


IST 718

Final Project

Life Expectancy

March 31st, 2023

Nick McFadden

Soham Nanavati

Sahil Nanavaty

Phuong (Adrianne) Van

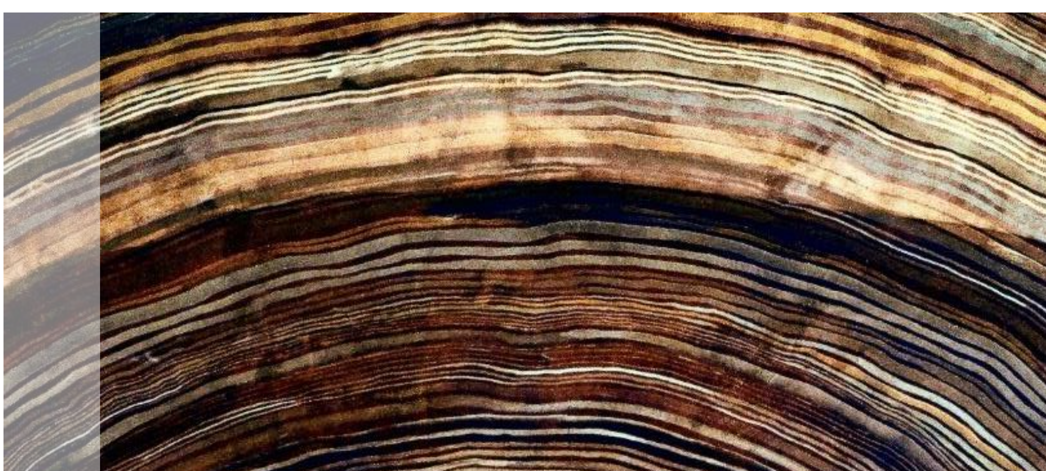
Cole Wood

Abstract/Exec Summary

This analysis was conducted to better understand factors that affect life expectancy. The main dataset was retrieved from Kaggle and then merged with 19 datasets from the WHO Mortality Database. The data tracked all countries' metrics from 2000 to 2015. The analysis determined that GDP_Healthcare, Adult Mortality (negative correlation), Income Composition of Resources (wealth classification), and Schooling were all highly correlated with life expectancy. Linear Regression and Neural Network models were applied to the data with high levels of accuracy and few insignificant values.

Unfortunately, we found that the original Kaggle data set had errors/inaccurate data values. This made some of the outputs unreliable/unusable but directional insights seemed meaningful. The variables above highlighted countries with higher healthcare spend did see some life expectancy gains but the potential factors are very broad and complex for any level of true certainty. However, the team believes that investment in socioeconomic mobility, education, and access to healthcare is a better use of funds. This translates into investment providing opportunities for the population and the benefits of those opportunities can trickle down to improved outlooks. A proper dataset/new factors could assist in clarifying the picture, but this is an elusive topic. If it was easy everyone would be living past 100!

In addition to the primary dataset, the team noted that additional data would be required to resolve the proposed business problems and to provide relevant recommendations based on the findings. The team sourced the other 19 datasets from the World Health Organization (WHO) and World Bank websites in CSV format as well. These additional sources provided information on the various types of diseases, illnesses, and injuries that are major contributing factors to life expectancy, as well as the healthcare expenditure percentages by country needed to address the business problems. These were later cleaned, manipulated, and merged with the original data source.



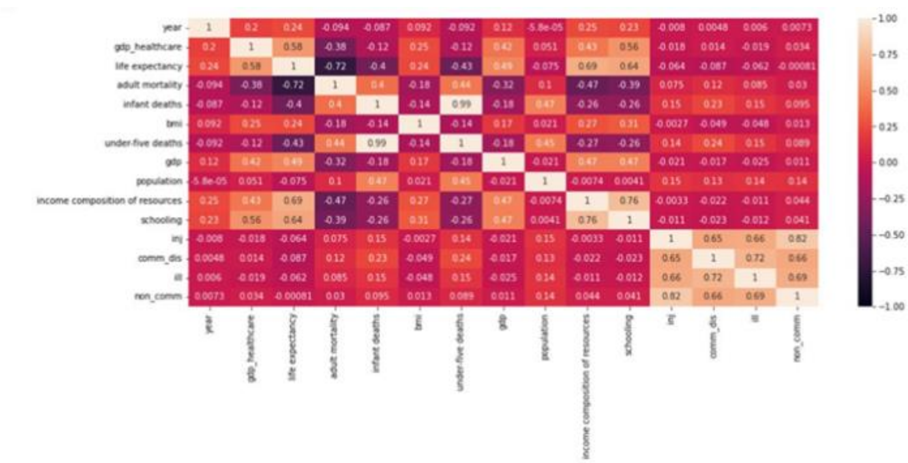
Specification

The team aimed to solve three key business problems pertaining to life expectancy data. The first of which was to determine the top contributing factors of life expectancy. Additionally, the team tackled the issue of whether life expectancy had a positive or negative relationship with the target variables (including population demographics, nutrition, and general health metrics). Lastly, the team tasked themselves with forming a recommendation for countries with relatively lower life expectancy values to potentially improve their average lifespans.

The original dataset was sourced from Kaggle in CSV format. It consisted of 22 variables and 2928 rows. The file contained country, year, status, life expectancy, adult mortality, infant deaths, alcohol related deaths, percentage expenditure, hepatitis B, measles, BMI, under-five deaths, polio, total expenditure, diphtheria, HIV/AIDS, GDP (Gross Domestic Product), population, thinness 1-19 years (malnourished), thinness 5-9 years (malnourished), income composition of resources (wealth class), and schooling. This data ranged from the years 2000 to 2015.

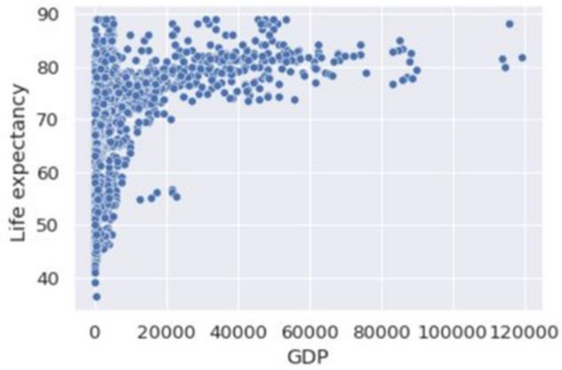
Observation

When seeking out what variables had the greatest effect on life expectancy, we first turned to making a correlation matrix to observe what variables were highly correlated to life expectancy (see below). We found that the variables GDP_Healthcare, Adult Mortality (negative correlation), Income Composition of Resources (wealth classification), and Schooling were all highly correlated with life expectancy.

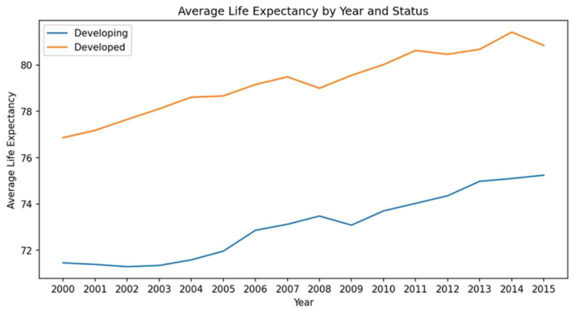


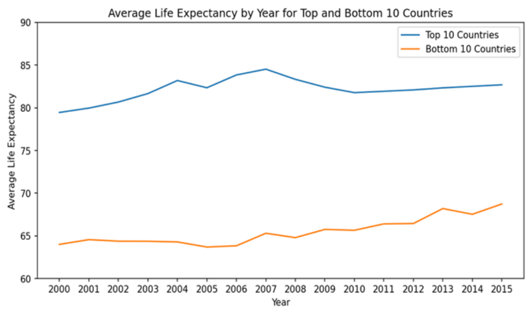
We also investigated the distribution of life expectancies within the data and found a mostly normal, slightly negatively skewed distribution of the histogram (see below). The most common age range was between 70 & 80 years old, with most falling directly in the middle at around 75 years of age.

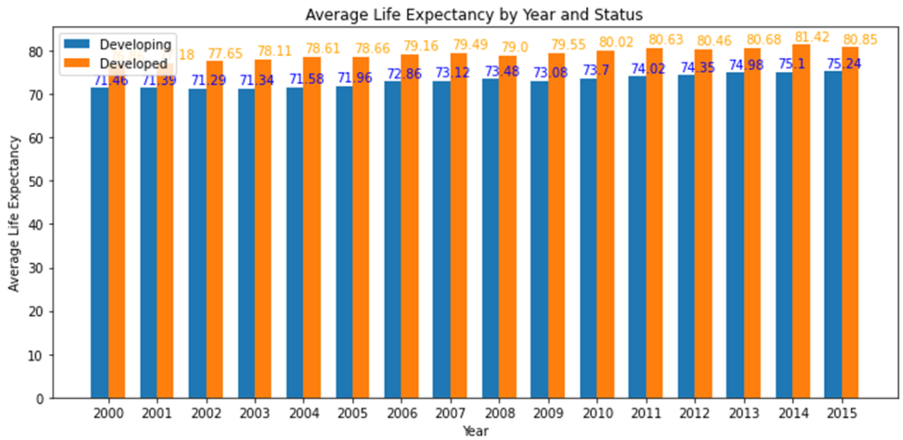
We observed that there is a slight correlation between life expectancy and GDP, with higher GDP countries recording higher life expectancies (see below).

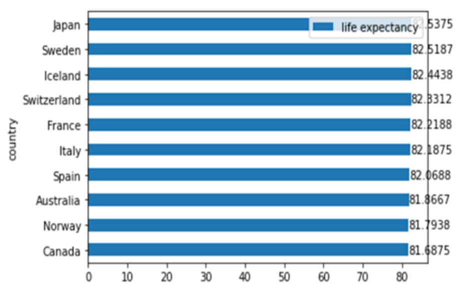


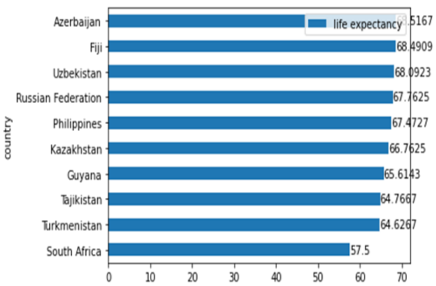
When looking at the status of the countries, we determined that both developed and developing countries average life expectancies were increasing over time from 2000-2015, but there are still some gaps between the two. When plotting a line graph that compares the top 10 countries with the longest life expectancy to the bottom 10 countries with the lowest life expectancy, we can see that countries with the lower life expectancy are trending slightly upward, while the top countries seemed to have plateaued since coming to a peak in 2007. There was a significant gap of 25.04 years between the country with the longest life expectancy (Japan, 82.5375 years) and the country with the lowest life expectancy (South Africa, 57.5 years). The top 10 and bottom 10 countries by life expectancy are shown below:



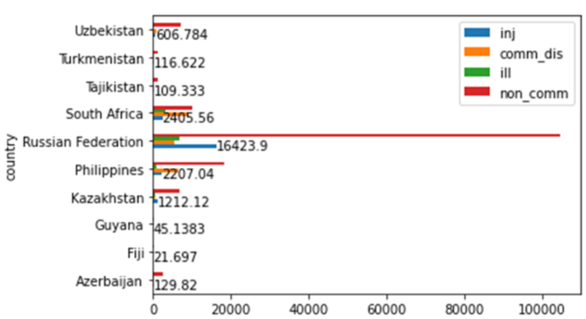








When our team originally gathered our data sources, we categorized the distinct types of diseases/illnesses based on the categories given by the WHO which were injuries, communicable diseases, illnesses, and non-communicable diseases. We found out that the type of ailments that were most prominent for the countries with the shortest life span were non-communicable diseases (see chart below). This is when we gathered the data sources for the 14 types of diseases found within non-communicable diseases and merged them to our current data to determine if we could pinpoint where these countries should concentrate their healthcare expenditure on.



Analysis

Linear Regression Model 1

The team initially conducted two separate linear regression models based on the life expectancy data sets. The first linear regression model contained all 21 variables from the original life expectancy data set and was expanded to include 20 additional variables. The 20 additional variables were created to break out the sex, age and status groups being into various buckets as seen in the screenshots on the following pages. Overall, the linear regression model surprisingly reflected a significantly low p-value below 0.05 for all variables. The low p-value indicates that all variables are statistically significant. However, the team was unable to determine why the model reflected a p-value near zero for all variables. While the team could not determine the root cause for the p-value, we believe there are potential issues within the data set that would need to be further explored next time if we had the opportunity to understand the issue. Nonetheless, the p-value from the linear regression model validated the four factors, including GDP healthcare, adult mortality, schooling and income composition of resources, that we zeroed on based on significant correlation were also statistically significant. We then evaluated the R-Squared value as seen in the screenshots on the following pages, which we determined was considered a satisfactory value at 0.768 since the value was above 0.70 and was not ideally above 0.80.

Initial Data frame (21 variables):

df_combln

	country	year	gdp_healthcare	status	life expectancy	adult mortality	infant deaths	bmi	under-five deaths	gdp	...	schooling	region code	region name	country code	sex	age group	inj	comm_dis	ill	no
0	Albania	2000	7.23337	Developing	72.6	11.0	1	45.0	1	1175.788981	...	10.7	EU	Europe	ALB	All	[All]	1416.0	1061.0	1679.0	
1	Albania	2000	7.23337	Developing	72.6	11.0	1	45.0	1	1175.788981	...	10.7	EU	Europe	ALB	All	[Unknown]	117.0	30.0	58.0	
2	Albania	2000	7.23337	Developing	72.6	11.0	1	45.0	1	1175.788981	...	10.7	EU	Europe	ALB	All	[0]	15.0	366.0	62.0	
3	Albania	2000	7.23337	Developing	72.6	11.0	1	45.0	1	1175.788981	...	10.7	EU	Europe	ALB	All	[1-4]	62.0	120.0	37.0	
4	Albania	2000	7.23337	Developing	72.6	11.0	1	45.0	1	1175.788981	...	10.7	EU	Europe	ALB	All	[5-9]	57.0	32.0	11.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
81412	South Africa	2015	8.79019	Developing	62.9	328.0	42	51.1	52	5769.772580	...	13.0	AF	Africa	ZAF	Unknown	[65-69]	2.0	2.0	4.0	
81413	South Africa	2015	8.79019	Developing	62.9	328.0	42	51.1	52	5769.772580	...	13.0	AF	Africa	ZAF	Unknown	[70-74]	1.0	2.0	4.0	
81414	South Africa	2015	8.79019	Developing	62.9	328.0	42	51.1	52	5769.772580	...	13.0	AF	Africa	ZAF	Unknown	[75-79]	1.0	2.0	1.0	
81415	South Africa	2015	8.79019	Developing	62.9	328.0	42	51.1	52	5769.772580	...	13.0	AF	Africa	ZAF	Unknown	[80-84]	0.0	2.0	2.0	
81416	South Africa	2015	8.79019	Developing	62.9	328.0	42	51.1	52	5769.772580	...	13.0	AF	Africa	ZAF	Unknown	[85+]	0.0	5.0	7.0	

70410 rows × 22 columns

Updated Data frame (41 total variables with the additional 20 variables):

df_combln2

	year	gdp_healthcare	status	life expectancy	adult mortality	infant deaths	bmi	under-five deaths	gdp	population	...	[50-54]	[55-59]	[60-64]	[65-69]	[70-74]	[75-79]	[80-84]	[85+]	[All]	[Un
0	2000	7.23337	0	72.6	11.0	1	45.0	1	1175.788981	38927.00	...	0	0	0	0	0	0	0	0	1	
1	2000	7.23337	0	72.6	11.0	1	45.0	1	1175.788981	38927.00	...	0	0	0	0	0	0	0	0	0	
2	2000	7.23337	0	72.6	11.0	1	45.0	1	1175.788981	38927.00	...	0	0	0	0	0	0	0	0	0	
3	2000	7.23337	0	72.6	11.0	1	45.0	1	1175.788981	38927.00	...	0	0	0	0	0	0	0	0	0	
4	2000	7.23337	0	72.6	11.0	1	45.0	1	1175.788981	38927.00	...	0	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
81412	2015	8.79019	0	62.9	328.0	42	51.1	52	5769.772580	5511976.68	...	0	0	0	1	0	0	0	0	0	
81413	2015	8.79019	0	62.9	328.0	42	51.1	52	5769.772580	5511976.68	...	0	0	0	0	1	0	0	0	0	
81414	2015	8.79019	0	62.9	328.0	42	51.1	52	5769.772580	5511976.68	...	0	0	0	0	0	1	0	0	0	
81415	2015	8.79019	0	62.9	328.0	42	51.1	52	5769.772580	5511976.68	...	0	0	0	0	0	0	1	0	0	
81416	2015	8.79019	0	62.9	328.0	42	51.1	52	5769.772580	5511976.68	...	0	0	0	0	0	0	0	1	0	

70410 rows × 41 columns

Linear Regression Model 1 Results:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        life expectancy   R-squared:                       0.768
Model:                            OLS   Adj. R-squared:                  0.768
Method:                 Least Squares   F-statistic:                     4916.
Date:                Sun, 26 Mar 2023   Prob (F-statistic):               0.00
Time:                        22:09:36   Log-Likelihood:            -1.3764e+05
No. Observations:               56328   AIC:                         2.754e+05
Df Residuals:                   56289   BIC:                         2.757e+05
Df Model:                          38
Covariance Type:            nonrobust
===============================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                          -64.7546      4.174    -15.515      0.000     -72.935     -56.574
year                             0.0729      0.003     26.863      0.000       0.068       0.078
gdp_healthcare                   0.6468      0.007     86.259      0.000       0.632       0.661
status                           0.7090      0.033     21.794      0.000       0.645       0.773
adult mortality                 -0.0231      0.000   -112.516      0.000      -0.023      -0.023
infant deaths                    0.5604      0.009     65.730      0.000       0.544       0.577
bmi                             -0.0110      0.001    -14.971      0.000      -0.012      -0.010
under-five deaths               -0.4920      0.007    -70.956      0.000      -0.506      -0.478
gdp                           2.794e-05   8.03e-07     34.773      0.000    2.64e-05    2.95e-05
population                   -4.178e-09   5.58e-10     -7.488      0.000   -5.27e-09   -3.08e-09
income composition of resources 11.1852      0.147     76.028      0.000      10.897      11.474
schooling                        0.2110      0.009     22.564      0.000       0.193       0.229
inj                          -3.729e-05   3.23e-06    -11.557      0.000   -4.36e-05    -3.1e-05
comm_dis                      1.818e-05   3.02e-06      6.026      0.000    1.23e-05    2.41e-05
ill                          -8.127e-07   4.07e-06     -0.200      0.842   -8.79e-06    7.17e-06
non_comm                      1.631e-06   4.82e-07      3.383      0.001    6.86e-07    2.58e-06
All                            -16.0793      1.044    -15.403      0.000     -18.125     -14.033
Female                         -16.1089      1.044    -15.432      0.000     -18.155     -14.063
Male                           -16.0643      1.044    -15.389      0.000     -18.110     -14.018
Unknown                        -16.5020      1.044    -15.800      0.000     -18.549     -14.455
[0]                             -3.1267      0.205    -15.218      0.000      -3.529      -2.724
[1-4]                           -3.0865      0.205    -15.022      0.000      -3.489      -2.684
[5-9]                           -3.1058      0.206    -15.112      0.000      -3.509      -2.703
[10-14]                         -3.0651      0.206    -14.892      0.000      -3.468      -2.662
[15-19]                         -3.0954      0.206    -15.053      0.000      -3.498      -2.692
[20-24]                         -3.0547      0.206    -14.852      0.000      -3.458      -2.652
[25-29]                         -3.0486      0.206    -14.829      0.000      -3.452      -2.646
[30-34]                         -3.0495      0.206    -14.827      0.000      -3.453      -2.646
[35-39]                         -3.0984      0.206    -15.066      0.000      -3.501      -2.695
[40-44]                         -3.0795      0.205    -14.995      0.000      -3.482      -2.677
[45-49]                         -3.0712      0.205    -14.947      0.000      -3.474      -2.668
[50-54]                         -3.0747      0.206    -14.947      0.000      -3.478      -2.671
[55-59]                         -3.0543      0.206    -14.845      0.000      -3.458      -2.651
[60-64]                         -3.0775      0.206    -14.964      0.000      -3.481      -2.674
[65-69]                         -3.1515      0.206    -15.318      0.000      -3.555      -2.748
[70-74]                         -3.0680      0.206    -14.923      0.000      -3.471      -2.665
[75-79]                         -3.0951      0.206    -15.054      0.000      -3.498      -2.692
[80-84]                         -3.1250      0.206    -15.188      0.000      -3.528      -2.722
[85+]                           -3.1095      0.205    -15.137      0.000      -3.512      -2.707
[All]                           -3.0542      0.207    -14.760      0.000      -3.460      -2.649
[Unknown]                       -3.0635      0.206    -14.895      0.000      -3.467      -2.660
==============================================================================
Omnibus:                     1998.306   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5064.741
```

Analysis

Linear Regression Model 2

We then conducted a second linear regression model to include the non-communicable diseases data sets, which expanded the total number of variables by 14 compared to the first model. Interestingly, the model also reflected a significantly low p-value below 0.05 for the additional 14 variables from the non-communicable diseases data set. Given the low p-value of these additional 14 variables, the model indicates that all variables are statistically significant. However, the linear regression model produced an undesirable low R-Squared value of 0.053 compared to the first model as seen in the screenshot below. Therefore, the team determined the non-communicable diseases data sets contained no factors that should be utilized for predicting life expectancy.

New Data Frame with Life Expectancy vs Non-Communicable Diseases:

	life expectancy	cardio_dis	cong_anom	dia_end	dig_dis	genit_dis	mal_neo	mus_skel_dis	neur_psy	oral_con	oth_neo	resp_dis	sen_org_dis	skin_dis	sids
0	72.6	1359.0	0.0	28.0	33.0	38.0	261.0	4.0	31.0	0.0	18.0	60.0	1.0	2.0	0.0
1	72.6	389.0	6.0	10.0	10.0	10.0	121.0	0.0	14.0	0.0	7.0	11.0	0.0	0.0	6.0
2	72.6	1214.0	0.0	16.0	39.0	26.0	170.0	3.0	40.0	0.0	18.0	66.0	0.0	0.0	0.0
3	72.6	7841.0	76.0	152.0	341.0	283.0	2465.0	26.0	404.0	1.0	199.0	454.0	9.0	4.0	76.0
4	72.6	33.0	64.0	0.0	5.0	2.0	3.0	3.0	13.0	0.0	1.0	14.0	1.0	0.0	64.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4613296	62.9	112.0	93.0	93.0	43.0	23.0	62.0	1.0	124.0	0.0	13.0	110.0	1.0	6.0	93.0
4613297	62.9	39.0	17.0	48.0	21.0	7.0	58.0	1.0	95.0	0.0	17.0	33.0	1.0	5.0	17.0
4613298	62.9	74.0	15.0	56.0	47.0	15.0	40.0	8.0	90.0	0.0	6.0	37.0	2.0	1.0	15.0
4613299	62.9	141.0	15.0	148.0	62.0	38.0	75.0	14.0	115.0	1.0	12.0	43.0	0.0	5.0	15.0
4613300	62.9	280.0	17.0	442.0	159.0	60.0	144.0	34.0	142.0	2.0	15.0	104.0	0.0	14.0	17.0

3996666 rows × 15 columns

Linear Regression Model 2 Results:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        life expectancy   R-squared:                       0.053
Model:                            OLS   Adj. R-squared:                  0.053
Method:                 Least Squares   F-statistic:                 1.200e+04
Date:                Sun, 26 Mar 2023   Prob (F-statistic):               0.00
Time:                        22:10:44   Log-Likelihood:            -8.6678e+06
No. Observations:             2797666   AIC:                         1.734e+07
Df Residuals:                 2797652   BIC:                         1.734e+07
Df Model:                          13
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           76.0167      0.003   2.31e+04      0.000      76.010      76.023
cardio_dis      -6.6e-05   6.21e-07  -106.349      0.000   -6.72e-05   -6.48e-05
cong_anom        0.0001   5.67e-06     21.099      0.000       0.000       0.000
dia_end       4.643e-05   4.25e-06     10.913      0.000    3.81e-05    5.48e-05
dig_dis          0.0002   6.28e-06     27.981      0.000       0.000       0.000
genit_dis    -4.531e-05   9.86e-06     -4.594      0.000   -6.46e-05    -2.6e-05
mal_neo       5.209e-05   1.23e-06     42.337      0.000    4.97e-05    5.45e-05
mus_skel_dis     0.0014   4.84e-05     29.881      0.000       0.001       0.002
neur_psy         0.0006   3.44e-06    184.521      0.000       0.001       0.001
oral_con        -0.2374      0.002   -126.525      0.000      -0.241      -0.234
oth_neo          0.0011   2.87e-05     40.101      0.000       0.001       0.001
resp_dis     -7.856e-05   4.61e-06    -17.036      0.000   -8.76e-05   -6.95e-05
sen_org_dis     -0.1667      0.002    -95.897      0.000      -0.170      -0.163
skin_dis        -0.0051    7.7e-05    -65.739      0.000      -0.005      -0.005
sids             0.0001   5.67e-06     21.099      0.000       0.000       0.000
==============================================================================
Omnibus:                   241760.975   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           533822.138
Skew:                          -0.553   Prob(JB):                         0.00
Kurtosis:                       4.832   Cond. No.                     1.11e+17
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 9.5e-20. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Analysis

Neural Network – MLP Regressor

The team also implemented the neural network model to conduct analysis on the life expectancy data. Specifically, we applied the Multi-Layer Perceptron (MLP) Regressor to predict life expectancy based on the following major factors with the most significant correlation and p-values per the linear regression model: GDP healthcare, adult mortality, schooling and income composition of resources. The data set was initially split 70/30 to create the training and test data sets. We then utilized the SciKit-Learn MLP Regressor model to perform the regression task. The following parameters were applied to the model: three hidden layers (100, 50 and 25 neurons) and the Rectified Linear Unit “reLU” activation. Below are two graphs of the results from the training and test data sets comparing the predicted to the actual life expectancy values based on the four factors. As seen in both graphs, the model accurately predicted life expectancy for both the training and test data sets.

Analysis

Neural Network – MLP Regressor *(continued)*

Given the extremely accurate results graphed for both the training and test data sets above, the model's accuracy and error metrics were as expected. The R-Squared value of the model was 0.9998, which means 99.98% of the variance in life expectancy is explained by our four factors. The Root Mean Square Error (RMSE) for the model was 0.0978, which calculated the minimal difference between the predicted and actual life expectancy values. The last metric we ran was the loss resulting from the model, which was 0.0045. Since the loss value is also significantly low, this tells us that the model optimally fits the training data set and maximized its regression performance. Therefore, the metrics overall tell us that the model can accurately predict life expectancy based on the four factors determined.

```
Number of inputs:  5
Number of outputs: 1
Number of layers:  5
Layer sizes: [(5, 100), (100, 50), (50, 25), (25, 1)]
Number of Iterations for Which Estimator Ran :  159
Number of Intercepts :  4
R_squared value:  0.9998772435677428
RMSE:  0.09783620173730562
Loss:  0.004467550119315629
```

However, our team believes the model is potentially reflecting overfitting issues given the significant accuracy and low error metrics. There is almost no error in the model, which seems unlikely for any model and given the complexity of the data set. While the neural network is known to be one of the more robust and sophisticated machine learning algorithms, the model can learn the training data set too well and overfit the training data set. While the team did not have the opportunity to address the overfitting issue, the team would complete the following next time to enhance the model: further exploring and cleaning up the data set structure, breaking down the data set into additional training and test data sets, and tuning the hyperparameters.

Recommendation

The analysis confirmed the difficulty of the undertaking. The topic is incredibly broad and impossible to account for all the potential influencing factors. Then to add insult to injury, we came to find our original Kaggle dataset has incorrect data (ex. France labeled as developing country). Therefore, obtaining the quantification of impact is unreliable but the team is confident on some of the relationships identified and the directionality of the outputs.

GDP_Healthcare, Adult Mortality (negative correlation), Income Composition of Resources (wealth classification), and Schooling were all highly correlated with life expectancy. Again, the dollar value of healthcare spending to life expectancy improvement is unreliable but the top 10 countries in gdp_healthcare spending is also in the top 25 of life expectancy. Juxtaposed with those insights is the possibility that death by illness is not a large enough contributing factor. The team's understanding of these results is that spending to mitigate illness/disease related deaths may not be the best allocation of resources in order to improve an entire country's average life expectancy.

Therefore, the team’s recommendation is to first fix or find a better dataset. Improved data will help quantify some of the impacts and narrow the significant inputs. This should improve the machine learning model and mitigate some of the overfitting issues too. Beyond that, the team recommends allocating funds to improving socioeconomic mobility, education, and access to healthcare than investment towards anyone department of healthcare. To our team the best “medicine” for life expectancy is opportunity. This is why public policy and the ecosystem around it is important. Improving opportunities and managing roadblocks to steer society is as much an art as a science.

Ultimately with an improved dataset, the regression model could help to narrow down variables and the neural network could help to predict the impact of spending on life expectancy. These models could help provide some more clarity on the subject, but the “truth” will always be elusive and complex. That is why any person/client seeking to answer these questions will have to be comfortable with a decent level of uncertainty, a willingness to iterate through implementation, and a continuous updating/understanding lifecycle.

References

Current health expenditure (% of GDP) Data (2022, January 30). *The World Bank*. Available at: https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS

Google Colaboratory (2023). *Google Colab*. Google. Available at: https://colab.research.google.com/drive/1Ttbv3bThFGCpBfa1JecHVRebpu4FSYRo#scrollTo=RiweulWYvrnF

Katari, K. (2020, October 15). Multiple Linear Regression Model using Python: Machine Learning, Medium. *Towards Data Science*. Available at: https://towardsdatascience.com/multiple-linear-regression-model-using-python-machine-learning-d00c78f1172a

Rajarshi, K. (2018, February 10). Life Expectancy (WHO). *Kaggle*. Available at: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Respiratory infections (2023). *World Health Organization*. Available at: https://platform.who.int/mortality/themes/theme-details/topics/topic-details/MDB/respiratory-infections

Sklearn Neural Network Regression Example – MLPRegressor (2018, June 13). *Vitalflux.* https://vitalflux.com/sklearn-neural-network-regression-example-mlpregressor/