Algorithmic fairness in finance and health

Group number: 10
Group members:
Adarsh Kudithipudi, Archita Basavaraju
Venkatesh Miriyala , Ramgopal tummala

Introduction:

In an era where algorithms are increasingly influencing crucial decisions in the banking and healthcare sectors, the quest of justice in their application is critical. The combination of artificial intelligence (AI) and machine learning (ML) models has transformed the way we examine data and make important decisions. This gain, however, is accompanied by a fundamental challenge: the possible continuation of biases inherent in the datasets and models used. This research digs into the complex environment of algorithmic fairness, focusing on money and health. We investigate the many elements of bias detection and reduction, as well as the critical necessity to prioritize fairness in decision-making processes. The research, divided into two main sections, investigates the intricacies of fairness in finance and health, explaining techniques to detect, address, and prevent bias inside algorithmic systems.

Algorithmic Fairness in Finance:

The "Fairness in Finance" section methodically dissects the influence of biases in financial statistics and subsequent decision-making. We investigate methods for identifying bias in datasets, emphasizing the vital importance of vigilance in algorithmic credit decisions. We explore the issue of age bias in credit evaluations in particular, and provide mitigating techniques to promote equal financial decisions for all demographics.

Motivation

Our project is inspired by the Apple Card controversy exposed in The New York Times in 2019. When women were given lower credit limits than men with similar finances, it highlighted a need to address biases in financial algorithms. This incident motivates us to focus on making financial decision-making fairer and more transparent.

Apple Card Credit Limit Controversy:

Neli mentioned The Apple Card was chastised in 2019 for its gender bias in credit limits. Some customers complained that women were given less credit than men. It was also discovered that women received less credit than their spouses, despite having the same income and credit score. However, Apple denied that gender played a role in their credit underwriting process.
They claimed that credit decisions were solely based on an individual's and that the algorithm was intended to be unbiased Article.
https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html
The following is a recent tweet about the issue:
https://twitter.com/RudolphRitz/status/1659421109593460736

Methodology:

Data Splitting:

The German credit dataset will be divided into training and test datasets to ensure a robust evaluation of

the model's performance.

Bias Detection:

Initial analysis will involve assessing bias in the training data. We will employ aif360 to detect any disparities in the model predictions with respect to the designated protected attribute, "Age."

Bias Mitigation:

Upon identifying bias, we will implement mitigation strategies using aif360. This may involve adjusting the model or dataset to alleviate the identified biases.
Reassessment:

Mitigation Strategies:

Data definition:

The first step in identifying and mitigating biases in financial datasets is to define the data parameters and criteria. This entails conducting a thorough examination of the variables, attributes, and characteristics used in decision-making processes. We can identify potential sources of bias and develop strategies to address them by establishing clear definitions and standards for data elements.

Data gathering:

The data collection process is critical to ensuring the dataset's inclusivity and representativeness. Using a variety of data sources and methodologies aids in capturing a broad range of demographic, socioeconomic, and geographical factors. This diversification allows for a more nuanced understanding of the population, lowering the possibility of skewed or biased datasets.

Data labeling:

In order to train machine learning models, accurate and unbiased data labeling is required. To avoid the perpetuation of unfairness in the algorithms, it is critical to use robust labeling methodologies that are free of subjective biases. Using a diverse set of annotators and validation processes helps to reduce label bias and ensure a more balanced and equitable dataset.

Data pre-processing:

Pre-processing techniques are critical in mitigating biases in the dataset. This entails scrutinizing, normalizing, and cleaning the data to eliminate potential sources of bias. Techniques such as feature scaling, outlier detection, and imputation methods aid in the correction of biases and the creation of a more balanced representation of the dataset.

These mitigation strategies form a solid framework for identifying and correcting biases in financial datasets. We hope to improve the fairness and equity of algorithmic decision-making processes in the finance domain by meticulously implementing these measures throughout the data lifecycle.

Data Preparation and Pre-processing for Financial Algorithm Fairness Assessment:

The initial steps of data preparation and pre-processing are critical in the pursuit of ensuring fairness within financial algorithms. This section is divided into three sections: loading the dataset, specifying the protected attribute, and splitting the dataset into training and testing subsets.

Loading the dataset:

We are now set to use AI Fairness 360 (aif360) to find and address bias. We'll work with a German credit dataset, dividing it into a training and test set. Our goal is to identify bias in creating a machine learning model that predicts whether an applicant should receive credit based on different features in a standard credit application. The protected attribute is "Age," where "1" represents individuals aged 25 or older (privileged group) and "0" represents those younger than 25 (unprivileged group). In this initial tutorial, we'll examine bias in the initial training data, address it, and then reassess. More advanced machine learning processes are detailed in the author's tutorials and demonstration notebooks in the codebase.

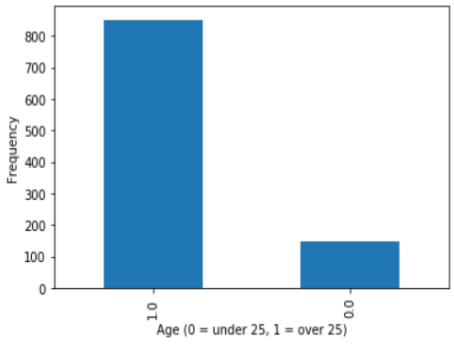
Specifying the protected attribute:

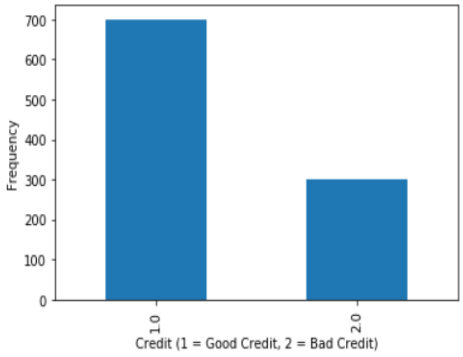
The identification and specification of the protected attribute within the dataset are critical in determining fairness. This attribute is typically used to represent sensitive demographic information (such as age, gender, and ethnicity) that must be protected from discrimination.

Splitting the dataset into training and testing subsets:

The training set is used to train the model, while the testing set is used to evaluate its performance. During this split, steps are taken to maintain the protected attribute's distributional parity across both subsets, ensuring equitable representation for comprehensive fairness assessment.

```
Original one hot encoded german dataset shape:  (1000, 57)
Train dataset shape:  (700, 57)
Test dataset shape:  (300, 57)
```





Compute fairness metric on original training dataset:

Disparate Impact measures the unequal effect of a policy on different groups. It is calculated by comparing the rates of favorable outcomes for a privileged group and an unprivileged group. A value less than 1 indicates a higher benefit for the privileged group, while a value greater than 1 suggests a higher benefit for the unprivileged group. The ideal value is 1, indicating equal treatment.

In [18]:
```
print("Original training dataset")
print("Disparate Impact = %f" % metric_orig_train.disparate_impact())
```

```
Original training dataset
Disparate Impact = 0.766430
```

Mitigate bias by transforming the original dataset

AI Fairness 360 is a tool that helps make artificial intelligence (AI) systems fairer. It has different methods to fix biases in data. One method it uses is called Reweighing. This method, found in the Reweighing class, changes the dataset to make sure both privileged and unprivileged groups have a fair chance of positive outcomes on a specific attribute.

Reweighing: Reweighing is a method to make sure that machine learning models treat different groups fairly. It involves assigning weights to training examples based on their group and label, aiming to eliminate bias. This helps in creating a discrimination-free training dataset. There are other methods like removing sensitive attributes or modifying labels, but reweighing is often more effective.

```
RW = Reweighing(unprivileged_groups=unprivileged_groups,
                privileged_groups=privileged_groups)
dataset_transf_train = RW.fit_transform(dataset_orig_train)
```

Compute fairness metric on transformed dataset

After transforming the dataset to reduce bias, we checked its effectiveness using the same metric as before. The transformation worked well, eliminating the bias and achieving equality in mean outcomes, with a difference of 0.0 compared to the initial 17% advantage for the privileged group.

```
In [29]: print("Transformed training dataset")
         print("Disparate Impact = %f" % metric_transf_train.disparate_impact())
```

```
Transformed training dataset
Disparate Impact = 1.000000
```

Results

The reweighing algorithm effectively mitigated the bias observed in the original training dataset, resulting in a transformed dataset where the mean outcomes are equal between privileged and unprivileged groups. The disparate impact metric is now 1.0, indicating a fair distribution of favorable outcomes between the two groups. The mitigation process successfully addressed the bias, achieving a more equitable distribution of outcomes in the transformed dataset.

Algorithmic Fairness in Healthcare:

In the realm of healthcare, algorithmic fairness plays a pivotal role in ensuring equitable outcomes for diverse patient populations. This investigation focuses on a heart attack prediction dataset where instances of heart attack are outnumbered by cases without such incidents. Acknowledging the imbalance, our approach involves leveraging the Synthetic Minority Over-sampling Technique (SMOTE) to harmonize the dataset. Additionally, we employ Local Interpretable Model-Agnostic Explanations (LIME) to enhance interpretability, shedding light on the nuanced decisions made by our predictive model. As a computer science master's student, this exploration delves into the crucial intersection of healthcare and artificial intelligence, striving for fairness and transparency in predictive analytics.

Motivation:

Our motivation stems from a critical examination of the article, "How machine-learning models can amplify inequities in medical diagnosis and treatment," by MIT researchers. This groundbreaking work, led by Marzyeh Ghassemi, highlights the potential biases existing in healthcare AI. Ghassemi's exploration into the amplification of disparities within underrepresented groups became a catalyst for our project, emphasizing the pressing need to address algorithmic fairness in healthcare.

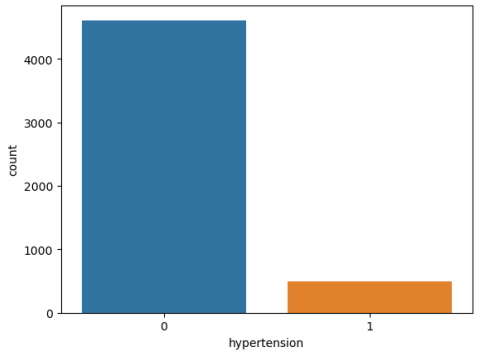
Methodology:

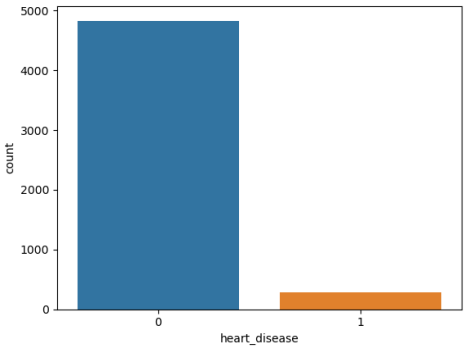
Dataset selection:

Dataset: Healthcare-dataset-stroke-data
Key parameters: id, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke.

Exploratory Data Analysis:

During our exploratory data analysis, a crucial observation emerged regarding the attributes related to health conditions in the dataset, particularly hypertension, stroke, and heart disease. In the context of our binary classification where cases with a heart attack are marked as '1' and those without as '0', a noteworthy imbalance was identified.
Specifically, the instances where the attributes hypertension, stroke, and heart disease are present (coded as '1') exhibit an imbalance compared to cases where these health conditions are absent (coded as '0'). The prevalence of '0' in these attributes is notably higher than '1'.
This imbalance suggests that within the dataset, a larger proportion of individuals tend to have the absence of hypertension, stroke, and heart disease compared to those who exhibit these health conditions. Understanding and addressing such imbalances are crucial steps in ensuring the robustness and fairness of our predictive models, particularly in the context of heart attack prediction. Future steps in our project will involve thoughtful strategies to handle these imbalances to enhance the overall performance and reliability of our predictive algorithms.





SMOTE Analysis:

In our pursuit of refining the heart attack prediction model, we leveraged the Synthetic Minority Over-sampling Technique (SMOTE) to address imbalances within the dataset. The SMOTE algorithm, a widely utilized method in handling class imbalances, focuses on augmenting the representation of the minority classes. It is a resampling technique designed to mitigate class imbalance by generating synthetic instances of the minority class (in this case, '1' representing cases with a heart attack. The augmented dataset helps the model generalize better to scenarios where the minority class attributes play a crucial role in predictions.
While SMOTE is a valuable tool for addressing imbalance, it is essential to apply it judiciously. Overly

aggressive oversampling may lead to model overfitting, impacting its performance on unseen data. Before applying the SMOTE, the dataset has about 195 counts of '1' and 3893 courts of case '0'. After applying SMOTE, there were equal instances of '0' and '1', about 3983 each.

Results:

We evaluated the performance of the model using various algorithms and considered RandomForest based on its metrics such as precision, recall and f1-score.

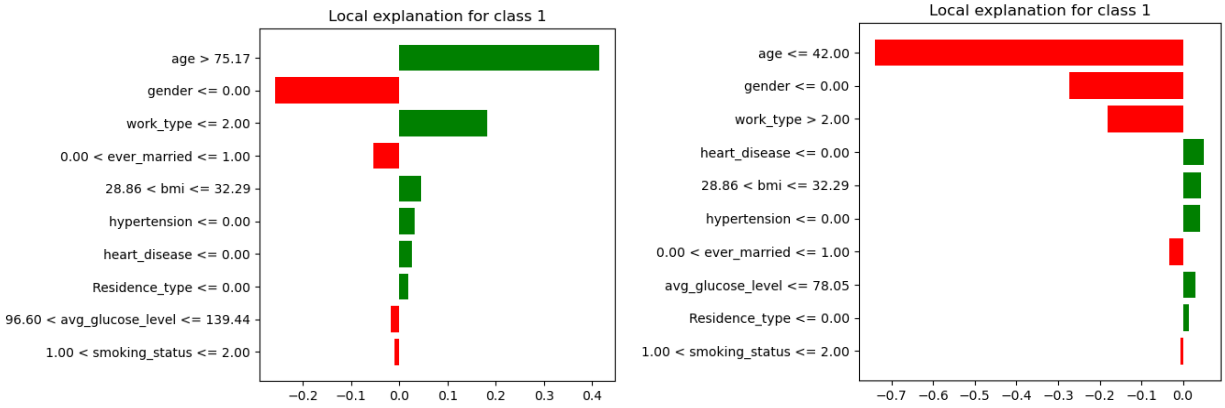
```
RandomForest: Average Fold Accuracy - 0.9571
GradientBoosting: Average Fold Accuracy - 0.9023
LogisticRegression: Average Fold Accuracy - 0.7962
SVM: Average Fold Accuracy - 0.7860
KNeighbors: Average Fold Accuracy - 0.8889
NaiveBayes: Average Fold Accuracy - 0.7972
DecisionTree: Average Fold Accuracy - 0.9362
AdaBoost: Average Fold Accuracy - 0.8639
XGBoost: Average Fold Accuracy - 0.9612
Overall OOF Accuracy: 0.3264
```

	precision	recall	f1-score	support
0	0.97	0.96	0.96	777
1	0.96	0.97	0.96	781
accuracy			0.96	1558
macro avg	0.96	0.96	0.96	1558
weighted avg	0.96	0.96	0.96	1558

```
[[745  32]
 [ 26 755]]
```

Interpretation using LIME:

LIME is a technique designed to enhance the interpretability of complex machine learning models. Its primary objective is to provide local and human-understandable explanations for individual predictions made by a model, especially in scenarios where the inherent complexity of the model might obscure the reasoning behind specific decisions.



Outcomes:

Graph1:

age > 75.17', 0.4103: If the age is greater than 75.17, it contributes positively (0.4103) to the prediction. Older ages are associated with a higher likelihood of the positive outcome.
'gender <= 0.00', -0.2722: If the gender is Male (0), it contributes negatively (-0.2722) to the prediction. Being male is associated with a lower likelihood of the positive outcome.
'work_type <= 2.00', 0.1874: If the work type is less than or equal to 2(Private job i.e if they work for Govt, self Never worked or Private sector), it contributes positively (0.1874) to the prediction. This could correspond to private employment or not having worked.
'hypertension <= 0.00', 0.0593: If hypertension is absent (0), it contributes positively (0.0593) to the prediction. The absence of hypertension is associated with a higher likelihood of the positive outcome.

Graph2:

age <= 42.00', -0.7392: If the age is less than or equal to 42.00, it contributes negatively (-0.7392) to the prediction. Younger ages are associated with a lower likelihood of the positive outcome.
'gender <= 0.00', -0.2741: If the gender is male (0), it contributes negatively (-0.2741) to the prediction. Being male is associated with a lower likelihood of the positive outcome.
'work_type > 2.00', -0.1800: If the work type is greater than 2, it contributes negatively (-0.1800) to the prediction. This could correspond to work types like private employment or self-employment.
'heart_disease <= 0.00', 0.0512: If the person doesn't have heart disease, it contributes positively (0.0512) to the prediction. The absence of heart disease is associated with a higher likelihood of the positive outcome.

Conclusion:

In the ever-evolving landscape of Artificial Intelligence (AI) in healthcare and finance, the transformative potential is undeniable, yet it is accompanied by ethical considerations that warrant meticulous attention. As AI assumes pivotal roles across diverse domains, the quality of the data it is fed becomes a critical factor in ensuring equitable and reliable outcomes. While our project successfully mitigated biases in datasets through adept preprocessing techniques, it emphasizes the paramount importance of addressing ethical concerns. As AI increasingly assumes critical roles in shaping human well-being, ensuring transparency, fairness, and ethical considerations is not just a technological imperative but a moral obligation with profound implications for human life.

Code: https://github.com/adarsh302/ethics-in-ai

Group contribution: Option 1: We agree that all group members made a valuable contribution and therefore believe it is fair that each member receive the same grade for the discussion.