Solving for Happiness

Isaac Wecht
Brown University
https://github.com/iwecht1/1030project

Introduction

Measuring the progress of a society can be a daunting challenge. Typical measurements vary from economic forecasts, such as GDP, GDP-per-capita, and CO2 emissions, to social trends, like population growth and life expectancy. In this paper I will argue that the true measure of a country's progress can only be assessed by the happiness of its citizens. The target variable of my machine learning model is the Happiness Index of 47 countries from 2005 to 2020. I will use the five aforementioned economic and social features to generate my model's predictions for the Happiness Index. Since I will be evaluating 47 countries over 16 years, I have a total of 752 data points within my dataset. My project is a regression based problem, as I will be predicting the value of the Happiness Index, a continuous variable. In an age of political and environmental instability, predicting society's future happiness standards is an important endeavor, ensuring that we are truly progressing towards a better future.

My target variable, the Happiness Index, was created by the UN and Gallup, a world polling company, for the purposes of publishing the World Happiness Report. The index was constructed by conducting social science surveys on a subsection of the population from each of the 47 countries. The survey consisted of several holistic questions, such as "how would you rate your work-life balance," which participants answered on a scale of 0-10. The scores of each country's participants were aggregated and averaged to derive the country's total Happiness Index for a given year. Additionally, the feature data was collected by each countries' governmental institutions for standard census and state assessment projects. A detailed description of each feature in my dataset can be found in the below table:

Feature Name	Feature Type	Description	Source
Country	Categorical	Name of Country (47 unique values)	Kaggle
GDP	Continuous	Gross Domestic Product [$]	World Bank
GDP-per-capita	Continuous	Share of Gross Domestic Product per citizen (GDP / population) [$]	World Bank
CO2 Emissions	Continuous	Amount of CO2 emitted during a calendar year [Tons]	Kaggle
Population	Continuous	Count of people living in a given country [#]	Kaggle
Life Expectancy	Continuous	Average lifespan of an individual living in a given country [Years]	Kaggle

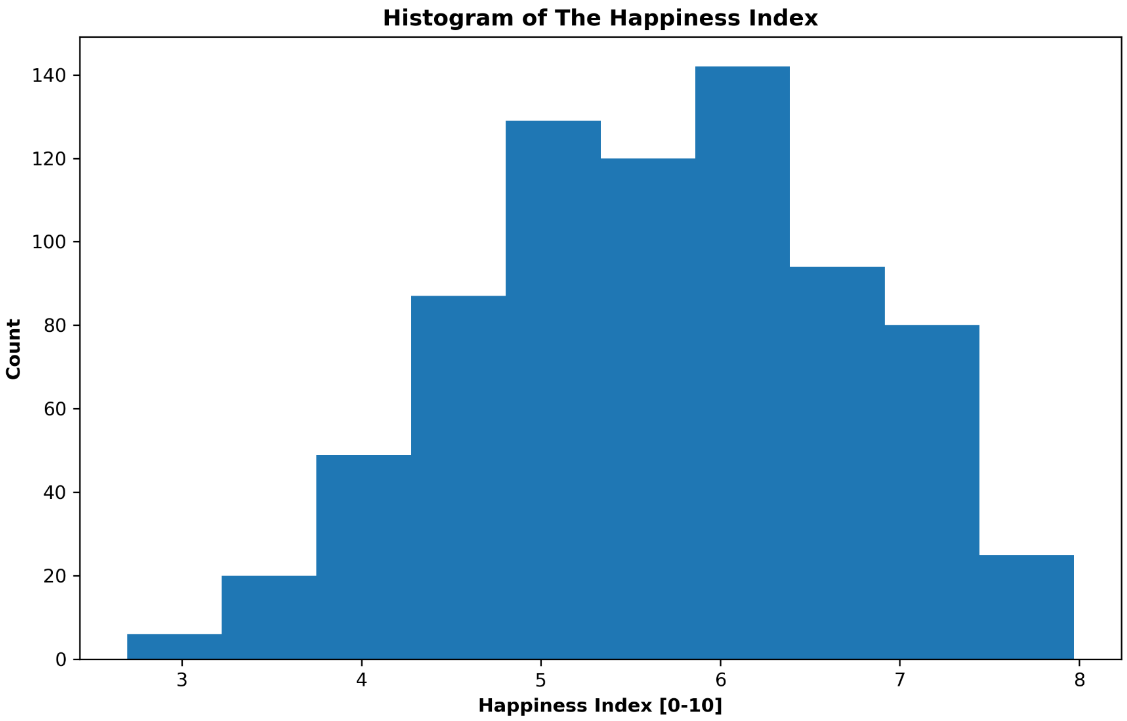
The data being used for my project has been analyzed in numerous other projects and publications, which can help refine the evaluation of my model. For instance, Filip Projcheski, a data scientist from a tech start-up, designed a “machine learning model that predicted the happiness index based on the GDP " of various countries (Projcheski). Projcheski found that his model’s “predicted index is quite different from the actual index in poor countries…where freedom of speech, women’s and minorities’ rights are suppressed " (Projcheski). This conclusion reveals an interesting bias in Projcheski’s model, which recorded a RMSE of 1.024. I believe I can improve on this error rate by including more features, such as population size and life expectancy, in addition to the GDP feature Projcheski analyzed.

Another project conducted by Chan Min Yi, a university student, utilized a random forest model to predict which features had the greatest influence on the Happiness Index of countries. Yi conducted this research by examining the correlation between question topics asked on the social science survey and the extent to which that question influenced the participant’s overall Happiness Index. Interestingly, Yi found questions pertaining to the participant’s age had little effect on the Happiness Index, indicating that happiness may not vary over age groups. Yi utilized the RMSE metric, finding an error rate of 1.356, to evaluate the model, which appeared “to be slightly more negative than positive in predicting the happiness levels” (Yi). I will need to consider Yi’s findings when implementing my model, ensuring that countries with lower Happiness Indexes do not inappropriately negatively skew the model’s results.

Exploratory Data Analysis

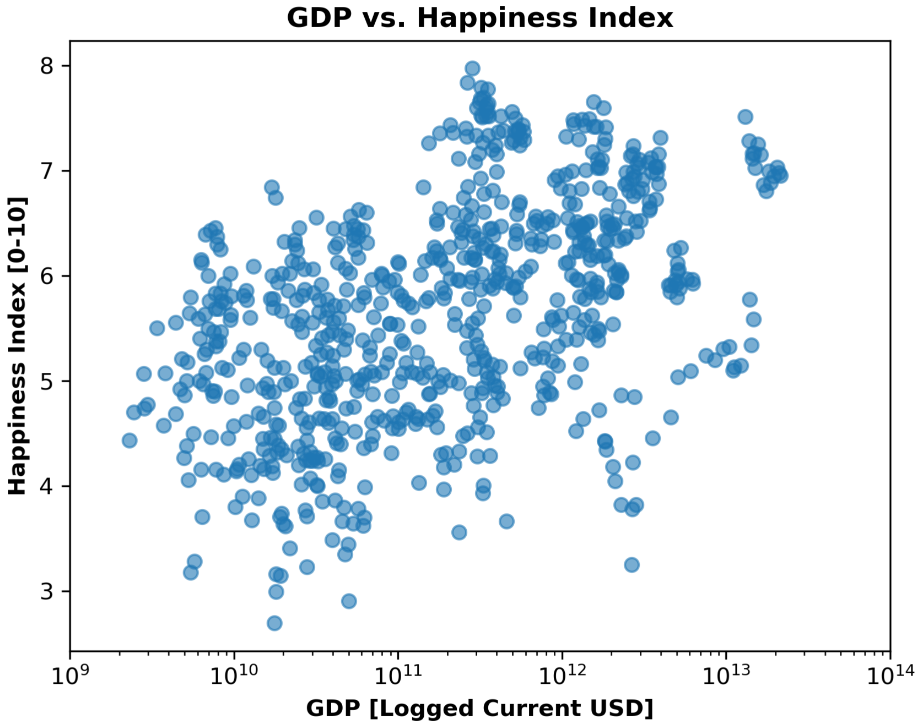
The below figures represent a sample of the graphics created during a thorough review of each of the features in my dataset.

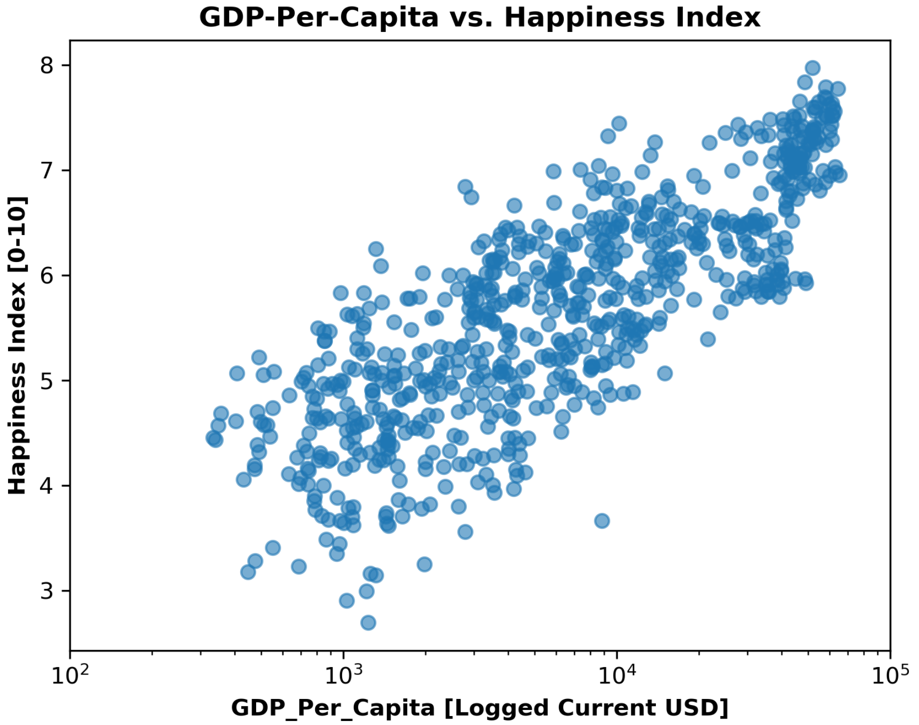
Analysis of the Target Variable



The histogram above shows the distribution of the target variable, the Happiness Index. From the plot we can see that the target is normally distributed, meaning that there is not a skew or obvious outliers for happiness across countries. Additionally, this histogram shows that the index does not vary over several magnitudes. The mean of the target is 5.67; the expected value for a score that ranges from 0-10. This histogram informs my splitting methodology as it confirms that a stratified split is not necessary since the index is evenly distributed.

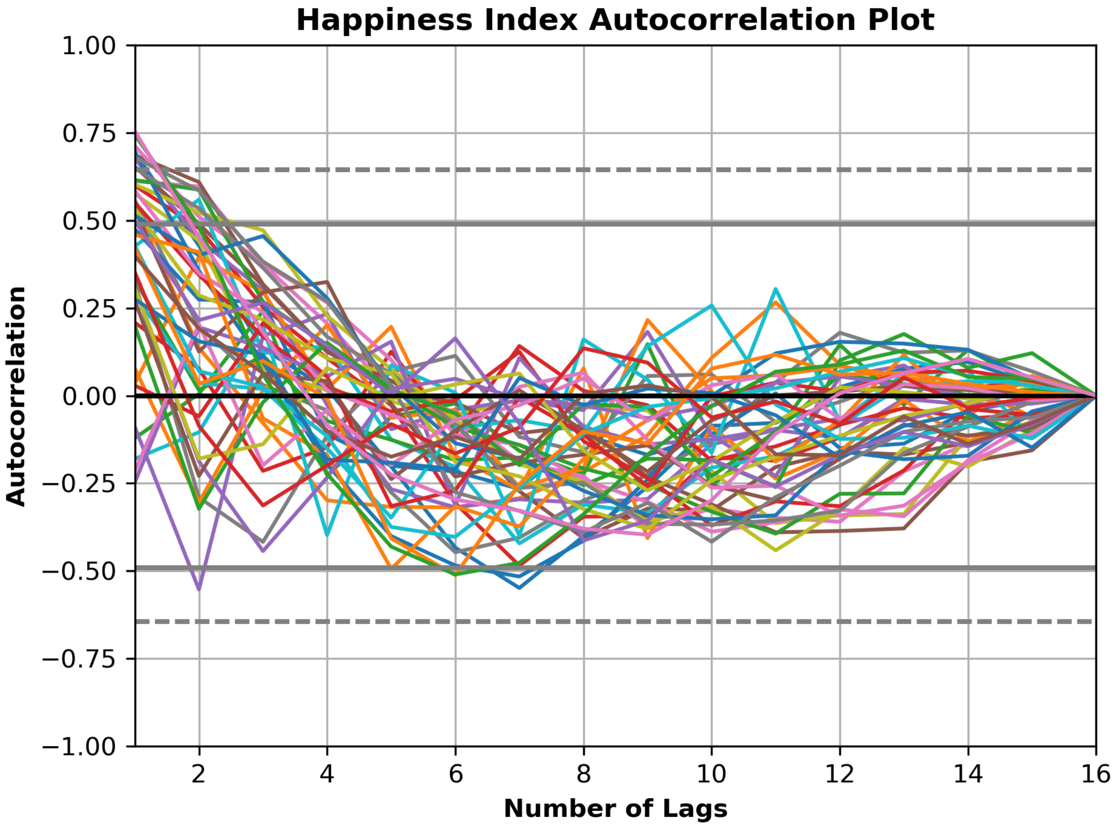
Comparison of GDP and GDP-Per-Capita





The scatter plots above illustrate the positive correlations between both GDP and GDP-per-capita with the Happiness Index. These relationships, as expected, indicate that wealthier countries tend to have happier citizens. Interestingly however, the GDP-per-capita plot displays a tighter fit of this positive correlation to the Happiness Index, meaning that GDP-per-capita explains more of the variance in the Happiness Index than GDP. This may tell us that the distribution of wealth in a country is a better indicator of its citizen's happiness, than the size of an economy alone.

Review of Target Variable's Autocorrelation



The above plot shows the autocorrelation of the Happiness Index for all 47 countries measured over 16 years of lag. This graphic reveals that, despite the differences in countries' wealth, population, and CO2 emissions, happiness among people of the world follows a similar correlation throughout time. For instance, the index is highly positively correlated before 4 years of lag, but slowly becomes negatively correlated as time progresses. This may tell us that happiness is an innate human feature which doesn't conform to political, economic, or social boundaries.

Methods

When developing my splitting strategy, I needed to avoid potential data leakage issues, since my data is time-series based and therefore the quantities in a given year are dependent on the previous years' information. Additionally, I decided to test different lag intervals on my dataset, in order to determine the best amount of past time to use for my model's predictions. Therefore, before splitting my dataset I created five separate dataframes which had time-lag amounts ranging from 1 to 5 years. Afterwards, I split my data chronologically into separate train (first valid year-2015), validation (2016-2017), and test (2018-2020) sets. This splitting method ensures that my model will not use future information to predict past indexes.

To preprocess my data, I used the OneHotEncoder on the country names, since this is a categorical variable that cannot be ordered. Additionally, I used the StandardScaler on the other five continuous features, as GDP and CO2 emissions were not reasonably bounded, and StandardScaler would scale the features to unit variance. My preprocessed datasets had a range of 60-80 features, depending on the amount of time-lag.

After splitting and preprocessing my data, I designed an ML pipeline to test various lagged datasets on five unique model types. For each model, I tuned several hyperparameters and calculated RMSE test scores to determine which models performed the best on different lagged time periods. I chose the RMSE metric because it describes the error rate of a regression model in the units of the target variable; this metric allowed me to compare different models' predictions in terms of the target variable. I measured uncertainty in my ML methods by comparing the standard deviations of test scores for the various lagged dataset, observing which lagged periods allowed for the greatest predictive power of my models. The below table outlines the different models and hyperparameters that were tested:

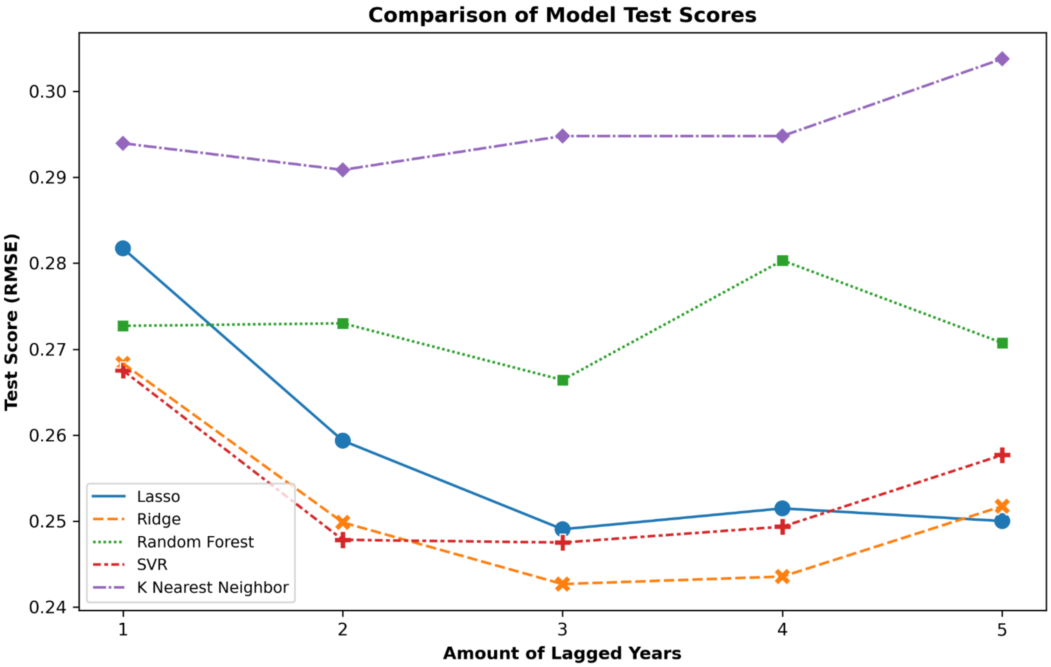
Model	Hyperparameter(s)	Values
Lasso Regression	Alpha	np.logspace(-30,10,21)
Ridge Regression	Alpha	np.logspace(-7,7,21)
Random Forest	Max Depth Max Features	[1, 3, 5, 7, 10,15,20] [0.15, 0.25, 0.5,0.75,1.0]
Support Vector Regressor	Gamma C	np.logspace(-8,0,15), np.logspace(-1,3,15)
K Nearest Neighbors	Neighbors Weights	np.linspace(1,10,10) ['uniform','distance']

Results

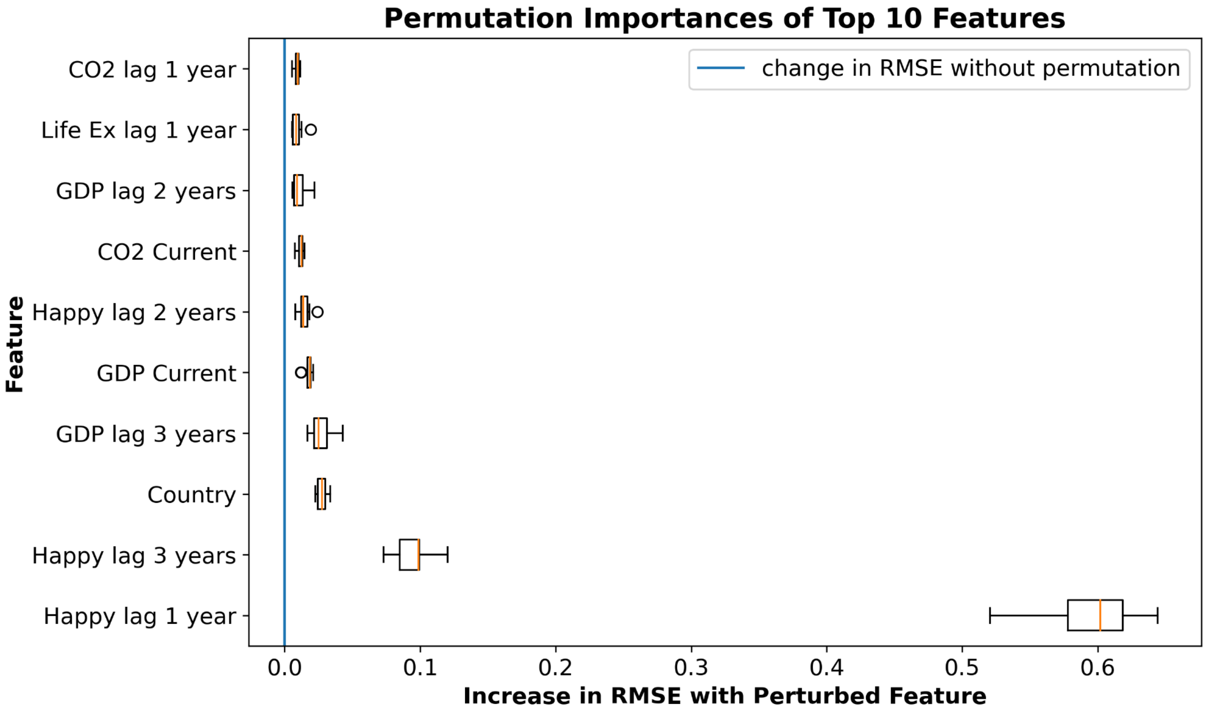
My model performed exceptionally well compared to the baseline RMSE score. To calculate the baseline score, I predicted the mean value of the happiness index for each datapoint, and found an RMSE of 0.988. Each of my models performed well below this baseline, although their test scores varied across lagged sets. The best model, the Ridge regression, performed 80 standard deviations below the baseline. The below table outlines the average RMSE scores and standard deviations for each model:

Model	Average RMSE	Standard Deviations	Std. Below Baseline
Lasso Regression	0.2583	0.0122	60
Ridge Regression	0.2512	0.0092	80
Random Forest	0.2735	0.0149	48
Support Vector Regressor	0.2539	0.0077	95
K Nearest Neighbors	0.2956	0.0043	161

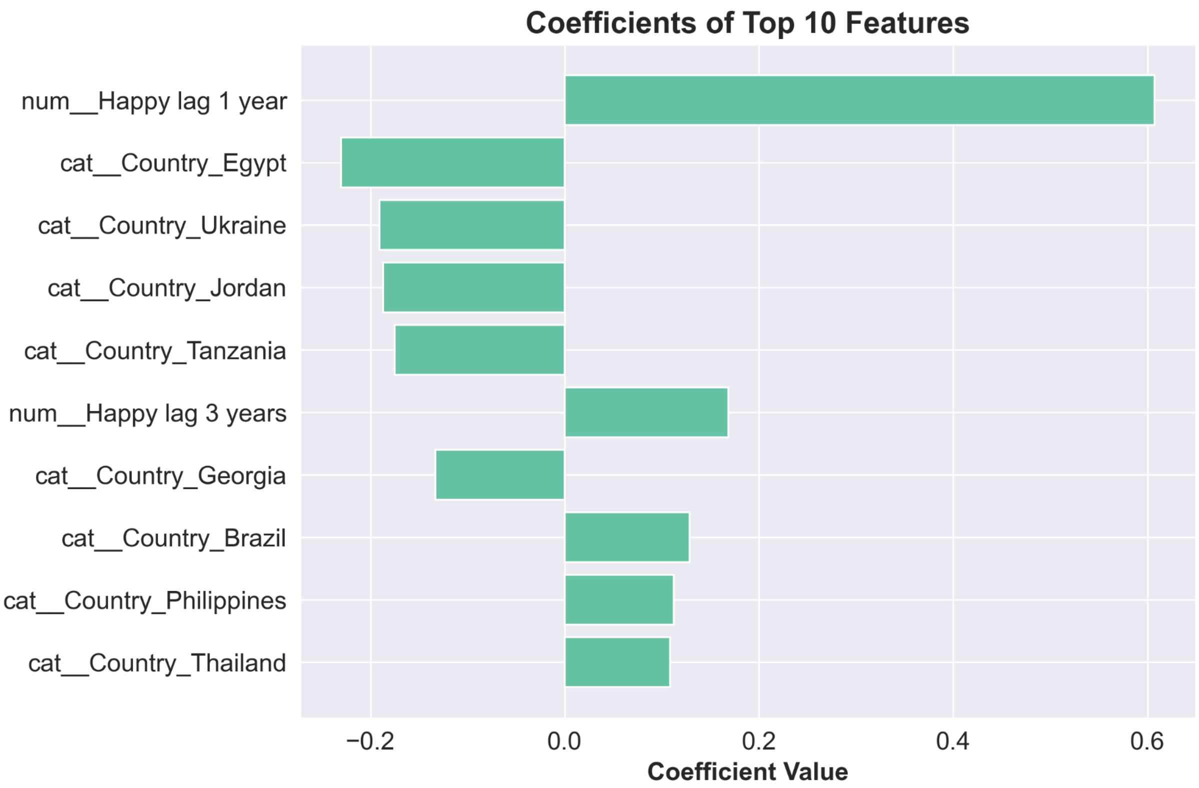
The below graphic displays the performance of each model over the five separate lagged datasets. From the plot we can see that the majority of the models, with the exception of K-neighbors and SVR, performed the best with the 3 year lagged dataset. This set offers the best compromise between having more feature data per datapoint while minimizing time loss from lagged periods. The best overall performer was the Ridge model on the 3 year lagged dataset.



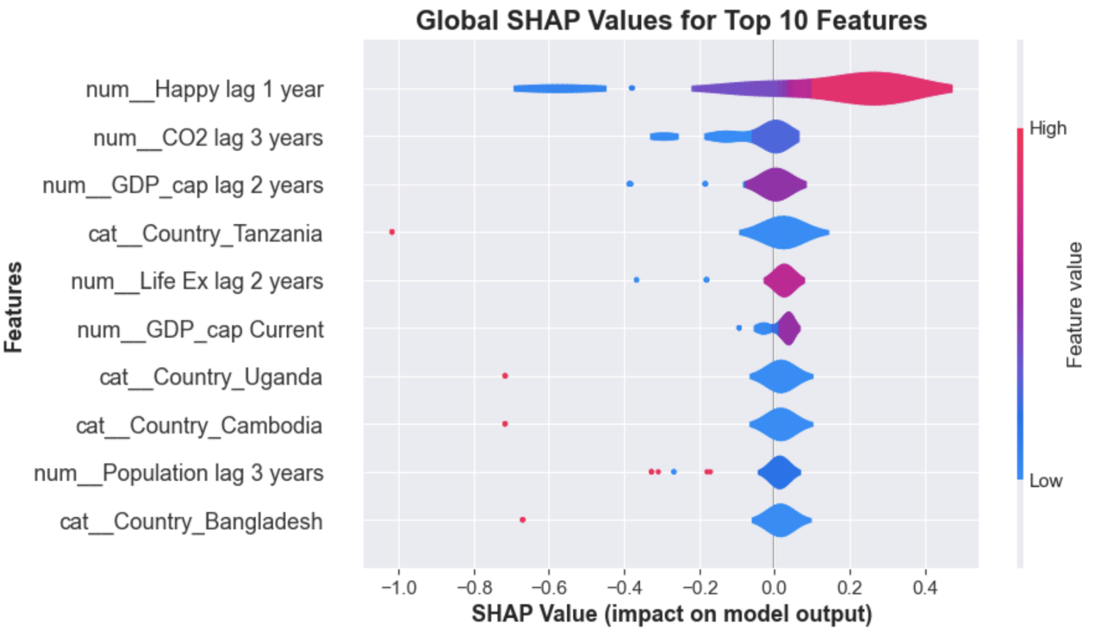
After preparing the models, I investigated the global impact of each feature on the overall predictions of the Ridge regressor. The below graphic shows the permutation importance of the top 10 features. This experiment measured the increase in RMSE when each of the features were individually randomly shuffled, thus breaking their relationship with the target variable. From the plot we can see that the most important feature is the 1 year lagged Happiness Index. This makes intuitive sense, as the happiness index of a given year should be relatively consistent with the previous year's index. On the other hand, CO2 lagged 1 year had a minimal effect on the predictive power of my model. This result is in line with our EDA, as the CO2 emissions of a country were not highly correlated with the happiness index.



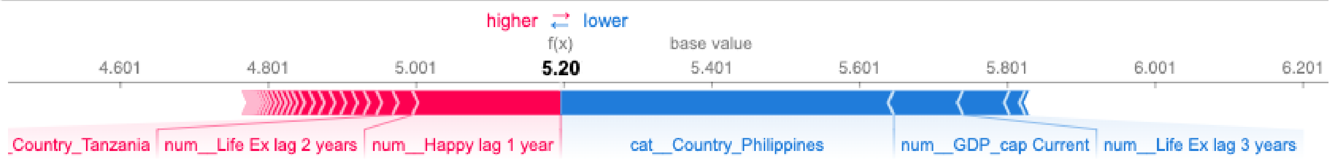
Similarly, I examined the model's coefficient values for all of the features to determine the predictive importance the model had assigned to each. From this investigation, I found that the 1 year lagged Happiness Index had the highest coefficient. Again, this confirms the intuition that the previous year's happiness index is strongly correlated with and predictive of the present year's index. Interestingly, many of the Country features also had high valued coefficients. This tells us that the model associates certain countries with higher or lower happiness levels, and uses this information to strongly influence its predictions for a given datapoint. The below graphic depicts the coefficient values for the top 10 most influential features.



Next, I looked at global SHAP values to obtain a better understanding of the predictive power of each feature. The figure below illustrates how high and low values of the top 10 features affect the model's predictions. Again we see that Happiness lagged 1 year is the most important predictor. Interestingly however, we see that low valued Happiness indexes are almost twice as predictive as the high valued indexes. This tells us that the model feels more confident about predicting a datapoint's current happiness index if the datapoint's previous index was relatively low.



Lastly, I calculated local SHAP values to explore the importance of each feature for a given datapoint. The below force plot illustrates the features that contribute most positively and negatively to the prediction of a specific datapoint. This plot confirms once again that the Happiness Index lagged 1 year is a strongly predictive feature, as it raised the predicted value for this point. Furthermore, we again see that the model utilizes certain countries to inform its predictions. For instance, the plot below shows that the model increased its prediction of the index's value as it noted that the given datapoint did not belong to Tanzania. This tells us that the model has made connections that span across countries, as it compares the historic happiness indexes of different regions to predict a score for a given country.



Outlook

Despite the model's success in predicting the target variable, there exist many weak spots in my modeling approach that can be improved for better results. For instance, the current model only considers a narrow array of economic and social features that fall far short of encapsulating the lives of individual citizens around the world. To improve relevancy of my model, more comprehensive data, on areas such as inequality and civil liberties, should be included in the feature matrix. Additionally, the interpretability of the model can be improved by enhancing the tuning techniques to include an examination of the non-lagged dataset. This exercise would provide a baseline understanding of how important each feature is to the prediction of the current year's target. Lastly, we must consider that the model is only capable of predicting the future up until the past. As economic and environmental realities continue to shift, the model's assumptions may no longer hold and the model may lose significant predictive power overtime.

References

Projcheski, Filip. “How to Predict Happiness Index Using Machine Learning Model: Laconic Ml.” *Laconic Machine Learning*, Laconic, 30 Apr. 2020, https://laconicml.com/predict-happiness-index-using-machine-learning-model/.

Yi, Chan Min. “Predicting Happiness Using Random Forest.” *Medium*, Towards Data Science, 14 Sept. 2020, https://towardsdatascience.com/predicting-happiness-using-random-forest

Ortiz-Ospina , Esteban. “Happiness and Life Satisfaction.” *Kaggle*, 27 Aug. 2022, https://www.kaggle.com/datasets/programmerrdai/happiness-and-life-satisfaction?select=life-satisfaction-vs-life-expectancy.csv.