Credit scoring models are supervised binary classifiers that predict the probability of default. Scoring models are trained on data from previously granted credit applications with the observed repayment behavior. This creates sample bias: a model is trained on the accepted cases only.

Figure 1a demonstrates the acceptance cycle. Figure 1b illustrates the bias after running the acceptance cycle for 300 iterations with synthetic data. Since labels of rejects are not missing at random, the bias degrades model performance when applying it to new applications as shown in Figure 1c.

We make two contributions: 1) Shallow Self-Learning to label rejected cases and mitigate sample bias; 2) Kickout Metric to improve model selection when true labels of rejects are unknown.

**Data description:** Since labels of rejected cases are not missing at random, sample bias degrades model performance when applying it to new applications.

**Scoring Model**
- **REJECTS**
- **GOOD**
- **BAD**

**Training Data**

**Shallow Self-Learning**

**Stage I**
- Filtering
  - Goal: removing rejects whose distribution is different from the accepts
  - **ACCEPTS**
  - **FILTERED REJECTS**

**Stage II**
- Labeling
  - Goal: labeling rejected cases that are classified with high confidence
  - **ACCEPTS**
  - **FILTERED REJECTS**

**Stage III**
- Scoring
  - Goal: training a scoring model on the extended data with smaller bias
  - **ACCEPTS**
  - **LABELLED REJECTS**
  - **NEW APPLICANTS**

**Kickout Metric**

**Intuition:** compare applicants accepted by two models: before and after reject inference

**Calculation:**
- train initial model on **TRAIN** and score cases in **VALID**
- split rejected cases into two subsets: R1 and R2
- use reject inference to label R1
- append labeled R1 to **TRAIN**; append R2 without labels to **VALID**
- train new model on **(TRAIN + R1)** and score cases in **(VALID + R2)**
- count **GOODs** and **BADs** that are “kicked out” - rejected after reject inference

Kicked-out cases are replaced with rejects with unknown labels:
- kicking out a **BAD** case has a positive expected value
- kicking out a **GOOD** case has a negative expected value
- hence, reject inference should aim at kicking out more **BADs** and less **GOODs**

**Figure 2a** depicts performance gains from using shallow self-learning on synthetic data. Compared to the scoring model trained on the unbiased sample, accepts-based model suffers from sample bias resulting in a 0.1 drop in AUC. Reject inference with shallow self-learning increases AUC by 0.05, recovering about 50% of the loss due to bias.

**Figure 2b** shows performance gains from using the kickout metric for model selection. Selecting a scoring model based on its AUC on the accepted cases leads to suboptimal performance on the unbiased holdout sample (AUC of 0.79). Selecting a model based on kickout improves AUC on the unbiased sample by 0.01.

Overall, ranking models based on their AUC on the unbiased sample better correlates with the kickout metric (r = 0.30) than with the AUC on the accepted cases (r = 0.12).

**Real-World Data**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Accepts</th>
<th>Rejects</th>
<th>Unbiased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cases</td>
<td>39,579</td>
<td>18,047</td>
<td>1,967</td>
</tr>
<tr>
<td>Number of features</td>
<td>2,410</td>
<td>2,410</td>
<td>2,410</td>
</tr>
<tr>
<td>BAD rate</td>
<td>0.39</td>
<td>-</td>
<td>0.66</td>
</tr>
</tbody>
</table>

**Performance Gains**

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ignore rejects</td>
<td>0.8007</td>
</tr>
<tr>
<td>Label all rejects as BAD</td>
<td>0.6797</td>
</tr>
<tr>
<td>Bureau score based inference</td>
<td>0.7911</td>
</tr>
<tr>
<td>Hard cutoff augmentation</td>
<td>0.7994</td>
</tr>
<tr>
<td>Parcelling</td>
<td>0.8041</td>
</tr>
<tr>
<td>Shallow Self-Learning + Kickout</td>
<td>0.8072</td>
</tr>
</tbody>
</table>

**Monetary Gains**

**Assumptions:**
- acceptance rate = 30% cases with the lowest score
- average loan size = $17,100, interest rate = 10.36%
- average loss given default = 25%

<table>
<thead>
<tr>
<th>Measure</th>
<th>Ignore rejects</th>
<th>SSL + Kickout</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision at 30%</td>
<td>0.7936</td>
<td>0.7996</td>
<td>0.006</td>
</tr>
<tr>
<td>BAD loans per 10,000 clients</td>
<td>2,064</td>
<td>2,004</td>
<td>60</td>
</tr>
<tr>
<td>Revenue loss</td>
<td>$9,737,725</td>
<td>$9,454,652</td>
<td>$283,073</td>
</tr>
</tbody>
</table>