Peer Review For Vincent Hughes

Matthew Keiser

I like that this is a pretty classic modeling problem with real implications for governments. This is a great example project, especially because you have to bring together several data sources. The approach seems logical from my perspective, looking at the change in immigration as it relates to social and economic metrics. With more time, this would be well served by a deeper dive into the statistics available to try and push back the starting year for the analysis, or to try and find more metrics to try and find more connections. I would also recommend trying some visualizations to look for some structure in the data you could use to improve the model - such as removing outliers.

Using the Lasso algorithm was very appropriate for this - the target is continuous and you are trying to focus in on specific features to try and see what really affects immigration. To gather more information from this type of analysis, I would recommend adding polynomial features using the preprocessing package in sklearn. Try several different orders (2,3,4…) to try and find which one reduces the mean squared error the best. As an extension, I would recommend implementing some of the more advanced algorithms, such as RandomForestRegressor. Another thing to look at would be to try and compare the training and testing error to determine which direction to take the model - more complexity/variance or less variance.