Brainteaser

CLEF 2023

Notebook for the iDPP Lab on Intelligent Disease Progression Prediction

Baseline Machine Learning Approaches to Predict Multiple Sclerosis Disease Progression

SBB team, UNIPD
Barbara Di Camillo
Alessandro Guazzo
Isotta Trescato
Enrico Longato
Erica Tavazzi
Martina Vettoretti

Alessandro Guazzo, PhD student
alessandro.guazzo@phd.unipd.it

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA101017598

iDPP@CLEF 2023 – Aims and methods

Task 1
Predict risk of **disease worsening** in MS

Task 2
Predict **cumulative probability of worsening** in MS

Disease worsening is defined in two ways for as many sub-tasks

Sub-task a
The patient **crosses the threshold EDSS ≥ 3** at least twice within a one-year interval

Sub-task b
EDSS worsening with respect to the first recorded value according to current **clinical practice guidelines**

iDPP@CLEF 2023 – Aims and methods

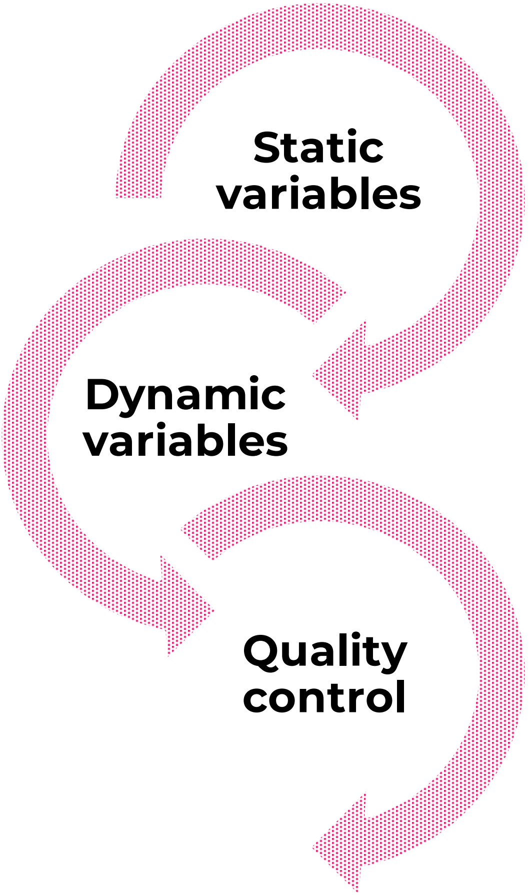
SURVIVAL ANALYSIS
(model time-to-event)

- ▷ Cox proportional-hazards model (Cox)
- ▷ Survival Support Vector Machines (SSVM)
- ▷ Random Survival Forest (RSF)

Outline

- ▷ **Preprocessing**
- ▷ **Model development framework**
- ▷ **Outcomes**
  - Models' outcome
  - Task 1 outcome
  - Task 2 outcome
- ▷ **Results**
  - Task 1 results
  - Task 2 results
- ▷ **Conclusion**

Preprocessing

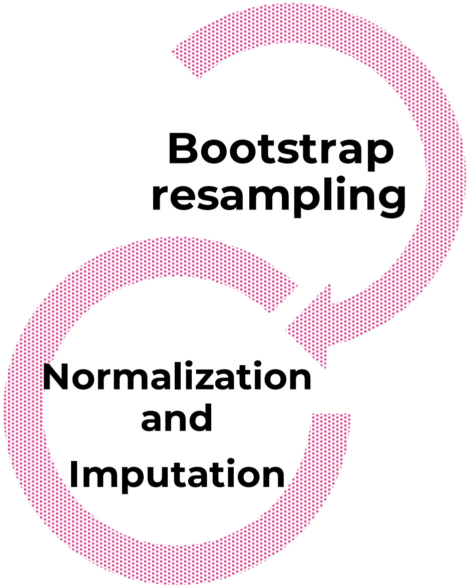


- ▷ ***Sex*** and ***centre*** mapped to binary variables
- ▷ **Residence** mapped to two dummy variables
- ▷ Only **Caucasian** subjects considered

- ▷ **EDSS:** min, max, first, and last values considered
- ▷ **Evoked potentials:** auditory, somatosensory, and visual
- ▷ **MRI measurements:** T1 gadolinium and T2 lesions for different anatomical regions, binary variables denoting presence or absence and numeric variable for maximum number of lesions

- ▷ All **variables** with more than **70%** missing removed
- ▷ No **subject** with more than **20%** missing

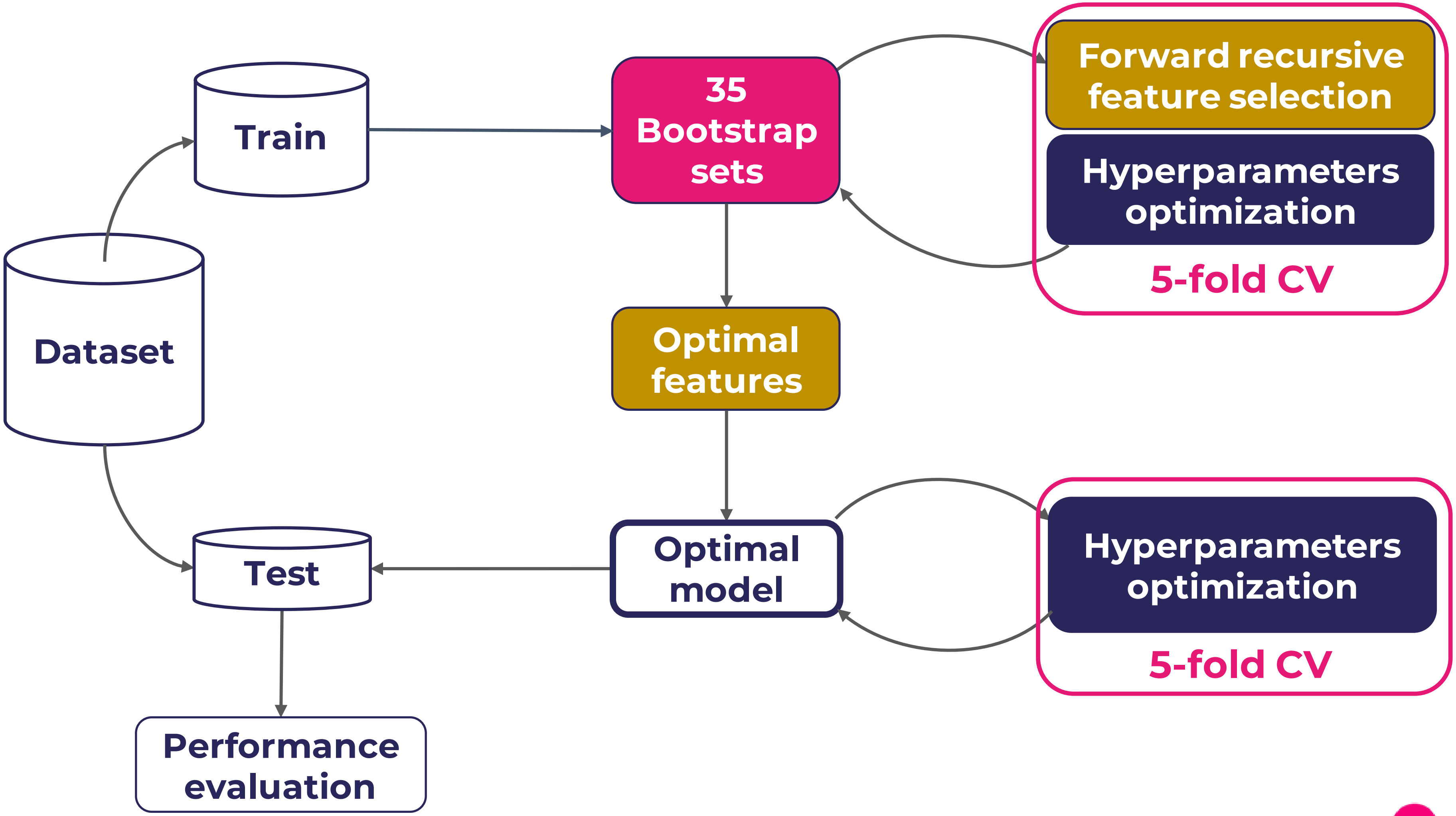
Preprocessing



- ▷ For **hyperparameters tuning** and **feature selection**
- ▷ **35 boots**: internal training set + validation set

- ▷ **Min max** scaling
- ▷ **MICE** imputation, 20 iterations

Model development framework



Dataset

Train

Test

35 Bootstrap sets

Forward recursive feature selection

Hyperparameters optimization

5-fold CV

Optimal features

Optimal model

Hyperparameters optimization

5-fold CV

Performance evaluation

Task 2 – Cumulative probability

Cox model

- Model outcome: **survival function**
- **Task 1:** $risk(t) = 1 - S(t)\ with\ t = 15\ years$
- **Task 2:** $risk(t) = 1 - S(t)\ with\ t \in (2,4,6,8,10)\ years$

SSVM

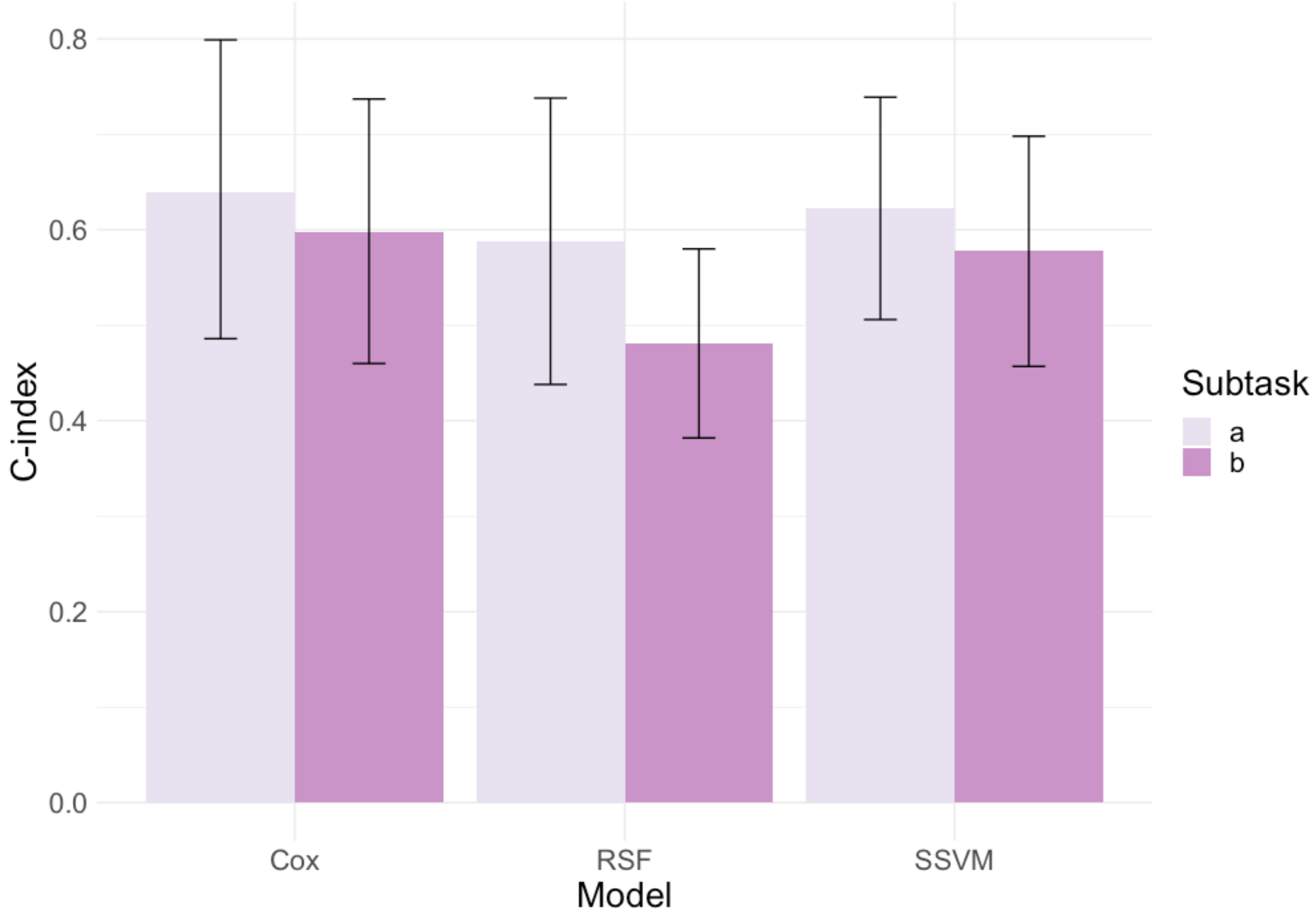
- Model outcome: **event time**
- **Task 1:** derived via Platt recalibration at 15 years
- **Task 2:** derived via Platt recalibration at (2,4,6,8,10) years

RSF

- Model outcome: **survival function**
- **Task 1:** $risk(t) = 1 - S(t) with\ t = 15\ years$
- **Task 2:** $risk(t) = 1 - S(t)\ with\ t \in (2,4,6,8,10)\ years$

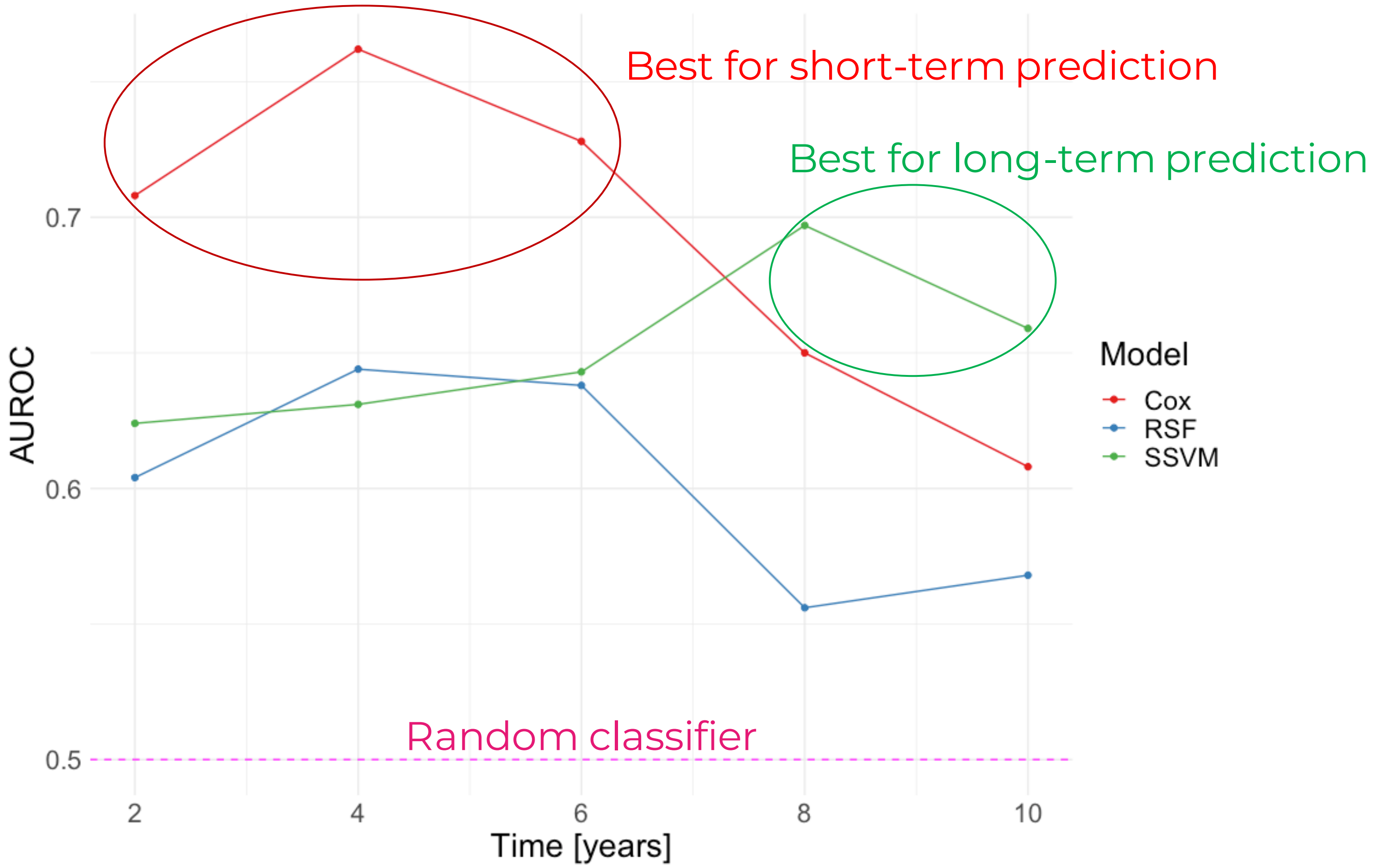
Results

Task 1: C-Index (CI)

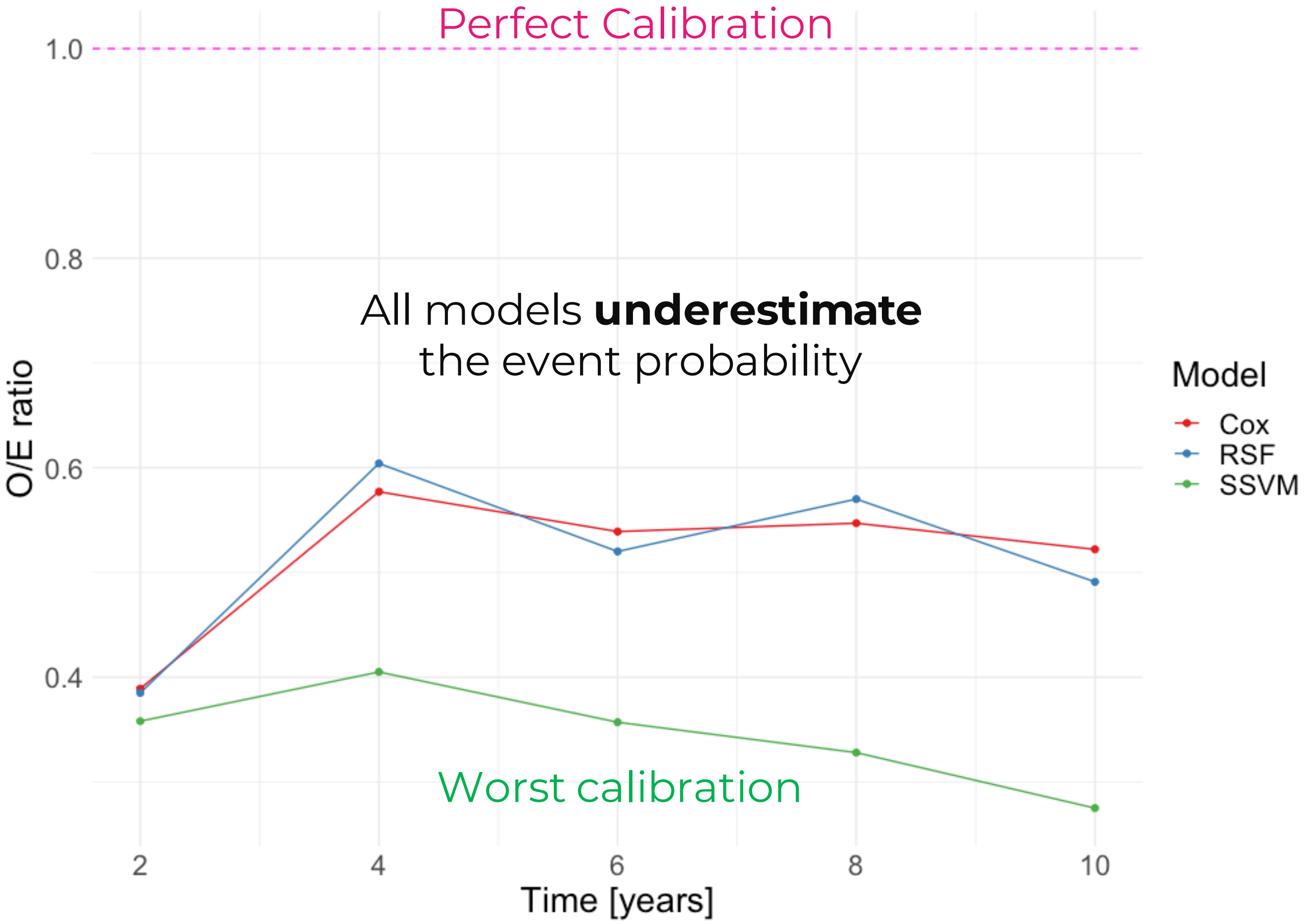


The **Cox model** achieves the **highest discrimination** in both sub-tasks

Task 2: sub-task a, AUROC

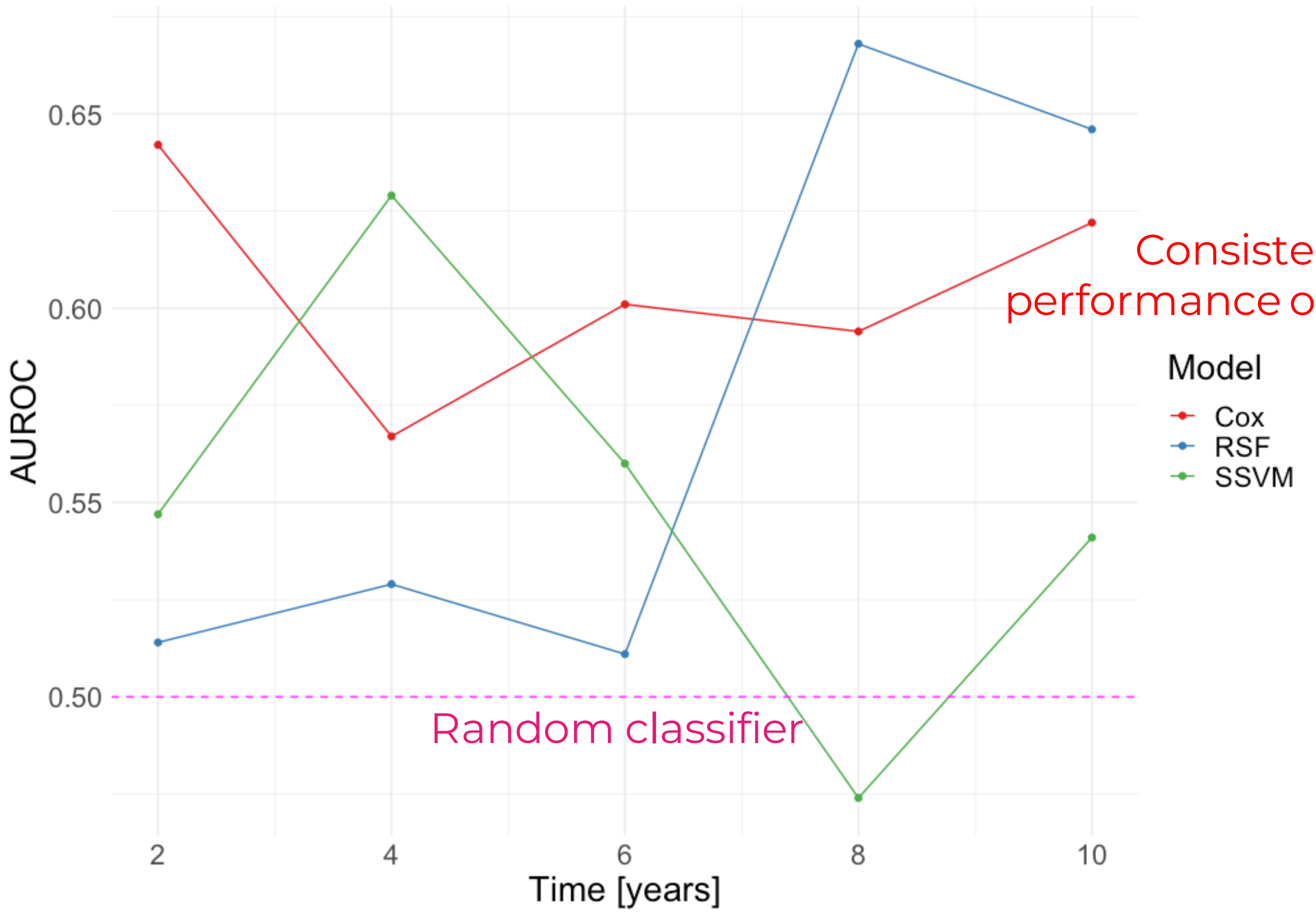


Task 2: sub-task a, O/E ratio



All models **underestimate** the event probability

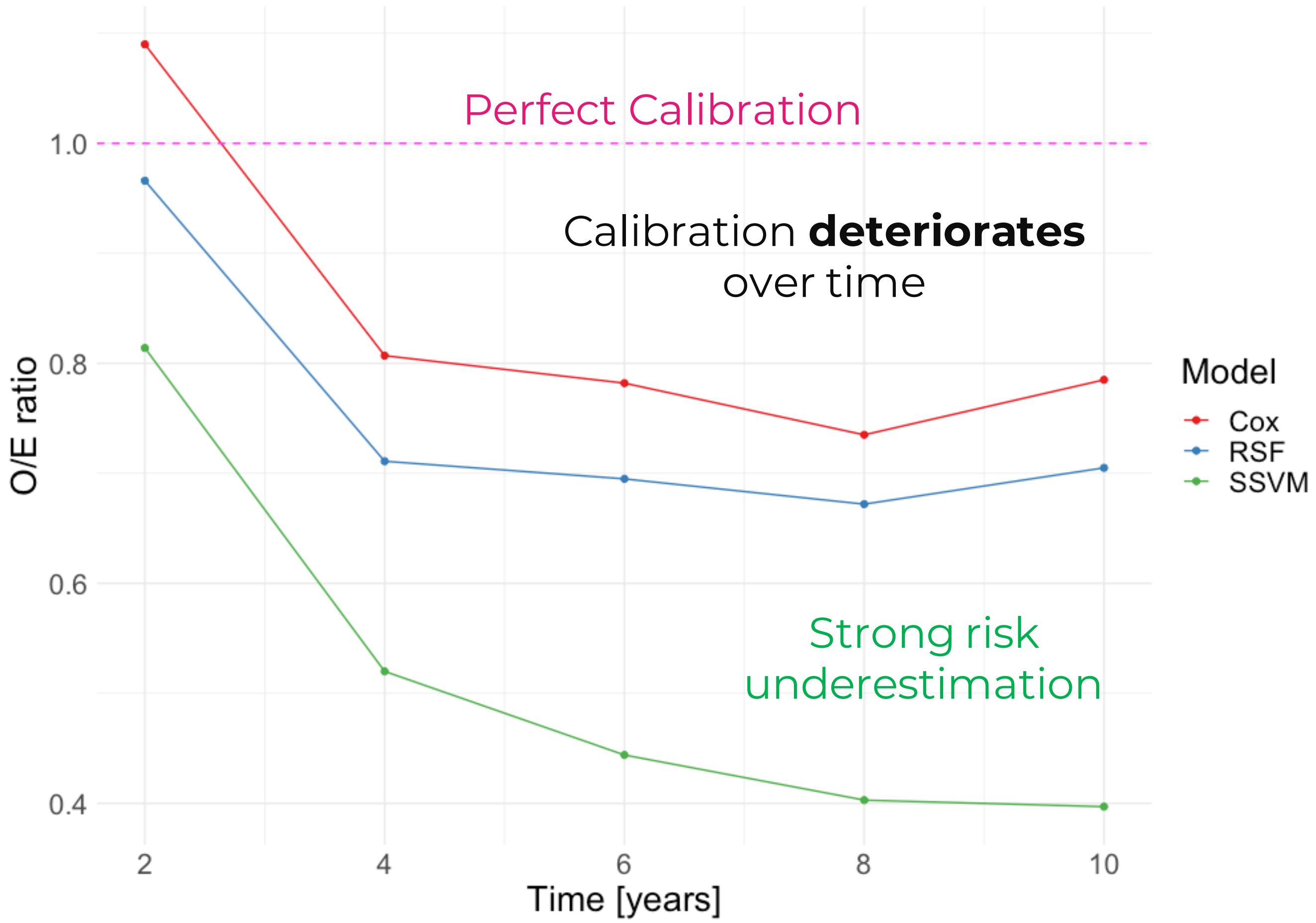
Task 2: sub-task b, AUROC



Consistent performance over time

Random classifier

Task 2: sub-task b, O/E ratio



Calibration **deteriorates** over time

Strong risk underestimation

Conclusion

Baseline Machine Learning Approaches to Predict Multiple Sclerosis Disease Progression
SBB Team

- **Task 1:** Poor discrimination across the board. Cox model is the best performing approach.
- **Task2:** Modest performance. Cox model is the best performing approach especially in terms of calibration.
- These results are **consistent** with what was previously observed in **similar studies.**
- Better results may be possible with **more sophisticated features extraction** processes concerning dynamic variables (i.e., **EDSS and MRI**)

Brainteaser

Thank You

Access the full text!

Alessandro Guazzo, PhD student
alessandro.guazzo@phd.unipd.it

SBB team, UNIPD
Barbara Di Camillo
Alessandro Guazzo
Isotta Trescato
Enrico Longato
Erica Tavazzi
Martina Vettoretti

@sysbiobigunipd

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No GA101017598