6372: Project 1

Megan Ball, Matt Farrow, Neddy Nyatome

# Introduction

Using the World Health Organization (WHO) data compiled by Kumar Rajarshi, Deeksha Russell, and Duan Wang, we developed three different models:

* The first model was designed to be easily interpreted using linear regression.
* The second model was designed to be used as a predictive tool using linear regression.
* The third model was developed using non-parametric methods for prediction.

# Data Description

The description and context of the Life Expectancy (WHO) data set can be found [here](https://www.kaggle.com/kumarajarshi/life-expectancy-who). Data has been compiled from several different data sets into a final data set that represents health factors for 193 countries between the years of 2000-2015.

Looking at the data, there are 2,938 observations and 22 variables that cover four broad factors: immunization-related, mortality, economic, and social. various social, economic, and health-related factors. Each record in the data contains measurements for a single year within the country being measured.

# Exploratory Data Analysis

We began by plotting life expectancy into a histogram as well as a Q-Q plot.

# 

Figure 1 and 2. Histogram and Q-Q Plot of Life Expectancy data.

As we would hope, life expectancy tends to skew towards the older side. The Q-Q plot shows some slight deviations from normality towards the edges, but after trying various transformations, the deviations from normality that are evident in the distribution did not seem severe enough to warrant a transformation and we proceeded using the original data.

Next, we began looking at correlation to narrow down our variable list before examining specific relationships (Figure 1). Based on a cut-off of > 0.9 for correlation, we removed the variable in each correlation pair that had the higher number of NA values (Figure 2). We then proceeded to look at what happens when we also remove population, since it has minimal correlation to life\_expectancy (Figure 3). In the end, we made the decision to remove under\_five\_deaths, gdp, thinness\_1\_19\_years, and population due to lack of correlation to the response variable or collinearity.

Our next task was to address the missing values in the data set (Figure 4). We then limited the scope of our analysis to not include those countries where life expectancy was missing (Figure 5). In doing that, we excluded the following countries from our scope: Cook Islands, Dominica, Marshall Islands, Monaco, Nauru, Niue, Palau, Saint Kitts and Nevis, San Marino, and Tuvalu.

Hepatitis B was now our variable with the most missing values. In looking at the relationship between Hepatitis B and Life Expectancy (Figure 6), our options with regards to the missing values were to drop them, impute them, or fill them in with 0’s. We chose different approaches based on each model.

### Interpretable Model

For our interpretable model, we made the decision to drop the hepatitis\_b variable along with the remainder of the NA’s (Figure 7). As a result of our feature engineering, we were left with only 2 records for 2015. After several looks at the data, we decided to only use the observations from the most recent four years (2011-2014).

### Linear Prediction Model

Knowing that we cannot have missing values for Ridge Regression or LASSO models, we examined the relationship of each variable that had more than 100 missing values to see which appeared to be significant.

* Hepatitis B (Figure 8)
* Total Expenditure (Figure 9)
* Alcohol (Figure 10)
* Income Composition of Resources (Figure 11)
* Schooling (Figure 12)

After reviewing the plots, we made the decision to remove Hepatitis B, total expenditure, and alcohol since the trend for those three variables was relatively flat. We then removed the remainder of the missing values from the data set before proceeding to modeling.

# Objective 1:

## Restatement of Problem

Model Selection

Type of Selection

Checking Assumptions

Compare Competing Models

## Interpretation of Regression Coefficients

We are 95% confident that the model’s intercept is between (45.41, 48.689) and the true regression coefficient’s for the predicted variables are: adult mortality (-0.016, -0.01), total expenditure (0.183, 0.359), HIV/AIDS (-1.062, -0.772), and income composition of resources (34.999, 38.894).

# Objective 2:

## Strategy

Once we had developed a linear regression model, we then looked at developing a non-parametric model for predicting the life expectancy. Although our ASE and adjusted R-squared metrics were very good for the linear regression models, there could be some question about the normality of our residuals, independence of the errors, and equal variance. By using a non-parametric model, it is no longer required that we meet those assumptions and we can proceed with as many variables as we think would be useful in our model.

As discussed in our EDA section, we used the imputed data set to run our non-parametric models, which already removed highly correlated variables and ones we suspected were not related to our response variable. For the KNN model, we used the caret package to iterate through several different K values to find the optimal K value using all of the variables we had left. We also ran a KNN model using the four variables we determined were significant from our interpretable model for comparison. We proceeded to do the same with a random forest model and our results will be discussed in the Metrics section below.

## Data Sets

There are limits to the levels of a factor that can be used in a KNN or tree model. For this reason, country was removed from the data sets used in our non-parametric models. Additionally, the model performance was tested on both an imputed data set and a data set with the NA values were removed instead of imputed.

## Metrics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model Type | Data Set | # of Predictors | | Train ASE | Train R2 | Test ASE | Test R2 |
| KNN | Imputed | 16 | 7.437 | | 0.919 | 5.438 | 0.943 |
| KNN | Removed <NA> | 16 | 7.241 | | 0.920 | 6.175 | 0.930 |
| KNN | Imputed | 4 | 6.559 | | 0.927 | 5.302 | 0.943 |
| Tree | Imputed | 4\* | 13.7 | | - | 12.823 | - |
| Random Forest | Imputed | 16 | 3.606 | | 0.96 | 2.727 | 0.971 |
| Random Forest | Imputed | 4 | 5.855 | | 0.935 | 4.169 | 0.957 |

Table 1. Metrics for multiple non-parametric models. \*See appendix for variables used.

Based on the metrics, as seen above, the ASE and R-squared values from each model were surprisingly good. Since our R-squared was over 90% on our training data set, there was concern about over-fit, but we did not see a decrease in performance once applying to our test data set. The models were run again with the same parameters but with a different train and test split to perform an additional check against over-fit and performed similarly. This indicates that our model is not overfitting *specific to our current data set*. We could further verify whether the data overfits by running this model against a larger data set (for example, with years up to 2019 populated). The plots of our predicted output versus our actual values can be seen in figures x – z in the appendix.

## Comparison to Objective 1

Even with a non-parametric model, both KNN and random forest performed optimally when using the same four predictors as our interpretable model from the first objective. Additionally, the KNN model that used all predictors performed about the same with an imputed data set and a data set with the NA values removed. Overall, with small changes to the data set and predictors, all the nonparametric models performed similarly with ASE generally below 10 and R-squared above 90%. This is interesting to note in comparison to our purely predictive linear regression model which deemed country as an important predictor, and we were not able to use it in our non-parametric models. As noted in our analysis, we did get very high R-squared values and low ASE values which indicates there is likely some overfit in our models.

# Conclusion & Final Recommendations

# Appendix

Chart, timeline, treemap chart

Description automatically generated

Figure : Correlation Matrix, original data

Chart, timeline, treemap chart

Description automatically generated

Figure : Correlation Matrix: excluding under 5 deaths, gdp, and thinness 1-19 years

Chart, treemap chart

Description automatically generated

Figure : Correlation Matrix: exluding all from Figure 2 along with population

Table

Description automatically generated

Figure : Missing Values, original data

Table

Description automatically generated

Figure : Missing values, removed life expectancy

Chart

Description automatically generated

Figure : Relationship between Hepatitis B and Life Expectancy

Table

Description automatically generated

Figure : Missing Data, remove Hep B & remaining NA's

Chart

Description automatically generated

Figure : Relationship between Life Expectancy & Hepatitis B

Chart, scatter chart

Description automatically generated

Figure : Relationship between Life Expectancy & Total Expenditure

Chart, scatter chart

Description automatically generated

Figure : Relationship between Life Expectancy & Alcohol

Chart, scatter chart

Description automatically generated

Figure : Relationship between Life Expectancy & Income Composition of Resources

Chart, scatter chart

Description automatically generated

Figure : Relationship between Life Expectancy & Schooling

# Remove country because this tree function has a maximum of 32 levels  
tree1 <- tree(life\_expectancy ~ ., data = tree\_train[, -1])  
summary(tree1)

##   
## Regression tree:  
## tree(formula = life\_expectancy ~ ., data = tree\_train[, -1])  
## Variables actually used in tree construction:  
## [1] "hiv\_aids" "income\_composition\_of\_resources"  
## [3] "adult\_mortality" "infant\_deaths"   
## Number of terminal nodes: 9   
## Residual mean deviance: 13.7 = 29990 / 2189   
## Distribution of residuals:  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -16.51000 -2.11000 -0.05483 0.00000 2.19300 16.19000

Figure : Output chunk for tree model showing the four variables used in the tree construction

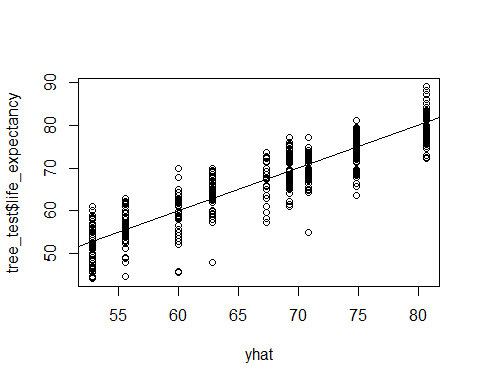


Figure : Model fit for our test data set using a tree regression

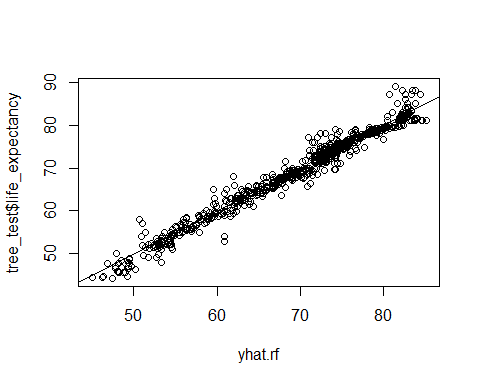


Figure : Model fit for our test data set using a random forest regression which used all predictors and 5 splits per node

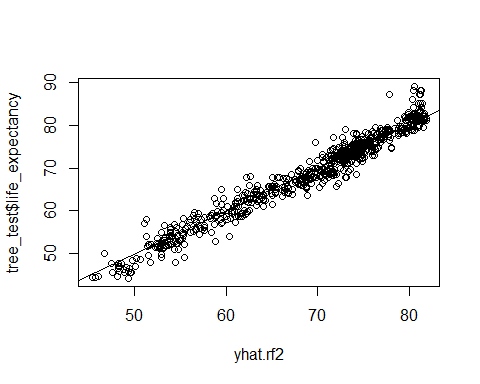


Figure : Model fit for our test data set using a random forest regression which used only the top four predictors and 1 split per node