Submission 2 – Team DL2

Introduction

Based on the exploratory data analysis in submission 1, we believe four variables (ZIP, rent, job stability and occupation) cause the bias. Thus, our hypothesis is: does removing these four variables reduce bias? This document aims to explain how the data was used to draw conclusions, which techniques and models were used to analyse and test our hypothesis. Our main goal was to identify bias and how to handle it. We started with building some models to analyse the data, evaluate the fairness of the dataset, introducing two methods to mitigate bias and finally what implications for businesses this work has.

First Step – Regression Models

To begin with, we built several regression models to see how the accuracy changes when removing variables that strongly correlate with the minority variable. We trained on the train.csv and tested it on the test.csv. The accuracy for the biased model was 37%, but when removing the variables *ZIP, rent, sex, job stability occupation,* it increased to 41%. We used a logistic regression as our base model to compare the results to the deep learning model.

Second Step – Deep Learning Model

In the second step we trained a 2-layer deep learning model on the biased dataset (12 columns) and trained another 2-layer deep learning model on the unbiased dataset (8 columns – excluded ZIP, rent, job stability, and occupation). The biased model scored an accuracy of 47% on the test dataset, while the unbiased model scored 85%.

The 2-layer deep learning model consists of 2 layers with ReLU activation and an output layer with sigmoid activation. We used a sigmoid to give us a 1D output instead of 2D with SoftMax. Also, the 1D output is easier for us to manipulate. A binary cross-entropy loss function was used to minimize the loss of our binary classification task.

We used StratifiedKFold to train/validate the model to ensure we had a reliable model. However, the issue seems to be more on the accuracy achieved when predicting using the test dataset, which ranged from 15% (predicting only 1s) to 85% (predicting only 0s). Therefore, we took an average of 10 models, resulting in 57% accuracy.

Third Step – Evaluating the Fairness of model

In the third step we analysed again the dataset to evaluate its "Fairness". Hereto, we applied three different approaches – "Demographic Parity", "Equal Opportunity", and "Equalized Odds".

Demographic Parity states that the proportion of each segment of a protected class (e.g. gender) should receive the positive outcome at equal rates. A positive outcome is the preferred decision, such as "getting to university", "getting a loan" or "being shown the ad". The difference should be ideally zero, but this is usually not the case. This means that the rate of (false positive + true positive) should be similar for both groups. For example, the model should predict a 50% default rate in both minority and non-minority. Overall, the biased

dataset does not achieve demographic parity since the difference between minority and non-minority are too large (9% vs 96%). Thus, removing the variables ZIP, rent, occupation and job stability from the dataset does ensure demographic parity (4% vs 4%).

Fig 1: Demographic Parity

Equal Opportunity states that each group should get the positive outcome at equal rates, assuming that people in this group qualify for it. The false negative and true positive rate should be the same for both groups. Overall, equal opportunity does not exist in the biased dataset (9% vs 94%). However, removing the variables ZIP, rent, occupation and job stability from the dataset does remove the bias in equal opportunity (4% vs 4%).

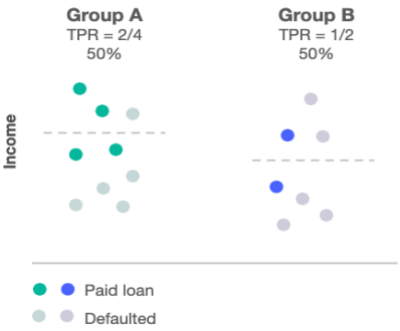

Fig 2 Equal Opportunity

Equalized Odds. This concept states that the model should: correctly identify the positive outcome at equal rates across groups *(same as in Equal Opportunity)*, but also miss-classify the positive outcome at equal rates across groups *(creating the same proportion of False Positives across groups)*

The biased dataset does not satisfy the constraints imposed by equalized odds suggesting the model is biased (94% vs 9%) and (3% vs 91%).

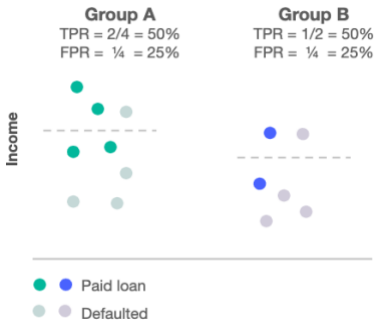

Fig 3: Equalized Odds

Thus, removing the variables ZIP, rent, occupation and job stability from the dataset does remove the bias in equalized odds (4% vs 4%) and (95% vs 95%).

Overall, removing the variables ZIP, rent, occupation and job stability from the dataset does make the model fairer i.e. less biased. However, removing variables is a drastic approach since the removed variables might contain signal i.e. less accuracy. Therefore, we tried other two approaches to mitigate the bias and retain the removed variables – down-sampling the train dataset and the usage of the "Disparate Impact Remover" algorithm.

Fourth Step – Down sample

To retain the variables ZIP, rent, occupation and job stability and at the same time reduce bias, we down sampled our train dataset. We aimed to have a 50% default rate in minority and non-minority. Thus, we reduced the train dataset to 1'000 rows since the minority only contained 250 cases of non-default. A deep learning model trained on the down-sampled dataset achieved a 52% accuracy, an improvement from 37% achieved with a logistic regression. Furthermore, the down-sampled model reduced bias in all metrics (demographic parity, equal opportunity and equalized odds) compared to the bias deep learning model. In our case down-sampling removed more than 99% of all cases. This is not ideal, therefore we present another solution - the Disparate Impact Remover.

Fifth Step – Disparate Impact Remover

After clearly identifying the bias and looking at the metrics we need to optimize (step 3), we used an algorithm called "Disparate Impact Remover". This algorithm tries to reduce the bias in the dataset by ensuring that the features in the un-privileged case (in our case minority) are similar to the features in the privileged case (non-minority). The disadvantage is a reduction in accuracy. We applied disparate impact to the biased dataset and the down-sampled dataset. In both cases, disparate impact remover changed the features to such an extent that it contained only noise, i.e. the model only predicted 0s in the biased dataset and 1s in the down-sampled dataset.

Sixth Step - Final results

We trained models on this dataset and compared the results with each other.

Model	Process	Accuracy on Test	Default	Precision	Recall	F-1 Score	Minority	Accuracy on test	Demographic Parity	Equal Opportunity	Equalised Odds
		0.37	0	0.82	0.33	0.47					
LR	Bias		1	0.14	0.58	0.22		Not calculated			
		0.41	0	0.84	0.38	0.53					
LR	No Bias		1	0.14	0.59	0.23		Not calculated			
		0.37	0	0.84	0.32	0.46					
LR	Down-S		1	0.14	0.65	0.24		Not calculated			
		0.47	0	0.85	0.47	0.60	0	0.79	0.96	0.93	0.03
DL	Bias		1	0.15	0.52	0.23	1	0.17	0.09	0.09	0.91
		0.85	0	0.85	0.95	0.90	0	0.82	0.05	0.05	0.95
DL	No Bias		1	0.15	0.04	0.07	1	0.82	0.05	0.05	0.95
		0.52	0	0.91	0.49	0.64	0	0.51	0.58	0.67	0.51
DL	Down-S		1	0.20	0.71	0.31	1	0.56	0.45	0.73	0.46

Source: Results from 4-Deep Learning.ipynb notebook

The fairest (least biased) model is deep learning trained on the non-bias dataset (excluding ZIP, rent, job stability and occupation) since the metrics demographic parity, equal opportunity and equalised odds are very similar to each other (less than 5% difference). This is also the model with the highest accuracy with 85%, while our base model (logistic regression) achieves only 41%. Down-sampling does improve the model's fairness compared to the biased model and has a higher accuracy at 52%.

Final Step – Implications for Businesses

Bias in Financial Data

Artificial intelligence (AI) is making rapid inroads into many aspects of our financial lives. Algorithms are being deployed to identify fraud, make trading decisions, recommend banking products, and evaluate loan applications. This is helping to reduce the costs of financial products and improve their efficiency and efficacy. However, there is growing evidence that AI systems are biased in ways that may harm consumers and employees. As regulators turn their attention to the impact of financial technology on consumers and markets, firms that deploy AI may be exposing themselves to unanticipated risks.

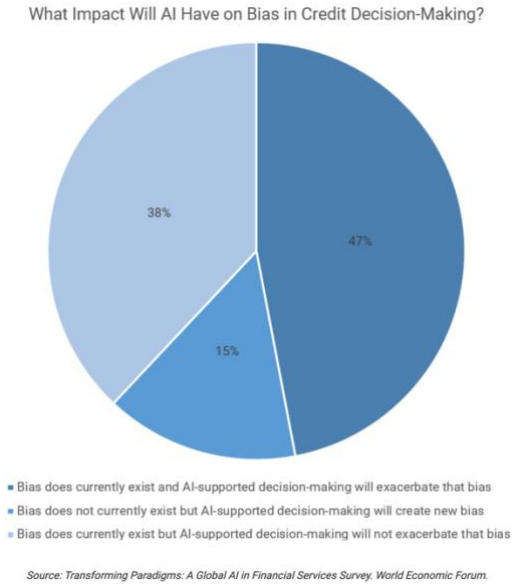

Fig. 4: AI Impact on Bias in Credit Decision-Making[1]

What Businesses can do to prevent

The figure shows that AI can have a very high impact on credit decision-making, if data is biased. This is due to the fact that the AI-models you create are trained on a biased dataset, thus leading to biased predictions and decision-making. The following 4 steps can help to prevent businesses from generating biased data.

1) *Maintain Transparency* – To know where the data comes from, how it is created and can be used will help to avoid bias since complete transparent data is understandable and reachable for everyone in the company and outside.

2) *Substantiate your Assumptions* – Always maintain quantitative evidence for validating your assumptions and hypotheses and approve the perceived impact of the application. Also, the model's intended use and appropriateness for that use should be accompanied by quantitative external evidence
3) *Ensue Data Quality and Security* – Training data needs to be vetted for accuracy, relevance, and freshness. Regularly updating your data is very important. Data security practices are frequently also needed to ensure that inappropriate datasets are not accidentally accessed by those building AI models.
4) *Perform Regular Bias Assessments* – It is important to regularly control your data for bias. Hereto, there are some tools that help to detect and mitigate it. Production AI systems should actively monitor for incomplete data. Two possible solutions have been presented in this paper – Down sampling and Disparate Impact Remover.

What businesses can do to mitigate bias

This work is focused on data biases and how that can affect the outcome of a system. For businesses nowadays it is crucial to fully understand the data they have and what can and can't be done with it. Businesses have two kinds of research approaches to handle bias and both focus on outcome fairness [1]. One is called "individual fairness," where the idea is to ensure that similar individuals are treated similarly. The second approach is focused on what is known as "group fairness." Here, the idea is to ensure that the error rates for different groups of people are similar. This type of approach is also focused on process fairness, i.e. when taking a decision to make sure that the process in place to make that decision is fair for everyone involved. All these kinds of fairness can be understood and evaluated with the metrics we tried to optimize; "Demographic Parity", "Equal Opportunity" and "Equalized Odds". We have shown that, though not the best method, down sampling provides a solution for reducing bias by achieving a 50% distribution rate of the target variable for the biased feature. The second method ensures that equalize the features that were biased i.e. minority vs non-minority. Unfortunately, this algorithm created only noise in its predictions.

Nonetheless, we have shown what metrics are important to analyse when assessing bias in data and two solutions that mitigate it. However, it is crucial to implement a company policy to try to prevent biased data in the first place.

Here is our Github Repository, with all the notebooks we created.

References

(1) https://www.intuition.com/disruption-in-financial-services-racist-robots-how-ai-bias-may-put-financial-firms-at-risk/