World Happiness: What Affects It?

Stats 415: Data Mining (Professor: Liza Levina)

Introduction

Our primary dataset from which we pulled our response variables is the World Happiness Report, a landmark survey on the state of global happiness. Global happiness is becoming an important factor in policy-making for governments, organizations, and civil society as they begin to use happiness indicators to inform decisions. The happiness score and ranking data that form the basis of this project source from a study of the Gallup World Poll.

It's a survey-based dataset, with the "Happiness Score" column being the response variable of interest, where six columns, economic production, social support, life expectancy, freedom, absence of corruption, and generosity, each affect that ultimate score. There is also a Dystopia Residual, which is a comparison to a hypothetical "worst country".

In our analysis, however, we decided not to use any of the other columns in the World Happiness Report except for the score and rank, as these have all been scaled already, provide little value in terms of interpretation, and are frequently used in prediction. Our main interest is determining ourselves which factors affect world happiness scores by pulling predictors from *other* separate datasets that correspond to countries.

We focused on the year 2016, which had data for 157 countries. These other datasets that we've included are climate data which contain the temperatures of each country for the years 2001-2013, peace agreements data covering data from 1990 to 2016, inflation data primarily from 2016, and schooling data from 2016. The question we posed to answer was whether we could predict happiness by fitting models on information from otherwise unrelated data sets. In addition, a follow up question was whether we could identify the significant predictors that contribute to a country's happiness score.

Methods and Data

With so many datasets from different sources, we had to spend a bit of time cleaning the data and making sure that we could merge them into a final dataset with unique countries and their corresponding predictors.

I. Climate data

Since the climate data did not include any years past 2013, we had two options - we could use the averages of 2001-2013 climate temperatures, or, since this is climate we're looking at over time, we could use time series analysis to extrapolate the temperatures for 2016. We decided to try this latter method, since our results did not seem too far off from what we got from averaging 2001-2013. We did observe that the temperatures were higher, but that was determined to be reasonable if one considers the probable effects of climate change. Time series analysis was conducted in R using the *forecast* library and going country by country to extrapolate 2016 from 2001-2013 data. The variables we decided to create based on

this data were: Average temperature, median temperature, high, low, and range for the 2016 year. We also acknowledge that outliers are certainly a major factor in high, low and range values, but decided to leave them in for this analysis.
Time series were not a focus of the class. As such, there may have been errors when attempting to predict an entirely new set of data. We trust that the extrapolation was done as accurately as possible to the best of our abilities and that the new 2016 climate data can be used for further analysis.

II. Peace Agreements dataset by the University of Edinburgh

The peace agreements data consisted of ~1500 peace agreements from 1990 to the beginning of 2016. This dataset only contained *resolved* conflicts(armed violence with +25 deaths) that were officiated by both sides. Additionally, former Yugoslavia consists of Serbia, Kosovo, Slovenia, Croatia, Bosnia, and Herzegovina, and we added that entry's count to constituent countries.
We utilized Power Query to separate data into one row per country (per conflict), then used pivot table to count the quantity of peace each country was involved in, ignoring the details of the given agreement. Possibly overloaded the individual countries that made up the Yugoslavia Republic, but it seemed necessary as the countries were still involved regardless of title. Also, a given country's number of conflicts aren't necessarily reflected by number of brokered peace agreements, so we had to be careful of making false assumptions about the data's implications. High number of peace agreements could have covered almost all the conflicts in a given country, or it could just be a symptom of many more ongoing, unresolved conflicts.

III. Inflation of Consumer Prices data

The inflation of consumer prices data consisted of the inflation in consumer prices of 264 countries/territories for each year from 1960 - 2017. The inflation in consumer prices is described as the change in cost to the average consumer of fixed goods and services. Often higher inflation rates are detrimental to an economy with increased interest rates and unemployment. We believed that this economic factor may contribute to a country's happiness.
We focused on the years of 2010 to 2016 for each country as they fall within our decade, representing the most recent information. Missing data within these ranges was extrapolated utilizing the average rate of change of the inflation rate of the previous 10 years. We used the 2016 inflation rates as well as the average inflation rate from 2010 to 2016 for each country for our project's data set.

IV. Schooling data

The schooling data set spanned the years of 2013 - 2017 with schooling and population data organized by ~180 countries. We focused on the 2016 data for 188 countries. Each country had data for expected years of schooling and mean years of schooling, but also had data for other population demographics such as GNI (gross national income) and life expectancy. This data set did not require any additional coding in order to format the data or fill in missing data. The data that we pulled to our project's data set were the

expected years of schooling, mean years of schooling, GNI per capita, and the country's human development index.

V. Condensing Data

After combining all the datasets we had the following columns: Country, happiness rank, happiness score, lower CI and upper CI for the scores, inflation, average inflation for that decade, HDI (Human Development Index), life expectancy, expected years of school, mean years of school, GNI (Gross National Income), GNI per capita rank, number of peace agreements, and finally, the average/median/high/low/range temperatures from the climate dataset.

We decided against including country names and happiness rank in the model estimation; while useful for identification and organizing data, they had no place in the regression analysis portion.

Additionally, we realized that the confidence intervals for the happiness score also had no place in this analysis. Life expectancy and HDI seemed closely related enough upon reviewing the data values, that we could likely drop one of the predictors and still have the data trend we needed. The high and low temperatures were closely related to temperature range, and deemed unnecessary for model fitting. We decided to focus on the range and mean temperatures. In addition, we decided to stick with gross national income and remove gross national income per capita rank.

While the inclusion of the missing variables may have aided the model estimation, the group believed that the original dataset contained too many predictors. The variables left were:

- 2016 Inflation
- Average Inflation
- HDI (Human Development Index)
- Expected Years in School
- Mean Years in School
- GNI (Gross National Income) per Capita
- Peace Agreements (count)
- Average Temperature
- Range Temperature

Data exploration

For our analysis, we first wanted to explore the data to see what correlations we could find that would prove useful for regression and overall prediction of happiness score.

As we can see from Figure 1, HDI (human development index), expected years in school, and expected mean years in school all appear to have a positive correlation with happiness score. Furthermore, from Figure 2 we also see correlation between happiness score and GNI (gross national income) per Capita as well as some correlation in both average and range temperatures. We will explore these further in our regression analysis.

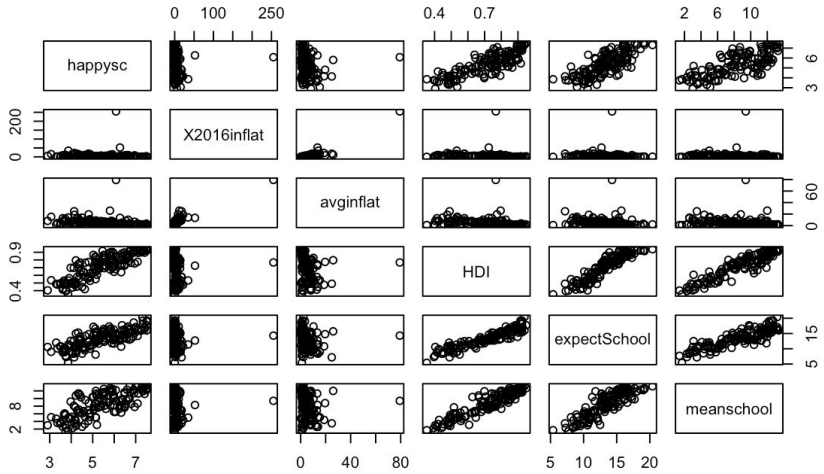

Figure 1. Looking at correlations between happiness score and first five predictors.

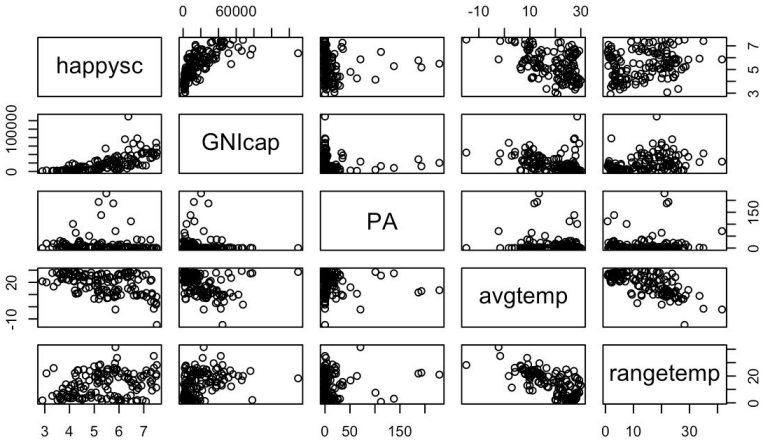

Figure 2. Looking at correlations between happiness score and last four predictors.

There are likely variables that are insignificant in our data, so the next step is to go through regression analysis methods and reduction techniques to try to build the predictive model for happiness score with the lowest test error.

Regression analysis methods and results

I. Linear Regression

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.000e-16  5.050e-02   0.000  1.00000
X2016inflat   7.123e-02  1.466e-01   0.486  0.62835
avginflat    -5.103e-02  1.487e-01  -0.343  0.73234
HDI           1.256e+00  2.489e-01   5.047 2.43e-06 ***
expectSchool -3.319e-01  1.567e-01  -2.118  0.03701 *
meanschool   -2.063e-01  1.620e-01  -1.273  0.20641
GNIcap        1.392e-01  8.368e-02   1.663  0.09983 .
PA           -4.084e-02  5.256e-02  -0.777  0.43924
avgtemp      -2.642e-01  9.323e-02  -2.834  0.00572 **
rangetemp    -3.099e-01  8.625e-02  -3.593  0.00054 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4974 on 87 degrees of freedom
Multiple R-squared:  0.7758,    Adjusted R-squared:  0.7526
F-statistic: 33.45 on 9 and 87 DF,  p-value: < 2.2e-16
```

Figure 3. Results from linear regression on all predictors.

The first thing we’re interested in observing is what variables ended up being marked as significant. Linear regression marked HDI, expected number of school years, average temperature, and range temperature. Train error was 1.07, while the test error was 1.27. The next observation we must take into account is the adjusted R-squared value so we know how accurate the fit was on the data - with an R-squared value of 0.7526, there is a high likelihood that there is a better model besides for linear regression that could more accurately predict happiness score.

II. Lasso Regression

When we ran lasso regression, our train error was 0.240 and our test error was 0.323. Comparing this to linear regression, this is a significant improvement. It also suggests that variable selection is a method that we should continue with - in this particular regression method, lasso removed the following variables: average inflation, expected years in school, and mean years in school. We look to best subset selection to see if there are other combinations of removing variables that yield lower errors.

III. Best Subset Selection

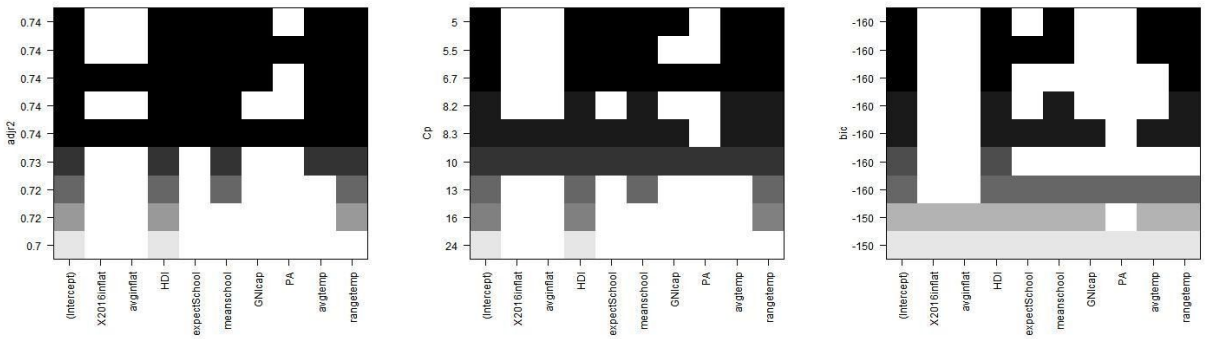

Figure 4. A visual comparison of the predictors selected by adjusted R^2, Mallow's CP, and BIC in that order.

```
## (Intercept)         HDI  meanschool     avgtemp   rangetemp
##  1.22275505  8.69507946 -0.13024699 -0.02603714 -0.03055460
```

Figure 5. The coefficients of a model fitted using BIC (recommended 4 predictors). **Training MSE** was: 0.3709279. **Test MSE** was: 0.3063137

Another attempt to fit a more accurate model involved selecting the best subset or combination of predictors. When running best subset selections with adjusted R^2, Mallow's Cp, and BIC, we decided to utilize BIC on the basis of Occam's Razor; BIC returned the simplest model with the least number of predictors and apparently fit a comparably accurate model.

It's worth noting that the three methods we utilized for best subset selection ended up sharing a number of predictors in common; they are as follows: HDI, the mean years of schooling, average temperature, and temperature range (coincidentally, this is the the model given by BIC).

IV. KNN Regression

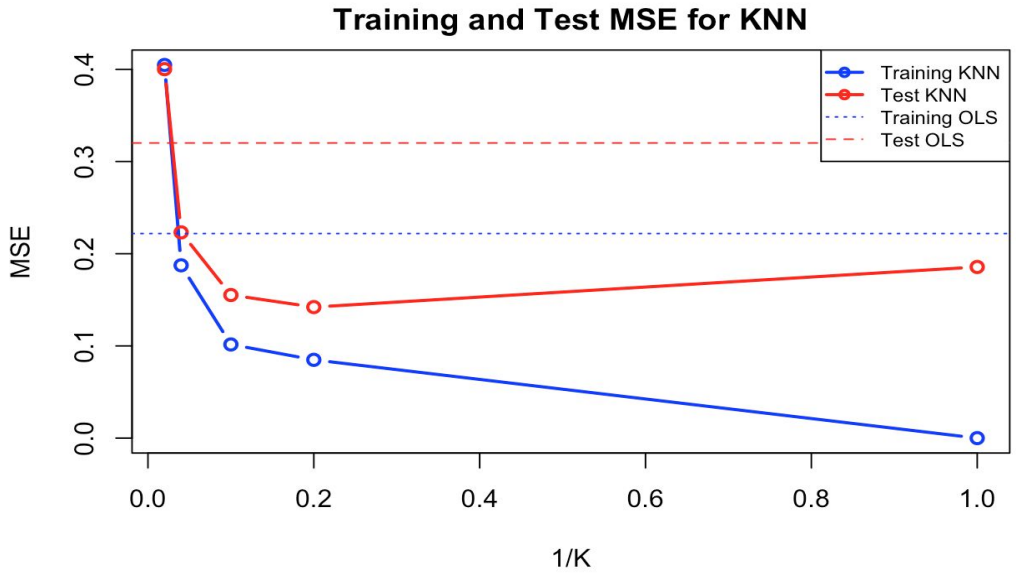

Figure 6. Choosing the optimal K.

When running K-Nearest-Neighbors, it was clear that the K value that yielded the lowest test error was K = 5. This indicates that a smaller value of K, which corresponds with a more flexible model, is more suitable for our data. More flexible models indicate that it fits to the noise better, and likely is non-parametric. Since the test error for KNN happened to be the lowest of all previous models, we can then make an assumption that predicting happiness score is more reliable with a non-linear model.

V. Best Method, Removing Variables

Results from all the models we created:

REGRESSION	TRAIN ERROR	TEST ERROR
Linear	1.07	1.27
Lasso	0.240	0.323
Best Subset Selection (BIC)	0.371	0.306
KNN	0.0849	0.142

As we can see, KNN performed the best out of the models that we tested out for regression - it performed significantly better than any of the other methods, and we can now try to reduce the test error even more by using the subset results from BIC and lasso and remove variables, re-running KNN to see if we can obtain lower test errors.

We now want to remove the variables suggested by Lasso and best subset selection (adjusted R-squared, CIP, and BIC) and rerun KNN to see if we can get lower test errors. The following will be removed: 2016 inflation, average inflation, expected years in school, and peace agreements. We are left with: HDI, mean years in school, GNI per Capita, average temperature and range temperature.

Upon running KNN again, we now end up with:
Train error: 0.051
Test error: 0.105

Thus, we've successfully reduced the error by removing variables suggested by reduction techniques. What this suggests about our data is that there are easily predictors that only created more noise in the data, and with further analysis we could more intricately pick and choose which predictors result in the most accurate predictive model for happiness score.

VI. Extension: Predicting Rank

Thus far our response variable has been a country's happiness score, however we wanted to see if our model was also accurate in predicting the overall happiness ranking of each country. Following our optimal model, a KNN regression was run with K = 5 on our training set and applied to our entire dataset.

Countries	Real_Happiness_Rank	Predicted_Happiness_Rank	Real_Happiness_Score	Predicted_Happiness_Score
Denmark	1	1	1.809249405	1.677619532
Switzerlanc	2	3	1.794737678	1.535713604
Iceland	3	5	1.78790863	1.532244425
Norway	4	4	1.785347737	1.535713604
Finland	5	6	1.712789102	1.527123257
Canada	6	12	1.705106423	1.406362799

Countries	Real_Happiness_Rank	Predicted_Happiness_Rank	Real_Happiness_Score	Predicted_Happiness_Score
Madagasca	130	122	-1.461010947	-1.254497246
Tanzania	131	128	-1.485766246	-1.360389795
Liberia	132	136	-1.523326009	-1.445467272
Guinea	133	135	-1.536130474	-1.407801904
Rwanda	134	138	-1.614664526	-1.465125951
Benin	135	132	-1.641127087	-1.394420786

Figure 7. Snapshots of our dataset with the predicted happiness scores and the predicted happiness rankings. *A reminder that the happiness scores were scaled.

The top ranking countries and bottom ranking countries were very similar between the predicted and and real rankings. On average, the predicted rank differed from the real ranking by 7.20863 countries. Interpreting this for the 139 countries, there was overall a 5.18% variation in predicting a country's rank utilized our best modeling method.

Conclusion & Pitfalls

We were able to develop a model that predicted a country's happiness score with a test error of ~10%. In addition our model was able to predict a country's happiness ranking to within ~5%. In addition, we were able to determine that the predictors having the most influence on a country's happiness score were HDI, mean years in school, GNI per Capita, average temperature, and range temperature. While our model did well in prediction, there are many ways in which the modeling can be improved.

When considering data about individual countries, it's important to think about the context behind specific data and why certain countries have a higher happiness score than others. Happiness score was chosen by the World Happiness Report, so it's a derived value that we have little context about. In terms of the data we gathered to try to predict this score, we never looked at our data in a spatial context, i.e. different regions of the world may appear "happier" because of certain cultural norms or political reasons. Japan, for example, was ranked 50, which is a country that would normally be assumed to have a much higher ranking based on their economy, life expectancy, high HDI and low inflation, but perhaps other social issues could be playing a role, and these were not captured by the data that we used to predict happiness score. Also, happiness is an ambiguous response variable that is ultimately subjective, making it difficult to know for sure whether a model is accurate in predicting a country's happiness, as well as the understanding that happiness is something that builds up over time, and we tended to just look at data from the year 2016. Perhaps a more extended analysis would include using time series analysis for multiple predictors and trying to track trends over time and seeing if that improves the predictions for happiness score. Retrospectively, more flexible, non-parametric models tended to be beneficial for the data in this project, and with more consideration for other factors that affect a country's overall happiness, there could potentially be more valuable results.