Assignment 10 - Report

Random forest performed on the dataset initially returned birth place #inv count as the most important feature with all others having their values substantially less. Although permutation importance also returned the same feature as important, it should be noted that two additional features also equivalently became important with this method having slight difference.

Two feature drill down methods where applied to the most important feature. First, winsorize method was applied to produce two bins with a threshold identified at 0.25. Evaluation resulted in a slight improvement of MSE/RMSE with addition of feature and deduced performance with replacement. Secondly, adaptive binning method based on quantiles was applied to obtain 2 bins at 75% quantile being threshold (value – 0.477). Evaluation resulted in a slight improvement of MSE/RMSE with replacement but degraded with addition of feature. A probable reason of improvement is that the continuous skewed distribution that contained outliers were regularized to avoid overfitting.

As a future recommendation, we can try doing similar drill down on other important features that are not discretized. Another approach would also be to remove features that affect model negatively using dimensionality reduction techniques.