Transfer Learning to Overcome Domain Shift in Football Analytics and Beyond

Luca Clissa, Antonio Macaluso
luca.clissa2@unibo.it

«Football laboratory»

Why football data?

«Football laboratory»

Why football data?

- Football data provide interesting learning opportunities
- Task: probability of scoring given sequence of actions
- Challenges:
  - Hard task
    - Extreme class imbalance (low score sport)
    - Complex interactions among features
  - Noisy labels
    - External (non-measurable) factors have an impact (e.g. pressure, form status, luck, ...)
  - Domain shift
    - Models break across competitions (season/leagues effect)

→ Very rich use case for stress-testing methodology

Domain shift in data analysis

Domain shift: how to leverage source knowledge on target data?

- X: input features, Y: target variable
- Source domain: training, sufficient available data
- Target domain: test/new data, typically no/few labels

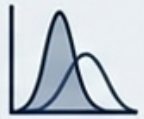
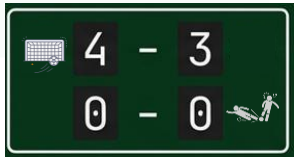
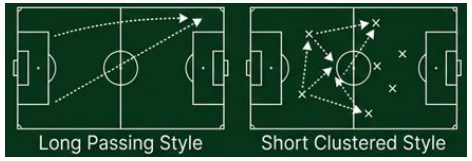
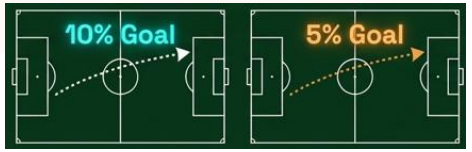
theory

Prior shift, P(Y)	Covariate shift, P(X)	Concept shift, P(Y|X)
Probability of target variable changes: $P_{target}(Y) \neq P_{source}(Y)$ 	Distribution of features changes: $P_{target}(X) \neq P_{source}(X)$ 	Relationship between features and target changes: $P_{target}(Y\|X) \neq P_{source}(Y\|X)$ 

Domain shift in data analysis

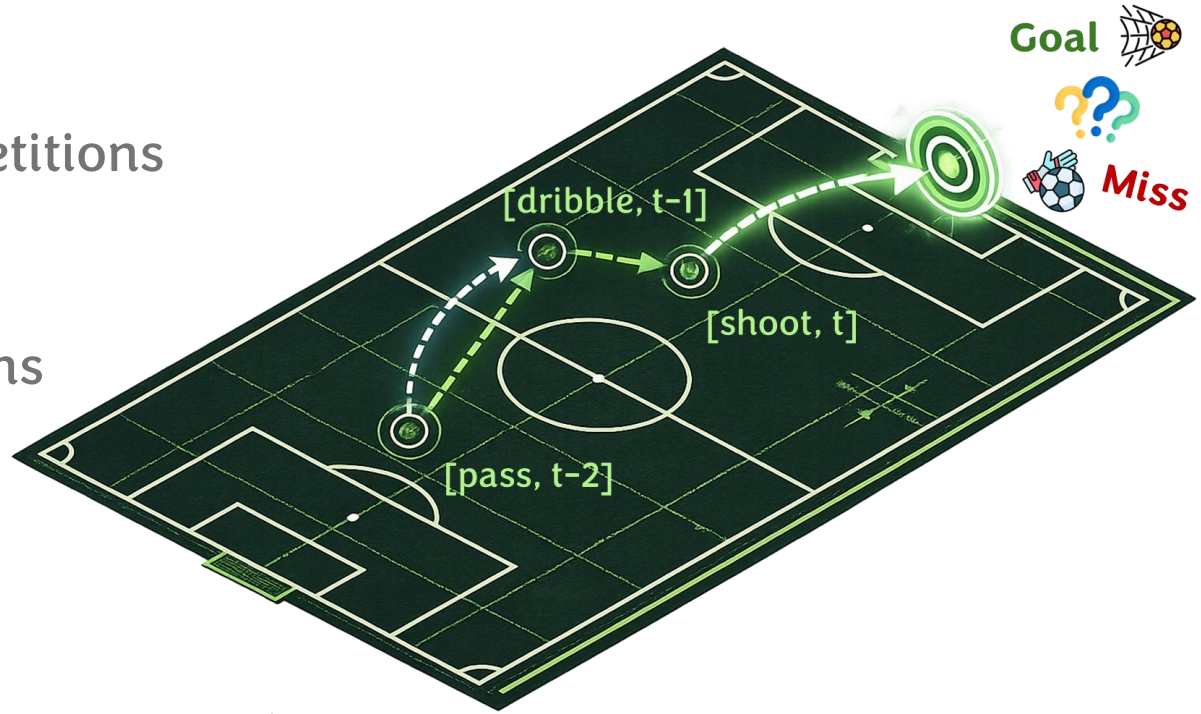
Domain shift: how to leverage source knowledge on target data?

- X: input features, Y: target variable
- Source domain: training, sufficient available data
- Target domain: test/new data, typically no/few labels

	Prior shift, P(Y)	Covariate shift, P(X)	Concept shift, P(Y|X)
theory	Probability of target variable changes: $P_{target}(Y) \neq P_{source}(Y)$ 	Distribution of features changes: $P_{target}(X) \neq P_{source}(X)$ 	Relationship between features and target changes: $P_{target}(Y\|X) \neq P_{source}(Y\|X)$ 
football	Probability of scoring changes across *league, season, or tournament types* (e.g. world cup VS Serie A) 	Playing style changes across *leagues,* and tactics evolve across *seasons* (e.g. long balls VS tiki-taka) 	Probability of scoring with given action changes based on competition-specific dynamics (e.g. difensive proficiency) 

Goal

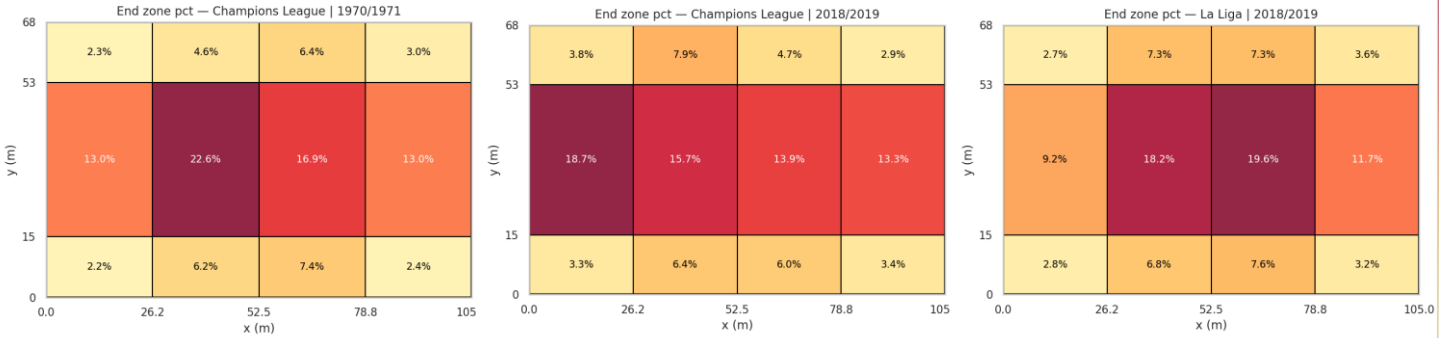
- Leverage StatsBomb open data
  - X leagues, including world-wide competitions
  - Several seasons per each league
  - Mixture of event and tracking data
  - SPADL conversion: focus only on actions

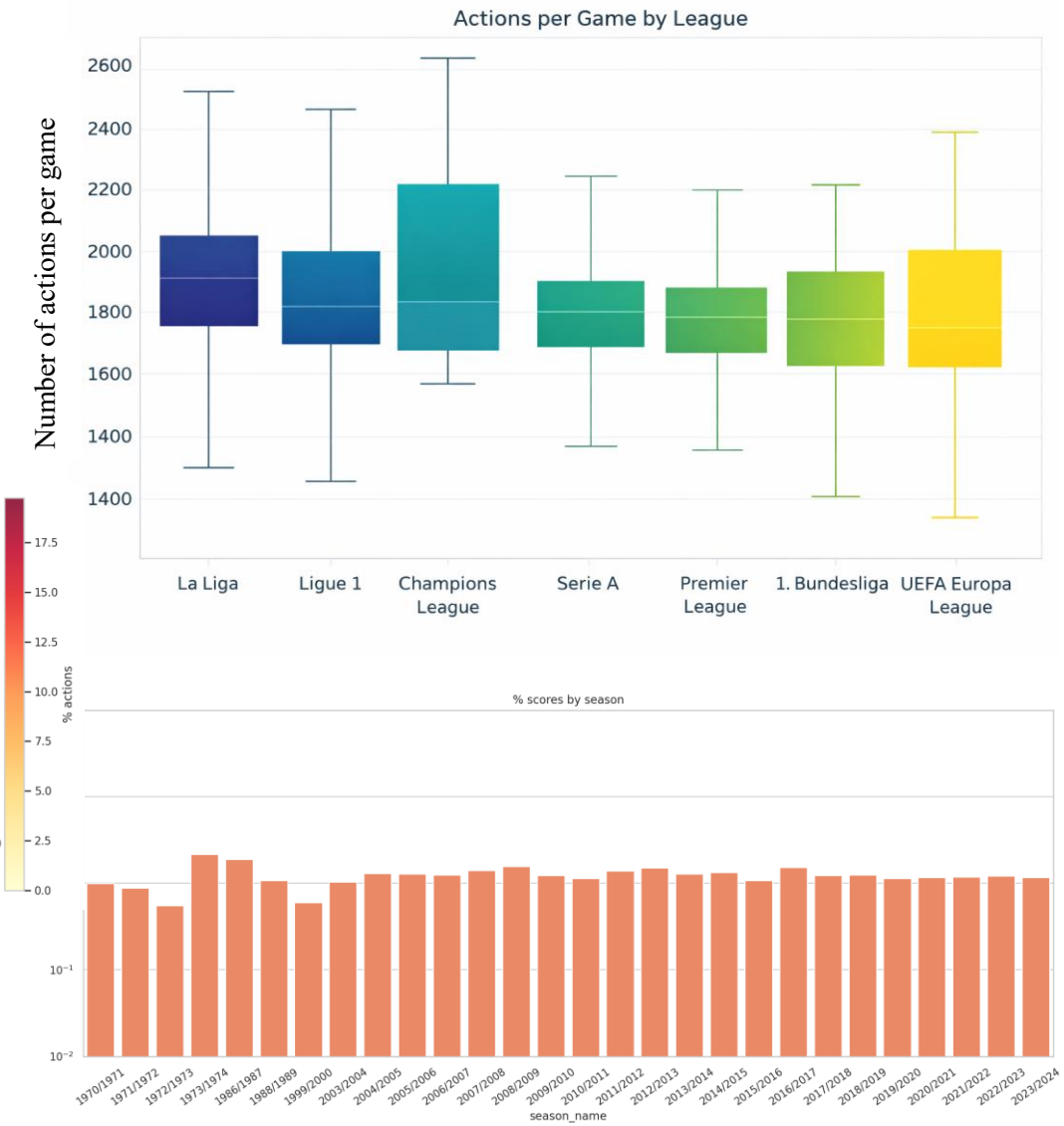


- Task: $P(score \mid a_t, a_{t-1}, a_{t-2}) = f(X_t, X_{t-1}, X_{t-2})$
- Explore transfer learning across competitions

Data visualization

- Playing style changes
  - By league
  - By season





Experimental design

Split into:

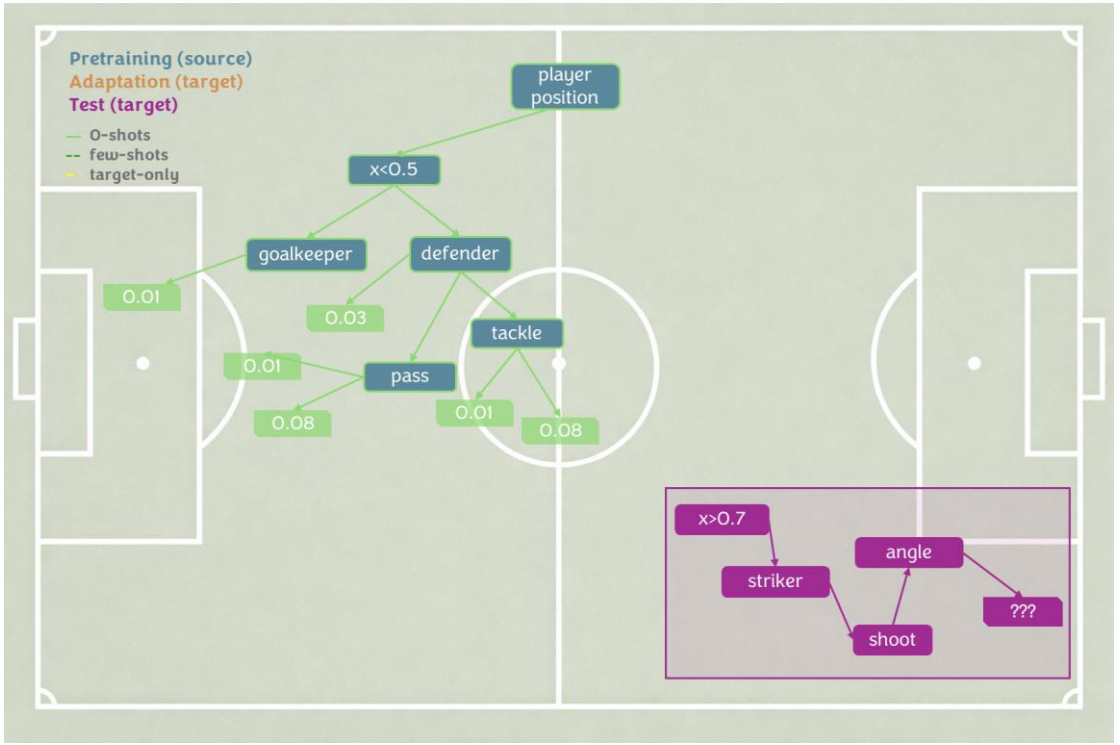
- **Source leagues**: European competitions
- **Target leagues:** i) other continents, ii) women's competition, iii) national teams
  - 2 splits: **adaptation set**, **test set**

Model: xgboost

- Strong baseline for tabular data
- Can be fine-tuned

Settings

- **0-shot:** pretrain on source, evaluate on test



Experimental design

Split into:

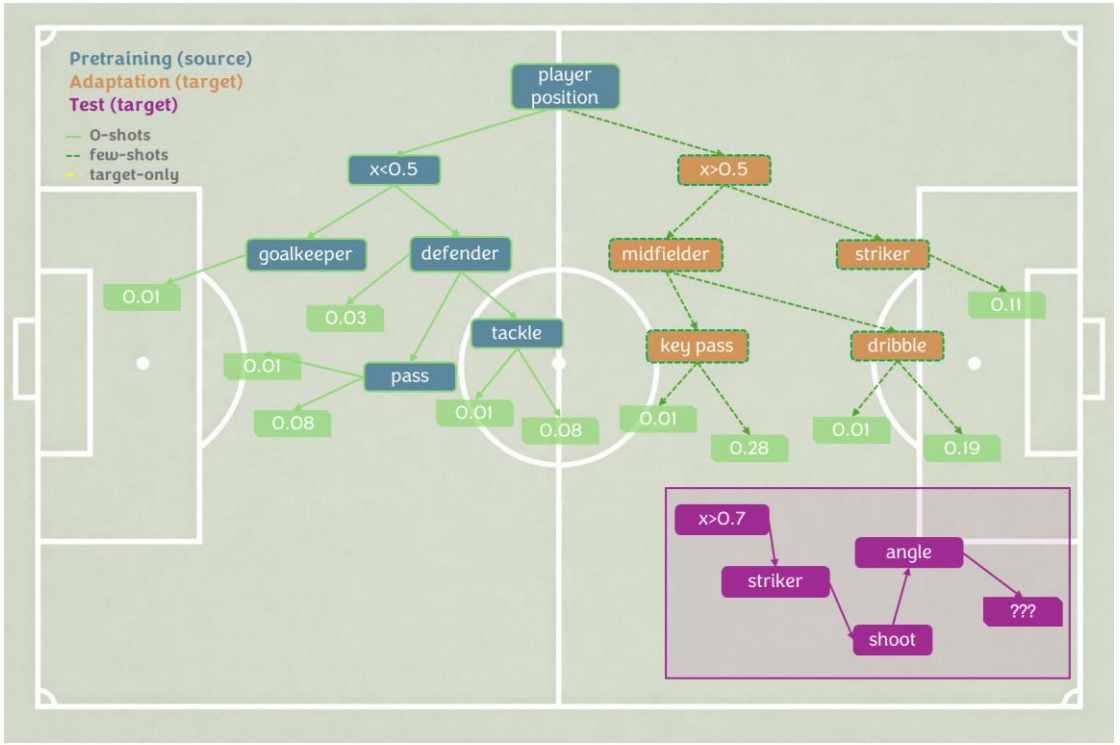
- **Source leagues**: European competitions
- **Target leagues:** i) other continents, ii) women's competition, iii) national teams
  - 2 splits: **adaptation set**, **test set**

Model: xgboost

- Strong baseline for tabular data
- Can be fine-tuned

Settings

- **0-shot:** pretrain on source, evaluate on test
- **Few-shots:** fine tuned on adaptation set, evaluate on test



Experimental design

Split into:

- **Source leagues**: European competitions
- **Target leagues:** i) other continents, ii) women's competition, iii) national teams
  - 2 splits: **adaptation set**, **test set**

Model: xgboost

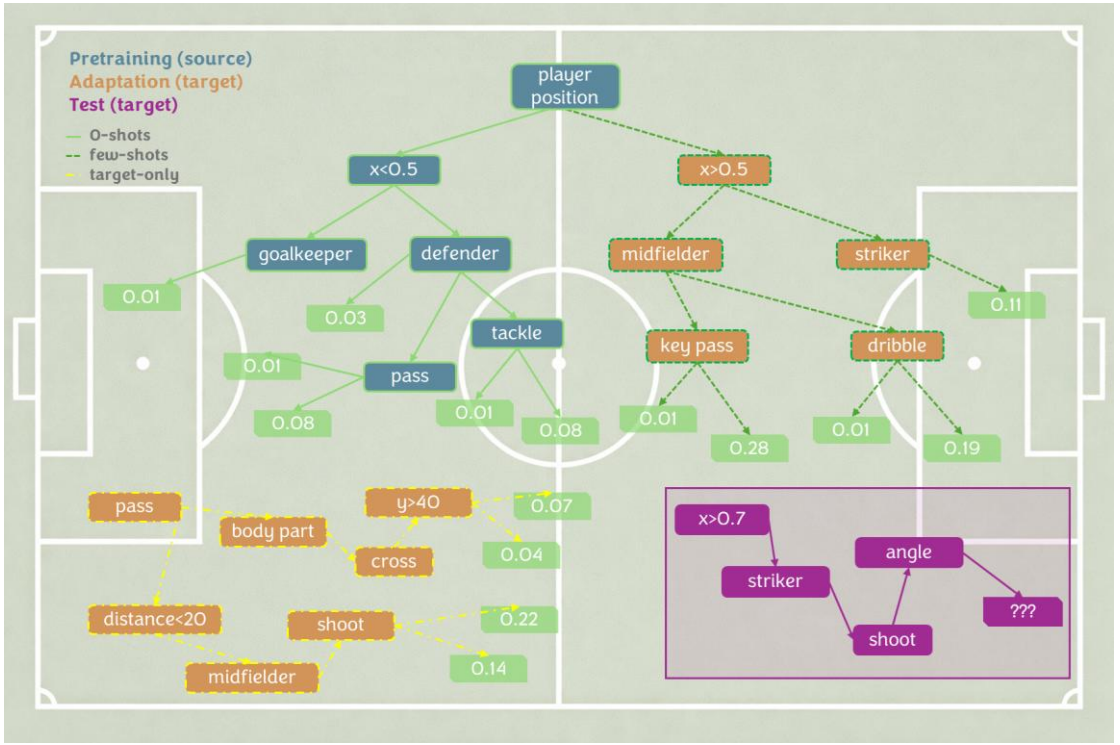
- Strong baseline for tabular data
- Can be fine-tuned

Settings

- **0-shot:** pretrain on source, evaluate on test
- **Few-shots:** fine tuned on adaptation set, evaluate on test
- **Target-only:** train on adaptation set, evaluate on test

Ablation studies

- 3 adaptation label budgets: 15%, 30% and 50%

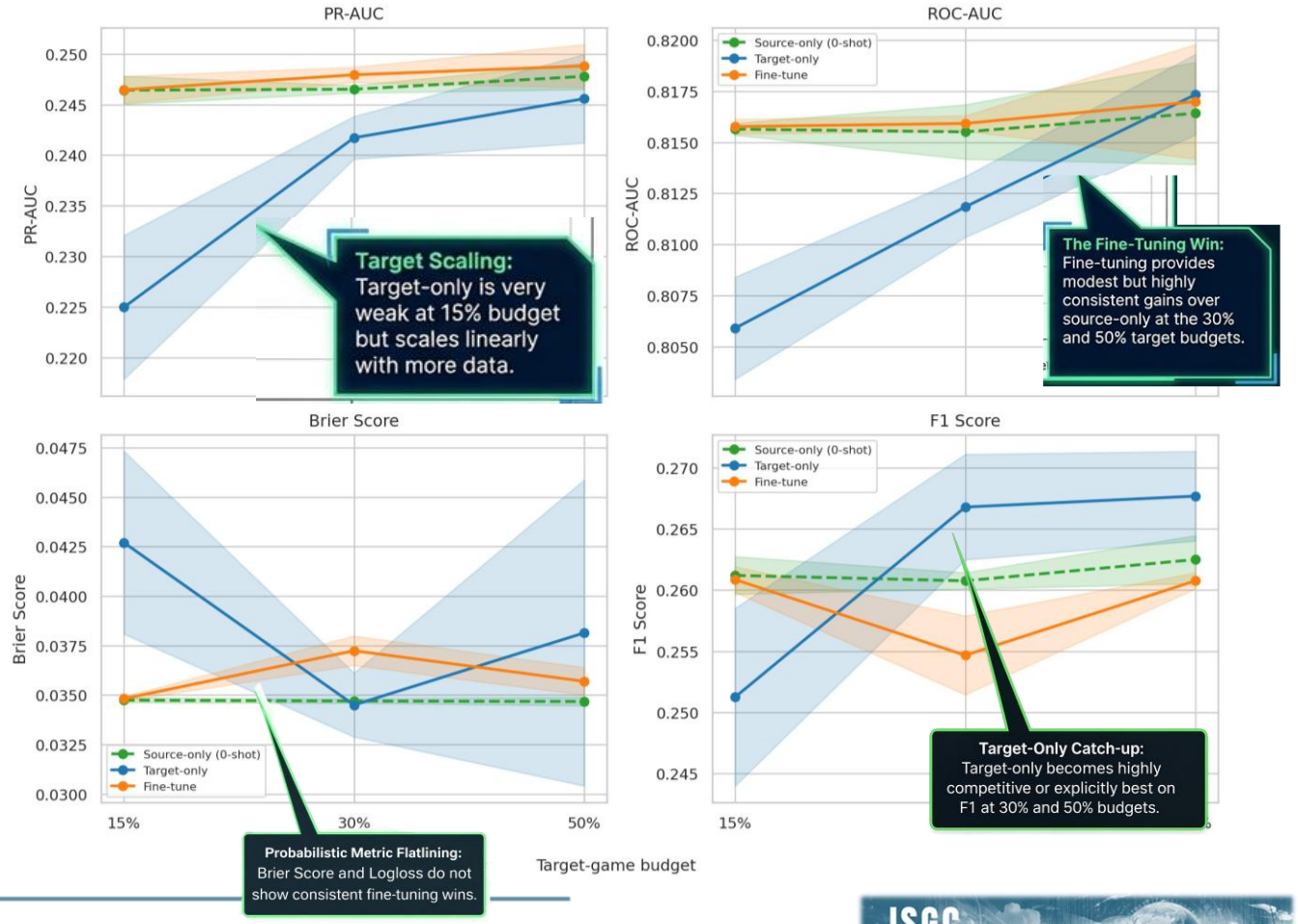


Preliminary results

- Metrics are not great → hard task!

	Positives (%)	Precision	Recall	F1	AUC	PR-AUC
Validation (in-sample)	~0.012	0.3780	0.2568	0.3059	0.8917	0.2833

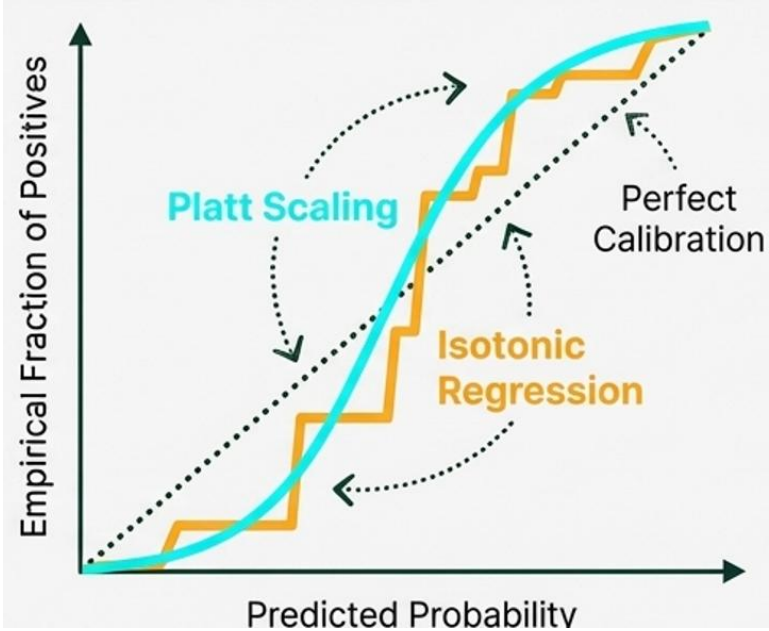
- Ranking signal: AUC, PR-AUC
  - few-shot gains are modest but consistent
  - More labels help target-only to catch up
- Classification: f1 score
  - Transfer does not help with classification
  - Target-only becomes highly competitive for larger label budgets
- Calibration: Brier score
  - Predicted probabilities far from observed frequencies (baseline ~0.0099 for 1% prevalence)



Conclusions & next steps

- Rich use case for transfer learning / domain adaptation studies
- Hard task: difficult to get nice metrics
  - Class imbalance
  - Noisy labels

- Promising results for fine-tuning evidence
- Future work:
  - More diagnostics: measure covariate shift via domain classifier
  - Calibrate domain shift directly (e.g. platt scaling, isotonic regression)
  - Exploration of more architectures and fine-tuning techniques



Questions?

Contacts: luca.clissa2@unibo.it

References

[1] StatsBomb (2019), "StatsBomb Open Data". Available at: https://github.com/statsbomb/open-data

[2] Decroos T, Bransen L, Van Haaren J, Davis J (2019), "Actions Speak Louder than Goals: Valuing Player Actions in Soccer". Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), pp. 1851–1861, doi: https://doi.org/10.1145/3292500.3330758

[3] Pan SJ, Yang Q (2010), "A Survey on Transfer Learning". IEEE Transactions on Knowledge and Data Engineering, Vol. 22 No. 10, pp. 1345–1359, doi: https://doi.org/10.1109/TKDE.2009.191

[4] Chen T, Guestrin C (2016), "XGBoost: A Scalable Tree Boosting System". Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), pp. 785–794, doi: https://doi.org/10.1145/2939672.2939785

[5] Niculescu-Mizil A, Caruana R (2005), "Predicting Good Probabilities with Supervised Learning". Proceedings of the 22nd International Conference on Machine Learning (ICML), pp. 625–632, doi: https://doi.org/10.1145/1102351.1102430

[6] Platt J (1999), "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". Advances in Large Margin Classifiers, MIT Press, pp. 61–74

Backup

Budget evaluation matrix

- Target-only struggles under strict data constraints (15%)
- Fine-tuning consistently edges out source prior in ranking

Model	Budget	target	PR-AUC	ROC-AUC	BRIER	F1	LOGLOSS
Source-only (0-shot)	15%		0.2464 ± 0.0014	0.8156 ± 0.0003	0.0347 ± 0.0001	0.2612 ± 0.0015	0.1565 ± 0.0004
Target-only	15%		0.2250 ± 0.0071	0.8059 ± 0.0025	0.0427 ± 0.0046	0.2513 ± 0.0073	0.2179 ± 0.0236
Fine-tune	15%	(few-shots)	0.2465 ± 0.0013	0.8158 ± 0.0004	0.0348 ± 0.0001	0.2609 ± 0.0011	0.1567 ± 0.0003
Source-only (0-shot)	30%		0.2465 ± 0.0004	0.8155 ± 0.0013	0.0347 ± 0.0001	0.2608 ± 0.0007	0.1564 ± 0.0004
Target-only	30%		0.2417 ± 0.0021	0.8118 ± 0.0015	0.0345 ± 0.0016	0.2668 ± 0.0043	0.1626 ± 0.0098
Fine-tune	30%	(few-shots)	0.2480 ± 0.0008	0.8159 ± 0.0004	0.0372 ± 0.0007	0.2547 ± 0.0032	0.1634 ± 0.0029
Source-only (0-shot)	50%		0.2478 ± 0.0013	0.8164 ± 0.0025	0.0347 ± 0.0002	0.2625 ± 0.0020	0.1563 ± 0.0007
Target-only	50%		0.2456 ± 0.0044	0.8173 ± 0.0020	0.0381 ± 0.0077	0.2677 ± 0.0037	0.1837 ± 0.0455
Fine-tune	50%	(few-shots)	0.2488 ± 0.0021	0.8170 ± 0.0028	0.0357 ± 0.0007	0.2608 ± 0.0006	0.1595 ± 0.0023