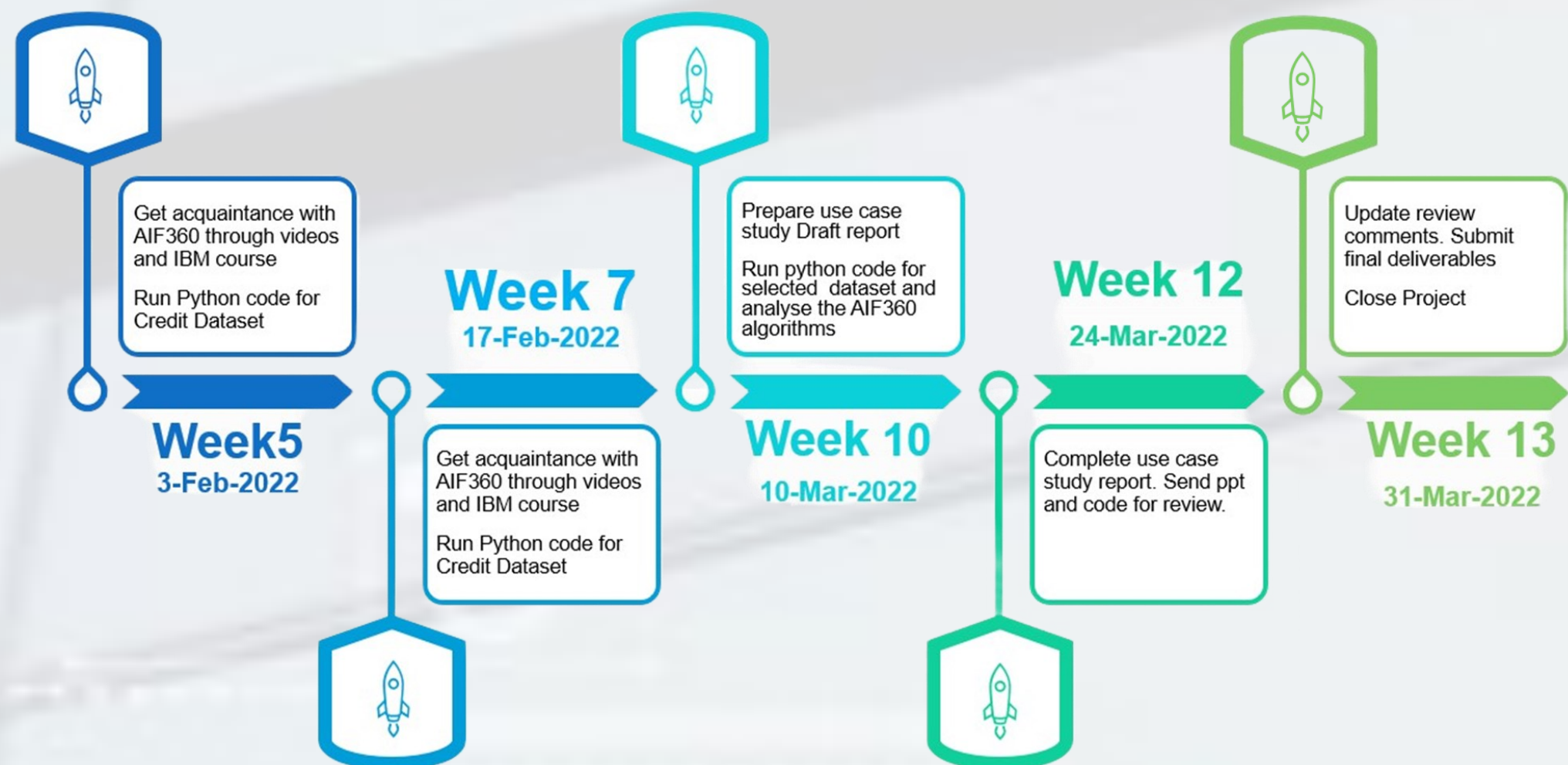
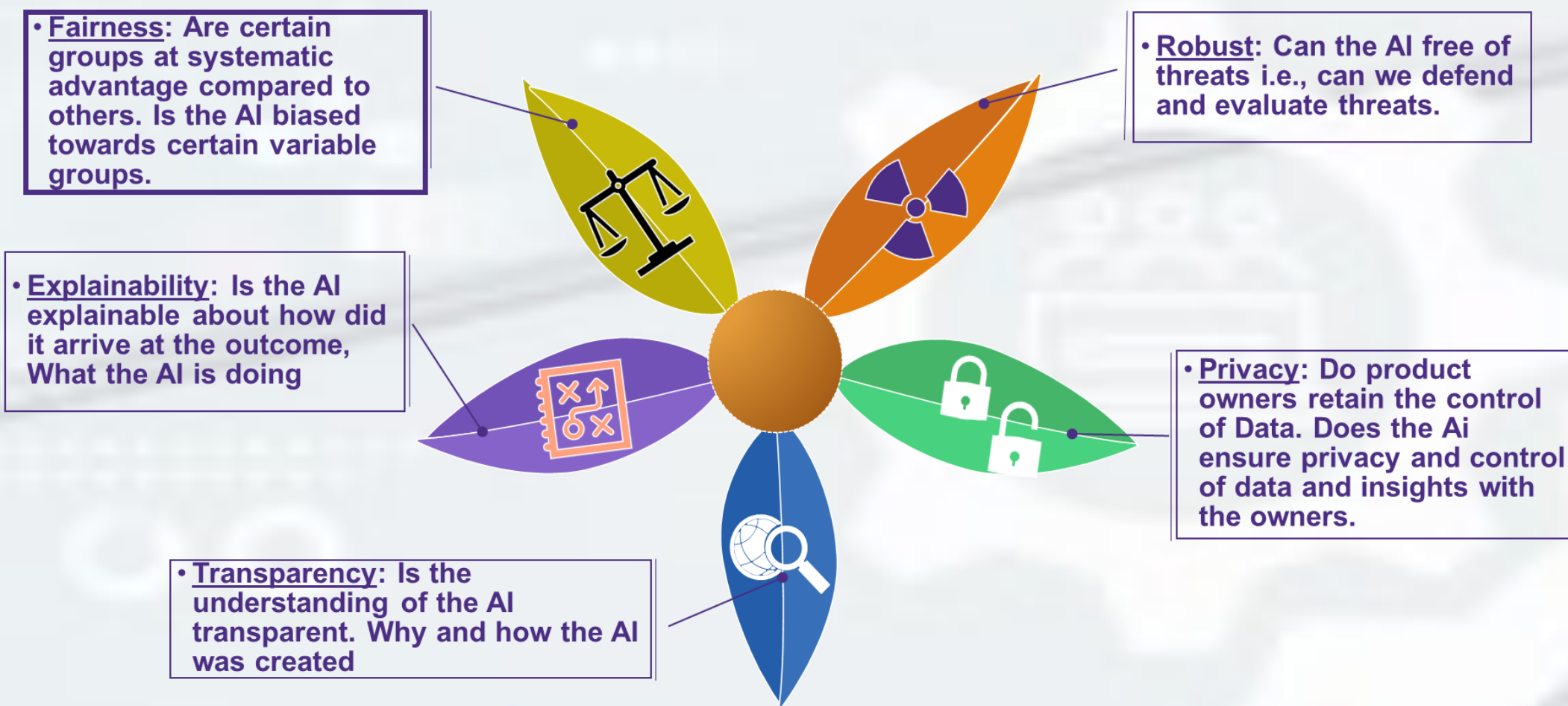




Project Timeline



AI Trust



AI FAIRNESS - AIF360 TOOLKIT

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

Teja Alluru

Bio
 9 years of experience in Aerospace and 2 years in data science. Experienced in building classification and prediction models and well-versed in the art of story telling through visualizations. I passionate to advance myself through gaining knowledge on new and advanced technologies in AI/ML domain. I believe that great ideas emerge from being empathetic, enthusiastic and by creating psychologically safe environments when communicating.



Monikuntala Saikia

Bio
 Seasoned financial service professional with 4 YOE providing data-driven solutions to clients in rural banking and making a difference. Here I am in Data Science sphere, challenging myself and exploring new learning experiences every day. Passionate about learning and growing, and enthusiastic about using BI tools. I want my work to create an impact. The good one at that!



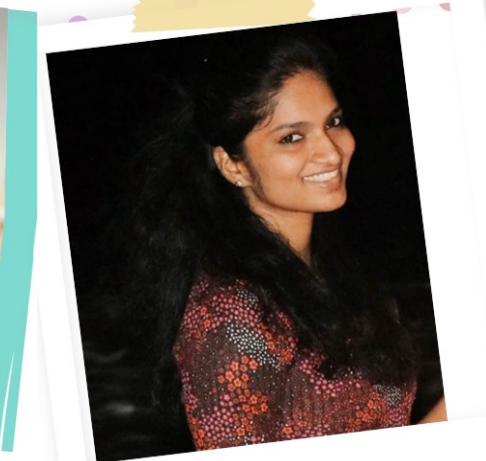
Shephali Jain

Bio
 I am a curious learner and enthusiast passionate towards Data Science. I carry 5 years of experience in data world focused on Business Intelligence and Analytics with a go-getter spirit solving problems and collaborating with high emotional intelligence leveraging people skills. Creating a long-lasting impact is something I chase for.



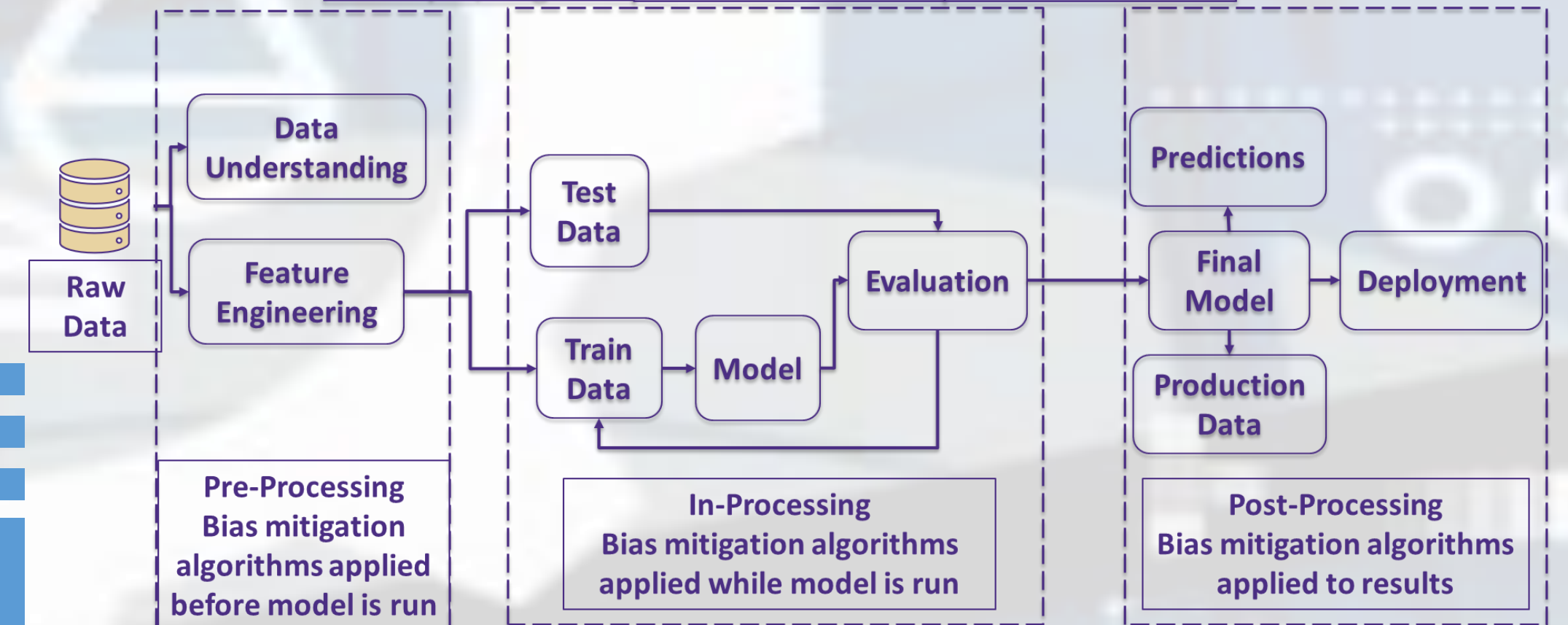
Nithisha Katasani

Bio
 I'm a data interpreter who has always been intrigued by Analytics and how data keeps playing a significant role in taking any management decisions which are merely based on data-based assessments. Exploring my career while pursuing my master's degree in business analytics with focus towards Data Science.



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

```

Calculating actual disparate impact on testing values from original dataset
Disparate impact is defined as the ratio of favorable outcomes for the unprivileged group divided by the ratio of favorable outcomes for the privileged group. The acceptable threshold is between .8 and 1.25, with .8 favoring the privileged group, and 1.25 favoring the unprivileged group.

actual_test = A_test.copy()
actual_test['Response_Actual'] = y_test
actual_test.shape

(2000, 10)

# Privileged group: Males (2)
# Unprivileged group: Females (0)
male_df = actual_test[actual_test['Gender'] == 1]
num_of_privileged = male_df.shape[0]
female_df = actual_test[actual_test['Gender'] == 0]
num_of_unprivileged = female_df.shape[0]

unprivileged_outcomes = female_df[female_df['Response_Actual'] == 1].shape[0]
unprivileged_ratio = unprivileged_outcomes/num_of_unprivileged
unprivileged_ratio

0.23609552108877

privileged_outcomes = male_df[male_df['Response_Actual'] == 1].shape[0]
privileged_ratio = privileged_outcomes/num_of_privileged
privileged_ratio

0.35180483381717

# Calculating disparate impact
disparate_impact = unprivileged_ratio / privileged_ratio
print("Disparate Impact, Sex vs. Predicted Loan Status: " + str(disparate_impact))
Disparate Impact, Sex vs. Predicted Loan Status: 1.534121261561542

Applying the Disparate Impact Remover to the dataset

import aif360
from aif360.algorithms.preprocessing import DisparateImpactRemover
# binaryLabelDataset = aif360.datasets.BinaryLabelDataset(
# df=yourDataFrameHere,
# label_names=['yourOutcomeLabelHere'],
# protected_attribute_names=['yourProtectedClassHere'])
# Must be a binaryLabelDataset
binaryLabelDataset = aif360.datasets.BinaryLabelDataset(
favorable_label=1,
unfavorable_label=0,
df=df,
label_names=['Exited'],
protected_attribute_names=['Gender'])
di = DisparateImpactRemover(repair_level=1.0)
dataset_transform_train = di.fit_transform(binaryLabelDataset)
transformed = dataset_transform_train.convert_to_dataframe()[0]
transformed.head()

# Privileged group: Males (1)
# Unprivileged group: Females (0)
male_df = transformed_output[transformed_output['Gender'] == 1]
num_of_privileged = male_df.shape[0]
female_df = transformed_output[transformed_output['Gender'] == 0]
num_of_unprivileged = female_df.shape[0]

unprivileged_outcomes = female_df[female_df['Exit_Status_Predicted'] == 1].shape[0]
unprivileged_ratio = unprivileged_outcomes/num_of_unprivileged
unprivileged_ratio

0.03191972008335946

privileged_outcomes = male_df[male_df['Exit_Status_Predicted'] == 1].shape[0]
privileged_ratio = privileged_outcomes/num_of_privileged
privileged_ratio

0.03071364046973883

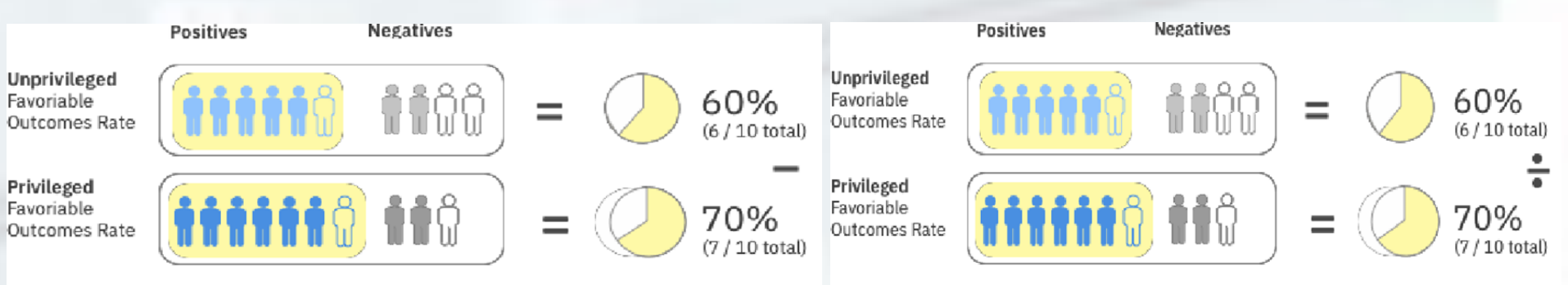
# Calculating disparate impact
disparate_impact = unprivileged_ratio / privileged_ratio
print("Disparate Impact, Sex vs. Predicted Churn Status: " + str(disparate_impact))
Disparate Impact, Sex vs. Predicted Churn Status: 1.2761817068799684

```

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

AI 360 Toolkit-Metrics/Algorithms



Pre-Processing Algorithms	In-Processing Algorithms	Post-Processing Algorithms
Reweighting	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

