FrenchData

Le Meilleur Data Scientist de France

2018

Informations générales

- Qscore : https://qscore.meilleurdatascientistdefrance.com
- Azure ML : https://studio.azureml.net/Home
- WIFI : MDSF / Mot de passe : MDSF2018
- Mot de passe de la compétition : **frenchdatarocks**

Préambule

Label Emmaüs propose à la vente en ligne des objets rénovés ou créés par le mouvement Emmaüs. Le but de ce challenge est d'estimer le **délai de vente** de chaque objet.

Description de la compétition

Label Emmaüs propose à la vente en ligne des objets rénovés ou créés par le mouvement Emmaüs. Son catalogue d'objets est en croissance régulière. L'ajout d'un objet au catalogue depuis la réception jusqu'à sa désignation (images, descriptif, entrepôt) n'est pas automatisé et prend de plus en plus de temps.

La détermination d'un prix et la rédaction d'une description ne sont pas toujours simples. Il faut un peu de temps et d'expérience pour traiter rapidement un objet. L'ajout d'un produit prend aujourd'hui **40 minutes** jusqu'à la mise en ligne. Beaucoup d'objets restent **longtemps** sur le site avant de trouver acquéreur. Et l'augmentation du catalogue nécessite de plus en plus d'espace de stockage, qui n'est pas toujours disponible.

Certains magasins vendent mieux que d'autres. Ceux qui vendent moins bien créent moins d'annonces : leur espace de stockage tend à saturer et l'activité n'est pas perçue comme assez performante pour investir. A la question "pourquoi mes produits ne se vendent pas", Label Emmaüs ne sait pas répondre (sauf par du ressenti métier).

L'objectif est donc de prédire **la durée entre la mise en ligne et la vente d'un objet**.

Les données contiennent environ **10 000 objets** (avec la description, le prix, la catégorie, etc.). Le délai de vente n'est connu que pour un sous-ensemble des produits (la base d'apprentissage).

Il faudra prédire la durée pour les produits de **la base de tests**.

La cible est catégorisée en **3 modalités** :

- **0 :** entre 0 et 10 jours
- **1 :** entre 10 jours et 60 jours
- **2 :** plus de 60 jours

Bonne chance à tous!

Métrique d'évaluation

Définissons les variables cibles comme étant des échantillons encodés comme 1, ..., K indicatrices donnant la matrice Y, i.e., $y_{i,k} = 1$ si l'observation i a le label k parmi un set de K labels de taille N.

Soit P, la matrice des probabilités estimées avec $p_{i,k} = \Pr(t_{i,k} = 1)$.
Alors la log loss est définie comme suit:

$$L_{\log}(Y, P) = -\log \Pr(Y|P) = -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{k=0}^{K-1} y_{i,k} \log p_{i,k}$$

Format de soumission

Le fichier soumis doit contenir **4 colonnes** :

- **colonne 1 :** id produit
- **colonne 2 :** probabilité que le produit appartient à la catégorie `delai_de_vente = 0` (entre 0 et 10 jours)
- **colonne 3 :** probabilité que le produit appartient à la catégorie `delai_de_vente = 1` (entre 11 et 60 jours)
- **colonne 4 :** probabilité que le produit appartient à la catégorie `delai_de_vente = 2` (plus de 60 jours)

Le séparateur de colonnes est la "`,`" (virgule).
Le séparateur décimal est le "`.`" (point).

Exemple:

```
id,0,1,2
0,0.05,0.81,0.14
```

Description des données

Nom variable	Description	Type
id	Identifiant du produit	Entier
nom_produit	Nom du produit	Texte
description_produit	Description du produit sur le site web	Texte
url_image	URL de l'image de présentation du produit	Texte
longueur_image	Longueur de l'image de présentation du produit	Entier
largeur_image	Largeur de l'image de présentation du produit	Entier
nb_images	Nombre d'image que contient la description du produit	Entier
taille	Taille du produit vestimentaire	Texte, Entier
pointure	Pointure du produit	Entier
matiere	Matière du produit (ex: Céramique, coton, verre, etc.)	Texte
garantie	Durée de la garantie du produit	Texte
annee	Année du produit	Entier
couleur	Couleur du produit	Texte
largeur_produit	Largeur du produit	Décimal
longueur_produit	Longueur du produit	Décimal
hauteur_produit	Hauteur du produit	Décimal
wifi	Si le produit électronique dispose de wifi ou non	Binaire
etat	Etat du produit	Texte
vintage	Si le produit est vintage ou non	Binaire
marque	Marque du produit	Texte
auteur	Auteur du produit (pour un livre)	Texte
editions	Edition du produit (pour un livre)	Texte
poids	Poids du produit	Décimal
nom_magasin	Nom du magasin dans lequel est vendu le produit	Texte
prix	Prix du produit	Décimal
categorie	Catégorie du produit	Texte
sous_categorie_1	Première sous-catégorie du produit	Texte
sous_categorie_2	Deuxième sous-catégorie du produit	Texte
sous_categorie_3	Troisième sous-catégorie du produit	Texte
sous_categorie_4	Quatrième sous-catégorie du produit	Texte

Règles

Règle 1

Le challenge est **individuel** (pas de travail en équipe).

Règle 2

Seules les soumissions réalisées par des personnes **inscrites avec leur prénom et nom réel** et **présentes physiquement** seront validées.

Règle 3

Il n'y a **aucune restriction** sur les technologies à utiliser, ni les plateformes.

Règle 4

Vous avez **2 heures** pour résoudre le challenge. Les soumissions réalisées avant et après la fin du concours ne seront pas prises en compte.

Règle 5

Donner le meilleur de soi-même et kiffer!

Et des coachs seront à votre disposition pendant la durée de la compétition pour répondre à toutes vos questions.

Engagement personnel de confidentialité des données fournies par Label Emmaüs

Considérant que les membres de chaque équipe auront vocation à recevoir des données considérées comme confidentielles ; afin d'assurer la protection de cette information et en considération du cadre défini pour l'événement, les participants conviennent de ce qui suit :

Article 1 – Définition des données

Les informations confidentielles (ci-après « Informations confidentielles ») divulguées par Label Emmaüs lors du Meilleur Data Scientist de France 2018 selon cet Engagement (ci-après « le Contrat ») recouvrent notamment l'ensemble de données fourni par Label Emmaüs, toutes les informations techniques liées à Label Emmaüs, l'aide liée aux données apportée aux candidats.

Article 2 – Définition des Organisateurs

Les organisateurs (ci-après désignés les « organisateurs ») recouvrent les entités et personnes ayant contribué à l'organisation de l'événement : Zelros, Microsoft, les coachs participant à l'événement.

Article 3 – Définition de la partie divulguante

Label Emmaüs, entreprise sociale immatriculée au Répertoire national des entreprises et des établissements sous le n° 821 489 002, dont le siège social est 104 Avenue de la Résistance, 93100 Montreuil, France ci-après dénommée « Label Emmaüs » ou « la Partie Divulguante ».

Article 4 – Durée de l'accord de confidentialité

Les dispositions de confidentialité prévues au présent accord s'appliqueront pendant toute la durée de celui-ci et pendant cinq (5) ans après son échéance ou sa résiliation quelle qu'en soit la cause.

Article 5 – Obligation de confidentialité

Le signataire pourra recevoir ou avoir accès à des données personnelles de clients de Label Emmaüs. Il reconnaît que ces informations sont protégées par la loi et qu'il doit les utiliser uniquement dans le cadre de sa participation à l'événement, en conformité avec les instructions qui lui sont données et de manière à ne pas placer Label Emmaüs ou les organisateurs en violation des lois relatives à la protection des données personnelles. Les données personnelles sont considérées comme des Informations Confidentielles et, par conséquent, ne doivent notamment pas être communiquées à un tiers ou transférées hors de France, sans l'autorisation préalable et écrite de Label Emmaüs.

Article 6 – Utilisation des données

1/ Le signataire devra utiliser, copier, reproduire et/ou divulguer les Informations confidentielles uniquement dans le cadre de sa participation au Meilleur Data Scientist de France 2018.
2/ Le signataire devra limiter la divulgation de ces Informations confidentielles aux membres de son équipe et ne devra, en aucun cas, divulguer ces Informations confidentielles à une tierce partie (particulier, entreprise, école, ou autre entité) sans l'accord écrit préalable de leur propriétaire.

Article 7 – Destruction des données

A la fin de l'événement, le présent Contrat impose au signataire l'obligation de supprimer les Informations Confidentielles reçues des différents appareils qu'il aura utilisés pour participer au Meilleur Data Scientist de France 2018, y compris toutes les copies réalisées, à la fin de l'événement.

Article 8 – Interprétation de l'accord

Ce Contrat ne doit pas être interprété comme créant, transmettant, transférant, consentant ou conférant au signataire quelque droit, licence ou autorité concernant l'ensemble de données fourni, à l'exception du droit d'usage limité spécifié à l'Article 2. De plus et en particulier, aucune licence ou transmission des droits de propriété intellectuelle n'est créée ou impliquée par ce Contrat.

Article 9 – Violation de l'accord

En cas de manquement ou de violation de quelque disposition que ce soit de cet engagement de confidentialité, le signataire comprend et accepte que les organisateurs disqualifieront l'équipe avant d'initier des poursuites judiciaires. Le signataire s'engage à avertir sans délai les organisateurs et Label Emmaüs de toute violation du présent Contrat.

Article 10 – Intégralité de l'accord

Le Contrat expose dans son intégralité l'accord né entre les parties concernant la divulgation d'Informations confidentielles et à ce titre, remplace et annule tout précédent accord, compréhension ou représentation à ce sujet.

Article 11 – Loi applicable

Le Contrat est fait sous et doit être interprété au regard de la loi française.

Article 12 – Juridictions compétentes

En cas de violation ou de manquement à ce Contrat, tout différend entre les parties devra être soumis aux juridictions françaises compétentes. En conséquence, les participants confirment avoir pris connaissance et compris le Contrat et acceptent librement les devoirs et obligations qui en découlent. Le signataire est informé que sa responsabilité pourrait être recherchée en cas de manquement à cet engagement.