Record Linkage Comparison Dataset

Par Etienne VINCENT et Charles XU

Dataset

Comme tout problème de *Record Linkage*, on veut savoir systématiquement si deux id provenant de deux tables avec une structure différente (qui ont été préalablement fusionnées) correspondent à une seule et même personne.

Ce dataset recense pour chaque ligne, les paires de prénoms et noms de famille, la date de naissance, le sexe et le zip code. Il est composé de 5 749 132 de lignes et 12 colonnes. On veut donc prédire si oui ou non (classification binaire supervisée) deux id sont liés à une personne en fonction des prédicteurs. Seul 0,36% des lignes matchent une personne, chaque paire dans les 99,64% restants ne font pas référence à une même personne.

Ces données proviennent du University Medical Center of the Johannes Gutenberg University Mainz en Allemagne.

Features

- id_1 et id_2 : ids internes correspondant à un individu

Mis à part id_1 et id_2 ainsi que la variable is_match (label booléen de supervision), chaque variable prend une valeur entre 0 et 1 et correspond à un score de confiance définis par ceux qui ont annoté le dataset. Cela permet à notre classification d'être plus efficace pour le traitement des 5M de lignes.

- cmp_fname1 et cmp_fname2 : correspond à la paire de prénoms extraits des deux dataset
- cmp_lname1 et cmp_lname2 : correspond à la paire de noms extraits des deux dataset

Features

Ces features peuvent-être égales à 0 ou à 1 respectivement si l'on ne peut pas du tout affirmer que la variable correspond effectivement à un individu et 1 si l'on est sûr que la valeur de cette variable est propre à l'individu. Bien sûr la valeur peut aussi se situer entre 0 et 1.

Par exemple si cmp_plz est égal à 1 alors on est certain que la paire id_1 et id_2 d'une ligne partagent la même adresse.

- cmp_sex : sexe
- cmp_bd : jour de naissance
- cmp_bm : mois de naissance
- cmp_by : année de naissance
- cmp_plz : zip code

Data exploration et prétraitement

Variables d'identification :

- id_1 : 97 283 valeurs uniques sur 5 749 133
- id_2 : 97 312 valeurs uniques sur 5 749 133
- is_match : 20 931 TRUE sur 5 749 133

Data exploration et prétraitement

	count	mean	std	min	25%	50%	75%	max
cmp_bd	5749133.0	0.224434	0.417209	0.0	0.000000	0.000000	0.000000	1.0
cmp_bm	5749133.0	0.488788	0.499874	0.0	0.000000	0.000000	1.000000	1.0
cmp_by	5749133.0	0.222718	0.416070	0.0	0.000000	0.000000	0.000000	1.0
cmp_fname_c1	5749133.0	0.712777	0.388839	0.0	0.285714	1.000000	1.000000	1.0
cmp_fname_c2	5749133.0	0.016234	0.125199	0.0	0.000000	0.000000	0.000000	1.0
cmp_lname_c1	5749133.0	0.315628	0.334234	0.0	0.100000	0.181818	0.428571	1.0
cmp_lname_c2	5749133.0	0.000136	0.010081	0.0	0.000000	0.000000	0.000000	1.0
cmp_plz	5749133.0	0.005516	0.074067	0.0	0.000000	0.000000	0.000000	1.0
cmp_sex	5749133.0	0.955001	0.207301	0.0	1.000000	1.000000	1.000000	1.0
is_match	5749132.0	0.003641	0.060228	0.0	0.000000	0.000000	0.000000	1.0

Tableau résumant les valeurs dans chaque variable

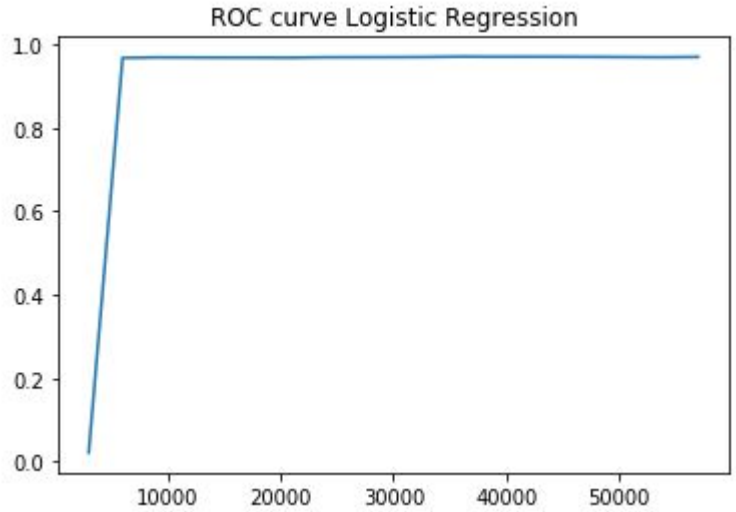
Modélisation - Approche Déterministe

Logistic Regression Classifier

Score : 96,95%

Matrice de confusion :

```
[[  19868    1063]
 [     93 5728108]]
```



Modélisation - Approche Probabiliste

Approche probabiliste : Ajout du poids sur chaque variable selon leur fréquence (frequencies.csv inclus dans le .zip)

Logistic Regression Classifier

Score : 96,95%

Matrice de confusion :

```
[[  19868    1063]
 [     93 5728108]]
```

Modélisation - Approche Probabiliste

Naive Bayes Classifier

Matrice de confusion	F Score	Précision	Rappel
[[19740 1191] [12 5728189]]	0,97	99,93%	94,23%

Modélisation - Approche Probabiliste

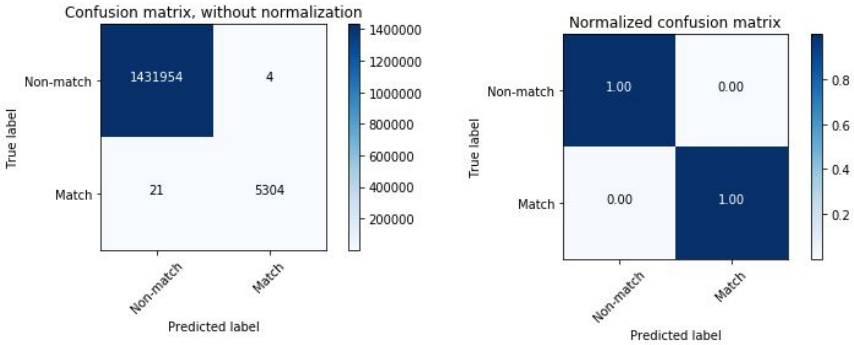
SVM: Support Vector Machine, SVM for Record Linkage

Matrice de confusion	F Score	Précision	Rappel
[[20719 212] [10 5728191]]	0,993	99,93%	98,83%

Modélisation - Approche Probabiliste

SVM: Support Vector Machine with Grid Search

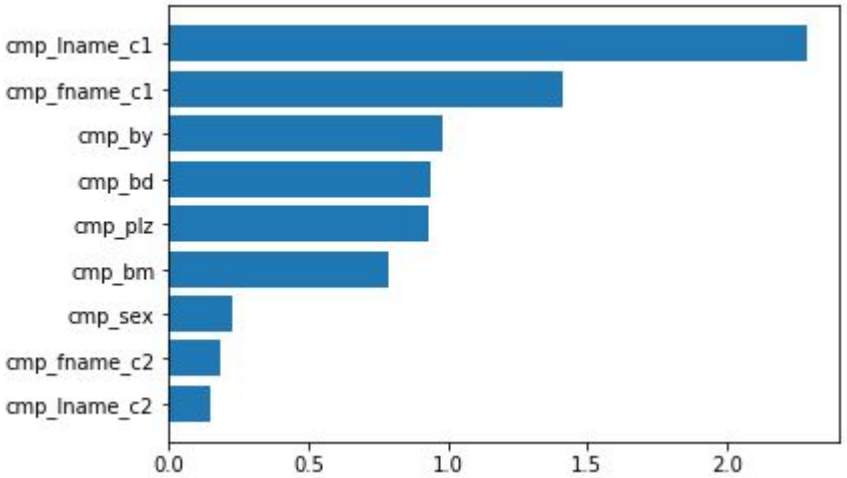
F Score	Précision	Rappel
0,997	99,99%	98,58%



Feature importance

Il est plus difficile de prédire avec les prénoms si les deux personnes ont le même sont la même personne qu'avec le nom de famille.

On remarque que les variables cmp_by, cmp_plz, cmp_bd, cmp_bm, cmp_sex sont les plus importants.



API

L’API est développée sous Django et permet de stocker et prédire par rapport aux informations entrées.

Elle permet d’insérer, de modifier ou de supprimer un élément dans la base.

Lorsqu’un objet est inséré, il est possible de prédire la variable cible avec le modèle SVM entraîné.