STRATOS theme 1 ideas

Combining machine learning with statistical methods

STRATOS Lorenz meeting
2024

UMC Utrecht

Overview ideas: Combining machine learning with statistical methods

- Model stacking/ensembling: Ensemble methods are commonly used in ML. They have the potential to improve prediction accuracy and robustness by combining the strengths of multiple models.
  - **Aim:** Review methods and provide practical guidance and demonstrate performance across a few selected case studies. Focus on predicting disease outcomes using healthcare datasets also comparing ensemble methods with single-model approaches.
- Machine learning for survival models: Comparison of ML methods with traditional methods.
  - **Aim:** Create an overview of available methods for survival data in the field of ML (Random forest, boosting, neural networks, regularized regression, SVM) and comparison with traditional regression models. Possible considerations on more complex settings about competing risk, dynamic prediction, multi-state models and the use of pseudo-values.
- Transformers:
  - **Aim:** Accessible explanation on transformer models for longitudinal data. Together with statistical approach close in spirit: joint models. Data for illustration: recurrent events (or multi state) with longitudinal measurements in between (>300 obs, >5 time points).
- Local fine-tuning vs external valildation: It is becoming increasingly common for prediction models to be pre-trained and then fine-tuned before implementation (for instance, developed in Hospital A and B, fine-tuned to “fit” in hospital C). This movement fits with a tendency to increasingly complex models that tend to not generalize well. Traditional thoughts on external validation and generalizability/transportability makes little sense in this setting.
  - **Aim:** Perspective paper discussing alternatives that quantify the value of fine-tuning versus external validation.
- Fairness for prediction models: Fairness has been acknowledged as an important problem for prediction models by both statistical (e.g., TRIPOD-AI, PROBAST-AI) and computer science fields. Methods have been proposed for identification and mitigation.
  - **Aim:** Review methods for identification and mitigation of fairness for prediction models.
- ML goes to explainability: In applied prediction model research the use of explainable AI techniques seems to be on the rise, perhaps in reaction to claims that ML/AI is “black-box” in nature. Popular approachers are SHAP and LIME values.
  - **Aim:** Discuss the limitations of these approaches: In terms of interactions/non-linearities, in terms of possible causal claims, and confidence intervals.

UMC Utrecht

Model stacking / ensemble

- **Rationale:** Ensemble methods are commonly used in ML. They have the potential to improve prediction accuracy and robustness by combining the strengths of multiple models.
- **Methods:**
- Model Averaging: Combine independent model outputs to reduce variance.
  - Each model predicts independently, and predictions are averaged.
  - Common strategies: equal weighting, performance-based weighting, Bayesian model averaging.
- Model Stacking: Base models generate predictions, which are combined by a meta-model.
  - Meta-model learns how to best combine model outputs for optimal accuracy.
  - Requires cross-validation to prevent overfitting in the meta-model.
- Super Learners: Cross-validated ensemble optimising model combination.
  - Can include diverse algorithms (e.g., regression, decision trees, neural networks) for a more comprehensive solution.
  - Assigns optimal weights to models, minimising prediction error.
- **Aim:** Review methods and provide practical guidance and demonstrate performance across a few selected case studies.
- Focus on predicting disease outcomes using healthcare datasets also comparing ensemble methods with single-model approaches.
- **Expected Outcomes:**
- A level 2 practical guidance for implementing ensemble methods.
- **People involved:** Aris (lead), Federico, Cécile, Thomas

Machine learning for survival models

Proposal: This could be a collaboration with topic group 8. Comparison of ML methods with traditional methods.

Overview of available methods for survival data in the field of ML (Random forest, boosting, neural networks, regularized regression, SVM) and comparison with traditional regression models. Possible considerations on more complex settings about competing risk, dynamic prediction, multi-state models and the use of pseudo-values.

While most works in this area primarily concentrate on metrics such as AUC and specific loss functions, calibration and accuracy are often overlooked. Variance estimation is usually ignored.

Case studies on a publicly available datasets (e.g. colon cancer, BCSG and Rotterdam data for training and validation, pbc, melanoma data, ...) will be provided.

People involved: Malka (lead), Federico, Cécile, Ben/Laure

Transformers

Proposal:

- Accessible explanation on transformer models for longitudinal data
- Together with statistical approach close in spirit: joint models
- Data for illustration: recurrent events (or multi state) with longitudinal measurements in between (>300 obs, >5 time points)

Aspects for comparison:

1) How to re-code data for the models (sequence of tokens for transformers; few event types for time-to-event)?

2) Advantages/disadvantages with missing information (e.g. interval censoring) and measurement errors (Does the transformer embedding automatically take care of this?)

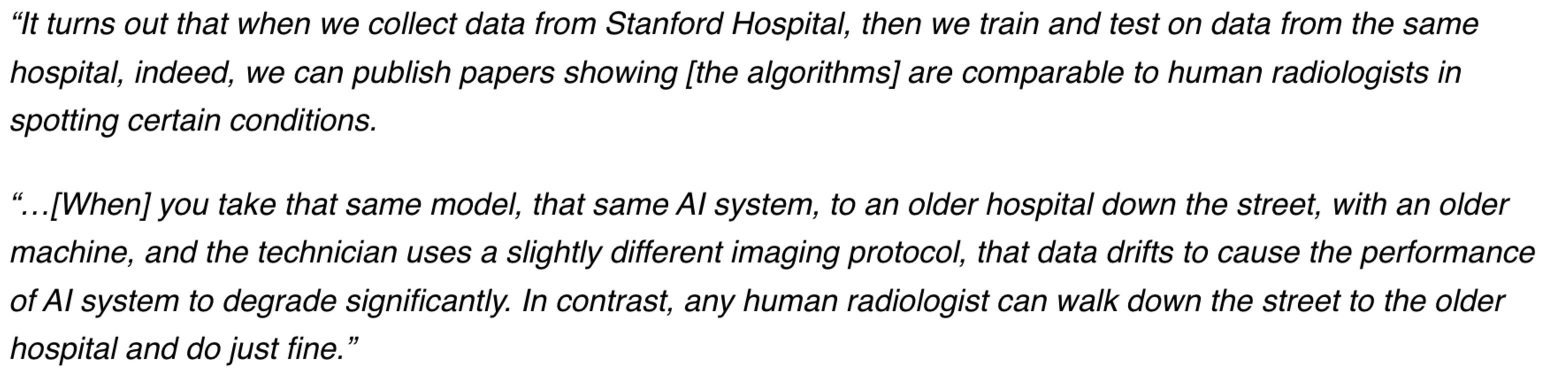
3) How do approaches deal with strong trends?

4) How are approaches affected by small sample size? Power gain by self-supervised learning?

People involved: Harald (lead), Cécile, Thomas

Local fine-tuning vs external validation

- It is becoming increasingly common for prediction models to be **pre-trained** and then **fine-tuned** before implementation (for instance, developed in Hospital A and B, fine-tuned to “fit” in hospital C)
- This movement fits with a tendency to increasingly complex models that tend to not generalize well



“It turns out that when we collect data from Stanford Hospital, then we train and test on data from the same hospital, indeed, we can publish papers showing [the algorithms] are comparable to human radiologists in spotting certain conditions.

“…[When] you take that same model, that same AI system, to an older hospital down the street, with an older machine, and the technician uses a slightly different imaging protocol, that data drifts to cause the performance of AI system to degrade significantly. In contrast, any human radiologist can walk down the street to the older hospital and do just fine.”

- Traditional thoughts on **external validation** and **generalizability/transportability** makes little sense in this setting
- Alternatives that quantify the **value of fine-tuning** are proposed, but what do they tell us?
- STRATOS level 1/2 paper? Close connection to **TG6**.

People involved: Maarten (lead), Ben, Ewout, Aris, Harald, Anne, Laure

Fairness for prediction models

Rationale: Fairness has been acknowledged as an important problem for prediction models by both statistical (e.g., TRIPOD-AI, PROBAST-AI) and computer science fields. Methods have been proposed for identification and mitigation.

Aim: Review methods for identification and mitigation of fairness for prediction models.

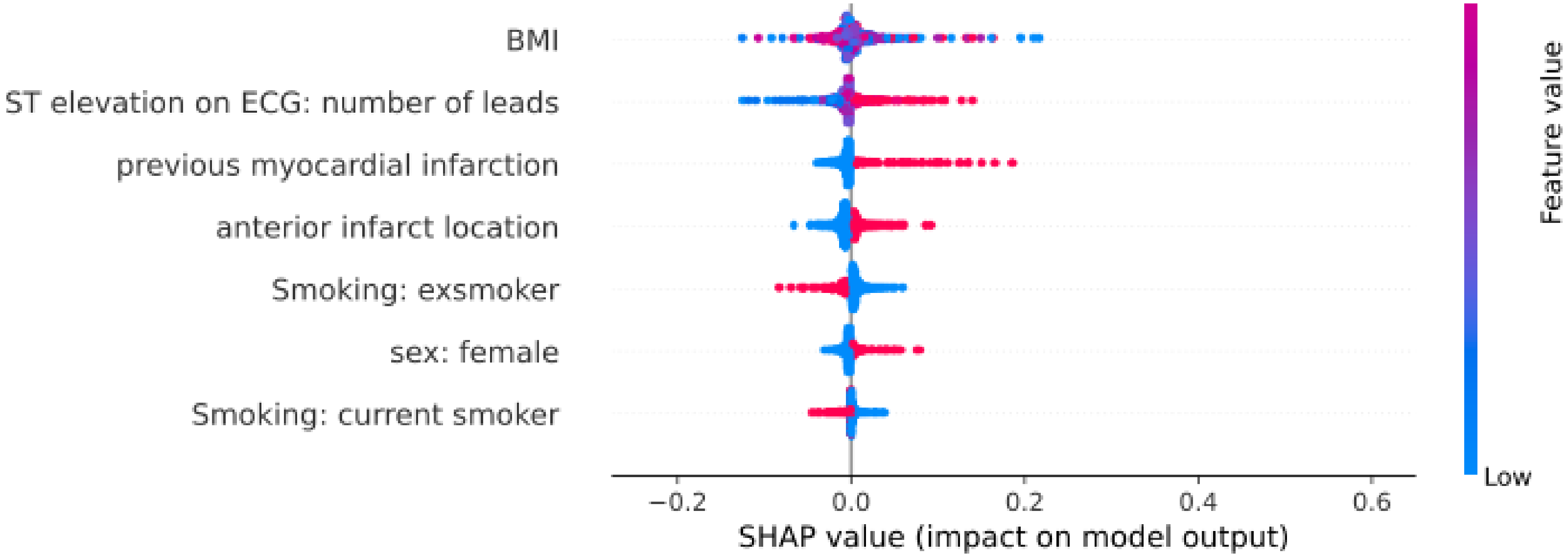
- Give broad overview of most popular methods (level 1).
- Provide guidance on how to identify bias (and how not to) (level 1).
- Illustrate the trade-off when applying different mitigation methods for discrimination/calibration (level 1).

People involved: Anne (lead), Laure, David van Klaveren, Ewout, Maarten, Ben, Tina

ML goes to explainability

- In applied prediction model research the use of explainable AI techniques seems to be on the rise, perhaps in reaction to claims that ML/AI is “black-box” in nature
- Popular approachers are SHAP and LIME values
- There are clear limitations what these approaches can achieve
  - In terms of interactions/non-lin
  - In terms of possible causal claims
  - Confidence intervals
- Level 1 paper? Connections to **TG2 and TG7**?

People involved: Maarten (interim lead), Anne, Ewout, Thomas, Cécile, Frederico



MC Utrecht