Course Project

July 11, 2021

1 Required

- **Main objective of the analysis**
  I will analyze the World Happiness Report. It quantified the level of happiness from 0 (Worst) to 10 (Happiest). To be specific, 1) we will get the best linear regression model to predict the level of happiness with features, and 2) we will take a closer look at what features have a significant impact on the level of happiness.

- **Brief description of the data set**
  The happiness scores and rankings use data from the Gallup World Poll. There are 149 rows (objects, countries), and 20 columns (features). The target is Ladder score which is the numeric indicator of the level of happiness. And there are two categorical features that are 'Country name', and 'Regional indicator'. (Figure 1a, 1b)

- **Brief summary of data exploration**

  1. Data cleaning, Delete unused features to predict the Ladder score.
  2. Plot the relationship between Ladder score and other variables and find the higher-correlation features (Figure 2). According to pairplot, the Generosity seems not to have a strong correlation.
  3. Change the categorical variable to numeric variables. (Figure 3)
  4. Standardize the features. (Figure 4)
  5. Check the normality of the target value and normalize if it's skewed.
     Ladder score's p-value is 0.526. So, we don't need to normalize the target value. (Figure 5)

	count	mean	std	min	25%	50%	75%	max
Ladder score	149.0	5.532839	1.073924	2.523	4.852	5.534	6.255	7.842
Standard error of ladder score	149.0	0.058752	0.022001	0.026	0.043	0.054	0.070	0.173
upperwhisker	149.0	5.648007	1.054330	2.596	4.991	5.625	6.344	7.904
lowerwhisker	149.0	5.417631	1.094879	2.449	4.706	5.413	6.128	7.780
Logged GDP per capita	149.0	9.432208	1.158601	6.635	8.541	9.569	10.421	11.647
Social support	149.0	0.814745	0.114889	0.463	0.750	0.832	0.905	0.983
Healthy life expectancy	149.0	64.992799	6.762043	48.478	59.802	66.603	69.600	76.953
Freedom to make life choices	149.0	0.791597	0.113332	0.382	0.718	0.804	0.877	0.970
Generosity	149.0	-0.015134	0.150657	-0.288	-0.126	-0.036	0.079	0.542
Perceptions of corruption	149.0	0.727450	0.179226	0.082	0.667	0.781	0.845	0.939
Ladder score in Dystopia	149.0	2.430000	0.000000	2.430	2.430	2.430	2.430	2.430
Explained by: Log GDP per capita	149.0	0.977161	0.404740	0.000	0.666	1.025	1.323	1.751
Explained by: Social support	149.0	0.793315	0.258871	0.000	0.647	0.832	0.996	1.172
Explained by: Healthy life expectancy	149.0	0.520161	0.213019	0.000	0.357	0.571	0.665	0.897
Explained by: Freedom to make life choices	149.0	0.498711	0.137888	0.000	0.409	0.514	0.603	0.716
Explained by: Generosity	149.0	0.178047	0.098270	0.000	0.105	0.164	0.239	0.541
Explained by: Perceptions of corruption	149.0	0.135141	0.114361	0.000	0.060	0.101	0.174	0.547
Dystopia + residual	149.0	2.430329	0.537645	0.648	2.138	2.509	2.794	3.482

(a) Numerical variables

	Country name	Regional indicator
0	Finland	Western Europe
1	Denmark	Western Europe
2	Switzerland	Western Europe
3	Iceland	Western Europe
4	Netherlands	Western Europe
...	...	...
144	Lesotho	Sub-Saharan Africa
145	Botswana	Sub-Saharan Africa
146	Rwanda	Sub-Saharan Africa
147	Zimbabwe	Sub-Saharan Africa
148	Afghanistan	South Asia

149 rows × 2 columns

(b) Categorical variables

Figure 1: Brief description of the data set

- **Summary of training at least three linear regression models** We tested 4 kinds of models such as *Linear Regression, Lasso, Ridge, ElasticNet* and in order to prevent overfitting we used cross validation. To evalueate the best model, we compare the results with rmse(root mean squared error). This is the result of rmse of the four models. (Figure 6)

- **Explanation of your final regressions model** Overall, all models showed the low rmse values. But the best one was Ridge ($ridgeCV.alpha\_ : 21.056578947368422 ridgeCV\_rmse : 0.501828479703691$). And the r2_score of the model was 0.7610082843481112.

- **Summary Key Findings and Insights** When we analyze the coefficients, it shows the interesting result. The regions are the main fators that people feel happy. I think it's resonable because happiness is decided based on the relationship. So, some regions might have a culture that put big emphasis on the relationship (Figure 7)

- **Suggestions for next steps** We may find more meaningful results if we exclude the regions to predict the target. Because regions can be the result of different features such as GDP, Healthy, Social support, etc. Then, we will know the factor that has the most impact on the happiness rather than the regions.

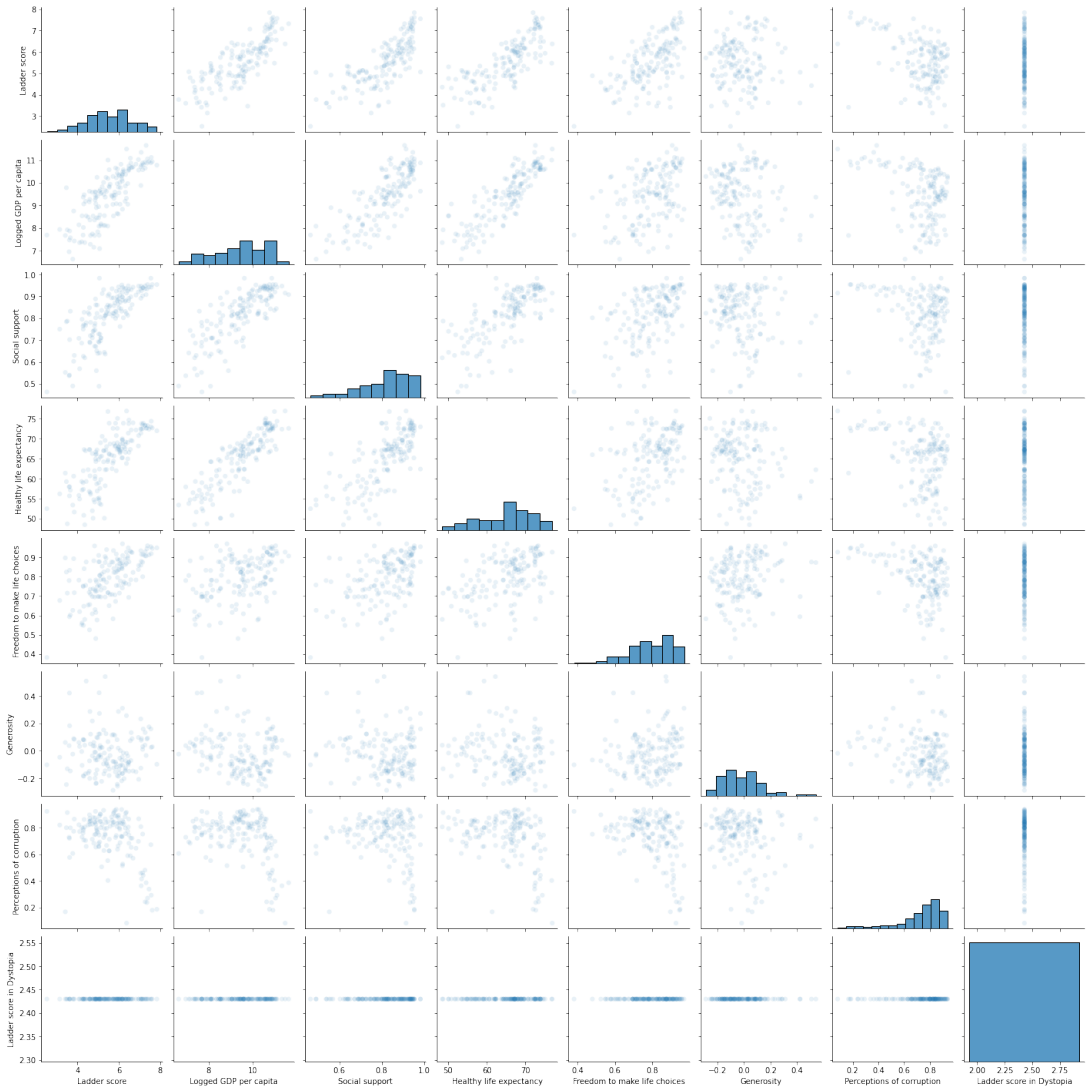

Figure 2: Pair plots among features

	Regional indicator_Central and Eastern Europe	Regional indicator_Commonwealth of Independent States	Regional indicator_East Asia	Regional indicator_Latin America and Caribbean	Regional indicator_Middle East and North Africa	Regional indicator_North America and ANZ	Regional indicator_South Asia	Regional indicator_Southeast Asia	Re indicato S
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	
144	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
145	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
146	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
147	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
148	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	

149 rows × 10 columns

Figure 3: OneHotEncoder

	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Ladder score in Dystopia	Regional indicator_Central and Eastern Europe	Regional indicator_Commonwealth of Independent States	Regional indicator_East Asia	Regional indicator_Latin America and Caribbean
0	0.018610	0.032016	-0.308185	0.703047	3.582410	0.786301	0.0	-0.410535	-0.307794	-0.172345	-0.343918
1	-1.388172	-0.356302	-0.802244	-0.304141	0.617315	0.192346	0.0	-0.410535	-0.307794	-0.172345	-0.343918
2	-1.570860	-2.207006	-1.306074	-1.627636	2.808907	-0.031788	0.0	-0.410535	-0.307794	-0.172345	2.907670
3	0.712320	0.792126	0.068466	-0.587153	-0.626736	0.663027	0.0	-0.410535	3.248931	-0.172345	-0.343918
4	-0.000753	-0.042343	0.267250	-2.568233	-0.343118	0.141916	0.0	-0.410535	-0.307794	-0.172345	-0.343918
...	...	...	...	...	...	...	...	...	...	...	...
99	0.502692	0.420333	0.678559	-0.670392	-0.130405	0.478116	0.0	2.435843	-0.307794	-0.172345	-0.343918
100	-0.227219	0.445119	0.373360	-1.444511	-1.000596	0.562167	0.0	-0.410535	-0.307794	-0.172345	2.907670
101	1.683009	1.155657	1.243612	0.752990	0.585086	-2.037787	0.0	-0.410535	-0.307794	-0.172345	-0.343918
102	0.149102	-0.909861	0.724362	-0.029454	-0.104621	0.976814	0.0	2.435843	-0.307794	-0.172345	-0.343918
103	0.050602	0.436857	-1.122253	-0.329113	-0.343118	0.747077	0.0	-0.410535	-0.307794	-0.172345	-0.343918

104 rows × 17 columns

Figure 4: Standardized values

```
1  from scipy.stats.mstats import normaltest
2
3  normaltest(Y.values)
```

```
NormaltestResult(statistic=masked_array(data=[1.2847842582602853],
                 mask=[False],
           fill_value=1e+20), pvalue=array([0.52603258]))
```

Figure 5: Check the normality of the target

	RMSE
Linear	0.504227
Ridge	0.501828
Lasso	0.588123
ElasticNet	0.570016

Figure 6: Check the rmse of the models

	0	1
0	Logged GDP per capita	-0.090020
1	Social support	-0.028215
2	Healthy life expectancy	-0.028215
3	Freedom to make life choices	-0.028215
4	Generosity	-0.028215
...	...	...
162	Regional indicator_North America and ANZ	0.039951
163	Regional indicator_South Asia	0.105882
164	Regional indicator_Southeast Asia	0.210569
165	Regional indicator_Sub-Saharan Africa	0.234631
166	Regional indicator_Western Europe	0.257199

167 rows × 2 columns

Figure 7: Coefficients. Regions are the main factors