Contributing Factors of Human's Life Expectancy

Alexandra Li, Ruby Wu

Abstract

Life expectancy is the number of years a person can be expected to live and is often used as an indicator of citizens' health. A collection of social, political, and public-health factors of a country, such as education level and vaccination rate, is believed to determine the country's average life expectancy. Therefore, a discrepancy can be typically observed between developing countries and developed countries. Previous research has applied machine learning methods to identify crucial contributors to the average lifespan but is limited in terms of their examination of the contributing factors' distribution and also places inadequate emphasis on the role of developmental status. Our research utilizes four machine learning models, linear regression with and without regularization, random forest, and XGBoost, to address the aforementioned insufficiency to provide a more straightforward and interpretable view of the factors that impact average lifespan by regarding developmental status as a key characteristic to life expectancy research.

1 Introduction

It might be a surprise for some people that it is only in the past one and a half centuries did human beings reach a life expectancy of more than 40 years [6] (see Figure 1). This scenario is largely due to not only medical discoveries but also economic and social progress that are only made possible by the industrial revolution. Thanks to the emergence and rapid development of modern technologies, human average life expectancy has nearly doubled for all continents over the last century. Unfortunately, there still exists a huge gap between the average life expectancy of developed countries and that of developing countries, with European nations reaching an average of more than 75 years while Africans have the lowest average lifespan of fewer than 65 years [6]. The social and economic environment in which individuals grow up and live in no doubt contributes a lot to their overall

physical fitness, vulnerability to certain diseases, and rate of having unnatural deaths. Instead of a straightforward cause-effect relationship, a country's average life expectancy is the result of the concatenation and multiplication of various factors spanning health and political realms. For instance, if a nation is involved in warfare for years like countries in modern post-colonial Africa, casualties in the war will immediately cause a drop in average lifespan, but will also result in a lower percentage of the educated population due to the lack of resources or opportunity due to becoming child soldiers. Both poverty and this deprivation of a chance to achieve a certain level of education will then lead to inaccessibility to healthcare and effective treatment due to economic constraints.

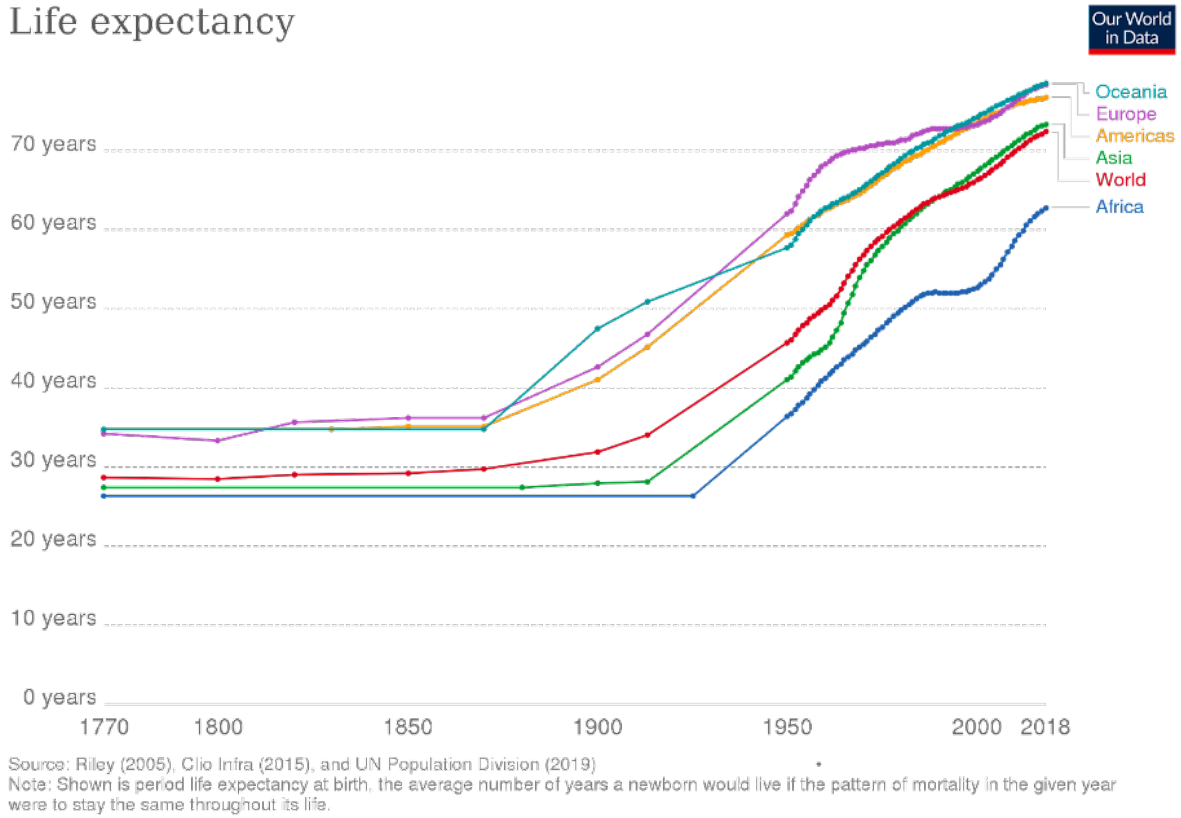

Figure 1: Human Life Expectancy Across Centuries [6]

However, the interaction between these factors remains unclear to us, and few studies have provided a more comprehensive picture in terms of the degree of influence individual factors have over life expectancy. In other

words, in order to understand what factors contribute more to increasing individuals' lifespan, what is in need of is a distribution of contributing factors. When different factors are intertwined with one another, they are capable of directly influencing each other, making it hard to tell which of them is a direct contributor to high life expectancy. Machine learning(ML) models make this task possible—-it is our objective to train ML models to obtain a list of weights corresponding to each contributing factor that provides a straightforward way to identify influential factors. Since the developmental status of a country is believed to be associated with such a combination of political factors related to life expectancy, it is crucial for researchers to devise an approach that categorizes countries based on these characteristics and goes from there to examine the distribution of contributing factors. Due to developing countries requiring more effort to make progress on increasing average lifespan and the limitation of data from developed countries, we compare the distributions of contributing factors of developing countries to that of all countries to see if they have different "most contributing" social attributes. With this information, we hope to offer government officials a more customized version of the specific areas or social indices countries will have to focus on in order to increase their citizens' health conditions and, therefore, average life expectancy.

The first phase of our pipeline consists of preprocessing and the usage of various supervised ML algorithms, by which we build models with our entire dataset. We then move to the second phase, where we build different models based on the developmental status of a country. By not only identifying factor's influence on life expectancy but also demonstrating how this influence differs for developing and developed countries, our research provides insights into the interplay between political factors and citizens' overall health as indicated by average lifespan such that more effective improvement can be made to specialized areas of the society.

2 Background

Background Previous work established a relationship between social attributes of the population, such as health conditions and happiness as a reflection of

mental health, with the population's corresponding life expectancy [2]. Educational factors have also been found to be explanatory toward longevity [3]. Though these studies have explored the contributors to life expectancy, they largely remain on the level of building a correlation instead of understanding what causes or "produces" human longevity as a health advantage as a species.

Other works that dig deeper into this relationship mainly focus on two aspects of life expectancy problems: prediction and the most influential factor. Some research has utilized tree-based ML methods to try to predict the mortality rate of individuals based on the social, economic, and environmental attributes of the environment they reside in [1]. Works of this nature are incapable of being self-explanatory, meaning that no examination or discussion has been built upon the distribution of contributing factors, but rather focus on a better prediction result. Other works that concentrate on finding the most influential factor are quite limited, with only some using relatively small amounts of data and the homogeneous population that does not really capture the variety of countries' developmental status [4]. Since a country's economic and social development is crucial to its citizen's overall health condition, we find it necessary to put countries with similar developmental statuses together to examine the distribution of influential factors.

3 Methods

In order to examine contributing factors to life expectancy, accounting for factors that influence one another, and adding interpretability to our model, we dealt with missing and unusual values by either discarding the data sample or setting them to nan. Then we set out to find factors that may be closely related and thus have a high covariance ratio, suggesting they may negatively interfere with the subsequent model application. By calculating Pearson scores which is a measurement of this correlation, we discard features that have strong associations with at least another variable, leaving only one in these similarity-based feature groups.

A linear regression model without regularization is used as our baseline model, which is a linear approach frequently used to model the relationship between a series of features and a dependent variable. By trying to find a line

that "best-fit" the training points, this approach helps make predictions based on known features. Linear regression with regularization ensures that the model that is built using the training set is not too large and thus overfitting to discourage complicated models so that it will give relatively good performance for test data. Random forest is where we build a forest of decision trees using random subsets of data and make predictions with respect to the decision boundaries included in the trained model. We use this ensemble method as our third due to its relatively high performance and interpretability brought by calculating feature importance. The last model we adopt for this research is XGBoost, another regression model that implements gradient boosting with parallel computation, which makes it relatively efficient and powerful.

This process will be repeated for developing countries to derive another distribution of contributing factors of life expectancy. To evaluate the performance of our models, we use the coefficient of determination, R-squared (R^2), which accounts for the percentage of the dependent variable's variance explainable by independent variables. K-fold cross-validation with grid search is the approach we used for hyperparameter tuning to find the optimal hyperparameters.

4 Experiments

4.1 Data

Life expectancy (WHO) is a dataset from Kaggle [5]. It records a total of 193 countries' average life expectancy from the year 2000 to 2015, including 21 features relevant to life expectancy. There is a total of 2938 samples, where each sample represents a country's statistics for a specific year. Countries' life expectancy ranges from 44 to 89 years old. The dataset includes health-related data (i.e., immunization rate and prevalence of thinness among different age groups) collected from the WHO data repository website, and economy-related data (i.e., GDP and income composition of resources) from the United Nation website.

4.2 Prepossessing

In the beginning, we conducted some data cleansing, including feature name standardization, country name standardization, one-hot code categorical variables, etc. We then ran some exploratory data analysis and found out that one major deficit of this dataset is the missing values. There are a total of 2563 missing cells (4% of all data), mainly observed in features such as population, Hepatitis B, GDP, and schooling for less-known countries like Tonga, Togo, Vanuatu, etc. In addition, we also noticed some counter-intuitive values; for example, the number of infant deaths per 1000 population has a value larger than 1000. Based on our knowledge, it is also unlikely for a developing country, especially one experiencing severe poverty, to have a value of 0 for the number of infant deaths per 1000 population. Therefore, before moving further to the modeling phase, we identified missing values in the dataset and dropped 10 samples where the value for life expectancy (y) is nan. We define the following to be nan:

- infant deaths: if a developing country has a value of 0 or if the value is larger than 1000
- percentage expenditure: has a value of 0
- measles: if a developing country has a value of 0 or if the value is larger than 1000
- under-five deaths: if a developing country has a value of 0 or if the value is larger than 1000

We identified 5509 missing cells (8.6% of all data), after which we ran Pearson correlation (Figure 2) to examine features sharing strong correlations and found that four pairs of features have an absolute Pearson score high than 0.8:

- infant death & under-five death: 0.99
- GDP per capita & percentage expenditure: 0.92
- thinness 1-19 years & thinness 5-9 years: 0.94
- schooling & income composition of resources: 0.80

Therefore we decided to drop one for each pair: "under-five deaths," "percentage expenditure," "thinness 5-9 years", and "income composition of resources."

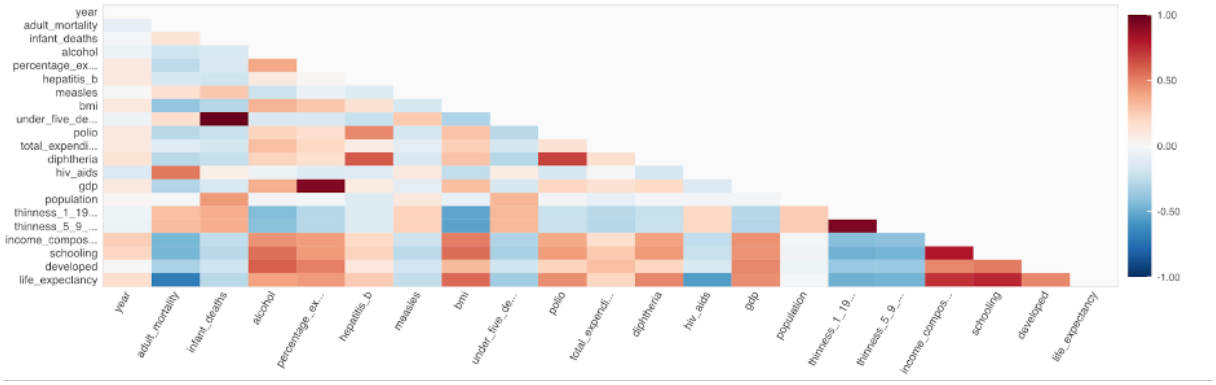

Figure 2: Heat-map for Pearson Score

Before the imputation, we split the data into training and testing data with train 70% and test 30%. To impute the missing values, we decided to use the KNN model since most of the missing values are from less developed countries and simply using mean or 0 is inappropriate. Then we train the min-max scaler on training data and transform both train and test data for KNN. Finally, we train the KNN model with k=16 on the training data since, in this dataset, each country has 16 samples. We end the whole process by imputing the missing value for both training and testing data.

The resulting training dataset for all countries has 16 features (excluding four features from feature selection and Country) with 2049 samples.

We also created another set of train and test datasets for developing countries only. After filtering out all samples for developed countries, we repeated the steps from the train/test split to the end. The resulting training dataset for developing countries has 15 features (further excluding development status) with 1691 samples. Due to the small number of developed countries in the current world and those included in our dataset, we did not train models specifically for developed countries but instead used models trained on the entire dataset as a benchmark against which models from developing countries can be compared.

4.3 Model

We select a total of 4 models: linear regression, linear regression with regularization, random forest, and XGBoost. The reason for our selection is that they provide good interpretability and can be instrumental for us in identifying vital features for life expectancy. For linear regression models, the coefficient of each feature can provide information on the exact impact the feature has on life expectancy (negative or positive, and the magnitude); for tree models, we calculate the feature importance using average information gain across all splits where the feature is used. Our following step is to train models using the two datasets respectively, and do hyperparameter tuning, whose result is determined by 10-fold cross-validation and grid search algorithms.

4.4 Result

As illustrated in Figures 3 and Figure 4, XGBoost and random forest obtain the highest and second-highest prediction accuracy, with the two linear regression models achieving a considerably smaller percentage of accurate prediction. Different models also share some common contributing factors, suggesting their significance to life expectancy. Factors of this kind include BMI, HIV/AIDS, schooling, and GDP per capita.

5 Discussion

Our research is divided into two parts: models that are trained using the whole data set, and models trained using only data from developing countries. In terms of both parts' prediction accuracy, part one outperforms part two for all four ML models, indicating the developmental state to be a crucial contributing factor to average life expectancy. Most of the other features found to either have a positive or negative effect on average lifespan for both parts are the same, with explicit ones including BMI, HIV/AIDS, schooling, and GDP per capita. Among these four features, the first two are directly related to fitness, nutrition, and susceptibility to diseases. A low BMI index is the result of malnutrition, indicating agricultural productivity to still be a huge issue in some parts of the world. This lack of food or richness of nutrients is frequently accompanied by low schooling and a higher rate of HIV/AIDS

	LR	LR with Regularization	Random Forest	XGBoost
Optimal Hyperparameter		Lasso with alpha = 0.1	max_depth=11 min_samples_leaf=5 n_estimators=500	colsample_bytree = 0.7 learning_rate = 0.1 max_depth = 7 min_child_weight = 3 n_estimators = 500 subsample = 0.7
R^2	0.821	0.802	0.949	0.953
Important Features	– Adult mortality – Infant deaths – Hiv aids + Schooling + GDP per capita	– Adult mortality – Hiv aids + Schooling + BMI	➢ Adult mortality ➢ BMI ➢ Hiv aids ➢ Schooling ➢ Thinness 1–19	➢ Hiv aids ➢ Developed ➢ Schooling ➢ Adult mortality

Figure 3: Model Results for All Countries

	LR	LR with Regularization	Random Forest	XGBoost
Optimal Hyperparameter		Ridge with alpha = 2	max_depth=11 min_samples_leaf=5 n_estimators=400	colsample_bytree = 0.7 learning_rate = 0.1 max_depth = 5 min_child_weight = 5 n_estimators = 500 subsample = 0.7
R^2	0.757	0.758	0.936	0.950
Important Features	– Adult mortality – Infant deaths – Hiv aids + Schooling + GDP per capita	– Adult mortality – Infant deaths – Hiv aids + Schooling + GDP per capita	➢ Adult mortality ➢ BMI ➢ Hiv aids ➢ Schooling	➢ Hiv aids ➢ Schooling ➢ Adult mortality ➢ Polio ➢ Diphtheria ➢ BMI

Figure 4: Model Results for Developing Countries

infection, all of which point out food insufficiency as the number one cause of low longevity.

One major difference between the distribution of contributing factors of

data including and excluding developed countries is two forms of vaccines are also found to be highly correlated with longevity in the latter's case but not in that of the former. It is clear that we will not be able to access this information if we haven't treated the developmental state of a country as an independent variable so that we can compare models trained by developing countries' data only with those by all countries. Polio and Diphtheria are both diseases preventable by vaccines and, therefore, rare in developed countries, but our results demonstrate that it is not the case for developing countries. Inaccessibility to polio and diphtheria vaccines increases the risk of infection for children living in poverty, and together with no access to advanced care, children's mortality increases, significantly lowering the region's average life expectancy.

One deficit of our research that we hope to improve upon during future studies is our current inability to gather information on the positivity or negativity of features' influence when using tree-based models. This is because, unlike the two linear regression models, tree splits are not associated with a positive or negative coefficient. Another issue is that we have to drop the data for developed countries due to the number of developed countries being rather small. If we can obtain more data by year for developed countries specifically, the model we trained by using these data can act as a reference to provide some valuable information regarding contributing factors of developing countries.

6 Conclusion

In this study, we accomplish two goals, first by examining the distribution of contributing factors in terms of their influence on counties' average life expectancy and second by using developmental status as a key variable to demonstrate the unique needs of the developing countries. We found that four factors directly related to poverty, which are GDP per capita, years in school, HIV/AIDS, and BMI, are key determinants for the average lifespan of all countries. And when extracting only samples from developing countries, we notice a low vaccination rate for polio and diphtheria to contribute to low longevity. By portraying a clearer picture of the relationship between

developmental status, other sociopolitical factors, and average lifespan, we offer insights into social realms that require the most international attention in order to help less developed countries' citizens obtain healthier life.

References

[1] Dorethe Skovgaard Bjerre. Tree-based machine learning methods for modeling and forecasting mortality. *ASTIN Bulletin: The Journal of the IAA*, 52(3):765–787, 2022.

[2] Preethi Chandirasekeran, Shreyaa Saravanan, Shreya Kannan, and V Pattabiraman. Analyzing implications of various social factors on life expectancy. *National Academy Science Letters*, pages 1–6, 2022.

[3] Khulood Faisal, Dareen Alomari, Hind Alasmari, Hanan Alghamdi, and Kawther Saeedi. Life expectancy estimation based on machine learning and structured predictors. In *2021 3rd International Conference on Advanced Information Science and System (AISS 2021)*, pages 1–8, 2021.

[4] Nittaya Kerdprasop, Kittisak Kerdprasop, and Paradee Chuaybamroong. Categorical modeling method to analyze factors relating to longevity of populations in the east and southeast asia. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*, pages 14–19, 2019.

[5] Kumar Rajarshi. Life expectancy (who), 2017.

[6] James C Riley. Estimates of regional and global life expectancy, 1800–2001. *Population and development review*, 31(3):537–543, 2005.

A Contributions

Our group has only two members, and we collaborate throughout the semester on the project proposal, spotlight, final presentations, technical works, and paper writing. Specifically, Alexandra focused more on the writing part (including presentations, proposal, and the final paper), and Ruby focused

more on the technical part (including data, preprocessing phase, model training and tuning, and evaluation.)

B Code

B.1 Code

Github Repository:`https://github.com/yunjiewu777/CS334`

B.2 Dataset

Github Repository: `https://github.com/yunjiewu777/CS334/blob/main/data/life_expectancy_data.csv`
Original Kaggle Link: `https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who`