Technical Report

Appendix of the Paper: Fair Classification with a Scalable Reductions Approach

Andrea Baraldi[1], Matteo Brucato[2], Miroslav Dudík[2], Francesco Guerra[1], Matteo Interlandi[2]
[1]University of Modena and Reggio Emilia, [2]Microsoft
[1]{andrea.baraldi96,francesco.guerra}@unimore.it, [2]{mbrucato,mdudik,mainterl}@microsoft.com

This appendix provides additional material on the experimental study in the submitted paper. This extra material has been omitted for space reasons in the submitted version. It mainly consist of the results of the experiments performed with the gradient-boosted decision tree classifier. The description of the experiments already included in the paper is also reported for the sake of completeness.

1 EXPERIMENTAL EVALUATION

We conduct four types of experiments that analyze ExpGrad^{++} from four perspectives: (1) **End-to-end usage.** The goal of the experiments in Section 1.2 is to show that ExpGrad^{++} is an effective, efficient, and easy-to-tune approach for the development of end-to-end ML pipelines. We compare ExpGrad^{++} with other baselines in terms of training time (quantifying efficiency) as well as accuracy and fairness (quantifying effectiveness). (2) **Contribution of the column generation.** In Section 1.3 we conduct an ablation study to evaluate the impact of the *column generation* on the effectiveness and efficiency of the reduction approach. (3) **Robustness to sampling.** In Section 1.4 we evaluate how the sampling ratio ρ impacts the effectiveness and efficiency of ExpGrad^{++}. We also demonstrate that adaptive sampling leads to a better use of data (meaning more effective solutions) than static sampling. (4) **Support of multi-valued sensitive attributes.** ExpGrad^{++} supports fairness metrics defined over multi-valued sensitive attributes, but many existing approaches only support binary sensitive attributes, so when the actual sensitive attribute is multi-valued the user needs to artificially binarize its values by partitioning them into two groups. The experiment in Section 1.5 shows that this procedure does not effectively mitigate unfairness with respect to the original attribute values.

1.1 Experimental Settings

System and implementation. We conducted the experiments on a machine equipped with Intel(R) Xeon(R) Platinum 8370C CPU @ 2.80GHz and 62.8 GB RAM, running on Ubuntu 20.04.5 LTS operating system. The experiments were coded in Python 3.9.12 using the *fairlearn* library version v0.9.0.dev0. All of the experiment code is publicly available in the project github.

Datasets. Our evaluation includes 5 real-world datasets, summarized in Table 1. Adult [3], COMPAS [9], and German [1] are small datasets usually adopted as a benchmark in many fairness studies [8, 10?]. ACSEmployment and ACSPublicCoverage [5] are large datasets that allow us to perform an evaluation in larger scenarios and to experiment with multi-valued sensitive attributes.

Adult describes demographic and occupational attributes of several thousand individuals extracted from the 1994 US Census database. We predict whether an individual has an income higher than 50K, with gender as the sensitive attribute.

COMPAS contains arrest records, demographic information, and criminal history of defendants arrested in 2013–2014. We predict reoffense within two years, with race as the sensitive attribute.

German contains records of individuals applying for credit or loan to a bank. We predict whether the default risk of an individual is high or low, using sex as the sensitive attribute.

ACSPublicCoverage and *ACSEmployment* have been recently proposed to create new and more challenging machine learning tasks with the aim to establish strong empirical evaluation practices within the algorithmic fairness community. They have been created starting from US Census data collected within the American Community Survey. The datasets include information related to ancestry, citizenship, education, race, employment, language proficiency, income, disability, etc. The two datasets differ in the prediction task: ACSPublicCoverage contains data for predicting whether an individual is covered by public health insurance, ACSEmployment contains data for predicting whether an individual is employed. We select race (RAC1P in the dataset) as the sensitive attribute. Unlike other datasets, in this case, the sensitive attribute is multi-valued.

Models, hyperparameters, and other settings. We perform all experiments with two different base models: logistic regression and gradient-boosted decision trees, implemented using *LogisticRegression* (LR) and *HistGradientBoostingClassifier* (LGBM) classes from the *scikit-learn* library. Their hyperparameters are tuned separately for each dataset, without fairness constraints, using 5-fold crossvalidation. We consider two types of fairness constraints: demographic parity and equalized odds. The violation of fairness constraints is quantified using the demographic parity difference and the equalized odds difference.

For each experiment configuration, we evaluate the performance of each algorithm using stratified 3-fold cross validation, executed twice with different random seeds, resulting in six replications of each experiment. Stratification is performed by jointly considering the sensitive attribute and the label, ensuring a consistent sampling approach for both the training and testing splits. Each fold constitutes one-third of the entire dataset; consequently, in each iteration, the training set encompasses two-thirds, while the test set comprises one-third of the dataset. We have verified the consistency between the performance of the baseline models in our pipeline with a 70/30 train-test split as performed in other prior works.

Table 1: The datasets used in the experiments. Size is the dimension of the dataset, n is the number of data points (#rows), d is the number of features (#columns), S and |S| are the name and the cardinality of the sensitive attribute.

Dataset	Size (MB)	n	d	S	|S|
ACSEmployment	319.47	3.2×10^6	99	RAC1P	9
ACSPublicCoverage	163.26	1.1×10^6	140	RAC1P	9
Adult	4.67	4.5×10^4	9	Sex	2
COMPAS	0.37	4.2×10^3	3	Race	2
German	0.05	1.0×10^3	9	Sex	2

Additional fairness measures and evaluation metrics for ExpGrad^{++} have been included in the csv file published in the projetc github[1]. Specifically, we reported metrics from the Fairlearn library, such as:

- **Fairness metrics:** 'DemographicParity', 'EqualizedOdds', 'demographic parity difference', 'demographic parity ratio', 'dp diff bg' (demographic parity difference with respect to the best group), 'dp ratio bg' (demographic parity ratio with respect to the best group), 'equalized odds difference', 'equalized odds ratio', 'eo diff bg' (equalized odds difference with respect to the best group), and 'eo ratio bg' (equalized odds ratio with respect to the best group). Additional fairness-related metrics: 'tnrb' (true negative ratio), 'tnrb bg' (with respect to the best group), 'tprb' (true positive ratio), and 'tprb bg' (with respect to the best group).
- **Performance metrics:** 'accuracy', 'f1', 'precision', 'recall', and 'error'.
- **Efficiency metrics:** 'time (train)'.

These measures consistently confirm the findings of the paper, so for clarity we decided to keep in the following only the demographic parity difference and the equalized odds difference to evaluate the fairness level and error and time to evaluate effectiveness and efficiency.

Baselines. We compare ExpGrad^{++} with six baselines. CALMON [4] and FELD [6] are selected as representative pre-processing approaches, ZAFAR, which proposes two methods to enforce disparate impact (DI) [11] and equalized odds (EO) [12], as a representative in-processing approach, and HARDT [7] as a representative post-processing approach. Moreover, we include ExpGrad [2], the approach we are extending with column generation and sampling, and UNMITIGATED, i.e., the fairness-unaware approach. Not all the baselines can be used for all the experiments. In particular, ZAFAR EO (with equalized odds does) not support multi-valued sensitive attributes (i.e., it cannot be run on the datasets ACS*), and CALMON requires specifying a problem-specific distortion function, which was not available for ACS* datasets.

1.2 End-to-end evaluation

The goal of our first set of experiments is to demonstrate that ExpGrad^{++} effectively and efficiently supports the development of ML pipelines. As its input, it requires a bound on the allowed amount of fairness violation. As we show, it returns an accurate classifier that satisfies the fairness constraint, while exhibiting satisfactory running time.

Implementation. ExpGrad^{++} requires specification of the allowed constraint violation. We consider five values for ϵ, i.e., 0.005, 0.01, 0.02, 0.05, 0.10, and 0.15 for our model. In ACS* datasets we do not evaluate at the value $\epsilon = 0.15$, because the test error of ExpGrad^{++} already matches the UNMITIGATED approach at $\epsilon = 0.10$. In ExpGrad^{++}, we use the sampling ratio of 0.25 with the datasets ACSPublicCoverage and ACSEmployment, and do not use subsampling on the three smaller datasets.

In Figures 1 (LR) and 2 (LGBM), we show a compact representation of the comparison of the performance of ExpGrad^{++} with baselines across the 5 datasets, for both demographic parity as well as equalized odds (Figure 1b). For each of the datasets and each of the constraint types, we plot the test error and the running time required to train the model as a function of the constraint violation on test data. ExpGrad^{++} and ExpGrad are plotted as curves because they allow various fairness-accuracy trade-offs by specifying the bound ϵ. Other methods optimize a fixed trade-off and are plotted as points. Detailed results are shown in Figure 3 (LR, Demographic Parity), Figure 4 (LR, Equalized Odds), Figure 5 (LGBM, Demographic Parity), Figure 6 (LGBM, Equalized Odds). At the URL https://github.com/softlab-unimore/fairnesseval/blob/main/table_results.csv a Table with the overall evaluation of ExpGrad^{++} with different fairness metrics is published.

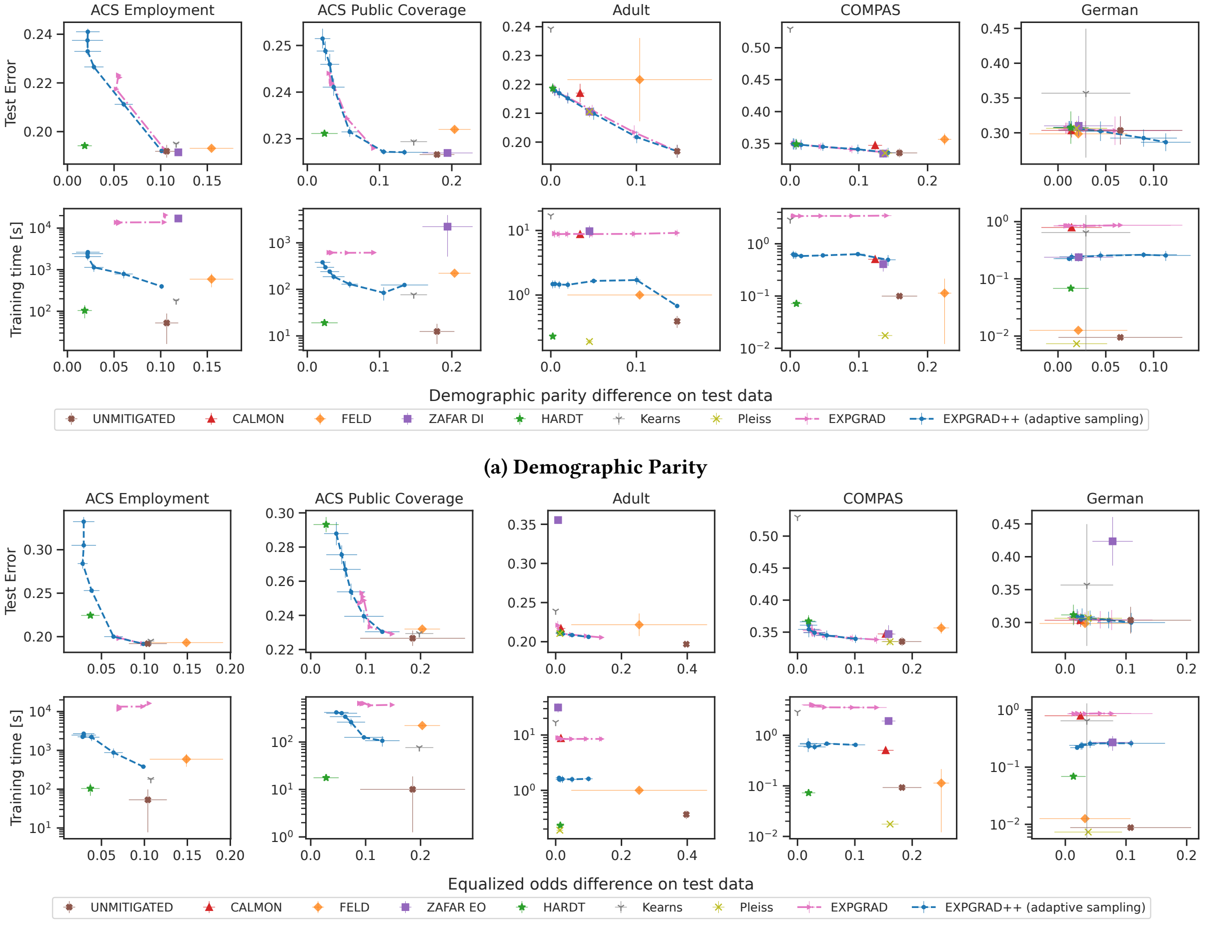
1.3 Impact of column generation effects

We next conduct an ablation study to evaluate the impact of the column generation component in the ExpGrad^{++} algorithm.

Implementation. For the sake of simplicity, we show the results on the Adult, COMPAS, and German datasets only, and we do not perform any subsampling. We compare the performance of ExpGrad^{++} with and without the column generation component; the version without column generation corresponds to the previous ExpGrad approach. In both versions, we set the allowed constraint violation to $\epsilon = 0.005$. We run ExpGrad^{++} with the default learning rate and default termination condition. For ExpGrad, we consider three different learning rates (specified in the *fairlearn* library via the hyperparameter `eta0` $\in \{0.5, 1.0, 2.0\}$), and for each, we take the best-performing solution among the three solutions (according to the duality gap). Note that we only report the time of one run, but not all three—this gives an advantage to ExpGrad.

In Figures 7 (LR), 8 (LGBM) we show how well the two algorithms optimize accuracy and fairness as a function of time. We focus on training metrics because these capture the progress of optimization (test metrics were reported in end-to-end evaluation). For ExpGrad^{++}, we plot the accuracy and fairness after reaching the termination condition based on the duality gap (since the algorithm always reaches the early termination condition), but for ExpGrad, we plot the performance at multiple different iterates, corresponding to different values of t. This again possibly gives ExpGrad some advantage, because we do not require it to know when to stop.

[1] https://github.com/softlab-unimore/fairnesseval/blob/main/table_results.csv



(a) Demographic Parity

(b) Equalized Odds

Figure 1: Logistic Regression Effectiveness and Efficiency on the test sets varying the level of violation (ϵ).

1.4 Robustness to sampling

In this Section, we investigate how the sampling ratio impacts the running time and the quality of solutions (i.e., test error and fairness violation) produced by ExpGrad^{++}.

Implementation. In all experiments, we set the allowed constraint violation to $\epsilon = 0.005$. We evaluate our *adaptive subsampling* approach for the sampling ratios $\rho \in \{0.001, 0.004, 0.016, 0.063, 0.251\}$. As a baseline, we also consider *static subsampling*, where the dataset is subsampled (uniformly at random) once at the beginning and then ExpGrad^{++} is run on the subsampled dataset (without any further subsampling). In both cases, we run ExpGrad^{++} with column generation and a default setting of optimization hyperparameters. As before, we consider both demographic parity and equalized odds. Additionally, we also report performance on unmitigated approach trained on the same subsampled dataset as the static ExpGrad^{++}.

We compute how the running time, test error, and test constraint violation vary as a function of sampling ratio: Figure 9 (LR, Demographic Parity), Figure 10 (LR, Equalized Odds), Figure 11 (LGBM, Demographic Parity), Figure 12 (LGBM, Equalized Odds)..

1.5 Managing multi-valued sensitive attributes

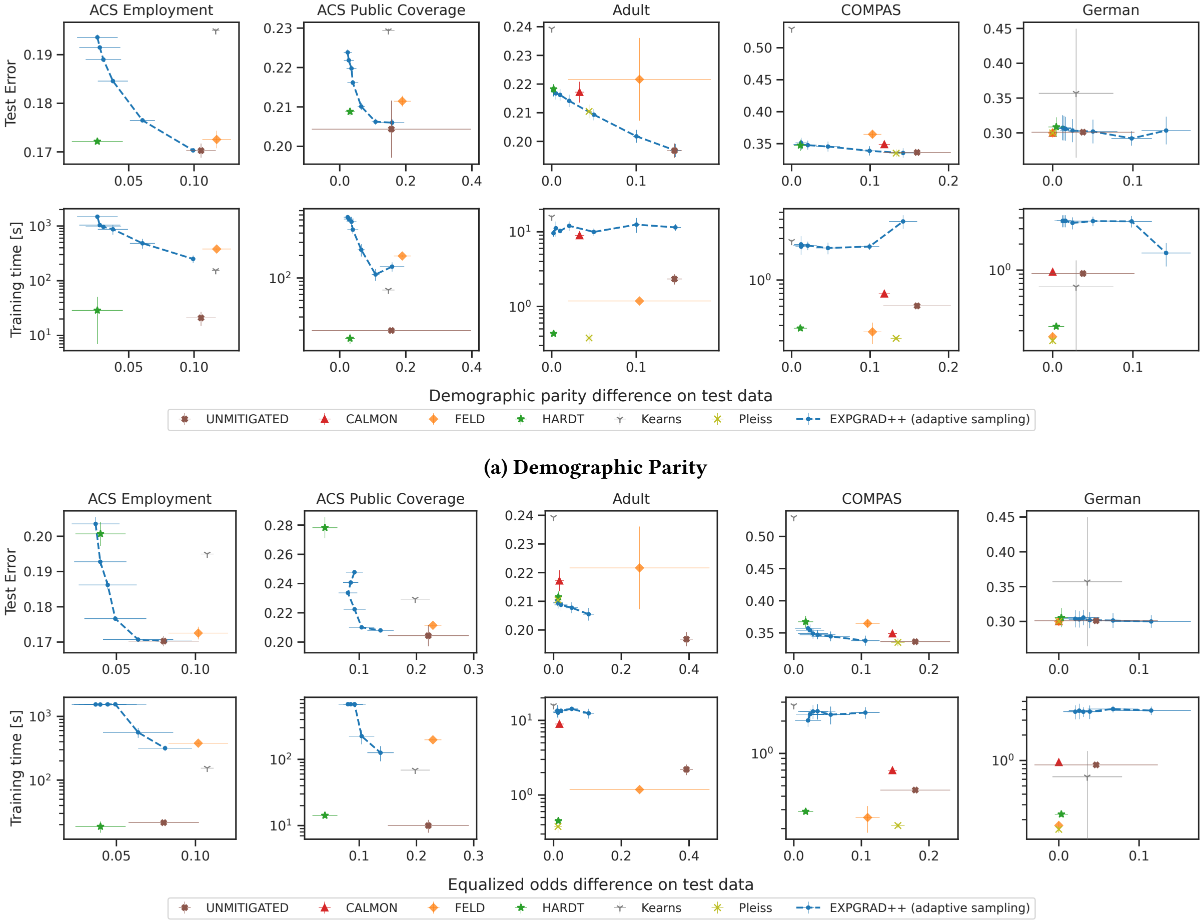
Here we demonstrate that although many fairness mitigation approaches assume that sensitive attributes are binary, full support of multi-valued sensitive attributes (as done by ExpGrad^{++}) is necessary to mitigate fairness violations.

Implementation. We perform experiments with the ACSEmployment dataset, a large dataset with a multi-valued sensitive attribute (RAC1P). We either consider the multi-valued sensitive attribute as is, or consider its binarized version. In binarized versions, we split 9 race categories into two groups based on the average employment rate in each group, relative to the median employment rate (across all 9 groups).

In Figure 13 we show the fairness violations using the *multi-valued* variants of demographic parity difference and equalized odds difference, where we consider training it using either multi-valued or binarized sensitive attribute for each method.

REFERENCES

[1] 2020. German credit risk. Kaggle. https://www.kaggle.com/uciml/ german-credit.



(a) Demographic Parity

(b) Equalized Odds

Figure 2: LGBM Effectiveness and Efficiency on the test sets varying the level of violation (ϵ).

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *International Conference on Machine Learning*. 60–69.

[3] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[4] Flávio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *NIPS*. 3992–4001.

[5] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *NeurIPS*. 6478–6490.

[6] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *KDD*. ACM, 259–268.

[7] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1610.02413.

[8] Maliha Tashfia Islam, Anna Fariha, Alexandra Meliou, and Babak Salimi. 2022. Through the Data Management Lens: Experimental Analysis and Evaluation of Fair Classification. In *SIGMOD Conference*. ACM, 232–246.

[9] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. ProPublica.

[10] Dana Pessach and Erez Shmueli. 2023. A Review on Fairness in Machine Learning. *ACM Comput. Surv.* 55, 3 (2023), 51:1–51:44.

[11] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *WWW*. ACM, 1171–1180.

[12] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS (Proceedings of Machine Learning Research)*, Vol. 54. PMLR, 962–970.

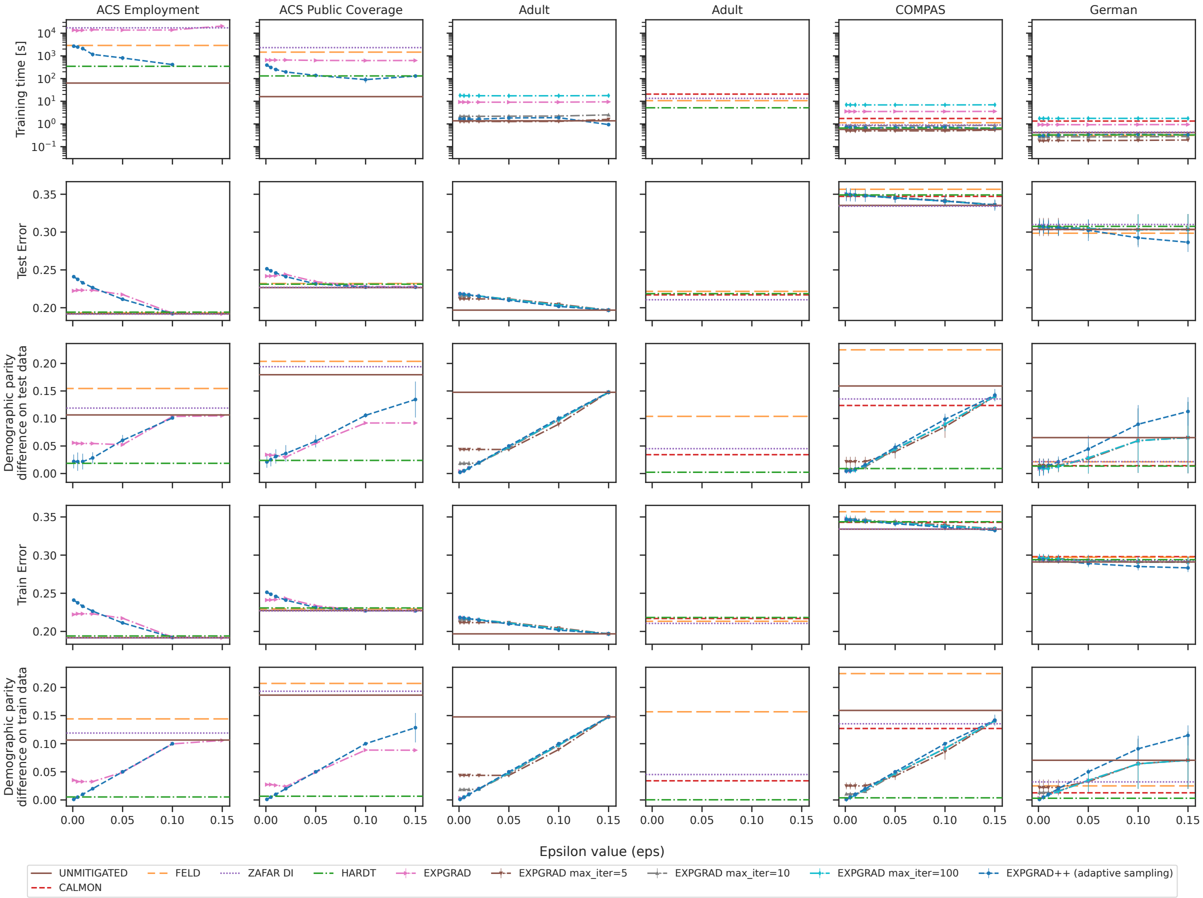


Figure 3: LR Effectiveness and Efficiency on train and test sets varying the level of violation (ϵ) for Demographic Parity constraint.

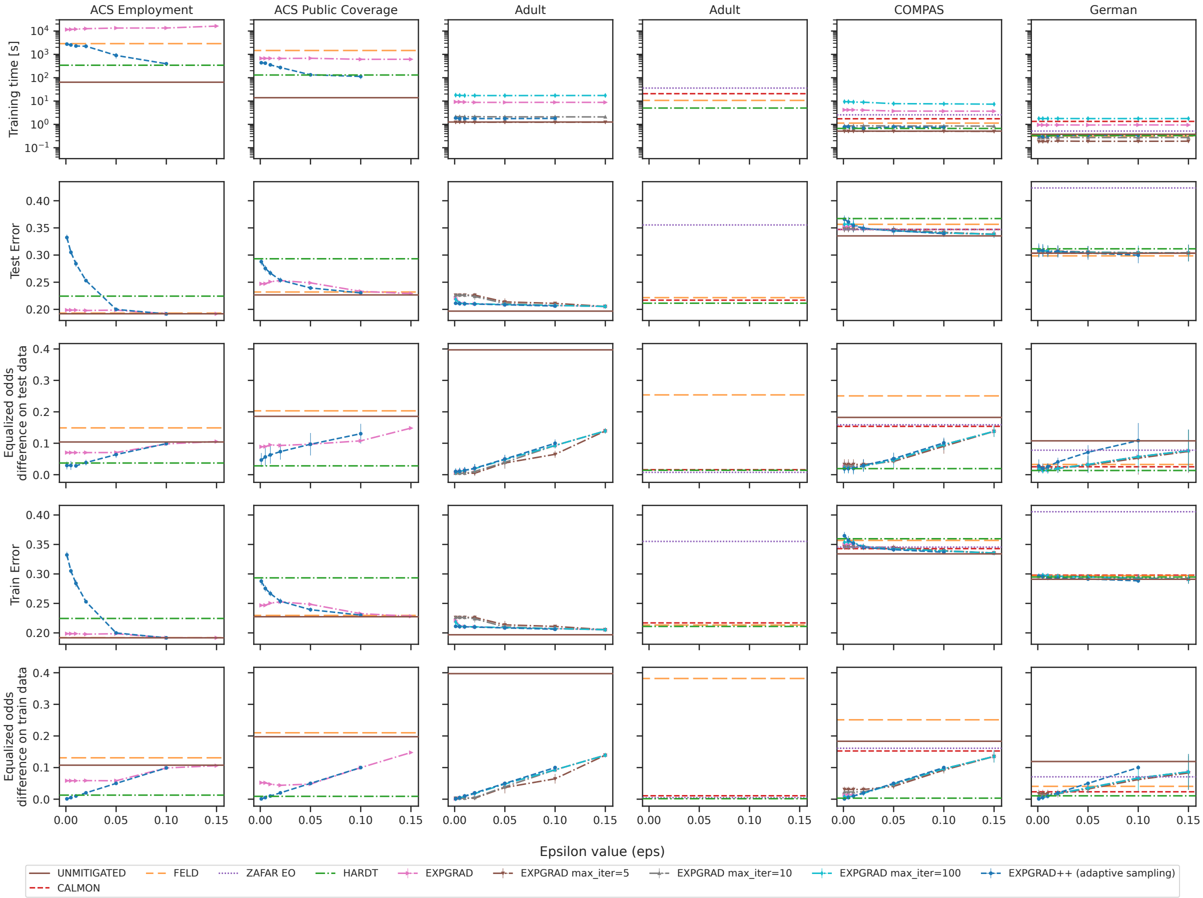


Figure 4: LR Effectiveness and Efficiency on train and test sets varying the level of violation (ϵ) for Equalized Odds constraint.

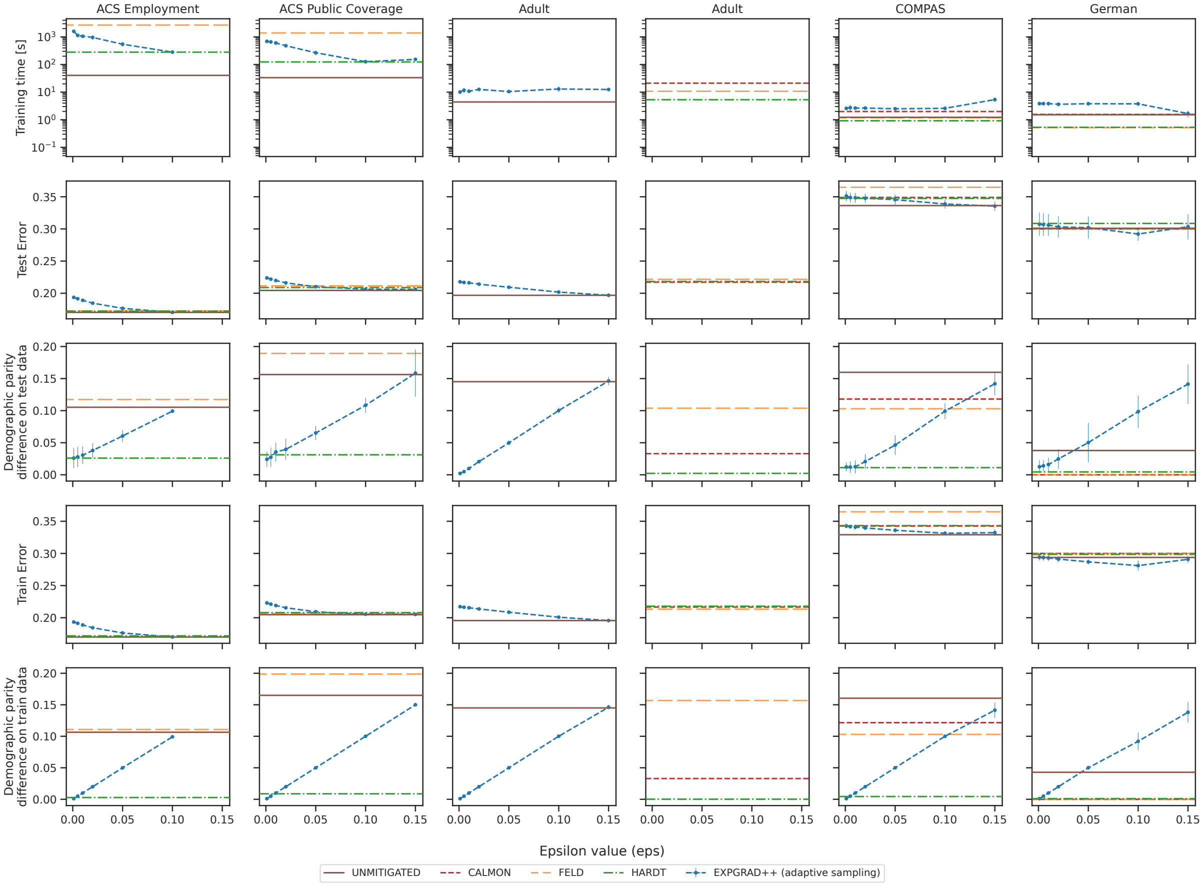


Figure 5: LGBM Effectiveness and Efficiency on train and test sets varying the level of violation (ϵ) for Demographic Parity constraint.

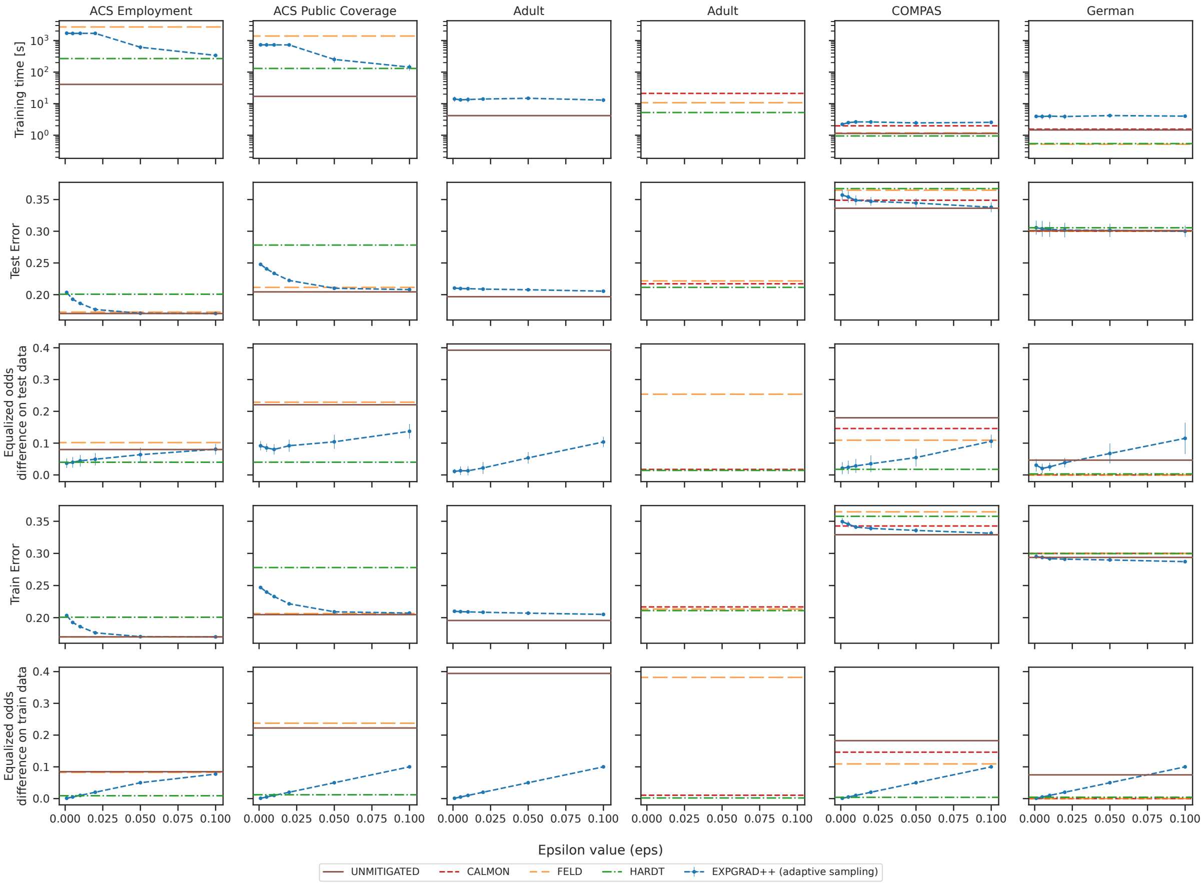
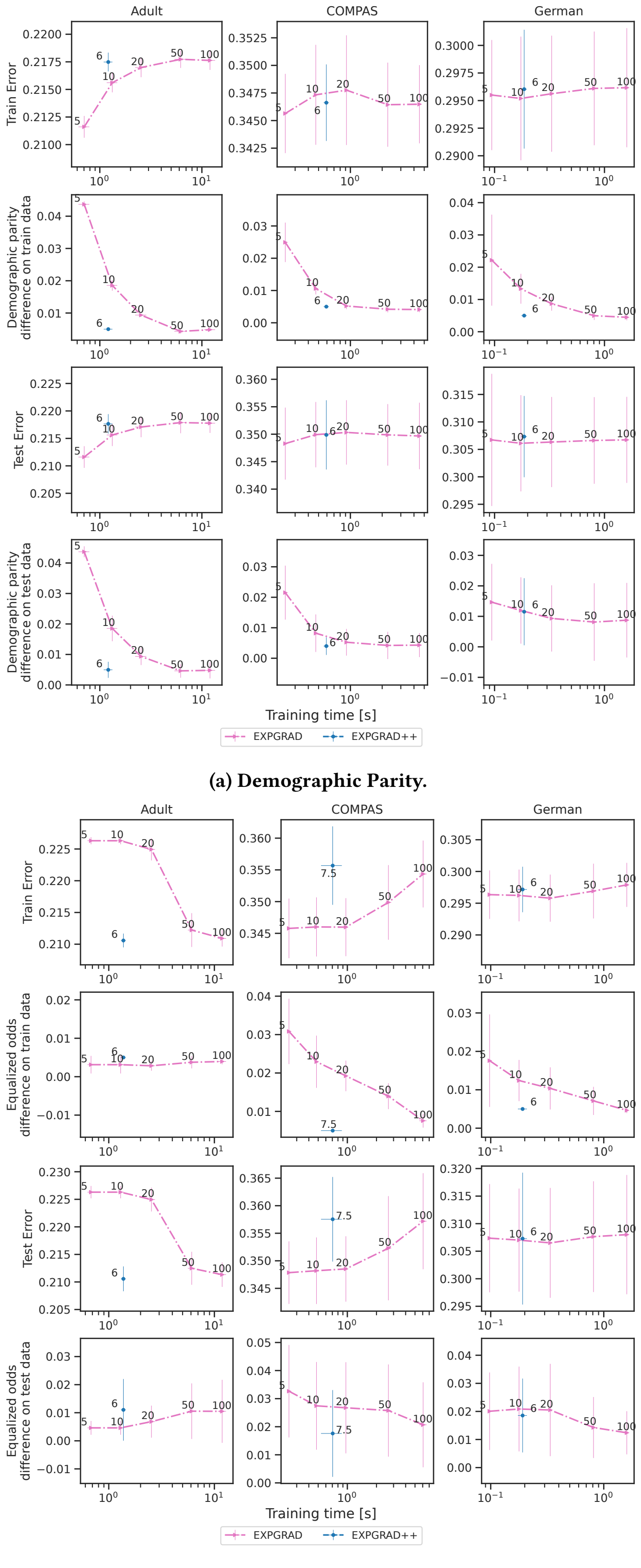


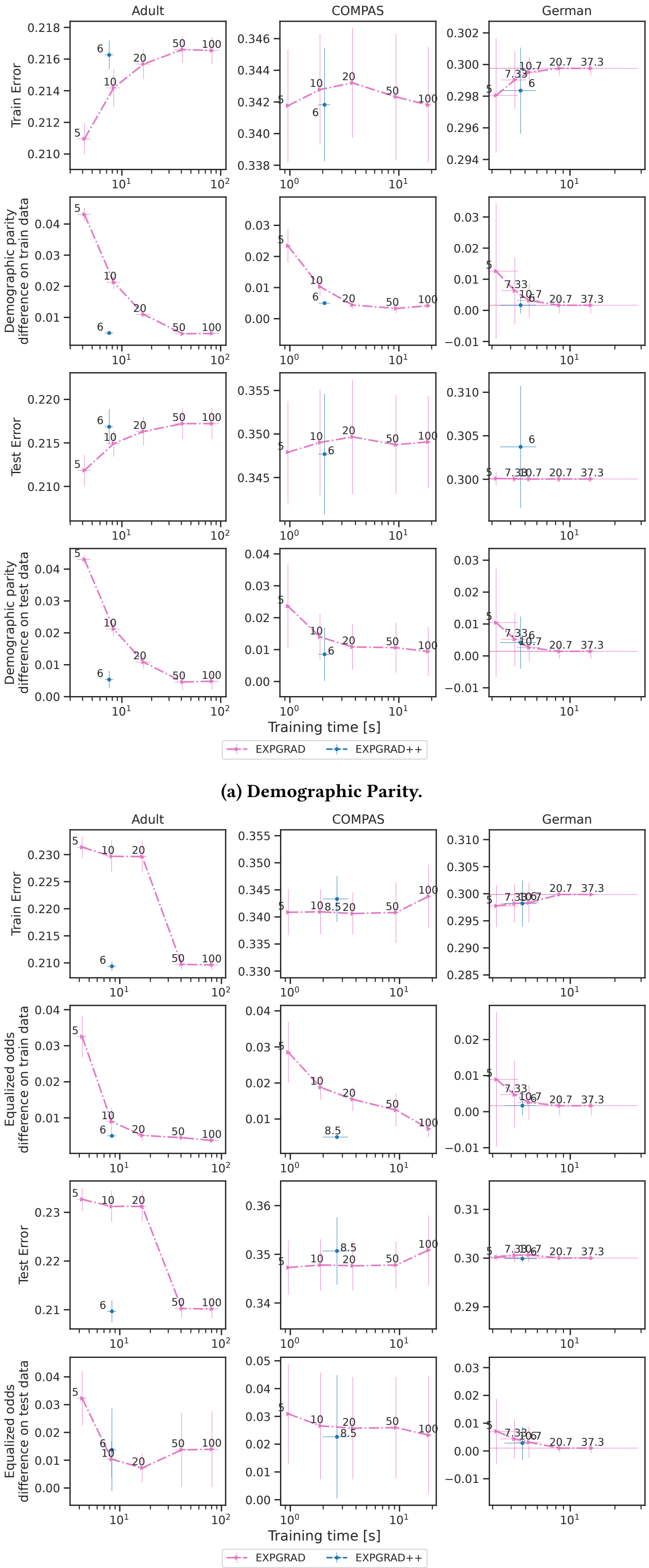
Figure 6: LGBM Effectiveness and Efficiency on train and test sets varying the level of violation (ϵ) for Equalized Odds constraint.



(a) Demographic Parity.

(b) Equalized odds.

Figure 7: Ablation of the column generation component with LR base model. The numbers represent the value of *max_iter*, i.e., the numbers of calls to the classifier to converge.



(a) Demographic Parity.

(b) Equalized odds.

Figure 8: Ablation of the column generation component with LGBM base model. The numbers represent the value of *max_iter*, i.e., the numbers of calls to the classifier to converge.

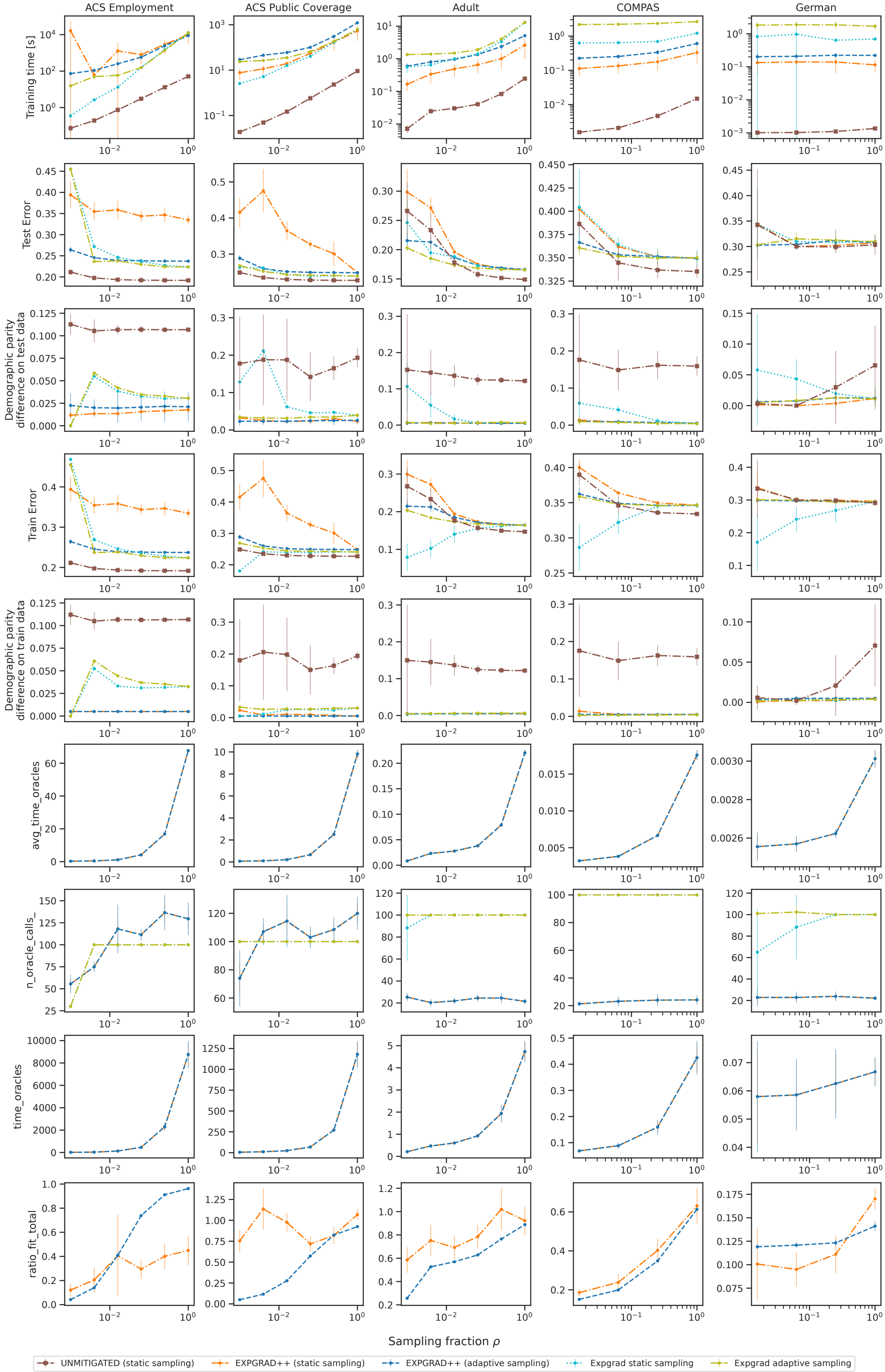


Figure 9: LR Effectiveness and Efficiency on the test sets varying the sampling size for Demographic Parity constraint.

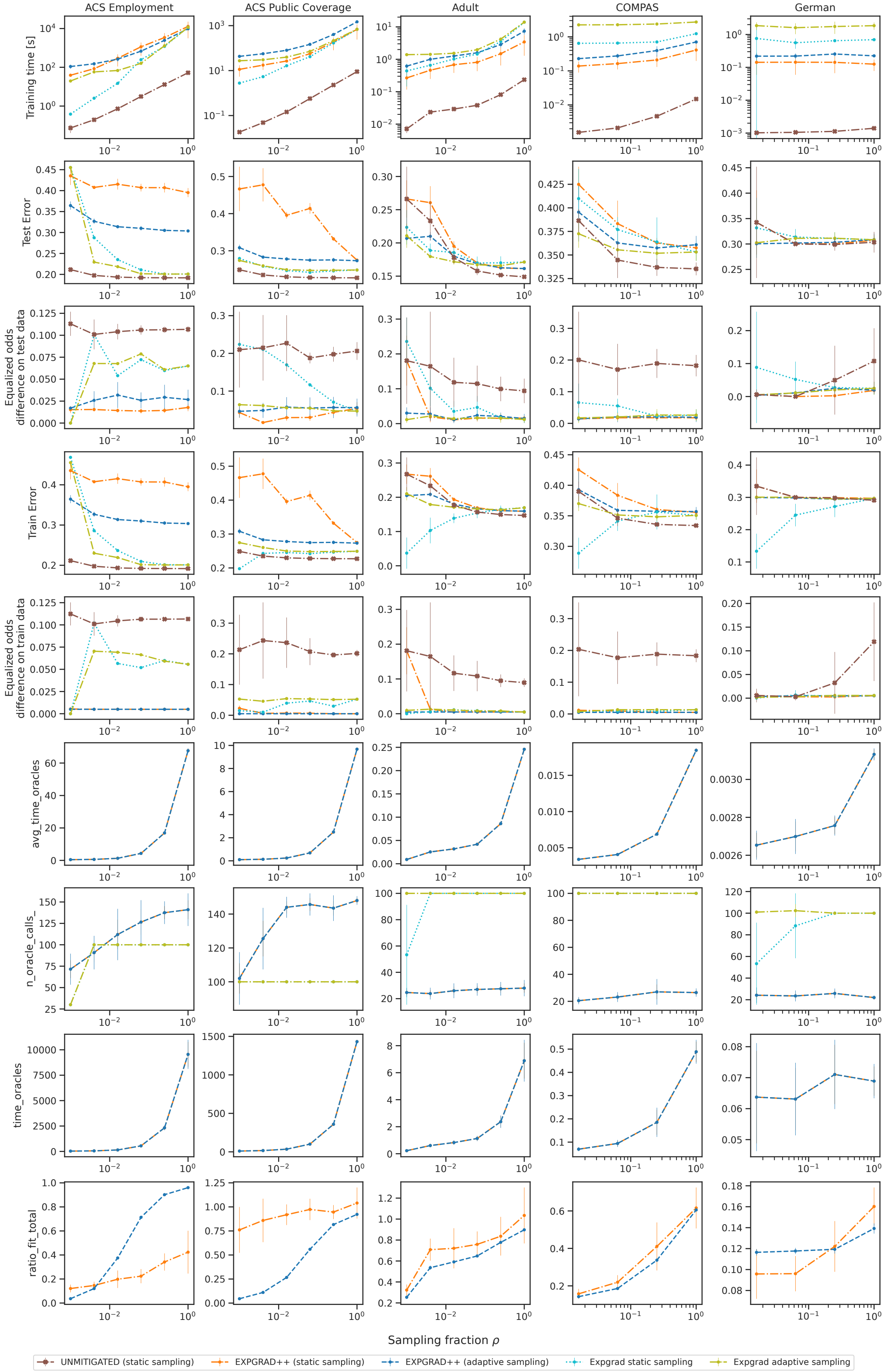


Figure 10: LR Effectiveness and Efficiency on the test sets varying the sampling size for Equalized Odds constraint.

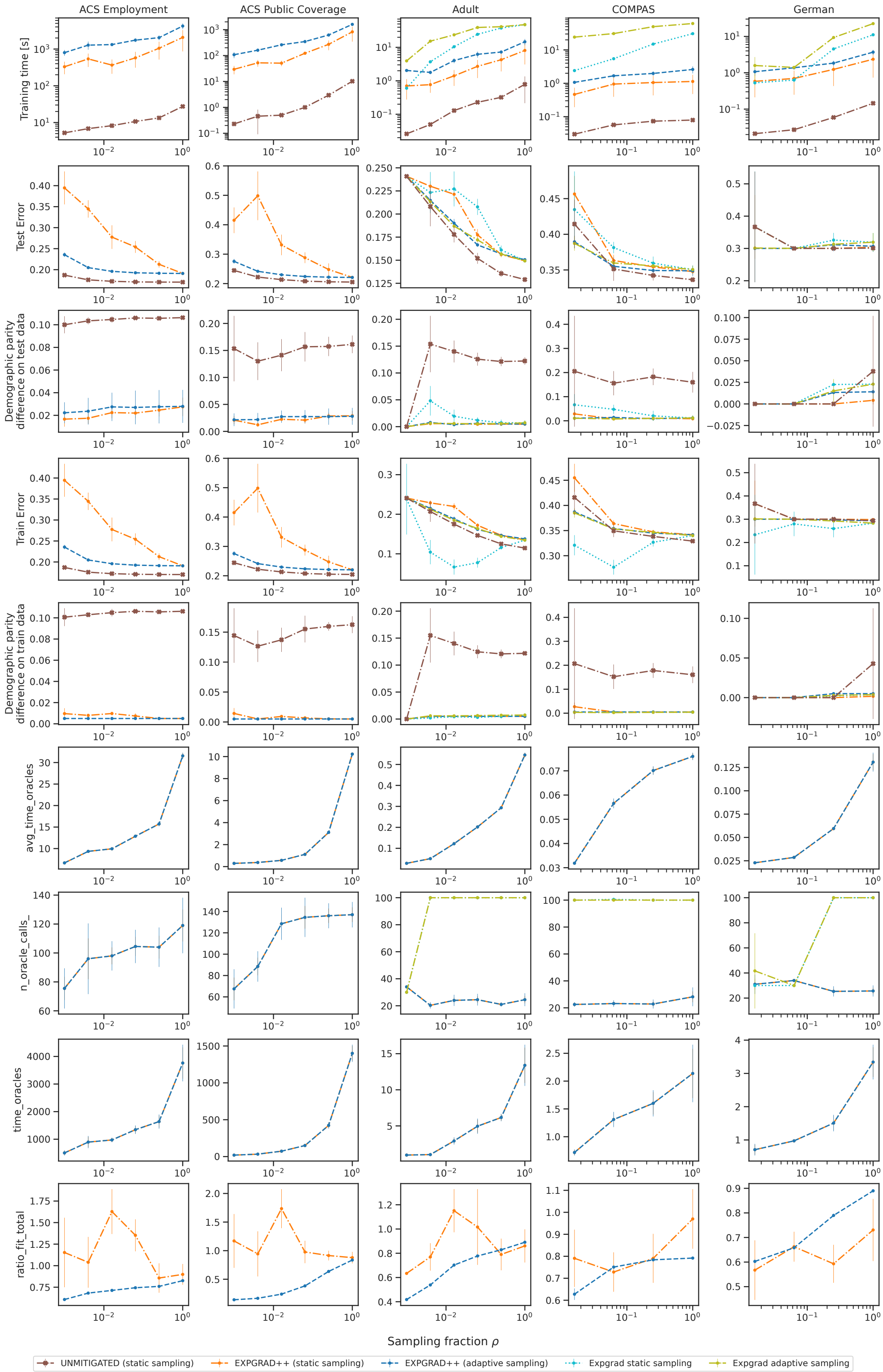


Figure 11: LGBM Effectiveness and Efficiency on the test sets varying the sampling size for Demographic Parity constraint.

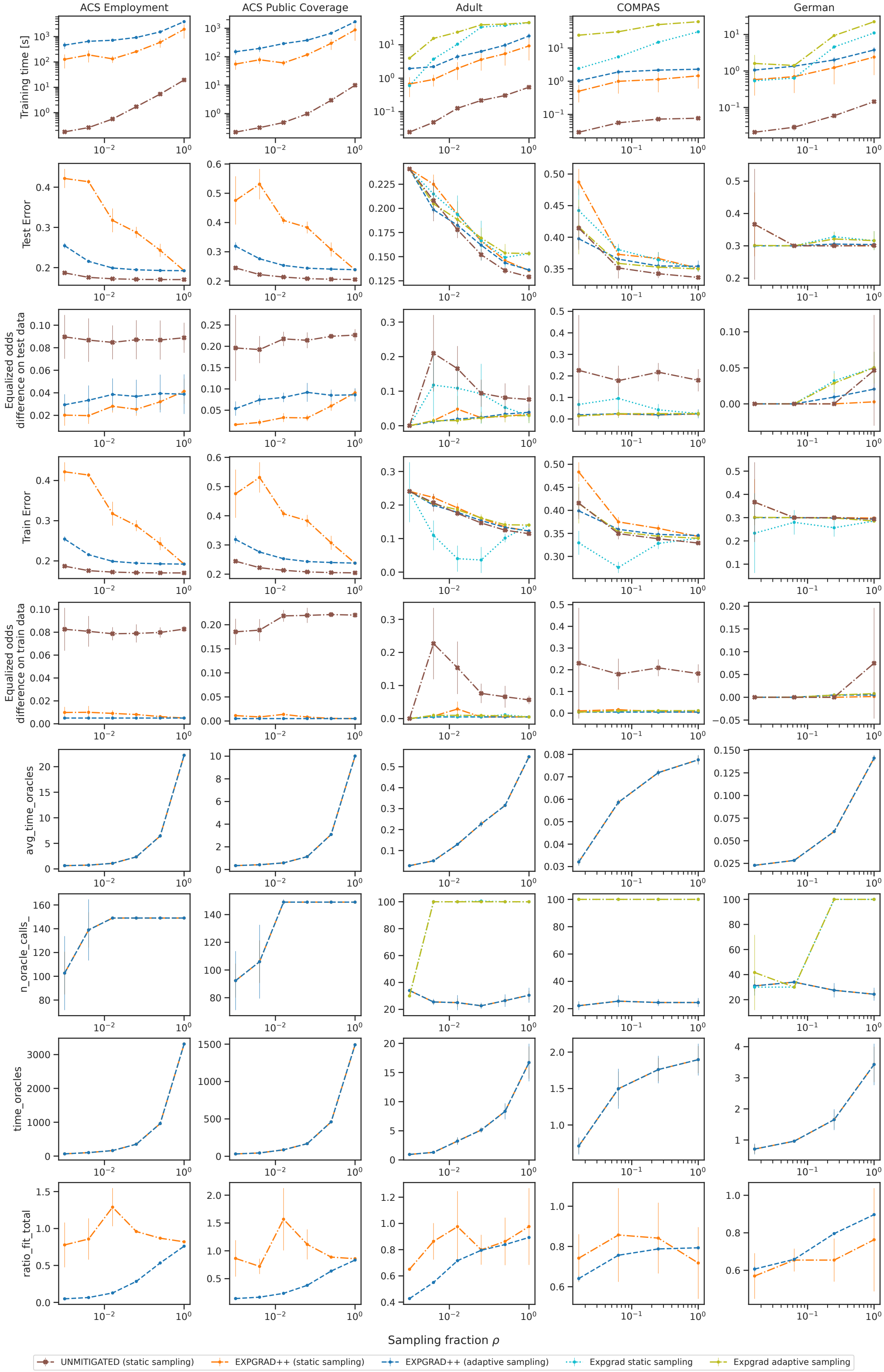
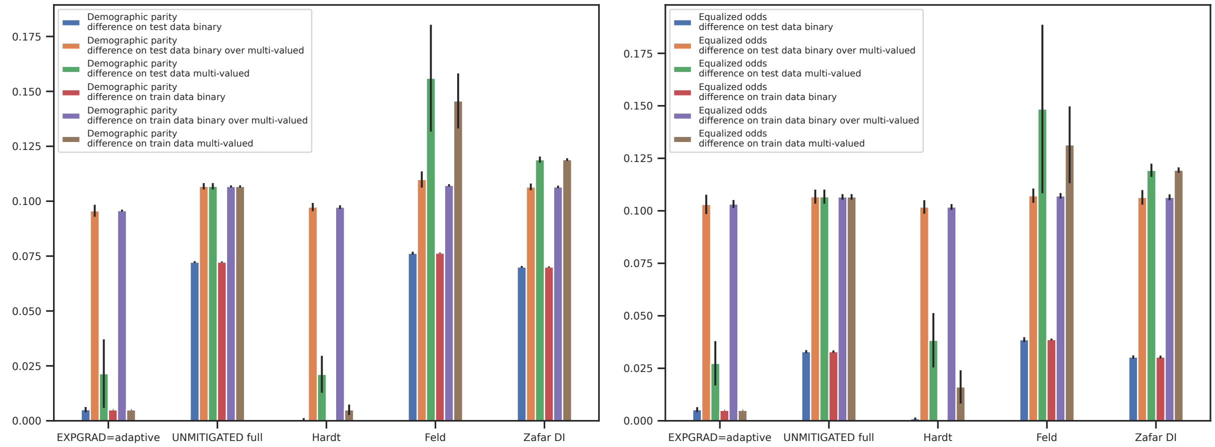


Figure 12: LGBM Effectiveness and Efficiency on the test sets varying the sampling size for Equalized Odds constraint.



(a) Demographic Parity.

(b) Equalized Odds.

Figure 13: Fairness constraint managing multi-valued as binary sensitive attributes.