STATS 415

DATA MINING PROJECT

An investigation into which features are most predictive of life expectancy.

Rahul Abraham, Parker Marcon, Paul Merica, & Elishua Shumpert

Stats 415
Data Mining Project 2019

Data

We discovered our dataset from Kaggle which is a website built to distribute datasets for data scientists. Our Kaggle data set on life expectancy was courtesy of the World Health Organization (WHO) statistics on life expectancy. Over the past 15 years, 2938 observations were collected throughout 190 countries. This study focused on many factors including immunization, mortality, economic, social and other health related factors. There are a total of 22 different variables that are included in this data set such as population, disease immunization coverage of polio, diphtheria, and measles, GDP per capita, BMI, schooling level, and many others. The response of interest in this data set is life expectancy which is measured by the number of years a person lives.

Among the data, there were a significant number of entries in our data set that contained missing data. Therefore, we used the feature engineering technique of omitting these observations from the analysis of the data. We ended up omitting 1289 observations from the dataset. Our main worry in performing this action was that the removed rows were going to be primarily from underdeveloped countries as they lacked funding to get values for all their metrics. This would skew our life expectancy estimates as we would only have numbers from developed countries primarily. When we analyzed the rows with missing values we found that 14.6% of the missing rows came from developed countries and 85.6% of the rows with missing values came from developing countries. This was close to the averages we found through our table before this filter (in the full dataset developing was 83% of the observations and developed was 17%), so it seems that the missing values were distributed relatively randomly, and did not have a disproportionate effect on developing countries.

Another challenge we faced was that there was a lot of collinearity among the predictors in the data set within the design matrix. As a result, it was decided that we omit some columns of the data set that were related to each other including under-five deaths (the number of five deaths per 1000 persons) as it was collinear with infant deaths and thinness 1-19 years (prevalence of thinness among children and adolescents for ages 10 to 19 (%)) because it was highly correlated with thinness 5-9 years (prevalence of thinness among children for ages 5 to 9 (%)). Likewise, the country, year, and country status predictors were removed from the data set because these columns were not useful in the analysis as we were looking for indicators of life expectancy based on health and wealth factors and not country origin or year.

For all models, we decided to split 70% of the data into the training set and 30% into the testing set to test the accuracy of every model we used in this report.

Overview

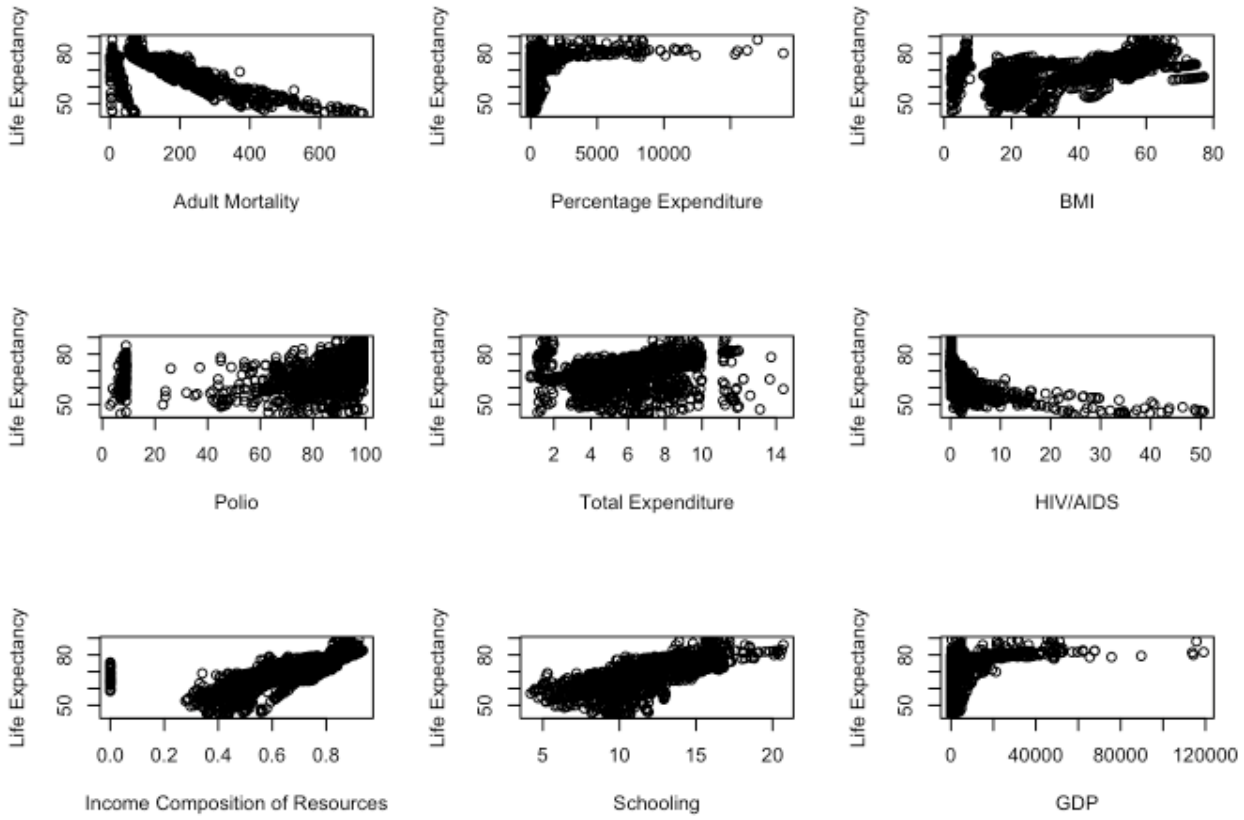
The goal of this study is to determine what predictors were important in predicting life expectancy and what model bests predicts life expectancy. We wanted to see what health and socioeconomic indicators might help to predict whether the life expectancy of a country might be high or low. For example, does a low HIV/AIDS rating and a large number of years the individual attended school mean a country is going to have a longer life expectancy for their citizens?

Stats 415
Data Mining Project 2019

Using a multitude of different methods we were able to select 5 variables that we believe most accurately predict life expectancy as well as optimize model complexity and goodness of fit. We also were able to select a model which we thought best fit the data and predicted life expectancy well. Before explaining the methods and models in more detail, we will first review the exploratory data analysis and diagnostic plots.

Exploratory Data Analysis and Feature Selection

Figure 1. *Scatter plots of Life Expectancy vs. other predictors*

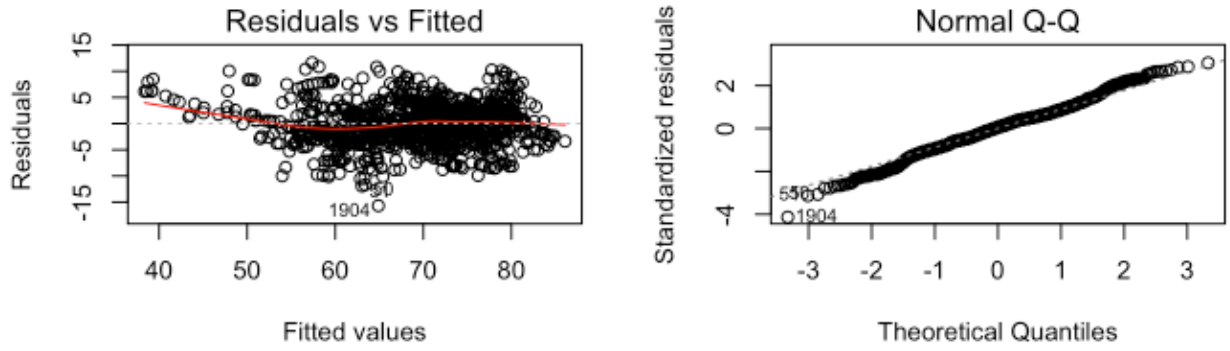


When plotting all the predictors against life expectancy in a scatter plot, life expectancy seems to be most correlated with adult mortality, alcohol consumption, BMI, diphtheria immunization coverage in percent, deaths from HIV/AIDS, income composition of resources, and years of schooling. In most of the scatterplots, there appears to be a lot of outliers. When looking at the Pearson's correlation coefficients for the relationship between these 8 predictors and life expectancy, the five most strongly predictive features of life expectancy are adult mortality, BMI, deaths from HIV/AIDS, income composition of resources, and years of schooling with years of schooling being the top-most correlated feature with life expectancy. When looking at the plot of life expectancy vs. percentage expenditure on health, it is clear that increases in healthcare expenditure help immensely at low levels of GDP in increasing life expectancy, but that at higher amounts it has diminishing returns and eventually does not help at all. The same pattern occurs for GDP per capita, suggesting that this also has diminishing

returns. These plots also suggest that there may not be a linear relationship between percentage expenditure on health and life expectancy and between GDP per capita and life expectancy because there is a logarithmic curve in the data for these predictors.

Another way to look at which predictors were most predictive of life expectancy was to look at the magnitude of each feature coefficient in the models. The higher the absolute value of the coefficient, the more important that variable is in determining life expectancy. When fitting the full model, it was observed that income composition of resources, years of schooling, deaths from HIV/AIDS, alcohol consumption, and total government expenditure on health had the highest coefficient estimates in magnitude. Lastly, the one step greedy algorithm of backward selection was used to choose a subset of predictors that would be useful in predicting life expectancy. AIC was used as the criteria of selecting the predictors, and backward selection chose 10 out of the 16 predictors in the data set which were adult mortality, the number of infant deaths, alcohol consumption, percentage expenditure on health, BMI, total government expenditure on health, diphtheria immunization coverage (%), HIV/AIDS, income composition of resources and years of schooling.

Figure 2. *Diagnostic Plots*



In addition to this analysis, we did a residual analysis to determine normality of the data. According to the diagnostic plots shown above, the data appears to be approximately normal from the normal Q-Q plot which would suggest that a linear fit for the model is appropriate, however, when looking at the residuals vs. fitted plot the residuals appear to be centered around zero, but there appears to be a little bit of curvature in the plot of the residuals which may be suspect for linearity of the data for this model. So, we decided to look at other methods that might fit the data better such as exponential regression and K-Nearest Neighbors (KNN).

Figure 3. *Correlation Matrix*

Above is a plot of the correlation matrix after the collinear columns were excluded from the data set. From the correlation matrix heatmap, the red represents higher and positive correlation while the blue represents negative correlation between predictors. This graphic also shows that there are more lighter colors than there are darker colors which suggest that there is not much collinearity between predictors in the design matrix.

1. Linear Regression

Linear regression was chosen because the response variable, life expectancy, that is being regressed upon, is a continuous variable. With linear regression, conclusions can be drawn on if there are relationships between the predictors and the response and how accurate we can predict life expectancy from a given level of predictors and evaluate the accuracy of that. As noted before, the diagnostic plots shown in figure 1 suggested that a linear model might be appropriate. The goal was to see what the significant predictors were in predicting life expectancy. We fit a linear regression model with the five most correlated predictors with life expectancy which were noted as adult mortality, BMI, deaths from HIV/AIDS, income composition of resources, and years of schooling. For this fitted model, all these predictors were significant in predicting life expectancy and had a great model fit according to its R^2 value of 0.8165. The test MSE for this model was 15.05 Then, a linear model was fitted using the predictors that were chosen by backward selection. All the predictors from this model were significant and also did well in predicting life expectancy with an R^2 value of 0.8324. The test MSE for this model was 15.47.

2. Exponential Regression

Next, we considered exponential regression to try and model life expectancy. This decision was made due to the fact that life expectancy is measured as a duration of time, and exponential models are appropriate for data that varies with time. We also found non-linear relationships in

the scatterplots specifically between percentage expenditure on health and life expectancy and between GDP per capita and life expectancy. An exponential regression model was fitted on two models: one predicting life expectancy from the five most correlated predictors and the other from the predictors chosen by backward selection. In both of these fitted models, all the predictors were significant with the model predicting life expectancy from the five most correlated predictors having an R^2 value of 0.8268 and the model predicting life expectancy from the predictors chosen by backward selection having an R^2 value of 0.8394. These models had a test MSE of 13.23 and 13.82 respectively. Therefore, it appears that exponential regression does slightly better than linear regression in predicting life expectancy, however, they fit the data almost the same.

3. K-Nearest Neighbors (KNN)

The KNN regression method was selected because we were curious as to whether this nonparametric method would outperform linear regression in terms of prediction accuracy of life expectancy since we now know that the underlying data might not have a linear relationship with life expectancy. In other words, we wanted to verify if the true relationship between the predictors of life expectancy were linear with life expectancy. Before fitting the KNN regression, the mean and standard deviation of the columns were checked, they varied for each predictor, so we standardized the columns to ensure that every predictor was on the same scale. We fitted the KNN regression on the same two models that we have been using with the five most correlated variables and the variables chosen by backward selection. The k value that minimizes the test MSE for the model regressed upon the five most correlated variables was 25 and was 7 for the model regressed upon the predictors chosen by backward selection. These models resulted in test error rates of 7.59 and 11.58 respectively. Compared to linear regression, these error rates are much lower and would suggest that the relationship between some features of life expectancy and life expectancy itself may be nonlinear.

4. Lasso

Figure 4. *Lasso Coefficients*

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                                    1
## (Intercept)            69.69436742
## Adult_Mortality        -2.13962342
## Infant_Deaths          -0.36684947
## Alcohol                -0.39340602
## Percentage_Expenditure  0.72292228
## HepatitisB             -0.23332672
## Measles                 0.19140804
## BMI                     0.60417121
## Polio                   0.13605709
## Total_Expenditure       0.45087118
## Diphtheria              0.68540427
## HIV.AIDS               -2.35780035
## GDP                     0.10156867
## Population              0.09073267
## thinness.5.9.years     -0.15673308
## Resource_Income         2.05358812
## Schooling               2.54216039
```

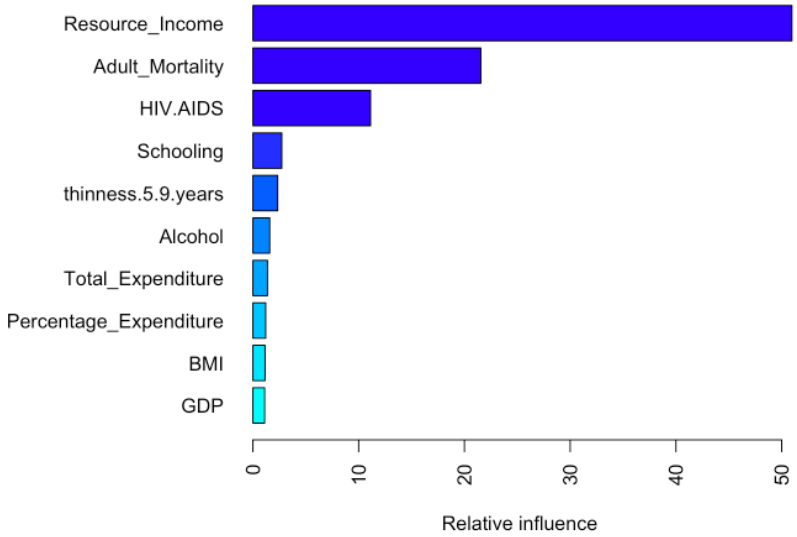
```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                                    1
## (Intercept)            69.69436742
## Adult_Mortality        -2.19091847
## Infant_Deaths           .
## Alcohol                 .
## Percentage_Expenditure  0.31789845
## HepatitisB              .
## Measles                 .
## BMI                     0.45371997
## Polio                   .
## Total_Expenditure       0.10662123
## Diphtheria              0.35272086
## HIV.AIDS               -1.96755712
## GDP                     0.13130218
## Population              .
## thinness.5.9.years     -0.02894945
## Resource_Income         1.96361800
## Schooling               2.53315167
```

Lasso is included in this data analysis because the goal of lasso regression is to obtain a subset of predictors that minimize prediction error for a response by imposing some constraints on the model parameters to cause some regression coefficients to shrink to zero. Variables with a regression coefficient shrunk to zero after the shrinkage process are excluded from the model and those with non-zero coefficients are most strongly associated with the response variable. This is why we decided to use this method to verify our choice of predictors for the two models we ran. Lasso also leads to a sparse model which is optimal for model complexity and interpretability. First, a lasso model was fitted on all the predictors and the coefficients were recovered by using the tuning parameter that gave the lowest MSE. This tuning parameter was 0.01 which was significantly small as a constraint allowing for a lot of flexibility in the selection of variables, therefore, lasso chose all the predictors of life expectancy (as shown in the left hand panel) in the data set which suggest that all the variables in the data set were greatly associated with life expectancy. Secondly, it was decided to recover the predictor coefficients that were chosen by lasso with a higher tuning parameter, specifically the higher value of lambda that gives an MSE within one standard error of the smallest to return a more-sparse model that would avoid model complexity and potential for overfitting. In the right-hand panel, 10 out of the 16 predictors were selected by lasso as useful predictors of life expectancy. The test MSE associated with lasso regression was 13.77.

5. Random Forest for Variable Importance

The last method that was used to determine which variables are important in predicting life expectancy is the random forest algorithm. While there is no interpretation in the random forest algorithm, the variables can be ranked by importance which is essential to answering the research question we posed. Random forest decorrelates the trees in tree-based methods. The random selection of variables within the random forest algorithm controls overfitting and reduces the variance of individual decision trees.

Figure 5. *Variable Importance Plot*



The variable importance is computed using the total amount that RSS is decreased due to the splits over a given predictor averaged over 5000 trees. We can see from this plot that the variables with the largest decrease in total RSS is income composition of resources, adult mortality, and HIV/AIDS. Additionally, our choices of predictors chosen to be useful in predicting life expectancy are verified with this plot. The five most correlated variables chosen are in the top five of the variable importance plot except for BMI. The test MSE associated with the random forest method is 7.27.

Conclusion

Figure 6. *Results*

	Test.MSE
Linear Regression Reduced Model	15.051400
Linear Regression Backward Selection Model	15.470319
Exponential Regression Reduced Model	13.229447
Exponential Regression Backward Selection Model	13.819892
KNN Reduced Model	7.585243
KNN Backward Selection Model	11.581118
Lasso	13.771425
Random Forest	7.266764

Overall, we found that five variables were essential in determining life expectancy because they continued to show up in the results throughout this analysis in the different models and algorithms we ran. These five variables are body mass index (BMI), adult mortality, income composition of resources, years of schooling, and HIV/AIDS. In figure 6, we refer to the reduced model as the model with the five most correlated predictors. When looking at the test error rates above, the random forest algorithm performed the best as it has the lowest test error rate. As noted before, KNN out performs the linear regression suggesting that there may be nonlinearity between life expectancy and its predictors. Therefore, it is more appropriate to use KNN for this data set rather than linear or exponential regression. Our hardest decision came from whether we should use KNN or random forest method for regression as they have similar testing errors. We decided that random forest was the best model to use for this data as it had the lower testing error, and is also useful in predicting which variables are important. The variable selection feature is useful because it could be used to help countries decide which areas should be given importance in order to efficiently improve the life expectancy of its population.