Software Engineering and Testing for AI Systems

Nicholas Wu

December 7, 2025

1 Assignment 1

1.1 Data Inspection - max 0.5 page

Describe at least 3 problems with the dataset, where at least 1 of these touches upon questions of validity.

1. Problem 1:
   The label **checked** shows who the organization decided to investigate, not who actually committed welfare fraud. It reflects past decisions influenced by rules, priorities, workload, and possible biases. A model trained on this dataset learns that the likelihood of being investigated rather than the likelihood of committing fraud, so the label has poor construct validity. Furthermore, the dataset does not contain any true ground-truth labels for fraud. We are only given Ja/Nee which seems to represent a ”risk” score calculated from a previous model where checked is the decision based off of thresholding. This highlights that the original dataset did not have a well established ground truth to base off of, rather, they somehow computed an arbitrary risk score with no correlation to fraud.

2. Problem 2:
   The feature **contacten_onderwerp_beoordelen_taaleis** mostly applies to people whose Dutch language skills are being assessed, which strongly correlates with having a migration background. Using it as a predictor allows the model to treat language assessment as a risk factor and to indirectly use a proxy for migration background. This can result in unfair over-selection of non-native speakers and reduces the fairness and validity of the model.

3. Problem 3:
   The feature **appointment_number_words** is not indicative or significant to reflect welfare risk because the number of ”words” to report an appointment between customer and municipality does not evaluate the ”risk” of an individual. The appointment could be anything from registering a new address to requiring actual social aid. It is not appropriate to use this feature without knowing the procedure or semantics that constitutes the ”word” count as there may be no significance between count and risk. For example of a regular address change, the report could even contain documents attachments of previous addresses/contracts/housing agreements which also ”skews” the word count.

2

2.1 Subgroup #1 (Nicholas and Ocean) Possible test cases - max 1 page

We have implemented the following **partitions** on the data set.

1. **Language proficiency:**
   Individuals are separated based on whether the language requirement has been met. This helps us see whether the model assigns systematically higher risk scores to people with limited language proficiency. Since lower language proficiency is more common among recent migrants, treating it as a signal may unfairly target a group already facing barriers rather than reflecting true risk.

2. **Address instability:**
   We use the number of address changes as a proxy for stability and split the data into a typical group and a high-instability group (based on the upper quartile). This allows us to test whether housing instability unfairly influences model predictions. Because housing instability is more common among recent migrants and lower-income groups, using address changes as a signal may unfairly amplify existing socioeconomic disadvantages rather than reflect true fraud risk.

3. **Neighborhood:** Using the recorded neighborhood information, we compare model scores across areas with enough data. Large differences may indicate geographic or socioeconomic bias in the predictions.

2.2 Metamorphic Testing for Bias Detection

To detect potential bias in the trained models, we employ metamorphic testing focused on language proficiency as a protected attribute. The test applies a simple transformation: all language proficiency values are set to "met" (value 1), regardless of their original status. This creates an unlikely dataset where everyone appears to meet the language requirement. If a model is unbiased, this transformation should have minimal impact on predictions, since language proficiency should not influence fraud risk assessment. We measure the impact by calculating the absolute difference between original and transformed predictions for each sample. For each model, we report mean change, median change, standard deviation, minimum change, and maximum change across all samples. Lower values indicate the model is more invariant to language proficiency and therefore exhibits less bias. Higher values suggest the model has learned to associate language proficiency with fraud risk, which is discriminatory behavior we do not want.

2.3 Subgroup #2 (Johnny and Emre) Possible test cases - max 1 page

Partitions:

1. **Age Groups:**
   Historically, age has often been associated with unfair treatment in both welfare and employment contexts. While a machine-learning model may evaluate numerical features in a seeminly more "systematic" way and therefore more *fair*, and does not have an inherent perception of age in the same way humans do - this does not guarantee that the underlying data is completely free from age-related biases. Partitioning by age helps us evaluate whether the models perform disproportionately worse for specific age groups. This is particularly important for younger individuals who may have less stable economic opportunities or fewer data administrative records in the system, making them more vulnerable to misclassifications. Considering that the datasets contain an abundance of features reflecting contact with the municpality or other institutional systems - in which these contacts usually accumulates with age - it is important to evaluate model behavior across age groups rather than trying to filter out through all of these "co-lineared" features.

2. **Gender:**
   While a model may no explicitly treat genders unfairly using gender as a direct feature, partitioning on gender helps us identify whether other features in the dataset functions as gender proxies and indirectly influence prediction of the models. Social welfare systems often have differing interactions and evaluations with men and women due to socioeconomic factors, caregiving roles, and gendered distributions of labor and requirements for support. These differences may be encoded in the underyling data (even unintentionally), and can cause performance disparities across genders.

3. **Medical Status (Exemption, Availability, Physical & Psychological Issues):**
   Medical-relataed features capture human limitations such as physical, psychological, and other medically documented restrictions that can hinder an individual's ability to work or participate in municpality-related activities. A model's prediction should therefore not be a consequence of health features that individuals, to an extent, cannot control. By partitioning on different medical features, we can detect whether the models will systematically disadvantage individuals with certain medical histories and constraints.

Metamorphic Tests:

For our model we decided that per test change on model prediction based on the morphed data less then 3% will be accepted as passed the test.

1. **Confidential Flag Invariance**
   The confidential flag is a data point that is completely unnecessary and unrelated for the purpose of the model. So any change in this field in the data should not affect the overall result.

2. **Age Invariance**
   Age is something very related to the data but based on our partitioning and grouping of ages, jittering the ages between the numbers allowed withing the groups should also have minimal to no impact on the model.

3. **Gender Invariance**
   The prediction of the model should not change based on the gender of the person. If this is the case this means the model is biased and problematic.

4. **Contact Channel Invariance**
   The models predition should not be effected at all by the changes in the contact channels. These values are complealty unrelated to the prediction if changes on these values effects the final prediction score, there is a problem in the model.

3

3.1 Training Good and Bad Models: Ocean Nicholas

We train two gradient boosting classifiers with contrasting fairness properties to evaluate our bias detection approach. The models differ in their training strategy but share several key characteristics:

- Both use identical architectures (gradient boosting with 100 estimators)
- Both receive the same 315 input features, including language proficiency
- Both achieve similar predictive accuracy (approximately 94.5%)
- Both are exported to ONNX format for consistent evaluation

Before training, we research the dataset to understand data bias patterns. The `analyze_language_bias.py` script examines how language proficiency correlates with the target variable. For each language status (not met, met, special), the script calculates the proportion of individuals who were checked and compares it to the overall base rate. This shows whether certain language groups are disproportionately represented in the positive class. The code also computes disparate impact. It is defined as the ratio of selection rates between language groups. A disparate impact ratio exceeding 1.25 points at potential bias that violates fairness thresholds. This initial analysis informs our training strategy and is for baseline measurements for evaluating model fairness.

The bad model uses uniform sample weights during training, allowing it to learn bias naturally from patterns in the data. This model treats all samples equally and can exploit correlations between language proficiency and the target variable.

The good model employs a reweighting strategy to reduce reliance on language proficiency. We calculate the positive class rate for each language group and compare it to the average across all groups. When a group is overrepresented in the positive class (exceeding the target rate by more than 10%), we down-weight its positive samples by a factor between 0.6 and 1.0. This balances the positive class distribution across language groups without removing the protected feature entirely, demonstrating that fairness can be achieved through training strategy rather than feature selection. The models are randomly assigned to file names after training to enable blind evaluation during metamorphic testing.

3.2 Subgroup #2 (Johnny and Emre) A purposefully 'good and 'bad' model - max 3 pages

Good Model:

- **Data Pre-processing:)** For the pre-processing of the the features, we deliberately removed features that are protected characteristics (ex: age & gender) as per our explanation in Section 2.

  Additionally, we handpicked and removed the following features that may induce bias closely related to those protected characteristics and biases in general:

  - `appointment_number_words` as highlighted in Part 1) #3, word count of documents is not an appropriate measurement for fraud detection as it is too broad and semantics associated with word count is unclear. There's no indication or justifiable means to say that higher or lower word count on a document indicates bad or good faith of an individual.
  - `address_unique_districts_ratio` reflects how many different districts an individual has lived in. This is directly related to protected characteristics such as age: where younger individuals may move more frequently due to schooling, unstable housing, shared accomodation and early-career transitions, while older adults generally maintain more stable long-term housing via financial and job security. It is linked with gender since women with caregiving responsibilities may relocate to access childcare, education, or safer environment for their childrens, or even additional relocation as a consequence of relationship changes or separation. In addition to that, this feature strongly correlates with socioeconmic instability, migration background, and neighborhood segregations, which can act as an indirect proxy to protected demographic characteristics as well. A model trained on this feature may incorrectly interpret housing mobility as behavioural risk associated with fraud.

  After removal of these specific features, the focus was on removing all features related to certain categories:

  - `address_latest_district_*`: 9 features, each corresponding to a specific district or 'other'
  - `address_latest_neighborhood_*`: 5 features, each corresponding to a neighborhood or 'other'
  - `address_latest_district_*`: 2 features, either Rotterdam or 'other'

    These district, neighbor, and place address related features encode geographic information that strongly correlates with socioeconoic status, migration background, and concentrations of demographics within Rotterdam. Including these features allows the model to learn geographic or demographic proxies rather than a legitimate means for evaluating fraud-behavior.
  - `address_latest_district_*`: 7 features, captures the communication medium used between individuals and the municipality, such as emails, phone calls, in-person visits, or exchanged documents. Communication preferences are often correlated with age (e.g., older individuals relying less on digital communication due to lower proficiency with technological channels), gender (e.g., difference in communication patterns or administrative burden), and disability status (physical, physiological, or cognitive limitations). As a result, these features do not reflect meaningful fraud-related behaviors.

Bad Model:

For the "bad" model we decided to created a modified version of the basic model as based on our tests the basic model already could be considered a "bad" model. So our "bad" model does not do any kind of data pre-processing or selection. To ensure we actually had a "bad" model not just based on the how we train the model but also based on the data we increased the bias of 2 specific features in the model. Those are:

- **Gender** The model specifically penalize people who are females, specifically flips nearly all of their positive values into negative ones.
- **Age** Similarly To gender, the model also penalize people over the age of 50, also flipping their positive values into negatives.

Model Accuracy Evaluation:
The Model Accuracy is as follows:

- **Good Model (Model 1)**: 91.6%
- **Bad Model (Model 2)**: 88.9%

While accuracy is not a good indication on the validity of ”fairness”, it is noteworthy to highlight that Model 1 achieved higher overall accuracy (∼2.7%) compared to Model 2 while using fewer features (∼25 features) - specifically as it removes sensitive attributes and clusters of correlated features. This shows that despite having access to less information, Model 1 generalizes better, while Model 2 performs worse even while using the full set of features which is an indication of Model 2 learning relationships that harm the quality of predictions.

Partitioning Test Evaluation:

Partition	Condition	N	Model 1	Model 2
Age Partitions				
Youth	age < 25	305	0.81	0.75
Middle Age	$25 \leq$ age ≤ 55	17936	0.92	0.88
Senior	age > 55	7759	0.91	0.91
Gender Partitions				
Male	gender = male	13481	0.91	0.92
Female	gender = female	12519	0.92	0.85
Medical Partitions				
Has Medical Exemption	med_exemption = 1	12901	0.93	0.90
No Medical Exemption	med_exemption = 0	12719	0.90	0.88
Physical Health Issues	physical_issue = 1	12378	0.92	0.90
No Physical Health Issues	physical_issue = 0	12889	0.91	0.88
Psychological Issues	psych_issue = 1	12273	0.92	0.90
No Psychological Issues	psych_issue = 0	13110	0.91	0.88
Medical Availability Restriction	med_availability = 1	977	0.92	0.91
No Medical Availability Restriction	med_availability = 0	25023	0.92	0.89

Table 1: Partition test accuracy for Model 1 (good) and Model 2 (bad) with sample sizes (N).

- **Age**: Model 1 shows relatively balanced accuracy performance across the different groups. The lower performance observed in the Youth sub-partition is attributed to the small sample size (305) in the partition. Comparatively, Model 2 exhibits a noticably larger performance gaps between the age groups. This would suggest that Model 2 is less robust and has learned to disproportionately misclassify certain age groups.
- **Gender**: Model 1 performs nearly identically for both men and women, showing no divergence between gender groups. This reinforces the idea that removing sensitive gender-related features in our pre-processing step does make the model more ”fair” and ”unbiased”. On the other hand, Model 2 which does not remove any sensitive features as a disproportion of 7% between the two gender groups, indicating that it has learned gender-correlated patterns in the dataset, leading to biased treatment across the two groups.

- **Medical Features**: Across all medical-related partitions, Model 1 remains highly consistent, with only minor variations of 1% between the subgroups. The accuracy of these detections remain above 90% whereas Model 2 does share consistentency across the partitions but accuracy fluctuates between 88% - 91%. The stability of Model 1 compared to Model 2 across the categories suggests a fairer and more reliable model in terms of medically sensitive contexts. It is, however, worth noting that for medical-related features, Model 2 is not necessarily biased; rather, it is simply less accurate and balanced compared to Model 1.

Metamorphic Test Evaluation: TBD

3.3 Independently testing the 'good' and 'bad' model - max 1 page

Based on your tests, can you tell which of the models is the better and worse one? Write a short summary of your main findings.clear Cross-compare and discuss your outcomes with the other subgroup (also learning their testing outcomes of your models).