Rendre explicable un processus opaque : Simuler l'admission sur Parcoursup à partir de données publiques

Martin Astyan par Prevision

Version révisée du 28 octobre 2025

Résumé

Nous proposons une méthodologie transparente pour *expliquer* et *approximer* un processus d'admission largement opaque — Parcoursup — en nous fondant uniquement sur des données publiques et quelques hypothèses parcimonieuses. Notre contribution est un modèle multiplicatif explicable qui transforme des agrégats (mentions, types de baccalauréat, spécialités) en un score calibré en percentile, accompagné de règles de décision (*refusé / en attente / accepté*). Nous détaillons le pipeline de données (y compris les cas où un scraping ponctuel a été nécessaire faute d'open data), les formules utilisées, les bornes appliquées, ainsi qu'une validation qualitative et un exemple chiffré reproductible.

1 Contexte et bref historique

L'ancienne plateforme APB a été remplacée par Parcoursup à partir de 2018 afin d'améliorer l'affectation des candidats et de répondre aux critiques (notamment autour des classements et du tirage au sort dans des filières en tension). Malgré des avancées, la granularité publique des critères et des pondérations reste limitée : les formations publient des *indications*, mais rarement des barèmes opérationnels complets. Cette opacité motive notre démarche : rendre le processus *intelligible* pour les candidats en exposant des hypothèses simples, des formules claires et des sorties interprétables.

2 Jeux de données, collecte et normalisation

2.1 Périmètre et provenance

Nous exploitons principalement les jeux de données agrégés publiés chaque année (admis par mentions, type de bac, parts de spécialités admises par famille de formation, *etc.*). Pour certaines formations, des informations complémentaires (classement des « doublettes » les plus fréquentes, coefficients dossier/concours) n'étaient pas disponibles dans l'open data lors de la préparation : elles ont été *scrapées ponctuellement* depuis les pages publiques de Parcoursup, aux mêmes millésimes que les fichiers utilisés, et consolidées avec un contrôle manuel. Lorsque l'open data a comblé ces manques sur des millésimes ultérieurs, nous privilégions les sources officielles et déprécions les extractions.

2.2 Nettoyage et variables dérivées

Les notebooks de préparation uniformisent le schéma (noms de colonnes, types, pourcentages), contrôlent les valeurs manquantes et dérivent les variables suivantes : (i) moyenne μ et dispersion proxy σ à partir de la distribution des mentions admises ; (ii) part d'admis par type de bac p_b ; (iii) indicateurs d'équité (genre, bourse) ; (iv) parts d'admis par *doublette* de

spécialités p_{spe} et moyenne des top doublettes $\bar{p}_{\text{top}}$. Les coordonnées géographiques éventuellement présentes sont séparées en colonnes lat/lon pour d'éventuels diagnostics, sans lien avec une intégration applicative.

3 Modèle : multiplicateurs interprétables et formules

Le score M d'un candidat pour une formation f est le produit de multiplicateurs centrés en 1 :

$$M = M_{\text{sexe}} \cdot M_{\text{bourse}} \cdot M_{\text{bac}} \cdot M_{\text{notes}} \cdot M_{\text{notes-spe}} \cdot M_{\text{doublette}} \cdot M_{\text{lycee}}.$$

Chaque composant est plafonné pour éviter les effets extrêmes. Nous convertissons ensuite M en un *percentile* $P \in [0, 1]$ (§3.7), puis en une décision (§3.8).

3.1 Effets démographiques (bornés)

Deux effets symétriques autour de 1, plafonnés à 2.5 % d'écart :

$$M_{\text{sexe}} = \exp(k_{\max}\,(2B_f - 1) \cdot s)\,,$$
$$M_{\text{bourse}} = \exp(k_{\max}\,(2B_b - 1) \cdot t)\,,$$

où $B_f, B_b \in [0, 1]$ proviennent des parts relatives observées parmi les admis, $s, t \in \{-1, +1\}$ encodent respectivement homme/femme et non-boursier/boursier, et $k_{\max} = \ln(1.025)$.

3.2 Notes générales (calibrées cohorte)

Soit g la moyenne du candidat, μ et σ la moyenne et la dispersion proxy de la formation :

$$z = \frac{g - \mu}{\sigma}, \quad p = \Phi(z), \quad M_{\text{notes}} = \begin{cases} \exp\big(1.25\,\ln(m_{\max})(2p-1)\big) & \text{si } g < \mu, \\ \exp\big(\ln(m_{\max})(2p-1)\big) & \text{sinon,} \end{cases}$$

avec $m_{\max} = 1.5$ et Φ la CDF normale standard.

3.3 Cohérence des spécialités

La moyenne des deux notes de spécialités $(s_1 + s_2)/2$ est comparée à μ :

$$n = \frac{(s_1 + s_2)/2 - \mu}{5}, \quad n \in [-1, 1], \qquad M_{\text{notes-spe}} = \begin{cases} \exp\big(2.5\,\ln(1.5)\,n\big) & \text{si } n < 0, \\ \exp\big(\ln(1.5)\,n\big) & \text{sinon.} \end{cases}$$

3.4 Type de bac (représentativité)

Pour les bacs non généraux, on applique un plancher $m_{\min} = 0.6$ et une mise à l'échelle par p_b :

$$M_{\text{bac}} = \begin{cases} 0 & \text{si } p_b \le 0, \\ \min(\max(3p_b, m_{\min}), 1) & \text{sinon.} \end{cases}$$

3.5 Rareté/adéquation de la doublette

Avec p_{spe} part d'admis ayant la doublette candidate et $\bar{p}_{\text{top}}$ la moyenne des N doublettes les plus fréquentes (typiquement N=3), on pose $r = \min(p_{\text{spe}}/\bar{p}_{\text{top}}, 1)$ et

$$M_{\text{doublette}} = \begin{cases} \exp\big(2\,\ln(1.15)\,(2r-1)\big) & \text{si } r < 0.5, \\ \exp\big(\ln(1.15)\,(2r-1)\big) & \text{sinon.} \end{cases}$$

Si un *top* explicite des doublettes est publié, un multiplicateur fixe $[1.03, 1.15]$ par rang peut être utilisé.

3.6 Lycée d'origine (contexte académique)

Pour refléter l'environnement du candidat sans introduire de biais excessif, nous intégrons un multiplicateur borné issu des indicateurs publics du lycée d'origine (taux de réussite et mentions au baccalauréat, effectif terminale, note moyenne de classement, etc.). Le score agrégé du lycée, noté $x \in [0,1]$, est normalisé selon :

$$z = \frac{x - \mu_x}{\sigma_x}, \quad p = \Phi(z), \qquad M_{\text{lycee}} = \text{clamp}\big(\exp(k\,(2p-1)),\, m_{\min},\, m_{\max}\big),$$

où $\mu_x = 0.548$, $\sigma_x = 0.182$, $m_{\min} = 0.85$, $m_{\max} = 1.15$ et $k = \ln(m_{\max})$. Ainsi, un lycée nettement au-dessus de la moyenne nationale augmente légèrement le score global (jusqu'à $+15\%$), tandis qu'un lycée moins favorisé réduit légèrement le multiplicateur (jusqu'à -15%). Si aucune information fiable n'est disponible (*infos*$= 0$ ou métrique manquante), on applique $M_{\text{lycee}} = 1.0$.

Ce facteur, bien que borné et symétrique, permet d'introduire un signal de contexte académique tout en limitant l'effet des écarts extrêmes.

3.7 De M au percentile calibré

Nous supposons M distribué autour de 1 et choisissons σ_M tel que le 97.5$^{\text{e}}$ centile corresponde à $M_{\max} \approx 2.5$:

$$\sigma_M = \frac{M_{\max} - 1}{\Phi^{-1}(0.975)}, \qquad z_M = \frac{M-1}{\sigma_M}, \qquad P = \Phi(z_M).$$

3.8 Décision explicable

Nous restituons un *pourcentage* $100 \times P$ et une décision à trois niveaux, explicite et stable :

$$\begin{aligned} \text{Refusé :} \quad & 100P < 20, \\ \text{En attente :} \quad & 20 \le 100P < 50, \\ \text{Accepté :} \quad & 100P \ge 50. \end{aligned}$$

4 Exemple chiffré (valeurs factices)

Formation avec $\mu = 13.2$, $\sigma = 1.8$; parts observées $B_f = 0.52$, $B_b = 0.30$, $p_b = 0.82$; doublette avec $p_{\text{spe}} = 0.35$ et $\bar{p}_{\text{top}} = 0.42$. Candidate femme ($s = +1$), boursière ($t = +1$), bac général, moyenne $g = 13.8$, notes de spécialités $s_1 = 14.0$ et $s_2 = 15.0$. Lycée d'origine : informations disponibles (`infos`=1) et métrique agrégée $x = 0.6696$ (environ 75$^{\text{e}}$ centile des lycées).

Composants :

- $z = \frac{g-\mu}{\sigma} = 0.333 \Rightarrow p = \Phi(z) = 0.631 \Rightarrow M_{\text{notes}} = 1.112$;
- $n = 0.260 \Rightarrow M_{\text{notes-spe}} = 1.111$;
- $M_{\text{sexe}} = 1.001$, $M_{\text{bourse}} = 0.990$, $M_{\text{bac}} = 1.000$;
- $r = \min(p_{\text{spe}}/\bar{p}_{\text{top}}, 1) = 0.833 \Rightarrow M_{\text{doublette}} = 1.098$.
- Lycée d'origine : $z_x = \frac{x-\mu_x}{\sigma_x} = \frac{0.6696-0.548}{0.182} = 0.668$, $p_x = \Phi(z_x) = 0.748$, $k = \ln(1.15)$, $M_{\text{lycee}} = \exp\big(k\,(2p_x - 1)\big) = \mathbf{1.072}$ (déjà borné dans $[0.85, 1.15]$).

Produit total : $M = 1.112 \times 1.111 \times 1.001 \times 0.990 \times 1.000 \times 1.098 \times \mathbf{1.072} = \mathbf{1.440}$ (arrondi). Avec $\sigma_M = 0.765$, on a $z_M = \frac{M-1}{\sigma_M} = \frac{0.440}{0.765} = 0.576$ et $P = \Phi(z_M) = \mathbf{0.718}$, soit $100P = \mathbf{71.8}\% \Rightarrow$ **Accepté**.

5 Validation qualitative, explicabilité et limites

Validation. Empiriquement, le modèle classe *plutôt justement* les profils en trois niveaux sur des jeux de cas exploratoires : les « Accepté » dominent pour des dossiers nettement au-dessus de la cohorte et des doublettes usuelles ; les « Refusé » concentrent des profils en fort décalage. *Cependant*, les seuils (20/50) doivent être **réajustés** sur des exemples précis (par formation et par année) pour améliorer l'alignement local. La simplicité des règles facilite ces calibrations itératives.

Explicabilité. Chaque multiplicateur est reportable au candidat (« notes sous la moyenne de la cohorte », « doublette moins fréquente », « type de bac peu représenté »), avec des bornes explicites limitant les effets indésirables.

Limites. Agrégation vs micro-décision réelle ; hypothèse log-normale implicite pour M ; critères non observables (projets, lettres, quotas spécifiques). Les effets démographiques étant bornés et symétriques, ils ne doivent pas être interprétés comme prescriptifs.

6 Reproductibilité

Organisation : `metric.py`, `admission.py`, `utils.py`, `variables.py` et notebooks de nettoyage. Procédure : exécuter les notebooks de préparation, configurer les chemins, puis appliquer les formules ci-dessus à des profils de test. **Aucune intégration applicative n'est décrite ici** : nous nous concentrons sur la modélisation et l'explicabilité.

7 Conclusion

Nous avons proposé une démarche transparente pour rendre intelligible un processus d'admission largement opaque. À partir de données publiques (complétées, lorsque nécessaire, par des extractions ponctuelles), nous avons construit un modèle multiplicatif explicable qui transforme des indicateurs observables (mentions, type de baccalauréat, spécialités, signaux d'équité) en un score normalisé puis en une décision à trois niveaux (*Refusé* < 20, *En attente* 20–49, *Accepté* ≥ 50). Chaque composant est borné, documenté et interprétable, ce qui permet de relier la décision à des facteurs concrets (*notes sous/sus la cohorte*, *doublette plus/moins fréquente*, *représentativité du bac*, etc.).

La validation qualitative suggère que le modèle classe *plutôt justement* les profils, tout en indiquant que les seuils doivent être révisés sur des exemples précis par formation et par année. Cette simplicité assumée favorise la reproductibilité, l'auditabilité et l'appropriation par les usagers, sans prétendre reproduire l'algorithme interne des formations. Les limites identifiées (agrégation des signaux, hypothèse de distribution pour le score, critères non observables) sont explicites, ce qui contribue à une lecture honnête des résultats.

Perspectives

(i) Calibrer empiriquement les seuils par filière et millésime ; (ii) quantifier l'incertitude (intervalles autour du pourcentage) ; (iii) tester la robustesse par analyses de sensibilité (notes, spécialités, pondérations) ; (iv) suivre les dérives temporelles (*drift*) des distributions de candidats ; (v) approfondir les audits d'équité sur sous-populations. Ces pistes visent à renforcer, dans la durée, l'explicabilité et l'utilité pratique du modèle.

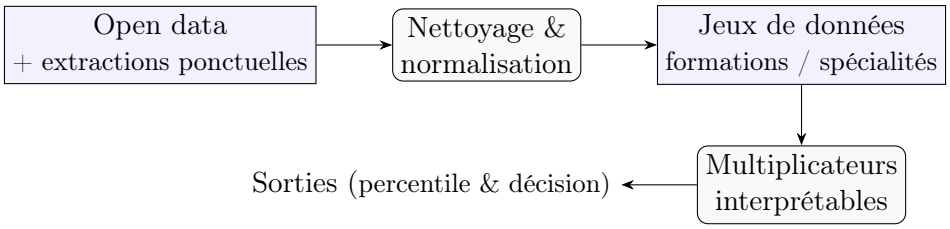

FIGURE 1 – Chaîne de traitement, orientée explicabilité (sans volet applicatif).

8 Schémas (synthèse)

Paramètres et bornes (référence rapide)

Composant	Paramètres
M_{sexe}, M_{bourse}	$m_{\max} = 1{,}025$
M_{notes}	$m_{\max} = 1{,}5$, pénalité $\times 1{,}25$ sous μ
$M_{\text{notes-spe}}$	$m_{\max} = 1{,}5$, pénalité $\times 2{,}5$ sous μ
M_{bac}	$m_{\min} = 0{,}6$
$M_{\text{doublette}}$	$m_{\max} = 1{,}15$ (ou $[1.03, 1.15]$ si top $1 \dots 5$)
M_{lycee}	$m_{\min} = 0{,}85$, $m_{\max} = 1{,}15$, μ_x=0,548, σ_x=0,182
Percentile	$M_{\max} = 2{,}5$ (97.5^{e})

Références

- Ministère de l'Enseignement supérieur. *Parcoursup – Plateforme nationale d'admission dans l'enseignement supérieur*, disponible en ligne : https ://www.parcoursup.gouv.frhttps ://www.parcour
- République Française. *Data.gouv.fr – Plateforme ouverte des données publiques françaises*, disponible en ligne : https ://www.data.gouv.frhttps ://www.data.gouv.fr
- Service Public. *Parcoursup : la plateforme nationale d'admission*, notice officielle, disponible en ligne : https ://www.service-public.fr/particuliers/vosdroits/F32446https ://www.service-public.fr/particuliers/vosdroits/F32446
- OpenDataFrance. *Réseau national des acteurs de l'open data*, disponible en ligne : https ://www.opendatafr