CIS 520 Fall 2019, Project Report

Classical Models to Predict Self-Reported Life Satisfaction Using United Nations World Happiness Report 2019

Team Members:

- Noah Hindes; Email: `hindesn@seas.upenn.edu`
- Victor Phun; Email: `phunv1@seas.upenn.edu`

Abstract

We sought to extend the analysis of a United Nations World Happiness Report article's findings on the six most significant factors driving the reported satisfaction (”Life Ladder” survey score) of respondents from around the world. In order to do so, we first replicated the OLS performed on a limited feature set in the article, achieving nearly identical regression parameters. However, we sought to provide a predictive model using all data supplied with the report, including independent consideration of feature selection and regularized representations of the data. In particular, we applied ElasticNet regularization before also regressing on reconstructed copies of our dataset generated using subsets of its principal components. We also implemented stepwise and streamwise feature selections and an adapted k-means clustering method as possible predictive models.

Findings: In general, our enhancements to the article's OLS yielded very modest improvements in the mean-squared error on predicted Ladder scores, suggesting the predictive (rather than purely retrospective) value of the data and the importance of careful feature selection and regularization on features beyond the six discussed in the article. However, based on hyperparameter tuning and feature selection, we see nothing that contradicts the importance that the article places on its six talking-point features–instead, our work strongly suggests that a few more features may be worthy of consideration and lays a sound framework for predicting future values in addition to commenting on recent trends.

1 Motivation

The United Nations Sustainable Development Network publishes an annual report on self-reported happiness using the Cantril Ladder assessment of overall life satisfaction.

In addition to comparing nations' scores on this metric and considering national and regional variations over time, the report also includes supplementary data on per capita GDP, life expectancy, and other scores from respondents that are believed to impact the overall satisfaction rating (e.g. average categorical response to ”have you donated to a charitable organization in the last month”?).

Although the article relies on a robust dataset and makes some comments on the impact of global trends (e.g. 2008 financial crisis) on self-reported satisfaction, work to understand predictors is limited. **Our objective is to employ classical machine learning techniques in an attempt to predict the Cantril Ladder score based on the other inputs provided.** The main data is available in this flat file, with a discussion of trends and comparison among countries and regions found here.

2 Related Work

The main article presents only one model, an ordinary least squares regression, in order to suggest the relative importance of six selected features (the only ones included in the model) [3]. Although the two appendices go into detail on why some variables are not included (e.g. near-zero correlation coefficient on measure of government effectiveness [2]), there is little discussion of predictive modeling **(no attempt to see how trained model generalizes)** and no in-depth analysis of covariance among features (only with dependent variable). Some additional regressions (e.g. country-specific or with a particular collection of inequality statistics) are presented in the first appendix–all are OLS[1].

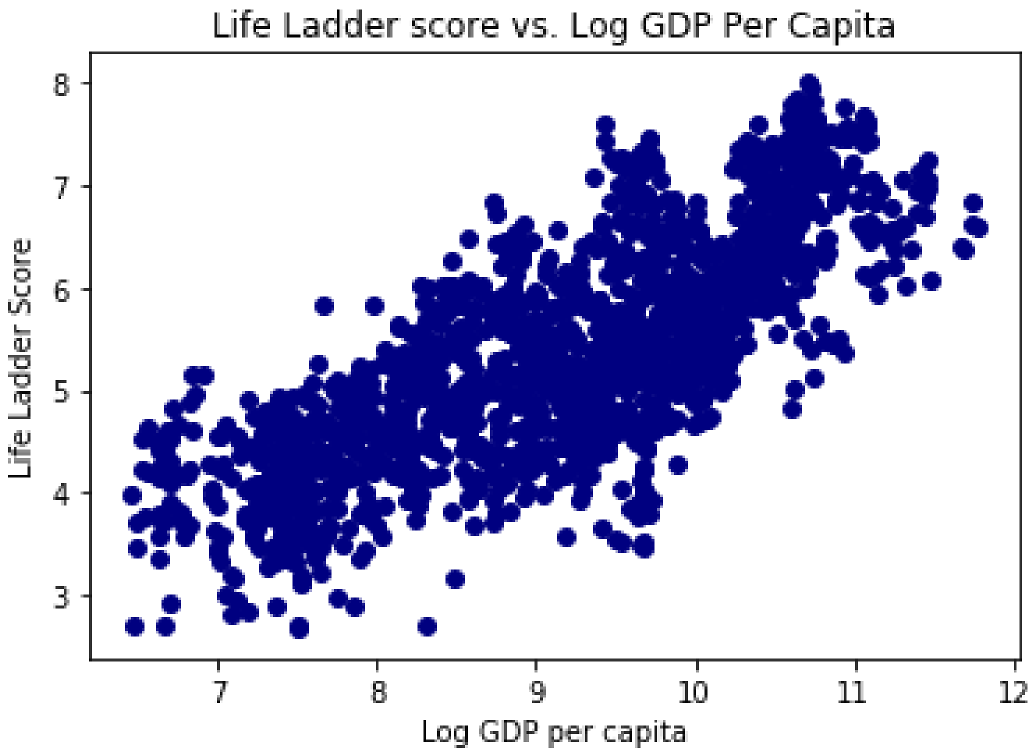
3 Dataset

Please find original file here.
n = 165, p = 11. Features are all continuous variables; Only the Ladder score is complete for the dataset; all other columns suffer varying degrees of sparseness.

Note that many columns will not appear in our analysis; we use only those 11 features listed in the notebook. In particular, we discard the yearly columns from 'people can be trusted' ratings which are not only likely to be duplicative but also extremely sparse, with only a minority of records containing these scores. Further, we discard columns that only contain standard deviations for other columns of interest.

Scales, Distributions: Our data come on a variety of scales and we make no assumptions about their distributions. In addition to disparate values like Log GDP and Healthy Life Expectancy, we see survey responses that may represent a rate (e.g. proportion of respondents reporting postive/negative affect) or a score (e.g. Democratic Quality of government), the latter of which may take on negative values.

Preprocessing: Choosing between data imputation and simply dropping missing values was the only major decision to make here. Given the small size of the dataset (1705 total entries, with 165 for 2018), we felt that dropping roughly 10% of our records due to missing values would be unacceptable, and instead chose to use scikit-learn's simple **mean imputation** (detailed in notebook).



Heterogeneity of Ladder Scores vs. GDP

This plot in particular suggests the need for a sophisticated model to analyze Ladder score drivers–although we can see the correlation between Log GDP and Ladder score, it is perhaps surprising to note the width of the distribution, i.e. number of country-year entries which share same approximate Ladder despite wildly different GDP.

4 Problem Formulation

This is a **supervised** task (all data used has corresponding Ladder scores). We will use **regression** techniques as well as a clustering method adapted to our continuous target values. The original article employs regression to illustrate broad importance of some selected features. We view this dataset as an opportunity to apply a variety of regression methods in which we **determine both a subset of measured features and appropriate weights that estimate reported life satisfaction in a manner that generalizes well.** By extension, we hope to identify the strongest predictors of higher life satisfaction *and* identify strong correlations among features. These may or may not coincide with intuition, and relative importances may or may not ultimately agree with those put forth in the World Happiness Report 2019.

Framing: Given that we want to predict continuous values, linear regression is the most intuitive model with which to start. In terms of alternatives, polynomial and other functional forms would present a high risk of overfitting even with much larger datasets, and we certainly do not have enough data to justify a neural network approach. For all of our regression training, we used the **ElasticNet** loss function. We wanted the benefit of the L1 norm in terms of our feature selection on an extended dataset and needed L2 to penalize large weights (small dataset, features on very different scales and not normalized)

5 Methods

Baseline method: We used **Ordinary Least Squares** as a baseline. This represents the simplest possible linear regression, and is also the method employed in the World Happiness Report article to comment on most significant features (replicating its approach was our first step).

Other methods:

- **ElasticNet** linear regression on additional columns with **grid search over penalty hyperparameter values**
  ElasticNet is the industry standard for regularization terms in linear regression. Because World Happiness Report article made no attempt to produce a generalizable model, regularization on both weights and number of features is a sensible first step to avoid overfitting. We iterate over an array of penalty weights, training with each one and assessing MSE of predicted Ladder values.

- **Feature selection methods (streamwise and stepwise)**
  Building upon the implementation of ElasticNet, we want to determine not only whether or not additional features enhance the model, but to determine an optimal subset to use. We code a streamwise (order-dependent) forward selection first as a naive way to see performance of a subset of features distinct from the 6 or 11 previously assessed.
  Next, we employed the **mlextend** library's **SequentialFeatureSelector** for a more thorough forward selection process (not just optimal number, but optimal subset of features).

- **Principal Component Regression**
  Another regularization measure that may or may not allow or model to generalize better to new data. In particular, we iterate over the hyperparameter k (# of components, all features available) and perform both OLS and ElasticNet regressions. In theory, we may be able to compress autocorrelated features into a lower number of meaningful dimensions and regress on the reconstructed data.

- **Modified K-Means Clustering**
  The only unorthodox approach we adopt here is a modified form of k-means clustering in which we:
  1) cluster on training data 2) assign mean of constituent points' Ladder score to each centroid
  3) use centroid's computed mean Ladder score to estimate Ladder score of new points assigned to (nearest in Euclidean terms) clusters
  4) Assess MSE of this estimate

Implementation Details: Please see associated notebook submission for full implementation.
Packages other than native Python:

- **mlxtend** for stepwise forward selection **numpy** to handle dataset and other arrays
- **pandas** for dataset ingestion
- **sklearn** for imputation, splitting of data into train and test sets, OLS, ElasticNet, cross validation, PCA and K-means clustering

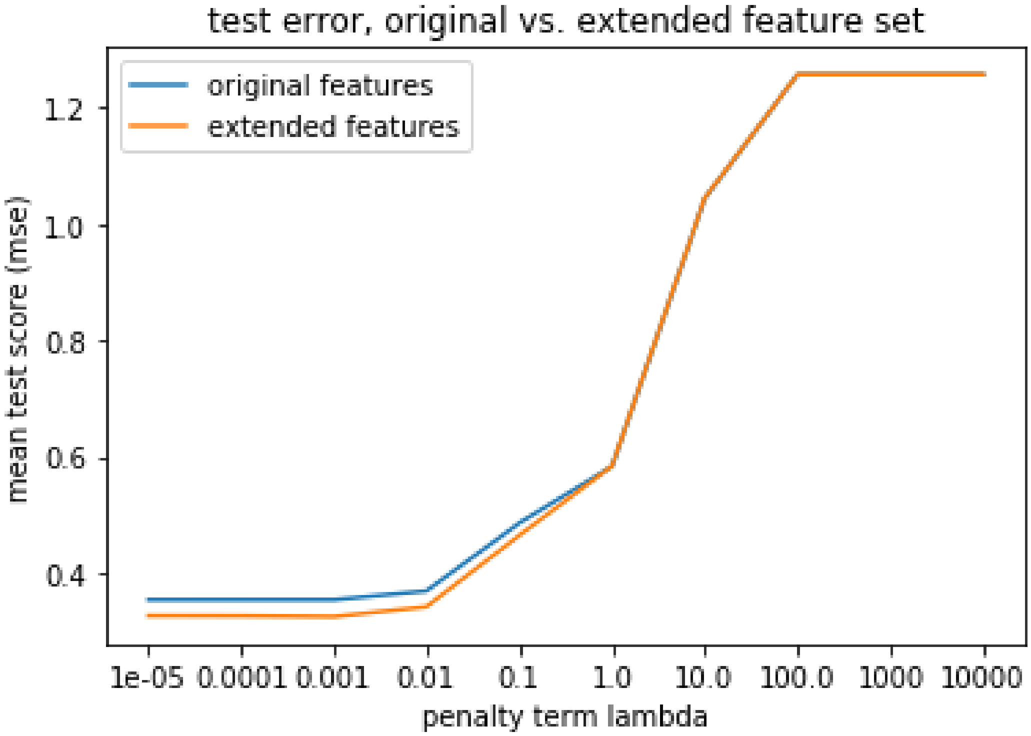
6 Experiments and Results

- **Recreating the UN's OLS findings**
  We used a standard Linear Regression model provided by scikit-learn to replicate the UN's findings with OLS. We fit the model with the mean imputed 6-feature dataset. We fit a second, identical model with the same dataset with rows containing null values dropped instead to compare.

Model	Coefficients
UN Model	[0.318 2.422 0.033 1.164 0.635 -0.540]
Mean Imputed	[0.32755251 2.4727108 0.03109864 1.01786092 0.68353747 -0.55043552]
Dropped Nulls	[0.32755251 2.4727108 0.03109864 1.01786092 0.68353747 -0.55043552]

  Our data was very similar to the UN's findings. The slight errors can be attributed to hyperparameter tuning, but our data resembles theirs close enough for us to use OLS as a baseline. However, we found that in this situation, mean imputation had no noticeable differences in results from dropping rows with null values.

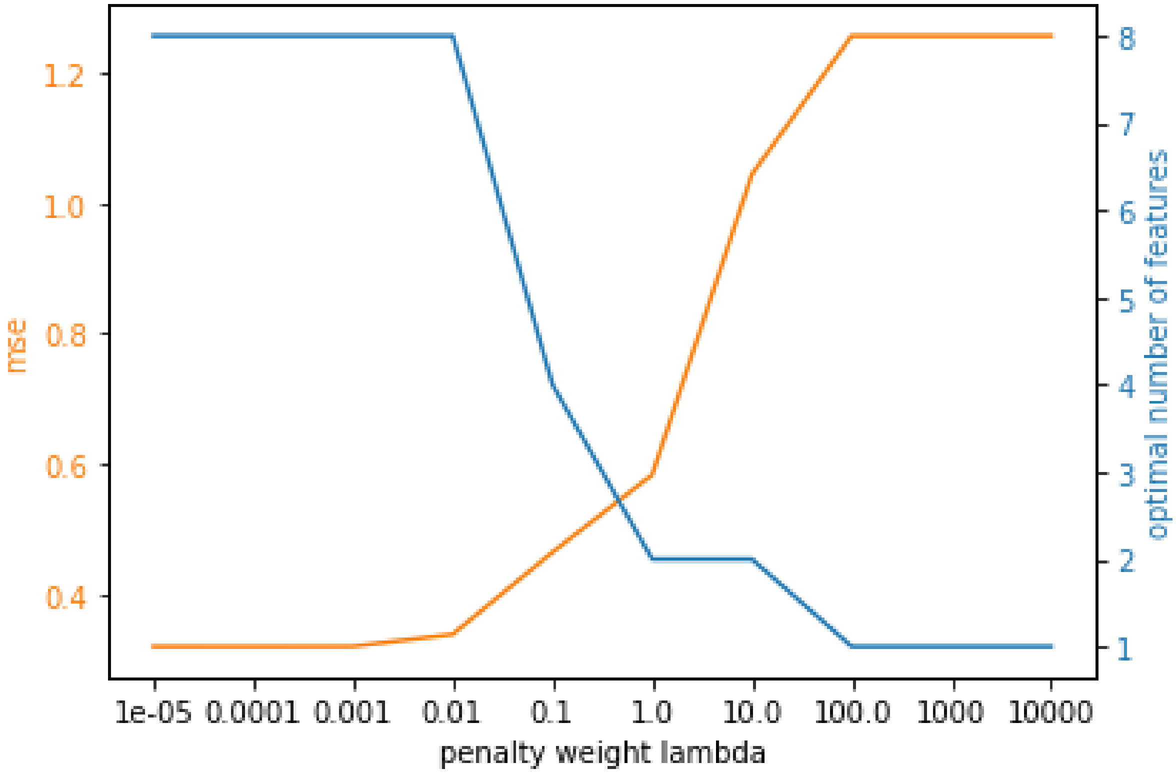
- **Checking value of extended feature set of ElasticNet**



Add'l features yield modest predictive improvement

Although there seems to be little harm in introducing an extended feature set, the UN authors do appear to be justified in highlighting only a few features. We will introduce PCA to look for a smaller feature set that reflects the apparent linear dependence we are finding here. There is also room to impute for $\tilde{5}$ more features, and we hope to add our own.

- **Forward feature selection** Prior section suggests that even with a penalty for additional features and weights, we are slightly better off in terms of prediction. Alternative models and finding meaningful additional features may be a more valuable use of time for now.



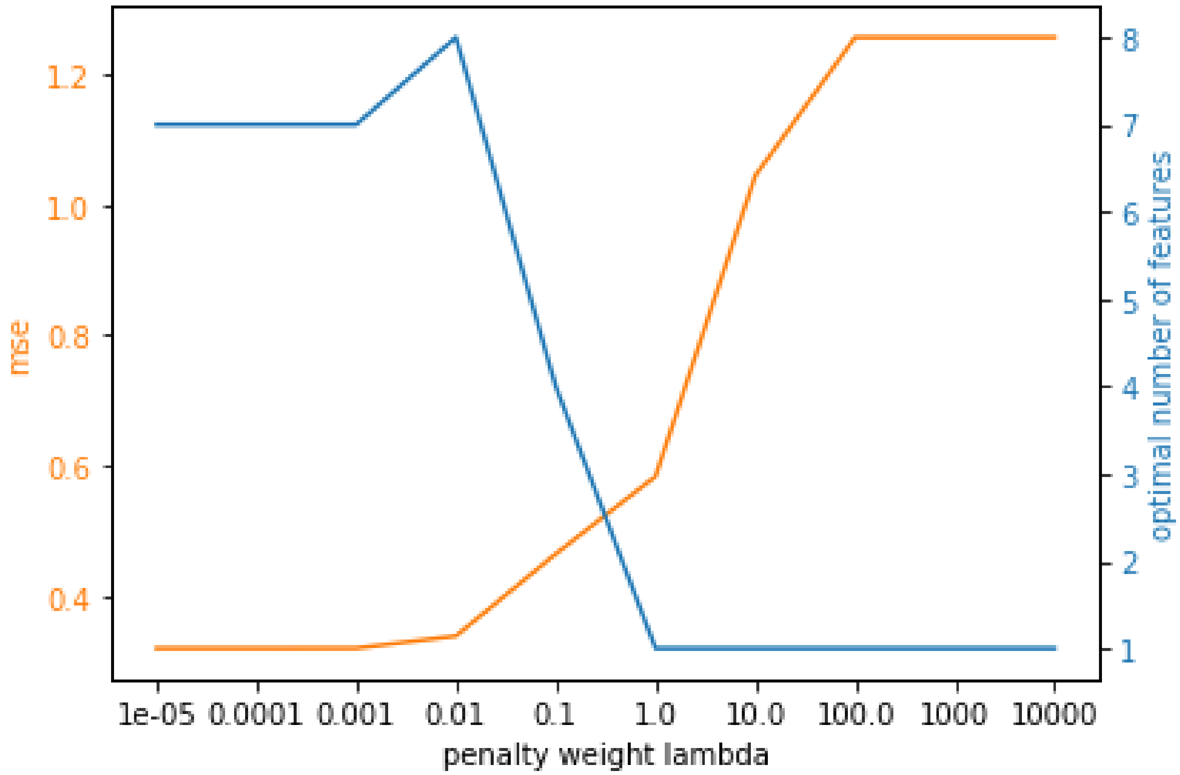
We do not start out using all 11 features–for small penalties there appear to be 8 (from an arbitary fixed order) that help us.

Note that even with this naive (depends on order in which features assessed) approach, we see slight improvement in performance: with lambda = 0.0001, we see the lowest mse achieved so far (0.3199). Moreover, note that of our 10 features only eight are included ('Negative affect', 'Delivery Quality' and 'GINI index' are tossed out).

'Affect' survey relates to respondents' reported mood, and it is perhaps unsurprising that we would only keep 'Positive affect'–by design, we would expect it to be more or less mutually exclusive with 'Negative affect'.

Inclusion of another government-effectiveness score, 'Democratic Quality' may allow respondents to explain away variation in government 'Delivery Quality'.

Perhaps the most perplexing is the exclusion of GINI coefficient (measure of income inequality); the fact that it appears last in this streamwise selection seems a likely explanation as to how it can be excluded.



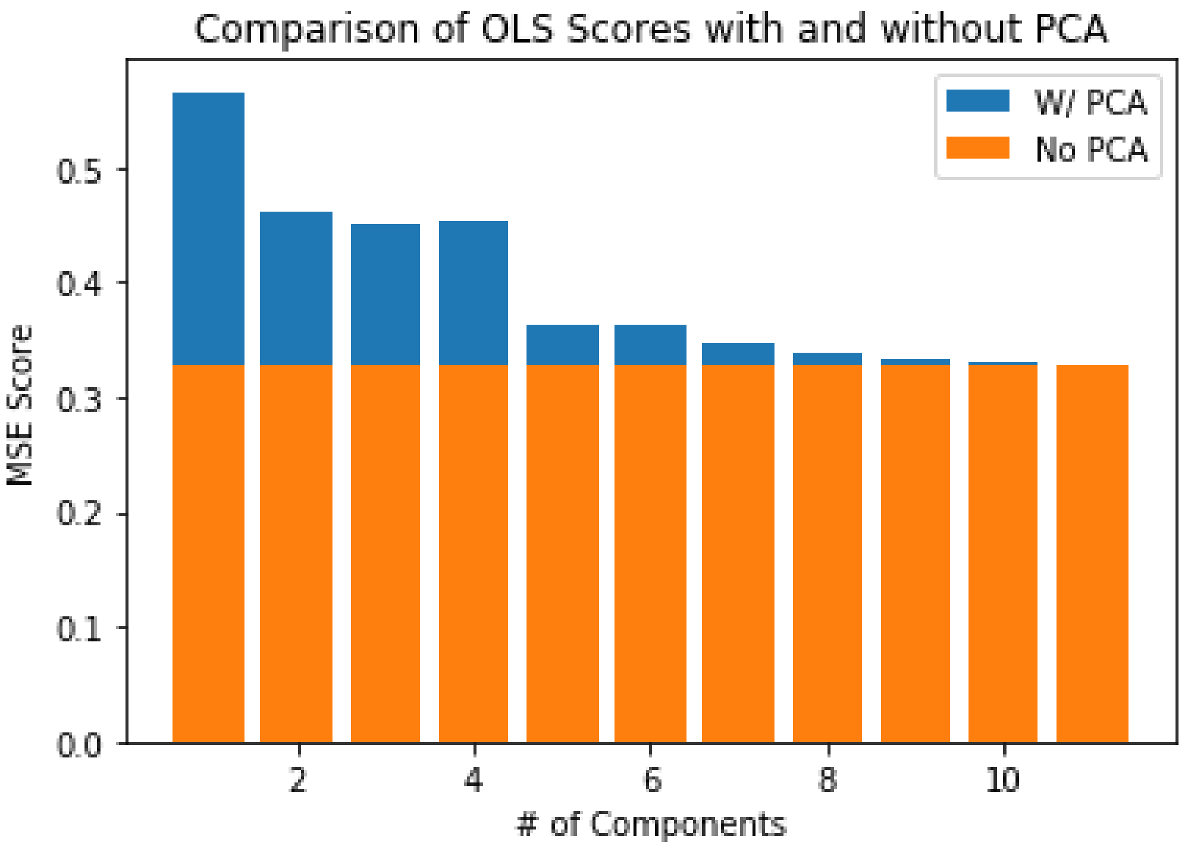
It is unsurprising that this method alights on a slightly smaller set of features (7 vs streamwise's 8 at small penalty weights) given that it can choose the 'next best' one with each iteration.

In terms of the spike to 8 features at lambda = 0.01 is difficult ot interpret. This is the only model that uses the "Freedom to Make Life Decisions" score and is otherwise indistinguishable from its predecessors' features.

We find our best MSE among all experimental methods here for lambda = 0.0001 (mse = 0.3198).

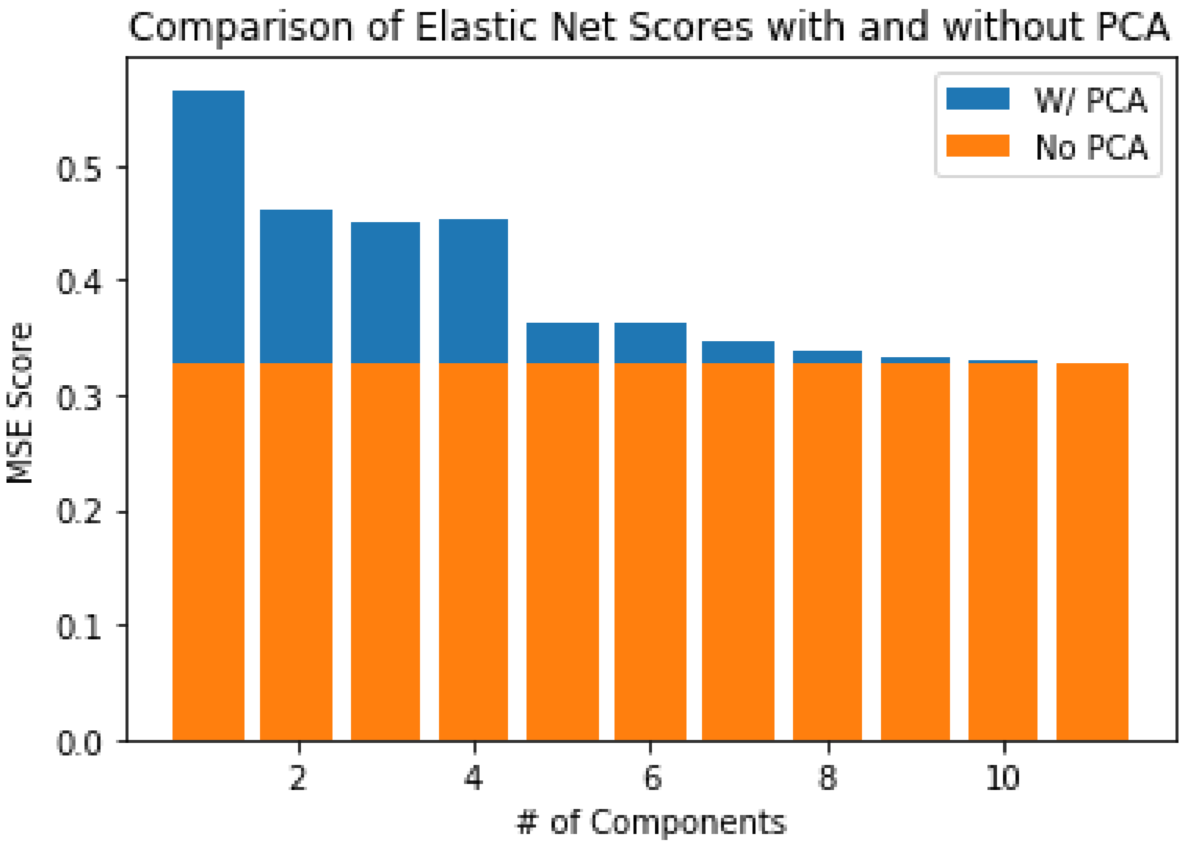
- **Regressing on principal components**
  We used Principal Component Analysis to reconstruct the dataset using 1,2,3,4,5,6,7,8,9,10, and 11 components, as we felt the results would be more meaningful if we used the feature set found by feature selection. We then calculated a percent error of the reconstructed dataset by comparing the Frobenius norm of the original to that of the reconstructed The linear regression model was identical to the one used earlier, with the reconstructed X and original y fitted to one linear regression model and the original X and y fitted to another. We calculated the Mean Squared Error using cross-validation over 4 folds for each model.

Number of Components	MSE Scores w/o PCA	MSE Scores w/ PCA
1	0.326527628038227	0.5657607077380816
2	0.326527628038227	0.46272059439185786
3	0.326527628038227	0.4510952914422206
4	0.326527628038227	0.4525330776610821
5	0.326527628038227	0.3617336520223621
6	0.326527628038227	0.3629866298896919
7	0.326527628038227	0.34710738405667363
8	0.326527628038227	0.3372110518325838
9	0.326527628038227	0.3339337487499395
10	0.326527628038227	0.3302010086790521
11	0.326527628038227	0.32652762803822705



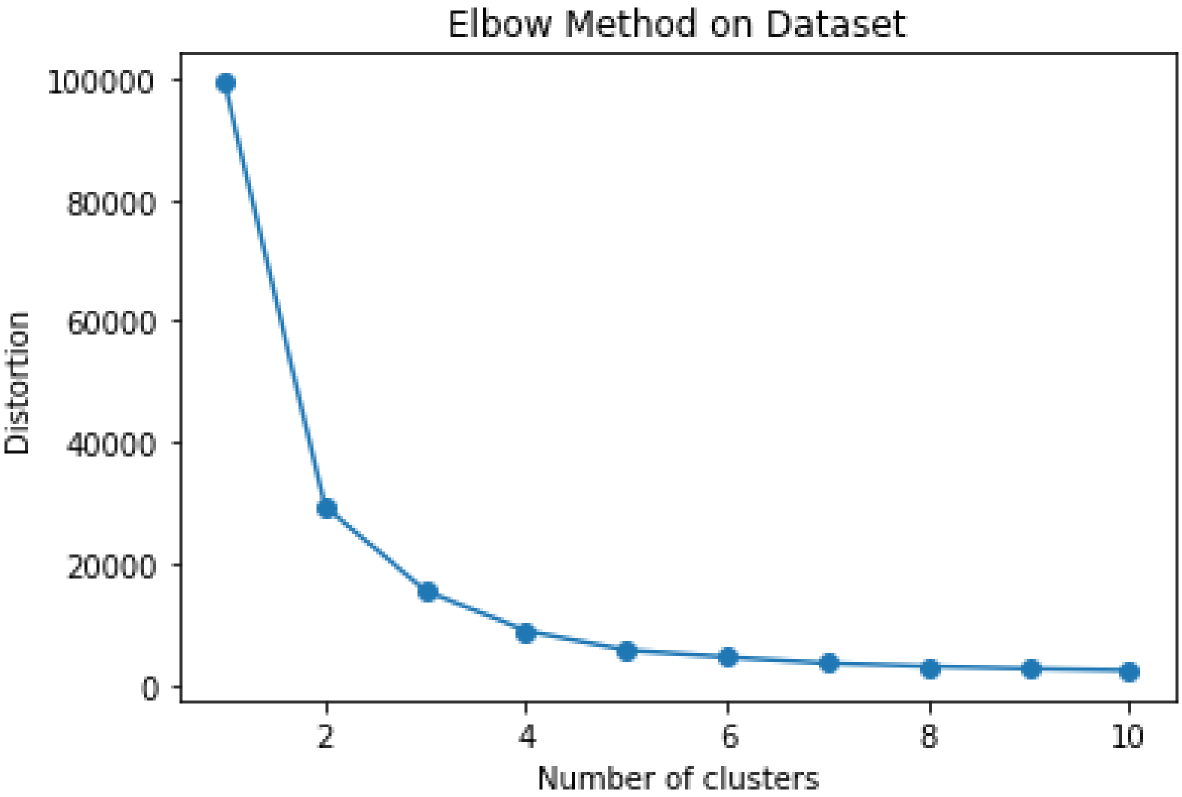
We found that the more components added, the closer the MSE resembled the original data. We also ran the same test on the Elastic Net model.

Number of Components	MSE Scores w/o PCA	MSE Scores w/ PCA
1	0.3264988924129414	0.5657581299575611
2	0.3264988924129414	0.46237086817140904
3	0.3264988924129414	0.45117978572479306
4	0.3264988924129414	0.4525432913466873
5	0.3264988924129414	0.3617314055204527
6	0.3264988924129414	0.3629847673585775
7	0.3264988924129414	0.347088017785669
8	0.3264988924129414	0.347088017785669
9	0.3264988924129414	0.33392553696129584
10	0.3264988924129414	0.33019123145788526
11	0.3264988924129414	0.3264988924129416



Both OLS and Elastic net show a sharp decrease in MSE at 5 components, with the MSE only decreasing with each added component.

- **Clustering with K-Means**
  We decided to use k-means clustering on the original 6-feature dataset to see if it yielded any interesting results. Before we started making the k-means model, we used an elbow method test to get an idea of what the optimal number of clusters is.



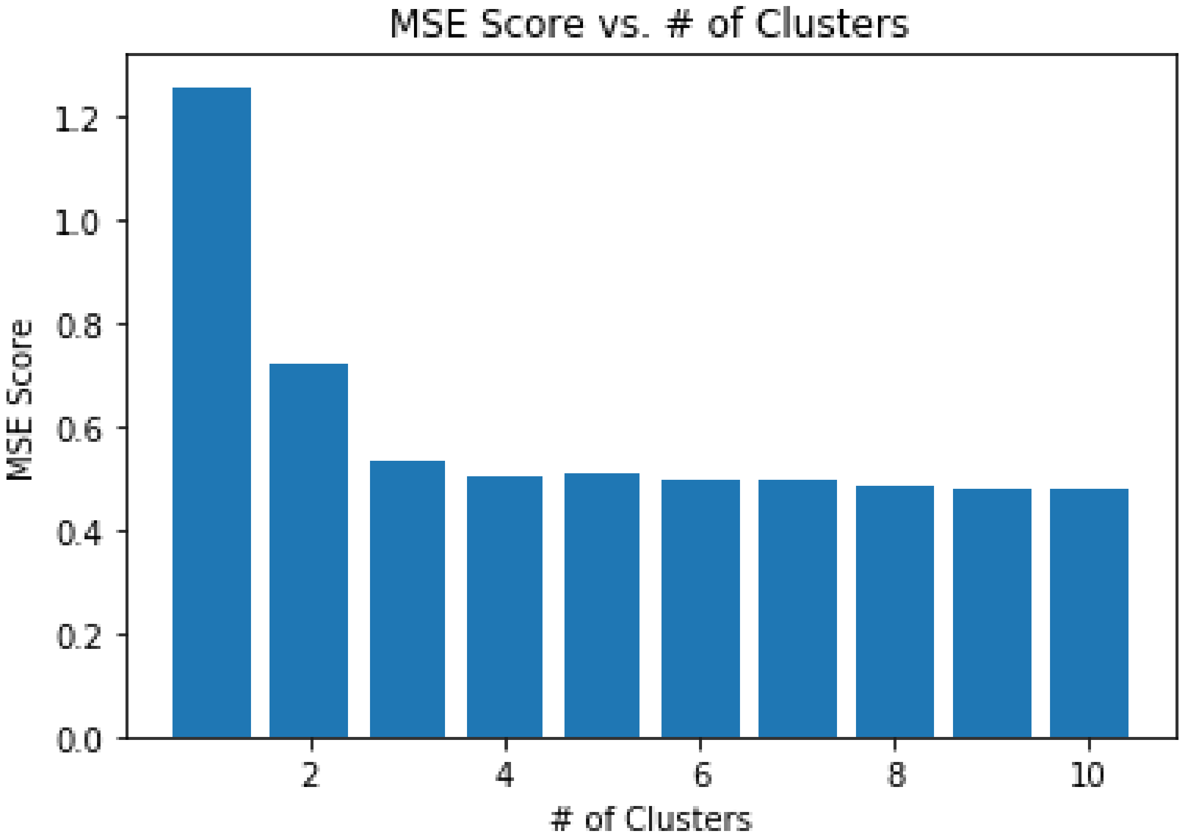
Based on the graph, the optimal number of clusters seems to be 4. We then proceeded to fit our X data into a k-means model with 4 clusters, and outputted the centroids of each cluster.

1	2	3	4	5	6
1.05141196e+01	9.06098623e-01	7.16659128e+01	8.33795974e-01	7.27984842e-02	6.03705766e-01
7.53325711e+00	6.91278395e-01	4.88898618e+01	6.58914096e-01	1.16356068e-02	8.08190699e-01
9.45736576e+00	8.31168022e-01	6.53741916e+01	7.18238632e-01	-5.05413693e-02	7.94521124e-01
8.30095915e+00	7.35562707e-01	5.75286326e+01	6.97182725e-01	1.51806485e-02	7.89639913e-01

Each centroid's values for Log GDP per capita, Social support, Healthy life expectancy at birth, Freedom to make life choices, Generosity, and Perceptions of corruption

Then, just for safe measure, we calculated the MSE of the true y value and the ys assigned by clustering, to ensure that 4 clusters made sense.

Number of Clusters	MSE Scores
1	1.2562375919499928
2	0.7220714874496833
3	0.5355094997823704
4	0.4996465416678599
5	0.5088020512866932
6	0.4950077594696217
7	0.49374963964883395
8	0.4812369423325605
9	0.4756215168994655
10	0.4763208516937037



As the data and the graphs show, the MSE seems to stagnate at and above 4 clusters, so our choice was logical.

Looking at the centroids' values allow us to see how the data is clustered for each feature. The first centroid has the highest values for Log GDP per Capita, social support, healthy life expectancy at birth, and generosity and the lowest value for perceptions of corruption. Looking at the predicted ladder score per cluster gives the first cluster an average score of 6.717, which is the highest of the clusters. The first cluster represents the first-world countries, as their values are higher than every other cluster. (other than perceptions of corruption, which is the minimum)

On the other hand, the second cluster seems to represent more underdeveloped countries. The centroid has the lowest values for Log GDP per Capita, social support, healthy life expectancy at birth, and freedom to make choices, the second lowest value for generosity, and highest value for perception of corruption. The average ladder score for this cluster was 4.215.

The third cluster looks like it represents more middle class countries. The third centroid has the second highest values for Log GDP, social support, healthy life expectancy at birth, and freedom to make choices, the lowest value for generosity, and the second highest perception of corruption. The average ladder score was 5.528. Whereas the other clusters were cut and dry in their values and corresponding ladder score, this cluster has positive and negative aspects. Even though countries in this cluster has bad values for perception of corruption and generosity, the ladder score is high because of the other factors.

The fourth cluster seems to also represent less developed countries. The fourth centroid has the third highest values for Log GDP per Capita, social support, healthy life expectancy, freedom to make life choices, and perceptions of corruption, and the second highest value for generosity. The average ladder score was 4.543.

The clusters show us some correlations between the data. Log GDP, social support, healthy life expectancy, and freedom to make life choices all appear to be directly correlated with the ladder score. However, generosity and perceptions of corruption tend not to be.

- **Experimental Summary**

Method	MSE Scores
Baseline OLS	0.3538104888363293
Elastic Net of extended dataset, alpha = 0.0001	0.3262568620537787
Streamwise Feature Selection	0.319998376955054
Stepwise Feature Selection	0.3197583996028711
OLS PCR w/ 5 components	0.3617336520223621
Elastic Net PCR w/ 5 components	0.3617314055204527
k-means Clustering w/ 4 clusters	0.4996465416678599

7 Conclusion and Discussion

SUMMARY OF FINDINGS AND MODEL COMPARISONS

We have demonstrated the feasibility of predicting Ladder Scores with relatively high accuracy, i.e. MSE around 0.32 on a 0-10 scale. Our model selection confirms the UN article's choice of six important variables and suggests that one or two more, e.g. 'Positive Affect' and 'Democratic Quality' merit inclusion.

As discussed thoroughly in the Experiments section, we **demonstrated modest improvement on the OLS baseline** by introducing a larger set of features, ElasticNet regularization and forward feature selection. In particular, each of these **incremental linear regression improvements coincide with our intuition**, and the stepwise feature selection with extended feature set and ElasticNet regularization offered our best predictive performance.

Modifying the basic linear regression framework (via PCA/PCR and via adapted K-means clustering) appears less suited to this task. PCA shows promise as a regularization method insofar as holdout MSE *nears* its minimum around 8 or 9 components instead of 11. However, unlike our standard ElasticNet (where the optimal model does not use all features), there are no cases where a reduced number of components generalized better than k = m on holdout data.

K-means was less fruitful. Although k = n of course converged to zero error, the 'elbow' occurred around 3-4 clusters. In this case, prediction error (MSE of $\tilde{0}.45$) is still dramatically higher than that achieved by more conventional regression techniques.

QUALITATIVE INTERPRETATION

It is clear that we have a set of strong predictors for Ladder score in seven features: 'Log GDP per capita' 'Social support' 'Healthy life expectancy at birth','Generosity','Perceptions of corruption','Positive affect' ,'Democratic Quality'. It is also worth noting that government 'Delivery Quality' does not feature in any of our well-tuned models. Why might this information be useful? At the very least, accepting these responses' strong correlation with Ladder means that it is possible to identify deficiencies closely related to lower life satisfaction. The United States, for instance, performs strongly along most included features but lags its higher-performing peers in Healthy Life Expectancy and Social Support, two features perennially included in our model alongside Per Capita GDP.

Domain knowledge of, say, population health management and further study of what drives high Social

Support scores seem like sensible resources to pursue in the context of considering new policies. As WHR and comparable studies nominally seek to measure human progress, a better understanding of what drives (Cantril Ladder-measured) life satisfaction may have implications for, say, attracting and retaining new talent in a given region.

LESSONS LEARNED

Our experience with this dataset confirms the value of beginning our analysis with the simplest possible approach–we saw the largest improvements to our baseline simply by introducing ElasticNet and feature selection. PCA, though elegant, does not appear to be meaningful for this very small dataset (to say nothing of totally inappropriate neural approaches). Clustering as part of our regression model appears to be too exotic.

It was surprising how little ElasticNet regularization improved our error in this case–the optimal penalty weight was routinely close to zero but error terms for lambda <0.01 were often nearly indistinguishable.

FUTURE WORK/EXTENSIONS

Our project could be extended in the future by using the years other than 2018. Doing so may yield new information from the clusters or different results from the regression techniques. We could also look at the older data to create some form of prediction for ladder scores and features for 2019. Another possible avenue we could explore is to try other regression techniques on the data, such as boosting.

This project has been an excellent hands-on exercise with linear regression techniques, including feature selection and tuning. Had we learned about AutoML earlier, we might have incorporated it as a benchmark near the high end of feasible predictive performance.

Finally, more sophisticated interpolation techniques (e.g. column regression) may be warranted given that all of our features have at least some missing data.

References

[1] John Helliwell Haifang Huang and Shun Wang. *Statistical Appendix 1 for Chapter 2 of World Happiness Report 2019*. Sustainable Development Solutions Network, 2019. https://s3.amazonaws.com/happiness-report/2019/WHR19_Ch2A_Appendix1.pdf.

[2] John Helliwell Haifang Huang and Shun Wang. *Statistical Appendix 2 for Chapter 2 of World Happiness Report 2019*. Sustainable Development Solutions Network, 2019. https://s3.amazonaws.com/happiness-report/2019/WHR19_Ch2A_Appendix2.pdf.

[3] John Helliwell Richard Layard and Jeffrey Sachs. *World Happiness Report 2019*. Sustainable Development Solutions Network, 2019. https://worldhappiness.report/ed/2019/changing-world-happiness/.