The Impact of Bias and Fairness on Machine Learning and Large Language Models

Sienna Rega and Gianluca Demartini

METR4911

Introduction

Aim: Improve fairness of synthetically generated tabular data generated from existing tabular datasets via gradient descent optimisation.

Motivation: Increasing interest in generating tabular data for industries such as healthcare, finance and legal, due to concerns of bias propagation and prediction fairness.

RQ1: Can tabular data synthetically generated by LLMs using gradient-descent optimisation improve fairness in ML classification tasks without compromising performance?
RQ2: How do ML models trained with gradient descent optimised synthetic tabular data perform versus models trained with traditional tabular data?

Method

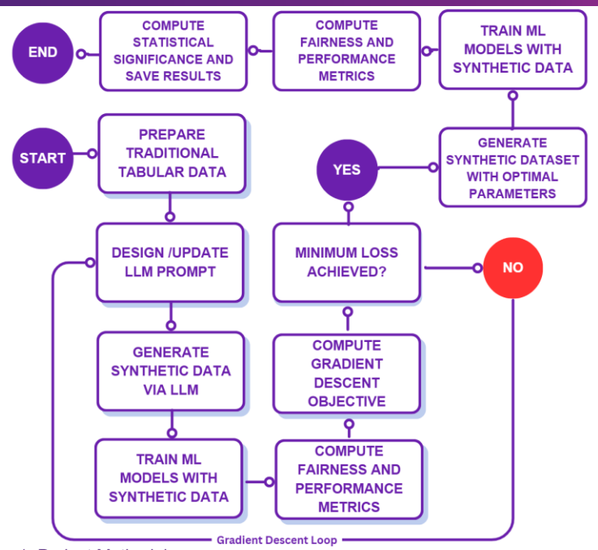

Figure 1: Project Methodology

Data:
- COMPAS – Sensitive group "African-American"
- Adult – Sensitive group "Black"

Metrics Investigated:
- **Fairness:** Statistical Parity (SP), Equality of Opportunity (EoO), Average Absolute Odds Difference (AAOD), Disparate Impact (DI)
- **Performance:** Accuracy (Acc.), F1, Area under Curve (AUC)

Design Parameters:
- LLM GPT- 4o mini with 3 few-shot examples per prompt
- Decision Trees (DT), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Extreme Gradient Boosting (XGB)
- 3 few-shot examples (FSE) from traditional tabular dataset
- Mini-Batch Gradient Descent with mini-batch size of 100
- Objective: $J = \alpha \times PerformanceScore + \beta \times (1 - FairnessScore)$
  - F1 used as consistent *PerformanceScore*
  - *FairnessScore* trialled as SP, AAOD, EoO and DI
  - Initial α and β as 0.5 and iteratively updated by 0.05 based on priority

Study 1: LR utilised in Gradient Descent loop for classification
Study 2: SVC utilised in Gradient Descent loop for classification

Results

	Acc.	F1	AUC	SP	EoO	AAOD	DI
DT	0.575366	0.351577	0.541364	0.116930	0.092033	0.106014	1.761421
LR	0.642956	0.505849	0.671826	0.204561	0.196826	0.181022	2.237841
RF	0.597691	0.400140	0.636414	0.107313	0.074093	0.090612	1.665530
SVM	0.642186	0.525100	0.636414	0.219762	0.213627	0.196086	2.166977
XGB	0.623711	0.465991	0.640585	0.183812	0.183628	0.165107	2.162361

Table 1: COMPAS Dataset Study 1 Statistically Significant Fairness Improvements with Maintained Performance Metrics For Trial on Fairness Metric DI

	Acc.	F1	AUC	SP	EoO	AAOD	DI
DT	0.583562	0.525438	0.583555	0.065364	0.074130	0.063448	1.143863
LR	0.588085	0.556294	0.616600	0.191352	0.154071	0.185967	0.817325
RF	0.618768	0.554529	0.630590	0.102843	0.116176	0.082520	1.071932
SVM	0.596425	0.445481	0.625075	0.064464	0.061325	0.053287	1.189794
XGB	0.610275	0.558164	0.645001	0.088518	0.102156	0.070717	1.074675

Table 2: COMPAS Dataset Study 2 Statistically Significant Fairness Improvements with Maintained Performance Metrics For Trial on Fairness Metric EoO

	Acc.	F1	AUC	SP	EoO	AAOD	DI
DT	0.578520	0.475198	0.572295	0.138048	0.158674	0.133302	0.606336
LR	0.543029	0.519939	0.568885	0.420232	0.374819	0.409414	0.176524
RF	0.596845	0.413797	0.653395	0.069906	0.091459	0.061578	0.670169
SVM	0.556019	0.206468	0.627937	0.004907	0.041907	0.026581	1.065098
XGB	0.595453	0.430148	0.638502	0.090168	0.109645	0.081679	0.630153

Table 3: Adult Dataset Study 1 Statistically Significant Fairness Improvements with Maintained Performance Metrics For Trial on Fairness Metric DI

	Acc.	F1	AUC	SP	EoO	AAOD	DI
DT	0.559267	0.567549	0.508357	0.159488	0.204191	0.169000	0.733300
LR	0.552308	0.209475	0.416213	0.022683	0.035128	0.022867	0.736421
RF	0.595917	0.545203	0.555874	0.183562	0.231901	0.177010	0.616930
SVM	0.551844	0.200995	0.465114	0.021201	0.034090	0.022062	0.734271
XGB	0.583623	0.554641	0.549947	0.132246	0.148221	0.127002	0.765137

Table 4: Adult Dataset Study 2 Statistically Significant Fairness Improvements with Maintained Performance Metrics For Trial on Fairness Metric EoO

	Acc.	F1	AUC	SP	EoO	AAOD	DI
DT	0.559267	0.512336	0.522752	0.090120	0.079476	0.086127	0.888787
LR	0.549988	0.321827	0.466588	0.163442	0.150843	0.158473	0.564730
RF	0.605196	0.513023	0.588964	0.157596	0.177029	0.146206	0.603887
SVM	0.551844	0.200177	0.542187	0.016059	0.031647	0.021792	0.847657
XGB	0.597773	0.520701	0.586146	0.137986	0.165206	0.139861	0.697443

Table 5: Adult Dataset Study 2 Statistically Significant Fairness Improvements with Maintained Performance Metrics For Trial on Fairness Metric DI

Note: In tables 1-5, = no statistically significant performance change, and = statistically significant fairness improvement.

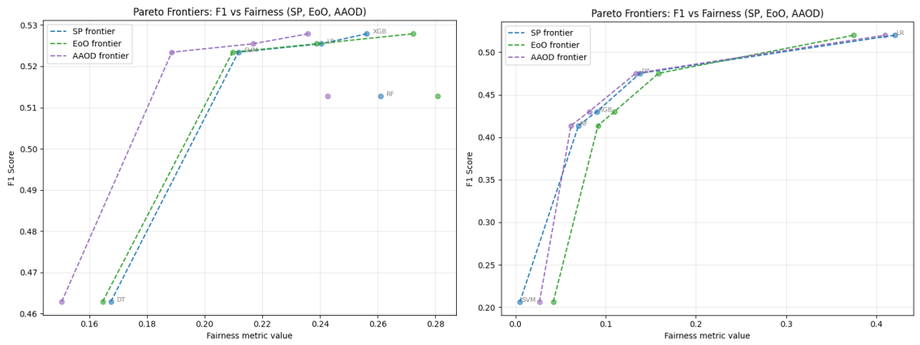

Figure 2: COMPAS SP Study 1 (Left) and Adult DI Study 1 (Right) Pareto Frontiers for F1 Score versus Fairness Metrics

Experiments				
	Study 1		Study 2	
	COMPAS with LR Training in GD	Adult with LR Training in GD	COMPAS with SVM Training in GD	Adult with SVM Training in GD
RQ1 Obs.	Fairness improvements do not coincide with maintained performance.	LR and DT show no significant changes in F1 score, with DI performance closer to 1.	EoO experiment shows improved fairness and maintained performance for DT, RF, SVM and XGB.	LR, DT, RF, XGB show stable F1 scores and improved DI, while SVM shows strongest fairness results across all trials, but does not maintain performance metrics.
RQ2 Obs.	Fairness improvement in SP using LR across most fairness metrics.	Improvement all fairness metrics (for SVM across SP and EoO trials, and for RF, SVM and XGB across AAOD and DI trials).	Improvement in fairness metrics SP, AAOD and DI in EoO experiment across DT, RF, SVM and XGB.	Improvement of SP, EoO and AAOD for SVM experiment, while DI improves across all models and training metrics.

Table 6: Experimental Observations in Terms of Research Questions

Conclusions

Insights and Observations:
- Adult Dataset provides more statistically significant results across RQ1 and RQ2
- SVM most consistent model for improving fairness
- DI most consistent metric with fairness improvement
- Overall trade-off between fairness and performance clear

Limitations:
- LLM constraints due to token windows
- Dataset sizes
- Limited trials and parameter changes

Future Work:
- Increased trials, sizes and datasets
- Alteration of differing GD parameters including
  - Objective function definition
  - Mini-batch size
  - Classification Model

Acknowledgements: Thank you to Prof. Gianluca Demartini for his continued support and contribution to the content and direction of this project. His expertise has been invaluable and has helped guide the projects success.

School of Electrical Engineering and Computer Science

s.rega@student.uq.edu.au
Sienna Rega
g.demartini@uq.edu.au
Gianluca Demartini