Case Studies

Bias and Fairness in ML: KDD 2020 Tutorial

# CASE STUDY A – Loans

**Description**: Acme Bank is developing a system to decide which loan applications to deny based on predicted risk of lender not paying back their loan in time.

**Goal**: Increase repayment rates for bank loans

**Data**: Historical loans and payments, credit reporting data, background checks

**Analysis**: Build model to predict risk of not repaying on time

**Actions**: Deny loan or increase interest rate/penalties

## Breakout Session 1 –– Sources of Bias:

### What are some potential sources of bias in the underlying data?

[The bullet points below are filled out answers from the participants in the KDD 2020 Tutorial]

Demographic discrimination - People from underpriviledged demographics might have inherent biases against them. They could also be underrepresented leading to sampling biases. Historical data would not consider the gain in economic situations post the loan.

Low education level

Historic prevention of homeownership in BIPOC Populations, religious and ethnic minorities

If the borrower is native speaker or not.

Zip code - There have been ways to infer race of people based on the zipcodes of their residential addresses

Historical redlining discriminated against minorities in terms of real-estate investment- claiming them as areas of high-risk investment areas without factual basis. This has also affected property values and thus quality of education (schools funded by property taxes)

Black and brown folks are disproportionately targeted and convicted by the justice system

Types of data available to the **lender**

* Demographic data
* ZIP code
* Income level, income type
* Reference letters
* Historical load default rates
* Interview with the borrower - notes
* Education level
* Credit report

Card type

neighborhood

Age

Historical data

Income

Gender

Ethnic groups

Occupation

The location of the banks where the data were collected.

Missing data

### How might biases be introduced in the data science pipeline? (Think about ETL, record linkage, feature engineering, labels, modeling, and model selection)

[The bullet points below are filled out answers from the participants in the KDD 2020 Tutorial]

* In an end to end approach, the model could learn the covariances between ethnic groups / classes of population and default payments. Carefully planned feature engineering could potentially remove these biases.
* If data contains general practices in the people and biases, those will be learned by the model.
* Imbalanced data
* Feature importance in model could be biased towards one and if there's a slight change then it would affect the decision.
* Sampling the data when creating the training dataset. If we don’t use stratified sampling there could be bias. E.g. we are including more males into training than females.
* Correlation analysis between training features and features we don’t want to discriminate against.
* Previous bias could reinforce new biased decisions that will be included in the dataset
* Some features might have null values that are not missing at random. For example, male has more missing values than female. When we impute them there could be bias.
* Low representativity of certain groups in the data might be exacerbated by imputation
* Incompleteness of certain records
* Data collection strategy is faulty - For instance, you only collect data about successful loan payment customers and not the defaulters (or vice versa)
* Some models do not accommodate ‘outliers’ as well
* Model selection could be biased in the way that we select model with highest accuracy but ignore fairness
* Lack of electronic records, or lack of records in English might disadvantage immigrants

### What are the risks to fairness in downstream applications and deployment of the model described?

[The bullet points below are filled out answers from the participants in the KDD 2020 Tutorial]

Leading to future bias as current decisions will bias the makeup of future datasets

Unexplainable rejection of loans for specific groups of people for no reason apart from the fact that they belong to that group. Such a system would also lead to strengthening human biases.

Bias perpetuation

Propagating inequity in loan determination

Contrarily, the AI/ML system could be inherently more fair since it would

* Eliminate direct human biases
* Correct biases through variable reweighting

## Breakout Session 2 –– Bias Metrics:

### How would you describe a **false positive in this problem** to a policymaker or business owner? What’s the potential harm/cost of one?

### How would you describe a **false negative** to a policymaker or business owner? What’s the potential harm/cost of one?

### What confusion matrix metric (e.g., FPR, FNR, TPR, FDR, etc.) would you choose to focus on in terms of equity for this case? Think of the fairness tree here.