Credit Default Prediction with Fairness in Mind:

A statistical and ethical exploration

Ibrahim AGOUNA ESAÏE MASSI, Jean-Eude GBADA, Grégoire MERCENIER

May 2025

Contents

List of Figures

1. Descriptive Statistics & Machine Learning Modeling

1.1. Introduction - Descriptive Statistics

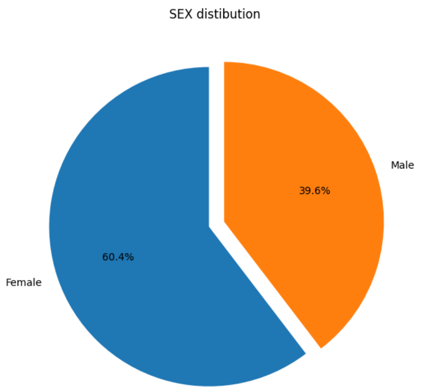
In today's algorithmically mediated financial landscape, credit scoring has become both a vital tool and a matter of ethical scrutiny. This project explores the development of a predictive model aimed at estimating the likelihood that a credit card holder will default on their payment in the following month. The dataset at the heart of our work comprises 30,000 clients from a Taiwanese financial institution, which captures a rich spectrum of demographic and financial behaviors.

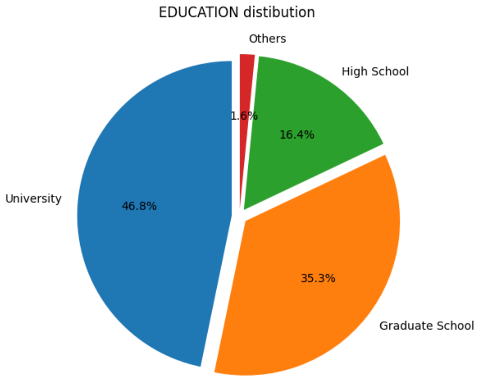
But predictive power alone is no longer sufficient. The central question guiding our approach is not just *how well* the model performs, but *for whom*. How can we ensure that the model does not simply reproduce existing inequalities under the guise of objectivity? In this context, we investigate not only classical machine learning approaches, but also their ethical ramifications—placing particular emphasis on education as a variable laden with socioeconomic meaning. Our objective is twofold: to build a model that is both accurate and just.

The credit card clients dataset (Yeh & Lien, 2009) investigates default payments among clients in Taiwan as of October 2005. The goal is to predict whether a customer will default on their payment in the upcoming month. This dataset has been widely used in multiple studies for default payment prediction.

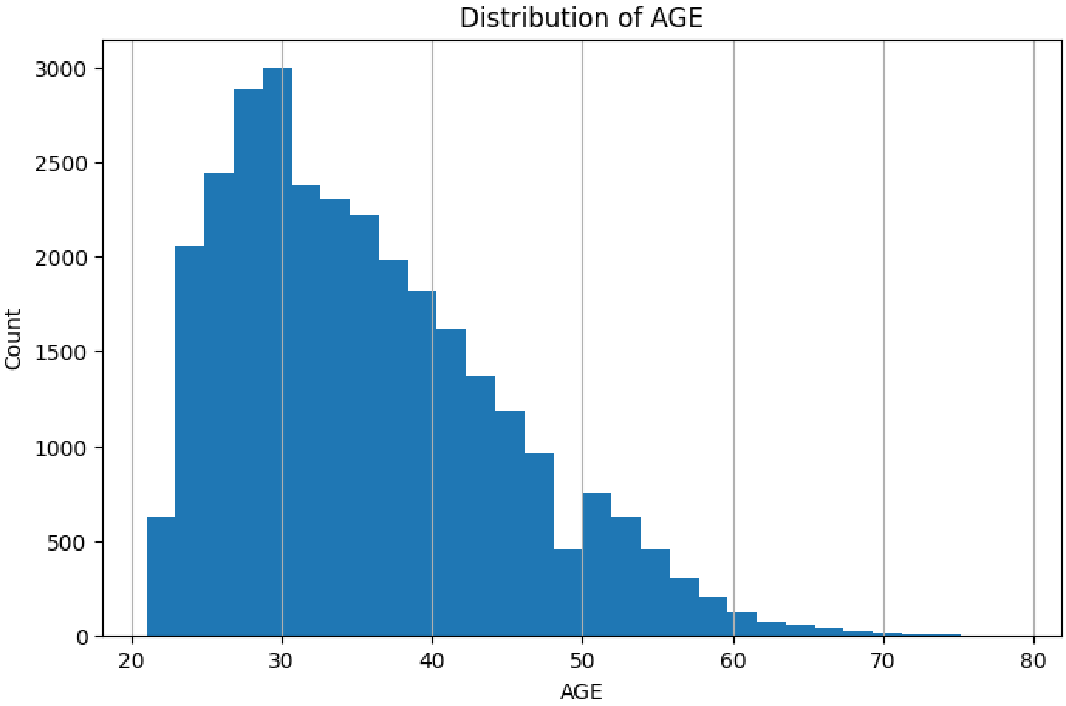
The sample includes 30,000 credit card clients, with a notable gender imbalance: women represent 60.37% of the sample compared to 39.63% men. This skew may influence the performance of the predictive model, particularly if repayment behaviors differ across gender lines—a point of caution to avoid algorithmic discrimination. Indeed, prior studies suggest that women may exhibit slightly better repayment performance, which must be factored into the fairness assessment.

In terms of education, the majority of clients hold a university degree: 46.77% are university graduates, while 35.28% possess advanced degrees (Master's or PhD). In contrast, only 16.39% have completed high school. This profile points to a socioeconomically privileged clientele, potentially correlating with better credit management. However, careful treatment of the education variable is needed to prevent bias against less educated borrowers.





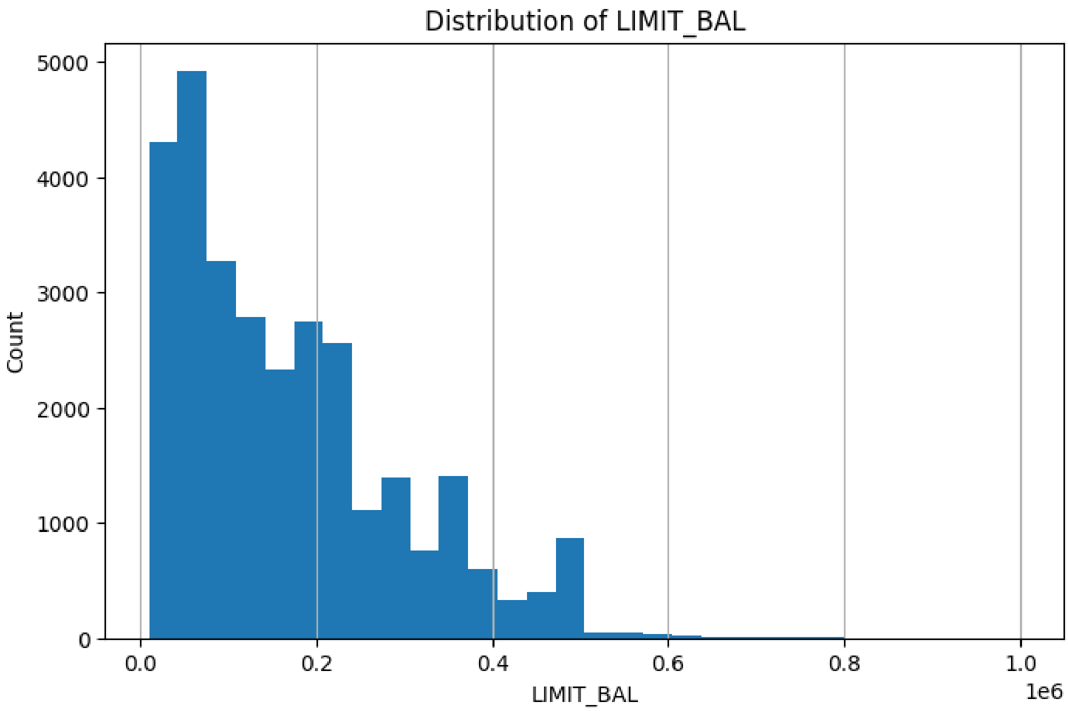
Marital status also reveals a slight dominance of single clients (53.21%) over married ones (45.53%). The average age is 35.5 years, with a strong concentration around early adulthood. Given that age is a sensitive attribute in fairness contexts, we must ensure that the model does not unjustly penalize specific age groups, especially if default rates vary across generations.



On the financial side, customer profiles vary widely. The average credit limit is NT[1]$167,484, ranging from NT$10,000 to NT$1,000,000, highlighting considerable disparity between small and large borrowers. Monthly billed amounts (BILL_AMT1 to BILL_AMT6) average between NT$40,000 and NT$50,000, while average payments (PAY_AMT1 to PAY_AMT6) are significantly lower—around NT$4,799 to NT$5,921. This imbalance suggests a common practice of partial repayments, typical of revolving credit usage, where debt accumulates over time.

Repayment behavior, captured by PAY_0 to PAY_6, shows mostly favorable histories: the majority of entries denote on-time payments or short delays (0 to 2 months). However, extreme cases of 7 to 8 months' delay are present and serve as strong indicators of default risk. These repayment history variables are expected to be central to our modeling efforts, as they reflect recent financial discipline—a key predictor of future behavior.

[1] New Taïwan Dollar



1.2. Machine Learning

1.2.1 Features importance

We began our modeling journey with a Random Forest classifier, selected for its capacity to handle non-linear relationships, manage feature interactions, and provide interpretable outputs. To ensure robustness and reduce variance, we used a stratified 50-fold cross-validation scheme—a choice that strengthened the statistical credibility of our results.

The analysis of feature importance confirms the primacy of repayment history, with PAY_0 and PAY_2 ranking highest. Financial variables such as LIMIT_BAL (amout of credit granted), the invoiced amounts (BILL_AMT1 to BILL_AMT6), and the amounts reimbursed (PAY_AMT1 to PAY_AMT6) also play a major role, emphasizing the relevance of credit utilization and repayment behavior in risk prediction. Socio-demographic variables—AGE, EDUCATION, SEX, MARRIAGE—have more moderate but still non-negligible impact. Given the reasonable number of variables and their overall contribution, we decide to retain all available variables for the final modeling in order to maximize the predictive capacity of the model without risking information loss.

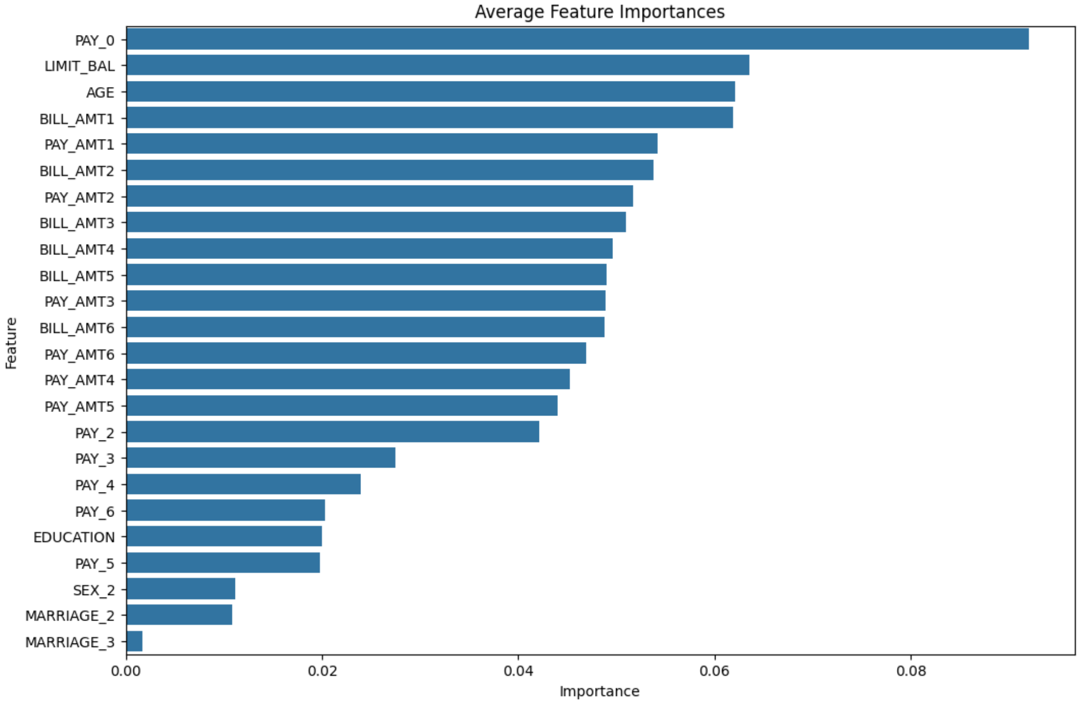


Figure 1: Average feature importance and standard deviation for all variables in the Random Forest model. Behavioral financial indicators dominate over socio-demographic features.

We chose to retain all variables[2] for final modeling to preserve predictive power while monitoring their ethical implications. The Random Forest model reached a mean accuracy of 81.4%, with low variance. However, recall was just 34.47%—a limitation given the cost of missing high-risk cases.

[2] Incorporating all available variables during the exploratory phase exemplifies a core principle of fairness-aware machine learning: prior to excluding potentially sensitive features, it is critical to quantify their true influence on both predictive efficacy and bias propagation.

1.2.2 Choice of model (KNN, XGBOOST, Random Forest)

The comparative results show that the **K-Nearest Neighbors (KNN)** model and the **Random Forest** model achieve high overall accuracies, around **79.36%** and **81.40%** respectively. However, despite this good performance in terms of *Accuracy*, their *Recall* rates remain relatively modest, at **36.78%** for KNN and **34.47%** for Random Forest.

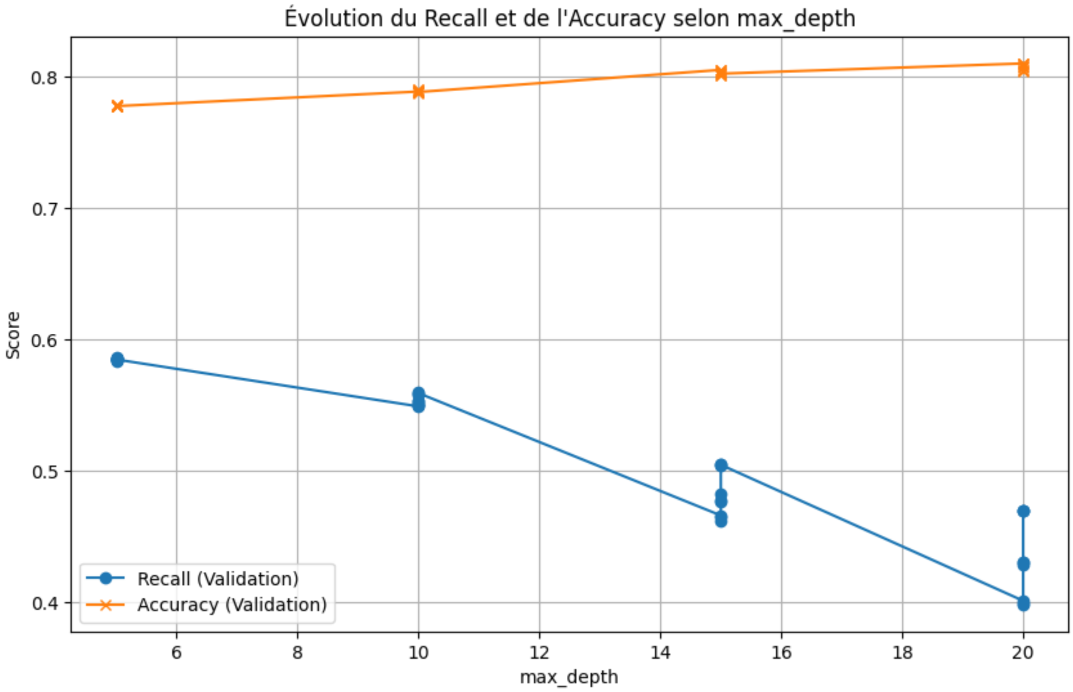
The **XGBoost** model, on the other hand, displays a significantly higher *Recall* of **57.28%**, but at the cost of a lower overall accuracy (**76.15%**).

In a logic of balancing overall robustness and operational stability, we decide to retain the **Random Forest** model for the continuation of the project, due to its good learning capacity and interpretability. Nevertheless, in order to significantly improve the detection of bad payers, which is of utmost priority in this credit scoring context, we will now focus our efforts on the specific **optimization of the Recall** of the Random Forest.

The findings illustrate the classical trade-off between model complexity, sensitivity to high-risk instances, and interpretability. In high-stakes domains such as credit scoring, prioritizing models with higher recall—even at the expense of overall accuracy—can be justified, as the cost of false negatives (missed defaults) often outweighs that of false positives. Thus, optimal model selection hinges not solely on standard metrics, but on domain-specific risk appetites and operational constraints.

1.2.3 Optimization of the model

In this section, we used a cross-validation optimization method with GridSearchCV to tune the hyperparameters of the Random Forest model. The main objective was to maximize recall, which is the model's ability to accurately detect defects. A search grid was defined with different values for the number of trees, maximum depth, and minimum node split threshold. At the end of the search, the best configuration identified was max_depth = 5, min_samples_split = 10, and n_estimators = 100, with an optimal recall of 58.62%. The comparative graph also shows that this configuration maintains relatively stable accuracy while improving defect detection.



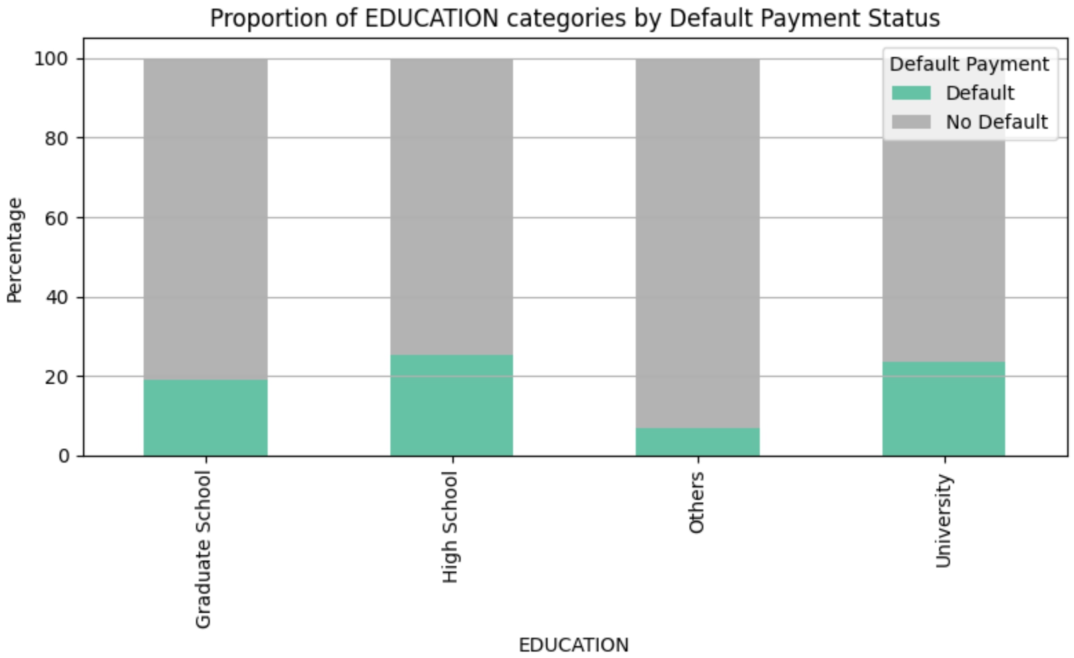
1.3. Saving pipeline: preprocessing + model

We have finally saved our complete pipeline under the name rf_pipeline_model.pkl, including data preprocessing and the Random Forest model with the best hyperparameters: n_estimators=100, max_depth=5, min_samples_split=10, and class_weight='balanced', in order to reuse it later in the fairness analysis.

2. Fairness Considerations

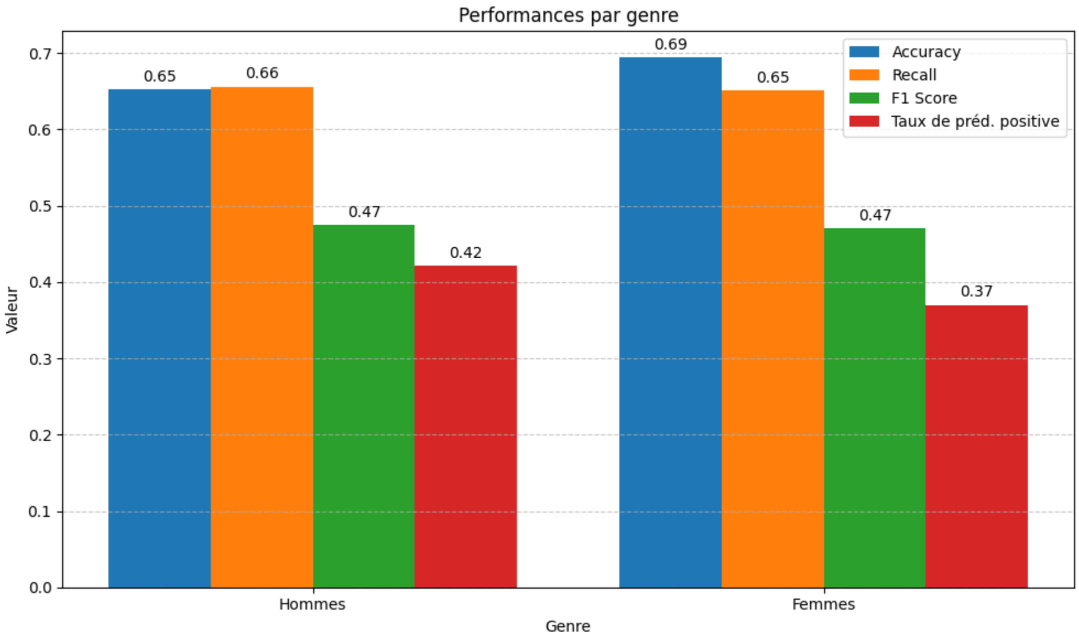
In order to deepen the analysis of algorithmic fairness, our focus will not be limited solely to traditionally protected variables such as **gender**. In line with the observations made in our dataset, and in accordance with the recommendations of *Hurlin et al. (2024)*, we will also include **education level** as a key protected attribute.

This choice is based on the evidence of significant discrepancies in default rates based on degree levels, and allows us to explore how the use of seemingly neutral variables can, in reality, reproduce or reinforce existing social inequalities.



We retrieve our Random Forest model and then perform parity tests on :

2.1. Gender

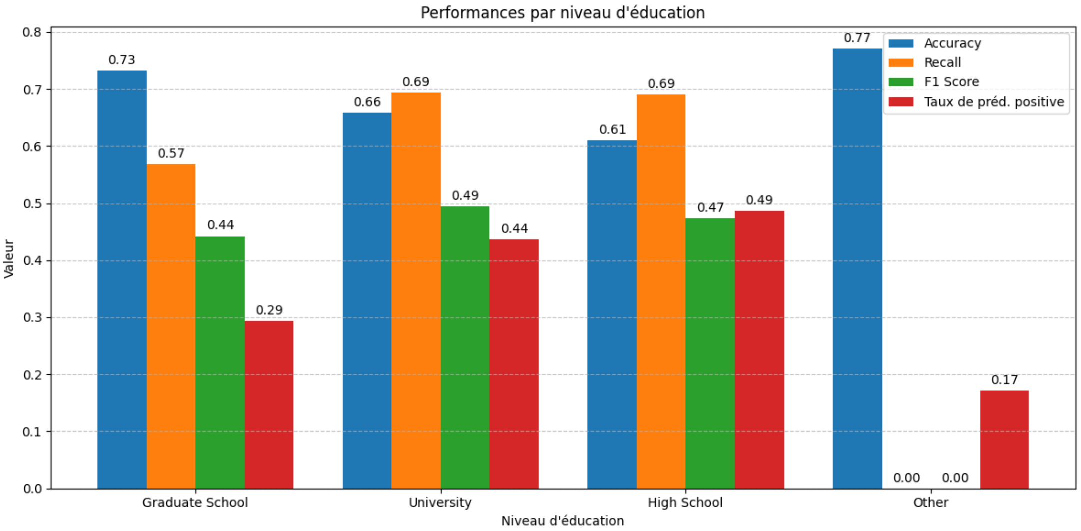


2.2. Education

The results reveal significant performance disparities of the model based on **education level**. Individuals with higher education (*Graduate School*) exhibit the best overall accuracy, with an accuracy of **73.2%**, but a relatively low defect detection rate (*Recall*) of **56.8%** and a positive prediction rate of only **29.3%**.

In contrast, clients with university or secondary education (*University* and *High School*) achieve much higher recall scores (**69.4%** and **69.0%** respectively), but at the cost of lower overall accuracy (**65.8%** and **61.0%** respectively). As a result, these two groups are assigned defect predictions more frequently: **43.6%** for university graduates and **48.6%** for high school graduates.

This imbalance is reinforced by the **Statistical Parity Difference**: the defect prediction rate among less educated individuals is significantly higher than that observed among higher education graduates.



2.2.1 Pipeline 2 with Education Included

To more rigorously evaluate the fairness of our model, we built a second pipeline explicitly incorporating the **EDUCATION** variable as a sensitive attribute. Additionally, we used the *stratify* option when splitting the dataset into training and test sets, in order to ensure a proportional distribution of default classes in each subset. This stratification helps prevent bias related to over- or under-representation of certain classes in the test sample, thereby improving the reliability of fairness metrics such as the *Statistical Parity Difference (SPD)*.

Thanks to these adjustments, the analysis by education level gains in precision and allows for a better identification of potential disparities in treatment across groups.

The results show a maximum **SPD of 0.221** between groups, indicating a significant disparity. The education levels *"High School"* and *"Other"* remain the most disadvantaged in terms of positive prediction rates, confirming the existence of an imbalance that mitigation strategies can aim to address.

3. Mitigation

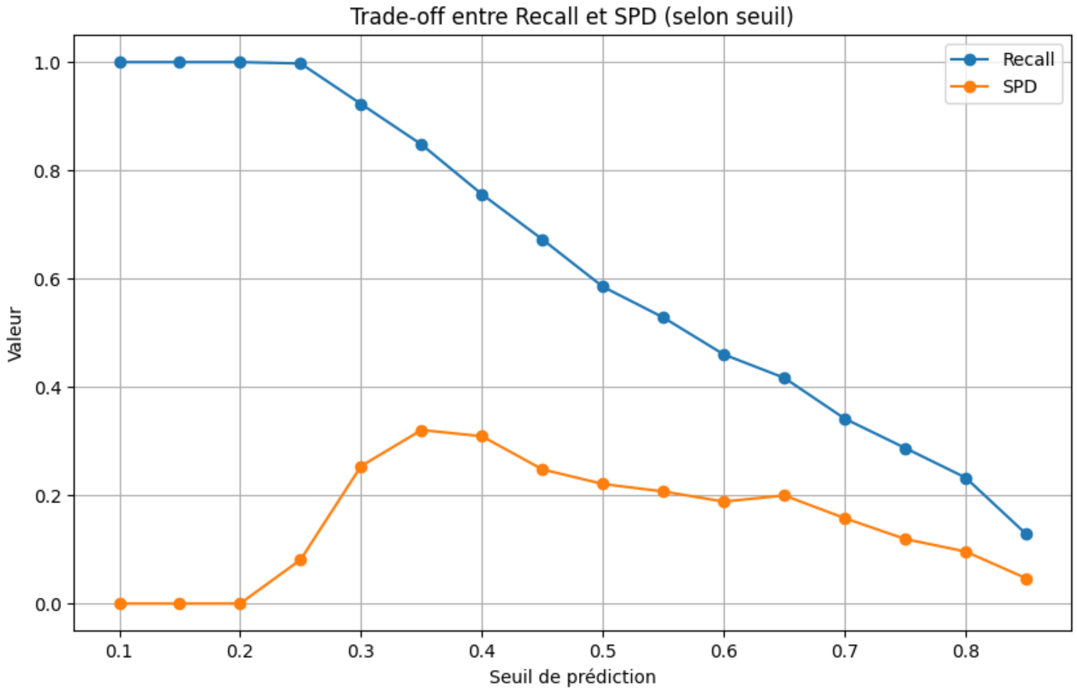
3.1. Attempt 1: ThresholdOptimizer

In this final step, we applied a post-processing method using the **ThresholdOptimizer**, based on the constraint of *Demographic Parity*, to mitigate the biases observed in model performance by education level. Following this mitigation, we re-evaluated the main metrics (accuracy, recall, selection rate) by group and visualized them using bar plots.

The results show a notable reduction in selection rate gaps between groups (ranging from **8.7% to 14.3%**), indicating an improvement in fairness. However, this correction came with a decrease in overall recall (e.g., **0.33** for Graduate School), highlighting a classic trade-off between performance and fairness. This outcome underscores the value of mitigation approaches while also emphasizing the need for careful balancing based on the model's objectives.

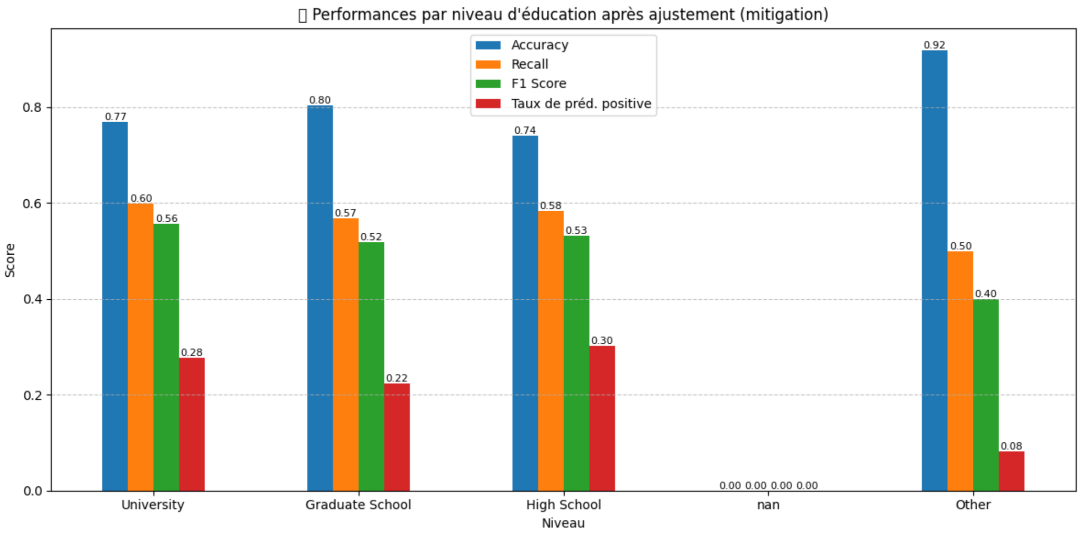
3.2. Attempt 2: Threshold Adjustment

This result illustrates that there is no single optimal threshold: increasing the detection of defaulters (*recall*) comes at the expense of fairness between groups (*SPD*). It is therefore essential to select a decision threshold according to the project's priorities (minimizing defaults vs. ensuring fairness). A threshold around **0.5** appears to offer a reasonable compromise between performance and fairness.



In this approach, we manually adjusted the decision threshold to **0.5** in order to optimize the trade-off between predictive performance—particularly *recall*—and fairness, as measured by the *Statistical Parity Difference (SPD).* After applying this threshold to the predicted probabilities, we evaluated the model's performance both overall and by education level.

In conclusion, the model achieves an overall **accuracy of 77.8%**, a **recall of 58.5%**, and an **F1-score of 53.8%**, with relatively balanced performance across groups. The maximum **SPD is reduced to 7.9 percentage points**, indicating a reasonable improvement in parity while maintaining a satisfactory ability to detect defaults.



4. Critical assessment and future perspectives

Our model is grounded in rigorous methodology, with robust cross-validation and strong baseline accuracy. Nevertheless, challenges remain—chief among them the modest recall rate. Improving this metric is imperative for effective default detection.

Our fairness approach, focused on removing education as a bias-laden variable, was a meaningful step, but we did not yet quantify fairness using metrics such as demographic parity or equalized odds. Future iterations must incorporate these to transform ethical principles into measurable outcomes.

We also acknowledge the temporal limitations[3] of static modeling. Credit behavior evolves, and incorporating time-aware methods (e.g., sequential models or survival analysis) could yield deeper insights. Additionally, the dataset's specificity to Taiwan limits generalizability—regional adaptation will be essential for broader deployment.

To bridge accuracy and transparency, future work will explore ensemble methods, fairness-aware algorithms, and explainability tools like SHAP. In credit modeling, the goal is not to replace human judgment but to augment it—ethically and responsibly.

[3] It is also critical to acknowledge the regional specificity of the dataset. While the findings hold within the Taiwanese context, broader applicability requires validation across more heterogeneous populations. This need for greater geographical and cultural diversity in fairness datasets aligns with recent calls in the literature (Le Quy et al., 2022)

5. Conclusion

At the crossroads of finance, ethics, and machine learning, credit default prediction represents both a technical challenge and a moral imperative. In developing our model, we aimed to reconcile accuracy with fairness, using Random Forest as a pragmatic choice for its balanced performance. More importantly, we adopted an ethical lens throughout—culminating in the removal of education as a proxy for structural bias.

But the stakes extend far beyond our dataset. As the World Bank reports, 31% of adults worldwide still lack access to formal financial systems. Models such as ours, if crafted with care, could support broader financial inclusion. Yet this promise can only be fulfilled if we build algorithms that are explainable, accountable, and subject to scrutiny—not just performant, but trustworthy.

In the words of George Box, *“All models are wrong, but some are useful.”* Our ambition was not perfection, but usefulness—with a conscience. The model we built is a first step, not a final answer. It invites iteration, critique, and collaboration. Because in the end, predictive systems should serve people—not define them.

References

[1] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

[2] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

[3] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository: German Credit Data. University of California, Irvine, School of Information and Computer Sciences.

[4] Hurlin, C., Pérignon, C., & Sobotka, F. (2024). Fairness in Credit Scoring: Revisiting the Use of Socio-Demographic Variables. *Journal of Banking & Finance*, *157*, 106852.

[5] Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33.

[6] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (pp. 3315–3323).

[7] European Parliament and Council. (2024). Regulation (EU) 2024/XXXX on Artificial Intelligence (AI Act).