Wanjing Anya Ma
wanjingm@stanford.edu

Léon Marbach
lmarbach@stanford.edu

Using Machine Learning to Predict Admissions to Higher Education

Stanford
GRADUATE SCHOOL OF
EDUCATION

Motivations

Problem: Different social groups have an unequal access to higher education [1].

Inputs: Information submitted by applicants to a selective pre-health studies program in 2020 in France.

Outputs: Ranking quintiles for admission established by the university's admission staff.

Finding: Multinomial logistic regression on selected features predicts with 75% accuracy the ranking quintile of a student.

Dataset

60 features on **9,575** students who were enrolled in the scientific track in high school at time of applying:

- Sociodemographic information (sex, age, financial aid status, country of birth)
- Academic background (high school public/private status, major and minor chosen in high school)
- Subject grades in each quarter of the last two years of high school and *Baccalauréat* tests grades
- Evaluation by high school teaching team on 5 items

Methodology

Model Comparison on all Features → **Feature Selection** → **Evaluation on Test Set**

Using 5-fold cross-validation on 80% of the data, compare:

- Multinomial logistic regression using softmax

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}$$

- Simple decision tree, random forest, adaBoost

Forward stepwise linear regression using individual rankings selects 24 features.

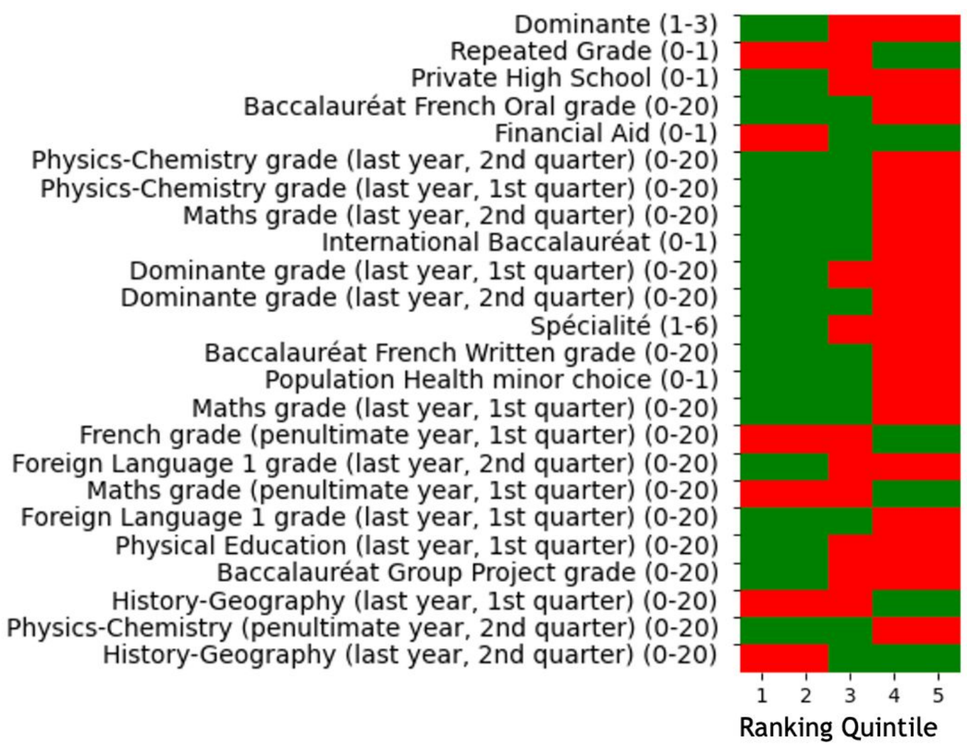
We re-fit our four models on selected features and evaluate most accurate model on the 20% held-out test set.

Results

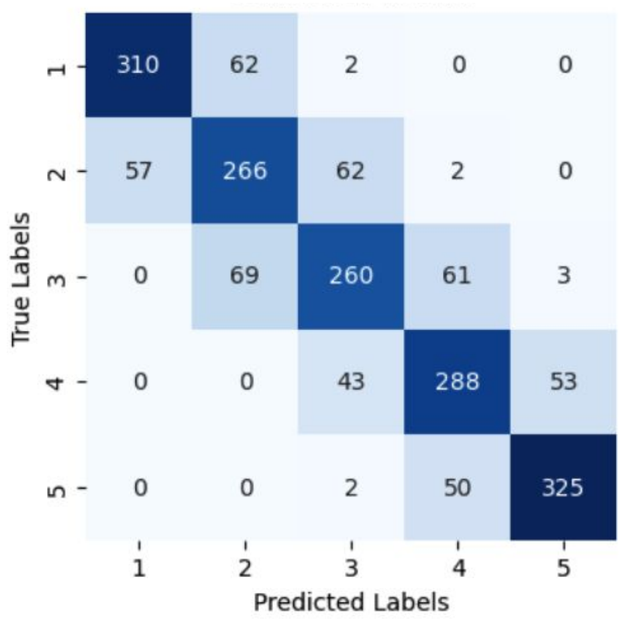
Accuracy performances comparison on all 60 / selected 24 features ↓

Model	Accuracy
Logistic regression	0.747 / 0.748
Decision tree	0.558 / 0.577
Random Forest	0.654 / 0.652
AdaBoost	0.655 / 0.693

Sign of coefficients by ranking quintile (green for positive, red for negative) in final logistic regression model →



Confusion matrix on test set of final logistic regression model ↓



Conclusion

Results reveal **social preferences** for students:

- From private high school and not on financial aid
- Who did not repeat a grade
- Who are good at maths and physics-chemistry, and not so much in social sciences
- Who chose certain majors/minors in high school, thereby highlighting effects of early tracking policies

Future Work

Explore **deep learning methods** and include **textual elements** that were required in applications, in particular personal statements [2].

References

[1] 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.*

[2] Ben Gebre-Medhin, et al.. 2022. Application Essays and the Ritual Production of Merit in US Selective Admissions. *Poetics*, 94:101706.