TP 5 et 6 - Mini challenge sur la classification de données clients pour des prêts bancaires

Olivier Goudet

March 5, 2020

Présentation du challenge

Contexte

Lors de ces deux dernières séances de TP, vous allez participer à un mini challenge de classification binaire. Les données qui sont fournies sont des données réelles relatives à une campagne de promotion de prêts bancaires réalisée pour une institution bancaire. Cette campagne marketing s'est déroulée sous la forme d'appels téléphoniques auprès d'un grand nombre de clients, de façon à savoir si ils étaient prêts à souscrire à un produit bancaire (dépôt à terme de type assurance vie). Pour chaque client interrogé au cours de cette étude, on sait s'il a finalement souscrit un prêt bancaire (réponse oui, classe 1) ou bien s'il n'a pas souscrit (réponse non, classe 0).

A partir des données relevées à propos des clients au cours de cette campagne marketing, le but de ce TP est de prédire si chaque client aura tendance à souscrire au produit bancaire ou non. Ce genre de modèle prédictif pourrait typiquement être utilisé par des banques ou des compagnies d'assurance pour mieux cibler des offres de produit à des clients.

Attributs

Les attributs relatifs à chaque clients qui ont été collectés lors de cette campagne marketing sont décrits en anglais dans le jeu de données. Ce sont des données mixtes (catégorielles et continues). Il y a des attributs propres aux caractéristiques du client (1-7), des données relatives aux derniers contacts obtenus lors de la campagne d'appel téléphonique (8-14) et enfin des attributs qui définissent le contexte socio-économique général au moment de l'appel (15-19). Les définitions de ces attributs sont les suivantes :

1. age : âge du client
2. job : type d'emploi (ouvrier, employé, chômeur,...)
3. marital : statut marital (célibataire, marié, ...)
4. education : niveau d'éducation
5. default: a déjà un crédit en défaut de paiement (oui ou non).
6. housing: a déjà un crédit immobilier (oui ou non).
7. loan: a déjà un crédit personnel (oui ou non)
8. contact: type de communication (sur fixe ou portable).
9. month: mois du dernier contact
10. day_of_week: jours de la semaine lors du contact.
11. duration: durée du dernier contact téléphonique.
12. campaign: nombre de contacts téléphoniques pour ce client durant cette campagne.
13. pdays: nombre de jours passés depuis la dernière campagne d'appel téléphonique.

14. previous: nombre de d'appels déjà effectués avant cette campagne et pour ce client.
15. poutcome: résultat de la dernière campagne de marketing pour ce client (succès, non existant, échec).
16. emp.var.rate: variation du taux de chômage (indicateur trimestriel).
17. cons.price.idx: indice du prix à la consommation (indicateur mensuel).
18. cons.conf.idx: indice de confiance des ménages.
19. euribor3m: taux interbancaire à trois mois (indicateur journalier).
20. nr.employed: nombre d'employés.

Données fournies et challenge

Le challenge est organisé sur la plateforme Codalab accesible à l'url suivante : https://competitions.codalab.org/competitions/23664.

Sur Moodle ou bien directement sur la plateforme de challenge codalab sont fournis les fichiers suivants :

- Un dossier zippé nommé starting_kit.zip qui contient trois fichiers :
  - un fichier "bank_train_data.csv" qui correspond aux données relevées pour 9211 clients.
  - un fichier "bank_train_labels.csv" qui correspond aux classes associées à ces 9211 clients (1 ou 0), correspondant au fait que le client ait accepté ou non le produit bancaire proposé à l'issu de la campagne de promotion.
  - un fichier "bank_test_results.csv" qui correspond à un exemple de fichier de résultats avec des prédictions pour les 1002 clients de la phase de test du challenge.
- Un dossier zippé nommé public_data.zip qui contient un fichier "bank_test_data.csv" correspondant aux données pour la phase de test du challenge.

Pour cette compétition, il s'agira d'entraîner différents modèles vus en cours pour la classification binaire en utilisant les données d'entraînement fournies. Une fois votre modèle mis au point vous pourrez l'utiliser pour classifier les clients dont les données se trouvent dans le fichier "bank_test_data.csv". Vous pourrez ensuite soumettre ce fichier de prédiction sur le site du challenge et obtenir ainsi votre score de précision pour ce problème de classification des 1002 clients de la base. Un affichage des meilleurs résultats obtenus par les participants est ensuite réalisé sur la plateforme. Attention à bien valider votre modèle avant de soumettre vos prédictions sur le site car le nombre de soumissions total est limité à 10 par personnes (de façon à limiter des phénomènes de sur-apprentissage des données de test).

Prise au main de la plateforme de challenge CodaLab

Questions

1. Créez un compte sur la plateforme Codalab avec un nom d'utilisateur compréhensible "prenom_nom".
2. Inscrivez-vous au challenge Codalab en suivant le lien https://competitions.codalab.org/competitions/23664.
3. Dans l'onglet *Participate* puis *Files* téléchargez le starting kit qui contient les données d'entraîenemet mais aussi un modèle de soumission des résultats du challenge nommé "bank_test_results.csv".
4. Compressez directement ce fichier "bank_test_results.csv" dans un zip (sans le mettre au préalable dans un dossier) et faites une soumission de ce fichier zippé à partir de l'onglet *Participate* puis *Submit* et visualiser les résultats que vous obtenez sur le leaderboard du challenge.

Phase d'entraînement et de validation

Lors de cette phase vous devez travailler avec les données d'entraînement contenues dans les deux fichiers "bank_train_data.csv" (pour les entrées) et "bank_train_labels.csv" pour les labels (1 ou 0). Il s'agit de données étiquetées pour 9211 clients.

Questions

1. Charger les dataframes correspond à ces deux fichiers avec pandas.
2. Certains attributs sont catégoriels comme le type d'emploi ou le statut marital par exemple. De façon à pouvoir traiter ces attributs avec les algorithmes vus en cours il s'agit tout d'abord de réaliser un encodage binaire de ces variables catégorielles ("one hot encoding"). Pour cela on peut directement utiliser la fonction "get_dummies" de pandas sur le dataframe correspondant aux entrées. Affichez le dataframe résultant. Est-ce que vous comprenez l'opération effectuée ? Quel est l'intérêt de faire cette opération ?
3. Proposer ensuite un algorithme vu en cours de votre choix pour effectuer de la classification binaire. Il sera judicieux de séparer l'ensemble des 9211 clients fournis en un ensemble d'apprentissage et un ensemble de validation (cf. exemple du cours) de façon à évaluer la performance de l'algorithme sur des données qui n'ont pas été utilisées pour l'apprentissage. Cela permettra de régler les différents hyperparamètres du modèle et contrôler d'éventuels phénomènes de sur-apprentissage.

Phase de test

Lors de cette phase, il s'agira de télécharger les données de test du challenge et d'effectuer une prédiction pour les 1002 clients de cet ensemble.

Questions

1. A partir du ou des meilleurs modèles obtenus à la question précédente, effectuez une prédiction et soumettez les résultats obtenus sur le site du challenge. Pour effectuer une soumission il faut fournir un fichier csv du même type que le fichier "bank_test_results.csv" fourni dans le starting kit et le compresser au format zip directement (sans le mettre au préalable dans un dossier).
2. Une fois soumis, si l'évaluation est terminée sans erreurs, vous pourrez soumettre le résultat sur le tableau des scores du challenge.
3. Après avoir testé des modèles vus en cours, n'hésitez pas à tester d'autres modèles qui vous semblent judicieux.
4. Une meilleure normalisation des données en entrée et gestion des données manquantes peut aussi être un moyen d'améliorer les résultats.

Soumission du mini projet

A la fin de ces deux séances, soumettez sur Moodle une archive contenant l'ensemble des codes que vous avez utilisés pour produire les meilleurs résultats que vous avez obtenus lors du challenge ainsi qu'un rapport très court d'une page qui décrit rapidement les méthodes que vous avez utilisées, les scores correspondant obtenus, ainsi qu'une rapide interprétation de vos résultats.