**Fondamentaux – Le Projet**

Qu’est-ce qu’un bon projet ?

* Machine Learning
  + Prédiction : prédire y (variable cible) en fonction de x

On va utiliser l’ensemble des couples x,y qu’on connait pour entrainer le modèle M

Du coup, avec ce modèle M, il sera possible de calculer les valeurs manquantes de y (M(x) = y)

Y peut être le Churn (quitte le service ou non), Credit-scoring (octroi ou non un crédit)

Prédire la température en fonction des émissions de CO², le salaire en fonction de XP, diplôme …, le prix d’une maison en fonction de l’historique des ventes, l’agrément (piscine, jardin, ..)

* + Trouver un jeu de données :
    - Sur GCP (BigQuery) :! Ressources > Accéder à des données publiques
    - Kaggle

Se limiter à un jeu de données :

* + - * à moins de 20 mo
      * avec au moins 30000 lignes
      * identifier son y, ce qu’on veut prédire

|  |  |  |  |
| --- | --- | --- | --- |
| **x1** | **x2** | **xn** | **Y** |
| fff | 134 | 12-mars | 10 |
| fgtr |  | 13-mars | 30 |
| gghr | 456 | 14-mars | 5 |

Entrainement :

D -> M : x1, x2, …,xn -> y

Evaluation performances :

- Train Set : on vérifie si y predicted = y sur la partie de la BDD utilisée pour l’entrainement du modèle

- Test Set : on vérifie si y predicted = y sur une partie de la BDD non utilisée pour l’entrainement du modèle

<https://www.kaggle.com/lava18/google-play-store-apps#googleplaystore.csv>

Nécessité d’être aidé pour la conversion de certaines variables (Size, Last updated, nb install -> mettre en binaire…)

Variable cible à prédire : à définir les 2 catégories : <1.000.000 et >=1.000.000 ??

Voir la répartition et peut-être checker pour une répartition 50/50

Overfitting

1/ Dataset plus gros

2/ Décomplexifier son modèle

Undefitting

* plus rare

Les régressions linéaire et continue fonctionnent très bien avec des variables continues

Si trop de var catégorielles : arbre de décision

**PRESENTATION DU PROJET :**

* présenter les spécificités / difficultés liées au dataset, problématique

1/ Contexte : dataset, que prédire

2/ Améliorations (transfo de var …)

3/ Interprétation des résultats (ex : matrice de confusion)

F1 score on training set : 0.6202967673555909

F1 score on test set : 0.5887445887445887

Autres métrics :

**Accuracy** : si Y pa trop déséquilibré

=(TP+TN)/ N

**Precision** : TP/(TP+FP) plus c’est élevé, plus le modele a eu raison parmi tous les positifs qu’il y avait

**Recall** : TP/(TP+TN) Combien de fois on a eu raison parmi

(cf Wikipedia où c’est bien expliqué)

**F1 Score**