

Multiple Sclerosis Survival Prediction Results from DSM-COMP BIO UNITO

Notebook for the iDPP Lab on Intelligent Disease Progression Prediction at CLEF 2023

Ivan Rossi¹, Giovanni Birolo¹ and Piero Fariselli¹

¹Dept. of Medical Sciences, University of Turin.

Abstract

Multiple Sclerosis (MS) is an inflammatory disease of the central nervous system in which gradual destruction of the myelin causes interruption or disordered transmission of nerve impulses. its cause remains uncertain and treatment is unsatisfactory. In the Intelligent Disease Progression Prediction challenge, participants were offered two prediction tasks focused on the progression of Multiple Sclerosis by using demographical and clinical features. We employ machine learning methods for survival analysis. All models were optimized through a cross-validation procedure and finally evaluated on an internal test set. The three best performing methods, namely Elastic-net-penalized Cox model(CoxNet), Component-wise Gradient Boosting Survival Analysis and an hybrid method combining both, have been submitted for evaluation. Our results show that linear survival-analysis models could reach C-index values greater than 0.77 and 0.62 respectively in predicting MS worsening and cumulative risk of worsening. However feature-importance analysis also suggests that usage of semi-quantitative features, such as the EDSS scale, mask the importance, and potential usefulness, of most of the other features.

Keywords

Survival Analysis, Machine Learning, Multiple Sclerosis

1. Introduction

Multiple Sclerosis (MS) is a chronic diseases characterized by progressive or alternate impairment of neurological functions (motor, sensory, visual, cognitive). Patients have to manage alternated periods in hospital with care at home, experiencing a constant uncertainty regarding the timing of the disease acute phases and facing a considerable psychological and economic burden that also involves their caregivers. Clinicians, on the other hand, need tools able to support them in all the phases of the patient treatment, suggest personalized therapeutic decisions, indicate urgently needed interventions.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

✉ ivan.rossi@unito.it (I. Rossi); giovanni.birolo@unito.it (G. Birolo); piero.fariselli@unito.it (P. Fariselli)


🌐 <https://github.com/compbio-med-unito> (I. Rossi); <https://github.com/compbio-med-unito> (G. Birolo);

<https://github.com/compbio-med-unito> (P. Fariselli)

🆔 0000-0002-2077-7496 (I. Rossi); 0000-0003-0160-9312 (G. Birolo); 0000-0003-1811-4762 (P. Fariselli)

© 2023 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

1.1. Predicting Risk of Multiple Sclerosis Worsening

The prediction tasks focused on ranking subjects based on the risk of worsening, setting the problem as a survival analysis (SA) task, where the event of interest is the risk of MS worsening. The event definition has been based on changes in the patients Expanded Disability Status Scale (EDSS)[1] score, accordingly to clinical standards, for which different definitions have been provided.

1.1.1. Tasks

Task 1 asks for the prediction of the risk of worsening according to different definitions. In Subtask 1a, the worsening event is defined as the patient surpassing the EDSS threshold of 3 at least twice within one year interval. In Subtask 1b, the definition of worsening depends on the baseline EDSS value accordingly to current clinical protocols, namely:

- if baseline EDSS < 1 , worsening occurs when an increase of EDSS by 1.5 points is first observed;
- if $1 \leq$ baseline EDSS < 5.5 , worsening occurs when an increase of EDSS by 1.0 point is first observed;
- if baseline EDSS ≥ 5.5 , worsening occurs when an increase of EDSS by 0.5 points is first observed.

Task 2 required to explicitly assign a probability of worsening at different time windows (after 2, 4, 6, 8 and 10 years) according to the worsening event definition provided for Tasks 1a and 1b respectively.

For each sub-task, participants were given a dataset containing 2.5 years of visits, with the occurrence of the worsening event and the time of occurrence pre-computed by the challenge organizers.

For more details on the iDPP Challenge at CLEF 2023 and the tasks please refer to the overview papers [2, 3].

2. Methods

All the methods tested are implemented within the *scikit-survival* [4] Python package.

We report results for three (linear) methods: Cox regression [5] with Elastic Net [6] regularization (CoxNet), Component-wise Gradient Boosting Survival Analysis [7, 8] (CWGBSA), and a hybrid method where the most important features selected by CWGBSA are used to build a CoxNet model (EvilCox).

We also tested non-linear methods such as Random Survival Forest and Gradient Boosting SA, but both showed a tendency to overfit the training data and never performed better than CoxNet even after parameter optimization. Their results is not reported.

Model performance has been estimated by taking the average and standard deviation across 50 reshuffling rounds of five-fold cross-validation.

For some of the methods hyper-parameter optimization has been performed. See Table 1 for their values.

Table 1

Non-default hyper-parameters.

Method	Modified Hyperparameters
CoxNet	none
CWGBSA	n_estimators = 300, subsample = 0.75
EvilCox	As CWGBSA + Permutation-importance selection threshold ≥ 0.001

2.0.1. EvilCox

During model development, we routinely performed Permutation-based Feature Importance Analysis[9], as implemented in the *scikit-learn*[10] package, on each trained model. Feature permutation importance is a general strategy to measure the contribution of each feature on the prediction score by scrambling the analyzed-feature values to destroy correlation.

We observed that CWGBSA is very resistant to over-fitting and does implicit feature selection, as Coxnet does, but it appears to be more aggressive with respect to feature elimination. Nonetheless CWGBSA cross-validated performance turned out to be almost on par with that of CoxNet, despite using a rather smaller set of features. Since CWGBSA approaches a linear least-square solution as the number of estimators grows, we tested a hybrid method where CWGBSA acts as a feature selector for CoxNet. This was done simply by selecting all features whose average (over 50 runs) contribution to the score was larger than 0.001 and using them to build a CoxNet model. In our tests, the cross-validated score of EvilCox outperformed that of the CoxNet model.

2.1. Feature engineering

Each data set provided had a training set comprising a table of static features, a table of longitudinal (temporal) features and a table of outcomes.

While static features are collected once for each individual, longitudinal features are collected repeatedly at different time points for the same individual. A variable had thus a different number of values (possibly also one or zero) for each individual. Since we did not use predictive methods that could handle longitudinal data directly, we re-coded the multiple values that were available for each individual for a longitudinal feature as several static features: the number, the mean, the standard deviation, the maximum, minimum, first (earliest collection time), second-to-last and last (most recent collection time) value for both EDSS values and time of EDSS evaluation. We also collected the number of EDSS evaluation above the three thresholds of 1.5, 3.0 and 5.5 (*over_t1*, *over_t2*, *over_t3*) for each patient.

Categorical features and boolean features with missing values were one-hot-encoded. All features with more than 40% of missing values in the training set were dropped. All the remaining missing values after feature dropping have been given a value equal to the median of the corresponding feature value.

The predicted time-to-event $t > 0$ risk was converted into a relative risk $0 < r < 1$ by rescaling it to fall within its maximum and minimum value

$$r_{rel}(t) = \frac{r(t) - r_{min}}{r_{max} - r_{min}}$$

3. Results

3.0.1. Task 1

As usual for survival-analysis tasks, Task 1 performance has been graded according to Harrell's concordance index[11] (C-index), both by cross-validation on the training set and on the challenge test set. Results are reported in table 2.

Our models performance are significantly better for task 1a than for task 1b, where the definition of the event changes according to the patient first EDSS evaluation. Furthermore, for task 1a, the training score is very consistent with respect to the testing results, showing very good generalization. Both the CWGBSA and EvilCox methods perform very well even when using a significantly smaller feature set.

Table 2

Task-1 results. The 90% confidence interval is reported within square brackets.

Task	Method	C-index (cross-validation)	C-index (test)
1a	CoxNet	0.727	0.802 [0.685-0.919]
1a	CWGBSA	0.729	0.771 [0.624-0.919]
1a	EvilCox	0.762	0.769 [0.631-0.908]
1b	CoxNet	0.664	0.634 [0.528-0.739]
1b	CWGBSA	0.673	0.613 [0.514-0.713]
1b	EvilCox	0.696	0.623 [0.526-0.721]

3.0.2. Feature importance

Feature importance analysis shows *Fig1, somewhat unsurprisingly, that the most important features for subtask a are different from those that describe subtask b, although all the methods consistently select similar features for the same task. In particular, subtask a feature importance is largely dominated by the most recent EDSS evaluation while the two most relevant feature for the subtask b are 1) the difference between the two most recent EDSS values and 2) the number of EDSS evaluation with score over 3.0.

In general, EDSS-related features dominate the feature-importance landscape.

3.0.3. Task 2

The survival methods we use in Task 1 can output their prediction both as a relative risk and as survival curve. So we took the same models fitted in Task 1 and used their predicted survival curve to get the Task 2 outcome.

Two metrics have been used in Task 2, namely the receiver operating characteristic area under the curve (ROC AUC) at each time threshold and the observed/expected events (O/E) ratio for each time interval.

Results are reported in table 3 and are in line with those from Task 1.

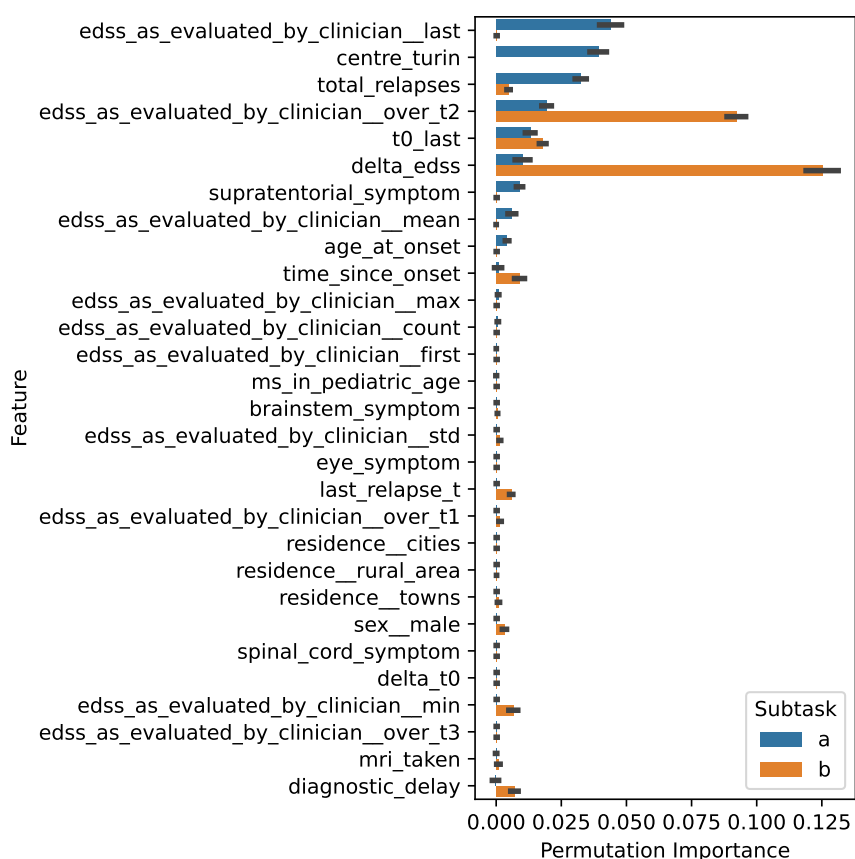


Figure 1: Permutation importance in the CoxnNet models fitted for subtasks a and b. Features are sorted by subtask a, error bars show the 95% confidence intervals.

4. Conclusion

We tried a series of standard with good results for subtask a and lower performances for subtask b. While subtask a has a fixed threshold, subtask b outcome is a relative increment in EDSS score, which seems to be more challenging to predict.

Beyond the methods we submitted to the challenge, which are all linear, we experimented also with non-linear models such as Random Survival Forest. However, linear models consistently outperformed non-linear ones, while non-linear ones suffered more from over-fitting.

While the best models were the CoxNet ones, it is possible to get a good model fit even with very few features: CWGBSA and EvilCox models used only six features. In particular, the feature-importance space is dominated by the EDSS most recent evaluations. While the importance of previous EDSS evaluations is not surprising given that the outcomes are also EDSS-derived, few non-EDSS features were found to be predictive. A rationale for this behaviour might be that, being EDSS a semi-quantitative feature, it summarizes the contribution of the other features through the clinician judgement, making difficult to factorize other con-

Table 3

Task 2: Relative risk after N years. The 90% confidence interval is reported within square brackets.

Method	Task	Years	ROC AUC	O/E
Coxnet	2a	2	0.890 [0.739 - 1.000]	0.443 [-0.018 - 0.904]
		4	0.900 [0.779 - 1.000]	0.627 [0.136 - 1.117]
		6	0.856 [0.722 - 0.991]	0.608 [0.184 - 1.031]
		8	0.787 [0.626 - 0.948]	0.652 [0.244 - 1.061]
		10	0.796 [0.640 - 0.952]	0.636 [0.257 - 1.016]
CWGBSA	2a	2	0.841 [0.618 - 1.000]	0.467 [-0.007 - 0.940]
		4	0.864 [0.691 - 1.000]	0.638 [0.143 - 1.134]
		6	0.821 [0.658 - 0.984]	0.619 [0.191 - 1.047]
		8	0.759 [0.587 - 0.931]	0.666 [0.253 - 1.079]
		10	0.765 [0.594 - 0.935]	0.643 [0.262 - 1.024]
EviCox	2a	2	0.854 [0.655 - 1.000]	0.449 [-0.015 - 0.913]
		4	0.867 [0.711 - 1.000]	0.620 [0.132 - 1.109]
		6	0.816 [0.655 - 0.978]	0.605 [0.182 - 1.027]
		8	0.749 [0.570 - 0.927]	0.649 [0.242 - 1.057]
		10	0.757 [0.585 - 0.929]	0.634 [0.255 - 1.012]
Coxnet	2b	2	0.676 [0.514 - 0.838]	1.082 [0.467 - 1.697]
		4	0.633 [0.486 - 0.780]	0.858 [0.430 - 1.286]
		6	0.635 [0.488 - 0.782]	0.811 [0.443 - 1.180]
		8	0.651 [0.503 - 0.800]	0.803 [0.465 - 1.141]
		10	0.686 [0.526 - 0.847]	0.845 [0.516 - 1.174]
CWGBSA	2b	2	0.632 [0.477 - 0.787]	1.101 [0.481 - 1.721]
		4	0.626 [0.484 - 0.768]	0.850 [0.424 - 1.276]
		6	0.655 [0.512 - 0.797]	0.809 [0.441 - 1.176]
		8	0.673 [0.530 - 0.816]	0.802 [0.464 - 1.140]
		10	0.709 [0.556 - 0.862]	0.850 [0.520 - 1.180]

tributions.

5. Acknowledgements

This work has been supported by the European Union's Horizon 2020 Brainteaser Project (GA101017598).

6. Bibliography

References

- [1] J. F. Kurtzke, Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS), *Neurology* 33 (1983) 1444–1452. doi:10.1212/wnl.33.11.1444.
- [2] G. Faggioli, A. Guazzo, S. Marchesin, L. Menotti, I. Trescato, H. Aidos, R. Bergamaschi, G. Birolo, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. M. Di Nunzio, P. Fariselli, J. M. García Domínguez, M. Gromicho, E. Longato, S. C. Madeira, U. Manera, G. Silvello,

- E. Tavazzi, E. Tavazzi, M. Vettoretti, B. Di Camillo, N. Ferro, Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2023, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, A. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023)*, Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2023.
- [3] G. Faggioli, A. Guazzo, S. Marchesin, L. Menotti, I. Trescato, H. Aidos, R. Bergamaschi, G. Birolo, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. M. Di Nunzio, P. Fariselli, J. M. García Dominguez, M. Gromicho, E. Longato, S. C. Madeira, U. Manera, G. Silvello, E. Tavazzi, E. Tavazzi, M. Vettoretti, B. Di Camillo, N. Ferro, Overview of iDPP@CLEF 2023: The Intelligent Disease Progression Prediction Challenge, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), *CLEF 2023 Working Notes*, CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073., 2023.
- [4] S. Pölsterl, scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn, *Journal of Machine Learning Research* 21 (2020) 1–6. URL: <http://jmlr.org/papers/v21/20-729.html>.
- [5] D. R. Cox, Regression Models and Life-Tables, *Journal of the Royal Statistical Society. Series B (Methodological)* 34 (1972) 187–220. URL: <https://www.jstor.org/bibliopass.unito.it/stable/2985181>.
- [6] H. Zou, T. Hastie, Regularization and Variable Selection Via the Elastic Net, *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67 (2005) 301–320. URL: <https://doi.org/10.1111/j.1467-9868.2005.00503.x>. doi:10.1111/j.1467-9868.2005.00503.x.
- [7] T. Hothorn, Survival ensembles, *Biostatistics* 7 (2005) 355–373. doi:10.1093/biostatistics/kxj011.
- [8] P. Bühlmann, T. Hothorn, Boosting algorithms: Regularization, prediction and model fitting, *Statistical Science* 22 (2007). doi:10.1214/07-sts242. arXiv:arXiv:0804.2752v1[stat.ME].
- [9] L. Breiman, Random Forests, *Machine Learning* 45 (2001) 5–32. URL: <https://doi.org/10.1023/A:1010933404324>. doi:10.1023/A:1010933404324.
- [10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* 12 (2011) 2825–2830.
- [11] F. E. Harrell, Jr, R. M. Califf, D. B. Pryor, K. L. Lee, R. A. Rosati, Evaluating the Yield of Medical Tests, *JAMA* 247 (1982) 2543–2546. URL: <https://doi.org/10.1001/jama.1982.03320430047030>. doi:10.1001/jama.1982.03320430047030.