Feature Engineering

The approach taken to find the most important feature contributing to the model predictions was found using sklearn's Random forest features_importance. Even though sklearn's feature importance is said to be biased, permutation importance is computationally expensive. With 380 features, it can become very expensive.

Using features_importances we find that the birthplace of a person is the most important predictor for the population label. I further scaled the data to range of 0-1, and divided the range of 0-1 into equally spaced threshold values. For each of the threshold value, built two models – one with the original feature being replaced and one with the feature being added to the dataset. In this result, the model with the feature replaced results in an MSE value lower than the base score at a threshold value of 0.22.

The improve in MSE can be attributed to the reduction of variance in the birthplace variable and reduction of outliers. Dichotomizing it also makes sense in the context of the problem. But, this could not always be the case. A lower MSE does not mean our model is better, we need to further test it using features_importances or permutation importance again to check the validity of our new feature.