Rapport d'Analyse Prédictive : Détection des Pages Publicitaires

Informations Générales

Nom : FAVEUR

Prénom : Kenson

Titre du projet : Prédiction des Pages Publicitaires

Objectif : Développer des modèles de machine learning pour distinguer les pages publicitaires ("pub") des pages non-publicitaires ("nonpub") à partir de 1554 variables numériques.

Introduction

Dans un contexte numérique où la publicité en ligne est omniprésente, il devient crucial de pouvoir distinguer de manière automatique les contenus publicitaires des contenus non-publicitaires. Cette capacité permet, par exemple, d'améliorer l'expérience utilisateur, de filtrer les contenus intrusifs ou encore de mieux cibler les analyses de données.

Le présent projet s'inscrit dans cette problématique. Il a pour objectif de développer des modèles d'apprentissage automatique capables d'identifier si une page web est publicitaire (pub) ou non (nonpub), à partir d'un ensemble de 1554 variables numériques décrivant leurs caractéristiques.

Pour cela, plusieurs modèles prédictifs ont été mis en œuvre, comparés et évalués à l'aide de métriques de performance classiques. Une attention particulière a été portée à la gestion du déséquilibre des classes, puisque la classe "nonpub" représente environ 86% des données disponibles, contre 14% pour la classe "pub".

Données et Prétraitement

L'ensemble de données utilisé comporte 3279 observations réparties de la manière suivante :

- 2624 observations (80%) pour l'entraînement des modèles
- 655 observations (20%) pour le test et l'évaluation finale

Chaque observation est décrite par 1554 variables numériques, toutes complètes (aucune valeur manquante). La variable cible, Y, est une variable catégorielle binaire prenant les modalités :

- pub : page publicitaire
- nonpub : page non-publicitaire

La distribution des classes est notablement déséquilibrée, ce qui justifie le recours à des indicateurs tels que la balanced accuracy, la sensibilité, et la spécificité, en plus de l'accuracy globale.

Le prétraitement a inclus :

- La normalisation des variables numériques
- La gestion du déséquilibre par évaluation sur plusieurs métriques
- La vérification de la colinéarité pour les modèles linéaires

Les modèles ont ensuite été entraînés sur l'échantillon d'apprentissage et testés sur l'échantillon de validation pour garantir la robustesse des résultats.

Méthodologie et Résultats des Modèles

Dans le but d'atteindre une classification efficace entre les pages publicitaires et non-publicitaires, plusieurs modèles d'apprentissage supervisé ont été implémentés et comparés. Chacun d'entre eux a été évalué selon les mêmes critères : accuracy, erreur globale, sensibilité, spécificité, balanced accuracy, ainsi que le coefficient Kappa, qui prend en compte le hasard dans les prédictions. Une attention particulière a été portée à la performance sur la classe minoritaire "pub".

Régression Logistique

La régression logistique a été utilisée comme modèle de base. Elle présente l'avantage d'être simple à interpréter et rapide à entraîner. Toutefois, en raison du très grand nombre de variables et de la colinéarité entre celles-ci, ses performances sont limitées, notamment sur la classe minoritaire.

- Accuracy : 91,6 %
- Erreur globale : 8,4 %
- Sensibilité (pub) : 94,5 %

- Spécificité (nonpub) : 73,6 %
- Balanced Accuracy : 84,1 %
- Kappa : 0,66

Interprétation : Le modèle détecte correctement la majorité des pubs, mais a tendance à mal classer les nonpubs. La colinéarité affecte négativement la spécificité.

Régression Lasso

La régression Lasso est une version pénalisée de la régression logistique, intégrant une régularisation de type L1. Cette approche permet une sélection automatique des variables les plus pertinentes, réduisant ainsi le risque de surajustement et améliorant l'interprétabilité du modèle.

- Accuracy : 96,6 %
- Erreur globale : 3,36 %
- Sensibilité (pub) : 98,8 %
- Spécificité (nonpub) : 83,5 %
- Balanced Accuracy : 91,1 %
- Kappa : 0,85

Interprétation : Le modèle offre d'excellentes performances globales et sélectionne un sous-ensemble efficace de variables. Cependant, il peut échouer dans des cas marginaux, comme observé sur un cas test mal classé.

Arbre de Décision (CART)

Le modèle CART (Classification And Regression Tree) construit une hiérarchie de décisions basée sur les variables les plus discriminantes. Il est particulièrement apprécié pour sa facilité d'interprétation et sa visualisation intuitive.

- Accuracy : 95,6 %
- Erreur globale : 4,42 %
- Sensibilité (pub) : 98,6 %
- Spécificité (nonpub) : 76,9 %

- Balanced Accuracy : 87,8 %
- Kappa : 0,80

Interprétation : L'arbre parvient à détecter très efficacement les pages publicitaires. Toutefois, il peut avoir tendance à sur-prédire cette classe, ce qui réduit légèrement sa spécificité.

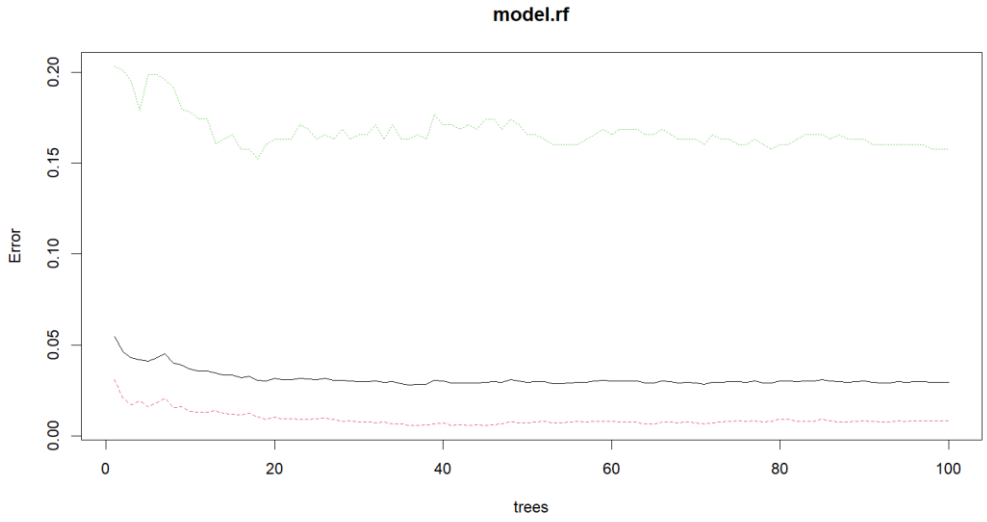
Forêt Aléatoire (Random Forest)

La Forêt Aléatoire repose sur l'agrégation de multiples arbres de décision construits sur des sous-échantillons aléatoires des données. Ce modèle est particulièrement adapté aux problèmes à haute dimension et présente une grande stabilité face au surapprentissage.

- Accuracy : 96,9 %
- Erreur globale : 3,05 %
- Sensibilité (pub) : 99,1 %
- Spécificité (nonpub) : 83,5 %
- Balanced Accuracy : 91,3 %
- Kappa : 0,87

Interprétation : C'est le modèle le plus performant de l'ensemble testé. Il combine une excellente détection des pubs à une bonne robustesse globale. L'erreur OOB (Out-of-Bag) diminue rapidement pour se stabiliser dès 30 arbres, démontrant sa fiabilité.

Graphe du forêt aléatoire



Étude de Cas et Résultat du Test

Un cas particulier de test, appartenant à la classe "pub", a été utilisé pour comparer les performances sur une observation concrète. Les prédictions sont les suivantes :

- Régression Logistique : correctement classé (pub)
- Régression Lasso : incorrectement classé (nonpub)
- Arbre de Décision : correctement classé (pub)
- Forêt Aléatoire : correctement classé (pub)

Conclusion : Seul le Lasso a échoué sur ce cas complexe, ce qui illustre que les modèles à base d'arbres, bien que moins explicables, restent plus robustes dans des cas limites.

Comparaison Globale des Modèles et Recommandations

Après l'évaluation individuelle de chaque modèle, une comparaison croisée permet de mettre en évidence leurs forces et leurs limites respectives. Cette étape est essentielle pour orienter le choix du modèle à utiliser en production ou pour des analyses explicatives.

Synthèse des Performances

Modèle	Erreur (%)	Sensibilité (pub)	Spécificité (nonpub)	Balanced Accuracy	Kappa
Régression Logistique	**8,4**	**94,5 %**	**73,6 %**	**84,1 %**	**0,66**
Régression Lasso	**3,36**	**98,8 %**	**83,5 %**	**91,1 %**	**0,85**
Arbre de Décision (CART)	**4,42**	**98,6 %**	**76,9 %**	**87,8 %**	**0,80**
Forêt Aléatoire	**3,05**	**99,1 %**	**83,5 %**	**91,3 %**	**0,87**

Cette synthèse met en évidence que tous les modèles testés sont capables de bien prédire la classe minoritaire "pub", avec des sensibilités très élevées (supérieures à 94 %). Cependant, seule la Forêt Aléatoire parvient à maintenir un équilibre optimal entre sensibilité et spécificité, avec une balanced accuracy supérieure à 91 % et la meilleure performance globale.

Avantages et Limites des Modèles

Régression Logistique: Simple et rapide, mais fortement affectée par la colinéarité des variables. Elle présente une bonne sensibilité mais une spécificité relativement faible, ce qui peut induire des faux positifs (classer à tort une nonpub comme pub).

Régression Lasso: Très performante et plus robuste face à la colinéarité. Elle sélectionne automatiquement les variables les plus pertinentes, ce qui la rend utile pour l'interprétation et la réduction de dimension. Néanmoins, elle peut échouer sur des cas atypiques, notamment les observations à la frontière entre les deux classes.

Arbre de Décision (CART): Offre une très bonne lisibilité et est facile à mettre en œuvre. Toutefois, en tant que modèle unique, il est sensible aux variations dans les données et tend à sur-prédire la classe minoritaire. Cela peut entraîner une légère baisse de spécificité.

Forêt Aléatoire: C'est le modèle le plus performant. Il combine robustesse, stabilité, et excellente capacité de généralisation. En contrepartie, son fonctionnement repose sur de nombreux arbres et il est donc plus difficile à interpréter.

Recommandations

Au vu des performances observées et de la nature du problème, les recommandations suivantes sont formulées :

Pour une mise en production : Le modèle Forêt Aléatoire est clairement recommandé. Il offre la meilleure précision, une bonne tolérance au déséquilibre des classes, et une grande robustesse même sur des données complexes ou bruitées.

Pour une analyse explicative ou exploratoire :La régression Lasso est une bonne alternative. Elle permet de comprendre les variables les plus influentes et de réduire la complexité du modèle tout en conservant de bonnes performances.

Pour l'amélioration des modèles :

- Mettre en place une réduction de dimension (par exemple via une sélection par importance ou PCA) afin de renforcer la stabilité des modèles.
- Ajouter des courbes ROC et des calculs d'AUC pour une évaluation plus fine des performances, notamment sur des cas limites.

- Appliquer une validation croisée (k-fold) pour obtenir une évaluation plus fiable et moins dépendante de la séparation initiale entre les jeux d'entraînement et de test.

Conclusion

Ce projet a permis de développer et d'évaluer plusieurs modèles d'apprentissage supervisé pour la prédiction automatique des pages publicitaires à partir d'un ensemble de 1554 variables numériques. Malgré un déséquilibre notable des classes, les modèles ont montré de très bonnes performances, en particulier sur la classe minoritaire "pub", souvent difficile à détecter.

La régression logistique, bien qu'intuitive, a montré ses limites en raison de la colinéarité entre les variables. Le Lasso, en revanche, s'est révélé particulièrement utile pour sélectionner les variables les plus discriminantes, tout en maintenant une performance élevée. L'arbre de décision, apprécié pour sa lisibilité, a bien performé mais reste sensible à la variabilité des données.

C'est cependant la Forêt Aléatoire qui s'est imposée comme le modèle le plus performant et le plus robuste, avec une précision élevée, un excellent équilibre entre sensibilité et spécificité, et une faible erreur de classification. Elle constitue donc le meilleur choix pour une mise en production.

En complément, il est recommandé d'intégrer des analyses ROC/AUC, de tester des techniques de réduction de dimension, et de mettre en œuvre une validation croisée afin de consolider la fiabilité des modèles à long terme.

Ce travail met en lumière l'efficacité des modèles d'ensemble, comme la Forêt Aléatoire, dans des environnements à haute dimension et à classes déséquilibrées. Il illustre également l'importance de combiner performance, robustesse et capacité d'explication selon les objectifs de l'analyse.