

Viktoriya Abakumova

DS4100 - Machine Learning and Data Mining 1

Spring 2018

Predicting Life Expectancy from Varied Country Statistics

Purpose and Goal

Life expectancy, as well as other statistics like maternal and child mortality rates, have not always had a clear correlation with how developed or urbanized a country is. For this project, I would like to predict life expectancy of a country in the world based on various statistics of that country. Specifically, I want to look at how unemployment, total population, population growth, poverty, health expenditures per GDP, the number of mobile users, the number of internet users, the ecosystem vitality, and the environmental health affect the life expectancy for a country.

The goal of this project is not so much geared towards using predictive modeling to predict life expectancy but using predictive modeling to see which of the features above make good predictors of life expectancy. These features could indicate that focusing on improving a specific category like health expenditures or decreasing cell phone use could potentially help elongate the life expectancy for a region.

Feature Selection and Data Sources

There was a lot of freedom in picking the descriptive variables for this particular topic. After seeing the variety of data that is available on a global scale, the chosen descriptive features are ones that could have a potential connection to life expectancy.

The data for unemployment, population statistics, health expenditures, and the number of mobile and internet users was taken from the CIA World Factbook which has data that is both updated to the most recent year available for that country and that accounts for a lot of the countries in the world. The mobile and internet user data was copy and pasted directly from the website as it was hard to parse the available download file.

The rest of the data was scraped through indexmundi, a website that displays the World Factbook data in a cleaner table the CIA website does.

The ecosystem vitality and environmental data was downloaded from 2014 Environmental Performance Index. According to the EPI, “environmental health measures the protection of human health from environmental harm” and “ecosystem vitality measures ecosystem protection and resource management”. These two categories include water and sanitation (access to drinking water, sanitation of water), air quality, health impacts (child mortality) data, and various other statistics about the climate such as wastewater treatment, co2 emissions, biodiversity, and protected areas. Instead of using just the EPI (Environmental Performance Index), both environmental health and ecosystem vitality are included as features to see if one affects life expectancy more than the other. Although including all of these features decreased the amount of countries/areas in the dataset, I considered it more important to include all of these features as they could have a significant impact on life expectancy.

Data Quality

A data quality report was first generated within R to look for missing values and to provide general statistics about each of the features in the dataset.

X <fctr>	non.missing <int>	missing <int>	missing.percent <int>	unique <int>	mean <dbl>	min <dbl>
unemployment	135	0	0	96	11.90	0.30
population	135	0	0	135	48061511.92	73897.00
population_growth	135	0	0	115	1.10	-1.08
poverty	135	0	0	114	27.39	0.20
health_expenditures	135	0	0	77	6.74	1.90
mobile_users	135	0	0	135	48544349.87	78444.00
internet_users	135	0	0	135	21104556.71	42552.00
environmental_health	135	0	0	133	63.92	27.42
ecosystem_vitality	135	0	0	134	40.84	5.82
life_expectancy	135	0	0	97	71.26	51.70

Table 1: A section of the generated data quality report that shows the number of missing values, as well as the mean and the min for each feature

The data quality report shows that there are no missing values in any of the features. This is because each feature for this dataset had no missing values in its original form.

The dataset was then analyzed for outliers. This dataset may have values that are far away from the mean. However, such data is kept in for a more accurate and realistic representation of real life statistics. Instead of checking for a number of standard deviations from the mean for outliers, it was checked whether there were illogical numbers in the dataset (i.e. negative total population, > 100 for % of gdp spent on health expenditures). For the population growth feature (has both positive and negative values), outliers were checked with the standard deviations from the mean method to ensure that there were no astronomical values. There were no illogical numbers in the features that are percentages. The population growth also contained no outliers.

Data Distributions and Transformations

For each feature, a histogram was built and looked at to see if it represented a normal distribution through a bell curve. Statistics like kurtosis and skewness were also calculated to account for the distribution. If such a bell curve was not seen and the statistics also pointed towards a skew, attempts were made to transform the data via transforms like the log, square root, etc. transforms.

For example,

```
[1] "st dev for unemployment 13.4140317860149"  
[1] "median for unemployment: 7.5"  
[1] "IQR for unemployment: 8.45"  
[1] "skewness for unemployment: 3.35006667077707"  
[1] "kurtosis for unemployment: 4.08770295442926"
```

Figure 1: Output of calculations for standard deviation, median, IQR, skewness, kurtosis for the unemployment feature

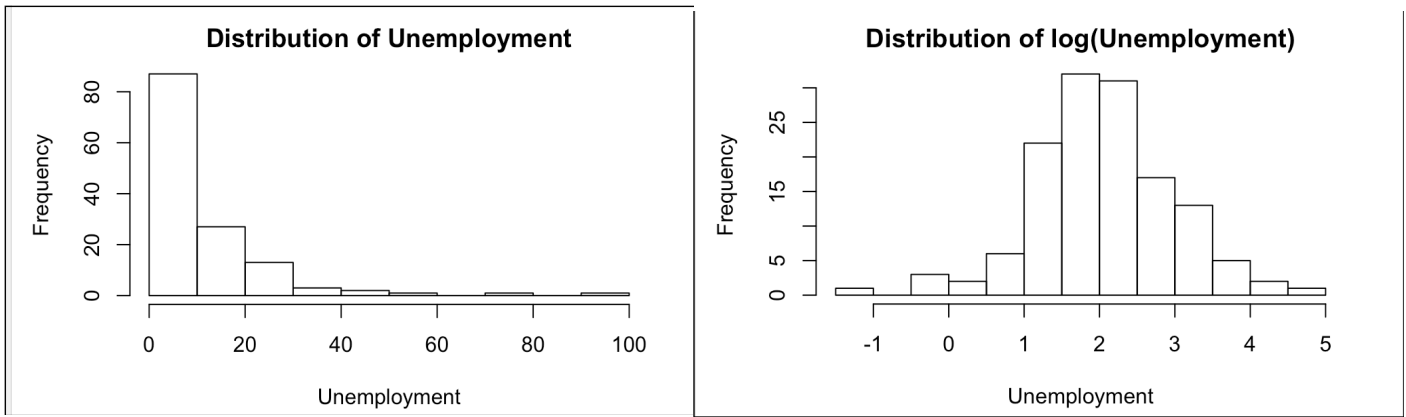


Figure 2: The before and after histograms for the unemployment feature, the original distribution on the left and the log transformed feature distribution on the right

There is a heavy positive/right skew for the unemployment feature as seen in the histogram and by the skewness calculation. The distribution has a kurtosis well above 3 and is therefore, leptokurtic. This is fixed with a log transform which makes the unemployment feature more normally distributed, less skewed, and lowers the kurtosis value, closer to 3.

Predictive Models

Prior to creating various predictive models, the dataset was split 50:50 into a training and testing set. The training set is composed of random sample of 50% of the dataset with the transformed features. The testing set contains the other half of the original dataset without any transformations.

Baseline Model

A baseline model was first created with the mean for the target feature. This model was created for comparison between a model that just uses the mean as its prediction to models that use different algorithm to accomplish the same goal.

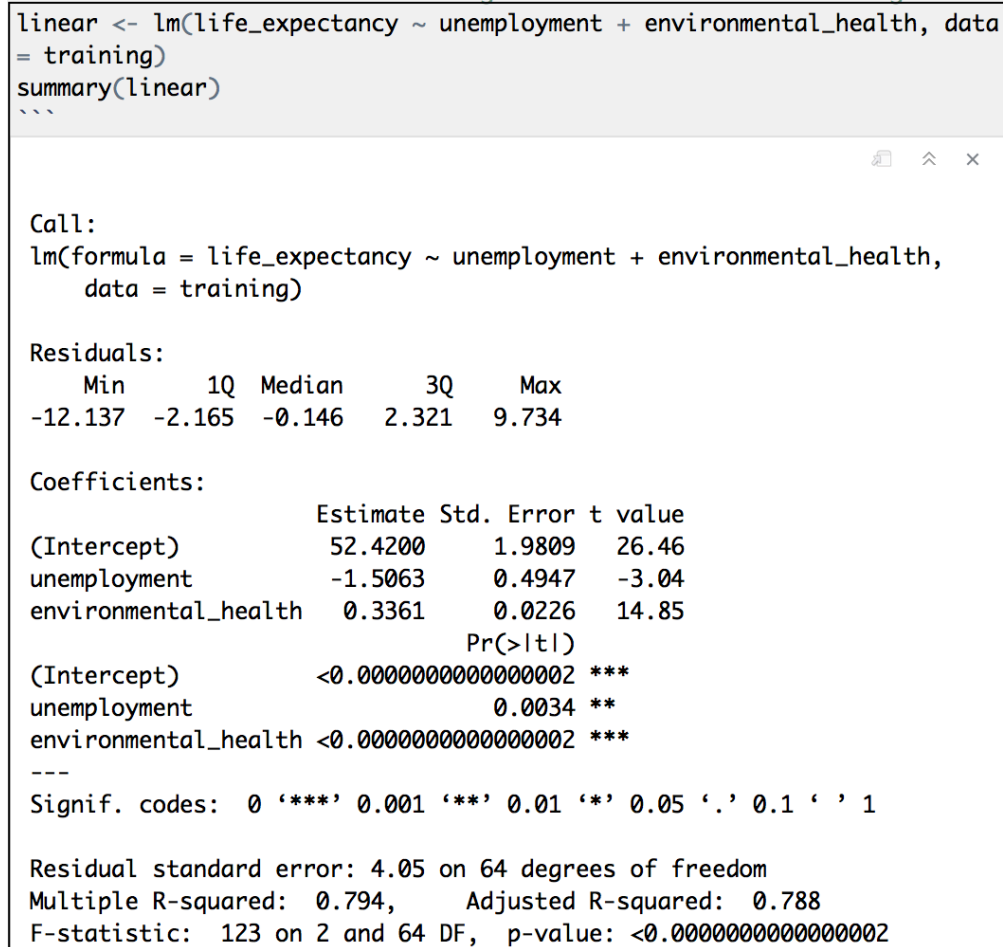
Model Type	RMSE	MAE
Baseline	7.09	5.99

Table 2: Calculated RMSE and MAE for the baseline model

Linear Regression Model

The first model to be created was a linear regression model because all of the features in this data are continuous and it is a classic model for continuous variables.

A multivariable linear regression model was created with the training set using the step function with backward fitting, which took away the feature with the highest p-value above 0.05 at each step until all of the feature p-values were above 0.05.

```
linear <- lm(life_expectancy ~ unemployment + environmental_health, data
= training)
summary(linear)
```  


Call:
lm(formula = life_expectancy ~ unemployment + environmental_health,
data = training)

Residuals:

Min	1Q	Median	3Q	Max
-12.137	-2.165	-0.146	2.321	9.734

Coefficients:

	Estimate	Std. Error	t value
(Intercept)	52.4200	1.9809	26.46
unemployment	-1.5063	0.4947	-3.04
environmental_health	0.3361	0.0226	14.85

Pr(>|t|)

(Intercept)	<0.0000000000000002	***
unemployment	0.0034	**
environmental_health	<0.0000000000000002	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.05 on 64 degrees of freedom
Multiple R-squared: 0.794, Adjusted R-squared: 0.788
F-statistic: 123 on 2 and 64 DF, p-value: <0.0000000000000002


```

*Figure 3: Generated linear regression model using backward fitting and its summary statistics*

| Model Type        | Accuracy | RMSE  | MAE   | Most Significant Feature(s)        |
|-------------------|----------|-------|-------|------------------------------------|
| Baseline          | -        | 7.09  | 5.99  | -                                  |
| Linear Regression | 55.6%    | 22.48 | 13.51 | Environmental Health, unemployment |

*Table 3: Calculated RMSE, MAE, prediction accuracy, and features that had the highest predictive strength for the baseline model and the linear regression model*

The linear regression model was statistically significant with p-values for both all of the selected features and the overall model below 0.05. The model also has an R squared above 0.7 which means that the model explains the variability of the data well. However, when tested on the testing dataset, the model did not perform as well. The prediction accuracy was only around 55%, which is deemed acceptable in some domains. The MAE and RMSE for this model were also higher than those for the baseline model. This means that the baseline model predicts better than the linear regression model. Because the R squared for this model is relatively high and the model is statistically significant but the error measurements indicate poor performance, it is possible that the model did overfit to the training data. This is also likely because the dataset is so small.

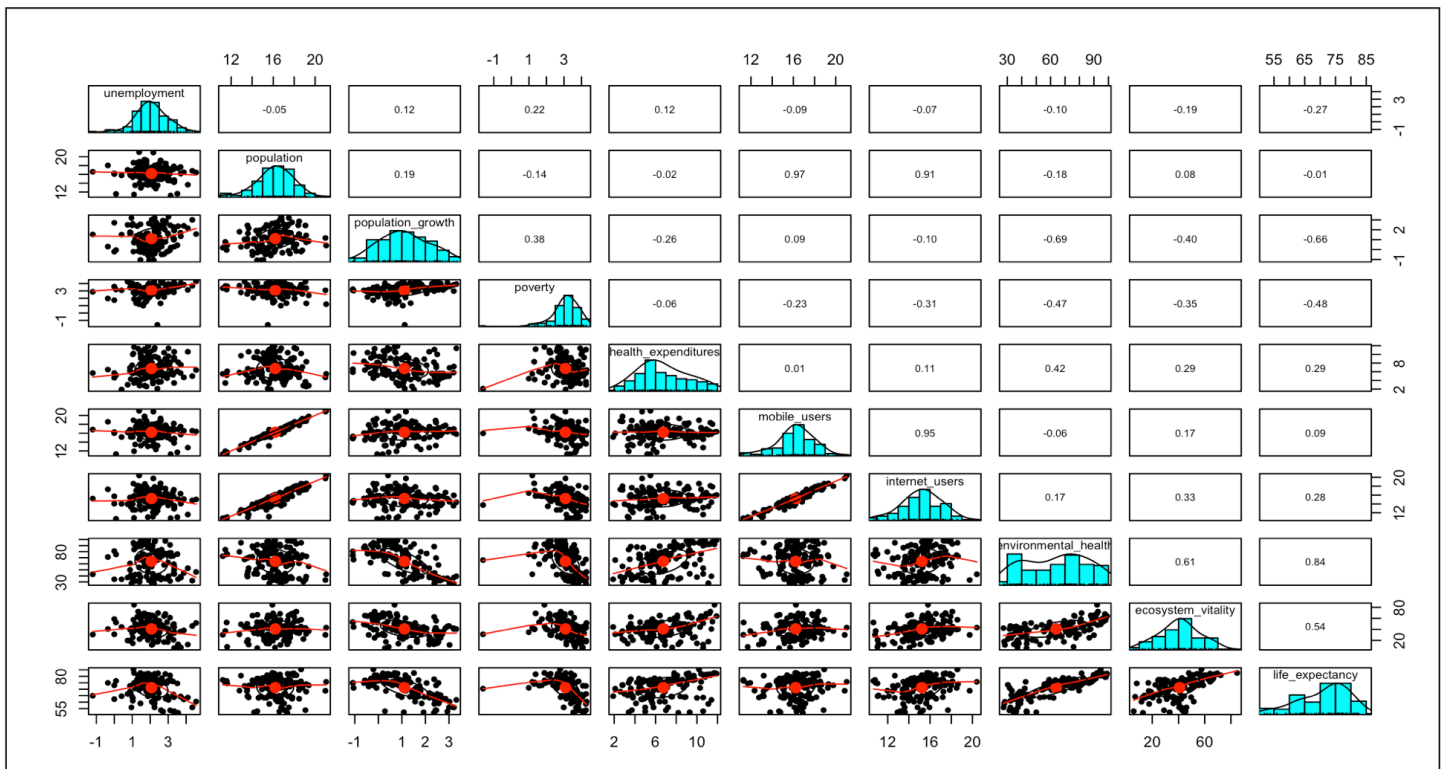


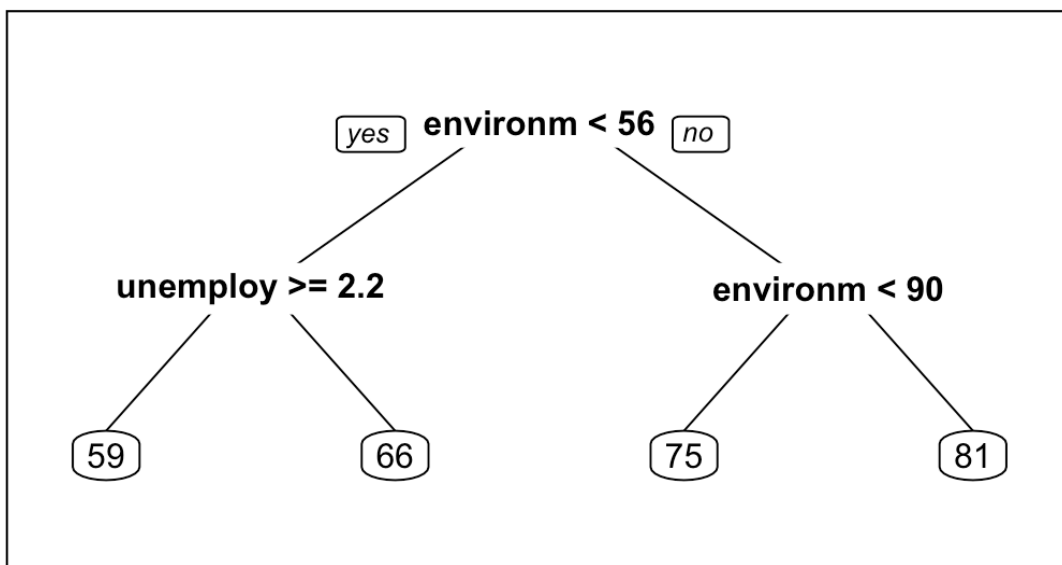
Figure 4: Scatterplot matrix with correlation for each pair of features in the dataset

The only statistically significant features that were selected for this model, out of 10 descriptive features, were unemployment (having a negative impact on life expectancy) and environmental health

(having a positive impact). Environmental health both had a strong positive correlation and a very linear relationship with life expectancy in the correlation and scatterplot matrix above. Therefore, it makes sense that it was selected for this model with a very small p-value as well. Unemployment did not display the same correlation. It was interesting that it did display enough significance to be retained in the model.

### Decision/Regression Tree Model

A regression tree model was also chosen as a predictive model for the data due to its ability to do feature selection which shows the importance of each variable. The variables that the tree does split on signify importance in predicting the target variable and the goal of the project was to find such variables.



*Figure 5: The resulting decision tree using the rpart function*



| Model Type        | Accuracy | RMSE  | MAE   | Most Significant Feature(s)        |
|-------------------|----------|-------|-------|------------------------------------|
| Baseline          | -        | 7.09  | 5.99  | -                                  |
| Linear Regression | 55.6%    | 22.48 | 13.51 | Environmental Health, unemployment |
| Decision Tree     | 78.3%    | 4.96  | 3.55  | Environmental Health, unemployment |

*Table 4: Calculated RMSE, MAE, prediction accuracy, and features that had the highest predictive strength for the baseline, linear regression, and decision tree models*

The decision tree was more successful than the baseline model and the linear regression models made above in predicting the target variable. The RMSE and the MAE were lower than the baseline and the linear regression models and its prediction accuracy was higher by over 20% than the linear regression model.

```
predict the test dataset using the pruned decision tree model
pred_tree_pruned <- predict(pruned_tree, test)

calculate the prediction accuracy, RMSE, MAE of the tree model
actuals_preds_tree_pruned <- cbind(data.frame(actuals =
test$life_expectancy, predicted = pred_tree_pruned))
correlation_accuracy_tree_pruned <- cor(actuals_preds_tree_pruned)
paste("Prediction Accuracy: ", correlation_accuracy_tree[1,2] * 100,
"% ", sep="")
rmse_tree_pruned <-
sqrt(mean((pred_tree_pruned-test$life_expectancy)^2))
mae_tree_pruned <- mean(abs(pred_tree_pruned-test$life_expectancy))
paste("RMSE:", rmse_tree_pruned)
paste("MAE:", mae_tree_pruned)
```


```
[1] "Prediction Accuracy: 78.3348632990714%"
[1] "RMSE: 4.96412335960695"
[1] "MAE: 3.54649982102374"
```


```

Figure 6: The code for pruning the above decision tree and the RMSE, MAE, and prediction accuracy for the pruned tree

The tree was also pruned to reduce overfitting and to remove the sections of the tree that provide little power in predictions. The pruned tree had the same structure as the regular regression tree, meaning that the optimal subtree is the original tree. This is confirmed by the error measures for the pruned tree being equivalent to those of the original tree.

The two variables that were selected to split on were unemployment and environmental health like they were in the linear regression model. This further confirms that unemployment and environmental health have a significant relationship with predicting life expectancy.

Random Forest Model

A random forest model was then built to even further improve the performance of the regression tree with more trees and improved accuracy as a result. Random forests give a good indicator of the importance of individual features as the quantity of trees also increases.

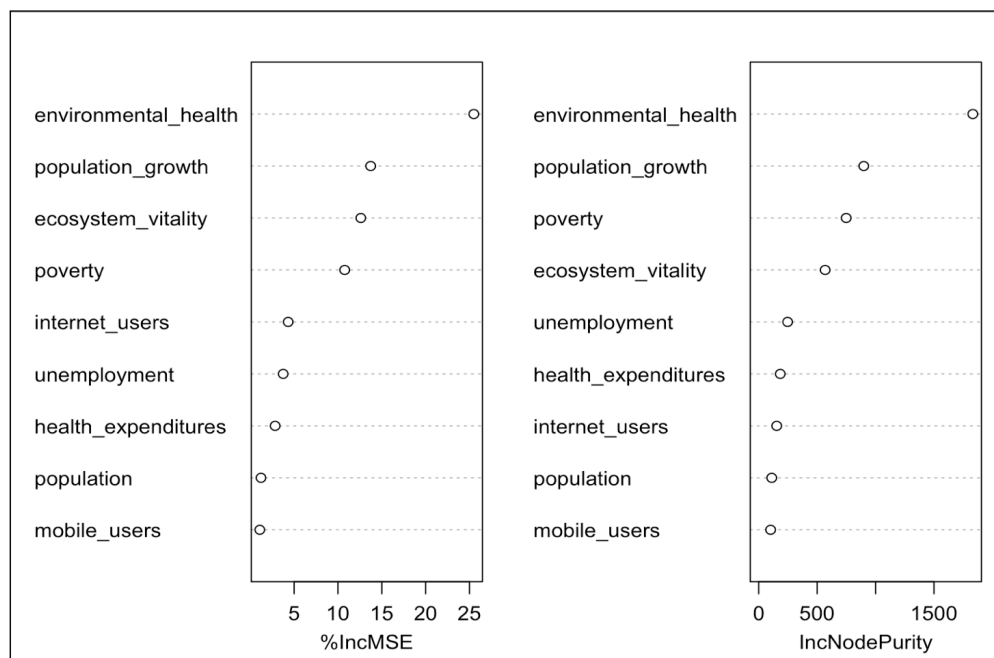


Figure 7: Importance graphs for the random forest, showing the increase in % MSE and Increase in Node Purity for each feature in the dataset

Feature	% Inc MSE	Inc Node Purity
Unemployment	3.76	247
Population	1.22	111
Population growth	13.73	899
Poverty	10.77	749
Health Expenditures	2.83	184
Mobile Users	1.09	101
Internet Users	4.32	153
Environmental Health	25.49	1833
Ecosystem Vitality	12.61	568

Table 5: Increase in % MSE and Increase in Node Purity measurements for comparison of the importance of each feature in the random forest model

From the plot and the importance table above, it can be seen that environmental health was the most important feature used by the random forest, followed by population growth, with unemployment not high on the light unlike in previous models.

Model Type	Accuracy	RMSE	MAE	Most Significant Feature(s)
Baseline	-	7.09	5.99	-
Linear Regression	55.6%	22.48	13.51	Environmental Health, unemployment
Decision Tree	78.3%	4.96	3.55	Environmental Health, unemployment
Random Forest	81.5%	5.54	4.85	Environmental health (highest % Inc MSE)

Table 6: Calculated RMSE, MAE, prediction accuracy, and features that had the highest predictive strength for the baseline, linear regression, decision tree, and random forest models

This model is also stronger than the other models with a higher rate of accuracy for predictions and a higher RMSE and MAE than both the baseline and linear regression model. However, the decision tree, although having a lower accuracy, has both a lower RMSE and MAE. Although both models do predict well, it was predicted that the random forest would perform better. This was predicted because random forests predict more accurately than decision trees. However, in this case, by a small margin of error, a decision tree predicted better than a random forest, perhaps due to the size of the data set.

Conclusion

Three different types of predictive models were built to predict life expectancy from poverty, health expenditures as a percent of GDP, total population, population growth, environmental health, ecosystem vitality, number of internet users, number of mobile users, and unemployment. Within all of these models, environmental health was consistently the strongest predictive variable and all of the models built had a decent predictive accuracy (the linear regression's predictive strength was subpar but still above 50%). In the decision tree and linear regression models, unemployment was a strong predictive variable while that was not the case in the random forest model. Therefore, from these

predictive models it can be concluded that environmental health, out of all of the other chosen features, definitely has the strongest effect on life expectancy and is a strong predictor.

For areas in which life expectancy is lower than desired, looking towards the environmental health would be a step towards improving life expectancy, as per the recommendation of these models. This makes sense because the environmental health measurement includes child mortality rate, air quality and pollution levels, access to water, and access to sanitation. All of these areas have a direct impact on the health and quality of life of a population. Therefore, it is not surprising that the combination of these areas would predict and impact the life expectancy in an area so strongly.

To further prove that environmental health or other features have an impact on life expectancy, different models could be built and with more data both in the training and testing set. This could be accomplished by looking at various regions on a country, which would result in a bigger aggregation of data. The problem that is presented with that approach is that figuring out the best predictive features from a large selection of features does not usually produce a large dataset. Not all countries actively track these statistics, which results in a sparsity of data, because these features are quite specific and it becomes difficult to make estimates and take surveys about them. However, the current models built still performed relatively well and made a good guess on which features are strongest in prediction. With an accuracy rate above 75% for two of the models, it's relatively safe to say that those models point to the right features that need to be focused on if a population wants to improve its life expectancy.

Data Sources

Central Intelligence Agency. (2018). The World Factbook. Retrieved from

<https://www.cia.gov/library/publications/the-world-factbook/>

IndexMundi. (2018). IndexMundi. Retrieved from <https://www.indexmundi.com/>

Yale University. (2014). Environmental Performance Index. Retrieved from

<https://www.indexmundi.com/>