Gia Musselwhite
SML 312: Research Projects in Data Science
Professor Jon Hanke
December 6, 2024

Credit Where Credit is Due: Mitigating racial unfairness in credit risk data with the Fairlearn toolkit

1 Abstract

Are unfairness mitigation strategies, like threshold optimizer algorithms, able to reduce unfair outcomes in the field of credit scoring classification tasks? This paper builds on a 2024 project by Chunyu Yang, which found the white identity variable to be of high importance in mortgage applications approvals/rejections, and tests the Fairlearn unfairness mitigation toolkit for binary classification. In this paper, I first conduct exploratory data analysis, feature importance visualizations, and robust classification modelling on the 1996 Wooldridge 'loanapp' dataset. Afterwards, I attempt to mitigate unfairness due to racial identity in 'loanapp' decisions, following the Fairlearn Credit Loan Decisions documentation. I find that postprocessing techniques using threshold optimization *reduce disparities* between white and non-white applicants for both false positive and false negative credit approval rates. This successfully minimizes the equalizer odds difference and mitigates unfairness in lightgbm dataset classification.

2 Introduction

2.1 Motivation

Racial bias is a well-documented area of concern in automatic loan, mortgage, and credit approval systems. Credit disparities persist today despite interventions like the US Equal Credit Opportunity Act, which was first passed in 1974 to prohibit lenders from discriminating against applicants on the basis of protected characteristics like race, sex, and age (Skowronski, 2017). In an influential 1996 study, researchers at the Federal Reserve Bank of Boston found that "even

after controlling for financial, employment, and neighborhood characteristics, black and Hispanic mortgage applicants in the Boston metropolitan area [were] roughly 60 percent more likely to be turned down than whites" (Munnell et al., 1996).

Decades later, in a 2019 analysis of Home Mortgage Disclosure Act (HMDA) data, the US Consumer Financial Protection Bureau found that applicants of color were still 20-120% more likely to be denied than white applicants with the same credit score (Martinez & Kirchner, 2021). These disparities already have concrete impacts on the housing opportunities available to Americans who fall into protected classes (Martinez & Kirchner, 2021), and will only rise in importance as lended institutions use more advanced machine learning classification techniques to designate applicants as good or bad credit risks and assess mortgage approval (Munkhdalai et al., 2019). Predictions to reject loan applications from applicants who would not have defaulted on their loans withhold loan opportunities from applicants, while predictions to approve loan applications from applicants who actually would have defaulted can lead to secondary harms like long-term effects on FICO credit scores (Dudík, 2020). Thus, ML strategies that mitigate unfairness along sensitive features are *critical steps* of the credit risk benchmarking process that should be robustly evaluated and refined for their role in algorithmic assessment.

To correct for these potential biases, I will use the Fairlearn library toolkit. Fairlearn is an open source project belonging to Microsoft and initially launched in 2018 that offers fairness assessments and various algorithms with the potential to mitigate fairness-related harms (Weerts et al., 2023). To my knowledge, the Fairlearn toolkit has not been used to measure or mitigate fairness-related harms on the 1996 Wooldridge 'loanapp' dataset thus far. The 'loanapp' dataset is a prime candidate for the toolkit's mitigation packages, especially given the notably high feature importance rating of white identity on mortgage approval in the dataset.

2.2 Objectives

Outside of the research of its creators, the evolving Fairlearn toolkit has not been widely tested or evaluated in the literature for binary classification, especially on work as impactful as credit scoring (an exception to this is Małowiecki & Chomiak-Orsa's 2024 paper, which evaluates the Fairlearn parity constraints that are most effective at countering gender bias in binary classification). Thus, I set out to test the Fairlearn unfairness mitigation techniques myself. Yang

(2024) found that white identity was one of the most important predictors of mortgage decisions across multiple different classification models, so I use ‘WHITE’ identity as my sensitive bias-prone predictor variable of interest. Using threshold optimization postprocessing techniques, I hope to see a difference between pre- and post-unfairness mitigation metric frames (especially as they relate to false positive and false negative rates). Results that minimize the equalized odds difference would be in line with the existing unfairness mitigation benefits of Fairlearn (Dudík et al., 2020).

3 Literature Review

Machine learning classification models have been robustly tested for credit scoring purposes in the literature, many using the Statlog (German Credit Data) multivariable dataset available on the UC Irvine Machine Learning Repository, provided by Hans Hofmann (1994). This dataset has been widely used because it is well-organized, simple, composed of multivariate categorical and integer variable features, and already optimized for tasks determining a target outcome of ‘good’ or ‘bad’ credit risk (including Random Forest Classification, Logistic Regression, and Support Vector Classification). In 2019, Lkhagvadorj Munkhdalai et al. compared ML approaches to credit risk analysis with the typical FICO credit scoring system, finding that a “random forest-based new approach” was very successful and more accurate than a multi-test approach. Moreover, Anna Stelzer affirms that ‘ensemble’ models like random forest or boosting perform better as compared to individual models (2019). Given this, the author promotes replacing the logistic regression method that is still standard among most credit scorers with these ensemble classifier strategies (Stelzer, 2019).

Despite its widespread use, however, the German Credit dataset is very limited (with only 20 attribute features), and a poor candidate for assessing model unfairness— it does not measure the typical sensitive variables in credit scoring that have historically led to disproportionate credit classification. The dataset lacks race as a feature entirely and poorly combines information on sex and marital status, rendering the sex variable unusable (Grömping, 1994). Though it is more sparingly used, the ‘loanapp’ dataset available from J. Wooldridge’s "Introductory Econometrics: A Modern Approach, 7e” collection (2020) is a strong candidate for racial bias assessment in mortgage approval algorithms. This dataset, which was originally used in a Boston Federal

Reserve Bank research study (Hunter & Walker, 1996), measures socio-economic variables like race and ethnicity (including white, Hispanic, and Black identities), along with measures for sex, race, and educational status. The 'loanapp' dataset was recently used in 2024 classification technique research by University of Melbourne academic Chunyu Yang, an analysis that I will point to as a comparison point throughout this paper.

The unfairness mitigation portion of my project expands on a 2020 Microsoft/Ernst & Young case study by Dudík et al. that uses the Fairlearn toolkit to measure gender-based unfairness and perform mitigation strategies in loan application assessment, comparing to baseline measures afterwards. I use the 'Credit Loan Decisions' Fairlearn notebook page modeled *after* that case study as a guide (performed on a large 2005 Taiwanese credit card default dataset). Dudík et al. use parity in model performance—like accuracy rate—or parity in selection rate as measures of fairness for binary classification (2020). In short, the Fairlearn toolkit works to first assess fairness in the existing model and then perform mitigation strategies on the model (namely postprocessing algorithms and reduction algorithms), so we may compare them to our pre-mitigation values and see if they have made a difference (Dudík et al., 2020).

In the following fairness assessment exercise, I am particularly interested in precision as an evaluation metric over accuracy or recall because in credit risk assessment, it is worse for the lender to classify applicants as good risks when they are bad than vice versa (Hofmann, 1994). In other words, credit scorers seek to minimize the false positive rate of application approval predictions (from applicants that should not have been approved) and capture the highest number of true negatives (or rejections of applicants that should have been approved). However, it should be noted that from the applicant perspective, both false positives and false negatives present harm—false negatives withhold loans and the opportunities that come with them, while false positives can have adverse effects on credit scores and future loan applications if the recipients do end up defaulting on their loans (Dudík et al., 2020).

4 Methods

Ensemble models, which combine the predictions of multiple classifiers to strengthen the prediction ability (Stelzer, 2019), perform particularly well on Hofmann's German Statlog Credit dataset. I use this to guide the methodology behind my own credit risk assessment on the

Wooldridge ‘loanapp’ dataset. The UCI Machine Learning Repository publicizes the baseline model performance for accuracy and precision of German Credit data across various classification models, including Xgboost, support vector classification (SVC), random forest, neural network, and logistic regression classification. Among these, the random forest model has the highest range of accuracy and shares the highest range of precision with the SVC model (UCI ML Repository). Using the established Python packages of pandas, numpy, seaborn, matplotlib, and scikit-learn, I first predict binary mortgage approval for 7 different classification models.

In the feature preparation stage, I perform an 80:20 train test split for training and testing, respectively. Unlike Yang (2024), I do not oversample rejected mortgage applications (following the logic of the Fairlearn Credit Loan Decisions procedure, in which the classification models are first trained using unbalanced testing datasets). Additionally, I use the StandardScaler() package to standardize the data to have a mean of 0 and a standard deviation of 1, given that the dataset includes a mix of continuous (non-binary) and binary numeric variables. Since all are numeric, label encoding of categorical variables is not necessary. I use a scikit-learn Simple Imputer (and the most_frequent strategy) to account for and transform the few NaN missing data points in my dataframe (most common in the ‘MIN30’ variable, with a few occurrences in ‘MALE’, ‘THICK’, ‘MULTI’, ‘DEP,’ and ‘MARRIED’ variables, as well). This choice differs from Yang's, who opted to remove all entries with NaN values (2024).

After this, using the decision tree, k-nearest neighbors, Naive Bayes, logistic regression, SVM, random forest, and AdaBoost classification model packages from scikit-learn, I have our 26 selected variables predict the positive ‘APPROVE’ outcome (where 1 = mortgage application approval). Since I am predicting application *approval* as my outcome variable as opposed to default, the outcome measure of interest in Dudík et al.’s 2020 white paper, I effectively assign false positive and false negative definitions that are inverse to them within my modeling and analysis. This paper takes false positive rates to be the prediction of loan approval where the true outcome was rejection and false negative rates to be the prediction of loan rejection where the true outcome was approval.

For evaluation metrics, I calculate accuracy, precision, and recall scores for both the test and train sets and draw confusion matrices. In addition, I plot the receiver operating characteristic (ROC) and area under the curve (AUC) for most models to further visualize the tradeoff between

false positive rate and true positive rate, also using scikit-learn models. Finally, I examine the feature importance scores in random forest, and compare them to the importances scores found by Yang (2024).

Once I substantially evaluate the classification performance on the the Wooldridge ‘loanapp’ dataset, I turn to the more novel part of my methodology—the unfairness mitigation procedure outlined in the Microsoft and EY paper describing the Fairlearn toolkit’s purpose (Dudík et al., 2020) and reproduced in the Fairlearn Credit Loan Decision example notebook. Unlike Dudík et al. (2020), who designate sex as their feature that may lead to fairness-related harms in the algorithmic assessment, I am selecting a race-related variable (‘WHITE’) as my feature of interest. Though the Fairlearn toolkit offers two approaches to mitigate unfairness, I am focusing on the postprocessing approach over the reductions one. This is because the postprocessing model better suits the Wooldridge ‘loanapp’ dataset that comes loaded with the sensitive features of race and sex already.

Here, for fairness assessment purposes, I do resample the training dataset (where loan approvals are imbalanced in the raw data) and create a balanced_ids variable using random choice. Following the Fairlearn example procedure, I also introduce a new type of classifier: the gradient-boosted tree lightgbm model. Having already explored feature importances, I define the fairness metrics based on harms in line with the Fairlearn tutorial, then use the MetricFame object to compute the disaggregated performance scores and plot the metric for each group (Dudík et al., 2020). This is performed with the intention of finding an equalized odds difference (the maximum of the false positive rate difference and false negative rate difference) for the unmitigated model. Minimizing the equalized odds difference will, in theory, minimize these metrics after processing if we take false positives and false negatives to have equal costs to the non-white group (a heuristic for this case study). Finally, Fairlearn’s ThresholdOptimizer algorithm uses our balanced accuracy subject to the equalized odds fairness constraint “resulting in a thresholded version of the underlying machine learning model” (Fairlearn Credit Loan Decisions Example Notebook), giving us metric frames to use for comparison.

4.1 Data

Drawing on Yang (2024), my dataset of interest is the ‘loanapp’ dataset from J. Wooldridge’s "Introductory Econometrics: A Modern Approach, 7e” collection (2020). This dataset, which was originally used in a Boston Federal Reserve Bank research study (Hunter & Walker, 1996), was loaded with 59 columns (variables) and 1,989 sample entries of mortgage approval decisions and predictor variables. One of these columns is the outcome/response variable, ‘APPROVE’, which is set to 1 if the application is approved and 0 if it is not, mirrored by the ‘REJECT’ variable). In his work with a Wooldridge dataset containing nearly identical variables, Yang found that “25 of these predictions were deemed irrelevant to the current analysis due to reasons such as minimal positive observations or a lack of logical correlation with the response variable 'APPROVE"’ (2024).

Generally, I supported Yang’s reasoning for variable selection and replicated his choice within the ‘loanapp’ dataset for my own models to perform classification tasks, measure feature importance, and mitigate fairness-related harms with Fairlearn. However, my variable choice differed slightly. The 2020 Wooldridge package I used did not provide ‘ASIAN’ or ‘PUBHIST’ as variables for the ‘loanapp’ dataset, so I chose to select ‘HISPAN’ (a different ethnicity-based classification) and ‘PUBREC’ (a similar measure of public debt) as predictor variables instead. Additionally, I chose to include ‘MIN30’ and ‘OLD‘ in my modeling and analysis as additional socio-economic predictors that could result in fairness discrepancies. The full list of select variables and their explanations I used is provided in Figure 1. Variables with ‘=1’ in their explanation are binary in the selected variables dataset, while the rest are continuous and non-binary.

Variable	Explanation	Variable	Explanation
APPROVE	=1 if the mortgage application is approved	DEP	number of dependents
APPINC	applicant’s annual income in thousands of dollars	MARRIED	=1 if applicant is married
HRAT	monthly housing expenditures to monthly income ratio	MALE	=1 if male is used as applicant’s identification
OBRAT	total monthly obligations to monthly	PUBREC	=1 if filed bankruptcy

	income ratio		
EMP	years employed in current line of work	SCH	=1 if the applicant has more than 12 years in school
SELF	=1 if applicant employed herself	LOANAMT	loan amount in thousands of dollars
LIQ	amount of liquid assets in thousands of dollars	LOANPRC	loan amount to purchase price ratio
NETW	net worth in thousands of dollars	THICK	=1 the application file is thick if more than two credit reports are observed
PRICE	purchase price of the property	WHITE	=1 if applicant is white
APR	appraised value of the property	BLACK	=1 if applicant is Black
MULTI	=1 if the property is a multi-bedroom family	HISPAN	=1 if applicant is Hispanic
OTHER	amount of other financing in thousands of dollars	MIN30	=1 if minority population > 30 percent
ATOTINC	total monthly income	OLD	=1 if applicant age > MSA median

Figure 1. Selected variables for model fitting, with outcome variable APPROVE in top left. Based on Yang's prediction selection table (2024), with additional variables found and explained in the Wooldridge 'loanapp' dataset description (2020).

5 Results & Discussion

The distribution between white and non-white is imbalanced; 84.5% of the 1,989 total applicants reflected identify as white, while 15.4% do not. When viewed on a count plot, as seen in Figure 2, we can see a discrepancy between white vs. non-white applicant classification and loan approval vs. rejection. This is further described in the cross-tabulation in Figure 3, where 90.80% of white-identifying applicants had their mortgage applications approved, while only 70.80% of non-white applicants were approved. Though the sample of non-white applicants is much smaller, we can already see there could be ample reason for the 'WHITE' variable to be scored highly in terms of feature importance.

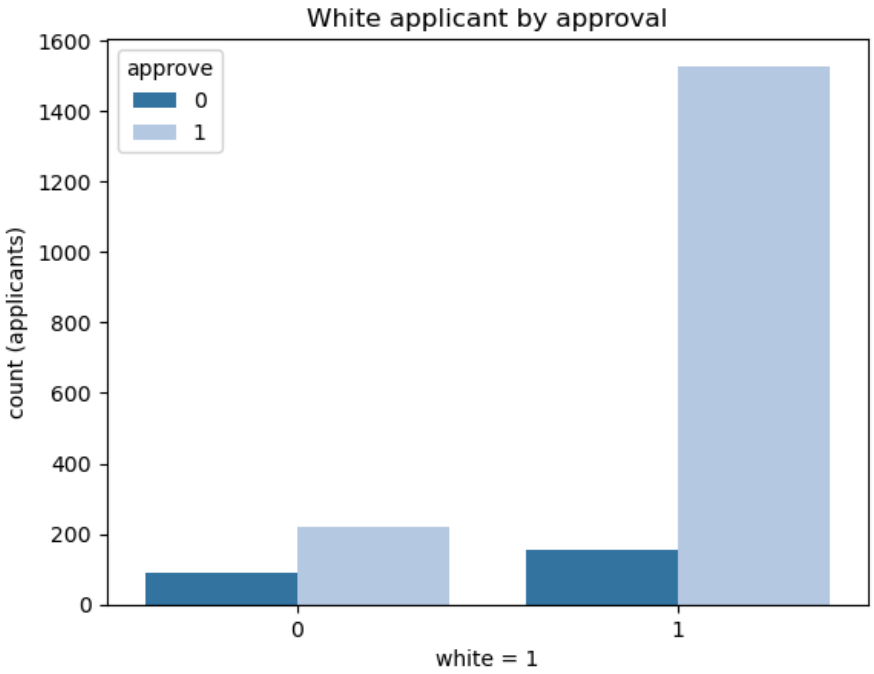


Figure 2. Seaborn count plot of counts, white vs. non-white, sorted by mortgage application approval hue colors for Wooldridge loanapp dataset. N = 1,989.

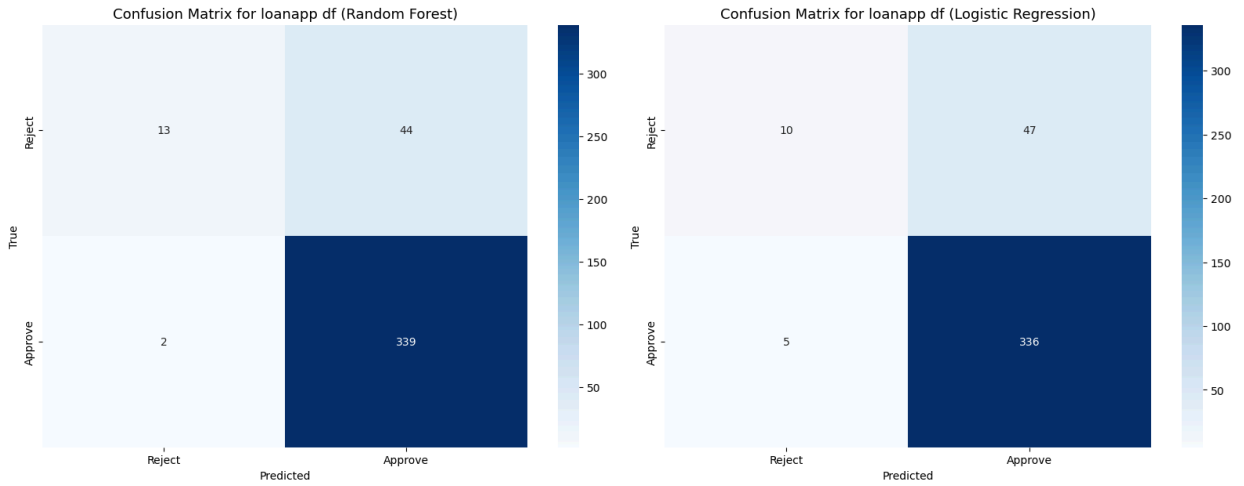
approve	**0 (not approved)**	**1 (approved)**
white		
0 (non-white)	29.20% (90)	70.80% (218)
1 (white)	9.20% (154)	90.80% (1527)

Figure 3. Table of 'WHITE' predictor percentages and counts cross-tabulated against 'APPROVE' outcome for Wooldridge loanapp dataset. N = 1,989.

As consistent with typical model evaluation metrics on the German Credit dataset (UCI ML Repository), random forest classification performs among the best in terms of accuracy (0.884) and precision (0.876). The logistic regression performance is close behind, with an accuracy score of 0.869 and a precision score of 0.772. The outcomes of the confusion matrices for these two classification models are presented as Figures 4 & 5 below. As we can see, both of the top-performing confusion matrices are heavily biased towards mortgage loan approval, which is fitting for the raw distribution of approvals as compared to rejections in the dataset. This is true for the rest of the classifiers in my project, as well. Due to Yang's powerful oversampling of rejected applications in his transformed 'loanapp' dataset, his confusion matrices have more variety (2024). His logistic regression model is quite balanced with similar false positive and

false negative rates, while most of the rest of his are biased towards loan rejection overall with a higher rate of false negatives (Yang, 2024).

Full figures of each classification model's accuracy, precision, and recall on the train and test sets for my project can be found in the appendix. My results differ from those of Yang (2024), who found that logistic regression had the highest test accuracy of 0.76 and SVM had the second highest accuracy of 0.73 (random forest was third, at 0.71). This may be due to the fact that Yang oversampled loan rejections to a high degree, while my project used the original distribution of 'APPROVE' data points.



Figures 4 & 5. Confusion matrices of test data for loanapp df Random Forest and Logistic Regression models, the two models with the highest test accuracy and precision scores. N = 1,989.

I was also able to evaluate the model performances using the ROC curves of each, which plots the true positive rate against the false positive rate and is a common measure of "the ability of models to distinguish between good and bad borrowers" (Munkhdalai et al., 2019). The ROC curves of all but one of my classification models used, as compared to that of a random classifier (the blue dashed line), can be viewed in Figure 6. The area under the curve scores for each can be viewed in the legend, a single numerical measure of performance (with random forest's AUC equal to 0.774 and logistic regression's AUC equal to 0.769). The highest AUC possible is 1, meaning that my top two classification models performed relatively well at classifying data points into the 'APPROVE' or 'REJECT' mortgage application outcomes.

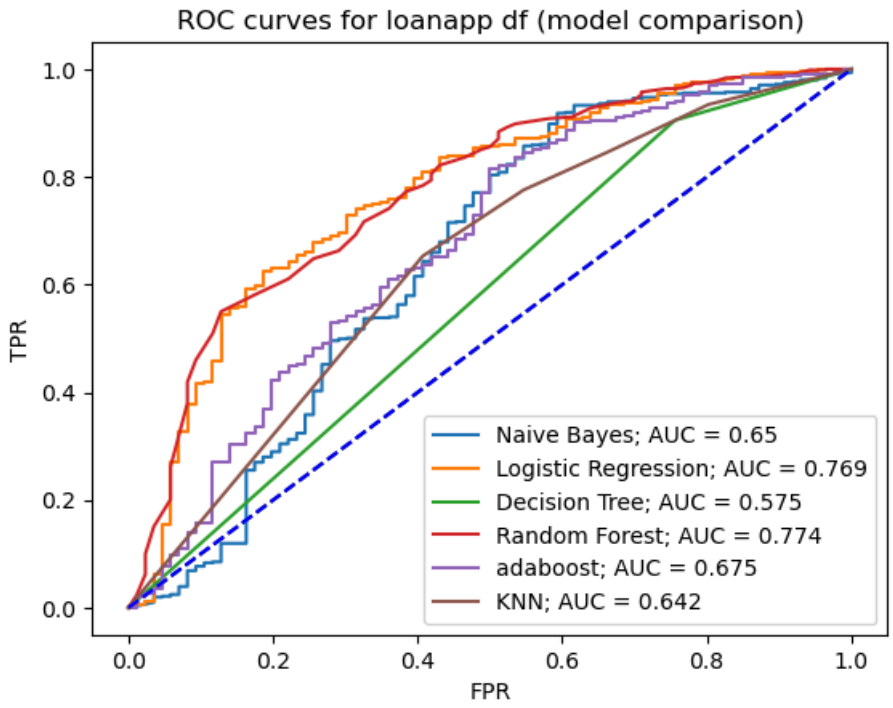


Figure 6. ROC curves for loanapp df (comparison across Naive Bayes, Logistic Regression, Decision Tree, Random Forest, AdaBoost, and K-Nearest Neighbor). AUC scores for each model are printed in the legend. N = 1,989.

A bar plot of feature importance scores calculated by RandomForest for my 'loanapp' data can be seen in Figure 7. Though the 'WHITE' variable does not present to be particularly important here (with a score that is ranked 13th out of 26 possible features), it is still the highest ranked of all socio-economic variables which may be especially prone to biased credit outcomes in the mortgage application process. Yang's feature selection graphs shows a far higher feature importance ranking for the 'WHITE' variable (it is the second-most important feature across multiple models), perhaps due to his choice to oversample 'REJECT' data points (2024).

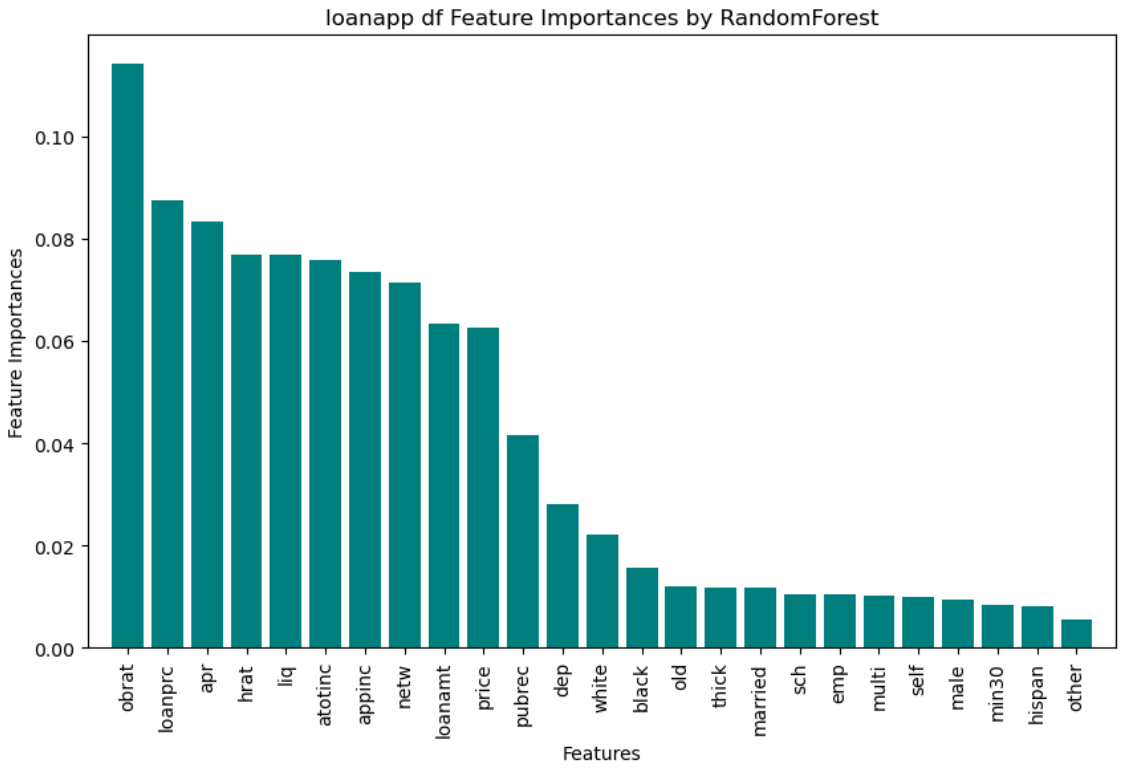


Figure 7. Random Forest feature importance scores. 'WHITE' is the socio-economic demographic variable with highest importance score. N = 1,989.

Finally, figures 8 and 9 below visualize the impact of the Fairlearn Credit Loan Decisions technique. They show metric frames (balanced accuracy, false positive rate, and false negative rate) before and after trying the Fairlearn unfairness mitigation techniques. Here, we see that even before mitigation, the balanced accuracy between the white and non-white group is near even (this marks a shift from unbalanced testing insights). Notably, before mitigating unfairness, the model produces a *lower* false positive rate for non-white applicants as compared to white ones, but a higher false negative rate for non-white applicants as compared to white ones. After using the ThresholdOptimizer algorithm and mitigating unfairness, the difference between the two groups is less noticeable, and the non-white group achieves slightly higher balanced accuracy than before (beating out the white group). It is evident that our Fairlearn unfairness mitigation approach, using the postprocessing model, works on this Wooldridge 'loanapp' dataset. Postprocessing for lightgbm classification reduces the false negative rate for the non-white group to be near the original rate of the white group, while raising the false positive rate of the non-white group so it approaches the original white group rate. Together, these postprocessing adjustments successfully minimize the equalized odds difference.

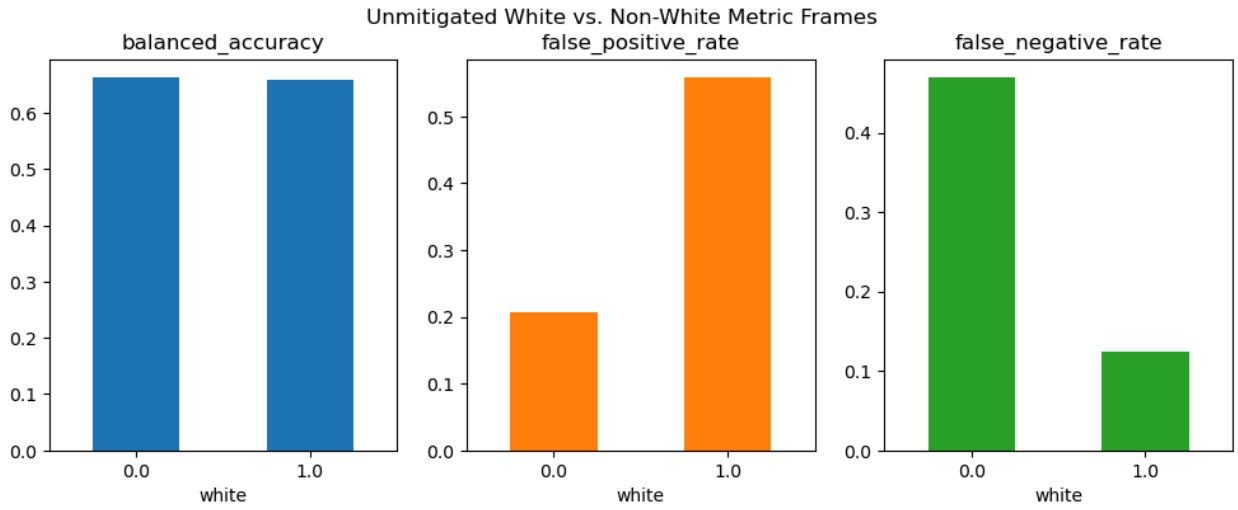


Figure 8. Metric frames (balanced accuracy, false positive rate, and false negative rate) for the 'WHITE' feature with *unmitigated* unfairness. The applicant identifies as 'white' where white=1.0. Based on Fairlearn Credit Loan Decisions example notebook.

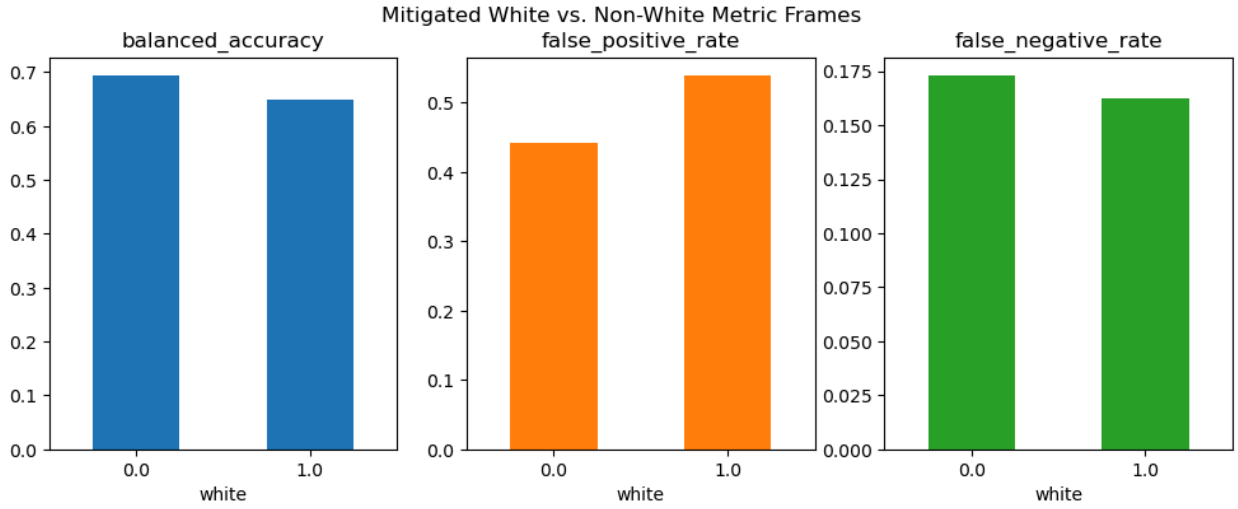


Figure 9. Metric frames (balanced accuracy, false positive rate, and false negative rate) for the 'WHITE' feature with *mitigated* unfairness. The applicant identifies as 'white' where white=1.0. Based on Fairlearn Credit Loan Decisions example notebook.

6 Limitations & Conclusion

The success of the Fairlearn toolkit in mitigating unfairness for credit risk assessment is a meaningful development, with applications in the consumer finance space and beyond to other fast-growing algorithmic assessment fields. The replication of Fairlearn success outside of their

development teams' papers and example notebooks is reassuring, proving the toolkit has merits outside of controlled settings. Implementation of Fairlearn and similar threshold optimization data processing techniques has the potential to rapidly open up borrowing opportunities for non-white applicants and other protected classes. Future research opportunities lie in multivariate unfairness mitigation among correlated sensitive features (like sex *and* race), multiclass classification bias mitigation, and further testing/replication of the alternate reductions approach.

However, there are some significant limitations present in the existing toolkit. First, the Fairlearn threshold optimization postprocessing model tested above is reliant on access to sensitive features at prediction time (Dudík et al., 2020). This fits for the Wooldridge 'loanapp' dataset, which includes variables like race, sex, and age, but would not be suited for many datasets that take a blind approach to sensitive features during training and prediction. As mentioned, the Fairlearn postprocessing model also takes false positives and false negatives to have equal costs to the non-white group (Dudík et al., 2020), but the actual calculus is less cut-and-dry. In reality, applicants might prefer false positives that damage their credit score over false negatives that withhold loan opportunities from them or vice versa, depending on their individual situation. Finally, the Fairlearn unfairness mitigation process is reliant on oversampling of unbalanced outcomes to create the balanced_ids variable, which might lead to overfitting and worsen the overall evaluation metric rates of the classification techniques. This technique's possible tendency to worsen performance is a serious harm, because it could push lenders and other algorithmic decision makers away from unfairness mitigation practices altogether.

7 Appendix

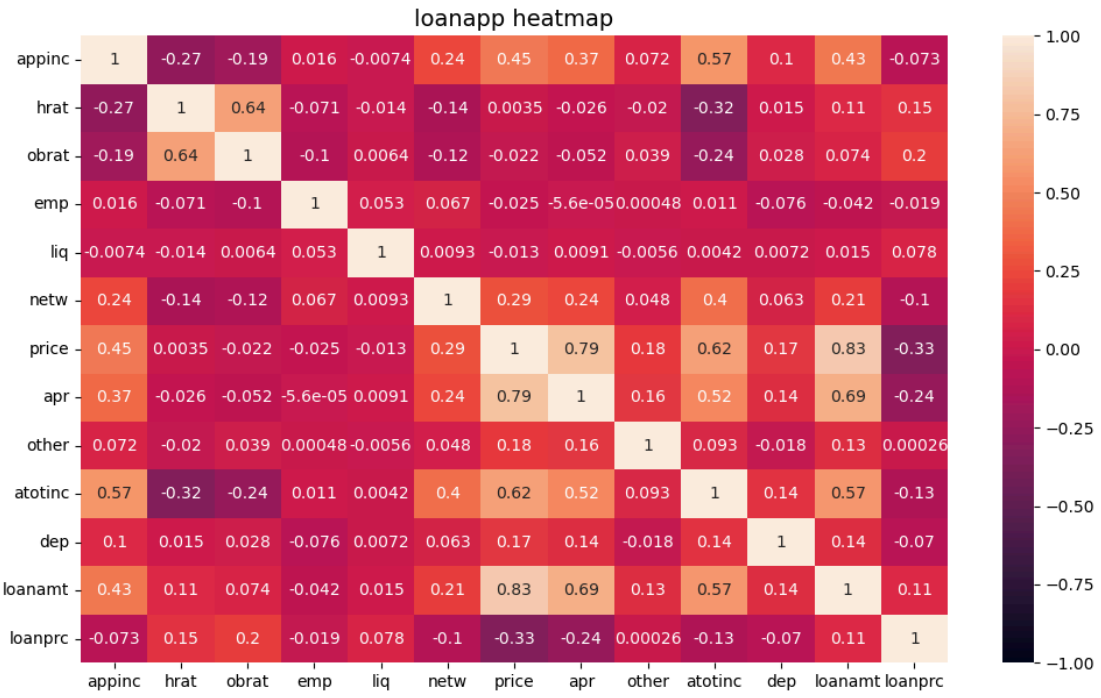
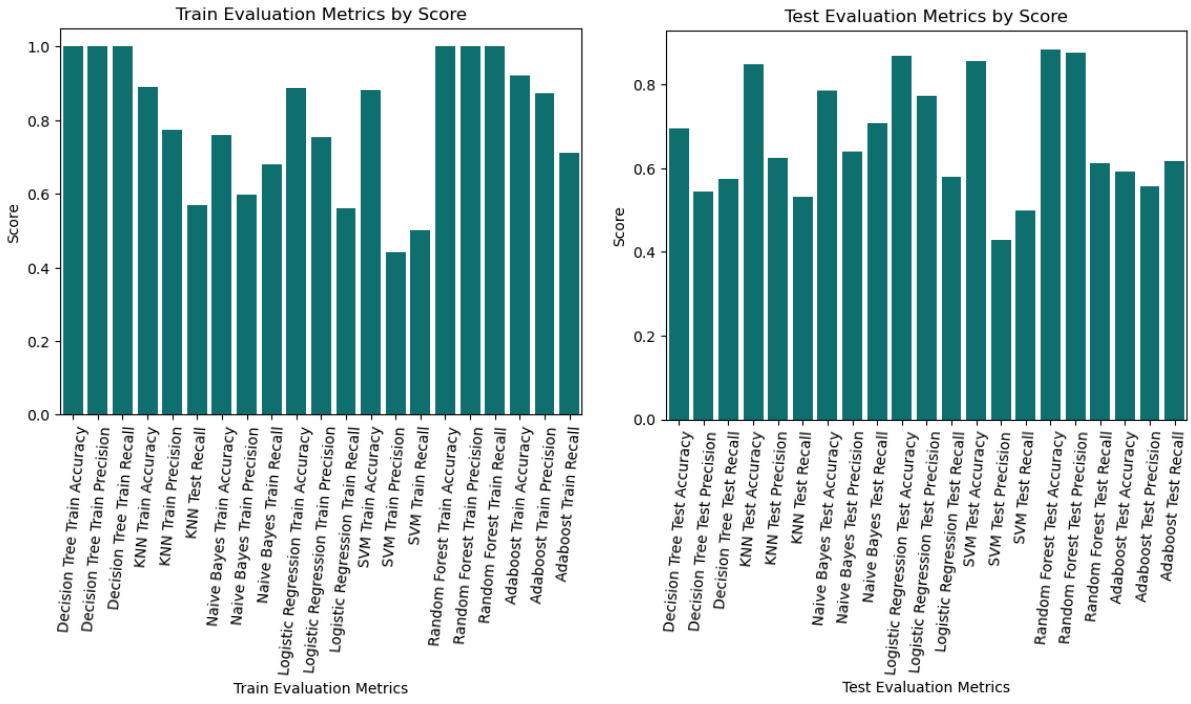


Figure A.1. Heat map for continuous (non-binary) selected variables in loanapp df. N = 1,989.



Figures A.2 and A.3. Bar plots of training and testing set evaluation metrics (accuracy, precision, and recall) for each classification model used.

8 References

Munkhdalai, L., Munkhdalai, T., Namsrai, O.-E., Lee, J. Y., & Ryu, K. H. (2019, January 29). *An empirical comparison of machine-learning methods on bank client credit assessments*. MDPI. https://www.mdpi.com/2071-1050/11/3/699.

Stelzer, A. (2019). Predicting credit default probabilities using machine learning techniques in the face of unequal class distributions. *ArXiv, abs/1907.12996*.

Hofmann, H. (1994). Statlog (German Credit Data) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5NC77.

South German Credit [Dataset]. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5QG88.

GeeksforGeeks. (2024, May 15). *How to visualize a decision tree from a random forest*. https://www.geeksforgeeks.org/ways-to-visualize-individual-decision-trees-in-a-random-forest/

Mahadevan, M. (2024, November 22). *Step-by-step exploratory data analysis (EDA) using Python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/

Predicting categorical data using classification algorithms - shiksha online. Shiksha Online. (n.d.). https://www.shiksha.com/online-courses/articles/predicting-categorical-data-using-classification-algorithms/

Grömping, U. (2019). South German Credit Data: Correcting a Widely Used Data Set. *Berliner Hochschule Für Technik Reports in Mathematics, Physics and Chemistry*. https://www1.beuth-hochschule.de/FB_II/reports/Report-2019-004.pdf.

ML: Extra tree classifier for feature selection. GeeksforGeeks. (2023, May 18). https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/

Brownlee, J. (2020, June 3). *Feature selection in python with scikit-learn*. MachineLearningMastery.com. https://machinelearningmastery.com/feature-selection-in-python-with-scikit-learn/

Eight ways to perform feature selection with scikit-learn. Shedload Of Code. (2023, August 5). https://www.shedloadofcode.com/blog/eight-ways-to-perform-feature-selection-with-scikit-learn

Feature importance. Feature importance - Scikit-learn course. (n.d.). https://inria.github.io/scikit-learn-mooc/python_scripts/dev_features_importance.html

Weerts, H., Dudík, M., Edgar, R., Jalali, A., Lutz, R., & Madaio, M. (2023, March 29). *Fairlearn: Assessing and improving fairness of AI Systems*. arXiv.org. https://arxiv.org/abs/2303.16626

Dudík, M., Chen, W., Barocas, S., Inchiosa, M., Lewins, N., Oprescu, M., Qiao, J., Sameki, M., Schlener, M., Tuo, J., & Wallach, H. (2020). Assessing and mitigating unfairness in credit models with the fairlearn toolkit. https://www.microsoft.com/en-us/research/uploads/prod/2020/09/Fairlearn-EY_WhitePaper-2020-09-22.pdf

Fairlearn. (n.d.). *Credit loan decisions documentation (Example Notebooks)*. Microsoft. https://fairlearn.org/main/auto_examples/plot_credit_loan_decisions.html#footcite-dudik2020assessing

Munnell, Alicia H, Geoffrey MB Tootell, Lynn E Browne, and James McEneaney, “Mortgage lending in Boston: Interpreting HMDA data,” The American Economic Review, 1996, pp. 25–53.

Martinez, E., & Kirchner, L. (2021, August 25). *The secret bias hidden in mortgage-approval algorithms – the Markup*. The Markup. https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms

Małowiecki, A., & Chomiak-Orsa, I. (2024). Fairlearn parity constraints for mitigating gender bias in binary classification models – comparative analysis. *Communications in Computer and Information Science*, 148–154. https://doi.org/10.1007/978-3-031-50485-3_13

Wooldridge, J.M. (2020). 115 Data Sets from "Introductory Econometrics: A Modern Approach, 7e.” rdrr.io. https://rdrr.io/cran/wooldridge/

Hunter, W.C. and Walker, M.B. (1996), “The Cultural Affinity Hypothesis and Mortgage Lending Decisions,” Journal of Real Estate Finance and Economics 13, 57-70.

Wooldridge. PyPI. (n.d.). https://pypi.org/project/wooldridge/0.1.0/

Yang, C. (2024). Research on loan approval and credit risk based on the comparison of machine learning models. *SHS Web of Conferences*, *181*, 02003. https://doi.org/10.1051/shsconf/202418102003

Skowronski, J. (2017, April 14). *All about the equal credit opportunity act*. Human Rights Campaign. https://www.hrc.org/news/all-about-the-equal-credit-opportunity-act#:~:text=It%20was%20passed%20back%20in,Federal%20Reserve%20Bank%20of%20Philadelphia.

This project represents my own work in accordance with University regulations.

/s/ Giorgia Musselwhite