THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

SOLASAI

Benchmark Machine Learning Model Development for Disparity Mitigation Methodologies

Abdulrahman Alsadun, Jenny Yazlovsky, Elena Huang, Jinni Yang, Chao Zhang, and Maximilian Smith-Uchida

TABLE OF CONTENTS

01 EXECUTIVE SUMMARY

02 PROBLEM UNDERSTANDING

03 DATA ANALYZED

04 XGBOOST

05 LOGISTIC REGRESSION

06 NEURAL NETWORK

07 EXPLAINABLE BOOST MACHINE

08 RESULTS

09 NEXT STEPS

Executive Summary

- SolasAI is a software company that helps customers deal with business issues and reduce bias in predictive decision models.
- The issue: There is lack of a benchmark model to validate bias mitigation methodologies. The project aims to come up with benchmark models to measure disparity and reduce bias in machine learning algorithms.
- Models used: XGBoost, logistic regression, explainable boosting machine, and neural networks.
- The expectation: To have a series of models that they can use in the future as a benchmark for novel bias mitigation methods.
- Result interpretation: Model performance and fairness evaluation

Problem Understanding

- Central Issue: Increase the number of benchmark models SolasAI has access to.
- Desired Business Outcomes: Build multiple different algorithm types that act as a baseline which can measure disparity: XGBoost, logistic regression, explainable boosting machine, and neural networks models.
- Client expectations: Different types of models that can be used repeatedly and function as an internal self-check when testing new bias mitigation techniques.

Risk of Disparity

- How do we consider disparity?
  - Disparity in a model refers to the systematic error of a model to consistently underestimate or overestimate the true values of a target variable. Some variables will affect the model fit with the model.
- The effect of bias
  - Results in unfair outcomes, which perpetuate inequities in our society.
  - Models can become inflexible and cause it to miss essential data or patterns, which leads to inaccurate predictions.
  - Can cause the model to be incorrect, which can't indicate the relationship between the independent variables and dependent variables.
- How to fix disparity
  - We should be careful with each variable, regularly check the model's performance and consider the appropriate features or predictors, increase model complexity, or use different modeling techniques.

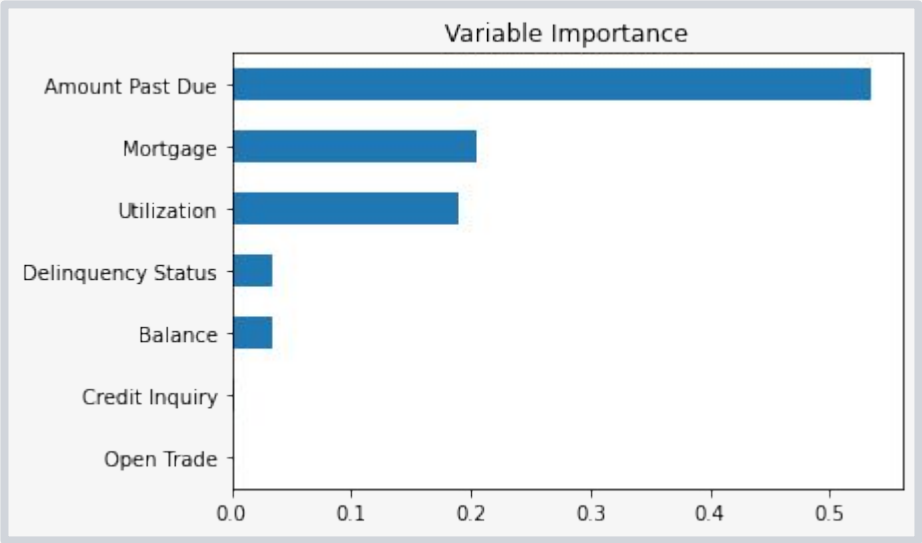
Data Analyzed

Dependent Variable	Independent Variables	Baseline	Demographics
Being denied for a loan.	• Mortgage • Balance • Amount Past Due • Delinquency Status • Credit Inquiry • Open Trade • Utilization	Baseline predictions that we used to compare our models against.	We extracted these from the data in order to use them for bias testing: • Race • Age • Gender

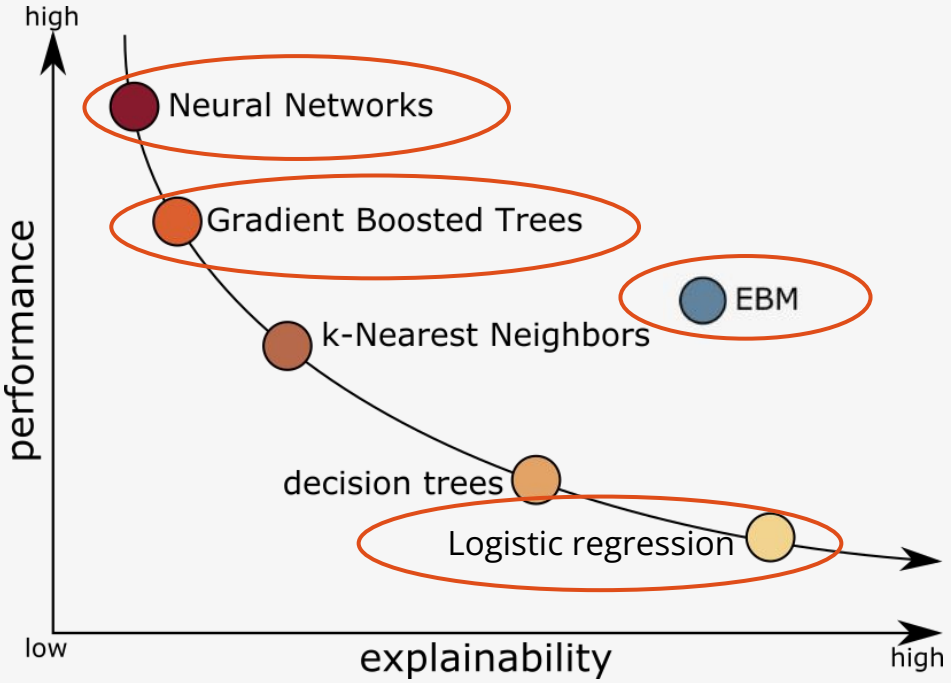
Note: Data is synthetic. No real customer or client data was used.

Exploratory Data Analysis

- Conducted **Decision Tree Analysis** and ranked variables based on its importance
- Surprisingly, the variable importance doesn't match correlations between target and predictor variables
  - EX) Delinquency Status Correlation = -0.296 while Amount Past Due Correlation = -0.188
- For Decision Tree Analysis
  - Test AUC of 0.73, which is lower than the baseline Test AUC of 0.75



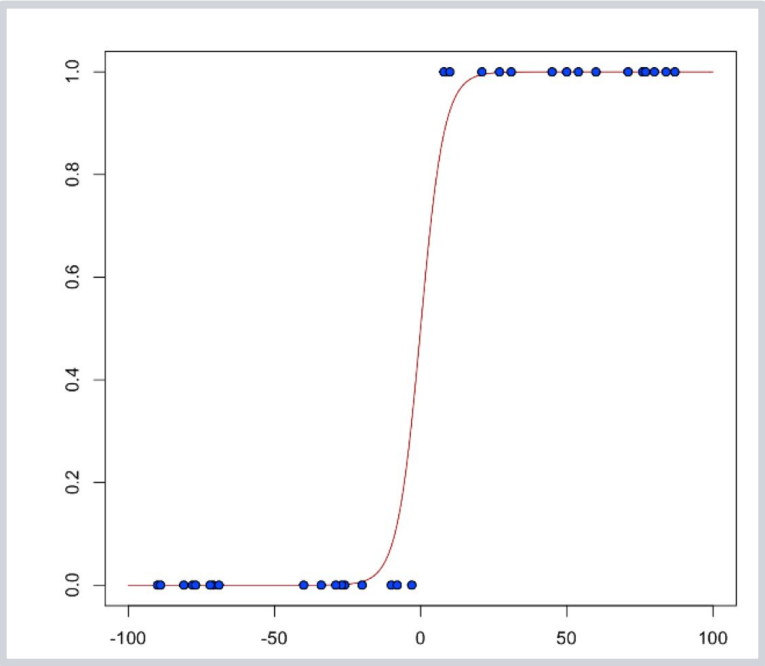
Key Model Types



https://blog.oakbits.com/ebm-algorithm.html

Logistic Regression

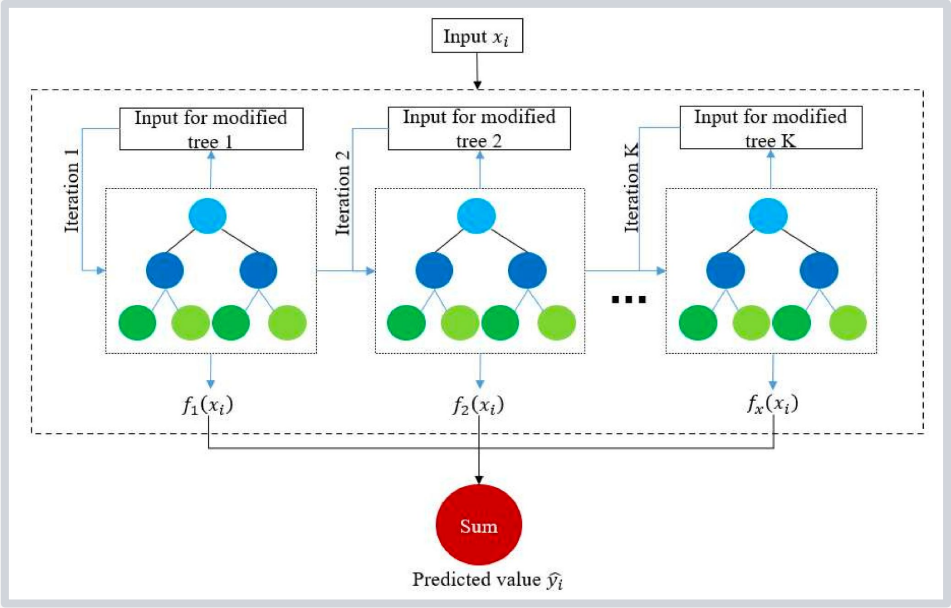
- Used to analyze the relationships between a binary or categorical dependent variable and one or more independent variables.
- In binary classification, the output of the model is binary, while in probability estimation, the output is a continuous value between 0 and 1 representing the probability of the binary event occurring.
- The goal of logistic regression is to model the probability of the binary outcome as a function of one or more independent variables.



https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/glm.html#logistic-regression-binomial-family

XGBoost

- Optimized distributed gradient boosting library designed to efficiently implement gradient boosting algorithms.
- Open-source machine learning library that uses a gradient boosting framework to implement a decision tree ensemble algorithm.
- Combines several weak models to create a stronger model. Each weak model tries to correct the errors made by the previous model, which improves the accuracy of the overall model.



https://www.mdpi.com/1996-1944/15/15/5298

Explainable Boost Machine (EBM)

- EBM is a tree based model where each feature (i.e., Mortgage, Balance, Amount Past Due) is trained by itself with 5,000 - 10,000 iterations.
- For each feature, the iterations are combined to show the effect of each feature on the outcome variable.
- EBMs make it easy to show the contribution of each feature to the final prediction.

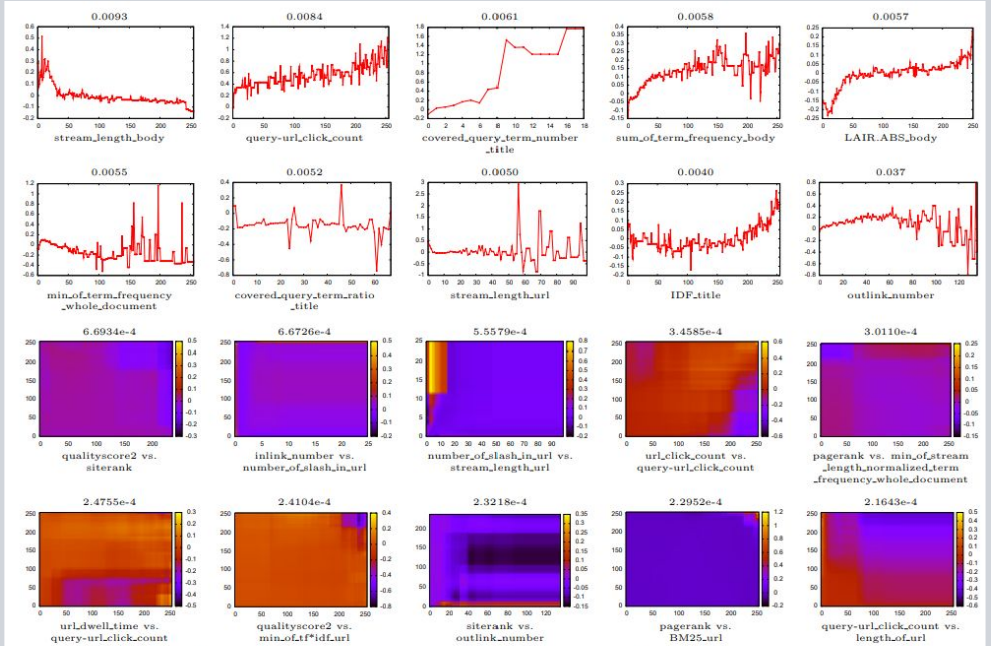
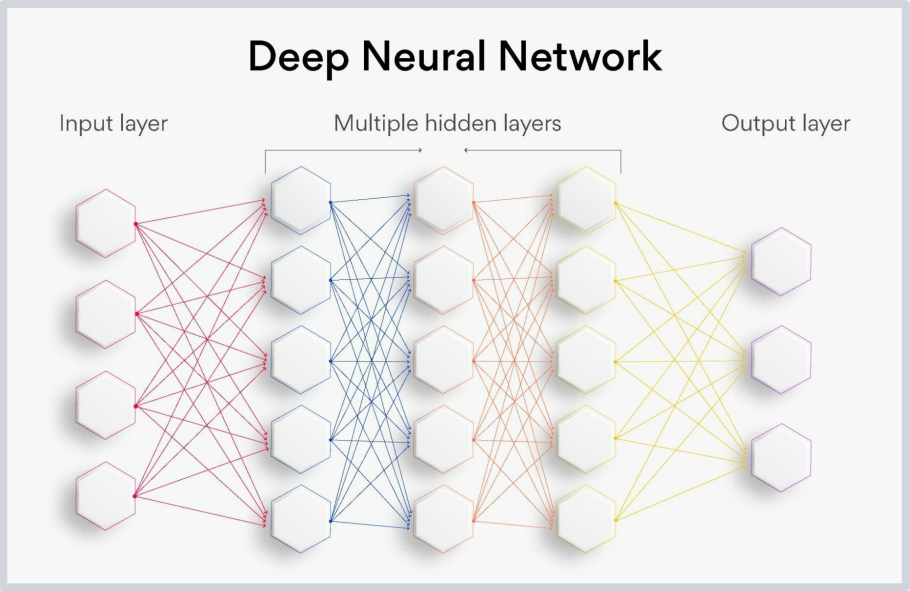


Figure 9: Shapes of features and pairwise interactions for the "MSLR10k" dataset with weights. Top two rows show top 10 strongest features. Next two rows show top 10 strongest interactions.

(GA2M, an early version of EBM) https://www.cs.cornell.edu/~yinlou/papers/lou-kdd13.pdf

Neural Networks

- Interconnected neurons that process and transmit information, inspired by the way the human brain works.
- Through a process called training, the network can learn to make accurate predictions or classifications based on input data.
- Useful for tasks involving complex and non-linear relationships between inputs and outputs.
- Neural networks can learn from raw data, which reduces the need for manual feature engineering.



https://www.turing.com/kb/mathematical-formulation-of-feed-forward-neural-network

Performance Metrics

AUC	RMSE	Accuracy
• Measures the area under the ROC curve, (i.e. the "AUROC") • 0.5: Random • 1.0: Perfect	• Measures the root mean square of the differences between the predicted values and the actual values	• Calculated by dividing correct predictions by total predicted number • 0.5: Random • 1.0: Perfect

Model Performance Results

	LR	XGB	EBM	NN
AUC	0.7276	0.7454	0.7505	0.7278
RMSE	0.4588	0.4512	0.4493	0.4646
Accuracy	0.6650	0.6784	0.6815	0.6672

Fairness Evaluation Metrics

Adverse Impact Ratio (AIR)	By-quantile AIR	Standardized Mean Difference
• Sets a threshold to test whether the proportion a of certain protected group is being underrepresented in loan selection. • Protected group categories: Race, Age, Gender	• Compares the acceptance rate of a certain protected group to that of a reference group across different score quantiles or ranges. • Protected group categories: Race, Age, Gender	• The difference in means between a certain protected group and a reference group. • Protected group categories: Race, Age, Gender

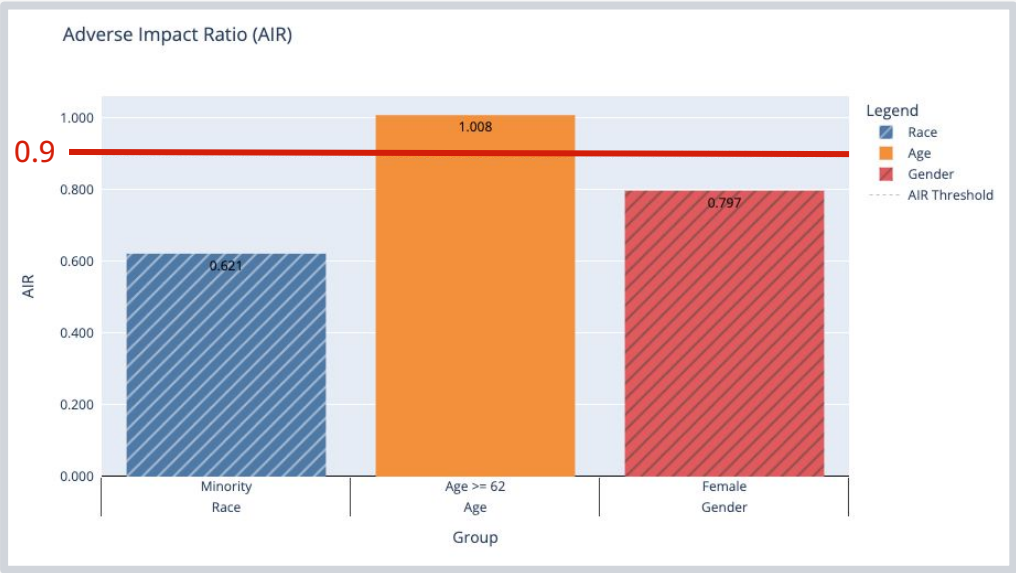
Note: the output used is from EBM using solas-ai 0.2.4 package

Fairness Evaluation Metrics cont.

Disparate Impact	Statistical Parity Difference
• Compares the proportion of being selected for a loan of the protected group to the majority group. • Protected group categories: Race, Age, Gender	• Compares the percentage of favorable outcomes for different demographic groups, such as race or gender. • Protected group categories: Race, Age, Gender

Note: IBM 360 methods were used to calculate these outputs

Fairness Evaluation Results: Adverse Impact Ratio (AIR)



- This shows that there is adverse disparities against protected race and gender groups. They are being selected for loans at a lower rate than the majority group of applicants.
- We used a practical significance threshold of 0.9.
- Gender and Race had AIR less than 0.9.
- Bias mitigation techniques are necessary to debias the model.

Note: the output used is from EBM using solas-ai 0.2.4 package

Fairness Evaluation Results: By-Quantile AIR

- Tested for AIR at each decile of the continuous model score
- Confirms results of AIR calculation
- We used a by-quantile practical significance threshold of 0.9
- Gender and Race had AIR less than 0.9

Note: the output used is from EBM using solas-ai 0.2.4 package

Fairness Evaluation Results: Standardized Mean Difference

- This shows that there is discrimination against protected race and gender groups
- We used a practical significance threshold of 30 (30th percentile)
- Gender and Race both were above the threshold

Note: the output used is from EBM using solas-ai 0.2.4 package

Fairness Evaluation Results: Minority Groups

	Disparate Impact	Statistical Parity Difference
Explainable Boost Machine	0.621	0.248

With a practical significance threshold of 0.9, this shows bias mitigation techniques need to be used on the model.

>24% differences in outcome probabilities typically require bias mitigation.

Note: IBM 360 methods were used to calculate these outputs

Fairness Evaluation Results: Elderly Groups

	Disparate Impact	Statistical Parity Difference
EBM	1.007	-0.004

With a threshold of 0.9, this shows that elderly groups are not discriminated against in the model.

Note: IBM 360 methods were used to calculate these outputs

Fairness Evaluation Results: Gender

	Disparate Impact	Statistical Parity Difference
EBM	**0.796**	**0.131**

With a threshold of 0.9, this shows bias mitigation techniques need to be used on the model.

>13% differences in outcome probabilities typically require bias mitigation.

Note: IBM 360 methods were used to calculate these outputs

Next Steps

BIAS MITIGATION

- Apply commonly used bias mitigation tools to benchmark models to evaluate effectiveness.
- Apply model development pipelines to real data.
- Regularly review the models performance and retrain the models as needed.

Appendix

- Models Colab link: SAI_Models
- Fairness evaluation Colab link: SAI_Fairness

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Thank You!

Abdulrahman Alsadun, Jenny Yazlovsky, Elena Huang, Jinni Yang, Chao Zhang, and Maximilian Smith-Uchida