Using Machine Learning to Predict Admissions to Higher Education

Wanjing Anya Ma
Graduate School of Education
Stanford University
wanjingm@stanford.edu

Léon Marbach
Graduate School of Education
Stanford University
lmarbach@stanford.edu

1 Introduction

In France, in order to get into dental, medical, midwifery or pharmacy school, students first need after high school to complete a one-year "common pre-health studies" program[1]. Our goal is to predict admissions, more precisely students' rankings for admissions, for this selective program in 2020 at one case study institution, "French University".

For this supervised learning task, we input information submitted by applicants to French University to output rankings established by the admissions staff (no *ex æquo*). Although we have the precise individual ranking of each student, we make the choice of categorizing rankings into quintiles and perform a classification rather than a regression task because we are interested in whether the performance of our models varies from the top- to the bottom-ranked students. After an exploratory data analysis, we first use Multinomial Logistic Regression as base model and compare its results with three tree-based classification algorithms: simple Decision Tree, Random Forest and AdaBoost. We then use Stepwise Regression to extract meaningful features.

In a national context of lasting inequalities in access to higher education based on social background but also type of high school attended and early tracking in secondary education, it is important to understand what elements of the applications are valued or not by the French University admission staff when sorting applicants. Recurrent criticism of the poor "human qualities" of health professionals in the French debate also make it critical to evaluate the role of universities that get to decide who can get into health professions.

Given the French tradition to place a high emphasis on academic criteria rather than on personality or extracurricular activities and accomplishments, our hypothesis is that ranking of applicants is mostly based on high school grades, in particular in "hard sciences" classes related to health (e.g. Physics-Chemistry), but also on the national standardized *Baccalauréat* tests[2] due to their high symbolic value in French culture.

2 Related Work

Most of literature on access to higher education has looked at what happens before and after admissions, but only a few studies have examined the actual process of sorting applications. The reason is the difficulty to obtain such data because of the reputational risks for universities. Among the researchers who have had this opportunity, some have used a qualitative approach with field observations and interviews of admission committees (Stevens, 2009; Posselt, 2016), while others have used machine learning.

The supervised learning models that have been tested in the past to predict admissions are Logistic Regression, Linear Support Vector Machines, Naive Bayes, Decision Trees, Random Forest, AdaBoost, Gradient Boost, and Multilayer Perceptron classifiers (Lux et al., 2016; Rees and Ryder, 2022; Neda and Gago-Masague, 2022; Lee et al., 2023). The perspective adopted by these studies was to evaluate the extent to which humans could be assisted or even replaced to perform admission decisions. Instead, we adopt a more sociological line of inquiry by focusing on how machine learning can help unveil social preferences and biases in admission decisions. Moreover, these previous studies only predicted admission/rejection (binary classification), while we predict ranking quintile, a more complex task.

[1] For details on the organization of health education in France, see van Zanten et al., 2021.

[2] The *Baccalauréat* is a series of oral and written tests all French high school students have to take at the end of both the last two years of high school. Only the grades obtained at the tests taken in the penultimate year are available because the others are taken after applications are submitted.

3 Dataset and Features

Our dataset[3] contains information submitted by the 15,167 applicants (14% admission rate), in particular: sociodemographic information (gender, age, country of birth, financial aid status), academic background (high school public/private status, *Dominante* (major) and *Spécialité* (minor) chosen in high school, grades out of 20 obtained in each subject in each quarter of the last two years of high school[4] as well as on the *Baccalauréat* tests, whether the student is doing an international *Baccalauréat*[5], whether the student repeated a grade in high school, as well as evaluations of the student by the high school teaching team on five items[6]) and minor chosen for the "pre-health studies" program[7]. This makes up a total of 60 features.

For the purpose of this project, because our main goal is to evaluate different machine learning methods to predict admission ranking quintiles, we only retain the $N = 9{,}575$ applicants that were enrolled in the last year of high school in France, who had the French citizenship, and who were enrolled in the scientific track[8] when applying. This ensures that the whole population is comparable (e.g. students in different high school tracks take different classes and *Baccalauréat* tests). Compared to the original application files, students' mandatory personal statements and teachers' remarks in transcripts are also not included in our analysis, while other variables were made less precise to preserve the anonymity of students.

4 Exploratory Data Analysis

In this section, we explore key variables that we hypothesize to contribute most to the ranking prediction. These include students' subject grades in their second quarter of their last year of high school (most recent grades available to the admission staff at the time of application review) and their grades on *Baccalauréat* tests. Table 1 shows that Physics-Chemistry, Maths, *Dominante* (chosen "major", which is necessarily a scientific subject in the scientific track), and French Oral in *Baccalauréat* are most correlated with the ranking. It is clear that the admissions committee valued mostly the "hard sciences" compared to the "social sciences" and Physical Education.

We now expand our exploratory analysis to all 45 grade variables (last two years of high school and *Baccalauréat* tests) using PCA. Figure 1 shows the first two PC scores for each observation colored

Subjects	Mean Grade (SD)	Correlation to Ranking
Physics-Chemistry	11.94 (3.35)	0.80
Maths	11.61 (3.79)	0.78
Dominante	13.32 (2.91)	0.71
Baccalauréat French Oral	13.72 (3.56)	0.67
History-Geography	13.20 (3.06)	0.58
Spécialité	13.57 (3.25)	0.56
Baccalauréat French Written	11.37 (3.46)	0.56
Philosophy	11.87 (2.81)	0.56
Baccalauréat Group Project	15.39 (3.22)	0.46
Physical Education	15.57 (2.66)	0.20

Table 1: Correlation of grades in the second quarter of the last year of high school and in *Baccalauréat* tests to ranking.

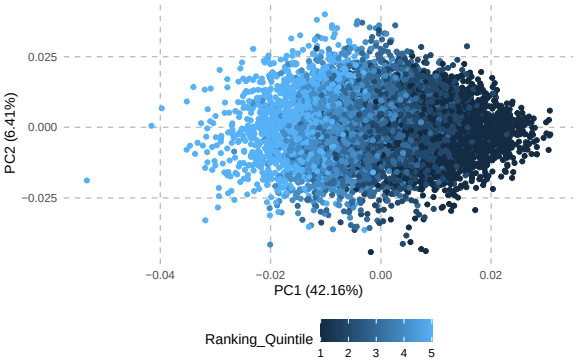

Figure 1: PC1 and PC2 scores from PCA on all grade variables, colored by ranking quintile.

Subjects	Coefficient
Maths (last year, second quarter)	0.225
Maths (last year, first quarter)	0.216
Maths (penultimate year, third quarter)	0.203
Physics-Chemistry (last year, second quarter)	0.198
Maths (penultimate year, second quarter)	0.196

Table 2: Top 5 loadings of PC1 from PCA on all grade variables.

[3]The dataset (not shareable) was obtained by Léon Marbach as part of a project directed by Agnès van Zanten at Sciences Po, see here.

[4]For the last year of high school, so-called *Terminale*, only the first two quarters of grades are available because students submit their applications before the end of the third quarter.

[5]Students doing an international *Baccalauréat* study the same curriculum but have certain classes and *Baccalauréat* tests in a foreign language.

[6]The five items are: working methods (scale 1-4), ability to work independently (scale 1-4), community engagement (yes-no), motivation (scale 1-4), and potential to succeed in higher education (scale 1-4)

[7]When applying, students had to choose the minor they were willing to pursue if admitted to the program among four possible: Biology-Physics-Chemistry, Health Population, Law or Economics.

[8]At the time of the study, French students were separated into three different high school tracks: scientific, economics and social sciences, and literature.

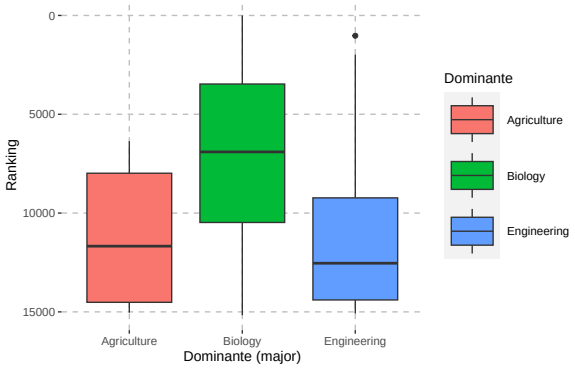

Figure 2: Box plot of original ranking by *Dominante*.

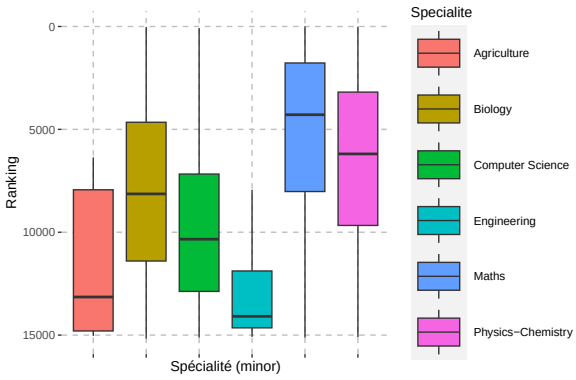

Figure 3: Box plot of original ranking by *Spécialité*.

by ranking quintile (label 1 corresponding to the top 20%). We do not scale our data because all grades are in the same unit out of 20. While the result in Figure 1 only shows one cluster, it appears that the PC1 scores, which explain 42.16% of the total variance, describe the ranking quintiles well. A one-way ANOVA relating PC1 scores to ranking quintiles confirms at the 0.001 level that we can reject the null hypothesis that the average values of PC1 are the same across all five ranking quintiles. Looking at the factor loadings, of which the top 5 are shown in Table 2, we can interpret that students' rankings is mostly influenced by Maths and Physics-Chemistry scores, especially the most recent grades available at the time of admission. At the bottom of the factor loadings are grades in Physical Education at any time with a coefficient close to 0.

We finally explore the relationship between students' chosen high school *Dominante* (major) or *Spécialité* (minor) and ranking. Among the three available majors in the scientific track, Figure 2 shows that students in Biology are more likely to get a higher ranking than those in Engineering or Agriculture. As for the choice of minor, Figure 3 reveals that students minoring in Maths and Physics-Chemistry are the best ranked on average.

5 Methods

5.1 Data Pre-Processing

Using previous findings, we convert the values of *Dominante* (major) and *Spécialité* (minor) as two numeric values. Specifically, we give students a major or minor score based on how likely the subjects contribute to the ranking positively. For example, students will receive 3 if they major in Biology, 2 in Engineering, and 1 in Agriculture.

Grade variables in our dataset contain missing values, in particular, on any given class, 1 to 3% of the students have no grade (maybe because the student was sick or, in the case of Physical Education, injured). This is a problem because most machine learning models can only be fitted on complete data. After trying around multiple methods such as using median, mean, mode or random values with the base model, we decide to use median in the following analysis as it provides slightly better accuracy scores than other methods.

5.2 Evaluation Methods

We use 5-fold cross-validation on 80% of our data for model comparison and evaluate final model on the held-out 20% data. We treat ranking as a classification problem, where we classify the rankings into quintiles. Label 1 means the top 20% and label 5 means bottom 20%.

5.3 Model Comparison

We first include all $p = 60$ features and fit four machine learning classifiers to explore which model is more suitable for our ranking prediction task.

We treat Multinomial Logistic Regression with softmax coding as our base model. We use Newton-CG solver because it is known for converging quickly and suitable for large-scale problem like our case.

We then compare this base model to Decision Tree and two other tree-based approaches: Random Forest and AdaBoost. We use all default parameters of *sklearn* library, as the goal is to find most suitable model first and then further tune. We use the Gini index as our purity criterion for making the binary splits in our decision trees. With Random Forest, we create 100 independent bootstrapped copies of the training data, fit a separate decision tree for each copy, and, for each split in a

Algorithm	Accuracy	Precision	Recall
Logistic Regression	0.747 / 0.748	0.747 / 0.748	0.747 / 0.748
Decision Tree	0.558 / 0.577	0.558 / 0.579	0.557 / 0.576
AdaBoost	0.654 / 0.652	0.644 / 0.643	0.654 / 0.651
Random Forest	0.655 / 0.693	0.656 / 0.696	0.655 / 0.694

Table 3: Model performances on all / selected features.

tree using again the Gini loss, we every time only consider $\sqrt{p}$ of the p features. With the AdaBoost classifier, we begin by fitting a decision stump, and then fit 50 additional copies on the same dataset but putting every time additional weight on incorrectly classified observations and using a learning rate of 1.

5.4 Feature Selection through Stepwise Regression

We use Forward Stepwise Linear Regression with p-value criteria set to 0.01 to select the most important variables that predict the original ranking (instead of the quintiles). The process involves incrementally adding new variables to the model and testing the statistical significance until stopping condition is reached. We have two goals to adapt Stepwise Regression. First, we aim to select the most meaningful features from all features. Second, although we understand this linear method can not be used for a classification task, we want to confirm if the selected features from this interpretable method can improve the performance of our classification models.

5.5 Model Comparison with Selected Features and Evaluation on Test Set

We re-fit the four models using only the selected features found in previous step and use the most accurate model to predict the ranking labels for the held-out test set.

6 Results

When including all variables, Table 3 shows that the Multinomial Logistic Regression model outperforms the other tree-based methods in accuracy, precision and recall. The confusion matrix in Figure 4 more precisely illustrates the model has relatively high recall and precision for the two extreme ranking categories (labels 1 and 5) but has limitations to differentiate students in middle ranking.

We also observe that Random Forest and AdaBoost perform better than simple Decision Tree. This is expected because both are ensemble learning methods that are usually less prone to overfitting. On the one hand, Random Forest aims at decorrelating the trees using bootstrapped samples, thereby making the average of the resulting trees having a lower variance than a simple decision tree. On the other hand, in AdaBoost, instead of fitting a single large decision tree, which corresponds to fitting the data hard and potentially overfitting, the model learns slowly.

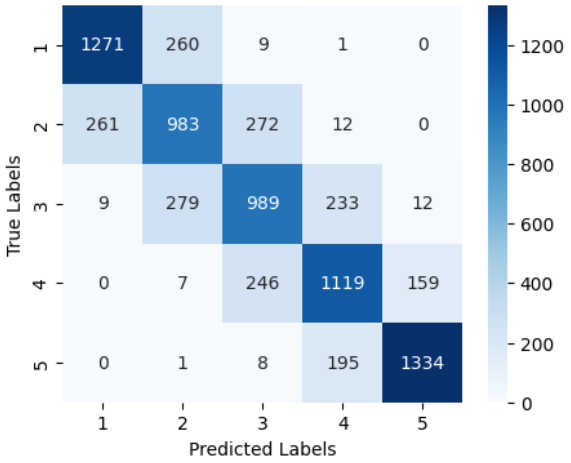

Figure 4: Confusion matrix for Multinomial Logistic Regression with cross-validation during training.

A Forward Stepwise Linear Regression is then used and selects 24 features. Besides variables we have hypothesized and confirmed to be predictive in the data exploration section, some interesting other features such as whether the student repeated a grade in high school, whether the student comes from a public or private high school, and whether the student receives financial aid appear important as well.

Results in Table 3 show that, with feature selection, Decision Tree and Random Forest slightly improve their performance, while AdaBoost and Multinomial Logistic Regression stay the same. These findings prove the usefulness of feature selection for certain models to better generalize to new data. Even if accuracy is not improved, feature selection still helps to make the new model more interpretable.

We use Multinomial Logistic Regression with the 24 selected features as our final model. We observe that using only the selected features greatly reduces the convergence time, probably because we eliminated variables that were highly correlated together. Figure 5 shows, on the left, the sign (green for positive, red for negative) of the coefficient associated to each selected feature for each ranking

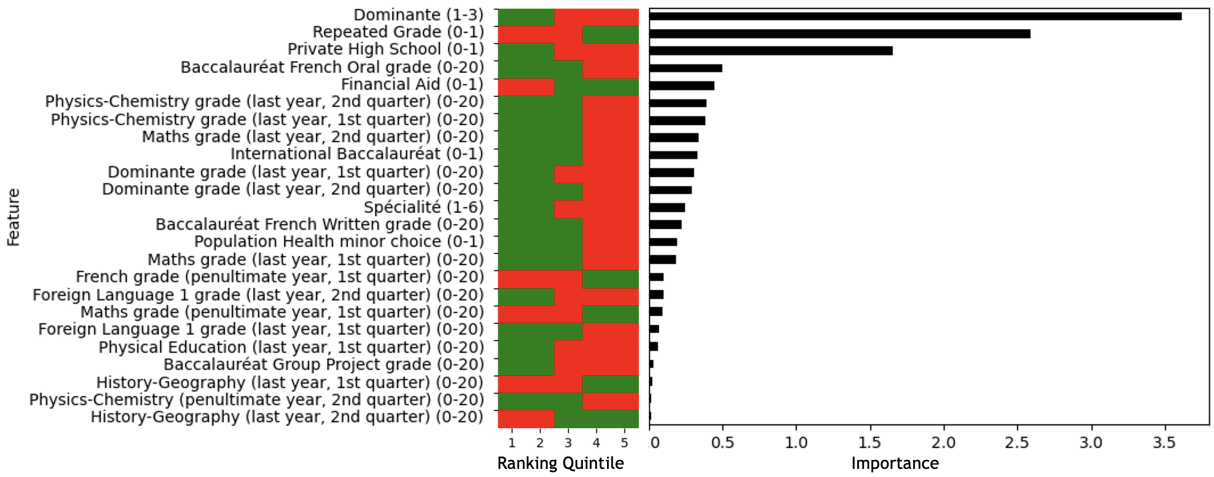

Figure 5: Sign of coefficients by ranking quintile (left, green for positive and red for negative) and feature importance (right) in final Multinomial Logistic Regression model. In parentheses are the range of values for each feature.

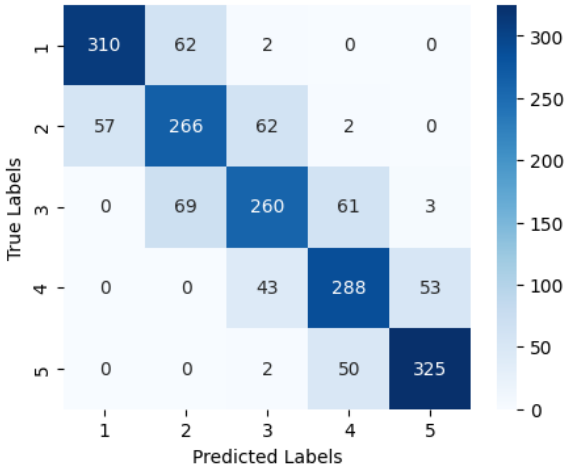

Figure 6: Confusion matrix for Multinomial Logistic Regression using selected features on held-out data.

quintile, and, on the right, the importance of each feature in the model, defined by the absolute value of the average of the coefficients for each class. Note, for interpretation, that we can only compare together features that are on the same scale.

Figure 6 presents the final results on the held-out dataset that the model can confidently predict the admission ranking for the two extreme labels and can at least predict the middle-level rankings correctly within one lower or upper label difference.

7 Conclusion and Future Work

Our results offer several explanations for social inequalities in admission to this program. Based on Figure 5, controlling for other variables, being on financial aid, coming from a public high school, not doing an International *Baccalauréat* (which is often only available in privileged high schools), or having repeated a grade (which is more common among lower-class students) has a negative impact on being ranked in the first two ranking quintiles. Because *Dominante* and *Spécialité* have been selected as important features, this means also that early tracking in high school (forcing students to choose a major and a minor while in high school) influences ranking. But making these important choices early on is socially biased because disadvantaged students have a less informed personal network and have less access to professional opportunities that could help them decide (Marbach and van Zanten, 2023). In terms of the disciplines valued, when we compare feature importances, we see that Maths grades have a high importance, even though Maths is not even a subject in the "common pre-health studies" program, along with Physics-Chemistry grades. On the other end of the spectrum, having a good grade in History-Geography or French class surprisingly seem to make a student less likely to be ranked in the first two quintiles. This reveals a preference for "hard sciences" by the admissions staff, while overlooking the potential contributions of students proficient in "social sciences" to the development of the "human qualities" that health professionals are perceived to lack.

As our models show, they do not fully explain the ranking done by the admissions staff. For future work, we could explore deep learning methods given we have a huge dataset, and include textual elements that were required in applications, in particular personal statements, as previous research has found their content and style significantly vary by student background (Gebre-Medhin et al., 2022).

Contributions

Anya worked on fitting the models and limitations. Léon worked on the exploratory data analysis and model interpretations. Both authors contributed to the writing.

References

Ben Gebre-Medhin, Sonia Giebel, AJ Alvero, Benjamin W. Domingue, and Mitchell Stevens. 2022. Application Essays and the Ritual Production of Merit in US Selective Admissions. *Poetics*, 94:101706.

Hansol Lee, René F. Kizilcec, and Thorsten Joachims. 2023. Evaluating a Learned Admission-Prediction Model as a Replacement for Standardized Tests in College Admissions. *L@S '23: Proceedings of the Tenth ACM Conference on Learning @ Scale*, pages 195–203.

Thomas Lux, Randall Pittman, Maya Shende, and Anil Shende. 2016. Applications of Supervised Learning Techniques on Undergraduate Admissions Data. *Proceedings of the ACM International Conference on Computing Frontiers*, pages 412–417.

Léon Marbach and Agnès van Zanten. 2023. With a little help from my family and friends: social class and contextual variations in the role of personal networks in students' higher education plans. *British Journal of Sociology of Education*.

Barbara Martinez Neda and Sergio Gago-Masague. 2022. Feasibility of Machine Learning Support for Holistic Review of Undergraduate Applications. *International Conference on Applied Artificial Intelligence (ICAPAI)*, pages 1–6.

Julie Posselt. 2016. *Inside Graduate Admissions*. Harvard University Press, Cambridge, MA.

Christiaan A. Rees and Hilary F. Ryder. 2022. Machine Learning for the Prediction of Ranked Applicants and Matriculants to an Internal Medicine Residency Program. *Teaching and Learning in Medicine*, pages 1–10.

Mitchell Stevens. 2009. *Creating a Class*. Harvard University Press, Cambridge, MA.

Agnès van Zanten, Alice Olivier, Christophe Birolini, Audrey Chamboredon, and Léon Marbach. 2021. The Reform of Access to Health Studies in France. *Cogito*.