Analysis & Prediction of Life-Expectancy

Abhijit Krishna Menon - Eda Aydin Oktay - Mansi Nagraj - Manya Raman

I. INTRODUCTION

Life expectancy (LE) is an important phenomenon that has important implications for society on various aspects. With the advent of new technologies and improvements in medical world our hope towards life expectancy is raising; American life expectancy will reach 79.8 years by 2040, compared to 78.7 years in 2016 [1]. On the other hand, some studies indicate lifespan may decrease due to detrimental effects of climate change, increased obesity as a result of processed food consumption or smoking and alcohol consumption[3].

Life expectancy has been studied with different focuses such as using different educational groups[8], health and social services in industrialized countries[4] or predicting survival time of cancer patients[6]. One of the well-known approaches is probabilistic projection of mortality which uses past age-specific death rates for each year in a country[5]. Another approach was a Bayesian hierarchical model for producing probabilistic forecasts of male period life expectancy at birth for all the countries of the world to 2100[7].

II. SUMMARY

In our work, we aim to design a machine learning model that helps analyze alternative scenarios to determine whether a country is more likely to have a worse/better Life Expectancy due to factors such as Accessibility, Economy, Environment, Education, and Health features. The project will look to enable the end user to then set goals on the various parameters and thus generate potential Life Expectancy goals.

We start with exploratory data analysis to uncover underlying trends between the predictors and Life Expectancy. We then used Linear Regression, ARIMA(Time-Series) Modelling, as well as a Neural Network model to perform predictions on our data-set to get potential life expectancy. We had two end goals, the first, was to understand how the various parameters acting as inputs affect the Life Expectancy of the country in question. The second goal was to perform an extended exploratory analysis of our data-set to be able to visualize the change in parameters as well as Life Expectancy across the years.

DATA-SET : For this project, we used a publicly available data set from https://databank.worldbank.org/data/home.aspx by the World Bank Group. The database contains 180 countries with around 50 indicator variables spanning over 1960 - 2016. We picked our own response variable as Life Expectancy and around 50 predictor variable from all the data that was available at the WorldBank data set.

III. METHODS

A. DATA CLEANING & TIDYING

We had a very comprehensive global data including information for most of the countries in the world with a large pool of variables. However, the biggest drawback was some of the countries were missing data for several years sometimes even for decades. For that reason, we started by filtering out the countries that lack data for more than 65% of the variables.

Secondly, we grouped our variables into 5 basic determinant categories; Education, Health, Accessibility, Environment, Economy and we kept the rest of the variables as Others. In the original data format variables were given in rows and repeated for each country, and the corresponding values were given in each column as year. We converted it to a format where each column is a variable and each row is a year and repeated for different countries. Lastly we created a subset for each variable category (e.g. Health data, Economy data etc.) to make the analysis more intuitive.

For data cleaning and tidying we took advantage of 'dplyr' and 'tidyr' libraries which make the data manipulation task very easy.

B. EXPLORATORY DATA ANALYSIS

Once the data was cleaned and transformed to a proper format, next step was to discover meaningful relations and summarize the main characteristics of the data set. For this purpose we performed a detailed Exploratory Data Analysis (EDA) on our data.

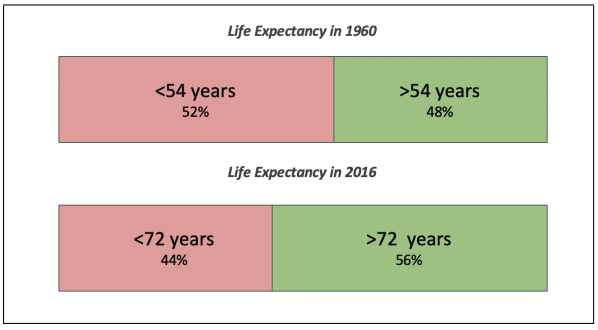

Fig. 1. Summary of Life Expectancy Change in 56 years

First we analyzed the average life expectancy in 1960, that was 54 years and the countries above average life

expectancy were 48%. But in 2016 the metric changed and the average life expectancy increased by 18 years ,i.e 72 years. But the interesting part we analyzed was that the countries above average life expectancy also increased, i.e 56%.

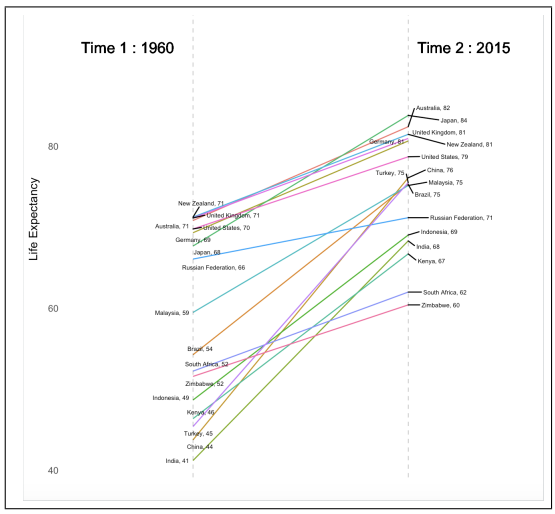

Fig. 2. Summary of Life Expectancy Change in 56 years

To understand this increase in percentage we did an in-depth analysis country-wise. Therefore, figure 2 and table I shows the change in life expectancy of 15 countries, from where we can see that countries like China, Turkey and Brazil which were below average in 1960 are above average in 2015, thus increasing the percentage of countries having above average life expectancy in 2015.

TABLE I
SUMMARY OF LIFE EXPECTANCY IN SELECTED COUNTRIES IN BETWEEN 1960 AND 2015

Countries	Initial(LE)	Final(LE)	Increase(LE)
Japan	68 Years	84 Years	16 Years
Australia	71 Years	82 Years	11 Years
United kingdom	71 Years	81 Years	10 Years
New-Zealand	71 Years	81 Years	10 Years
United States	70 Years	79 Years	9 Years
China*	44 Years	76 Years	32 Years
Turkey*	45 Years	75 Years	30 Years
Brazil*	54 Years	75 Years	21 Years
India	41 Years	68 Years	27 Years

We also inferred from the country-wise analysis of life expectancy that Japan′s has the highest life expectancy of any major country. The reasons for its longevity have been a source of debate, but due to genetic, social and lifestyle issues it is hard to find a single causal factor. Research suggests that nations high living standards, medical advances, diet and the universal and accessible health care system are some of the major contributors.

Secondly, U.S. had comparable life expectancy to major countries in 1980, however, in 2016 the LE has not increased with respect to other countries as it should. This anomaly can be attributed to a rising death rate among younger age groups. CDC research identified drug overdoses as a leading factor, with the age-adjusted rate of drug overdose deaths in 2016 (19.8 per 100,000) being 36% higher than the rate in 2014 (14.7).

Fig. 15 and Fig. 16 in Appendix shows snapshots of LE around the world in 1960 and 2016 respectively. It shows the increase in LE around the globe from 1960 to 2016 but also the increase continent-wise. From that it can be inferred that Africa as a continent has lowest LE in both the years.

To understand the reasoning behind it we analyzed continent-wise LE over the years as shown in Fig.18. The finding was that Asia has the highest GDP vs Life Expectancy growth rate among all the continents for the entire period and Africa shows lowest.

1) **Variable Selection:** After that we created correlation plots for each domain to find out the parameters that are significantly correlated with LE. One example correlation plot in Health domain can be seen from Figure 3 where we can see that Death Rate and Fertility Rate have highest correlation with Life Expectancy.

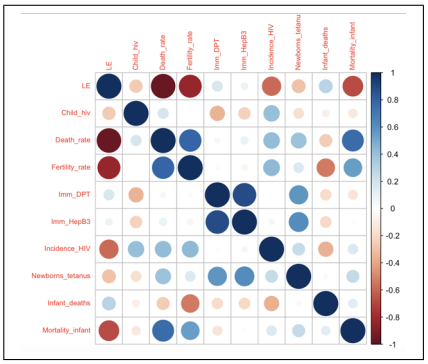

Fig. 3. Correlation of variables from Health data-set with Life Expectancy

In the second step of variable selection, we created a line plot for each potential predictor variable with a set of randomly selected countries. We aimed to find the relationship between LE and the potential predictor. An example observation can be seen from Figure 4.

When comparing the Fertility rate, Death rate and Mortality Rate with LE we observed strong correlations. However we did not include these parameters in the model assuming that Life Expectancy is inherently dependent on the mortality and fertility.

After finding the most significant parameters with the help of EDA we narrowed down our predictor variables. The list of most significant attributes and there correlation with Life Expectancy is given in the Table II below.

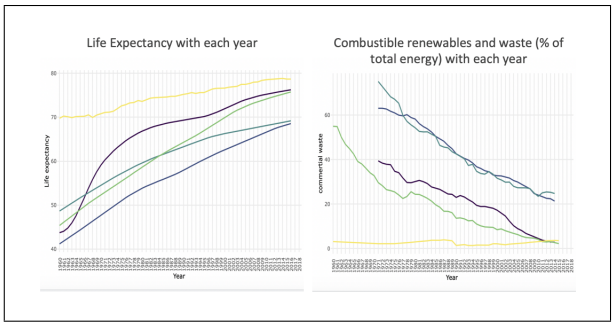

Fig. 4. Relationship between LE and Combustible Renewables and Waste

TABLE II
SIGNIFICANT VARIABLES AND CORRELATION WITH LE

Variables	Correlation	Domain
GDP per Capita	Positive	Economy
Food Deficiency	Negative	Economy
Death Rate	Negative	Health
Fertility Rate	Negative	Health
Incidence of HIV	Negative	Health
Electric Access	Positive	Accessibility
School Enrollment	Positive	Accessibility
Adjusted Education Savings	Positive	Education
Primary Education Expenditure	Negative	Education
CO2 Emission	Positive	Environment
Combustible Renewable & Waste	Negative	Environment

C. INTERACTIVE VISUALIZATION

In the last part of the project, we created an app for interactive visualization of our data-set. For this purpose we used Shiny package which provides elegant and powerful web framework to build applications with R.

The interactive interface gives option to select the data-set, minimum and maximum years to define the time frame, country of interest and lastly the x-axis and y-axis to plot the desired relation. As we have a time series data, animated visualizations are very effective in exhibiting the change across multiple time periods (which is year in or case). Figure 24 shows how the interface looks with 'scatter plot' tab.

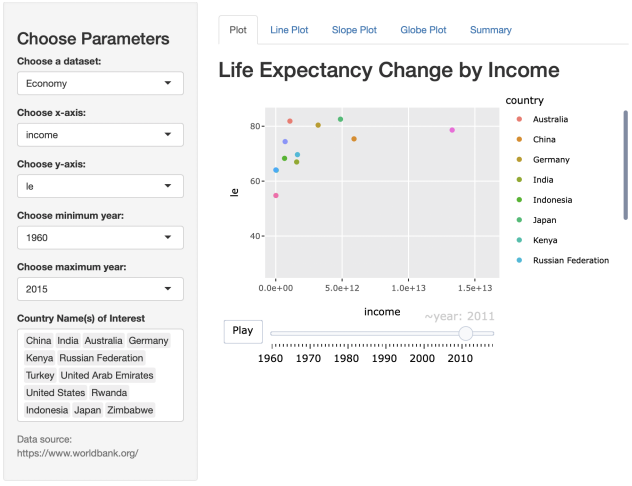

Fig. 5. Animated scatter plot tab in Shiny interface

Similarly with 'line plot' tab, users can visualize the change of single parameter with respect to time. This plot is helpful in understanding the general trend of a specific parameter.

Third tab in the interface is the 'slope plot' which is an ideal visualization of overall LE change between specific years for the selected countries. Similar to other plots the user can select the countries of interest and the time range.

Another animated visualization is 'globe plot'. With this tab, the user can see the global picture of LE change across the selected time range.

Lastly with the summary tab displays the top observations from the selected data-set.

D. CANDIDATE MODELS

1) **Linear Regression:** Regression analysis is a set of tools for building mathematical models that can be used to predict the value of one variable from another. Multiple linear regression analysis, allows the modeling of two or more independent variables to predict one dependent variable. A multiple linear regression model with k predictor variables X1, X2, ..., Xk and a response Y, can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...\beta_k x_k + \varepsilon \quad (1)$$

The are the residual terms of the model and the distribution assumption we place on the residuals allows us to infer from the remaining model parameters. The step wise regression consists of iteratively adding and removing predictors, in the predictive model, in order to find the subset of variables in the data set resulting in the best performing model, that is a model that lowers prediction error. Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.

2) **Time Series Forecasting Model:** A stationary time series is one whose statistical properties such as mean, variance, auto correlation, etc. are all constant over time. Most business, health and economic time series that we come across are not stationary in their original units of measurement and exhibit some trends, cycles and other non-stationary behavior.

Time series, which contain trend and seasonal patterns, are also non-stationary in nature. Thus from application view point ARMA models are inadequate to properly describe non-stationary time series, which are frequently encountered in practice. For this reason the ARIMA model is proposed, which is a generalization of an ARMA model to include the case of non-stationarity as well [2].

In ARIMA models a non-stationary time series is made stationary by applying finite difference of the data points. The mathematical formulation of the ARIMA (p,d,q) model using lag polynomials is given below :

$$\varphi(L)(1-L)^d y_t = \theta(L)\varepsilon_t \quad (2)$$

TABLE III
SIGNIFICANT ATTRIBUTE DESCRIPTIONS

Domain	Attribute	Description
Economy	income	Adjusted net national income (current US$)
	foodDef	Depth of the food deficit (kilo calories per person per day)
Access	Elec. access	Access to electricity (% of population)
	Gdp	Gross Domestic Product (current US$)
	Indv. using Net	Individuals using the Internet (% of population)
Education	Adj_Savings	Adjusted savings: education expenditure (current US$)
	prim_edu_expenditure	Expenditure on primary education (% of government expenditure on education)
Health	Fertility_rate	Fertility rate, total (births per woman)
	Incidence_HIV	Incidence of HIV (% of uninfected population ages 15-49)
	Death_rate	Death rate, crude (per 1,000 people)
	Fertility_rate	Fertility rate, total (births per woman)
Environment	co2emit	CO2 emissions (metric tons per capita)
	comwaste	Combustible renewable and waste (% of total energy)
	Fossil	Fossil fuel energy consumption (% of total)

$$(1-\sum_{i=1}^{p}\varphi_i L^i)(1-L)^d y^t = (1+\sum_{j=1}^{q}\theta_j L^j)\varepsilon_t$$

- Here, p, d and q are integers greater than or equal to zero and refer to the order of the auto regressive, integrated, and moving average parts of the model respectively
- The integer d controls the level of difference. Generally d=1 is enough in most cases. When d=0, then it reduces to an ARIMA(p,q) model
- An ARIMA(p,0,0) is nothing but the AR(p) model and ARIMA(0,0,q) is the MA(q) model [2]

3) **Neural Networks:** : Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. In the case of our data set, the neural network recognizes patterns between the input parameters and thus generates the output parameter by assigning weights. The basic formula that a neural network function upon is

$$y = f(x) \tag{3}$$

Where the 'y' can be generated from any relationship 'f' that the neural network assigns to the input parameters 'x'. A neural network consists of several layers. The layers are made of nodes. A node is just a place where computation happens, loosely patterned on a neuron in the human brain, which fires when it encounters sufficient stimuli.

A node combines input from the data with a set of coefficients, or weights, that either amplify or dampen that input, thereby assigning significance to inputs with regard to the task the algorithm is trying to learn. These input-weight products are summed and then the sum is passed through a nodes so-called activation function, to determine whether and to what extent that signal should progress further through the network to affect the ultimate outcome, say, an act of classification. If the signals passes through, the neuron has been 'activated'.

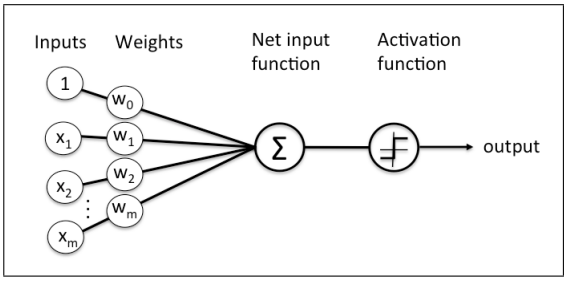

Fig. 6. Architecture of a Neural Network

Pairing the models adjustable weights with input features is how we assign significance to those features with regard to how the neural network classifies and clusters input.

For our data-set we use what is known as a feed-forward neural network. Since a neural network is born out of ignorance, the network learns with every passing iteration. The journey towards getting to the right weight to each of the input parameters are given by 3 pseudo mathematical equations.

1) A random weight is assigned to each input and a guess is generated.

*input * weight = guess*

2) The guess is compared with the output and the error is realized.

ground truth - guess = error

3) The error is then used to calculate the amount of adjustment in weight is needed to the initial values of each input.

*error * weight's contribution to error = adjustment*

The weights contribution to the error is calculated by the formula:

$$\frac{dError}{dWeight} = \frac{dError}{dActivation} + \frac{dActivation}{dWeight}$$

The activation function generates the kind of output the layer generates. For our model we have employed the 'RELU' function, or the 'Rectified Linear Unit' activation function.

Also, for calculating accuracy we have used the 'Mean Squared Error' function.

E. MODELLING

Linear Regression: We have created separate models for each of the five domains and analyzed the prediction and residual plots. For Accessibility domain, Electricity Access and Individuals using internet are the significant factors impacting life-expectancy as observed by the correlation plot and also in the Linear-Regression model with the least AIC.

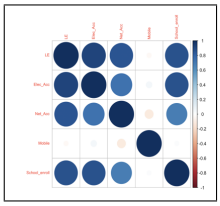

Fig. 7. Correlation plot for accessibility domain

```
Start:  AIC=820.61
LE ~ `Elec. access` + `Indv. using Net`

                    Df Sum of Sq     RSS     AIC
<none>                           4456.9  820.61
- `Indv. using Net`  1    2553.3  7010.2  955.84
- `Elec. access`     1    9007.2 13464.1 1153.60

Call:
lm(formula = LE ~ `Elec. access` + `Indv. using Net`, data = htrain)

Coefficients:
      (Intercept)     `Elec. access`  `Indv. using Net`
          47.1076             0.2472             0.1001
```

Fig. 8. Summary of Linear Regression Model

The prediction plot for LE with Electricity access is as shown below.

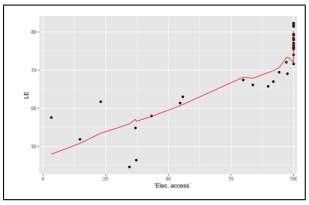

Fig. 9. Predictor plot for LE with Access to Electricity

Observation: We clearly observe that the predictions are not very accurate as the time-series nature of the observations was ignored while formulating the linear model. In order to improve the predictions, we proceeded with the time-series forecasting.

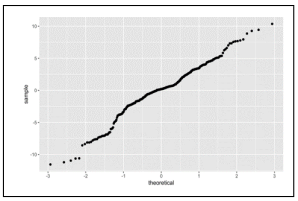

Fig. 10. Q-Q Plot of Residuals

Time Series Forecasting: The Life-Expectancy data from 1960 to 2015 showed a linearly increasing trend over the time but did not contain any seasonal or cyclic component. Also,

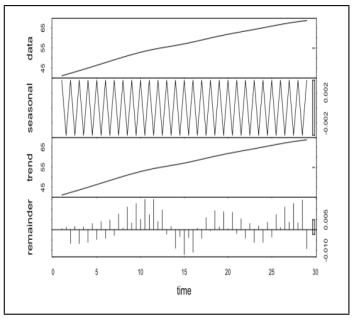

Fig. 11. Components of Time-Series Life-Expectancy Data

the graphs for auto correlation and Partial-auto-correlation function were gradually decreasing with time indicating the need for a combination of Auto-regressive (AR) and Moving-Average (MA) Model. In order to fulfill the requirement and handle the non-stationary data, we decided to proceed with ARIMA Model for modelling LE. The data was partitioned

```
time_Series_model <- auto.arima(trainLE, approximation=T, trace=TRUE,ic="aic", allowdrift=T)

predicted_values <- predict(time_Series_model,trainLE)
```

Fig. 12. Time series Forecasting (ARIMA) Model

into training and test set before modelling. The graph below shows the variation in LE from 1960-2015 and also predicts the values of LE for an upcoming decade for the country India.

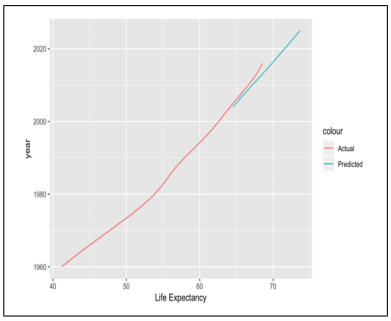

Fig. 13. Prediction plot for India's Life expectancy for upcoming decade

The histogram-plot for residuals shows that the residuals are centered around zero, which clearly indicates a good-model fit.

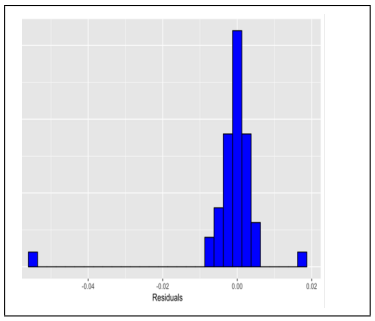

Fig. 14. Plot of Residues

Neural Networks: After performing the Linear Regression, and time series modelling using the Auto-Arima model we realized that those two methods either were not accurate enough or were not satisfying the objectives we set for ourselves at the beginning of the project. Our linear regression model did not take into consideration the temporal component of our time series data which provided us inaccurate results.

The Auto-Arima prediction model while giving us pretty accurate results, was not giving us the relationship between Life Expectancy and our predictor variables which is what we would have liked for our model to determine. Hence, we decided to move to a neural network model.

For the neural network model, we still had to work with our 5 different facets of our data, i.e. Accessibility, Education, Environment, Economy and Health as separate models. We had to go about the modelling this way because our data-set was not uniform in terms of data for each facet. Not all countries had data for all facets in the same year, hence in a combined data-set when we looked to remove rows with null values a lot of data was lost.

To begin modelling our data we first split our data set into a testing and training set. We then proceeded to scale our data-set which is basically to normalize all our values in between a certain range. This helps with the optimization of our algorithm.

We then proceeded to design our neural network. We built a feed forward neural network in a sequential format. The model had one hidden layer and the activation function for the hidden layer as well as the output was a Rectified Linear Unit. Our model works as a regression model that gives out the life expectancy on the basis of the input parameters. We then set the number of epochs(iterations) to 260.

```
neural_model = keras_model_sequential() %>%
  layer_dense(units = 64, activation = "relu", input_shape = ncol(x_train)) %>%
  layer_dense(units = 64, activation = "relu") %>%
  layer_dense(units = ncol(y_train))
```

Fig. 15. Neural Network Model

Our predictions from the Neural Network were fairly accurate.

We got a Mean Absolute Error of 2.4007 and an accuracy of 96%.

To generate accuracy, we used the following formula:

$$Accuracy(\%) = \frac{\sum \frac{|ActualValues - PredictedValues|}{ActualValues}}{N} * 100 \quad (4)$$

where N is number of predicted values.

IV. RESULTS

During our modelling phase of the project, we employed three different algorithms. Namely, Linear Regression, Time Series, and Neural Networks.

Of the three of them, the linear regression model gave us inaccurate results, because we were not considering the temporal component of our data.

In the ARIMA(time-series) model, while the results were promising, only a single parameter could be predicted while, on the basis of the time series component. That would mean we were not predicting our life expectancy on the basis of other parameters, but just on the pattern of the progress of Life-Expectancy itself, over the multiple years.

However, we needed the model to consider more than just one parameter. Thus, we decided to employ a neural network model that takes in all the parameters as inputs for the model facet wise and then generates a life- expectancy on the basis of the input parameters.

Our final product from the Neural network is a function. The function takes in new values of the input parameters from the user and then generates the Life Expectancy on the basis of those parameters.

```
```{r neural_net function}

# Insert values for -> co2 emisions,co2 fuel
consumptions,Combustible renewables and waste, fossil fuel
consumption , greenhouse gas emmissions

to_predict <- matrix(c(1.84900669,  27.84810, 10,
75.397463,7632.098), nrow = 1)

neural_function <- function(predict){
                                    to_predict <-
scale(predict, center = col_means_train, scale =
col_stddevs_train)
                                    new_test <-
neural_model_env %>% predict(to_predict)
                                    return(new_test)
                                    }


neural_function(to_predict)
```

          [,1]
[1,] 69.72309
```

Fig. 16. Predicted Life Expectancy for User Defined Values

While we faced with some unexpected issues mainly the objectives of the project were met.

A. *Forecasting*

Lastly, we perform a worldwide Life Expectancy forecasting with two alternative scenarios based on the change of Combustible Renewable and Waste (CRW). CRW comprises solid biomass, liquid biomass, biogas, industrial waste, and municipal waste, measured as a percentage of total energy use.

As can be seen from the Figure 17, with the pessimistic scenario, if we fit a model considering the last 55 years, we see that the CRW has increased approximately 100% in the last 55 years. With the optimistic scenario, if we fit a model considering the last 35 years, we see that the CRW had dropped approximately 60% in the last 35 years.

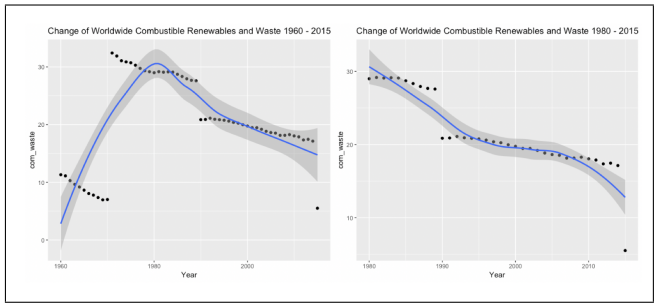

Fig. 17. Combustible Renewable and Waste with different time ranges

Assuming the same rate of changes in the future we generate two LE forecasting. The optimistic scenario forecasting results are illustrated in Figure 18. With this scenario LE is expected to increase to 78 by 2090s.

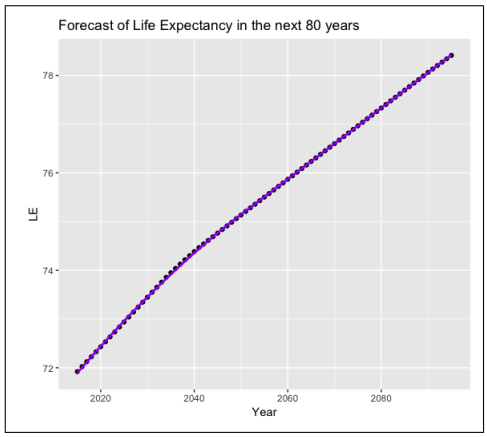

Fig. 18. Life Expectancy forecast with optimistic scenario

Similarly the pessimistic scenario forecasting results are illustrated in Figure 19. With this scenario LE is expected to increase to 74.5 by 2090s.

Although a potential increase in CRW negatively effects the LE, linear

V. DISCUSSION

A. *SHORTCOMINGS*

While the results from the neural network were accurate,there were certain issues with the model.Due to the

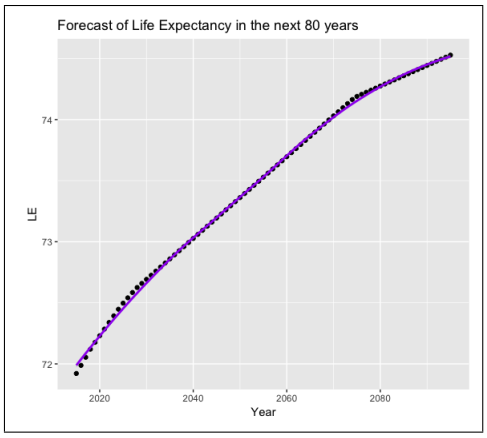

Fig. 19. Life Expectancy forecast with pessimistic scenario

presence of missing values for different time-periods in the data-set, separate models were generated for different domains such as health, education, environment etc. As we had to split our data into the various facets, the model could not account for all our parameters before assigning a weight. As a result, we observed some spurious outcomes such as increasing life expectancy with the increase in carbon emissions, as the model did not take into account other parameters like health, environment etc.

B. *FUTURE SCOPE*

- Once all the data is collected, a time-series based neural network can be developed across the whole data-set, as opposed to just having certain facets of inputs.
- A model can be built for each of the predictor variables to generate their values over the next decade.
  A graphical output can be generated from predicted values of variables and hence predicted life expectancy of the next decade.
  The model can then be used to set goals on certain parameters, by tweaking them and generating a required Life-Expectancy by government agencies.

VI. STATEMENT OF CONTRIBUTIONS

Each Team member had inputs and overlaps in each of the areas of our project. We worked together on the majority of the project (as a whole). However, each of us took responsibility of completing several specifics:

Abhijit Krishna Menon - Data Collection & Transformation, Research - Linear Modelling - Time Series forecasting - Neural-net modelling, Insights & Conclusion

Eda Aydin Oktay - Data Collection & Transformation, Exploratory Data Analysis, Research- Plotly - RShiny Dashboard, Insights & Conclusion

Mansi Nagraj - Data Collection & Transformation, Research- Linear Modelling - Time-series forecasting - Neural-net modelling, Insights & Conclusion

Manya Raman - Data Collection & Transformation, Exploratory Data Analysis, Research- Plotly - RShiny

VII. ACKNOWLEDGEMENT

To Professor Kylie Bemis, the Teaching Assistants and our classmates, who inspired us through their original work.

REFERENCES

[1] Institute for health metrics and evaluation, http://www.healthdata.org/news-release/how-healthy-will-we-be-2040.

[2] Ratnadip Adhikari and Ramesh K Agrawal. An introductory study on time series modeling and forecasting. *arXiv preprint arXiv:1302.6613*, 2013.

[3] Kyle J Foreman, Neal Marquez, Andrew Dolgert, Kai Fukutaki, Nancy Fullman, Madeline McGaughey, Martin A Pletcher, Amanda E Smith, Kendrick Tang, Chun-Wei Yuan, et al. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016–40 for 195 countries and territories. *The Lancet*, 392(10159):2052–2090, 2018.

[4] Vasilis Kontis, James E Bennett, Colin D Mathers, Guangquan Li, Kyle Foreman, and Majid Ezzati. Future life expectancy in 35 industrialised countries: projections with a bayesian model ensemble. *The Lancet*, 389(10076):1323–1335, 2017.

[5] Ronald D Lee and Lawrence R Carter. Modeling and forecasting us mortality. *Journal of the American statistical association*, 87(419):659–671, 1992.

[6] J Llobera, M Esteva, J Rifa, E Benito, J Terrasa, C Rojas, O Pons, G Catalan, and A Avella. Terminal cancer: duration and prediction of survival time. *European journal of cancer*, 36(16):2036–2043, 2000.

[7] Adrian E Raftery, Jennifer L Chunn, Patrick Gerland, and Hana Ševčíková. Bayesian probabilistic projections of life expectancy for all countries. *Demography*, 50(3):777–801, 2013.

[8] Pieter Van Baal, Frederik Peters, Johan Mackenbach, and Wilma Nusselder. Forecasting differences in life expectancy by education. *Population studies*, 70(2):201–216, 2016.

APPENDIX

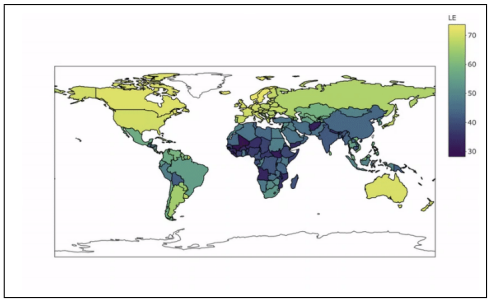

Fig. 20. Life Expectancy 1960

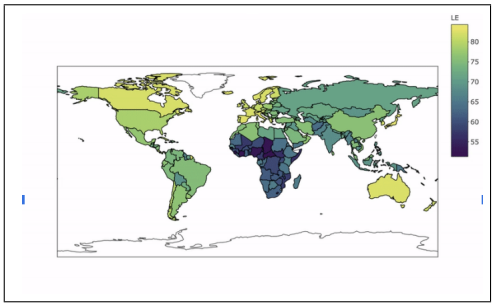

Fig. 21. Life Expectancy 2016

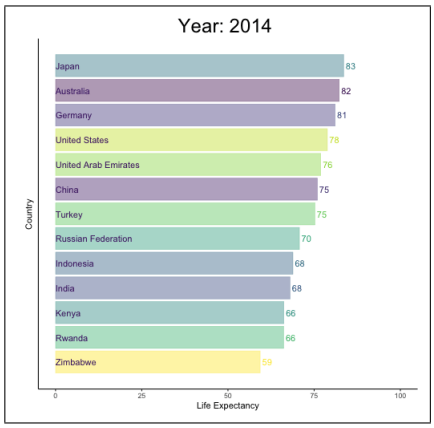

Fig. 22. Animation Chart of Life Expectancy Over Time for 15 Countries

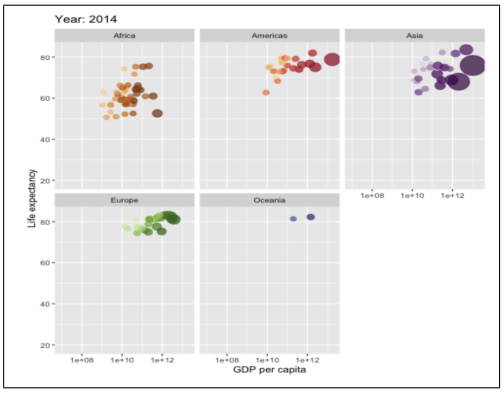

Fig. 23. GDP vs Life Expectancy Continent-Wise

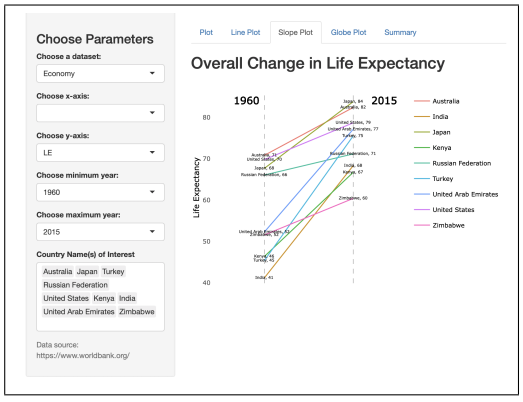

Fig. 24. Slope plot tab in Shiny interface