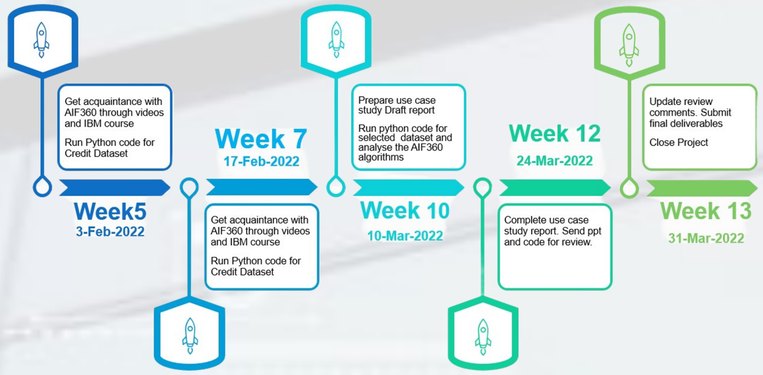
educate | explore | engage
MILGARD SCHOOL OF BUSINESS
UNIVERSITY of WASHINGTON TACOMA

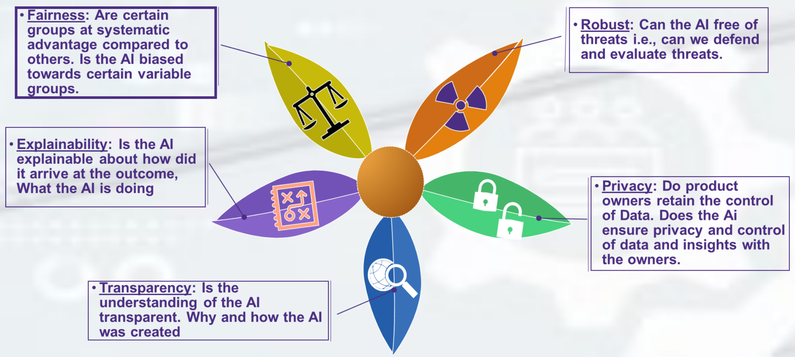
AI FAIRNESS - AIF360 TOOLKIT

UNIVERSITY of WASHINGTON TACOMA

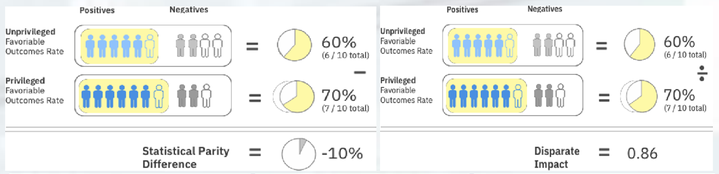
Project Timeline



AI Trust



AI 360 Toolkit-Metrics/Algorithms



Pre-Processing Algorithms	In-Processing Algorithms	Post-Processing Algorithms
Reweighing	Adversarial Debiasing	Reject Option Classification
Disparate Impact Remover	Prejudice Remover	Calibrated Equalized Odds
Optimized Pre-processing	Meta Fair Classifier	Equalized Odds
Learning Fair Representations		

AI 360 Toolkit Guidelines

Work with your stakeholders early to define fairness, protected attributes & thresholds → Apply the earliest mitigation in the pipeline that you have permission to apply → Check for bias as often as possible using any metrics that are applicable

Caveat: AIF360 should only be used with well defined data sets & well defined use cases

PROJECT SCOPE

The scope is to understand AI trust, AIF 360 toolkit, and the metrics and algorithms available with the AIF360 toolkit to identify and mitigate bias in datasets. The objective is to run AIF360 metrics and algorithms to identify and mitigate bias in datasets and to create a use case presentation.

PROJECT TIMELINE

We completed the project in 12 weeks. The major hurdle being the exposure to python which was overcome with the help of python courses offered by Dr. Sergio from the CBA team.

AI TRUST

AI is slowly becoming an integral part of any industry as all the industries are extensively relying on AI for decisions. So we need to ensure that any responsible AI system has to be trustworthy. Five aspects make the AI trustworthy which are 1. Fairness, 2. Robustness, 3. Explainability, 4. Transparency and 5. Privacy. In this project we dealt with AI Fairness as such will be discussing only Fairness.

AI FAIRNESS

Bias can occur at different stages of the AI decision making environment—Through Data, Through algorithms trained on biased datasets & Through decision making. A biased AI/ML decision-making system might not always fall under legal boundaries, but it might fall under ethical boundaries which may lead to mistrust in the organization. Fairness has over 21 definitions which makes it very difficult to understand it, as different definitions might have different outcomes.

AIF360 TOOLKIT

The AIF360 is a comprehensive toolkit consisting of over 70 fairness metrics and over 10 bias mitigation algorithms. The toolkit does many jobs from detecting bias, and understanding bias to mitigating bias through various algorithms. Also, the toolkit is open source. AIF360 bias reduction algorithms can be applied at varied states of the ML system, namely pre-process (before the model is run on Dataset), in-process (while running Model-on classifier), and post-process (after running the model on results).

AIF360 TOOLKIT GUIDELINES

First, identify the protected and privileged attributes based on the fairness definitions applicable to the project. Second, test for bias through available metrics and apply respective bias mitigation algorithms. Finally check the bias metrics after applying the mitigation algorithm. A general rule of thumb is that the AIF360 algorithms should only be applied to well-defined data and use case. A little understanding of bias is required for success of the AIF360 algorithms.

PROJECT RESULTS & RECOMMENDATIONS

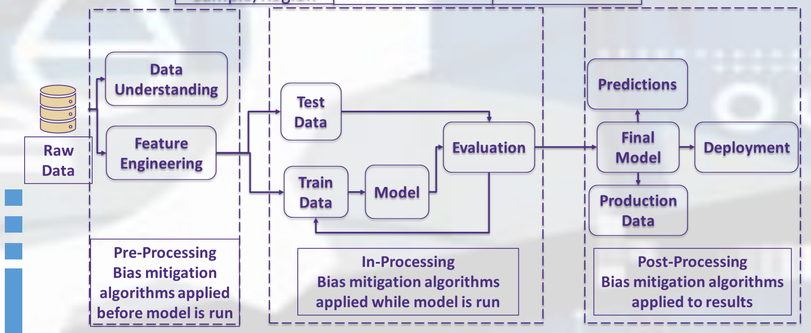
AIF360 algorithms & metrics were run on two datasets, the banking churn, and heart disease dataset. The results showed that the bias (disparate impact values) was reduced without effecting the accuracy of the model much.

Any industry using a decision-making system can leverage the AIF360 toolkit capabilities to ensure the decision made by the system are not biased. The trade -off is that the dataset should be well-defined and clean and the business case should be well-defined to identify protected/privileged attributes to identify and mitigate bias.

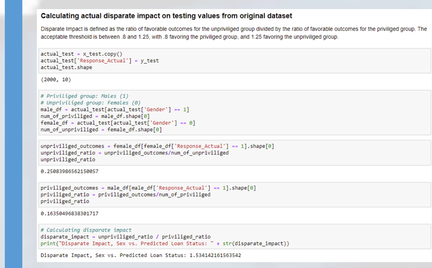
Project Team

AI Fairness

DATA BIAS	AI/ML BIAS	HUMAN BIAS
Data Collection Historical Data Imbalanced Data Sample/Region	Algorithm Aggregation Evaluation	Social Behavioural Unconscious



Python Notebook & Results



BANK CHURN DATASET

ENTITY	ORIGINAL	TRANSFORMED
DISPARATE IMPACT	1.53	1.27
ACCURACY - LOGISTIC	0.789	0.789

HEART DISEASE DATASET

ENTITY	ORIGINAL	TRANSFORMED
DISPARATE IMPACT	1.81	1.07
ACCURACY - LOGISTIC	0.859	0.878

CENTER FOR BUSINESS ANALYTICS
UNIVERSITY of WASHINGTON | TACOMA

OpenDS4All