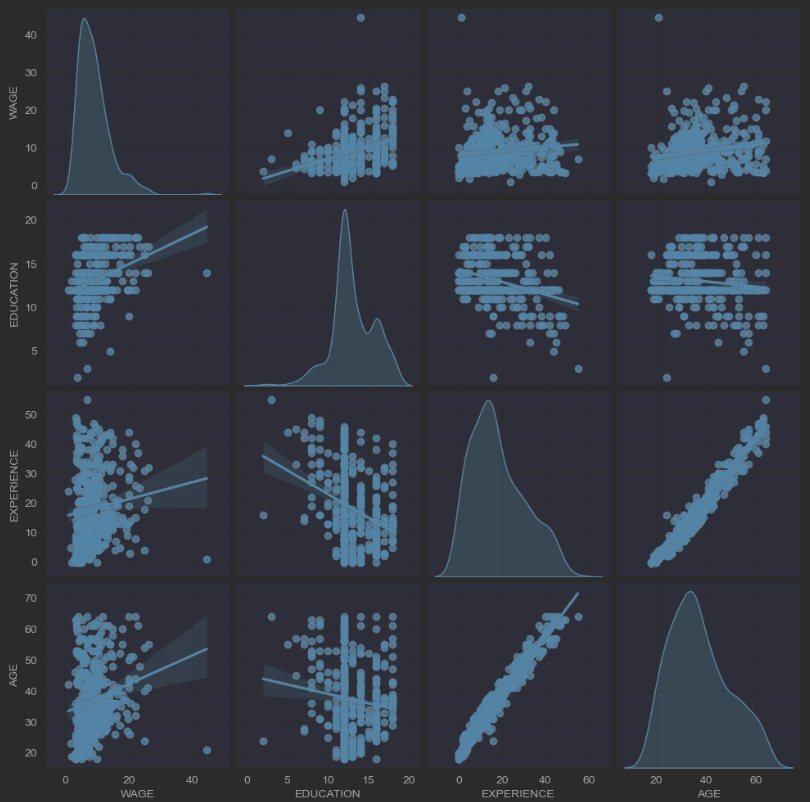
In this work I will concentrate my analysis on the ‘Wages’ dataset by using Ridge regularization with TransformedTargetRegressor estimator.

I have done a detailed analysis of LASSO regularization and LASSO with sequential feature selection in my previous work, so here I will a little bit deeper analysis with Ridge regularization

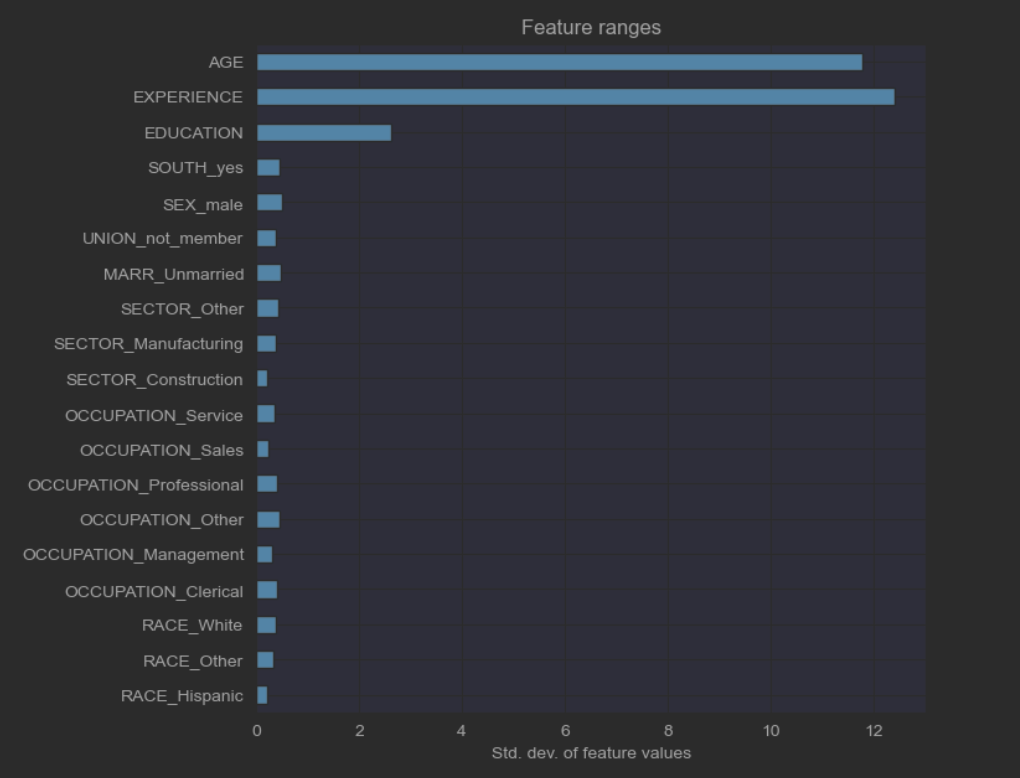
* Correlation

By looking at the variable distributions and at the pairwise relationships between them. Only numerical variables will be used. In the following plot, each dot represents a sample.

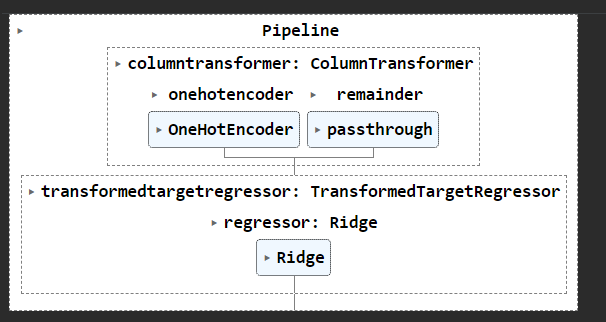


Looking closely at the WAGE distribution reveals that it has a long tail. For this reason, we should take its logarithm to turn it approximately into a normal distribution (linear models such as ridge or lasso work best for a normal distribution of error).

The WAGE is increasing when EDUCATION is increasing. Note that the dependence between WAGE and EDUCATION represented here is a marginal dependence, i.e., it describes the behaviour of a specific variable without keeping the others fixed. Also, the EXPERIENCE and AGE are strongly linearly correlated.

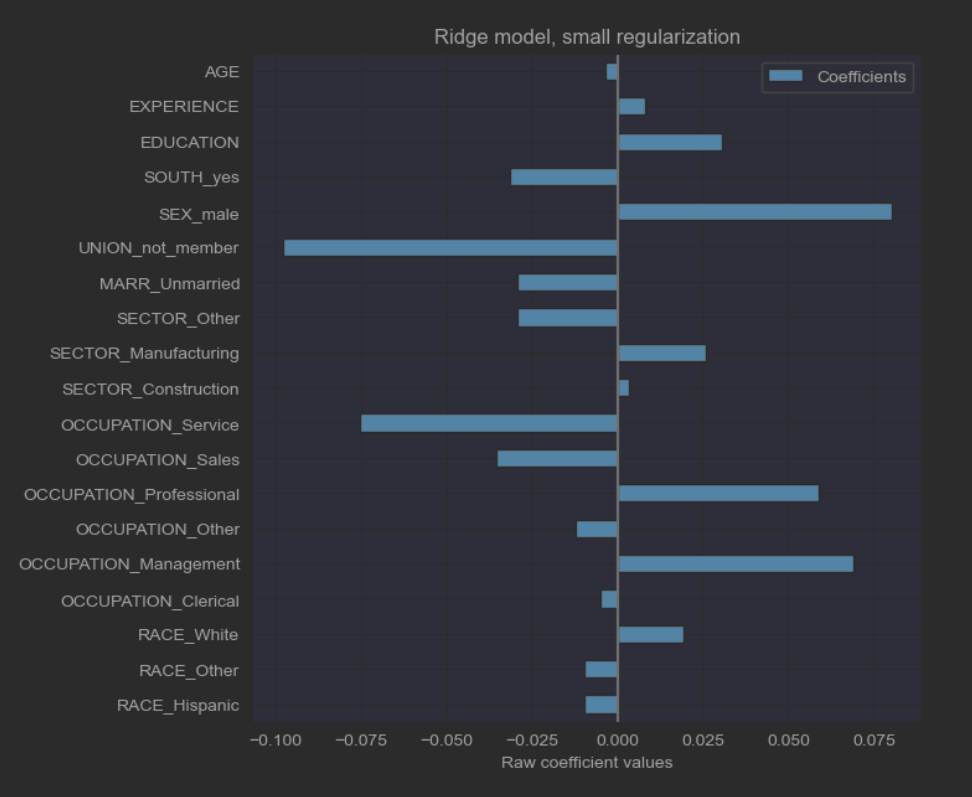


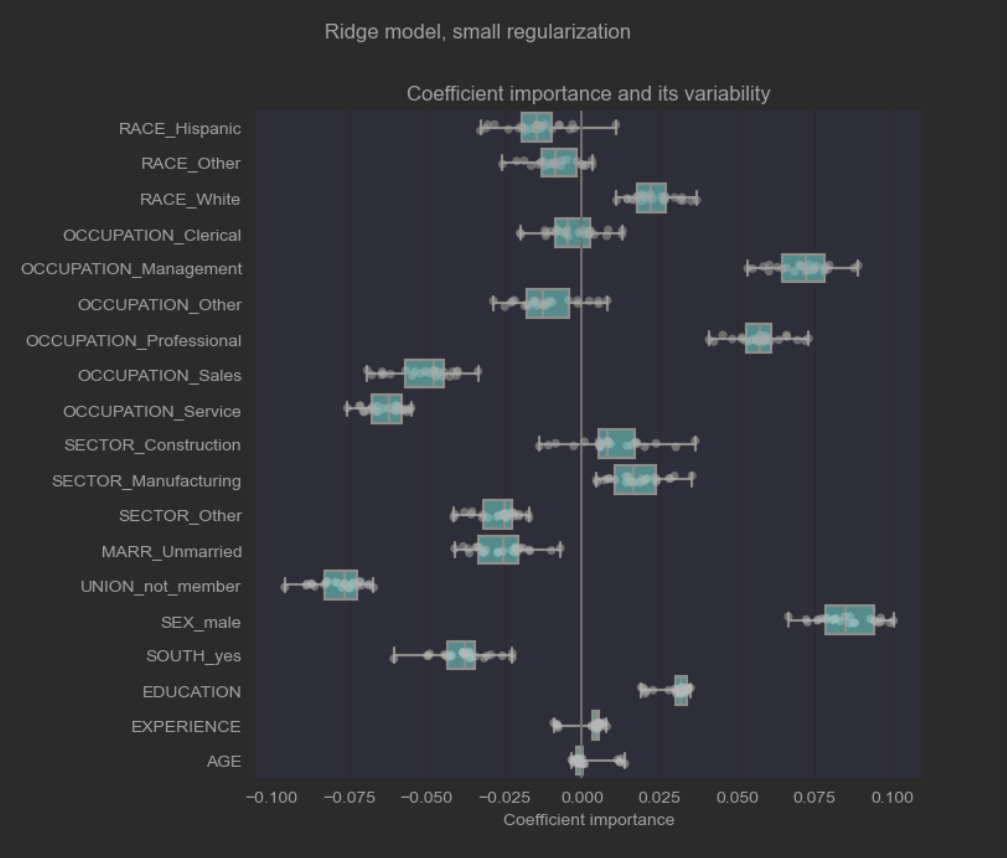
* Built a pipeline with Ridge Model with small regularization



Looking at the coefficient plot to gauge feature importance can be misleading as some of them vary on a small scale, while others, like AGE, varies a lot more, several decades

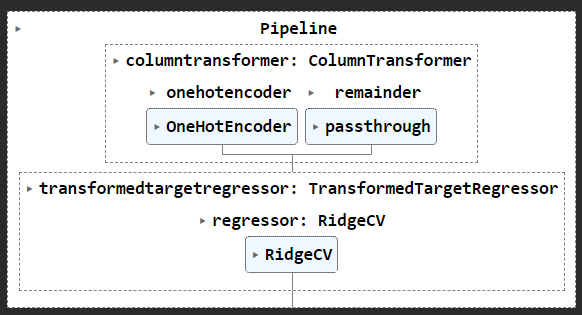
If coefficients vary significantly when changing the input dataset their robustness is not guaranteed, and they should probably be interpreted with caution.



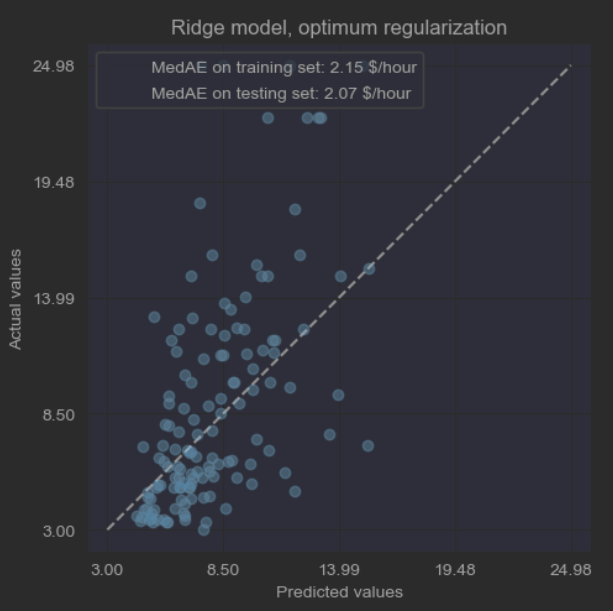


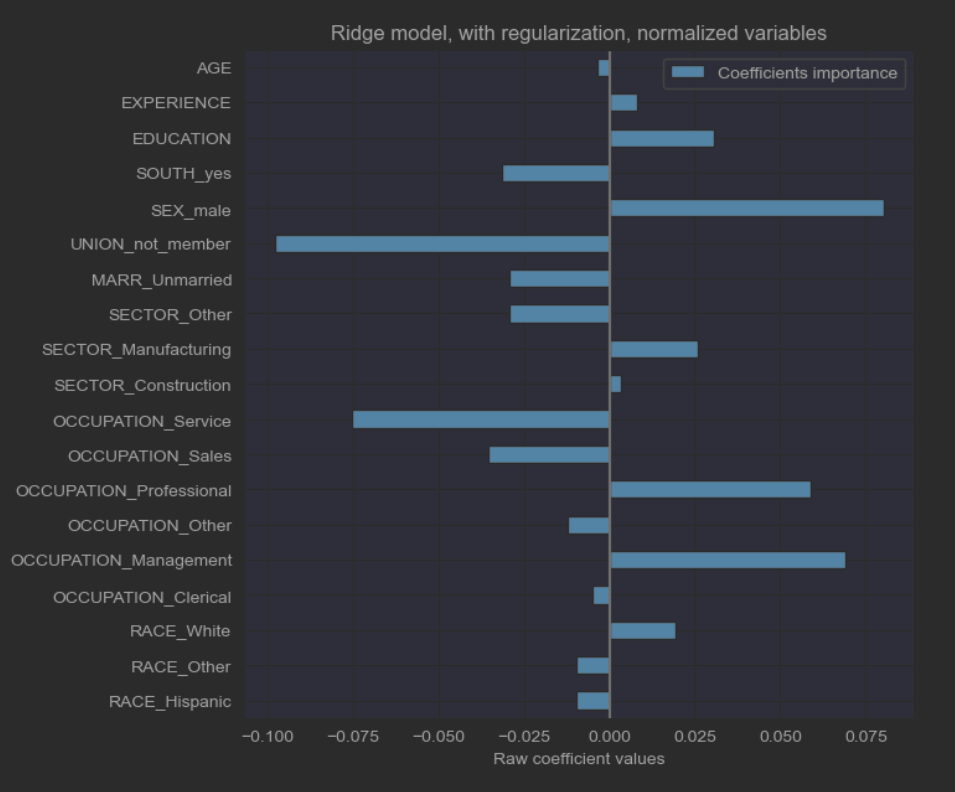
* Built a pipeline of RidgeCV Model with regularization

We limited this regularization to a very little amount. Regularization improves the conditioning of the problem and reduces the variance of the estimates. RidgeCV applies cross validation in order to determine which value of the regularization parameter (alpha) is best suited for prediction. For alpha we use 10.

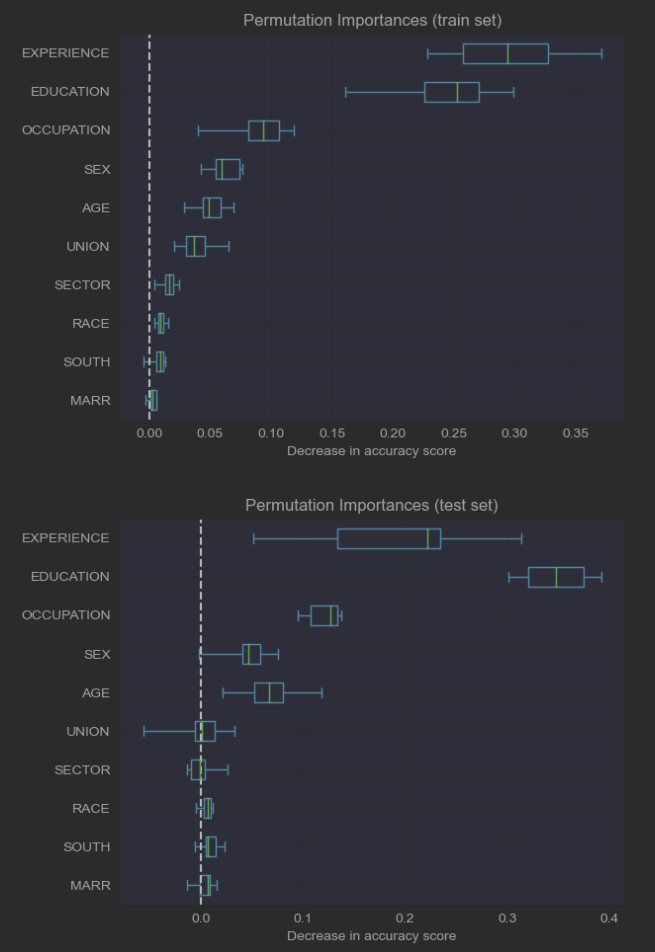


We check the performance of the computed model plotting its predictions on the test set and computing, for example, the median absolute error of the model.



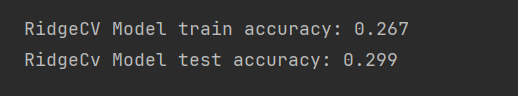


The permutation importance’s of pipeline model are computed on a held out test set.



This shows that the low cardinality categorical feature, Experience and Education are the most important feature. Indeed, permuting the values of these features will lead to most decrease in accuracy score of the model on the test set.

Observing the accuracy score on the training and testing set, we observe that the two metrics are different. Therefore, our model is overfitting.



The coefficients are significantly different from the above small regularization model . AGE and EXPERIENCE coefficients are both positive but they now have less influence on the prediction.

The regularization reduces the influence of correlated variables on the model because the weight is shared between the two predictive variables, so neither alone would have strong weights.

On the other hand, the weights obtained with regularization are more. This increased stability is visible from the plot, obtained from data perturbations, in a cross-validation. This plot can be compared with the previous one.



Hi Vani,

Good work, so much time spent and at the we get a low model accuracy, mine is even lower.

Hi Lisha,

Good work, Nice analysis on Ridge model with small regularization, I started with Ridge Model and finished with RidgeCV model. It looks we have some similarities with Ridge model approach.

Hi Jim,

Nice work, I also noticed the effect of alpha with Ridge regression in my previous work, even I run alpha = 100 vs alpha=1, and the higher alpha model had better accuracy.  Though higher values of alpha reduce overfitting, significantly high values can cause underfitting. For the current model I run RidgeCV with alpha=10, and it shows the model is overfitting, no time to change alpha again.

Thanks Lisha, much appreciated It looks a good structured work, but it consumed most of the time of the week