Sample Size considerations for the KIPRUN study

Márton Kiss, MD

2025-01-21

Table of contents

[*Executive summary* 2](#_Toc188358415)

[Variables of interest 2](#_Toc188358416)

[AUC investigation 4](#_Toc188358417)

[Random Forest 5](#_Toc188358418)

[Remarks 8](#_Toc188358419)

[References 8](#_Toc188358420)

# *Executive summary*

It is with clarity — and a measure of regret — that we must conclude that the proposed clinical trial evaluating the utility PCT (Procalcitonin) change monitoring should not proceed. The preliminary data at our disposal, which includes both a conventional Area Under the Curve (AUC) analysis for the change and a random forest–based approach, offers no credible evidence that comparing a new PCT test to the last known one could reliably guide clinical decision-making or reduce inappropriate therapy.

The fundamental premise—that measurements taken well in advance of an actual infection event would inform clinical suspicion—appears unsound in light of the data. Indeed, these results are akin to attempting to predict influenza infection from body temperature changes hours before the virus could plausibly manifest a fever. Previous attempts hinted that a reliable classification system would need data from 16-24 hours **after** the suspicion of infection arose.

Despite the theoretical promise of PCT change as a biomarker, the data were lacking a predictive signal which would justify further investment of time or resources. The random forest model, which typically excels at extracting subtle patterns, also failed to produce an acceptable method for forgoing therapy in case of a suspicion. On the whole, the lack of a signal in the preliminary shared data does not justify a repeat clinical trial.

## Variables of interest

The distribution of the changes between the last known (T-1) PCT value and the current (T0) PCT value is shown in Figure 1.

In an attempt to provide a better contrast between subjects with a marked change and the ones who had less marked changes, the rate of change over the absolute difference (CONTR variable) is also reported.

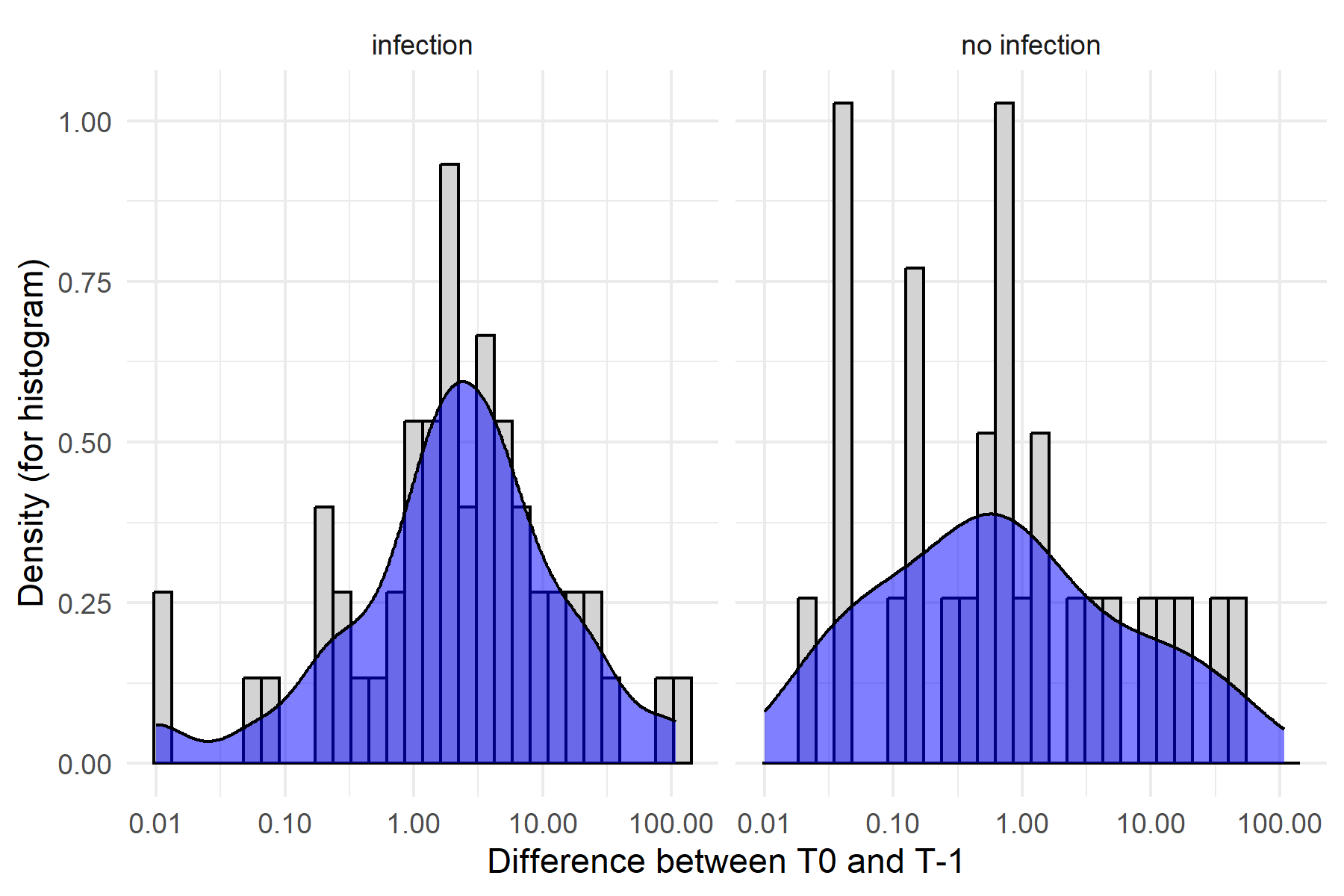


Figure 1

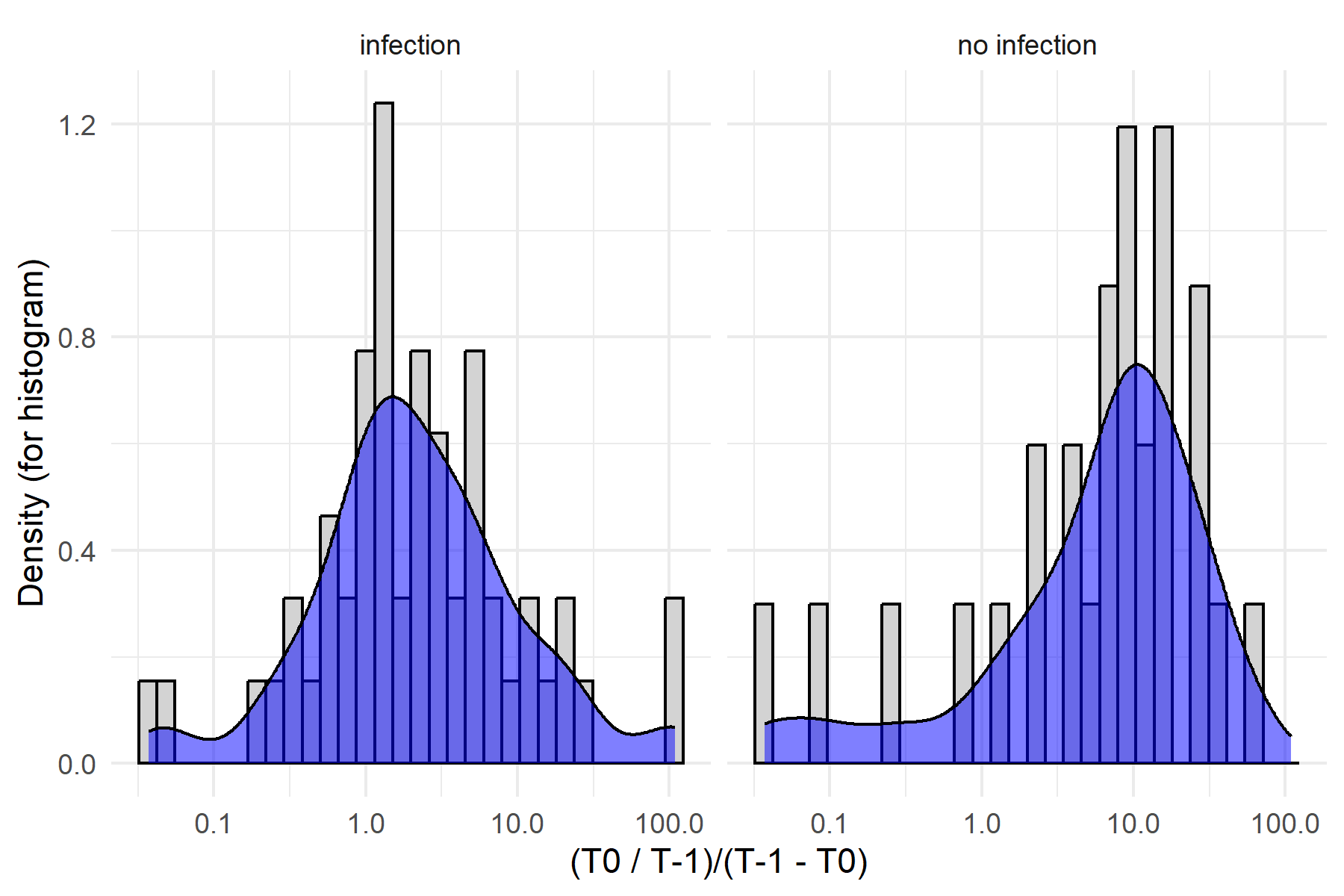


Figure 2

## AUC investigation

The ROC curve in Figure 3 shows the performance of the CONTR variable in predicting the infection status. The area under the curve (AUC) is 0.5984 which is less than optimal. The AUC curve shows that the variable is practically useless in predicting negative infection status with confidence.

Fixing the cutoff value as -1.4056 or at the 10th percentile (making only 10 percent of the sample to be prediced ‘negative’) the ‘negative’ cases had a 50% NPV, meaning that in 50% of cases which were predicted negative (predicted ‘0’ as per the coding used), there was an infection. Given the seriousness of the condition, this does not seem to be an acceptable rate of false negatives.

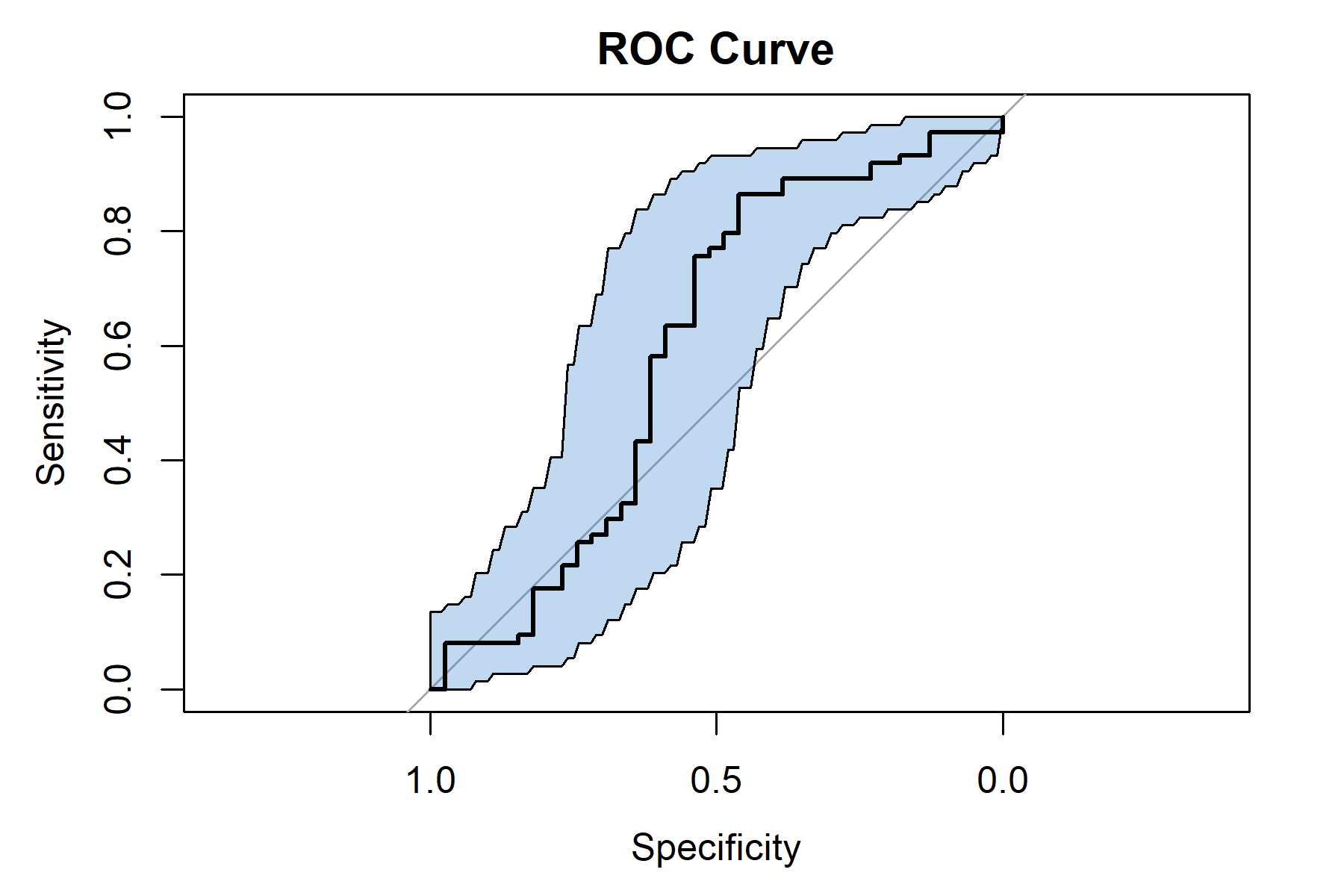


Figure 3

Confusion Matrix and Statistics  
  
 Reference  
Prediction 1 0  
 1 68 33  
 0 6 6  
   
 Accuracy : 0.6549   
 95% CI : (0.5596, 0.7418)  
 No Information Rate : 0.6549   
 P-Value [Acc > NIR] : 0.5434   
   
 Kappa : 0.087   
   
 Mcnemar's Test P-Value : 0.00003136   
   
 Sensitivity : 0.9189   
 Specificity : 0.1538   
 Pos Pred Value : 0.6733   
 Neg Pred Value : 0.5000   
 Prevalence : 0.6549   
 Detection Rate : 0.6018   
 Detection Prevalence : 0.8938   
 Balanced Accuracy : 0.5364   
   
 'Positive' Class : 1

## Random Forest

A Random forest approach was used in order to make sure that the maximum amount of information was utilized in the classification.

A 10-fold crossvalidation setup was used to ensure that the model was not overfitting the data and to better investigate the generalizability of the results.

The model was optimized on the number of predictors available per tree. 2000 trees per forest were used to ensure the stability of the results.

The optimized NPV for the holdout sets was only 30%, meaning that if the model signalled a negative result, there was a 70% chance that the patient was actually infected (!)

This is not an acceptable rate of false negatives for this condition. As this approach ensured the best possible classification mirroring the intended use, it can be stated that the T-1 vs. T0 PCT values could **not** be reliably used to predict the lack of infection if the suspicion for an infection is present as determined by the attending physician.

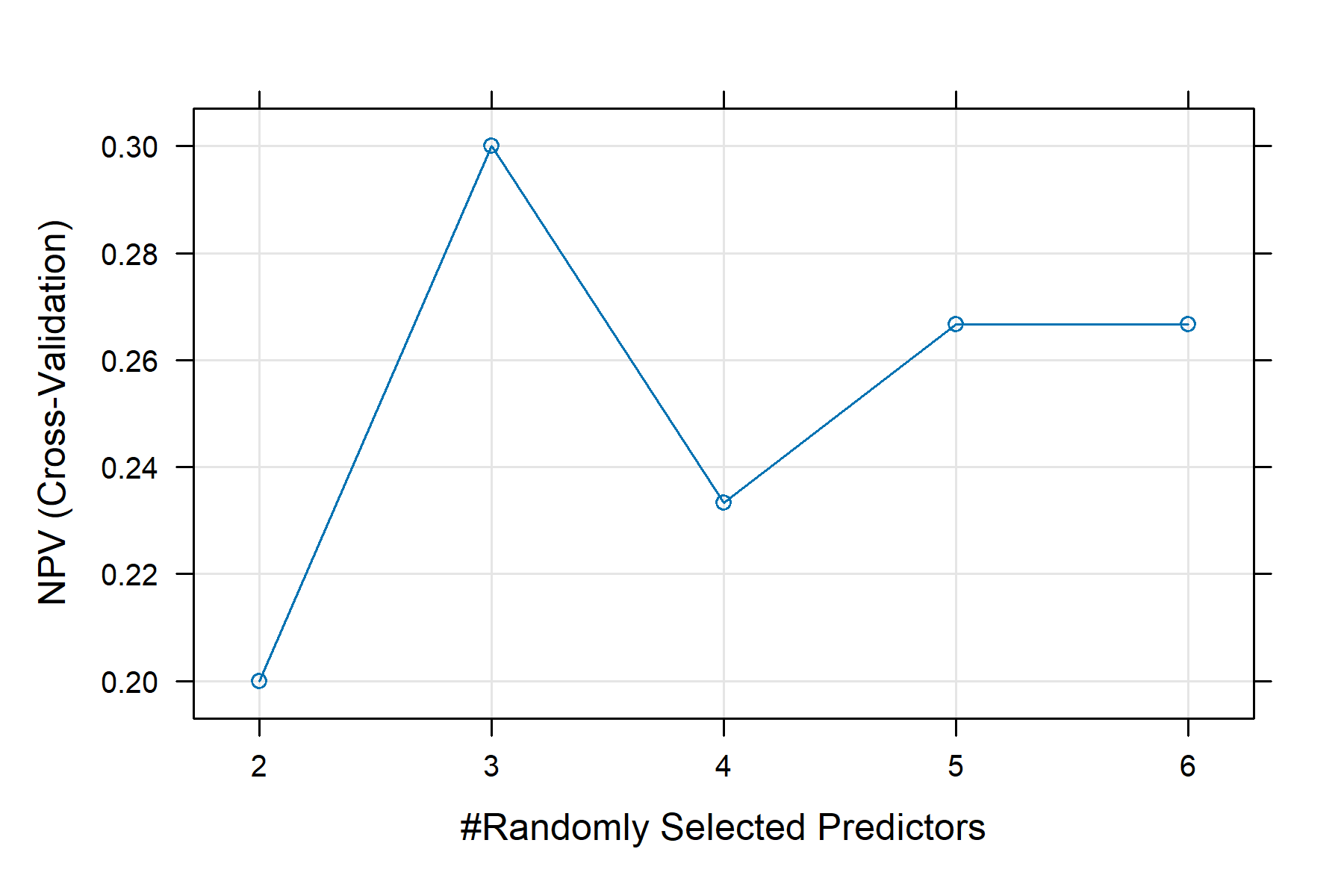


Figure 4 - NPV optimization in the Random Forest model

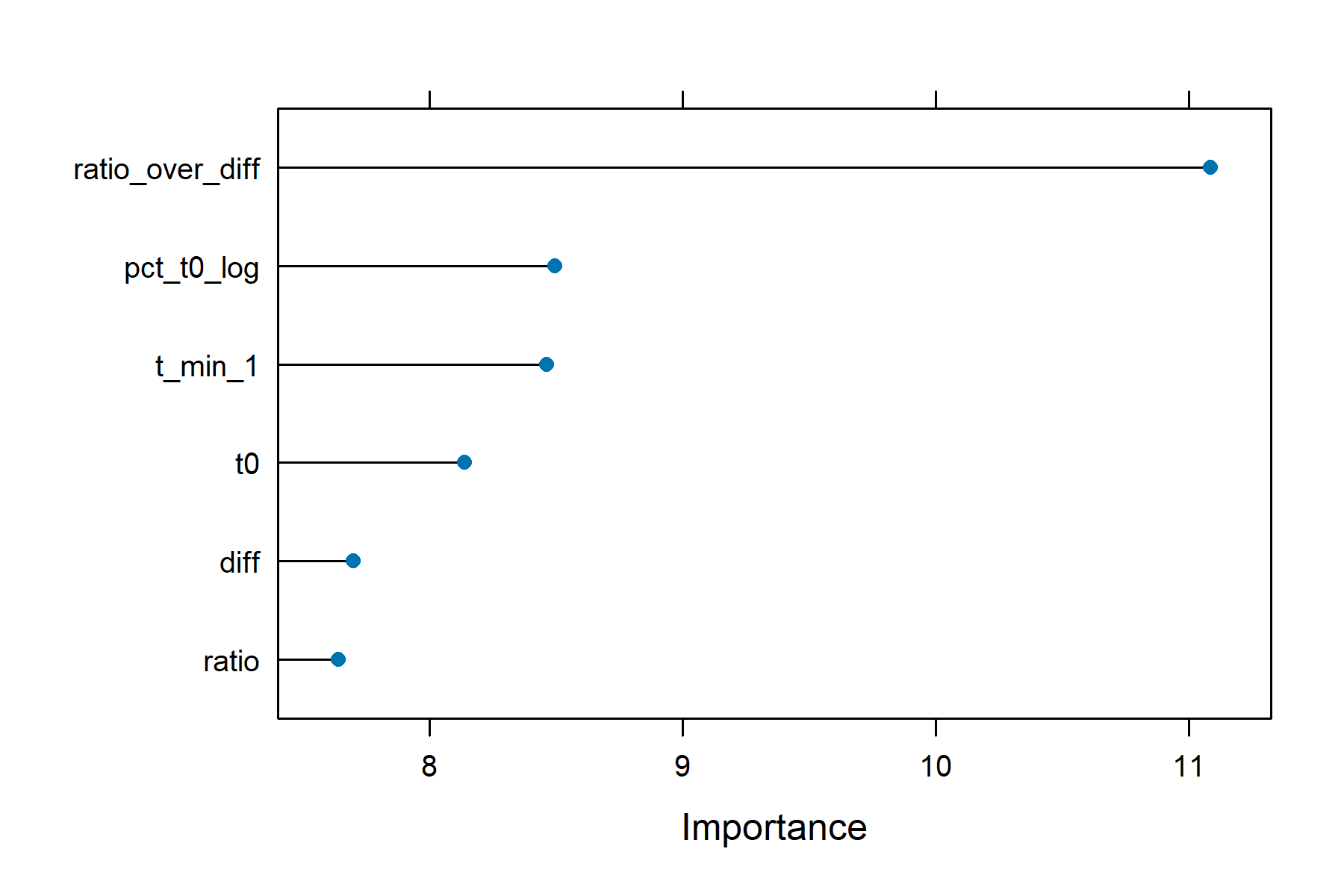


Figure 5 - Variable Importance plot

# Remarks

### Information regarding the document’s compilation

Analyses were conducted using the R Statistical language (version 4.4.1; R Core Team, 2024) on Windows 11 x64 (build 26100), using the packages caret (version 6.0.94; Kuhn, Max, 2008), randomForest (version 4.7.1.2; Liaw A, Wiener M, 2002), pROC (version 1.18.5; Robin X et al., 2011), lattice (version 0.22.6; Sarkar D, 2008), rpart (version 4.1.23; Therneau T, Atkinson B, 2023), ggplot2 (version 3.5.1; Wickham H, 2016) and dplyr (version 1.1.4; Wickham H et al., 2023).

## References

* Kuhn, Max (2008). “Building Predictive Models in R Using the caret Package.” Journal of Statistical Software, 28(5), 1–26. doi:10.18637/jss.v028.i05 <https://doi.org/10.18637/jss.v028.i05>, <https://www.jstatsoft.org/index.php/jss/article/view/v028i05>.
* Liaw A, Wiener M (2002). “Classification and Regression by randomForest.” R News, 2(3), 18-22. <https://CRAN.R-project.org/doc/Rnews/>.
* R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
* Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, Müller M (2011). “pROC: an open-source package for R and S+ to analyze and compare ROC curves.” BMC Bioinformatics, 12, 77.
* Sarkar D (2008). Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5, <http://lmdvr.r-forge.r-project.org>.
* Therneau T, Atkinson B (2023). rpart: Recursive Partitioning and Regression Trees. R package version 4.1.23, <https://CRAN.R-project.org/package=rpart>.
* Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, <https://ggplot2.tidyverse.org>.
* Wickham H, François R, Henry L, Müller K, Vaughan D (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.4, <https://CRAN.R-project.org/package=dplyr>.

### Time of compilation

2025-01-21 13:26:24.403906