Prédiction du sexe à partir de données textuelles

Fynch Meynent
juliette.meynent@ensae.fr
Github du Projet

Avril 2024

1 Présentation de la tâche

Les recensements sont des données qui intéressent grandement les démographes et archivistes. Dans le cadre du projet Scorface, ils s'associent à des informaticiens pour en extraire les informations à grande échelle. Le problème est que, les données ayant été collectées durant un siècle, les nomenclatures ont pu varier selon les vagues, ce qui rend plus difficile l'exploitation des données.

En effet, dans certains recensements, le sexe n'est pas renseigné. Cependant, c'est une donnée sociologique essentielle pour l'analyse des ménages ainsi que des trajectoires genrées des individus. Ainsi, l'objectif sera de prédire le sexe d'une personne à partir de ses données personnelles pour combler ce manque afin de permettre d'uniformiser les données et les analyses.

2 Présentation et Description des données

Nous utiliserons deux bases de données différentes : un fichier de prénoms avec le nombre d'hommes et de femmes ayant porté les prénoms, ainsi qu'un fichier contenant des extraits de recensement retranscrits de manière manuelle et automatique. Nous n'avons pas ajouté de base de données externe, parce que nous n'estimions pas pouvoir obtenir davantage d'informations sans décrire des cas trop spécifiques, ce qui risquait d'impliquer de l'overfitting.

2.1 Base 1 : Prénom par sexe

Cette base de données nous donne pour 6 946 prénoms, le nombre de personnes l'ayant porté par sexe. La plupart des prénoms sont soit clairement féminins, soit clairement masculins (Voir Figure 1 et Figure 2). Les prénoms rares sont également présents dans cette base (Exemple : Zéraphin).

2.2 Base 2 : Extrait de recensement

Cette base comporte 241 individus issus du recensement, dont 125 hommes, 107 femmes et 9 personnes ambiguës. Nous avons retiré les 9 personnes ambiguës, car elles n'apportaient pas d'information. Leurs informations ont été retranscrites de manière automatique et manuelle. Nous avons conservé seulement les données manuelles, car elles étaient davantage complètes et semblaient plus propres. Nous connaissons, pour chaque individu, leur nom de famille, leur sexe, leur prénom, leur emploi, certains liens familiaux... Nous n'avons pas conservé toutes les données parce que toutes ne donnent pas d'indice sur le sexe, par exemple l'âge. Cependant, nous observons un grand nombre de valeurs manquantes pour les catégories autres que le prénom qui pourrait pourtant nous apporter des informations sur le sexe de la personne(Voir Table 1).

2.3 Objectifs de l'analyse

Notre objectif est d'appliquer ce modèle à l'ensemble des données de recensement de 1836 à 1936. De 1836 à 1936, la population est recensée dans son intégralité tous les 5 ans. On peut

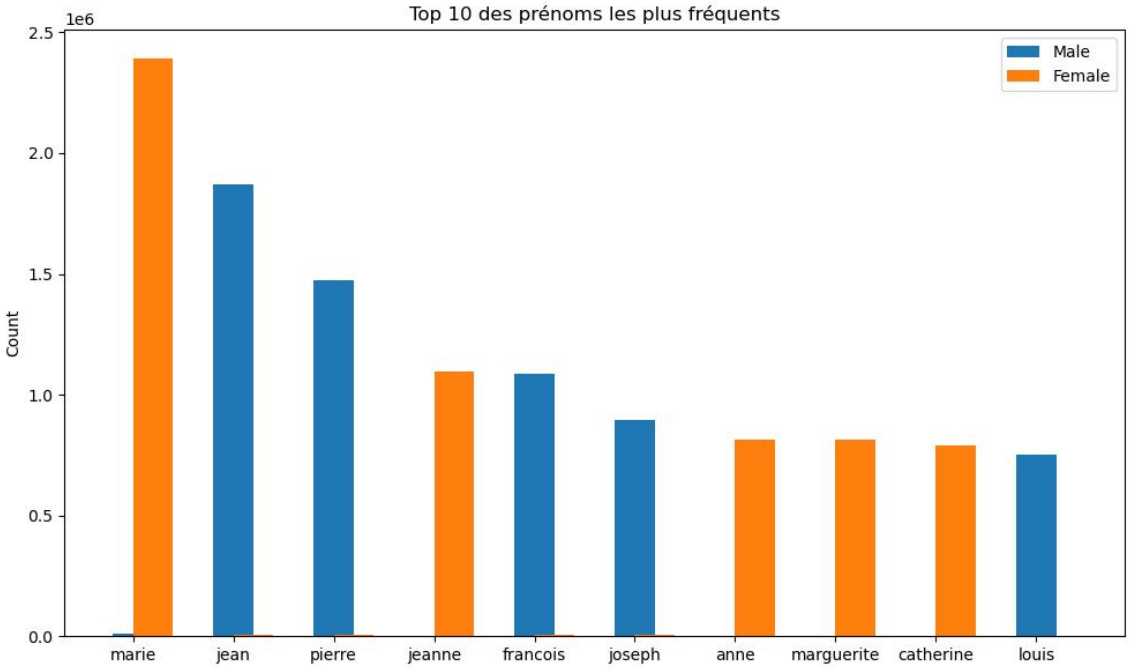


FIGURE 1 – 10 prénoms les plus fréquents par base répartis par sexe

Column	Non-Null Count
surname	231 non-null
firstname	231 non-null
occupation	190 non-null
link	221 non-null
employer	46 non-null
civil_status	3 non-null
sex	232 non-null

TABLE 1 – Valeurs manquantes dans la base 2

donc estimer qu'il y a eu une vingtaine de recensements durant la période. [1] De plus, la population française comptait 34 293 000 habitants en 1836 et 41 500 000 en 1936. [2]. On va donc estimer qu'il y avait environ 39 millions d'individus en moyenne à chaque recensement (les chiffres étant souvent aux alentours de 40 millions en tombant parfois en dessous surtout avant 1896 et immédiatement après la Première Guerre mondiale). On va donc supposer que le nombre d'observations visées est donc :

$$taille = 39000000 * 20 = 780000000 \quad (1)$$

L'objectif est donc de créer un modèle pour prédire le sexe d'environ 780 000 000 observations [3].

3 Analyse des modèles

Notre objectif est de prédire le sexe de la personne. Il s'agit donc d'un problème de classification binaire. Un premier arbitrage est nécessaire. Si nous utilisons que la base 2, nous aurons seulement 233 observations. Cela peut donc conduire donc à des difficultés de généralisation du modèle.

1. Histoire du recensement en France, Wikipédia
2. Histoire Démographique de la France, Wikipédia
3. Nous ne parlons pas de personnes car celles-ci sont redondantes entre les recensements.

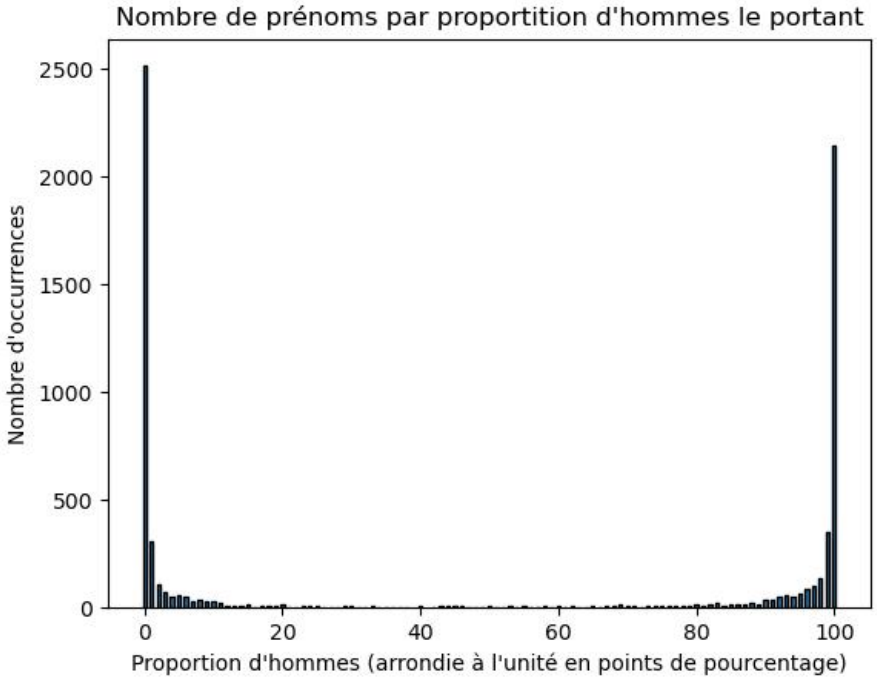


FIGURE 2 – Nombre de prénoms par la proportion d'hommes le portant

Mais si nous utilisons que la base 1, nous n'aurons que les prénoms en guise d'informations, même si cela permet d'avoir 6946 observations d'entraînement. Pourtant, l'emploi ou le lien familial pourrait être un indice. En effet, certaines personnes sont par exemple labellisées "épouse" ce qui signifie que l'observation est une femme.

Étant donné le faible nombre d'observations dans la base 2 et les valeurs manquantes en son sein, nous avons préféré nous restreindre au prénom afin de compléter la base 2 par la base 1 en y ajoutant les prénoms qui ne se trouvaient pas dans la base 2. De plus, cela rendrait le modèle applicable à davantage de données, car le prénom est systématiquement recensé, alors que les autres nomenclatures peuvent varier sur un siècle.

Pour labelliser un prénom à chaque sexe dans la base 1, nous avons calculé la probabilité que la personne soit d'un sexe à partir de la formule suivante pour chaque sexe :

$$proba_{sexe} = \frac{nb_{sexe}}{nb_h + nb_f} \tag{2}$$

où :

— nb_h correspond au nombre d'hommes portant ce prénom.

— nb_f est le nombre de femmes portant ce prénom.

Nous avons créé deux mesures du sexe possibles, car cela n'est pas explicité dans la base 1 comme tel. Le premier labellisant « homme » si la probabilité d'être un homme est supérieure ou égale à 95%, femme si la probabilité d'être un homme est supérieure ou égale à 95%, « ambigu » sinon.

Également, nous avons aussi fait un label majoritaire qui prend « homme » si la probabilité d'être un homme est plus forte que d'être une femme, « femme » sinon.

Ayant peu de données textuelles, les modèles de Deep Natural Language Processing ne nous paraissaient pas adaptés. Connaissant les labels, nous nous tournerons vers des méthodes classifications supervisées. Des modèles telles que des régressions linéaires ne sont pas non plus idéales pour la classification binaire, car nous sommes contraints de choisir un seuil brut à partir duquel on classifie « homme » plutôt que « femme ». De plus, ayant peu de variables, les modèles tels que des arbres de décision ou random forest ne semblent pas convenir.

Par conséquent, nous avons opté pour une régression logistique dont nous détaillerons les modalités plus tard. Cela estimera la probabilité qu'une personne soit un homme et la classifiera en conséquence. Sa structure permet d'avoir un seuil plus "soft" dans l'établissement d'une classification qu'avec une forme quelconque de régression linéaire. C'est pour cela qu'on a choisi ce modèle.

4 Expérimentation

4.1 Protocole

Nous allons chercher à prédire le sexe de chaque observation. Nous comparerons en choisissant dans un cas le label au seuil de 95%, et dans un cas le label majoritaire. Dans la base à 95%, nous retirerons les valeurs ambiguës. Pour la partie de la base partant de la base 2, le label sera le même dans la modélisation 95% et dans la modélisation majoritaire.

Les variables explicatives seront définies à partir du prénom. En effet, le prénom en soit ne constitue pas une information que l'on peut traiter de manière mathématisée. Ainsi, nous avons tenté de définir les prénoms par des variables binaires reprenant les informations suivantes.

— Le nombre de lettres. En effet, nous supposons que les prénoms féminins sont souvent plus longs du fait qu'on les accorde par rapport à leur équivalent masculin. (Exemple : "René" devient "Renée")

— La dernière lettre. En effet, nous supposons que les prénoms féminins finissent plus souvent par un "e" ou un "a" que les prénoms masculins. D'une part, du faire de l'accord féminin en "e" et d'autre part, du fait que les prénoms en "a" paraiss"nt souvent féminins (exemple : Theodora, Magdalena. Bien que ce soient plutôt des prénoms rares).

— Les deux et trois dernières lettres. En effet, certaines structures semblent récurrentes pour les prénoms féminins. Par exemple, la fin en "-ine" (Exemples : Claudine, Caroline...). Également, nous avons le cas double consonne + e liée à des accords féminins. (Exemple : Jeanne). Les trois dernières lettres semblent représentatives, mais les deux dernières peuvent aussi contenir des patterns intéressants. Par exemple, peut-être que la fin des prénoms en -ne peut également être corrélé au fait d'être un prénom féminin.

Toutefois, cette approche pourra présenter des limites. Par exemple, le prénom "Etienne" sera probablement mal classifié au vu de nos hypothèses. Pour obtenir des valeurs numériques pour implémenter notre modèle, nous avons transformé en variable binaire les terminaisons. Ainsi, la colonne "finit par un A" indiquera True pour Magdalena et False pour Henri. Et ainsi de suite pour "fini par ne" "finit par ine" (Voir Table 2).

Index	last_with_e	last_with_ne	last_with_ine
1	False	False	False
2	True	False	False
3	False	False	False
4	False	False	False
5	True	True	False

TABLE 2 – Extrait de notre base de données

Nos variables explicatives seront donc le nombre de lettres combiné aux nombreuses variables binaires définies. Nous avons essayé de réduire la dimensionalité par une Analyse en Composantes Principales (PCA) mais elle donnait une précision plus faible (voir Annexe), nous avons donc décidé de ne pas la conserver parce que l'algorithme prenait peu de temps à tourner même sans réduction de dimensionalité.

Nous avons associé à homme la valeur 1 à « homme » et -1 à « femme ». L'estimation s'effectue de la manière suivante. Puisqu'on a fait une régression logistique, on estime d'abord :

$$\hat{p} = \frac{1}{1 + \exp^{tX\theta}} \tag{3}$$

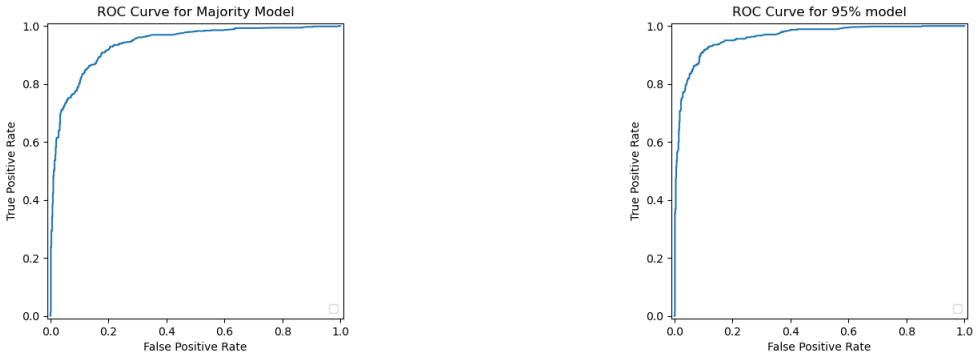
Où
— X correspond aux variables explicatives.
— θ correspond aux poids optimaux associés aux variables que l'on cherche à estimer durant l'apprentissage.

Si $\hat{p} \geq 0.5$, on associe alors '1' à la variable, '-1' sinon.

On divise notre jeu de données de la manière suivante : 80% seront des données d'entraînement, et 20% des données tests. Ainsi, on estimera les poids sur 80% du jeu de données et on estimera la précision que cela engendre sur les données tests. Nous procéderons à une validation croisée pour vérifier que la précision ne change pas selon les parties de la base de données sélectionnées. Nous essaierons également d'autres partitions (par exemple 66% de données d'entraînement, 33% de données test).

4.2 Résultats

Pour le modèle majoritaire, nous obtenons une précision de 86%, et ce, quels que soient la partition et le découpage effectué entre les jeux de données d'entraînement et de tests. Pour le modèle 95%, nous obtenons une précision de 91%. Cependant, nous perdons 1000 observations dans le second modèle, car nous ne gardons pas les labels ambigus (7 033 dans le premier modèle contre 6 035 ici).



(a) Modèle majoritaire

(b) Modèle 95%

FIGURE 3 – Comparaison des ROC Curve

L'aire sous la ROC Curve du modèle 95% semble plus élevée, ce qui rend ce modèle plus convaincant pour la tâche demandée (Voir Figure 3). Pour ces deux modèles, la validation croisée, quelle que soit la manière de découper les données d'entraînement, semble indiquer des taux de précision similaires.

5 Conclusion

Pour étendre une prédiction à l'ensemble du corpus, nous recommanderons d'utiliser le modèle 95%. En effet, cela devrait être possible, car le prénom est présent dans chaque corpus et c'est celui qui offre la meilleure précision parmi nos deux propositions.

Cependant, notre méthode présente de nombreuses limites : elle ne traite pas les prénoms légèrement épicènes et classifiera mal également les personnes portant un prénom associé au sexe opposé, ce qui pourrait être corrigé avec d'autres données professionnelles ou familiales que nous n'avons pas pris en compte ici.

De plus, les données étant larges, elles devront nécessairement passer par la transcription automatique. Or, nous avons conservé uniquement les données manuelles qui semblaient mieux retranscrites. Cela peut donc poser problème en fonction de la méthode de récolte des données et des algorithmes de reconnaissance d'écriture.

A Régression Logistique avec PCA : Résultat

En appliquant le modèle 95% sur des variables explicatives sous forme de PCA expliquant 95% de la variance des données, nous obtenons une précision de 90%. Cela n'est pas extrêmement différent du modèle sans PCA, mais étant donné que le modèle donne un résultat rapidement sans PCA, nous avons préféré conserver le 1% de précision supplémentaire en prenant davantage de variables.

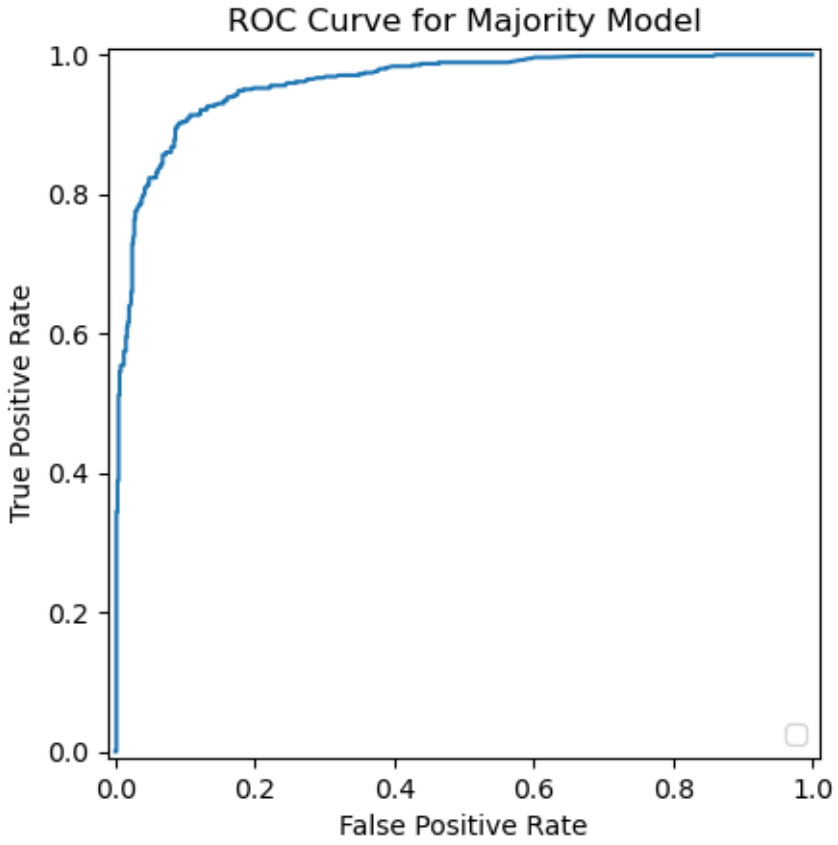


Figure 4 – ROC Curve pour le modèle 95% avec PCA