CV - Comparing Happiness Models

Vid Stropnik, 63200434

I. Introduction

We are given a dataset, from the World Happiness report, describing the happiness index of different world countries, along with their GDP per capita and the government corruption fraction, as percieved by their population. In this report, we try to model the happineess score, using different approaches. We also pay some attention to raking and comparing the developed models, paying special attention to new bayesian approaches of doing so. All work is reproducible and made available on GitHub.

II. Data Exploration & Modelling

A natural instinct when considering geographical data is to investigate potential regional clusterings of the observed features. Figure 1 visualizes the three dataset variables on the world map. From there, we can easily see that some regional patterns are present in all three. For example, a high trend of happiness if present in Scandinavia. The entirety of South America is colored in a shade, lighter than that of Africa. Notice, also, how the GDP progresses from the inland to the coast of Africa in almost a gradient fashion. We use these coordinates in two of our models: *Continents* and *Subregions*.[1]

Another assumption is that the Happiness score is a composite of a country's economic power and its percieved corruption. We will evaluate this claim with two models.*Linear*, is a simple linear model with feature interactions, while *Poly (3)* is a more complex polynomial model of (max) third degree, with (max) 2nd degree interactions. See more about the models used in Table I. In it, an additional and our best performing model, *Poly-Reg*, is described as well.

All approaches in this work are homoscedastic linear/polynomial models. In all of them, all trainable parameters, except the intercepts, are initialized using *Cauchy, (0,1)* priors. All models, describing country's GDP per Capita and Corruption, also train weights for feature interactions.

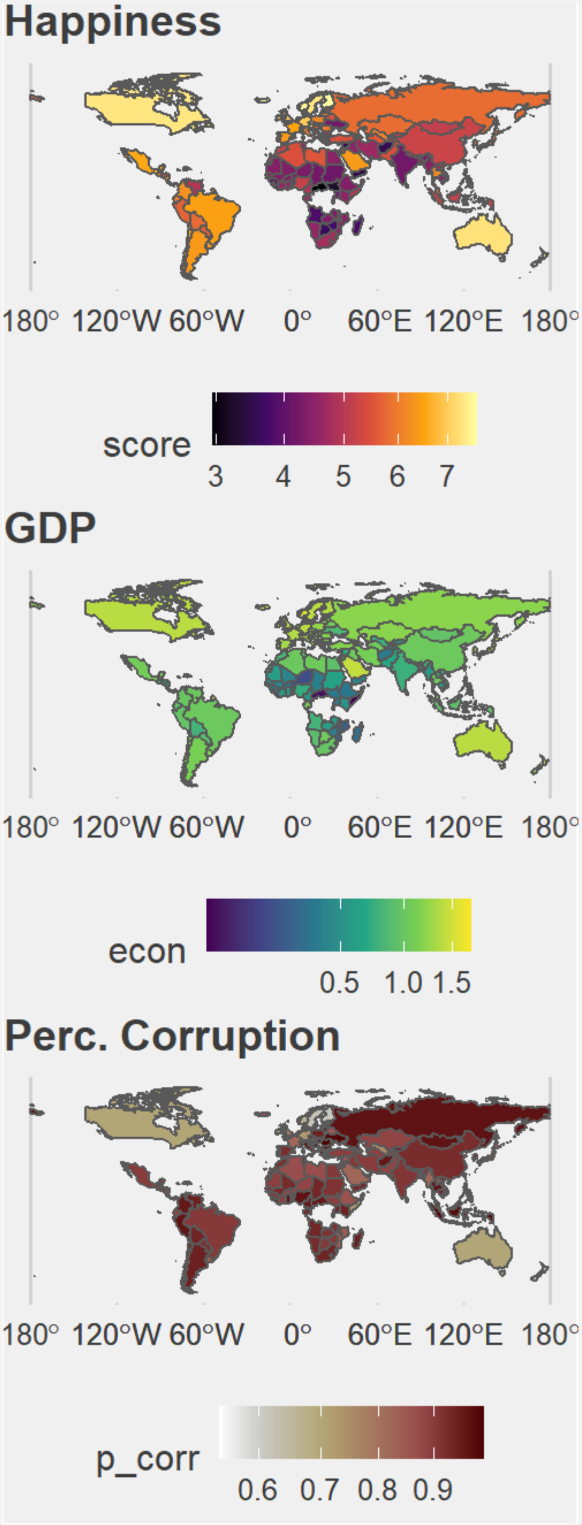
Happiness
180° 120°W 60°W 0° 60°E 120°E 180°
score
3 4 5 6 7
GDP
180° 120°W 60°W 0° 60°E 120°E 180°
econ
0.5 1.0 1.5
Perc. Corruption
180° 120°W 60°W 0° 60°E 120°E 180°
p_corr
0.6 0.7 0.8 0.9

Fig. 1. The geographical plot of the WH dataset country-averages shows clear regional patterns, which we used in some of our models.

III. Model training & evaluation

A. Model Selection

We train the models **on only the 2017 and 2018** data from the World Happiness Report dataset,

[1] Geometric coordinates and continent/subregion categorization courtesy of the `rnaturalearthdata` package for R.

Model	Features, terms	# TP	OOS MSE	LOOIC	Akaike W.
Continents	One-hot encoding of continents	8	0.50 ± 0.01	640 ± 24	$0\|0$
Subregions	One-hot encoding of UN Subregions	20	0.37 ± 0.01	553 ± 25	$0\|0$
Linear	GDP per Capita, percieved corruption	4	0.35 ± 0.01	519 ± 24	$0\|0$
Poly (3)	Polynomial of degree $d \in [0, 3]$ of GDP & corruption	11	0.32 ± 0.01	498 ± 25	$0\|1$
Poly-Reg	*summed terms of Poly (3) and Regions*	30	0.32 ± 0.01	363 ± 24	$1\|$NA

TABLE I
Detailed descriptions of the compared models. The *#TP* column shows the number of trainable parameters. The *Akaike W.* shows the weights, to be assigned to each model, when used in a composite scenario. Given the strong performance of Poly-Reg, an Akaike weiging run withot it included was also performed. Results of the tests are shown on each side of the '|' symbol.

while we reserve the data from 2019 for testing.[2] We measure our model's predictions using mean squared error (MSE). The measurements are shown in Table I and the top subplot of Figure 2. In it, we can observe that the *Poly-Reg* composite model performs very well, while both methods of determining the happiness score from GDP and percieved corruption surpass the pure location-based approaches.

In practice, however, ground truth values of the target variable (here - the happiness score) might not be as easy to come by after using our data for training. This is why we turn to bayesian approaches of model evaluation and comparison. Here, we use the Leave One Out Information Criterion (LOOIC). When using LOOIC, we strive to weight posterior draws in a way that would adjust the posterior to what it would have looked like if the target (eg. happiness score) had not yet been observed. Particullary, the weight w_i should be equal to the smoothed inverse of the likelihood that the observed posterior yielded this draw, or

$$w_i = \frac{1}{p(y_i|\theta)}. \tag{1}$$

The full explanation of the method exceeds the scope of this text and can be found here. The LOOIC values of our models are visualized in the bottom subplot of Figure 2. Notice that we were able to accurately re-rank the models in virtually the same way as we did with the holdout error estimation method, while not having to withold any data from the training process.

B. Model Ensembling

Akaike weighting is a process for determining how decisions made by multiple models should be mixed together in an ensembling scenario. The Akaike weight of a model *i* is

$$w_i = \frac{\exp(-\frac{1}{2}\Delta\mathbf{IC}_i)}{\sum_{j=1}^{m}\exp(-\frac{1}{2}\Delta\mathbf{IC}_j)}, \tag{2}$$

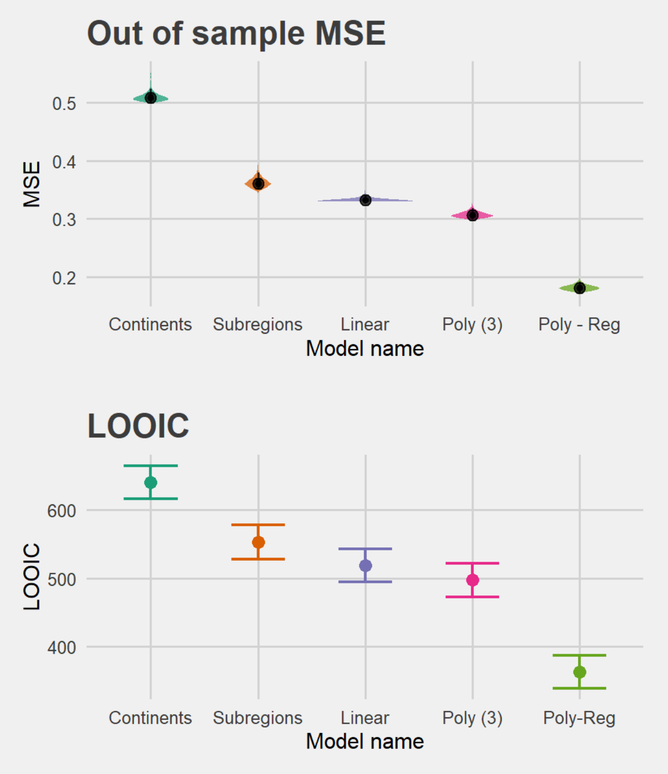

Fig. 2. Comparison of results of Holdout estimation using MSE and LOOIC. We can see that the trends revealed are similar, whereas the second approach is entirely Bayesian.

where $\Delta\mathbf{IC}_i$ is the difference in any information criterion (here, LOOIC, but results were the same when using the WAIC) between any of *m* models and *the best model*. These weights can be interpreted as the likelihood of the model *i* making the best prediction on unseen data. The derived Akaike weights from two different weight calculations are shown in Table I (see caption). We can see that, even when not considering the expectedly best model, *Poly-Reg*, the weighing always resulted in a single fully-weighted model. This concludes our work with a valuable insight about Akaike weighing: it does not correspond to understanding the *semantic* preference of any given model to the observed posterior sample.

[2] We are sacrificing the model's expressive power by limiting its number of training samples. We do this intentionally, to showcase the power of LOOIC and compare it with traditional frequentist methods.