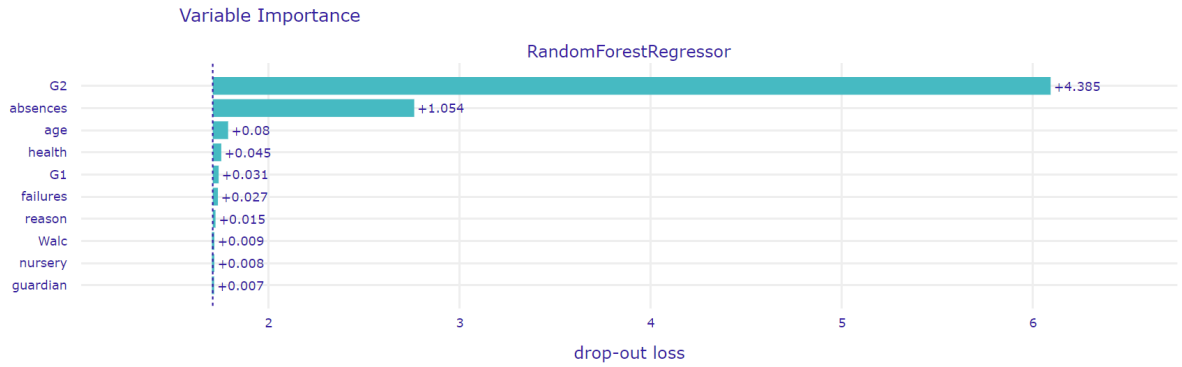
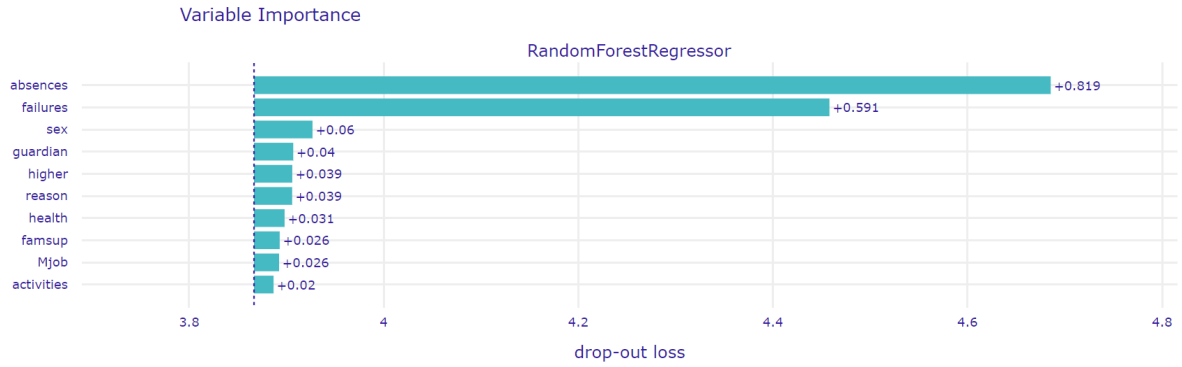
1. For the selected data set, train at least one tree-based ensemble model. Calculate Permutation-based Variable Importance for the selected model.

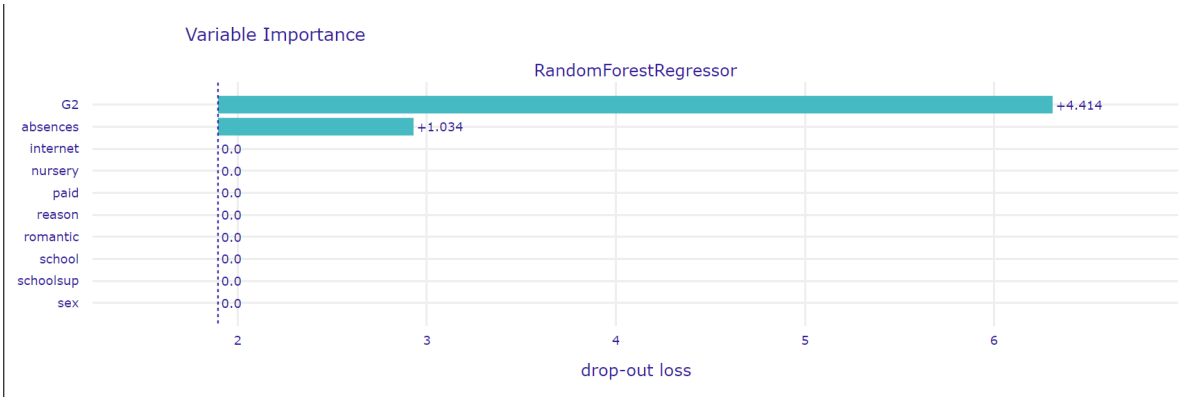


As expected from previous analyses, the model values most the grades from 2nd semester, however, the grade from 1st semester is not the second most important value, rather it is the number of school absences. Weekly alcohol consumption appears 8th on the list.

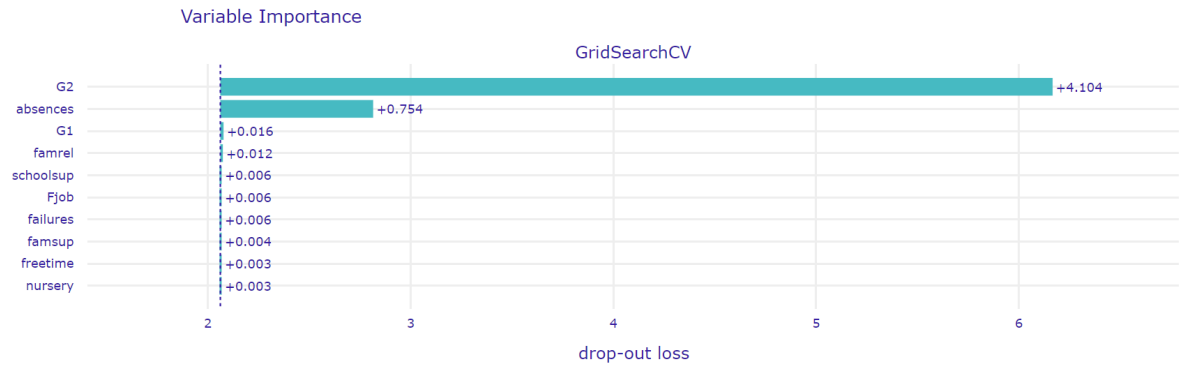
2. Train three more candidate models and compare their rankings of important features using PVI. What are the differences? Why?



The first model was stripped of the G1 and G2 features. As expected, the number of absences takes the first place as the most important feature, but instead of age, the second most important feature is the number of failures, which appeared only as 6th in the unmodified model.



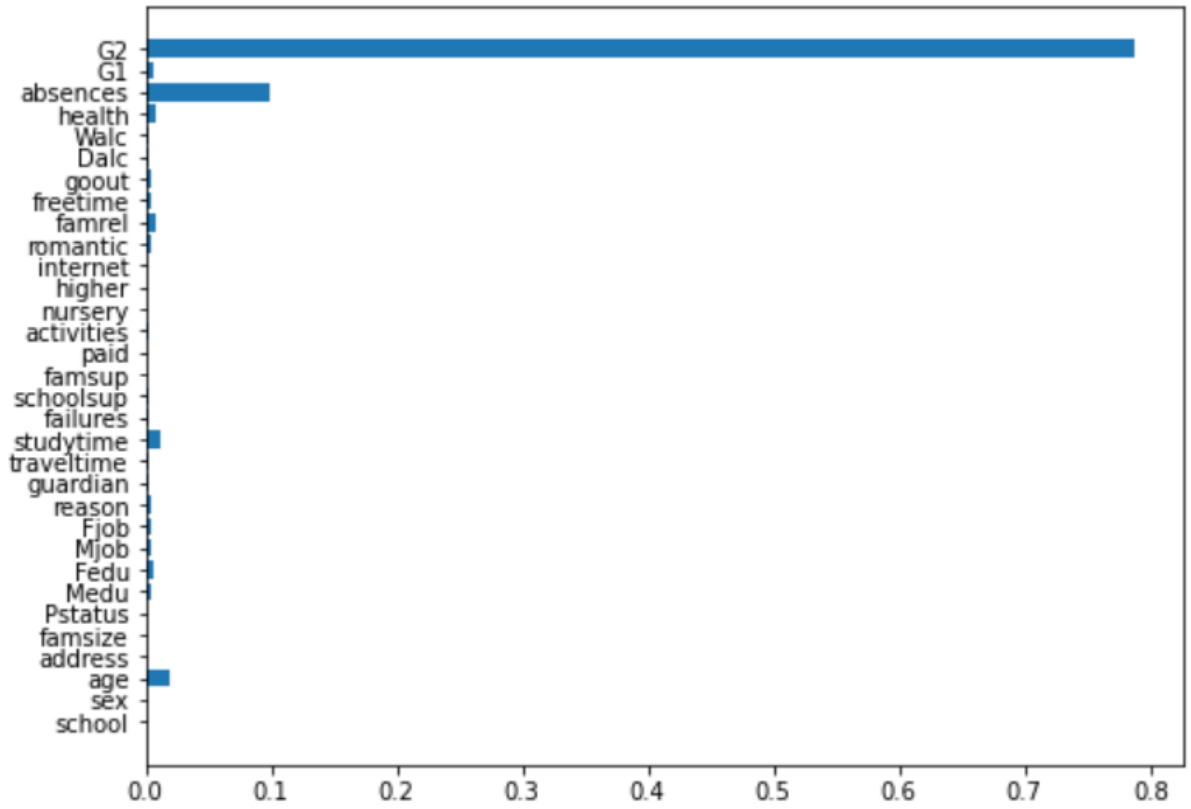
The second model has parameters changed to be more rough and general. Not surprisingly, it only took into account the two most important features from the unmodified model.



The 3rd model had automatically optimized parameters. It resembles the unmodified model's plot, the only difference seems to be that all of the features beside the two most important have even smaller values than before.

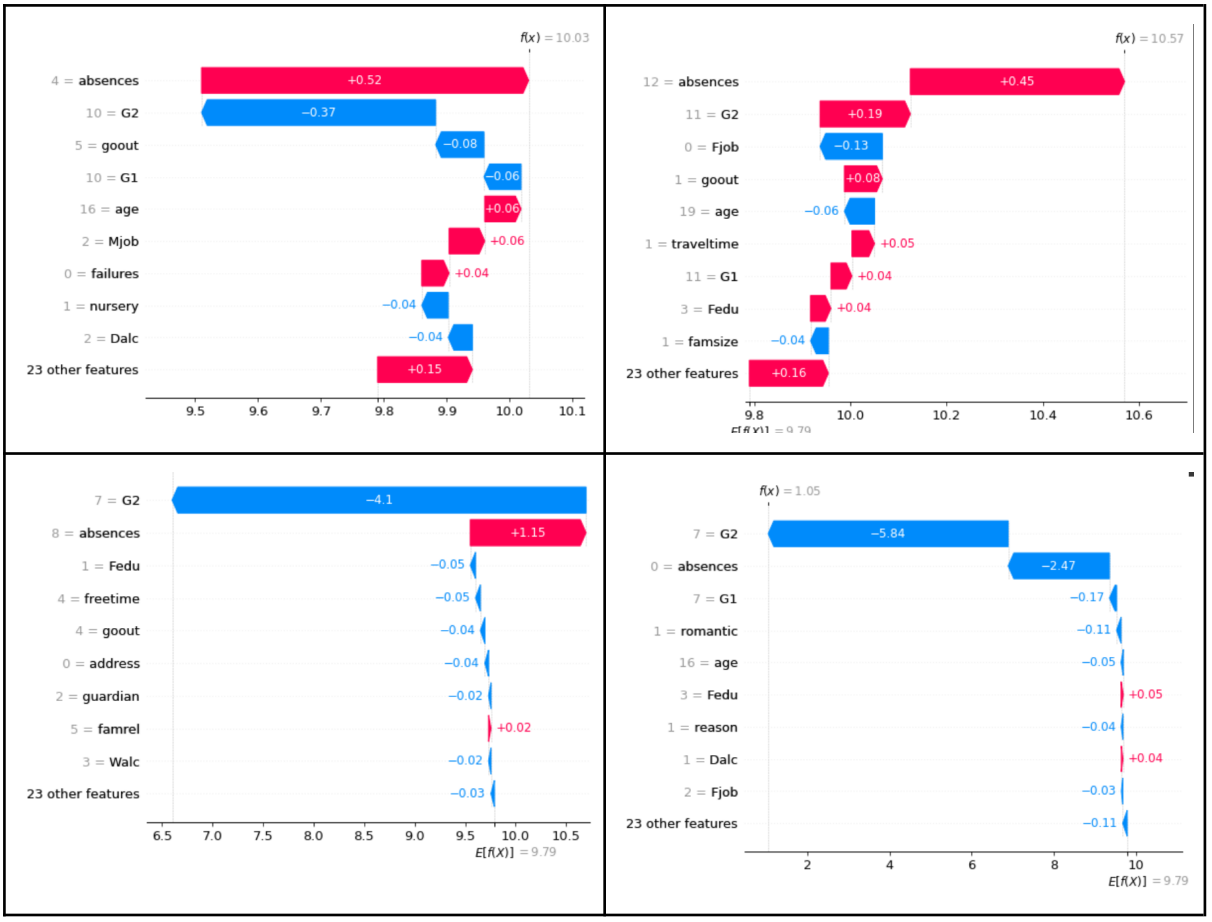
3. For the tree-based model from (1), compare PVI with:

a) the traditional feature importance measures for trees,



The random forest's own measure of importance is consistent with the output of the PVI method.

b) [in Python] SHAP variable importance based on the TreeSHAP algorithm available in the shap package.



Of the 4 visualized samples most all have the same most important features (“absences”, “G2”) as the basic model’s PVI results, however, the first two samples have them in reversed order - the number of absences is more influential on the samples’ value than 2nd semester grades.