Predicting Life Expectancy

Amber Aposhian, Sam Carpenter, Spencer Halverson, and Kate Wall

December 2021

1 Introduction

Humans care a lot about how long they have to live. As such, there is plentiful research about predicting and modeling life expectancy. Some of the most cited of these papers show that record national life expectancy (highest average life expectancy in a country) is growing linearly with time, seemingly without limit [2, 5].

Our research questions are focused on two main aspects: First, given a dataset with certain statistics about countries, can we predict national life expectancy? Second, which factors or combinations of factors are most correlated with life expectancy? Naturally, if we cannot make an accurate model to predict national life expectancy, it will be virtually impossible to answer the second question.

2 Data

Our dataset comes from information gathered by the WHO, and it is found at Kaggle.com [3]. Our code for this analysis can be found on our GitHub repository [1].

The dataset lists countries of the world by year, along with the life expectancy and other contributing factors. Specifically, the features are: `Country`, `Year`, `Status`, `Life Expectancy`, `Adult Mortality`, `Infant Deaths`, `Alcohol`, `Percentage Expenditure`, `Hepatitis B`, `Measles`, `BMI`, `Polio`, `Under 5 Deaths`, `Total expenditure`, `Diptheria`, `HIV/AIDS`, `Population`, `GDP`, `Thinness 10-19 years`, `Thinness 5-9 years`, `Income Comp`, and `Schooling`. The data measures national averages by country and year. It contains nearly 3,000 samples. We hypothesize the most important predictors will include: infant mortality, alcohol consumption, prevalence of specific diseases, and GDP. We also suppose that population and education levels will not be great predictors of life expectancy. One thing to consider is that many of the columns contain data that might be strongly correlated (for example infant deaths and under five deaths), so we will need to examine which features are collinear.

The data spans 16 years (from 2000 to 2015) and includes 133 unique countries. There are 109 countries with at least 10 data points in the sample. The mean national life expectancy in this dataset is 68.89 years, with a standard deviation of 9.30 years.

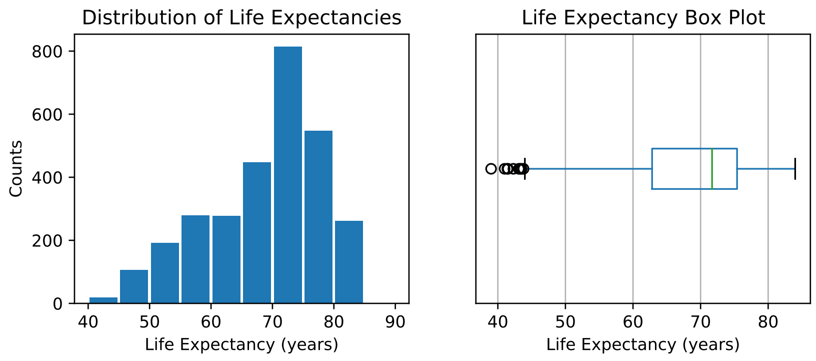

Figure 1: (a) Histogram of every life expectancy value in the data; (b) Box plot with min, median and max of life expectancy values.

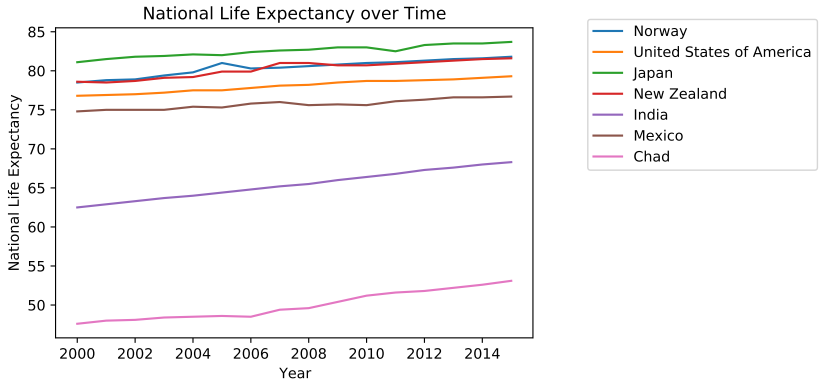

Figure 2: Line plot of average life expectancy in specific countries over the date range.

One limitation of our dataset is that it does not have separate average life expectancy for males and females. Life expectancy estimates for individuals are highly influenced by sex. Accordingly, life expectancy increases linearly with a slightly larger slope for national female averages than national male averages [5]. However, since we are predicting national life expectancy and not individual life expectancy our predictions will still be helpful.

3 Data Cleaning

We examined our data to ensure it was reliable. By looking at scatter plots of life expectancy for each country, we found about 200 data points that were outliers, and filled them with the correct value from World Bank [4]. We chose to spend the time doing this because we felt having correct data for the value we were trying to predict would improve our accuracy. Indeed, this data cleaning improved the mean squared error (MSE) for our models. However, it didn't change the accuracy by much.

Another step we took to validate our data was writing a function that found data points that were further than ±3 standard deviations away from the mean of each country for each feature. These values were then changed to NaNs (not a number) and were corrected using methods described below. We found that values for these features remained relatively constant for each country. We concluded that when they were more than three standard deviations away from the mean they were incorrect. There were 379 of these outliers we were able to correct.

Our dataset had 2,563 NaN or missing values. There are two functions we primarily used to fill these values in the life expectancy dataset. These functions handle differences in the nature of these NaN values. For some columns, including country, year, and status, there was not any missing data. All the missing data were from numerical features.

The first function filled NaN values by country. We made the assumption that all of these values were generally stable within a country from year to year. If a specific year was missing data, but the neighboring years had values, the missing data was filled with the average of the previous and following year. When this was not the case (if there were several NaN values in a row or the missing data was in the first or last year recorded for the country) the NaN value was filled with the average of the entire column. This function was able to fill 865 of the approximately 2,500 missing data points we originally had. This left us with 1,700 data points that were still missing; these could not be filled because the data was missing for the entire column for that country.

By looking at summary statistics, we decided to group by Status to fill these NaN values. In total, developing countries had 1,300 of these missing values. The features missing the most data were GDP, Population, Income Composition, and Schooling. When comparing summary statistics on the GDP column, we found developing countries had a lower standard deviation than the full dataset by approximately 3,000, and the mean was lower by 5,000. The other three factors had similar shifts for developing countries versus the whole dataset.

This lower variance among countries with the same Status motivated our second function, which filled NaN values by Status and by Year. We made the assumption that values are comparable for countries with the same Status, developing or developed. The missing data was filled with the average of other countries with the same Status for the same Year. This function was able to fill all the remaining missing values.

Developed countries only had 300 missing data points, and the most came from Hepatitis B and Population. For both of these factors, the standard deviation for developed countries was significantly lower than it was for the whole dataset. For these reasons we felt confident filling with averages by status. For random forests, which can handle missing data, we did not use this method because these values are less reliable than the previous method.

Each of these three functions (for detecting outliers and filling NaN values) are robust enough to easily transfer to other problems. Many other datasets have natural groupings like country, year, or gender. The names of the columns to group by can easily be replaced in these functions, and then they will be applicable to other problems.

We considered creating new features. One obvious choice was GDP per capita. We created this by dividing GDP by Population and dropping GDP. Through further analysis, we found that this new feature had high collinearity with other features and did not enhance our models. We felt that the features we had were sufficient in predicting life expectancy, and did not use new features in our final models.

4 Methods

4.1 Scoring

We tried classifiers, we really did. But they just aren't the right thing for predicting a specific number. We found regressors outperformed classifiers.

We used both Mean Squared Error (MSE) and Mean Absolute Error (MAE) to score our regression models. Mean Squared Error is a standard metric that penalizes larger deviations. Mean Absolute Error tells us (on average) how far off our model's prediction was from the true value. For classifiers, we found the average accuracy, whether or not our test data is correctly classified. Regressors were scored by R^2, which measures the percentage of the variation in our test data that is explained by our model. We fit the models, and averaged these metrics ten times to account for randomness in choosing training and test data. This average was our final reported value. These metrics allowed us to compare the performance of our different models. (See 1).

4.2 Classifiers

Life expectancy is not categorical by nature. In order to use classifiers, we decided to separate the life expectancies into 5-year bins. The classifier models predicted a specific bin given the input data.

4.2.1 Naive Bayes and Gaussian Naive Bayes

As mentioned above, classifying is probably not the optimal approach to this problem. Despite that, we wanted to see if Naive Bayes and Gaussian Naive Bayes could be effective. We hypothesized that because life expectancy is likely normally distributed across countries, that Gaussian Naive Bayes could be a successful model. Upon executing the models we saw Gaussian Bayes had about 17% higher accuracy than Naive Bayes. The Gaussian model also had much better MSE and MAE (see Table 1). Regardless, both models under performed compared to regression models.

4.2.2 Logistic Regression

We also decided to try Logistic Regression (which is actually not a regression model, as it predicts categories). We used a regularization strength coefficient of 0.001 on our L2 regularized model. We found this coefficient by conducting a simple grid search to find the regularization strength that yields the highest scoring model. This gave us an MSE of 8.129, MAE of 1.363, and a classification bin score of 74.5%.

4.3 Regressors

4.3.1 Random Forest

Unsurprisingly, a random forest did well on predicting life expectancy. We used a random forest regressor, and got an average out-of-bag score (score on data that wasn't used for training) of 0.940, see Figure 3. After running a grid search to find the optimal parameters, we constructed a forest with 250 trees, with a max depth of 6, and at least 5 samples in each leaf.

We have some concerns that trees with these parameters could be overfitting. However, random forests usually avoid overfitting because each tree can only train on a subset of features and data points. Additionally, as long as we can be reasonably confident in our model, we trust the random forest to tell us which features were most important in the predictions. This feature importance is extremely useful in answering our research question. So random forests were a great tool, offering both an accurate model and rankings for the importance of various features.

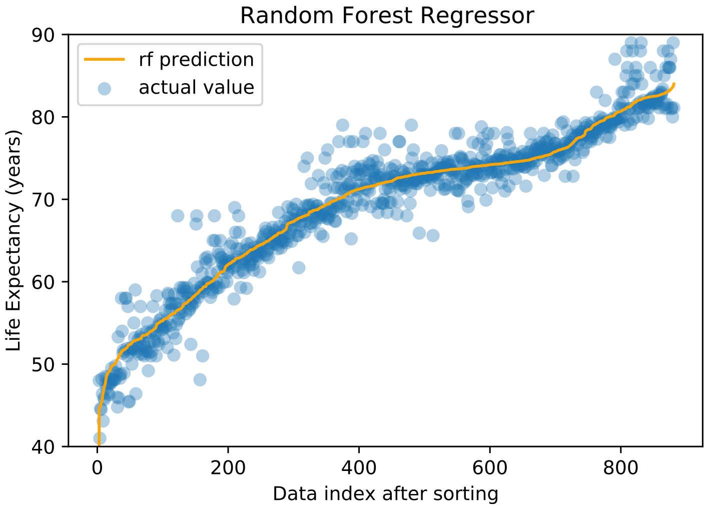

Figure 3: Performance of our random forest model. This forest was trained on 70% of our data. The blue points represent the remaining 30% of the data, while the orange line represents the model's prediction. Ordering on the x-axis is arbitrary, but sorted from least to greatest predicted value for visual convenience.

4.3.2 Linear and LASSO Regression

Of all the regression and classification models we used, Linear Regression performed the best with MSE of 1.747, MAE of 0.793, and R^2 value of 0.98. Perhaps the success of this model comes from the previously mentioned observation that our life expectancy increases linearly by year, and other diseases like Diphtheria, AIDS, Measles, and Hepatitis B decrease by year as well. This makes a Linear Regression model a good fit for the data (see Figure 4). Adding LASSO regularization, our metrics improved slightly with MSE of 1.53, MAE of 0.793, and R^2 value of 0.982. Ridge Regression and Elastic Net did not perform as well. One important thing to note is that these training methods were quite fast. Linear Regression (without regularization) took about 0.31 seconds to run and LASSO regression took 2.82 seconds to run.

4.3.3 Gradient Boosted Trees

Gradient boosted trees did fairly well, but not as well as linear regression. Our model parameters were 250 trees, a learning rate of 0.02, and a max depth of 3. Using a similar method for training and testing we obtained the following results: MSE of 3.46, MAE of 1.40, and R^2 value of 0.961. While these results are decent, this method takes significantly longer to run than linear regression, about 24 seconds. We concluded that the gradient boosted trees method is not as efficient as the previously mentioned models for this particular problem.

5 Feature Importance

Both random forests and gradient boosted trees allow you to identify which features are the most important. This is done by randomizing the values for a given feature, passing those data points in, and comparing accuracy to the normal data. Both types of models placed `HIV/AIDS` as the most important feature for determining life expectancy. This was closely followed by `Income Composition`, `Adult Mortality`, `Under 5 deaths`, and `Thinness 5-9 years`, which were ranked by both models in the top 7 most important features. (`Income Composition` is Gross National Income per capita, `Adult Mortality` is deaths between the ages of 15 and 60.)

The importance of these features was not surprising to us. We were surprised, however, to see that `Country` did not rank highly; in fact, the accuracy of random forests remained about the same after dropping the

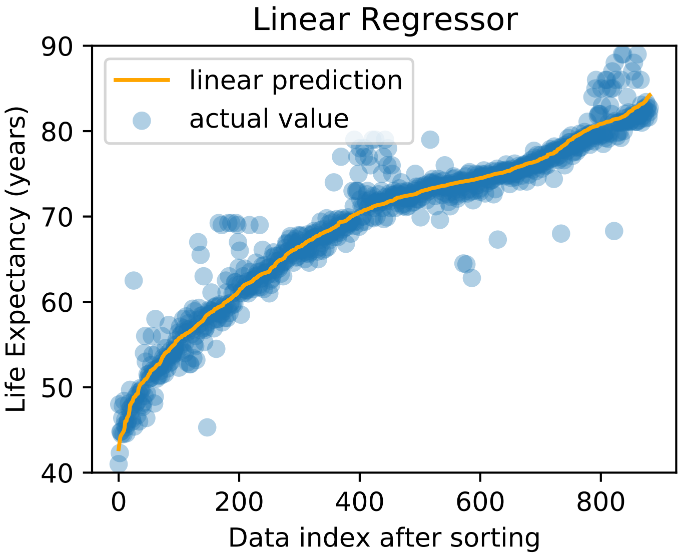

Figure 4: Performance of our linear regression model, with no penalty. The model was trained on 70% of our data. The blue points represent the remaining 30% of the data, while the orange line represents the model's prediction. Ordering on the x-axis is arbitrary, but sorted from least to greatest predicted value for visual convenience.

country column. Once countries were dropped, we were surprised to see that random forests reported `status` (IE developing or developed) as the least important feature. We hypothesize that this is because a binary classification is not nuanced enough to give insights beyond those of other features like immunization against diseases, other health indicators, or year.

We also used linear regression to find the least important features. To do so, we ran the model, looked at summary stats, and removed the feature with the highest p-value of $p >= 0.05$. We repeated this until all features had p-value of $p < 0.05$.

We found that dropping `Population`, `Hepatitis B`, `Thinness 5-9 years`, `Year` and `Total expenditure` lowered both the AIC and BIC on the linear regression. Dropping `Thinness 10-19 years` further lowered the BIC. This suggests that these features may be correlated to other remaining features, and they are not necessary for predicting life expectancy.

Model	MSE	MAE	Score
Naive Bayes	91.20	6.96	0.296
Gaussian NB	45.67	4.09	0.464
Logistic Regression	8.13	1.36	0.745
Random Forest	6.37	1.76	0.940
Gradient Boost	2.99	1.32	0.965
Linear Regression	1.75	0.79	0.985
LASSO	1.53	0.79	0.982

Table 1: Summary of the mean squared error (MSE), mean absolute error (MAE), and score for each model. Notice that Naive Bayes, Gaussian NB, and logistic regression are classifiers, and are scored based on whether they classified a point into the correct age interval [i,i+5] for i in $\{40,45,\ldots,85\}$. Scores for the regressors are the R^2 coefficient. Each statistic is an average across 10 train-test splits.

6 Ethics

Our dataset was created from publicly available data, and it is aggregated at a national level, so we do not see privacy concerns related to the data collection. However, because these models make predictions based on national averages, they should not be used to predict individual life expectancies. For example, it would be both unethical and unhelpful to use this model in assessing risk for life insurance.

When considering feature importance, it's important to remember that even if our models say certain features are excellent predictors of life expectancy, it doesn't mean those features are good tools to change life expectancy. More research should be done to determine which factors actually affect life expectancy, rather than simply predicting it. Additionally, the correlation may not follow what our societal biases would predict. For example, although western cultures think of being thin as an indicator of health, thinness among children 10 to 19 years old was negatively correlated with life expectancy.

7 Conclusion and Future Work

As shown in Table 1, our best models were random forests, gradient boosted trees, and linear regression. We were impressed that a simple Linear Regression model was able to perform so well. The R^2 values from these models were all above 0.9. Since our primary goal was to examine which features were the greatest indicators of life expectancy, we were pleased to see that these features did a very good job at predicting life expectancy. Although these results do not reveal what causes life expectancy to increase or decrease, they provide insights into factors that are correlated.

We found deaths caused by HIV/AIDS per thousand in children 0-4 years old was the strongest predictor of life expectancy. We again emphasize that this statistic does not necessarily mean that these deaths greatly affect life expectancy, but rather are a strong indicator of it. We hypothesize that this is more reflective of a country's access to and effective administration of life-saving medicines. Still, the fact that the aftereffect of the AIDS pandemic is a strong indicator of life expectancy should not be minimized. We would be remiss not to mention that in the future, the aftereffects of the current COVID-19 pandemic could be strong indicators of national life expectancy, for similar reasons of access to health care and administration of vaccines.

There are several techniques we could test to improve the performance of our models. When we find correlation matrices or run linear regressions, the coefficients between features are skewed based on the magnitude of the features. To remedy this, we could try scaling each feature by its maximum value so that every feature is between 0 and 1. Then larger coefficients may give additional evidence of which features are more important predictors.

We would also like to find new data to research related problems. It would be great to find data split between male and female populations, since sex is a very important feature for predicting average life expectancy. Other features we would like to predict on include race/ethnicity, types of national healthcare systems, or diet. It would also be interesting to compare these results on national life expectancy to individual life expectancy. Would factors such as national income and schooling scale down and play a similar role on an individual's life expectancy (using household income and individual schooling)? Also, it would be interesting to examine whether national life expectancy is correlated with quality of life.

Future research in this field could significantly improve health outcomes in many countries throughout the world. This could lead to efforts in increasing vaccine availability in developing countries, and improved efforts to educate on disease prevention.

References

[1] Amber Aposhian, Sam Carpenter, Spencer Halverson, and Kate Wall. Life Expectancy GitHub Repo, 2021. Available at `https://github.com/karsmars/life-expectancy`.

[2] Jim Oeppen and James W Vaupel. Broken limits to life expectancy. *Science*, 296(5570):1029–1031, May 2002.

[3] Kumar Rajarshi. Life expectancy (WHO), 2017. Available at `https://www.kaggle.com/kumarajarshi/life-expectancy-who`.

[4] The World Bank. Life expectancy at birth, total (years), 2019. Available at `https://data.worldbank.org/indicator/SP.DYN.LE00.IN`.

[5] Kevin M. White. Longevity advances in high-income countries, 1955–96. *Population and Development Review*, 28(1):59–76, March 2008.