Microsoft Azure

Module 9: Responsible Machine Learning

Mohammed Arif

Agenda

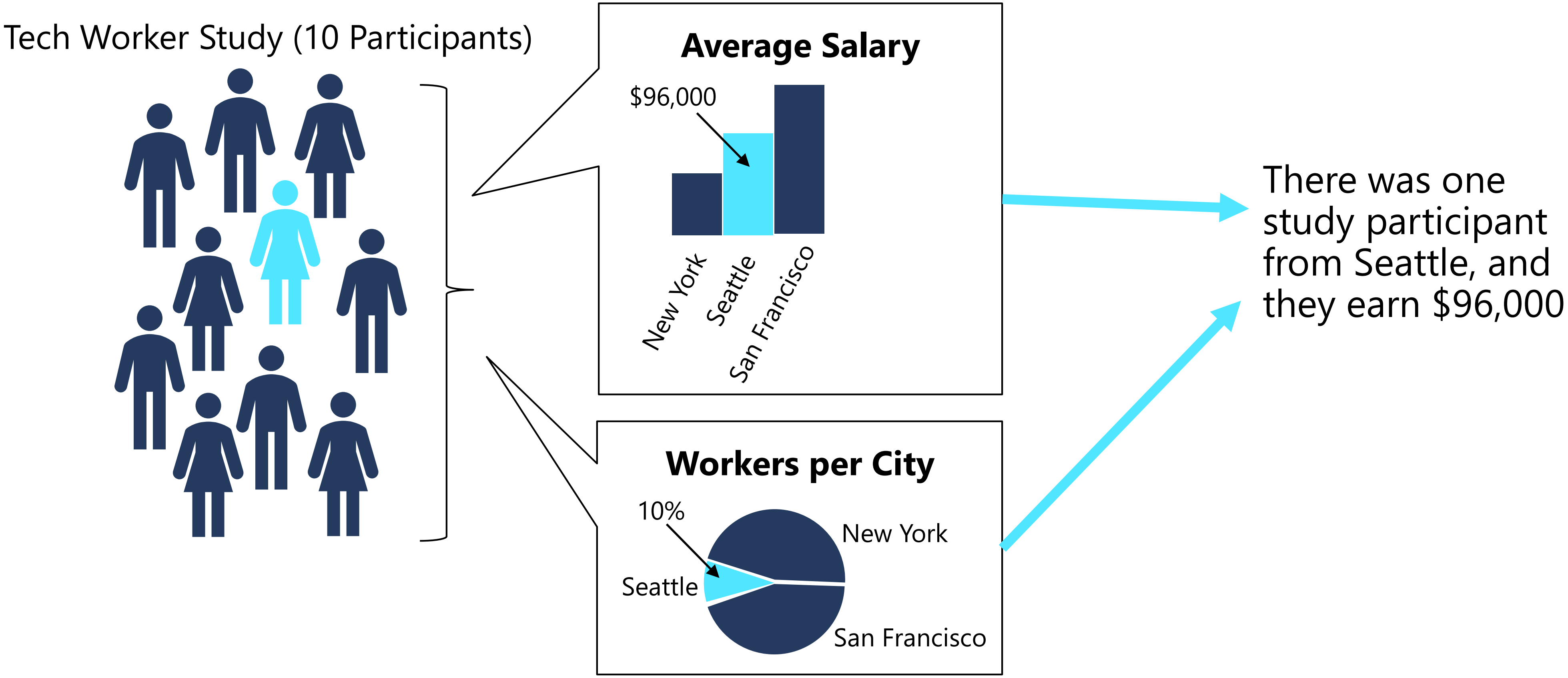
- Differential Privacy
- Model Interpretability
- Fairness

Differential Privacy

The Data Privacy Problem

Studies are ethically and legally required to protect personal information

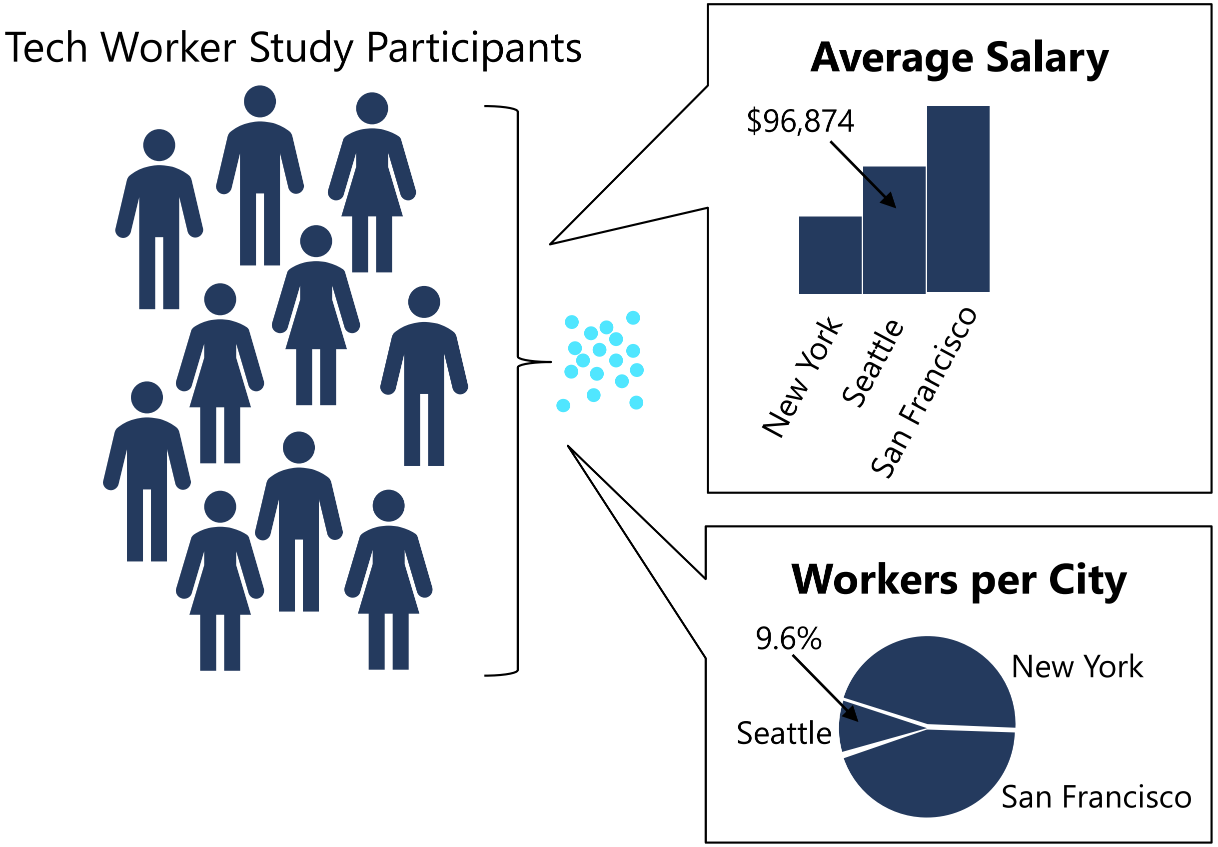
Repeated analyses of aggregated results can reveal details about individuals



What is Differential Privacy?

The analysis function adds random "noise" to the data

Results are statistically consistent, non-deterministic approximations



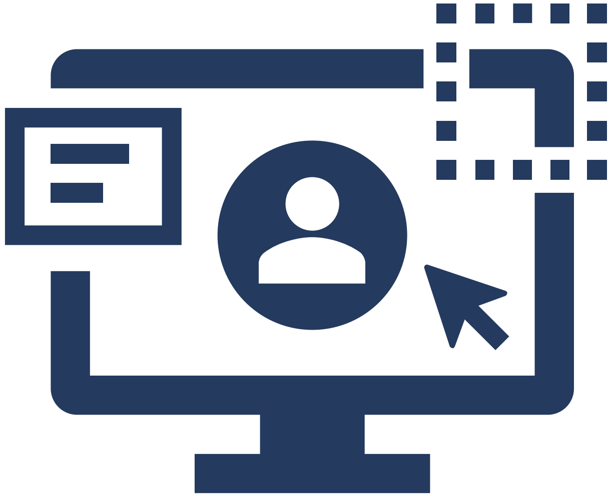
- Each analysis produces slightly different results due to random noise
- Results are statistically consistent with true data distribution allowing for random deviation based on probability
- Individual contributions to the aggregated values are not identifiable

Epsilon - The Privacy Loss Parameter

- **To minimize risk of personal identification, an individual could *opt out* of a study**
  - To be effective for all individuals, they would all need to opt out - so the study would be useless
- **Differential privacy adds noise so the maximum impact of an individual on the outcome of an aggregative analysis is at most *epsilon* (ϵ)**
  - The incremental privacy risk between opting out vs participation for any individual is governed by ϵ
  - Lower ϵ values result in greater privacy but lower accuracy
  - Higher ϵ values result in greater accuracy with higher risk of individual identification

Privacy $\longleftarrow$ ϵ $\longrightarrow$ Accuracy

Lab: Explore Differential Privacy



1. View the lab instructions at https://aka.ms/mslearn-dp100
2. Complete the **Explore differential privacy** exercise

Model Interpretability

Model Interpretability in Azure Machine Learning

Statistical explanation of feature importance

Quantifies the influence of each feature on prediction

Important to identify bias or unintended correlation in the model

Based on the Open Source *Interpret-Community* package

Includes explainers based on common model interpretation algorithms like:

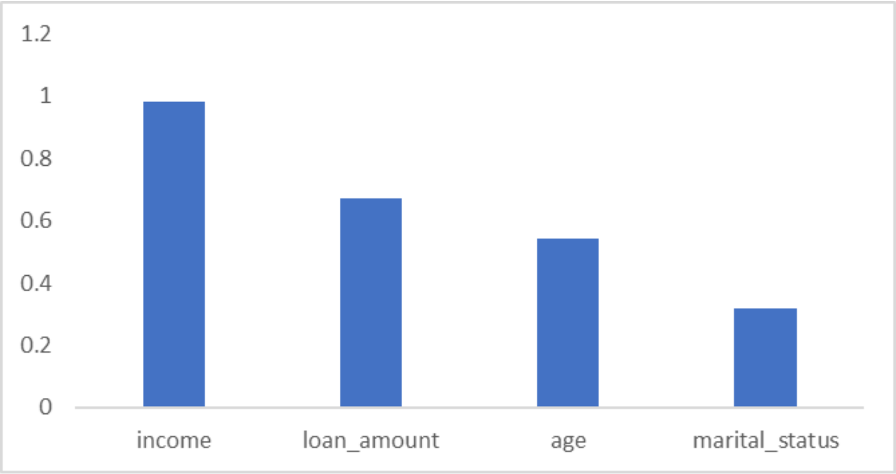
- Shapely Additive Explanations (SHAP)
- Local Interpretable Model-Agnostic Explanations (LIME)

Global and Local Feature Importance

Global Feature Importance

Overall feature importance for all test data

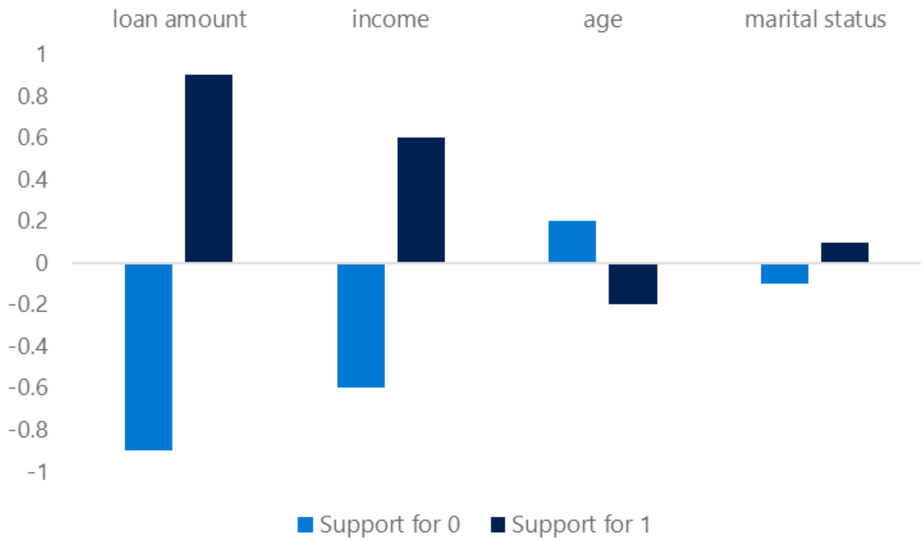
Indicates the relative influence of each feature on the predicted label



Local Feature Importance

Feature importance for an individual prediction

In classification, this shows the relative support for each possible class per feature



Explainers

Use the azureml-interpret package

Create an explainer:

MimicExplainer – global surrogate model that approximates your model

TabularExplainer – Invokes direct SHAP explainer based on model architecture

PFIExplainer – Permutation Feature Importance based on feature shuffling

Get global or local feature explanations

```
from interpret.ext.blackbox import TabularExplainer

tab_explainer = TabularExplainer(model, X_train, features=features, classes=labels)
global_explanation = tab_explainer.explain_global(X_train)
```

Adding Explanations to Training Experiments

In the training script, import the ExplanationClient class

Generate explanations and upload them to the run

```
explain_client = ExplanationClient.from_run(run)
explainer = MimicExplainer(model, X_train, LinearExplainableModel,
                           features=features, classes=labels)
explanation = explainer.explain_global(X_test)
explain_client.upload_model_explanation(explanation, comment='Model Explanation')
```

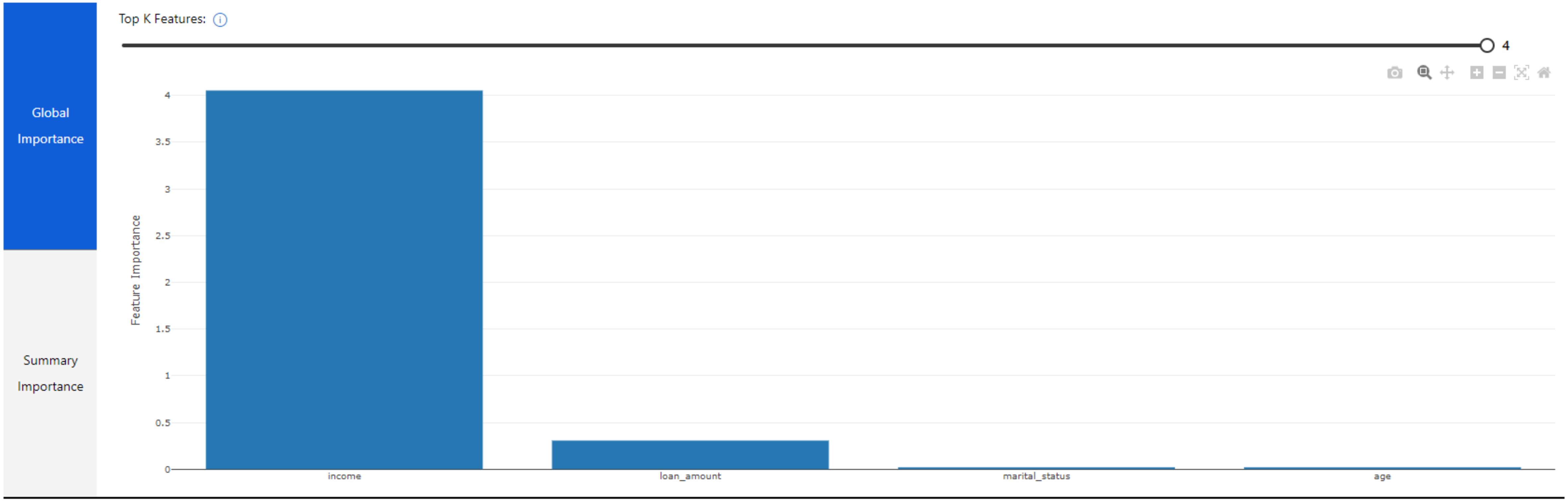
Use ExplanationClient to download explanations

```
from azureml.interpret.explanation_client import ExplanationClient

client = ExplanationClient.from_run_id(workspace=ws,
                                       experiment_name=experiment.experiment_name,
                                       run_id=run.id)
explanation = client.download_model_explanation()
```

Visualizing Model Explanations

View the Explanations tab for the run in Azure Machine Learning studio



Interpretability During Inferencing

Register a lightweight scoring explainer with the model

```
scoring_explainer = KernelScoringExplainer(explainer)
save(scoring_explainer, directory='dir', exist_ok=True)
Model.register(ws, model_name='model', model_path='dir/model.pkl')
Model.register(ws, model_name='explainer', model_path='dir/scoring_explainer.pkl')
```

Use the model and the explainer in the service scoring script

```
def run(raw_data):
    data = json.loads(raw_data)['data']
    predictions = model.predict(data)
    local_importance_values = explainer.explain(data)
    return {"predictions":predictions.tolist()), "importance":local_importance_values}
```

Deploy a service with the model and explainer

```
service = Model.deploy(ws, 'classify_svc', [model, explainer], inf_config, dep_config)
```

Lab: Interpret Models

1. View the lab instructions at https://aka.ms/mslearn-dp100
2. Complete the **Interpret models** exercise

Fairness

What is Fairness?

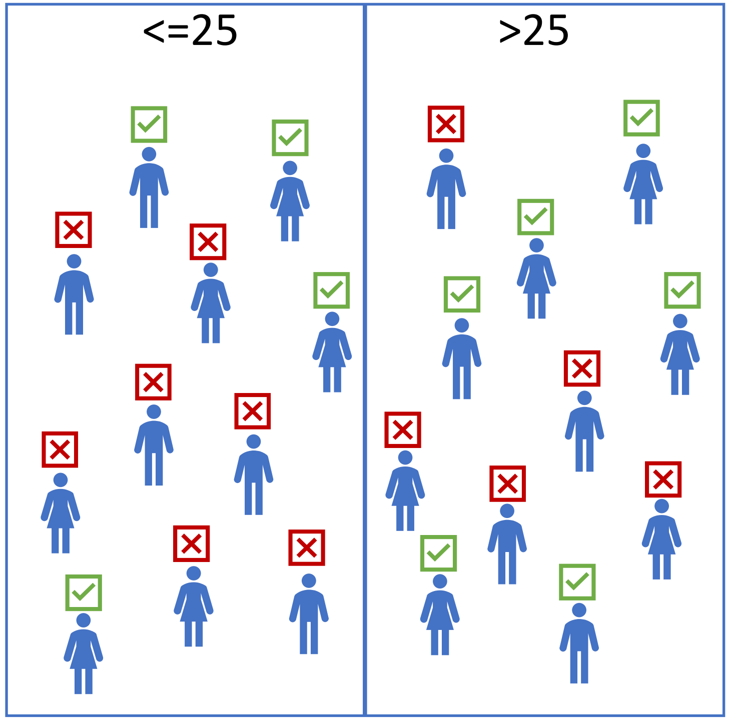
Absence of negative impact on groups based on:

- Ethnicity
- Gender
- Age
- Physical disability
- other sensitive features

Evaluating Model Fairness

Example: Loan repayment binary classification for two age groups

Selection Rate Disparity



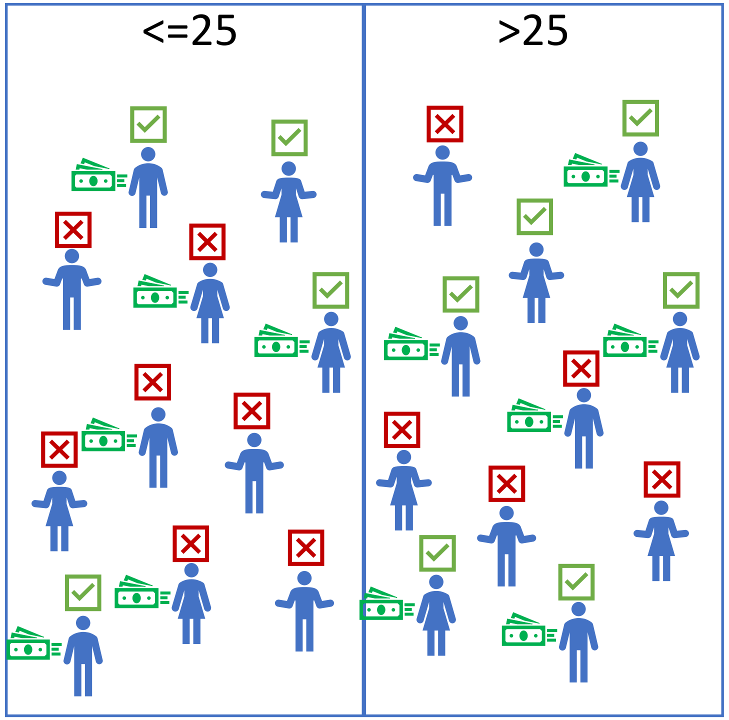
Overall *selection rate* = 10/22 (45%)

25 & under *selection rate* = 4/11 (36%)

Over 25 *selection rate* = 6/11 (54%)

Disparity = 18%

Prediction Performance Disparity



Overall *recall* = 8/12 (67%)

25 & under *recall* = 3/6 (50%)

Over 25 *recall* = 5/6 (83%)

Disparity = 33%

Mitigating Unfairness

Create models with *parity constraints*:

- **Demographic parity**: Minimize disparity in the selection rate across sensitive feature groups.
- **True positive rate parity**: Minimize disparity in *true positive rate* across sensitive feature groups
- **False positive rate parity**: Minimize disparity in *false positive rate* across sensitive feature groups
- **Equalized odds**: Minimize disparity in combined *true positive rate* and *false positive rate* across sensitive feature groups
- **Error rate parity**: Ensure that the error for each sensitive feature group does not deviate from the overall error rate by more than a specified amount
- **Bounded group loss**: Restrict the loss for each sensitive feature group in a regression model

Lab: Detect and Mitigate Unfairness

1. View the lab instructions at https://aka.ms/mslearn-dp100
2. Complete the **Detect and mitigate unfairness** exercise

Knowledge check

In a differential privacy solution, what is the effect of setting an *epsilon* parameter?

- [x] A lower epsilon reduces the impact of an individual's data on aggregated results, increasing privacy and decreasing accuracy
- [ ] A lower epsilon reduces the amount of noise added to the data, increasing accuracy and decreasing privacy

---

You have trained a model, and you want to quantify the influence of each feature on a specific individual prediction. What kind of feature importance should you examine?

- [ ] Global feature importance
- [x] Local feature importance

You are training a binary classification model to support admission approval decisions for a college degree program.

How can you evaluate if the model is fair, and doesn't discriminate based on ethnicity?

- [ ] Evaluate each trained model with a validation dataset and use the model with the highest *accuracy* score.
- [ ] Remove the ethnicity feature from the training dataset.
- [x] Compare disparity between selection rates and performance metrics across ethnicities.

References

Microsoft Learn: Explore differential privacy
https://docs.microsoft.com/learn/modules/explore-differential-privacy

Microsoft Learn: Explain machine learning models with Azure Machine Learning
https://docs.microsoft.com/learn/modules/explain-machine-learning-models-with-azure-machine-learning

Microsoft Learn: Detect and mitigate unfairness in models with Azure Machine Learning
https://docs.microsoft.com/learn/modules/detect-mitigate-unfairness-models-with-azure-machine-learning

Azure Machine Learning responsible ML documentation
https://docs.microsoft.com/azure/machine-learning/concept-responsible-ml

Microsoft Azure