Online Shoppers Purchasing Intention Dataset

Christian AHADJI

Valentin FOSSE

Source : *Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018)*

Description

Ce dataset porte sur l'intention d'achat des acheteurs en ligne.

Il a été formé sur une période d'un an. Ceci pour éviter toute tendance à une campagne spécifique, un jour spécial, un profil utilisateur ou une période.

Chaque **ligne** appartient à une **session** d'un utilisateur.

Quelques points importants :

- Parmis les **12 330 sessions**, 84.5% (10 422) se terminent sans achat
- Il y a **18 attributs** : 10 numériques et 8 catégoriques
- Aucune valeur n'est manquante

Problème

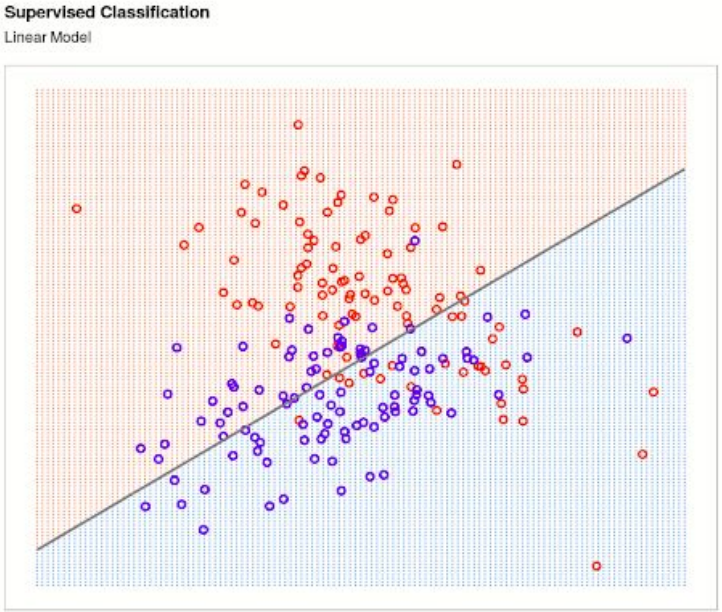
C’est un problème de classification.

Chaque session a un attribut nommé *Revenue*, il traduit le fait que l’utilisateur ait procédé à un achat durant cette session.

On peut donc, grâce à un modèle, prédire l’intention d’achat de chaque nouveau visiteur sur le site, d’où ce nom de dataset.

Cette classification est dite supervisée car la donnée est bien labellisée. Sa target value est *Revenue*.

Les autres variables de ce problème sont décrites par la suite.



Variables numériques

Administrative
Administrative_Duration
Informational
Informational_Duration
ProductRelated
ProductRelated_Duration

représentent le nombre de différents types de pages visitées par le visiteur au cours de cette session et le temps total passé dans chacune de ces catégories de pages

BounceRates
ExitRates
PageValues

représentent les statistiques mesurées par "Google Analytics" pour chaque page du site de e-commerce :

- *BounceRates* : pourcentage de visiteurs qui accèdent au site à partir de cette page, puis la quittent ("rebond")
- *ExitRates* : pourcentage que la page soit la dernière page visitée de la session
- *PageValues* : valeur moyenne d'une page Web visitée par un utilisateur avant d'effectuer une transaction

SpecialDay

indique la proximité de la date de visite du site avec un jour spécial spécifique (par ex. la St Valentin, entre le 2 février et le 12 février elle est maximale le 8 et est nulle avant et après cette période)

Variables catégoriques

Month : le mois durant lequel le visiteur se connecte

OperatingSystems : le système d'exploitation

Browser : le navigateur internet

Region : la région de connexion

TrafficType : le type de trafic

VisitorType : le type de visiteur, s'il est nouveau ou s'il est déjà venu

Weekend : un booléen, vrai si c'est le weekend

Revenue : notre variable de classification

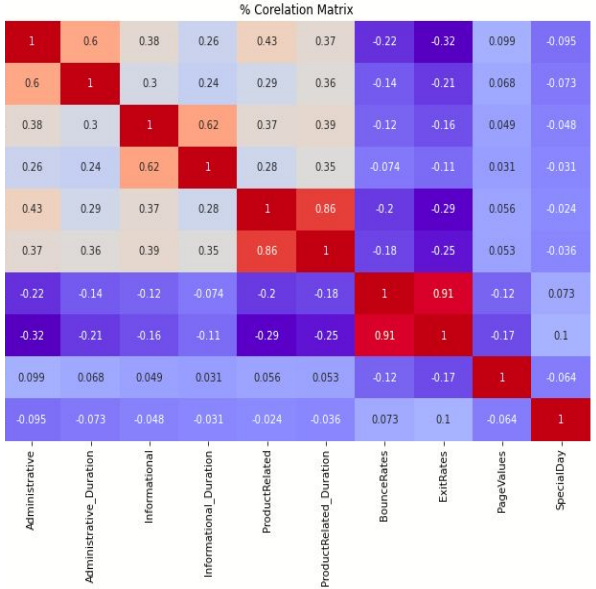
Variables créées

Deux variables ont été créées car l'étude menée a montré que certains attributs étaient très corrélés :

- *Administrative_Duration* et *Administrative* → ***ratio_administrative*** : elle peut être interprétée comme le temps passé par page dans la catégorie *Administrative*.

- *Informational_Duration* et *Informational* → ***ratio_informational*** : celle-ci par le temps passé par page dans la catégorie *Informational*.

Cela permet d'avoir un modèle plus simple, sans dédoublement de variables qui influe sur la précision et la stabilité du modèle.



Choix du modèle

Après un pre-processing des données (fusion de variables corrélées, création de dummies variables, partitionnement du dataset en train et test set, normalisation des données), elles sont maintenant exploitables.

Quatre algorithme ont été appliqués pour trouver le meilleur modèle, avec une cross validation et/ou une grille de réglage des hyperparamètres suivant les cas. Voici leur score final :

- *KNN* : 0.8691
- *Régression logistique* : 0.8758
- *Arbre de décision de classification* : 0.8933
- ***Forêt aléatoire*** : 0.9015

Nous retenons donc le modèle de forêt aléatoire, qui obtient le meilleur score.

Résultat

On a appliqué au modèle de forêt aléatoire une grille de recherche pour optimiser les hyperparamètres du modèle afin d'obtenir le meilleur score.

Celui-ci est obtenu pour ces paramètres :

- proportion d'attributs à prendre en compte à chaque split : 0.387904466210252
- nb minimum d'échantillons pour chaque nœud : 15
- nb d'estimateurs : 200

Le score retenu est **0.9015**.