Optimization Aware Active Learning

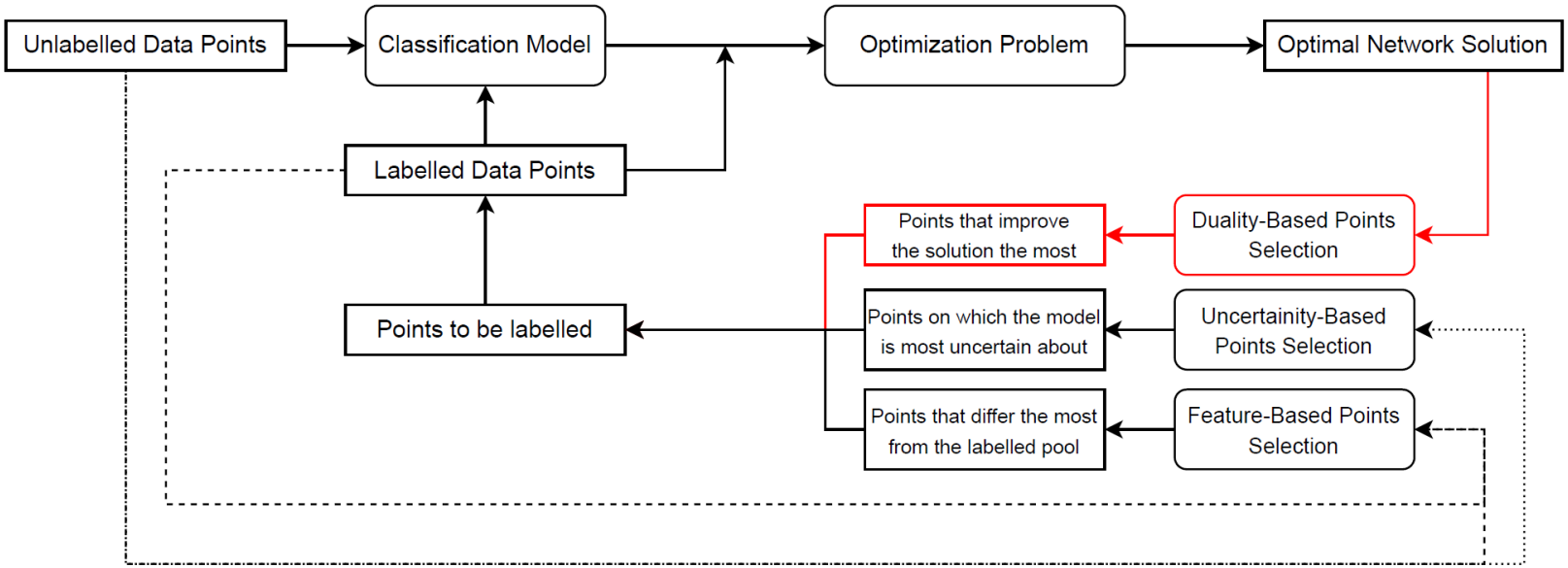
Relocation of Red Pandas in more Suitable Habitats

MIT MANAGEMENT BUSINESS ANALYTICS

Sara Pasquino, Giorgio Demarchi

Framework

General Framework



Optimization Model

- Network Flow Optimization model

$$\min \quad \sum_i \sum_j c_{ij} x_{ij} \tag{1}$$

$$\text{s.t.} \quad [\ldots] \tag{2}$$

$$\sum_j x_{ij} \leq z_i M \quad \forall i \tag{3}$$

$$x_{ij} \geq 0 \quad \forall (i,j) \tag{4}$$

- z_i is the **output** of the **classification model** and an **array of nodes** in the optimization problem
- (2) are classic **Network Flow constraints** (demand, supply, capacity)
- (3) is a **linking constraint**: constrains to 0 inflow to nodes z_i classified as 0

Optimization-Aware Heuristic

Optimization-Aware Heuristic

- Split the Classifier output z_i into two sets:

$$O = \{z_i \in \hat{\mathbf{z}} \mid z_i = 0\}$$
$$I = \{z_i \in \hat{\mathbf{z}} \mid z_i = 1\}$$

- Select the $z_i \in O$ with the highest associated **shadow price**

$$U = \{z_i \in O \mid p_i \geq \bar{p}\} \quad (5)$$

- Select the $z_i \in I$ with the **highest inflow** $\sum_j x_{ij}$

$$L = \{z_i \in I \mid q_i \geq \bar{q}\} \quad (6)$$

- Label $U \cup I$

Experiments and results

Performance Metrics

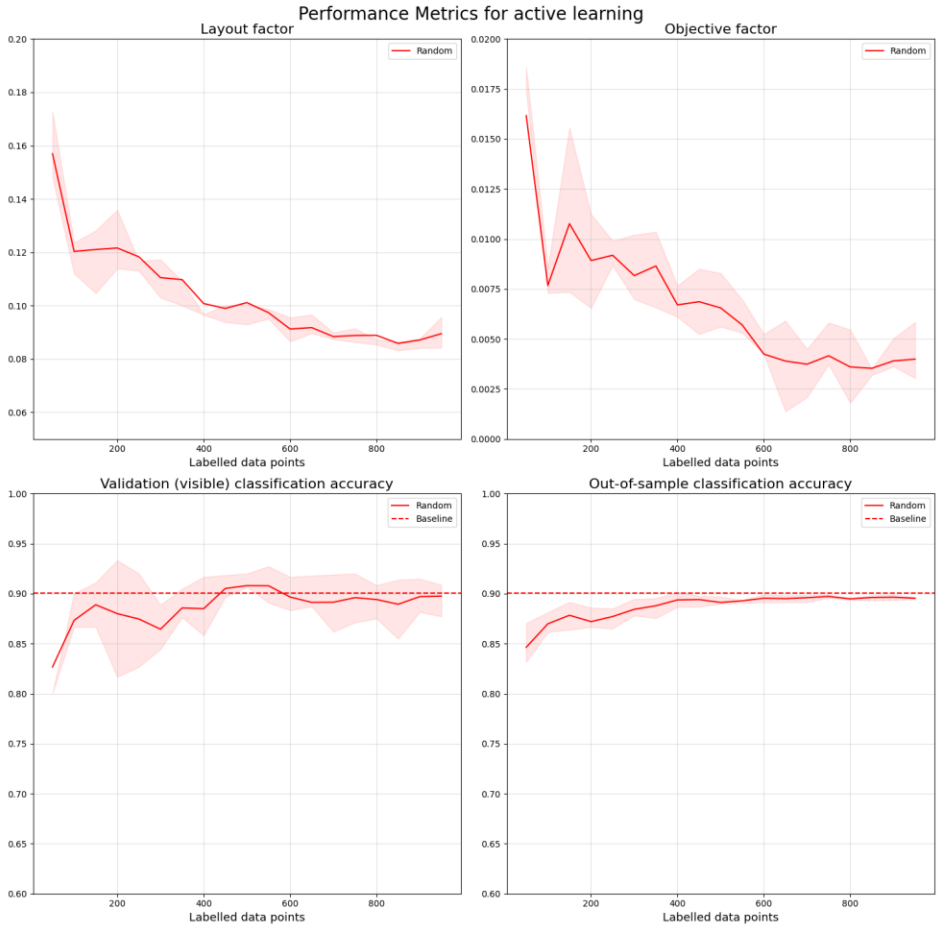
- Validation Classification Accuracy
  - Accuracy in the **Labelled Points**
- Out of Sample Classification Accuracy
  - Accuracy in the **Non-Labelled Points**
- Layout Factor
  - Absolute Deviation from Full-Information **Optimization Framework**

$$LF = \frac{\|\hat{X} - X_{FI}\|_1}{2 \cdot S} \tag{17}$$

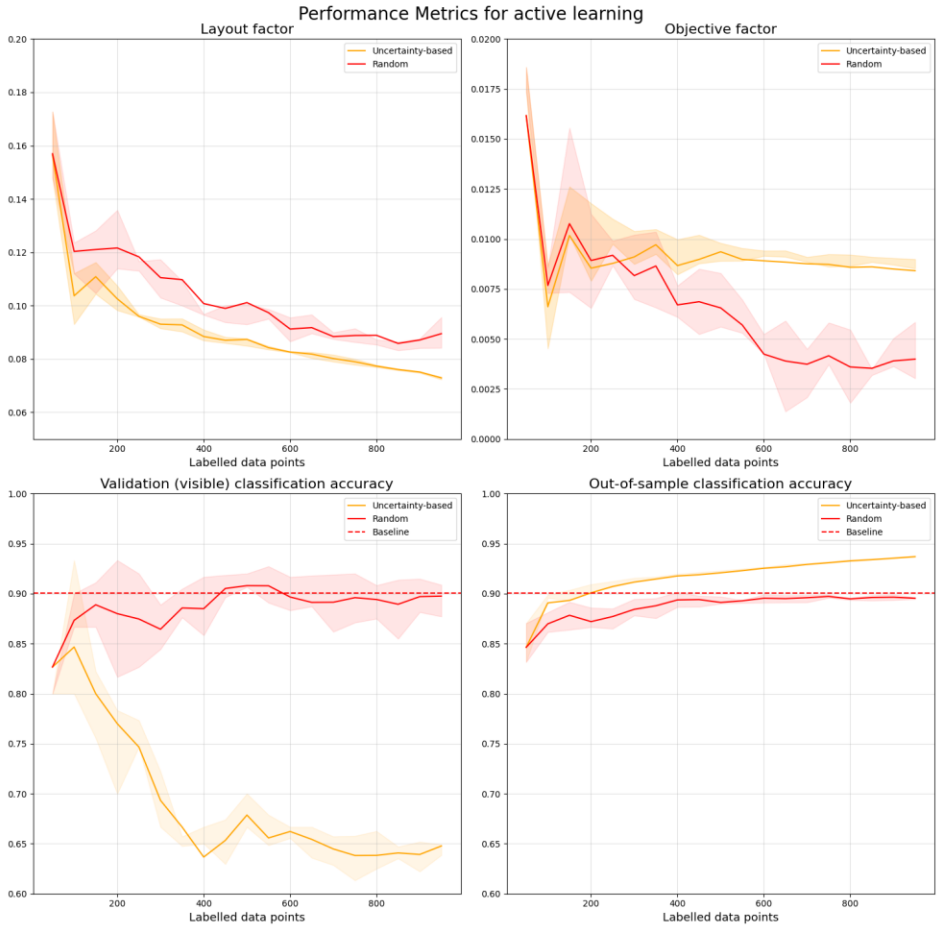
- Objective Factor
  - Absolute Deviation from Full-Information **Objective Value**

$$OF = |1 - \frac{Z^*_{FI}}{\hat{Z}^*}| \tag{18}$$

Baseline: Random Batch Selection



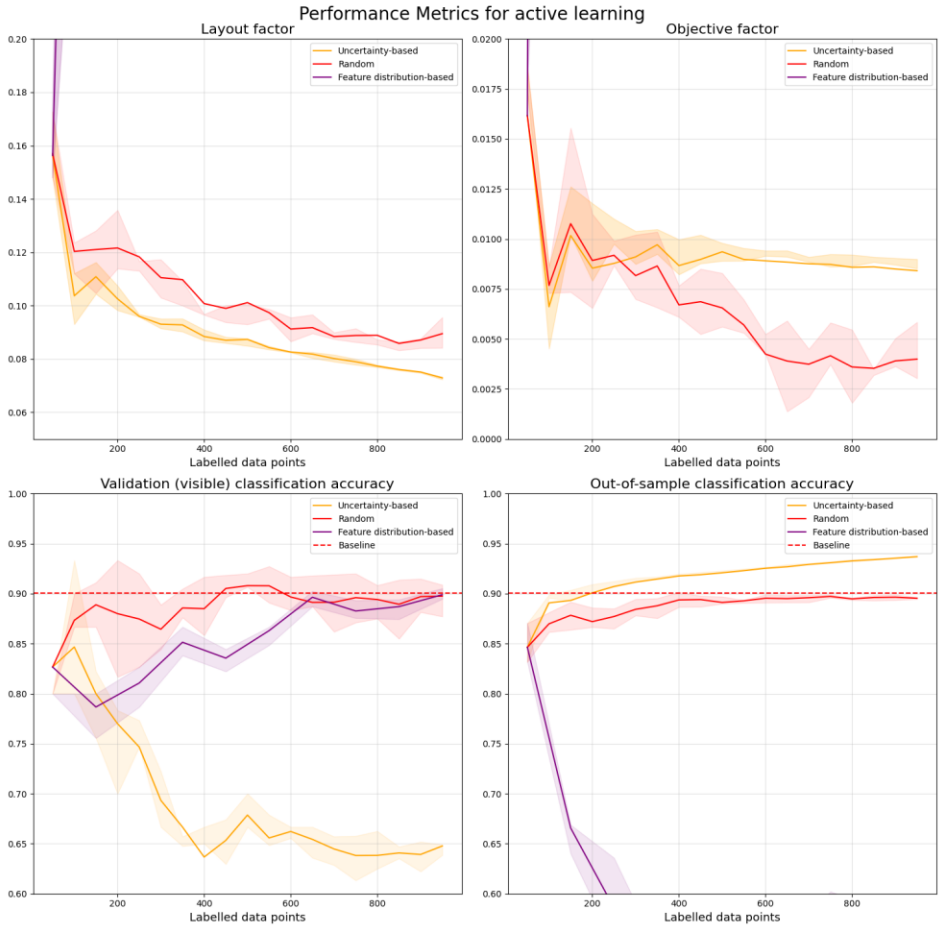
Uncertainty-Based AL



- Strong in layout factor as confirmed by performance in out-of-sample accuracy
- Bad in objective factor: it classifies most points correctly, but it misses the important ones

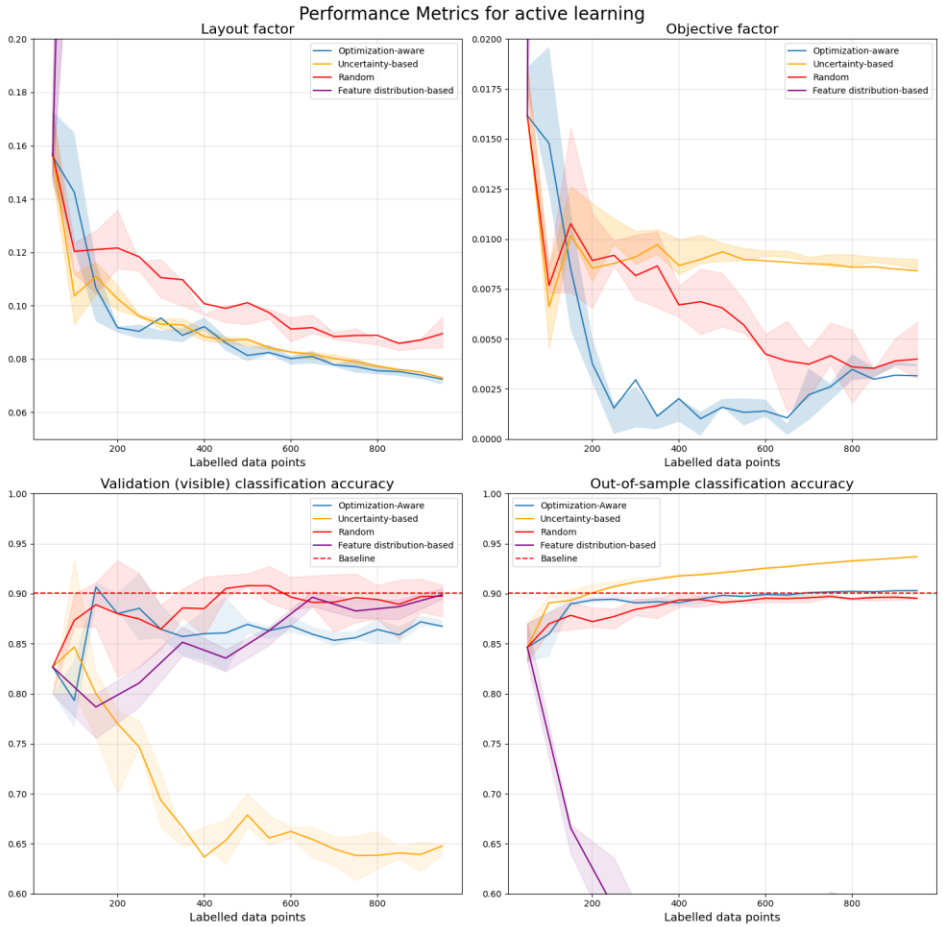
- Strong performance in out-of-sample accuracy
- Very weak in validation accuracy (as it trains on weak points)

Feature Distribution-Based AL



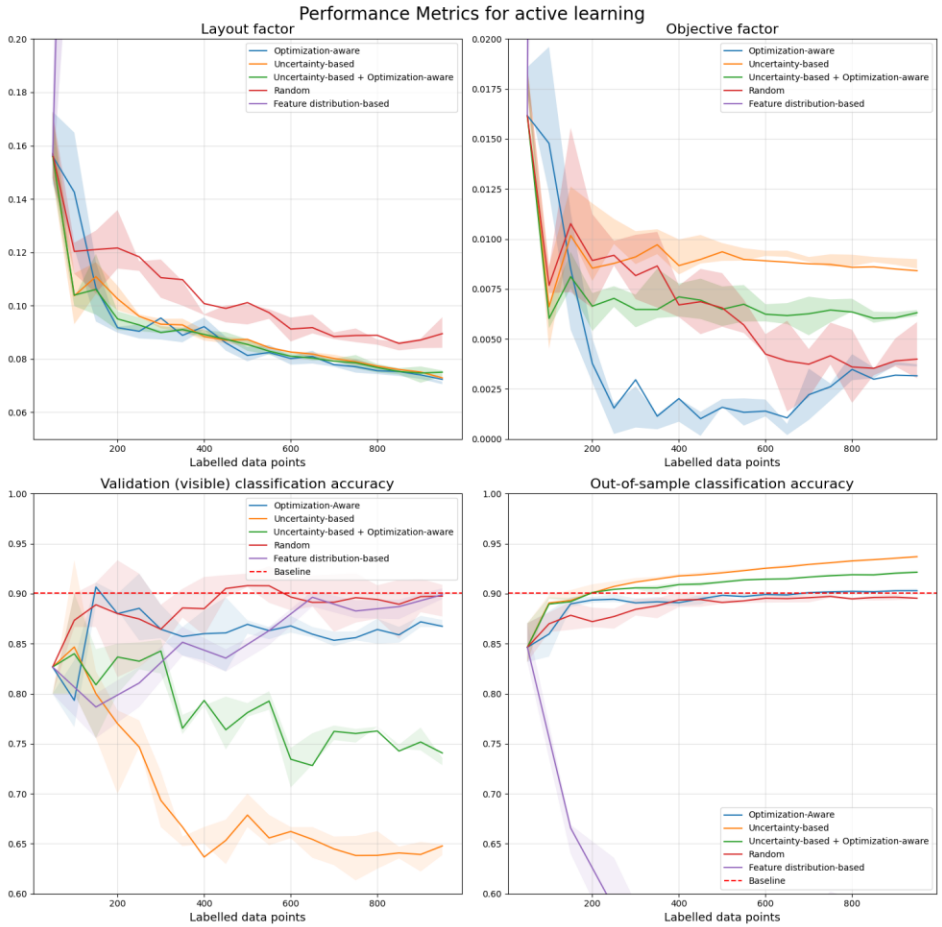
- Optimization problem becomes infeasible after few iterations with feature selection
- Terrible out-of-sample performance explains infeasibility
- Explanation as to why this is the case require further investigation

Optimization-Aware AL



- Same layout factor performance as uncertainty-based AL, despite lower out-of-sample accuracy
  - Less accurate than uncertainty, but accurate on most important points
- Great performance in objective factor: labels correctly the most important points
- Worse than uncertainty-based AL in out-of-sample accuracy, but better than random
- Better in "visible" / validation performance

Uncertainty + Optimization AL



- Averages the performance of the uncertainty-based and the optimization-aware method in all four KPIs
- Best of both worlds

Thank you!