Steven Alnemri

Professor David Biron

DATA 22100: Introduction to Machine Learning

22 May 2024

Predicting Political Instability

Data Preparation

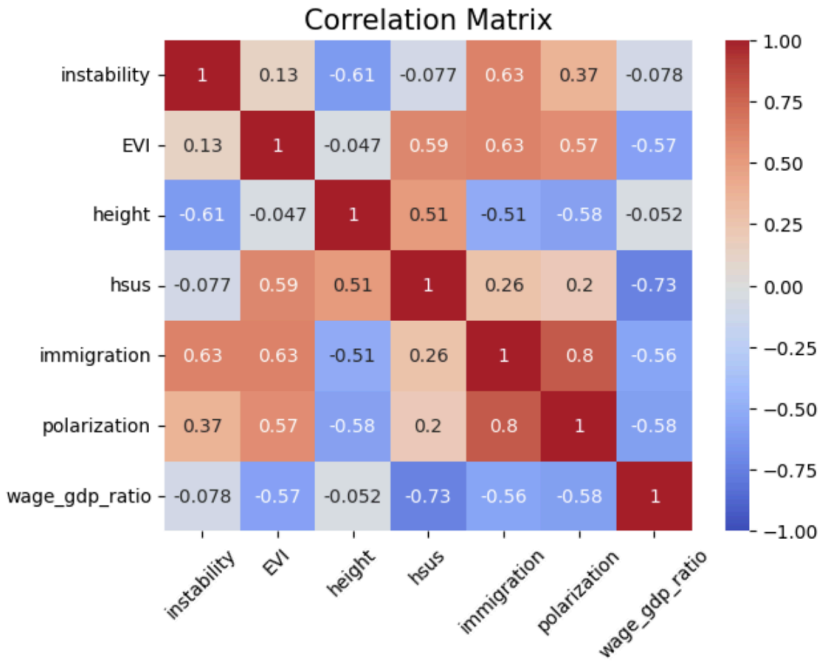
To interpolate missing observations for the csv files, I created a function that rounds years with a floor function, grouped by year, reset the index with a range of years, then interpolated missing values linearly. All fatality subtypes were considered when calculating instability. All log(0) values were replaced with NaN. Only rows with complete data were kept in the final data frame as the range of years varied significantly for some csv files; interpolating seemed too arbitrary and excluding missing rows allows easier model fitting.

	index	Year	Population	instability	EVI	height	hsus	immigration	polarization	wage_gdp_ratio
0	119	1901	76212168.0	0.556828	1474.485794	170.327901	30.705119	13.799276	0.845472	0.651275
1	120	1902	76212168.0	0.586460	1566.712894	170.500641	31.292457	13.896828	0.860217	0.645180
2	121	1903	76212168.0	0.177338	1667.193907	170.672486	31.879503	13.987653	0.868912	0.639086
3	122	1904	76212168.0	0.593733	1739.168108	170.887293	32.216537	14.093615	0.877794	0.627711
4	123	1905	76212168.0	0.629323	1850.340634	171.104785	32.525996	14.199577	0.890111	0.606085
5	124	1906	76212168.0	0.709904	1965.170701	171.330332	33.112025	14.340019	0.898414	0.596688
6	125	1907	76212168.0	0.670425	2051.363070	171.562144	33.724216	14.483824	0.891646	0.607063
7	126	1908	76212168.0	1.142942	2182.491979	171.782321	34.027114	14.576331	0.882256	0.610940
8	127	1909	76212168.0	0.959784	2325.601346	172.015625	34.346872	14.673883	0.873668	0.608747
9	128	1910	92228496.0	0.943791	2419.599594	172.188067	34.935227	14.620061	0.860964	0.602996

Exploratory Data Analysis

The appearance of non-linear relationships between some of the predictors and political instability was the most interesting insight from exploring the data set, such as with height and immigration. The predictors that had the strongest correlation with political instability were immigration (0.63) and height (-0.61). There were also strong correlations between predictors,

such as percent of wealth held by the top 1% (hsus) and wage to GDP ratio (-0.73). EVI, immigration, and polarization were positively correlated with instability, while height, hsus, and wage to GDP ratio were negatively correlated.



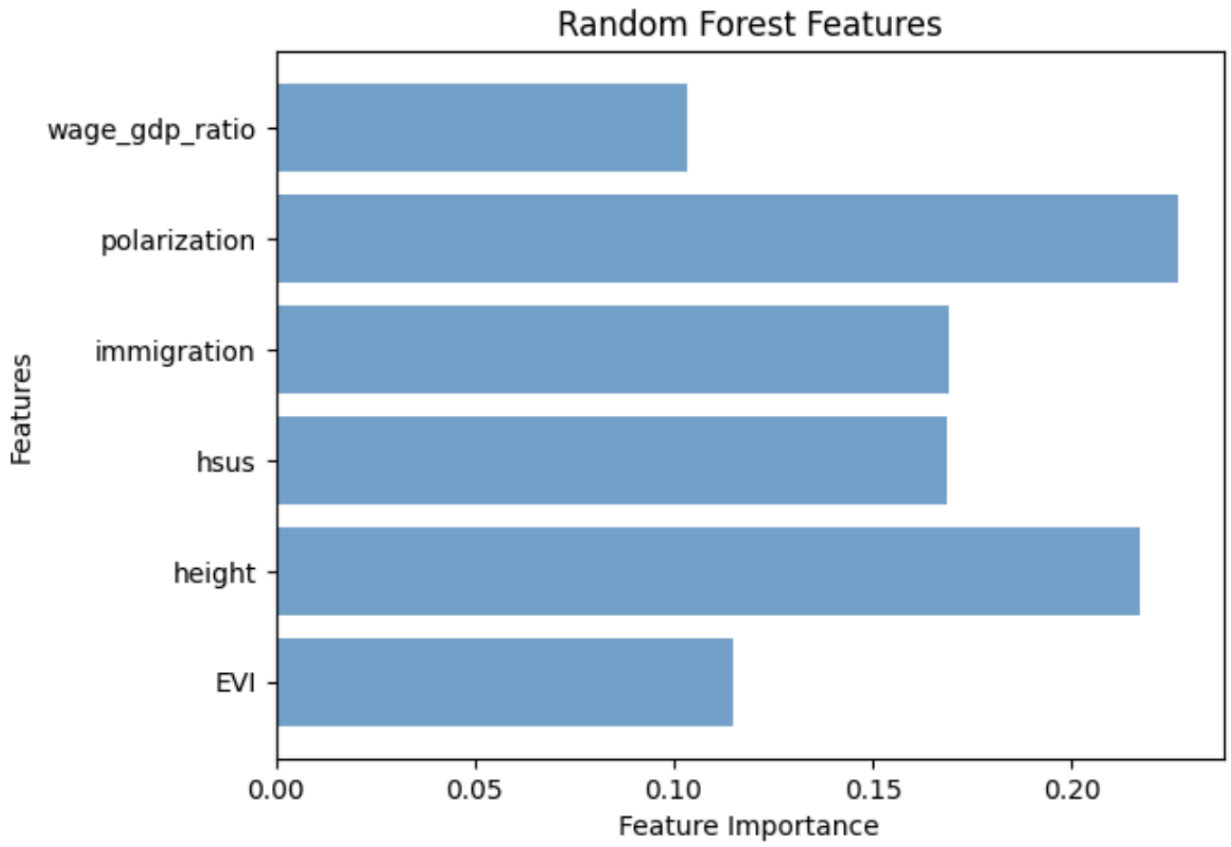
Finding Best Regressor

Models were fitted both for binary classification (logistic Regression and random forest) and linear regression (ridge and lasso). Binary classification models were evaluated using recall score, since incorrectly predicting stability (less than 50th percentile of instability) would leave a society ill-prepared for preventing fatalities. Linear regression models were evaluated using R^2, a measure of variability explained by predictors. From the two best-performing models which we will compare, lasso regression and random forest, the latter produced the most favorable results (see code book for other models). Parameters for both models were optimized using GridSearchCV or RandomSearchCV.

After the regularization strength and feature selection parameters for lasso were optimized to 1e-05 and “cyclic,” no features converged to 0. Regularization was manually altered

for convergence, which left the predictors polarization, immigration, height, and EVI; although the R^2 decreased from 0.64 to 0.62, the latter model is favored for simplicity. Still, this is a relatively weak R^2.

After optimizing the random forest, it produced a recall score of 0.92 and F1 score of 0.91, which performs very well. The most significant features were polarization and, surprisingly, height. The random forest was the best model.



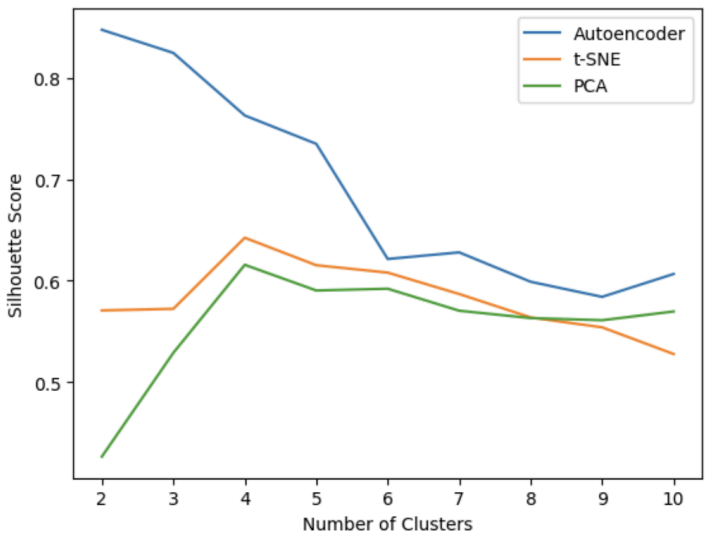
Dimensionality Reduction for Regressor

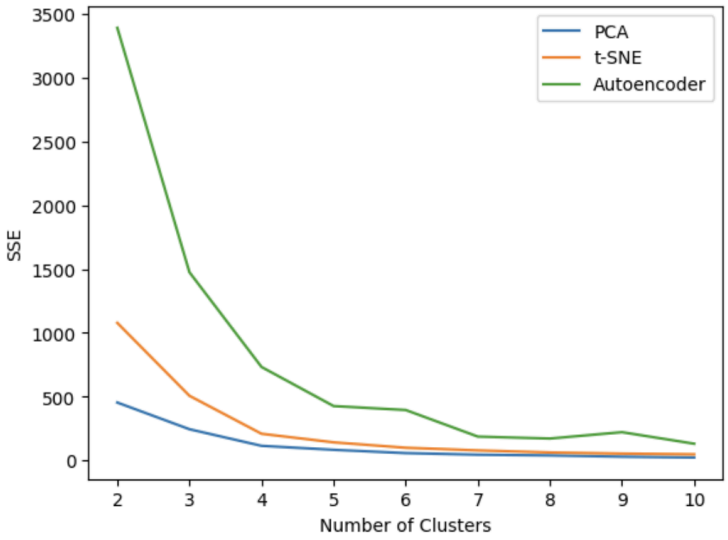
Data was reduced to two dimensions for PCA, autoencoder, and t-SNE reduction. All features were used in the reduction since they either showed importance in the random forest, or we might expect them to contribute to instability without necessarily being as important (wage:GDP). The best dimensionality reduction was t-SNE with Recall 0.95 and F1 0.844. ,

which can be considered an improvement from the random forest without reduction (recall 0.923 and F1 0.9056). This improved recall score could be due to t-SNE capturing more non-linear relationships in the data mentioned earlier, which might be why it performs better on our model. Autoencoders and PCA provide linear changes in the data, which may be less suited for some features as Recall and F1 scores dropped for both of these transformations. The drop in F1 score can be due to loss of information with the reduction.

Dimensionality Reduction for Unsupervised Learning

From the three reductions, the linear autoencoder provided the best results for unsupervised classification for a k-means model. It was found that for each reduction, the optimal number of clusters was 4 (using elbow method). Autoencoder had the smalles SSE and Silhouette score from the three reductions at any amount of clusters.

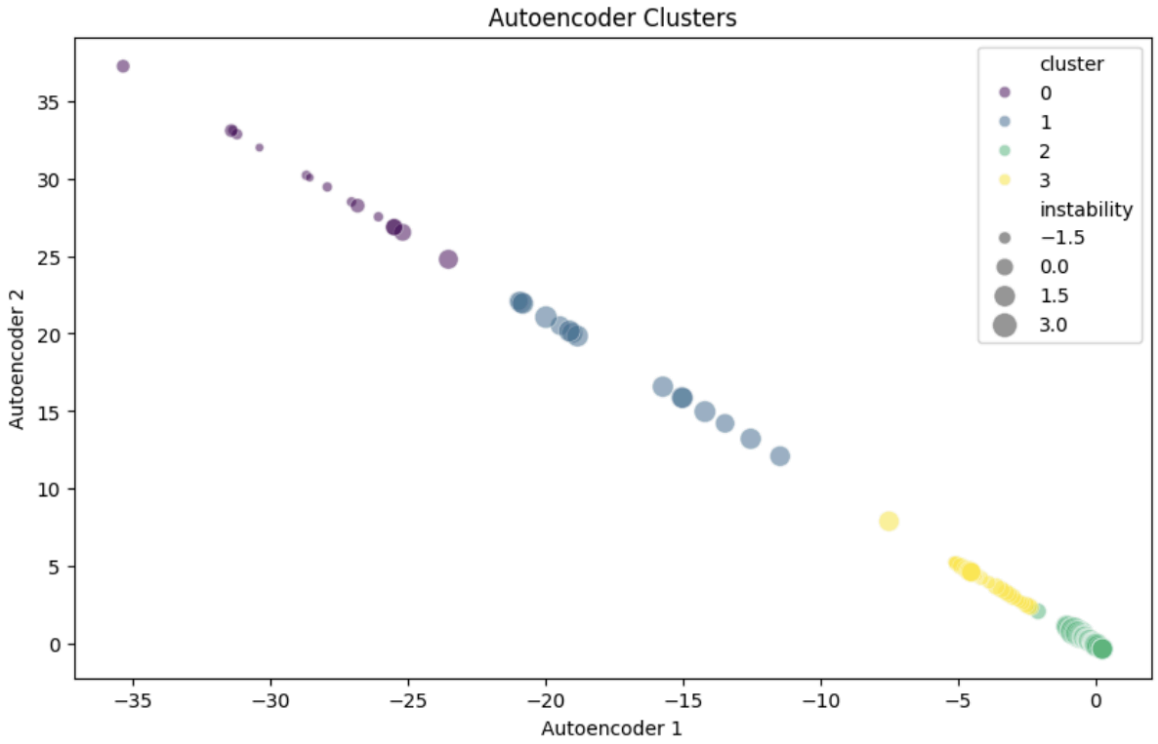




Clusters of Instability Scores

The scatter plot below with color indicating cluster and size indicating instability, there is a clear spatial separation between clusters that. The purple, blue, and yellow clusters have a larger spread, but there are still noticeable difference in Autoencoder 1 values; there is a high density of points in the green cluster, indicating good cluster identification. Instability seems to

increase with Autoencoder 1, but decreases as Autoencoder 2 increases. Difference in instability between clusters is apparent from the varying point sizes.



Real Life Modelling

One potential danger of comparing hundreds of models to find the optimal one is that a model may perform well due to chance rather than actually indicate meaningful conclusions; there are issues of overfitting. One way to mitigate this is using k-fold cross validation and changing the random state when splitting data to can ensure results are generalizeable.

A benefit to this approach of fitting many models is the ability to identify common results across models. Some data sets can have hundreds of features, which makes it difficult to identify the most significant ones to include in a model. Features that persist through many models after lasso regularization or other forms of simplification can inform which features to include in this "best" model and effectively reduce noise in the data when not using PCA, t-SNE, or autoencoders; however, keeping track of this with 100+ models is time-consuming.