Not so Fair: Machine Learning and its Impacts

Mackenzie Jorgensen[1], Elizabeth Black[1], Natalia Criado[2], Jose Such[1,2]
[1]Dept. of Informatics, King's College London, London, UK
[2]Dept. of Computer Systems and Computing, Universidad Politècnica de València, Valencia, Spain

BACKGROUND

- Problem: bias in ML models
- Fairness != Positive Impacts
- Lack of consensus of robust and effective mitigation methods

QUESTION

- How do fairness interventions and ML models affect impact on protected groups?

DATASET

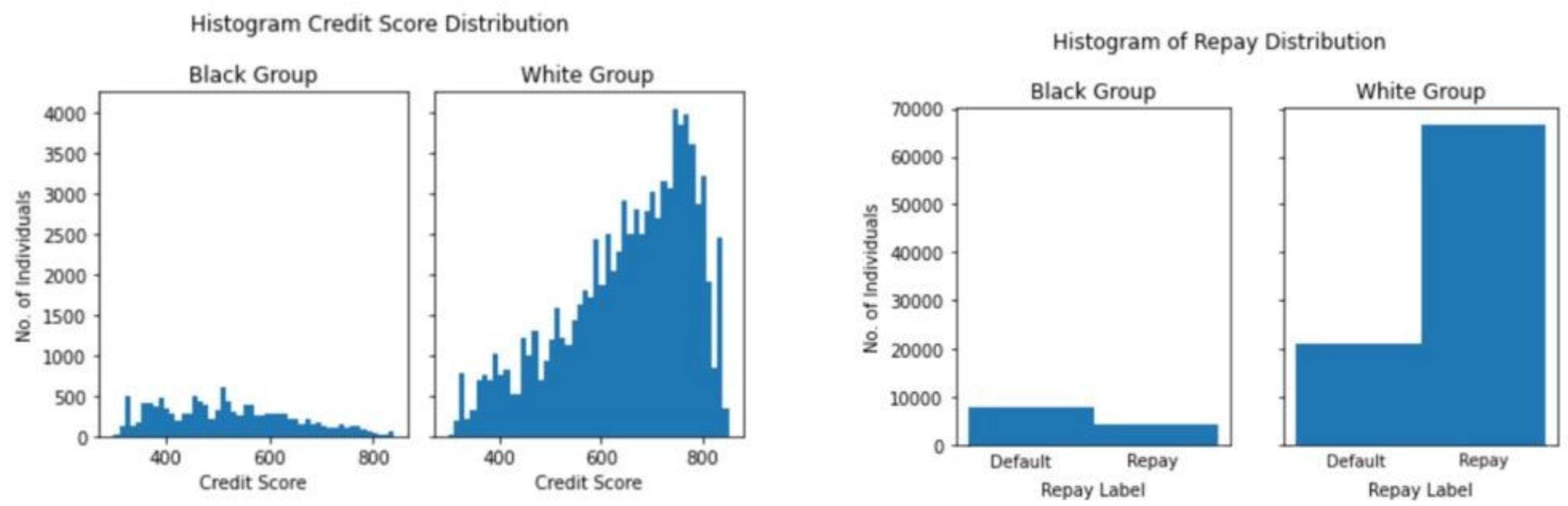
Domain: Loan Granting

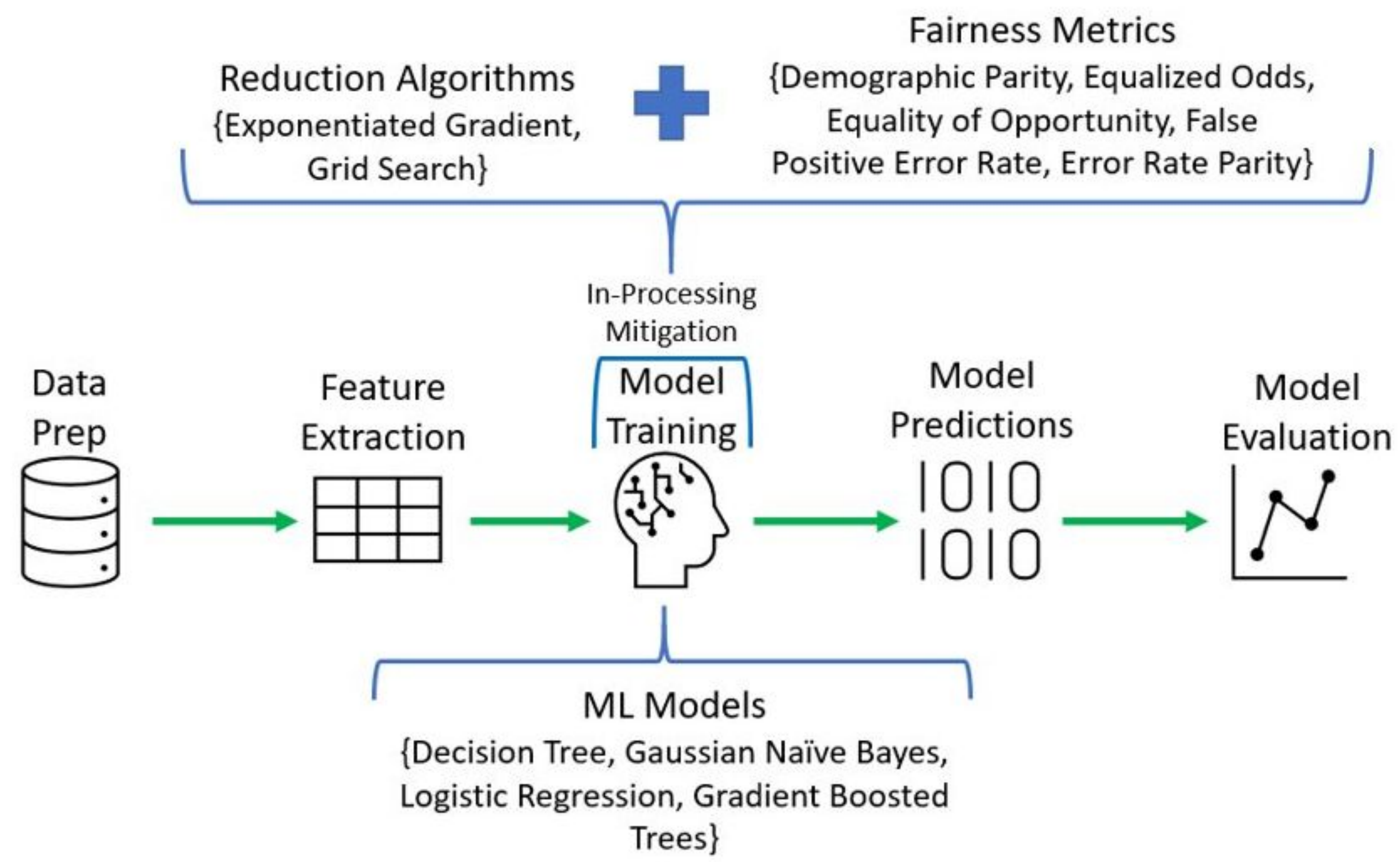
- Simulated 100k applicant dataset from 301,536 TransUnion TransRisk scores from 2003, originally used in Hardt et al. 2016
- Sensitive feature: race (black or white)
- Labels: 0 (default) or 1 (repay)

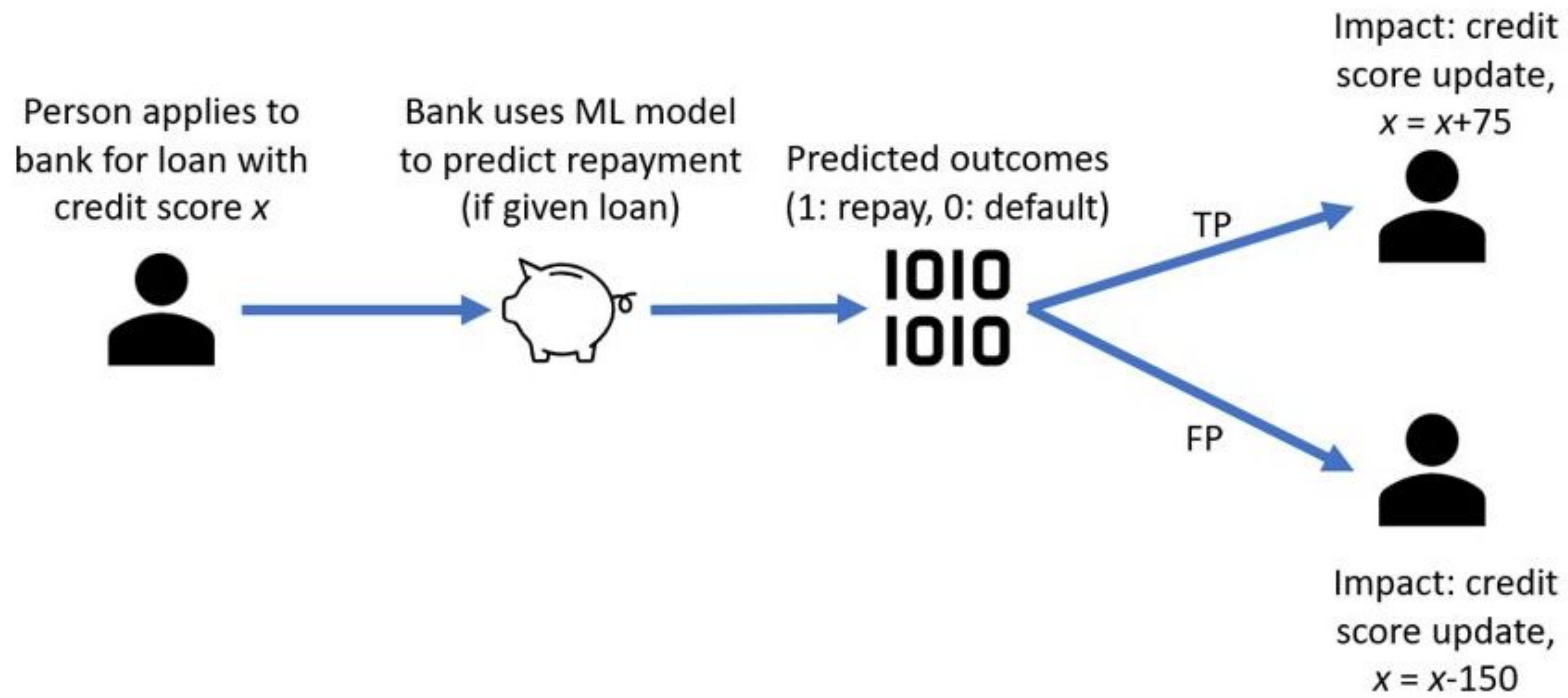
IMPACT

- Impact: the effect of a model prediction after its been made
- Variable affected: credit score
- Average change in credit score measured by group
- True Positive outcome: +75 points
- False Positive outcome: -150 points

METHODS







RESULTS

- Most "fair" predictions fail to improve impact for the disadvantaged group
- Improvement (credit score increase) for the disadvantaged group, when achieved, is quite modest

CONCLUSIONS

- "Fair" predictions can result in worse impacts for advantaged and disadvantaged groups
- Impact is more sensitive to the fairness intervention method than to ML model choice

FUTURE WORK

- Synthetic dataset variations for different scenarios
- Statistical significance testing
- Variations of impact functions

ACKNOWLEDGEMENTS

- The PhD studentship for Mackenzie Jorgensen is funded by the EPSRC through the EP/S023356/1 grant
- This work was presented at the AIofAI: 2nd Workshop on Adverse Impacts and Collateral Effects of AI Technologies at IJCAI '22