THE FAIRNESS AND EXPLAINABILITY OF CREDIT CARD DEFAULTS IN TAIWAN

GROUP 7

Abdelrahaman Shehata, Ektoras Manouselis, Sonia Horváthová

THIS REPORT IS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE METHODS FOR RESPONSIBLE AI COURSE OF BACHELOR PROGRAM ON COGNITIVE SCIENCE AND ARTIFICIAL INTELLIGENCE AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES OF TILBURG UNIVERSITY

INSTRUCTORS
dr. Seyed Mostafa Kia
dr. Juan Sebastien Olier

LOCATION
Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands

DATE
November 29th, 2023

WORD COUNT
1795

The fairness and explainability of credit card defaults in Taiwan

Group 7

Abdelrahaman Shehata, Ektoras Manouselis, Sonia Horváthová

Introduction

In our exploration of the fairness and explainability of credit card defaults in Taiwan, the dataset "Default of Credit Card Clients" contains a compilation of Taiwanese customers' payment behaviour. With 23 explanatory variables and a binary response variable indicating credit card payment defaults or successful loan payment, our approach involves nested cross-validation, robust cross-validation, and hyperparameter tuning using Grid_search_cv and random_search_cv on a random forest classifier. Crucially, our focus extends beyond model performance to emphasise two key facets: explainability and fairness. Leveraging the random forest classifier for its feature importance enhances model interpretability, while fairness considerations, employing statistical parity and equalised odds metrics give us an equitable predictive model.

Dataset

The dataset used for this project is “Default of Credit Card Clients”, obtained from the UCI Machine Learning Repository(Yeh, 2016). The dataset captures detailed information about Taiwanese customers' default payments, including 30000 instances and 25 features that include demographic and financial information about the customers. Furthermore, it contains 24 variables, one response variable and 23 explanatory variables. The response variable, Y, represents the default payment of clients (1 = Yes, 2 = No). The explanatory variables consist of X1, amount of clients and client's families credit in New Taiwan dollar; X2, gender (1 = male; 2 = female); X3, education (1 = graduate school; 2 = university; 3 = high school; 4 = others); X4, marital status (1 = married; 2 = single; 3 = others); and X5 representing age in years. Further, X6-X11 are the repayment statuses for each month in 2005, and they are in reverse chronological order having X6 represent September and X7 is August and so on .
X12-X17 stand for the amounts of the bill statement, so the amount of money the credit card owner has to pay the bank for that month corresponds with the repayment status from X6-X11. The order of the months follows the same reverse chronological order as for X6-X11. Lastly, X18-X23 are amounts that the credit card owner has paid back to the bank for that month, it also corresponds with the previous variables.

Methods:

1-Data preprocessing:

The data preprocessing done was removing the first column which only contained ids, as well as removing the first row as it was the name of each column in a string type which would affect the training and test. According to the OpenML dataset description, values 3 and 4 in both X3 and X4 were originally intended to represent "others". However, the dataset contained additional values, including 4, 5, and 6. To prevent confusion and simplify the input for the models, we combined these values under the "others" category in their respective columns.

2-Preliminary data exploration:

Initially, an exploratory data analysis was done in order to discover this dataset's class balances, the distribution of values of its features and a correlation matrix. The three categorical features that will be our main focus going forward are gender, Education and marital status. The feature distribution showed in gender that the dataset has more instances of females than males, while education was distributed with university as the most occurring, followed by graduate school,then high school and the least was others. The distribution of marital status showed that single label was the highest closely followed by married and at a very low count was others .The class balance of the default of credit card decision revealed that there are 4 times as many samples of not defaulting compared to defaulted accounts at roughly 200 thousand and 50 thousand respectively. The correlation matrix did not show any significant relationship between the three features and the target variable.

2-Baseline fairness and explainability:

The two main aspects of responsible AI that we believe are worth exploring in this dataset are fairness and explainability; therefore, we conducted a baseline for both aspects. The fairness tests showed that according to the statistical parity and equalised odds metrics,the gender had low bias and was relativity fair. Meanwhile, education exhibited the lowest metrics, significantly underperforming with a Statistical Parity Difference of 0.12, Statistical Parity Ratio of 0.29, Equalized Odds Difference of 0.11, and Equalized Odds Ratio of 0.34. Similarly, marital status demonstrated suboptimal performance , with a Statistical Parity Difference of 0.12, Statistical Parity Ratio of 0.49, Equalized Odds Difference of 0.16, and Equalized Odds Ratio of 0.68.The baseline explainability results were achieved using LIME(Zafar & Khan, 2021) tabular explainer and feature importance which will be used at the end of the project to compare with our model after applying fairness methods to it.

4-Finding best performing model:

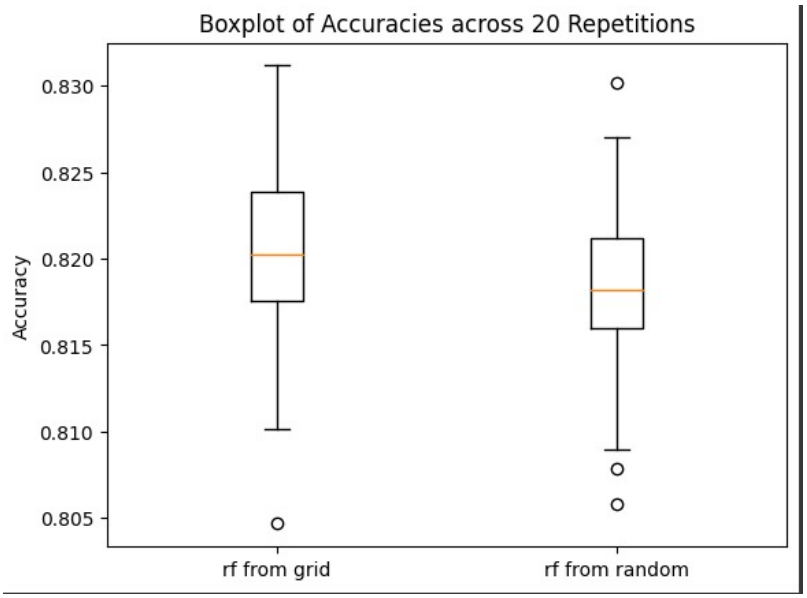

Figure 1. Conducting robust cross validation to find the best performing model

In the aim of finding the best performing model on our dataset,we employed nested cross validation (Refaeilzadeh et al., 2009) after which we did robust cross validation on the two best performing models. Grid_search_cv and random_search_cv

were used on a random forest classifier using three values to have one higher and one lower from the default for the following hyperparameters: n_estimators max_depth, min_samples_split, and min_samples_leaf. One of the main reasons we decided to use random forest was that it didn't require us to perform one hot encoding on the data, it can be given as it is. Another reason is that random forest contains feature importance which aids in the model's explainability, understanding its decision-making by revealing the contribution of each feature. The best performing models was the grid search model, with parameters {'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 150}, achieving a 82.33% accuracy on the test set. Surprisingly, the randomised search identifies a strong performer with parameters {'n_estimators': 50, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_depth': 40}, achieving the same accuracy of 82%. The solution to find the better model was to perform robust cross validation as you can see above in Figure 1. that the grid search hyperparameters model was better over 20 iterations.

Results:

1-Fairness results:

To ensure fairness in the dataset, fairness mitigation strategies were tested on the dataset. To determine the optimal fairness method suited for each of the demographic features, a variety of mitigation strategies were used, such as data resampling, bias removing, data reweighting, exponentiated gradient, and lastly threshold adjustment. After a thorough analysis of gender equity mitigation approaches, we found that threshold optimizer and reweighting were the most effective approaches.The improvement from the baseline gender fairness was quite small since it was already fair relatively speaking. However, for the education-based fairness mitigation approaches, the use of threshold optimizer and exponentiated gradient(Majidi, 2021) were the most effective. The improvement for education was the most significant for the exponentiated gradient as the Statistical Parity Ratio increased by approximately 159.48% from the baseline value of 0.29 to 0.7525 , as well as,the Statistical Parity Difference decreased by approximately 73.25% from the baseline value of 0.12 to 0.0321. Finally, for the marital status it was revealed that exponentiated gradient and correlation elimination approaches were the most suitable. The correlation-based model shows a significant improvement over the baseline, with a 176.56% higher Statistical Parity Ratio from 0.29 to 0.8011 and a 76.69% lower Statistical Parity Difference from 0.12 to 0.027975, indicating improved fairness in achieving a more balanced group representation and reduced disparate impact.

2-Explainability results:

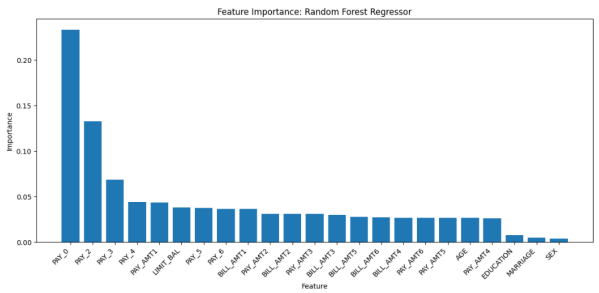

Figure 2. Baseline feature importance

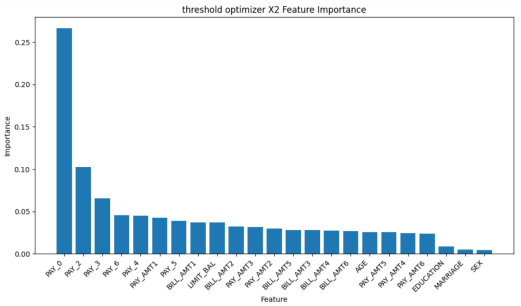

Figure 3. Feature importance of gender using threshold optimizer method

The aim of the following tests done using feature importance(Saarela & Jauhiainen, 2021) was to see if any observable changes occurred to the models after applying each fairness method to each of the three features. As shown in Figure 3., the feature importance shows that for the gender feature importance, there is an emphasis on pay_0, the payment status of the nearest month. This would make sense that the model would rely on the last payment as an important feature as it may be highly correlated with previous ones and it simply makes there no need for the model to consider payments statuses further back in time. The same feature importances were found for all features with the applied fairness methods. When comparing the baseline feature importance plot in Figure 2. To the plot that represents all the post fairness technique models in Figure 3. it can be seen as almost identical.

Discussion :

The implication of the results found was a great improvement to the dataset's fairness metrics which benefits many credit card users who are significantly affected by the use of this biassed and unfair dataset to decide whether their credit cards would get defaulted or not. This especially would affect people of lower education and people with a non-preferred marital status according to the model's point of view which is very harmful to the lives of many. The use of the explainability methods allowed us to gain insight into the model's decision making process making it more interpretable for its general users. The finding of the correlation remover making the feature more equally important to the model is a valuable discovery that is worth exploring further as it may benefit many other models and datasets. The limitations and possible improvements that could be made to our project was the testing of only random forest classifiers which might have prevented us from finding a better performing model. Another addition for later research to use more iteration utilising more random states as to have our results more robust.

Conclusion:

In our study on credit card defaults in Taiwan, we analysed how well a model predicts whether customers' credit cards will get defaulted based on previous payments. We investigated the model's fairness and explainability. We considered it as our main feature that might contain biases: age, education, and marital status. We found that the model was not very fair in education and marital status. To mitigate these issues, we tried different ways to make the model fairer. The best model we found was a random forest with specific hyperparameters. Our work shows the possible improvements to the models and dataset's banking companies use contain harmful bias which can be resolved using certain fairness and explainability techniques.

References

Link to dataset: https://www.openml.org/search?type=data&status=active&id=42477&sort=runs
Link to github with code:
https://github.com/aaref123/methods_project_fairness_-_explainability_of_credit_card_defaults/tree/main

Majidi, N. (2021, April 3). Exponentiated gradient reweighting for robust training under label noise and beyond. arXiv.org. https://arxiv.org/abs/2104.0149

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In Springer eBooks (pp. 532–538). https://doi.org/10.1007/978-0-387-39940-9_565

Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. SN Applied Sciences, 3(2). https://doi.org/10.1007/s42452-021-04148-9

Vishnu, M. K., Vishal Rupak, V. R., Vedhapriyaa, S., Sangeetha, M., Manjuladevi, R., & Sagana, C. (2023). Recurrent gastric cancer prediction using randomized search CV optimizer. *2023 International Conference on Computer Communication and Informatics (ICCCI)*. https://doi.org/10.1109/iccci56745.2023.10128409

Yeh,I-Cheng. (2016). default of credit card clients. UCI Machine Learning Repository https://doi.org/10.24432/C55S3H.

Zafar, M. R., & Khan, N. (2021b). Deterministic local interpretable Model-Agnostic explanations for stable explainability. Machine Learning and Knowledge Extraction, 3(3), 525–541. https://doi.org/10.3390/make3030027