Nolan Chai
Sanghyun Hwang
Shirin Saifuddin
Cogs 109 Spring 2021: Final Project

2020 World Happiness Report: An Exploratory Data Analysis

Introduction

The World Happiness Report is a survey that ranks each country with the goal of assessing global happiness, or contentment based on a variety of factors. Life satisfaction is derived from a combination of how content the country's citizens are within their social, urban, and natural environments. Using the Gallup World Poll, researchers asked respondents to rate their current lives from 1 to 10 with respect to what they believe a conceivable life to be. Essentially, the World Happiness Score is explained by the following six factors: Economic production, social support, life expectancy, freedom, generosity, and absence of corruption. By measuring which variables have the most significant impact on overall happiness, researchers can contribute to policy-making by determining the most impactful way a country can advance and better the lives of their citizens. The data we will use is from the World Happiness Report 2020, and the number of observations (n) is 153, and predictors (p) is 6. The hypothesis that we will test is predicting a country's happiness score based on the statistical significance of the features reported by each country. We hypothesize that no one factor significantly drives the magnitude of a country's reported happiness scores. By understanding which variables contribute the most to overall reported happiness, we can make predictions for what a country may report in 2021. Essentially, what we have is an inference-based model driving a prediction hypothesis.

Methods/CV

The type of cross-validation we used to randomize the split of our observations into training and testing was k-fold cross validation. We split our data using k=5 and k=10 folds, used

k-1 folds as a training set and the remainder of each split as a testing, repeated this k times using a loop, ultimately to choose the model with the lowest MSE. Advantages of using k-fold cross-validation is a reduced bias and variance, and because we are able to eventually use all our data to build a model, we will eventually yield a more predictive, accurate model. Another advantage of K-fold is because the training data is separate from the testing data, we get to prevent overfitting since each subset is a shuffled, randomized set of data points. In K-fold, the model parameters we generate at the end of each fold will eventually be averaged to build the final model.

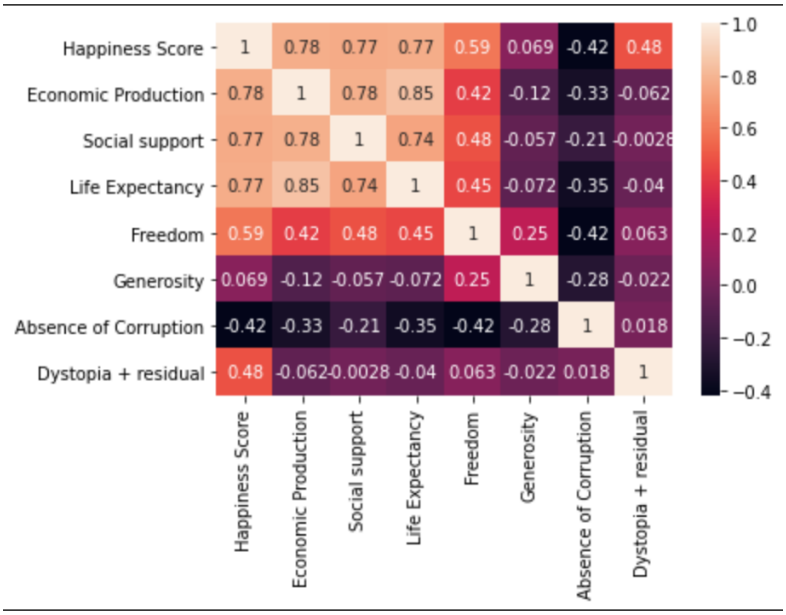
Methods

The data analysis approach that we decided to opt for was using multiple linear regression as well as backward stepwise subset selection. The reasoning for why we chose multiple linear regression was to be able to observe the relationships that the predictors (Economic Production, Social Support, Life Expectancy, Freedom, Generosity, Absence of Corruption) have on happiness scores of nations. Through these observations, we will be able to make an inference on the happiness scores of nations in 2021, by analyzing the relationships that each individual predictor shares to happiness levels, and make predictions to determine which predictor has the most significant effect on happiness scores. Multiple linear regression is appropriate for our dataset as we are able to run OLS regression with our predictors, and observe whether or not the predictors are significant in relation to happiness scores. This methodology allows us to make predictions with the data that is present. The reasoning for why we chose backward stepwise subset selection was to determine which model will be most applicable by observing the lowest Mean Standard Error of training vs test data using k-fold cross validation.

With this new model, we incorporated it into multiple linear regression to perfectly align our happiness score with our predicted happiness score.

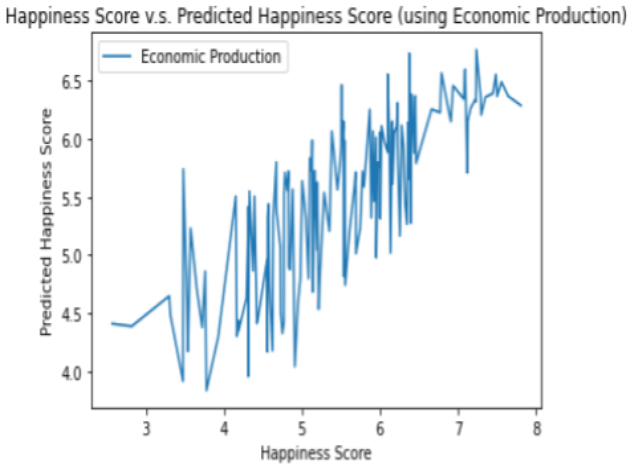
Results - model selection

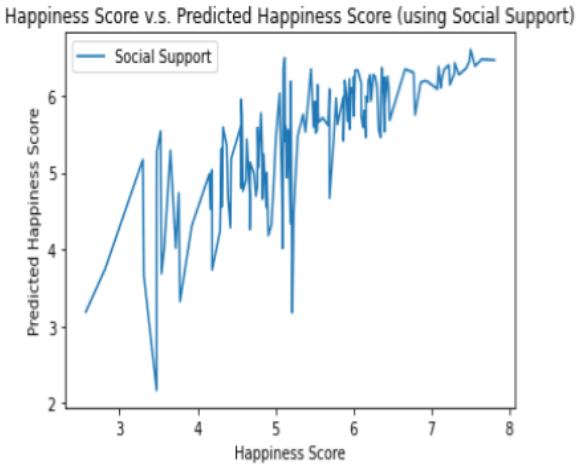
To begin, we found the Pearson's R between each variable (including Happiness Score) and generated a correlation matrix to give us an idea of the relationships between our beta variables.

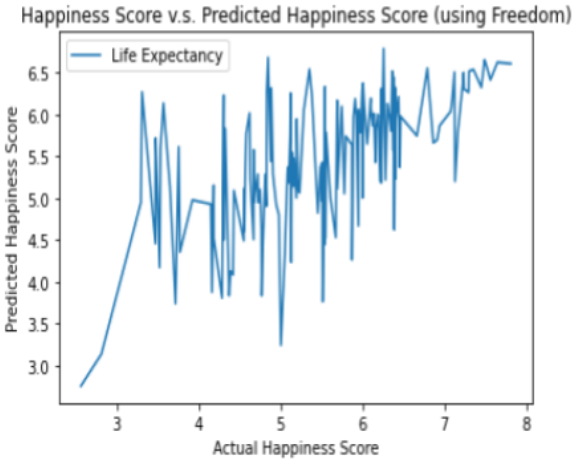


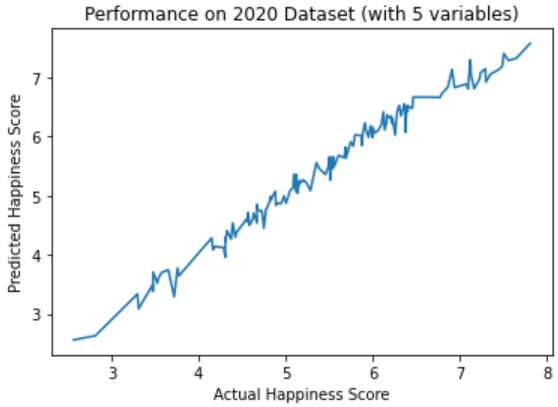
Keeping this in mind, we started off the full model with all 7 beta variables: 'Economic Production', 'Social support', 'Life Expectancy', 'Freedom', 'Absence of Corruption', 'Dystopia + residual', and 'Generosity.' Using backwards stepwise regression, we simplified the model as much as we could and compared the two models with the lowest K-Fold Cross Validation MSEs. During each step, we removed a single variable from the model and compared the performance of all model combinations of the remaining variables by calculating the MSE between training and test sets using K-fold cross validation. The model with the lowest MSE was selected for the next step, and we continued until there was one single variable left.

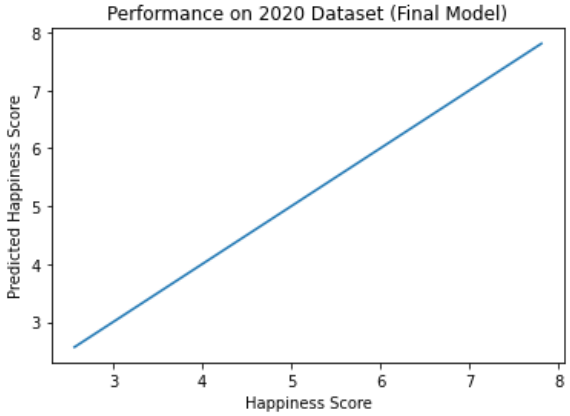
Looking at the most statistically significant 6 variables in our dataset, we were able to see the relationships and performance of each relative variable.









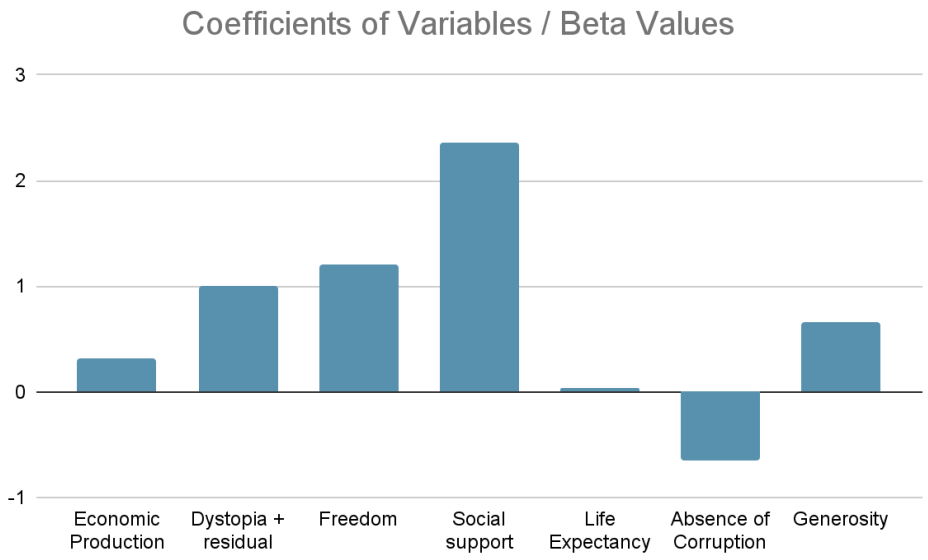


In our first step of iterating through our models, we removed generosity, followed by absence of corruption, life expectancy, social support, freedom, dystopia + residual, until we were finally left with the single variable: economic production, selected based on the MSE performance on K-Fold Cross Validation.

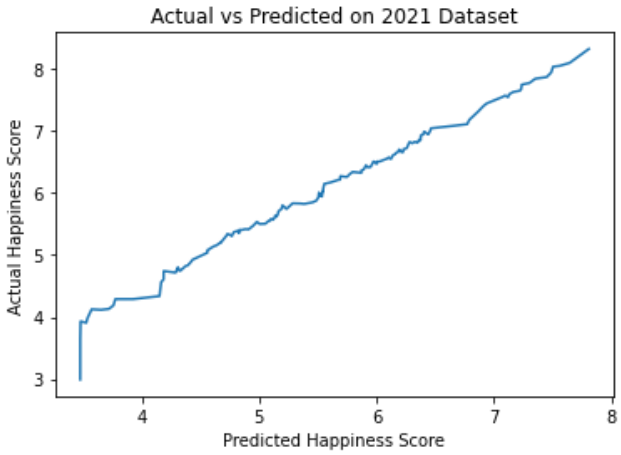
Although our final model had the lowest MSE on our 2020 dataset, it is important to note that it is also the model with the highest complexity of the 6 we generated, meaning there is a higher chance of overfitting.

Results - model estimation

Our final parameter estimates (coefficients) for our best fitted model were 0.30993547, 1.00000283, 1.19860363, 2.36193766, 0.0360018, -0.64561715, and 0.66134567, in respect to the variables Economic Production, Dystopia + residual, Freedom, Social support, Life Expectancy, Absence of Corruption, and Generosity.



Using MSE (mean squared error) to evaluate the accuracy of our model's performance on the test dataset using K-Fold Cross Validation returned an average MSE of 2.9943566272131014e-05. When tested on the 2021 dataset, the model's predicted v.s actual returned an average MSE of 0.2416528121416543. Considering the fact that the model's performance on the 2021 dataset was orders of magnitude greater than the 2020 dataset, our final model seems to be overfitted.



Discussion

After applying the research methods that were mentioned above on our data, we were able to discover that Economic Production (0.78), Social Support (0.77), and Life Expectancy (0.77) were the three most significant predictors towards Happiness Score. Economic Production most likely had the most significant effect on a nation's happiness score not only due to the fact that countries with more stable economies attract higher levels of satisfaction, but also the fact that countries with better economies are able to improve the standard of life for citizens. From our data from 2020, we were able to see that the higher ranking nations, in terms of happiness scores, tended to always have high economic production, social support, and life expectancy. From this, we concluded that this occurred because high economic production leads to better social support from the government, which gives longer life expectancies for citizens. Likewise, we can predict that nations in 2021 (the next year) with higher happiness scores will share the same traits (higher pearson's r for economic production, social support, and life expectancy) as those from 2020. However, it can also be noted that happiness scores in general from 2021 could be much lower than those from 2020 as COVID-19 crippled many economies, including the USA.

For researchers who are interested in this topic, we think that it is critical to understand that happiness scores change from year to year; some countries that were placed in the top 10 from 2019 are no longer in the top 10 for 2020. There are many outlying factors that can cause a shift in happiness scores that were not discussed in our research. For example, in our research, our model looked as if it overfitted as our test MSE for our predicted 2021 happiness scores were high (0.24162812416), while our test MSE for 2020 happiness scores were low (2.994356627213e-05). However, there were many external factors involved that could have affected our model's performance, such as the pandemic and election. We think that it would be interesting if future researchers reverse the question, and try to discover which features cause a decay in happiness scores, in order to maybe solve possible real-world issues that may arise in the future.