

In the original dataset, the RandomForest was trained and evaluated. We used Permutation Importances to find the most important feature for our model. We obtained the BirthPlace column with the highest feature importance score and it has continuous values with basic statistics described in cell-10.

Later, we did FE in two different methods. Both R2 score and accuracy are rarely different from the non-FE result. In method 1, we added the new feature column that was engineered after the BirthPlace column by changing values in column based on threshold. While in method 2, we did the same FE with BirthPlace column; however, we replaced original BirthPlace column with our new engineered column using the same threshold as method 1.

For threshold values, we trained model with different thresholds based on basic statistics like mean, median, and quantiles of BirthPlace column. We found out that both R2 score and accuracy stay almost the same.

The very similar results may be because of very simple dataset and our powerful model as we can observe even Linear Regressor can obtain very high accuracy in the last cell. Therefore, FE may not be helpful in this case as our dataset is very clean.