que la régression logistique, plus simple, permet de mieux comprendre l'impact des variables comme l'éducation, le chômage et les décès. En résumé, bien que Random Forest soit le meilleur pour la précision, la régression logistique offre une meilleure interprétation des facteurs influençant les comportements électoraux. Pour une campagne efficace, il est crucial de se concentrer sur l'éducation, l'emploi et la santé.


LES ENSEIGNANTS
DATAGONG
Paul Thavenot
Guillaume Clément
Amed Coulibaly
Paul Leydier
Justine Charley
Romain Avouac
Alexandra Lorenzo
Jules Bonogo
Hippolyte
Louise Rodriguez

MERCI