Homework No. 5 - Permutation-based Variables Importance

Paulina Kaczynska

November 22, 2022

The goal of this homework is to explain the models trained on the Alcohol Effects on Study Dataset from Kaggle with Permutation-based Variables Importance method. It bases on the data preprocessed in previous homeworks.

1 Permutation-based Variable Importance for random forest regressor

Firstly, I train Random Forest Regression using function from sklearn package. This method is a tree ensemble method.

The plots of the features' importance can be seen on the Figure 1. On the horizontal axis there is drop in the loss after permutation of a given variable and on the vertical axis - features.

Changing model hypermaparameters result in a small change in variables' importance, which, in turn, contributes to the different order of the less important variables. The hyperparameters are: 300 estimators and depth of 4 for the first model and 100 estimators and depth of 8 for the second one.

2 Permutation-based Variable Importance for linear regression and stochastic gradient descent

Two models with different architectures were trained: linear regression and stochastic gradient descent.

Although stochastic gradient descent's features' importance is similar to random forest regressor's both in the aspect of average drop-out loss and feature ranking, the linear regression profile is very different. The loss is bigger, but change in it after the permutation is smaller than for the other models.

3 Comparison of PVI, impurity based feature importance and SHAP values for random forest regressor

Different methods of estimating features' contribution were compared. Impurity based feature importance is provided as the attribute of the random forest regressor in the sklearn package. It was plotted on the Figure 3 for the two random forest regressors trained in the earlier steps.

Shap value was obtained using SHAP package and can be seen on the Figure 4.

The most important variables are the same for all three methods. SHAP shows the feature attribution to the score and PVI is based on the drop of the model's performance when there is noise in the given feature, so they measure different things and do not have to be so similar as they are here. Impurity based feature importance score should measure the probability of misclassifying when impurity is introduced in the given feature, so they focus on the model performance similarilyas the PVI.

References

[1] Alcohol Effects on Study Dataset: *https://www.kaggle.com/datasets/whenamancodes/alcohol-effects-on-study*

[2] eXplainable Machine Learning - Wyjaśnialne Uczenie Maszynowe - 2023; *https://github.com/mim-uw/eXplainableMachineLearning-2023/*

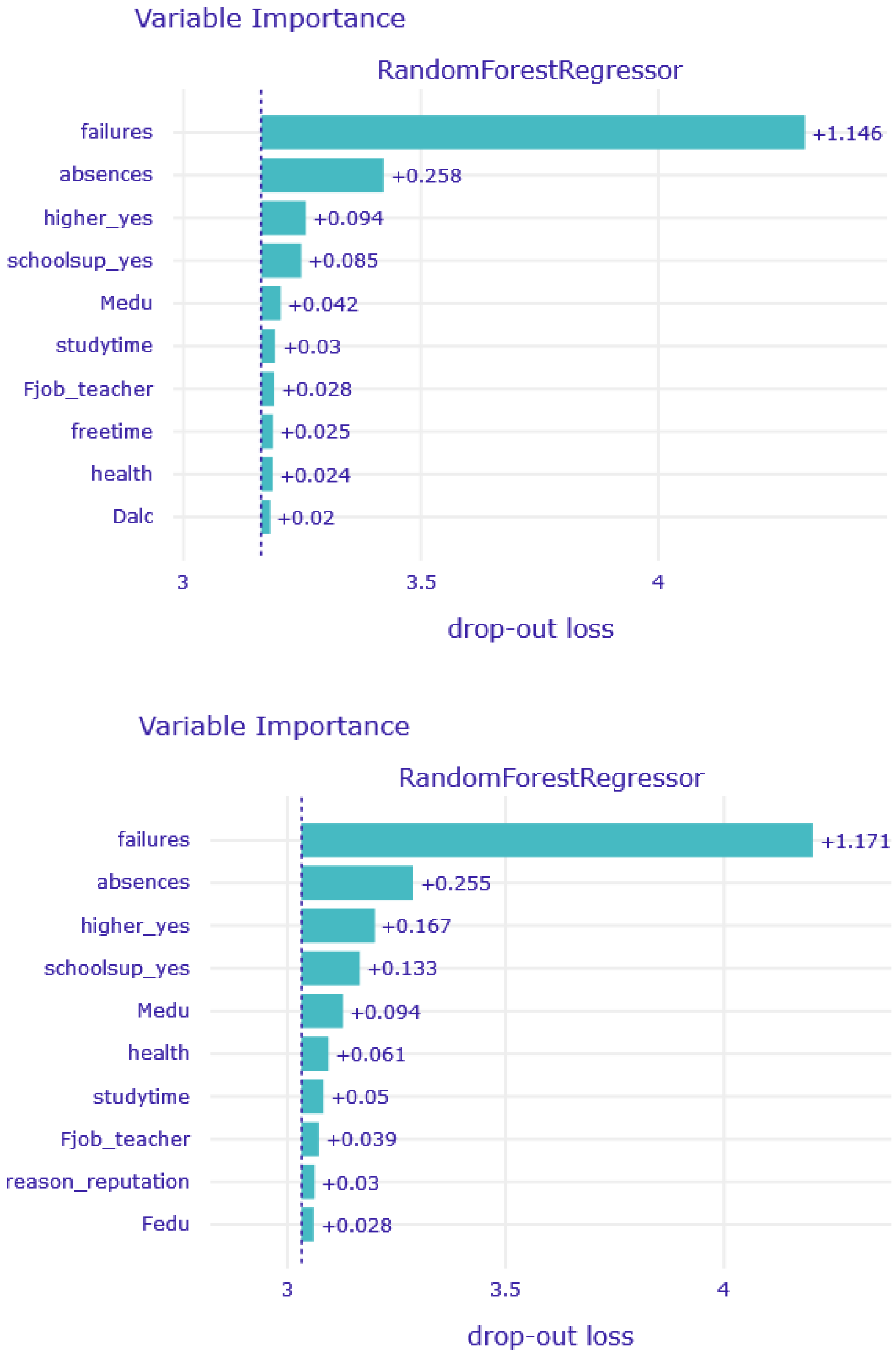
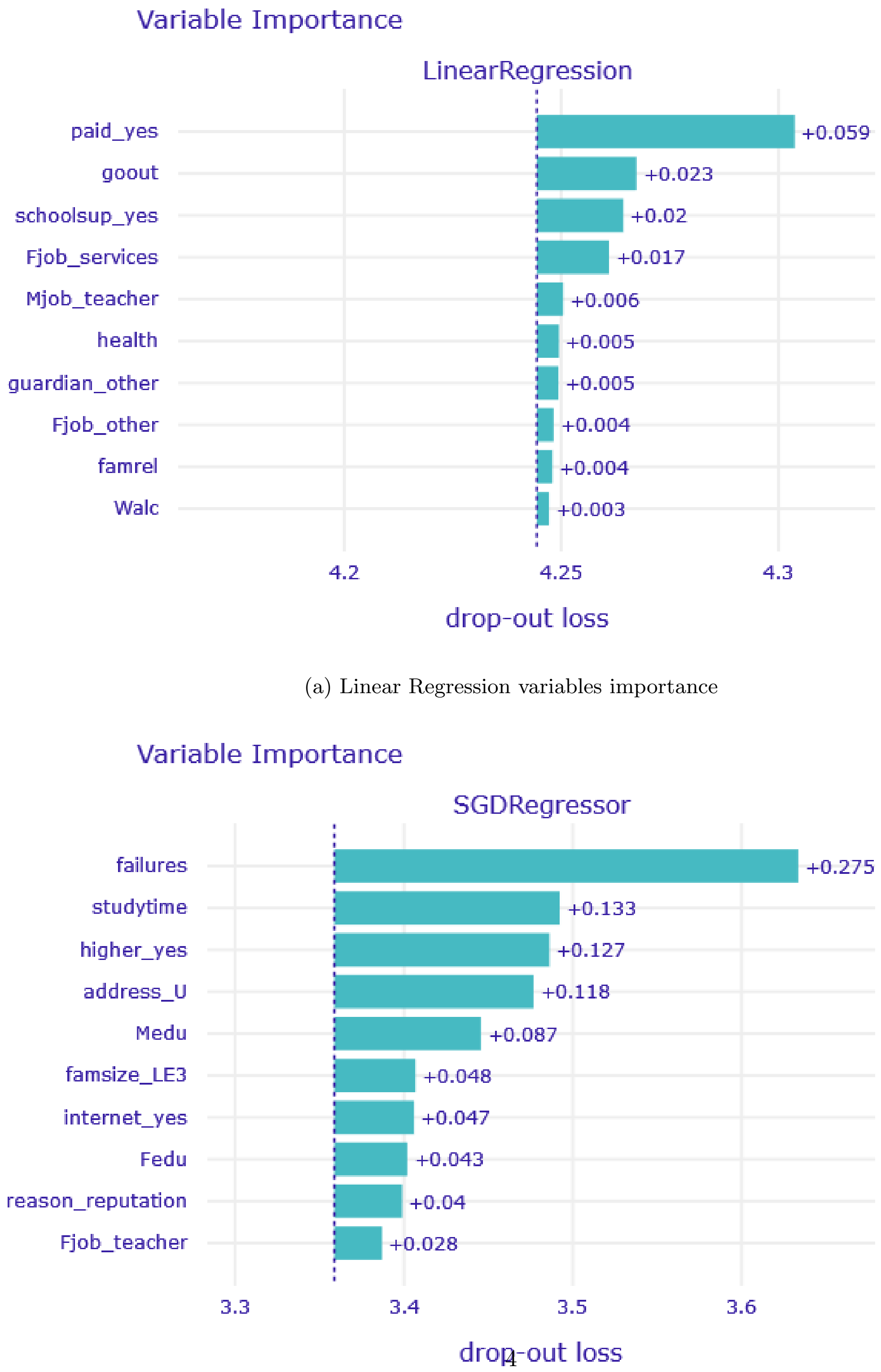


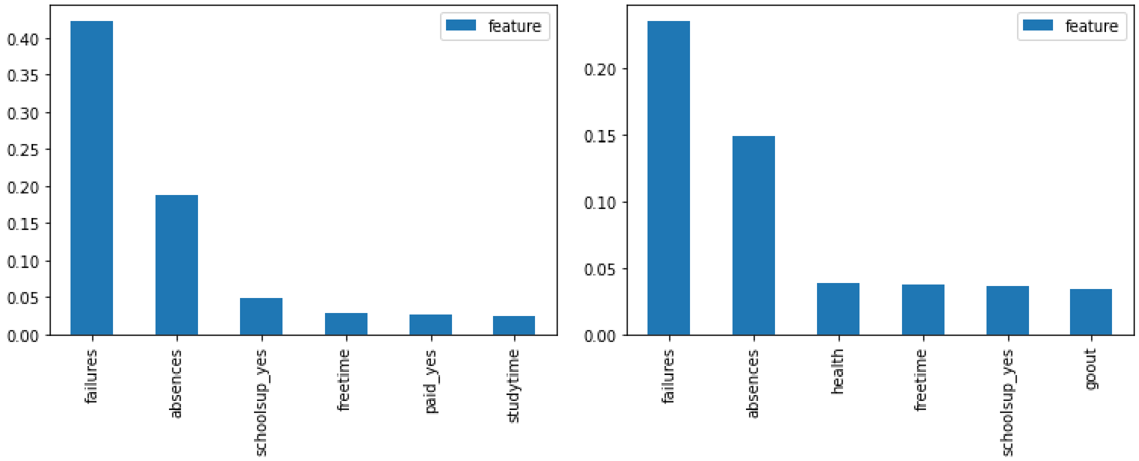
Figure 1: Permutation-based variable importance explanation for Random Forest Regressor with different hyperparameters



(a) Linear Regression variables importance

(b) Stochastic gradient descent variables importance

Figure 2: Permutation-based Variable Importance for linear regression and stochastic gradient descent



(a) number of estimators: 300, depth: 4

(b) number of estimators: 100, depth: 8

Figure 3: Impurity based feature importance of Random Forest Regressors with different hyperparameters.

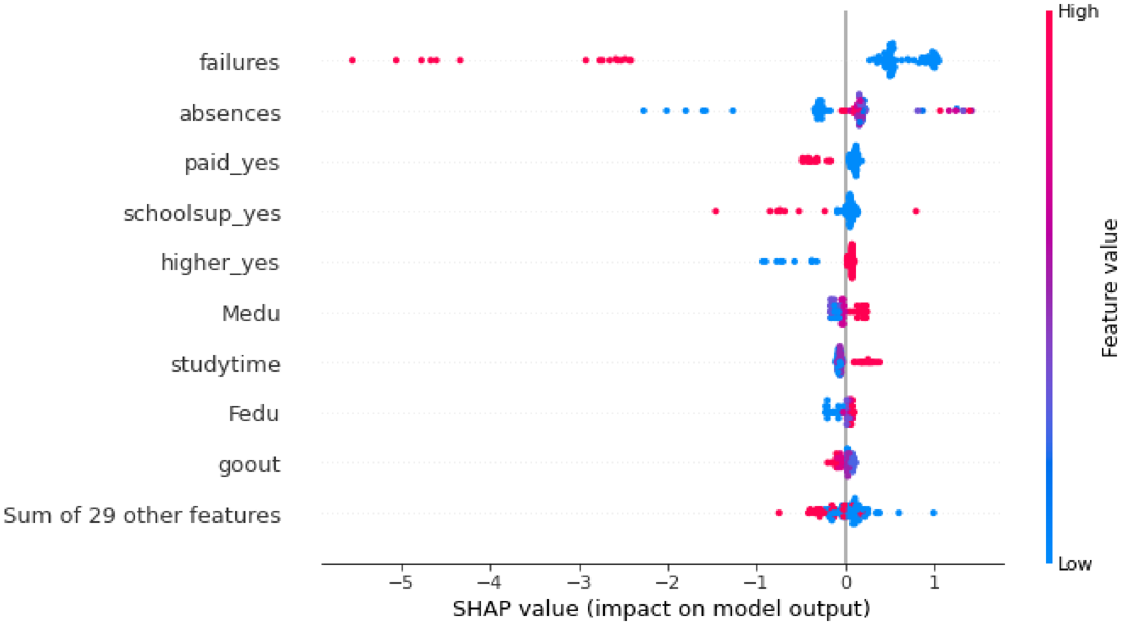


Figure 4: SHAP value for random forest regressor